SMaC: Statistics, Math, and Computing

NYU Applied Statistics for Social Science Research

Eric Novik | eric.novik@nyu.edu

23 Jan 2025

1 Course description

This course aims to prepare students for the Applied Statistics for Social Science Research program at NYU. We will cover basic programming using the R language, including data manipulation and graphical displays; some key ideas from Calculus, including differentiation and integration, basic matrix algebra, including vector and matrix arithmetic; some core concepts in Probability, including random variables, discrete and continuous distributions, and expectations; and a few simple regression examples.

2 Where and when

The course will run from August 19 through August 30 from 9 am to 12 pm. We will meet on weekdays from 9 AM to 12 PM at 19 West 4th Street, Room 102, and for those who can't attend in person on Zoom. Students are strongly encouraged to attend in person, as this course will be highly interactive.

3 Course prerequisites

The students should be fluent in basic algebra and have seen exponential and logarithmic functions. Some programming experience would be helpful but not required. Students should be prepared to write many small R programs during the course.

4 Course materials and references

Students are expected to have access to R and RStudio. Use these instructions on downloading and installing the required software. In addition, you may find the following free resources helpful as you continue your studies.

4.1 Programming and Data Visualization

- Hands-On Programming with R, Grolemund (2014)
- R for Data Science, Wickham, Çetinkaya-Rundel, and Grolemund (2023)
- Data Visualization, A practical introduction, Healy (2018)

4.2 Calculus

- YouTube: Essence of Calculus, Sanderson (2018a)
- Calculus Made Easy, Thompson (1980)
- Calculus, Herman, Strang, and OpenStax (2016)

4.3 Probability

- YouTube: Probability Animations
- YouTube: Statistics 110 @ Harvard
- Intoduction to Probability, Blitzstein and Hwang (2019)
- Introduction to Probability Cheatsheet v2, Chen (2015)

4.4 Linear Algebra (for those who need it)

- YouTube: Essense of Linear Algebra, Sanderson (2018b)
- Introduction to Linear Algebra, Boyd and Vandenberghe (2018)
- Matrix Cookbook, Petersen and Pedersen (2012)

5 Tentative schedule

The following tentative schedule assumes ten 3-hour sessions, during which we plan to cover the following material.

- **(1) The Big Picture and Introduction to R**

  We will briefly discuss the motivation behind the topics in the class and give some examples of the types of regression problems you may encounter. We will use Shuttle O-Rings damage and John Snow's (the other one) Cholera datasets as motivating examples. We will introduce programming principles in R, including RStudio IDE, R objects, built-in functions, and how to write your own functions.

  - Hands-On Programming with R, *Part 1.*
  - R for Data Science, Introduction.
  - YouTube: Essence of Calculus, Videos 1, 2, 3.

- **(2) Plotting, Exponentials, Logs, and Derivatives**

  We will review linear, exponential, and logarithmic functions and develop some intuition using compound interest. We will explain and code a softmax function using the log-sum-exp trick. We will introduce a derivative geometrically and symbolically. We will cover basic rules for differentiating functions and use noisy measurements of motion in a straight line as a motivating example. We will also practice plotting lines and curves using base `plot` and `ggplot2`.

  - Hands-On Programming with R, *Part 2*
  - R for Data Science, Data visualisation
  - YouTube: Essence of Calculus, Videos 4, 5, 6.

- **(3) Reshaping Data, Loops, and Maps; Introduction to Integration**

  We will practice data wrangling using `dplyr` and `tidyr` packages. We will cover the intuition behind integration, approximating areas, the definite integral, and review the basic integration rules. We will learn how to do numerical and symbolic integration (in one dimension) using R.

  - Hands-On Programming with R, *Part 3.*
  - R for Data Science, Data transformations.
  - YouTube: Essence of Calculus, Videos 7, 8, 9.

