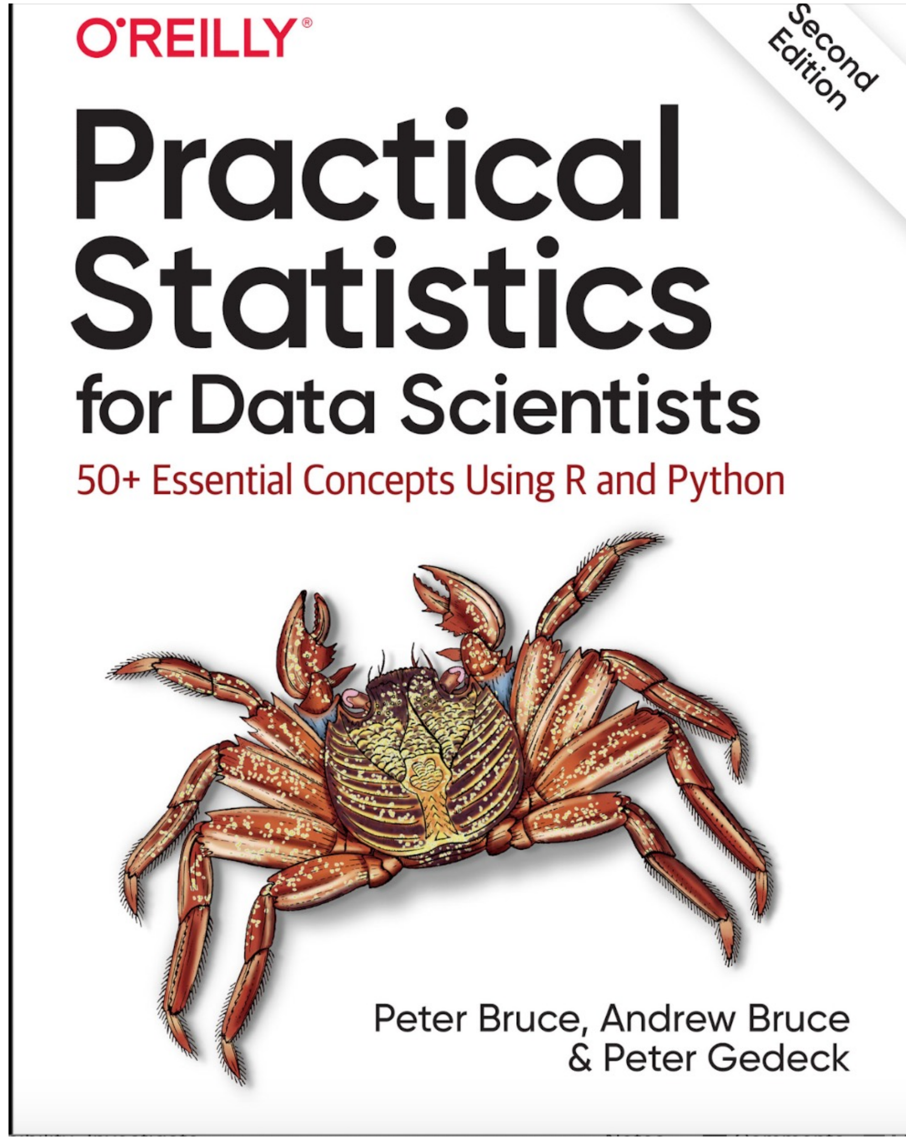
CSCI 692: LECTURE 3 DATA FRAMES, OUTLIERS, ERROR AND BIAS

Professor David Harrison

TODAY

- Remainder of Chapter 1
- Data Frames
- Review:
  - Distributions
  - Central Tendency
  - Dispersion



HOMEWORK 1

Due tonight!

11 pm.

Focuses on

- setting up accounts,
- using github and Databricks
- Notebooks.

Submission:

- Submit archived Databricks Notebook to Blackboard.
- NOTE: Submission only needs to be the notebook. No README is necessary.

