The Fault in Our Stats

November 25, 2014

Just as advancements in technology have revolutionized the collection and measurement of data, new statistics offer an approach to analysis that allows results to be examined and applied in a broader scientific context. A growing demand for accountability and reliability in the social sciences has spurred an increase in studies reporting *effect size*, which measures the magnitude of a relationship, and *confidence intervals*, which provide an estimate of error. Unlike hypothesis testing and *p* values, these parameters are not as skewed by population size and thus can be used to compare results across multiple studies (meta-analysis), serving as a standard measure. With increasing use of effect size and confidence intervals in the field of psychological science, it is important for young researchers to master these "new statistics."

As an undergraduate at a small public liberal arts college, I discovered new statistics not directly from classes, but by heeding the advice of my research methods professor. My professor imbued in me the notion that mastering statistics could help me enter the world of academia, which seemed inaccessible to me as a freshman at an undergraduate institution. However, like many of my extremely supportive professors, he could offer me only encouragement to discover more, as this new generation of statistics was not included decades ago in his training as a psychologist. To build upon the basics of my stats class, I turned to my generation's fountain of infinite wisdom: the Internet.

My first priority was tackling Cohen's *d* and confidence intervals. I turned to Wikipedia to learn about the language of effect size and the movement that encouraged new statistics as a standard of empiricism. Then, I searched for straightforward papers that demystified the math, such as Robert Coe's 2002 paper "It's the Effect Size, Stupid" and Geoff Cumming's seminal 2001 article. I practiced using data that were both easily available and familiar to me — for example, I reran regressions for a project with my dedicated mentor, Joan Zook, investigating how adolescents' social goals and peer-nominated popularity influenced academic self-presentation strategies. To replicate estimation parameters for ANOVAs, I examined how school-specific factors predicted academic outcomes for the Rochester City School District beyond the effect of income — data I analyzed to contribute to a white paper for a local state senator (data available here).

At first, I sought to replicate the results I had already established: New to SPSS software, I clicked my way through the descriptive statistics I needed, plugged them into the equations and calculated Cohen's *d* by hand for several observations. This mundane procedure reinforced the meaning of each descriptive statistic and the logic of effect size as opposed to hypothesis testing. Eventually, I discovered Exploratory Software for Confidence Intervals (ESCI), a free plug-in for Microsoft Excel available at tinyurl.com/free-esci. I simply plugged my data into Excel, calculated ANOVAs, and generated confidence intervals. To double-check my regressions, I reran them in SPSS to include ?2, the total variance of a given effect (obtained by checking the option "Estimates of Effect Size"). As many trained in quantitative methods warn, there is danger in merely clicking your way through statistics without being mindful of context, limitations, and assumptions, such as treating count data as continuous.

However, the benefits of expanding the accessibility of estimate parameters to a wider audience outweigh this pitfall. Ultimately, adoption of effect size as standard practice will increase participation in meta-analyses and highlight discrepancies between significance and effect size.

I proceeded to check these boxes as part of my routine data analyses, retaining the value of estimation techniques as I acquired increasingly advanced statistical prowess. I conducted a literature review to ground my newfound skills in the theory and larger context of applied statistics. The advantage of estimation techniques is that they are more informative and less susceptible to sample size errors than traditional hypothesis testing. Confidence intervals can be interpreted like *p* values: The range of the confidence interval is statistically significant if it does not include the null values. Confidence limits further provide an estimate of precision: A narrow range of values indicates a more accurate estimate, while a wider range of values, sometimes corresponding to a small sample size, provides a less precise measure. The result is a more reliable measure of effect than *p* values (Cumming, 2008): The center of the confidence interval is the most likely value for the population mean, such as the sample mean for ANOVAs or the beta coefficient for regression.

Confidence intervals are reported in the same units as the outcome measure, making them easier to interpret. Likewise, effect sizes lead to consistent interpretations, have practical implications, and encourage collaboration and comparison among scholars within a discipline. When estimation parameters are combined with bootstrapping techniques, we can be more confident that results are not an artifact of selective sampling or reporting. That said, more thorough statistics cannot compensate for a faulty study design or sampling bias.

Learning new statistics encouraged me to investigate the methods behind the content in my psychology classes. However, the slow VPN connection necessary to access SPSS, coupled with a growing hobby of programming, led me to discover the magic of R, the open-source statistical software. My college happened to offer two advanced seminars that required R and thus, for the first 2 weeks of my semester, learning R became my (third) part-time job. All the rumors are true: The learning curve is steep, there is no built-in graphical interface, and the error messages are frustrating. However, the freedom, efficiency, and computing power of R yield great returns. R demands that one think to achieve an outcome; above and beyond the syntax of other statistical software, the precision R requires has taught me to appreciate the strengths and limitations of statistical science.

Those who endeavor to learn R are never alone: A unique advantage of R is that all the answers are a Google search away and provided by the growing, active open-source community of R users and developers. Thus, information is freely (both in cost and in quantity) dispersed, from the most basic to the most advanced statistical procedures. I strongly suggest new adopters of R try Quick-R, located at statmethods.net. In addition, I was lucky to have a brilliant statistics pro for a roommate, whom I could pester when my code refused to run. We would edit each other's codes for typos (like a misplaced parenthesis); search through solutions from stackoverflow.com, a help forum for programmers; or, if the angry red error messages would not cease, suggest a Netflix break. Further, the academic culture at my school is immensely supportive, with professors who generously pored over my code, nurtured my ambitions, and offered constructive criticism to hone my interest in statistics into a refined skill.

My foray into advanced statistical methods was intended to bolster my odds at getting into graduate school; however, in the process, I found a love for data science I never knew I had. As an "Advanced

Algebra" dropout, I could never have envisioned that I would conquer estimation statistics, the dreaded R, and structural equation modeling. This semester, I'll be shaping my honors thesis for publication, applying to grad school, and spearheading a data-driven, campus-wide weekly newsletter that surveys public opinion from students and faculty and models the results.

Now is a critical turning point in data analytic science: Interdisciplinary academics, journalists, and programmers are uniting in the interest of improving the reliability of information. Psychologists are responsible for advancing our science in this vein — whether through a few clicks or a few lines of code, new statistical methods offer a more accessible, accurate, and widely applicable alternative to *p* values, empowering researchers to raise our standards of empiricism.

References and Further Reading

Coe, R. (2002, September). *It's the Effect Size, Stupid: What Effect Size Is and Why It Is Important*. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, United Kingdom. Retrieved from http://www.leeds.ac.uk/educol/documents/00002182.htm

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286–300.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(4), 532–574.

Kalinowski, P. (2010, April). Understanding confidence intervals (CIs) and effect size estimation. *Observer, 23*(4). Retrieved from http://www.psychologicalscience.org/index.php/publications/observer/2010/april-10/understanding-confidence-intervals-cis-and-effect-size-estimation.html

Greenwald, A., Gonzalez, R., Harris, R., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33*(2), 175–183.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods, 16*(2), 93.

Sullivan, G. M., & Feinn, R. (2012). Using effect size — or why the P value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–32.