International Psychogeriatrics, Vol. 12, No. 1, 2000, pp. 3-7

Guest Editorial

Some Basic Issues for Clinicians Concerning Things Statistical

From the title, you may be wondering whether this is another one of those esoteric, unintelligible statistics articles. Or whether the author is another pedantic statistician telling you how much you don't know. I hope that you will find that neither of these fears applies. There are no formulas, derivations, or complex figures here. And I am no "statistician," at least not in the formal sense, although I do teach statistics. Rather, the focus of this article is to identify and explain, for the clinician consumer of the literature, four basic but often overlooked issues in statistics (as practiced in today's medical and social science journals). Although it is true that most clinicians are misinformed about things statistical (Wulff et al., 1987), it is also true that most academic psychologists and physician researchers—and a surprising number of statisticians—are too (Cohen, 1994; Falk & Greenbaum, 1995; Goodman, 1999a; Oakes, 1986).

ISSUE #1: STATISTICAL INFERENCE AND THE *p* VALUE

The most important and pervasive error with regard to statistical inference is the notion that significance tests tell us something about the status of the null hypothesis (i.e., the thing we are trying to reject in our hypothesis testing that says the effect or relationship is zero) for a given study. As many authors have clearly demonstrated, standard hypothesis testing does not tell us whether the null hypothesis is true or false (Cohen, 1990, 1994; Falk & Greenbaum, 1995; Goodman, 1999a). It can't, most obviously because the probabilities against which we evaluate our results—the source of the infamous *p* value—*assume that the null hypothesis is true.* A *p* value of .05 does not mean that the null hypothesis is false with a degree of confidence of 95%. Rather, it means that, assuming the null hypothesis to be true in the population, the obtained results are unlikely (i.e., they would occur only 5 times out of 100 in a *theoretical, population-based* sampling distribution containing all possible results when the null hypothesis is true).

Unfortunately, by the rules of logic, finding unlikely results under the assumption of "no difference" is not the same thing as a false null hypothesis (Cohen, 1994). Probability values, no matter how low, cannot confirm the truth or falsity of our theories and hypotheses.

https://doi.org/10.1017/S1041610200006153 Published online by Cambridge University Press

In standard hypothesis testing, we simply deductively calculate the frequency of every possible outcome of our study (Goodman, 1999a). The results of any single experiment are judged either "likely" or "unlikely," with the *p* value functioning as the index of likelihood (but without any concern for the strength of this "likely" or "unlikely" event—an important omission, as detailed in Issue #2 below).

There are (at least) two other common errors about statistical inference. First, a *p* value of .05 does not mean that 100 replications of the experiment would yield 95 significant results. The power of a study to detect differences must be considered (see Issue #3 below), because power has a strong impact on the likelihood of significant results across replications (Cohen, 1994). Second, the *p* value obtained in a given study should not be confused with the false-positive (Type I) error rate under the null hypothesis (which is set prior to data collection [in practice, generally assumed at .05] and signified by "alpha"). As Goodman (1999a) notes, hypothesis testing was designed to limit errors, including false-positive errors, "over the long run." The *p* value, then, becomes only a tool for evaluating statistical significance relative to alpha; it is *not* a data-specific false-positive error probability. Both of these errors contribute to the apotheosis of the *p* value: the mistaken notion that it tells us everything we need to know about the results of the study (e.g., replicability, error vulnerability).

So for a single study, the obtained *p* value is neither an index of truth, nor an index of replication probability, nor the probability of a false-positive error. It is neither permissible nor desirable, then, to accept conclusions from a study that reflect a "mere linguistic transformation" of the *p* value into a verbal statement (Goodman, 1999a). Although alternatives to the *p* value have been touted that address its limitations, they are beyond the scope and focus of this article (Browner & Newman, 1987; Goodman, 1999b).

ISSUE #2: EFFECT SIZE

What difference does a low *p* value make anyway? The old dictum is worth repeating: Statistical significance does not equal practical significance. Beware of phrases like "highly significant," "marginally significant," or "a trend toward significance." These meaningless phrases are due to our mechanical, slavish devotion to the .05 *p* value cutoff as an indicator of the "realness" of a specific result (Cohen, 1994; Goodman, 1999a), this devotion being a product of the misinterpretations noted in Issue #1 above. As Goodman (1999a) points out, what scientific meaning can be attached to *p* values when a "close" *p* value (e.g., .08) is interpreted as "no difference" when comparing nonequivalent groups at baseline, but "marginally significant" when reporting an expected (and therefore essential) relationship? This dominion of the *p* value leads to a disregard for effect size.

Regrettably, effect size estimates are still rare in the published literature, although appeals for their inclusion have been made repeatedly (Cohen, 1994). Indices of effect size include eta-squared (η^2), omega-squared (ω^2), Cohen's *d*, and—believe it or not—confidence intervals. The first two are "proportion of variance" indices. They range from 0%

to 100% (theoretically), and give us an indication of, for example, the amount of variance in the outcome measure that is associated with (or explained by or shared with) the independent or treatment variable. If you are likening this to a squared correlation coefficient or R^2 in regression, you understand the concept. The latter two indices are based on the standard deviation, and essentially express the outcome in standard deviation units: How big is the obtained difference (e.g., between drug and placebo group) relative to the variability of the outcome measure? This signal-to-noise concept is also represented in the width of confidence intervals.

Effect sizes are not silver bullets, however. Eta-squared and omega-squared are, like means, subject to the vagaries of sampling, and should themselves be reported within confidence intervals (Jaccard & Becker, 1990). Also, effect size estimates tend to overestimate the population effect size; they, like the *p* value, cannot be taken as "truth" (Cohen, 1994).

