A HANDBOOK FOR DATA ANALYSIS IN THE BEHAVIORAL SCIENCES: Methodological Issues

Edited by

Gideon Keren

Free University of Amsterdam

Charles Lewis

Educational Testing Service

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS

1993 Hillsdale, New Jersey Hove & London

11 The Superego, the Ego, and the Id in Statistical Reasoning

Gerd Gigerenzer
University of Chicago

Piaget worked out his logical theory of cognitive development, Köhler the Gestalt laws of perception, Pavlov the principles of classical conditioning, Skinner those of operant conditioning, and Bartlett his theory of remembering and schemata—all without rejecting null hypotheses. But, by the time I took my first course in psychology at the University of Munich in 1969, null hypothesis tests were presented as *the* indispensable tool, as the sine qua non of scientific research. Post–World War 2 German psychology mimicked a revolution of research practice that had occurred between 1940 and 1955 in American psychology.

What I learned in my courses and textbooks about the logic of scientific inference was not without a touch of morality, a scientific version of the 10 commandments: Thou shalt not draw inferences from a nonsignificant result. Thou shalt always specify the level of significance before the experiment; those who specify it afterward (by rounding up obtained p values) are cheating. Thou shalt always design thy experiments so that thou canst perform significance testing.

THE INFERENCE REVOLUTION

What happened between the time of Piaget, Köhler, Pavlov, Skinner, and Bartlett and the time I was trained? In Kendall's (1942) words, statisticians "have already overrun every branch of science with a rapidity of conquest rivalled only by Attila, Mohammed, and the Colorado beetle" (p. 69).

What has been termed the *probabilistic revolution in science* (Gigerenzer et

al., 1989; Krüger, Daston, & Heidelberger, 1987; Krüger, Gigerenzer, & Morgan, 1987) reveals how profoundly our understanding of nature changed when concepts such as chance and probability were introduced as fundamental theoretical concepts. The work of Mendel in genetics, that of Maxwell and Boltzmann on statistical mechanics, and the quantum mechanics of Schrödinger and Heisenberg that built indeterminism into its very model of nature are key examples of that revolution in thought.

Psychology did not resist the probabilistic revolution, and psychologists in turn actively contributed to the growth of statistics. But psychology is nonetheless a peculiar case. In psychology and in other social sciences, probability and statistics were typically not used to revise the understanding of our *subject matter* from a deterministic to some probabilistic view (as in physics, genetics, or evolutionary biology), but rather to mechanize the *experimenters'* inferences—in particular, their inferences from data to hypothesis. Of course, there have been several attempts to revise our theories as well—for example, to transform Piaget's logical determinism into a more Darwinian view, where variability and irregularity are seen as the motor of evolution rather than as an annoyance (Gruber, 1977; Gruber & Vonèche, 1977), or to transform Skinner's theory into a probabilistic learning theory (Estes, 1959). But the real, enduring transformation came with statistical inference, which became institutionalized and used in a dogmatic and mechanized way. This use of statistical theory contrasts sharply with physics, where statistics and probability are indispensable in theories about nature, whereas mechanized statistical inference such as null hypothesis testing is almost unknown.

So what happened with psychology? David Murray and I described the striking change in research practice and named it the *inference revolution* in psychology (Gigerenzer & Murray, 1987). It happened between approximately 1940 and 1955 in the United States, and led to the institutionalization of one brand of inferential statistics as *the* method of scientific inference in university curricula, textbooks, and the editorials of major journals.[1]

---

[1]The ground for the inference revolution was prepared by a dramatic shift in experimental practice. During the 1920s, 1930s, and 1940s, the established tradition of experimenting with *single* subjects—from Wundt to Pavlov—was replaced in the United States by the *treatment group experiment*, in which group means are compared. For instance, between 1915 and 1950, the percentage of empirical studies reporting only group data in the *American Journal of Psychology* rose from 25% to 80%, and the reporting of only individual data decreased from 70% to 17% (Danziger, 1990). Danziger argued that this shift was in part due to the pressure felt by United States academic psychologists to legitimize their work through showing its practical utility. The Wundtian type of experiment was useless to educational administrators, the largest market for psychological products. The treatment group experiment, however, appeared to fit their needs exactly, for example, by allowing them to compare mean performance in two classrooms that were using different instruction methods. After this change in experimental practice, null hypothesis testing of group means appeared to be tailor-made to the new unit of research, the group aggregate. Consistent with Danziger's argument, the institutionalization of both the treatment group and null hypothesis testing spread from

The figures are telling. Before 1940, null hypothesis testing using analysis of variance or *t* test was practically nonexistent: Rucci and Tweney (1980) found only 17 articles in all from 1934 through 1940. By 1955, more than 80% of the empirical articles in four leading journals used null hypothesis testing (Sterling, 1959). Today, the figure is close to 100%. By the early 1950s, half of the psychology departments in leading U.S. universities had made inferential statistics a graduate program requirement (Rucci & Tweney, 1980). Editors and experimenters began to measure the quality of research by the level of significance obtained. For instance, in 1962, the editor of the *Journal of Experimental Psychology,* A. W. Melton (1962, pp. 553–554), stated his criteria for accepting articles. In brief, if the null hypothesis was rejected at the .05 level but not at the .01 level, there was a "strong reluctance" to publish the results, whereas findings significant at the .01 level deserved a place in the journal. The *Publication Manual of the American Psychological Association* (1974) prescribed how to report the results of significance tests (but did not mention other statistical methods), and used, as Melton did, the label *negative* results synonymously with "not having rejected the null" and the label *positive* results with "having rejected the null."

It is likely that Piaget's, Köhler's, Bartlett's, Pavlov's, and Skinner's experimental work would have been rejected under such editorial policies—these men did not set up null hypotheses and try to refute them. Some of them were actively hostile toward institutionalized statistics. For his part, Skinner (1972) disliked the intimate link Fisher established between statistics and the design of experiments: "What the statistician means by the design of experiments is design which yields the kind of data to which his techniques are applicable" (p. 122). And, "They have taught statistics in lieu of scientific method" (p. 319). Skinner continued to investigate one or a few pigeons under well-controlled conditions, rather than run 20 or more pigeons under necessarily less well-controlled conditions to obtain a precise estimate for the error variance. In fact, the Skinnerians were forced to found a new journal, the *Journal of the Experimental Analysis of Behavior,* in order to publish their kind of experiments (Skinner, 1984, p. 138). Their focus was on experimental control, that is, on minimizing error beforehand, rather than on large samples, that is, on measuring error after the fact.

This is not an isolated case, nor one peculiar to behaviorists. The *Journal of Mathematical Psychology* is another. One of the reasons for launching this new

---

the applied fields to the laboratories (Lovie, 1979). The contrast with Germany is telling. German academic psychologists of the early 20th century had to legitimize their work before a different tribunal, the values of a well-entrenched intellectual elite (Danziger, 1990). In contrast to the United States, the German educational system, run by tradition rather than by experimentation, provided only a limited market for psychologists. No comparable shift in experimental practice happened in German psychology. It was only after World War II that a new generation of German psychologists began to assimilate the methodological imperatives imported from their colleagues in the United States.

journal was again to escape the editors' pressure to perform institutionalized null hypothesis testing.[2] One of its founders, Luce (1988), called the institutionalized practice a "wrongheaded view about what constituted scientific progress" and "mindless hypothesis testing in lieu of doing good research: measuring effects, constructing substantive theories of some depth, and developing probability models and statistical procedures suited to these theories" (p. 582).

Who is to blame for the present state of mindless hypothesis testing? Fisher was blamed by Skinner, as well as by Meehl: "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 . . . is . . . one of the worst things [that] ever happened in the history of psychology" (Meehl, 1978, p. 817).

I share the sentiments expressed by Luce and Meehl. But to blame Fisher, as Meehl and Skinner did, gives us at best a spurious understanding of the inference revolution. Fisher declared that a significance test of a null hypothesis is only a "weak" argument. That is, it is applicable only in those cases where we have very little knowledge or none at all. For Fisher, significance testing was the most primitive type of argument in a hierarchy of possible statistical analyses (see Gigerenzer et al., 1989, chap. 3). In this chapter I argue the following points:

1. What has become institutionalized as *inferential statistics* in psychology is not Fisherian statistics. It is an incoherent mishmash of some of Fisher's ideas on one hand, and some of the ideas of Neyman and E. S. Pearson on the other. I refer to this blend as the "hybrid logic" of statistical inference. Fisher, Neyman, and Pearson would all have rejected it, although for different reasons.

