11 Experimental Design in Clinical Trials

The design and analysis of clinical trials is fertile soil for statistical applications. The use of sound statistical principles in this area is particularly important because of close FDA involvement, in addition to crucial public health issues that are consequences of actions based on the outcomes of clinical experiments. Principles and procedures of experimental design, particularly as applied to clinical studies, are presented. Relatively few different experimental designs are predominantly used in controlled clinical studies. In this chapter, we discuss several of these important designs and their applications.

11.1 INTRODUCTION

Both pharmaceutical manufacturers and FDA personnel have had considerable input in constructing guidelines and recommendations for good clinical protocol design and data analysis. In particular, the FDA has published a series of guidelines for the clinical evaluation of a variety of classes of drugs. Those persons involved in clinical studies have been exposed to the constant reminder of the importance of design in these studies. Clinical studies must be carefully devised and documented to meet the clinical objectives. Clinical studies are very expensive indeed, and before embarking, an all-out effort should be made to ensure that the study is on a sound footing. Clinical studies designed to "prove" or demonstrate efficacy and/or safety for FDA approval should be controlled studies, as far as is possible. A controlled study is one in which an adequate control group is present (placebo or active control), and in which measures are taken to avoid bias. The following excerpts from *General Considerations for the Clinical Evaluation of Drugs* show the FDA's concern for good experimental design and statistical procedures in clinical trials [1]:

1. Statistical expertise is helpful in the planning, design, execution, and analysis of clinical investigations and clinical pharmacology in order to ensure the validity of estimates of safety and efficacy obtained from these studies.
2. It is the objective of clinical studies to draw inferences about drug responses in well-defined target populations. Therefore, study protocols should specify the target population, how patients or volunteers are to be selected, their assignment to the treatment regimens, specific conditions under which the trial is to be conducted, and the procedures used to obtain estimates of the important clinical parameters.
3. Good planning usually results in questions being asked that permit direct inferences. Since studies are frequently designed to answer more than one question, it is useful in the planning phase to consider listing of the questions to be answered in order of priority.

The following are general principles that should be considered in the conduct of clinical trials:

1. Clearly state the objective(s).
2. Document the procedure used for randomization.
3. Include a suitable number of patients (subjects) according to statistical principles (see chap. 6).
4. Include concurrently studied comparison (control) groups.
5. Use appropriate blinding techniques to avoid patient and physician bias.
6. Use objective measurements when possible.
7. Define the response variable.
8. Describe and document the statistical methods used for data analysis.

11.2 SOME PRINCIPLES OF EXPERIMENTAL DESIGN AND ANALYSIS

Although many kinds of ingenious and complex statistical designs have been used in clinical studies, many experts feel that *simplicity* is the key in clinical study design. The implementation of clinical studies is extremely difficult. No matter how well designed or how well intentioned, clinical studies are particularly susceptible to Murphy's law: "If something can go wrong, it will!" Careful attention to protocol procedures and symmetry in design (e.g., equal number of patients per treatment group) often is negated as the study proceeds, due to patient dropouts, missed visits, carelessness, misunderstood directions, and so on. If severe, these deviations can result in extremely difficult analyses and interpretations. Although the experienced researcher anticipates the problems of human research, such problems can be minimized by careful planning.

We will discuss a few examples of designs commonly used in clinical studies. The basic principles of good design should always be kept in mind when considering the experimental pathway to the study objectives. In *Planning of Experiments*, Cox discusses the requirements for a good experiment [2]. When designing clinical studies, the following factors are important:

1. absence of bias;
2. absence of systematic error (use of controls);
3. adequate precision;
4. choice of patients;
5. simplicity and symmetry.

11.2.1 Absence of Bias

As far as possible, known sources of bias should be eliminated by blinding techniques. If a double-blind procedure is not possible, careful thought should be given to alternatives that will suppress, or at least account for possible bias. For example, if the physician can distinguish two comparative drugs, as in an open study, perhaps the evaluation of the response and the administration of the drug can be done by other members of the investigative team (e.g., a nurse) who are not aware of the nature of the drug being administered.

In a double-blind study, both the observer and patient (or subject) are unaware of the treatment being given during the course of the study. Human beings, the most complex of machines, can respond to drugs (or any stimulus, for that matter) in amazing ways as a result of their psychology. This is characterized in drug trials by the well-known "placebo effect." Also, a well-known fact is that the observer (nurse, doctor, etc.) can influence the outcome of an experiment if the nature of the different treatments is known. The subjects of the experiment can be influenced by words and/or actions, and unconscious bias may be manifested in the recording and interpretation of the experimental observations. For example, in analgesic studies, as much as 30% to 40% of patients may respond to a placebo treatment.

The double-blind method is accomplished by manufacturing alternative treatment dosage forms to be as alike as possible in terms of shape, size, color, odor, and taste. Even in the case of dosage forms that are quite disparate, ingenuity can always provide for double blinding. For example, in a study where an injectable dosage form is to be compared to an oral dosage form, the *double-dummy technique* may be used. Each subject is administered both an oral dose and an injection. In one group, the subject receives an active oral dose and a placebo injection, whereas in the other group, each subject receives a placebo oral dose and an active injection. There are occasions where blinding is so difficult to achieve or is so inconvenient to the patient that studies are best left "unblinded." In these cases, every effort should be made to reduce possible biases. For example, in some cases, it may be convenient for one person to administer the study drug, and a second person, unaware of the treatment given, to make and record the observation.

Examples of problems that occur when trials are not blinded are given by Rodda et al. [3]. In a study designed to compare an angiotensin converting enzyme (ACE) inhibitor with a beta-blocker, unblinded investigators tended to assign patients who had been previously unresponsive to beta-blockers to the ACE group. This allocation results in a treatment bias. The ACE group may contain the more seriously ill patients.

An important feature of clinical study design is randomization of patients to treatments. This topic has been discussed in chapter 4, but bears repetition. The randomization procedure

as applied to various designs will be presented in the following discussion. Randomization is an integral and essential part of the implementation and design of clinical studies. Randomization will help to reduce potential bias in clinical experiments, and is the basis for valid calculations of probabilities for statistical testing.

11.2.2 Absence of Systematic Errors

Cox gives some excellent examples in which the presence of a systematic error leads to erroneous conclusions [2]. In the case of clinical trials, a systematic error would be present if one drug was studied by one investigator and the second drug was studied by a second investigator. Any observed differences between drugs could include "systematic" differences between the investigators. This ill-designed experiment can be likened to Cox's example of feeding two different rations to a group of animals, where each group of animals is kept together in separate pens. Differences in pens could confuse the ration differences. One or more pens may include animals with different characteristics that, by chance, may affect the experimental outcome. In the examples above, the experimental units (patients, animals, etc.) are not independent. Although the problems of interpretation resulting from the designs in the examples above may seem obvious, sometimes the shortcomings of experimental procedures are not obvious. We have discussed the deficiencies of a design in which a baseline measurement is compared to a post-treatment measurement in the absence of a control group. Any change in response from baseline to treatment could be due to changes in conditions during the intervening time period. To a great extent, systematic errors in clinical experiments can be avoided by the inclusion of an appropriate control group and random assignment of patients to the treatment groups.

11.2.3 Adequate Precision

Increased precision in a comparative experiment means less variable treatment effects and more efficient estimate of treatment differences. Precision can always be improved by increasing the number of patients in the study. Because of the expense and ethical questions raised by using large numbers of patients in drug trials, the sample size should be based on medical and statistical considerations that will achieve the experimental objectives described in chapter 6.

Often, an appropriate choice of experimental design can increase the precision. Use of baseline measurements or use of a crossover design rather than a parallel design, for example, will usually increase the precision of treatment comparisons. However, in statistics as in life, we do not get something for nothing. Experimental designs have their shortcomings as well as advantages. Properties of a particular design should be carefully considered before the final choice is made. For example, the presence of carryover effects will negate the advantage of a crossover design as presented in section 11.4.

Blocking is another way of increasing precision. This is the basis of the increased precision accomplished by use of the two-way design discussed in section 8.4. In these designs, the patients in a block have similar (and relevant) characteristics. For example, if age and sex are variables that affect the therapeutic response of two comparative drugs, patients may be "blocked" on these variables. Thus if a male of age 55 years is assigned to drug A, another male of age approximately 55 years will be assigned Treatment B. In practice, patients of similar characteristics are grouped together in a block and randomly assigned to treatments.

11.2.4 Choice of Patients

In most clinical studies, the choice of patients covers a wide range of possibilities (e.g., age, sex, severity of disease, concomitant diseases, etc.). In general, inferences made regarding drug effectiveness are directly related to the restrictions (or lack of restrictions) placed on patient eligibility as described in the study protocol. This is an important consideration in experimental design, and great care should be taken to describe that patients may be qualified or disqualified from entering the study.

11.2.5 Simplicity and Symmetry

Again we emphasize the importance of *simplicity.* More complex designs have more restrictions, and a resultant lack of flexibility. The gain resulting from a more complex design should be

weighed against the expense and problems of implementation often associated with more sophisticated, complex designs.

Symmetry is an important design consideration. Often, the symmetry is obvious: In most (but not all) cases, experimental designs should be designed to have equal number of patients per treatment group, equal number of visits per patient, balanced order of administration, and an equal number of replicates per patient. Some designs, such as balanced incomplete block and partially balanced incomplete block designs, have a less obvious symmetry.

11.2.6 Randomization

Principles of randomization have been described in chapter 4. Randomization is particularly important when assigning patients to treatments in clinical trials, ensuring that the requirements of good experimental design are fulfilled and the pitfalls avoided [4]. Among other qualities, proper randomization avoids unknown biases, tends to balance patient characteristics, and is the basis for the theory that allows calculation of probabilities. Randomization ensures a balance in *the long run.* In any given experiment, two groups may not have similar characteristics due to chance. Therefore, it is important to carefully examine properties of the groups to assess if group differences could affect the experimental outcome. Use of covariance analysis can help overcome differences between groups as discussed in section 8.6.

In section 4.2, the advantages of randomization of patients in blocks is discussed. Table 11.1 is a short table of random permutations that gives random schemes for block sizes of 4, 5, 6, 8, and 10. This kind of randomization is also known as restricted randomization and allows for an approximate balance of treatment groups throughout the trial. As an example of the application of Table 11.1, consider a study comparing an active drug with placebo using a parallel design, with 24 patients per group (a total of 48 patients). In this case, a decision is made to group patients in blocks of 8, that is, for each group of eight consecutive patients, four will be on drug and four on placebo. In Table 11.1, we start in a random column in the section labeled "Blocks of 8," and select six sequential columns. Because this is a short table, we would continue into the first column if we had to proceed past the last column. (Note that this table is meant to illustrate the procedure and should not be used repeatedly in real examples or for sample sizes exceeding the total number of random assignments in the table. For example, there are 160 random assignments for blocks of size 8; therefore for a study consisting of more than 160 patients, this table would not be of sufficient size.) If the third column is selected to begin the random assignment, and we assign Treatment A to an odd number and Treatment B to an even number, the first eight patients will be assigned treatment as follows:

$$B\ B\ A\ B\ B\ A\ A\ A.$$

11.2.7 Intent to Treat

In most clinical studies, there is a group of patients who have been administered drug who may not be included in the efficacy data analysis because of various reasons, such as protocol violations. This would include patients, for example, who (a) leave the study early for nondrug-related reasons, (b) take other medications that are excluded in the protocol, or (c) are noncompliant with regard to the scheduled dosing regimen, and so on. Certainly, these patients should be included in summaries of safety data, such as adverse reactions and clinical laboratory determinations. Under FDA guidelines, an analysis of efficacy data should be performed with these patients included as an "intent to treat" (ITT) analysis [5]. Thus, both an efficacy analysis including only those patients who followed the protocol, and an ITT analysis, which includes all patients randomized to treatments (with the possible exception of inclusion of ineligible patients, mistakenly included) are performed. In fact, the ITT analysis may take precedence over the analysis that excludes protocol violators. The protocol violators, or those patients who are not to be included in the primary analysis, should be identified, with reasons for exclusion, prior to breaking the treatment randomization code. The ITT analysis should probably not result in different conclusions from the primary analysis, particularly if the protocol violators and other "excluded" patients occur at random. In most circumstances, a different conclusion may occur for the two analyses only when the significance level is close to 0.05.

Table 11.1 Randomization in Blocks

BLOCKS OF 4																			
1	3	3	2	4	4	1	I	1	2	1	3	3	1	2	4	2	3	1	4
2	2	4	3	3	2	2	2	2	3	4	2	2	4	4	2	4	2	4	3
3	1	1	4	2	1	3	3	3	1	2	1	1	2	3	1	1	4	3	2
4	4	2	1	1	3	4	4	4	4	3	4	4	3	1	3	3	1	2	1
BLOCKS OF 5																			
4	4	1	3	5	5	4	2	5	5	3	5	4	3	2	2	3	2	5	4
2	5	3	5	2	3	5	5	1	1	2	2	2	4	3	5	4	3	1	2
3	3	5	4	1	2	1	3	4	3	5	4	1	5	4	3	2	4	4	3
1	2	4	2	3	1	3	4	2	4	4	3	5	2	5	1	1	1	2	1
5	1	2	1	4	4	2	1	3	2	1	1	3	1	1	4	5	5	3	5
BLOCKS OF 6																			
1	5	2	5	3	2	5	1	5	1	1	2	5	2	6	4	3	4	2	2
2	6	5	3	2	1	2	6	6	3	4	4	1	1	3	5	6	2	6	5
5	9	4	4	1	3	3	5	4	4	2	6	6	6	1	3	2	5	3	1
6	1	1	2	5	5	4	2	3	6	5	1	2	3	2	1	4	6	4	3
3	4	6	1	6	6	1	3	2	5	3	3	3	4	4	6	5	3	1	6
4	3	3	6	4	4	6	4	1	2	6	5	4	5	5	2	1	1	5	4
BLOCKS OF 8																			
7	4	2	4	1	2	1	5	3	4	4	8	5	3	5	2	2	5	1	6
8	2	4	5	8	5	5	2	4	5	6	6	4	5	4	7	8	3	7	7
4	3	1	6	3	6	3	4	5	2	7	5	1	1	3	6	6	6	8	5
1	5	6	3	2	7	8	8	2	1	3	1	3	8	6	3	3	8	5	1
2	8	8	1	7	8'	4	3	8	7	5	7	7	6	1	4	4	2	3	3
3	1	5	8	6	1	2	7	7	6	2	3	2	2	2	5	5	1	6	2
6	7	3	7	5	4	7	1	6	8	8	2	8	4	7	8	7	4	2	4
5	6	7	2	4	3	6	6	1	3	1	4	6	7	8	1	1	7	4	8
BLOCKS OF 10																			
1	9	4	1	3	4	1	4	6	8	9	9	10	9	5	5	6	6	4	3
4	6	5	8	2	7	4	5	3	9	7	6	6	1	1	4	3	2	9	2
5	2	3	4	7	8	5	9	9	2	10	8	10	7	4	3	9	7	10	9
9	8	6	10	8	9	8	10	5	7	2	4	4	4	10	10	4	1	2	7
2	10	8	9	1	6	6	8	4	10	5	2	9	2	6	1	1	9	7	5
10	3	9	5	6	2	9	1	8	1	1	3	5	8	8	8	7	3	3	10
8	4	7	7	9	3	10	7	1	4	3	7	3	3	2	9	2	5	1	8
3	5	2	2	5	1	7	6	7	5	8	1	7	5	3	6	5	8	5	1
6	7	10	3	10	5	3	3	2	6	4	10	8	6	9	2	8	4	6	6
7	1	1	6	4	10	2	2	10	3	6	5	2	10	7	7	10	10	8	4

If the conclusions differ for the two analyses, ITT results are sometimes considered to be more definitive. Certainly, an explanation should be given when conclusions are different for the two analyses. One should recognize that the issue of using an ITT analysis vis-à-vis an analysis including only "compliant" patients remains controversial.

11.3 PARALLEL DESIGN

In a parallel design, two or more drugs are studied, drugs being randomly assigned to different patients. Each patient is assigned a single drug. In the example presented here, a study was proposed to compare the response of patients to a new formulation of an antianginal agent and a placebo with regard to exercise time on a stationary bicycle at fixed impedance. An alternative approach would be to use an existing product rather than placebo as the comparative product. However, the decision to use placebo was based on the experimental objective: to demonstrate that the new formulation produces a measurable and significant increase in exercise time. A difference in exercise time between the drug and placebo is such a measure. A comparison of the new formulation with a positive control (an active drug) would not achieve the objective directly.

In this study, a difference in exercise time between drug and placebo of 60 seconds was considered to be of clinical significance. The standard deviation was estimated to be 65 based

on change from baseline data observed in previous studies. The sample size for this study, for an alpha level of 0.05 and power of 0.90 (beta = 0.10), was estimated as 20 patients per group (see Exercise Problem 7). Therefore 40 patients were entered into the study, 20 each randomly assigned to placebo and active treatment. A randomization that obviates a long consecutive run of patients assigned to one of the treatments was used as described in section 11.2.6. Patients were randomly assigned to each treatment in groups of 10, with 5 patients to be randomly assigned to each treatment. This randomization was applied to each of the 4 subsets of 10 patients (40 patients total). From Table 11.1 starting in the fourth column, patients are randomized into the two groups as follows, placebo if an odd number appears and new formulation if an even number appears:

	Placebo	New formulation
Subset 1	1, 5, 6, 7, 9	2, 3, 4, 8, 10
Subset 2	11, 13, 15, 17, 18	12, 14, 16, 19, 20
Subset 3	22, 24, 27, 28, 29	21, 23, 25, 26, 30
Subset 4	31, 33, 36, 38, 39	32, 34, 35, 37, 40

The first subset is assigned as follows. The first number is 1; patient 1 is assigned to placebo. The second number (reading down) is 8; patient 2 is assigned to the new formulation (NF). The next two numbers (4, 10) are even. Patients 3 and 4 are assigned to NF. The next number is odd (9); patient 5 is assigned to Placebo. The next two numbers are odd and Patients 6 and 7 are assigned to Placebo. Patients 8, 9, and 10 are assigned to NF, placebo, and NF, respectively, to complete the first group of 10 patients. Entering column five, patient 11 is assigned to placebo, and so on.

An alternative randomization is to number patients consecutively from 1 to 40 as they enter the study. Using a table of random numbers, patients are assigned to placebo if an odd number appears, and assigned to the test product (NF) if an even number appears. Starting in the eleventh column of Table IV.1, the randomization scheme is as follows:

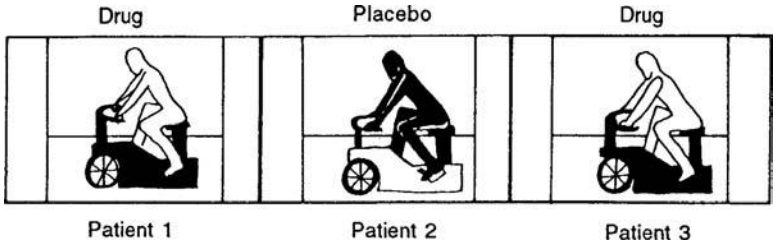
Placebo	New formulation
1, 6, 7, 8	2, 3, 4, 5
12, 13, 14	9, 10, 11
15, 18, 20	16, 17, 19
21, 22, 26	23, 24, 25
27, 28	29, 30, 31
32, 34, 35	33, 38, 39
36, 37	40

For example, the first number in column 11 is 7; patient number 1 is assigned to placebo. The next number in column 11 is 8; the second patient is assigned to the NF; and so on. A problem with this approach is that by chance we may observe a long string of consecutive of odd or even numbers, which would negate the purpose of the randomization as noted above.

Patients were first given a predrug exercise test to determine baseline values. The test statistic is the time of exercise to fatigue or an anginal episode. Tablets were prepared so that the placebo and active drug products were identical in appearance. Double-blind conditions prevailed. One hour after administration of the drug, the exercise test was repeated. The results of the experiment are shown in Table 11.2.

The key points in this design are as follows:

1. There are two independent groups (placebo and active, in this example). An equal number of patients are randomly assigned to each group.
2. A baseline measurement and a single post-treatment measurement are available.



This design corresponds to a one-way analysis of variance, or in the case of two treatments, a two independent groups t test. Since, in general, more than two treatments may be included in the experiment, the analysis will be illustrated using ANOVA.

When possible, pretreatment (baseline) measurements should be made in clinical studies. The baseline values can be used to help increase the precision of the measurements. For example, if the treatment groups are compared using differences from baseline, rather than the post-treatment exercise time, the variability of the measurements will usually be reduced. Using differences, we will probably have a better chance of detecting treatment differences, if they exist (increased power) [6]. "Subtracting out'' the baseline helps to reduce the between-patient variability that is responsible for the variance (the "within mean square") in the statistical test. A more complex, but more efficient analysis is *analysis of covariance.* Analysis of covariance [6] takes baseline readings into account, and for an unambiguous conclusion, assumes that the slope of the response versus baseline is the same for all treatment groups. See "Analysis of Covariance" (sect. 8.6) for a more detailed discussion. Also, the interpretation may be more difficult than the simple "difference from baseline" approach.

To illustrate the results of the analysis with and without baseline readings, the data in Table 11.2 will be analyzed in two ways: (a) using only the post-treatment response,

Table 11.2 Results of the Exercise Test Comparing Placebo to Active Drug: Time (Seconds) to Fatigue or Angina

Placebo				Active drug (new formulation)			
	Exercise time				Exercise time		
Patient	Pre	Post	Post–Pre	Patient	Pre	Post	Post–Pre
1	377	345	−32	2	232	372	140
6	272	310	38	3	133	120	−13
7	348	347	−1	4	206	294	88
8	348	300	−48	5	140	258	118
12	133	150	17	9	240	340	100
13	102	129	27	10	246	393	147
14	156	110	−46	11	226	315	89
15	205	251	46	16	123	180	57
18	296	262	−34	17	166	334	168
20	328	297	−31	19	264	381	117
21	315	278	−37	23	241	376	135
22	133	124	−9	24	74	264	190
26	223	289	66	25	400	541	141
27	256	303	47	29	320	410	90
28	493	487	−6	30	216	301	85
32	336	309	−27	31	153	143	−10
34	299	281	−18	33	193	348	155
35	140	186	46	38	330	440	110
36	161	125	−36	39	258	365	107
37	259	236	−23	40	353	483	130
Mean	259	256	−3.05	Mean	226	333	107.2
s.d.	102	95	36.3	s.d.	83	106	51.5

post-treatment exercise time, and (b) comparing the difference from baseline for the two treatments. The reader is reminded of the assumptions underlying the t test and ANOVA: the variables should be independent, normally distributed with homogeneous variance. These assumptions are necessary for both post-treatment and difference analyses. Possible problems with lack of normality will be less severe in the difference analysis. The difference of independent non-normal variables will tend to be closer to normal than are the original individual data.

Before proceeding with the formal analysis, it is prudent to test the equivalence of the baseline averages for the two treatment groups. This test, if not significant, gives some assurance that the two groups are "comparable." We will use a two independent groups t test to compare baseline values (see sect. 5.2.2).

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{1/N_1 + 1/N_2}}$$
$$= \frac{259 - 226}{S_p\sqrt{1/20 + 1/20}} = \frac{33}{93\sqrt{1/10}} = 1.12.$$

Note that the pooled standard deviation (93) is the pooled value from the baseline readings, $\sqrt{(102^2 + 83^2)/2}$. From Table IV.4, a t value of approximately 2.03 is needed for significance (38 d.f.) at the 5% level. Therefore, the baseline averages are not significantly different for the two treatment groups. If the baseline values are significantly different, one would want to investigate further the effects of baseline on response in order to decide on the best procedure for analysis of the data (e.g., covariance analysis, ratio of response to baseline, etc.).

11.3.1 ANOVA Using Only Post-Treatment Results

The average results for exercise time after treatment are 256 seconds for placebo and 333 seconds for the NF of active drug, a difference of 77 seconds (Table 11.2). Although the averages can be compared using a t test as in the case of baseline readings (above), the equivalent ANOVA is given in Table 11.3. The reader is directed to Exercise Problem 1 for the detailed calculations. According to Table IV.6A1, between groups (active and placebo) is significant at the 5% level.

11.3.2 ANOVA of Differences from the Baseline

When the baseline values are taken into consideration, the active drug shows an increase in exercise time over placebo of 110.25 seconds [107.2 − (−3.05)]. The ANOVA is shown in Table 11.4. The data analyzed here are the (post–pre) values given in Table 11.2. The F test for treatment differences is 61.3! There is no doubt about the difference between the active drug and placebo. The larger F value is due to the considerable reduction in variance as a result of including the baseline values in the analysis. The within-groups error term represents *within-* patient variation in this analysis. In the previous analysis for post-treatment results only, the within-groups error term represents the *between*-patient variation, which is considerably larger than the within-patient error. Although both tests are significant ($p < 0.05$) in this example, one can easily see that situations may arise in which treatments may not be *statistically* different based on a significance test if between-patient variance is used as the error term, but would be significant based on the smaller within-patient variance. Thus, designs that use the smaller within-patient variance as the error term for treatments are to be preferred, other things being equal.

Table 11.3 ANOVA Table for Post-Treatment Readings for the Data of Table 11.2

Source	d.f.	SS	MS	F
Between groups	1	59,213	59,213	$F_{1,38} = 5.86^*$
Within groups	38	383,787	10,099.7	
Total	39	443,000		

$^*p < 0.05$.

Table 11.4 Analysis of Variance for Differences from Baseline (Table 11.1)

Source	d.f.	SS	MS	*F*
Between groups	1	121,551	120,551	$F_{1,38} = 61.3$*
Within groups	38	75,396	1984	
Total	39	196,947		

*$p < 0.01$.

11.4 CROSSOVER DESIGNS AND BIOAVAILABILITY/BIOEQUIVALENCE STUDIES

In a typical crossover design, each subject takes each of the treatments under investigation on different occasions. Comparative bioavailability* or bioequivalence studies, in which two or more formulations of the same drug are compared, are usually designed as crossover studies. Perhaps the greatest appeal of the crossover design is that each patient acts as his or her own control. This feature allows for the direct comparison of treatments, and is particularly efficient in the presence of large interindividual variation. However, caution should be used when considering this design in studies where carryover effects or other interactions are anticipated. Under these circumstances, a parallel design may be more appropriate.

11.4.1 Description of Crossover Designs: Advantages and Disadvantages

The crossover (or changeover) design is a very popular, and often desirable, design in clinical experiments. In these designs, typically, two treatments are compared, with each patient or subject taking each treatment in turn. The treatments are typically taken on two occasions, often called *visits*, *periods*, or *legs.* The order of treatment is randomized; that is, either *A* is followed by *B* or *B* is followed by *A*, where *A* and *B* are the two treatments. Certain situations exist where the treatments are not separated by time, for example, in two visits or periods. For example, comparing the effect of topical products, locations of applications on the body may serve as the visits or periods. Product may be applied to each of two arms, left and right. Individuals will be separated into two groups, (1) those with Product *A* applied on the left arm and Product *B* on the right arm, and (2) those with Product *B* applied on the left arm and Product *A* on the right arm.

$A \longrightarrow B$ (First week, Second week) or $B \longrightarrow A$ (First week, Second week)

This design may also be used for the comparison of more than two treatments. The present discussion will be limited to the comparison of two treatments, the most common situation in clinical studies. (The design and analysis of three or more treatment crossovers follows.) Crossover designs have great appeal when the experimental objective is the comparison of the performance, or effects, of two drugs or product formulations. Since each patient takes each product, the comparison of the products is based on *within*-patient variation. The *within*- or intrasubject variability will be smaller than the *between*- or intersubject variability used for the comparison of treatments in the one-way or parallel-groups design. Thus, crossover experiments usually result in greater precision than the parallel-groups design, where different patients comprise the two groups. Given an equal number of observations, the crossover design is more powerful than a parallel design in detecting product differences.

The crossover design is a type of Latin square. In a Latin square, the number of treatments equals the number of patients. In addition, another factor, such as order of treatment, is included in the experiment in a balanced way. The net result is an $N \times N$ array (where N is the number of treatments or patients) of N letters such that a given letter appears only once in a given row or column. This is most easily shown pictorially. A Latin square for four subjects taking four drugs is shown in Table 11.5. For randomizations of treatments in Latin squares, see Ref. [6].

* A bioavailability study, in our context, is defined as a comparative study of a drug formulation compared to an optimally absorbed (intravenous or oral solution) formulation.

Table 11.5 4 × 4 Latin Square: Four Subjects Take Four Drugs

	Order in which drugs[a] are taken			
Subject	**First**	**Second**	**Third**	**Fourth**
1	*A*	*B*	*C*	*D*
2	*B*	*C*	*D*	*A*
3	*C*	*D*	*A*	*B*
4	*D*	*A*	*B*	*C*

[a]Drugs are designated as *A, B, C, D*.

For the comparison of two formulations, a 2 × 2 Latin square ($N = 2$) consists of two patients each taking two formulations (A and B) on two different occasions in two ''orders'' as follows:

	Occasion period	
Patient	**First**	**Second**
1	*A*	*B*
2	*B*	*A*

The balancing of order ($A - B$ or $B - A$) takes care of time trends or other "period" effects, if present. (A period effect is a difference in response due to the occasion on which the treatment is given, independent of the effect due to the treatment.)

The 2 × 2 Latin square shown above is familiar to all who have been involved in bioavailability/bioequivalence studies. In these studies, the 2 × 2 Latin square is repeated several times to include a sufficient number of patients (see also Table 11.6). Thus, the crossover design can be thought of as a repetition of the 2 × 2 Latin square.

The crossover design has an advantage, previously noted, of increased precision relative to a parallel-groups design. Also, the crossover is usually more economical: one-half the number of patients or subjects have to be recruited to obtain the same number of observations as in a parallel design. (Note that each patient takes *two* drugs in the crossover.) Often, a significant part of the expense in terms of both time and money is spent recruiting and processing patients or volunteers. The advantage of the crossover design in terms of cost depends on the economics of patient recruiting, cost of experimental observations, as well as the relative within-patient/between-patient variation. The smaller the within-patient variation relative to the between-patient variation, the more efficient will be the crossover design. Hence, if a repeat observation on the same patient is very variable (nonreproducible), the crossover may not be very much better than a parallel design, cost factors being equal. This problem is presented and quantitatively analyzed in detail by Brown [7].

There are also some problems associated with crossover designs. A crossover study may take longer to complete than a parallel study because of the extra testing period. It should be noted, however, that if recruitment of patients is difficult, the crossover design may actually save time, because fewer patients are needed to obtain equal power compared to the parallel design. Another disadvantage of the crossover design is that missing data pose a more serious problem than in the parallel design. Since each subject must supply data on *two occasions* (compared to a single occasion in the parallel design), the chances of observations being lost to the analysis are greater in the crossover study. If an observation is lost in one of the legs of a two-period crossover, the data for that person carry very little information. When data are missing in the crossover design, the statistical analysis is more difficult and the design loses some efficiency. Finally, the administration of crossover designs in terms of management and patient compliance is somewhat more difficult than that of parallel studies.

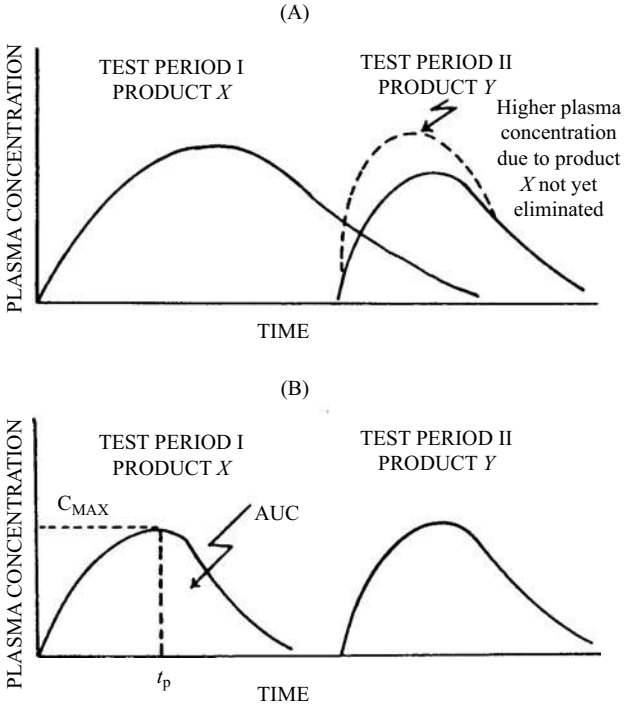


Figure 11.1 Carryover in a bioequivalence study.

Perhaps the most serious problem with the use of crossover designs is one common to all Latin square designs, the *possibility of interactions.* The most common interaction that may be present in crossover design is a *differential carryover* or residual effect. This effect occurs when the response on the second period (leg) is dependent on the response in the first period, and this dependency differs depending on which of the two treatments is given during the first period. Carryover is illustrated in Figure 11.1(A), where the short interval between administration of dosage forms *X* and *Y* is not sufficient to rid the body of drug when formulation *X* is given first. This results in an apparent larger blood level for formulation *Y* when it is given subsequent to formulation *X*. In the presence of differential carryover, the data cannot be properly analyzed except by the use of more complex designs (see replicate crossover designs in sect. 11.4.7). These special designs are not easily accommodated to clinical studies [8].

Figure 11.1(B) illustrates an example where a sufficiently long washout period ensures that carryover of blood concentration of drug is absent. The results depicted in Figure 11.1(A) show a carryover effect that could easily have been avoided if the study had been carefully planned. This example only illustrates the problem; often, carryover effects are not as obvious. These effects can be caused by such uncontrolled factors as psychological or physiological states of the patients, or by external factors such as the weather, clinical setting, assay techniques, and so on.

Grizzle has published an analysis to detect carryover (residual) effects [9]. When differential carryover effects are present, the usual interpretation and statistical analysis of crossover studies are invalid. Only the first period results can be used, resulting in a smaller, less sensitive experiment. An example of Grizzle's analysis is shown in this chapter in the discussion of bioavailability studies (sect. 11.4.2). Brown concludes that most of the time, in these cases, the parallel design is probably more efficient [7]. Therefore, if differential carryover effects are suspected prior to implementation of the study, an alternative to the crossover design should be considered (see below).

