15 Nonparametric Methods

Nonparametric statistics, also known as distribution-free statistics, may be applicable when the nature of the distributions are unknown, and we are not willing to accept the assumptions necessary for the application of the usual statistical procedures. For most of the statistical tests described in this book, we have assumed that data are normally distributed. This assumption, although never exactly realized, is bolstered by the central limit theorem (sect. 3.4.2) when we are testing hypotheses concerning the means of distributions. However, occasions arise in which data are clearly too far from normal to accept the assumption of normality. The data may deviate so much from that expected for a normal distribution that to assume normality, even when dealing with means, would be incorrect. In these situations, a data transformation may be used, chapter 10, or nonparametric methods may be applied for statistical tests. As we shall see, many of the nonparametric tests are easy to compute, and can be used for a quick preliminary approximation of the level of significance when parametric tests may be more appropriate. Although some people believe that any kind of data, no matter what the distribution, can be correctly analyzed using nonparametric methods, a kind of panacea, this is not true. Many if not most nonparametric methods require that the distributions be continuous and symmetrical, and that data be independent, for example. These are among the assumptions underlying parametric analyses, as exemplified by the normal t, and F tests.

15.1 DATA CHARACTERISTICS AND AN INTRODUCTION TO NONPARAMETRIC PROCEDURES

Before proceeding, a review of the different kinds of data that are usually encountered in scientific experiments will be useful for the understanding of the applications of nonparametric methods.

1. Perhaps the most elementary kinds of data are *categorical* or *attribute* measurements. These are also known as *nominal* observations (i.e., the observation is given a *name*). Thus, a person is observed to be a "male" or a "female" or "black," "white," or "yellow." Some other examples are given in Table 15.1. The assignment of a number to such nominal data may be useful to differentiate the categories, perhaps for computer usage. However, actual values, a number assigned to these categories where the numbers have meaning in terms of rank, would not make sense. For example, we could assign the number 1 to a male and 2 to a female, but this does not imply that a female is *larger* (or, for that matter, smaller) than a male. Data that comprise two classes and consist of such attribute measurements may be analyzed using the binomial distribution. As discussed in chapter 5, Chi-square tests may be used to test the significance of differences of the proportion of attributes in comparative groups if the sample size and incidences are sufficiently large. These kinds of data are usually presented in the form of contingency tables, such as the 2×2 table for proportions discussed in chapter 5.
2. The next, perhaps more "sophisticated" level of measurement involves data that can be *ranked* in order of magnitude. That is, we can say that one measurement is equal to, less than, or greater than another. These kinds of ordered data are known as *ordinal* measurements. Continuous variables are ordinal measurements according to this definition, but here, we usually think of ordinal data as arising from some arbitrary scale, as constructed for rating scales. For example, patients receiving antidepressant medication, may be rated according to attributes such as "sociability." A high score will be assigned to a patient performing well on this criterion. If the patient shows characteristics of "withdrawal," a low score will result. Intermediary scores reflect various degrees of response. These are ordinal measurements. A

Table 15.1 Examples of Nominal Data

Products categorized as acceptable and unacceptable in quality control
Side effects in a clinical study
Males and females in a clinical study
Various descriptions of "feel" of an ointment preparation, or taste of a product (tart, biting, sharp, etc.)
Concomitant diseases or medicaments in a clinical study

patient with a score of zero after one week of medication, and a score of 3 after two weeks of medication can be said to have improved during the period between one and two weeks of treatment. A score of 3 is *better* than a score of zero. Some examples of this kind of data are shown in Table 15.2. Many nonparametric tests are based on *ranking* data. Certainly, data derived from a continuous distribution, such as the normal distribution, can be ranked in order of magnitude. (Ordinal data, by definition, can be ranked.) The nonparametric tests that will be discussed here, which use ranks for the analysis, require that the data have a continuous distribution. One might question the validity of nonparametric tests using data derived from an arbitrary ordinal rating scale such as that described above. If we understand (or assume) that the rating scale has an underlying continuity, the discreteness and arbitrary nature of the scale can be considered acceptable for nonparametric tests. The condition of the "depressed" patient is a continuum. The condition can vary from one extreme to another with infinitely small gradations, in theory. It is not possible practically to measure the subjective condition with its infinite subtleties, and therefore we substitute an ordered scale that approximates the condition of the patient. Controversy exists regarding the analysis of this kind of data. Some people believe that data derived from rating scales, as described above, should not be analyzed by parametric methods such as the t test. One reason for this position is that the intervals in these rating scales are not equal in terms of the degree of response; that is, the scores do not represent an equi-interval scale. In fact, the scale points do not precisely correspond to the description of the condition. The points are usually arbitrarily defined. Thus, there is not an exact correspondence of the numbers on the rating scale to the patients' conditions, as defined by an arbitrary description based on an assumed underlying continuous distribution (Fig. 15.1). For example, if a score of 3 represents "marked improvement" in sociability, 2 represents "moderate improvement," and 1 represents "no improvement," one usually cannot say that the difference between scores of 3 and 2 is equal in magnitude to the difference of 2 and 1. Yet the data analysis of such scores usually treats a difference between 3 and 2 as equivalent to a difference between 2 and 1. Perhaps, if the psychological aspects of depression were known to a sufficient extent, and the observer could discern subtle differences, the scoring system could be shown to be better represented by 3, 2.5, and 0.8 for the conditions corresponding to "marked improvement," "moderate improvement," and "no improvement," respectively.

Although we can and do analyze data from a rating scale using nonparametric methods (as presented below), the typical parametric methods (ANOVA, t tests) are also commonly applied to such data. The use of parametric methods to analyze rating scale data is considered to be acceptable by many statisticians, including members of the FDA. Snedecor and Cochran discuss the analysis of this kind of data using a modified t test [1].

3. When comparing ages using a "ranking" scale, one person may be said to be older than another without regard to the magnitude of the difference in age. One can also specify the numerical differences with such data (e.g., one person is *two* years older than another). This

Table 15.2 Examples of Ordinal Data

Rating scales for sensory attributes (degree of liking)
Degree of effectiveness of therapeutic agent (pain relief, joint swelling, etc.)
Dichotomization of a continuous variable (underweight and overweight)
Number of anginal attacks in one week
Number of ulcers in skin-diseased patient

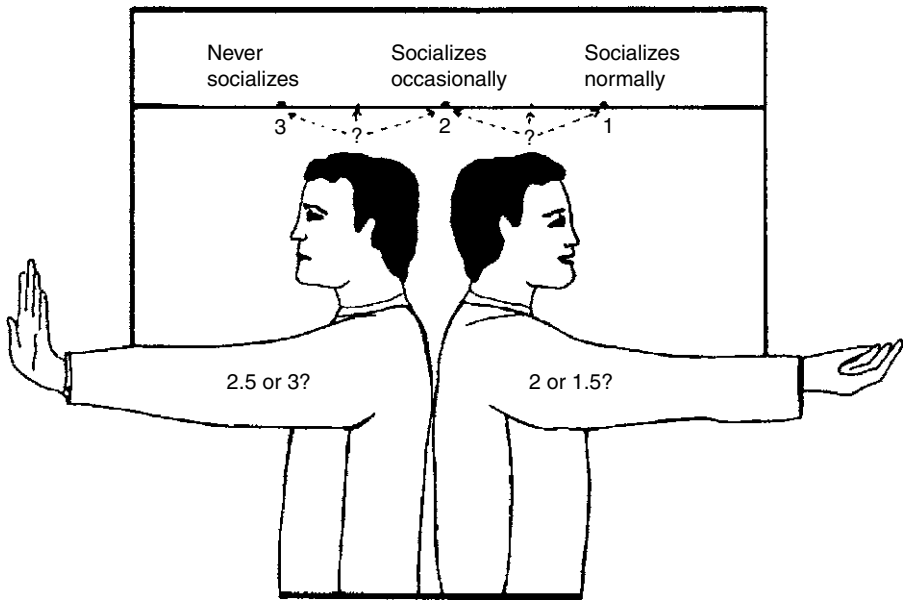


Figure 15.1 Problems with correspondence of a number and a subjective condition.

is an example of numerical data, often encountered in scientific experiments, where the distances between the values representing experimental outcomes have physical meaning. These data have a precise, better-defined meaning than data that are only ranked. Such data are often categorized as *interval* or *ratio* scaled data, depending on whether or not a true "zero point" exists. Age, weight, and concentration are examples of ratio scales. A person who weights 200 pounds is twice as heavy as one who weighs 100 pounds. Temperature does not have a true zero point (according to the concept above) and is an example of an interval scale. A temperature designated as "zero" is an arbitrary position on the scale and does not represent the lack of temperature. We cannot say that 40°C is twice as hot as 20°C. Ratio and interval-type data are the kinds of numbers that usually are subjected to the typical parametric tests. If these data are not normally distributed, they may be appropriately analyzed using nonparametric methods. One should understand that, in general, nonparametric tests can be applied to most of the data that we usually encounter, including that from continuous data distributions. Hence, data that are normally distributed may also be analyzed using these methods. A disadvantage of using nonparametric methods rather than the usual analyses for normally distributed data is that nonparametric methods are less sensitive (i.e., they are less powerful). Nevertheless, some nonparametric methods are surprisingly sensitive and are able to differentiate treatments that are normally distributed with efficiency almost equal to that of the parametric tests.

Nonparametric tests are most effectively used for data that consist of only classified (nominal) variables or ranked variables that are considered to have an underlying continuous distribution. Data derived from continuous distributions are particularly amenable to nonparametric methods when the distributions deviate greatly from normality. The reader should be aware that many nonparametric tests assume a symmetric distribution and equality of variance in the comparative groups. A marked disadvantage of the simpler nonparametric techniques is the lack of flexibility of the design and analysis. Elementary designs may be readily analyzed using nonparametric methods, but more complex designs in which interactions and other ANOVA components are present cannot be simply analyzed with these techniques, particularly when sample sizes are small.

Most of the nonparametric methods for data that are not categorical use *ranking* procedures. The observations in the various treatment groups are ranked according to specific procedures, and the ranks that replace the raw data are then analyzed. These analyses use

simpler statistical computations than the corresponding parametric analyses. The transformation to ranks results in simple whole or fractional numbers of relatively small magnitude.

15.2 SIGN TEST

The sign test is probably the simplest of the nonparametric tests. The sign test is a test of the equality of the medians of two comparative groups. This test is used for *paired* data with an underlying continuous distribution, and can be applied to ranked or higher level data such as continuous *interval* and *ratio*-type data. The pairs are matched, and *differences* of the measurements for each pair tabulated. The differences are then categorized only with regard to the *sign* of the difference. That is, we count the number of times one treatment has a higher value than the other. *Ties* are not counted for this test. Ties give no information regarding which treatment has the higher median value. Theoretically, with continuous variables, there should be no ties.* However, with limited measuring instruments or the use of a crude rating scale, ties do occur.

As noted above, the sign test is a test of equal medians. If the test shows "significance," we can say that two comparative populations have different medians at the α level of significance. Under the null hypothesis that the medians of the two comparative distributions are the same, the probability of observing a value for Treatment A being larger or smaller than an observation for Treatment B is *one-half*; that is, the probability that an observation for Treatment A will be greater than a paired observation for Treatment B is one-half. Having recorded the differences, we compute the proportion of observations where the difference of treatment pairs is positive (or negative), disregarding ties (i.e., zero differences).

If positive and negative signs are observed to occur with approximately equal frequency, we can conclude that the treatments have a similar median. If either positive (+) or negative (−) signs predominate, there is evidence that one treatment has a higher median than the other. The statistical test is based on the binomial distribution. When applying two treatments to the same person, there are two possible outcomes: either Treatment A is favored or Treatment B is favored. Under the null hypothesis, the probability of A being favored is one-half; $H_0 : p = 0.5$. We compare the observed proportion to one-half (0.5). With N small and $p = 0.5$, the probabilities of various experimental outcomes can be calculated using computer software, or from the expansion of the binomial [Eq. (3.9)], or from tables of the binomial distribution (Table IV.3). For sample sizes of 6 to 20, inclusive, the number of positive or negative signs needed for significance at the 5% level for the sign test is given in Table IV.12. For sample sizes greater than 20, the normal approximation to the binomial, with a continuity correction, will suffice (see sect. 5.2.4). The normal approximation test is

$$Z = \frac{|p - 0.5| - 1/(2N)}{0.5/\sqrt{N}}, \tag{15.1}$$

where p is the observed proportion and N is the sample size. If Z is greater than 1.96, the treatments differ at the 5% level (two-sided test). The calculation can be simplified as follows:

$$Z = \frac{\left|\text{number of } + \text{'s} - \text{number of } - \text{'s}\right| - 1}{\sqrt{\text{number of } + \text{'s} + \text{number of } - \text{'s}}}. \tag{15.2}$$

Remember that ties are discarded and that N, the sample size, does *not* include ties.

Example 1. Because of its simplicity, the sign test may be used for a fast look at data from comparative experiments before applying a more sensitive parametric test such as the t test (if appropriate). This was the case for the data in Table 15.3, which were obtained to compare the "time to peak" plasma level for two oral formulations of the same drug. These data would usually be analyzed using a more sensitive nonparametric test (see sect. 15.3) or a t test for paired data (or ANOVA for a crossover design). Values were obtained by administering both drugs

* With continuous measurements, the probability of two values being identical is zero.

Table 15.3 Paired Data Obtained from the Bioavailability Experiment: Time to Peak Plasma Concentration

	Time to peak (hr)		Difference
Subject	***A***	***B***	***B − A***
1	2.5	3.5	+ 1
2	3.0	4.0	+ 1
3	1.25	2.5	+ 1.25
4	1.75	2.0	+ 0.25
5	3.5	3.5	0
6	2.5	4.0	+ 1.5
7	1.75	1.5	−0.25
8	2.25	2.5	+ 0.25
9	3.5	3.0	−0.5
10	2.5	3.0	+ 0.5
11	2.0	3.5	+ 1.5
12	3.5	4.0	+ 0.5

to each of 12 persons on two different occasions. Although these data would ordinarily result from a crossover design, and ANOVA techniques might be more appropriate, for the present purposes, we will consider an example where treatments have been assigned in random order. We will, therefore, not analyze "order" effects, and we will assume that no carryover effects are present.

From Table 15.3, tabulation of the differences ($B - A$) results in nine positive signs and two negative signs. One subject showed no difference between Treatments A and B. Referring to Table IV.12, 10 of 11 positive (or negative) signs are needed to obtain significance at the 5% level. Thus, according to the sign test, the difference just misses significance, although product B appears to take a longer time to peak than does product A.

If the differences can be assumed to have a normal distribution, the paired t test would be a more sensitive test than the sign test. For any given, specific example, one could not predict that the t test would result in a "more significant" difference; but on the average, the t test will be more discriminating. In this example, the t test results in a highly significant difference between the two formulations ($t = 3.02$; see Exercise Problem 1).

15.3 WILCOXON SIGNED RANK TEST

For the comparison of two treatments in a paired design, a more sensitive nonparametric test than the sign test is the Wilcoxon signed rank test. In the Wilcoxon test, the magnitude of the difference between the paired results is taken into consideration in addition to the sign. This feature results in a more powerful test, the sign test still retains its advantage for a very quick assessment of the experimental results.

The Wilcoxon test is based on the assumption that the distributions of the comparative treatments are symmetrical. Therefore, we are testing the equality of the means or the medians; the mean and median are equal in a symmetrical distribution.

The initial calculations are the same as in the sign test. We first take differences between the treatment pairs as in Table 15.3. Again, when the values for a treatment pair are equal (a difference of zero), a tie, these data are discarded for purposes of the test. As in the sign test, a zero difference does not contribute information regarding the differentiation of treatments in the Wilcoxon signed rank test. The differences of the untied pairs are then *ranked* in order of magnitude, *disregarding sign.* For the data in Table 15.3, the comparison of the time to peak plasma concentration for two formulations, A and B, the ranking of the absolute values of the differences is shown in Table 15.4. Differences of equal magnitude (disregarding sign) are given the *average rank.* The three subjects, 4, 7, and 8, all showed a difference (*absolute value*) equal to 0.25. Each of the differences are given a rank of 2, since these are the three smallest differences observed; 2 is the average of ranks 1, 2, and 3.

Table 15.4 Data from Table 15.3: Ranking Differences Without Regard to Sign for the Wilcoxon Signed Rank Test

Subject	Value	Rank	Assigned rank	Assigned rank with sign
7	−0.25	1	2	−2
4	0.25	2	2	2
8	0.25	3	2	2
9	−0.5	4	5	−5
10	0.5	5	5	5
12	0.5	6	5	5
1	1.0	7	7.5	7.5
2	1.0	8	7.5	7.5
3	1.25	9	9	9
6	1.5	10	10.5	10.5
11	1.5	11	10.5	10.5

Ranks with positive signs	Ranks with negative signs
2	2
2	5
5	Sum = 7
5	
7.5	
7.5	
9	
10.5	
10.5	
Sum = 59	

After ranking (disregarding sign) is completed, the signs corresponding to the signs of the original differences are reassigned to the ranks. For example, for subject 7 (originally given a rank of 2), the rank is changed to −2, because the difference for this subject was negative. The ranks with *like signs are summed* as shown following Table 15.4. The sum of the positive ranks is 59, and the sum of the negative ranks is 7. These are known as the *rank sums.* Table IV.13 gives the values of the *smaller* of the two rank sums needed for significance at the 5% level for various sample sizes, N (N is the sample size, the number of pairs, less the number of ties). The smaller rank sum must be equal to or less than that designated in Table IV.13 for the two means to be significantly different at the 5% level. In our example, Table IV.13 shows that the means are significantly different. The table shows that a rank sum of 10 or less for the smaller rank sum is significant at the 5% level for $N = 11$. In our example, the smaller rank sum is 7. Therefore the difference is significant at the 0.05 level ($p \sim 0.02$) [2]. This test gives very similar conclusions to that obtained by the t test. The Wilcoxon signed rank test is 95% as efficient as a t test for the comparison of normal populations. This means that a sample size of 100 that is analyzed using the Wilcoxon test would have equal sensitivity to a sample size of 95 using the t test. Considering the less restrictive assumptions of the Wilcoxon test compared to the t test, there is much to recommend it.

For sample sizes larger than those shown in Table IV.13, a normal approximation is available to compare two population means using the Wilcoxon signed rank test

$$Z = \frac{\left|R - N(N+1)/4\right|}{\sqrt{[N(N+1/2)(N+1)]/12}}, \tag{15.3}$$

where R is the sum of ranks (either the larger or smaller rank sum can be used) and N is the sample size (disregarding ties). This formula works well also for smaller sample sizes. In our

example, $N = 11$ and $R = 59$.

$$Z = \frac{|59 - 11(12)/4|}{\sqrt{[11(11.5)(12)]/12}} = 2.31.$$

From Table IV.2, $p = 0.02$, which is very close to the exact probability, if the data are normally distributed.

15.3.1 Nonparametric Confidence Intervals for Crossover Studies and Bioequivalence

If the assumptions of ANOVA (and *t* test) are violated, particularly the assumption of normality, a confidence interval can be formed based on a nonparametric approach. The method is based on ordering or ranking the outcomes and is relevant to bioequivalence studies, being introduced in this context. For the analysis of bioequivalence, a controversy concerning the nature of the data distribution recently polarized regulatory agencies. For many years, bioequivalence parameters were analyzed as the raw, untransformed values. For a two-period crossover design, this would be analogous to analyzing the differences of the treatments for each individual in the absence of period and carryover effects. Recently, agreement appears to have been reached, in the spirit of international harmonization, that a log transformation of AUC and C_{max} values is appropriate prior to the statistical analysis. This is analogous (but not the same) to an analysis of the ratio of the estimated parameters. However, one can use a nonparametric test in which the error structure and distribution assumptions are less rigid. A nonparametric confidence interval for ratios (or differences of logs) is given in Hollander and Wolfe [3] and is expounded in a paper by Steinijens and Diletti [4]. In this method, as opposed to parametric techniques, period and sequence (carryover) effects are assumed to be absent, and no adjustment is made for these effects.

The example in Steinijens and Diletti uses logs that would be appropriate in light of current practice. The method is described for N subjects in a two-period crossover design (or paired designs). First, compute the difference for each subject (e.g., test–reference). For the case of a log transformation for AUC, compute the difference of the product responses,

$$\log \mathrm{AUC}_t - \log \mathrm{AUC}_r = \log\left(\frac{\mathrm{AUC}_t}{\mathrm{AUC}_r}\right) = R$$

for each subject. (One may also calculate the ratios $\mathrm{AUC}_t/\mathrm{AUC}_r$ because of the one–one relationship of ranks to the differences of logs. Compute R', the average (geometric mean for ratios) of all possible pairs of the N individual ratios (R), where N is the number of subjects. There are $N(N+1)/2$ such pairs, including the ratio, R, for the same subject. (This will be clarified in the example below.) The values of R' are then ranked in order from low to high. The lower and upper nonparametric 90% and 95% confidence limits are given in Table 15.5. "C" is defined as the value of the R' that has the rank given in the table. For example, if "C" for the lower limit in Table 15.5 is 11, this means that the 11th ranked R' is given as the lower limit of the confidence interval. The details of the theory and the computations of C are given in Refs. [3,4,5]. In practice, it is not necessary to compute the logs because we are really interested in the ratios of test to reference. If we compute the ratios and use the geometric mean of the $N(N+1)/2$ pairs for the ranks, we will obtain the confidence interval for the ratio of test/reference directly. Again, this is a result of the monotonic relationship between the ratio and difference of the logs. The following example clarifies the procedure. In this example, both the parametric and nonparametric confidence intervals are calculated for purposes of comparison.

Example 2. Data for 12 subjects comparing two products for C_{max} is shown in Table 15.6. The ratio of the C_{max} for the products (B/A) is also calculated for each subject. In Table 15.7, the geometric mean of each pair of ratios (B/A) is shown in rank order. There are $N(N+1)/2$ such combinations (pairs) including each ratio with itself. The geometric mean is simply the square root of the product of 2 ratios. Thus, the ratio for subject 1 combined with itself is $102/135 =$

Table 15.5 Nonparametric Confidence Intervals Based on Wilcoxon's Signed Rank Test

	Rank for lower limit		Rank for upper limit	
Subjects (*N*)	**95%**	**90%**	**95%**	**90%**
6	1	3	21	19
7	3	4	26	25
8	4	6	33	31
9	6	9	40	37
10	9	11	47	45
11	11	14	56	53
12	14	18	65	61
13	18	22	74	70
14	22	26	84	80
15	26	31	95	90
16	30	36	107	101
17	35	42	119	112
18	41	48	131	124
19	47	54	144	137
20	53	61	158	150
21	59	68	173	164
22	66	76	188	178
23	74	84	203	193
24	82	93	219	208

0.756, and the geometric mean is the square root of $0.756 \times 0.756 = 0.756$. For subject 1 combined with subject 2, the geometric mean is the square root of $0.756 \times 0.821 = 0.788$, and so on.

