CHAPTER 2

Completely Randomized Design

In this chapter we consider the use of completely randomized designs, both with and without restrictions on randomization and with and without the use of Analysis of Means (ANOM).

2.1 COMPLETELY RANDOMIZED DESIGN

As stated in Chapter 1, complete randomization is not always possible, and when it isn't possible, the ramifications of restricted randomization must be understood (see Section 4.19). We will initially assume, however, that complete randomization *is* possible, and later in the chapter will relax that assumption and discuss the consequences of restricted randomization.

When we have a single factor, complete randomization means that (1) the levels of the factor are assigned to the experimental units in a random fashion, and (2) the order in which the experiment is carried out after the assignment has been made is also random. The latter is important when one is conducting a physical experiment, in particular, but may not even be applicable for many other types of experiments. For example, for the teacher experiment described in Section 1.4, there would be no separate runs within each level of the factor "method of instruction," as all students would be subject to a particular method of instruction at the same time. If, however, the factor was "temperature," it might be highly impractical, if not impossible, to randomly change temperatures during an experiment. If, however, three levels of temperature were used with several values of the response variable recorded at each temperature level and the temperatures in the experiment changed successively from the lowest level to the highest level, an apparent temperature effect could be confounded with the effect of one of more extraneous variables (i.e., lurking factors) that could be influencing the value of the response variable over time.

If the temperatures were randomized, a plot of the response variable over time should not show almost strictly increasing values of the response variable. If such a

Modern Experimental Design By Thomas P. Ryan

plot did occur, this would almost certainly mean that one or more extraneous factors were affecting the results. But if the temperatures were not randomized, we couldn't tell from such a plot whether the trend was due primarily to the temperature effect (with extraneous factors perhaps having a small effect), or was the trend due almost exclusively to the effect of extraneous factors.

Of course if there were a positive temperature effect, we might observe level shifts, with points randomly scattered about the midline for each level, but we would not expect to observe the response values strictly increasing. Such a plot would likely suggest that the response values are getting a boost from the effect of at least one extraneous factor.

2.1.1 Model

The model for a completely randomized design with a single factor can be written as

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \qquad i = 1, 2, \ldots, k \quad j = 1, 2, \ldots, n_i \tag{2.1}$$

with μ the overall mean (which would be estimated by the mean of all the observations), Y_{ij} is the jth observation for the ith level of the single factor, A_i is the effect of the ith level of the factor, with n_i observations for each level, and ϵ_{ij} is the corresponding error term, which is assumed to have a normal distribution with a mean of zero and a variance that is the same for each level. The errors are also assumed to be independent. Stated compactly, the assumption is, for each i, $\epsilon_{ij} \sim \text{NID}(0, \sigma_\epsilon^2)$. (Hereinafter, σ_ϵ^2 will generally be written simply as σ^2.) As discussed in Section 1.6.2.2, the levels of the factor may be either selected at random from a range of levels that is of interest, which would make the factor a *random factor*, or specific levels of interest might be used, which would make the factor a *fixed factor.* For a single factor the analysis is the same regardless of how the factor is classified, although the inference that is drawn from the data is different.

That is, for the model given by Eq. (2.1), if the factor is fixed, the null hypothesis that is tested is H_0: $A_i = 0$ for $i = 1, 2, \ldots, k$, which is the same as H_0: $\mu_1 = \mu_2 = \cdots \mu_k = \mu$. This becomes clear if we recognize that $\mu_i = \mu + A_i$. Whether H_0 is true or not, the side condition $\sum_{i=1}^{k} A_i = 0$ is imposed. In words, this means that one or more levels of the factor will affect the response value over and above the overall mean, μ, whereas one or more levels will cause the response value to be less than μ. The need for the side condition should be apparent if we sum both sides of Eq. (2.1). We would logically estimate μ by $\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} / \sum_{i=1}^{k} n_i$ and this estimator will occur from Eq. (2.1) only if the side condition is imposed in addition to the mean of ϵ_{ij} being zero so that the sum of ϵ_{ij} is zero.

If the factor were random, then the appropriate test would be H_0: $\sigma_A^2 = 0$. That is, in the fixed effects case the interest is solely on the effect of the levels of the factor used in the experiment, whereas in the random effects case the null hypothesis states that there is no effect of *any* level of A within the range of interest. The form of these hypothesis tests also applies more generally when there is more than one factor.

Assumptions should, of course, always be checked, and the normality and constant variance assumptions should therefore be checked. Methods for doing so are

illustrated in subsequent chapters. From a practical standpoint, if the means differ considerably, the variances may differ more than slightly as the variance is generally related to the magnitude of the numbers. Thus, if we suspect that the means may differ considerably, it is especially important to test the equal variances assumption.

As far as parameter estimation is concerned, it would seem appropriate for A_i to be estimated by $\overline{y_i} - \overline{\overline{y}}$, with the latter denoting the overall average, and this is how A_i is estimated. The sum of the $\widehat{A_i}$, the estimates of the A_i, is zero when the n_i are equal, as the reader is asked to show in Exercise 2.1. (When the n_i differ, $\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{\overline{y}}) = 0$.) The variance of the error term, σ_ϵ^2, is estimated analogous to the way it is estimated in an independent sample t-test: by pooling the variances within each level.

2.1.2 Example: One Factor, Two Levels

Assume that an experiment is performed involving 120 patients. The objective of the experiment is to test a blood pressure medication against a purported placebo, but the placebo is actually garlic, suitably disguised. Sixty patients are randomly assigned to the medication, with the other 60 patients assigned to the placebo. The study is double blinded so that the investigators do not know the patient–medication/placebo assignment, and of course the patients don't know this either. The correct assignment is known only by the person who numbered the bottles, with this information later used to properly guide the computer analysis.

Assume further that deviation from diastolic blood pressure at the start of the experiment is used for the analysis, with three measurements taken for each of three consecutive days starting 60 days after the experiment began, with the average of those nine measurements used as a single number for each of the 120 patients. Thus there are 120 numbers. We revisit this problem briefly in Section 11.1.1 in which repeated measures designs are presented. With "P" denoting the placebo and "M" denoting the medication, the results are given below.

```
M -2 -7  -4 -8  -7 -4 -4 -1 -2 -10 -3  0  1 -10 -4 -4  -7 -3 -7
P -2 -5  -8 -4  -2 -2  1 -8  1  -2 -3 -8 -1  -7 -1  0  -8 -7 -7
M -9 -2  -1  1  -3 -7 -9 -6 -4  -8  1 -1  2  -2 -1 -8  -8 -1 -9
P -1 -5  -7 -2  -8 -2 -4 -4 -6  -5 -6 -3  0  -5 -8 -4  -8 -1 -4
M -4 -6 -10  0 -10 -3 -6 -1 -5  -7  1 -4  2  -5 -7 -9 -10 -6 -8
P  1  1  -5  1   0 -2  1  0 -7  -4 -2 -5 -8  -7  1 -7  -5 -6 -8
M -1 -9  -5
P -7 -8  -3
```

2.1.2.1 Assumptions

As stated previously, the assumptions must be kept in mind when the data are analyzed. One assumption that is rarely addressed is that the observations within each level of a factor must be independent. This is because the variance of an average, such as the average deviation for the medication over the 30 patients, is assumed to be σ^2/n, but this will be true only if the observations that comprise the average are independent. As Czitrom (2003) stated, "Perhaps the single most important issue related to the application of statistics in the semiconductor industry is the frequent *lack*

of independence of observations . . . . The lack of independence affects the application of such basic statistical tools as *t*-tests, confidence intervals, analysis of variance, and control charts." By no means is this problem confined to the semiconductor industry, so independence *between observations within a group* must be checked so that the application of methods such as those given in this book will not be undermined.

