Statistical Power Analysis

A Simple and General Model for Traditional and Modern Hypothesis Tests

THIRD EDITION

Kevin R. Murphy
rett Myors
llen Wolach

Statistical Power Analysis

A Simple and General Model for Traditional and Modern Hypothesis Tests

THIRD EDITION

Kevin R. Murphy
Pennsylvania State University

Brett Myors
Griffith University

Allen Wolach
Illinois Institute of Technology

Routledge
Taylor & Francis Group
New York London

Routledge
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Routledge
Taylor & Francis Group
2 Park Square
Milton Park, Abingdon
Oxon OX14 4RN

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

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-84169-774-1 (Softcover) 978-1-84169-775-8 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Murphy, Kevin R., 1952-
Statistical power analysis : a simple and general model for traditional and modern hypothesis tests / Kevin R. Murphy, Brett Myron, Allen Wolach. -- 3rd ed.
p. cm.
ISBN 978-1-84169-774-1 (pbk.) -- ISBN 978-0-415-96555-2 (hardback)
1. Statistical hypothesis testing. 2. Statistical power analysis. I. Myron, Brett. II. Wolach, Allen H. III. Title.

QA277.M87 2008
519.5'6--dc22 2008037988

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

and the Psychology Press Web site at
http://www.psypress.com

Statistical Power Analysis

A Simple and General Model
for Traditional and Modern Hypothesis Tests

THIRD EDITION

Preface

One of the most common statistical procedures in the behavioral and social sciences is to test the hypothesis that treatments or interventions have no effect, or that the correlation between two variables is equal to zero, etc.—i.e., tests of the null hypothesis. Researchers have long been concerned with the possibility that they will reject the null hypothesis when it is in fact correct (i.e., make a Type I error), and an extensive body of research and data-analytic methods exists to help understand and control these errors. Less attention has been devoted to the possibility that researchers will fail to reject the null hypothesis, when in fact treatments, interventions, etc., have some real effect (i.e., make a Type II error). Statistical tests that fail to detect the real effects of treatments or interventions might substantially impede the progress of scientific research.

The statistical power of a test is the probability that it will lead you to reject the null hypothesis when that hypothesis is in fact wrong. Because most statistical tests are conducted in contexts where treatments have at least some effect (although it might be minuscule), power often translates into the probability that your test will you lead to a correct conclusion about the null hypothesis. Viewed in this light, it is obvious why researchers have become interested in the topic of statistical power and in methods of assessing and increasing the power of their tests.

This book presents a simple and general model for statistical power analysis that is based on the widely used F statistic. A wide variety of statistics used in the social and behavioral sciences can be thought of as special applications of the "general linear model" (e.g., t-tests, analysis of variance and covariance, correlation, multiple regression), and the F statistic can be used in testing hypotheses about virtually any of these specialized applications. The model for power analysis laid out here is quite simple, and it illustrates how these

analyses work and how they can be applied to problems of study design, to evaluating others' research, and even to problems such as choosing the appropriate criterion for defining "statistically significant" outcomes.

In response to criticisms of traditional null hypothesis testing, several researchers have developed methods for testing what we refer to as "minimum-effect" hypotheses—i.e., the hypothesis that the effect of treatments, interventions, etc., exceeds some specific minimal level. Ours is the first book to discuss in detail the application of power analysis to both traditional null hypothesis tests and minimum-effect tests. We show how the same basic model applies to both types of testing and illustrate applications of power analysis to both traditional null hypothesis tests (i.e., tests of the hypothesis that treatments have no effect) and to minimum-effect tests (i.e., tests of the hypothesis that the effects of treatments exceeds some minimal level).

Most of the analyses presented in this book can be carried out using a single table, the One-Stop *F* Table presented in Appendix B. Appendix C presents a comparable table that expresses statistical results in terms of the percentage of variance (PV) explained rather than the *F* statistic. These two tables make it easy to move back and forth between assessments of statistical significance and assessments of the strength of various effects in a study. The One-Stop *F* Table can be used to answer many questions that relate to the power of statistical tests. A computer program, the One-Stop *F* Calculator, is on the book's website www.psypress.com/statistical-power-analysis. The One-Stop *F* Calculator can be used as a substitute for the One-Stop *F* Table. This computer program allows users more flexibility in defining the hypothesis to be tested, the desired power level, and the alpha level than is typical for power analysis software. The One-Stop *F* Calculator also makes it unnecessary to interpolate between values in a table.

