CSCI 443 LECTURE 4: BIAS AND ERROR

Professor David Harrison

TODAY

- Tradeoffs Pandas, Databricks
- Bias
- Error

HOMEWORK 1

Due Thursday, January 30.

11 pm.

Focuses on

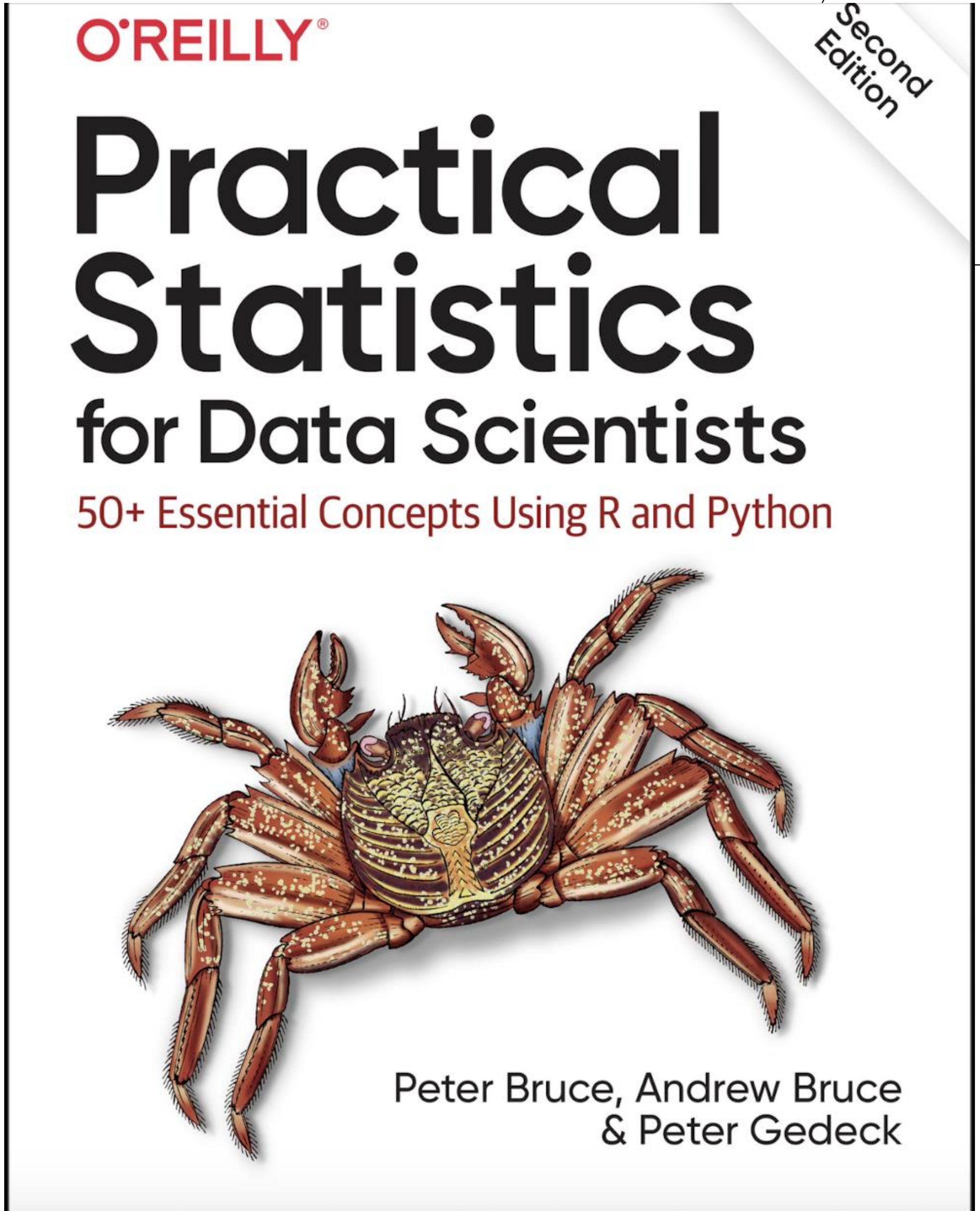
- setting up accounts,
- using github and Databricks
- Notebooks.

Submission:

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

READING!

- Book provides examples in Python and R. We are using Python.
- Read Chapter 1: Exploratory Data Analysis.



OFFICE HOURS

Due to scheduling conflict, office hours updated

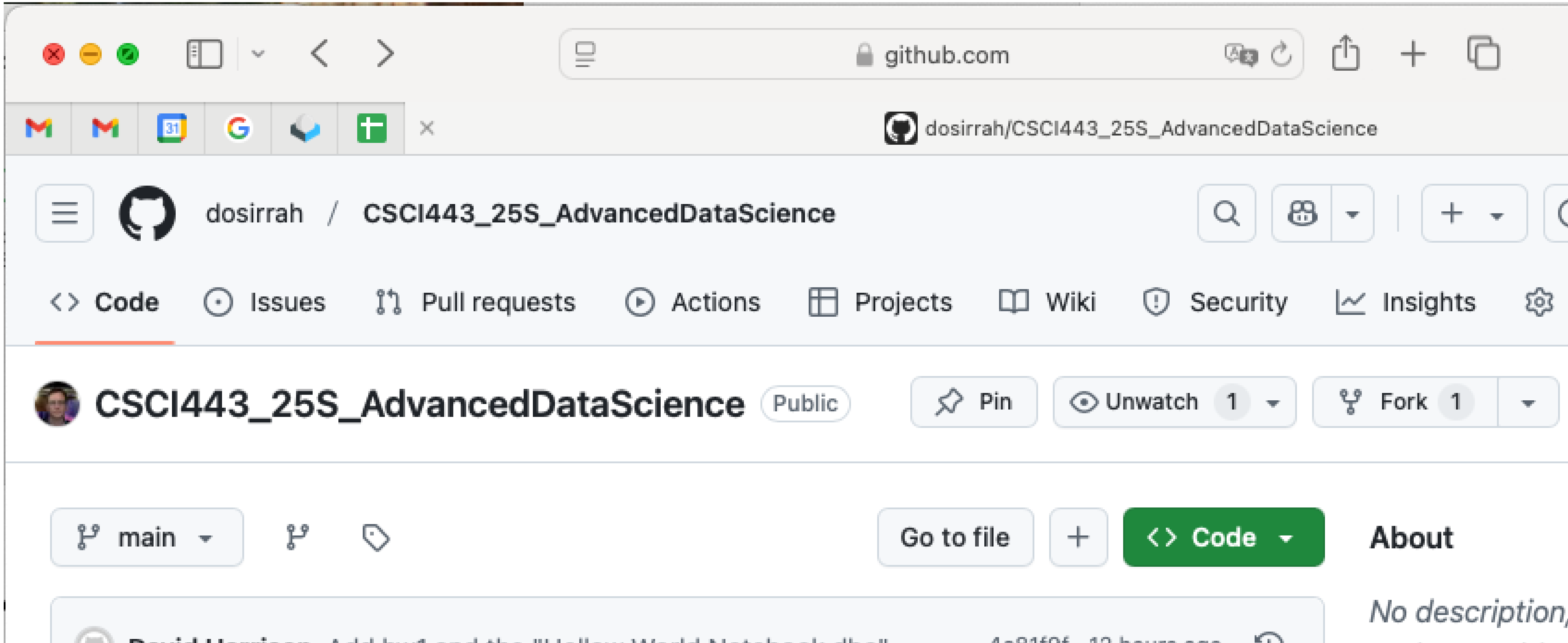
Monday	1:00-2:00 PM
Tuesday	4-5 PM

GITHUB

Lecture slides and examples have been committed to GitHub for lectures 1, 2 and 3.