We would not expect that assumption to be violated for this example because the experimental units for each level of the factor are different people. However, this does not preclude the possible effect of a lurking variable, so the data for the medication and the placebo should each be plotted over time.

2.1.2.1.1 Checking the Assumptions

Our analysis proceeds as follows. Time sequence plots for each set of 60 measurements do not exhibit any unusual patterns, and an autocorrelation plot of each set reveals no significant autocorrelations. So there appears to be independence and stability within each set of measurements. The normal probability plots for each set exhibit clear evidence of nonnormality. Neither these plots nor a histogram of each set provides strong evidence of skewness, however, so with moderately large sample sizes we can proceed by realizing that the sample means will be approximately normally distributed, and by realizing that a *t*-test will be robust to small-to-moderate departures from normality of the sample means. Levene's test for equal variances has a *p*-value of .27; since this is much larger than .05 or .01, we can proceed to perform a pooled *t*-test. If the *p*-value had been, say, .02, the generalized *F*-test for the unequal variances case given by Weerahandi (1994) would have to be used.

The results of the test are given below.

```
Two-Sample T-Test and CI: M, P

Two-sample T for M vs P

    N   Mean   StDev   SE Mean
M   60  -4.57   3.53      0.46
P   60  -3.92   3.08      0.40

Difference = mu M − mu P
Estimate for difference: -0.650
95% CI for difference: (−1.848, 0.548)
T-Test of difference = 0 (vs not =): T-Value = -1.07
  P-Value = 0.285. DF = 118
Both use Pooled StDev = 3.31
```

The results show that there is not sufficient evidence to reject the null hypothesis of an equal mean difference from the starting blood pressures for the medication group and the placebo group. Can we then conclude that the medication is ineffective? Since garlic is known to lower blood pressure, such a conclusion cannot be drawn, and in fact a one-sample *t*-test for the medication yields a *p*-value of less than .001, thus providing strong evidence that the medicine was effective. Thus, a completely erroneous conclusion could be drawn from the two-sample test, if the placebo is not a true placebo.

```
One-way ANOVA: M, P

Analysis of Variance
Source      DF       SS      MS         F        P
Factor       1     12.7    12.7       1.15    0.285
Error      118   1295.3    11.0
Total      119   1308.0
                                Individual 95% CIs For Mean
                                Based on Pooled StDev
Level        N     Mean   StDev -------+---------+---------+---
M           60   -4.567   3.529    (----------- * -----------)
P           60   -3.917   3.082             (---------- * ----------)
                                -------+---------+---------+---------
Pooled  StDev =   3.313              -4.90     -4.20     -3.50
```

Equivalently, the two-sample data can be analyzed using Analysis of Variance (ANOVA), with the output given above. In fact, we may state, loosely speaking, that the ANOVA table contains the square of the information/data from the two-sample t-test, with the same null hypothesis tested. For example, the F-statistic value of 1.15 is, without rounding, equal to $(1.07)^2$. Similarly, the square of the "pooled standard deviation" of 3.31 from the output for the two-sample t-test is, within rounding, equal to 11.0, the MS_{error} value from the ANOVA table, which estimates σ^2.

The manner in which the numerical values for the components of the ANOVA table are computed is explained in detail in Section 2.1.3.3, in addition to a detailed discussion of what ANOVA provides.

2.1.3 Examples: One Factor, More Than Two Levels

When there is a single factor with more than two levels, the experimenter does not have the option of using a t-test, since a t-test is applicable only when there are at most two levels. This is due to the fact that with a t-test either the equality of two means is being tested or a specified value of a single mean is tested. Analysis of Variance can be used, however, when there are any number of levels.

It is of course important to understand what ANOVA provides. To some extent, the term "Analysis of Variance" is a misnomer because what is analyzed is variation—variation due to different sources. To illustrate, consider the following hypothetical data.

1	2	3
8.01	8.03	8.04
8.00	8.02	8.02
8.02	8.01	8.02
8.01	8.02	8.03
8.01	8.02	8.04
8.02	8.03	8.02
8.00	8.01	8.04

The averages for the three levels are 8.01, 8.02, and 8.03, respectively. Although these averages differ only slightly, when we test the hypothesis that the three population means (μ_1, μ_2 and μ_3) are equal (using a methodology to be given shortly), we easily reject that hypothesis because the p-value for the test is .002, with the computer output as follows. (The computations that underlie the analysis are explained in Section 2.1.3.3; statistical significance results because of the very small variability within each level.)

```
     Analysis of Variance
Source     DF      SS         MS         F        P
Factor      2   0.0014000  0.0007000   9.00    0.002
Error      18   0.0014000  0.0000778
Total      20   0.0028000
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level       N      Mean      StDev -------+---------+---------+---------
1           7   8.01000    0.00816   (----------*----------)
2           7   8.02000    0.00816             (---------*---------)
3           7   8.03000    0.01000                  (--------*-------)
                                   -------+---------+---------+-------
Pooled StDev = 0.00882                  8.010     8.020     8.030
```

2.1.3.1 Multiple Comparisons

In general, there is a need to determine which of the three population means differ since we are rejecting the null hypothesis of equality of the means. If we observed actual data like this, we would suspect that the difference in the means would not likely be of any practical significance. To finish the example, however, since the confidence intervals for μ_1 and μ_3 do not overlap, as the computer output shows, we would logically conclude that these two means differ, and this is why the hypothesis of equality of the three means was rejected.

Looking to see if confidence intervals for means overlap is essentially an ad hoc approach. Nevertheless, such an approach is not necessarily a bad idea because it is both simple and intuitive. Another approach that is similarly intuitive is the *sliding reference distribution* approach given by Box, Hunter, and Hunter (1978, p.191). The general idea is to construct a t-distribution with the appropriate scale factor of $\frac{s}{\sqrt{n}}$ for equal sample sizes, and an approximate scale factor of $\frac{s}{\sqrt{\overline{n}}}$ with $\overline{n}$ denoting the average sample size if the sample sizes do not differ greatly. A dotplot of the means is constructed and the idea is to see if the t-distribution can be positioned in such a way as to cover as many of the means as possible. The means that cannot be covered by the (sliding) t-distribution are said to be different from the other means. Of course it would be cumbersome to have a cutout that would be used to physically slide along a dotplot, but of course this could be handled rather easily with a Java applet, although this has apparently not been done.

For the present example, $\frac{s}{\sqrt{n}} = \frac{0.00882}{\sqrt{7}} = 0.00333$. If we center the sliding t-distribution at the mean of 8.02, then $8.02 \pm t\frac{s}{\sqrt{n}}$ are at almost exactly 8.01 and

8.03 when $t = 3$. The latter is an extreme value for 18 degrees of freedom for the error term, as in this example, with $P(t_{18} > 3) = .004$. Since the latter is a small value, we would conclude that all three means differ using this approach.

There are various *multiple comparison procedures*, as they are called, from which an experimenter can select. Some of these procedures are for use when the comparisons to be made are selected before the data are collected, and others are for use after the experimenter looks at the data. Some are conservative, some are not conservative; one method (Dunnett's procedure) is used when testing against a control. It is interesting to note that Box et al. (1978) did not present any of these methods, however, instead opting for their sliding reference distribution approach. (The same is true of Box, Hunter, and Hunter, 2005.)