Nevertheless, for the clinician consumer of the published literature, a consideration of effect size cannot be underestimated. A "highly significant" *p* value (e.g., one study reported a *p* value to the seventh decimal place [Nilsson & Lindahl, 1986]) may in fact be linked with a very weak effect size (e.g., in the case of a very large sample size). Conversely, a "no difference" finding may be linked with a large effect size (e.g., when the sample size is too small, reliability of measurement is low, or the data do not fit the statistic of choice [see Issue #4 below]). There are numerous published studies that report very small *p* values, but when the effect size is calculated, one is left to wonder whether a two-tenths standard deviation difference (or a treatment that leaves 97% of the variance in the outcome measure unexplained) would ever be detectable (or meaningful) in the clinic. This neglect of effect size is inextricably linked with the entrenched but mistaken notion that a low *p* value, by itself for a single study, means something concrete. Effect size must also be considered in light of power (to which I alluded when discussing issues of sample size), which is the next issue.

ISSUE #3: POWER

Like the *p* value, power too is a probability. Empirically, it is the probability that a study will yield a *p* value of less than alpha (at most .05, by convention), for any given effect size. Conceptually, it is the probability of correctly rejecting a false null hypothesis. Although this conceptual definition sounds important, Cohen and others (Cohen, 1990, 1994; Tukey, 1991) have pointed out that the null hypothesis is, for all intents and purposes, always false! This is meant only partly facetiously, and illustrates the issue of power: No matter how tiny the effect size for a given analysis, there is some sample size large enough to yield a *p* value of less than .05 (in the nearly universal case where the null hypothesis assumes an effect size of 0). But there is more: Conversely, a study can have too little power, which is a problem not when the effect size is tiny but when it is substantial.

The important message for the clinician consumer of published research is to understand that a study can be too powerful or too weak. Both of these

conditions are testament to the often (of necessity) haphazard nature of clinical and social science research. The clear benefit of an a priori power analysis is that it forces the researcher to take effect size into consideration in the planning of the study (Berry et al., 1998). By convention (Cohen, 1988), a study should have an 80% probability of generating a p value of less than .05 (or .01 or wherever alpha is set) for an effect size that: (a) is suggested by previous research, and/or (b) represents a minimum meaningful difference (Aron & Aron, 1994) (i.e., what is the smallest effect size that has clinical relevance, meaning, or practical application?). A power analysis, through its association with effect size, also forces the researcher to face the fact that—in the often cross-sectional, correlational world of clinical and social science research—lots of things are related to lots of other things, often spuriously and arbitrarily (and substantially, by social science standards) (Cohen, 1994). Effect size estimates across studies and power analyses within studies help us to more carefully consider what of true interest might actually be happening in research.

ISSUE #4: RESPECTING THE DATA

The final issue is about matching the statistical technique to the data at hand. Too often, the statistical techniques that are most familiar or most available (the advent of statistical analysis packages for the PC has done more to produce awful—but very sophisticated—data analyses than any other single event) are forced on an unsuspecting data set, inevitably justified with recourse to the notion that such-and-such a statistic "is robust to violations of its assumptions." On a simpler level, the characteristics of our data, particularly in relation to our study design (e.g., sample/cell sizes), are often ignored, for example with respect to skewness, outliers, variance equality, or ceiling/floor effects (Tabachnik & Fidell, 1996). We ignore these characteristics to the detriment of the interpretability of our results, which is also a symptom of our reliance on the p value to tell us the truth. There are numerous published articles that report the mean as an indicator of central tendency, but where the standard deviation exceeds the mean by a factor of 2 or more. Other articles report means for scales that can, for example, theoretically range from 0 to 30, but where the mean is 1. I recall reviewing a study using logistic regression that made all manner of conclusions about risk factors for the disease under consideration (and where statistically significant odds ratios were found), but where, after the multi-way frequencies were worked out (by me, not by the authors), the entire edifice of conclusions was based on a single case! Issues of robustness aside, how interpretable are these results? The clinician, as consumer of the published literature, must clearly be aware of the nature of the data being analyzed before accepting any of the study conclusions.

SUMMARY

It is hoped that by consideration of these four issues, the interested clinician can more critically and knowledgeably digest what appears in the published literature. Rest assured, there are no perfect

studies, and outlining the rules as I have done above does not mean that I consistently adhere to them (far from it). Nevertheless, knowing the issues and critiquing the research forces a recognition of our limitations, and is the next best thing to actually putting all of this into practice on a routine basis.

John T. Chibnall, PhD
Department of Psychiatry
Saint Louis University
School of Medicine
St. Louis, MO, USA

REFERENCES

Aron, A., & Aron, E. N. (1994). *Statistics for psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Berry, E. M., Coustere-Yakir, C., & Grover, N. B. (1998). The significance of non-significance. *QJM, 91,* 647-653.

Browner, W. S., & Newman, T. B. (1987). Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *Journal of the American Medical Association, 257,* 2459-2463.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cohen, J. (1994). The world is round ($p < .05$). *American Psychologist, 49,* 997-1003.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5,* 75-98.

Goodman, S. N. (1999a). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine, 130,* 995-1004.

Goodman, S. N. (1999b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine, 130,* 1005-1013.

Jaccard, J., & Becker, M. A. (1990). *Statistics for the behavioral sciences* (2nd ed.). Belmont, CA: Wadsworth.

Nilsson, N., & Lindahl, O. (1986). The effect of nitrogen fertilization on forest blueberries. *Nutrition and Health, 4,* 151-153.

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

Tabachnik, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100-116.

Wulff, H. R., Andersen, B., Brandenhoff, P., & Guttler, E. (1987). What do doctors know about statistics? *Statistics in Medicine, 6,* 3-10.