How Many Subjects?

Second Edition

Statistical Power Analysis in Research

Helena Chmura Kraemer
Christine Blasey

How Many Subjects?

Statistical Power Analysis in Research

Second Edition

Helena Chmura Kraemer
Stanford University

Christine Blasey
Palo Alto University

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Printed in the United States of America

Cataloging-in-publication data is available for this title from the Library of Congress.

ISBN 978-1-4833-1954-4

This book is printed on acid-free paper.

Acquisitions Editor: Helen Salmon
Assistant Editor: Katie Guarino
Editorial Assistant: Anna Villarruel
Production Editor: Laura Barrett
Copy Editor: Karin Rathert
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Jennifer Grubba
Indexer: Jennifer Pairan
Cover Designer: Anupama Krishnan
Marketing Manager: Nicole Elliot

15 16 17 18 19 10 9 8 7 6 5 4 3 2 1

How Many Subjects?

Second Edition

SAGE was founded in 1965 by Sara Miller McCune to support the dissemination of usable knowledge by publishing innovative and high-quality research and teaching content. Today, we publish more than 750 journals, including those of more than 300 learned societies, more than 800 new books per year, and a growing range of library products including archives, data, case studies, reports, conference highlights, and video. SAGE remains majority-owned by our founder, and after Sara's lifetime will become owned by a charitable trust that secures our continued independence.

Los Angeles | London | Washington DC | New Delhi | Singapore | Boston

List of Greek Symbols

Greek letters are here used to indicate population values that are sometimes estimated in the samples. It is useful to keep the distinction between an unknown population parameter and a sample of researcher-selected values.

Use	Greek Letter	Population Value	Sample Estimate
alpha	alpha	α	
regression parameter	beta	β	b
chi-square test statistic or distribution	chi-square	χ^2	
delta	delta lower case	δ	d
indicator of effect size/ design parameters	delta upper case	Δ	
error term	epsilon	ε	
lambda	lambda	λ	
mean	mu	μ	$\bar{x}$
indicator of necessary sample size	nu	ν	
standard normal cumulative distribution	phi upper case	Φ	
proportion	pi	π, $\pi'=1-\pi$	p, $q=1-p$
correlation coefficient	rho	ρ	r
standard deviation	sigma	σ	s
variance	sigma squared	σ^2	s^2

Preface to the Second Edition

In the preface to the first edition of this book (1987), we acknowledged the valuable contributions of Jacob Cohen[1], whose book on statistical power, even today, some 20 years after his death, remains arguably the best text on that issue. His influence on statistical methods in both psychology and medicine, particularly his efforts to emphasize the crucial role that statistical power plays on success in testing research hypotheses, remains strong. The first edition of this book was written on his suggestion, in an effort to simplify the application of statistical power.

Shortly before his death, I (HCK) met Jack face-to-face for the first time at a meeting of a subcommittee of the American Psychological Association that was jocularly called "The committee to ban the *p*-value." The committee did not actually recommend banning the *p*-value but emphasized how poorly statistical hypothesis testing was often done; how often *p*-values were overused, misused, and abused; and made recommendations to help repair the situation (summarized in the Wilkinson et al. paper[2]). In a side conversation with me, Jack lamented that his book on power and mine had probably done more harm than good in promoting better-designed studies. He pointed out how often those proposing research projects did power computations for one test but used another or did multiple power computations, apparently just to show they knew how to do such computations, that had no relevance to the design of the study proposed or simply misused the entire concept of power, often citing our books. Before our books were available, he suggested, researchers used common sense and knew, for example, that one could not draw valid inferences about heterogeneous populations with a sample size of 20. After our books became available, the same researchers would incorrectly use power calculations to propose a sample size of 20, citing our books for justification.

