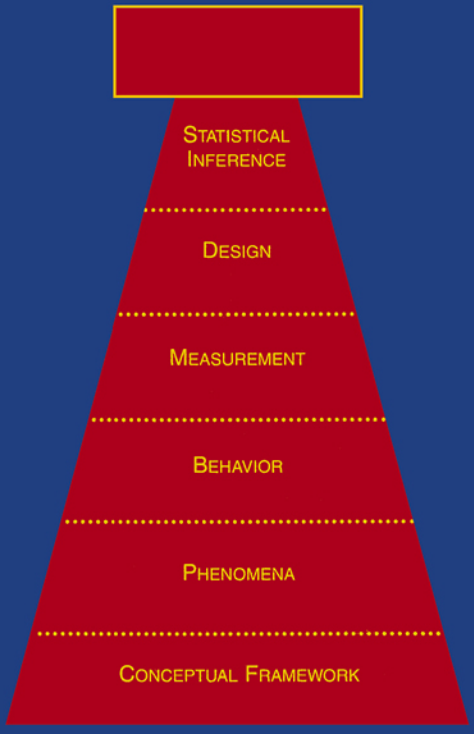
Empirical Direction in Design and Analysis

SCIENTIFIC PSYCHOLOGY SERIES



Norman H. Anderson

Empirical Direction in Design and Analysis

This page intentionally left blank

Empirical Direction in Design and Analysis

Norman H.Anderson
University of California, San Diego

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2004 by
Lawrence Erlbaum Associates, Inc.

Published 2014 by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
711 Third Avenue, New York, NY, 10017, USA

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

President/CEO:	Lawrence Erlbaum
Executive Vice-President, Marketing:	Joseph Petrowski
Senior Vice-President, Book Production:	Art Lizza
Director, Editorial:	Lane Akers
Director, Sales and Marketing:	Robert Sidor
Director, Customer Relations:	Nancy Seitz
Senior Editor:	Debra Riegert
Textbook Marketing Manager:	Marisol Kozlovski
Editorial Assistant:	Jason Planer
Cover Design:	Kathryn Houghtaling Lacey
Textbook Production Manager:	Paul Smolenski
Text and Cover Printer:	Hamilton Printing Company

Library of Congress Cataloging-in-Publication Data

Anderson, Norman H.
Empirical direction in design and analysis/Norman H.Anderson
p. cm.
Includes bibliographical references and index.
1. Psychometrics. 2. Psychology—Methodology. I. Title.

BF39 .A49 2001
150'.1'5195–dc21 2001033377
CIP

ISBN 13: 978-0-805-83978-4 (hbk)
ISBN 13: 978-0-805-84083-4 (pbk)

DEDICATION

This book is dedicated to the many students who have been in my courses for the last 40 years. You have taught me a lot; I deeply appreciate this unique learning experience. I hope your lives have been rewarding. This book seeks to continue as coach and aide, passing on to new students what we together have learned.

This book has benefited from comments of numerous persons. I owe special thanks to Richard Bogartz, Edward Karpp, Pamela Moses, Charles Reichardt, Anne Schlottmann, Ewart Thomas, and James Zalinski. Many others have made helpful comments on various issues and sections. Among these are James Alexander, Gwendolyn Alexander, James Anderson, Mark Appelbaum, Margaret Armstrong, Ann Norman Atkinson, Rita Atkinson, Eileen Beier, Michael Birnbaum, Donnie Bocko, Lyle Bourne, Gordon Bower, Clifford Butzin, Robert Calfee, Edward Carr, Jenny Cantor, John Clavadetscher, Diane Cuneo, Claire Ernhart, Robert Farber, Arthur Farkas, Philip Gallo, William Gaver, Anthony Greenwald, Reid Hastie, Wilfried Hommers, Stephen Hubert, James Jaccard, Beth Jaworski, Martin Kaplan, Eileen Karsh, Lucille Kirsch, James Kulik, Andrius Kulikauskas, Anita Lampel, Irwin Levin, Stephen Link, Frank Logan, Lola Lopes, Jordan Louviere, Tracy Love-Geffen, Donald MacLeod, Irving Maltzman, George Mandler, Jean Mandler, Sergio Masin, Mark McDaniel, Jennifer McDowell, William McGill, William McGuire, Craig McKenzie, Colleen Surber Moore, Philip Moore, Etienne Mullet, Gregg Oden, Allen Osman, Allen Parducci, Mary Pendery, Joan Prentice, Mike Rinck, Maria Teresa Sastre, Shlomo Sawilowsky, Sandra Scarr, Laura Schreibman, John Shaughnessy, Juliet Shaffer, James Shanteau, Ling-Po Shiu, Ramadhar Singh, Cheryl Graesser Stecher, Jeff Steinberg, Saul Sternberg, Billy Vaughn, John Verdi, Mingshen Wang, David Weiss, Ben Williams, Wendy Williams, Friedrich Wilkening, John Wixted, Yuval Wolf, Chungfang Yang, Gregory Zarow, and Shu-Hong Zhu.

FOREWORD

Statistics should be an organic component of substantive investigation. This is how statistics should be learned—and how it should be taught.

A text should aim to give students what they will later need to know. What students will later need to know is how to utilize statistics in their empirical work. To get such transfer requires that statistics be embedded within a framework of substantive inquiry.

Substantive investigation rests on *extrastatistical inference*—substantive considerations concerning validity of task-procedure and generality of results. Practical understanding—transfer to empirical analysis—requires that statistics be integrated into a larger framework of extrastatistical inference.

This extrastatistical theme is embodied in the *Experimental Pyramid* of Figure 1.1. The six levels of the Pyramid portray a hierarchy of considerations involved in empirical investigation. Statistics, to be effective, needs to be integrated into the substantive considerations at each level of the Pyramid.

The main value of statistics is in planning the investigation, long before the data are collected. Contrary to the standard stereotype, the main function of statistics is to get more information *into* the data.

Current texts pursue two largely incompatible goals: To be a text for first-year graduate students and to serve as a reference handbook for advanced researchers. Both audiences suffer thereby, especially first-year students. Facing a plethora of formulas, uncertain which are basic, doing exercises largely devoted to numerical calculations, first-year students are hindered and side-tracked from developing understanding and research judgment.

Chapters 1–12 present a core intended for first-year graduate students. Far fewer formulas are presented than in other texts, which is intended to facilitate conceptual and empirical understanding. Two novelties are the heavy emphasis on confidence intervals and the separate chapters on confounding and single subject design.

Chapters 13–21 serve in part a reference handbook function, for they take up more specialized topics: within-versus-between design, Latin squares, multiple regression, analysis of covariance, quasi-experimental design, multiple comparisons, and the difficult problem of measuring effect size. Also included are chapters on the foundations of statistics, on mathematical models in psychology, and on psychological measurement theory. These chapters aim to give conceptual understanding that will facilitate empirical analysis.

The "Empirical Direction" in this book is not essentially new. It is a return to and unification of statistics with the extrastatistical nature of empirical science. Further discussion is given in Chapter 23, *Lifelong Learning.*

CONTENTS

PREFACE

Generality is a prime goal of scientific inference because science depends on evidence from **samples**. In psychology, experimental results are typically obtained from a small sample of subjects tested in one narrow experimental situation, with a very small sample of stimulus conditions and usually a single task and a single measure of behavior. These results have value only to the extent that they generalize. No one is interested, for example, in those particular infant monkeys in Harlow's experiments on mother love; their behavior is of interest only insofar as it generalizes to other infants, especially human infants, across a wider range of test situations than Harlow used. The four kinds of generality implicit in the previous sentence are discussed on pages 20–24.

Reliability, or replicability, is one aspect of generality. Different samples of subjects will yield different results. Perhaps the effect observed in your sample is merely a chance accident of which subjects chanced to get into your particular sample. Any claim for a real effect should be prefaced by evidence that it is reliable—not likely to be produced by chance alone. Statistics can help assess reliability. Not less important, statistics can help you plan your experiments to get more reliability for less cost (pages 16–20).

Validity, which is far more important than reliability, is primarily an extrastatistical issue—which must be answered in terms of substantive knowledge. The ubiquitous threat to validity is **confounding**. Confounding arises because the experimenter employs some concrete stimulus manipulation that is intended to elicit a specified process. But this manipulation may also elicit some other process that undercuts the interpretation. The classic example of confounding is the placebo effect, in which the suggestion produced by giving a medicine has beneficial effects even though the medicine itself is worthless. Validity is more complex, however, as discussed on pages 8–16.

Generality, reliability, and validity are mainly extrastatistical problems. Statistics can furnish valuable assistance with some aspects of these problems, but effective use of statistics depends on integration of statistics with extrastatistical, empirical knowledge.

Six levels of knowledge are distinguished in the **Experimental Pyramid** (page 3). Each lower level is more important. Statistics, although mainly applicable at the top level, can also help at each lower level. Your labors will be more productive the more you learn how to integrate statistical inference into an empirical framework of extrastatistical, scientific inference. This **empirical direction** is the main theme of this initial chapter and of the entire book.

Chapter 1

SCIENTIFIC INFERENCE

Statistics teaching should be integrated with empirical substance. Statistics is not varnish, to be applied after the experiment is done. Statistics is an integral component of the research plan—beginning with choice of problem for investigation. Statistics is important at every intermediate level, including apparatus and procedure, and concluding in the interpretation of the results.

Students almost inevitably come to view the significance test as the be-all and end-all. This test condenses the entire study into a single number of pivotal importance for making claims about what the study shows. A positive answer to the question, "Is it significant?" thus comes to be considered the ultimate goal. *"It is significant!"* seems all the more potent because "significant" carries undertones of everyday meaning.

The essential question, of course, is what the results mean. The significance test has the necessary—but minor—function of providing evidence of whether there is a result to interpret. What this result may mean depends on considerations at deeper levels.

The more important functions of statistics are in these deeper levels. These more important functions apply at the planning stage, before the data are collected. These more important functions condition the meaning and interpretation of any result that may be obtained. These more important functions of statistics need to be understood in organic relation to the substantive inquiry.

The significance test, in contrast, applies after the data have been collected. It is then too late to correct missed opportunities and shortcomings in the research plan. Finding a significant decrease in felt pain may not be worth much if the placebo control was overlooked. The placebo control may not be valid if the treatment was not "blind." And even the most careful procedure may founder if you or your research assistant stumbled with the random assignment. These three

examples illustrate vital functions of statistics that operate in the planning stage, before the data are collected. This is where statistics is most needed—and most effective.

Some writers argue strenuously against using significance tests. There is something to be said for their argument; it would force people to look more closely at their data. Many statisticians voice similar complaints about the fixation on significance tests and seek to emphasize the more important functions of statistics by avoiding that term and speaking of experimental design, data analysis, and so forth. The significance test has an essential role, however, a role that it performs reasonably well.

The real difficulty is how to integrate statistics-design with empirical inquiry. A very modest number of statistical ideas and formulas will cover most situations that arise in experimental research, as this book will show. What is not so easy is to develop the research judgment to integrate statistical considerations into the planning stage of an experiment. A conceptual framework that puts statistics in its proper place—as an aide to scientific inference—is given by the Experimental Pyramid taken up next.

1.1 EXPERIMENTAL PYRAMID

Scientific inference is central to empirical research. Our empirical observations are clues to deeper reality. How we make inferences from these clues may be considered within the *Experimental Pyramid* of Figure 1.1. Each level of the Experimental Pyramid corresponds to different aspects of empirical investigation. These levels range from statistical inference and experimental design at the top to the conceptual framework at the bottom.

The lower levels are more fundamental. Validity at lower levels is prerequisite to validity at higher levels. The different levels are not separate and distinct, as the dashed lines might suggest; all levels interrelate as facets of an organic whole. Each following section comments briefly on one level of the Experimental Pyramid.

1.1.1 STATISTICAL INFERENCE

The significance test is a minor concern of scientific investigation. It is needed as evidence whether the result you observe is real, rather than chance. Unless you have reasonable evidence that your result is real, there is little point in trying to decide what it means. Before expounding your result, therefore, you owe it to your readers, and to yourself, to show that you have something to expound. This is why the significance test is ubiquitous.

Despite its ubiquity, the significance test is only a minor aspect of substantive inference. What your result means depends on substantive considerations: What experimental task you choose, what response measure you use, your control

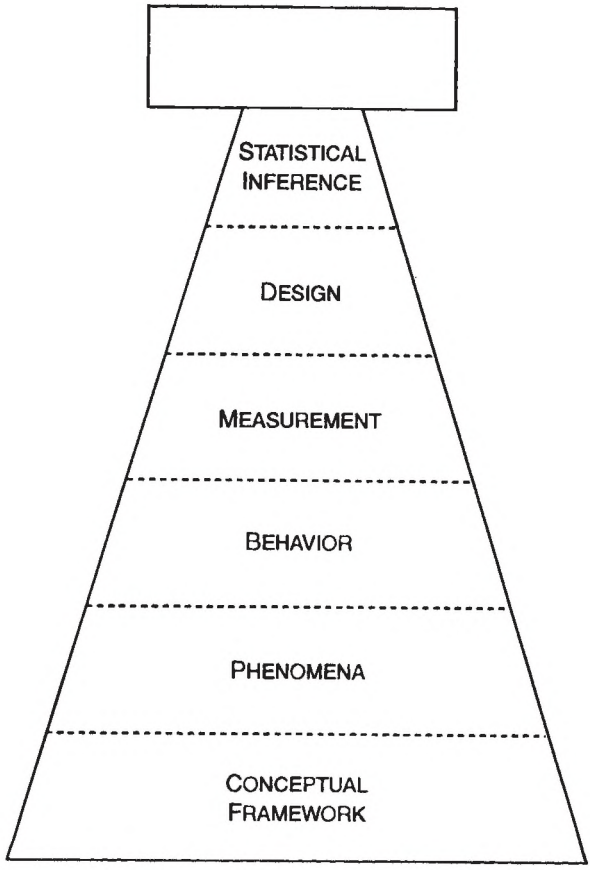


Figure 1.1. Experimental Pyramid

conditions, and so forth. Such substantive inference depends on considerations at more basic levels of the Pyramid.

Even within statistical inference, the significance test has a minor function. More important functions of statistical analysis concern other aspects of *reliability,* especially confidence intervals that describe a sample mean as an interval of likely error. Even more important is *validity.* When measuring each subject under multiple conditions, for example, statistical inference is essential for dealing with sticky problems of practice, adaptation, and transfer from initial conditions that confound the response to conditions that follow.

Validity, reliability, and other functions of statistical analysis share one common property. Unlike the test of significance, these other functions operate before or during the data collection. This is because these other functions are

interwoven with substantive considerations at the lower levels. Some of these are taken up in the following discussions of the other levels of the Pyramid. Others will appear in the later sections on validity, reliability, and samples.

1.1.2 EXPERIMENTAL DESIGN

The experimental design mirrors and embodies questions being asked by the investigator. Almost the simplest design involves two treatments, experimental and control, the question being whether the experimental treatment has real effects on the behavior. The significance test aims to provide an objective answer. But the design is more basic—it determines what the data mean.

It is at this design stage—before any subject is run—that statistics has its greatest value. Most valuable is the function of controlling variables that might confound the interpretation. One notable example is the principle of random assignment discussed in Section 1.4.1, which has the vital function of controlling unknown variables.

Statistics can also help at the design stage by calculating the probability of success/failure, formally called the *power* of the experiment. If this power calculation indicates the experiment is too weak to detect the expected effect, design changes may be possible that will yield adequate power.

Still another design function of statistics arises in analysis of multiple determination. Multiple determination is fundamental in psychological science because most behavior depends on joint action of two or more variables. Among the questions of interest are whether two variables "interact," and if so, in what way. One major achievement of twentieth century statistics has been the development of tools to study multiple determination (Section 1.5).

Although the questions asked by the investigator are defined formally in the design, their substantive meaning depends on what is measured. Substantive meaning requires consideration of the next two levels of the Pyramid.

1.1.3 MEASUREMENT

Measurement has a unique role in the Experimental Pyramid. It is the link between the world of behavior and the world of science. Measurement is thus a transformation, or mapping, from the real world of objects and events to a conceptual world of ideas and symbols.

This measurement transformation is a vital feature of science. Our measurements are produced by our experimental task, apparatus, and procedure. Measurement is thus grounded in experimental specifics that define the transformation from the behavioral world to the conceptual world.

This empirical grounding of measurement will be emphasized in the later discussions of *validity* and *reliability*. These two concepts subsume virtually all of measurement. Reliability represents intrinsic informational content of

our measurements; validity represents substantive or conceptual informational content. Both depend on the three lower levels of the Pyramid, beginning with the level of behavior.

1.1.4 BEHAVIOR

Behavior is the central level of the Experimental Pyramid. Behavior, however, is not autonomous. It is partly created by the investigator's choices in the experimental setup, which include organism, task, apparatus, procedure, response measure, and so forth. These choices determine what the measured data mean.

Progress in any science depends on development of "good" experimental setups. Among the criteria of a good setup are importance of the behavior, its simplicity and generalizability, statistical properties of the response measure, and cost, including time and trouble.

Pavlov's studies of conditioned salivary reflexes in dogs are famous because of the seeming simplicity of the behavior and its presumed generality, not merely as a base for psychological theory, but also as a model and tool for analysis of behavior. The white rat is more popular, partly because of cheapness and convenience, but also because the rat exhibits a broad spectrum of behaviors common with us humans.

The importance of choices in the experimental setup is visible in controversies in the literature. Many, perhaps most, are concerned with confoundings that may undercut the interpretation of the results. These controversies provide useful lore for newcomers in any field.

The choices in the experimental setup are in mutual interaction with the upper levels of the Pyramid. These choices are determiners of the quality and validity of the response measure, as well as its reliability. Mutually, requirements at upper levels guide choices at the behavioral level. The final setup requires compromises between aspiration and practicality, compromises that not infrequently must be made without adequate information. Early work on some problems can look strangely crude until it is recognized how subsequent work transformed our knowledge system.

Experimental setups should be treated as a matter of continuing development. A major impetus to such development stems from arguments over confounding and validity. Similar arguments over reliability would also be useful. In experimental psychology, however, they remain infrequent—in dark contrast to the attention lavished on the significance test.

1.1.5 PHENOMENA

We usually aim to study some phenomenon—information integration, memory, color vision, intuitive physics, language, social attitudes, and so forth. What we actually study is some observable behavior, which we hope is a good measure of

the phenomenon. The difference is one of kind: between the fact of behavior and the name given the phenomenon, which usually carries a conceptual interpretation of the behavior.

We usually conflate the behavior and the phenomenon, presuming that the name we impose on the behavior is warranted. This presumption is more than a convenience; it is the most important determinant of our choices in the experimental setup. But this presumption may be unwarranted. The innumerable arguments over confounding in the literature demonstrate the difference between behavior and phenomenon. This central issue of confounding is discussed in Section 1.2.3 and Chapter 8 is devoted to it.

A related reason for distinguishing behavior from phenomenon is that performance in any setup involves other abilities besides the focal behavior. This is a recurrent difficulty in studying young children's development of concepts such as time and number. Younger children may be handicapped by lesser development of verbal ability, for example, that interferes with their performance on the focal concept. Such confounding is frequent in Piaget's work, to take one example from Chapter 8. This example has statistical relevance because statistical design techniques can remove some of these confoundings.

