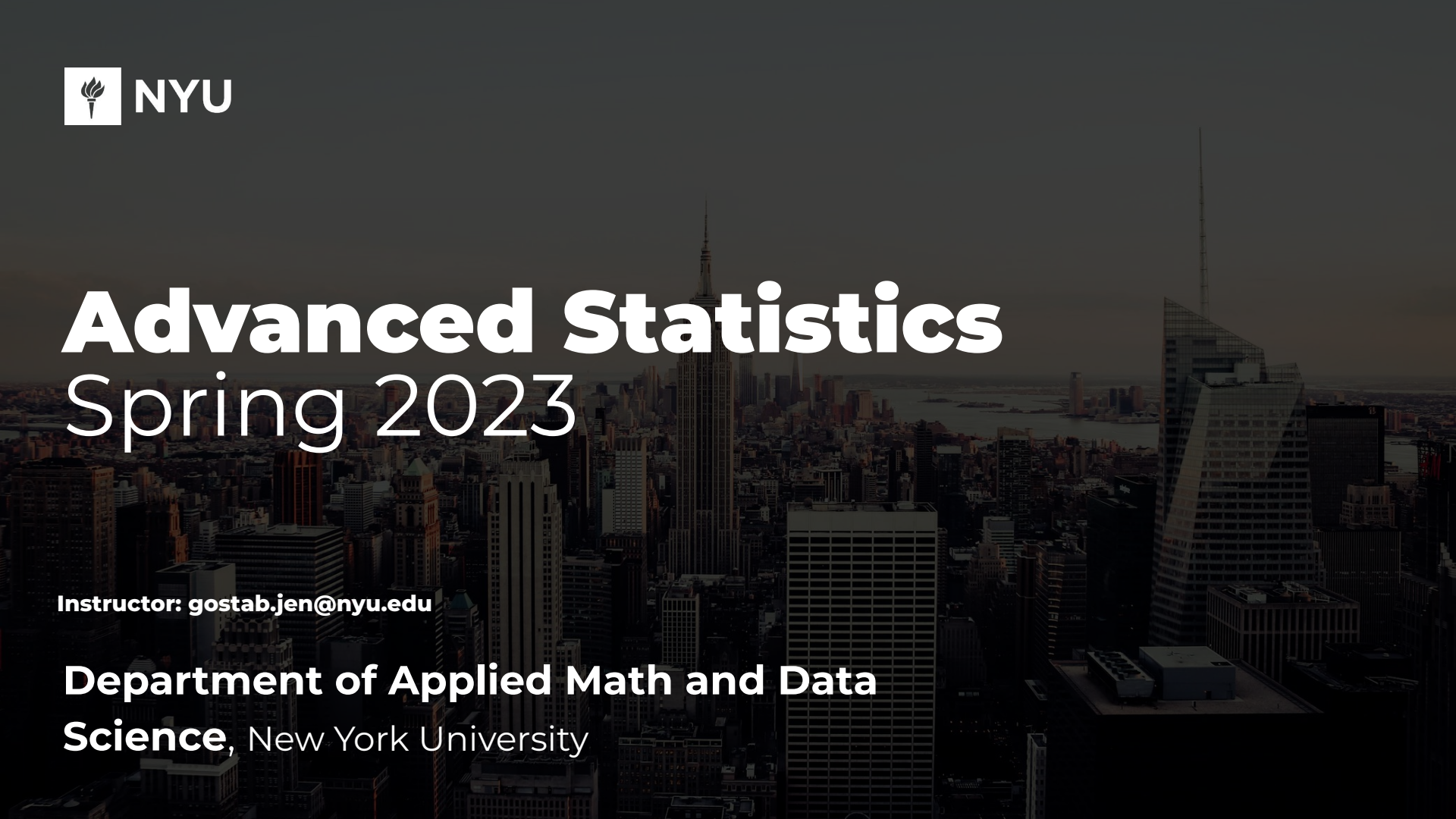


NYU

Advanced Statistics
Spring 2023

Instructor: gostab.jen@nyu.edu

Department of Applied Math and Data Science, New York University

Changes made for Monday's class

- Logistics
- Course Components
- Assignment
- Week -3

NYU

Changes made for Monday's class

From class 2, there are some changes in our course components: Section 003 will focuses more on Homework Supporting as we meet before the class.