This book is intended for a wide audience, including advanced students and researchers in the social and behavioral sciences, education, health sciences, and business. Presentations are kept simple and nontechnical whenever possible. Although most of the examples in this book come from the social and behavioral sciences, general principles explained in this book should be useful to researchers in diverse disciplines.

Changes in the New Edition

This third edition includes expanded coverage of power analysis for multifactor analysis of variance (ANOVA), including split-plot and randomized block factorial designs. Although conceptual issues for power analysis are similar in factorial ANOVA and other methods of analysis, special features

of ANOVA require explicit attention. The present edition of the book also shows how to calculate power for simple main effects tests and *t* tests that are performed after an analysis of variance is performed, and it provides a more detailed examination of *t* tests than was included in our first and second editions.

Perhaps the most important addition to this third edition is a set of examples, illustrations, and discussions included in Chapters 1 through 8 in boxed sections. This material is set off for easy reference, and it provides examples of power analysis in action and discussions of unique issues that arise as a result of applying power analyses in different designs.

Other highlights of the third edition include the following:

- A completely redesigned, user-friendly software program that allows users to carry out all of the analysis described in this book and to conduct a wide range of tests; this new program allows users to conduct significance tests, power analyses, and assessments of N and alpha needed for traditional and minimum-effects tests
- New chapters (Chapters 7 and 8) demonstrating the application of POWER in complex ANOVA designs, including randomized block, split-plot, and repeated measures designs
- A separate chapter (chapter 3) describing the rationale and operation of minimum effects tests
- Worked examples in all chapters
- Expanded coverage of the concepts behind power analysis (Chapters 1 and 4) and the application of these concepts in correlational studies (Chapter 5)

Using the One-Stop *F* Calculator

A book-specific website www.psypress.com/statistical-power-analysis includes the One-Stop F Calculator, which is a program designed to run on most Windows-compatible computers. Following the philosophy that drives our book, the program is simple to install and use. Visit this website, and you will receive instructions for quickly installing the program. The program asks you to make some simple decisions about the analysis you have in mind, and it provides information about statistical power, effect sizes, *F* values, and/or significance tests. Chapter 2 illustrates the use of this program.

Acknowledgments

We are grateful for the comments and suggestions of several reviewers, including Stephen Brand, University of Rhode Island; Jaihyun Park, Baruch College–CUNY; Eric Turkheimer, University of Virginia; and Connie Zimmerman, Illinois State University.

Contents

1

The Power of Statistical Tests

In the social and behavioral sciences, statistics serve two general purposes. First, they can be used to describe what happened in a particular study (descriptive statistics). Second, they can be used to help draw conclusions about what those results mean in some broader context (inferential statistics). The main question in inferential statistics is whether a result, finding, or observation from a study reflects some meaningful phenomenon in the population from which that study was drawn. For example, if 100 college sophomores are surveyed and it is determined that a majority of them prefer pizza to hot dogs, does this mean that people in general (or college students in general) also prefer pizza? If a medical treatment yields improvements in 6 out of 10 patients, does this mean that it is an effective treatment that should be approved for general use? The goal of inferential statistics is to determine what sorts of inferences and generalizations can be made on the basis of data of this type and to assess the strength of evidence and the degree of confidence one can have in these inferences.

The process of drawing inferences about populations from samples is a risky one, and a great deal has been written about the causes and cures for errors in statistical inference. Statistical power analysis (Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990) falls under this general heading. Studies with too little statistical power can lead to erroneous conclusions about the meaning of the results of a particular study. In the example cited above, the fact that a medical treatment worked for 6 out of 10 patients is probably insufficient evidence that it is truly safe and effective; and if you have nothing more than this study to rely on, you might conclude that the treatment had not been proven effective. Does this mean that you should abandon the treatment or that it is unlikely to work in a broader population? The conclusion that the treatment has not been shown to be effective may

say as much about the low level of statistical power in your study as about the value of the treatment.

In this chapter, we will describe the rationale for and applications of statistical power analysis. In most of our examples, we describe or apply power analysis in studies that assess the effect of some treatment or intervention (e.g., psychotherapy, reading instruction, performance incentives) by comparing outcomes for those who have received the treatment to outcomes of those who have not (nontreatment or control group). However, power analysis is applicable to a very wide range of statistical tests, and the same simple and general model can be applied to virtually all of the statistical analyses you are likely to encounter in the social and behavioral sciences.

