Introduction to Statistical Methods

Fall 2019 Session 2

Lectures: Mon-Wed 2:45 pm - 5:15 pm Seminar 1A 1079
Lab: Tue 4:15 pm-5:30 pm Seminar 1A 1079
Academic credit: 4 DKU credits
Course format: Lectures, Labs

Instructor's Information

My contact information:
Andrew MacDonald
andrew.macdonald@dukekunshan.edu.cn
Office: IB 1025
Office Hours: Mon, Wed 1:00 pm – 2:30 pm; Tue, Thur 2:30 pm – 4:00 pm or by appointment

Head Tutor:
Yunqiu (Maisie) Zhang
yunqiu.zhang@dukekunshan.edu.cn
Tutoring hours: ARC Schedule

About Me

My interest in statistics comes from a passion in answering difficult questions in Chinese politics using data in innovative ways. I invite you to view examples of my work at: www.andrewmacdonald.org

What is this course about?

How can we use data to shed light on age-old and new human problems such as pollution, discrimination, and economic growth? How can we be "sure" that the evidence we have points us in the right direction? How meaningful are our findings? Do our results suggest the relationships we find between factors such smoking and cancer are meaningful or meaningless? How would we know? How should one properly display and explain your statistical results to these important issues?

This class introduces you to the tools and concepts that begin to tackle these questions. We will cover topics such as data summaries, sampling, data analysis, production of graphical displays, and regression techniques. The goal at the end of the course is that you will be able to conduct basic data manipulation, know how to properly summarize and display data, and make basic statistical inferences using real datasets.

The emphasis in the course will not be on learning mathematical formulas related to statistics but rather to develop an intuitive understanding of statistical inference and measures of uncertainty. For those interested in a more mathematical treatment of statistics, the math department offers a course on the mathematical foundations of statistics that I also encourage you to attend, should you have the interest (calculus required).

A third set of goals for the course is that you will also be able to read more fluently research literature that employs statistics. During this course I will reference, in class, a number of historically important academic articles and we will analyze the data from those articles. Doing so should help you understand how data is used (and misused) to construct social science arguments.

What background knowledge do I need before taking this course?

None

What will I learn in this course?

Upon completing the course, you will develop the following abilities:

- Intuitively interpret statistics in course materials and in the larger world
- Become a statistics results producer in addition to a statistics consumer
- Assess when and how to use statistics to answer specific questions in the social sciences
- Analyze how previously learned problems can be answered with statistical methods
- Apply statistical methods to future social science coursework and capstone project
- Judge how appropriately statistics are used in everyday life when reading the news, business reports, and other real-world applications

In support of this you will be able to:

- Understand and interpret basic statistical properties of data (confidence intervals, t-tests, etc.)
- Identify when various statistical tests are appropriate given a specific dataset
- Formulate testable hypotheses in the data and learn how to execute those tests
- Interpret statistical results to understand both significance of the results and their substantive impact
- Illustrate statistical results with appropriate and clear graphical displays that provide meaning to the reader
- Evaluate critically other, published, statistical work with the skills and techniques learned in class
- Propose an independent research project that integrates statistical methods with their research interest for their capstone project

What will I do in this course?

In general, class will generally take the following structure:

Lectures will generally consist of:

- Discussion of pre-class quiz results & homework
- Lecturing on course concepts for a time
- Demonstrative activities during class putting concepts into practice

A brief break (10 min) will be offered after the first 1:10 min of class

Labs:

- After some discussion of the purpose of the lab, instructions will be handed out and pairs/groups assigned according to the needs of the activity.
- After a period of working on the labs together, there will be another 5-10 minutes of discussion at the end of the lab session on main takeaways.

How can I prepare for the class sessions to be successful?

The main goal of this class is to learn by doing. After class, if you are still confused on how certain statistical properties work, I strongly encourage you to download the dataset we discussed in class (all of them will be available on Sakai) and manipulate the data yourself. Practice exploring the data and graphing the results as much as possible.

To the extent possible, for the substantive readings I will try to make available for downloading datasets on which the articles are based. Download these datasets and interact with them as much as you can!

What required texts, materials, and equipment will I need?

Technical Preparation

For class, you will need to install RStudio on your own personal computer (we will discuss how to do that in class).

You will also need an iPhone, Android phone or other device capable of running the class clicker software.

Textbooks

The primary required textbook for the class is *Intro Stats* by De Veaux, Velleman, and Bock (5th Edition) – notated IS in the schedule section. You should have an electronic copy already sent to your book account. If, for some reason, you have an older edition of the text, that is probably fine but all page numbers will be given for the 5th Edition.