The full demonstration of your homework will be possible

We will link our class to your upcoming lecture.

I need more time as I cannot repeat efforts of instructor Moty and I need to have two version slides.

NYU

Logistics

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

Office hour: over Zoom

Assignment supporting session: in class

Time made busy: Tuesday and *Thursday Other time outside of class*

You should not expect a response during this time

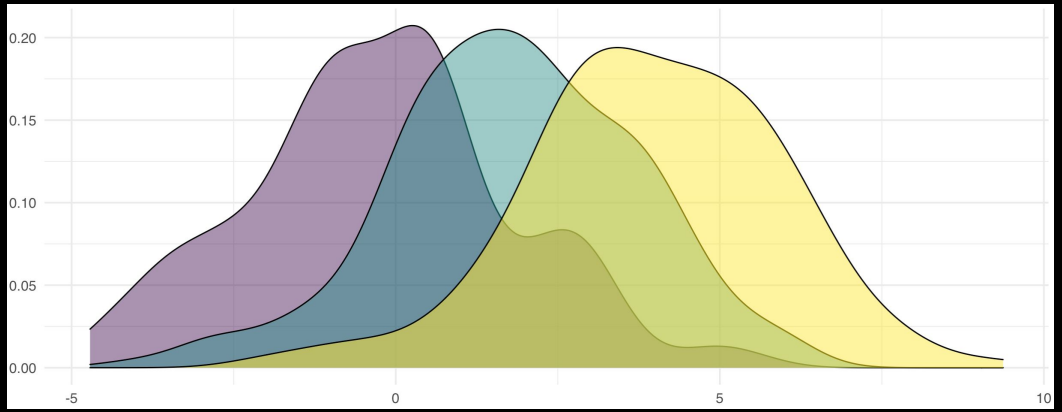
Grading: Three exams, 6 data assignments, research report

Week 3 Plans

Review: descriptive statistics

Start with R Markdown

→**Practice for R:**
estimator : mean, variance



NYU

Estimator-R exploration（¼)

Before our today's class:

R installed(search Cran; choose mirror)

R Studio installed(search Posit)

---

check your current environment: `getwd()`

Install some key packages: `install.packages('')`

Call the library : `library()`

You need `arm, dplyr,ggplot2` **minimum**

Estimator-R exploration（2/5）

Mean has some valuable properties and we will be exploring using **R**

Imagine you are addressing a data problem, you are given the following data structure and you need to simulate the data for your research/project:

Create a data set that represents students' score and meets the following requirements:

the score will not exceed 100, and no lower than 0

There is a cut off 60 for which students are considered pass the exam

If the student passed the exam, there should be letter grade P, otherwise F.

If you want to control of grade-level proportion, any insights? -Using sample()

Estimator-R exploration（3/5）

The basic in-built function mean(), summary(), dim() can tell us some facts about the variables

For example>

`mean(score)` will return a average score based on the vector you create

`summary()` can return a dimension-wise data information

Estimator-R exploration（4/5）

Move to R studio and get the data

We try to get some data from the population, so there must be some logics behind it

Now the student's score should be like a **belt-shaped** distribution(you learn at UA-10)

However, to be general, let's just randomly create the data without any patterns

There are three ways to do that:

`runif(); rep/replicate; sample()` **just play around it!**

Estimator-R exploration（5/5）

the score will not exceed 100, and no lower than 0 🔺

There is a cutoff 60 for which students are considered pass the exam

If the student passed the exam, there should be letter grade P, otherwise F.

If you want to control of grade-level proportion, any insights? -Using sample()

Additional Data

Now let's consider you get some data from an external institution whose records match your original data. For example, you got the cumulative GPA that corresponds to your other columns in the data frame. **YOU NEED TO COMBINE!**

`rbind()`

`cbind()`

`rbind.data.frame()`

`cbind.data.frame()`

Be careful when combining!!

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