STATISTICAL POWER ANALYSIS FOR THE BEHAVIORAL SCIENCES

Revised Edition

Jacob Cohen

Statistical Power Analysis for the Behavioral Sciences

Revised Edition

This page intentionally left blank

Statistical Power Analysis for the Behavioral Sciences

Revised Edition

Jacob Cohen
Department of Psychology
New York University
New York, New York

ACADEMIC PRESS New York San Francisco London 1977
A Subsidiary of Harcourt Brace Jovanovich, Publishers

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1

Library of Congress Cataloging in Publication Data

Cohen, Jacob, Date
Statistical power analysis for the behavioral sciences.

Bibliography: p.
Includes index.
1. Social sciences–Statistical methods.
2. Probabilities. I. Title.
HA29.C66 1976 300'.1'82 76-19438
ISBN 0–12–179060–6

PRINTED IN THE UNITED STATES OF AMERICA

to Marcia and Aviva

This page intentionally left blank

Contents

This page intentionally left blank

Preface to the Revised Edition

The structure, style, and level of this edition remain as in the original, but three important changes in content have been made:

1. Since the publication of the original edition, multiple regression/correlation analysis has been expanded into a very general and hence versatile system for data analysis, an approach which is uniquely suited to the needs of the behavioral sciences (Cohen and Cohen, 1975). A new chapter is devoted to an exposition of the major features of this data-analytic system and a detailed treatment of power analysis and sample size determination (Chapter 9).

2. The effect size index used for chi-square tests on frequencies and proportions (Chapter 7) has been changed from e to $w(=\sqrt{e})$. This change was made in order to provide a more useful range of values and to make the operational definitions of "small," "medium," and "large" effect sizes for tests of contingency tables and goodness of fit consistent with those for other statistical tests (particularly those of Chapters 5 and 6). The formulas have been changed accordingly and the 84 look-up tables for power and sample size have been recomputed.

3. The original treatment of power analysis and sample size determination for the factorial design analysis of variance (Chapter 8) was approximate and faulty, yielding unacceptably large overestimation of power for main effects and underestimation for interactions. The treatment in this edition is materially changed and includes a redefinition of effect size for interactions.

The new method gives quite accurate results. Further insight into the analysis of variance is afforded when illustrative problems solved by the methods of this chapter are addressed and solved again by the multiple regression/correlation methods of the new Chapter 9.

Thus, this edition is substantially changed in the areas for which the original edition was most frequently consulted. In addition, here and there, some new material has been added (e.g., Section 1.5.5, "Proving" the Null Hypothesis) and some minor changes have been made for updating and correction.

In the seven years since the original edition was published, it has received considerable use as a supplementary textbook in intermediate level courses in applied statistics. It was most gratifying to note that, however slowly, it has begun to influence research planning and the content of textbooks in applied statistics. Several authors have used the book to perform power-analytic surveys of the research literature in different fields of behavioral science, among them Brewer (1972) in education (but see Cohen, 1973), Katzer and Sodt (1973) and Chase and Tucker (1975) in communication, Kroll and Chase (1975) in speech pathology, Chase and Baran (1976) in mass communication, and Chase and Chase (1976) in applied psychology; others are in preparation. Apart from their inherent value as methodological surveys, they have served to disseminate the ideas of power analysis to different audiences with salutary effects on them as both producers and consumers of research. It is still rare, however, to find power analysis in research planning presented in the introductory methods section of research reports (Cohen, 1973).

As in the original edition, I must first acknowledge my students and consultees, from whom I have learned so much, and then my favorite colleague, Patricia Cohen, a constant source of intellectual excitement and much more. I am grateful to Patra Lindstrom for the exemplary fashion in which she performed the exacting chore of typing the new tables and manuscript.

NEW YORK
JUNE 1976

JACOB COHEN

Preface to the Original Edition

During my first dozen years of teaching and consulting on applied statistics with behavioral scientists, I became increasingly impressed with the importance of statistical power analysis, an importance which was increased an order of magnitude by its neglect in our textbooks and curricula. The case for its importance is easily made: What behavioral scientist would view with equanimity the question of the probability that his investigation would lead to statistically significant results, i.e., its power? And it was clear to me that most behavioral scientists not only could not answer this and related questions, but were even unaware that such questions were answerable. Casual observation suggested this deficit in training, and a review of a volume of the *Journal of Abnormal and Social Psychology* (JASP) (Cohen, 1962), supported by a small grant from the National Institute of Mental Health (M-5174A), demonstrated the neglect of power issues and suggested its seriousness.

