Cornell University
Cornell Statistical Consulting Unit

Stats Happening – May 2016

1) **CSCU Summer Schedule**
2) **Data Carpentry Workshop at Cornell**
3) **Summer 2016 CSCU Workshops**
4) **The ASA's statement on p-values**
5) **How does R Treat Missing Values?**
6) **Nonparametric Pairwise Comparisons in SPSS**
7) **When Random Effects are ignored in JMP 12**
8) **Why Propensity Scores should not be used for Matching**
9) **Assessing Preference when Items are Presented in Pairs**
10) **Mixed-Effects Parametric Survival Time Model in Stata 14**

1) CSCU Summer Schedule

CSCU will be available for appointments and walk-in consulting throughout the summer as usual.

- Appointments: Scheduling of appointments is encouraged for matters that require more than 10 minutes of time. To schedule an appointment, contact a staff statistician or visit: http://www.cscu.cornell.edu/about/appointment.php.

- Walk-in Consulting: Staff statisticians will be available for questions that take ten minutes or less daily
  - 11:00-11:30am in B11 and B18 Savage Hall
  - 1:30- 2:00pm in B07, B09, and B13 Savage Hall.
  - Mann walk-in hours will not be held in this summer, but will resume with classes in the fall.

2) Data Carpentry Workshop at Cornell

The Cornell Statistical Consulting Unit (CSCU) is presenting a Data Carpentry workshop on June 13th and 14th. We will be using Data Carpentry's teaching materials, which are designed to teach **fundamental concepts, skills and tools for working more effectively with data**. Our workshop is aimed at academic researchers in all fields and at all career stages. We will cover the following topics:

- Using spreadsheet programs (such as Excel) more effectively
- Cleaning large and messy data sets with OpenRefine
- Using databases, including managing and querying data in SQL
- Aggregating and analyzing data with dplyr in R
- Visualizing data with ggplot in R
- Programming in R
- Generating automatic reports using R Markdown

To allow for coverage of more advanced R topics, we require that participants be familiar with R or attend CSCU's free workshops *Introductory Statistical Analysis in R* and *Intermediate Statistical Analysis in R* offered on June 9th and 10th.

Participants will be able to work on library computers, but are encouraged to bring and use their own laptops to ensure the proper setup of tools for an efficient workflow once you leave the workshop.

To help defray the costs there will be a $40 registration fee, with some allowances for scholarships. For more information on workshop content, prerequisites, and to register, see http://erdavenport.github.io/2016-06-13-cornell/ or contact cscu@cornell.edu .

3) Summer 2016 CSCU Workshops

CSCU will also be offering the following free workshops this summer**.** For more information and to register visit https://cscu.cornell.edu/workshops/schedule.php.

Date and Time	Workshop	Location
Monday, June 6 10:00am – 12:00pm	**Basic Data and Research Skills**	200 Savage Hall
Thursday, June 9 10:00am – 12:00pm	**Introductory Statistical Analysis Using R**	Mann Library MannB30A
Friday, June 10 10:00am – 12:00pm	**Intermediate Statistical Analysis Using R**	Mann Library MannB30A
Part 1: Wednesday, June 15 10:00am – 11:30am	**Interpreting Linear Models** **(A two part workshop)**	Mann Library MannB30B

Part 2: Thursday, June 16 10:00am – 11:30am		
Friday, June 17 10:00am – 12:00pm	**Introductory Statistical Analysis with a User-Friendly Software**	Mann Library MannB30A
Part 1: Tuesday, June 21 10:00am – 12:00pm	**Introduction to Multilevel Modeling** **Part 1:** Key Concepts and Applications to Clustered Cross-sectional Data	100 Savage Hall
Part 2: Tuesday, June 21 2:00pm – 4:00pm	**Part 2:** Multilevel Models for Longitudinal Data	100 Savage Hall
Part 3: Wednesday, June 22 2:00pm – 4:00pm	**Part 3:** Hands-on	Mann Library MannB30A

4) The ASA's statement on p-values

Many recent discussions and published articles question the validity of practicing science using a strict cutoff of p=0.05 to denote significance of effects or predictors. The board of the American Statistical Association (ASA) recently developed a policy statement on p-values and statistical significance. While this statement does not reflect new attitudes about these issues from the field of statistics, the ASA board hopes it will draw attention to changing the practice of science regarding statistical inference. In non-technical terms, the statement articulates a few select principles surrounding p-values and their use, which we have included below as they are a good reminder:

- P-values indicate how incompatible the data are with a specified model (usually the null hypothesis of no difference, no relationship, etc…).
- P-values do not measure the probability that a hypothesis is true, or the probability the data were produced by random chance.
- Research conclusions should not be based on whether a p-value passes a certain threshold (this represents a false dichotomy).
- The concept of statistical significance does not directly relate to the magnitude of an effect or importance of a result.
- A p-value alone does not provide good measure of evidence regarding a model or hypothesis.
- Proper inference requires full reporting of the results of all analyses conducted (no cherry-picking), as well as all initial hypotheses explored and all decisions about data collection and management.

