Guillaume Pagnier

February 27, 2019

Multivariate Stats HW5

Formally learning statistics in a classroom is not straightforward. It’s not until someone starts being frustrated by only having access to a small sample and needing to make claims about a population (the overall purpose of statistics anyways), that they start developing an intuition for how hypothesis test-derived conclusions could be possible/plausible. Oftentimes it requires more than one pass through the same material for the facts to start to stick; statistics is inherently an abstract science, where you’re trying to make reasonable claims about a ‘theoretical’ uncalculatable population. I was fortunate enough to undertake my very own research project as a senior in college where I was exposed to stats in a very real way. My project revolved around having two continuous dependent measures (skin conductance responses and reaction time) and multiple binary and continuous predictors and making assumptions as to which predictors were important. Though I hadn’t formally learned how regression works, my advisor, a very patient psychological faculty member, spoke to me about a program called R and how it could be used to calculate which predictors significantly affect those two DVs. Not having a sense of statistics yet, I made plenty of mistakes, treating RT as normally distributed and analyzing RT and SCRs totally separately though they were linked via subject. However, this is when more straightforward concepts became clear; what a t test actually computes and why an ANOVA doesn’t tell you which groups significantly differ from one another. Subsequently, the flaws of classical frequentist statistics started to emerge, how in a best-case scenario, 5% of t tests results could be acquired even though the null hypothesis was true (due to the concepts we discussed in class, it’s probably much higher than 5% of studies anyway).

I continued my statistics education in graduate school in 2014, by opting to pursue a fast track masters. I essentially redesigned my undergraduate project started conducting my own unsupervised analyses. This is when I learned about hierarchical modeling and why, if predictors come from the same subject, must be related somehow. My advisor was one of the original lme4 contributors and as a result, understood the need for rigorous statistics. He also pushed for atypical analyses such as Bayesian statistical testing which at the time, wasn’t as popular as it is now. Trying to make meaningful conclusions about a relatively complicated dataset (two continuous dependent variables with autocorrelated binary and continuous predictors) is what opened my eyes about statistical misuse and how easy it was to reach an arbitrary level of significance and thus, importance. It’s during this time when I realized how important it was to come up with conclusions before looking at the data set and why finding significant results post-hoc is dangerous.

After my masters, I stayed in research, working in a neuroscience lab in Boston studying Alzheimer’s disease with mice as models. After learning all about lme4 and the importance of rigorous statistics I was frankly shocked at how little people thought about statistics. I know that not everything that people do is bad but I honestly believe that I saw the full range of statistical misuse during my two years there. I don’t believe this was intentional, I think these very smart and successful scientists considered the nuances of statistics to be below them. At this time, I considered myself at least proficient in how to account for all the common pitfalls and now I was surrounded by extremely prominent researchers using t test after t test post-hoc in minitab and just checking off the different boxes (parametric, non-parametric etc.) until they got the much sought-after star of significance, regardless of any assumption they were violating. The converse also occurred, when scientists would assume that a non-significant hypothesis test would indicate that two groups were equivalent (this happens A LOT in neuroscience labs). After bringing up these concerns (and my attempt to explain why a non-significant hypothesis test doesn’t produce any information) it was explained to me that the effect sizes we were looking for were so extreme and the descriptive statistics alone would indicate reality so it was a ‘waste of time’ to conduct the appropriate statistics: ‘since everyone does it like this why would we change?’. Fortunately, there was another post-doc who shared my concerns and when clearly more sophisticated statistics were needed, we convinced the PI to send the data to the neighboring biostatistics lab so we could be confident that we were reaching the right conclusions using the right methodology.

Another instance where I’ve been subject to questionable research practices is in the same lab when I witnessed someone classifying each cell as independent, totally disregarding that multiple animals contributed to the cell count. To make matters worse there were unequal numbers of cells in an unequal number of animals in each condition so it was by no means a balanced design. They had three predictors that they wanted to check if they were significant and the way they did this is by conducting three independent t tests, ignoring any collinearity among predictors, never checking any assumption, and capitalized on the high cell count to conclude that all three predictors significantly contributed to the outcome measure via 3 independent p scores. I’m not sure I can think of a worse way to do this and unfortunately, this is not a unique case. “A solution to dependency: using multilevel analysis to accommodate nested data” (Aarts et al., 2014) is a review of neuroscience studies that fail to do what my lab does: fail to account for multilevel data. Aarts et al, found that 53% of 314 reviewed papers contained were guilty of such an error. This incorrect usage of statistics is also compounded by omitting effect sizes. ‘Power failure: why small sample size undermines the reliability of neuroscience’ (Button et al., 2013) echoes what we discussed in class, omitting effect sizes can create extremely misleading results by indicating that there’s a significant difference when in reality the p value is artificially deflated via the large sample size. This is especially problematic in neuroscience since outcomes are measured via cells, synapses, or trials all of which are relatively easy to acquire a lot of.

There are also other popularized cases of research in my field being done poorly that I haven’t seen personally. Nieuwenhuis, Forstmann, & Wagenmakers found in a 2011 Nature Neuro study that nearly half of reviewed articles in 5 top neuro journals do not appropriately describe differences in effect size. Essentially, they found that if one effect is significant, and one is not, many labs don’t then compare the two differences to see if they themselves are significant. There is an easy solution to this, another hypothesis test needs to be done that analyzes the two differences. As the authors point out, this was vastly decrease authors’ chance of type I error.

Fortunately, there are ample resources available that offer solutions to the problems I’ve described above. Aarts et al., found that using multilevel analyses for multi level data decreases the chance of type I error but doesn’t decrease the power. An example of this would be using linear mixed effects modeling and accounting for random effects that can occur per grouping variable (such as subject or animal). Another solution is to simply stop aggregating ALL of the data and running a t-test.

Another solution is to be as transparent as possible. Nord, Valton, Wood, and Roiser describe in their 2017 paper that researchers should document everything they do as a single summary statistic doesn’t provide enough information to the readers about the study. Always including the effect size and p value for instance will grow the field of neuroscience by leaps and bounds. Top tier journals and NIH grants, undoubtedly aware that they have been guilty of enabling incorrect statistics (see Nord et al., 2017, and Aarts et al., 2014 for empirical evidence of this) are now including statistics reporting checklists when researchers submit manuscripts. Such checklists enforce basic statistical upkeep such as checking assumptions, justifying appropriate degrees of freedom, including effect sizes, and demanding a priori power analyses and rationales. To supplement this, Wilcox & Rousselet in their 2018 study ‘A Guide to Robust Statistical Methods in Neuroscience’ lays out what I think is the cleanest protocol of appropriate statistics which describe what to do given certain datasets. Paul Kording, in a 2014 paper argues that Bayesian statistics are also useful in making meaningful conclusions about the brain, a domain that has historically not been known to use bayes theorem. As he points out, bayesian and frequentist statistics can be combined to draw identical conclusions and make applicable statements about the population. These are relatively recent developments for the field of neuroscience, but I am optimistic that even these basic tenets of hypothesis testing will become mandatory for all neuroscience research groups and that the science, as a result, will significantly improve.