Soci 252-002 – Data Analysis

Spring 2017

Professor François Nielsen

The Big Picture of Data Analysis (SOURCE: OLI)

SOCI 252 Data Analysis (3). Required of sociology majors. Methods of data analysis: descriptive statistics, elements of probability, and inferential statistics and multivariate analysis to permit causal inference.

This class is an introduction to data analysis and statistical inference in a sociological perspective. The course has no prerequisite. (In particular, you do not need to have previously taken Soci 251 to enroll.) The course satisfies the Quantitative Intensive Research Intensive (QI) requirement of the College. This semester we cover the materials over fourteen classes Fri 9:00am–11:30am. The class meets in Hamilton 151.

The format of the class is unusual in that we make extensive use of online materials developed by the Open Learning Initiative (OLI) project at Carnegie Mellon University. During a typical class period students progress through the online curriculum, take short “checkpoint” tests online, and carry out data analysis activities (online and off) using the powerful statistical program R, which can be downloaded free of charge (there are versions for Windows, the Mac, and Unix). There is an in-class final exam, also online.

There is no textbook for the class, but there is a $25 charge for registration to the OLI course site using the course code that will be provided by the instructor prior to the beginning of class. Students are required to bring to class a laptop capable of accessing the internet and running the R program.

1See [http://oli.cmu.edu/courses/free-open/statistics-course-details/](http://oli.cmu.edu/courses/free-open/statistics-course-details/) for more details. Click on Enter Course (twice) to access all course materials (except graded tests) without registering.
Contents

1 Times, Places & Contacts 2
2 Course Description and Goals 2
3 Three Things to Do ASAP 4
4 Requirements and Grading 5
5 How to Do Well in This Course 6
6 Frequently Asked Questions (FAQ) 7
7 Outline and Schedule 7

1 Times, Places & Contacts

Classes Meet

- Fri 9:00am–11:30pm in Hamilton 151.

Instructor

- Professor François Nielsen – Email: francois_nielsen@unc.edu Office: Hamilton 163 Hours: by appointment.
  
  The best way to contact me is to talk to me during class; we can step out into the hallway for a private conversation.

- Web site http://www.unc.edu/~nielsen Click on the Soci252 link to access this syllabus and other materials.

- OLI web site http://oli.cmu.edu You will go to this site to do most of the online work in the course.

Teaching Assistant

- There is no TA for this class.

2 Course Description and Goals

Overview

The course covers descriptive statistics and graphs for exploratory analysis of data distributions, including descriptive uses of normal distributions; exploratory approaches to association for both categorical and continuous variables; elements of probability theory up to and including Bayes’ Theorem, random variables and the algebra of expectations; probability distributions for continuous variables with special emphasis on the normal distribution; sampling distributions; principles of statistical inference, including confidence intervals and hypothesis testing, in large and small samples; inference for population means and proportions; the use of the chi-square statistic for tests of goodness-of-fit, independence, and homogeneity for categorical variables; analysis of variance; simple and multiple linear regression including statistical inference, some discussion of data transformations, and the analysis of residuals. The course includes learning how to use the microcomputer for data entry and analysis using the statistical program R.
Goals

Our society is besieged by seemingly unlimited information. The ubiquitous mass media, the computer, the Internet, and the worldwide web have contributed to the exponential growth of information. A significant proportion of the information is quantitative in nature. This mass of quantitative information has not gone unnoticed. There is scarcely a day that goes by that a newspaper, magazine, a television or radio program fails to use statistical information as part of their coverage of events. Debates are influenced by numerical data. Statistical reasoning is becoming a basic skill of an educated person.

The primary purpose of this course is to increase your data analysis and statistical literacy. Even if you do not continue your formal schooling after college, such training will make you a more informed citizen and a more intelligent consumer of information. If you do further your education, many professional and graduate schools expect their applicants to have quantitative training. So this course also provides important preparation for advanced education and for carrying out your own empirical research in any field of study.