A further aspect of the behavior-phenomenon distinction concerns generality. Any given setup must be restricted to one or a few exemplars of the phenomenon. Generality for other exemplars is a primary desideratum. Studies of learning, for example, usually concentrate on a single task, hoping the results will hold for other tasks. Sometimes this happens, sometimes not. Pavlov's salivary reflex, surprisingly, yielded findings and principles of considerable generality. Ebbinghaus' rote memory tasks, on the other hand, were a disappointment in the search for general principles of memory.

The most important problems in experimental analysis are at the interface and interaction between the levels of behavior and phenomena. This is widely understood, but this understanding remains largely localized in the lore of particular substantive areas. Such lore represents general problems of method that deserve more focused and systematic discussion than they receive. How the investigator resolves these problems is a primary component of scientific inference.

1.1.6 CONCEPTUAL FRAMEWORK

The base of the Experimental Pyramid is the conceptual framework of the investigator. This framework is most apparent in the interpretation of results in the discussion section of an article. This framework is a major determinant of choices at all upper levels. The experimental design, as one example, is often constructed specifically to support one theoretical interpretation and eliminate alternatives. Studies that are primarily observational or exploratory generally stem from and embody preconceptions about what constitutes interesting aspects of behavior.

Conceptual frameworks are strong determiners of what phenomena are studied. Learning theory was the dominating framework during 1930–1960, but suffered eclipse in the later cognitive movement. Even in its dominant period, however, learning theory concentrated on a few narrow tasks. This led to neglect of the field of education, for one, which suffered further from trying to make do with theories developed for animal conditioning and rote learning.

Conceptual frameworks also strongly influence choice of behavioral task and measure. The memory domain, for example, has been almost totally dominated by a reproductive conception of memory, epitomized in rote learning. The stimulus material is given; the subject's task is to remember it correctly. The hallmark of reproductive memory is reliance on accuracy as the basic measure. Everyday life, however, operates largely through functional memory; reproductive memory has a relatively minor role. Different tasks and measures are needed to study functional memory.

These brief remarks indicate the pervasive role of conceptual frameworks at every level of investigation. A conceptual framework is a complex of knowledge systems, with interconnected levels of generality. Some of the broadest levels appears in the classic polarities of psychology, or dualities, as they might better be called: observational-experimental, association-gestalt, molecular-molar, central-peripheral, behavioral-cognitive, nature-nature, and individual-society, to name some of the more prominent. More specific conceptual frameworks underlie the choice and definition of phenomena, as with the example of memory in the previous paragraph.

The issue of conceptual frameworks has been discussed in various forms by many writers, who have emphasized their appearance and change in the various movements that take place in science. Present concern, however, is more limited and specific, namely, to emphasize the pervasive influence of conceptual frameworks in every aspect and level of experimental analysis.

Conceptual frameworks are personal knowledge systems. Each investigator has his or her own, undergoing continuous change and development. These personal differences are desirable, for they mean that different ones of us will study different phenomena and pursue different directions in prospecting Nature's boundless riches.

1.1.7 THE CAP ON THE PYRAMID

The cap on the Experimental Pyramid represents the theoretical interpretation put on the results of a given investigation. This is what appears in *Discussion* sections. These discussions integrate considerations from all levels of the Pyramid: statistical significance; possible confounds; the quality of the response measure and its relation to phenomena at issue; how far the results may be expected to generalize; alternative conceptual interpretations; and so forth. Discussion sections of published articles exhibit science at work.

Discussion sections are concerned with substantive inference. Much of this is extrastatistical in nature, but it is all concerned with the multiple aspects of validity of the results and their interpretation. The Discussion section thus involves integration of statistical and extrastatistical inference, considered further in Section 1.2.

1.1.8 THE EXPERIMENTAL PYRAMID

The Experimental Pyramid calls attention to important aspects of research endeavor. It makes explicit a multiplicity of considerations that arise in even the simplest experiment. This brief overview emphasizes the functional role of statistics as a way of thinking—a knowledge system—that integrates statistical and substantive considerations at every level of investigation.

Problems of reliability and validity, in particular, occur at all levels of the Experimental Pyramid; they need to be handled in relation to substantive considerations. A seeming exception is the significance test itself. It is applied after the data have been collected, as already observed, and it has been largely separated and divorced from substantive considerations. It is a significant commentary on current outlooks that the significance test came to be the apotheosis of statistics, not that more substantive and more important pair of concepts, **validity** and **reliability**. To these, we now turn.

1.2 VALIDITY

Validity is a primary concern at every level of the Experimental Pyramid. The most important aspect of validity is whether your measures are a veridical representation of the phenomenon or process you seek to study. No less important is the generality of your results. Validity questions are ubiquitous. Concerns about validity underlie most audience questions to colloquium speakers, questions to you by your Ph.D. committee, and criticisms by reviewers of the papers you submit for publication.

A useful research perspective on validity is given by the two distinctions of the validity diagram of Figure 1.2. The *outcome-process* distinction reflects two foci of concern: with observable results or with underlying process.

The *internal-external* distinction refers to generality: internally within the particular research setting, and externally to other settings, more or less different. This distinction will reappear later in the related distinction of *statistical-extrastatistical* inference.

Both distinctions are important for effective research, as indicated in the next two sections. Subsequent sections take up some intertwined aspects of validity that deserve specific discussion, especially confounding and measurement.

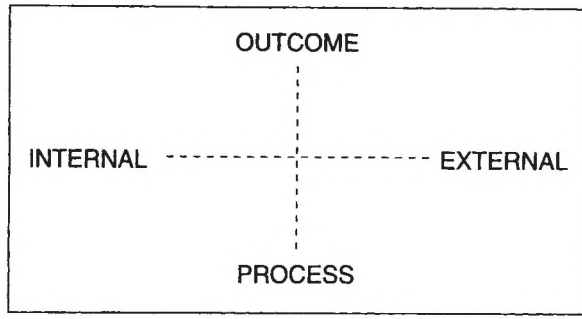


Figure 1.2. Validity diagram: Two continuous validity distinctions.

1.2.1 PROCESS VALIDITY AND OUTCOME VALIDITY

Process and outcome correspond to two quite different research goals. *Outcome validity* is mainly concerned with the observable level of behavior and is a prime goal in applied research. In an accident prevention program, for example, observable frequency of accidents has direct interest in itself. In a Head Start program, similarly, vocabulary size and social skills are outcomes desirable in themselves. The main validity issue concerns how far a given outcome can be generalized beyond the particular situation in which it was obtained.

Process validity is mainly concerned with conceptual interpretation. The observed behavioral outcome is taken to reflect some underlying process. In Pavlov's work on the salivary reflex, for example, there is little interest in salivation in dogs or in humans. Outcome generality is not at issue. Instead, the goal is to illuminate a general process, associative conditioning, dramatically portrayed in Aldous Huxley's novel, *Brave New World.* Process analysis is the focus of much psychological research: perceptual development in children, intuitive statistics, person cognition, and numerous others. Even phenomena of everyday life, such as self-esteem and family interaction, are often, perhaps too often, more focused on process than outcome.

Outcome validity and process validity both have two levels: internal and external. The *internal level* refers to the particular setting in which the study is performed; the *external level* refers to generality across other settings. Internal validity is, of course, prerequisite to external validity (see Section 18.4.1).

Assessment of internal validity begins by assessing whether the result is real, rather than chance, within the specific situation at hand. This is the function of the statistical significance test. A statistically significant result implies internal outcome validity. To assess internal process validity, in contrast, is far more difficult, an issue considered further in Section 1.2.3 on confounding.

The usefulness of the significance test for assessing internal outcome validity suggests that it should be equally useful for assessing external outcome validity. It would be, except that it assumes random samples, and random samples are uncommon. With rare exceptions, experimental analysis employs *handy samples* (see Section 1.4.1). Although statistical inference is central in assessing outcome validity within the handy sample, external outcome validity requires something more.

This something more is *extrastatistical inference:* Inference based on substantive considerations about whether or how much a result obtained in one particular situation will generalize to other situations. The necessity for extrastatistical inference is clear in generalizing from animal studies to humans.

Extrastatistical inference is essential for establishing process validity, both internal and external. This is because process validity is not factual in nature, but conceptual.

Extrastatistical inference is thus the backbone of science, both in the laboratory and in field applications. This fundamental fact of scientific life is glossed over in statistics instruction. Statistics texts and teachers, striving to motivate often ambivalent, statistics-averse students, shrink from emphasizing its limits and inadequacies. Statistical theory itself is elegant and self-contained, wellsuited to autonomous instruction, and it comes to be taught this way. A hazy illusion thus develops that the internal outcome validity certified by the significance test somehow confers more or less internal process validity, and external validity to boot. Quite naturally, but unfortunately, the fact that the main validity questions require extrastatistical inference is avoided. Both kinds of inference, statistical and extrastatistical, will be more effective when they are employed as a team. Statistics should, in short, be appreciated and understood within the context of the Experimental Pyramid.

1.2.2 PROCESS-OUTCOME DISCORDANCE

Choices of research problem and research method are primary determinants of achievement, whether the goal be process validity or outcome validity. These two goals, regrettably, are typically discordant at the external level. Design and procedure that facilitate outcome validity are likely to undercut process validity, and vice versa. Attempts to achieve both goals are likely to achieve neither.

External outcome validity is best obtained when the situation studied is "realistic," that is, similar to or representative of a target situation to which generalization is desired. These target situations are typically complex; behavior is governed by numerous determinants, some not measurable, some not even known. Similarity between experimental and target situations helps ensure that similar determinants are operative and hence that similar outcomes will be obtained.

Outcome-oriented research is typically concerned with behavior change and social action. Educational psychology is a prime example. Comparative studies

of teaching methods may not make sense outside of classroom situations. The one-hour experiments so common in process studies may sometimes be useful for pilot work, but classroom life contains numerous determinants of learning that are not simulated in the laboratory.

The goal of process validity requires an almost opposite strategy. A primary consideration is to simplify the situation in order to eliminate confounding from other processes. The usual aim is to study processes that are sufficiently basic that generality can reasonably be expected. Generality of specific outcomes is not usually of much concern in the process orientation. It is notable how so many of the famous tasks in psychology, the salivary reflex, the rat bar-press, and optical illusions, for example, are so far from everyday experience.

Process—outcome discordance deserves careful consideration. This discordance is regrettable, for it is desirable to derive social value from process research. However, the two orientations impose different constraints on design, measurement, and analysis. It is hard enough to attain even one kind of validity, as may be seen in the criticisms of articles in the literature. To seek both kinds of validity together will typically require compromises in research problem and research method that will compromise both goals.

Choice of research problem and research method are complicated by the fact that the process-outcome distinction is blurred in practice. Few outcome studies are without some process considerations, which may obscure an underlying outcome orientation. Even workers who emphasize social relevance often appeal to process explanations. Many process studies, on the other hand, are under pressure to justify themselves by dubious analogies of outcome generality. The difficulty, as already indicated, is that pursuing two discordant goals is a recipe for failure (see examples in Section 20.2.3; Anderson & Shanteau, 1977; Cook & Campbell, 1979, Chapter 8).

Outcome and process studies are both important. They can be mutually beneficial. They will work together better if the nature of their discordance is recognized. Outcome studies frequently raise new and important questions for process analysis. Such questions can ameliorate the dead-ending and trivialization that beset the process orientation. In return, process studies can be indispensable for testing and developing ideas suggested by outcome studies. The emphasis on process-outcome discordance is an argument against good-intentioned but ineffectual compromise in design and procedure.

1.2.3 CONFOUNDING

The main threat to process validity is *confounding.* We seek to manipulate some designated factor in an experiment, planning to attribute any observable effect to the operation of that factor. Some additional factor may intrude, however, that may also cause the observed effect, undercutting our planned interpretation. The

two factors are thus confounded in operation and, using the term in a related sense, this confounds our interpretation.

Two important examples of confounding are noted here. The first appears in within subject design, in which each subject serves in two or more treatment conditions. Within subject design has notable advantages (Chapter 6), but it suffers potential confounding from order effects. Some condition must come second; thereby it is confounded with possible practice and other transfer from the first. The second treatment may seem good only because it benefits from practice on the first. Or it may seem bad only because it suffers from interference. Such confounding with order of presentation is a stumbling block in attempts to exploit the several advantages of within subject design. Statistical theory provides helpful ways of dealing with this form of confounding, discussed in later chapters.

The second example of confounding appears in the classical concept of *control group.* Suppose we give a treatment to an experimental group, measuring response before and after treatment. The natural measure of the treatment seems to be the change score, (after—before). We might be shot down, of course, if we did not include a control group, that is, a comparable group of subjects, measured similarly, who did not receive the treatment.

The control group is needed because the experimental group might change even if no treatment was given. Such autonomous change can occur for various reasons, including extraneous happenstance and natural growth and development. Such change is then confounded with our experimental treatment.

This confounding is resolved with the control group. By comparing the experimental group with the control, we assess the effect of the treatment itself, freed from whatever confounding was operative.

A more subtle aspect of confounding appears with the *placebo effect.* In medical psychology, the placebo effect refers to improvements reported by patients given a neutral treatment that has no medicinal properties in its own right. Patients' beliefs that they are getting a medicine is enough to make them feel better in surprisingly many cases.

The control problem is more difficult with the placebo effect. Unlike the autonomous changes just considered, the placebo effect is part of the treatment. A control group that received no treatment would not generally do. What is desired is a control treatment similar to the experimental treatment in all respects except the one under test. Sometimes this is as easy as giving the controls a sugar tablet or saline injection. Sometimes this is hardly feasible, as with surgical treatments.

Furthermore, "double blind" treatments may be necessary. The patients, obviously, should not know whether they receive the experimental or control treatment. Less obvious, but hardly less necessary, those who administer the

treatments should also be blind to which treatment they are giving lest their expectations influence the treatment or bias their recording of the response.

A notable example of nonblind confounding appeared in the much-publicized claim that personality type is highly correlated with bodily build, or somatotype. This was an old idea, but it gained new respectability from a careful, rational method for measuring somatotype. These body-mind correlations were exceptionally impressive. For some time they were a staple of introductory texts. But the high correlations turned out to be merely personal bias of the investigators. They knew the somatotype of each subject when they evaluated his personality type; they worked their theory into their data (page 222). This wasteful commotion would have been avoided with rudimentary respect for experimental design. So important is the issue of confounding that a separate core chapter is devoted to it.

1.2.4 MEASUREMENT VALIDITY

Measurement is the link between two very different worlds: a real world of objects and events and a conceptual world of ideas and symbols. This function of measurement appeared in the Experimental Pyramid. The substantive virtue of our theories, accordingly, is bound up with the validity of our measurements. Measurement is thus an integral part of substantive theory.

Measurement has two aspects: *quantity* and *quality*. Quantity is concerned with "how much," which physics made into the essential currency of science and which is also important in psychology. Psychological science, however, faces difficult questions whether our observable response measure, score on a test, for example, or latency in reaction to an emotional stimulus, is a true quantitative measure of the phenomenon under investigation. This question of quantity underlies the concern with linear (equal interval) scales considered in Section 7.2 and in Chapter 21 on psychological measurement theory.

The more important aspect of measurement concerns the quality, or conceptual nature, of what is measured. Quality is the primary issue in measurement validity—whether the numbers are a veridical representation of the quality or process at issue.

Measurement has a twofold empirical base: in the task and procedure on one hand, and in the organism on the other hand. These interact to determine what is observed and what its quality may be. The main function of pilot work is to develop and shape both parts of this empirical base. These empirical operations determine the validity of the measurement transformation.

Measurement also has a conceptual base, manifest in the names we give to our measures: learning, memory, intelligence, expectancy, blame, and so forth. These are not simple object properties, like mass or length. Most refer to complex

phenomena, whose definition involves successive approximation within an ever-developing conceptual framework.

This issue may be illustrated with the concept of intelligence. Scientific understanding of intelligence began with Binet's search for objective tests to improve on teacher's decisions about which children should be sent to special schools for mentally retarded. The conceptual framework of that era dictated the use of such tests as reaction time and sensory acuity. These proved useless. Not without trouble, Binet developed better tests, forerunners of the IQ tests of today. Even today, of course, the concept of intelligence is by no means agreed upon. It undergoes continued change and development, as do many other psychological concepts.

A basic validity consideration in psychological measurement is that many entities are complex, only partially representable in terms of a numerical scale. Verbal and quantitative abilities, for example, both operate in intelligent behavior. Some individuals may be high in one, yet low in the other. Hence a single "IQ" score cannot fully represent the concept of intelligence.

But multiple measures to represent multiple abilities are still only partial measures. Basic processes and even a quality itself are in good part not representable in quantitative terms. Analogous complexity appears with concepts in many areas, from visual perception to self-cognition.

A partial measure of a complex entity may be entirely adequate for the purpose at hand. A test of sentence understanding may be a fine indicator of usefulness of a Head Start program, even though it leaves much unassessed.

Partial measures are more common than is ordinarily realized. We are accustomed to considering number of drops of saliva in the conditioned dog as a measure of learning. The same applies to bar press rate by the rat and to error frequency in a human memory task. Obviously, these are not measures of the learning itself, only one particular product or output of that learning.

Although partial measures are invaluable in life science, two limits on their usefulness should be kept in mind. The name given the measure often becomes reified; the measure is identified with the concept. Part of the confusion over intelligence tests resulted from failure to realize that one-dimensional conceptions are inherently too narrow to represent the phenomena. Similar problems trouble the study of beliefs and values, standardly measured with one-dimensional scales, even though they are complex knowledge systems.

Partial measures can also be dangerous by delimiting the scope of inquiry. A single task and measure can thus become autonomous, surviving long after its usefulness has dwindled away. Memory research thus became trapped within a reproductive framework, defined by reliance on accuracy measures. Much work on social attitudes, to take another example, rests on superficial measures obtained in short experiments that have little relevance to social attitudes of everyday life.

Because of the unique role of measurement in the Experimental Pyramid, measurement validity is entwined with all other aspects of validity. Confounding, in particular, is essentially a problem of measurement validity. The process-outcome distinction, similarly, implies rather different measurement criteria for outcome validity than for process validity. Much of progress consists of improvements in measurement, both in its empirical base and in its conceptual base.

1.2.5 CONCEPTUAL VALIDITY

Conceptual validity refers to the interpretation we place on particular results. More generally, it may refer to our overall conceptual framework, which generates such particular interpretations.

In the Experimental Pyramid, the conceptual framework constitutes the foundation. Your conceptual framework is the primary determinant of your judgments and decisions in any investigation: About which phenomena are important, which are tractable, which particular tasks will pay off, which aspects of behavior you measure, what confounds are likely, how you analyze the data, and how you interpret the outcome.

Many major advances in psychology consisted mainly of broadening our conceptual frameworks, as with the once-foreign idea that the study of animals could be part of our field. The modern cognitive movement, similarly, made its main contribution by liberalizing and broadening the scope of inquiry, as with consciousness and mother love. Equally instructive is the work of the ethologists, who showed the insufficiency of the prevailing experimental framework by demonstrating the effectiveness of field observation for revealing the nature of animal behavior.

