Null Hypothesis Significance Testing:
History, Criticisms and Alternatives

Daniel J. Denis

A thesis submitted to the Faculty of Graduate Studies in
partial fulfillment of the requirements
for the degree of
Master of Arts

Graduate Programme in Psychology
York University
Toronto, Ontario

October 1999
The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author’s permission.

L’auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L’auteur conserve la propriété du droit d’auteur qui protège cette thèse. Ni la thèse ni desextraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-59127-1
Null Hypothesis Significance Testing:  
History, Criticisms and Alternatives

by  
Daniel J. Denis

a thesis submitted to the Faculty of Graduate Studies of York University in partial fulfillment of the requirements for the degree of  
MASTER OF ARTS

©1999

Permission has been granted to the LIBRARY OF YORK UNIVERSITY to lend or sell copies of this thesis, to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film, and to UNIVERSITY MICROFILMS to publish an abstract of this thesis. The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.
Abstract

Despite years of criticism, null hypothesis significance testing (NHST) continues to be psychology's most widely-used model for the purposes of statistical inference. Since Fisher (1925), psychologists have increasingly adopted the model in ruling out chance from their conclusions. The purpose of the present investigation was to address two questions: 1) what are the problems associated with NHST and are they due primarily to inherent difficulties of the model, or due to misunderstanding and misuse, and 2) given the problems that accompany the use of NHST, can any other model successfully replace it, and thereby overcome these problems? It was found that although many problems can be attributed to user misuse, there are nonetheless serious theoretical and methodological flaws inherent in the model itself that warrant the search for a substitute model of inference. Upon considering alternatives, it was found that confidence intervals, Loftus' plot-plus-error-bar procedure (PPE) and Serlin & Lapsley's "good-enough principle" could feasibly serve as replacements to NHST. The Bayesian model of statistical inference, despite its alleged problem of prior probabilities, is recommended as the "best" alternative to NHST. Furthermore, the fact that posterior probabilities converge is used as support for why even frequentists should approve of the Bayesian approach.
Acknowledgments

I wish to sincerely thank my supervisor, Dr. Christopher D. Green, for his invaluable help and guidance in the completion of this thesis. I would like to especially thank him for introducing me to the controversy surrounding null hypothesis significance testing. I would also like to thank members of my committee, Dr. Michael Cowles and Dr. Raymond Fancher for their comments on earlier drafts of this thesis. Finally, I would like to thank my colleague Dan Aalbers for providing me with interesting insight into the world of hypothesis testing.
Contents

Introduction 1

Chapter 1: A Historical Overview of Null Hypothesis Significance Testing: 9
Is Today's NHST Attributable to R. A. Fisher?

Chapter 2: Problems with NHST: Model or Misuse? 37

Chapter 3: Proposed Alternatives to NHST 71

Chapter 4: The Bayesian Alternative to NHST 114

Conclusion 153

References 161
List of Tables

Table 1 -- p. 91.
Data in "tabular" form, -- contrast to graphical representation shown in Figure 2.
Reprinted from Loftus, 1996.

Table 2 -- pp. 148-149.
Bayesian probability revisions of Scientist A and Scientist B for hypothesis H.
List of Illustrations

Figure 1 -- p. 89.
Graphical representation of statistical power. Reprinted from Loftus, 1996.

Figure 2 -- p. 92.
Data in graphical form (i.e., linear functions). Reprinted from Loftus, 1996.

Figure 3 -- p. 96.
Example of a power table. The y-axis represents the number of observations required per treatment for a test of specified power. Reprinted from Feldt & Mahmoud, 1958.

Figure 4 -- p. 131.
Joint probability set; shaded area represents the probability that $\mu_1 \neq \mu_2$. Reprinted from Iversen, 1984.

Figure 5 -- p. 151.
Bayesian posterior probability convergence (Scientist A and Scientist B).
Introduction

Null hypothesis significance testing (NHST) has been at the receiving end of much criticism since its presumed origin with R. A. Fisher in 1925. It is rare to find a psychology journal that hasn't at least once published material specifically aimed at the defamation of psychology's most popular data-analytic tool. Bakan (1966), Meehl (1967, 1978) and Cohen (1990, 1994) are among many who have pointed out the inherent difficulties that accompany using NHST as a medium for drawing inferences. According to Gigerenzer (1993) however, the procedure is still employed in close to 100% of published empirical studies, despite its deplorable reputation among methodologists. Some (e.g., Hunter, 1997) have even advocated the extreme position that the procedure be banned completely from psychology journals. Indeed, the problem has been regarded serious enough that the Board of Scientific Affairs of the American Psychological Association has recently appointed a committee (called "Task Force on Statistical Inference") dedicated to re-evaluating the current status of statistics in psychology.

To quickly review, null hypothesis significance testing is a procedure in which the goal is to reject an unlikely hypothesis (i.e., the null) and infer what is thought by the researcher to be the more probable hypothesis. This latter hypothesis is often referred to as the alternative, or research hypothesis. Although the null hypothesis is often rejected using a conventional standard (i.e., $p < .05$), the alternative hypothesis
must be assumed, since no specific probability can be assigned to it. Contained in the null is often a statement of no difference, or that any difference obtained is due to chance factors. Contained in the alternative hypothesis is often a statement of what the researcher expects will account for the observed data, given that the data are not due to chance. This latter hypothesis is often termed the conceptual or substantive alternative. As will be discussed in chapter 2, it differs significantly from the statistical alternative. Briefly, the statistical alternative is a statement of "not the null." Inference of the statistical alternative merely asserts that the results are unlikely due to chance. The conceptual alternative serves as somewhat of an explanation that accounts for the statistical alternative. Given an inference of the statistical alternative, the question then becomes one of asking -- if the results are not due to chance, then what are they due to? This question is most often answered by positing a single conceptual alternative (also called "research hypothesis"), although theoretically, many may exist.

As a simple example of null hypothesis significance testing, consider the researcher who wants to know if group A is significantly different from group B. He compares the means of each using a simple between-groups t-test. The null hypothesis is that any observed difference is due to chance factors; the alternative hypothesis is that any observed difference is due to something other than chance. If we reject the null hypothesis, we conclude that chance is not solely responsible for the discrepancy in means. That is, something else must be responsible for a difference this large. That "something" is stated in the alternative hypothesis, and is
held as accounting for the difference in means between group A and group B. Thus, NHST routinely involves the inference of a research hypothesis given that the data are unlikely to be attributable to mere chance.

The Legacy of NHST

Why is NHST still commonly employed? Why do psychological researchers continue to use a tool that has been consistently shown to be inadequate, if not methodologically "immoral"? More generally, why would anyone use anything if it were continually "black-balled" and held as inappropriate or even harmful? Psychologists, in general, are interested in answering such questions. Indeed, the neuro-psychologist is interested in understanding why one continually abuses a substance, despite its devastating long-term consequences. The developmental psychologist is interested in answering why an adolescent begins smoking, despite a wealth of presumably discouraging influences. As in the present case, it is the role of the psychologist to determine the "why" behind such popular use of a "substance" (NHST) in light of its consequences. Just as the cocaine addict continually abuses a substance the experts claim to be detrimental, researchers continually abuse the use of NHST despite what the experts profess -- NHST is methodologically unsound. The clinician may attempt to explain cocaine abuse as a result of addiction. The developmental psychologist may explain teenage smoking as resulting largely from peer pressure and rebellious attitudes. How is one to best explain NHST abuse?

Providing an answer to the question of why researchers continue to use NHST is a difficult process. Many factors may play a role in its overuse. For instance, it is
quite possible that the latest generation of NHST users have adopted the model not because of its theoretical import for psychology, but rather because their teachers have taught the model void of critical analysis. Or, it is possible that researchers have adopted NHST because they have learned that to obtain tenure, one must reject at the p < .05 level. These such reasons, although each may have some validity, will not be considered in the present thesis. In other words, I will not entertain the sociological/political factors that likely play a big part in the overuse of NHST. Rather, I will restrict my discussion to methodological factors alone. That is, why, on a purely methodological basis, is NHST still used? Is it the only model available? Are there no alternatives to this presumably hopeless model? If NHST is as bad as its critics claim, can we not use something else, or is NHST the "best" of alternatives? Is a bad apple better than no apple?

The present thesis constitutes an effort to provide educated answers to questions such as those posed above. Over the next four chapters, I hope to provide answers to these questions through two means: 1) an analysis and evaluation of today's NHST, and 2) by examining and appraising alternatives to the model. In doing so, I hope to provide sufficient information to allow the reader to make an informed decision with regards to NHST. At the same time, I will find myself in a position to substantiate a recommendation with regards to what I believe the future status of NHST should be, again, from a methodological point of view. Should we continue using NHST, or are there competing alternatives to the model? Do these alternatives solve the problems characteristically associated with NHST, or do they create more problems than they
solve? If the latter, are these problems then resolvable in their own right? One thing is worth emphasizing at this point. I write this thesis holding the point of view than every model can, on its own, be considered problematic. It is only through comparing one model to others available that one can fully appreciate a given model. In short, nothing exists in a vacuum, and inferential statistical models are no different. I take the position that a given model should be recommended for use in psychology unless another model deems itself more appropriate. In this, I implicitly assume the necessity of inferential statistics in psychology, and hence, discuss more of what model should be used, rather than whether a model should be used at all.

Chapter 1 serves as a general overview to null hypothesis significance testing in a historical context. The primary goal of the chapter is to compare today's NHST to the model once proposed by Fisher (1925). Hence, some preliminary "roots" of NHST are discussed. Where appropriate, the Neyman-Pearson (1928) model of hypothesis-testing is mentioned, particularly in those instances where it deviates substantially from Fisher's original model. My concluding argument emanating from this chapter is that today's NHST is hardly attributable to Fisher. Although many parallels exist, Fisher's model made specifications that are not followed at all by today's users. Today's model is largely a hybridized, misinterpreted approach to statistical inference. Although Fisher is rightly named the "father of statistics," it is doubtful as to whether he should be associated as truly with today's brand of significance testing. Furthermore, it is hardly fair to fault Fisher with today's NHST problems.
Chapter 2 outlines the fundamental criticisms charged against the NHST model. Criticisms against NHST are traceable soon after Fisher introduced the procedure. Since Berkson (1938) noted of an over-reliance on p values, many professionals from various fields have written specifically with regards to the problems that accompany using NHST. Criticisms against NHST can be categorized into eight main areas of difficulty: 1) sample-size sensitivity, 2) the prior falsity of the null hypothesis, 3) the testing of the wrong hypothesis, 4) difficulties in construing the alternative hypothesis, 5) the "effects/non-effects dichotomy, 6) arbitrary alpha, 7) the "significance" of "insignificant" results, and 8) the violation of assumptions underlying the procedure. It is important to decide between those criticisms that are correctly charged against the model, and those that are more a product of user misuse rather than any inherent faults with NHST. As will be seen, although NHST is misused, it has fundamental difficulties that lie at the heart of the model. The criticisms of NHST are for the most part, well founded.

Chapter 3 presents and reviews a number of alternatives to null hypothesis significance testing. Can NHST be replaced by an alternative model? More specifically, is there a comprehensive alternative that can avoid the many problems that plague NHST, and still provide the same if not more information? The answer to this question, as will be seen, is an overwhelming "yes." Indeed, there are numerous alternatives to traditional NHST, many of which are methodologically equipped to stand as replacements to the model. Granted that any model of statistical inference will present with its own difficulties, the issue is whether another model is "better"
than NHST in dealing with these difficulties. Assuming psychologists are rightly not prepared to abandon statistical inference, the question becomes that of asking which of the alternatives, including NHST, is "best" for psychology? What method of data analysis will tell us most of what we want to know? As will be discussed, some of the procedures are best viewed as complements to NHST, while others are capable of fully replacing the model. For instance, effect size measures along with power analysis are best employed to complement NHST rather than to serve as complete replacements. These tools are likely to make NHST more useable rather than to replace the model. On the other hand, confidence intervals, Loftus' PPE (plot-plus-error-bar) procedure, and Serlin and Lapsley's "good-enough" principle can be regarded as complete replacements to NHST.

Chapter 4 presents the Bayesian model of statistical inference. Also considered an alternative, Bayesianism represents a shift in philosophy about inference and is a rival to the frequentist-based NHST approach. Contrary to the frequentist, who regards probability as a relative frequency based on theoretical sampling distributions, the Bayesian holds that probability is best regarded as representing one's belief or opinion regarding the likelihood of an event. Advantages and disadvantages of the Bayesian model will be spelled out in detail, and they will be contrasted against NHST. The appealing feature of Bayesian statistics is that they are capable of directly testing the research hypothesis rather than merely refuting a null hypothesis as in NHST. However, deriving a probability of the research hypothesis necessitates the estimation of its prior distribution before the data are analyzed. The
way in which prior probability is quantified is the topic of much controversy, and is a popular reason why Bayesian statistics are not employed in psychology. As will be demonstrated however, given the same data (i.e., likelihood ratio), varying prior probabilities eventually converge to similar if not identical posterior distributions, and hence the researcher's opinion is secondary to the data. In other words, prior probabilities are not overstated in the posterior distribution. I will argue that even with a Bayesian approach, prior opinion is minimal in determining posterior probability. In the end, the data do speak for themselves.
Chapter 1

A Historical Overview of Null Hypothesis Significance Testing: Is Today's NHST Attributable to R. A. Fisher?

The purpose of this chapter is to provide an historical overview of null hypothesis significance testing (NHST), focusing primarily on Fisher. The components of Fisher's model will be drawn out in detail, for the purpose of staging a contrast and comparison between his original model, and later modifications that were added to this early configuration. The Neyman-Pearson approach to hypothesis testing will be discussed insomuch as it presents novel contributions or alterations to the Fisherian model. This brief review of the "hybridization" of the two approaches will prepare for my final objective; that of offering a comparison between the model currently used, and that proposed almost 75 years ago. I intend to evaluate the claim that today's null hypothesis significance testing is attributable to Fisher. The following will show that while psychologists use a similar model to that once proposed by Fisher, today's researchers use anything but a pure Fisherian approach. Differences between Fisher's original model and today's model will be highlighted and will support the claim that we can hardly give Fisher's name to the model as currently applied today. As a result of this misattribution, Fisher has been on
occasion unjustly criticized for a model that he did not advocate. I will close the chapter by citing and discussing one such occurrence.

Cowles (1989) is also aware of this pseudo-Fisherian model and makes a comment that so aptly describes the position I intend to defend:

Perhaps we should spare a thought for Sir Ronald Fisher, curmudgeon that he was. He must indeed be constantly tossing in his grave as lecturers and professors across the world, if they remember him at all, refer to the content of most current curricula as Fisherian statistics. (p. 189)

Null Hypothesis Significance Testing: Fisher's Original Paradigm

Before diving into Fisher's significance testing principles, it is perhaps wise to first comment on Fisher's view of induction and inference in the context of experimental design. In Design of Experiments (1966), Fisher's introductory chapter delineates his view regarding mathematical induction. He argues for the estimation of population parameters based on small sample data. While results may be probabilistic, Fisher sees no problem with this:

... many mathematicians, if pressed on the point, would say that it is not possible rigorously to argue from the particular to the general; that all such arguments must involve some sort of guesswork, which they might admit to be plausible guesswork, but the rationale of which, they would be unwilling, as mathematicians, to discuss. We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous,
for the nature and degree of the uncertainty may itself be capable of rigorous expression . . . The mere fact that inductive inferences are uncertain cannot, therefore, be accepted as precluding perfectly rigorous and unequivocal inference. (pp. 3-4)

Hence, Fisher argues for the rigorous quantification of uncertainty when drawing inferences from a sample to a population. He believes that scientific inference can be exact, even if uncertain. In other words, an uncertain (i.e., probabilistic) inference can be as precise as one that is certain. Fisher's philosophy of science held that we learn from experience, yet knowledge must always remain provisional. Knowledge is uncertain, but this uncertainty can be quantified using appropriate statistical measures (Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989). It is the development of these measures that would occupy Fisher throughout much of his life.

In considering now the components of Fisherian significance testing, I must begin with a basic prescription made by Fisher; that of always forecasting beforehand all possible results of the experiment. Furthermore, he asserts that we must know in advance the interpretation of each of the given possibilities. As Fisher (1966) states:

... it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each one of them. Further, we must know by what argument this interpretation is to be sustained. (p. 12)
Fisher required that the experimenter know in advance the possible outcomes of the given experiment. This would simply require the experimenter to calculate the probability of a given result occurring by chance alone. This is typically accomplished using mathematical permutations and combinations. It is a relatively straightforward task. For example, a correct hand grab from a subject claiming to be able to "psychically" select the marked ball from a urn containing a total of just two balls, would not impress in the least since the probability of selecting the marked ball is .5. On the other hand, should there be a total of 1000 balls in the urn and the subject successfully selects the correct ball, this may (according to most minds) be cause to reject the hypothesis that the selection was due to chance. This outcome is presumably more likely to be used as justification for refuting the chance hypothesis because the probability of selecting the one ball from a total of 1000 balls is equal to 0.001 (or 1 in 1000), making it an extremely unlikely event.

Fisher's second requirement noted above, while somewhat more difficult, is just as important. In saying that the experimenter must know in advance the interpretation of each possible outcome, Fisher places great weight on fully describing the design of the experiment before the data are collected and analyzed. In the above example, this requirement would have the experimenter state in advance his interpretation of possible results before the subject reaches in the urn to choose a

---

1 I say "according to most minds" for the sake of highlighting the often rejected subjective platform on which even classical statistics is based. I argue simply that the decision to reject a hypothesis is made on subjective grounds, the probability level (in this case, .001) being a major criterion on which many scientists base their rejections. This statement obviously has Bayesian overtones and
ball. For Fisher then, the experimenter must have adequately designed the experiment before the data are collected, and for Fisher, "design" meant the anticipation of possible outcomes, along with their respective interpretations. After the data are collected, there shouldn't be any surprises. Later in his career however, Fisher (1956) did recommend that the exact significance level be reported after the analysis of the data. This contradicts his earlier recommendation (Fisher, 1935) that the significance level be determined before the experiment is executed. A more thorough treatment of significance levels is discussed later in this chapter.

The second major component of Fisherian significance testing is that of randomization. Fisher was adamant with regards to the randomization of variables and treatments if an experiment were to be considered at all valid. His recommendations for randomization were strict, with each treatment having to be randomly chosen for each individual. The random assignment of subjects to conditions would likely be different should the experimenter have allocated them, for he might subconsciously let his opinions influence the allocation of variables (Gigerenzer et al., 1989). Although in some cases, error could actually be reduced by systematic allocation ("Student," 1937), Fisher was more concerned with the validity of the estimates of error, rather than the quantity of error.

For Fisher, randomization was necessary to satisfy the assumption that should the null hypothesis fail to be rejected, the experimental result is better explained as

---

serves to foreshadow the "frequentist vs. subjectivist" debate that will be presented later in this thesis.
being governed by chance. Although randomization did not eliminate all possible sources of bias in the experiment, it did minimize potential error. As Fisher (1966) summarizes:

Apart, therefore, from the avoidable error of the experimenter himself introducing with his test treatments, or subsequently, other differences in treatment, the effects of which the experimenter is not intended to study, it may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged. (p. 21)

A third component of Fisher's theory was that the population could not be known per se. That is, when a sample is drawn, it is impossible for the researcher to have specified beforehand the population from which the sample was chosen. Rather, the population is hypothetical. As Fisher (1955) argues:

... there is always, as Venn (1876) in particular has shown, a multiplicity of populations to each of which we can legitimately regard our sample as belonging; so that the phrase "repeated sampling from the same population" does not enable us to determine which population is to be used to define the probability level, for no one of them has objective reality, all being products of the statistician's imagination. (p. 71)

---

2 Of course, this doesn't include surprises of what the experimenter expected. My point is only that the experimenter should have *foreseen* the possibility of the actual outcome, even if unexpected.
Exactly what Fisher meant by infinite hypothetical population, is not at all clear. Kendall (1943), obviously confused by Fisher's postulate, states, "This is, to me at all events, a most baffling conception" (p. 17). Gigerenzer (1993) has said, "the concept of an unknown hypothetical infinite population has puzzled many" (p. 321). The logic behind Fisher's argument has been questioned by some. Opposition to this construct will be discussed later in the chapter.

A fourth component of Fisher's statistical theory is the testing of just one hypothesis -- the null hypothesis. The addition of an alternative was introduced by Neyman and Pearson, and formed an integral part of their model of hypothesis testing. Fisher was adamantly against testing an alternative hypothesis to account for experimental results not explained by the null. As Fisher (1966) explains:

It might be argued that if an experiment can disprove the hypothesis that the subject possesses no sensory discrimination between two different sorts of object, it must therefore be able to prove the opposite hypothesis, that she can make some discrimination. But this last hypothesis, however reasonable or true, it may be, is ineligible as a null hypothesis to be tested by experiment, because it is inexact. (p. 16)

In this, Fisher holds that the opposite hypothesis, or alternative hypothesis, can never be staged as a hypothesis to be "nullified" because it is not precise enough to be under test. The null hypothesis, as Fisher (1966) explains, "must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the 'problem of distribution,' of which the test of significance is the solution" (p. 16). Since the
alternative hypothesis does not exhibit these characteristics, it is invalid to test it in any way with a significance test, and it is questionable whether one can infer it when the null is shown to be false. For Fisher then, the alternative hypothesis simply has no place in significance testing when the purpose is that of scientific inference.

Regarding the treatment of the null, Fisher (1966) says:

... it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis. (p. 16)

Therefore, no matter how many times a null fails to be rejected, it never in itself is proved. A null hypothesis can never be shown to be true. All the experimenter can hope for is to possibly reject the null hypothesis. That for Fisher is the purpose of using significance testing in an experiment. As noted by Gigerenzer (1993), Fisher later said that, "It is a fallacy ... to conclude from a test of significance that the null hypothesis is thereby established; at most it may be said to be confirmed or strengthened" (p. 73). From this it would appear that Fisher is leaning towards a confirmation theory of the null, yet this inference depends on how one interprets his use of the term "established" as being different from the term "confirmed". As Gigerenzer (1993) notes, Fisher never explained further how a non-significant result might possibly act as support for the null hypothesis. The reader is left confused by Fisher's writings.
A fifth component of early Fisherian statistics is that of significance levels. Fisher was vague as to what level of significance the researcher should adopt in testing the null hypothesis. This vagueness is a hardly surprising feature of his work on significance testing. As Gigerenzer (1993) notes, his writings on significance testing "had a remarkably elusive quality, and people have read his work quite differently" (p. 316). His recommendations are often conflicting. For instance, early on in Design of Experiments (1966), Fisher says, "It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result" (p. 13). Later however, on the same page, Fisher says the following:

It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore [emphasis added] all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. (p. 13)

Since then, Fisher (1956) has said, much in response to the "alpha" definition proposed by Neyman and Pearson, that "... no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" (p. 42). Through these passages, Fisher gives an ambiguous instruction as to which significance levels to use and when to use them. I stress
however that Fisher's most extreme recommendation for probability values at or below the .05 level, was that "it is usual and convenient"; he never implied that a paper's scientific value should be judged on this basis alone, or that publication decisions should be made on meeting this sole criterion. Indeed, as noted by his daughter, Joan Fisher-Box, later in his career, Fisher himself regarded the significance test to be a rather "weak" argument. As Fisher-Box (1978) comments, "much of his early work [Fisher's work] has been devoted to what he came to regard as the lowest level of scientific inference -- to tests of significance which make a dichotomy between hypotheses that are discredited by the data and those that are not" (pp. 447-448).

In summarizing Fisher's notion of significance levels, Gigerenzer (1993) argues for two categories for his ideas. The first was that of a standard level of significance, which consisted of a conventional standard (i.e., .05) that could be adopted by researchers. This was Fisher's early position. The second position became apparent near the end of Fisher's career; that of an exact level of significance, for which the level (the exact level, e.g., .03) was noted in publication. It would appear that psychologists adopted Fisher's early view, despite what he had to say later in his career. The concept of significance levels remains perhaps the most important feature of Fisherian significance testing. Yet because of Fisher's ambiguity in explaining this all-important concept, they remain quite possibly the most misunderstood and controversial component of his entire statistical theory.
Related to significance levels, were Fisher's ideas regarding publication policies. According to Gigerenzer et al. (1989), with whom I agree, Fisher's discussion of the relation between a significant result and the demonstration of a phenomenon suggests that both significant and non-significant results should be published, for the purpose of being able to compare the relative frequency of the significant to the non-significant results. This in turn would supply the literature with a relative comparison and through this, the establishment of a phenomenon would become apparent. However, a precise ratio of "significant vs. non-significant" results that would serve to demonstrate the existence of a phenomenon was never outlined by Fisher. More recently, the problem of not accounting for non-significant results has been called the "file drawer problem". As noted by Rosenthal (1979), the problem arises when one considers the possibility that journals are filled with the 5% of studies that constitute Type I errors, while those studies not published (i.e., the file drawers) are filled with the 95% that show non-significant results. Had significance testing remained Fisherian, the "file drawer problem" would likely not exist today.