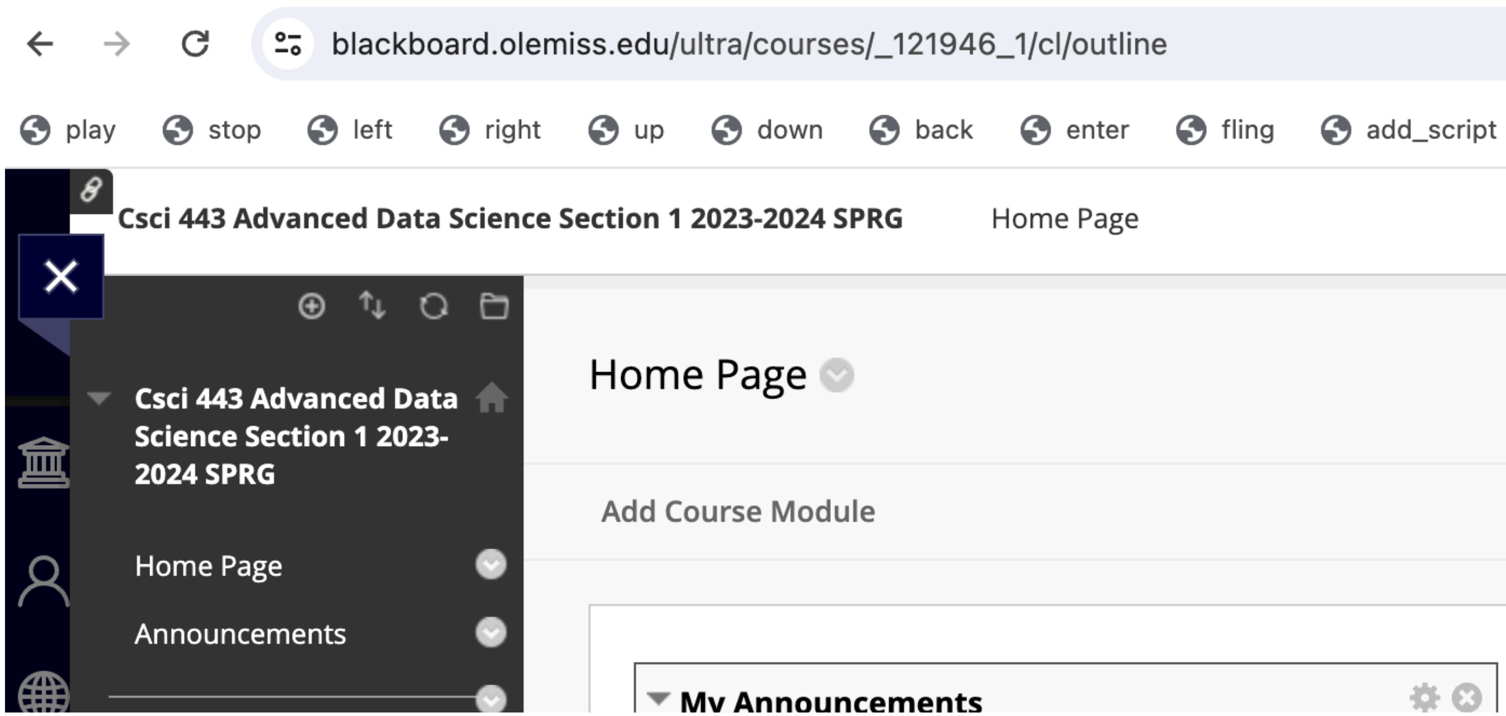
OFFICE HOURS

Due to scheduling conflict, office hours updated

Tuesday	4:00-5:00 PM
Wednesday	12:30-2:30 PM

BLACKBOARD

All lecture slides, homeworks, and solutions will appear on blackboard.

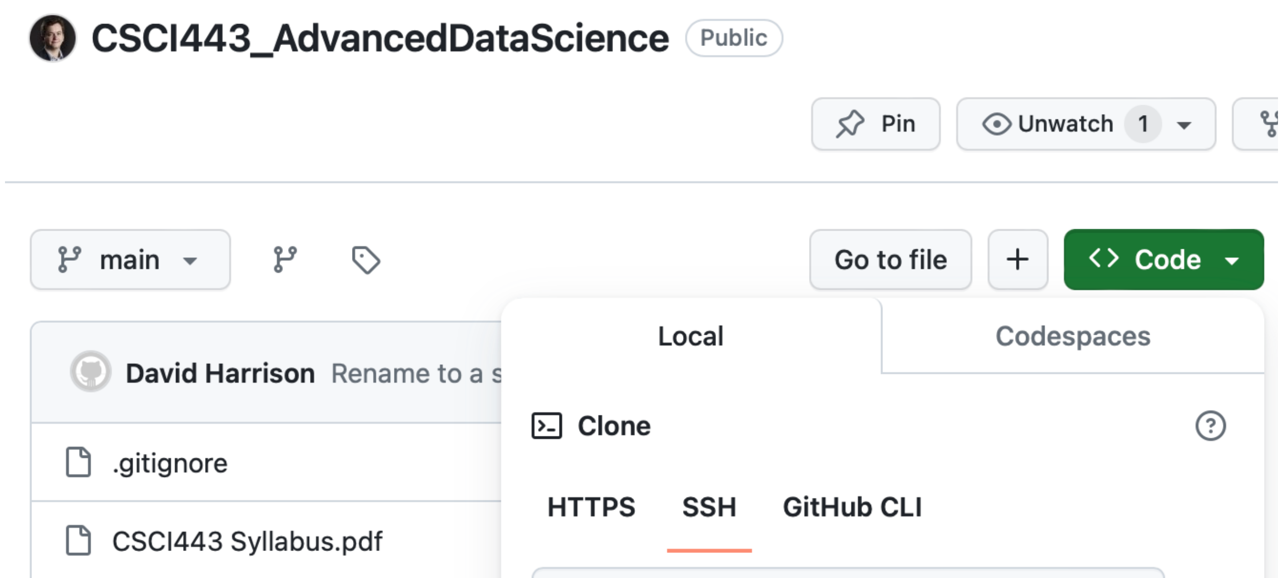


GITHUB

Lecture slides and examples have been committed to GitHub for lectures 1 and 2.

The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience



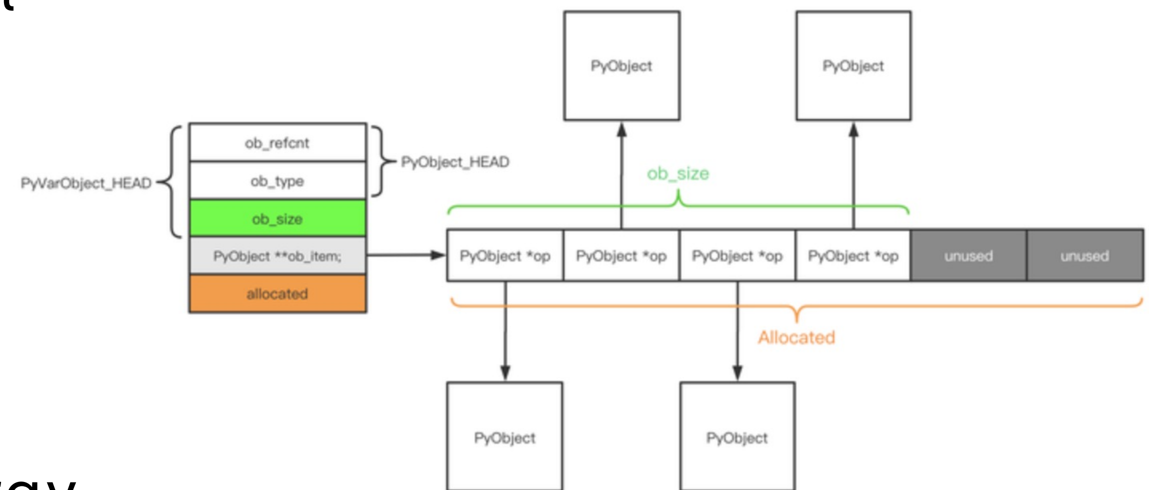
WHY NUMPY?

