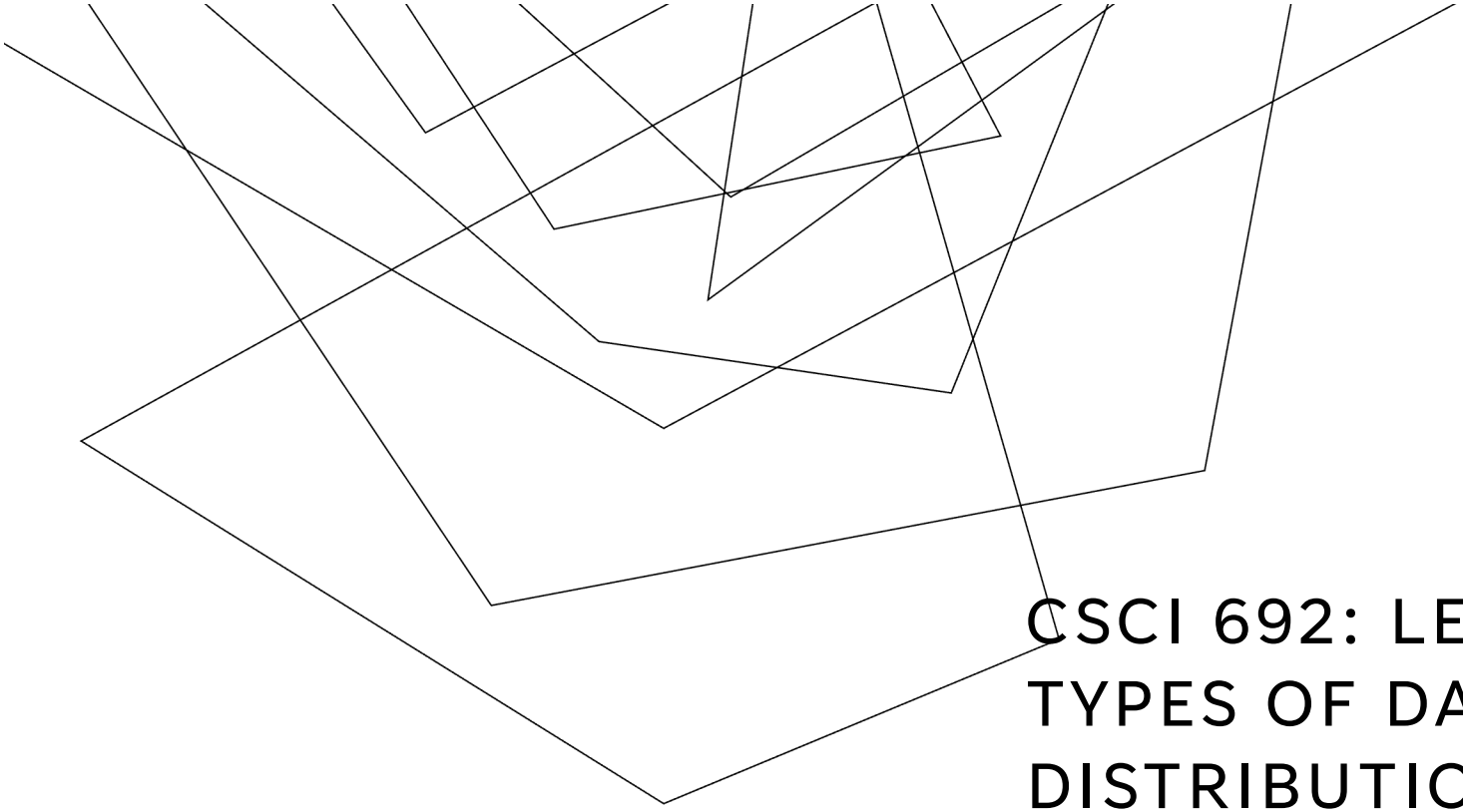


CSCI 692: LECTURE 3 TYPES OF DATA, DISTRIBUTIONS, DISPERSION

Professor David Harrison

OFFICE HOURS

Due to scheduling conflict, office hours updated

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

.

HOMEWORK 1

Some delay in grading because I misconfigured the role of the Teaching Assistant in Blackboard.

I hadn’t given her access to the submissions.