A more specific example of this foundation role of conceptual framework in experimental analysis appears with *simplification strategy.* In developmental psychology, to take one instance of simplification strategy, knowledge of the external physical world is a basic component of cognition, and much effort sought to trace out the development of concepts of time, speed, and so forth. A popular question asked at what ages different concepts emerged, whether 5-year-olds, for example, have concepts of time or speed, and whether one concept is necessary as a precursor to the other.

A recurrent objection to such studies was that the tasks used for concept assessment depended on auxiliary abilities, verbal abilities, for example, which are confounded with the concept under study. Confounding from inadequate verbal ability is especially likely at early ages. This makes it uncertain which concepts are primitive, which are derived. The simplification strategy was intended to ameliorate this objection by finding tasks that required minimal auxiliary abilities.

Implicit in simplification strategy, however, is an assumption that the concepts are autonomous, well-defined, and can be studied in isolation. When made

explicit, this assumption becomes questionable. Instead, the concept may begin as a tenuous component of other abilities, only gradually and partially becoming autonomous. Under this alternative conceptual framework, the basic assumption of simplification strategy, namely, well-defined autonomous concepts, becomes increasingly inappropriate at younger ages. Simplification strategy thus becomes increasingly difficult to apply the more it is needed.

One alternative to simplification strategy is to study the focal concept in tandem with auxiliary abilities. This strategy of multiple determination, as it may be called, has some effectiveness (see e.g., Figures 1.3 and 20.1). The main point here, however, is that these two strategies embody very different conceptual frameworks and lead to very different experimental approaches to cognitive development.

A second example appears in attempts to apply standard statistical tools of "interaction" to study joint action of multiple determinants. These begin from a conceptual framework that takes the additive model as basic and also assumes linearity ("equal intervals") of the measuring scale. Both assumptions are often false (Chapter 7). Here again, the validity of particular studies depends on the larger validity of the overall conceptual framework.

We tend to take our conceptual frameworks for granted. To some extent, we are unaware of them until they are challenged. It is at this level, however, that new ideas can have greatest impact.

1.3 VARIABILITY

That variability can serve as a yardstick to assess real effects is a basic statistical principle. Although this principle may seem paradoxical, it makes perfect sense. To claim a real effect of the experimental treatment, it must be shown that the mean difference *between* experimental and control treatments is greater than expected by chance from the variability *within* each treatment—this variability thus constitutes the yardstick. Almost miraculously, statistical theory transformed this qualitative principle to exact formulas (Sections 2.2.0 and 3.2).

This statistical principle has an important empirical implication:

Reduce Variability!

Variability obscures the signals Nature sends us. Such variability, accordingly, is called *error variability* or just *error*. If this variability can be reduced, Nature's signal will be clearer. Ways to reduce error variability are considered in the first three following sections.

More closely considered, variability is seen to be a substantive phenomenon. It arises from causal factors, especially individual differences, that deserve substantive consideration. The last section focuses on individual differences as a desirable phenomenon, not to be reduced, but to be studied in their own right and even to serve as a tool for theory construction.

1.3.1 INDIVIDUAL DIFFERENCES

The main source of variability in many investigations is individual differences. If their effect can be reduced, the results will be more precise. This is an important consideration in designing an investigation.

One way to reduce the effect of individual differences is through experimental hygiene. Subjects find it surprisingly easy to misunderstand instructions, for example, so instructions should be carefully developed in pilot work. The way in which they are presented should also be designed to minimize misunderstanding (see further Section 14.1.2, page 405).

More generally, experimental procedure should follow the principle of task-subject congruence. A good start with a child is to appreciate its name; a good start with a rat is to "gentle" it to reduce its fear in the hands of this huge monster. In addition to reducing error variability, experimental hygiene should yield a cleaner measure of response.

A second way to reduce the effect of individual differences is by stratification of the subject population. One form of stratification is seen in screening tests, in which certain classes of subjects are screened out of the investigation. Screening is common in studies of sensory function, usually to screen out persons with sensory defects such as color weakness or hearing loss. Sometimes, however, the normals may be screened out in order to study the sensory defect. Screening criteria may be similarly used with patient groups or children to ensure some acceptable degree of task functioning. Thus, a pretest may be given to assess whether the subject possesses the background skills or knowledge to perform acceptably in the assigned task.

Screening has more potential than is often realized, especially for situations in which occasional extreme subjects appear. Even one extreme score in a small sample can markedly increase the error variance and decrease power to detect treatment effects. For continuing work in such situations, development of a screening test could be very useful. This is one reason for always looking through your data for extreme scores. Such scores may also point to improvements in procedure and should be reported as a guide to other workers.

Subject stratification can also be incorporated into the experimental design. This can provide a double benefit. In one experimental study of two methods of language learning, for example, subjects were stratified on the basis of their grades in Spanish I (A and B students in one group, C and D students in the other). The two learning methods were then tested with both groups of subjects. One benefit was a substantial reduction in error variability. This came about because part of individual differences that correlated with grade in Spanish I was fractionated out of the error (detailed in Section 14.2.2, page 409).

More important, this stratified design yielded information on the generality of the result. The same teaching method was superior for both the better and the poorer students. Stratification thus yielded a substantive as well as a statistical

benefit. Stratified design is discussed further under block design in Section 14.2 and analysis of covariance in Chapter 13.

A third way of dealing with individual differences is to give several treatments to each subject. Subjects thus serve as their own controls, so to speak. Statistically, the main effect of individual differences is fractionated out, typically decreasing manyfold the variability yardstick. This tremendous statistical advantage has made within subject designs widely popular (Chapter 6).

1.3.2 EXTRANEOUS VARIABLES

Aside from individual differences, extraneous variables also contribute to variability. One major class of extraneous variables consists of attentional factors. Human subjects should generally be run one at a time, not together in batches or classroom groups, for example, because their attention is then under poor control. This threatens the validity of the experiment more than its reliability. In particular, success in working with young children often requires experimenter skills to maintain a forward pace that keeps the child's interest and attention centered on the task.

One way to control extraneous variables is to reduce them through experimental hygiene, as with the attentional factors just noted. Also important is preliminary practice to adjust the subject to the task. Experimental procedure and apparatus, similarly, should be monitored for trouble-free functioning.

A different way to control extraneous variables is to incorporate them in the design. To illustrate, suppose subjects are to be run by two experimenters. It could be unfortunate if one of them was obtuse with subjects or careless with data recording. One function of the pilot work, accordingly, would be to train the experimenters so that both used reasonably similar procedure. In addition, experimenters would be included as a variable in the experimental design, each running half of the subjects in each treatment condition. This design has a double benefit, for it avoids confounding the experimenters with treatments, as would happen if each ran half of the treatments, for example, and it provides an assessment of their performance (see *Factorial Design,* Chapter 5).

This same technique may be used to control other minor variables. One common minor variable is position of a goal object, as in two-choice discrimination learning in children. Right and left are routinely balanced to avoid bias from the child's position preference. In within subject design, similarly, the treatments may be presented to different subjects in different orders. Order and serial position of treatments may then be included as variables in the experimental design. Even though these minor variables may often be expected to have minor effects, this design technique provides insurance for the investigator and assurance for reviewers and readers of the written report that the experimental procedure was indeed under control.

Not all extraneous variables are truly extraneous. Some represent aspects of the experimental situation that are also important for validity. Some stimulus variables, in particular, may represent substantive variability that should be incorporated into the design (Section 1.4.3).

1.3.3 RESPONSE MEASURE

Variability also depends on how the response is measured. The time it takes a rat to run a maze or a child to do a sum may be expressed instead as a speed score, that is, the reciprocal of the time score. Galvanic skin resistance, similarly, may be measured instead as the reciprocal conductance. In each case, the two measures are essentially equivalent, but one may yield lower variability.

Alternatively, the task itself may be changed to get a less variable response process. One major function of pilot work is to shape the task toward lower variability, much discussed under the concept of reliability in test theory. Analogous general methodology does not seem possible in experimental psychology because of the great diversity of experimental tasks. Much can be learned, however, from *Method* sections of published articles in your field.

In general, the investigator has some freedom of choice with the response measure and with the task. These choices are opportunities for decreasing variability and increasing power.

1.3.4 VARIABILITY AS SUBSTANTIVE PHENOMENON

In life sciences, unlike physical sciences, the main source of variability is real individual differences rather than error of measurement. This variability has basic substantive interest. In physical sciences, at least above the quantum level, variability typically resides in the measuring apparatus, not in the entity to be measured. The mean electric charge measured for different electrons represents a universal constant. In life sciences, however, a mean over a group of individuals has mainly a conventional significance, not corresponding to any natural entity. The statistical principle that a mean is meaningless without reference to variability thus has a second, deeper significance in life sciences.

Indeed, an experimental treatment may have opposite effects for different individuals; it may be beneficial to the group as a whole, yet harmful to a minority. Some medical drugs, for example, have noxious side effects to which a few individuals are susceptible. This problem of opposite effects is also a concern in educational psychology, as in the praise-blame study of Note 5.2.2a, and has general relevance in practical applications.

In most theory-oriented research, in contrast, individual differences are ignored. The general run of experimental process studies take for granted either that essentially all individuals in a given condition will exhibit the same directional effect or that any opposite effects by a minority of subjects would not trouble the

face value of the means. This may usually be justified, but the massive unconcern with obtaining such justification is disturbing.

This ignoral of individual differences in experimental analysis has been criticized in other areas of psychology. Educational psychology is concerned with all learners; optimal methods of instruction may differ across abilities and aptitudes. Personality psychology emphasizes an idiographic approach that insists on the uniqueness of the individual, as may be seen in the popularity of case histories. In both these fields, the consignment of individual differences to the statistical error term so common in experimental psychology seems horrifying.

One way to get evidence on individual differences uses stratification technique. In educational psychology, for example, different instruction methods are commonly expected to be more effective with different levels of aptitude. Accordingly, subjects would be stratified on aptitude prior to testing the different methods, as in the foregoing example of language teaching. Stratification thus yields a win-win design, all outcomes being instructive.

A quite different approach employs individual response patterns. Each individual is tested under multiple conditions, and the analysis focuses on the pattern of response for each individual. This approach may be seen in fields as diverse as perception, operant behavior, judgment-decision, and person cognition (see *Single Subject Design,* Chapter 11). Individual differences in response pattern have potential as a tool for general theory, as demonstrated with algebraic models of psychological process (Chapters 20 and 21).

1.4 SAMPLES AND GENERALITY

All scientific inference rests on samples, but seldom are these samples random as statistical theory requires. Scientific inference is mainly extrastatistical. Statistical inference must be integrated into extrastatistical inference.

The primacy of extrastatistical inference is underscored by the fact that sampling occurs at many levels in our experiments. Subject sampling, the usual referent, is only one of these levels. Three other kinds of sampling, no less important, are also discussed here.

1.4.1 SUBJECT SAMPLES

Subject samples are historical accidents. They seldom have intrinsic interest. Their value lies in whatever generality they may have beyond the accidental. Scientific truth must thus be founded on evanescent behavior.

In fact, subject samples are nearly always samples of convenience—*handy samples.* The college students who serve in so many research studies are not random samples from their own colleges, much less from all colleges in the English speaking countries, and much much less from students yet unborn. Handy

samples are also standard in animal research. Monkeys are hard to get, and those who work with them are grateful for every one they can lay hands on. Those who work with rats take what the supply company dispenses.

Scientific inference must thus be founded on handy samples. Statistical inference must do the same, if it is to be helpful. In fact, statistical inference can cope with handy samples through a ingenious device.

This ingenious device is *randomization:* Subjects in the handy sample are randomly assigned to experimental conditions. A statistical significance test can then be used to assess whether the observed effect can reasonably be considered real for this handy sample. This test provides a solid base for extrastatistical inference to larger classes of subjects (see further Section 3.3.2).

Extrastatistical inference, although more or less uncertain, does command considerable consensus. Many experiments on visual perception, such as the line-box optical illusion of Figure 8.1 (page 226), use two or three subjects with confidence that similar results would be found in China and Poland. Experiments on language, on the other hand, or on social attitudes, may be expected to have much less generality.

The issue of generality is usually left implicit, however, except for the usual disclaimer that generality must be assessed in future work. This approach seems reasonable. Requiring arguments about generality would produce much bootless speculation in journal articles, a nuisance for writer and reader alike.

Some additional aspects of extrastatistical inference also require discussion. One appeared in the distinction between outcome validity and process validity of Section 1.2.1. Three others relate to three other kinds of sampling, noted briefly in the following sections.

1.4.2 ORGANISM SAMPLES

Much psychological research is done with animals in the expectation that the results may have some generality beyond the particular species being studied. Indeed, most prominent learning theories have been based primarily on animal data, especially from dogs, rats, and pigeons.

Organism sampling is also governed by considerations of convenience and generality. Fruit flies have long been a favorite in genetic research for practical reasons, including rapid breeding, with good hope of finding panspecies properties of genes. Indeed, Mendel's classic work foreshadowing gene theory was based on a shrewd choice of peas. The popularity of the white rat rests on an impressive list of experimental advantages, such as hardiness and small size, as well as similarities to humans in motivation and learning.

Generalization across species is necessarily extrastatistical. Some investigators demur at attempting cross-species generalization, advocating a philosophy of "rat for rat's sake." One virtue of this philosophy is that it removes the pressure of attempting to justify rat studies by forced, dubious analogies to humans. Another

virtue is that the rat does deserve to be studied for its own sake, as one of Nature's functional biological systems. Moreover, development of interlocking knowledge systems about each species separately is an important base for extrastatistical generalization. Even those who emphasize "rat for rat's sake" expect some degree of cross-species generality.

1.4.3 STIMULUS SAMPLES

The stimuli used in an experiment are often more or less arbitrary, a sample from some larger class of stimuli. In studies of face memory, for example, the stimulus faces may be a sample from some student yearbook. In person cognition, similarly, person descriptions may be constructed from a standardized set of personality trait adjectives. The outcome, however, may be peculiar to the arbitrary choice of stimuli. The examples given in Chapter 8 on confounding show that this problem of stimulus generality should not be taken lightly.

Stimulus generality can be assessed with stimulus replication: Include stimulus materials as an additional variable in the experimental design. Thus, two or more different sets of faces could be constructed by using yearbooks from different schools. Similarly, two or more person descriptions could be constructed, equivalent except for using different trait adjectives. If it turns out that different stimulus materials yield different results, generalization from a study that included only one would have been dangerous. On the other hand, generality is enhanced if different stimulus materials yield equivalent results. Stimulus replication is thus a win-win design. Statistically, stimulus replication may be handled with factorial design (Section 1.5.2).

In some areas, background knowledge will suggest that different stimulus samples will yield equivalent results. A single list of nonsense syllables might be considered enough in a learning study, partly on the basis of common sense, partly because the items of the list themselves constitute a sample. A single example of an optical illusion, similarly, would usually suffice to demonstrate the illusion, since background knowledge indicates size-independence, in particular. Even so, it may be advisable to use more than one set of stimulus materials, expecting to demonstrate that they do yield equivalent results. This can often be done with little cost to the investigator and helpful reassurance to the reader.

In other areas, such as psycholinguistics and personality-social, stimulus replication seems more generally necessary. A single metaphor would hardly suffice as a base for a general theory of metaphor; and an entire class of ambiguous expressions would only illuminate a small part of the general issue of ambiguity in communication. In personality-social, it would be risky to attempt any general conclusion with a single person description, for example, or a single communication in attitude research, or a single moral dilemma in moral judgment. Here again, the method of factorial design is useful.

The stimulus itself is complex in many experimental situations. In education, for example, the teacher is perhaps the most important aspect of the stimulus situation, a primary determinant of relative effectiveness of different teaching methods. It is certainly desirable, if not essential, therefore, to replicate by including more than one teacher for each teaching method. Similar replication is desirable with studies of counseling, therapy, and behavior change generally. Experimental studies with a single teacher, counselor, or therapist can be useful, but they are severely limited.

This too-brief discussion suggests that stimulus sampling may approach subject sampling in importance. This matter is often overlooked, as various writers have complained. Often, it is true, stimulus replication will merely verify an obvious expectation that different stimulus materials yield equivalent results. But such verification may help the reader's evaluation of the study; and failure to verify may be important for the investigator to learn about. In other cases, the behavior may depend as much on the stimulus sample as on the primary experimental variable. Although this can be distressing, it is the better part of workmanship to uncover this dependence.

1.4.4 TASK-BEHAVIOR SAMPLES

The behavior we observe in any study is always a sample, often several levels removed from our focal concern. Classic examples are Pavlov's salivary reflex and Ebbinghaus' rote learning of nonsense syllables. Neither of these behaviors had much interest in its own right. Both investigators considered their findings about learning of these peculiarly specific behaviors to hold far more generally. Most experimental studies throughout psychology are similarly based on very specific behaviors in very specific tasks.

Choice of experimental task is thus a basic determinant of the behavior sampling process. Choice of experimental task is also a basic determinant of the meaningfulness and generality of the results.

Among the criteria for choosing a task are interest and importance of the task behavior, convenience, simplicity, unitariness, and generalizability. Convenience, a practical virtue, includes cost and availability of subjects, apparatus, and so forth. Convenience especially includes a body of prior work that indicates what confounds are harmful, what controls are important, and what kinds of substantive inferences seem tenable.

Simplicity is a universal virtue. A simple task has less to go wrong and is easier to understand. The simplicity of many psychological tasks deserves reflection. The T-maze, for example, is a descendant of the complex, real-life Hampton Court maze. The rat bar-press seems even simpler, one reason for its great popularity.

The related virtue of unitariness refers to the fact that the focal behavior in any task always involves auxiliary abilities that may confound the interpretation.

This problem of auxiliary abilities can be especially difficult in developmental psychology. Stage theories of moral judgment, for example, relied almost entirely on extensive interviews about moral dilemmas that have little relevance to everyday life, even for adults, and can hardly be used below 10–12 years of age. Simpler tasks, such as the blame task shown later in Figure 1.3, are applicable down to at least 4 years. That results with such tasks infirmed the stage theories is only a sign of their more important accomplishment of being applicable with far younger children.

The final criterion, generalizability, is not independent of the others, but it is often the most important. The observed data are only a sample of the behavior of the sample of organisms in the sample task, thus involving at least three levels of generalization. And this behavior itself is usually of interest only as an index of some phenomenon or process at each level of generalization.

Choice of task for eliciting behavior is a central determinant of achievement in research. Most workers begin with existing tasks. This is sensible because developing any new task is often arduous and sometimes unrewarding. This is one reason for attending to *Method* sections in published articles, which often contain useful hints about pitfalls to be avoided. Similarly, existing tasks offer opportunities to tie into current issues and problems.

At the same time, an existing task often develops intellectual inertia that continues its life long past its usefulness, as happened with Piaget's standard choice task in developmental psychology. Similarly, the monolithic conception of memory as reproductive, descended from Ebbinghaus' pioneering work, helped fix this field in an admittedly unproductive rut, and obstructed recognition of the very different functional conceptions, which study memory as it functions in school learning and in everyday life.

