13 Validation

Although validation of analytical and manufacturing processes has always been important in pharmaceutical quality control, recent emphasis on their documentation by the FDA has resulted in a more careful look at the implementation of validation procedures. The FDA defines process validation as "... a documented program which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specification and quality attributes" [1]. Pharmaceutical process validation consists of well-documented, written procedures ensuring that a specific pharmaceutical technology is capable of and is attaining what is specified in official or in-house specifications, for example, a specified precision and accuracy of an assay procedure or the characteristics of a finished pharmaceutical product. Validation can be categorized as either *prospective* or *retrospective.* Prospective validation should be applied to new drug entities or formulations in anticipation of the product's requirements and expected performance. Berry [2,3] and Nash [4] have reviewed the physical–chemical and pharmaceutical aspects of process validation.

Retrospective validation may be the most convenient and effective way of validating processes for an existing product. Data concerning the key in-process and finished characteristics of an existing product are always available from previously manufactured batches. Usually, there is sufficient information available to demonstrate whether or not the product is being manufactured in a manner that meets the specifications expected of it.

13.1 PROCESS VALIDATION

In order to achieve a proper validation, an in-depth knowledge of the pharmaceutical process is essential. Since the end result of the process is variable (e.g., sterility, potency assay, tablet hardness, dissolution characteristics), statistical input is essential to validation procedures. For example, experimental design and data analysis are integral parts of assay and process validation.

For new products, prospective process validation studies are recommended based on GMP guidelines. Products already marketed may not have been validated for various reasons, for example, products marketed before the formal introduction of validation procedures. Retrospective or concurrent validation methods are used to validate processes for products that have not been validated previously. To recommend specific procedures for validation is difficult because of the variety of products and conditions used during their manufacture. However, there are some common procedures, including issues of sampling, assaying, and statistical analysis, that deserve some discussion.

13.1.1 Retrospective Validation

The GMP guidelines referring to validation [1] suggest that either retrospective or prospective validation may be used to validate a process. Retrospective validation would be applicable for a product that has been on the market for which adequate data are available for evaluation. Although there is no theoretical lower limit on the number of lots needed for such an evaluation, 20 lots have been suggested as an approximate lower limit [5]. In fact, judgment is needed when deciding what constitutes an adequate number of batches. For example, for a product that is made infrequently, or for a product that has an impeccable history of quality, a small number of batches (perhaps 5–10) may be sufficient. Retrospective validation consists of an evaluation of product characteristics over time. The characteristics should consist of attributes that reflect the consistency, accuracy, and safety of the product. For solid dosage forms, the chief characteristics

to be evaluated are typically blend uniformity, content uniformity, final assay, dissolution, and hardness (for tablets). The most simple and direct way of evaluating and displaying these characteristics is via control charts (see sect. 12.2.2). Each attribute could be charted giving a visual display of the batch history. All batches that were released to the market should be included. However, if a rejected batch is clearly out of specifications, inclusion in the charting calculations could bias the true nature of the process. Certainly, if not included in the control charts, the absence of such batches should be clearly documented. Again, scientific judgment would be needed to make decisions regarding the inclusion of such batches in the control charts. The control chart not only allows an evaluation of the consistency of the process, but also can be helpful in identifying problems and as an aid in setting practical in-house release limits. Thus, retrospective validation is a useful evaluative procedure, and, representing a relatively large number of batches over a long period of time gives detailed information on the product performance.

13.1.2 Prospective Validation

On the other hand, prospective validation must always be performed for a new product during initial development and production. Usually, the first three production batches are evaluated in great detail in order to demonstrate consistency and accuracy. The important feature of prospective validation is that the attributes measured reflect the important or critical characteristics of the process. This requires a knowledge of the process. Having identified these features, an experimental design and sampling plan that captures the relevant measurements is needed. Each type of dosage form or product is different and may require different considerations.

One should be careful to distinguish process validation from formulation development and optimization. The validation process follows the formulation and processing conditions (such as mixing) "optimization," critical attributes having been evaluated and determined for the manufacturing process. At this point, the question of whether or not the process results in a consistent, reproducible product is the primary concern.

13.1.3 Sampling in Process Validation

Sampling is an important consideration during process validation. The answers to where, when, and how to take samples, as well as sample size and how many samples to be taken, are often not obvious. Judgment is important and no firm rules can be given. For example, during the validation procedure for solid dose forms, samples are taken (1) during the blend stage, (2) when core tablets are produced if applicable, and (3) from the finished tablets. We speak of random samples during these procedures, but, in fact, it is not possible, or practical, to take samples randomly during production. Rather, we try to take samples that are representative of the material being tested. For example, during the blend testing, we sample from the mixer or drums, ensuring that the samples are taken from locations that are representative. Samples taken from a blender or mixer, for purposes of validation, should include areas where good mixing may be suspect because of the geometry of the blender. The number and size of samples to be taken depend on the purpose of the study. For purposes of testing drug content or potency, one or more well selected large size, composited samples may suffice. For purposes of testing uniformity during a validation study, many smaller size samples are necessary. A sample size equal to no more than three dosage units has been suggested in a recent court decision [5], but sample sizes as small as one dosage unit are now becoming routine. Where electrostatic effects may cause the assay of small samples to be biased, single dose weights washed out of the thief, or larger sample sizes may give more reliable results. Although there are no rules for the number of samples to be taken, certainly 10 suitably selected samples should be sufficient when performing time-mix studies to determine the optimal mixing time. Having validated the process, for routine blend testing, assay of three to six samples, selected representatively, should be sufficient. The number of samples, if any, to be taken during production depends on the product and the process. For many products, blend testing may be eliminated for production lots after the process has been validated. A product that has a history of performing well will need less extensive sampling than one that has shown a propensity for problems during its history.

During routine production, if blend assays are indicated, samples are typically taken from drums rather than the mixer, as a matter of convenience. Also, sampling from the drums represents a further step in the process, so that if the blend is satisfactory at this stage, one has more confidence in the integrity of the batch than if samples are taken only from the blender. When drums are tested for blend uniformity and drug content during validation, each drum may be sampled. In addition, some or all of the drums may be sampled more than once; top, middle, and bottom, for example. Some companies sample the first and last drums extensively, from top, middle, and bottom, and the intervening drums only once. Again, as with sampling from the mixer, the size of the sample requires judgment, based on the nature of the product and the objective of the test.

The assays obtained from the drums can be analyzed for drug potency and uniformity. These assays should show a relatively small RSD, so that one has confidence that the RSD of the final product will be within limits, based on content uniformity assays. Although we cannot ensure that every portion of the mix is identical since the product is by nature not uniform, we would hope that the uniformity is good based on RSD requirements for the finished product (less than 6%). For example, if the RSD at an intermediate stage (such as a blend) were greater than 5%, some doubt would exist about the adequate uniformity of the mix.

In addition to blend testing for purposes of process validation or routine QC sampling, sampling for intermediate product testing (e.g., tablet cores) and final product testing is important. Some sampling procedures have been reviewed in chapter 4. Product is usually sampled by selecting units throughout the run from the production lines, by QC personnel. The sample is then submitted for assay as a composite of, for example, tablets over the entire run. This is a form of stratified sampling, tablets being selected every hour during the run, for example. Since final product analysis is usually specified in official documents (content uniformity, dissolution, and a potency composite assay), the number of samples to be analyzed is prespecified. The samples to be analyzed are taken, at random, from the units supplied by the QC department. For validation, additional assays are usually performed to ensure uniformity and compliance. For example, content uniformity and dissolution tests may be performed on dosage units selected from the beginning, middle, and end of the run. Coated products may be tested from different coating pans. For all of these tests, in validation studies, analysis of both average drug potency and uniformity is important. Statistical tests are less useful than statistical estimation in the form of confidence intervals or point estimates. For example, in a validation study, if content uniformity tests are run for tablets at the beginning, middle, and end of the run, we would look at the results from each of the three content uniformity tests to ensure that the RSD was similar and comfortably within the official limits without subjecting the data to a rigorous statistical test.

There are no specific rules. The GMPs and validation guidelines are only recommendations. If a standard procedure is implemented within a company, the procedure should be examined with regard to each product, to ensure that a particular product is not unique in some way that would require a variation in the testing procedure. Careful testing, based on good judgment and science, benefits both the consumer and the manufacturer.

Statistical analysis of the data is useful. However, statistical methods should be used to aid in an understanding of the data only. Hypothesis testing may not be useful, in part because of power considerations. Scientific judgment should prevail.

Several examples will be given with solutions to illustrate the "validation" train of thought. There is no unique statistical approach to any single problem in most practical situations. In validation procedures, in particular, there will be more than one way of attacking a problem. What is most needed is a clear idea of the problem and some common sense. In all of the following examples, statistical methods will be used that have been discussed elsewhere in this book.

Example 1: Retrospective validation. Quality control data are available for an ointment that has been manufactured during a period of approximately one year. The in-process (bulk) product is assayed in triplicate for each batch (top, middle, and bottom of the mixing tank). The finished product consists of either a 2-oz container or a 4-oz container, or both. A single assay is performed on each size of the finished product. The assay results for the eight batches manufactured are shown in Table 13.1.

Table 13.1 Results of Bulk and Finished Tablet Assays of Eight Batches

	In-process bulk material (%)				Finished product (%)		
Batch	**Top**	**Middle**	**Bottom**	**Average**	**2 oz**	**4 oz**	**Average**
1	105	106	106	105.7	104	101	102.5
2	105	107	103	105.0	108	107	107.5
3	102	109	105	105.3	—	107	107.0
4	105	104	104	104.3	105	107	106.0
5	106	104	107	105.7	107	102	104.5
6	110	108	107	108.3	108	107	107.5
7	103	105	105	104.3	102	104	103
8	108	112	114	111.3	113	—	113
Avg.	105.5	106.9	106.4	106.24	106.7	105	106.38
s.d.	2.56	2.75	3.38	2.40	—	—	3.31

The following questions must be answered to pursue the process validation of this product:

1. Are the assays within limits as stated in the in-house specifications?
2. Do the average results differ for the top, middle, and bottom of the bulk? This can be considered as a measure of drug homogeneity. If the results are (statistically or practically) different in different parts of the bulk container, mixing heterogeneity is indicated.
3. Are the average drug concentration and homogeneity in the bulk mix different from the average concentration and homogeneity of the finished product?
4. Are batches in control based on the charting of averages using control charts?

Answers:

Question 1. The in-house specifications call for an average assay between 100% and 120%. All batches pass based on the average results of both the bulk and finished products. Batch 8 has a relatively high assay, but still falls within the specifications.

Question 2. A two-way analysis of variance (chap. 8) is used to test for equality of means from the top, middle, and bottom of the bulk container. The average results are shown in Table 13.1, and the ANOVA table is shown in Table 13.2. The F test shows lack of significance at the 5% level, and the product can be considered to be homogeneous. The assays of top, middle, and bottom are treated as *replicate assays* for purposes of determining within-batch variability. (Some statisticians may not recommend a two-step procedure where a preliminary statistical test is used to set the conditions for a subsequent test. However, in this case for purposes of validation in the absence of true replicates, there is little choice.) Note that if the average results of top, middle, and bottom showed significant differences (from both a practical and statistical point of view), a mixing problem would be indicated. This would trigger a study to optimize the mixing procedure and/or equipment to produce a relatively homogeneous product. We understand that a heterogeneous system, as exemplified by an ointment, can never be perfectly homogeneous. The aim is to produce a product that has close to the same concentration of material in each part. From Table 13.2, the within-batch variation is obtained by pooling the between position (top, middle, bottom) sum of squares and the error sum of squares. The within-batch error (variance) estimate is $64.67/16 = 4.04$ with 16 d.f. The standard deviation

Table 13.2 ANOVA for Top, Middle, and Bottom of Bulk

Source	**d.f.**	**SS**	**MS**	***F***
Batches	7	121.8	17.4	—
Top–middle–bottom	2	7.75	3.88	0.95
Error	14	56.92	4.07	—
Total	23	186.5		

Table 13.3 Paired t Test for Comparison of 2- and 4-oz Containers (Omit Batches 3 and 8)

Average of 2 oz	=	105.67
Average of 4 oz	=	104.67
t	=	$1/(2.76\sqrt{1/6}) = 0.89$

is the square root of the variance, 2.01. This would be the same error that would have been obtained had we considered this a one-way ANOVA with eight batches and disregarded the "top–middle–bottom" factor. Again, statistical analysis of the data is useful. However, statistics should be used to help understand the data only. Hypothesis testing may not be useful because of power considerations, and scientific judgment should always prevail.

Question 3. The comparison of the variability in the bulk and finished product would be a test of change in homogeneity due to handling from the bulk to the finished product. Although this may not be expected in a viscous, semisolid product such as an ointment, a test to confirm the homogeneity of the finished product should be carried out if possible. In powdered mixes such as may occur in the bulk material for tablets, a disruption of homogeneity during the transformation of bulk material into the final tablet is not an unlikely occurrence. For example, movement of the material in the containers during transport, or vibrations resulting in the settling and sifting of particles in the storage containers prior to tableting, may result in preferential settling of the materials comprising the tablet mix.

In order to compare the within-batch variability of the bulk and finished product, a within-batch error estimate for the finished product is needed. We can use a similar approach to that used for the bulk. Compare the average results for the two different containers (when both sizes are manufactured) and if there is no significant difference, consider the results for the two finished containers as duplicates. The analysis comparing the average results for the 2- and 4-oz containers for the six batches where both were manufactured is shown in Table 13.3. The paired t test shows no significant difference ($p > 0.05$). The within-batch variation is obtained by pooling the error from each of the six pairs, considering each pair a duplicate determination.

$$\text{Within} - \text{mean square} = \frac{(104-101)^2 + (108-107)^2 + \ldots + (102-104)^2}{2 \times 6} = 3.67.$$

This estimate of the within-batch variation is very close to that observed for the bulk material (3.67 vs. 4.04). If a doubt concerning variability exists, a formal statistical test may be performed to compare the within-batch variance of the bulk and finished products for the six batches (F test; sect. 5.3) where estimates of the variability of both the bulk and finished product are available. (We can assume that all variance estimates are independent for purposes of this example.) The results show no evidence of a discrepancy in homogeneity between the bulk and finished product. Although this approach may seem complex and circuitous, in retrospective, undesigned experiments, one often must make do with what is available, making reasonable and rational use of the available data.

