Data Analysis

Spring Semester, 2023

January 26, 2023

Lecture 2

Introduction and Overview

Welcome to AMS 315. This is a second course with AMS 310 or its equivalent as a prerequisite. The course requires that you attend using a relatively high level of computer and computer network.

Zoom Hours

I will hold 3 hours of Zoom contact weekly: on Mondays from 2:30 to 4:00 pm and on Wednesdays from 3:30 to 5:00 pm Stony Brook time. The TAs will also hold 2 hours of Zoom or office contact at the times specified on the Class Blackboard. The Student Accessibility Services Center may also be able to provide you with additional resources.

Free academic support services including one-on-one and small group course-based tutoring, one-on-one skill-based tutoring, peer assisted learning (Supplemental Instruction), and public speaking courses are available for you. Learn more about these services by visiting www.stonybrook.edu/tutoring.

Syllabus Quiz

There is a quiz worth 25 points on the class syllabus available for taking now and due before January 30 at 11:59 pm. To get credit for this quiz, you must download the Respondus browser and use it to take this quiz. You will not get any points if you do not use the Respondus browser. You will be penalized additional points if you do not use the Respondus browser for the first midterm. There will be additional penalties if you do not use the Respondus browser in subsequent examinations.

Examinations

Examinations will be open book and open notes. You may use any calculator that you wish. The pdf file of statistical tables is available on the class Blackboard in the Assignments section. I recommend that you print it out for use in the examinations. This pdf file will also be attached to your examination file. However, in past semesters some students had trouble accessing this file during an examination.

The dates and tentative content of the examinations are:

Feb 23: Examination One: Chapters 3, 4, 5, 6, and 7

Mar 30: Examination Two: Chapters 11 and 12, in addition to Chapters 3, 4, 5, 6, and 7.

Apr 27: ***Examination Three: Chapters 8 and 9, in addition to Chapters 3, 4, 5, 6, 7, 11, and 12.***

May 5: ***Make-up examination for students who missed one mid-term: Chapters 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12.***

May 15: ***5:30 pm-8:00 pm, Stony Brook time. Final examination: Chapters 3-12.***

Projects

The dates of the projects are below. I will liberally grant extensions for project 1. The project 2 due date can be extended a bit, but not much. Please note that the project 2 due date is near the end of the semester.

Thursday, Mar 9:	Project One data posted.
Tuesday, Apr 11:	Project One report is due at 11:59:00 Stony Brook time. I will liberally allow extensions to this deadline.
Tuesday, Apr 18:	Project Two data posted.
Tuesday, May 2:	Project Two report is due at 11:59:00 pm Stony Brook time.

Course Blackboard

Course materials posted to Blackboard will include:

- The syllabus.
  - An introductory video lecture discussing the syllabus and course; please view this video as soon as possible.
  - Zoom videos of class lectures.
  - Pre-recorded Zoom videos of specific problem types.
  - Supplemental readings.
  - Instructions for accessing online tools and components.
  - Information on assignments and assessments.

Supplemental Readings

Scientific papers and other material will be posted on the Class Blackboard. These papers provide case studies and more detailed discussion and illustration of the application of statistical techniques than your text or I can give. They are an important supplement to your studies. You should study them to enhance your understanding of the applicability of the material in the course and how the material of the course is part of quantitative research.

Tentative Schedule Spring 2023 AMS 315
Lectures are from 4:45 to 6:05 pm Eastern Time

