CONTENTS

CHAPTER 1

GENERALIZED INFERENCE IN REPEATED MEASURES: EXACT METHODS FOR MANOVA AND MIXED MODELS

PREFACE

This book presents some recent developments and classical methods in Repeated Measures involving Mixed Models, Multivariate Analysis of Variance (MANOVA), and Growth Curves, in particular. Repeated Measures models and Growth Curves models are in fact special classes of mixed models. A substantial fraction of applications in Biomedical data analysis, many clinical trials, and industrial experiments involve observations taken from experimental units at a number of time points. That is, they involve repeated measurements taken over time. In fact, even in such applications as marketing research, market analysts need to deal with repeated measurements. For example, after a price increase of a product, market analysts do not just analyze the price effect at one time point, but rather follow the change in consumer demand of the product over time so that corrective action can be taken if necessary. Data from such experiments as well as from clinical trials can be analyzed by alternative methods, especially by methods in MANOVA, Repeated Measures, and Growth Curves, depending on the assumptions deemed appropriate and the objectives of the study. Perhaps due to lack of knowledge of these methods, sometimes practitioners resort to simpler methods that do not take the full advantage of the data. It seems that this is mostly due to inadequate college courses that cover methods in Analysis of Repeated Measurements. Hence, it is the belief of many experts in the area that there is a

great need to promote the extent of teaching, research, and practice of specialized methods for analysis of repeated measures. This book is intended to contribute in meeting that need.

As demonstrated in almost all chapters of this book, the classical approach to solving these problems provides exact solutions to only a fraction of the problems in this context. Conventional methods alone do not always provide exact solutions to even some simple problems. For example, in the univariate Analysis of Variance, the classical approach fails to provide exact tests when the underlying population variances are unequal. In some widely used Growth Curves models, there are no exact classical tests even in the case equal variances. As a result, practitioners often resort to asymptotic results in search of approximate solutions even when such approximations are known to perform rather poorly with typical sample sizes. In view of this fact, the book starts out with an introduction to the generalized inference approach that can help tackle such problems. In particular, an introduction to the notions of generalized p-values and generalized confidence intervals and related methods are presented in Chapter 1. Then, these techniques are applied in each problem undertaken in the book. It should be emphasized that a practitioner can utilize results obtained by applying such new notions and concepts with or without deviating from the conventional philosophy of statistical inference.

Solutions to the statistical problems addressed in this book are presented as extensions, as opposed to alternatives, to conventional methods of statistical inference. In fact, each class of problems is started with a simple model under special assumptions that are necessary for the classical approach to work. After discussing solutions available for such special cases, we will relax such assumptions when they are considered to be too restrictive or unreasonable in some applications, especially when they are known to have poor size (Type I error) performance or poor power performance. For example, in fixed effects ANOVA, the problem is first considered under the homoscedastic variance/covariance assumption and then later we drop the assumption which is usually made for simplicity and mathematical tractability rather than anything else. According to simulation studies, the equal variances assumption has been found to be much more serious than the assumption of normality itself. When the assumption of homoscedasticity is not reasonable, conventional F-tests frequently lead to wrong conclusions and often fail to detect significant experimental results. In higher-way ANOVA, the drawback becomes even more serious in that one can even come to the opposite conclusion in detecting significant effects. Such lack of power of tests and erroneous conclusions can result in serious repercussions in practical applications, especially in biomedical research. In the case of mixed models, some widely used methods such as MLE-based tests and confidence intervals on variance components are now known to have very serious size problems. In each application, where variance components are encountered, this book will present inference procedures that do not suffer from such false-positive problems.

This book is suitable as a textbook on Analysis of Repeated Measures or as a reference book. The book is useful to teachers, researchers, and experimenters such as those in the agricultural and biomedical fields. It is also useful to industrial statisticians and to many other practicing statisticians including those in Market and Business Research. In the opinion of some researchers and practitioners of Repeated Measures (RM), many articles and books on repeated measures and growth curves are difficult to read. Sometimes, even with extensive formulas and equations found in such literature, it is difficult to figure out the correct formula for setting up a simple ANOVA table in RM, perhaps due to experts' assumption that readers are already familiar with RM ANOVA. Therefore, in view of the fact that courses in RM are yet to be taught at many colleges, I have taken special care and effort to try to make this book easy to read, concise, and yet self-contained with essential formulas. Some of the formulas are presented without formal arguments and derivations, especially when they distract the presentation of an intuitive result while adding little value to the understanding of the material. Nevertheless, some references to the literature providing proofs of such results are provided. The mathematical level of the book is kept somewhat low, and some prior knowledge of statistical notions in the analysis of linear models is assumed. More emphasis is placed on concepts, methods, and applications rather than on mathematical rigor. For the benefit of those readers who might be interested in formal proofs of theorems and further details of results, a fairly comprehensive set of references is provided.

As a special feature of this book, exact statistical methods are provided for each application considered in the book, including the problem of comparing a number of normal populations. The methods are exact in the sense that the tests and the confidence intervals developed in the book are based on exact probability statements rather than on asymptotic approximations. This means that inferences based on them can be made with any desired accuracy, provided that assumed parametric model and/or other assumptions are correct. To make this possible, solutions to problems of testing various hypotheses considered in this book are presented in terms of p-values. There is readily available computer software to implement these exact statistical methods.

All practitioners and researchers of statistical methods can benefit from parametric exact methods presented in this book, irrespective of their philosophy and belief, because they are developed as extensions to the classical methods as opposed to alternatives, thus providing solutions to a wider class of problems. Practitioners who prefer to carry out tests without new concepts and notions should also find all of the methods presented in the book useful. This is because exact p-values and confidence intervals obtained with extended definitions also serve to provide excellent approximate solutions in the classical sense. According to simulation studies reported in the literature, Type I error and power performance of these approximations are usually much better than the performance of more complicated approximate tests obtained by

other means. Moreover, unlike some approximations such as those available on functions of variance components, the approach that one needs to take in deriving a solution does not change from one situation to another in a class of similar problems. Therefore, exact significance tests and generalized confidence intervals reported in this book are useful to any practitioner regardless of whether or not he/she insists on classical fixed level tests and conventional confidence intervals.

The book is written with all potential readers in mind. For their benefit, a fairly comprehensive set of formulas is given for all important results presented in each chapter. A large number of illustrative examples and exercises is provided for the benefit of teachers and students. They range from simple numerical examples to extended versions of methods presented in each chapter. The exercises given at the end of each chapter are intended to further illustrate the concepts and methods covered in the chapter. Some of the exercises are intended not only to stimulate further thoughts on the material covered in the book, but also to stimulate much needed further research in the context of generalized inference to solve inference problems in more complicated repeated measures designs that are not addressed in the book.

The first chapter of the book is devoted to present some new notions and concepts in statistical inference, in particular, generalized p-values. Exact statistical methods based on such concepts as well as conventional ones are then presented in rest of the chapters covering the topics of the book. Chapters 2−4 provide an introduction to ANOVA and Mixed Models, which are needed in later chapters. Chapters 5 and 6 extend results presented in Chapter 3 to the multivariate case. Chapters 7−10 address some basic models widely used in the Analysis of Repeated Measures and provide solutions under alternative assumptions and design matrices. A large number of numerical examples are provided in each chapter to illustrate applications of methods based on the normal theory. The computations of the major exact parametric methods presented in this book can be easily performed with the XPro statistical software package, which specializes in Exact Parametric Inference. Most p-values and generalized confidence intervals involve some simple numerical integrations. So, with some coding to implement the new formulas, the inferences can also be carried out with widely used major statistical software packages such as SAS, SPSS, and SPlus. In fact, one can use any statistical package that provides capabilities to generate simulated samples from normal and chi-squared distributions so that the numerical integrations involved in the computation of p-values could be performed by Monte Carlo methods.

I would like to thank the editorial staff at Wiley for their cooperation, patience, and support given during the course of this project.

I am very grateful to a number of people who read the manuscript and made useful comments and suggestions. I wish to thank and extend my appreciation to M. M. Ananda, Y. Ho, K. G. Jinadasa, J. Lee, S. Lin, T. Mathew, and H. Weerahandi for reading the manuscript. I extend my gratitude also to S. Dalal, R. A. Johnson, J. Kettenring, M. Koschat, T. Spacek, K. W. Tsui, and J. V. Zidek for their help, guidance, and support provided to me at various occasions.

My special thanks are extended to Hongying Mo for proof reading the whole manuscript and for the support given to me throughout the preparation of the manuscript.

Samaradasa Weerahandi
Edison, New Jersey
May 2004

CHAPTER 2

EXACT GENERALIZED INFERENCE

2.1 INTRODUCTION

In this book, by exact generalized inference we simply mean various procedures of hypothesis testing and confidence intervals that are based on exact probability statements. Here we confine our attention to the problems of making inferences concerning parametric linear models with normally distributed error terms. In particular, this book does not address exact nonparametric methods that are discussed, for instance in Good (1994) and Weerahandi (1995). The purpose of this chapter is to provide a brief introduction to the notions and methods in generalized inference that enable one to obtain parametric analytical methods that are based on exact probability statements.

There is a wide class of problems for which classical fixed-level tests based on sufficient statistics do not exist. For specific examples of simple problems in which conventional fixed-level tests do not exist, the reader is referred to Chapter 5 of Weerahandi (1995). Actually, this is the case even with widely used linear models. For example, in the problem of comparing the means of two or more normal populations, exact fixed-level tests and conventional confidence intervals based on sufficient statistics are available only when the

(Generalized Inference in Repeated Measures, Edition 2). By (Weerahandi)

population variances are equal or when some additional information is available about the variances. The situation only gets worse in more complicated problems such as the two-way ANOVA, the MANOVA, Mixed Models, and in Repeated Measures models including Crossover Experiments and Growth Curves that we will address in the following chapters. In each of these models, exact conventional tests and confidence intervals are available only in special cases. The limited availability of fixed-level tests is a very serious problem. For example, widely used classical F-tests used in linear models sometimes fail to detect significant differences in treatments being compared even when the available data provide sufficient evidence to do so. In applications such as those in biomedical experiments, this drawback of classical F-tests could substantially delay the time of getting a good drug into the market and incur substantially greater research cost. Application of the classical F-test in such an application could even result in discontinuation of good research due to erroneous conclusions that treatments being tested are not effective. In this chapter we will extend the classical definition of p-values and confidence intervals so that one could obtain testing procedures in wide class of applications including a variety of linear models addressed in this book. Unlike the approach in Fraser (1979), we shall do this extension without affecting the interpretations of classical tests when they do exist.

Kempthorne and Folks (1971) indicated how tests of significance could be obtained in situations where the classical approach fails, but did not give explicit definitions. Without formal definitions and derivations, Bernard (1984) and Rice and Gaines (1989) gave formulae for computing exact p-values for the Behrens−Fisher type tests. As it will become clear later, unconventional p-values reported in these articles are in fact generalized p-values. In the application of comparing two regression models, Weerahandi (1987) gave the first introduction to the notion of generalized $p-$value and showed that it is an exact probability of an unbiased extreme region, a well-defined subset of the sample space formed by sufficient statistics. Motivated by that application, Tsui and Weerahandi (1989) provided formal definitions and methods of deriving generalized p-values. In a Bayesian treatment, Meng (1994) introduced a Bayesian p-value (posterior predictive p-value) which is, under the noninformative prior, numerically equivalent to the generalized p-value. Weerahandi and Tsui (1996) showed how Bayesian p-values could be obtained for ANOVA type problems, which are numerically equivalent to the generalized p-values.

As discussed at length in Weerahandi (1995), exact probability statements are not necessarily related to the classical repeated sampling properties. In special cases the former may have such implications on the latter, but this is not something that one should take for granted. For example, in applications involving discrete distributions, often one can compute exact $p-$values, but not exact fixed-level tests. Rejecting a hypothesis based on such p-values, say at the 5% level if $p < 0.05$, does not imply that the false positive rate in repeated sampling is 5%. Simply, such a p-value is a measure of false positive error and hence one can indeed rejects the null hypothesis when it is

less than a certain threshold such as the .05 level. Nevertheless, in most practical applications, fixed-level tests based on p-values, including the generalized p-values we discuss below, do provide excellent approximate fixed-level tests that are better than asymptotic tests. In fact, according to a number of simulation studies reported in the literature [cf. Gamage and Weerahandi (1998), and Park and Burdick (2004)], generalized tests based on exact probability statements tend to outperform, in terms of Type I error or power, more complicated approximate tests. Moreover, in many situations, Type I error of generalized tests do not exceed the intended level. Therefore, procedures based on probability statements, that are exact for any sample size, are useful for all practitioners, regardless of they insist on repeated sampling properties or not. Also the practitioners and researchers who insist on classical procedures, and anyone who has difficulties with the meaning of exactness, can just consider the generalized approach as a way of finding good approximate tests and confidence intervals, which are expected to perform better than asymptotic methods. In summary, all practitioners and researchers of statistical methods can benefit from the generalized approach to statistical inference, irrespective of their philosophy and belief, because it is an extension to the classical approach to inference as opposed to an alternative, providing solutions to a wider class of problems. Therefore, obviously the best choice of methods available from the extended class is as good as or better than that of the original class.

Inferences on discrete and categorical variable models, nonlinear models, and models based on non-normal distributions are beyond the scope of this book. In particular, the readers interested in nonparametric exact methods based on permutation and randomization tests are referred to Mehta and Patel (1983), Agresti (1990), Good (1994), Weerahandi (1995), and Berger (2000). For applications of generalized inference in non-normal distributions, the reader is referred to Ananda and Weerahandi (1996), Ananda (1999), Krishnamoorthy and Mathew (2003), and Mathew and Roy (2004). For nonlinear models in repeated measures the reader is referred to Davidian and Giltinan (1995) and Vonesh and Chinchilli (1997).

2.2 TEST STATISTICS AND P-VALUES

Classical p-values as well as testing at a fixed nominal level, such as the 0.05 level, are based on what is known as test statistics. Basically, a test statistic is a statistic with some special properties, a function of some observable data set from an experiment. The function should not depend on any unknown parameters to qualify to be a test statistic. In the classical approach to testing of hypotheses, this is an important requirement because, given a set of data, we should be able to compute such a statistic and compare against some threshold or a critical value, which used to be available from most textbooks in Statistics in the form of percentiles of widely used statistical distributions.

To give a formal definition of test statistics, consider an observable random vector $\mathbf{Y}$ representing a certain population. The type of the distribution of $\mathbf{Y}$ is assumed to be known except for certain unknown parameters, say $\boldsymbol{\zeta} = (\theta, \boldsymbol{\delta})$ where θ is a parameter of interest and $\boldsymbol{\delta}$ is a vector of other parameters, which are sometimes called nuisance parameters. Let Ψ be the sample space of possible values of $\mathbf{Y}$, and let Θ be the parameter space of θ. The observed value of the random vector $\mathbf{Y}$ is denoted by $\mathbf{y}$. Consider the problem of testing the hypotheses

$$\theta \in \Theta_0 \text{ versus } \theta \in \Theta_1, \tag{2.1}$$

where Θ_0 and Θ_1 are two disjoint subsets of the parameter space Θ.

Definition 1.1. A *test statistic* is a real-valued function of $\mathbf{y}$ and a prespecified value θ_0 of θ, of the form $T(\mathbf{y}; \boldsymbol{\theta}_0)$ satisfying the following two properties:

1. The distribution of $T = T(\mathbf{y}; \boldsymbol{\theta}_0)$ does not depend on the nuisance parameters $\boldsymbol{\delta}$.

2. T is stochastically monotonic (increasing or decreasing) in θ, that is, the cumulative distribution function (abbreviated as cdf) of T is a monotonic function of θ.

In many applications Property 2 above is too restrictive or not well defined. In such situations, a statistic satisfying the following milder condition is considered acceptable to be qualified as a test statistic:

2′. T satisfies the condition that

$$\Pr(T \leq t | \theta \in \Theta_0) \leq \Pr(T \leq t | \theta \in \Theta_1) \text{ for all } t. \tag{2.2}$$

In a given application, there may be multiple test statistics satisfying above conditions. Then one can attempt to find a unique test statistic by imposing certain optimality criteria. Often one can find unique test statistics by requiring the test statistic to be based on minimal sufficient statistics and to be invariant with respect to certain transformations of the parameters and the statistics. The reader is referred to Lehmann (1986) and Weerahandi (1995) for formal definitions and details on these concepts.

Definition 1.2. A subset of the sample space $C(\mathbf{Y}\,; \mathbf{y})$ is said to be an *extreme region* if

1. the observed value $\mathbf{y}$ of $\mathbf{Y}$ falls on the boundary of the region,

2. $C(\mathbf{Y}\,; \mathbf{y})$ does not depend on nuisance parameters $\boldsymbol{\delta}$,

3. its probability does not depend on unknown parameters $\boldsymbol{\zeta}$ when θ has been specified,

4. its probability increases for deviations from the null hypothesis; that is,

$$\Pr(C(\mathbf{Y}\,;\mathbf{y})\,|\,\theta \in \Theta_0 \in \Theta_0) \leq \Pr(C(\mathbf{Y}\,;\mathbf{y})\,|\,\theta \in \Theta_1)\,. \tag{2.3}$$

Definition 1.3. The *p-value* of a test based on an extreme region C_y is defined as

$$p = \text{Sup}\{\Pr(C(\mathbf{Y}\,;\mathbf{y})\,|\theta \in \Theta_0)\}. \tag{2.4}$$

Of particular interest is the problem of testing hypotheses of the form

$$H_0 :\ \theta \leq \theta_0 \text{ versus } H_1 :\ \theta > \theta_0. \tag{2.5}$$

If the test statistic T being used is stochastically increasing in θ, then the p-value can be conveniently computed using the formula

$$p = \Pr(T \geq t_{obs}|\theta = \theta_0), \tag{2.6}$$

a result that follows from above definition, where t_{obs} is the observed value of the test statistic. Similarly the p-value for testing hypotheses of the form $H_0 :\ \theta \geq \theta_0$ versus $H_1 :\ \theta < \theta_0$ is computed using the formula $p = \Pr(T < t_{obs}|\theta = \theta_0)$. For other related definitions, such as that of power functions, and for some useful theorems in this context, the reader is referred to Weerahandi (1995). Also of interest, especially in problems involving multiple location parameters including a variety of ANOVA problems that we will study in the following chapters, are hypothesis problems of the form

$$H_0 :\ \theta = \theta_0 \text{ versus } H_1 :\ \theta \neq \theta_0. \tag{2.7}$$

The literature on p-values provides alternative definitions of p-values for this case except in special cases. Especially in this situation Property 2´ given above is considered adequate. In this case the p-value is computed as

$$p = \Pr(C(\mathbf{Y}\,;\mathbf{y})|\theta = \theta_0) \tag{2.8}$$

$$= \Pr(T \in C_{obs}|\theta = \theta_0), \tag{2.9}$$

where C_t is a subset of the sample space, an extreme region, defined by $t = T(\mathbf{y})$.

The p-values serve to measure the evidence in favor or against the null hypothesis. The smaller the p-value, the greater the evidence against the null hypothesis. Of course one can reject the null hypothesis when the p-value falls below a certain threshold, which is sometimes known as the critical point in fixed-level testing of hypotheses.

In search of testing procedures with certain optimum properties, one can confine the search to extreme regions that are based on minimal sufficient statistics. In fact, test statistics based on sufficient statistics provide a convenient way of constructing extreme regions and p-values.

Example 1.1. Testing the normal mean

Let $Y_1, \ldots, Y_n$ be a random sample from a normal population with the distribution

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\mu)^2},$$

where μ and σ^2 are the mean and the variance of the population. Let $\overline{Y}$ and S^2 be the unbiased estimators of μ and σ^2, respectively, where

$$\overline{Y} = \frac{\sum Y_i}{n} \text{ and } S^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-1}.$$

Consider the problem of testing the hypotheses

$$H_0 : \ \mu \leq \mu_0, \quad H_1 : \ \mu > \mu_0 \tag{2.10}$$

It is known from the theory of sampling from a normal distribution that

$$Z \ = \ \frac{\overline{Y} \ - \ \mu}{\sigma/\sqrt{n}} \ \sim N(0,1) \text{ and } U = \frac{(n-1)S^2}{\sigma^2} \ \sim \chi^2_{n-1}$$

and that they are independently distributed. In view of these distributional results, we can define a potential test statistic as

$$T \ = \ \frac{\overline{Y} \ - \ \mu_0}{S/\sqrt{n}} \ .$$

When $\mu = \mu_0$, the random variable T has a t-distribution with $n-1$ degrees of freedom. Otherwise it has a noncentral t-distribution, which is stochastically increasing in μ. This fact is also seen from the expression

$$\Pr(T \geq t) = \Pr(\frac{\overline{Y} \ - \ \mu}{S/\sqrt{n}} + \frac{\mu - \ \mu_0}{S/\sqrt{n}} \geq t),$$

because

$$\frac{\overline{Y} \ - \ \mu}{S/\sqrt{n}} \sim t_{n-1}$$

is free of unknown parameters. Hence, T is indeed a test statistic and therefore we can proceed to compute the p-value using (2.6) as

$$\begin{aligned} p &= \Pr(T \geq \frac{\overline{y} \ - \ \mu_0}{s/\sqrt{n}} | \mu = \mu_0) \\ &= 1 - G_{n-1}(\frac{\overline{y} \ - \ \mu_0}{s/\sqrt{n}}), \end{aligned}$$

where G_{n-1} is the cumulative distribution function (cdf) of the Student's t-distribution with $n-1$ degrees of freedom. The null hypothesis is rejected for smaller values of the p-value.

2.3 TEST VARIABLES AND GENERALIZED P-VALUES

Test statistics discussed in the previous section provides a convenient way of constructing extreme regions, on which p-values and tests can be based. But, as discussed extensively in Weerahandi (1995), that approach works only in a very limited set of applications. For example in the problem of sampling from a normal population as in Example 1.1, it is not clear how a test statistic could be constructed if the parameter of interest were a function such as, $\theta = \mu + \sigma^2$. The Behrens−Fisher problem is a well-known example of a situation where a test statistic based on sufficient statistics does not exist. This limitation extends well into all types of linear models including ANOVA, regression models, and all types of repeated measures problems. The limitation is caused by the fact that the test statistic is required to be a single quantity regardless of the number of minimal sufficient statistics available in a statistical problem. This was a requirement even in multivariate problems, perhaps because during the times of limited computing facilities, we had to rely on tabulated critical points with which we can compare the observed values of such test statistics and draw conclusions about the validity of hypotheses. With readily available statistical software packages and ever-increasing computing power, today we no longer need to rely on test statistics and tabulated critical values.

In the spirit of Fisher's original treatment [cf. Fisher (1956)] of hypothesis testing, what is basically needed to device a testing procedure is to specify a subset of the sample space, which can be considered as an extreme region if the null hypothesis is true. It is not necessary that such an extreme region is represented by a test statistic. Yet, as before we can insist that, in order to be an extreme region, the subset must be unbiased and should contain the observed sample point on its boundary. We can search for extreme regions with desired properties within the sample space formed by minimal sufficient statistics, regardless of whether or not they can be expressed in terms of a single test statistic. Nevertheless, search for such subsets of sample space can be facilitated by what is known as *test variables* in the context of generalized inference. The notion of test variables was formerly introduced by Tsui and Weerahandi (1989). Although it is possible to define extreme regions based on a number of statistics (preferably minimal sufficient statistics) without invoking the notion of test variables, as done by Weerahandi (1987), test variables do provide a convenient way of defining extreme regions as they play the role of test statistics in the generalized setting. In particular, often extreme regions can be defined using a single test variable just as in the case of a test statistic. In early simulation studies, Thursby (1992) and Griffiths and Judge (1992) found the generalized p-value given by Weerahandi (1987) for the problem of comparing regressions to have good size and power performance. Performance of generalized p-values in a number of other applications are provided by Zhou and Mathew (1994), Weerahandi and Amaratunga (1999), and Gamage and Weerahandi (1998).

To provide formal definitions, consider a random vector $\mathbf{Y}$ with the cumulative distribution function $F(\mathbf{y};\boldsymbol{\zeta})$, where $\boldsymbol{\zeta} = (\theta, \boldsymbol{\delta})$ is a vector of unknown parameters. Recall that θ is the parameter of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters. Let $\mathbf{y}$ be the observed value of the random vector $\mathbf{Y}$. An extreme region with the observed sample point on its boundary can be denoted as $C(\mathbf{y};\theta,\delta)$. By taking the classical approach we defined extreme regions using test statistics in the previous section. That approach allowed us to define extreme regions having representations such as $\{\mathbf{Y} \mid T(\mathbf{Y};\boldsymbol{\theta}_0) \geq T(\mathbf{y};\boldsymbol{\theta}_0)\}$. There is no reason why an extreme region should always have such a simple structure. The boundary of extreme regions could be allowed to be any function of the quantities $\mathbf{y}$, θ, and $\boldsymbol{\delta}$, and therefore, we need to allow test variables to depend all these quantities. However, an extreme region is of practical use only if its probability does not depend on $\boldsymbol{\zeta}$. Moreover, a subset of the sample space obtained by more general methods should truly be an extreme region in that its probability should be greater under the alternative hypothesis than under the null hypothesis, as defined more formerly below.

Definition 1.4. A *generalized test variable* is a random variable of the form $T = T(\mathbf{Y};\mathbf{y},\boldsymbol{\zeta})$ having the following three conditions:

1. The observed value $t = T(\mathbf{y};\mathbf{y},\boldsymbol{\zeta})$ of T does not depend on unknown parameters.

2. The probability distribution of T does not depend on nuisance parameters.

3. Given t, $\mathbf{y}$ and $\boldsymbol{\delta}$, $\Pr(T \leq t;\theta)$ is a monotonic function of θ.

In ANOVA problems and in some other applications, Property 3 is too restrictive or not well-defined. In such situations, the following milder condition, which is referred to as the unbiasedness property, is considered adequate in place of Property 3.

3´. T satisfies the condition that, given t, $\mathbf{y}$ and $\boldsymbol{\delta}$,

$$\Pr(T \leq t|\theta \in \Theta_0) \leq \Pr(T \leq t|\theta \in \Theta_1) \text{ for all } t. \tag{2.11}$$

One can argue that Condition 1 is redundant, because if the property is not satisfied, then we can define an alternative generalized test variable $\tilde{T}$ as $\tilde{T} = T(\mathbf{Y};\mathbf{y},\boldsymbol{\zeta}) - T(\mathbf{y};\mathbf{y},\boldsymbol{\zeta})$ and then require it to have Conditions 2 and 3. These conditions are also referred to as properties of a test variable. Property 2 ensures that p-values based on generalized test variables are of practical use in that they can be computed for decision making. Property 3 or Property 3´ impose the requirement that the resulting testing procedure is unbiased.

It should be pointed out that any test statistic is also a test variable, and hence Definition 1.4 is indeed a generalization of Definition 1.1. The main differences in the two definitions are that the test variables can be functions of nuisance parameters and that the former allows the observed value $\mathbf{y}$ of

$\mathbf{Y}$ to appear on both sides of probability statements such as $\Pr(\mathbf{Y}|T(\mathbf{Y};\mathbf{y},\boldsymbol{\zeta}) \leq T(\mathbf{y};\mathbf{y},\boldsymbol{\zeta}))$. Note that such probability statements are based on proper subsets of the sample space Ξ with the observed value $\mathbf{y}$ falling on its boundary. In defining generalized p-values we need to make such probability statements, because, as we will see throughout this book, many unbiased extreme regions leading to tests and confidence regions cannot be obtained by taking the classical approach that relies on Definition 1.1 or its variations.

Since test variables are extensions of test statistics, they inherit all advantages and drawbacks of test statistics as well. In particular, usually there are multiple test variables for a given problem of hypothesis testing and a particular member might have certain undesirable properties. In other words, just because a test variable satisfy above conditions does not necessarily mean that it would lead to testing procedures having good size and power performance. Nevertheless, in many applications we can drop candidates having poor performance by confining the search for generalized test variables within the class of complete sufficient statistics. Moreover, if there are still multiple test variables, then the number of available can be further reduced by imposing additional optimality conditions and other desirable conditions such as those suggested by invariance properties.

Example 1.2. Testing the ratio of the normal distribution parameters

Let $X_1, \ldots, X_n$ be a random sample from the normal population with mean μ and variance σ^2. Suppose the parameter of interest is the ratio of the mean and the standard deviation, namely $\theta = \mu/\sigma$. Consider the problem of testing the hypothesis

$$H_0 : \frac{\mu}{\sigma} \leq \theta_0 \text{ against } H_1 : \frac{\mu}{\sigma} > \theta_0.$$

Without losing the power performance, testing procedures can be based on the sufficient statistics

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \text{ and } S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n},$$

whose distributions are given by

$$Z = \sqrt{n}(\frac{\overline{X}}{\sigma} - \theta) \sim N(0\ ,1) \quad , \quad and \quad U = \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1},$$

which are mutually independent. In this problem, it is does not seem easy to find a test statistic having Properties 1 and 2 of Definition 1.1. So, in view of the structure of above random variables, consider the potential test variable

$$T = \frac{\overline{x}S}{\sigma s} - \frac{\overline{X}}{\sigma},$$

a function of sufficient statistics, their observed values, and the parameters of the problem. Obviously, the observed value of T is zero and hence does not depend on any unknown parameters. When expressed as

$$T = \frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}} - \theta$$

it is also clear that the distribution of T does not depend on the nuisance parameter σ. Finally, it follows from the above identity that the probability distribution function of T is an increasing function of θ, because

$$\begin{aligned} F(t) &= \Pr(T \leq t) \\ &= \Pr(\frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}} \leq t + \theta) \\ &= F_V(t + \theta) \end{aligned} \tag{2.12}$$

is an increasing function of θ, where

$$V = \frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}}$$

is a random variable free of unknown parameters. Hence, T satisfies all three conditions of Definition 1.4 and so it is indeed a test variable.

2.3.1 Generalized extreme regions and generalized p-values

Now we are in a position to generalize the definition of extreme regions and facilitate their construction using generalized test variables defined above.

Definition 1.5. A subset of the sample space $C(\mathbf{Y}\,; \mathbf{y}, \boldsymbol{\zeta})$ is said to be a *generalized extreme region* if

1. the observed value $\mathbf{y}$ of $\mathbf{Y}$ falls on the boundary of the region,

2. the probability $\Pr(C(\mathbf{Y}\,; \mathbf{y}, \boldsymbol{\zeta}))$ does not depend on the nuisance parameters $\boldsymbol{\delta}$,

3. the probability satisfy the property

$$\Pr(C(\mathbf{Y}\,; \mathbf{y}, \boldsymbol{\zeta}) \,|\, \theta \in \Theta_0) \leq \Pr(C(\mathbf{Y}\,; \mathbf{y}, \boldsymbol{\zeta}) \,|\, \theta \in \Theta_1)\,. \tag{2.13}$$

Definition 1.6. The *generalized p-value* of a test based on a generalized extreme region is defined as

$$p = \mathrm{Sup}\{\Pr(C(\mathbf{Y}\,; \mathbf{y}, \boldsymbol{\zeta}) \,|\theta \in \Theta_0)\}.$$

The boundary of a generalized extreme region is denoted as $C_{\mathbf{y}}(\boldsymbol{\zeta})$. Test variables provide a convenient way of constructing generalized extreme regions. In fact, when there is only one parameter of interest, one-sided hypotheses of the form

$$H_0 : \ \theta \leq \theta_0 \text{ versus } H_1 : \ \theta > \theta_0 \tag{2.14}$$

or

$$H_0 : \ \theta \geq \theta_0 \text{ versus } H_1 : \ \theta < \theta_0,$$

the generalized p-value can be computed conveniently. If a test variable T is stochastically increasing in θ, then the p-value can be conveniently computed using the formulas

$$p = \Pr(T(\mathbf{Y}; \mathbf{y}, \boldsymbol{\zeta}) \ \geq t_{obs} | \theta = \theta_0) \tag{2.15}$$

and

$$p = \Pr(T(\mathbf{Y}; \mathbf{y}, \boldsymbol{\zeta}) \ \leq t_{obs} | \theta = \theta_0),$$

respectively, where $t_{obs} \ = T(\mathbf{y}; \mathbf{y}, \boldsymbol{\zeta})$. For other related definitions, such as that of power functions, and for some useful theorems in this context, the reader is referred to Weerahandi (1995) and Ananda and Weerahandi (2002).

Also of interest, especially in problems involving multiple location parameters including a variety of ANOVA problems that we will study in the following chapters, are hypothesis problems of the form

$$H_0 : \ \theta = \theta_0 \ versus \ H_1 : \ \theta \neq \theta_0. \tag{2.16}$$

In this situation Property 3´ of a test variable given above is considered adequate. In this case the p-value is computed as

$$\begin{aligned} p &= \Pr(C(\mathbf{Y} \ ; \mathbf{y}, \boldsymbol{\zeta}) | \theta = \theta_0) \\ &= \Pr(T \in C_t | \theta = \theta_0), \end{aligned} \tag{2.17}$$

where C_t is a subset of the sample space„ an extreme region, defined by $t = t_{obs} = T(\mathbf{y}; \mathbf{y}, \boldsymbol{\zeta})$.

Just like conventional p-values, generalized p-values serve to measure the evidence in favor or against the null hypothesis. The smaller the p-value, the greater the evidence against the null hypothesis, as implied by Property 3. Here also, in a fixed-level setting one can reject the null hypothesis when the p-value falls below a certain critical value such as the typical values 0.05 and 0.01.

For other important notions in this context such as the invariance, similarity, and unbiasedness, the reader is referred to Weerahandi (1995). Notions such as the invariance is particularly useful when there are multiple test variables satisfying the necessary conditions of a test variable.

Example 1.3. Testing the ratio of the parameters of the normal distribution (continued)

Consider again the problem discussed in Example 1.2 involving the ratio of the mean μ and the standard deviation σ. In Example 1.2 we saw that

$$\begin{aligned} T &= \frac{\overline{x}S}{\sigma s} - \frac{\overline{X}}{\sigma} \\ &= \frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}} - \theta \end{aligned}$$

is a generalized test variable, where $\overline{X}$ and S^2 are the maximum likelihood estimators of μ and σ^2. Also recall that

$$Z \;=\; \sqrt{n}(\frac{\overline{X}}{\sigma} - \theta) \sim N(0\ ,1) \quad ,\quad \textit{and}\ \ U = \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Consider again the problem of testing the hypotheses

$$H_0 :\ \theta \leq \theta_0, \quad H_1 :\ \theta > \theta_0\ ,$$

where $\theta \;=\; \mu/\sigma$ is the ratio of the parameters. Although it is possible to obtain a p-value for this problem by the classical approach, the generalized approach provides a more convenient way of constructing extreme regions and p-values. Consider the natural choice for a potential extreme region defined as

$$C = \{(\overline{X}, S \,|\, \frac{\overline{x}}{s} \leq \frac{\overline{X}}{S})\}.$$

This particular region is free of unknown parameters, but its probability depends on the parameter of interest. In general, the generalized approach allows the extreme region to depend even on the unknown parameters, as we will illustrate later. The above subset of the sample space can be expressed in terms of the test variable as

$$\begin{aligned} C &= \{(\overline{X}, S \,|\, \frac{\overline{x}S}{\sigma s} \leq \frac{\overline{X}}{\sigma})\} && (2.18) \\ &= \{(\overline{X}, S \,|\, T \leq 0)\}. && (2.19) \end{aligned}$$

It now follows from the properties of test variables that C is a proper extreme region.

Hence, the p-value for testing the hypothesis can be computed as

$$\begin{aligned} p = \;& \Pr(\ T \ \leq\ 0 \mid \theta = \theta_0\ ) \\ &= \Pr(\frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}} \leq \theta_0) \\ &= \Pr(\frac{\overline{x}}{s} \leq \frac{\theta_0 + Z/\sqrt{n}}{\sqrt{U/n}}). \qquad (2.20) \end{aligned}$$

The p-value in (2.20) can be computed by numerical integration with respect to the independent standard normal and chi-squared random variables Z and

U. The probability of the inequality in (2.20) can also be evaluated by the Monte Carlo method by generating a large number of random numbers from Z and U, and then finding the fraction of pairs of random numbers for which the inequality is satisfied. As it will become clear from Example 1.6, the p-value can also be computed using the cdf of the noncentral t-distribution with $n-1$ degrees of freedom and the noncentrality parameter $\theta_0\sqrt{n}$.

2.4 SUBSTITUTION METHOD

Example 1.3 demonstrates how easily one can obtain testing procedures for functions of parameters even in situations where the classical approach can produce exact tests, but the derivation is not as simple. In sampling from a normal distribution, one can in fact obtain generalized p-values for any complicated function such as $\theta = (\mu + \sigma)/(\mu^2 + \sigma^2)$. Except for a few special functions, the classical approach to testing of hypotheses fails to provide solutions to such problems. The problems get more and more complicated in sampling from a number of populations as we will have to address in the following chapters. Although the generalized approach can provide solutions in a wide class of problems, finding appropriate test variables in many applications is not a trivial task. So, it is desirable to have a systematic approach that we can take in solving at least some of the problems. One such method was proposed by Berger, Peterson, and Weerahandi (2003), which we refer to as the *substitution method.*

The substitution method assumes that there is a set of observable statistics with known distributions that is equal in number to the number of unknown parameters of the problem, say $(\alpha_1, \alpha_2, \ldots, \alpha_k)$. Let $(X_1, X_2, \ldots, X_k)$ be the set of observable sufficient statistics and let $(x_1, x_2, \ldots, x_k)$ be their observed values. In many applications a set of minimal sufficient statistics will serve this purpose. For example, in sampling from a normal population as we discussed in the previous section, the two statistics $\overline{X}$ and S^2 will satisfy this requirement in tacking any specified function of μ and σ^2. It is also assumed that through a set of random variables having distributions free of unknown parameters, the statistics are related to the unknown parameters. Continuing with the illustrative example, the random variables

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ and } U = \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

are used to meet this requirement in the case of sampling from a single normal population. Let $(V_1, V_2, \ldots, V_k)$ be the set of random variables with distributions free of unknown parameters. Though desirable, it is not necessary that these random variables are mutually independent. However, their joint distribution is assumed to be known. Then the substitution method is carried out in the following steps:

- By writing the parameter of interest, θ, in terms of the parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ or otherwise, express θ in terms of the sufficient statistics $(X_1, X_2, \ldots, X_k)$ and the random variables $(V_1, V_2, \ldots, V_k)$.

- Replace the statistics $(X_1, X_2, \ldots, X_k)$ by their observed values $(x_1, x_2, \ldots, x_k)$ and substrate θ from the result to define a representation of a generalized test variable T, say Representation 1.

- Rewrite $(V_1, V_2, \ldots, V_k)$ terms appearing in T in terms of $(X_1, X_2, \ldots, X_k)$ and $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ and obtain a second representation of T, say Representation 2.

- Use Representation 2 to show that T satisfies Property 1 of a test variable and to show that an extreme region defined in terms of T have the observed sample point on its boundary.

- Use Representation 1 to show that T satisfies Property 2 and to check whether T satisfies Property 3.

- Compute the generalized p-value based on T.

It should be emphasized that various replacements of parameters by random variables and substitution of random variables by their observed values as appearing in the above steps are merely for the purpose of finding a potential test variable. They are by no means parts of a derivation. Of course one cannot even do such substitutions in a derivation. Therefore, after obtaining the form of the potential test variable, we need to prove that it is indeed a test variable leading to a well-defined extreme region.

Example 1.4. Testing the parameter $\theta = \mu + \sigma^2$

As discussed in Weerahandi (1995), $\theta = \mu + \sigma^2$ is a function of the parameters of the normal distribution that arise in some practical applications. The parameter can be expressed in terms of the sufficient statistics and random variables as

$$\theta = \overline{X} - Z\,\sigma/\sqrt{n} + \sigma^2 \tag{2.21}$$

$$= \overline{X} - Z\,\frac{S}{\sqrt{U}} + \frac{nS^2}{U}\,, \tag{2.22}$$

where Z and U are the normal and chi-squared random variables defined above. Having obtained the identity that relates the parameter to the sufficient statistics and random variables that are free of unknown parameters, we can now follow step 3 and 4 to obtain the potential test variable as

$$
\begin{aligned}
T &= \overline{x} \; - Z \; \frac{s}{\sqrt{U}} + \frac{ns^2}{U} - \theta \\
&= \; \overline{x} \; - \frac{\overline{X} \; - \; \mu}{\sigma/\sqrt{n}} \; \frac{s\sigma/\sqrt{n}}{S} + \frac{s^2\sigma^2}{S^2} - \theta \\
&= \overline{x} \; - \frac{s(\; \overline{X} \; - \; \mu)}{S} \; + \frac{s^2\sigma^2}{S^2} - \theta. \qquad (2.23)
\end{aligned}
$$

Obviously, the observed value of T is zero and by construction its distribution does not depend on nuisance parameters. It also follows from (2.23) that it is stochastically decreasing in the parameter of interest θ. Hence, T is indeed a test variable. So, for instance hypotheses of the form $H_0 : \; \theta \leq \theta_0$ can be tested based on the p-value

$$
\begin{aligned}
p &= \Pr(\; T \leq 0 \mid \theta = \theta_0 \;) \qquad (2.24) \\
&= \Pr(\overline{x} \; - \theta_0 \leq Z \; \frac{s}{\sqrt{U}} - \frac{ns^2}{U}) \qquad (2.25)
\end{aligned}
$$

Figure 1.1 shows the form of the extreme region on which the p-value is based. The figure is drawn for the particular observed values $\overline{x} \; = 5.5, s = 1$, and parameters $\mu = 5, \sigma = 1$. The figure shows a part of the sample space formed by $\overline{X}$ and S^2. The region below the curve is the extreme region. Although this is a well-defined subset of the sample space with $(\overline{x},\, s^2)$ on its boundary, the classical approach based on test statistics cannot produce this extreme region.

