RESEARCH DESIGN *and* STATISTICAL ANALYSIS

Third Edition

SAMPLE CHAPTER

JEROME L. MYERS • ARNOLD D. WELL
ROBERT F. LORCH, JR.

Published in 2010
by Routledge
270 Madison Avenue
New York, NY 10016
www.psypress.com
www.researchmethodsarena.com

Published in Great Britain
by Routledge
27 Church Road
Hove, East Sussex BN3 2FA

Routledge is an imprint of the Taylor & Francis Group, an Informa business

Typeset in Times by RefineCatch Limited, Bungay, Suffolk, UK
Printed and bound by Sheridan Books, Inc. in the USA on acid-free paper
Cover design by Design Deluxe

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data
Myers, Jerome L.
Research design and statistical analysis / Jerome L. Myers, Arnold D. Well. –3rd ed. / Robert F. Lorch, Jr.
p. cm.
Includes bibliographical references.
1. Experimental design. 2. Mathematical statistics. I. Well, A. (Arnold). II. Lorch, Robert Frederick, 1952– . III. Title.
QA279.M933 2010
519.5—dc22 2009032606

ISBN: 978–0–8058–6431–1 (hbk)

Contents

Preface

Like the previous editions, this third edition of *Research Design and Statistical Analysis* is intended as a resource for researchers and a textbook for graduate and advanced undergraduate students. The guiding philosophy of the book is to provide a strong conceptual foundation so that readers are able to generalize concepts to new situations they will encounter in their research, including new developments in data analysis and more advanced methods that are beyond the scope of this book. Toward this end, we continue to emphasize basic concepts such as sampling distributions, design efficiency, and expected mean squares, and we relate the research designs and data analyses to the statistical models that underlie the analyses. We discuss the advantages and disadvantages of various designs and analyses. We pay particular attention to the assumptions involved, the consequences of violating the assumptions, and alternative analyses in the event that assumptions are seriously violated.

As in previous editions, an important goal is to provide coverage that is broad and deep enough so that the book can serve as a textbook for a two-semester sequence. Such sequences are common; typically, one semester focuses on experimental design and the analysis of data from such experiments, and the other semester focuses on observational studies and regression analyses of the data. Incorporating the analyses of both experimental and observational data within a single textbook provides continuity of concepts and notation in the typical two-semester sequence and facilitates developing relationships between analysis of variance and regression analysis. At the same time, it provides a resource that should be helpful to researchers in many different areas, whether analyzing experimental or observational data.

CONTENT OVERVIEW

Also like the previous editions, this edition can be viewed as consisting of four parts:

1. Data exploration and basic concepts such as sampling distributions, elementary probability, principles of hypothesis testing, measures of effect size, properties of estimators,

and confidence intervals on both differences among means and on standardized effect sizes.

2. Between-subject designs; these are designs with one or more factors in which each subject provides a single score. Key elements in the coverage are the statistical models underlying the analysis of variance for these designs, the role of expected mean squares in justifying hypothesis tests and in estimating effects of variables, the interpretation of interactions, procedures for testing contrasts and for controlling Type 1 error rates for such tests, and trend analysis—the analysis and comparison of functions of quantitative variables.
3. Extension of these analyses to repeated-measures designs; these are designs in which subjects contribute several scores. We discuss nesting and counterbalancing of variables in research designs, present quasi-*F* ratios that provide approximate tests of hypotheses, and consider the advantages and disadvantages of different repeated-measures and mixed designs.
4. The fourth section provides a comprehensive introduction to correlation and regression, with the goal of developing a general framework for analysis that incorporates both categorical and quantitative variables. The basic ideas of regression are developed first for one predictor, and then extended to multiple regression. The expanded section on multiple regression discusses both its usefulness as a tool for prediction and its role in developing explanatory models. Throughout, there is an emphasis on interpretation and on identifying common errors in interpretation and usage.

NEW TO THIS EDITION

Although the third edition shares the overall goals of the previous editions, there are many modifications and additions. These include: (1) revisions of all of the chapters from the second edition; (2) seven new chapters; (3) more examples of the use of SPSS to conduct statistical analyses; (4) added emphasis on power analyses to determine sample size, with examples of the use of G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to do this; and (5) new exercises. In addition to the modifications of the text, there is a substantial amount of additional material at the website, www.psypress.com/research-design. The website contains the following: (1) SPSS syntax files to perform analyses from a wide range of designs and a hotlink to the G*Power program; (2) all of the data files used in the text and in the exercises in SPSS and Excel formats; (3) technical notes containing derivations of some formulas presented in the book; (4) extra material on multiple and on logistic regression; and (5) a solutions manual and the text's figures and tables for instructors only.

Additional chapters

There are seven new chapters in the book. Chapters 1, 13, and 27 were added to provide more emphasis on the connections between design decisions, statistical analyses, and the interpretation of results from a study. In addition, a chapter has been added at the end of each of the four sections noted above to provide integrated examples of applications of the principles and procedures covered in the sections.

Planning the Research. The first chapter provides a schema for thinking about the major steps involved in planning a study, executing it, and analyzing and interpreting the results. The

emphasis is on the implications of decisions made in the planning study for subsequent analyses and interpretation of results. The chapter establishes a critical theme for the rest of the book; namely, that design and statistical analyses go hand-in-hand.

Comparing Experimental Designs. Chapter 13, the first chapter in the third section on repeated-measures, provides a bridge between the second and third sections. It introduces blocking in research designs, the analysis of covariance, and repeated-measures and Latin square designs. Advantages and disadvantages of these designs and analyses are discussed. In addition, the important concept of the relative efficiency of designs is introduced, and illustrated with data and the results of computer sampling studies. This chapter reinforces the theme of the intimate connection between design and statistical analysis.

Review of Important Points and Cautions About Common Errors. Chapter 27 is intended to remind readers of points discussed in the book—points we believe to be important but sometimes overlooked—and to warn against common errors in analyzing and interpreting results. For example, the chapter reminds readers of the importance of carefully choosing a research design and the need for *a priori* power calculations. As another example, the chapter again emphasizes the distinction between statistical and practical, or theoretical, significance.

Integrated Analysis Chapters. Each chapter in the book covers a lot of conceptual and procedural territory, so the integrated analysis chapters provide opportunities to see how the concepts and analyses come together in the context of a research problem. In these chapters, we consider the design of a study and the analysis of the resulting data. The presentation includes discussion of the pros and cons of possible alternative designs, and takes the analysis through exploration of the data to inferential procedures such as hypothesis tests, including, where applicable, tests of contrasts, estimates of effect size, and alternatives to the standard analyses in consideration of possible violations of assumptions. These chapters also serve as a review of the preceding chapters in the section and, in some cases, are used to introduce additional methods.

Use of Statistical Software

We assume that most readers will have access to some statistical software package. Although we have used SPSS for most of our examples, the analyses we illustrate are available in most packages. At several points, we have indicated the relevant SPSS menu options and dialog box choices needed to carry out an analysis. In cases where certain analyses are not readily available in the menus of current versions of SPSS (and possibly other statistical packages), we have provided references to Internet sites that permit free, or inexpensive, downloads of relevant software; for example, programs for obtaining confidence intervals for various effect size measures. Also, although their use is not required in the book, we have provided a number of SPSS syntax files available at the book's website (see below) that can be used to perform analyses for a wide range of designs. We note that the syntax files were written using SPSS 17 and that the analyses reported in the book are also based on that version. Future versions may provide slightly different output, other options for analysis, or somewhat different syntax.

We have used G*Power 3 for power analyses in many of the chapters and in some exercises. This very versatile software provides both *a priori* and post hoc analyses for many designs and analyses, as well as figures showing the central and noncentral distributions for the test and

parameters under consideration. The software and its use are described by Faul, Erdfelder, Lang, and Buchner (2007). G*Power 3 can be freely downloaded from the website (http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/). Readers should register there in order to download the software and to be notified of any further updates. Description of the use of G*Power 3 and illustrations in the current book are based on Version 3.0.9. An excellent discussion of the use of that program has been written by Faul, Erdfelder, Lang, and Buchner (2007).

Exercises

As in previous editions, each chapter ends with a set of exercises. Answers to odd-numbered exercises are provided at the back of the book; all answers are available in the password-protected Instructor's Solution Manual available at the book's website. There are more than 40 exercises in the four new integrated-analysis chapters to serve as a further review of the material in the preceding chapters.

The Book Website

For the third edition, a variety of materials are available on the website, www.psypress.com/research-design. These include the following.

Data Files

A number of data sets can be accessed from the book's website. These include data sets used in analyses presented in the chapters, so that these analyses can be re-created. Also, there are additional data sets used in the exercises. All data files are available both in SPSS and Excel format, in order to make them easily accessible. Some of these data sets have been included in order to provide instructors with an additional source of classroom illustrations and exercises. For example, we may have used one of the tables in the book to illustrate analysis of variance, but the file can also be used to illustrate tests of contrasts that could follow the omnibus analysis. A listing of the data files and descriptions of them are available on the website.

SPSS Syntax Files

As mentioned above, a number of optional syntax files are provided on the website that can be used to perform analyses for a wide range of designs. These include analyses involving nesting of variables having random effects; tests involving a variety of contrasts, including comparisons of contrasts and of trend components; and a varied set of regression analyses. These files, together with a Readme Syntax file that describes their use, are available at the website.

Technical Notes

For the sake of completeness, we wanted to present derivations for some of the expressions used in the book—for example, standard errors of regression coefficients. Because these derivations are not necessary to understand the chapters, and may be intimidating to some readers, we have made them available as optional technical notes on the website.

Additional Chapters

Here we present two supplementary chapters in pdf format that go beyond the scope of the book. One is a brief introduction to regression analysis using matrix algebra. The other is an introduction to logistic regression that is more comprehensive than the brief section that we included in Chapter 23. Other material will be added at various times.

Teaching Tools

There is information on the website for instructors only. Specifically, there is a solutions manual for all the exercises in the book and electronic files of the figures in the book.

Errata

Despite our best intentions, some errors may have crept into the book. We will maintain an up-to-date listing of corrections.

ACKNOWLEDGMENTS

We wish to express our gratitude to Axel Buchner for helpful discussions about G*Power 3; to J. Michael Royer for permission to use the data from his 1999 study; to Jennifer Wiley and James F. Voss for permission to use the Wiley-Voss data; to Melinda Novak and Corrine Lutz for permission to use the self-injurious behavior data set; and to Ira Ockene for permission to use the Seasons data. The *Seasons* research was supported by National Institutes of Health, National Heart, Lung, and Blood Institute Grant HL52745 awarded to University of Massachusetts Medical School, Worcester, Massachusetts.

Special thanks go to those individuals who reviewed early chapters of the book and made many useful suggestions that improved the final product: Jay Parkes, University of New Mexico; John Colombo University Kansas; William Levine, University of Arkansas; Lawrence E. Melamed, Kent State University; and one anonymous reviewer. We also wish to thank our colleagues, Alexander Pollatsek and Caren Rotello, and the many graduate assistants who, over the years, have contributed to our thinking about the teaching of statistics.

We also wish to express our gratitude to several individuals at Routledge. We have been greatly helped in the development and publication of this book by Debra Riegert, Senior Editor, and her Editorial Assistant, Erin Flaherty, as well as by Nicola Ravenscroft, the Project Editor for this book, and Joseph Garver, who was responsible for the final technical editing. We also wish to thank the American Statistical Association, the Biometric Society, and the Biometrika Trustees for their permission to reproduce statistical tables.

Finally, as always, we are indebted to our wives, Nancy A. Myers, Susan Well, and Elizabeth Lorch, for their encouragement and patience during this long process.

Chapter 8
Between-Subjects Designs: One Factor

8.1 OVERVIEW

In Chapter 8, we consider a basic research design in which there is a single independent variable with several levels that make up the conditions of the study, no subject is tested in more than one condition, and each subject contributes one score to the data set. Like any experimental design, the one-factor, between-subjects design has advantages and disadvantages.

