COMMUNICATION METHODS AND MEASURES, *1*(4), 291–299

Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses

Daniel J. O'Keefe
Northwestern University

This brief note hopes to dispel some apparent confusions concerning statistical power, especially ones stemming from power computations based on treating an observed sample effect size as the population effect of interest ("post hoc," "observed," or "retrospective" power). Some after-the-fact power analyses can be useful for presentation purposes but only when based on population effect sizes of independent interest. It is recommended that labels such as "post hoc power," "observed power," "retrospective power," and "a priori power" be avoided, that reported power figures be accompanied by specification and justification of the values used to compute power, and that results characteristically be described with effect sizes, confidence intervals, and *p* values.

BACKGROUND

The power of a statistical test is the likelihood that it will produce a statistically significant result. Broadly put, the power of a test is a function of three factors:

Thanks to Tim Levine, Rene Weber, and Hee Sun Park for useful commentary. Portions of this article appeared in CRTNET #9692 (archived at: http://lists1.cac.psu.edu/cgi-bin/wa?S1=crtnet).

Correspondence should be addressed to Professor Daniel J. O'Keefe, Department of Communication Studies, Frances Searle Building, Northwestern University, 2240 Campus Drive, Evanston, IL 60208-3545. E-mail: d-okeefe@northwestern.edu

1. the significance criterion (e.g., everything else being equal, a given test has more power if the significance level is .05 rather than .01),
2. the sample size (everything else being equal, as *N* increases, the power of the test increases), and
3. the population effect size (everything else being equal, as the population effect size increases, so does the power of a test—strong signals are easier to pick up than weak ones are).

These four variables—power, significance criterion, sample size, and population effect size—are related such that when the values of three of these are fixed, the fourth is determined (Cohen, 1988, p. 14).[1]

A question that naturally arises is: Given that one doesn't *know* what the population effect size is, how can power be computed? The answer is that power is computed for a given *potential* population effect size—and indeed can be computed for several different potential population effect sizes. For example, with .05 alpha and a two-tailed test, a sample size of 50 provides power of .83 given a population correlation of .40 but provides power of only .29 given a population correlation of .20.

Thus it can be misleading to speak of *the* power of a given statistical test, because any particular test (i.e., with a specified sample size) actually has many different power values. A sample size of 50, for example, gives power of .03 (given a population correlation of .10, .01 alpha, and a two-tailed test), .99 (given a population correlation of .50, .10 alpha, and a one-tailed test), .57 (given a population correlation of .30, .05 alpha, and a two-tailed test), and so on.

POST HOC, RETROSPECTIVE, OBSERVED, ACHIEVED, A PRIORI, AND PROSPECTIVE POWER

One sometimes sees reference made to "post hoc power," "observed power," "retrospective power," "achieved power," "prospective power," and "a priori power"—for a variety of purposes and with a variety of meanings (compare, e.g., Gillett, 1994, and Onwuegbuzie & Leech, 2004). If the phrase post hoc power (or retrospective power) is understood as referring to "the power of the test when

[1]In the interest of simplifying matters, this list of factors influencing power is incomplete; for example, it omits influences such as measurement unreliability (see, e.g., Cohen, 1988, pp. 535–537; Williams & Zimmerman, 1989). Although such other factors are put aside here, it should be remembered that any factor (such as unreliability) that "increases the variability of the observations beyond their necessary 'true' variability also . . . reduces power" (Cohen, 1988, p. 8). (The present discussion is simplified in other ways as well. For example, it runs together null hypotheses and nil hypotheses, because the complexities attendant to that distinction might obscure the larger points under discussion.)

computed after the test is done," the phrase is potentially misleading. The power of a test is the same (ceteris paribus) no matter when the power is computed—before or after the test is done. Understood in this way, the power of a test after-the-fact (post hoc or retrospective power) is exactly the same as the power of the test before-the-fact (a priori or prospective power), everything else being equal (same significance criterion, same sample size, same population effect size).

But sometimes post hoc power is used with a more specific meaning, namely, "the power of the test assuming a population effect size exactly equal to the effect size observed in the current sample." (In SPSS, this is called "observed power." One can also see "retrospective power" and "achieved power" used with this meaning.) That is, a researcher collects data, obtains a nonsignificant effect, computes the sample effect size, and then computes power on the basis of (1) the significance criterion that was used, (2) the sample size that was used, and (3) a population effect size equal to that observed in the sample (e.g., Onwuegbuzie & Leech, 2004). So, for example, imagine a study with $N=80$ that found a nonsignificant sample correlation of .13, two-tailed $p=.25$; the investigator would report post hoc (observed, achieved, retrospective) power of .21 (assuming .05 alpha, two-tailed test).