After my initial shock, I reluctantly came to agree with his assessment and long puzzled why this would be so. I finally came to the realization that we had put the primary, almost exclusive, emphasis on calculation of power, but calculation of power is meaningless and even misleading unless it is done within

the proper context of statistical hypothesis testing and, even further, statistical hypothesis testing is meaningless and even misleading unless it is done within the proper context of the scientific method. Moreover, not only our books on power but, what is worse, many courses in statistics given to researchers in training put the primary, almost exclusive, emphasis on calculation. Students learn to "do" a two-sample t-test, a regression analysis, a 2 by 2 chi-square test, and so forth, with very little understanding of when each such test is appropriate or not.

This is like teaching a child to play baseball by teaching him to bat, throw, run, and field but never exposing that child to the rules of baseball or allowing him to interact with eight other players in an actual game. If, on the basis of such training, one were to ask that child to play in a game, he's not likely to do very well. Having all the basic skills without knowledge of the context or interactions necessary to the game will not serve him well. In the same way, teaching statistical test calculations in absence of knowledge of the "rules" of the scientific method or the "interaction" with sampling, design, and measurement issues does not equip researchers well to do valid and powerful scientific research. The problems are particularly salient in biobehavioral research, where research is done using living (human or animal) subjects rather than tissue samples or chemical reactions.

Consequently in this second edition, we begin with two new chapters, updating and replacing the original Chapters 1 (Introduction) and 2 (General Concepts) with one chapter setting statistical hypothesis-testing in the context of the scientific method as applied to biobehavioral research and another spelling out all the components of statistical hypothesis testing and where power considerations play a role. In these chapters, special attention is paid to the most common mistakes that we see in reviewing proposal or paper submissions or in publications.

The organization of the remaining chapters parallels those in the first edition, but often with new examples, with reference, where appropriate, to contextual issues, and with addition, where appropriate, to mistakes often made.

What we haven't changed are the tables—one set of tables that can be used for a variety of different common tests by modifying the relationship of the effect size, design parameters, and sample size to the row and column definitions. While there are many ways to implement power computations, including tables, nomograms, and computer packages, when times come for "what if" thinking in designing a study, being able to compare power with different designs all referring to the same set of tables still seems the easiest. However, even if users find they prefer using different tables, nomograms, or computer packages to do power computations (to be honest, both of the present authors use these tables for teaching but not usually in designing studies), the logic of the materials in each chapter will, we think, serve them well in using such alternative methods.

One reviewer of the first edition of this book suggested that readers borrow a library copy of the book, read through the book, then copy out the dozen pages of tables, in which case they would need not to buy the book. The reviewer was right. We hope that with the second edition many will note that the text is valuable whether or not one uses the tables.

In his foreword to the first edition, the late Victor H. Denenberg, then at the University of Connecticut, commented, "If this book only presented the reader with a straight-forward set of procedures for determining N (sample size) for any particular research design, it would have fulfilled its mission successfully. But the book does more. In the course of discussing different designs, the authors make note of important points that are of value to the empirical researcher. These include: the conditions under which a repeated measures design will be more or less efficient than a cross-sectional design; the considerations involved in deciding to match or stratify subjects; the selection of variables for a multiple regression analysis; the value of equal (or near equal) N in analysis of variance designs; how to insure, in a correlational study, that the study will be valid; and the N required to make a reasonably rigorous test of one hypothesis using the chi-square technique." In short, one of the advantages of our approach is what would correspond to "differential diagnosis" in medicine, the process here of sifting through various valid options available for testing a hypothesis and choosing the one most likely to succeed. Even greater stress will be placed on such "differential design diagnosis" in this edition.

We wish to thank the many researchers at Stanford University (particularly, but not exclusively, those in the Department of Psychiatry and Behavioral Sciences), at the University of Pittsburgh (Department of Psychiatry), and the researchers at other universities and in research organizations with whom we have worked over the years, who have made us aware of the importance of cost-effective research and the challenges inherent in trying to produce such research. We continue to acknowledge the influence of Jacob Cohen and of all those who read, commented on, and often criticized the limitations of the first edition. We would also especially acknowledge and appreciate the contribution of Sue Thiemann, who coauthored the first edition. Without her contributions to the first edition, clearly, this second edition would be impossible.