The reason for this neglect in the applied statistics textbooks became quickly apparent when I began the JASP review. The necessary materials for power analysis were quite inaccessible, in two senses: they were scattered over the periodical and hardcover literature, and, more important, their use assumed a degree of mathematical sophistication well beyond that of most behavioral scientists.

For the purpose of the review, I prepared some sketchy power look-up tables, which proved to be very easily used by the students in my courses at New York University and by my research consultees. This generated the

idea for this book. A five-year NIMH grant provided the support for the program of research, system building, computation, and writing of which the present volume is the chief product.

The primary audience for which this book is intended is the behavioral or biosocial scientist who uses statistical inference. The terms "behavioral" and "biosocial" science have no sharply defined reference, but are here intended in the widest sense and to include the academic sciences of psychology, sociology, branches of biology, political science and anthropology, economics, and also various "applied" research fields: clinical psychology and psychiatry, industrial psychology, education, social and welfare work, and market, political polling, and advertising research. The illustrative problems, which make up a large portion of this book, have been drawn from behavioral or biosocial science, so defined.

Since statistical inference is a logical-mathematical discipline whose applications are not restricted to behavioral science, this book will also be useful in other fields of application, e.g., agronomy and industrial engineering.

The amount of statistical background assumed in the reader is quite modest: one or two semesters of applied statistics. Indeed, all that I really assume is that the reader knows how to proceed to perform a test of statistical significance. Thus, the level of treatment is quite elementary, a fact which has occasioned some criticism from my colleagues. I have learned repeatedly, however, that the *typical* behavioral scientist approaches applied statistics with considerable uncertainty (if not actual nervousness), and requires a verbal-intuitive exposition, rich in redundancy and with many concrete illustrations. This I have sought to supply. Another feature of the present treatment which should prove welcome to the reader is the minimization of required computation. The extensiveness of the tables is a direct consequence of the fact that most uses will require no computation at all, the necessary answers being obtained directly by looking up the appropriate table.

The sophisticated applied statistician will find the exposition unnecessarily prolix and the examples repetitious. He will, however, find the tables useful. He may also find interesting the systematic treatment of population effect size, and particularly the proposed conventions or operational definitions of "small," "medium," and "large" effect sizes defined across all the statistical tests. Whatever originality this work contains falls primarily in this area.

This book is designed primarily as a handbook. When so used, the reader is advised to read Chapter 1 and then the chapter which treats the specific statistical test in which he is interested. I also suggest that he read all the relevant illustrative examples, since they are frequently used to carry along the general exposition.

The book may also be used as a supplementary textbook in intermediate level courses in applied statistics in behavioral/biosocial science. I have been

using it in this way. With relatively little guidance, students at this level quickly learn both the concepts and the use of the tables. I assign the first chapter early in the semester and the others in tandem with their regular textbook's treatment of the various statistical tests. Thus, each statistical test or research design is presented in close conjunction with power-analytic considerations. This has proved most salutary, particularly in the attention which must then be given to anticipated population effect sizes.

Pride of place, in acknowledgment, must go to my students and consultees, from whom I have learned much. I am most grateful to the memory of the late Gordon Ierardi, without whose encouragement this work would not have been undertaken. Patricia Waly and Jack Huber read and constructively criticized portions of the manuscript. I owe an unpayable debt of gratitude to Joseph L. Fleiss for a thorough technical critique. Since I did not follow all his advice, the remaining errors can safely be assumed to be mine. I cannot sufficiently thank Catherine Henderson, who typed much of the text and all the tables, and Martha Plimpton, who typed the rest.

As already noted, the program which culminated in this book was supported by the National Institute of Mental Health of the Public Health Service under grant number MH-06137, which is duly acknowledged. I am also most indebted to Abacus Associates, a subsidiary of American Bioculture, Inc., for a most generous programming and computing grant which I could draw upon freely.

NEW YORK
JUNE 1969

JACOB COHEN

This page intentionally left blank

CHAPTER 1

The Concepts of Power Analysis

The power of a statistical test is the probability that it will yield statistically significant results. Since statistical significance is so earnestly sought and devoutly wished for by behavioral scientists, one would think that the *a priori* probability of its accomplishment would be routinely determined and well understood. Quite surprisingly, this is not the case. Instead, if we take as evidence the research literature, we find that statistical power is only infrequently understood and almost never determined. The immediate reason for this is not hard to discern—the applied statistics textbooks aimed at behavioral scientists, with few exceptions, give it scant attention.