The primary advantage of the design is that it is simple in several respects. First, data collection is simple. Only one observation is taken from each subject. No additional measures are required for the purpose of matching subjects in different conditions. Nor is there a need to be concerned about the order of presentation of treatments, or the interval between tests, as in designs in which subjects are tested in several conditions. Second, there are fewer assumptions underlying the data analysis than in most other research designs. More complex designs involve additional assumptions that, if violated, increase the likelihood of drawing incorrect conclusions from our data. Finally, there are fewer calculations than in other designs, and decisions about how to draw inferences based on those calculations are less complicated.

One disadvantage of the between-subjects design is that it requires more subjects than designs in which subjects are tested in several conditions. A second disadvantage is that there is less control of nuisance variables, and therefore the error variance is larger than in other designs. In particular, because subjects in different conditions differ in characteristics such as ability and motivation, it is more difficult to assess the effects of conditions than in designs in which such individual differences are better controlled.

In between-subjects designs, subjects may either be sampled from existing populations, or be assigned randomly to one of several experimental conditions, or treatment levels. An example of the former is the *Seasons* study[1] in which individuals were sampled from populations differing with respect to various factors, including gender, educational level, and occupation. Strictly speaking,

[1] See the *Seasons* data set on the website for this book.

that study would be classified as an observational study. True experiments involve random assignment of subjects to levels of an independent variable; the independent variable is said to be manipulated and the design is often referred to as *completely randomized*. Whether the levels of the independent variable are observed or manipulated, the data analysis has much the same form and the underlying assumptions are the same.

We view each group of scores as a random sample from a *treatment population*. The first question of interest is whether the means of these treatment populations vary. To address this question, we introduce the *analysis of variance*, or *ANOVA*, in which the total variability in the data set is partitioned into two components, one reflecting the variance of the treatment population means, and a second that reflects only the effects of nuisance variables.

In addition to testing whether the treatment population means are equal, we want some way of evaluating the practical or theoretical importance of the independent variable. Therefore, following the development of the ANOVA, we focus on several measures of importance. We also consider the role of statistical power in the research design and relate power to measures of importance.

Throughout this chapter, we will have made certain assumptions to justify the calculations presented. When those assumptions are violated, Type 1 and Type 2 error rates may increase. Therefore, the chapter also discusses such consequences, and alternative procedures that may improve the situation.

In summary, the main concerns of this chapter are:

- *Testing the null hypothesis that the treatment population means are equal*. This involves the ANOVA for the one-factor between-subjects design.
- *Measures of the importance of the independent variable*. These are derived from the ANOVA table.
- *The power of the test of the null hypothesis* and the relationship between power and the decision about sample size.
- *The assumptions underlying the ANOVA, measures of importance, and power of the significance test*, including the consequences of violations of assumptions and alternative methods that can be used in the face of violations.

8.2 AN EXAMPLE OF THE DESIGN

An example of an experiment, together with a data set, will make subsequent developments more concrete. Table 8.1 presents data from a hypothetical memory study in which 40 subjects were randomly divided into four groups of 10 each. Each subject studied a list of 20 words and was tested for recall a day later. Ten subjects were taught and instructed to use a memory strategy called the method of loci, in which each object on the list was associated with a location on campus; 10 subjects were told to form an image of each object on the list; 10 others were told to form a rhyme with each word; and 10 others—the control group—were just told to study the words.[2]

Fig. 8.1 presents the group means and 95% confidence intervals for those means. The three groups that were instructed to use a memory strategy had higher average recall scores than the control group, although the widths of the confidence intervals indicate that the data were quite variable. There is also some indication that the method of loci may be superior to the other two

[2] The data are also in a file labeled *Table 8_1 Memory Data*; a link is on the *Tables* page on the website for this book.

Table 8.1 Recall scores from a hypothetical memory study

	Control	Loci	Image	Rhyme	
	11	10	13	16	
	4	18	16	9	
	8	6	3	7	
	3	20	6	10	
	11	15	13	9	
	8	9	10	14	
	2	8	13	16	
	5	11	9	3	
	8	12	5	9	
	5	12	19	12	
$\bar{Y}_{.j} =$	6.5	12.1	10.7	10.5	$\bar{Y}_{..} = 9.95$
$s_j^2 =$	10.056	19.433	25.567	16.722	

experimental methods. However, differences among the four means may just reflect differences in the effects of nuisance variables. By chance, the average ability or motivational level in one group of students may be higher than in the others; or other differences between individuals or between the conditions in which they were tested (e.g., the time of day, the temperature in the room) may account for the apparent differences among experimental conditions. A major goal of the data analysis is to separate out the effects of the instructional method from the effects of nuisance variables.

At this point, it would be wise to explore the data further, calculating additional statistics and plotting other graphs as described in Chapter 2. However, we will leave that as an exercise for the reader and proceed to address the question of whether the differences in Fig. 8.1 reflect true differences in the effects of the four study methods, or merely error variance. We begin by developing a framework for the analysis of variance.

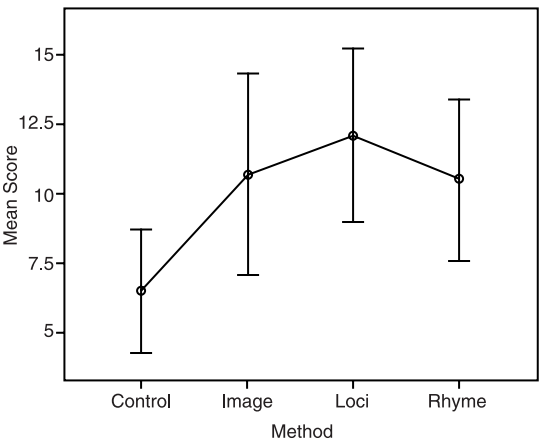

Fig. 8.1 Means and confidence interval bars for the data of Table 8.1.

8.3 THE STRUCTURAL MODEL

We view the various groups of scores in a study as random samples from populations selected for investigation. Then the question of whether the four study methods of Table 8.1 differ in their effectiveness can be rephrased as: Do the means of the four treatment populations differ? To answer this question, we need a way of linking the observed data to the hypothetical populations, of relating sample statistics to population parameters. We begin by constructing a *structural model*, a model that relates the scores to population parameters.

We start by assuming that the subjects in the experiment are identical in ability, motivation, and any other characteristics that would affect their scores. We further assume that they are identically treated—e.g., tested at the same moment in time, and under the exact same conditions. Under these very unrealistic assumptions, everyone in an instructional population, and therefore everyone in an instructional group, would have the same score; that score would be the treatment population mean. We can represent this state of affairs with the following notation:

$$Y_{11} = Y_{21} = Y_{31} = \ldots = Y_{i1} = \ldots = Y_{n1} = \mu_1$$
$$Y_{12} = Y_{22} = Y_{32} = \ldots = Y_{i2} = \ldots = Y_{n2} = \mu_2$$
$$Y_{13} = Y_{23} = Y_{33} = \ldots = Y_{i3} = \ldots = Y_{n3} = \mu_3$$

where there are n subjects in a group, Y_{ij} represents the i^{th} score in the j^{th} group, and μ_j is the mean of the *jth* population of scores. For example, Y_{52} would refer to the score of the fifth subject in the second group.

Of course, this is not realistic; the scores of individuals in an instructional group will vary, and therefore will differ from the instructional population mean, because of nuisance variables such as ability level, prior relevant experience, interest in the topic, or conditions at the time of testing. We can represent this complication by saying that the score of the i^{th} subject in the j^{th} group will differ from the treatment population mean, μ_j, by some amount—an *error component,* ε_{ij}. This means that an individual's score equals the mean of the treatment population plus an error component. That is,

$$\begin{aligned} Y_{ij} &= \mu_j + (Y_{ij} - \mu_j) \\ &= \mu_j + \varepsilon_{ij} \end{aligned} \tag{8.1}$$

Note that ε_{ij} can be positive or negative; that is, nuisance variables can raise the score above the population mean, or lower it below that mean.

We can rewrite Equation 8.1 in a way that more directly expresses the relation between a score and the effect of the condition under which that score was obtained. First, we define one more population parameter, μ, the mean of all the treatment populations; i.e., $\mu = \Sigma\mu_j/a$, where a is the number of levels of the independent variable. Equation 8.1 is unchanged if we add and subtract μ from the right side:

$$Y_{ij} = \mu + (\mu_j - \mu) + \varepsilon_{ij} \tag{8.2}$$

Let $\alpha_j = (\mu_j - \mu)$; because α_j (Greek alpha) is the difference between the mean of the j^{th} treatment population and the grand mean of all the populations, it represents the effect of the j^{th} treatment on the scores in the j^{th} population. Therefore, we can rewrite Equation 8.2:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \tag{8.3}$$

Equation 8.3 is a *structural equation*; it defines the structure of a score obtained in a one-factor between-subjects experiment. In words, the structure of a score is

score = grand mean + treatment effect + error component

The parameters in Equation 8.3 are rather abstract and not very useful unless we tie them to statistics that we can calculate from our data. To do this, we need to estimate the population means, the treatment effects, and the errors. We have the following parameter estimates:[3]

Parameter	μ	μ_j	α_j	ε_{ij}
Estimate	$\overline{Y}_{..}$	$\overline{Y}_{.j}$	$\overline{Y}_{.j} - \overline{Y}_{..}$	$Y_{ij} - \overline{Y}_{.j}$

where Y_{ij} is the score of the i^{th} person in the j^{th} group, $\overline{Y}_{.j}$ is the mean of all the scores in the j^{th} group, and $\overline{Y}_{..}$ is the mean of all the scores in the data set. For example, in Table 8.1, Y_{23} is 16, the score of the second person in the image condition; $\overline{Y}_{.4}$ is 10.5, the mean of the rhyme condition; and $\overline{Y}_{..}$ is 9.95, the grand mean.

With the structural equation as a basis, we now can begin to calculate the terms we need in order to draw inferences from our data.

8.4 THE ANALYSIS OF VARIANCE (ANOVA)

The ANOVA involves partitioning the variability of all the scores into two components, or *sums of squares*. These in turn are divided by their degrees of freedom to form *mean squares*, estimates of population variances. The ratio of mean squares, the *F ratio*, provides a test of the hypothesis that all the treatments have the same effect. In what follows, we consider each of these aspects of the ANOVA.

8.4.1 Sums of Squares

As Equation 8.3 implies, scores can vary because of the effects of the independent variable and because of the effects of uncontrolled nuisance variables. If we can separate those two sources of variance in our data, we will have the basis for deciding how much, if any, of the variance is due to the independent variable.

The structural equation suggests an approach to partitioning variability. If we rewrite Equation 8.3 by subtracting μ from both sides, we can express the deviation of a score from the grand mean of the population as consisting of a treatment effect and error; that is,

$$Y_{ij} - \mu = \alpha_j + \varepsilon_{ij}$$

Replacing the parameters in the preceding equation by the estimates we presented earlier, we have

$$Y_{ij} - \overline{Y}_{..} = (\overline{Y}_{.j} - \overline{Y}_{..}) + (Y_{ij} - \overline{Y}_{.j}) \tag{8.4}$$

The next step in partitioning the total variability is to calculate the terms in Equation 8.4, and then square and sum them. The results are the sums of squares. For the data set of Table 8.1, the left side of Equation 8.4 leads to the *total sum of squares*:

$$SS_{total} = \sum_{j=1}^{4} \sum_{i=1}^{10} (Y_{ij} - \overline{Y}_{..})^2 = (11 - 9.95)^2 + \ldots + (12 - 9.95)^2 = 819.9$$

[3] These are *least-squares estimators*; that is, we can show that if these statistics are calculated for many samples, their variance about the parameter being estimated is less than that for any other estimator.

