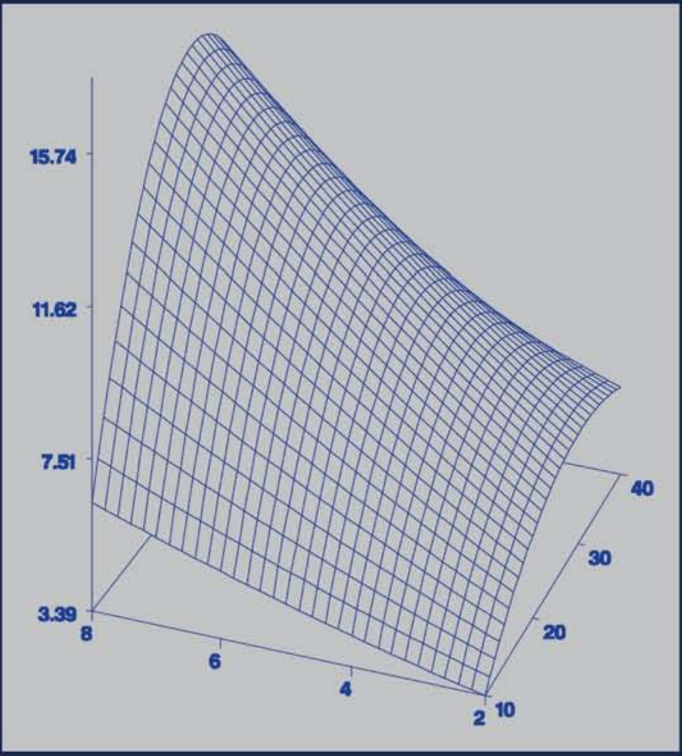
SECOND EDITION

ANALYSIS OF MESSY DATA

VOLUME 1

DESIGNED EXPERIMENTS



George A. Milliken
Dallas E. Johnson

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

ANALYSIS OF MESSY DATA

VOLUME 1

DESIGNED EXPERIMENTS

SECOND EDITION

ANALYSIS OF MESSY DATA

VOLUME 1

DESIGNED EXPERIMENTS

SECOND EDITION

George A. Milliken
Dallas E. Johnson

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130920

International Standard Book Number-13: 978-1-4200-1015-2 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

1

The Simplest Case: One-Way Treatment Structure in a Completely Randomized Design Structure with Homogeneous Errors

Suppose an experimenter wants to compare the effects of several different treatments, such as the effects of different drugs on people's heart rates or the yields of several different varieties of wheat. Often the first step in analyzing the data from such experiments is to use a statistical method, known as a one-way analysis of variance model, to describe the data. The model on which the one-way analysis of variance is based is one of the most useful models in the field of statistics. Many experimental situations are simply special cases of this model. Other models that appear to be much more complicated can often be considered as one-way models. This chapter is divided into several sections. In the first two sections, the one-way model is defined and the estimation of its parameters is discussed. In Sections 1.3 and 1.5, inference procedures for specified linear combinations of the treatment effects are provided. In Sections 1.7 and 1.9, we introduce two basic methods for developing test statistics. These two methods are used extensively throughout the remainder of the book. Finally, in Section 1.11, we discuss readily available computer analyses that use the above techniques. An example is used to demonstrate the concepts and computations described in each section.

1.1 Model Definitions and Assumptions

Assume that a sample of N experimental units is selected completely at random from a population of possible experimental units. An experimental unit is defined as the basic unit to which a treatment will be applied and independently observed. A more complete description of experimental units can be found in Chapters 4 and 5.

In order to compare the effects of t different treatments, the sample of N experimental units is randomly divided into t groups so that there are n_i experimental units in the ith group, where $i = 1, 2, \ldots, t$, and $N = \sum_{i=1}^{t} n_i$. Grouping the experimental units at random into t groups should remove any systematic biases. That is, randomness should ensure that the t groups of experimental units are similar in nature before the treatments are applied. Finally, one of the t treatments should be randomly assigned to each group of experimental units. Equivalently, the experimental units could be randomly assigned to the t treatment groups using some randomization device such as placing n_1 tags in a bowl with treatment 1, n_2 tags in a bowl with treatment $2, \ldots, n_t$ tags in a bowl with treatment t, mixing the tags and then randomly selecting tags from the bowl to determine the treatment assigned to each experimental unit. This process of using tags in a bowl can obviously be carried out using software that has random number generation possibilities.

Let y_{ij} denote a response from the jth experimental unit assigned to the ith treatment. The values $y_{11}, y_{12}, \ldots, y_{1n_1}$ can be thought of as being a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2, the values $y_{21}, y_{22}, \ldots, y_{2n_2}$ can be thought of as being a random sample of size n_2 from a population with mean μ_2 and variance σ_2^2, and similarly for $i = 3, 4, \ldots, t$. The parameters μ_i and σ_i^2 represent the population mean and population variance if one applied treatment i to the whole population of experimental units.

The simplest case is considered in this chapter in that the variances are assumed to be homogeneous or equal across treatments or $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$. That is, it is assumed that the application of the ith treatment to the experimental units may affect the mean of the responses but not the variance of the responses. The equal variance assumption is discussed in Chapter 2 as well as the analysis of variance with unequal variances.

The basic objectives of a good statistical analysis are to estimate the parameters of the model and to make inferences about them. The methods of inference usually include testing hypotheses and constructing confidence intervals.

There are several ways to write a model for data from situations like the one described above. The first model to be used is called the μ_i model or the means model. The means model is:

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, 2, \ldots, t, \; j = 1, 2, \ldots, n_i$$

where it is assumed that

$$\varepsilon_{ij} \sim i.i.d.\ N(0, \sigma^2) \quad i = 1, 2, \ldots, t, \; j = 1, 2, \ldots, n_i \tag{1.1}$$

The notation $\varepsilon_{ij} \sim i.i.d.\ N(0, \sigma^2)$ is used extensively throughout this book. It means that the ε_{ij} $(i = 1, 2, \ldots, t;\ j = 1, 2, \ldots, n_i)$ are independently and identically distributed and that the sampling distribution of each ε_{ij} is the normal distribution with mean equal to zero and variance equal to σ^2.

1.2 Parameter Estimation

The most important aspect of a statistical analysis is to get a good estimate of the error variance per experimental unit, namely σ^2. The error variance measures the accuracy of an experiment—the smaller the σ^2, the more accurate the experiment. One cannot make any statistically valid inferences in any experiment or study without some knowledge of the experimental error variance.

In the above situation, the ith sample, $i = 1, 2, \dots, t$, provides an estimate of σ^2 when $n_i > 1$. The estimate of σ^2 obtained from the data from the ith treatment is

$$\hat{\sigma}_i^2 = \sum_{j=1}^{n_j} \frac{(y_{ij} - \bar{y}_{i\cdot})^2}{n_i - 1}$$

which is an unbiased estimate of σ^2 where

$$\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

The estimate of σ^2 from the ith treatment is $\hat{\sigma}_i^2$, which is based on $n_i - 1$ degrees of freedom, and the sampling distribution of $(n_i - 1)\hat{\sigma}_i^2/\sigma^2$ is a chi-square distribution with $n_i - 1$ degrees of freedom.

A weighted average of these t independent estimates of σ^2 provides the best estimate for σ^2 possible for this situation, where each estimate of the variance is weighted by its corresponding degrees of freedom. The best estimate of σ^2 is

$$\hat{\sigma}^2 = \sum_{i=1}^{t} (n_i - 1)\hat{\sigma}_i^2 \Big/ \sum_{i=1}^{t} (n_i - 1)$$

For computational purposes, each variance times its weight can be expressed as

$$(n_i - 1)\hat{\sigma}_i^2 = \sum_{i=1}^{t} (y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{t} y_{ij}^2 - n_i \bar{y}_{i\cdot}^2 = \sum_{i=1}^{t} y_{ij}^2 - (y_{i\cdot})^2/n_i = SS_i$$

where $y_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}$. Then the pooled estimate of the variance is

$$\hat{\sigma}^2 = \frac{SS_1 + SS_2 + \cdots + SS_t}{(n_1 - 1) + (n_2 - 1) + \cdots + (n_t - 1)} = \frac{\sum_{i=1}^{t} SS_i}{N - t}$$

The pooled estimate of the variance $\hat{\sigma}^2$ is based on $N - t$ degrees of freedom and the sampling distribution of $(N-t)\hat{\sigma}^2/\sigma^2$ is a chi-square distribution with $N - t$ degrees of freedom; that is, $(N-t)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-t}$.

The best estimate of each μ_i is $\hat{\mu}_i = \bar{y}_{i\cdot}$, $i = 1, 2, \dots, t$.

Under the assumption given in Equation 1.1, the sampling distribution of $\hat{\mu}_i$ is normal with mean μ_i and variance σ^2/n_i. That is,

$$\hat{\mu}_i \sim N\left(\mu_i, \frac{\sigma^2}{n_i}\right) \quad i = 1, 2, \dots, t \tag{1.2}$$

Using the sampling distributions of $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ then

$$t_i = \frac{\hat{\mu}_i - \mu_i}{\sqrt{\hat{\sigma}^2/n_i}} \sim t_{N-t} \quad i = 1, 2, \ldots, t \tag{1.3}$$

That is, the sampling distribution of t_i is the t-distribution with $N - t$ degrees of freedom. In addition, $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_t$ and σ_i^2 are statistically independent.

1.3 Inferences on Linear Combinations—Tests and Confidence Intervals

This section provides tests of hypotheses and confidence intervals for linear functions of the parameters in the means model. The results in the previous section can be used to test hypotheses about the individual μ_i. Those results can also be used to test hypotheses about linear combinations of the μ_i or to construct confidence intervals for linear combinations of the μ_i.

For an experiment involving several treatments, the investigator selects the treatments to be in the study because there are interesting hypotheses that need to be studied. These interesting hypotheses form the objectives of the study. The hypotheses involving the treatment means most likely will involve specific linear combinations of the means. These linear combinations will enable the investigator to compare the effects of the different treatments or, equivalently, the means of the different treatments or populations. The hypotheses about the means the experimenter has selected can be of the following types of hypotheses:

$$H_{01}\text{: } \sum_{i=1}^{t} c_i\mu_i = a \text{ vs } H_{a1}\text{: (not } H_{01}\text{:)}$$

for some set of known constants $c_1, c_2, \ldots, c_t$ and a,

$$H_{02}\text{: } \mu_1 = \mu_2 = \cdots = \mu_t \text{ vs } H_{a2}\text{: (not } H_{02}\text{:)}$$

and

$$H_{03}\text{: } \mu_i = \mu_{i'} \text{ for some } i \neq i' \text{ vs } H_{a3}\text{: (not } H_{03}\text{:)}$$

For a linear combination such as that given in H_{01}, one can show that

$$\frac{\sum_{i=1}^{t} c_i\hat{\mu}_i - \sum_{i=1}^{t} c_i\mu_i}{\sqrt{\hat{\sigma}^2 \sum_{i=1}^{t} c_i^2/n_i}} \sim t_{(N-t)} \tag{1.4}$$

This result can be used to make inferences about linear combinations of the form $\sum_{i=1}^{t} c_i\mu_i$. Since the hypothesis in H_{03} can be written as H_{03}: $\mu_i - \mu_{i'} = 0$, it is a special case of H_{01} with

$c_i = 1$, $c_{i'} = -1$, and $c_k = 0$ if $k \neq i$ or i'. A test for H_{02} is given in Section 1.5. The estimated standard error of $\sum_{i=1}^{t} c_i \hat{\mu}_i$ is given by

$$\widehat{s.e.}\left(\sum c_i \hat{\mu}_i\right) = \sqrt{\hat{\sigma}^2 \sum \frac{c_i^2}{n_i}} \tag{1.5}$$

To test H_{01}: $\sum_{i=1}^{t} c_i \; \mu_i = a$ vs H_{a1}: (not H_{01}:) compute the t-statistic

$$t_c = \frac{\sum c_i \hat{\mu}_i - a}{\widehat{s.e.}\left(\sum c_i \hat{\mu}_i\right)} \tag{1.6}$$

If $|t_c| > t_{\alpha/2,\nu}$, where $\nu = N - t$, then H_{01} is rejected at the $\alpha = 100\%$ significance level, where $t_{\alpha/2,\nu}$ is the upper $\alpha/2$ critical point of a t-distribution with ν degrees of freedom. A $(1 - \alpha)$ 100% confidence interval for $\sum_{i=1}^{t} c_i \mu_i$ is provided by

$$\sum c_i \hat{\mu}_i \pm t_{\alpha/2,\nu} \; \widehat{s.e.}\left(\sum c_i \hat{\mu}_i\right) \tag{1.7}$$

1.4 Example—Tasks and Pulse Rate

The data in Table 1.1 came from an experiment that was conducted to determine how six different kinds of work tasks affect a worker's pulse rate. In this experiment, 78 male workers were assigned at random to six different groups so that there were 13 workers in each group. Each group of workers was trained to perform their assigned task. On a selected day after training, the pulse rates of the workers were measured after they had performed their assigned tasks for 1 h. Unfortunately, some individuals withdrew from the experiment during the training process so that some groups contained fewer than 13 individuals. The recorded data represent the number of heart pulsations in 20 s where there are $N = 68$ observations and the total is $y = 2197$.

For the tasks data, the best estimate of σ^2 is

$$\hat{\sigma}^2 = \sum_{i=1}^{6} SS_i \Big/ (N - t) = 1{,}916.0761/62 = 30.9045$$

which is based on 62 degrees of freedom. The best estimates of the μ_i are $\hat{\mu}_1 = 31.923$, $\hat{\mu}_2 = 31.083$, $\hat{\mu}_3 = 35.800$, $\hat{\mu}_4 = 38.000$, $\hat{\mu}_5 = 29.500$, and $\hat{\mu}_6 = 28.818$.

For illustration purposes, suppose the researcher is interested in answering the following questions about linear combinations of the task means:

a) Test H_0: $\mu_3 = 30$ vs H_a: $\mu_3 \neq 30$.

b) Find a 95% confidence interval for μ_1.

TABLE 1.1

Pulsation Data and Summary Information for Six Tasks

	Task					
	1	**2**	**3**	**4**	**5**	**6**
	27	29	34	34	28	28
	31	28	36	34	28	26
	26	37	34	43	26	29
	32	24	41	44	35	25
	39	35	30	40	31	35
	37	40	44	47	30	34
	38	40	44	34	34	37
	39	31	32	31	34	28
	30	30	32	45	26	21
	28	25	31	28	20	28
	27	29			41	26
	27	25			21	
	34					
$y_{i.}$	415	373	358	380	354	317
n_i	13	12	10	10	12	11
$\bar{y}_{i.}$	31.9231	31.0833	35.8000	38.0000	29.5000	28.8182
SS_i	294.9231	352.9167	253.6000	392.0000	397.0000	225.6364

c) Test H_0: $\mu_4 = \mu_5$ vs H_a: $\mu_4 \neq \mu_5$.

d) Test H_0: $\mu_1 = (\mu_2 + \mu_3 + \mu_4)/3$ vs H_a: $\mu_1 \neq (\mu_2 + \mu_3 + \mu_4)/3$.

e) Obtain a 90% confidence interval for $4\mu_1 - \mu_3 - \mu_4 - \mu_5 - \mu_6$.

These questions can be answered by applying the results of this section.

Part a result: A t-statistic for testing H_0: $\mu_3 = 30$ is obtained by substituting into Equation 1.6 to obtain

$$t_c = \frac{\hat{\mu}_3 - 30}{\widehat{s.e.}(\hat{\mu}_3)} = \frac{\hat{\mu}_3 - 30}{\sqrt{\hat{\sigma}^2/n_3}} = \frac{35.8 - 30.0}{\sqrt{30.9045/10}} = 3.30$$

The significance probability of this calculated value of t is $\hat{\alpha} = \Pr\{|t_c| > 3.30\} = 0.0016$ where $\Pr\{|t_c| > 3.30\}$ is the area to the right of 3.30 plus the area to the left of –3.30 in a t-distribution with 62 degrees of freedom. The above value of $\hat{\alpha}$ was obtained from computer output, but it can also be obtained from some special hand-held calculators. Readers of this book who lack access to a computer or a calculator should compare $t_c = 3.30$ to $t_{\alpha/2,62}$ for their choice of α.