This component may be argued to have little to do with Fisher's theory per se, and everything to do with publication policy. I would venture to disagree with this and argue that because significance levels are so influential in publication decisions, publication should be discussed as part of Fisherian theory. If publication is to include those documents that are part of "knowledge" in general, then what is allowed to be included in that category has serious implications for what is acknowledged as progress in a field of study. In other words, publication is a
derivative of the word "public," and it is assumed that anything not published, is for all purposes, not known to the community of researchers. Having said this, I quote Fisher (1966): "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment that will rarely fail to give us a statistically significant result" (p. 14). In this, Fisher implies that both significant and non-significant results should be published. How else can we account for both positive and negative results if they are not published? How are we supposed to know how many "failures" occur if we do not document them, as we do positive results? Fisher would have it that a ratio of positive publications be contrasted with negative publications that would in turn represent the existence or non-existence of a phenomenon. However, as I already mentioned, he did not specify what magnitude this ratio shall take before a phenomenon is reputed as "existing." There is no question that Fisher would not regard a single significant result as evidence for the existence of a phenomenon, but we are unfortunately left with an incomplete account of how Fisher would address various ratios. This leads to the question of what ratio of significant to non-significant results would Fisher accept as deeming a phenomenon "significant"? Unfortunately, he provided us with no answer except to say that any result is provisional upon further experimentation.

A seventh feature of Fisherian significance testing concerns the sensitivity of an experiment. What Neyman and Pearson (1928) called power, I argue, is closely allied to Fisher's sensitivity. I say this with some reservation because Fisher never used, nor liked, the term power. Despite this, Fisher did recognize the utility of
considering the size of the sample used in relation to the desired effect size, however further elaboration on this point had to wait until Neyman and Pearson properly defined "power." Cohen (1962) later contributed to the concept of power by providing relatively easy methods to calculate it. A full discussion of power is included in the following two chapters, but for now it suffices to say that Fisher did acknowledge the importance of sample size (which is the major determinant of statistical power) and estimating effect when designing the typical experiment. Although Fisher rejected the concept of power as propounded by Neyman and Pearson, I argue that while he may have disregarded the way Neyman and Pearson wanted to use the concept, he did not disregard the meaning or utility behind it. In Design of Experiments (1966), Fisher says the following:

By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or, in other words, of a quantitatively smaller departure from the null hypothesis. . . . We may say that the value of the experiment is increased whenever it permits the null hypothesis to be more readily disproved. (pp. 21-22)

Based on the above quote, it is clear that Fisher places great emphasis on sensitivity. He claims the value of an experiment to be increased when sensitivity is taken into consideration. Thus, it is clear that Fisher did not fundamentally disagree with the idea of Neyman and Pearson's power, but only that he did not approve of the concept being used in decision-making. It is likely that Fisher rejected the concept of
power largely because he rejected the very foundations of Neyman and Pearson's Type II error, which forms the basis of power calculations. Fisher (1955) argued that calculations of power reflected the "mental confusion" between technology and scientific inference (p. 73). He did not however, explain his reasoning for this. Surely, the concept of power as applied to statistical inference does not diminish the quality of scientific inference. I suspect that Fisher disliked the power advocated by Neyman and Pearson simply because they were applying it to quality control and to a decision-making process, and that because of this, it could not be used as a part of scientific inference. Since Fisher was strongly opposed to Neyman and Pearson's decision-making for the purposes of scientific inference, it is likely that Fisher rejected the concept of power because it originated in the "decision-making" context. However, as has been shown by Cohen, power does have a place in scientific experiments, and in no way diminishes their "scientific" value. I argue also that Fisher rejected power in totality more because of the personal conflict with Neyman and Pearson than because of a genuine dislike for it. Surely, Fisher must have recognized the value of power in addressing his concerns of experimental sensitivity, not in the context of decision-making, but in the context of scientific inference. As Gigerenzer (1993) has said, "The concept of power makes explicit what Fisher referred to as 'sensitivity'." It is thus surprising that Fisher would have wholeheartedly rejected Neyman and Pearson's power based solely on theoretical issues. A possible answer is that since the development of power tables had to wait until the 1960s, Fisher was blind to the utility of such a tool in his statistical theory. Fisher did
note however ways in which the sensitivity could be increased without necessarily increasing sample size. These methods included qualitative procedures\textsuperscript{3}, such as structural organization of the experiment and refinements of technique. He did however, mention these second to sample size increments.

Fisher's position with regards to sensitivity can be summarized to mean that each experimenter should take into consideration the sample size used when seeking a particular effect. Fisher wanted us to consider the sensitivity of the experiment\textsuperscript{4}, something that psychologists to a large extent, have neglected in designing their research.

The following then, summarizes the main components of the NHST model as proposed by Fisher.

1. Forecast all possible results \textbf{before} the experiment is executed; know beforehand the interpretation of each possibility, should it arise.

2. It is imperative that subjects be assigned to conditions \textit{randomly}. Only through randomization can error be properly estimated. Estimation of error presides over quantity of error.

3. The sample is drawn from an \textit{unknown hypothetical infinite population}. Fisher held that we could not \textit{know} the population from which our sample was drawn, but rather could only "imagine" the given population.

\textsuperscript{3} Chow (1996) includes a discussion of these qualitative adjustments as a way of increasing power. Thus, power is not solely a term used in relation to sample size.
4. There is only one hypothesis to be tested -- the null hypothesis. The null hypothesis cannot be established, yet it possibly can be "confirmed" or "strengthened."

5. Using a significance level of .05 is convenient, but not mandatory (early Fisher). The exact significance level should be reported in publication (late Fisher).

6. Both statistically significant, and non-significant results should be published, as to yield a relative frequency from which a phenomenon could be shown to exist.

7. A researcher must consider the sensitivity of an experiment by either enlarging the number of repetitions [i.e., sample size], or by qualitative methods. Power as advocated by Neyman and Pearson was inappropriate if the goal was scientific inference.

Today's Current Model of Null Hypothesis Significance Testing: Comparable to Fisher's?

In considering the current model of NHST, I am immediately confronted with a problem. Shall I outline the model recommended by textbook authors, or shall I outline the model that is actually used by researchers? What is true in textbooks isn't of course, what is always practiced. I will attempt therefore to draw on information from both sources, placing more importance on one source over another where I see fit, in an effort to account best for today's model.

Components 1 through 3 are roughly satisfied by today's model, if only in theory. By this I mean that these components, especially 1 and 2, are recognized as

---

4 It should be noted that while Fisher spoke of the sensitivity of the experiment, Neyman and Pearson related power to that of the statistical test. This difference may or may not have had an influence on Fisher rejecting the concept of power, and implicitly denying its relation to sensitivity.
being important even if they cannot be completely satisfied to Fisher's approval. The remaining 4 points (4 through 7) constitute strong evidence for my claim that we cannot name today's model "Fisherian." I will expand on those points shortly. First however, a brief discussion of components 1 through 3 is required.

The first component, that of forecasting all possible outcomes and knowing beforehand the interpretation of each of these, is at least partly satisfied by today's model. Most researchers design their experiments with conscious attention to predicting possible outcomes, thus the process of specifying hypotheses. Whether today's investigators are prepared to interpret all possible results is debatable. This may be largely due to the fact that there are often infinitely many possible outcomes when continuous measures are used. In relation to whether we follow Fisher's model, I score today's NHST a "yes-no" on this component; "yes" in that we specify our hypotheses beforehand in effort to forecast possible results, and "no" in that we are often unprepared to account for results that deviate from our predicted outcomes. If results do not follow as expected, we are often left formulating post-hoc explanations to account for these unexpected findings. Fisher stresses that the theory must precede experimentation, and while today's psychologists attempt to fulfill this requirement, it is many times left unfulfilled.

The second component of Fisher's model is that of randomization. Every member of the population should have an equal chance of being included in the sample. While today's textbooks do stress the randomization of samples from populations, in practice, this is seldom fulfilled. That is not to say that researchers are
not (most of the time) properly ensuring that each member of the population is randomly selected. The biggest problem occurs when researchers generalize their sample-based results to populations from which a random sample was not drawn. Often, experimental studies will use a very restricted sample base, then over-generalize the results to a wider population than that from which the sample was drawn. For instance, a sample of university students is only generalizable to a very narrow band of population parameters. However, we often see discussion sections generalizing to wider populations.

The way in which subjects are recruited today would likely not satisfy Fisher. Even the most common and valid (however, this is of course debatable) method of random selection, that of telephone number sampling, can only be generalized to the population consisting of those subjects that both have a phone, and are listed in the phone book used in sampling. Because Fisher's methods were intended for agriculture, an analogy of this would be an investigator selecting those plots of land that were listed in the community's property listings. Such an investigation is still valid and useful, but the results can only be generalized for those land estates listed. Perhaps the land not listed was not fertile enough and therefore was not worth listing? This is similar to the individual who is not listed in the telephone directory because he suffers from major depression, which in turn results in him not being able to work, which in turn results in not having the funds to afford a telephone. If the study were recruiting a sample to study the proportion of the population that suffers from major depression, the methodological problem is obvious. Even more
methodologically unsound is recruiting subjects by advertisement, then attempting to generalize the study's results to a wider population than those subjects that were volunteers. But how many times do we read results in the form, "These results suggest that men volunteers significantly differ from women volunteers on variable X (please note that these results can only be generalized to the 'volunteer-type' subject)?" As stated succinctly by Howell (1989), "one person's sample might be another person's population" (p. 4). I argue that Fisher would have charged us with failing to recognize this. Thus to conclude this point, Fisher would advocate that while we are often randomly selecting, we are not randomly selecting from the population to which we generalize our results. It should be noted however, that Fisher may have well understood this to be a practical problem, and not a "true" departure from his original model.

We Randomly Select, Only From the Wrong Populations

The third component of Fisher's model consists of a hypothetical assumption that in practicality, does not influence modern research practices to a significant degree. Fisher basically holds that we cannot begin to specify the population from which our sample is drawn because we are unaware of it. If we were aware of it, then why would we have to sample in the first place? Hacking (1965) believes the idea of hypothetical infinite populations contributes to unnecessary confusion. He argues that chance set-ups should be used to describe long-term frequency. He says, "However much they [hypothetical infinite populations] have been a help, I shall argue that hypothetical infinite populations only hinder full understanding of the very
property von Mises and Fisher did so much to elucidate" (p. 7). Later, Hacking continues, "One hopes our logic need not explicitly admit an hypothetical infinite population of tosses with this coin, of which my last ten tosses form a sample. Chance-set-ups at least seem a natural and general introduction to the study of frequency" (p. 25). By "chance-set-up," Hacking is referring to a system in which there are conducted experimental trials, of which each single trial is a member of a more complete class of possibilities. Thus for Hacking, having a class of possibilities is more enlightening and logical than having the population be infinite, as Fisher holds. The population is a long-run frequency of possibilities, yet not infinite as Fisher would have. The debate of whether we sample from finite populations or hypothetical populations is a philosophical one, and shall be left to the philosophers of science to grapple with. The implication of either has no direct influence on how we practice hypothesis testing today. I would argue that few practicing researchers have given such a topic much thought, so while this component is a part of Fisherian NHST, whether it is even acknowledged by today's practitioners of NHST is unknown. Either way, in the world of significance testing, as the expression goes "we have much larger fish to fry."

Hardly Fisherian

The following 4 components of Fisher's model are almost completely disregarded by today's researchers and journal editors. Hence, these following 4 points strongly show that today's NHST is dissimilar to yesterday's Fisherian model.
Component 4, that of positing only one hypothesis (the null) before analysis is not followed in the least by today's researchers, or textbook writers. Today's procedure is that of setting up a null hypothesis and an alternative hypothesis. Should the null hypothesis be rejected, the investigator infers the substantive alternative as the most plausible argument to account for the data. As will be discussed in chapter 2, the substantive, or conceptual hypothesis, is the hypothesis that is held to best account for the data, given that the null is false. Usually, one substantive hypothesis is specified. The statistical alternative on the other hand, can be stated merely as "not the null," in that it suggests a distribution other than the null to account for the data. The primary difference between the statistical and the substantive hypothesis is that while the statistical hypothesis is simply a statement of "not the null," the substantive hypothesis constitutes an effort to explicitly account for the rejection of a null hypothesis. Although given a rejection of the null, the statistical alternative is likely true, the substantive hypothesis may be only one of many hypotheses that can best account for why the null was rejected. In this respect, the substantive alternative serves as something of an "explanation" of why the null was rejected. As argued above, Fisher was skeptical about inferring an alternative hypothesis. The introduction of an alternative hypothesis is a Neyman-Pearson innovation, and was applied in the context of decision-making -- what Fisher would refuse for the purposes of scientific inference. In Neyman-Pearson terms, the user of statistical methods needs to make a decision between two alternatives, not simply reject an unlikely hypothesis. While Fisher was not against using an alternative hypothesis
when making decisions in industry, he totally rejected them for use in the field of pure scientific investigation. It is worth quoting Fisher (1966) extensively here for an acute sense of his position on what he called "Acceptance Procedures":

The situation is entirely different in the field of Acceptance Procedures, in which irreversible action may have to be taken, and in which, whichever decision is arrived at, it is quite immaterial whether it is arrived at on strong evidence or on weak. All that is needed is a Rule of Action which is to be taken automatically, and without thought devoted to the individual decision. The procedure as a whole is arrived at by minimizing the losses due to wrong decisions, or to unnecessary testing, and to frame such a procedure successfully the cost of such faulty decisions must be assessed in advance; equally, also, prior knowledge is required of the expected distribution of the material in supply. In the field of pure research no assessment of the cost of wrong conclusions, or of delay in arriving at more correct conclusions can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence; moreover, accurately assessable prior information is ordinarily known to be lacking. Such differences between the logical situations should be borne in mind whenever we see tests of significance spoken of as "Rules of Action". *A good deal of confusion has certainly been caused by the attempt to formalise the exposition of tests of significance in a logical framework different from that which they were in fact first developed* [emphasis added]. (pp. 25-26)
In the above quote, Fisher leaves little doubt of how he feels with regards to his tests being used in the field of so-called "Acceptance Procedures" or used as "Rules of Action." According to him, these procedures are not to be used for the purpose of judging scientific evidence -- period. Today however, researchers continue to utilize these "Acceptance Procedures" as a model for establishing scientific evidence. I would venture to suggest that on this fact alone, Fisher would want little to do with today's NHST.

Component 5, that of the "convenient" use of a significance level of .05, is totally dismissed by today's textbooks and by practically most researchers and journal editors. Fisher stated that the .05 level of significance is "convenient," not gospel. Today, we seldom find an experimental study in which the .05 level has not been used. An exception to this occurs when we read a significant result at the .01 level. Rarely, if ever, do we find a significance level of over .05, even if the difference is slight. Journal editors have been found to be quite rigid in their demand for the null to be rejected at least at .05 (e.g., see Melton, 1962). As noted above, Fisher required experimenters to state precisely the significance level when reporting results, the exact probability. This procedure is not followed in the least today, where the common "p < .05" is reported. Huberty (1993) found that many recent textbook authors suggested choosing a significance level prior to data collection. The data may yield a probability of .03, however today's journals list it as merely lower than .05, despite the fact that the most recent edition of the APA Publication Manual (1994) cites both ways as acceptable. Therefore, the combination of adhering to a
rigid probability level (contra early Fisher), and not reporting exact significance levels (contra late Fisher), are two extremely important elements in significance testing that run opposite to Fisher's recommendations. Furthermore, as noted by Huberty, despite Fisher's rejection of a fixed level of significance for all experiments (see Fisher, 1959), some researchers still cite him as support for their choice of the .05 level of significance. The issue surrounding significance levels is perhaps the most compelling reason why we can hardly attribute today's model to Fisher -- we simply do not do significance testing as Fisher prescribed.

Component 6, that of publishing both significant and non-significant results, is also completely dismissed by today's researchers and journal editors. As noted by Gigerenzer et al. (1989), Fisher would have wanted us to be publishing both significant and non-significant results, so that a relative comparison between the two groups of experiments could be made to ascertain the existence of a given phenomenon. Indeed, only by a comparison of the two groups can we claim with any kind of accuracy that a phenomenon exists. Today's methods overlook the importance of accounting for negative results, and these are typically not published. There have been some efforts to change this state of affairs. Neher (1967) for instance, argued that psychological research falls victim to what is referred to as "probability pyramiding," a process in which the significant outcomes are reported more faithfully than the insignificant outcomes. As Neher demonstrates:

In an extreme case, for example, 20 different analyses might be done; 19 may be insignificant at the 5 per cent level and the 1 analysis that is significant may be
reported as a finding. Obviously, this is about what one would expect to find on the basis of chance alone, so that this is likely to be a spurious finding (a Type I error). (p. 257)

This of course, is a major consequence of not counting negative results. A way around this problem would be to account for both significant and non-significant results, thus being able to directly compare the frequency of each. Smart (1964) has also noted the importance of negative results in psychological research. He found that only 9 per cent of the aggregate of papers published were projecting negative findings. I would argue that today, this figure is probably even lower. Journal editors do not want to publish research that barely fails to meet the .05 probability level, never mind being outright negative! Fisher would disagree with today's "positive-only" publishing. Also, implicit in point 6, is the idea that experiments should be replicated. Only by replicating an exact experiment could we possibly arrive at a ratio of "positive vs. negative" results. Today, psychologists do neither. Negative results are not accounted for, and exact replication is almost non-existent. Fisher would not approve.

Component 7, that of the sensitivity of a test, is again, largely disregarded by today's research psychologists. Fisher wanted us to consider the sensitivity of an experiment in being able to reject the null hypothesis. As mentioned earlier, Fisher advocated a similar construct as the Neyman-Pearson idea of power. Today, researchers neither concern themselves with the sensitivity or the power of their research. Both Cohen (1962) and Rossi (1990) have found the power of most studies
they reviewed to be quite low. Power in most textbooks is seldomly discussed adequately, if at all. Furthermore, as with many other statistical concepts, power is poorly defined in some texts (Brewer, 1985).

As mentioned earlier, it is tempting to argue that Fisher's sensitivity is similar to Neyman and Pearson's power. Except for the idea of estimating the Type II error rate, Fisher's idea of sensitivity is very similar to the idea of power. Again, perhaps Fisher would have been more accepting of power after Cohen devised ways of actually calculating it. Either way however, it is a well known fact that today's researchers devote little attention to either the sensitivity or the power of their experiments. As will be more thoroughly discussed in chapter 3, power surveys have suggested the calculation of power to be almost non-existant in psychological journals, and when calculated, to be quite low (Cohen, 1962; Rossi, 1990). How many times have you heard a discussion of sensitivity in journal articles? Fisher (1966) wanted us to think of sensitivity as a way of improving our experiments, as he said "it [sample size increase] will allow . . . of a quantitatively smaller departure from the null hypothesis" (p. 22). Researchers today however, seem almost indifferent as to the number of subjects they use. Instead, they use rough guidelines and hope for the best. This has resulted in some post-hoc sample size studies which make it almost impossible to reject the null hypothesis (Tversky & Kahneman, 1971). A full discussion of power is included in chapter 3. For now, it is enough to say that Fisher's sensitivity, or Neyman and Pearson's power, are rarely addressed in
present psychological research, hence again further distancing today's model from Fisher's original model of significance testing.

**Fisherian Felony?**

Because many psychologists still believe we employ a Fisherian model when analyzing data, this has resulted in unjust criticisms directed towards Fisher -- laying blame on Fisher for things he never supported in the first place. In ending this chapter, I think it appropriate to cite a typical instance where Fisher was wrongly faulted for ideas that were not his own. I hope that the above analysis of NHST will help make obvious the misattribution in the following accusation.

Meehl (1978) is perhaps the most harsh with Fisher. He blames Fisher outright for our misuse of significance testing. He says:

I suggest to you that Sir Ronald has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas [of psychology] is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology. (p. 817)

Meehl is correct, the over-reliance on refuting the null in psychology is detrimental to our discipline -- but this is not Fisher's fault. First, Fisher did not lead us anywhere, we (i.e., psychologists) lead ourselves astray. Meehl's critique is that much less credible when one considers that Fisher never recommended his procedures for psychology! I dare ask how he could have led us down the primrose
path when he never even suggested we follow him! Meehl automatically associates modern significance testing with Fisher's name, and if I've shown nothing else is this chapter, doing this is misleading, and hardly fair to Fisher's legacy.

In closing, I have argued I hope successfully, that to ascribe Fisher's name with today's NHST, is nothing short of an academic misdemeanor. Today, researchers do something vastly different than what Fisher once proposed, and to call today's statistics "Fisherian," does not do justice to the statistical genius. Today's model is much too hybridized, misused, and misunderstood to be attributed with the likes of a statistical pioneer as Fisher. Therefore with Cowles, I hope you will concur -- to call today's procedures "Fisherian" is indeed likely to cause Fisher to unduly shuffle in his tomb. Let's allow the Master to rest, once and for all.
Chapter 2

Problems with NHST: Model or Misuse?

The purpose of this chapter is two-fold. First, it is intended to serve as a descriptive and analytical survey of the problems associated with current null hypothesis significance testing. A primary goal of this chapter will be to thoroughly review and evaluate the objections raised against NHST since its origin. In evaluating these objections, it will prove beneficial to differentiate those criticisms that are correctly directed toward the model from those that reflect more the lack of understanding of its users. The latter constitutes the second goal of the chapter. For instance, Brewer (1985) found many misconceptions in statistics textbooks, including many regarding the nature of hypothesis testing. Huberty (1993) has gone as far as to say, "it is not statistical testing itself that is at fault; rather, some of the textbook presentation, teaching practices, and journal editorial reviewing may be questioned" (p. 317). Do these misunderstandings lie at the "root" of common criticisms against NHST? Cohen (1994) has noted the "wishful thinking" that many researchers are guilty of when interpreting NHST. He argues that researchers are forever trying to get NHST to tell them something that it cannot -- the probability of the research hypothesis.

The differentiation between inherent problems of the model and misunderstandings is a necessary prerequisite to considering alternatives to NHST, a topic that will be covered in length in chapter 3. If a criticism is indeed illegitimate
(i.e., one that is misguided), then it may be that a further understanding of the model is required rather than a competing alternative. As Chow (1996) appropriately remarks:

That someone misuses a tool does not necessarily speak ill of the tool. The problem may be due to the user. For example, the user may have misread the instruction accompanying the tool. Alternatively, the user may have been misled by the instruction provided by the tool manufacturer. (p. 11)

More recently, some have suggested (e.g., Estes, 1997) that a neglect in the teaching of the mathematical foundations of statistical design and analysis is to blame for this lack of understanding. Surely, a model cannot be deemed inappropriate and unworthy merely as a result of sloppy teaching and poor interpretation by its users -- or can it? Can this fact alone be reason enough to fault the model? Shall we adopt the view advocated by Sohn (1993): "What is the role of the null hypothesis test? The risk of misusing the test is so great and the benefits so problematic that the answer must be none" (p. 1174)? If a model is so resistant to human understanding, then should this not qualify itself as a drawback and criticism against the model? Some space will be allotted for a brief discussion of this question, and an appropriate answer will be provided.

The Problems with NHST

The criticisms of NHST go back almost as far as the presumed origin of the model. For instance, Berkson (1938) criticized NHST as an inappropriate means of decision-making based on the $p$ value alone. Since Berkson, most of the criticisms
directed at the model have been repeated over and over throughout the literature and thus, few can be regarded as original. There exists a plethora of criticisms aimed specifically at NHST, many of them similar (or the same) in content. Without a doubt, NHST is a well-beaten horse.

This chapter will detail and evaluate the criticisms against NHST since Fisher. A categorical review of the problems of NHST is best, in that most individual critiques fit into one of the following eight categories listed below.

1. Sample Size Sensitivity

Sample size sensitivity is perhaps the most common criticism against the NHST model. That significance tests are partially a function of sample size, for many critics, constitute a major problem. Berkson (1938), Nunnally (1960), Bakan (1966), Lykken (1968), among others, have pointed out that any effect can prove significant given an ample sample size. In other words, a significant result is always achievable -- if you have a large enough sample. Nunnally (1960) for instance, says the following when discussing sample size and its relatedness to statistical significance:

Experience shows that when large numbers of subjects are used in studies, nearly all comparisons of means are "significantly" different and all correlations are "significantly" different from zero. . . . The point of view taken here is that if the null hypothesis is not rejected, it usually is because the N is too small. If enough data is gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data. (p. 643)
Nunnally provides an example in support of the above argument. By taking a sample of 700 subjects in a study of public opinion, he found correlations as small as .08 to be statistically significant at \( p < .05 \). He concluded that many of these correlations were of little practical or theoretical interest.

Bakan (1966) also criticizes NHST based on the problem of "sample size sensitivity." Like Nunnally, he also provides a "real-life" demonstration. Running significance tests on a battery of tests comprising of over 60,000 subjects, Bakan found every test to come up significant. Like Nunnally, he notes that even when differences between means are quite small, significant \( p \) values are found. More recently, Kirk (1996) has summarized the problem: "One researcher, for example, might obtain a \( p \) value of .06 and decide to not reject the null hypothesis. The other researcher uses slightly larger samples and obtains a \( p \) value of .05, which leads to a rejection. What is troubling here is that identical treatment effects can lead to different decisions" (p. 748).