The project is at

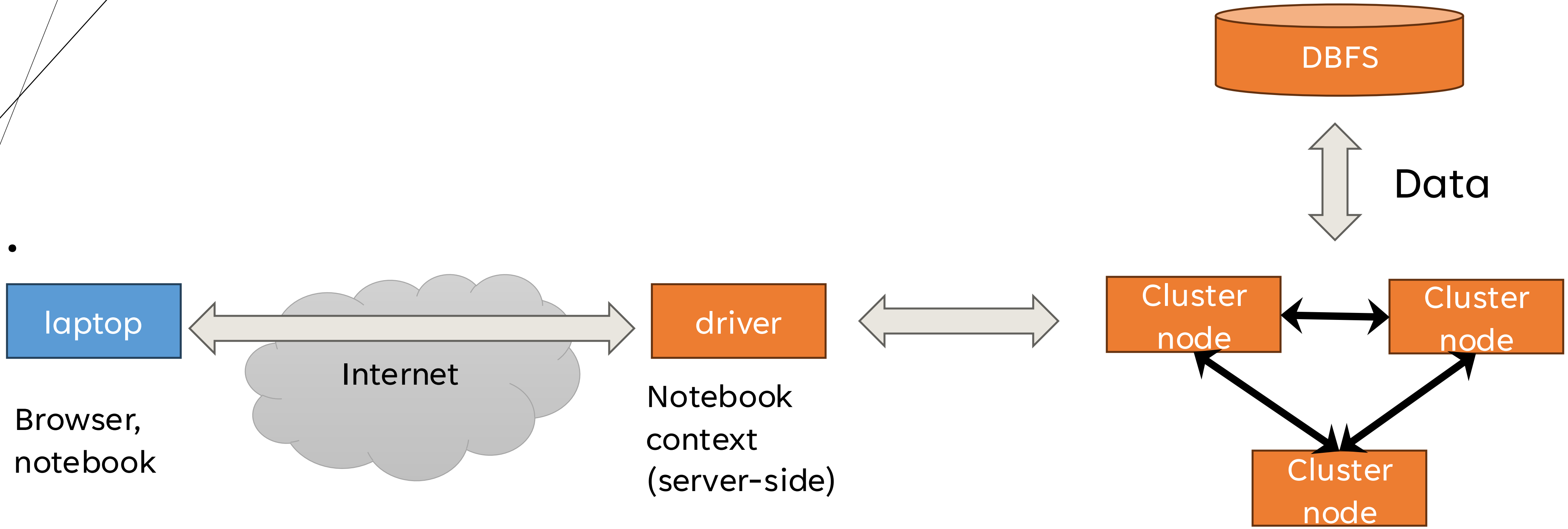
https://github.com/dosirrah/CSCI443_25S_AdvancedDataScience



REVIEW OF TRADEOFFS

	Numpy	Pandas DataFrame	Spark DataFrame
Best for	Numerical computations	Tabular data.	Tabular. Big data.
Scalability	🚫 Limited. Single machine, RAM-bound	🚫 Limited. Single machine, RAM-bound	✅ Distributed processing
Parallelism	Single-threaded but vectorized	Single-threaded but vectorized (via Numpy)	Vectorized and distributed
Lazy	🚫 No	🚫 No	✅ Optimized execution planning
Memory Usage	🚫 Entire data set must fit in RAM	🚫 Entire data set must fit in RAM	✅ **Optimized** – Uses disk, caching, and partitions
Persistence	🚫	✅ in-memory but can save to disk	✅ Distributed storage (DBFS on top of HDFS, S3, Azure Blob, GCS)

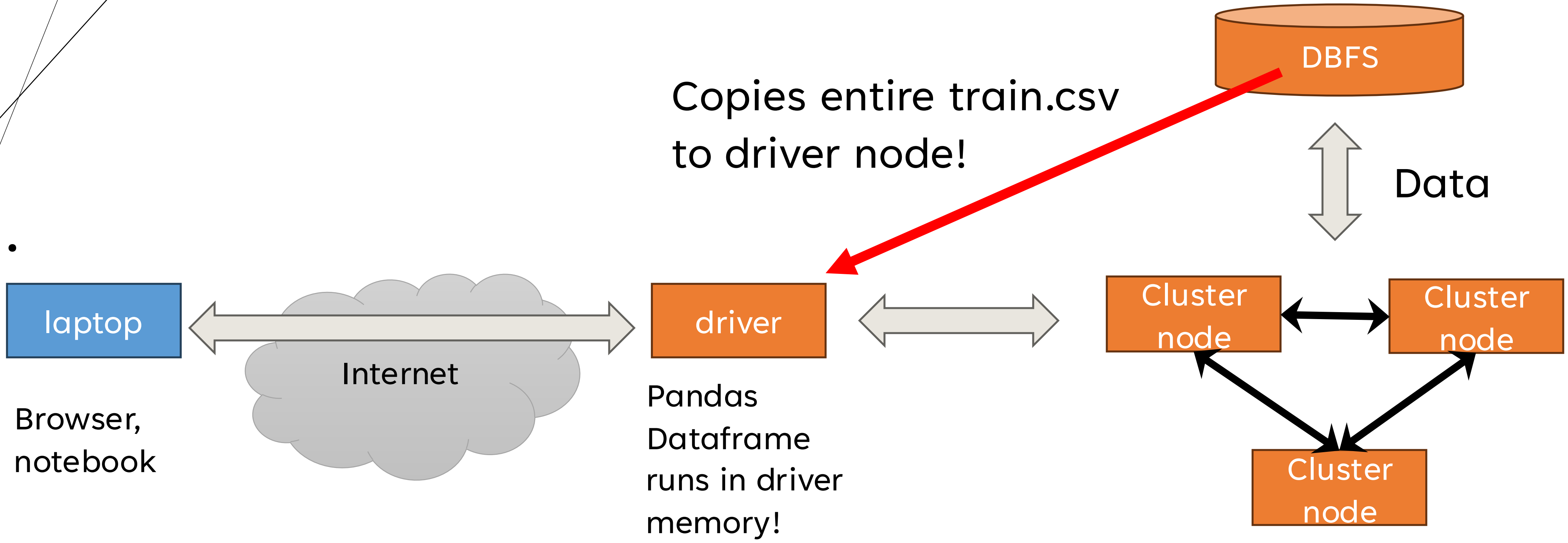
SPARK ARCHITECTURE (1000 FT)



HOMEWORK 1 PART 4 TRADEOFFS

	.toPandas()	pyspark.pandas	pyspark.sql
Best for	Local manipulation of small tables. Exploring a subset of data.	Pandas users moving to Spark	Massive-scale, distributed data processing
Scalability	🚫 Limited. Single machine, RAM-bound	✅ Moderate- Spark with some overhead.	✅ Good. Distributed processing
Performance	🚫 Slow for large datasets.	✅ Mostly good but sometimes slower than native Spark	✅ Good
Lazy	🚫 No	🚫 Mostly No, some optimizations but limited to Pandas semantics	✅ Optimized execution planning
Memory Usage	🚫 Entire data set must fit in RAM	✅ Moderate – uses Spark but retains Pandas semantics	✅ **Optimized** – Uses disk, caching, and partitions
SQL Integration	🚫 No.	🚫 No.	✅ Full SQL!!!

TOPANDAS()


DBFS
Copies entire train.csv
to driver node!
Data
laptop
Internet
driver
Cluster node
Cluster node
Cluster node
Browser,
notebook
Pandas
Dataframe
runs in driver
memory!

FROM CHAPTER 1

From Chapter 1, you should know (or learn quick)

- Types of data
  - Numerical, categorical
- Outcomes, records, …
- Estimates of Location (entire section)
  - Mean, median, percentile, weighted mean, trimmed mean.
- Estimates of Variability (entire section)
  - Variance, standard deviation, mean absolute deviation, median absolute deviation, range, order statistics, interquartile range

FROM CHAPTER 1: EXPLORING DATA DISTRIBUTION

From Chapter 1, you should know (or learn quick)

- Percentile and Boxplots
- Frequency Tables, Histograms
- Density Plots and Estimates
- Mode, Expected value, Bar charts, Pie charts

We will cover all of these in homework 2.

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

CAUTIONARY TALE: WAKEFIELD 1998

Cherry-picking:

- Ignored children who received the vaccine without developing autism.
- Ignored multiple data sets contradicting his hypothesis.