The average results of the bulk and finished product can be compared using a paired t test. For this test, we first compute the average result of the bulk and finished material for each batch. The average results are shown in Table 13.1. The t test (Table 13.4) shows no significant difference between the average results of the bulk and finished material.

If either or both of the tests comparing bulk and finished product (average result or variance) show a significant difference, the data should be carefully examined for outliers or

Table 13.4 Paired t Test Comparing the Average of the Bulk and Finished Product

Average of bulk	=	106.24
Average of finished	=	106.38
t	=	$0.14/1/(2.03\sqrt{1/8}) = 0.20$

Table 13.5 Computations for the Moving Average Control Chart Shown in Figure 13.2

Moving average ($N = 3$)	Moving range
105.33	0.7
104.87	1
105.1	1.4
106.1	4
106.1	4
107.97	7
Av. 105.91	3.02

obviously erroneous data, or research should be initiated to find the cause. In the present example, where data for only eight batches are available, if the cause is not apparent, further data may be collected as subsequent batches are manufactured to ensure that conclusions based on the eight batches remain valid.

Question 4. A control chart can be constructed to monitor the process based on the data available. This chart is preliminary (only eight batches) and should be updated as new data become available. In fact, after a few more batches are produced, the estimates and comparisons described above should also be repeated to ensure that bulk and finished product assays are behaving normally. The usual Shewhart control chart for averages uses the within-batch variation as the estimate of the variance (see chap. 12). Sometimes in the case of naturally heterogeneous products, such as ointments, tablets, and so on, a source of variation between batches is part of the process that is difficult to eliminate. In these cases, we may wish to use between-batch variation as an estimate of error for construction of the control chart. As long as this approach results in limits that are reasonable in view of official and/or in-house specifications, we may feel secure. However, to be prudent, one would want to find reasons why batches cannot be more closely reproduced. The within-batch variation for the bulk material was estimated as (s.d.) 2.01. A control chart with 3 sigma limits could be set up as $\bar{X} \pm 3 \times 2.01/\sqrt{3} = 106.2 \pm 3.5$ based on the average of the top, middle, and bottom assays. Because of the presence of between-lot variability, a moving average control chart may be appropriate for this data. This chart is constructed from the averages of the three bulk assays for the eight batches (Table 13.1) using a control chart with a moving average of size 3. Table 13.5 shows the calculations for this chart.

For samples of size 3, from Table IV.10, the control chart limits are $105.91 \pm 1.02(3.02) = 105.91 \pm 3.08$. Figures 13.1 and 13.2 show the control charts based on within-batch variation and that based on the moving average. Note that the moving average chart show no out-of-control values and would include batch number 8 within the average control chart limits. The control chart based on within-batch variation finds batch number 8 out of limits.

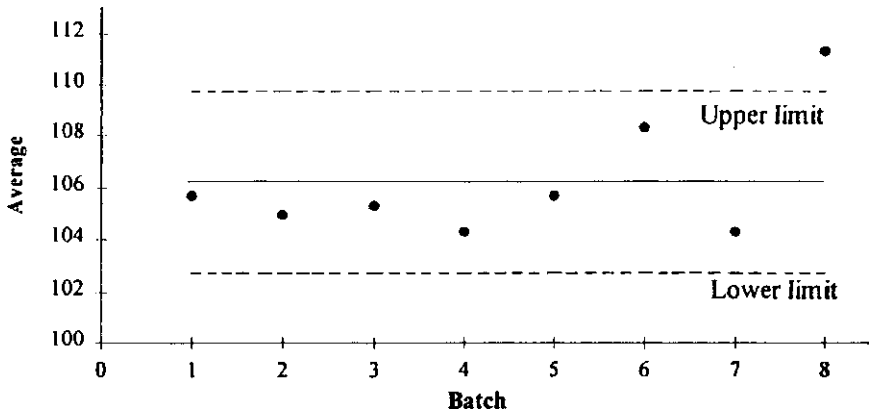


Figure 13.1 Control chart for Table 13.1 data using within-batch variation to construct limits.

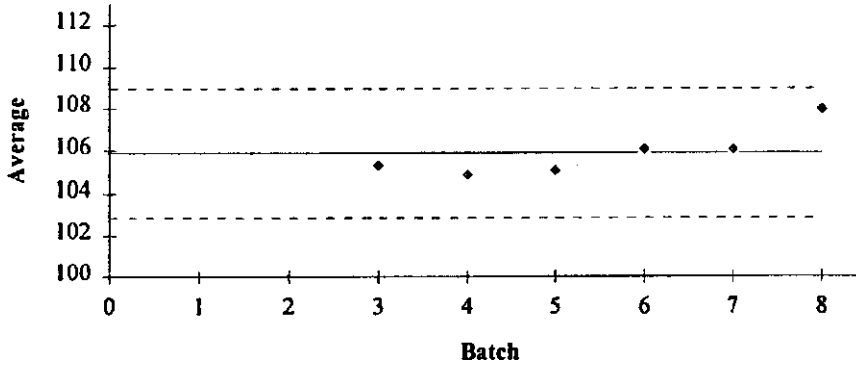


Figure 13.2 Moving average control chart for data of Table 13.1.

Within-batch variation appears to underestimate the inherent variation that includes between-batch variability. Until other sources of variability can be discovered, the moving average chart, which includes between-batch variation, appears to accomplish the objective, that is, to set up a control chart that allows monitoring of the average result of the manufacturing process.

A control chart for the moving range can be constructed using the factor for samples of size 3 in Table IV.10. The upper limit is $2.57 \times 3.02 = 7.76$.

Another control chart of interest is a "range chart" that monitors within-batch variability. If top, middle, and bottom assays are considered to be replicates, we can chart the range of assays within each batch, a monitoring of product homogeneity. The construction of range charts is discussed in chapter 12. Figure 13.3 shows the range chart for the bulk data from Table 13.1 (see also Exercise Problem 1).

A control chart for the finished product is less easily conceived. Different batches may have a different number of assays depending on whether one container or two different size containers are manufactured. There are several alternatives here including the possibility of using (1) separate control charts for the two different sizes, (2) a control chart based on an average result, or (3) a chart with varying limits that depend on the sample size. Note that only a single assay was performed for each finished container. If separate control charts are used for each product, one may wish to consider assays from duplicate containers for each size container so that a range chart to monitor within-batch variability can be constructed. In the present case, limits for the average control chart for the finished product would be wider than that for the bulk average chart since each value is derived from a single (or duplicate) reading rather than the three readings from the bulk. (Note that if within variation is appropriate for construction of the control chart, as may occur with other products, one might use the pooled within variation from both the finished and bulk assays as an estimate of the variance to construct limits.) Exercise Problem 2 asks for the construction of a control chart for finished containers.

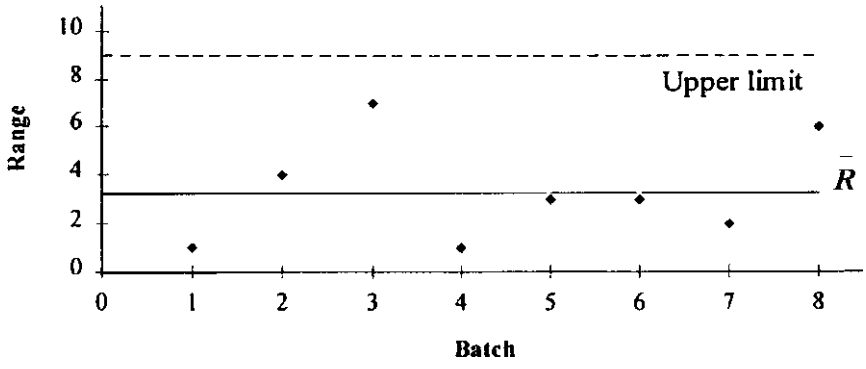


Figure 13.3 Range chart for Table 13.1 data.

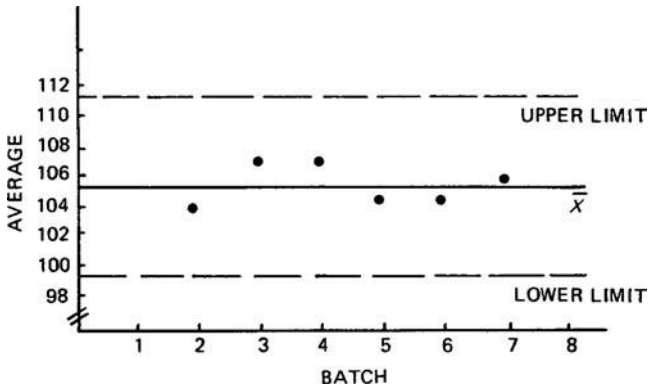


Figure 13.4 Moving average chart for a 4-oz container from Table 13.1 data.

A preliminary control chart using a moving average of size 2 is shown for the 4-oz container in Figure 13.4.

Should any values fall outside the control chart limits, appropriate action should be taken. Refer to the discussion on control charts in chapter 12.

Example 2: An example of a prospective validation study. In this example, a new manufacturing process is just underway for a tablet formulation of either (1) a new drug entity or (2) reformulation of an existing product. Since it would be difficult to generate data from many batches in a reasonable period of time, a recommended procedure is to carefully collect and analyze data from at least three consecutive batches. Of course, this procedure does not negate the necessity of keeping careful in-process and finished-product quality control records to ensure that the quality of the product is maintained.

Prior to the design of the validation procedure, a review of the critical steps in the manufacturing process is necessary. The critical steps will vary from product to product. For the manufacture of tablets, critical steps would include (1) homogeneity and potency after mixing and/or other processes in the preparation of bulk powder prior to tableting, (2) maintenance of homogeneity after storage of the bulk material prior to tableting, (3) the effect of the tableting process on potency as well as other important tablet characteristics such as content uniformity, hardness, friability, disintegration, and dissolution.

In this example, we consider a product in which potency and homogeneity are to be examined as indicators of the validation of the manufacturing process. To this end, both the bulk material and final product are to be tested. We will assume that the critical steps have been identified as (1) the mixing or blending step prior to compression and (2) the manufacture of the finished tablet. Therefore, the product will be sampled both prior to compression in the mixing equipment and after compression, the final manufactured tablet. Three mixing times will be investigated to determine the effect of mixing time on the homogeneity of the mix.

If many variables are considered to be critical, the number of experiments needed to test the effects of these variables may not be feasible from an economic point of view. In these cases, one can restrict the range of many of the variables based on a "knowledge" of their effects from experience. Other options include the use of fractional factorial designs or other experimental screening designs [6].

The question of how many samples to take, as well as where and how to sample is not answered easily. The answer will depend on the nature of the product, the manufacturing procedure, as well as a certain amount of good judgment and common sense. We are interested in taking sufficient samples to answer the questions posed by the validation process.

1. Does the process produce tablets that are uniform?
2. Does the process produce tablets that have the correct potency?
3. Does the variability of the final tablet correspond to the variation in the precompression powdered mix?

Table 13.6 Analysis of Bulk Mix in Blender and Final Tablets

Location	1	2	3	4	5	6
			5 minutes mixing time			
	101	104	101	104	101	109
	93	110	104	100	105	103
	102	106	96	94	99	105
Average	98.7	106.7	100.3	99.3	101.7	105.7
s.d.	4.93	3.06	4.04	5.03	3.06	3.06
			10 minutes mixing time			
	101	105	100	104	99	103
	103	102	99	100	103	104
	103	104	103	101	102	103
Average	102.3	103.7	100.7	101.7	101.3	103.3
s.d.	1.15	1.53	2.08	2.08	2.08	0.58
			20 minutes mixing time			
	102	100	101	99	101	103
	101	102	104	100	101	98
	104	103	100	102	105	102
Average	102.3	101.7	101.7	100.3	102.3	101.0
s.d.	1.53	1.53	2.08	1.53	2.31	2.65

	Final tablets		
	Beginning	Middle	End
	102	99	102
	98	100	103
	103	105	100
	100	101	100
	103	97	104
	103	102	102
	101	98	100
	100	103	97
	99	102	105
	104	100	101
Average	101.3	100.7	101.4
s.d.	2.00	2.41	2.32

Usually, samples are taken directly from the mixing equipment to test for uniformity. Samples may be taken from different parts of the mixer depending on its geometry and potential trouble spots. For example, some mixers, such as the ribbon mixer, are known to have "dead" spots where mixing may not be optimal. Such "dead" spots should be included in the samples to be analyzed. The finished tablets can be sampled at random from the final production batch, or sampled as production proceeds. In the present example, 10 samples (tablets) will be taken at each of the beginning, middle, and end of the tableting process.

Data for the validation of this manufacturing process are shown in Table 13.6. Triplicate assay determinations were made at six different locations in the mixer after 5, 10, and 20 minutes of mixing. In this example, six locations were chosen to represent different parts of the mixture. In other examples, samples may be chosen by a suitable random process. For example, the mixer may be divided into three-dimensional sectors, and samples taken from a suitable number of sectors at random. In the present case, each sample assayed from the bulk mix was approximately the same weight as the finished tablet. During tablet compression, 10 tablets were chosen at three different times in the tablet production run and drug content measured on individual tablets. This procedure was repeated for three successive batches to ensure that the process continued to show good reproducibility. We will discuss the analysis of the results of a single batch.

Table 13.7 ANOVA for Table 13.6

Description	Source	d.f.	MS	*F*
5-minute mix	Between	5	33.79	2.16
	Within	12	15.67	—
10-minute mix	Between	5	4.10	1.45
	Within	12	2.83	—
20-minute mix	Between	5	1.82	0.46
	Within	12	3.94	—
Tablets	Between	2	1.43	0.28
	Within	27	5.06	—

Analysis of variance can be used to estimate the variability and to test for homogeneity of sample averages from different parts of the blender or from different parts of the production run (Table 13.7).

For the bulk mix, none of the F ratios for between sampling locations mean squares are significant. This suggests that drug is dispersed uniformly to all locations after 5, 10, and 20 minutes of mixing. However, the within-MS is significantly larger in the 5-minute mix compared to the 10- and 20-minute mixes. A test of the equality of variances can be performed using Bartlett's test or a simple F test, whichever is appropriate (see Exercise Problem 3). The data suggest a minimum mixing time of 10 minutes. The homogeneity of the finished tablets is not significantly different from the bulk mixes at 10 and 20 minutes as evidenced by the within-MS error term. The tablet variance is somewhat greater than that in the mix (5.06 compared to 2.83 and 3.94 in the 10- and 20-minute bulk mixes). This may be expected, a result of moving and handling of the mix subsequent to the mixing and prior to the tableting operation.

