STATISTICALLY Speaking

The Importance of Meaning

In the last incarnation of this column, we were reminded of the importance of accuracy, narrowly operationalized as understanding the difference between the population parameter (β) and our estimate of that parameter through regression. Here, we will remind ourselves of the importance of meaning, operationalized as understanding the meaning of the magnitude of the difference in addition to its statistical significance.

In 1951, a quarter century after the "invention" of Fisherian inference, Frank Yates commented on perhaps the unintended consequences of such an approach. In a section of his paper called "Present Trends," he writes that the emphasis given to formal tests of significance in Fisherian inference "has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data . . . and too little to the estimates of the magnitude of the effect they are investigating."[1(p32)] A half century after that, Joe Fleiss et al. reminded us that "no matter how small the difference is . . . —as long as it is non-zero—samples of sufficiently large size can virtually guarantee statistical significance."[2(p64)] So my discourse here is not new, but rather, a reminder that "eternal vigilance" is necessary to keep our science as strong as possible.

The roots of this discussion begin in the concepts of power analysis, which is usually expressed through a family of conceptually simple formulas consisting of only 4 ingredients: f(α), the numerical "cutoff" value of the appropriate distribution corresponding to the prespecified type I error rate (often set at 5%); f(β), the numerical "cutoff" value of the appropriate distribution corresponding to the prespecified type II error rate (often set at 20%); Δ, the level of change in the outcome measure that is important to detect; and n, the number of experimental units (often, human subjects). The investigators' garden-variety encounter with a statistician is to ask him or her to do a "power analysis," which is typically interpreted as solving the equation by putting n on the left side of the equation, and the other 3 variables to the right of the equal sign. The equation is then solved for the number of units necessary in order to detect the effect size, with prespecified power, while keeping type I error to 5% (or less).

However, those same investigators usually do not come back once the study is over, after only half or so of the required patients were actually recruited, or after the online survey produced 10 000 responses rather than the 1000 originally specified, to recalculate the detectable effect size. These examples leave the former investigator to perhaps declare that their intervention "didn't work," even if it produced meaningful change, and leading the latter investigator to declare that even trivial improvements or associations are "significant" and, therefore, important—and so, our discussion of the importance of meaning!

We are probably more sympathetic to the first investigator but see it less in the

"We'll double our chances of recovery if we buy two lottery tickets."

Courtesy of The New Yorker collection at cartoonbank.com

Source. Cartoon by Leo Cullum. Available at: http://www.cartoonbank.com.

literature. Although Rosnow and Rosenthal might have been right, that "surely God loves the .06 nearly as much as the .05,"[3(p1277)] reviewers and journal editors tend to love $P<.05$ more. The plight of these investigators deserves an example. Let's take a common measure of association: the correlation coefficient.

Let's assume that Investigator A observes a sizable correlation between 2 variables of, say, $r=0.50$, meaning that one quarter (25%) of the variability is shared between the 2, or that one variable "explains" one quarter of the variance in the other. We would probably agree that this is a meaningful effect size, and need measurement of these variables on only 20 or so units (i.e., people) before we achieve a small *P* value and declare it "statistically significant." So we have *observed a meaningful effect size* and declared it to be significant. But if Investigator B replicated that experiment with only 15 people and observed the exact same correlation ($r=0.50$), he or she would compute a *P* value of .06 and fail to reject the null hypothesis, referring to this as nonsignificant, and perhaps including an "NS" in the significance column of the table. But is this effect size any less meaningful?

Now let's change the correlation that these 2 investigators observe, and change the number of people they enroll. Using our Internet data collection example, let's say that Investigator A collects information on $n=1000$ respondents and observes a correlation of $r=0.02$, meaning that only 4/100ths of 1% of the variance is shared between the 2 variables. We would probably agree that this is not a meaningful effect size. This investigator would have also computed a very large *P* value and would not have declared this correlation significant. Enter Investigator B, lucky enough to have collected information from around 10 000 respondents and observed the same $r=0.02$. However, with a sample size this big, the computed *P* value would enable this investigator to declare this same correlation ($r=0.02$) significant and to possibly discuss it as important! But the effect size is the same 4/100ths of 1% of the variance, and the declaration of significance is merely a function of the very large sample size. It did not become meaningful just because it is now significantly different from zero.

So where do these examples lead us? To avoid declaring trivial associations as "important," we should begin our investigations by determining (and publishing) the magnitude of an effect that we and our colleagues consider "important"; that is, will the observed change or association benefit public health, medicine, policy, or advance any theory? Clearly, this determination is strongly situation dependent—what is considered "meaningful" for an increase in knowledge may not be the same as what is considered "meaningful" for an increase in lifespan. Then, design a study with the appropriate sample size required to assure a reasonable probability of detecting that difference. It is also worth the effort to calculate the minimum effect size that a study using your actual (vs planned) sample size could detect, so that the interpretation of your results can be in context, keeping *meaning* separate from *significance*.

The cartoon on the previous page allows us to laugh at the absurdity of the businessmen's hopes of saving their company by winning the lottery, even though they are indeed "doubling" their chances of winning by buying two tickets, just as students of public health realize that the magnitude of a calculated odds ratio is relatively meaningless without the context of the underlying prevalence. I hope that laughing at this cartoon will enable us to ensure that we are not replicating this hilarity in our own research. ■

Roger D. Vaughan, DrPH, MS

About the Authors

The author is the Journal's associate editor for statistics and evaluation.

Request for reprints should be sent to Roger D. Vaughan, DrPH, MS, Columbia University, Mailman School of Public Health, Department of Biostatistics, 722 West 168th Street, 6th Floor, New York, NY 10032 (e-mail: rdv2@columbia.edu).

doi:10.2105/AJPH.2006.105379

References

1. Yates F. The influence of Statistical Methods for Research Workers on the development of the science of statistics. *J Am Stat Assoc.* 1951;46:19–34.

2. Fleiss J, Levin B, Paik MC. *Statistical Methods for Rates and Proportions.* 3rd ed. Hoboken, NJ: John Wiley and Sons; 2003.

3. Rosnow R, Rosenthal R. Statistical procedures and the justification of knowledge in psychological science. *Am Psychol.* 1989;44:1276–1284.