The statement includes a list of references for further exploration of these issues. Read the full statement here: http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108

5) How does R Treat Missing Data?

Different functions in R have different ways of dealing with the presence of missing data. We recently were reminded that some of the function do not deal with missing values in a way that one would expect. For example:

- *rank()* will assign the missing values the highest rank; if there is more than one missing value present, each missing value will be assigned a unique rank determined by the order in which they appear in the dataset.
- *mean()* will output NA if any missing values are present—but this problem can be solved using the na.rm=TRUE option or by using *summary()*.
- *lme()* (from the nlme package) will not execute if you have missing data, however *lmer()* (from the lme4 package) will run if there is missing data.

Before you do statistical analysis, it is important to explore your data and take note of any missing values and check how various functions deal with missing data.

6) Nonparametric Pairwise Comparisons in SPSS

SPSS can do multiple comparisons for the nonparametric Kruskal-Wallis test, the non-parametric equivalent to the one-way ANOVA test. Within the Nonparametric tests menu, there is an option for "Independent Samples" and a legacy dialog of "K Independent Samples". Both will perform a Kruskal-Wallis test; however, the legacy dialog is unable to calculate pairwise comparisons whereas when using the "Independent Samples" option you can obtain them by double clicking on the output titled *Hypothesis Test Summary*. This will open a new window called a model viewer. Here, pairwise comparisons will be calculated but only if the Kruskal-Wallis test was found to be significant. For more information, visit http://www-01.ibm.com/support/docview.wss?uid=swg21479073.

7) When Random Effects are ignored in JMP 12

Be aware that JMP is unable to incorporate random effects in generalized linear models such as logistic regression and Poisson regression despite the fact that you can specify random effects. When executing a generalized linear model with specified random effects, the random effects will be simply ignored but no error message will appear. Note that for continuous responses, JMP is able to estimate linear mixed effects models appropriately. Random effects are also ignored when using the response screening personality.

8) Why Propensity Scores should not be used for Matching

Propensity score matching (PSM) is an incredibly popular method for reducing the imbalance between treatment and control groups in observational studies prior to testing for causal effects. The goal of PSM is to try to mimic the results from a randomized study by selecting a control group that is well matched to the treatment group with respect to a large number of covariates. A recent article by Gary King (Harvard University) and Richard Nielson (MIT) suggests that in practice, the use of propensity scores for matching purposes may actually increase imbalance, inefficiency, model dependency, researcher discretion, and statistical bias, the exact opposite of its intended goal. Through simulation studies and the re-analysis of previously published papers, the authors demonstrate the issues with PSM matching and how the problems are magnified when the data is pruned following the matching procedure, a phenomenon they term the *PSM paradox*. The paper provides recommendations for preferred matching procedures as well as suggestions for researchers that would like to use PSM despite its flaws. The paper can be accessed online at: http://gking.harvard.edu/files/gking/files/psnot.pdf?m=1456683191 and a nice presentation of the paper by Gary King can be viewed online at: https://www.youtube.com/watch?v=rBv39pK1iEs

9) Assessing Preference when Items are Presented in Pairs

The Bradley-Terry model can be used when assessing the preference, or performance, among items that were presented in pairs. This model is applicable when analyzing a sports tournament, seed preferences of insects, or color preferences of children. The Bradley-Terry model implements a logistic regression assuming that the multiple trials of a given pair are independent with fixed probability of preferring one item and that the evaluations of different pairs are independent. In SAS, this model can be estimated using Proc Logistic or Proc Genmod. For more information on the execution of this model, and how the data must be organized, refer to this example from SAS's support website: http://support.sas.com/kb/24/992.html.

10) Mixed-Effects Parametric Survival Time Model in Stata 14

In Stata 14, the new **mestreg** command can estimate a mixed-effects parametric survival time model. The conditional distribution of the response given the random effects can be specified as Weibull, lognormal, loglogistic, or gamma. The model can be formulated using the proportional-hazards parameterization or the accelerated failure-time (AFT) parameterization. A likelihood ratio test is provided at the end of the output, which compares the fit of the model with the specified random effects to the fit of the model

with only the fixed effects. For more information and examples on this type of model, refer to http://www.stata.com/manuals14/memestreg.pdf.