Because of the "built-in" individual-by-individual comparisons of products provided by the crossover design, the use of such designs in comparative clinical studies often seems very attractive. However, in many situations, where patients are being treated for a disease state, the design is either inappropriate or difficult to implement. In acute diseases, patients may be cured or improved so much after the first treatment that a "different" condition or state of illness is being treated during the second leg of the crossover. Also, psychological carryover has been observed, particularly in cases of testing psychotropic drugs.

The longer study time necessary to test two drugs in the crossover design can be critical if the testing period of each leg is of long duration. Including a possible *washout period* to avoid possible carryover effects, the crossover study will take at least *twice* as long as a parallel study to complete. In a study of long duration, there will be more difficulty in recruiting and maintaining patients in the study. One of the most frustrating (albeit challenging) facets of data analysis is data with "holes," missing data. Long-term crossover studies will inevitably have such problems.

11.4.2 Bioavailability/Bioequivalence Studies†

The assessment of "bioequivalence" (BE) refers to a procedure that compares the bioavailability of a drug from different formulations. Bioavailability is defined as the rate and extent to which the active ingredient or active moiety is absorbed from a drug product and becomes available at the site of action. For drug products that are not intended to be absorbed into the bloodstream, bioavailability may be assessed by measurements intended to reflect the rate and extent to which the active ingredient or active moiety becomes available at the site of action. In this chapter, we will not present methods for drugs that are not absorbed into the bloodstream (or absorbed so little as to be unmeasurable), but may act locally. Products containing such drugs are usually assessed using a clinical endpoint, using parallel designs discussed elsewhere in this chapter. Statistical methodology, in general, will be approached in a manner consistent with methods presented for drugs that are absorbed.

Thus, we are concerned with measures of the release of drug from a formulation and its availability to the body. BE can be simply defined by the relative bioavailability of two or more formulations of the same drug entity. According to 21 CFR 320.1, BE is defined as "the absence of a significant difference in the rate and extent to which the active ingredient or active moiety . . . becomes available at the site of drug action when administered . . . in an appropriately designed study."

BE is an important part of an NDA in which formulation changes have been made during and after pivotal clinical trials. BE studies, as part of Abbreviated New Drug Application (ANDA) submissions, in which a generic product is compared to a marketed, reference product, are critical parts of the submission. BE studies may also be necessary when formulations for approved marketed products are modified.

In general, most BE studies depend on accumulation of pharmacokinetic (PK) data that provide concentrations of drug in the bloodstream at specified time points following administration of the drug. These studies are typically performed, using oral dosage forms, on volunteers who are incarcerated (housed) during the study to ensure compliance with regard to dosing schedule as well as other protocol requirements. This does not mean that BE studies are limited to oral dosage forms. Any drug formulation that results in measurable blood concentrations after administration can be treated and analyzed in a manner similar to drugs taken orally. For drugs that act locally and are not appreciably absorbed, either a surrogate endpoint may be utilized in place of blood concentrations of drug (e.g., a pharmacodynamic response) or a clinical study using a therapeutic outcome may be necessary. Also, in some cases where assay methodology in blood is limited, or for other relevant reasons, measurements of drug in the urine over time may be used to assess equivalence.

To measure rate and extent of absorption for oral products, PK measures are used. In particular, model independent measures used are (a) area under the blood concentration versus

† Additional discussion of designs and analyses are given in Appendix X.

time curve (AUC) and the maximum concentration (C_{max}), which are measures of the amount of drug absorbed and the rate of absorption, respectively.

The time at which the maximum concentration occurs (t_{max}) is a more direct measure as an indicator of absorption rate, but is a very variable estimate.

Bioavailability/bioequivalence studies are particularly amenable to crossover designs. Virtually all such studies make use of this design. Most BE studies involve single doses of drugs given to normal volunteers, and are of short duration. Thus the disadvantages of the crossover design in long term, chronic dosing studies are not apparent in bioavailability studies. With an appropriate washout period between doses, the crossover is ideally suited for comparative bioavailability studies.

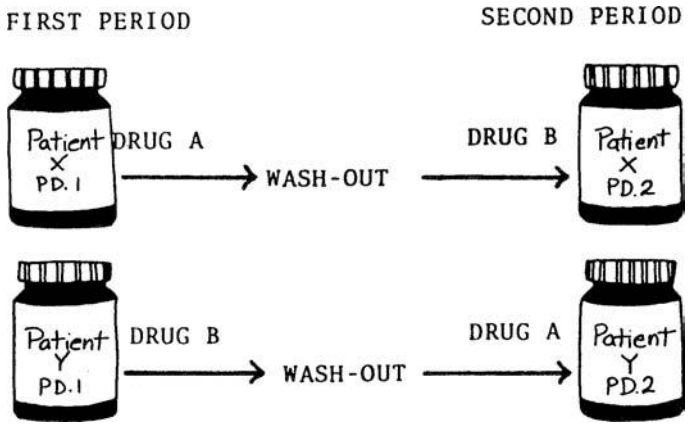
Statistical applications are essential for the evaluation of BE studies. Study designs are typically two-treatment, two-period (tttp) crossover studies with single or multiple (steady state) dosing, fasting or fed. Designs with more than two periods are now becoming more common, and are recommended in certain cases by the FDA. For long half-life drugs, where time is crucial, parallel designs may be desirable, but these studies use more subjects than would be used in the crossover design, and the implementation of parallel studies may be difficult and expensive. The final evaluation is based on parameter averages derived from the blood level curves, AUC, C_{max}, and t_{max}. Statistical analyses that have been recommended are varied, and the analyses presented here are typical of those recommended by regulatory agencies.

This section discusses some designs, their properties, and statistical evaluations.

Although crossover designs have clear advantages over corresponding parallel designs, their use is restricted, in general, as previously noted, because of potential differential carryover effects and confounded interactions. However, for BE studies, the advantages of these designs far outweigh the disadvantages. Because these studies are typically performed in healthy volunteers and are of short duration, the potential for carryover and interactions is minimal. In particular, the likelihood of differential carryover seems to be remote. Carryover may be observed if administration of a drug affects the blood levels of subsequent doses. Although possible, a carryover effect would be very unusual, particularly in single-dose studies with an adequate washout period. A washout period of at least seven half-lives is recommended. Even more unlikely, would be a differential carryover, which suggests that the carryover from one product is different from the carryover from the second product. A differential carryover effect can invalidate the second period results in a two-period crossover (see below). Because BE studies compare the same drug in different formulations, if a carryover exists at all, the carryover of two different formulations would not be expected to differ. This is not to say that differential carryover is impossible in these studies, but to this author's knowledge, differential carryover has not been verified in results of published BE studies, single or multiple dose. In the typical tttp design, differential carryover is confounded with other effects, and a test for carryover is not definitive. Thus, if such an effect is suspected, proof would require a more restrictive or higher order design, that is, a design with more than two periods. This problem will be discussed further as we describe the analysis and inferences resulting from these designs.

The features of the tttp design are as follows:

1. N subjects recruited for the study are separated into two groups, or two treatment *sequences.* N_1 subjects take the treatments in the order AB, and N_2 in the order BA, where $N_1 + N_2 = N$. For example, 24 (N) subjects are recruited and 12 (N_1) take the Generic followed by the Brand product, and 12 (N_2) take the Brand followed by the Generic. Note that the product may be taken as a single dose, in multiple doses, fasted or fed.
2. After administration of the product in the first period, blood levels of drug are determined at suitable intervals.
3. A washout period follows, which is of sufficient duration to ensure the "total" elimination of the drug given during the first period. An interval of at least nine drugs half-lives should be sufficient to ensure virtually total elimination of the drug. Often, a minimum of seven half-lives is recommended.
4. The alternate product is administered in the second period and blood levels determined as during Period 1.



Crossover designs are planned so that each treatment is given an equal number of times in each period. This is most efficient and yields unbiased estimates of treatment differences if a period effect is present.

The blood is analyzed for each subject with both first and second periods analyzed concurrently (the same day). To detect possible analytical errors, the samples are usually analyzed chronologically (starting from the time 0 sample to the final sample), but with the identity of the product assayed unknown (sample blinding).

After the blood assays are complete, the blood level versus time curves are analyzed for the derived parameters, AUC_t (also noted as AUC_{0-t}), $AUC_{0-\infty}$, C_{max}, and t_{max} (t_p), for each analyte. AUC_t is the area to the last quantifiable concentration, and AUC_{inf} is AUC_t augmented by an estimate of the area from time t to infinity (C_t/K_e). This is shown and explained in Figure 11.2. A detailed analysis follows.

The analysis of the data consists of first determining the maximum blood drug concentration (C_{max}) and the area under the blood level versus time curve (AUC) for each subject, for each product. Often, more than one analyte is observed, for example, metabolites or multiple ingredients, all of which may need to be separately analyzed.

AUC is determined using the trapezoidal rule. The area between adjacent time points may be estimated as a trapezoid (Fig. 11.3). The area of each trapezoid, up to and including the final time point, where a measurable concentration is observed, is computed, and the sum of these areas is the AUC, designated as AUC_t. The area of a trapezoid is $^1/_2$ (base) (sum of two sides). For example, in Figure 11.3, the area of the trapezoid shown in the blood level versus time curve is 4. In this figure, C_{max} is 5 ng/mL and t_{max}, the time at which C_{max} occurs, is two hours. Having performed this calculation for each subject and product, the AUC and C_{max} values are transformed to their respective logarithms. Either natural logs (ln) or logs to the base

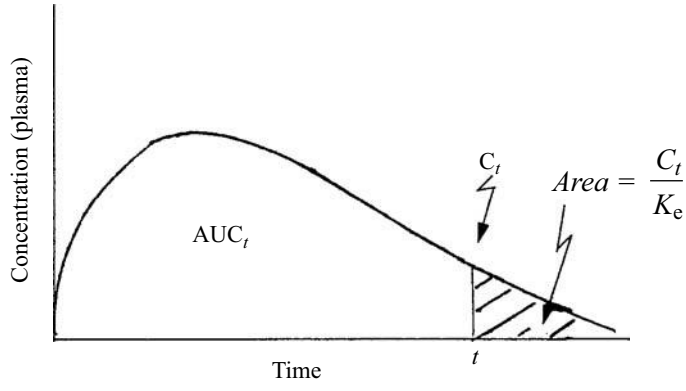


Figure 11.2 Derived parameters from bioequivalence study.

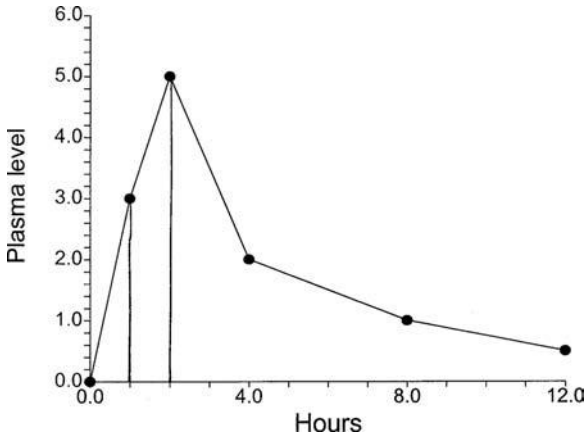


Figure 11.3 Illustration of trapezoidal rule.

10 (log) may be used. Typically, one uses the natural log, or ln. The details of the analysis are described later in this chapter. The analysis of AUC and C_{max} was not always performed on the logs of these values. Originally, the actual, observed (nontransformed) values of these derived parameters were used in the analysis. (This history will be discussed in more detail below.) However, examination of the theoretical derivations and mathematical expression of AUC and C_{max}, as well as the statistical properties, has led to the use of the logarithmic transformation. In particular, data appear to show that these values follow a log-normal distribution more closely than they do a normal distribution. The form of expression for AUC suggests a multiplicative model

$$\text{AUC} = FD/VK_e,$$

where F is fraction of drug absorbed, D is dose, V is volume of distribution, and K_e is elimination rate constant.

The distribution of AUC is complex because of the nonlinearity; it is a ratio. Ln(AUC) is equal to $\ln(F) + \ln(D) - \ln(V) - \ln(K_e)$. This is linear, and the statistical properties are more manageable. A similar argument can be made for C_{max}.

The present FDA requirement for equivalence is based on product ratios using a symmetric 90% confidence interval for the difference of the average parameters, after a log transformation. Earlier, according to FDA guidelines, the AUC and C_{max} were analyzed using the untransformed values of these derived parameters. Note that when using a clinical or pharmacodynamic endpoint (such as may be used in a parallel study when drug is not absorbed), the nontransformed data may be more appropriate and the "old" way of forming the confidence interval may still be used. This analysis is described below. However, at the present time, FDA is leaning toward an analysis based on Fieller's Theorem (Locke's Method). (These analyses, along with a log-transformed analysis, are described in the example following this discussion.)

11.4.2.1 Statistical Analysis

It is convenient to follow the statistical analysis and estimation of various effects by looking at the two sequences in the context of the model for this design:

Let
μ = overall mean
G_i = Effect of sequence group i ($i = 1, 2$)
S_{ik} = Effect of subject k in sequence i ($k = 1, 2, 3 \ldots N$)
P_j = Effect of period j ($j = 1, 2$)
$T_{t(ij)}$ = treatment effect t ($t = 1,2$) in sequence i and period j
$Y_{ijk} = \mu + G_i + S_{ik} + P_j + T_t(_{ij}) + e_{ijk}$

Table 11.6 Design for Two-Way Crossover Study

	Period I	Period II
Sequence I	*A*	*B*
Sequence II	*B*	*A*

The sequence × period interaction is the treatment effect (sequence × period is the comparison Period I–Period II for the two sequences; see Table 11.6).

e.g.,

$$\frac{\left[(A-B)_{\text{seq I}} - (B-A)_{\text{seq II}}\right]}{2} = A - B.$$

Suppose that carryover is present, but carryover is the same for both products. We can show that this would not bias the treatment comparisons. For the sake of simplicity, suppose that there is no period effect ($P_1 = P_2$). Also suppose that the direct treatment effects are $A = 3$ and $B = 2$. Both products have a carryover that adds 2 to the treatment (product) in the second period. (This would result in an additional value of 2 for the period effect.) Finally, assume that the effects for the sequences are equal; Sequence I = Sequence II. This means that the average results for subjects in Sequence I are the same as that for Sequence II. Based on this model, Product *B* in Period II would have a value of 2 + 2 for carryover = 4. Product *A* in Period II has a value of 3 + 2 = 5. Thus, the average difference between *A* and *B* is 1, as expected ($A = 3$ and $B = 2$). Table 11.7 shows these simulated data.

This same reasoning would show that equal carryover effects do not bias treatment comparisons in the presence of a period effect. (See Exercise Problem 11 at the end of this chapter.)

Differential carryover, where the two products have different carryover effects, is confounded with a sequence effect. This means that if the sequence groups have significantly different average results, one cannot distinguish this effect from a differential carryover effect in the absence of more definitive information. For example, one can show that if there is a sequence effect and no differential carryover, a differential carryover in the absence of a sequence effect could give the same result.

To help explain the confounding, assume that the difference between treatments is 0 (treatments are identical) and that Sequence I averages 2 units more than Sequence II (e.g., Sequence I = Sequence II + 2). Since subjects are assigned to the sequence groups at random, the differences should not be significant except by chance. With no carryover or period effects, the average results could be something like that shown in Table 11.8.

If Sequence I is the order *A* followed by *B* (*AB*) and Sequence II is the order *BA*, the treatment differences, $A - B$, would be $6 - 6 = 0$ in Sequence I, and $4 - 4 = 0$ in Sequence II. Treatment *A* is the same as Treatment *B* in Sequence I, and in Sequence II. However, this same result could occur as a result of differential carryover in the presence of treatment differences.

Table 11.9 shows the same results as Table 11.8 in a different format.

The data from Table 11.9 can be explained by assuming that *A* is 2 units higher than *B* (see Period I results), a carryover of +2 units when *B* follows *A*, and a carryover of −2 units when *A* follows *B*. The two explanations, a sequence effect or a differential carryover, cannot be separated in this two-way crossover design. The sequence effect is $G_1 - G_2$. The differential carryover is $\left\{[T_A(2) - T_A(1)] - [T_B(2) - T_B(1)]\right\}/2 = \left\{[T_A(2) + T_B(1)] - [T_B(2) + T_A(1)]\right\}/2$, which is exactly the sequence effect (average results in Sequence II − average results in Sequence I). The subscript *B*(l) refers to average result for Product *B* in Period I.

Table 11.7 Simulated Data Illustrating Equal Carryover

	Period I	Period II
Sequence I	$A = 3$	$B = 4$
Sequence II	$B = 2$	$A = 5$

Table 11.8 Example of Sequence Effect

	Treatment *A*	Treatment *B*	Average
Sequence I	6	6	6
Sequence II	4	4	4
Average	5	5	

In practice, an ANOVA is performed, which results in significance tests for the sequence effect and an estimate of error for computing confidence intervals (see later in sect. 11.4.3).

The results of a typical single-dose BE study are shown in Table 11.10. These data were obtained from drug plasma level versus time determinations similar to those illustrated in Figure 11.1(B). Area under the plasma level versus time curve (AUC, a measure of absorption), time to peak plasma concentration (t_p), and the maximum concentration (C_{max}) are the parameters that are usually of most interest in the comparison of the bioavailability of different formulations of the same drug moiety.

The typical ANOVA for crossover studies will be applied to the AUC data to illustrate the procedure used to analyze the experimental results. In these analyses, the residual error term is used in statistical computations, for example, to construct confidence intervals. An ANOVA is computed for each parameter based on the model. The ANOVA table is not meant for the performance of statistical hypothesis tests, except perhaps to test the sequence effect, which uses the between subject within sequences mean square as the error term. Rather, the analysis removes some effects from the total variance to obtain a more "efficient" or pure estimate of the error term. It is the error term, or estimate of the within-subject variability (assumed to be equal for both products in this analysis), that is used to assess the equivalence of the parameter being analyzed. A critical assumption for the correct interpretation of the analysis is the *absence* of differential carryover effects, as discussed previously. Otherwise, the usual assumptions for ANOVA should hold. FDA statisticians encourage a careful statistical analysis of crossover designs. In particular, the use of a simple *t* test that ignores the possible presence of period and/or carryover effects is not acceptable.[‡] If period effects are present, and not accounted for in the statistical analysis, the analysis will be less sensitive. The error mean square in the ANOVA will be inflated due to inclusion of the period variance, and the width of the confidence interval will be increased. If differential carryover effects are present, the estimate of treatment differences will be biased (see sects. 11.4.1 and 11.4.2).

The usual ANOVA separates the total sum of squares into four components: subjects, periods, treatments, and error (residual). In the absence of differential carryover effects, the statistical test of interest is for treatment differences. The subject and period sum of squares are separated from the error term which then represents "intrasubject" variation. The subjects sum of squares (SS) can be separated into sequence SS and subject within sequence SS to test for the sequence effect. The sequence effect is confounded with carryover, and this test is described following the analysis without sequence effect.

Some history may be of interest with regard to the analysis recommended in the most recent FDA guidance [10]. In the early evolution of BE analysis, a hypothesis test was used at the 5% level of significance. The raw data were used in the analysis; that is, a logarithmic transformation was not recommended. The null hypothesis was simply that the products were equal, as opposed to the alternate hypothesis that the products were different. This had the obvious problem with regard to the power of the test. Products that showed nearly the same average

Table 11.9 Example of Differential Carryover Effect

	Period I	Period II	Average
Sequence I	$A = 6$	$B = 6$	6
Sequence II	$B = 4$	$A = 4$	4

[‡] In bioavailability studies, carryover effects are usually due to an inadequate washout period.

Table 11.10 Data for the Bioequivalence Study Comparing Drugs *A* and *B*

		AUC		Peak concentration		Time to peak	
Subject	**Order**	***A***	***B***	***A***	***B***	***A***	***B***
1	*AB*	290	210	30	18	8	8
2	*BA*	201	163	22	19	10	4
3	*AB*	187	116	18	11	6	6
4	*AB*	168	77	20	14	10	3
5	*BA*	200	220	18	21	3	3
6	*BA*	151	133	25	16	4	6
7	*AB*	294	140	27	14	4	10
8	*BA*	97	190	16	23	6	6
9	*BA*	228	168	20	14	6	6
10	*AB*	250	161	28	19	6	4
11	*AB*	293	240	28	18	6	12
12	*BA*	154	188	16	20	8	8
	Mean	209.4	167.2	22.3	17.3	6.4	6.3
	Sum	2513	2006	268	207	77	76

results, but with very small variance, could show a significant difference, which may not be of clinical significance, and be rejected. Alternatively, products that showed large differences with large variance could show a nonsignificant difference, and be deemed equivalent. Similarly, products could be shown to be equivalent if a small sample size was used resulting in an undetected difference that could be clinically significant. Because of these problems, an additional caveat was added to the requirements. If the products showed a difference of less than 20%, was not statistically significant ($p > 0.05$), and the power of the study to detect a difference of 20% exceeded 80%, the products would be considered to be equivalent. This helped to avoid undersized studies and prevent products with observed large differences from passing the BE study. The following examples illustrate this problem.

Example 1. In a BE two-period, crossover study, with eight subjects, the test product showed an average AUC of 100, and the reference product showed an average AUC of 85. The observed difference between the products is (100–85)/85, or 17.6%.

The error term from the ANOVA (see below for description of the analysis) is 900, s.d. = 30. The test of significance (a t test with 6 d.f.) is

$$\frac{|100 - 85|}{\left[900\left(\frac{1}{8} + \frac{1}{8}\right)\right]^{1/2}} = 1.00.$$

This is not statistically significant at the 5% level (a t value of 2.45 for 6 d.f. is needed for significance). Therefore, the products may be deemed equivalent.

However, this test is underpowered based on the need for 80% power to show a 20% difference. A 20% difference from the reference is $0.2 \times 85 = 17$. The approximate power is (Eq. 6.11)

$$Z = [17/42.43]\,[6]^{1/2} - 1.96 = -0.98.$$

Referring to a Table of the Cumulative Standard Normal Distribution, the approximate power is 16%. Although the test of significance did not reject the null hypothesis, the power of the test to detect a 20% difference is weak. Therefore, this product would not pass the BE requirements.

Example 2. In a BE two-period, crossover study, with 36 subjects, the test product showed an average AUC of 100, and the reference product showed an average AUC of 95. The products

differ by approximately only 5%. The error term from the ANOVA is 100, s.d. = 10. The test of significance (a t test with 34 d.f.) is

$$\frac{|100 - 95|}{\left[100\left(\frac{1}{36} + \frac{1}{36}\right)\right]^{1/2}} = 2.12.$$

This is statistically significant at the 5% level (a t value of 2.03 for 34 d.f. is needed for significance). Therefore, the products may be deemed nonequivalent.

This test passes the criterion based on the need for 80% power to show a 20% difference. A 20% difference from the reference is $0.2 \times 95 = 19$. The approximate power is (see chap. 6)

$$Z = [19/14.14]\,[34]^{1/2} - 1.96 = 5.88.$$

The approximate power is almost 100%. Although the power of the test to detect a 20% difference is extremely high, the test of significance rejected the null hypothesis that the products were equal. Therefore, this Product would fail the BE requirements. In some cases, a Medical review would rule such a small difference as clinically non-significant and the product would be approved.

Other requirements at that time included the 75/75 rule [11]. This rule stated that 75% of the subjects in the study should have ratios of test/reference between 75% and 125%. This was an attempt to include a variability criterion in the assessment of study results. Unfortunately, this criterion has little statistical basis, and would almost always fail with highly variable drugs. In fact, if a highly variable drug (CV greater than 30–40%) is tested against itself, it would most likely fail this test. Eventually, this requirement was correctly phased out.

Soon after this phase in the evolution of BE regulations, the hypothesis test approach was replaced by the two one-sided t test or, equivalently, the 90% confidence interval approach [12]. This approach resolved the problems of hypothesis testing, and assumed that products that are within 20% of each other with regard to the major parameters, AUC and $C_{\max}$, are therapeutically equivalent. For several years, this method was used without a logarithmic transformation. However, if the study data conformed better to a log-normal distribution than a normal distribution, a log transformation was allowed. An appropriate statistical test was applied to test the conformity of the data to these distributions.

The AUC data from Table 11.10 are analyzed below. To ease the explanation, the computations for the untransformed data are detailed. The log-transformed data are analyzed identically, and these results follow the untransformed data analysis. The sums of squares for treatments and subjects are computed exactly the same way as in the two-way ANOVA (see sect. 8.4). The new calculations are for the "period" (1 d.f.) and "sequence" (1 d.f.) sums of squares. We first show the analysis for periods. The analysis for sequence is shown when discussing the test for differential carryover. Two new columns are prepared for the "period" calculation. One column contains the data from the first period, and the second column contains data from the second period. For example, for the AUC data in Table 11.10, the data for the first period are obtained by noting the order of administration. Subject 1 took Product A during the first period (290); subject 2 took B during the first period (163); and so on. Therefore, the first period observations are

290, 163, 187, 168, 220, 133, 294, 190, 168, 250, 293, and 188 (sum = 2544).

The second period observations are

210, 201, 116, 77, 200, 151, 140, 97, 228, 161, 240, 154 (sum = 1975).

The "period" SS may be calculated as follows:

$$\frac{(\sum P_1)^2 + (\sum P_2)^2}{N} - \mathrm{CT}, \tag{11.2}$$

where $\sum P_1$ and $\sum P_2$ are the sums of observations in the first and second periods, respectively, N is the number of subjects, and CT is the correction term. The following ANOVA and Table 11.10 will help clarify the calculations.

Calculations for ANOVA

$\sum X_t$ is the sum of all observations $= 4519$
$\sum X_A$ is the sum of observations for Product $A = 2513$
$\sum X_B$ is the sum of observations for Product $B = 2006$
$\sum P_1$ is the sum of observations for Period $1 = 2544$
$\sum P_2$ is the sum of observations for Period $2 = 1975$
$\sum X_t^2$ is the sum of the squared observations $= 929{,}321$
CT is the correction term $\frac{(\sum X_t)^2}{N_t} = \frac{(4519)^2}{24} = 850,890.04.$
Total SS $= \sum X_t^2 - \text{CT} = 78,430.96$
$\sum S_i$ is the sum of the observations for subject i (e.g., 500 for first subject)
Subject SS

$$= \frac{\sum (\sum S_i)^2}{2} - \text{CT} = \frac{500^2 + 364^2 + \ldots + 342^2}{2} - \text{CT} = 43{,}560.46$$

$$\text{Period sum of squares} = \frac{2544^2 + 1975^2}{12} - \text{CT} = 13{,}490.0$$

$$\text{Treatment sum of squares} = \frac{2513^2 + 2006^2}{12} - \text{CT} = 10{,}710.4$$

$$\begin{aligned}\text{Error SS} &= \text{total SS} - \text{subject SS} - \text{period SS} - \text{treatment SS} \\ &= 78{,}430.96 - 43{,}560.46 - 13{,}490 - 10{,}710.38 \\ &= 10{,}670.1.\end{aligned}$$

Note that the d.f. for error are equal to 10. The usual two-way ANOVA would have 11 d.f. for error (subjects $-$ 1) $\times$ (treatments $-$ 1). In this design, the error SS is diminished by the *period SS*, which has 1 d.f.

Again, the ANOVA is typically performed using appropriate computer programs. A General Linear Models (GLM) program is suitable with factors, sequence, subjects within sequence, treatment, and period.

11.4.2.2 Test for Carryover Effects

Dr. James Grizzle published a classic paper on analysis of crossover designs and presented a method for testing carryover effects (sequence effects in his notation) [9]. Some controversy exists regarding the usual analysis of crossover designs, particularly with regard to the assumptions underlying this analysis. Before using the Grizzle analysis, the reader should examine the original paper by Grizzle as well as the discussion by Brown, in which some of the problems of crossover designs are summarized [7].

One of the key assumptions necessary for a valid analysis and interpretation of crossover designs is the absence of differential carryover effects as has been previously noted. Data from Table 11.10 were previously analyzed using the typical crossover analysis, assuming that differential carryover was absent. Table 11.10 is reproduced as Table 11.11 (AUC only) to illustrate the computations needed for the Grizzle analysis.

The test for carryover, or sequence, effects is performed as follows:

1. Compute the SS due to carryover (or sequence) effects by comparing the results for group I to group II. (Note that these two groups, groups I and II, which differ in the order of treatment are designated as treatment "sequence" by Grizzle.) It can be demonstrated that in the absence of sequence effects, the average result for group I (*A* first, *B* second) is expected to be equal to the average result for group II (*B* first, *A* second). The SS is calculated as

$$\frac{\left(\sum \text{group I}\right)^2}{N_1} + \frac{\left(\sum \text{group II}\right)^2}{N_2} - \text{CT}.$$

Table 11.11 Data for AUC for the Bioequivalence Study Comparing Drugs *A* and *B*

Group I (Treatment *A* first, *B* second)				Group II (Treatment *B* first, *A* second)			
Subject	***A***	***B***	**Total**	**Subject**	***A***	***B***	**Total**
1	290	210	500	2	201	163	364
3	187	116	303	5	200	220	420
4	168	77	245	6	151	133	284
7	294	140	434	8	97	190	287
10	250	161	411	9	228	168	396
11	293	240	533	12	154	188	342
Total	1482	944	2426	Total	1031	1062	2093

In our example the sequence SS is (1 d.f.)

$$\frac{(2426)^2}{12} + \frac{(2093)^2}{12} - \frac{(2426+2093)^2}{24} = 4620.375.$$

2. The proper error term to test the sequence effect is the within-group (sequence) mean square, represented by the SS between subjects within groups (sequence). This SS is calculated as follows:

$$\frac{1}{2}\sum (\text{subject total})^2 - (\text{CT})_{\text{I}} - (\text{CT})_{\text{II}},$$

where CT_{I} and CT_{II} are the correction terms for groups I and II, respectively. In our example, the within-group SS is

$$\tfrac{1}{2}(500^2 + 303^2 + 245^2 + \ldots + 364^2 + 420^2 + \ldots 342^2) - \frac{(2426)^2}{12} - \frac{(2093)^2}{12} = 38{,}940.08.$$

This within-group (or subject within-sequence) SS has 10 d.f., 5 from each group. The mean square is $38{,}940/10 = 3894$.

3. Test the sequence effect by comparing the sequence mean square to the within-group mean square (F test).

$$F_{1,10} = \frac{4620.375}{3894} = 1.19$$

Referring to Table IV.6, the effect is not significant at the 5% level. (Note that in practice, this test is performed at the 10% level.) If the sequence (carryover) effect is not significant, one would proceed with the usual analysis and interpretation as shown in Table 11.12.

Table 11.12 Analysis of Variance Table for the Crossover Bioequivalence Study (AUC) Without Sequence Effect

Source	**d.f.**	**SS**	**MS**	***P***
Subjects	11	43,560.5	3960.0	
Period	1	13,490.0	13,490.0	
Treatment	1	10,710.4	10,710.4	$F_{1,10} = 12.6^*$
Error	10	10,670.1	1067.0	
Total	23	78,430.96		$F_{1,10} = 10.0^*$

$^*p < 0.05$.

If the sequence (carryover) effect is significant, the usual analysis is not valid. The recommended analysis uses only the first period results, deleting the data contaminated by the carryover, the second period results. Grizzle recommends that the preliminary test for carryover be done at the 10% level (see also the discussion by Brown [7]). For the sake of this discussion, we will compute the analysis as if the data revealed a significant sequence effect in order to show the calculations. Using only the first-period data, the analysis is appropriate for a one-way ANOVA design (sect. 8.1). We have two "parallel" groups, one on Product *A* and the other on Product *B*. The data for the first period are as follows:

Subject	*A*	Subject	*B*
1	290	2	163
3	187	5	220
4	168	6	133
7	294	8	190
10	250	9	168
11	293	12	188
Mean	247		177
S^2	3204.8		870.4

The ANOVA table is as follows:[§]

	d.f.	SS	MS	*F*
Between treatments	1	14,700	14,700	7.21
Within treatments	10	20,376	2037.6	

Referring to Table IV.6, an *F* value of 4.96 is needed for significance at the 5% level (1 and 10 d.f.). Therefore, in this example, the analysis leads to the conclusion of significant treatment differences.

The discussion and analysis above should make it clear that sequence or carryover effects are undesirable in crossover experiments. Although an alternative analysis is available, one-half of the data are lost (second period) and the error term for the comparison of treatments is usually larger than that which would have been available in the absence of carryover (within-subject versus between-subject variation). One should thoroughly understand the nature of treatments in a crossover experiment in order to avoid differential carryover effects if at all possible. (Note: Although at one time the presence of a sequence effect could cause rejection of a BE submission by FDA, at the present time if there are no circumstances that could cause carryover, the FDA review would take this into consideration as a spurious event.)

Since the test for carryover was set at 5% a priori, we will proceed with the interpretation, assuming that carryover effects are absent. (Again, note that this test is usually set at the 10% level in practice). Both period and treatment effects are significant ($F_{1,10} = 12.6$ and 10.0, respectively). The AUC values tend to be higher during the first period (on the average). This period (or order) effect does not interfere with the conclusion that Product *A* has a higher average AUC than that of Product *B*. The balanced order of administration of the two products in this design compensates equally for both products for systematic differences due to the period or order. Also, the ANOVA subtracts out the SS due to the period effect from the error term, which is used to test treatment differences.

If the design is not symmetrical, because of missing data, dropouts, or poor planning, a statistician should be consulted for the data analysis and interpretation. In an asymmetrical design, the number of observations in the two periods is different for the two treatment groups. This will always occur if there is an odd number of subjects. For example, the following scheme shows an asymmetrical design for seven subjects taking two drug products, *A* and *B*. In such situations, computer software programs can be used, which adjust the analysis and mean results for the lack of symmetry [13].

[§] This analysis is identical to a two-sample independent groups *t* test.

Subject	Period 1	Period 2
1	*A*	*B*
2	*B*	*A*
3	*A*	*B*
4	*B*	*A*
5	*A*	*B*
6	*B*	*A*
7	*A*	*B*

The complete ANOVA is shown in Table 11.13.