The Structure of Statistical Tests

To understand statistical power, you must first understand the ideas that underlie statistical hypothesis testing. Suppose 100 children are randomly divided into two groups. Fifty children receive a new method of reading instruction, and their performance on reading tests is on average 6 points higher (on a 100-point test) than the other 50 children who received standard methods of instruction. Does this mean that the new method is truly better? A 6-point difference *might* mean that the new method is really better, but it is also possible that there is no real difference between the two methods, and that this observed difference is the result of the sort of random fluctuation you might expect when you use the results from a single sample to draw inferences about the effects of these two methods of instruction in the population.

One of the most basic ideas in statistical analysis is that results obtained in a sample do not necessarily reflect the state of affairs in the population from which that sample was drawn. For example, the fact that scores averaged 6 points higher in this particular group of children does not necessarily mean that scores will be 6 points higher in the population, or that the same 6-point difference would be found in another study examining a new group of students. Because samples do not (in general) perfectly represent the populations from which they were drawn, you should expect some instability in the results obtained from each sample. This instability is usually referred to as "sampling error." The presence of sampling error is what makes drawing inferences about populations from samples difficult. One of the key goals of statistical theory is to estimate the amount of sampling error that is likely to be present in different statistical procedures and tests and thereby gaining some idea about the amount of risk involved in using a particular procedure.

Statistical significance tests can be thought of as decision aids. That is, these tests can help you reach conclusions about whether the findings of your particular study are likely to represent real population effects or whether they fall within the range of outcomes that might be produced by random sampling error. For example, there are two possible interpretations of the findings in this study of reading instruction:

1. The difference between average scores from the two programs is so small that it might reasonably represent nothing more than sampling error.

versus

2. The difference between average scores from the two programs is so large that it cannot be reasonably explained in terms of sampling error.

The most common statistical procedure in the social and behavioral sciences is to pit a null hypothesis (H_0) against an alternative (H_1). In this example, the null and alternative hypotheses might take the forms:

H_0—Reading instruction has no effect. It doesn't matter how you teach children to read, because in the population there is no difference in the average scores of children receiving either method of instruction.

versus

H_1—Reading instruction has an effect. It does matter how you teach children to read, because in the population there *is* a difference in the average scores of children receiving different methods of instruction.

Although null hypotheses usually refer to "no difference" or "no effect," it is important to understand that there is nothing magic about the hypothesis that the difference between two groups is zero. It might be perfectly reasonable to evaluate the following set of possibilities:

H_0—In the population, the difference in the average scores of those receiving these two methods of reading instruction is 6 points.

versus

H_1—In the population, the difference in the average scores of those receiving these two methods of reading instruction is *not* 6 points.

Another possible set of hypotheses is:

H_0—In the population, the new method of reading instruction is *not* better than the old method; the new method might even be worse.

versus

H_1—In the population, the new method of reading instruction *is* better than the old method.

This set of hypotheses leads to what is often called a "one-tailed" statistical test, in which the researcher not only asserts that there is a real difference between these two methods, but also describes the direction or the nature of this difference (i.e., that the new method is not just different from the old one, it is also better). We discuss one-tailed tests in several sections of this book, but in most cases we will focus on the more widely used two-tailed tests that compare the null hypothesis that nothing happened with the alternative hypothesis that something happened. Unless we specifically note otherwise, the traditional null hypothesis tests discussed in this book will be assumed to be two-tailed. However, the minimum effect tests we introduce in Chapter 2 and discuss extensively throughout the book have all of the advantages and few of the drawbacks of traditional one-tailed tests of the null hypothesis.

Null Hypotheses Versus Nil Hypotheses

The most common structure for tests of statistical significance is to pit the null hypothesis that treatments have no effect, or that there is no difference between groups, or that there is no correlation between two variables against the alternative hypotheses that there is *some* treatment effect. In fact, this structure is so common that most people assume that the "null hypothesis" is essentially a statement that there is no difference between groups, no treatment effect, no correlation between variables, etc. This is not true. The null hypothesis is simply the hypothesis you actually test, and if you reject the null, you are left with the alternative. That is, if you reject the hypothesis that the effect of an intervention of treatment is *X*, you are left to conclude that the alternative hypothesis that the effect of treatments is *not-X* must be true. If you test and reject the hypothesis that treatments have no effect, you are left with the conclusion that they must have some effect. If you test and reject the hypothesis that a particular diet will lead to a 20% weight loss, you are left with the conclusion that the diet will *not* lead to a 20%

weight loss (it might have no effect; it might have a smaller effect; it might even have a larger effect).

