# Statistics: Understanding and Using Data

DHSS-GA.1100

*(Note: This course will not be offered in the academic year 2017/2018.)*

---

## Course description:

The course will give the student a project-based understanding of how to draw inferences from data, or the science of statistics. Statistics studies how data can be used to help answer relevant questions. The course is about how to pose such questions and how to interpret the answers returned. The course will introduce students to a powerful tool for dealing with data: the programming language R. The course will stress applications of interest to students with a liberal arts background. The design of the course is highly interactive, and stresses the completion of projects which use the various skills that are taught. There is no prerequisite to this course.

Students will use R to work on a variety of data-based projects. These projects may include: racial discrimination in employment, forecasting election results, changing minds on gay marriage, civilian victimization during wartime, programs to reduce poverty, predicting authorship of the Federalist Papers, analyzing the preambles of constitutions, marriage networks in Renaissance Florence, international trade networks, election fraud in Russia, the 1932 German election in the Weimar Republic, human rights around the world, and others.

## Sample Projects Using R

Data analysis approach on human rights; visualizing data from the IMDB data set; election forecasting; gender studies; text analysis for authorship attribution; modeling citation networks in academic publishing; studying urban indicators; assessing data based on the climate system; and others.

## Course Philosophy

Learning statistics involves learning some things that are best taught using traditional lecture and text, but most learning takes place when students actually apply and interpret what is taught in the context of topics and data of interest. This is accomplished in this course in various ways and using various existing technologies. In particular:

a) The course will emphasize understanding of statistical models rather than the computational methods needed to implement those models.

b) In learning about R, students will work through available online tutorials and the various code examples provided in the texts

c) In working with the readings, students will be expected to read and annotate texts before class, and to discuss the material with other students before class (with submissions of such using Perusall, https://perusall.com)

d) Class meetings will largely take the form of discussions, sometimes of the class as a whole and sometimes dividing into small groups, rather than lectures.

e) While students may attend lab sessions in person, it is expected that most students will attend the lab sessions virtually, with the instructor and students being able to share screens so that the labs can be maximally useful and so that students can use their own computers and need not travel to the NYU campus to attend the lab sessions

f) The instructor will hold a weekly “virtual office hour” wherein s/he can work with small groups of students with the students and instructor sharing computer screens and interacting using standard software such as Skype

## Format

The class will meet weekly for 2.5 hours, with 1.25 hours allocated to discussion of class readings and 1.25 hours allocated to a project-based lab. There is an additional weekly “virtual office hour” wherein students and the instructor can share screens.

## Software

By using R, the student will learn the structure of programming (particularly object oriented programming) as well as a very flexible tool which can serve many future needs. Since R consists of a common language and a variety of packages, students can learn how to work with a variety of data using this common language. In general R provides more flexibility than various proprietary packages, and students who know R will be at an advantage relative to those who only have worked with specific proprietary packages. In addition, R is free and works on any operating system. Along with Python, R gives the student an excellent base to pursue any data oriented career. R can be obtained for all relevant operating systems at CRAN (https://www.r-project.org). Students will access R through R-Studio (https://www.rstudio.com) which allows for easy access to R packages and various other features of R.

## Readings

The principal course text is Imai, A First Course in Quantitative Social Science, Princeton University Press, 2016 (QSS in the weekly assignment list). This text will be supplemented with Arnold and Tilton, Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text, Springer 2015, and with various R based tutorials and manuals which are available either freely online (url provided in weekly assignment list) or as e-books or pdf’s available online through Bobst Library, referred to as Bobst in the weekly assignment list).