The first term to the right of the equal sign in Equation 8.4 is also squared and summed for each individual, yielding the *method sum of squares*:

$$SS_{method} = 10 \sum_{j=1}^{4} (\bar{Y}_{.j} - \bar{Y}_{..})^2 = 10[(6.5 - 9.995)^2 + \ldots + (10.5 - 9.95)^2] = 173.9$$

and finally we obtain the *residual sum of squares* which can be calculated either directly as

$$SS_{residual} = \sum_{j=1}^{4} \sum_{i=1}^{10} (Y_{ij} - \bar{Y}_{.j})^2 = (11 - 6.5)^2 + \ldots + (12 - 10.5)^2 = 646.0$$

or as the difference between the total and method sum of squares:

$$SS_{residual} = SS_{total} - SS_{method} = 819.9 - 173.9 = 646.0$$

The preceding results are based on the example of Table 8.1. In general, we designate the independent variable by the letter A, and we assume a levels of A with n scores at each level. Then,

$$\sum_{j=1}^{a} \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_{..})^2 = n \sum_{j=1}^{a} (\bar{Y}_{.j} - \bar{Y}_{..})^2 + \sum_{j=1}^{a} \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_{.j})^2 \tag{8.5}$$

$$SS_{total} \qquad = SS_A \qquad + SS_{S/A}$$

where S/A represents "subjects within levels of A" to remind us that the residual term reflects the variability of the scores within each level of A. A general proof that $SS_{total} = SS_A + SS_{S/A}$ is presented in Appendix 8.1.

8.4.2 Degrees of Freedom (*df*)

The three terms in Equation 8.5 are numerators of variances and, as such, must be divided by their corresponding degrees of freedom (*df*) in order to be converted into variances, or mean squares. The *df* associated with a particular *SS* term is the number of independent observations contributing to that estimate of variability. For our three *SS* terms, we have the following *df*:

1. *The total degrees of freedom, df_{total}.* The *total sum of squares, SS_{total},* is the numerator of the variance of all *an* scores. Therefore, $df_{total} = an - 1$.
2. *The between-groups degrees of freedom, df_A,* scores. The *between-groups sum of squares, SS_A,* is n times the numerator of the variance of the a group means about the grand mean and is therefore distributed on $a - 1$ *df*.
3. *The within-groups degrees of freedom, $df_{S/A}$.* The *within-groups sum of squares, $SS_{S/A}$,* is the sum, or "pool" of the numerators of each of the group variances. Because each of the a group variances is distributed on $n - 1$ *df, $SS_{S/A}$* is distributed on $a(n - 1)$ *df*. Note that

$$an - 1 = a(n - 1) + (a - 1)$$
$$df_{tot} = df_{S/A} + df_A \tag{8.6}$$

Equation 8.6 demonstrates that the degrees of freedom are partitioned into two parts that correspond to the sums of squares. This partitioning of the degrees of freedom provides a partial check on the partitioning of the total variability. Although the partitioning in a one-factor design is simple, keeping track of the number of distinguishable sources of variance can be difficult in more

complex designs. There are also designs in which it is a challenge to analyze the relations among the factors in the design. Therefore, when designs have many factors, it is wise to find the degrees of freedom associated with each source of variability and to check whether the *df* sum to the total number of scores minus one.

8.4.3 Mean Squares, Expected Mean Squares, and the *F* Ratio

The ratio of a sum of squares to degrees of freedom is called a *mean square*. In the one-factor design, the relevant mean squares are the A mean square, where

$$MS_A = SS_A/df_A$$

and the S/A mean square,

$$MS_{S/A} = SS_{S/A}/df_{S/A}$$

Under the assumptions summarized in Box 8.1, the ratio $MS_A/MS_{S/A}$ has a sampling distribution

Box 8.1 Parameter Definitions and Assumptions

1. *The parent population mean,* μ. This is the grand mean of the treatment populations selected for this study and is a constant component of all scores in the a populations. It is the average of the treatment population means:

$$\mu = \sum_{j=1}^{a} \mu_j / a$$

2. The effect of treatment A_j, α_j. This equals $\mu_j - \mu$ and is a constant component of all scores obtained under A_j but may vary over treatments (levels of j).
- 2.1 Because the deviation of all scores about their mean is zero, $\Sigma_j \alpha_j = 0$.
- 2.2 If the null hypothesis is true, all $\alpha_j = 0$.
- 2.3 The population variance of the treatment effects is $\sigma_A^2 = \sum_{j=1}^{a} \alpha_j^2 / a$.

3. *The error,* ε_{ij}. This is the deviation of the i^{th} score in group j from μ_j and reflects uncontrolled, or chance, variability. It is the only source of variation within the j^{th} group, and if the null hypothesis is true, the only source of variation within the data set. We assume that
- 3.1 The ε_{ij} are independently distributed; i.e., the probability of sampling some value of ε_{ij} does not depend on other values of ε_{ij} in the sample.
- 3.2 The ε_{ij} are normally distributed in each of the a treatment populations. Also, because $\varepsilon_{ij} = Y_{ij} - \mu_j$, the mean of each population of errors is zero; i.e., $E(\varepsilon_{ij}) = 0$.
- 3.3 The distribution of the ε_{ij} has variance σ_e^2 (error variance) in each of the a treatment populations; i.e., $\sigma_1^2 = \ldots = \sigma_j^2 = \ldots = \sigma_a^2$. This is the assumption of *homogeneity of variance*. The error variance is the average squared error; $\sigma_e^2 = E(\varepsilon_{ij}^2)$.

called the F distribution *if the null hypothesis is true*. It provides a test of the null hypothesis that the treatment population means are all equal; that is, the F statistic tests the *null hypothesis*:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_a = \mu$$

or, equivalently,

$$H_0\colon \alpha_1 = \alpha_2 = \ldots = \alpha_j = \ldots \alpha_a = 0$$

To understand the logic of the *F* test, we need to consider the relationship of the mean squares to the population variances. This requires us to determine the expected values of our two mean square calculations.

Suppose we draw a samples of n scores from their respective treatment populations, and calculate MS_A and $MS_{S/A}$. Now suppose that we draw another a samples of n scores, and again calculate MS_A and $MS_{S/A}$. We could repeat this sampling experiment many times and generate two sampling distributions, one for MS_A and another for $MS_{S/A}$. The means of these two sampling distributions are the expected values of the mean squares, or the *expected mean squares* (*EMS*). Given the structural model of Equation 8.3, and assuming that the ε_{ij} are independently distributed with variance, σ_e^2, the *EM*S of Table 8.2 can be derived (Kirk, 1995; Myers & Well, 1995). Consider each expected mean square in turn to understand the information provided by MS_A and $MS_{S/A}$.

Table 8.2 Sources of variance (*SV*) and expected mean squares (*EMS*) for the one-factor between-subjects design

SV	*EMS*
A	$\sigma_e^2 + n\theta_A^2$
S/A	σ_e^2

Note: $\theta_A^2 = \sum_j (\mu_j - \mu)^2/(a-1)$. We use the θ^2 (theta squared) notation rather than σ^2 to remind us that the treatment component of the *EMS* involves division by degrees of freedom; the variance of the treatment population means would be $\sigma_A^2 = \sum_j (\mu_j - \mu)^2/a$.

$E(MS_A)$ states that the *between-groups mean square,* MS_A, estimates *error variance*, σ_e^2, plus n times the variance in the treatment population means, θ_A^2 (if there is any effect of the treatment). This result should make intuitive sense when you examine the formula for MS_A:

$$MS_A = \frac{n\sum_j^a (\bar{Y}_{.j} - \bar{Y}_{..})^2}{a-1} \tag{8.7}$$

Equation 8.7 states that MS_A is the variance of the condition means times the sample size, n. Even if there were no differences among the treatment population means, the sample means would differ just by chance because there are different individuals with different characteristics in each group. The error variance, σ_e^2, reflects this. If there is also an effect of the treatment, the μ_j will differ and their variability will also be reflected in the value of MS_A.

$E(MS_{S/A})$ states that the *within-groups mean square,* $MS_{S/A}$, is an estimate of error variance. Again, this result may be understood intuitively by examining how $MS_{S/A}$ is calculated:

$$MS_{S/A} = \frac{SS_{S/A}}{a(n-1)} = \frac{\sum_j \sum_i (Y_{ij} - \bar{Y}_{.j})^2}{a(n-1)} \tag{8.8}$$

Equation 8.8 may be rewritten as

$$MS_{S/A} = \left(\frac{1}{a}\right) \sum_j \left[\frac{\sum_i (Y_{ij} - \bar{Y}_{.j})^2}{n-1} \right]$$

The expression in the square brackets on the right side is the variance of the j^{th} group of scores, and the entire right side is an average of the a group variances. Because subjects within a condition are treated identically, they should differ only due to error (see Eq. 8.1). If we assume that error variance is equal in each treatment population, $MS_{S/A}$ is an average of a estimates of the population variance, σ_e^2.

Given our understanding of what MS_A and $MS_{S/A}$ estimate, we are in a position to understand the logic of the F test, where $F = MS_A/MS_{S/A}$. First, *assume that the null hypothesis is true* and, also, that there is *homogeneity of variance*; that is

$$\mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_a \text{ and } \sigma_1 = \sigma_2 = \ldots = \sigma_j = \ldots = \sigma_a$$

Under these assumptions, MS_A is an estimate of the error variance common to the a treatment populations. In terms of $E(MS_A)$, $\theta_A^2 = 0$ so $E(MS_A)$ is an estimate of error variance, σ_e^2. Thus, if the null hypothesis is true, MS_A and $MS_{S/A}$ both estimate the same population error variance and their ratio should be about 1. Of course, it would be surprising if two independent estimates of the same population variance were identical; that is, the ratio of MS_A to $MS_{S/A}$ has a distribution of values. More precisely, *if H_0 is true*, the ratio, $MS_A/MS_{S/A}$, is distributed as F on $a - 1$ and $a(n-1)$ *df*. Critical values of F are tabled in Appendix Table C.5 and can also be obtained from various software packages and websites.

But what if the null hypothesis is, in fact, false? For example, suppose that the method of study does affect recall in the example of Table 8.1. Then the means of the groups of scores in Table 8.1 will differ not only because the scores in the different groups differ by chance, but also because the groups were studied by different methods. In other words, if H_0 is false, $\theta_A^2 > 0$ so $E(MS_A) = \sigma_e^2 + n\,\theta_A^2$. The situation with respect to the within-group variance does not change: $MS_{S/A}$ should not be affected by the independent variable because all subjects in a group receive the same treatment. Therefore, when H_0 is false, the ratio $MS_A/MS_{S/A}$ should be greater than 1.

In summary, under the assumptions of the null hypothesis, homogeneity of variance, and independently distributed scores, MS_A and $MS_{S/A}$ are two independent estimates of the population error variance, σ_e^2. If we also assume that the population of scores is normally distributed, the ratio of two independent estimates of the same population variance has an F distribution. Therefore, under the assumptions summarized in Box 8.1, the ratio $MS_A/MS_{S/A}$ is distributed as F. Because the numerator is based on an estimate of the variance of a population means, it has $a - 1$ *df*. The denominator has $a(n-1)$ *df* because the variance estimate for each group is based on $n - 1$ *df* and $MS_{S/A}$ is an average of a variance estimates.

Appendix Table C.5 presents critical values of the F distribution. As an example of its use, suppose we have three groups of 11 subjects each. Then the numerator $df = a - 1 = 2$, and the

denominator $df = a(n - 1) = 30$. Turning to the column headed by 2 and the block labeled 30, if $\alpha = .05$, we would reject the null hypothesis of no difference among the three treatments if the F we calculate is greater than 3.32. Interpolation may be needed for degrees of freedom not listed in the table. However, the critical F value is not necessary if the analysis is performed by any of several software packages. These packages usually calculate the F based on the data and provide the exact p-value for that F and the df for the design used.