Much of the value of the recent cognitive movement consists of greater openness to phenomena and tasks to which the behavioral outlook was uncongenial or inimical. Much is made of the return of consciousness from exile, perhaps too much. No less important, perhaps more, are other "soft" phenomena, such as beliefs and attitudes, probability concepts, self-esteem, excuses and other forms of ego defense, health, play, and many, many other aspects of everyday life. Development of new issues and tasks is a continual source of fundamental contributions.

1.5 MULTIPLE DETERMINATION

Behavior characteristically depends on many determinants, not just one. This multiple determination appears everywhere in psychology, from contrast effects in psychophysics to motive conflicts in everyday life, and from food preferences to marital satisfaction. Prediction of behavior thus requires capabilities for dealing with multiple determinants. Understanding behavior similarly involves reference to its multiple determinants. The fact of multiple determination is thus a central problem for psychological theory and application.

There is a qualitative difference between experiments with one variable and experiments with two. The two-variable experiment is more than two one-variable experiments joined together. Two variables have a combined action that may not be determinable from their separate actions. No number of one-variable experiments can suffice to understand behavior or to predict it. The fact of multiple determination must be addressed in any attempt to develop psychological theory or to apply it.

Statistical theory provides two general methods for analysis of multiple determination: *regression analysis* and *analysis of variance.* These two methods, especially the latter, are the ground for most of the statistical material in this book. A preview is given here.

1.5.1 REGRESSION ANALYSIS

The psychology department at a major university realized that large amounts of time were spent every year evaluating the hundreds and hundreds of applications for graduate study. Despite earnest efforts by the admissions committee, a considerable fraction of those admitted did not do well. Perhaps, someone thought, a statistical prediction equation could do as well or better.

Two predictor variables were the applicant's grade point average (GPA) and score on the graduate record exam (GRE). The response measure Y was success in graduate school. The prediction equation was very simple, just a weighted sum of the two predictors:

$$Y = w_1 \text{GPA} + w_2 \text{GRE}.$$

The weights, w_1 and w_2, represent the predictive importance of the two variables. These weights are related to the correlations between these predictor variables and the specified response measure.

If these weights are known, they may be applied to each applicant's GPA and GRE scores to determine his/her predicted Y. To find these weights requires a set of calibration cases, for which GPA, GRE, and Y are known. In the present example, these calibration cases were graduate students from previous years, whose success in graduate school was known. A regression analysis was applied to these calibration data to obtain the weights. With these weights, Y can be

predicted for new applicants. Applicants for each new year can then be admitted in rank order of their predicted success.

How well does this statistical method compare with human judgment? In the given example, the statistical equation did somewhat better than the admissions committee. It was adopted, accordingly, freeing the committee members for other work. This illustrates a useful social resource, considering that even in 1970 there were more than seven million applications for graduate school, counting multiple applications by the same person to different schools. Furthermore, the predictive weights were later made public so potential applicants could calculate their own Y score. If this was unduly low, they might prefer not to apply, saving time, trouble, and the application fee.

Similar comparisons of statistical method with expert judgment have been made in numerous fields, including personnel selection, pilot trainee selection, diagnosis of disease from medical tests, parole decisions, and clinical diagnosis. A rough summary is that the statistical method ranges from perhaps a little inferior to the best experts in a few fields, such as radiography, to somewhat superior in most fields, such as personnel selection and clinical psychology.

Experts often vehemently disbelieve that an unthinking, statistical method can outpredict them. This outcome, however, is not surprising. One reason is that cognitive integration often follows an averaging process rather than the statistically optimal adding process. A second reason is that the statistical equation usually has, in effect, much more experience: It is constructed from an extensive set of calibration cases with known outcomes (see Section 16.1.3).

1.5.2 FACTORIAL DESIGN

The concept of factorial design is one of the fundamental ideas of experimental analysis. It opens a way to systematic study of causal inference when multiple determinants are operative. An example of factorial design is shown in Figure 1.3, based on a study of moral development. Children were told about a boy who interfered with some workmen painting a house. They judged how much blame or punishment the boy deserved, depending on two specified variables: his intent to harm and how much damage he caused.

The plan, or *design,* of the experiment is shown in the design table at the left of Figure 1.3. Intent of the harmdoer and amount of damage done are the variables, or *factors,* of this design. There were three levels of intent, listed as the rows of the table. Similarly, there were four graded levels of damage, listed as the columns of the table. As this design table indicates, each level of intent was paired with each level of damage to yield a total of $3 \times 4 = 12$ experimental conditions. Each cell of this factorial design thus corresponds to one of the 12 intent-damage combinations.

INTENT	DAMAGE A	B	C	D
MALICE	7	10	12	13
DISPLACE	5	8	10	11
ACCIDENT	2	2	2	2

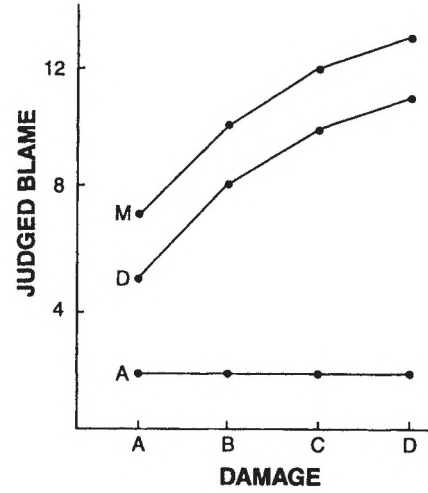


Figure 1.3. Integration schema for blame. Subject assigns blame for harmful actions, given the *intent* behind the action and the *damage* caused by the action. Left panel shows *factorial design,* with three levels of intent (purposive malice, displaced aggression, accident) combined with four levels of damage, graded from A to D. Entry in each cell is the blame assigned for the corresponding levels of intent and damage. Right panel shows *factorial graph* of data in left panel. After Leon (1980).

The number in each cell is the response to the intent-damage combination for that cell. Thus, the top row represents blame judgments for all four cases of malicious intent to harm; blame is least when no damage is done, and increases steadily as damage increases. A parallel pattern is visible with displaced aggression, listed in the second row of the table. For accidental damage in the last row, blame is constant, independent of damage. These data are slightly idealized from an actual experiment (see Figure 11.3, page 318).

The *factorial graph* at the right of Figure 1.3 presents the same data. Each row of data in the design table corresponds to one curve in the graph. Graph and table are thus isomorphic forms of the same data.

The trends in the data table are visible in the factorial graph. The upward trend of the top curve represents the relative effects of the four levels of damage. The elevations of the three curves represent the relative effects of the three levels of intent.

In this particular study, the *pattern* in the factorial graph has primary interest. The top two curves are parallel. Parallelism suggests an *addition rule:* Blame=Intent+Damage. Since the two curves are a constant distance apart, it is as though increasing the level of intent adds a constant amount of blame, regardless of level of damage.

In contrast, the bottom curve, for accidental damage, is flat, independent of damage. Taken together, the three curves thus exhibit a *configural effect,* in which the response depends on the configuration of the stimulus variables.

This factorial graph thus suggests the operation of two cognitive processes in moral judgment, the addition rule and the accident-configural rule. Neither process could have been revealed without joint manipulation of both variables.

In this example, the factorial graph stands up and tells us what the data mean. In practice, the factorial pattern will be more or less variable and the interpretation less definite. It is often desirable, accordingly, to augment the visual inspection with statistical assessment. Analysis of variance is useful for this purpose, as will be seen in Chapter 5 on factorial design.

1.5.3 MULTIPLE DETERMINATION

Two kinds of multiple determination were illustrated in the two foregoing examples. The first involves uncontrolled observational data, the second involves controlled experimental data. In each example, additional variables are also important. The prediction equation for graduate selection actually involved a third variable, namely, quality of the applicant's undergraduate institution. The study of blame needs to consider other determinants such as extenuating circumstances, apology, and atonement, as well as age and gender of the harmdoer.

These two kinds of multiple determination are ubiquitous in psychological science, a consequence of the basic fact that perception, thought, and action are generally determined by multiple factors. Statistical methods for studying these two forms of multiple determination are presented in the following chapters. Basic ideas and techniques are most easily presented for a single variable, as is done in Chapters 3, 4, and 9. Following chapters extend these ideas and techniques to multiple determinants.

1.5.4 TOWARD UNIFIED THEORY

Perception, thought, and action are always determined by multiple factors; multiple determination is a basic fact of psychology. Unified science depends on development of methods to analyze processes of multiple determination. Some assistance has been provided by the statistical tools of regression analysis and factorial design, as just illustrated.

But multiple determination is more subtle. These statistical tools, despite their usefulness, have strong limitations, as will be shown when discussing measurement and models in Chapter 7. In some situations, fortunately, these limitations can be transcended by incorporating these tools within a larger theory of functional cognition, a theory that has had some success in developing unified theory of psychological science (Chapter 21).

This page intentionally left blank

PREFACE

The sample mean should be considered an interval of uncertainty. Although the sample mean appears as a specific number, this appearance is misleading because it is only an uncertain clue to the population mean. To understand the sample mean, this uncertainty must be included.

Statistical theory shows how to make precise this idea of the sample mean as a range of uncertainty. This range of uncertainty can be quantified with two simple formulas; the information in the sample can be used to construct an interval within which the population mean is included with specified confidence. This *confidence interval* properly represents the sample mean.

As a confidence interval, the sample lifts itself by its bootstraps to the level of the population. Thereby, it provides a foundation for statistical inference.

The logic of the confidence interval is developed along a seven-step road in Section 2.2. Each step on the road involves some fundamental concept. Taken together, the seven steps constitute a chain of reasoning that deserves admiration for internal beauty as well as for practical value. The confidence interval reveals order in disorder; the variability and uncertainty of the sample data are transformed into a precision tool.

The confidence interval can be used as a test of statistical significance—whether experimental manipulations have real effect. Significance tests, however, involve additional concepts, of null hypothesis, false alarm, power, and so forth. Limitations and misuses of significance tests are also discussed.

BASIC CONCEPTS

sample and population
sampling distribution
variance and standard deviation
confidence interval
law of sample size
central limit theorem
significance test
null hypothesis
2 × 2 decision table
false alarm and power
individual differences
principle of replication

Chapter 2

STATISTICAL INFERENCE*

All science rests on evidence from samples. A sample, however, cannot provide an exact picture of the population from which it came. A sample mean, in particular, will always differ from the population mean.

It follows that a sample mean should be represented as an *interval of uncertainty*. It is only meaningful when accompanied by its *likely error* from the population mean.[a]

This view of the sample mean as an interval of likely error is vital in principle. In practice, this view may seem too vague, too indeterminate to be useful. Is it really possible to specify an exact value of likely error?

Remarkably, the answer is *yes:* Statistical theory has found a way to measure uncertainty. The sow's ear of variability in the sample can be transformed into the silk purse of a confidence interval. The confidence interval epitomizes statistical inference—and provides a tool for empirical analysis.

2.1 SAMPLE AND POPULATION

The prototypical problem of statistics is to use *sample* data to make inferences about *populations*. This requires an idealization in which we consider a *random sample* of elements drawn from some specified population. The elements are assumed to be *independent:* Knowledge of any one element tells nothing about any other element; each added element carries equal information.

* A review of basic concepts is given in the last chapter of this book.

2.1.1 RANDOM SAMPLE

Statistical theory rests squarely on *randomness.* A random sample, in its simplest form, includes each element of the population with equal probability. Randomness has two interrelated consequences.

First, randomness avoids *bias,* that is, systematic or long-run misrepresentation of the population. Nonrandom samples are susceptible to bias, and many notorious examples of such bias have been recorded. Some investigators have sought to avoid bias by selecting representative samples. However, when people try to construct a sample that is representative of some population, they invariably introduce bias. Random sampling, accordingly, has come to be considered vital.

Second, randomness allows the laws of probability to be applied. These laws enable us to determine the likely error of the sample mean.

The best-known case of random sampling is in election polls, which customarily include a warning about "margin of error." A notable statistical result for large populations is that the likely error of the sample depends only on the sample size—regardless of population size. A random poll of 1200 voters will be equally reliable with populations of 20 thousand and 20 million voters. This would yield a typical margin of error near 1.4%, which makes random polling an expensive but cost-effective tool for social issues (Section 0.1.4).

In experimental work, randomness is usually obtained by *randomization,* that is, by assigning subjects at random to the experimental conditions (see Section 3.3.2, page 69). Our subjects are usually handy samples, chosen for convenience. Randomization not only avoids bias in assigning subjects to conditions, but also allows statistical inference to be applied. Randomization confers empirical reality on the foregoing idealization of random sample.

2.1.2 SAMPLE MEAN AS INTERVAL OF UNCERTAINTY

The most common statistical inference uses sample data to estimate the population mean. This sample-population inference faces a basic difficulty: Different samples from the same population will yield different sample means; all will differ from the population mean; all will be more or less in error. We must live with this error, and statistical theory gives us a rational way to do so.

The rational way to live with sampling error is to specify its likely size. We take the sample mean as an estimate of the population mean, but we know it will be in error. How far wrong is it likely to be? Statistical theory formalizes this goal as follows: Specify a range about the sample mean such that the population mean will lie within this range for 95% (say) of all such sample means. For our one particular sample, we may then have 95% confidence that the population mean lies within the specified range of our sample mean. This range may be considered the *likely error* of our sample mean.

How is it possible to calculate the likely error of our sample mean? If we could draw many samples, the variability among their means would give us an indication of their likely error. But—we have only one sample. How can we use this one sample to determine the likely error of its own mean?

The answer comes from common sense: Look at the variability within our one sample. If our sample data show little variability among themselves, we expect their mean to be close to the population mean. If our sample data are highly variable, we fear their mean may be far away from the population mean. With statistical theory, this commonsense reasoning can be made precise.

This discussion of sample-population inference brings out an interesting subtlety. A sample is a historical accident; it has no meaning in itself. All its meaning is limited to what it can tell us about the population. But all it can tell us is an interval that is likely to contain the population mean.

From this perspective, the customary sample mean is misleading because it is a single number. Central tendency from the sample data should not be identified with the sample mean or any other single number. Instead, the sample mean should be seen as an interval of uncertainty. Of course, the same applies to any other measure of central tendency, such as the sample median. This interval conception of central tendency is taken up next.

2.2 CONFIDENCE INTERVAL

The *confidence interval* is an epitome of sample-population inference. It summarizes the information in the sample in such a way as to put likely bounds on the population mean. This remarkable result is summarized in the following seven sections.

This seven-step road to the confidence interval involves concepts that have general importance in statistical thinking. These concepts will appear repeatedly in later chapters. Each section thus has the collateral purpose of explaining one or more basic concepts.

This seven-step road to the confidence interval also illustrates the nature of statistical reasoning. All seven steps are essential; every bit of available information is needed; there is nothing to spare. The goal is just attainable, and the last step comes from an unexpected beneficence. Everything fits together in perfect harmony. This seven-step road is a godsend; besides its practical value, it deserves your understanding and appreciation for its simplicity and beauty.

2.2.0 TWO FORMULAS FOR CONFIDENCE INTERVALS

Before setting off on the seven-step road, a glance at the final destination may be helpful. Accordingly, two formulas for confidence intervals are given here. Both formulas assume normal distributions, which seems a severe limitation. However,

the last step on the seven-step road will show that they are practically correct for most nonnormal distributions.

The first formula specifies the 95% confidence interval for the mean of a sample of size *n:*

$$\bar{Y} \pm t^* s/\sqrt{n}, \qquad \text{df} = n - 1 \tag{1}$$

where $\bar{Y}$ is the sample mean,
s is the sample standard deviation,
t^* is the .05 criterial value of Student's t on $n-1$ df (see page 813).

We may have 95% confidence that the true mean, μ, lies within this interval, between $\bar{Y} - t^* s/\sqrt{n}$ and $\bar{Y} + t^* s/\sqrt{n}$.

Confidence intervals are narrower for larger samples. The main reason is $\sqrt{n}$ in the denominator of Expression 1. If n is increased by a factor of 4, $\sqrt{n}$ is increased by a factor of 2, and this halves the width of the confidence interval. A secondary reason is that the value of t^* decreases with larger n. For a single mean, t^* equals 2.26, 2.09 and 2.00, for n=10, 20 and 60, respectively. The confidence interval shortens because s is less variable with larger n.

The second formula gives the confidence interval for the difference between two means, $\bar{Y}_1$ and $\bar{Y}_2$, each from a sample of size n. This is simple. In Expression 1, just replace the standard deviation for a single mean by the standard deviation for the difference between two means—replace s by $\sqrt{2}\,s$.

$$\bar{Y}_1 - \bar{Y}_2 \pm t^* \sqrt{2}\, s/\sqrt{n}. \qquad \text{df} = 2(n - 1) \tag{2}$$

If 0 lies outside this interval, we can be 95% confident the difference between the two groups is real (Section 2.3). Student's t ratio is closely related to confidence intervals; its equations are given in Note 2.2.0a (page 53).[a]

We now start down the seven-step road to these confidence intervals.

2.2.1 MEAN, VARIANCE, AND STANDARD DEVIATION

We assume a random sample of n independent scores from some population with mean μ and variance σ^2. Let Y_i denote the score for the ith subject or case, with $\bar{Y}$ the mean of the Y_i:[a]

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{3}$$

We wish to use this sample mean $\bar{Y}$ as an estimate of the population mean μ. Since $\bar{Y}$ will differ from μ because of sampling variability, we desire some indication of the likely error.

To get the likely error of the sample mean, use the variability within the sample itself. The common sense of this approach has already been stated: If the individual Y_i are close together, we think $\bar{Y}$ will be close to μ; if the individual Y_i are far apart, we think $\bar{Y}$ may be far from μ. Our problem is to make this idea precise. To do this, we need a formula to quantify the variability among the Y_i.

A natural measure of sample variability begins with the deviation of each score from the sample mean, namely, $(Y_i - \bar{Y})$. If these deviations are small, variability is low; if large, high. However, averaging these deviations to get a single overall measure of variability goes nowhere because this average is equal to 0. Averaging the magnitude of these deviations, ignoring the minus signs, is attractive but runs into technical difficulties.

Surprisingly, but happily, averaging the squared deviations leads to simple, effective statistical theory. The *sample variance,* accordingly, is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right]. \tag{4}$$

Dividing by $n-1$ instead of n makes s^2 an unbiased estimate of the population variance, σ^2. The second sum is easier for hand calculation.

The variance has the great virtue of *additivity:* The variance of a sum of independent scores is the sum of their variances. For any two independent variables, X and Y, the variance of their sum, and of their difference, is given by the addition formula

$$\sigma^2_{X+Y} = \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y. \tag{5}$$

This additive property leads to the law of sample size below. Also, it underlies the *analysis of variance,* taken up in the next chapter.

For assessing the likely error of a mean, it is preferable to use the *standard deviation,* which is the square root of the variance. Thus, $\sigma = \sqrt{\sigma^2}$, and $s = \sqrt{s^2}$ denote the standard deviations for population and sample.

The name, *standard* deviation, is apt because s lies on the same scale with the same unit as the measured data. If any constant is added to each sample score, the standard deviation remains unchanged. If each score is multiplied by any positive constant, the standard deviation will be multiplied by the same constant. Hence the standard deviation always has the same unit as the response scale. For example, if you measure in inches (centimeters), your standard deviation will also be in inches (centimeters). Because of these properties, the standard deviation is indeed a standard unit of variability.