The purpose of this book is to provide a self-contained comprehensive treatment of statistical power analysis from an "applied" viewpoint. The purpose of this chapter is to present the basic conceptual framework of statistical hypothesis testing, giving emphasis to power, followed by the framework within which this book is organized.

1.1 General Introduction

When the behavioral scientist has occasion to don the mantle of the applied statistician, the probability is high that it will be for the purpose of testing one or more null hypotheses, i.e., "the hypothesis that the phenomenon to be demonstrated is in fact absent [Fisher, 1949, p. 13]." Not that he hopes to "prove" this hypothesis. On the contrary, he typically hopes to "reject" this hypothesis and thus "prove" that the phenomenon in question is in fact present.

Let us acknowledge at the outset the necessarily probabilistic character of statistical inference, and dispense with the mocking quotation marks

about words like *reject* and *prove*. This may be done by requiring that an investigator set certain appropriate probability standards for research results which provide a basis for rejection of the null hypothesis and hence for the proof of the existence of the phenomenon under test. Results from a random sample drawn from a population will only approximate the characteristics of the population. Therefore, even if the null hypothesis is, in fact, true, a given sample result is not expected to mirror this fact exactly. Before sample data are gathered, therefore, the investigator working in the Fisherian framework selects some prudently small value **a** (say .01 or .05), so that he *may* eventually be able to say about his sample data, "*If the null hypothesis is true,* the probability of the obtained sample result is no more than **a**," i.e. a statistically significant result. *If* he can make this statement, since **a** is small, he is said to have rejected the null hypothesis "with an **a** significance criterion" or "at the **a** significance level." If, on the other hand, he finds the probability to be greater than **a**, he cannot make the above statement and he has failed to reject the null hypothesis, or, equivalently finds it "tenable," or "accepts" it, all at the **a** significance level.

We have thus isolated one element of this form of statistical inference, the standard of proof that the phenomenon exists, or, equivalently, the standard of disproof of the null hypothesis that states that the phenomenon does not exist.

Another component of the significance criterion concerns the exact definition of the nature of the phenomenon's existence. This depends on the details of how the phenomenon is manifested and statistically tested, e.g., the directionality/nondirectionality ("one tailed"/"two tailed") of the statement of the alternative to the null hypothesis.[1] When, for example, the investigator is working in a context of comparing some parameter (e.g., mean, proportion, correlation coefficient) for two populations A and B, he can define the existence of the phenomenon in two different ways:

1. The phenomenon is taken to exist if the parameters of A and B differ. No direction of the difference, such as A larger than B, is specified, so that departures in either direction from the null hypothesis constitute evidence against it. Because either tail of the sampling distribution of differences may contribute to **a**, this is usually called a two-tailed or two-sided test.

2. The phenomenon is taken to exist only if the parameters of A and B differ in a direction specified in advance, e.g., A larger than B. In this

[1] Some statistical tests, particularly those involving comparisons of more than two populations, are naturally nondirectional. In what immediately follows, we consider those tests which contrast two populations, wherein the experimenter ordinarily explicitly chooses between a directional and nondirectional statement of his alternate hypothesis. See below, Chapters 7 and 8.

circumstance, departures from the null hypothesis only in the direction specified constitute evidence against it. Because only one tail of the sampling distribution of differences may contribute to **a**, this is usually called a one-tailed or one-sided test.

It is convenient to conceive of the significance criterion as embodying both the probability of falsely rejecting the null hypothesis, **a**, and the "sidedness" of the definition of the existence of the phenomenon (when relevant). Thus, the significance criterion on a two-tailed test of the null hypothesis at the .05 significance level, which will be symbolized as $\mathbf{a}_2 = .05$, says two things: (*a*) that the phenomenon whose existence is at issue is understood to be manifested by any difference between the two populations' parameter values, and (*b*) that the standard of proof is a sample result that would occur less than 5% of the time if the null hypothesis is true. Similarly, a prior specification defining the phenomenon under study as that for which the parameter value for A is larger than that of B (i.e., one-tailed) and the probability of falsely rejecting the null is set at .10 would be symbolized as a significance criterion of $\mathbf{a}_1 = .10$. The combination of the probability and the sidedness of the test into a single entity, the significance criterion, is convenient because this combination defines in advance the "critical region," i.e., the range of values of the outcome which leads to rejection of the null hypothesis and, perforce, the range of values which leads to its nonrejection. Thus, when an investigator plans a statistical test at some given significance criterion, say $\mathbf{a}_1 = .10$, he has effected a specific division of all the possible results of his study into those which will lead him to conclude that the phenomenon exists (with risk **a** no greater than .10 and a one-sided definition of the phenomenon) and those which will not make possible that conclusion.[2]

The above review of the logic of classical statistical inference reduces to a null hypothesis and a significance criterion which defines the circumstances which will lead to its rejection or nonrejection. Observe that the significance criterion embodies the risk of mistakenly rejecting a null hypothesis. The entire discussion above is conditional on the truth of the null hypothesis.