Allow me for a moment to expand on the logic of the examples cited above. It should be noted that every variable imaginable is correlated with every other variable imaginable. The correlations differ only in degree. While it may be foolish to consider the correlation between number of taxi cabs in a large city, with that of number of jam jars sold in that same city, one can be sure a correlation does exist, if only a minute one. As the number of taxi cabs increase, jam jars likely increase or decrease to some degree, even if very small. Mathematically, a correlation between taxi cabs and jam jars will always exist. What concerns critics is that such a trivial
correlation between two variables has the potential to be deemed significant if a large enough sample is chosen for the analysis. What information can such a small correlation yield? Very little to be sure. The moral of the story is that if p values are used as the sole criterion for judging the presence of an effect or the importance of a result, then any experiment can yield "significant" values given an ample sample size.

The problem of sample-size-sensitivity is opposite to the problem of low power. If a sample N is very low, the statistical test may be unable to reject a null hypothesis if indeed it is false. Therefore, with NHST we have a two-way problem. If the sample is too small, significance cannot be achieved, regardless of whether or not variations of the independent variable are having an effect on the dependent variable. At the opposite end, if sample size is too large, significance will be obtained, regardless of how large the effect is.

Related to sample size, is the fact that statistical significance does not necessarily imply a large effect size. Although a result may prove statistically significant, its effect size measure may not be large. As Oakes (1986) remarks:

... test statistics are functions both of effect size and of sample size. Hence a statistically significant effect may result from the testing of a large effect with a small sample or from the testing of a small effect with a large sample. (p. 51)

Thus, as stated by Oakes, a large sample may produce a statistically significant result, but the effect may be small. Many argue that effect size measures should be
used to assess results, one reason being that they are "insensitive" to sample size. Indeed, as even Hays (1963) has made clear many years ago:

The occurrence of a significant result says nothing at all about the strength of the association between treatment and score. A significant result leads to the inference that some association exists, but in no sense does this mean that an important degree of association necessarily exists. (p. 342)

Evaluation

The problem of sample-size-sensitivity can be regarded as both an inherent problem with the model and a "misuse" issue. What makes it problematic is that the mathematical derivation of the p value is partially a function of sample size. The reality highlighted by Kirk (1996) that two identical experiments can lead to different conclusions because of differing sample sizes used, is problematic. However, this is only the case because most researchers pay undue attention to the p value and disregard other potentially meaningful measures (e.g., effect sizes). Thus, two studies performed identically but yielding differing p values should not be automatically interpreted as presenting differing effects; if they are, this is the fault of the interpreter, not the model. Hence, the "sample size problem" should be regarded both as a "misuse" problem and an inherent problem of the model. Researchers misinterpret sample-size issues, but even with a favorable sample size, many critics would argue that NHST still does not tell us what we want to know.
2. The Prior Falsity of the Null Hypothesis

A second problem cited in the literature, is that typically, the hypothesis to be tested (i.e., the null hypothesis) is false before analysis even begins (Savage, 1957; Cohen, 1990; Thompson, 1996). The general argument is that we have rejected the null hypothesis long before analysis is run, so why bother with the experiment? If the goal of a given analysis is to reject the null hypothesis, critics claim that this can be done long before experimentation takes place. They argue that the null can never be true, and therefore must always be false. Using a simple independent t-test as an example, critics claim that there is always a difference between the two means before we even collect data; the means will always be different to some decimal place. If this is the case, then why test a hypothesis of no difference, when we already know there is a difference?

I argue that this criticism stems from a lack of understanding of how to properly define the null hypothesis. Typically, again using a t-test, the null hypothesis is defined as the hypothesis of no difference or the hypothesis stating \( \mu_1 = \mu_2 \). The critique is that we all know beforehand that the mean of population 1 cannot equal that of population 2. As Loftus (1996) says, "rejecting a typical null hypothesis is like rejecting the proposition that the moon is made of green cheese. The appropriate response would be 'Well, yes, okay... but so what?'." Meehl (1967), provides a good explanation of the problem:

In order for two groups which differ in some identified properties (such as social class, intelligence, diagnosis, racial or religious background) to differ not at all
in the "output" variable of interest, it would be necessary that all determiners of
the output variable have precisely the same average values in both groups, or
else that their values should differ by a pattern of amounts of difference which
precisely counterbalance one another to yield a net difference of zero. Now our
general background knowledge in the social sciences, or, for that matter, even
"common sense" considerations, makes such an exact equality of all determining
variables, or a precise "accidental" counterbalancing of them, so extremely
unlikely that no psychologist or statistician would assign more than a negligibly
small probability to such a state of affairs. (p. 108)

Meehl's comments are well-taken and are perfectly logical. However, I am going
to venture to suggest that the problem of "prior falsity" would disappear if the null
hypothesis were defined properly in the first place. Furthermore, I will argue that this
misunderstanding has existed since Fisher's innovation of the NHST model in the
first quarter of the twentieth century. As I will demonstrate, Fisher himself may be
responsible for this ill-defining of the null hypothesis.

As previously mentioned, the null is traditionally defined as the hypothesis of no
difference, the hypothesis that two means (in the case of a simple t-test) are exactly
the same. I would go one step further than Meehl, and suggest that even if you
weren't a statistician or psychologist, but had only an elementary grounding in
mathematics, you would know that two means cannot be exactly the same to an
infinite number of decimals. Thus, Meehl's argument is perfectly valid if the null
hypothesis is indeed defined as the "hypothesis of no difference." However, I suggest
that a change in definition of the null obliterates the "prior falsity" problem. Consider the following argument.

Most would agree that the null hypothesis represents the distribution such that chance is the governing influence. That is, the null hypothesis represents the circumstances for which any observed differences in the data are due to chance, or as formally known in statistics, "sampling error." If this is agreed upon, then why define the null hypothesis as an "equality" among means? Why do we in one instance define the null as the distribution that would result if only sampling error were the cause of difference, then in the next say there is no difference? As Meehl, Loftus, Cohen, and many others have pointed out, everyone knows there is a difference. The correct conception of the null hypothesis is to state it as the hypothesis for which any differences found are due to chance. We know there will be differences, the point is whether these differences are large enough (and thus, "significant" enough) to discount the plausibility of chance, and thereby, infer a hypothesis that chance factors did not produce the observed difference. This latter hypothesis would of course be construed as the alternative hypothesis, in accordance to the Neyman-Pearson model of hypothesis-testing. Hence, the appropriate question to ask is not whether the mean of population 1 is different from the mean of population 2, but rather whether the mean of population 1 is different enough from that of population 2 such that we are prepared to discount chance factors as producing this difference. The null hypothesis has nothing to do with equality, but has everything to do with chance. If the probability of the data are low enough given that the results are due to
chance (i.e., under the null hypothesis), then we reject the null, not because we have learned that there is a difference between means, but rather that this difference (which we always knew existed in the first place, as Meehl, Loftus, and others continually point out) is too rare to be attributed to chance. The further the means are away from each other in a t-test, the more likely we are to reject the null, because the less likely a difference this large can "reasonably" be attributed to chance factors. There must be something else producing this difference -- namely, treatment effects (or so we hope). The bigger the difference, the less likely it is due to simple chance factors. Rejecting the null hypothesis is not rejecting the hypothesis that no difference exists. It is rather rejecting the hypothesis that the differences, which of course are sure to exist, are due to chance.

Hagen (1997) is one of the few authors aware of the misunderstanding that results in defining the null hypothesis:

... the null hypothesis says nothing about samples being equal, nor does the alternative hypothesis say that they are different. Rather, when addressing group differences, the null hypothesis says that the observed samples, given their differences, were drawn from the same population, and the alternative hypothesis says that they were drawn from different populations. (p. 20)

Essentially, what Hagen is saying, is what I am trying to clarify. That is, the null hypothesis asserts that differences found between two groups (again, in the case of a t-test) are small enough to not conclude they come from different populations. Otherwise said, these differences are small enough to attribute to chance. Reject the
null, and in doing so, we infer that these samples were drawn from two different populations.

So, if this definition of "equality" is so problematic, why is it used as such? To know why, we must read Fisher's explanation regarding the nature of the null. In Design of Experiments (1966), Fisher says the following when describing the null hypothesis:

The two classes of results which are distinguished by our test of significance are, on the one hand, those which show a significant discrepancy from a certain hypothesis; namely, in this case, the hypothesis that the judgments given are in no way influenced by the order in which the ingredients have been added; and on the other hand, results which show no significant discrepancy from this hypothesis. This hypothesis, which may or may not be impugned by the result of an experiment, is again characteristic of all experimentation . . . In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null is never proved or established, but is possibly disproved, in the course of experimentation. (pp. 15-16)

In explaining the null hypothesis above, Fisher alludes to the famous tea experiment in which a woman claims to be able to differentiate between cups of tea. She professes to be able to differentiate between those in which the tea infusion was first added from those in which the milk was first added. Fisher implies that the null tested in this case is that the judgments made by the woman are in no way associated to whether the tea or milk was first added to each cup. We can't begin to claim that
her judgments are so influenced if the results "show no significant discrepancy from this hypothesis." Should the results show a discrepancy from the null hypothesis, then we may begin to speak of such an influence. What Fisher doesn't explain here is how we are able to ascertain whether the results do show a discrepancy from the null hypothesis. The answer is of course, whether or not the result is better accounted for by chance alone. If the results show a discrepancy that is large enough to discount chance factors, then we shall reject the null hypothesis.

A few pages later, again referring to the typical experiment, he says, "the most general statement of our null hypothesis is, therefore, that the limits to which these two averages tend are equal" (p. 35). It is in such a statement where the notion of equality among means became (unfortunately) a defining property of the null hypothesis. However, even in this explanation, Fisher only describes the "tendency" of the means, not a static equality. This is to mean that the hypothesis tested is that the two means "tend" to be equal; whether they are exactly equal is not at issue. Of course, this is easily misinterpreted by today's users, evidenced by many testing the hypothesis of no difference or that of equality among means.

**Evaluation**

In evaluating the criticism of "prior falsity," I suggest that the problem would disappear if the null hypothesis were defined as the "chance hypothesis" rather than the hypothesis of "no difference". This ill-defining may be attributed to a general misunderstanding of the null beginning with Fisher. Until this is clarified, the "literal" interpretation of the null hypothesis is likely to continue to cause problems
for students and researchers. In the end however, this is hardly a problem inherent to
the NHST model.

3. NHST Tests the Wrong Hypothesis

A third criticism against NHST is that it does not test the hypothesis of interest
to the investigator. As Cohen (1990, 1993) and others (e.g., Sohn, 1993) have so
often argued, NHST tells us nothing with regard to the hypothesis of interest, termed
the "research" or "alternative" hypothesis. It is not that it tells us little about the
alternative, it is not that it tells us something indirect about the alternative -- it tells
us nothing about the alternative hypothesis. As Cohen (1990) remarks, "If the p
value with which we reject the Fisherian null hypothesis does not tell us the
probability that the null hypothesis is true, it certainly cannot tell us anything about
the probability that the research or alternate hypothesis is true" (p. 1307).

To clarify, the basic logic of the NHST set-up is as follows. The researcher
states a null hypothesis to be tested, and construes an alternative hypothesis that he
thinks will account for the data, should chance be ruled out as producing any
observed differences. If the null is rejected, he infers the alternative hypothesis,
hoping that all extraneous variables were controlled in the experiment. It is this
element of experimental control that gives the alternative any plausible sense of
being a "correct" or "valid" hypothesis in the NHST procedure. Theoretically, if
every variable were able to be controlled, then inferring the alternative hypothesis
would not be such risky business. But the truth of the matter is that regardless of how
much experimental control we impose in our experiments, there is always the chance
that we've overlooked a crucial variable that may be causing differences on the
dependent variable. Cowles (1989) gives a good historic example of how this can
occur:

The names malaria [i.e., "bad air"], marsh fever, and paludism all reflect the
view that the cause of the disease was the breathing of damp, noxious air in
swamp lands. The relationship between swamp lands and the incidence of
malaria is quite clear. The relationship between swamp lands and the presence
of mosquitoes is also clear. But it was not until the turn of the century that it was
realized that the mosquito was responsible for the transmission of the malarial
parasite and only 20 years earlier, in 1880, was the parasite actually observed. . .

This episode is an interesting example of the control of a concomitant or
correlated bias or effect that was the direct cause of the observations. (p. 149)

In the preceding selection, we have a perfect example of how the logic of NHST
typically works. The null is that diseases such as malaria are caused by chance, that
is, that the disease occurs in individuals randomly, governed by mere chance factors.
The alternative hypothesis is that the disease is caused by breathing in swamp air. In
this case, had an actual experiment been performed with these two hypotheses, then
we would surely reject the null hypothesis, since after all, those individuals living
close to the swamp would have a higher incidence of malaria than those individuals
living further away from the swamp. Hence, we know that malaria is caused by
something, and can be explained in a more reliable way than by just chance factors.
After running the analysis, we notice that the probability of the data occurring given
that the null is true is very low (i.e., $p < .05$), so, we infer the alternative. In doing so, we have potentially overlooked other possible predictors (such as mosquitoes in this case) that may have produced the observed difference in our dependent variable of interest, that of incidence of malaria.

What I am trying to clarify is that the inferred alternative is nothing more than a "next best guess explanation" given the falsity of the null hypothesis. That was Fisher's main problem with positing an alternative, that of it not being "exact." The alternative is not exact, and that is why so many critics of NHST recommend alternative means of hypothesis-testing. NHST can be said to be a kind of "backward" procedure; we test a hypothesis that is not of great interest to us (i.e., the null), while not testing the hypothesis that is many more times of interest to the investigator. This, without a doubt, is the most serious difficulty with NHST -- that of not being able to test the research hypothesis. As will be discussed in chapter 4, Bayesianism offers a solution to this profound problem. In the Bayesian paradigm, one tests the research hypothesis directly, and thus conclusive statements can be made regarding the alternative, rather than "giant-leap" inferences, as is the case in classical statistics.

At this point, I should note the difference between the statistical versus the scientific or conceptual alternative hypotheses. As noted by Bolles (1962), although rejection of the null hypothesis may imply the statistical alternative, it may not necessarily imply the scientific alternative. These are two different hypotheses. This difference stems largely from the fact that statistical inference is not equal to
scientific inference (Morrison & Henkel, 1969). More recently, Chow (1996) has argued that although the statistical alternative may be quite easily inferred, this inference is much more difficult with regards to the conceptual alternative, because of what he calls the "Reality of Multiple Explanations." As discussed previously, a multitude of explanations could exist if the null hypothesis is rejected. According to the Neyman-Pearson model of hypothesis-testing, if the null is rejected, the alternative hypothesis is inferred — that is, the statistical alternative. However, it should be noted that merely because the statistical alternative is inferred, this does not directly imply the inference of the conceptual alternative. As Chow notes, "multiple conceptual alternative hypotheses give rise to their respective statistical alternative hypotheses" (p. 55). This would mean that in order to have any statistical alternative, you first need to "conceive" of an alternative conceptual hypothesis. Once the conceptual alternative hypothesis is stated, you may deduce its statistical alternative, but this does not necessarily imply the truth or confirmation of the conceptual hypothesis. In sum, if the null is rejected, the statistical alternative is sure to be inferred — this much is obvious.\(^1\) What is not obvious is the inference of the conceptual alternative hypothesis. Consider the following example illustrating this point.

In a coin-flip paradigm, a null hypothesis could be that the coin is fair. After many trials, if indeed the null is rejected, what shall we infer? Well, assuming we

---

\(^1\) This is only obvious according to the Neyman-Pearson model of hypothesis-testing. This would not apply to Fisher's model since he rejected the conceptual reality of an alternative hypothesis.
have an alternative conceptual hypothesis, this necessarily implies a statistical alternative hypothesis. However, although the statistical hypothesis may be easily inferred (upon rejection of the null), this does not mean the conceptual alternative can be inferred with equal ease. If $p < .05$, I will reject the null and infer the statistical alternative. This follows from the Neyman-Pearson model of hypothesis-testing. However, I may not be so willing to infer the conceptual alternative, especially if it is something that is not a plausible explanation for why the null was rejected. For instance, I would hardly infer an alternative such as the coin is governed by spiritual agents (and thus the consecutive heads) that are invisible in this room. This conceptual alternative is unlikely² in that it probably would not be suggested by a social scientist. The scientist would more likely infer something of the nature, "this coin is biased, due to a jagged edge." This would be a more "common-sense" alternative to chance factors producing the successive heads.

However, it should be noted that the alternative could comprise almost anything and is not restricted to a particular number of hypotheses. Indeed, the logical possibility of alternative hypotheses is practically infinite.

To summarize, the statistical alternative can be regarded as a "numerical" alternative (e.g., see McClure & Suen, 1994; Harcum, 1990) to the null hypothesis, while the conceptual alternative may be regarded as an "explanation" or "theory" as

---

² By "unlikely," this is in no way ascertained by statistical or scientific means, but rather is a judgment made by the researcher according to his paradigm. See my discussion on the previous page of how the conceptual alternative is born out of controlling variables rather than by statistical or scientific means.
to the reason why the null hypothesis has been rejected. Bolles (1962) offers an excellent explanation of this distinction:

The statistician is confronted with just two hypotheses [i.e., the null and the statistical alternative], and the decision which he makes is only between these two. Suppose he has two samples and is concerned with whether the two means differ. The observed difference can be attributed either to random variation (the null hypothesis) or to the alternative hypothesis that the samples have been drawn from two populations with different means. Ordinarily these two alternatives exhaust the statistician's universe. The scientist, on the other hand, being ultimately concerned with the nature of natural phenomena, has only started his work when he rejects the null hypothesis. (p. 639)

Said somewhat differently, but still grasping the concept, is Cowgill (1977), "We are not interested in evidence about $H_1$ because $H_1$ itself explains anything, but because we want to know if there is enough evidence in support of $H_1$ to make it worth trying to explain why $H_1$ seems to be true" (p. 353).

Evaluation

Without a doubt, the fact that NHST cannot directly test the alternative hypothesis constitutes the model's most serious difficulty, and should be considered in light of a suitable alternative, reason enough to abandon the model. Science is interested in confirming hypotheses (or gathering confirmatory strength), not merely rejecting the plausibility of chance. Bayesianism presents a competing alternative,
providing a procedure that directly tests the research hypothesis. Bayesianism will be examined later in chapter 4.

Despite the fact that NHST cannot tell us a thing about the conceptual alternative hypothesis, many researchers continue to think it does. As Gigerenzer (1993) has called it, the "Bayesian Id's wishful thinking" amongst researchers is pervasive in that many researchers misconstrue the p value to tell us something about the alternative hypothesis. This difficulty however, results from the misunderstanding of researchers and cannot in itself be held as an inherent fault of the NHST model.

4. Construing the Alternative Hypothesis

Another problem closely allied to the one just discussed, is that of choosing an appropriate alternative hypothesis (i.e., conceptual or scientific) that will be inferred should the null hypothesis be rejected. There is no formal logic (and certainly no statistical logic) as to how to select the alternative hypothesis. As already mentioned, the selection of the alternative arises out of a given research question and by ensuring a certain degree of experimental control. However, as noted, rarely is any experiment controlled enough to ascertain the one-and-only true alternative. Logically, an experiment could never be controlled as such simply because there are infinitely many possible alternatives. Indeed, how and why scientists narrow the range of seriously entertained options is a matter of ongoing dispute among philosophers of science. As noted by Rozeboom (1960), and more recently Chow (1996), a simple rejection of the null hypothesis does not necessarily indicate which
of the potential alternatives is true, or should be inferred. Again, NHST simply does
not provide the researcher with the knowledge he/she most desires, that of knowing
the truth or probability of the alternative hypothesis. Since NHST cannot test the
single alternative most often posited, it can't of course test a multitude of alternatives.

Referring again to the "malaria" example will help clarify the problem of many
alternatives. Suppose that it is found that mosquitoes are responsible for infecting
individuals with malaria. This then is the alternative that has been inferred upon the
rejection of the null hypothesis. Is this the correct alternative? How do we know?
Well, for one, we can't test it, so we are left unsure. That is, all we can do is infer the
alternative given a rejection of the null. Moreover, unlike the null hypothesis, we
cannot even assign the alternative a probability of being true. Secondly, another
hypothesis may better account for the data, such as only those mosquitoes of a
certain species, or only mosquitoes that were born from an infected mother, carry
malaria. Thus, the problem of specifying an alternative is a big one that is not
resolved by NHST. Nor can this be expected from NHST, given its logical set-up. A
single rejection of the null hypothesis does not lend evidence for the conceptual
alternative, it merely rejects chance as a possibility, or rather "plausibility".
Following from this, rejecting the null does not "point" to the correct alternative.
There could be any one of a number of causes or reasons why the effect is present.
Once more, NHST fails miserably when one tries to accumulate even minimal

---

3 I am here suggesting the possibility of a single infected "mother mosquito" giving birth to infected
offspring. This would imply that the "mosquito proper" is generally harmless and does not carry
malaria.
knowledge regarding the research hypothesis. True that NHST tells us that we are better off with an alternative explanation to the null (i.e., statistical alternative), it does not inform us in any way of which alternative best accounts for the observed data. In short, NHST does not test the conceptual alternative hypothesis. Indeed, it is this difficulty that may prove the most detrimental to NHST's future in the behavioral sciences.

5. The "Effects/Non-Effects Dichotomy"

A fifth alleged difficulty with NHST (noted by Loftus, 1991, 1996), is that of creating a false dichotomy between what is and what is not considered to be an effect. Despite the fact that a significant $p$ value tells us little with regards to the size of effect, many researchers continue to distinguish a result reaching $p < .05$ as representing a significant effect. As already noted however, any result can be significant if sample size is increased substantially. Statistical significance is not indicative of effect size. For instance, a given experiment may yield a statistically significant result at $p < .05$, yet may have a very small effect size. In short, the $p$ value is minimally informative when ascertaining any estimation regarding effect size. However, the illusion persists among investigators. Loftus attributes this illusionary quality of significance tests to researchers' need for simplicity in their decision-making:

... it is a problem that arises because investigators, like all humans, are averse to making decisions that are both complicated and weak, such as "we fail to conclude that the null hypothesis is false." Rather, people prefer simple, strong
decisions, such as "the null hypothesis is true." This fact of human nature fosters an artificial dichotomy that revolves around the arbitrary nature of the .05 $\alpha$ level . . . . Accordingly, the world of perceived psychological reality tends to become divided into "real effects" ($p < \alpha = .05$) and "non-effects" ($p > .05$). Statistical conclusions about such real effects and non-effects made in Results sections then somehow are sanctified and transmuted into conclusions that endure into Discussion sections and beyond, where they insidiously settle in and become part of our discipline's general knowledge structure. (Loftus, 1996, pp. 163-164)

Overwhelming evidence for Loftus' "effects/non-effects" critique is evident by merely reviewing a random sample of journal articles. Relatively few experiments boast a finding with a $p$ value over .05. This may not be direct evidence for researchers and editors dichotomizing significant vs. nonsignificant results as "effects/non-effects," but the leap in suggesting this is a small one. If a significance level of over .05 is still possibly meaningful, then why are these results not published? The answer is that most researchers and editors take the .05 level to be the "cut-off" between that which is meaningful and that which is not. Somewhat amusingly, Cohen (1990) ridicules this historical over-reliance: "In governing decisions about the status of null hypotheses, it came to determine decisions about the acceptance of doctoral dissertations and the granting of research funding, and about publication, promotion, and whether to have a baby just now" (p. 1307).
Empirical findings support the existence of the "cliff-effect" (Rosenthal & Gaito, 1963). Indeed, research suggests that nonsignificant results are more than three times less likely to be accepted by journal editors compared to statistically significant results (Atkinson, Furlong, & Wampold, 1982;^{4} Sterling, 1959). This over-reliance on significance values as decision criteria, as argued by Hubbard (1995), may encourage researchers to "make" significant results: "This mystique encourages data mining and other dubious practices that have the effect of calling into question the integrity of published empirical findings" (p. 1098).

**Evaluation**

The "effects/non-effects" distinction is largely a product of misunderstanding. Journal editors and researchers alike, should accustom themselves to a new understanding that the p value is but only one indicator of a statistical result. At the same time, the "effects/non-effects" problem constitutes a model issue as well. The model requires one to chose a level of significance, and this must count as a "strike" against the model. Moreover, this single value is insufficient in differentiating "effects" from "non-effects." I conclude this problem to be largely one of the model and of misuse and misunderstanding.

**6. Arbitrary Alpha**

A sixth difficulty with NHST is that the p value is chosen arbitrarily by the investigator. As Chow (1996) notes, "statistical significance is not a stable property

---

^{4} This research was based on the "publication probability" of three versions of the same article, differing only with respect to attained significance level. Only two journals were used in the study: Counseling Psychology, and Journal of Consulting and Clinical Psychology.
or characteristic of the data" (p. 5). A result significant at $p < .05$ may not be significant at $p < .01$. The issue is whether such "arbitrariness" should be permitted in conducting "scientific" investigations. Rozeboom (1960) has also criticized NHST based on this arbitrary selection of the $p$ value.

