Comment

Naomi S. ALTMAN

P-values and power estimates are both required in order to understand reproducibility of results. Improving the power of a study improves both the false discovery and false nondiscovery rates at any p-value threshold. An idea from the multiple testing literature, π_0, the proportion of truly null hypotheses among the tested hypotheses, can be used to reinterpret the p-value of a single test as the false discovery rate if the null is rejected. Use of bench-mark estimates of π_0 based on whether the hypothesis is a well-supported primary aim of the study, a secondary aim formulated as part of the study design, or a hypothesis proposed after examining the data can assist in interpreting the importance of observed p-values

KEY WORDS: FDR; π_0; Power; Reproducibility; Reproducible research.

Ideas from multiple testing of high dimensional data provide insights about reproducibility and false discovery rates of hypotheses supported by p-values.

Much of the statistical testing literature focuses on false rejection of the null hypothesis, also called false discovery. Individual p-values of prespecified hypotheses provide protection against a single false discovery if rejection of the null is done at some prespecified threshold such as $p < 0.05$. In many contexts, however, false nondiscovery (i.e. failure to reject the null hypothesis when it is false) is equally important, as it may lead to lack of follow-up of important hypotheses or failure to understand all the determinants of a system under study.

Both false positives and false negatives lead to irreproducibility of results. For example, sample sizes are often determined by the specification of "80% power (probability of false nondiscovery) when rejecting at $p < 0.05$." In this case, with two independent studies of the process both of which achieve the specified power and use rejection threshold $p < 0.05$, the probability of discordant results is 9.5% if the null hypothesis is true and 32% if it is false. The probability of two rejections in two trials when the null hypothesis is false is the square of the power—64%. For a fixed sample size, decreasing the significance threshold improves the probability of concordant results if the null hypothesis is true, but decreases the power to detect true discoveries. If the power is reduced to 70%, then the probability of two rejections in two trials in which the null distribution is false is only 49% and the probability of discordant results increases.

Improved study design, including larger sample size, increases power for any fixed significance threshold. What is less appreciated is that if some proportion π_0 of the hypotheses under test are truly null (and the remainder are truly not) then increased power reduces both the false discovery rate (FDR, the expected proportion of rejections for which the null is true) and the false nondiscovery rate (the expected proportion of failures to reject that are actually nonnull). If m independent hypotheses are tested, with significance threshold α and power β, we expect $\alpha\pi_0 m$ false discoveries and $\beta(1-\pi_0)m$ rejections, so that the FDR is $\alpha\pi_0/(\alpha\pi_0+\beta(1-\pi_0))$ which is a decreasing function of β. Similarly, the expected proportion of nondiscoveries that are false is $(1-\beta)(1-\pi_0)/((1-\alpha)\pi_0+(1-\beta)(1-\pi_0))$ which is also a decreasing function of β.

For high-dimensional data such as "omics" data, we can often obtain an estimate of π_0 from the observed p-values (e.g., Storey 2002; Pounds and Cheng 2004). This can be used to determine a significance threshold that produces a reasonable estimated FDR and is the basis behind methods such as Storey's q-value (Storey 2003).

For studies with much smaller numbers of test statistics, there are two reasonable ways to proceed using this paradigm. We could target a particular FDR (assuming known power) and determine whether the π_0 needed to achieve this FDR for the significance threshold is reasonable given what is known about the system under study or we could determine a benchmark value for π_0 and determine an appropriate threshold. For example, if we select FDR $= 0.05$, then at $\alpha = 0.05$ and $\beta = 0.80$, π_0 needs to be 46% or less, which implies that fewer than half of the hypotheses we expect to test are actually null. If we expect π_0 to be 90%, then to obtain FDR $= 0.05$, we need to use the threshold $\alpha < 0.0046$—however, recall that as α decreases so does β, which has not been accounted for here.

My suggestion for studies with too few hypotheses for estimation of π_0 is to establish some benchmark levels of π_0 which can be used to estimate the false discovery rate when rejecting with the observed p-value. This could be viewed as analogous to a proposal of Rosenthal to assess the "file drawer problem" in meta-analysis by computing the number of null results needed to overturn the proposed conclusion.

For example, for primary hypotheses from a study with adequate preliminary data, we might expect π_0 to be 50%—that is, equipoise . For exploratory results, found by multiple tests or fitting multiple models after the data were collected, we might expect π_0 to be much higher, say 95%. In principle, appropriate benchmarks could be determined for each discipline by a literature search, but lack of details in the literature about nonsignificant results might make this difficult. Instead, I would propose benchmarks of 50% for well-supported primary hypotheses, 75% for secondary hypotheses proposed when the study is first designed, and 95% for fortuitous findings and findings that require selection of covariates to attain statistical significance. The observed p-values can be converted into false discovery rates using the benchmark values and power appropriate to the hypothesis being tested. As explained in Storey (2003) the false

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Naomi S. Altman, Department of Statistics, The Pennsylvania State University, (Email: nsa1@psu.edu).

discovery rate associated with a hypothesis behaves much like the posterior probability that the null is true. Use of benchmark values of π_0 yields many of the good properties of posterior probabilities for interpreting test results without requiring formulation of a full Bayesian model for each hypothesis.

References

Pounds, S., and Cheng, C. (2004), "Improving False Discovery Rate Estimation," *Bioinformatics*, 20, 1737–1745.

Rosenthal, R. (1979), "The 'File Drawer Problem' and Tolerance for Null Results," *Psychological Bulletin*, 86, 638–641.

Storey, J.D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479–498.

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *Annals of Statistics*, 31, 2013–2035.