In this example also the p-value can be computed by numerical integration with respect to the independent standard normal and chi-squared random variables Z and U. The probability of the inequality in appearing in the formula can also be evaluated by the Monte Carlo method. This is accomplished by generating a large number of random numbers from Z and U, and then finding the fraction of pairs of random numbers for which the inequality is satisfied.

When there is more than one parameter of interest, as usually the case in most linear models that we will undertake, the substitution method can be modified to obtain a potential test variable. The modified approach will be discussed in Chapter 3. Other more formal methods deriving test variables using properties such as invariance are discussed by Tsui and Weerahandi (1989). The reader is also referred to Weerahandi (1995) for a detailed discussion of the notions of *invariance*, *unbiasedness*, and *similarity.* It also provides numerous illustrations of how these notions can be utilized in deriving testing procedures in the context of the generalized inference.

.

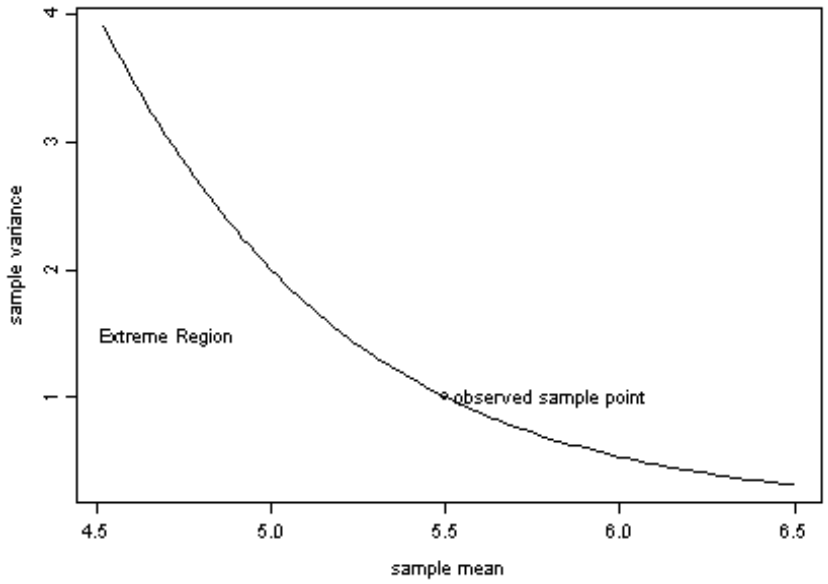

Figure 2.1 Extreme region of sufficeient statistics

2.5 FIXED-LEVEL TESTING

Since the class of test statistics is a particular case of test variables, there is a generalization of conventional fixed-level tests as well. In fact, despite the unconventional approach to driving generalized tests, the generalized p-values defined and illustrated in previous sections are functions of data only. This means that the generalized p-value itself can be employed as a test statistic. In fact, fixed-level tests based on generalized p-values can be performed as suggested by the definition below.

Definition 1.7. Let $p(\mathbf{y})$ be a generalized p-value based on a generalized test variable $T = T(\mathbf{Y};\ \mathbf{y}, \boldsymbol{\zeta})$ with the observed value $t = T(\mathbf{y};\ \mathbf{y}, \boldsymbol{\zeta})$. Assume that $p(\mathbf{y})$ is continuous in $\mathbf{y}$. Then, the test based on the rule

$$\text{reject } H_0 \ \text{ if } \ p(\mathbf{y}) \ < \ \alpha \tag{2.26}$$

is said to be a *generalized fixed-level test* of level α for testing the null hypothesis $H_0 :\ \theta \in \Theta_0$ against the alternative $H_1 :\ \theta \in \Theta_1$.

Since $p(\mathbf{y})$ is being used as a test statistic, a generalized fixed-level of level α is also classical fixed-level test of approximate level α. A major advantage of fixed-level tests obtained using generalized p-values compared to other methods of obtaining fixed-level tests is that, as illustrated in Chapter 4, they enable one to specify a general testing rule for a wide range of similar problems.

Moreover, not only do these tests tend to be simpler both in form and in the derivation, but also they often outperform more complicated approximations, in power and/or in size. Therefore, the generalized p-values and generalized fixed-level tests should be of interest to decision theorists, researchers, and practitioners who insist on classical fixed-level tests, as well as of interest to practitioners who prefer to report p-values. It should be noted, however, that when we resort to generalized p-values in situations where exact classical tests do not exist, one cannot always expect them to have other desirable properties such as exact frequency coverage in repeated sampling.

2.5.1 Frequency Properties

Except in special cases, the generalized p-values has only unconventional frequency interpretations implied by the definition. For example, if the experiment is repeated with new samples, the proportion of samples that fall into the current extreme region is equal to p. But this result does not really have any implication on the Type I error.

To shed more light on this issue, consider the test statistic P defined as $P = p(\mathbf{Y})$, which is obtained by replacing the observed $\mathbf{y}$ appearing in the generalized p-value by the random vector $\mathbf{Y}$. When P is a continuous random variable, it follows from the probability integral transform that the cumulative distribution function $F(P; \boldsymbol{\zeta})$ of P has a uniform distribution over the interval $[0, 1]$; that is

$$F(P) \sim U(0, 1).$$

This implies, for instance, that if we reject the null hypothesis $H_0 : \theta = \theta_0$ for the observed values of p of P that satisfies the inequality $F(p \mid \theta = \theta_0) < \alpha$, then in repeated sampling, the probability of rejecting H_0 when H_0 is true exactly α. But the problem is that $F(p|\theta = \theta_0)$ is not always free of nuisance parameters. In some applications such as that in Example 1.2 it is free of nuisance parameters, as further discussed in Example 1.6. In fact when the test variable is of the form $T = T(\mathbf{Y}; \boldsymbol{\zeta})$, from the stochastic monotonicity of T it follows that, $F(P|\theta = \theta_0) = \Pr(P \leq p)$ can be computed using the original test variable or using the p-value as $p(\mathbf{y}) = \Pr(T \leq t)$ or $\Pr(T \geq t)$, which is free of nuisance parameters, where $t = T(\mathbf{y}; \boldsymbol{\zeta})$. When the test variable is of the form $T = T(\mathbf{Y}; \mathbf{y}, \boldsymbol{\zeta})$, we can still compute the generalized p-value, but rejecting H_0 when $p(\mathbf{y}) < \alpha$ is only an approximate fixed-level test in the classical sense.

Nevertheless, according to various simulation studies reported in the literature, in many applications the generalized fixed-level tests often do not exceed the desired level. In other situations they provide excellent approximate tests in the classical sense. According to the simulation studies reported by Griffiths and Judge (1992), Thursby (1992), Weerahandi and Johnson (1992), and Zhou and Mathew (1994), Ananda and Weerahandi (1996, 1997), Gamage and Weerahandi (1998), and Weerahandi and Amaratunga (1999), in a

variety of applications the generalized fixed-level tests did not exceed the intended level. Some of these studies also reported that power performance of generalized fixed-level tests were as good as or better than other approximate tests which are not exact in any sense.

Even when the actual size of tests given by the rule in Definition 1.7 exceeds the intended level, usually it is possible to construct tests based on generalized p-values, at the cost of a little loss of power. As Weerahandi (1995) argued, resorting to such tests is worthwhile only in situations where repeated sampling properties of fixed-level testing is considered to be practically useful as opposed to a matter of convention or habit. The author also argued that, if the same experiment can indeed be repeated, one should rather combine the data to perform a more powerful test. Except perhaps for applications involving statistical quality control, there is no common agreement about the practical use of repeated sampling properties. Some statisticians advocate the use of procedures having desirable properties with respect to the current sample rather than other possible samples that could have been observed, but were not. Moreover, as arguments of Pratt (1961) and Kiefer (1977) imply, insisting on the repeated sampling property with the same experiment can sometimes lead to procedures with highly undesirable features.

2.6 GENERALIZED CONFIDENCE INTERVALS

The generalized p-values have implications in interval estimation as well. In this section we first define a counterpart of generalized p-values in interval estimation and then show how they can be derived directly or deduced from generalized p-values. Just like the generalized p-values, the generalized confidence intervals will prove useful to all practitioners regardless of whether or not they insist on conventional confidence intervals with frequency interpretations.

The classical approach to interval estimation suffer from more difficulties than that of hypothesis testing. Even when the problem does not involve nuisance parameters and there are exact confidence intervals, in some applications they lead to results that contradict the very meaning of *confidence*. Probably Pratt (1961) was the first author to provide a very simple example of a uniformly most accurate confidence interval having highly undesirable properties. Weerahandi (1995) showed how such undesirable confidence intervals can be avoided by expanding the class of intervals available to choose from. Just as in the case of testing of hypotheses, here we extend the class of available procedures for any given problem by insisting on exact probability statements rather than on repeated sampling properties. This will enable us to solve such problems as the Behrens−Fisher problem for which exact classical confidence intervals do not exist. As in the Bayesian approach, the idea is to do the best with the observed data at hand instead of talking about other samples that could have been observed. The generalized confidence intervals

are nothing but the enhanced class of interval estimates obtained from exact probability statements with no special regard to repeated sampling properties that are of little practical use [cf. Weerahandi (1995)].

2.6.1 Classical confidence intervals

Consider a population represented by an observable random variable Y. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a random sample available from the population. Suppose the distribution of the random variable Y is known except for a vector of parameters $\boldsymbol{\zeta} = (\theta, \boldsymbol{\delta})$, where θ is a parameter of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters. In general, θ could be a vector of parameters and we might be interested in finding a confidence region, but for the sake of simplicity and to be specific, let us first assume that there is only one parameter of interest and we are interested in finding an interval estimate of θ based on observed values of $\mathbf{Y}$. The problem is to construct generalized confidence intervals of the form $[A(\mathbf{y}), B(\mathbf{y})] \subset \Theta$, where $A(\mathbf{y})$ and $B(\mathbf{y})$ are functions of $\mathbf{y}$, the observed data.

In the classical approach to interval estimation we find two functions of the observable random vector, say $A(\mathbf{Y})$ and $B(\mathbf{Y})$ such that the probability statement

$$\Pr[\, A(\mathbf{Y}) \leq \theta \leq B(\mathbf{Y}) \,] = \gamma \tag{2.27}$$

is satisfied, where γ is specified by the desired confidence level. For example, to construct 95% confidence intervals we set γ at 0.95. If it is possible to find two such functions, $A(\mathbf{Y})$ and $B(\mathbf{Y})$, that do not depend on unknown parameters, then we compute $a = A(\mathbf{x})$ and $b = B(\mathbf{y})$ using the observed value $\mathbf{y}$ of $\mathbf{Y}$ and call $[a, b]$ a $100\gamma\%$ confidence interval. The interval obtained in this manner has the property that, in repeated sampling, the interval would contain the parameter θ, $100\gamma\%$ of the times. It of course has no implication about the coverage of θ with the sample that we have actually observed. In fact Pratt (1961) and Kiefer (1977) provide examples where the current intervals violating the very meaning of *confidence*. In particular, they showed that in those applications the so called exact confidence intervals do not contain the parameters at all. The only thing truly exact about a confidence interval is the probability statement on which the interval is based.

Therefore, we can extend the class of candidates eligible to be interval estimators by insisting on the probability statement only. This will allow us to find interval estimates for situations where it is not easy or impossible to find $A(\mathbf{Y})$ and $B(\mathbf{Y})$ satisfying (2.27) for all possible values of the nuisance parameters. Weerahandi (1993) showed how this can be accomplished by making probability statements relative to the observed sample, as done in the Bayesian approach, but without having to treat unknown parameters as random variables. More precisely, we can allow $A()$ and $B()$ to depend on the observable random vector $\mathbf{Y}$ and the observed data $\mathbf{y}$ both. When there are

a number of parameters of interest, in general we could allow subsets of the sample space possibly depending on the current sample point $\mathbf{y}$ of $\mathbf{Y}$. The construction of such regions can be facilitated by generalizing the classical definition of pivotal quantities.

Definition 1.8. A random variable of the form $R = R(\mathbf{Y}; \mathbf{y}, \boldsymbol{\zeta})$, a function of $\mathbf{Y}$,$\mathbf{y}$, and $\boldsymbol{\zeta}$, is said to be a *generalized pivotal quantity* if it has the following two properties:

Property A: The probability distribution of R does not depend on unknown parameters.

Property B: The observed pivotal, $r_{obs} = R(\mathbf{y}; \mathbf{y}, \boldsymbol{\zeta})$ does not depend on nuisance parameters, $\boldsymbol{\delta}$.

Property A allows us to write probability statements leading to confidence regions that can be evaluated regardless of the values of the unknown parameters. Property B ensures that when we specify the region with the current sample point $\mathbf{y}$, then we can obtain a subset of the parameter space that can be computed without knowing the values of the nuisance parameters.

Example 1.5. The ratio of the parameters of the normal distribution (continued)

Let $X_1, \ldots, X_n$ be a random sample from the normal population with mean μ and variance σ^2. Suppose $\theta = \mu/\sigma$, the ratio of the mean and the standard deviation, is the parameter of interest. In view of the results in Example 1.2, consider the potential generalized pivotal quantity

$$R = R((\overline{X}, S); \overline{x}, s, \theta, \sigma) = \frac{\overline{x} S}{\sigma s} - \frac{\overline{X}}{\sigma} + \theta,$$

based on the sufficient statistics, their observed values, and the parameters of the problem. Obviously, the observed value of R is θ and so it satisfies Property B of a generalized pivotal. When expressed as

$$R = \frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}},$$

it is also clear that the distribution of R is free of unknown parameters, where

$$Z = \sqrt{n}(\frac{\overline{X}}{\sigma} - \theta\ ) \sim N(\ 0\ ,\ 1\ )\ ,\ \text{and}\ \ U = \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Hence R is indeed a generalized pivotal quantity.

Suppose we have constructed a generalized pivotal $R = R(\mathbf{Y}; \mathbf{y}, \boldsymbol{\zeta})$ for a parameter θ of interest and we wish to construct a confidence region at confidence coefficient γ. Consider a subset C_γ of the sample space chosen such that

$$\Pr(\ R\ \in\ C_\gamma\ )\ =\ \gamma\ . \tag{2.28}$$

The region defined by (2.28) also specifies a subset $C(\mathbf{y};\theta)$ of the original sample space satisfying the equation $\Pr(\mathbf{Y} \in C(\mathbf{y};\theta)) = \gamma$. Unlike classical confidence intervals, this region depends not only on γ and θ, but also on the current sample point $\mathbf{y}$. With this generalization we can obtain interval estimates on θ relative to the observed sample with no special regard to samples that could have been observed but were not. Although the generalized approach shares the same philosophy of the Bayesian approach that the inferences should be made with special regard to the data at hand, here we do not treat parameters as random variables and hence the probability statements are made with respect to the random vector $\mathbf{Y}$. Having specified a subset of the sample space relative to the current sample point, we can evaluate the region at the observed sample point and proceed to solve (2.28) for θ and obtain a region as defined below

Definition 1.9. A subset Θ_c of the parameter space is said to be a $100\gamma\%$ *generalized confidence interval* for θ if it satisfies the equation

$$\Theta_c(r) \;=\; \{\theta \in \Theta \mid R(\mathbf{y};\mathbf{y},\boldsymbol{\zeta}) \;\in C_\gamma\}, \tag{2.29}$$

where the subset C_γ of the sample space ρ of R satisfies equation (2.28).

It should be noted that generalized confidence intervals are not alternatives, but rather extensions of classical confidence intervals. In fact, for a given problem there is usually a class of confidence intervals satisfying the probability statement (2.28), a feature of classical intervals as well. Weerahandi (1995) discussed how the choice of appropriate generalized pivotals could be facilitated by invoking the principals of sufficiency and invariance. Even after we have obtained a particular pivotal quantity we could construct a variety of confidence regions. Depending on the application, a left-sided interval, a right-sided interval, a two-sided interval symmetric around the parameter, a shortest confidence interval, or some other interval might be preferable.

Being an extension, the generalized confidence intervals also inherit all desirable and undesirable features of confidence intervals. Weerahandi (1995) shows how the extended class of intervals help avoid some of the undesirable features of classical intervals. In those applications where the classical approach fails to provide exact confidence intervals and yet the classical Bayesian approach works, the generalized confidence intervals are often (but not always) numerically equivalent to the Bayesian confidence intervals under certain noninformative priors. At present there are no Bayesian results available for most of the ANOVA type problems that we will encounter throughout this book. In fact ANOVA can be considered as a class of problems in which the classical Bayesian approach fails and hence the Bayesian school has been silent in a wide class of problems in the area of repeated measures. In a Bayesian treatment of the one-way ANOVA problem, however, Weerahandi and Tsui (1996) showed how Bayesian procedures equivalent to generalized p-values can be derived using the posterior predictive p-value approach introduced by Meng (1994) [see also Gelman, Meng, and Stern (1996)]. In fact

their exposition provides a promising Bayesian approach that one can take in solving problems in the context of repeated measures.

2.6.2 Intervals with frequency interpretations

Neither the generalized approach nor the classical approach is truly based on repeated sampling considerations. The probability statements such as (2.27), however, have implications in repeated sampling from the same sample space. To be specific, consider the problem of interval estimation of a parameter θ. If the same experiment is repeated a large number of times to obtain new sets of observations, $\mathbf{y}$, of the same sample size n, then the confidence intervals given by the probability statement (2.27) will include the true value of the parameter θ, $100\gamma\%$ of the time. The generalized confidence intervals, Bayesian intervals, and classical intervals on parameters of discrete distributions do not have this property exactly, but they all do so approximately.

The above property has no implication on the coverage of the parameter by the current sample. In fact Pratt (1961) showed situations where the current sample itself implies that the confidence interval does not contain the parameter. Weerahandi (1993) argued that a confidence interval we report does not even remain valid if we can indeed repeat the experiment to obtain more samples. Even if one attempts to justify its practical use by imagining a large number of practitioners analyzing different sets of data from identical experiments (which is unlikely even for two practitioners, except perhaps in quality control problems, where the same experiment is repeated periodically), the fraction of practitioners with intervals containing the parameter will only be approximately $100\gamma\%$. The only thing exact about a confidence interval is the probability statement on which the interval is based, a property that generalized confidence intervals also have. Weerahandi (1995) also provided some unconventional repeated sampling properties that the generalized confidence intervals possess. Although the repeated sampling experiments are all hypothetical, the repeated sampling property has one practical use when one needs to compare the performance of any type of confidence intervals by simulation regardless of the way they are constructed.

The generalized confidence intervals can also serve as approximate classical confidence intervals, and so they are useful to all practitioners regardless of the unconventional way they are constructed. A number of studies have reported findings on the repeated sampling properties of generalized procedures. The simulation studies carried out by Griffiths and Judge (1992), Thursby (1992), Weerahandi and Johnson (1992), Zhou and Mathew (1994), Gamage and Weerahandi (1998), and Weerahandi and Amaratunga (1999) show that in most applications, the generalized confidence intervals preserve the confidence level. That is, actual coverage of generalized confidence intervals is less than or equal to the intended level. In other applications, they not only hold the confidence level approximately but also tend to perform better than other approximations available in the literature. Therefore, practitioners who

prefer to make inferences without deviating from the classical philosophy of statistical inference can also take advantage of generalized confidence intervals to find solutions in wide class of problems.

Example 1.6. The ratio of the parameters of the normal distribution (continued)

In Example 1.5 we showed that, in the problem of sampling from a normal distribution with mean μ and standard deviation and σ,

$$\begin{aligned} R &= \frac{\overline{x}S}{\sigma s} - \frac{\overline{X}}{\sigma} + \theta \\ &= \frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}}, \end{aligned} \tag{2.30}$$

is a generalized pivotal quantity for $\theta = \mu/\sigma$, the ratio of the mean and the standard deviation, where Z is standard normal random variable and U is a chi-squared random variable with $n-1$ degrees of freedom. Suppose the problem is to construct $100\gamma\%$ lower confidence intervals for θ. Consider the probability statement

$$\begin{aligned} \gamma &= \Pr(R \le k) \\ &= \Pr(\frac{\overline{x}}{s}\sqrt{\frac{U}{n}} - \frac{Z}{\sqrt{n}} \le k) \\ &= \Pr(\frac{\overline{x}}{s} \le \frac{k\sqrt{n}+Z}{\sqrt{U}}) \\ &= 1 - \Pr(\frac{k\sqrt{n}+Z}{\sqrt{U/(n-1)}} \le \frac{\overline{x}}{s}\sqrt{n-1}) \\ &= 1 - F_W(\frac{\overline{x}}{s}\sqrt{n-1}), \end{aligned} \tag{2.31}$$

where F_W is the cdf of the random variable W, which has noncentral t-distribution with $n-1$ degrees of freedom and noncentrality parameter $k\sqrt{n}$. Let $k(\overline{x}/s;\gamma)$ be the value of k that satisfies the above equation. Since the observed value of R is θ, it is now clear that $(\theta \le k(\overline{x}/s;\gamma)$ is a $100\gamma\%$ lower confidence interval for θ.

Note that in this application $R = \frac{\overline{x}S}{\sigma s} - \frac{\overline{X}}{\sigma} + \theta$ is not a pivotal quantity in the classical sense. Yet the generalized confidence interval is also a classical confidence interval having the repeated sampling property. To see this more clearly, notice that

$$\begin{aligned} \Pr(\frac{\overline{X}}{S} \le c) &= \Pr(\frac{Z+\theta\sqrt{n}}{\sqrt{U}} \le c) \\ &= 1 - F_W(c\sqrt{n-1}) = \gamma \end{aligned} \tag{2.32}$$

and hence the generalized confidence intervals given by R is equivalent to the classical interval given by the statistic $V = \overline{X}/S$. Yet V is not a pivotal quantity in the classical sense. In repeated sampling, $\overline{x}/s$ would fall below c, $100\gamma\%$ of the times and the above confidence interval would contain $\theta, 100\gamma\%$ of the times. Not in all situations can a generalized confidence interval also be attained using a statistic or a pivotal quantity. The Behrens−Fisher problem is the well-known problem where such statistics do not exist. This is also the case in simpler problems such as the ones undertaken in the illustrations of the next two sections.

2.7 SUBSTITUTION METHOD IN INTERVAL ESTIMATION

In Example 1.6, we obtained a generalized pivotal quantity by using the generalized test variable we had obtained before. It is desirable to have an independent and a systematic method of finding generalized pivotals when the parameter of interest is not a simple function of the parameters of the underlying distribution. Peterson, Berger, and Weerahandi (2003) proposed one such method for a class of applications. As we discussed its counterpart in the context of testing of hypotheses, the procedure involves various substitutions of random variables by their observed values and parameters and hence they called the method the *substitution approach*.

This method requires that there is a set of observable statistics with known distributions that is equal in number to the number of unknown parameters of the problem, say $(\alpha_1, \alpha_2, \ldots, \alpha_k)$. Again consider a set of observable statistics $(X_1, X_2, \ldots, X_k)$ with the observed values $(x_1, x_2, \ldots, x_k)$. It is assumed that through a set of random variables having distributions free of unknown parameters, the statistics are related to the unknown parameters. In many applications this would be a set of minimal sufficient statistics with known distributions that can be transformed into distributions free of unknown parameters. Recall that, for the problem of sampling from a normal population, the two statistics $\overline{X}$ and S^2 will serve this need in constructing interval estimates for a certain function of .μ and σ^2. In that situation the sufficient statistics can be transformed into a standard normal random variable and a chi-squared random variable.

Let $\mathbf{V} = (V_1, V_2, \ldots, V_k)$ be the set of random variables with distributions free of unknown parameters. It is assumed that the joint distribution of the random vector $\mathbf{V}$ is known. Although the substitution method can work when θ is a vector of parameters and we need to find a generalized confidence region for θ, for the sake of simplicity, here we present the method when there is just one parameter of interest. In Chapter 6 we deal with situations where there is a vector of parameters of interest and hence we will leverage the substitution method in such a way that substitutions implemented in terms of various matrix inversions and multiplications are valid in the context of matrix

algebra. In finding generalized pivotal quantities, the substitution method is carried out in the following steps:

- Express the parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ and then θ in terms of the sufficient statistics $(X_1, X_2, \ldots, X_k)$ and the random variables $(V_1, V_2, \ldots, V_k)$.
- Define a potential generalized pivotal quantity, say R, by replacing the statistics $(X_1, X_2, \ldots, X_k)$ by their observed values $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ and argue that the distribution of is free of unknown parameters.
- Rewrite $(V_1, V_2, \ldots, V_k)$ terms appearing in R in terms of $\mathbf{X}$ and $\boldsymbol{\alpha}$ and show that when $\mathbf{X} = \mathbf{x}$, the observed value of the quantity $R(\mathbf{x}; \mathbf{x}, \boldsymbol{\alpha})$ does not depend on the nuisance parameters, where $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$.

It should be emphasized that various substations and use of alternative replacements in the above steps are merely steps in obtaining a potential generalized pivotal of the form $R = R(\mathbf{X}; \mathbf{x}, \boldsymbol{\alpha})$. Then we can establish that it is indeed a generalizer pivotal quantity and proceed to construct generalized confidence intervals having desirable features. Moreover, as Berger, Peterson, and Weerahandi (2003) state, just because R is a generalized pivotal quantity, there is no guarantee that it will always lead to inference procedures having optimum properties. Unless the pivotal based on minimal sufficient statistics is unique up to equivalent pivotals, a generalized pivotal quantity obtained by another method may have better properties. In many applications, the choice of available pivotals can be minimized by requiring them to have further desirable properties such as the invariance.

Example 1.7. Interval estimation of $\theta = (\mu + \sigma)/(\mu^2 + \sigma^2)$

The substitution method is especially useful in finding confidence intervals of complicated functions of parameters such as θ in this example, where μ and σ are the mean and the standard deviation of the normal distribution. Consider the problem of finding intervals for θ based on the sufficient statistics $\overline{X}$ and S^2. Recalling that

$$Z = \sqrt{n}(\frac{\overline{X} - \mu}{\sigma}) \sim N(0,1) \text{ and } U = \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}.$$

we can base the substitution method on the identity

$$\begin{aligned} \theta &= (\mu + \sigma)/(\mu^2 + \sigma^2) \\ &= \frac{\overline{X} - Z\ \sigma/\sqrt{n} + \sigma}{(\overline{X} - Z\ \sigma/\sqrt{n})^2 + \sigma^2} \\ &= \frac{\overline{X} - Z\ S/\sqrt{U} + S\sqrt{n/U}}{(\overline{X} - Z\ S/\sqrt{U})^2 + nS^2/U}. \end{aligned} \tag{2.33}$$

Now replacing the observable random variables by the observed values, we get two representations of the potential generalized pivotal as

$$R = \frac{\overline{x} \ - Z \ s/\sqrt{U} + s\sqrt{n/U}}{(\overline{x} \ - Z \ s/\sqrt{U})^2 + ns^2/U} \tag{2.34}$$

$$= \frac{\overline{x} \ - \ s(\overline{X} - \mu)/S + s\sigma/S}{(\overline{x} \ - s(\overline{X} - \mu)/S)^2 + (s\sigma/S)^2} \tag{2.35}$$

It is evident from the first of the above representations that the distribution R is free of unknown parameters, and the second representation implies that $r_{obs} = R(\overline{X} \ , S; \overline{\mathbf{x}}, \mathbf{s}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \boldsymbol{\theta}$ does not depend on nuisance parameters. Hence R is indeed a generalized pivotal quantity, and so we can construct any type of generalized confidence intervals for the parameter of interest θ. For example $100\gamma\%$ upper confidence bound can be obtained by finding k from the probability statement

$$\begin{aligned} \gamma &= \Pr(R \leq k) \\ &= \Pr(\overline{x} \ - Z \ s/\sqrt{U} + s\sqrt{n/U} \leq k((\overline{x} \ - Z \ s/\sqrt{U})^2 + ns^2/U)). \end{aligned} \tag{2.36}$$

The probability can be evaluated by numerical integration with respect to (Z, U) or by Monte Carlo integration. If k_γ is the value of k that satisfies the above inequality, then the $100\gamma\%$ upper confidence bound for $\boldsymbol{\theta}$ is k_γ.

2.8 GENERALIZED P-VALUE-BASED INTERVALS

Generalized confidence intervals can also be deduced from generalized p-values. This method is specially useful when we have already dealt with a hypothesis testing problem concerning the parameter of interest, say θ, a function of the parameters of underlying distribution. Suppose we have obtained a p-value, say $p(t; \theta_0)$ for testing a one-sided null hypothesis. Assume that the p-value is based on a generalized test variable T, which is stochastically increasing in θ. When the form of the p-value is known in terms of θ_0, consider the function $p(t; \theta)$ of the test is obtained simply by replacing θ_0 by θ. In fact, in repeated sampling with fixed x, $p(T(X; x, \boldsymbol{\zeta}); \theta)$ serves as a generalized pivotal quantity having a uniform distribution on $[0, 1]$, provided that T is a continuous random variable. Therefore, suppressing t in $p(t; \theta)$ inequalities such as

$$[\ \ p(\theta) \leq \gamma] \tag{2.37}$$

would yield one-sided $100\gamma\%$ generalized confidence intervals for θ. Similarly, statements such as

$$[\ \gamma_1 \ \leq \ p(\theta) \ \leq \gamma_2 \ ] \tag{2.38}$$

can be used to obtain bounded $100\gamma\%$ generalized confidence intervals for θ, where γ_1 and γ_2 are numbers between 0 and 1 chosen appropriately subject to the condition $\gamma = \gamma_2 - \gamma_1$.

Example 1.8. Interval estimation of $\theta = \mu + \sigma^2$

In Example 1.4 we obtained generalized p-values for testing the parameter, $\theta = \mu + \sigma^2$, the sum of the mean and the variance of the normal distribution $N(\mu, \sigma^2)$. Consider the problem of constructing confidence intervals for θ based on the sufficient statistics $\overline{X}$ and S^2. Recall that in Example 1.4 by using the generalized test variable

$$T = \overline{x} \; - \frac{s(\; \overline{X} \; - \; \mu)}{S} \; + \frac{s^2\sigma^2}{S^2} - \theta$$

we obtained the generalized p-value

$$p = \Pr(\; T \leq 0 \mid \theta = \theta_0 \;) \tag{2.39}$$

$$= \Pr(\overline{x} \; - \theta_0 \leq Z \; \frac{s}{\sqrt{U}} - \frac{ns^2}{U}) \tag{2.40}$$

for testing hypotheses of the form $H_0 : \; \theta \leq \theta_0$, where the probability is computed with respect to the independently distributed standard normal random variable Z and the chi-squared random variable U. Let

$$W = \overline{x} \; - Z \; \frac{s}{\sqrt{U}} + \frac{ns^2}{U}$$

so that we can use

$$p(\theta) = \Pr(W \leq \theta)$$

to deduce generalized confidence intervals from generalized p-values. Since W is a continuous random variable, $P = p(W)$ has a uniform distribution on the [0,1] interval when $(\overline{x}, s)$ fixed at the current values. Therefore if θ^* is chosen such that

$$p(\theta^*) = \Pr(\overline{x} \; - Z \; \frac{s}{\sqrt{U}} + \frac{ns^2}{U} \leq \theta^*) = .95,$$

then $(P(\theta) \leq P(\theta^*)]$ is a 95% generalized confidence interval for θ. Since $p(W)$ is an increasing function of W, the 95% generalized confidence interval is in fact $(\theta \leq \theta^*]$. Moreover, if θ_* is chosen such that

$$p(\theta_*) = \Pr(\overline{x} \; - Z \; \frac{s}{\sqrt{U}} + \frac{ns^2}{U} \leq \theta_*) = .05,$$

then $[\theta_*, \theta^*]$ is a 90% confidence interval for θ.

Exercises

2.1 Let $Y_1, ..., Y_n$ be a random sample from the normal distribution with mean μ and variance σ^2. Let $\overline{Y}$ and S^2 be the unbiased estimators of μ and σ^2, respectively, where

$$\overline{Y} = \frac{\sum Y_i}{n} \text{ and } S^2 = \frac{\sum (Y_i - \overline{Y})^2}{n-1}.$$

Show that

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ and } U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Also show that the random variables Z and U are independently distributed.

2.2 Consider a random sample $Y_1, ..., Y_n$ from a population with mean μ and variance σ^2. Suppose $\theta = \mu/\sigma^2$ is the parameter of interest. By applying the substitution method or otherwise,

(a) obtain a generalized test statistic for making inferences about the parameter θ,
(b) obtain generalized p-values for testing hypotheses of the form $H_0 : \theta \leq \theta_0$ *against* $H_1 : \theta > \theta_0$,
(c) deduce left-sided and right-sided $100\gamma\%$ generalized confidence intervals for θ.

2.3 Consider the generalized test variable obtained in the previous problem.

(a) Deduce a generalized pivotal quantity appropriate in interval estimation for the parameter θ.
(b) Obtain the equal tail $100\gamma\%$ generalized confidence intervals for θ.
(c) Discuss whether or not this generalized confidence interval has the repeated sampling property.

2.4 Consider a random sample from a normal population. Establish procedures for making inferences about the second moment of its distribution.

2.5 Consider again a random sample from a normal population. Establish procedures for making inferences about the third moment of the distribution.

2.6 Suppose $Y_1, ..., Y_n$ is a random sample from a normal population with mean μ and variance σ^2. In a certain statistical problem, suppose

$$\frac{1}{\mu} + \frac{1}{\sigma}$$

is the parameter of interest. By applying the substitution method or otherwise,

(a) obtain a generalized pivotal quantity for interval estimation of the parameter θ,
(b) construct $100\gamma\%$ generalized confidence intervals for θ,
(c) deduce a test variable from the generalized pivotal,
(d) find the generalized p-value given by that test variable for testing hypotheses of the form $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$.

2.7 Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution with density

$$f_X(x) = \frac{1}{\beta - \alpha} \quad \text{for } \alpha \leq x \leq \beta .$$

Suppose that $\theta = \alpha/(\alpha + \beta)$ is the parameter of interest. Let $X_{(1)}, ..., X_{(n)}$ be the order statistic.

(a) Show that $(X_{(1)}, X_{(n)})$ is a sufficient for making inference about any function of α and β.
(b) Obtain the joint distribution of $(X_{(1)}, X_{(n)})$.
(b) Construct generalized confidence intervals for the parameter $\theta = \alpha/(\alpha + \beta)$.
(d) Does this interval have the repeated sampling property?

2.8 Consider again the problem of sampling from the uniform distribution given in the previous exercise. Deduce from the above results generalized procedures for testing hypotheses concerning the parameter θ.

2.9 Let $Y_1, \ldots, Y_n$ be a random sample from a distribution with the density function

$$f(y) = \frac{\lambda\mu}{\lambda - \mu}\frac{1}{y^2} \text{ for } \mu \leq y \leq \lambda,$$

where λ and μ are positive parameters.

(a) Find two sufficient statistics for making inferences about any function of (λ, μ).
(b) By considering the distribution of $X = 1/Y$ or otherwise, find a generalized pivotal quantity for making inferences about the parameter $\theta = \lambda/(\lambda - \mu)$.
(c) Find $100\gamma\%$ generalized confidence intervals for θ.
(d) Find the generalized p-value for testing left-sided and right-sided null hypotheses concerning the parameter θ.

2.10 Suppose $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent sets of independent exponential random variables with means α and β, respectively. Establish generalized procedures for making inferences about the difference in means, $\theta = \alpha - \beta$.

2.11 Let $Y_1, \ldots, Y_n$ be a random sample from the two-parameter exponential distribution

$$f(y) = \frac{1}{\beta}e^{-(y-\alpha)/\beta} \quad \text{for } \alpha < y \text{ , } \beta > 0.$$

(a) Express the mean μ and the variance σ^2 of the distribution in terms of α and β.
(b) Find a set of minimal sufficient statistics for making inferences about functions of the parameters of the distribution.
(c) Establish procedures for making inferences about the mean μ.
(d) Establish procedures for making inferences about the variance σ^2.

CHAPTER 3

METHODS IN ANALYSIS OF VARIANCE

3.1 INTRODUCTION

Many biomedical, socioeconomic, and industrial experiments, and even business and marketing trials, involve comparison of two or more populations. Data from such experiments may pertain to observations taken at a single point in time or repeated measures taken over time. In either case, appropriate statistical methods available for analysis of data from such experiments include a procedure known as the *Analysis of Variance*, which is abbreviated as ANOVA. The purpose of this chapter is to introduce some widely used procedures for comparing a number of univariate normal populations. Later in Chapters 5−10, we will extend the main results presented in this chapter to the case of multivariate normal populations under various models. Here we confine our attention primarily to the problem of comparing population means. The readers interested in inferences concerning other parameters such as the variance and the reliability parameter are referred to Weerahandi (1995). This chapter also presents some latest developments in the problem of comparing normal means.

(Generalized Inference in Repeated Measures, Edition 2). By (Weerahandi)

In practical applications, often we need to compare several population means. In biomedical experiments, one often has to deal with the problem of comparing available and potential treatments for a certain disease. The problem also arises in almost every field of research including industrial experiments, agricultural experiments, socioeconomic experiments, and so on. For example, an investigator might wish to compare several types of a product, or some diet plans, or several brands of fertilizers, or some methods of teaching, or advertising methods, and so on. When the means are computed using samples of data taken from the populations being compared, almost always the sample means will be different of course regardless of whether the population means are equal or not. The question that we need to answer is whether or not such differences are due to real differences in the population means or they are just artifacts of sampling variation. Once we conclude that population means are different, then we also need to address the question that how different the means are.

3.2 COMPARING TWO POPULATION MEANS

The purpose of this section is to provide a brief overview of some important results available in the literature for comparing two normal populations. The results valid for the two-sample problem will prove to be useful even in multiple comparisons involving a number of normal populations. Here we consider only the problem of comparing the means of two normal populations. For details of the problem and for related problems concerning two populations, such as the problem of comparing the variances of two normal populations or comparing the means of two exponential populations, the reader is referred to Weerahandi (1995).

Consider two populations that we would like to compare. Let

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$$

be a random sample of size n_1 from the first population and let

$$Y_{21,} Y_{22}, \ldots, Y_{2n_2}$$

be a random sample of size n_2 from the second population. Assume that the data are normally distributed. More specifically assume that

$$Y_{1j} \sim N(\mu_1, \sigma_1^2);\ j = 1, \ldots, n_1 \tag{3.1}$$

$$Y_{2j} \sim N(\mu_2, \sigma_2^2);\ j = 1, \ldots, n_2\ . \tag{3.2}$$

Let

$$\overline{Y}_1 = \frac{\sum Y_{1j}}{n_1},\ \overline{Y}_2 = \frac{\sum Y_{2j}}{n_2}$$

and

$$S_1^2 = \frac{\sum (Y_{1j} - \overline{Y}_1)^2}{n_1}, S_2^2 = \frac{\sum (Y_{2j} - \overline{Y}_2)^2}{n_2}$$

be the sample means and sample variances of the two data sets, which are the Maximum Likelihood Estimates for the population means and variances. It is well known that $(\overline{Y}_1, \overline{Y}_2, S_1^2, S_2^2)$ is a set of sufficient statistics for the parameters of the populations, namely for $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. The distributions of these statistics are given by

$$\overline{Y}_i \sim N(\mu_i, \frac{\sigma_i^2}{n_i}),\ i = 1, 2 \tag{3.3}$$

and

$$\frac{n_i S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1}\ ,\ i = 1, 2\ . \tag{3.4}$$

Moreover, these four random variables are mutually independent.

3.2.1 Case of equal variances

Assume first that the variances of the two populations are equal, an assumption that we will relax later in this chapter. Let σ^2 be the common variance. In this case the set of sufficient statistics further reduces to $(\overline{Y}_1, \overline{Y}_2, S^2)$, where

$$S^2 = \frac{\sum(Y_{1j} - \overline{Y}_1)^2 + \sum(Y_{2j} - \overline{Y}_2)^2}{n_1 + n_2 - 2}$$

is referred to as the pooled unbiased sample variance. Its distribution that follows from (3.4) is

$$\frac{(n_1 + n_2 - 2)S^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}.$$

Inferences on the difference in the two means can be based on the random variable $T_\delta = (\overline{Y}_1 - \overline{Y}_2 - \delta)/S$, where $\delta = \mu_1 - \mu_2$. Since

$$(\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)) \sim N(0, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})),$$

its distribution is given by

$$\frac{T_\delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}\ . \tag{3.5}$$

This result can be employed to construct testing procedures and confidence intervals for δ. First consider the problem of testing hypotheses of the form

$$H_0 : \mu_1 - \mu_2 \leq \delta_0, \tag{3.6}$$

where δ is a specified constant. Clearly $T = T_{\delta_0}$ is a test statistic appropriate for testing (3.6), because its distribution is free of unknown parameters and

its distribution is stochastically increasing in δ. Hence, the p-value for testing H_0 is computed as

$$\begin{aligned} p &= \sup_{\delta \le \delta_0} \Pr(T \ge t_{obs}) \\ &= 1 - \Pr(T < \frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{s}) \\ &= 1 - G_{n_1+n_2-2}(\frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}), \end{aligned} \tag{3.7}$$

where $(\overline{y}_1, \overline{y}_2, s)$ is the set of observed values of $(\overline{Y}_1, \overline{Y}_2, S)$ and $G_{n_1+n_2-2}$ is the cumulative distribution function (cdf) of the t distribution with n_1+n_2-2 degrees of freedom. The null hypothesis is rejected if the p-value is too small, say less than a certain nominal value such as 0.05. The p-value for testing two-sided hypotheses of the form $H_0 : \mu_1 - \mu_2 = \delta_0$ can be obtained in a similar manner as

$$p = 2G_{n_1+n_2-2}(-\frac{|\overline{y}_1 - \overline{y}_2 - \delta_0|}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}). \tag{3.8}$$

To outline the approach to constructing various confidence intervals, consider the problem of finding equal-tail confidence intervals for the difference in the two population means, $\delta = \mu_1 - \mu_2$. Let t_ν denote the νth quantile of the t distribution with $n_1 + n_2 - 2$ degrees of freedom. To construct $100\gamma\%$ confidence intervals, consider the probability statement

$$\begin{aligned} \gamma &= \Pr(-t_{\frac{1+\gamma}{2}} \le \frac{T_\delta}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le t_{\frac{1+\gamma}{2}}) \\ &= \Pr(-t_{\frac{1+\gamma}{2}} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \overline{Y}_1 - \overline{Y}_2 - \delta \le t_{\frac{1+\gamma}{2}} S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}), \end{aligned} \tag{3.9}$$

which follows from the definition of the quantiles. It is now evident from (3.9) that the $100\gamma\%$ equal tail confidence interval for δ is

$$\left[(\overline{y}_1 - \overline{y}_2) - t_{\frac{1+\gamma}{2}} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \delta \le (\overline{y}_1 - \overline{y}_2) + t_{\frac{1+\gamma}{2}} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{3.10}$$

Similarly, right-sided confidence intervals for δ can be obtained using the formula

$$\left[\delta \le (\overline{y}_1 - \overline{y}_2) + t_\gamma s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{3.11}$$

Example 2.1. Testing the effectiveness of a promotional campaign

In order to quantify the effect of a promotional display of a product, weekly sales data are obtained from 10 stores before and after the start of the promotion. These stores are referred to as test stores. Any increase in sales from one week to another could occur due to reasons other than the promotion, say due to improved weather conditions in the second week. Therefore, weekly sales are obtained from a sample of control stores without the promotion as well. Logarithmic weekly sales from the test stores and the control are shown in Table 2.1.

