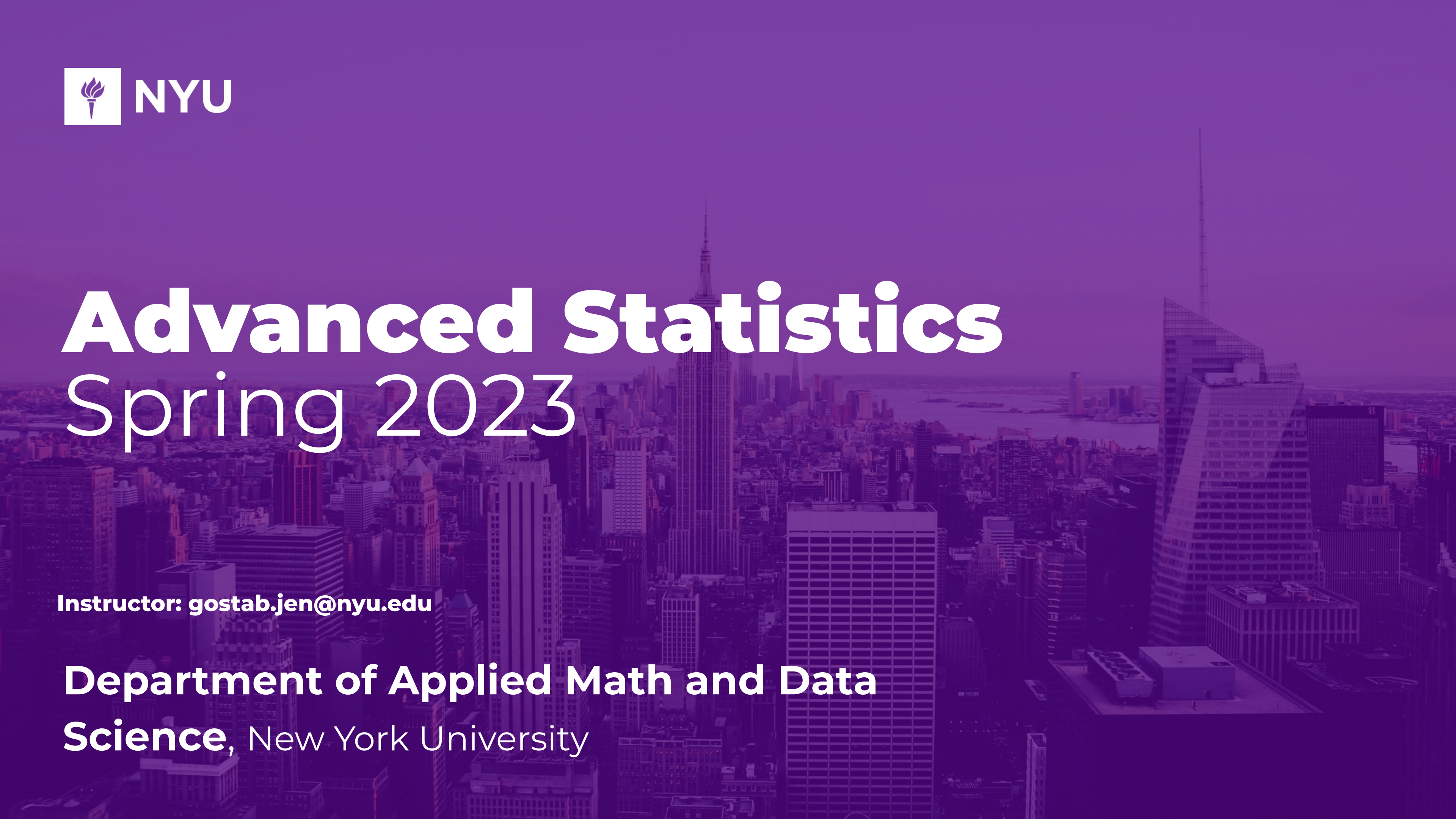


NYU

Advanced Statistics
Spring 2023

Instructor: gostab.jen@nyu.edu

Department of Applied Math and Data Science, New York University

Course Roadmap

- Logistics
- Course Components
- Assignment
- Week -2

Logistics

Time we meet: **9:30** am each Monday(003)/**Thursday(007)**

Office hour: 3-4 pm Friday or over Zoom

Assignment supporting session: in class

Time made busy: Tuesday and *Thursday*
‘S other time outside of class

You should not expect a response during this time

Three exams, 6 data assignments, research report

Course Components

Each class will be divided into two parts:

Part 1 : Lecture Review(**30** mins)

Part 2: Practice and Workshop(**40** mins)

NYU

About me

You can call me: **Gostab Jen**

I work mainly on : **Causal inference, statistical learning in Data Science**

Program: Pre-doctoral

Applied Mathematics and Engineering, Tandon school of Engineering and Courant institute of Mathematics

Courses I teach: Statistics for Data Analysts, Statistics for Behav-Science, Advanced Statistics, Introduction to machine learning

R

The goal of this course is to teach students to get the basic skills in processing data, we will cover **data simulation**, **statistical methods**, **model diagnostic**.

If possible, some students who need more practice will use R to **explore some important properties of our model.**

R Markdown will be discussed and mostly used in my recitation.

However, you can **just copy and paste** the codes in your doc/pdf documents.

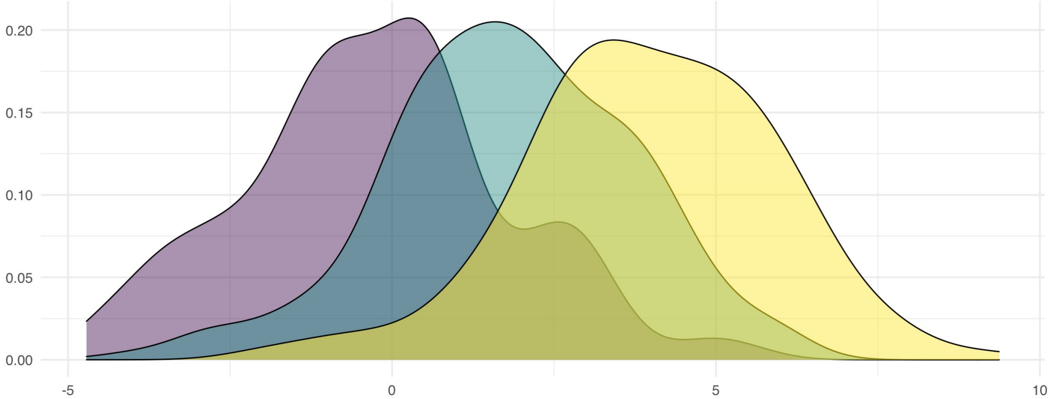
Also, choose R script

Week 2 Plans

Review: descriptive statistics

Start with R Markdown

Practice for R:



Mean

Perhaps the most important measure of location is the mean.

- The mean provides a measure of central location.
- The mean of a data set is the average of all the data values.
- The sample mean

$\bar{x}$ is the point estimator of the population mean μ.

Sample Mean $\bar{x}$

$$\bar{x} = \frac{\sum x_i}{n}$$

where:

$\sum x_i$ = sum of the values of the n observations

n = number of observations in the sample

Median (1 of 2)

- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Median (2 of 2)

For an odd number of observations:

7 observations

26	18	27	12	14	27	19

12	14	18	19	26	27	27

In ascending order

Median is the middle value

Median = 19

Mode (1 of 2)

- The mode of a data set is the value that occurs with greatest frequency.

- The greatest frequency can occur at two or more different values.

- If the data have exactly two modes, the data are bimodal.

- If the data have more than two modes, the data are multimodal.

Measures of Variability (1 of 2)

- It is often desirable to consider measures of variability (dispersion), as well as measures of location.

- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each but also the variability in delivery time for each.

Measures of Variability (2 of 2)

- **Range**
- Interquartile Range
- **Variance**
- **Standard Deviation**
- Coefficient of Variation

Simulation for artificial data

Process real data

Any insights?

Create data set using simulation.

Simulation distributions

Randomization

arm, dplyr, ggplot2

NYU

- This is part, we will use another larger data set that contains more than 200 days texting data and try applying the SAME SIMULATION PROCESS.

  The mostly possible changing date is around 130-150, relatively proportional to the sample size and consistent to our posterior.

What kind of method can reasonably be used to model the effects?