For 12 subjects, the lower and upper cut-off points for a 95% confidence interval are the values ranked 14 and 65 (Table 15.5). For the data in this example, these values correspond to the ratios 0.800 and 1.247, respectively. The 90% confidence interval refers to the 18th and 61st rankings in Table 15.7, corresponding to an interval of 0.804 to 1.065. The 90% confidence interval would just pass the lower limits of the FDA requirements of 0.8 for the ratio.

Using a parametric analysis of variance (two-way ANOVA, assuming no period or sequence effects), with a log transformation (see Exercise Problem 15), the 90% interval is 0.79 to 1.26. The wider interval observed using the parametric approach is due to the "outlying" ratio for subject 3.

Table 15.6 Results for C_{max} from Bioequivalence Study

	Product		Ratio
Subject	***A***	***B***	***B/A***
1	135	102	0.7555556
2	179	147	0.8212290
3	101	385	3.8118813
4	109	106	0.9724771
5	138	189	1.3695653
6	135	105	0.7777778
7	158	130	0.8227848
8	156	125	0.8012821
9	174	144	0.8275862
10	147	133	0.9047619
11	145	114	0.7862069
12	147	167	1.1360544

Table 15.7 Ranks for Determining Confidence Interval for Bioequivalence Study

Rank	Subjects *A*, *B*	*R′* geometric mean	Rank	Subjects *A*, *B*	*R′* geometric mean
1	1, 1	0.75556	40	2, 4	0.89366
2	1, 6	0.76659	41	4, 7	0.89451
3	1, 11	0.77073	42	4, 9	0.89711
4	6, 6	0.77778	43	10, 10	0.90476
5	1, 8	0.77808	44	1, 12	0.92647
6	6, 11	0.78198	45	4, 10	0.93801
7	11, 11	0.78621	46	6, 12	0.94000
8	1, 2	0.78771	47	11, 12	0.94508
9	1, 7	0.78845	48	8, 12	0.95410
10	6, 8	0.78944	49	2, 12	0.96590
11	1, 9	0.79075	50	7, 12	0.96681
12	8, 11	0.79371	51	9, 12	0.96963
13	2, 6	0.79921	52	4, 4	0.97248
14	6, 7	0.79996	53	10, 12	1.01383
15	8, 8	0.80128	54	1, 5	1.01724
16	6, 9	0.80230	55	5, 6	1.03209
17	2, 11	0.80353	56	5, 11	1.03767
18	7, 11	0.80429	57	5, 8	1.04757
19	9, 11	0.80663	58	4, 12	1.05109
20	2, 8	0.81119	59	2, 5	1.06053
21	7, 8	0.81196	60	5, 7	1.06154
22	8, 9	0.81433	61	5, 9	1.06463
23	2, 2	0.82123	62	5, 10	1.11316
24	2, 7	0.82201	63	12, 12	1.13605
25	7, 7	0.82278	64	4, 5	1.15407
26	2, 9	0.82440	65	5, 12	1.24736
27	7, 9	0.82518	66	5, 5	1.36957
28	1, 10	0.82680	67	1, 3	1.69708
29	9, 9	0.82759	68	3, 6	1.72186
30	6, 10	0.83887	69	3, 11	1.73116
31	10, 11	0.84340	70	3, 8	1.74768
32	8, 10	0.85145	71	2, 3	1.76930
33	1, 4	0.85718	72	3, 7	1.77098
34	2, 10	0.86198	73	3, 9	1.77614
35	7, 10	0.86280	74	3, 10	1.85711
36	9, 10	0.86531	75	3, 4	1.92535
37	4, 6	0.86970	76	3, 12	2.08099
38	4, 11	0.87440	77	3, 5	2.28487
39	4, 8	0.88274	78	3, 3	3.81188

15.4 WILCOXON RANK SUM TEST (TEST FOR DIFFERENCES BETWEEN TWO INDEPENDENT GROUPS)

The sign test and Wilcoxon signed rank test are nonparametric tests for the comparison of paired samples. These data result from designs where each treatment is assigned to the same person or object (or at least subjects that are very much alike). If two treatments are to be compared where the observations have been obtained from two independent groups, the nonparametric *Wilcoxon rank sum* test (also known as the Mann–Whitney U test) is an alternative to the two independent sample *t* test. The Wilcoxon rank sum test is applicable if the data are at least ordinal (i.e., the observations can be ordered). This nonparametric procedure tests the equality of the *distributions* of the two treatments. Thus, this procedure tests for both the location and spread of the distributions.

The calculations for the Wilcoxon rank sum test are similar to those for the signed rank test discussed above. First, the observations from both groups are pooled and ranked, regardless of group designation. Identical observations are given a rank equal to the average of the ranks. In this procedure, the signs of the observations are taken into account for ranking. For example, a

Table 15.8 Results of a Dissolution Test Using the Original Dissolution Apparatus and a Modification: Amount Dissolved in 30 Minutes

Original apparatus		Modified apparatus	
Amount dissolved	**Rank**	**Amount dissolved**	**Rank**
53	3	58	11
61	14	55	5.5
57	9	67	21
50	1	62	15.5
63	17	55	5.5
62	15.5	64	18.5
54	4	66	20
52	2	59	12.5
59	12.5	68	22
57	9	57	9
64	18.5	69	23
		56	7
Sum of ranks	105.5		[170.5]

value of −1 has a lower rank than 0.5, which has a lower rank than 1. After ranking the pooled data, the observations are returned to their respective treatment groups. The observations are then replaced by their corresponding ranks. The *sum of the ranks of the smaller sample* is the basis for the statistical test. If the sample sizes are equal in the two treatment groups, the sum of the ranks in either group can be used as the statistic for the Wilcoxon rank sum test.

Table 15.8 shows tablet dissolution results observed in the original dissolution apparatus and a modification of the apparatus. The objective of this experiment was to compare the performance of the two pieces of apparatus. Twelve individual tablets were used for each ''treatment'' (apparatus). The amount of drug dissolved in 30 minutes was determined for each tablet. One tablet assay, determined in the original apparatus, is not included in the results (Table 15.5) because of an overt error during the assay procedure for this tablet.

Note how the ranks are obtained. The original apparatus has the four smallest values, 50, 52, 53, and 54, which are ranked 1, 2, 3, and 4, respectively. The next two highest values are from the modified apparatus, both equal to 55. These values are both given the *average* rank of 5 and 6, equal to 5.5. The next value, 56, from the modified apparatus is given a rank of 7. The next highest value is 57, which occurs twice in the original and once in the modified apparatus. These are each given a rank of 9, the average of the three ranks which these values occupy, 8, 9, and 10, and so on.

For moderate-sized samples, the statistical test for equality of the distribution means may be approximated using the normal distribution. This approximation works well if the smaller sample is equal to or greater than 10. For samples less than size 10, refer to Table IV.16 for exact significance levels [2]. The normal approximation is

$$Z = \frac{\left|T - N_1(N_1 + N_2 + 1)/2\right|}{\sqrt{N_1 N_2 (N_1 + N_2 + 1)/12}}, \tag{15.4}$$

where N_1, is the smaller sample size, N_2 is the larger sample size, and T is the sum of ranks for the smaller sample size. If Z is greater than or equal to 1.96, the two treatments can be said to be significantly different at the 5% level (two-sided test). In our example

$$Z = \frac{\left|105.5 - 11(11 + 12 + 1)/2\right|}{\sqrt{(11)(12)(11 + 12 + 1)/12}} = \frac{26.5}{16.25} = 1.63.$$

A value of Z equal to 1.63 is not large enough to show significance in a two-sided test at the 5% level ($p = 0.11$; Table IV.2). Therefore, these data do not provide sufficient evidence to show that the two different pieces of apparatus give different dissolution results.

One should appreciate, as noted previously, that in ranking tests, *ties* result only because of measurement limitations, because the distributions are assumed to be continuous. Too many ties result in erroneous probabilities with regard to the test of significance. The error is on the "conservative" side. For data with many ties (*more than 10%* of the data result in ties, as is the case in our example) statistical tests will tend to give results that overestimate α (i.e., the α error is larger than it should be). Hence, we tend to miss significant differences more often than we should when too many ties appear in the data. A correction for ties is available, but in most applications the difference between the corrected and uncorrected Z value is negligible.

It would also be of interest to compare the two pieces of equipment using the two independent sample t test in order to see how the conclusions might differ. Of course, in general, one cannot determine what would be expected to occur from a single example. The t test is more efficient than the nonparametric rank sum test if the assumptions for the t test are valid (see sect. 5.2.4). Similar to the signed rank test, the Wilcoxon rank sum test is very efficient, approximately 95% compared to the corresponding t test. A two independent groups t test for the data of Table 15.8 results in a t value of 1.84 with 21 d.f. ($p < 0.10$)

$$t = \frac{61.3 - 57.45}{5.05\sqrt{1/12 + 1/11}} = 1.84.$$

The probability level is somewhat less for the t test compared to the Wilcoxon rank sum test in this example. However, the conclusions are similar for the two statistical procedures.

The tests described above may replace the *paired t test* (use the *sign test* or *signed rank test*) or the *two independent groups t test* (use the *rank sum test*) when the assumptions required for the validity of the t tests are questionable. For the comparison of more than two groups, nonparametric tests, analogous to the analysis of variance parametric methods, are available. However, simple nonparametric tests are not available for the analysis of more advanced designs or for tests of interaction. The tests to be described below can be used to test for treatment effects for a simple one-way or two-way analysis of variance. These tests are widely used, and are recommended when ANOVA assumptions regarding normality are suspect and/or cannot be easily tested. The nonparametric tests are useful in experiments where the data consist of values derived from a rating scale with an underlying continuous distribution.

15.4.1 Nonparametric Analysis of Two-Way Crossover (Bioequivalence Designs)

Some people have advocated the use of nonparametric analyses for crossover designs or for pharmacokinetic parameters from bioequivalence studies. As presented earlier (sect. 15.3.1), the reason for this is the less restrictive assumptions of the nonparametric analysis. In particular, this would, apparently, resolve the problem of violations of certain assumptions inherent in ANOVA, for example, linearity, normality, and variance assumptions, although the theory is quite complex. One problem with nonparametric techniques for the analysis of these data is that we cannot account simultaneously for effects due to periods or carryover in the nonparametric model. Cornell [6] has presented a lucid discussion of methods to analyze these data taken from Koch [7]. One can demonstrate that specific sums and differences of observations in the two-way crossover design are equivalent to effects of interest, that is, treatment, period, and carryover effects (see discussion of parametric analysis in sect. 11.4.2). Applying a model that includes treatment, period, carryover, and random effects, the principles of the analysis are as follows:

1. Total the data for each subject over both periods and compare the totals for group (Sequence) I (subjects taking test followed by reference) to the totals for Group II (subjects taking reference followed by test). This comparison is a test for unequal carryover effects (Note that this is

the same procedure as in the parametric analysis, where the sequence effect is confounded with a carryover).
2. Take differences of Period 1 and Period 2 for each subject. Compare the differences for Group 1 to Group 2. This is a test of treatment differences.
3. Take differences of Treatment 1 and Treatment 2 for each subject. Compare the differences for Group 1 to Group 2. This is a test of period differences.

To see how this works, consider the estimate of treatment effects (item 2 above). In Group 1, Treatment 2 follows Treatment 1; in Group 2, Treatment 1 follows Treatment 2. The expected value for each subject is

$$\text{In Group 1, Period 1} : \mu + P_1 + T_1$$

$$\text{In Group 1, Period 2} : \mu + P_2 + T_2 + C_1$$

$$\text{In Group 2, Period 1} : \mu + P_1 + T_2$$

$$\text{In Group 2, Period 2} : \mu + P_2 + T_1 + C_2$$

where P_i, T_i, and C_i refer to the effects due to period i, treatment i, and carryover due to treatment i, respectively.

The expected values of the differences between Period 1 and 2 for the two groups are

$$\text{Group 1} : P_1 + T_1 - P_2 - T_2 - C_1$$

$$\text{Group 2} : P_1 + T_2 - P_2 - T_1 - C_2$$

If carryover has been shown to be nonsignificant, $C_1 = C_2$ (see next paragraph), the difference between the expected values for Groups 1 and 2 is equal to $2(T_1 - T_2)$, or twice the treatment effect. The same approach can be used to demonstrate the results of the calculations for sequence and period effects, items 1 and 3 above (see Exercise Problem 16).

The nonparametric statistical tests may be applied sequentially. If the sequence effect is significant, we may have to use only the Period I data as is the case for the parametric analysis, as has been noted in chapter 11. For bioequivalence studies, real carryover effects are very rare because of the nature of the design (short period of dosing and washout period). Therefore, significant carryover effects may be dismissed if there is no rational or reasonable explanation for their existence. (The FDA has accepted carryover effects as spurious for single dose studies, in some cases, if the sponsor can demonstrate no obvious cause. However, in the nonparametric test, no adjustment is made for the treatment differences in the presence of period or carryover effects. Therefore, one should be cautious when applying these tests in the presence of a significant "carryover.") We can then apply the usual nonparametric tests. The data in Table 15.9, taken from Wallenstein and Fisher [8] as also presented by Cornell, are used to illustrate the procedure.

To test for significance, the Wilcoxon rank sum test is applied to the ranks of each of the differences (Period 1–Period 2 and Treatment 1–Treatment 2) in Table 15.9. Eight subjects are in Sequence I, Treatment 1 followed by Treatment 2. Nine subjects are in Sequence II. The comparison of treatment totals for Sequences I and II is a test for a sequence or carryover effect. The sequence effect is not significant [see Exercise Problem 17 and Eq. (15.4)]. The treatment effect can be tested by comparing the Period I to Period 2 differences for the two sequences [(Treatment 1–Treatment 2)$_1$ − (Treatment 2–Treatment 1)$_2$]. The sum of ranks for Period 1 to

Table 15.9 AUC (log) Data for Crossover Study [11] to Illustrate Nonparametric Analysis

Sequence (Group I)	Pd. 1	Pd. 2	Total	Rank	Pd. 1–Pd. 2	Rank	Tr1–Tr2	Rank
Tr1–Tr2	2.60	2.16	4.76	16	0.44	1.5	0.44	1.5
	2.81	2.53	5.34	5	0.28	4	0.28	4
	3.02	2.69	5.71	1	0.33	3	0.33	3
	2.59	2.50	5.09	9	0.09	9	0.09	12
	2.70	2.45	5.15	8	0.25	5	0.25	5
	2.01	2.49	4.50	17	−0.48	17	−0.48	17
	2.71	2.27	4.98	10	0.44	1.5	0.44	1.5
	2.67	2.55	5.22	7	0.12	8	0.12	10
	Total			73		49		54
Sequence (Group II)								
Tr2–Tr1	2.57	2.38	4.95	11	0.19	6	−0.19	16
	2.36	2.50	4.86	13.5	−0.14	13	0.14	13
	2.73	2.75	5.48	2	−0.02	11	0.02	9
	2.38	2.55	4.93	12	−0.17	14	0.17	8
	2.64	2.75	5.39	3	−0.11	12	0.11	11
	2.52	2.71	5.23	6	−0.19	15	0.19	7
	2.46	2.32	4.78	15	0.14	7	−0.14	15
	2.57	2.79	5.36	4	−0.22	16	0.22	6
	2.46	2.40	4.89	13.5	0.06	10	−0.06	14
	Total			80		104		99

Period 2 for Sequence I is 49. Applying Eq. (15.4),

$$Z = \frac{|49 - 8(8+9+1)/2|}{\sqrt{8 \times 9(8+9+1)/12}} = 2.21 (p < 0.05).$$

The test for period effects is based on the comparison of the ranks in the two sequences in the last column of Table 15.9 (see Exercise Problem 17). Of course, the current test for bioequivalence is not based on statistical significance. Nevertheless, the nonparametric approach to this problem is instructive. See section 15.3.1 for an illustration of a nonparametric confidence interval to bioequivalence data.

15.5 KRUSKAL–WALLIS TEST (ONE-WAY ANOVA)

The Kruskal–Wallis test is an extension of the rank sum test to more than two treatments, and is basically a test of the *location* of the distributions, assuming variance symmetry, that is, equality of variances in the different groups. Significant differences can be interpreted as meaning that the averages of at least two of the comparative treatments are different. The computations and analysis will be illustrated using an experiment in which data were obtained from a preclinical experiment in which rats, injected with two doses of an experimental compound and a control (a known sedative), were observed for sedation. The time for the animals to fall asleep after injection was recorded. If an animal did not fall asleep within 10 minutes of the drug injection, the time to sleep was arbitrarily assigned a value of 15 minutes. The experimental results are shown in Table 15.10. One data point was lost from the control group because of an illegible recording, obliterated in the laboratory notebook.

The analysis for treatment differences is not dependent on equal numbers of observations per group, although, as in most experiments, equal sample sizes are most desirable (optional). The analysis consists of first combining all of the data, as in the Wilcoxon rank sum test. To obtain the ranks, one lists all observations in order of magnitude, identifying each value by its group designation. The observations are then reclassified into their original groups, similar to the Wilcoxon rank sum test procedure. The ranks corresponding to each observation are retained and summed for each group as shown in Table 15.10. Note that ties are given the average rank

Table 15.10 "Time to Sleep" for a Control and Two Doses of an Experimental Compound (minutes)

Control	Rank	Low dose	Rank	High dose	Rank
8	22	10	26	3	10
1	3.5	5	13	4	12
9	24.5	8	22	8	22
		6	15	1	3.5
9	24.5	7	18.5	1	3.5
6	15	7	18.5	3	10
3	10	15	28	1	3.5
15	28	1	3.5	6	15
1	3.5	15	28	2	7.5
7	18.5	7	18.5	2	7.5
Sum of ranks	149.5		191.0		94.5

as in the previously described rank sum test. In addition to the usual analysis, we will present a procedure that corrects the analysis for tied observations [3].

The test statistic for the Kruskal–Wallis test, as described below, is approximately distributed as Chi-square with $k-1$ d.f., where k is the number of treatments (groups) in the experiment. For small sample sizes, tables to determine the treatment rank sums needed for significance are available [3]. The Chi-square approximation is good if the number of observations in each group is greater than five. The computation of the Chi-square statistic is as follows:

$$\chi^2_{k-1} = \frac{12}{N(N+1)}\left(\sum \frac{R_i^2}{n_i}\right) - 3(N+1), \tag{15.5}$$

where N is the total number of observations in all groups combined, R_i the sums of ranks in ith group, n_i the number of observations in ith group, and k the number of groups.

In our example, $N = 29, R_1 = 149.5, R_2 = 191, R_3 = 94.5, n_1 = 9, n_2 = 10, n_3 = 10$, and $k = 3$. Applying Eq. (15.5), we have

$$\chi^2_2 = \frac{12}{(29)(30)}\left(\frac{149.5^2}{9} + \frac{191^2}{10} + \frac{94.5^2}{10}\right) - 3(29+1) = 6.89.$$

The value of Chi-square with 2 d.f. must be equal to or greater than 5.99 to be significant at the 5% level (Table IV.5). Therefore, the average "time to sleep" differs for at least two of the three treatment groups (control, high dose, and low dose) at the 5% level of significance.

As in the parametric tests, if statistically significant differences among treatments are found, one usually would want to know which treatments are different. For individual (pairwise) comparisons, Table IV.17 tabulates the differences between rank sums needed for significance at the 5% level, given the number of treatments in the design and the sample size [2]. To perform the pairwise treatment comparisons, the number of observations per treatment must be the same. For example, in the case of three treatments, each with a sample size of 10, a difference between the rank sums of two of the treatments (groups) must exceed 92 in order for the two treatments to be considered different at the 5% level. In our example, had the control group had 10 observations instead of 9, we could apply the pairwise test. However, if an additional observation had been included in the control group, the greatest difference between the rank sums of the control group and one of the doses of the experimental drug in this experiment could not exceed 92.[†] The observed difference between the high and low doses is $(191 - 94.5) = 96.5$, which exceeds 92. Thus, the pairwise comparison criterion shows a significant difference

[†] The largest difference between the control and one of the experimental drug doses would occur if the tenth value in the control group were the highest observation. The rank sum of the control group would be increased by 30, resulting in a rank sum of 179.5 (Table 15.10). The difference between the rank sums of the control and high-dose groups would be $179.5 - 94.5 = 85$, which is not significant at the 5% level.

between the high and low doses of the experimental drug ($p < 0.05$), agreeing with the significant Chi-square test. For more details concerning multiple comparisons in the Kruskal–Wallis test, see Refs. [2,3].

As in the ranking procedures previously described, tied values are given the average rank. A correction for ties can be used that increases the value of Chi-square. Therefore, if the null hypothesis is rejected (significant treatment differences), the correction only increases the degree of significance. If Chi-square just misses significance, the correction may result in statistically significant differences. The correction is as follows:

$$\text{Correction} = \frac{\chi^2}{1 - \sum(t_i^3 - t_i)/(N^3 - N)},$$

where t_i is the number of tied observations in group i and N is the total number of observations. The calculations are illustrated below. There are eight groups of ties in the data shown in Table 15.10. For example, there are six values equal to 1. For this group of ties, $t^3 - t$ *is* equal to $6^3 - 6 = 210$. Another group of ties are the two values equal to 2. There are two values of 2 in the data, and for this group, $t = 2$ and $t^3 - t = 6$. The other ties occurred for values of 3, 6, 7, 8, 9, and 15. The reader can verify that the sum of T (where $T = \Sigma(t_i^3 - t_i)$) is 378. The correction for Chi-square is

$$\frac{6.89}{1 - 378/(29^3 - 29)} = \frac{6.89}{0.984} = 7.00.$$

(Note that $N = 29$ in this example.) The correction for ties is usually very small. Of course, in this example, the correction does not change the conclusion of significant differences among treatment means.

15.6 FRIEDMAN TEST (TWO-WAY ANALYSIS OF VARIANCE)

The Friedman test is a nonparametric test applied to data that is, at least, ranked and that is in the form of a two-way ANOVA design (randomized blocks). This test, which may be applied to ranked or interval/ratio-type data, is used when more than two treatment groups are included in the experiment. For two groups in a paired (two-way) design, the rank sum test may be used. In the Friedman test, the treatments are ranked *within each block* (e.g., animal or person), disregarding differences between blocks. The procedure will be illustrated using the data from Table 15.11. These data describe the results of a *validation* experiment to test the performance of four tablet presses, with regard to tablet hardness. The average hardness of 10 tablets was computed for five different tablet products manufactured on four presses. The tablets are a random selection of five typical tablet products. The presses were identically set for the same pressure for each tablet formulation.