The average results and homogeneity of the final tablets appear to be adequate. Nevertheless, it would be prudent to continue to monitor the average results and the within variation of both the bulk mix and finished tablet during production batches using appropriate control charts. Again, a moving average chart, where between-batch rather than within-batch variance is the measure of variability, may be necessary in order to keep results for the average chart within limits.

13.2 ASSAY VALIDATION

Validation is an important ingredient in the development and application of analytical methodology for assaying potency of dosage forms or drug in body fluids. Assay validation must demonstrate that the analytical procedure is able to accurately and precisely predict the concentration of unknown samples. This consists of a "documented program which provides a high degree of assurance that the analytical method will consistently result in a recovery and precision within predetermined specifications and limits." To accomplish this, several procedures are usually required. A calibration "curve" is characterized by determining the analytical response (optical density, area, etc.) over a suitable range of known concentrations of drug. Unknown samples are then related to the calibration curve to estimate their concentrations. During the validation procedure, calibration curves may be run in duplicate for several days to determine between- and within-day variation. In most cases, the calibration curve is linear with an intercept close to 0. The proof of the validity of the calibration curve is that known samples, prepared independently of the calibration samples, and in the same form as the unknown samples (tablets, plasma, etc.), show consistently good recovery based on the calibration curve. By "good," we mean that the known samples show both accurate and precise recovery. These known samples are called quality control (QC) samples and are used in both the assay validation and in real studies where truly unknown samples are to be assayed. Typically, the QC samples are prepared in three concentrations that cover the range of concentrations expected in the unknown samples, and are run in duplicate. The QC samples are markers and as long as they show good recovery, the assay is considered to be performing well, as intended.

In general, specific statistical procedures are not recommended by the FDA. This is not necessarily negative as judgment is needed for the many different scenarios that are possible

when developing new assays. For example, linearity "should be evaluated by visual inspection." If linearity is accepted, then standard statistical techniques can be applied, such as fitting a regression line by least squares (see sect. 7.5). Transformations to achieve linearity are encouraged. "The correlation coefficient, y-intercept, slope of the regression line and residual sum of squares should be submitted." An analysis of residuals is also recommended.

Some definitions used in assay methodology and validation follow:

Accuracy: Closeness of an analytical procedure result to the true value.
Precision: Closeness of a series of measurements from the same homogeneous sample.
Repeatability: Closeness of results under the same conditions over a short period of time (intra-assay precision).
Interlaboratory (collaborative studies): Studies comparing results from different laboratories. This is not recommended for approval of marketing. This is used more for defining standardization of official assays.
Detection limit: Lowest level that can be detected but not necessarily quantified. The signal-to-noise ratio is used when there is baseline noise. Compare low concentration samples with a blank. "Establish the minimum concentration at which the analyte can be reliably detected." A signal-to-noise ratio of 2/1 or 3/1 is considered acceptable. The detection limit may be expressed as

$$\mathrm{DL} = \frac{3.3\,\mathrm{sigma}}{S},$$

where sigma is the s.d. of response and S is the slope of calibration curve.
Quantitation limit (QL): The QL is determined with "known concentrations of analyte, and by establishing the minimum level at which the analyte can be quantified with acceptable accuracy and precision."

A typical calculation of QL is

$$\mathrm{QL} = \frac{10\,\mathrm{sigma}}{\mathrm{Slope}}.$$

Good experimental design should be carefully followed in the validation procedure. Careful attention should be paid to the use of proper replicates and statistical analyses. In the following example, the calibration curve consists of five concentrations and is run on three days. Separate solutions are freshly prepared each day for construction of the calibration curve. A large volume of a set of QC samples at three concentrations is prepared from the start to be used throughout the validation and subsequent analyses. A complete validation procedure can be rather complicated in order to cover the many contingencies that may occur to invalidate the assay procedure. In this example, only some of the many possible problems that arise will be presented. The chief purpose of this example is to demonstrate some of the statistical thinking needed when developing and implementing assay validation procedures.

The results of the calibration curves run in duplicate on three days are shown in Table 13.8 and Figure 13.5.

As is typical of analytical data, the variance increases with concentration. For the fitting and analysis of regression lines, a weighted analysis may be used with each value weighted by $1/X^2$, where X is the concentration. For analysis of variance, either a weighted analysis or a log transformation of the data can be used to get rid of the variance heterogeneity (heteroscedasticity). Analyses will be run to characterize the reproducibility and linearity of these data. The calibration lines are at the heart of the analytical procedure as these are used to estimate the unknown samples during biological (e.g., clinical) studies or for QC.

ANOVA: Table 13.9 shows the analysis of variance for the data of Table 13.8 after a log (ln) transformation. The analysis is a three-way ANOVA with factors days (random), replicates (fixed), and concentration (fixed). The two replicates from Table 13.8 are obtained by running all concentrations at the beginning of the day's assays and repeating the procedure at the end of the day. Although the ANOVA for three factors has not been explained in any detail in this

Table 13.8 Calibration Curve Data for Validation (Peak Area)

Day	Concentration	Replicate 1	Replicate 2	Average
1	0.05	0.003	0.004	0.0035
	0.20	0.016	0.018	0.017
	1.00	0.088	0.094	0.092
	10.0	0.920	0.901	0.9105
	20.0	1.859	1.827	1.843
2	0.05	0.006	0.004	0.005
	0.20	0.024	0.020	0.022
	1.00	0.108	0.116	0.112
	10.0	1.009	1.055	1.032
	20.0	2.146	2.098	2.122
3	0.05	0.005	0.008	0.0065
	0.20	0.019	0.023	0.021
	1.00	0.099	0.105	0.102
	10.0	1.000	0.978	0.989
	20.0	1.998	2.038	2.018

book, the interpretation of the ANOVA table follows the same principles presented in chapters 8 and 9, "Analysis of Variance" and "Factorial Designs," respectively.

The terms of interest in Table 13.9 are replicates and replicate × concentration (BC) interaction. If the assay is performing as expected, neither of these terms should be significant. A significant replicate term indicates that the first replicate is giving consistently higher (or lower) results than the second. This suggests some kind of time trend in the analysis and should be corrected or accounted for in an appropriate manner. A replicate × concentration interaction suggests erroneous data or poor procedure. This interaction may be a result of significant differences between replicates in one direction at some concentrations and opposite differences at other concentrations. For example, if the areas were 1.0 and 1.2 for replicates 1 and 2, respectively, at a concentration of 10.0, and 2.3 and 2.1 at a concentration of 20.0, a significant interaction may be detected. Under ordinary conditions, this interaction is unlikely to occur.

A least squares fit should be made to the calibration data to check for linearity and outliers. A weighted regression is recommended as noted above (see also sect. 7.7). This analysis is performed if the ANOVA (Table 13.9) shows no problems. A single analysis may be performed

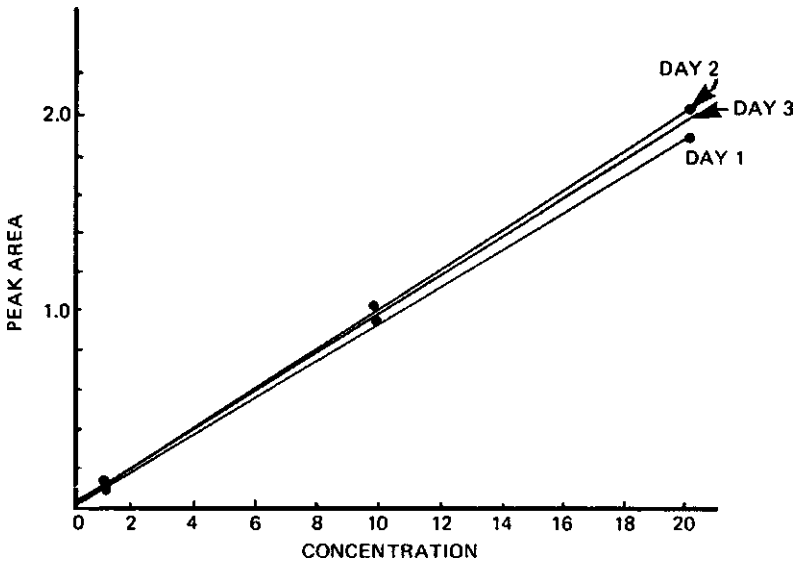


Figure 13.5 Calibration curves from Table 13.8 (weighted least squares fits).

Table 13.9 Analysis of Variance for Calibration Data (Log Transformation)

Source	d.f.	SS	MS	*F*
Days (A)	2	0.3150	0.1575	—
Replicates (B)	1	0.01436	0.01436	0.36
Concentrations (C)	4	155.78	38.945	1528*
AB	2	0.0803	0.04016	—
AC	8	0.2038	0.0255	—
BC	4	0.0155	0.0387	0.18
ABC	8	0.1742058	0.02178	—
Total	29	156.5834		

$*p < 0.01$.

for all three (days) calibration curves, but experience suggests that calibration curves may often vary from day to day. (This is the reason for the use of QC samples, to check the adequacy of each calibration curve.) In the present case, regression analysis is performed separately for each day's data. Table 13.10 shows the analysis of variance for the weighted least squares fit for the calibration data on day 1 (weight $= 1/X^2$). Each concentration is run in duplicate. The computations for the analysis are lengthy and are not given here. Rather, the interpretation of the ANOVA table (Table 13.10) is more important.

The important feature of the ANOVA is the test of deviations from regression (deviations). This is an *F* test (deviation-MS/within-MS) with 3 and 5 d.f. The test shows lack of significance (Table 13.10) indicating that the calibration curve can be taken as linear. This is the usual, expected conclusion for analytical procedures. If the *F* test is significant, the regression plot (Fig. 13.5) should be examined for outliers or other indications that result in nonlinearity (e.g., residual plots, chap. 7). Sometimes, even with a significant *F* test, examination of the plot will reveal no obvious indication of nonlinearity. This may be due to a very small within-MS error term, for example, and in these cases, the regression may be taken as linear if the other days' regressions show linearity. If curvature is apparent as indicated by inspection of the plot and a significant *F* test, the data should be fit to a quadratic model, or an appropriate transformation applied to linearize the concentration–response relationship. The test for linearity is discussed further in Appendix II.

A control chart may also be constructed for the slope and intercept of each day's calibration curve, starting with the validation data. This will be useful for detecting trends or outlying data.

A critical step in the assay validation procedure is the analysis of the performance of the QC samples. These samples provide a constant standard from day to day to challenge the validity of the calibration curve. In the simplest case, large volumes of QC samples at three concentrations are prepared to be used both in the validation and in the real studies. The concentrations cover the greater part of the concentration range expected for the unknown samples. The QC samples are run in duplicate (a total of six samples) throughout each day's assays. Usually, the samples will be run at evenly spaced intervals throughout the day with the three concentrations (low, medium, and high) run during the first part of the day and then run again during the latter part of the day. Each set of three should be run in random order. For example, the six QC samples may be interspersed with the unknowns in the following random order:

Medium . . . Low . . . High . . . Low . . . High . . . Medium

Table 13.10 ANOVA for Regression Analysis for Calibration Data from Day 1

Source	d.f.	SS	MS	*F*
Slope	1	0.056889	0.056889	1653.0
Error	8	0.000275	0.0000344	—
Deviations from regression	3	0.000004	0.0000013	0.02
Within (duplicates)	5	0.000271	0.0000542	—
Total	9	0.057164		
Slope (weighted regression) = 0.09153				
Intercept = −0.00109				

Table 13.11 Data for Quality Control Samples (% Recovery)

Day	Concentration	Replicate 1	Replicate 2	Average
1	0.50	106.5	103.9	105.2
	1.50	97.8	102.4	100.1
	15.0	101.6	97.2	99.4
2	0.50	99.4	107.6	103.5
	1.50	104.0	105.4	104.7
	15.0	96.9	100.7	98.8
3	0.50	97.4	100.2	98.8
	1.50	100.6	99.2	99.9
	15.0	104.2	101.8	103.0

Table 13.11 shows the results for the QC samples, in terms of percent accuracy, during the validation procedure. Percent accuracy is used to help equalize the variances for purposes of the statistical analysis. The first step is to perform an ANOVA for the QC results using all of the data. In this example, the factors in the ANOVA are days (3 days), concentrations (3 concentrations), and replicates (2, beginning of run vs. end of run). The ANOVA table is shown in Table 13.12.

Table 13.12 should not indicate problems if the assay is working as expected. No effect should be significant. A significant replicates effect indicates a trend from the first set of QC samples (beginning of run) to the second set. A significant replicate × concentration interaction is also cause for concern, and the data should be examined for errors, outliers, or other causes. Table 13.12 shows no obvious evidence of assay problems.

To test that the assay is giving close to 100% accuracy, a t test is performed comparing the overall average of all the QC samples versus 100%. This is a two-sided test

$$t = \frac{|\text{Overall average} - 100|}{\sqrt{\text{Days MS}/3}}, \tag{13.1}$$

where 3 = number of days. This is a weak test with only 2 d.f. If no significant effects are obvious in the ANOVA, one may perform the t test on all the data disregarding days and replicates ($N = 18$), and the t test would be

$$t = \frac{|\text{Overall average} - 100|}{\sqrt{S^2/18}}. \tag{13.2}$$

Table 13.12 Analysis of Variance for Quality Control Samples

Source	d.f.	SS	MS	*F*
Days (A)	2	9.418	4.709	—
Replicates (B)	1	5.556	5.556	0.44
Concentrations (C)	2	13.285	6.642	0.31
AB	2	25.498	12.749	—
AC	4	84.675	21.169	—
BC	2	11.231	5.616	0.73
ABC	4	30.956	7.739	—
Total	17	180.618		

The interpretation of this test should be made with caution because of the assumption of the absence of day, replicate, concentration, and interaction effects. For the data of Table 13.11, the t tests [Eqs. (13.1) and (13.2)] are

$$t = \frac{|101.489 - 100|}{\sqrt{4.709/3}} = 1.188 \tag{13.3}$$

$$t = \frac{|101.489 - 100|}{\sqrt{(180.618/17)/18}} = 1.938. \tag{13.4}$$

We can conclude that the assay is showing close to 100% accuracy. Should the t test show significance, at least one of the three QC concentrations is showing low or high assay accuracy. The data should be examined for errors or outliers, and if necessary, each concentration analyzed separately. The t tests would proceed as above but the data for a single concentration would be used. For the low concentration in Table 13.11, the t test (ignoring the day and replicate effects), would be

$$t = \frac{|102.5 - 100|}{4.12\sqrt{1/6}} = 1.486.$$

To monitor the assay performance, control charts for QC samples may be constructed starting with the results from the validation data. Control charts may be used for each QC concentration separately or, if warranted, all QC concentrations during a day's run can be considered replicates. In the example to follow, we examine the control chart for each QC concentration separately and use the medium concentration as an example. Probably, the best approach is to use a control chart for individuals or a moving average chart (see chap. 12). The validation data cover only three days. Following the validation, data were available for six more days using unknown samples from a clinical study. The data for the medium QC sample from the three validation days and the six clinical study days are shown in Table 13.13.