Table 3.1 Log sales by promotion

Test stores		Control stores	
Before	After	Before	After
8.2	11.7	24.5	23.1
10.1	12.4	17.5	16.5
13.5	12.8	21.3	21.7
7.0	10.2	16.5	18.2
13.6	12.9	16.2	15.9
8.6	11.8	13.3	14.1
11.1	11.8	8.7	9.8
13.3	12.5	20.6	19.1
13.9	14.1	17.9	18.9
10.3	12.6	25.6	24.9

This is in fact a particular case of repeated measures experiments that we will study in great detail in Chapter 5−10. Nevertheless, in the case of two periods as in this example, the data can be analyzed by applying the methods in this section under the assumption that the log sales data are normally distributed, a reasonable assumption as often seen in practical applications. This is accomplished by analyzing the increase in store sales during the promotional period. This is indeed appropriate thing to do, because the mean increase in sales that can be attributed to the promotion is the quantity of primary interest. Table 2.2 shows the two sets of increased sales figures that are appropriate for analysis by two sample methods.

Table 3.2 Increase in log sales

Test stores:	3.5, 2.3, -0.7, 3.2, -0.7, 3.2, 0.7, -0.8, 0.2, 2.3
Control stores:	-1.4, -1.0, 0.4, 1.7, -0.3, 0.8, 1.1, -0.7, 1.0, -0.7

Let μ_t and μ_c be the change in mean log sales for the test and control sales, respectively. Their estimates given by the sample means are 1.32 and 0.09, respectively, indicating the possibility that the promotion has worked.

The question is whether or not this is an artifact of sampling variation of the data across stores. To establish the statistical significance of the effectiveness of the promotion, let us apply the t-test given by (3.7). Nevertheless, it should be noted here that the assumption of equal variances might not be reasonable, because the sample variances of the two sets of data are 2.80 and 0.99, respectively, and therefore we will revisit the problem later in this chapter. If we formulate the null hypothesis as $H_0 : \mu_t - \mu_c = 0$, then the p-value appropriate for testing the hypothesis can be computed using (3.8). The resulting p-value is 0.07, indicating some evidence against the null hypothesis. In fixed-level testing, the result is not significant at the 0.05 level, but it is significant at the 0.1 level. The 95% equal-tail confidence interval of $\delta = \mu_t - \mu_c$ computed from (3.10), namely the interval [-0.1331, 2.593], implies the same conclusion. If we formulate the hypothesis as $H_0 : \mu_t - \mu_c \leq 0$ of course, the $p-$value computed from (3.7) becomes 0.037, which allows us to reject the null hypothesis at the 0.05 level of fixed-level testing.

3.3 CASE OF UNEQUAL VARIANCES

Now let us drop the assumption of equal variances, and consider again the problem of comparing the means of two normal populations. The problem of finding inference procedures based on sufficient statistics in this context is often referred to as the Behrens−Fisher problem. Tsui and Weerahandi (1989) solved this problem by taking the generalized p-value approach and their result is also equivalent to that of Bernard (1984). It is also numerically equivalent to the Bayesian solution under the natural non-informative prior. A formal derivation of the generalized t-test for the Behrens−Fisher problem could be found in Weerahandi (1995). Here we provide an intuitive argument using the substitution method.

Notice first of all from (3.3) that if the two variances were known, we would base our inference procedures on the result

$$Z = \frac{\overline{Y}_1 - \overline{Y}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1), \tag{3.12}$$

where $\delta = \mu_1 - \mu_2$. When the variances are unknown, as usually the case, we cannot compute the Z statistic. In this case we can tackle the unknown variances by taking advantage of the distributional result

$$V_i = \frac{n_i S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1}.$$

The substitution approach suggests that, when the variances are unknown, the p-value should be computed using the formula

$$\Pr\left(\frac{\overline{Y}_1 - \overline{Y}_2 - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{\sqrt{\frac{s_1^2}{Y_1} + \frac{s_2^2}{Y_2}}}\right).$$

To establish this result, consider the test variable given by the substitution method

$$\begin{aligned} T &= (\overline{y}_1 - \overline{y}_2) - Z\sqrt{\frac{s_1^2}{V_1} + \frac{s_2^2}{V_2}} - \delta \\ &= (\overline{y}_1 - \overline{y}_2) - \frac{(\overline{Y}_1 - \overline{Y}_2 - \delta)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\sqrt{\frac{\sigma_1^2 s_1^2}{n_1 S_1^2} + \frac{\sigma_2^2 s_2^2}{n_2 S_2^2}} - \delta \end{aligned} \tag{3.13}$$

with the observed value $t_{obs} = 0$. It is clear that the distribution of T and its observed value are both free of nuisance parameters. Moreover, it is stochastically decreasing in δ, the parameter of interest. Hence, the p-value for testing left-sided null hypotheses of the form $H_0 : \mu_1 - \mu_2 \leq \delta_0$ can be based on the generalized p-value

$$\begin{aligned} p &= \operatorname*{Sup}_{\delta \leq \delta_0} \Pr(T \leq 0) \\ &= \operatorname*{Sup}_{\delta \leq \delta_0} \Pr(Z\sqrt{\frac{s_1^2}{V_1} + \frac{s_2^2}{V_2}} \geq \overline{y}_1 - \overline{y}_2 - \delta) \\ &= \Pr(Z \geq \frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{\sqrt{\frac{s_1^2}{V_1} + \frac{s_2^2}{V_2}}}), \end{aligned} \tag{3.14}$$

as expected. Although this representation of the p-value is sufficient for numerical computation of the p-value, the dimension of numerical integration can be reduced by defining a beta random variable and an independent chi-squared random variable as

$$B = \frac{V_1}{V_1 + V_2} \sim \text{Beta}(\frac{n_1 - 1}{2}, \frac{n_2 - 1}{2}),\ V = V_1 + V_2 \sim \chi^2_{n_1+n_2-2}. \tag{3.15}$$

In terms of B and Y random variables we can express the p-value as

$$\begin{aligned} p &= \Pr(Z \geq \frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{\sqrt{\frac{s_1^2}{BV} + \frac{s_2^2}{(1-B)V}}}) \\ &= \Pr\left(\frac{Z}{\sqrt{V/(n_1 + n_2 - 2)}} \geq \frac{\overline{y}_1 - \overline{y}_2 - \delta_0}{\sqrt{(\frac{s_1^2}{B} + \frac{s_2^2}{(1-B)})/(n_1 + n_2 - 2)}}\right) \\ &= 1 - EG_{n_1+n_2-2}\left((\overline{y}_1 - \overline{y}_2 - \delta_0)\sqrt{\frac{n_1 + n_2 - 2}{\frac{s_1^2}{B} + \frac{s_2^2}{(1-B)}}}\right), \end{aligned} \tag{3.16}$$

where $G_{n_1+n_2-2}$ is the cdf of the t-distribution with $n_1 + n_2 - 2$ degrees of freedom and the expectation is taken with respect to the beta random variable

B. The test based on this p-value is referred to as the generalized t-test. Apart from being a one dimensional integration, this representation of the p-value has further advantages. In particular, it has a form similar to that of the classical F-test, and the integration that is over the [0,1] interval as opposed to the $[0,\infty)$ interval is better behaved.

The counterpart of the p-value in interval estimation can be deduced directly from the p-value itself or derived from the test variable defined by (3.13). It is easily seen that the $100\gamma\%$ left-sided generalized confidence interval and the $100\gamma\%$ equal-tail generalized confidence interval obtained in this manner are of the form

$$\delta \geq (\overline{y}_1 - \overline{y}_2) - c_\gamma(s_1^2, s_2^2) \tag{3.17}$$

and

$$\left[(\overline{y}_1 - \overline{y}_2) - c_{\frac{1+\gamma}{2}}(s_1^2, s_2^2) \leq \delta \leq (\overline{y}_1 - \overline{y}_2) + c_{\frac{1+\gamma}{2}}(s_1^2, s_2^2)\right], \tag{3.18}$$

respectively, where $c_\nu(s_1^2, s_2^2)$ is the solution of the equation

$$EG_{n_1+n_2-2}(c_\nu \sqrt{\frac{n_1+n_2-2}{\frac{s_1^2}{B} + \frac{s_2^2}{(1-B)}}}) = \nu \ , \tag{3.19}$$

where the expectation is taken with respect to the Beta random variable B.

Akahira (1999) provided a formula for actual size of confidence intervals of above form with more general $c(s_1^2, s_2^2)$ having the property $c(ks_1^2, ks_2^2) = \sqrt{k}c(s_1^2, s_2^2)$ for positive real number k. For example, the actual size of confidence intervals of the form (3.17), in terms of $\rho = \frac{\sigma_1^2/n_1}{\sigma_1^2/n_1+\sigma_2^2/n_2}$ is given by

$$\begin{aligned} p(\rho) &= \Pr((\overline{Y}_1 - \overline{Y}_2) - c(S_1^2, S_2^2) \leq \delta) \\ &= EG_{n_1+n_2-2}(\sqrt{n_1+n_2-2}c(\rho B, (1-\rho)(1-B))) \ . \end{aligned} \tag{3.20}$$

Akahira (1999) discussed some simple functions $c(s_1^2, s_2^2)$, which ensured the confidence level; i.e., $p(\rho) \geq \gamma$ for all $\rho \in (0,1)$. According to empirical results, $c(s_1^2, s_2^2)$ given by (3.19) also satisfies the desired property $p(\rho) \geq \gamma$. This means that, while the above intervals are exact generalized confidence intervals, they also preserve the confidence level in the classical sense; i.e., the actual level of generalized confidence intervals is at least as large as the intended level.

Example 2.2. Testing the effectiveness of a promotional campaign (continued)

Recall that in Example 2.1, in order to apply the classical t-test we had to make the assumption that the variances of change in log sales data are the same for the test stores and the control stores. The estimated variances of 2.80 and 0.99 suggest that the assumption is not a reasonable one. Now we are in

a position to test the hypotheses of interest without making any assumption on the variances. The p-value for testing the null hypothesis $H_0 : \mu_t - \mu_c \leq 0$ as given by (3.19) is 0.046. This p-value also allows us to reject the null hypothesis at the 0.05 level. The advantage of this p-value is that we can come to the same conclusion without relying on an unreasonable assumption. The equal-tail confidence interval for $\delta = \mu_t - \mu_c$ in the unequal variances case is [-0.2297 , 2.69], suggesting that the result is not significant if the hypothesis is formulated as $H_0 : \mu_t - \mu_c = 0$. In this type of applications, the length of the confidence interval can be tightened and hence the type I error can be further reduced by taking a larger sample of control stores.

3.4 ONE-WAY ANOVA

Now we are in a position to undertake the simplest case of ANOVA where we have k univariate populations to compare. The populations will be sometimes referred to as the treatment groups. As in the previous section, the population means are the quantities of primary interest. The means of the populations are not necessarily equal, and the problem of primary interest is to test the equality of the means. For a related problem of making inferences about the common mean of several normal populations, the reader is referred to Krishnamoorthy and Lu (2003).

In testing the equality of several means, it is tempting to carry out pairwise comparisons based on results from the previous section. The main drawback of this approach is that it will seriously increase the chance of rejecting the hypothesis of equal means even when it is true. In fact if one performs a large number of pairwise comparisons, say at the 0.05 level, the chances are that some differences will come out to be significant even when the true means are all equal. For example, consider the problem of comparing means of four populations. If we are to test the hypothesis that all four means are equal by pairwise comparison of means, we will have to run a sequence of tests on each of the six possible pairs of means. If we perform each test at fixed size 0.05, the size of the combined test will be considerably larger than 0.05. The larger the numbers of populations being compared, the greater the chance of making such misleading conclusions. The point is that, when we perform a large number of comparisons, the chance of observing a pair of significantly different sample means can be fairly large even when there are no differences in the population means. Therefore, even though the ultimate goal might be to identify the population with the largest significant mean, the hypothesis of equal means should be first tested by a single testing procedure that takes advantage of all the information in the data.

Suppose we have a sample of size n_i from Population i. Let $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ denote the sample of data available from ith population. Sometimes the random variable Y will be referred to as the *response variable*. Let μ_i be the mean of the ith population and let σ_i^2 be its variance. Assume further

that each population is normally distributed so that $Y_{ij} \sim N(0, \sigma_i^2)$. The underlying linear model can also be written as

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij}, \\ i &= 1, \ldots, k, \; j = 1, \ldots, n_i, \end{aligned} \tag{3.21}$$

where ϵ_{ij} is a residual term distributed as

$$\epsilon_{ij} \sim N(0, \sigma_i^2). \tag{3.22}$$

Now the problem is to compare the μ_i's appearing in (3.21) based on the sample data. Before we do pairwise comparisons or inferences on any other function of the means, it is important that first we do one test to see whether or not the means are different at all. That is, we need to test whether or not we have sufficient evidence to reject the null hypothesis that population means are all equal. More specifically, the problem is to test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \tag{3.23}$$

against the alternative H_1 : not all the population means are equal. If the null hypothesis can be rejected at a small risk of making a wrong conclusion, which is also known as the false positive error or Type I error, the differences in means are said to be statistically significant. It is important that first we establish the statistical significance of the difference in means, because when we go about doing various multiple comparisons, the larger the number of comparisons we do, the greater the false positive error we will commit.

It follows from one sample inference, or is easily seen directly from the likelihood function, that the maximum likelihood estimates (abbreviated as MLE) of the means and the variances are the sample means and variances given by

$$\begin{aligned} \widehat{\mu}_i &= \overline{Y}_i \\ &= \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \end{aligned} \tag{3.24}$$

and

$$\begin{aligned} \widehat{\sigma}_i^2 &= S_i^2 \\ &= \frac{\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)}{n_i}, \end{aligned} \tag{3.25}$$

respectively. Unlike the MLE μ_i, the MLE of σ_i^2 is not quite unbiased in the classical sense, but $\widetilde{S}_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)/(n_i - 1)$ is unbiased. It is also known from one sample inference that the distributions of MLEs are given by

$$\overline{Y}_i \sim N(\mu_i, \frac{\sigma_i^2}{n_i}) \tag{3.26}$$

and

$$n_i \, S_i^2/\sigma_i^2 \sim \chi^2_{n_i \, -1}, \tag{3.27}$$

respectively. Moreover, the statistics $\overline{Y}_i$ and S_i^2 are independently distributed and they are sufficient for making inferences about all unknown parameters of the problem. Yet, testing the null hypothesis (3.23) based on these sufficient statistics is not a trivial matter.

3.4.1 Case of equal variances

In the classical treatment of the above statistical problem, it is assumed that the population variances are all equal, an assumption made for simplicity and mathematical tractability of the classical approach to solving problems of hypotheses testing. It is not really a natural assumption. In fact, it is often seen in real-world applications that the variances tend to be substantially different especially when the means are substantially different. It has also been observed that the assumption of equal variances is much more serious than the assumption of normally distributed populations, in that the former has greater chance of leading to wrong conclusions. It should also be pointed out that in most applications, despite a common belief, it is not possible to transform data to achieve the approximate normality and equal variances simultaneously.

Nevertheless, there are situations where the variances are not very different to impact the false positive error or the power of a test. Moreover, first tackling the simpler problem will give us insight into the approach that we could take in handing the more difficult problem posed by unequal variances. So, let us first consider the problem of testing (3.23) when all the populations have a common variance, say σ^2. In this case (3.26) and (3.27) imply that the random variables

$$\overline{Y}_i \sim N(\mu_i, \frac{\sigma^2}{n_i}), \;\; i = 1, \ldots, k$$

and

$$\sum_{i=1}^{k} n_i \, S_i^2/\sigma^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2/\sigma^2 \sim \chi^2_{N \, - \, k}, \tag{3.28}$$

are independently distributed, where $N = \sum_{i=1}^{k} n_i$. Define

$$S_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2,$$

an important statistic, which we will refer to as the *error sum of squares* or *within group sum of squares*. It follows from (3.28), or is seen directly from

$$E(S_E) \;=\; \sum_{i=1}^{k} n_i E(S_i^2) \;=\; (N \;-\; k)\sigma^2,$$

that

$$\text{MSE} \;=\; \frac{S_E}{N \;-\; k} \tag{3.29}$$

is an unbiased estimator of the error variance σ^2.

In ANOVA, the decomposition of the total sum of squares,

$$S_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2,$$

into independent components will prove to be insightful and helpful in deriving appropriate statistical tests for a variety of linear models we will encounter in the following chapters including the analysis of repeated measures. The decomposition in the present problem is

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2 + \sum_{i=1}^{k} n_i(\overline{Y}_i \;-\; \overline{Y})^2, \tag{3.30}$$

which is an implication of the identity,

$$Y_{ij} - \overline{Y}_i = (Y_{ij} - \overline{Y}_i) + (\overline{Y}_i \;-\; \overline{Y}).$$

The sum of squares,

$$S_B = \sum_{i=1}^{k} n_i(\overline{Y}_i \;-\; \overline{Y})^2, \tag{3.31}$$

is referred to as the *between group sums of squares*. Sometimes it is also called the *among group sum of squares*. In terms of the foregoing sums of squares, the total sum of squares can be expressed as

$$S_T = S_B + S_E. \tag{3.32}$$

In various types of ANOVA we will encounter in the development below and in the following chapters, the sums of squares of this nature, and the decomposition of total sum of squares into a number of components as in (3.32) will play an important role in the analysis.

Example 2.3. Comparing the mean strength of reinforcing bars

An engineer at a construction company wishes to compare four brands of

Table 3.3 Strength of four brands of reinforcing bars

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

reinforcing bars for their strength. The four brands of reinforcing bars were tested for their strength, and results were reported in Table 2.3 in certain units.

In this example the strength of reinforcing bars is the variable of interest. The problem is to compare the strengths of brands in terms of mean values of this variable. Table 2.4 displays the summary statistics, namely the sample sizes, sample means, and sample standard deviations computed from the data.

Table 3.4 Sample statistics

Population	n_i	$\overline{y}_{i.}$	s_i
Brand A	7	17.96	3.07
Brand B	8	20.68	5.28
Brand C	7	18.10	1.39
Brand D	9	20.83	1.48

In this application, the number of populations being compared is $k = 4$ and the total sample size is $N = 31$. By applying the above equations, the error sum of squares and the between sum of squares can be computed using the information given in Table 2.2 as $s_E = 322.526$ and $s_B = 57.638$, respectively. The total sum of squares is then $s_T = s_E + s_B = 380.164$.

To establish procedures for testing the null hypothesis H_0, note that the expected value S_B can be expressed as

$$\begin{aligned} E(S_B) &= E(\sum_{i=1}^{k} n_i(\overline{Y}_i - \overline{Y})^2) \\ &= E\sum_{i=1}^{k} n_i\left((\mu_i - \overline{\mu}) + (\overline{\epsilon}_i - \overline{\epsilon})\right)^2 \\ &= \sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2 + (k-1)\sigma^2, \end{aligned} \tag{3.33}$$

where $\overline{\mu} = \sum_{i=1}^{k} n_i\mu_i/N$. Hence, depending on H_0 is true or not, the expected value of the mean between group sum of squares

$$\text{MSB} = \frac{S_B}{k-1} \tag{3.34}$$

is equal to the error variance σ^2 or greater than σ^2. The larger the discrepancy between the means, the greater the deviation of the expected value from σ^2 tends to be. The results (3.29) and (3.34) provide a basis for deriving an unbiased test of H_0. To compute the p-value of the resulting test, we also need the distributions of S_E and S_B. From (3.28), it is evident that the error sum of squares S_E is related to the chi-squared distribution

$$S_E/\sigma^2 \sim \chi^2_{N\,-\,k},$$

whereas from (3.33) and from results known for one sample inference we can deduce that the between group sum of squares s S_B is related to the non-central chi-squared distribution

$$S_B/\sigma^2 \sim \chi^2_{N\,-\,k}(\delta),$$

where $\delta = \sum_{i=1}^{k} n_i(\mu_i-\overline{\mu})^2/((k-1)\sigma^2)$ is the noncentrality parameter, a quantity that becomes 0 under H_0, thus reducing the above distribution to a central chi-squared distribution. The results can also be derived from (3.30) and the properties of the normal distribution. Moreover, (3.30) implies that these statistics are independently distributed. Hence, under H_0, we have

$$F = \frac{S_B/(k-1)}{S_E/(N-k)} \sim F_{k-1,N-k}. \tag{3.35}$$

This means that under the null hypothesis the ratio of the two mean sums of squares has an F distribution with $k-1$ and $N-k$ degrees of freedom (abbreviated as DF). If the null hypothesis is not true, then the distribution is a noncentral F distribution with $k-1$ and $N-k$ degrees of freedom and noncentrality parameter δ. That is, under H_1,

$$F \sim F_{k-1,N-k}(\delta).$$

This means that the greater the deviation of the individual means from one another, the larger the F-statistic tends to be. On the other hand, the left-tail probabilities corresponding to the observed value of F would take small values under H_1 than under H_0, thus suggesting the rejection of the null hypothesis for smaller values of such tail probabilities. Hence, the left-tail probabilities of the F distribution can be used to base unbiased tests for H_0.

The p-value of the F-test suggested by the above observations can be computed as

$$p = \Pr\left(\frac{S_B/(k-1)}{S_E/(N-k)} \geq \frac{s_B/(k-1)}{s_E/(N-k)}\right)$$
$$= 1 - H_{k-1,N-k}\left(\frac{s_B/(k-1)}{s_E/(N-k)}\right), \tag{3.36}$$

where s_B and s_E are the observed values of the sums of squares S_B and S_E, respectively, and $H_{k-1,N-k}$ is the cumulative distribution function (cdf) of the F distribution with $k-1$ and $N-k$ degrees of freedom. The null hypothesis is rejected for small values of p, say $p < 0.05$, in testing at the fixed-level 0.05.

In ANOVA, it is customary and insightful to set out various quantities leading to this F-statistic in an Analysis of Variance Table as displayed below in Table 2.5. In the ANOVA table, the sum of squares and the mean sums of squares columns are abbreviated as SS and MS, respectively.

Table 3.5 ANOVA table

Source	DF	SS	MS	F-Statistic
Between Groups	$k-1$	S_B	$S_B/(k-1)$	MSB/MSE
Error	$N-k$	S_E	$S_E/(N-k)$	
Total	$N-1$	S_T		

Example 2.4. Comparing the mean strength of reinforcing bars (continued)

Continuing from Example 2.3 let $\mu_i,\ i = 1,2,3,4$ be the mean strengths of four brands of reinforcing bars. Consider the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. Assume that the variances of the strength distribution for the four brands are equal. We can apply the foregoing results to test whether or not the data in Table 2.3 supports the null hypothesis. Using the sums of squares (SS) computed in Example 2.3, we can compute the mean sums of squares (MS) and the F-statistic. The resulting ANOVA table is displayed in Table 2.6.

Under the assumption of equal error variances we can use the F-test to compare the four brands. The 95th percentile of the F distribution with 3 and 27 degrees of freedom is 2.96. Therefore, the null hypothesis of equal means cannot be rejected at the 0.05 level. The p-value computed from 3.35 is $p = 1 - H_{3,27}(1.6084) = 0.211$, which leads us to conclude that the data does not provide sufficient evidence to doubt the null hypothesis H_0. Figure

Table 3.6 ANOVA for comparing the reinforcing bars

Source of	DF	SS	MS	F-Statistic
Between	3	57.638	19.213	1.6084
Error	27	322.526	11.945	
Total	30	380.164		

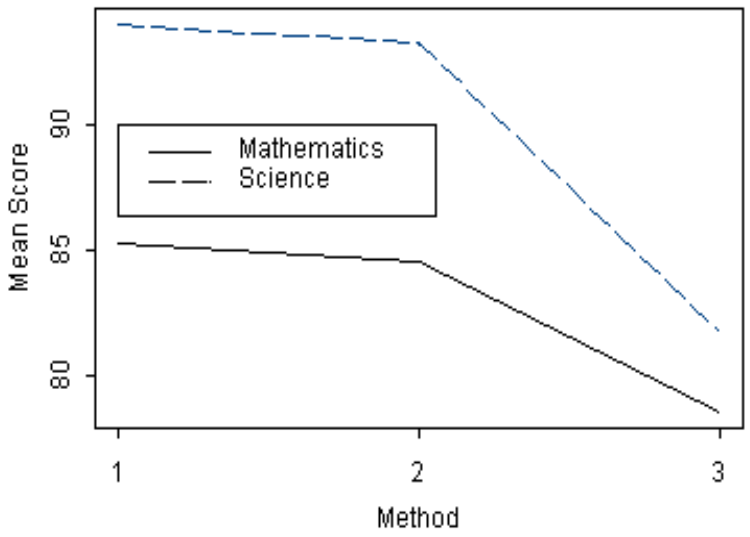

Figure 3.1 Error bars for the four brands

2.1 provides error bars for each of the four brands. The midpoint of each bar represents the corresponding sample mean $\overline{y}$, and the half-width of the bar represents its standard error, $s/\sqrt{n}$. Although the distributions are not well separated, notice that error bar of Brand D falls fairly above those of Brands A and B suggesting that perhaps the difference in means is significant. Indeed in this application, the assumption of equal variances is not reasonable and so the above results lack credibility. Therefore, we will revisit this problem again in the next section and try to perform the mean comparison without that assumption.

3.4.2 Case of unequal variances

The p-value given above is valid only if the variances are equal, and the test is not appropriate if the variances are significantly different. But, in many

situations this assumption is not reasonable. As demonstrated later in this section, the classical F-test can lead to very serious repercussions if applied when the assumption is not reasonable. There are numerous tests available in the literature [cf. Krutchkoff (1988, 1989)] that do not rely on the assumption of equal variances. Here we confine our attention only to those tests that are based on an exact probability statements and sufficient statistics. Weerahandi (1994) generalized the F-test to be valid for the unequal variances case, and he argued that it is also equivalent to the test given by Rice and Gains (1989), who extended an argument due to Bernard (1984). The particular representation of the test introduced by Weerahandi (1994) is referred to as the generalized F-test.

3.4.3 Substitution method in ANOVA

Finding generalized tests in ANOVA problems can also be accomplished by a variation of the substitution method introduced in Chapter 1. This approach requires that a test is available from the classical theory for the case of known nuisance parameters (usually the unknown variances), say $(\alpha_1, \alpha_2, \ldots, \alpha_k)$, . It also requires that the nuisance parameters can be expressed in terms of a set of sufficient statistics $(X_1, X_2, \ldots, X_k)$ and a set random variables $(V_1, V_2, \ldots, V_k)$ with distributions free of unknown parameters. Though desirable, it is not necessary that these random variables are mutually independent. However, their joint distribution is assumed to be known. Then the substitution method is carried out in the following steps:

- From the classical ANOVA, obtain the test statistic (usually a between group sum of squares), say $S(\alpha_1, \alpha_2, \ldots, \alpha_k)$, that is unbiased for testing the null hypothesis of interest when the nuisance parameters are known. Suppose it tends to take larger values for deviations from the null hypothesis.

- Express the nuisance parameters $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ in terms of the sufficient statistics $(X_1, X_2, \ldots, X_k)$ and the random variables

$$\mathbf{V} = (V_1, V_2, \ldots, V_k).$$

- Replace the statistics $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ by their observed values $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ and define a generalized test variable T as

$$T = S(\alpha_1, \alpha_2, \ldots, \alpha_k) - s(\mathbf{V}; \mathbf{x}),$$

  where s is the observed value of S.

- Show that $\{T \geq 0\}$ is a proper generalized extreme region.

- Compute the generalized p-value as the probability of the generalized extreme region.

To present the generalized F-test in One-Way ANOVA under unequal variances, consider again the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

under the weaker assumptions,

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij}, \\ \text{where } \epsilon_{ij} &\sim N(0, \sigma_i^2), \quad i = 1, \ldots, k, \ j = 1, \ldots, n_i, \end{aligned} \tag{3.37}$$

where σ_i^2 are the variances of the populations. In view of (3.31), define the standardized between group sum of squares

$$\widetilde{S}_B = \widetilde{S}_B(\sigma_1^2, \ldots, \sigma_k^2) = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} (\overline{Y}_i - \overline{Y}_\sigma)^2 , \tag{3.38}$$

where

$$\overline{Y}_\sigma = \frac{\sum_{i=1}^{k} n_i \overline{Y}_i / \sigma_i^2}{\sum_{i=1}^{k} n_i / \sigma_i^2}.$$

This is the between group sum of squares that we would have used to derive a test for H_0 if the variances were known. Under the null hypothesis of equal means we have

$$\widetilde{S}_B \sim \chi^2_{k-1}. \tag{3.39}$$

The substitution method allows us to use this quantity to obtain testing procedures when the variances are unknown. This is possible because chi-squared distributions

$$Y_i = n_i S_i^2 / \sigma_i^2 \sim \chi^2_{n_i - 1} \tag{3.40}$$

relate them to the observable sufficient statistics. The substitution approach indicates that the appropriate procedure is to replace σ_i^2 by $n_i s_i^2 / Y_i$ and then construct the extreme region. Let $\widetilde{s}_B$ be the observed value of $\widetilde{S}_B$ obtained by replacing σ_i^2 by $n_i s_i^2 / Y_i$, $i = 1, \ldots, k$. This amounts to using the well-defined extreme region

$$\left\{ \mathbf{Y} | \widetilde{S}_B(\sigma_1^2, \ldots, \sigma_k^2) \geq \widetilde{s}_B(\frac{s_1^2}{S_1^2}\sigma_1^2, \ldots, \frac{s_k^2}{S_k^2}\sigma_k^2) \right\}.$$

Notice that the probability of this subset of the sample space increases for any deviation from the null hypothesis and that the observed sample point falls on the boundary of the subset. Hence, the p-value for testing the null hypothesis of equal means can be obtained as

$$p = \Pr(\widetilde{S}_B \geq \widetilde{s}_B(\frac{n_1 s_1^2}{Y_1}, \frac{n_2 s_2^2}{Y_2}, \ldots, \frac{n_k s_k^2}{Y_k}) = 1 - EG_{k-1}(\widetilde{s}_B(\frac{n_1 s_1^2}{Y_1}, \frac{n_2 s_2^2}{Y_2}, \ldots, \frac{n_k s_k^2}{Y_k})), \quad (3.41)$$

where G_{k-1} is the cdf of the chi-squared distribution with $k-1$ degrees of freedom and the expectation is taken with respect to the independent Y_i random variables. The p-value serves to measure the evidence in favor of H_0. The p-value given by (3.41) does not quite have the form of the classical F-test. However, by change of variables (see Appendix for a derivation) the p-value can be expressed to take the form of a generalized F-test as

$$p = 1 - E(H_{k-1,N-k}(\frac{N-k}{k-1}\widetilde{s}_B[\frac{n_1 s_1^2}{B_1 B_2 \cdots B_{k-1}}, \frac{n_2 s_2^2}{(1-B_1)B_2 \cdots B_{k-1}}, \frac{n_3 s_3^2}{(1-B_2)B_3 \cdots B_{k-1}}, \ldots, \frac{n_k s_k^2}{(1-B_{k-1})}])), \quad (3.42)$$

where $H_{k-1,N-k}$ is the cdf of F distribution with $k-1$ and $N-k$ degrees of freedom, and the expectation is taken with respect to the independent Beta variables

$$B_j \sim \text{Beta}(\sum_{i=1}^{j} \frac{(n_i - 1)}{2}, \frac{n_{j+1} - 1}{2}), \quad j = 1, 2, \ldots, k-1. \quad (3.43)$$

Being an average of familiar F probabilities, not only is this representation of the generalized p-value more appealing, but also it is computationally more efficient in exact numerical integration.

If the number of treatments being compared is very large, the expectation in (3.41) or (3.42) can also be well approximated by a Monte Carlo method. In this case perhaps the representation (3.41) is more convenient. The computation using the representation (3.41) is carried out in the following steps:

- Generate a set of large number of random numbers from each chi-squared random variable $Y_i \sim \chi^2_{n_i-1}$,
- For each set compute the cdf $g = G_{k-1}(\widetilde{s}_B(\frac{n_1 s_1^2}{Y_1}, \frac{n_2 s_2^2}{Y_2}, \ldots, \frac{n_k s_k^2}{Y_k})$,
- Compute their average, say $\overline{g}$,
- Estimate the generalized p-value by $1 - \overline{g}$.

The accuracy of the approximation can also be assessed. For example, if the sample (simulated) standard deviation of g values is s_g, then with probability 0.999, the estimated p-value is accurate up to about $3s_g/\sqrt{L}$, where L is the

number of simulated sample sets. Both the exact numerical integration and the Monte Carlo procedure are available from the one-way ANOVA tools of the XPro software package. With a little coding to implement (3.41) or (3.42), the computation can also be performed with widely used software packages such as SAS and SPlus.

In generalized fixed-level testing at level α, H_0 is rejected if $p < \alpha$. In significance testing, the generalized p-value computed using (3.41) or (3.42) provides an exact unbiased test based on sufficient statistics. Exact conventional fixed-level tests based on sufficient statistics do not exist. In this case, as demonstrated by Gamage and Weerahandi (1998), the generalized F-test provides an excellent approximate test and its size does not exceed the intended value of α. Therefore, the generalized F-test is extremely useful regardless of whether one prefers conventional fixed-level testing or significance testing based on p-values.

For the problem of comparing two normal populations, the test given by the generalized approach is unique (up to equivalent p-values) among all affine invariant and unbiased procedures based on minimal sufficient statistics. In view of the results in Khuri and Littell (1987), Khuri (1990), and Zhou and Mathew (1994), however, the uniqueness of the test is not expected beyond the case of two means, unless further conditions are imposed.

Also note that, for large sample sizes, the degenerated beta random variables give rise to the same p-value as the one implied by (3.35) for the case of known error variances. That the p-value given by (3.42) is symmetric with respect to the population indices is clear from the form (3.41), which is equivalent to (3.42). In other words, we will get the same p-value irrespective of how the treatments are indexed.

Example 2.5. Comparing the mean strength of reinforcing bars (continued)

Consider again the data in Table 2.3 and the summary statistics in Table 2.4. Recall that in Example 2.4 we concluded that there is no sufficient evidence to reject the null hypothesis that the mean strengths of the four brands are equal. Let us now drop the assumption of equal error variances and retest the hypothesis. In this application the p-value given by (3.42) reduces to $p = 1 - E(H_{3,27}(9\widetilde{s}_B[\frac{7\times 9.41}{B_1B_2B_3}, \frac{8\times 27.93}{(1-B_1)B_2B_3}, \frac{7\times 1.93}{(1-B_2)B_3}, \frac{9\times 2.19}{(1-B_3)}])) = 0.021$, where the expectation is taken with respect to the beta random de variables $B_1 \sim \text{Beta}(3, 3.5)$, $B_2 \sim \text{Beta}(6.5, 3)$ and $B_3 \sim \text{Beta}(9.5, 4)$. This means that although the classical F-test failed to detect the statistical significance, the current data set does provide fairly strong evidence to conclude that the observed differences of the mean strengths of the four brands of reinforcing bars are actually significant and cannot be attributed to just the sampling variation. Therefore, the engineer can proceed with multiple comparisons and deal with hypotheses concerning absolute differences in mean yields.

Recall that the p-value given by the classical F-test was 0.211, which is ten times as large as the p-value given by the generalized F-test. This means that

when the assumption of equal variances is not reasonable, the test given by (3.42) is much more powerful than the test given by (3.36). The failure of the classical F-test to detect truly significant mean differences is considered very serious. In this example by applying the classical F-test the engineer would have concluded that there is no difference in the four brands and would have recommended the use of perhaps a cheap brand, which could have been the worst brand in terms of strength.

This example demonstrates the importance of avoiding unreasonable assumptions which are made for mathematical simplicity. This is especially true in biomedical experiments where one does not usually get large samples and cannot afford to resort to less efficient statistical procedures for the sake of simplicity. It should also be noted that assumption of equal variances can lead to misleading conclusions even with fairly large samples. In fact, the assumption of equal variances is much more serious than the assumption of normality, because F-tests including the generalized F-test are robust against the distributional assumption. When the sample variances indicate that the assumption of equal variances is not reasonable, use of the generalized F-test can prove to be a win−win situation because not only does it depend on milder assumptions, but also it can be substantially more powerful than the classical F-test. Therefore, the classical F-test is recommended only when the assumption of equal variances is very reasonable. For various procedures for testing the assumption of equal variances, the reader is referred to Weerahandi (1995).

For various simulation studies concerning the power and the size of the generalized F-test compared with the classical F-test and some other approximate tests, the reader is referred to Gamage and Weerahandi (1998). Anderson and McLean (1974) provide and in-depth investigation of the performance of the classical F-test. According to results in Gamage and Weerahandi (1998), the actual size of the generalized F-test tends to be as good as or better than most approximate tests available in the literature. It has the added advantage that its p-value is the exact probability of an extreme region of the sample space. Moreover, as Weerahandi and Tsui (1996) demonstrated, the p-value is numerically equivalent to a Bayesian posterior predictive p-value [cf. Meng (1994)], a property that no other competing test has.

3.5 MULTIPLE COMPARISONS: CASE OF EQUAL VARIANCES

Foregoing results allow us to test only the equality of all population means. In many applications, the problem is not completely solved yet. Suppose a certain hypothesis of equal treatment means has been rejected at a desired nominal level and we still need to identify the means which are significantly different from others. One can of course carry out a set of pairwise comparisons. Although the concerns about the Type I error of pairwise comparisons expressed in the introduction of this chapter still hold, the issue is no longer

very serious since the null hypothesis has already been rejected. In any event, the appropriate procedure to carry out multiple comparisons depends on the way we want to control the Type I error. For a discussion on various pre-planned and post-hoc procedures valid under different control conditions, the reader is referred to Woolson (1987). Here we outline some of the procedures available for controlling the error rate for (1) a few pre-planned experiments, (2) all possible pairwise comparisons of means, and (3) all possible linear combinations of means, in which Type I error is on a per-experiment basis.

3.5.1 Bonferroni method

First consider the problem of making some pre-planned comparisons. To be specific suppose we wish to carry out some pairwise comparisons. In making just one comparison when variances are equal, we can employ the results in Section 2.2. When the population variances are all equal, the common variance can be estimated as in (3.29). Let $S^2 = S_E/(N-k)$ be the unbiased estimator. The distribution of the random variable S^2 is given by

$$\frac{(N-k)S^2}{\sigma^2} \sim \chi^2_{N-k}.$$

Inferences on any pair of the means can be made by taking the approach in the two sample problem treated above. For example, inferences on $\delta = \mu_1 - \mu_2$ can be made based on the result

$$\frac{\overline{Y}_1 - \overline{Y}_2 - \delta}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{N-k.}$$

For instance, one-sided hypotheses of two means, say $H_0 : \mu_1 - \mu_2 \leq \nu$, can thus be tested on the basis of the p-value

$$p = 1 - G_{N-k}\left(\frac{\overline{y}_1 - \overline{y}_2 - \nu}{s\sqrt{(1/n_1 + 1/n_2)}}\right), \tag{3.44}$$

where G_{N-k} is the cdf of the Student's t distribution with $N-k$ degrees of freedom. The p-value for testing two-sided hypotheses of the form $H_0 : \mu_1 - \mu_2 = \nu$ can be obtained in a similar manner as

$$p = 2G_{N-k}\left(-\frac{|\overline{y}_1 - \overline{y}_2 - \nu|}{s\sqrt{(1/n_1 + 1/n_2)}}\right). \tag{3.45}$$

At fixed-level α, the null hypothesis is rejected if $p < \alpha$.

Now suppose we wish to carry out r number of such comparisons. Then, the Bonferroni procedure is to use significance level α/r in place of α in fixed-level testing. In other words, each hypothesis is rejected at the α level if its p-value is less than α/r. The Bonferroni procedure is very general in that it can be applied for any situation where we have a procedure for testing a

single test. The procedure ensures the size of combined tests to be no larger than α. Obviously this procedure would become conservative as r increases.