Part b result: A 95% confidence interval for μ_1 is given by

$$\begin{aligned} \hat{\mu}_1 \pm t_{0.025,62}\, \widehat{s.e.}(\hat{\mu}_1) &= 31.923 \pm 2.00\sqrt{30.9045/13} \\ &= 31.923 \pm 2.00 \times 1.542 \end{aligned}$$

Thus the 95% confidence interval about μ_1 is $28.839 < \mu_1 < 35.007$ and we are 95% confident that this interval contains the true, but unknown value of μ_1.

Part c result: To test H_0: $\mu_4 = \mu_5$, let $l_1 = \mu_4 - \mu_5$, then $\hat{l}_1 = \hat{\mu}_4 - \hat{\mu}_5 = 38.0 - 29.5 = 8.5$ and

$$\widehat{s.e.}(\hat{l}_1) = \sqrt{\hat{\sigma}^2 \sum_{i=1}^{6} c_i^2/n_i} = \sqrt{30.9045\left(\frac{1}{10} + \frac{1}{12}\right)} = 2.380$$

since $c_1 = c_2 = c_3 = c_6 = 0$, $c_4 = 1$, and $c_5 = -1$.

The t-statistic for testing H_0: $\mu_4 = \mu_5$ is

$$t_c = \frac{8.5}{2.380} = 3.57$$

The significance probability for this test is $\hat{\alpha} = 0.0007$.

Part d result: A test of H_0: $\mu_1 = (\mu_2 + \mu_3 + \mu_4)/3$ is equivalent to testing H_0: $\mu_1 - \frac{1}{3}\mu_2 - \frac{1}{3}\mu_3 - \frac{1}{3}\mu_4 = 0$ or testing H_0: $3\mu_1 - \mu_2 - \mu_3 - \mu_4 = 0$. By choosing the last version, the computations are somewhat easier and the value of the t_c test statistic is invariant with respect to a constant multiplier.

Let $l_2 = 3\mu_1 - \mu_2 - \mu_3 - \mu_4$, then

$$\hat{l}_2 = 3\hat{\mu}_1 - \hat{\mu}_2 - \hat{\mu}_3 - \hat{\mu}_4 = 3(31.923) - 31.083 - 35.8 - 38.0 = -9.114$$

The estimate of the standard error of $\hat{l}_2$ is

$$\widehat{s.e.}(\hat{l}_2) = \sqrt{30.9045\left(\frac{9}{13} + \frac{1}{12} + \frac{1}{10} + \frac{1}{10}\right)} = 5.491$$

A t-statistic for testing H_0: $3\mu_1 - \mu_2 - \mu_3 - \mu_4 = 0$ is

$$t_c = \frac{-9.114}{5.491} = -1.66$$

The significance probability corresponding to t_c is $\hat{\alpha} = 0.1020$.

Part e result: Let $l_3 = 4\mu_1 - \mu_3 - \mu_4 - \mu_5 - \mu_6$. Then $\hat{l}_3 = -4.426$ and $\widehat{s.e.}(\hat{l}_3) = 7.0429$. A 90% confidence interval for l_3 is

$$\hat{l}_3 \pm t_{0.05,62}\, \widehat{s.e.}(\hat{l}_3) = -4.426 \pm 1.671 \times 7.043 = -4.426 \pm 11.769$$

Thus, a 90% confidence interval is $-16.195 < 4\mu_1 - \mu_3 - \mu_4 - \mu_5 - \mu_6 < 7.343$.

1.5 Simultaneous Tests on Several Linear Combinations

For many situations the researcher wants to test a simultaneous hypothesis about several linear combinations of the treatment's effects or means. For example, the general hypothesis involving k linearly independent linear combinations of the treatment means can be expressed as

$$H_0: \begin{array}{l} c_{11}\mu_1 + c_{12}\mu_2 + \cdots + c_{1t}\mu_t = a_1 \\ c_{21}\mu_1 + c_{22}\mu_2 + \cdots + c_{2t}\mu_t = a_2 \\ \vdots \\ c_{k1}\mu_1 + c_{k2}\mu_2 + \cdots + c_{kt}\mu_t = a_k \end{array} \quad \text{vs} \quad H_a\text{: (not } H_0) \tag{1.8}$$

The results presented in this section are illustrated using vectors and matrices. However, knowledge of vectors and matrices is not really necessary for readers having access a computer with matrix manipulation software, since most computers allow even novice users to easily carry out matrix computations.

The hypothesis in Equation 1.8 can be written in matrix notation as

$$H_0\text{: } \boldsymbol{C\mu} = \boldsymbol{a} \text{ vs } H_a\text{: } \boldsymbol{C\mu} \neq \boldsymbol{a} \tag{1.9}$$

where

$$\boldsymbol{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1t} \\ c_{21} & c_{22} & \cdots & c_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1} & c_{k2} & \cdots & c_{kt} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_t \end{bmatrix}, \quad \text{and} \quad \boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix} \tag{1.10}$$

It is assumed that the k rows in $\boldsymbol{C}$ were chosen such that they are linearly independent, which means that none of the rows in $\boldsymbol{C}$ can be expressed as a linear combination of the remaining rows. If the k rows in $\boldsymbol{C}$ are not linearly independent, a subset of the rows that are linearly independent can always be selected so that they contain all the necessary information about the required hypothesis.

For example, suppose you have three treatments and you wish to test

$$H_0\text{: } \mu_1 - \mu_2 = 0,\ \mu_1 - \mu_3 = 0 \quad \text{and} \quad \mu_2 - \mu_3 = 0$$

the corresponding $\boldsymbol{C}$ matrix is

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

but the third row of $\boldsymbol{C}$ is the difference between the second row and the first row, hence the three rows are not linearly independent. In this case, an equivalent hypothesis can be stated as H_0: $\mu_1 - \mu_2 = 0$ and $\mu_1 - \mu_3 = 0$, since if $\mu_1 - \mu_2 = 0$ and $\mu_1 - \mu_3 = 0$, then $\mu_2 - \mu_3$ must be equal to 0. The following discussion uses the assumption that the rows of $\boldsymbol{C}$ are linearly independent.

Denote the vector of sample means by $\hat{\boldsymbol{\mu}}$, then the sampling distribution of $\hat{\boldsymbol{\mu}}$ in matrix notation is

$$\hat{\boldsymbol{\mu}} \sim N_t(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\boldsymbol{D}) \quad \text{where } \boldsymbol{D} = \begin{bmatrix} 1/n_1 & 0 & \cdots & 0 \\ 0 & 1/n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/n_t \end{bmatrix}$$

This equation is read as follows: The elements of the $t \times 1$ vector $\hat{\boldsymbol{\mu}}$ have a joint sampling distribution that is the t-variate normal distribution with means given by the vector $\boldsymbol{\mu}$ and with variances and covariances given by the elements in the matrix $\boldsymbol{\sigma}^2\boldsymbol{D}$. The ith diagonal element of $\boldsymbol{\sigma}^2\boldsymbol{D}$ is the variance of $\hat{\mu}_i$ and the (i, j)th $i \neq j$ off-diagonal element gives the covariance between $\hat{\mu}_i$ and $\hat{\mu}_j$.

The sampling distribution of $\boldsymbol{C}\hat{\boldsymbol{\mu}}$ is

$$\boldsymbol{C}\hat{\boldsymbol{\mu}} \sim N_k(\boldsymbol{C\mu}, \boldsymbol{\sigma}^2\boldsymbol{CDC}')$$

The sum of squares due to deviations from H_0 or the sum of squares for testing H_0: $\boldsymbol{C\mu} = \boldsymbol{a}$ is given by

$$SS_{H0} = (\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a})'(\boldsymbol{CDC}')^{-1}(\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a}) \tag{1.11}$$

and is based on k degrees of freedom, the number of linearly independent rows of $\boldsymbol{C}$. Using the assumption of normality, the sampling distribution of SS_{H0}/σ^2 is that of a noncentral chi-square with k degrees of freedom. If H_0 is true, then $SS_{H0}/\sigma^2 \sim \chi^2_k$. The statistic for testing H_0 is

$$F_c = \frac{SS_{H0}/k}{\hat{\sigma}^2}$$

The hypothesis H_0: $\boldsymbol{C\mu} = \boldsymbol{a}$ is rejected at the significance level of α if $F_c > F_{\alpha,k,N-t}$ where $F_{\alpha,k,N-t}$ is the upper α critical point of the F-distribution with k numerator degrees of freedom and $N - t$ denominator degrees of freedom. The result given here is a special case of Theorem 6.3.1 in Graybill (1976).

When H_0 is true, then SS_{H0}/k is an unbiased estimate of σ^2, which is then compared with $\hat{\sigma}^2$, which in turn is an unbiased estimate of σ^2 regardless of whether H_0 is true or not. Thus the F-statistic given above should be close to 1 if H_0 is true. If H_0 is false, the statistic SS_{H0}/k is an unbiased estimate of

$$\sigma^2 + \frac{1}{k}(\boldsymbol{C\mu} - \boldsymbol{a})'(\boldsymbol{CDC}')^{-1}(\boldsymbol{C\mu} - \boldsymbol{a})$$

Thus, if H_0 is false, the value of the F-statistic should be larger than 1. The hypothesis H_0 is rejected if the calculated F-statistic is significantly larger than 1.

1.6 Example—Tasks and Pulse Rate (Continued)

The following is a summary of the information from the example in Section 1.4 with the sample size and mean for each of the six tasks.

Task *i*	**1**	**2**	**3**	**4**	**5**	**6**
n_i	13	12	10	10	12	11
$\bar{y}_{i\cdot}$	31.9231	31.0833	35.8000	38.0000	29.5000	28.8182

The pooled estimate of the variance is $\hat{\sigma}^2 = 30.9045$ and it is based on 62 degrees of freedom. The ***D*** matrix associated with the sampling distribution of vector of estimated means is

$$\boldsymbol{D} = \begin{bmatrix} \frac{1}{13} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{10} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{12} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{11} \end{bmatrix}$$

Suppose the researcher is interested in simultaneously testing the following hypothesis involving two linear combinations of the task means:

$$H_0\colon \mu_4 - \mu_5 = 4 \text{ and } 3\mu_1 - \mu_2 - \mu_3 - \mu_4 = 0 \text{ vs } H_a\colon (\text{not } H_0)$$

The ***C*** matrix consists of two rows, one for each of the linear combinations in H_0, and the vector ***a*** has two elements as

$$\boldsymbol{C} = \begin{bmatrix} 0 & 0 & 0 & 1 & -1 & 0 \\ 3 & -1 & -1 & -1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{a} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

Preliminary computations needed to provide the value of SS_{H0} are:

$$\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a} = \begin{bmatrix} 8.5 - 4 \\ -9.114 - 0 \end{bmatrix} = \begin{bmatrix} 4.500 \\ -9.114 \end{bmatrix}$$

$$\boldsymbol{CDC}' = \begin{bmatrix} \frac{1}{10} + \frac{1}{12} & -\frac{1}{10} \\ -\frac{1}{10} & \frac{9}{13} + \frac{1}{12} + \frac{1}{10} + \frac{1}{10} \end{bmatrix} = \begin{bmatrix} 0.1833 & -0.1000 \\ -0.1000 & 0.9756 \end{bmatrix}$$

$$(\boldsymbol{CDC}')^{-1} = \begin{bmatrix} 5.7776 & 0.5922 \\ 0.5922 & 1.0856 \end{bmatrix}$$

and

$$SS_{H0} = (\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a})'(\boldsymbol{CDC}')^{-1}\,(\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a}) = 158.602$$

with 2 degrees of freedom. The test statistic is

$$F_c = \frac{158.602/2}{30.9045} = 2.566$$

The significance probability of this F-statistic is $\hat{\alpha} = \Pr\{F > 2.566\} = 0.0850$.

1.7 Testing the Equality of All Means

Often the first hypothesis of interest to most researchers is to test that the means are simultaneously equal. The hypothesis is H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$ vs H_a: (not H_0). Two basic procedures are examined for testing the equal means hypothesis. For the particular situation discussed in this chapter, the two procedures give rise to the same statistical test. However, for most messy data situations (for treatment structures other than one-way), the two procedures can give rise to different tests. The first procedure is covered in this section, while the second is introduced in Section 1.9.

The equal means hypothesis, H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$ is equivalent to a hypothesis of the form, H_0: $\mu_1 - \mu_2 = 0$, $\mu_1 - \mu_3 = 0, \ldots, \mu_1 - \mu_t = 0$, or any other hypothesis that involves $t - 1$ linearly independent linear combinations of the μ_i. The $\boldsymbol{C}$ matrix and $\boldsymbol{a}$ vector corresponding to the set of $t - 1$ pairwise differences are:

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & -1 & \cdots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & -1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The $\boldsymbol{C}$ matrix corresponding to following set of $t - 1$ linearly independent linear combinations of the μ_i; H_0: $\mu_1 - \mu_2 = 0$, $\mu_1 + \mu_2 - 2\mu_3 = 0$, $\mu_1 + \mu_2 + \mu_3 - 3\mu_4 = 0, \ldots, \mu_1 + \mu_2 + \cdots - (t - 1)\,\mu_t = 0$ is:

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & -2 & 0 & \cdots & 0 \\ 1 & 1 & 1 & -3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & 1 & 1 & 1 & \cdots & t-1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Many other matrices exist, so that $\boldsymbol{C\mu} = \mathbf{0}$ if and only if $\mu_1 = \mu_2 = \cdots = \mu_t$; however, all such matrices produce the same sum of squares for deviations from H_0 and the same degrees of freedom, $t - 1$, and hence the same F-statistic. For this special case Equation 1.11 always reduces to

$$SS_{H0:\mu_1=\mu_2=\cdots=\mu_t} = \sum_{i=1}^{t} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{t}\left(\frac{y_{i\cdot}^2}{n_i}\right) - \frac{y_{\cdot\cdot}^2}{N} \tag{1.12}$$

1.8 Example—Tasks and Pulse Rate (Continued)

For the task and pulse rate data in Section 1.4, the $SS_{H0:\mu_1=\mu_2=\cdots=\mu_t}$ is computed using Equations 1.11 and 1.12.

Using the formula in Equation 1.12, provides

$$\begin{aligned} SS_{H0} &= \frac{415^2}{13} + \frac{373^2}{12} + \frac{358^2}{10} + \frac{380^2}{10} + \frac{354^2}{12} + \frac{317^2}{11} - \frac{2197^2}{68} \\ &= 694.4386 \end{aligned}$$

with $t - 1 = 5$ degrees of freedom. The value of the F_c statistic is

$$F_c = \frac{694.4386/5}{30.9045} = 4.49$$

and the significance probability is $\hat{\alpha} = 0.0015$.

Next, using Equation 1.11, the matrix $\boldsymbol{C}$, vector $\boldsymbol{a}$, and matrix $\boldsymbol{D}$ are

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \quad \boldsymbol{a} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

and

$$\boldsymbol{D} = \begin{bmatrix} 0 & \begin{matrix} \frac{1}{13} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{10} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{10} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{12} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{11} \end{matrix} \end{bmatrix}$$

Next compute the individual quantities in Equation 1.11 as

$$\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a} = \begin{bmatrix} 0.844 \\ -3.877 \\ -6.077 \\ 2.423 \\ 3.105 \end{bmatrix} \quad \text{and} \quad \boldsymbol{CDC}' = \begin{bmatrix} \frac{25}{156} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} \\ \frac{1}{13} & \frac{23}{130} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} \\ \frac{1}{13} & \frac{1}{13} & \frac{23}{130} & \frac{1}{13} & \frac{1}{13} \\ \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{25}{156} & \frac{1}{13} \\ \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{1}{13} & \frac{24}{143} \end{bmatrix}$$

The inverse of **CDC′** is

$$(\boldsymbol{CDC}')^{-1} = \begin{bmatrix} 9.882 & -1.765 & -1.765 & -2.118 & -1.941 \\ -1.765 & 8.529 & -1.471 & -1.765 & -1.618 \\ -1.765 & -1.471 & 8.529 & -1.765 & -1.618 \\ -2.118 & -1.765 & -1.765 & 9.882 & -1.941 \\ -1.941 & -1.618 & -1.618 & -1.941 & 9.221 \end{bmatrix}$$

Finally, the value of the sum of squares is

$$SS_{H0} = (\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a})' \, (\boldsymbol{CDC}')^{-1} \, (\boldsymbol{C}\hat{\boldsymbol{\mu}} - \boldsymbol{a}) = 694.4386$$

which is the same as the sum of squares computed using Equation 1.12.