By itself, however, the standard deviation is not enough to tell us the likely error of our sample mean. Ideally, we could state that 95% of the sample means lie within two standard deviations of the population mean. We would then have

95% confidence that the population mean lay within two standard deviations of our sample mean. But this 95% figure depends on the shape of the population. It applies specifically to the bell-shaped populations called *normal.*

Some empirical populations are approximately normal, but many are not. In fact, we are typically far more uncertain about the shape of a population than about the value of its mean. Unless we know the population shape, we cannot determine the likely error of a sample mean.

We thus face a seemingly insuperable obstacle. Statistical theory, however, has engineered a road through this obstacle. There are six more steps along this road, taken in the next six sections. The last step will bring us to our goal: likely bounds on the population mean.

2.2.2 SAMPLING DISTRIBUTION

The *sampling distribution* of any sample statistic gives the probability of each possible sample outcome. This distribution may be portrayed as a curve with the possible values of the sample statistic on the horizontal axis and their probability on the vertical axis. The sampling distribution thus summarizes the outcomes of all possible samples, giving the probability of each outcome.

All sample inferences rest on sampling distributions. All general properties of a sample statistic are contained in its sampling distribution:

> *Inference from our particular sample is only justified if the same inference would follow from most of the possible samples—as represented in the sampling distribution.*

This statement reemphasizes that the sample is important only as a clue, an uncertain clue, about the population. Sampling distribution is really a basic concept of common sense, therefore, as indicated under *Sample Principle* in Section 0.1.1 (page 783). This common sense concept of sampling distribution can be given a precision edge with statistical theory.

The concept of sampling distribution brings us face to face with a formidable problem. Inference is from the sampling distribution, as the italicized sentence indicates, not from the sample itself. But in practice, we have only one single sample. How can we possibly get beyond our one single sample to the sampling distribution of all possible samples? That this can be possible, even easy, is a stellar achievement of statistical theory.

2.2.3 SAMPLING DISTRIBUTION OF THE MEAN

Since the sample mean is such a useful statistic, its sampling distribution deserves special consideration. Conceptually, the *sampling distribution of the sample mean* is itself a population, namely, a population of means of samples (of size n) from

the parent population. For brevity, it will be called the sample mean distribution. This *sample mean distribution* specifies the probability of obtaining each possible sample mean when we take a random sample of given size from the population. We must rely on this sample mean distribution to tell us about the likely error of our sample mean.

We now have two populations under consideration: The parent population from which we draw our samples—and the population of sample means. Each of these two populations has a population mean, denoted by μ and μ_{mean}, respectively. It is not hard to prove that these two are equal:

$$\mu_{mean} = \mu. \tag{6}$$

Although equal numerically, these two means are quite different conceptually. Hence we use the subscript to distinguish the mean of the sample mean distribution from the mean of the population.

The variability of our sample mean is represented by *its* sampling distribution. There is a natural tendency, even for statisticians, to take a sample mean at face value, without adequate appreciation of its variability (see Section 1.3, pages 16–20). Such appreciation may be gained empirically: Take repeated samples and observe the variability among successive sample means. Even a few samples will give some feeling for the likely error of a single mean.

With numerous samples, a histogram would approximate the shape of the sampling distribution of the sample mean. Such empirical experience can be illuminating and is highly recommended. In experimental work, of course, repeated sampling would require repeating the experiment, which is not often practicable. Even when practicable, it is generally inefficient.

Statistical theory can do as well—even better. Statistical analysis of a single sample can yield better information than empirical analysis of several samples. The next four sections show how we can use the variability within a single sample to determine its likely error.

2.2.4 LAW OF SAMPLE SIZE

What happens to the sampling distribution of the mean when we use larger samples? Intuitively, larger samples should give more reliable means, closer to the population mean. In other words, the spread, or variance, of the sample mean distribution should be smaller for larger samples.

This intuition, fortunately for us, leads to a simple formula. Let σ^2_{mean} be the variance of the sample mean distribution. This variance equals the population variance divided by the sample size:

$$\sigma^2_{mean} = \sigma^2 / n. \tag{7a}$$

This formula follows from the additivity rule for variance (Equation 5).

Equation 7a for σ^2_{mean} has great generality. It holds almost regardless of the shape of the population distribution. Equation 7a thus provides a simple, direct relation between the variance of the population, σ^2, and the variance of the sample mean σ^2_{mean}.

The standard deviation is in some ways more meaningful than the variance as an index of average deviations from the mean. From Equation 7a

$$\sigma_{mean} = \sigma/\sqrt{n}. \tag{7b}$$

Parallel to Equation 7b, a similar formula holds for s_{mean}, the standard deviation of a sample mean. First calculate the standard deviation s of the sample, taking the square root in Equation 4. Then

$$s_{mean} = s/\sqrt{n}. \tag{7c}$$

Equation 7c, the *square root law of sample size,* shows how the standard deviation of our one particular sample estimates the standard deviation of the sample mean distribution for all possible samples. This law of sample size is a critical step on our road to the confidence interval.

It is worth pausing to comment on one practical consequence of this square root law of sample size: Increasing sample size yields decreasing benefits. In the denominator of s_{mean}, $\sqrt{n}$ corresponds to a law of diminishing returns for sample size. To halve the standard deviation of the sample mean requires not doubling but quadrupling the size of the sample. Thus, to get a sample standard deviation around half the size of the population standard deviation would require a sample size of 4. To halve that would require increasing sample size from 4 to 16; halving that would require a further increase from 16 to 64.

This square root law means that each additional subject you run adds less benefit. This is one reason why so much experimental analysis relies on small samples, mainly in the range from 6 to 24. Some lines of inquiry do require samples of size 50, 200, and above, but most experimental studies employ small samples, as you can verify by looking at Method sections of current articles. If you need samples of size 30 to get statsig results, you should think about changing your problem or your task—procedure. For the same reason, two related experiments with 16 subjects per condition might well be preferable to a single experiment with 32 subjects per condition.

The conceptual distinction between s and s_{mean} deserves heavy emphasis. The former refers to the parent population; the latter refers to the sample mean distribution. These two distributions are easily confused. This confusion appears in the not uncommon statement that variability can be reduced by increasing sample size. “The” variability refers to the parent population, which consists mainly of individual differences in typical experiments. It is obviously incorrect to think that sampling more individuals reduces the variability among them. This

statement is quasi-correct—s_{mean} decreases as n increases—but this refers to a distribution different from the parent population.

There is, in fact, a different sample mean distribution for each sample size. The foregoing discussion slurred over this by considering the general case of sample size n. Each n, however, yields a different sampling distribution. These distributions are all alike in one respect, for all have a mean equal to μ, as shown in Equation 6. They all differ, however, in their standard deviations, as shown in Equation 7b.

To emphasize the distinction between s and s_{mean}, the latter standard deviation is often called the *standard error of the mean.* These standard errors are sometimes presented as *error bars* in published articles (Section 4.1.2, page 92).

2.2.5 CONFIDENCE INTERVALS: NORMAL DISTRIBUTION

Given a normal distribution, likely error of a sample mean can be specified exactly with its standard deviation, namely, $\sigma/\sqrt{n}$. The mathematical formula for the normal distribution shows that 95% of the probability in the sampling distribution of the mean lies less than 1.96 standard deviations from the mean of this sampling distribution. More succinctly, 95% of the sample means will lie within 1.96 standard deviations of μ.

Equivalently, μ will lie less than 1.96 standard deviations from $\bar{Y}$ for 95% of the samples. Accordingly, the expression

$$\bar{Y} \pm 1.96\sigma/\sqrt{n} \tag{8}$$

is called the 95% *normal confidence interval.* By this formula, the sample mean, together with its standard deviation, provide likely bounds on the true mean. The smaller the standard deviation, the tighter these bounds.

Two serious obstacles stand in the way of using the normal confidence interval of Expression 8. First, the population a is unknown. The natural step is to substitute the sample estimate s. But s is subject to sampling error, with greater error for smaller samples. Because of this uncertainty about s, we can no longer have 95% confidence. The 1.96 figure is too small; it yields more confidence than the sample warrants.

This obstacle was resolved by Student with his t distribution. This distribution gives us a t^* value to use in place of 1.96, depending on the size of our sample (Table A4, page 813). If our parent population is normal, we can thus obtain the confidence interval for 95% or any other desired degree of confidence. Correct confidence intervals can thus be calculated from the sample data in the manner already specified in Expressions 1 and 2.

The second obstacle has a higher order of difficulty. The 1.96 figure in Expression 8 assumes samples from a normal distribution; so does Student's t. But many populations are nonnormal. Perhaps the 1.96 figure is far off for nonnormal

populations. This obstacle, most fortunately, can be surmounted with the central limit theorem, which is taken up next.

2.2.6 CENTRAL LIMIT THEOREM

We now come to a powerful result—the *central limit theorem.* This theorem shows that the normal distribution has a central place in statistical theory: It is a limit toward which other distributions converge. Specifically, the central limit theorem shows that the shape of the sample mean distribution becomes more and more normal as sample size increases—almost regardless of the shape of population from which the samples are drawn (but see Section 4.1.5, page 94).

Furthermore, this convergence toward normality is usually rapid. Even with rather nonnormal populations, the sample mean distribution may be practically normal for n as small as 5 or 10. With a flat parent population, for example, the distribution of sample means will be nearly normal even for n=4.

2.2.7 CONFIDENCE INTERVALS: NONNORMAL DISTRIBUTIONS

The central limit theorem resolves the second obstacle for constructing confidence intervals, namely, that many data distributions are nonnormal.

> *We obtain the confidence interval for the sample mean distribution, not for the parent data distribution from which our sample is drawn.*

By the central limit theorem, this sample mean distribution will be approximately normal in most applications. Hence the normal confidence interval will be approximately correct. It will provide good likely bounds on the true mean of the sample mean distribution. But this true mean equals the mean of the parent population, by Equation 6. We have thus obtained the confidence interval that we sought. This seven-step road of reasoning justifies the formulas for confidence intervals listed in Expressions 1 and 2 at the beginning.

This confidence interval is a powerful tool. It means that the sample can lift itself by its bootstraps to make precise statements about the population—with specifiable limits of likely error. The variability among the sample elements is essential information; this variability allows the likely error to be estimated. With this simple, elegant chain of reasoning, statistical theory arrives at a fundamental goal. The crucial link in the chain is the central limit theorem, but each other link is also necessary.

It is worth pausing to appreciate how valuable the confidence interval is. Conceptually, the confidence interval emphasizes that the sample mean should be considered a range or interval of uncertainty. It is really an illusion that the sample mean has a specific numerical value. This specific value is meaningless without information about variability in the sample. Viewed in this way, the sample

mean is indeed a range of likely values. Conceptually, therefore, the appropriate estimate of the population mean is not the sample mean as a single number, but the confidence interval about the sample mean.

For many experimental applications, moreover, the confidence interval can be employed as a significance test. To illustrate, suppose Y_i is a measure of how much subject i improves under some experimental treatment. Construct the 95% confidence interval for the sample mean $\overline{Y}_.$. If 0 lies outside this interval, we may have 95% confidence that the true mean is not 0. In this way, the confidence interval quantifies the sample uncertainty, thereby transforming variability into a precision tool.

One conceptual peculiarity of the confidence interval deserves notice. Different samples will give different confidence intervals, of which 95% will contain the true mean. Before we draw a sample, therefore, we can rightly say there is .95 probability that the confidence interval we will get after we have drawn the sample will contain the true mean. Now suppose we draw a sample and construct the confidence interval. This interval is now fixed; the true mean lies either inside or outside this one specific interval. We do not know whether it lies inside or out, of course, but the matter is no longer probabilistic. This is why we use the term *confidence* rather than probability (Section 4.1.6, page 94).

The seven-step road has thus solved the "formidable problem" noted on page 36. The confidence interval constitutes a tool for going beyond our one single sample to the sampling distribution of the means for all possible samples.

2.3 STATISTICAL SIGNIFICANCE TEST

The statistical significance test assesses whether some observed difference is reasonable evidence for a real difference. The classic example involves comparison of experimental and control treatments, E and C. Subjects are randomly assigned to two groups, the treatments are administered, and some response is measured. The question is whether E is better than C.

Implicit in—and essential to—this question is the sample-population distinction. The observed means are only sample results. It is not enough to show that the observed mean response is higher for E than for C. This can readily happen by chance in the sample, even when both treatments have identical effects in the population. The question becomes whether the higher performance of E is reliable.

This question of reliability is answered by the statistical significance test. By looking at the sample data, the significance test tells you whether the observed difference in the sample is reasonable evidence for a real difference.

The following sections are concerned with the conceptual basis and framework for the statistical significance test. Equations and formulas are deferred to the next chapter.

2.3.1 LOGIC OF SIGNIFICANCE TEST

The *significance test* assesses whether the sample data provide reasonable evidence to believe a real effect. The question is: Does the difference between *sample* means warrant the conclusion of a real difference between the corresponding *population* means?

The difficulty of this question about real effects becomes clear upon considering individual differences, prominent in every area of psychology. In any task, some subjects are better, some poorer. If we randomly assign different subjects to the two conditions, chance may put more of the better subjects in group E, more of the poorer subjects in group C. Chance alone could thus produce the difference between the sample means. Unless this operation of chance can reasonably be ruled out, it is not reasonable to decide that the experimental treatment had a real effect. The significance test assesses whether the sample information implies that chance is an unlikely explanation of the observed difference.

The idea of a significance test is straightforward and simple. Small differences between sample means are likely to occur by chance alone. Large differences, in contrast, are unlikely to occur by chance—but likely if there is a substantial real effect. If the observed sample difference is "large enough," therefore, we infer a real effect.

This "large enough" idea of a significance test can be expressed as the symbolic ratio:

$$\text{Test Ratio} = \frac{\text{Observed Difference}}{\text{Chance Difference}}. \tag{9a}$$

If there is no real difference between E and C, the Observed Difference in the numerator will be only chance. The Test Ratio is then expected to be around 1. But if there is a real difference between E and C, the Observed Difference is expected to be greater than chance. The Test Ratio will then be greater than 1, at least on average.

A large Test Ratio thus suggests a real difference between groups. If the Test Ratio is large enough, we infer a real difference. This same logic also holds with more than two groups.

How large is "large enough?" This question has a quantitative answer. To get this answer, replace the symbolic numerator and denominator of the Test Ratio by algebraic formulas for variability, namely, the variances.

The Chance Difference in the denominator of the Test Ratio is the variability *within* each condition. This variability is mainly individual differences; since all subjects within each condition are treated alike, the observed differences in their scores reflect individual differences. Because subjects were assigned randomly to conditions, these individual differences represent the prevailing chance variability, also called *Within Variability.*

The Observed Difference in the numerator of the Test Ratio is the variability between the sample means for the experimental conditions. This is accordingly called the *Between Variability.* With no real effect, the Between Variability will be merely chance, just like the Within Variability, and the Test Ratio will be around 1. Any real effect makes the sample means differ more than chance; the bigger the real effect, the bigger the Between Variability. A large Test Ratio thus argues for a real effect.

In short, a significance test compares variability *between groups* to variability *within groups.* This comparison may be done with Fisher's *F* ratio:

$$F = \frac{\text{Between variance}}{\text{Within variance}}. \tag{9b}$$

If *F* is greater than some criterial value F^*, the difference between conditions is "large enough" to justify the conclusion of a real effect. Formulas for this *F* ratio are given in the next chapter. Here, however, these numerical details are not of concern; Equations 9a and 9b express the logic of the significance test.

Fisher's *F* ratio also applies with three or more experimental conditions, a notable "first" in statistical theory. Student's *t* test is limited to one or two groups, so *F* is used instead.[a]

2.3.2 LOGIC OF THE NULL HYPOTHESIS

The significance test involves a *null hypothesis*—that there is no real effect. The null hypothesis asserts that the observed differences between the sample means are merely chance variability. Our *experimental hypothesis,* in contrast, typically asserts that there is a real effect, over and above chance effects. To accept this experimental hypothesis, it is necessary (though not sufficient) to show that chance alone is not a likely cause of the observed difference.

The null hypothesis is denoted H_0. For the example of Experimental versus Control group, H_0 may be written

$$H_0: \mu_E = \mu_C \qquad \text{or} \qquad H_0: \mu_E - \mu_C = 0.$$

We reject H_0, as already noted, if the *F* ratio is large enough. The significance test, accordingly, is said to be a test of the null hypothesis.

This null hypothesis logic may seem odd, as though we are setting up a straw hypothesis to knock down in order to accept our experimental hypothesis. A little reflection, however, shows this logic is essential. Chance alone will produce some differences among the sample means. To claim a real effect, we should begin by showing that chance alone is not likely to cause a difference as large as that observed. This null hypothesis logic is just common sense.

2.3.3 TWO SAMPLING DISTRIBUTIONS

The null hypothesis logic may be elucidated with the two curves of Figure 2.1. The curve labeled "H_0 true" is the sampling distribution of F under the assumption of no real difference between groups. This sampling distribution gives the probability density (vertical axis) of obtaining various values of F (horizontal axis)—**given** H_0 true. The point labeled F^*_α is chosen so the area under the curve to the right of F^*_α equals the proportion a of the total area under this curve. This figure shows F^* for α of .05 and .01.

Even more important is the second curve in Figure 2.1, labeled "H_0 false." This curve is also a sampling distribution of the F ratio—calculated under the assumption that there is a real effect, that is, that H_0 is false. This curve is shifted to the right of the first curve, reflecting the fact that F tends to be larger when H0 is false. The bigger the real effect, the bigger the rightward shift.

The logic of the null hypothesis may be summarized in terms of these two sampling distributions:

> If our observed F is greater than F^*, we call it *statsig*,
> reject H_0, and decide there is a real effect.
> This decision is justified on the ground that
> a. If H_0 is true, a statsig F is unlikely (a probability);
> b. If H_0 is false, a statsig F is more likely than α;
> the more false is H_0, the more likely is a statsig F.

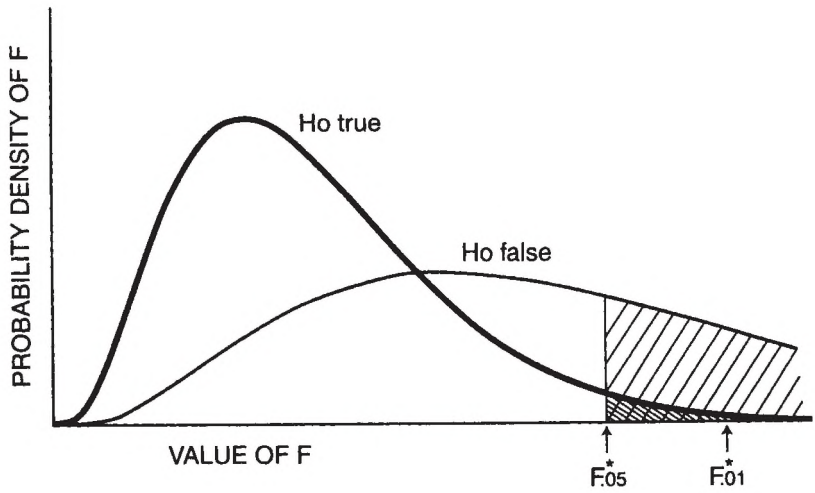


Figure 2.1. Two sampling distributions of F ratios: H_0 true (heavier curve); H_0 false (lighter curve). Vertical axis plots probability density of each value of F on horizontal axis. Points labeled F^* cut off .05 or .01 tail areas under the curve for H_0 true.

