intro_stats

Introduction to Statistics: School of Criminal Justice - RU-N

View the Project on GitHub f-edwards/intro_stats

Introduction to Statistics	27:202:542
Lecture: Monday, Wednesday 1:00 - 2:20	Room: CLJ 574
Instructor: Frank Edwards	frank.edwards@rutgers.edu
Office hours: Tuesday 9:30-12:30	Room: CLJ 579B

Quick links

Lecture slides

Homework assignments

Course description

This is the course syllabus for Introduction to Statistics, Fall 2024. It is a graduate-level introduction to conducting quantitative social science research, and is the first part of a two-semester sequence.

For computing and data analysis workflow, we will cover the foundations of statistical computing with a heavy emphasis on data visualization using the R programming language and tidyverse suite of packages. You will also learn how to write professional reports on statistical findings using the RMarkdown format for fusing code and plain text writing together.

For statistics, we will review core mathematical concepts in algebra, linear algebra, and calculus, then proceed to build foundations in core probability theory. From there, we will learn foundational principles and techniques in statistical inference and conclude the class with a detailed unit on linear regression.

Course goals

1. Become comfortable fundamentals of probability and statistics. By the end of the course, they should be able to interpret and use common statistical measures of central tendency and variability, and be able to describe and interpret random events using probability statements.

2. Learn how to describe and estimate relationships for continuous outcomes using linear regression.

3. Use command-line interfaces for interacting with a computer and its file structure.

4. Design and write basic data analysis programs using the R programming language.

5. Produce univariate and bivariate data visualizations using the ggplot2 library in R.

Books

- Open Intro to Statistics. 2019. https://www.openintro.org/book/os/
- Healy, Data Visualization. 2018. https://socviz.co/
- Alexander, Rohan. Telling Stories with Data. 2023. https://tellingstorieswithdata.com/

Communication

We will use Canvas for course discussion and communication. Email is my preferred mode of one-on-one communication.

Expectations

- Attendance is strongly recommended. We move fast, it'll be hard to keep up if you miss lecture.
- Bring a computer - we'll be writing code in class.
- Complete homework on time. Homework should take between 4-8 hours to complete. *Don't start them the day before they are due.* All students are granted one no-questions-asked extension on homework assignments. Please notify me if you are using it for the week.
- Be respectful and professional. Be mindful of the space you take up in the classroom.
- Collaborate with your colleagues. Social science is a team sport. I encourage you all to work together to complete assignments. However, you DO need to submit your own work. We will penalize work that is copy/pasted from other students or online sources.
- Document your code. Explain what your code does in lots of detail. It helps you and helps us to evaluate your work.
- Don't use AI tools. If you want to learn to become a data analyst, you must learn to code. AI tools will make you sloppy and will often produce wrong code. Plus they are burning the planet.

Prerequisites

No prior statistics or programming experience is assumed. Statistics requires a basic grasp of algebra, geometry, matrix algebra, and calculus. We'll be reviewing the foundational math throughout the semester.

Software

All instruction will be conducted in the R statistical programming language. R is free and open-source, and can be downloaded here.

We will be using the RStudio integrated development environment. RStudio provides a powerful text editor and a range of very useful utilities.

In addition to writing code, it is a great tool for writing reports, papers, and slides using RMarkdown. This syllabus, most of my course materials, and most of my academic papers are based on Markdown.

You are required to submit assignments using RMarkdown.

Lastly, I recommend learning some form of version control to ensure your work is a) backed up, b) easily accessible to collaborators and c) reproducible. Git and GitHub are great and flexible tools for software development that have powerful applications for researchers. Here's a useful intro to GitHub for R users.

Assignments and grading

Course grading is based entirely on homework assignments. I grade assignments with a simple 2 point scale, and am generally a forgiving grader. If your work indicates a serious effort to complete the assignment, you can expect to receive full 2 points of credit. If you submit incomplete or sloppy work, you can expect 1 point of credit. Incomplete work will receive a zero.

All students who work hard and complete the assignments can expect to receive an A as their final grade.

Homeworks

I will assign homework each week. Assignments are Fridays by 7pm. Email your homework assignments (output and source code) to me and the course TA.

Problem sets provide you an opportunity to directly apply what we've learned to real-world data analysis and statistical problems. Don't wait until the last minute to get started. These homeworks should take you on average between 2 and 6 hours of work to complete. Space that work out and give yourself time to ask for help from your peers and your instructor.

Group work is strongly encouraged for homework. I recommend scheduling a time to meet with your classmates to work on the problem sets. Each week, I will open a channel on the course Slack page for you to ask coding and technical questions. Quantitative research is a team sport, but I still do expect you to write your own code and interpretation. Don't just copy/paste from your peers, the internet, or a chatbot.

Homework should be submited via canbas with attached code and code output. Generally, this means I want to see two files: your script and your rendered output.

Life happens. All students are granted two free extensions on homework, no questions asked. Just email prior to the due date and let me know you'll be taking an extension and when I should expect your submission.