Example 2.6. Pairwise comparisons

To illustrate the procedure, consider the set of summary data given in Table 2.7 for comparing four treatments P, Q, R, and S. Shown in the three columns of the table are the sample sizes, the sample means, and the sample variances (MLEs of variances), respectively.

Table 3.7 Treatment statistics

Treatment	n_i	$\overline{Y}_i$	s_i^2
P	10	23.7	2.4
Q	15	24.5	8.9
R	20	23.3	1.6
S	15	21.2	2.8

The sum of squares and the F-Statistic computed from the summary data in Table 2.7 are given in Table 2.8.

Table 3.8 ANOVA for comparing the treatments

Source	DF	SS	MS	F-Statistic
Between	3	87.8	29.3	7.08
Error	56	231.5	4.13	
Total	59	319.4		

The observed p-value that follows from the F-test in the ANOVA table is 0.0004, and therefore we have very strong evidence to reject the hypothesis of equal means. Now suppose we need to establish the significance of possible differences in means between R and S, Q and R, and Q and P. To perform the pairwise comparisons, consider the hypotheses $\mu_R = \mu_S$, $\mu_Q = \mu_P$, and $\mu_Q = \mu_R$. Let us carry out these tests at $0.05/3 = 0.017$ level to ensure that the size of the combined test is less than the desired level 0.05. By applying (3.44) it is seen that the p-value for testing the three hypotheses are 0.0037, 0.3393, 0.0895, respectively. Hence, we can conclude at the 0.05 Type I error level that, the mean of Treatment R is greater than that of treatment S, but the difference in means between Q and P or Q and R are not quite statistically significant.

3.5.2 Scheffe method

The Bonferroni method ensures Type I error for the pre-planned multiple comparisons only. A procedure that ensures Type I error for any number of post-hoc comparisons of linear contrasts was introduced by Scheffe (1953). A linear contrast is a linear combination of the means of the form $\sum_{i=1}^{k} c_i \mu_i$, with the coefficients of means satisfying the condition $\sum_{i=1}^{k} c_i = 0$. Scheffe test is closely related to the F-test developed above for testing the equality of all means. Of all multiple comparison methods available in the literature, the Scheffe test has the advantage that it detects one or more significant linear contrasts if and only if the F-test is significant. Its main disadvantage is that even if we are interested only in some contrasts, the test ensures the size for all possible contrasts, and hence tends to be less powerful in detecting significant contrasts.

For details of the Scheffe test including derivations, the reader is referred to Scheffe (1953) and Weerahandi (1995). To briefly describe the testing procedure, consider the null hypothesis

$$H_0 : \sum_{i=1}^{k} c_i \mu_i = 0 \text{ for all } c_i \text{ such that } \sum_{i=1}^{k} c_i = 0. \tag{3.46}$$

Then Scheffe showed that the null hypothesis all zero contrasts can be tested on the basis of the p-value

$$p = 1 - H_{k-1,N-k}\left(\frac{N-k}{k-1}\frac{(\sum_{i=1}^{k} c_i \overline{y}_i)^2}{s_E \sum_{i=1}^{k} c_i^2/n_i}\right). \tag{3.47}$$

A set of $100\gamma\%$ simultaneous confidence intervals for the linear contrasts is given by

$$\sum_{i=1}^{k} c_i \overline{y}_i - g \leq \sum_{i=1}^{k} c_i \mu_i \leq \sum_{i=1}^{k} c_i \overline{y}_i + g\,, \tag{3.48}$$

where

$$g^2 = (k-1)\widehat{\sigma}^2 \left(\sum_{i=1}^{k} c_i^2/n_i\right) F_{k-1,N-k}, \tag{3.49}$$

$\widehat{\sigma}^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2/(N-k)$, and $F_{k-1,N-k}$ is the γth quantile of the F distribution with $k-1$ and $N-k$ degrees of freedom. In testing a few

contrasts, the length of the confidence interval given by (3.48) can be much larger than the $100\gamma\%$ confidence interval given by Bonferroni method.

Example 2.7. Pairwise comparisons (continued)

Consider again the data in Table 2.7 concerning the problem of comparing four treatments. In the previous subsection we tested the hypotheses $\mu_R = \mu_S$, $\mu_Q = \mu_P$, and $\mu_Q = \mu_R$ by the Bonferroni method. To test the same hypothesis by the Scheffe method, we need to apply (3.47) repeatedly with contrasts, $\mathbf{c} = (0, 0, 1, -1)$, $(1, -1, 0, 0)$, and $(0, 1, -1, 0)$, respectively. The resulting p-values for testing the three hypothesis are 0.0360, 0.9289, 0.1801, respectively. Thus we come to the same conclusion that only the first contrast is significantly different from zero. Notice that all thee p-values are much larger than before, indicating the lack of power of the Scheffe test. This is true even after the necessary size adjustment required by the Bonferroni test. To see this, suppose we were to test the hypotheses at the 0.02 level of significance. Then none of the hypotheses can be rejected by the Scheffe test, whereas the first hypothesis is still rejected by the Bonferroni method because $0.0037 < 0.02/3$.

3.5.3 Generalized Tukey—Kramer method under equal variances

This is a procedure for controlling the Type I error rate for all possible pairwise mean comparisons as opposed to all possible contrasts. It is indeed the case in many applications that we really need to make only pairwise comparisons. In situations where only pairwise multiple comparisons are needed, the Tukey-Kramer procedure would produce more powerful tests compared to the Scheffe procedure. Original results due to Tukey (1953) in this context are valid only for the case of equal sample sizes. Kramer (1956) provided an extension to the case of unequal sample sizes. However, Kramer's extension was too conservative and hence there were a number of attempts to extend the Tukey−Kramer results. Of particular interest is a natural generalization of Tukey−Kramer procedure presented by Hsu (1995). Here we take the approach of Chang, Huang, and Wong (2002) and present the generalized Tukey−Kramer procedure by taking the generalized p-value approach.

Consider the problem of constructing simultaneous confidence intervals for $\delta_{ij} = \mu_i - \mu_i$ for all $i \neq j$. Chang, Huang, and Wong (2002) showed that

$$(\overline{y}_i - \overline{y}_j) - \xi\lambda_{ij} \leq \delta_{ij} \leq (\overline{y}_i - \overline{y}_j) + \xi\lambda_{ij} \tag{3.50}$$

provides a set of $100\gamma\%$ simultaneous confidence intervals for the mean differences δ_{ij}, where

$$\lambda_{ij} = \left\{ \frac{(N-k)\widehat{\sigma}^2}{N-k-2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right) \right\}^{1/2}$$

and ξ is the solution of the equation

$$\sum_{i=1}^{k} E\left(\prod_{i\neq j}\left[\Phi\left(\sqrt{\frac{n_j}{n_i}}Z\right)-\Phi\left(\sqrt{\frac{n_j}{n_i}}Z-\sqrt{\frac{n_jV}{(N-k)\widehat{\sigma}^2}}\xi\lambda_{ij}\right)\right]\right)=\gamma\ ,$$

where Φ is the cumulative distribution function of standard normal distribution and the expectation is taken with respect to the normal and chi-squared random variables

$$Z\sim N(0,1)\ \text{ and } V\sim\ \chi^2_{N-k}.$$

Hypotheses of equal means are rejected at α level if the $100(1-\alpha)\%$ confidence intervals do not contain them.

For a simple derivation of this procedure, the reader is referred to Chang, Huang, and Wong (2002). Their derivation provides an example of how very complicated inference procedures by classical approach could be derived by a few simple arguments in the context of generalized inference. The simultaneous confidence intervals given by (3.50) is equivalent to that of Hsu (1995). The above representation has the advantage that it naturally extends to the case of unequal variances. Moreover, under the usual noninformative prior, it is numerically equivalent to the Bayesian credible region.

3.6 MULTIPLE COMPARISONS: CASE OF UNEQUAL VARIANCES

We assumed in the previous section that the variances of all the populations being compared are equal. This assumptions tends to be unreasonable especially when the population means being compared are unequal. As briefly discussed below, the generalized approach allows us to extend each of the three methods discussed above to the case of unequal variances.

3.6.1 Generalized Bonferroni method

Note first of all that if the variances are not assumed equal, the generalized t-test developed in Section 2.2 can be employed to make inferences about the difference of one pair of the means. In this case, since no additional information about the variances are available, the procedure remains the same, and the p-values are computed exactly the same way as in the two sample case. In other words, we can still apply (3.19) if there were only one preplanned pair-wise comparison. When we need to carry out r number of such comparisons, we simply apply the generalized t-test with the Bonferroni size adjustment; that is, we compute the p-value using the same formula and perform each generalized t-test with the significance level α/r in place of α. That is, if p is the p-value obtained using formula (3.19) for comparing a certain pair of means, then the underlying null hypothesis is rejected if $p<\alpha/r$.

Example 2.8. Pairwise comparisons (continued)

Continuing with the numerical example used in the case of equal variances, consider again the hypotheses $\mu_R = \mu_S$, $\mu_Q = \mu_P$, and $\mu_Q = \mu_R$ without the assumption of equal variances. If we compute the p-values by applying the generalized t-tests, we get the p-values 0.0009, 0.4271, 0.1801, respectively for the three hypotheses. Comparing these $p-$values with adjusted level of 0.017, we come to the same conclusion that only the mean difference of Treatments R and S are significant. In fact, the p-value for testing the difference in R and S is less than before, indicating that the generalized t-test is somewhat more powerful than the classical t-test due to unequal variances. Of course this is not always the case and one should not expect to see smaller p-values whenever the means are truly different and the variances are unequal. In fact, unlike the F-test in ANOVA, the lack of power of the two sample t-test is not very serious.

3.6.2 Generalized Scheffe method

Now consider the problem of ensuring Type I error for all possible linear contrasts when the variances are not necessarily equal. This can be accomplished by considering the standardized observations $\widetilde{Y}_{ij} = Y_{ij}/\sigma_i,\ i = 1, \ldots, k,\ j = 1, \ldots, n_i$, defining a test variable parallel to the one used in the derivation of the generalized F-test in terms of $\sum_{i=1}^{k} c_i \overline{Y}_i$. Weerahandi (1994) showed that the p-value appropriate for testing (3.46) is given by

$$p = 1 - E(H_{k-1,N-k}(\frac{N-k}{k-1}\frac{(\sum_{i=1}^{k} c_i \overline{y}_i)^2}{\sum_{i=1}^{k} c_i^2 s_i^2/U_i})), \tag{3.51}$$

where the expectation is taken with respect to the random variables

$$U_i = (1 - B_{i-1})B_i \cdots B_{k-1}, \quad i = 2, \ldots, k-1,$$

$$U_1 = B_1 B_2 \cdots B_{k-1} \quad \text{and} \quad U_k = (1 - B_{k-1}),$$

where $B_i,\ i = 1, \ldots, k-1$ are the beta variables defined by (3.43). A test based on this p-value is referred to as the generalized Scheffe test. The statistical software package XPro provides tools to obtain the p-value for the generalized Scheffe test as well as for the conventional Scheffe test.

Example 2.9. Testing of contrasts

Consider again the data in Table 2.7. Suppose among other multiple comparisons of interest we need to test the hypothesis,

$$H_0 : \mu_P + \mu_R = \mu_Q + \mu_S.$$

If the Scheffe test had been used to perform all prior comparisons, we can proceed to test this hypotheses with no concern about the overall Type I error due to multiple comparisons. The current hypothesis is represented by the contrast given by $c_1 = 1$, $c_2 = -1$, $c_3 = 1$, and $c_4 = -1$. The $p-$values, for testing this contrast obtained with and without the assumption of equal variances, obtained by applying (3.47) and (3.51) are, 0.9982 and 0.9993, respectively. Both p-values suggest that there is no evidence against the null hypothesis.

Simultaneous generalized confidence intervals for contrasts of the means can also be deduced from (3.51). A set of $100\gamma\%$ simultaneous confidence intervals for the linear contrasts obtained in this manner are

$$\sum_{i=1}^{k} c_i \overline{y}_i \; - \; h \; \leq \; \sum_{i=1}^{k} c_i \mu_i \; \leq \; \sum_{i=1}^{k} c_i \overline{y}_i \; + \; h \; , \tag{3.52}$$

where h is chosen such that

$$E(H_{k-1,N-k}(\frac{(\sum_{i=1}^{k} c_i \overline{y}_i)^2}{h^2 \sum_{i=1}^{k} c_i^2 s_i^2/U_i})) = \gamma, \tag{3.53}$$

where $H_{k-1,N-k}$ is the cdf of the F distribution with $k-1$ and $N-k$ degrees of freedom and the expectation is taken with respect to U_i random variables.

3.6.3 Generalized Tukey—Kramer method under heteroscedasticity

The Scheffe tests are conservative by design because they control the error rate for all possible linear contrasts. The Bonferroni tests tend to be conservative when the number of comparisons is large. Due to these reasons, when one is interested only in pairwise mean comparisons, Tukey−Kramer type procedures are most desirable. Here we present the generalized Tukey−Kramer test for the heteroscedastic variances case developed by Chang, Huang, and Wong (2002).

Consider again the problem of constructing simultaneous confidence intervals for $\delta_{ij} = \mu_i - \mu_i$ for all $i \neq j$. Chang, Huang and Wong (2000) showed that

$$(\overline{y}_i - \overline{y}_j) - \xi\tau_{ij} \leq \delta_{ij} \leq (\overline{y}_i - \overline{y}_j) + \xi\tau_{ij} \tag{3.54}$$

provides a set of $100\gamma\%$ confidence intervals for the mean differences δ_{ij}, where

$$\tau_{ij} = \left\{ \frac{(n_i - 1)s_i^2}{n_i(n_i - 3)} + \frac{(n_j - 1)s_j^2}{n_j(n_j - 3)} \right\}^{1/2}$$

and ξ is the solution of the equation

$$\sum_{i=1}^{k} E\left(\prod_{i\neq j}\left[G_{n_i-1}\left(\sqrt{\frac{n_j s_i^2}{n_i s_j^2}}T\right) - G_{n_i-1}\left(\sqrt{\frac{n_j s_i^2}{n_i s_j^2}}T - \sqrt{\frac{n_j}{s_j^2}}\xi\tau_{ij}\right)\right]\right) = \gamma \ ,$$

where the expectation is taken with respect to the random variable distributed as

$$T \sim t_{n_i-1} \ ,$$

and G_{n_i-1} is the cumulative distribution function of the t distribution with $n_i - 1$ degrees of freedom. Hypotheses of equal means are rejected at the α level if the $100(1-\alpha)\%$ confidence intervals do not contain them. A proof of this result is given in the Appendix.

Chang, Huang, and Wong (2002) argued that the simultaneous confidence intervals given by (3.54) have the desirable property that they are numerically equivalent to Bayesian credible regions under the natural noninformative prior. They also performed a simulation to compare the probability coverage of (3.54) with the approximate simultaneous confidence intervals proposed by Dunnett (1980) and discussed the conditions under which the probability coverage of one is better than the other. While they found no clear winner and no substantial differences in probability coverage in repeated sampling, the generalized Tukey−Kramer interval has the added advantage of being based on an exact probability statement.

3.7 TWO-WAY ANOVA UNDER EQUAL VARIANCES

Foregoing results can be easily extended to various higher way ANOVA models where subjects or experimental units are observed under multiple levels of two or more factors of classification. In this book we consider only the case of two factors. To outline the approach in extending results for the one-way ANOVA to higher way layouts, consider the two-way layout with or without replications. Let us first consider the two-way cross classified designs under the assumption that the error variances are all equal, an assumption that we will relax in the next section.

In analyzing data from a two-way layout, suppose we are interested in the fixed effects of two factors, A and B. The case of the two-way layout with no replicates is treated in Appendix A.2. When replications are available we would be interested in the interactions between A and B as well their main effects. Let $A_1,\ A_2, \ldots,\ A_k$ be the levels of factor A, and $B_1,\ B_2, \ldots,\ B_n$ be the levels of factor B. Distinct values of a factor are called the levels of the factor. In most practical applications we would also have multiple data corresponding to each combination of factor levels. Appendix A.2 provides results for the special case where each combination of levels of A and B is represented by a single value of data. In this section we assume that we have replicates from each of the factor level combinations A and B. Also

assume that we have at least two data points from each pair $(A_i,\ B_j)$ of factor levels. The number of data available from $(A_i,\ B_j)$ is referred to as the cell frequency of the (i,j)th cell. Let $y_{ijk},\ i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,n_{ij}$ be the available data, where n_{ij} is the cell frequency, the number of observations available from the factor level combination (i,j). Define the vectors $\mathbf{y}_{ij} = (y_{ij1}, y_{ij2},\ldots,y_{ijn_{ij}}),\ i = 1,\ldots,I,\ j = 1,\cdots,J$. Then, the data can be set out just like in Table A.1 except that each $\mathbf{y}_{ij}$ now represents a vector of observations.

If there is no interaction between the two factors, we can make inferences by assuming the model given in Appendix A.2. When we have replicates as assumed in this section, instead assume the linear model

$$\begin{aligned} Y_{ijk} &= \mu_{ij} + \epsilon_{ijk}, \\ &\quad i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,n_{ij}, \end{aligned} \tag{3.55}$$

where γ_{ij} terms represent the interactions. We assume in this section that the error terms are normally distributed with constant variance; that is,

$$\epsilon_{ijk} \sim N(0,\sigma^2).$$

The assumption of common error variance will be relaxed later. Let μ_{ij} be mean of the random variables representing observations in the ijth cell in Table A.1. The decomposition of μ_{ij}, namely the mean of Y_{ijk} given in (3.55), is not unique. Widely used constraints to make the decomposition unique are presented below under various scenarios.

3.7.1 Case of equal sample sizes

First consider the case where we have equal number of observations, say K observations, from each cell; that is, $n_{ij} = K$ for all $i = 1,\ldots,I,\ j = 1,\ldots,J$. Define

$$\mu = \sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}/IJ\ ,\ \mu_{i.} = \sum_{j=1}^{J}\mu_{ij}/J\ ,\ \text{and}\ \mu_{.j} = \sum_{i=1}^{I}\mu_{ij}/I\ .$$

In terms of these parameters the model can be expressed as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{3.56}$$

where $\alpha_i = \mu_{i.} - \mu$ and $\beta_j = \mu_{.j} - \mu$ are the standardized main effects, respectively, due to A and B. With the constraints $\sum_{i=1}^{I}\gamma_{ij} = 0$ and $\sum_{j=1}^{J}\gamma_{ij} = 0$ we can represent interactions in terms of various means as

$$\begin{aligned} \gamma_{ij} &= \mu_{ij} - \mu - \alpha_i - \beta_j, \\ &\text{for } i = 1,\ldots,I \ \text{ and } \ j = 1,\ldots,J. \end{aligned}$$

Consider the problems of testing the hypotheses,

$$H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0, \tag{3.57}$$

$$H_{0B}: \beta_1 = \beta_2 = \cdots = \beta_n = 0 \tag{3.58}$$

and

$$H_{0AB}: \gamma_{11} = \gamma_{12} = \cdots = \gamma_{kn} = 0 \tag{3.59}$$

against the obvious alternative hypotheses.

Let $\overline{Y}_{ij} = \overline{Y}_{ij}.$ be the sample mean of the observations available from the factor level combination (A_i, B_j). Table 2.12, which is similar to Table A.1, displays all the sample means obtained in this manner along with their marginals. Shown also in the table by $\overline{Y}_{i.}$ are the mean of all the data corresponding to factor A_i. Similarly $\overline{Y}_{.j}$ is the mean of all the data corresponding to factor B_j. The average of observations available from all IJ cells is denoted by $\overline{Y}$, which is sometimes referred to as the grand mean.

Table 3.9 Sample means by factor

Levels	$\mathbf{B}_1$	...	$\mathbf{B}_j$	...	$\mathbf{B}_J$	Row means
A_1	$\overline{y}_{11}$	...	$\overline{y}_{1j}$	...	$\overline{y}_{1J}$	$\overline{y}_{1.}$
A_2	$\overline{y}_{21}$	...	$\overline{y}_{2j}$	...	$\overline{y}_{2J}$	$\overline{y}_{21.}$
...	...	...	...	...	...	...
A_i	$\overline{y}_{i1}$	...	$\overline{y}_{ij}$	...	$\overline{y}_{iJ}$	$\overline{y}_{i.}$
...	...	...	...	...	...	...
A_I	$\overline{y}_{I1}$	...	$\overline{y}_{Ij}$	...	$\overline{y}_{IJ}.$	$\overline{y}_{I.}$
Column means	$\overline{y}_{.1}$	...	$\overline{y}_{.j}$	...	$\overline{y}_{.J}$	$\overline{y}$

It is easily seen that the MLEs as well as the LSEs of the parameters α_i, β_j, and γ_{ij} are

$$\widehat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}, \ \widehat{\beta}_j = \overline{Y}_{.j} - \overline{Y}$$

and

$$\widehat{\gamma}_{ij} = \overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y},$$

respectively. These estimates are also unbiased.

Now proceeding to hypotheses testing, first consider the problem of testing the hypothesis of zero interactions, namely H_{0AB}. If this hypothesis is not rejected, it is also of interest to test the equality of the main effects—that is, the null hypotheses H_{0A} and H_{0B} defined by (3.57) and (3.58), respectively. As in the case of the two-way ANOVA with no replications, testing of various hypotheses can be facilitated by appropriate sums of squares and the mean sums of squares in the ANOVA table given below. The derivation of the

results are provided in Appendix A.3. Each mean-squared term appearing in the fourth column of the table is obtained by dividing the corresponding sum of squares by its degrees of freedom.

Table 3.10 Two-way ANOVA for cross classified design

Source	DF	SS	MS	F-Statistic
A	$I - 1$	S_A	MSA	MSA/MSE
B	$J - 1$	S_B	MSB	MSB/MSE
Interac.	$(I-1)(J-1)$	s_I	MSI	MSI/MSE
Error	$N - IJ$	s_E	MSE	
Total	$N - 1$	s_T		

The definitions of the sums of squares appearing in the ANOVA table are as follows:

$$S_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y})^2,$$

$$S_A = JK\sum_{i=1}^{I}(\overline{Y}_{i.} - \overline{Y})^2,\ S_B = IK\sum_{j=1}^{J}(\overline{Y}_{.j} - \overline{Y})^2,$$

$$S_I = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y})^2,$$

and

$$S_E = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij})^2 = K\sum_{i=1}^{I}\sum_{j=1}^{J}S_{ij}^2.$$

Under the hypothesis of zero effects, each F-statistic has an F distribution with the degrees of freedom suggested by the second column of the ANOVA table. For example, if H_{0AB} is true, then the F-statistic

$$F_I = \frac{\text{MSI}}{\text{MSE}} = \frac{S_I/(I-1)(J-1)}{S_E/(N-IJ)} \sim F_{(I-1)(J-1),N-IJ}; \tag{3.60}$$

has an F distribution with $(I-1)(J-1)$ and $N-IJ$ degrees of freedom. In fixed-level testing, H_{0AB} is rejected at α level if the observed value of the F-statistic is greater than the $(1-\alpha)$th quantile of the F distribution. In testing the hypothesis, the strength of the evidence in favor or against H_{0AB} can be better reported by the corresponding p-value

$$p_I = 1 - H_{i,e}\left(\frac{s_I/i}{s_E/e}\right), \tag{3.61}$$

where $i = (I-1)(J-1)$, $e = N-IJ$, and in general the notation $H_{a,b}$ stands for the cdf of the F distribution with a and b degrees of freedom. Similarly, the hypothesis H_{0A} is can be tested based on the $p-$value $p_A = 1 - H_{a,e}(\frac{s_A/a}{s_E/e})$. A derivation of the results are given Appendix A.3.

Example 2.10. Comparing teaching methods

In order to study the effect of three teaching methods, a research scientist have the methods tried out at 16 high schools. The teaching methods were tried out in Mathematics and Science in different classes of students and mean scores at the end of the marking period were recorded. The results of the experiment are shown in Table 2.11.

Table 3.11 Mean scores by teaching method

Class:	Mathematics	Science
Method 1	84, 87, 82, 88	90, 93, 96, 97
Method 2	88, 89, 77, 84	88, 92, 97, 96
Method 3	79, 84, 71, 80	84, 86, 79, 78

Table 2.12 presents the sum of squares computed using the data in Table 2.11. Corresponding mean sums of squares, the F-Values, and the p-values are also shown in the ANOVA table.

Table 3.12 ANOVA for comparing teaching methods

Source	DF	SS	MS	F-Value	p-value
Method	2	446.33	223.17	12.313	.00043
Class	1	287.04	287.04	15.837	.00088
Interaction	2	40.33	20.17	1.113	.35027
Error	18	326.25	18.13		
Total	23	1099.96			

It is evident from the F-Values and the p-values in the above ANOVA table that the differences in teaching methods are highly significant, mainly due to teaching method 1. This is also clear from Figure 2.2, which suggests that in fact Method 1 scores for each subject are much higher than that of Method 3. Perhaps there is no significant difference between Methods 1 and Method 2. It seems that mean score that students obtained for Science is significantly larger than that for Mathematics. In Figure 2.2 the two curves are not quite parallel, suggesting some interaction between the subject and the method of teaching. But, according to the ANOVA, the interaction terms are not statistically significant, suggesting that the effect of different teaching

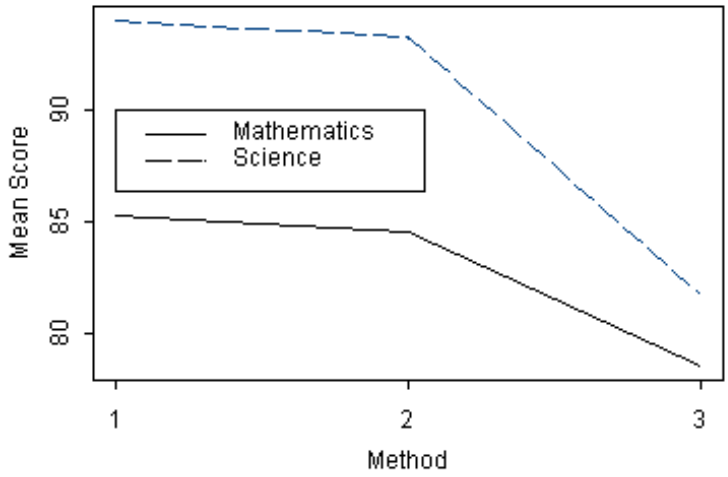

Figure 3.2 Scores by subject and method of teaching

methods is the same regardless of whether they are practiced in Science or in Mathematics.

3.7.2 Case of unequal sample sizes

Next consider the model given in (3.55) without the assumption of equal cell frequencies. In this section we also continue to assume that the error variances are all equal. We need to decompose the means into main effects and interactions in the form $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ so that we can specify the testing problem. Further, making the decomposition unique requires some constrains. Consider the general linear constraints

$$\sum_i w_i \alpha_i = 0, \ \sum_j v_j \beta_j = 0, \ \sum_i w_i \gamma_{ij} = 0, \ \sum_j v_j \gamma_{ij} = 0. \tag{3.62}$$

The choice of weights and their impact on testing procedures will be discussed later.

Testing the interactions

First consider the problem of testing the equality of interaction effects, namely the hypothesis

$$H_{0AB} : \ \gamma_{ij} = 0 \quad \text{for all} \quad i = 1, \ldots, I, \ j = 1, \ldots, J.$$

In this case, tests of equal interaction terms can be based on the standardized interaction sum of squares

$$S_I = \sum_i \sum_j n_{ij}(\overline{Y}_{ij} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j)^2 \tag{3.63}$$

and the error sum of squares

$$S_E \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \overline{Y}_{ij})^2, \tag{3.64}$$

where $\overline{Y}_{ij} = \sum_{k=1}^{n_{ij}} Y_{ijk}/n_{ij}$ is the sample mean of the data from (i,j)th cell, and $(\widehat{\mu}\ ,\ \widehat{\alpha}_i\ ,\ \widehat{\beta}_j)$ is the set of values of parameters $(\mu,\ \alpha_i\ ,\ \beta_j)$ that minimizes the quadratic form

$$f = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \mu - \alpha_i - \beta_j)^2$$

subject to the constraints in equation (3.62).

Despite the unequal cell frequencies, testing the of equality of interactions can still be based on the result

$$F_I \;=\; \frac{\text{MSI}}{\text{MSE}} \;=\; \frac{S_I/(I\ -\ 1)(J-1)}{S_E/(N-IJ)} \;\sim\; F_{(I-1)(J-1),N-IJ}$$

and the resulting p-value

$$p_I = H_{(I\ -\ 1)(J-1),(N-IJ)}\left(\frac{(N-IJ)s_I}{(I\ -\ 1)(J-1)s_E}\right), \tag{3.65}$$

where $H_{(I\ -\ 1)(J-1),(N-IJ)}$ is the cdf of the F distribution with $(I\ -\ 1)(J-1)$ and $(N-IJ)$ degrees of freedom, and $N = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$ is the total sample size. The result is valid regardless of the weights appearing in (3.62); the reader is referred to Arnold (1981) for a detailed discussion of these issues. In view of this result, it is convenient to choose the weights as $w_i = n_{i.} = \sum_{j=1}^{J} n_{ij}$ and $v_j = n_{.j} = \sum_{i=1}^{I} n_{ij}$. Then, $(\widehat{\mu}\ ,\ \widehat{\alpha}_i\ ,\ \widehat{\beta}_j)$ required in the computation of the interaction sum of squares can be found by solving the system of linear equations

$$\sum_{j=1}^{J} n_{ij}(\overline{y}_{ij} - \widehat{\mu}\ - \widehat{\alpha}_i - \widehat{\beta}_j\ ) = 0, \quad i = 1,\ldots,I$$

$$\sum_{i=1}^{I} n_{ij}(\overline{y}_{ij} - \widehat{\mu}\ - \widehat{\alpha}_i - \widehat{\beta}_j\ ) = 0, \quad j = 1,\ldots,J$$

Testing the main effects

Now consider the problem of testing the main effects. To be specific and to describe the nature of testing procedures available for testing main effects, consider the problem of testing the hypothesis

$$H_{0A}: \ \alpha_1 = \alpha_2 = \ \cdots = \alpha_k = 0.$$

Define the sum of squares due to factor A as

$$S_A = \sum_i \sum_j n_{ij}(\overline{Y}_{ij} - \widehat{\mu} - \widehat{\beta}_j - \widehat{\gamma}_{ij})^2,$$

where $(\widehat{\mu}\ ,\ \widehat{\beta}_j\ ,\widehat{\gamma}_{ij})$ are the estimates of the nuisance parameters $(\mu, \beta_j, \gamma_{ij})$. In this case, they are estimated by minimizing the function

$$g = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \mu - \beta_j - \gamma_{ij})^2.$$

Given any set of weights, under H_{0A}, S_A has a chi-squared distribution with $I-1$ degrees of freedom and it is distributed independently of S_E. The reader is referred to Arnold (1981) for details of these and related results. It is now clear that, given a set of user specified weights, the hypothesis can be tested based on the p-value

$$p_A = H_{(I\ -\ 1),(N-IJ)} \left(\frac{(N - IJ)s_A}{(I\ -\ 1)s_E} \right). \tag{3.66}$$

Similarly, with the obvious definition of S_B, the equality of B main effects can be tested based on the p-value

$$p_B = H_{(J\ -\ 1),(N-IJ)} \left(\frac{(N - IJ)s_B}{(J\ -\ 1)s_E} \right).$$

As discussed by Fujikoshi (1993), with the widely used choice of using n_{ij} for weights with the constraints $\sum_i n_{i.}\alpha_i \ = 0$, $\sum_j n_{.j}\beta_j = 0$, $\sum_i n_{ij}\gamma_{ij} = 0$, $\sum_j n_{ij}\gamma_{ij} = 0$, the two sum of squares S_A and S_B can be conveniently computed as

$$S_A = \sum_i \sum_j n_{ij}(\overline{Y}_{ij} - \overline{Y})^2 - S_I - \sum_j n_{.j}(\overline{Y}_{.j} - \overline{Y})^2 \tag{3.67}$$

and

$$S_B = \sum_i \sum_j n_{ij}(\overline{Y}_{ij} - \overline{Y})^2 - S_I - \sum_j n_{i.}(\overline{Y}_{i.} - \overline{Y})^2, \tag{3.68}$$

where $\overline{Y}_{i.} = \sum_j \overline{Y}_{ij}/J$ and $\overline{Y}_{.j} = \sum_i \overline{Y}_{ij}/I$ are the sample means corresponding to the levels of the two factors, and $\overline{Y} = \sum_i \sum_j \sum_k Y_{ijk}/N$ is the grand mean of all the observations.

3.8 TWO-WAY ANOVA UNDER HETEROSCEDASTICITY

Let us now drop the assumption of equal variances made in previous sections, and consider the testing problem of a factorial design with two factors. Specifically, consider the linear model (3.55) under the milder assumption

$$\epsilon_{ijk} \ \sim \ N(0, \sigma_{ij}^2),\ i = 1, \ldots, I,\ j = 1, \ldots, J.$$

Inferences can still be based on the sufficient statistics $\overline{Y}_{ij},\ S_{ij}^2,\ i = 1, \ldots, I,$ $j = 1, \ldots, J$. As in one-way ANOVA, suitable tests for this case can be derived from results valid for the known variances case. Appropriate sums of squares leading to F-Statistics can be deduced, for instance, from results of Fujikoshi (1993) for the unbalanced model or from the sums of squares decomposition we get when the variances are known.

Consider again the null hypothesis

$$H_{0AB}:\ \gamma_{ij} \ = \ 0 \ \text{ for all } \ i = 1, \ldots, I,\ j = 1, \ldots, J.$$

for testing the interaction between A and B. To test this hypothesis, consider the variance weighted interaction sum of squares and the error sum of squares

$$\widetilde{S}_I(\sigma_{11}^2, ..., \sigma_{IJ}^2) \ = \ \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}(\overline{Y}_{ij} \ - \ \widehat{\alpha}_i \ - \ \widehat{\beta}_j \ - \ \widehat{\mu})^2, \tag{3.69}$$

and

$$\widetilde{S}_E \ = \ \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \sigma_{ij}^{-2}(Y_{ijk} \ - \ \overline{Y}_{ij})^2 \ = \ \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}S_{ij}^2/\sigma_{ij}^2, \tag{3.70}$$

where $(\widehat{\mu},\ \widehat{\alpha}_i,\ \widehat{\beta}_j)$ is the set of values of parameters $(\mu,\ \alpha_i,\ \beta_j)$ that minimizes the quadratic form

$$f = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \sigma_{ij}^{-2}(Y_{ijk} - \mu - \alpha_i - \beta_j)^2$$

If n_{ij} is used to define the constraints as in the previous section, we get

$$\widehat{\mu} \ = \ \frac{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}\overline{Y}_{ij}}{\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}}$$

and $\widehat{\alpha}_i$ and $\widehat{\beta}_j$ become the solutions of the linear equations

$$\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}(\overline{Y}_{ij} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j) = 0, \text{ for } i = 1, \ldots, I$$

and

$$\sum_{i=1}^{I} n_{ij}\sigma_{ij}^{-2}(\overline{Y}_{ij} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j) = 0, \text{ for } j = 1, \cdots, J.$$

The estimates also satisfy the constraints

$$\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}\widehat{\alpha}_i = 0 \text{ and } \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}\widehat{\beta}_j = 0$$

so that $\widehat{\beta}_j$ can be eliminated from the first equation and $\widehat{\alpha}_i$ can be eliminated from the second equation, thus enabling their computation by matrix manipulations. For additional details of the problem the reader is referred to Ananda and Weerahandi (1997).

From results available for the known variances case we have

$$\widetilde{S}_I \sim \chi^2_{(I-1)(J-1)} \tag{3.71}$$

and

$$V_{ij} = \frac{n_{ij}S_{ij}^2}{\sigma_{ij}^2} \sim \chi^2_{n_{ij}-1}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J. \tag{3.72}$$

Equation (3.72) can be employed to tackle unknown variances in

$$\widetilde{S}_I(\sigma_{11}^2, ..., \sigma_{IJ}^2).$$

By applying the method of substitution, a p-value appropriate for testing the hypothesis of zero interaction effects can be computed as

$$\begin{aligned} p &= Pr(\widetilde{S}_I(\sigma_{11}^2, ..., \sigma_{IJ}^2) \geq \widetilde{s}_I(\frac{n_{11}s_{11}^2}{V_{11}}, \frac{n_{12}s_{12}^2}{V_{12}}, ..., \frac{n_{IJ}s_{IJ}^2}{V_{IJ}})) \\ &= 1 - E(G_i(\widetilde{s}_I(\frac{n_{11}s_{11}^2}{V_{11}}, \frac{n_{12}s_{12}^2}{V_{12}}, ..., \frac{n_{IJ}s_{IJ}^2}{V_{IJ}}))), \end{aligned} \tag{3.73}$$

where G_i is the cdf of the chi-squared distribution with $i = (I-1)(J-1)$ degrees of freedom and the expectation is taken with respect to the chi-squared random variables V_{ij}. The hypothesis H_{0AB} is rejected for small values of p. The p-value can be computed by exact numerical integration or by Monte Carlo method using on a large number of random numbers generated from each of the independent chi-squared random variables, V_{ij}.

For details and for a formal derivation of foregoing results, the reader is referred to Ananda and Weerahandi (1997). To show that this is actually

a generalization of the classical F-test, they also expressed (3.73) in terms of independent Beta random variables and the cdf of an F distribution with $(I-1)(J-1)$ and $N-IJ$ degrees of freedom. The p-value can be conveniently computed using the XPro software package. With a script to implement the Monte Carlo integration, it can also be computed using widely used statistical software such as SAS and SPlus.

Procedures for testing the hypotheses H_{0A} and H_{0B} can be derived in a similar manner. It should be noted however that, as in unbalanced models with unequal cell frequencies [cf. Lindman (1992)], the two-way ANOVA model with unequal and known variances does not yield orthogonal terms leading to a sums of squares decomposition. Moreover, the F-Statistics can be defined in alternative ways using different constraints for α_i and β_j main effects. Consequently, there is no common agreement about how the main effects should be tested in the presence of interactions. Here we employ a widely used method of computing the sums of squares due to main effects when the model is unbalanced.

To obtain the appropriate variance weighted sums of squares for the main effects, first define the sums of squares as

$$S_A(\sigma_{11}^2,...,\sigma_{IJ}^2) \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}(\overline{Y}_{ij} \;-\; \overline{Y}_{i.})^2 \tag{3.74}$$

and

$$S_B(\sigma_{11}^2,...,\sigma_{IJ}^2) \;=\; \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}(\overline{Y}_{ij} \;-\; \overline{Y}_{.j})^2 \tag{3.75}$$

Also define

$$\widetilde{S}_{T-E} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}(\overline{X}_{ij} \;-\; \widehat{\mu})^2. \tag{3.76}$$

Then, the standardized sums of squares due to the main effects A and B, in the presence of the other are defined as

$$\widetilde{S}_A(\sigma_{11}^2,...,\sigma_{IJ}^2) \;=\; \widetilde{S}_{T-E} \;-\; \widetilde{S}_I \;-\; S_B \tag{3.77}$$

and

$$\widetilde{S}_B(\sigma_{11}^2,...,\sigma_{IJ}^2) \;=\; \widetilde{S}_{T-E} \;-\; \widetilde{S}_I \;-\; S_A, \tag{3.78}$$

respectively. When the variances are unknown, the main effects can be tested by chi-squared tests based on the known results

$$\widetilde{S}_A \sim\; \chi^2_{(I-1)} \;\text{ and }\; \widetilde{S}_B \sim\; \chi^2_{(J-1)}\;.$$

When the variances are unknown as usually the case, the tests can be derived by taking the same approach we took in testing H_{0A}. The p-value for testing H_{0A} can be expressed in two alternative forms as before. The simpler expression of the two is

$$p = 1 - E(G_a(\widetilde{s}_A(\frac{n_{11}s_{11}^2}{V_{11}}, \frac{n_{12}s_{12}^2}{V_{12}}, ..., \frac{n_{1j}s_{IJ}^2}{V_{IJ}}))), \tag{3.79}$$

where $\widetilde{s}_A$ is the observed value of S_A and $a = I - 1$. This p-value also can be expressed in terms of Beta random variables and the cdf of an F distribution with a and $e = (I-1)(J-1)$ degrees of freedom. Similarly, the p-value for testing H_{0B} is obtained by replacing $\widetilde{s}_A$ by $\widetilde{s}_B$ and replacing a by the corresponding degrees of freedom $b = J - 1$. For detailed results concerning the unbalanced models due to unequal cell frequencies and heteroscedasticity, the reader is referred to Ananda and Weerahandi (1997). Since, the sum of squares decomposition is no longer orthogonal, they also discuss how one main effect can be tested in the presence and absence of the other. In application, the spirit of the above procedures remains the same except for the use of appropriate sum of squares, for a given factor (e.g. S_A or $\widetilde{S}_A$ for factor A), before or after the other one is added and the total sum of squares is balanced. This is illustrated by Example 2.11.

Example 2.11. Comparing teaching methods (continued)

Consider again the data presented in Table 2.11. In Example 2.7 we tested the effect of teaching methods under the assumption that data from each cell have equal variances. Now we are in a position to carry out the tests without that assumption. The p-values for computing the three hypotheses of interest are displayed in the table below. Also shown are the p-values for testing the incremental main effect of one factor before the other factor.