2. The institutionalized hybrid carries the message that *statistics is statistics is statistics*, that is, that statistics is a single integrated structure that speaks with a single authoritative voice. This entails the claim that the problem of inductive inference in fact *has* an algorithmic answer (i.e., the hybrid logic) that works for all contents and contexts. Both claims are wrong, and it is time to go beyond this institutionalized illusion. We must write new textbooks and change editorial practices. Students and researchers should be exposed to different approaches (not one) to inductive inference, and be trained to use these in a constructive (not mechanical) way. A free market of several good ideas is better than a state monopoly for a single confused idea.

3. Statistical tools tend to turn into theories of mind. We can find the dogma "statistics is statistics is statistics" reappearing in one of the most interesting research areas in cognitive psychology: intuitive statistics and judgments under uncertainty. One statistical theory is confused with rational inductive inference per se.

---

[2]R. Duncan Luce, personal communication, April 4, 1990. See also Luce's (1989) autobiography, on p. 270 and pp. 281–282.

THE "PARENTS" AND THEIR CONFLICTS

In order to understand the structure of the hybrid logic that has been taught in psychology for some 40 years, I briefly sketch those ideas of Fisher, on the one hand, and Neyman and Pearson on the other, that are relevant to understanding the hybrid structure of the logic of inference.

Fisher's first book, *Statistical Methods for Research Workers*, published in 1925, was successful in introducing biologists and agronomists to the new techniques. It had the agricultural smell of issues like the weight of pigs and the effect of manure, and, such alien topics aside, it was technically far too difficult to be understood by most psychologists.

Fisher's second statistical book, *The Design of Experiments*, first published in 1935, was most influential on psychology. At the very beginning of this book, Fisher rejected the theory of inverse probability (Bayesian theory) and congratulated the Reverend Bayes for having been so critical of his own theory as to withhold it from publication (Bayes' treatise was published posthumously in 1763). Bayes' theorem is attractive for researchers because it allows one to calculate the probability $p(H|D)$ of a hypothesis H given some data D, also known as *inverse probability*. A frequentist theory, such as Fisher's null hypothesis testing or Neyman–Pearson theory, however, does not. It deals with the probabilities $p(D|H)$ of some data D given a hypothesis H, such as the level of significance.

Fisher was not satisfied with an approach to inductive inference based on Bayes' theorem. The use of Bayes' theorem presupposes that a prior probability distribution over the set of possible hypotheses is available. For a frequentist, such as Fisher, this prior distribution must theoretically be verifiable by actual frequencies, that is, by sampling from its reference set. These cases are rare. But if we are ignorant and have no a priori distributional information, then every researcher can express that ignorance in different numbers leading, for Fisher, to an unacceptable subjectivism. As we shall see, however, Fisher wanted to both reject the Bayesian cake and eat it, too.

Fisher proposed several alternative tools for inductive inference. In *The Design of Experiments*, he started with *null hypothesis testing*, also known as *significance testing*, and he gave that tool the most space in his book. It eventually became the backbone of institutionalized statistics in psychology. In a test of significance, one confronts a null hypothesis with observations, to find out whether the observations deviate far enough from the null hypothesis to conclude that the null is implausible. The specific techniques of null hypothesis testing, such as the t test (devised by Gossett, using the pseudonym "Student", in 1908) or the F test (F for Fisher, e.g., in analysis of variance) are so widely used that they may be the lowest common denominator of what psychologists today do and know.

The topic of this chapter is the *logic* of inference rather than specific tech-

niques. Just as with Bayes' theorem, the problems we encounter do not concern the formula—the theorem is a simple consequence of the definition of conditional probability. The problems arise with its application to inductive inference in science. To what aspect of inductive inference does a particular algorithm, or technique, refer? What do the calculations mean? These are questions that pertain to what I call the *logic* of inference.

Concerning my account of Fisher's logic of significance testing, one thing must be said in advance: Fisher's writings and polemics had a remarkably elusive quality, and people have read his work quite differently. During Fisher's long and acrimonious controversy with Neyman and Pearson, which lasted from the 1930s to his death in 1962, he changed, and sometimes even reversed, parts of his logic of inference. Thus, the following brief account of Fisher's logic of inference represents one possible reading (for a more detailed analysis, see Gigerenzer et al., 1989, chap. 3).

How Do We Determine the Level of Significance?

In the *Design*, Fisher suggested that we think of the level of significance as a *convention:* "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 all results which fail to reach this standard" (1935/1951, p. 13). Fisher's assertion that 5% (in some cases, 1%) is a *convention* that is adopted by all experimenters and in all experiments, and nonsignificant results are to be ignored, became part of the institutionalized hybrid logic.

But Fisher had second thoughts, which he stated most clearly in the mid-1950s. These did not become part of the hybrid logic. One of the reasons for that revision was his controversy with Neyman and Pearson, and Neyman's (e.g., 1950) insistence that one has to specify the level of significance (which is denoted as α in Neyman–Pearson theory) *before* the experiment, in order to be able to interpret it as a long-run frequency of error. Neyman and Pearson took the frequentist position more seriously than Fisher. They argued that the meaning of a level of significance such as 5% is the following: If the null hypothesis is correct, and the experiment is repeated many times, then the experimenter will wrongly reject the null in 5% of the cases. To reject the null if it is correct is called an *error of the first kind* (Type I error) in Neyman–Pearson theory, and its probability is called *alpha* (α). In his last book, *Statistical Methods and Scientific Inference* (1956), Fisher ridiculed this definition as "absurdly academic, for in fact 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). Fisher rejected the Neyman–Pearson logic of repeated experiments (repeated random sampling from the same population), and thereby rejected his earlier proposal to have a conventional standard level of significance, such as .05 or .01. What

researchers should do, according to Fisher's second thoughts, is to publish the *exact level of significance*, say, $p = .03$ (not $p < .05$), and communicate this result to their fellow research workers. This means that the level of significance is determined *after* the experiment, not, as Neyman and Pearson proposed, *before* the experiment.

Thus the phrase "level of significance" has three meanings: (a) the *standard level of significance*, a conventional standard for all researchers (early Fisher), (b) the *exact level of significance*, a communication to research fellows, determined after the experiment (late Fisher), and (c) the *alpha level, the relative frequency of Type I errors in the long run*, to be decided on using cost-benefit considerations *before* the experiment (Neyman & Pearson). The basic difference is this: For Fisher, the exact level of significance is a property of the data (i.e., a relation between a body of data and a theory); for Neyman and Pearson, alpha is a property of the test, not of the data. Level of significance and alpha are not the same thing.

Neyman and Pearson thought their straightforward long-run frequentist interpretation of the significance test—and the associated concepts of power and of stating two statistical hypotheses (rather than only one, the null)—would be an improvement on Fisher's theory and make it more consistent. Fisher disagreed. Whereas Neyman and Pearson thought of mathematical and conceptual consistency, Fisher thought of ideological differences. He accused Neyman, Pearson, and their followers of confusing technology with knowledge: Their focus on Type I and Type II errors, on cost-benefit considerations that determine the balance between the two, and on repeated sampling from the same population has little to do with scientific practice, but it is characteristic for quality control and acceptance procedures in manufacturing. Fisher (1955, p. 70) compared the Neyman-Pearsonians to the Soviets, their 5-year plans, and their ideal that "pure science can and should be geared to technological performance." He also compared them to Americans, who confuse the process of gaining knowledge with speeding up production or saving money. (Incidentally, Neyman was born in Russia, and went to Berkeley, CA, after Fisher made it difficult for him to stay on at University College in London).

What Does a Significant Result Mean?

The basic differences are these: Fisher attached an epistemic interpretation to a significant result, which referred to a particular experiment. Neyman rejected this view as inconsistent and attached a behavioral meaning to a significant result that did not refer to a particular experiment, but to repeated experiments. (Pearson found himself somewhere in between.)

In the *Design*, Fisher talked about how "to disprove" a null hypothesis (e.g., pp. 16–17). Whatever the words he used, he always held that a significant result affects our confidence or degree of belief that the null hypothesis is false. This is

what I refer to as an *epistemic interpretation:* Significance tells us about the truth or falsehood of a particular hypothesis in a particular experiment. Here we see very clearly Fisher's quasi-Bayesian view that the exact level of significance somehow measures the confidence we should have that the null hypothesis is false. But from a more consistent frequentist viewpoint, as expressed by Neyman, a level of significance does not tell us anything about the truth of a particular hypothesis; it states the relative frequency of Type I errors in the long run.

