591

Intro

General summary, nothing new. But to begin with, it should be noted that Wilcox wrote an article 10 years ago on the same thing in this same journal (Link on the class webpage). Not sure why folks keep insisting on keeping SPSS in the picture right now. The only way it is able to do this stuff is via the R plugin and Zumastat, which itself requires R, which after what seemed like years the guy finally got the robust thing completed only to have version 16 break it. I'd actually inquired with him about it around a year or so ago and it wasn't done then because of his applied research obligations: so basically nothing, then finally working, then broken. No word on whether it will work with 17, which in the little I've used it is performing better than 16. They'll change, if there's money in it. But I just read their financial report from last year the other day, they think they're leading in general offerings and that only SAS is a notable competitor.

Problems With Classic Parametric Methods

FYI, the Micceri article has been available to you guys on the website from day 1. The Wilcox text is the supplemental reading from your syllabus.

592

He looks like a combination of Stephen Malkmus and Matt Kadane. I'm not sure what that means statistically, but it would be a quite the mix musically. I will say this, I've always found this practice of putting pictures in articles by AmerPsyc pretty poor from a scientific point of view. If psychologists know anything, it's that people judge books by their cover. Oddly, I could not find him at the U. of W.A. psych page.

Problems With Classic Parametric Methods cont'd.

Shift of focus from non-normality to the heteroscedasticity issue. Play with this why don't cha? http://onlinestatbook.com/stat_sim/robustness/index.html

593

Problems With Classic Parametric Methods cont'd.

"An ANOVA is run in SPSS or SAS, and the *p* value reported is .05. This should mean that if the null hypothesis is rejected, there is less than a 5% chance that a Type I error has been made." See? Even when people mean well they confuse observed p-values and alpha, or perhaps it's just that the whole thing is so confusing that there is no good way to write about it. SPSS and SAS give the observed p-value $p(D|H_0)$, and that is not alpha. If a study is designed with alpha = .05, and the observed p-value is .0001, type I error rate is still (no more than) .05. A p-value can be calibrated to represent a lower bound on alpha, but if you don't do that, it doesn't inform you about alpha is. See the Hubbard & Bayarri article on the class website.

Why Are Modern Methods Underused?

Lack of Familiarity With Modern Methods, Assumption Testing Issues, The Robustness Argument

This all sounds vaguely familiar...

http://www.unt.edu/benchmarks/archives/2008/june08/rss.htm "The reasons for such poor practice probably stem from two things primarily: the first is simple lack of knowledge regarding the underlying issues, and the second is using poor software for the analysis." I later quote Dennis Hopper and use the phrase 'ol' school'. I debated about spelling it with a k. See what APA keeps us from?

594

The Robustness Argument

"We are not trying to blame researchers for not being familiar with modern statistical methods. Researchers are busy in their own areas of expertise and cannot be expected to be familiar with cutting-edge developments within statistics." Every time I try to tell myself that I feel queasy.

Cutting edge? No, certainly not. Stuff from 30-40 years ago? That's just laziness and/or ineptitude.

Transformations
"We strongly recommend using modern robust methods instead of conducting classic parametric analyses on transformed data." Another reason for bootstrapping, but this is a nice list of the problematic approach of transforming your data.

Classic Nonparametric
It's true, they do ok, but have problems and don't extend to complex settings.

595

Misconceptions
I would say I've heard of SAS's robust capabilities for awhile, the bigger problem is SPSS. Stats guys can talk all they want, but until people see it magically appear in whatever poor choice of software they use, they won't do any particular analysis. For some this magic has to happen in the output stage even (i.e. done for them), but no one is paying attention to their research anyway.
The key word here compromise. If you desire a measure of 'central tendency', the mean does not provide a good one for data with outliers, and no amount of wishing will make it do so. If it takes trimming some data for a robust estimate then so be it. Remember that the trimmed mean is estimating a trimmed mean, not a mean. You aren't 'fudging' the data or anything like that. The median is a 50% trimmed mean, the middle value. Does it have no important information because it's not as sufficient as the mean? The mean is as poor at resistance as the median is at sufficiency, why the preference for one over the other in every situation? Where else do we behave like that that has ever benefited us? It's called *adaptation* folks, it works.

Practical Introduction to Modern Methods
Of course this begs the question of 'practical' and 'nontechnical' to *whom*. It certainly is nontechnical- you guys, with little to no research experience, picked up this stuff pretty easily. However, I'm not sure how practical this is. This article doesn't give you enough info on how to actually do an analysis, not even a simple example of code to show how easy it can be, and does the software part comes late in the article. If part of the 'practical' nature of the article is to show ease of implementation, the ordering of topics is a little odd in my opinion.

596

Bootstrapping
You guys are expert at this already! Yawn.

Robust Hypothesis Testing, Software, and Resources
Finally, some info on how to actually do this stuff. Not sure why the recent Keselman article is the cited source of the trimmed t-test, though Keselman is definitely a good resource in general. Yuen, K.K. (1974). The *two-sample trimmed* t for unequal population variances, Biometrika. It's at least that old *that* old. And that ADF solution is just the Welch correction (1938) under no trimming. Psychologists need to be reminded how far behind their typical methods are, I'll talk more of this later. The Wilcox text is a good and readable introduction, I even used it as a required text.