Source	p-value
Interaction	0.5266
SS Decomposition: $\widetilde{S}_{T-E} = \widetilde{S}_I + S_A + \widetilde{S}_B$	
Method (A)	0.0051
Class (B)	0.0043
SS Decomposition: $\widetilde{S}_{T-E} = \widetilde{S}_I + S_B + \widetilde{S}_A$	
Method (A)	0.0058
Class (B)	0.0587

These p-values, especially those based on the incremental sums of squares, $\widetilde{S}_I$, $\widetilde{S}_A$, and $\widetilde{S}_B$, also lead to the same conclusion as before that the differences in the effects of teaching methods, the class effect are statistically significant and that the interaction term is not significant.

Ananda and Weerahandi (1997) and Bao and Ananda (2002) provided examples and simulations to demonstrate the importance of addressing the prob-

lem of heteroscedasticity. In two-way ANOVA the assumption of equal variances not only has a severe adverse effect on the power of the test, but also can lead to concluding that factor A is significant when in fact factor B is the one that is significant. Although this is not always the case, depending on the variances, the classical F-test might lead to such misleading conclusions in other situations as well. Example 2.13 shows another problematic situation with a hypothetical data set. This by no means is a reasonable simulation study, and it simply serves to illustrates how the classical F-test can lead to wrong conclusions as a result of ignoring the unequal variances.

Example 2.12. Misleading implications of classical F-test

Table 2.12 shows the sample means and sample standard deviations (MLEs) computed from a balanced two-way layout with sample cell frequency 5. The hypothetical data set in this example were generated from a model having unequal means for the levels of factor A and normally distributed errors with unequal variances.

Table 3.13 Sample means by factor

Levels	B_1	B_2	B_3	B_4	B_5
A_1	16.3	14.1	14.1	13.6	13.5
A_2	15.9	15.8	15.7	18.6	14.1
A_3	17.4	16.9	14.0	15.1	14.1
A_4	16.9	17.7	14.5	13.9	14.9

Table 3.14 Sample variances (MLE) by factor

Levels	B_1	B_2	B_3	B_4	B_5
A_1	11.3	0.9	4.5	4.1	0.9
A_2	4.1	0.9	4.5	3.7	13.3
A_3	6.9	16.1	6.9	6.5	8.1
A_4	2.1	6.5	5.7	13.3	6.5

It is evident from the sample variances that the assumption of equal variances is not a reasonable one in this situation. If we ignore this fact and proceed with the classical approach, we get the following ANOVA table.

Source	DF	SS	MS	F-value	p-value
Interaction $A \times B$	12	96.335	8.028	1.013	0.4452
Factor A	3	39.63	13.22	1.667	0.1808
Factor B	4	85.39	21.35	2.694	0.0367
Error	80	634	7.925		
Total	99	855			

The classical ANOVA suggests that the differences in the levels of factor A are significant, but those of factor B are not. There is also no interaction between the two factors. Now let us drop the assumption of equal variances and retest the hypotheses using generalized p-values. The p-values for computing the three hypotheses are shown in the table below. Also shown are the p-values for testing the incremental main effect of one factor in the presence of the other factor.

Source	p-value
Interaction	0.4509
SS Decomposition: $\widetilde{S}_{T-E} = \widetilde{S}_I + S_A + \widetilde{S}_B$	
Factor A	0.0001
Factor B	0.1028
SS Decomposition: $\widetilde{S}_{T-E} = \widetilde{S}_I + S_B + \widetilde{S}_A$	
Factor A	0.0052
Factor B	0.0226

Observe that according to the above p-values, while we come to the same conclusion about interactions, it is factor A, not B, that is highly significant both in the presence or absence of factor B. In fact, in the presence of factor A, factor B is not quite significant at the 0.05 level, but may become significant with additional data. Since the classical F-test relied on an unreasonable assumption, its results are not reliable. This example demonstrates how the classical F-test can mislead us to making erroneous conclusions. Therefore, the classical F-test is not recommended unless the assumption of equal variances is reasonable.

3.9 TWO-FACTOR NESTED DESIGN

Procedures for analyzing data from other types of designs including higher-way designs and nested designs can be established by taking an approach similar to that in the above sections. To illustrate the nature of testing procedures available for nested designs, to be specific, consider a two-factor nested design with factors A and B. Ananda (1995) provided procedures for testing the usual hypotheses in this context with unequal cell frequencies and unequal variances. To outline the main results, suppose Factor A has I levels

and Factor B is nested within A having $J_1, J_2, \ldots, J_I$ levels so that the total number of levels of factor B is $J = \sum J_i$. Suppose a random sample of size n_{ij} is available from (i,j)th level of B, $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J_i$. Hence the total sample size is

$$N = \sum_{i=1}^{I} \sum_{j=1}^{J_i} n_{ij}.$$

Let Y_{ijk}, $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J_i$; $k = 1, 2, \ldots, n_{ij}$ be the random variables representing the observations available from each cell.

Assuming a linear model, we can consider the true cell mean of the (i,j) level of factor B, say μ_{ij}, as the sum of a general mean θ, the main effect α_i of the ith level of A, and an effect δ_{ij} of the (i,j)th level of factor B representing the interaction effects confounded with the main effect of B; that is

$$\mu_{ij} = \theta + \alpha_i + \delta_{ij}.$$

Let $\bar{Y}_{ij}$ and S_{ij}^2, $i = 1, \ldots, I$; $j = 1, \ldots, J_i$ denote the sample mean and the sample variance of the (i,j)th treatment; that is,

$$\bar{Y}_{ij} = \sum_{k=1}^{n_{ij}} Y_{ijk}/n_{ij},\ S_{ij}^2 = \sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \bar{Y}_{ij}\right)^2 /n_{ij}.$$

Their observed sample values are denoted by $\bar{y}_{ij}$ and s_{ij}^2, $i = 1, \ldots, I$; $j = 1, \ldots, J_i$ respectively. Assuming normally distributed observations, consider the model

$$Y_{ijk} = \theta + \alpha_i + \delta_{ij} + \epsilon_{ijk}, \tag{3.80}$$

where

$$\epsilon_{ijk} \sim N\left(0, \sigma_{ij}^2\right), \quad i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J_i;\ k = 1, 2, \ldots, n_{ij}.$$

and σ_{ij}^2 is the variance of data taken from (i,j)th cell. To make θ, α_i, and δ_{ij} unique, the usual linear constraints

$$\sum_{i=1}^{I} v_i \alpha_i = 0,\ \sum_{j=1}^{J_i} w_{ij} \delta_{ij} = 0$$

are imposed, where v_i and w_{ij} are nonnegative weights such that $\sum_{i=1}^{I} v_i > 0$ and $\sum_{j=1}^{J_i} w_{ij} > 0$.

3.9.1 Testing interactions

Consider the problem of testing hypothesis

$$H_{0\delta} : \delta_{ij} = 0, \quad i = 1, \ldots, I, \quad j = 1, \ldots J_i$$

against the natural alternative. Testing of $H_{0\delta}$ can be considered as a problem of testing whether or not the true cell means μ_{ij} depend only on i. For the unbalanced case, Arnold (1981) provided an F-test under the usual assumption of equal error variances. The p-value of the $F-$test is

$$p = 1 - H_{(J-I),(N-J)} \left[\frac{(N-J)\sum_{i=1}^{I}\sum_{j=1}^{J_i} n_{ij}(\bar{y}_{ij}^2 - \bar{y}_{i.})^2}{(J-I)\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij})^2} \right], \qquad (3.81)$$

where $H_{(J-I),(N-J)}$ is the cumulative distribution function of the F distribution with $(J-I)$ and $(N-J)$degrees of freedom and $\bar{y}_{i.} = \sum_j n_{ij}\bar{y}_{ij} / \sum n_{ij}$. When the variances are unequal, a generalized F-test can be obtained as before by first considering the solution based on the sum of squares $\tilde{S}_\delta$

$$\tilde{S}_\delta\left(\sigma_{11}^2, \sigma_{12}^2, \cdots, \sigma_{I,J_I}^2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{n_{ij}(\bar{X}_{ij})^2}{\sigma_{ij}^2} - \frac{\sum_{i=1}^{I}\left(\left(\sum_{j=1}^{J_i} \frac{n_{ij}\bar{X}_{ij}}{\sigma_{ij}^2}\right)^2\right)}{\sum_{j=1}^{J_i} \frac{n_{ij}}{\sigma_{ij}^2}} \qquad (3.82)$$

when the variances are known and then tackling the unknown variances by their estimates having the distribution

$$n_{ij}S_{ij}^2/\sigma_{ij}^2 \sim \chi_{n_{ij}-1}^2.$$

Then, it is straightforward to show (Exercise 2.7) that the generalized p-value appropriate for testing $H_{0\delta}$ is

$$p = 1 - E\left\{ G_{J-I}\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{\bar{x}_{ij}^2 R_{ij}}{s_{ij}^2} - \frac{\sum_{i=1}^{I}\left(\sum_{j=1}^{J_i} \frac{\bar{x}_{ij}R_{ij}}{s_{ij}^2}\right)^2}{\sum_{j=1}^{J_i} \frac{R_{ij}}{s_{ij}^2}} \right] \right\} \qquad (3.83)$$

where G_{J-I} is the cdf of the chi-squared distribution with $J-I$ degrees of freedom and the expectation is taken with respect to the independent chi-squared random variables

$$R_{ij} \sim \chi_{n_{ij}-1}^2, \quad i = 1, \ldots, I, \; j = 1, \ldots J_i.$$

With most statistical software packages, the $p-$value can be computed by Monte Carlo integration by generating chi-squared random numbers and then estimating the expected value appearing in the formula by the sample mean of the simulated data

$$g_l = G_{J-I}\left[\sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{\bar{x}_{ij}^2 r_{ijl}}{s_{ij}^2} - \frac{\sum_{i=1}^{I}\left(\sum_{j=1}^{J_i} \frac{\bar{x}_{ij}r_{ijl}}{s_{ij}^2}\right)^2}{\sum_{j=1}^{J_i} \frac{r_{ijl}}{s_{ij}^2}} \right], \quad l = 1, ..., L,$$

where r_{ijl} is the lth chi-squared random number generated from R_{ij}. The XPro software automatically performs the necessary Monte Carlo integrations.

3.9.2 Testing main effects

Now consider the problem of testing the main effects of Factor $A-$ i.e., the problem of testing the hypothesis $H_{0\alpha}$. Unlike the problem of the interactions, appropriate tests in this situation depend on the weights chosen for w_{ij}. Therefore, the weights must be specified prior to testing. In the case of equal variances, the hypothesis can be tested using an F-test. In particular, with the weights $w_{ij} = n_{ij}$ this F-Statistic leads to the p-value

$$p = 1 - H_{(I-1),(N-J)} \left[\frac{(N-J)\sum_{i=1}^{I} n_{i.}(\bar{x}_{i.}^2 - \bar{x}_{..})^2}{(I-1)\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{n_{ij}}(x_{ijk} - \bar{x}_{ij})^2} \right]. \qquad (3.84)$$

Ananda (1995) provided generalized tests for the case of unequal variances and general weights w_{ij} and also for proportional weights $w_{ij} = n_{ij}/\sigma_{ij}^2$. The generalized test for the case of general weights is given by

$$p = 1 - E\left\{ G_{I-1}\left[\tilde{s}_\alpha \left(\frac{n_{11}s_{11}^2}{R_{11}}, \frac{n_{12}s_{12}^2}{R_{12}}, \cdots, \frac{n_{I,J_I}s_{I,J_I}^2}{R_{I,J_I}} \right) \right] \right\}, \qquad (3.85)$$

where G_{I-1} is the cdf of the chi-squared distribution with $(I-1)$ degrees of freedom, the expectation is taken with respect to the independent random variables $R_{ij} \sim \chi^2_{n_{ij}-1}$, and $\tilde{s}_\alpha$ is the observed value of

$$\tilde{S}_\alpha\left(\sigma_{11}^2, \sigma_{12}^2, \cdots, \sigma_{I,J_I}^2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J_i} \frac{n_{ij}}{\sigma_{ij}^2}\left(\bar{X}_{ij} - \hat{\theta} - \hat{\delta}_{ij}\right)^2, \qquad (3.86)$$

where

$$\hat{\theta} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left[w_{ij}\left(\sum_{k=1}^{J_i} w_{ik}\bar{X}_{ik}\right) / \left(\sum_{k=1}^{J_i} w_{ik}^2\sigma_{ik}^2/n_{ik}\right)\right]}{\sum_{i=1}^{I}\sum_{j=1}^{J_i}\left[w_{ij}\left(\sum_{k=1}^{J_i} w_{ik}\right) / \left(\sum_{k=1}^{J_i} w_{ik}^2\sigma_{ik}^2/n_{ik}\right)\right]}$$

and

$$\hat{\delta}_{ij} = \bar{X}_{ij} - \hat{\theta} - \frac{w_{ij}\sigma_{ij}^2\left(\sum_{k=1}^{J_i} w_{ik}\bar{X}_{ik} - \hat{\theta}\sum_{k=1}^{J_i} w_{ik}\right)}{n_{ij}\sum_{k=1}^{J_i}\left(w_{ik}^2\sigma_{ik}^2/n_{ik}\right)}$$

for all $i = 1, 2, \ldots, I; j = 1, 2, \ldots, J_i$. Ananda (1995) also expressed the generalized test in the form of a generalized F-test.

The weights of the form $w_{ij} = n_{ij}/\sigma_{ij}^2$ is the counterpart of the weights leading to the simple F-Statistic that Arnold (1981) discussed. With these weights the generalized sum of squares $\tilde{S}_\alpha$ reduces to

$$\tilde{S}_\alpha^* = \frac{\sum_{i=1}^{I}\left(\sum_{j=1}^{J_i} n_{ij}\bar{X}_{ij}/\sigma_{ij}^2\right)^2}{\sum_{j=1}^{J_i} n_{ij}/\sigma_{ij}^2} - \frac{\left(\sum_{i=1}^{I}\sum_{j=1}^{J_i} n_{ij}\bar{X}_{ij}/\sigma_{ij}^2\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J_i} n_{ij}/\sigma_{ij}^2} \qquad (3.87)$$

Hence, the p-value for testing $H_{0\alpha}$ can be conveniently computed as

$$p = 1 - E\left\{ G\left[\sum_{i=1}^{I} \left(\left(\sum_{j=1}^{J_i} \bar{x}_{ij} R_{ij}/s_{ij}^2 \right)^2 \left(\sum_{j=1}^{J_i} R_{ij}/s_{ij}^2 \right)^{-1} \right) - \left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} \bar{x}_{ij} R_{ij}/s_{ij}^2 \right)^2 \left(\sum_{i=1}^{I} \sum_{j=1}^{J_i} R_{ij}/s_{ij}^2 \right)^{-1} \right] \right\}. \quad (3.88)$$

Example 2.13. Power deficiency of the classical F-test in nested designs.

Ananda (1995) used the data shown in table below to demonstrate the lack of power of the classical F-test under heteroscedasticity. He simulated the data from an exact model in which factor A levels are different and the interactions are not. The following table shows the summary statistics computed from the simulated data.

A Level	B Level	Sample Size	Mean	Standard Deviation
A_1	B_1	10	51.13	1.29
A_1	B_2	7	49.15	2.49
A_2	B_3	6	50.01	2.58
A_2	B_4	9	49.26	1.19
A_2	B_5	8	48.99	0.99

It is evident from the sample variances that the assumption of equal variances is not a reasonable one in this situation. If we ignore this fact and proceed with the classical approach, we get the following ANOVA table, which leads us to conclude that none of the effects are statistically significant.

Source	DF	SS	MS	F-value	p-value	Gen. p-value
$B(A)$	3	19.863	6.621	1.922	0.144	0.334
A	1	8.877	8.877	2.577	0.117	0.010
Error	35	120.566	3.445			
Total	39	149.306				

Also included in the ANOVA table are the p-values we get using the generalized test that does not rely on the unreasonable assumption of homoscedastic variances. While we come to the same conclusion about the interactions (confounded with B effects), we now find strong evidence to conclude that the differences in factor A levels are highly significant, which we know to be the right conclusion from the simulated experiment. This example demonstrates the lack of the classical F-test in the presence of unequal error variances, just as was the case in One-Way ANOVA.

Exercises

3.1 Let $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ be a random sample of size n_1 from one population and let $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ be a random sample of size n_2 from a second population. Assume that

$$Y_{1j} \sim N(\mu_1, \sigma_1^2), \quad j = 1, \ldots, n_1$$
$$Y_{2j} \sim N(\mu_2, \sigma_2^2), \quad j = 1, \ldots, n_2 .$$

Let $\overline{Y}_i$ and S_i^2 be the sample mean and the sample variance (MLEs) of the ith population. Show that

$$\overline{Y}_i \sim N(\mu_i, \frac{\sigma_i^2}{n_i}), \ \frac{n_i S_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1}, \ i = 1, 2$$

and that they are independently distributed.

3.2 Consider the two normal samples in Exercise 2.1. Let a and b be two known constants. Assuming that the two variances are equal, derive $100\gamma\%$ right-sided confidence intervals for $\theta = a\mu_1 + b\mu_2$. Also derive right-sided generalized confidence intervals without the equal variances assumption.

3.3 Consider again the two normal samples in Exercise 2.1. Establish a procedure for testing the hypothesis

$$H_0 : \frac{\mu_1 - \mu_2}{\mu_1 + \mu_2} \leq \delta_0$$

when (i) the variances are equal, (ii) the variances are unequal.

3.4 Let $Y_i, \ i = 1, 2$ be two independent random variables distributed as

$$Y_i \ \sim \ \text{Gamma}(\ \alpha_i \ , \ \beta \ ), \quad i \ = \ 1, 2 \ .$$

(a) Show that the random variables

$$B \ = \ \frac{Y_1}{Y_1 \ + \ Y_2} \ \text{ and } \ S \ = \ Y_1 + Y_2$$

are independent.
(b) Show that

$$B \sim \text{Beta}(\ \alpha_1 \ , \ \alpha_2 \ ),$$

and that

$$S \sim \text{Gamma}(\ \alpha_1 + \alpha_2 \ , \ \beta).$$

3.5 Extend the results in Exercise 2.4 to the case of k gamma random variables. Hence show that formulas (3.42) and (3.41) are equivalent.

3.6 By considering the identity

$$\begin{aligned} Y_{ijk} - \overline{Y} &= (Y_{ijk} - \overline{Y}_{ij}) + (\overline{Y}_{i\cdot} - \overline{Y}) + (\overline{Y}_{\cdot j} - \overline{Y}) \\ &\quad + (\overline{Y}_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y}), \end{aligned}$$

squaring and summing it, and then showing that the sum of cross product of any two terms on the right-hand side is equal to zero, prove that

$$S_T = S_A + S_B + S_I + S_E, \tag{3.89}$$

where

$$S_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y})^2,$$

$$S_A = JK\sum_{i=1}^{I}(\overline{Y}_{i\cdot} - \overline{Y})^2,\ S_B = IK\sum_{j=1}^{J}(\overline{Y}_{\cdot j} - \overline{Y})^2,$$

$$S_I = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij} - \overline{Y}_{i\cdot} - \overline{Y}_{\cdot j} + \overline{Y})^2,$$

and

$$S_E = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij})^2\ .$$

3.7 By using (3.82) or otherwise, prove that the generalized p-value for testing the interactions of two-way nested model can be based on the p-value (3.83).

3.8 An agricultural research scientist is interested in comparing four hybrids of corn. The four corn hybrids were planted in a random order in 22 plots of equal size and fairly homogeneous soil conditions. A set of data on yield from corn hybrids obtained from the experiment are shown below:

Hybrid A	7.4, 6.6, 6.7, 6.1, 6.5, 7.2
Hybrid B	7.1, 7.3, 6.8, 6.9, 7.0
Hybrid C	6.8, 6.3, 6.4, 6.7, 6.5, 6.8
Hybrid D	6.4, 6.9, 7.6, 6.8, 7.3

(a) Assuming that the variances of yield from all four hybrids are equal, test whether there is a significant difference between the mean yields.
(b) Carry out the above hypothesis without the assumption of equal variances.
(c) Test whether data provide sufficient evidence to indicate that the variances are not the same.

3.9 Consider the summary data in Table 2.7. Carry out multiple comparisons by the Bonferroni and Scheffe methods to compare pairs of corn hybrids and discuss your findings. Construct generalized Tukey−Kramer intervals under the assumption of equal error variances and without that assumption to compare the three means. Discuss your findings.

3.10 In order to test whether there is no difference in average intelligence of students in two school districts, an IQ test is administered. Suppose only the mean test scores are available to an analyst. The mean test scores by education level and the school district are shown below:

Education level	1	2	3	4	5	6
District 1	68	64	71	74	67	73
District 2	69	72	70	71	77	75

Establish procedures for testing whether there is any difference between the effects of school district and the education level on the IQ scores. Compute $p-$value for testing each hypothesis and discuss your findings.

3.11 In a two-way factorial design, subjects of three age groups were allocated to one of three diet plans during a study period. The reductions in blood pressure due to the diets after the study period are shown in the following table:

	Group 1	Group 2	Group 3
Diet 1	3, 4, -2	4, 3, 5	2, 3, -2
Diet 2	-3, 0, 2	2,-1, 1	2,-1, 0
Diet 3	4, 1, 5	-2, 2, 4	-1, 4, 2

Perform an analysis of variance of this data under the assumption of equal error variances. Compute p-values for testing main effects and the interaction and discuss your findings. Repeat the analysis without the assumption of equal variances and compare the results.

3.12 In order to study the effect of two catalysts and the temperature on the yield of a chemical process, an experiment is carried out using one catalyst at a time, each under three temperatures. The following table shows the results of the experiment, the yields obtained in four runs under each temperature.

	Catalyst A	Catalyst B
Temp. 1	53, 56, 62, 58	59,63, 65, 57
Temp. 2	58, 59, 57, 64	58,62, 67, 66
Temp. 3	59, 54, 61, 60	64,55, 61, 58

Perform an analysis of variance of this data under the assumption of equal error variances. Compute p-values for testing the effects of the catalysts, the

temperature, and their interaction. Repeat the analysis without the assumption of equal variances, discuss the findings, and compare the results.

3.13 Consider the data summarized below from a two-factor design.

A Level	B Level	Sample Size	Mean	Standard Deviation
A_1	B_1	8	72.36	4.87
A_1	B_2	10	63.34	2.69
A_1	B_3	7	64.56	3.02
A_2	B_4	10	69.87	3.78
A_2	B_5	10	70.56	4.19
A_2	B_6	10	65.79	1.99

Perform an analysis of variance of this data under the assumption of equal error variances. Repeat the analysis without the assumption of equal variances, discuss the findings, and compare the results.

CHAPTER 4

INTRODUCTION TO MIXED MODELS

4.1 INTRODUCTION

The purpose of this chapter is to provide an introduction to Mixed Models which will play an important role in the analysis of Repeated Measures. In fact, except for the MANOVA approach, all the models that we will develop in this book for repeated measures analyses fall under the class of mixed models. In the case of Growth Curves we undertake in Chapter 10, except for the GANOVA approach, we will also employ mixed models.

The models that we studied in Chapter 2 are called fixed effects models. In that setting, the levels of each factor were considered as deliberate choices of an experimenter and the assumed models allowed the experimenter to compare the levels. In some other applications, the levels of each factor used in an experimental design are not of particular interest and they are selected at random from a large population of potential levels. Random effects models incorporate this feature explicitly. Yet in another class of application each factor level is of interest, but they are treated as if they are realized values of a distribution and hence treated as random effects. In some other applications, the levels of some of the factors are treated as fixed effects while the levels of

(Generalized Inference in Repeated Measures, Edition 2). By (Weerahandi)

the other factors are treated as random effects. Models having some random effects and some fixed effects are referred to as mixed models.

Random effects models and mixed models arise in the design and analysis of many industrial, biomedical, and agricultural model. Lately mixed models have become very popular among practitioners in Sales & Marketing as well, since use of the Best Linear Unbiased Predictor (BLUP), a notion introduced by Henderson (1975), can provide more accurate estimates of factor levels than the LSE when we treat them as realized values of a certain distribution. As an example of the former, consider the problem of estimating the power consumption of a certain brand of refrigerators. Quantities of primary importance in this application might be the average power consumption and its variance. In designing an experiment to estimate these quantities and in modeling the data from such an experiment, we need to take in to account a number of factors affecting the power consumption. Although individual levels of such factors are important and usually not reported, they contribute to the variance and its structure. Some of the factors affecting the power consumption in this application are the temperature setting in the refrigerator, the external temperature, the refrigerator load, and so on. Table 3.1 shows a hypothetical data set from an experiment which is designed to quantify the effect of just one factor, namely the refrigerator load. In this design, perhaps the temperature setting is set fixed at the average level that households are expected to use. In Chapter 4, we will revisit the problem when the temperature setting is also varied and modeled in the setting of a higher-way mixed model. In this example the number of observations available at different load conditions are not equal. We will refer to such designs as unbalanced designs.

Table 4.1 Energy consumption

Low	Medium	High
9.80	12.13	15.58
10.57	7.84	14.01
10.47	12.54	15.98
8.59	15.02	12.70
8.62	13.17	11.97
11.02	12.60	18.27
7.31	9.92	14.04
12.83	12.84	9.94
10.99	14.26	
8.18	12.05	
	10.66	
	11.55	

As an example of the latter, when one needs to estimate the effect of a TV campaign by Market, one can obtain more accurate estimates of consumer

response to the TV advertisements if the response by market (factor levels) are treated as random effects distributed around the average national response, and using BLUP instead of the LSE as we will discuss later. Although the treatment of factor levels of interest as random effects and use of the BLUP is most desirable in dealing with noisy data, they are also useful in designed experiments when one has to work with small samples.

The variances of random effects models and in mixed models are referred to as variance components. In random effects models, we are mainly interested making inferences about variance components. In mixed models, we would be interested in making inferences about both the means of fixed effect terms and the values of the variance components. For further discussion on variance components and mixed models the reader is referred to Khuri, Mathew, and Sinha (1998).

4.2 RANDOM EFFECTS ONE-WAY ANOVA

First consider the simplest possible random effects model involving just one factor, say factor A, and no fixed effect terms. In other words, this model has only two sources of variation, namely the variation due to randomly selected factor levels and the overall sampling variation. Let

$$A_1, A_2, \ldots, A_i, \ldots, A_k$$

be the factor levels. In the balanced case of the problem, which has implications on higher-way mixed models, we have n observations corresponding to each of the k factor levels. Later in this chapter we will consider the case of the unequal number of observations available from different levels of the factor. Let $Y_{i1}, Y_{i2}, \ldots, Y_{in}$ denote the sample of data available from the ith factor level. They are also known as the observed values of the response variable. Let α_i be the random effects corresponding level A_i of factor A. Assume that

$$\alpha_i \ \sim \ N(0, \sigma_\alpha^2), \tag{4.1}$$

where σ_α^2 is the population variance of the random factor. In the random effects model, while individual α_i terms are of no particular interest in some applications they are considered important in other applications. The variance term σ_α^2, called the factor *variance component,* is of practical importance in both applications. Let μ denote the population mean of all responses; i.e., $E(Y_{ij}) \ = \mu$. Further assume a linear model of the form

$$Y_{ij} \ = \mu \ + \ \alpha_i \ + \ \epsilon_{ij}, \text{ for } i = 1, \ldots, k, \ j = 1, \ldots, n, \tag{4.2}$$

where ϵ_{ij} is the error term representing the deviation of the response of the jth observation from the mean of observations from A_i. Assume that

$$\epsilon_{ij} \ \sim \ N(0, \sigma_\epsilon^2) \tag{4.3}$$

and that α_i, ϵ_{ij}; $i = 1, \ldots, k,\ j = 1, \ldots, n$ are mutually independent. Note that an assumption underlying the above model is that

$$E(Y_{ij} \mid \alpha_i\ ) \ = \mu \ + \alpha_i.$$

The model also implies that

$$\mathrm{Var}(Y_{ij}) \ = \ \sigma_\alpha^2 \ + \ \sigma_\epsilon^2\ .$$

The terms σ_α^2 and σ_ϵ^2 are referred to as the variance components of the model. Moreover, σ_α^2 is sometimes referred to as the *factor variance* and σ_ϵ^2 is referred to as the *error variance*.

A related model that arise in the analysis of repeated measures is based on the less restrictive assumption that

$$E(Y_{ij} \mid \alpha_i\ ) \ = \mu_j \ + \alpha_i.$$

We defer further discussion and problem of making inferences on μ_j parameters and the variance components until Chapter 7.

Consider the problem of making inferences about the common mean μ and the variance components σ_α^2 and σ_ϵ^2. Let $\overline{Y}_i,\ i = 1, \ldots, k$, be the sample means corresponding to the k random effects and let $\overline{Y}$ be the mean of all data. As in the one-way ANOVA fixed effects model, the decomposition of the total sum of squares given by

$$S_T \ = \ S_E \ + \ S_B \ = \ \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} \ - \ \overline{Y})^2$$

play an important role in random effects as well, where

$$S_B \ = \ n\sum_{i=1}^{k}(\overline{Y}_i. \ - \ \overline{Y})^2 \text{ and } S_E \ = \ \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} \ - \ \overline{Y}_i)^2 \qquad (4.4)$$

are the between-group sum of squares and the error sum of squares. It is easily seen (see Appendix A.5) that they are independently distributed as

$$\frac{S_E}{\sigma_\epsilon^2} \sim \ \chi^2_{N-k} \qquad (4.5)$$

and

$$\frac{S_B}{\sigma_\epsilon^2 \ + n\sigma_\alpha^2} \sim \ \chi^2_{k-1}, \qquad (4.6)$$

and $N = nk$.

These distributional results will play a key role in all types of inferences about the parameters of the model, as we will see later in this section. When taking the generalized approach, there is also no difficulty doing inferences more generally when the design is unbalanced in the sense that we do not

have an equal number of observations from the k groups. If the sample size available from group i is n_i, then in place of (4.6), we could use a result due to Wald (1940). When the sample sizes are different, we would use the result

$$\frac{S_{wB}}{\sigma_\epsilon^2} \sim \chi^2_{k-1} \tag{4.7}$$

and redefine the sums of squares and the total sample size as

$$S_{wB} = \sum_{i=1}^{k} w_i \left(\overline{Y}_i - \frac{\sum_{i=1}^{k} w_i \overline{Y}_i}{\sum_{i=1}^{k} w_i} \right)^2, \tag{4.8}$$

$$S_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2, \tag{4.9}$$

and $N = \sum n_i$, where

$$w_i = \frac{n_i}{1 + n_i \rho_\alpha}, \ \overline{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i, \quad \text{and} \quad \rho_\alpha = \sigma_\alpha^2 / \sigma_\epsilon^2.$$

The design with unequal samples sizes will be referred to as the unbalanced design.

4.3 POINT ESTIMATION

The grand mean μ appearing in the linear model (4.2) is estimated by the sample mean $\overline{Y}$, an unbiased estimator that can be derived using the MLE or LSE methods. When the random effects are of particular interest as we discussed above they are not estimated, but rather predicted using the BLUP formula given by Henderson (1975) as

$$\widehat{\alpha}_i = E(\alpha_i | \overline{Y}_i) = \varpi \mu + (1 - \varpi) \bar{y}_1, \text{ where } \varpi = \frac{\sigma_e^2}{\sigma_\epsilon^2 + n\sigma_\alpha^2}. \tag{4.10}$$

The above formula of the BLUP involve unknown parameters and so they need to be estimated. The LSE estimates of the variance components σ_e^2 and σ_α^2 are

$$\widehat{\sigma}_e^2 = MS_E$$

and

$$\widehat{\sigma}_a^2 = \frac{MS_B - MS_E}{n}, \tag{4.11}$$

respectively, where

$$MS_B = \frac{S_B}{k-1} \text{ and } MS_E = \frac{S_B}{N-k}. \tag{4.12}$$

While the estimator of the error variance have all desirable properties, that is not the case with the estimate of the factor variance, because with a fraction of possible values of the sample space the estimator could become negative. The reader is referred to Weerahandi (2012) for a detailed discussion of the issues and for a class of estimators that do not suffer from this drawback.

The most popular and widely used methods of estimation of the BLUP are ML and REML discussed in Searle, Casella, and McCulloch (1992). However, as Yu et al (2013) argued, these methods have serious drawbacks when the number of factor levels is not large. The reader is referred to Yu et al (2013) and Gamage et al (2013) for improved point estimation methods of the BLUP and interval estimation methods, respectively. Next we consider the problem of making inferences about the variance components, which are important regardless of whether or not a practitioner considers random effects are of particular interest or not.

4.4 INFERENCE ABOUT VARIANCE COMPONENTS

It is straightforward to make inferences about the error variance σ_ϵ^2 based on the distributional result (4.7). For example, $100\gamma\%$ upper confidence bound for σ_e^2 is obtained from the probability statement

$$\Pr(\frac{S_E}{\sigma_\epsilon^2} \geq c_\gamma) = \gamma.$$

Obviously the resulting confidence interval for σ_ϵ^2 is $[0, s_E/c_\gamma]$, and c_γ is the $(1-\gamma)$th quantile of the chi-squared distribution with $N-k$ degrees of freedom. The result is valid regardless of whether or not we have equal sample sizes from the k groups, provided that formula (4.8) is used to compute the error sum of squares S_E. In testing, hypotheses of the form $H_0 : \ \sigma_e \leq \ \sigma_0$ are rejected at α level if $\sigma_0 > s_E/c_\alpha$.

$$H_0 : \ \sigma_e \leq \ \sigma_0 \ \ .$$

Finally, an unbiased estimate of the error variance can also be obtained from (4.6) as

$$\widehat{\sigma}_\epsilon^2 = \frac{S_E}{N-k},$$

a result that follows from the fact that the mean of a chi-squared distribution is the same as its degrees of freedom.

Inferences on the ratio $\rho_\alpha = \sigma_\alpha^2 / \ \sigma_\epsilon^2$ of the two variance components can also be obtained by classical methods. In general for the unbalanced case, tests

and confidence intervals concerning ρ_α can be obtained from the distributional result

$$\frac{S_{wB}/(k-\ 1)}{S_E/(N\ -\ k))}\ \ \sim\ F_{k-1,N-k}, \tag{4.13}$$

which follows from (4.5) and (4.7). For example, equal-tail $100\gamma\%$ confidence bound for ρ_α is obtained from the probability statements

$$\Pr(\frac{S_{wB}/(k-\ 1)}{S_E/(N\ -\ k))} \leq F_{\frac{1-\gamma}{2}}) = \frac{1-\gamma}{2}$$

and

$$\Pr(\frac{S_{wB}/(k-\ 1)}{S_E/(N\ -\ k))} \geq F_{\frac{1+\gamma}{2}}) = \frac{1-\gamma}{2},$$

where F_γ is the γth quantile of the F distribution with $k-1$ and $N-k$ degrees of freedom. Let $s_{wB}(\rho)$ denote the observed value of sum of squares S_{wB} when $\sigma_\alpha^2/\ \sigma_\epsilon^2 = \rho$. Now it is clear that, equal-tail $100\gamma\%$ confidence bound for $\rho_\alpha$can be expressed as

$$s_{wB}^{-1}(\frac{k-1}{N-k}s_E F_{\frac{1-\gamma}{2}}) \leq \rho_\alpha \leq s_{wB}^{-1}(\frac{k-1}{N-k}s_E F_{\frac{1+\gamma}{2}}), \tag{4.14}$$

where $s_{wB}^{-1}()$ is the inverse function of $s_{wB}()$. This confidence interval is referred to as the *Wald interval*. Its computation involve numerically solving non-linear equations, but it can be conveniently obtained using the XPro software package. In the balanced case of equal sample sizes from the k groups (4.13) reduces to

$$F\ = \frac{1}{1+n\rho_\alpha}\frac{S_B/(k-\ 1)}{S_E/(N\ -\ k)}\ \ \sim\ F_{k-1,N-k},$$

and thus the confidence interval reduces to

$$\frac{1}{n}\left(\frac{(k-1)s_E}{(N-k)s_B F_{\frac{1-\gamma}{2}}} - 1\right) \leq \rho_\alpha \leq \frac{1}{n}\left(\frac{(k-1)s_E}{(N-k)s_B F_{\frac{1+\gamma}{2}}} - 1\right). \tag{4.15}$$

As before, testing of hypotheses can be based on the confidence interval or derived directly from the F distribution. To be specific consider the problem of testing

$$H_0:\ \rho_\alpha \leq\ \rho\ \text{ versus } H_1:\ \rho_\alpha\ >\ \rho\ .$$

First consider the balanced case. Define

$$T\ =\ \frac{S_B/(k-\ 1)}{S_E/(N\ -\ k)}.$$

Obviously, T is stochastically increasing in ρ_α, the parameter of interest. Hence, the p-value for testing H_0 can be obtained as

$$\begin{aligned} p &= \Pr(T \geq t_{obs} \mid \rho_\alpha = \rho) \\ &= 1 - Pr(F \leq t_{obs}/(1+n\rho)) \\ &= 1 - H_{k-1,N-k}\left(\frac{(N-k)s_B}{(1+n\rho)(k-1)s_E}\right) \end{aligned}$$

where $H_{k-1,k(n-1)}$ is the cdf of the F distribution with $k-1$ and $k(n-1)$ degrees of freedom. More generally, regardless whether or not we have equal samples sizes from the k groups, we could use the corresponding extreme region given by the statistic S_{wB}/S_E and compute the p-value as

$$p = 1 - H_{k-1,N-k}\left(\frac{(N-k)s_{wB}(\rho)}{(k-1)s_E}\right). \tag{4.16}$$

Example 3.1. Variance due to refrigerator load. Consider the data set in Table 3.1. Suppose an analyst is interested in testing the hypothesis that the variance due to the refrigerator load is less than 50% of the total variance. This hypothesis testing problem here is equivalent to the hypothesis

$$H_0 : \rho_\alpha \leq 1 \; versus \; H_1 : \rho_\alpha > 1 .$$

When $\rho_\alpha = 1$, the weighted between group sum of squares is computed using (4.8) and the error sum of squares is computed using (4.9). Then, the p-value computed using (4.16) is $p = 0.39$. With this p-value there is no reason to doubt the null hypothesis. Moreover, confidence intervals for the percent variance,

$$\begin{aligned} \phi &= 100\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2} \\ &= 100\frac{\rho_\alpha}{\rho_\alpha + 1}, \end{aligned}$$

can be computed by first obtaining the confidence interval for ρ_α using (4.14) and then converting it to a confidence interval for ϕ. The 95% confidence interval for the percent factor variance due to refrigerator load obtained in this manner is

$$12.4 \leq \phi \leq 97.7.$$

This confidence interval also leads to the same conclusion concerning the hypothesis of interest.

4.4.1 Testing the factor variance

Testing the hypothesis of zero factor variance, namely $H_0 : \sigma_\alpha^2 = 0$, follows from (4.16), because the hypothesis is equivalent to $H_0 : \rho_\alpha = 0$. It could also be derived from (4.5) and (4.6). In fact, that it leads to an F-test follows from the analysis of variance table shown below. The fourth column of the table shows the expected values of the mean sum of squares.

Table 4.2 Random effects ANOVA: Expected values

Source	DF	SS	E(SS)
Between	$k - 1$	S_B	$(k-1)(\sigma_\epsilon^2 + n\sigma_\alpha^2)$
Within	$N - k$	S_E	$(N-k)\sigma_\epsilon^2$
Total	$N - 1$	S_T	

It is easily deduced from the expected values of the ANOVA table that an unbiased estimate of the variance component σ_a^2 based on the sufficient statistics is

$$\widehat{\sigma}_a^2 = \frac{MS_B - MS_E}{n}, \tag{4.17}$$

where

$$MS_B = \frac{S_B}{k-1} \text{ and } MS_E = \frac{S_B}{N-k}. \tag{4.18}$$

As discussed before a major drawback of the unbiased estimate is that with some data sets the estimate can become negative. The MLE of the variance component basically have the same drawback except for that with such data sets the estimate could become zero. Therefore, in making inferences with variance components it is better to report confidence intervals. In fact, 50% confidence bound of σ_α^2 can be treated as a better point estimate.

The computation of the F-Value for testing the hypothesis can also be conveniently set out in the ANOVA table as shown below. In fact, under $H_0 : \sigma_\alpha^2 = 0$, $w_i = n_i$ and so regardless of whether or not the design is balanced, a test of H_0 can be based on the F-Statistic

$$\frac{MS_{wB}}{MS_E} \backsim F_{k-1,N-k},$$

where

$$MS_{wB} = \frac{\sum_{i=1}^{k} n_i \left(\overline{Y}_i - \frac{\sum_{i=1}^{k} n_i \overline{Y}_i}{\sum_{i=1}^{k} n_i} \right)^2}{k-1}.$$

Table 4.3 Random effects ANOVA: F-Values.

Source	DF	SS	MS	F-value
Between	$k-1$	S_B	$MS_B = \frac{S_B}{k-1}$	$\frac{MS_B}{MS_E}$
Within	$N-k$	S_E	$MS_E = \frac{S_E}{N-k}$	
Total	$N-1$	S_T		

4.4.2 General hypotheses and interval estimation

Except for the particular test, $H_0: \sigma_\alpha^2 = 0$, making inferences about the variance component σ_α^2 is not an easy task. For instance, the above F-test has no implication in interval estimation. In fact classical approach fails to provide exact solutions to testing problems of the form $H_0: \sigma_\alpha^2 \geq \sigma_0^2$ and hence in interval estimation. A number of authors including Satterthwaite (1946), Welch (1956), Bulmer (1957), and Samaranayake and Bain (1988) provided approximate confidence intervals for σ_α^2. Weerahandi (1993, 1995) provided tests and confidence intervals using the generalized approach.