NumPy provides

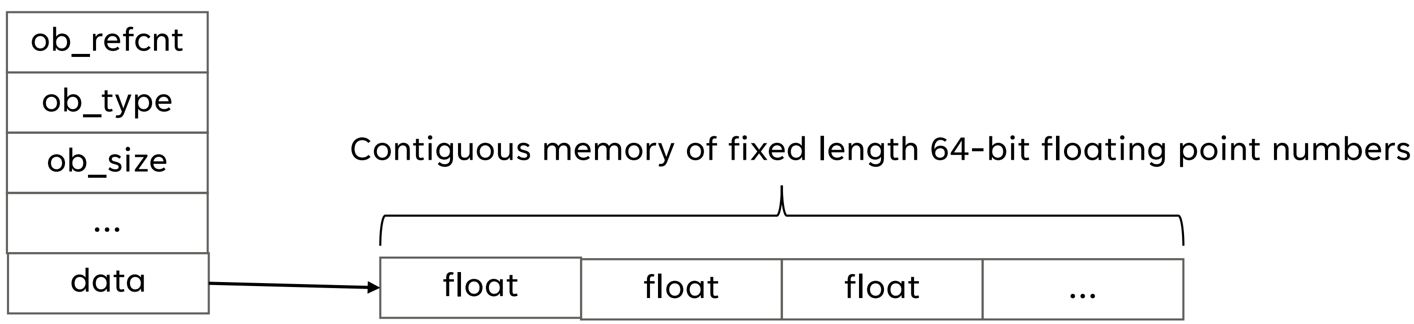
- large, memory-efficient, multi-dimensional arrays
- Fixed size integers and floats without
  - reference count field
  - type field
  - object size field
- Array and matrix operations
  - Utilizes vector operations when supported by the hardware. Thus FAST.

PYTHON LIST VS. NUMPY ARRAY

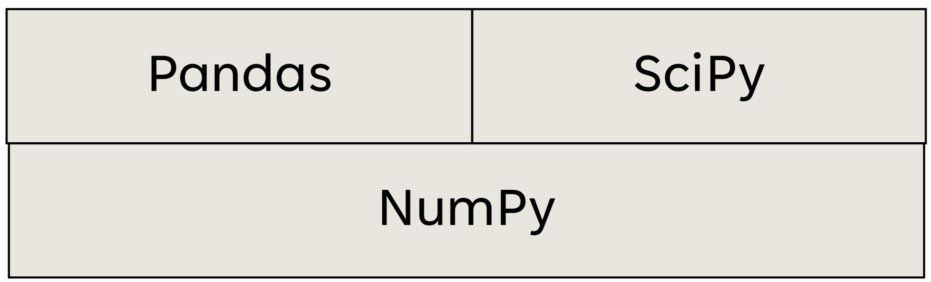
Python list



Numpy array



MANY OTHER LIBRARIES BUILD ON NUMPY



WHY USE DATA FRAMES?

NumPy great for raw arrays, but doesn’t provide an annotated tabular data type.

```
# lecture03/example1_data_frames.py
import pandas as pd

# Data to be represented in the DataFrame
data = {
    'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Eva'],
    'Age': [20, 21, 19, 22, 20],
    'Grade': [88, 92, 85, 90, 95]
}

# Create a DataFrame from the data
students_df = pd.DataFrame(data)

# Display the DataFrame
print(students_df)
```

CAN PERFORM MATH ON COLUMNS

Can add columns, multiply columns, ...

- With other columns
- With a constant

```
x = np.arange(n)  # Generate index values for x
y = np.random.rand(n)  # Generates n random numbers
df = pd.DataFrame( data: {'y': y}, index=x)

start_time = time.time()
df["newy"] = df["y"] + c
end_time = time.time()
```

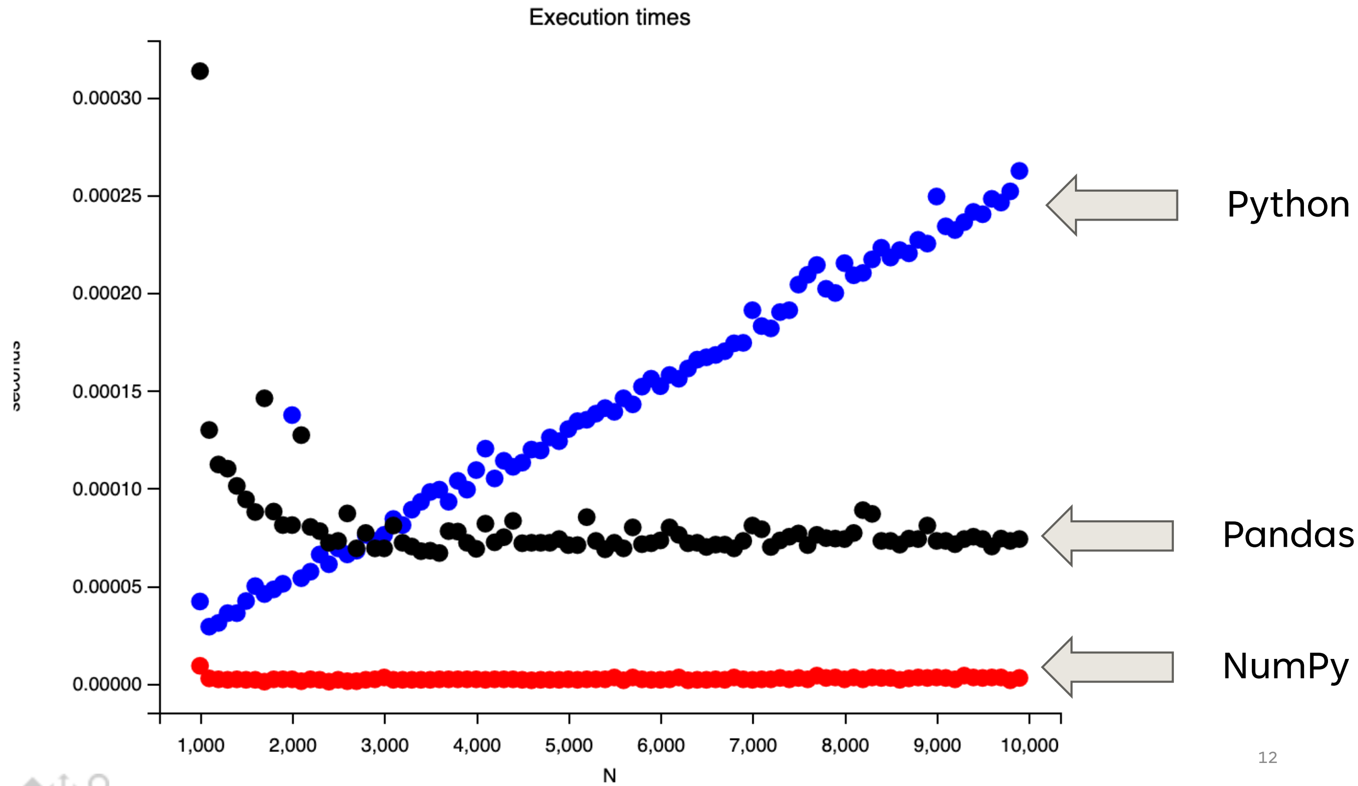
ARE DATA FRAMES FAST?