But what if, indeed, the phenomenon *does* exist and the null hypothesis is *false*? This is the usual expectation of the investigator, who has stated the null hypothesis for tactical purposes so that he may reject it and conclude that the phenomenon exists. But, of course, the fact that the phenomenon exists in the population far from guarantees a statistically significant result,

[2] The author has elsewhere expressed serious reservations about the use of directional tests in psychological research in all but relatively limited circumstances (Cohen, 1965). The bases for these reservations would extend to other regions of behavioral science. These tests are however of undoubted statistical validity and in common use, so he has made full provision for them in this work.

i.e., one which warrants the conclusion that it exists, for this conclusion depends upon meeting the agreed-upon standard of proof (i.e., significance criterion). It is at this point that the concept of statistical power must be considered.

The power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis, i.e., the probability that it will result in the conclusion that the phenomenon exists. Given the characteristics of a specific statistical test of the null hypothesis and the state of affairs in the population, the power of the test can be determined. It clearly represents a vital piece of information about a statistical test applied to research data (cf. Cohen, 1962). For example, the discovery, during the planning phase of an investigation, that the power of the eventual statistical test is low should lead to a revision in the plans. As another example, consider a completed experiment which led to nonrejection of the null hypothesis. An analysis which finds that the power was low should lead one to regard the negative results as ambiguous, since failure to reject the null hypothesis cannot have much substantive meaning when, even though the phenomenon exists (to some given degree), the *a priori* probability of rejecting the null hypothesis was low. A detailed consideration of the use of power analysis in planning investigations and assessing completed investigations is reserved for later sections.

The power of a statistical test depends upon three parameters: the significance criterion, the reliability of the sample results, and the "effect size," that is, the *degree* to which the phenomenon exists.

1.2 Significance Criterion

The role of this parameter in testing null hypotheses has already been given some consideration. As noted above, the significance criterion represents the standard of proof that the phenomenon exists, or the risk of mistakenly rejecting the null hypothesis. As used here, it directly implies the "critical region of rejection" of the null hypothesis, since it embodies both the probability of a class of results given that the null hypothesis is true (**a**), as well as the definition of the phenomenon's existence with regard to directionality.

The significance level, **a**, has been variously called the error of the first kind, the Type I error, and the alpha error. Since it is the rate of rejecting a true null hypothesis, it is taken as a relatively small value. It follows then that the smaller the value, the more rigorous the standard of null hypothesis rejection or, equivalently, of proof of the phenomenon's existence. Assume that a phenomenon exists in the population to some given degree. Other things equal, the more stringent the standard for proof, i.e., the lower the value of **a**, the poorer the chances are that the sample will provide results

which meet this standard, i.e., the lower the power. Concretely, if an investigator is prepared to run only a 1% risk of false rejection of the null hypothesis, the probability of his data meeting this standard is lower than would be the case were he prepared to use the less stringent standard of a 10% risk of false rejection.

The practice of taking **a** very small ("the smaller the better") then results in power values being relatively small. However, the complement of the power (1 − power), here symbolized as **b**, is also error, called Type II or beta error, since it represents the "error" rate of failing to reject a false null hypothesis. Thus it is seen that statistical inference can be viewed as weighing, in a manner relevant to the substantive issues of an investigation, these two kinds of errors. An investigator can set the risk of false null hypothesis rejection at a vanishingly small level, say **a** = .001, but in so doing, he may reduce the power of his test to .10 (hence beta error probability, **b**, is 1 − .10 = .90). Two comments may be made here:

1. The general neglect of issues of statistical power in behavioral science may well result, in such instances, in the investigator's failing to realize that the **a** = .001 value leads in his situation to power = .10, **b** = .90 (Cohen, 1962). Presumably, although not necessarily, such a realization would lead to a revision of experimental plans, including possibly an upward revision of the **a** level to increase power.

