Data science for everyone

Prof. Jones-Rooy & Prof. Policastro

March 9, 2020

7.1: Midterm Review

ANNOUNCEMENTS

- **Assignments**
  - Lab 4 due today, March 9, 8p
  - HW 2 due today, March 9, 8p
  - Project description available on JupyterHub shared folder
- **Midterm**
  - TA office hours review session: Tuesday, March 10, 2-5p CDS rooms 660 & 665
  - I will hold extra office hours (virtual) on Wednesday 12-2p in CDS 640
  - Exam during lecture time Wednesday, March 11 (60 minutes)
- No sections at all this week

GRADES & FEEDBACK SO FAR

- **Lab 0**: All grades released
- **Lab 1**: All grades released
- **Lab 2**: All grades released
- **Lab 3**: All grades will be released tonight
- **Lab 4**: All grades for *those who submit on time* will be released tonight
- **Homework 1**: All grades released
- **Homework 2**: All auto-grader grades for *those who submit on time* will be released tonight. We will **not** ask questions on the midterm related to questions on which you do not yet have feedback
- **All regrade requests except Homework 2 must be submitted by Friday, March 13**
- After spring break, your course grade overall will be on Classes, where **late penalties** and **academic integrity violations** for assignments will also be applied

MIDTERM ON WEDNESDAY

- Remote
- Online
- Instructions will follow, likely will be conducted on **Classes**
- Still during the same exact timeframe. Once you open the assignment you have 60 minutes to complete it, and **no more**.
- The course academic integrity policy still applies
  - We are building in safeguards against cheating & to catch cheating
  - Anyone caught cheating will **immediately fail the entire course**
- TA office hours are still in person on Tuesday

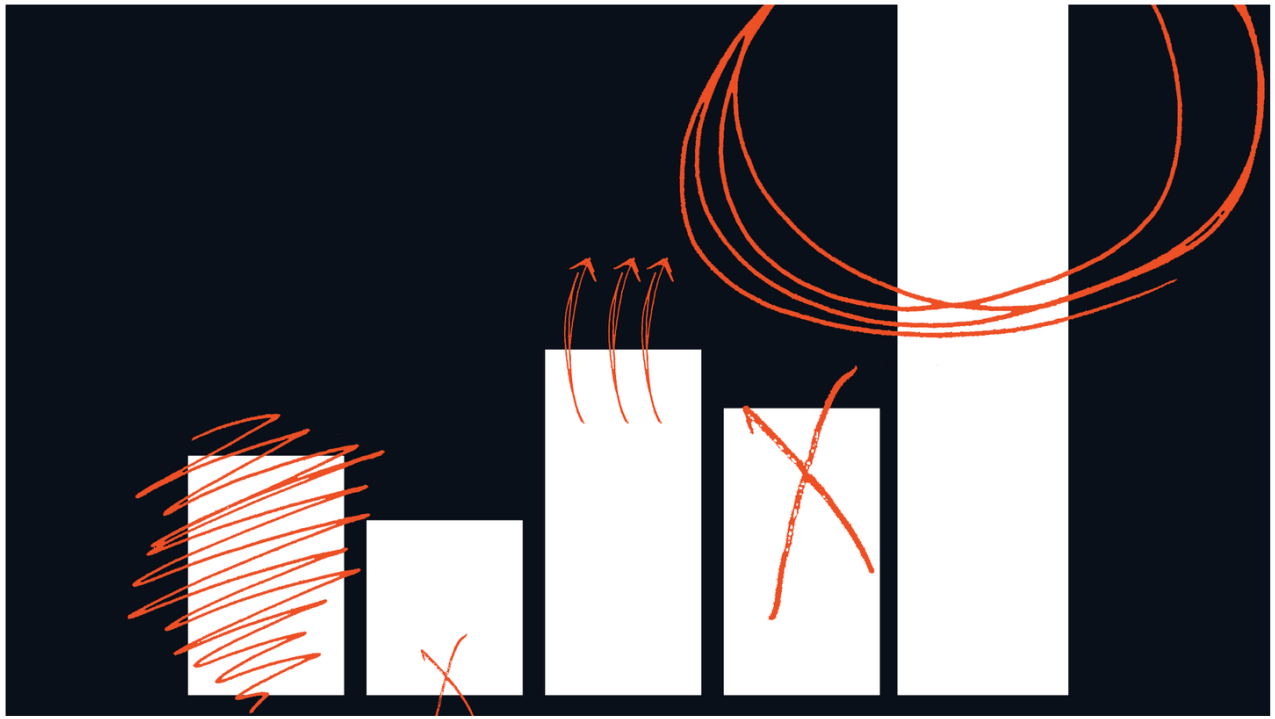
The Atlantic

TECHNOLOGY

The Official Coronavirus Numbers Are Wrong, and Everyone Knows It

Because the U.S. data on coronavirus infections are so deeply flawed, the quantification of the outbreak obscures more than it illuminates.