Especially if you are a Sociology major you may have already taken Soci 251 (although Soci 251 is not a prerequisite for this class) and learned about research design, the collection of data, and measurement. In this course our attention is directed to the analysis of data. At the most general level any type of empirical analysis consists of the four steps represented in the figure The Big Picture of Data Analysis shown on the front of this syllabus.

1. Data Production In this step data are produced, either in an observational study based on a sample from a population, or through an experimental design with random allocation of subjects to treatments. We treat related issues of sampling and experimental designs relatively briefly in this course. If you have taken Sociology 251 you have covered these materials more thoroughly so this unit will be mainly review.

2. Exploratory Data Analysis We first distinguish the major measurement types of variables that you will encounter, namely categorical and quantitative variables. We follow this with an extended discussion of univariate descriptive statistics and graphs. These are techniques to describe the distribution of a single categorical or quantitative variable. From one variable we move to descriptive statistics and graphs to describe bivariate relationships, where we simultaneously consider two variables and their association. Our treatment of descriptive statistics concludes with a look at multiple variables.

3. Probability Our ultimate goal beyond descriptive analysis is to be able to draw conclusions from our data about the population to which we wish to generalize our descriptive findings. To do this we need to focus first on developing some important principles of probability theory. This part of the course may feel more abstract compared to what came before, but the concepts presented here are very important in all sorts of fields and, I believe, of general and profound intellectual interest.

4. Statistical Inference The remainder of the course concerns the use of models and statistical inferences. Here we study the problem of making inferences beyond the data at hand to characteristics of the population from which the data is obtained. Using the principles of probability that we learned in the previous step we can understand what happens when we estimate a characteristic of a population (such as a mean, or a proportion, or a regression coefficient) from a random sample, and how we can assess the reliability of this estimate. This process is called statistical inference. We first consider making inferences about single variables and then move on to inferences about bivariate and multivariate relations. Throughout the course
we will make frequent use of sets of real data.

These four steps of data analysis will constitute the overall outline of the course, with one modification: we will begin with Step 2 (Exploratory Data Analysis) before moving on the Step 1 (Data Production). The detailed outline is found below under Outline and Schedule.

3 Three Things to Do ASAP

As soon as possible at the start of the semester you should (1) open a student account and pay the $25 course registration fee at the OLI Carnegie Mellon site (http://oli.cmu.edu/), (2) check your system setup at the course site, and (3) install the free R statistical program on your laptop.

(1) Creating an OLI Student Account

Follow the steps below to register for the course and pay the $25 registration fee.

1. Go to the Open Learning Initiative (OLI) website: http://oli.cmu.edu/
   In the upper right hand corner of the site, click “Sign Up” and fill out the form.

2. On the “Confirm Your Account Information” page, review the account information you entered. If everything is correct, click the “Confirm Account” button. If not, click “Edit Account” to make your changes.
   Important Note: the only account setting that can’t be changed after you confirm your account is your Account ID.

3. Read the statements in the “Online Consent Form” and select “I Agree” or “I Do Not Agree” then select “Submit.”

4. Under “My Academic Courses” enter your Course Key and click “Go”.
   Your Course Key is soci252s17

5. Review the course details to make sure that you are registering for the correct course, and click “Register”.

6. You will reach the “Payment Required” page. Click on the “pay by credit card” button.

7. On the Carnegie Mellon credit card payment page enter your billing information and click “Submit”. If your transaction is successful, you will see a receipt page. Save this page for your records.

8. Click the return link at the bottom of the receipt page and you will be taken back to your OLI home page.

9. Under “My Courses” you will see your registered course. You can double-check to make sure the correct course appears by confirming that the correct Course Key and instructor are listed.
   Important Note: After you complete registration you will not have immediate access to the course, as I will have to authorize your registration. For your registration to be authorized your name will have to be on the class roster. I will be checking for pending authorizations at least every evening for one week after I post this syllabus at the beginning of the semester.