The statistical analysis in the example above was performed on AUC, which is a measure of relative absorption. The FDA recommends that plasma or urine concentrations be determined out to at least three half-lives, so that practically all the area under the curve will be included when calculating this parameter (by the trapezoidal rule, for example). Other measures of the rate and extent of absorption are time to peak and peak concentration. Often, more than one analyte is observed, for example, metabolites or multiple ingredients.

Much has been written and discussed about the expression and interpretation of bioequivalency/bioavailability data as a measure of rate and extent of absorption. When are the parameters AUC, t_p, and C_{max} important, and what part do they play in bioequivalency? The FDA has stated that products may be considered equivalent in the presence of different rates of absorption, particularly if these differences are designed into the product [14]. For example, for a drug that is used in chronic dosing, the extent of absorption is probably a much more important parameter than the rate of absorption. It is not the purpose of this presentation to discuss the merits of these parameters in evaluating equivalence, but only to alert the reader to the fact that BE interpretation need not be fixed and rigid.

The ANOVA for log AUC (AUC values are transformed to their natural logs) is shown in Table 11.14. Exercise Problem 9 at the end of this chapter requests the reader to construct this table. The procedure is identical to that shown for the untransformed data. *Analysis of the log-transformed parameters is currently required by the FDA. The critical parameters are AUC and* C_{max}.

11.4.3 Confidence Intervals in BE Studies

The scientific community is virtually unanimous in its opposition to the use of hypothesis testing for the evaluation of BE. Hypothesis tests are inappropriate in that products that are very close, but with small variance, may be deemed different, whereas products that are widely different, but with large variance, may be considered equivalent (not significantly different). (See previous discussion in sect. 11.4.2). The use of a confidence interval, the present criterion for equivalence, is more meaningful and has better statistical properties. (See chap. 5 for a discussion of confidence intervals.) Given the lower and upper limit of the ratio of the parameters, the user or prescriber of a drug can make an educated decision regarding the equivalence of alternative products. The confidence limits must lie between 0.8 and 1.25 based on the difference of the back-transformed averages of the log-transformed AUC and C_{max} results. This computation for AUC is shown below. For historical purposes and purposes of comparison, the confidence interval is computed using the nontransformed data (the old method) and the log-transformed

Table 11.13 ANOVA for Untransformed Data from Table 11.10 for AUC

Variable (source)	d.f.	SS	MS	*F* ratio	Prob > *F*
Sequence	1	4620.4	4620.4	1.19	0.3016
Subject (sequence)	10	38,940.1	3894.0	3.65	0.0265
Period	1	13,490.0	13,490.0	12.64	0.0052
Treat	1	10,710.4	10,710.4	10.04	0.0100
Residual	10	10,670.1	1067.0		
Total	23	78,430.96			

Table 11.14 ANOVA for Log-Transformed Data from Table 11.10 for AUC

Variable (source)	d.f.	SS	MS	*F* ratio	Prob > *F*
Sequence	1	0.0613	0.0613	0.46	0.5128
Subject (sequence)	10	1.332	0.1332	2.96	0.0507
Period	1	0.4502	0.4502	10.02	0.0101
Treat	1	0.2897	0.2897	6.44	0.0294
Residual	10	0.44955	0.04496		
Total	23	2.58307			

data (the current method). Note that a ratio based on the untransformed data may be used in certain special circumstances where a log transformation may be deemed inappropriate, such as data derived from a clinical study, where the data consist of a pharmacodynamic response or some similar outcome. (See also, the currently recommended analysis using Locke's Method based on Fieller's Theorem below.)

11.4.3.1 Locke's Method of Analysis (Confidence Interval for the Ratio of Two Normally Distributed Variables)

The confidence interval for the ratio of two variables is described in "Guidance for Industry, Center for Drug Evaluation and Research, Appendix V, Feb 1997 [15]." The computations assume normality of the variables. The example uses data supplied in the FDA document referenced above.

If two variables are both normally distributed, it is not statistically valid to place a confidence interval on ratios. Ratios of normally distributed variables are not normal. For example, the data in the FDA document are as follows:

Subject	Test	Reference	Ratio
2	−48.52	−22.2	2.19
3	−38.99	−18.65	2.09
4	−7.62	−22.42	0.34
7	0.98	−10.96	−0.09
9	−32.05	−37.4	0.86
11	−26.18	−26.73	0.98
12	−11.62	−12.56	0.93

The average ratio is 1.04 with a s.d. of 0.84 (the reader may verify these calculations). The 90% confidence interval is $1.04 \pm 1.94 \times 0.84 \times \sqrt{1/7}$ = approximately 0.42 to 1.66.

This is not correct. The correct calculations are as follows:

Calculate the mean and variance of the test and reference products

Mean of test = $AV_t = -23.43$
Mean of reference = $AV_r = -21.56$
Variance test = $\sigma^2{}_T = 323.13$
Variance reference = $\sigma^2{}_R = 80.10$.

Since the two variables are related or correlated (crossover design), calculate the covariance $= = \sigma_{TR} = \sum (t - AV_t)(r - AV_r)(N - 1)$, where N = sample size = 7 and covariance = 78.83.

A variable is defined as "*G*," where G must be greater than zero in order for the calculations to be valid.

$$G = \frac{(t^2\sigma_R^2)}{(N \times AV_r^2)},$$

where N = sample size = 7

$$= \frac{(1.9432 \times 80.10)}{(7 \times 21.56^2)}$$
$$G = 0.093.$$

Then, apply the following formulas to calculate the confidence interval. (Note the similarity of these equations to the inverse equation to calculate the confidence interval for X, given Y in regression. This is a similar application of the calculation of a confidence interval for the ratio of two normal variables.)

$$K = \{AV_t^2/AV_r^2\} + \{\sigma_2^T/\sigma_R^2\}(1 - G) + \{\sigma_{TR}/\sigma_R^2\}[G(\sigma_{TR}/\sigma_R^2) - 2(AV_t/AV_r)]$$
$$= \{-23.43/-21.56\}^2 + \{323.13/80.1\}(1 - 0.093) +$$
$$\{78.83/80.1\}[0.093(78.3/80.1) - 2(-23.43/-21.56)]$$
$$K = 2.791.$$

Finally, calculate the 90% confidence interval (t = 1.943) as follows:

$$[(AV_t/AV_r) - G(\sigma_{TR}/\sigma_R^2)] \pm [(t/AV_r)\sqrt{\sigma_R^2 K/N}]/(1 - G).$$

$$= [(-23.43/21.56) - 0.0929(78.83/80.1)] \pm [(1.943/21.56) \text{ sqrt } (80.1 \times 2.791/7)]/(1 - 0.0929).$$

The 90% confidence interval is approximately 54% to 166%.

11.4.3.2 Nontransformed Data

The following discussion refers to the approach to the analysis of confidence intervals for BE *prior to the present use of the logarithmic transformation. See also, above, the preferred method using Fieller's (Locke) Theorem.*

90% confidence interval for AUC difference for data in Table 11.10

$$= \bar{\Delta} \pm t\sqrt{\text{EMS}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$42.25 \pm 1.81\sqrt{\frac{1067}{6}} = 42.25 \pm 24.14 = 18.11 \text{ to } 66.39,$$

where 42.25 is the average difference of the AUCs, 1.81 the t value with 10 d.f., 1067 the variance estimate (Table 11.12), and $1/6 = 1/N_1 + 1/N_2$. The confidence interval can be expressed as an approximate percentage relative bioavailability by dividing the lower and upper limits for the AUC difference by the average AUC for Product B, the reference product as follows:

Average AUC for drug Product $B = 167.2$

Approximate 90% confidence interval for A/B

$= (167.2 + 18.11)/167.2$ to $(167.2 + 66.39)/167.2$

$= 1.108$ to 1.397.

Product A is between 11% and 40% more bioavailable than Product B. The ratio formed for the nontransformed data, as shown in the example above, has random variables in both the numerator and denominator. The denominator (the average value of the reference) is considered fixed in this calculation, when, indeed, it is a variable measurement. Also, the decision rule is not symmetrical with regard to the average results for the test and reference. That is, if the reference is 20% greater than the test, the ratio test/reference is not 0.8 but is $1/1.2 = 0.83$. Conversely, if the test is 20% greater than the reference, the ratio will be 1.2. Nevertheless, at one time this approximate calculation was considered satisfactory for the purposes of assessing BE. Note that the usual concept of power does not play a part in the approval process. It behooves

the sponsor of the BE study to recruit sufficient number of subjects to help ensure approval based on this criterion. If the products are truly equivalent (the ratio of test/reference is truly between 0.8 and 1.2), the more subjects recruited, the greater the probability of passing the test. Note again that in this scenario the more subjects, the better the chance of passing. In practice, one chooses a sample size sufficiently large to make the probability of passing reasonably high. This probability may be defined as power in the context of proving equivalence. Sample size determination for various assumed differences between the test and reference products for various values of power (probability of passing the confidence interval criterion) has been published by Diletti et al. [20] (see Table 6.5).

The conclusions based on the confidence interval approach are identical to two one-sided t tests each performed at the 5% level [12,17]. The null hypotheses are

$$H_0 : \frac{A}{B} < 0.8 \quad \text{and} \quad H_0 : \frac{A}{B} > 1.25.$$

Note that with the log transformation, the upper limit is set at 1.25 instead of 1.2. This results from the properties of logarithms, where log (0.8) = −log (1/0.8). If both tests are rejected, the products are considered to have a ratio of AUC and/or C_{max} between 0.8 and 1.25 and are taken to be equivalent. If either hypothesis (or both) is not rejected, the products are not considered to be equivalent.

The test product (A in Table 11.10) would not pass the FDA equivalency test because the upper confidence limit exceeds 1.25. For the two one-sided t tests, we test the observed difference versus the hypothetical difference needed to reach 80% and 125% of the standard product.

If the test product had an average AUC of 175 and the error were 1067, the product would pass the FDA criterion using the "old" method. The 90% confidence limits would be

$$175 - 167.2 \pm 1.81\sqrt{\frac{1067}{6}} = -16.34 \text{ to } 31.94.$$

The 90% confidence limits for the ratio of the AUC of test product/standard product are calculated as

$$\frac{(167.2 - 16.34)}{167.2} = 0.902$$

$$\frac{(167.2 + 31.94)}{167.2} = 1.191.$$

The limits are within 0.8 and 1.25.
The two one-sided t tests are

$$H_0 : \frac{A}{B} < 0.8 \quad t = \frac{175 - 167.2 - [-33.4]}{\sqrt{1067/6}} = 3.09$$

$$H_0 : \frac{A}{B} > 1.25 \quad t = \frac{175 - 167.2 - [41.8]}{\sqrt{1067/6}} = 2.55,$$

where −33.4 represents 20% and 41.8 represents 25% of the reference.¶ Since both t values exceed 1.81, the table t for a one-sided test at the 5% level, the products are deemed to be equivalent.

¶ The former FDA criterion for the confidence interval was 0.8 to 1.20 based on nontransformed data. Therefore this presentation is hypothetical.

Westlake has discussed the application of a confidence interval that is symmetric about the ratio 1.0, the value that defines equivalent products. The construction of such an interval is described in section 5.1.

11.4.3.3 Log-Transformed Data (Current Procedure)

The log transform appears to be more natural when our interest is in the ratio of the product outcomes. The antilog of the difference of the average results gives the ratio directly [18].

Note that the *difference* of the logarithms is equivalent to the logarithm of the *ratio* [i.e., $\log A - \log B = \log (A/B)$]. The antilog of the average difference of the logarithms is an estimate of the ratio of AUCs.

The ANOVA for the ln-transformed data is shown in Table 11.14.

The averages ln values for the test and standard products are

$$\overline{A} = 5.29751$$
$$\overline{B} = 5.07778$$
$$\overline{A} - \overline{B} = 5.29751 - 5.0778 = 0.21973.$$

The anti-ln of this difference, corresponding to the geometric mean of the individual ratios, is 1.246. This compares to the ratio of A/B for the untransformed values of 1.252.

$$0.21973 \pm 1.81\sqrt{0.045/6} = 0.06298 \text{ to } 0.37648.$$

The anti-ln of these limits are 1.065 to 1.457. The 90% confidence limits for the untransformed data are 1.108 to 1.397.

It is not surprising that both analyses give similar results and conclusions. However, in situations where the confidence interval is close to the lower and/or upper limits, the two analyses may result in different conclusions. A nonparametric approach has been recommended (but is not currently accepted by the FDA) if the data distribution is far from normal (see chap. 15). As discussed earlier, at one time, the FDA suggested an alternative criterion for proof of BE: at least 75% of the subjects should show the availability for a test product compared to the reference or standard formulation to be between 75% and 125%. This is called the *75/75 rule.* If 75% of the population *truly* shows at least 75% relative absorption of the test formulation compared to the standard, a sample of subjects in a clinical study will have a 50% chance of failing the test based on the FDA criterion. This criterion has little statistical basis and has fallen into disrepute. The concept of individual BE (sect. 11.4.6) is concerned with assessing the equivalence of products on an individual basis based on a more statistically based criterion.

11.4.4 Sample Size and Highly Variable Drug Products

Phillips [19] published sample sizes as a function of power, product differences, and variability. Diletti et al. [20] have published similar tables where the log transformation is used for the statistical analysis. These tables are more relevant to current practices. Table 6.4 shows sample sizes for the multiplicative (log-transformed) analysis, reproduced from the publication by Diletti. This table as well as more details on sample size estimation is given in section 6.5. (See also Excel program on the accompanying disk to calculate sample size under various assumptions.)

When the variation is large because of inherent biologic variability in the absorption and/or disposition of the drug (or due to the nature of the formulation), large sample sizes may be needed to meet the confidence interval criterion. Generally, using results of previous studies, one can estimate the within-subject variability from the residual error term in the ANOVA. This can be assumed to be the average of the within-subject variances of the two products. These variances cannot be separated in a two-period crossover design, nor can the variability be separately attributed to the drug itself or to the formulation effects. Thus, the variability estimate is some combination of both the drug and the formulation variances. A drug product is considered to be highly variable if the error variance shows a coefficient of variation (CV) of 30% or greater. There are many drug products that show such variability. CV's of 100% or more

have been observed on occasion. To show equivalence for highly variable drug products, using the FDA criterion of a 90% confidence interval of parameter ratios of 0.8 to 1.25 requires a very large sample size.

For example, from Table 6.5, if the CV is 30% and the products differ by only 5%, a sample size of 40 is needed to have 80% power to show the products are equivalent. The FDA has been considering the problems of designing studies and interpreting results for variable drugs and/or drug products. This problem has been debated for some time, and a few recommendations have been proposed to deal with this problem. Although there is no single solution, possible alternatives include widening of the confidence interval criterion from 0.8 to 1.25 to 0.75 to 1.33 [21] and use of replicated or sequential designs. The European Agency for the Evaluation of Medicinal Products also makes provision for a wider interval provided it is prospectively defined and can be justified accordingly [22]. Another recommendation by Endrenyi [23] is to scale the ratio using the reference CV as the scaling factor. At the time of this writing, the FDA has published a guidance that includes a scaled analysis. This approach may be recommended for BE studies of highly variable products. This scaled analysis is described below. Individual BE in a replicate design to assess BE is also supposed to result in smaller sample sizes for highly variable drug products as compared to the corresponding two-period design. This solution to the problem is yet to be fully confirmed. Currently, products with large CVs require large studies, with an accompanying increased expense. Because these highly variable drugs have been established as safe and effective and have a history of efficacy and safety in the marketplace, increasing the confidence interval would be congruent with the drug's variability in practice. Scaled BE may provide an economical way of evaluating these drug products.

Note that for the determination of BE based on the final study results, power (computed a posteriori) plays no role in the determination of equivalence. However, to estimate the sample size needed before initiating the study, power is an important consideration. The greater the power one wishes to impose, where power is the probability of passing the 0.8 to 1.25 confidence interval, the more subjects will be needed. Usually, a power of 0.8 is used to estimate sample size. However, if cost is not important (or not excessive), a greater power (0.9, for example) can be used to gain more assurance of passing the study, assuming that the products are truly bioequivalent.

Equation (11.3) can be used to approximate the sample size needed for a specified power.

$$N = 2(t_{\alpha,\,2N-2} + t_{\beta,\,2N-2})^2 \left[\frac{\mathrm{CV}}{(\mathrm{V}-\delta)}\right]^2, \tag{11.3}$$

where N is the total number of subjects required to be in the study; t the appropriate value from the t distribution (approximately 1.7); α the significance level (usually 0.1); $1-\beta$ the power, usually 0.8; CV the coefficient of variation; V the BE limit ($\ln 1.25 = 0.223$); and δ the difference between the products (for 5% difference, delta equals [$\ln(1.05) = 0.0488$]).

If we assume a 5% difference between the products being compared, the number of subjects needed for a CV of 30% and power of 0.8 is: $N = 2\,(1.7 + 0.86)^2\,[0.3/(0.223 - .0488)]^2 =$ approximately 39 subjects, which is close to the 40 subjects from Table 6.5.

If the CV is 50%, we need approximately 108 subjects!

$$N = 2(1.7 + 0.86)^2 \left(\frac{0.5}{0.223 - 0.0488}\right)^2 = \text{approximately 108 subjects.}$$

It can be seen that with a large CV, studies become inordinately large.

BE studies are usually performed at a single site, where all subjects are recruited and studied as a single group. On occasion, more than one group is required to complete a study. For example, if a large number of subjects are to be recruited, the study site may not be large enough to accommodate the subjects. In these situations, the study subjects may be divided into two cohorts. Each cohort is used to assess the comparative products individually, as might be done in two separate studies. Typically, the two cohorts are of approximately equal size. The final assessment is based on a combination of both groups. The totality of data is analyzed with

a new term in the ANOVA, a Treatment-by-Group interaction term.** This is a measure (on a log scale) of how the ratios of test to reference differ in the groups. For example, if the ratios are very much the same in each group, the interaction would be small or negligible. If interaction is large, as tested in the ANOVA, then the groups statistically should not be combined. However, if at least one of the groups individually passes the confidence interval criteria, then the test product might be acceptable. If interaction is not statistically significant ($p > 0.10$), then the confidence interval based on the pooled analysis, after dropping the interaction term, will determine acceptability. It is an advantage to pool the data, as the larger number of subjects increases power and there is a greater probability of passing the BE confidence interval, if the products are truly bioequivalent.

An interesting question arises if more than two groups are included in a BE study. As before, if there is no interaction, the data should be pooled. If interaction is evident, it is implied that at least one group is different from the others. Usually, it will be obvious which group is divergent from a visual inspection of the treatment differences in each group. The remaining groups may then be tested for interaction. Again, as before, if there is no interaction, the data should be pooled. If there is interaction, the aberrant group may be omitted, and the remaining groups tested, and so on. In rare cases, it may not be obvious which group or groups are responsible for the interaction. In that case, more statistical treatment may be necessary, and a statistician should be consulted. In any event, if any single group or pooled groups (with no interaction) passes the BE criteria, the test should pass. If a pooled study passes in the presence of interaction, but no single study passes, one may still argue that the product should pass, if there is no apparent reason for the interaction. For example, if the groups are studied at the same location under the identical protocol, and there is overlap in time among the treatments given to the different groups, as occurs often, there may be no obvious reason for a significant interaction. Perhaps, the result was merely due to chance. One may then present an argument for accepting the pooled results.

The following statistical models have been recommended for analysis of data in groups:

Model 1: GRP SEQ GRP∗SEQ SUBJ(GRP∗SEQ) PER(GRP) TRT GRP∗TRT.
If the GRP∗TRT term is not significant ($p > 0.10$), then reanalyze the data using Model 2.
Model 2: GRP SEQ GRP∗SEQ SUBJ(GRP∗SEQ) PER(GRP) TRT,

where GRP is the group, SEQ the sequence, GRP∗SEQ the group-by-sequence, SUBJ(GRP∗SEQ) the subject nested within group-by-sequence, PER(GRP) the period nested within group, TRT the treatment, and GRP∗TRT the group-by-treatment interaction.

11.4.5 Outliers in BE Studies

An outlier is an observation far removed from the bulk of the observations.

The problems of dealing with outlying observations is discussed in some detail in section 10.2. These same problems exist in the analysis and interpretation of BE studies. Several kinds of outliers occur in BE studies. Analytical outliers may occur because of analytical errors, and these can usually be rectified by reanalyzing the retained blood samples. Another kind of outlier is a value that does not appear to fit the PK profile. If repeat analyses verify these values, one has little choice but to retain these values in the analysis. If such values appear rarely, they will usually not affect the overall conclusions since the individual results are a small part of the overall average results, such as in the calculation of AUC. An exception may occur if the aberrant value occurs at the time of the estimated C_{max}, where the outlier could be more influential. The biggest problem with outliers is when the outlier arises from a derived parameter (AUC or C_{max}) for an individual subject. The current FDA position is to disallow the exclusion of an outlier from the analysis solely on a statistical basis. However, if a clinical reason can be determined as a potential cause for the outlier and when the outlier appears to be due to the reference product, an outlier may be omitted from the analysis at the discretion of the FDA. The FDA also suggests

** Currently, FDA requires this only when groups are not from the same population or are dosed widely separated in time.

that the outlier be retested in a sample of 6 to 10 subjects from the original study to support the anomalous nature of the suspected outlier. Part of the reasoning for not excluding outliers is that one or two individual outliers suggest the possibility of a subpopulation that shows a difference between the products. Although theoretically possible, this author's opinion is that this is a highly unlikely event without definitive documentation. Also, using this reasoning, an outlying observation due to the reference product would suggest that the reference did not act uniformly among patients, suggesting a deficiency in the reference product. Another possible occasion for discarding an individual subject's result is the case where very little or no drug is absorbed. Explanations for this effect could be product-related or subject-related, but the true cause is unlikely to be known. Zero blood levels, in the absence of corroborating evidence for product failure, are most likely due to a failure of the subject. These problems remain controversial and should be dealt with on a case-by-case basis.

A more creative approach is possible in the case of replicate designs (see below). In these situations, the estimates of within-subject variability can be used to identify outliers. For example, if the within-subject variance for a given treatment is 0.61, but reduces to 0.04 when omitting the subject with the suspected outlier value, an F test can be performed comparing variances for the suspect data and the remaining data. The F ratio, in this example, is

$$F = \frac{0.61}{0.04} = 15.3.$$

The d.f. for the numerator are those for the variance estimate obtained using the results from all subjects and those for the denominator are those for the variance estimate obtained from the results omitting the suspected outlier. In the above example, if the numerator and denominator d.f. were 30 and 28, respectively, then an F value of 15.3 is highly significant ($p < 0.01$). An alternative analysis could be an ANOVA with and without the suspected outlier. An F test with 1 d.f. in the numerator and appropriate d.f. in the denominator would be:

[SS (all data) − SS (without outlier data)]/residual SS (all data)<

Another approach that has been used is to compare results for periods 1 and 2 versus periods 3 and 4 in a four-period fully replicated design.

Of course, if there is an obvious cause for the outlier, a statistical justification is not necessary. However, further evidence, even if only suspicious, is helpful.

If an outlier is detected, as noted above, the most conservative approach is to find a reason for the outlying observation, such as a transcription error, or an analytical error, or a subject that violated the protocol, and so on. In these cases, the data may be reanalyzed with the corrected data, or without the outlying data if due to analytical or protocol violation, for example.

If an obvious reason for the outlier is not forthcoming, one may wish to perform a new small study, replicating the original study, including the outlying subject along with a number of other subjects (at least five or six) from the original study. The results from the new study can be examined to determine if the data for the outlier from the original study are anomalous. It should be noted that the data from the small study are not used as a replacement for any of the original data, but serve only to confirm, or refute, that the suspected outlier subject is reproducibly an outlier. The procedure here is not fixed, but should be reasonable, and make sense. One can compare the test to reference ratios for the outlying subject in the two studies, and demonstrate that the data from the new study show that the outlying subject is congruent with the other subjects in the new study, for example.

11.4.6 Replicate Designs for BE Studies**

Replicate crossover designs may be defined as designs with more than two periods where products are given on more than one occasion. In the present context such replicate studies are studies in which individuals are administered one or both products on more than one occasion. FDA gives sponsors the option of using replicate design studies. Replicate studies can isolate

**A more advanced topic.

the within-subject variance of each product separately, as well as potential product-by-subject interactions.

The FDA recommends that submissions of studies with replicate designs be analyzed for average BE. The following (Table 11.15) is an example of the analysis of a two-treatment–four period replicate design to assess average BE. The design has each of two products, balanced in two sequences, *ABAB* and *BABA*, over four periods. Table 11.16 shows the results for C_{max} for a replicate study. Eighteen subjects were recruited for the study and 17 completed the study. An analysis using the usual approach for the two-treatment, two-period design, as discussed above, is not recommended. The FDA recommends use of a mixed model approach as in SAS PROC MIXED [9]. The recommended code is

```
PROC MIXED;

CLASSES SEQ SUBJ PER TRT;

MODEL LNCMAX = SEQ PER TRT/DDFM = SATTERTH;

RANDOM TRT/TYPE = FAO (2) SUB = SUBj G;

REPEATED/GRP = TRT SUB = SUBJ;

LSMEANS TRT;

ESTIMATE "T VS. R" TRT 1 − 1/CL ALPHA = 0.1;

RUN;
```

We will concentrate on the comparison of two products in three- or four-period designs. The FDA recommends using only two sequence designs because the interaction variability estimate, subject × formulation, will be otherwise confounded (see Ref. 24 for a comparison of the 2 and 4 sequence designs). The subject × formulation interaction is crucial because if this effect is substantial, the implication is that subjects do not differentiate formulations equally, that is, some subjects may give higher results for one formulation, and other subjects respond higher on the other formulation. Two sequence designs for three- and four-period studies are shown below. Although there are other designs available, these seem to have particularly good properties [16,24].

Three-period design	
Sequence	Period 1 2 3
1	*A B B*
2	*B A A*

Four-period design	
Sequence	Period 1 2 3 4
1	*A B B A*
2	*B A A B*

With replicate designs, carryover effects, within-subject variances and subject × formulation interactions can be estimated, unconfounded with other effects. Nevertheless, an unambiguous acceptable analysis is still not clear. Do we include all the effects in the model simultaneously or do we perform preliminary tests for inclusion in the model? What is the proper error term to construct a confidence interval on the average BE parameter (e.g., AUC)? Some estimates may

not be available if all terms are included in the model. Therefore, preliminary testing may be necessary. These questions are not easy to answer and, despite their advantages, make the use of replicate designs problematic at the time of this writing.

The following is one way of proceeding with the analysis: Test for differential carryover. This term may be included in the model (along with the usual parameters) using a dummy variable, that is, 0 if treatment in Period 1, if Treatment B follows Treatment A, and 2 if Treatment A follows Treatment B. If differential carryover is not significant, remove it from the model. Include a term for subject $\times$ formulation interaction, and if this effect is large, the products may be considered bioinequivalent (see sect. 11.4.6.1). Another problem that arises here is concerned with what error term should be used to construct the confidence interval for the average difference between formulations. The choices are among the within-subject variance (residual), the interaction term, or the residual with no interaction term in the model (pooled residual and interaction). The latter could be defended if the interaction term is small or not significant.

The analysis of studies with replicate designs would be very difficult without access to a computer program. Using SAS GLM, the following program can be used. (See below for FDA recommended approach.)

```
proc glm;
class sequence subject product period co;
model auc = period subject (sequence) product co;
lsmeans product/stderr;
estimate 'test-ref'product -11;
```

∗co is carryover∗

Using the data from Chow and Liu [16], a four-period design with nine subjects completing the study, the SAS output is as follows:

Dependent variable: AUC

Source	d.f.	Sum of squares	Mean square	*F* value	Pr > *F*
Model	13	40895.72505	3145.82500	8.25	0.0001
Error	22	8391.03801	381.41082		
Corrected total	35	49286.76306			

Dependent variable: AUC

Source	d.f.	Type I SS	Mean square	*F* value	Pr > *F*
SEQ	1	9242.13356	9242.13356	24.23	0.0001
SUBJECT (SEQ)	7	25838.61700	3691.23100	9.68	0.0001
PRODUCT	1	1161.67361	1161.67361	3.05	0.0949
PERIOD	3	4650.60194	1550.20065	4.06	0.0193
CO	1	2.69894	2.69894	0.01	0.9337

Source	d.f.	Type III SS	Mean square	*F* value	Pr > *F*
SEQ	1	8311.37782	8311.37782	21.79	0.0001
SUBJECT (SEQ)	7	25838.61700	3691.23100	9.68	0.0001
PRODUCT	1	975.69000	975.69000	2.56	0.1240
PERIOD	2	2304.85554	1152.42777	3.02	0.0693
CO	1	2.69894	2.69894	0.01	0.9337

Parameter	Estimate	T for HO:	Pr > |T|	Std error of parameter estimate
test-ref	−10.98825000	−1.60	0.1240	6.87019569

Because carryover is not significant ($p > 0.9$), we can remove this term from the model and analyze the data with a subject $\times$ formulation (within sequence) term included in the model. The SAS output is as follows:

General linear models procedure Dependent variable: AUC					
Source	d.f.	Sum of squares	Mean squares	*F* value	Pr > *F*
Model	19	42490.87861	2236.36203	5.27	0.0008
Error	16	6795.88444	424.74278		
Corrected total	35	49286.76306			
Source	d.f.	Type I SS	Mean square	*F* value	Pr > *F*
SEQ	1	9242.13356	9242.13356	21.76	0.0003
SUBJECT (SEQ)	7	25838.61700	3691.23100	8.69	0.0002
PRODUCT	1	1161.67361	1161.67361	2.74	0.1177
PERIOD	3	4650.60194	1550.20065	3.65	0.0354
SUBJECT * PRODUCT(SEQ) (SEQ)	7	1597.85250	228.26464	0.54	0.7940
Source	d.f.	Type III SS	Mean square	*F* value	Pr > *F*
SEQ	1	9242.13356	9242.13356	21.76	0.0003
SUBJECT (SEQ)	**7**	25838.61700	3691.23100	8.69	0.0002
PRODUCT	1	1107.56806	1107.56806	2.61	0.1259
PERIOD	2	4622.20056	2311.10028	5.44	0.0157
SUBJECT * PRODUCT (SEQ)	7	1597.85250	228.26464	0.54	0.7940

The subject × product interaction is not significant ($p > 0.7$). Again the question of which error term to use for the confidence interval is unresolved. The choices are (a) interaction = 228, within-subject variance = 425, or pooled residual = 365. The d.f. will also differ depending on the choice. The simplest approach seems to be to use the pooled variance if the interaction term is not significant (the level must be defined). If interaction is significant, use the interaction term as the error. In the example given above, the analysis without interaction and carryover may be appropriate (also see sect. 11.4.6.1). The following analysis has an error term equal to 365.

Dependent variable: AUC **Source**	**d.f.**	**Sum of squares**	**Mean square**	***F* value**	**Pr > *F***
Model	12	40893.02611	3407.75218	9.34	0.0001
Error	23	8393.73694	364.94508		
Corrected total	35	49286.76306			
Source	d.f.	Type III SS	Mean square	*F* value	Pr > *F*
SEQ	1	9242.13356	9242.13356	25.32	0.0001
SUBJECT (SEQ)	7	25838.61700	3691.23100	10.11	0.0001
PRODUCT	1	1107.56806	1107.56806	3.03	0.0949
PERIOD	3	4650.60194	1550.20065	4.25	0.0158
	PRODUCT	AUC LSMEAN	Std err LSMEAN	Pr > |T| HO : LSMEAN = 0	
	1	87.7087500	4.5308014	0.0001	
	2	76.5462500	4.5308014	0.0001	
Parameter		Estimate	T for HO: Parameter = 0	Pr > |T|	Std error of estimate
test-ref		−11.16250000	−1.74	0.0949	6.40752074

The complete analysis of replicate designs can be very complex and ambiguous, and is beyond the scope of this book. An example of the analysis as recommended by the FDA is shown later in this section. For an in-depth discussion of the analysis of replicate designs including estimation of sources of variability (see Refs. [16,24,25]).

The four-period design will be further discussed in the discussion of individual bioequivalence (IB), for which it is recommended. In a relatively recent guidance, the FDA [10] gives sponsors the option of using replicate design studies for all BE studies. However, at the time of this writing, the agency has ceased to recommend use of replicate studies although they may be useful in some circumstances. The purpose of these studies was to provide more information about the drug products than can be obtained from the typical, nonreplicated, two-period

design. The FDA was interested in obtaining information from these studies to aid them in evaluation of the need for IB. In particular, replicate studies provide information on within-subject variance of each product separately, as well as potential product × subject interactions. As noted previously, the use of these designs and assessment of IB have been controversial, and its future in its present form is in doubt.

The FDA recommends that submissions of studies with replicate designs be analyzed for average BE [10]. Any analysis of IB will be the responsibility of the FDA, but will be only for internal use, not for evaluating BE for regulatory purposes.