Neyman (1957) called his frequentist interpretation *behavioristic:* To accept or reject a hypothesis is a decision to take a particular action. Imagine a typical application of Neyman–Pearson theory: quality control. Imagine you have chosen the probability of Type I errors (false alarms) to be .10 and that of Type II errors (misses) to be .01, because misses are much more costly to your firm than false alarms. Every day you take a random sample from the firm's production. Even if the production is normal, you will expect a significant result (false alarm) in 10% of all days. Therefore, if a significant result occurs, you will act as if the null hypothesis were false, that is, stop the production and check for a malfunction; but you will not necessarily believe that it is false—because you expect a lot of false alarms in the long run.

Fisher rejected Neyman's arguments for "inductive behavior" as "childish" (1955, p. 75), stemming from "mathematicians without personal contact with the Natural Sciences" (p. 69). And he maintained his epistemic view: "From a test of significance . . . we have a genuine measure of the confidence with which any particular opinion may be held, in view of our particular data" (p. 74). For all his anti-Bayesian talk, Fisher adopted a very similar-sounding line of argument (Johnstone, 1987).

Does "Significant" Imply that There Is a Causal Effect?

Of course not. It is useful to distinguish between the *statistical* null hypothesis and the *substantive* null hypothesis.[3] Only the latter refers to the absence of a particular cause. What is rejected in significance testing is the statistical hypothesis, not the existence or absence of a cause. But in Fisher's writings we can read both "yes" and "no" as answers to the aforementioned question. Sometimes Fisher formulated the null hypothesis as "the treatment has no effect, period," whereas in other places he formulated it as a statistical null hypothesis (see Gigerenzer et al., 1989, pp. 95–97). In the famous Tea-Tasting Experiment in the *Design,* for instance, he stated clearly that we cannot conclude from a significant result (disproving the null) that the opposite hypothesis (which is not formulated

[3]On the distinction between statistical and substantive hypotheses, see Hager and Westermann (1983) and Meehl (1978).

as an exact statistical hypothesis in null hypothesis testing) is proven. (This experiment was designed to test a lady's claim that she could tell whether the milk or the tea infusion was first added to a cup.) That is, we cannot infer the existence of a causal process from a significant result—here, that the lady can discriminate between whether the milk or the tea infusion was first added to the cup. For instance, there exist other causal mechanisms (someone told the lady in which cups the tea infusion had been poured first) that are consistent with rejecting the null hypothesis.

What Does a Nonsignificant Result Mean?

In the *Design*, Fisher proposed asymmetry: A null hypothesis can be disproved, but "never proved or established" (p. 16), so "experimenters . . . are prepared to ignore all [nonsignificant] results" (p. 13). This has been understood by many textbook writers as saying that no conclusions can be drawn from a nonsignificant result. And several textbook authors laid down the commandment that I was taught "Thou shalt not draw inferences from a nonsignificant result." This made nonsignificance appear a negative, worthless, and disappointing result. In Neyman–Pearson theory, in contrast, there is symmetry, and a conclusion is drawn from nonsignificance: Act as if the null hypothesis were true. The reason is that Neyman and Pearson start with a disjunction of two symmetric hypotheses (either H_0 or H_1 is true), and proceed by induction through elimination.

Fisher (1955) again had second thoughts: "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). Thus, although nonsignificant results cannot establish null hypotheses, according to his second thoughts, we can do more than just "ignore" them: We may say that a nonsignificant result "confirms," but does not "establish," the null hypothesis. Now Fisher suggested that a nonsignificant result might indeed support the null hypothesis, but he did not explain how.

Power

In null hypothesis testing, only one kind of error is defined: rejecting the null hypothesis when it is in fact true. In their attempt to supply a logical basis for Fisher's ideas and make them consistent (see Gigerenzer et al., 1989, pp. 98–106), Neyman and Pearson replaced Fisher's single null hypothesis by a *set* of rival hypotheses. In the simplest case, two hypotheses, H_0 and H_1, are specified, and it is assumed that one of them is true. This assumption allows us to determine the probability of both Type I errors and Type II errors, indicated in Neyman–Pearson theory by α and β respectively. If H_1 is rejected although H_1 is true, a Type II error has occurred. α is also called the *size* of a test, and $1 - \beta$ is called its *power*. The power of a test is the long-run frequency of accepting H_1, if it is

true. The concept of power makes explicit what Fisher referred to as "sensitivity."

In the *Design*, Fisher pointed out two ways to make an experiment more sensitive: by enlarging the number of repetitions, and by qualitative methods, such as experimental refinements that minimize the error in the measurements (pp. 21–25). Nevertheless, he rejected the concept of Type II error and calculations of power on the grounds that they are inappropriate for scientific induction. In his view, calculations of power, although they look harmless, reflect the "mental confusion" between technology and scientific inference (Fisher, 1955, p. 73). If someone designs a test for acceptance procedures in *quality control*, where the goal is to minimize costs due to decision errors, calculations of power based on cost-benefit considerations in situations of repetitive tests are quite appropriate. But *scientific inference* and discovery, in Fisher's view, are about gaining knowledge, not saving money.

Fisher always rejected the concept of *power*. Neyman, for his part, pointed out that some of Fisher's tests "are in a mathematical sense 'worse than useless,'" because their power is less than their size (see Hacking, 1965, p. 99). Even in the Tea Tasting Experiment, used by Fisher to introduce the logic of null hypothesis testing in the *Design*, the power is only a little higher than the level of significance (.05), or cannot be calculated at all, depending on the conditions (see Neyman, 1950).

Random Sampling from Known Populations?

Acceptance procedures involve random sampling from a known population (say, a firm's daily production). They also allow for repeated random sampling (every day a random sample may be taken). Recall that Neyman and Pearson based their theory on the concept of repeated random sampling, which defined the probability of Type I and Type II errors as long-run frequencies of wrong decisions in repeated experiments.

Fisher, in contrast, held that in scientific applications there is no known population from which repeated sampling can be done. There are always many populations to which a sample may belong. "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" (Fisher, 1955, p. 71). Fisher proposed to view any sample (such as the sample of subjects in a typical psychological experiment, which is not drawn randomly from a known population) as a random sample from an *unknown hypothetical infinite population*. "The postulate of randomness thus resolves into the question, 'Of what population is this a random sample?' which must frequently be asked by every practical statistician" (Fisher, 1922, p. 313). But how can the practical statistician find out? The

concept of an unknown hypothetical infinite population has puzzled many: "This is, to me at all events, a most baffling conception" (Kendall, 1943, p. 17).

Mechanical Scientific Inference

One reading of the *Design* is that null hypothesis testing is a fairly mechanical procedure: Set up a null hypothesis, use a conventional level of significance, calculate a test statistic, and disprove the null hypothesis, if you can. Fisher later made clear that he did not mean it to be so. For instance, he pointed out that the choice of the test statistic, and of deciding which null hypotheses are worth testing, cannot be reduced to a mechanical process. You need constructive imagination and much knowledge based on experience (Fisher, 1933). Statistical inference has two components: informed judgment and mathematical rigor.

Similarly, Neyman and Pearson always emphasized that the *statistical* part has to be supplemented by a *subjective* part. As Pearson (1962) put it: "We left in our mathematical model a gap for the exercise of a more intuitive process of personal judgment in such matters—to use our terminology—as the choice of the most likely class of admissible hypotheses, the appropriate significance level, the magnitude of worthwhile effects and the balance of utilities" (pp. 395–396).

In Neyman and Pearson's theory, once all judgments are made, the decision (reject or accept) falls out mechanically from the mathematics. In his later writings, Fisher opposed these mechanical accept/reject decisions, which he believed to be inadequate in science where one looks forward to further data. Science is concerned with communication of information, such as exact levels of significance. Again, Fisher saw a broader context, the freedom of the Western world. Communication of information (but not mechanical decisions) recognizes "the right of *other* free minds to utilize them in making *their own* decisions" (Fisher, 1955, p. 77).

But Neyman reproached Fisher with the same sin—mechanical statistical inference. As a statistical behaviorist, Neyman (1957) looked at what Fisher actually did in his own research in genetics, biology, and agriculture, rather than at what he said one should do. He found Fisher using .01 as a conventional level of significance, without giving any thought to the choice of a particular level dependent on the particular problem or the probability of an error of the second kind; he accused Fisher of drawing mechanical conclusions, depending on whether or not the result was significant. Neyman urged a thoughtful choice of the level of significance, *not* using .01 for all problems and contexts.