A conservative multiple comparison procedure is one that is based on Bonferroni's inequality, which was named after an Italian mathematician Carlo Bonferroni (1892–1960). The inequality stated that for events $B_1, B_2, \ldots, B_q$ (which in this case will be confidence intervals),

$$P\Big(\bigcap B_i\Big) \geq 1 - \sum_{i=1}^{q} [1 - P(B_i)]$$

or

$$P\Big(\bigcap B_i\Big) \geq 1 - \sum_{i=1}^{q} P(\overline{B}_i) \tag{2.2}$$

with $P(\overline{B}_i) = [1 - P(B_i)]$ denoting the probability that event B_i does not occur.

Applying this result to experimental design, let B_i denote the event that a confidence interval for a treatment effect or a linear combination of treatment effects contains the treatment effect or linear combination of treatment effects, so that $P(B_i)$ is the probability of this occurrence, and $P(\overline{B}_i)$ is the probability that the confidence interval does not include the unknown treatment effect.

It follows from the inequality given in expression (2.2) that if the objective were to have the entire set of confidence intervals cover the respective treatment effects with probability of at least $1 - \alpha$, each interval could be a $1 - \alpha/q$ confidence interval, so that the probability that each interval does not cover the treatment effect (i.e., $P(\overline{B}_i)$) would be α/q. Then from expression (2.2), $P(\bigcap B_i) \geq 1 - q(\alpha/q)$ so that $P(\bigcap B_i) \geq 1 - \alpha$.

Each confidence interval would be of the general form

$$\sum_{i=1}^{k} c_i \widehat{A}_i \pm t_{n-\nu,\alpha/2q} \sqrt{\widehat{\text{Var}}\left(\sum_{i=1}^{k} c_i \hat{A}_i\right)}$$

with A_i denoting the effect of the ith treatment, as in Eq. (2.1), $\widehat{A}_i$ denoting the estimator of that effect, and ν denoting the degrees of freedom for the error term. The

value of k is the number of treatment effects involved in the confidence interval. For two treatment effects, a logical comparison would be $A_1 - A_2$, for which the constants would be $c_1 = 1$ and $c_2 = -1$. In general, $\sum_{i=1}^{k} c_i = 0$ for each comparison, which must be planned before collecting the data. To do otherwise would be to bias the results.

Since $\text{Var}(\widehat{A}_i) = \sigma^2/n_i$ and $\widehat{\sigma}^2 = \text{MS}_{\text{error}}$, as stated in Section 2.1.2.1.1, and $\widehat{A}_i = \overline{y}_i$, we may write the confidence interval as

$$\sum_{i=1}^{k} c_i \bar{y}_i \pm t_{n-v,\alpha/2q} \sqrt{\text{MS}_{\text{error}} \sum_{i=1}^{k} c_i^2/n_i}$$

The probability that each of the q comparisons contains $\sum_{i=1}^{k} c_i A_i$ for $i = 1, 2, \ldots, q$ is at least $1 - \alpha$ and quite likely much greater than $1 - \alpha$.

One of the best known multiple comparison procedures is due to Scheffé (1953), which is for every possible contrast $\sum_{i=1}^{k} c_i \mu_i$, with $\mu_1, \mu_2, \ldots, \mu_k$ denoting the k treatment means and the c_i being arbitrary constants but with the restriction that $\sum_{i=1}^{k} c_i = 0$. The Scheffé procedure, which can be used when decisions about which comparisons to make are made after examining the data, gives a set of simultaneous $100(1 - \alpha)\%$ confidence intervals with the objective being to see if any of the intervals do not cover zero. If so, the corresponding contrast (such as $\mu_1 - \mu_2$) is significant and in this example the conclusion would be that $\mu_1 \neq \mu_2$. The confidence intervals are of form

$$\sum_{i=1}^{k} c_i \overline{y}_i \pm \sqrt{(v-1)F_{v-1,n-v,\alpha}} \sqrt{\text{MS}_{\text{error}} \sum_{i=1}^{k} c_i^2/n_i}$$

where v is defined as it was for the Bonferroni intervals. Notice that the form for the Scheffé intervals differs from the form for the Bonferroni intervals only in terms of the first component after the $\pm$ sign. There are various other multiple comparison procedures, including those due to Tukey (1953), Duncan (1955, 1975), Dunnett (1955), Fisher (1935), Kramer (1956), and Hsu (1984).

These methods are discussed in detail in Hsu (1996) and Hochberg and Tamhane (1987), and most are also discussed in detail in Dean and Voss (1999). It is worth noting that some of these papers have been extensively cited, with Duncan (1955) being the third most cited paper in the list of the 25 most cited papers given by Ryan and Woodall (2005), whereas Dunnett (1955) is 14th, Dunnett (1964) is 21st, and Kramer (1956) is 22nd.

In general, multiple comparison procedures are fraught with problems and controversies. When used, a procedure should be carefully selected and used appropriately. A good, relatively recent online treatise on multiple comparison procedures is Dallal (2001), which is available at http://www.tufts.edu/∼gdallal/mc.htm.

At the other extreme, consider the following data.

1	2	3
7.06	4.91	4.87
3.13	9.04	8.01
4.12	5.95	6.76
5.59	3.86	7.98
5.10	6.24	7.38

Here the averages are 5.0, 6.0, and 7.0, respectively, but the hypothesis of equal means is not rejected, as the p-value is .184. The computer output is as follows.

```
Analysis of Variance
Source    DF       SS       MS       F        P
Factor     2    10.00     5.00    1.96    0.184
Error     12    30.66     2.56
Total     14    40.66
                                Individual 95% CIs For Mean
                                Based on Pooled StDev
Level      N     Mean   StDev  -------+---------+--------+--------
1          5    5.000   1.489    (----------*----------)
2          5    6.000   1.941           (---------*---------)
3          5    7.000   1.296                   (--------*-------)
                                -------+---------+---------+-------
Pooled StDev =  1.598              4.5       6.0       7.5
```

Notice that the confidence intervals overlap, and notice that the standard deviations of the means in this example are larger, relative to the difference in the means, than are the standard deviations in the previous example. This helps explain why the conclusions differ. Using the applet that was used in Section 1.4.4, which is found at http://www.stat.uiowa.edu/∼rlenth/Power/index.html, the probability of detecting a difference of $\Delta = 2.005$ (which is essentially the difference between the first and third means above) is only .4454 when $\sigma = 1.6$ and $\alpha = .05$. The less than 50–50 chance of detecting the stated difference is due largely to the fact that σ is large relative to the magnitude of the numbers. (Another method of comparing means is ANOM, which is described in Section 2.2.)

Thus, we reject the null hypothesis when the means differ by only 0.01, but we fail to reject the null hypothesis when the means differ by 1.0. The reason for this is that in the second example there is considerable within-level variability, which drowns out the between-level variability. The reverse occurs in the first example as the within-level variability is so small that it does not offset the between-level variability, although the latter is obviously small.

2.1.3.2 Unbalanced and Missing Data

Although the examples given in this chapter have the same number of observations for each level of the factor, this is not a requirement as the number of observations

per level could differ, as is implied by model (2.1). Unbalanced data can be easily handled when a completely randomized design is used, although we would naturally prefer that the mean for each level of the factor be estimated from the same number of observations. Unbalanced data do present a problem when there is more than one factor, however, and imputation methods are discussed briefly in Section 4.12.2. If such methods are not used, then methods for analyzing unbalanced data must be employed, and the reader is referred to Searle (1987) for the proposed methods of analysis. Since unbalanced data present no problem with a completely randomized design, it follows that missing data that cause unbalanced data also do not present a problem unless the missing data are numerous and/or are not missing at random.