To obtain exact tests and generalized confidence intervals for σ_α^2 more generally for the unbalanced case, consider the potential generalized test variable suggested by the results,

$$U = \frac{S_E}{\sigma_\epsilon^2} \sim \chi^2_{N-k} \quad \text{and} \quad V = \frac{S_{wB}}{\sigma_\epsilon^2} \sim \chi^2_{k-1},$$

namely

$$\begin{aligned} T &= \frac{s_E S_{wB}}{S_E s_{wB}\left(\frac{\sigma_\alpha^2}{\sigma_\epsilon^2}\frac{S_E}{s_E}\right)} \\ &= \frac{V}{U}\frac{s_E}{s_{wB}\left(\frac{U\sigma_\alpha^2}{s_E}\right)}. \end{aligned} \tag{4.19}$$

It is easily seen that the distribution of the test variable is free of nuisance parameters, and it reduces 1 to at the observed values of the random variables. Moreover, except in extreme cases of unbalances, the T is stochastically increasing in the variance component. In the balanced case of equal sample sizes, it reduces to

$$\begin{aligned} T &= \frac{V}{U}\frac{s_E}{s_B}\frac{1+n\frac{U\sigma_\alpha^2}{s_E}}{n} \\ &= \frac{s_E}{ns_B}(\frac{1}{U}+\frac{n\sigma_\alpha^2}{s_E}), \end{aligned}$$

making it clear that T is stochastically increasing. Hence the generalized p-value for testing $H_0: \sigma_\alpha^2 \geq \sigma_0^2$ can be obtained as

$$
\begin{aligned}
p &= \Pr(T \leq 1) \\
&= \Pr(V \leq \frac{U}{s_E} s_{wB}(\frac{U\sigma_0^2}{s_E})) \\
&= EG_{k-1}\left(\frac{U}{s_E} s_{WB}(\frac{U\sigma_0^2}{s_E})\right),
\end{aligned} \tag{4.20}
$$

where G_{k-1} is the cdf of the chi-squared distribution with $k-1$ degrees of freedom and the expectation is taken with respect to the random variable $U \sim \chi^2_{N-K}$. In the balanced case, the test reduces to

$$
p = EG_{k-1}\left(\frac{s_B}{\frac{s_E}{U} + n\sigma_\alpha^2}\right). \tag{4.21}
$$

Various generalized confidence intervals for the factor variance component can be derived from a generalized pivotal or deduced directly from the above p-value. It is evident that if σ_1^2 and σ_2^2 are chosen such that

$$
EG_{k-1}\left(\frac{U}{s_E} s_{WB}(\frac{U\sigma_1^2}{s_E})\right) = \frac{1-\gamma}{2} \tag{4.22}
$$

and

$$
EG_{k-1}\left(\frac{U}{s_E} s_{WB}(\frac{U\sigma_2^2}{s_E})\right) = \frac{1+\gamma}{2}, \tag{4.23}
$$

then $[\sigma_1^2, \sigma_2^2]$ is a $100\gamma\%$ generalized confidence interval for σ^2. This confidence interval can be conveniently computed using the XPro software package.

Example 3.2. Variance due to refrigerator load (continued)

Consider again the data set in Table 3.1. In view of the fact that the point estimates of variance components are not reliable (e.g., widely used unbiased estimates could even become negative) and do not provide adequate information about the magnitude of the variance component, suppose the analyst wishes to report a lower confidence bound for the variance due to the refrigerator load. This can be accomplished by applying the formula

$$
EG_{k-1}\left(\frac{U}{s_E} s_{WB}(\frac{U\sigma_0^2}{s_E})\right) = 1-\gamma,
$$

where σ_0^2 represents the desired lower confidence bound. The 95% upper confidence interval obtained with data from Table 3.1 is $(\sigma_\alpha^2 \geq .981]$. Moreover, the generalized p-value for testing the hypothesis $H_0: \sigma_\alpha^2 \leq 1$ is 0.05, indicating that we have sufficient evidence to conclude that the magnitude of the variance component is not smaller than 1.

4.5 FIXED-LEVEL TESTING

Suppose a practitioner is interested in testing hypotheses of σ_α^2 at a certain fixed nominal level such as the 0.05 level or constructing 95% confidence intervals for σ_α^2. Also suppose that the practitioner is interested in ensuring that in repeated sampling, the tests will have the intended Type I error and the intervals will have the intended frequency coverage. Unfortunately there are no tests (and hence there are no confidence intervals) that can attain the intended level exactly for all possible values of the parameter σ_ϵ^2. As a result, practitioners resort to asymptotic and other types of approximate solutions. However, there are tests that do not exceed the intended Type I error for all possible values of the nuisance parameter. Perhaps the only class of classical confidence intervals (and hence the corresponding tests) having this property is the one proposed by Tukey (1951) and Williams (1962). The main drawback of Tukey−Williams intervals is that they are highly conservative in that the intervals tend to be too wide. According to simulation studies [cf. Weerahandi and Amaratunga (1999)], one can obtain much more powerful procedures having the desired property by taking the generalized approach. In fixed-level testing with the generalized p-value given in the previous section, one simply rejects the null hypothesis if the p-value is less than the specified Type I error level. For other ways of obtaining more powerful procedures based on the Tukey−Williams intervals, the reader is referred to Samaranayake and Bain (1988) and Wang (1990).

Despite the availability of these procedures having excellent frequency properties, the most widely used procedures in making inferences about variance components are perhaps the likelihood based methods, especially the ML (maximum likelihood) and REML (restricted maximum likelihood) discussed by Searle, Casella, and McCulloch (1992). The reason for wide use of these procedures is perhaps because they are readily available from SAS PROC MIXED. Unfortunately, the ML- and REML-based inference is probably the worst choice one can make in most applications of mixed models including applications involving higher-way mixed models. This is because the ML- and REML-based tests and confidence intervals have very poor size performance. Shown in the table below is a set of simulation results carried out by Weerahandi and Amaratunga (1999) to compare the size performance of ML-based method against the generalized test. When σ_ϵ^2 is fixed at 1.0 (without loss of generality) and σ_a^2 takes on values 0.01, 0.1, 1.0, and 10.0, they estimated the actual size of each procedure using 10,000 simulated samples. Since the actual size of ML-based test is so large, the estimates are reported here only up to 2 decimal places. Observe that the true size of ML-based tests can be as large as 0.6 and that of REML can be as large as 0.4, highly prohibitive levels by any standard. Also note that the performance of ML-based tests is worse than that of REML despite the fact that the advocates of likelihood-based procedures prefer ML over REML. The size of the generalized test is

0.05 up to the reported accuracy of two decimal points; when computed up to 3 decimal points they range from about 0.048 to 0.050.

Although Weerahandi and Amaratunga (1999) reported the results for the one-way random effects model , the results have implications for variance components in any higher-way mixed model having the canonical form that we discuss in the next chapter. For example, the results for the case applies to any situation with the canonical form having chi-squared random variables with 4 and 45 degrees of freedom, which are quite typical especially in higher way mixed models. Notice that ML-based tests becomes somewhat reasonable only when the first degree of freedom is very large, a result that is also seen from the asymptotic variance of ML estimates. Unfortunately, this is a parameter that cannot be increased by increasing the number of replicates in a cell. In fact the first degree of a freedom is usually related to the levels of a factor, something that should be set at a low value (to avoid even more serious practical problems) at the design stage of experiments. Burdick and Larsen (1997) provide some simulation results for another class of applications where ML-based procedures have very poor size performance. In view of these considerations and due to very serious size problems of ML-based procedures in variance components, their use in mixed models is highly discouraged.

Actual sizes of tests with intended level 0.05.

σ_a^2:	0.01	0.1	1.0	10.0
Method\Case:	$k = 2$, $n = 10$			
ML	0.51	0.57	0.58	0.59
REML	0.33	0.38	0.40	0.40
Generalized	0.05	0.05	0.05	0.05
Method\Case:	$k = 5$, $n = 10$			
ML	0.25	0.30	0.31	0.30
REML	0.17	0.20	0.21	0.21
Generalized	0.05	0.05	0.05	0.05
Method\Case:	$k = 5$, $n = 1000$			
ML	0.31	0.31	0.30	0.31
REML	0.21	0.20	0.20	0.21
Generalized	0.05	0.05	0.05	0.05
Method\Case:	$k = 10$, $n = 10$			
ML	0.17	0.19	0.20	0.20
REML	0.12	0.13	0.14	0.14
Generalized	0.05	0.05	0.05	0.05
Method\Case:	$k = 100, n = 100$			
ML	0.07	0.07	0.07	0.06
REML	0.06	0.06	0.06	0.06
Generalized	0.05	0.05	0.05	0.05

4.6 INFERENCE ABOUT THE MEAN

In some applications the parameter of primary interest might be the mean μ while in some other applications σ_α^2 might be of more importance. In all applications the mean and the two variance components σ_α^2 and σ_ϵ^2 all are of some importance. In any case first consider the problem of making inferences about the mean μ. Consider the problem in general for the unbalanced case where we have unequal sample sizes from the k groups.

From (4.2) we get

$$\overline{Y}_i = \mu \ + \alpha_i + \overline{\epsilon}_i,$$

and in turn that

$$\overline{Y}_i \sim \ N(\mu, \ \sigma_\alpha^2 \ + \ \sigma_\epsilon^2/n_i), \quad i = 1, \cdots, k \ . \tag{4.24}$$

It is obvious (see also Exercise 3.1) from (4.24) that if the weights were known, then the MLE of μ is the weighted sample mean

$$\overline{Y} = \frac{\sum_{i=1}^{k} w_i \overline{Y}_i}{\sum_{i=1}^{k} w_i}.$$

But the parameter ρ_α appearing in the definition of w_i is unknown, and it needs to be estimated with the aid of (4.6). Before we address this issue, note that in the balanced case the weighted sample mean reduces to a simple average of sample means, a quantity having no unknown parameters. In the balanced case, making inferences on μ is trivial as they follow from the results

$$\overline{Y} \sim N(\mu, \frac{1}{k}(\sigma_\alpha^2 + \sigma_\epsilon^2/n))$$

and (4.6), which implies that

$$(\overline{Y} - \mu)\sqrt{\frac{N(k-1)}{S_B}} \sim t_{k-1}. \tag{4.25}$$

4.6.1 The unbalanced case

When the sample sizes are unequal, procedures for making inferences about μ can be obtained by taking the generalized approach. To do this we can start with the distribution of $\overline{Y}$ given by

$$\overline{Y}(\rho_\alpha) \sim N(\mu, \frac{\sigma_\epsilon^2}{\sum_{i=1}^{k} w_i(\rho_\alpha)}). \tag{4.26}$$

To make inferences on μ by taking the generalized approach, we can use the distributional results

$$Z = \frac{\overline{Y}(\rho_\alpha) - \mu}{\sigma_\epsilon}\sqrt{\sum_{i=1}^{k} w_i(\rho_\alpha)},$$

$$W_1 = \frac{S_E}{\sigma_\epsilon^2} \sim \chi^2_{k-1} \quad \text{and} \quad W_2 = \frac{S_{wB}(\rho_\alpha)}{\sigma_\epsilon^2} \sim \chi^2_{k-1}. \tag{4.27}$$

Hence,

$$\sigma_\epsilon^2 = \frac{S_E}{W_1} \text{ and } \rho_\alpha = S_{wB}^{-1}(\frac{W_2 S_E}{W_1}).$$

Now by applying the substitution method we can obtain a potential generalized pivotal quantity as

$$R = \overline{y}(s_{wB}^{-1}(\frac{W_2 s_E}{W_1})) - Z\sqrt{\frac{\frac{s_E}{W_1}}{\sum_{i=1}^{k} w_i(s_{wB}^{-1}(\frac{W_2 s_E}{W_1})}},$$

$$= \overline{y}(s_{wB}^{-1}(\frac{S_{wB}(\rho_\alpha)s_E}{S_E})) - (\overline{Y} - \mu)\sqrt{\frac{s_E\sum_{i=1}^{k} w_i(\rho_\alpha)}{S_E\sum_{i=1}^{k} w_i(s_{wB}^{-1}(\frac{S_{wB}(\rho_\alpha)s_E}{S_E})}}. \quad (4.28)$$

From the first representation of R above it is clear that the distribution R is free of unknown parameters. From the second representation it is clear that its observed value, $R_{obs} = \mu$, does not depend on nuisance parameters. Therefore, R is indeed a generalized pivotal quantity. Moreover $T = R - \mu$ is a generalized test variable, which is stochastically decreasing in μ.

Now any type of generalized confidence interval or test could be based on R. For example, if a constant k is chosen to satisfy the equation

$$\gamma = \Pr(R \leq r_\gamma), \quad (4.29)$$

then $\mu \leq r_\gamma$is a $100\gamma\%$ generalized confidence interval for μ. The quantiles of the distribution of R can be easily found by simulating the distribution of R using random numbers from the standard normal random variable Z and the chi-squared random variables W_1 and W_2. It can be deduced from this confidence interval or derived from T that the generalized p-value for testing hypotheses of the form $H_0 : \mu \leq \mu_0$ is

$$p = 1 - \Pr(R \leq \mu_0). \quad (4.30)$$

4.7 TWO-WAY MIXED MODEL WITHOUT REPLICATES

When the design is balanced, the above results can be easily extended to the case of one fixed effect and one random effect. To do this, suppose we have data from a two-way layout without replicates as shown in the table below. Suppose in the design of the experiment, the levels of factor A are randomly chosen from a population of possible levels. Suppose we are interested in the fixed effects of the factor B. This design is used, for instance, in the classical randomized block design, where blocking is introduced for the purpose of reducing the error variance.

Two-way layout

A and B Levels	B_1	...	B_j	...	B_n	Means
A_1	y_{11}	...	y_{1j}	...	y_{1n}	$\overline{y}_{1.}$
A_2	y_{21}	...	y_{2j}	...	y_{2n}	$\overline{y}_{2.}$
...	...	...	...	...	...	...
A_i	y_{i1}	...	y_{ij}	...	y_{in}	$\overline{y}_{i.}$
...	...	...	...	...	...	...
A_k	y_{k1}	...	y_{kj}	...	y_{kn}	$\overline{y}_{k.}$
Means	$\overline{y}_{.1}$	...	$\overline{y}_{.j}$	...	$\overline{y}_{.n}$	$\overline{y}$

Let $B_1,\ B_2, \ldots,\ B_n$ be the levels of factor B and we have k levels of the random effect A. In the two-way layout with no replicates we have just one data for each combination of factor levels. Suppose the observations obtained from this design follow the linear model

$$Y_{ij} \ = \ \mu \ + \ \alpha_i \ + \ \beta_j \ + \ \epsilon_{ij}, \quad i = 1, \ldots, k, \ j = 1, \ldots, n, \tag{4.31}$$

where α_i is the ith effect of random factor A and β_j is the jth effect of factor B standardized such that $\sum_{j=1}^{n} \beta_j \ = \ 0$. Assume that

$$\epsilon_{ij} \ \sim \ N(0, \sigma_\epsilon^2) \text{ and that } \alpha_i \ \sim \ N(0, \sigma_\alpha^2), \tag{4.32}$$

We are interested in comparing the treatment means and also making inference about the variance component σ_α^2. Let $\overline{Y}_{i.}$and $\overline{Y}_{.j}$ denote the column and row means of the above table. Let $\overline{Y}$ be the average of all kn data. As in the case of the two-way ANOVA fixed effects given in the Appendix, here also we can base our analysis on the sum of squares decomposition

$$S_T = S_A \ + S_B \ + S_E,$$

where

$$S_T \ = \ \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} \ - \ \overline{Y})^2, \tag{4.33}$$

$$S_A \ = \ n \sum_{i=1}^{k} (\overline{Y}_{i.} \ - \ \overline{Y})^2, \tag{4.34}$$

$$S_B \ = \ k \sum_{j=1}^{n} (\overline{Y}_{.j} \ - \ \overline{Y})^2, \tag{4.35}$$

and

$$S_E = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y})^2 . \tag{4.36}$$

It is easily derived or deduced from the normal distribution theory that each of these random variables can be transformed into chi-squared random variables and that S_A, S_B, and S_E are independently distributed. For example, averaging (4.31) j over, we get

$$\overline{Y}_{i.} \sim N(\mu, \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n}),$$

which implies that

$$\frac{S_A}{\sigma_\epsilon^2 + n\sigma_\alpha^2} \sim \chi^2_{k-1}. \tag{4.37}$$

Similarly, the distribution of S_B is obtained by first averaging over i, and then over j to get

$$\overline{Y}_{.j} - \overline{Y} = \beta_j + (\overline{\epsilon}_{.j} - \overline{\epsilon}).$$

This equation implies that

$$\frac{S_B - k\sum_{j=1}^{n}\beta_j^2}{\sigma_\epsilon^2} \sim \chi^2_{n-1}. \tag{4.38}$$

Similarly, we can obtain the distribution of S_E as

$$\frac{S_E}{\sigma_\epsilon^2} \sim \chi^2_{(n-1)(k-1)}. \tag{4.39}$$

As displayed in Table 3.4 and Table 3.5, these results can be summarized in the form of an ANOVA table.

Table 4.4 Two-way ANOVA mixed model: Expected values

Source	DF	SS	E(SS)
α	$k-1$	s_A	$(k-1)(\sigma_\epsilon^2 + n\sigma_\alpha^2)$
β	$n-1$	s_B	$(n-1)\sigma_\epsilon^2 + k\sum_{j=1}^{n}\beta_j^2$
Error	$(n-1)(k-1)$	s_E	σ_ϵ^2
Total	$nk-1$	s_T	

It is clear from the ANOVA table that the hypothesis of zero variance due to Factor A and the hypothesis of no effect due to Factor B, namely $H_0 : \beta_1 = \beta_2 = ... = \beta_n$ both can be tested using F-tests. The computation of the F-Values are summarized in Table 3.6. For example, H_0

is rejected at the α level if $F_B > F_\alpha$, where F_α is the $(1-\alpha)$th quantile of the F-distribution with $n-1$ and $(n-1)(k-1)$ degrees of freedom. Testing of nonzero variance components could be carried out as in Section 3.3 by taking the generalized approach. In fact tests and intervals on σ_α^2 could be deduced from the results that we will establish in Chapter 4 for any variance component that follows the canonical form of the distributional structure. In the current problem, the distributional results (4.37) and (4.39) implies that the variance component follows the canonical form.

Two-way ANOVA mixed model: F-values

Source	DF	SS	MS	F-Value
α	$k-1$	s_A	$MS_A = \frac{s_A}{k-1}$	$F_A = \frac{MS_A}{MS_E}$
β	$n-1$	s_B	$MS_B = \frac{s_B}{n-1}$	$F_B = \frac{MS_B}{MS_E}$
Error	$(n-1)(k-1)$	s_E	$MS_E = \frac{s_E}{(n-1)(k-1)}$	
Total	$nk-1$	s_T		

Example 3.3. Mileage variance due to driver

The table below shows a set of hypothetical data on highway mileage (average miles per gallon) obtained by a random sample of seven operators driving a certain model of a compact car on three type of traffic conditions, light, average, and heavy (yet without full congestion at any part of the highway). In the table, the three levels of traffic conditions are denoted as T_1, T_2, T_3, respectively.

Driver mileage by traffic condition

Driver\Traffic	T_1	T_2	T_3
D_1	30.3	28.7	24.7
D_2	29.9	28.9	25.4
D_3	31.4	27.6	26.2
D_4	29.2	30.1	29.1
D_5	28.9	23.9	20.7
D_6	30.7	32.1	27.2
D_7	28.4	30.2	27.4

It is obvious in this example that the mileage under different traffic conditions are different and that it is decreasing with the traffic condition. We can carry out a formal ANOVA assuming a two-way mixed effects model with random effects due to the drivers and fixed effects due to the traffic conditions. The ANOVA with F-Values for testing the significance of each of these factors is shown below.

ANOVA for driver and traffic effects

Source	DF	SS	MS	F-Value
Driver	6	56.63	9.45	3.43
Traffic	2	60.74	30.37	11.02
Error	12	33.08	2.76	
Total	21	s_T		

The p-value for testing the hypothesis of zero variance due to driver is $p = 1-H_{6,12}\,(3.43) = 0.033$, implying that we have sufficient evidence to reject the hypothesis and conclude that there is a variation of mileage due to the driver. The point estimate of its variance component is $\widehat{\sigma}_{\alpha}^{2} = (MS_A - MS_E)/n = (9.45 - 2.76)/2 = 3.35$. The p-value for testing the equality of traffic effects is $p = 1-H_{2,12}\,(11.02) = 0.002$. As expected, the hypothesis needs to be rejected since the traffic effects on the mileage is highly significant. The estimated traffic effects at the light, the average, and the heavy traffic conditions are 29.8, 28.8, and 25.8, respectively.

Exercises

4.1 Let Y_i be an observation from the normal distribution

$$Y_i \sim N(\mu, \ \sigma_\epsilon^2/w_i), \quad i = 1, \ldots, k.$$

If $Y_i, \ i = 1, \ldots, k$ are independently distributed and $w_i, \ i = 1, \ldots, k$, is a set of known weights, show that

$$\overline{Y} = \frac{\sum_{i=1}^{k} w_i Y_i}{\sum_{i=1}^{k} w_i}$$

is the maximum likelihood estimate (MLE) of the parameter μ. Also derive the MLE of σ_ϵ^2.

4.2 Consider again the random sample given in Exercise 3.1. Find the distributions of the estimates of μ and σ_ϵ^2. Are they independently distributed? Establish the form of confidence intervals for the parameter $\theta = \mu + \sigma_\epsilon$.

4.3 Consider the one-way random effects model with unequal sample size considered in Section 3.1. Establish a procedure for testing hypotheses concerning the parameter

$$\theta = \frac{\mu}{\sigma_\epsilon + \sigma_\alpha}.$$

Deduce $100\gamma\%$ generalized confidence intervals for θ.

4.4 Show that the error sum of squares defined by (4.36) has the chi-squared distribution

$$\frac{S_E}{\sigma_\epsilon^2} \ \sim \ \chi^2_{(n-1)(k \ - \ 1)}.$$

4.5 The table below shows the highway mileage under average traffic conditions obtained by a random sample of four operators by driving a certain model of a midsize car on 15 different test drives. Assuming a one-way random effects model with the variance components σ_ϵ^2 and σ_α^2,

(a) construct the equal-tail 99% confidence interval for the error variance σ_ϵ^2,
(b) find the unbiased estimate of σ_α^2 based on the sufficient statistics.
(c) find the estimate of σ_α^2 based on its 50% confidence bound,
(d) find the equal-tail 99% generalized confidence interval for the factor variance σ_α^2,
(e) test the null hypothesis, $H_0 : \ \sigma_\alpha^2 \ \leq \ 2$,
(f) construct the equal-tail 95% confidence interval for the mean mileage of the midsize car.

Driver 1	Driver 2	Driver 3	Driver 4
27	22	23	23
25	24	24	25
28	24	19	21
26	28	27	23
22	21	19	18
29	20	24	24
25	21	28	22
27	19	18	20
23	21	22	25
28	20	24	24
28	28	24	20
26	18	18	17
28	25	26	25
29	28	23	24
23	30	19	29

4.6 Weerahandi (1995) reported the data shown in table below. They represent the sample means and the sample variances weights of 20 babies born in each of the eight hospitals.

	Hospital							
	H1	H2	H3	H4	H5	H6	H7	H8
$\overline{y}$	7.6	8.3	7.5	7.8	8.5	7.9	7.8	7.2
s_i^2	2.1	2.3	2.2	1.9	2.1	2.0	2.1	2.0

Assuming a one-way random effects model, establish the ANOVA table for making inferences about the variance components.

4.7 Consider again the data set in Exercise 3.6. Assuming a one-way random effects model,

(a) construct the equal-tail 95% confidence interval for the error variance σ_ϵ^2,
(b) construct the left-sided 95% generalized confidence interval for the factor variance σ_α^2,
(c) test the null hypothesis, $H_0 : \sigma_\alpha^2 \leq 1$,
(d) construct the equal-tail 95% confidence interval for the mean weight of babies.

4.8 Consider the data in Table 3.1. Assuming a one-way random effects model construct a 95% equal-tail generalized confidence interval for the mean power consumption.

4.9 The data given in table below are reported by Swallow and Searle (1978). They are the weights of a sample bottles from five groups of vegetable oil.

Group A:	15.75	15.82	15.75	15.71	15.84
Group B:	15.70	15.68	15.64	15.60	
Group C:	15.68	15.66	15.59		
Group D:	15.69	15.71			
Group E:	15.65	15.60			

Assuming a one-way random effects model,

(a) construct 95% confidence intervals for each variance component, (b) construct a 95% confidence interval for the mean weight, (c) test the hypothesis that the mean weight is at least 15.7.

4.10 Consider again the data set in Exercise 3.9. If the groups are not randomly selected from a population of vegetable oil bottles, discuss the underlying statistical problem and analyze the data.

4.11 The table below shows a hypothetical data set from an experiment designed to estimate the effect of temperature setting T on the power consumption of a certain brand of refrigerators. Assuming a two-way mixed effects model with the fixed effect T, the variance component σ_ϵ^2 due to unit-to-unit variation, and the variance component σ_α^2 due to the refrigerator load L,

(a) construct the equal-tail 95% confidence interval for the error variance σ_ϵ^2,
(b) find the unbiased estimate of σ_α^2 based on the sufficient statistics,
(c) find the estimate of σ_α^2 based on its 50% confidence bound,
(d) estimate the mean power consumption at each of the three temperature settings,
(e) test the hypothesis that the temperature setting has no significant effect.

Load	T_1	T_2	T_3
L_1	9.9	10.2	12.8
L_2	10.5	9.8	13.1
L_3	10.6	11.4	15.8
L_4	9.2	15.2	12.70
L_5	8.9	13.2	10.7
L_6	10.7	12.0	11.2
L_7	8.4	10.2	11.4

CHAPTER 5

HIGHER-WAY MIXED MODELS

5.1 INTRODUCTION

In many practical applications we need to deal with a number of factors of random effects and fixed effects. Hence we will have the need to make inferences about a number of variance components. As we will discuss below, in some applications we will also have to make inferences about, not only on the individual variance components, but also on functions of a number of variance components. The purpose of this chapter is to develop methodologies to tackle such problems in a general manner so that results would apply to a wide class of higher-way mixed models. To make this possible in this chapter we assume that we have observations from a balanced higher-way design.

As an example of a higher-way mixed model, consider again the problem of making inferences about the factors affecting the power consumption of a certain brand of refrigerators. To make inferences about the mean power consumption and factors affecting the power consumption, we could carry out an experiment under various conditions on the refrigerator load, internal temperature levels, and external temperature levels simulating ceratin household temperature conditions during various seasons. Although individual levels of

such conditions are important and usually not reported they do contribute to the variance and its structure. They constitute random effects representing, unit to unit variation, variations due to the two sets of temperatures, the variation due to the refrigerator load, and so on. To estimate the variance components and the mean power consumption, we could setup the experiment according to a certain design to enable estimation of each individual variance component or some of them. Table 4.1 shows a set of hypothetical data in a scaled unit that would allow us to estimate just three variance components, namely the variation due to internal temperature setting, the load, and the unit to unit variation (including the random unexplained variation). The levels of the refrigerator load is chosen at random from a set of typical loads. The particular values of the load are of no particular interest as they are not usually included in energy efficiency reports. In designing the experiment, the temperature settings can be chosen at some desires levels such as low, medium, and high (very cold) or at some randomly chosen levels depending on the need. If the levels are randomly selected, the data in Table 4.1 should be modeled as a two-way random effects model and otherwise they should be modeled as a two-way mixed effects model. If external temperature is also controlled and observations are taken when it is set at a number of levels, then the data from the experiment should be treated as a three-way random effects model or a three-way mixed model depending on the way the levels of the factors are chosen.

Table 5.1 Power consumption of refrigerators

	Temperature		
Load	T_1	T_2	T_3
L_1	11.9, 11.3, 9.90	12.6, 12.2, 12.4	12.5, 12.2, 11.4
L_2	11.7, 12.2, 10.9	12.7, 13.9, 13.9	11.6, 13.1, 12.5
L_3	11.8, 10.1, 11.1	12.9, 12.2, 12.9	12.8, 13.4, 12.8

5.2 CANONICAL FORM OF THE PROBLEM

In most balanced random effects models, the underlying inference problem on variance components can be reduced by means of a sum of squares decomposition into a problem having a common structure. Therefore methods of inference can be developed and presented in a form that is valid for any variance component in a higher-way balanced mixed model following the structure. A variance component σ_a^2 in a model is said to have the *canonical form of a variance component* if the sufficient statistics for making inferences about the

variance component have distributions of the form

$$V = \frac{S_a}{\sigma^2 + A\sigma_a^2} \sim \chi_a^2 \quad \text{and} \quad W = \frac{S_b}{\sigma^2} \sim \chi_b^2 , \tag{5.1}$$

where σ^2 is a nuisance parameter, typically a linear combination of some other variance components including or excluding the error variance, S_a and S_b are some sums of squares of deviations, A is a known constant, and a and b are the degrees of freedom of the two chi-squared distributions. Similarly, in a balanced mixed model, a factor of fixed effects with levels $B_1, B_2, \ldots, B_n$ having fixed effects $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)'$ is said to have the *canonical form of a fixed effect* if the sufficient statistics for making inferences about the variance component have distributions of the form,

$$U = \frac{S_c - B\sum_{j=1}^{n} \beta_j^2}{\sigma^2} \sim \chi_c^2 \quad \text{and} \quad W = \frac{S_b}{\sigma^2} \sim \chi_b^2 , \tag{5.2}$$

where σ^2 is a nuisance parameter, typically a linear combination of the variance components, S_c and S_b are some sums of squares of deviations, B is a known constant, and c and b are the degrees of freedom of the two chi-squared distributions. These canonical forms are further illustrated by the following applications.

5.2.1 Two-Way random effects model

In many industrial experiments, often we need to deal with designs involving two-way random effects. As a specific example, consider the data set in Table 4.2 reported by Montgomery (1991) and Montgomery and Runger (1993a) pertaining to an important class of applications in the assessment of measurement systems. In this class of applications, an experiment known as a gauge R & R study is performed. As applied to the case of ideal setting of this class, p parts from a population of parts made by a certain process are randomly chosen. Then, it involves choosing o operators at random from a population of operators and having each operator measure each part n times. In the data set presented in Table 4.2, there are $p = 20$ parts, $o = 3$ operators, and $n = 2$ measurements leading to 60 observations.

Let Y_{ijk} denote the observations from a two-way random effects model with two random factors A and B. Suppose in the design of the experiment A is set to take I random levels and B is set to take J random levels. Allowing interaction between the two factors as well, assume the linear model

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \\ i &= 1, \ldots, I; \;\; j = 1, \ldots, J; \;\; k = 1, \ldots, K, \end{aligned} \tag{5.3}$$

Table 5.2 Assessing measurement systems

	Operator					
	1		2		3	
	Meas. 1		**Meas. 2**		**Meas. 3**	
Part	1	2	1	2	1	2
1	21	20	20	20	19	21
2	24	23	24	24	23	24
3	20	21	19	21	20	22
4	27	27	28	26	27	28
5	19	18	19	18	18	21
6	23	21	24	21	23	22
7	22	21	22	24	22	20
8	19	17	18	20	19	18
9	24	23	25	23	24	24
10	25	23	26	25	24	25
11	21	20	20	20	21	20
12	18	19	17	19	18	19
13	23	25	25	25	25	25
14	24	24	23	25	24	25
15	29	30	30	28	31	30
16	26	26	25	26	25	27
17	20	20	19	20	20	20
18	19	21	19	19	21	23
19	25	26	25	24	25	25
20	19	19	18	17	19	17

where α_i is the effect due to the ith level of A, where β_j is the effect due to the jth level of B, and γ_{ij} is due to their interactions. Also assume that

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad \beta_j \sim N(0, \sigma_\beta^2),$$
$$\gamma_{ij} \sim N(0, \sigma_\gamma^2), \quad \text{and } \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2).$$

As in the two-way fixed effects model it is also assumed that, α_i, β_j, γ_{ij}, and ϵ_{ijk} are independently distributed.

An ANOVA for this model can also be established similar to the results in Chapter 3 by first decomposing the total sums of squares of deviations into orthogonal components as

$$S_T = S_\alpha + S_\beta + S_\gamma + S_e, \tag{5.4}$$

where

$$S_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y})^2, \tag{5.5}$$

$$S_\alpha = JK\sum_{i=1}^{I}(\overline{Y}_{i.} - \overline{Y})^2, \tag{5.6}$$

$$S_\beta = IK\sum_{j=1}^{J}(\overline{Y}_{.j} - \overline{Y})^2, \tag{5.7}$$

$$S_\gamma = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y})^2, \tag{5.8}$$

and

$$S_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \overline{Y}_{ij})^2. \tag{5.9}$$

Table 4.3 displays the ANOVA table based on these sums of squares, with which we can make inferences about all of the variances components. The sums of squares of deviations are denoted in the table by S with appropriate subscripts. The mean sums of squared deviations are denoted by MS, which are defined as the sum of squares divided by the associated degrees of freedom. More specifically, they are defined as

$$MS_\nu = S_\nu/DF_\nu.$$

Table 5.3 ANOVA for the two-way random effects model

Source	DF	SS	E(MS)
Factor A	$I-1$	S_α	$\theta_1 = JK\sigma_\alpha^2 + K\sigma_\gamma^2 + \sigma_\epsilon^2$
Factor B	$J-1$	S_β	$\theta_2 = IK\sigma_\beta^2 + K\sigma_\gamma^2 + \sigma_\epsilon^2$
A × B	$(I-1)(J-1)$	S_γ	$\theta_3 = K\sigma_\gamma^2 + \sigma_\epsilon^2$
Error	$IJ(K-1)$	S_e	$\theta_4 = \sigma_\epsilon^2$
Total	$IJK-1$	S_T	

It is easy to establish the expected values appearing in the ANOVA table from the distribution of sums of squares given by the normal theory (see Appendix A.6 more generally for the three-way mixed model) or by direct evaluation. For example, starting from the equations

$$\overline{Y}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \overline{\epsilon}_{ij},$$

$$\overline{Y}_{i.} = \mu + \alpha_i + \overline{\beta} + \overline{\gamma}_i + \overline{\epsilon}_i,\ \ \overline{Y}_{.j} = \mu + \overline{\alpha} + \beta_j + \overline{\gamma}_j + \overline{\epsilon}_j,$$

and

$$\overline{Y} = \mu + \overline{\alpha} + \overline{\beta} + \overline{\gamma} + \overline{\epsilon},$$

we can obtain the expected value of S_γ, and then that of MS_γ, as follows:

$$\begin{aligned}
E(S_\gamma) = \; & KE(\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij} \; - \; \overline{Y}_{i.} \; - \; \overline{Y}_{.j} \; + \; \overline{Y})^2), \\
= \; & KE(\sum_{i=1}^{I}\sum_{j=1}^{J}((\beta_j + \gamma_{ij} + \overline{\epsilon}_{ij} - \overline{\beta} - \overline{\gamma}_i - \overline{\epsilon}_i) \\
& \qquad\qquad - \; (\beta_j + \overline{\gamma}_j + \overline{\epsilon}_j - \overline{\beta} - \overline{\gamma} - \overline{\epsilon}))^2) \\
= \; & KE(\sum_{i=1}^{I}\sum_{j=1}^{J}((\gamma_{ij} - \overline{\gamma}_i \; - \; \overline{\gamma}_j + \overline{\gamma}) + (\overline{\epsilon}_{ij} - \overline{\epsilon}_i - \overline{\epsilon}_j + \overline{\epsilon}))^2) \\
= \; & K\sum_{i=1}^{I}\sum_{j=1}^{J}E(\gamma_{ij} - \overline{\gamma}_i \; - \; \overline{\gamma}_j + \overline{\gamma})^2 + E(\overline{\epsilon}_{ij} - \overline{\epsilon}_i - \overline{\epsilon}_j + \overline{\epsilon})^2 \\
= \; & K(I-1)(J-1)({\sigma^2}_\gamma + {\sigma^2}_\epsilon/K),
\end{aligned}$$

where the last equation follows from the known results from the classical two-way ANOVA for fixed effects. Note that the expected mean squares denoted by θ involve some or all the variance components ${\sigma^2}_\beta$, ${\sigma^2}_\alpha$, ${\sigma^2}_\gamma$, and ${\sigma^2}_\epsilon$ thus providing a basis for deriving estimates and significance tests for the variance components. As in Chapter 3, it is easy to show that (see Appendix A.5) their distributions are related to the chi-squared distributions as

$$S_\alpha/\theta_1 \sim \chi^2_{I-1}, \tag{5.10}$$

$$S_\beta/\theta_2 \sim \chi^2_{J-1}, \tag{5.11}$$

$$S_\gamma/\theta_3 \sim \chi^2_{(I-1)(J-1)}, \tag{5.12}$$

$$S_e/\theta_4 \sim \chi^2_{IJ(K-1))}. \tag{5.13}$$

The distributions can be derived directly form the normal theory or deduced from known results available for the conventional two-way ANOVA. For example, the distribution of S_γ follows from the identity

$$\overline{Y}_{ij} \; - \; \overline{Y}_{i.} \; - \; \overline{Y}_{.j} \; + \; \overline{Y} = \overline{e}_{ij} - \overline{e}_i - \overline{e}_j + \overline{e},$$

and from known results in the two-way ANOVA model without replications, where

$$e_{ij} = \gamma_{ij} + \overline{\epsilon}_{ij} \sim N(0, \sigma^2_\gamma + \frac{\sigma^2_\epsilon}{K}).$$

Now it is evident that each of the three variance components ${\sigma^2}_\alpha$, ${\sigma^2}_\beta$, and ${\sigma^2}_\gamma$ have the canonical form of a variance component. In making inferences

about the variance components, we set the quantities of the canonical form in (5.1) to the values shown in the following table.

Canonical forms for the variance components

Inference on $\sigma_a^2 =$	${\sigma^2}_\gamma$	${\sigma^2}_\alpha$	${\sigma^2}_\beta$
$\sigma^2 =$	${\sigma^2}_\epsilon$	$K{\sigma^2}_\gamma + {\sigma^2}_\epsilon$	$K{\sigma^2}_\gamma + {\sigma^2}_\epsilon$
$S_a =$	S_γ	S_α	S_β
$S_b =$	S_e	S_γ	S_γ
$A =$	K	JK	IK
$a =$	$(I-1)(J-1)$	$I-1$	$J-1$
$b =$	$IJ(K-1)$	$(I-1)(J-1)$	$(I-1)(J-1)$

5.2.2 Two-Way mixed effects model

Suppose the data in Table 4.1 is taken when the internal temperature is set at three desired levels, say T_1 representing the low level (cold), T_2 representing the medium level (colder), and T_3 representing the high level (coldest). Then the design constitutes a two-way mixed model. In general, suppose we have an experimental design with one random factor A taking on I random levels and a second factor B taking on J fixed-levels. Again we assume that we have obtained data from a balanced two-way cross-classified design with Y_{ijk}, $k = 1, \ldots, K$ denoting the observations taken from the ijth cell. Assume the linear model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{5.14}$$

with β_j now representing main effect due to jth level of factor B. Further, as before, assume that all the random effects are normally distributed with zero means. More specifically, assume that

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad \gamma_{ij} \sim N(0, \sigma_\gamma^2), \quad \text{and} \quad \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2),$$

Without loss of generality assume that the fixed effects are measured as deviations from the overall mean so that they satisfy the equation $\sum_{j=1}^{J} \beta_j = 0$.

It is easily verified that the above model also satisfy the orthogonal decomposition (5.4) of sums of squares of deviations with the same definition of the sums of squares. In this case, however, the ANOVA table needs to be changed as shown in Table 4.4 below.

For example, starting from the equations $\overline{Y}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \overline{\epsilon}_{ij}$ and $\overline{Y}_{.j} = \mu + \overline{\alpha} + \beta_j + \overline{\gamma}_j + \overline{\epsilon}_j$, the expected value of S_β, and then that of MS_β,

Table 5.4 ANOVA for the two-way mixed model

Source	DF	SS	E(MS)
Factor A	$I-1$	S_α	$\theta_1 = JK\sigma_\alpha^2 + \theta_3$
Factor B	$J-1$	S_β	$\theta_2 = IK\frac{\sum_{j=1}^{J}\beta_j^2}{J-1} + \theta_3$
A × B	$(I-1)(J-1)$	S_γ	$\theta_3 = K\sigma_\gamma^2 + \sigma_\epsilon^2$
Error	$IJ(K-1)$	S_e	$\theta_4 = \sigma_\epsilon^2$
Total	$IJK-1$	S_T	

could be obtained using the result

$$\begin{aligned} S_\beta &= IK\sum_{j=1}^{J}(\overline{Y}_{.j} - \overline{Y})^2 \\ &= IK\sum_{j=1}^{J}[\beta_j + (\overline{\gamma}_j - \overline{\gamma}) + (\overline{\epsilon}_j - \overline{\epsilon})]^2. \end{aligned}$$

Moreover, the orthogonal decomposition of the sums of squares of deviations also lead to the distributional results

$$\frac{S_\alpha}{\theta_1} \sim \chi^2_{I-1}, \tag{5.15}$$

$$\frac{S_\beta - IK\sum_{j=1}^{J}\beta_j^2}{\theta_3} \sim \chi^2_{J-1}, \tag{5.16}$$

$$\frac{S_\gamma}{\theta_3} \sim \chi^2_{(I-1)(J-1)}, \tag{5.17}$$

$$\frac{S_e}{\theta_4} \sim \chi^2_{IJ(K-1))}. \tag{5.18}$$

Now it is evident that each of the three variance components ${\sigma^2}_\alpha$, ${\sigma^2}_\beta$, and ${\sigma^2}_\gamma$ have the canonical form of a variance component. In making inferences about the variance components, we set the quantities of the canonical form in (5.1) to the values shown in the following table.