Both camps in the controversy accused the other party of mechanical, thoughtless statistical inference, thus I conclude that here at least they agreed—statistical inference should not be automatic.

These differences between what Fisher proposed as the logic of significance testing and what Neyman and Pearson proposed as the logic of hypothesis testing

suffice for the purpose of this chapter. Both have developed further tools for inductive inference, and so did others, resulting in a large toolbox that contains maximum likelihood, fiducial probability, confidence interval approaches, point estimation, Bayesian statistics, sequential analysis, and exploratory data analysis, to mention only a few. But it is null hypothesis testing and Neyman–Pearson hypothesis-testing theory that have transformed experimental psychology and part of the social sciences.

THE OFFSPRING: HYBRID LOGIC

The conflicting views presented earlier are those of the parents of the hybrid logic. Not everyone can tolerate unresolved conflicts easily and engage in a free market of competing ideas. Some long for the single truth or search for a compromise that could at least repress the conflicts. Kendall (1949) commented on the desire for peace negotiations among statisticians:

> If some people asserted that the earth rotated from east to west and others that it rotated from west to east, there would always be a few well-meaning citizens to suggest that perhaps there was something to be said for both sides, and maybe it did a little of one and a little of the other; or that the truth probably lay between the extremes and perhaps it did not rotate at all. (p. 115)

The denial of the existing conflicts and the pretense that there is only one statistical solution to inductive inference were carried to an extreme in psychology and several neighboring sciences. This one solution was the *hybrid logic of scientific inference*, the offspring of the shotgun marriage between Fisher and Neyman and Pearson. The hybrid logic became institutionalized in experimental psychology (see Gigerenzer, 1987), personality research (see Schwartz & Dangleish, 1982), clinical psychology and psychiatry (see Meehl, 1978), education (see Carver, 1978), quantitative sociology (see Morrison & Henkel, 1970), and archaeology (see Cowgill, 1977; Thomas, 1978), among others. Nothing like this happened in physics, chemistry, or molecular biology (see Gigerenzer et al., 1989).

The Hybrid Logic Is Born

Before World War 2, psychologists drew their inferences about the validity of hypotheses by many means—ranging from eyeballing to critical ratios. The issue of statistical inference was not of primary importance. Note that this was not because techniques were not yet available. On the contrary; already in 1710, John Arbuthnot proved the existence of God by a kind of significance test, astronomers had used them during the 19th century for rejecting outliers (Swij-

tink, 1987), and Fechner (1897) wrote a book on statistics including inference techniques—to give just a few examples. Techniques of statistical inference were known and sometimes used, but experimental method was not yet dominated by and almost equated with statistical inference.

Through the work of the statisticians Snedecor at Iowa State College, Hotelling at Columbia University, and Johnson at the University of Minnesota, Fisher's ideas spread in the United States. Psychologists began to cleanse the Fisherian message of its agricultural smell and its mathematical complexity, and to write a new genre of textbooks featuring null hypothesis testing. Guilford's *Fundamental Statistics in Psychology and Education*, first published in 1942, was probably the most widely read textbook in the 1940s and 1950s. In the preface, Guilford credited Fisher for the new logic of hypothesis testing taught in a chapter that was "quite new to this type of text" (p. viii). The book does not mention Neyman, E. S. Pearson, or Bayes. What Guilford teaches as the logic of hypothesis testing is Fisher's null hypothesis testing, deeply colored by "Bayesian" terms: Null hypothesis testing is about the probability that the null hypothesis is true. "If the result comes out one way, the hypothesis is probably correct, if it comes out another way, the hypothesis is probably wrong" (p. 156). Null hypothesis testing is said to give degrees of doubt such as "probable" or "very likely" a "more exact meaning" (p. 156). Its logic is explained via headings such as "Probability of hypotheses estimated from the normal curve" (p. 160).

Guilford's logic is not consistently Fisherian, nor does it consistently use "Bayesian" language of probabilities of hypotheses. It wavers back and forth and beyond. Phrases like "we obtained directly the probabilities that the null hypothesis was plausible" and "the probability of extreme deviations from chance" are used interchangeably for the same thing: the level of significance. And when he proposed his own "somewhat new terms," his intuitive Bayesian thinking becomes crystal clear. A p value of .015 for a hypothesis of zero difference in the population "gives us the probability that the true difference is a negative one, and the remainder of the area *below* the point, or .985, gives us the probability that the true difference is positive. The odds are therefore .985 to .015 that the true difference is positive" (p. 166). In Guilford's hands, p values that specify probabilities $p(D/H)$ of some data (or test statistic) D given a hypothesis H turn miraculously into Bayesian posterior probabilities $p(H/D)$ of a hypothesis given data.

Guilford's logic is not an exception. It marks the beginning of a genre of statistical texts that vacillate between the researcher's "Bayesian" desire for probabilities of hypotheses and what Fisher is willing to give them.

This first phase of teaching Fisher's logic soon ran into a serious complication. In the 1950s and 1960s, the theory of Neyman and E. S. Pearson also became known. How were the textbook writers to cope with two logics of scientific inference? How should the ideological differences and personal insults be dealt with? Their solution to this conflict was striking. The textbook writers

did not side with Fisher. That is, they did not go on to present null hypothesis testing as scientific inference and add a chapter on hypothesis testing outside science, introducing the Neyman–Pearson theory as a logic for quality control and related technological problems. Nor did they side with Neyman and Pearson, teaching their logic as a consistent and improved version of Fisher's and dispensing entirely with Fisherian null hypothesis testing.

Instead, textbook writers started to add Neyman–Pearsonian concepts on top of the skeleton of Fisher's logic. But acting as if they feared Fisher's revenge, they did it without mentioning the names of Neyman and Pearson. A *hybrid logic* of statistical inference was created in the 1950s and 1960s. Neither Fisher nor Neyman and Pearson would have accepted this hybrid as a theory of statistical inference. The hybrid logic is inconsistent from both perspectives and burdened with conceptual confusion. Its two most striking features are (a) it hides its hybrid origin and (b) it is presented as *the* monolithic logic of scientific inference. Silence about its origin means that the respective parts of the logic are not identified as part of two competing and partly inconsistent theoretical frameworks. For instance, the idea of testing null hypotheses without specifying alternative hypotheses is not identified as part of the Fisherian framework, and the definition of the level of significance and the power of a test as long-run frequencies of false and correct decisions, respectively, in repeated experiments is not identified as part of the Neyman–Pearson framework. And, as a consequence, there is no mention of the fact that each of these parts of the hybrid logic were rejected by the other party, and why, and what the unresolved controversial issues are.

The Structure of Hybrid Logic

In order to capture the emotional tensions associated with the hybrid logic, I use a Freudian analogy.[4]

The Neyman–Pearson logic of hypothesis testing functions as the Superego of the hybrid logic. It demands the specification of precise alternative hypotheses, significance levels, and power in advance to calculate the sample size necessary, and it teaches the doctrine of repeated random sampling. The frequentist Superego forbids epistemic statements about particular outcomes or intervals, and it outlaws the interpretation of levels of significance as the degree of confidence that a particular hypothesis is true or false.

The Fisherian theory of significance testing functions as the Ego. The Ego gets things done in the laboratory and gets papers published. The Ego determines the level of significance after the experiment, and it does not specify power nor calculate the sample size necessary. The Ego avoids precise predictions from its

[4]Here I am elaborating on a metaphor suggested by Acree (1978). In a different context, Devereux (1967) talked about the relation between anxiety and elimination of subjectivity by method.

research hypothesis; that is, it does not specify the exact predictions of the alternative hypothesis, but claims support for it by rejecting a null hypothesis. The Ego makes abundant epistemic statements about particular results. But it is left with feelings of guilt and shame for having violated the rules.

Censored by both the frequentist Superego and the pragmatic Ego are statements about probabilities of hypotheses given data. These form the Bayesian Id of the hybrid logic. Some direct measure of the validity of the hypotheses under question—quantitatively or qualitatively—is, after all, what researchers really want.