ALEXIS C. MADRIGAL MARCH 3, 2020



THE ATLANTIC

We know, irrefutably, one thing about the coronavirus in the United States: The

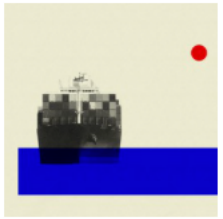
Read the full article here (still optional, but highly recommended).

We know, irrefutably, one thing about the coronavirus in the United States: The number of cases reported in every chart and table is far too low.

The data are untrustworthy because the processes we used to get them were flawed. The Centers for Disease Control and Prevention's testing procedures missed the bulk of the cases. They focused exclusively on travelers, rather than testing more broadly, because that seemed like the best way to catch cases entering the country.

The Coronavirus Is a Data Time Bomb

ALEXIS C. MADRIGAL



ONE MORE IN THE NEWS (RELEVANT TO MIDTERM REVIEW!)

Moderate Drinking Tied to Lower Levels of Alzheimer's Brain Protein

Compared with abstainers, those who drank up to 13 standard drinks a week had a 66 percent lower rate of beta amyloid deposits in their brains.

The study controlled for age, sex, education, socioeconomic status, body mass index, vascular health and many other factors.

Dr. Dong Young Lee, the senior author and a professor of psychiatry at Seoul National University College of Medicine, cautioned that this was an observational study that looked at people at one point in time, and does not prove cause and effect.

Still, he said, "In people without dementia and without alcohol abuse or dependency, moderate drinking appears to be helpful as far as brain health is concerned."

Read here

Outline

1. Structure of the exam
2. Example questions
3. Review of what we've covered so far

THE EXAM

- The exam is **one hour**
- **Three** parts:
  - **True/false** – 5-10
  - **Multiple choice** – 5-10
  - **Short answer** – 10-20
- Questions are similar in style to the homework questions
- No calculators or other devices allowed or required
- Like the course and assignments, the focus on the exam is on application of concepts to real life examples and data

SCOPE

- Everything from the first lecture to the present, **except**
  - We won't ask about the syllabus, uploading assignments, other course logistics (hopefully that's obvious)
  - We won't ask you about functions, ranges, or conditional (if/else) statements
  - We won't ask about anything on HW 2 on which you have not received feedback
- Everything else is **fair game**!
  - For **prioritization**: the more we talked about something in lecture, section, and on labs and homeworks, the more likely we are to ask about it
  - For **facts**: You do **not** need to memorize facts (e.g., the details of John Snow and the Broad Street Pump) but you do need to know how to talk about concepts related to it (e.g., what is a natural experiment?). If we reference an example from class, we will give you all the context you need. (Thus, it helps if you've been reading & paying attention, but you **do not** need to memorize historical or example details.)
  - For **code**: You do **not** need to memorize code syntax (e.g., whether a list is in square brackets, whether something has a ( ) at the end, etc.). But you do need to be **literate** in code (e.g., we may show you a line or a few lines of code and you need to be able to tell us what it's doing or what the output tells us. Example:

```
1  pd.set_option('display.max_columns', 50)
```

Outline

1. Structure of the exam
2. Example questions
3. Review of what we've covered so far

EXAMPLE TRUE/FALSE QUESTIONS

- **As we gather more data, we can prove a theory correct or incorrect.**
  - False
- **A sample may be bigger than the population**
  - False
- **The standard deviation of X is stated in the original units of X**
  - True
- **Endogeneity is when we suspect X causes Y, but Y causes Z**
  - False

EXAMPLE MCQ 1

- The measure of centrality that represents the most common value of a variable is the:

A. Mean

B. Median

C. Mode

D. Range

E. None of the above

EXAMPLE MCQ 2

- A recipe blogger is making the case for organic food. He argues that while eating organic food is better for a person's health, it's neither required for health nor a guarantee of good health. This person is making which argument?

A. Eating organic food is necessary for health

B. Eating organic food is sufficient for health

C. Eating organic food is both necessary and sufficient for health

D. Eating organic food is neither necessary nor sufficient for health but there is still a causal relationship

E. Eating organic food is neither necessary nor sufficient for health and there is no causal relationship

EXAMPLE MCQ 3

- A recent study by two economists explored the effect of education on what they called "deaths of despair", by which they meant deaths by suicide, alcoholism, and drug abuse ("SAD"). Defining deaths of despair in terms of "SAD" represents which stage of the measurement process?