Course schedule, topics, and readings

Week 1

Reading: Alexander Ch1; OI 1

- 9/3: Lab - Introduction to the course and intro to R
  - Review the syllabus and course format
  - Installation and FOSS principles
  - Familiarizing yourself with the RStudio workspace
  - File formats: R Scripts, RMarkdown
  - R: basic math operations, creating atomic objects

- HW 1: Due 9/10 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw1.Rmd)

Week 2

Reading: OI 2; Healy 1 and 2

- 9/8: Lecture - Math review

    - Algebra: order of operations, exponents and logarithms, polynomials, plotting functions on a cartesian plane
    - Linear algebra: vectors and matrices, scalar operations
    - Functions and limits
    - The basics of derivitaves and integrals

- 9/9: Lab - Working with vectors in R
    - Vectors, matrices, and data.frames
    - Indexing
    - Vector operations
    - Commonly used functions for vectors

- HW 2: Due 9/17 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw2.Rmd)

Week 3

Reading: OI 3

- 9/15: Lecture - Probability
    - Making probability statements, computing probabilities (marginal, conditional, joint)
    - Basics of set theory

- 9/16: Lab - group_by and ggplot basics
    - Theory: grammar of graphics
    - Importing tabular data
    - Basic univariate visuals: densities, histograms, barplots
    - Introduction to plain text editing with markdown
    - Writing math with LaTeX

- HW 3: Due 9/24 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw3.Rmd)

Week 4

Reading: OI 4; Alexander 2, 3 (recommended)

- 9/22: Random variables and moments
  - The Normal distribution and Normal random variables
  - Simulating Normal variables
  - Measures of central tendency: mean, median, mode
  - Measures of dispersion: variance, standard deviation, quantiles

- 9/23: Lab - Visualizing more than one variable, basics of tidy data
  - Tidy data principles
  - Manipulating data frames with mutate(), select(), and filter()
  - Scatterplots

- HW 4: Due 10/1 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw4.Rmd)

Week 5

Reading: Healy 3

- 9/29: Lecture - Types of variables and measures of association
  - Continuous, integer, binary, categorical, and ordinal measures
  - Correlation
  - Assessing bivariate distributions with crosstabs and scatterplots

- 9/30: Lab - Summary operations
  - Intermediate vector operations: group_by() and summarize()
  - Comparing means for causal inference
  - Visualization for comparing groups

- HW 5: Due 10/8 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw5.Rmd)

Week 6

Reading: Alexander 4

- 10/6: Lecture - Causal inference in experimental designs
  - The fundamental problem of causal inference
  - Thinking counterfactually
  - Experimental design and randomization
  - Comparing means and the sample average treatment effect

- 10/7: Lab - Summary operations
  - Intermediate vector operations: group_by() and summarize()
  - Comparing means for causal inference
  - Computing the SATE
  - Visualization for comparing groups

- HW 6: Due 10/15 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw6.Rmd)

Week 7

Reading: Healy 4, Alexander 7, 11 (6, 8 recommended)

- 10/13: Lecture - Observational data and bivariate association
  - Correlation
  - Kinds of observational data: cross-sectional, panel, longitudinal
  - Natural experiments and basic causal inference
  - Descriptive vs causal analysis

- 10/14: Lab - Data visualization with more than two variables

    - Using additional ggplot aesthetics: color and fill
    - Adding multiple geoms to visuals

- HW 7: Due 10/22 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw7.Rmd)

Week 8

Reading: OI 5, Alexander 9-10

- 10/20: Lecture - Generalizing from samples to populations
    - Measurement: theoretical constructs and operational measures
    - Basic principles of inference: parameters and statistics
    - Internal and external validity
    - Ethics of measurement and inference in social science, epistemic humility

10/21: Lab - working with more than one object - Harmonizing tables - Joins

- HW 8: Due 10/29 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw8.Rmd)

Week 9

Reading: OI 8.1,

- 10/27: Lecture - Introducing the linear regression model
    - Basic theory of linear regression
    - The structural component of the model
    - Theorizing and visualizing relationships as linear
    - Deterministic predictions

- 10/28: Lab - lm()
    - Introduction to lm() syntax

    - Estimating and interpreting the model

- HW 9: Due 11/5 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw9.Rmd)

Week 10

Reading: OI 8.2-8.4, Alexander 12

- 11/3: Lecture - The stochastic (random) component of the linear regression model
    - Anatomy of a linear regression model
    - What is error?
    - Ordinary Least Squares as an estimation approach
    - The meaning of stochastic error
    - Prediction with error

- 11/4: Lab - more lm()
    - OLS diagnostics
    - Using fit to compare models
    - predict()

- HW 10: Due 11/21 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw10.Rmd)

Week 11

- 11/12 and 11/14 No class, ASC meetings

Week 12

Reading: OI 6, 7

- 11/17: Inference and regression

    - The law of large numbers
    - The central limit theorem
    - Standard errors of parameters
    - The logic of frequentist hypothesis testing
    - t-tests for OLS parameters

- 11/18: Inference for OLS in R
    - summary.lm()
    - Interpretation and writing up OLS results

- HW 11: Due 11/26 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw11.Rmd)

Week 13

Reading: Alexander 12

- 11/24: Confidence intervals, prediction with error
    - The logic of confidence intervals
    - Interpretation of confidence intervals (danger!)
    - Using regression models for expected values with error
    - Using regression models for prediction with error

- 11/25: Lab - inference and uncertainty
    - Using predict() for uncertainty
    - Expected value intervals, prediction intervals
    - The bootstrap

- HW 12: Due 12/3 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw12.Rmd)

Week 14

Reading: OI 9.1, Healy 6

- 12/1: Multiple regression
  - Confounding
  - Visualizing causal relationships with Directed Acyclic Graphs
  - OLS with multiple additive predictors

- 12/2: Lab
  - lm() with multiple predictors
  - prediction with multiple predictors
  - Visualizing OLS predictions

- HW 13: Due 12/10 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw13.Rmd)

Week 15

Reading: OI 9.2-9.4, Healy 6

- 12/8: Multiple regression part 2
  - Interactions
  - Regression as a tool for modeling the data generating process

- 12/9: Advanced lm()
  - Estimating interactions
  - Visualizing interactions

- HW 14: Due 12/16 (https://github.com/f-edwards/intro_stats/blob/master/hw/hw14.Rmd)

This project is maintained by f-edwards