In evaluating the merit of this criticism, one must consider the often overlooked subjective nature of science in general. We may be hard-pressed to find any model that does not, at some point, require some "cut-off" if we are to decipher some results over others. Even if we were to use a model that measures the nearness to a specific prediction, at some point, we would have to decide what counts as "near enough" to be considered meaningful. Although selection of the alpha level may be considered "arbitrary," it has been argued that it may make much sense to choose such a cut-off value. The .05 level may simply represent what most people would say is "more than chance." In fact, preliminary evidence for this has been gathered by Cowles & Davis (1982) who found the .05 level to be subjectively reasonable. As they explain:

It might be concluded from these data that, on average, people do have doubts about the operation of chance when the odds reach about 9 to 1, and are pretty well convinced when the odds are 99 to 1. Now, the mean equivalent probability of these two values is .055, or odds which are close to 19 to 1. Nineteen to 1 or 1 in 20 is the conventional level for rejection of the null hypothesis. . . . the 5% level would appear to have the appealing merit of having some grounding in common sense. (p. 251)
Consider the following example as a similar detailing of the experiment conducted by Cowles & Davis (1982), demonstrating why the .05 level may be a reasonable choice for alpha.

Suppose I present a coin to a subject and ask him/her whether it is a fair coin or a biased coin. The individual asks me to flip the coin (i.e., provide evidence) so he can make an informed decision. I flip the coin, it turns up heads. The individual states that there is no reason for him to disbelieve that it is a fair coin. I flip the coin again, and again it turns up heads. Again, the individual does not yet have enough evidence to declare it to be unfair. I continue the process, until I'm up to the eighth flip, and again, the eighth and all subsequent flips have turned up heads. The individual scratches his head, but still refuses to reject the "fairness" of the coin. I flip the coin a ninth time, and upon the ninth outcome of "heads," the individual exclaims, "That's it, it can't be a fair coin, that coin is definitely not fair!"

The above example is of course a parallel to what actually transpires in null hypothesis testing. The null hypothesis above is that the coin is fair. In other words, the individual will assume chance to govern results unless evidence encourages him to reject this possibility. The individual believes the coin is fair right up until the ninth trial, at which point because of the successive heads on the past eight trials, and

---

5 I use the word "encourages" here purposefully in hopes of highlighting the subjective nature of rejecting the null hypothesis. I argue that "evidence" does not reject the null, but rather the influence of the evidence on the interpreter is what rejects the null. This clarifies why some may choose to reject the null at .05, while some may choose to reject the null at .06, or .07 etc. The point to remember is that rejection of the null is a function of subjective certainty that the results obtained are not due to chance, rather than some mere "objective quality."
an additional one on the ninth, the individual rejects the possibility that chance is responsible for these outcomes. Rather, something else must be producing these events, the most common explanation being that the coin is not fair (i.e., alternative hypothesis).\(^6\) By this example, it is somewhat reasonable to suggest that the subject would reject the "chance hypothesis" if there is less than a 5% probability that the results are due to chance. Should results such as found by Cowles & Davis be replicated and further research support this idea, then we may consider this as a basis for choosing .05 as an appropriate alpha level. However, in making this choice, we must guard against committing the error that if people generally act as though they're using the .05 level, then we are justified in setting it as a standard. This is commonly referred to as the naturalistic fallacy. And for those who still have a problem with science operating based on "choices," there are numerous instances where investigators choose the facets of their experiments; what to investigate, sample selection, alternative hypotheses to be inferred, just to name a few. Objectivity in statistical analysis may indeed be an illusion (Berger & Berry, 1988), in the way that choices must always be made with regards to the tools we use. Using statistics does not imply a lack of researcher-based decision-making. As Thompson (1993) has recently asserted, "Like it or not, empirical science is inescapably a subjective business" (p. 365).

\(^6\) As noted previously however, the alternative is "inexact," and therefore any explanation can be considered a "reason" for 9 successive heads. The most common of these explanations is that the coin is not a fair one, but others could be posited. Logically, the alternative hypothesis stating that super-powers are influencing the tosses, while not very intuitive to most, cannot be ruled out based solely on the NHST procedure alone, since the alternative is never tested directly.
Indeed, Fisher (1959) alluded to a similar "reasonability" of the significance level, stating, "the feeling induced by a test of significance has an objective basis in that the probability statement on which it is based is a fact communicable to and verifiable by other rational minds" (p. 214). This quote suggests then, that Fisher too saw the significance level as a probability level that signifies when a "rational" mind would reject the hypothesis that results obtained are due to mere chance.

Evaluation

The criticism of the arbitrary selection of a p value is not on its own, enough to condemn the model. Far from it, it is yet merely another circumstance where researchers (and indeed the scientific community as a whole) are obligated in choosing how to make their science as "objective" as possible. Although the p value is arbitrary, one would likely run into the same problem employing any other procedure. At some point, at some level, the individual or scientific community as a whole, must "choose" what counts as "significant." Indeed, even Fisher⁷ (1935) did not deny each experimenter the opportunity of choosing his/her appropriate level of significance for his/her research: "It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result" (p. 13).

---

⁷ I say "even Fisher" because he adamantly rejected the opinion-based prior probability of Bayesianism, yet did not appear to have difficulty in making the recommendation "open to the experimenter" with regards to when the null should and should not be rejected.
7. Statistically Insignificant Results May be "Significant"

A seventh, but less frequently documented problem with NHST, is that even if a result is not significant at $p < .05$ or $p < .01$, it may still be socially or practically significant. And contrary to this, as noted by Tyler (1931), and more recently Thompson (1999), the mere statistical significance of a result does not by itself imply importance or relevance. I will focus here on the former possibility, simply because it is of much greater consequence to disregard results when they may be of some use than it is to not evaluate the social significance of a significant result. Indeed, one would hope that researchers are conscientious enough to realize a socially relevant result from one that is not, regardless of the $p$ value. The idea of "throwing away" statistically insignificant (i.e., $p > .05$) results that may be useful is indeed a "scary" possibility.

Let us turn then to the greater problem, that of dismissing statistically insignificant results that may be worthy of attention. Hays (1963) years ago, clarified the lack of a necessary relation between what is statistically significant, and what is strongly associated: "Evidence of a strong statistical association can occur in data even when the results are not significant" (p. 342). Rosnow and Rosenthal (1989) cite an example in which a result, while not statistically significant, was nonetheless "practically" significant. They found a recent study investigating the effectiveness of aspirin as reducing the likelihood of heart attacks to be successful in finding an effect

---

8 Tyler (1931) also discusses an example pertaining to socially relevant results that were not found statistically significant. Thus, Tyler argues for both misinterpretations of "statistical significance" vs. "social significance."
for aspirin. However, this much was not obvious by merely interpreting the p value. They explain:

In the aspirin condition 4.8% had a fatal heart attack, whereas in the placebo condition 9.5% had a fatal heart attack. It appears that mortality from myocardial infarction decreased by approximately one half as a result of aspirin taken every other day. When we compute the effect size, we find it to be more than twice the size ($r = .08$) of that computed for the overall results. However, even though the effect size for the smaller subset is more impressive than the effect size for the entire sample, were we to rely on dichotomous significance testing for a yes-or-no-decision we would be led to not reject the null hypothesis. (pp. 1279-1280)

In this example, it is clear that interpreting a p value was insufficient in judging results. But why? The authors go on to explain that because the sample size used was so small, power was minimal, and therefore a significant value could not be achieved. Their conclusion was to continue the investigation with a larger sample size, and only then, could informed conclusions be made.

Again, we have a situation that could have possibly been resolved if power analysis were performed to determine a sufficient sample size. Effect size measures are not sensitive to sample size, and therefore they may suggest the presence of a "finding" even if sample size is small (i.e., low power). The optimal solution would be to ensure reliable power before the analysis is run. Then, the problem of
disregarding nonsignificant results would likely disappear. In quoting Cowgill (1977), I foreshadow my discussion of power in the following chapter, as well as the importance I place on its utility: "Especially common is the failure even to consider the question of 'power,' and without this reports of 'significance' are almost meaningless" (p. 352).

**Evaluation**

In evaluating this criticism, I come to the conclusion that it is valid, if power analysis is to be disregarded by researchers. If sample size is not given due attention, then the interpretability of \( p \) values is decreased, and that might lead to the dismissing of potentially informative results as non-meaningful. However, as will be expanded on in the following chapter, power analysis appears to be a cure for problems of ensuring adequate sample size, and thus enhancing the interpretability of the \( p \) value.

8. Violation of Assumptions

An eighth criticism of NHST is that many times it is performed without first ensuring that the assumptions crucial to the procedure are satisfied. Clark (1963) notes some difficulties that arise when this is done. More recently, Loftus (1996) has also argued this point, claiming "off-the-shelf assumptions produce off-the-shelf conclusions" (p. 164). Loftus takes this critique one step further and says that these assumptions, satisfied or not, dictate psychological theory. Referring to this problem as "The hypothesis-testing tail wags the theory-construction dog," he continues:
... the nature of the data analysis technique generally dictates the nature of psychological theory, which, in turn, engenders strong biases against formulating theories incorporating other perhaps more realistic or interesting assumptions. (p. 164)

Cochran (1947) analyzed the potential difficulties that result when the assumptions underlying ANOVA are not satisfied. Overall, he found it to be very difficult for any investigator to satisfy all the assumptions:

Failure of any assumption will impair to some extent the standard properties on which the widespread utility of the technique depends. Since an experimenter could rarely, if ever, convince himself that all the assumptions were exactly satisfied in his data, the technique must be regarded as approximative rather than exact. ... The factors that are liable to cause the most severe disturbances are extreme skewness [i.e., non-Gaussian distribution], the presence of gross errors, anomalous behavior of certain treatments or parts of the experiment, marked departures from the additive relationship, and changes in the error variance, either related to the mean or to certain treatments or parts of the experiment. (p. 37)

Therefore, according to Cochran, violation of assumptions can indeed impair the attainment of valid results. He does believe however, that such gross violations are detectable before the investigation is undertaken. This in turn permits the investigator to properly select the statistical test most appropriate for his/her data.
Others find ANOVA fairly robust to assumption violations (e.g., Boneau, 1960; Gaito, 1959). For instance, Feir-Walsh and Toothaker (1974) compared the ANOVA F-test to a non-parametric test, the Kruskal-Wallis. They found empirical evidence to support the use of the F-test over the Kruskal-Wallis even if assumptions were not satisfied. As they state, "When normality and/or homogeneity of variance is doubtful, the ANOVA F-test is the recommended procedure for testing hypotheses about means" (p. 797). They note however, that when testing medians, the Kruskal-Wallis test is still recommended over the F-test.

Although ANOVA may be robust to most assumption violations, it appears not to be to unequal Ns (Rogan & Keselman, 1977). That is, as noted by Alexander and Govern (1994), when sample sizes in an ANOVA are not equal, one should not assume homogeneity of variance: "It is now well established that the conventional ANOVA for the comparison of means is not robust to heterogeneity of variance when sample sizes are unequal" (p. 91). However, these authors offer a means of adjusting for this violation (see also Gaito, 1972).

Evaluation

I don't see the "violation of assumptions" problem to be severe. As Cochran claims, the conscientious researcher should be able to anticipate and investigate problems of normality etc., and deal with them effectively. Moreover, as noted, many find ANOVA robust to violations, and for those violations of which it is not, adjustments are offered to rectify the problem. While NHST may not prove as applicable to as wide a range of investigations as one would like, this is no reason to
refute it as a model of inference. Unless there exists a model that deals more effectively with these violations, we have no reason to reject NHST on this merit alone.

Is a Correct Comprehension of NHST Impossible?

Evident from the preceding discussions and evaluations, many criticisms stem from misunderstandings and misconceptions regarding the nature and applicability of the NHST model. I will now proceed to address the question of whether these misconceptions and misunderstandings warrant a problem in their own right against NHST.

In line with my evaluation of the many criticisms discussed in this chapter, I conclude that a great deal of alleged problems of NHST result from "human error" — that is, from a misunderstanding and misuse of the model. Do these misunderstandings warrant a "fault" against NHST? Is NHST that stubborn that it defies human understanding? Certainly not. I argue that most, if not all misconceptions and misuses of NHST result from poor training and poor attention with respect to its major issues. The tool isn't simple, but it isn't overwhelmingly complex either. Should we continue to use the model, what is required is a re-training with respect to the procedures, both technically and theoretically, so that the wealth of misconceptions and misuses can be rid of once and for all. I dare suggest that we would dismiss the utility of quantum physics based on it being difficult to grasp. Nor should we dismiss NHST solely based on its need for greater intellectual attention.
Conclusion

In appraising the validity of the above mentioned criticisms and problems of NHST, I am faced with a major difficulty -- that of evaluating in "absolute" terms. Having not yet explored alternative models of hypothesis testing, it becomes next to impossible to pass judgment on whether the criticisms directed at NHST merit abandonment of the model. Although many of them are misdirected, and result from a general misunderstanding/misuse of the model, the few that are legitimate are serious enough to consider abandoning NHST altogether. As argued, the fact that NHST cannot yield new information pertaining to the research hypothesis must be considered a prime concern. This again, I repeat, is the most troubling aspect of NHST. However, it is only troubling so much as other model(s) are able to rectify this, and other problems. In order to properly appraise the NHST model as a whole, I must be able to relate it to alternatives, and answer the question of "what would we do instead?" NHST may be problematic, but it may be the only or "best" way. The only way to truly give NHST a fair trial is to thoroughly review, and where appropriate, recommend alternative means of hypothesis testing. It is to that task that I now turn, in chapter 3.
Chapter 3

Proposed Alternatives to NKST

Judging from the discussion in chapter 2, it is fair to conclude that problems abound with regards to NHST. Indeed, years of criticism have made clear the fact that NHST is very problematic. Even the most loyal advocate would surely admit that the model is questionable as a means of statistical inference. The problems with NHST are so severe that some have argued for it to be totally banned from scholarly journals (Shrout, 1997). Given this reputation, the question then becomes that of determining a suitable alternative to NHST as currently used. Although it is fine to point out the difficulties with today's hypothesis testing, it is also necessary to contemplate alternatives that could either complement or replace today's model. I write this chapter with the view that a model of hypothesis testing is favorable if it constitutes the "best" of available alternatives, even if exhibiting problems of its own. Indeed, even the best model may require amendments. However, until a superior model is agreed upon, the "best alternative" shall be considered ideal and will be promoted for use in psychological research. As with any theory, a theory of statistical inference is favorable until another proves itself more desirable.

The goal of this chapter is to address the following question -- given that NHST is problematic, what other procedures or models can be used in its place? Are there any feasible alternatives to null hypothesis significance testing? Many alternatives to
NHST have been proposed and will be considered below. Each will be evaluated for its ability to either replace or complement today's NHST procedure.

I. Effect Size

Perhaps the most commonly recommended alternative to solely interpreting the p value is to determine the magnitude of effect, more commonly known as "effect size." The employment of effect size measures has been recommended repeatedly throughout the psychological literature (e.g., see Dunnette, 1966; Vaughan & Corballis, 1969; Huberty, 1987; Rosenthal, 1992). Magnitude of effect measures are indicators of the strength of association between the independent variable(s) and the dependent variable(s). As noted by Huberty (1993), magnitude of effect measures have existed in statistics textbooks since the early 1930s. It wasn't until the 1970s however, that the name "effect size" was popularized. Since Hays (1963) discussed the use of omega squared ($\omega^2$), many articles have been published on the topic (Haase, Waechter, & Solomon, 1982). However, according to Huberty, few textbook authors presently discuss effect size measures.¹ This is despite the fact that evidence suggests that effect size reporting, especially to a younger audience, tends to increase researchers' confidence in results (Nelson, Rosenthal, & Rosnow, 1986). Although the tide is currently shifting and more researchers are beginning to report it in results sections (Cohen, 1994), effect size has traditionally been ignored by most investigators. For instance, in a survey of articles published in British Journal of Psychology and British Journal of Social and Clinical Psychology between the years
1969 and 1972, none reported or discussed strength of relationship among the variables under study (Cochrane & Duffy, 1974). Furthermore, the survey found $\eta^2$ (once it was calculated) to be very weak in 56% of the studies, and approximately one quarter of the studies to have relationships accounting for less than 10% of shared variance. Today, investigators appear to be aware that effect size measures reveal important information, despite the fact that still few are reported in the psychological literature. Prior to considering effect size measures as alternatives to hypothesis testing, a brief review of their nature is in order.

What is Effect Size?

An effect size measure is an indicator of the association that exists between two or more variables. An exception to this is Cohen's $d$, which is a measure of the distance between means. These definitions translate into how much variance in one variable is accounted for by knowledge of another variable. As noted by some (e.g., Cohen, 1968; Kerlinger & Pedhazur, 1973), the increase in interest in effect size measures among psychologists is associated with their increased awareness of the similarity between ANOVA and regression (Haase, Waechter, & Solomon, 1982). Hence, realizing that assessing group differences is but only one way of employing basic correlational techniques may have sparked the interest of psychologists in

---

1 An exception to this is Howell (1989) who provides a good explanation of effect size.

2 Three other journals were also surveyed and showed more promising reports of magnitude of effect measures. Journal of Applied Psychology had an effect size reporting rate of 77%, Journal of Educational Psychology had an effect size reporting rate of 55% and Journal of Personality and Social Psychology had an effect size reporting rate of 47%.
effect size measures. In a word, effect size is the difference between two population means divided by the standard deviation of either population (Howell, 1989).

Determining what constitutes a "big" effect is troubling for work in the social sciences. As noted by Haase et al. (1982), answering the question of "how big is big?" is not as difficult in the natural sciences as it is for psychological science:

Effect sizes and related values of strength of association found in engineering, for example, are likely to be much larger than those found in the typical experiment in the behavioral sciences. Thus all investigators are left to develop their own intuitive standards as to what constitutes a small, medium, or large effect. (p. 59)

So, what constitutes a "big" effect in the average psychology experiment? Cohen (1977), without a doubt one of the strongest advocates of reporting effect size statistics, has issued guidelines as to what constitutes small, medium, and large effect sizes. According to Cohen, for $d$, .20, .50, and .80 constitute small, medium, and large effects respectively.

Effect Size vs. $p$ Value

Critics of significance testing argue that magnitude of effect measures provide more reliable and informative information than do $p$ values. Moreover, they note that associating strength of effect with a $p$ value is wholly misleading. As Oakes (1986) states, "the fact that two experiments may yield sharply varying test statistics (i.e., $p$ values) is no guarantee that the underlying strengths-of-effect are likewise variant" (p. 50). Hence, it is important not to confuse the significance of a test with a healthy
magnitude of effect. Contrary to the p value, effect size measures offer the advantage of not being sensitive to sample size. Therefore, unlike p values, effect size indicators yield a stable numerical value that will not increase or decrease based solely on the total number of subjects used in the research. That is not to say that an increase in sample size may not facilitate finding new effects (and thereby increasing effect size), but only that effect size measures are not "functions" of sample size. A substantial increase in sample size does not necessarily guarantee an increased effect. However, as was noted in chapter 2, an increase in sample size is almost always associated with a smaller p value.

In illustrating the above, consider an example offered by Chow (1996) in which two means are compared. In study A, the mean of the experimental group is 6 while the mean of the control group is 5. Degrees of freedom are equal to 22. Given a t-test, statistical significance is obtained, with an effect size of .1. In study B, the mean of the experimental group is 25 while the mean of the control group is 24. Degrees of freedom in this case are equal to 8. Again given a t-test, but contrary to study A's results, study B did not obtain a significant p value, but had the same magnitude of effect as that found in study A (i.e., effect size = .1). The point of this example is that similar studies although showing an identical degree of shared variance among variables, are nonetheless different in terms of statistical significance. Seeing that statistical significance and effect size measures lead to different "conclusions," the

\[3\] I use the word "conclusions" hesitantly. My goal is not to imply that conclusions should be based on either the p value or effect size. Rather, as discussed in the text, if NHST is to be used, both indicators
question then becomes one of evaluating the importance and relevance of each indicator. Which is the more important to interpret, the p value or effect size?

Perhaps the above question is a bit too narrowing. After all, why should we not be able to interpret both? For the purpose of this discussion however, I will try to answer my original question, and I will assume for a moment that we had to choose one or the other. This in mind, I ask again — which is more important for our purposes, the p value or effect size? Although both indicators might (and should) be used in evaluating the merit of a study, which of the two is the more "informative"? Otherwise said, which is more relevant for the goals of our research? Cohen (1990) claims to have learnt that the primary product of research is that of effect size measures, and not p values. As already mentioned, in solely interpreting significance levels, we are left with an inadequate account of the degree of association that exists between variables. Similarly however, as noted by Chow (1996), by interpreting only effect size measures, we are left with an inadequate account of the probability of the data. Whereas a healthy effect size measure may appear illuminating, it may lose some of its shine when a glance at the p value informs us that the probability of the results being due to chance is still quite high. Hence, a large effect may exist in the sample, but at the same time, results may stand a substantial probability of being due to chance. Can healthy effects be found in the proximity of the null hypothesis? Presumably yes. As noted earlier, there is little or no relation between the size of the p value to that of effect magnitude. Thus, the compatibility of the two statistics may

should be employed in data analysis and interpretation.
indeed be questioned and even the most liberal researcher (or editor) may find himself with conflicting statistics. He may therefore have to chose the one he prefers based on his own theoretical orientation when drawing conclusions from data.

**Criticisms of Effect Size**

Although criticisms directed toward effect size measures exist, they are not severe enough to seriously question the utility of these measures. For instance, Favreau (1997) criticizes effect size based on the idea that they do not represent the size of difference that distinguishes each member of one group from those of another. She points out that given the effect size, we are still left unknowing whether all individuals differ moderately or substantially from those in a comparison group. Levin (1967) also notes the problem in interpreting the variation shared among variables. As an example, he notes an experiment done in which the shared variance was equal to a substantial 37%, but that by closer inspection, one would find that over 85% of the total shared variance resulted from one of the six experimental groups. Thus, if you have one superior group, this "success" may blanket over to the other groups and give the illusion of a large effect.

Although the above critique is worthwhile, the difficulties they highlight must be considered secondary for the purposes of our discussion. Although it is true that we cannot see where the differences lie simply by interpreting effect size, effect size doesn't purport to answer this question anyway. This is not to disregard the importance of knowing where these differences lie. Graphical techniques (e.g., linear functions as described by Loftus, 1996) would be ideal in visually interpreting where
differences lie and discerning patterns of means. The use of error bars could also emphasize the degree to which differences between individual means are substantial.

Favreau (1997) also argues that a limitation of effect size is its dependence on how the dependant variable is operationalized. She gives the example that in a comparison of helping behavior, the difference may depend on the type of help required. More recently, Prentice and Miller (1992) have reviewed studies in which small effects were regarded as very impressive due largely to the operationalization of the variables studied. The way a variable is operationalized can have much influence on how effect size is interpreted, and hence one can say that an effect size estimate is at least partly a function of how variables are defined. However, it is difficult to pin this problem solely on effect size. It is obvious that how a variable is operationalized has much to do with what differences are found, but we can't begin to say that this is a limitation of effect size. This is more of a methodological issue, and not a statistical one. This critique brings to light a limitation of effect size, rather than a problem with the statistic.

Another critique of effect size is provided by Dooling and Danks (1975). These researchers argue that psychology, because of the nature of its experimental designs, is simply not ready to begin adequately interpreting effect size statistics. The argument centers around how levels of the independent variable are chosen. They argue that this choice is arbitrary and that various levels may lead to various effect size estimations. For instance, a researcher wanting to increase the likelihood of obtaining a significant result may choose levels that are at opposite ends of values of
the independent variable. Another researcher, however, performing the same study, may choose levels that are not so diametrically opposed to one another. Dooling and Danks argue that values of estimated $\omega^2$ depend on the way in which these levels are determined. Thus, while one experiment may reveal a particular effect size magnitude, another one, identical except for chosen levels of the independent variable, may reveal different effect size values. Furthermore, as noted by Dooling and Danks, most psychologists will interpret these results as resulting from a random effects model, rather than from a fixed effects model, and in doing so, will generalize effects across all levels of the independent variable rather than restrict them to levels actually tested. This possibility greatly reduces the interpretability of effect size measures in psychological research:

For such purposes of statistical inference, $\omega^2$ is either useless, misleading, or both because the strength of an experimental effect will depend directly on ill-defined decisions made by the E [i.e., experimenter]. The E who chooses extreme values of his independent variable is likely to obtain strong effects (assuming, of course, that there is some relationship between independent and dependent variables). Another E might choose closely related values of the same independent variable that would in turn yield a smaller $\omega^2$. Hence, a correct inference based on the $\omega^2$ is impossible because the underlying population of values on the independent variable has not been randomly sampled. Since most readers of research articles are likely to drawn unwarranted inferences from a
reported $\omega^2$, we suggest that its widespread use be discouraged. (Dooling & Danks, 1975, p. 16)

Dooling and Danks entertain the possibility of keeping $\omega^2$ and changing the designs typically employed in psychological research, but claim that this is likely to be an impossible assignment since "psychology is a science that still does not know what its major independent variables are" (p. 16). Thus, changing the designs of psychology may not be possible in order to better interpret $\omega^2$. As an improvement, they recommend the use of random effects models instead of fixed effects for levels of the independent variable. They continue to argue however, that for the most part, this cannot be accomplished:

We are simply not ready to force ill-defined variables into standard molds. Nor would there be any wide agreement on standard experimental procedures in the various areas of psychology. Such efforts at standardization would eliminate many potentially interesting variables that would now be considered "extraneous" and thereby limit the scope of psychological research. Clearly, there is no realistic hope that experiments can be redesigned in order to make $\omega^2$ a useful measure. (p. 16)

The criticisms offered by Dooling and Danks, although legitimate, are not enough to seriously challenge magnitude of effect measures. Choosing levels of the independent variable is more of a methodological concern and although it may have an influence on the statistics that are interpreted, this problem cannot be considered a
"drawback" of effect size indicators. It is rather up to the well-trained researcher to appropriately choose the levels of the independent variable and to report these decisions in methods sections of articles. Likewise, it is up to the knowledgeable interpreter of such literature to realize the limitations of the study when reading that a fixed-effects model was used. In such a case, the results would only be applicable to those levels of the independent variable that were actually tested in the experiment, and could not be generalized to all possible levels.