Canonical form for the variance components

Inference on $\sigma_a^2 =$	${\sigma^2}_\gamma$	${\sigma^2}_\alpha$
$\sigma^2 =$	${\sigma^2}_\epsilon$	$K{\sigma^2}_\gamma + {\sigma^2}_\epsilon$
$S_a =$	S_γ	S_α
$S_b =$	S_e	S_γ
$A =$	K	JK
$a =$	$(I-1)(J-1)$	$I-1$
$b =$	$IJ(K-1)$	$(I-1)(J-1)$

Similarly, in making inferences about the fixed effects, we set the quantities of the canonical form in (5.2) to the values shown in the following table.

Canonical form for the fixed effects.

Inference on:	β
$\sigma^2 =$	$K{\sigma^2}_\gamma + {\sigma^2}_\epsilon$
$\sum_{j=1}^{J} \beta_j^2 =$	$\sum_{j=1}^{J} \beta_j^2$
$S_c =$	S_β
$S_b =$	S_γ
$B =$	IK
$c =$	$J-1$
$b =$	$(I-1)(J-1)$

Example 4.1. Analysis of energy consumption

Consider the data set reported in Table 4.1. The load levels are randomly chosen and so they leads to random effects. Depending on the way the temperature levels are chosen they are fixed effects or random effects. Then, the model appropriate for analyzing the data is either a two-way random effects model or a two-way mixed model, respectively. In either case the analysis can be based on the ANOVA table shown below; the last column of the table represents the usual mean sums of squares defined as

$$MS = \frac{SS}{DF}.$$

Source	DF	SS	MS
Load	2	2.090	1.045
Temperature	2	13.354	6.677
Load × Temperature	4	1.997	0.499
Error	18	7.820	0.434
Total	26	25.261	

5.2.3 Three-way mixed effects model

Above results can be easily extended to higher-way random effects models and mixed effects models. For example, in the experiment on the power consumption of refrigerators the design becomes a three way mixed model if we simulate room temperature as well. This is also the case if power consumption is measured by a number of operators using the same or a number of units of the same brand of a measuring device. Table 4.5 shows the nature of data from such experiments. In this type of application, while factors such as the internal temperature setting should be treated as fixed effects, the load and the operator effects should be treated as random effects. In conducting the experiment, the levels of the factors should be selected accordingly.

Table 5.5 Energy consumption by three factors

		Temperature		
Load	Operator	T_1	T_2	T_3
L_1	O_1	11.8, 11.2, 10.9	12.4, 12.1, 12.6	12.4, 12.2, 11.6
L_1	O_2	10.9, 10.8, 9.80	11.6, 11.3, 12.1	12.1, 12.5, 12.4
L_2	O_1	10.7, 11.3, 10.8	12.5, 13.8, 12.9	12.6, 12.4, 12.7
L_2	O_2	11.8, 11.2, 9.90	11.7, 13.1, 12.9	11.8, 12.1, 12.1
L_3	O_1	11.9, 11.1, 10.3	11.9, 12.4, 12.6	12.4, 13.8, 13.1
L_3	O_2	11.4, 10.5, 10.1	12.1, 13.2, 12.4	13.2 13.1, 12.9

Let A, B, and C be the three factors of a three-way cross classified design. First suppose we have more than one observation, say L observations, from each combination of factor levels. Let Y_{ijkl} be lth observation from the factor level combination (A_i, B_j, C_k). Suppose in the design of the experiment A is set to take I random levels and B is set to take J fixed-levels and C is set to take K random levels. For example, in Table 4.5, $A = L$, $B = T$, $C = O$, $I = 3$, $J = 3$, $K = 2$, and $L = 3$. Allowing all possible interactions, assume the linear model

$$\begin{aligned} Y_{ijkl} = \mu + (\alpha_i + \beta_j + \theta_k) + (\alpha\beta_{ij} + \alpha\theta_{ik} + \beta\theta_{jk}) \\ + \alpha\beta\theta_{ijk} + \epsilon_{ijkl}, \\ i = 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K, \, l = 1, \ldots, L, \end{aligned} \tag{5.19}$$

where α_i is the effect due to the ith random level of A, β_j is the fixed effect due to the jth level of B, and θ_k is the random effect due to kth level of C. The terms such as $\alpha\beta_{ij}$ represent the two way intersections and $\alpha\beta\theta$ denotes the interaction between all three factors. As in the two-way mixed effects model, assume that the fixed effects are measured as deviations from the overall mean so that they satisfy the equation $\sum_{j=1}^{J} \beta_j = 0$.

It should be emphasized that, unless there are a large number of replicates, it is a not necessarily a good idea to allow too many interaction terms. For instance, too many interaction terms could lead to loss of power in making inferences about the main effects. When it is reasonable to assume that there is no significant interaction between certain factor levels, they can be dropped from the model. For example in the above example, on one hand there is no reason why there should be any interaction between the operators and loads, and on the other hand we expect that the effect of the temperature on the power consumption to depend on the load. In that case the appropriate model could be expressed as

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \theta_k + \alpha\beta_{ij} + \epsilon_{ijkl}. \tag{5.20}$$

In applications with no replicates also−that is, when $L = 1-$ we need to work with model (5.20) and replace Y_{ijkl} by Y_{ijk}.

In any case we can proceed with the general model (5.19) and develop the ANOVA. As it will become clear later, we will be able to obtain tests appropriate for model (5.20) by pooling sums of squares and degrees of freedoms available from the general ANOVA. Proceeding with the general model, assume that

$$\begin{aligned} \alpha_i &\sim N(0, \sigma^2_\alpha), \\ \theta_k &\sim N(0, \sigma^2_\theta), \end{aligned} \tag{5.21}$$

$$\begin{aligned} \alpha\beta_{ij} &\sim N(0, \sigma^2_{\alpha\beta}), \\ \alpha\theta_{ik} &\sim N(0, \sigma^2_{\alpha\theta}), \\ \beta\theta_{jk} &\sim N(0, \sigma^2_{\beta\theta}), \end{aligned}$$

$$\alpha\beta\theta_{ijk} \sim N(0, \sigma^2_{\alpha\beta\theta}),$$

and

$$\epsilon_{ijkl} \sim N(0, \sigma^2_\epsilon),$$

which define the underlying variance components as well. Continuing with the usual terminology, let us denote various sample means of the data by $\overline{Y}_{i..}, \overline{Y}_{.j.}, \overline{Y}_{..k}, \overline{Y}_{ij.}, \overline{Y}_{.jk}, \overline{Y}_{i.k}$, and $\overline{Y}_{ijk}$, depending on the indices with respect to which the average is taken. For example, $\overline{Y}_{i..}$is defined as

$$\overline{Y}_{i..} = \frac{\sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{l=1}^{L} Y_{ijkl}}{JKL}$$

and $\overline{Y}_{.jk}$ is defined as

$$\overline{Y}_{.jk} = \frac{\sum_{i=1}^{I} \sum_{l=1}^{L} Y_{ijkl}}{IL}.$$

Let us continue to denote the grand average of all the data by $\overline{Y}$. To present the ANOVA table for the three-way mixed model, consider as usual the decomposition of the total sums of squares of deviations

$$S_T = (S_\alpha + S_\beta + S_\theta) + (S_{\alpha\beta} + S_{\alpha\theta} + S_{\beta\theta}) + S_{\alpha\beta\theta} + S_e \tag{5.22}$$

suggested by the model (5.19) and the identity

$$\begin{aligned}(Y_{ijkl} - \overline{Y}) &= (\overline{Y}_{i..} - \overline{Y}) + (\overline{Y}_{.j.} - \overline{Y}) + (\overline{Y}_{..k} - \overline{Y}) \\ &+ (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}) + (\overline{Y}_{i.k} - \overline{Y}_{i..} - \overline{Y}_{..k} + \overline{Y}) \\ &+ (\overline{Y}_{.jk} - \overline{Y}_{,j.} - \overline{Y}_{..k} + \overline{Y}) + (\overline{Y}_{ijk} - (\overline{Y}_{ij.} + \overline{Y}_{i.k} + \overline{Y}_{.jk}) \\ &+ (\overline{Y}_{i..} + \overline{Y}_{..k} + \overline{Y}_{.j.}) - \overline{Y}) + (Y_{ijkl} - \overline{Y}_{ijk}),\end{aligned}$$

where

$$S_T = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L}(Y_{ijkl} - \overline{Y})^2,$$

$$\begin{aligned}S_\alpha &= JKL\sum_{i=1}^{I}(\overline{Y}_{i..} - \overline{Y})^2, \\ S_\beta &= IKL\sum_{j=1}^{J}(\overline{Y}_{.j.} - \overline{Y})^2, \\ S_\theta &= IJL\sum_{k=1}^{K}(\overline{Y}_{..k} - \overline{Y})^2,\end{aligned}$$

$$\begin{aligned}S_{\alpha\beta} &= KL\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y})^2, \\ S_{\alpha\theta} &= JL\sum_{i=1}^{I}\sum_{k=1}^{K}(\overline{Y}_{i.k} - \overline{Y}_{i..} - \overline{Y}_{..k} + \overline{Y})^2, \\ S_{\beta\theta} &= IL\sum_{j=1}^{J}\sum_{k=1}^{K}(\overline{Y}_{.jk} - \overline{Y}_{.j.} - \overline{Y}_{..k} + \overline{Y})^2,\end{aligned}$$

$$\begin{aligned}S_{\alpha\beta\theta} = L\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\{&\overline{Y}_{ijk} - (\overline{Y}_{ij.} + \overline{Y}_{i.k} + \overline{Y}_{.jk}) \\ &+ (\overline{Y}_{i..} + \overline{Y}_{..k} + \overline{Y}_{.j.}) - \overline{Y}\}^2,\end{aligned}$$

and

$$S_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{l=1}^{L} (Y_{ijkl} - \overline{Y}_{ijk})^2.$$

Table 4.6 displays the ANOVA table based on these sums of squares, with which we can make inferences about all of the variances components.

Table 5.6 ANOVA for the three-way mixed model

Source	DF	SS	E(MS)
Factor A	$I-1$	S_α	$\phi_3 + JKL\sigma^2_\alpha$
Factor B	$J-1$	S_β	$\phi_2 + IKL\frac{\sum_{j=1}^{J}\beta_j^2}{J-1}$
Factor C	$K-1$	S_θ	$\phi_1 + IJL\sigma^2_\theta$
A $\times$ B	$(I-1)(J-1)$	$S_{\alpha\beta}$	$\phi_1 = \phi_4 + KL\sigma^2_{\alpha\beta}$
A $\times$ C	$(I-1)(K-1)$	$S_{\alpha\theta}$	$\phi_2 = \phi_4 + JL\sigma^2_{\alpha\theta}$
B $\times$ C	$(J-1)(K-1)$	$S_{\beta\theta}$	$\phi_3 = \phi_4 + IL\sigma^2_{\beta\theta}$
A $\times$ B $\times$C	$(I-1)(J-1)(K-1)$	$S_{\alpha\beta\theta}$	$\phi_4 = \sigma^2_\epsilon + L\sigma^2_{\alpha\beta\theta}$
Error	$IJK(L-1)$	S_e	σ^2_ϵ
Total	$IJKL-1$	S_T	

Moreover, as suggested by the ANOVA table, each sum of squares is related to a chi-squared distribution:

$$\frac{S_\alpha}{\phi_3 + JKL\sigma^2_\alpha} \sim \chi^2_{I-1},$$

$$\frac{S_\beta - IKL\sum_{j=1}^{J}\beta_j^2}{\phi_2} \sim \chi^2_{J-1},$$

$$\frac{S_\theta}{\phi_1 + IJL\sigma^2_\theta} \sim \chi^2_{K-1},$$

$$\frac{S_{\alpha\beta}}{\phi_1} \sim \chi^2_{(I-1)(J-1)},$$

$$\frac{S_{\alpha\theta}}{\phi_2} \sim \chi^2_{(I-1)(K-1)},$$

$$\frac{S_{\beta\theta}}{\phi_3} \sim \chi^2_{(J-1)(K-1)},$$

$$\frac{S_{\alpha\beta\theta}}{\phi_4} \sim \chi^2_{(I-1)(J-1)(K-1)},$$

$$\frac{S_e}{\sigma^2_\epsilon} \sim \chi^2_{IJK(L-1)}.$$

The derivations of the results are similar to the two-way mixed model. For example, the expected value and the distribution of S_α is easily seen from the identity

$$(\overline{Y}_{i..} - \overline{Y}) = (\alpha_i - \overline{\alpha}) + (\overline{\alpha\beta}_i - \overline{\alpha\beta}) + (\overline{\alpha\beta\theta}_i - \overline{\alpha\beta\theta}) + (\overline{\epsilon}_i - \overline{\epsilon}).$$

Details of the derivation is given in Appendix A.6. The sums of squares are independently distributed due to the orthogonal decomposition (5.22).

Now it is evident that all the variance components of the model have the canonical form. To present them in a compact manner, define $I_1 = I-1$, $J_1 = J-1$, $K_1 = K-1$, and $L_1 = L-1$. Then, in making inferences about the variance components of a three way mixed model, various quantities appearing in the canonical form in (5.1) are set to the values shown in the following table..

Canonical form for the variance components

Inference on $\sigma_a^2 =$	$\sigma^2{}_\alpha$	σ^2_θ	$\sigma^2_{\alpha\beta}$	$\sigma^2_{\alpha\theta}$	$\sigma^2_{\beta\theta}$	$\sigma^2_{\alpha\beta\theta}$
$\sigma^2 =$	ϕ_3	ϕ_1	ϕ_4	ϕ_4	ϕ_4	$\sigma^2{}_\epsilon$
$S_a =$	S_α	S_θ	$S_{\alpha\beta}$	$S_{\alpha\theta}$	$S_{\beta\theta}$	$S_{\alpha\beta\theta}$
$S_b =$	$S_{\beta\theta}$	$S_{\alpha\beta}$	$S_{\alpha\beta\theta}$	$S_{\alpha\beta\theta}$	$S_{\alpha\beta\theta}$	S_e
$A =$	JKL	IJL	KL	JL	IL	L
$a =$	I_1	K_1	I_1J_1	I_1K_1	J_1K_1	$I_1J_1K_1$
$b =$	J_1K_1	I_1J_1	$I_1J_1K_1$	$I_1J_1K_1$	$I_1J_1K_1$	$IJKL_1$

Similarly, in making inferences about the fixed effects, we set the quantities of the canonical form in (5.2) to the values shown in the following table.

Canonical form for the fixed effects

Inference on:	β
$\sigma^2 =$	$\phi_2 = \sigma^2_\epsilon + L\sigma^2_{\alpha\beta\theta} + JL\sigma^2_{\alpha\theta}$
$\sum_{j=1}^{J} \beta_j^2 =$	$\sum_{j=1}^{J} \beta_j^2$
$S_c =$	S_β
$S_b =$	$S_{\alpha\theta}$
$B =$	IKL
$c =$	$J-1$
$b =$	$(I-1)(K-1)$

5.3 TESTING FIXED EFFECTS

Consider the problem of testing the difference in fixed effects of a balanced mixed model. Suppose the fixed effects of interest have the canonical form

$$U = \frac{S_c - B\sum_{j=1}^{n} \beta_j^2}{\sigma^2} \sim \chi_c^2 \quad \text{and} \quad W = \frac{S_b}{\sigma^2} \sim \chi_b^2 ,$$

where σ^2 is a nuisance parameter, S_c and S_b are certain sums of squares of deviations, and B is a known constant. Consider the problem of testing the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

against the natural alternative hypothesis. Under the null hypothesis, the distribution of the random variable U becomes

$$U \;=\; \frac{S_c}{\sigma^2} \sim\; \chi^2_c$$

and hence

$$F = \frac{U/c}{W/b} = \frac{S_c/c}{S_b/b} \sim F_{c,b.}$$

Moreover, if the null hypothesis is not true, then F has a noncentral F distribution. It is now evident that the p-value appropriate for testing H_0 is

$$p = \;\; 1 \;-\; H_{c,b}\left(\frac{s_c/c}{s_b/b}\right), \tag{5.23}$$

where $H_{c,b}$ is the cdf of the F distribution with c and b degrees of freedom.

Example 4.2. Analysis of energy consumption (continued)

Consider again the data set reported in Table 4.1 and assume that the temperature levels had been set at some desired levels so that the model appropriate for analyzing the data is a two-way mixed model. The hypothesis of equal temperature effects can be tested based on the ANOVA given in Example 4.1. The p-value for testing the hypothesis computed using (5.23) is then

$$\begin{aligned} p &= 1 - H_{2,4}\left(\frac{6.677}{0.499}\right) \\ &= 1 - H_{2,4}\,(13.38) = 0.0169. \end{aligned}$$

Hence, we can conclude that different temperature settings have different effects on the power consumption−colder the temperature setting, greater the power consumption tend to be.

5.4 ESTIMATING VARIANCE COMPONENTS

Consider the problem of estimating variance component σ^2_a and the variance function σ^2 when it has the canonical form

$$V \;=\; \frac{S_a}{\sigma^2 \;+\; A\sigma^2_a} \sim\; \chi^2_a \tag{5.24}$$

and

$$W \;= \frac{S_b}{\sigma^2} \sim\; \chi^2_b, \tag{5.25}$$

where S_a and S_b are certain sums of squares, A is a known constant. It is straightforward to make any type of inference about the variance σ^2 based on (5.25). For example, $100\gamma\%$ equal-tail confidence bound for σ^2 is obtained from the probability statements

$$\Pr(\frac{S_b}{\sigma^2} \leq c_{\frac{1-\gamma}{2}}) = \frac{1-\gamma}{2} \quad \text{and} \quad \Pr(\frac{S_b}{\sigma^2} \leq c_{\frac{1+\gamma}{2}}) = \frac{1+\gamma}{2},$$

where c_α is the αth quantile of the chi-squared distribution with b degrees of freedom. Obviously the resulting confidence interval is

$$\frac{S_b}{c_{\frac{1+\gamma}{2}}} \leq \sigma^2 \leq \frac{S_b}{c_{\frac{1-\gamma}{2}}}. \tag{5.26}$$

Moreover, the expression (5.25) implies that

$$E\,(S_b) = b\sigma^2,$$

and hence

$$\widehat{\sigma}^2 = MS_b = \frac{S_b}{b} \tag{5.27}$$

is the natural unbiased estimate of σ^2. Obviously it is also the MLE of σ^2. The unbiased estimate of σ_a^2 can be obtained from (5.27) and from

$$E\left(\frac{S_a}{a}\right) = \sigma^2 \ + \ A\sigma_a^2, \tag{5.28}$$

which implies that $MS_a = S_a/a$ is an unbiased estimate for the parameter $\sigma^2 \ + \ A\sigma_a^2$. Equations (5.27) and (5.28) imply that

$$\widehat{\sigma}_a^2 \ = \frac{MS_a \ - \ MS_b}{A} \tag{5.29}$$

is an unbiased estimate of σ_a^2. A major drawback of this unbiased estimate is that it could become negative with some data sets. So, the estimate cannot be considered reliable even when it is slightly positive. Since σ_a^2 is supposed to be a nonnegative parameter, one can of course simply set $\widehat{\sigma}_\alpha^2 = \ 0$ in (5.29) whenever $MS_a \ - \ MS_b < 0$, but this would make the estimate no longer unbiased. The unbiased estimate is also closely related to the RMLE (restricted maximum likelihood estimate) and the MLE [cf. Searle, Casella, and McCulloch (1992)] of σ_a^2. Despite these properties, all these estimates suffer from serious drawbacks. Neither the MLE nor the RMLE addresses the true underlying problem which causes negative or zero estimates. In fact, as we discussed in Section 3.4., compared with the unbiased estimate, MLE based inferences actually have more serious problems in other types of inferences such as testing of hypotheses and in interval estimation. Later in this chapter we will address the problem of interval estimation of σ_a^2, which will give a clearer picture of the situation.

Example 4.3. Analysis of Energy Consumption (continued). Consider again the data set reported in Table 4.1 and assume that both the factors, the load levels and temperature levels, are chosen randomly so that the model appropriate for analyzing the data is a two-way random effects model. For both variance components we have $A = 9$, and the mean sums of squares were presented in Example 4.1. The unbiased estimate of the variance component of the load can be computed from the ANOVA table given in Example 4.1:

$$\begin{aligned} \widehat{\sigma}_L^2 &= (1.045 - 0.499)/9 \\ &= 0.061. \end{aligned}$$

Similarly, the unbiased estimate of the variance component of the temperature is computed as

$$\begin{aligned} \widehat{\sigma}_T^2 &= (6.677 - 0.499)/9 \\ &= 0.686. \end{aligned}$$

5.5 TESTING VARIANCE COMPONENTS

Especially in dealing with variance components of random effects models and mixed models it is important to make inferences beyond the point estimation. To outline the testing procedures, consider again the canonical form of the model. The problem is not a trivial one, except for the case of testing zero variance component, namely the case of testing $H_0 : \sigma_a^2 = 0$, in which the $p-$value is easily obtained from (5.24) as

$$p = 1 - H_{a,b}\left(\frac{s_a/a}{s_b/b}\right). \tag{5.30}$$

Of more interest and importance are one-sided null hypotheses of the form

$$H_0 : \sigma_a^2 \geq \sigma_0^2. \tag{5.31}$$

Exact tests for (5.31) based on a single test statistic do not exist. Exact tests such as the one proposed by Healy (1961) involve an artificial randomization device in addition to the experimental data. However, exact tests based on an extreme region in the sample space formed by two test statistics, namely both the statistics, S_a and S_b do exist. To see this, consider the subset of the sample space defined by

$$C = \{(S_a, S_b) \mid \frac{S_a(\sigma^2 \frac{s_b}{S_b} + A\sigma_a^2)}{(\sigma^2 + A\sigma_a^2)} \leq s_a\}. \tag{5.32}$$

Note that the observed sample point (s_a, s_b) falls on the boundary of the sample space. To see that this is an extreme region and that its probability

does not depend on nuisance parameters, we can express the subset in terms of the chi-squared random variables V and W as

$$C \;=\; \{\, V(\frac{s_b}{W} \;+\; A\sigma_a^2) \;\leq\; s_a\}.$$

Obviously, the probability of C increases for deviations from the null hypothesis and its probability does not depend on σ^2, the nuisance parameter in the current problem. The generalized test variable underlying this extreme region is

$$T \;=\; V(s_b/W \;+\; A\sigma_a^2). \tag{5.33}$$

The generalized p-value is the maximum probability of the extreme region under the null hypothesis. It is computed as

$$\begin{aligned} p = \;& Pr(C|\; \sigma_\alpha^2 \;=\; \sigma_0^2) \\ &= Pr(V \leq\; s_a/(\frac{s_b}{W} \;+\; A\sigma_0^2)\;) \\ &= E\left(G_a(\frac{s_a}{A\sigma_0^2 \;+\; s_b/W})\right), \end{aligned} \tag{5.34}$$

where G_a is the cdf of the chi-squared distribution with a degrees of freedom and the expectation is taken with respect to the chi-squared random variable $W \;\sim\; \chi_b^2$. The uniqueness (up to equivalent p-values) of this test can be established by invoking the principle of invariance. A formal derivation of the result could be deduced from results in Weerahandi (1991). The expectation appearing in (5.34) can also be expressed as an integral over a subset of the interval [0, 1] with respect to a beta random variable. To do this, define the random variables

$$S = V + W \;\sim\; \chi_{a+b}^2 \text{ and } \;\; B = \; W/(V \;+\; W) \;\sim\; \text{Beta}\left(\frac{b}{2}, \frac{a}{2}\right), \tag{5.35}$$

which are also independently distributed. Then, the p-value can be expressed in terms of a well behaved integral as

$$\begin{aligned} p &= \Pr(V(\frac{s_b}{W} \;+\; A\sigma_0^2) \leq\; s_a) \\ &= \Pr\left(S(1-B)(\frac{s_b}{SB} + A\sigma_0^2) \leq s_a\;\right) \\ &= \Pr(\frac{s_b}{B} + SA\sigma_0^2 \leq \frac{s_a}{1-B}\;) \\ &= \int\limits_{\frac{s_b}{s_a+\;s_b}}^{1} G_{a+b}\left(\frac{1}{A\sigma_0^2}(s_a/(1-B) \;-\; s_b/B)\right) f_B(B)\, dB, \end{aligned} \tag{5.36}$$

where G_{a+b} is the cdf of the chi-squared distribution with $a + b$ degrees of freedom and $f_B(B)$ is the density of function of the beta random variable B.

Example 4.4. Analysis of energy consumption (continued)

Consider the problem of testing the variance due to the refrigerator load using the data in Table 4.1. The hypothesis of no variation in power consumption due to the load can also be tested based on the ANOVA given in Example 4.1. The p-value for testing zero variance is

$$\begin{aligned} p &= 1 - H_{2,4}\left(\frac{1.045}{0.499}\right) \\ &= 1 - H_{2,4}\,(2.09) \\ &= 0.239. \end{aligned}$$

This $p-$value does not support rejection of the hypothesis. To illustrate formula (5.34), also consider the problem of testing the hypothesis

$$H_0 :\ \sigma_L^2\ \geq 2\ .5,$$

where σ_L^2 is the variance due to the load. The generalized p-value for testing this hypothesis is computed as

$$\begin{aligned} p &= E\left(G_2(\frac{2.099}{9 \times 2.5\ +\ 1.997/W})\right) \\ &= 0.0437, \end{aligned}$$

where the expectation is taken with respect to $W\ \sim\ \chi_4^2$. The p-value suggests that the null hypothesis can be rejected at the 0.05 level. It can also be seen that there is not quite sufficient evidence to reject the null hypothesis $H_0 :\ \sigma_L^2\ \geq 1$ since its generalized p-value is 0.101.

5.6 CONFIDENCE INTERVALS

Since the unbiased estimate and the MLE of a variance component is not necessarily positive, it is important to provide various interval estimates for the variance component so that one can get clearer picture of the situation. As Weerahandi (1995) argued, when the point estimate is too small or negative, it is more informative to report confidence intervals rather than point estimates. If it is necessary to provide a point estimate, the 50% lower confidence bound, which could be treated as a median unbiased estimate, could be reported as a point estimate.

Consider the problem of constructing confidence intervals for a variance component σ_a^2 following the canonical form given by (5.1). The classical

approach does not provide confidence intervals based on exact probability statements and so there are many articles including those of Satterthwaite (1946) and Samaranayake and Bain (1988) providing approximate confidence intervals for such variance components. There is no such inference problem in generalized inference. We can derive generalized confidence intervals using a generalized pivotal quantity or they can be deduced from generalized p-values given in the previous section. In fact, the generalized test variable T define by (5.33) itself could be used to obtain a generalized pivotal. A more convenient generalized pivotal that reduces to the variance component at the observed values of the statistics is

$$\begin{aligned} R &= \frac{1}{A}\{(\sigma^2 \ + \ A\sigma_a^2)\frac{s_a}{S_a} - \sigma^2\frac{s_b}{S_B}\} \\ &= \frac{1}{A}\{\frac{s_a}{V} - \frac{s_b}{W}\}. \end{aligned}$$

Hence, the generalized confidence intervals for σ_a^2 can be obtained by writing various probability statements about R or they can be deduced from the generalized p-value given by (5.34). By the former approach a $100\gamma\%$ generalized lower confidence bound σ_0^2 for σ_a^2 is found such that

$$\begin{aligned} \Pr(R \geq \sigma_0^2) &= \Pr(\frac{s_a}{V} - \frac{s_b}{W} \geq A\sigma_0^2) \\ &= \Pr(V \leq \frac{s_a}{A\sigma_0^2 \ + \ s_b/W}) \\ &= E\left(G_a(\frac{s_a}{A\sigma_0^2 \ + \ s_b/W})\right) = \gamma. \end{aligned} \tag{5.37}$$

Similarly, if σ_1^2 and σ_2^2 are chosen such that

$$E\left(G_a(\frac{s_a}{A\sigma_1^2 \ + \ s_b/W})\right) = \frac{1+\gamma}{2} \tag{5.38}$$

and

$$E\left(G_a(\frac{s_a}{A\sigma_2^2 \ + \ s_b/W})\right) = \frac{1-\gamma}{2}, \tag{5.39}$$

then $[\sigma_1^2, \ \sigma_2^2]$ is an equal-tail $100\gamma\%$ generalized confidence interval for σ^2. The generalized interval can be computed conveniently using the XPro software package. For certain values of s_a and s_b, one may not be able to find a positive confidence bound to satisfy equation (5.38). In such situations, other confidence statements with low coefficients and asymmetric confidence intervals are considered more informative. By invoking the method of conditional inference, Weerahandi (1995) described how to find positive confidence limits in all situations.

Example 4.5. Analysis of energy consumption (continued)

Consider again the problem of making inferences about the variance due to

the refrigerator load with the data set in Table 4.1. Recall that with the data in Table 4.1 we had $s_a = 2.099$, $s_b = 1.997$, and $A = 9$. The upper 95% generalized confidence bound for σ_L^2 is 2.162. It can also be seen that the 95% lower confidence bound become negative suggesting that an exact confidence statement with a large coefficient is not possible with the observed data unless the probability statement is conditional. A 50% generalized confidence interval based on an exact probability statement is [0.0034, 0.3165]. Moreover, the 50% confidence bound used as a point estimate for the variance component is 0.088. These statements provide us with a better sense of the magnitude of the variance component as opposed to reporting the unbiased estimate of 0.061.

5.7 FUNCTIONS OF VARIANCE COMPONENTS

The purpose of this section is to present a class of applications of variance components, in which the generalized approach to solving underlying inference problems is very convenient and appealing. Here the problem is to make inferences about various sums and ratios of variance components. Although various other functions of variance components could also be tackled by the generalized approach, we confine our attention to the sums and ratios since they arise most often in practical applications. The reader is referred to Burdick and Graybill (1992) for various such functions and for approximate solutions for the problem of making inferences about such functions. Here we do not consider problems involving functions of both means and variance components of mixed models as arised, for instance, in applications of bioequivelence testing. For generalized inference in such applications the reader is referred to Peterson (2000) and McNally, Ijer, and Mathew (2003).

An important class of applications in this context arise in measurement systems. Of particular interest is the so-called gauge repeatability and reproducibility R & R studies. The reader is referred to Montgomery and Runger (1993a, 1993b) for a discussion of gauge R & R studies dealing with designed experiments. Tian and Cappelleri (2004) discuss another class of applications involving the problem of testing the reliability of expert evaluations and judgements.

The data models used for gauge R & R studies are mixed models with certain number of variance components. These variance components, individually, as well as sums, as ratios, and as ratios of sums, characterize the quality of a measurement system. Montgomery and Runger (1993a, 1993b) and Burdick (1994) stress the practical importance of using confidence intervals in this context since the point estimates by themselves can be misleading. Burdick and Larsen (1997) demonstrated the problems with the ANOVA-based confidence intervals. In particular, as they pointed out, the actual coverage of quantities of interest by such intervals in repeated sampling can be well below

the intended level. They also provide good alternative solutions based on intricate distributional approximations. The size performance of the confidence intervals they proposed and the generalized intervals provided by Hamada and Weerahandi (2000) are substantially better than ML/REML-based methods, which suffer from very serious size problems even in applications involving a single variance component, as seen in Chapter 3.

A major drawback of classical methods of finding good confidence intervals for functions of variance components, including those reported by Burdick and Graybill (1992), is that there is no single approach to deriving solutions and there is no general formula that can yield solutions to a wide class of problems. For example, different sums of variance components in a single model require different methods to obtain such approximate solutions and they cannot be deduced from a single formula. As we will see later in this chapter, the generalized approach provide general solutions to classes of problems in this context. Therefore, even a practitioner who is looking for a reasonable approximate solution with no regard to whether or not the solution is based on exact probability statements can benefit from the generalized procedures provided in this section. The main objective of this section is to provide a general approach to obtaining generalized tests and confidence intervals for sums of variance components and ratios of variance components, regardless of the composition of such functions.

5.7.1 Variance functions in the two-way model

To better describe the nature of the underlying problem, applications, and the nature of the solutions provided by the generalized approach, first consider the two-way random effects model that we studied above:

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \\ i &= 1, \ldots, I; \quad j = 1, \ldots, J; \quad k = 1, \ldots, K \end{aligned} \tag{5.40}$$

where $\alpha_i \sim N(0, {\sigma^2}_\alpha)$, $\beta_j \sim N(0, {\sigma^2}_\beta)$, $\gamma_{ij} \sim N(0, {\sigma^2}_\gamma)$, and $\epsilon_{ijk} \sim N(0, {\sigma^2}_\epsilon)$, being the random effects of the model, are all random variables. As usual in the case of the basic two-way layouts, α_i, β_j, γ_{ij}, and ϵ_{ijk} are assumed to be independently distributed.

In the previous sections, we addressed the problem of making inferences on individual variance components such as confidence intervals on the main effect ${\sigma^2}_\alpha$. The purpose of this section is to provide solutions to the problem of making inferences about sums and ratios of the variance components such as ${\sigma^2}_\beta + {\sigma^2}_\beta$ and ${\sigma^2}_\beta/({\sigma^2}_a + {\sigma^2}_\beta + {\sigma^2}_\gamma + {\sigma^2}_\epsilon)$. Before we undertake this task, let us consider a class of applications of special interest in which one needs to deal with various sums and ratios of variance components. Montgomery (1991) and Montgomery and Runger (1993a) described a class of important and interesting applications in the area of assessing measurement systems. In

this class of applications, an experiment known as a gauge R & R study is performed. As applied to the case of ideal setting of this class, p parts from a population of parts made by a certain process are randomly chosen. Then, it involves choosing o operators at random from a population of operators and having each operator measure each part n times. The parts are randomly presented to the operators so that the operators do not know which part is being measured.

In terms of the terminology used in the literature in this context, and yet keeping some of the above terminology in variance components as well, the data from such a study can be analyzed based on the variance components model:

$$\begin{aligned} y_{ijk} &= \mu + O_i + P_j + OP_{ij} + \epsilon_{ijk}, \\ i &= 1, \ldots, p; \quad j = 1, \ldots, o; \quad k = 1, \ldots, n, \end{aligned} \tag{5.41}$$

where $O_i \sim N(0, {\sigma^2}_\alpha)$, $P_j \sim N(0, {\sigma^2}_\beta)$, $OP_{ij} \sim N(0, {\sigma^2}_\gamma)$, and $\epsilon_{ijk} \sim N(0, {\sigma^2}_\epsilon)$ are independent. The parameters of the model are the variance components ${\sigma^2}_\alpha$, ${\sigma^2}_\beta$, ${\sigma^2}_\gamma$ and ${\sigma^2}_\epsilon$. Using the terminology in this class of applications, let ${\sigma^2}_{repeatability} = {\sigma^2}_\epsilon$ and ${\sigma^2}_{process} = {\sigma^2}_\alpha$. There are many quantities of interest that become important in various aspects of assessing a measurement system. Some of the important sums in various applications are

$$\begin{aligned} {\sigma^2}_{reproducibility} &= {\sigma^2}_\beta + {\sigma^2}_\gamma, \\ {\sigma^2}_{gauge} &= {\sigma^2}_{reproducibility} + {\sigma^2}_{repeatability}, \end{aligned}$$

measurement system variability, and

$$ {\sigma^2}_{total} = {\sigma^2}_{process} + {\sigma^2}_{gauge}, $$

the total variability.
Some ratios of interest widely addressed in the literature are

$$\frac{{\sigma^2}_{gauge}}{{\sigma^2}_{total}}, \frac{{\sigma^2}_{repeatability}}{{\sigma^2}_{total}}, \frac{{\sigma^2}_{reproducibility}}{{\sigma^2}_{total}},$$
$$\frac{{\sigma^2}_{gauge}}{{\sigma^2}_{process}}, \frac{{\sigma^2}_{repeatability}}{{\sigma^2}_{process}}, \frac{{\sigma^2}_{reproducibility}}{{\sigma^2}_{process}},$$

and

$$\frac{{\sigma^2}_{repeatability}}{{\sigma^2}_{gauge}}.$$

Hamada and Weerahandi (2000) discussed various specific inference problems involving some of these quantities. For example, according to the Automotive Industry Action Group standards, the ratio $\frac{{\sigma^2}_{gauge}}{{\sigma^2}_{process}}$ should be less than 0.2 in order for system to be considered adequate. Tsai (1988) reported

an application in which $\frac{5.15\sigma_{gauge}}{tolerance\ \ range}$ should be less than 0.1. In some applications, a measurement system's variation is also measured by $6\sigma_{gauge}$, which covers 99.73% of the measurement system variation. According to Montgomery (1991), the parameter $\frac{6\sigma_{gauge}}{tolerance\ range}$ should be less than 0.1 for a measurement system to be adequate.

5.7.2 The general problem

The problems involving functions of variance components arise in various mixed models involving higher-way layouts including the nested designs that we discussed above. Here we consider only those problems that can be expressed in the canonical form and concentrate only on sums and ratios of variance components. Suppose there are N variance components, say σ_1^2, σ_2^2, ..., σ_N^2 , in a certain model and let s_1, s_2, ..., s_N be the observed sums of squares that can provide adequate information to enable all inferences concerning the variance components. The random variables and the degrees of freedom corresponding to these sums of squares are denoted as S_1, S_2, ..., $S_N$and d_1, d_2,...,d_N, respectively.

As in the above two-way ANOVA, in the canonical form, the distributions of the sums of squares are related to the chi-squared random variables

$$\begin{aligned} Y_1 = S_1/\theta_1 &\sim \chi^2_{d_1} \\ Y_2 = S_2/\theta_2 &\sim \chi^2_{d_2} \\ &\vdots \\ Y_N = S_N/\theta_N &\sim \chi^2_{d_N}, \end{aligned} \tag{5.42}$$

where θ_1, θ_2, ..., θ_N, are the expected mean sum of squares that arise in the ANOVA. As can be seen from model (5.41), these are certain linear functions of the variance components, but are not necessarily the functions that we are interested in. Let

$$\Sigma = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \vdots \\ \sigma_N^2 \end{pmatrix}$$

be the vector of variance components and let

$$\Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{pmatrix}$$

be the vector of equations formed by expected mean sums of squares. The linear functions that relates former to the latter can be expressed as

$$\Theta = A\Sigma \tag{5.43}$$

where A is an $N \times N$ matrix. For example, in the two-way random effects model (5.41),

$$A = \begin{pmatrix} JK & 0 & K & 1 \\ 0 & IK & K & 1 \\ 0 & 0 & K & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

is an easily invertible triangular matrix. We assume that A is invertible, as is the case in typical problems. Then, we can express Σ in terms Θ of as

$$\Sigma = A^{-1}\Theta \tag{5.44}$$

In the above example, the matrix A^{-1} is also a triangular matrix:

$$A^{-1} = \begin{pmatrix} J^{-1}K^{-1} & 0 & -J^{-1}K^{-1} & 0 \\ 0 & I^{-1}K^{-1} & -I^{-1}K^{-1} & 0 \\ 0 & 0 & K^{-1} & -K^{-1} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.45}$$

5.7.3 Inference on linear functions of variance components

Hamada and Weerahandi (2000) considered the problem of making inferences of a linear combination of variance components of the two-way random effects model. More generally, consider the problem of making inferences concerning a certain linear combination of variance components of the above general model, say

$$\theta = \mathbf{c}^{'}\Sigma \tag{5.46}$$

where $\mathbf{c}$ is an $N \times 1$ vector. For example, in the Gauge R & R study, the $\mathbf{c}$ vectors appearing in the table below are all important.

Table 5.7 The $\mathbf{c}$ vectors for the R & R study

quantity	$\mathbf{c}^{'}$
$\sigma^2_{repeatability}$	(0,0,0,1)
$\sigma^2_{reproducibility}$	(0,1,1,0)
σ^2_{gauge}	(0,1,1,1)
$\sigma^2_{process}$	(1,0,0,0)
σ^2_{total}	(1,1,1,1)

Using (5.44), the parameter of interest θ can be expressed as

$$\theta = \mathbf{k}^{'}\Theta, \tag{5.47}$$

where

$$\mathbf{k}^{'} = \mathbf{c}^{'}A^{-1} \tag{5.48}$$

is a vector of constants with no unknown parameters. Continuing to illustrate with the two-way random effects model, the $\mathbf{k}$ particular vector in that case is

$$\mathbf{k} = \begin{pmatrix} c_1 J^{-1}K^{-1} \\ c_2 I^{-1}K^{-1} \\ K^{-1}(-c_1 J^{-1} - c_2 I^{-1} + c_3) \\ K^{-1}(-c_3 + c_4) \end{pmatrix} \tag{5.49}$$

Now consider, in particular, the problem of testing of hypotheses of the form

$$H_0 : \theta \leq \theta_* \tag{5.50}$$

where is θ_* a specified constant. Since the generalized confidence intervals can be deduced from generalized p-values, let us consider the problem of deriving the latter so that we can address the former in a later section of this chapter. To construct a generalized procedure for testing this hypothesis, consider the potential extreme region

$$C = \left(\mathbf{S}|\mathbf{k}^{'}\Theta(\mathbf{S};\mathbf{s}) \leq \theta\right) = \left(\mathbf{S}|W(\mathbf{S};\mathbf{s},\mathbf{k}) \leq \theta\right), \tag{5.51}$$

where $W(\mathbf{S};\mathbf{s}) = \mathbf{k}^{'}\Theta(\mathbf{S};\mathbf{s})$ is the potential test variable**,** $\mathbf{S} = (S_1, S_2, \ldots, S_N)$ and $\mathbf{s}$ is the observed value of $\mathbf{S}$, and $\Theta(\mathbf{S};\mathbf{s})$ is defined as

$$\Theta(\mathbf{S};\mathbf{s}) = \begin{pmatrix} \theta_1 s_1/S_1 \\ \theta_2 s_2/S_2 \\ \vdots \\ \theta_N s_N/S_N \end{pmatrix}. \tag{5.52}$$

When $\mathbf{S}$ takes on its observed value $\mathbf{s}$, it follows from (5.47) and (5.52) that $W = \mathbf{k}^{'}\Theta = \theta$, and hence the observed $\mathbf{s}$ falls on the boundary of the subset of the sample space defined by (5.51). Moreover, using (5.42) and (5.48) W can be expressed as

$$W = \mathbf{c}^{'}A^{-1}\begin{pmatrix} s_1/Y_1 \\ s_2/Y_2 \\ \vdots \\ s_N/Y_N \end{pmatrix} \tag{5.53}$$

$$= \frac{k_1 s_1}{Y_1} + \frac{k_2 s_2}{Y_2} \cdots + \frac{k_N s_N}{Y_N}, \tag{5.54}$$

where Y quantities are independent chi-squared random variables defined in (5.42) and k's are the components of k vector defined in (5.48). Hence, the probability of Y is free of unknown parameters and $\Pr(C)$ is an increasing function of θ. Therefore, the generalized p-value for testing (5.50) can be computed as

$$\begin{aligned} p &= \max_{H_0} \Pr(C) \\ &= F_W(\theta_*), \end{aligned} \tag{5.55}$$

where F_W is the cdf of the distribution of W.