The Freudian metaphor suggests that the resulting conceptual confusion in the minds of researchers, editors, and textbook writers is not due to limited intelligence. The metaphor brings the anxiety and guilt, the compulsive and ritualistic behavior, and the dogmatic blindness associated with the hybrid logic into the foreground. It is as if the raging personal and intellectual conflicts between Fisher and Neyman and Pearson, and between these frequentists and the Bayesians were projected into an "intrapsychic" conflict in the minds of researchers. And the attempts of textbook writers to solve this conflict by denying it have produced remarkable emotional, behavioral, and cognitive distortions.

Anxiety and Guilt

Editors and textbook writers alike have institutionalized the level of significance as a measure of the quality of research. As mentioned earlier, Melton, after 12 years editing one of the most prestigious journals in psychology, said in print that he was reluctant to publish research with significance levels below .05 but above .01, whereas $p < .01$ made him confident that the results would be repeatable and deserved publication (1962, pp. 553–554). In Nunnally's *Introduction to Statistics for Psychology and Education* (1975) the student is taught similar values and informed that the standard has been raised: "Up until 20 years ago, it was not uncommon to see major research reports in which most of the differences were significant only at the 0.05 level. Now, such results are not taken very seriously, and it is more customary today to see results reported only if they reach the 0.01 or even lower probability levels" (p. 195). Not accidentally, both Melton and Nunnally show the same weak understanding of the logic of inference and share the same erroneous belief that the level of significance specifies the probability that a result can be replicated (discussed later). The believers in the divinatory power of the level of significance set the standards.

The researcher's Ego knows that these publish-or-perish standards exist in the outside world, and knows that the best way to adapt is to round up the obtained p value after the experiment to the nearest conventional level, say to round up the value $p = .006$ and publish $p < .01$. But the Superego has higher moral standards: If you set alpha to 5% before the experiment, then you must report the same finding ($p = .006$) as "significant at the 5% level." Mostly, the Ego gets its

way, but is left with feelings of dishonesty and of guilt at having violated the rules. Conscientious experimenters have experienced these feelings, and statisticians have taken notice. The following comment was made in a panel discussion among statisticians; Savage remarked on the statisticians' reluctance to take responsibility for once having built up the Superego in the minds of the experimenters:

> I don't imagine that anyone in this room will admit ever having taught that the way to do an experiment is first carefully to record the significance level then do the experiment, see if the significance level is attained, and if so, publish, and otherwise, perish. Yet, at one time we must have taught that; at any rate it has been extremely well learned in some quarters. And there is many a course outside of statistics departments today where the modern statistics of twenty or thirty years ago is taught in that rigid way. People think that's what they're supposed to do and are horribly embarrassed if they do something else, such as do the experiment, see what significance level would have been attained, and let other people know it. They do the better thing out of their good instincts, but think they're sinning. (Barnard, Kiefer, LeCam & Savage, 1968, p. 147)

Statistics has become more tolerant than its offspring, the hybrid logic.

Denial of the Parents

The hybrid logic attempts to solve the conflict between its parents by denying its parents. It is remarkable that textbooks typically teach hybrid logic without mentioning Neyman, E. S. Pearson, and Fisher—except in the context of technical details, such as specific tables, that are incidental to the logic. In 25 out of 30 textbooks I have examined, Neyman and E. S. Pearson do not appear to exist. For instance, in his *Statistical Principles in Experimental Design* (1962; 2nd ed., 1971), Winer credited Fisher for the "logic of scientific method" (p. 3), and a few pages later, introduced the Neyman–Pearson terminology of Type I error, Type II error, power, two precise statistical hypotheses, cost-benefit considerations, and *rejecting* and *accepting* hypotheses. Nowhere in the book do the names of Neyman and E. S. Pearson appear (except in a "thank you" note to Pearson for permission to reproduce tables), although quite a few other names can be found in the index. No hint is given to the reader that there are different ways to think about the logic of inference. Even in the exceptional case of Hays's textbook (1963), where all parents are mentioned by their names, the relationship of their ideas is presented (in a single sentence) as one of cumulative progress, from Fisher to Neyman and Pearson (p. 287).[5] Both Winer's and Hays's are among the best texts, without the confusions that abound in Guilford's, Nunnally's, and a

[5]In the 3rd edition (1981), however, Hays's otherwise excellent text falls back to common standards: J. Neyman and E. S. Pearson no longer appear in the book.

mass of other textbooks. Nevertheless, even in these texts the parents' different ways of thinking about statistical inference and the controversial issues are not pointed out.

Denial of Conflicts Between Parents

Thus the conflicting views are almost unknown to psychologists. Textbooks are uniformly silent. (Some statistics teachers protest that airing these disputes would only confuse students. I believe that pointing out the conflicting views would make statistics much more interesting to students who enjoy thinking rather than being told what to do next.) As a result of this silence, many a text muddles through the conflicting issues leaving confusion and inconsistency in its wake—at least, among the more intelligent and alert students. For instance, Type I and Type II errors are often defined in terms of long-run frequencies of erroneous decisions in repeated experiments, but the texts typically stop short of Neyman's behavioral interpretation, and fall back to epistemic interpretations of the two errors as levels of confidence about the validity of the hypotheses. In fact, the poorer texts overflow with amazing linguistic contortions concerning what a level of significance means. For instance, within three pages of text, Nunnally explained that "level of significance" means all of the following: (a) "If the probability is low, the null hypothesis is improbable" (p. 194); (b) "the *improbability* of observed results being due to error" (p. 195); (c) "the probability that an observed difference is real" (p. 195); (d) "the *statistical confidence* . . . with odds of 95 out of 100 that the observed difference will hold up in investigations" (p. 195); (e) the degree to which experimental results are taken "seriously" (p. 195); (f) "the danger of accepting a statistical result as real when it is actually due only to error" (p. 195); (g) the degree of "faith [that] can be placed in the reality of the finding" (p. 196); (h) "the null hypothesis is rejected at the 0.05 level"; and (i) "the investigator can have 95 percent confidence that the sample mean actually differs from the population mean" (p. 196). And, after the last two versions, the author assured his readers: "All of these are different ways to say the same thing" (p. 196).

Nunnally did not spell out the differences between the logics of Fisher, Neyman and Pearson, and the Bayesians. He avoided the conflicting interpretations by declaring that everything is the same. The price for this is conceptual confusion, false assertions, and an illusory belief in the omnipotence of the level of significance. Nunnally was a pronounced but not an atypical case.

Obsessive-Compulsive and Mechanical Behavior

As previously mentioned, statisticians have emphasized the indispensable role of personal judgment, although with respect to different parts of their logics. For Fisher, informed judgment was needed for the choice of the statistical model, the

test statistics, and a null hypothesis worth investigating. For Neyman and Pearson, personal judgment was needed for the choice of the class of hypotheses (two hypotheses, in the simplest case), and the cost-benefit considerations that lead to the choice of Type I error, power, and sample size. For Bayesians such as de Finetti, finally, "subjectivism" and "relativism" are the very cornerstones of 20th-century probability theory (de Finetti, 1931/1989; Jeffrey, 1989).

The need for these kinds of informed judgments was rarely a topic in the textbooks. Rather, a mass of researchers must have read the textbooks as demanding the mindless, mechanical setting up of null hypotheses and recording of p values. Journals filled with p values, stars, double stars, and triple stars that allegedly established replicable "facts" bear witness to this cookbook mentality.

Guilford's misunderstanding that to set up a null hypothesis means to postulate a *zero* difference or a *zero* correlation was perpetuated. "Null" denotes the hypothesis to be "nullified," not that it is necessary to postulate a zero effect. Rarely were null hypotheses formulated that postulated something other than a zero effect (such as "the difference between the means is 3 scale points"). Rarely were precise alternative hypotheses stated, and even if there were two competing precise hypotheses, as in Anderson's information integration theory, only one of them was tested as the null hypothesis, sometimes resulting in tests with a power as low as .06 (Gigerenzer & Richter, 1990). Reasons for using a particular level of significance were almost never given, and rarely was a judgment about the desired power made and the sample size calculated. As a result, the power of the tests is typically quite low (below .50 for a medium effect), and pointing this out (Cohen, 1962) has not changed practice. Two-and-a-half decades after Cohen's work, the power of the null hypothesis tests was even slightly worse (Sedlmeier & Gigerenzer, 1989). Rather, null hypotheses are set up and tested in an extremely mechanical way reminiscent of compulsive handwashing. One can feel widespread anxiety surrounding the exercise of informed personal judgment in matters of hypothesis testing. The availability of statistical computer packages seems to have reinforced this mechanical behavior. A student of mine once tested in his thesis the difference between two means, which were numerically exactly the same, by an F test. Just to say that the means are the same seemed to him not objective enough.