However, it is not clear why such a power value should be of interest to anyone. First, given a nonsignificant result, one already knows that the observed statistical power (the power for detecting a population effect equal to the obtained sample effect) is low. As Hoenig and Heisey (2001) point out, "because of the one-to-one relationship between p values and observed power, nonsignificant p values always correspond to low observed powers" (p. 20). Thus, "the claim that a study is 'underpowered' with respect to an observed nonsignificant result" is "tautological and uninformative" (Goodman & Berlin, 1994, p. 202). It does not make sense to suggest that "If statistical significance is not reached, then the researcher should conduct a post hoc power analysis in an attempt to rule in or to rule out inadequate power (e.g., power < .80) as a threat to the internal validity of the finding" (Onwuegbuzie & Leech, 2004, p. 219), because the nonsignificant result guarantees that the power was inadequate for detecting a population effect equal to the sample effect.

Second, this power value provides an answer to a question that doesn't much matter, namely, "What chance was there of producing a statistically significant result, assuming that the population effect is exactly equal to the observed sample effect?" This question is different from one asking about the chances of producing a statistically significant result assuming some population value of independent interest (one that is theoretically motivated, based on prior research results, identifiable as a practically important effect, and so forth). The answer to the latter sort of question, even if addressed after-the-fact, can potentially be useful: "Previous researchers found effects averaging about $r=.40$, and we had good power (a good chance of finding statistically significant results) assuming

a population effect of .40, so the fact that we didn't find significant effects is meaningful . . ."

So where post hoc power refers to "the power of the test assuming a population effect size exactly equal to the observed sample effect size," such power figures do not provide much helpful information. But where after-the-fact power analyses are based on population effect sizes of independent interest (as opposed to a population effect size exactly equal to whatever happened to be found in the sample at hand), they can potentially be useful.

POWER ANALYSES BASED ON POPULATION EFFECT SIZES OF INDEPENDENT INTEREST

One might wonder whether after-the-fact power analyses based on population effect sizes of independent interest can actually be of much value. After all, such analyses do not appear to provide any information that is not implicitly provided by p values and confidence intervals (as indicated by, e.g., Colegrave & Ruxton, 2003). Consider, for example, the hypothetical researcher whose study with N=80 yielded a sample correlation of .13. This effect is not significantly different from zero (two-tailed p=.25); the 95% confidence interval around the observed correlation extends from −.09 to .34. So if one were interested in seeing whether the data were consistent with a population correlation as large as .40 (perhaps because previous research suggested such a value), the confidence interval would make it plain that a population effect of .40 is unlikely.

But appropriate power analyses can provide another way of displaying the relevant point. With N=80 (given .05 alpha, two-tailed test), the hypothetical study had power of .96 for detecting a population correlation of .40—which is to say that if the population effect really is as large as .40, the study was almost certain to yield a statistically significant result. Thus the fact that no such significant result was obtained makes it correspondingly unlikely that the population effect is as large as .40.[2] In a sense, this way of expressing the point provides no more information than does the confidence interval (we already know that given these data, the population effect is unlikely to be .40, because .40 lies outside the 95% confidence interval), and indeed the confidence interval gives even more information than the power analysis does. But the power analysis does provide

[2]As a reader pointed out, this inference may be suspect in cases where factors such as measurement unreliability have reduced power; the failure to find a significant effect might reflect measurement problems or other research shortcomings rather than the lack of a genuine effect. Of course, such research weaknesses (e.g., unreliability, poor realizations of desired experimental contrasts, and so forth) condition the interpretation not only of power estimates but also of (for example) effect sizes and related confidence intervals.

another way of displaying the idea that these data indicate that it is improbable that the population effect is as large as .40.

Appropriate after-the-fact power analyses can also permit one to assess whether a study was well-designed for examining various hypotheses about the population effect. The hypothetical study under discussion (N=80, .05 alpha, two-tailed test) had power in excess of .995 given a population correlation of .50, power of .96 given a population correlation of .40, power of .78 given a population correlation of .30, power of .43 given a population correlation of .20, and power of .14 given a population correlation of .10. Thus for detecting relatively large population effects, the study was very sensibly planned—but it was not well-designed for finding (establishing the existence of) smaller effects.

