Cornell University
Cornell Statistical Consulting Unit

Stats Happening – Fall 2017

1) **Important Statistical Concepts for Reproducible Research**
2) **Debate on P-value Threshold**
3) **New Handout on SEM Fit Indices**
4) **New R package 'emmeans'**
5) **New User Written Functions in Stata for Longitudinal Data Analysis**
6) **Multicollinearity in R with the mctest package**
7) **New tools in SAS for Propensity Score Analysis and for Assessing Causality**
8) **New in Stata 15**
9) **New in SPSS 25**

---

1) Important Statistical Concepts for Reproducible Research

Reproducibility is one of the latest buzz words related to research. What does it really mean? Below we are presenting its definition along with related terms. If you want to know more about this look for our workshop on reproducibility that CSCU will be offering again next semester.

Reproducible: "Given a population, hypothesis, experimental design, experimenter, data, analysis plan, and code you get the same parameter estimates in a new analysis" (*1*)
Replicable study: "Given a population, hypothesis, experimental design, and analysis plan you get consistent estimates when you recollect data and redo the analysis" (*1*)
HARKing: Hypothesizing After the Results are Known–"Presenting a post Hoc hypothesis (i.e., one based on or informed by one's results) in one's research report as if it were, in fact, an a priori hypotheses." (*2*)
P-hacking: "Given a population, hypothesis, experimental design, experimenter, data, analysis plan, and analyst, the code changes to match a desired statement." (*1*)(*3*)
Fishing Expedition: "A willingness to look hard for patterns and report any comparisons that happen to be statistically significant." (*4*)
Garden of forking paths: "Given a population, hypothesis, experimental design, experimenter, data, analysis plan, and analyst, the code changes given the data you observe." (*5*)

Keep these useful concepts in mind for your own research and publications.

1. http://biorxiv.org/content/early/2016/07/29/066803.abstract.
2. http://journals.sagepub.com/doi/abs/10.1207/s15327957pspr0203_4.
3. https://fivethirtyeight.com/features/science-isnt-broken/#part1.
4.http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html
5. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

2) Debate on P-value Threshold

A debate has been waging on whether the threshold for p-values should be lowered to combat spurious relationships. Most researchers use a cut-off of 0.05, which was first established by Fisher in 1925 in his *Statistical Methods for Research Workers*. The rationale for setting the cut-off at 0.05 was that this would result in a coincidental finding only once out of every twenty trials. A survey of *Nature* readers published on September 19th, concluded that 69% of readers (out of 6,938 respondents) would be in favor of lowering the threshold from 0.05 to 0.005. (References: **http://www.nature.com/news/one-size-fits-all-threshold-for-p-values-under-fire-1.22625?WT.mc_id=SFB_NNEWS_1508_RHBox; http://www.stat.columbia.edu/~gelman/research/unpublished/abandon.pdf)**

3) New Handout on SEM Fit Indices

Structural equation models, which include path analysis and confirmatory factor analysis, are becoming more widely used across fields. We have created a useful handout of the common fit indices that are reported in structural equation models and confirmation factor analyses. To access this handout, visit https://www.cscu.cornell.edu/news/Handouts/SEM_fit.pdf.

4) New R Package 'emmeans'

The authors of the popular R package 'lsmeans' have recently released a new package called 'emmeans' which is intended to replace 'lsmeans.' Both packages can be used to obtain estimated marginal means for many linear, generalized linear, and mixed models. Additionally, post-hoc pairwise comparisons between estimated marginal means, general linear contrasts, and comparisons of slopes can be performed. Compact letter displays can be constructed to group pairs of means that are not significantly different. The 'emmeans' package allows graphical representations of the estimated marginal means using 'ggplot2'. This package is sure to become a staple for R users. For an example of a dataset utilizing this package, visit https://cran.r-project.org/web/packages/emmeans/vignettes/basics.html.

5) **New User Written functions in Stata for Longitudinal Data Analyses**

The user written Stata command **xthybrid** can decompose cluster-varying covariates on an outcome into a within-cluster effect and a between-cluster effect. This type of model is a hybrid between a fixed effects model, which can only calculate within-cluster effects, and a random effects model, which assumes the within-cluster effects and between-cluster effects are equal. For more information, visit https://ideas.repec.org/c/boc/bocode/s458146.html.

The user written Stata command **xtdpdml**, uses a structural equation model framework for estimating dynamic panel data. Cross-lagged models can be executed easily using this command. For more information on these types of models, visit https://www3.nd.edu/~rwilliam/dynamic/ and https://statisticalhorizons.com/lagged-dependent-variables.

6) **Multicollinearity in R with the mctest package**

Models that have two or more correlated predictor variables might have unstable coefficient estimates because of multicollinearity. Most researchers use VIF as their sole diagnostic measure for multicollinearity; however, VIF is neither a necessary nor a sufficient measure of multicollinearity. The R package mctest calculates multicolinearity diagnostic measures at the model level such as the determinant of the correlation matrix, condition index, Theil's indicator, Farrar Chi-square. Individual diagnostic measures such as VIF, tolerance, Klein's rule, and Leamer's method can be used to determine which variables might be associated with multicollinearity. **(**Reference**: https://journal.r-project.org/archive/2016/RJ-2016-062/RJ-2016-062.pdf)**

7) **New tools in SAS for Propensity Score Analysis and for Assessing Causality**

Randomized controlled trials (RCTs) are considered the gold standard in many field as they can establish causality. RCT are however often very difficult, if not impossible to implement. On the other hand, nonrandomized or observational studies do offer many advantages in their implementation, but one of their drawbacks is that it is more difficult to show causality, rather than simple associations. In the latest versions of SAS/STAT three new procedure are introduced that can assist in establishing causality with observational studies. The CAUSALTRT procedure allows the estimation of the average treatment effect (ATE) also known as average causal effect (ACE) and the average treatment effect for the treated (ATT or ATET) of a binary treatment variable on a continuous or binary outcome. The PSMATCH procedure provides a variety of approaches for propensity score analysis. The CAUSALMED procedure allows estimation of mediation under the counterfactual framework and to decompose the overall effect into the 4 components.

For further information see:

1. http://support.sas.com/resources/papers/proceedings17/SAS0374-2017.pdf
2. http://support.sas.com/documentation/onlinedoc/stat/142/causaltrt.pdf
3. https://support.sas.com/documentation/onlinedoc/stat/142/psmatch.pd
4. http://documentation.sas.com/?docsetId=statug&docsetTarget=statug_causal

med_overview01.htm&docsetVersion=14.3&locale=en

8) **New in Stata 15**

The newest release of Stata presents users with many new statistical techniques. There are two new features that might be especially useful for Cornell researchers. The generalized structural equation modeling command **gsem** now can be used to conduct latent class analysis. Bayesian generalized linear mixed-effects models, can now be run with the **bayes** prefix. The new command **menl** can estimate non-linear mixed-effects models. For more information about Stata's new capabilities, visit https://www.stata.com/new-in-stata/.

9) **New in SPSS 25**

Among the new features in the latest edition of SPSS 25 is a suite of Bayesian Procedures, which allow uses to run several different types of general linear models through a Bayesian framework. (Reference: https://www.ibm.com/support/knowledgecenter/SSLVMB_25.0.0/statistics_casestudies_project_ddita/spss/tutorials/bayesian_intro.html)