597

Do you guys think R code is like SPSS syntax? "SAS/STAT 9 includes some inbuilt capability for performing robust analyses." Again, people insisting on default (and even advanced) SPSS menu options as their primary approach (and are not using the R plugin to do the real work) are

not taking necessary steps to deal with problematic assumptions, such that if they arise, they are reporting incorrect statistics, probabilities, and incorrect conclusions based on them. Some may be bending over backward to force SPSS to take those steps, but unless that's reported, you can no more assume they have than those researchers can assume normality, homogeneity of variance etc.

Modern Rank Statistics

Rank Transform

I like this "RT is only mentioned here so that researchers know to avoid it." Enough said.

Anova-type Statistic

The describe the null hypothesis that the distributions are the same, which is what lies at the heart of 'nonparametric' analyses. They are distribution free. I actually had a psych graduate student ask about this the other day (and no I hadn't taught it to them). I am curious how testing 'relative treatment effect' would tell something different than testing the mean ranks. I went to the website, there isn't anything for SPSS in case you were wondering. But it does say ACTUNG! regarding the SAS macro so be advised. In any case, this is nice but the Wilcox functions have bootstrap/trim approaches that will allow you to keep things on the original scale.

598

Other rank based statistics

The Wilcoxon analysis appears to only deal with one type of outlier (those on the DV) and as such, I'd recommend a different approach. That RGLM website is very odd to me. Why go to all the trouble to make a webpage-program instead of just providing macros/functions or writing a package in R. Furthermore, it got stuck on its own regression example for me. Not recommended. I like people doing their own thing and making it available, but making a web stats program is more for demonstration, not for actual analysis (at least for now). Note however "Journals that regularly feature articles about modern robust methods include *Psychological Methods*, *Educational and Psychological Measurement*, and the *Journal of Modern Applied Statistical Methods*." If you consider yourself a researcher, these are the types of journals that need to be looked through from time to time, not just the ones related to your esoteric research area.

599

Effect Size

Effect sizes are probably the most glaring example of how far behind psychological research is. If η^2 wasn't obtained with a click in SPSS we wouldn't see that as often as we do, which still may not even be most of the time in ANOVA analyses, but standardized mean differences are still vastly underreported given how many studies do group comparisons. But the methods folk aren't talking about simple effect sizes anymore (that was *so* 80s), they've moved on to interval estimates of them and robust versions with intervals on those. However the probability superiority index? I like it but check your notes: that's the *common language effect size* (CL) we talked about in the discussion of 'case-level' effect sizes, and offered by McGraw & Wong (1992, the stochastic superiority was a response to it). However, this variant here might be better under nonnormal situations. Note that if you want yours to have the same range as the CL, i.e. from .50-1.0, you'd always have to make the larger mean 'first'. The thing to know that if you use their formulation and it is below .50 that probability refers to the smaller-mean group randomly exceeding the larger mean group. However, recognize again the easy transformation from *r* to *d* to case-level measures in group comparisons. *Effect size* is a measure of impact of one variable on another, but one can talk about in different ways.

I predict this effect size section will have little to no effect of its own. How do you talk to applied researchers about improvements to statistics they aren't even reporting right now? It's like trying to talk about a car to a civilization that hasn't invented the wheel.

Summary
"Most researchers analyze data using outdated methods."
"Most researchers analyze data using outdated methods."
"Most researchers analyze data using outdated methods."

"We recommend that researchers bypass classic parametric statistics in favor of modern robust methods. Modern methods perform well in a much larger range of situations than do classic techniques. The use of modern methods will result in researchers finding more statistically significant results when real effects exist in the population. Using modern methods will also reduce the number of Type I errors made by researchers and result in more accurate confidence intervals around robust effect size statistics."

More power and accuracy? Who needs that?

My overall summary is that I think this is a nice article for folks such as yourself just getting started, had you not just finished a class where all that conceptual content was covered, but even then you saw some new stats if nothing else. I personally get a sense of 'new to the game' from these authors, not that there's anything wrong with that, but this should not be taken as authoritative in my opinion (though most of their citations could be). Furthermore, this would have been greatly improved had they supplied actual macros/code in the appendix.

But really? This article is evidence of the failure to get the point across. Ten years after Wilcox puts his out there in the exact same journal talking about the exact same things, how much improvement has been made? If people are not taught it, either by applied researchers who keep up with this things (rare, since they aren't rewarded for it) or quantitative psychologists outright, they aren't going to use it. And furthermore, it is clear many are simply ignoring these advancements. I mean if this type of stuff consistently makes Psy Bull, Amer Psych, etc. and people still aren't using them, and more to the point, are using the *wrong* methods, then something must be fundamentally wrong with the discipline's approach to science. Quite frankly, psychology should be *embarrassed* for not keeping up with the last 30+ years of statistical and computing improvements. It is not alone as a discipline with regard to this, but there is almost a sense of arrogance and/or stubbornness concerning what for all intents and purposes seems to be a willful ignorance of the problems associated with typical research. One cannot trust the school, the person, or the journal its printed in to define good research. Bad research abounds, and you will have to take each article on its own merit (as it should be anyway).

For those of you that will that be 'purely applied' in your pursuits, i.e. practitioners of psychological science and 'evidence-based' therapy, you know what to look for in articles and are armed to guard against overreaching based on poor science. For those that will do their own research, you've been given both the tools and the research backing for how and why to use them, along with evidence of its use already in psych research. There is no excuse for continuing bad habits.