2.3.4 THE 2×2 DECISION TABLE

The point F^*_α divides each curve of Figure 2.1 into two regions, each with its own meaning. Each region corresponds to one of the four outcome cells in the 2×2 decision table of Figure 2.2.

In the right column of this decision table, H_0 is false. In this case, a statsig result is good, for then we reject H_0, correctly claiming a real effect. This is a *hit,* in other words, a correct claim for a real effect. A nonstatsig result is bad, for this is a *miss,* that is, a failure to detect a real effect.

In the left column of the decision table, H_0 is true. In this case, a statsig result is bad, for then we reject H_0, falsely claiming a real effect. This is called a *false rejection, false positive,* or *false alarm,* denoted F.A. The good outcome is a nonstatsig result, for then we do not reject H_0.

The probability of false alarm is α—**if H_0 is true.**

The probability of false alarm is 0—**if H_0 is false.**

The probability of a miss (given H_0 false) is denoted β. Correspondingly, the probability of a hit is 1−β, which is also called *power.* The power of your experiment has paramount importance. *If you have low power, don't do the experiment.* Even power as high as .70 means you have 30% chance of a miss.

Every experiment is predicated on an assumption that it has adequate power, that enough subjects are being run. Most investigators, however, still decide how many subjects to run by guess and by God. But a simple power formula is available to help with this important, difficult decision. This problem of estimating power before doing the experiment is taken up in Chapter 4.

		H_0	
		true	false
Decision:	statsig	F. A.	Hit
Decision:	nonstatisg	o. k.	Miss

Figure 2.2. 2×2 decision table for significance test. Each decision yields one good outcome, one bad outcome, depending on whether H_0 is true or false.

A false alarm is usually called a *Type I error* and a miss is called a *Type II error*. This statistical terminology requires learning two arbitrary terms when two meaningful terms are available, as in Figure 2.2.

The 2×2 decision table may seem overly constrictive in that it entails yes-no decisions about marginal results. This yes-no view can be made more flexible by including a band near the .05 level, from .03 to .06, for example, that may be called marginal. Such marginal results are often best handled by replication, as discussed in the last section of this chapter. Similar ideas have been suggested by a number of writers, and they seem sensible to me. Accordingly, this yes-marginal-no view will be implicitly adopted in this book.

Of course, any decision based on sample data is only provisional. "Reject H_0" merely indicates reasonable evidence for a tentative decision.

2.3.5 α-β TRADEOFF DILEMMA

You set the false alarm parameter a at whatever value you desire. Perhaps you feel that α=.05 is too weak, too risky, for it means you have 1 chance in 20 of falsely claiming a real effect when there is none. You can reduce this risk to 1 in 100 merely by changing α to .01. This certainly gives you more reassurance about false alarms.

But—you pay a big price in power. To see this price, look back at Figure 2.1. To change α from .05 to .01 is equivalent to changing from $F^*_{.05}$ to $F^*_{.01}$ in the figure. This increases the miss rate, β, which corresponds to the area under the H_0 false curve to the *left* of the new F^*. Power is thus decreased. Other things being equal, α and β stand in a see-saw, tradeoff relation.

The choice of α depends on the *costs and benefits* of the outcomes listed in the decision table of Figure 2.2. In screening numerous drugs for potential activity, false alarms might be inexpensive because they could be detected in follow-up tests. To avoid a miss of some useful drug, accordingly, a weak a of .10 or higher might be used. Costs and benefits are quite different in preparing to manufacture and market a drug. The cost of a false alarm could be enormous, and a single experiment even with α=.001 would seem inadequate.

Scientific studies use α=.05 for most purposes. I consider .05 somewhat weak, but it seems to work fairly well in practice. Further discussion of this issue is given in Section 19.2, but the conventional .05 will be used throughout this book for simplicity of exposition.

2.3.6 DO NOT ACCEPT H_0

Failure to establish a real effect does not mean the real effect is zero. The 2×2 decision table allows two conclusions: "Reject H_0" and "Do not reject H_0." Of course, "Do not reject H_0" does not mean "Accept H_0" any more than "Not guilty"

means "Innocent." Trial evidence may incline jurors toward "Guilty," but not beyond reasonable doubt.

Do not accept H_0 needs emphasis. When the effect of A is not statsig, it seems natural to say "A had no effect." This is a loose way of speaking, convenient but often harmful (see, e.g., *"Inconsistencies" in the Literature* in Section 4.3.1, page 103). That the sample evidence is not strong enough to meet your criterion for a real effect does not imply zero effect. For scientific communication, such loose speaking seems better avoided.

A deeper, substantive reason for not accepting H_0 is that most experimental manipulations involve observable surface variables, intended to manipulate some underlying process. Failure to demonstrate an effect may merely reflect inappropriate choice of experimental task and procedure; some alternative task-procedure might have succeeded. Human infants, for example, have shown striking perceptual-cognitive abilities in the last two decades, previously unsuspected, through development of new kinds of experimental tasks.

2.3.7 REDUCE VARIABILITY!

Variability is an eternal companion and enemy of experimental analysis. The confidence interval and significance test seize variability by the horns and incorporate it into the data analysis. In doing so, they conceptualize the problem to make clear there is no alternative way.

At the same time, the confidence interval and significance test show that reducing variability has valuable benefits. These benefits were signaled in Equation 9b, in which the denominator is a measure of chance variability. The smaller this denominator, the larger is F and the stronger is the evidence for a real effect. Another sign of these benefits appeared in Expressions 1 and 2, in which lower variability, s, yields tighter confidence intervals.

Variability needs to be reduced before the first subject is run. One function of pilot work is to find ways to reduce variability. Good method is important. Make sure you are in tune with your subjects, that your subjects are in tune with the task, and that your procedure is in smooth working order. The experimental design, similarly, can sometimes confer substantial reductions in variability (see further Section 4.3.4, *Nine Ways to Increase Power,* pages 107ff).

2.4 BEFORE AND BEYOND SIGNIFICANCE TESTS

Statistics has its most important functions in designing an investigation, long before any significance test is made. Significance tests are a minor part of statistical analysis. This basic fact inevitably gets obscured because a significance test is generally essential as a minimal indication that the observed effect is real, not chance. *Is it statsig?* thus comes to seem all-important, whereas it is least important.

It is so easy to fall into viewing the significance test as the be-all and end-all of scientific investigation. Statistics texts have difficulty going beyond the significance test because doing so involves substantive considerations, different for different substantive areas. Indeed, statistical inference so easily lends itself to abstract, context-free presentation that it comes to be taught that way.

Such misconceptions about significance tests are aggravated by the common language meanings of "significant" and "highly significant." These surplus meanings divert attention from more basic problems of data analysis and substantive inference. This is why the term "statsig" is used in this book.

This point was foreshadowed in the discussion of confidence intervals, which go beyond the significance test in an important way by setting likely bounds on the size of the effect. Other limitations and inadequacies of the significance test are taken up in the next few sections. These are not criticisms of a useful tool, but guidance in its use.

2.4.1 SIZE AND IMPORTANCE OF EFFECTS

The significance test applies mainly to just one major class of questions, namely, hypothesis testing. The most glamorous hypotheses are critical predictions from some theory. Other sources of hypotheses are practical assessment of some remedial treatment and hunches that some variable will meaningfully affect some behavior. When the first goal is to show that the effect is not zero, the null hypothesis logic is appropriate (see also Sections 4.1.3 and 19.2).

A second class of investigations is more concerned with estimation. The main interest is to measure the size of some effect, not merely to show it is nonzero, as in hypothesis testing. Observational studies are often of this class, as in studying age trends in vocabulary, or in estimating size of wildlife populations, as with the census of bowhead whales in Section 19.1.3.

As a concrete example of estimation, consider the fact that male births outnumber female births by 106 to 100 in humans. This curious near-equality might be illuminated by investigating nonhuman species. The main concern, however, would be to estimate the size of the inequality in different species, not to show it was nonzero. In statistical terms, the main concern would be the center—and width—of the confidence interval, not whether it excluded zero.

Even for hypothesis testing, the significance test might almost be considered a necessary evil. It is an important first step, for it gives a reasonable modicum of evidence for a real effect. This is surely minimal, as a rule, for asking others to spend their time evaluating your method and procedure, scrutinizing your data analysis, pondering your theoretical interpretation, and relating all this to their own work. The evils of the significance test do not arise from its limitations, but from users who fail to appreciate these limitations.

One major limitation is that a statsig result says little about the size of the effect or about its importance. A small effect can be made "highly significant"

by running a large number of subjects. Conversely, a marginally statsig result may represent a large effect. Of course, a statsig result says even less about the importance of the effect than about its size (see further Section 18.1).

2.4.2 INDIVIDUAL DIFFERENCES

In life sciences, the significance test has a limitation that deserves serious consideration. This concerns individual differences. A test of the mean may be misleading if the treatment induces opposite shifts in a minority of cases. A medicine that cures 90% of the patients is a fine discovery, but it would be bad medicine for you if you are one of the 10% who suffer harmful side-effects.

The significance test is not at fault for neglecting individual differences in this way—that is the substantive business of the investigator. The test performs its proper function on whatever data the investigator sees fit to give it. The investigator should be sensitive to the possibility that a significance test on means may be misleading—*because the means themselves may be misleading by masking individual differences.*

How individual differences should be handled depends on a complex of substantive and statistical considerations. This issue will appear repeatedly in this book. This issue is a never-ending concern in empirical analysis.

2.4.3 MISUNDERSTANDING *p* VALUES

The *p* value of an *F* test can be visualized in the "H_0 true" distribution shown in Figure 2.1. The *p* value is the area under this curve and to the right of the observed *F.* An *F* that large or larger occurs with probability *p*—**if H_0 is true.** This *p* value is printed out by the computer. If $p<\alpha$, the result is statsig. Other than this, the *p* value has little use.

It is tempting but fallacious to think that $1-p$ is the probability that H_0 is false. This fallacy is undeniably attractive: A statsig result is unlikely if H_0 is true; hence a statsig result *seems* to imply some other cause is indeed likely. This seeming, however, implicitly relies on extrastatistical belief that some other cause is indeed likely.

The implicit operation of extrastatistical belief can be made explicit with an experiment to test whether praying over plants will increase their growth. Suppose we get devout persons to pray over 20 corn seedlings, with a proper control condition, and find a statsig result, with $p=.05$. By the given reasoning, the null hypothesis would have probability of .95 of being false. Even devout persons would hesitate to believe prayer–plus–statistics can yield such potent conclusions. And virtually any outcome would yield a value of $(1-p)$ substantially greater than the presumably true value of 0.[a]

In practice, of course, we do take a statsig result to imply that the null hypothesis is false, that is, that our result was caused by a real effect. But the main justification

is empirical, not statistical. This justification lies in our prior, extrastatistical belief that there is a real effect, a belief that led us to do the experiment. Experiments are not done at random, but emerge from our knowledge systems, as indicated in the Experimental Pyramid of Chapter 1. The practical effectiveness of the significance test rests squarely on implicit appeal to our prior belief in a nonchance cause. The significance test works in practice because our prior beliefs generally have some measure of truth (see further Section 19.2.2, page 628).

You can easily avoid this logical fallacy. Don't say a word about the probability of the null hypothesis or the alternative hypothesis. Just say the test was statsig, and that you provisionally consider the effect to be real. Then proceed to the more important issue of what it means.

The exact value of p should ordinarily be ignored for the same reason. Some texts and some articles are enamored of such phrases as "highly significant" for $p<.01$, indexing it as **, and even using *** for $p<.001$. Such asterisking is specious glitter.

The significance test is a simple-minded technique, useful for making provisional reject-do not reject decisions. A statsig result represents a provisional claim that the results are not due to chance alone. Any conclusion about a real effect must integrate other, extrastatistical information.

A decision to publish ordinarily represents a reject H_0 decision, more precisely, a claim that the results constitute sufficient evidence to warrant expense of editorial time and reader time. Once the result appears in the literature, it has a claim on the time and attention of other investigators. Given this, primary concern should be with more important issues beyond the p value.

Among these more important issues are effect size and power, of which the p value is a poor measure (Note 18.1.3c, page 590). Even more important are validity considerations, especially confounding, to which the p value is irrelevant. Few arguments in the literature concern the value of p itself; most concern validity issues of experimental design, procedure, and confounding. Not only do p values have little bearing on these issues, they obscure them.

2.4.4 POWER

Power is a basic consideration in experimental design. For the H_0 false curve in Figure 2.1, power is markedly less than ½ for $\alpha=.05$. This seems far too low to make the experiment advisable. Even granted that there is a real effect, doing the experiment would seem a mistake for there is less than one chance in two of getting reasonable evidence for this real effect. If the result did chance to be statsig, moreover, it would seem a poor bet for follow-up work. Will-o'-the-wisp phenomena are generally unpromising for scientific inquiry.

Low power does not mean you must abandon your experiment. Rather, it suggests you seek to increase power. The power calculation is thus a guide to improving your experiment.

A formula for calculating power is given in Section 4.3, together with nine ways to increase power. It is prudent, of course, to apply the power formula before doing the experiment. This can provide a warning to avoid weak experiments. At the same time, it may suggest ways to increase power. A power calculation is the ounce of foresight that prevents the pound of regret.

2.4.5 EXPERIMENTAL DESIGN: VALIDITY

A statsig result says nothing whatever about substantive meaning or validity. Validity of any statsig result may be undercut by *confounding*. Confounding means that a manipulation of some specified variable is accompanied by variation in some other variable. Then the observed effect may be due to the other variable, not to the specified variable.

A standard example of confounding appears in before-after design. Some response is measured on each of a group of subjects, they are given an experimental treatment, and the response is measured again. The response might be reaction time, say, or degree of felt pain. The experimental hypothesis is that the treatment will yield a faster response or less pain. But a statsig change in response can hardly be interpreted as an effect of the experimental treatment; it might instead be an effect of practice with the reaction time or an effect of suggestion with the pain.

Another class of confounding arises from failure to randomize, which allows bias to enter. A classic example is the large-scale study to test whether supplementary nutrition would increase weight gain in British school children, retrospectively discussed by Student (1931). This was a large, well-planned investigation, impressive even today, with an initial random assignment of children to experimental conditions. Understandably, but most unfortunately, the investigators were nervous about leaving this assignment entirely to chance. Accordingly, they gave teachers in each school some flexibility to adjust the randomization for apparent mischance. But what evidently happened is that the teachers humanitarianly assigned more of the needier-looking children to the experimental group, which got the supplementary nutrition. These children would weigh less, and in fact the experimental group did weigh markedly less than the control before the experiment began. This bias was confounded with the experimental variable. This violation of the randomization largely destroyed the value of this expensive experiment (see page 81 of Chapter 3). This and other violations of the principle of randomization are discussed in Section 8.1.5 (pages *235ff*).

Confounding can take many, many forms, as will appear in later chapters. So important is confounding that Chapter 8 is devoted to it. Here it may be reemphasized that the significance test addresses only the issue of reliability. The significance test is oblivious to substantive confounding and other issues

of validity. Validity depends on extrastatistical inference, based on substantive considerations at the lower levels of the Experimental Pyramid.

2.4.6 PRINCIPLE OF REPLICATION

The principle that results must be replicable is basic to experimental science. One reason is substantive. Any single experiment must be carried out under specific choices of task, stimulus materials, response measure, and subjects. The results may be peculiar to these specific choices, lacking generality (Section 1.4, pages 20–24). Moreover, the results may arise from some confounded variable within the specific situation that compromises the conceptual interpretation. Such considerations underlie the general consensus that knowledge develops gradually and solidifies as an interlocking network of results.

The term *replication* refers here to follow-up studies that pursue aims and findings of an initial study. Although follow-up studies may well include some literal replication of an initial study, completely literal replication is not usually desirable. Replication is most useful when it extends the scope of the initial work.

Replication might include some additional variable, for example, to extend the generality of the results. Or some aspect of procedure might be changed to reduce variability, especially to reduce extreme scores, or to rule out some objection about confounding.

Indeed, potential for such extended replication should be a major consideration already in planning the initial experiment, for it is through replication that an interlocking network of results is built up. A task-procedure that lacks potential for extended replication is usually inadvisable.

A second reason for replication is statistical. A result from any single experiment is only provisional. It cannot be well trusted unless it has been replicated, a point emphasized by Fisher long ago. A significance test provides provisional confidence, justification for further work, but such further work is necessary for solidity. An associated reason is that a result can be statsig without being especially replicable (see *Are Statsig Results Reliable?* on page 104).

Replication has a notable advantage over a test of significance. The latter is only a hopeful promise of replicability, the former is empirical reality.

Furthermore, replication obviates much of the concern and anxiety over significance testing. One concern is with statistical assumptions, such as normality, that are required in the formal statistical analysis, but may be poorly satisfied in the data (Section 3.3). Another concern is with false alarm escalation that occurs when making multiple tests (Section 4.2.2, pages 99ff). These and other statistical perplexities can be greatly eased through replication. A bird in the hand is worth a covey in the statistical bushes.

In general, single experiments are weak in solidity and in informativeness as well. This line of thought implies that the normal unit for publication is an integrated set of experiments. Even a single replication experiment can add valuable solidity. Replication has become increasingly required for publication in psychological journals.

There are, of course, various exceptions. Some studies provide valuable qualitative information, as with observational or case studies. In some studies, the behavior may be sufficiently important that the results deserve publication regardless of whether they are statsig, especially if the study is expensive. Some studies use a factorial design that provides, in effect, substantial replication as part of a single large experiment. Some studies present useful new methods or techniques. Allowing for such exceptions, however, the principle of replication says that a statsig result in a single experiment is not generally enough to warrant consideration by the scientific community.

NOTES

2a. The sample mean could equal the population mean in special cases of little practical importance, for example, when the population has only one element. In this case, as in others, a nonconsequential technical misstatement markedly simplifies the exposition.

2.2.0a. The formula for Student's *t* is closely related to Expressions 1 and 2 for confidence intervals. To test H_0: $\mu=0$ for a single group:

$$t = \bar{Y} \div s/\sqrt{n}. \qquad \mathrm{df} = n - 1.$$

To test H_0: $\mu_E - \mu_C = 0$ for two groups, make the denominator larger by $\sqrt{2}$:

$$t = (\bar{Y}_E - \bar{Y}_C) \div \sqrt{2}\, s/\sqrt{n}. \qquad \mathrm{df} = 2(n-1).$$

If 0 lies just at either end of the confidence interval, then *t* equals *t**.

2.2.1a. I use *Y,* not *X,* as the response measure to be consistent with near-universal usage in regression analysis, in which *X* denotes the stimulus variable or the predictor, and *Y* denotes the response.