REFERENCES

1. Cohen J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.
2. Wilkinson, L. and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist. 54*, 594–604.

Acknowledgments

SAGE and the authors gratefully acknowledge feedback from the following reviewers:

- Glen C. Gamst, University of La Verne
- Bryan Rooney, Concordia University College of Alberta
- Richard E. Strauss, Texas Tech University
- Rebecca Warner, University of New Hampshire

About the Authors

Helena Chmura Kraemer received her bachelor's degree in Mathematics from Smith College (Summa cum Laude, 1958), did her first year of graduate study in Statistics as a Fulbright scholar at the University of Manchester, England, and then completed her doctoral studies in the Department of Statistics, Stanford University (1963). She joined the Department of Psychiatry and Behavioral Sciences at Stanford in 1964. Her primary interests concern the applications of biostatistics in the behavioral areas of medicine. In 1964, that seemed largely concentrated in Psychiatry, but in the years since, she has worked in Cardiology, Pediatrics, Radiology, Oncology, etc., as behavioral issues have become more prominent in all areas of medicine. She is a Fellow of the American Statistical Association, and of the American College of Neuropsychopharmacology. She was elected a member of the Institute of Medicine, Academy of Sciences, in 2003. She was also the recipient of the Harvard Prize in Psychiatric Biostatistics and Epidemiology in 2001, and Andrew C. Leon Distinguished Career Award, ISCTM, 2014, and a Honorary Doctor of Science, Wesleyan University, 2014.

She has published more than 300 papers in peer-reviewed journals, numerous chapters in books, and 6 books. At various times, she has served as associate editor or on the editorial boards of, for example, *Statistics in Medicine*, *Psychological Methods*, *Archives of General Psychiatry*, *Medical Decision Making*, and is a frequent reviewer for journals in Statistics, Psychiatry, and other fields of medicine.

Over the years before retirement in 2007, she mentored many young investigators both at Stanford, Pittsburgh, and other universities, providing training in research methods, as well as consultation on their proposals.

Her major current research interests concern the use of statistical methods in risk research, specifically the focus on moderators and mediators, the use of effect sizes to indicate clinical or practical significance to replace the overuse and abuse of statistical significance, and, in general, identifying and trying to

rectify common problems in the application of statistical methods in medicine. She became Emerita in 2007, but continued to be active, serving on the NIMH council until 2008, and on the DSM-5 Task Force until 2012.

Christine Blasey is a Professor of Psychology at Palo Alto University and a Research Psychologist at Stanford University School of Medicine Department of Psychiatry and Behavioral Sciences. Christine received a Ph.D. in Psychology from the University of Southern California and an M.S. in Epidemiology from Stanford University. Christine provides statistical consultation in academic settings (e.g., Departments of Psychiatry, Cardiovascular Medicine, Education) and in the private sector for pharmaceutical companies testing new medicines and medical devices. Christine's consultation area of expertise is the interaction between pharmaceutical companies and the United States Food and Drug Administration (FDA). She has co-authored over fifty peer-reviewed journal articles and book chapters and serves as a statistical reviewer for several psychology and psychiatry journals. Christine teaches statistics, research methods, and psychometrics in the PGSP-Stanford University Consortium for Clinical Psychology. Her primary interest is mentoring future psychologists.

Table of Contents

1

The "Rules of the Game"