Small sample size

- 12 children. Not statistically significant
- No control group

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

ERROR

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

- Systematic error (Bias)
    - Observer bias: **researcher's beliefs influence observations**
    - Selection bias
    - Measurement bias
    - Confounding factors
- Random error (Noise)
    - Measurement error
    - Heisenberg uncertainty

ERROR

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

- Systematic error (Bias)
  - Observer bias: researcher's beliefs influence observations
  - Selection bias: **selection of data points is not random.**
  - Measurement bias
  - Confounding factors
- Random error (Noise)
  - Measurement error
  - Heisenberg uncertainty

ERROR

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

- Systematic error (Bias)
  - Observer bias: researcher's beliefs influence observations
  - Selection bias: selection of data points is not random.
  - Measurement bias: **tools introduce systematic error.**
  - Confounding factors
- Random error (Noise)
  - Measurement error
  - Heisenberg uncertainty

ERROR

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

- Systematic error (Bias)
  - Observer bias: researcher’s beliefs influence observations
  - Selection bias: selection of data points is not random.
  - Measurement bias: tools introduce systematic error.
  - Confounding factors: **affect both independent and dependent variable**
- 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.

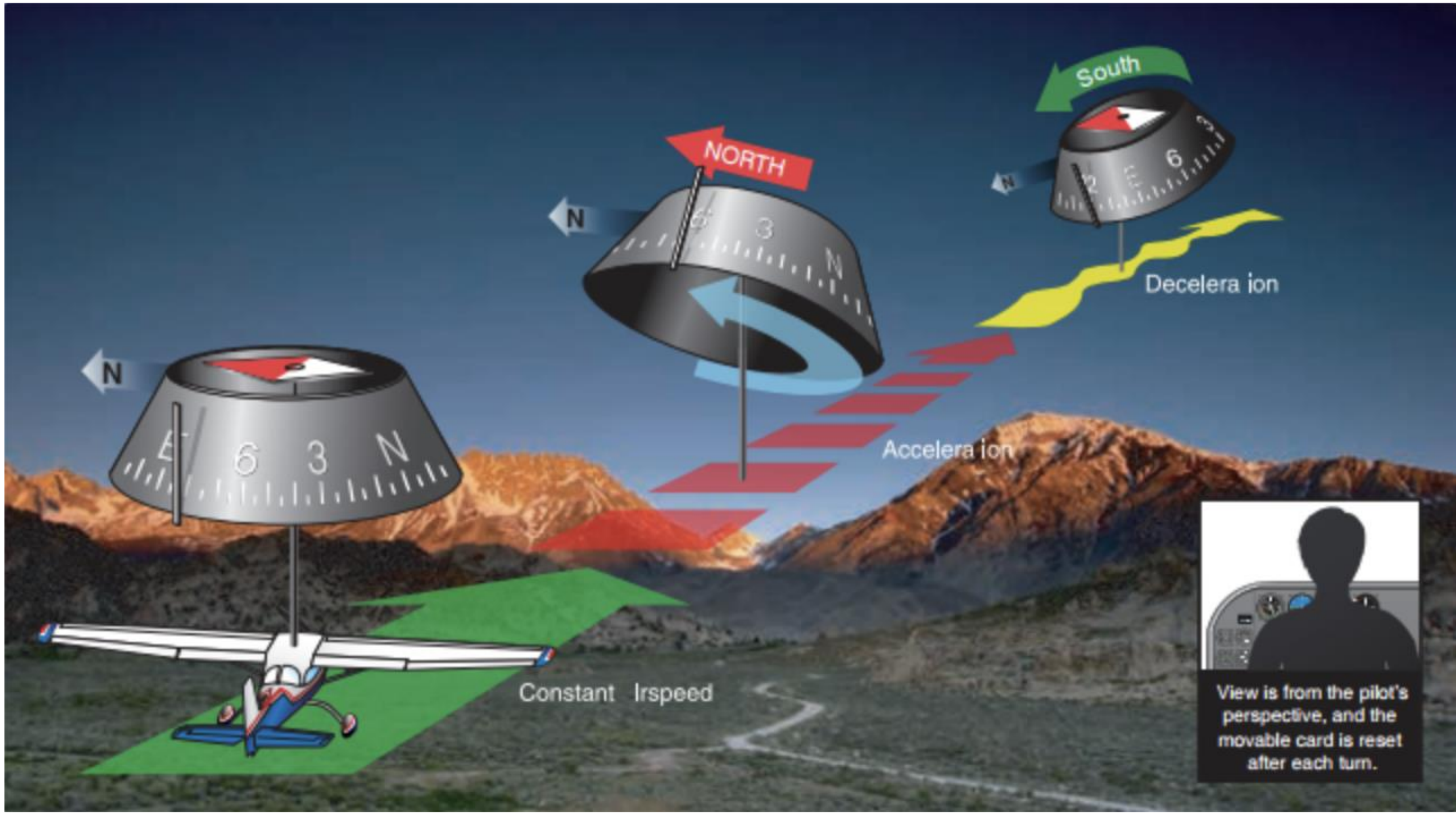
2^{ND} CAUTIONARY TALE: KELLER

- 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?

MEASUREMENT BIAS



- Instrument bias:
  - Failure to tare a scale
  - Acceleration error in airplane compasses (ANDS)

MEASUREMENT BIAS

- Social Desirability Bias
  - Also examples of self-reporting bias
    - Answer in way that will be perceived favorably by others.
    - Self-reported dietary intake
    - Self-reported exercise
    - TV consumption avoiding guilty pleasures or reality TV

CONFOUNDING FACTORS

A confounding factor, also known as a confounder, is a variable that influences both the dependent variable and independent variable. This can lead to misleading conclusions about the relationship between the variables of interest.

Examples:

- Socioeconomic Status (SES) and health
  - Are people healthier because they have higher SES?
  - Or do people of higher SES tend to have better access healthy food and can afford a gym?
  - Or better access to doctors?
  - [Foster, Polz, et al 2020] shows the issue is complex, but does not refute the clear correlation between unhealthy lifestyles and various conditions, non-communicable diseases, and mortality.

CONFOUNDING FACTORS

- Exercise and weight loss
  - **Exercise reduces weight!**
- Confounding factor
  - **Diet.**
- 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?