2.3.1a. Student's *t* ratio also has the symbolic form of Between Variability divided by Within Variability, namely, $(\bar{Y}_E - \bar{Y}_C) \div \sqrt{2}\, s/\sqrt{n}$. But $t^2=F$, so *F* always applies, even with two groups. Since *F* is more general, it is used throughout this book.

2.4 3a. For a more direct argument against the probability interpretation of *p* value, look at the H_0 true curve in Figure 2.1. For any *F,* $(1-p)$ equals the area under this curve to the left of that *F*. But roughly half of these *Fs* have $(1-p)$ at least ½. Virtually none have $(1-p)$ equal to its true value of 0.

HOW TO DO EXERCISES

Exercises can help you develop research judgment. This is part of the "Empirical Direction" pursued in this book. Exercises in other texts on psychological statistics are mostly numerical calculations, fossils from the precomputer age. Exercises in this book aim instead at integration of statistical and extrastatistical knowledge (see *Exercises* in the concluding chapter, *Lifelong Learning*).

The more important exercises concern conceptual issues. They deserve thought and reflection. Answers to some exercises depend on your personal research judgment, and should be discussed with classmates. There need not be a single best answer, nor are those I give in the Instructor's Manual necessarily the best.

What is important is the understanding you gain from thinking and writing your answers, not your answers themselves. A wrong answer can be a good learning experience; a hasty correct answer can be a poor learning experience.

Development of good exercises should be a cooperative effort by all of us, *students above all.* I shall greatly appreciate hearing from you which exercises you do and do not consider useful. Suggestions for improvement and for new exercises will be most welcome. You can reach me at

Norman H.Anderson
Psychology—0109
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093–0109
nanderson@ucsd.edu

EXERCISES FOR CHAPTER 2

NOTE. Some exercises are intended to help develop skills of visual inspection and familiarity with certain formulas. To this end, exercises indicated "by hand" (which includes hand calculator), are intended to be done without a packaged program, even if available on the hand calculator. Mere drudgery of getting sums of squares, of course, is appropriate for a hand calculator.

1. By 1890, it was accepted that air was composed solely of oxygen and nitrogen; assiduous chemical investigations had revealed no other chemically active components. To measure density of nitrogen, Lord Rayleigh used hydrogen to burn out the oxygen from a sample containing both oxygen and nitrogen, eliminated the resultant water with a dessicant, and weighed the remainder gas (overcoming considerable difficulties). (With thanks to Tukey, 1977, pp. *49ff.*)

a. Graph Lord Rayleigh's data in some meaningful way.
b. Find the "anomaly" by visual inspection.

c. How did you recognize this anomaly? What is the statistical principle?
d. Argue against application of a t test to verify the anomaly.
e. From the table, what seems to be the immediate cause of this anomaly?
f. What might underlie the immediate cause of (e)?
g. What other features of the data strike your eye?

Rayleigh's values of density of nitrogen

Date	Source	Density
29 Nov. 1893	Nitrous oxide	2.30143
2 Dec. 1893	Nitrous oxide	2.29890
5 Dec. 1893	Nitrous oxide	2.29816
6 Dec. 1893	Nitrous oxide	2.30182
12 Dec. 1893	Air	2.31017
14 Dec. 1893	Air	2.30986
19 Dec. 1893	Air	2.31010
22 Dec. 1893	Air	2.31001
26 Dec. 1893	Nitric oxide	2.29869
28 Dec. 1983	Nitric oxide	2.29940
9 Jan. 1894	Ammon. nitrite	2.29849
13 Jan. 1894	Ammon. nitrite	2.29889
27 Jan. 1984	Air	2.31024
30 Jan. 1894	Air	2.31010
1 Feb. 1894	Air	2.31028

NOTE: Data from "On an anomaly encountered in determinations of the density of nitrogen gas" by Lord Rayleigh (1894), *Proceedings of the Royal Society of London*, *55*, 340-344.

2. An experiment with n=10 subjects in each of two groups yields $\bar{Y}_E = 8$ and $\bar{Y}_C = 4$ with s=3.
   a. Show that the 95% confidence interval for the mean difference is 4±2.82.
   b. Show that the t ratio for the mean difference is 2.98 (see Note 2.2.0a).
   c. Compare and contrast the implications of (a) and (b).

3. By visual inspection of Figure 2.1, estimate power for α=.05 and also for α=.01, assuming the given curve for H_0 false applies.

4. Consider the sample {1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5}.

a. Show by visual inspection that the sample mean is 3.
b. Calculate the sample variance by hand from Equation 4.
c. Show that the 95% confidence interval for the sample mean is 3±.81.
d. Construct the 99% confidence interval for the sample mean.
e. Explain the difference in width of these two confidence intervals.
f. In your opinion, is the 4% increase in confidence worth the increase in width of the interval?

5. Prove the assertion of Section 2.2.7 that if 95% of the sample means lie within 2 standard deviations of the population mean, then an interval of width ±2 standard deviations about the sample mean assures you 95% confidence that this interval contains the population mean.

6. What does the text mean by saying in Section 2.3.4 that "Every experiment is predicated on the assumption it has adequate power"?

7. You are TA in an undergraduate class on research methods.

a. Write a paragraph for your students giving an intuitive rationale why larger samples have narrower confidence intervals for the sample mean.
b. Show how this intuition is quantified with a formula in the text.

8. You are TA in undergraduate research methods. Write a paragraph giving your students an intuitive explanation about the common sense of "likely error" as discussed in the fourth paragraph of Section 2.1.2.

9. In what way does the standard deviation of the sample mean distribution differ from the standard deviation of the population distribution?

10. a. In what *qualitative* ways will the sample mean distribution for samples of size 3 differ from that for samples of size 2?

b. In what way will they be the same?

11. What relation is there between the distribution of heights of a population of adult women and the corresponding sample mean distribution for:

a. samples of 1 woman?
b. samples of 2 women?
c. samples of n women?

12. This exercise concerns similarities and differences between a jury trial and a significance test.

a. What is the legal analogue of the false alarm parameter, α?
b. What are the legal analogues of miss and false alarm?
c. P says the analogue of the null hypothesis is "The defendant is guilty." Q argues for "The defendant is innocent." What do you say?
d. What is the legal analogue of the stricture, "Do not accept H_0?
e. What is the legal analogue of the α-β tradeoff?

How does the legal system handle this tradeoff?

f. What is the legal analogue of power?
g. How does power affect the behavior of prosecuting attorneys?
h. How does power affect the behavior of defense attorneys?
i. Do you see any logical difference between the decision of a jury and the decision of the journal editor on your thesis you submit for publication?
j. Jurors may judge whether the defendant deserves to be punished by taking extenuating circumstances into account, not merely whether he or she is guilty of having performed a certain action. In your opinion, how much does this consideration change the foregoing decision analysis?

(As a case with historic interest, Daniel Sickles shot and killed his wife's lover, not under the compulsion of momentary emotion, but with deliberate premeditation. There was no doubt about his action; nor that such action was criminal. Yet he was acquitted by the jury. Perhaps the jury would not have bent the law so far had they foreseen that in his later career as a brash, political general in the Civil War, Sickles would put the Union army in dire peril at the Peach Orchard at Gettysburg on 2 July, 1863. Some truly wonderful letters by his wife are quoted in *Sickles, the Incredible,* Swanberg, 1956.)

13. You study efficacy of prayer by having devout persons pray over corn seedlings, using a *t* test to compare their growth amount with a control.

a. Your data analysis yields $t(60)=2.00$. How confident are you *personally* that *prayer* had the observed effect? How confident are you *statistically*?
b. You do an exact replication of the experiment and get similar results. Now how confident are you personally that prayer had the observed effect?

14. An undergraduate in your class on research methods measures the stocking-foot height of all persons in a certain class and constructs a confidence interval. She finds mean height statsig greater for females than males. What different explanations would you consider possible?

15. What mental model do you think underlies the intuitive feeling that larger populations require larger samples to get same accuracy?

PREFACE

A useful way to look at data is in terms of *variance:* If the differences *between* our treatment means are large—compared to individual differences *within* each treatment condition—this is a sign of real treatment effects.

This intuitive decision process is made precise in this chapter through analysis of variance (Anova). Formulas are presented for the variance between treatment means ($MS_{between}$) and for the variance of the individual responses within a single treatment (MS_{within}). The ratio,

$$F = MS_{between} / MS_{within},$$

becomes our decision guide. A larger F is stronger evidence for real differences between our treatments. If F is "large enough," we have a statistically significant result, provisional evidence for real treatment effects.

Anova makes "large enough" precise. In terms of Chapter 1, your F ratio is an index of reliability, that is, the reliability of the observed differences between the treatment means.

In practice, a significance test is easy. Just give your data to the computer. It will calculate your F ratio and tell you whether it is "large enough."

The F test applies to two *or more* conditions; it includes the t test as a special case. Further, $MS_{within}=s^2$, which may be used to construct confidence intervals using the expressions in Chapter 2. Confidence intervals can be very helpful to your reader when you describe your data.

Anova depends on certain assumptions. Two of these, *normal distribution* and *equal variance,* are not usually problematic in experimental studies. *Independence* is critical, but can usually be ensured through careful procedure and random assignment. Practical aspects of *How to Randomize* are discussed in the appendix beginning on page 77.

It cannot be emphasized too much that the statistical significance test says nothing whatever about substantive significance of your results. The significance test merely tells whether your result has some minimum degree of reliability. This is a minimum first step; unless the result is reliable, there is little point in worrying what it might mean.

Questions of meaning, however, are primarily extrastatistical. Questions of meaning involve considerations at lower levels of the Experimental Pyramid of Chapter 1. Statistics can help with some of these questions, as will be seen in later chapters, but this help requires going beyond the significance test to issues of experimental design.

Chapter 3

ELEMENTS OF ANALYSIS OF VARIANCE I

This chapter and the next give the elements of analysis of variance (Anova). Different subjects are assumed in each experimental condition, with a single score for each subject.

3.1 ALGEBRAIC MODEL FOR ANALYSIS OF VARIANCE

All Anova rests on some algebraic model. This model represents each response as a sum of empirical quantities: effects of the experimental manipulations, together with response variability.

3.1.1 POPULATION MODEL

The experimental variable is denoted *A,* with *a* specific *levels,* A_j, and with *n* different subjects assigned to each A_j. The A_j are experimental *treatments* or *conditions,* and the *n* subjects or scores for each treatment condition are sometimes called a *group.*

The score of individual *i* in condition *j* is denoted Y_{ij}. The population mean for condition *j* is denoted μ_j, the sample mean by $\bar{Y}_j$. The mean over all the conditions is denoted $\bar{\mu}$ for population and $\bar{\bar{Y}}$ for sample.

For a single variable, the Anova model is so simple it hardly deserves the dignity of being called a model. Its simplicity, however, underlies its usefulness. For the population, the model represents each score as the sum of the treatment mean, μ_j, plus an individual subject deviation from that mean, ε_{ij}:

$$Y_{ij} = \mu_j + \varepsilon_{ij}. \tag{1a}$$

By rewriting the treatment mean as $\mu_j = \bar{\mu} + (\mu_j - \bar{\mu})$, we get

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

$$= \bar{\mu} + \alpha_j + \varepsilon_{ij}. \quad (1b)$$

The α_j in this Anova model represent the relative treatment effects: $\alpha_j = (\mu_j - \bar{\mu})$ is the *difference,* or *deviation,* of the condition mean μ_j from the overall mean $\bar{\mu}$. In this deviation form, α_j represents the *relative effect* of treatment A_j. If all treatments had equal effect, all μ_j would be equal—each μ_j would equal $\bar{\mu}$—and each α_j would be zero.

Differences in treatment effects are thus represented as nonzero α_j. The purpose of the experiment is to estimate these relative effects.

The ε_{ij} in the Anova model measure variability between individuals. From Equation 1a, $\varepsilon_{ij}=Y_{ij}-\mu_j$, that is, the *deviation* of individual *i* in condition A_j from the mean for that condition. These ε_{ij} thus represent variability between individuals who receive the same experimental treatment. The variance of ε_{ij} for group *j* is denoted $\sigma^2_{\varepsilon j}$. Equal variance is assumed, unless otherwise noted, so we drop the *j* subscript and simply write σ^2_{ε}.

This population model is true by definition. It is imposed on the data because it turns out to be useful. Each deviation score in Equation 1b represents a pertinent empirical quantity. This idea of deviation scores is basic, and will reappear continually in later chapters.

3.1.2 SAMPLE MODEL

Exactly parallel to the foregoing population model is the model for the observed sample of data,

$$Y_{ij} = \bar{Y}_j + e_{ij}. \quad (2a)$$

The error term is denoted *e* for the sample model rather than ε for the population model. In terms of deviation scores, the sample model has exactly the same form as the population model of Equation 1b:

$$Y_{ij} = \bar{Y} + (\bar{Y}_j - \bar{Y}) + e_{ij}$$

$$= \bar{Y} + \hat{\alpha}_j + e_{ij}. \quad (2b)$$

The "hat" on α_j denotes the sample estimate of the population parameter, α_j.

We rely—perforce—on these sample data to make inferences about the population. Specifically, we wish to test whether the $\hat{\alpha}_j$ from the sample are large enough to infer real differences among the α_j.

3.2 SIGNIFICANCE TESTS

Inferences about real differences between treatments employ the idea of Section 2.3.1. Real effects are measured by the variability *between* groups; chance effects are measured by the variability *within* groups. Comparison of these two yields the *F* ratio:

$$F = \frac{\text{Between variability}}{\text{Within variability}}.$$

Real effects tend to yield larger *F* ratios. Conversely, larger *F* ratios suggest the operation of real effects. We need to put this idea into quantitative form, which can be done with mean squares, or variances, of the sample data.

3.2.1 MEAN SQUARES AND *F*

The denominator of the *F* ratio is just the within group variance. It is computed separately for each group and averaged across groups. This is conveniently done in terms of *sum of squares* (SS) and *mean square* (MS). For group *j,* the within group SS is

$$SS_{within,j} = \sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2, \tag{3}$$

Each $(Y_{ij} - \bar{Y}_j)$ represents the deviation of individual *i* from the mean of all individuals who get the same treatment. The SS is the sum of these deviations squared. This is a general rule: Every SS is a sum of squared deviations.

Summing Equation 3 over groups yields

$$SS_{within} = \sum_{j=1}^{a} SS_{within,j} = \sum_{j=1}^{a}\sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2. \tag{4a}$$

Dividing SS_{within} by $(na-a)=a(n-1)$ yields the *within group variance,*

$$MS_{within} = SS_{within}/a(n-1). \tag{4b}$$

This within groups variance is the denominator of *F*.

The numerator of *F* is obtained using analogous formulas for the *between groups variance.* In this case, the deviations are between the group means, $\bar{Y}_j$, and the overall mean, $\bar{Y}$. Thus,

$$SS_{between} = \sum_{j=1}^{a}\sum_{i=1}^{n} (\bar{Y}_j - \bar{Y})^2 = n\sum_{j=1}^{a} (\bar{Y}_j - \bar{Y})^2; \tag{5a}$$

$$MS_{between} = SS_{between}/(a-1). \tag{5b}$$

These two between formulas are directly analogous to the two preceding within formulas of Equations 4.

Finally, F is obtained as the ratio of the MSs for between and within:

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}}. \qquad \text{df} = (a-1)/(na-a) \tag{6}$$

Here "df" is an abbreviation for *degrees of freedom,* or information units, discussed in Section 3.2.4. The df for the numerator of F are $(a-1)$, the number of groups minus 1. The df for the denominator are $(na-a)$, the total number of scores minus the number of groups.

A larger F ratio is stronger evidence for real effects. If the real effects are zero, the F ratio tends to be near 1. If there are real effects, F tends to be larger than 1, by an amount related to their size. If the observed F is large enough, accordingly, we conclude that our treatments did have real effects.

How large is "large enough"—that is the question. We desire a *criterial value,* F^*, such that an observed F greater than F^* warrants a decision that our treatments really did have different effects. It is astonishing that this question has an answer that is both exact and general, regardless of the substantive nature of the experiment.[a]

3.2.2 FINDING F^*

How to find F^* is the concern of this section. In practice, you can simply look up F^* in Table A5 (page 815), corresponding to the df listed in Equation 6. Note that F^* is smaller for greater df. The df for $\text{MS}_{\text{within}}$, namely, $(na-a)$, equals the number of information units used to estimate this within variability. More df means a more reliable estimate. The same size effect is thus more reliable with more information units, a fact that is quantified through the concept of df. A related argument holds for the numerator df.

Statistically, the value of F^* is obtained as shown in Figure 3.1, repeated here from the previous chapter. As explained in Section 2.3.3 (pages 43*f*), F^* is chosen so that a of the area under the H_0 true distribution lies to the right of F^*.

To get the values of F^* in Table A5, Fisher had to derive a mathematical formula for the curve labeled "H_0 true" in Figure 3.1, which is the sampling distribution of F, given H_0 true. With this formula, we can choose F^* so that the proportion of the area under the curve that lies to the right of F^* is equal to α. This sets the false alarm parameter at α.

The formula for this sampling distribution can be derived with four assumptions: random sample, independence, and equinormality (normal distribution and equal variance), discussed in Section 3.3. Fisher's derivation of this formula involved a fundamentally new conceptual framework for statistical theory. In this book, however, we do not need to know this formula, only that it was the basis for constructing Table A5, which gives us F^* with no effort.[a]

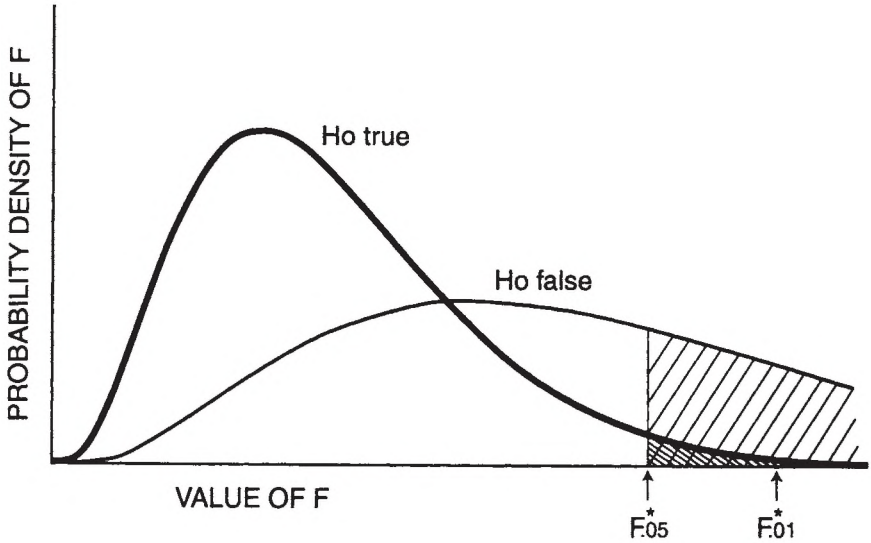


Figure 3.1. Two sampling distributions of F ratios: H_0 true (heavier curve); H_0 false (lighter curve). Vertical axis plots probability density of each value of F listed on horizontal axis. Points labeled F^* cut off .05 or .01 of the area under curve for H_0 true.