8.4.4 The ANOVA Table

Panel *a* of Table 8.3 summarizes the developments so far, presenting the formulas for sums of squares, degrees of freedom, mean squares, and the F ratio for the one-factor between-subjects design. For any data set, most statistical software packages present this table with numerical results in some form. Panel *b* presents the output for the data of Table 8.1. The results are significant at the .05 level, indicating that there are differences among the means of the populations defined by the four different study methods. Fig. 8.1 suggests that this is due to the poorer performance of the control condition. However, there are a number of interesting questions that the omnibus F test leaves unanswered. Are all three experimental methods significantly superior to the control method? Do the means of the three experimental methods differ from each other? We will consider such comparisons of means within subsets of conditions in Chapter 10.

We might also ask whether the differences among the four population means are large enough to be of practical significance. As we noted when discussing effect size measures in Chapter 6, statistical significance is not the same as practical or theoretical significance.

Table 8.3 The analysis of variance for the one-factor between-subjects design

(a) General form of the ANOVA

Source	*df*	*SS*	*MS*	*F*
Total	$an - 1$	$\sum_{j=1}^{a}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{..})^2$		
A	$a - 1$	$n\sum_{j=1}^{a}(\bar{Y}_{.j} - \bar{Y}_{..})^2$	SS_A/df_A	$MS_A/MS_{S/A}$
S/A	$a(n - 1)$	$\sum_{j=1}^{a}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_{.j})^2$	$SS_{S/A}/df_{S/A}$	

(b) ANOVA of the data of Table 8.1

Source	Sum of squares	*df*	Mean square	*F*	*p*-value
Method	173.90	3	57.967	3.230	.034
Error	646.00	36	17.944		
Total	819.90	39			

8.5 MEASURES OF IMPORTANCE

The *p*-value in Table 8.3 informs us that the effects of the method of memorization are statistically significant, assuming that we had set alpha at .05. In addition, however, we need some indication of the practical or theoretical importance of our result. Generally, we seek a measure that assesses the magnitude of the effect of our treatment, *A*, relative to error variance. We will find that the *EMS* analyses that guided the logic of the *F* test will also be very useful in thinking about appropriate ways in which to assess the importance of an effect. We will consider several possible measures in this section. Several sources also present discussions of these and other measures (e.g., Kirk, 1996; Maxwell, Camp, & Arvey, 1981; Olejnik & Algina, 2000, 2003).

8.5.1 Measuring Strength of Association in the Sample: η^2 (Eta-Squared)

Greek-letter designations are usually reserved for population parameters, but η^2 is actually a sample statistic that is often used as a measure of association between the dependent and independent variables (Cohen, Cohen, West, & Aiken, 2003). It describes the proportion of variability in the sample as

$$\eta^2 = \frac{SS_A}{SS_{total}} \tag{8.9}$$

Referring to Table 8.3, we have

$$\eta^2_{method} = 173.9/819.9 = .212$$

Using SPSS, η^2 can be obtained by selecting *Analyze*, then *General Linear Model*, then *Univariate*. SPSS reports the R^2, which for the one-factor design is also the same as η^2. It also reports an *adjusted* R^2 as .146.[4]

Eta-squared has the advantage of being easily calculated and easily understood as a proportion of sample variability. However, the value of η^2 is influenced not only by the relative magnitudes of the treatment effect and error variance, but also by *n*, df_A, and $df_{S/A}$. In addition, σ_e^2 contributes to the numerator of η^2. For these reasons, other statistics that measure importance are often preferred. We turn to such estimates now, bearing in mind that our results rest on the assumptions underlying the derivation of the *EMS*; i.e., independence of the ε_{ij} and homogeneity of variance.

8.5.2 Estimating Strength of Association in the Population: ω^2 (Omega-Squared)

Whereas η^2 describes the strength of association between the dependent and independent variables by forming a ratio of sample sums of squares, ω^2 is a measure of the strength of association in the population; unlike η^2, it is a ratio of population variances:

$$\omega^2 = \frac{\sigma_A^2}{\sigma_e^2 + \sigma_A^2} \tag{8.10}$$

The numerator of the ratio is the variance of the treatment population means (the μ_j) or, equivalently, the variance of the treatment effects (the α_j):

[4] The adjusted $R^2 = [SS_A - (a-1)MS_{S/A}]/SS_{total}$

$$\sigma_A^2 = \frac{\sum_j^a (\mu_j - \mu)^2}{a} = \frac{\sum_j^a \alpha_j^2}{a} \tag{8.11}$$

The denominator of ω^2 is the total population variance; that is, the treatment population error variance, σ_e^2, plus the variance of the treatment population means, σ_A^2. Thus, ω^2 assesses the magnitude of the treatment effect relative to the total variance in the design. We cannot know the ratio described by Equation 8.10 but we can derive estimates of σ_A^2 and σ_e^2 and therefore of ω^2. We begin with the *EMS* equations of Table 8.2:

$$E(MS_A) = \sigma_e^2 + n\theta_A^2 \tag{8.12}$$

and

$$E(MS_{S/A}) = \sigma_e^2 \tag{8.13}$$

To obtain an estimate of σ_A^2 we first subtract Equation 8.13 from Equation 8.12, and divide by n; then we have

$$\frac{MS_A - MS_{S/A}}{n} = \hat{\theta}_A^2$$

where the "hat" above θ_A^2 means "is an estimate of." Because the numerator of ω^2 as defined by Equation 8.10 involves σ_A^2, not θ_A^2, and noting that $\sigma_A^2 = [(a-1)/a] \times \theta_A^2$, our estimate of σ_A^2 is

$$\hat{\sigma}_A^2 = \left(\frac{a-1}{a}\right)\left(\frac{MS_A - MS_{S/A}}{n}\right) \tag{8.14}$$

We now have estimates of the numerator and denominator of ω^2, therefore, substituting into Equation 8.10, we have an estimate of ω^2 for the one-factor, between-subjects design:

$$\hat{\omega}^2 = \frac{[(a-1)/a](1/n)(MS_A - MS_{S/A})}{[(a-1)/a](1/n)(MS_A - MS_{S/A}) + MS_{S/A}} \tag{8.15}$$

We may write Equation 8.15 in a different form, one which allows us to calculate $\hat{\omega}^2$ from knowledge of the F ratio, a, and n. The advantages are that the expression is somewhat simpler and, perhaps more importantly, because most research reports contain this information, we can estimate the strength of association for data collected by other investigators. We begin by defining $F_A = MS_A/MS_{S/A}$. Then, multiplying the numerator and denominator of Equation 8.15 by an, and dividing by $MS_{S/A}$, we have

$$\hat{\omega}^2 = \frac{(a-1)(F_A - 1)}{(a-1)(F_A - 1) + na} \tag{8.16}$$

Let's review what Equations 8.15 and 8.16 represent. If we replicate the experiment many times, the average value of the right-hand term will approximately equal ω^2, the proportion of the total

variance in the a treatment populations that is attributable to the variance of their means. We say "approximately equal" because the expected value of a ratio is not the same as the ratio of expected values. The approximation is reasonably accurate and the expression is much simpler than that for the exact expression.

One other aspect of Equation 8.16 should be noted. Because the numerator and denominator of the F reflect two independent estimates of the population error variance, when the null hypothesis is true or the effects of A are very small, the F may be less than 1. Then, $\hat{\omega}^2$ would be less than 0. Because a variance cannot be negative, we conclude that $\omega^2 = 0$; that is, none of the total population variance is attributable to the independent variable.

We can apply Equation 8.16 to the memory data in Table 8.1. In that experiment, $a = 4$, $n = 10$, and (from Table 8.3) $F = 3.230$. Then, inserting these values into Equation 8.16,

$$\hat{\omega}^2 = \frac{(3)(2.23)}{(3)(2.23) + 40} = .143$$

This is very close to the value of .146 noted earlier for adjusted R^2. That the values of R^2_{adj} and ω^2 are so close is not unusual; Maxwell, Camp, and Arvey (1981) reported that the two rarely differ by more than .02. With respect to assessing the importance of either measure, Cohen (1988) suggested that values of .01, .06, and .14 may be viewed as small, medium, and large, respectively. According to those guidelines, the proportion of variability accounted for by the study method may be judged to be large. Again, however, we caution that the importance attached to any value must be assessed in the context of the research problem and the investigator's knowledge of the research literature.

8.5.3 Cohen's *f*

In Chapter 6, we presented Cohen's d, a measure of the standardized effect for designs in which two means are compared. Cohen's f (1988) is a similar measure for situations in which the variance of more than two means is of interest. The parameter f is defined as

$$f = \sigma_A/\sigma_e \tag{8.17}$$

We can estimate f by substituting the estimate in Equation 8.14 in the numerator and $MS_{S/A}$ in the denominator. Then we have

$$\hat{f} = \sqrt{\frac{(a-1)(MS_A - MS_{S/A})}{anMS_{S/A}}} \tag{8.18}$$

which can also be written as

$$\hat{f} = \sqrt{(a-1)(F_A - 1)/an} \tag{8.19}$$

For the data of Table 8.1, substituting the F value from Table 8.3 into Equation 8.19, we have

$$\hat{f} = \sqrt{(3)(2.23)/40} = .409$$

Cohen has suggested that values of f of .1, .25, and .4 be viewed as small, medium, and large, respectively. Therefore, as with ω^2, the guidelines for f suggest that the standardized variance of the four reading method estimates is large. That ω^2 and f lead to the same conclusion about the size of the variance of effects follows from the relationship between them; given an estimate of f, we can also calculate an estimate of ω^2, and vice versa. The relations are

$$f^2 = \frac{\omega^2}{1-\omega^2} \quad \text{and} \quad \omega^2 = \frac{f^2}{1+f^2}$$

A useful property of f is that it is closely related to the noncentrality parameter of the F distribution; specifically,

$$\lambda = N f^2 \tag{8.20}$$

The parameter, λ (lambda), determines areas under the noncentral F distribution, and therefore the power of the F test. Smithson (2001) provides an SPSS syntax file for obtaining a confidence interval on λ, and by making use of its relation to f and ω^2, confidence intervals on those measures can be obtained. The relation between f, λ, and power will be developed further in Section 8.8.

8.5.4 Measures of Importance: Limitations

In an introductory chapter to an edited collection aptly titled, "What if there were no significance tests?", Harlow (1997, pp. 5–6) reported that 11 of the book's other 13 chapters "were very much in favor" of reporting measures such as R^2, ω^2, and f, and the remaining two contributors "at least mildly endorsed such use." Similar support for measures such as these can be found in the American Psychological Association's guidelines for statistical usage (Wilkinson & Task Force, 1999), which urge researchers to report effect size statistics. Nevertheless, there are potential pitfalls. Values of these statistics may depend on the experimental design, the choice and number of levels of the independent variable, the dependent variable, and the population sampled. Estimates of ω^2 and f imply homogeneous variances and independence of observations (cf. Grissom & Kim, 2001, who discuss the variance assumption, and suggest alternative approaches for the two-group case). Another concern is that squared coefficients tend to be small and it is sometimes easy to dismiss an effect as trivial because of a small value of ω^2.

These arguments suggest that we must be careful when interpreting these measures, or when generalizing the results of any one study, or of making comparisons across studies that differ with respect to the factors just cited. In addition, we should treat guidelines such as those set forth by Cohen (1988) as suggestions, not as definitive boundaries between important and unimportant effects. Even a very small advantage of one therapy over another may be important. In theoretical work, a small effect predicted by a theory may be important support for that theory. In summary, if care is taken in interpreting measures of strength, statistics such as $\hat{f}$ and $\hat{\omega}^2$ are useful additions to the test statistics usually computed.