Clearly, this formula is not easy to use if one must do the calculations by hand. However, in many messy data situations, formulas such as this one are necessary in order to obtain the statistic to test meaningful hypotheses. Fortunately, by utilizing computers, **C** matrices can be constructed for a specific hypothesis and then one can allow the computer to do the tedious calculations.

1.9 General Method for Comparing Two Models—The Principle of Conditional Error

A second procedure for computing a test statistic compares the fit of two models. In this section, the two models compared are $y_{ij} = \mu_i + \varepsilon_{ij}$, which is the general or unreduced model, and $y_{ij} = \mu + \varepsilon_{ij}$, which is the model one would have if H_0: $\mu_1 = \mu_2 = \cdots = \mu_t = \mu$ (say) were true. The first model is called the full model or the unrestricted model, while the second model is called the reduced model or the restricted model.

The principle known as the principle of conditional error is used to compare two models where one model is obtained by placing restrictions upon the parameters of another model. The principle is very simple, requiring that one obtain the residual or error sums of squares for both the full model and the reduced model. Let ESS_F denote the error sum of squares after fitting the full model and ESS_R denote the error sum of squares after fitting the reduced model. Then the sum of squares due to the restrictions given by the hypothesis or deviations from the null hypothesis is $SS_{H0} = ESS_R - ESS_F$. The degrees of freedom for both ESS_R and ESS_F are given by the difference between the total number of observations in the data set and the number of (essential) parameters to be estimated (essential parameters will be discussed in Chapter 6). Denote the degrees of freedom corresponding to ESS_R and ESS_F by df_R and df_F, respectively. The number of degrees of freedom corresponding to SS_{H0} is $df_{H0} = df_R - df_F$. An F-statistic for testing H_0 is given by

$$F_c = \frac{SS_{H0}/df_{H0}}{ESS_f/df_F}$$

One rejects H_0 at the significance level if $F_c > F_{\alpha,df_{H0},df_F}$.

For the case discussed above, $y_{ij} = \mu_i + \varepsilon_{ij}$ is the full model and $y_{ij} = \mu + \varepsilon_{ij}$ is the reduced model. The error sum of squares for the full model is

$$ESS_F = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = (N-t)\hat{\sigma}^2$$

with $df_F = N - t$, and the error sum of squares for the reduced model is

$$ESS_R = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$$

with $df_R = N - 1$. Thus the sum of squares due to deviations from H_0 is

$$SS_{H0:\mu_1=\mu_2=\cdots=\mu_t} = ESS_R - ESS_F = \sum_{i=1}^{t} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

with $t - 1$ degrees of freedom. This is the same sum of squares as was obtained in Equation 1.12.

The sums of squares that are of interest in testing situations are often put in a table called an analysis of variance table. Such a table often has a form similar to that in Table 1.2. The entries under the column "Source of variation" are grouped into sets. In a given situation only one of the labels in each set is used, with the choice being determined entirely by the experimenter.

TABLE 1.2

Analysis of Variance Table for One-Way Model to Test Equality of the Means

Source of Variation	*df*	SS	MS	*F*-test
$H_0\,\mu_1 = \mu_2 = \cdots \mu_1$ *Treatments* *between samples*	$t - 1$	SS_{H0}	$\frac{SS_{H0}}{t-1}$	$\frac{SS_{H0}/t-1}{\hat{\sigma}^2}$
Error *within samples*	$N - t$	SS_F	$\hat{\sigma}^2 = \frac{ESS_F}{N-t}$	

Note: *df* = degrees of freedom, SS = sum of square, and MS = mean square. These standard abbreviations are used throughout the book.

The principle of conditional error is also referred to as the model comparison procedure and the process is quite flexible. For example, if you are interested in testing a hypothesis for the task and pulse rate data, like H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_a: (not H_0), then the model under the conditions of H_0 has the form

$$\begin{aligned} y_{ij} &= \mu_0 + \varepsilon_{ij} \quad \text{for } i = 1, 2, 3 \\ y_{ij} &= \mu_i + \varepsilon_{ij} \quad \text{for } i = 4, 5, 6 \end{aligned}$$

that is, the model has equal means for the first three tasks and different means for the last three treatments. Such a model can be fit using most software packages where a qualitative or class variable is defined to have the value of 0 for tasks 1, 2, and 3 and the value of task for tasks 4, 5, and 6.

1.10 Example—Tasks and Pulse Rate (Continued)

The principle of conditional error is applied to the task and pulse rate data of Section 1.4 to provide a test of the equal means hypothesis, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ vs H_a: (not H_0). The error sum of squares for the full model is $ESS_F = 1916.076$ with $df_F = 62$. The error sum of squares for the reduced model is $ESS_R = 73{,}593 - (2197)^2/68 = 2610.545$ with $df_R = 67$. Hence $SS_{H0} = 2610.545 - 1916.076 = 694.439$ with $df_{H0} = 67 - 62 = 5$. The analysis of variance table summarizing these computations is displayed in Table 1.3.

1.11 Computer Analyses

This chapter concludes with some remarks about utilizing computers and statistical computing packages such as SAS®, BMDP®, SYSTAT®, JMP®, and SPSS®. All of the methods and formulas provided in the preceding sections can easily be used on most computers. If the computer utilizes a programming language such as MATLAB, SAS-IML, or APL, the required matrix calculations are simple to do by following the matrix formulas given in the preceding sections. SAS, JMP, BMDP, SYSTAT, and SPSS each contain procedures that enable users to generate their own linear combinations of treatment means about which to test hypotheses. In addition, these packages all provide an analysis of variance table, treatment means, and their standard errors. Table 1.4 contains SAS-GLM code with estimate and contrast statements needed to test hypotheses described for the task and pulse data. The estimate statement is used to evaluate one linear combination of the means and the

TABLE 1.3

Analysis of Variance Table for Test Equality of the Means for the Task and Pulse Rate Data

Source of Variation	*df*	SS	MS	*F*	$\hat{\alpha}$
Due to H_0	5	694.439	138.888	4.49	0.0015
Error	62	1,916.076	30.9045		

TABLE 1.4

Proc GLM Code to Fit the Task and Pulse Rate Data with Estimate and Contrast Statements Needed to Provide the Analysis Described in the Text

```
PROC GLM DATA=EX1; CLASS TASK;
MODEL PULSE20=TASK/NOINT SOLUTION E;
ESTIMATE ‘Ho: M4=M5’ TASK 0 0 0 1 –1 0;
ESTIMATE ‘Ho: 3M1=M2+M3+M4’ TASK 3 –1 –1 –1 0 0;
ESTIMATE ‘Ho: 3M1=M2+M3+M4_mn’ TASK 3 –1 –1 –1 0 0/DIVISOR=3;
ESTIMATE ‘4M1–M3–M4–M5–M6_mn’ TASK 4 0 –1 –1 –1 –1/DIVISOR=4;
CONTRAST ‘4M1–M3–M4–M5–M6_mn’ TASK 4 0 –1 –1 –1 –1;
CONTRAST ‘M4=M5 & 3M1=M2+M3+M4’ TASK 0 0 0 1 –1 0, TASK 3 –1 –1 –1 0 0;
CONTRAST ‘EQUAL MEANS 1’
  TASK 1 –1 0 0 0 0, TASK 1 0 –1 0 0 0, TASK 1 0 0 –1 0 0,
    TASK 1 0 0 0 –1 0, TASK 1 0 0 0 0 –1;
```

TABLE 1.5

Proc IML Code to Carry Out the Computations for the Task and Pulse Data in Section 1.6

```
proc iml;
dd={13 12 10 10 12 11};
d=diag(dd);
c={0 0 0 1 –1 0, 3 –1 –1 –1 0 0};
muhat={31.9231 31.0833 35.8000 38.0000 29.5000 28.8182}`;
s2=30.90445;
a={4,0};
cmua=C*muhat - a;
cdc=c*inv(D)*c`;
cdci=inv(cdc);
ssho=cmua`*cdci*cmua;
f=ssho/(2*s2);al=1-probf(f,2,62);
print dd d cmua cdc cdci ssho f al;
```

provided results are the estimate of the contrast, its estimated standard error, and the resulting *t*-statistic with its corresponding significance level. The contrast statement is used to evaluate one or more linear combinations of the means and the provided results are the sums of squares, degrees of freedom, and the resulting *F*-statistic. For both the estimate and contrast statements in SAS-GLM, the only values of $\boldsymbol{a}$ in the hypotheses are zero, that is, one can only test the linear combinations of means that are equal to zero.

Table 1.5 contains SAS-IML code to provide the computations for the hypotheses being tested in Section 1.6. By constructing the code in a matrix language, one can obtain a test of any hypothesis of the form $\boldsymbol{C\mu} = \boldsymbol{a}$.

1.12 Concluding Remarks

In this chapter, the analysis of the one-way analysis of variance model was described. General procedures for making statistical inferences about the effects of different treatments were provided and illustrated for the case of homogeneous errors. Two basic procedures for obtaining statistical analyses of experimental design models were introduced. These procedures are used extensively throughout the remainder of the book for more complex models used to describe designed experiments and for messier data situations. A test for comparing all treatment effect means simultaneously was also given. Such a test may be considered an initial step in a statistical analysis. The procedures that should be used to complete the analysis of a data set could depend on whether the hypothesis of equal treatment means is rejected.

1.13 Exercises

1.1 A company studied five techniques of assembling a part. Forty workers were randomly selected from the worker population and eight were randomly assigned to each technique. The worker assembled a part and the measurement was the amount of time in seconds required to complete the assembly. Some workers did not complete the task.

Data for Comparing Techniques of Assembling a Part for Exercise 1.1

Technique 1		**Technique 2**		**Technique 3**		**Technique 4**		**Technique 5**	
Worker	**Time**	**Worker**	**Time**	**Worker**	**Time**	**Worker**	**Time**	**Worker**	**Time**
1	45.6	7	41.0	12	51.7	19	67.5	26	57.1
2	41.0	8	49.1	13	60.1	20	57.7	27	69.6
3	46.4	9	49.2	14	52.6	21	58.2	28	62.7
4	50.7	10	54.8	15	58.6	22	60.6		
5	47.9	11	45.0	16	59.8	23	57.3		
6	44.6			17	52.6	24	58.3		
				18	53.8	25	54.8		

1) Write down a model appropriate to describe the data. Describe each component of the model.

2) Estimate the parameters of the model in part 1.

3) Construct a 95% confidence interval about $\mu_1 - \mu_2$.

4) Use a t-statistic to test H_0: $\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$ vs H_a: (not H_0).

5) Use a F-statistic to test H_0: $\mu_1 + \mu_2 - \mu_3 - \mu_5 = 0$ vs H_a: (not H_0).

6) Use a t-statistic to test H_0: $(\mu_1 + \mu_2 + \mu_3)/3 = (\mu_4 + \mu_5)/2$ vs H_a: (not H_0).

7) Use a F-statistic to test H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_a: (not H_0).

8) Use a F-statistic to test H_0: $(\mu_1 + \mu_2 + \mu_3)/3 = (\mu_4 + \mu_5)/2$, $(\mu_1 + \mu_2 + \mu_6)/3 = (\mu_3 + \mu_4 + \mu_5)/3$, and $(\mu_1 + \mu_4 + \mu_5)/3 - (\mu_3 + \mu_6)/2$ vs H_a: (not H_0).

1.2 Five rations were evaluated as to their ability to enable calves to grow. Thirty-one calves were used in the study. A mistake in the feeding of the rations produced unbalanced distributions of the calves to the rations. The data recorded was the number of pounds of weight gained over the duration of the study.

1) Write down a model appropriate to describe the data. Describe each component of the model.

Gain Data for Comparing Rations of Exercise 1.2

Ration 1		Ration 2		Ration 3		Ration 4		Ration 5	
Calf	**Gain**	**Calf**	**Gain**	**Calf**	**Gain**	**Calf**	**Calf**	**Calf**	**Gain**
1	825	10	874	19	861	21	829	23	837
2	801	11	854	20	856	22	814	24	851
3	790	12	883					25	824
4	809	13	839					26	781
5	830	14	836					27	810
6	825	15	839					28	847
7	839	16	840					29	826
8	835	17	834					30	832
9	872	18	894					31	830

2) Estimate the parameters of the model in part 1.

3) Construct a 95% confidence interval about $\mu_1 + \mu_2 - 2\mu_5$.

4) Use a t-statistic to test H_0: $\mu_1 + \mu_2 - 2\mu_3 = 0$ vs H_a: (not H_0).

5) Use an F-statistic to test H_0: $2\mu_2 - \mu_4 - \mu_5 = 0$ vs H_a: (not H_0).

6) Use a t-statistic to test H_0: $(\mu_1 + \mu_2 + \mu_3)/3 = (\mu_4 + \mu_5)/2$ vs H_a: (not H_0).

7) Use an F-statistic to test H_0: $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$ vs H_a: (not H_0).

8) Use an F-statistic to test H_0: $\mu_1 + \mu_2 - 2\mu_3 = 0$, $2\mu_2 - \mu_4 - \mu_5 = 0$, $(\mu_1 + \mu_2 + \mu_3)/3 = (\mu_4 + \mu_5)/2$ vs H_a: (not H_0).

1.3 A study was conducted to evaluate the effect of elevation on the lung volume of birds raised at specified elevations. Thirty-five environmental chambers which could simulate elevations by regulating the air pressure were used. The five effective elevations were each randomly assigned to seven chambers and 35 baby birds were randomly assigned to the chambers, one per chamber. When the birds reached adult age, their lung volumes were measured. The data table contains the effective elevations and the volumes of the birds. Three birds did not survive the study, thus producing missing data.

Lung Volumes for Birds Raised at Different Simulated Elevations

Elevation 1000 ft		Elevation 2000 ft		Elevation 3000 ft		Elevation 4000 ft		Elevation 5000 ft	
Bird	**Volume**	**Bird**	**Volume**	**Bird**	**Volume**	**Bird**	**Volume**	**Bird**	**Volume**
1	156	8	160	15	156	22	168	29	177
2	151	9	160	16	173	23	167	30	170
3	161	12	154	18	165	24	171	31	169
4	152	13	152	19	172	25	173	32	176
5	164	14	153	20	169	26	167	33	183
6	153			21	168	27	167	34	178
7	163					28	173	35	174

1) Write down a model appropriate to describe the data. Describe each component of the model.

2) Estimate the parameters of the model in part 1.

3) Determine if there is a linear trend in the lung volume as elevation increases by testing H_0: $-2\mu_1 - \mu_2 - 0\mu_3 + \mu_4 + 2\mu_5 = 0$ vs H_a: (not H_0) (coefficients were obtained from a table of orthogonal polynomials for equally spaced values (Beyer, 1966, p. 367)).
4) Determine if there is a quadratic trend in the lung volume as elevation increases by testing H_0: $2\mu_1 - \mu_2 - 2\mu_3 - \mu_4 + 2\mu_5 = 0$ vs H_a: (not H_0).
5) Determine if the assumption of a linear/quadratic response to elevation is appropriate by simultaneously testing the cubic and quadratic trends to be zero by testing H_0: $-1\mu_1 + 2\mu_2 + 0\mu_3 - 2\mu_4 + 1\mu_5 = 0, 1\mu_1 - 4\mu_2 + 6\mu_3 - 4\mu_4 + 1\mu_5 = 0$ vs H_a: (not H_0).
6) Use a *t*-statistic to test H_0: $(\mu_1 + \mu_2 + \mu_3)/3 = (\mu_4 + \mu_5)/2$ vs H_a: (not H_0).
7) Use a *F*-statistic to test H_0: $\mu_1 = \mu_2 = \mu_3$ and $\mu_4 = \mu_5$ vs H_a: (not H_0).