One may notice that this last set of after-the-fact power figures functionally identifies a range of plausible population effects given the obtained nonsignificant results (e.g., a population correlation of .20 is plausible, but one of .40 is not). Of course, this range is more transparently (and, arguably, more accurately; see Hoenig & Heisey, 2001) provided by the 95% confidence interval (−.09, .34). Perhaps if researchers (and research consumers) understood confidence intervals as well as one might like, reporting such power analyses would be indeed pointless. At least at the moment, however, such analyses might be helpful to at least some readers.

In any case, the larger point is that after-the-fact power analyses can sometimes be a useful supplement to p values and confidence intervals, but only when based on population effect magnitudes of independent interest. (For some other discussions relevant to observed [post hoc, retrospective, achieved] power, see Froman & Shneyderman, 2004; Goodman & Berlin, 1994; Levine & Ensom, 2001; Smith & Bates, 1992; Yuan & Maxwell, 2005; and Zumbo & Hubley, 1998).

GPOWER AND SPSS

The widely available GPower program (Erdfelder, Faul, & Buchner, 1996; Faul, Erdfelder, Lang, & Buchner, 2007) is perhaps one source of confusion here, given the way in which the program's computational options are labeled. The "a priori" computation option has the user enter a specific effect size, significance criterion, and power; the program returns the needed sample size (i.e., the sample size that will provide the specified power, given the specified population effect size and specified significance criterion). The "post hoc" computation option has the user enter a specific effect size, significance criterion, and sample size; the program returns the power (i.e., the power of a test for detecting a population effect of the specified size, given the specified sample size and the specified significance criterion).

In the post hoc option in GPower, one can enter any population effect size (or, sequentially, a variety of different population effect sizes). That is, the post hoc power computation does not necessarily involve entering the specific effect size observed in some particular sample. In GPower, "post hoc" is just the name for the option that outputs a power value (on the basis of the sample size, effect size, and significance criterion specified by the user). "A priori" is just the name for the GPower option that outputs a sample size value (on the basis of the effect size, significance criterion, and power specified by the user).

To see the potential for confusion here, imagine a researcher who is planning a study, knows the approximate likely *N*, has identified a population effect size of interest, and wants to have an estimate of statistical power in advance (perhaps to decide whether to seek additional participants). Notwithstanding the GPower labels, that researcher can use the GPower post hoc option to compute power as part of research planning.

In SPSS, the power terminology "observed power" is similarly unhelpful by virtue of being less transparent than one might like. Unsuspecting users might well suppose that an observed power figure is informative because somehow it is the "actual" (observed, measured) power of the test as opposed to some imagined or hypothetical power. These users might not realize that SPSS's observed power figure is based on treating the obtained sample effect size as the population effect size. Such power figures are at a minimum not especially helpful and potentially can be badly misunderstood.

REASONING ABOUT NONSIGNIFICANT RESULTS

When a statistical test returns a nonsignificant result, bad thinking of various kinds—often related to statistical power—seems invited, so it may be useful to consider explicitly some of the pitfalls here.

One kind of bad reasoning, perhaps less common than it once was, is to think "My result was nonsignificant, therefore the null hypothesis is true: The population effect is zero." One useful effect of greater attention to issues of statistical power and effect sizes has been to encourage the recognition that a nonsignificant result can occur even with a nonzero population effect. A nonsignificant result does not mean that the population effect is in fact zero; it means only that a population effect of zero cannot be ruled out.

However, a second, related form of bad reasoning about nonsignificant results is sometimes seen: "My result was nonsignificant, but my power was large (for some population effect size), therefore the null hypothesis is true: The population effect is zero." So, for example, a researcher reasons "With N=200, my obtained sample r of .12 is not significant, but my power was .99 (for detecting a population correlation of .30, which is a medium-sized effect) so the null

hypothesis must be true." But, again—no matter what the power figure is for any specified population effect—a nonsignificant result does not mean that the null hypothesis is true; it means only that the null hypothesis cannot be ruled out.

A third kind of bad reasoning is: "My result was nonsignificant, but with a larger *N* it would have been statistically significant. So I'm entitled to think that there really is a nonzero population effect—it's just that I couldn't establish the existence of that population effect (I couldn't get a statistically significant result) because my power was so low." This sort of reasoning can seem to be buttressed by observed power analyses: "My sample *r* of .13 was nonsignificant given my *N* of 80, but my observed power was only .21, so there probably is a positive population correlation." If additional data were in fact collected (and so the *N* made larger), there is no guarantee that the sample effect size would remain unchanged. Given additional data, the size of the new effect might be quite different from the initially-observed value. So for a researcher to think wistful thoughts here ("If only we had had more participants, we would have had significant results") constitutes genuinely wishful thinking.