There are many measures that can be used in estimating effect size. As already discussed above, $\omega^2$ and $\eta^2$ are common means for estimating magnitude of effect. However, as noted by Howell (1989), $\eta^2$, although relatively simple to calculate, often yields an overestimate of the present effect. Cohen's $d$, the standardized mean difference, has also been criticized as a measure of effect size (see Richardson (1996), for a review). Other measures for estimating magnitude of effect, reviewed by Snyder and Lawson (1993), include Epsilon squared, Wherry, Herzberg, and Lord. Ray and Shadish (1996) caution however against interchanging effect size estimates. They found that, especially for meta-analytic studies, different estimators yield various estimates, and are not often equal to $d$. Methods for estimating effect size for various models (i.e., one-way ANOVA, two-way ANOVA etc.) are offered by Vaughan and Corballis (1969) and more recently Kirk (1996). Rosenthal and Rubin (1982) also offer a method for calculating effect magnitude to determine improvements in success rates. Kirk (1996) has found the reporting of effect size measures to be quite low. In a survey of articles published in 1995, Kirk found only
12% of articles in *Journal of Experimental Psychology, Learning & Memory* to use a magnitude of effect measure when reporting statistical significance.

**Effect Size: Complement or Replacement to Significance Testing?**

In evaluating effect size measures as an alternative to significance testing, I conclude that they are best used in conjunction with p values. They are especially useful for conducting meta-analyses of prior research when the goal is that of finding overall effects across many research studies (Schmidt, 1996). They are helpful in that instead of merely reporting whether a given result is due to chance, they indicate the strength of any relationship that may be present. Again, however, as noted by Chow (1996), effect size measures are limited in that they do not indicate the probability that the association between variables (however large) is due to chance factors. As Chow says:

Effect size advocates are prepared to apply the non-statistical criterion even when it is not possible to exclude chance influences as an explanation of the result. Why is statistics used at all? Should the researcher embark on a course of action when the research outcome may actually be the result of chance influences? (p. 117)

In addition to Chow's remark, it should be noted that effect sizes are only "descriptive" statistics, and not "inferential" ones. That is, they only reveal the magnitude of effect in the sample, and offer no information of how likely this estimate is in the population. One needs NHST to reveal the latter information. Clearly, the use of either effect size measures or significance values is not a
sufficient means of appraising data. It is difficult to choose one over the other in that crucial and useful information is lost when such a selection is made. As noted by Richardson (1996), "the important point is that measures of effect size are simply another part of the composite picture that a researcher builds when reporting data that indicate that one or more variables are helpful in understanding a particular behavior" (p. 21). Taking an even more extreme position, Snyder and Lawson (1993) conclude: "These aids [effect size measures] are merely tools to assist the researcher in gaining a more informed analysis of data. The ultimate responsibility for developing a comprehensive analysis of the meaning of results rests with the researcher" (p. 347). Thus, under the aegis of null hypothesis testing, I recommend using both significance testing and effect size when interpreting data.

II. Confidence Intervals

The use of confidence intervals is another means of making sense of our data, and consequently, are a candidate for either replacing or complementing NHST. Many have recommended the use of confidence intervals when interpreting experimental findings (Bakan, 1966; Chandler, 1957; Cohen, 1990; Grant, 1962; Hammond, 1996; Kirk, 1996; Rozeboom, 1960; Schmidt, 1996). A confidence interval provides a medium for assessing the probability that a parameter lies within a particular range of values. As Kirk (1996) notes, providing a confidence interval requires no more information (i.e., calculations) than conducting a null hypothesis significance test. Furthermore, a confidence interval tells us all and more than a null
hypothesis significance test. He strongly recommends the use of intervals in advancement of our science:

A rejection [of the null hypothesis] means that the researcher is pretty sure of the direction of the difference. Is this any way to develop psychological theory? I think not. How far would physics have progressed if their researchers had focused on discovering ordinal relationships? What we want to know is the size of the difference between A and B and the error associated with our estimate; knowing that A is greater than B is not enough. (p. 754)

Kirk offers an excellent example of how a mere significance test can be greatly misleading and that instead, a confidence interval can inform us of what we are really searching for in scientific investigations. He considers a researcher who wants to know whether a medication will improve intelligence test performance for Alzheimer patients. Twelve patients are assigned to the experimental condition while 12 are assigned to the control group. Using a t-test, the researcher tests whether there is a significant difference in IQ scores between the two groups. The test reports a non-significant value, t(10) = 1.61, p = .14. Although most researchers would conclude that a reliable difference does not exist and that the medication does not have a substantial effect, Kirk argues that reliance on the p value alone is very misleading, and that the medication does have a great effect. By calculating a 95% confidence interval, the researcher would have found the mean difference to be between 6.3 and 32.3 IQ points. He would have also found that an unbiased estimate of Cohen's d is .86, suggesting that the difference represents a large effect.
According to Kirk, it is clear the data provide good evidence for the scientific hypothesis, and hence the study should not be discarded merely because the .05 level was not reached. He argues that "the nonsignificant t test does not mean that there is no difference between the IQs; all it means is that the researcher cannot rule out chance or sampling variability as an explanation of the observed difference" (p. 755).

Kirk advises researchers to inspect the "practical" significance of results rather than continually limit data appraisal to mere p values. With Kirk I concur. Again, as has been noted many times before by many methodologists, data analysis that ends with the p value is inadequate. Additional tools are required, if not even good old fashioned experimenter judgment.

Cohen (1990) suggests building 80% confidence intervals around sample effect sizes. Such a measure would inform the researcher as to the probability that a given range includes the given effect size. For instance, an 80% confidence interval for the effect size of .3 would imply that we are 80% sure that the true effect size lies between say, .1 and .5. Confidence intervals represent a way of quantifying our degree of certainty with regards to the true population parameter.

Loftus (1991) has also been a strong advocate of using confidence intervals in reporting results. For him, the provision of intervals constitutes a useful and necessary means of accounting for error and provides the researcher with a much more meaningful answer to the question he is really probing:

The emphasis on hypothesis testing produces a concomitant de-emphasis of an alternative technique for coping with statistical error that is simple, direct, and
intuitive and that has wide acceptance in the natural sciences: the use of confidence intervals. Whereas hypothesis testing emphasizes a very narrow question ("Do the population means fail to conform to a specific pattern?"), the use of confidence intervals emphasizes a much broader question ("What are the population means?"). Knowing what the means are, of course, implies knowing whether they fail to conform to a specific pattern, although the reverse is not true. In this sense, use of confidence intervals subsumes the process of hypothesis testing. (p. 103)

More recently, Hunter (1997) has remarked in the high use of confidence intervals in the quantitative sciences (i.e., physics, chemistry, electronics), and that adoption of these measures in psychology would greatly reduce the error rates present with sheer significance testing:

The older quantitative sciences, such as physics, chemistry, and electronics do not now use and have never used the significance test as it is currently used in psychology. This fact strongly suggests that there are alternatives to the significance test that are used in the other sciences. The dominant technique used in the quantitative sciences and the dominant technique used by mathematical statisticians is the technique of confidence intervals. I have studied the use of confidence intervals in other fields, and I am convinced that these fields do not suffer the high error rate that we suffer in psychology. (p. 6)

In advocating the use of confidence intervals, Hunter highlights the fact that a "95% confidence interval always has only a 5% error rate; it is not context dependent
like the significance test....furthermore, the various confidence intervals are all compatible with each other; there is no need for a social convention such as 'α = 5%" (p. 6). Many sources exist with good explanations of how to construct and use confidence intervals (for example, see Hays, 1963).

Evaluation of Confidence Intervals

In evaluating confidence intervals, I find them to be an indispensable tool in evaluating research results. They allow the researcher to estimate the degree to which (i.e., probability) the observed value (or difference) is likely to be the true value. Unlike effect sizes, I find confidence intervals better able to replace significance testing, again if we had to choose one over the other. What makes confidence intervals so appealing is that they can provide both a hypothesis test, and produce an estimate of the observed parameter (Becker, 1991). This is true, despite the fact that the hypothesis test and confidence interval are still regarded as separate steps in the process of data appraisal (Serlin, 1993). As will be discussed later in this chapter, confidence intervals work extremely well with the "good-enough" model proposed by Serlin & Lapsley (1985). The use of confidence intervals is summarized quite well by Serlin (1993) in that "we use them, in the same way an astronomer would use a telescope, to determine if the theoretical prediction, fortified by a good-enough principle belt, obtains" (p. 352). What this essentially results in, is using a "range null hypothesis" rather than a "point-null hypothesis," the range represented by the "belt" that surrounds the obtained estimate. The good-enough principle will be described in more detail later in this chapter. However, for now it is enough to recognize that
from the preceding discussion, confidence intervals constitute an indispensable tool in analyzing psychological data. In conjunction with hypothesis-testing (which as mentioned, is really one process), they offer much and are highly recommended (by yet another author, and for the billionth time!) for use in the behavioral sciences. On methodological grounds, confidence intervals can easily replace significance tests. Trying to convince this to the research community is another story.

III. Interval Estimation as Significance Test

Expanding on the preceding discussion, an alternative to merely rejecting null hypotheses is to provide an interval estimation of a parameter, and treat it as a significance test. Loftus (1996) recommends this technique as a replacement to hypothesis-testing. But how exactly is this done? According to Loftus, the null hypothesis could be rejected if the confidence intervals suggest that there is some sort of effect in the data. He uses the example of a psychologist comparing his newly discovered treatment to a traditional treatment, hypothesizing that his treatment is just as effective as the traditional one. The psychologist tests this hypothesis by comparing scores from his treatment versus scores obtained on the traditional treatment. A simple t-test shows the difference to be non-significant, thus suggesting both treatments are equally effective. However, Loftus argues that this report can be grossly misleading if confidence intervals are not provided. As shown in Figure 1, high power is indicated by smaller error bars, while low power is indicated by larger error bars.
Loftus claims that the top representation gives us reason to reject the null hypothesis because of the large degree of power, while the lower representation represents little power and hence the null should not be rejected based on these data alone. The lower panel suggests that the means are more likely to vary than are the means in the top panel. One advantage of Loftus' approach is the ease in which results can be interpreted. By visually inspecting the two graphs, results are readily apparent, and tentative conclusions regarding the data can immediately be made.
IV. Plot-Plus-Error-Bar Procedure (PPE Procedure)

Another alternative to traditional NHST, advocated by Loftus (1993, 1996) is called the "plot-plus-error-bar" approach to data analysis. In general, Loftus (1996) finds present-day reporting of F ratios and p values cumbersome and confusing. He argues:

Most people find such tabular-cum-text data presentation difficult to assimilate.... Decades of cognitive research, plus millennia of common sense, teach us that the human mind is not designed to integrate information that is presented in this form. There is too much of it, and it cannot be processed in parallel. (p. 166)

Loftus is referring to the data in Table 1. Rather than presenting data in this "confusing" way, Loftus recommends the use of graphical techniques. He argues that such techniques convey the same information, but in a more clear, and concise fashion. Using his example, consider the data presented in Table 1, representing results obtained from an experiment investigating memory for visually presented material.

Loftus argues that such information is difficult to interpret and at most, causes overdue confusion. He suggests rather, to present such information as is depicted in Figure 2. By plotting confidence intervals around each group mean, this allows the interpreter to assess the degree of statistical power present in the investigation. Loftus argues that by visual inspection, the reader can obtain as much information from Figure 2 as he could from Table 1 -- the only difference being that the graphical
Table 3

Pe+rnmce
us
amion
of
expumre
duruthotr
firi
cmdith
"LI_
Egposure
duration
(in
mi
Cliseconds)

Table 1. Performance as a function of exposure duration for four conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>Exposure duration (in milliseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
</tr>
<tr>
<td>NVE,HU</td>
<td>0.287</td>
</tr>
<tr>
<td>VE,HU</td>
<td>0.461</td>
</tr>
<tr>
<td>NVE,LU</td>
<td>0.099</td>
</tr>
<tr>
<td>VE,LU</td>
<td>0.683</td>
</tr>
</tbody>
</table>

Note. NVE = no verbal encoding; VE = verbal encoding; HU = high spatial uncertainty; LU = low spatial uncertainty.

Table 1 -- Data in "tabular" form, -- contrast to graphical representation shown in Figure 2. (Reprinted from Loftus, 1996.)

representation of results is much easier to interpret. Taking an extreme position, he holds that "creative use of such procedures allows one to jettison NHST entirely" (Loftus, 1996).

Although Loftus' approach offers much in the way of reporting results with maximal clarity, his model has been criticized heavily as a way of replacing NHST. Specifically, Morrison & Weaver (1995) argue that although the PPE procedure is very useful in attaining an advanced understanding of a data set, it cannot and should not totally supplement standard hypothesis testing procedures. Firstly, they argue that in Loftus' example, it is unclear as to how he obtained the error bars, since his design is a completely within-subject design. They charge Loftus with "pooling" error estimates (i.e., error bars) across all means drawn in the linear function (see Figure 2). More specifically, they argue:
Figure 2 -- Data in graphical form (i.e., linear functions). (Reprinted from Loftus, 1996.)
There is no way that a single value can represent the standard error of the mean in this case. Inspection of Figures 1A and 1B, [Figure 2] however, reveals that the plotted error bars are all equal, which suggests the use of a common error term. The use of error bars in this case -- especially equal ones -- is both meaningless and misleading. (p. 54)

Loftus' (1995) reply to this criticism is that since the error terms are sufficiently similar, they can be pooled, and little information is lost in doing so. He argues that should the error terms be dissimilar, this would cast the opportunity to question why the error terms are different across levels of various factors. Therefore, contrary to Morrison and Weaver (1995), Loftus holds that using the same error term is acceptable, unless error terms are vastly different, in which case this difference could be studied in itself as a meaningful source of data.

Morrison and Weaver (1995) also charge Loftus with "selecting results to demonstrate his points" (p. 54). They argue that the linear functions plotted in Figure 2 may be interpreted differently by different researchers. They claim that it is entirely possible to believe that while the mean increases, so does the variances (i.e., the error bars become larger). By merely viewing the functions, one could question whether the variances of the means would eventually become so large as to overlap, and thus essentially making the linear functions "parallel" for larger means. In response to this criticism, Loftus (1995) admits the possible "ambiguity" of this problem, but claims that the critics' implication that an ANOVA would help clarify this ambiguity is false. As Loftus says:
It is the lot of many data sets reported in the social sciences to be similarly ambiguous. Morrison and Weaver seem to imply, however, than a simple ANOVA would clear up this ambiguity, since the ANOVA would cleanly result in a "reject" or a "don't reject" decision. (p. 58)

This last criticism by Morrison and Weaver and reply by Loftus highlights perhaps the greatest misuse of NHST -- the tendency to "decide" based on some statistical model. In the above, I would side with Loftus' reasoning that although results may be ambiguous, this does not call for some artificial means of making sense of the data. We may have to accept the ambiguity of our data if we are to represent them in an honest way. To impose some statistical decision where we "fool" ourselves into thinking we have clarity where there is none, is at best, an exercise in triviality. Too often, researchers deceive themselves into a "yes-no" answer when using NHST. Indeed, it is foolish to think of NHST as being able to provide this degree of information regarding a hypothesis. Rather, the function of NHST is to provide the experimenter with the probability of the data, given a true null hypothesis. It is then up to him/her to make a decision based on the data, and not the other way around.

Overall, Loftus' techniques represent a welcome shift in data appraisal. His PPE procedure is to be applauded for improving the clarity of analysis. However, it is doubtful that Loftus is completely "jettisoning" NHST by using his PPE procedure. After all, do error bars distant from one another not suggest rejecting the null hypothesis? It would appear that although presented in a much clearer fashion, the
concept of NHST may still be present in Loftus' model; only instead of $F$-ratios, Loftus uses linear functions.

V. Power Analysis

Jacob Cohen, the spokesman for power analysis, first recommended the procedure for psychology in 1962, and later provided an extensive handbook on the subject (Cohen, 1969). Although he wasn't the first to advocate power (Neyman & Pearson (1928) first originated the concept), he nonetheless can be considered a pioneer in its development. Before Cohen, Feldt and Mahmoud (1958) published power tables that depicted power graphically and could serve as a tool in determining desired sample size, along with setting $\alpha$.\(^4\) Cohen however, provided a means for calculating power that was relatively easy, mathematically non-threatening, and very comprehensive. In doing this, he addressed the concerns of both Neyman and Pearson, and also Fisher's "sensitivity" with regards to detecting an effect.

\(^4\) See Figure 3 for an example of a power table.
Figure 3 -- Example of a power table. The y-axis represents the number of observations required per treatment for a test of specified power. (Reprinted from Feldt & Mahmoud, 1958.)
Cohen's pioneering work originated from his own perception that power analysis was largely neglected in psychological research. In a survey of articles published in *Journal of Abnormal and Social Psychology* in the year 1960, Cohen (1962) found the mean power to detect medium effect sizes to be .48. From this, Cohen concluded that "the chance of obtaining a significant result was about that of tossing a head with a fair coin." Hoping to correct for this neglect in statistical power, Cohen would publish his most popular work, *Statistical power analysis for the behavioral sciences* (1969).

**What is power?**

The statistical power of a significance test is the long-term probability, given the population effect size, alpha, and sample size, of rejecting the null hypothesis. The influence of sample size on significance has long been understood. For instance, Hay (1953) noted that "sometimes the characteristics of a sample are such that no amount of treatment will bring about a satisfactory result" (p. 445). The present discussion is a very brief explanation of the logic of power analysis. For an extensive review, see Cohen's handbook. For a brief overview and very "user-friendly" publication, see his more recent article, "A Power Primer" (1992).

Chandler (1957) explains through the following example, that without the consideration of power, one chances having a very inflated \( \beta \) value:

Although texts in psychological statistics do not seem to place a great deal of emphasis upon the power of a test, power is the basic concept responsible for one's employing statistical tests as a basis for taking action on an \( H \). If this were
not so, to test an H at the 5% level of significance, one could simply draw from
a box of 100 beads -- 95 white and 5 red -- a bead at random and adopt the
convention that he would reject the H whenever a red bead appeared. With such
a test, one can readily see that not only $\alpha$ but also $1-\beta$ always equals .05, or $\beta = .95$. It is this large value of $\beta$ that precludes one's employing the bead-box test.

(p. 430)

Another technique, also related to sample size, has been recommended by some
authors (e.g., Snyder and Lawson, 1993). The technique is called the critical n, and
although similar to power analysis, it is much simpler to interpret. It is somewhat of
a "watered-down" version of Cohen's power analysis. In using the technique, the
researcher makes an estimate of effect size based on the critical value of the test
statistic. Results with low critical ns represent larger effect sizes than results with
high critical ns. Thus, a test achieving significance with a low number of subjects
would automatically suggest a prominent effect, compared to a test barely achieving
significance with many subjects. However, as pointed out earlier in the present
chapter, the relation between a significant p value and effect size is not clear, and
thus drawing conclusions regarding magnitude of effect from significance values is
risky. Power analysis provides a much more accurate estimate of these parameters
and it should be used in its place.

Despite Cohen's efforts, power analysis has historically been given minimal
attention by psychological researchers. Only a few textbooks have discussed sample
size in its relation to statistical power (e.g., Glass & Stanley, 1970; Howell, 1989;
Furthermore, as noted by Olejnik (1984), many of these sources restrict the discussion to one-way ANOVA designs. Reviews by Sedlmeier and Gigerenzer (1989) report little increase in power since Cohen first recommended it for social and life sciences. Out of 54 articles reviewed, only two mentioned power, and none actually calculated it. Furthermore, Sedlmeier and Gigerenzer found the median power in 1960, .46 for a medium size effect, to have dropped to .37 in 1984. Only 3 power surveys have been conducted specifically for psychology since Cohen (1962). In addition to the study by Sedlmeier and Gigerenzer (1989), a survey by Chase and Chase (1976) found mean power to be .66 for medium effects, quite higher than that found by Cohen and that found by Sedlmeier and Gigerenzer (1989). A more recent study by Rossi (1990) suggests power to be higher than results in Cohen's first survey. In a survey of three journals, Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology, and Journal of Personality and Social Psychology, all selected from the year 1982, average power was found to be .17 for small effects, .57 for medium effects, and .83 for large effects. These results suggest an increase in power since Cohen's 1962 survey. However, as noted by Rossi, these results are in no way reason for great optimism:

The general character of the statistical power of psychological research remains the same today as it was then: Power to detect small effects continues to be poor, power to detect medium effects continues to be marginal, and power to detect large effects continues to be adequate. (p. 650)
Thus, power estimates continue to be on average, low for psychological research studies, as well as research in other fields. Research scientists continue to pay little attention to sample size when conducting experiments. As remarked by Tversky and Kahneman (1971), this makes for sloppy research practices: "Cohen's analysis shows that the statistical power of many psychological studies is ridiculously low. This is a self-defeating practice; it makes for frustrated scientists and inefficient research" (p. 107).

Why the Lack of Power Analysis?

Attempts to discern why psychologists continue to ignore the calculation of power have produced as of yet, no definite answer. There seems to be hardly any "rational" reason why the technique is ignored to the extent it is. As Cohen (1977) has remarked, likely in frustration, "since statistical significance is so earnestly sought and devoutly wished for by the behavioral scientists, one would think that the \textit{a priori} probability of its accomplishment would be routinely determined and well understood" (p. 1).

Despite Cohen's push for power, it remains poorly understood by today's investigators, and is seldom used in psychological studies. If power analysis is such a wonderful tool, then why is it not used? I suggest three potential reasons as to why power analysis is practically ignored by today's community of psychological researchers: 1) ignorance, 2) laziness, and 3) lack of responsibility for proper research.
In considering the first reason, that of ignorance, I suggest that few researchers are aware of the basic necessity of conducting power-analytic procedures, given the use of significance testing. In other words, few psychologists may be truly aware of the "meaningfulness" of their studies if a power analysis is not performed to ascertain a good probability of rejecting a false null hypothesis. Because it is well-known that many psychologists misuse and misinterpret the statistical tools they use daily (Brewer, 1985), it would surely be of no surprise to find that the methodological requirement of power is generally poorly understood as well. As Rossi (1990) notes, "It is probably not an exaggeration to assert that most researchers know little more about statistical power than its definition, even though a routine consideration of power has several beneficial consequences" (p. 646).

Or, perhaps the mathematical calculations in power analysis (as elementary as they are) prove intimidating to psychologists, who, on the average, have extremely little mathematical training compared to members of the natural sciences. As recently argued by Estes (1997), the community of psychology students and researchers are not adequately prepared or trained in understanding the statistical tools they use. The implication here is that better-trained researchers (mathematically) might be more willing to investigate power analysis. However, even to gain a moderate understanding of power analysis does not require extensive

---

5 By "meaningfulness," I imply that a good study may have been performed, but because power was not calculated, the experiment might be dismissed as not providing evidence for the theory under test. Thus, although the study may indeed be "meaningful," (for instance, large effect size) a researcher might mistakenly consider it to be meaningless because of the lack of power in rejecting
mathematical ability. Also, many computer software programs exist that will perform the math and determine sample size for the researcher. Goldstein (1989) provides an excellent review of available computer programs used for power analysis and his paper is highly recommended for consultation for those seeking a computer package.

The second possible reason for the lack of concern for power is that of laziness. Of the few researchers that are aware of the theoretical necessity of power, few may be prepared to bother with the technique. After all, any new calculations are time consuming and provide an opportunity cost against other activities that may occupy the researcher, such as summarizing results of other "power-deficient" studies. Furthermore, since editors (or APA) do not require power analysis, many researchers probably don't consider it a necessity to calculate. Why make life more complicated when editors don't require it to be?