The average moving range is 2.62 based on samples of size 2. The overall average (of the "average" column in Table 13.13) is 101.17. The 3 sigma limits are $101.17 \pm 3(2.62/1.128) = 101.17 \pm 6.97$. The control chart is shown in Figure 13.6. All the results fall within the control chart limits. Another control chart can be constructed for the range for the duplicate assays performed each day. The average range is 2.38. The upper limit for the range chart is 7.78 (see Exercise Problem 4). As for all control charts, the average and limits should be updated as more data become available.

The control chart for the individual daily averages of the QC samples and the control charts for the slope and intercept, if desired, are used to monitor the process for the analysis of the unknown samples submitted during the clinical studies or for QC. If QC samples fall out of

Table 13.13 Data for Medium QC Sample (Concentration = 1.50) for Control Chart

Day	Replicate 1	Replicate 2	Average	Moving range
1	97.8	102.4	100.1	—
2	104.0	105.4	104.7	4.6
3	100.6	99.2	99.9	4.8
4	99.3	97.8	98.55	1.35
5	103.8	101.4	102.6	4.05
6	103.4	103.0	103.2	0.60
7	99.6	102.4	101.0	2.2
8	99.4	103.8	101.6	0.6
9	100.1	97.6	98.85	2.75

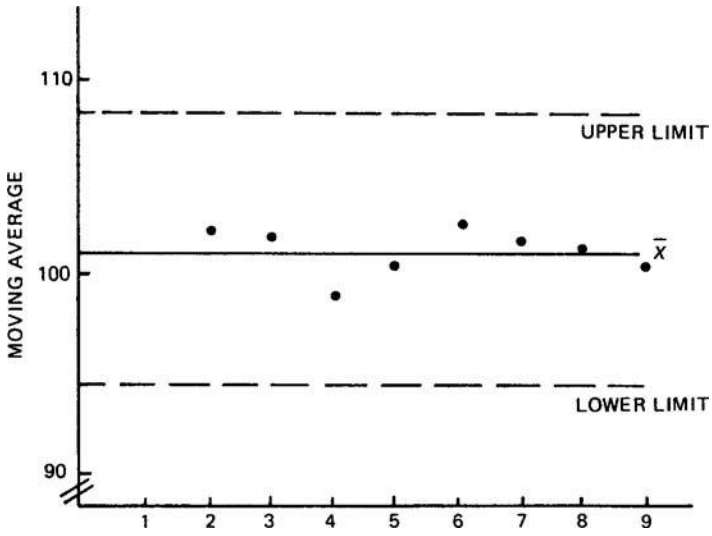


Figure 13.6 Moving average chart for Table 13.13 data.

limits and no obvious errors can be found, the analyses of the samples during that run may be suspect.

A detailed description of analytical validation for blood assays in bioavailability studies has been published by Lang and Bolton [7,8].

For further discussion of assay validation, see Ref. [9].

13.3 CONCLUDING REMARKS

In this chapter, some examples of statistical analysis and design of validation studies have been presented. As we have noted, statistical input is a necessary part of the design and analysis of validation procedures. The statistical procedures that may be used to analyze such data are not limited to the examples given here, but are dependent on the design of the procedures and the characteristics of the data resulting from these experiments. The design of the experiments needed to validate processes will be dependent on the complexity of the process and the identification of critical steps in the process. This is a most important part of validation and the research scientist should be very familiar with the nature of the process, for example, a manufacturing process or assay procedure [1,2,4]. The steps in the validation and statistical analysis are best implemented with the cooperation of a scientist familiar with the physical and chemical processes and a statistician. This is one of the many areas where such a joint venture can greatly facilitate project completion.

KEY TERMS

Assay validation
Average control chart
Calibration curve
Control chart
Critical steps
Moving average control chart
Moving range control chart
Process validation
Prospective validation
Quality control samples
Range control chart
Refractive validation
Weighted analysis

EXERCISES

1. Construct the range chart using within-batch variation for the bulk material in Table 13.1. Assume that the 3 readings within each batch are true replicates.
2. Construct a moving average chart ($n = 3$) for the 2-oz finished container in Example 2, Table 13.1.

3. Compare the variances during the mixing stage in Example 2 using Bartlett's test. (The variances are estimated from the within-MS terms in the ANOVAs in Table 13.7.)
4. Construct a range chart for the data of Table 13.12. Use the range of the daily duplicates to construct this chart.
5. Construct a control chart for individuals based on the data for three days for the low QC concentration from Table 13.11.

REFERENCES

1. Guideline of General Principles of Process Validation. Rockville, MD: FDA, 1991.
2. Berry IR. Process validation: practical applications to pharmaceutical products. Drug Dev Ind Pharm 1988; 14:377.
3. Berry IR, Nash R. Pharmaceutical Process Validation. New York: Marcel Dekker, 1993.
4. Nash R. In: Lieberman HA, Lachman L, Schwartz J, eds. Pharmaceutical Dosage Forms: Tablets, Vol 3, 2nd ed. New York: Marcel Dekker, 1990.
5. United States v. Barr Labs., Inc., Consolidated Docket No. 92–1744 (AMW) (Court Ruling 2/4/93).
6. Plackett RL, Burman J P. The design of optimum multifactorial experiments. Biometrika 1946; 33:305–325.
7. Lang JR, Bolton S. A comprehensive method validation strategy for bioanalytical applications in the pharmaceutical industry, Part I. J Pharm Biomed Anal 1991; 9:357.
8. Lang JR, Bolton S. A comprehensive method validation strategy for bioanalytical applications in the pharmaceutical industry–2. Statistical analyses. J Pharm Biomed Anal 1991; 9:435.
9. Schofield T. Assay validation. In: Chow S-C, ed. Encyclopedia of Pharmaceutical Statistics. New York: Marcel Dekker, 2000:21–30.

14 Computer-Intensive Methods

The widespread availability of powerful computers has revolutionized society. The field of statistics is no exception. Twenty years ago, the typical statistician had a collection of tables and charts, without which he or she would have been lost. Today one can perform complex statistical analyses without ever referring to a printed table, relying instead on the statistical functions available in many standard software packages. The ubiquitous availability of personal computing power and sophisticated programming packages permit us to approach statistics today in a less mathematical, more intuitive manner than is possible with the traditional formula-based approach.

The study of statistics by those lacking a strong mathematical background can be a daunting task. The traditional approach usually begins with the introduction of basic probability theory followed by a presentation of standard statistical distributions. To this point, nothing beyond algebra is required. Unfortunately, the progression to real-life problems and the development of inferential methods often involve the derivation of formulas. In many cases, this is accomplished through application of the calculus. The resulting formulas are generally neither intuitive nor simple to comprehend. Too often, the study of statistics is relegated to a process of memorization of these formulas that are then used in cookbook fashion. While the formula-based method of problem solving has an important place in statistics, it is often intimidating to the nonstatistician. For the statistician, this standard approach can become so automatic that the art of data analysis is lost and important characteristics of the data may go unrecognized. Using computer-intensive methods, we approach the solution of statistical problems through a logical application of basic principles applied to a computer-based experiment.

Computer simulations can let us explore the behavior of probability-based processes without becoming overly concerned about the underlying mathematics. When a real-life process can be formulated to follow, or to approximately follow, a known statistical distribution, its characteristics can be explored using Monte Carlo simulation. The Bootstrap Method [1] is a form of computer simulation that is applied to a specific set of data (a sample) without assuming any specific underlying statistical distribution. Bootstrap methods complement standard nonparametric statistical analyses. These are used when we do not know, or do not want to assume, what underlying statistical distribution is operative.

14.1 MONTE CARLO SIMULATION

Monte Carlo simulation enables exploration of complex, probability-based processes, many of which would be difficult to understand by even the most astute statistician using standard formula-based methods. In simulation, the computer performs a large number of experiments, such as the random drawing of balls from an urn, the tossing of a fair or biased coin, or the drawing of random samples from a Normal distribution. Solving a problem using computer simulation involves reducing it to a simple probability-based model, designing a sampling experiment based on the model, and then conducting the experiment, via the computer, a large number of times. The cumulative frequency distribution of the experimental outcomes is viewed as the cumulative probability distribution for the outcomes.

A simple example of how Monte Carlo simulation can be used instead of, or to complement, formula-based methods can be demonstrated using the antibiotic example of chapter 3. In this example, the cure rate for an antibiotic treatment is stated to be 0.75. The question posed is, what is the probability that three of four treated patients will have a cure? The analysis tool add-in of Microsoft Excel provides a convenient way to simulate an answer to this question. To activate this Excel option, if it is not already available in your installation, choose the

Tools option from the Main Menu and then select Add-ins. From the choices in the drop-down Add-ins menu, select (click on) both the Analysis ToolPak and the Analysis ToolPak-VBA. Both choices should show a check mark in their respective boxes.

To answer our antibiotic question, open a new Excel worksheet.

	A	B	C	D	E	F	G
1							
2							
3							
4							
5							
6							
7							
8							

Execute the following commands to simulate 30,000 flips of a biased coin, expected to land on heads 75% of the time and on tails 25% of the time:

From the Main Menu bar, choose Tools.
From the options listed under Tools, choose Data Analysis.
From the Data Analysis options, choose the Random Number Generator.

In the drop-down Dialog Box, enter the following:

For Number of Variables, enter 6.
For Number of Random Numbers, enter 30000.
For Distribution, select Binomial from the choices in the pop-up menu.
Enter 0.75 for the p value and 4 for the Number of Trials.
Enter 12345 for the Random Seed. (Any random number can be used for the seed.)
Click on Output Range and enter A1 in the area to the right of this option.
Click OK to start the simulation.

The commands instructed Excel to generate entries in the cells of the first six columns of the first 30,000 rows of the worksheet. The entry in each of these 180,000 cells represents the simulated number of successes (heads) observed in four independent Bernoulli trials (four flips of a biased coin). The possible outcome of each trial (flip) is either a 0 (tail) or a 1 (head), with the probability of getting a 1 (success) being 0.75 and the probability of getting a 0 (failure) being 0.25. (The coin is biased toward heads.) We might also have flipped a balanced tetrahedron with three sides labeled success and one labeled failure. The possible cell values are 0, 1, 2, 3, or 4 (number of heads in four flips of the coin). The following shows partial results of one simulation and the set of commands used to obtain these results.

	A	B	C	D	E	F	G
1	3	3	3	4	3	4	
2	4	3	3	3	3	3	
3	3	4	3	3	3	3	
4	2	3	4	1	3	2	
5	4	1	4	3	4	4	
6	2	3	3	3	2	4	
7	3	1	3	3	4	4	
8	3	1	1	4	4	3	

Commands in Simulation

Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	Enter 6	Generate 6 variables
Number of Random Numbers	Enter 30,000	Generate 6 variables 30,000 times
Distribution	Select Binomial	Simulate flips of a coin
p Value	Enter 0.75	Coin comes up heads 3 out of 4 flips
Number of Trials	Enter 4	Each variable is # heads in 4 flips
Random Seed	Enter 12345	Can enter desired value here
Output Range	Enter A1	Simulated values in cells A1 − F30000
OK	Click on this to perform the random numbers generation	

We need to determine the proportion of the 180,000 cells that have a simulated value of exactly 3 (three heads from four flips of the coin). The final set of commands to obtain a solution to our question is

Final Commands in Simulation

Into:	*Enter:*	*Result:*
Cell H1	= IF(A1 = 3,1,0)	Places a 1 if Al is a 3, 0 otherwise
Cells I1 through M1	Copy the formula from H1	Determines if 3 heads are in cells B1:F1
Cells H2 through M30000	Copy formulas from row 1 to rows 2 through 30,000	Determines where 3 heads occur in remaining cells
Cell G1	= AVERAGE(H1: M30000)	Proportion (probability) of 3 heads in 4 flips

	G	H	I	J	K	L	M
1	0.4235	1	1	1	0	1	0
2		0	1	1	1	1	1
3		1	0	1	1	1	1
4		0	1	0	0	1	0
5		0	0	0	1	0	0

The probability, 0.4235, observed in the simulation, compares favorably to the exact value, 0.4219, calculated using the formula for the binomial expansion. We can increase the accuracy of the simulation estimate by including more columns in the simulation or repeating it a number of times and using the average result of all the simulations. Performing the simulation, with the same seed, 12345, but using 10 columns (variables) instead of 6 gave a probability of 0.4222.

The Central Limit Theorem, used extensively in statistics, indicates that the shape of the distribution of sample means tends toward normality as the sample size increases. This occurs regardless of the underlying statistical distribution from which the sample is drawn. This important concept is not particularly intuitive. Computer simulation is a simple way to demonstrate the impact that the Central Limit Theorem has on the sampling process.