---

If you are not ready to take this step, remember that you can access all the course materials (except graded “checkpoints”) without registering (see previous footnote). However the work you do will not be saved. Registration gives you access to graded quizzes, records your progress in the curriculum, and gives the instructor access to these data.
(2) Perform a System Check

Once you have registered to OLI you should perform a system check of your computer. To do this go to the OLI site at [http://oli.cmu.edu/](http://oli.cmu.edu/) and sign in to your account. Click on the System Check link, and then on the Test Your System button. OLI will check the settings of your browser and inform you of any aspect of your system that may create problems in doing the course work online.

(3) Install the R Program

For statistical work we will use the program R that is freely available. You should install R on the laptop that you bring to class as early as possible at the beginning of the class. To install the program go to [http://cran.r-project.org/](http://cran.r-project.org/) Click on the appropriate link and follow the instructions to download the version of R for the Mac (OS X), Windows or Linux.

In the course of installation you will be requested to choose a mirror site. Choose a U.S. site geographically close to here, probably the MD site which is in Bethesda, MD.

The installation program will automatically create an icon for the program on your desktop. (If you have a 64 bits system there will be two icons, one for 32 bits version and one for the 64 bits version. You can use either.)

I recommend that you also install John Fox's car package. (If you are using Windows you should first close R; then restart R in “administrator mode” by right-clicking on the desktop icon and selecting Run as administrator on the local menu.) To install the car package enter the following command at the prompt in the R console

```
> install.packages("car", dependencies=TRUE)
```

Note that you can simply select and copy the command above (without the prompt) with Ctrl-C and enter it in R at the prompt (>) with Ctrl-V and press Enter. You will be able to access the car package later with the command

```
> library(car)
```

For most of the computer work in the class we will simply access the data we need by double-clicking on a data link in the OLI curriculum and allowing R to open the data set. Additional instructions on how to use R for the course will be provided in class and in a document available at [http://www.unc.edu/~nielsen/soci252/software/soci252_RwithOLI.pdf](http://www.unc.edu/~nielsen/soci252/software/soci252_RwithOLI.pdf) (under construction).

4 Requirements and Grading

Attendance in class is required. Class work will consist for the most part of students individually progressing through the materials online. Studies have shown that the OLI curriculum is highly efficient in terms of time. It is possible for the faster students to complete the course work entirely during class time. Catch up work and additional assignments can be completed at home in the evenings and weekend. You can work on most materials outside of class, however all graded tests (checkpoints) are to be completed exclusively in class. Each checkpoint is set up so that there is only one attempt and a time limit of 30 minutes. Class activities will also include discussions by instructors of topics that may be causing difficulty for students and special presentations.

Work on the curriculum and on the StatTutor labs done before the evening of the final day of class will be credited in your grade. On the evening of the final I will download the gradebook from the OLI site and work done after that will not be credited.

To summarize, your grade in the course will be based on the following items.
1. Class Work (30%). You will get credit for all class materials in all 17 modules (including Do I get This? and Learn By Doing items) that you have completed with a reasonable degree of quality. You can make mistakes and still get a maximum score for this component.

   **Guidelines.** (1) You must do the class work yourself, but you can receive help from anyone in completing it. (2) You can work on the materials, in and out of class, until the evening of the final day of class; work completed later will not be credited.

2. Stat'Tutor Labs (15%). This component of the grade is evaluated as the percentage completed of all seven Stat'Tutor Labs, as recorded in the OLI gradebook. The labs consist of guided analyses of a real data set, where you are prompted through all the stages of the analysis, from deciding on the appropriate method through the actual statistical analysis using R, to drawing substantive conclusions from the results.

   **Guidelines.** (1) You must do the lab work yourself, but you can receive help from anyone in completing it. (2) In order to get credit for Lab work you should copy the R text output that you produced and paste it as part of your answer in your answer window. (You will not be able to insert graphs in the window, however.) (3) You can work on the labs, in and out of class, until the evening of the final day of class; work completed later cannot be credited.