2.1.3.3 Computations

If we were to devise a measure of the variability between the levels of a factor, one obvious choice is to use some function of the difference between the average value for each level and the overall average. We can't sum those differences, however, because the sum will always be zero, as was stated in Section 2.1.1. We can, however, sum the squares of the differences, and that sum is multiplied by the number of observations for each level, if that number is constant across all levels. (The reason for the multiplier will soon become apparent.)

For a measure of variability within the levels, the obvious choice is to compute, for each level, the square of each observation from the level average and sum the squares over the observations in each level, and then sum over the levels.

An obvious choice as a measure of the total variability would be the sum of the squares of the observations from the overall average. Adopting the notation that was used in Section 2.1.1, we have the following *Analysis of Variance Identity* when the number of observations per level (i.e., the n_i) is constant and equal to n, as in the two examples in Sections 2.1.3 and 2.1.3.1.

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij}-\overline{\overline{y}})^2 \equiv n\sum_{j=1}^{k}(\overline{y}_i-\overline{\overline{y}})^2+\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij}-\overline{y}_i)^2 \tag{2.3}$$

(Only a slight modification to this expression is needed if the n_i differ.) This equivalence is very easy to derive and can be accomplished by subtracting and adding $\overline{y}_i$ within the parentheses on the left side, squaring, and then simplifying, as the reader is asked to show in Exercise 2.2.

The first term on the right side of Eq. (2.3) gives what was labeled "Factor" in the computer output, and the second term gives what was labeled "Error." It could be shown with a small amount of algebra that this term is the extension to k levels of what the numerator of s_p^2 would be in Section 1.4.4 if that numerator were written in terms of the appropriate summation expressions rather than in terms of the two variances. If this is accepted, which the reader is asked to show in Exercise 2.17, it could then be shown (also Exercise 2.17) that MS_{error} is simply the average of the variances for each level when there is an equal number of observations in each level.

Notice that *if* hand computation were performed, the error sum of squares would be obtained by subtracting the factor sum of squares from the total sum of squares

that is given on the left side of the equation. Obviously only two terms in the identity would have to be computed, with the other obtained by addition or subtraction. (The total sum of squares is computed the same way for every model and experimental design; only the form of the right side is model/design dependent.)

The *degrees of freedom* (df) can be similarly partitioned. Since $\sum_{i=1}^{n}\sum_{j=1}^{k}(y_{ij} - \overline{\overline{y}}) = 0$, the sum does not contain kn independent pieces of information. Rather, only $kn - 1$ components are independent since the sum is zero. Accordingly, only $kn - 1$ components on the left side of Eq. (2.3) are independent and "free to vary." Thus, that sum has $kn - 1$ df. Similarly, $\sum_{j=1}^{k}(\overline{y}_i - \overline{\overline{y}}) = 0$, so only $k - 1$ components of the sum are free to vary; thus, the first term on the right side of the equation has $k - 1$ df. Since degrees of freedom are additive, it follows that the second sum on the right side of the equation must have $kn - k$ df.

We may summarize the degrees of freedom breakdown as follows, assuming an equal number of observations per treatment.

Analysis of Variance

Source	df
Treatments	$k - 1$
Error	$k(n - 1)$
Total	$kn - 1$

If the n_i are not all the same, then the total degrees of freedom is $\sum_{i=1}^{k} n_i^{-1}$ and the error degrees of freedom is $\sum_{i=1}^{k} n_i - k$.

2.1.4 Example Showing the Effect of Unequal Variances

Weerahandi (2004) gave an example that showed how the F-test for the equality of treatment means can produce a result that differs from the result obtained using a heteroscedastic ANOVA approach when the population variances are apparently unequal.

An engineer at a construction company was interested in testing the comparative strength of four brands of reinforcing bars. Although Weerahandi (2004) didn't give the unit of measurement (and the data are presumably hypothetical), the data are as follows.

BRAND A	21.4	13.5	21.1	13.3	18.9	19.2	18.3		
BRAND B	27.3	22.3	16.9	11.3	26.3	19.8	16.2	25.4	
BRAND C	18.7	19.1	16.4	15.9	18.7	20.1	17.8		
BRAND D	19.9	19.3	18.7	20.3	22.8	20.8	20.9	23.6	21.2

The means for BRANDS A–D are 17.96, 20.68, 18.10, and 20.83, respectively, whereas the variances are 3.07, 5.28, 1.39, and 1.48, respectively.

It is useful to begin the analysis by looking at boxplots for the four brands, which are given in Figure 2.1. (Reese (2005) stated, "Make it a rule: never do ANOVA without

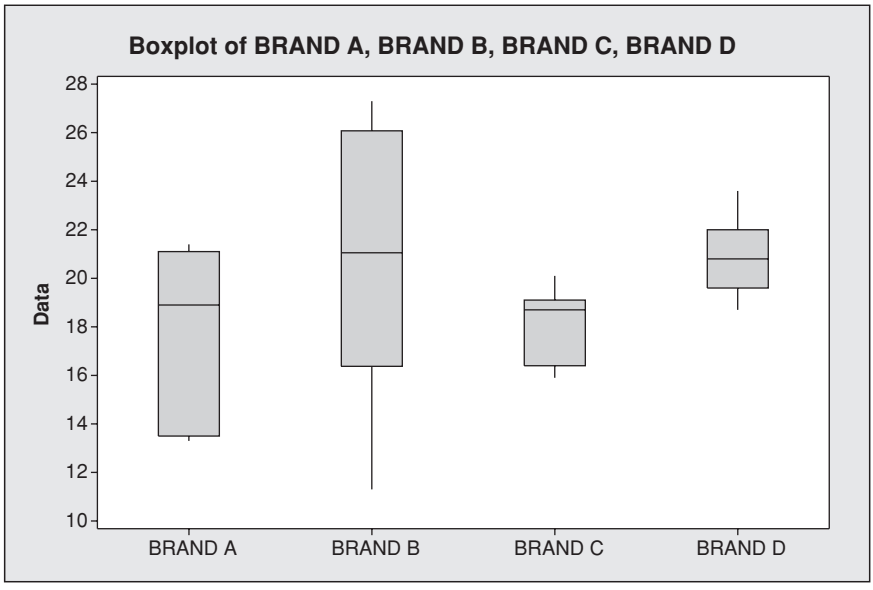

Figure 2.1 Boxplots of brand data.

a boxplot." Although such a strong statement will not be made here, especially since boxplots are of very limited value when there is more than one factor, boxplots for designs with single factors are very useful.)

The heteroscedasticity is apparent by comparing the last two brands with the first two brands.

The means are not directly comparable because the variability is so much less for the last two brands than it is for the first two brands. If this disparity is ignored and an F-test performed, $F = 1.61$ is obtained, which has a p-value of .211. Thus, the conclusion is that the means do not differ. Certainly Figure 2.1 shows that the medians do not differ greatly. (We may note that the Kruskal–Wallis nonparametric test is not applicable here, since it is based on the assumption that the populations have the same continuous distribution except for possibly different medians.) Weerahandi (2004, p. 51) obtained a p-value of .021 when applying the generalized F-test given by Weerahandi (1994).

2.2 ANALYSIS OF MEANS

Although ANOM has been in use for decades and has been part of the MINITAB software for many years, and is also included in SAS/QC 9.0 and 9.1 from SAS Software, it is still not well known to people who use designed experiments. In fact, many people would undoubtedly confuse ANOM with ANOVA, since the latter also

involves an analysis involving means. It is apt to have more appeal to engineers and other industrial personnel than does ANOVA, however, since ANOM is inherently a graphical procedure and is in terms of the original unit(s) of measurement, whereas ANOVA is in the square of the original unit(s), as was stated previously in Section 1.9.