The following is another example of the analysis of a two-treatment–four-period replicate design to assess average BE, as recommended by the FDA. This design has each of two products,

Table 11.15 Results of a Four-Period, Two-Sequence, Two-Treatment, Replicate Design (C_{max})

Subject	Product	Sequence	Period	C_{max}	Ln(C_{max})
1	Test	1	1	14	2.639
2	Test	1	1	16.7	2.815
3	Test	1	1	12.95	2.561
4	Test	2	2	13.9	2.632
5	Test	1	1	15.6	2.747
6	Test	2	2	12.65	2.538
7	Test	2	2	13.45	2.599
8	Test	2	2	13.85	2.628
9	Test	1	1	13.05	2.569
10	Test	2	2	17.55	2.865
11	Test	1	1	13.25	2.584
12	Test	2	2	19.8	2.986
13	Test	1	1	10.45	2.347
14	Test	2	2	19.55	2.973
15	Test	2	2	22.1	3.096
16	Test	1	1	22.1	3.096
17	Test	2	2	14.15	2.650
1	Test	1	3	14.35	2.664
2	Test	1	3	22.8	3.127
3	Test	1	3	13.25	2.584
4	Test	2	4	14.55	2.678
5	Test	1	3	13.7	2.617
6	Test	2	4	13.9	2.632
7	Test	2	4	13.75	2.621
8	Test	2	4	13.25	2.584
9	Test	1	3	13.95	2.635
10	Test	2	4	15.15	2.718
11	Test	1	3	13.15	2.576
12	Test	2	4	21	3.045
13	Test	1	3	8.75	2.169
14	Test	2	4	17.35	2.854
15	Test	2	4	18.25	2.904
16	Test	1	3	19.05	2.947
17	Test	2	4	15.1	2.715
1	Reference	1	2	13.5	2.603
2	Reference	1	2	15.45	2.738
3	Reference	1	2	11.85	2.472
4	Reference	2	1	13.3	2.588
5	Reference	1	2	13.55	2.606
6	Reference	2	1	14.15	2.650
7	Reference	2	1	10.45	2.347
8	Reference	2	1	11.5	2.442
9	Reference	1	2	13.5	2.603
10	Reference	2	1	15.25	2.725

(*Continued*)

Table 11.15 Results of a Four-Period, Two-Sequence, Two-Treatment, Replicate Design (C_{max}) *Continued*

11	Reference	1	2	11.75	2.464
12	Reference	2	1	23.2	3.144
13	Reference	1	2	7.95	2.073
14	Reference	2	1	17.45	2.859
15	Reference	2	1	15.5	2.741
16	Reference	1	2	20.2	3.006
17	Reference	2	1	12.95	2.561
1	Reference	1	4	13.5	2.603
2	Reference	1	4	15.45	2.738
3	Reference	1	4	11.85	2.472
4	Reference	2	3	13.3	2.588
5	Reference	1	4	13.55	2.606
6	Reference	2	3	14.15	2.650
7	Reference	2	3	10.45	2.347
8	Reference	2	3	11.5	2.442
9	Reference	1	4	13.5	2.603
10	Reference	2	3	15.25	2.725
11	Reference	1	4	11.75	2.464
12	Reference	2	3	23.2	3.144
13	Reference	1	4	7.95	2.073
14	Reference	2	3	17.45	2.859
15	Reference	2	3	15.5	2.741
16	Reference	1	4	20.2	3.006
17	Reference	2	3	12.95	2.561

balanced in two sequences, *ABAB* and *BABA*, over four periods. Table 11.15 shows the results for C_{max} for a replicate study. Eighteen subjects were recruited for the study and 17 completed the study. An analysis using the usual approach for the tttp design, as discussed above, is not recommended. The FDA [10] recommends use of a mixed model approach as in SAS PROC MIXED [13]. The recommended code is

```
PROC MIXED;
CLASSES SEQ SUBJ PER TRT;
MODEL LNCMAX = SEQ PER TRT/DDFM = SATTERTH;
RANDOM TRT/TYPE = FAO (2) SUB = SUBj G;
REPEATED/GRP = TRT SUB = SUBJ;
LSMEANS TRT;
ESTIMATE "T VS. R" TRT 1 − 1/CL ALPHA = 0.1;
RUN;
```

The abbreviated output is shown in Tables 11.16 and 11.17. Table 11.16 shows an analysis of the first two periods for ln (C_{max}) and untransformed C_{max}. Table 11.17 shows the output for the analysis of average BE using all four periods. Note that the confidence interval using the complete design (0.0592–0.1360) is not much different from that observed from the analysis of the first two periods (see Exercise at the end of the chapter), 0.0438, 0.1564. This should be expected because of the small variability exhibited by this product.

11.4.6.1 Individual Bioequivalence[††]

Another issue that has been introduced as a relevant measure of equivalence is "individual" bioequivalence (IB). This is in contrast to the present measure of "average" BE. Note that

[††] FDA has never accepted, nor currently endorses, this method, despite its having devoted resources to its development over a period >5 years. It is presented here due to its elegant statistical derivation from basic principles of drug interchangeability and its place in the history of bioequivalence testing in the U.S.

the evaluation of data from the tttp design results in a measure of average BE. Average BE addresses the comparison of average results derived from the tttp BE study, and does not consider differences of within-subject variance and interactions in the evaluation.

The IB approach is an attempt to evaluate the effect of changing products (brand to generic, for example) for an individual patient, considering the potential for a change of therapeutic effect

Table 11.16 ANOVA for Data from First Two Periods of Table 11.15

(A) LN TRANSFORMATION
Dependent variable: LNCMAX

Source	d.f.	Sum of squares	Mean square	*F* value	Pr > *F*
Model	18	1.65791040	0.09210613	10.34	0.0001
Error	15	0.13359312	0.00890621		
Corrected Total	33	1.79150352			

R square	CV	Root MSE	LNCMAX mean
0.925430	3.528167	0.09437271	2.67483698

Source	d.f.	Type I SS	Mean square	*F* value	Pr > *F*
SEQ	1	0.09042411	0.09042411	10.15	0.0061
SUBJ(SEQ)	15	1.48220203	0.09881347	11.09	0.000
PER	1	0.00039571	0.00039571	0.04	0.8359
TRT	1	0.08488855	0.08488855	9.53	0.0075

Least squares means

TRT	LNCMAX LSMEAN
Reference	2.62174427
Test	2.72185203

T for HO: Parameter	Pr > |T| Estimate	Std error of Parameter=0	Estimate	
T VS.R	0.10010777	3.09	0.0075	0.03242572

(B) Dependent variable: CMAX

Source	d.f	Sum of squares	Mean square	*F* value	Pr > *F*
Model	18	381.26362847	21.18131269	9.07	0.0001
Error	15	35.01637153	2.33442477		
Corrected total	33	416.28000000			

R square	CV	Root MSE	CMAX mean
0.915883	10.25424	1.52788245	14.90000000

Source	d.f.	Type I SS	Mean square	*F* value	Pr > *F*
SEQ	1	18.59404514	18.59404514	7.97	0.0129
SUBJ(SEQ)	15	346.22095486	23.08139699	9.89	0.0001
PER	1	0.24735294	0.24735294	0.11	0.7493
TRT	1	16.20127553	16.20127553	6.94	0.0188

Least squares means

TRT	CMAX LSMEAN
Reference	14.1649306
Test	15.5479167

Dependent variable: CMAX

T for HO: Parameter	Pr > |T| Estimate	Std Error of Parameter=0	Estimate	
T VS. R	1.38298611	2.63	0.0188	0.52496839

Table 11.17 Analysis of Data from Table 11.15 for Average Bioequivalence

ANALYSIS FOR LN-TRANSFORMED CMAX
The MIXED procedure
Class level information

Class	Concentrations values
SEQ	2 1 2
SUBJ	17 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
PER	4 1 2 3 4
TRT	2 12

Covariance parameter estimates (REML)

Cov Parm	Subject	Group	Estimate
FA(1,1)	SUBJ		0.20078553
FA(2,1)	SUBJ		0.22257742
FA(2,2)	SUBJ		−0.00000000
DIAG	SUBJ	TRT 1	0.00702204
DIAG	SUBJ	TRT 2	0.00982420

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
SEQ	1	13.9	1.02	0.3294
PER	3	48.2	0.30	0.8277
TRT	1	51.1	18.12	0.0001

ESTIMATE statement results

Parameter T VS. R

Alpha = 0.1	Estimate		Std error	d.f.	*t*	Pr > |*t*|
	0.09755781	0.02291789		51.1	4.26	0.0001
	Lower		0.0592	Upper	0.1360	

Least squares means

Effect	TRT	LSMEAN	Std Error	d.f.	t	Pr > |t|
TRT	1	2.71465972	0.05086200	15	53.37	0.0001
TRT	2	2.61710191	0.05669416	15.3	46.16	0.0001

or increased toxicity when switching products [38]. This is a very difficult subject from both a conceptual and statistical point of view. Statistical methods and meaningful differences must be established to show differences in variability between products before this criterion can be implemented. Whether or not a practical approach can be developed, and whether or not such approaches will have meaning in the context of BE remains to be seen. Some of the statistical problems to be contemplated when implementing this concept include recommendations of specific replicated crossover designs to measure both within- and between-variance components as well as subject × product interactions, and definitions of limits that have clinical meaning. The issue is related to variability. Assuming that the average bioavailability is the same for both products as measured in a typical BE study, the question of IB appears to be an evaluation of formulation differences. Since the therapeutic agents are the same in the products to be compared, it is formulation differences that could result in excessive variability or differences in bioavailability that are under scrutiny. Some of the dilemmas are related to the inherent biologic variability of a drug substance. If a drug is very variable, we would expect large variability in studies of interchangeability of products. In particular, taking the same product on multiple occasions would show a lack of "reproducibility." The question that needs to be addressed is whether the new (generic) product would cause efficacy failure or toxicity when exchanged with the reference or brand product due to excessive variability. The onus is on the generic product. Product failure could be due to a change in the rate and extent of drug absorption as well as an increase in inter- and intrapatient variability. The FDA has spent some energy in addressing

the problem of how to define and evaluate any changes incurred by the generic product. This is a difficult problem, not only in identifying the parameters to measure the variability, but also to define the degree of variability that would be considered excessive. For example, drugs that are very variable may be allowed more leniency in the criteria for "interchangeability" than less variable, narrow-therapeutic-range drugs.

The FDA has proposed an expression to define IB

$$\theta = \frac{[\delta^2 + \sigma_I^2 + (\sigma_T^2 - \sigma_R^2)]}{\sigma_R^2} \tag{11.4}$$

where δ is the difference between means of test and reference, σ_I^2 the subject $\times$ treatment interaction variance, σ_T^2 the within-subject test variance, and σ_R^2 the within-subject reference variance.

Equation 11.4 makes sense in that the comparison between test and reference products is scaled by the within-reference variance, thereby not penalizing very variable drug products when testing for BE. In addition, the expression contains a term for testing the mean difference, the interaction, and the difference between the within-subject variances. If the test product has a smaller within-subject variance than the reference, this favors the test product.

Before IB was to be considered a requirement from a regulatory point of view, data were accumulated from replicate crossover studies (three or more periods) to compile a database to assess the magnitude and kinds of intrasubject and formulation $\times$ subject variability that exist in various drug and product classes. The design and submission of such studies were more or less voluntary, and were analyzed for average BE. However, this gave the regulatory agency the opportunity to evaluate the data according to IB, and to evaluate the need for this new kind of criterion for equivalence. At the time of this writing, the FDA has rejected further development of this approach. The details of the design and analysis of these studies are presented below.

In summary, IB is an assessment that accounts for product differences in the variability of the PK parameters, as well as differences in their averages. IB evaluation is based on the statistical evaluation of the metric [Eq. (11.4)], which represents a "distance" between the products. In average BE, this distance can be considered the square of the difference in average results. In IB, in addition to the difference in averages, the difference between the within-subject variances for the two products, and the formulation $\times$ subject interaction (FS) are evaluated. In this section, we will not discuss the evaluation of population BE. The interested reader may refer to the FDA guidance [10].

The evaluation of IB is based on a 95% upper confidence limit on the metric, where the upper limit for approval, theta (θ), is defined as 2.4948. Note that we only look at the upper limit because the test is one-sided; that is, we are only interested in evaluating the upper value of the confidence limit, upon which a decision of passing or failing depends. A large value of the metric results in a decision of inequivalence. Referring to Eq. (11.4), a decision of inequivalence results when the numerator is large and the denominator is small in value. Large differences in the average results, combined with a large subject $\times$ formulation interaction, a large within-subject variance for the test product and a small within-subject variance for the reference product, will increase the value of theta (and vice versa).

Using the within-subject variance of the reference product in the denominator as a scaling device allows for a less stringent decision for BE in cases of large reference variances. That is, if the reference and test products appear to be very different based on average results, they still may be deemed equivalent if the reference within-subject variance is large. This can be a problem in interpretation of BE, because if the within-subject variance of the test product is sufficiently smaller than the reference, an unreasonably large difference between their averages could still result in BE [see Eq. (11.4)]. This could be described as a compensation feature or trade-off; that is, a small within-subject variance for the test product can compensate for a large difference in averages. To ensure that such apparently unreasonable conclusions will not be decisive, the FDA guidance has a proviso that the observed T/R ratio must be not more than 1.25 or less than 0.8.

11.4.6.2 Constant Scaling

The FDA guidance [10] also allows for a constant scaling factor in the denominator of Eq. (11.4). If the variance of the reference is very small, the IB metric may appear very large, even though the products are reasonably close. If the within-subject variance for the reference product is less than 0.04, a value of 0.04 may be used in the denominator, rather than the observed variance. This prevents an artificial inflation of the metric for cases of a small within-subject reference variance. This case will not be discussed further, but is a simple extension of the following discussion. The reader may refer to the FDA guidance for further discussion of this topic [10].

11.4.6.3 Statistical Analysis for IB

For average BE, the distribution of the difference in average results (log transformed) is known based on the assumption of a log-normal distribution of the parameters. One of the problems with the definition of BE based on the metric, Eq. (11.4), is that the distribution of the metric is complex, and cannot be easily evaluated. At an earlier evolution in the analysis of the metric, a bootstrap technique, a kind of simulation, was applied to the data to estimate its distribution. The nature of the distribution is needed to construct a confidence interval so that a decision rule of acceptance or rejection can be determined. This bootstrap approach was time consuming, and not exactly reproducible. An approximate "parametric" approach was recommended [26], which results in a hypothesis test that determines the acceptance rule. We refer to this approach as the ''Hyslop'' evaluation. This will be presented in more detail below.

To illustrate the use of the Hyslop approach, the data of Table 11.18 will be used. This data set has been studied by several authors during the development of methods to evaluate IB [27].

The details of the derivation and assumptions can be found in the FDA guidance [28] and the paper by Hyslop et al. [26].

The following describes the calculations involved and the definitions of some terms that are used in the calculations. The various estimates are obtained from the data of Table 11.18, using SAS [13], with the following code:

```
proc mixed data = Drug;

class seq subj per trt;

model ln C_max = seq per trt;

random int subject/subject = trt;

repeated/grp = trt sub = subj;

estimate "t vs.r" trt 1 - 1/cl alpha = 0.1;

run;
```

Table 11.19 shows the estimates of the variance components and average results for each product from the data of Table 11.18.

Basically, the Hyslop procedure obtains an approximate upper confidence interval on the sum of independent terms (variables) in the IB metric equation [Eq. (11.4)]. However, the statistical approach is expressed as a test of a hypothesis. If the upper limit of the CI is less than 0, the products are deemed equivalent, and vice versa. The following discussion relates to the scaled metric, where the observed reference within-subject variance is used in the denominator. An analogous approach is used for the case where the reference variance is small and the denominator is fixed at 0.04 (see Ref. [28]).

The IB criterion is expressed as

$$\theta = \frac{[\delta^2 + \sigma_d^2 + (\sigma_T^2 - \sigma_R^2)]}{\sigma_R^2}. \tag{11.5}$$

Table 11.18 Data from a Two-Treatment, Two-Sequence, Four-Period Replicated Design [20]

Subject	Sequence	Period	Product	Ln C_{max}
1	1	1	1	5.105339
1	1	3	1	5.090062
2	1	1	1	5.594340
2	1	3	1	5.459160
3	2	2	1	4.991792
3	2	4	1	4.693181
4	1	1	1	4.553877
4	1	3	1	4.682131
5	2	2	1	5.168778
5	2	4	1	5.213304
6	2	2	1	5.081404
6	2	4	1	5.333202
7	2	2	1	5.128715
7	2	4	1	5.488524
8	1	1	1	4.131961
8	1	3	1	4.849684
1	1	2	2	4.922168
1	1	4	2	4.708629
2	1	2	2	5.116196
2	1	4	2	5.344246
3	2	1	2	5.216565
3	2	3	2	4.513055
4	1	2	2	4.680278
4	1	4	2	5.155601
5	2	1	2	5.156178
5	2	3	2	4.987025
6	2	1	2	5.271460
6	2	3	2	5.035003
7	2	1	2	5.019265
7	2	3	2	5.246498
8	1	2	2	5.249127
8	1	4	2	5.245971

It can be shown that

$$\sigma_1^2 = \sigma_d^2 + 0.5(\sigma_T^2 + \sigma_R^2), \tag{11.6}$$

where σ_d^2 is the pure estimate of the subject × formulation interaction component. We can express this in the form of hypothesis test, where the IB metric is linearized as follows:

Substituting Eq. (11.6) into Eq. (11.5), and linearizing

$$\text{Let } \eta = (\delta)^2 + \sigma_1^2 + 0.5\,\sigma_T^2 - \sigma_R^2(-1.5 - \theta). \tag{11.7}$$

Table 11.19 Parameter Estimates from Analysis of Data of Table 4 with Some Definitions

μ'_T = mean of test; estimate = 5.0353
μ'_R = mean of reference; estimate = 5.0542
δ = difference between observed mean of test and reference = −0.0189
$\mu'_t{}^2$ = interaction variance; estimate = M_I = 0.1325
$\mu'_T{}^2$ = within-subject variance for the test product; estimate = M_T = 0.0568
$\mu'_R{}^2$ = within-subject variance for the reference product; estimate = M_R = 0.0584
n = degrees of freedom
s = number of sequences

Table 11.20 Computations for Evaluation of Individual Bioequivalence

Hq = (1 − alpha) level upper Confidence limit	**Eq = point estimate**	**Uq = (Hq − Eq)^2**
$H_D = \left[\lvert\delta\rvert + t(1-\alpha,\ n-s)(1/s^2 \sum n_i^{-1}\ M_I)^{1/2}\right]^2$	$E_D = \delta^2$	U_D
$H_I = [(n-s)\cdot M_I]/\chi^2(\alpha, n-s)$	$E_I = M_I$	U_I
$H_T = [0.5\cdot(n-s)\cdot M_T]/\chi^2(\alpha,\ n-s)$	$E_T = 0.5{\cdot}M_T$	U_T
$H_R = [-(1.5+\theta_1)\cdot(n-s)\cdot M_R]/\chi^2(1-\alpha, n-s)$ [a]	$E_R = -(1.5+\theta_1){\cdot}M_R$	U_R

[a]Note that we use the $1-\alpha$ percentile here because of the negative nature of this expression. $n = \sum n_j$; s = number of sequences; n_i = the number of subjects in sequence i.

We then form a hypothesis test with the hypotheses

$H_0: \eta > 0 \qquad H_a: \eta > 0.$

Howe's Method (Hyslop) effectively forms a CI for η by first finding an upper or lower limit for each component in η. Then, a simple computation allows us to accept or reject the null hypothesis at the 5% level (one-sided test). This is equivalent to seeing if an upper CI is less than the FDA-specified criterion, θ. Using Hyslop's Method, if the upper confidence limit is less than θ, the test will show a value less than 0, and the products are considered to be equivalent.

The computation for the method is detailed below.

We substitute the observed values for the theoretical values in Eq. (11.7). The observed values are shown in Table 11.19.

The next step is to compute the upper 95% confidence limits for the components in Eq. (11.7). Note that δ is normal with mean, true delta, and variance $2\sigma_d^2/N$. The variances are distributed as $(\sigma^2)\cdot\chi^2_{(n)}/n$ (where n = d.f.). For example, $M_T \sim \sigma_T(n)^2\,\chi^2_{(n)}/n$.

The equations for calculations are given in Table 11.20 [26].

$$H = \sum(E_i) + \sum(U_i)^{0.5} = -0.0720 + 0.3885 = 0.3165.$$

Table 11.21 shows the results of these calculations.

Examples of calculations

$$H_D = \left[\lvert-0.0189\rvert + 1.94\cdot((1/4)\cdot 0.1325/2)^{1/2}\right]^2 = 0.07213$$

$$H_I = ((6)\cdot 0.1325)/1.635 = 0.4862$$

$$H_T = (0.5\cdot(6)\cdot 0.0568)/1.635 = 0.1042$$

$$H_R = (-(1.5+2.4948)\cdot(6)\cdot 0.0584)/12.59 = -0.1112$$

Table 11.21 Results of Calculations for Data of Table 11.20

H_i = confidence limit	E_i = point estimate	$U_i = (H-E)^2$
$H_d = 0.07213$	$E_d = 0.00357$	0.0052
$H_i = 0.4862$	$E_i = 0.1325$	0.1251
$H_t = 0.1042$	$E_t = 0.0284$	0.0057
$H_r = -0.1112$	$E_r = -0.2333$	0.0149
SUM	−0.0720	0.1509

If the upper CI exceeds zero, the hypothesis is rejected, and the products are bioinequivalent. This takes the form of a one-sided test of hypothesis at the 5% level.

Since this value (0.3165) exceeds 0, the products are considered to be inequivalent.

An alternative method to construct a decision criterion for IB based on the metric is given in Appendix IX.

11.4.6.4 The Future

At the present time, the design and analysis of BE studies use tttp designs with a log transformation of the estimated parameters. The 90% CI of the back-transformed difference of the average results for the comparative products must lie between 0.8 and 1.25 for the products to be deemed equivalent. Four-period replicate designs have been recommended on occasion for controlled-release products and, in some cases, very variable products. However, FDA recommends that these designs be analyzed for average BE. The results of these studies were analyzed for IB by the FDA to assess the need for IB; that is, is there a problem with formulation × subject interactions and differences between within-subject variance for the two products? The result of this venture showed that replicate designs were not needed, that is, the data does not show significant interaction or within-subject variance differences. IB may be reserved for occasions where these designs will be advantageous in terms of cost and time. In fact, recent communication with FDA suggests that IB requirements are not likely to continue in the present form. Some form of IB analysis may be optimal for very variable drugs, requiring less subjects than would be required using a tttp design for average BE. On the other hand, in the future if IB analysis shows the existence of problems with interaction and within-subject variances, it is possible that the four-period replicate design and IB analysis will be considered for at least some subset of drugs or drug products that exhibit problems. For very variable drug products, a scaled analysis has been proposed that would reduce the sample size relative to the usual crossover analysis (see below, sect. 11.4.9). Also, FDA is investigating the use of sequential designs, or add-on designs in the implementation of BE studies.

See Appendix X for a discussion of designs used in BE studies.

11.4.7 Sample Size for Test for Equivalence for a Dichotomous (Pass–Fail) Outcome

Tests for BE are usually based on an analysis of drug in body fluids (e.g., plasma). However, for drugs that are not absorbed, such as topicals and certain local acting gastrointestinal products (e.g., sucralfate), a clinical study is necessary. Often the outcome is based on a binomial outcome such as cured/not cured. See section 5.2.6 for confidence intervals for a proportion. A continuity correction is recommended. Makuch and Simon [29] have published a method for determining sample size for these studies, as well as other kinds of clinical studies where the objective is to determine equivalence. This reference is concerned particularly with cancer treatments where a less intensive treatment is considered to replace a more toxic treatment if the two treatments can be shown to be therapeutically equivalent. As for the case of BE studies with a continuous outcome, one needs to specify both alpha and beta errors in addition to a difference between the treatments that is considered important to estimate the required sample size.

In this approach, we assume a parallel-groups design (two independent groups), typical of these studies. To estimate the number of subjects required in the two groups, we will assume an equal number to be assigned to each group. An estimate of (1) the value of the proportion of subjects who will be "cured" or have a positive outcome for each treatment (P_1 and P_2), and (2) the difference between the treatments that are not clinically meaningful is needed. Makuch and Simon have shown that the number of subjects per group can be calculated from Eq. (11.4):

$$N = [P_1(1 - P_1) + P_2(1 - P_2)] \times \left\{ \frac{[Z_\alpha + Z_\beta]}{[\Delta - [P_1 - P_2]} \right\}^2, \tag{11.8}$$

where delta (Δ) is the maximum difference between treatments considered to be of no clinical significance.

If we assume that the products are not different a priori, $P_1 = P_2 = P$, Eq. (11.4) reduces to

$$N = 2P(1 - P)\left\{\frac{[Z_\alpha + Z_\beta]}{\Delta}\right\}^2. \tag{11.9}$$

In a practical example, a clinical study is designed to compare the efficacy of a generic sucralfate to the brand product. The outcome is the healing of gastrointestinal ulcers. How many subjects should be entered in a parallel study with a dichotomous endpoint (healed/ not healed) if the expected proportion healed is 0.80 and the CI of the difference of the proportions should not exceed ±0.2? We wish to construct a two-sided 90% CI with a beta of 0.2 (power = 0.8). This means that with the required number of patients, we will be able to determine, with 90% confidence, if the healing rates of the products are within ±0.2. If indeed the products are equivalent, with a beta of 0.2, there is 80% probability that the CI for the difference between the products will fall within ±20%.

The values of Z for beta can be obtained from Table 6.2.

Note that if the products are not considered to be different with regard to proportion or probability of success, the values for beta will be based on a two-sided criterion. For example, for 80% power, use 1.28 (not 0.84). From Eq. (11.5),

$$N = 2(0.8)(1 - 0.8)\left\{\frac{[1.65 + 1.28]}{0.2}\right\}^2 = 69.$$

Sixty-nine subjects per group are required to satisfy the statistical requirements for the study.

If the criterion is made more similar to the typical BE criterion, we might consider the difference (delta) to be 20% of 0.8 or 16%, rather than the absolute 20%. If delta is 16%, the number of subjects per group will be approximately 108. (See Exercise Problem 12 at the end of this chapter.) The BE subject number calculator on the CD included with this book provides for the calculation of these subject numbers with the inclusion of a continuity correction often requested by FDA.

11.4.8 SCALED CRITERION FOR BE

The scaled criterion is currently endorsed by FDA for highly variable drug products [30]. A within-subject CV of 30% or greater is considered "highly variable." The recommended design is a three-period crossover with three sequences, TRR, RTR and RRT, where R is the reference and T is the test product. Thus, only the reference is replicated, and the within-subject variance can be estimated for the reference product. Although a minimum sample size of 24 is recommended, the appropriate sample size is determined by the sponsor. After a log transformation, the parameters (AUC and C_{max}) are calculated in addition to the within-subject variance of the reference product.

The statistical null hypothesis is

$$H_o : (X_T - X_R)^2/S_R^2 > \theta$$

The alternative hypothesis is

$$H_1 : (X_T - X_R)^2/S_R^2 \leq \theta,$$

where θ is the scaled average BE limit, $X_T - X_R$ is the difference between the average parameter (AUC or C_{max}) after a log transformation, and $S_R{}^2$ is the calculated within-subject variance for the reference product.

θ is defined as $(\ln \Delta)^2/\sigma^2{}_{wo}$, where $\Delta = 1.25$ and $\sigma_{wo} = 0.25$.

Therefore, $\theta = 0.7967$.

BE is accepted if the null hypothesis is rejected and *the ratio of test to reference is between 0.8 and 1.25. Both criteria must be satisfied to declare* BE.

A 95% upper bound for $(X_T - X_R)^2/S_R^2$ from the BE study must be $\leq \theta$ in addition to the restriction of the ratio of test to reference parameters (0.8–1.25). As of this writing, a method for computing the upper bound is not forthcoming. Use of the "Hyslop" Method for individual BE, previously discussed and modified for this application, has been proposed.

11.4.9 NONINFERIORITY TRIALS

Noninferiority trials are related to BE studies in that in both cases we are not testing for differences. For noninferiority trials, we are testing that a test product is not worse than a reference product based on results of a clinical study. Again, we must define a value such that if the lower confidence bound (usually 95%) of the test treatment compared to the reference exceeds that value, the test treatment will be considered noninferior. This value should be defined in the protocol prior to seeing the study results, and is a value such that any value lower than the specified value would result in a conclusion of inferiority.

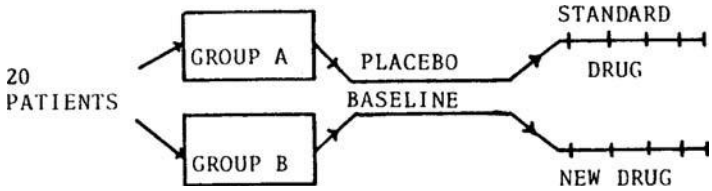
For example, comparing Test Drug *X* to Reference Drug *Y*, it was determined that a difference in average response of 2 units would be acceptable for purposes of noninferiority. That is, if study results showed that Drug *X* was no more than 2 units less than Drug *Y*, Drug *X* would be considered to be noninferior to Drug *Y*. The study showed that Drug *X* was 1 unit less than Drug *Y*. The 95% lower bound of this difference was 2.1, that is, based on the lower bound, Drug *X*, could be as much as 2.1 units less than Drug *Y*. Therefore, Drug *X* failed the noninferiority test. The lower confidence bound showed more than a 2 unit difference, and we can not conclude that Drug *X* is noninferior to Drug *Y*.

11.5 REPEATED MEASURES (SPLIT-PLOT) DESIGNS

Many clinical studies take the form of a baseline measurement followed by observations at more than one point in time. For example, a new antihypertensive drug is to be compared to a standard, marketed drug with respect to diastolic blood pressure reduction. In this case, after a baseline blood pressure is established, the patients are examined every other week for eight weeks, a total of four observations (visits) after treatment is initiated.

11.5.1 Experimental Design

Although this antihypertensive drug study was designed as a multiclinic study, the data presented here represent a single clinic. Twenty patients were randomly assigned to the two treatment groups, 10 to each group (see sect. 11.2.6 for the randomization procedure). Prior to drug treatment, each patient was treated with placebo, and blood pressure determined on three occasions. The average of these three measurements was the baseline reading.



The baseline data were examined to ensure that the three baseline readings did not show a time trend. For example, a placebo effect could have resulted in decreased blood pressure with time during this preliminary phase.

Treatment was initiated after the baseline blood pressure was established. Diastolic blood pressure was measured every two weeks for eight weeks following initiation of treatment. (The dose was one tablet each day for the standard and new drug.) Two patients dropped out in the "standard drug" group, and one patient was lost to the "new drug" group, resulting in eight and nine patients in each treatment group. The results of the study are shown in Table 11.22 and Figure 11.4.

The design described above is commonly known in the pharmaceutical industry as a *repeated measures* or *split-plot* design. (This design is also denoted as an incomplete three-way or a partially hierarchical design.) This design is common in clinical or preclinical studies, where

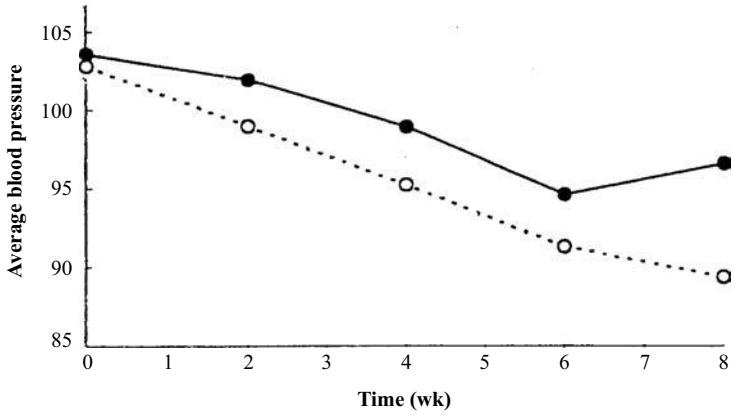


Figure 11.4 Plot of mean results from antihypertensive drug study. •—standard drug; O—new drug.

two or more products are to be compared with multiple observations over time. The design can be considered as an extension of the one-way or parallel-groups design. In the present design (repeated measures), data are obtained at more than one time point. The result is two or more two-way designs, as can be seen in Table 11.22, where we have two two-way designs. The two-way designs are related in that observations are made at the same time periods. *The chief features of the repeated measures design* as presented here are as follows:

1. Different patients are randomly assigned to the different treatment groups, that is, a patient is assigned to only one treatment group.
2. The number of patients in each group need not be equal. Equal numbers of patients per group, however, result in optimum precision when comparing treatment means. Usually, these studies are designed to have the same number of patients in each group, but dropouts usually occur during the course of the study.
3. Two or more treatment groups may be included in the study.
4. Each patient provides more than one measurement over time.
5. The observation times (visits) are the same for all patients.
6. Baseline measurements are usually available.
7. The usual precautions regarding blinding and randomization are followed.

Although the analysis tolerates lack of symmetry with regard to the number of patients per group (see feature 2), the statistical analysis can be difficult if patients included in the study

Table 11.22 Results of a Comparison of Two Antihypertensive Drugs

Standard drug						New drug					
		Week						Week			
Patient	Baseline	2	4	6	8	Patient	Baseline	2	4	6	8
1	102	106	97	86	93	3	98	96	97	82	91
2	105	103	102	99	101	4	106	100	98	96	93
5	99	95	96	88	88	6	102	99	95	93	93
9	105	102	102	98	98	8	102	94	97	98	85
13	108	108	101	91	102	10	98	93	84	87	83
15	104	101	97	99	97	11	108	110	95	92	88
17	106	103	100	97	101	12	103	96	99	88	86
18	100	97	96	99	93	14	101	96	96	93	89
						16	107	107	96	93	97
Mean	103.6	101.9	98.9	94.6	96.6	Mean	102.8	99.0	95.2	91.3	89.4

have missing data for one or more visits. In these cases, a statistician should be consulted regarding data analysis [31].

The usual assumptions of normality, independence, and homogeneity of variance for each observation hold for the split-plot analysis. In addition, there is another important assumption with regard to the analysis and interpretation of the data in these designs. The assumption is that the data at the various time periods (visits) are not correlated, or that the correlation is of a special form [32]. Although this is an important assumption, often ignored in practice, moderate departures from the assumption can be tolerated. Correlation of data during successive time periods often occurs such that data from periods close together are highly correlated compared to the correlation of data far apart in time. For example, if a person has a high blood pressure reading at the first visit of a clinical study, we might expect a similar reading at the subsequent visit if the visits are close in time. The reading at the end of the study is apt to be less related to the initial reading. The present analysis assumes that the correlation of the data is the same for all pairs of time periods, and that the pattern of the correlation is the same in the different groups (e.g., drug groups) [32]. If these assumptions are substantially violated, the conclusions based on the usual statistical analysis will not be valid. The following discussion assumes that this problem has been considered and is negligible [31].

11.5.2 ANOVA

The data of Table 11.22 will be subjected to the typical repeated measures (split-plot) ANOVA. As in the previous examples in this chapter, the data will be analyzed, corrected for baseline, by subtracting the baseline measurement from each observation. The measurements will then represent *changes from baseline.* The more complicated analysis of covariance is an alternative method of treating such data [31, 32]. More expert statistical help will usually be needed when applying this technique, and the use of a computer is almost mandatory. Subtracting out the baseline reading is easy to interpret and, generally, results in conclusions very similar to that obtained by covariance analysis. Table 11.23 shows the "changes from baseline" data derived from Table 11.22. For example, the first entry in this table, two weeks for the standard drug, is $106 - 102 = 4$.