3. Graded Tests (Checkpoints) (40%). This component of the grade represents your total score on all the (30) checkpoints tests distributed throughout the course. Note that each test is short (some 10 multiple choice questions), but the total number of items is quite large so this score is likely to reliably reflect your understanding of the materials.

   **Guidelines.** (1) You must complete the checkpoints yourself; help from another person is not allowed, although using a calculator or any notes is allowed. (2) Checkpoints are to be completed exclusively in class and during class time. (3) There is a time limit of 30 minutes to complete each checkpoint, with the quiz shutting off automatically at the end of the allocated period. (4) The instructor is available for questions and may provide clarifications or hints; however you remain solely responsible for your answer to a quiz question.

4. Final (15%). The final exam takes place in class during the last day of class. Presence in class at the final is required. The final is online so you need to bring your laptop as usual. You will have one hour and 15 minutes to complete the final, which consists of 40 multiple-choice questions.

   **Guidelines.** (1) You must complete the final yourself; help from another person (including remote help) is not allowed, although using a calculator or any notes is allowed. (2) The final must be completed in class. (3) There is a maximum time of 75 minutes to complete the final.

5. Class Non-attendance Penalty. Presence in class is required; attendance will be recorded. There will be a penalty levied for insufficient attendance not to exceed 5 percentage points taken off the total grade.

### 5 How to Do Well in This Course

- Pages 1–5 of the OLI curriculum includes a discussion of effective strategies for working through the course, and of the “big picture” of statistics. Do not skip these materials!

- The first unit of the course (Exploratory Data Analysis) covers techniques such as the histogram and the boxplot that you may have learned before in high school, or even earlier, and you may be tempted to rush through that part to get to more advanced materials sooner. Resist that temptation! The concepts taught in these first modules are absolutely

---

*Excluding Module 3, which has no interactive component.*
crucial for understanding everything that follows. For example, to make sense of the rest of the course you will need a thorough, expert understanding of how a histogram or a boxplot describe the distribution of a quantitative variable, or how to use the standard deviation rule (also called the 68%-95%-99.7% rule) to scale a normally-distributed variable.

- Course materials are inherently cumulative. Concepts learned in one module are needed to understand later modules. Thus it is most effective to work through the modules one after another, completing all activities for one topic before moving on to the next.

- Some of the materials are presented as optional. Even if you feel that you do understand a concept, I would recommend that you do these optional materials anyway, as this will test your knowledge and contribute to strengthen it. Also, you'll get credit for them in your course grade.

- There are no useful short cuts. The character of Cheryl Strayed in the movie Wild decides to hike a thousand miles along the Pacific Crest Trail. Her attitude is that you get on the trail at the beginning and then put one foot in front of the other until you reach the end. The best attitude in approaching the course is the same: you get into the curriculum at one end (the beginning) and work your way through, one step at a time, until you get out at the other (the end). During the trip, relax! Enjoy the new insights and intellectual vistas that open up. When you get off the trail you will have learned a great deal.

- Plan to make optimal use of class time. There is nothing wrong in checking your email periodically, but it is not a good idea to spend all your time on Facebook. When you get tired of working on class materials, take a break; take a nap if you need to. Don't get far behind the schedule suggested in the outline.

- Don’t use the graded checkpoints to test yourself. Instead use the non-graded materials (the Did I Get This? and Learn by Doing activities) to test your own knowledge. Do not take the graded test until you have completed all the activities that precede it (including StatTutor labs) and you feel you have the best understanding of the materials you can achieve at that point.

- One approach that does not work is to jump to a graded test, discover that one does not know what it is all about, and then frantically going back over the course trying to figure out where the relevant passages are. This does not work well in general, but it is guaranteed to fail within the 30 minutes time limit of the checkpoints.

6 Frequently Asked Questions (FAQ)

This section is in preparation. Topics to be discussed here in the future include the following.

