Things I Have Learned (So Far)

Jacob Cohen *New York University*

ABSTRACT: This is an account of what I have learned (so far) about the application of statistics to psychology and the other sociobiomedical sciences. It includes the principles "less is more" (fewer variables, more highly targeted issues, sharp rounding off), "simple is better" (graphic representation, unit weighting for linear composites), and "some things you learn aren't so." I have learned to avoid the many misconceptions that surround Fisherian null hypothesis testing. I have also learned the importance of power analysis and the determination of just how big (rather than how statistically significant) are the effects that we study. Finally, I have learned that there is no royal road to statistical induction, that the informed judgment of the investigator is the crucial element in the interpretation of data, and that things take time.

What I have learned (so far) has come from working with students and colleagues, from experience (sometimes bitter) with journal editors and review committees, and from the writings of, among others, Paul Meehl, David Bakan, William Rozeboom, Robyn Dawes, Howard Wainer, Robert Rosenthal, and more recently, Gerd Gigerenzer, Michael Oakes, and Leland Wilkinson. Although they are not always explicitly referenced, many of you will be able to detect their footprints in what follows.

Some Things You Learn Aren't So

One of the things I learned early on was that some things you learn aren't so. In graduate school, right after World War II, I learned that for doctoral dissertations and most other purposes, when comparing groups, the proper sample size is 30 cases per group. The number *30* seems to have arisen from the understanding that with fewer than 30 cases, you were dealing with "small" samples that required specialized handling with "small-sample statistics" instead of the critical-ratio approach we had been taught. Some of us knew about these exotic small-sample statistics—in fact, one of my fellow doctoral candidates undertook a dissertation, the distinguishing feature of which was a sample of only 20 cases per group, so that he could demonstrate his prowess with small-sample statistics. It wasn't until some years later that I discovered (mind you, not invented) power analysis, one of whose fruits was the revelation that for a two-independent-group-mean comparison with $n = 30$ per group at the sanctified two-tailed .05 level, the probability that a medium-sized effect would be labeled as significant by the most modern methods (a t test) was only .47. Thus, it was approximately a coin flip whether one would get a significant result, even though, in reality, the effect size was meaningful. My $n = 20$ friend's power was rather worse (.33), but of course he couldn't know that, and he ended up with nonsignificant results—with which he proceeded to demolish an important branch of psychoanalytic theory.

Less Is More

One thing I learned over a long period of time that *is* so is the validity of the general principle that *less is more*, except of course for sample size (Cohen & Cohen, 1983, pp. 169–171). I have encountered too many studies with prodigious numbers of dependent variables, or with what seemed to me far too many independent variables, or (heaven help us) both.

In any given investigation that isn't explicitly exploratory, we should be studying few independent variables and even fewer dependent variables, for a variety of reasons.

If all of the dependent variables are to be related to all of the independent variables by simple bivariate analyses or multiple regression, the number of hypothesis tests that will be performed willy-nilly is at least the product of the sizes of the two sets. Using the .05 level for many tests escalates the experimentwise Type I error rate—or in plain English, greatly increases the chances of discovering things that aren't so. If, for example, you study 6 dependent and 10 independent variables and should find that your harvest yields 6 asterisks, you know full well that if there were no real associations in any of the 60 tests, the chance of getting one or more "significant" results is quite high (something like $1 - .95^{60}$, which equals, coincidentally, .95), and that you would expect three spuriously significant results on the average. You then must ask yourself some embarrassing questions, such as, Well, which three are real?, or even, Is six significant *significantly* more than the chance-expected three? (It so happens that it isn't.)

And of course, as you've probably discovered, you're not likely to solve your multiple tests problem with the Bonferroni maneuver. Dividing .05 by 60 sets a per-test significance criterion of $.05/60 = 0.00083$, and therefore a critical two-sided t value of about 3.5. The effects you're dealing with may not be large enough to produce any interesting ts that high, unless you're lucky.

Nor can you find salvation by doing six stepwise multiple regressions on the 10 independent variables. The amount of capitalization on chance that this entails is more than I know how to compute, but certainly more than would a simple harvest of asterisks for 60 regression coefficients (Wilkinson, 1990, p. 481).

In short, the results of this humongous study are a muddle. There is no solution to your problem. You

 0003-066X/90/$00.75
Vol. 45, No. 12, 1304-1312

wouldn't, of course, write up the study for publication as if the unproductive three quarters of your variables never existed. . . .

The irony is that people who do studies like this often start off with some useful central idea that, if pursued modestly by means of a few highly targeted variables and hypotheses, would likely produce significant results. These could, if propriety or the consequences of early toilet training deemed it necessary, successfully withstand the challenge of a Bonferroni or other experimentwise-adjusted alpha procedure.

A special case of the too-many-variables problem arises in multiple regression–correlation analysis with large numbers of independent variables. As the number of independent variables increases, the chances are that their redundancy in regard to criterion relevance also increases. Because redundancy increases the standard errors of partial regression and correlation coefficients and thus reduces their statistical significance, the results are likely to be zilch.