The institutionalization of the hybrid logic as the sine qua non of scientific method is the environment that encourages mechanical hypothesis testing. The *Publication Manual of the American Psychological Association*, for instance, called "rejecting the null hypothesis" a "basic" assumption (1974, p. 19) and presupposes the hybrid logic. The researcher was explicitly told to make mechanical decisions: "Caution: Do not infer trends from data that fail by a small margin to meet the usual levels of significance. Such results are best interpreted as caused by chance and are best reported as such. Treat the result section like an income tax return. Take what's coming to you, but no more" (p. 19; this passage was deleted in the 3rd ed., 1983). This prescription sounds like a Neyman–

Pearson accept–reject logic, where it matters for a decision only on which side of the criterion the data fall, not how far. Fisher would have rejected such mechanical behavior (e.g., Fisher, 1955, 1956). Nevertheless, the examples in the manual that tell the experimenter how to report results use p values that were obviously determined *after* the experiment and rounded up to the next conventional level, such as $p < .05$, $p < .01$, and $p < .001$ (pp. 39, 43, 48, 49, 70, 96). Neyman and Pearson would have rejected this practice: These p values are not the probability of Type I errors—and determining levels of significance after the experiment prevents determining power and sample size in advance. Fisher (e.g., 1955, 1956) would have preferred that the exact level of significance, say $p = .03$, be reported, not upper limits, such as $p < .05$, which look like probabilities of Type I errors but aren't.

Distorted Statistical Intuitions

Mechanical null hypothesis testing seems to go hand-in-hand with distorted statistical intuitions. I distinguish distorted statistical intuitions from the confusion and inconsistency of the hybrid logic itself. The latter results from mishmashing Fisher and Neyman and Pearson without making the conflation explicit, as I argued earlier. The conceptual confusion of the hybrid logic provided fertile ground for the growth of what I call *distorted statistical intuitions*. The distortions all seem to go in one direction: They exaggerate what can be inferred from a p value.

The framework of distorted intuitions makes the obsessive performance of null hypothesis testing seem quite reasonable. Therefore, distorted intuitions serve an indispensable function. These illusions guide the writings of several textbook authors and editors, but they seem to be most pronounced in the users of null hypothesis testing, researchers in psychology and neighboring fields. Some distorted intuitions concern the frequentist part of the hybrid logic, others the Bayesian Id. I give one example of each (there is a larger literature on distorted statistical intuitions taught in statistical textbooks and held by experimenters; see Acree, 1978; Bakan, 1966; Brewer, 1985; Carver, 1978; Guttman, 1977, 1985; Lykken, 1968; Pollard & Richardson, 1987; Rozeboom, 1960; Tversky & Kahneman, 1971).

Replication Fallacy. Suppose α is set as .05 and the null hypothesis is rejected in favor of a given alternative hypothesis. What if we replicate the experiment? In what percentage of exact replications will the result again turn out significant? Although this question arises from the frequentist conception of repeated experiments, the answer is unknown. The α we chose does not tell us, nor does the exact level of significance.

The *replication fallacy* is the belief that the level of significance provides an answer to the question. Here are some examples: In an editorial of the *Journal of*

Experimental Psychology, the editor stated that he used the level of significance reported in submitted papers as the measure of the "confidence that the results of the experiment would be repeatable under the conditions described" (Melton, 1962, p. 553). Many textbooks fail to mention that the level of significance does not specify the probability of a replication, and some explicitly teach the replication fallacy. For instance, "The question of statistical significance refers primarily to the extent to which similar results would be expected if an investigation were to be repeated" (Anastasi, 1958, p. 9). Or, "If the statistical significance is at the 0.05 level . . . the investigator can be confident with odds of 95 out of 100 that the observed difference will hold up in future investigations" (Nunnally, 1975, p. 195). Oakes (1986, p. 80) asked 70 university lecturers, research fellows, and postgraduate students with at least 2 years' research experience what a significant result ($t = 2.7$, $df = 18$, $p = .01$) means. Sixty percent of these academic psychologists erroneously believed that these figures mean that if the experiment is repeated many times, a significant result would be obtained 99% of the time.

In Neyman and Pearson's theory the level of significance (alpha) is defined as the relative frequency of rejections of H_0 if H_0 is true. In the minds of many, 1 − alpha erroneously turned into the relative frequency of rejections of H_0, that is, into the probability that significant results could be replicated.

The Bayesian Id's Wishful Thinking. I mentioned earlier that Fisher both rejected the Bayesian cake and wanted to eat it, too: He spoke of the level of significance as a measure of the degree of confidence in a hypothesis. In the minds of many researchers and textbook writers, however, the level of significance virtually turned into a Bayesian posterior probability.

What I call the *Bayesian Id's wishful thinking* is the belief that the level of significance, say .01, is the probability that the null hypothesis is correct, or that 1 − .01 is the probability that the alternative hypothesis is correct. In various linguistic versions, this wishful thinking was taught in textbooks from the very beginning. Early examples are Anastasi (1958, p. 11), Ferguson (1959, p. 133), Guilford (1942, pp. 156–166), and Lindquist (1940, p. 14). But the belief has persisted over decades of teaching hybrid logic, for instance in Miller and Buckhout (1973, statistical appendix by Brown, p. 523), Nunnally (1975, pp. 194–196), and the examples collected by Bakan (1966) and Pollard and Richardson (1987). Oakes (1986, p. 82) reported that 96% of academic psychologists erroneously believed that the level of significance specifies the probability that the hypothesis under question is true or false.

The Bayesian Id has got its share. Textbook writers have sometimes explicitly taught this misinterpretation, but more often invited it by not specifying the difference between a Bayesian posterior probability, a Neyman–Pearsonian probability of a Type I error, and a Fisherian exact level of significance.

Dogmatism

The institutionalization of *one* way to do hypothesis testing had its benefits. It made the administration of the social science research that exploded since World War 2 easier, and it facilitated editors' decisions. And there were more benefits. It reduced the high art of hypothesis construction, of experimental ingenuity and informed judgment, into a fairly mechanical schema that could be taught, learned, and copied by almost anyone. The informed judgments that remain are of a low-level kind: whether to use a one- or a two-tailed significance test. (But even here some believed that there should be no room for judgment, because even this simple choice seemed to threaten the ideal of mechanical rules and invite cheating.) The final, and perhaps most important, benefit of the hybrid logic is that it provides the satisfying illusion of *objectivity:* The statistical logic of analyzing data seemed to eliminate the subjectivity of eyeballing and wishful distortion. To obtain and maintain this illusion of objectivity and impartiality, the hybrid logic had to deny its parents—and their conflicts.

The danger of subjective distortion and selective reading of data exists, to be sure. But it cannot be cured by replacing the distortions of particular experimenters by a collective distortion. Note that the institutionalized practice produces only selective and limited objectivity, and hands other parts of scientific practice over to rules of thumb—even parts for which the statistical methods would be applicable. For example, during the 19th century, astronomers used significance tests to reject *data* (so-called outliers), assuming, at least provisionally, that their hypothesis was correct (Swijtink, 1987). Social scientists today, in contrast, use significance tests to reject *hypotheses,* assuming that their data are correct. The mathematics does not dictate which one the scientists should trust and which one they should try to refute. Social scientists seem to have read the statistical textbooks as saying that statistical inference is indispensable in selecting good from bad hypotheses, but not for selecting good from bad data. The problem of outliers is dealt with by rules of thumb.[6]

The dogmatism with which the hybrid logic has been imposed on psychology researchers by many textbook writers and editors and by researchers themselves has lasted for almost half a century. This is far too long. We need a knowledgeable use of statistics, not a collective compulsive obsession. The last two decades suggest that things are, although very slowly, changing in the right direction.

---

[6]So is the problem of how many replications (subjects) an experiment should use. Sedlmeier and Gigerenzer (1989) found no use of Neyman–Pearsonian calculations of sample size in published work. Some statistical texts have explicitly encouraged this: "Experienced researchers use a rule of thumb sample size of approximately twenty. Smaller samples often result in low power values while larger samples often result in a waste of time and money" (Bruning & Kintz, 1977, p. 7).