Jan 24:	Introduction, Chapter Three, Data Description
Jan 26:	Chapter Four, Probability and Probability Distributions (simulation issues)
Jan 31:	Chapter Five, Inferences about Population Central Values
Feb 2:	Chapter Five, Inferences about Population Central Values
Feb 7:	Chapter Six, Inferences Comparing Two Population Central Values
Feb 9:	Chapter Six, Inferences Comparing Two Population Central Values
Feb 14:	Chapter Seven, Inferences about Population Variances
Feb 16:	Chapter Seven, Inferences about Population Variances
Feb 21:	Chapter Eleven, Linear Regression and Correlation
Feb 23:	***Examination One: Chapters 3, 4, 5, 6, and 7***
Feb 28:	Chapter Eleven, Linear Regression and Correlation
Mar 2:	Chapter Eleven, Linear Regression and Correlation
Mar 7:	Chapter Eleven, Linear Regression and Correlation
Mar 9:	Chapter Twelve, Multiple Regression and the General Linear Model *Project 1 Data Posted*
Mar 14:	Spring recess, no class
Mar 16:	Spring recess, no class
Mar 21:	Chapter Twelve, Multiple Regression and the General Linear Model
Mar 23:	Chapter Twelve, Multiple Regression and the General Linear Model
Mar 28:	Chapter Twelve, Multiple Regression and the General Linear Model
Mar 30:	***Examination Two: Chapters 11 and 12, in addition to Chapters 3, 4, 5, 6, and 7.***
Apr 4:	Chapter Eight, Inferences about More than Two Population Central Values
Apr 6:	Chapter Eight, Inferences about More than Two Population Central Values
Apr 11:	Chapter Eight, Inferences about More than Two Population Central Values *Project 1 Due at 11:59:00 pm Eastern U.S. Time*
Apr 13:	Chapter Nine, Multiple Comparisons
Apr 18:	Chapter Nine, Multiple Comparisons *Project 2 Data Posted*
Apr 20:	Chapter Nine, Multiple Comparisons
Apr 25:	Chapter Nine, Multiple Comparisons
Apr 27:	***Examination Three: Chapters 8, and 9 in addition to Chapters 3, 4, 5, 6, 7, and 12.***
May 2:	Chapter Ten, Categorical Data *Project 2 Due at 11:59:00 pm Eastern U.S. Time*
May 4:	Chapter Ten, Categorical Data
May 5:	***Make-up examination for students who missed one mid-term, 1:00 pm.***
May 15:	***5:30 pm-8:00 pm Eastern US Time. Final examination: Chapters 3-12.***

Grading

The target grade distribution is roughly 25% A, 25% B, and 25% C+, and the remainder C or lower. That is, the target course GPA is 2.75. An examination score in the upper quartile is roughly an A grade; a score in the second quartile is roughly a B grade; a score in the third quartile is roughly a C+ grade; and a score in the lowest quartile is a C or lower grade.

Your final grade will be determined using your final examination grade and your total point score. Your total point score is based on three components. One is your computer project component, which is the sum of the scores received for the two projects. The second is the examination component, *E*, which will be calculated by the sum of your examination and quiz scores:

$$E = E_1 + E_2 + E_3 + E_F + Q.$$

Each in-class examination will have about 6 questions, each worth about 50 points. Each midterm examination will be worth approximately 300 points. The final examination will have about 14 problems and will be worth approximately 600 points.

The third component *P*, class participation, is based upon your usage of the class Blackboard. The class participation component is the sum of four components:

$$P = P_1 + P_2 + P_3 + P_{Cumulative},$$

where P_i is your class participation score for the *i*th examination and $P_{Cumulative}$ is your class participation for the whole semester. The relative weight of the participation scores are $0 \leq P_i \leq 60$ and $0 \leq P_{Cumulative} \leq 120$.

Your total point score is the sum of your examination scores, project scores, and class participation score:

$$TP = E_1 + E_2 + E_3 + E_F + Q + CP_1 + CP_2 + P.$$

Total point boundaries for an *A*, *B*, *C+*, *C*, *D*, and *F* grades will be set. Similarly, final examination boundaries for *A*, *B*, *C+*, *C*, *D*, and *F* grades will be set. Each student will have a total point grade and a final examination grade. A student with satisfactory computer projects will have a course grade that is the higher of the final examination grade and the total point grade. The grade of a student without satisfactory computer projects scores will be the total point grade.

That is, if you have at least minimally satisfactory work on the computer reports, a strong performance on the final will be a major factor in the final grading decision. A student with at least minimally satisfactory computer performance who gets an *A* on the final gets an *A* on the course; a student with satisfactory computer performance who gets a *B* in the final gets at least a *B* in the course; and so on.

AMS 315
Data Analysis
Spring Semester, 2023
Lecture 2

Human Subjects Research

- Must have institutional approval (IRB approval, Institutional Review Board)
- Participants must give informed consent; the researchers must explain to the participant the goals of the research and the risks.
- Participants may withdraw from the study at any time.