I have so heavily emphasized the desirability of working with few variables and large sample sizes that some of my students have spread the rumor that my idea of the perfect study is one with 10,000 cases and no variables. They go too far.

A less profound application of the less-is-more principle is to our habits of reporting numerical results. There are computer programs that report by default four, five, or even more decimal places for all numerical results. Their authors might well be excused because, for all the programmer knows, they may be used by atomic scientists. But we social scientists should know better than to report our results to so many places. What, pray, does an $r = .12345$ mean? or, for an IQ distribution, a mean of 105.6345? For $N = 100$, the standard error of the r is about .1 and the standard error of the IQ mean about 1.5. Thus, the *345* part of $r = .12345$ is only 3% of its standard error, and the *345* part of the IQ mean of 105.6345 is only 2% of its standard error. These superfluous decimal places are no better than random numbers. They are actually worse than useless because the clutter they create, particularly in tables, serves to distract the eye and mind from the necessary comparisons among the meaningful leading digits. Less is indeed more here.

Simple Is Better

I've also learned that simple is better, which is a kind of loose generalization of less is more. The simple-is-better idea is widely applicable to the representation, analysis, and reporting of data.

If, as the old cliché has it, a picture is worth a thousand words, in describing a distribution, a frequency polygon or, better still, a Tukey (1977, pp. 1–26) stem and leaf diagram is usually worth more than the first four moments, that is, the mean, standard deviation, skewness, and kurtosis. I do not question that the moments efficiently summarize the distribution or that they are useful in some analytic contexts. Statistics packages eagerly give them to us and we dutifully publish them, but they do not usually make it possible for most of us or most of the consumers of our products to see the distribution. They don't tell us, for example, that there are no cases between scores of 72 and 90, or that this score of 24 is somewhere in left field, or that there is a pile-up of scores of 9. These are the kinds of features of our data that we surely need to know about, and they become immediately evident with simple graphic representation.

Graphic display is even more important in the case of bivariate data. Underlying each product–moment correlation coefficient in an acre of such coefficients there lies a simple scatter diagram that the r presumes to summarize, and well it might. That is, it does so if the joint distribution is more-or-less bivariate normal—which means, among other things, that the relationship must be linear and that there are no wild outlying points. We know that least squares measures, like means and standard deviations, are sensitive to outliers. Well, Pearson correlations are even more so. About 15 years ago, Wainer and Thissen (1976) published a data set made up of the heights in inches and weights in pounds of 25 subjects, for which the r was a perfectly reasonable .83. But if an error in transcription were made so that the height and weight values for one of the 25 subjects were switched, the r would become $-.26$, a rather large and costly error!

There is hardly any excuse for gaps, outliers, curvilinearity, or other pathology to exist in our data unbeknownst to us. The same computer statistics package with which we can do very complicated analyses like quasi-Newtonian nonlinear estimation and multidimensional scaling with Guttman's coefficient of alienation also can give us simple scatter plots and stem and leaf diagrams with which we can see our data. A proper multiple regression/correlation analysis does not begin with a matrix of correlation coefficients, means, and standard deviations, but rather with a set of stem and leaf diagrams and scatter plots. We sometimes learn more from what we see than from what we compute; sometimes what we learn from what we see is that we shouldn't compute, at least not on those data as they stand.

Computers are a blessing, but another of the things I have learned is that they are not an unmixed blessing. Forty years ago, before computers (B.C., that is), for my doctoral dissertation, I did three factor analyses on the 11 subtests of the Wechsler-Bellevue, with samples of 100 cases each of psychoneurotic, schizophrenic, and brain-damaged patients. Working with a pad and pencil, 10-to-the-inch graph paper, a table of products of two-digit numbers, and a Friden electromechanical desk calculator that did square roots "automatically," the whole process took the better part of a year. Nowadays, on a desktop

This invited address was presented to the Division of Evaluation, Measurement, and Statistics (Division 5) at the 98th Annual Convention of the American Psychological Association in Boston, August 13, 1990.

I am grateful for the comments on a draft provided by Patricia Cohen, Judith Rabkin, Raymond Katzell, and Donald F. Klein.

Correspondence concerning this article should be addressed to Jacob Cohen, Department of Psychology, New York University, 6 Washington Pl., 5th Floor, New York, NY 10003.

computer, the job is done virtually in microseconds (or at least lickety-split). But another important difference between then and now is that the sheer laboriousness of the task assured that throughout the entire process I was in intimate contact with the data and their analysis. There was no chance that there were funny things about my data or intermediate results that I didn't know about, things that could vitiate my conclusions.

I know that I sound my age, but don't get me wrong—I love computers and revel in the ease with which data analysis is accomplished with a good interactive statistics package like SYSTAT and SYGRAPH (Wilkinson, 1990). I am, however, appalled by the fact that some publishers of statistics packages successfully hawk their wares with the pitch that it isn't necessary to understand statistics to use them. But the same package that makes it possible for an ignoramus to do a factor analysis with a pull-down menu and the click of a mouse also can greatly facilitate with awesome speed and efficiency the performance of simple and informative analyses.