Pandas uses NumPy underneath.

Pandas = add constant to all elements in a column containing N elements.

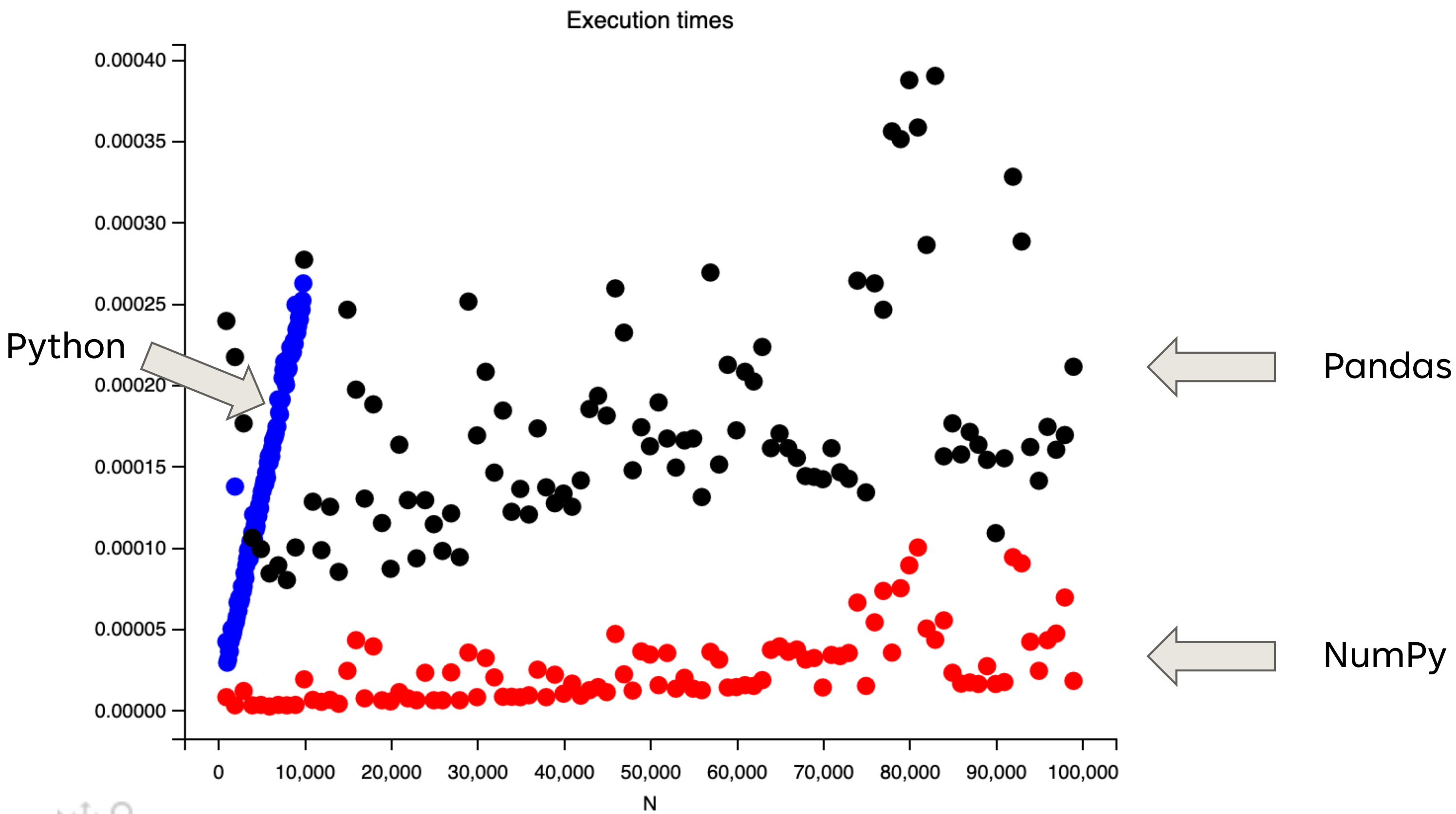
Python = add constant to all elements in a list of length N.

NumPy = add constant to all elements in a NumPy array of length N.



ARE DATA FRAMES FAST?

The tabular abstraction provided py Pandas comes with a cost.



COMMONLY USED OPERATIONS PROVIDED BY DATAFRAMES

DataFrames provides many operations that can be computed efficiently over columns:

- min
- max
- median
- percentiles
- standard deviation

(see lecture03/example2_df_notebook.dbc in class github repository)

DEFINITIONS FROM CHAPTER 1

Feature

A column within a table is commonly referred to as a *feature*.