2. If the investigator proceeds as originally planned, he implies a conception of the relative seriousness of Type I to Type II error (risk of false null rejection to risk of false null acceptance) of **b**/**a** = .90/.001 = 900 to 1, i.e., he implicitly believes that mistakenly rejecting the null hypothesis under the assumed conditions is 900 times more serious than mistakenly accepting it. In another situation, with **a** = .05, power = .80, and hence **b** = 1 − .80 = .20, the relative seriousness of Type I to Type II error is **b**/**a** = .20/.05 = 4 to 1; thus mistaken rejection of the null hypothesis is considered four times as serious as mistaken acceptance.

The directionality of the significance criterion (left unspecified in the above examples) also bears on the power of a statistical test. When the null hypothesis can be rejected in *either* direction so that the critical significance region is in *both* tails of the sampling distribution of the test statistic (e.g., a **t** ratio), the resulting test will have less power than a test at the same **a** level which is directional, *provided that* the sample result is in the direction predicted. Since directional tests cannot, by definition, lead to rejecting the null hypothesis in the direction *opposite* to that predicted, these tests have no power to detect such effects. When the experimental results are in the predicted direction, all other things equal, a test at level $\mathbf{a}_1$ will have power equal for all practical purposes to a test at $2\mathbf{a}_2$.

Concretely, if an experiment is performed to detect a difference between the means of populations A and B, say $\mathbf{m}_A$ and $\mathbf{m}_B$, in *either* direction at the $\mathbf{a}_2 = .05$ significance criterion, under given conditions, the test will have a certain power. If, instead, an anticipation of $\mathbf{m}_A$ greater than $\mathbf{m}_B$ leads to a test at $\mathbf{a}_1 = .05$, this test will have power approximately equal to a two-tailed test with $\mathbf{a}_2 = .10$, hence greater power than the test at $\mathbf{a}_2 = .05$, provided that in fact $\mathbf{m}_A$ is greater than $\mathbf{m}_B$. If $\mathbf{m}_B$ is greater than $\mathbf{m}_A$, the test at $\mathbf{a}_1 = .05$ has *no* power, since that conclusion is inadmissible. The temptation to perform directional tests because of their greater power at the same **a** level should be tempered by the realization that they preclude finding results opposite to those anticipated. There are occasional circumstances where the nature of the decision is such that the investigator does not need to know about effects in the opposite direction. For example, he will take a certain course of action if $\mathbf{m}_A$ is greater than $\mathbf{m}_B$ and not otherwise. If otherwise, he does not need to distinguish between their equality and $\mathbf{m}_B$ greater than $\mathbf{m}_A$. In such infrequent instances, one-tailed tests are appropriate (Cohen, 1965, pp. 106–111).

In the tables in this book, provision is made for tests at the .01, .05, and .10 significance levels. Where a statistical test may ordinarily be performed either nondirectionally or directionally, both $\mathbf{a}_2$ and $\mathbf{a}_1$ tables are provided. Since power for $\mathbf{a}_1 = .05$ is virtually identical with power for $\mathbf{a}_2 = .10$, a single power table suffices. Similarly, tables for $\mathbf{a}_1 = .01$ provide values for $\mathbf{a}_2 = .02$, and tables for $\mathbf{a}_1 = .10$ values for $\mathbf{a}_2 = .20$; also, tables for $\mathbf{a}_2 = .01$ provide values for $\mathbf{a}_1 = .005$, tables at $\mathbf{a}_2 = .05$ provide values for $\mathbf{a}_1 = .025$.

1.3 Reliability of Sample Results and Sample Size

The reliability (or precision) of a sample value is the closeness with which it can be expected to approximate the relevant population value. It is necessarily an estimated value in practice, since the population value is generally unknown. Depending upon the statistic in question, and the specific statistical model on which the test is based, reliability may or may not be directly dependent upon the unit of measurement, the population value, and the shape of the population distribution. However, it is *always* dependent upon the size of the sample.

For example, one conventional means for assessing the reliability of a statistic is the standard error (SE) of the statistic. If we consider the arithmetic mean of a variable $\mathbf{X}$ ($\overline{\mathbf{X}}$), its reliability may be estimated by the standard error of the mean,

$$\mathrm{SE}_{\overline{\mathbf{X}}} = \sqrt{\frac{\mathbf{s}^2}{\mathbf{n}}},$$

where $\mathbf{s}^2$ is the usual unbiased estimate (from the random sample) of the