A prime example of the simple-is-better principle is found in the compositing of values. We are taught and teach our students that for purposes of predicting a criterion from a set of predictor variables, assuming for simplicity (and as the mathematicians say, "with no loss of generality"), that all variables are standardized, we achieve maximum linear prediction by doing a multiple regression analysis and forming a composite by weighting the predictor z scores by their betas. It can be shown as a mathematical necessity that with these betas as weights, the resulting composite generates a higher correlation with the criterion in the sample at hand than does a linear composite formed using any other weights.

Yet as a practical matter, most of the time, we are better off using unit weights: +1 for positively related predictors, −1 for negatively related predictors, and 0, that is, throw away poorly related predictors (Dawes, 1979; Wainer, 1976). The catch is that the betas come with guarantees to be better than the unit weights only for the sample on which they were determined. (It's almost like a TV set being guaranteed to work only in the store.) But the investigator is not interested in making predictions for that sample—he or she *knows* the criterion values for those cases. The idea is to combine the predictors for maximal prediction for *future* samples. The reason the betas are not likely to be optimal for future samples is that they are likely to have large standard errors. For the typical 100 or 200 cases and 5 or 10 correlated predictors, the unit weights will work as well or better.

Let me offer a concrete illustration to help make the point clear. A running example in our regression text (Cohen & Cohen, 1983) has for a sample of college faculty their salary estimated from four independent variables: years since PhD, sex (coded in the modern manner—1 for female and 0 for male), number of publications, and number of citations. The sample multiple correlation computes to .70. What we want to estimate is the correlation we would get if we used the sample beta weights in the population, the cross-validated multiple correlation, which unfortunately shrinks to a value smaller than the shrunken multiple correlation. For $N = 100$ cases, using Rozeboom's (1978) formula, that comes to .67. Not bad. But using unit weights, we do better: .69. With 300 or 400 cases, the increased sampling stability pushes up the cross-validated correlation, but it remains slightly smaller than the .69 value for unit weights. Increasing sample size to 500 or 600 will increase the cross-validated correlation in this example to the point at which it is larger than the unit-weighted .69, but only trivially, by a couple of points in the *third* decimal! When sample size is only 50, the cross-validated multiple correlation is only .63, whereas the unit weighted correlation remains at .69. The sample size doesn't affect the unit weighted correlation because we don't estimate unstable regression coefficients. It is, of course, subject to sampling error, but so is the cross-validated multiple correlation.

Now, unit weights will not always be as good or better than beta weights. For some relatively rare patterns of correlation (suppression is one), or when the betas vary greatly relative to their mean, or when the ratio of sample size to the number of predictors is as much as 30 to 1 and the multiple correlation is as large as .75, the beta weights may be better, but even in these rare circumstances, probably not much better.

Furthermore, the unit weights work well outside the context of multiple regression where we have criterion data—that is, in a situation in which we wish to measure some concept by combining indicators, or some abstract factor generated in a factor analysis. Unit weights on standardized scores are likely to be better for our purposes than the factor scores generated by the computer program, which are, after all, the fruits of a regression analysis for that sample of the variables on the factor as criterion.

Consider that when we go to predict freshman grade point average from a 30-item test, we don't do a regression analysis to get the "optimal" weights with which to combine the item scores—we just add them up, like Galton did. Simple *is* better.

We are, however, *not* applying the simple-is-better principle when we "simplify" a multivalued graduated variable (like IQ, or number of children, or symptom severity) by cutting it somewhere along its span and making it into a dichotomy. This is sometimes done with a profession of modesty about the quality or accuracy of the variable, or to "simplify" the analysis. This is not an application, but rather a perversion of simple is better, because this practice is one of willful discarding of information. It has been shown that when you so mutilate a variable, you typically reduce its squared correlation with other variables by about 36% (Cohen, 1983). Don't do it. This kind of simplification is of a piece with the practice of "simplifying" a factorial design ANOVA by reducing all cell sizes to the size of the smallest by dropping cases. They are both ways of throwing away the most precious commodity we deal with: information.

Rather more generally, I think I have begun to learn how to use statistics in the social sciences.

The atmosphere that characterizes statistics as ap-

plied in the social and biomedical sciences is that of a secular religion (Salsburg, 1985), apparently of Judeo-Christian derivation, as it employs as its most powerful icon a six-pointed cross, often presented multiply for enhanced authority. I confess that I am an agnostic.

The Fisherian Legacy

When I began studying statistical inference, I was met with a surprise shared by many neophytes. I found that if, for example, I wanted to see whether poor kids estimated the size of coins to be bigger than did rich kids, after I gathered the data, I couldn't test this research hypothesis, but rather the null hypothesis that poor kids perceived coins to be the same size as did rich kids. This seemed kind of strange and backward to me, but I was rather quickly acculturated (or, if you like, converted, or perhaps brainwashed) to the Fisherian faith that science proceeds only through inductive inference and that inductive inference is achieved chiefly by rejecting null hypotheses, usually at the .05 level. (It wasn't until much later that I learned that the philosopher of science, Karl Popper, 1959, advocated the formulation of falsifiable *research* hypotheses and designing research that could falsify *them*.)