We use Excel to simulate samples drawn from the Uniform distribution, whose shape is markedly different from that of the Normal distribution. In the Uniform distribution, every value has an equal probability of occurrence. A histogram of independent, single samples (sample size of 1) is expected to be represented by a series of bars of equal height (frequency). As a result of the Central Limit Theorem, a histogram of the sample means, where the sample consists of a sufficient number of values drawn from the Uniform distribution, should have a pattern approximating the familiar bell-shaped curve of the Normal distribution. We can show this by performing a Monte Carlo simulation. We simulate the sampling of six values randomly and independently drawn from the Uniform distribution with range 0 to 1. We then determine the mean of the six values in the sample. The sampling is then repeated a large number of times. Histograms are constructed for both the first value from each set of six independent values and for the mean of the six independent values in each sample. The histogram of the single values shows how distinctly different the shape of the Uniform distribution is from that of the Normal distribution. The histogram of the sample mean demonstrates the power of the Central Limit Theorem, even when dealing with a relatively small number of values, only six, sampled from a distribution whose shape is extremely non-normal.

Open an Excel Worksheet and enter the labels shown in row 1

Commands in the Simulation

Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	6	Generate 6 variables in each trial
Number of Random Numbers	1000	Generate 1000 trials
Distribution	Uniform	Simulate the Uniform distribution
Between	0 and 1	Distribution range
Random Seed	12345	Can enter a different seed value if desired
Output Range	A2	Simulated values placed in cells A2–F1001
OK		Click to perform simulation
Cell G2	= Average(A2:F2)	Calculate mean of simulated values, trial 1
Cells G3:G100l	Copy G2 formula	Calculate mean for remaining trials
Cells H2:Hl2	0,0.1,0.2,. . ., 0.9,1.0	Bins for histogram bars
Cells 12:114	0.20,0.25,. . ., 0.75,0.80	Bins for histogram
Main Menu	Tools → Data Analysis → Histogram	
Dialog Box		
Input Range	A2:A1001	Use variable 1 values
Bin Range	H2:H12	Bin range
New Worksheet Ply	Check this option	
Chart Output	Check this option	
OK	Click to create histogram	
In New Worksheet Click on Histogram Chart		
Main Menu	Chart → Location	
As new sheet	Click this option and enter "Graph 1"	
Double Click on one of the histogram bars		
Options Tab	Click to open	
Gap Width	10	
Sheet 1	Click on this to return to simulation results	

Main Menu	Tools → Data Analysis → Histogram	
Dialog Box		
Input Range	G2:G1001	Use mean of the 6 values in each trial
Bin Range	I2:I14	Bin range
New Worksheet Ply	Check this option	
Chart Output	Check this option	
OK	Click to create histogram	

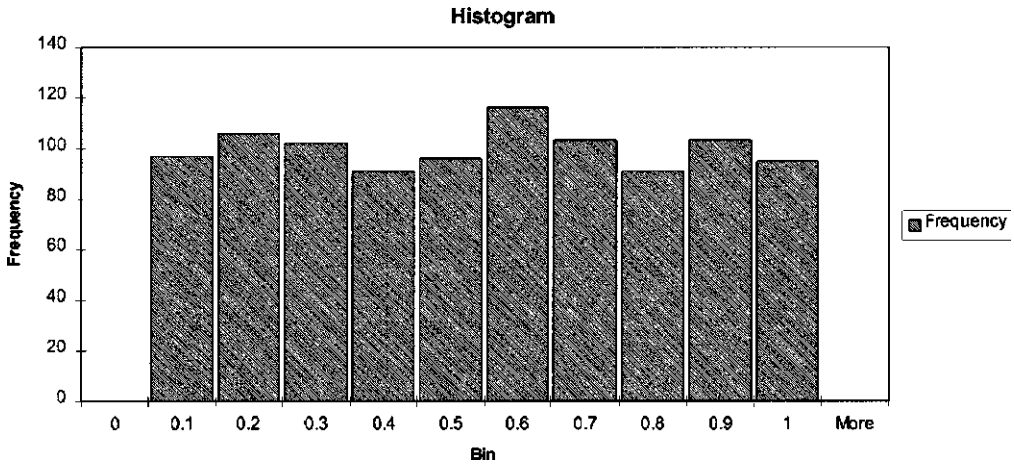
In New Worksheet Ply Click on Histogram Chart	
Main Menu	Chart → Location
As new sheet	Click this option and enter "Graph Mean"
Double Click on a Bar	
Options Tab	Click to open
Gap Width	10

	A	B	C	D	E	F	G	H	I
1	**Var 1**	**Var 2**	**Var 3**	**Var 4**	**Var 5**	**Var 6**	**Mean**	**Bins1**	**Bins Mean**
2	0.23	0.58	0.79	0.68	0.18	0.71	0.53	0.0	0.20
3	0.84	0.17	0.89	0.71	0.46	0.45	0.59	0.1	0.25
4	0.62	0.66	0.29	0.43	0.67	0.31	0.50	0.2	0.30
5	0.80	0.12	0.31	0.25	0.34	0.47	0.38	0.3	0.35
6	0.08	0.32	0.65	0.11	0.04	0.84	0.34	0.4	0.40
7	0.65	0.06	0.40	0.97	0.14	0.69	0.48	0.5	0.45
8	0.20	0.64	0.48	0.87	0.46	0.42	0.51	0.6	0.50
9	0.68	0.78	0.38	0.89	0.05	0.23	0.50	0.7	0.55
10	0.09	0.55	0.75	0.86	0.57	0.23	0.51	0.8	0.60
11	0.56	0.22	0.11	0.81	0.11	0.05	0.31	0.9	0.65
12	0.03	0.14	0.81	0.72	0.02	0.28	0.33	1.0	0.70
13	0.85	0.24	0.76	0.54	0.46	0.67	0.59		0.75
14	0.26	0.80	0.86	0.45	0.57	0.44	0.56		0.80

One of the most useful applications of computer simulation is in dealing with a complex probability problem. This can be demonstrated by an example based on FDA's guidance for industry entitled "Bioanalytical Method Validation," May 2001, copies of which are available at http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm064964.htm. The prescribed procedure for monitoring the accuracy and precision of a validated bioanalytical method, in routine use, involves the measurement of quality control (QC) samples, processed in duplicate, at each of three different concentrations. The QC samples are prepared in the same matrix (serum, plasma, blood, urine, etc.) as the samples with unknown concentrations to be analyzed. The three concentration levels of the QC samples cover the working range of the bioanalytical method, one in the lower region, a second at midrange, and the third in the upper region of the standard curve. QC samples are to be analyzed with each batch run of unknown samples. The run is acceptable if at least four of the six QC sample values are within 20% of their nominal concentrations. Two of the six samples may be outside the ± 20% acceptance region, but not two at the same concentration level.

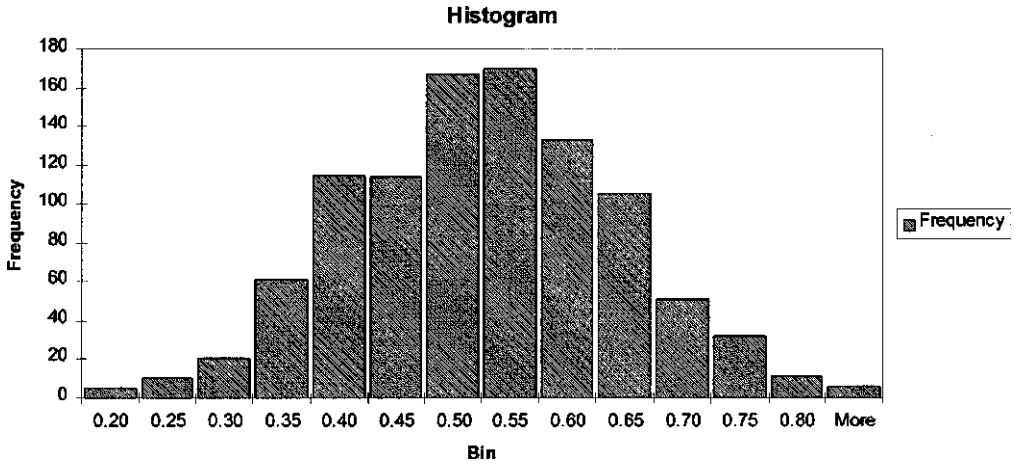
Assume that we have QC levels at 10, 250, and 750 ng/mL. Our assay method has a 15% CV (% relative standard deviation) over its entire working range. What proportion of batch

runs do we expect to reject when the assay is running as validated? Also assume that we have accurately prepared our QC samples and that any deviations in their assayed values are random errors that follow a Normal distribution (i.e., the mean deviation = 0%, standard deviation = CV% of the assay).



Histogram of simulated values (sample of size 1) from the Uniform distribution.

The traditional formula-based calculations rely on the known properties of the Normal and Binomial distributions. The probability that a single QC value will be within the acceptance region is equal to the proportion of the Standard Normal distribution, which lies between $-Z$ and $+Z$, where $Z = 20\%/\text{CV}\%$. With a CV% equal to 15%, the probability that any *single* QC value will be acceptable is the proportion of the Standard Normal distribution that lies between Z values of -1.33 and $+1.33$, or $p = 0.8176$. [$Z = (\text{X} - \mu)/\sigma = (20 - 0)/15 = 1.33$; see chap. 3]. The probability that a single QC value will fail to be accepted is $1 - \text{P}$ or $q = 0.1824$. The batch run is acceptable if all six QC values pass the criteria, five of six pass, or four of six pass. According to the binomial expansion, this probability is $p^6q^0 + 6p^5q^1 + 15p^4q^2$ (see chap. 3). However, three of the 15 ways that four QC values pass involve two failures at the same concentration level. This is not permitted by the FDA acceptance criteria. Therefore, this reduces the 15 possible ways of 4 QC values passing to 12 ways. The probability of run acceptance, based on the QC results, is $p^6 + 6p^5q^1 + 12p^4q^2$, or 0.88. We expect that 12% of our runs ($1 - 0.88 = 0.12$) will fail due to random error alone.



Histogram of the sample mean (n = 6) simulated from the Uniform distribution.

The simulation to evaluate this same question is easily accomplished using Excel. Open an Excel Worksheet and place the labels in the cells as shown in row 1.

	A	B	C	D	E	F	G
1	**QC1**	**QC1**	**QC2**	**QC2**	**QC3**	**QC3**	**Prob. Pass**
2							
3							
4							
5							

Commands in Simulation

Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	6	Generate 6 QC values for each run (row)
Number of Random Numbers	5000	Generate the values for 5000 runs
Distribution	Normal	Sample is from the Normal Dist.
Mean	0	True QC deviation is 0% (100% accurate)
Standard Deviation	15	CV for QC deviation is 15%
Random Seed	Enter 12345	Can enter a different seed if desired
Output Range	Enter A2	Variable values in cells A2 − F5001
Click OK		
Cell H2	= IF(ABS(A2) < 20,1,0)	If QCl deviation is < 20%, it passes (1)
Cells I2, H3:I500l	Copy H2 formula	Evaluates remaining QCl values
Cell J2	= IF((H2 + I2) >0,l,0)	QCl passes (1) if either replicate passes
Cells J3:J5001	Copy J2 formula	Evaluates runs 2–5000 for QCl passing
Cell K2	= IF(ABS(C2)<20,l,0)	If QC2 deviation is <20%, it passes (1)
Cells L2, K3:L5001	Copy K2 formula	Evaluates remaining QC2 values
Cell M2	= IF((K2 + L2)>0,l,0)	QC2 passes (1) if either replicate passes
Cells M3:M5001	Copy M2 formula	Evaluates runs 2–5000 for QC2 passing
Cell N2	= IF(ABS(E2)<20,1,0)	If QC3 deviation is <20%, it passes (1)
Cells O2, N3:O5001	Copy N2 formula	Evaluates remaining QC3 values
Cell P2	= IF((N2 + O2)>0,1,0)	QC3 passes (1) if either replicate passes
Cells P3:P5001	Copy P2 formula	Evaluatea runs 2–5000 for QC3 passing
Cell Q2	= J2∗M2∗P2	Flag is 1 if each QC level passes
Cells Q3:Q5001	Copy Q2 formula	Evaluates runs 2–5000 for passing each level
Cell R2	= IF((H2 + I2 + K2 + L2 + N2 + O2) >3,1,0	Flag is 1 if ≥4 QC passing

Cells R3:R5001	Copy R2 formula	Evaluates runs 2–5000 for ≥4 passing
Cell S2	= Q2∗R2	Flag value is 1 if all QC criteria are met
Cells S3:S5001	Copy S2 formula	Evaluates runs 2—5000 for meeting criteria
Cell G1	= Average(S2:S5001)	Probability of run passing (here, 0.8754)

	A	B	C	D	E	F	G
1	**QC1**	**QC1**	**QC2**	**QC2**	**QC3**	**QC3**	**Prob. Pass**
2	−11.01	3.22	11.95	6.82	−13.85	8.49	0.8754
3	14.83	−14.60	18.18	8.49	−1.64	−1.82	
4	4.74	6.32	−8.26	−2.66	6.61	−7.36	
5	12.40	−17.59	−7.36	−9.93	−6.22	−1.26	
6	−21.03	−7.02	5.91	−18.44	−26.17	15.19	
7	5.74	−22.78	−3.95	28.12	−16.39	7.45	
8	−12.51	5.25	−0.90	16.61	−1.51	−2.88	
9	7.18	11.53	−4.59	18.02	−25.29	−11.08	
10	−20.20	1.84	9.95	16.37	2.66	−11.17	

In this simulation of 5000 runs, 87.5% passed (probability = 0.8754) and 12.5% failed. These results are in close agreement with the theoretical values of 88% passing and 12% failing.

In the QC example, it would have been easier to apply the normal and binomial formulas rather than conducting the Excel simulation to answer our question. Had we wanted to investigate a more complex and perhaps more realistic situation, a simulation approach might be far simpler, and considerably more intuitive, than the formula-based approach. As an example, consider the situation where the standard deviation is 18% at the lowest concentration QC, 15% at the next higher concentration, and only 12% at the highest concentration. In addition, if the highest QC value exceeds the highest standard curve concentration, it cannot be reported so it is considered a failing value. It would be difficult to deal with this using the formula-based approach, but only marginally more difficult than our previous example if solved by simulation. The more complicated (realistic) our scenario, the more likely it is that computer simulation will prove to be the easier methodology to implement.