Since W is a linear combination of independent chi-squared random variables, this probability can be evaluated by exact numerical integration. As in the generalized p-value applications in ANOVA and mixed models, it is computed more easily by Monte Carlo integration. In that approach, the p-value is evaluated by simulating the $\mathbf{Y}$ random vector a large number of times, say 100,000, and then calculating the proportion of times that $W(\mathbf{Y}; \mathbf{s}, \mathbf{c})$ variable is less than or equal to θ_*.

As before, the p-value given by (5.55) can also be expressed as an integral with respect to chi-squared random variables, or more preferably in terms of fewer number of beta random variables leading to well behaved exact numerical integrations. In the former representation, the generalized p-value is expressed as

$$\begin{aligned} p &= \Pr\left(\frac{k_1 s_1}{Y_1} + \frac{k_2 s_2}{Y_2} + \cdots + \frac{k_N s_N}{Y_N} \leq \theta_*\right) \\ &= E\left\{ I\left(\theta_* - \frac{k_1 s_1}{Y_1} + \frac{k_2 s_2}{Y_2} + \cdots + \frac{k_N s_N}{Y_N}\right)\right\}, \end{aligned} \tag{5.56}$$

where $I(x)$is an indicator variable, which takes on the values 1 or 0 depending on x is positive or not and the expectation is taken with respect to the chi-squared random variables. Clearly, the p-value can also be computed using the formula (cf. Appendix A.1)

$$\begin{aligned} p = 1 - E\{ & F_{\chi^2_{d_1+d_2+\cdots+d_N}} \\ & \left(\frac{1}{\theta_*}\left(\frac{k_1 s_1}{B_1 B_2 \cdots B_{N-1}} + \frac{k_2 s_2}{(1-B_1) B_2 \cdots B_{N-1}} + \cdots + \frac{k_N s_N}{(1-B_{N-1})}\right)\right)\}, \end{aligned} \tag{5.57}$$

where the expectation is taken with respect to the beta random variables

$$\begin{aligned} B_1 &= \frac{Y_1}{Y_1+Y_2} \sim \text{Beta}\left(\frac{d_1}{2}, \frac{d_2}{2}\right) \\ B_2 &= \frac{Y_1+Y_2}{Y_1+Y_2+Y_3} \sim \text{Beta}\left(\frac{d_1+d_2}{2}, \frac{d_3}{2}\right), \cdots, \\ B_{N-1} &= \frac{Y_1+Y_2+\cdots+Y_{N-1}}{Y_1+Y_2+\cdots+Y_N} \sim \text{Beta}\left(\frac{d_1+d_2+\cdots+d_{N-1}}{2}, \frac{d_N}{2}\right), \end{aligned}$$

and $F_{\chi^2_{d_1+d_2+\cdots+d_N}}$ is the cdf of the chi-squared distribution with $d_1 + d_2 + \cdots + d_N$ degrees of freedom. It should be noted that if some of the k's are zero, then the p-value could be expressed in terms of a lesser number of beta random variables corresponding to the non-zero k's.

5.7.4 Testing reproducibility

To illustrate the use of foregoing formulas in specific applications, consider the problem of testing the reproducibility in the Gauge R & R study of the form $H_0 : \theta = {\sigma^2}_{reproducibility} = {\sigma^2}_{\beta} + {\sigma^2}_{\gamma} \leq \theta_*$. In this application

$$\Sigma = \begin{pmatrix} \sigma^2_{\alpha} \\ \sigma^2_{\beta} \\ \sigma^2_{\gamma} \\ {\sigma^2}_{\epsilon} \end{pmatrix}$$

so that $\mathbf{c} = (0, 1, 1, 0)$ and

$$\begin{aligned} \Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} &= \begin{pmatrix} on{\sigma^2}_{\alpha} + n{\sigma^2}_{\gamma} + {\sigma^2}_{\epsilon} \\ pn{\sigma^2}_{\beta} + n{\sigma^2}_{\gamma} + {\sigma^2}_{\epsilon} \\ n{\sigma^2}_{\gamma} + {\sigma^2}_{\epsilon} \\ {\sigma^2}_{\epsilon} \end{pmatrix} \\ &= \begin{pmatrix} on & 0 & n & 1 \\ 0 & pn & n & 1 \\ 0 & 0 & n & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma^2_{\alpha} \\ \sigma^2_{\beta} \\ \sigma^2_{\gamma} \\ {\sigma^2}_{\epsilon} \end{pmatrix} \\ &= A\Sigma \end{aligned}$$

are the expected mean sums of squares.. By directly solving the linear equations, or using the inverse of A given by (5.45), for the sake of further illustration of the notations, we can express Σ in terms of Θ as

$$\Sigma = \begin{pmatrix} (\theta_1 - \theta_3)/on \\ (\theta_2 - \theta_3)/pn \\ (\theta_3 - \theta_4)/n \\ \theta_4 \end{pmatrix}$$

and hence

$${\sigma^2}_{reproducibility} = (\theta_2 - \theta_3)/pn + (\theta_3 - \theta_4)/n = \mathbf{k}^{'}\Theta,$$

where $\mathbf{k}^{'} = \begin{pmatrix} 0 & 1/pn & 1/n(1-1/p) & -1/n \end{pmatrix}$. This means that in this application the test variable $W(\mathbf{S}; \mathbf{s}) = \mathbf{k}^{'}\Theta(\mathbf{S}; \mathbf{s})$ becomes

$$W = \frac{s_o}{npY_2} + \frac{1}{n}(1 - \frac{1}{p})\frac{s_{op}}{Y_3} - \frac{s_e}{nY_4},$$

and the p-value can be computed by exact numerical integration

$$p = 1 - E\left\{F_\chi\left(\frac{1}{\theta_*}\left(\frac{s_o}{pnB_1B_2} + \frac{1}{n}(1-\frac{1}{p})\frac{s_{op}}{(1-B_1)B_2} - \frac{s_N}{n(1-B_2)}\right)\right)\right\}, \tag{5.58}$$

where F_χ is the cdf of the chi-squared distribution with $o-1+(o-1)(p-1)+op(n-1)$ degrees of freedom, and the expectation is taken with respect to the beta random variables

$$B_1 = \frac{Y_2}{Y_2+Y_3} \sim \text{Beta}\left(\frac{o-1}{2}, \frac{(o-1)(p-1)}{2}\right),$$
$$B_2 = \frac{Y_2+Y_3}{Y_2+Y_3+Y_4} \sim \text{Beta}\left(\frac{o-1+(o-1)(p-1)}{2}, \frac{op(n-1)}{2}\right).$$

5.7.5 Comparing variance components

As another illustration consider the problem of comparing variance components addressed by Zhou and Mathew (1994) to compare a new tube against a control tube used for firing ammunition from tanks. In their application the response variable was the miss distance and the tube-to-tube variance due to the new tube is compared with the control tube. They considered two independent balanced mixed models with variance components σ_x^2 and σ_y^2, among other variance components. It is of interest to compare the two variance components on the basis of the statistics S_x, S_y, S_G, and S_H with distributions:

$$Y_1 = \frac{S_G}{\sigma_g^2} \sim \chi_g^2, \quad Y_2 = \frac{S_H}{\sigma_h^2} \sim \chi_h^2,$$

$$Y_3 = \frac{S_x}{\sigma_g^2 + M\sigma_x^2} \sim \chi_x^2, \quad \text{and} \quad Y_4 = \frac{S_y}{\sigma_h^2 + N\sigma_y^2} \sim \chi_y^2,$$

where σ_g^2 and σ^2 are nuisance parameters, M and N are some known positive constants. The problem is to test hypotheses of the form

$$H_0 : \sigma_x^2 \leq \rho\sigma_y^2 \text{ versus } H_1 : \sigma_x^2 > \rho\sigma_y^2, \tag{5.59}$$

where ρ is a constant that is specified prior to testing the hypothesis. For the case $\rho = 1$, Zhou and Mathew (1994) described how unbiased tests can be obtained in this kind of situation. More generally, tests of (5.59) could be deduced from the above general results. In this application the parameter of interest is

$$\theta = \sigma_x^2 - \rho\sigma_y^2$$
$$= \frac{\theta_3 - \theta_1}{M} - \rho\frac{\theta_4 - \theta_2}{N}.$$

This means that in this application the test variable $W(\mathbf{S};\mathbf{s}) = \mathbf{k}^{'}\Theta(\mathbf{S};\mathbf{s})$ becomes

$$W = -\frac{s_G}{MY_1} + \rho\frac{s_H}{NY_2} + \frac{s_x}{MY_3} - \rho\frac{s_y}{NY_4}$$

and thus the generalized p-value for testing hypotheses is obtained as

$$\begin{aligned} p = &\ \Pr(W \ \leq \ 0) \\ &= \Pr(\rho M\frac{s_H}{Y_2} + N\frac{s_x}{Y_3} \leq N\frac{s_G}{Y_1} + \rho M\frac{s_y}{Y_4}). \end{aligned} \tag{5.60}$$

The p-value could also be computed using the formula

$$p = \Pr\left(\rho M\frac{s_H}{(1-B_3)} + N\frac{s_x}{B_1B_2B_3} \leq N\frac{s_G}{(1-B_2)B_3} + \rho M\frac{s_y}{(1-B_1)B_2B_3}\right), \tag{5.61}$$

which involve only 3 variables of integration, namely the Beta variables

$$B_1 \sim \text{Beta}(\frac{x}{2}, \frac{y}{2}),\ B_2 \sim \text{Beta}(\frac{x\ +\ y}{2}, \frac{g}{2}),$$

and

$$B_3\ \sim\ \text{Beta}(\frac{x\ +\ y\ +\ g}{2}, \frac{h}{2}),$$

which are independently distributed.

Example 4.6. Comparing measurements from two labs.

Two labs A and B test whether weight of a certain packaged food conform to a certain standard. Table below shows a set of data appropriate for testing the consistency of the labs. The illustrative data shown in the table represent 3 weight measurements taken by 7 operators randomly chosen from each of the two labs.

Table 5.8 Weight measurements by lab and operator

Operator	Lab A			Operator	Lab B		
A_1	15.63	16.38	16.64	B_1	15.65	15.60	15.84
A_2	14.87	15.08	15.01	B_2	15.82	16.02	15.46
A_3	15.68	15.17	15.59	B_3	15.69	15.71	15.21
A_4	15.87	16.02	15.99	B_4	15.73	15.56	15.47
A_5	15.75	15.82	15.99	B_5	15.45	15.23	15.39
A_6	15.69	15.82	16.24	B_6	16.03	15.67	15.78
A_7	14.68	15.14	15.13	B_7	15.48	15.91	16.12

Assuming a one-way random effects model for data from each program, let μ_A and μ_B be the two grand means of the weights of the product and σ_A^2 and σ_B^2 be the variances of random effects due to operators from lab A and B, respectively. Let δ^2 and σ^2 be the error variances of the two models. In this kind of applications, the problems of comparing the means and the variances are both important. The former can be accomplished by procedures discussed in Chapter 2 and is left as an exercise (see Exercise 4.15). Consider the null hypothesis $H_0 : \sigma_A^2 \leq \sigma_B^2$ that the among operator variability of the measurements at Lab A is no larger than that of Lab B. The ANOVA tables computed using data from Table 4.8 for the two labs are shown below.

ANOVA for lab A measurements.

Source	DF	SS	MS	E(MS)
Operators	6	4.319	0.720	$\delta^2 + 3\sigma_A^2$
Error	14	1.068	0.076	δ^2
Total	20	5.387		

ANOVA for lab B measurements

Source	DF	SS	MS	E(MS)
Operators	6	0.553	0.092	$\sigma^2 + 3\sigma_B^2$
Error	14	0.695	0.050	σ^2
Total	20	1.248		

Based on the quantities in the two ANOVA tables, we can test the desired hypothesis using the generalized p-value given by Formula (5.60). The p-value is computed as

$$\begin{aligned} p &= \Pr\left(\frac{s_A}{Y_2} + \frac{s_x}{Y_3} \leq \frac{s_G}{Y_1} + \frac{s_y}{Y_4}\right) \\ &= \Pr\left(\frac{0.695}{Y_2} + \frac{4.319}{Y_3} \leq \frac{1.068}{Y_1} + \frac{0.553}{Y_4}\right), \\ &= 0.017, \end{aligned}$$

where the probability is computed with respect to the chi-squared random variables

$$Y_1 = \frac{S_{AE}}{\delta^2} \sim \chi^2_{14}, \quad Y_2 = \frac{S_{BE}}{\sigma^2} \sim \chi^2_{14}$$

and

$$Y_3 = \frac{S_{AO}}{\delta^2 + 3\sigma_A^2} \sim \chi^2_6, \quad \text{and} \quad Y_4 = \frac{S_{BO}}{\sigma^2 + 3\sigma_B^2} \sim \chi^2_6 .$$

Hence, we have sufficient evidence to reject the null hypothesis and conclude that the measurements taken at Lab A has a higher variation than that of Lab B.

5.7.6 Inference on ratios of linear functions of variance components

Hamada and Weerahandi (2000) also considered the problem of making inferences of specific ratios of linear combinations of variance components in a two-way random effects model. More generally, consider the problem of making inferences concerning a certain linear combination of variance components of the general model in canonical form, say

$$\rho = \frac{\mathbf{c}_1^{'}\Sigma}{\mathbf{c}_2^{'}\Sigma}, \tag{5.62}$$

where $\mathbf{c}_1$and $\mathbf{c}_2$ are vectors of known constants derived from the ratio of interest. Hamada and Weerahandi (2000) gave the values of these vectors for many quantities that are important in Gauge R & R studies. The quantities that they reported along with corresponding $\mathbf{c}_1$and $\mathbf{c}_2$ vectors are shown in Table 4.9.

Table 5.9 The **a** and **b** vectors for the ratios

quantity	$\mathbf{c}_1$	$\mathbf{c}_2$
${\sigma^2}_{gauge}/{\sigma^2}_{total}$	(0,1,1,1)	(1,1,1,1)
${\sigma^2}_{repeatability}/{\sigma^2}_{total}$	(0,0,0,1)	(1,1,1,1)
${\sigma^2}_{reproducibility}/{\sigma^2}_{total}$	(0,1,1,0)	(1,1,1,1)
${\sigma^2}_{gauge}/{\sigma^2}_{process}$	(0,1,1,1)	(1,0,0,0)
${\sigma^2}_{repeatability}/{\sigma^2}_{process}$	(0,0,0,1)	(1,0,0,0)
${\sigma^2}_{reproducibility}/{\sigma^2}_{process}$	(0,1,1,0)	(1,0,0,0)
${\sigma^2}_{repeatability}/{\sigma^2}_{gauge}$	(0,0,0,1)	(0,1,1,1)

Using equation (5.47), the parameter of interest can also be expressed in terms of Θ as

$$\rho = \frac{\mathbf{k}^{'}\Theta}{\mathbf{l}'\Theta}, \tag{5.63}$$

where $\mathbf{k}^{'} = \mathbf{c}_1^{'}A^{-1}$and $\mathbf{l}^{'} = \mathbf{c}_2^{'}A^{-1}$. Since the generalized confidence intervals can be deduced from generalized confidence intervals, first consider the problem of testing of hypotheses of the form

$$H_0 : \rho \leq \rho_*, \tag{5.64}$$

where is ρ_* a specified constant. The problem of generalized confidence intervals for linear combinations of variance components and their ratios will be jointly addressed in the next section.

To construct a generalized procedure for testing this hypothesis, in view of (5.51), consider now a potential extreme region of the form

$$\begin{aligned} C &= \left(\mathbf{S}|\frac{\mathbf{k}^{'}\Theta(\mathbf{S};\mathbf{s})}{\mathbf{l}'\Theta(\mathbf{S};\mathbf{s})} \leq \rho\right) \\ &= \left(\mathbf{S}|\frac{W(\mathbf{S};\mathbf{s},\mathbf{k})}{W(\mathbf{S};\mathbf{s},\mathbf{l})} \leq \rho\right). \end{aligned} \tag{5.65}$$

where $\Theta(\mathbf{S};\mathbf{s})$ is as defined by equation (5.52). The p-value given by this extreme region can be expressed as

$$\begin{aligned} p &= \Pr\left(\frac{W(\mathbf{S};\mathbf{s},\mathbf{k})}{W(\mathbf{S};\mathbf{s},\mathbf{l})} \leq \rho_*\right) \\ &= \Pr\left(\frac{\frac{k_1 s_1}{Y_1} + \frac{k_2 s_2}{Y_2} \cdots + \frac{k_N s_N}{Y_N}}{\frac{l_1 s_1}{Y_1} + \frac{l_2 s_2}{Y_2} \cdots + \frac{l_N s_N}{Y_N}} \leq \rho_*\right). \end{aligned} \tag{5.66}$$

In typical applications, the denominator of (5.62) is a positive linear combinations of variance components. In this case, equation (5.63) reduces to

$$H_0 : \theta \leq 0,$$

where $\theta = \mathbf{c}_1^{'}\Sigma - \rho_*\mathbf{c}_2^{'}\Sigma$, a special case of (5.50). Then, the computation of p-value can be greatly simplified as

$$p = \Pr\left(\frac{m_1 s_1}{Y_1} + \frac{m_2 s_2}{Y_2} + \cdots + \frac{m_N s_N}{Y_N} \leq 0\right), \tag{5.67}$$

where $m_i = k_i - l_i\rho_*$, $i = 1, \cdots, N$.

This p-value can be evaluated by Monte Carlo integration by simulating $\mathbf{Y}$ random vector a large number of times, say 100,000, and then calculating the proportion of times the inequality in (5.67) is satisfied. This p-value can also be evaluated by exact numerical integration in terms of chi-squared random variables using the form (5.56) or in terms of beta random variables using (5.57). Actually, the exact numerical integration of (5.67) can be further simplified when we know the sign of m_i of variables.

5.7.7 Illustration

To illustrate the application of (5.66), consider problems of testing the repeatability and reproducibility in the Gauge R & R study. Consider, in particular hypotheses of the form

$$H_{01} : \frac{{\sigma^2}_{repeatability}}{{\sigma^2}_{total}} \leq \rho_* \tag{5.68}$$

where ρ_* is a specified known quantity. It follows from formula (5.49) that

$$\begin{aligned}
{\sigma^2}_{repeatability} &= \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix}\Theta,\\
{\sigma^2}_{reproducibility} &= \begin{pmatrix} 0 & 1/pn & 1/n(1-1/p) & -1/n \end{pmatrix}\Theta,\\
{\sigma^2}_{total} &= \begin{pmatrix} 1/on & 1/pn & 1/n(1-1/o-1/p) & 1-1/n \end{pmatrix}\Theta.
\end{aligned}$$

In testing the hypothesis (5.68), since ${\sigma^2}_{total}$ is a positive quantity, equation (5.66) reduces to

$$\begin{aligned}
p &= \Pr\left(\frac{\frac{s_e}{Y_4}}{\frac{s_p}{onY_1}+\frac{s_o}{pnY_2}+\frac{s_{op}(1-1/p-1/o)}{nY_3}+\frac{s_e(1-1/n)}{Y_4}} \le \rho_*\right)\\
&= \Pr\left(\frac{s_e}{Y_4}-\rho_*\left(\begin{array}{c}\frac{s_p}{onY_1}+\frac{s_o}{pnY_2}+\frac{s_{op}(1-1/p-1/o)}{nY_3}\\ +\frac{s_e(1-1/n)}{Y_4}\end{array}\right)\le 0\right)\\
&= \Pr\left(\frac{s_e}{Y_4}-\rho_*\left(\frac{s_o}{pnY_2}+\frac{s_{op}(1-1/p-1/o)}{nY_3}+\frac{s_e(1-1/n)}{Y_4}\right)\le\frac{\rho_* s_p}{onY_1}\right)\\
&= \Pr\left(\frac{s_e}{Y_4}(1/\rho_*-1+1/n)\le\frac{s_p}{onY_1}+\frac{s_o}{pnY_2}+\frac{s_{op}(1-1/p-1/o)}{nY_3}\right)\\
&= \Pr\left(P(m_4 s_e\left[\frac{m_1 s_p}{Y_1}+\frac{m_2 s_o}{Y_2}+\frac{m_3 s_{op}}{Y_3}\right]^{-1}\le Y_4\right), \qquad (5.69)
\end{aligned}$$

where $m_1 = 1/on$, $m_2 = 1/pn$, $m_3 = (1-1/p-1/o)/n$, and $m_4 = 1/\rho_* - 1 + 1/n$ are all positive constants with practical values of p, o, and ρ_*, which is a ratio. Since Y_4 is a chi-squared random variable with $d_4 = op(n-1)$ degrees of freedom, we can express (5.69) as

$$\begin{aligned}
p = 1 - \int\int F_{\chi^2_{op(n-1)}}&\left(m_4 s_e\left[\frac{m_1 s_p}{Y_1}+\frac{m_2 s_o}{Y_2}+\frac{m_3 s_{op}}{Y_3}\right]^{-1}\right)\\
&\times f_{Y_1}(Y_1) f_{Y_2}(Y_2) f_{Y_3}(Y_3) dY_1 dY_2 dY_3, \qquad (5.70)
\end{aligned}$$

where the integration is taken with respect to the random variables, Y_1, Y_2, Y_3 with chi-squared distributions

$$\begin{aligned}
Y_1 &\sim \chi^2_{d_1},\\
Y_2 &\sim \chi^2_{d_2},\\
Y_3 &\sim \chi^2_{d_3}, \qquad (5.71)
\end{aligned}$$

where $d_1 = p-1$, $d_2 = o-1$ and $d_3 = (o-1)(p-1)$. To obtain the numerically more stable integral representation of (5.70) involving beta random variables,

as before, define jointly independent random variables,

$$\begin{aligned} B_1 &= \frac{Y_1}{Y_1+Y_2} \sim \text{Beta}(\frac{d_1}{2}, \frac{d_2}{2}) \\ B_2 &= \frac{Y_1+Y_2}{Y_1+Y_2+Y_3} \sim \text{Beta}(\frac{d_1+d_2}{2}, \frac{d_3}{2}) \\ B_3 &= \frac{Y_1+Y_2+Y_3}{Y_1+Y_2+Y_3+Y_4} \sim \text{Beta}(\frac{d_1+d_2+d_3}{2}, \frac{d_4}{2}) \\ V &= Y_1+Y_2+Y_3+Y_4 \sim \chi^2_{d_s}, \end{aligned} \tag{5.72}$$

where $d_s = d_1+d_2+d_3+d_4$. Let $B_{-1} = (1-B_1)$ and $B_{-2} = (1-B_2)$. It is now evident that the p-value can be expressed as

$$\begin{aligned} p &= \Pr\left(\frac{m_4 s_e}{Y_4} \leq \frac{m_1 s_p}{Y_1} + \frac{m_2 s_o}{Y_2} + \frac{m_3 s_{op}}{Y_3}\right) \\ &= \Pr\left(\frac{m_4 s_e}{Y_4} \leq \frac{1}{V}\left[\frac{m_1 s_p}{B_1 B_2 B_3} + \frac{m_2 s_o}{B_{-1} B_2 B_3} + \frac{m_3 s_{op}}{B_{-2} B_3}\right]\right) \\ &= 1 - E\left\{F_{F_{d_s,d_4}}\left(\frac{m_4 s_e d_4}{d_s}(\frac{m_1 s_p}{B_1 B_2 B_3} + \frac{m_2 s_o}{B_{-1} B_2 B_3} + \frac{m_3 s_{op}}{B_{-2} B_3})\right)\right\}. \end{aligned} \tag{5.73}$$

As another illustration, consider the problem of testing the hypothesis

$$H_{02} : \frac{{\sigma^2}_{reproducibility}}{{\sigma^2}_{total}} \leq \rho_* \tag{5.74}$$

To illustrate the direct application of (5.67), first of all note that the hypothesis of interest can be rewritten as

$$H_{02} : \sigma^2_{diff} \leq 0, \tag{5.75}$$

where $\sigma^2_{diff} = {\sigma^2}_{reproducibility} - \rho_* {\sigma^2}_{total}$. This parameter can be expressed as

$$\begin{aligned} \sigma^2_{diff} &= \begin{pmatrix} 0 & 1/pn & 1/n(1-1/p) & -1/n \end{pmatrix} \Theta \\ &\quad - \rho_* \begin{pmatrix} 1/on & 1/pn & 1/n(1-1/o-1/p) & 1-1/n \end{pmatrix} \Theta \\ &= \begin{pmatrix} \frac{-\rho_*}{on} & \frac{\rho_{-*}}{pn} & \frac{(1-1/p)\rho_{-*}}{n} + \frac{\rho_*}{on} & -\rho_* - \frac{\rho_{-*}}{n} \end{pmatrix} \Theta, \end{aligned}$$

where $\rho_{-*} = (1-\rho_*)$. Then, it follows from (5.67) that

$$\begin{aligned} p &= \Pr\left(-\frac{m_1 s_p}{Y_1} + \frac{m_2 s_o}{Y_2} + \frac{m_3 s_{op}}{Y_3} - \frac{m_4 s_e}{Y_4} \leq 0\right) \\ &= \Pr\left(\frac{m_2 s_o}{Y_2} + \frac{m_3 s_{op}}{Y_3} \leq \frac{m_1 s_p}{Y_1} + \frac{m_4 s_e}{Y_4}\right), \end{aligned} \tag{5.76}$$

where $m_1 = 1/on$, $m_2 = (1/\rho_* - 1)/pn$, $m_3 = (1-1/p)(1/\rho_* - 1)/n + 1/on$, and $m_4 = 1 + (1/\rho_* - 1)/n$. The p-value could be easily computed by Monte

Carlo integration. As in previous cases, the probability can also be evaluated by numerical integration with respect to beta random variables, by careful choice of variables to keep the inequality valid:

$$\begin{aligned} p &= \Pr\left(-\frac{m_1 s_p}{Y_1} + \frac{m_2 s_o}{Y_2} + \frac{m_3 s_{op}}{Y_3} \leq \frac{m_4 s_e}{Y_4}\right) \\ &= \Pr\left(\frac{1}{V}\left[-\frac{m_1 s_p}{B_1 B_2 B_3} + \frac{m_2 s_o}{B_{-1} B_2 B_3} + \frac{m_3 s_{op}}{B_{-2} B_3}\right] \leq \frac{m_4 s_e}{Y_4}\right) \\ &= 1 - E\left\{F_{F_{d_s,d_4}}\left(\frac{m_4 s_e d_4}{d_s}\left(\frac{-m_1 s_p}{B_1 B_2 B_3} + \frac{m_2 s_o}{B_{-1} B_2 B_3} + \frac{m_3 s_{op}}{B_{-2} B_3}\right)\right)\right\}, \quad (5.77) \end{aligned}$$

where the beta random variables are as defined in (5.72).

5.7.8 Generalized confidence intervals

Generalized confidence intervals can be constructed using the above generalized p-values. To do this, let τ be the linear combination or the ratio of such linear combinations of interest. Recall that we were able to find test variable $T(\mathbf{Y};\mathbf{s}) = W(\mathbf{S};\mathbf{s},\mathbf{k})$ or $W(\mathbf{S};\mathbf{s},\mathbf{k},\mathbf{l})$ such that $P(T(\mathbf{Y};\mathbf{s},) \leq \tau_*)$ is the p-value for left-sided null hypotheses and similarly $P(W(\mathbf{S};\mathbf{s},\mathbf{k}) \geq \tau_*)$ is the p-value for right-sided null hypotheses. Therefore, we can obtain $100(1-\alpha)\%$ equal-tail generalized confidence intervals of the form (τ_L, τ_U) by solving $P(T(\mathbf{Y};\mathbf{s}) \geq \tau_L) = 1 - \alpha/2$ and $P(T(\mathbf{Y};\mathbf{s}) \geq \tau_U) = \alpha/2$. For example, τ_L is the root of the equation $P(T(\mathbf{Y};\mathbf{s}) \geq \tau_L) - (1 - \alpha/2) = 0$.

A simple way to obtain confidence intervals is by simulation. In the implementation of this method, we simulate the random vector $\mathbf{Y}$ N times, calculate $T(\mathbf{Y};\mathbf{s})$ for each $\mathbf{Y}$, and then use the $100\alpha/2$th and $100(1-\alpha/2)$th quantiles of the simulated T values. For example, $\mathbf{Y}$ could be simulated 100,000 times resulting in 100,00 T's. Then the 95% confidence intervals would be the 2500th and 97500th values of the ordered T's. Note that τ_L or δ_L may be negative in which case they should be set to zero since variances are non-negative. An alternative is to obtain confidence intervals from the positive T's. The distributions of T describe the plausible values of τ, that are consistent with the $\mathbf{s}$ data. The (1-α)100% confidence intervals then are the central (1-α)100% of these distributions. As Hamada and Weerahandi (2000) pointed out, the medians of the T distribution could be used to obtain a point estimate of τ.

Example 4.5. Inference in measurement systems

To illustrate the utility of above procedures, consider the data reported in Table 4.2 involving a measurement system with 20 parts, 3 operators, and 2 measurements taken from each combination. For this data set, $(p, o, n) = (20, 3, 2)$ and the vector of observed sums of squares is $\mathbf{s} = (s_p, s_o, s_{op}, s_e) = (1185.43, 2.62, 27.05, 59.50)$, which is the third column of the ANOVA table.

A number of articles including those of Montgomery (1991) and Hamada and Weerahandi (2000) considered various problems involving functions of

variance components of the two-way ANOVA model (5.41). For example, to answer the question of how good the measurement system is, they considered the following hypothesis:

$$H_{02} : \frac{\sigma^2_{gauge}}{\sigma^2{}_{total}} \leq 0.1$$

It can be deduced from (5.67) that the generalized p-value for testing this hypothesis is (see Exercise 4.12)

$$p = P(Y_1 \geq \frac{0.1 s_P}{6}(\frac{d_1 s_o}{Y_2} + \frac{d_2 s_{op}}{Y_3} + \frac{d_3 s_e}{Y_4})^{-1}), \tag{5.78}$$

where $d_1 = (1-\delta_*)/np = 0.0225$, $d_2 = (1/n)(\delta_*/o+(1-\delta_*)(1-1/p)) = 0.4442$ and $d_3 = (1-\delta_*)(n-1) = 0.9$. We can conveniently evaluate the p-value by first generating a large number random numbers from $Y_1 \sim \chi^2_{19}, Y_2 \sim \chi^2_2, Y_3 \sim \chi^2_{38}$, and $Y_4 \sim \chi^2_{119}$, say 100,000 from each, and then calculating the proportion of times that the inequality in (5.78) is satisfied. The p-value obtained in this manner is 0.67, suggesting that the data provides no support to reject the null hypothesis. It can also be shown that the generalized confidence interval for the parameter $\rho = \sigma^2_{gauge}/\sigma^2{}_{total}$ is [0.04,0.19]. This confidence interval also leads to the same conclusion. Hamada and Weerahandi (2000) computed the 95% generalized confidence intervals by simulating 100,000 random numbers. They are presented in the table below along with their point estimates. The table also provides classical confidence intervals when available; for many quantities classical intervals are not available and are left blank. Note that the generalized confidence intervals are wider than those of the ANOVA-based confidence intervals especially for σ_{gauge}. This is a result of overcoming the undercoverage problem of ANOVA-based confidence intervals.

Confidence intervals for quantities in R & R studies

	Generalized Inference		Classical ANOVA	
Quantity	Estimate	C.I.	Estimate	C.I.
$\sigma^2{}_{repeatability}$	1.00	(0.72, 1.48)	0.88	(0.68, 1.19)
$\sigma^2{}_{reproducibility}$	0.12	(.004, 3.78)	0.011	(0*, 1.28)
$\sigma^2{}_{gauge}$	0.94	(0.70, 2.18)	0.89	(0.69, 1.19)
$\sigma^2{}_{process}$	10.65	(5.86, 21.98)		
$\frac{\sigma^2{}_{gauge}}{\sigma^2{}_{total}}$	0.08	(0.04, 0.19)		
$\frac{\sigma^2{}_{repeatability}}{\sigma^2{}_{total}}$	0.09	(0.04, 0.16)		
$\frac{\sigma^2{}_{reproducibility}}{\sigma^2{}_{total}}$	0.01	(0.003, 0.25)		
$\frac{\sigma^2{}_{gauge}}{\sigma^2{}_{process}}$	0.09	(0.04, 0.24)		
$\frac{\sigma^2{}_{repeatability}}{\sigma^2{}_{process}}$	0.09	(0.04, 0.19)		
$\frac{\sigma^2{}_{reproducibility}}{\sigma^2{}_{process}}$	0.01	(0.0003, 0.37)		

Exercises

5.1 Consider the two-way random effects model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$
$$i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K.$$

Assuming that

$$\alpha_i \sim N(0, \sigma_\alpha^2), \beta_j \sim N(0, \sigma_\beta^2),$$
$$\gamma_{ij} \sim N(0, \sigma_\gamma^2), \text{ and } \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2),$$

show that

$$E(S_\alpha) = (I-1)(JK\sigma_\alpha^2 + K\sigma_\gamma^2 + \sigma_\epsilon^2),$$

where $S_\alpha = JK \sum_{i=1}^{I} (\overline{Y}_{i.} - \overline{Y})^2$.

5.2 Consider again the two-way random effects model in Exercise 4.1. Show that

$$\frac{S_\alpha}{JK\sigma_\alpha^2 + K\sigma_\gamma^2 + \sigma_\epsilon^2} \sim \chi_{I-1}^2.$$

5.3 Consider the two-way random effects model. Find the distribution of the grand mean $\overline{Y}$ if there are no fixed effects in the model. Establish procedures for testing point null hypotheses concerning the mean μ. Also establish the form of lower confidence bounds for μ.

5.4 Repeat Exercise 4.3 if α is a fixed effect and β and γ are random effects.

5.5 Consider the following random effects model that arise in a certain hierarchical classification:

$$Y_{ijk} = \mu + \epsilon_i + \epsilon_{ij} + \epsilon_{ijk},$$
$$i = 1, \ldots, I; \ j = 1, \ldots, J; \ k = 1, \ldots, K,$$

where the random error terms ϵ_i, ϵ_{ij}, and ϵ_{ijk} are all independently and normally distributed with zero means and certain variances. Discuss whether or not the variance components have the canonical form. Establish procedures for testing and constructing confidence intervals for the variance components.

5.6 Consider the mixed model

$$Y_{ijk} = \mu_i + \gamma_{ij} + \epsilon_{ijk},$$
$$i = 1, \ldots, I; \ j = 1, \ldots, J; \ k = 1, \ldots, K,$$

where $\mu_i, \ i = 1, \ldots, I$ are I fixed effects, $\gamma_{ij} \sim N(0, \sigma_\gamma^2)$, and ϵ_{ijk} are normally and independently distributed error terms. Discuss whether or not the variance component σ_γ^2 has the canonical form. Establish procedures for making inferences about the parameters μ_i's and the variance component σ_γ^2.

5.7 Consider the data set given in Table 4.2 involving a measurement system. Test each of the following hypotheses:

(a) $H_0 : {\sigma^2}_{repeatability} < 2* {\sigma^2}_{gauge}$,
(b) $H_0 : {\sigma^2}_{gauge}/{\sigma^2}_{reproducibility} > 2.5$.

5.8 Consider again he data in Table 4.2. Construct 95% confidence intervals for each of the following quantities:

(a) $\phi_1 = \frac{{\sigma^2}_{repeatability}}{{\sigma^2}_{gauge}}$,
(b) $\phi_2 = \frac{{\sigma^2}_{gauge}}{{\sigma^2}_{reproducibility}}$.

5.9 Consider a problem of hypothesis involving $\phi_3 = \phi_1 + \phi_2$ in the previous problem. Using the generalized approach, derive the generalized p-value for testing the hypothesis. Construct the form of confidence intervals for ϕ_3. Apply your formula in constructing a 95% confidence interval for ϕ_3 with data from Table 4.2.

5.10 Consider the two-way random effects model (5.41) and suppose σ^2_{gauge} is the parameters of interest in a certain application.

(a) Express σ^2_{gauge} in terms of ${\sigma^2}_{\alpha}$, ${\sigma^2}_{\beta}$, ${\sigma^2}_{\gamma}$, and ${\sigma^2}_{\epsilon}$.
(b) Express σ^2_{gauge} in terms of the vector Θ.
(c) Find generalized p-values for testing left-sided hypotheses and right-sided hypotheses concerning the parameter.
(d) Find the equal-tail 95 % generalized confidence interval for σ^2_{gauge}.

5.11 Consider the two-way random effects model (5.41) and suppose σ^2_{total} is the parameter of interest in a certain application.

(a) Express σ^2_{total} in terms of the vector Θ
(b) Find the generalized p-value for testing right-sided hypotheses concerning the parameter.
(c) Find the equal-tail 95% generalized confidence interval for σ^2_{total}.
(d) Find the one-sided 95% generalized confidence intervals for σ^2_{total}.

5.12 Suppose that in a certain application the parameter of interest in above model is the ratio

$$\rho = \frac{\sigma^2_{gauge}}{{\sigma^2}_{total}}.$$

(a) Find the generalized p-value for testing the hypothesis

$$H_0 : \rho \leq \rho_*.$$

(b) Deduce the right-sided 99% generalized confidence intervals for ρ.

5.13 Consider the general model in canonical form involving N chi-squared random variables.

(a) Express the parameter $\theta = {\sigma^2}_1 + {\sigma^2}_2 + \cdots + {\sigma^2}_M$, in terms of Θ, where $M < N$.
(b) Find the generalized p-value for testing the hypothesis $H_0 : \theta \leq \theta_*$.
(c) Find the generalized p-value for testing the hypothesis

$$H_0 : \frac{{\sigma^2}_i}{\theta}$$

(d) Find the generalized p-value for testing the hypothesis

$$H_0 : \frac{\theta}{{\sigma^2}_1 + {\sigma^2}_2 + \cdots + \sigma^2_N}.$$

5.14 Consider the random effects model described in Exercise 9.6. Let ${\sigma^2}_1, {\sigma^2}_2$, and ${\sigma^2}_3$ be the variance components corresponding to ε_i, ε_{ij}, and ε_{ijk}.

(a) Establish procedures for making inference concerning linear combinations of the form $c_1{\sigma^2}_1 + c_2{\sigma^2}_2 + c_3{\sigma^2}_3$, where c_1, c_2,and c_3 are positive constants.
(b) Establish procedures for making inference concerning ratios of linear combinations of the form

$$\rho = \frac{c_1{\sigma^2}_1 + c_2{\sigma^2}_2 + c_3{\sigma^2}_3}{d_1{\sigma^2}_1 + d_2{\sigma^2}_2 + d_3{\sigma^2}_3},$$

where (c_1, c_2, c_3) and (d_1, d_2, d_3) are both vectors of positive constants.
(c) Construct the equal-tail 95% confidence interval for σ^2_ϵ based on the chi-squared distribution given by (3.50).
(d) Construct the left-sided 95% confidence interval for σ^2_α.
(e) Test the null hypothesis, $H_0 : \sigma^2_\alpha \leq 7$.
(f) Test the null hypothesis, $H_0 : \sigma^2_\gamma \leq 1$.

5.15 Weerahandi (1995) considered the data set shown below to compare the productivity of some factory workers under two musical programs.

Productivity under two musical programs

Program X								
Worker	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Mean	93.2	98.1	89.6	88.4	96.2	95.0	99.6	97.9
Variance	23.4	27.6	18.6	22.1	15.4	26.2	33.1	29.8
Program Y								
Worker	*I*	*J*	*K*	*L*	*M*	*N*	*O*	*P*
Mean	90.3	85.1	99.4	98.4	86.2	82.5	103.9	96.7
Variance	32.4	26.3	16.8	23.7	18.4	25.6	34.1	28.3

Assuming a one-way random model for the data from each program, let μ_x and μ_y be the two grand means, and let σ_x^2 and σ_y^2 be the variances of random effects due to workers, under the programs X and Y, respectively.

(a) Construct 90% confidence intervals for each of the variance components.
(b) Construct a 95% confidence interval for $\mu_x - \mu_y$.
(c) Test the null hypothesis $H_0 : \sigma_x^2 \leq 0.4\sigma_y^2$.

5.16 Consider the data given in Table 4.8 on the measurements taken by two labs. Assuming a one-way random effects model for the data from each lab, test the hypothesis that there is no difference between the mean weights of the packaged food as measured by the two labs. Also construct a 95% confidence interval for the difference in the two means.