The fact that Fisher's ideas quickly became *the* basis for statistical inference in the behavioral sciences is not surprising—they were very attractive. They offered a deterministic scheme, mechanical and objective, independent of content, and led to clear-cut yes–no decisions. For years, nurtured on the psychological statistics textbooks of the 1940s and 1950s, I never dreamed that they were the source of bitter controversies (Gigerenzer & Murray, 1987).

Take, for example, the yes–no decision feature. It was quite appropriate to agronomy, which was where Fisher came from. The outcome of an experiment can quite properly be the decision to use this rather than that amount of manure or to plant this or that variety of wheat. But we do not deal in manure, at least not knowingly. Similarly, in other technologies—for example, engineering quality control or education—research is frequently designed to produce decisions. However, things are not quite so clearly decision-oriented in the development of scientific theories.

Next, consider the sanctified (and sanctifying) magic .05 level. This basis for decision has played a remarkable role in the social sciences and in the lives of social scientists. 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. Its arbitrary unreasonable tyranny has led to data fudging of varying degrees of subtlety from grossly altering data to dropping cases where there "must have been" errors.

The Null Hypothesis Tests Us

We cannot charge R. A. Fisher with all of the sins of the last half century that have been committed in his name (or more often anonymously but as part of his legacy), but they deserve cataloging (Gigerenzer & Murray, 1987; Oakes, 1986). Over the years, I have learned not to make errors of the following kinds:

When a Fisherian null hypothesis is rejected with an associated probability of, for example, .026, it is *not* the case that the probability that the null hypothesis is true is .026 (or less than .05, or any other value we can specify). Given our framework of probability as long-run relative frequency—as much as we might wish it to be otherwise—this result does not tell us about the truth of the null hypothesis, given the data. (For this we have to go to Bayesian or likelihood statistics, in which probability is not relative frequency but degree of belief.) What it tells us is the probability of the data, given the truth of the null hypothesis—which is not the same thing, as much as it may sound like it.

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. In fact, there *is* no alternate hypothesis in Fisher's scheme: Indeed, he violently opposed its inclusion by Neyman and Pearson.

Despite widespread misconceptions to the contrary, the rejection of a given null hypothesis gives us no basis for estimating the probability that a replication of the research will again result in rejecting that null hypothesis.

Of course, everyone knows that failure to reject the Fisherian null hypothesis does not warrant the conclusion that it is true. Fisher certainly knew and emphasized it, and our textbooks duly so instruct us. Yet how often do we read in the discussion and conclusions of articles now appearing in our most prestigious journals that "there is no difference" or "no relationship"? (This is 40 years after my $N = 20$ friend used a nonsignificant result to demolish psychoanalytic theory.)

The other side of this coin is the interpretation that accompanies results that surmount the .05 barrier and achieve the state of grace of "statistical significance." "Everyone" knows that all this means is that the effect is not nil, and nothing more. Yet how often do we see such a result to be taken to mean, at least implicitly, that the effect is *significant*, that is, *important, large*. If a result is *highly* significant, say $p < .001$, the temptation to make this misinterpretation becomes all but irresistible.

Let's take a close look at this null hypothesis—the fulcrum of the Fisherian scheme—that we so earnestly seek to negate. A null hypothesis is any precise statement about a state of affairs in a population, usually the value of a parameter, frequently zero. It is called a "null" hypothesis because the strategy is to nullify it or because it means "nothing doing." Thus, "The difference in the mean scores of U.S. men and women on an Attitude Toward the U.N. scale is zero" is a null hypothesis. "The product–moment r between height and IQ in high school students is zero" is another. "The proportion of men in a population of adult dyslexics is .50" is yet another. Each is a precise statement—for example, if the population r

between height and IQ is in fact .03, the null hypothesis that it is zero is false. It is also false if the r is .01, .001, or .000001!

A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is *always* false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false). If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it?

Another problem that bothered me was the asymmetry of the Fisherian scheme: If your test exceeded a critical value, you could conclude, subject to the alpha risk, that your null was false, but if you fell short of that critical value, you couldn't conclude that the null was true. In fact, all you could conclude is that you *couldn't* conclude that the null was false. In other words, you could hardly conclude anything.

And yet another problem I had was that if the null were false, it had to be false to some degree. It had to make a difference whether the population mean difference was 5 or 50, or whether the population correlation was .10 or .30, and this was not taken into account in the prevailing method. I had stumbled onto something that I learned after awhile was one of the bases of the Neyman–Pearson critique of Fisher's system of statistical induction.