Synonyms

attribute, input, predictor, variable

Outcome

Many data science projects involve predicting an *outcome* — often a yes/no outcome (in Table 1-1, it is “auction was competitive or not”). The *features* are sometimes used to predict the *outcome* in an experiment or a study.

Synonyms

dependent variable, response, target, output

Records

A row within a table is commonly referred to as a *record*.

Synonyms

case, example, instance, observation, pattern, sample

Table 1-1. A typical data frame format

HYPOTHESIS TESTING

Often in Data Science we are trying to find the answer to a question:

- Is a drug safe?
- Is a drug effective?

Or prove a hypothesis

- H1: Drug A is safe.
- H2: Drug A is effective.

Or find a statistic.

- GDP increased by 3.5% in the 3rd quarter.

OUTCOMES

In a clinical trial testing the safety of Drug X, outcomes might include:

- Incidence of specific side effects,
- Changes in vital signs (blood pressure, heart rate),
- Laboratory test results (liver enzyme levels, blood cell counts),
- Reports of adverse events,
- Patient-reported symptoms or quality of life measures.

The hypothesis "Drug X is safe" is a statement that is tested against the collected outcome data.

The specific outcomes measured in answering a hypothesis like safety are called “endpoints.”

Data scientists (or researchers) analyze these outcomes to determine whether they support or refute the hypothesis.

OUTLIERS

All real-world data is subjected to noise.   Noise can result in samples that land far from most of the other samples.

Some real-world processes also generate infrequent results far from the other samples.

Both are called ***outliers***.

Can we remove such outliers?

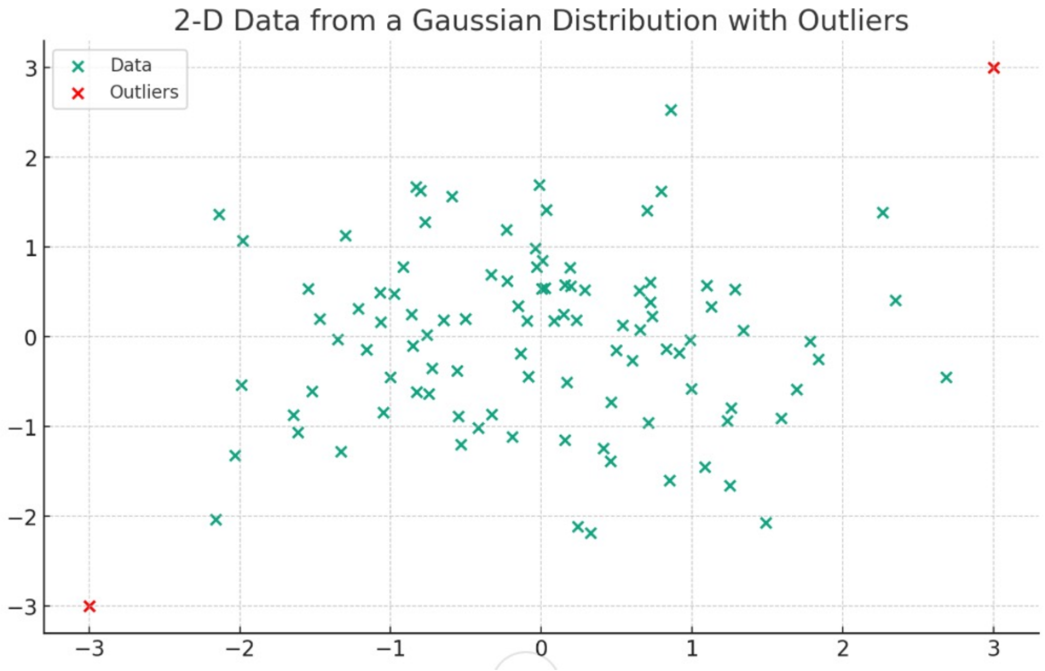
Are they due to ***natural variability***?

Or are they due to ***error*** in the data collection?

OUTLIERS

Are the outliers really due to error?

Can we remove them as spurious?



CAUTIONARY TALE: WAKEFIELD 1998

Andrew Wakefield published a paper showing a link between

- The Measles, Mumps, Rubella (MMR) vaccine and
- autism.

Wakefield had undisclosed funding from lawyers representing parents suing multiple vaccine manufacturers.

Brian Deer made allegations of cherry-picking that were eventually published in the British Medical Journal.

- General Medical Council stripped Wakefield of his license.
- Lancet retracted the paper in 2010.

Andrew Wakefield

Brian Deer

ERROR

All real-world data is subjected to error.   Error can be categorized as

- Systematic error (Bias)
  - Observer bias
  - Selection bias
  - Measurement bias
  - Confounding factors
- Random error (Noise)
  - Measurement error
  - Heisenberg uncertainty

OBSERVER BIAS: LAETRILE

- Biochemist Dr. Ernst T. Krebs, Jr often credited for popularizing Laetrile (Amygdalin/B17) in 1950s through 70s as a cancer treatment.
  - **Most of the support came from anecdotal evidence.**
  - **Known for showcasing testimonials**
- National Cancer Institute in 1982 published clinical trial in New England Journal of Medicine concluding that data did not support the case for efficacy of Laetrile.
- FDA has refused to approve Laetrile as a cancer treatment.
- Still significant support today for Laetrile.

Ernst T. Krebs, Jr.

OBSERVER BIAS: CLEVER HANS

OBSERVER BIAS: CLEVER HANS

- In early 20th century, math teacher Wilhelm von Osten claimed his horse Clever Hans could do math and spelling.
- Hans would tap his hoof to give his answer.
- In 1907, psychologist Oskar Pfungst performed experiments in which:
  - **Clever Hans could not see any observers**
- When Hans could not see the questioner, he didn't know the answer.

- Good reason to use blinding!