The parenthetical values in Table 15.11 are the ranks of the average hardness for each formulation over the four presses. For formulation 1, the lowest value, 6.9, is assigned a rank of 1, and the highest value, 7.5, is assigned a rank of 4. Although no ties occurred in this example, if ties were observed, the average rank would be assigned to the tied observations as discussed in the preceding sections. If one of the presses consistently had the highest (or lowest) rank, one would conclude that the press (treatment) produced harder (or less hard) tablets than the other presses. In our example, tablet press *C* had the highest hardness value for all formulations with the exception of formulation 1, where it had the next-to-largest value. The test of significance is an objective assessment of whether or not the data of Table 15.11 provide sufficient evidence to say that tablet press *C* is, indeed, producing harder tablets than the other presses.

Table 15.11 Average Hardness of 10 Tablets for Five Different Tablet Formulations Prepared on Four Presses[a]

	Tablet press			
Tablet formulation	***A***	***B***	***C***	***D***
1	7.5 (4)	6.9 (1)	7.3 (3)	7.0 (2)
2	8.2 (3)	8.0 (2)	8.5 (4)	7.9 (1)
3	7.3 (1)	7.9 (3)	8.0 (4)	7.6 (2)
4	6.6 (3)	6.5 (2)	7.1 (4)	6.4 (1)
5	7.5 (3)	6.8 (2)	7.6 (4)	6.7 (1)
R_i	14	10	19	7

[a]Parenthetical values are the within-tablet press ranks.

If the sample sizes are sufficiently large, a Chi-square distribution can be used to approximate the test of significance. The Chi-square test is

$$\chi^2_{c-1} = \frac{12}{rc(c+1)}\left(\sum R_i^2\right) - 3r(c+1), \tag{15.6}$$

where χ^2_{c-1} is the χ^2 statistic with $c-1$ d.f., r the number of rows (blocks), c the number of columns (treatments), and R_i the sums of ranks in the ith group (column).

In our example, the Chi-square statistic has 3 d.f.

$$\chi^2_3 = \frac{12}{(5)(4)(4+1)}(14^2+10^2+19^2+7^2) - 3(5)(5) = 9.72.$$

A Chi-square value of 7.81 or larger is needed for significance at the 5% level (Table IV.5). We can conclude that at least two of the tablet presses differ with regard to tablet hardness. Examination of Table 15.11 shows that tablet press *C* produces harder tablets than those produced by the other presses. Table IV.18 shows that a difference of 11 is needed for significance ($p < 0.05$) for individual comparisons between pairs of means for 4 treatments ($k = 4$) and 5 rows ($n = 5$). Therefore, press *C* produces significantly harder tablets than press *D* with a sum of ranks of 19 and 7, respectively.

For small samples, exact probabilities for the Friedman test are given in *Nonparametric Statistical Methods* [3]. This test also describes a test that corrects Chi-square for tied observations.

15.6.1 Modified Friedman Test

Conover [9,10] recommends a statistic that has an approximate F distribution with $(c-1)$, $(c-1)(r-1)$ d.f. (where r is the number of rows and c is the number of columns in the RXC matrix of data). This method of analysis has been shown to be superior to the Chi-square distribution for the Friedman nonparametric analysis (sect. 15.6) of a two-way ANOVA model. The statistic T_2 is calculated as follows:

Compute $A_2 = \sum(x_{ij})^2$, where the x_{ij} are the individual ranks.

A_2 is equal to $cr(c+1)(2c+1)/6$ if there are no ties (ties are given the value of the average rank).

c = number of columns and r = number of rows

Compute $B_2 = (1/r)\sum(C_i)^2$

where C_i is the sum of observations in column i. Then

$$T_2 = \frac{[(r-1)\{B_2 - rc(c+1)^2/4\}]}{A_2 - B_2}$$

Refer T_2 to an F table with $c-1$ and $(r-1)(c-1)$ d.f.

Example 3. The computations for this analysis are shown below for the data from Table 15.11

$$A_2 = \frac{cr(c+1)(2c+1)}{6} = \frac{4 \times 5 \times (4+1)(2 \times 4+1)}{6} = 150$$

$$B_2 = \left(\frac{1}{r}\right)\sum (C_i)^2 = \left(\frac{1}{5}\right)(14^2 + 10^2 + 19^2 + 7^2) = \frac{706}{5} = 141.2$$

$$T_2 = \frac{[(5-1)(141.2 - 4 \times 5(4+1)^2/4)]}{(150 - 141.2)} = 7.364$$

Compare 7.364 to the tabled value of F with 3 and 12 d.f. at the 5% level (App. IV, Table IV.6A)

$$F_{3,\,12,\,0.05} = 3.49.$$

Since the observed F (7.364) is larger than the tabled F (3.49) at the 5% level, the differences among tablet presses are significant ($p = 0.005$). The usual Friedman test that uses a Chi-square statistic shows a level of 0.02 (sect. 15.6). See Exercise Problem 18 at the end of this chapter for the application of ANOVA to this data.

15.6.1.1 Multiple Comparisons for the Modified Friedman Test

If the null hypothesis of equal treatment means is rejected, the following formula can be used to calculate a least significant difference between pairs of treatments:

$$[C_j - C_i] > t\sqrt{[2r(A_2 - B_2]/[(r-1)(c-1)}$$

where t is the tabled t value with $(r-1)(c-1)$ d.f. at the specified alpha level.

Applying this formula to the data in Table 15.11 for tablet press differences at the 5% level.

$$\left[C_j - C_i\right] > 2.18\sqrt{[2 \times 5(150 - 141.2)]/[(5-1)(4-1)]} = 5.90.$$

Any difference in rank sums $\geq$5.9 is significant at the 0.05 level. Inspection of the results shown in Table 15.11 shows that Tablet Press C gives higher results ($p < 0.05$) than B and D, and A is higher than D. In this example, we see more significant differences with the modified test compared to the Friedman test described in section 15.6.

15.6.2 Quade Test for Randomized Block Design

Conover [9] presents another test (Quade Test) that is still valid in the presence of many ties. In addition to the usual computations as shown in the Friedman Test, a further computation is needed. The range of values (largest minus smallest value) is calculated for each block (row). The blocks are ranked in order from the smallest to the largest with regard to the range of values within a block. Call these ranks $Q_1 \ldots Q_r$, where r is the number of rows (blocks). Let $R(X_{ij})$ be the rank of each observation, where ranks are within each row or block. Compute for each observation

$$S_{ij} = Q_i[R(X_{ij}) - (k+1)/2],$$

where k is the number of treatments (columns).

Let $S_i = \sum S_{ij}$ for each treatment.
Calculate $A = \sum S_{ij}^2$ (for all observations).
Calculate $B = \sum S_i^2/r$.

For Table 15.11, the calculations for the "Quade" test are as follows:

Range of row 1 = 0.6 (7.5 − 6.9)
Range of row 2 = 0.6 (8.5 − 7.9)
Range of row 3 = 0.7 (8.0 − 7.3)
Range of row 4 = 0.7 (7.1 − 6.4)
Range of row 5 = 0.9 (7.6 − 6.7)

Q_i is the rank of row i.

$Q_1 = 1.5$

$Q_2 = 1.5$

$Q_3 = 3.5$

$Q_4 = 3.5$

$Q_5 = 5$

(Note: as usual, compute the average rank for ties.) As an example, the calculation of S_{11} follows:

S_{11}= value for formulation 1 on press $1 = 1.5[4 - (4 + 1)/2] = 2.25.$

The values of S_{ij} derived from the data in Table 15.11 are shown in Table 15.12.

$$A = \sum \left(S_{ij}\right)^2 = 270$$

$$B = \sum \frac{S_i^2}{r} = \left[2^2 + (-5.5)^2 + 21^2 + \left(-17.5^2\right)\right] 5 = 156.3$$

The test statistic is

$$T = \frac{(r-1)B}{A - B}$$

$$T = \frac{4(156.3)}{270 - 156.3} = 5.499.$$

Refer T to an F distribution with $(c - 1)$ and $(r - 1)(c - 1)$ d.f. at the appropriate alpha level (App. IV, Table IV.6A).

Table 15.12 Table to Aid Computations for Quade Test *(S_{ij})* Press

	A	*B*	*C*	*D*	Range	Rank
Formulation						
1	2.25	−2.25	0.75	−0.75	0.6	1.5
2	0.75	−0.75	2.25	−2.25	0.6	1.5
3	−5.25	1.75	5.25	−1.75	0.7	3.5
4	1.75	−1.75	5.25	−5.25	0.7	3.5
5	2.5	−2.5	7.5	−7.5	0.9	5
Sum	2.00	−5.50	21.00	−17.50		

Tabled $F_{3,12} = 3.49$ at the 5% level.

Therefore, at least two of the presses are significantly different ($p = 0.013$). Multiple comparisons can be made if the F test shows significance. The difference between the sums of any two treatments is significant if the absolute value of the difference exceeds

$$t \times \left[\frac{2r(A-B)}{(r-1)(c-1)}\right]^{1/2},$$

where t is the appropriate tabled t value at the alpha level with $(r - 1)(c - 1)$ d.f. In this example, $t_{0.05,12} = 2.18$, and the least significant difference is

$$2.18\left[\frac{10(270-156.3)}{(4)(3)}\right]^{1/2} = 21.22.$$

We conclude that press C gives higher results than presses B and D at the 5% level of significance.

This analysis is identical to an analysis of variance on the ranks in Table 15.12. The least significant difference is computed as in Fisher's LSD, based on the analysis of the "adjusted" ranks, as computed from Eq. (8.7) (see Exercise Problem 19).

Three different tests applied to these data give somewhat different overall conclusions. This is caused by the fact that some of the comparisons are close to significant (C vs. A and C vs. B). As always the test to be applied and the level of significance should be clearly defined at the initiation of the experiment.

15.7 NONPARAMETRIC ANALYSIS OF COVARIANCE

Quade has proposed a simple and neat nonparametric analysis of covariance (ANCOVA) [11]. The procedure is described in detail using the data of Table 15.13.

Rank each of X and Y (raw material and product assays, respectively) disregarding treatment. Let the lowest value have rank 1 up to the highest value, rank N, where there are a total of N observations. Correct the rankings so the mean of the ranks $= 0$, by subtracting the average rank, $(N + 1)/2$, from each rank Y. In this example, $N = 8$. The lowest value of Y (product assay) is 95.4 and is given a rank of 1. Subtract $(N + 1)/2 = 9/2 = 4.5$ from 1, resulting in $R_y = 1 - 4.5 = -3.5$. Similarly, the largest assay is 98.5, with an adjusted rank of $8 - 4.5 = +3.5$. The ranks of the raw material assays are calculated similarly. Use average ranks in case of ties.

Table 15.13 Data for Quade Nonparametric Covariance Analysis (ANCOVA)

Final assay Y	Raw material X	Ranks R_y	-4.5 R_x	Predicted	Residual
Method I					
98.00	98.40	2.50	−3.00	1.4451220	1.0548780
97.80	98.60	1.50	−1.00	0.4817073	1.0182927
98.50	98.60	3.50	−1.00	0.4817073	3.0182927
97.40	99.20	−0.50	2.50	−1.2042683	0.7042683
Sum					5.795732
Method II					
97.60	98.70	0.50	0.50	−0.2408537	0.7408537
95.40	99.00	−3.50	1.50	−0.7225610	−2.7774391
96.10	99.30	−2.00	3.50	−1.6859756	−0.3140245
96.10	98.40	−2.00	−3.00	1.4451220	−3.4451220
Sum					−5.795732

Next, perform a regression of the adjusted ranks of $Y(R_y)$ on the adjusted ranks of $Y(R_x)$ for all data, to obtain the residuals. Remember, the residuals are the difference between the observed values of R_y and the predicted values of R_y based on the calculated regression parameters (Table 15.13).

An analysis of variance is performed on the residuals (Group I vs. Group II). Note that there is no correction for the mean because the mean of the residuals is 0.

Quade used the following formula that has an F distribution with $k-1$ and $N-k$ d.f., where k is the number of groups and N is the total number of observations. In our example, we have two groups and eight observations, resulting in an $F_{1,6}$ distribution.

$$F_{k-1,N-k} = \frac{(N-k)\sum (Z_{ij})^2/n_i}{(k-1)[\sum_i \sum_j Z_{ij}^2 - \sum (Z_{ij})^2/n_i]}$$

$$N = 8$$

$$k = 2$$

$$n_1 = n_2 = 4$$

$$\sum \frac{(Z_{ij})^2}{n_i} = \frac{(5.795732^2 + \{-5.795732\}^2)}{4} = 16.795255$$

$$\sum_i \sum_j Z_{ij}^2 = 31.98628$$

$$F_{k-1,N-k} = (8-2)(16.795255/[(2-1)(31.98628 - 16.795255)]$$

$$F_{1,6} = 6.634$$

$p = 0.042$ (this result may be compared to $p = 0.037$ using a parametric analysis, sect. 8.6).

An assumption for this test is that the variables be on an ordinal scale, not necessarily continuous (dichotomous variables may be used). We do not have to assume normality or linearity of y on x. However, the distribution of X should be the same in each group, a requirement not needed for the parametric analysis.

15.8 RUNS TEST FOR RANDOMNESS

When performing an experiment (or observing a process) where values are observed sequentially, it may be of interest to determine whether the observations are randomly varying about the central value (i.e., the median). If the process is not random, we might expect to see trends in the data, perhaps a consecutive series of high or low values, which are unlikely to occur by chance. The *runs* test is a simple method of investigating the "random" nature of such a process. Tests for runs were introduced in section 12.2.5, the discussion of control charts. A run is a series of *uninterrupted, like* observations. For example, suppose that the median weight of 20 tablets, sequentially taken during a batch run, is 200 mg. Twenty consecutive tablets were weighted with the following results:

The first six tablets weighed more than 200 mg.
The next five tablets weighed less than 200 mg.
The next four tablets weighed more than 200 mg.
The next (remaining) five tablets weighed less than 200 mg.

If we designate tablet weights less than 200 mg by a minus (−), and tablet weights more than 200 mg by a plus (+), the 20 weights can be described by the following sequence:

$$200\,\text{mg} \rightarrow \underline{++++++}\ \underline{------}\ \underline{++++}\ \underline{-----}$$

The first six values, +'s, represent a *run.* Each time that a series of like signs change, a new run begins. There are *four* runs in these data: *six* pluses, *five* minuses, *four* pluses, and *five* minuses. If the tablet weights follow a random process, one might suspect that the sequence of values described above is unlikely. It appears that the pluses and minuses come in "bunches." One might guess that the sequence of pluses and minuses could have been due to too-frequent weight adjustments on the tablet press. For example, the first tablets sampled were over the median weight of 200 mg. The tablet press may then have been adjusted down, more than necessary, resulting in too-low tablet weights (the next five tablets were underweight), and so on.

To test for randomness for sample sizes as large as 40, we can refer to Table IV.14. The table gives the lower and upper limits for the number of runs that would be expected to occur in a random process in a sample of size N. An observed number of runs equal to or less than the critical lower number or greater than the critical upper number shown in Table IV.14 is an indication that the process is not random at the 5% level. The runs test is usually a two-sided test; either too few or too many runs lead to significance (nonrandomness). In some cases, for example, control charts, only relatively few runs may be considered to suggest problems with a process. In these situations, critical values for a one-sided test as shown in Table IV.14 are appropriate. According to Table IV.14, for a sample size of 20, between 7 and 16 runs would be expected to occur if the null hypothesis of randomness is true. We observed four runs in the sample of 20 tablets ($N = 20$) in our example. Therefore, we conclude that the process is not random ($p < 0.05$). The clusters of high and low values are probably due to some malfunctioning of the tableting process.

Consider the following as a further example of an application of the *runs* test. A *standard* is analyzed every 20th sample in an automated analytical procedure. A record of the readings for the standard in chronological order derived from one day's assay results are shown in Table 15.14. The median value for the data in the table is 0.7985 (the 20th and 21st ordered values are 0.798 and 0.799). As in the previous example, we label values greater than the median as + and values less than the median as −. The sequence of pluses and minuses is as follows (Samples 1 and 2 are below the median; 3 and 4 are above the median, etc.):

$$\underline{--}\,\underline{++}\,\underline{-}\,\underline{+}\,\underline{-}\,\underline{+}\,\underline{-}\,\underline{+}\,\underline{---}\,\underline{+++++++}\,\underline{---}$$

$$\underline{+++++}\,\underline{--}\,\underline{+++}\,\underline{--------}$$

The runs are underlined in the previous sequence. There are 15 runs. For sample sizes of 40 or more, a normal approximation to the distribution of runs is available, under the null hypothesis that the observed values occur in a random manner.

$$Z = \frac{|r - (N/2 + 1)|}{\sqrt{N(N-2)/4(N-1)}}, \tag{15.7}$$

where r is the number of runs and N is the sample size.

Values of Z equal to or greater than 1.96 are unlikely ($p \leq 0.05$) if the observations are random. In our example $N = 40$ and $r = 15$. Therefore,

$$Z = \frac{|15 - (40/2 + 1)|}{\sqrt{40(40-2)/4(40-1)}} = \frac{6}{3.12} = 1.92.$$

The value of Z is not quite large enough for the data to be considered nonrandom at the 5% level. Table IV.14 shows that for a sample of size 40, an observation of between 15 and 26 runs leads to acceptance of the null hypothesis of randomness, agreeing with the conclusion of

Table 15.14 Readings of a Standard Solution in Chronological Order (Optical Density)

Sample	Reading
1	0.795
2	0.796
3	0.804
4	0.801
5	0.792
6	0.816
7	0.791
8	0.819
9	0.796
10	0.815
11	0.782
12	0.795
13	0.798
14	0.800
15	0.800
16	0.802
17	0.799
18	0.805
19	0.820
20	0.802
21	0.796
22	0.797
23	0.795
24	0.802
25	0.800
26	0.801
27	0.802
28	0.820
29	0.788
30	0.780
31	0.813
32	0.804
33	0.801
34	0.793
35	0.790
36	0.791
37	0.784
38	0.791
39	0.788
40	0.794

the normal approximation [Eq. (15.7)]. Had 14 runs been observed, we would have concluded that the data were not random ($p < 0.05$).

15.9 CONTINGENCY TABLES

Chi-square tests for contingency tables (e.g., 2×2 tables) are often categorized as nonparametric tests. The analysis of 2×2 tables using a Chi-square test was described in section 5.2.5. The Chi-square test can be applied to nominal or categorical data that cannot be analyzed using the ranking techniques discussed above. These data cannot be ordered (the data are not ordinal or on an interval/ratio scale). Nominal data are usually available in the form of *counts*, such as 25 males and 12 females entered into a clinical study; or the *number* of tablets categorized as acceptable, chipped, cracked, and so on. For large samples, Chi-square methods can be used to compare "statistically" the relative frequency of such events that occur in two or more groups. Here we will briefly expand the case of the fourfold table, discussed in Chapter 5, to the analysis

Table 15.15 Examples of $R \times C$ Tables

2×2 table (Fourfold table)

Treatment	Cured	Not cured
A		
B		

2×3 table

Treatment	Unsuccessful	Moderately successful	Successful
A	25	10	40
B	27	23	25

$R \times C$ table

	Outcome
Treatment	1 2 3 ⋯ *C*
1	
2	
3	
.	
.	
R	

of $R \times C$ tables, R rows and C columns. We will then examine the case of 2×2 tables with small expected frequencies, followed by different tests of hypotheses for fourfold tables.

15.9.1 *R* × *C* Tables

In the binominal case, data are dichotomized, resulting in the 2×2 table, for example, comparison of success rates of two treatments as shown in Table 15.15. When experiments consist of more than two comparative groups and/or more than two possible outcomes, we are, in general, confronted with an $R \times C$ table (Table 15.15).

In the experiments involving contingency tables, we are usually interested in testing group differences with regard to proportions or the distribution of counts in the various outcome categories. Consider the data in the 2×3 table in Table 15.15. Two treatments have been compared where the outcomes are categorized as "unsuccessful," "moderately successful," and "successful." Inspection of the data indicates that Treatment *A* has a greater incidence of "successful" events and less "moderately successful" events than Treatment *B*.

Equivalently the hypothesis in contingency tables is often stated in terms of the relationship between rows and columns. "Acceptance" of the null hypothesis suggests that the rows and columns are independent. For example, in the 2×3 contingency table in Table 15.15, lack of rejection of the null hypothesis would be interpreted, in this context, as meaning that the experimental outcomes are independent of the treatment (i.e., the treatments do not differ with respect to the experimental outcome).

The relationship of the rows and columns in an $R \times C$ contingency table may be tested by means of the Chi-square distribution with $(R - 1)(C - 1)$ d.f. Note that for a 2×2 table, we have 1 d.f., agreeing with the analysis of 2×2 tables described in chapter 5. The Chi-square statistic is calculated as

$$\chi^2_{(R-1)(C-1)} = \sum \frac{(O - E)^2}{E}, \tag{15.8}$$

where O is the observed count and E is the expected count. The summation in Eq. (15.8) is for all $R \times C$ cells in the contingency table.

The Chi-square test is an approximate test and should be used only when the expected values are sufficiently large. The usually recommended minimum expected value of five for each cell, as described in section 5.2.5, is conservative [1]. If most of the cells have an expected value of five or more, the test should be reliable. If there is doubt about using the Chi-square test, the exact test (multinomial) may be computed [12]. The calculations for the exact test solution are usually very tedious.

Table 15.16 Patients Categorized by Severity of Disease Entered into Two Treatment Groups in a Clinical Study

	Very severe	Moderately severe	Mildly severe	Total
Treatment *A*	13	24	18	55
Treatment *B*	19	20	12	51
Total	32	44	30	106

Table 15.16 shows data from a clinical study in which patients entering the study were categorized according to the severity of disease. Severity was divided into three classes: very severe, moderately severe, and mildly severe. The categorization was made to ensure that the severity of disease was similar for patients in the two treatment groups. Thus, the question addressed by these data is "Is the severity of disease similar for patients entered into the two treatment groups?" or "Is there a relationship between 'treatment' and 'severity of disease'?" In a sense, this test is a confirmation of the randomization procedure used to assign patients to the two treatment groups. We would expect that, "on the average," the severity would be similar in Groups *A* and *B*.

The Chi-square calculation is similar to that for the fourfold (2×2) table (chapter 5). The *expected* values for each cell are obtained by multiplying the row and column totals corresponding to the cell, and dividing this result by the grand total (row total $\times$ column total/grand total). In the example in Table 15.16, this calculation needs to be done for only two cells (note the 2 d.f.), because the remaining four expected values can be obtained by subtraction from the fixed row and column totals. The sum of the expected values must equal the row and column totals of the raw data. In the table the expected value for the cell with 13 patients (Treatment *A*, very severe) is $(32)(55)/(106) = 16.60$. For the cell defined by Treatment *A*, moderately severe, the expected value is $(44)(55)/(106) = 22.83$. The expected values are shown in Table 15.17.