A. Operationalization
B. Capturing
C. Conceptualization
D. Observation
E. None of the above
F. All of the above

EXAMPLE SHORT ANSWER QUESTION 1

- The previous question refers to a real, ongoing study by two economists at Princeton University. They argue that white people without four-year degrees are more likely to die from deaths of despair than those with four-year degrees. Without reading the actual article, do you expect a reasonable takeaway is that in order to prevent such deaths, more people should get college degrees? Why or why not?

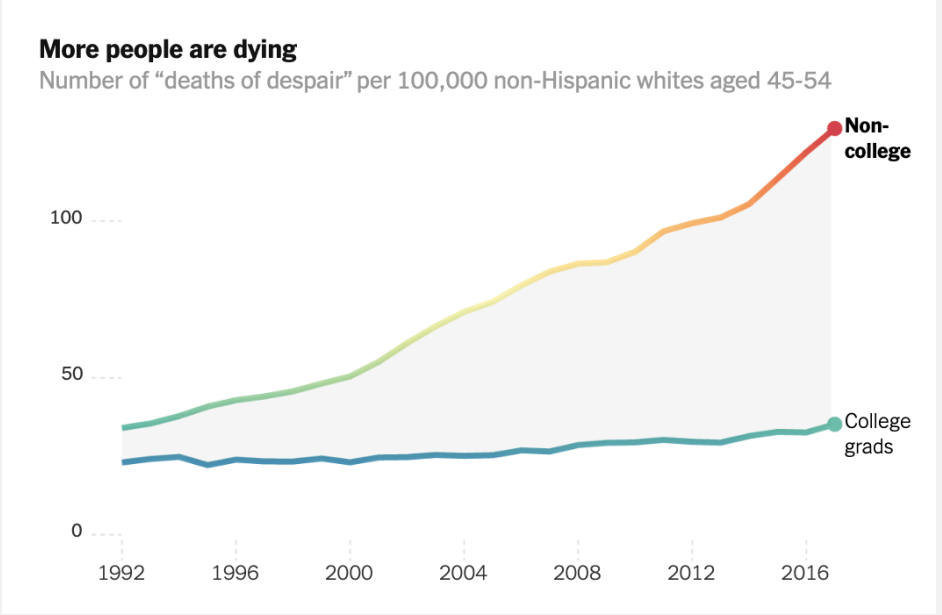
A: No, it could be a confounder of wealth/opportunity/privilege causes both education & lack of deaths of despair

(Read the story here, if you're interested)

EXAMPLE SHORT ANSWER QUESTION 2

- As part of their evidence, the researchers from the previous question provide some visualizations of their findings. What is the visualization type that they have chosen here? Is this a good choice for what they seem to be trying to convey here? Briefly explain why or why not.

A: Line graph; yes, because they are best for visualizing trends over time



SHORT ANSWER QUESTION 3

- A researcher is interested in what determines a child's height by age 5. He wants to run an experiment where the treatment is the child's daily caloric intake in the first year of life. What are some challenges with conducting an experimental study? Provide your answer in terms of one each of ethical, practical, and compliance reasons.

A: Ethical challenge = you can't starve (or overfeed) children
Practical = even if you could, how would you convince parents to participate
Compliance = even if the parents agreed to participate, odds are if the child is starving & crying the parent will sneak them some food!

SHORT ANSWER QUESTION 4

- The researcher argues that they can compare two populations of children within one country – one where there was a famine, and one where there was not – in order to draw conclusions about the effect of childhood nutrition on height. The researcher is proposing to do what kind of study?

  A: Natural experiment

- In order for this study to work, what needs to be true about these two populations?

  A: They need to be otherwise identical, at least with respect to variables of interest (e.g., nothing else can be different that would explain the famine or lower access to food)

Outline

1. Structure of the exam
2. Example questions
3. Review of what we've covered so far

SIX WEEKS OF DS4E

1. Introduction to data science & thinking like a scientist
2. Causality
3. Programming preliminaries
4. Working with data
5. Organizing data in Python
6. Data visualizations

WEEK 1: INTRO. TO DS & THINKING LIKE A SCIENTIST

At minimum you need to know:

- Data science is both a science and an art
- Data is not Truth and doesn't "say anything" on its own
- Science is about disproving not proving
- The scientific method
- Causal mechanisms

WEEK 2: CAUSALITY

At minimum you need to know:

- What is a theory? Why do we form them?
- Evaluating theories (what makes one "better" than another, and how do you know)
- Observable implications, falsifiable, useful, parsimonious
- Testing theories & testable hypotheses
- Repeating & extending tests (last step of the scientific method)
- Establishing association vs. causality (why the former is easier)
- Experiments (treatment and control, randomly assigned, double-blind design)
  - vs. observational studies
- Causal mechanisms
- Natural experiments
- Confounders
- Independent and dependent variables
- Selection on the dependent variable
- Survivor bias
- Principles of causality (covariation, temporal precedence, control for third variables)
- Causal conditions (necessary vs. sufficient)
- Deterministic vs. probabilistic causality
- Endogeneity (reciprocal causation)

WEEK 3: PROGRAMMING PRELIMINARIES

At minimum you need to know:

- Computer program
  - Set of instructions for the computer to solve a problem
  - Usually divided into sub-problems
- Python building blocks
  - Arithmetic (expressions in textbook), naming objects, built-in functions, modules & packages (what they do generally, what the ones we've used in this course so far do specifically)
- Data types
  - int, float, string, bool
  - Inspecting and changing types (reading the code)
- Sequences: lists and arrays and their different functionality
- What it means to turn the world into a data science problem

WEEK 4: WORKING WITH DATA

At minimum you need to know:

- Samples vs. population
- What makes a good sample (random, sufficiently large)
- Randomness = non-deterministic way, by chance, or at least orthogonal to what we are trying to understand
  - Randomness is important for both experimental and observational studies
- Measurement – what it is, why we do it, two steps (conceptualization & operationalization)
- Evaluating measurement and data
  - Random errors, systematic errors (selection bias), errors of validity, errors of exclusion (invisibility bias)

WEEK 5: ORGANIZING DATA IN PYTHON

At minimum you need to know:

- Tables: the fundamental unit of analysis in data science
  - Rows = observations, columns = variables
- Dummy variables
- Finding and evaluating a dataset, locating and understanding the codebook
- Be able to **read** and **understand** code for importing and inspecting/cleaning/organizing and making sense of a dataset, including:
  - Import dataset, view certain number of observations, import necessary packages/modules
  - Organize columns and rows
  - Generate and interpret **variable types** (discrete, continuous, categorical)
  - Generate and interpret descriptive statistics (measures of frequency, centrality, dispersion/variation, position)
  - Evaluate the **distribution** of a variable from descriptive statistics (later, visualizations)

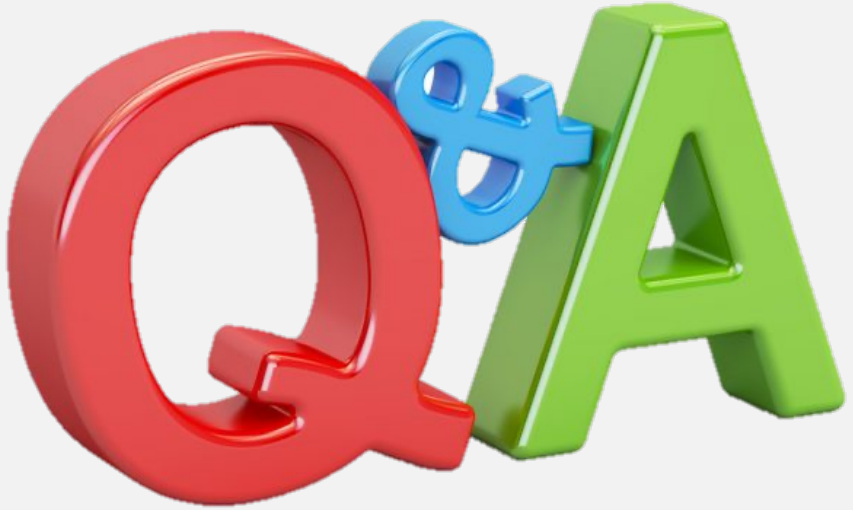
```
import pandas as pd
data = pd.read_csv('horses.csv')
```

WEEK 6: VISUALIZATIONS

At minimum you need to know:

- Why we create visualizations (understand and communicate)
  - And also it's a science and an art! (notice a theme in this course?)
- Key elements of a good visualization (e.g., title, labels)
- Types of data visualizations, why we use them, what types of data or variables they are typically best suited for
  - Histograms, bar charts, line graphs, scatter plots
- Using scatterplots to visually evaluate the association between two variables
- Using histograms to visually evaluate the distribution of a variable
- Be able to read and talk about output of code for the four main types of visualizations we are focusing on
- Be able to read, understand, and discuss all code in the 6.2 Example Code Jupyter notebook (on Classes and JupyterHub, as ever)

ANY ADDITIONAL QUESTIONS?



Outline

1. Structure of the exam
2. Example questions
3. Review of what we've covered so far