A fourth kind of bad reasoning about nonsignificant results involves using observed power to make inferences about sample-size requirements for detecting large (important) population effects: "I found a large, important effect size but my result was nonsignificant; my observed power was low. So now I have learned that my sample size was too small to reliably detect an important effect." It's certainly true that if a sample effect size is large but nonsignificant, then (ceteris paribus) the sample size was too small to have provided a good chance of yielding a statistically significant result given a large population effect (a population effect equal to the large observed sample effect). But one can (and should) know *in advance* that a sample size is too small to reliably detect a large population effect; one needn't wait until data are collected. If a correlation of .40, for example, is known to be substantively important, then a before-the-fact power analysis can indicate the sample size needed to have reasonable power (or, alternatively, will indicate the power provided by one's contemplated sample size) for detecting such a population effect.

Against this backdrop, one might wonder about those editorial policies (in a number of communication journals) requiring authors to provide power estimates for nonsignificant effects. Such policies have surely had the salutary effect of encouraging greater attention to issues of statistical power and effect sizes, with correspondingly more widespread understanding that (for instance) a nonsignificant result does not mean that the null hypothesis is true. However, as should be plain, when nonsignificant results are obtained, only certain sorts of power estimates are likely to be valuable, namely, those based on population effect sizes of independent interest. If authors comply with these editorial policies by computing power figures assuming a population effect size equal to the observed effect size (or by mindlessly reporting "observed power" figures from SPSS),

then confusion and misunderstanding will be invited. When based on population effect sizes of independent interest, such after-the-fact power computations can be useful (in helping readers understand the obtained results, or in planning future research), but not otherwise.

CONCLUSION

The implications of the considerations discussed here can perhaps be summarized in three brief directives:

1. When power estimates are reported, specify (and, as necessary, justify) the values for each of the variables used to compute power. This practice, if followed consistently, would (a) help remove the misunderstanding that a given statistical test has a single particular power figure associated with it and (b) give greater attention to the reasoning behind the choice of population effect(s) to be detected. One imagines that researchers will not often be called on to justify, for instance, .05 alpha or the use of a two-tailed test, but they will presumably need to underwrite their expectations about the range of plausible population effects.
2. Avoid labels such as "post hoc" power, "observed" power, "retrospective" power, "achieved" power, "prospective" power, and "a priori" power. These are potentially confusing shorthand expressions that do not encourage specification of the particular values underlying reported power figures. If one's power computations use the observed sample effect size as the basis of the population effect, say so; do not simply call this post hoc power.
3. Use effect sizes, confidence intervals, and p values to describe and interpret results; if after-the-fact power analyses seem useful for didactic purposes, ensure that they are based on population effect sizes of independent interest.

These practices may over time encourage some broader reformation in our understandings of research methods. Consider, for example, that (in appropriate circumstances) researchers are commonly advised to strive for as large an N as possible. For the moment, many will no doubt think of this as a matter of "increasing my statistical power" and hence "increasing my chances of finding statistically significant results." But this can alternatively be seen as a matter of "making my confidence interval narrower" and thereby providing a better estimate of the population value of interest. To the extent that null hypothesis significance testing becomes displaced by considerations of effect sizes and confidence intervals, to that same extent concerns about statistical power will evaporate, to be replaced by concerns about accurate effect size estimation.

REFERENCES

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

Colegrave, N., & Ruxton, G. D. (2003). Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology, 14*, 446–447.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers, 28*, 1–11.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Froman, T., & Shneyderman, A. (2004). Replicability reconsidered: An excessive range of possibilities. *Understanding Statistics, 3*, 365–373.

Gillett, R. (1994). Post hoc power analysis. *Journal of Applied Psychology, 79*, 783–785.

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121*, 200–206.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*, 19–24.

Levine, M., & Ensom, M. H. H. (2001). Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy, 21*, 405–409.

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics, 3*, 201–230.

Smith, A. H., & Bates, M. N. (1992). Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology, 3*, 449–452.

Williams, R. H., & Zimmerman, D. W. (1989). Statistical power analysis and reliability of measurement. *Journal of General Psychology, 116*, 359–369.

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30*, 141–167.

Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician, 47*, 385–388.