The Chi-square statistic is calculated according to Eq. (15.8).

$$\chi_2^2 = \frac{(13-16.60)^2}{16.60} + \frac{(24-22.83)^2}{22.83} + \frac{(18-15.57)^2}{15.57} + \frac{(19-15.4)^2}{15.40} + \frac{(20-21.17)^2}{21.17} + \frac{(12-14.43)^2}{14.43} = 2.54.$$

For significance at the 5% level, a value of 5.99 is needed for Chi-square with 2 d.f. (Table IV.5). Since the observed Chi-square is 2.54, we conclude that there is not sufficient evidence to show that severity and treatment are related; that is, the two treatment groups cannot be shown to differ with regard to the distribution of severity of disease.

Another example of an $R \times C$ table is shown in Table 15.18. This differs from the previous example in that we have three treatments each with a dichotomous outcome, rather than two treatments with three categories of outcome. The analysis tests for differences among the three treatments. This data is derived from a clinical study in which three treatments were randomly assigned to 60 patients. Only 54 patients successfully completed the study. Patients were classified as success or failure, depending on their response to treatment.

The analysis proceeds exactly as in the preceding example. The value of Chi-square with 2 d.f. is 7.76. Since the table Chi-square with 2 d.f. is 5.99, the treatments are significantly different. To test for differences suggested by the data (a posteriori tests), perform a Chi-square test for two

Table 15.17 Expected Values for the Data of Table 15.10

	Very severe	Moderately severe	Mildly severe	Total
Treatment *A*	16.60	22.83	15.57[a]	55
Treatment *B*	15.40[a]	21.17[a]	14.43[a]	51
Total	32	44	30	106

[a]Obtained by subtraction from total; see the text (e.g., $55 - 16.60 - 22.83 = 15.57$).

Table 15.18 Number of Successes and Failures Following Three Treatments

Treatment	Successes	Failures	Total
A	9	6	15
B	8	11	19
C	17	3	20
Total	34	20	54

treatments (a 1 d.f. test), but use the Chi-square cut-off point for 2 d.f., 5.99, for significance. For example, the Chi-square value for the comparison of Treatments *B* and *C* is 7.79, and Treatments *B* and *C* are significantly different (see Exercise Problem 13).

For a further discussion of multiple comparisons and other topics in the analysis of categorical data, the book *Statistical Methods for Rates and Proportions* by Fleiss [13] is highly recommended.

15.9.2 Fisher's Exact Test

In the Chi-square analysis of 2×2 contingency tables, if the expected values are too small, the Chi-square test may not be appropriate. Dichotomous data with small expected values are commonly encountered in pharmaceutical research, particularly in preclinical toxicology studies. For example, in preclinical animal carcinogenic studies, when comparing control and treatment groups with respect to some characteristic that occurs infrequently, the comparison of the frequencies may not be amenable to a Chi-square analysis. Fisher's exact test for 2×2 tables can be used to compute the exact probabilities. This test can be used, for example, to compare proportions for two independent groups (treatments), a binomial test, where expected values are very small.

Fisher's exact test makes use of the fact that the probability of a given configuration in a fourfold table with *fixed margins*[‡] can be computed using the *hypergeometric* distribution. The probability calculation will be described with reference to the notation in Table 15.19 to help clarify the procedure.

The probability of the values found in Table 15.19, given the four *fixed* margins, $(A + C)$, $(B + D)$, $(A + B)$, and $(C + D)$, is

$$\frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!} \tag{15.9}$$

The numerator of Eq. (15.9) is obtained by multiplying the factorials of the marginal totals. The denominator is the product of the factorials of the individual cells of the fourfold table, multipled by $N!$, the factorial of the total number of observations.

Table 15.20 shows data typically analyzed using the Fisher's exact test. One group of animals was administered a placebo preparation consisting of all components of the drug formulation with the exception of the active ingredient (placebo group). Another group of animals (drug group) was administered the drug formulation. After a fixed period of time, the incidence of a particular type of carcinoma was noted. The probability of the fourfold table shown in Table 15.20 with fixed margins (12, 14, 5, and 21) is calculated using Eq. (15.9).

$$\frac{5!21!12!14!}{26!1!4!11!10!} = 0.183.$$

[‡] Theoretically, Fisher's exact test is appropriate when marginal totals are fixed. In the example in Table 15.20, this means that before the initiation of the experiment, we decided to use 12 animals on placebo and 14 animals on drug; a total of five carcinomas will be observed in both groups. The latter result is clearly not under our control (although in some experiments, the marginal totals can be controlled). There exists some controversy whether data, in which two independent groups are to be compared (as in Table 15.19), where the margins are not fixed, are appropriate for Fisher's exact test. However, the test is commonly used to analyze such data.

Table 15.19 Fourfold Table as an Aid to the Calculation of Fisher's Exact Test

		Column		
		I	**II**	**Total**
Row	I	A	C	$A + C$
	II	B	D	$B + D$
Total		$A + B$	$C + D$	$A + B + C + D = N$

Thus, the probability of the results shown in Table 15.14 are *not* very unlikely. However, this is not the entire statistical test. In Fisher's test, we compute the probability of the observed configuration *plus* the probabilities of *all less likely* configurations (a cumulative probability). If the sum of the observed configuration plus all less likely configurations is *less than* α (0.05, for example), we conclude that the rows and columns (treatment and carcinoma) are *not* independent; that is, the treatments differ with respect to the incidence of carcinomas. If the sum of these probabilities exceeds α (0.05, for example), we accept the null hypothesis of independence, concluding that the evidence is not sufficient to conclude that the treatments differ. In the example (Table 15.20), the sum of probabilities obviously exceeds 0.183. (The probability of the observed table is 0.183). Therefore, there is insufficient data to show conclusively that the incidence of carcinoma is greater in the drug group compared to the placebo group.

To clarify this procedure further, we will work out an example in more detail based on the data shown in Table 15.21. These data are similar to that in Table 15.20, except that no carcinomas were observed in the placebo group and five were observed in the drug group. Thus, the marginal totals are the same in Tables 15.20 and 15.21. The probability of Table 15.21 is calculated as before, using Eq. (15.9).

$$\frac{5!21!12!14!}{26!0!5!12!9!} = 0.03043.$$

In order to assess the possible "statistical" significance of this table, we must compute the probability of all less likely configurations as discussed above. What constitutes less likely tables is not always obvious without some "trial and error" calculations. With experience, good, educated guesses can be made as to what constitutes a less likely table.

If a configuration is mistakenly chosen with a higher probability than the observed table, the calculation is discarded. Possible "less likely" tables are shown in Table 15.22 with the probability of each table. The only table with a lower probability than the observed table (Table 15.21) is the one with all five carcinomas appearing in the placebo group.

$$\frac{5!21!12!14!}{26!5!0!7!14!} = 0.01204.$$

Table 15.20 Incidence of Carcinoma in Drug- and Placebo-Treated Animals: Example 1

	Number of animals with:		
	Carcinoma	**No carcinoma**	**Total**
Placebo	1	11	12
Drug	4	10	14
Total	5	21	26

Table 15.21 Incidence of Carcinoma in Drug- and Placebo-Treated Animals: Example 2

	Carcinoma present	Carcinoma absent	Total
Placebo	0	12	12
Drug	5	9	14
Total	5	21	26

The sum of the probabilities of the observed table and all less likely (or equally likely) tables is $0.03043 + 0.01204 = 0.0425$. Therefore, Table 15.15 is "significant" at the 5% level ($p < 0.05$); the drug appears to result in an increased incidence of carcinomas.

Note that Fisher's exact test requires that the probabilities of tables with fixed margins be computed for all possible configurations. If we calculate all possible configurations, the sum of the probabilities of the different tables would be equal to 1. Among all of these probabilities will be the probability of the observed table, in addition to possible probabilities equal to or smaller than that of the observed table. If the sum of these probabilities is less than or equal to 0.05, for example, the treatments are said to be "significantly" different at the 5% level, in the context of the present example.

The computations are often very tedious. For cases where the computations are unduly long and tedious, the use of computer programs or tables to determine significance points in fourfold tables are recommended [14].

15.9.3 Fourfold Tables with Related Samples

The examples of 2×2 contingency tables previously discussed in this chapter and chapter 5 have involved the comparison of proportions or frequencies in two or more *independent* groups. A similar problem that occurs less frequently in pharmaceutical research is the comparison of two groups where the observations are *related*, also known as matched pairs. For example, Table 15.23 shows the results of two versions of an allergy test, *A* and *B*, applied to 50 persons. The test reagents were applied at the same time at different sites for each subject, and either a positive or negative reaction was observed. In this design, the total sample size is specified in advance, but the marginal totals are not fixed. We cannot anticipate the total positive and negative for test *B* in Table 15.23, for example. In the previous example, the size of the two treatment groups can be fixed in advance. Note that in this example each person is subjected to both treatments (allergy tests). In the previous examples of fourfold tables, each person is subjected to a single treatment and a dichotomous response is observed (e.g., *cured* or *not cured*).

The objective of this experiment is contained in the question: "Does the proportion of positive reactions for test *A* differ from that for test *B*?" (i.e., H_0: $p_a = p_b$, where p_a and p_b are the proportion of positive reactions in tests *A* and *B*, respectively). Note that test *A* has 32 positive reactions ($23 + 9$), and test *B* has 29 positive reactions ($23 + 6$). It can be shown that the statistical test for the equality of positive reactions for the two tests is equivalent to the test for the equality of the counts in the diagonal cells designated by an *a* in Table 15.23 (9 and 6) [13]. The counts (or proportions) in these two cells represent the *untied* responses (*positive A* and *negative B*, and *negative A* and *positive B*, 9 and 6, respectively). The counts in the other two cells do not differentiate the two allergy tests. For example, the upper left-hand cell shows the 23 patients who were positive on *both* tests.

Table 15.22 Some "Unlikely" Tables with Margins Identical to Table 15.21

	Carcinoma present	Carcinoma absent	Total	Carcinoma present	Carcinoma absent	Total
Placebo	5	7	12	4	8	12
Drug	0	14	14	1	13	14
Total	5	21	26	5	21	26
	Probability = 0.0120			Probability = 0.1054		

Table 15.23 Frequency of Positive and Negative Reactions to Two Allergy Tests Applied to Two Sites in 50 Persons

		Test *B*		
		Positive	**Negative**	**Total**
Test *A*	Positive	23	9[a]	32
	Negative	6[a]	12	18
	Total	29	21	50

[a]Patients who were positive on one test and negative on the other test.

Under the null hypothesis that the probability of a positive reaction is equal for both tests, the diagonal counts, 9 and 6, should be equal. The test of significance is a binomial test, as in the sign test (sect. 15.2). In the latter procedures, the observed proportion is compared to 0.5, the expected proportion if both treatment groups have an equal probability of being positive. The statistical test in this example makes use of the normal approximation to the binomial distribution [Eq. (15.1)].

$$Z = \frac{\left|\text{observed proportion} - 0.5\right| - 1/(2N)}{\sqrt{(0.5)(0.5)/N}}. \tag{15.1}$$

If Z is greater than 1.96, the difference is significant at the 5% level and we conclude that the probability of a positive response is different for the comparative treatments. (As in other examples where the normal approximation to the binomial is used, the sample size should be sufficiently large, approximately 10 for this test.) The observed proportion in the example in Table 15.23 is $9/15 = 0.60$. $N = 15$, the number of untied pairs. Therefore,

$$Z = \frac{|0.60 - 0.5| - 1/30}{\sqrt{(0.5)(0.5)/15}} = 0.52.$$

Since Z is not equal to or greater than 1.96, the difference is not significant at the 5% level. The difference is not sufficiently large to conclude that the two tests differ with regard to the frequency (proportion) of positive responses. This test is also known as McNemar's test.

The data shown in Table 15.23 can also answer a different question that requires a different analysis. In the previous example, we inquired if the proportion of positive reactions was different in the two tests. Another question that is often relevant to such data is: "Are the allergy tests independent, that is, is the probability of a positive response for test *B* independent of the outcome for test *A*?" This question implies that if *A* and *B* are independent, there should be an equal proportion of positive results to test *A* in both patients with a positive test to *B* and in patients with a negative test to *B*. Table 15.24 shows the *expected results* if, in fact, tests *A* and *B* are independent. Note that the expected proportion of positive *A*'s in patients who had a positive test for *B* is 0.64, 18.56/29. This is the same expected proportion of positive *A*'s as that for patients who had a negative test for *B*, 13.44/21.

Table 15.24 Expected Values from Table 15.17 if Allergy Tests *A* and *B* Are Independent

		Test *B*		
		Positive	**Negative**	**Total**
Test *A*	Positive	18.56	13.44	32
	Negative	10.44	7.56	18
	Total	29	21	50

Table 15.25 Fourfold Table for Treatment and Placebo

	Improvement		
Treatment	**None**	**Some or marked**	**Total**
Active	13	28	41
Placebo	29	14	43
Total	42	42	84

The test for independence is the same Chi-square test as that used for the comparison of proportions in two independent samples, although the question to be answered is different (see sect. 5.2.5). We apply Eq. (15.8)

$$\chi_1^2 = \sum \frac{(O-E)^2}{E}, \tag{15.8}$$

where O is the observed count and E is the expected count. The expected values for the Chi-square test are shown in Table 15.24 (see sect. 5.2.5 for calculation of expected values). Applying Eq. (15.8) to the data of Tables 15.23 and 15.24 (including the continuity correction discussed in sect. 5.2.5), we have

$$\chi_1^2 = \frac{(4)^2}{13.44} + \frac{(4)^2}{18.56} + \frac{(4)^2}{7.56} + \frac{(4)^2}{10.44} = 5.70.$$

To obtain significance at the 5% level, a Chi-square value of 3.84 is needed (Table IV.5). Clearly, the test is significant and we conclude that the results of this test warrant rejection of the null hypothesis (i.e., the results of tests A and B are dependent). This significant result suggests that tests A and B are related; a positive test for A is associated with a positive test for B; and a negative test for A is associated with a negative test for B.

15.9.4 Analysis of Combined Sets of 2 × 2 Tables

Two situations may arise in which the analysis of combined fourfold tables is needed. Consider a clinical study in which two treatments are to be compared with regard to a dichotomous variable where the data are collected from more than one center. Rather than pooling all the data to form one combined table, the analysis is performed with the data stratified by center. In a second example, a study may be performed at a single center, but there may be a variable within the center that needs further clarification with respect to interpretation of the results. The data are then stratified by this variable. Koch and Edwards [15] give an example of a clinical study at a single center comparing a test drug and placebo in a study of arthritis. The outcome of the treatment is dichotomized into either no improvement or (some or marked) improvement. The overall results are shown in Table 15.25. Table 15.26 stratifies Table 15.25 into two groups, results for males and females. The following discussion summarizes part of their presentation (for more detail, see Ref. [15]).

Note that males appear to be less responsive than females to both active drug and placebo. If the distribution of males and females to treatment groups is unbalanced, the experimental results can be biased.

The Chi-square test for significance for the data of Table 15.25 is 10.7 with a correction factor, and 12.3 without the correction factor. The Mantel–Haenszel method [16] tests for significance, taking into account the sex-adjusted response (Table 15.26).

If treatments are equally effective, the expected value of n_{111} and n_{211} in Table 15.25 are

$$E(n_{111}) = m_{111} = \frac{(n_{11+})(n_{1+1})}{n_1}$$

$$E(n_{211}) = m_{211} = \frac{(n_{21+})(n_{2+1})}{n_2}.$$

Table 15.26 Table 15.25 with Two Subgroups

		Improvement		
Sex	**Treatment**	**None**	**Some or marked**	**Total**
Female	Test drug	$n_{111} = 6$	$n_{112} = 21$	$n_{11+} = 27$
Female	Placebo	$n_{121} = 19$	$n_{122} = 13$	$n_{12+} = 32$
Female total		$n_{1+1} = 25$	$n_{1+2} = 34$	$n_1 = 59$
Male	Test drug	$n_{211} = 7$	$n_{212} = 7$	$n_{21+} = 14$
Male	Placebo	$n_{221} = 10$	$n_{222} = 1$	$n_{22+} = 11$
Male total		$n_{2+1} = 17$	$n_{2+2} = 8$	$n_2 = 25$

The variances of n_{111} and n_{211} are

$$\text{Var}\,(n_{111}) = \frac{n_{11+}n_{12+}n_{1+1}n_{1+2}}{[(n_1^2)(n_1 - 1)]}$$
$$= \frac{(27)(32)(25)(34)}{(59)^2(58)} = 3.63748$$

$$\text{Var}(n_{211}) = \frac{n_{21+}n_{22+}n_{2+1}n_{2+2}}{[(n_2^2)(n_2 - 1)]}$$
$$= \frac{(14)(11)(17)(8)}{(25)^2(24)} = 1.39627.$$

The Mantel–Haenszel statistic is calculated as

$$\frac{\left[\sum_{h=1}^{2}(n_{h1+}n_{h2+}/n_h)(p_{h11} - p_{h21})\right]^2}{\sum_{h=1}^{2} v_{h11}}, \tag{15.9}$$

where $h = 1, 2$ and $p_{hi1} = (n_{hi1}/n_{hi+})$ the proportion of patients in each sex and treatment group who show no improvement.

$$p_{111} = 6/27$$

$$p_{211} = 7/14$$

$$p_{121} = 19/32$$

$$p_{221} = 10/11$$

For the data of Table 15.26, the calculation of the Mantel–Haenszel statistic (eq. 15.9) is

$$Q_{MH} = \frac{[\{(27)(32)/59\}(6/27 - 19/32) + \{(14)(11)/25\}(7/14 - 10/11)]^2}{3.63748 + 1.39627} = 12.59.$$

Q_{MH} is distributed approximately as Chi-square with 1 d.f. Therefore, the conclusion is that after adjustment for sex differences, the treatments are significantly different ($p < 0.05$).

This analysis summarizes an elementary but common occurrence in the analysis of clinical studies. For more detail of the application of the Mantel–Haenszel statistic, see Refs. [13,15].

Table 15.27 Treatments with Binomial Outcome in Randomized Blocks[a]

	Treatment			
Subject	**I**	**II**	**III**	B_j
1	1	1	0	2
2	0	0	0	0
3	1	0	1	2
4	1	0	1	2
5	0	1	1	2
6	1	0	1	2
7	1	0	0	1
8	1	0	0	1
9	1	0	1	2
10	1	0	0	1
T_i	8	2	5	15

[a] 1 = success, 0 = failure.

15.9.5 Randomized Blocks with Binomial Outcome

For data with a binomial outcome that is in the form of a randomized block, the following test to compare treatments [17] may be used:

Compute:

$$Q = \frac{\left[c(c-1)\sum(T_i^2) - (c-1)N^2\right]}{\left[cN - \sum(B_j^2)\right]},$$

where c is the number of treatments (columns), T_i the total for treatment i, B_j the total for block j, and N is the grand total.

For large samples, Q has an approximate Chi-square distribution with $c - 1$ d.f.

Example 3. Ten subjects were treated with a topical product for a fungus infection. Subjects were evaluated as cured (1) or not cured (0) (Table 15.27).

$$Q = \frac{\left[c(c-1)\sum(T_i^2) - (c-1)N^2\right]}{\left[cN - \sum(B_j^2)\right]}$$

$$Q = \frac{\left[3(2)(64+4+25) - 2(15)^2\right]}{(3\times 15 - 27)} = \frac{108}{18} = 6.$$

The tabled value of Chi-square with 2 d.f. at the 5% level is 5.99. Therefore, we can conclude that the differences are significant at the 0.05 level. (Treatment 1 is different from Treatment 2.)

15.10 NONPARAMETRIC TOLERANCE INTERVAL

If the data set appears to be non-normal, the usual tolerance interval calculation assuming a normal distribution may be inappropriate (see sect. 5.1). In this case, a nonparametric tolerance interval can be constructed. The nonparametric interval can be considered conservative, and would be wider, on average, than that usually calculated assuming normality, if the data truly are normal. The computation quantifies the intuition [18] that most future observations would lie within the minimum and maximum of an observed sample from the distribution. The calculation is as follows [18]. Given a sample of size n from a distribution, we can state the probability of the proportion of samples that are within the minimum and maximum values

observed. Let $p =$ the proportion of samples within the maximum and minimum. Let Q be the probability space covered by min, max; n is the sample size.

$$P(Q > p) = 1 - n(p)^{n-1} + (n-1)p^n$$

$P(Q > p)$ is the probability that the minimum and maximum values will cover a proportion p of all values in the distribution.

An example should clarify the calculation. Suppose that we assay 50 individual tablets randomly chosen from a batch, with a minimum of 96% and a maximum of 103%. Furthermore, we have reason to believe that the data are not normally distributed. We wish to compute a tolerance interval that will give a probability that p proportion of the tablets assay between 96% and 103%. In this example, $n = 50$. Suppose, we wish to set the proportion of tablets within 96 to 103 to be 95%. We then calculate the probability.

$$P(Q > p) = 1 - 50(0.95)^{50-1} + (50-1)0.95^{50} = 0.72.$$

Therefore, we say that the probability is 0.72 (72%) that at least 95% of the tablets in the batch are within 96% to 103%.

We might want to compare this result with the tolerance interval assuming a normal distribution (sect. 5.1). The average is 100.2% and the standard deviation of the 50 tablets is 2.5%. Referring to Table IV.19 in the appendix, a 75% probability tolerance interval containing 95% of the tablets is

$$100.2 \pm 2.138 \times 0.025 = 100.2 \pm 5.3 = 94.9 - 105.5.$$

That is, the probability is 75% that at least 95% of the tablets are between 94.9% and 105.5%.

KEY TERMS

ANCOVA
Attribute
Categorical data
Confidence interval
Contingency table (R x C table)
Continuous data
Distributions
Efficiency
Fisher's exact test
Friedman's test
Hypergeometric distribution
Independence
Interval or ratio scale
Kruskal–Wallis test
Mantell–Haenszel test
McNemar's test
Multinominal distribution
Nominal data
Normal approximation
Ordered data
Ordinal data
Quade test
Rating scale
Run
Runs test
Sensitive
Sign test
Ties
Tolerance interval
Wilcoxon rank sum test
Wilcoxon signed rank test

EXERCISES

1. Perform a t test to compare treatments for the data from Table 15.3. Compare the results of this test to the nonparametric test presented in the text.
2. The following data were observed comparing two assays using 12 batches of material:
   (a) Use the sign test to determine if the two tests are different.
   (b) Compare the two tests (A and B) using the t test.