Analysis of Means is not a full substitute for Analysis of Variance, however, as ANOM can be used only for fixed factors, whereas ANOVA can be used for fixed or random factors, or for a combination of the two.

The reader will recall that with ANOVA the experimenter concludes either that all of the means are equal, or that at least one of the means differs from the others. One procedure need not be used to the exclusion of the other, however. As Ott (1967) indicates, ANOM can be used either alone or as a supplement to ANOVA.

2.2.1 ANOM for a Completely Randomized Design

It was stated previously that with ANOM one compares $\overline{x}_i$ against the average of the $\overline{x}_i$, which will be denoted by $\overline{\overline{x}}$, analogous to the notation used for an $\overline{X}$ chart. The original ANOM methodology given by Ott (1958, 1967) was based upon the multiple significance test for a group of means given by Halperin, Greenhouse, Cornfield, and Zalokar (1955), which was based upon the studentized maximum absolute deviate. That approach provided an upper bound for the unknown critical value, but will not be discussed here since it is no longer used. The interested reader is referred to Schilling (1973) for more details, including the theoretical development. The current approach is based upon the exact critical value, h, and is described in L. S. Nelson (1983).

If we were testing for the significance of a single deviation, $\overline{x}_1 - \overline{\overline{x}}$, it would stand to reason that we would want to look at some test statistic of the form

$$\frac{\overline{x}_i - \overline{\overline{x}} - E(\overline{x}_i - \overline{\overline{x}})}{s_{\overline{x}_i - \overline{\overline{x}}}} \tag{2.4}$$

where E stands for expected value. If $\mu_i = (\mu_1 + \mu_2 + \cdots + \mu_k)/k$, then $E(\overline{x}_i - \overline{\overline{x}}) = 0$, and since the former is what would be tested, we take $E(\overline{x}_i - \overline{\overline{x}})$ to be zero. (We should note that some authors have indicated that the null hypothesis that is tested with ANOM is H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$. Certainly if each μ_i is equal to the average of all the means, then it follows that the μ_i must all be equal, since they are equal to the same quantity. But stating the null hypothesis in this alternative way obscures the testing that is done.)

It can be observed that Eq. (2.4) becomes a t-test when $k = 2$ since $\overline{x}_i - \overline{\overline{x}}$ is then $\overline{x}_1 - (\overline{x}_1 + \overline{x}_2)/2 = (\overline{x}_1 - \overline{x}_2)/2$ for $i = 1$ (and $(\overline{x}_2 - \overline{x}_1)/2$ for $i = 2$) so that

$$t = \frac{(\overline{x}_1 - \overline{x}_2)/2 - 0}{s_{(\overline{x}_1 - \overline{x}_2)/2}} = \frac{\overline{x}_1 - \overline{x}_2}{s_{\overline{x}_1 - \overline{x}_2}}$$

since the 2s cancel.

The two deviations $\overline{x}_1 - \overline{\overline{x}}$ and $\overline{x}_2 - \overline{\overline{x}}$ are thus equal, so we conclude that the two means differ if

$$t = \frac{|\overline{x}_1 - \overline{x}_2|}{s_{\overline{x}_1 - \overline{x}_2}} > t_\alpha$$

for a selected value of α.

When $k > 2$, the t-distribution cannot be used, however, so another procedure is needed. It can be shown that, assuming equal sample sizes, the deviations $\overline{x}_i - \overline{\overline{x}}$ are equally correlated with correlation coefficient $\rho = -1/(k-1)$. If we let $T_i = (\overline{x}_i - \overline{\overline{x}})/s_{\overline{x}_i - \overline{\overline{x}}}$, the joint distribution of $T_1, T_2, \ldots, T_k$ is an equicorrelated multivariate noncentral-t distribution, assuming that the sample averages are independent and normally distributed with a common variance (see P. R. Nelson 1982, p. 701).

Exact critical values for $k > 2$ were first generated by P. R. Nelson (1982), with a few tabular values subsequently corrected, and the corrected tables published in L. S. Nelson (1983). More complete and more accurate critical values given to two decimal places were given by P. R. Nelson (1993). These values differ by one in the second decimal place from some of the critical values given in L. S. Nelson (1983).

The general idea is to plot the averages against *decision lines* obtained from

$$\overline{\overline{x}} \pm h_{\alpha,k,\nu}\, s\sqrt{(k-1)/(kn)} \tag{2.5}$$

where n is the number of observations from which each average is computed, ν is the degrees of freedom associated with s, the estimate of σ, k is the number of averages, and $h_{\alpha,k,\nu}$ is obtained from the tables in P. R. Nelson (1993) for a selected value of α, with those tabular values given in this book in Table D. It is demonstrated in the appendix to this chapter that $s\sqrt{(k-1)/(kn)}$ is the estimate of $\sigma_{\overline{x}_i - \overline{\overline{x}}}$.

Analysis of Means can be used when the sample sizes are unequal; the expression for the decision lines is just slightly different. Specifically, the radicand is $\frac{N-n_i}{Nn_i}$, as is shown in the chapter Appendix. This causes the decision lines to be somewhat aesthetically unappealing, since the distance between the lines varies as the n_i vary. An example of an ANOM graph with varying decision lines for unequal sample sizes is given by Nelson, Coffin, and Copeland (2003, p. 262).

The value of α is the probability of (wrongly) rejecting the hypothesis that is being tested when, in fact, the hypothesis is true. (Here we are testing that each mean is equal to the average of all the k means, as indicated previously.)

2.2.1.1 Example

We will use the example in Section 2.1.3.1 for illustration, assuming the factor to be fixed. If we were using hand computation, the first step would be to compute the overall average, $\overline{\overline{x}}$, and then compute the decision lines from Eq. (2.5) for a selected value of α and plot the averages for the factor levels. Of course we know from Section 2.1.3.1 that the averages are 5, 6, and 7, respectively, and of course the overall average is 6. An ANOVA would generally be performed to obtain the value of s, and that output showed the value to be 1.598. The value of $h_{\alpha,k,\nu}$ is the value of

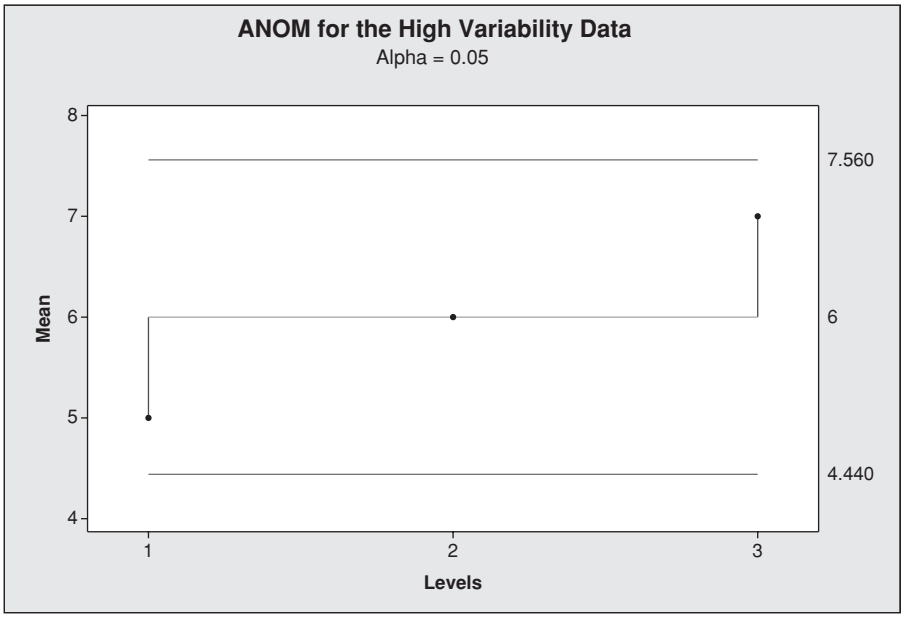

Figure 2.2 ANOM display for the high variability data.

