Cornell University
Cornell Statistical Consulting Unit

Stats Happenings – May 2017

---

1) Data Carpentry Workshop at Cornell

The staff of the Cornell Statistical Consulting Unit (CSCU) will be presenting a Data Carpentry workshop on June 14th and 15th. 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:

- Best practices for data management in spreadsheets
- Cleaning big, messy data in OpenRefine
- Manipulating and rearranging data in R with dplyr and tidyr
- Visualizing data in R with ggplot
- Reproducible research in R: loops, functions, and automatic reports with knitr and R markdown
- Introduction to data analysis and visualization in Python

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 12th and 13th.

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.

Registration is on a first come first served basis. Space is limited! To help defray the costs, there will be a $40 registration fee, with some limited allowances for scholarships. For more information on workshop content, prerequisites, and to register, see https://www.cscu.cornell.edu/workshops/data_carpentry.php or contact cscu@cornell.edu.

2) Summer 2017 Workshops

In addition to the data carpentry workshop, CSCU will be offering the following workshops this summer.

Workshop	Date and Time
Basic Data and Research Skills	June 7, 2pm-4pm, Savage 200
Introductory Statistical Analysis with a User-Friendly Software	June 9, 10am-12pm, Mann B30B
Introductory Statistical Analysis Using R	June 12, 10am-12pm, Mann B30A
Intermediate Statistical Analysis Using R	June 13, 10am-12pm, Mann B30A
Interpreting Linear Models **(A two part workshop)**	June 21, 1pm-3pm, Mann B30B June 22, 1pm-3pm, Mann B30B
Introduction to Multilevel Modeling **(A three part workshop)**	June 19, 10am-12pm, Savage 100 June 20, 10am-12pm, Savage 100 June 20, 1pm-3pm, Mann B30A

For more information about our workshops, and to register, please visit https://cscu.cornell.edu/workshops/schedule.php.

3) The American Statistical Association's Recommendations for Reproducible Research

In research, reproducibility refers to the ability to reproduce all numerical findings from a study, given the same set of data. This is typically achieved by writing computer code for data management and analysis. The American Statistical Association (ASA) has developed a set of guidelines for researchers to improve the reproducibility of their studies. The guide outlines barriers researchers may face in making their research workflow reproducible, including limited skills, time, and incentives. Additionally, recommendations for reproducibility training, changes to funding mechanisms, and changes to incentive systems to promote reproducible research are provided. (Resource: **https://www.amstat.org/ASA/News/ASA-Develops-Reproducible-Research-Recommendations.aspx**)

4) **New CSCU Handout: Estimating Marginal Means in Various Statistical Software Packages**
After running a model, conducting post-hoc analysis by estimating marginal means for categorical variables often follows. Certain post-hoc analyses may be difficult to execute in some statistical programs. We have created a reference guide for users of R, Stata, SPSS, SAS, and JMP on the various post-hoc analyses, which can be accessed by visiting: http://cscu.cornell.edu/news/Handouts/Post.pdf.

5) **Creating Attractive Tables from Model Output in R and Stata**
The *stargazer* package in R can be used to create well-formatted univariate and bivariate summary tables as well as tables containing parameter estimates from multiple models. Tabout is a command in Stata that can also be used to create high quality tables from model output. Both can produce LaTeX code and ASCII text for the tables.

6) **Using Complex Survey Weights in Stata 14**
Stata 14 now allows researchers to use complex survey weights in multilevel models. This allows weights to be assigned to the different levels of the sampling design. Generalized linear models, parametric survival models, and generalized structural equation models support complex survey weights.
(Reference: http://www.stata.com/new-in-stata/multilevel-models-survey-data/)

7) **New in JMP 13: Latent Class Analysis**
Latent class analysis allows researchers to find underlying latent clusters from a set of categorical variables. Each subject is then assigned to a particular cluster based on predicted probabilities, which can then be used in further analyses.
(Reference: https://www.jmp.com/en_dk/software/data-analysis-software/new-in-jmp.html#Latent-Class-Analysis)

8) **Caution When Using *as.numeric()* in R**
The as.numeric() command in R should be used with great caution! Researchers often have a variable that R recognizes as a character or factor variable, but which they would like to be recognized as a numerical (continuous) variable for analysis purposes. If the variable is currently a character variable, the *as.numeric()* function is appropriate, but if

the variable is a factor variable, *as.numeric()* will reassign each value in the variable with its alphanumeric level. Instead, the nested functions *as.numeric(as.character())* must be used. To illustrate the issue, consider this simple example of a numeric variable forced to be a character and then a factor. The *as.numeric()* command is appropriate for the character but not the factor.

9) What We Are Reading

The following recently published books discuss the ins-and-outs of data analysis. The topics in these books range from big data exploration in linguistics, criminal justice, and neuroeconomics to the history of mathematical and statistical theorems and everything in between.

- Nabokov's Favorite Word is Mauve (Ben Blatt)
- Living by Numbers (Steven Connor)
- Dear Data (Giorgia Lupi)
- The Theory That Would Not Die (Sharon McGrayne)
- Weapons of Math Destruction (Cathy O'Neil)
- The Art of Risk (Kayt Sukel)
- The Seven Pillars of Statistical Wisdom (Stephen Stigler)

10) 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: https://www.cscu.cornell.edu/consulting/appointment_form.php.

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