Batch	Test *A*	Test *B*
1	8.1	9.0
2	9.4	9.9
3	7.2	8.0
4	6.3	6.0
5	6.6	7.9
6	9.3	9.0
7	7.6	7.9
8	8.1	8.3
9	8.6	8.2
10	8.3	8.9
11	7.0	8.3
12	7.7	8.8

3. Use the Wilcoxon signed rank test to compare the two assay methods to determine if the methods are significantly different for the data in Exercise Problem 2. Use Table IV.13 and the normal approximation.
4. Blood glucose uptake for corresponding halves of rat diaphragms for compounds *A* and *B* are as follows (adapted from Ref. [2]):

	Rat								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
A	9	9.5	5.7	3.9	6.7	5	8.6	3	8
B	8	9.7	5.1	3.6	7.1	5	8.4	4.2	7.1

Use a nonparametric procedure to compare the two compounds.

5. Twenty patients were randomly allocated to two treatment groups, 10 patients per group. The following data are the change in serum chloride after treatment.

Treatment *A*	Treatment *B*
4.3	6.1
6.2	0.9
4.4	0.7
8.2	0.8
0.5	1.3
2.6	3.1
4.2	1.9
4.1	3.9
5.6	2.1
3.4	0.1

Test for treatment differences using a nonparametric test and a *t* test.

6. Dissolution is compared for three experimental batches with the following results (each point is the time in minutes to 50% dissolution for a single tablet):
Batch 1: 15, 18, 19, 21,23, 26
Batch 2: 17, 18, 24, 20
Batch 3: 13, 10, 16, 11,9
Is there a significant difference among batches?

7. A bioavailability study was conducted in which three products were compared: a standard product and two new formulations, *A* and *B*. The peak blood concentrations were as follows:

Subject	Standard	*A*	*B*
1	14	12	17
2	12	18	9
3	11	17	8
4	17	15	14
5	20	16	16
6	16	12	13
7	14	11	10
8	16	16	10
9	18	17	19
10	15	10	8
11	22	15	15
12	14	13	14

Use Friedman's test to determine if there is a difference among the three treatments.

8. In a test for pain relief, two drugs are compared where the outcome is 0, 1, or 2, where 0 = no relief, 1 = partial relief, 2 = complete relief. With drug *A*, 50 had a score of 0, 50 scored 1, and 75 scored 2. With drug *B*, 20 had a score of 0, 60 scored 1, and 60 scored 2. Use a Chi-square test to compare drugs *A* and *B*. How would you interpret a significant effect?

9. The following fourfold table was constructed from data for inspection of 1000 tablets in quality control.
   (a) Are "specks" and "capping" independent?
   (b) Are the proportion of tablets specked and capped different in this batch of tablets?

		Capped	
		Yes	No
Specked	Yes	13	45
	No	18	924

§10. In a preclinical study, the following incidence of tumors was observed in control and treated animals:
Controls: 0 of 12 animals
Treated: 5 of 14 animals
Use Fisher's exact test to determine if the incidence is significantly different in the two groups. Compare the results to a Chi-square test with continuity correction.

11. The following assay results were observed from sequential readings from a control chart. Using the runs test, determine if these values conform to a "random" sequence. Use a two-sided test. What would be your conclusion if the test were one-sided? 300.1, 300.5, 300.7, 308.2, 304.4, 303.9, 302.1, 303.1, 300.9, 303.4, 305.6, 306.2, 304.1, 306.1, 306.8, 301.3, 304.3, 301.9, 304.2, 302.6

12. Confirm that the corrected χ^2 is 7.0 by computing the correction for ties for the analysis of the data in Table 15.10.

13. For the 3 × 2 table at the end of section 15.9.1 (Table 15.18), compute the χ^2 value for the entire table and for the comparison of Treatments *B* and *C*.

§§ Optional, more advanced problems.

14. Analyze the following data, using the combined data from two centers. Use the Mantel–Haenszel test.

	Success	Failure	Total
Center I			
Drug *A*	12	6	18
Drug *B*	9	9	18
Total	21	15	36
Center II			
Drug *A*	14	3	17
Drug *B*	9	11	20
Total	23	14	37

Are the two treatments significantly different?

15. Compute the parametric two-way ANOVA and confidence intervals for the data of Table 15.6, using a log transformation.
16. In section 15.4.1, show that the comparison of Treatment 1 with Treatment 2 for Sequence 1 and Sequence 2 is equal to twice the period effect (no carryover).
17. Compute the tests for sequence and period effects for Table 15.9; use Eq. (15.4).
18. Perform a two-way analysis of variance on the data of Table 15.11. Assuming no interaction, what is the probability associated with the *F* test (tablet press MS/error MS).
19. Perform a two-way ANOVA on ranks in Table 15.11. Show that Fisher's LSD is the same as the nonparametric multiple comparison.

REFERENCES

1. Snedecor GW, Cochran WG. Statistical Methods, 7th ed. Ames, IA: State University Press, 1980.
2. Wilcoxon F, Wilcox RA. Some Rapid Approximate Statistical Procedures. Pearl River, NY: Lederle Laboratories, 1964.
3. Hollander M, Wolfe DA. Nonparametric Statistical Methods. New York: Wiley, 1973.
4. Steinijens VW, Diletti E. Statistical analysis of bioavailability studies: parametric and nonparametric confidence intervals. Ear J Clin Pharmacol 1983; 24:127–136.
5. Jones B, Kenward MC. Design and Analysis of Cross-Over Trials. London: Chapman and Hall, 1989.
6. Cornell RG. Evaluation of bioavailability data using nonparametric statistics. In: Albert KS, ed. Drug Absorption and Disposition, Statistical Considerations. Washington, DC: American Pharmaceutical Association, Academy of Pharmaceutical Sciences, 1980:51.
7. Koch G. The use of nonparametric methods in the statistical analysis of the two period change-over design. Biometrics 1972; 28:577–584.
8. Wallenstein S, Fisher AC. The analysis of the two-period repeated measurements crossover design with application to clinical trails. Biometrics 1977; 33:261.
9. Conover W. Practical Nonparametric Statistics, 2nd ed. New York: John Wiley and Sons, 1980.
10. Sprent P. Applied Nonparametric Statistical Methods. New York: Chapman and Hall, 1989.
11. Quade D. J Am Stat Assoc 1967:1187.
12. Sokal RR, Rohlf FJ. Biometry. San Francisco: W.H. Freeman, 1969.
13. Fleiss JL. Statistical Methods for Rates and Proportions. New York: Wiley, 1980.
14. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969.
15. Koch GG, Edwards S. Summarization analysis, and monitoring of adverse experiences. In: Peace KE, ed. Statistical Issues in Drug Research and Development. New York: Marcel Dekker, 1990: 19–161.
16. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Canc Inst 1959; 22:719.
17. Sprent P. Applied Nonparametric Statistical Methods. New York: Chapman and Hall, 1989:128.
18. Rice JA. Mathematical Statistics and Data Analysis. Pacific Grove, CA: Wadsworth and Brooks/Cole, Statistics/Probability Series, 1988.

16 Optimization Techniques and Screening Designs*

The* optimization of pharmaceutical formulations with regard to one or more attributes has always been a subject of importance and attention for those engaged in formulation research. Product formulation is often considered an art, the formulator's experience and creativity providing the "raw material" for the creation of a new product. Given the same active ingredient and a description of the final marketed product, two different scientists will very likely concoct different formulations. Certainly, human input is an essential ingredient of the creative process. In addition to the *art* of formulation, techniques are available that can aid the scientist's choice of formulation components that will optimize one or more product attributes. These techniques have been traditionally applied in the chemical and food industries, for example, and in recent years have been applied successfully to pharmaceutical formulations. In this chapter, we describe the application of factorial designs (and modified factorials) and simplex lattice designs to formulation optimization. When the effects of factors on a pharmaceutical process or response are unknown, the use of screening designs to estimate factor effects may be indicated.

16.1 INTRODUCTION

The pharmaceutical scientist has the responsibility to choose and combine ingredients that will result in a formulation whose attributes conform with certain prerequisite requirements. Often, the choice of the nature and quantities of additives (excipients) to be used in a new formulation is based on experience, for example, similar products previously prepared by the scientist or his or her colleagues. To break habits based on experience and tradition is difficult. Although there is much to be said for the practical experience of many years, we often become caught in the web of the past. The application of formulation optimization techniques is relatively new to the practice of pharmacy. When used intelligently, with common sense, these "statistical" methods will broaden the perspective of the formulation process.

Although several optimization procedures are available to the pharmaceutical scientist, a few frequently used methods will be presented in this chapter. The objective is to produce a mathematical model that describes the responses. In general, the procedure consists of preparing a series of formulations, varying the concentrations of the formulation ingredients in some systematic manner. These formulations are then evaluated according to one or more attributes, such as hardness, dissolution, appearance, stability, taste, and so on. Based on the results of these tests, a particular formulation (or series of formulations) may be predicted to be optimal. The ''proof of the pudding," however, is actually to prepare and evaluate the predicted *optimal* formulation.

If the formulation is optimized according to a single attribute, the optimization procedure is relatively uncomplicated. To optimize on the basis of two or more attributes, dissolution and hardness, for example, may not be possible. The formulation that is optimal for one attribute very well may be different from the formulation needed to optimize other attributes. In these cases, a compromise must be made, depending on the relative importance of each attribute. The final formulation, therefore, is suitably modified to attain an acceptable performance of all relevant attributes, if possible. We will discuss the optimization procedure based on a single attribute. More complex situations may require more complex designs, and the advice of an experienced statistician is recommended in these cases. Therefore, the use of the term, "optimization" may be a misnomer. An optimal response may not be a single response, but a region of responses that

* This is an advanced topic.

satisfy the requirements of the formulation. Once such a region is defined, the desired response may be defined using a range of factors.

In general, an advanced understanding of statistics is not necessary. One should be familiar with the following concepts as described elsewhere in this book.

16.1.1 Planning Experiments

Common sense should prevail. Design and choice of variables are discussed later in this chapter. In most cases, we have a reasonable idea of which variables are important, and their effective ranges. But, we may be surprised. If everything were known, we would not have to experiment. Also, we should be careful not to neglect potentially important variables. Screening designs may be useful if little is known of the system.

16.1.2 Variables

Variables may be considered as Independent and Dependent (X, Y). Dependent variables (Y) are outcome variables (e.g., dissolution). Independent variables (X) are set in advance (e.g., lubricant level). Variables can be continuous or discrete. The number of experiments should be kept at a reasonable level. The more variables used, the more knowledge is gained, but expense and time should be taken into consideration.

16.1.3 Variability or Experimental Error

It is important to have an idea about variability of response (Y) and/or "predicted response." Replication is typically needed to estimate variability, but this adds time and cost to the study. Estimates of variance can be obtained from replication, from ANOVA or from experience.

16.1.4 Regression

For our purposes, regression is used to predict Responses, and/or to describe relationships. Either simple linear or multiple regression may be used to obtain optimized systems. We derive a response equation from the data (as described in this chapter), and predict a response within the bounds of the fixed independent variables, X. Prediction outside of the bounds of the independent variables is unreliable. Consider the following example.

Suppose that the theoretical response relationship (Y as a function of X_1 and X_2, where we have two independent variables) is $Y = 5 + 6X_1 + 7X_1^2 + 3X_2$. We obtain six values of Y as follows:

X_1	X_2	Y
1	1	21
2	1	48
1	2	24
2	2	57
3	1	89
1	3	45

Using multiple regression, we obtain the following equation relating Y to the independent variables.

$$Y = -7 + 7.2X_1 + 7X_1^2 + 11.4X_2$$

This works well within the experimental space. But predictions outside are questionable. For example, if $X_1 = 4$ and $X_2 = 4$

Predicted = 179.4
Actual = 153

16.2 OPTIMIZATION USING FACTORIAL DESIGNS

The basic principles of factorial designs have been presented in chapter 9. In factorial designs, levels of factors are independently varied, each factor at two or more levels. The effects that can be attributed to the factors and their interactions are assessed with maximum efficiency in factorial designs. Also, factorial designs allow for the estimation of the effects of each factor and interaction, unconfounded by the other experimental factors. Thus, if the effect of increasing stearic acid by 1 mg is to decrease the dissolution by 10%, in the absence of interactions, this effect is independent of the levels of the other factors. This is an important concept. If the levels of factors are allowed to vary haphazardly, as in an undesigned experiment, the observed effect due to any factor is dependent on the levels of the other varying factors. Generalities, or predictions, based on results of an undesigned experiment will be less reliable than those that would be obtained in a designed experiment, in particular, a factorial design. Screening designs use less runs, and estimate the main effects of factors. The latter part of this chapter will introduce screening designs. These designs are useful when a relatively large number of factors may affect the response or process. From a regulatory viewpoint, the data derived from factorial designs can be useful to predict responses when confronted with formulation or manufacturing modifications.

The optimization procedure is facilitated by construction of an equation that describes the experimental results as a function of the factor levels. A *polynomial* equation can be constructed, in the case of a factorial design, where the coefficients in the equation are related to the effects and interactions of the factors. For the present, we will restrict our discussion to factorial designs with factors at only two levels, called 2^n factorials, where n is the number of factors (see chap. 9). These designs are simplest and often are adequate to achieve the experimental objectives. These designs estimate only linear effects. That is, if there is a curved response as a function of factor levels or combination, such effects will be missed. Sometimes, use of these smaller designs is imperative, for the sake of economy. Increasing the number of factor levels dramatically increases the number of formulations that are needed to complete the design. With a large number of factors, even designs where factors are restricted to two levels may result in a very large number of formulations to be prepared and tested. In such cases, *fractional* factorial designs may be used. Some information is lost when using fractional factorial designs, but one-half, one-fourth, or less of the formulations are needed compared to those needed to run a full factorial design. A brief description of fractional factorial designs is presented in section 9.5. The theory and construction of these designs are presented in detail in *The Design and Analysis of Industrial Experiments*, edited by Davies [1]. Also see Ref. [2] for an example of optimization applied to an HPLC analytical method.

As noted above, the optimization procedure is facilitated by the fitting of an empirical polynomial equation to the experimental results. The equation constructed from a 2^n factorial experiment is of the following form:

$$\begin{aligned} Y = {} & B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_{12}X_1X_2 \\ & + B_{13}X_1X_3 + B_{23}X_2X_3 + \ldots + B_{123}X_1X_2X_3 + \ldots \end{aligned} \tag{16.1}$$

where Y is the measured response, X_i is the level (e.g., concentration) of the ith factor, B_i, B_{ij}, B_{ijk}, . . .represent coefficients computed from the responses of the formulations in the design, as will be described below. (B_0 represents the intercept.)

For example, in an experiment with three factors, each at two levels, we have eight formulations, a total of eight responses. The eight coefficients in Eq. (16.1) will be determined from the eight responses in such a way that each of the responses will be exactly predicted by the polynomial equation. For the present, to illustrate this concept we will look at the problem in reverse. Suppose that we already have an equation to predict the experimental results derived from a factorial design as follows:

$$\begin{aligned} Y = {} & 5 + 2(X_1) + 3(X_2) + X_3 - 0.6(X_1X_2) - 0.4(X_1X_3) \\ & + 0.7(X_2X_3) + 0.12(X_1X_2X_3) \end{aligned} \tag{16.2}$$

From Eq. (16.2), we can reconstruct the original data from the 2^3 experiment. Suppose that the levels (in mg) of the three factors in the design were as follows:

	Low level	High level
X_1 = stearate	0	2
X_2 = colloidal silica	0	1
X_3 = drug	0	5

Based on Eq. (16.2), the formulation with all factors at the low level will have a response of five. All factors are equal to 0, and all terms containing X_1, X_2, or X_3 are equal to 0. If X_1 is at the high level (2 mg), and X_1, and X_3 are at the low level (0), the predicted response is $Y = 5 + 2(X_1) = 5 + 2(2) = 9$. All other terms are equal to 0. If X_1 and X_2 are at the high level, and X_3 is at the low level, the response is

$$5 + 2(X_1) + 3(X_2) - 0.6(X_1 X_2) = 5 + 2(2) + 3(1) - 0.6(2)(1) = 10.8.$$

The results for all eight combinations (formulations) as predicted from Eq. (16.2) are shown in (Table 16.1).

Table 16.1 shows the results of the factorial experiment that were used to construct Eq. (16.2). The practical, more realistic problem is to construct the polynomial equation, given the experimental results. To solve this problem, we find the solution to eight equations with eight unknowns [the unknowns are the eight coefficients in Eq. (16.2)]. For example, in formulation 1 (Table 16.1),

$$X_1 = X_2 = X_3 = 0.$$

Substituting $X_1 = X_2 = X_3 = 0$ into the general equation [Eq. (16.1)] results in

$$Y = B_0(\text{all other terms are } 0).$$

Since the response (Y) for formulation 1 (where $X_1 = X_2 = X_3 = 0$) is equal to 5,

$$Y = B_0 = 5.$$

This is the simple solution for the first of the simultaneous equations.
In the second formulation, $X_1 = 2$, X_2 and X_3 are equal to 0 and Eq. (16.1) reduces to

$$Y = B_0 + B_1 X_1(\text{all other terms are } 0). \tag{16.3}$$

Table 16.1 Results of the 2^3 Factorial Experiment That Led to the Construction of the Polynomial Equation (16.2)

	Factor level			
Formulation	X_1	X_2	X_3	Predicted response, Y
1	0	0	0	5
2	2	0	0	9
3	0	1	0	8
4	2	1	0	10.8
5	0	0	5	10
6	2	0	5	10
7	0	1	5	16.5
8	2	1	5	16.5

The response, Y, for formulation 2 is 9 (Table 16.1). We can solve for B_1, using Eq. (16.3) ($B_0 = 5$ and $X_1 = 2$)

$$9 = 5 + B_1(2) \quad B_1 = 2.$$

This procedure is continued, until we solve for all coefficients, B_i, B_{ij}, B_{ijk}, and so on.

In the example above, the solution for the coefficients for the polynomial equation is very simple, because the low level of all factors is zero. In general, the solution would be more difficult if the low level of all factors is not equal to zero. However, the general solution for the polynomial coefficients is not difficult for 2^n factorial designs, because of the independence (orthogonality) inherent in factorial designs. The first step in the solution is to code the levels of the factors so that the high level of each factor is $+1$, and the low level of each factor is -1. This procedure requires a transformation of each of the three variables, X_1, X_2, and X_3 to X_1', X_2', and X_3',respectively, as follows:

For X_1, *let* $X_1' = X_1 - 1$. Note that when $X_1 = 2$ (the high level), $X_1' = +1$, and when $X_1 = 0$ (the low level), $X_1' = -1$.

$$For\ X_2,\ let\ X_2' = 2X_2 - 1.$$

$$For\ X_3,\ let\ X_3' = \frac{2X_3 - 5}{5}.$$

In general, the formula for the transformation is

$$\frac{X - \text{the average of the two levels}}{\text{one-half the difference of the levels}}. \tag{16.4}$$

After the transformation, the levels of the factors are as shown in Table 16.2 (see also chap. 9).

Table 16.2 also contains "transformed" values for the interactions, represented by $+1$ or -1. These values are obtained by multiplying the values in the appropriate columns of X_1, X_2, and X_3. For example, in formulation 1, X_1X_2 is represented by $+1$, the product of -1 for X_1 and -1 for X_2 [$X_1X_2 = (-1)(-1) = +1$]. $X_1X_2X_3$ is represented by the product of $(-1)(-1)(-1) = -1$, derived from the values in the columns headed by X_1, X_2, and X_3. (See also chap. 9 to clarify this procedure.) The "total" column contains only the value $+1$, and is used to calculate the intercept, B_0.

The coefficients for the polynomial equation (16.1) are calculated as $\Sigma XY/8$ ($\Sigma XY/2^n$, in general), where X is the value ($+1$ or -1) in the *column* appropriate for the coefficient being calculated, and Y is the response. An example should make the calculation clear. For the coefficient corresponding to X_1 (B_1), the calculation is performed as follows. We multiply each

Table 16.2 Transformed Levels of Factors Showing Signs to Be Used to Determine Effects and Polynomial Coefficients

Formulation	X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	$X_1X_2X_3$	Total	Y
1[a]	−1	−1	−1	+1	+ 1	+1	−1	+1	5
2	+1	−1	−1	−1	−1	+1	+1	+1	9
3	−1	+1	−1	−1	+1	−1	+1	+1	8
4	+1	+1	−1	+1	−1	−1	−1	+1	10.8
5	−1	−1	+1	+1	−1	−1	+1	+1	10
6	+1	−1	+1	−1	+1	−1	−1	+1	10
7	−1	+1	+1	−1	−1	+1	−1	+1	16.5
8	+1	+1	+1	+1	+1	+1	+1	+1	16.5

[a]Note that X_1, X_2, and X_3 are at their low levels (0). Transformed values are -1, -1, and -1.

value in the column headed X_1 (+1 or −1) by the corresponding response, Y. The sum of these products (ΣXY) divided by 8 (2^n) is the coefficient, B_1.

$$[(-1)(5) + (+1)(9) + (-1)(8) + (+1)(10.8) + (-1)(10) + (+1)(10) + (-1)(16.5) + (+1)(16.5)] = \frac{6.8}{8} = 0.85.$$

The coefficient, B_2, is calculated using the values (+1 or −1) in the second column, the X_2 column.

$$[(-1)(5) + (-1)(9) + (+1)(8) + (+1)(10.8) + (-1)(10) + (-1)(10) + (+1)(16.5) + (+1)(16.5) = \frac{17.8}{8} = 2.225.$$

The coefficient for $X_1X_2X_3$ is B_{123}, and is calculated using the values in the column headed by $X_1X_2X_3$ as follows:

$$[(-1)(5) + (+1)(9) + (+1)(8) + (-1)(10.8) + (+1)(10) + (-1)(10) + (-1)(16.5) + (+1)(16.5)] = \frac{1.2}{8} = 0.15.$$

All of the coefficients are calculated in this manner. B_0 is the sum of all of the observations, Y, divided by 8 (10.725).[†] (Note that all of the values in the "total" column are +1; this column is used to obtain B_0 in the same manner as the other coefficients.) The final polynomial equation for predicting the response, Y, is

$$\begin{aligned} Y = {} & 10.725 + 0.85(X_1) + 2.225(X_2) + 2.525(X_3) \\ & -0.15(X_1X_2) - 0.85(X_1X_3) + 1.025(X_2X_3) + 0.15(X_1X_2X_3) \end{aligned} \tag{16.5}$$

This equation looks entirely different from Eq. (16.2), which also predicts the responses in this experiment. However, the two equations predict the same response. Equation (16.5) uses the transformed levels of X_1, X_2, and X_3 (+ 1 or −1), and Eq. (16.2) uses the actual, observed, untransformed values. For example, if X_1 and X_2 are at their high levels, and X_3 is at the low level, we can solve for the response, Y, using Eq. (16.5) and the transformed values, +1, +1, and −1 for X_1, X_2, and X_3, respectively.

$$\begin{aligned} Y = {} & 10.725 + 0.85(+1) + 2.225(+1) + 2.525(-1) - 0.15(+1)(+1) \\ & -0.85(+1)(-1) + 1.025(+1)(-1) + 0.15(+1)(+1)(-1) = 10.8. \end{aligned}$$

The response with X_1 and X_2 at the high level is 10.8, exactly equal to the value obtained from Eq. (16.2), where X_1, X_2, and X_3 are the actual levels, 2, 1, and 0 mg, respectively.