BEYOND DOGMATISM: TOWARD A THOUGHTFUL USE OF STATISTICS

Here are a few first principles: Do not replace the dogmatism of the hybrid logic of scientific inference by a new, although different one (e.g., Bayesian dogmatism). Remember the obvious: The problem of inductive inference has no universal mathematical solution. Use informed judgment and statistical knowledge. Here are several more specific suggestions:

1. *Stop teaching hybrid logic as the sine qua non of scientific inference.* Teach researchers and students alternative theories of statistical inference, give examples of typical applications and teach the students how to use these theories in a constructive (not mechanical) way. Point out the confused logic of the hybrid, the emotional, behavioral, and cognitive distortions associated with it, and insist on consistency (Cohen, 1990). This will lead to recognizing the second point.

2. *Statistical inference (Fisherian, Neyman–Pearsonian, or Bayesian) is rarely the most important part of data analysis.* Teach researchers and students to look at the data, not just on p values. Computer-aided graphical methods of data display and exploratory data analysis are means toward this end (Diaconis, 1985; Tukey, 1977). The calculation of descriptive statistics such as effect sizes is a part of data analysis that cannot be substituted by statistical inference (Rosnow & Rosenthal, 1989). A good theory predicts particular curves or effect sizes, but not levels of significance.

3. *Good data analysis is pointless without good data.* The measurement error should be controlled and minimized before and during the experiment; instead one tends to control it after the experiment by inserting the error term in the F ratio. Teach researchers and students that the important thing is to have a small real error in the data. Without that, a significant result at any level is, by itself, worthless—as Gosset, who developed the t test in 1908, emphatically emphasized (see Pearson, 1939). Minimizing the real error in measurements may be achieved by an iterative method: First, obtain measurements and look at the error variance, then try methods to minimize the error (e.g., stronger experimental control, investigating each subject carefully in a single-case study rather than in a classroom), then go back and obtain new measurements and look at the new error variance, and so on, until improvements are no longer possible. Axiomatic measurement theory that focuses on ordinal rather than numerical judgments may help (Krantz, Luce, Suppes, & Tversky, 1971). It is all too rarely used.

4. *Good data need good hypotheses and theories to survive.* We need rich theoretical frameworks that allow for specific predictions in the form of precise research hypotheses. The null hypothesis of zero difference (or zero correlation) is only one version of such a hypothesis—perhaps only rarely appropriate. In

particular, it has become a bad habit not to specify the predictions of a research hypothesis, but to specify a different hypothesis (the null) and to try to reject it and claim credit for the unspecified research hypothesis. Teach students to derive competing hypotheses from competing theoretical frameworks, and to test their ordinal or quantitative predictions *directly,* without using the null as a straw man.

EPILOGUE: MORE SUPEREGOS

Around 1840, the classical theory of probability dissolved and the frequentist interpretation of probability emerged (Daston, 1988; Porter, 1986). Today, teaching in statistics departments is still predominantly in the frequentist tradition, and Fisher's and Neyman and Pearson's theories are two variants thereof. But this century has witnessed the revival of subjective probability, often referred to as *Bayesian statistics,* largely through the writings of the Italian actuary de Finetti and the English philosopher Ramsey in the 1920s and 1930s, and in the 1950s by the American statistician Savage. For a Bayesian, probability is about subjective degrees of belief, not about objective frequencies. A degree of belief of 1/10 that the next president of the United States will be a woman can be interpreted as the willingness to take either side of a nine to one bet on this issue. Bayesians are still a minority in statistics departments, but the Bayesian model of rationality has found a role in theoretical economics (mainly microeconomics), cognitive psychology, artificial intelligence, business, and medicine.

In 1963, Edwards, Lindman, and Savage argued that psychologists should stop frequentist null hypothesis testing and do Bayesian statistics instead (their counterparts in Europe were, among others, Kleiter, 1981; Tholey, 1982). Edwards and his colleagues also started a research program on whether intuitive statistical judgments follow Bayes' theorem. Their suggestion that psychologists should turn Bayesian fell on deaf ears, both in the United States and in Europe. Researchers already had their hybrid logic, which seemed to them the objective way to do scientific inference, whereas Bayesian statistics looked subjective. And given the distorted statistical intuitions of many, there was actually no need; the level of significance already seemed to specify the desired Bayesian posterior probabilities.[7]

The second of Edwards's proposals, in contrast, caught on: To study whether and when statistical intuitions conform to Bayes' theorem (e.g., Edwards, 1968). More than in Edwards's research, the heuristics and biases program of the 1970s and 1980s (e.g., Tversky & Kahneman, 1974) focussed on what were called *fallacies* and *errors* in probabilistic reasoning: *discrepancies* between human judgment and Bayes' formula.

[7] I know of only a handful of studies published in psychological journals where researchers used Bayesian statistics instead of the hybrid logic. Even Hays, who included a chapter on Bayesian statistics in the second edition of his statistics text, dropped it in the third edition.

The New Bayesian Superego

The Bayesian Id of the hybrid logic had turned into the new Superego of research on intuitive statistics. Frequentist theories were suppressed. Bayesian statistics (precisely, one narrow version thereof) was seen as *the* correct method of statistical reasoning, whether it was about the subjective probability that a particular person was an engineer (Kahneman & Tversky, 1973) or that a cab involved in a hit-and-run accident at night was blue (Tversky & Kahneman, 1980). However, if one applies Neyman–Pearson theory to the cab problem, or alternative Bayesian views, one obtains solutions that are strikingly different from Tversky and Kahneman's Bayesian calculations (Birnbaum, 1983; Gigerenzer & Murray, 1987, pp. 167–174; Levi, 1983). The objections of Fisher and Neyman to the universal use of Bayesian statistics seemed to be buried below the level of consciousness, and so was the most basic objection of a frequentist: Probability is about frequencies, not about single events (such as whether a particular cab was blue or Linda is a bank teller).

A striking result demonstrates the importance of that objection: So-called fallacies frequently disappear when subjects are asked for frequency judgments rather than for single-event probabilities (Gigerenzer, 1991a, 1991b; Gigerenzer, Hoffrage, & Kleinbölting, 1991). Within the heuristics and biases program, the frequentist Superego of the hybrid logic, who had banned probability statements about particular events or values, was no longer heard. Nor was the frequentist Barnard (1979), who commented thus on subjective probabilities for single events: "If we accept it as important that a person's subjective probability assessments should be made coherent, our reading should concentrate on the works of Freud and perhaps Jung rather than Fisher and Neyman" (p. 171).

Suddenly, the whole psychic structure of statistical reasoning in psychology seemed to be reversed. Now Bayesian statistics (precisely, a narrow version thereof) was presented as the sine qua non of statistical reasoning, as *the* normative standard. Against this standard, all deviating reasoning seemed to be a fallacy. Neyman had warned of "the *dogmatism* which is occasionally apparent in the application of Bayes' formula" (1957, p. 19). He meant the conviction "that it is possible to devise a formula of universal validity which can serve as a normative regulator of our beliefs" (p. 15). Similarly, for Fisher, only some uncertain inferences, but not all kinds, can be adequately dealt with by probability theory. Bayesian theory "is founded upon an error, and must be wholly rejected" (Fisher, 1925, p. 9).

Good statistical reasoning has been once more equated with the mechanical application of some statistical formula.

It seems to have gone almost unnoticed that this dogmatism has created a strange double standard. Many researchers believe that their subjects must use Bayes' theorem to test hypotheses, but the researchers themselves use the hybrid logic to test their hypotheses—and thus themselves ignore base rates. There is

the illusion that one kind of statistics normatively defines objectivity in scientific inference, and another one rationality in everyday inference. The price is a kind of "split brain," where Neyman–Pearson logic is the Superego for experimenters' hypothesis testing and Bayesian statistics is the Superego for subjects' hypothesis testing.

CONCLUSIONS

Statistical reasoning is an art and so demands both mathematical knowledge and informed judgment. When it is mechanized, as with the institutionalized hybrid logic, it becomes ritual, not reasoning. Many colleagues of mine have argued that it is not going to be easy to get researchers in psychology and other sociobiomedical sciences to drop this comforting crutch unless one offers an easy-to-use substitute. But this is exactly what I want to avoid—the substitution of one mechanistic dogma for another. It is our duty to inform our students about the many good roads to statistical inference that exist, and to teach them how to use informed judgment to decide which one to follow for a particular problem. At the very least, this chapter can serve as a tool in arguments with people who think they have to defend a ritualistic dogma instead of good statistical reasoning. Making and winning such arguments is indispensable to good science.

