Leveraging Statistical Methods to Improve Validity and Reproducibility of Research Findings

Jennifer Urbano Blackford, PhD

**Scientific discoveries** have the profound opportunity to impact the lives of patients. They can lead to advances in medical decision making when the findings are correct, or mislead when not. We owe it to our peers, funding sources, and patients to take every precaution against false conclusions, and to communicate our discoveries with accuracy, precision, and clarity. With the National Institutes of Health’s new focus on rigor and reproducibility, scientists are returning attention to the ideas of validity and reliability. At *JAMA Psychiatry*, we seek to publish science that leverages the power of statistics and contributes discoveries that are reproducible and valid. Toward that end, I provide guidelines for using statistical methods: the essentials, good practices, and advanced methods.

**The Essentials**

Choosing the appropriate statistical approach is an essential first step. Traditional parametric statistics are based on assumptions of normally distributed data and linear associations. While many commonly used statistical tests and estimation procedures are fairly robust to violations of normality, both validity and statistical power can be substantially reduced when data are not normally distributed.1,2 This is especially important in psychiatry, where many phenomena are not normally distributed or linearly associated, for example, symptom ratings, duration of untreated illness, number of medications, and even education level. Early adoption of nonparametric statistics was thwarted by the requirement of substantial computing power; however, this is no longer an issue. Permutation tests, nonparametric tests and estimates of effect sizes, and bootstrap approaches to generate confidence intervals are now readily available in most statistical packages.

Scientists often create categories from continuous data. For example, categories of depressed/not depressed are created from continuous measures of depression, like the Patient Health Questionnaire-9 (PHQ-9) or Beck Depression Inventory (BDI). While categories can be helpful for diagnosis and contribute discoveries that are reproducible and valid. Toward that end, I provide guidelines for using statistical methods: the essentials, good practices, and advanced methods.

While categories can be helpful for diagnosis and contribute discoveries that are reproducible and valid. Toward that end, I provide guidelines for using statistical methods: the essentials, good practices, and advanced methods.

**Good Practices**

In the era of huge samples—combined data sets from collaborative projects, epidemiological studies that report on the populations of entire countries, and genome-wide association studies of tens of thousands individuals—it is crucial to report effect sizes. Effect sizes, estimates of population parameters that are independent of sample size and other design decisions, provide a tool for determining whether a finding is not only statistically significant, but also clinically significant. Despite the prominence of the Cohen *d* (or Hedges *g*), there are many different effect sizes. For 2-groups comparisons, besides *d* or *g*, are the success rate difference and its reciprocal number needed to treat (or take). For correlation, there are the Pearson *r*, Spearman *r*, and Kendall *τ*. Of note, some fields, like neuroimaging, are still considering the most appropriate way to compute effect sizes.

As lucidly described in a recent *JAMA Psychiatry* editorial by Statistical Editor Helena Chmura Kraemer, PhD,6 covariates must be justified and selected a priori. Including covariates to adjust for baseline differences between groups is inherently post hoc and will increase the likelihood of false-positive results. When using covariates, both the unadjusted and adjusted models must be presented so that the impact of the covariates on the final model can be fully evaluated.

**Advanced Methods**

The most unbiased and valid estimates of effect size are generated with independent replication across multiple samples.
While external validity is the ultimate goal, an important first step—internal validity—can be established within a single sample using cross-validation methods. Cross-validation is commonly used in model prediction analyses, but these methods have yet to proliferate to other fields. In the simplest form of cross-validation, data are randomly split into 2 subsets; the model is generated from the first set and then validated with the second set. Another variation, the leave-one-out (LOO) uses the N-1 sample to generate the estimate and tests the model on the left-out observation. However, 5-fold and 10-fold cross-validations tend to be more stable and are therefore preferred to LOO. Cross-validation methods provide final parameter estimates that are less biased and more likely to replicate in independent future studies and therefore should be used whenever possible.

The proliferation of affordable and readily available statistical software has supported the exponential increase in sophisticated data analyses methods that have led to novel discoveries. However, with the ease of these sophisticated “black-box” programs comes extra caution, as illustrated by recent reports of bugs in 2 widely used neuroimaging software packages.7,8 Russ Poldrack recently shared his experience in “Reproducible Analyses in the MyConnectome Project”9 where he describes how a bug in the Linux version of an R package produced different results despite the same data and analytic pipeline. He concludes, “the experience showed just how fragile our results can be when we rely upon complex black-box analysis software.” Thus, when using black-box programs, let the user beware.

Selective reporting creates bias. The registration of clinical trials before enrollment of participants is an important step forward. However, other types of studies also suffer from bias. Prospective registration of meta-analyses (eg, Cochrane Prospective Meta-Analysis Methods group or PROSPERO) and observational studies will further enhance transparency. Reporting the results for all variables examined—either in the manuscript or Supplement—allows an evaluation of the significant findings within the context of all findings. Increasing transparency will provide for a growing collection of all findings ultimately leading to more accurate meta-analyses and valid conclusions.

Conclusions
In summary, it is our collective responsibility to ensure that we attend to the essentials, use good practices, and strive to use methods that will enhance both reproducibility and validity. Over the next several months, we will publish several Viewpoint articles that address critical issues related to statistics and research methods. With proper attention to data and statistical methods, we will advance psychiatry.