To reiterate, the reason for the transformation (also called coding) is to allow for calculation of the coefficients in the polynomial equation.[‡] The transformation of the high and low factor levels to +1 and −1 also results in easy calculation of the variance of the coefficients. Using the transformed levels, the variance of a coefficient is $\sigma^2/\Sigma(X - \overline{X})^2 = \sigma^2/8$. With an estimate of the variance, S^2, each coefficient can be tested for significance, using a t test. These tests are exactly equivalent to the testing of the effects of the ANOVA of a factorial design as explained in chapter 9. If, for example, the X_1X_2 interaction were found to be nonsignificant in an ANOVA, the coefficient of X_1X_2, −0.15 in this example, will also be nonsignificant. Usually, when constructing the polynomial equation, only those terms that are statistically "significant" are retained. In the

[†] $B_0 = \overline{Y}$.

[‡] The coded values also result in orthogonality (independence) of effects.

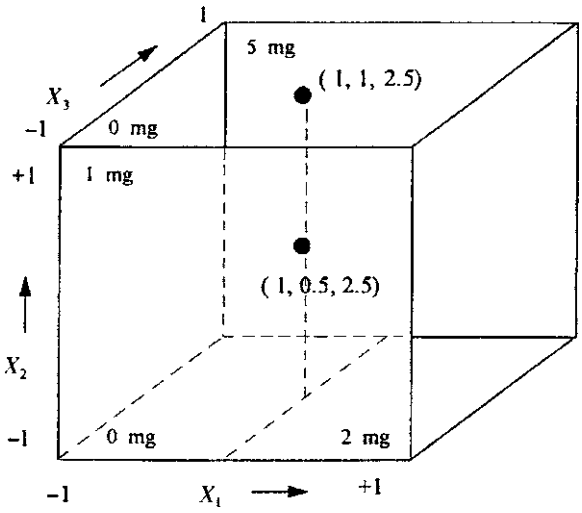


Figure 16.1 Factor space for experiment with factor levels shown in Table 16.1.

experiment above, an estimate of the standard deviation was available from previous similar experiments; s.d. = 0.32 with 16 d.f. Therefore, the coefficients B_{12} and B_{123} (0.15) are not significant.

$$t = \frac{|0.15|}{0.32/\sqrt{8}} = 1.3(P > 0.05).$$

Omitting the "nonsignificant" B_{12} and B_{123} terms, the final equation is

$$Y = 10.725 + 0.85(X_1) + 2.225(X_2) + 2.525(X_3) \\ -0.85(X_1X_3) + 1.025(X_2X_3). \tag{16.6}$$

An advantage of the transformation described above is that the omission of the two coefficients, B_{12} and B_{123} does not affect the values of the remaining coefficients, that is, recalculation of the polynomial equation results in the same coefficients. This result would not occur if Eq. (16.2) were used to describe the data. Equation (16.2) used the untransformed factor levels and would necessitate extensive computations if some terms were omitted, probably requiring use of a computer as a computing aid. Using the transformed values ensures that the factors are orthogonal. This means that the estimates of the coefficients are independent.

Having derived an equation (16.6) that describes the experimental system based on the results of the experimental formulations, we consider this equation to approximately predict the response within the experimental space. Figure 16.1 shows the space described by this design. The *prediction* of the response, Y, at $X_1 = 1$ mg, $X_2 = 1$ mg, and $X_3 = 2.5$ mg is 12.95 [Eq. (16.6)] (see Exercise Problem 1). How do we know that Eq. (16.6) will be a good predictor for responses other than those included in the factorial design? Without actually testing some "extra-design" formulations, we have no way of knowing that the derived empirical equation will be adequate to predict the results of yet-to-be-tested formulations. If the response is "well behaved," the in-between points should be able to be accurately predicted from the response equation.

Usually, it is a good idea to test at least one formulation, not included in the design, as a *checkpoint*. The observed results of the checkpoint formulation can then be compared to the predicted value to test the equation. In our example, a formulation was prepared with $X_1 =$ 1 mg, $X_2 = 0.5$ mg, and $X_3 = 2.5$ mg. The transformed values are equal to zero for the three variables [see the transformation equation (16.4)]. Using Eq. (16.6), the predicted response is 10.725 (only the intercept term is not equal to 0). The factor values for the checkpoint are the average of the low and high levels of the factors (X variables), and lie in the center of the cube in Figure 16.1. This is called a "Center Point." The actual observation made on this formulation was 10.5, very close to the predicted value. Extrapolation of predicted results outside the factor

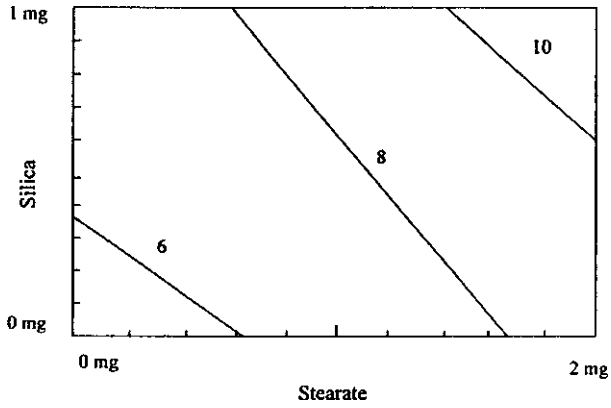


Figure 16.2 Response surface with drug (X_3) constant (low level) [Eq. (16.6)].

space, as shown in Figure 16.1, is not recommended. A two-level design can make predictions only in a linear fashion, usually a gross approximation. If curvature is present, the response may be misrepresented both inside and outside the confines of the design.

Once the polynomial-response equation has been established, an optimum formulation (or a region of optimum formulations) can be found by various techniques. Sometimes, inspection of the experimental results may be sufficient to choose the desired product. In the example above, if large values of the response are desirable, formulations 7 and 8 may be chosen as "best" (Table 16.1). With the use of computers (programmable calculators will often do), a "grid" method may be used to identify optimum regions, and response surfaces may be depicted (Fig. 16.2). The response surface is a geometrical representation of the response and the factor levels, similar to a contour map. For more than two factors, response surfaces cannot be easily represented in two-dimensional space. However, one can take slices of the surface, with all but two factors at fixed levels, as shown in Figure 16.2. A computer can calculate the response, based on Eq. (16.1), at many combinations of the factor levels. The formulation(s) whose response has optimal characteristics based on the experimenter's specifications can then be chosen. To illustrate the grid method, a very rough grid with predicted responses based on Eq. (16.6) is shown in Table 16.3.

The experimental system analyzed above is a very simple example, but is a typical approach to the optimization process. More sophisticated designs may be used, such as the composite designs to be described below (sect. 16.3), or fractional factorial designs. The principles are the same. All of these designs have orthogonal properties to allow for clear and simple estimation of the polynomial coefficients. For these designs, the magnitude of the coefficients is directly related to the magnitude of the response.

The polynomial coefficients may be calculated by techniques such as described here, or by using a multiple regression computer program (see App. III). For two-level experiments

Table 16.3 Grid Solutions for Responses (Y) Based on Eq. (16.6)

X_1[a]	X_2	X_3	Y	X_1	X_2	X_3	Y	X_1	X_2	X_3	Y
−1	−1	−1	5.3	0	−1	−1	7	+1	−1	−1	8.7
−1	−1	0	7.65	0	−1	0	8.5	+1	−1	0	9.35
−1	−1	+1	10	0	−1	+1	10	+1	−1	+1	10
−1	0	−1	6.5	0	0	−1	8.2	+1	0	−1	9.9
−1	0	0	9.875	0	0	0	10.725	+1	0	0	11.575
−1	0	+1	13.25	0	0	+1	13.25	+1	0	+1	13.25
−1	+1	−1	7.7	0	+1	−1	9.4	+1	+1	−1	11.1
−1	+1	0	12.1	0	+1	0	12.95	+1	+1	0	13.8
−1	+1	+1	16.5	0	+1	+1	16.5	+1	+1	+1	16.5

[a] Transformed values.

(2^n factorials), the factor levels should be transformed so that the low level is equal to -1 and the high level equal to $+1$, according to Eq. (16.4). (Experiments with factors at more than two levels should be analyzed with the help of a statistician.) The transformation considerably reduces the complexity of the computations, and aids in the interpretation of the results. Each coefficient may be tested for significance discarding those coefficients that are not significant, although there are no firm rules regarding this procedure. In addition to the statistical criteria, scientific judgment may be used in making decisions about the "significance" of the coefficients. In order to statistically test the coefficients for significance, an estimate of the experimental error is required. This error estimate may be obtained from previous experience, but is best estimated by replicating runs. Replication, however, may result in a large number of experiments, which could be very costly. Replication, accomplished by performing duplicate assays on the same sample, for example, is usually *not* sufficient. The best procedure for replication consists of preparing each formulation or experiment in duplicate (or more), and randomizing the order of the experiments, if all formulations cannot be prepared and tested simultaneously. Methods are available to obtain an estimate of error from an unreplicated factorial experiment (e.g., halfnormal plots [3,4], or from higher order interactions as discussed in chap. 9, but these procedures will not be discussed here).

16.2.1 Replication (Sample Size)

We may only want to find optimum conditions, or we may want to know that effects are real, and not just due to random error. In the latter case, we may want to perform statistical tests (or confidence intervals). To determine the sample size for hypothesis tests, we may use the approximate formula, $N = 4(S^2/\text{delta}^2)(10)$, where N is the sample size for the comparative groups ($N = 4$ for the 2^3 design), where alpha $= 0.05$ and beta $= 0.8$. Usually a sample size between 10 and 20 should be sufficient.

Note that for two-level designs, the variance of an effect is $4S^2/N$, where N is the number of runs.

EXAMPLE:

A difference in response of 2.5 units is meaningful in a 2^3 experiment. The s.d. is expected to be 1.5. What size sample should we use?

$$N = 4(2.25/6.25)(10) = \text{approximately } 16.$$

16.2.2 Extra (Center) Points

Often, it is useful to include an extra run as a "prediction" point, or to estimate curvature. A center point should be equal to the average of the "run" points if there is no curvature. If curvature is present, more runs will be needed to model the data.

The ANOVA for the following data set is shown below to illustrate the analysis of replicated data.

		Level		
Experiment	***A,B***	***P***	***D***	**Response**
1 (1)	*A*	1	0.1	5,6
2 *P*	*B*	1	0.1	7,11
3 *D*	*A*	2	0.1	4,6
4 *PD*	*B*	2	0.1	8,11
5 *A*	*A*	1	0.2	12,12
6 *PA*	*B*	1	0.2	16,21
7 *DA*	*A*	2	0.2	11,12
8 *PDA*	*B*	2	0.2	24,29
9 Checkpoint	*B*	1.5	0.15	22

Analysis of variance table

Source term	d.f.	Sum of squares	Mean square	*F* ratio	Prob. level
P	1	162	162	40.50	0.000380*
D	1	5.555555	5.555555	1.39	0.277097
PD	1	10.88889	10.88889	2.72	0.142947
A	1	304.2222	304.2222	76.06	0.000052*
AP	1	26.88889	26.88889	6.72	0.035802*
AD	1	5.555555	5.555555	1.39	0.277097
APD	1	5.555555	5.555555	1.39	0.277097
S	7	28	4		
Total	14	456.9333			

*$p < 0.05$.

In the absence of replication, there is no proper error term to test significance of the effects. Sometimes we can use an estimate of error from previous experiments or pool the higher order interaction terms. If the runs are replicated, we would have a new term in the ANOVA, residual or error. Then, we can perform F (or t) tests to test for significance.

We could also construct an equation to predict the response (assuming a linear response with factors at two levels). This will be discussed later.

Fractional factorial designs use a fraction of the full factorials (e.g., $^1/_2$, $^1/_4$). The gain is that we use less runs in the experiment. The loss is that we confound some effects. We try to confound effects that we feel are not significant (or very small) with effects that we wish to measure. In this example, the smallest fractional design is a $^1/_2$ replicate, using four of the eight runs. In four runs, we can only measure three effects. The logical choice of effects to measure are A, P, and D. We assume that all interactions are negligible. If our assumption is wrong, the measure of the main effects will be biased.

16.2.3 Optimization of a Combination Drug Product

The following example of a 2^2 factorial experiment is another illustration of the technique of "optimization" using factorial designs. In this experiment, a *combination* drug product was tested to obtain the dose of each drug that would result in an optimal response. The product contained two drugs, $A(X_1)$ and $B(X_2)$. The experiment consists of formulating combinations containing each drug at two dose levels. The doses for A were 5 and 10 mg; B was chosen at doses of 50 and 100 mg. These levels were carefully selected to cover a range of doses that would include an appropriate dose to be chosen as the prime candidate for the final marketed product. The full factorial consists of the four experiments shown in Table 16.4.

The product is a local anesthetic, and the response (Y) is the average time to anesthesia for 12 patients per group. The high and low levels of drug A and drug B are transformed to $+1$ and -1 [Eq. (16.4)]. For drug A, the transformation is

$$\frac{\text{Potency} - 7.5}{2.5} \text{(high level is 10; low level is 5).}$$

Table 16.4 Factorial Design for the Drug Combination Study

	Potency (mg)		Potency (transformed)			
Formulation	$A\,(X_1)$	$B\,(X_2)$	$A\,(X_1)$	$B\,(X_2)$	$AB\,(X_1X_2)$	Response, Y (min)
1	5	50	−1	−1	+1	9.7
2	10	50	+1	−1	−1	7.2
3	5	100	−1	+1	−1	8.4
4	10	100	+1	+1	+1	4.1

For drug B, the transformation is

$$\frac{\text{Potency} - 75}{25} (\text{high level is 100; low level is 50}).$$

The response equation has the form

$$Y = B_0 + B_1(X_1) + B_2(X_2) + B_{12}(X_1)(X_2). \tag{16.7}$$

The coefficients are computed as described earlier in this section. For example, referring to Table 16.4, B_1 is

Column A (X_1)	Y	X_1Y
−1	9.7	−9.7
+1	7.2	+7.2
−1	8.4	−8.4
+1	4.1	+4.1
		−6.8/4 = −1.7

(B_1 is the sum of $X_1Y/4 = -1.7$.) The polynomial equation is calculated as

$$Y = 7.35 - 1.7(X_1) - 1.1(X_2) - 0.45(X_1X_2). \tag{16.8}$$

The response, Y, is the time to anesthesia. Formulation 4, which has the high levels of both drugs, has the shortest time to anesthesia, and formulation 1 or 4 would be chosen as optimal if either a long time or a short time to anesthesia is desired. However, an intermediate time might be more desirable. For example, suppose that a time of 5 minutes is the most desirable time based on considerations such as the administration of the product and the type of conditions that are meant to be treated with the aid of the product. Table 16.5 is a rough grid of the predicted responses based on Eq. (16.8). Based on a time to anesthesia of approximately 5 minutes, a formulation containing 0.5 of A and 1 of B would be a candidate. Decoding the values results in a formulation containing 8.75 mg of A and 100 mg of B.

16.3 COMPOSITE DESIGNS TO ESTIMATE CURVATURE

In general, when looking for optimality, the response equation will be more reliable if it contains terms that reflect curvature. Physical systems are less satisfactorily described by empirical equations containing only linear terms. Figure 16.3 shows an example of a single factor, X, at two levels. Clearly, to interpolate the response, Y, at values of X between the low and high levels requires an assumption of linearity. These predictions would be very much in error if the response is curved, as shown in Figure 16.4.

In order to estimate curvature, more than two levels of the factor must be included in the experiment. The presence of curvature would be reflected in the presence of terms with a power

Table 16.5 Predicted Values of Response to Anesthetic Combinations of Drugs *A* and *B* Based on Eq. (16.8)

		Dose of drug A[a]				
		−1	−0.5	0	+0.5	+1
Dose of drug B[a]	−1	9.7	9.075	8.45	7.825	6.2
	0	9.05	8.2	7.35	6.5	6.65
	+1	8.4	7.325	6.25	5.17	4.1

[a]Coded values of drug potency.

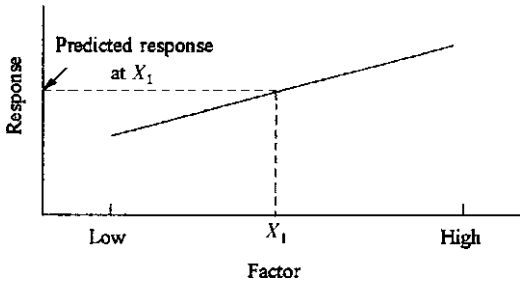


Figure 16.3 Figure showing linear response as a function of a single variable (factor).

greater than 1 (e.g., X_1^2) in the response equation. Such equations are known as polynomials of order 2, and have the following form for a two-factor design:

$$Y = B_0 + B_1 X_1 + B_{11} X_1^2 + B_2 X_2 + B_{22} X_2^2 + B_{12} X_1 X_2 + \ldots \tag{16.9}$$

Composite designs are effective designs to estimate second-order terms. These designs have a number of desirable features. In addition to allowing an estimate of curvature, composite designs give orthogonal estimates of the polynomial coefficients, and allow for the possibility of proceeding with the experiment in a stepwise fashion rather than performing the entire experiment at once. The theory underlying composite designs is beyond the scope of this book. An excellent description of this design and optimization procedure can be found in chapter 11 of Ref. [1].

Although the following discussion is somewhat more advanced than the bulk of material presented in this book, for those who are interested in this subject, an example of a two-factor composite design will be presented to illustrate the technique. A two-factor composite design is identical to a 3^2 factorial design, that is, two factors each at three levels, a total of nine combinations (Table 16.6).

In general, composite designs are not full factorials of the class 3^n, where n is the number of factors. These full factorial designs require a larger number of experiments. For example, a 3^n design with three factors requires 27 runs (27 formulations, for example), 3^3. With more than two factors, composite designs consist of the 2^n design, plus *extra-design* points. The extra points include a *center point* and 2^n extra points, appropriately chosen to maintain orthogonality of the design [1]. The two-factor composite design is shown in Figure 16.5.

The coded values -1, 0, and $+1$ in Table 16.6 for the factor levels represent three *equally spaced* levels of each factor. The coded values in the column headed X_1X_2 are obtained by multiplying the corresponding values in the first two columns (X_1, X_2) as previously described. The values in the columns $X_1^2 - 2/3$ and $X_2^2 - 2/3$ are derived so that the product of corresponding values in any two columns of Table 16.6 sum to zero, resulting in orthogonality (independence) of effects. The special orthogonality obtained by transforming X_i^2 to $X_i^2 - 2/3$ allows for easy

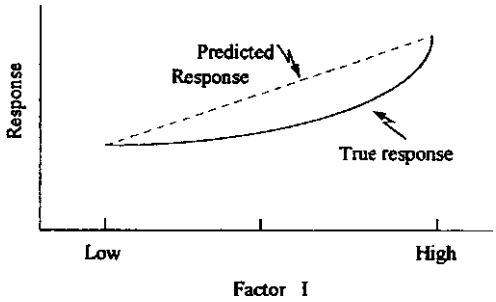


Figure 16.4 Figure showing curved response as a function of a single variable (factor).

Table 16.6 Orthogonal Composite Design with Two Factors (3^2 Design)

Formulation	Coded level X_1	X_2	X_1X_2	$X_1^2 - \frac{2}{3}$	$X_2^2 - \frac{2}{3}$	Response, Y	Predicted response
1	−1	−1	+1	+1/3	+1/3	9.7	9.3
2	−1	0	0	+1/3	−2/3	9.0	9.4
3	−1	+1	−1	+1/3	+1/3	8.4	8.4
4	0	−1	0	−2/3	+1/3	5.3	5.6
5	0	0	0	−2/3	−2/3	4.8	5.0
6	0	+1	0	−2/3	+1/3	3.8	3.3
7	+1	−1	−1	+1/3	+1/3	8.2	8.3
8	+1	0	0	+1/3	−2/3	7.5	6.9
9	+1	+1	+1	+1/3	+1/3	4.1	4.6

calculation of the coefficients and their variances. With this transformation, Eq. (16.9) is modified to

$$Y = B_0 + B_1X_1 + B_{11}\left(X_1^2 - \frac{2}{3}\right) + B_2X_2 + B_{22}\left(X_2^2 - \frac{2}{3}\right) + B_{12}X_1X_2 + \ldots \tag{16.10}$$

The data in Table 16.6 consist of the four formulations from Table 16.4 plus five new runs to complete the composite design. The doses of each drug (X_1 and X_2) were chosen such that the three doses are at equally spaced intervals. Thus the third dose, in addition to the two doses chosen for the 2^2 factorial, is 7.5 mg for $X_1(A)$ and 75 mg for $X_2(B)$. The experiment consists of evaluating the nine combinations of doses, 5, 7.5, and 10 mg for $X_1(A)$ and 50, 75, and 100 mg for $X_2(B)$. Note that the *center point* for the composite design is the combination 7.5 and 75 mg of X_1 and X_2, respectively.

The results of the nine runs are shown in Table 16.6. The results are shown schematically in Figure 16.6(A). The plane at the bottom of the figure shows the combinations of X_1 and X_2. The vertical "sticks" are the responses at each combination of X_1 and X_2. We will compute an equation of the form of Eq. (16.10) that represents a smooth curved surface based on the experimental data. In general, the equation can be obtained through the use of a multiple regression computer program.

The coefficients can also be calculated by "hand" (calculator) using the coded values in Table 16.6. The sum of the products of the coded values times the responses divided by the sum of the squared coded values in the column of interest gives the coefficient. For example, the

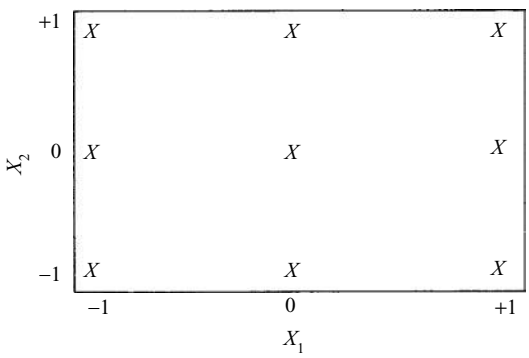


Figure 16.5 Two-factor composite design (3^2 factorial).