CENTRAL TENDENCY

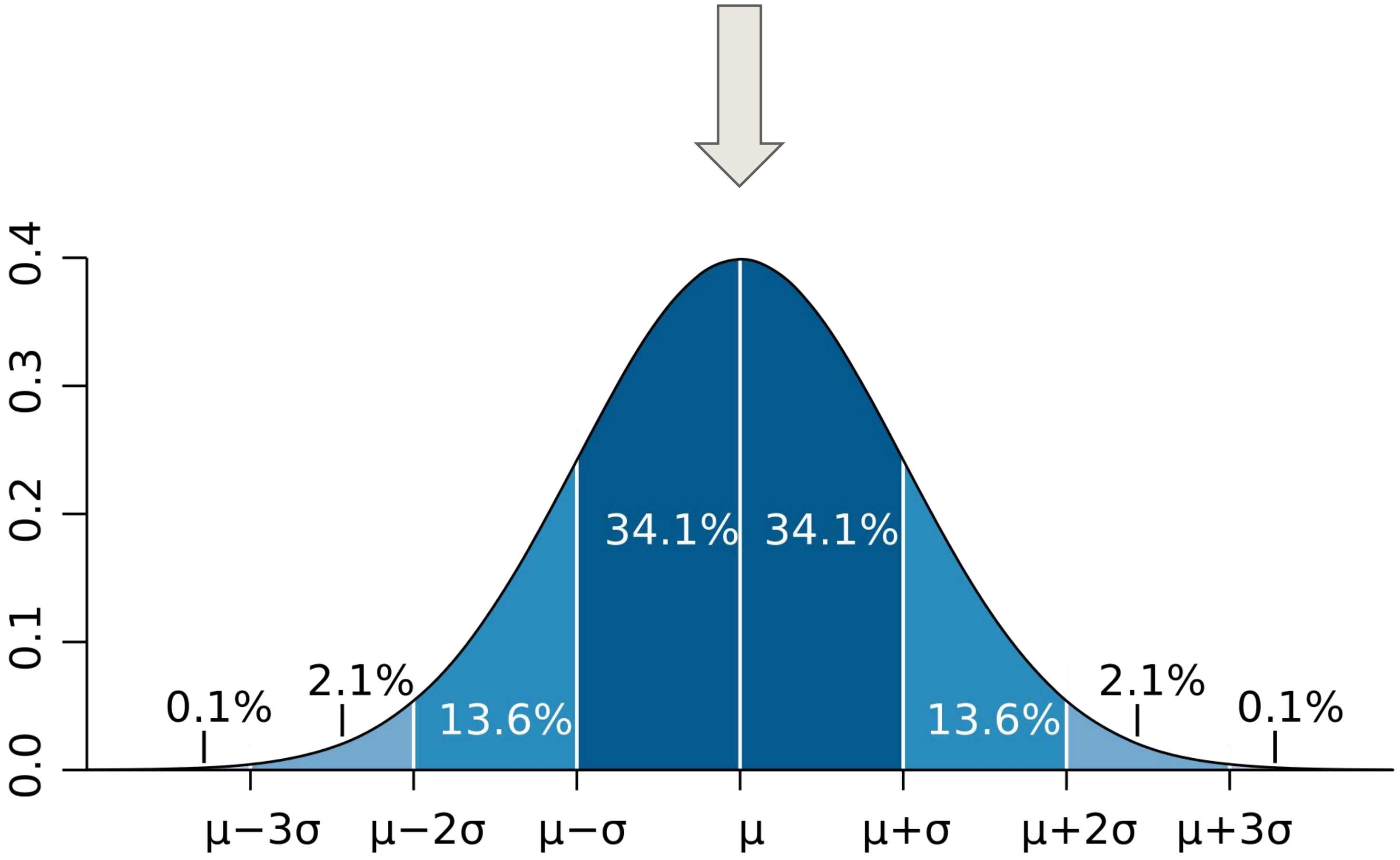
A measure of central tendency is a “typical value” for a [probability distribution](#).

Covered in Chapter 1

Means, medians, truncated means

When to not use mean?

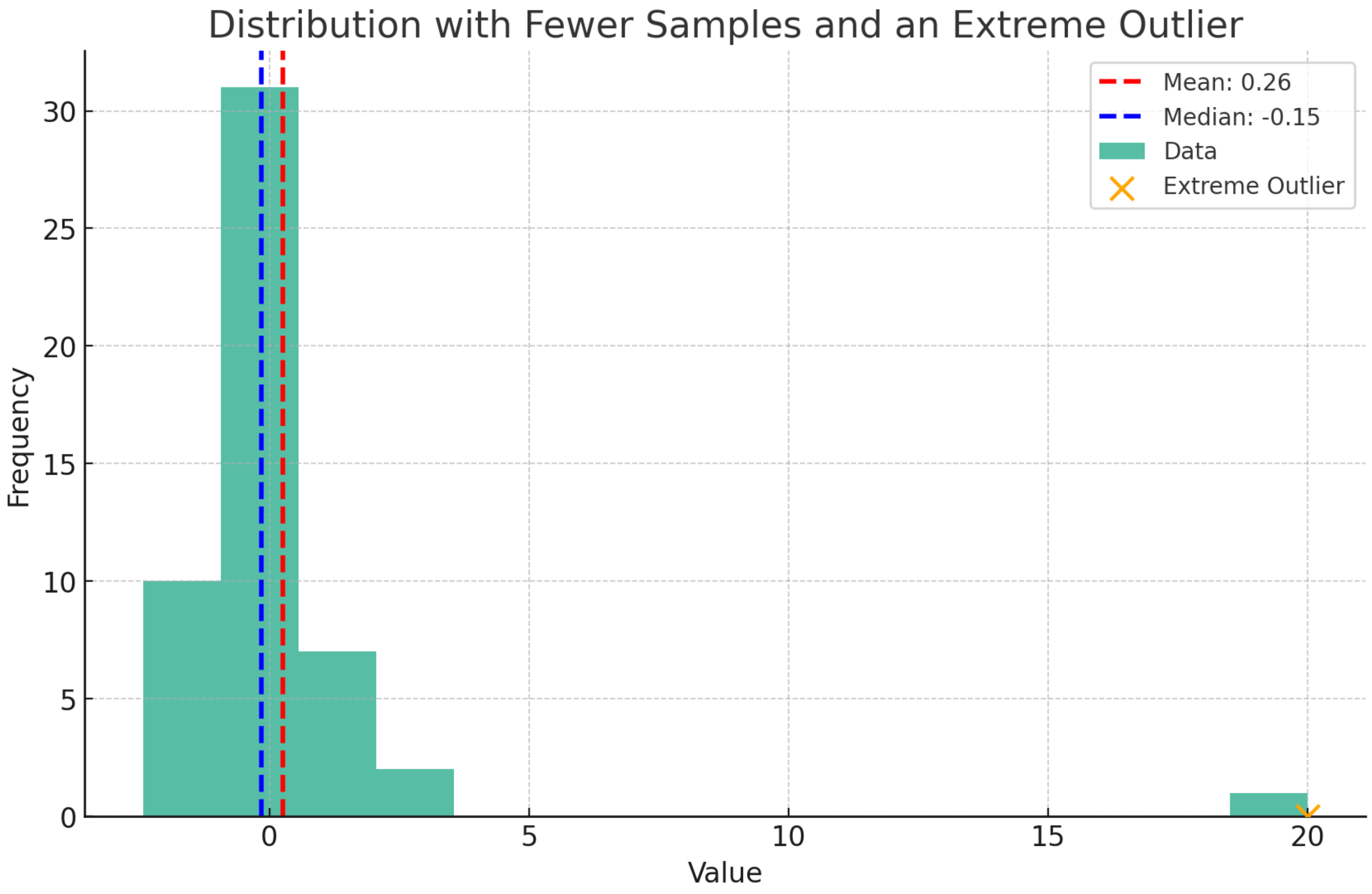
Both mean and median are good metrics of central tendency for a symmetric distribution.