- I am not very good at math. How difficult will this course be for me?
- Why no textbook?
- If I wanted to have a textbook as a reference, which should I use?
- I have learned about histograms and scatterplots before – in high school or college. Is it OK to skip this material? (BTW, the answer is NO! See How to Succeed in this Course above.)
- Why is class attendance required?
- Where can I find help using R?

7 Outline and Schedule

Activities listed in this suggested schedule are to be completed by the end of the corresponding class. Modules are prefixed with the corresponding unit name as EDA = Exploratory Data
Analysis; PD = Producing Data; PROB = Probability; INF = Inference. Module names and checkpoints refer to the syllabus on the OLI course site.

Class 1 – Fri 13 Jan

• Module 1 Introduction
• Module 2 Learning Strategies
• Module 3 The Big Picture
• Module EDA 4 – Examining Distributions
  ◦ Checkpoint 4.1
  ◦ Lab – Drinking Habits and Integrity of College Students
  ◦ Checkpoint 4.2

Class 2 – Fri 20 Jan

• Module EDA 5 – Examining Relationships
  ◦ Checkpoint 5.1
  ◦ Lab – Body Image and Academic Performance of College Students
  ◦ Checkpoint 5.2

Class 3 – Fri 27 Jan

• Module PD 6 – Sampling
  ◦ Checkpoint 6.1
• Module PD 7 – Designing Studies
  ◦ Checkpoint 7.1
  ◦ Lab – Treating Depression: A Randomized Clinical Trial
  ◦ Checkpoint 7.2

Class 4 – Fri 3 Feb

• Module PROB 8 – Introduction (Probability)
  ◦ Checkpoint 8.1
• Module PROB 9 – Finding Probability of Events
  ◦ Checkpoint 9.1
  ◦ Checkpoint 9.2

Class 5 – Fri 10 Feb

• Module PROB 10 – Conditional Probability and Independence
  ◦ Checkpoint 10.1
  ◦ Checkpoint 10.2
• Module PROB 11 – Random Variables
  ◦ Checkpoint 11.1

Class 6 – Fri 17 Feb

• Module PROB 11 (cont’d)
  ◦ Checkpoint 11.2
  ◦ Checkpoint 11.3
  ◦ Checkpoint 11.4
Class 7 – Fri 24 Feb

- Module PROB 12 – Sampling Distributions
  - Checkpoint 12.2
  - Checkpoint 12.3

Class 8 – Fri 3 Mar

- Module INF 13 – Introduction (Inference)
- Module INF 14 – Estimation
  - Checkpoint 14.1
- Module INF 15 – Hypothesis Testing
  - Checkpoint 15.1

Class 9 – Fri 10 Mar

- Module INF 15 (cont’d)
  - Checkpoint 15.2
  - Lab – Cell Phones
  - Checkpoint 15.3

NO CLASS – Fri 17 Mar (Spring Break)

Class 10 – Tue 24 Mar

- Module INF 15 (cont’d)
  - Checkpoint 15.4
  - Checkpoint 15.5
- Module INF 16 – Inference for Relationships
  - Checkpoint 16.1
  - Checkpoint 16.2

Class 11 – Fri 31 Mar

- Module INF 16 (cont’d)
  - Checkpoint 16.3
  - Lab – Analyzing Data from a Course’s Grade Book
  - Checkpoint 16.4

Class 12 – Fri 7 Apr

- Module INF 17 – Inference for Relationships Continued
  - Lab – Risk Factors for Low Birth Weight
  - Lab – Monthly Premiums of Auto Insurance
  - Checkpoint 17.1
  - Checkpoint 17.2

NO CLASS – Fri 14 Apr (Good Friday)

Class 13 – Fri 21 Apr

- Special presentation: Multiple Regression
- Continue and complete all modules as needed.
Class 14 – Fri 28 Apr

- Continue and complete all modules as needed.
- Final exam 10:15–11:30. To take the exam please go to the [ARTIST Site](#). The access code for the exam will be provided in class.