The third reason why power is generally ignored, follows from that of the second. Researchers may be unwilling to bother with the calculation of power, even if they know it to be an important and necessary inclusion in any research report. However, as mentioned, since journal editors do not require its inclusion, many do not bother with its calculation. Knowing that power should be calculated, but refusing to do so because you won't be penalized for it, I refer to as lack of responsibility for proper research.
From the above discussion, one important consideration remains. If knowing the power of an experiment and calculating it appropriately can only increase the likelihood of publication, why wouldn't researchers rejoice in a method of ensuring to reach the revered .05 level of significance if in fact the tested null is false? As Rossi (1990) correctly puts it:

Knowledge of power of a statistical test indicates the likelihood of obtaining a statistically significant result. Presumably, most researchers would not want to conduct an investigation of low statistical power. The time, effort, and resources required to conduct research are usually sufficiently great that a reasonable chance of obtaining a successful outcome is at least implicitly desirable. Thus, if a priori power estimates are low, the researcher might elect either to increase power or to abandon the proposed research altogether if the costs of increasing power are too high, or if the costs of conducting research of low power cannot be justified. (p. 646)

The truth of the matter is however, as pointed out by Schmidt (1996), that if every researcher were conscientious about power, many studies wouldn't be conducted at all. The reason for this is the high number of subjects needed to have an even moderately powerful test. Schmidt notes correctly that even with power at .80, this still allows a 20% Type II error rate when the null hypothesis is false. In order to obtain power this high, the number of subjects needed are often in the order of 1,000 or more (Schmidt & Hunter, 1978). That in itself could account for why most researchers ignore power; they know just enough about it to know they never achieve
high power, so why diminish the credibility of their research studies with calculations pointing out their weaknesses! Although power may be considered as something of a life-saver if NHST is retained as the dominant model of statistical inference, the difficulty in achieving reasonable power may be problematic, and thus constitutes another problem directed at the NHST model. Conclusion -- power doesn't save NHST, it only makes it more bearable.

VI. Multiple Models

Another alternative to traditional NHST is that of specifying how well the data fit various models. This approach, advocated by Edwards (1965) and later discussed by Wilson, Miller, and Lower (1967), is one in which the data are not automatically assumed to derive from the null distribution, but are left open for any model that best accounts for it. As Wilson et al. argue, traditional classical statistics usually limit the specification of models upon considering the data:

Multiple models seem thoroughly desirable, and it seems worthwhile to note that in a traditional two-tailed test, classical statistics always implies three families of models: one predicting no difference, one predicting a positive difference, and one predicting a negative difference. (p. 196)

By comparing models to discern which is most accountable for the data, this would presumably avoid the need to "reject" a null hypothesis, or at least minimize its importance. The null hypothesis, which merely represents the distribution (i.e., model) that the data are distributed according to chance would indeed be rejected if a more plausible model were found to better account for the data. As Edwards (1965)
illustrates through an example, comparing models is a necessary prerequisite to
determining the plausibility of the null hypothesis:

A man from Mars, asked whether or not your suit fits you, would have trouble
answering. He could notice the discrepancies between its measurements and
yours, and might answer no; he could notice that you did not trip over it, and
might answer yes. But give him two suits and ask him which fits you better, and
his task starts to make sense. (p. 402)

Dunnette (1966), in addition to defaming NHST, has recommended that
psychology advance by testing not one grand theory, but rather by testing a multitude
of hypotheses. As Dunnette suggests:

The approach entails devising multiple hypotheses to explain observed
phenomena, devising crucial experiments each of which may exclude or
disprove one or more of the hypotheses, and continuing with the retained
hypotheses to develop further subtests or sequential hypotheses to refine the
possibilities that remain....However, in psychology, the approach is little used,
for, as we have said, the commitments are more often to a theory than to the
process of finding out. (p. 350)

More recently, Dixon and O'Reilly (1999) have recommended that psychologists
use the maximum likelihood ratio in comparing two or more models to the obtained
data. They conclude that since most investigations discuss alternative plausible
hypotheses that could account for their findings, why not include these models in the
main analysis? They argue that as an alternative to null testing, a comparison of
alternative hypotheses doesn't simply reject a "meaningless" null hypothesis, but rather compares how the data fit any one of the posited alternatives by means of the maximum likelihood ratio. According to the authors, a likelihood ratio of greater than 10 constitutes substantial evidence for siding with one model over another. They stress however that decisions regarding the importance of the research should not be limited solely to the value of the likelihood ratio. In effect, these researchers are arguing for a similar philosophy of inference as Edwards and Dunnette, but at the same time are recommending the maximum likelihood ratio as a reputable means of comparing alternative hypotheses. The maximum likelihood ratio is the component of a Bayesian analysis that allows the researcher to update belief in light of evidence, and will be more thoroughly discussed in the following chapter.

Such recommendations by the preceding authors make much sense. Although the nature of the problem discussed is not merely a statistical one, it may be precipitated by psychologists' obsession to state one alternative instead of trying to specify several substantive hypotheses that could just as well account for the data. Customarily, researchers are interested in confirming the hypothesis of interest, and not in comparing that hypothesis against rival hypotheses within the same investigation. As Dunnette has said, we are hardly interested in "finding out" the answer as we are in "finding support" for a given theory.

VII. The Good-Enough Principle

Serlin and Lapsley (1985), in response to psychology's methodological problems highlighted by Meehl (1967, 1978), have proposed the "good-enough" principle.
Their model adheres to Popperian falsificationism, and does not always reject the null hypothesis given a large enough sample size. Well-deserving of proper space in a chapter of alternatives, the details of their proposal will now follow.

According to Popperian falsificationism, scientists, whether physical or social, should specify \textit{in advance} what they will accept as a falsifying instance for a theory under test. Implied in this is that the scientist will also predict what data, or range of data will be accepted as support for a given theoretical position. The former is what Serlin & Lapsley refer to as the "good-enough" principle. The principle requires that scientists predict in advance what results will count as "good enough" for support of a theory. The main tenet of the model rests on the assumption that when specifying a hypothesis under test, the hypothesized "true value" can never be \textit{exactly} true in theory, but must rather have some margin (or width) to vary. Thus, the predicted value becomes a value \textit{plus} a given width around that value, much in the sense that a confidence interval will specify a "margin of error." Once the results of the experiment are available, a statistical test is used to determine if the observed value is in the range of the "good-enough belt," that is, the area around the predicted value that was formerly specified as "good enough" to act as confirming evidence for the theory. Following this procedure, the problems of "sample-size-sensitivity" are eliminated, because as precision increases (i.e., sample size increases, resulting in a wider estimate), the imprecision involved in estimating the population value is reduced. One would then find theoretical support for a theory \textit{outside} of the previous "good enough" range. Hence, as Serlin & Lapsley argue, "even with an infinite
sample size, the point-null hypothesis, fortified with a good-enough belt, is not always false" (p. 79).

As noted by these authors, this principle holds just as well for what occurs in physics as for what transpires in psychology, in that, "nature is just as unkind to physicists as it is to psychologists" (pp. 79-80). This is to say that even with the most precise of experiments (i.e., of the kind conducted in physics), the obtained value will never be exactly identical to the predicted value. There always must be a "belt" to account for the margin for which the data will still be considered to support the theory. Serlin and Lapsley argue that it is only through the use of such a "belt" that a precise experiment (i.e., one that predicts a point-value) can yield important information, especially considering that the most probable outcome is that the observed value will not equal the predicted or theoretical values.

The theoretical logic behind the principle is best appreciated when considered for directional hypotheses in disciplines such as psychology. As argued by Serlin and Lapsley, "under the aegis of the good-enough principle, one may not merely predict a direction. One also must specify in advance the magnitude of the change in that direction that is good enough" (p. 80). Good enough for what? "Good enough to still count as confirming evidence for the theory. As they continue, "if the statistical test indicates a possible increase less than that which is specified as good enough, the directional null hypothesis is retained. With infinite precision, one does not always reject the directional null hypothesis, and this is especially advantageous when the result is in the correct direction but only infinitesimally so" (p. 80). The authors also
argue that such a specification of the magnitude of change accords well with
Lakatosian principles of scientific methodology in that a theory that specifies the
degree of change in a specified direction is better than a theory that specifies merely
the direction (i.e., such as traditional null hypothesis testing). Serlin and Lapsley
(1999) provide a good example of how to use a statistical test in conjunction with the
good-enough principle (see their 1985 article for details).

Is the Good-Enough Principle Good Enough?

The good-enough principle is excellent and has much to offer in terms of how
we interpret psychological data. More similar to confidence intervals than
hypothesis-testing, the good-enough principle allows for an estimate of a mean (or
mean difference), while specifying what results will count as "significant" should the
precise prediction not be the observed value (i.e., the theoretical prediction does not
match up to the observed value). This method of hypothesis-testing is advantageous
over traditional NHST in that it specifies the range of values that will count toward
support for a theory rather than merely reject chance as a possibility for the data. If
there is a disadvantage to the good-enough philosophy, it is that it remains unknown
how to specify the "good enough" region, and therefore this constitutes a similar
problem as that of specifying the alpha level in hypothesis testing. According to
Serlin (1993), the width of the belt is determined by "the state of the art of the theory
and by the strengths of the auxiliary theories on which the prediction is based" (p.
351). In other words, if the theory is very successful and prominent, the width of the
belt would tend to be smaller because of our prior "confidence" in the theory. We
would be prepared to "zero in" on some phenomenon. On the contrary, if we are
testing a rather lose and non-established theory, we would tend to specify a larger
belt. What is key to the good enough test is that the magnitude of effect must be
specified before the procedure is run. Each theory has various expected effect
magnitudes, and these are determined by what the researcher would except in light of
the success or failure of the given theory based on past research. The concept sounds
similar to that of estimating a prior probability within the context of a Bayesian
analysis. As emphasized by Serlin (1993), the goal of psychological research is to
estimate parameters of the population, not merely to reject chance:

Simply concluding that the population results seem to be or seem not to be good
enough does not do us justice. Rather, if theory is supported, we would like to
know how big the effect really is; if theory is not supported, we want to know
how close we came. In either case, the magnitude of the effect can be used to
gauge theoretical progress. It is for these reasons that most of the critics of
significance testing suggested the use of confidence intervals as a way of
improving the scientific utility of statistical methodology. (p. 352)

Evaluation

The good-enough principle shines over NHST because of its ability to deal
effectively with large samples, thus increasing the precision of estimate, without
inflating the degree to which the chance hypothesis will be rejected. The success of
this model owes much to the fact that it is similar in theory to confidence intervals, in
that point-estimation is the goal, not merely the rejection of an unlikely theoretical
distribution. The good-enough model represents much advancement over null testing, and can be considered as a complete substitute to the NHST model.

Alternatives Within the NHST Paradigm

Some authors recommend keeping NHST but changing the way in which we report findings and utilize the procedure. I will now briefly discuss recommendations based on "modifications" to NHST rather than complete abandonment of the statistical model. Rather than specifying alternatives to NHST, the following recommendations attempt to save (at least somewhat) NHST from complete failure through sharpening the means by which we use the model. Thus, the question to answer in the following is whether these recommendations do salvage NHST to any degree or whether they are meaningless attempts to revive a doomed model. Two recommendations for improvement are discussed below.

Among the popular recommendations is that the exact significance level be reported in results sections rather than merely reporting whether \( p \) exceeds a given alpha level. As Morrison and Henkel (1970) argue, following this procedure fosters a view to report only the data, and not imply conclusions based solely on the data. Implied in their argument is that too many times decisions are reached with regards to experiments (and maybe even theories) based on the "greater" or "lesser" than alpha level. Indeed, as noted in chapter 1, the reporting of the specific probability at which the null is rejected has been recommended elsewhere (e.g., APA, 1994, and even Fisher himself). Although I personally agree that it is wise to report exact \( p \) values, this is hardly enough to account for the problems of NHST. Therefore, it is
concluded that although a nice amendment, including exact p values in results sections is like topping off a poorly baked cake with high-quality frosting.

A second recommendation to "save" null hypothesis testing, is to restructure the vocabulary used in reporting results, specifically, the use of the term "significance." As argued by some, (e.g., Carver, 1993; Scarr, 1997; Thompson, 1996), the term "reliability" should be used in place of the term "significance" for the purpose of eliminating the supposed confusion that exists with "significant" results implying a degree of importance. Although I would agree that some researchers may confuse the significance of a result with its importance, the well-educated and conscientious psychologist would not make such an error. As concluded in chapter 2, there are much more urgent problems when it comes to our statistical procedures than a mere semantic difficulty. I therefore declare the "significance" distinction to be a non-issue.

Conclusion and Comment

The primary goal of this chapter was to review alternatives that could be used in place of traditional null hypothesis significance testing. Whereas I hope this goal was accomplished, I hope even more that an additional purpose was achieved -- that of bringing awareness of just how many good, feasible alternatives there are to NHST. Given these attractive alternatives, one could easily write a book in exploring the reasons why the psychological community as a whole has been so resistant to alter its practices with respect to statistical inference.
In concluding this chapter, I will not do as I thought I would originally do from the outset -- that of recommending a single alternative to NHST. To make such a suggestion would be to imply that all psychological data should be analyzed the same way, whether coming from social, cognitive, or general experimental domains. What I can recommend however, based on my review, is that NHST as we know it, be either 1) abandoned, or 2) complemented by one or more of the procedures outlined above. I would venture as far as to say that NHST is not even required for most studies, and assuming the regular inclusion of confidence intervals, one or more of the alternatives listed above would more than satisfy our expectations of what a data-analytic tool should provide.
Chapter 4

The Bayesian Alternative to NHST

Alternatives to null hypothesis significance testing abound, as demonstrated by the previous chapter. Another alternative, the Bayesian model, will be considered in the present chapter. Contrary to other alternatives, Bayesianism is deserving of a separate chapter for the reason that it represents a shift in philosophy rather than merely a shift in statistical tooling. In this regard, I agree with Chow (1996): "The arguments in favor of using Bayesian statistics do not so much constitute a rejection of NHSTP as they suggest a new approach to empirical research" (p. 144). The various approaches presented in chapter 3 all have one basic commonality -- they are all based on the "frequentist" approach to statistical inference. That is, they all assume probability to be a relative frequency. The Bayesian approach measures probability not in terms of frequency, but in terms of subjective belief. The typical Bayesian would likely agree with Savage (1972), in that "personal (or subjective) probability is a good key, and the best yet known, to all our valid ideas about the applications of probability" (pp. iii-iv). The Bayesian model relates the concept of probability to one's degree of belief in a hypothesis rather than as a frequency. Indeed, it is this interpretation of probability that makes Bayesian statistics distinct from classical statistics. The classical statistician argues that statistical inference should be based on "objective" frequencies originating from sampling theory (Keren
& Lewis, 1993), while the Bayesian statistician holds that this objectivity is something of an illusion, and that probability is a function of one's "subjective" belief. Although the Bayesian may use relative frequency in quantifying his or her belief, using it is but only one option for estimating his degree of certainty in a hypothesis.

The goal of this chapter is to evaluate the Bayesian statistical model as a candidate for replacing current null hypothesis significance testing. In discussing the Bayesian approach, I will first provide a brief overview of Bayesian fundamentals. This overview will review key Bayesian mechanics. I will then present the three most significant advantages in employing a Bayesian statistical model for psychology. As will be seen, Bayesianism is highly recommended by many philosophers of science and social scientists alike. I will then present some arguments against the Bayesian model, leading to a discussion of its most severe presumed drawback, that of prior probabilities. Although prior probabilities constitute the strength of the Bayesian model for Bayesians, they constitute a serious weakness for frequentists. Whether the problem of prior probabilities is reason enough to discount the model will be discussed in light of the importance placed on subjective probabilities. If the Bayesian model is in most respects superior to the frequentist approach, and thus constitutes a viable alternative to NHST, can the problem of prior probabilities be dealt with successfully, or are the problems of prior probabilities too much for the Bayesian model to overcome? In approaching this
problem both from a Bayesian and frequentist position,¹ I will draw the conclusion that given either interpretation of probability, the Bayesian model is superior to NHST. I will close in arguing that prior probabilities do not constitute the significant problem that frequentists claim, and thus do not warrant discounting the Bayesian model. "Posterior probability convergence" will be shown to be the winning argument for Bayesianism, whether one initially assumes a frequentist or Bayesian position.

The Bayesian Approach to Statistical Inference

The Bayesian model of hypothesis testing originated with Thomas Bayes² in the 18th century (Bayes, 1763). As noted by Brenner-Golomb (1993), Bayes developed his theory in attempting to deal with the uncertainties of induction. Unlike deduction, induction always involves an amount of guesswork or uncertainty. Bayes' goal was to account for this "unknown" quantity (i.e., uncertainty) through probabilities. For him, as for Fisher over one and a half centuries later, "the nature and degree of the uncertainty may itself be capable of rigorous expression" (Fisher, 1947, p. 4). It was how this estimate of uncertainty was derived that divided Bayes' subjective approach to probabilistic inference from Fisher's relative-frequency approach. Fisher (1966) completely opposed the Bayesian axiom of inverse probability, that of the probability of the hypothesis given the data:

¹ Obviously, subjective priors are not problematic from the Bayesian viewpoint, but rather constitute the strength of the subjectivist approach. Consequently, the topic of "problematic priors" will be approached from a frequentist position. That is, priors will be critically evaluated from a frequentist perspective.
The axiom leads to apparent mathematical contradictions. In explaining these contradictions away, advocates of inverse probability seem forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes. (pp. 6-7)

Early objections to inverse probability, as noted by Gregson (1997), may not have originated with Bayes' contribution, but rather with Venn in the 19th century. According to Oakes (1986), Venn was the most "vehement" opponent of the classical account of probability. Venn was among the first to draw a distinction between the frequentist and subjectivist interpretations of probability:

Is our natural belief in the happening of two different events in direct proportion to the frequency with which those events happen in the long run? There is a lottery with 100 tickets and ten prizes; is man's belief that he will get a prize fairly represented by one-tenth of certainty? The mere reference to a lottery should be sufficient to disprove this. Lotteries have flourished at all times, and have never failed to be abundantly supported, in spite of the most perfect conviction, on the part of many, if not of most, of those who put into them, that in the long run all will lose. (Venn, 1886, p. 128)

Although Bayes and Venn were the early advocates of inverse probability, as remarked by Dale (1991), the contemporary rise in Bayesian interest may be

---

2 Despite Bayes being the founder of a major statistical movement, as noted by Dale (1991), little biographical information is known. Even his date of birth is subject to debate.
attributed as a reply to Fisher's ideas: "One can perhaps view the rigorous development of Bayes's work into a statistical tool to be reckoned with as a reaction to Fisher's evolution of sampling theory" (p. x). For Bayes, if a hypothesis is to precede evidence, then the set of hypotheses includes all possible circumstances that may precede the evidence. In other words, it is assumed that given the evidence, all plausible "causes" (Brenner-Golomb, 1993) that may account for the evidence are contained in the set of mutually exclusive hypotheses. These hypotheses are then evaluated and assigned probabilities by the investigator. In the Bayesian paradigm, the truth of a hypothesis is partly based on the subjective belief of the person doing the hypothesizing, rather than on an observable frequency as in the frequentist Fisherian camp.

The Bayesian approach to inference revolves around updating subjective belief. The mechanism used for incorporating new data with prior information is Bayes' theorem (Winkler, 1993). Although other means of updating subjective belief exist (Diaconis & Zabell, 1982; Jeffrey, 1983), Bayes' rule is the most frequently used in achieving this end. Contrary to the Fisherian/Neyman-Pearson approach to statistical inference, in which data D is evaluated in light of hypothesis H, in a Bayesian

---

3 This may account for why significant developments in Bayesian statistics are relatively recent (Keren & Lewis, 1993). These developments may have arisen in opposition to the Fisherian model of significance testing. According to Winkler (1993), many classical statisticians viewed the Bayesian model as controversial during the late 1950s and 1960s. However, controversy has died down and work in Bayesian statistics has steadily increased. Presently there is much Bayesian activity in statistical departments in various countries (Winkler, 1993).

4 As will be demonstrated later in this chapter, Bayesian priors prove to have minimal influence across successive Bayesian trials.

5 The "complement," represented in the denominator, may consist of "~H" (i.e., not H) or any other hypothesis that is competing against H.
analysis, the probability of $H$ is evaluated in light of $D$. The Bayesian model tests the hypothesis that is of most interest to the researcher (i.e., research hypothesis) and evaluates its probability in light of new information. Thus, Bayes' theorem, in which the truth (i.e., probability) of one hypothesis is set in terms of its complement, \(^5\) is represented as:

$$
P(H|D) = \frac{p(H) \cdot p(D|H)}{p(\neg H) \cdot p(D|\neg H) + p(H) \cdot p(D|H)}$$

The term on the left-hand side of the equation, $p(H|D)$, is equal to the probability of the hypothesis given the data. The first term in the numerator, $p(H)$, called the prior probability, represents the prior odds that $H$ is true in light of the information that is known before $D$ is available. The second term in the numerator, $p(D|H)$, is called the likelihood ratio for $H$. This is the probability of $D$ given that $H$ is true. The first term in the denominator, $p(\neg H)$, the complement of $p(H)$, represents the probability that $H$ is not true. This probability may represent a competing hypothesis to $H$ or it may merely represent the negation of $H$. The second term in the denominator, $p(D|\neg H)$, represents the likelihood of the data given the negation of $H$. The comparing of likelihood ratios (i.e., $p(D|H)$ to $p(D|\neg H)$) constitutes the most crucial component of the Bayesian theorem. It is this ratio that essentially determines whether the posterior probability of $H$ will be different from the prior probability of $H$. If $p(D|H)$ is the same as its complement, $p(D|\neg H)$, then
the posterior probability will undergo no change and will be the same as the prior probability. This is simply because the data remain the same regardless of the truth of the hypothesis. However, given any discrepancy between \( p(D|H) \) and \( p(D|\neg H) \), the likelihood ratios will be different, and the posterior probability will be different from the prior probability. The likelihood ratio is not exclusively a Bayesian principle and is used in virtually all statistical methods (Phillips, 1973).

Consequently, as noted by Clark (1963), "many statisticians, including Fisher, have argued that the likelihood ratio on its own, provides sufficient information for data interpretation" (p. 464).

It should be noted that although the Bayesian model is often perceived as diametrically opposed to the frequentist model, as remarked by Grayson, Pattison and Robins (1997), the Bayesian and frequentist views may overlap. For instance, given a fair coin, the frequentist may believe that if tossed many times, the relative frequency of "heads" will be approximately 0.5. And it is just for this "relative frequency" reason that the Bayesian will likely set his/her prior probability of a "head" at 0.5. Thus, while the Bayesian obtained his probability through "subjective" means, his opinion can be influenced by frequentist or analytical assumptions regarding the nature of probability.

Advantages of Bayesian Statistics Over NHST

How does the Bayesian model stack up against NHST? I will now review three reasons why Bayesian statistics are held by its supporters to be superior to NHST. These are: 1) use of prior information, 2) testing of research hypothesis, and 3) few
Bayesian problems. A basic question that needs to be answered is whether or not the Bayesian model solves the problems that trouble NHST, or whether it exacerbates more of its own problems than it solves. In addition to evaluating each advantage, I will, where applicable, critically evaluate these advantages on their own merit, independent of comparing them to NHST.

1. Use of Prior Information

A first posited advantage of Bayesian statistics over classical statistics (i.e., NHST) is their cumulative feature. Using null hypothesis testing, psychology has been unable to cumulate evidence effectively (Rosenthal, 1993). The Bayesian model allows for prior information to be used in the analysis of future data. As argued by some (e.g., Berry, 1996; Iversen, 1984), prior probabilities allow for past experimental information to be reflected in future investigations. If hypothesis H has been confirmed many times in past research, this information may be used in determining the prior probability of the hypothesis currently under test. For instance, if an experimenter conducts an investigation of which ten of the past studies have yielded confirming results, then the investigator may use this evidence in estimating a high prior probability. Granted that this new prior probability is still expressed as the experimenter's subjective opinion, Bayesians argue that this information can nonetheless be used efficiently rather than having to "start over" each time an experiment is performed. Although the prior opinion specified by each researcher is subjective, it is nonetheless based on the available evidence and thus according to the Bayesian, constitutes a very rational way of estimating a distribution. Bayesians
claim that this quality makes their model one for which knowledge of a hypothesis is cumulative. This is not so in classical statistics, where each time an investigation is performed, researchers are not required to begin with an "overt" prior distribution. At this point, something requires clarification. That Bayesians specify prior probabilities "subjectively," means only that the probability is derived from the person. "Subjective" in this sense is not meant to imply that the assignment of Bayesian priors is irrational or arbitrary. For this reason, Bayesianism is often referred to as "personalism," to avoid these negative connotations. Indeed, as argued by the Bayesian, Bayesian probabilities follow a rigorous pattern, one they would argue is worthy of the term "rational."