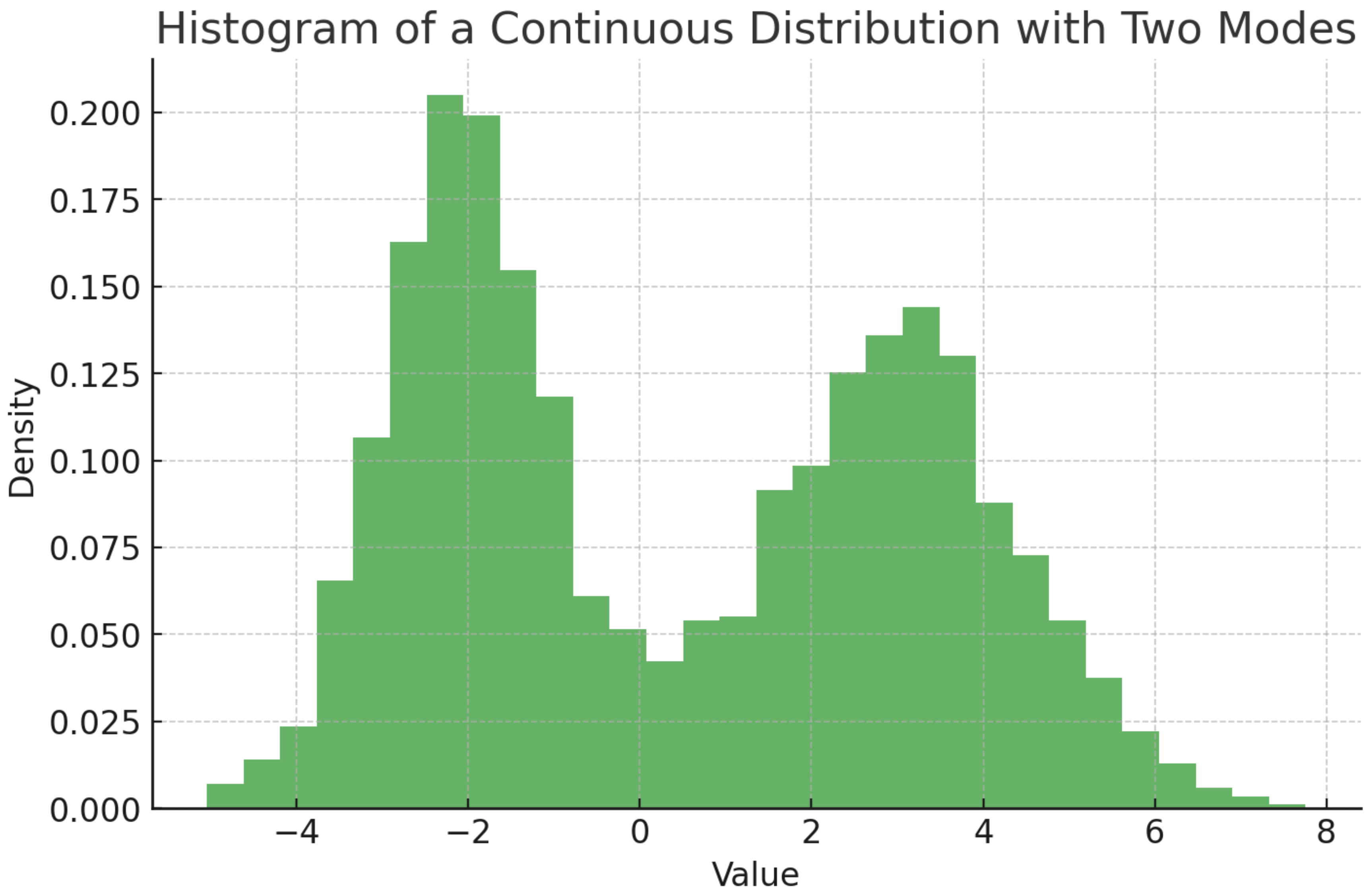
EXTREME OUTLIERS: BAD FOR MEAN

Particularly important when few samples or noisy data.

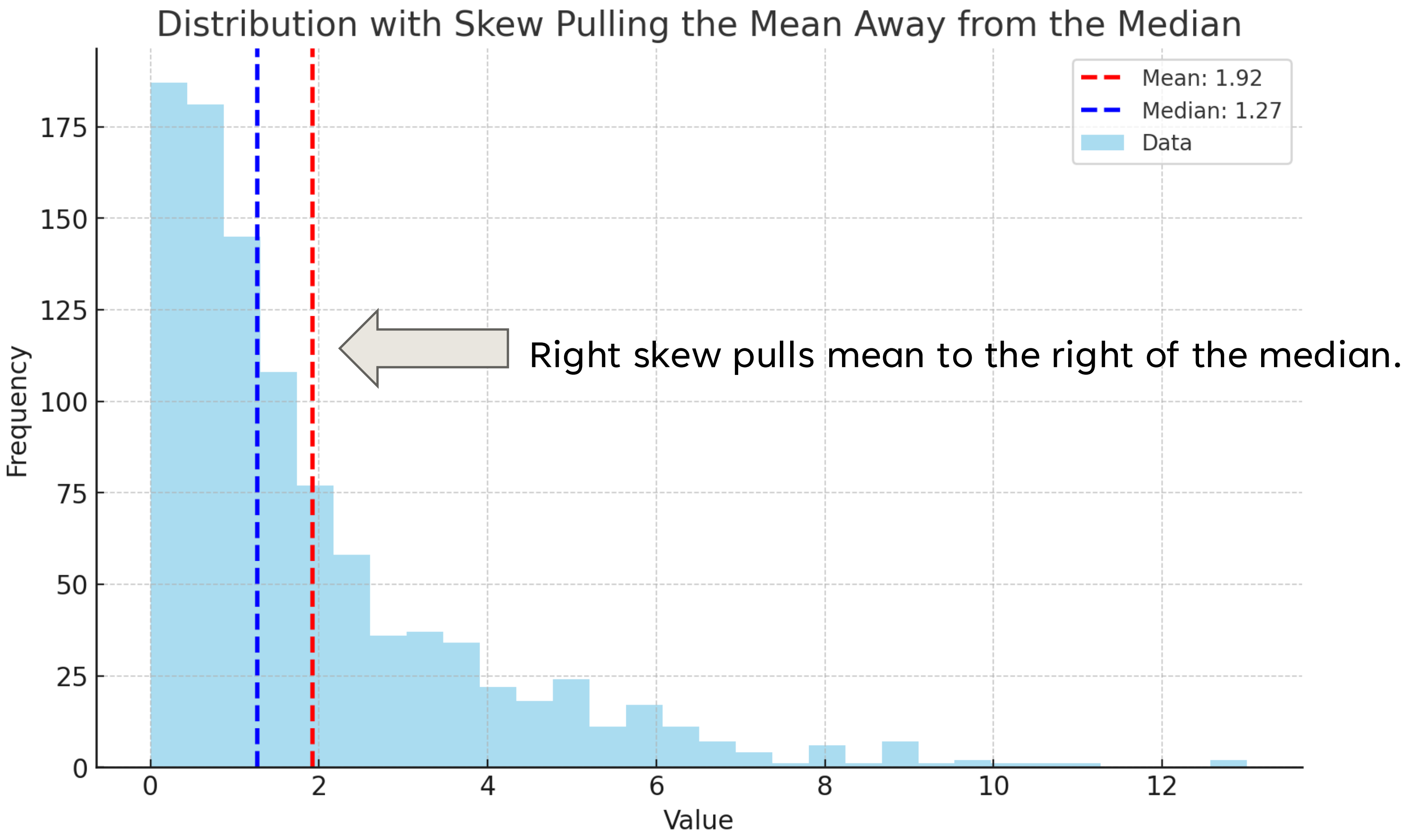
A single extreme outlier can throw off the mean making mean no longer a good metric for central tendency.



MULTIPLE MODES: MISLEADING MEAN?