$h_{.05,3,12}$, which is 2.67. Thus, the decision lines, as they are called, are obtained from $6 \pm 2.67(1.598)\sqrt{\frac{2}{3(5)}} = (4.44, 7.56)$, which are the numbers in the display (Fig. 2.2).

It can be observed that the means are well inside the $\alpha = .05$ decision lines, with .05 being the experiment-wise error rate (i.e., for the three tests that each of the three means is equal to the average of all of the means), so the conclusion is that no population mean differs from the average of the population means. Ott (1975) showed multiple sets of decision lines on ANOM displays, such as for .01, .05, and .10, but that option is not available with MINITAB, which is being used to produce these displays.

The ANOM display for the low variability data is shown in Figure 2.3. The conclusion is that the first and third population means differ from the averages of all the means since the first and third averages plot outside the .05 decision lines. Of course this result is very intuitive as the analysis using ANOVA showed a significant result and the two "extreme" averages of 8.01 and 8.03 are equidistant from the overall average of 8.02.

2.2.2 ANOM with Unequal Variances

There is also an ANOM procedure, due to Nelson and Dudewicz (2002), that can be used when there is evidence of unequal variances. This might be viewed as being analogous to the t-test that does not assume equal variances, which can be used when the pooled t-test cannot be used because of evidence of (highly) unequal variances.

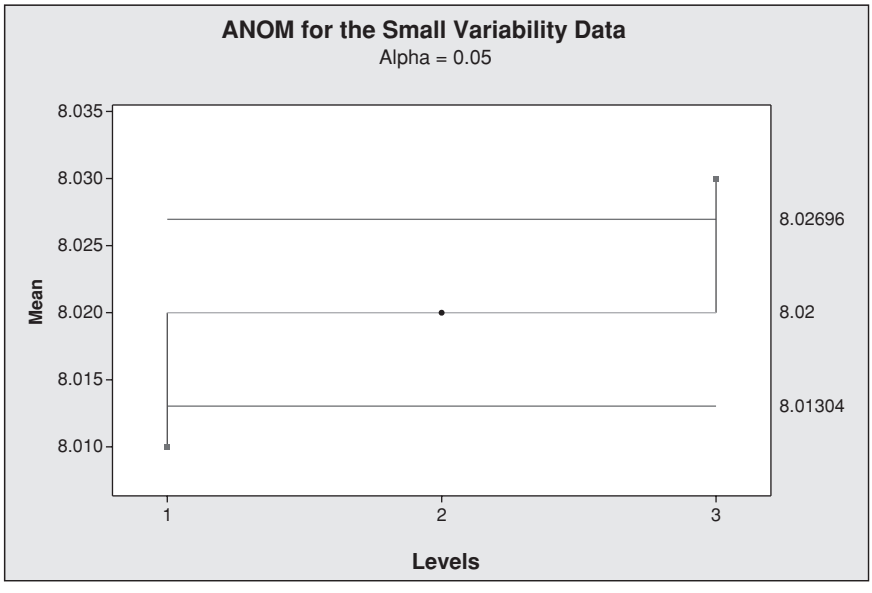

Figure 2.3 ANOM display for the low variability data.

Nelson and Dudewicz (2002) dubbed their procedure HANOM (heteroscedastic analysis of means). Before this procedure is used, the variances should be tested, and in the one-factor case it is quite possible that there will be enough observations per factor level for such a test to have reasonable power for detecting unequal variances.

One possible test would be Levene's or the modification of it, both of which were mentioned in Section 1.4.3, or if one wanted to stick strictly with ANOM procedures, the analysis of means for variances (ANOMV), due to Wludyka and Nelson (1997), might be used.

If the results from one of these methods provided evidence that the variances are more than slightly unequal, HANOM could be used. It should be noted, however, that unlike the ANOM procedure, HANOM is not a one-stage procedure. Instead, an initial sample is taken from the relevant populations for the purpose of computing the decision lines. A second sample is then taken and the sample means are computed. The latter are then compared against the decision lines. Thus, HANOM cannot be viewed as an alternative to heteroscedastic ANOVA unless the experimenter is willing to collect additional data.

If obtaining more data is not practical or feasible, heteroscedastic ANOVA (see, e.g., Bishop and Dudewicz, 1978) could be used, as could the generalized F-test due to Weerahandi (1994). (See also Weerahandi (2004), pp. 48–51). A second option would be to use Kruskal-Wallis ANOVA, a nonparametric method that assumes neither normality nor equal variances. It would be preferable to use heteroscedastic ANOVA if approximate normality seems to exist, however, since that would be a more powerful procedure than the Kruskal-Wallis procedure under (approximate) normality.

2.2.2.1 Applications

The fact that an initial sample is required reduces the practical value of HANOM. Although experimentation should be sequential, the number of factors investigated in the second stage is generally less than the number of factors investigated in the first stage. It is unlikely that experimenters will very often collect preliminary data just so that HANOM can be used. Although Nelson and Dudewicz (2002) do give an example, there is no evidence that real data were used.

2.2.3 Nonparametric ANOM

Analysis of Means is, like ANOVA, robust to small-to-moderate departures from normality. When the response values are strongly nonnormal, one possibility would be to try to transform the data to approximate normality. If that is unsuccessful, a nonparametric ANOM approach might be used. Bakir (1989) developed such a procedure based on ranks for use with a completely randomized design. The populations were assumed to be the same except for having possibly different means. The procedure is discussed in detail and illustrated by Nelson, Wludyka, and Copeland (2005, Section 9.3).

Another nonparametric ANOM approach is a permutation test. This can be done using either symmetric decision lines or asymmetric decision lines. For the former, N random permutations of the data are made and $D_{\max}^{(q)} = \max_i |\overline{Y}_{i(q)} - \overline{Y}_{\text{all}}|$ is computed for the qth permutation with $q = 1, \ldots, N$, with $\overline{Y}_{i(q)}$ denoting the ith treatment mean for the qth permutation and $\overline{Y}_{\text{all}}$ denoting the average of all the observations. This forms the randomization reference distribution for $\max_i |\overline{Y}_i - \overline{Y}_{\text{all}}|$. The treatment means, the $\overline{Y}_i$, are then plotted against decision lines given by $\overline{Y}_{\text{all}} \pm k_\alpha$, with k_α denoting the upper α quantile of the distribution of $D_{\max}^{(q)}$.

2.2.4 ANOM for Attributes Data

Although this book is primarily concerned with measurement data, it should be noted in passing that ANOM can be used advantageously with attributes data (see Ryan, 2000; Nelson et al., 2003; or Chapter 2 of Nelson et al., 2005). When this is done, it may be necessary to transform the attribute random variable to approximate normality if ANOM is to be used and if the normal approximation to the appropriate distribution is inadequate. It is well known that the rules of thumb for the adequacy of the normal approximation to the Poisson and binomial distributions, respectively, that are given in introductory statistics books fail in control chart applications, but their applicability in ANOM has apparently not been investigated, at least in the literature. The problem in control chart applications is that extreme tails are involved, which is not the case in ANOM. However, if we show .01 decision lines for proportions data, as recommended by Tomlinson and Lavigna (1983), we may be far enough out in the tails that there will often be problems relative to those decision lines, even if the binomial were the appropriate model and not even considering such possible problems as extrabinomial variation. Similar problems may exist for count data. This is something that should be researched.