In 1928 (when I was in kindergarten), Jerzy Neyman and Karl Pearson's boy Egon began publishing papers that offered a rather different perspective on statistical inference (Neyman & Pearson, 1928a, 1928b). Among other things, they argued that rather than having a single hypothesis that one either rejected or not, things could be so organized that one could choose between two hypotheses, one of which could be the null hypothesis and the other an alternate hypothesis. One could attach to the precisely defined null an alpha risk, and to the equally precisely defined alternate hypothesis a beta risk. The rejection of the null hypotheses when it was true was an error of the first kind, controlled by the alpha criterion, but the failure to reject it when the alternate hypothesis was true was also an error, an error of the second kind, which could be controlled to occur at a rate beta. Thus, given the magnitude of the difference between the null and the alternate (that is, given the hypothetical population effect size), and setting values for alpha and beta, one could determine the sample size necessary to meet these conditions. Or, with the effect size, alpha, and the sample size set, one could determine the beta, or its complement, the probability of rejecting the null hypothesis, the power of the test.

Now, R. A. Fisher was undoubtedly the greatest statistician of this century, rightly called "the father of modern statistics," but he had a blind spot. Moreover, he was a stubborn and frequently vicious intellectual opponent. A feud with Karl Pearson had kept Fisher's papers out of *Biometrika*, which Karl Pearson edited. After old-man Pearson retired, efforts by Egon Pearson and Neyman to avoid battling with Fisher were to no avail. Fisher wrote that they were like Russians who thought that "pure science" should be "geared to technological performance" as "in a five-year plan." He once led off the discussion on a paper by Neyman at the Royal Statistical Society by saying that Neyman should have chosen a topic "on which he could speak with authority" (Gigerenzer & Murray, 1987, p. 17). Fisher fiercely condemned the Neyman–Pearson heresy.

I was of course aware of none of this. The statistics texts on which I was raised and their later editions to which I repeatedly turned in the 1950s and 1960s presented null hypothesis testing à la Fisher as a done deal, as *the* way to do statistical inference. The ideas of Neyman and Pearson were barely or not at all mentioned, or dismissed as too complicated.

When I finally stumbled onto power analysis, and managed to overcome the handicap of a background with no working math beyond high school algebra (to say nothing of mathematical statistics), it was as if I had died and gone to heaven. After I learned what noncentral distributions were and figured out that it was important to decompose noncentrality parameters into their constituents of effect size and sample size, I realized that I had a framework for hypothesis testing that had four parameters: the alpha significance criterion, the sample size, the population effect size, and the power of the test. For any statistical test, any one of these was a function of the other three. This meant, for example, that for a significance test of a product–moment correlation, using a two-sided .05 alpha criterion and a sample size of 50 cases, if the population correlation is .30, my long-run probability of rejecting the null hypothesis and finding the sample correlation to be significant was .57, a coin flip. As another example, for the same $\alpha = .05$ and population $r = .30$, if I want to have .80 power, I could determine that I needed a sample size of 85.

Playing with this new toy (and with a small grant from the National Institute of Mental Health) I did what came to be called a meta-analysis of the articles in the 1960 volume of the *Journal of Abnormal and Social Psychology* (Cohen, 1962). I found, among other things, that using the nondirectional .05 criterion, the median power to detect a medium effect was .46—a rather abysmal result. Of course, investigators could not have known how underpowered their research was, as their training had not prepared them to know anything about power, let alone how to use it in research planning. One might think that after 1969, when I published my power handbook that made power analysis as easy as falling off a log, the concepts and methods of power analysis would be taken to the hearts of null hypothesis testers. So one might think. (Stay tuned.)

Among the less obvious benefits of power analysis was that it made it possible to "prove" null hypotheses. Of course, as I've already noted, everyone knows that one can't actually prove null hypotheses. But when an inves-

tigator means to prove a null hypothesis, the point is not to demonstrate that the population effect size is, say, zero to a million or more decimal places, but rather to show that it is of no more than negligible or trivial size (Cohen, 1988, pp. 16–17). Then, from a power analysis at, say, $\alpha = .05$, with power set at, say, .95, so that $\beta = .05$, also, the sample size necessary to detect this negligible effect with .95 probability can be determined. Now if the research is carried out using that sample size, and the result is *not* significant, as there had been a .95 chance of detecting this negligible effect, and the effect was *not* detected, the conclusion is justified that no nontrivial effect exists, at the $\beta = .05$ level. This does, in fact, probabilistically prove the intended null hypothesis of no more than a trivially small effect. The reasoning is impeccable, but when you go to apply it, you discover that it takes enormous sample sizes to do so. For example, if we adopt the above parameters for a significance test of a correlation coefficient and $r = .10$ is taken as a negligible effect size, it requires a sample of almost 1,300 cases. More modest but still reasonable demands for power of course require smaller sample sizes, but not sufficiently smaller to matter for most investigators—even .80 power to detect a population correlation of .10 requires almost 800 cases. So it generally takes an impractically large sample size to prove the null hypothesis as I've redefined it; however, the procedure makes clear what it takes to say or imply from the failure to reject the null hypothesis that there is no nontrivial effect.