	H	I	J	K	L	M	N	O	P	Q	R	S
1	**P1_1**	**P1_2**	**Pass1**	**P2_1**	**P2_2**	**Pass2**	**P3_1**	**P3_2**	**Pass3**	**No 2**	**Pass4**	**Run**
2	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1
6	0	1	1	1	1	1	0	1	1	1	1	1
7	1	0	1	1	0	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	0	1	1	1	1	1
10	0	1	1	1	1	1	1	1	1	1	1	1

Monte Carlo simulation also offers an intuitive approach to hypothesis testing. In Table 5.9 of Chapter 5, the percent dissolutions after 15 minutes for two different tablet formulations, *A* and *B*, are listed. The distributions of the mean values for the two samples (10 values for each formulation) are assumed to follow Normal distributions. Is the average dissolution of

formulation *A* at 15 minutes different from that of formulation *B*? The formula-based approach relies on the application of the *t* test for the difference between two independent means as described in section 5.2.2. The calculated *t* statistic, 1.99, indicates that the probability of seeing a difference as large as that observed for these two formulations, if the two formulations are actually equivalent, is 0.062. The simulation approach requires applying only our knowledge that the variance of the sample mean is equal to the variance of the individual values divided by *n*, the size of the sample. The square root of this variance is the standard error of the mean. According to the Central Limit Theorem, the sample mean will tend to be normally distributed about its true mean value with a variability equal to the standard error.

Our question is formulated for a simulation solution by the following null hypothesis and its alternative:

H_0: The difference actually observed between the *A* and *B* means, 5.7, occurs by chance at least 5% of the time from two independent samples, each of size 10, taken from the same Normal distribution with mean and standard deviation equivalent to those in the combined (*A* + *B*) sample.

H_a: The difference observed between the sample means, 5.7, occurs less than 5% of the time by chance, indicating that it is unlikely that the two formulations represent the same population (i.e., their means are not equal).

The following is a simulation to evaluate our hypotheses:

	A	B	C	D	E	F	G	H
1	**Form**	**Percent**		**Sim *A***	**Sim *B***	**Abs(Diff)**	**GE 5.7**	**Prob. (Diff GE 5.7)**
2	***A***	68		72.7	74.7	2.1	0	0.065
3	***A***	84		76.0	75.2	0.7	0	
4	***A***	81		72.2	75.5	3.2	0	
5	***A***	85		76.4	72.1	4.3	0	
6	***A***	75		76.9	75.5	1.4	0	
7	***A***	69		74.0	74.0	0.0	0	
8	***A***	80		74.9	75.2	0.2	0	
9	***A***	76		73.1	73.9	0.8	0	
10	***A***	79		75.2	73.2	2.0	0	
11	***A***	74		76.1	71.7	4.4	0	
12	***B***	74		73.2	72.8	0.4	0	
13	***B***	71		73.3	74.1	0.7	0	
14	***B***	79		71.2	73.2	2.0	0	
15	***B***	63		75.1	71.6	3.5	0	
16	***B***	80		70.4	76.5	6.0	1	
17	***B***	61		75.1	70.9	4.1	0	
18	***B***	69		73.7	78.3	4.7	0	
19	***B***	72		71.9	75.3	3.5	0	
20	***B***	80		72.4	75.0	2.6	0	
21	***B***	65		74.1	76.7	2.5	0	
22				74.0	73.8	0.2	0	
23	**Mean**	74.25		75.3	75.9	0.6	0	
24	**Variance**	47.46053		73.6	76.9	3.3	0	
25	**Stderr 10**	2.178544		70.6	72.6	2.1	0	

Commands in Simulation

Cells B2–B21	Enter the 15-minute dissolution values from Table 5.9	
Cell B23	= AVERAGE(B2:B21)	Mean of combined A and B values
Cell B24	= VAR(B2:B21)	Variance of combined values
Cell B25	= SQRT(B24/10)	Standard error of mean for $n = 10$ values
Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables 2		Generate simulated means for two samples
Number of Random Numbers	30,000	Perform 30,000 simulations
Distribution	Normal	The means are Normally distributed.
Mean	74.25	Actual mean from combined $A + B$ sample
Standard Deviation	2.178544	Standard error of a mean for a sample ($n = 10$)
Random Seed	Enter 12345	Can enter a different seed if desired
Output Range	Enter D2	Simulated means in cells D2–E30001
Click OK		
Cell F2	= ABS(D2–E2)	Absolute value of difference between the two simulated means
Cells F3-F30001	Copy formula from F2	
Cell G2	= IF(F2 < 5.7, 0, 1)	If simulated means differ as much as what we saw for the actual sample then value is 1, otherwise it is 0
Cells G3-G30001	Copy formula from G2	
Cell H2	= AVERAGE(G2: G30001)	Proportion of times that results are as extreme as what we saw with actual sample (probability)

Our estimated probability for the difference between the formulation A and B means is 0.065, which is very similar to the result obtained with the t test, $p = 0.062$. We can further refine our estimate by repeating the simulation multiple times (using different seed values each time) and using the average probability. The results from a second simulation using a seed value 5555 gave a probability estimate of 0.062. The estimated probability obtained from averaging those from the two simulation estimates $p = 0.0635$.

The next example again uses the data in Table 5.9. Having observed a difference of 5.7 between the mean 15-minute dissolution values of formulations A and B, what is the 95% confidence interval for the true mean difference between the formulations? Using Monte Carlo simulation, the answer can be obtained in a very intuitive way. Assume that the means from the two samples are normally distributed, a reasonable assumption given the Central Limit theorem. The variance of the difference between two sample means is the sum of the two samples' variances divided by the number of observations (n) in the samples. It is assumed that there is a common variance (VAR) for the two formulations. The variance for the difference between the sample means is $(\text{VAR}/n_a + \text{VAR}/n_b)$, where n_a and n_b are the number of values

in the *A* and *B* samples, respectfully. As both samples consist of 10 values, the variance for the difference between means is equal to (2 × VAR/10). The standard error for the difference is equal to the square root of this value.

Applying the Central Limit Theorem, we can assume that the difference between the two means will be approximately normally distributed with $\mu = 5.7$ (our observed mean difference) and standard deviation equal to our estimated standard error. Simulating 30,000 mean differences, we can easily estimate the lower and upper 95% confidence limits. The 95% confidence limits encompass values between the 2.5th and 97.5th percentiles of the distribution describing the mean difference between the two samples (see chap. 5). These limits, for our Monte Carlo simulation of 3000 mean differences, are simply the 750th sorted value (2.5th percentile) and the 29,250th sorted value (97.5th percentile). The confidence interval obtained from the simulation, −0.34% to 11.79%, is comparable to that calculated using the *t*-distribution method, −0.32% to 11.72% (see chap. 5 for a description of how to apply the *t*-distribution method).

	A	B	C	D	E	F	G
1	**Form**	**Percent**	**Sim Diff**	**Sorted**	**Position**	**95% Cl Lo**	**95% Cl Hi**
2	***A***	68	2.14	−23.68		−0.34	11.79
3	***A***	84	−1.95	−23.68	**2.5%**		
4	***A***	81	4.83	−6.65	751		
5	***A***	85	3.79	−6.14			
6	***A***	75	7.35	−5.83	**97.5%**		
7	***A***	69	3.48	−5.83	29251		
8	***A***	80	5.44	−5.60			
9	***A***	76	7.76	−5.15			
10	***A***	79	2.14	−5.15			
11	***A***	74	4.97	−5.04			
12	***B***	74	6.55	−4.95			
13	***B***	71	5.83	−4.78			
14	***B***	79	7.84	−4.64			
15	***B***	63	2.10	−4.64			
16	***B***	80	4.55	−4.51			
17	***B***	61	5.04	−4.41			
18	***B***	69	3.15	−4.31			
19	***B***	72	7.07	−4.26			
20	***B***	80	0.07	−4.18			
21	***B***	65	4.57	−4.03			
22			5.96	−3.96			
23	**Mean *A***	77.1	11.29	−3.93			
24	**Mean *B***	71.4	1.74	−3.90			
25	**Difference**	5.7	10.80	−3.79			
26	**Variance**	47.46053	5.64	−3.59			
27	**Stderr Diff**	3.080926	5.70	−3.59			

Commands in Simulation

Cells B2–B21	Enter the formulation *A* and *B* dissolution values from Table 5.9	
Cell B23	= AVERAGE(B2:B11)	Mean of formulation *A* values
Cell B24	= AVERAGE(B12:B21)	Mean of formulation *B* values
Cell B25	= B23–B24	Difference between *A* and *B* means
Cell B26	= VAR(B2:B21)	Variance of combined *A* and *B* values
Cell B27	= SQRT(2∗B26/10)	Standard error for difference between means
Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables 1	Generate a simulated difference between means	
Number of Random Numbers 30,000	Generate 30,000 mean differences	
Distribution	Normal	Value is from the Normal dist.
Mean	5.7	Observed difference between *A* and *B* means
Standard Deviation	3.080926	Standard error of the difference
Random Seed	Enter 12345	Can enter a different seed if desired
Output Range	Enter C2	Simulated differences in cells C2–C30001
Click OK		
Select column C by clicking at the top of the column and then from Main Menu choose Edit → Copy		
Click at the top of column D and from Main Menu choose Edit → Paste		
Cell D1	Change label to Sorted	
Click on column D to select it		
Main Menu	Data → Sort Choose to sort only the selection in ascending order.	
Cell E4	= 1 + 0.25∗30,000	Column D cell with 2.5th percentile value
Cell E7	= 1 + 0.975∗30,000	Column D cell with 97.5th percentile value
Cell F2	= D751	Lower 95% confidence limit value
Cell G2	= D29251	Upper 95% confidence limit value

The next example comes from section 5.2.6. In two groups of patients, the incidences of headaches are evaluated to obtain a 95% confidence interval on the true difference in headache rates between the groups. In Group 1, there were 46 patients with headaches among the 196 patients, for a rate (proportion) of 0.2347. In the second group, 35 of the 212 patients experienced headaches, for a rate of 0.1651. The following Excel worksheet shows how to obtain the 95% confidence interval on the difference between the incidence proportions in the two groups by simulation.

We start by generating random values of the number of headaches in the two groups based on the binomial distribution. For Group 1, $N = 196$ and $p = 0.2347$. For Group 2, $N = 212$ and $p = 0.1651$. For each generated number of headaches for the two groups (simulated trial), we calculate the proportion of headaches "observed" in the groups and then find the difference between these proportions (Groups 1–Group 2). Thus, we generate 30,000 trials (see below). From the distribution of the Group l-to-2 differences in theses trials, we find the 2.5th and 97.5th percentiles as in the previous example. This is the 95% confidence interval.

	A	B	C	D	E	F	G	H	I	J
1	**S1**	**S2**	**Sim S1**	**Sim S2**	**p1**	**p2**	**Diff**	**Sorted**	**Position**	**95% Cl**
2	**46**	**35**	47	41	0.240	0.193	0.046	−0.079	**2.5%**	**Low**
3			51	40	0.260	0.189	0.072	−0.073	751	−0.009
4	**N1**	**N2**	33	30	0.168	0.142	0.027	−0.072		
5	**196**	**212**	39	27	0.199	0.127	0.072	−0.072	**97.5%**	**Hi**
6			50	38	0.255	0.179	0.076	−0.070	29251	0.147
7	**P1**	**P2**	40	30	0.204	0.142	0.063	−0.070		
8	**0.2347**	**0.1651**	48	47	0.245	0.222	0.023	−0.069		
9			55	33	0.281	0.156	0.125	−0.064		
10			51	40	0.260	0.189	0.072	−0.063		

Commands in Simulation

Cells A2 and B2	Enter the number of patients in Group 1 and 2, respectively, who experienced headaches	
Cells A5 and B5	Enter the total number of patients in Groups 1 and 2, respectively	
Cell A8	= A2/A5	Group 1 observed proportion of headaches
Cell B8	= B2/B5	Group 2 observed proportion of headaches
Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	1	Simulate Group 1 headache number
Number of Random	30,000	Generate 30,000 simulated trials
Numbers Distribution	Binomial	Simulated number is from binomial distribution
p Value	0.234694	Group 1 observed proportion of headaches
Total Number of Trials	196	Number of patients in Group 1
Random Seed	Enter 12345	Enter a different seed if desired
Output Range	Enter C2	Simulated headache numbers in C2−C30001
Click OK		
Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	1	Simulate Group 2 headache number
Number of Random	30,000	Generate 30,000 simulated trials
Numbers Distribution	Binomial	Simulated number is from binomial distribution
p Value	0.165094	Group 2 observed proportion of headaches
Total Number of Trials	212	Number of patients in Group 2

Random Seed	Enter 12345	Enter a different seed if desired
Output Range	Enter D2	Simulated headache numbers in D2–D30001
Click OK		
Cell E2	= C2/196	Proportion of Group 1 simulated headaches in trial 1
Cell F2	= D2/212	Proportion of Group 2 simulated headaches in trial 1
Cell G2	= E2–F2	Difference between group headache rates for 1st trial
Cells E3:G30001	Copy E2:G2	Calculates proportions and differences for other trials
Cells H2:H30001	Copy column G values, using the Paste Special, Values method	
	Click at top of column H to choose it (column is highlighted)	
Main Menu	Tools → Data → Sort	
	Click option to continue without expanding current selection	
	Click OK to sort the column with a header, in ascending order	
Cell I3	= 1 + 0.025∗30,000	Row with the 2.5th percentile difference
Cell I6	= 1 + 0.975∗30,000	Row with the 97.5th percentile difference
Cell J3	= H751	95% CI low limit = 2.5th percentile value
Cell J6	= H29251	95% CI hi limit = 97.5th percentile value

The 95% confidence interval limits based on the simulations, −0.009 to 0.147, are in close agreement with the limits of −0.008 to 0.148 calculated using normal approximation methods and with the limits −0.012 to 0.152 obtained using the more conservative continuity-corrected, normal approximation. Running the simulation a number of times, using different seed values each time, and then averaging the results should provide values closer to the exact limits.

One area where Monte Carlo methods are extremely useful is in determining the sizes of samples needed to obtain a desired power for a given statistical evaluation. A number of formulas are presented in chapter 6 that can be used for these calculations. In many situations, while the formulas are easily applied, their derivations are not so easily understood. Simulation provides an extremely intuitive approach in this area. As discussed in chapter 6, to determine the sample size needed for a given study we need to state the alpha level (e.g., 0.05), the beta level (e.g., 0.2 = power of 0.8), and a difference between treatments of a specified magnitude (usually a difference of practical significance). To determine the probability of obtaining a given outcome from a particular statistical test (e.g., the probability of getting a P value ≤ 0.05 in an independent group t test) we simply simulate a large number of random samples, calculate the statistic for each simulation, and then determine the proportion of times the statistic had the desired outcome. The more complicated the problem, the more intuitive and useful is the simulation method.