2.3 SOFTWARE FOR EXPERIMENTAL DESIGN

Software must be used in analyzing data from designed experiments. The well-known software that can be used for design construction and analysis do differ somewhat, and Reece (2003) provides a very detailed and extensive comparison, although such comparisons become at least slightly outdated rather quickly, since companies introduce new releases of their software quite frequently. Nevertheless, the comparison given by Reece (2003) is worth reading, especially since it includes some software that are not widely known.

The software packages that are used to produce graphs and numerical output in this book are MINITAB, JMP, and Design-Expert, with the last two each receiving the highest possible rating by Reece (2003). Neither has ANOM capability, however, and SAS Software does not have the capability to directly (i.e., without programming) produce the ANOM displays that are used in Chapter 4. The ANOM capability in MINITAB is also somewhat limited, although an ANOM macro available at the MINITAB, Inc. Web site extends the capability that is provided by the ANOM command. The capabilities of another software package, D. o. E. Fusion Pro, are discussed in certain chapters, including Chapters 4 and 5, but output from the software is not used in the book.

2.4 MISSING VALUES

Unlike many of the designs that are presented in subsequent chapters, a missing value or two does not generally create a serious problem when a completely randomized design is used because data from an experiment that used such a design can be analyzed with unequal n_i, using either ANOVA or ANOM. Although missing values might be estimated (and methods for doing so have been proposed), an analysis using estimated values will be only approximate and thus not entirely satisfactory. If possible, the experimental run(s) that resulted in the missing value(s) might simply be repeated.

2.5 SUMMARY

A completely randomized design is a frequently used design that is attractive because of its simplicity. The numerical analysis is performed the same way regardless of whether the single factor is fixed or random, and the number of observations per factor level need not be the same. The user must check for possible heteroscedasticity, however, as this could undermine the results. Nonnormality could also be a problem, but only if it is moderate to severe.

The data from an experiment with a completely randomized design may be analyzed using either ANOVA or ANOM. The assumptions are the same for each and both can be used with unequal sample sizes.

APPENDIX

It was stated in Section 2.2.1 that $s\sqrt{(k-1)/(kn)}$ is the estimated standard deviation of $\overline{X}_i - \overline{\overline{X}}$, with the assumption that each $\overline{X}_i$ is computed from n observations. This can be demonstrated as follows:

$$\begin{aligned}
\mathrm{Var}(\overline{X} - \overline{\overline{X}}) &= \mathrm{Var}\left(\overline{X}_i - \frac{\overline{X}_1 + \cdots + \overline{X}_i + \cdots + \overline{X}_k}{k}\right) \\
&= \mathrm{Var}(\overline{X}_i) - 2\,\mathrm{Cov}\left(\overline{X}_i, \frac{\overline{X}_i}{k}\right) + \mathrm{Var}(\overline{\overline{X}}) \\
&= \frac{\sigma^2}{n} - \frac{2}{k}\left(\frac{\sigma^2}{n}\right) + \frac{\sigma^2}{kn} \\
&= \frac{\sigma^2(k-1)}{kn}
\end{aligned}$$

The result then follows after the square root of the last expression is taken, and s is substituted for σ.

Now assume that the $\overline{X}_i$ are computed from n_i observations, with the n_i not all equal. Then the corresponding derivation is

$$\begin{aligned}
\mathrm{Var}(\overline{X} - \overline{\overline{X}}) &= \mathrm{Var}\left(\overline{X}_i - \frac{n_1\overline{X}_1 + \cdots + n_i\overline{X}_i + \cdots n_k\overline{X}_k}{N}\right) \\
&= \mathrm{Var}(\overline{X}_i) - 2\,\mathrm{Cov}\left(\overline{X}_i, \frac{n_i\overline{X}_i}{N}\right) + \mathrm{Var}(\overline{\overline{X}}) \\
&= \frac{\sigma^2}{n_i} - \frac{2n_i}{N}\left(\frac{\sigma^2}{n_i}\right) + \frac{\sigma^2}{N} \\
&= \frac{\sigma^2}{n_i} - \frac{\sigma^2}{N} \\
&= \sigma^2\left(\frac{N - n_i}{Nn_i}\right)
\end{aligned}$$

REFERENCES

Bakir, S. T. (1989). Analysis of means using ranks. *Communications in Statistics—Simulation and Computation*, **18**(2), 757–776.

Bishop, T. A. and E. J. Dudewicz (1978). Exact analysis of variance with unequal variances: Test procedures and tables. *Technometrics*, **20**, 419–430.

Box, G. E. P., J. S. Hunter, and W. G. Hunter (2005). *Statistics for Experimenters: Design, Innovation, and Discovery.* Hoboken, NJ: Wiley.

Box, G. E. P., W. G. Hunter, and J. S. Hunter (1978). *Statistics for Experimenters.* New York: Wiley.

Czitrom, V. (2003). Statistics in the semiconductor industry. In *Handbook of Statistics*, Vol. 22, pp. 459–498 (R. Khattree and C. R. Rao, eds.). Amsterdam: Elsevier Science B. V.

Dallal, G. E. (2001). Multiple comparison procedures. Online article available at http://www.tufts.edu/∼gdallal/mc.htm.

Dean, A. and D. Voss (1999). *Design and Analysis of Experiments.* New York: Springer-Verlag.

Duncan, D. B. (1955). Multiple range and multiple F tests. *Biometrics*, **11**, 1–42.

Duncan, D. B. (1975). t-Tests and intervals suggested by the data. *Biometrics*, **31**, 739–759.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096–1121.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics*, **20**, 482–491.

Fisher, R. A. (1935). *The Design of Experiments.* London: Oliver & Boyd.

Halperin, M., S. W. Greenhouse, J. Cornfield, and J. Zalokar (1955). Tables of percentage points for the studentized maximum absolute deviate in normal samples. *Journal of the American Statistical Association*, **50**, 185–195 (March).

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Procedures.* New York: Wiley.

Hsu, J. C. (1984). Ranking and selection and multiple comparisons with the best. In *Design of Experiments: Ranking and Selection (Essays in Honor of Robert E. Bechhofer)*, pp. 22–33 (T. J. Santner and A. C. Tamhane, eds.). New York: Marcel Dekker.

Hsu, J. C. (1996). *Multiple Comparisons: Theory and Methods.* New York: Chapman & Hall.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal sample sizes. *Biometrics*, **12**, 307–310.

Nelson, L. S. (1983). Exact critical values for use with the analysis of means. *Journal of Quality Technology*, **15**(1), 40–44.

Nelson, P. R. (1982). Exact critical points for the analysis of means. *Communications in Statistics—Part A, Theory and Methods*, **11**(6), 699–709.

Nelson, P. R. (1993). Additional uses for the analysis of means and extended tables of critical values. *Technometrics*, **35**(1), 61–71.

Nelson, P. R., M. Coffin, and K. A. F. Copeland (2003). *Introductory Statistics for Engineering Experimentation.* San Diego, CA: Academic Press.

Nelson, P. R. and E. J. Dudewicz (2002). Exact analysis of means with unequal variances. *Technometrics*, **44**(2), 152–160.

Nelson, P. R., P. S. Wludyka, and K. A. F. Copeland (2005). *The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions.* Philadelphia: American Statistical Association and Society for Industrial and Applied Mathematics.

Ott, E. R. (1958). Analysis of means. Technical Report #1, Rutgers University.