A salutary effect of power analysis is that it draws one forcibly to consider the magnitude of effects. In psychology, and especially in soft psychology, under the sway of the Fisherian scheme, there has been little consciousness of how big things are. The very popular ANOVA designs yield F ratios, and it is these whose size is of concern. First off is the question of whether they made the sanctifying .05 cut-off and are thus significant, and then how far they fell below this cut-off: Were they perhaps *highly significant* (p less than .01) or *very highly significant* (less than .001)? Because science is inevitably about magnitudes, it is not surprising how frequently p values are treated as surrogates for effect sizes.

One of the things that drew me early to correlation analysis was that it yielded an r, a measure of effect size, which was then translated into a t or F and assessed for significance, whereas the analysis of variance or covariance yielded only an F and told me nothing about effect size. As many of the variables with which we worked were expressed in arbitrary units (points on a scale, trials to learn a maze), and the Fisherian scheme seemed quite complete by itself and made no demands on us to think about effect sizes, we simply had no language with which to address them.

In retrospect, it seems to me simultaneously quite understandable yet also ridiculous to try to develop theories about human behavior with p values from Fisherian hypothesis testing and no more than a primitive sense of effect size. And I wish I were talking about the long, long ago. In 1986, there appeared in the *New York Times* a UPI dispatch under the headline "Children's Height Linked to Test Scores." The article described a study that involved nearly 14,000 children 6 to 17 years of age that reported a *definite* link between height (age- and sex-adjusted) and scores on tests of both intelligence and achievement. The relationship was described as significant, and persisting, even after controlling for other factors, including socioeconomic status, birth order, family size, and physical maturity. The authors noted that the effect was small, but *significant*, and that it didn't warrant giving children growth hormone to make them taller and thus brighter. They speculated that the effect might be due to treating shorter children as less mature, but that there were alternative biological explanations.

Now this was a newspaper story, the fruit of the ever-inquiring mind of a science reporter, not a journal article, so perhaps it is understandable that there was no effort to deal with the actual size of this small effect. But it got me to wondering about how small this significant relationship might be. Well, if we take significant to mean $p < .001$ (in the interest of scientific tough-mindedness), it turns out that a correlation of .0278 is significant for 14,000 cases. But I've found that when dealing with variables expressed in units whose magnitude we understand, the effect size in linear relationships is better comprehended with regression than with correlation coefficients. So, accepting the authors' implicit causal model, it works out that raising a child's IQ from 100 to 130 would require giving the child enough growth hormone to increase his or her height by 14 ft (more or less). If the causality goes the other way, and one wanted to create basketball players, a 4-in. increase in height would require raising the IQ about 900 points. Well, they said it was a small effect. (When I later checked the journal article that described this research, it turned out that the correlation was much larger than .0278. It was actually about .11, so that for a 30-point increase in IQ it would take only enough growth hormone to produce a 3.5-ft increase in height, or with the causality reversed, a 4-in. increase in height would require an increase of only 233 IQ points.)

I am happy to say that the long neglect of attention to effect size seems to be coming to a close. The clumsy and fundamentally invalid box-score method of literature review based on p values is being replaced by effect-size-based meta-analysis as formulated by Gene Glass (1977). The effect size measure most often used is the standardized mean difference d of power analysis. Several book-length treatments of meta-analysis have been published, and applications to various fields of psychology are appearing in substantial numbers in the *Psychological Bulletin* and other prestigious publications. In the typical meta-analysis, the research literature on some issue is surveyed and the effect sizes that were found in the relevant studies are gathered. Note that the observational unit is the study. These data do not only provide an estimate of the level and variability of the effect size in a domain based on multiple studies and therefore on many observations, but by relating effect size to various substantive and methodological characteristics over the stud-

ies, much can be learned about the issue under investigation and how best to investigate it. One hopes that this ferment may persuade researchers to explicitly report effect sizes and thus reduce the burden on meta-analysts and others of having to make assumptions to dig them out of their inadequately reported research results. In a field as scattered (not to say anarchic) as ours, meta-analysis constitutes a welcome force toward the cumulation of knowledge. Meta-analysis makes me very happy.

Despite my career-long identification with statistical inference, I believe, together with such luminaries as Meehl (1978) Tukey (1977), and Gigerenzer (Gigerenzer & Murray, 1987), that hypothesis testing has been greatly overemphasized in psychology and in the other disciplines that use it. It has diverted our attention from crucial issues. Mesmerized by a single all-purpose, mechanized, "objective" ritual in which we convert numbers into other numbers and get a yes–no answer, we have come to neglect close scrutiny of where the numbers came from. Recall that in his delightful parable about averaging the numbers on football jerseys, Lord (1953) pointed out that "the numbers don't know where they came from." But surely *we* must know where they came from and should be far more concerned with why and what and how well we are measuring, manipulating conditions, and selecting our samples.