For sample size determination, we usually calculate the proportion of times we get a significant difference under the null hypothesis, which causes us to reject it in favor of the alternative hypothesis where the meaningful difference is specified. The following example uses both this approach and a modification needed when we want to test for noninferiority (or equivalence) rather than testing for a difference.

In this example, we want to conduct a clinical trial on a new drug developed to treat a certain disease. Preliminary animal studies indicate that the new drug will be at least as effective as the current treatment for the disease and is likely to have fewer serious side effects. The FDA has indicated that it wants to see a placebo-controlled, noninferiority trial. This trial

will compare the new Drug, A, the current treatment B, and placebo, in the treatment of subjects with the disease. The primary efficacy measure will be the proportion of subjects who show improvement. We intend to show that the new drug is at least as effective as (noninferior to) the current Drug B. To demonstrate noninferiority we must construct a 95% confidence interval for the difference between the Drug A and Drug B proportions and then show that this difference is no worse than 20% (i.e., Drug A is no more inferior to Drug B than 20%). In addition to showing noninferiority, we must simultaneously demonstrate that the clinical trial had adequate sensitivity to detect true differences in efficacy had they existed. This is established by showing that both Drugs A and B have superior efficacy to that of the placebo.

From prior experience, we know that 25% of patients left untreated will improve spontaneously (placebo success proportion is expected to be 0.25) and improvement is seen in 45% of those treated with Drug B (success rate for B is expected to be 0.45). We believe that the new drug will be successful in treating at least 50% of the patients (cure rate for Drug A is conservatively set at 0.50).

The statistical evaluation comparing Drug A to Drug B involves the construction of the 95%, continuity-corrected, confidence interval on the difference between their success proportions. If the lower limit of this confidence interval is greater than -0.20, then noninferiority of A to B will be established. Note that while our interest is only with the lower confidence interval limit (i.e., it is one-sided), the FDA usually requires the use of the more conservative, two-sided, confidence interval (critical Z value of 1.96 is used instead of 1.645). Had our intention been to show therapeutic equivalence of Drug A to Drug B, rather than noninferiority, then we would need to show that the entire confidence interval falls within the equivalence interval -0.20 to $+0.20$. For the trial to be successful, we must also show that the two-sided, continuity-corrected, Z tests on the differences between the success proportions for Drug A compared to placebo and for Drug B compared to placebo show statistical superiority for the active treatments (i.e., differences > 0 and $p < 0.05$). The following equations will be used (see chap. 5):

$$95\%\,\mathrm{CI} = (p_a - p_b) \pm \left[1.96 * \left(\frac{p_a * q_a}{n_a} + \frac{p_b * q_b}{n_b}\right)^{1/2} + 0.5 * \left(\frac{1}{n_a} + \frac{1}{n_b}\right)\right]$$

$$Z\ \text{test}\,1 = \frac{\left[(p_a - p_\mathrm{p}) - 0.5 * \left((1/n_a) + (1/n_\mathrm{p})\right)\right]}{\left[(p_0 * q_0)\left((1/n_a) + (1/n_\mathrm{p})\right)\right]^{1/2}}$$

$$Z\ \text{test}\,2 = \frac{\left[(p_b - p_\mathrm{p}) - 0.5 * \left((1/n_b) + (1/n_\mathrm{p})\right)\right]}{\left[(p_0 * q_0)\left((1/n_b) + (1/n_\mathrm{p})\right)\right]^{1/2}}$$

where p_a is the observed success rate for Drug A, $q_a = 1 - p_a$ failure rate for Drug A, p_b the observed success rate for Drug B, $q_b = 1 - p_b$ failure rate for Drug B, p_p the observed success rate for placebo, and $q_\mathrm{p} = 1 - p_\mathrm{p}$ failure rate for placebo; n_y the number of patients receiving treatment Y; Y $= A$, B, or placebo, $p_0 = (n_\mathrm{y} * p_\mathrm{y} + n_\mathrm{p} * p_\mathrm{p})/(n_\mathrm{y} + n_\mathrm{p})$, pooled success rate; Y $=$ A for Z Test 1, B for Test 2; $q_0 = 1 - p_0$ pooled failure rate for Z test.

We will determine our sample size by trial and error. First we specify a given sample size. We assume that the two active products' success rates actually differ by 5% (0.50 vs. 0.45) and that the placebo success rate is that known to occur in untreated patients (0.25). We then randomly generate (simulate) success/failure results for treating patients with Drug A, Drug B, and placebo. From these results, we calculate the proportions of patients with success in each treatment group and calculate the above statistics. Our probability of trial success (power) is the proportion of times that our simulated samples meet the criteria for noninferiority and superiority.

Our initial evaluation uses a 2:2:1 randomization (A:B:placebo) in about 350 patients (a number consistent with our initial budget allocation). We propose to use 340 patients, 136 in each active treatment group and half that number, 68, in the placebo group. We want to estimate the probability that our trial will show both noninferiority of Drug A compared to Drug B,

and superiority of both *A* and *B* over placebo. We determine this easily using Monte Carlo simulation.

As shown in the following Excel worksheet, we simulate the results for 30,000 trials each involving the treatment of 136 patients for Drugs *A* and *B*, and 68 patients for placebo. The number of successfully treated patients for Drug *A* is placed in column A, for Drug *B* in column B, and for placebo in column C. Columns D, E, and F contain the

	A	B	C	D	E	F	G	H	I	J	K
1	**Drug *A***	**Drug *B***	**Placebo**	$\mathbf{n_a}$	$\mathbf{n_b}$	$\mathbf{n_p}$	$\mathbf{p_a}$	$\mathbf{p_b}$	$\mathbf{p_p}$	***p*01**	***p*02**
2	66	61	10	136	136	68	0.485	0.449	0.147	0.373	0.348
3	65	57	13	136	136	68	0.478	0.419	0.191	0.382	0.343
4	63	59	17	136	136	68	0.463	0.434	0.250	0.392	0.373
5	66	62	18	136	136	68	0.485	0.456	0.265	0.412	0.392
6	66	57	14	136	136	68	0.485	0.419	0.206	0.392	0.348
7	63	60	16	136	136	68	0.463	0.441	0.235	0.387	0.373
8	67	63	14	136	136	68	0.493	0.463	0.206	0.397	0.377
9	67	66	18	136	136	68	0.493	0.485	0.265	0.417	0.412
10	73	61	20	136	136	68	0.537	0.449	0.294	0.456	0.397
11	73	70	19	136	136	68	0.537	0.515	0.279	0.451	0.436
12	82	61	20	136	136	68	0.603	0.449	0.294	0.500	0.397

total number of treated patients (136, 136, and 68) for Drug *A*, Drug *B*, and placebo, respectively. The calculated success proportions for Drug *A*, Drug *B*, and placebo are placed in columns G, H, and I, respectively. The pooled success proportions for the Drug *A* and placebo comparisons and for the Drug *B* and placebo comparisons, under the null hypothesis of no difference between treatment success proportions, are placed in columns J and K, respectively. A portion of the worksheet with these results is shown above along with the following commands used to obtain them.

Commands in Simulation

Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	1	Simulate number of successes for Drug *A*
Number of Random	30,000	Generate 30,000 simulated trials
Numbers Distribution	Binomial	Numbers come from binomial distribution
p Value	0.50	Expected Drug *A* success proportion
Total Number of Trials	136	Number of patients in treatment group
Random Seed	Enter 1234	Enter a different seed if desired
Output Range	Enter A2	Drug *A* number of successes in A2–A30001
Click OK		
Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	1	Simulate number of successes for Drug B
Number of Random	30,000	Generate 30,000 simulated trials
Numbers Distribution	Binomial	Numbers come from binomial distribution

p Value	0.45	Expected Drug B success proportion
Total Number of Trials	136	Number of patients in treatment group
Random Seed	Enter 2341	Enter a different seed if desired
Output Range	Enter B2	Drug B number of successes in B2−B30001
Click OK		
Main Menu	Tools→ Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	1	Simulate number of successes for placebo
Number of Random	30,000	Generate 30,000 simulated trials
Numbers Distribution	Binomial	Numbers come from binomial distribution
p Value	0.25	Expected placebo success proportion
Total Number of Trials	68	Number of patients in treatment group
Random Seed	Enter 3412	Enter a different seed if desired
Output Range	Enter C2	Placebo number of successes in C2–C30001
Click OK		
Cells D2, E2, F2	Enter number of patients in treatment groups A, B and placebo	
Cell G2	= A2/D2	Proportion of successes for Drug A
Cell H2	= B2/E2	Proportion of successes for Drug B
Cell I2	= C2/F2	Proportion of successes for placebo
Cell J2	= (A2 + C2)/(D2 + F2)	Pooled proportion for A and placebo
Cell K2	= (B2 + C2)/(E2 + F2)	Pooled proportion for B and placebo
Cells D3: K30001	Copy formulas from cells D2 through K2	

Next we calculate the continuity correction, $0.5 \times (1/n_a + 1/n_b)$, for the noninferiority calculation and place it in column L. We do the same for the superiority comparisons of Drug A to placebo and Drug B to placebo, and place these values in columns M and N. The 90% confidence interval lower limit for each trial (row) is calculated and placed in column O. The Z test value for the comparison of Drug A to placebo is calculated and placed in column P and that for Drug B to placebo is placed in column Q. Flags in columns R, S, and T are set to 1 if we pass the noninferiority test, the A-to-placebo superiority test, and the B-to-placebo superiority test, respectively. If all three tests are passed, then a 1 is placed in column U indicating that the trial was successful. A failed test is designated by a flag value of 0 placed in its respective column.

	L	M	N	O	P	Q	R	S	T	U
1	**CCAB**	**CC1**	**CC2**	**95%CI LO**	**Z test1**	**Z test2**	**Flag1**	**Flag2**	**Flag3**	**Flag All**
2	0.007	0.011	0.011	−0.089	4.557	4.105	1	1	1	1
3	0.007	0.011	0.011	−0.067	3.820	3.076	1	1	1	1
4	0.007	0.011	0.011	−0.096	2.789	2.406	1	1	1	1
5	0.007	0.011	0.011	−0.097	2.867	2.484	1	1	1	1
6	0.007	0.011	0.011	−0.059	3.701	2.858	1	1	1	1
7	0.007	0.011	0.011	−0.104	2.998	2.714	1	1	1	1
8	0.007	0.011	0.011	−0.097	3.794	3.421	1	1	1	1

9	0.007	0.011	0.011	−0.119	2.962	2.867	1	1	1	1
10	0.007	0.011	0.011	−0.037	3.131	1.973	1	1	1	1
11	0.007	0.011	0.011	−0.104	3.333	3.045	1	1	1	1
12	0.007	0.011	0.011	0.030	4.010	1.973	1	1	1	1
13	0.007	0.011	0.011	−0.082	3.888	3.328	1	1	1	1
14	0.007	0.011	0.011	−0.029	3.397	2.161	1	1	1	1
15	0.007	0.011	0.011	−0.163	3.491	3.960	1	1	1	1
16	0.007	0.011	0.011	−0.082	3.169	2.598	1	1	1	1
17	0.007	0.011	0.011	−0.015	2.139	0.664	1	1	0	0
18	0.007	0.011	0.011	−0.148	3.678	3.960	1	1	1	1

Commands in Simulation (Continued)

Cell L2	= 0.5∗(1/D2 + 1/E2)	Continuity correction (*A* vs. *B*)
Cell M2	= 0.5∗(1/D2 + 1/F2)	Continuity correction (*A* vs. placebo)
Cell N2	= 0.5∗(l/E2 + 1/F2)	Continuity correction (*B* vs. placebo)
Cell O2	= (G2–H2) – ((1.96∗SQRT(G2∗(1–G2)/D2 + H2∗(1–H2)/E2) + L2))	
Cell P2	= (G2–I2 – M2)/SQRT(J2∗(1–J2)∗(1/D2 + 1/F2))	
Cell Q2	= (H2–I2 – N2)/SQRT(K2∗(1–K2)∗(1/E2 + 1/F2))	
Cell R2	= IF(O2>–0.2,1,0)	Flag = 1 if 95% CI is above –0.20
Cell S2	= IF(P2>1.96,1,0)	Flag = 1 if *A* versus placebo Z test is significant
Cell T2	= IF(Q2> 1.96,1,0)	Flag = 1 if *B* versus placebo Z test is significant
Cell U2	= R2∗S2∗T2	Flag = 1 if all three tests pass
Cells L3:U30001	Copy formulas from cells L2 through U2	

Our probability (power) of showing noninferiority, superiority, or passing all three required tests is simply the average of the 0/1 entries in the corresponding flag column, the proportion of simulated trials in which we observed a successful (1) outcome.

	V	W	X	Y
1	***p* (noninf)**	***p* (super*A*)**	***p* (super*B*)**	***p* (trial)**
2	0.9830	0.9206	0.7668	0.7353

Final Commands in Simulation

Cell V2 = Average(R2:R30001)	Proportion where *A* was noninferior to *B*
Cell W2 = Average(S2:S30001)	Proportion where *A* was superior to placebo
Cell X2 = Average(T2:T30001)	Proportion where *B* was superior to placebo
Cell Y2 = Average(U2:U30001)	Proportion where *A* was noninferior to *B* and both *A* and *B* were each superior to placebo (overall probability of success)

The probability of showing noninferiority, 0.983, and the probability of showing the superiority of Drug *A* over placebo, 0.921, are both high with assumed proportions of 0.5 and 0.25 for Drug *A* and placebo, respectively. The probabilities of showing superiority of Drug *B* (assumed proportion 0.45) over placebo, 0.767, and for the overall success of the trial, 0.735, are unacceptably low.