Clever Hans and Wilhelm von Osten

SELF-SELECTION BIAS: KELLER

In the 1960s, Psychologist Fred Keller developed the “Personalized System of Instruction”

Emphasized:

- Self-paced learning
- Master material before moving forward
- Use of proctors

Fred S. Keller

SELF-SELECTION BIAS: KELLER

Problems in Keller's studies:

- Self-Selection bias:
  - Significantly above average students tended to volunteer.
  - Skewed results in favor of PSI.
- Lack of blinding
  - Both students and instructors knew they were using PSI.
- Instructor enthusiasm
  - Another source of self-selection bias, but on the part of the teachers.
  - More enthusiastic teachers were more likely to implement PSI.
  - More enthusiastic teachers leads to better performance even when NOT using PSI.

2ND CAUTIONARY TALE: KELLER

- Failure to recognize limitations of a study can backfire.
- Keller was derided for some of the limitations in his studies
- Research in PSI diminished over time, but interest remained particularly in math.
  - **Kumon**
- Resurgence when computers allowed us to overcome some of the limitations:
  - **Self-paced learning with active / interactive learning**
    - **Codeacademy**
    - **Brilliant**
  - **Repetition of similar questions until demonstration of mastery**
    - **Khan Academy**
  - **Gamification**
    - **Duolingo**

3RD CAUTIONARY TALE: KELLER

- Sometimes self-selection bias is itself important and can be used to identify a cohort for which a strategy is more effective.
- Self-paced courses may work better for those that naturally self-select.
  - **Self-motivated**
  - **Academically capable within the scope of the material.**
- Is the existence of self-selection bias a reason to abandon self-paced courses just because they don't work for some people?

HOW DO WE DEAL WITH OUTLIERS

1. Remove outliers.
   - **Better have a good explanation as to why the data is erroneous.**
2. Use metrics that are less affected by outliers.

CENTRAL TENDENCY

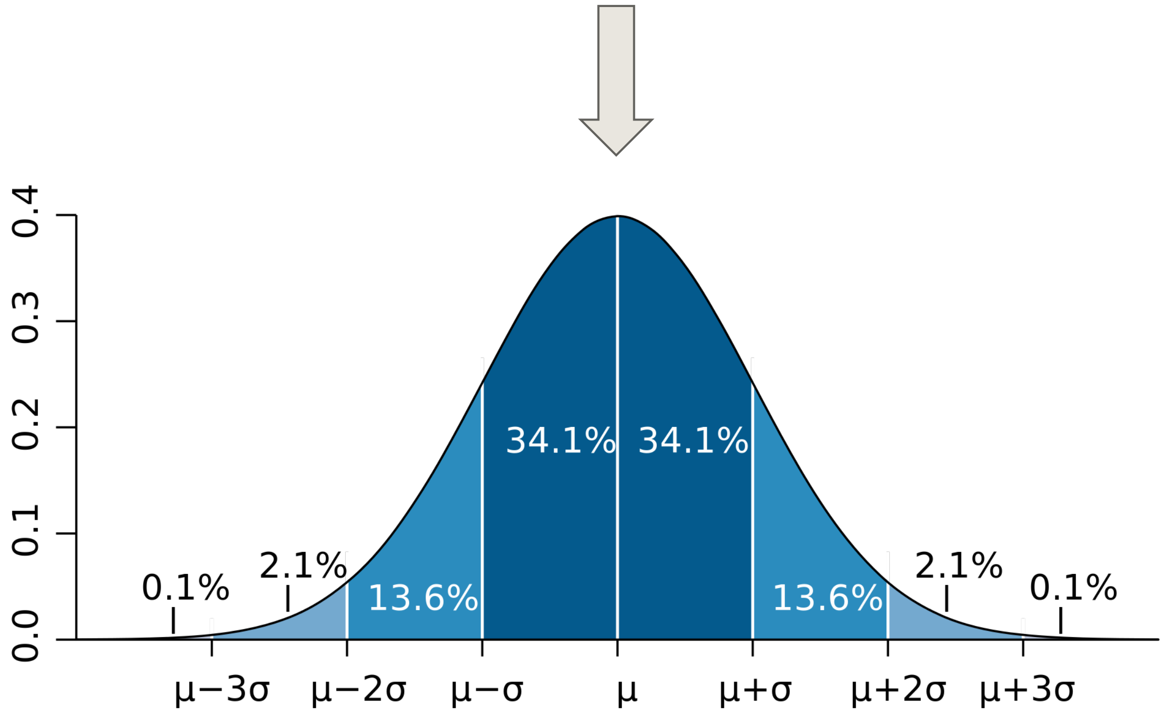
A measure of central tendency is a “typical value” for a probability distribution.

Covered in Chapter 1

Means, medians, truncated means

When to not use mean?