8.6 WHEN GROUP SIZES ARE NOT EQUAL

In the developments so far, our formulas have been based on the assumption that there are the same number of scores in each group. In this section, we present an example with unequal ns, and present formulas for sums of squares, expected mean squares, and measures of importance for this case.

The ns in conditions in a study may vary for one of several reasons. The populations may be equal in size but data may be lost from some conditions, perhaps because of a malfunction of equipment, or a subject's failure to complete the data-collection session. Usually, individuals can be replaced but sometimes this is impossible. In other instances, the treatments may affect the availability of scores; for example, animals in one drug condition may be less likely to survive the experiment than animals in another condition. In still other instances, usually when we collect data

from existing populations, conditions may differ naturally in the availability of individuals for participation. For example, in clinical settings, there may be different numbers of individuals in different diagnostic categories.

Unequal n complicates calculations in the one-factor design, which might tempt some researchers to discard scores from some conditions to equate n. This is not a good idea for a couple of reasons. Discarding subjects to equalize the group ns will reduce error degrees of freedom and, consequently, power. Discarding subjects also may misrepresent the relative size of the populations sampled. If so, the effects of some conditions may be weighted too heavily or too lightly in the data analysis. Finally, computational ease should not be a consideration when software programs are available to handle the calculations.

8.6.1 The ANOVA with Unequal *n*

The ANOVA for unequal group sizes is a straightforward modification of the equal-n case, at least in the one-factor between-subjects design. (Complications arise when more than one factor is involved; these will be treated in Chapters 9 and 24.) Table 8.4 presents the ANOVA formulas and expected mean squares for the unequal-n case; the squared deviations in the SS_A formula and the n_j are weighted by the group size. Note that if the n_j are equal, these formulas reduce to the formulas in Table 8.3.

Table 8.4 The analysis of variance for the one-factor between-subjects design with unequal group sizes

Source	df	*SS*	*MS*	*F*	*EMS*
A	$a-1$	$\sum_{j=1}^{a} n_j(\overline{Y}_{.j}-\overline{Y}_{..})^2$	SS_A/df_A	$MS_A/MS_{S/A}$	$\sigma_e^2+\frac{1}{a-1}\sum_j n_j\alpha_j$
S/A	$N-a$	$\sum_{j=1}^{a}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{.j})^2$	$SS_{S/A}/df_{S/A}$		σ_e^2
Total	$N-1$	$\sum_{j=1}^{a}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_{..})^2$			

Note: n_j is the number of scores in the j^{th} group and $N=\sum_{j=1}^{a} n_j$.

Table 8.5 presents statistics based on Beck Depression scores for four groups of males who participated in the University of Massachusetts Medical School research on seasonal effects; the statistics are based on scores averaged over the four seasons. For the purposes of this example, we excluded some subjects (those having no or only some high-school education, and those with vocational training or an associate's degree). The remaining groups are HS (high-school diploma only), C (some college), B (bachelor's degree), and GS (graduate school).[5]

The statistics of Table 8.5 and the box plots of Fig. 8.2 indicate that the groups differ in their average depression score. Both means and medians are noticeably higher for those subjects who had only a high-school education; subjects with a graduate school education have lower scores but they

[5] The data may be found in the *Table 8_5 Male_educ* file; go to the *Tables* page of the book's website.

Table 8.5 Summary statistics for Beck Depression scores in four educational levels (the data are in the *Male-educ* file; go to the *Seasons* page on the book's website)

	Level of education			
	HS	C	B	GS
No. of cases	19	33	37	39
Median	6.272	2.875	2.265	3.031
Mean	6.903	3.674	3.331	4.847
Variance	34.541	5.970	9.861	26.218
Skewness ($g1$)	.824	.368	2.047	1.270
Kurtosis ($g2$)	.168	−.919	5.837	.745

Note: HS = high-school diploma only; C = some college; B = bachelor's degree; GS = graduate school.

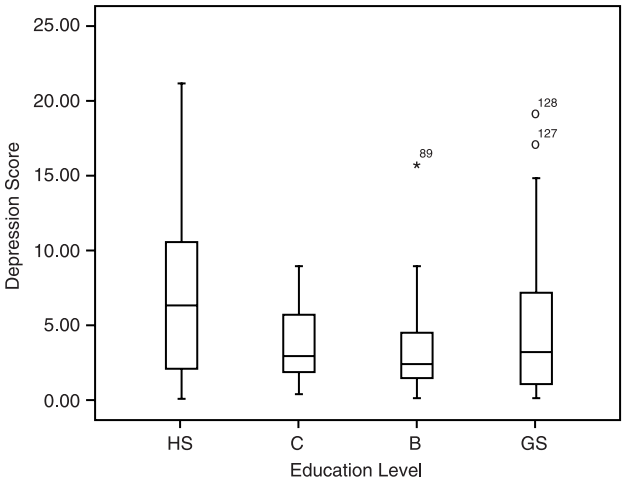

Fig. 8.2 Box plot of Beck Depression scores as a function of educational level.

are higher than those in the remaining two categories. Variances are also highest in the *HS* and *GS* groups; the differences among the variances as well as among the *H*-spreads in the figure warn us that heterogeneity of variance may be an issue. We also note that both the skew statistics and the long tails at the high end of depression scores in the figure indicate that the populations are unlikely to be normally distributed. A *Q*–*Q* plot (see Chapter 2) would confirm this.

Applying the formulas in Table 8.4, we obtain the ANOVA results in Table 8.6; these reveal that the means of the four groups differ significantly. However, the characteristics of the data revealed by

Table 8.6 ANOVA of the depression means in Table 8.5

Source	*SS*	*df*	*MS*	*F*	*p*
Education	186.501	3	62.167	3.562	.016
Error	2,164.061	124	17.452		
Total	2,350.562	127			

our preliminary exploration (Table 8.5, Fig. 8.2) indicate that the assumptions of the analysis of variance are violated. In Section 8.8, we discuss those assumptions, consider alternative analyses that respond to violations of the assumptions, and apply one such analysis to the depression scores.

8.6.2 Measures of Importance with Unequal *n*

As in the equal *n* design, $\eta^2 = SS_A/(SS_A + SS_{S/A})$. For the Beck Depression data, substituting values from Table 8.6, $\eta^2 = 186.501/2{,}350.562$, or .079. The formulas for ω^2 and Cohen's *f* undergo a very slight modification. We replace the *n* in Equation 8.14 by the average *n*, $\bar{n}$, where $\bar{n} = \sum_j n_j/a = N/a$. We can simplify things further by replacing $a\bar{n}$ by *N*, the total sample size. Then the equations estimating, σ_A^2, ω^2, and Cohen's *f* apply with no further changes.

To estimate the population variance of the Beck Depression means as a function of educational level, substitute the mean squares from Table 8.6 into Equation 8.14, and with $\sum_j n_j = 128$,

$$\hat{\sigma}_A^2 = (3)(62.167 - 17.452)/128 = 1.048$$

We now can estimate ω^2 using Equation 8.16 with *N* replacing *an*. Then

$$\hat{\omega}^2 = \frac{(a-1)(F_A - 1)}{(a-1)(F_A - 1) + N}$$

$$= \frac{(3)(2.562)}{(3)(2.562) + 128} = .057$$

We use the same variance estimate to estimate Cohen's *f*:

$$\hat{f} = \hat{\sigma}_A/\hat{\sigma}_e$$

$$= \sqrt{1.048/17.245} = 0.245$$

Whether we view ω^2 or *f*, Cohen's guidelines suggest that the effect is of medium size.

In leaving the analysis of the effects of educational level on Beck Depression scores, we again caution that our exploration of the data suggested that both the normality and homogeneity of variance assumptions were violated, making suspect the results of significance tests and estimates of effect size. We will return to the general issue of violations of assumptions, and possible remedies, in Section 8.8.

8.7 DECIDING ON SAMPLE SIZE: POWER ANALYSIS IN THE BETWEEN-SUBJECTS DESIGN

Together with the practical constraints imposed by the available access to subjects and time, statistical power should be a primary consideration in deciding on sample size. In order to incorporate this into our decision about sample size, we need to decide on a value of power and we need a value of the minimum effect size we wish to detect. As we saw in our treatment of *t* tests in Chapter 6, there are several ways that we might proceed.

One possibility is that we use Cohen's guidelines to establish an effect size. For example, suppose that in designing the memory experiment described in Section 8.2, we had decided that we want power equal to at least .90 to reject an effect that was large by Cohen's guidelines; then $f = .4$.

How many subjects should we have included in the study? An easy way to answer this question is to use G*Power 3, available on the book's website. Fig. 8.3 shows the screen involved in the calculations. We selected *F tests* from the *Test Family* menu, the ANOVA for the one-way design from the *Statistical test* menu, and the *a priori* option from the *Type of power analysis* menu. We set the *Input*

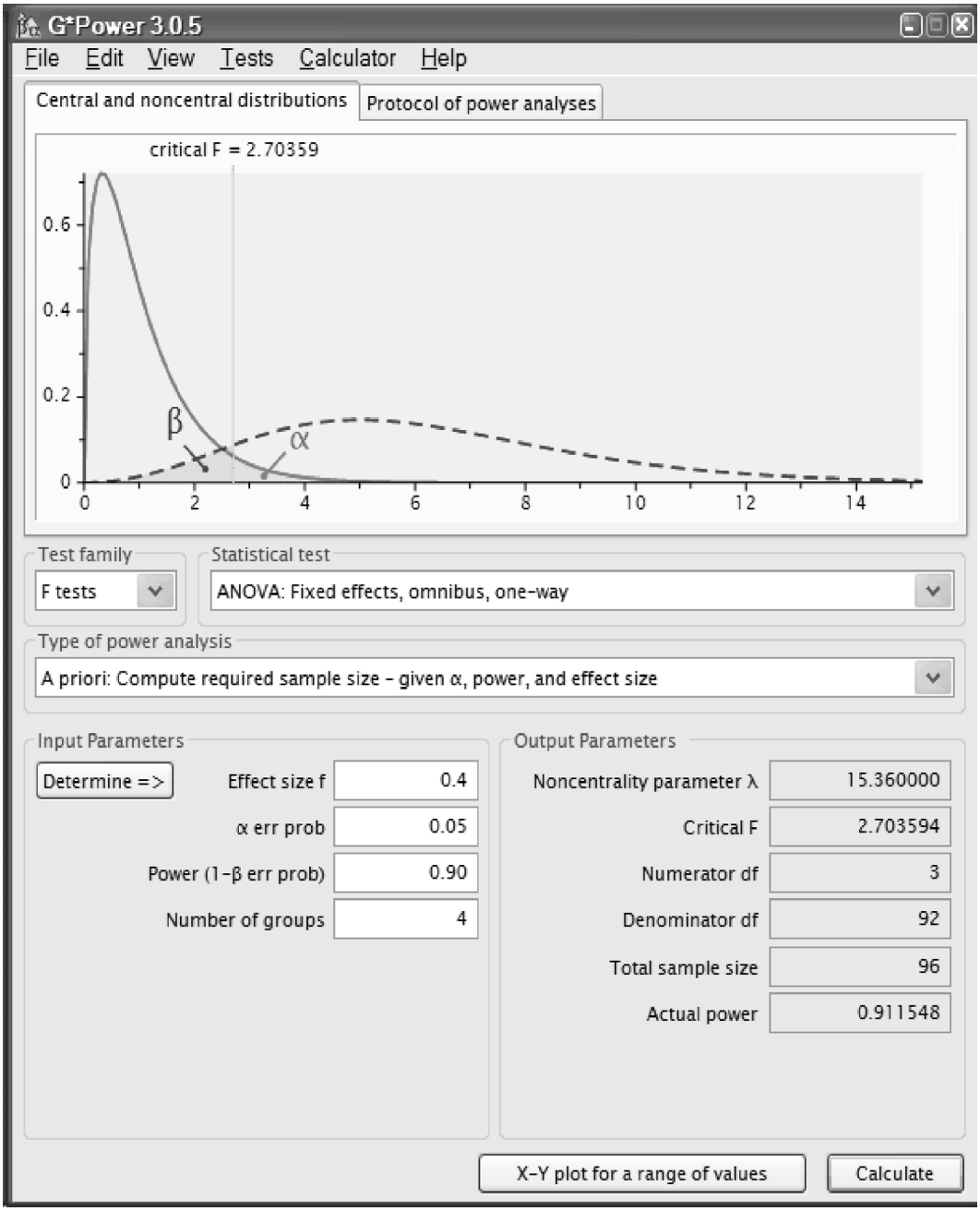

Fig. 8.3 G*Power 3 calculation of the *N* needed to have power = .90 when the effect size, *f*, equals .4.