When computing the ANOVA by hand (use a calculator), the simplest approach is to first compute the two-way ANOVA for each treatment group, "standard drug" and "new drug." The calculations are described in section 8.4. The results of the ANOVA are shown in Table 11.24. Only the sums of squares need to be calculated for this preliminary computation.

The final analysis combines the separate two-way ANOVAs and has two new terms, "weeks × drugs" interaction and "drugs," the variance represented by the difference between the drugs. The calculations are described below, and the final ANOVA table is shown in Table 11.25.

Table 11.23 Changes from Baseline of Diastolic Pressure for the Comparison of Two Antihypertensive Drugs

Standard drug					New drug				
	Week					Week			
Patient	2	4	6	8	Patient	2	4	6	8
1	4	−5	−16	−9	3	−2	−1	−16	−7
2	−2	−3	−6	−4	4	−6	−8	−10	−13
5	−4	−3	−11	−11	6	−3	−7	−9	−9
9	−3	−3	−7	−7	8	−8	−5	−4	−17
13	0	−7	−17	−6	10	−5	−14	−11	−15
15	−3	−7	−5	−7	11	2	−13	−16	−20
17	−3	−6	−9	−5	12	−7	−4	−15	−17
18	−3	−4	−1	−7	14	−5	−5	−8	−12
					16	0	−11	−14	−10
Mean	−1.75	−4.75	−9	−7	Mean	−3.8	−7.6	−11.4	−13.3
Sum	−14	−38	−72	−56	Sum	−34	−68	−103	−120

Table 11.24 ANOVA for Changes from Baseline for Standard Drug and New Drug

	Standard drug		New drug	
Source	**d.f.**	**Sum of squares**	**d.f.**	**Sum of squares**
Patients	7	57.5	8	114.22
Weeks	3	232.5	3	486.97
Error	21	255.5	24	407.78
Total	31	545.5	35	1008.97

Patients' SS: Pool the SS from the separate ANOVAs ($57.5 + 114.22 = 171.72$ with $7 + 8 = 15$ d.f.).

Weeks' SS: This term is calculated by combining all the data, resulting in four columns (weeks), with 17 observations per column, 8 from the standard drug and 9 from the new drug. The calculation is

$$\frac{\sum C^2}{R_1 + R_2} - \text{CT},$$

where C is the column sums of combined data and $R_1 + R_2$ is the sum of the number of rows,

$$= \frac{(-48)^2 + (-106)^2 + (-175)^2 + (-176)^2}{17} - \frac{(-505)^2}{68}$$

$$= 4420.1 - 3750.4 = 669.7.$$

Drug SS:

$$\text{Drug SS} = (\text{CT}_{\text{SP}}) + (\text{CT}_{\text{NP}}) - (\text{CT}_{\text{T}}),$$

where CT_{sp} is the correction term for the standard drug, CT_{NP} the correction term for the new product, and CT_{T} the correction term for the combined data.

$$\text{Drug SS} = \frac{(-180)^2}{32} + \frac{(-325)^2}{36} - \frac{(-505)^2}{68}$$

$$= 196.2.$$

Table 11.25 Repeated Measures (Split-Plot) ANOVA for the Antihypertensive Drug Study

Source	**d.f.**[a]	**SS**	**MS**	
Patients	15	171.7	11.45	
Weeks	3	669.7	223.23	
Drugs	1	196.2	196.2	$F_{1,15} = \frac{196.2}{11.45} = 17.1^*$
Weeks × drugs	3	49.8	16.6	
Error (within treatments)	45	663.3	14.74	$F_{1,15} = \frac{16.6}{14.74} = 1.1$
	67	1750.6		

[a]Degrees of freedom for "patients" and "error" are the d.f. pooled from the two-way ANOVAs. For "weeks" and "drugs," the d.f. are (weeks − 1) and (drugs − 1), respectively. For "weeks × drugs," d.f. are (weeks − 1) × (drugs − 1).

$^*p < 0.01$.

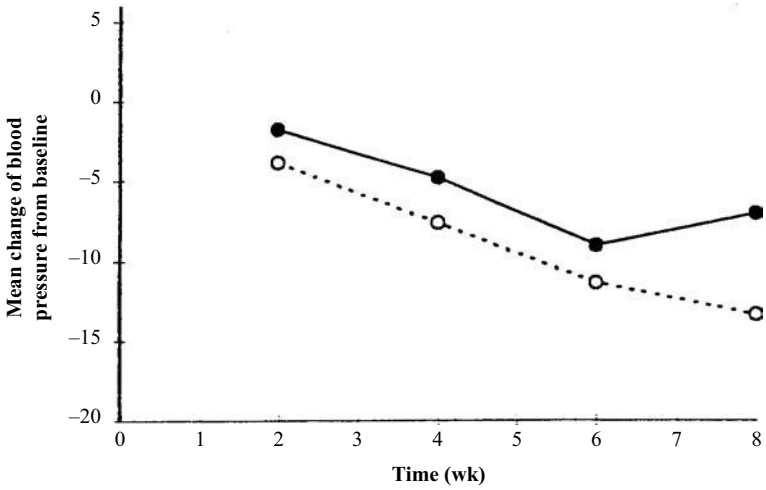


Figure 11.5 Plot from the data of Table 11.23 showing lack of significant interaction of weeks and drugs in experiment comparing standard and new antihypertensive drugs. •—standard drug; o—new drug.

Weeks × drugs SS: This interaction term (see below for interpretation) is calculated as the pooled SS from the "week" terms in the separate two-way ANOVAs above, minus the week term for the final combined analysis, 669.7.

$$\text{Weeks} \times \text{drug SS} = 232.5 + 486.97 - 669.7 = 49.8.$$

Error SS: The error SS is the pooled error from the two-way ANOVAs, 255.5 + 407.8 = 663.3.

11.5.2.1 Interpretation and Discussion

The terms of most interest are the "drugs" and "weeks × drugs" components of the ANOVA. "Drugs" measures the difference between the overall averages of the two treatment groups. The average reduction of blood pressure was (180/32) = 5.625 mm Hg for standard drug, and (325/36) = 9.027 mm Hg for the new drug. The F test for "drug" differences is (drug MS)/(patients MS) equal to 17.1 (1 and 15 d.f.; see Table 11.25). This difference is highly significant ($p < 0.01$). The significant result indicates that on the average, the new drug is superior to the standard drug with regard to lowering diastolic blood pressure.

The significant difference between the standard and new drugs is particularly meaningful if the difference is constant over time. Otherwise, the difference is more difficult to interpret. "Weeks × drugs" is a measure of interaction (see also chap. 9). This test compares the parallelism of the two "change from baseline" curves as shown in Figure 11.5. The F test for "weeks × drugs" uses a different error term than the test for "drugs." The F test with 3 and 45 d.f. is 16.6/14.74 = 1.1, as shown in Table 11.25. This nonsignificant result suggests that the pattern of response is not very different for the two drugs. A reasonable conclusion based on this analysis is that the new drug is effective (superior to the standard drug), and that its advantage beyond the standard drug is approximately maintained during the course of the experiment.

A significant nonparallelism of the two "curves" in Figure 11.5 would be evidence for a "weeks × drugs" interaction. For example, if the new drug showed a lower change in blood pressure than the standard drug at two weeks, and a higher change in blood pressure at eight weeks (the curves cross one another), interaction of weeks and drugs would more likely be significant. Interaction, in this example, would suggest that drug differences are dependent on the time of observation.

If interaction is present or the assumptions underlying the analysis are violated (particularly concerning the form of the covariance matrix) [31], a follow-up or an alternative is to perform p one-way ANOVAs at each of the p points in time. In the previous example, analyses

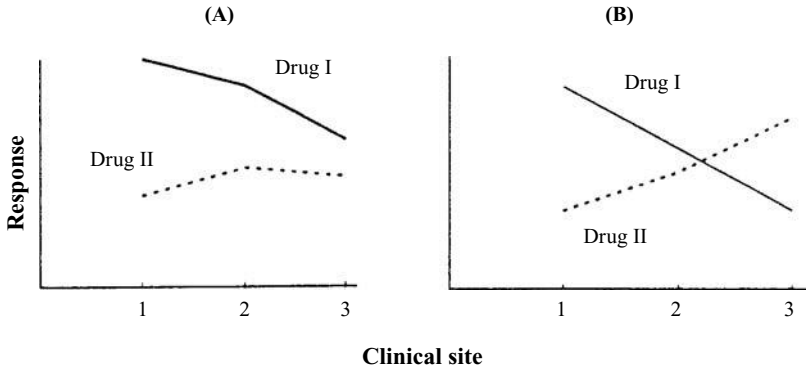


Figure 11.6 Two kinds of interaction: (**A**) one drug always better than another, but the difference changes for different clinical sites; (**B**) one drug better than another at sites 1 and 2 and worse at site 3.

would be performed at each of the four post-treatment weeks. A conclusion is then made on the results of these individual analyses (see Exercise Problem 8).

11.6 MULTICLINIC STUDIES

Most clinical studies carried out during late phase 2 or phase 3 periods of drug testing involve multiclinic studies. In these investigations, a common protocol is implemented at more than one study site. This procedure, recommended by the FDA, serves several purposes. It may not be possible to recruit sufficient patients in a study carried out by a single investigator. Thus multiclinic studies are used to "beef up" the sample size. Another very important consideration is that multiclinic studies, if performed at various geographic locations with patients representing a wide variety of attributes, such as age, race, socioeconomic status, and so on, yield data that can be considered representative under a wide variety of conditions. Multiclinic studies, in this way, guard against the possibility of a result peculiar to a particular single clinical site. For example, a study carried out at a single Veterans' Administration hospital would probably involve older males of a particular economic class. Also, a single investigator may implement the study in a unique way that may not be typical, and the results would be peculiar to his or her methods. Thus, if a drug is tested at many locations and the results show a similar measure of efficacy at all locations, one has some assurance of the general applicability of the drug therapy. In general, one should attempt to have more or less equal numbers of patients at each site, and to avoid having too few patients at sites.

However, there are instances where a drug has been found to be efficacious in the hands of some investigators and not for others. When this occurs, the drug effect is in some doubt unless one can discover the cause of such results. This problem is statistically apparent in the form of a treatment $\times$ site interaction. The comparative treatments (drug and placebo, for example) are not differentiated equally at different sites. A treatment $\times$ site interaction may be considered very serious when one treatment is favored at some clinical sites and the other favored at different sites. Less serious is the case of interaction where all clinics favor the same treatment, but some favor it more than others. These two examples of interaction are illustrated in Figure 11.6.

When interaction occurs, the design, patient population, clinical methods, protocol, and other possible problems should be carefully investigated and dissected, to help find the cause. The cause will not always be readily apparent, if at all. See section 8.4.3 for a further example and discussion of interactions in clinical studies. An important feature of multiclinic studies, as noted above, is that the same protocol and design should be followed at all sites.

Since one can anticipate missing values due to dropouts, missed visits, recording errors, and so on, an important consideration is that the design should not be so complicated that missing data will cause problems with the statistical interpretation or that the clinicians will have difficulty following the protocol. A simple design that will achieve the objective is to

be preferred. Since parallel-groups designs are the most simple in concept, these should be preferred to some more esoteric design. Nevertheless, there are occasions where a more complex design would be appropriate providing that the study is closely monitored and the clinical investigators thoroughly educated.

11.7 INTERIM ANALYSES

Under certain conditions, it is convenient (and sometimes prudent) to look at data resulting from a study prior to its completion in order to make a decision to change the protocol procedure or requirements, or to abort the study early or to increase the sample size, for example. This is particularly compelling for a clinical study involving a disease that is life-threatening, is expensive, and/or is expected to take a long time to complete. A study may be stopped, for example, if the test treatment can be shown to be superior early on in the study. However, if the data are analyzed prior to study completion, a penalty is imposed in the form of a lower significance level to compensate for the multiple looks at the data (i.e., to maintain the overall significance level at alpha). The more occasions that one looks at and analyzes the data for significance, the greater the penalty, that is, the more difficult it will be to obtain significance at each analysis. The penalty takes the form of an adjustment of the alpha level to compensate for the multiple looks at the data. The usual aim is to keep the alpha level at a nominal level, for example 5%, considering the multiple analyses; this fixes the probability of declaring the treatments different when they are truly the same at, at most, 5%, taking into account the fact that at each look we have a chance of incorrectly declaring a significant difference. For example, if the significance level is 0.05 for a single look, two looks will have an overall significance level of approximately 0.08.

In addition to the advantage (time and money) of stopping a study early when efficacy is clearly demonstrated, there may be other reasons to shorten the duration of a study, such as stopping because of a drug failure, modifying the number of patients to be included, modifying the dose, and so on. If interim analyses are made for these purposes in phase 3 pivotal studies, an adjusted p level will probably be needed for regulatory purposes. Davis and Huang discuss this in more detail [33]. In any event, the approach to interim analyses should be clearly described in the study protocol, a priori; or, if planned after the study has started, the plan of the interim analysis should be communicated to the regulatory authorities (e.g., FDA). Even if interim looks do not affect the study procedure or outcome, such procedures should be clearly documented either in the study protocol or as an amendment to the protocol. One of the popular approaches to interim analyses was devised by O'Brien and Fleming [34], an analysis known as a group sequential method. The statistical analyses are performed after a group of observations have been accumulated rather than after each individual observation. The analyses should be clearly documented and should be performed by persons who cannot influence the continuation and conduct of the study.

The procedure and performance of these analyses must be described in great detail in the study protocol, including the penalties in the form of reduced "significance" levels. A very important feature of interim analyses is the procedure of breaking the randomization code. One should clearly specify who has access to the code and how the blinding of the study is maintained. It is crucial that the persons involved in conducting the study, clinical personnel and monitors alike, not be biased as a result of the analysis. This is of great concern to the FDA. Interim analyses should not be done willy-nilly, but should be planned and discussed with regulatory authorities. Associated penalties should be fixed in the protocol. As noted previously, this does not mean that interim analyses cannot and should not be performed as an afterthought if circumstances dictate their use during the course of the study. A Pharmaceutical Manufacturer's Association (PMA) committee [35] suggested the following to minimize potential bias resulting from an interim analysis. (1) "A Data Monitoring Committee (DMC) should be established to review interim results." The persons on this committee should not be involved in decisions regarding the progress of the study. (2) If the interim analysis is meant to terminate a study, the details should be presented in the protocol, a priori. (3) The results of the interim analysis should be confidential, known only to the DMC.

Sankoh [25] discusses situations where interim analyses have been used incorrectly from a regulatory point of view. In particular, he is concerned with unplanned interim analyses.

Table 11.26 Significance Levels for Two-Sided Group Sequential Studies with an Overall Significance Level of 0.05 (According to O'Brien/Fleming)

Number of analysis (stages)	Analysis	Significance level
2	First	0.005
	Final	0.048
3	First	0.0005
	Second	0.014
	Final	0.045
4	First	0.0005
	Second	0.004
	Third	0.019
	Final	0.043
5	First	0.00001
	Second	0.001
	Third	0.008
	Fourth	0.023
	Final	0.041

These include (a) the lack of reporting these analyses and the consequent lack of adjustment of the significance level, (b) inappropriate adjustment of the level and inappropriate stopping rules, (c) interim analyses inappropriately labeled "administrative analyses," where actual data analyses have been carried out and results disseminated, (d) lack of documentation for the unplanned interim analysis, (e) and the importance of blinding and other protocol requirements.

An interim analysis may also be planned to adjust sample size. In this case, a full analysis should not be done. The analysis should be performed when the study is not more than half done, and only the variability should be estimated (not the treatment differences). Under these conditions, no penalty need be assessed. However, if the analysis is done near the end of the trial or if the treatment differences are computed, a penalty is required [25].

Table 11.26 shows the probability levels needed for significance for k looks (k analyses) at the data according to O'Brien and Fleming [34], where the data are analyzed at equal intervals during patient enrollment. For example, if the data are to be analyzed three times ($k = 3$, where k is the number of analyses or stages, including the final analysis), the analysis should be done after $^1/_3$, $^2/_3$ and all of the patients have been completed [36]. There are other schemes for group sequential interim analyses, including those that do not require analyses at equal intervals of patient completion [37].

For example, a study with 150 patients in each of two groups is considered for two interim analyses. This corresponds to three stages, two interim and one final analysis. The first analysis is performed after 100 patients are completed (50 per group) at the 0.0005 level. To show statistically significant differences, the product differences must be very large or obvious at this low level. If not significant, analyze the data after 200 patients are completed. A significance level of 0.014 must be reached to terminate the study. If this analysis does not show significance, complete the study. The final analysis must meet the 0.045 level for the products to be considered significantly different.

One can conjure up reasons as to why stopping a study early based on interim analysis is undesirable (less information on adverse effects or less information for subgroup analyses, for example). One possible solution to this particular problem in the case where the principle objective is to establish efficacy, is to use the results of the interim analysis for regulatory submission, if the study data meet the interim analysis p level, but to continue the study after the interim analysis, and then analyze the results for purposes of obtaining further information on adverse effects, and so on. However, in this procedure, one may face a dilemma if the study fails to show significance with regard to efficacy after including the remaining patients.

KEY TERMS

Analysis of covariance
AUC (area under curve)
Balance
Baseline measurements
Between-patient variation (error)
Bias
Bioavailability
Bioequivalence
Blinding
Carryover
Changeover design
C_{max}
Controlled study
Crossover design
Differential carryover
Double blind
Double dummy
75–75 rule
Experimental design
Grizzle analysis
Incomplete three-way ANOVA
Individual bioequivalence
Intent to treat
Interaction
Interim analyses
Latin square
Locke's Method
Log transformation
Multiclinic
Objective measurements
Parallel design
Period (visit)
Placebo effect
Positive control
Randomization
Repeated measures
Replicate designs
Scaled bioequivalence analysis
Sequences
Split plot
Symmetry
Systematic error
t_p
Washout period
Within-patient variation (error)
80% power to detect 20% difference

EXERCISES

1. (a) Perform the calculations for the ANOVA table (Table 11.3) from the data in Table 11.2.
   (b) Perform a t test comparing the differences from baseline for the two groups in Table 11.2. Compare the t value to the F value in Table 11.3.

2. Using the data in Table 11.10, test to see if the values of t_p are different for formulations A and B (5% level).

3. (a) Using the data in Table 11.10, compare the values of C_{max} for the two formulations (5% level). Calculate a confidence interval for the difference in C_{max}.
   (**b) Analyze the data for C_{max} using the Grizzle Method. Is a differential carryover effect present?

4. Analyze the AUC data in Table 11.10 using ratios of AUC (A/B). Find the average ratio and test the average for significance. (Note that H_0 is $AUC_A/AUC_B = 1.0$.) Assume no period effect.

5. Analyze the AUC data in Table 11.10 using logarithms of AUC. Compare the antilog of the average difference of the logs to the average ratio determined in Problem 4. Put a 95% confidence interval on the average difference of the logs. Take the antilogs of the lower and upper limit and express the interval as a ratio of the AUCs for the two formulations.

6. ** In a pilot study, two acne preparations were compared by measuring subjective improvement from baseline (10-point scale). Six patients were given a placebo cream and six different patients were given a cream with an active ingredient. Observations were made once a week for four weeks. Following are the results of this experiment:

**This is an optional, more difficult problem.

Placebo					Active				
	Week					Week			
Patient	1	2	3	4	Patient	1	2	3	4
1	2	2	4	3	1	2	2	3	3
2	3	2	3	3	2	4	4	5	4
3	1	4	3	2	3	1	3	4	5
4	3	2	1	0	4	3	4	4	7
5	2	1	3	2	5	2	2	3	6
6	4	4	5	3	6	3	4	6	5

A score of 10 is complete improvement. A score of 0 is no improvement (negative scores mean a worsening of the condition). Perform an ANOVA (split plot). Plot the data as in Figure 11.4. Are the two treatments different? If so, how are they different?

7. For the exercise study described in section 11.3, the difference considered to be significant is 60 minutes with an estimated standard deviation of 55 minutes. Compute the sample size if the Type I (alpha) and Type II (beta) error rates are set at 0.05 and 0.10, respectively.
8. From the data in Table 11.23, test for a difference ($\alpha = 0.05$) between the two drugs at week 4.
9. Perform the ANOVA on the ln transformed bioavailability data (sect. 11.4.2, Table 11.10).
10. A clinical study is designed to compare three treatments in a parallel design. Thirty patients are entered into the study, 10 in each treatment group. The randomization is to be performed in groups of six. Show how you would randomize the 30 patients.
11. In the example in Table 11.7, suppose that a period effect of 3 existed in this study. This means that the observations in Period 2 are augmented by 3 units. Show that the difference between treatments is not biased, that is, the difference between *A* and *B* is 1.
12. Exercise: Compute the sample size for the example in section 11.4.8, assuming that a difference of 0.16 (16%) is a meaningful difference.
13. Compute the confidence interval using Locke's Method as described in section 11.4.3.

REFERENCES

1. Department of Health, Education and Welfare. General Considerations for the Clinical Evaluation of Drugs (GPO 017–012–00245–5) (Pub. HEW (FDA) 77–3040). Washington, D.C.: FDA Bureau of Drugs Clinical Guidelines, U.S. Government Printing Office, 1977.
2. Cox DR. Planning Experiments. New York: Wiley, 1958.
3. Rodda BE, Tsianco MC, Bolognese JA, et al. Clinical Development. In: Peace KE, ed. Statistical Issues in Drug Research and Development. New York: Marcel Dekker, 1990.
4. Buncher CR, Tsay J-Y. Statistics in the Pharmaceutical Industry, 2nd ed. New York: Marcel Dekker, 1994.
5. Fisher LD, et al. Intention to treat in clinical trials. In: Peace KE, ed. Statistical Issues in Drug Research and Development. New York: Marcel Dekker, 1990: 331–349.
6. Snedecor GW, Cochran WG. Statistical Methods, 7th ed. Ames, IA: Iowa University Press, 1980.
7. Brown BW Jr. The crossover experiment for clinical trials. Biometrics 1980; 36:69.
8. Cochran WG, Cox GM. Experimental Designs, 2nd ed. New York: Wiley, 1957.
9. Grizzle JE. The two-period change-over design and its use in clinical trials. Biometrics 1965; 21:467, 1974; 30:727.
10. Guidance for Industry (Draft). Bioavailability and Bioequivalence Studies for Orally Administered Drug Products, General Considerations. Rockville, MD: FDA, CDER, 2003.
11. Haynes JD. Change-over design and its use in clinical trials. J Pharm Sci 1981; 70:673.
12. Schuirman DL. On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. Biometrics 1981, 37:617.

13. The SAS System for Windows, Release 6.12. Cary, NC: SAS Institute Inc.
14. FDA bioavailability/bioequivalence regulations. Fed Regist 1977; 42:1624.
15. FDA. Guidance for Industry, Appendix V. New York: Center for Drug Evaluation and Research, 1997.
16. Chow S-C, Liu J-P. Design and Analysis of Bioavailability and Bioequivalence Studies. New York: Marcel Dekker, 1992:280.
17. Schuirman DL. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J Pharmacokinet Biopharm 1987; 15:657–680.
18. Westlake WJ. Bioavailability and bioequivalence of pharmaceutical formulations. In: Peace KE, ed. Biopharmaceutical Statistics for Drug Development. New York: Marcel Dekker, 1988:329–351.
19. Phillips KE. Power of the two one-sided tests procedure in bioequivalence. J Pharmacokinet Biopharm 1990; 18:137.
20. Diletti E, Hauschke D, Steinijans VW. Sample size determination for bioequivalence assessment by means of confidence intervals. Int J Clin Pharmacol Ther Toxicol 1991; 29(1):1.
21. Boddy AW, Snikeris FC, Kringle RO, et al. An approach for widening the bioequivalence acceptance limits in the case of highly variable drugs. Pharm Res 1995; 12:1865.
22. Committee For Proprietary Medicinal Products (CPMP). Note for Guidance on the Investigation of Bioavailability and Bioequivalence. London: The European Agency for the Evaluation of Medicinal Products, Evaluation of Medicines for Human Use, 2001.
23. Tothfalusi L, Endrenyi L, Midha KK, et al. Evaluation of the bioequivalence of highly variable drugs and drug products. Pharm Res 2001; 18:728.
24. Jones G, Kenward MG. Design and Analysis of Cross-over Trials. London: Chapman and Hill, 1989.
25. Sankoh AJ. Interim analyses: an update of an FDA reviewer's experience and perspective. Drug Inf J 1995; 29:729.
26. Hyslop T, Hsuan F, Hesney M. A small sample confidence interval approach to assess individual bioequivalence. Stat Med 2000; 19:2885–2897.
27. Eckbohm, Melander H. The subject-by-formulation interaction as a criterion of interchangeability of drugs. Biometrics 1989; 45:1249–1254.
28. US Department of Health and Human Services. Statistical Approaches to Establishing Bioequivalence. Rockville, MD: FDA CDER, 2001.
29. Makuch R, Simon R. Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep 1978; 62(7):1037.
30. Haidar SH, et al. Pharm Res 2008; 25:237–240.
31. Chinchilli VM. Clinical efficacy trials with quantitative Data. In: Peace KE, ed. Biopharmaceutical Statistics for Drug Development. New York: Marcel Dekker, 1988:353–394.
32. Winer BJ. Statistical Principles in Experimental Design, 2nd ed. New York: McGraw-Hill, 1971.
33. Davis R, Huang I. Interim analysis. In: Buncher CR, Tsay J-Y, eds. Statistics in the Pharmaceutical Industry, 2nd ed. New York: Marcel Dekker, 1994:267–285.
34. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35:549.
35. Summary from "Issues with Interim Analysis in the Pharmaceutical Industry," The PMA Biostatistics and Medical Ad Hoc Committee on Interim Analysis, 1990.
36. Berry DA. Statistical Methodology in the Pharmaceutical Sciences. New York: Marcel Dekker, 1990:294.
37. Geller N, Pocock S. Design and analysis of clinical trials with group sequential stopping rules. In: Peace KE, ed. Biopharmaceutical Statistics for Drug Development. New York: Marcel Dekker, 1988, Chapter 11.
38. Hauck WH, Anderson S. Measuring switchability and prescribability: When is average bioequivalence sufficient? J Pharmacokinet Biopharm 1994; 22:551.

12 Quality Control

The science of quality control is largely statistical in nature, and entire books have been devoted to the application of statistical techniques to quality control. Statistical quality control is a key factor in process validation and the manufacture of pharmaceutical products. In this chapter, we discuss some common applications of statistics to quality control. These applications include Shewhart control charts, sampling plans for attributes, operating characteristic curves, and some applications to assay development, including components of variance analysis. The applications to quality control make use of standard statistical techniques, many of which have been discussed in previous portions of this book.

12.1 INTRODUCTION

Starting from raw materials to the final packaged container, quality control departments have the responsibility of assuring the integrity of a drug product with regard to safety, potency, and biological availability. If each and every item produced could be tested (100% testing), there would be little need for statistical input in quality control. Those individual dosage units that are found to be unsatisfactory could be discarded, and only the good items would be released for distribution. Unfortunately, conditions exist that make 100% sampling difficult, if not impossible. For example, if every dosage unit could be tested, the expense would probably be prohibitive both to manufacturer and consumer. Also, it is well known that attempts to test individually every item from a large batch (several million tablets, for example), result in tester fatigue, which can cause misclassifications of items and other errors. If testing is destructive, such as would be the case for assay of individual tablets, 100% testing is, obviously, not a practical procedure. However, 100% testing is not necessary to determine product quality precisely. Quality can be accurately and precisely estimated by testing only part of the total material (a sample). In general, quality control procedures require relatively small samples for inspection or analysis. Data obtained from this sampling can then be treated statistically to estimate population parameters such as potency, tablet hardness, dissolution, weight, impurities, content uniformity (variability), as well as to ensure the quality of attributes such as color, appearance, and so on.

In various parts of this book, we discuss data from testing finished products of solid dosage forms. The details of some of these tests are explained at the end of this chapter, section 12.7.

Statistical techniques are also used to monitor processes. In particular, control charts are commonly used to ensure that the average potency and variability resulting from a pharmaceutical process are stable. Control charts can be applied during *in-process* manufacturing operations, for *finished* product characteristics, and in *research and development* for repetitive procedures. Control charts are one of the most important statistical applications to quality control.

12.2 CONTROL CHARTS

Probably the best-known application of statistics to quality control that has withstood the test of time is the Shewhart control chart. Important attributes of the control chart are its simplicity and the visual impression that it imparts. The control chart allows for judgments based on an easily comprehended graph. The basic principles underlying the use of the control chart are described below.

12.2.1 Statistical Control

A process under statistical control is one in which the process is susceptible to variability due only to inherent, but unknown and uncontrolled *chance* causes. According to Grant [1]:

"Measured quality of manufactured product is always subject to a certain amount of variation as a result of chance. Some stable system of chance causes is inherent in any particular scheme of production and inspection. Variation within this stable pattern is inevitable. The reasons for variation outside this stable pattern may be discovered and corrected."

Using tablet manufacture as an example, where tablet weights are being monitored, it is not reasonable to expect that each tablet should have an identical weight, precisely equal to some target value. A tablet machine is simply not capable of producing identical tablets. The variability is due, in part, to (a) the variation of compression force, (b) variation in filling the die, and (c) variation in granulation characteristics. In addition, the balance used to weigh the tablets cannot be expected to give exactly reproducible weighings, even if the tablets could be identically manufactured. Thus, the weight of any single tablet will be subject to the vagaries of chance from the foregoing uncontrollable sources of error, in addition to other identifiable sources that we have not mentioned.

12.2.2 Constructing Control Charts

The process of constructing a control chart depends, to a great extent, on the process characteristics and the objectives that one wishes to achieve. A control chart for tablet weights can serve as a typical example. In this example, we are interested in ensuring that tablet weights remain close to a target value, under "statistical control." To achieve this objective, we will periodically sample a group of tablets, measuring the mean weight and variability. The mean weight and variability of each sample (*subgroup*) are plotted sequentially as a function of time. The control chart is a graph that has time or order of submission of sequential lots on the X axis and the average test result on the Y axis. The process average together with upper and lower limits are specified as shown in Figure 12.1. The preservation of order with respect to the observations is an important feature of the control chart. Among other things, we are interested in attaining a

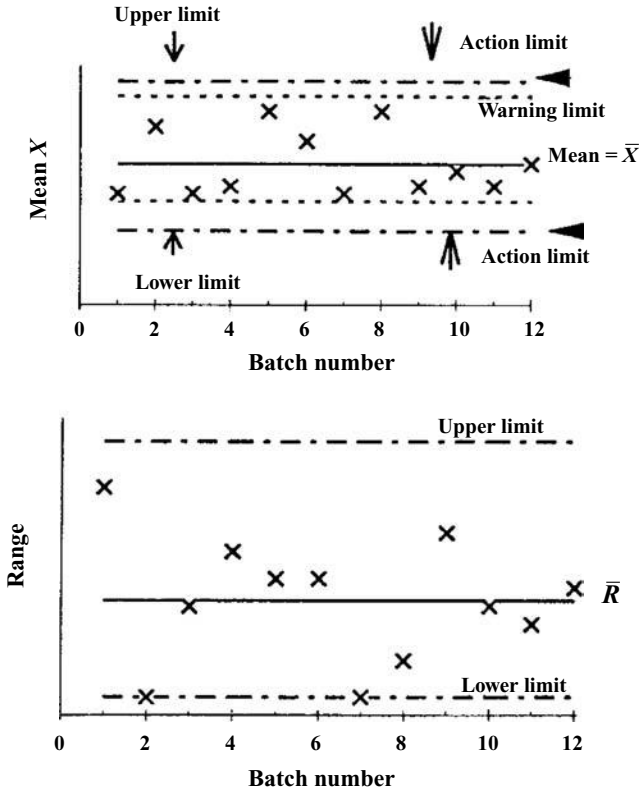


Figure 12.1 Quality control $\overline{X}$ and range charts.

state of statistical control and detecting *trends* or changes in the process average and variability. One can visualize such trends (mean and range) easily with the use of the control chart. The "consistency" of the data as reflected by the deviations from the average value is not only easily seen, but the chart provides a record of batch performance. This record is useful for regulatory purposes as well as for an in-house source of data.

As will be described subsequently, variability can be calculated on the basis of the standard deviation or the range. The range is easier to calculate than the standard deviation. Remember: The range is the difference between the lowest and highest value. If the sample size is not large (<10), the range is an efficient estimator of the standard deviation. Figure 12.1 shows an example of an "$\overline{X}$" (X bar or average) and "range" chart for tablet weights determined from consecutive tablet production batches.

12.2.2.1 Rational Subgroups

The question of how many tablets to choose at each sampling time (*rational subgroups*) and how often to sample is largely dependent on the nature of the process and the level of precision required. The larger the sample and the more frequent the sampling, the greater the precision, but also the greater will be the cost. If tablet samples are taken and weights averaged over relatively long periods of time, significant fluctuations that may have been observed with samples taken at shorter time intervals could be obscured. The subgroups should be as homogeneous as possible relative to the overall process. Subgroups are usually (but not always) taken as units manufactured close in time. For example, in the case of tablet production, consecutively manufactured tablets may be chosen for a subgroup. If possible, the subgroup sample size should be constant. Otherwise, the construction and interpretation of the control chart is more difficult. Four to five items per subgroup is usually an adequate sample size. Procedures for selecting samples should be specified under SOPs (standard operating procedures) in the quality control manual. In our example, 10 consecutive tablets are individually weighted at approximately one-hour intervals. Here the subgroup sample size is larger than the "usual" four or five, principally because of the simple and inexpensive measurement (weighing tablets). The average weight and range are calculated for each of the subgroup samples. One should understand that under ordinary circumstances the variation between individual items (tablets in this example) within a subgroup is due only to chance causes, as noted above. In the example, the 10 consecutive tablets are made almost at the same time. The granulation characteristics and tablet press effects are similar for these 10 tablets. Therefore, the variability observed can be attributed to causes that are not under our control (i.e., the inherent variability of the process).