2

One-Way Treatment Structure in a Completely Randomized Design Structure with Heterogeneous Errors

In this chapter, the case is considered where the treatments assigned to the experimental units may affect the variance of the responses as well as the mean. Start with the one-way means model, $y_{ij} = \mu_i + \varepsilon_{ij}$, for $i = 1, 2, \ldots, t$, $j = 1, 2, \ldots, n_i$. In Chapter 1 it was assumed that the experimental errors all had the same variance; that is, the treatments were expected to possibly change the mean of the population being sampled, but not the variance. In this chapter, some methods are described for analyzing data when the treatments affect the variances as well as the mean. The types of questions that the experimenter should want to answer about the means in this setting are similar to those in Chapter 1. That is,

1) Are all means equal?
2) Can pairwise comparisons among the means be made?
3) Can a test of the hypothesis of the form $\sum_{i=1}^{t} c_i\mu_i = a$ be tested and can confidence intervals be constructed about $\sum_{i=1}^{t} c_i\mu_i$?

In addition, there are also questions about the variances that may be of interest, such as

1) Are all of the variances equal?
2) Are there groupings of the treatments where within a group the variances are equal and between groups the variances are not equal?

Before questions about the means of the model can be answered, an appropriate description of the variances of the treatments must be obtained.

Tests of homogeneity of variances are used to answer questions about the variances of the data from the respective treatments. If there are two treatments, the problem of comparing means when there are unequal variances is usually known as the Behrens–Fisher problem. Also, heterogeneous error variances pose a much more serious problem when ignored than non-normality of the error variances. The procedures in Chapter 1 are robust with respect to non-normality, but not quite so robust with respect to heterogeneous error variances. In the analyses previously considered, it was assumed that the population variances were all equal, which is a reasonable assumption in many cases. One method for analyzing data when variances are unequal is simply to ignore the fact that they are unequal and calculate the same F-statistics or t-tests that are calculated in the case of equal variances. Surprisingly perhaps, simulation studies have shown that these usual tests are quite good, particularly if the sample sizes are all equal or almost equal. Also, if the larger sample sizes correspond to the treatments or populations with the larger variances, then the tests computed with the equal variance assumption are also quite good. The usual tests are so good, in fact, that many statisticians do not even recommend testing for equal variances. Others attempt to find a transformation that will stabilize the treatment variances, that is, transform the data such that the treatment variances are equal. When the variances are not equal, there are techniques to make comparisons about the means in the framework of the unequal variance model.

Procedures for testing the equality of treatment variances are described for the one-way model and procedures for analyzing the treatment means when the variances are unequal are described in the following sections. These procedures should be used when the usual techniques are suspect. The unequal variance model is described next.

2.1 Model Definitions and Assumptions

The unequal variance model is

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{for } i = 1,2,\dots,t,\ j = 1,2,\dots,n_i \text{ and } \varepsilon_{ij} \sim \text{independent } N(0, \sigma_i^2) \tag{2.1}$$

The notation ε_{ij}-independent $N(0, \sigma_i^2)$ means that the errors, ε_{ij}, are all independent, normally distributed and the variance of each normal distribution depends on i and may be different for each population or treatment.

2.2 Parameter Estimation

The best estimates of the parameters in the model are:

$$\hat{\mu}_i = \sum_{j=1}^{n_i} y_{ij}/n_i = \bar{y}_{i.}, \quad i = 1,2,\dots,t$$

and

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1}, \quad i = 1,2,\dots,t$$

The sampling distributions associated with the parameter estimates are

$$\hat{\mu}_i \sim \text{independent } N(\mu_i, \sigma_i^2/n_i), \quad i = 1,2,\ldots,t$$

and

$$\frac{(n_i - 1)\,\hat{\sigma}_i^2}{\sigma_i^2} \sim \text{independent } \chi^2_{n_i - 1}, \quad i = 1,2,\ldots,t$$

These sampling distributions are used as the basis for establishing tests for equality of variances and for providing the analysis of the means when the variances are unequal.

2.3 Tests for Homogeneity of Variances

In this section, five procedures are described for testing the equal variances hypothesis,

$$H_0\colon \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2 \text{ vs } H_a\colon (\text{not } H_0\colon)$$

Before the analysis of the means is attempted, the equal variance hypothesis should be investigated. If there is not enough evidence to conclude the variances are not equal, then the equal variance model in Chapter 1 can be used to investigate the means. If there is sufficient evidence to believe the variances are unequal, then the procedures described in Section 2.5 should be used to provide an analysis of the means in the unequal variance framework. The recommendation is to use the unequal variance model when the equal variance hypothesis is rejected at $\alpha \le 0.01$.

2.3.1 Hartley's *F*-Max Test

The first test described is known as Hartley's F-max test (1950). This test requires that all samples be of the same size, that is, $n_1 = n_2 = \cdots = n_t$. The test is based on the statistic

$$F_{\max} = \frac{\max\limits_i\{\hat{\sigma}_i^2\}}{\min\limits_i\{\hat{\sigma}_i^2\}}$$

Percentage points of $F_{\max}$ are provided in the Appendix in Table A.1 for $\alpha = 0.05$ and 0.01. The null hypothesis, H_0, is rejected if $F_{\max} > F_{\max,\alpha,\nu,k}$ where $\nu = n - 1$, the degrees of freedom associated with each of the k individual treatment variances. If the n_i are not all equal, a "liberal" test of H_0 vs H_a can be obtained by taking $\nu = \max_i\{n_i\} - 1$. This test is liberal in the sense that one is assuming all treatments have the same (maximum) sample size and so you are going to reject the null hypothesis more often than specified by the choice of α. When the sample sizes are not too unequal, this process provides a reasonable test. It also protects one from doing the usual analysis of variance when there is even a remote chance of it being inappropriate. An example illustrating the use of this test is found in Section 2.4.

2.3.2 Bartlett's Test

A second test for testing for homogeneity of variances is a test proposed by Bartlett (1937), which has the advantage of not requiring the n_i to be equal. Bartlett's test statistic is

$$U = \frac{1}{C}\left[v \log_e(\hat{\sigma}^2) - \sum_{i=1}^{t} v_i \log_e(\hat{\sigma}_i^2) \right] \tag{2.2}$$

where

$$v = n_i - 1, \quad v = \sum_{i=1}^{t} v_i, \quad \hat{\sigma}^2 = \sum_{i=1}^{t} v_i \hat{\sigma}_i^2 / v$$

and

$$C = 1 + \frac{1}{3(t-1)}\left[\sum_{i=1}^{t} \frac{1}{v_i} - \frac{1}{v} \right]$$

The hypothesis of equal variances is rejected if $U > \chi^2_{\alpha,t-1}$. One of the disadvantages of the preceding two tests for homogeneity of variance is that they are quite sensitive to departures from normality as well as to departures from the equal variances assumption. Most of the following tests are more robust to departures from normality.

2.3.3 Levene's Test

Levene (1960) proposed doing a one-way analysis of variance on the absolute values of the residuals from the one-way means or effects model. The absolute values of the residuals are given by $z_{ij} = |y_{ij} - \bar{y}_{i.}|$, $i = 1,2,\ldots,t$; $j = 1,2,\ldots,n_i$. The F-test from the analysis of variance is providing a test of the equality of the treatment means of the absolute values of the residuals. If the means are different, then there is evidence that the residuals for one treatment are on the average larger than the residuals for another treatment. The means of the absolute values of the residuals can provide a guide as to which variances are not equal and a multiple comparison test (see Chapter 3) can be used to make pairwise comparisons among these means. One modification of Levene's test is to use the squared residuals in the analysis of variance.

2.3.4 Brown and Forsythe's Test

Brown and Forsythe (1974) used Levene's process and modified it by doing a one-way analysis of variance on the absolute values of the deviations of the observations from the median of each treatment. The absolute values of the deviations from the medians are given by $u_{ij} = |y_{ij} - y_{i\,\mathrm{med}}|$, $i = 1,2,\ldots,t$; $j = 1,2,\ldots,n_i$. The F-test from the analysis of variance provides a test of the equality of the treatment means of the absolute values of the deviations. If the means are different, then there is evidence that the deviations for one treatment are on the average larger than the deviations for another treatment. The means of the absolute values of the deviations from the medians can provide a guide as to which variances are not equal as a multiple comparison tests can be used to make pairwise comparisons among these means. This use of the deviations from the medians provides more powerful tests than Levene's when the data are not symmetrically distributed.

2.3.5 O'Brien's Test

O'Brien (1979) computed scores as

$$r_{ij} = [(w + n_i - 2)n_i(y_{ij} - \bar{y}_{i.})^2 - w\hat{\sigma}_i^2(n_i - 1)]/[(n_i - 1)(n_i - 2)] \tag{2.3}$$

where w is a weight parameter. The procedure is to carry out an analysis of variance on the computed score values. When $w = 0.5$, the means of the scores are the sample variances, $\hat{\sigma}_i^2$, thus the comparison of the means of the scores is a comparison of the variances of the data.

There are several other procedures that can be used to test the equality of variances or the equality of scale parameters using parametric and nonparametric methods (Conover et al., 1981; Olejnik and Algina, 1987). McGaughey (2003) proposes a test that uses the concept of data depth and applies the procedure to univariate and multivariate populations. Data depth is beyond the scope of this book.

2.3.6 Some Recommendations

Conover et al. (1981) and Olejnik and Algina (1987) conducted simulation studies of homogeneity of variance tests that included the ones above as well as numerous others. The studies indicate that no test is robust and most powerful for all situations. Levene's test was one of the better tests studied by Conover et al. O'Brien's test seems to provide an appropriate size test without losing much power according to Olejnik and Algina. The Brown–Forsythe test seems to be better when distributions have heavy tails. Based on their results, we make the following recommendations:

1) If the distributions have heavy tails, use the Brown–Forsythe test.
2) If the distributions are somewhat skewed, use the O'Brien test.
3) If the data are nearly normally distributed, then any of the tests are appropriate, including Bartlett's and Hartley's tests.

Levene's and O'Brien's tests can easily be tailored for use in designed experiments that involve more than one factor, including an analysis of covariance (Milliken and Johnson, 2002). Levene's, O'Brien's and Brown–Forsythe's tests were shown to be nearly as good as Bartlett's and Hartley's tests for normally distributed data, and superior to them for nonnormally distributed data. Conover et al. and Olejnik and Algina discuss some nonparametric tests, but they are more difficult to calculate and the above recommended tests perform almost as well. An example follows where each of the tests for equality of variances is demonstrated.

2.4 Example—Drugs and Errors

The data in Table 2.1 are from a paired-association learning task experiment performed on subjects under the influence of two possible drugs. Group 1 is a control group (no drug), group 2 was given drug 1, group 3 was given drug 2, and group 4 was given both drugs.

TABLE 2.1

Data from Paired-Association Learning Task Experiment

	No Drug	Drug 1	Drug 2	Drugs 1 and 2
	1	12	12	13
	8	10	4	14
	9	13	11	14
	9	13	7	17
	4	12	8	11
	0	10	10	14
	1	—	12	13
	—	—	5	14
n	7	6	8	8
Sum	32	70	69	110
Median	4	12	9	14
Mean	4.5714	11.6667	8.6250	13.750
Variance	16.2857	1.8667	9.6964	2.786

The sample sizes, sums, medians, means and variances of each group's data are included in Table 2.1.

The F-max statistic is $F_{\max} = 16.286/1.867 = 8.723$. The liberal 5% critical point is obtained from Table A.1 with $k = t = 4$ and $\nu = 7$. The critical point is 8.44 and since $8.723 > 8.44$, one rejects H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$ versus H_a: (not H_0:) with significance level 0.05, but cannot reject at the $\alpha = 0.01$ level.

The computations for Bartlett's test are:

$$C = 1 + \frac{1}{3\times 3}\left(\frac{1}{6} + \frac{1}{5} + \frac{1}{7} + \frac{1}{7} - \frac{1}{25}\right)$$

and

$$\hat{\sigma}^2 = \frac{6(16.2857) + 5(1.8667) + 7(9.6964) + 7(2.7860)}{25} = 7.7769$$

Thus

$$\begin{aligned} U &= \frac{1}{C}\left(\nu \log_e \hat{\sigma}^2 - \sum_{i=1}^{4} \nu_i \log_e \hat{\sigma}_i^2\right) \\ &= \frac{1}{1.068}[25 \log_e(7.7769) - 6 \log_e(16.2857) - 5 \log_e(1.8667) \\ &\quad - 7 \log_e(9.6964) - 7 \log_e(2.7860)] \\ &= 7.8111 \end{aligned}$$

The asymptotic sampling distribution associated with U is a that of a chi-square distribution based on three degrees of freedom. The significance level of the test is 0.0501 and one would again conclude that the variances are unequal at an approximate 5% significance level.

The computations for Levene's test begin with the computation of the residuals or the deviations of the observations from the treatment means. Next the absolute values of the residuals are computed as illustrated in Table 2.2. Finally, a one-way analysis of variance is carried out on these absolute values of the residuals. The value of the resulting *F*-statistic is 6.97, which is based on 3 and 25 degrees of freedom. The observed significance level of Levene's test is 0.0015. The squared deviations or squared residuals version of Levene's test can be obtained by squaring the items in Table 2.2 before doing the analysis of variance. In this case, the value of the *F*-statistic is 7.36 and the observed significance level is 0.0011 (also based on 3 and 25 degrees of freedom).

The Brown–Forsythe test statistic is obtained by computing the absolute value of the deviations of the observations from the treatment median (medians are in Table 2.1). Table 2.3 contains the absolute values of the deviations from the medians. Next, the one-way analysis of variance provides an *F*-statistic of 5.49 and the observed significance level is 0.0049 (also based on 3 and 25 degrees of freedom).

Table 2.4 contains the values of r_{ij} computed using Equation 2.3 with $w = 0.5$. The O'Brien test statistic is obtained by carrying out an analysis of variance. The value of the *F*-statistic

TABLE 2.2

Values of $z_{ij} = |y_{ij} - \bar{y}_{i.}|$ for Computing Levene's Test Where y_{ij} Values are from Table 2.1

No Drug	Drug 1	Drug 2	Drugs 1 and 2
3.571	0.333	3.375	0.750
3.429	1.667	4.625	0.250
4.429	1.333	2.375	0.250
4.429	1.333	1.625	3.250
0.571	0.333	0.625	2.750
4.571	1.667	1.375	0.250
3.571	—	3.375	0.750
—	—	3.625	0.250

TABLE 2.3

Absolute Values of Deviations of the Observations from the Treatment Medians

No Drug	Drug 1	Drug 2	Drugs 1 and 2
3	0	3	1
4	2	5	0
5	1	2	0
5	1	2	3
0	0	1	3
4	2	1	0
3	—	3	1
—	—	4	0

TABLE 2.4

Scores Using $w = 0.5$ for O'Brien's Test

Obr1	Obr2	Obr3	Obr4
14.740	–0.083	13.295	0.464
13.457	3.517	25.676	–0.155
23.540	2.167	6.176	–0.155
23.540	2.167	2.461	12.845
–1.210	–0.083	–0.324	9.131
25.190	3.517	1.533	–0.155
14.740	—	13.295	0.464
—	—	15.461	–0.155

is 6.30 and the observed significance level is 0.0025. The value of the F-statistic using $w = 0.7$ (computations not shown) is 5.90 and the observed significance level is 0.0035. There are 3 and 25 degrees of freedom associated with each of O'Brien's tests.

Each of the test statistics indicates that there is sufficient evidence to conclude that the variances are not equal. The group means of the absolute values of the residuals are shown in Table 2.5. Pairwise comparisons among these treatment absolute residual means are shown in Table 2.6. The means of the absolute values of the residuals for no drug and drug 2 are not different, for drug 1 and drugs 1 and 2 are not different, but there are differences between these two sets. A simple model with two variances could be used to continue the analysis of the treatment means. Using a simple variance model will improve the power of some of the tests about the means. The two variance model and the corresponding comparisons of means will follow the discussion of the analysis using four variances.