Parameters as shown in the left-hand column. In the *Output Parameters* column, we find that the required total N is 96, or 24 in each group. The other output results should be self-explanatory except for the noncentrality parameter λ; $\lambda = N f^2$, or $96 \times .16$. This parameter is an index of the noncentral F distribution's distance from the central F distribution; when $\lambda = 0$, power equals the Type 1 error rate and as λ increases, so does the power of the test.

An alternative to using Cohen's guidelines is to base our assumed effect size on results from a pilot study. In this case, we would recommend using Equation 8.18 or 8.19 to obtain an estimate of f. We would then insert this into G*Power. For the data of Table 8.1, this estimate was .409. If we require power = .9, set $\alpha = .05$; then with four groups the required N is 92, slightly less than when we entered $f = .4$.

Finally, in some cases, we might have no single data set on which to base an estimate of f. However, practical or theoretical considerations, or a review of published results from several related studies, might suggest reasonable values of the treatment population means and of the population standard deviation. For example, in planning an experiment involving three groups, we might decide that the most likely values of the population means are 50, 60, and 70, and that the population standard deviation is about 20. Using G*Power 3, we enter $\alpha = .05$, power = .9, and the number of groups = 3, and then select the *determine* button. This brings up a panel in which we enter the hypothesized population means and the population standard deviation. Select "Calculate and transfer to main window." G*Power calculates

$$\sigma_A = \sqrt{[(50 - 60)^2 + (60 - 60)^2 + (70 - 60)^2]/3} = 8.165$$

and divides by the standard deviation of 20 to yield $f = .408$. Transferring this value to the main window, the required total N is 81, or 27 in each of the three groups.

Post hoc, or retrospective power, is also available in G*Power, as well as in SPSS; in the latter, *Observed power* is an option in the univariate analysis program. However, as we discussed in Chapter 6, we have reservations about reporting power based on the observed set of statistics. Confidence intervals on the raw and standardized effects will prove more informative and be less misleading.

8.8 ASSUMPTIONS UNDERLYING THE *F* TEST

In Chapter 6, we discussed the consequences of violations of assumptions of independence, normality, and homogeneity of variance for the t test, and considered several possible remedies, including transformations, t tests that take heterogeneous variances into account, t tests based on trimming the data set, and tests based on ranks. The F test in the one-factor, between-subjects design rests on the same assumptions, and in addition the data should be consistent with the underlying structural model. The consequences of violations of assumptions, as well as the remedies proposed, parallel those discussed in Chapter 6. In the following sections, we consider each assumption in turn, describing both the consequences of violations and possible remedies.

8.8.1 The Structural Model

The analysis of variance for the one-factor design begins with the structural model of Equation 8.3. This equation implies that only one factor systematically influences the data, and the residual variability ($MS_{S/A}$) represents random error. However, researchers sometimes ignore factors that have been manipulated but are not of interest in themselves. If those factors contribute significant variability, the one-factor model is not valid for the research design. Common examples arise when

half of the subjects are male and half are female, or when subject running is divided equally between two experimenters, or when the position of an object is counterbalanced in an experiment involving a choice. Although these variables may be irrelevant to the purpose of the research, they may affect the scores. If so, the $MS_{S/A}$ represents both error variance and variance due to gender, experimenter, or position. However, the variance due to these "irrelevant" variables will not contribute to MS_A. For example, if each method in our earlier example has an equal number of male and female subjects, sex of subject will not increase the *method* mean square. The analysis based on the one-factor model then violates the principle that the numerator and denominator of the F ratio should have the same expectation when H_0 is true. In such situations, the denominator has a larger expectation than the numerator because the irrelevant variable makes a contribution only to the denominator. The result is a loss of power, which can be considerable if the irrelevant variable has a large effect. We say that the F test is *negatively biased* in this case, meaning that the Type 1 error rate will be less than its nominal value if the null hypothesis is true. As a general rule, the researcher should formulate a complete structural model, one which incorporates all systematically varied factors, even those thought to be irrelevant or uninteresting. In the examples cited, this would mean viewing the study as involving two factors, the independent variable of interest and gender (or experimenter, or position), and carrying out the analysis described in Chapter 9.

8.8.2 The Independence Assumption

When only one observation is obtained from each subject, and subjects are randomly assigned to treatments or randomly sampled from distinct populations, the assumption that the scores are independently distributed is likely to be met. However, there are exceptions that sometimes are not recognized by researchers. For example, suppose we want to compare attitudes on some topic for males and females. Further suppose that before being tested, subjects participate in three-person discussions of the relevant topic. The scores of individuals who were part of the same discussion group will tend to be positively correlated. If this failure of the independence assumption is ignored (and it has been in some studies; see Anderson and Ager, 1978, for a review), there will be a *positive bias*—an inflation of Type 1 error rate—in an F test of the gender effect (Myers, DiCecco, & Lorch, 1981; Myers & Well, 1995). A class of analyses referred to as multilevel, or hierarchical (e.g., Raudenbush & Bryk, 2002), provides a general approach to this and other data analysis issues.

Another potential source of failure of the independence assumption is the "bottom-of-the-barrel" problem. Researchers at universities often feel that as the semester progresses, the performance of volunteer subjects in experiments tends to become poorer because less motivated subjects usually volunteer for research credit late in the semester. In this case, scores obtained close in time will tend to have higher correlations than those further apart in time.

8.8.3 The Normality Assumption

Violations of the normality assumption are relatively common and merit attention because they can reduce the power of the F test.

Consequences of Violating the Normality Assumption. As with the t test, the Type 1 error probability associated with the F test is little affected by sampling from non-normal populations unless the samples are quite small and the departure from normality extremely marked (e.g., Donaldson, 1968; Lindquist, 1953, pp. 78–90; Scheffé, 1959). This is true even when the independent variable is discretely distributed, as it is whenever rating data or response frequencies are analyzed. In all but the most skewed distributions, computer sampling studies indicate that Type 1

error rates are relatively unaffected when such measures are submitted to an analysis of variance (Bevan, Denton, & Myers, 1974; Hsu & Feldt, 1969; Lunney, 1970).

Although in most instances the Type 1 error rate is relatively unaffected by departures from normality, loss of power is a concern when distributions are long-tailed, skewed, or include outliers. In each of these situations, variability is high relative to the normal distribution. Thus, procedures that address the increased variability often have more power than the conventional *F* test. Several potential remedies are considered next.

Dealing with Violations of the Normality Assumption: Tests Based on Trimmed Data. As we explained in Chapter 6, merely deleting scores is not a valid procedure. However, a ratio of mean squares distributed approximately as *F* can be constructed by an approach similar to the trimmed *t* test illustrated in Chapter 7. An example should help us understand how this is done. Table 8.7 presents three groups of 11 scores each; the *Y* scores are the original values, sorted in ascending order within each group. Recognizing the presence of some outlying scores in each group's tail, we trimmed the lowest and highest two scores in each group, yielding the *T* data. We then replaced the deleted scores with the closest remaining scores to create the winsorized, *W*, set. We chose to delete two from each group's tail because this is roughly 20% of 11, the number of *Y* scores. Two ANOVAs, one on the *T* scores and one on the *W* scores, provide the values needed for the trimmed *F* test. The test is described and illustrated in Box 8.2.

Table 8.7 An example of original (*Y*), trimmed (*T*) and winsorized data (*W*)

Group 1	*Y*	5	6	8	8	10	10	10	10	10	11	14
	T			8	8	10	10	10	10	10		
	W	8	8	8	8	10	10	10	10	10	10	10
Group 2	*Y*	5	9	11	11	11	11	12	13	13	13	16
	T			11	11	11	11	12	13	13		
	W	11	11	11	11	11	11	12	13	13	13	13
Group 3	*Y*	9	10	10	10	10	11	11	11	12	12	15
	T			10	10	10	11	11	11	12		
	W	10	10	10	10	10	11	11	11	12	12	12

Box 8.2 The Trimmed *F* Test Applied to the Data of Table 8.7

After performing the ANOVAs on the *T* and *W* scores, do the following:

1. From the ANOVA of the *T* data, get the between-groups mean square, $MS_{BG,tk}$ (tk refers to trimming *k* scores from each tail); $MS_{BG,tk} = 9.19$.
2. From the ANOVA of the *W* data, get the within-groups sum of squares, $SS_{WG,wk}$; $SS_{WG,wk} = 27.455$. This is divided by the degrees of freedom for the trimmed data set, $a(n - 1 - 2k) = (3)(10 - 4) = 18$. Therefore, the winsorized error mean square, $MS_{WG,wk}$, is $27.455/18 = 1.523$.
3. The trimmed *F* statistic is $F = MS_{BG,tk}/MS_{WG,wk} = 27.455/1.523 = 18.269$, and is distributed on 2 and 18 *df*. Then $p = .000$.

For comparison purposes, the F statistic on 2 and 30 *df* for the Y data was 2.531; the corresponding p-value was .096. Clearly, in this example, the trimmed t test led to a considerably lower p-value. However, a word of caution is in order. In our example, the variances were roughly homogeneous and the distributions were roughly symmetric. Under other conditions, particularly if the group sizes vary, Type 1 errors may be inflated (Lix & Keselman, 1998). We will consider such cases in Section 8.8.4.

Dealing with Violations of the Normality Assumption: Tests Based on Ranked Data. In the *Kruskal–Wallis H Test* (1952) and the *rank-transform F test* (Conover & Iman, 1981), all scores are ordered with a rank of 1 assigned to the lowest score in the data set and a rank of N assigned to the largest, where N is the total number of scores. In case of ties, the median rank is assigned; for example, if the five lowest scores are 1, 4, 7, 9, and 9, they would receive ranks of 1, 2, 3, 4.5, and 4.5 respectively. The H test is available from the nonparametric menu of statistical packages such as SPSS, Systat, and SAS. In SPSS, select *Analyze*, then *Nonparametric*, and then *k independent groups*. Applying the Kruskal–Wallis H test to the data of Table 8.7, $p = .039$. As with the trimmed F test, the result is a lower value than that associated with the usual F test of the Y data.

In the rank-transform F test, the usual one-way ANOVA is performed on the ranks and the test statistic, F_R, is evaluated on $a - 1$ and $N - a$ *df*. H and F_R will generally result in similar p-values. This is not surprising given that they are related by the following equation:

$$F_R = \frac{(N-a)H}{(a-1)(N-1-H)}$$

Both tests are more powerful alternatives than the usual F test when the populations are not normally distributed but have the same values of variance, skewness, and kurtosis; that is, when the populations have the same, non-normal distribution. Furthermore, they are only slightly less powerful than the F test when the distributions are normal. However, if the treatment populations do not have identical distributions, then H and F_R tests may reject the null hypothesis because of differences in the shapes or variances of the distributions. Therefore, the tests are not appropriate as tests of location when heterogeneity of variance is suspected (Oshima & Algina, 1992; Vargha & Delaney, 1998).

As we discussed in Chapter 6, when the data are skewed, transformations may provide a solution. Data transformations are also a possible response to heterogeneity of variance. We will discuss this option in the context of our treatment of the assumption of homogeneous variances.