There is also a supplemental text that may help many of you for whom a more intuitive explanation of statistical concepts would be valuable. That text is *Naked Statistics: Stripping the Dread from Data* by Charles Wheelan – notated as NS in the schedule section. His explanations of statistical concepts do not have any mathematical formulas and are written in a funny and engaging way. Even if you feel the main textbook adequately explains a topic, you still might find value in reading the corresponding chapter in this text.

We will also consider some supplementary readings that I will post on Sakai from time to time. While these are in addition to the textbook, they are not optional and will be an important part of class discussion and homework.

In addition, the library has a number of good statistical textbooks in both English and Chinese should you want to consult additional sources. The librarian will be happy to assist you in locating those texts.

How will my grade be determined?

Assessment: Grades in the class will be determined by:

- *Participation (10% of grade):* This class is designed to be highly interactive. We will spend significant amounts of time in pair, group, and class discussions. You do not need to be a frequent speaker to get a good grade but you do need to meaningfully participate in all discussions, particularly during the lab sessions.

- *Pre-Class Quiz (10% of grade);* Before each class there will be a series of short online questions or a puzzle to prepare you for the day's material. Completing them before class is mandatory and we will discuss the answers to the quizzes in class. I will drop your lowest quiz grade.
- *Homework (40% of grade):* During each week we will work with some statistical material in class. Each week, based on that material, there will be homework that will require you to further refine what we have worked on in class and prepare a brief analysis and graphical presentation of your findings. Homeworks will be split into two parts, problem sets (40% of the grade) and Labs (60% of the grade). Problem sets are due on Thursdays at 11:59 pm, Labs are due on Saturday at 11:59 pm.
- *Midterm Project (15% of grade):* The exam will consist of questions that cover the core topics of Unit 1 of the class and to use these core concepts to solve specific real world examples.
- *Final Project (25% of grade):* You should go out into the world at DKU and in Kunshan and find data relevant to you. It could be data on anything you find relevant to your own life, whether it relates to your life at DKU, something you gather in Kunshan, or information about a hobby or passion of yours. We will meet individually and consider what a good research question will be and then you should write a research report, using appropriate statistical techniques, graphs, and visual displays, to answer the question.

Grades

The grading scale is as follows:

A+	4.0	100- 98
A	4.0	97 - 93
A-	3.7	92-90
B+	3.3	89 - 87
B	3.0	86 - 83
B-	2.7	82 - 80
C+	2.3	79 - 77
C	2.0	76 - 73
C-	1.7	72 - 70
D+	1.3	69 - 67
D	1.0	66 - 63
D-	1.0	62 - 60
F	0	59 and below

In addition to specific grades I will provide as much feedback and information about your progress on the road to mastery of statistics as I possibly can. I care much more that you can successfully master the course objectives and are able to take what you learn in this class and use it in your future life than whatever grade I assign you (though the two are, of course, related).

What are the course policies?

Laptop/Tablet Policy:

There has been significant research showing that unfocused computer use in class hinders learning (I know you all like to spend time messaging on WeChat and shopping on Taobao) not only for you but for everyone else in class that can see your screen. However, I recognize some people do have specific learning situations that require a computer. If you have a specific situation that requires a computer use in class please either contact me directly with your needs or speak to the relevant advising department and ask them to contact me.

Therefore, unless you or someone from the advising center have contacted me in advance, during lecture time you may not use your laptop or tablet during class.

During activity/lab time, you will need make use of your laptop or tablet to class to complete class assignments; usually we will break into teams or pairs and students will take turns being the "computer operator" while others think and advise.

Late Homework Policy:

Homework forms an essential part of our in-class discussion and so not having done the homework will make it difficult for you to participate. You will need to turn in your homework (and submit your online pre-class quiz) before the start of class. Late homework will be penalized 50% not because I want to be a jerk but because the homework is such an essential part of our class environment. I will only accept late homeworks up to 48 hours after the due date.

I will drop your lowest quiz grade but otherwise I will not grant any other exceptions to the lateness policy unless you have a documented serious personal or medical emergency.

Collaboration Policies:

You are expected to strictly adhere to the Duke Kunshan University Community Standard in all of your work and participation, and violations will be enforced. More details can be found here: https://dukekunshan.edu.cn/en/advising/academic-integrity.

All work must be done exclusively by the individual to whom it has been assigned. You should assume that collaboration on assignments, the use of previously-assigned homework, quizzes and answer keys, outside sources or outside aids (both written and electronic) are not allowed unless explicitly noted in the assignment guidelines. All cases of suspected cheating will be referred for adjudication to the Dean's Office. Any violation for which a student is found responsible is considered grounds for failure in the course.