TABLE 2.5

Means of the Absolute Values of the Residuals

Group	Estimate	Standard Error	*df*	*t*-Value	Pr > |*t*|
Both drugs	1.0625	0.4278	25	2.48	0.0201
Drug 1	1.1111	0.4940	25	2.25	0.0336
Drug 2	2.6250	0.4278	25	6.14	<0.0001
No drug	3.5102	0.4574	25	7.67	<0.0001

TABLE 2.6

Pairwise Comparisons between the Group Means of the Absolute Values of the Residuals

Group	_Group	Estimate	Standard Error	*df*	*t*-Value	Pr > |*t*|
Both drugs	Drug 1	–0.04861	0.6535	25	–0.07	0.9413
Both drugs	Drug 2	–1.5625	0.6050	25	–2.58	0.0161
Both drugs	No drug	–2.4477	0.6263	25	–3.91	0.0006
Drug 1	Drug 2	–1.5139	0.6535	25	–2.32	0.0290
Drug 1	No drug	–2.3991	0.6732	25	–3.56	0.0015
Drug 2	No drug	–0.8852	0.6263	25	–1.41	0.1699

2.5 Inferences on Linear Combinations

The problems of testing hypotheses about and constructing confidence intervals for an arbitrary linear combination of the treatment means, $\sum_{i=1}^{t} c_i \mu_i$, are discussed in this section when the variances σ_i^2 are too unequal to apply the tests and confidence intervals discussed in Chapter 1. It is recommended that you use the procedures in this section and the next if the equality of variance hypothesis is rejected at the 0.01 or 1% level. If there is not sufficient evidence to believe that the variances are unequal, then one can use the results in Chapter 1 to make inferences about the treatment means.

The best estimate of $\sum_{i=1}^{t} c_i \mu_i$ is $\sum_{i=1}^{t} c_i \hat{\mu}_i$ and the sampling distribution is

$$\sum_{i=1}^{t} c_i \hat{\mu}_i \sim N\left(\sum_{i=1}^{t} c_i \mu_i, \sum_{i=1}^{t} c_i^2 \sigma_i^2 / n_i\right)$$

and thus,

$$z = \frac{\sum_{i=1}^{t} c_i \hat{\mu}_i - \sum_{i=1}^{t} c_i \mu_i}{\sqrt{\sum_{i=1}^{t} c_i^2 \sigma_i^2 / n_i}} \sim N(0, 1)$$

An obvious statistic to use for making inferences about $\sum_{i=1}^{t} c_i \mu_i$, when the variances are not known and are unequal, is

$$z = \frac{\sum_{i=1}^{t} c_i \hat{\mu}_i - \sum_{i=1}^{t} c_i \mu_i}{\sqrt{\sum_{i=1}^{t} c_i^2 \hat{\sigma}_i^2 / n_i}}$$

If the n_i corresponding to nonzero c_i are all very large, one can reasonably assume that Z has an approximate $N(0, 1)$ distribution, and hence Z can be used to make inferences about $\sum_{i=1}^{t} c_i \mu_i$. In this case, an approximate $(1 - \alpha)100\%$ confidence interval for $\sum_{i=1}^{t} c_i \mu_i$, is provided by

$$\sum_{i=1}^{t} c_i \hat{\mu}_i \pm z_{\alpha/2} \sqrt{\sum_{i=1}^{t} c_i^2 \hat{\sigma}_i^2 / n_i}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ critical point of the standard normal probability distribution.

To test H_0: $\sum_{i=1}^{t} c_i \mu_i = a$ vs H_a: $\sum_{i=1}^{t} c_i \mu_i \neq a$, where a is a specified constant, one could calculate

$$z = \frac{\sum_{i=1}^{t} c_i \hat{\mu}_i - a}{\sqrt{\sum_{i=1}^{t} c_i^2 \hat{\sigma}_i^2 / n_i}}$$

and if $|z| > z_{\alpha/2}$, then reject H_0 at a significance level of α.

In other instances, note that z can be written as

$$z = \frac{\left(\sum_{i=1}^{t} c_i\hat{\mu}_i - \sum_{i=1}^{t} c_i\mu_i\right)\Big/\sqrt{\sum_{i=1}^{t} c_i^2\sigma_i^2/n_i}}{\sqrt{\sum_{i=1}^{t} c_i^2\hat{\sigma}_i^2/n_i}\Big/\sqrt{\sum_{i=1}^{t} c_i^2\sigma_i^2/n_i}}$$

The numerator of z has a standard normal distribution and the numerator and denominator of z are independently distributed. The distribution of z could be approximated by a $t(v)$ distribution if v could be determined such that

$$V = v \times \frac{\sum_{i=1}^{t} c_i^2\hat{\sigma}_i^2/n_i}{\sum_{i=1}^{t} c_i^2\sigma_i^2/n_i}$$

is approximately distributed as $\chi^2(v)$. In order to get a good chi-square approximation to the distribution of V when the variances are unequal, select a chi-square distribution that has the same first two moments as V. That is, to find v for the case of unequal variances, find v so that the moments of V are equal to the first two moments of a $\chi^2(v)$ distribution (this is known as Satterthwaite's method). This results in determining that the approximate number of degrees of freedom is

$$v = \frac{\left(\sum_{i=1}^{t} c_i^2\sigma_i^2/n_i\right)^2}{\sum_{i=1}^{t}[c_i^4\sigma_i^4/n_i^2(n_i-1)]}$$

Unfortunately, since v depends on $\sigma_1^2, \sigma_2^2, \ldots, \sigma_t^2$ it cannot be determined exactly. The usual procedure is to estimate v by

$$\hat{v} = \frac{\left(\sum_{i=1}^{t} c_i^2\hat{\sigma}_i^2/n_i\right)^2}{\sum_{i=1}^{t}[c_i^4\hat{\sigma}_i^4/n_i^2(n_i-1)]} \tag{2.4}$$

Summarizing, one rejects H_0: $\sum_{i=1}^{t} c_i\mu_i = a$ vs H_a: $\sum_{i=1}^{t} c_i\mu_i \neq a$, if

$$|t_c| = \frac{\left|\sum_{i=1}^{t} c_i\hat{\mu}_i - a\right|}{\sqrt{\sum_{i=1}^{t} c_i^2\hat{\sigma}_i^2/n_i}} > t_{\alpha/2,\hat{v}}$$

where $\hat{\nu}$ is determined using Equation 2.4. An approximate $(1 - \alpha)100\%$ confidence interval for $\sum_{i=1}^{t} c_i \mu_i$ is given by

$$\sum_{i=1}^{t} c_i \hat{\mu}_i \pm t_{\alpha/2,\hat{\nu}} \sqrt{\sum_{i=1}^{t} c_i^2 \hat{\sigma}_i^2 / n_i}$$

Unfortunately, every time one wants to test a new hypothesis or construct another confidence interval, the degrees of freedom $\hat{\nu}$ must be re-estimated. It can be shown that $n_* - 1 \le \hat{\nu} \le t(n^* - 1)$ where $n_* = \min\{n_1, n_2, \ldots, n_t\}$ and $n^* = \max\{n_1, n_2, \ldots, n_t\}$. Thus, if $|t_c| > t_{\alpha/2, n_* - 1}$, one can be assured that $|t_c| > t_{\alpha/2,\hat{\nu}}$, and if $|t_c| < t_{\alpha/2, t(n^*-1)}$, one can be assured that $|t_c| < t_{\alpha/2,\hat{\nu}}$. In these cases, one can avoid calculating $\hat{\nu}$. When $t_{\alpha/2, t(n^*-1)} < |t_c| < t_{\alpha/2, n^*-1}$ the value of $\hat{\nu}$ must be calculated in order to be sure whether one should reject or fail to reject the null hypothesis being tested. For confidence intervals, $\hat{\nu}$ should always be calculated. Next, the preceding results are demonstrated with the drug errors example.

2.6 Example—Drugs and Errors (Continued)

Consider the data in Table 2.1, and suppose the experimenter is interested in answering the following questions:

1) On average, do drugs have any effect on learning at all?
2) Do subjects make more errors when given both drugs than when given only one?
3) Do the two drugs differ in their effects on the number of errors made?

To answer the first question, one might test the hypothesis that the mean of the three drug groups is equal to the control mean. That is, one would test

$$H_{01}\colon l_1 = \mu_1 - \frac{(\mu_2 + \mu_3 + \mu_4)}{3} = 0 \text{ vs } H_{a1}\colon l_1 \neq 0$$

The estimate of this linear combination is

$$\hat{l}_1 = \hat{\mu}_1 - \frac{\hat{\mu}_2 + \hat{\mu}_3 + \hat{\mu}_4}{3} = 4.571 - \frac{1}{3}(34.042) = -6.776$$

and the estimate of the corresponding standard error of $\hat{l}_1$ is

$$s.e.(\hat{l}_1) = \sqrt{\sum_{i=1}^{4} \left(\frac{c_i^2 \hat{\sigma}_i^2}{n_i} \right)}$$

$$= \sqrt{\frac{\hat{\sigma}_1^2}{7} + \frac{1}{9}\left(\frac{\hat{\sigma}_2^2}{6}\right) + \frac{1}{9}\left(\frac{\hat{\sigma}_3^2}{8}\right) + \frac{1}{9}\left(\frac{\hat{\sigma}_4^2}{8}\right)} = \sqrt{2.535} = 1.592$$

The approximate degrees of freedom associated with this estimated standard error are obtained by using

$$\sum_{i=1}^{4} \frac{c_i^4 \hat{\sigma}_i^4}{n_i^2(n_i - 1)} = 0.9052$$

so that

$$\hat{\nu} = \frac{(2.535)^2}{0.9052} = 7.10$$

The value of the test statistic is $t_c = -6.776/1.992 = -4.256$ with the observed significance level $\hat{\alpha} = 0.0038$.

A 95% confidence interval for l_1 is

$$\hat{l}_1 \pm t_{\alpha/2,\hat{\nu}} \times \widehat{s.e.}(\hat{l}_1) = -6.776 \pm (2.365)(1.592)$$

which simplifies to

$$-10.54 < \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3} < -3.01$$

Next test to see if the mean of the group given both drugs is equal to the mean of the average of the means of the two groups given a single drug. That is, test

$$H_{02}\colon l_2 = \mu_4 - \frac{\mu_2 + \mu_3}{2} = 0 \text{ vs } H_{a2}\colon l_2 \neq 0$$

The estimate of this linear combination is

$$\hat{l}_2 = \hat{\mu}_4 - \frac{\hat{\mu}_2 + \hat{\mu}_3}{2} = 3.6042$$

and its estimated standard error is

$$\widehat{s.e.}(\hat{l}_2) = \sqrt{\sum_{i=1}^{4}\left(\frac{c_i^2\hat{\sigma}_i^2}{n_i}\right)}$$

$$= \sqrt{\frac{1}{4}\left(\frac{\hat{\sigma}_2^2}{6}\right) + \frac{1}{4}\left(\frac{\hat{\sigma}_3^2}{8}\right) + \left(\frac{\hat{\sigma}_4^2}{8}\right)} = \sqrt{0.7290} = 0.8538$$

The value of the test statistic is $t_c = 3.6042/0.8538 = 4.221$, which is significant at $\alpha = 0.01$ since $|t_c| > t_{0.005,5}$. In this case, the value of $\hat{\nu}$ need not be computed using $n_* - 1$ as the approximating degrees of freedom. The computed value of $\hat{\nu}$ is 16.8, which would be needed if one wanted to construct a confidence interval about l_2.

Finally, to test the hypothesis to see if the two drug means differ, test H_0: $l_3 = \mu_2 - \mu_3 = 0$ vs H_a: $l_3 = \mu_2 - \mu_3 \neq 0$. The estimate of this linear combination is $\hat{l}_3 = \hat{\mu}_2 - \hat{\mu}_3 = 3.042$ and its estimated standard error is

$$\widehat{s.e.}(\hat{l}_3) = \sqrt{\sum_{i=1}^{4}\left(\frac{c_i^2\hat{\sigma}_i^2}{n_i}\right)} = \sqrt{\left(\frac{\hat{\sigma}_2^2}{6}\right) + \left(\frac{\hat{\sigma}_3^2}{8}\right)} = \sqrt{1.523} = 1.234$$

The approximate number of degrees of freedom is computed using

$$\sum_{i=1}^{4}\frac{c_i^4\hat{\sigma}_i^4}{n_i^2(n_i-1)} = 0.229$$

so that

$$\hat{v} = \frac{(1.523)^2}{0.229} = 10.1$$

Thus, $t_c = 3.042/1.234 = 2.465$, which has an observed significance level of $\hat{\alpha} = 0.0334$.

2.7 General Satterthwaite Approximation for Degrees of Freedom

The Satterthwaite approximation to the number of degrees of freedom associated with estimated standard error is obtained from

$$v = \frac{2 * \left(E\{[\widehat{s.e.}(\hat{l})]^2\}\right)^2}{\mathrm{Var}\{[\widehat{s.e.}(\hat{l})]^2\}}$$

where $[\widehat{s.e.}(\hat{l})]^2$ is used to estimate $E[s.e.(\hat{l})]^2$ and the $\mathrm{Var}[\widehat{s.e.}(\hat{l})]^2$ is estimated by $\sum_{i=1}^{t} c_i^4\hat{\sigma}_i^4/[n_i^2(n_i - 1)]$. For more complex models, $\mathrm{Var}[\widehat{s.e.}(\hat{l})]^2$ can be approximated by using a first-order Taylor's series (Kendall and Stuart, 1952) as $\boldsymbol{q'Mq}$ where $\boldsymbol{M}$ is the estimated asymptotic covariance matrix of the estimates of the variances and the elements of the vector $\boldsymbol{q}$ are the first derivatives of $E[s.e.(\hat{l})]^2$ with respect to the individual variances, that is,

$$q_i = \frac{E[(s.e.(\hat{l})]^2}{\partial\sigma_i^2}, \quad i = 1,2,\ldots,t$$

The q_i are evaluated at the estimated values of each treatment's variances (Montgomery and Runger, 1993, 1994). When the data from each of the samples are normally distributed, then

$$\frac{(n_i-1)\hat{\sigma}_i^2}{\sigma_i^2}$$

is distributed as a central chi-square random variable. Thus $E(\hat{\sigma}_i^2) = \sigma_i^2$ and $\mathrm{Var}(\hat{\sigma}_i^2) = 2\sigma_i^4/(n_i - 1)$. Let the linear combination of interest be $l = \sum_{i=1}^{t} c_i\mu_i$, which has variance $\sigma_l^2 = \sum_{i=1}^{t} c_i^2\sigma_i^2/n_i$. The partial derivative of σ_l^2 with respect to σ_i^2 is

$$\frac{\partial\sigma_l^2}{\partial\sigma_i^2} = \frac{c_i^2}{n_i}$$

The approximate variance of σ_l^2 obtained using the Taylor's series first-order approximation is

$$\text{Var}(\sigma_l^2) = \sum_{i=1}^{t}\left[\left[\frac{c_i^2}{n_i}\right]^2\left[\frac{2\sigma_i^4}{n_i-1}\right]\right]^2$$

The next step is to replace the population variances with their corresponding sample estimates providing the approximating degrees of freedom

$$\hat{v} = \frac{2 * \left(E\{[\widehat{s.e.}(\hat{l})]^2\}\right)^2}{\text{Var}\{[\widehat{s.e.}(\hat{l})]^2\}} = \frac{\left(\sum_{i=1}^{t} c_i^2\hat{\sigma}_i^2\right)^2}{\sum_{i=1}^{t} c_i^4\hat{\sigma}_i^4/[n_i^2(n_i-1)]}$$

the same as that provided by the Satterthwaite approximation above.