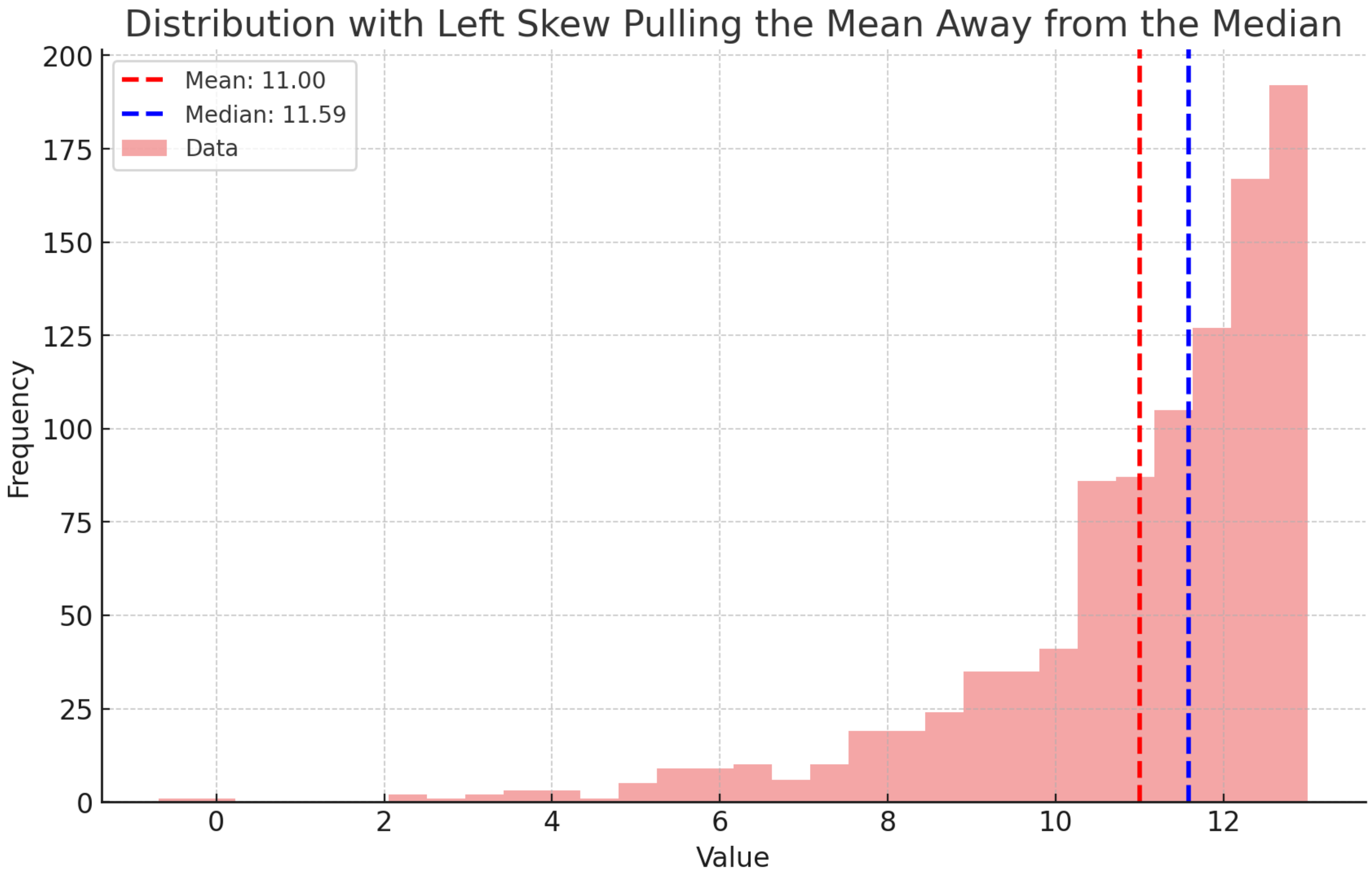


SKEW



Skew can cause significant difference between the mean and median

SKEW



Left skew

RANDOM VARIABLE

Random variable assigns numbers to outcomes.

- T = 0
- H = 1

For dice:

- Roll 1 = 1
- Roll 2 = 2
  - …
- Roll 6 = 6

We can then assign probabilities to each value the random variable can take.

Wikipedia says,

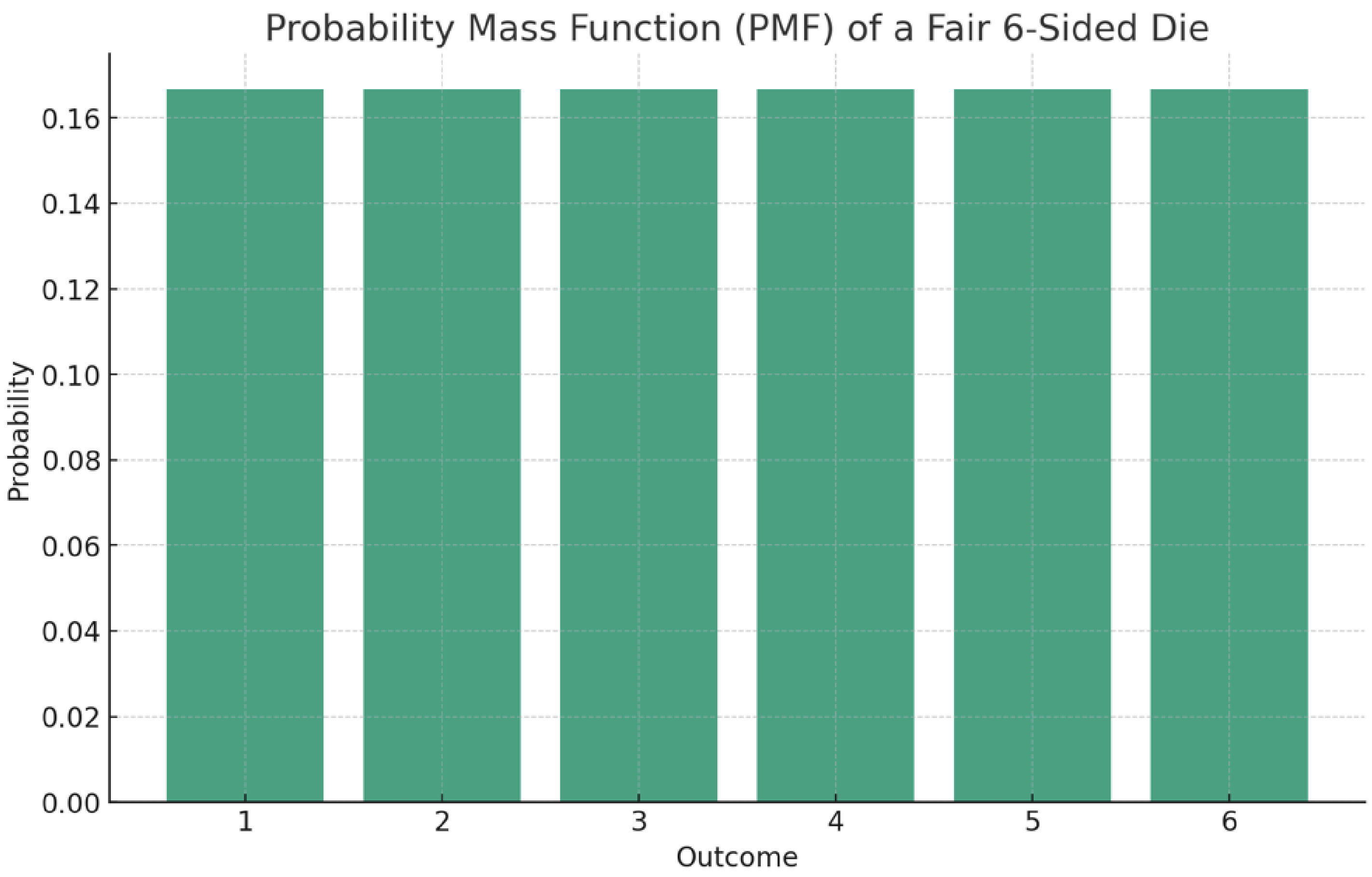
In probability theory and statistics, a **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible **outcomes** for an experiment.[1][2] It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).[3]

PROBABILITY MASS FUNCTION

Describes the probability of each discrete outcome.