8.8.4 The Homogeneity of Variance Assumption

Variances may differ across conditions for one of several reasons. One possible cause of heterogeneity of variance is an interaction of an experimental treatment with individual characteristics. For example, a drug tested for its effects on depression may result in a higher variance than, but the same mean score as, a placebo. This would suggest that some individuals had improved but others had been adversely affected by the drug. A second possible reason for unequal variances is that some populations are more variable than others on a particular task. For example, although boys may have higher average scores on some measure of mathematical ability, they may also have a higher variance. Still another factor in findings of heterogeneity of variance are floor, or ceiling, effects. Variability may be reduced in one condition relative to another because of a lower, or upper, limit on performance due to the measuring instrument. Finally, variances tend to be correlated with

means, usually positively; the normal distribution is the sole exception in which the means and variances are independently distributed. For all of these reasons, variances are often unequal, or *heterogeneous*, in the populations sampled in our research. In what follows, we summarize some consequences of the failure of this assumption and we then consider alternatives to the standard *F* test.

Consequences of Heterogeneity of Variance. When there are the same number of scores in all conditions, heterogeneous variances usually will cause Type 1 error rates to be slightly inflated. The inflation is usually less than .02 at the .05 level, and less than .005 at the .01 level, provided the ratio of the largest to smallest variance is no more than 4 to 1, and n is at least 5. Even larger ratios may not be a problem, but this will depend upon sample size, the number of groups, and the shape of the population distributions. The results of computer simulations employing these factors are discussed in articles by Clinch and Keselman (1982) and Tomarken and Serlin (1986).

When there are different numbers of scores in each condition, simulation studies clearly demonstrate that heterogeneous variances are a problem. Sampling from heavy-tailed and skewed distributions, and using variance ratios of largest to smallest as high as 16:1, Lix and Keselman (1998) found that error rates were as high as .50 in some conditions. Sampling from sets of either three or four normally distributed populations, Tomarken and Serlin found that at a nominal .05 level, the actual Type 1 error rate was as low as .022 when the group size was positively correlated with the variance (i.e., larger groups associated with greater variances) and as high as .167 when the correlation was negative. This is because the average within-group variance (i.e., the error term, $MS_{S/A}$) is increased when the largest groups have the largest variances and, conversely, is decreased when the largest groups have the smallest variances. Therefore, a positive relation between group size and variance will negatively bias the *F* test whereas a negative relation will positively bias the test.

There is evidence in the research literature that extreme variance ratios do occur (Wilcox, 1987), and simulation studies make clear that heterogeneity of variance can inflate Type 1 error rates or deflate power, depending upon various factors such as sample sizes and the type of distribution sampled. That leaves us with two questions. First, for a given data set, how do we decide whether to abandon the standard ANOVA for some remedial procedure? Second, If we do decide that unequal variances are a threat to the validity of the standard *F* test, what alternative should we use? We consider these questions next.

Detecting Heterogeneity of Variance. As always, we urge that researchers begin the data analysis by examining summary statistics and plots of the data. Computer programs such as SPSS's Explore module are very helpful in this respect. Typically, they provide descriptive statistics, tests of homogeneity of variance, and box plots. The box plot for the Beck Depression data as a function of educational level was presented in Fig. 8.2 and, as we noted there, differences among the groups in shape and spread are quite evident. The range of variances in Table 8.5 suggest that the alpha level reported in Table 8.6 may not be the actual probability of a Type 1 error. For confirmation of this, we may wish to test whether the variances are homogeneous. Several tests of homogeneity of variance have been proposed. Some are overly sensitive to violations of the normality assumption (Bartlett, 1937; Cochran, 1941; Hartley, 1950; Levene, 1960) or lack power relative to other procedures (Box, 1953; Games, Keselman, & Clinch, 1979; Scheffé, 1959).

We recommend the Brown–Forsythe test (Brown & Forsythe, 1974*a*) based on deviations from the median. Sampling studies indicate it has only a slightly inflated Type 1 error rate and good power relative to various competitors even when *n*s are not equal and distributions depart markedly from the normal (Games, Keselman, & Clinch, 1979). In this test, the absolute residual of each

score from its group median, $|Y_{ij} - \overline{Y}_{.j}|$, is computed, and these residuals are then submitted to the analysis of variance. Although these residuals do not directly represent the variance, their variance is an index of the spread of scores. For the depression scores summarized in Table 8.5, SPSS's Explore module reports the value of this statistic as 4.511 which, on 3 and 124 *df*, is very significant ($p = .005$). This indicates that the mean absolute residual varies significantly as a function of education level, confirming our sense that the spread of scores was indeed a function of the educational level.

Once we conclude that the population variances are not equal, the next question is: What shall we do about it? One possible response is to seek a transformation of the data that yields homogeneity of variance on the scale resulting from the transformation. A second possibility is to compute an alternative to the usual *F* test. We consider each of these approaches next.

Dealing with Heterogeneity of Variance: Transformations of the Data. Transformation of the data can sometimes result in variances that are more nearly similar on the new scale. Typical data transformations include raising scores to a power, or taking the natural logarithm of each score. These and other transformations have been used (1) to transform skewed distributions into more nearly normal distributions; (2) to reduce heterogeneity of variance; and (3) to remedy a condition known as "nonadditivity" in designs in which each subject is tested on several trials or under several treatment levels. A transformation which best achieves one purpose may not be equally suitable for other purposes, although it is true that transformations that equate variances do tend to yield more normally distributed scores. Our focus here will be on transformations designed to achieve homogeneous variances.

We begin by noting that transformations are not always a good option for a researcher. One potential problem is that values on the new scale may be less easily interpreted than on the original scale. For example, the percent correct on a test (y) is easily understood and communicated, but this is less true of the arc sine transformation ($\sin^{-1}\sqrt{y}$, the angle whose sine is the square root of y), often recommended to stabilize the variances of percentage scores. Another potential problem is that although variance-stabilizing transformations will usually maintain the ordering of the group means, the relative distances among means may change, creating problems when interpreting the effects of the factors manipulated. Suppose a researcher has predicted that response time will vary as a linear function of the levels of the independent variable. A test of linearity on a transformed scale will probably not support the prediction because the means on the new scale are likely to fall on a curve.

If the measurement scale is arbitrary, however, transfoming the data is one strategy for dealing with heterogeneous variances. In that case, how does the researcher identify a useful transformation for reducing the range of variances across experimental conditions? One possibility is to try several transformations. However, it is not appropriate to conduct a significance test after each transformation and choose the data scale that yields the largest *F* value. Such a procedure is bound to increase the probability of a Type 1 error if the population means do not differ. A more principled approach to identifying a variance-stabilizing transformation depends on observing a functional relation between the cell variances and cell means (Smith, 1976).

Emerson and Stoto (1983) described an approach that will frequently produce more nearly equal variances. The technique involves plotting the log of the *H*-spread (or interquartile range; see Chapter 2) as a function of the log of the median and then finding the slope of the best-fitting straight line. SPSS provides such a *spread versus level plot* in its Explore module if the "Plots" option is chosen and "Power Transformations" is checked. The output includes the value of the slope, which is used to transform the original score, Y, into the transformed score, Z, by the following *power transformation:*

$$Z = Y^{1-slope} \tag{8.24}$$

We obtained this plot for data from Royer's (1999) study of arithmetic skills in elementary schoolchildren.[6] Using SPSS's Explore module to analyze multiplication response times (*RT*), the slope of the spread-versus-level plot was 2.227 and the recommended power was therefore –1.227. We rounded this, letting $Z = Y^{-1} = 1/Y$, thus re-expressing response time as response speed, a measure that is easily understood. Table 8.8 presents the group means and variances on the original and new data scales. On the original *RT* scale, the ratio of largest to smallest variance is almost 15 to 1; on the speed scale, that ratio is only 1.4 to 1. We might also point out that the procedure illustrated here of rounding the recommended power to the nearest "meaningful" number makes sense from the perspective of communicating the transformation to an audience. For example, transforming by taking an inverse (power = –1) or square root (power = .5) or square (power = 2) will be more easily communicated and is unlikely to produce very different results than some "odd" number, such as –.742 or 1.88.

Often the researcher will not wish to transform the data because of the difficulty of interpreting effects (or lack of effects) on the new scale, or because a strong theory dictates the dependent variable. In other instances, it may be impossible to find a variance-stabilizing transformation.[7] Fortunately, there are other solutions that often can solve the heterogeneity problem. We turn now to consider modifications of the standard *F* test.

Dealing with Heterogeneity of Variance: Welch's F test. Several alternatives to the standard *F* test of the equality of the *a* population means have been proposed (Alexander & Govern, 1994; Brown & Forsythe, 1974b; James, 1951, 1954; Welch, 1951), but no one test is best under all conditions. When the data are normally distributed and *n*s are equal, most of the procedures are reasonably robust with respect to Type 1 error rate; however, the standard *F* is slightly more powerful if the population variances are equal. When the variances are not equal, the choice of test depends upon the degree of skew and kurtosis, whether outliers are present, the degree of heterogeneity of variance, the relation between group sizes and group variances, and the total *N* (Clinch & Keselman, 1982; Coombs, Algina, & Oltman, 1996; Grissom, 2000; Lix, Keselman, & Keselman, 1996; Tomarken & Serlin, 1986).

Table 8.8 Means and variances of multiplication *RT* and speeds from the *Royer* data

	Grade			
	5	6	7	8
RT mean	3.837	1.998	1.857	1.935
RT variance	4.884	.612	.328	.519
Speed mean	.350	.560	.586	.583
Speed variance	.033	.028	.031	.038

[6] These data are in the file *Royer_RT*; go to the *Royer* data set on the book's website.

[7] Not all measures can be successfully transformed. For example, on the basis of the spread-vs-level plot, we found that the best transformation of the depression scores was to raise them to the –.106 power. However, following this transformation, and also after a log transform, variances still differed significantly by some tests, and the normality assumption was still violated.

Although there is rarely a clear-cut choice for any given data set, in a review of several simulation studies, Lix et al. (1996) concluded that the *Welch test,* F_w, provided the best alternative when both Type 1 error rates and power were considered. F_w performs well relative to various competitors except when the data are highly skewed (skew > 2.0) or group sizes are less than 10 (Lix et al., 1996; Tomarken & Serlin, 1986). Furthermore, the test is available in several statistical packages. In SPSS, both the standard test results and those for the Welch test are in the output of the *One-Way ANOVA* (*Compare Means*) option.

If you lack the appropriate software, Box 8.3 presents the necessary formulas. Substituting values from Table 8.5 in the equations in the box, we have:

	HS	C	B	GS
$w_j =$	.550	5.528	3.752	1.488
$\bar{Y}_{.j} =$	6.903	3.674	3.331	4.847

Then, $u = 11.318$, $\bar{Y}_{..} = 3.871$, $A = 2.594$, $B = 1.024$, $F = 2.533$, $df_1 = 3$, $df_2 = 1/.018 = 55$, and $p = .066$. The resulting p-value is considerably higher than the .016 we obtained using the standard F calculations. The discrepancy can be accounted for by noting that the correlation between n_j and s_j^2 is negative, −.59. There are only 19 subjects in the group having only a high-school education (HS) whereas the other groups all have at least 33 subjects. Because the larger groups have smaller variances, they have more weight in the denominator of the F test; that small denominator contributes to a larger F statistic with a resulting inflated probability of a Type 1 error. The Welch test has compensated for this by taking the inequalities in group sizes and variances into account.