ACKNOWLEDGMENTS

This chapter was written while I was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, CA. I am grateful for financial support provided by the Spencer Foundation and the Deutsche Forschungsgemeinschaft (DFG 170/2-1). Leda Cosmides, Lorraine Daston, Raphael Diepgen, Ward Edwards, Ruma Falk, Gideon Keren, Duncan Luce, Kathleen Much, Zeno Swijtink, and John Tooby helped to improve the present chapter.

REFERENCES

Acree, M. C. (1978). *Theories of statistical inference in psychological research: A historicocritical study.* Ann Arbor, MI: University Microfilms International. (University Microfilms No. H790 H7000)

Anastasi, A. (1958). *Differential psychology* (3rd ed.). New York: Macmillan.

Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society, 27,* 186–190.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Barnard, G. A. (1979). Discussion of the paper by Professors Lindley and Tversky and Dr. Brown. *Journal of the Royal Statistical Society* (A), *142,* 171–172.

Barnard, G. A., Kiefer, J. C., LeCam, L. M., & Savage, L. J. (1968). Statistical inference. In D. G. Watts (Ed.), *The future of statistics* (p. 147). New York: Academic Press.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society, 53*, 370–418.

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *American Journal of Psychology, 96*, 85–94.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10*, 252–268.

Bruning, J. L., & Kintz, B. L. (1977). *Computational handbook of statistics* (2nd ed.). Glennview, IL: Scott, Foresman.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378–399.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312.

Cowgill, G. L. (1977). The trouble with significance tests and what we can do about it. *American Antiquity, 42*, 350–368.

Danziger, K. (1990). *Constructing the subject*. Cambridge: Cambridge University Press.

Daston, L. (1988). *Classical probability in the Enlightenment*. Princeton, NJ: Princeton University Press.

Devereux, G. (1967). *From anxiety to method in the behavioral sciences*. Paris: Mouton.

Diaconis, P. (1985). Theories of data analysis: From magical thinking through classical statistics. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends and shapes* (pp. 1–36). New York: Wiley.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193–242.

Estes, W. K. (1959). The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 2, pp. 380–491). New York: McGraw-Hill.

Fechner, G. T. (1897). *Kollektivmasslehre* (G. F. Lipps, Ed.). Leipzig: W. Engelmann.

Ferguson, L. (1959). *Statistical analysis in psychology and education*. New York: McGraw-Hill.

Finetti, B. De (1989). Probabilism. *Erkenntnis, 31*, 169–223. (Original work published 1931)

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A, 222*, 309–368.

Fisher, R. A. (1925). *Statistical methods for research workers* (8th ed., 1941). Edinburgh: Oliver & Boyd.

Fisher, R. A. (1933). The contributions of Rothamsted to the development of the science of statistics. *Annual Report of the Rothamsted Station*, 43–50. (Reprinted in *Collected papers*, Vol. 3, 84–91)

Fisher, R. A. (1935). *The design of experiments* (5th ed., 1951; 7th ed., 1960; 8th ed., 1966). Edinburgh: Oliver & Boyd.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society* (B), *17*, 69–77.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution. Vol. 2. Ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Gigerenzer, G. (1991a). From tools to theories. A heuristic of discovery in cognitive psychology. *Psychological Review, 98*, 252–267.

Gigerenzer, G. (1991b). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology, 2,* 83–115.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98,* 506–528.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., & Richter, H. R. (1990). Context effects and their interaction with development: Area Judgments. *Cognitive Development, 5,* 235–264.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance. How probability changed science and everyday life.* Cambridge: Cambridge University Press.

Gruber, H. E. (1977). The fortunes of a basic Darwinean idea: Chance. In R. W. Rieber & K. Salzinger (Eds.), *The roots of American psychology: Historical influences and implications for the future* (pp. 233–245). New York: New York Academy of Sciences.

Gruber, H. E., & Vonèche, J. J. (Eds.). (1977). *The essential Piaget.* New York: Basic Books.

Guilford, J. P. (1942). *Fundamental Statistics in Psychology and Education* (3rd ed., 1956, 6th ed., 1978, with B. Fruchter). New York: McGraw-Hill.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26,* 81–107.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1,* 3–10.

Hacking, I. (1965). *Logic of statistical inference.* Cambridge: Cambridge University Press.

Hager, W., & Westermann, R. (1983). Zur Wahl und Prüfung statistischer Hypothesen in psychologischen Untersuchungen. *Zeitschrift für experimentelle und angewandte Psychologie, 30,* 67–94.

Hays, W. L. (1963). *Statistics for psychologists* (2nd ed.). New York: Holt, Rinehart & Winston.

Jeffrey, R. (1989). Reading Probabilismo. *Erkenntnis, 31,* 225–237.

Johnstone, D. J. (1987). Tests of significance following R. A. Fisher. *British Journal of the Philosophy of Science, 38,* 481–499.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Kendall, M. G. (1942). On the future of statistics. *Journal of the Royal Statistical Society, 105,* 69–80.

Kendall, M. G. (1943). *The advanced theory of statistics* (Vol. 1). New York: Lippincott.

Kendall, M. G. (1949). On the reconciliation of theories of probability. *Biometrika, 36,* 101–116.

Kleiter, G. D. (1981). *Bayes Statistik.* Berlin: De Gruyter.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Krüger, L., Daston, L., & Heidelberger, M. (Eds.). (1987). *The probabilistic revolution: Vol. 1. Ideas in history.* Cambridge, MA: MIT Press.

Krüger, L., Gigerenzer, G., & Morgan, M. S. (Eds.). (1987). *The probabilistic revolution: Vol. 2. Ideas in the sciences.* Cambridge, MA: MIT Press.

Levi, I. (1983). Who commits the base rate fallacy? *Behavioral and Brain Sciences, 6,* 502–506.

Lindquist, E. F. (1940). *Statistical analysis in educational research.* Boston: Houghton Mifflin.

Lovie, A. D. (1979). The analysis of variance in experimental psychology: 1934-1945. *British Journal of Mathematical and Statistical Psychology, 32,* 151–178.

Luce, R. D. (1988). The tools-to-theory hypothesis. Review of G. Gigerenzer and D. J. Murray, "Cognition as intuitive statistics." *Contemporary Psychology, 33,* 582–583.

Luce, R. D. (1989). Autobiography. In G. Lindzey (Ed.), *Psychology in autobiography* (Vol. 8, pp. 245–289). Stanford: Stanford University Press.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology, 64,* 553–557.

Miller, G. A., & Buckhout, R. (1973). *Psychology: The science of mental life.* New York: Harper & Row.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy.* Chicago: Aldine.

Neyman, J. (1950). *First course in probability and statistics.* New York: Holt.

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *International Statistical Review, 25,* 7–22.

Nunnally, J. C. (1975). *Introduction to statistics for psychology and education.* New York: McGraw-Hill.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.

Pearson, E. S. (1939). "Student" as statistician. *Biometrika, 30,* 210–250.

Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics, 33,* 394–403.

Pollard, P., & Richardson, J. T. E. (1987). On the probability of making Type I errors. *Psychological Bulletin, 102,* 159–163.

Porter, T. M. (1986). *The rise of statistical thinking, 1820-1900.* Princeton, NJ: Princeton University Press.

Publication Manual of the American Psychological Association. (1974) (2nd ed.). Baltimore: Garamond/Pridemark Press.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Rucci, A. J., & Tweney, R. D. (1980). Analysis of variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin, 87,* 166–184.

Schwartz, S., & Dangleish, L. (1982). Statistical inference in personality research. *Journal of Research in Personality, 16,* 290–302.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Skinner, B. F. (1972). *Cumulative record.* New York: Appleton-Century-Crofts.

Skinner, B. F. (1984). *A matter of consequences.* New York: New York University Press.

Sterling, R. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

"Student" [W. S. Gosset]. (1908). The probable error of a mean. *Biometrika, 6,* 1–25.

Swijtink, Z. G. (1987). The objectification of observation: Measurement and statistical methods in the nineteenth century. In L. Krüger, L. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Vol. 1. Ideas in history* (pp. 261–285). Cambridge, MA: MIT Press.

Tholey, P. (1982). Signifikanztest und Bayessche Hypothesenprüfung. *Archiv für Psychologie, 134,* 319–342.

Thomas, D. H. (1978). The awful truth about statistics in archaeology. *American Antiquity, 43,* 231–244.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (Vol. 1, pp. 49–72). Hillsdale, NJ: Lawrence Erlbaum Associates.

Winer, B. J. (1962). *Statistical principles in experimental design* (2nd ed., 1971). New York: McGraw-Hill.