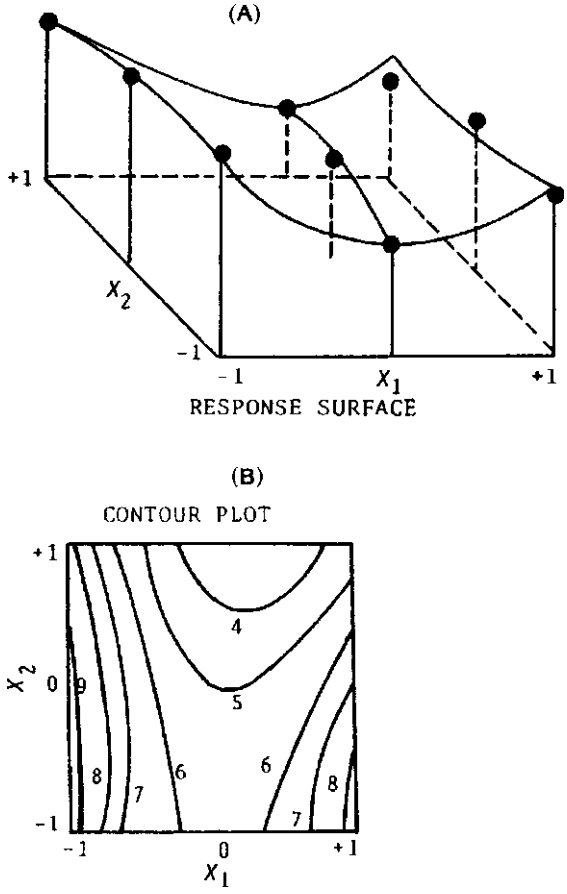


Figure 16.6 Results of composite design experiment from Table 16.6 and response surface computed from Eq. (16.11).

coefficient B_{11} in Eq. (16.10) is calculated as follows:

$X_1'^2 = X_1^2 - 2/3$	Y	$(X_1'^2)(Y)$
+1/3	9.7	3.23
+1/3	9.0	3.00
+1/3	8.4	2.80
−2/3	5.3	−3.53
−2/3	4.8	−3.20
−2/3	3.8	−2.53
+1/3	8.2	2.73
+1/3	7.5	2.50
+1/3	4.1	1.37
$\sum X_1'^2 = 2$		Sum = 6.37

The sum of squared values in the $(X_1^2 - 2/3)$ column is 2. Therefore, the coefficient, B_{11}, is $6.37/2 = 3.18$. The intercept, B_0, is the average of the nine responses, $\overline{Y}$, equal to 6.756. The response equation is

$$Y = 6.756 - 1.22(X_1) + 3.18\left(X_1^2 - \frac{2}{3}\right) - 1.15(X_2) \\ -0.52\left(X_2^2 - \frac{2}{3}\right) - 0.7(X_1 X_2) \tag{16.11}$$

Note that Eq. (16.11) is not an exact fit to the experimental data, as was the case with the polynomial fit described for factorial designs in section 16.2. Had we included three more terms representing various interactions, the equation would exactly fit the data. Equation (16.11) is computed with the assumption that interactions are negligible. Because of the larger number of experiments and the estimation of only six coefficients, we have 2 d.f. for error. Although such an error estimate is not very reliable, it does give us some information, albeit small. The response surface described by Eq. (16.11) is shown in Figure 16.6(A). If this equation does not adequately represent the experimental observations, more terms may be needed in the polynomial equation [Eq. (16.9)] to improve the fit.

The contour plot (similar to contour maps) shown in Figure 16.6(B) allows the selection of combinations of X_1 and X_2 to satisfy given levels of the response. If a maximum response is desired, the X_1, X_2 combinations are limited to a small area of the $X_1 - X_2$ space. If a response of approximately 5 minutes is desired, various combinations of X_1 and X_2 will satisfy the requirements. The ultimate choice will probably depend on other factors, as well, such as cost, toxicity, and so on.

Use of factorial designs in tablet formulation optimization has been presented by Schwartz et al. [5], Fonner et al. [6], and Lindberg et al. [7]. These papers discuss designs somewhat more complex than that presented here. However, for those interested in pursuing this topic further, these papers and the books *The Design and Analysis of Industrial Experiments* [1] and *Statistics for Experimenters* [4] are recommended.

16.4 THE SIMPLEX LATTICE [12]

Response surfaces and optimal regions for *formulation* characteristics are frequently obtained from the application of simplex lattice designs. This class of designs is particularly appropriate in formulation optimization procedures where the *total* quantity of the different ingredients under consideration must be *constant*. Therefore, these are also called "Mixture Designs." For example, suppose that in a liquid formulation, the active ingredient and solvent compose 90% of the product. The remaining 10% of the formulation consists of preservatives, coloring agents, and a surfactant. We wish to prepare a formulation with a certain optimal attribute(s) that is dependent on the relative concentrations of preservative, color, and surfactant. In order to determine optimal regions, we vary the concentrations of these three ingredients in a systematic manner, with the restriction that the total concentration of these ingredients is 10%. This approach differs from the previous procedures (sects. 16.2 and 16.3) in that a constraint is imposed on the total amount of the varying ingredients. In this example, the total amount of the varying components is maintained at 10%. Given the concentration of two of the ingredients, the third ingredient is fixed where in this example $C = 10\% - A - B$.

Implementation of the simplex design consists of preparing various formulations containing different combinations of the variable ingredients. The combinations are prepared in a manner such that the experimental data can be used to predict the responses over the simplex space[§] in a simple and efficient manner. The combinations (formulations) in a simplex design are chosen to cover the space of interest in a symmetrical manner. The experimental results are used to compute a polynomial (simplex) equation that can be used to estimate the response surface. As is true with all optimization and so-called response surface procedures, extrapolation to combinations outside the range included in the experimental design is not recommended. The equation resulting from the experiment, the simplex equation, is an empirical equation that approximately describes the response pattern in the simplex space. There is no reason to believe that the equation has any physical meaning, other than the fact that the complex response patterns resulting from the varying formulations can often be approximated by simple polynomial equations.

Figure 16.7 representing a two-component system (A and B) is useful to help clarify some concepts of simplex designs. One can consider components A and B to be two solvents, which

[§] The simplex space is the region enclosed by the various combinations of ingredients chosen for the experiment. See Figure 16.8, for example.

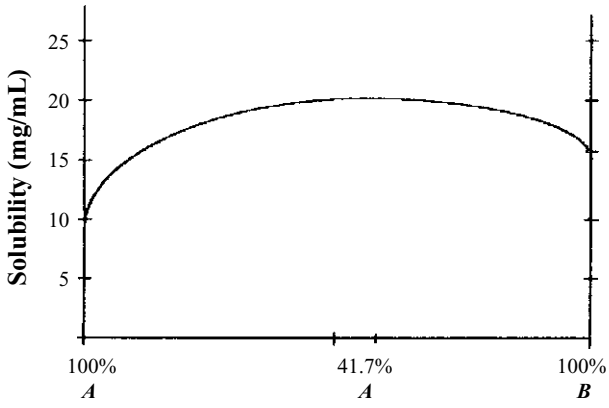


Figure 16.7 Two-component solvent system used to illustrate the simplex approach to optimization.

together comprise the entire solvent system of a drug product. We wish to mix A and B in the correct proportion to optimize the solubility of the drug.

Figure 16.7 is familiar as a solubility phase diagram. This system can also be visualized as an elementary simplex system. The constraint is that the concentrations of A and B must add to 100%. This experiment consists of observing responses (solubility) at three points, *100% A*, *100% B*, and a *50–50 mixture of A and B*, an elementary simplex design. According to Figure 16.7, the solubilities of the drug at the three simplex points, 100% A, 100% B, and 50% A to 50% B, are 10, 15, and 20 mg/mL, respectively. In the simplex approach, we construct an equation of the form

$$Y = B_1(A) + B_2(B) + B_{12}(A)(B), \tag{16.12}$$

where Y is the response (solubility in this example), and (A) and (B) are the concentrations (proportions) of A and B, respectively. The coefficients, B_1, B_2, and B_{12}, are calculated from the experimental observations. The response, Y, can then be predicted for all combinations of A and B, where $(A) + (B) = 1.0$ (100%). (The proportion of each component is usually indicated as a decimal rather than as a percentage.) The form of the simplex design allows for easy calculation of the coefficients. In this example, the coefficients are simply calculated as follows:

$B_1 =$ response at (A) equal to $1.0(100\%) = 10$

$B_2 =$ response at (B)equal to $1.0(100\%) = 15$

$B_{12} = 4$(response at $0.5 - 0.5$ mixture of $A - B$)
$\quad - 2$ (sum of responses at $A = 1.0$ and $B = 1.0$)

$B_{l2} = 4(20) - 2(10 + 15) = 30$

The response equation is

$$Y = 10(A) + 15(B) + 30(A)(B). \tag{16.13}$$

The solution above for the three coefficients is a result of the solution of three simultaneous equations:

With $A = 1.0$ and $B = 0$, from Eq. (16.12), B_1** $= 10$

With $A = 0$ and $B = 1.0$, from Eq. (16.12), $B_2 = 15$

**The response, Y, with A equal to 1.0 (100%) is 10.

With $A = 0.5$ and $B = 0.5$, from Eq. (16.12),

$$20 = 0.5B_1 + 0.5B_2 + 0.25B_{12} \text{ or } B_{12} = 4(20) - 2(B_1 + B_2) = 30.$$

We will see that in more complex simplex designs, the polynomial coefficients are, similarly, easily calculated as linear combinations of experimental results.

Equation (16.13) exactly predicts the observed points: a fit of a polynomial with three terms to three experimental points. We can always construct an equation with N coefficients that will exactly pass through N points. For example, for the 50–50 mixture,

$$Y = 10(0.5) + 15(0.5) + 30(0.5)(0.5) = 20.$$

The *response equation* predicts responses at extra-design points, those formulations not included in the experiment but that lie within the simplex space, 100% A to 100% B in this example. For example, what solubility would be predicted in a solvent system containing 75% A and 25% B? (Note that $A + B$ must equal 100%.) Applying Eq. (16.13), we have

$$Y = 10(0.75) + 15(0.25) + 30(0.75)(0.25) = 16.875.$$

See also Figure 16.7. The entire response may be sketched in by predicting solubilities along the curve, as shown in the figure.

The primary experimental objective in experiments such as that described above may be the determination of the solvent combination that results in maximum drug solubility. The optimum solubility can be computed by calculating the predicted solubility at many solvent combinations so as to clearly define the response over the solvent mixture continuum. This may seem an indirect and tedious approach, but with the ready availability of computers, this is often the most expeditious route. The maximum solubility is predicted to occur at 41.67% A. In this simple example, the maximum can easily be calculated by setting the first derivative of Eq. (16.13) equal to 0 (see Exercise Problem 6).

In general, the simplex design is usually applied to formulation problems in which a mixture of three or more components is to be investigated. The design is conveniently represented by regular-sided figures, which can be visualized for three- or four-component systems. For more than four components, a single figure cannot be conveniently constructed, but can be theoretically conceived as an N-sided figure in $(N - 1)$-dimensional space. For example, Figure 16.8 shows the three-component system that is represented as an equilateral triangle in two-dimensional space. A regular simplex design for a three-component mixture system consists of six or seven formulations.

Three formulations, one each at each vertex, *A*, *B*, and *C*. These formulations represent formulations with the pure components, *A*, *B*, and *C*, respectively.

Three formulations are prepared with 50−50 mixtures of each pair of components, *AB*, *AC*, and *BC*.

A *seventh* formulation may be prepared with one-third of each component. This lies in the *center* of the design.

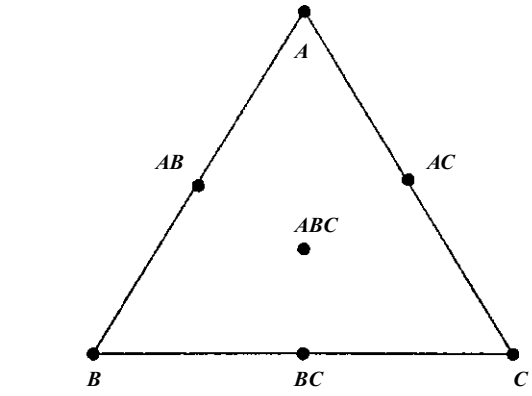
An example of a simplex design for four components consisting of 15 formulations is shown in Figure 16.8. The 15 formulations consist of

Four formulations each with 100% of each of the four pure components *Six* formulations of 50−50 mixtures of component pairs (*AB*, *AC*, *AD*, *BC*, *BD*, and *CD*).

Four formulations consisting of one-third mixtures of combinations of three components *(ABC*, *ABD*, *ACD*, *BCD*).

A mixture containing 25% of each of the four components (*ABCD*).

The simplex design is arranged so that the experimental space is well covered in a symmetrical fashion. In addition, the symmetrical spacing of the points allows for an easy computation



Four-component simplex

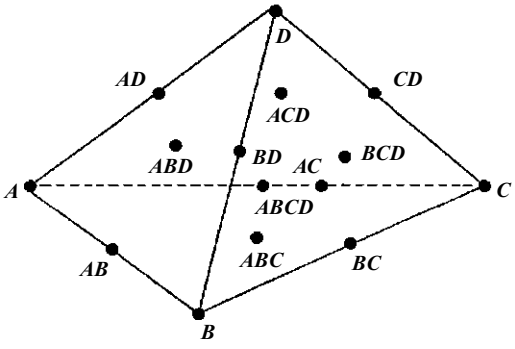


Figure 16.8 Three-component simplex lattice design and four-component simplex lattice design.

of the response equation coefficients. The general equation for the response based on a simplex design contains terms for pure components and all mixtures of components as follows:

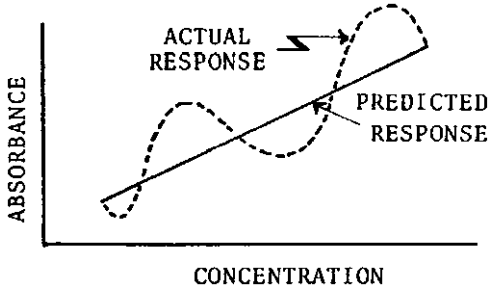
$$Y = B_a(A) + B_b(B) + B_c(C) + \ldots + B_{ab}(A)(B) + B_{ac}(A)(C) + B_{bc}(B)(C) + \ldots + B_{abc}(A)(B)(C) + \cdots \tag{16.14}$$

where (A), (B), and (C) are the proportions of components A, B, and C, and $(A) + (B) + (C) + \ldots$ is equal to 1.0.

The subscripted B's (e.g., B_a) are coefficients that can be easily calculated from the responses, Y, or using a multiple regression computer program.

After the coefficients have been calculated, the response equation [Eq. (16.14)] may be used to predict the response of combinations of the N components in the system. With the aid of a computer, responses may be calculated over the simplex space, and contour diagrams printed (see also Fig. 16.6). The contour plot is a graphic description of the response surface resulting from data derived from experimental designs such as the simplex. For the two-component system (Fig. 16.7), the response surface is simply the solubility curve. With three components, a three-dimensional figure would be necessary to show the response surface. A contour plot is a means of illustrating the response on a two-dimensional surface, as is familiar to those who have been exposed to contour maps. A computer may be programmed to produce two-dimensional figures (commercial programs are also available) that are slices through the three-dimensional figure for a three-component system. The slices are taken at a constant concentration of one of the components. In computer outputs, the regions of equal response are indicated by a common symbol, such as a letter or a figure. An example of a contour plot was shown in Figure 16.6. The contour plot will be discussed further in the example that follows. Examination of the contour plot(s) allows the experimenter to choose formulations that have predicted responses of some specified magnitude.

When constructing an empirical response equation based on a limited number of experimental observations, one should understand that predicted values based on the equation may be in error for several reasons. For example, the empirical equation (or model, as it is often called) rarely exactly defines the experimental system. The equation is an approximation to the system. To understand this important concept, note that the same problem would exist if we had only two points in the experimental space. The empirical equation derived from the two points could only relate the observations by a straight line. In-between points could only be predicted on the basis of the straight-line relationship (Figs. 16.3 and 16.4).



If the true relationship of the X, Y variables were curved, the linear interpolation would be in error. In the simplex design, we used a limited number of points to define a relatively large region of response. Even if the model represented by the empirical equation is a reasonable representation of the true surface, other sources of variation can contribute to error in the prediction equation and predicted responses (e.g., error in measuring the response). Thus, in these systems, we have at least two obvious sources of variability: that due to the empirical model and that due to observational errors.

How can we protect ourselves from inadvertently proceeding with predictions when the derived equation is indeed inaccurate? As insurance against such a possibility, it is a good idea to run one or more extra-design points. These points are not used to estimate the coefficients in the simplex equation [Eq. (16.14)] but will be used as checkpoints. Once the simplex equation is derived, the result at the extra-design checkpoint(s) is predicted based on the equation, and its agreement with the observed value assessed. If the agreement is close, we have increased faith in the predictive power of the response equation (see sect. 16.2). If we have an estimate of error from replication or other means, we may wish to perform a statistical test to test the adequacy of the model (a statistician may be consulted for this calculation).

The calculation of the simplex equation coefficients is easily accomplished using the following formulas. These formulas are an extension of those discussed previously for the two-component system as applied to a three-component system. The general formulas for calculation of coefficients for an N-component system may be found in Ref. [7].

$$
\begin{aligned}
B_1 &= Y_1, \text{ the response at 100\% } A \\
B_2 &= Y_2, \text{ the response at 100\% } B \\
B_3 &= Y_3 \text{ the response at 100\% } C \\
B_{12} &= 4(Y_{12}) - 2(Y_1 + Y_2), \text{ where } Y_{12} \text{ is the response at } 50 - 50\ AB \\
B_{13} &= 4(Y_{13}) - 2(Y_1 + Y_3), \text{ where } Y_{13}, \text{ is the response at } 50 - 50\ AC \\
B_{23} &= 4(Y_{23}) - 2(Y_2 + Y_3), \text{ where } Y_{23} \text{ is the response at } 50 - 50\ BC \\
B_{123} &= 27(Y_{123}) - 12(Y_{12} + Y_{13} + Y_{23}) + 3(Y_1 + Y_2 + Y_3), \\
&\quad \text{where } Y_{123} \text{ is the response at } 1/3A, 1/3B, \text{ and } 1/3C
\end{aligned}
\tag{16.15}
$$

The discussion above has been based on an experimental situation where the components being varied in the simplex design comprise the entire mixture (100%). In pharmaceutical formulations, a more common situation is one in which part of the formulation must remain

fixed (e.g., drug concentration in a tablet). The remaining components, which may be varied, therefore do not make up 100% of the mixture. In addition, the lower limit for the varying components is often not equal to 0. For example, some components must be present in some minimal quantity in order that a marketable product can be manufactured. This is known as a design with constraints. For tablets, some minimal amount of a lubricating agent may be necessary in order to obtain an acceptable product. These modifications in the simplex design present no problem, however, because we can restrict the treatment of the simplex to those components that are varied, and with suitable transformations, treat the data in exactly the same way as described above. For example, if the components to be varied make up 60% of the total formulation ingredients, we can appropriately transform the actual percentages of these components so that the transformed percentages total 100%. In a three-component mixture containing 20% of each of three components, each component can be transformed to 33.3% (1/3) for purposes of the simplex analysis. Transformations can also be made where the components have a lower limit greater than 0% and an upper limit less than 100%, as will be explained in the following worked example.

The example presented below is an experiment in which a simplex design was used to obtain a formulation with optimal properties. This example should clarify the concepts and procedures described above. This experiment was prompted by problems with tablet hardness for a large-volume marketed product. Although the reason for the problem was not obvious, the pharmaceutical product development scientists felt that the cause could be traced to three components of the tablets, which we will denote as ingredients *A*, *B*, and *C*. Together, these components consisted of 25% of the original formulation, or 75 mg of the total tablet weight of 300 mg. A careful evaluation of the product ingredients indicated that the three components had to be present in an amount equal to at least 10 mg each in order for the tablet to be satisfactorily compressed. Thus, the recommended simplex design to obtain a satisfactory tablet hardness consisted of varying the three components with the constraint that the sum of the components must be 75 mg, and that each component be present in an amount equal to at least 10 mg.

In order to apply the simplex equation to be derived from this experiment in a convenient manner, the actual concentrations used should be *transformed* such that the minimum concentration (10 mg) corresponds to 0% and the highest concentration corresponds to 100%.[††] In our example, the transformation is the same for all three components because each component is subject to the same restrictions. The minimum quantity is 10 mg and the maximum is 55 mg. (The other two components, each at 10 mg, make up the 20-mg difference, a total of 75 mg.) The transformation is as follows:

$$\begin{aligned}\text{Transformed proportion} &= \frac{\text{Amount used} - \text{minimum}}{\text{maximum} - \text{minimum}} \\ &= \frac{\text{Amount used} - 10}{55 - 10}.\end{aligned} \tag{16.16}$$

Thus, a formulation prepared with a 50–50 mixture of components *A* and *B* would actually contain 32.5 mg of *A*, 32.5 mg of *B*, and 10 mg of *C*. Note that from Eq. (16.16), if a component is at a concentration of 32.5 mg, the transformed proportion is $(32.5 - 10)/(55 - 10) = 0.5$. A formulation with "100%" *A* would actually contain 55 mg of *A*, 10 mg of *B*, and 10 mg of *C*.

The three-component simplex design was run with one checkpoint, as shown in Table 16.7. The hardness values represent the average hardness of 20 tablets taken at random from the experimental batches. The simplex coefficients are computed as described previously [Eq. (16.15)], resulting in the following equation:

$$\begin{aligned}Y = {} & 6.1(A) + 7.5(B) + 5.3(C) \\ & -0.8(A)(B) + 2.8(A)(C) + 2.0(B)(C) + 15(A)(B)(C).\end{aligned} \tag{16.17}$$

[††]If there are no constraints on the upper and lower limits, the highest concentration would ordinarily be 100% and the lowest 0%.

Table 16.7 Results of a Three-Component Simplex System for Tablet Hardness

Formulation components			Transformed proportion			
A	***B***	***C***	***A***	***B***	***C***	**Average hardness, *Y***
55	10	10	1.0	0	0	6.1
10	55	10	0	1.0	0	7.5
10	10	55	0	0	1.0	5.3
32.5	32.5	10	0.5	0.5	0	6.6
32.5	10	32.5	0.5	0	0.5	6.4
10	32.5	32.5	0	0.5	0.5	6.9
25	25	25	0.33	0.33	0.33	7.3
32.5[a]	21.25	21.25	0.5	0.25	0.25	7.2

[a]Extra-design checkpoint.