Box 8.3 Formulas for the Welch (F_w) Test

$$F_W = \frac{A}{B}$$

where $$A = \frac{1}{a-1}\sum w_i(\bar{Y}_{.i} - \bar{Y}_{..})^2$$

$$B = 1 + \left[\frac{2(a-2)}{a^2-1}\right]\sum\frac{[1-(w_i/u)]^2}{n_i - 1}$$

and $$w_i = n_i / s_i^2;\ u = \sum w_i;\ \bar{Y}_{..} = \sum w_i\bar{Y}_{.i}/u$$

$$df_1 = a - 1$$

$$\frac{1}{df_2} = \left[\frac{3}{a^2-1}\right]\sum\frac{[1-(w_i/u)]^2}{n_i - 1}$$

A Robust F Test. The normality and homogeneity of variance assumptions often are violated in the same data set. This is particularly a problem when *n*s are unequal, as in the Beck Depression data we analyzed. A promising approach is to apply the Welch F test to means based on data from which the highest and lowest 20% have been trimmed.(Keselman, Wilcox, Othman, & Fradette, 2002; Lix & Keselman, 1998). The test is described in Box 8.4. Because it uses trimmed means and

winsorized variances, it may be helpful to refer to Chapter 7 where trimmed means and winsorized variances were defined and illustrated.

Box 8.4 Welch's (1951) *F* Test with Trimmed Means and Winsorized Variances

1. Replace the n_j in Box 8.2 by $h_j = n_j - 2k_j$, where k_j is the number of scores trimmed from each tail of the jth group. For example, in the HS group of the depression analysis, $n_j = 19$. If we trim approximately 20% from each tail, $k = 4$ and $h = 19 - (2)(4) = 11$.
2. The $\overline{Y}_{.j}$ are replaced by the trimmed means.
3. The s_j^2 are replaced by the winsorized group variances.
4. With these substitutions, the formulas for the Welch F test in Box 8.3 apply directly.

8.9 SUMMARY

This chapter introduced the analysis of variance in the simplest possible context, the one-factor, between-subjects design. The developments in this chapter will be relevant in the analyses of data from other designs. These developments included:

- *The analysis of variance*. We illustrated the idea of a structural model that underlies the data and directs the partitioning of variability in the data. The structural model is the basis for determining what the variance calculations estimate in terms of population variance parameters. The *EMS*, in turn, justify the error terms for tests of the null hypothesis and are involved in estimating measures of the magnitude of effects.
- *Measures of importance*. We defined and applied to data several statistics that indicate the importance of the independent variable. Confidence intervals also were presented that provide a range of plausible values of the parameter being estimated.
- *A priori* power and sample size. We illustrated how sample size for a multi-group study can be determined, once the values of α, the desired power, and the effect size of interest are selected.
- *Assumptions underlying the significance test and the estimates of measures of importance.* We reviewed these assumptions, discussed the consequences of their violation, cited tests of the assumptions that are available in many software packages, and described procedures that respond to violations.

APPENDIX 8.1

Partitioning the Total Variability in the One-Factor Design

The following developments involve two indices of summation: i indexes a value from 1 to n within each group, where n is the number of individuals in a group; j indexes a value from 1 to a, where a is the number of groups. Appendix A provides an explanation of the use of this notation, using several examples.

Squaring both sides of Equation 8.1 yields

$$(Y_{ij} - \bar{Y}_{..})^2 = (Y_{ij} - \bar{Y}_{.j})^2 + (\bar{Y}_{.j} - \bar{Y}_{..})^2 + 2(Y_{ij} - \bar{Y}_{.j})(\bar{Y}_{.j} - \bar{Y}_{..})$$

Summing over i and j, and applying the rules of Appendix A, we have

$$\sum_{j}^{a}\sum_{i}^{n}(Y_{ij} - \bar{Y}_{..})^2 = \sum_{j}^{a}\sum_{i}^{n}(Y_{ij} - \bar{Y}_{.j})^2 + n\sum_{j}^{a}(\bar{Y}_{.j} - \bar{Y}_{..})^2 + 2\sum_{j}^{a}\sum_{i}^{n}(Y_{ij} - \bar{Y}_{.j})(\bar{Y}_{.j} - \bar{Y}_{..})$$

Rearranging terms, we can show that the rightmost (cross-product) term equals 0:

$$2\sum_{j}^{a}\sum_{i}^{n}(Y_{ij} - \bar{Y}_{.j})(\bar{Y}_{.j} - \bar{Y}_{..}) = 2\sum_{j}^{a}(\bar{Y}_{.j} - \bar{Y}_{..})\sum_{i}^{n}(Y_{ij} - \bar{Y}_{.j})$$

$$= 2\sum_{j}^{a}(\bar{Y}_{.j} - \bar{Y}_{..})(0) = 0$$

The last result follows because the sum of deviations of scores about their mean is zero.

EXERCISES

8.1 A data set has three groups of five scores each. Because the scores involve decimal values, each score is multiplied by 100.

- **(a)** How will the mean squares and F ratio be affected (relative to an analysis of the original data set)?
- **(b)** In general, what happens to a variance when every score is multiplied by a constant?
- **(c)** Suppose we just added a constant, say, 10, to all 15 scores. How would that affect the mean squares and F ratio?
- **(d)** Suppose we added 5 to all scores in the first group, 10 to all scores in group 2, and 15 to all scores in group 3? Should MS_A change? $MS_{S/A}$? Explain.

8.2 Following are summary statistics from a three-group experiment. Present the ANOVA table when (a) $n_1 = n_2 = n_3 = 10$ and (b) $n_1 = 6$, $n_2 = 8$, and $n_3 = 10$; the totals, or sums of scores, for the groups, the $T_{.j}$, and the variances are:

	A_1	A_2	A_3
Totals	30	48	70
Variances	3.2	4.1	5.7

8.3 The data are:

A_1: 27 18 16 33 24 $\qquad$ A_2: 23 33 26 19 38

- **(a)** Perform the ANOVA.
- **(b)** Next, do a t test. How are the results of parts (a) and (b) related?

8.4 The F ratio is basically a test of the equality of the population variances estimated by its numerator and denominator. Therefore, it is applicable to the following problem. We have samples of reading scores from 5 boys and 11 girls. We form a ratio of the variances of the two samples, s_B^2/s_G^2.

(a) If many samples of sizes 5 and 11 are drawn, (i) what is the proportion of F values greater than 2.61 that we should expect? (ii) less than 4.47?

(b) What assumptions are implied in your approach to answering part (a)?

8.5 The file *EX8_5* at the website contains three groups of 15 scores.

(a) Explore the data; examine statistics and graphs relevant to assessing the normality and homogeneity of variance assumptions. What are the implications for a significance test?

(b) Calculate the F and Kruskal–Wallis H tests for these data and comment on the outcome, relating your discussion to your answer to part (a).

8.6 **(a)** A nonparametric test is only one way to reduce the effect of the straggling right tail of the data in Exercise 8.5. Explore the data after transformation by taking (1) the square root of each score and (2) the natural log of each score. Does either one better conform to the assumptions underlying the F test? Explain.

(b) Carry out the ANOVA with the transformation you selected in part (a). How do the results compare with those for the original F test in Exercise 8.5?

(c) Find the confidence intervals for the three means, using the Y data. Then do the same with the group means for the transformed scores. Transform the means of the transformed scores to the original scale. For example, if you had selected the square-root transformation, you would square the transformed means; if you had selected the log transformation, you would raise e to the power of the mean on the log scale (for example, if the mean on the log scale = 3, on the original scale we would have $e^3 = 20.09$). Do the same for the 95% confidence limits for each of the three means. Compare the results for the original and transformed data.

8.7 The following are the results of two experiments, each with three levels of the independent variable.

Table 1

SV	*df*	*MS*
A	2	80
S/*A*	27	5

Table 2

SV	*df*	*MS*
A	2	42.5
S/*A*	12	5

(a) For each of the two tables, calculate the Fs, and estimates of ω_A^2.

(b) What does a comparison of the two sets of results suggest about the effect of the change in n upon these two quantities?

(c) Calculate η_A^2 for each table. How does the change in n affect the value of η_A^2?

(d) Suppose $F = 1$. (i) What must the value of ω_A^2 be? (ii) What must the value of η_A^2 be (as a function of a and n)?

(e) Comment on the relative merits of the various statistics calculated as indices of the importance of A.

8.8 The result of an ANOVA of a data set based on three groups of 10 scores each is:

SV	*df*	*SS*	*MS*	*F*
A	2	192	96	3.2
S/*A*	27	810	30	

(a) Is there a significant A effect if $\alpha = .05$?

(b) Estimate Cohen's f for these results.

(c) Assuming this is a good estimate of the true effect of A, what power did the experiment have?

(d) How many subjects would be required to have power = .8 to detect a medium-sized effect? Use Cohen's guidelines.

8.9 According to a mathematical model for the experiment in Exercise 8.8, the predicted means are 10 in condition 1, 14 in condition 2, and 18 in condition 3. If the theory is correct, what sample size would be needed to achieve .8 power to reject the null hypothesis of no difference among the means? Assume that the error mean square is a good estimate of the population variance, and $\alpha = .05$.

8.10 In a study of the relative effectiveness of three methods of teaching elementary probability, students read one of three texts: the Standard (S), the Low Explanatory (LE), and the High Explanatory (HE). The data—scores on a test after a single reading—are in the file *EX8_10* on the website.

(a) Explore the data. Are there any indications of departures from the underlying assumptions?

(b) Test the null hypothesis that the texts do not differ in their effects.

(c) Estimate ω^2 and Cohen's f. Verify that $\omega^2 = f^2/(1 + f^2)$ and $f^2 = \omega^2/(1 - \omega^2)$.

(d) Based on these results, if you were to replicate the study, how many subjects would you run to have power = .8?

8.11 The *Sayhlth* file linked to the *Seasons* page on the website contains *Sayhlth* scores (self-ratings of health) of 1–4 (excellent to fair; three subjects with poor ratings in the *Seasons* file are not included). The four categories will be the independent variable in this exercise and the *Beck_D* score will be the dependent variable in the following analyses. The *Beck_D* score is an average of the four seasonal Beck Depression scores and is available only for those subjects whose scores were recorded in all four seasons. The distribution of *Beck_D* scores tends to be skewed and, as in most non-normal distributions, heterogeneity of variance is often a problem.

(a) Explore the *Beck_D* (seasonal mean) data, using any statistics and plots you think are relevant, and comment on the relative locations, shapes, and variabilities of the scores in the four categories.

(b) Using the four *Sayhlth* categories, plot the spread vs level; as stated in Chapter 8, this is the log of the H-spread plotted against the log of the median. Several statistical software packages make this plot available. Find the best-fit regression line for this plot and transform the *Beck_D* scores by raising them to the power, 1 – slope.

(c) Explore the distribution of the transformed scores at each *Sayhlth* category. Has the transformation had any effect on the shape of the distributions or on their variances? Test for homogeneity of variance.

(d) Next try a different transformation. Calculate log(*Beck_D* + 1) and discuss the effects of this transformation.

(e) What might be the advantages of transforming data to a scale on which they are normally distributed with homogeneous variances?

8.12 The *Sayhlth* file also categorizes individuals by employment category; 1 = employed full time; 2 = employed part-time; 3 = not employed.

(a) Explore the *Beck_D* data in each *Employ* category, looking at relevant graphs and statistics. Comment on the validity of the ANOVA assumptions.

(b) In Exercise 8.11, we considered transformations of the *Beck_D* data, one of which appeared to provide results more in accord with the ANOVA model. Use that

transformation and again explore the data. Are the results more in accord with the ANOVA model?

(c) Do ANOVAs on the *Beck_D* scores and the transformed scores as a function of employment status. What do you conclude about the effects of employment?

(d) Does the Welch *F* test confirm or contradict your conclusion?

(e) Calculate Cohen's *f* for both the original and the transformed data. How would you characterize the effect sizes? In general, what can you say about the effect of employment status on depression scores?

8.13 Continuing with the *Sayhlth* file,

(a) Using the four *Sayhlth* categories as your independent variable, do separate *ANOVA*s of the *Beck_D* data for men and for women.

(b) Calculate Cohen's *f* for each sex and compare the effect sizes.