Following Cohen's (1994) suggestion, we think it is useful to distinguish between the null hypothesis in general and its very special and very common form, the "nil hypothesis" (i.e., the hypothesis that treatments, interventions, etc., have no effect whatsoever). The nil hypothesis is common because it is very easy to test and because it leaves you with a fairly simple and concrete alternative. If you reject the nil hypothesis that nothing happened, the alternative hypothesis you should accept is that something happened. However, as we show in this chapter and in the chapters that follow, there are often important advantages to testing null hypotheses that are broader than the traditional nil hypothesis.

Most treatments of power analysis focus on the statistical power of tests of the nil hypothesis (i.e., tests of the hypothesis that treatments or interventions have no effect whatsoever). However, there are a number of advantages to posing and testing substantive hypotheses about the size of treatment effects (Murphy & Myors, 1999). For example, it is easy to test the hypothesis that the effects of treatments are negligibly small (e.g., they account for 1% or less of the variance in outcomes, or that the standardized mean difference is .10 or less). If you test and reject this hypothesis, you are left with the alternative hypothesis that the effect of treatments is *not* negligibly small, but rather large enough to deserve at least some attention. The methods of power analysis described in this book are easily extended to such minimum-effect tests and are not limited to traditional tests of the null hypothesis that treatments have no effect.

What determines the outcomes of statistical tests? There are four outcomes that are possible when you use the results obtained in a particular sample to draw inferences about a population; these outcomes are shown in Figure 1.1.

As Figure 1.1 shows, there are two ways to make errors when testing hypotheses. First, it is possible that the treatment (e.g., a new method of instruction) has no real effect in the population, but the results in your sample might lead you to believe that it does have some effect. If the results of this study lead you to incorrectly conclude that the new method of instruction does work better than the current method, when in fact there were no differences, you would be making a *Type I* error (sometimes called an *alpha* error). Type I errors might lead you to waste time and resources by pursuing what are essentially dead ends, and researchers have traditionally gone to great lengths to avoid Type I errors.

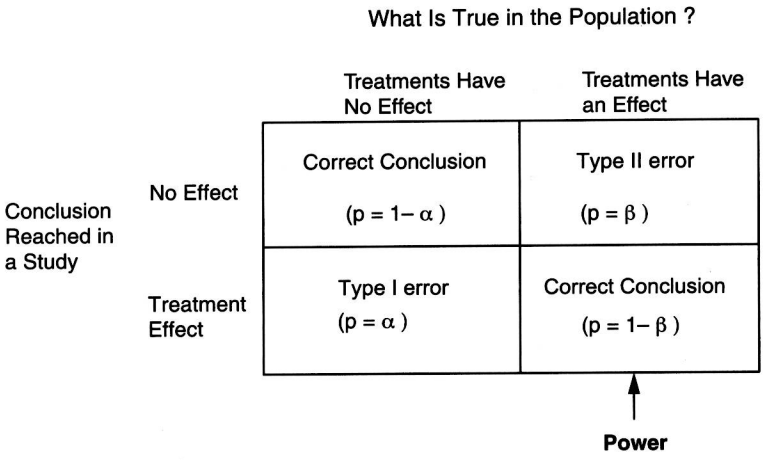

Figure 1.1 Outcomes of statistical tests.

There is an extensive literature dealing with methods of estimating and minimizing the occurrence of Type I errors (e.g., Zwick & Marascuilo, 1984). The probability of making a Type I error is in part a function of the standard or decision criterion used in testing your hypothesis (often referred to as alpha, or α). A very lenient standard (e.g., if there is *any* difference between the two samples, you will conclude that there is also a difference in the population) might lead to more frequent Type I errors, whereas a more stringent standard might lead to fewer Type I errors.[1]

A second type of error (referred to as *Type II* error, or *beta* error) is also common in statistical hypothesis testing (Cohen, 1994; Sedlmeier & Gigerenzer, 1989). A Type II error occurs when you conclude in favor of H_0, when in fact H_1 is true. For example, if you conclude that there are no real differences in the outcomes of these two methods of instruction, when in fact one really is better than the other in the population, you have made a Type II error.

Statistical power analysis is concerned with Type II errors (i.e., if the probability of making a Type II error is β, power = $1 - \beta$). Another way of saying this is to note that power is the (conditional) probability that you will

[1] It is important to note that Type I errors can only occur when the null hypothesis is actually true. If the null hypothesis is that there is no true treatment effect (a nil hypothesis), this will rarely be the case. As a result, Type I errors are probably quite rare in tests of the traditional null hypothesis, and efforts to control these errors at the expense of making more Type II errors might be ill advised (Murphy, 1990).