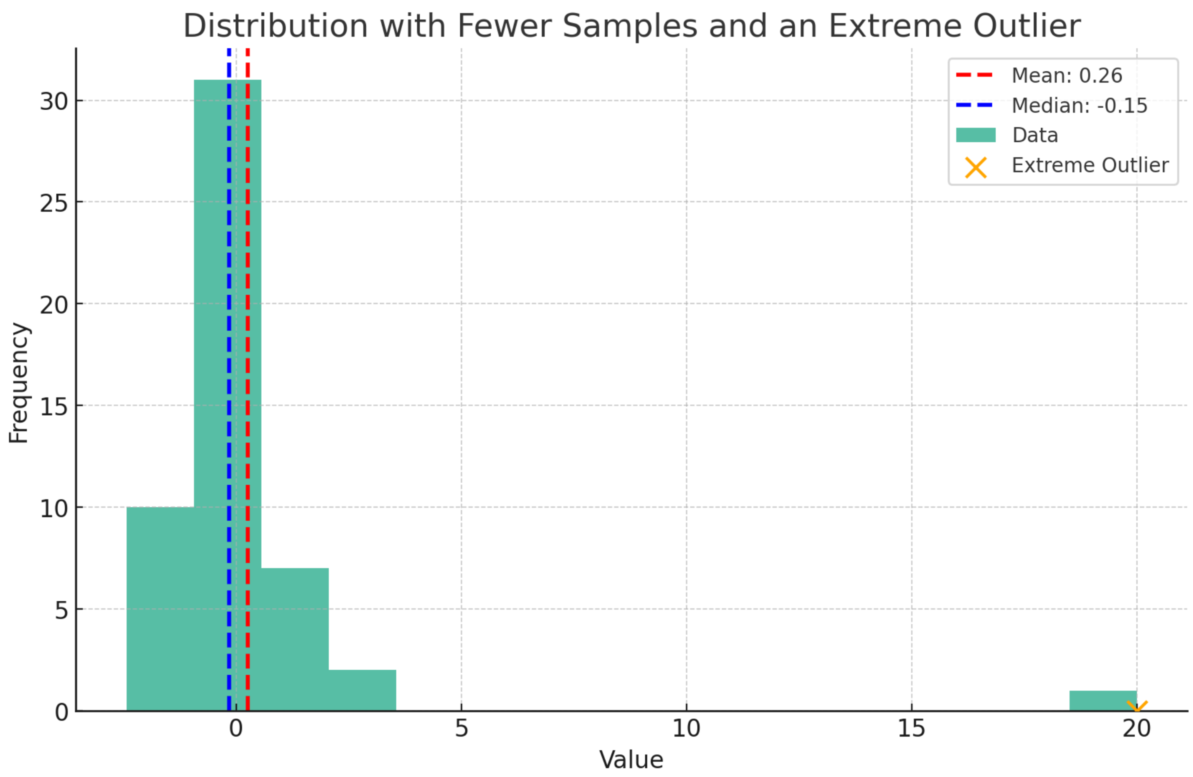
Both mean and median are good metrics of central tendency for a symmetric distribution.



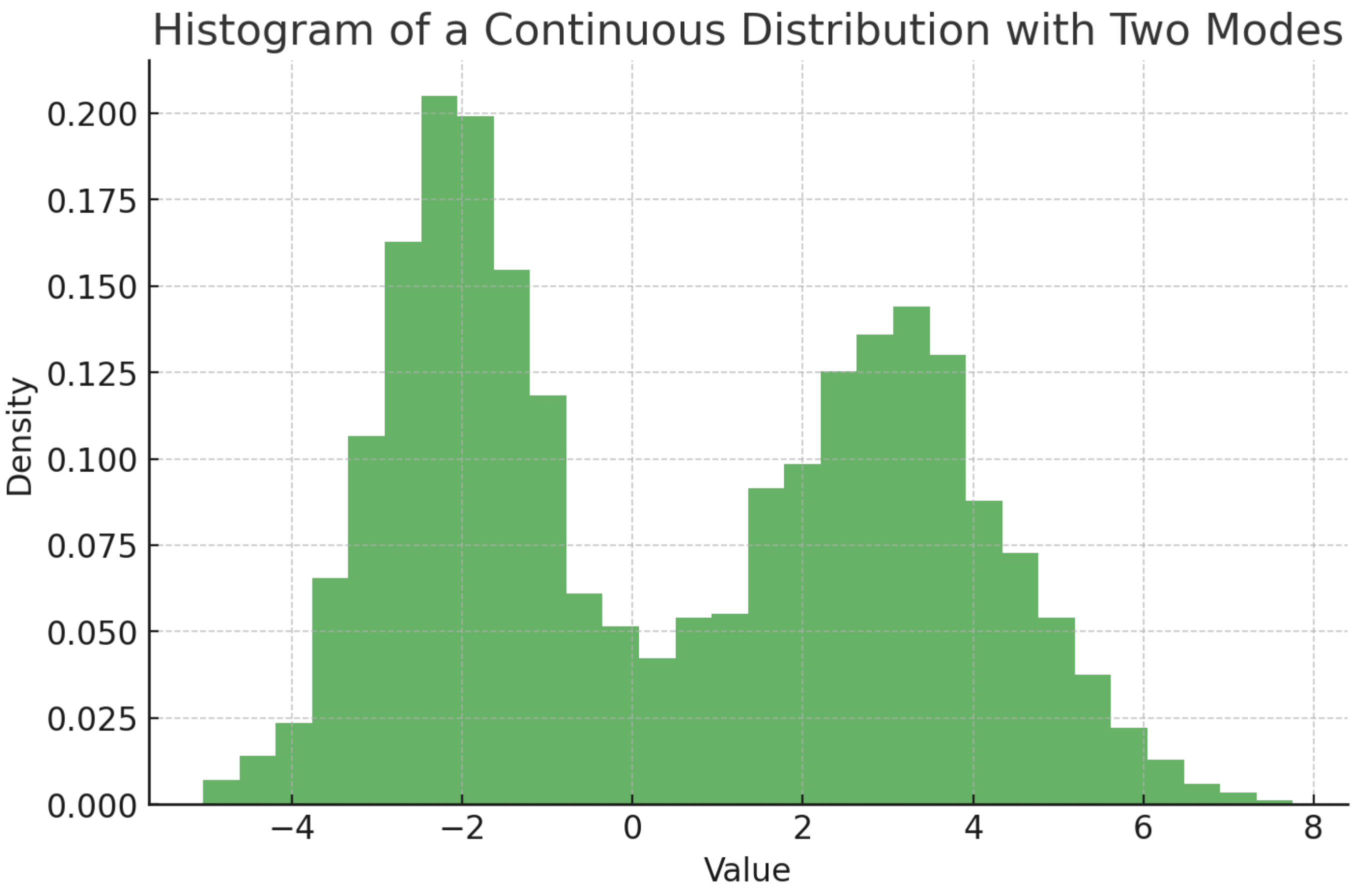
EXTREME OUTLIERS: BAD FOR MEAN

Particularly important when few samples or noisy data.