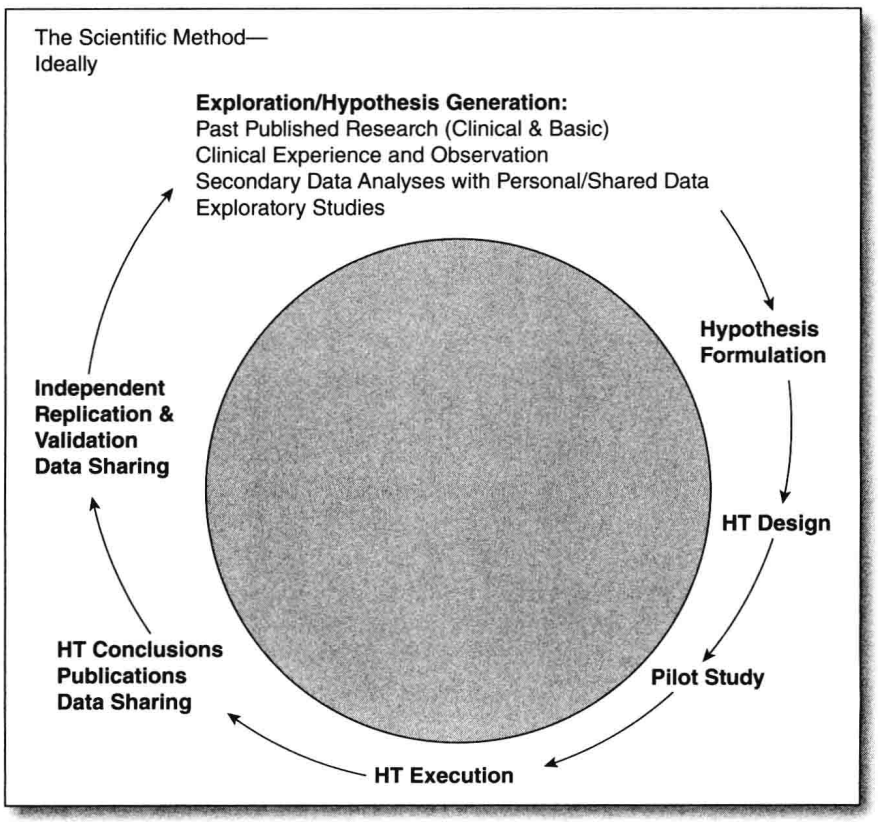
The "game" of interest here is the use of the scientific method to establish scientific facts, at least as it might be applied in biobehavioral research. According to the Oxford English Dictionary, the scientific method is defined as "a method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses." The focus here is on the testing of hypotheses using the methods of statistical hypothesis testing that have been in use since early in the 20th century, but statistical hypothesis testing can only be successful within the framework of the scientific method in general. One view of the scientific method, as applied to biobehavioral research, is shown in Figure 1.0.

1.1 Exploratory Studies

Every study begins with exploration. Exploratory studies include review of the relevant literature, consideration of theories current in the relevant field, incorporation of clinical experiences and observations, secondary data analysis of data from earlier studies, and when one is close to the cutting edge of science and little is known about the relevant field, perhaps even research studies designed and executed specifically for exploration. Exploratory studies are efforts to find out what is going on in a particular area and are not designed to address specific a priori hypotheses. Usually there is no data analytic plan—the analyses done are inspired and guided by data. It is not unusual that one analyzes data in a variety of different ways, trying out different models and approaches. It is not unusual that one sees patterns that are completely unexpected that inspire ideas completely different from those that initiated the study.

What emerges from exploratory studies are not conclusions. The primary goal of such exploratory studies is to generate the theoretical rationale and the

Figure 1.0 The Process of the Scientific Method



HT = Hypothesis Testing

empirical justification for proposing a certain hypothesis to be tested in a subsequent hypothesis-testing study designed for that purpose. From an exploratory study, there should be enough evidence to make it reasonable that the hypothesis proposed is true, and if true, of some importance. But there should not be enough evidence to assure its truth. This balance is often called *equipoise* (Freedman, 1987).

Equipoise is crucial, particularly in dealing with human subjects in research studies. Clinical equipoise is often presented as an ethical issue, because, regardless of the research question, participation in a research study places a burden on human subjects, can, in some cases, endanger their health and well-being, and wastes time and money. However, equipoise is also a scientific issue. It is very difficult for a researcher who already "knows" what the "right" answer to a research question should be to design, execute, analyze, and interpret the results without a bias generated by that "knowledge."