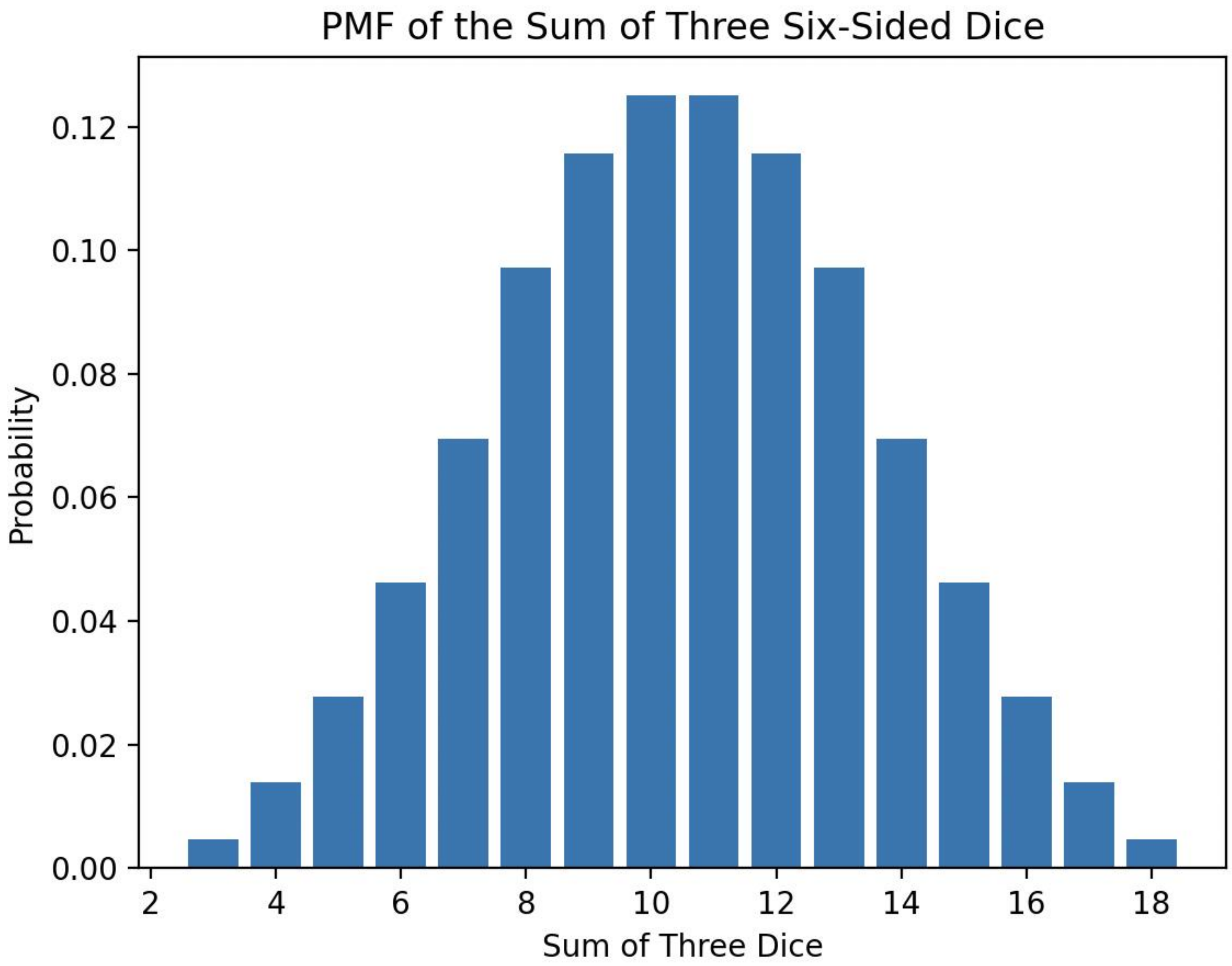
For discrete random variables, a PMF loloks like a histogram where each bin refers to a single outcome.

Sum of probabilities must be 1.