2.8 Comparing All Means

As previously stated, the usual F-test is very robust when the variances are unequal, provided that the sample sizes are nearly equal or provided that the larger sample sizes correspond to the samples from populations with the larger differences of variances. In this section, two additional tests of the hypothesis of equal means are provided. The first test of the equal means hypothesis, H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$ vs H_a: (not H_0:), is given by Welch (1951), and is known as Welch's test. Define weights $W_i = n_i/\hat{\sigma}_i^2$, let $\bar{y}^* = \sum_{i=1}^{t} W_i\bar{y}_{i\cdot}/\sum_{i=1}^{t} W_i$ be a weighted average of the sample means, and let

$$\Lambda = \sum_{i=1}^{t}\frac{(1-W_i/W.)^2}{n_i-1}$$

where $W. = \sum_{i=1}^{t} W_i$. Then Welch's test statistic is

$$F_c = \frac{\sum_{i=1}^{t} W_i\dfrac{(\bar{y}_{i\cdot}-\bar{y}^*)}{(t-1)}}{1+2(t-1)\Lambda/(t^2-1)} \tag{2.5}$$

which has an approximate F-distribution with numerator and denominator degrees of freedom, $v_1 = t-1$ and $v_2 = (t^2-1)/3\Lambda$, respectively. Thus, the null hypothesis H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$ is rejected if $F_c > F_{\alpha,v_1,v_2}$. The numerator of Equation 2.5 can also be computed as

TABLE 2.7

Quantities for Computing Welch's Test

i	Drug 1	Drug 2	Drug 3	Drug 4
n_i	7	6	8	8
$\bar{y}_{i\cdot}$	4.5714	11.6667	8.62500	13.7500
$\hat{\sigma}_i^2$	16.2857	1.8667	9.69643	2.7857
W_i	0.4298	3.2143	0.82505	2.8718

$[\sum_{i=1}^{t}(W_i \bar{y}_{i\cdot}^2) - W.\bar{y}^{*2}]/(t-1)$. The procedure is demonstrated using the data from Section 2.4 and the preliminary computations are provided in Table 2.7.

From the above information compute $W. = 7.341$, $\bar{y}^* = 11.724$,

$$\Lambda = \frac{(1-0.430/7.341)^2}{6} + \frac{(1-3.214/7.341)^2}{5} + \frac{(1-0.825/7.341)^2}{7} + \frac{(1-2.872/7.341)^2}{7} = 0.376$$

and $\sum_{i=1}^{t} W_i \bar{y}_{i\cdot}^2 - \overline{W}\,\bar{y}^{*2} = 1050.8069 - 1009.0954 = 43.7114$.

The value of Welch's test statistic is

$$F_c = \frac{41.7114/3}{1 + 2\times 2\times 0.376/15} = \frac{13.9038}{1.1003} = 12.6355$$

with $\nu_1 = 3$ and $\nu_2 = 15/(3 \times 0.376) = 13.283$ degrees of freedom. The observed significance probability corresponding to F_c is $\hat{\alpha} = 0.00035$. For comparison purposes, the usual F-statistic is $F_c = 14.91$ with 3 and 25 degrees of freedom. Welch's test can be obtained using SAS®-GLM by specifying WELCH as an option on the MEANS statement. Table 2.8 contains the

TABLE 2.8

SAS-GLM Code to Provide the Brown–Forsythe's Test of Equality of Variances and to Provide the Welch Test of Equal Means with the Unequal Variance Model

```
proc glm data=task;
class group;
model errors=group;
means group/HOVTEST=BF WELCH;
format group druggrps.;
```

Welch's Test

Source	*df*	*F*-Value	Pr > *F*
Group	3	12.64	0.0003
Error	13.2830		

Brown and Forsythe's Test for Homogeneity of Errors Variance
ANOVA of Absolute Deviations from Group Medians

Source	*df*	Sum of Squares	Mean Square	*F*-Value	Pr > *F*
Group	3	31.3762	10.4587	5.49	0.0049
Error	25	47.5893	1.9036		

GLM code used to provide BF test for equality of variances and Welch's test for equality of means. The important parts of the output are in the second part of Table 2.8. Other tests for equality of variances can be obtained by specifying O'Brien, Levene or Bartlett.

The second procedure for testing the equality of the treatment means is obtained from generalizing the process of testing a hypothesis about a set of linear combinations of the μ_i. Suppose a hypothesis is formed involving r independent linear combinations of the μ_i, such as H_0: $\sum_{i=1}^{t} c_{1i}\mu_i = 0, \sum_{i=1}^{t} c_{2i}\mu_i = 0, \ldots, \sum_{i=1}^{t} c_{ri}\mu_i = 0$ vs H_a: (not H_0). Let $\boldsymbol{C}$ be a $r \times t$ matrix where the kth row contains the coefficients of the kth linear combination. If one assumes the data from each of the populations or treatments are normally distributed, then the joint sampling distribution of the vector of treatment means is $\hat{\boldsymbol{\mu}} \sim N[\boldsymbol{\mu}, \boldsymbol{V}]$ where $\boldsymbol{V}$ is a diagonal matrix whose ith diagonal element is σ_i^2/n_i. The joint sampling distribution of the set of linear combinations $\boldsymbol{C\mu}$ is $\boldsymbol{C\hat{\mu}} \sim N[\boldsymbol{C\mu}, \boldsymbol{CVC'}]$. The sum of squares due to deviations from the null hypothesis is $SSH_0 = [\boldsymbol{C\hat{\mu}}]'[\boldsymbol{C\hat{V}C'}]^{-1}[\boldsymbol{C\hat{\mu}}]$, which is asymptotically distributed as a chi-square distribution with r degrees of freedom. An approximate small sample size statistic is $F_c = SSH_0/r$ with the approximating distribution being F with r and ν degrees of freedom where ν needs to be approximated (Fai and Cornelius, 1996; SAS Institute, Inc., 1999, p. 2118). The computation of the approximate degrees of freedom starts with carrying out a spectral decomposition on $\boldsymbol{C\hat{V}C'} = \boldsymbol{QDQ'}$ where $\boldsymbol{D}$ is an $r \times r$ diagonal matrix having the characteristic roots of $\boldsymbol{C\hat{V}C'}$ as diagonal elements and where $\boldsymbol{Q}$ is a $r \times r$ orthogonal matrix of the corresponding characteristic vectors of $\boldsymbol{C\hat{V}C'}$. Let z_k' be the kth row of $\boldsymbol{QC}$, and let

$$\nu_k = \frac{2(d_k)^2}{b_k' M b_k}$$

where d_k is the kth diagonal element of $\boldsymbol{D}$, b_k contains the partial derivatives of $\boldsymbol{z_k' V z_k}$ with respect to each of the variance parameters in $\boldsymbol{V}$ evaluated at the estimates of the variances, and $\boldsymbol{M}$ is the asymptotic covariance of the vector of variances. Let

$$S = \sum_{k=1}^{r} \frac{\nu_k}{\nu_k - 2} I[\nu_k > 2]$$

where $I[\nu_k > 2]$ is an indicator function with the value of 1 when $\nu_k > 2$ and 0 otherwise. The approximate denominator degrees of freedom for the distribution of F_c are

$$\nu = \begin{cases} \dfrac{2S}{S-r} & \text{if } S > r \\ 0 & \text{if } S \le r \end{cases}$$

The above process can be used to provide a test of the equal means hypothesis by selecting a set of $t - 1$ linearly independent contrasts of the μ_i.

The SAS-Mixed procedure implements a version of this approximation to the denominator degrees of freedom associated with an approximate F statistic with multiple degrees of freedom in the numerator. SAS-Mixed can be used to fit models with unequal variances per treatment group or unequal variances in some other prespecified pattern using the REPEATED statement and specifying the GROUP = option. The Mixed code in Table 2.9 was used to fit the unequal variance model to the data in Table 2.1. The REPEATED statement is used to specify that a different variance (each value of group) is to be estimated for

TABLE 2.9

SAS-Mixed Code to Fit the Unequal Variance Model to the Data in Table 2.1

```
proc mixed cl covtest data=task;
class group;
model errors=group/ddfm=kr;
repeated/group=group;
estimate “part(1)” group −1 −1 −1 3/divisor=3 cl alpha=0.05;
estimate “part(2)” group 2 −1 −1 0/divisor=2 cl alpha=0.05;
estimate “part(3)” group 0 1 −1 0/cl alpha=0.05;
lsmeans group/diff cl;
```

each treatment. The three Estimate statements are used to provide the computations corresponding to the three questions in Section 2.6.

The results from the Mixed procedure are given in Table 2.10, where the Covariance Parameter Estimates are the estimates of the four treatment variances, AIC in the Fit Statistics is the Akaike Information Criteria (Akaike, 1974), the Null Model Likelihood Ratio Test provides a test of the equal variance hypothesis, the type III tests of fixed effects provides the test of the equal means hypothesis using the second statistic and the corresponding

TABLE 2.10

Results of Fitting the Unequal Variance Model to the Data in Table 2.1

Covariance Parameter Estimates

Covariance			Standard					
Parameter	Group	Estimate	Error	*Z*-Value	Pr *Z*	α	Lower	Upper
Residual	Both drugs	2.7857	1.4890	1.87	0.0307	0.05	1.2178	11.5394
Residual	Drug 1	1.8667	1.1806	1.58	0.0569	0.05	0.7273	11.2286
Residual	Drug 2	9.6964	5.1830	1.87	0.0307	0.05	4.2388	40.1658
Residual	No drug	16.2857	9.4026	1.73	0.0416	0.05	6.7625	78.9710

Fit Statistics

AIC (smaller is better)	129.8

Null Model Likelihood Ratio Test

df	Chi-Square	Pr > Chi-Square
3	8.34	0.0394

Type III Tests of Fixed Effects

Effect	Num *df*	Den *df*	*F*-Value	Pr > *F*
group	3	11.8	12.53	0.0006

Estimates

Label	Estimate	Standard Error	*df*	*t*-Value	Pr > |*t*|	α	Lower	Upper
Part 1	−6.7758	1.5920	7.1	−4.26	0.0036	0.05	−10.5299	−3.0217
Part 2	3.6042	0.8538	16.8	4.22	0.0006	0.05	1.8011	5.4073
Part 3	3.0417	1.2342	10.1	2.46	0.0332	0.05	0.2962	5.7871

TABLE 2.11

Estimates of the Drug Group Means and Pair Wise Comparisons Using the Unequal Variance Model

Least Squares Means

Effect	Group	Estimate	Standard Error	*df*	*t*-Value	Pr > |*t*|	α	Lower	Upper
Group	Both drugs	13.7500	0.5901	7	23.30	<0.0001	0.05	12.3546	15.1454
Group	Drug 1	11.6667	0.5578	5	20.92	<0.0001	0.05	10.2329	13.1005
Group	Drug 2	8.6250	1.1009	7	7.83	0.0001	0.05	6.0217	11.2283
Group	No drug	4.5714	1.5253	6	3.00	0.0241	0.05	0.8392	8.3037

Differences of Least Squares Means

Effect	Group	_Group	Estimate	Standard Error	*df*	*t*-Value	Pr > |*t*|	α	Lower	Upper
Group	Both drugs	Drug 1	2.0833	0.8120	11.9	2.57	0.0249	0.05	0.3117	3.8550
Group	Both drugs	Drug 2	5.1250	1.2491	10.7	4.10	0.0018	0.05	2.3668	7.8832
Group	Both drugs	No drug	9.1786	1.6355	7.78	5.61	0.0006	0.05	5.3886	12.9685
Group	Drug 1	Drug 2	3.0417	1.2342	10.1	2.46	0.0332	0.05	0.2962	5.7871
Group	Drug 1	No drug	7.0952	1.6241	7.55	4.37	0.0027	0.05	3.3109	10.8796
Group	Drug 2	No drug	4.0536	1.8811	11.3	2.15	0.0536	0.05	–0.07507	8.1822

approximate degrees of freedom for the denominator, and the Estimates contain the results corresponding to the three questions in Section 2.6, where *t*-statistics, approximate denominator degrees of freedom, and 95% confidence intervals are provided. Table 2.11 contains the estimated treatment means with their corresponding estimated standard errors. The denominator degrees of freedom are the degrees of freedom corresponding to their respective variances. The second part of Table 2.11 contains the pairwise comparisons of the treatment means including the approximate denominator degrees of freedom for each comparison. This model could be simplified by using one variance for drug 1 and both drugs and one variance for drug 2 and no drug. This can be accomplished by defining a variable, say *T*, to be 1 for drug 1 and both drugs and 0 for the other two treatments. Then place *T* in the class statement and use Repeated/Group = T; in the model specification. The estimates of the two variances are 2.4028 and 12.7376 and the AIC is 126.4, which is a smaller AIC value than that for the four variance model, indicating the two variance model is adequate to describe the data. Using a model with fewer variances in the model specification provides more degrees of freedom for the respective standard errors and thus provides more powerful tests of hypotheses concerning the fixed effects in the model.

2.9 Concluding Remarks

In summary, for comparing all means, the following are recommended:

1) If the homogeneity of variance test is not significant at the 1% level, do the usual analysis of variance test.

2) If the homogeneity of variance test is significant at the 1% level use either Welch's test or the mixed models test and the corresponding approximate denominator degrees of freedom.
3) If the homogeneity of variance is significant at the 1% level, use the AIC to determine if a simpler or fewer number of variances can be used to adequately describe the data in order to increase the power of tests concerning the means.

Many text books and articles have been written about using transformations on data in order to achieve equal treatment variances so that the usual analysis of variance can be used to compare the treatments. With the ability to fit an unequal variance model to provide estimated standard errors of means and comparisons of means, many situations will not require the use of transformations. One major benefit of not having to use a transformation to achieve equal variances is that the units of the means are in the units of measurement, thus simplifying interpretations.

This chapter contains discussion about the statistical analysis of a one-way analysis of variance model with heterogeneous errors. The discussion included several statistical tests for determining homogeneity of the error variances and recommendations on when to use each test. Procedures appropriate for making statistical inferences about the effects of different treatments upon discovering heterogeneous error variances as well as examples illustrating the use of these procedures were also reviewed.

2.10 Exercises

2.1 The following data are body temperatures of calves that were vaccinated and then challenged to determine if the vaccination protected the animal. Test the equality of variances of the treatment groups using two or more techinques. Based on the results of the test of equality of variances, test the equality of the treatment means using both Welch's and the mixed model *F*-statistics and make all pairwise comparisons.

Data for Exercise 2.1

Vaccine A	Vaccine B	Vaccine C	Vaccine D	Vaccine E	Vaccine F	Vaccine G
101.5	96.3	101.8	97.3	97.5	96.9	97.3
100.5	97.2	97.4	96.8	96.4	97.1	100.7
104.5	99.3	104.9	97.1	98.6	96.8	103.3
102.3	98.0	104.0	97.0	96.6	97.0	100.2
100.6	97.6	103.7	97.1		96.2	103.5
97.7	96.8	104.5	96.9		96.6	
	99.1	100.4	96.1			
	96.7	102.2	96.3			
	96.4	100.2	96.7			
			97.1			

2.2 Use the data in Table 1.1 and test the equality of variances using several of the methods described in Section 2.2. What is your conclusion?

2.3 The data in the following table are times required for a student to dissolve a piece of chocolate candy in their mouth. Each time represents one piece of candy dissolved by one student. Provide a detailed analysis of the data set and provide tests of the following hypotheses:

1) The mean of the Blue Choc = the mean of the Red Choc.
2) The mean of the Buttons = the mean of the means of the Blue Choc and Red Choc.
3) The mean of the ChocChip = the mean of the WchocChip.
4) The mean of the Small Choc = ½ the mean of the means of the Blue Choc and Red Choc.
5) The mean of the Blue Choc and Red Choc = the mean of the ChocChip and WchocChip.

Data for Exercise 2.3

Buttons	**Blue Choc**	**Small Choc**	**ChocChip**	**WChocChip**	**Red Choc**
69	57	28	52	35	47
76	41	27	50	37	70
59	70	28	60	38	48
55	66	30	55	40	51
68	48	29	57	34	42
34	62	28	49	35	
35		24		36	

2.4 The following data are the amount of force (kg) required to fracture a concrete beam constructed from one of three beam designs. Unequal sample sizes occurred because of problems with the pouring of the concrete into the forms for each of the designs.

1) Write out an appropriate model to describe the data and describe each component of the model.
2) Estimate the parameters of the model in part 1.
3) Use Levene's, O'Brien's, and Brown–Forsythe's methods to test the equality of the variances.
4) Use Welch's test to test H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_a: (not H_0).
5) Use the mixed model F-test to test H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_a: (not H_0).

Data for Exercise 2.4

Design	**Beam 1**	**Beam 2**	**Beam 3**	**Beam 4**	**Beam 5**	**Beam 6**	**Beam 7**	**Beam 8**	**Beam 9**	**Beam 10**
1	195	232	209	201	216	211	205			
2	231	215	230	221	218	227	218	219		
3	223	226	223	224	224	226	227	224	226	226

2.5 Four rations with different amounts of celluose were evaluated as to the amount of feed required for a chicken to gain one pound during the trial. Twenty-four chickens were randomly assigned to the four rations (six chickens per ration) and the chickens were raised in individual cages.

