**AML 612: Statistical Methods for Students in the Life and Social Sciences**

**(aka: How to be a Data Boss)**

**Instructor: Dr. Sherry Towers** ([smtowers@asu.edu)](mailto:smtowers@asu.edu))

**Objectives:**

This course is meant to introduce students in the life and social sciences to the skill set needed to do well-executed and well-explicated statistical analyses. The course is aimed at students with little prior experience in statistical analyses. The course will be almost entirely based on material posted on the website [www.sherrytowers.com](http://www.sherrytowers.com)

Upon completing this course:

Students will have an understanding of basic statistical methods, including hypothesis testing, linear regression and generalized regression methods, and will understand common pitfalls in statistical analyses, and how to avoid them.

Students will have basic functionality in R, and will learn how to read in, manipulate, and export data in R, and will be able to create publication-quality plots in R.

**The skills needed to produce publication-quality written presentation of work will be strongly, and repeatedly, stressed throughout the course, as will verbal presentation skills. Demonstration that the student has learned and become adept at these skills will be an important factor in their final grade.**

There will be regular homework projects assigned throughout the course, which will be worth 50% of the grade. Students are strongly encouraged to work together in groups to discuss issues related to the course and help each other. However, plagiarism of code will not be tolerated.

There are usually homework assignments every two weeks or so. There is also the potential for short, in-class pop quizzes during the semester; if these occur, they will be included in the homework grades.

The culmination of the course will be a group term project (worth 50% of the grade) that involves statistical analysis of data sets identified by the students. The students will write-up the results of their project in a format suitable for publication (submission for publication is not required, but encouraged if the analysis is novel), and will perform an in-class presentation of the results. Students are responsible for developing an appropriate research question, so this should be something they begin to think about early in the course.

**Students are expected to bring their laptops to class. Before the course begins, students are expected to have downloaded the R programming language onto their laptop from http://www.r-project.org/ (R is open-source free software).**

**Course topics:**

Basics of the R programming language

Locating free data:

* Online data
* Datathief

Overview of probability distributions

* Uniform
* Normal
* Poisson
* Exponential
* Gamma
* Negative Binomial
* Binomial
* Beta Binomial
* Student’s t
* Chi-squared

Hypothesis testing

Ordinary least squares linear regression

Visual methods to check fit assumptions:

* QQplots to check for Normality of fit residuals
* Plots to check linearity assumptions are upheld
* Checking for homoscedasticity

Transformation of regression variables (log transforms, logit, etc) to ensure linearity and homoscedasticity assumptions of regression methods are upheld.

Including categorical variables in a fit

Including population standardization in linear regression fits when fitting per-capita incidence rates

Generalized linear models:

* Poisson linear regression for count data
  + Population standardization when using Poisson linear regression
* Negative Binomial regression for over-dispersed count data
* Logistic regression for Binomial data
* Regression methods for over-dispersed Binomial data

Be in the data analysis Hall of Fame, by learning about and applying these:

* Always checking to see if model assumptions are upheld
* Avoiding overfitting
* Checking for multi-collinearity, and dimensionality reduction
* Model selection (AIC, likelihood ratio test for nested models)
* The critical importance of model validation

Don’t be in the data analysis Hall of Shame, by learning about and avoiding these:

* p-hacking (aka ``fishing for significance’’)
* Subtle, and not-so-subtle methods for ``cherry picking data’’.

Producing well-written academic papers

How to give a good presentation

There will be other statistical analysis topics covered, if there is time, including:

Harmonic linear regression methods for periodic data

Kolmogorov-Smirnoff test for testing similarity of two distributions

Kmeans clustering

R Shiny