We have also lost sight of the fact that the error variance in our observations should challenge us to efforts to reduce it and not simply to thoughtlessly tuck it into the denominator of an F or t test.

How To Use Statistics

So, how would I use statistics in psychological research? First of all, descriptively. John Tukey's (1977) *Exploratory Data Analysis* is an inspiring account of how to effect graphic and numerical analyses of the data at hand so as to understand *them*. The techniques, although subtle in conception, are simple in application, requiring no more than pencil and paper (Tukey says if you have a handheld calculator, fine). Although he recognizes the importance of what he calls confirmation (statistical inference), he manages to fill 700 pages with techniques of "mere" description, pointing out in the preface that the emphasis on inference in modern statistics has resulted in a loss of flexibility in data analysis.

Then, in planning research, I think it wise to *plan* the research. This means making tentative informed judgments about, among many other things, the size of the population effect or effects you're chasing, the level of alpha risk you want to take (conveniently, but not necessarily .05), and the power you want (usually some relatively large value like .80). These specified, it is a simple matter to determine the sample size you need. It is then a good idea to rethink your specifications. If, as is often the case, this sample size is beyond your resources, consider the possibility of reducing your power demand or, perhaps the effect size, or even (heaven help us) increasing your alpha level. Or, the required sample may be smaller than you can comfortably manage, which also should lead you to rethink and possibly revise your original specifications. This process ends when you have a credible and viable set of specifications, or when you discover that no practicable set is possible and the research as originally conceived must be abandoned. Although you would hardly expect it from reading the current literature, failure to subject your research plans to power analysis is simply irrational.

Next, I have learned and taught that the primary product of a research inquiry is one or more measures of effect size, not p values (Cohen, 1965). Effect-size measures include mean differences (raw or standardized), correlations and squared correlation of all kinds, odds ratios, kappas—whatever conveys the magnitude of the phenomenon of interest appropriate to the research context. If, for example, you are comparing groups on a variable measured in units that are well understood by your readers (IQ points, or dollars, or number of children, or months of survival), mean differences are excellent measures of effect size. When this isn't the case, and it isn't the case more often than it is, the results can be translated into standardized mean differences (d values) or some measure of correlation or association (Cohen, 1988). (Not that we understand as well as we should the meaning of a given level of correlation [Oakes, 1986, pp. 88–92]. It has been shown that psychologists typically overestimate how much relationship a given correlation represents, thinking of a correlation of .50 not as its square of .25 that its proportion of variance represents, but more like its cube root of about .80, which represents only wishful thinking! But that's another story.)

Then, having found the sample effect size, you can attach a p value to it, but it is far more informative to provide a confidence interval. As you know, a confidence interval gives the range of values of the effect-size index that includes the population value with a given probability. It tells you incidentally whether the effect is significant, but much more—it provides an estimate of the range of values it might have, surely a useful piece of knowledge in a science that presumes to be quantitative. (By the way, I don't think that we should routinely use 95% intervals: Our interests are often better served by more tolerant 80% intervals.)

Remember that throughout the process in which you conceive, plan, execute, and write up a research, it is on your informed judgment as a scientist that you must rely, and this holds as much for the statistical aspects of the work as it does for all the others. This means that your informed judgment governs the setting of the parameters involved in the planning (alpha, beta, population effect size, sample size, confidence interval), and that informed judgment also governs the conclusions you will draw.

In his brilliant analysis of what he called the "inference revolution" in psychology, Gerd Gigerenzer showed how and why no single royal road of drawing conclusions from data is possible, and particularly not one that does not strongly depend on the substantive issues concerned—that is, on everything that went into the research besides the number crunching. An essential ingredient in the re-

search process is the judgment of the scientist. He or she must decide by how much a theoretical proposition has been advanced by the data, just as he or she decided what to study, what data to get, and how to get it. I believe that statistical inference applied with informed judgment is a useful tool in this process, but it isn't the most important tool: It is not as important as everything that came before it. Some scientists, physicists for example, manage without the statistics, although to be sure not without the informed judgment. Indeed, some pretty good psychologists have managed without statistical inference: There come to mind Wundt, Kohler, Piaget, Lewin, Bartlett, Stevens, and if you'll permit me, Freud, among others. Indeed, Skinner (1957) thought of dedicating his book *Verbal Behavior* (and I quote) "to the statisticians and scientific methodologists with whose help this book would never have been completed" (p. 111). I submit that the proper application of statistics by sensible statistical methodologists (Tukey, for example) would not have hurt Skinner's work. It might even have done it some good.

The implications of the things I have learned (so far) are not consonant with much of what I see about me as standard statistical practice. The prevailing yes–no decision at the magic .05 level from a single research is a far cry from the use of informed judgment. Science simply doesn't work that way. A successful piece of research doesn't conclusively settle an issue, it just makes some theoretical proposition to some degree more likely. Only successful future replication in the same and different settings (as might be found through meta-analysis) provides an approach to settling the issue. How much more likely this single research makes the proposition depends on many things, but not on whether p is equal to or greater than .05; .05 is not a cliff but a convenient reference point along the possibility–probability continuum. There is no ontological basis for dichotomous decision making in psychological inquiry. The point was neatly made by Rosnow and Rosenthal (1989) last year in the *American Psychologist*. They wrote "surely, God loves the .06 nearly as much as the .05" (p. 1277). To which I say amen!