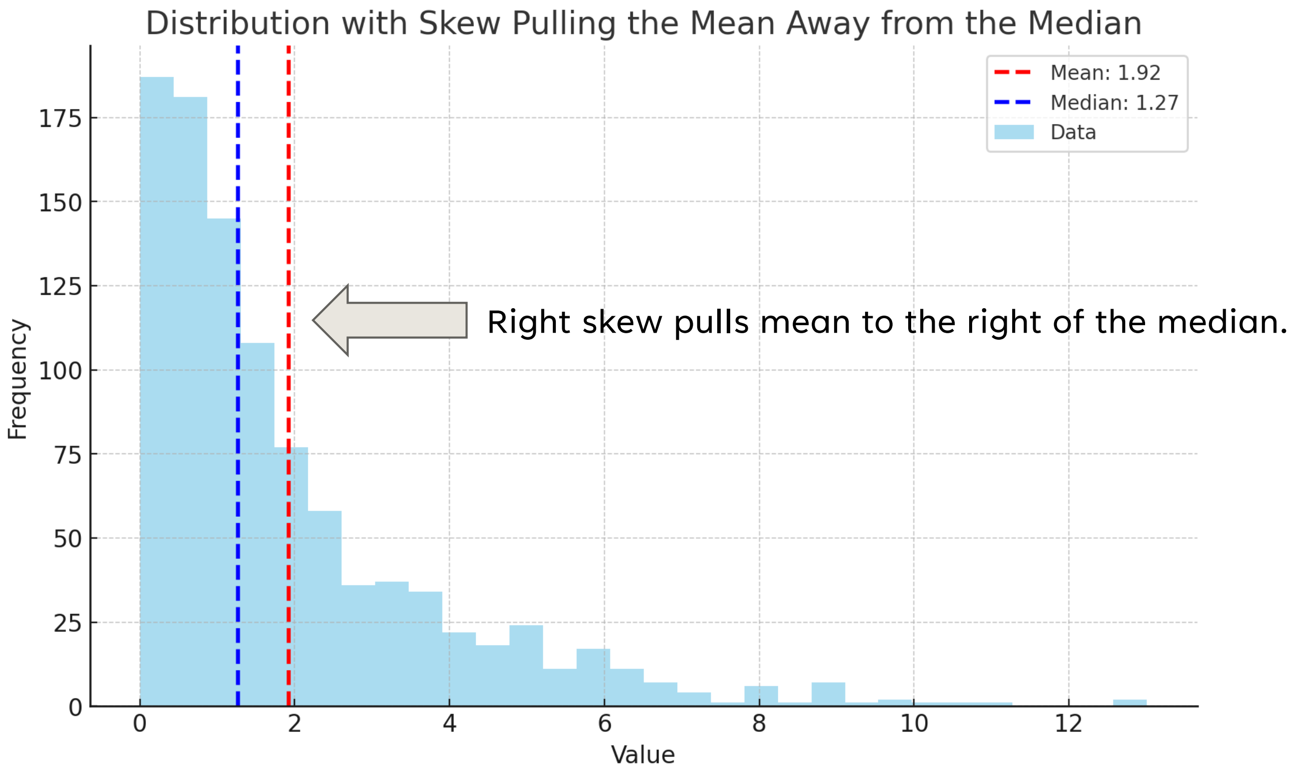
A single extreme outlier can throw off the mean making mean no longer a good metric for central tendency.



MULTIPLE MODES: MISLEADING MEAN?

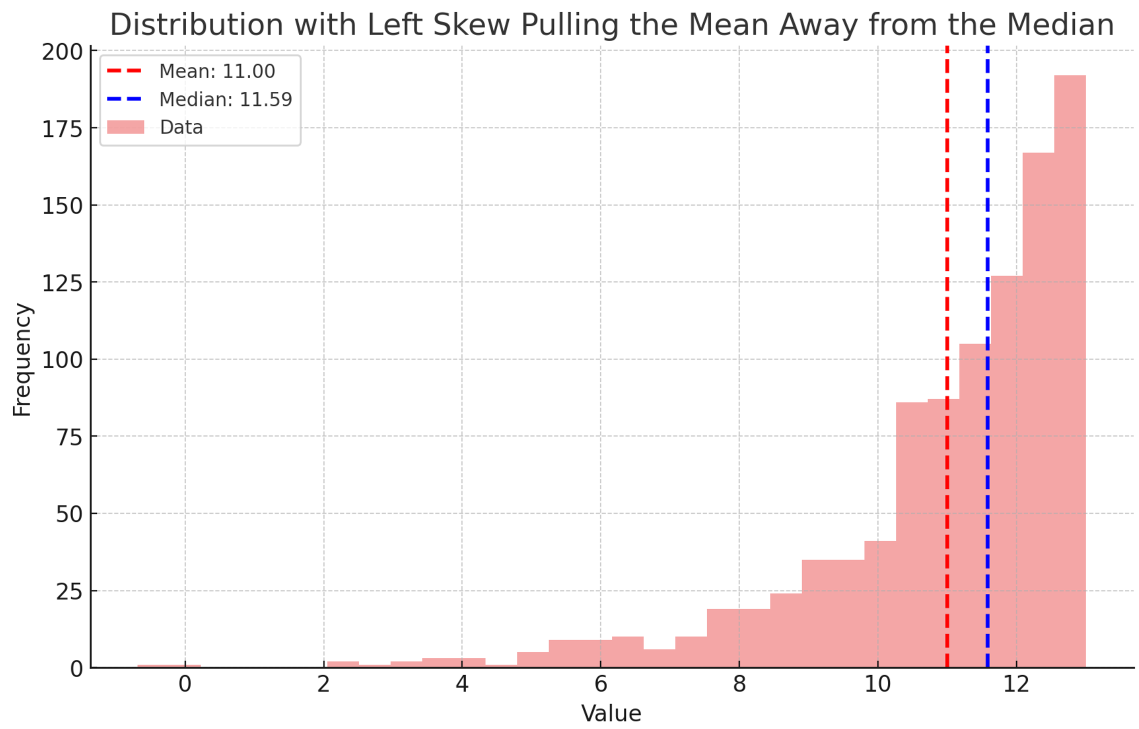


SKEW



Skew can cause significant difference between the mean and median

SKEW



Left skew

NEXT TIME

- Dispersion
- More on visualization

THANK YOU

David Harrison

Harrison@cs.olemiss.edu