12.2.2.2 Establishing Control Chart Limits

The principal use of the control chart is as a means of monitoring the manufacturing process. As long as the mean and range of the 10 tablet samples do not vary "too much" from subgroup to subgroup, the product is considered to be in control. To be "in control" means that the observed variation is due only to the random, uncontrolled variation inherent in the process, as discussed previously. We will define *upper and lower limits* for the mean and range of the subgroups. Values falling outside these limits are cause for alarm. The construction of these limits is based on normal distribution theory. We know, from chapter 3, that individual values from a normal distribution will be within 1.96 standard deviations of the mean 95% of the time, and within 3.0 (or 3.09) standard deviations of the mean 99.73% (or 99.8%) of the time (see Table IV.2). Therefore, the probability of observing a value outside these limits is small; only 1 in 20 in the former case and 2.7 in 1000 in the latter case. Two limits are often used in the construction of X (mean) charts as "warning" and "action" limits, respectively (Fig. 12.1). The warning limits are narrower than the action limits and do not require immediate action. If a process is subject only to random, chance variation, a value far from the mean is unlikely. In particular, a value more than 3.0 standard deviations from the mean is highly unlikely (2.7/1000), and can be considered to be probably due to some *systematic*, *assignable* cause. Such a "divergent" observation should signal the quality control unit to modify the process and/ or initiate an investigation into its cause. Of course, the "aberrant" value may be due only to chance. If so, subsequent means should fall close to the process average as expected. In some circumstances, one may wisely

Table 12.1 Tablet Weights and Ranges from a Tablet Manufacturing Process[a]

Date	Time	Mean, $\overline{X}$	Range
3/1	11 a.m.	302.4	16
	12 p.m.	298.4	13
	1 p.m.	300.2	10
	2 p.m.	299.0	9
3/5	11 a.m.	300.4	13
	12 p.m.	302.4	5
	1 p.m.	300.3	12
	2 p.m.	299.0	17
3/9	11 a.m.	300.8	18
	12 p.m.	301.5	6
	1 p.m.	301.6	7
	2 p.m.	301.3	8
3/11	11 a.m.	301.7	12
	12 p.m.	303.0	9
	1 p.m.	300.5	9
	2 p.m.	299.3	11
3/16	11 a.m.	300.0	13
	12 p.m.	299.1	8
	1 p.m.	300.1	8
	2 p.m.	303.5	10
3/22	11 a.m.	297.2	14
	12 p.m.	296.2	9
	1 p.m.	297.4	11
	2 p.m.	296.0	12

[a]Data are the average and range of 10 tablets.

make an observation on a new subgroup before the scheduled time, in order to verify the initial result. If two successive averages are outside the acceptable limits, chances are extremely high that a problem exists. An investigation to detect the cause and make a correction may then be initiated.

The procedure for constructing control charts will be illustrated using data on tablet weights as shown in Table 12.1 and Figure 12.2. Note that the $\overline{X}$ chart consists of an "average" or "standard" line along with upper and lower lines that represent the *action* lines. The *average* line may be determined from the history of the product, with regular updating, or may be determined from the product specifications. In this example, the average line is defined by the quality control specifications (standards) for this product, a target value of 300 mg. The *action* lines are constructed to represent ± 3 standard deviations from the target value. This is also known as "3σ limits." Observations that lie outside these limits are a cause for action. Adjustments or other corrective action should *not* be implemented if the averages are within the action limits. Tampering with equipment and/or changing other established procedures while the process remains within limits should be avoided. Such interference will often result in increased variation.

In order to establish the *upper* and *lower* limits for the mean ($\overline{X}$), we need an estimate of the standard deviation, if it is not previously known. The standard deviation can be obtained from the replicates (10 tablets) of the subgroup samples that generate the means for the control chart. By pooling the variability from many subgroups ($N = 10$), a very good estimate of the true standard deviation, σ, can be obtained (see App. I). Note that an estimate of the standard deviation or range is needed before limits for the $\overline{X}$ chart can be established. If a "range" chart is used in conjunction with the $\overline{X}$ chart, the upper and lower limits for the $\overline{X}$ chart can be obtained from the range according to Table IV.10 (column A). These factors are derived from theoretical calculations relating the range and standard deviation. For example, in the long run, the range can be shown to be equal to 3.078 times the standard deviation for samples of size 10. If we wish

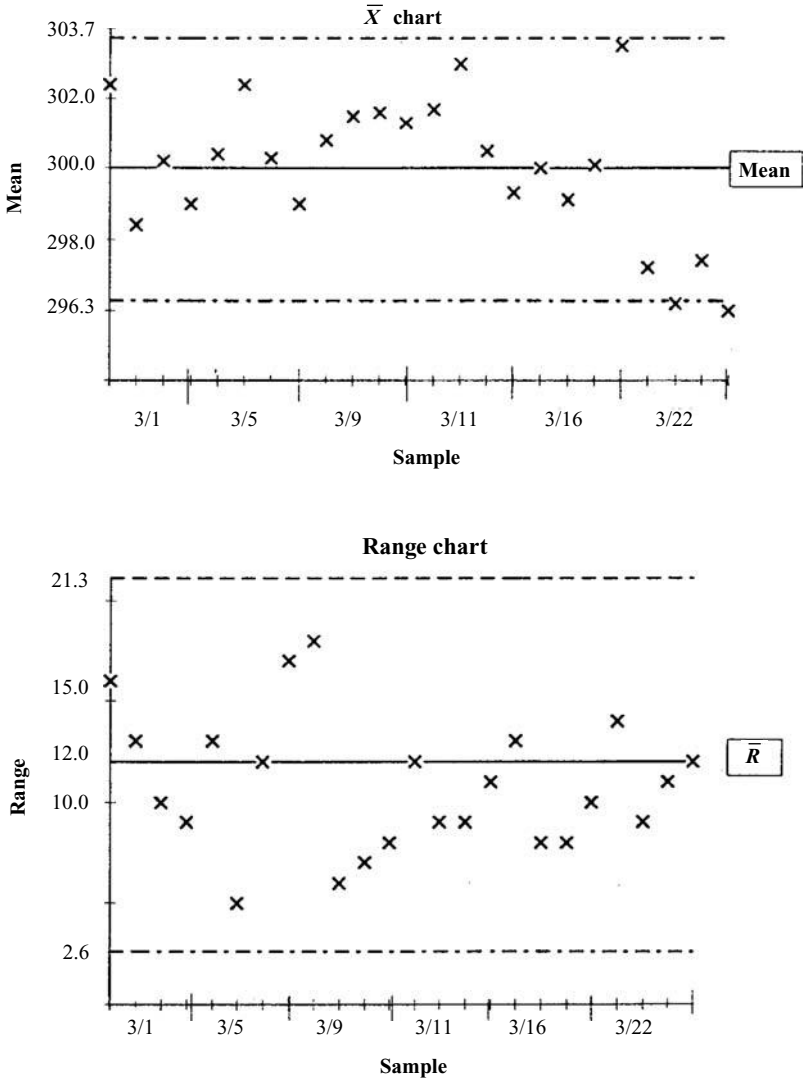


Figure 12.2 Control chart for tablet averages and range data from Table 12.1.

to establish 3σ limits about the *mean* of samples of size 10 (s.d. $= \sigma/\sqrt{10}$) using the range, the following relationship leads to the value 0.31 in Table IV.10 (see column A):

$$\overline{X} \pm \frac{3\sigma}{\sqrt{10}} = \overline{X} \pm \frac{3(\overline{R})}{(3.078)\sqrt{10}} = \overline{X} \pm 0.31\overline{R},$$

where $\overline{R}/3.078$ is the average range divided by 3.078, which on the average is equal to σ. Thus, if the average range is 12 for samples of size 10, the upper and lower control chart limits for $\overline{X}$ are

$$\overline{X} \pm 0.31\overline{R} = \overline{X} \pm 0.31(12) = \overline{X} \pm 3.72. \tag{12.1}$$

Note that the average range is simply the usual average of the range values, obtained in a manner similar to that for calculating the process average. Ranges obtained during the control charting process are averaged and updated as appropriate.

Table IV.10 also has factors for upper and lower limits for a *range chart*. The values in columns D_L and D_U are multiplied times the average range to obtain the lower and upper limits for the range. Usually, a range that exceeds the upper limit is a cause for action. A small value of the range shows good precision and may be disregarded in many situations. In the present example, the average range is set equal to 12 based on previous experience. For samples of size 10, D_L and D_U are 0.22 and 1.78, respectively. Therefore, the lower and upper limits for the range are

$$\text{Lower limit: } 0.22 \times 12 = 2.6$$

$$\text{Upper limit: } 1.78 \times 12 = 21.3. \tag{12.2}$$

These limits are shown in the control chart for the range in Figure 12.2. See Figure 12.1 for another example of a range chart. Ordinarily, the sample size should be kept constant. If sample size varies from time to time, the limits for the control chart will change according to the sample size. If the sample sizes do not vary greatly, one solution to this problem is use an average sample size [2].

Having established the *mean* and the average *range*, the process is considered to be under control as long as the average and range of the subgroup samples fall within the lower and upper limits. If either the mean or range of a sample falls outside the limits, a possible "assignable" cause is suspected. The reason for the deviation should be investigated and identified, if possible. One should appreciate that a process can change in such a way that (a) only the average is affected, (b) only the variability is affected, or (c) both the average and variability are affected. These possibilities are illustrated in Figure 12.3.

In the example of tablet weights, one might consider the following as possible causes for the results shown in Figure 12.3. A change in average weight may be caused by a misadjustment of the tablet press. Increased variability may be due to some malfunction of one or more punches. Since 10 consecutive tablets are taken for measurement, if one punch gives very low weight tablets, for example, a large variability would result. A combination of lower weight and increased variability probably would be quickly detected if half of the punches were *sticking* in a random manner. Under these circumstances, the average ($\overline{X}$) would be substantially reduced and the range would be substantially increased relative to the values expected under *statistical control*.

The control charts shown in Figure 12.2 are typical. For the $\overline{X}$ chart, the mean was taken as 300 mg based on the target value as set out in the quality control standards. The upper and lower action limits were calculated on the basis of an average range of 12 and factor A in Table IV.10. The lower and upper action limits are 300 ± 3.72 mg or approximately 296.3 to 303.7 mg,

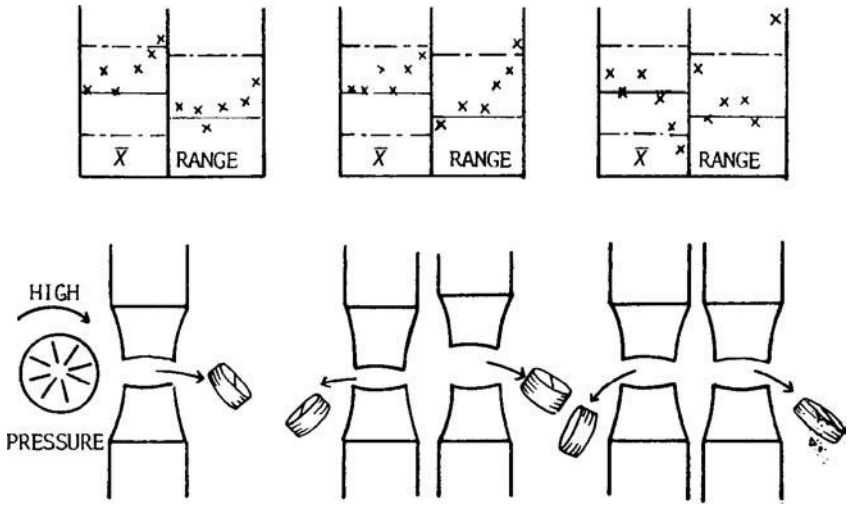


Figure 12.3 Representation of possible process changes as may be detected in a control chart procedure.

respectively. The process is out of control during the production of the batch produced on 3/22. This will be discussed further below. The range control chart shows that the process is in control with respect to this variable.

When the *standard deviation* rather than the range is computed for purposes of constructing control charts, the factors for calculating the limits for the $\overline{X}$ chart are different. The variability is monitored via a chart of the standard deviation of the subgroup rather than the range. Factors for setting limits for both $\overline{X}$ charts and "sigma" (standard deviation) charts may be found in Ref. [1].

If an outlying observation ($\overline{X}$, R) is eliminated because an assignable cause has been found, that observation should be eliminated from future updating of the $\overline{X}$ and R charts.

12.2.3 Between-Batch Variation as a Measure of Variability (Moving Averages)

The discussion of control charts above dealt with a system that is represented by a regular schedule of production batches. The action limits for $\overline{X}$ were computed using the "within"-batch variation as measured by the variability between items in a "rational subgroup." The subgroup consists of a group of tablets manufactured under very similar conditions. For the manufacture of unit dosage forms with inherent heterogeneity, such as tablets, attempts to construct control charts that include different batches, based on within-subgroup variation, may lead to apparently excessive product failure and frustration. Sometimes, this unfortunate situation may result in the discontinuation of the use of control charts as an impractical statistical device. However, the nature of the manufacture of a heterogeneous mixture, such as the bulk granulations used for manufacturing tablets, lends itself to new sources of uncontrolled error. This error resides in the variability due to the different (uncontrolled) conditions under which different tablet batches are manufactured. One would be hard put to describe exactly why batch-to-batch differences should exist, or to identify the sources of these differences. Perhaps the dies and punches of the tablet press are subject to wear and erosion. Perhaps a new employee involved in the manufacturing process performs the job in a slightly different manner from his or her predecessor. Whatever the reason, such interbatch variation may exist.* In these cases, the within-subgroup variation underestimates the variation, and many readings will appear out of control. This is exemplified by the last batch in Table 12.1 and Figure 12.2.

Thus, when significant interbatch variation exists, the usual control chart will lead to many batches being out of control. If the cause of this variation cannot be identified or controlled, and the product consistently passes the official quality control specifications, other methods than the usual control chart may be used to monitor the process.

Use of the "Control Chart for Individuals" [1,2] seems to be one reasonable approach to monitoring such processes. The limits for the $\overline{X}$ chart are based on a moving range using two consecutive samples (Table 12.2). For example, the first value for the two-batch moving range is the range of batches 1 and $2 = 1.1(399.5 - 398.4)$. The second moving range is $399.5 - 398.8 = 0.7$, and so on. The average moving range is 1.507. The average tablet weight of the 30 batches is 400.01. The average range is based on samples of 2. To estimate the standard deviation from the average range of samples of size 2, it can be shown that we should divide the average range by 1.128 (Table IV.10). The 3 sigma limits are $\overline{X} \pm 3(\overline{R}/1.128) = 400.01 \pm 3(1.507/1.128) = 400.01 \pm 4.01$. The range chart has an upper limit of $3.27(1.507) = 4.93$. These charts are shown in Figure 12.4. Batch 13 is out of limits based on both the average and range charts.

The moving average method is another approach to construct control charts that can be useful in the presence of interbatch variation. In this method, we use only a single mean value for each batch, ignoring the individual values within the subgroup, if they are available. Thus, the data consist of a series of means over many batches as shown in Table 12.2. A three-batch moving average consists of averaging the present batch with the two immediately preceding batches. For example, starting with batch 3, the first value for the moving average chart is

$$\frac{398.4 + 399.5 + 398.8}{3} = 398.9.$$

* Process validation investigates and identifies such variation.

Table 12.2 Average Weight of 50 Tablets from 30 Batches of a Tablet Product: Example of the Moving Average

Batch	Batch average (mg)	Two-batch moving range	Three-batch moving average	Three-batch moving range
1	398.4	—	—	—
2	399.5	1.1	—	—
3	398.8	0.7	398.9	1.1
4	397.4	1.4	398.6	2.1
5	402.7	5.3	399.6	5.3
6	400.5	2.2	400.2	5.3
7	401.0	0.5	401.4	2.2
8	398.5	2.5	400.0	2.5
9	399.5	1.0	399.7	2.5
10	400.1	0.6	399.4	1.6
11	399.0	1.1	399.5	1.1
12	401.7	2.7	400.3	2.7
13	395.4	6.3	398.7	6.3
14	400.7	5.3	399.3	6.3
15	401.6	0.9	399.2	6.2
16	401.4	0.2	401.2	0.9
17	401.5	0.1	401.5	0.2
18	400.4	1.1	401.1	1.1
19	401.0	0.6	401.0	1.1
20	402.1	1.1	401.2	1.7
21	400.9	1.2	401.3	1.2
22	400.8	0.1	401.3	1.3
23	401.5	0.7	401.1	0.7
24	398.6	2.9	400.3	2.9
25	398.4	0.2	399.5	3.1
26	398.8	0.4	398.6	0.4
27	399.9	1.1	399.0	1.5
28	400.9	1.0	399.9	2.1
29	399.9	1.0	400.2	1.0
30	399.5	0.4	400.1	1.4

The second value is $(399.5 + 398.8 + 397.4)/3 = 398.6$. The calculation is similar to that used for the two-batch moving range in the example of the Control Chart for Individuals. The moving average values are plotted as in the ordinary control chart. Limits for the control chart are established from the moving range, which is calculated in a similar manner. The range of the present and the two immediately preceding batches is calculated for each batch. The average of these ranges is $\overline{R}$, the limits for the control chart are computed from Table IV.10. The computations of the moving average and range for samples of size 3 are shown in Table 12.2, and the data charted in Figure 12.5. The average weight was set at the targeted weight of 400 mg. The average moving range (from Table 12.2) is 2.35. The limits for the moving average chart are determined using the average range and the factor from Table IV.10 for samples of size 3.

$$400 \pm 1.02(2.35) = 400 \pm 2.4.$$

All of the moving average values fall within the limits based on the average moving range. In this analysis, the suspect batch number 13 is "smoothed" out when averaged with its neighboring batches. The upper limit for the range chart is $2.57(2.35) = 6.04$, which would be a cause to investigate the conditions under which batch number 13 was produced (Table 12.2). For further details of the construction and interpretation of moving average charts, see Refs. [1,3].

Another approach to the problem of between-batch variation is the difference chart. A good standard lot is set aside as the control. Each production lot is compared to the standard lot

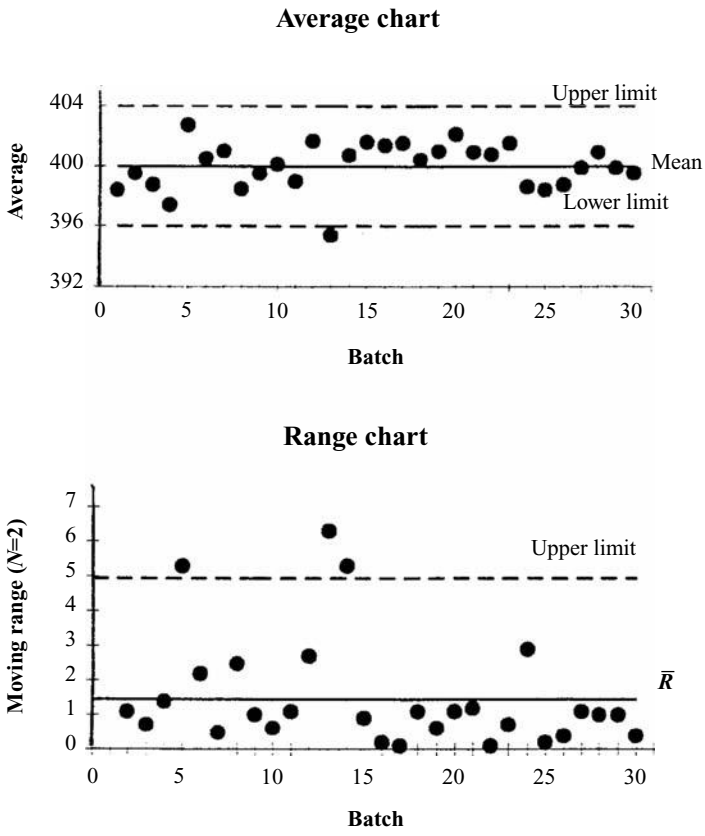


Figure 12.4 Control charts for individuals from Table 12.2.

by taking samples of each. Both the control and production lots are measured and the difference of the means is plotted. The limits are computed as

$$0 \pm \frac{3}{\sqrt{n}}\sqrt{S_c^2 + S_p^2},$$

where S_c^2 and S_p^2 are the estimates of the variances of the control and production lots, respectively.

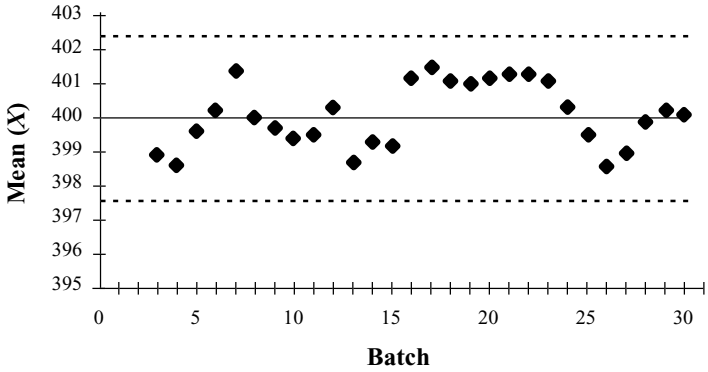


Figure 12.5 Moving average plot for tablet weight means from Table 12.2.

12.2.4 Quality Control Charts in Research and Development

Control charts may be advantageously conceived and used during assay development and validation, in preliminary research or formulation studies, and in routine pharmacological-screening procedures. During the development of assay methodology and validation, for example, by keeping records of assay results, an initial estimate of the assay standard deviation is available. The initial estimate can then be updated as data accumulate.

The following example shows the usefulness of control charts for control measurements in a drug-screening procedure. This test for screening potential anti-inflammatory drugs measures improvement of inflammation (guinea pig paw volume) by test compounds compared to a control treatment. A control chart was established to monitor the performance of the control drug (a) to establish the mean and variability of the control, and (b) to ensure that the results of the control for a given experiment are within reasonable limits (a validation of the assay procedure). The average paw volume difference (paw volume before treatment–paw volume after treatment) and the average range for a series of experiments are shown in Table 12.3. The control chart is shown in Figure 12.6.

As in the control charts for quality control, the mean and average range of the "process" were calculated from previous experiments. In this example, the screen had been run 20 times previous to the data of Table 12.3. These initial data showed a mean paw volume difference of 40 and a mean range ($\overline{R}$) of 9, which were used to construct the control charts shown in Figure 12.6. The subgroups consist of four animals each. Using Table IV.10, the action limits for the $\overline{X}$ and range charts were calculated as follows:

$$\overline{X} \pm 0.73\overline{R} = 40 \pm 0.73(9) = 33.4 \text{ to } 46.6 \ (\overline{X} \text{ chart})$$

$$\overline{R}(2.28) = 9(2.28) = 20.5 \text{ the upper limit for the range.}$$

Note that the lower limit for the range of subgroups consisting of four units is zero. Six of the 20 means are out of limits. Efforts to find a cause for the larger intertest variation failed. The procedures were standardized and followed carefully, and the animals appeared to be homogeneous. Because different shipments of animals were needed to proceed with these tests

Table 12.3 Average Paw Volume Difference and Range for a Screening Procedure (Four Guinea Pigs Per Test Group)

Test number	Mean	Range
1	38	4
2	43	3
3	34	3
4	48	6
5	38	24
6	45	4
7	49	5
8	32	9
9	48	5
10	34	8
11	28	12
12	41	10
13	40	22
14	34	5
15	37	4
16	43	14
17	37	6
18	45	8
19	32	7
20	42	13

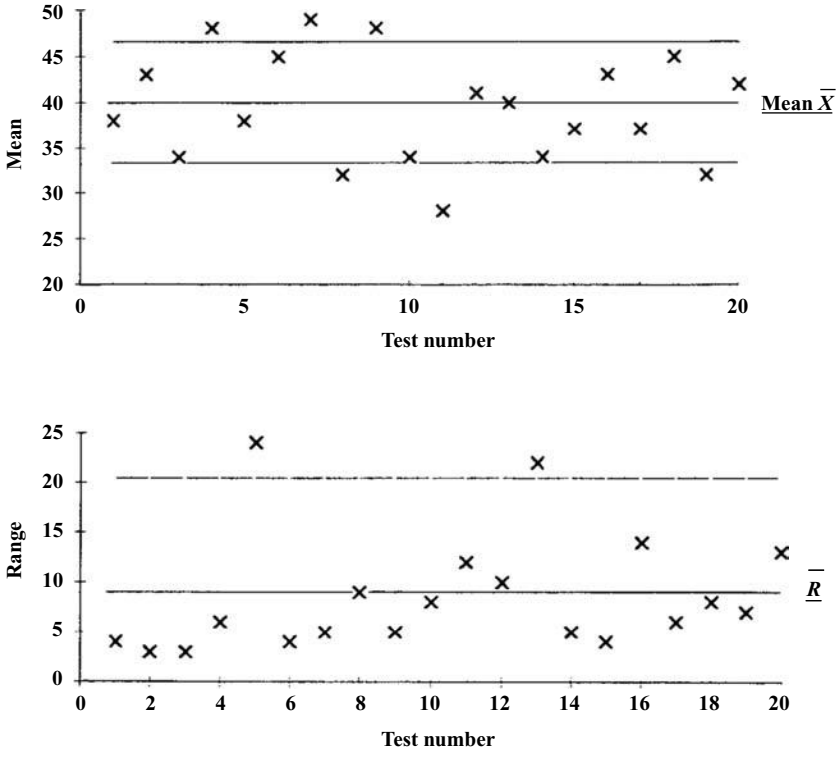


Figure 12.6 Control chart for means and range for control group in a pharmacological-screening procedure.

over time, the researchers felt that there was no way to "tighten up" the procedure. Therefore, as in the tablet weight example discussed in the preceding section, a new control chart was prepared based on the variability between test means. A moving average was recommended using *four* successive averages. Based on historical data, $\overline{X}$ was calculated as 39.7 with an average moving range of 12.5. The limits for the moving average graph are

$$39.7 \pm 0.73(12.5) = 30.6 \text{ to } 48.8.$$

The factor 0.73 is obtained from Table IV.10 for subgroup samples of size 4.

12.2.5 Control Charts for Proportions

Table 12.4 shows quality control data for the inspection of tablets where the measurement is an attribute, a binomial variable. Three hundred tablets are inspected each hour to detect various problems, such as specks, chips, color uniformity, logo, and so on. For this example, the defect

Table 12.4 Proportion of Chipped Tablets of 300 Inspected During Tablet Manufacture

	Time			
Batch	**10 a.m.**	**11 a.m.**	**12 p.m.**	**1 p.m.**
1	0.060	0.053	0.087	0.055
2	0.073	0.047	0.060	0.047
3	0.040	0.067	0.033	0.053
4	0.033	0.040	0.030	0.027
5	0.040	0.013	0.023	0.040
6	0.025	0.000	0.027	0.013

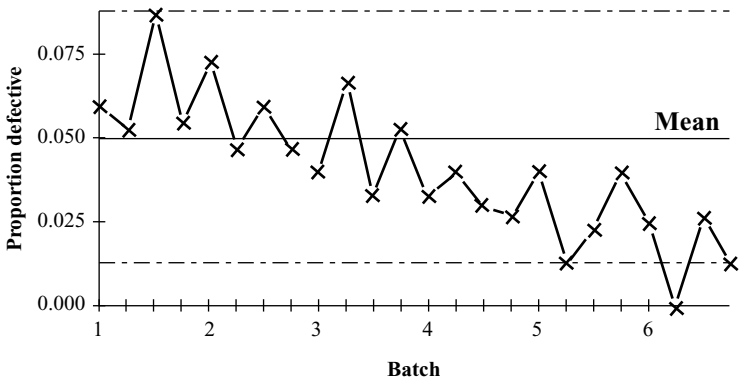


Figure 12.7 Control chart for proportion of tablets chipped.

under consideration is a chipped tablet. According to quality control specifications, this type of defect is considered of minor importance and an average of 5% chipped tablets is tolerable. This problem of chipped tablets was of recent origin, and the control chart was implemented as an aid to the manufacturing and research and development departments, who were looking into the cause of this defect. In fact, the 5% average had been written into the specifications as a result of the persistent appearance of the chipped tablets in recent batches. The data in Table 12.4 represent the first six batches where this attribute was monitored.

For the control chart, 5% defects was set as the average value. The action limits can be calculated from the standard deviation of a binomial. In this example, where 300 tablets were inspected, $N = 300$, $p = 0.05$, and $q = 0.95$ [$\sigma = \sqrt{pq/N}$, Eq. (3.11)].

$$\sigma = \sqrt{\frac{(0.05)(0.95)}{300}} = 0.0126.$$

The limits are $0.05 \pm 3\sigma = 0.05 \pm 3(0.0126) = 0.012$ to 0.088. Proportions below the lower limit indicate an improvement in the process in this example. Note that we can use the normal approximation to the binomial when calculating the 3σ limits, because both NP and Nq are greater than 5 (see sect. 3.4.3). The control chart is shown in Figure 12.7.

The chart clearly shows a trend with time toward less chipping. The problem seems to be lessening. Although no specific cause was found for this problem, increased awareness of the problem among manufacturing personnel may have resulted in more care during the tableting process.

12.2.6 Runs in Control Charts

The most important feature of the control chart is the monitoring of a process based on the average and control limits. In addition, control charts are useful as an aid in detecting trends that could be indicative of a lack of control. This is most easily seen as a long consecutive series of values that are within the control limits but (a) stay above (or below) the average or (b) show a steady increase (or decline). Statistically, such occurrences are described as "runs." For example, a run of 7 successive values that lie above the average constitutes a run of size 7. Such an event is probably not random because if the observed values are from a symmetric distribution and represent random variation about a common mean, the probability of 7 successive values being above the mean is $(1/2)^7 = 1/128$. In fact, the occurrence of such an event is considered to be suggestive of a trend and the process should be carefully watched or investigated.

In general, when looking for runs in a long series of data, the problem is that significant runs will be observed by chance when the process is under control. Nevertheless, with this understanding, it is useful to examine data to be forewarned of the possibility of trends and potential problems. The test for the number of runs above and below the median of a consecutive series of data is described in section 15.7. For the consecutive values 9.54, 9.63, 9.42, 9.86, 9.40, 9.31, 9.79, 9.56, 9.2, 9.8, and 10.1, the median is 9.56. The number of runs above and below the

median is 8. According to Table IV.14, this is not an improbable event at the 5% level. If the consecutive values observed were 9.63, 9.86, 9.79, 9.8, 10.1, 9.56, 9.54, 9.42, 9.40, 9.31, and 9.2, the median is till 9.56, but the number of runs is 2. This shows a significant lack of randomness ($p < 0.05$). Also see Exercise Problem 12.

Duncan [2] describes a runs test that looks at the longest run occurring above or below the median. The longest run is compared to the values in Table IV.15. If the longest run is equal to or greater than the table value, the data are considered to be nonrandom. For the data of Table 12.1, starting with the data on the date 3/5 (ignore the data on 3/1 for this example), the median is 300.35. The longest run is 7. There are seven consecutive values above the median starting at 11 a.m. on 3/9. For $N = 20$, the table value in Table IV.15 is 7, and the data are considered to be significantly nonrandom ($p < 0.05$). Note that this test allows a decision of lack of control at the 5% level if a run of 7 is observed in a sequence of 20 observations.

For other examples of the application of the runs test, see Ref. [2]. Also see section 15.7 and Exercise Problem 11 in chapter 15.

In addition to the aforementioned criteria, that is, a point outside the control limits, a significant number of runs, or a single run of sufficient length, other rules of thumb have been suggested to detect lack of control. For example, a run of 2 or 3 outside the 2σ limits but within the 3σ limits, and runs of 4 or 5 between $l\sigma$ and 2σ limits can be considered cause for concern.

Cumulative sum control charts (cusum charts) are more sensitive to process changes. However, the implementation, construction, and theory of cusum charts are more complex than the usual Shewhart control chart. Ref. [4] gives a detailed explanation of the use of these control charts.

For more examples of the use of control charts, see chapter 13.

12.3 ACCEPTANCE SAMPLING AND OPERATING CHARACTERISTIC CURVES

Finished products or raw materials (including packaging components) that appear as separate units are inspected or analyzed before release for manufacturing purposes or commercial sale. The sampling and analytical procedures are specified in official standards or compendia (e.g., the USP), or in in-house quality control standards. The quality control procedure known as *acceptance sampling* specifies that a number of items be selected according to a scheduled sampling plan, and be inspected for attributes or quantitatively analyzed. The chief purpose of acceptance sampling is to make a decision regarding the acceptability of the material. Therefore, based on the inspection, a decision is made, such as "the material or lot is either accepted or rejected." Sampling plans for variables (quantitative measurements such as chemical analyses for potency) and attributes (qualitative inspection) are presented in detail in the U.S. government documents MIL-STD-414 and MIL-STD-105E, respectively [3,5].

A single *sampling plan for attributes* is one in which N items are selected at random from the population of such items. Each item is classified as defective or not defective with respect to the presence or absence of the attribute(s). If the sample size is small relative to the population size, this is a binomial process, and the properties of sampling plans for attributes can be derived using the binomial distribution. For example, consider the inspection of finished bottles of tablets for the presence of an intact seal. This is a binomial event; the seal is either intact or it is not intact. The sampling plan states the number of units to be inspected and the number of defects which, if found in the sample, leads to rejection of the lot. A typical plan may call for inspection of 100 items; if two or more are defective, reject the lot (batch). If one or less are defective, accept the lot. (The acceptance number is equal to one.) Theoretically, "100% inspection" will separate the good and defective items (seals in our example). In the absence of 100% inspection, there is no guarantee that the lot will have 0% (or any specified percentage) defects. Thus, underlying any sampling plan are two kinds of risks:

1. The *producer's or manufacturer's risk.* This is the risk or probability of rejecting (not releasing) the product, although it is really good. By "good" we mean that had we inspected every item, the batch would meet the criteria for release or acceptance. This risk reflects an unusually high number of defects appearing in the sample taken for inspection, by chance. The producer's risk can be likened to the α error, that is, rejecting the batch, even though it is good.