3.2.3 SIGNIFICANCE TEST

Testing the Null Hypothesis. The F ratio tests the overall null hypothesis that all treatments have the same effect. More precisely, this null hypothesis asserts that the true means of the a treatment groups are all equal:

$$\text{overall } H_0\text{: } \mu_1 = \mu_2 = \ldots = \mu_a. \tag{7}$$

If our observed F is larger than our criterial F^*, we reject the overall H_0. The logic of this significance test has been given in Section 2.3 (pages 41*ff*)

The formula for the sampling distribution of F in Figure 3.1, given H_0 true, depends only on the degrees of freedom (see Note 3.2.2a). Most fortunately, it does not depend on the error variance. For given df, a single formula holds—regardless of the particulars of the experiment. This formula is used to construct the curve for H_0 true in Figure 3.1 and thereby to determine the criterial F^* for the chosen α. In this way, the probability of getting a larger F—if chance alone is operative—is set at α.

Actually, you seldom have to bother looking up F^* in Table A5. Computer programs print out the "p value," which is the area in the upper tail to the right of the F obtained from the data. If $p<\alpha$, then the observed F is greater than F^*. In this way, the p value provides a significance test. This is almost the only use of the

p value. The fallacy that *p* equals the probability that H_0 is false is a treacherous cognitive illusion (Section 2.4.3, pages 49*f*).

Numerical Example. The Anova formulas may be illustrated with the following example of two groups, each with three subjects.

$$\text{group 1: } \{1,\ 3,\ 5\} \qquad \text{group 2: } \{7,\ 7,\ 10\}$$

Our first step is to calculate the variance *within* groups, MS_{within}. Since $\bar{Y}_1 = 3$ and $\bar{Y}_2 = 8$, Equation 3 yields

$$SS_{within,\,1} = (1-3)^2 + (3-3)^2 + (5-3)^2 = 8;$$

$$SS_{within,\,2} = (7-8)^2 + (7-8)^2 + (10-8)^2 = 6.$$

By Equations 4a and 4b,

$$MS_{within} = (8+6)/2(3-1) = 3.5.$$

This is the variance within groups, or error variance.

Our second step is to calculate the variance *between* groups, $MS_{between}$. Since $\bar{Y} = 5.5$, Equations 5a and 5b yield

$$MS_{between} = 3\left[(3-5.5)^2 + (8-5.5)^2\right]/(2-1) = 37.5.$$

Our third and final step is to calculate *F* with Equation 6:

$$F = 37.5/3.5 = 10.71; \quad df = 1/4.$$

From Table A5, F^* on 1/4 df is 7.71 for α=.05. Since 10.71>7.71, the sample gives reasonable evidence for a real difference between the two groups.[a]

Why the .05 Level? The standard α of .05 is obviously a convenient round number and has been the target of various criticisms. One common concern is that false alarms tend to get published whereas failures to replicate generally do not. Hence, it is argued, the literature is filling up with false alarms. Student and Fisher could as easily have constructed their tables for α=.02, and perhaps they should have done so.

In practice, however, the .05 level seems to work reasonably well. One reason lies in the principle of replication. When results agree across a series of experiments, the danger of false alarms is greatly reduced (Section 2.4.6).

A second reason lies in the role of extrastatistical background knowledge, which acts as a primary filter of false alarms. No experimental result stands alone. All are evaluated relative to a network of background knowledge, based in part on results from similar experiments, in part on general empirical sense. A prudent investigator will hesitate to publish a statsig result that appears to disagree with prevailing knowledge without replicating it. Reviewers and editors, similarly, may request replication in such cases.

Much the same point appears in the fact that experimental studies grow out of prevailing knowledge and expectation. Many results thus possess reasonable prior validity. With such prior belief against false alarms, α need not be so stringent (see also *One-Tailed Tests* on page 98). Although I personally consider the .05 level somewhat weak, this convention evidently works fairly well, not only in psychology, but also in other sciences.

3.2.4 MEAN SQUARES AS VARIANCES

The mean squares in the preceding section are actually variances, a statistical fact that is relevant to experimental design.

Error Variance. MS_{within} of Equation 4b is the error variance s^2 of the sample data—differences among subjects treated alike. By Equation 4 of Chapter 2, the sample variance for group j is the SS of Equation 3 above divided by $(n-1)$:

$$MS_{within,\,j} = SS_{within,\,j}/(n-1) = s^2_{e,j}. \tag{8a}$$

Since $MS_{within,\,j}$ measures variability between subjects treated alike, it serves as an error yardstick.

Pooling these MSs across groups yields MS_{within} of Equation 4b. Assuming equal true variance, we may drop the j subscript to get

$$MS_{within} = s^2_e. \tag{8b}$$

Hence $\sqrt{MS_{within}} = s_e$, which may be used as s in the two expressions for confidence intervals on page 34.

Although MS_{within} varies from sample to sample, its average value, also called its *expected value,* equals the population variance. The symbol E is used to denote expected values so

$$E(MS_{within}) = E(s^2_e) = \sigma^2_\varepsilon. \tag{8c}$$

Treatment Variance. The variance of the sample treatment means is given similarly by Equation 5b. It is just the average of the squared deviations of the sample means from their overall mean. It may seem odd to average across the a groups by dividing by $(a-1)$. This is a little statistical device that makes the sample variance an unbiased estimate of the population variance. This device applies generally, as in Equation 8a for error variance.

The treatment variance, however, has two components. One is the error variability, which will cause the sample means to differ even if the true means are equal. The other is real differences, if any, among the true means. This latter component is the variance among the population means, written as[a]

$$\sigma^2_A = \sum(\mu_j - \bar{\mu})^2/a = \sum\alpha^2_j/a.$$

The two components of the treatment variance are additive. In terms of expected values,

$$E(\mathrm{MS}_{\mathrm{between}}) = \sigma_\varepsilon^2 + n\sigma_A^2. \tag{8d}$$

***F* as Ratio of Variances.** From the two foregoing subsections, it follows that the *F* ratio of Equation 6 is the ratio of two variances: the variance between treatment means divided by the error variance. The meaning of *F* may be appreciated by looking at its expected value (see also Note 5.1.4a):

$$\begin{aligned} E(F) &= E\left[\frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}}\right] \\ &= \frac{\mathrm{df}_{\mathrm{denom}}}{\mathrm{df}_{\mathrm{denom}}-2} \times \left[1 + \frac{\mathrm{df}_{\mathrm{num}}+1}{\mathrm{df}_{\mathrm{num}}} \times \frac{E(\mathrm{MS}_{\mathrm{between}}) - E(\mathrm{MS}_{\mathrm{within}})}{E(MS_{\mathrm{within}})}\right] \\ &= \frac{\mathrm{df}_{\mathrm{denom}}}{\mathrm{df}_{\mathrm{denom}}-2} \times \left[1 + \frac{\mathrm{df}_{\mathrm{num}}+1}{\mathrm{df}_{\mathrm{num}}} \times \frac{n\sigma_A^2}{\sigma_\varepsilon^2}\right]. \end{aligned} \tag{9}$$

Equation 9 makes explicit three major determinants of the power of your experiment. All three appear in the last term, $n\sigma_A^2/\sigma_\varepsilon^2$. As this expression shows, two ways to increase power are to increase sample size, *n,* and to increase the real effect, σ_A^2. The third way to increase power is to decrease error variability, σ_ε^2. In fact, reducing error variability is often the most effective way (see *Nine Ways to Increase Power,* pages 107*ff*).

Note that *E(F)* will ordinarily be near 1 when H_0 is true. For then $\sigma_A^2 = 0$ and $E(F)=\mathrm{df}_{\mathrm{denom}}/(\mathrm{df}_{\mathrm{denom}}-2)$, which will be near 1 unless the error df are small. An observed *F* near 1 thus indicates nonstatsig results (Note 3.2.3a).

Degrees of Freedom. A degree of freedom may be considered a statistical unit of information—the amount of information in a single score. With *N* independent scores, there are *N* information units, or *N* df. These *N* df are divided among the Anova sources to index the amount of information in each source.

Each mean represents one unit of information. With *a* independent groups, there are *a* df for their means. But we are usually interested in differences among these means, namely, the relative effects, $\alpha_j = \mu_j - \bar{\mu}$, in the Anova model. Only $a-1$ of these α_j are independent, since their sum is always zero. This is why the between groups source has $a-1$ df in the Anova table.

A similar argument holds for a single group of *n* scores. Their mean has 1 df, leaving $n-1$ df for deviations from their mean. These deviations yield the variance

within group *j,* shown in Equation 8a, which thus has $n-1$ df. Pooling across the a groups yields $a(n-1)$ df for MS_{within}.

The df for numerator and denominator in Equation 6 are parameters in the theoretical formula for the F ratio (Note 3.2.2a). They govern the shape of the H_0 true distribution in Figure 3.1. Thereby they also determine F^*.

The error df reflect the reliability of the error mean square. Each error df represents one unit of information about reliability used to calculate MS_{error}. The more df, the more reliable is the estimate of MS_{error}, and the smaller is the criterial F^*.

Error as Yardstick. The use of error variability as a yardstick or reference standard for real effects is basic in data analysis. The smaller this yardstick, the better. Reducing error variability through experimental procedure and statistical design should thus be a primary concern. In this endeavor, it is helpful to recognize that the error term has multiple components: individual differences, variability within individuals, apparatus variability, procedural shortcomings, and ambient environmental influences.

Individual differences is usually the largest component of error by far, so reducing their effect is most desirable. Good procedure can help. The subject should be put at ease, whether rat, child, patient, or college sophomore. Task and instructions should be carefully piloted to remove the ambiguities that can trip up an occasional subject and produce an extreme score. Statistical designs to reduce effect of individual differences include screening subjects, stratification, repeated measures, and single subject design, discussed in later chapters.

Error variability is almost more important than the treatment means. Error variability determines the width of your confidence intervals, for example, which confer meaning on the mean. Moreover, your error variability mirrors the effectiveness of your task and procedure; it stands as a measure of the quality of your work.

3.2.5 HAND CALCULATION OF SS

Hand calculation of an SS is sometimes necessary, but Equations 3–5 should not generally be used for this purpose. The deviation scores in Equations 3 and 5a have conceptual meaning as expressions of within and between variability, respectively. Deviation scores exhibit the logic of Anova.

Deviation scores, however, typically involve decimal places, which makes them time-consuming, not to mention error-producing, for hand calculation. The following hand formulas are considerably easier.[a]

Hand Formulas for Equal *n*. The hand formula for $SS_{between}$ is

$$SS_{between} = n \sum_{j=1}^{a} \bar{Y}_j^2 - SS_{mean}. \tag{10a}$$

This equation exhibits a pattern that appears repeatedly in Anova. It is easiest to remember its schema:

Square each treatment mean.
Multiply each by the number of scores in that mean (in this case n).
Add.
Subtract SS_{mean} (given by Equation 10b just below).

Two other formulas complete the Anova:

$$SS_{mean} = an\bar{Y}_{\cdot}^2 \tag{10b}$$

$$SS_{within} = \sum_{j=1}^{a}\sum_{i=1}^{n} Y_{ij}^2 - SS_{between} - SS_{mean}. \tag{10c}$$

These equations can be useful in reanalysis of published data. You may wish to test between a subset of treatment means for which only the overall F is reported. Using the means for this subset of treatments from the published graph or table, apply Equations 10ab to get $SS_{between}$ for this subset. If MS_{error} is not given in the article, you can reconstruct it from the reported means and F (Exercise 4).

Hand Formulas for Unequal *n*. The foregoing schema for hand formulas also applies when the groups have unequal size. Thus,

$$SS_{between} = (n_1\bar{Y}_1^2 + n_2\bar{Y}_2^2 + \ldots + n_a\bar{Y}_a^2) - SS_{mean}, \tag{11a}$$

where n_j is the number of scores that go into the mean for group j. This equation reduces to Equation 10a when all n_j are equal to n, as you can readily show. The df for $SS_{between}$ are $a-1$ in either case.

Similarly, the total number of subjects is $N=n_1+n_2+\ldots+n_a$, so

$$SS_{mean} = N\bar{Y}^2, \qquad (\bar{Y} = \sum\sum Y_{ij}/N = \sum n_j\bar{Y}_j/N). \tag{11b}$$

The formula for SS_{within} is still given by Equation 10c. The error df are $N-a$.

3.3 VIOLATIONS OF ASSUMPTIONS

The four statistical assumptions of Anova listed in Section 3.2.2, page 62, will never be exactly true. It is necessary, accordingly, to inquire how sensitive the statistical analysis is to deviations from these assumptions. Independence is vital, so this assumption is considered first. Hardly less important is the assumption of random sampling.

Nonnormal distribution, on the other hand, often has little adverse effect. The same also holds for unequal variance (with equal n). Understanding the causes and consequences of violations of these two distributional assumptions can help you improve your experimental design and your data analysis.

3.3.1 INDEPENDENCE

Independence is violated when one score carries information about some other score. One common source of nonindependence is natural groupings among the subjects: classroom groups or discussion groups in educational psychology, work groups in industrial psychology, and families in social or counseling psychology. Subjects' responses may be expected to be more similar within groups than between groups, thereby violating independence.

Nonindependence can also arise when otherwise independent subjects are run in batches or ad hoc groups. Within each group, subjects may be influenced similarly by the same extraneous factor. In one group, for example, some subject may ask a question about the instructions that affects other subjects in that group. Such effects may usually be small and unimportant, but it seems generally preferable to avoid danger by using appropriate design.[a]

Independence can also be violated when treatments are applied in blocks rather than randomly. Apparatus settings, for example, may not be easy to change from one subject to the next, especially when each change requires recalibration. It then seems attractive to set the apparatus once and run several subjects before changing the setting. Unless done appropriately, however, this can introduce serious nonindependence.[b]

3.3.2 RANDOM SAMPLES AND HANDY SAMPLES

The foundation of statistical theory is the assumption that the observed data are a random sample from some well-defined population. But actual samples are nearly always handy, convenience samples from some ill-defined population. How is it possible to apply statistical theory to handy samples?

The answer lies in *randomization*—random assignment of subjects to experimental conditions. Randomization allows a statistical inference from the data to a well-defined population, namely, the population of all possible random assignments of subjects to conditions. A statsig result means our observed effect is not limited to our one random assignment, but would have been observed for most other random assignments. We may thus reasonably reject the null hypothesis and decide our result is real for our handy sample.

Statistical inference thus takes us no further than our handy sample. This limitation may seem disappointing, but doing this much is a notable achievement. Further, it makes explicit the deeper importance of extrastatistical judgment in the Experimental Pyramid and helps avoid the common confusion of statistical with extrastatistical inference. Establishing the reality of the difference for our handy sample provides a firm base for extrastatistical inference.

With nonrandom groups, statistics cannot control confounds from preexperimental differences. A significance test may merely reflect these differences. Without randomization, a statsig result for a handy sample does not imply a real treatment effect (Sections 13.2, 15.5, and 16.2).

3.3.3 NONNORMAL SHAPE

With most common nonnormal distributions, Anova keeps α under good control. This is a valuable result because it means that departures from the normal shape do not seriously bias the false alarm parameter.

Several studies have investigated this issue with rating scales, which often have five or even fewer steps. As an extreme case, consider a rating scale with just two response steps, 0 and 1. This two-point distribution is certainly far from the normal bell shape, yet Anova handles it quite well. As long as the frequencies of the two responses are not too extreme, say between 20% and 80%, even 20 df for error yields an effective α close to the assigned .05 or .01. The reason, of course, is that the sampling distribution of the *mean* is fairly normal by virtue of the central limit theorem.

Power, on the other hand, can be suboptimal with distributions that are long-tailed or heavy-tailed, both fairly common. Even one or two extreme scores can markedly increase the variability of the sample mean. Three alternative analyses that may yield more power have been used: Trimming, which reduces the influence of tail scores; transforming the data to a more normal shape, followed by regular Anova; and analysis based on the ranks of the data, which does not require normality (see Chapter 12).

3.3.4 UNEQUAL VARIANCE

Much of the concern expressed about unequal variance in different conditions is unnecessary. For randomized experiments, the most important reason is that unequal variance implies real treatment effects.

Randomized Experiments. Unequal variance is not usually a problem in randomized experiments. The main reason is simple: Unequal variance cannot occur unless there are real treatment effects. If subjects are assigned at random and the treatments have identical effects, then the population distributions for all treatment groups will be identical. Hence all groups will have equal true means and equal true variances.

If treatments do have differential effects, central tendency and variability both will generally be affected. Typically, however, means are affected considerably more than variances. With equal *n,* moreover, the α level is not much affected by unequal variance. Even when the true variances differ by a factor of 2 or 3, the effective a may change only from .05 to .06. With unequal variance, as with nonnormality, accordingly, the effective false alarm parameter of the overall *F* remains close to the nominal value specified by the investigator.

This conclusion, however, has an important limitation: It may not apply to analyses beyond the overall *F*. Confidence intervals, in particular, can be adversely affected by unequal variance (pages 93 and 97). Extensions of Anova to handle these cases are discussed in Section 12.5.

Nonrandom Groups. Natural groups, in contrast to randomized groups, will differ naturally in variability. One common example involves comparison of age groups in developmental psychology. In many tasks, younger children will be more variable. Anova that allows for unequal variance may be necessary, especially for two-mean comparisons and other follow-up tests (Section 12.5).

3.3.5 A WIDER NULL HYPOTHESIS

The Anova assumptions can be viewed from a different perspective, in which *F* is considered a joint test of all assumptions: the standard null hypothesis of equal means; the other two distributional assumptions (normal distribution and equal variance); and the independence assumption. Independence is critical, as already noted, but it can ordinarily be secured through experimental procedure and will be assumed to hold in what follows.

A statsig *F,* accordingly, implies violation of one or more of the three distributional assumptions: equal means, equal variances, and normal distributions. The null hypothesis of equal means is just one of these three assumptions. Yet we single it out, claiming that a statsig *F* implies unequal means, whereas logically a statsig *F* could imply nonnormal distribution or unequal variance. This asymmetrical treatment is justified because *F* is sensitive to unequal means and relatively insensitive to nonnormality and unequal variance.

This property of *F* is a piece of good fortune, one that could hardly have been anticipated. To appreciate just how fortunate this is, it may be noted that this same approach falls flat for testing the null hypothesis of equal variances. Most tests of variances are so sensitive to nonnormality as to be almost useless. If our null hypotheses had generally to refer to equal variances, as happens in some experiments, statistical analysis would have been frustrating. The actual outcome, however, justifies the asymmetrical treatment of the three distributional assumptions.

This justification can be pushed one step further by considering *F* as a test of the "wider hypothesis" of equal means *and* equal variances (see Fisher, 1960, pp. 44*ff*). In this wider null hypothesis, equal variance is no longer a statistical assumption, but goes hand-in-hand with equal means.

This wider null hypothesis is entirely sensible. Under random sampling or randomization, means and variances both must be equal, unless there are real treatment effects. Hence unequal variances imply real treatment effects, just as unequal means do, although the interpretation may differ in the two cases. Confidence intervals, as already noted, can be adversely affected by unequal variance. This wider null hypothesis, accordingly, ameliorates much of the concern with unequal variance in randomized experiments.