Five Most Important Contributions of Statistics

1. Randomized Experiment (Clinical Trial)
2. Genetic Statistics
3. Methods to Analyze Observational Studies
4. Quality Control
5. Opinion Surveys Using Random Samples

1. Randomized Experiment:

In a randomized experiment, a difference in response between the group given an experimental treatment and the group given a control treatment is ***CAUSED*** by the experimental treatment or is a chance event.

What is randomization of assignment of participant to treatment group?

The two groups are roughly balanced on all independent variables, whether important or not.

❖ Fuller discussion in Chapters 6, 8, and 9.

2. Genetic Statistics

Not covered in this course.

❖ Example 1: statistics identified the gene in the mechanism for familial hypercholesterolemia, which led to the development of the statin class of medicines. See Tobert review paper and Brown and Goldstein Nobel Prize speech.

3. Methods to Analyze Observational Studies

Chapters 10, 11, and 12

In an observational study, correlation shows association, not necessarily causation.

Example of Observational Study

- Reported association of prior mononucleosis infection and subsequent contraction of multiple sclerosis.

4. Quality Control

Chapters 5 and 6

5. Opinion Surveys Using Random Samples

Chapter 10.

Please follow a news source (New York Times, Wall Street Journal, Newsday, CNN, …) regularly to identify stories and issues that are relevant to this course.

Chapter One, Statistics and the Scientific Method:

There will be no examination questions directly about this chapter. You should read and review it. It contains fundamental background material. I recommend that you focus on the schematic for the scientific method (Figure 1.1): 1. formulate research goal; 2. plan the study (specifically identify variables); 3. collect data; 4. inferences; 5. conclusions (decisions); 6. formulate new goals and return to step 2. You should learn the definition of populations and sample.

Chapter Two, Using Surveys and Experimental Studies to Gather Data:

There will be no examinations questions directly about this material. It contains important definitions. The most important point is the distinction between a randomized experiment and an observational study. You should know the distinctions between comparative and descriptive studies, between prospective and retrospective studies, and case-control studies. You should also know about confounding variables and have examples at hand. Other terms to master: target population, samples population, sample, observation unit, sampling unit, sampling frame, simple random sample, and systematic sample. You should know that two problems associated with random sampling surveys are non-response and measurement problems.

Chapter Three, Data Description:

You may use a calculator but not a computer in examinations. The only way to perform the computational assignments is to use a computer and data analysis package. I recommend R, SAS, SPSS, or Minitab. For those of you who are going to seek a position as a computer oriented quantitative analyst, expertise in SAS should be helpful as a qualification for most of these positions.

Your text does not discuss scales of measurement. Wikipedia and your favorite search engine are effective tools to get a fuller definition of scales of measurement. The ***nominal scale*** of measurement is the simplest. The value of a nominal scale variable is actually a verbal characterization. For example, ethnicity or hair color are nominal variables. Traditionally, numbers are used in databases rather than the words. Pie charts, bar charts, and contingency tables are tools used to deal with a nominal scale variable. The ***ordinal scale*** of measurement has ranked values. For example, course grade with values A, B, …, is an example of an ordinal scale variable. Psychology statistics texts often recommend nonparametric statistical procedures. The common statistical procedures that you studied in AMS 310 can be

viewed as an approximation to a permutation procedure. The permutation procedure is considered a gold standard statistical procedure in scientific research. In an ***interval scale***, differences have meaning. For example, the net balance in a bank account (positive for asset and negative for liability) is an example of an interval scale variable. The techniques that you studied in AMS 310 (such as t-tests) can be applied to interval scale variables. A ***ratio scale*** variable is one in which ratios have meaning. For example, the weight of a study participant is a ratio scale variable. Often, a monotonic transformation of a ration scale variable is helpful.

I expect you to know about histograms and that they can be classified as: 1. unimodal, bimodal, or multimodal; 2. symmetric or skewed. I expect you to know the definitions of mean, mode, and median (see page 85). I also expect you to know the definition of standard deviation (which is the square root of the variance), range, percentiles, and interquartile range.

Chapter 4, Probability and Probability Distributions

.A key result that is of importance here is the theorem that the probability of the union of two events, *A* and *B*, is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. This leads immediately to Boole's inequality, which states that $P(A \cup B) \leq P(A) + P(B)$ for any two events *A* and *B*. This will be used extensively in lecture when we discuss multiple comparisons and in your second computing project.

Conditional Probability and Bayes' Theorem

Please see Part B of Lecture 2.