It is also worth highlighting that the Bayesian prior probability is designed to reflect not only a "personal" probability, representing one's personal belief, but also a probability estimate that is grounded in something more than "subjectivity." The Bayesian prior should reflect what is known regarding the hypothesis of interest. As argued above, prior information is assumed to have been considered in arriving at a prior probability. Thus, again, I emphasize that Bayesian "subjective" probability does not reflect arbitrariness or an irrational assignment of probability, but rather denotes that probability is a function of a person's belief (thus the term "personalism"), and should be grounded in the prior knowledge of the researcher.

---

6 The classical statistician may still be said to postulate a prior distribution before determining the probability of the data. However, as discussed later in the chapter, this distribution is rectangular.
In defense of NHST, Wilson, Miller, and Lower (1967) argue that classical statistics, similar to Bayesian statistics, use a prior probability estimate:

For one thing, a choice of significance level can, under some circumstances, be construed as a prior probability estimate — not the subjective one of the individual scientist but an admittedly arbitrary attempt to standardize a bias against alternative hypotheses. It appears to be a deliberate attempt to offer a standardized, public method for objectifying an individual scientist’s willingness to make an inference. (p. 191)

In response to this claim, it is not evident how a significance level can be construed as a prior probability estimate. Estimate of what? Certainly not of the research hypothesis, as is the case in the Bayesian model. An imposed significance level has little resemblance to a Bayesian prior. At most, one could say that a significance level, if stated beforehand (i.e., before the data are collected), can be considered an estimate of when to reject the null hypothesis. However, even this does not explain what Wilson et al. mean by "estimate." Indeed, a significance level is nothing more than a pre-established convention that is used to rule out the possibility that the data were obtained by chance, not from a research hypothesis. The parallel that Wilson et al. try to draw between classical statistics and Bayesian statistics consequently, does not work.

Prior information has also been found to be useful in decision-making. Consider an excellent example offered by Phillips (1973) in which prior information is crucial to the decision-making process. In his example, a clinician is interested in diagnosing
patients as either functionally ill, or brain-damaged. In doing so, he relies heavily on research reported by Page et al. (1957) who found that 85% of people who are functionally ill report seeing the "spiral after-effect" illusion, while only 40% of brain-damaged patients report seeing the illusion. Given only these figures, the clinician is likely to diagnose functional illness if the patient sees the illusion, and diagnose brain damage if the patient does not see the illusion. Based on the data provided by Page et al., and limited to diagnosing based on the "illusion test," this would at first sight appear to be a wise decision rule. However, as Phillips notes, when prior information is considered, the following is found. In consulting the records of the hospital, it was found that 90% of the patients admitted in the past suffered from functional illness, while only 10% suffered from brain damage. Using these figures as prior probabilities, and doing the necessary calculations using Bayes' theorem, it was found that given the person sees the illusion, the posterior probability that they are functionally ill is 0.95, while the posterior probability that they are brain-damaged is 0.05. Indeed, as Phillips remarks, the clinician would be better diagnosing functional illness over brain damage if the patient sees the illusion, as the clinician first thought. However, recall that the clinician would have diagnosed brain damage if the patient did not see the illusion. Calculating Bayesian posteriors, the probability of brain damage given the patient does not see the illusion is only 0.31, while the posterior is 0.69 for diagnosing a patient as functionally ill.⁷ Thus, based on Bayesian posteriors, it is concluded that the clinician would be better off

⁷ See Phillips (1973), pp. 72-74 for calculations.
diagnosing functional illness irregardless of whether the person sees the illusion or not. In other words, the utility of the test is questioned since it does not result in a higher probability of making a correct diagnosis.

Prior Information: Exclusive to Bayesianism?

Is prior information really an exclusive advantage of Bayesian statistics over classical statistics? Yes and no. Yes, because as argued above, data analysis and consequently hypothesis-evaluation would appear to be more "cumulative" than in classical statistics. However, it should be noted that even with NHST, science proceeds in a cumulative fashion, especially if replications are performed. Many replications of an experiment do constitute "cumulative confidence" in any hypothesis whether it is analyzed through Bayesian or classical means. The point is that for a research paradigm to be cumulative, it must build on prior research. Both the Bayesian model and NHST do this, but in different ways. The Bayesian model measures prior information in terms of a probability, while the classical model must assume this cumulative progress through the originating of hypotheses and research interests. Surely, if a hypothesis has been refuted many times, a researcher will be unlikely to test the exact one again. Thus, he uses "prior information" in determining not to test it again. He may also rely on previous published research in making his research decisions. The difference in Bayesian terms is that the investigator may decide to test the hypothesis again, but assign it a low probability, based on prior research. In the NHST model however, the investigator might discard the hypothesis.
and not test it. In sum, the Bayesian approach makes the cummulation across experiments explicit, whereas it is only implicitly assumed in NHST.

In terms of decision-making, the Bayesian model lends itself particularly to this end, whereas NHST offers no feasible alternative. As cited in the example above by Phillips, Bayesian reasoning may be very useful when making a decision where prior information is available. However, a distinction needs to be made. Is the Bayesian model as applied to research the same model as applied to decision-making?

According to Howson and Urbach (1993), the two are not compatible with each other: "The intellectual tools one needs to assess such things as empirical support, weight of evidence, and probability itself, have nothing to do with making decisions, in any but a purely trivial sense" (p. 439). Decision-making would appear to have little to do with evaluating the probability of a research hypothesis, or gathering substantive support for a theory. However, caution must be warranted against such a conclusion. The line between Bayesian hypothesis-evaluation and decision-making is sometimes not clearly drawn. According to Winkler (1993), "the motivation for much of the recent interest in Bayesian methods has been decision-theoretic in nature" (p. 228). Is it not possible then that researchers use Bayesian decision-making in evaluating their hypotheses? For instance, in considering the example by Phillips, how are we to differentiate between it being a decision-making process, and it representing a long set of individual hypotheses? What can be considered a decision-making process can also be considered hypothesis-evaluation in this case.

For example, each case taken separately, the research hypothesis could be stated as
the presenting patient having a functional illness, while its complement would be that the patient has brain-damage. Given the Bayesian analysis, it becomes clear that the probability for functional illness is high, and therefore a decision is based on this probability. Indeed, each individual decision for each patient would be said to result from the research hypothesis (again, that of functional illness) boasting a high probability. Though I generally agree with Howson & Urbach, I merely note that the boundary between what is regarded as a decision and what is regarded as hypothesis-evaluation is not always clear. For the purposes of this chapter however, I shall presume Bayesian decision theory to be separate from that of Bayesian confirmation. I will assume as Howson & Urbach do, that when it comes to empirical support and weight of evidence, "decision theory presupposes these tools, not they it" (p. 439). However, as shown, should a paradigm be centered around decision-making, the Bayesian model comes highly recommended (Winkler, 1993).

2. Testing of Research Hypothesis

A second significant advantage of Bayesian statistics over traditional NHST is the testing of the research hypothesis. As argued by some (e.g., Cohen, 1994), the Bayesian model tests the hypothesis that is of most interest to the researcher, rather than a mere null hypothesis whose "truth" has been the topic of much debate. Indeed, as expressed by Cosmides and Tooby (1996), the testing of the research hypothesis constitutes the main interest in Bayesian statistics:

In science, what everyone really wants to know is the probability of a hypothesis given data -- \( p(H|D) \). That is, given these observations, how likely is this
theory to be true?… The strong appeal of Bayes' theorem arises from the fact that it allows one to calculate this probability. (p. 4).

In addition, the Bayesian model tests the substantive hypothesis, rather than infer a statistical alternative as is the case with NHST. Bakan (1966) has long recommended that psychologists begin testing the research hypothesis rather than continue with NHST:

We need to get on with the business of generating psychological hypotheses and proceed to do investigations and make inferences which bear on them; instead of, as so much of our literature would attest, testing the statistical null hypothesis in any number of contexts in which we have every reason to suppose that it is false in the first place. (p. 436)

As first noted by Gigerenzer (1993), and later recalled by Cohen (1994), many researchers interpret NHST results in a Bayesian fashion, what Gigerenzer first called the "Bayesian Id's wishful thinking." Many investigators mistakenly conclude something about the research hypothesis based on the data. However, as Cohen points out, NHST tells us the probability of D given $H_0$, not the probability of $H_0$ given D, which is what a Bayesian analysis reveals. Cohen sides with Bayesian logic:

What is always the real issue, is the probability that $H_0$ is true, given the data, $p( H_0|D)$, the inverse probability. When one rejects $H_0$, one wants to conclude that $H_0$ is unlikely, say, $p < .01$. The very reason the statistical test is done is to be able to reject $H_0$ because of its unlikelihood! But that is the posterior
probability, available only through Bayes's theorem, for which one needs to know $p(H_0)$, the probability of the null hypothesis before the experiment, the "prior" probability. (p. 998)

As pointed out in chapter 2, the fact that NHST cannot assign a probability to the alternative hypothesis constitutes its greatest weakness and consequently constitutes the Bayesian model with its greatest strength. Can both models arrive at similar conclusions, given the same data? Johnstone and Lindley (1995) show that the compatibility of classical statistics with Bayesian statistics is less than desired. For instance, these researchers found that point-null hypotheses that were reported significantly rejected at a particular alpha level should not be considered false when evaluated under the Bayesian model. Furthermore, Falk (1998) has recently noted that rarely do Bayesian posteriors justify rejecting the null hypothesis, and recommends the Bayesian model: "Instead of relying on a lucky match between the two inverse conditional probabilities, we would better compute our target probability. $p(D|H_0)$ is problematic as an estimate of $p(H_0|D)$. Rare mathematical conditions have to hold for the two to be equal" (p. 316). Others however, have noted that regardless of the model, both classical and Bayesian statistics will yield similar results (Berger & Berry, 1988; Iversen, 1984). For example, Iversen (1984) compared a classical analysis to that of a Bayesian one. The research problem was to determine whether 5 population means were equal. Using classical statistics, the F value was found to be 5.20 at $p < .01$. Thus, using the classical model, the null hypothesis is rejected (i.e., there is a difference in means between the 5 groups).
Iversen then describes how such a problem could be approached from the Bayesian perspective. Assuming rectangular distributions as prior probabilities, the joint posterior distribution for the population means must be found. This distribution contains in it the most likely set of values of the 5 population means. To illustrate this concept using only 2 means (using 5 requires 5 dimensions), consider Figure 4. The oval shaded area represents the probability set obtained from the joint posterior distribution. Assuming this set to be equal to 99%, any point located on the line $\mu_1 = \mu_2$ has a probability of 0.01 of being outside the shaded area. Thus, the probability that $\mu_1 = \mu_2$ is 0.01. Or, conversely, the probability is 0.99 that $\mu_1 \neq \mu_2$. This is the same result as found when rejecting the null hypothesis at $p < .01$, only using NHST alone, we cannot attach a probability of 0.99 to $\mu_1 \neq \mu_2$. In applying this concept to the present problem of 5 means, the line would contain $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ instead of just $\mu_1 = \mu_2$. In relation to the classical model, the line (i.e., subset containing equal means) will only intersect the probability set (i.e., that area representing inequality among means) when $F$ is small.
Figure 4 -- Joint probability set; shaded area represents the probability that $\mu_1 \neq \mu_2$.

(Reprinted from Iversen, 1984.)

Conversely, when $E$ is large, the line will not intersect the probability set. Hence, this example demonstrates how a problem can be approached from both the classical and Bayesian viewpoints, and yield similar results.

Based exclusively on the advantage of directly testing the research hypothesis, one could make a very strong appeal for using Bayesian statistics instead of classical statistics in psychology. Indeed, isn't that what researchers want to know, the probability that their hypotheses are true? To calculate the probability of some data
given a hypothetical situation in which we really aren't that interested in anyways, it seems counter-intuitive. Bayesian analysis seems much more logical than classical statistical models. As Iversen (1984) argues, the true uncertainty lies with regards to the research hypothesis, not with the data:

Bayesian inference uses the concept of probability to express uncertainty about the quantities of which we are truly uncertain, namely the unknown parameters. Classical inference, on the other hand, computes the probability of the observed data. But there is nothing uncertain about the observed data. They are known and are there for everyone to see. (p. 76)

3. Few Bayesian Problems

A third advantage of the Bayesian model is that it does not share the difficulties that are usually associated with NHST. Difficulties such as the sample-size-problem, the priori falsity of the null hypothesis, and the inferring of a research hypothesis, are non-existent in the Bayesian model. This is even more encouraging since Bayesian statistics can provide similar information as found in a typical NHST analysis, namely interval estimates, tests of hypotheses and point estimates (Winkler, 1993). Also, Bayesian methods for assessing effects have been suggested (Rouanet, 1996). This is not to say that Bayesianism does not have its own problems (as will be discussed shortly), but only that given its scarcity of problems relative to NHST, this by itself can be considered an advantage. Once past its most controversial issue, that

---

8 A clarification is needed here. Although we are interested in rejecting chance from our data, this should be regarded as only one of many interests. In other words, our interest in rejecting the null should be secondary to our interest in the probability of the research hypothesis.
of prior probabilities (to be discussed at length later), the Bayesian model appears to be methodologically sound, is tailored to the research process (Iversen, 1984), and avoids most of the problems associated with NHST.

4. Other "Conditional" Advantages

It should be noted that many other advantages exist within the Bayesian model of hypothesis-testing. However, most of these are conditional upon accepting probability as a subjective entity. That is, one must first accept the basic axiom that probability is a function of one's belief, and not a function of theoretical sampling distributions if one is going to discuss further advantages. As an example, Winkler (1993) cites an advantage of the Bayesian model in that it "enables the statistician to find unconditional probabilities for future observables" (p. 230). Although this is true, a frequentist would have little difficulty in countering this argument simply by stating that he does not accept the assumption of subjective probability, hence, end of story. A list of advantages could be deduced from the Bayesian vantage point, but these are conditional upon accepting the foundational Bayesian postulate of probability being a function of one's belief. Posited as another advantage of the Bayesian model, is that it allows the calculation of the probability of a single event. Cosmides & Tooby (1996) give the example of calculating one's probability of having breast cancer given they have tested positive for the disease. Again however, this is only an advantage if one accepts the notion of subjective and prior probabilities. For one to conclude a probability of the event that one has cancer given a positive test, one must first assign an initial probability to the event. Then, and only
then, can the data (i.e., positive test) be used to revise the prior probability and produce a posterior probability. What I am pointing out can likewise be analogously expressed in terms of a conditional probability -- the Bayesian model works extremely well given (i.e., conditional upon) the acceptance of probability as a subjective entity. This topic will be more thoroughly discussed later in this chapter, when prior probabilities are analyzed in detail.

Criticisms of the Bayesian Model

As with any model of inference, there are critics. As mentioned earlier, the primary criticism of Bayesian statistics is that of specifying subjective prior opinion, and thus supposedly making the inferential process "non-objective." I will treat the issue of priors separately from that of other claimed disadvantages of the Bayesian model. What is often at issue with priors is not the priors themselves, but rather how the priors are conceived. Before addressing prior probabilities, I will first discuss other criticisms, and will render a "tentative" evaluation and conclusion of the Bayesian model minus the problem of priors. As will become evident, most criticisms of the Bayesian model are relatively easily resolved and dismissed.

1. Probabilistic Truth

A first rather harsh criticism against Bayesianism is made by Chow (1996). He implicitly faults Bayesianism because it uses probability to represent the "truth" of tested hypotheses. He says that "Bayesian analysis is not about the truth of hypotheses at all. Editor E does not collect data to accept or reject any of the
hypotheses" (p. 149). He then quotes Earman (1992) in defending this position:
"theories are not chosen or accepted but merely probabilified" (p. 150).

What Chow may consider to be a drawback of Bayesianism, I consider to be a advantage. Let me explain. Chow claims that in the Bayesian model, hypotheses are probabilistic and thus do not represent "truth." If Chow could provide a model for hypothesis-testing that could show a hypothesis to be "true," then his criticism of Bayesianism may have some merit. In quoting Earman, it would appear Chow is implying that since the Bayesian does not "accept" or "reject" any of the hypotheses, it is to be faulted for not providing the user with a built-in "decision-making" mechanism. What Chow fails to realize is that, as is the case with NHST, the Bayesian model is a probabilistic one, and therefore definite truth cannot be known regarding any hypothesis. As Rozeboom (1960) has pointed out, "belief in (i.e., acceptance of) a proposition is not an all-or-none affair; rather, it is a matter of degree" (p. 420). Of course, this fact which is well known to philosophers of science and methodologists, is somewhat disguised in NHST, where one is able to "accept" or "reject" a hypothesis, and in doing so, give the illusion that a "probable" result is not what has been achieved. The "accept or reject" NHST rule fosters an illusion of certainty. By contrast however, when we reject a null hypothesis and infer an alternative, we are not sure that the null is false, and we are even less sure that the alternative is true. What is important to realize is that regardless of the model used,

---

9 A criticism of the Bayesian approach similar to Chow's is the following: given the probability of a hypothesis H, it is debatable whether H could ever be confirmed, and hence represent a probability of one
NHST or Bayesian statistics, results are probabilistic. Chow's concern seems to lie more with Bayesianism making this probability overt, rather than artificially accepting or rejecting a hypothesis. Results are no less probabilistic in NHST than they are in Bayesianism.

2. Computational Concerns

A second criticism of the Bayesian model is that of computational difficulty. As noted by Grayson et al. (1997), "one of the difficulties inherent in the Bayesian approach has always been that of performing the necessary computations" (p. 69). According to Sheu and O'Curry (1998), this difficulty in computation can be present especially when estimating model parameters: "In real applications involving many parameters, Bayesian computation requires the evaluation of complex, high-dimensional integrals" (p. 232). Given these difficulties however, as in classical statistics, computational software has been made available to rectify these problems (Grayson et al., 1997; Gregson, 1997; Howson & Urbach, 1993; Sheu & O'Curry, 1998). With this increased technological advance, such computational problems are likely to disappear, and as claimed by Sheu & O'Curry, had this software been available much earlier, this may have resulted in a stronger resurgence of the Bayesian model: "We suspect that the ground-breaking work done by Edwards et al. (1963) three decades ago would have had a much greater impact on psychological research had it come bundled with computer software" (p. 237). There now exist (Vineberg, 1996). If Chow were basing his argument on this problem, it would be more credible. 10 For an example of such software, see Sheu and O'Curry (1998) for an illustration of the applicability of "BUGS" computer software for psychology.
tutorial programming software designed for teaching Bayesian reasoning, as well as performing the necessary calculations (Sedlmeier, 1997). Although the criticism of computational difficulty may have been merited in the past, today it is not. With the advance of statistical computer packages tailored to Bayesian inference, one need not be intimidated by the mathematical complexity of Bayesian statistics.

3. Scientific Gambling

Novick and Jackson (1974) note a third criticism directed at the Bayesian model, then quickly (and correctly) rebuke the critique. The issue is whether the idea of betting is relevant to scientific inference. The Bayesian model is said to be inappropriate by some because its main foundation (that of conditional probability) is allied to the concept of "gambling," that is, making a bet with a specified probability. The authors are quick to note however, that given any scientific inference, this issue cannot be avoided:

Only from the most remote ivory-tower position (scientific or moral) can this objection be sustained. If any action is to be taken as a result of our inferences, now or at some future date (and if not, why make them?), a little reflection shows that we are inescapably involved in a betting situation. (p. 145)

I would take this line of reasoning one step further and argue that Bayesian "gambling" is hardly different than "classical gambling." Indeed, when one posits a significance level of .05, one is essentially saying that given many of these results, the long-term frequency at which we will incorrectly reject the null hypothesis is
This is equivalent to saying that given these results, we will bet a 5% chance that we are making a Type I error. Thus, we cannot ever be "sure" when we reject a null hypothesis, we always do so with an associated probability (or gamble) that we are making the wrong decision. Bayesianism gambles no more than classical statistics.

4. Lack of Instruction

A fourth criticism is offered by Sohn (1998) directly in response to those (namely, Cohen, 1994; Falk, 1998; Rozeboom, 1960) advocating Bayesian statistics. Sohn accuses these researchers to have "proclaimed a position on the goal of science, namely to be able to estimate the probability of a hypothesis in the light of data, without making an effort to present the position in a way that has any chance of being understood by the uninitiated" (p. 332). It should be noted that this criticism has little to do with the Bayesian model per se, but rather with how it is promoted for use in psychology. Sohn then goes on to say that important issues raised by the Bayesian model, some of them controversial, "are not discussed by proponents of Bayesianism in psychology....the validity of Bayesian inference is simply proclaimed" (p. 332). Such issues include whether or not the Bayesian model constitutes an ideal means for making inductive inferences.

Sohn's criticism is unfounded. Cohen, Falk, and Rozeboom, among others, namely Savage (1954), Edwards, Lindman and Savage (1963), and Iversen (1984),

\footnote{It should be noted that this is only true for the Neyman-Pearson model of hypothesis-testing, and not for Fisher's original model. See chapter 1 for a fuller discussion of the differences between}
have provided excellent discussions that render a more than adequate "chance of being understood," to quote Sohn. These discussions apply to both psychologists and social scientists alike. Furthermore, Fischhoff and Beyth-Marom (1983) provide an outstanding review of issues centered around Bayesian hypothesis evaluation, and is highly recommended for reading. Hence, Bayesianism is not taken as "self-evidently more reasonable than alternative views" (p. 333), but rather has been propounded by Bayesian supporters as both a theoretically correct and practically useful approach to statistical inference in psychology.

Preliminary Evaluation of the Bayesian Model

The Bayesian model would appear to be superior to null hypothesis testing. It tells us what we want to know (i.e., the probability of the research hypothesis), and for that reason alone, is very appealing. In addition, as noted by Grayson et al. (1997), a Bayesian approach may also be applied to interval estimation, a topic discussed in chapter 3. As noted by some (e.g., Gregson, 1997; Pankoff & Roberts, 1968; Sedlmeier, 1997), and as previously argued, prior information (i.e., prior probabilities) provide meaningful data for inference in clinical settings, and using a Bayesian approach can assist the clinician in decision-making. Classical statistical models are simply not equipped in dealing with prior information as is the Bayesian model.

---

Fisher's model and the model used today, which is essentially based in Neyman-Pearson "decision theory" logic.
So, why aren't we using Bayesian statistics in psychology today? That is a difficult question and I will not attempt to provide an answer here. To entertain this question would certainly call for a complete publication dedicated to detailing the reasoning and social behavior of psychologists, and that is not my goal. I will however launch a discussion of a popular reason for Bayesian dislike by frequentists -- that of estimating prior probabilities.

Prior Probabilities

Without doubt, the most controversial aspect of the Bayesian approach to statistical inference revolves around the notion of prior probabilities. It is controversial because its application requires a basic acceptance and agreement that probability is to be interpreted as subjective belief, a fundamental that frequentists are strictly opposed. Frequentists generally hold that probability is best interpreted as a relative frequency based on theoretical sampling distributions rather than as a degree of belief about a hypothesis. It is this difference in assumptions regarding the nature of probability that clearly makes prior probabilities problematic for the frequentist.

Are frequentists too hasty in their rejection of prior probabilities? Can the frequentist be convinced that the Bayesian prior is a valid way of making explicit one's degree of uncertainty? Are there any grounds for the frequentist to accept the Bayesian prior? Answering these questions is the goal of the following discussion. I will first present arguments in support of the Bayesian prior, in hopes of convincing the frequentist that probability is best viewed as a subjective entity. In short, I will
try to convert the frequentist into a Bayesian. Assuming the frequentist is still
unmoved by these arguments, I will present what I believe to be the Bayesian's
winning argument for the tolerance of prior probability, that of "posterior probability
convergence." It will be demonstrated that given diverse prior opinion, in light of
identical data, posterior probabilities converge to the point of agreement. I will show
that what is most important to the frequentist is indeed what is the product of
Bayesian analysis -- that is, the data in the end, do speak for themselves.

**Defending Bayesian Prior Probability**

Bayesians hold that stating one's prior belief regarding a hypothesis is essential
for the purpose of scientific inference. For the Bayesian advocate, it is only
reasonable to begin with a probability estimate of the hypothesis under
consideration. As argued by Good (1950), the specifying of priors is beneficial in
evaluating a hypothesis:

> When a scientific memoir is concerned with experimental evidence for a
> hypothesis, it is helpful if something is stated about the subjective initial
> probability of the hypothesis. To omit such a statement gives only a superficial
> appearance of objectivity. The uninitiated are liable to be misled into regarding
> the probability as higher than would be claimed by the writer of the memoir (p.
> v).