It may sound cliché to say, but if you cheat and borrow other's code or answers you are only cheating yourself; you will not learn how to do statistics and doing so will mean you will do worse on the midterm and the final anyway. Cheating is ultimately self-defeating so for both of our benefit, please, don't do it. If you are having trouble completing the assignment and feel tempted to cheat, please contact me directly instead with the difficulties you are having.

What campus resources can help me during this course?

If you are having problems with the technology, IT may be able to help you if you are unable to install R Studio (though they will not be able to help you with using R Studio, see me or the tutors for help with that), email them at: service-desk@dukekunshan.edu.cn

If you are having problems with the course material and feel you could use additional tutoring, please contact the Academic Resource Center: dku-arc@dukekunshan.edu.cn

What is the expected course schedule?

Overall, we will try to fully investigate one complete concept per class period so it is essential that you do not miss class!

Simple Data Analysis

Basic Descriptions in Statistics (October 28)
Readings: IS Chapter 1, 2.1 and 2.2, and 3
Topics covered:
- What are data and variables?
- How to display quantitative and qualitative variables
- Contingency tables

Lab 1: Introduction to R (October 29)

Distributions

Characteristics of Distributions (October 30)
Readings: IS Chapter 2.3-5 and 4, NS Chapter 2
Topics covered:
- How to describe the shape, center, and spread of a distribution
- How to compare distributions
- Dealing with problem distributions (outliers, reexpression)

Problem Set 1 Due October 31

Homework 1 Due November 2

The Normal Distribution (November 4)
Readings: IS Chapter 5, NS Chapter 3
Topics covered:
- Standard deviation and standardizing values
- Normal models
- Normal percentiles

Lab 2: Using ggplot in R to Describe Distributions (November 5)

Relationships Between Variables

Association and Correlation (November 6)
Readings: IS Chapter 6
Topics covered:
- Scatterplots
- Correlations
- Does correlation imply causation?

Problem Set 2 Due November 7

Homework 2 Due November 9

Simple Linear Regression (November 11)
Readings: IS Chapter 7, NS Chapter 11
Topics covered:
- Line of best fit: least squares

- The linear model
- What are residuals
- Regression assumptions

Lab 3: Correlations and Regressions in R (November 12)

Regression Wisdom (November 13)
Readings: IS Chapter 8, NS Chapter 7
Topics covered:
- Beware extrapolation
- Outliers and leverage
- Lurking variables
- Straightening scatterplots

Problem Set 3 Due November 14

Homework 3 Due November 16

Multiple Regression (November 18)
Readings: IS Chapter 9
Topics covered:
- What is multiple regression?
- Interpreting multiple regression coefficients
- Partial regression plots
- Indicator variables

Lab 4: Multiple Regression in R (November 19)

Measuring Uncertainty

Confidence Intervals - Proportions (November 20)
Readings: IS Chapter 13, NS Chapters 8 and 10
Topics covered:
- What is a sampling distribution?
- When does the normal model apply?
- Constructing a confidence interval
- Interpreting a confidence interval

******* Midterm Project Due on November 23 at 11:59 pm *******

Confidence Intervals – Means (November 25)
Readings: IS Chapter 14
Topics covered:
- The Central Limit Theorem
- Confidence interval for means
- Interpreting a confidence interval
- Final thoughts on confidence intervals

Lab 5: Sampling Methods in R (November 26)
Readings: IS Chapter 10, NS Chapter 7

Hypothesis Testing (November 27)

Readings: IS Chapter 15, NS Chapter 9

Topics covered:

- What are hypotheses?
- P-values
- P-values and decisions – how to make a decision

Problem Set 4 Due November 28

Homework 4 Due November 30

Hypothesis Testing Wisdom (December 2)

Readings: IS Chapter 16, NS Chapter 9

Topics covered:

- Interpreting p-values
- Alpha and critical values
- Practical vs. statistical significance
- Type I and II errors
- Power of a test

Lab 6: Recoding Data in R (December 3)

Statistical Inference

Comparing Groups (December 4)

Readings: IS Chapter 17

Topics covered:

- Confidence intervals for comparing two samples
- Assumptions and conditions for two-sample hypothesis tests
- Two-sample *z* test
- Two-sample *t* test

Problem Set 5 Due December 5

Homework 5 Due December 7

Returning to Regression (December 9)

Readings: IS Chapter 20, NS Chapter 12

Topics covered:

- Regression inference and intuition
- The regression table
- Confidence and prediction intervals

Lab 7: Regression Confidence Intervals in R (December 10)

Regression in Practice (December 11)

Readings: *see Sakai*

Topics covered:

- How to read academic statistical results
- Locating the model
- Interpreting the test

- Understanding possible weaknesses of the model

Extra Credit Due December 15

*******Final Project Due on December 18 at 11:59 pm*******