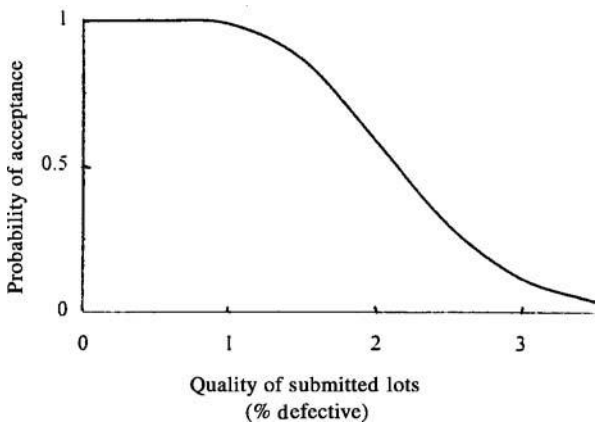


Figure 12.8 Operating characteristic curve for sampling plan *N*: sample 500 items—accept if 10 or less defective.

2. The *consumer's risk.* This is the probability that the product is considered acceptable (released), although, in truth, it would not be acceptable were it 100% inspected. The consumer's risk can be likened to the β error, that is, the batch is accepted even though it has a more than the acceptable number of defects.

There are any number of possible plans that, in addition to economic considerations, depend on

1. the number of items sampled;
2. the producer's risk;
3. the consumer's risk.

MIL-STD-105E is an excellent compilation of such plans [3]. Each plan gives the number of items to be inspected, and the number of defects in the sample needed to cause rejection of the lot. Each plan is accompanied by an *operating characteristic* (OC) *curve.* The OC curve shows the probability of accepting a lot based on the sampling plan specifications, given the true proportion of defects in the lot. A typical OC curve is shown in Figure 12.8.

The OC curve is a form of power curve (see sect. 6.5). The OC curve in Figure 12.8 is derived from a sampling plan (plan *N* from MIL-STD-105E) in which 500 items (bottles) are inspected from a lot that contains 30,000 items. If 11 or more items inspected are found to be defective, the lot is rejected. Inspection of Figure 12.8 shows that if the batch truly has 1% defect, the probability of accepting the lot is close to 99% when plan *N* is implemented. This plan is said to have an *acceptable quality level* (AQL) of 1%. An AQL of 1% means that the consumer will accept most of the product manufactured by the supplier if the level of defects is not greater than 1%, the specified AQL (i.e., 1%). In this example, with the AQL equal to approximately 1%, about 99% of the batches will pass this plan if the percent defects is 1% or less.

The plan actually chosen for a particular product and a particular attribute depends on the lot size and the nature of the attribute. If the presence (or absence) of an attribute (such as the integrity of a seal) is critical, then a stringent plan (a low AQL) should be adopted. If a defect is considered of minor importance, inspection for the presence of a defect can make use of a less stringent plan. MIL-STD-105E describes various plans for different lot (population) sizes, which range from less stringent for minor defects to more stringent for critical defects. These are known as levels of inspection, level I, II, or III. This document also includes criteria for contingencies for switching to more or less tight plans depending on results of prior inspection. A history of poor quality will result in a more stringent sampling plan and vice versa. If 2 of 2, 3, 4 or 5 consecutive lots are rejected, the normal plan is switched to the tightened plan. If five consecutive lots are accepted under the tightened plan, the normal plan is reinstated. If quality remains very good, reduced plans may be administered as described in MIL-STD-105E. The characteristics of the plan are defined by the AQL and the OC curve. For example,

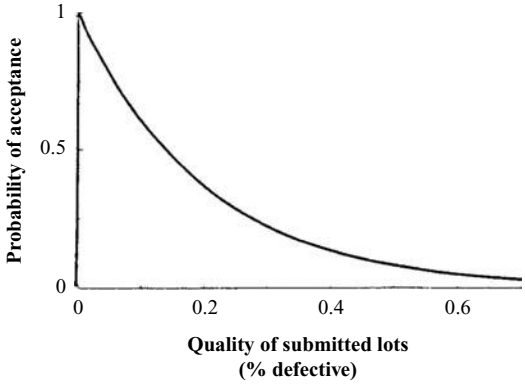


Figure 12.9 Operating characteristic curve for plan *N*: AQL = 0.025%.

for lot sizes of 10,001 to 35,000, the following are two of the possible plans recommended by MIL-STD-105E:

		Reject number[a] if AQL =	
Plan	**Sample size**	**0.4%**	**1%**
K	125	2	4
N	500	6	11

[a]Reject the lot if the number of defects (or more) are observed.

Plan *N* is a more "discriminating" plan than plan *K*. The larger sample size results in a greater probability of rejecting lots with more than AQL percentage of defects. For plan *N*, if there are 2% defects in the lot, the lot will be accepted approximately 57% of the time. For plan *K*, with 2% defects in the lot, the lot will be accepted 75% of the time. (See MIL-STD-105E [3] for OC curves. The OC curve for an AQL of 1% for plan *N* is shown in Fig. 12.8.)

In the present example, a defective seal is considered a critical defect and plan *N* will be implemented with an AQL of 0.025%. This means that lots with 0.025% (25 defects per 100,000 bottles) are considered acceptable. According to MIL-STD-105E, if one or more defects are found in a sample of 500 bottles, the lot is rejected.† This means that the lot is passed only if all 500 bottles are good. The OC curve for this plan is shown in Figure 12.9.

The calculations of the probabilities needed to construct the OC curve are not very difficult. These calculations have been presented in the discussion of the binomial distribution in chapter 3. As an illustration, we will calculate the probability of rejecting a lot using plan *N* with an AQL of 0.025%. As noted above, the lot will be rejected if one or more defects are observed in a sample of 500 items. Thus, the probability of accepting a lot with 0.025% defects is the probability of observing zero defects in a sample of 500. This probability can be calculated from Eq. (3.9)

$$\binom{N}{X} P^X q^{N-X} = \binom{500}{0} P^0 q^{500} = (0.00025)^0 (0.99975)^{500} = 0.88,$$

where 500 is the sample size, P the probability of a defect (0.00025), and q the probability of observing a bottle with an intact seal (0.99975). Thus, using this plan, lots with *0.025% defects will be passed 88% of the time.* A lot with 0.4% (4 defects per 1000 items) will be accepted with a probability of

$$\binom{500}{0} (0.004)^0 (0.996)^{500} = 0.13 \text{ (i.e., 13\%)}.$$

† If the result of inspection calls for rejection, 100% inspection is a feasible alternative to rejection.

Copies of sampling plans *K* and *N* from MIL-STD-105E are shown in Tables 12.5 and 12.6.

In addition to the sampling plans discussed above, MIL-STD-105E also presents *multiple-sampling plans*. These plans use less inspection than single sampling plans, on the average. After the first sampling, one of three decisions may be made:

1. Reject the lot
2. Accept the lot
3. Take another sample

In a double-sampling plan, if a second sample is necessary, the final decision of acceptance or rejection is based on the outcome of the second sample inspection.

The theory underlying acceptance sampling for *variables* is considerably more complex than that for sampling for attributes. In these schemes, actual measurements are taken, such as assay results, dimensions of tablets, weights of tablets, measurement of containers, and so on. Measurements are usually more time consuming and more expensive than the observation of a binomial attribute. However, quantitative measurements are usually considerably less variable. Thus, there is a trade-off between expense and inconvenience, and precision. Many times, there is no choice. Official procedures may specify the type of measurement. Readers interested in plans for variable measurements are referred to MIL-STD-414 [5] and the book, "Quality Control and Industrial Statistics" [2] for details.

12.4 STATISTICAL PROCEDURES IN ASSAY DEVELOPMENT

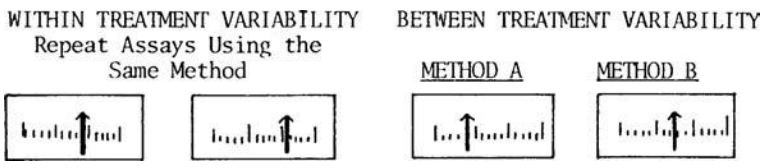
Statistics can play an important role in assisting the analytical chemist in the development of assay procedures. A subcommittee of PMA (Pharmaceutical Manufacturers Association) statisticians developed a comprehensive scheme for documenting and verifying the equivalence of alternative assay procedures to a standard [6]. The procedure is called the *Greenbriar procedure* (named after the location where the scheme was developed). This approach includes a statistical design that identifies sources of variation such as that due to different days and different analysts. The design also includes a range of concentration of drug. The Greenbriar document emphasizes the importance of a thoughtful experimental design in assay development, a design that will yield data to answer questions raised in the study objectives. The procedure is too detailed to present here. However, for those who are interested, it would be a good exercise to review this document, a good learning experience in statistical application.

For those readers interested in pursuing statistical applications in assay and analytical development, two books, *Statistical Methods for Chemists* by Youden [7] and *The Statistical Analysis of Experimental Data*, by Mandel [8], are recommended. Both of these statisticians had long tenures with the National Bureau of Standards.

In this book, we have presented some applications of regression analysis in analytical methodology (see chaps. 7 and 13). Here, we will discuss the application of sample designs to identify and quantify factors that contribute to assay variability (components of variance).

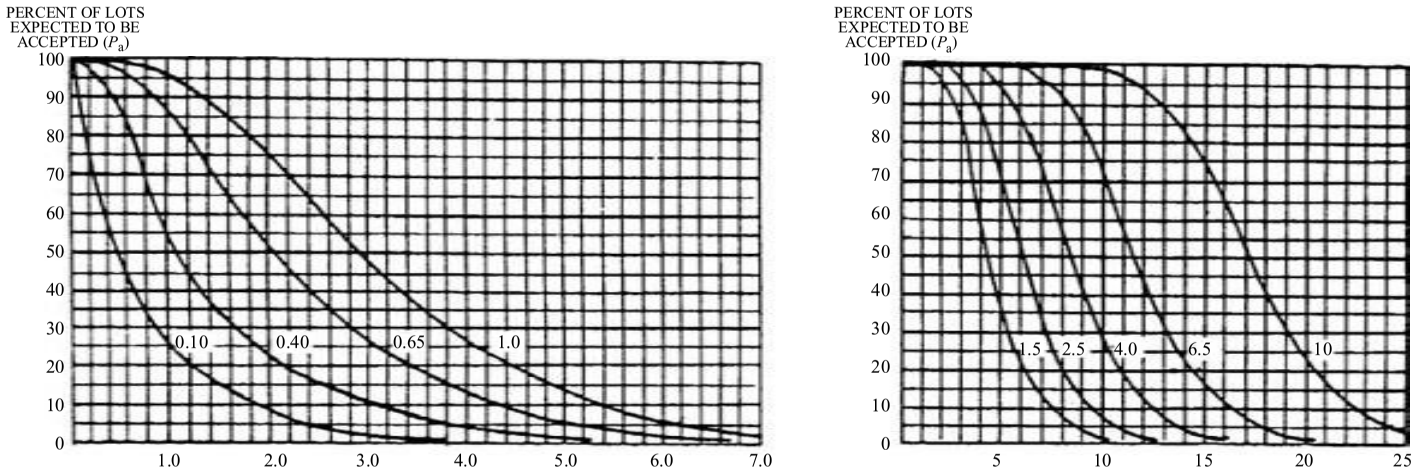
12.4.1 Components of Variance‡

During the discussion of the one-way ANOVA design (sect. 8.1), we noted that the "between-treatment mean square" is a variance estimate that is composed of two different (and *independent*) variances: (a) that due to variability among units *within* a treatment group, and (b) that due to variability due to differences *between* treatment groups. If treatments are, indeed, identical, the ANOVA calculations are such that observed differences between treatment means will probably be accounted for by the within-treatment variation. In the ANOVA table, the ratio of the between-treatment mean square to the within-treatment mean square ($F = \text{BMS/WMS}$) will be approximately equal to 1 on the average when treatments are identical.



‡ A more advanced topic [16].

Table 12.5 Sample Size Code Letter: *K* (Chart Shows Operating Characteristic Curves for Single Sampling Plans)[a]



Tabulated values for operating characteristic curves for single sampling plans

	Acceptable quality levels (normal inspection)											
P_a	0.10	0.40	0.65	1.0	1.5	2.5	*X*	4.0	*X*	6.5	*X*	10
99.0	0.0081	0.119	0.349	0.658	1.43	2.33	2.81	3.82	4.88	5.98	8.28	10.1
95.0	0.0410	0.284	0.654	1.09	2.09	3.19	3.76	4.94	6.15	7.40	9.95	11.9
90.0	0.0840	0.426	0.882	1.40	2.52	3.73	4.35	5.62	6.92	8.24	10.9	13.0
75.0	0.230	0.769	0.382	2.03	3.38	4.77	5.47	6.90	8.34	9.79	12.7	14.9
50.0	0.554	1.34	2.14	2.94	4.54	6.14	6.94	8.53	10.1	11.7	14.9	17.3
25.0	1.11	2.15	3.14	4.09	5.94	7.75	8.64	10.4	12.2	13.9	17.4	20.0
10.0	1.84	3.11	4.26	5.35	7.42	9.42	10.4	12.3	14.2	16.1	19.8	22.5
5.0	2.40	3.80	5.04	6.20	8.41	10.5	11.5	13.6	15.6	17.5	21.4	24.2
1.0	3.68	5.31	6.73	8.04	10.5	12.8	18.3	16.1	18.3	20.4	24.5	27.5
	0.15	0.65	1.0	1.5	2.5	–	4.0	–	6.5	–	10	–
	Acceptable quality levels (tightend inspection)											

Sampling plans for sample size code letter *K*

Type of sampling plan	Cumulative sample size	Acceptable quality levels (normal inspection)																														Cumulative sample size
		Less than 0.10		0.10		0.15		–		0.25		0.40		0.65		1.0		1.5		2.5		–		4.0		–		6.5		–		10		Higher than 10	
		Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	
Single	125	▽		0	1							1	2	2	3	3	4	5	6	7	8	8	9	10	11	12	13	14	15	18	19	21	22	△		125
Double	80	▽		*								0	2	0	3	1	4	2	5	3	7	3	7	5	9	6	10	7	11	9	14	11	16	△		80
	160					Use letter *J*		Use letter *M*		Use letter *L*		1	2	3	4	4	5	6	7	8	9	11	12	12	13	15	16	18	19	23	24	26	27			160
Multiple	32	▽		*								#	2	#	2	#	3	#	4	0	4	0	4	0	5	0	6	1	7	1	8	2	9	△		32
	64											#	2	0	3	0	3	1	5	1	6	2	7	3	8	3	9	4	10	6	12	7	14			64
	96											0	2	0	3	1	4	2	6	3	8	4	9	6	10	7	12	8	13	11	17	13	19			96
	128											0	3	1	4	2	5	3	7	5	10	6	11	8	13	10	15	12	17	16	22	19	25			128
	160											1	3	2	4	3	6	5	8	7	11	9	12	11	15	14	17	17	20	22	25	25	29			160
	192											1	3	3	5	4	6	7	9	10	12	12	14	14	17	18	20	21	23	27	29	31	33			192
	224											2	3	4	5	6	7	9	10	13	14	14	15	18	19	21	22	25	26	32	33	37	38			224
		Less than 0.15		0.15		–		0.25		0.40		0.65		1.0		1.5		2.5		–		4.0		–		6.5		–		10		–		Higher than 10	
		Acceptable quality levels (tightened inspection)																																	

[a] Quality of submitted lots (p, in percent defective for AQLs ≤ 10; in defects per hundred units for AQLs > 10).

Note: Figures on curves are acceptable quality levels (AQLs) for normal inspection. Curves for double and multiple sampling are matched as closely as practicable.

△ = Use next preceding sample size code letter for which acceptance and rejection numbers are available.

▽ = Use next subsequent sample size code letter for which acceptance and rejection numbers are available.

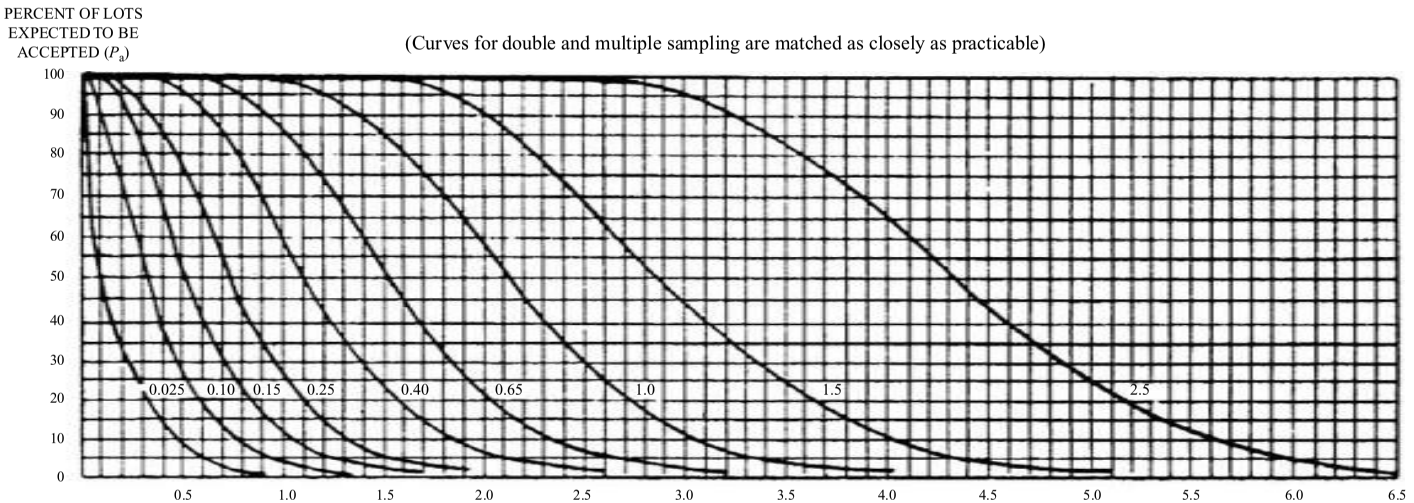
Ac = Acceptance number.

Re = Rejection number.

* = Use single sampling plan above (or alternatively use letter *N*).

= Acceptance not permitted at this sample size.

Table 12.6 Sample Size Code Letter: *N* (Chart *N* Shows Operating Characteristic Curves for Single Sampling Plans)[a]



Tabulated values for operation characteristic curves for single sampling plans

	Acceptable quality levels (normal inspection)											
P_a	0.025	0.10	0.15	0.25	0.40	0.65	*X*	1.0	*X*	1.5	*X*	2.5
99.0	0.0020	0.030	0.087	0.165	0.357	0.581	0.701	0.954	1.22	1.50	2.07	2.51
95.0	0.0103	0.071	0.164	0.273	0.523	0.796	0.939	1.23	1.54	1.85	2.49	2.98
90.0	0.0120	0.106	0.220	0.349	0.630	0.931	1.09	1.40	1.73	2.06	2.73	3.25
75.0	0.0576	0.192	0.345	0.507	0.844	1.19	1.37	1.72	2.08	2.45	3.18	3.74
50.0	0.139	0.336	0.535	0.734	1.13	1.53	1.73	2.13	2.53	2.93	3.73	4.33
25.0	0.277	0.539	0.784	1.02	1.48	1.94	2.16	2.60	3.04	3.48	4.35	4.99
10.0	0.461	0.778	1.06	1.34	1.86	2.35	2.60	3.08	3.56	4.03	4.95	5.64
5.0	0.599	0.949	1.26	1.55	2.10	2.63	2.89	3.39	3.89	4.38	5.34	6.05
1.0	0.921	1.328	1.68	2.01	2.62	3.20	3.48	4.03	4.56	5.09	6.12	6.87
	0.040	0.15	0.25	0.40	0.65	–	1.0	–	1.5	–	2.5	–
	Acceptable quality levels (tightened inspection)											

Sampling plans for sample size code letter *N*

Acceptable quality levels (normal inspection)

Type of sampling plan	Cumulative sample size	Less than 0.025		0.025		0.040		–		0.065		0.10		0.15		0.25		0.40		0.65		–		1.0		–		1.5		–		2.5		Higher than 2.5		Cumulative sample size
		Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	Ac	Re	
Single	500	▽		0	1							1	2	2	3	3	4	5	6	7	8	8	9	10	11	12	13	14	15	18	19	21	22	△		500
Double	315	▽		*								0	2	0	3	1	4	2	5	3	7	3	7	5	9	6	10	7	11	9	14	11	16	△		315
	630											1	2	3	4	4	5	6	7	8	9	11	12	13	15	16	18	19	23	24	26	27	27			630
Multiple	125	▽		*		Use letter *M*		Use letter *Q*		Use letter *P*		#	2	#	2	#	3	#	4	0	4	0	4	0	5	0	6	1	7	1	8	2	9	△		125
	250											#	2	0	3	0	3	1	5	1	6	2	7	3	8	3	9	4	10	6	12	7	14			250
	375											0	2	0	3	1	4	2	6	3	8	4	9	6	10	7	12	8	13	11	17	13	19			375
	500											0	3	1	4	2	5	3	7	5	10	6	11	8	13	10	15	12	17	16	22	19	25			500
	625											1	3	2	4	3	6	5	8	7	11	9	12	11	15	14	17	17	20	22	25	25	29			625
	750											1	3	3	5	4	6	7	9	10	12	12	14	14	17	18	20	21	23	27	29	31	33			750
	875											2	3	4	5	6	7	9	10	13	14	14	15	18	19	21	22	25	26	32	33	37	38			875
		Less than 0.040		0.040		–		0.065		0.10		0.15		0.25		0.40		0.65		–		1.0		–		1.5		–		2.5		–		Higher than 2.5		

Acceptable quality levels (tightened inspection)

[a] Quality of submitted lots (p, in percent defective for AQLs ≤ 10; in defects per hundred units for AQLs > 10).

Note: Figures on curves are acceptable quality levels (AQLs) for normal inspection. Curves for double and multiple sampling are matched as closely as practicable.

△ = Use next preceding sample size code letter for which acceptance and rejection numbers are available.

▽ = Use next subsequent sample size code letter for which acceptance and rejection numbers are available.

Ac = Acceptance number.

Re = Rejection number.

* = Use single sampling plan above (or alternatively use letter *R*).

\# = Acceptance not permitted at this sample size.

Table 12.7 Design to Analyze Components of Variance for the Tablet Assay

	Tablets (treatment groups)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Assay	48	49	49	55	48	54	45	47	53	50
Results	51	50	52	55	47	52	49	49	50	51
Mean	49.5	49.5	50.5	55	47.5	53	47	48	51.5	50.5
	Grand average = 50.2									

In certain situations (particularly when treatments are a random effect), one may be less interested in a statistical test of treatment differences, but more interested in separately estimating the variability due to different treatment groups *and* the variability within treatment groups. We will consider an example of a quality control procedure for the assay of finished tablets. Here, we wish to characterize the assay procedure by estimating the sources of variation that make up the variability of the analytical results performed on different, distinct tablets. This variability is composed of two parts: (a) that due to analytical error, and (b) that due to tablet heterogeneity. A oneway ANOVA design such as that shown in Table 12.7 will yield data to answer this objective. In the example shown in the table, 10 tablets are each analyzed in duplicate. Duplicate determinations were obtained by grinding each tablet separately, and then weighing two portions of the ground mixture for assay. The manner in which replicates (duplicates, in this example) are obtained is important, not only in the present situation, but also in most examples of statistical designs. Here we can readily appreciate that analytical error, the variability due to the analytical procedure only, is represented by differences in the analytical results of the two "identical" portions of a homogeneously ground tablet. This variability is represented by the "within" error in the ANOVA table shown in Table 12.8. The "within"-mean square is the pooled variance *within* treatment groups, where a group, in this example, is a single tablet.

The *between-tablet* mean square is an estimate of both *assay* (analytical *error*) and the *variability of drug content in different tablets* (tablet heterogeneity) as noted above. If tablets were identical, individual tablet assays would not be the same because of analytical error. In reality, in addition to analytical error, the drug assay is variable due to the inherent heterogeneity of such dosage forms. Variability between tablet assays is larger than that which can be accounted for by analytical error alone. This is the basis for the F test in the ANOVA [(between-mean square)/(within-mean square)]. Large differences in the drug content of different tablets result in a large value of the between-tablet mean square. This concept is illustrated in Figure 12.10, which shows an example of the distribution of *actual* drug content in a theoretical batch of tablets. The distribution of tablet assays is more spread out than the drug content distribution, because the variation based on the assay results of the different tablets include components due to *actual drug content variation plus assay error*.

Based on the theoretical model for the one-way ANOVA, section 8.1 (random model), it can be shown that the between-mean square is a combination of the assay error and tablet variability as follows:

$$\text{BMS} = n\sigma_{\text{T}}^2 + \sigma_{\text{w}}^2, \tag{12.3}$$

where n is the number of replicates in the design (based on equal replication in each group, two assays per tablet in our example), σ_{T}^2 the variance due to tablet drug content heterogeneity,

Table 12.8 Analysis of Variance for the Tablet Assay Data from Table 12.7

Source	**d.f.**	**SS**	**MS**
Between tablets	9	112.2	12.47
Within tablets	10	27.0	2.70
Total	19	139.2	

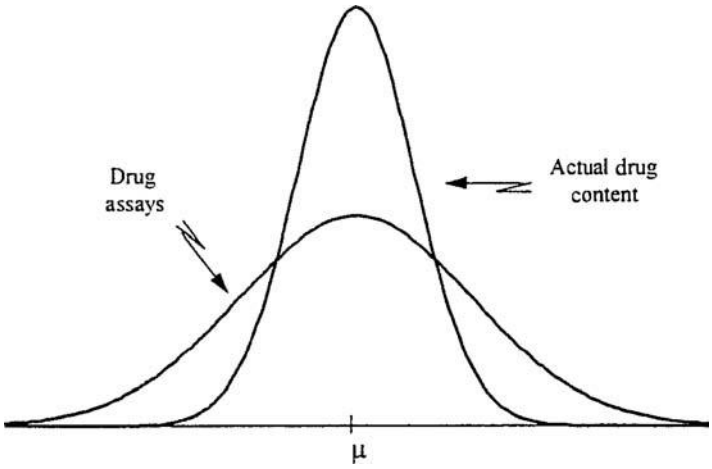


Figure 12.10 Distribution of actual drug content compared to distribution of analytical results of tablets (these are theoretical, hypothetical distributions).

and σ_W^2 is the within-treatment (assay) variance. In our example, $n = 2$, and the between-mean square is an estimate of $2\sigma_T^2 + \sigma_W^2$. The *within*-tablet mean square is an estimate of σ_W^2, equal to 2.70 (Table 12.8). The estimate of σ_T^2 from Eq. (12.3) is $(\text{BMS} - \sigma_W^2)/n$

$$\text{Estimate of } \sigma_T^2 = \frac{\text{between MS} - 2.70}{2} = \frac{12.47 - 2.70}{2} = 4.9.$$

In this manner we have estimated the *two components* of the between-treatment mean square term

$$\sigma_W^2 = 2.7 \text{ and } \sigma_T^2 = 4.9.$$

The purpose of the experiment above, in addition to estimating the components of variance, would often include an estimation of the overall average of drug content based on the 20 assays (Table 12.7). The average assay result is 50.2 mg. The estimates of the variance components can be used to estimate the variance of an average assay result, consisting of m tablets with n assay replicates per tablet. We use the fact that the variance of an average is equal to the variance divided by N, where N is equal to mn, the total number of observations. According to Eq. (12.3), the variance of the average result can be shown to be equal to

$$\frac{n\sigma_T^2 + \sigma_W^2}{mn}. \tag{12.4}$$

The variance estimate of the average assay result (50.2) for the data in Table 12.7, where $m = 10$ and $n = 2$, is

$$\frac{2(4.9) + 2.7}{10(2)} = 0.62.$$

Note that this result is exactly equal to the between-mean square divided by 20.

According to Eq. (12.4), the variance of *single* assays performed on *two* separate tablets, for example, is equal to $(m = 2, n = 1)$

$$\frac{4.9 + 2.7}{2} = 3.8.$$

Note that the variance of a *single assay* of a *single tablet* is $\sigma_T^2 + \sigma_W^2$. Similarly, the variance of the average of two assays performed on a single tablet ($m = 1$, $n = 2$) is $(2\sigma_T^2 + \sigma_W^2)/2$ (see Exercise Problem 11). The former method, where *two tablets* were each *assayed once*, has greater precision than duplicate assays on a single tablet. Given the same number of assays, the procedure that uses more tablets will always have better precision. The "best" combination of the number of tablets and replicate assays will depend on the particular circumstances, and includes time and cost factors. In some situations, it may be expensive or difficult to obtain the experimental material (e.g., obtaining patients in a clinical trial). Sometimes, the actual observation may be easily obtained, but the procedure to prepare the material for observation may be costly or time consuming. In the case of tablet assays, it is conceivable that the grinding of the tablets, dissolving, filtration, and other preliminary treatment of the sample for assay might be more expensive than the assay itself (perhaps automated). In such a case, replicate assays on ground material may be less costly than assaying separate tablets, where each tablet must be crushed and ground, dissolved, and filtered prior to assay. However, such situations are exceptions. Usually, in terms of precision, it is cost effective to average results obtained from different tablets.

The final choice of how many tablets to use and the total number of assays will probably be a compromise depending on the precision desired and cost constraints. The same precision can be obtained by assaying different combinations of numbers of tablets (m) with different numbers of replicate determinations (n) on each tablet. Time-cost considerations can help make the choice. Suppose that we have decided that a sufficient number of assays should be performed so that the variance of the average result is equal to approximately 1.5. In our example, where the variance estimates are $S_T^2 = 4.9$ and $S_W^2 = 2.7$, the average of five single-tablet assays would satisfy this requirement

$$S_T^2 = \frac{4.9 + 2.7}{5} = 1.52.$$

As noted above, the variance of a single-tablet assay is $S_T^2 + S_W^2$. An alternative scheme resulting in a similar variance of the mean result is to assay four tablets, each in duplicate

$$(m = 4, n = 2).$$

$$S_{\bar{X}}^2 = \frac{2(4.9) + 2.7}{8} = 1.56.$$

The latter alternative requires eight assays compared to five assays in the former scheme. However, the latter method uses only four tablets compared to the five tablets in the former procedure. The cost of a tablet would probably not be a major factor with regard to the choice of the alternative procedures. In some cases, the cost of the item being analyzed could be of major importance. In general, for tablet assays, in the presence of a *large assay variation*, if the analytical procedure is automated and the preparation of the tablet for assay is complex and costly, the procedure that uses less tablets with more replicate assays per tablet could be the best choice.

12.4.1.1 Nested Designs

Designs for the estimation of variance components often fall into a class called *nested* or completely hierarchical designs. The example presented above can be extended if we were also interested in ascertaining the variance due to differences in average drug content between *different batches* of tablets. We are now concerned with estimating (a) between-batch variability, (b) between-tablet (within batches) variability, and (c) assay variability. Between-batch variability exists because, despite the fact that the target potency is the same for all batches, the actual mean potency varies due to changing conditions during the manufacture of different batches. This concept has been discussed under the topic of control charts.

A design used to estimate the variance components, including batch variation, is shown in Table 12.9 and Figure 12.11. In this example, four batches are included in the experiment, with

Table 12.9 Nested Design for Determination of Variance Components

Batch	*A*			*B*			*C*			*D*		
Tablet	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
	50.6	49.1	51.1	50.1	51.0	50.2	51.4	52.1	51.1	49.0	47.2	48.9
	50.5	48.9	51.1	49.0	50.9	50.0	51.7	52.0	51.9	49.0	47.6	48.5
	50.8	48.5	51.4	49.4	51.6	49.8	51.8	51.4	51.6	48.5	47.6	49.2

ANOVA

Source	d.f.	SS	MS	Expected MS[a]
Between batches	3	48.6875	16.229	$\sigma_W^2 + 3\sigma_T^2 + 9\sigma_B^2$
Between tablets (within batches)	8	17.52	2.190	$\sigma_W^2 + 3\sigma_T^2$
Between assays (within tablets)	24	2.50	0.104	σ_W^2

[a]Coefficient for σ_T^2 = replicate assays; coefficient for σ_B^2 = replicate assays times the number of tablets per batch.

three tablets selected from each batch (tablets nested in batches), and three replicate assays of each tablet (replicate assays nested in tablets). This design allows the estimate of variability due to batch differences, tablet differences, and analytical error. The calculations for the ANOVA will not be detailed (see Ref. [9]) but the arithmetic is straightforward and is analogous to the analysis in the previous example.

The *mean squares* (MS) calculated from the *ANOVA estimate* the *true variances* indicated in the column "expected MS." The coefficients of the variances from the expected mean squares and the estimates of the three "sources" of variation can be used to estimate the components of variance. The variance components, σ_B^2, σ_T^2, and σ_W^2 may be estimated as follows from the mean square and expected mean square columns in Table 12.9.

$$S_W^2 = 0.104$$
$$S_W^2 + 3S_T^2 = 2.190 \quad S_T^2 = 0.695$$
$$S_W^2 + 3S_T^2 + 9S_B^2 = 16.229 \quad S_B^2 = 1.56$$

An estimate of the variance of single-tablet assays randomly performed within a single batch is $S_W^2 + S_T^2 = 0.799$. If tablets are randomly selected from different batches, the variance estimate of single-tablet assays is $S_W^2 + S_T^2 + S_B^2 = 2.36$.

Nested designs should be symmetrical to be easily analyzed and interpreted. The symmetry is reflected by the equal number of tablets from each batch, and the equal number of replicates per tablet. Missing or lost data result in difficulties in estimating the variance components [10].

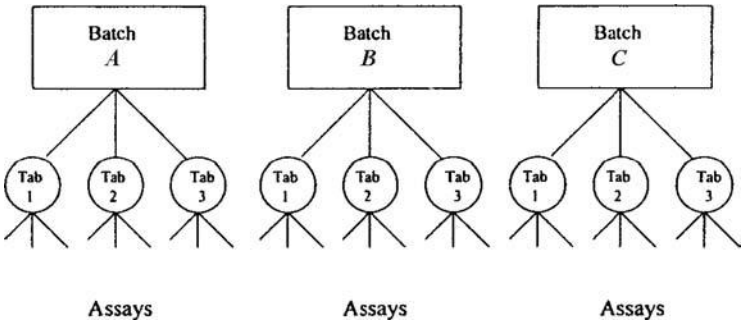


Figure 12.11 A nested or completely hierarchical design to estimate variance components (three of four batches are shown).