1) Write out an appropriate model to describe the data and describe each component of the model.

2) Estimate the parameters of the model in part 1.
3) Use Levene's, O'Brien's, and Brown–Forsythe's methods to test the equality of the variances.
4) Use Welch's test to test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_a: (not H_0).
5) Use the mixed model F test to test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_a: (not H_0).
6) Construct 90% confidence intervals about c_1, c_2, and c_3 where $c_1 = \mu_1 - \mu_2 + \mu_3 - \mu_4$, $c_2 = \mu_1 + \mu_2 - \mu_3 - \mu_4$, and $c_3 = \mu_1 - \mu_2 - \mu_3 + \mu_4$.

Data for Exercise 2.5

	Chick 1	**Chick 2**	**Chick 3**	**Chick 4**	**Chick 5**	**Chick 6**
Ration 1	2.60	2.54	2.87	2.33	2.45	2.77
Ration 2	3.87	3.18	2.59	3.62	2.71	3.08
Ration 3	2.69	5.31	2.08	4.00	3.12	4.19
Ration 4	4.43	5.59	5.06	4.17	5.17	4.47

3

Simultaneous Inference Procedures and Multiple Comparisons

Often an experimenter wants to compare several functions of the μ_i in the same experiment, leading to a multiple testing situation. Experimenters *should* consider all functions of the μ_i that are of interest; that is, they should attempt to answer all questions of interest about relationships among the treatment means. The overriding reason to include more than two treatments in an experiment or study is to be able to estimate and/or test hypotheses about several relationships among the treatment means. Often the treatments are selected to provide a structure of comparisons of interest (see, for example, the drug experiment in Section 2.4). At other times, the experimenter may be interested in comparing each treatment to all other treatments, that is, making all pairwise comparisons. This would be the case, for example, when one is comparing the yields of several varieties of wheat or for any other set of treatments that have been selected for a comparative study.

One concern when making several comparisons in a single experiment is whether significant differences observed are due to real differences or simply to making a very large number of comparisons. Making a large number of comparisons increases the chance of finding differences that appear to be significant when they are not. For example, if an experimenter conducts 25 independent tests in an experiment and finds one significant difference at the 0.05 level, she should not put too much faith in the result because, on average, she should expect to find (0.05)(25) = 1.25 significant differences just by chance alone. Thus, if an experimenter is answering a large number of questions with one experiment (which we believe one should do), it is desirable to have a procedure that indicates whether the differences might be the result of chance alone. Fisher (1949) addressed this problem when he put forward the protected least significant difference (LSD) procedure. Since then, many authors have contributed to the area of multiple testing where procedures for numerous settings have been developed.

In this chapter, several well-known and commonly used procedures for making multiple inferences are discussed and compared. Some of the procedures are primarily used for testing hypotheses, while others can also be used to obtain simultaneous confidence intervals; that is, a set of confidence intervals for a set of functions of the μ_i can be derived for which one can be 95% confident that all the confidence intervals simultaneously contain their respective functions of the μ_i.

3.1 Error Rates

One of the main ways to evaluate and compare multiple comparison procedures is to calculate error rates. If a given confidence interval does not contain the true value of the quantity being estimated, then an error occurs. Similarly, if a hypothesis test is used, an error is made whenever a true hypothesis is rejected or a false hypothesis is not rejected. Next four kinds of error rates are defined.

Definition 3.1: The comparisonwise error rate is equal to the ratio of the number of incorrect inferences made to the total number of inferences made in all experiments analyzed.

Definition 3.2: The experimentwise error rate (EER) is equal to the ratio of the number of experiments in which at least one error is made to the total number of experiments analyzed. It is the probability of making at least one error in an experiment when there are no differences between the treatments. The EER is also referred to as the experimentwise error rate under the complete null hypothesis (EERC).

Definition 3.3: The familywise error rate (FWER) (Westfall et al., 1999) is the probability of making at least one erroneous inference for a predefined set of k comparisons or confidence intervals. The set of k comparisons or confidence intervals is called the family of inferences.

Definition 3.4: The false discovery rate (FDR) (Benjamini and Hochberg, 1995) is the expected proportion of falsely rejected hypotheses among those that were rejected.

The EER controls the error rate when the null hypothesis is that all of the treatments are equally effective, that is, there are no differences among the treatment means. But many experiments involve a selected set of treatments where there are known differences among some treatments. Instead of an all means equal null hypothesis, there may be a collection of k null hypotheses, $H_{01}, H_{02}, \ldots, H_{0k}$ about the set of t means. These k null hypotheses are called partial null hypotheses and the error rate is controlled by using a method that controls the FWER (Westfall et al., 1999). For example, the set of treatments in Exercise 2.3 are six candy types, buttons, blue choc, red choc, small choc, chocChip and WchocChip. It is known at the start of the study that the time required to dissolve the small choc is much less than the time required to dissolve any of the other candies. The null question could be: Is the time it takes to dissolve a small choc equal to one-half of the mean times to dissolve the red and blue chocs? In this case a method that controls the FWER is in order since the condition of using a method that controls the EER does not hold; that is, it is known that the mean times are not all equal from the start. The FDR is very useful in the context of microarray experiments in genetics.

In order to avoid finding too many comparisons significant by chance alone in a single experiment, one quite often attempts to fix the experimentwise error rate, when applicable, or the FWER when needed at some prescribed level, such as 0.05. Whenever an experimenter is trying to answer many questions with a single experiment, it is a good strategy to control the FWER.

3.2 Recommendations

There are five basic types of multiple comparison problems: 1) comparing a set of treatments to a control or standard; 2) making all pairwise comparisons among a set of t means; 3) constructing a set of simultaneous confidence intervals or simultaneous tests of hypotheses; 4) exploratory experiments where there are numerous tests being conducted; and 5) data snooping where the comparisons are possibly data-driven. In the first four situations, the number of comparisons or confidence intervals or family of inferences is known before the data are analyzed. In the last situation, there is no set number of comparisons of interest and the final number can be very large. The recommendations given in this chapter are based on information from Westfall et al. (1999), SAS Institute, Inc. (1999), and Westfall (2002).

1) If the experiment is an exploratory or discovery study and the results are going to be used to design a follow-up or confirmatory study, then possibly no adjustment for multiplicity is necessary, thus use t-tests or unadjusted confidence intervals based on LSD values.
2) Use Dunnett's procedure for comparing a set of treatments with a control. There are two-sided and one-side versions of Dunnett's procedure, so one can select a version to fit the situation being considered.
3) For pairwise comparisons, if there is an equal number of observations per treatment group, use Tukey's method. If the data are unbalanced, then use the method that simulates (Westfall et al., 1999) a percentage point, taking into account the pattern of unequal numbers of observations.
4) If the set of linear combinations is linearly independent, then the multivariate t can be used to construct confidence intervals or to test hypotheses. If the linear combinations are uncorrelated or orthogonal, the multivariate t works well. If the linear combinations are not uncorrelated, then a simulation method that incorporates the correlation structure should be used instead of the multivariate t. Most cases with unequal numbers of observations per treatment group provide correlated linear combinations and the simulation method should be used.
5) The Bonferroni method can be used to construct simultaneous confidence intervals or tests about a selected number of linear combinations of the means, but if the number of combinations of interest is large (say 20 or more), the Scheffé procedure can often produce shorter confidence intervals, so check it out. For a set of hypotheses, the methods of Šidák (1967), Holm (1979), or Šidák–Holm can be used effectively. When the linear combinations are uncorrelated these bounds are quite good, but when there are correlations among the linear combinations, the realized FWER can be much less than desired. SAS®-MULTTEST can be used carry out bootstrap and simulated percentage points for a given set of comparisons that takes into account the correlation among the comparisons within the set.
6) For data snooping or for data-driven comparisons or hypotheses, use Scheffé's procedure as one can make as many comparisons as one wants and still control the EER or FWER.
7) For studies such as genetic studies that involve thousands of comparisons, use a method that controls the FDR, such as the method suggested by Benjamini and Hochberg (1995).

8) For studies that involve evaluating the safety of a treatment as compared with a control or placebo for possible adverse effects, use a method that does not correct for multiple tests or comparisons. Adjustment for multiplicity may not be needed for safety studies, where it is much more serious to make a type II error than it is to make a type I error.
9) Once the type of comparison is determined and the desired level of error rate control is specified, select the method satisfying these conditions that provides the smallest *p*-values or smallest critical differences or shortest confidence interval widths.

Each of the recommended multiple comparison procedures as well as a few other popular procedures available for the one-way treatment structure of Chapter 1 are examined in the following discussion. Each of the procedures can also be used in much more complex situations, as will be illustrated throughout the remainder of this book. The parameter ν used during the remainder of this book represents the degrees of freedom corresponding to the estimator of σ^2. For the one-way case of Chapter 1, the error degrees of freedom are $\nu = N - t$.

3.3 Least Significant Difference

The LSD multiple comparison method has possibly been used more than any other method, perhaps because it is one of the easiest to apply. It is usually used to compare each treatment mean with every other treatment mean, but it can be used for other comparisons as well. The LSD at the 100% significance level for comparing μ_i to μ_j is

$$\text{LSD}_\alpha = t_{\alpha/2,\nu}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} \tag{3.1}$$

One concludes that $1\mu_i \neq \mu_j$ if $|\hat{\mu}_i - \hat{\mu}_j| > \text{LSD}_\alpha$. This procedure has a comparisonwise error rate equal to α. A corresponding $(1-\alpha)100\%$ confidence interval for $\mu_i - \mu_j$ is

$$\hat{\mu}_i - \hat{\mu}_j \pm t_{\alpha/2,\nu}\hat{\sigma}\sqrt{\frac{1}{n_i}+\frac{1}{n_j}} \tag{3.2}$$

If all sample sizes are equal (to n, say), then a single LSD value can be used for all pairwise comparisons. In this case, the single LSD_α value is given by

$$\text{LSD}_\alpha = t_{\alpha/2,\nu}\,\hat{\sigma}\sqrt{\frac{2}{n}} \tag{3.3}$$

Suppose a study includes t treatment means and that all possible pairwise comparisons at the 5% significance level are going to be made. Comparisons of the comparisonwise and

TABLE 3.1

Simulated Error Rates for the LSD Procedure

Number of treatments	2	3	4	5	6	8	10	20
Comparisonwise error rate	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Experimentwise error rate	0.05	0.118	0.198	0.280	0.358	0.469	0.586	0.904

experimentwise error rates for experiments with different values of t are displayed in Table 3.1. The information in the table applies to cases where all treatment means are equal. Table 3.1 shows that, in an experiment involving six treatments, 35.8% of the time one would find at least one significant difference, even when all the treatment means were equal to one another. Obviously, using the LSD procedure could be very risky without some additional protection. When there is more than one test or parameter or linear combination of parameters of interest, meaning there is a multiplicity problem, some adjustment should be taken into account in order to eliminate discovering false results. Westfall et al. (1999) present an excellent discussion of all of the problems associated with the multiplicity problem and/or the multiple comparison problem. The following discussion attempts to describe those procedures that are useful or have been used in the analysis and interpretation of the results from designed experiments.

3.4 Fisher's LSD Procedure

Fisher's recommendation offers some protection for the LSD procedure discussed in the preceding section. In Fisher's procedure, LSD tests are made at the α100% significance level by utilizing Equation 3.1, but only if H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$ is first rejected at that level of α by the F-test discussed in Chapter 1.

This gives a rather large improvement over the straight LSD procedure since the experimentwise error rate is now approximately equal to α. However, it is possible to reject H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$ and not reject any of H_0: $\mu_i = \mu_j$ for $i \neq j$. It is also true that this procedure may not detect some differences between pairs of treatments when differences really exist. In other words, differences between a few pairs of treatments may exist, but equality of the remaining treatments may cause the F-test to be nonsignificant, and this procedure does not allow the experimenter to make individual comparisons without first obtaining a significant F-statistic. The other problem with this procedure is that many experiments contain treatments where it is known there are unequal means among some subsets of the treatments. In this case, one expects to reject the equal means hypothesis and the LSD would be used to make all pairwise comparisons. If a subset of the treatments has equal means, then more of the pairwise comparisons will detected as being significantly different than expected. Thus the FWER is not maintained. Fisher's LSD can be recommended only when the complete null hypothesis is expected to be true.

These two LSD procedures are not recommended for constructing simultaneous confidence intervals on specified contrasts of the μ_i because the resulting confidence intervals obtained will generally be too narrow.

Each of the above LSD procedures can be generalized to include several contrasts of the treatment means. The generalization is: conclude that $\Sigma_{i=1}^{t} c_i\mu_i \neq 0$ if

$$\left|\sum_{i=1}^{t} c_i\hat{\mu}_i\right| > t_{\alpha/2,v}\hat{\sigma}\sqrt{\sum_{i=1}^{t} c_i^2/n_i} \tag{3.4}$$

Examples are given in Sections 3.10, 3.12, 3.14, and 3.16.

3.5 Bonferroni's Method

Although this procedure may be the least used, it is often the best. It is particularly good when the experimenter wants to make a small number of comparisons. This procedure is recommended for planned comparisons whenever it is necessary to control the FWER. Suppose the experimenter wants to make p such comparisons. She would conclude that the qth comparison $\Sigma_{i=1}^{t} c_{iq}\mu_i \neq 0$, $q = 1, 2, \ldots, p$, if

$$\left|\sum_{i=1}^{t} c_{iq}\hat{\mu}_i\right| > t_{\alpha/2p,v}\hat{\sigma}\sqrt{\sum_{i=1}^{t} \frac{c_{iq}^2}{n_i}} \tag{3.5}$$

These p-tests will give a FWER less than or equal to α and a comparisonwise error rate equal to α/p. Usually the FWER is much less than α. Unfortunately, it is not possible to determine how much less. Values of $t_{\alpha/2p,v}$ for selected values of α, p, and v are given in the Appendix in Table A.2. For example, if $\alpha = 0.05$, $p = 5$, and $v = 24$, then from Table A.2 one gets $t_{\alpha/2p,v} = 2.80$. The examples in Sections 3.10, 3.12, 3.14, and 3.16 demonstrate the use of the Bonferroni method. The tables m is equivalent to our p.

Simultaneous confidence intervals obtained from the Bonferroni method, which can be recommended, have the form:

$$\sum_{i=1}^{t} c_{iq}\hat{\mu}_i \pm t_{\alpha/2p,v}\hat{\sigma}\sqrt{\sum_{i=1}^{t} \frac{c_{iq}^2}{n_i}}, \quad q = 1, 2, \ldots, p \tag{3.6}$$

The Bonferroni method can be applied to any set of functions of the parameters of a model, including variances as well as means.