Finally, I have learned, but not easily, that things take time. As I've already mentioned, almost three decades ago, I published a power survey of the articles in the 1960 volume of the *Journal of Abnormal and Social Psychology* (Cohen, 1962) in which I found that the median power to detect a medium effect size under representative conditions was only .46. The first edition of my power handbook came out in 1969. Since then, more than two dozen power and effect-size surveys have been published in psychology and related fields (Cohen, 1988, pp. xi–xii). There have also been a slew of articles on power-analytic methodology. Statistics textbooks, even some undergraduate ones, give some space to power analysis, and several computer programs for power analysis are available (e.g., Borenstein & Cohen, 1988). They tell me that some major funding entities require that their grant applications contain power analyses, and that in one of those agencies my power book can be found in every office.

The problem is that, as practiced, current research hardly reflects much attention to power. How often have you seen any mention of power in the journals you read, let alone an actual power analysis in the methods sections of the articles? Last year in *Psychological Bulletin*, Sedlmeier and Gigerenzer (1989) published an article entitled "Do Studies of Statistical Power Have an Effect on the Power of Studies?". The answer was no. Using the same methods I had used on the articles in the 1960 *Journal of Abnormal and Social Psychology* (Cohen, 1962), they performed a power analysis on the 1984 *Journal of Abnormal Psychology* and found that the median power under the same conditions was .44, a little worse than the .46 I had found 24 years earlier. It was worse still (.37) when they took into account the occasional use of an experimentwise alpha criterion. Even worse than that, in some 11% of the studies, research hypotheses were framed as null hypotheses and their nonsignificance interpreted as confirmation. The median power of these studies to detect a medium effect at the two-tailed .05 level was .25! These are not isolated results: Rossi, Rossi, and Cottrill (in press), using the same methods, did a power survey of the 142 articles in the 1982 volumes of the *Journal of Personality and Social Psychology* and the *Journal of Abnormal Psychology* and found essentially the same results.

A less egregious example of the inertia of methodological advance is set correlation, which is a highly flexible realization of the multivariate general linear model. I published it in an article in 1982, and we included it in an appendix in the 1983 edition of our regression text (Cohen, 1982; Cohen & Cohen, 1983). Set correlation can be viewed as a generalization of multiple correlation to the multivariate case, and with it you can study the relationship between anything and anything else, controlling for whatever you want in either the anything or the anything else, or both. I think it's a great method; at least, my usually critical colleagues haven't complained. Yet, as far as I'm aware, it has hardly been used outside the family. (The publication of a program as a SYSTAT supplementary module [Cohen, 1989] may make a difference.)

But I do not despair. I remember that W. S. Gosset, the fellow who worked in a brewery and appeared in print modestly as "Student," published the t test a decade before we entered World War I, and the test didn't get into the psychological statistics textbooks until after World War II.

These things take time. So, if you publish something that you think is really good, and a year or a decade or two go by and hardly anyone seems to have taken notice, remember the t test, and take heart.

REFERENCES

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program*. Hillsdale, NJ: Erlbaum.

Children's height linked to test scores. (October 7, 1986). *New York Times*, p. C4.

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

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.
Cohen, J. (1982). Set correlation as a general multivariate data-analytic method. *Multivariate Behavioral Research, 17,* 301–341.
Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Cohen, J. (1989). *SETCOR: Set correlation analysis, a supplementary module for SYSTAT and SYGRAPH.* Evanston, IL: SYSTAT.
Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.
Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Erlbaum.
Glass, G. V. (1977). Integrating findings: The meta-analysis of research. In L. Shulman (Ed.), *Review of research in education* (Vol. 5, pp. 351–379). Itasca, IL: Peacock.
Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.
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.
Neyman, J., & Pearson, E. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20A,* 175–240.
Neyman, J., & Pearson, E. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika, 20A,* 263–294.
Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences.* New York: Wiley.
Popper, K. (1959). *The logic of scientific discovery.* New York: Basic Books.
Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.
Rossi, J. S., Rossi, S. R., & Cottrill, S. D. (in press). Statistical power in research in social and abnormal psychology. *Journal of Consulting and Clinical Psychology.*
Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin, 85,* 1348–1351.
Salsburg, D. S. (1985). The religion of statistics as practiced in medical journals. *The American Statistician, 39,* 220–223.
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. (1957). *Verbal behavior.* New York: Appleton-Century-Crofts.
Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.
Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213–217.
Wainer, H., & Thissen, D. (1976). When jackknifing fails (or does it?). *Psychometrika, 41,* 9–34.
Wilkinson, L. (1990). *SYSTAT: The system for statistics.* Evanston, IL: SYSTAT.