12.5 ESTABLISHING IN-HOUSE LIMITS

An important consideration in establishing standards is to evaluate limits for release of products. The two important kinds of release limits are "official" limits, such as stated in the USP or in regulatory submissions, and "in-house" limits that are narrower than the "official" limits. The purpose of in-house limits is to obtain a greater degree of assurance that the true attributes of the product are within official limits when the product is released. Thus, in-house limits decrease the consumer risk. If a product shows measurable decomposition during its shelf life, the in-house release specifications must be more narrow than the official limits to compensate for the product instability.

In the absence of instability, in-house limits should be sufficiently within the official limits to ensure the integrity of the product considering the variability of the measurement (assay). For the case of a *homogeneous sample* (e.g., solutions or a composite sample of a solid dosage form), the variability of the assay may be accounted for by analytical error. An important consideration is to use a proper estimate of the analytical variability. A distinction should be made between within-day variability and between-day variability. For this application, the variability of the analytical method should be estimated as between-day variability. The reason for this is that the variability of an assay on any given day will be dependent on assay conditions on that day, and is apt to be larger than the within-day variability (differences among replicate assays on the same day). For solid dosage forms, the variability of the final assay is a combination of analytical error and tablet heterogeneity (that is, in the absence of analytical error, two separate samples will differ in drug content due to the fact that perfect mixing is not possible in a powder mix). In this case, the estimate of assay variability should not ignore these components of variance. (See discussion of components of variance.)

The examples below show the calculation for a lower limit for in-house release specifications, but the same reasoning will apply for an upper in-house release specification.

$$\begin{aligned} \text{LRL} &= \text{Lower official limit} + t \times S \\ \text{LRL} &= \text{Lower release specification} \end{aligned} \tag{12.5}$$

For a 95% one-sided confidence interval, t is determined from a t table with d.f. based on the estimate of the assay standard deviation, S. The standard deviation is obtained from between-day replicates during assay development or from a standard product assayed on different days. For tablets, the proper standard deviation should include tablet heterogeneity, that is, replicate assays on different composites. A standard deviation estimated from replicates done on the same day (sometimes estimated from control charts) is not the correct standard deviation.

If, according to SOPs, the assay for release is done in duplicate, one might be tempted to divide the last term in Eq. (12.5) by $\sqrt{2}$. This is not strictly correct because the duplicates refer to within-day variability. If the duplicates were done on two separate days (an unlikely procedure) and on separate composites, then the division by $\sqrt{2}$ would be more correct. If replicates are used for the final assay, one could estimate the correct error if an estimate of the within- and between-day components of variance (based on assay of different composites) is available.

$$S^2_{\text{total}} = \frac{S^2_{\text{between}} + S^2_{\text{within}}}{n},$$

where n = number of replicates (separate sets of composites). In this case, the number of d.f. can be estimated using Saterthwaite's (see below) approximation. An alternative way of estimating the s.d., if product heterogeneity is not a factor, is to perform replicate determinations on a standard product over time and compute the s.d. of the average results. Some examples should clarify the procedure.

Example 1. Single assays on a portion of a cough syrup are performed as one of the tests for the release of the product. The assay has a s.d. of 2.1 based on the results of the assay performed on a single stable batch on 15 different occasions (days). From Table IV.4, the value of t with 14 d.f.

for a one-sided 95% confidence interval is 1.76. If the official limits are 90% to 110%, in-house limits of

$$90\% + 1.76 \times 2.1 = 93.7$$
$$110\% - 1.76 \times 2.1 = 106.3$$

mean that if the assay falls within 93.7% and 106.3%, the probability that the true batch mean is out of official specifications (90%–110%) is less than 5%.

Example 2. Single assays on a composite of 20 tablets are performed as one of the tests for the release of a product. During development of the product and the assay, an experimental batch of tablets was assayed on 20 different days (a different composite each day). This assay was identical to the composite assay, a 20 tablet composite. The drug in the dosage form is very stable. The s.d. (19 d.f.) is 2.1. From Table IV.4, the value of t with 19 d.f. for a one-sided 95% confidence interval is 1.73. If the official limits are 90% to 110%, the in-house limits are

$$90\% + 1.73 \times 2.1 \qquad \text{and} \qquad 110\% - 1.73 \times 2.1$$
93.63% to 106.37%.

Example 3. Consider the situation in Example 2 where the assay is performed in duplicate and the average result is reported as a basis for releasing the batch. The duplicate determination is performed on two portions of the same 20 tablet composite on the same day. The variability of the result is a combination of tablet content heterogeneity, and within- and between-day assay variability. Since the same composite is assayed twice, the variance is

$$\frac{[S^2_{\text{tablet heterogeneity}}]}{20} + S^2_{\text{(assay) between}} + \frac{[S^2_{\text{(assay) within}}]}{2}. \tag{12.6}$$

If one considers the first term to be small relative to the last two terms, the s.d. can be computed with estimates of the within- and between-day variance components. These estimates could be obtained from historical data, including data garnered during the assay development. The important point to remember is that the computation is not straightforward because of the need to estimate variance components and the d.f. based on these estimates. Assuming that the between-day variance component of the assay is 0, we could calculate the limits as follows.

Assume that the first two terms in Eq. (12.6) are small and that the assay variability has been estimated based on 15 assays with s.d. = 2.1. The average of duplicate assays on the same composite would have in-house limits of

$$90\% + 1.76 \times 2.1/\sqrt{2} \text{ and } 110\% - 1.76 \times 2.1\sqrt{2}$$
92.6 to 107.4%.

If the tablet variability, $[S^2_{\text{tablet heterogeneity}}]/20$, is large compared to assay variability (probably a rare occurrence), performing duplicate assays on the same composite will not yield much useful information. In this case, to get more precision, one can assay separate 20 tablet composites (see Exercise Problem 13 at the end of this chapter).

Allen et al. [4] discuss the setting of in-house limits when a product is susceptible to degradation. This situation is complicated by the fact that the in-house limits must now take into consideration an estimate of the rate of degradation with its variability, as well as the variability due to the assay. Obviously, the in-house release limits should be within the official limits. In particular, for the typical case where the slope of the degradation plot is negative, we are concerned with the lower limit. If the official lower limit is 90%, the in-house release limit should be greater than 90% by an amount equal to the estimated amount of drug degraded during the shelf life plus another increment due to assay variability. The following notation

is somewhat different from Allen et al., but the equations are otherwise identical. The lower release limit (LRL) can be calculated as shown in Eq. (12.7).

$$\text{LRL} = \text{OL} - \text{DEGRAD} + t \times \left(\frac{S_\text{d}^2 + S_\text{a}^2}{n} \right)^{1/2} \tag{12.7}$$

where OL is the official lower limit; DEGRAD the predicted amount of degradation during shelf life = average slope of stability regression lines × shelf life; S_d^2 the variance of total degradation = shelf life^2 × S_slope^2.

Note: Variance of slope = $S_{y.x}^2 / \sum (X - \overline{X})^2$
Var (k × variable) = $k^2 \times S^2$ (variable) where k is a constant
S_a^2 = variance of assay
Note: S_a^2 is added because the assay performed at release is variable.

Another problem in computing the LRL is computation of d.f. for the one-sided 95% t distribution. The problem results from the fact that d.f. are associated with two variance estimates. When combining independent variance estimates, Satterthwaite approximation can be used to estimate the d.f. associated with the combined variance estimate [Eq. (12.8)].

For the linear combination, L, where

$$L = a_1 S_1^2 + a_2 S_2^2 + \ldots$$

the d.f. for L are approximately

$$\text{d.f.} = \frac{(a_1 S_1^2 + a_2 S_2^2 + \ldots)^2}{(a_1 S_1^2)^2 / v_1 + (a_2 S_2^2)^2 / v_2 + \ldots} \tag{12.8}$$

where v_i is d.f. for variance i.

The following example (from Allen) illustrates the calculation for the release limits.

OL = 90%
Average slope = −0.20%/month
shelf life = 24 months
DEGRAD = $-0.20 \times 24 = -4.8\%$
$S_\text{a} = 1.1\%$
Standard error of the slope = 0.03%
$S_\text{d} = 0.03 \times 24 = 0.72\%$
d.f. = 58
$t = 1.67$
$n = 2$ (duplicate assays)

If more than one lot is used for the computation, the lots should not be pooled without a preliminary test. Otherwise, an average slope may be used. In the case of multiple lots, the computations are not as straightforward as illustrated, and statistical assistance may be necessary.

Note the precautions on the variance of duplicate assays as discussed above.

$$\begin{aligned} \text{LRL} &= 90 + 4.8 + 1.67 \times \left(\frac{0.72^2 + 1.1^2}{2} \right)^{1/2} \\ &= 96.6\%. \end{aligned}$$

The lower release specification is set at 96.6%.

12.6 SOME STATISTICAL ASPECTS OF QUALITY AND THE "BARR DECISION"

The science of quality control is largely based on statistical principles, in part because we take small samples and make inferences about the large population (e.g., a batch). Following is a discussion of a few topics that illustrate some statistical ways of looking at data.

What is a good sample size? The FDA often seeks information on the rationale for sample sizes in SOPs. Are we taking enough samples? How many samples should we use for analysis? Actually, this is not an easy question to answer in many cases and that is why the question is asked so often. To answer this question from a statistical point of view, one has to answer a few questions, not all of them easy (chap. 6). For example, we need an estimate of the s.d. and definitions of alpha and beta levels for a given meaningful difference, if the data suggest some comparison.

Often the sample size is fixed based on other considerations such as official specifications. Cost is a major consideration. As an example, consider the composite assay for tablets as one of the QC release criteria. Twenty tablets are assayed to represent a million or more tablets in many cases.

Is this sample large enough? The sample size needed to make such an estimate depends on the precision (s.d.) of the data and the desired precision of the estimate in which we are interested, the mean of the 20 tablets in this case. For the composite assay test, we are required to assay at least 20 tablets. Suppose that tablet variability (RSD) as determined from CU tests is about 3% and the analytical error (RSD) is 1%. Based on this information, we can estimate the variability of the composite assay. The content uniformity variation is due to tablet heterogeneity, which includes weight variation and potency variation, in addition to analytical error.

$$S^2_{\text{content uniformity}} = S^2_{\text{weight}} + S^2_{\text{potency}} + S^2_{\text{analytical}}.$$

The tablet heterogeneity variance is the content uniformity variance minus the analytical variance.

$$S^2_{\text{potency}} + S^2_{\text{weight}} = S^2_{\text{content uniformity}} - S^2_{\text{analytical}} = (3)^2 - (1)^2 = 8.$$

We could even estimate the potency variation separately from weight variation if an estimate of weight variation is available (from QC tests for example).

The variability of the average of 20 tablets (without analytical error) is

$$S^2_{\text{composite}} = \frac{8}{20} = 0.4.$$

If we assay a mixture of 20 tablets, the variance including analytical error is

$$S^2 = 0.4 + 1 = 1.4$$
$$S = 1.18.$$

Do you think that the average of a randomly selected sample of 20 tablets gives an accurate representation of the batch? We might answer this question by looking at a confidence interval for the average content based on these data. Assume that the analytical error is well established and, for this calculation, 9 d.f. (based on CU data) are reasonable for the t value needed for the calculation of the confidence interval. If the observed composite assay is 99.3%, a 95% confidence interval for the true average is

$$99.3\% \pm 2.262 \times 1.18 = 96.6 \text{ to } 102.0.$$

If this is not satisfactory (too wide), we could reduce the interval width by performing replicate assays of the composite or, perhaps, by using more tablets in the composite. For

example, duplicate assays from a single composite may be calculated as follows:

$$S^2 = 0.4 + \frac{1}{2} = 0.9$$
$$S = 0.95.$$

Note that the assay variance is reduced by half, but the variance due to tablet heterogeneity is not changed because we are using the same composite. The confidence interval for the duplicates is

$$99.3 \pm 2.262 \times 0.95 = 97.2\% \text{ to } 101.4\%.$$

Using more than 20 tablets would decrease the CI slightly. If we used 40 tablets with a single assay, the variance would be

$$S^2 = \frac{8}{40} + 1 = 1.2$$

and the CI would be 96.8 to 101.8.

When combining independent variance estimates, Satterthwaite approximation can be used to estimate the d.f. associated with the combined variance estimate. The formula [Eq. (12.8)] is presented in section 12.5

$$\text{d.f.} = \frac{(a_1 S_1^2 + a_2 S_2^2 + \ldots)^2}{(a_1 S_1^2)^2/v_1 + (a_2 S_2^2)^2/v_2 + \ldots}, \tag{12.8}$$

where v_i is d.f. for variance i.
For example, suppose the estimates of variance have the d.f. as follows:

$$S^2_{\text{analytical}} = 2 \text{ with } 15 \text{ d.f.}$$
$$S^2_{\text{weight}} = 9 \text{ with } 9 \text{ d.f.}$$
$$S^2_{\text{potency}} = 1 \text{ with } 6 \text{ d.f.}$$

The d.f. for an estimate of content uniformity are based on the following linear combination:

$$1 \times S^2_{\text{analytical}} + 1 \times S^2_{\text{weight}} + 1 \times S^2_{\text{potency}}.$$

From Eq. (12.8),

$$\text{d.f.} = \frac{(9 + 2 + 1)^2}{(4/15 + 81/9 + 1/6)} = 15.3.$$

Estimating the d.f. using this approximation is less good for the differences of variances as compared to the sum of variances.

Example. Limits based on analytical variation are to be set for release of a product. The lower limit is 90%. In-house limits are to be sufficiently above 90% so the probability of an assay being below 90% is less than 0.05. Calculate the release limits where a single assay is done on a composite of 20 tablets. The assay RSD is 3% based on 25 d.f. Tablet heterogeneity (RSD) is estimated as 1% based on 9 d.f.

The estimated variance of the composite assay is ($a_1 = 1/20$, $s_1^2 = 1$, $a_2 = 1$, $S_2^2 = 3$)

$$\frac{1}{20} + 3^2 = 9.05\%$$
$$\text{d.f.} \approx (3^2 + 1^2/20)^2/[3^2 \times (1/25) + (1/20)^2 \times (1/9)] = 25.3$$

Assuming 26 d.f., $t = 1.71$
The lower limit is $90 + 1.71 \times \sqrt{9.05} = 95.1$.
Therefore, the lower in-house limit is 95.1%.

Blend Samples. What are some properties of three dose weight samples for blend testing? This has been interpreted in different ways, such as (a) take three sample weights and assay the whole sample. (b) Take three sample weights and assay a single dose weight without mixing the sample. (Tread lightly when transferring the sample to the laboratory.) (c) Take three sample weights, mix thoroughly and assay a single sample. Based on the Barr decision [11], the latter (c) appears to be preferable. Some firms have been requested to sample the blend (3 dose weights) and to impose limits of 90% to 110% for each sample. One might ask if this standard is too restrictive, too liberal, or just right? To help evaluate this procedure, consider the case of a firm that assays three samples, each of single dosage weights. How might the above criterion for acceptance compare to that for 10 dosage units in which all must be between 85% and 115%?

Some approximate calculations to see if the 90% to 110% limits are fair for the blend samples can shed some light on the nature of the specifications. Suppose that the assay is right on, at 100%. Suppose, also, that 99.9% of the tablets in the batch are between the 85% to 115% CU limits. The probability of 10 of 10 tablets passing if each has a probability of 0.999 of passing is (binomial theorem)

$$0.999^{10} = 0.99.$$

If the tablets are distributed normally, the s.d. is about 4.6. This is based on the fact that a normal distribution with a mean of 100 and a s.d. of 4.6 will have 99.9% of the values between 85 and 115. This same distribution will have 97% of the tablets between 90 and 110. The probability of 3 of three units being between 90 and 110 is

$$0.97^3 = 0.91,$$

which is less than the probability of passing for the final tablet content uniformity test.

The FDA has recommended that the limits for the blend samples be 90% to 110%. Since the probability of passing the final tablet CU test is 0.99 under these circumstances, the chances of failing the blend uniformity test may not seem fair, unless you believe that the blend should be considerably more uniform than the final tablets.

What limits would be fair to make this acceptance criterion (3/3 must pass) equivalent to the USP test given the above estimates. Let the probability of passing the blend test = 0.99 to make the test equivalent to that for the finished tablets.

$$P^3 = 0.99,$$

where p is the probability of a single blend sample passing.

$$p = 0.9967$$

That is, to make the probability of passing (3/3) the same as the final CU test, we would assume that 99.67% of the samples should be within limits. Assuming a normal distribution with a RSD of 4.6%, this corresponds to acceptance limits of about 87.5 to 112.5. This would seem fair. However, what are the consequences if the 3-dosage unit weight is composited? In

this case, we are assaying the average of 3 tablet weights. These assays should be less variable with a s.d. less than 4.6, the s.d. of single unit weights. Although the variability of the average of 3 dosage weights will be smaller than a single dosage weight, the exact s.d. cannot be defined because the nature of tablet heterogeneity cannot be defined. For the sake of this example, let us assume that the s.d. is 2.66 ($4.6/\sqrt{3}$). Would 90% to 110% be fair limits for each of three blend samples, each consisting of 3 dosage weights? We can compute the probability of a single sample (3 tablet weights) passing using normal distribution theory.

$$Z = 10/2.66 = 3.76$$

probability ($90 < \text{assay} < 110$) = 0.99983.

The probability of three samples passing is

$$0.99983^3 = 0.999.$$

Although this test would be easier to pass than the final tablet content uniformity test, it is based on an assumption of the value of the s.d. for the three unit weight samples, an unknown!

12.7 IMPORTANT QC TESTS FOR FINISHED SOLID DOSAGE FORMS (TABLETS AND CAPSULES)

Important finished solid dosage form tests include

1. content uniformity;
2. assay;
3. dissolution.

In this section, a description of these tests is presented. Included also is the f_2 test for comparing dissolution profiles of two different products with the same active ingredient (as is often done when comparing the dissolution of generic and brand products).

12.7.1 Content Uniformity

The content uniformity is a test to assess and control the variability of solid dosage forms. Although the sampling of the batch for these tests is not specified, good statistical practice recommends some kind of random or representative sampling [12]. This test consists of two stages. For tablets, 30 units are set aside to be tested. In the first stage, individually assay 10 tablets. If all tablets assay between 85% and 115% of label claim and the RSD is less than or equal to 6, the test passes. If the test does not pass, and no tablet is outside 75% or 125%, assay the remaining individual 20 tablets (Stage 2). The test passes if, of the 30 tablets, not more than one tablet is outside 85% to 115% of label claim, no tablet is outside 75% to 125%, and the RSD is less than or equal to 7.8%.

For capsules, the first stage is the same as for tablets, except that one capsule may lie outside of 85% to 115%, but none outside 75% to 125%. The second stage assays 20 more capsules and of the total of 30 capsules, no more than three capsules can be outside 85% to 115%, none outside 75% to 125% and the *RSD not more than 7.8%*.

12.7.2 Assay

The potency of the final product is based on the average of (at least) 20 dosage units. Twenty random or representative units are ground into a "homogeneous" mix using a suitable method. A sample(s) of this mix is assayed. This assay must be within the limits specified in the USP or a specified regulatory document. Typically, but not always, the assay must be within 90% to 110% of label claim.

12.7.3 Dissolution

The FDA guidance for "Dissolution Testing of Immediate Release Oral Dosage Forms" succinctly describes methods for testing and evaluating dissolution data [13]. Dissolution testing evaluates the dissolution behavior of the drug from a dosage form as a function of time. Thus,

Table 12.10 USP Dissolution Test Acceptance Criteria

Stage	Number tested	Criteria
Stage 1 (S1)	6	Each unit not less than $Q + 5\%$
Stage 2 (S2)	6	Average of 12 units (S1 + S2) equal to or greater than and no unit less than $Q - 15\%$
Stage 3 (S3)	12	Average of 24 units (S1 + S2 + S3) equal to or greater than Q; and not more than 2 units are less than $Q - 15\%$, and no unit is less than $Q - 25\%$

Q is the dissolution specification in percent dissolved.

the typical dissolution-vs.-time curve shows the cumulative dissolution of drug over time. Provided a sufficient quantity of solvent is available, 100% of the drug should be dissolved, given enough time. The procedure for dissolution testing is described in the USP. Briefly, the procedure requires that individual units of the product (for solid dosage forms) be placed in a dissolution apparatus that typically accommodates six separate units. The volume and nature of the dissolution medium is specified (e.g., 900 mL of 0.1 N HCl), and the containers, rotating basket or paddle (USP), are then agitated at a prescribed rate in a water bath at 37°C. Portions of the solution are removed at specified times and analyzed for dissolved drug. Usually, dissolution specifications for immediate-release drugs are determined as a single point in time. Table 12.10 shows the USP Dissolution Test Acceptance Criteria [14], which may be superseded by specifications in individual drug monographs. For controlled-release products and during development, dissolution at multiple time points, resulting in a dissolution profile (Fig. 12.12) is necessary.

The principal purposes of dissolution testing are threefold: (1) for quality control, dissolution testing is one of several tests to ensure the uniformity of product from batch to batch. (2) Dissolution is used to help predict bioavailability for formulation development. For the latter purpose, it is well known that dissolution characteristics may predict the rate and extent of absorption of drugs in some cases, particularly if dissolution is the rate-determining step for drug absorption. Thus, although not always reliable, dissolution is probably the best predictor of bioavailability presently available. (3) Finally, dissolution may be used as a measure of change when formulation or manufacturing changes are made to an existing formulation.

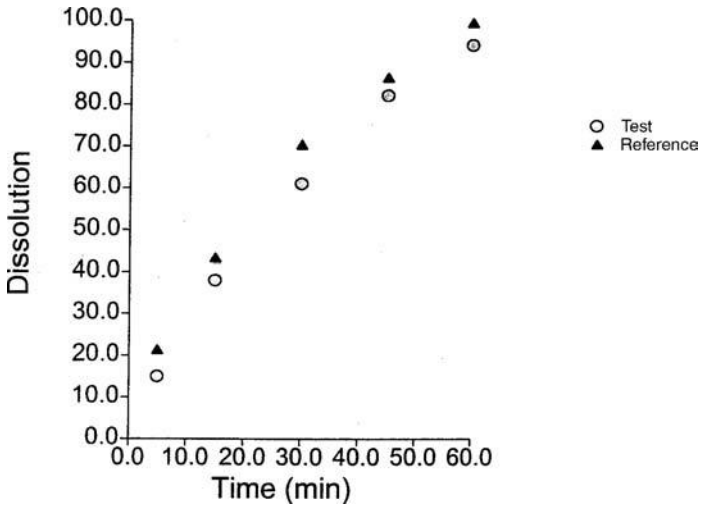


Figure 12.12 Dissolution profile comparing test to reference products.

Table 12.11 Comparison of Test and Reference Dissolution Profiles

	% Dissolution		
Time (min)	**Test**	**Reference**	**Difference ($R_i - T_i$)**
5	15	21	6
15	38	43	5
30	61	70	9
45	82	86	4
60	94	99	5

The so-called f_2 method can be used to compare two dissolution profiles. The formula for the computation of f_2 is as follows:

$$f_2 = 50 \log \left\{ \left[1 + \left(\frac{1}{N} \right) \sum (\mathrm{R}_i - \mathrm{T}_i)^2 \right]^{-0.5} \times 100 \right\},$$

where N is the number of time points; R_i and T_i are the dissolution of reference and test products at time i.

Consider the following example (Table 12.11 and Fig. 12.12).

$$\begin{aligned} f_2 &= 50 \log \left\{ \left[1 + \frac{1}{N} \sum (\mathrm{R}_i - \mathrm{T}_i)^2 \right]^{-0.5} \times 100 \right\} \\ &= 50 \log \left\{ \left[1 + \frac{1}{5} \times (36 + 25 + 81 + 16 + 25) \right]^{-0.5} \times 100 \right\} \\ &= 50 \log \left\{ \left[1 + \frac{1}{5} \times 183 \right]^{-0.5} \times 100 \right\} = 60.6 \end{aligned}$$

f_2 must be greater than 50 to show similarity.

f_2 should not be absolute. There are situations where the use of this test does not give results that give reasonable conclusions. For example, with rapidly dissolving drugs, large differences at early time points could result in an f_2 value less than 50 when the dissolution profiles seem to be similar. Also, the method should be used and interpreted with care when few data points are available.

Consider another example (Table 12.12 and Fig. 12.13).

$$f_2 = 50 \log \left\{ \left[1 + \left(\frac{1}{4} \right) (576 + 36 + 9 + 1) \right]^{-0.5} \times 100 \right\} = 45.$$

These products differ only at the very early, and probably variable, time point. Yet, they are not considered similar using this test. As noted, an interpretation of these kinds of data should be made with caution.

Table 12.12 A Second Comparison of Test and Reference Dissolution Profiles

	% Dissolution	
Time (min)	**Test**	**Reference**
5	51	75
10	89	95
15	93	96
30	97	98

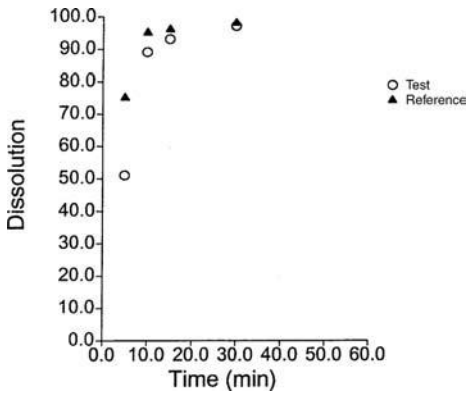


Figure 12.13 Dissolution profile comparing test to reference products for fast dissolving products.

12.8 OUT OF SPECIFICATION (OOS) RESULTS

A discussion of OOS results (failing assay) is presented in Appendices V and VI. These articles were prompted by the Barr decision and FDA's interpretation of Judge Wolin's decision [11]. Since these articles were published, the FDA has published a guidance for "Investigating Out of Specification (OOS) Test Results for Pharmaceutical Production," which addresses these problems and more clearly defines procedures to be followed if an OOS result is observed [15].

The following is a synopsis of the document and comments on topics relevant to this book. All OOS results should be investigated, whether or not the batch is rejected. It is important to find causes that would help maintain the integrity of the product in future batches. The laboratory data should first be inspected for accuracy before any test solutions are discarded. If no errors are apparent, a "complete failure investigation should follow." If an obvious error occurs, the analysis should be aborted, and immediately documented. The thrust of the investigation is to distinguish between a laboratory error and problems with the manufacture of the product. Of course, the optimal procedure would be to have the opportunity to retest the suspect sample if it is still available. In any event, if a laboratory error is verified, the OOS result will be invalidated.

In the laboratory phase of the investigation, various testing procedures are defined. Retesting is a first option if there is not an obvious laboratory error. This is a test of the same sample that yielded the OOS result. For example, for a solution, an aliquot of that same solution may be tested. For a powdered composite, a new weighing from the same composite may be tested. The analysis should be performed by a person other than the one who obtained the OOS result. This retesting could confirm a mishandling of the sample or an instrumental error, for example. The SOPs should define how many assays are necessary to confirm a retesting result. The number of retests should be based on sound scientific and statistical procedures. (See Appendix for an example of a basis for retesting.) However, an OOS result that cannot be documented as a laboratory error, in itself, may not be sufficient to reject the batch. All analytical and other QC results should "be reported and considered in batch release decisions."

Resampling is sampling not from the original sample, but from another portion of the batch. This may be necessary when the original sample is not available or was not prepared properly, for example. These results may further indicate manufacturing problems, or may help verify the OOS result as an anomoly.

The document also discusses averaging (see also App. VII). Averaging is useful when measuring several values from a homogeneous mixture. If the individual results are meant to measure variability, it is clear that averaging without showing individual values is not tolerable. In any event when reporting averages, the individual values should be documented. All of these procedures should be clearly spelled out in the appropriate SOPs. It is of interest that the document discusses the case where three assays yield values of 89, 89, and 92, with a lower limit of 90. Clearly, this should raise some questions, although the FDA document states that this by itself does not necessarily mean that the batch will be failed.

Finally, the FDA does allow the use of outlier tests as long as the procedure is clearly documented in the SOPs. As a final comment, common sense and good scientific judgment are required to make sensible decisions in this controversial area.

KEY TERMS

Acceptance sampling
Action limits
AQL
Batch variation
Between- and within-batch variation
Chance variation
Components of variance
Nested designs
Operating characteristic (OC)
OOS (out of specification)
Power curve
Producer's (manufacturer's) risk
Proportion (p) charts
Range chart
Rational subgroups
Release limits
Resampling
Consumer's risk
Control chart
Control chart for differences
Control chart for individuals
Expected mean square
f_2
Moving average chart
Runs
Sampling for attributes
Sampling for variables
Sampling plan
Statistical control
Upper and lower limits
Warning limits
$\overline{X}$ charts
100% inspection

EXERCISES

1. Duplicate assays are performed on a finished product as part of the quality control procedure. The average of assays over many batches is 9.95 and the average range of the duplicates is 0.10 mg. Calculate upper and lower limits for the $\overline{X}$ chart and the range chart.
2. Past experience has shown the percentage of defective tablets to be 2%. What are the lower and upper 3σ limits for samples of size 1000?
3. A raw material assay shows an average percentage of 47.6% active with an average range of 1.2 based on triplicate assay. Construct a control chart for the mean and range.
4. What is the probability of rejecting a batch of product that truly has 1.0% rejects (defects) if the sampling plan calls for sampling 100 items and rejecting the batch if two or more defects are found?
5. The initial data for the assay of tablets in production runs are as follows (10 tablets per batch):

Batch	Mean	Range
1	10.0	0.3
2	9.8	0.4
3	10.2	0.4
4	10.0	0.2
5	10.1	0.5
6	9.8	0.4
7	9.9	0.2
8	9.9	0.5
9	10.3	0.3
10	10.2	0.6

   Construct an $\overline{X}$ and range chart based on this "initial" data. Comment on observations out of limits.
6. A sampling plan for testing sterility of a batch of 100,000 ampuls is as follows. Test 100 ampuls selected at random. If there are no rejects, pass the batch. If there are one or more rejects, reject the batch. If 50 of the 100,000 ampuls are not sterile, what is the probability that the batch will pass?

7. A new method was tried by four analysts in triplicate.§

1	2	3	4
115	105	131	129
120	130	152	121
112	106	141	130

Perform an analysis of variance (one-way). Estimate the components of variance (between-analyst and within-analyst variance). What is the variance of the mean assay result if three analysts each perform four assays (a total of 12 assays)? What is the variance if four analysts each perform duplicate assays (a total of eight assays)? If the first analysis by an analyst costs $5 and each subsequent assay by that analyst costs $1, which of the two alternatives is more economical?

8. Construct an $\overline{X}$ chart for the data of Table 12.2, using the moving average procedure. Use the moving average to obtain $\overline{X}$ and $\overline{R}$ for the graph, from the first 15 batches. Plot results for first 15 batches only.

9. Duplicate assays were run for quality control purposes for production batches. The first 10 days of production resulted in the following data: (a) 10.1, 9.8; (2) 9.6, 10.0; (3) 10.0, 10.1; (4) 10.3, 10.3; (5) 10.2, 10.8; (6) 9.3, 9.9; (7) 10.1, 10.1; (8) 10.4, 10.6; (9) 10.9, 11.0; (10) 10.3, 10.4.
   (a) Calculate the mean, average range, and average standard deviation.
   (b) Construct a control chart for the mean and range and plot the data on the chart.

10. What are the lower and upper limits for the range for the example of the moving average discussed at the end of section 12.2.3?

11. What is the variance of the average of duplicate assays performed on the same tablet where the between-tablet variance is 4.9 and the within tablet variance is 2.7? Compare this to the variance of the average of singles assays performed on two different tablets.

12. How did 8 runs arise from the data in the example discussed in section 12.2.5?

13. For an assay that is being used to determine in-house limits, the within- and between-day variances are estimated as 0.3% and 0.5%, respectively. Tablet heterogeneity is 4%. The assay is performed in duplicate on the same day from the same composite.
   (a) Compute the in-house limits if the official specifications are 90% and 110% and there are 25 d.f. for the assay.
   (b) Compute in-house limits if single assays are performed on two different composites on the same day.

REFERENCES

1. Grant EL. Statistical Quality Control, 4th ed. New York: McGraw-Hill, 1974.
2. Duncan AJ. Quality Control and Industrial Statistics, 5th ed. Homewood, IL: Irwin, 1986.
3. U.S. Department of Defense. MIL-STD-105E, Military Sampling Procedures and Tables for Inspections by Attributes. Washington, DC: Superintendent of Documents, U.S. Government Printing Office, 1989.
4. Allen PV, Dukes GR, Gerger ME. Determination of release limits: a general methodology. Pharm Res 1991; 8:1210.
5. U.S. Department of Defense. MIL-STD-414, Sampling Plans. Washington, DC: Superintendent of Documents, U.S. Government Printing Office, 1989.
6. Haynes JD, Pauls J, Platt R. Statistical Aspects of a Laboratory Study for Substantiation of the Validity of an Alternate Assay Procedure, "The Green-Briar Procedure." Final Report of the Standing Committee on Statistics to the PMA/QC Section, March 14, 1977.
7. Youden WJ. Statistical Methods for Chemists. New York: Wiley, 1964.

§ Optional more difficult problem.

8. Mandel J. The Statistical Analysis of Experimental Data. New York: Interscience, 1964.
9. Bennet CA, Franklin NL. Statistical Analysis in Chemistry and the Chemical Industry. New York: Wiley, 1963.
10. Steel R, Torrie J. Principles and Procedures of Statistics. New York: McGraw-Hill Book Co., Inc., 1960.
11. United Stales v. Barr Labs, Inc., Consolidated Docket No. 92–1744 (AMW).
12. Bergum J, Utter M. Process validation. In: Chow S-C, ed. Encyclopedia of Pharmaceutical Statistics. New York: Marcel Dekker, 2000:422–440.
13. US Dept of Health and Human Services, FDA, CDER, 1997.
14. United States Pharmacopeia, 23, 1995 pp. 1791–1793.
15. Investigating Out of Specification (OOS) Test Results for Pharmaceutical Production US Dept of Health and Human Services, FDA, CDER, Sept 1998.
16. Box GE, Hunter WG, Hunter JS. Statistics for Experimenters. New York: Wiley, 1978.