- **(4) Introduction to Probability 1**

  We will review sample spaces, counting, and introduce the axiomatic definition of probability. We will discuss how to solve probability problems using simulations and will go through a few famous paradoxes. We will extensively use R's `sample` and `replicate` functions.

  - R for Data Science, Data Tidying
  - Intoduction to Probability, Chapter 1
  - YouTube: Essence of Calculus, Videos 10, 11, 12.

- **(5) Introduction to Probability 2**

  We will introduce conditional probability, the law of total probabilities, and independence and discuss a few famous paradoxes. We will write R simulations to check analytic solutions.

  - Intoduction to Probability, Chapter 2

- **(6) Introduction to Probability 3**

  We will introduce Random Variables, PDFs, and CDFs. We will cover Bernoulli, Binomial, Uniform, Normal, and Exponential RVs. We will discuss the concept of expectations. We will use R's distribution functions to simulate the realizations of RVs and compute their properties.

  - Intoduction to Probability, Chapter 3

- **(7) Introduction to Matrix Algebra**

  We will review the basics of Linear Algebra, including vector and matrix addition and multiplication. We will practice matrix and vector arithmetic in R. Time permitting, we will solve the over-determined system of linear equations for the motion in a straight-line problem from Lecture 2.

  - YouTube: Essence of Linear Algebra, Videos 1, 2, 3

- **(8) Exploratory Data Analysis (EDA)**

  We will take a break from math and turn our attention to EDA. We will start with an existing dataset and learn how to create various numerical summaries. We will then produce several plots focusing on the purpose of each graph and the comparison you are trying to make. We will discuss the grammar of graphics, what makes a good graph, and the concept of small multiples. We will also learn how to make simple interactive graphs with Shiny.

  - R for Data Science, Exploratory Data Analysis
  - Data Visualization, A practical introduction, Look at Data

- **(9) Analysis Workflow and Linear Regression**

  We will introduce an analysis workflow using a dataset containing ratings of red wines. We will use the `rstanarm` package but the same model can be done in with R's `lm()` function. We will attempt to determine what makes a good wine. After doing some basic EDA, we will fit several regressions and evaluate how well our model performs.

  - Intoduction to Probability, Chapter 3

- **(10) Review and Discussion**

  We will quickly review the material and answer questions. Before this session, each person will submit a one-page description of what they found particularly interesting or relevant in the class, and ask one or two questions pertaining to the material or to statistics in general. We will discuss these questions and comments as a group and try to answer them in real-time.

6 Assessment

- A short take-home test covering the material from the course

References

Blitzstein, Joseph K., and Jessica Hwang. 2019. *Introduction to Probability*. Second edition. Boca Raton: crc Press/Taylor & Francis Group.

Boyd, Stephen P., and Lieven Vandenberghe. 2018. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge, UK ; New York, NY: Cambridge University Press.

Grolemund, Garrett. 2014. *Hands-on Programming with r*. First edition. Sebastopol, CA: O'Reilly. https://rstudio-education.github.io/hopr/.

Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press.

Herman, Edwin, Gilbert Strang, and OpenStax. 2016. *Calculus Volume 1*. https://d3bxy9euw4e147.cloudfront.net/oscms-prodcms/media/documents/CalculusVolume1-OP.pdf.

Petersen, Kaare Brandt, and Michael Syskind Pedersen. 2012. "The Matrix Cookbook." https://www.freetechbooks.com/the-matrix-cookbook-t435.html.

Sanderson, Grant. 2018a. "Essence of Calculus," November. https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr.

———. 2018b. "Essence of Linear Algebra," November. https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab.

Thompson, Silvanus P. 1980. *Calculus Made Easy: Being a Very-Simplest Introduction to Those Beautiful Methods of Reckoning Which Are Generally Called by the Terrifying Names of the Differential Calculus and the Integral Calculus*. 3d ed. New York: St. Martin's Press.

Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund. 2023. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition. O'Reilly Media.