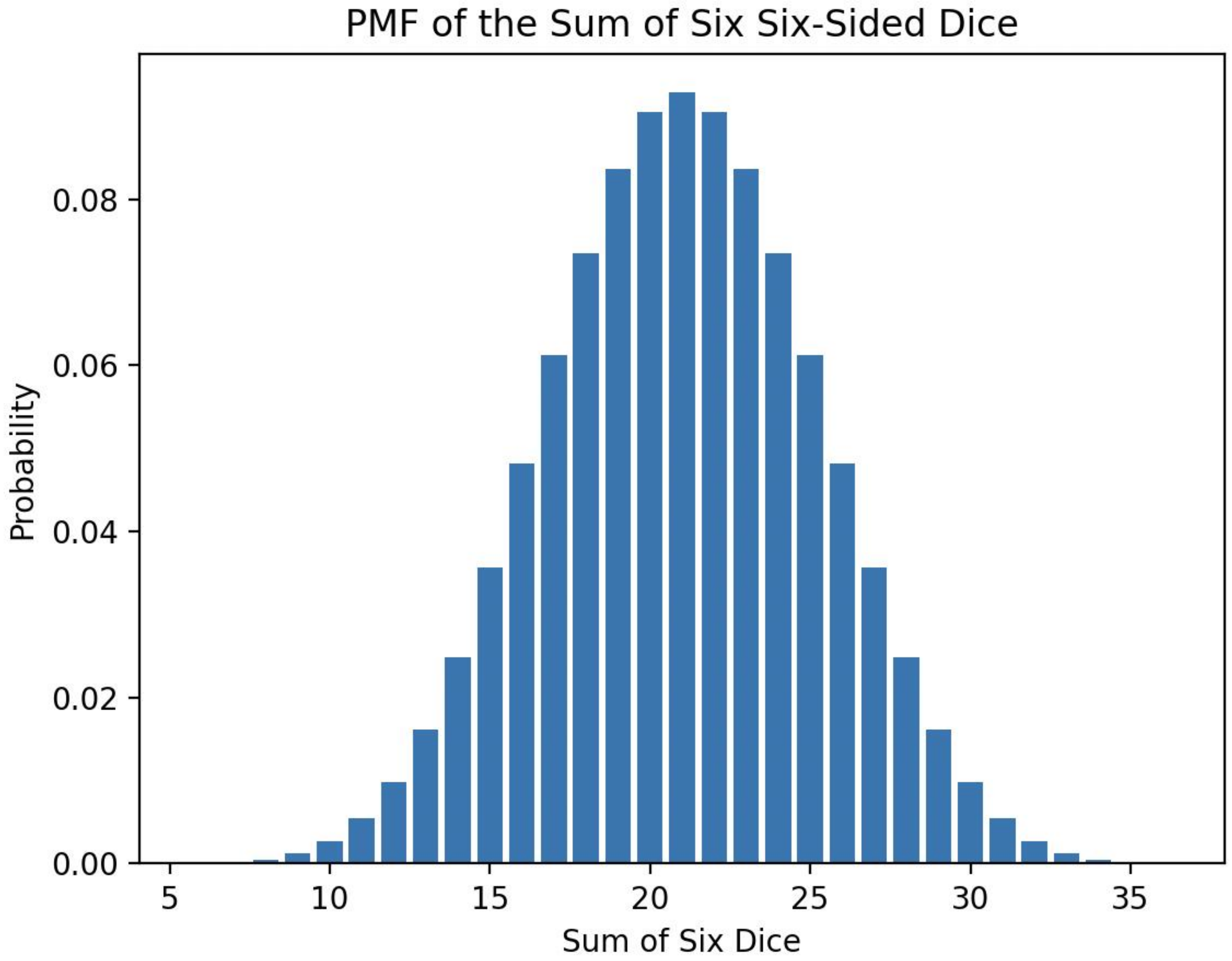
PROBABILITY MASS FUNCTION

Sum of three dice



PROBABILITY MASS FUNCTION

Sum of six six-sided dice

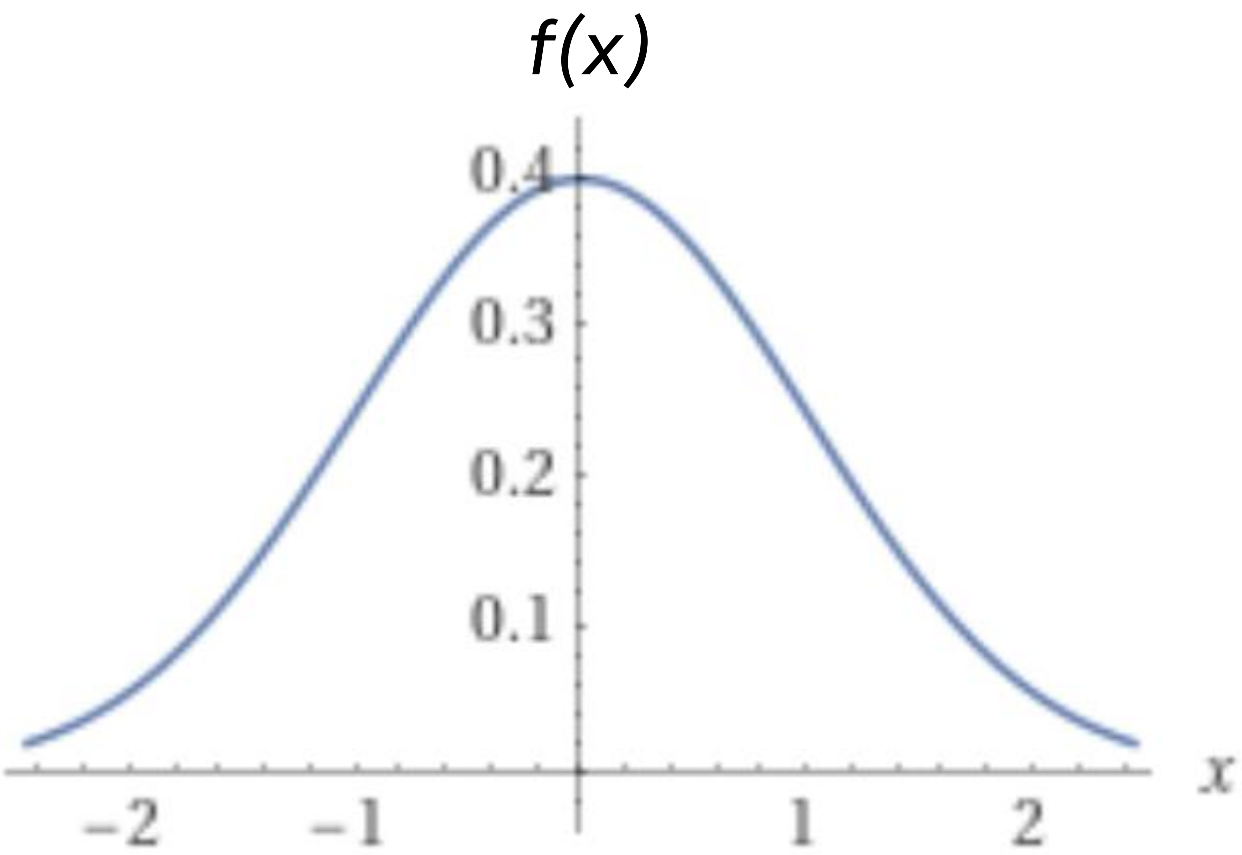


PROBABILITY DENSITY FUNCTION (PDF)

Is the analog of the PMF for continuous random variables.

Ex: Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



CENTRAL LIMIT THEOREM

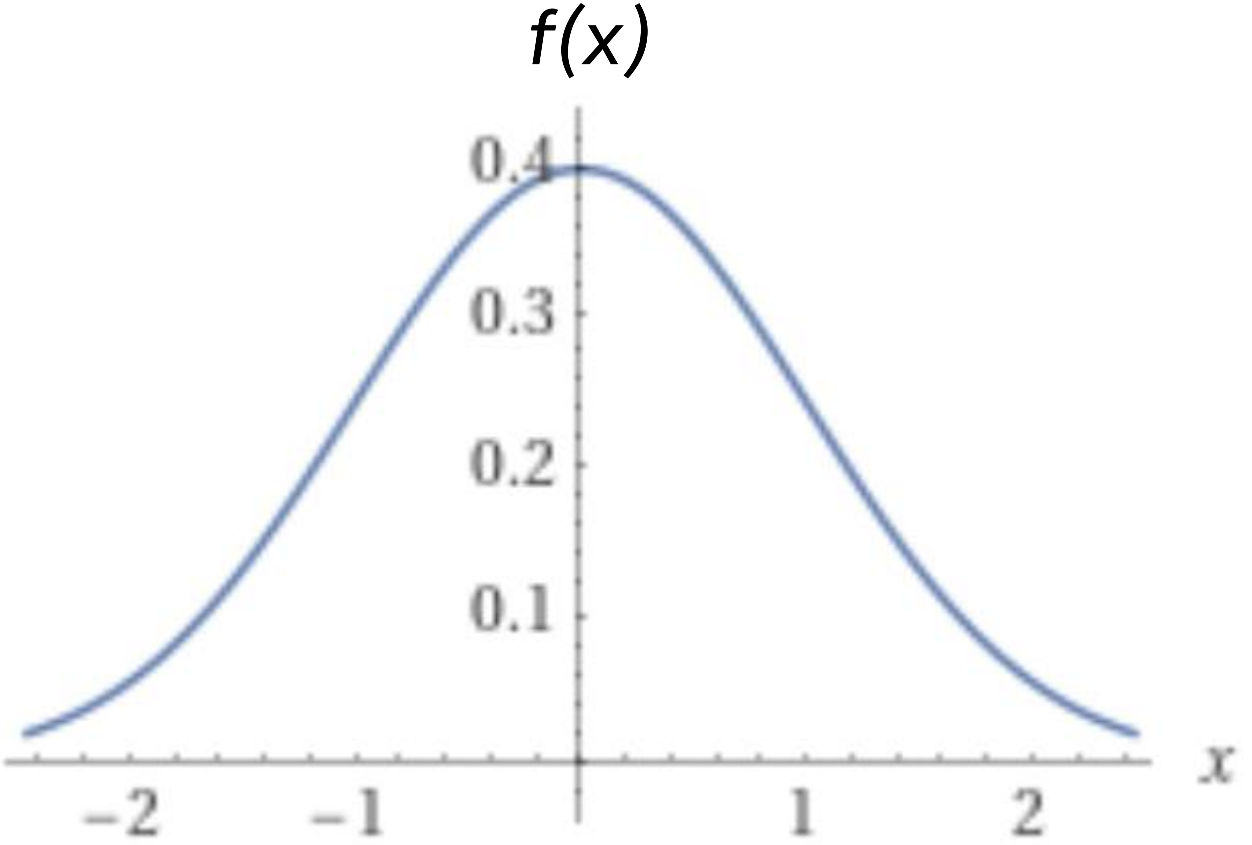
Sum of independent random variables with

- Finite mean
- Finite variance

Tend toward a Gaussian

Gaussian

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

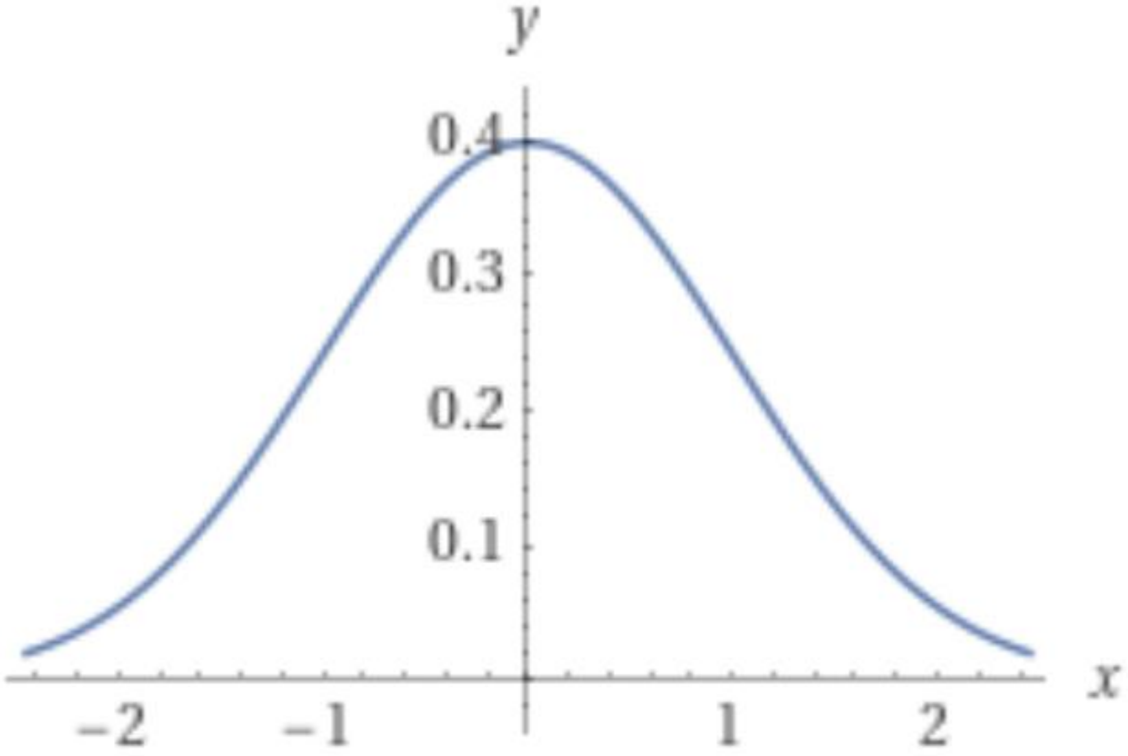


PROBABILITY DENSITY FUNCTION

For all probability density functions (PDFs):

- Function is non-negative for all x.
- The integral over the entire range is 1.

f(x)



THANK YOU

David Harrison

Harrison@cs.olemiss.edu