Ott, E. R. (1967). Analysis of means—a graphical procedure. *Industrial Quality Control*, **24**(2), 101–109.

Ott, E. R. (1975). *Process Quality Control: Troubleshooting and Interpretation of Data.* New York: McGraw-Hill.

Reece, J. E. (2003). Software to support manufacturing systems. In *Handbook of Statistics*, Vol. 22, chap. 9 (R. Khattree and C. R. Rao, eds.). Amsterdam: Elsevier Science B.V.

Reese, A. (2005). Boxplots. *Significance*, **2**(3), 134–135.

Ryan, T. P. (2000). *Statistical Methods for Quality Improvement*, 2nd ed. New York: Wiley.

Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics*, **32**, 461–474.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104.

Schilling, E. G. (1973). A systematic approach to the analysis of means, Part I. Analysis of treatment effects. *Journal of Quality Technology*, **5**(3), 93–108.

Searle, S. R. (1987). *Linear Models for Unbalanced Data.* New York: Wiley.

Tomlinson, L. H. and R. J. Lavigna (1983). Silicon crystal termination—an application of ANOM for percent defective data. *Journal of Quality Technology*, **15**, 26–32.

Tukey, J. W. (1953). The problem of multiple comparisons. Originally unpublished manuscript that appears in *Collected Works of J. W. Tukey*, Vol. VII (H. Braun, ed.). New York: Chapmans & Hall, 1994.

Weerahandi, S. (1994). ANOVA under unequal error variances. *Biometrics*, **51**, 589–599.

Weerahandi, S. (2004). *Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models.* Hoboken, NJ: Wiley.

Wludyka, P. S. and P. R. Nelson (1997). An analysis of means type test for variances from normal populations. *Technometrics*, **39**, 274–285.

EXERCISES

2.1 For the model given in Eq. (2.1), show that the sum of the $\widehat{A}_i$ must be zero for any value of i, assuming n observations for each value of i.

2.2 Derive the equivalence given by Eq. (2.3) in Section 2.1.3.3, using the suggestion that was given below the equation.

2.3 Fill in the blanks in the following output for data from a completely randomized, one-factor design with three levels. Do you need to know whether the factor is fixed or random for completing the table? Why, or why not?

One-way Analysis of Variance

Analysis of Variance for Y

Source	DF	SS	MS	F	P
Factor	—	—	186.3	—	0.005
Error	12	—	22.9		
Total	—	647.4			

2.4 A completely randomized design was used and part of the ANOVA table is as follows.

Source	d.f.
Factor	3
Error	27
Total	30

(a) Explain why there could not have been an equal number of observations per factor level.
(b) Give one possible combination of the number of observations per factor level.

2.5 A study is to be undertaken to compare the coagulation times for samples of blood from 16 animals receiving four different diets, so that a set of four animals receive one diet.
(a) Does a completely randomized design seem appropriate for this experiment or would there likely be any extraneous factors that would have to be accounted for in the design?
(b) Note that each diet average would be computed from four observations. Does this seem adequate? Explain.
(c) Explain in detail how the experiment would be performed if a completely randomized design were used.
(d) What are the assumptions that must be made and explain how they would be tested when the data are analyzed?

2.6 Consider the following data for a completely randomized design with four levels of a fixed factor in a one-factor experiment:

1	2	3	4
17	16	16	19.6
18	20	19	18.6
15	17	14	23.6
19	18	18	17.6
13	14	14	22.6

Assume that you decide to analyze these data using both ANOVA and ANOM (with $\alpha = 0.05$), remembering that it is reasonable to use the two methods together. Are the assumptions that must be made for each the same, or do they differ? Do the assumptions appear to be met or is it even practical to test the assumptions with this amount of data? Explain why the two procedures produce different results. Since the results differ, which result would you go by? Explain.

2.7 In a completely randomized design with unequal group numbers, that is, $n_1 = 5$, $n_2 = 7$, and $n_3 = 6$, what is the degrees of freedom for the error term?

2.8 Construct an example for a single (fixed) factor with three levels and five observations per level for which the overall F-test shows a significant result but the averages for the three levels are 19.2, 19.3, and 19.5, respectively, and all 15 numbers are different.

2.9 The following data are available for a completely randomized design: $T_1 = 20$, $T_2 = 30$, and $T_3 = 40$, with T_i denoting the total of the observations for treatment i. In like manner, $n_1 = n_2 = n_3 = 5$. If the F-statistic for testing the equality of the three treatment means equals 4,

(a) What does SS_{total} equal?

(b) Would the null hypothesis be rejected for $\alpha = .05$?

2.10 Assuming the following data have come from a completely randomized design, compute the treatment sum of squares.

Treatment

1	2	3
4	6	1
3	5	4
5	5	4
4	4	4
4		2
		5

2.11 One of the sample datasets that comes with MINITAB is RADON.MTW. The dataset consists of 80 measurements of radiation in an experimental chamber. There were four different devices: filters, membranes, open cups, and badges and 20 devices of each type were used. The response variable is the amount of radiation that each device measured and the objective is to determine if there is any difference between the devices relative to the response variable. Can these data be analyzed using one or more of the methods given in this chapter? Why, or why not?

2.12 Consider the second example in Section 2.1.3. The conclusion was that the means do not differ, despite the fact that the sample means differ by far more than in the first example in that section. Apply the sliding t-distribution approach to that example. Do you also conclude that the population means do not differ? Explain.

2.13 Assume that a one-factor design is to be used and there are two levels of the factor.

(a) What is the smallest possible value of the F-statistic and when will that occur? What is the largest possible value?
(b) Construct an example with six observations at each level that will produce this minimum value.

2.14 Consider Exercises 2.3 and 2.9. Could an ANOM display be constructed for the data in the first exercise after the blanks have been filled in? Why, or why not? Could an ANOM display be constructed using the data summary in Exercise 2.9? Why, or why not? If either or both of the ANOM displays can be constructed from the information given, construct the display(s).

2.15 It was stressed in Section 1.7 that, ideally, processes should be in a state of statistical control when statistically designed experiments are performed. Assume that a manufacturing process was improved slightly with an eye toward improving the process. The change was made near the middle of the process and management wants to see if there has been any improvement. The product must go through one of two (supposedly identical) machines near the end of the process. An experiment is performed using both the standard process and the improved process, with product from each assigned to one of the two machines in a semirandom fashion such that each machine handles the same number of production units. Now assume that the older of the two machines malfunctions in such a way that is not obvious but does affect a key measurement characteristic.

(a) If the malfunction results in a higher-than-normal reading for the measurement characteristic, will this likely affect the conclusions that are drawn from the experiment if the reading is inflated by 20% and the variance is inflated by 5%? Explain.
(b) Would your answer be different if each machine received output from only one of the two processes? Explain.
(c) Could the semirandomization be improved so that true randomization is employed? If so, how?

2.16 (Harder problem) The raw data from a study are freqently summarized after collection with the consequence that data in terms of means and variances may be all that is available to an analyst. Assume that there is interest in comparing different types of paints in terms of drying times, with three types to be compared. Four walls in a building are painted with each type or paint, with the walls considered to be essentially the same. The average drying times in hours and the variances of the drying times are given below.

	Paint Type		
	1	2	3
Average	7.23	8.44	8.67
Variance	2.34	1.97	2.21

Perform an Analysis of Variance. What do you conclude and what would you recommend?

2.17 Assume that there is an equal number of observations, n, per level. Show that for k levels the second term on the right side of Eq. (2.3) is equal to the sum of the variances of the levels, multiplied by n. Then show that MS_{error} is equal to the average of the variances.