Good suggests that to omit the reporting of prior probability fosters a miscast
illusion of objectivity. Understanding this point is crucial to understanding why
priors do not constitute the "de-objectifying" of science. To the contrary, they
constitute a more objective science. A central tenet of the Bayesian model is that prior opinion exists whether we quantify it or not. Most frequentists would agree that given any experiment, the investigator is not void of biases in relation to the hypothesis under test. Indeed, if the investigator had no prior opinion of the hypothesis, why would he be testing it in the first place? The point is, is that the researcher carries with him biases regarding his own research, and this fact is undeniable. Given this, it would only seem beneficial to have some way of quantifying this bias instead of leaving it unaccounted for. This is the purpose of prior probability. McClure and Suen (1994) emphasize the inevitability of personal bias in research paradigms:

This bias is there whether one formally employs the Bayesian paradigm or not because researchers use their prior knowledge to formulate research objectives, design experiments, and select analytical methods. Also, other methods of evaluating data are not free from a priori assumptions although they may not be explicitly stated. For example, even in a totally objective study, the researcher implicitly does not consider one event more likely than another; thus all outcomes are of equal probability. This is essentially making a prior assumption that the probability distribution is a uniform one (i.e., a rectangular distribution). The use of the Bayes theorem provides a rational way in which to incorporate the prior knowledge of the researcher into a decision-making process. (p. 95)

McClure and Suen make reference to the implicit use of rectangular prior distributions by frequentists in assessing their degree of uncertainty for a given
hypothesis. However, as evinced by the following discussion, this does not make their prior probability estimate any more "objective" than that made by the Bayesian. Frequentists argue that given no prior distribution, a scientist should assume a rectangular distribution, that is, make the assumption that all values are equally likely. Bayesians remind us however that all the frequentist is doing is stating a prior opinion; that all values have an equal chance of occurring. As argued by Lewis (1993), assuming a rectangular distribution is unrealistic: "Prior beliefs about a mean may be assumed to be uniformly distributed over the entire real line. There is, of course, no such probability distribution: it is merely a convenient fiction" [italics added] (p. 234). Furthermore, as Rozeboom (1960) points out, applying the rectangular distribution to Bayesian prior probability may not extend well to all possible alternatives:

The traditional assumption (made hesitantly by Bayes, less hesitantly by his successors) has been the "principle of insufficient reason," namely, that given no knowledge at all, all alternatives are equally likely. But not only is it difficult to give a convincing argument for this assumption, it does not even yield a unique a priori probability distribution over a continuum of alternative hypotheses, since there are many ways to express such a continuous set, and what is an equilikelihood a priori distribution under one of these does not necessarily transform into the same under another. (p. 427)

Hence, there appears no escaping the notion of prior probability, whether it is approached from a Bayesian or frequentist approach -- prior opinion exists and
should be accounted for. Furthermore, prior opinion should not be limited to a rectangular distribution. As Gigerenzer, Swijtink, Porter, Daston, Beatty, and Krüger (1989) note, Bayes himself made this clear in his ground-breaking essay: "in the case of an event concerning the probability of which we absolutely know nothing antecedently to making any trials concerning it, I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another" (Bayes, 1763, p. 143). In this, Bayes strongly asserted that given no prior distribution, we have no more reason to assume a uniform rectangular distribution than we do to assume a skewed distribution. In other words, "ignorance should not be converted into a uniform distribution" (Gigerenzer et al., 1989, p. 31).

A further concern that prevents the frequentist from accepting prior probabilities is that the postulation of a subjective prior is presumably not in accord with how science should proceed. More specifically, Cosmides and Tooby (1996) argue that by accepting probability as subjective, this jeopardizes a common goal to science which can be said to arrive at conclusions based on methods that most rational scientists would agree upon. They argue that "science strives for intersubjective agreement and consensual methods for arriving at knowledge; to accept Bayes' theorem is to renounce this goal" (p. 5). This is a harsh attack on Bayesianism and their claim is inaccurate. After all, if this were true, would anyone claim to be Bayesian? Most Bayesians would undoubtedly disagree with Cosmides and Tooby's conclusion. The Bayesian would likely argue that consensual methods for arriving at knowledge should include that initial uncertainty be quantified in the form of a prior probability.
To say that the acceptance of the Bayesian axiom necessarily implies renouncing the goal of science is simply incorrect.

What may be at issue with most Bayesian critics is that there are no established methods for estimating prior probabilities, and hence mere unfounded opinion should not be permitted to influence a research result. Such critics would undoubtedly find Jeffrey's (1983) "purist" view of the Bayesian prior surely inadequate: "the numerical probabilities and desirabilities are meant to be subjective in the sense that they reflect the agent's actual beliefs and preferences, irrespective of factual or moral justification" (p. 1). Must this extreme position be taken when estimating priors? Apparently not. Models for establishing priors are being sought and implemented in Bayesian analysis. Some attempts to quantify prior probability deal mainly with comparing one's probability of a hypothesis to that of hypothetically drawing a particular colored ball from a given number of colored balls (Phillips, 1973). Other approaches to estimating prior probabilities have produced somewhat reliable means of estimation. Research by Winkler (1967) resulted in a model for plotting subjects' prior distributions regarding the probability of a particular event. By employing an interview technique, a probability distribution is plotted that reflects one's prior opinion with regards to an event's likelihood. Mathematical models have also been developed for the purpose of measuring prior opinion (see Osherson, Smith, Shafir, Gualtierotti, and Biolsi, 1995).

Posterior Probability Convergence: A Demonstration
In concluding this chapter, I will present what I believe to be the "winning argument" in favor of adopting a Bayesian model in psychology -- that of posterior probability convergence. I will argue that even a frequentist should accept the Bayesian model based on this characteristic. Given identical data (i.e., likelihood ratios), divergent prior probabilities will eventually converge into similar, if not identical posterior probabilities. This behavior of Bayesian posteriors has been noted before. Falk (1998), (C. Green?) McClure and Suen (1994), and Phillips (1974) are among those noting "posterior probability convergence." As Phillips explains:

Where more than one observation is possible we will find that as more and more data are collected, initially divergent opinion comes more into agreement....It is this feature of Bayes's theorem that save Bayesian statistics from being wholly subjective. Initially subjective opinion is brought into contact with data through the operation of Bayes' theorem, and with enough data differing prior opinions are made to converge. This comes about because the prior opinions become less and less relevant to posterior opinion as more and more data are observed. Prior opinion is swamped out by the data, so that posterior opinion is controlled solely by the data. [italics added] For a Bayesian, this is the only way in which data can 'speak for themselves'. (pp. 76, 78)

To demonstrate this effect, consider the following hypothetical example. Two scientists debate whether a given hypothesis H is true or false. Scientist A assigns H a prior probability of 0.3 of being true, while scientist B assigns 0.8 as a prior
probability that H is true. Subsequent probability revisions are represented in Table 2.

Assume now that the likelihood, which is usually an agreed upon item (see Phillips, p. 78 for an explanation), is 0.50 for "H is True," and 0.30 for "H is not True." Given this data, consider one round of hypothesis revision, as represented by "Revision 1." From the table, it can be seen that Scientist A's initial prior, that of 0.30, in light of the data, increased to a posterior probability of 0.42. For scientist B, his initial prior probability of 0.80 increased to 0.87. The initial discrepancy between scientists' priors was 0.50 (i.e., scientist B's prior, 0.80 - scientist A's prior, 0.30) before the data. Given one round of the data however, the difference between probabilities is now 0.45 (i.e., scientist B's posterior, 0.87 - scientist A's posterior, 0.42). Thus, as a result of the same data, posterior probabilities are now more similar than they initially were as prior probabilities. The probabilities are beginning to converge. Will this convergence continue over many rounds of data? Using the latest posteriors as priors, and keeping the likelihoods constant, "Revision 2" displays the results of a 2nd round of hypothesis revision.
Table 2

Bayesian Probability Revisions of Scientist A and Scientist B for Hypothesis H.

<table>
<thead>
<tr>
<th>Prob. Revision</th>
<th>Prior Probability</th>
<th>Likelihood Ratio</th>
<th>Prior x Likelihood</th>
<th>Posterior Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scientist A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H is True</td>
<td>0.30</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>H is not True</td>
<td>0.70</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Initial Prob.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientist B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H is True</td>
<td>0.80</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>H is not True</td>
<td>0.20</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Revision 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientist B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H is True</td>
<td>0.80</td>
<td>0.50</td>
<td>0.40</td>
<td>0.86956</td>
</tr>
<tr>
<td>H is not True</td>
<td>0.20</td>
<td>0.30</td>
<td>0.06</td>
<td>0.13043</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientist A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H is True</td>
<td>0.41666</td>
<td>0.50</td>
<td>0.20833</td>
<td>0.54349</td>
</tr>
<tr>
<td>H is not True</td>
<td>0.58333</td>
<td>0.30</td>
<td>0.17499</td>
<td>0.45651</td>
</tr>
<tr>
<td>Revision 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientist B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H is true</td>
<td>0.86956</td>
<td>0.50</td>
<td>0.43478</td>
<td>0.91743</td>
</tr>
<tr>
<td>H is not true</td>
<td>0.13043</td>
<td>0.30</td>
<td>0.03913</td>
<td>0.08257</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientist A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H is true</td>
<td>0.54349</td>
<td>0.50</td>
<td>0.27175</td>
<td>0.66491</td>
</tr>
<tr>
<td>H is not true</td>
<td>0.45651</td>
<td>0.30</td>
<td>0.13695</td>
<td>0.33509</td>
</tr>
<tr>
<td>Revision 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientist B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H is true</td>
<td>0.91743</td>
<td>0.50</td>
<td>0.45871</td>
<td>0.94877</td>
</tr>
<tr>
<td>H is not true</td>
<td>0.08257</td>
<td>0.30</td>
<td>0.02477</td>
<td>0.05123</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2 (con't)
Bayesian Probability Revisions of Scientist A and Scientist B for Hypothesis H.

<table>
<thead>
<tr>
<th>Prob. Revision</th>
<th>Scientist A</th>
<th></th>
<th>Scientist B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prior Probability</td>
<td>Likelihood Ratio</td>
<td>Prior x Likelihood</td>
<td>Posterior Probability</td>
</tr>
<tr>
<td><strong>H is true</strong></td>
<td>0.66491</td>
<td>0.50</td>
<td>0.33246</td>
<td>0.77055</td>
</tr>
<tr>
<td><strong>H is not true</strong></td>
<td>0.33509</td>
<td>0.30</td>
<td>0.09900</td>
<td>0.22945</td>
</tr>
<tr>
<td><strong>Revision 4</strong></td>
<td></td>
<td></td>
<td></td>
<td>Posterior Differential = 0.19807</td>
</tr>
<tr>
<td><strong>H is true</strong></td>
<td>0.94877</td>
<td>0.50</td>
<td>0.47439</td>
<td>0.96862</td>
</tr>
<tr>
<td><strong>H is not true</strong></td>
<td>0.05123</td>
<td>0.30</td>
<td>0.01537</td>
<td>0.03138</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H is true</strong></td>
<td>0.77055</td>
<td>0.50</td>
<td>0.38775</td>
<td>0.84923</td>
</tr>
<tr>
<td><strong>H is not true</strong></td>
<td>0.22945</td>
<td>0.30</td>
<td>0.06884</td>
<td>0.15077</td>
</tr>
<tr>
<td><strong>Revision 5</strong></td>
<td></td>
<td></td>
<td></td>
<td>Posterior Differential = 0.13171</td>
</tr>
<tr>
<td><strong>H is true</strong></td>
<td>0.84923</td>
<td>(0.50)^5</td>
<td>0.02654</td>
<td>0.98625</td>
</tr>
<tr>
<td><strong>H is not true</strong></td>
<td>0.15077</td>
<td>(0.30)^5</td>
<td>0.00037</td>
<td>0.01375</td>
</tr>
<tr>
<td><strong>Revision 6-10</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H is true</strong></td>
<td>0.98094</td>
<td>(0.50)^5</td>
<td>0.03065</td>
<td>0.99837</td>
</tr>
<tr>
<td><strong>H is not true</strong></td>
<td>0.01906</td>
<td>(0.30)^5</td>
<td>0.00005</td>
<td>0.00163</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Posterior Differential = 0.01212</td>
</tr>
<tr>
<td><strong>H is true</strong></td>
<td>0.98625</td>
<td>(0.50)^5</td>
<td>0.03082</td>
<td>0.99903</td>
</tr>
<tr>
<td><strong>H is not true</strong></td>
<td>0.01375</td>
<td>(0.30)^5</td>
<td>0.00003</td>
<td>0.00097</td>
</tr>
<tr>
<td><strong>Revision 11-15</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>H is true</strong></td>
<td>0.99837</td>
<td>(0.50)^5</td>
<td>0.03120</td>
<td>0.99987</td>
</tr>
<tr>
<td><strong>H is not true</strong></td>
<td>0.00163</td>
<td>(0.30)^5</td>
<td>0.000004</td>
<td>0.00013</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Posterior Differential = 0.00084</td>
</tr>
</tbody>
</table>
Again, as can be seen from the table, posteriors are now closer in proximity than they were after the 1st round of data. Scientist A's prior of 0.42 increased to 0.54, while scientist B's prior of 0.87 increased to 0.92. The difference in posteriors is now 0.37, down from the difference of 0.45 after the 1st round of data. Hence, the probabilities continue to converge. Revisions 3-5 show the results of 3 more consecutive rounds of the same data. As can be seen, the difference between scientist A's posterior and scientist B's posterior probability continues to decrease. Indeed, after a complete 5 rounds of data, the difference between posteriors is only 0.13. Revisions 6-10 show the results of rounds of data 6 through 10. By the end of the 10th round, the posterior probabilities of scientist A and scientist B are different by only 0.012. Revisions 11-15 show the results of rounds of data 11 through 15 which result in posteriors that are practically indistinguishable from each other. This "probability convergence" is best seen by viewing Figure 5. Indeed, as data become available, initially divergent prior probabilities have a tendency to converge to yield identical posterior probabilities.

As has been demonstrated, differing prior probabilities, given the same data, will converge. In the above example, the posteriors will necessarily converge to practically 1.0 because each round of revision constitutes "positive" data. That is, for each round of data, the likelihood for H is consistently larger than the likelihood for \sim H. What determines the rate at which they converge to 1.0 depends on the difference between the likelihoods. For instance, referring back to equation (1), if \( p ( D|H ) \) is much greater than \( p ( D|\sim H ) \), the rate at which they will reach 1.0 will be faster than if \( p ( D|H ) \) was only slightly greater than \( p ( D|\sim H ) \). Contrarily, if
\( p(D|H) \) was less than \( p(D|\sim H) \), then each scientist's probability would move away from 1.0 and slowly decrease.

---

**Figure 5** -- Bayesian posterior probability convergence (Scientist A and Scientist B)
The important point to note is that given enough data, prior probabilities are not significant in determining posterior probabilities. Regardless of priors, posteriors will reflect the data, not prior opinion. This constitutes the winning argument for Bayesianism. I argue that whether you consider yourself a frequentist or a Bayesian, prior probabilities do not constitute a problem. The Bayesian model is a sound, logical approach to statistical inference, and given that posteriors reflect the data, the Bayesian model should be appealing to the frequentist as it is to the Bayesian. Indeed, what the frequentist has always stressed is true of the Bayesian model -- in the end, the data speak for themselves.

Summary and Conclusion

The goal of this chapter was to present the Bayesian model as an alternative to traditional null hypothesis significance testing. An additional goal was to argue for its application and logical compatibility for use in psychology. Through pointing out its advantages, and subduing its disadvantages, I believe these two goals have been met. Furthermore, by demonstrating the posterior probability convergence argument, I hope to have shown that Bayesianism is incorrectly faulted for being overly subjective. I have argued that rejecting the Bayesian model on this facet alone is incorrect. The fact that posterior probabilities converge calls on the frequentist who usually dismisses the Bayesian model on account of its subjectivist philosophy, to reconsider. The Bayesian model represents a logically sound approach to statistical inference, and indeed stands as a successful competing alternative to null hypothesis significance testing.
Conclusion

The primary goal of this thesis was to answer two questions: 1) what are the problems associated with NHST and are they due primarily to inherent difficulties of the model, or due to misunderstanding and misuse, and 2) given the problems that accompany the use of NHST, can any other model successfully replace it, and thereby overcome these problems? Answering the first question is a necessary prerequisite to answering the second. That is, if it is determined that the problems associated with NHST are indeed due more to misuse than to fundamental difficulties with the model, then perhaps what is necessary is better education and attention with regards to its use. As concluded in chapter 2, a good model should not be discarded based solely on its level of complexity. However, if the answer to the first question is that NHST is a poor model, with inherently severe difficulties beyond simple misunderstanding, then it naturally follows to review and evaluate various other models for a comparison of how they stack up to NHST. As was stated at the outset of this thesis, most if not all models of inference will be accompanied by at least some methodological and statistical difficulties. To expect otherwise would be unrealistic. The aim should be to choose a model that is well-suited for psychology, and at the same time minimizes methodological difficulties. Hence, the overall goal of this thesis was through an analysis of NHST and alternatives, to deduce the "best" of available inferential models for use in psychology.
Providing a uniform answer to the first question is difficult. As discussed in chapter 2, NHST is methodologically flawed and is a victim of misuse. Chapter 2 addressed whether problems associated with NHST are due primarily to inherent difficulties, or due to misuse. The answer is that the problems can be attributed approximately equally to both the model and to misuse. However, the problems associated with the model are much more severe than those associated with misuse. Misuse issues can theoretically be dealt with, in that by bringing more awareness to these misunderstandings, researchers may eventually interpret NHST results correctly. For instance, the problem of sample-size-sensitivity is held as a "misuse" issue, primarily because given power analysis, the problem all but disappears. The "effects/non-effects dichotomy" problem can also be attributed to researcher ignorance and misuse. A conscientious researcher should know not to conclude effects based on p values alone, regardless if they reach p < .05 or lower. If they do not know this, then future generations of researchers need to be taught these basic principles. Finally, the fact that statistically insignificant results may be significant is also attributable to model misuse and misinterpretation. Again, one cannot fault a model based on the user drawing incorrect conclusions.

What is true regarding problems of "misuse" is not true with regards to "model" issues. That is, although problems of misuse can theoretically be rectified, problems centered within the model are difficult to solve, primarily because they are "hardwired" into the model. For instance, that NHST tests only a null hypothesis and not the research hypothesis is hardly "fixable" if one is to keep using NHST. It is
doubtful that one could devise an improvement for NHST that would have the research hypothesis tested directly. Such attempts would likely prove futile, simply because NHST is, and never was, construed to test the research hypothesis. This is an inherent component of the model, and if this component is problematic, then it follows that the difficulty is inherent and hardly amenable to change. One cannot "make" NHST test the research hypothesis, it simply can't. Likewise is true with regards to the "arbitrary alpha" problem. As long as one uses NHST, one must choose a level at which to set alpha. There is no getting away from this, and if this fact is problematic, then it counts again as an inherent "strike" against NHST. Simply said, one cannot alter fundamentals on which a model is based, and still hope to use that model. NHST is NHST, and if we cannot tolerate that it does not test the research hypothesis, that it requires the setting of an arbitrary alpha, and that the null hypothesis is likely always priori false, then indeed we will have difficulty in accepting the model. In sum, we can work with "misuse," but we can hardly work with the fundamental roots of the model without jeopardizing it entirely.

Having said this, what shall we conclude with respect to the "model-misuse" distinction? A reasonable conclusion would be to say that although the misuse problems may\(^1\) be dealt with successfully, they are hardly worth addressing given the inherent and unalterable problems of the model. In other words, the inherent problems associated with NHST are severe enough to overshadow mere

---

\(^1\) I emphasize "may" for the reason that given almost 75 years of methodological instruction, NHST is still misused. Indeed, this fact casts some doubt as to whether NHST could ever be used correctly. In the present discussion, I make the assumption that it can.
misunderstanding. Using NHST correctly does not save NHST. By all means, if one is committed to using NHST, then misuse issues need to be addressed. However, solving misuse issues will not solve NHST's bigger problems, and unfortunately these latter ones are much more serious, and are integrated into the very nature of the model. The same way it is illogical to train (or re-train) a pilot on how to operate an aircraft if the design of its engines is faulty, it is illogical to re-train a researcher on using NHST if its bigger problems will cause the model to "crash" anyway.

Given the most severe problems associated with NHST, (i.e., those inherently rooted in the model), we require a means of appropriately evaluating their degree of severity. As mentioned in the conclusion of chapter 2, it is difficult to fully appraise and appreciate the problems of NHST in "absolute" terms. That is, if we are interested in using the "best" of alternatives, problems associated with NHST should not be interpreted in a vacuum. Rather, they should be compared to difficulties that are a part of other models of inference. Such a comparison will allow one to correctly conclude the "best" of alternatives, while minimizing potential drawbacks and difficulties. As concluded from chapter 3, there exist many alternatives to null hypothesis testing, and hence answering the question of why NHST is still currently used suggests reasons other than ones purely methodological in nature. Do these alternatives solve the "severe" (i.e., those inherent to the model) problems associated with NHST? Do they introduce their own problems? These are questions that must be addressed when appraising alternatives. In chapter 3, it was concluded that some
of the reviewed alternatives best serve as "auxiliaries" to NHST; that is, they are best employed as complements rather than whole replacements. For instance, effect size measures are best employed for use in conjunction with NHST. Power analysis cannot even be considered a true alternative, but rather makes NHST more tolerable by providing a means for calculating sufficient sample size. Confidence intervals on the other hand, do constitute an ideal alternative to null hypothesis testing largely because they are able to estimate the degree to which a particular value is the true value. An NHST analysis alone does not reveal this information. Confidence intervals also offer more than merely rejecting an implausible hypothesis -- they offer a direct estimation of the mean, and provide a margin of error. In addition, one can still perform hypothesis testing within the context of confidence interval estimation (Kirk, 1996). This last advantage makes using NHST seem trivial if we can use a procedure that can yield the same and more information. Loftus' PPE procedure and Serlin and Lapsley's good-enough principle are also highly qualified in replacing NHST. Although still suffering drawbacks of their own, they constitute major advancements over simply rejecting a null hypothesis -- that is what makes these alternatives attractive; they accomplish the basics of NHST, and at the same time provide much more information for the researcher to use in the data-analytic process.

The above mentioned alternatives constitute advancements over NHST, that much is clear. However, as already argued, all alternatives must be compared to other competing alternatives. To restate, a model is "best" if it triumphs over other
models. In chapter 3, the goal was to evaluate alternatives to NHST within a frequentist paradigm. As seen, confidence intervals, the PPE procedure, and the good-enough principle are advantageous to NHST and could feasibly replace it. However, a fourth alternative, that of the Bayesian model, is even more favorable. The Bayesian model constitutes the ideal alternative to NHST. What is so attractive about Bayesianism is its ability to directly test the research hypothesis. The Bayesian model tests what NHST can only infer. Through the typical analysis, the researcher can directly quantify the probability of the hypothesis under test, rather than inferring it from a rejection of a null hypothesis. As demonstrated in chapter 4, what is most objectionable to the frequentist, that of stating prior probabilities, is solved by the fact that posterior probabilities converge given initially discrepant opinions. At the heart of Bayesianism, is the likelihood ratio, a ratio that is stable and does not change with regards to the scientist's prior opinion. In the end, no matter what the prior probabilities were, the likelihood ratio behaves as the major influence in determining the posterior probability. Compared to NHST, the Bayesian model is the more logical and intuitive approach to statistical inference, and is ideally the model of choice for psychology.

What then, shall be the "final word" of this thesis? Is it that NHST is a "poison" for psychology and should be avoided at all costs? Is it that Bayesianism represents the "saving grace" for psychology and will forever solve the discipline's methodological difficulties? Certainly not. The present thesis has focused on the use of statistical models for the purpose of inference. I have tried to offer an unbiased
evaluation and appraisal of our currently used model, NHST, and other competing alternatives that may be employed in its place. I have implicitly assumed that psychology requires a statistical model, and have set out by trying to conclude which of the many alternatives is best for psychology. Based on the evidence, I have concluded the Bayesian model to be the "best" of available models for statistical inference in psychology.

At this point, something needs to be made explicit. Although I've examined many models, it should be stressed that a statistical result should not in any way imply an "end" to what psychology is most interested in -- theory confirmation. A statistical model, and hence the statistical result, should be but one "tool" used in this process. In other words, although a statistical model may prove extremely helpful in reaching decisions regarding theories, it should not by itself provide the decision. Theory corroboration is a complex process, and more than anything else, requires the insight and judgment of the researcher. Assuming that psychology's major purpose is to evaluate theories regarding psychological phenomena, it must be stressed that statistical inference is but one step in this process. A statistical result should not exhaust the psychologist's range of flexibility in evaluating theories. Such a philosophy leads to the current state of the editorial process, that of deeming a paper publishable only if it meets an arbitrary criterion (i.e., that of $p < .05$). Thus, in the end, the data may speak for themselves, but this end should be regarded as preliminary in the world of theory evaluation. The job is not yet finished. The true "end" is where the researcher, upon evaluating many things, a statistical result being
only one of them, comes to an educated conclusion regarding his research problem. Thus in sum, although I personally advocate the use of the Bayesian model in psychology, my final word is that a Bayesian posterior probability should only be one tool among many in arriving at a decision with regards to a psychological theory. In this, I advocate a similar viewpoint to that recently expressed by Wilkinson and the Task Force on Statistical Inference (1999), "Statistical methods should guide and discipline our thinking but should not determine it" (p. 603).
References


Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161-171.


Sohn, D. (1993). Psychology of the scientist: LXVI. The idiots savants have taken over the psychology labs! Or why in science using the rejection of the null hypothesis as the basis for affirming the research hypothesis is unwarranted. *Psychological Reports, 73*, 1167-1175.