For example, the coefficient B_{123} is calculated as follows:

$$27(7.3) - 12(6.6 + 6.4 + 6.9) + 3(6.1 + 7.5 + 5.3) = 15.$$

(A), (B), and (C) in Eq. (16.17) are the transformed proportions. The extra-design checkpoint (the final formulation in Table 16.7) has a response of 7.2. The predicted value based on Eq. (16.17) is 7.09, very close to the observed value, 7.2. This is some confirmation of the adequacy of Eq. (16.17) as a predictor of tablet hardness. Figure 16.9 shows a contour plot of the results of the experiment based on Eq. (16.17). Tablets with high hardness are found in the region with relatively larger amounts of component B. If a tablet hardness of 7 or more is satisfactory, the pharmaceutical scientist has a choice of formulations. The final composition may then be dependent on other factors, such as cost or other tablet properties.

The following example shows data (Table 16.8) and analysis from a replicated simplex design that gives an estimate of experimental error. The design is a basic three-component (A, B, and C) simplex design with a center point consisting of 1/3 of each of the three components. This example is set up for a computer analysis. Note that the interaction term coefficients are the product of the main effect coefficients. For example for Run #7, the ABC interaction is

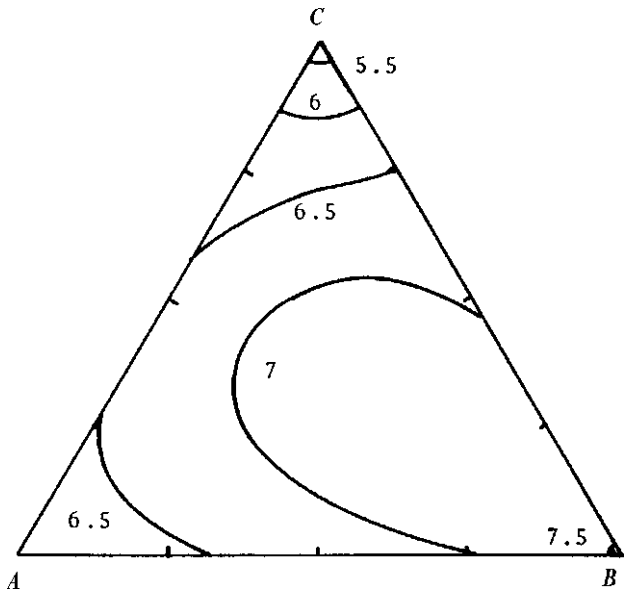


Figure 16.9 Contour plot of three-component simplex system (Table 16.7).

Table 16.8 Example of a Replicated Simplex Design

Run	*A*	*B*	*C*	*AB*	*AC*	*BC*	*ABC*	Response
1	1	0	0	0	0	0	0	38
2	0	1	0	0	0	0	0	27
3	0	0	1	0	0	0	0	46
4	0.5	0.5	0	0.25	0	0	0	33
5	0.5	0	0.5	0	0.25	0	0	51
6	0	0.5	0.5	0	0	0.25	0	32
7	0.333	0.333	0.333	0.111	0.111	0.111	0.037	48
8	1	0	0	0	0	0	0	42
9	0	1	0	0	0	0	0	28
10	0	0	1	0	0	0	0	41
11	0.5	0.5	0	0.25	0	0	0	35
12	0.5	0	0.5	0	0.25	0	0	47
13	0	0.5	0.5	0	0	0.25	0	32
14	0.333	0.333	0.333	0.111	0.111	0.111	0.037	50

Independent variable	Regression coefficient	Standard error	Lower 95% CL	Upper 95% CL
A	40	1.535299	36.36959	43.63041
B	27.5	1.535299	23.86959	31.13041
C	43.5	1.535299	39.86959	47.13041
AB	1	7.521398	−16.78528	18.78528
AC	29	7.521398	11.21472	46.78528
BC	−14	7.521398	−31.78528	3.78528
ABC	277.1	52.90734	151.9937	402.2056

		Analysis of variance section			
Source	d.f.	Sum of squares	Mean square	Prob. *F* ratio	Level
Intercept	0	0	0		
Model	7	22461	3208.714	680.6364	0.000000
Error	7	33	4.714286		

$0.333 \times 0.333 \times 0.333 = 0.037$. The computer analysis gives the regression coefficients for the response equation, and an ANOVA to estimate the experimental error. The variance estimate is 4.71.

A checkpoint was run at $A = 0.25$, $B = 0.25$, and $C = 0.5$ with a response of 46. The model predicted 49.2.

In my experience, this approach gives excellent results.

16.5 SEQUENTIAL OPTIMIZATION**

Sequential optimization was developed as a means to optimize a process in a stepwise fashion. Evolutionary operation (EVOP) uses factorial type designs and usually requires a large number of experiments [8]. A relatively simple approach to sequential optimization is a stepwise application of the simplex procedure [9,10]. The procedure consists of first generating data from $n + 1$ experiments where n is the number of independent variables or factors. Based on the $n + 1$ responses and predetermined rules, one result is eliminated from the set and a new experiment is performed. A decision is made as a result of the most recent experiment, generating another new experiment, and so on, eventually terminating the design at an "optimal" response. Thus, each new experiment leads the researcher on a path toward an optimum. The procedure and rules are illustrated in the following example. For further details and illustrations, the reader is encouraged to study Refs. [9–11].

**A more advanced topic.

Table 16.9 Initial Four Experiments for Simplex Experiment

Experiment	Disintegrant	Lubricant	Fill weight	Response
1	+(50)[a]	−(0.2)	−(100)	37
2	−(0)	+(2.2)	−(100)	58
3	−(0)	−(0.2)	+(400)	46
4	+(50)	+(2.2)	+(400)	40

[a]Parenthetical value is the amount of ingredient in the formulation.

16.5.1 An Example of Sequential Simplex Optimization

This example is based on the presentation by Shek et al. [11] using the simplex procedure to optimize properties of a capsule formulation. They were interested in optimizing dissolution and compaction rates as a function of the factors (or variables) drug, disintegrant, lubricant, and fill weight. In this synthetic example, we will look at a single response, dissolution at 30 minutes, as a function of three variables: disintegrant, lubricant, and fill weight.

We start with four experiments (we have three variables). There are no firm rules regarding the design of these experiments, but principles of good experimental design should prevail. For example, a 1/2 replicate of a 2^3 factorial design can be used for the initial four experiments. This requires setting low (−) and high (+) levels for each factor; see Table 16.9.

Let W = vector of worst response
Let S = vector of second worst response
Let B = vector of best response
Let R_w = worst response
Let R_s = second worst response
Let R_b = best response
Let P = average vector after elimination of worst response among formulations under consideration.

Note that since Formula 2 shows the worst response (the longest dissolution time) $\overline{P}$ is the average of experiments 1, 3, and 4 and is equal to (33.3, 0.87, 300). For example, the first vector element refers to the average disintegrant = $(+50 - 0 + 50)/3 = 33.3$.

Procedure:

Step 1. Eliminate W, the vector of the worst response from the data set and compute R [Eq. (16.18) below], the formulation for the new experiment.

$$\begin{aligned} &R = \overline{P} + (\overline{P} - W) \\ &(33.3,\ 0.87,\ 300) + (33.3,\ -1.33,\ 200) = (66.6,\ -0.46,\ 500). \end{aligned} \tag{16.18}$$

In this example, we need 66.6 of disintegrant, −0.46 of lubricant and a fill weight of 500. We will interpret this result after the rules are specified and we proceed with the optimization.

If the response from experiment R, R_r, is better than the second-worst response, R_s, but worse than the best response, retain R_r and proceed to Step 1, evaluating a new formulation with the new set of four formulations.

If the response to R_r is better than the best response, proceed to Step 2.
If the response to R_r is worse than the second-worst response, go to Step 3.
If the response to R_r is worse than the worst response, go to Step 4.

Step 2. Compute E [Eq. (16.19) below] and evaluate R_e.

$$E = \overline{P} + 2(\overline{P} - W) \tag{16.19}$$

If R_r is better than the response to E, R_e, retain R. If R_e is better than R_r, retain E.

Step 3. Compute C_r [Eq. (16.20) below] and evaluate the response to C_r, R_{cr}.

$$C_r = \overline{P} + 0.5(\overline{P} - W) \tag{16.20}$$

Retain C_r. However, if R_{cr} is worse than R_s (the next-to-worst response), then set $R_w = R_s$ and $W = S$. (This means that the worst response is set equal to the next-to-worst response.) Set R_{cr} as the next-to-worst response, that is, $S = C_r$ and $R_s = R_{cr}$.

Step 4. Compute C_w [Eq. (16.21) below] and evaluate R_{cw}. Retain C_w. However, if R_{cw} is worse than R_s (the next-to-worst response), then set $R_{cw} = R_s$ and $W = S$ (this means that the worst response is set equal to the next-to-worst response). Set R_{cw} as the next-to-worst response, that is, $S = C_w$ and $R_s = R_{cw}$.

Summary of calculation of new formulations

1. $R = \overline{P} + (\overline{P} - W)$
   $R_r =$ The response to formula R (16.18)
2. $E = \overline{P} + 2(\overline{P} - W)$
   $R_e =$ The response to formula E (16.19)
3. $C_r = \overline{P} + 0.5(\overline{P} - W)$
   $R_{cr} =$ The response to formula C_r (16.20)
4. $C_w = \overline{P} - 0.5(\overline{P} - W)$
   $R_{cw} =$ The response to formula C_r (16.21)

Although this procedure may appear confusing, if one follows the example, the process will be clarified.

We have already calculated the vector for the first new formulation using Step 1 above: (66.6, −0.46, 500). The response to this formulation will replace the worst formulation, W, which is formulation 2. Unfortunately, we cannot prepare this formulation because of the negative quantity of lubricant. We will make a rule that in such impossible situations we consider the response to this new formulation to be worse than the remaining formulations under consideration (formulations 1, 3, and 4).

This sends us to Step 4 according to our rules. The formulations under consideration are 1, 3, 4, and 5 in Table 16.10. According to Eq. (16.21)

$$C_w = (33.3, 0.87, 300) - 0.5(-33.3, 1.33, -200)$$
$$= (50, 0.20, 400).$$

The response, R_{cw}, to C_w is 44. According to Step 4 above, we retain this result. This is shown as experiment 6 in Table 16.9. We now operate on experiments 1, 3, 4, and 6; experiment 3 is the new worst result.

We go to Step 1 and compute our new formulation R from Eq. (16.18)

$$R = (50, 0.87, 300) + (50, 0.67, -100) = (100, 1.54, 200).$$

Table 16.10 Sequential Experiments in Optimization Process

Experiment	Disintegrant	Lubricant	Fill weight	Response
1	50	0.2	100	37
2	0	2.2	100	58(W_1)[a]
3	0	0.2	400	46(W_3)
4	50	2.2	400	40
5	66.6	−0.46	500	(W_2)
6	50	0.20	400	44(W_4)
7	100	1.54	200	42(W_6)
8	83.3	2.42	67	43(W_5)
9	58.4	0.75	316	36
10	8.5	0.07	416	41(W_7)
11	39	0.56	344	44(W_8)
12	56.2	0.8	308	35

[a] W_1 means that this result was eliminated after the first evaluation.

The response R, is 42 (represented by experiment 7 in Table 16.9). This is better than the second worst response (44 for experiment 6) and we retain R_r as directed in Step 1 above. We recompute R for the set of experiments 1, 4, 6, and 7

$$R = (66.7, 1.31, 233) + (16.7, 1.11, -167) = (83.3, 2.42, 67).$$

The response, R_r, is 43. This is worse than the second-to-worst response, 42. Therefore we go to Step 3

$$C_r = \overline{P} + 0.5(\overline{P} - W)$$

$$C_r = (66.7, 1.31, 233) + 0.5(-16.7, -1.11, 167)$$
$$= (58.4, 0.75, 316).$$

The new response (experiment 9) is 36.
According to our rules, we go to Step 2

$$E = \overline{P} + 2(\overline{P} - W)$$

$$E = (69.5, 1.05, 272) + 2(-30.5, -0.49, 72)$$
$$= (8.5, 0.07, 416).$$

The response to E is 41. According to Step 2, we retain R in lieu of E because R gave the better response. We compute a new R from Step 1

$$R = (69.5, 1.05, 272) + (-30.5, -0.49, 72)$$
$$= (39, 0.56, 344).$$

The response is 44. Our new set of four experiments is numbers 1, 4, 9, and 11, with number 11 the worst.
We go to Step 4 and compute C_w because the value of R is worse than R_w

$$C_w = (69.5, 1.05, 272) - 0.5(30.5, 0.49, -72)$$
$$= (54.2, 0.8, 308).$$

The response was 35 (see experiment 12).
The experiments may continue as described above until repeated experiments do not show improvement. We are searching for an optimal response in the presence of variability. In the present case, a formula containing approximately 55 of disintegrant and 0.75 of lubricant with a fill weight of 300 mg appeared to show minimal dissolution time; the study was stopped after experiment 12.
As with other optimization procedures presented in this chapter, studying details in the literature references is essential to understand the procedure and calculations [8–11].

16.6 SCREENING DESIGNS

Usually, we know the factors that we wish to investigate, from our experience. However, in new, unknown, situations, it is possible that we may consider a number of factors to investigate, to see if any of these may affect the response or outcome. If there are only a few such variables (or factors), we may wish to use a factorial or fractional factorial design. If there are many potential factors of interest, screening designs are available that use less runs, but do give us insight into effects of interest. The most popular of such designs are the Plackett–Burman designs.
Screening designs may be useful if little is known of the system. In most cases, one should have a reasonable idea of which variables are important, and their effective ranges. But, we may be surprised. If everything were known, experimentation would not be necessary. Also, one should be careful not to neglect potentially important variables.

Table 16.11 Twelve Run Plackett–Burman Design

Run	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	+	+	−	+	+	+	−	−	−	+	−
2	+	−	+	+	+	−	−	−	+	−	+
3	−	+	+	+	−	−	−	+	−	+	+
4	+	+	+	−	−	−	+	−	+	+	−
5	+	+	−	−	−	+	−	+	+	−	+
6	+	−	−	−	+	−	+	+	−	+	+
7	−	−	−	+	−	+	+	−	+	+	+
8	−	−	+	−	+	+	−	+	+	+	−
9	−	+	−	+	+	−	+	+	+	−	−
10	+	−	+	+	−	+	+	+	−	−	−
11	−	+	+	−	+	+	+	−	−	−	+
12	−	−	−	−	−	−	−	−	−	−	−

Screening designs, in general, are fractional factorials of 2^n designs that estimate main effects, but not interactions. If results of such experiments point to specific factors, one can follow up with more complete designs to evaluate specific interactions.

A 12-run design is shown in Table 16.11. Note that the − and + signs refer to the low and high levels of the factor, respectively. Thus, for example, factor 1 in run 1 is at the high level. (See chap. 9 for further explanation of terminology.) For other designs, for example, higher order or more complex designs, a statistician should be consulted. In general, variability cannot be estimated without replication (run the design in duplicate, for example) or partial replication. This would increase the size and cost of the experiment. As in other design considerations, the cost and time considerations must be weighed against the information gained from expanded experiments. If less factors than runs are used, an estimate of variability can be provided. This is shown in the following example.

An example of a 12-run Plackett–Burman design is shown in Table 16.12. This design estimates the main effects of six variables. This leaves 5 d.f. for estimating the error. The estimates based on columns 7 to 11, inclusive, are only used to compute the variability, and are not related to the six factors in the experiment. An example of an experiment using this design could be as follows. The effect of six variables on the dissolution of a tablet is to be investigated. The six factors are (X_1) hardness, (X_2) level of disintegrant, (X_3) time of mixing granulation, (X_4) level of lubricant, (X_5) type of coating, and (X_6) tablet press pressure. The response is the percent dissolution in 30 minutes. Each factor is set at a low (−1) and a high (+1) level. (Note that "type of coating" is arbitrarily set at −1 and +1.)

The analysis is most easily accomplished using a multiple regression computer program. When designating values for the model in the computer program, it is convenient to input −1 for the low level and + 1 for the high level. Table 16.13 shows an example of relevant computer output.

Table 16.12 Example of Twelve Run Plackett–Burman Design

Dissolution	X_1	X_2	X_3	X_4	X_5	X_6	Error	Error	Error	Error	Error
75	1	1	−1	1	1	1	−1	−1	−1	1	−1
104	1	−1	1	1	1	−1	−1	−1	1	−1	1
57	−1	1	1	1	−1	−1	−1	1	−1	1	1
54	1	1	1	−1	−1	−1	1	−1	1	1	−1
46	1	1	−1	−1	−1	1	−1	1	1	−1	1
58	1	−1	−1	−1	1	−1	1	1	−1	1	1
3	−1	−1	−1	1	−1	1	1	−1	1	1	1
98	−1	−1	1	−1	1	1	−1	1	1	1	−1
80	−1	1	−1	1	1	−1	1	1	1	−1	−1
12	1	−1	1	1	−1	1	1	1	−1	−1	−1
100	−1	1	1	−1	1	1	1	−1	−1	−1	1
13	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

Table 16.13 Multiple Regression Computer Output of Data in Table 16.12

Independent variable	Regression coefficient	*T* value (H_0: $B = 0$)	Prob. level	Decision (5%)
Intercept	58.33	13.3748	0.000042	Reject H_0
X_1	−0.167	−0.0382	0.970996	Accept H_0
X_2	10.33	2.3692	0.064013	Accept H_0
X_3	12.5	2.8660	0.035158	Reject H_0
X_4	−3.167	−0.7261	0.500353	Accept H_0
X_5	27.5	6.3052	0.001477	Reject H_0
X_6	−2.667	−0.6114	0.567651	Accept H_0

Analysis of variance Source	d.f.	Sum of squares	Mean square	*F* ratio	Prob. level
Intercept	1	40833.33	40833.33		
Model	6	12437.33	2072.889	9.0810	0.014
Error	5	1141.33	228.267		
Total	11	13578.67	1234.424		

Note that only main effects are estimated. The error term comprises the five columns that were not assigned to factors (columns 7–11). If only five factors were investigated, columns 6 to 11 would be used to estimate error with 6 d.f. The estimate of error allows us to test the main effects for significance. This is a conservative test because the error will be, if anything, estimated on the high side. That is, if any interactions are present, the error estimate will be too high. This means that we may miss some significant effects if interaction is present. In this example, X_2 just misses significance, and X_3 and X_5 are significant. Again, the six factors are (X_1) hardness, (X_2) level of disintegrant, (X_3) time of mixing granulation, (X_4) level of lubricant, (X_5) type of coating, and (X_6) tablet press pressure. Therefore, we might wish to consider the level of disintegrant, time of mixing, and type of coating if we wish to modify the dissolution. The type of coating seems to have the greatest effect.

KEY TERMS

Checkpoint
Coding
Composite designs
Contour plot
Extra-design points
Factorial designs
Fractional factorial designs
Grid
Independence
Model
Model error
Multiple regression
Optimization
Orthogonality
Plackett–Burman
Polynomial equation
Replication
Response equation
Response surface
Screening designs
Sequential optimization
Simplex design
Simplex space
Transformation

EXERCISES

1. Calculate the predicted response from Eq. (16.6) for
   (a) $X_1 = 1$ mg, $X_2 = 1$ mg, $X_3 = 2.5$ mg
   (b) $X_1 = 2$ mg, $X_2 = 1$ mg, $X_3 = 4$ mg
   Note that Eq. (16.6) uses coded values; see Eq. (16.4).] For example, the coded value for $X_1 = 1$ mg is $0 = (1 - 1)/1$.
2. Show that the transformed values of $X_1 = 1$, $X_2 = 0.5$, and $X_3 = 2.5$ are all equal to zero for the three variables in Exercise Problem 1.
3. Calculate the coefficients for the polynomial equation, (16.8). The coefficients are calculated from the data in Table 16.4.

4. Show that decoded values of A and B equal to 0.5 and 1, respectively, are equal to 8.75 mg of A and 100 mg of B, for the data of Table 16.4 and Eq. (16.8). Calculate the expected response of this combination of A and B using Eq. (16.8).
5. A formulation was to be prepared to optimize dissolution time. (The formulation with the dissolution time of approximately 15 minutes is "optimal.") Stearic acid and mixing time were varied according to a 2^2 factorial design with the following results:

		Stearic acid	
		0.25%	1%
Mixing time (min)	15	10	23
	30	21	25

   (a) Construct a polynomial response equation [see Eq. (16.8)].
   (b) What concentration of stearic acid and mixing time would you choose for the final product?

‡‡6. Calculate the maximum solubility based on Eq. (16.13), using procedures of calculus. [Hint: Set the first derivative equal to 0 after substituting $(1.00 - A)$ for B.]

7. A total of 100 mg of three components, stearic acid (A), starch (B), and DCP (C), are to be added to a tablet formulation. Dissolution time was measured in a simplex design with the following results:

100% *A*:	292.0 min
100% *B*:	5.6 min
100% *C*:	50.4 min
50% *A*, 50% *B*:	25.6 min
50% *B*, 50% *C*:	15.6 min
50% *A*, 50% *C*:	124.5 min
1/3 *A*, 1/3 *B*, and 1/3 *C*:	37.0 min

   (a) Compute the simplex equation coefficients.
   (b) Give a combination with very fast dissolution.
   (c) Give a combination that has a dissolution time of 90 minutes.

REFERENCES

1. Davies OL. The Design and Analysis of Industrial Experiments. New York: Hafner, 1963.
2. Ahmed S, Bolton S. Factorial design in the study of the effects of selected liquid chromatographic conditions on resolution and capacity factors. J Liq Chromatogr 1990; 13:525.
3. Daniel C. Use of half normal plots in interpreting factorial two-level experiments. Technometrics 1959; 1:311.
4. Box GE, Hunter WG, Hunter JS. Statistics for Experimenters. New York: Wiley, 1978.
5. Schwartz JB, Flamholtz JR, Press RH. Computer optimization of pharmaceutical formulations. I. General procedure. J Pharm Sci 1973; 62:1165.
6. Fonner DE Jr, Buck JR, Banker GS. Mathematical optimization techniques in drug product design and process analysis. J Pharm Sci 1970; 59:1587.
7. Lindberg N-O, Jonsson C, Holmquist B. Optimization of disintegration time and crushing strength of a tablet formulation. Drug Dev Ind Pharm 1985; 11(4):931–943.
8. Box GEP, Draper NR. Evolutionary Operations. New York: Wiley, 1969.
9. Spendley W, Hext GR, Himsworth FR. Sequential application of simplex designs in optimization and evolutionary operation. Technometrics 1962; 4:441.
10. Nelder JA, Mead R. A simplex method for function minimization. Comput J 1965; 7:308.
11. Shek E, Ghani M, Jones RE. A new attempt to solve the scale-up problem for granulation using response surface methodology. J Pharm Sci 1980; 69:1135.
12. Gorman JW, Hinman JE. Simplex lattice designs for multicomponent systems. Technometrics 1962; 4:463.

‡‡Optional, more advanced problem.