3.6 Scheffé's Procedure

This procedure is recommended whenever the experimenter wants to make a large number of "unplanned" comparisons. Unplanned comparisons are comparisons that the experimenter had not thought of making when planning the experiment. These arise frequently, since the results of an experiment frequently suggest certain comparisons to the experimenter. This procedure can also be used when there are a large number of planned comparisons, but the widths of the confidence intervals are generally wider than for other procedures, although not always. Consider testing H_0: $\sum_{i=1}^{t} c_i\mu_i = 0$ for a given contrast vector c. It is true that

$$\Pr\left\{\frac{\left(\sum_{i=1}^{t} c_i\hat{\mu}_i - \sum_{i=1}^{t} c_i\mu_i\right)^2}{\sum_{i=1}^{t} c_i^2/n_i} \le (t-1)F_{\alpha,t-1,v}\hat{\sigma}^2 \quad \text{for all contrast vectors } c\right\} = 1 - \alpha$$

Thus a procedure with an FWER equal to α for comparing all possible contrasts of the μ_i to zero is as follows: Reject H_0: $\sum_{i=1}^{t} c_i\mu_i = 0$ if

$$\left|\sum_{i=1}^{t} c_i\hat{\mu}_i\right| > \sqrt{(t-1)F_{\alpha,t-1,v}}\,\hat{\sigma}\sqrt{\sum_{i=1}^{t} c_i^2/n_i} \tag{3.7}$$

This procedure allows one to compare an infinite number of contrasts to zero while maintaining an experimentwise error rate equal to α. However, most experimenters will usually not be interested in an infinite number of comparisons; that is, only a finite number of comparisons are of interest. Scheffé's procedure can still be used, but in this case, the FWER will generally be much smaller than α. Bonferroni's method or the multivariate t-method when appropriate will often be better (narrower confidence interval or more powerful test) than Scheffé's procedure for a finite number of comparisons. That is, a smaller value of $\sum_{i=1}^{t} c_i\hat{\mu}_i$ can often enable one to declare that $\sum_{i=1}^{t} c_i\mu_i$ is significantly different from zero using Bonferroni's method or the multivariate t-method than can be declared significant by Scheffé's method. However, if one is going to "muck around" in the data to see if anything significant turns up, then one should use Scheffé's method, since such comparisons are really unplanned comparisons rather than planned comparisons. It should be noted that Scheffé's method will not reveal any contrasts significantly different from zero unless the F-test discussed in Chapter 1 rejects H_0: $\mu_1 = \mu_2 = \cdots = \mu_t$. Scheffé's procedure can also be used to obtain simultaneous confidence intervals for contrasts of the μ_i. The result required is that, for any set of contrasts $c_1, c_2, \ldots$, one can be at least $(1-\alpha)100\%$ confident that $\sum_{i=1}^{t} c_{iq}\mu_i$ will be contained within the interval given by

$$\sum_{i=1}^{t} c_{iq}\hat{\mu}_i \pm \sqrt{(t-1)F_{\alpha,t-1,v}}\;\hat{\sigma}\sqrt{\sum_{i=1}^{t} c_{iq}^2/n_i} \quad \text{for all } q = 1,2,\ldots \tag{3.8}$$

If one wants to consider all linear combinations of the μ_i rather than just all contrasts, then $\sqrt{[(t-1)F_{\alpha,t-1,v}]}$ must be replaced by $\sqrt{[tF_{\alpha,t,v}]}$ in Equations 3.7 and 3.8.

Examples can be found in Sections 3.10, 3.12, and 3.14.

3.7 Tukey–Kramer Method

The preceding procedures can be used regardless of the values of the n_i. Tukey's (Tukey, 1952, 1953; Kramer, 1956) honest significant difference (HSD) procedure was designed to make all pairwise comparisons among a set of means. The procedure, however, requires equal n_i. Tukey (1953) and Kramer (1956) provided a modification for the case where one has unequal sample sizes. Hayter (1984) provided proof that the Tukey–Kramer method provides FWER protection, although an approximate procedure can be used if the n_i are not too unequal. The Tukey–Kramer method is to reject H_0: $\mu_i = \mu_{i'}$ for $i \neq i'$ if

$$|\hat{\mu}_i - \hat{\mu}_{i'}| > q_{\alpha,t,v}\sqrt{\frac{\hat{\sigma}^2}{2}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} \tag{3.9}$$

where $q_{\alpha,t,v}$ is the upper percentile of the distribution of the Studentized range statistic. Values of $q_{\alpha,t,v}$ for selected values of α, t, and v are given in Appendix Table A.4.

If the sample sizes are all equal to n, then the decision is to reject H_0: $\mu_i = \mu_{i'}$ for $i \neq i'$ if

$$|\hat{\mu}_i - \hat{\mu}_{i'}| > q_{\alpha,t,v}\sqrt{\frac{\hat{\sigma}^2}{n}}$$

Tukey's general procedure for equal sample sizes is to reject H_0: $\sum_{i=1}^{t} c_i\mu_i = 0$ for a contrast if

$$\left|\sum_{i=1}^{t} c_i\hat{\mu}_i\right| > q_{\alpha,t,v}\frac{\hat{\sigma}}{\sqrt{n}}\left(\frac{1}{2}\sum_{i=1}^{t}|c_i|\right)$$

3.8 Simulation Methods

For unequal sample size problems, for problems where the comparisons are other than pairwise comparisons, and for problems where the comparisons are not linearly independent, the above methods provide FWER significance levels that are less than desired. In this case, the percentage points for the appropriate set of comparisons can be simulated.

Suppose you are interested in p linear combinations of the μ_i such as $\sum_{i=1}^{t} c_{iq}\mu_i$, $q = 1, 2, \ldots, p$ and it is desired to provide a procedure that controls the FWER for either the set of hypotheses H_0: $\sum_{i=1}^{t} c_{iq}\mu_i = 0$, $q = 1, 2, \ldots, p$ or a set of simultaneous confidence intervals for $\sum_{i=1}^{t} c_{iq}\mu_i$. The process is:

1) Generate a sample of data in the same structure of the data set at hand. If there are five treatments with sample sizes, 5, 9, 3, 6, and 7, generate data with those sample sizes.
2) Carry out the analysis of the generated data set as is to be done with the actual data set and compute the p t-statistics:

$$t_q = \frac{\sum_{i=1}^{t} c_{iq}\hat{\mu}_i}{\sqrt{\hat{\sigma}^2\sum_{i=1}^{t} c_{iq}^2/n_i}} \qquad q = 1, 2, \ldots, p$$

3) Compute the maximum of the absolute values of the t_q, T_s= max($|t_1|, |t_2|, \ldots, |t_p|$).
4) Repeat steps 1, 2 and 3 a very large number of times, keeping track of the computed values of T_s. Determine the upper α100 percentile of the distribution of the T_s, and denote this percentile by T_α.
5) For the actual data set, compute t_q, $q = 1, 2, \ldots, p$ and reject the qth hypothesis if $|t_q| > T$, $q = 1, 2, \ldots, p$ or construct simultaneous confidence intervals as

$$\sum_{i=1}^{t} c_{iq}\hat{\mu}_i \pm T_\alpha \sqrt{\hat{\sigma}^2 \sum_{i=1}^{t} c_{iq}^2/n_i}, \quad q = 1,2,\ldots,p$$

The accuracy of the simulation can be specified by using the method of Edwards and Berry (1987). SAS-MULTTEST can be used to obtain simultaneous inferences using the bootstrap method (Westfall et al., 1999). Bootstrap methodology is beyond the scope of this book.

3.9 Šidák Procedure

Šidák (1967) provided a modification of the Bonferroni method by using a different percentage point for each of the comparisons. The process is to compute a *t*-statistic for each of the comparisons:

$$t_q = \frac{\sum_{i=1}^{t} c_{iq}\hat{\mu}_i}{\sqrt{\hat{\sigma}^2 \sum_{i=1}^{t} c_{iq}^2/n_i}}, \quad q = 1,2,\ldots,p$$

Compute the significance level for each comparison and order the significance levels from smallest to largest as $p_1, p_2, \ldots, p_p$. For a FWER of α, reject the individual comparison if $p_q \le 1 - (1 - \alpha)^{1/p}$ or equivalently if $\alpha \ge 1 - (1 - p_q)^p$.

3.10 Example—Pairwise Comparisons

The task data in Section 1.6 is used to demonstrate the results of the above multiple comparisons procedures. Table 3.2 contains the SAS-Mixed code to fit the one-way means model and the LSMeans statements are used to extract several of the multiple comparison procedures. Table 3.3 contains the percentage points used to provide confidence differences or significant differences for the simulate, Tukey–Kramer, Bonferroni, Šidák, Scheffé, and *t* (unadjusted) methods. Excluding the unadjusted *t*, the other methods provide 0.05 type I FWER for all pairwise comparisons. The simulate and Tukey–Kramer methods use the smallest quantiles with the Šidák and Bonferroni methods in the middle, while the Scheffé method is largest. Table 3.4 contains the critical significant differences for each of

TABLE 3.2

SAS System Code Using Proc Mixed to Request the Computation of Several Multiple Comparisons Procedures for All Pairwise Comparisons

```
PROC mixed DATA=EX1; CLASS TASK;
MODEL PULSE20=TASK/NOINT SOLUTION;
LSMEANS TASK/ DIFF CL;
LSMEANS TASK/ DIFF ADJUST=TUKEY CL;
LSMEANS TASK/ DIFF ADJUST=BON CL;
LSMEANS TASK/ DIFF ADJUST=SCHEFFE CL;
LSMEANS TASK/ DIFF ADJUST=SIDAK CL;
LSMEANS TASK/ DIFF ADJUST=SIMULATE (REPORT SEED=4938371) CL;
```

TABLE 3.3

Percentage Points Used for All Pairwise Comparisons of the Six Task Means

Simulation Results

Method	95% Quantile	Estimated α	99% Confidence Limits	
Simulate	2.932480	0.0500	0.0450	0.0550
Tukey–Kramer	2.940710	0.0486	0.0436	0.0535
Bonferroni	3.053188	0.0359	0.0316	0.0401
Šidák	3.044940	0.0370	0.0326	0.0413
Šcheffé	3.437389	0.0131	0.0105	0.0157
t	1.998972	0.3556	0.3446	0.3666

TABLE 3.4

Critical Differences Used to Compare the Differences between Pairs of Means for the Unadjusted *t* and Several Multiple Comparison Procedures

TASK	_TASK	Estimate	Standard Error	*t*	Bonferroni	Tukey–Kramer	Scheffé	Šidák	Simulate
1	2	0.840	2.225	4.449	6.795	6.544	7.650	6.776	6.526
1	3	−3.877	2.338	4.674	7.139	6.876	8.038	7.120	6.857
1	4	−6.077	2.338	4.674	7.139	6.876	8.038	7.120	6.857
1	5	2.423	2.225	4.449	6.795	6.544	7.650	6.776	6.526
1	6	3.105	2.277	4.553	6.953	6.697	7.828	6.935	6.679
2	3	−4.717	2.380	4.758	7.267	7.000	8.182	7.248	6.980
2	4	−6.917	2.380	4.758	7.267	7.000	8.182	7.248	6.980
2	5	1.583	2.270	4.537	6.929	6.674	7.801	6.911	6.655
2	6	2.265	2.321	4.639	7.085	6.824	7.977	7.066	6.805
3	4	−2.200	2.486	4.970	7.591	7.311	8.546	7.570	7.291
3	5	6.300	2.380	4.758	7.267	7.000	8.182	7.248	6.980
3	6	6.982	2.429	4.855	7.416	7.143	8.349	7.396	7.123
4	5	8.500	2.380	4.758	7.267	7.000	8.182	7.248	6.980
4	6	9.182	2.429	4.855	7.416	7.143	8.349	7.396	7.123
5	6	0.682	2.321	4.639	7.085	6.824	7.977	7.066	6.805

TABLE 3.5

Adjusted Significance Levels to Test the Equality of All Pairwise Comparisons of TASK Minus _TASK Obtained from Six Procedures Where *t* Corresponds to the Unadjusted *t*

TASK	_TASK	*t*	Bonferroni	Tukey–Kramer	Scheffé	Šidák	Simulate
1	2	0.7072	1.0000	0.9990	0.9996	1.0000	0.9990
1	3	0.1024	1.0000	0.5642	0.7378	0.8021	0.5646
1	4	0.0117	0.1751	0.1129	0.2552	0.1615	0.1111
1	5	0.2805	1.0000	0.8840	0.9446	0.9928	0.8804
1	6	0.1777	1.0000	0.7484	0.8661	0.9469	0.7501
2	3	0.0520	0.7795	0.3645	0.5642	0.5509	0.3657
2	4	0.0051	0.0761	0.0546	0.1506	0.0735	0.0545
2	5	0.4880	1.0000	0.9815	0.9923	1.0000	0.9813
2	6	0.3328	1.0000	0.9238	0.9651	0.9977	0.9234
3	4	0.3796	1.0000	0.9488	0.9772	0.9992	0.9474
3	5	0.0103	0.1543	0.1014	0.2364	0.1437	0.0985
3	6	0.0055	0.0831	0.0590	0.1596	0.0799	0.0584
4	5	0.0007	0.0104	0.0087	0.0366	0.0104	0.0090
4	6	0.0004	0.0053	0.0046	0.0219	0.0053	0.0052
5	6	0.7699	1.0000	0.9997	0.9999	1.0000	0.9998

the pairwise comparisons. The observed differences for task 1 to task 4, task 2 to task 4, task 3 to task 4, task 3 to task 5, task 3 to task 6, task 4 to task 5 and task 4 to task 6 all exceed the critical differences for the *t* or LSD, which controls the comparisonwise error rate, but not the experimentwise error rate. Only the comparisons of task 4 to task 5 and task 4 to task 6 exceed the critical differences for the other five methods, all of which provide experiment wise error rate protection. The magnitudes of the critical differences are smallest for the uncorrected *t* or LSD method. The simulate and Tukey–Kramer critical differences are similar in magnitude while the simulate values are a little smaller. The Šidák and Bonferroni differences are similar in magnitude, with the Šidák values slightly smaller. The Scheffé critical differences are largest, as is expected since they control the FWER for an infinite number of comparisons and only 15 pairwise comparisons are made. A set of simultaneous confidence intervals about all pairwise comparisons can be constructed by adding and subtracting the critical difference from the estimated difference. For example, the simultaneous 95% confidence interval about $\mu_1 – \mu_2$ using the simulate method is 0.840 ± 6.526. Table 3.5 contains the adjusted *p*-values for each of the methods. The *p*-values provide the same decision as the 5% critical differences in Table 3.4.

3.11 Dunnett's Procedure

One really interesting case is that of comparing all treatments with a control. This type of inference is important in safety studies, where it is of interest to compare different doses of a treatment with the control or placebo. Dunnett's test is to declare a treatment mean μ_i to be significantly different from the mean of the control μ_0 if

$$|\hat{\mu}_i - \hat{\mu}_0| > d_{\alpha,t,v}\sqrt{\hat{\sigma}^2\left(\frac{1}{n_i} + \frac{1}{n_0}\right)}$$

where $d_{\alpha,t,\nu}$ is the upper $\alpha 100$ percentile of the "many-to-one *t*-statistic" (Miller, 1967). Dunnett's method controls the FWER. If the sample sizes are unequal, a simulate procedure can take into account the sample size structure and possibly provide a shorter bound.

3.12 Example—Comparing with a Control

The task data in Section 1.6 is used to demonstrate the process of comparing each treatment with a control. In this study, assume that task 2 is the control task and the other five tasks are the experimental tasks. Table 3.6 contains the SAS-Mixed code to use the unadjusted *t*, Bonferroni, Dunnett, Scheffé, Šidák, and Simulate methods to compare all of the other tasks with task 2. The option on the LSMean statement DIFF=CONTROL('2') requests that task 2 be considered as the control and is compared with each of the other tasks in the study. Table 3.7 contains the 95% quantiles for each of the methods. The Dunnett quantile is less than the others (except for the unadjusted *t*) with the simulate method very close. There are five comparisons being made, which dictates the magnitude of the Bonferroni and Šidák quantiles. The Scheffé quantile is the same as in Table 3.4, which is useful for an infinite number of comparisons. The only comparison where the observed difference exceeds the critical difference is for comparing task 4 to the control or task 2. A set of simultaneous confidence intervals about all differences between the treatment and control means can be constructed by adding and subtracting the critical difference in Table 3.8

TABLE 3.6

SAS System Code Using Proc Mixed to Request the Computation of Several Multiple Comparisons Procedures for Comparing Each Task to the Means of Task 2 (Control)

```
PROC mixed DATA=EX1; CLASS TASK;
MODEL PULSE20=TASK/NOINT;
LSMEANS TASK/ DIFF=CONTROL('2') CL;
LSMEANS TASK/ DIFF=CONTROL('2') ADJUST=BON CL;
LSMEANS TASK/ DIFF=CONTROL('2') ADJUST=DUNNETT CL;
LSMEANS TASK/ DIFF=CONTROL('2') ADJUST=SIDAK CL;
LSMEANS TASK/ DIFF=CONTROL('2') ADJUST=SIMULATE (REPORT SEED=4938371) CL;
LSMEANS TASK/ DIFF=CONTROL('2') ADJUST=scheffe CL;
```

TABLE 3.7

Percentage Points Used for Comparing Each Task Mean to the Mean of Task 2 (Control)

Simulation Results

Method	95% Quantile	Exact α
Simulate	2.590707	0.0494
Dunnett, two-sided	2.585505	0.0500
Bonferroni	2.657479	0.0418
Šidák	2.649790	0.0427
Scheffé	3.437389	0.0048
t	1.998972	0.1831