We would like to know if there is a way to increase the probability of overall success without increasing our costs (i.e., patient numbers). We decide to explore the question by looking to a different randomization scheme. Perhaps a 1:1:1 randomization would increase the probability of trial success. We will evaluate using an equivalent number of patients in each treatment group to see if this improves our expected outcome. By setting our sample sizes to 110 patients in each treatment (330 total) and performing the simulation and calculations again, we find that the probability of showing noninferiority decreases slightly to 0.947, the probability of showing superiority of Drug *A* over placebo increases slightly to 0.963, and the probability of showing Drug *B* to be superior to placebo significantly increases to 0.849. The overall effect is that the probability of a successful trial is now increased to 0.789. By adding a few more patients to each treatment group and using the 1:1:1 randomization scheme, we can bring the overall probability of trial success to 0.80, a typical level of power used in designing a clinical trial. This is accomplished by using essentially the same number of subjects that would provide only a 0.74 probability of trial success with the 2:2:1 randomization scheme. Using computer simulations, these types of what-if evaluations are easy to conduct and to understand.

Another important application of Monte Carlo simulation is estimating the properties of a certain statistic when there is no known formula for doing so. For example, we might want to determine the probability distribution for the difference between two sample medians when the samples are drawn from similar, or dissimilar, statistical distributions. When there are no standard formulas to evaluate the distributional properties of a complex or unusual statistic, computer simulation is often the only tool available.

14.2 BOOTSTRAPPING

Bootstrapping (sometimes called resampling) encompasses a group of computer simulation methods in which samples are repeatedly drawn not from some hypothesized statistical distribution, but from the set of values that come from an actual sample obtained from some real population. These methods typically assume only that the sample was randomly selected from the population, thereby ensuring that it is likely to be representative of the population from which it was drawn. The theory behind Bootstrap methods proposes that the probabilistic information contained in the sample is reflective of corresponding information contained in the actual population. This same assumption is also required for most standard inferential methods. The primary difference between bootstrapping and standard inferential methods is how we use this information contained in the sample.

Standard inferential methods rely on our knowledge of the distribution of a statistic or parameter (e.g., mean, standard deviation, etc.) that we calculate from a sample collected from a population with some assumed statistical distribution. As an example, it is known that the average value calculated from a sample whose underlying population is assumed to be normally distributed with mean μ and variance σ^2 will follow a Normal distribution with mean μ and variance σ^2/n, regardless of the size of the sample, n. When neither μ nor σ is known, we estimate these parameters from the sample average and its standard deviation. A 95% confidence interval on μ is calculated using the standard equation: average $\pm t_{\alpha/2,n-1} \times$ SE, where SE is the sample standard deviation divided by the square root of n. The value $t_{\alpha/2,\,n-1}$ is obtained from student's t distribution. In the standard method, using statistics calculated from the sample (e.g., mean and standard deviation) we infer back to the values of the unknown parameters (e.g., μ and σ) of the underlying population.

In bootstrapping, we make no assumptions about the statistical distribution of the population from which the sample was collected or about the distributional properties of the sample itself. Instead, we treat the sample as if it was the population and repetitively take samples (resamples) from it using computer simulation. The distribution of statistics calculated from these computer-generated samples theoretically mimics the distribution in the

population. Using the frequency distribution of the statistic in the computer-generated samples, we make inferences about the corresponding distribution in the underlying population. One of the simplest bootstrapping methods will be used to provide a brief introduction to these powerful, computer-intensive simulation methods. The method is known as the percentile method and is one of the most intuitive ones available.

Table 5.1 shows the assay results for 10 randomly selected tablets. The average value for these results is 103.0 mg and the standard deviation is 2.218. If we assume that the sample comes from a population that is normally distributed, or that based on the Central Limit Theorem the sample average is normally distributed, then we can calculate a 95% confidence interval on μ. This interval is determined to be 101.4 to 104.6, as shown in chapter 5. If we do not want to make distributional assumptions about the underlying population or about the sample, then a Bootstrap method can be used to obtain a confidence interval on μ (the population average value) as shown below.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Mean	Stdev	Number:	1	2	3	4	5	6	7	8	9	10
2	103.0	2.218	Sample:	101.8	102.6	99.8	104.9	103.8	104.5	100.7	106.3	100.6	105.0

In row 1, columns D to M, we enter the numbers 1 to 10 to identify each observed assay value in the sample. The observed sample values are entered into row 2, immediately below their corresponding identification numbers. The mean (cell A2) and standard deviation (cell B2) of the values are calculated using the Excel formulas = AVERAGE(D2:M2) and = STDEV(D2:M2). We now go to column Y, reserving columns N to W for use later. We next generate 10 random numbers from the Uniform distribution for each of our 3001 simulated trials (rows) and place these numbers in columns Y to AH. The numbers will be rounded to integer values, and placed into columns N to W, to be used in obtaining our Bootstrap sample for each simulated trial.

	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1										
2										
3	3.08	6.26	8.09	7.08	2.60	7.43	8.55	2.49	8.99	7.43
4	5.11	5.07	6.62	6.97	3.62	4.87	7.03	3.81	8.16	2.08
5	3.81	3.29	4.05	5.20	1.72	3.88	6.88	1.99	1.36	8.60

Commands in Simulation

Main Menu	Tools → Data Analysis → Random Number Generator	
Dialog Box		
Number of Variables	10	Simulate 10 values for each trial
Number of Random Numbers	3001	Perform the simulation for 3001 trials
Distribution	Uniform	Values come from the Uniform distribution
Parameters Between	1,10	Generate equally probable values between 1 and 10
Random Seed	12345	Enter a different seed if desired
Output Range	Y3:AH3003	Place values in cells Y3 through AH3003
Click OK		

To convert the simulated Uniform distribution values into integer, index numbers, enter the following equation into cell N3: = ROUND(Y3,0). Next, copy this formula to all cells within the range N3: W3003.

	N	O	P	Q	R	S	T	U	V	W
1	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
2										
3	3	6	8	7	3	7	9	2	9	7
4	5	5	7	7	4	5	7	4	8	2
5	4	3	4	5	2	4	7	2	1	9

Using one of Excel's table lookup functions (HLOOKUP), we select values (resample) from the original sample whose assigned numbers in row 1 of columns D to M match the corresponding index values found in cells N3–W3003. In this way, each of the 3001 rows, representing a trial, contains a computer-generated sample of size 10 drawn from the original sample. As each original value can appear more than once in the Bootstrap sample, the method involves sampling with replacement. For each of the 3001 Bootstrap samples (rows 3–3003), we calculate the mean and standard deviation for its 10 values in columns D to M. These are the Bootstrap sample means and standard deviations whose frequency distributions will be used to make inferences to the characteristics of the underlying population from which our original sample was obtained.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Mean	Stdev	Number	1	2	3	4	5	6	7	8	9	10
2	103.0	2.2181	Sample	101.8	102.6	99.8	104.9	103.8	104.5	100.7	106.3	100.6	105.0
3	101.6	2.17		99.8	104.5	106.3	100.7	99.8	100.7	100.6	102.6	100.6	100.7
4	103.2	1.99		103.8	103.8	100.7	100.7	104.9	103.8	100.7	104.9	106.3	102.6
5	102.7	1.93		104.9	99.8	104.9	103.8	102.6	104.9	100.7	102.6	101.8	100.6

Commands in Simulation

Cell D3	= HLOOKUP (N3,$D$1:$M$2,2)	From the sample values in row 2, section D1 to M2, select the value whose number in row 1 matches the random index number in N3
Cells E3:M3	Copy formula from cell D3	Generate first bootstrap sample
Cells D4:M3003	Copy formulas from cells D3:M3	Generate the remaining 3000 samples
Cells A3:A3003	Copy formula from cell A2	Calculate the Bootstrap samples' means
Cells B3:B3003	Copy formula from cell B2	Calculate the samples' standard deviations

Now we will estimate a 95% confidence interval on μ and test the hypothesis that μ is less than 102. These results will be compared to those obtained by standard formulas that assume that the sample is normally distributed. The procedure that we use in the percentile method is similar to that shown in previous examples, but here we apply them to the Bootstrap samples rather than samples simulated from some underlying assumed statistical distribution.

We start by opening a new worksheet and transferring the mean (sample average) values from our existing column A to column A in the new worksheet. Wanting to transfer only the values and not the formulas, we use the Edit → Copy → Edit → Paste Special → Values sequence in Excel. We delete the first entry in the new column A, as this is the mean for the original sample, not for a Bootstrap sample. All remaining values in the column shift upwards. The Bootstrap estimate of μ is the average of the Bootstrap sample means. The 95% confidence interval lower limit is the 2.5th percentile sorted mean and the upper limit is the 97.5th percentile sorted mean. The probability that $\mu < 102$ is simply the frequency that we have a Bootstrap mean value that is less than 102. The analyses and the commands to conduct the analyses are shown in the following.

	A	B	C	D
1	**Mean**	**<102**		
2	100.78	1	**Bootstrap**	
3	100.78	1	Mean	103.0
4	100.85	1	2.5% observation	76
5	101.00	1	97.5% observation	2927
6	101.00	1	95% Cl Lower	101.7
7	101.03	1	95% Cl Upper	104.2
8	101.10	1	Prob(mu<102)	0.076
9	101.10	1		
10	101.14	1		
11	101.14	1	**Normality Assumed**	
12	101.19	1	95% Cl Lower	101.4
13	101.29	1	95% Cl Upper	104.6
14	101.29	1	Z < 102	−1.426
15	101.29	1	Prob(μ<102)	0.077

Commands in Simulation

New Column A	Data → Sort	Sort the Bootstrap means in ascending order
Cell B2	= IF(A2 < 102,1,0)	1 if bootstrap mean is < 102, 0 otherwise
Cells B3:B3002	Copy formula from B2	
Cell D3	= AVERAGE(A2:A3002)	Bootstrap estimate of μ
Cell D4	= 1 + 0.025∗3001	Row number for 2.5th percentile mean value
Cell D5	= 1 + 0.975∗3001	Row number for 97.5th percentile value
Cell D6	= A76	2.5th percentile value is Bootstrap CI lower limit
Cell D7	= A2927	97.5th percentile value is Bootstrap CI upper limit
Cell D8	= AVERAGE(B2:B3002)	Proportion of the means that are < 102
Cells D12 and D13	95% confidence limits	Obtain from text assuming Normal distribution
Cell D14	= (102 − 103)/(2.2181/SQRT(10))	Z value for standard test of $\mu <$ 102
Cell D15	= Normsdist(D14)	Probability from Standard Normal distribution

A similar evaluation for standard deviation can be conducted from the Bootstrap sample values. Copy the standard deviation values from column B of our original worksheet into column A of a new worksheet. Delete the standard deviation value for the original sample, leaving only those for the Bootstrap samples. As with the analysis of the means, we sort the column of values. The average of the Bootstrap values is our estimate of σ. The 2.5th and 97.5th percentile values are our lower and upper 95% confidence interval limits.

For the standard method, we rely on the Chi-square distribution (see chap. 5) as the assumed statistical distribution of the variance. The square root of the variance, the standard deviation of the original sample values, is the estimate of σ. By using the 2.5th percentile and 97.5th percentile critical Chi-square values, we can construct a 95% confidence interval on σ using standard formulas.

The average standard deviation value of our Bootstrap samples is 2.122. The 2.5th percentile and the 97.5th percentile standard deviations are the 76th and 2927th sorted values. The Bootstrap 95% confidence interval limits are 1.4 and 2.7.

The 95% confidence interval limits based on the Chi-square distribution are derived from the distribution's critical values of 2.70 (0.025 probability level, 9 d.f.) and 19.02 (0.975 probability level, 9 d.f.). The lower limit, 1.5, is calculated as SQRT($(2.218^2 \times 9)/19.02$) and the upper limit, 4.0, is calculated as SQRT($(2.218^2 \times 9)/2.70$).

It is notable that while the lower limits from both the Bootstrap method and the formula-based method are quite similar, those for the upper limit are not. This may be due to the small size of the original sample, resulting in a biased bootstrap estimate of variability. If this was the case, then taking a second sample and combining it with the original sample, then repeating the Bootstrap process, might improve the estimate. It is also possible that the actual underlying statistical distribution for the sample variance is not that of the assumed Chi-square distribution. In this case, the Bootstrap confidence interval may be closer to reality than that obtained by the standard formulas. Only additional actual sampling would help us evaluate the cause of the discrepancy between the two estimates.

Both Monte Carlo simulation and bootstrapping methods are powerful tools for solving problems. Monte Carlo simulation, carrying out repeated computer-simulated experiments based on simple statistical principles, is a process that has an intuitive appeal to many scientists.

	A	C	D
1	**Stdev**		
2	0.703	**Bootstrap**	
3	0.811		
4	0.891	stdev	2.122
5	0.982	2.5% obs	76
6	1.001	97.5% obs	2927
7	1.001	95% Cl Low	1.4
8	1.020	95% Cl Hi	2.7
9	1.020		
10	1.105		
11	1.116	**Chi-Sq on sample**	
12	1.116		
13	1.128	**95% Cl Low**	1.5
14	1.132	**95% Cl Hi**	4.0
15	1.132		
16	1.160		
17	1.171		

While less intuitive in its theoretical underpinnings, Bootstrapping provides a simple nonparametric method for solving problems when we are unable to make assumptions about the underlying statistical properties that govern the process of interest.

While the examples presented have relied upon the computing power of Microsoft Excel, there are other packages that may provide more accessible simulation and bootstrapping capabilities. The author is familiar with two such packages provided by the company Resampling Stats, Inc. (*www.resample.com*). One is marketed as an Excel Add-in [2] that enhances the built-in simulation capabilities in Excel and provides a considerably easier way to perform bootstrapping in Excel. The second, Resampling Stats [3,4] is a self-contained simulation and bootstrapping package with extremely intuitive commands and easy to use programming wizard interface. The reader who wishes to further pursue simulation methods would be well advised to consider one of these computer packages.

REFERENCES

1. Mooney CZ, Duval RD. Bootstrapping: A Nonparametric Approach to Statistical Inference. Newbury Park, CA: Sage University Paper Series on Quantitative Applications in the Social Sciences, Sage Publications, Inc., 1993.
2. Blank S, Seiter C, Bruce P. Resampling Stats in Excel. Arlington, VA: Resampling Stats, Inc., 1999.
3. Simon JL. Resampling: The New Statistics, 2nd ed. Arlington, VA: Resampling Stats, Inc., 1998.
4. Simon JL. Resampling Stats: User's Guide. Arlington, VA: Resampling Stats, Inc., 1999.