HOMEWORK 2

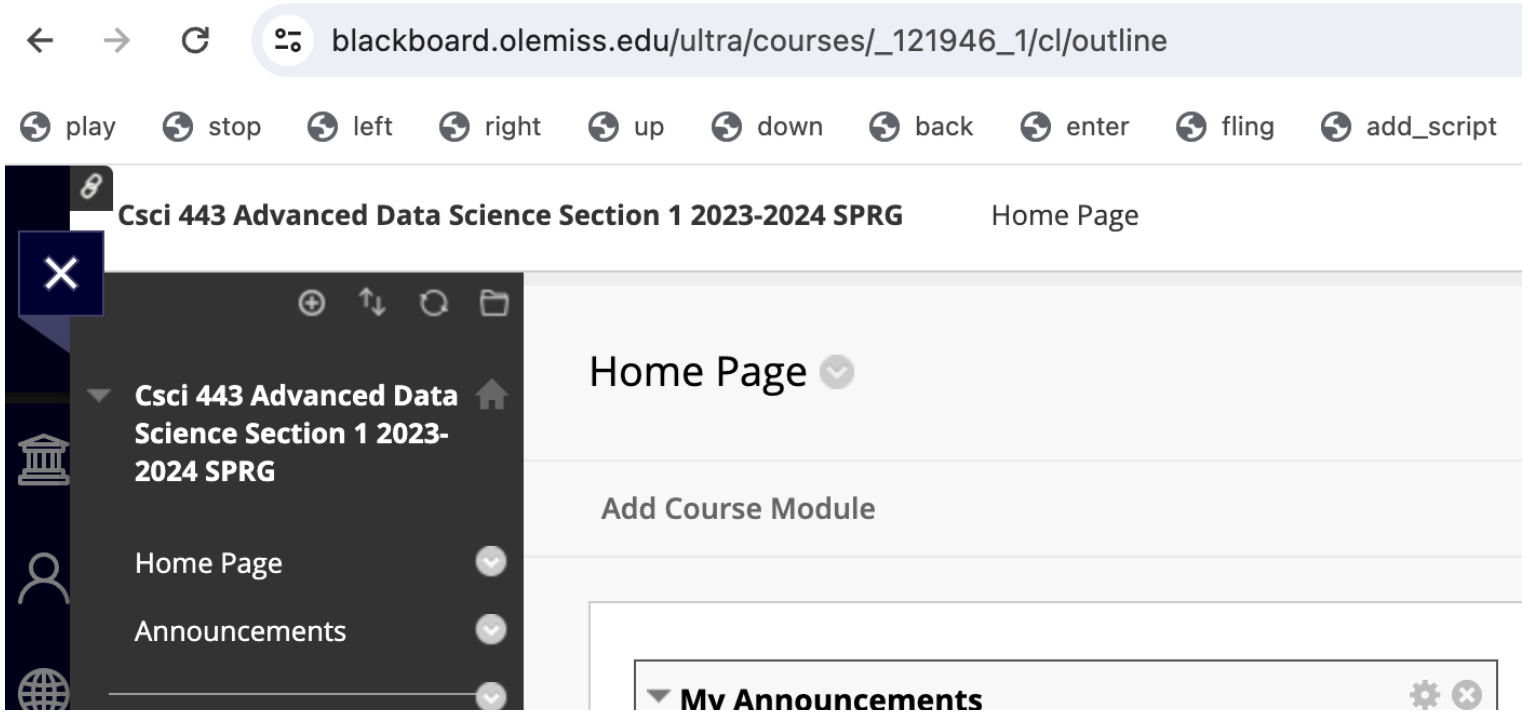
Will be handed out on Thursday and due the following Thursday (February 22)

DATES OF INTEREST

February 8	HW2 handed out
February 15	HW2 due, HW3 handed out
February 22	HW3 due
February 27	Review
February 29	Midterm (must be before progress reports)
March 4	Progress Reports
March 8	Deadline for Withdrawal
March 9-17	Spring Break

BLACKBOARD

Slides up through lecture 3 on blackboard.

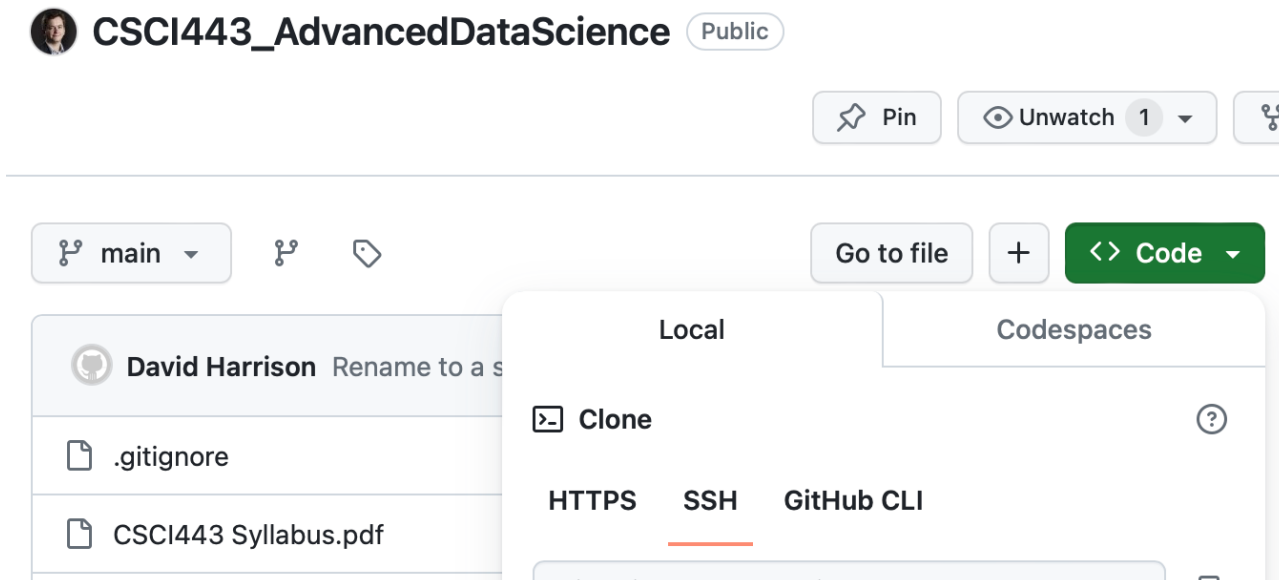


GITHUB

Lecture slides and examples committed to GitHub also up through lecture 3.

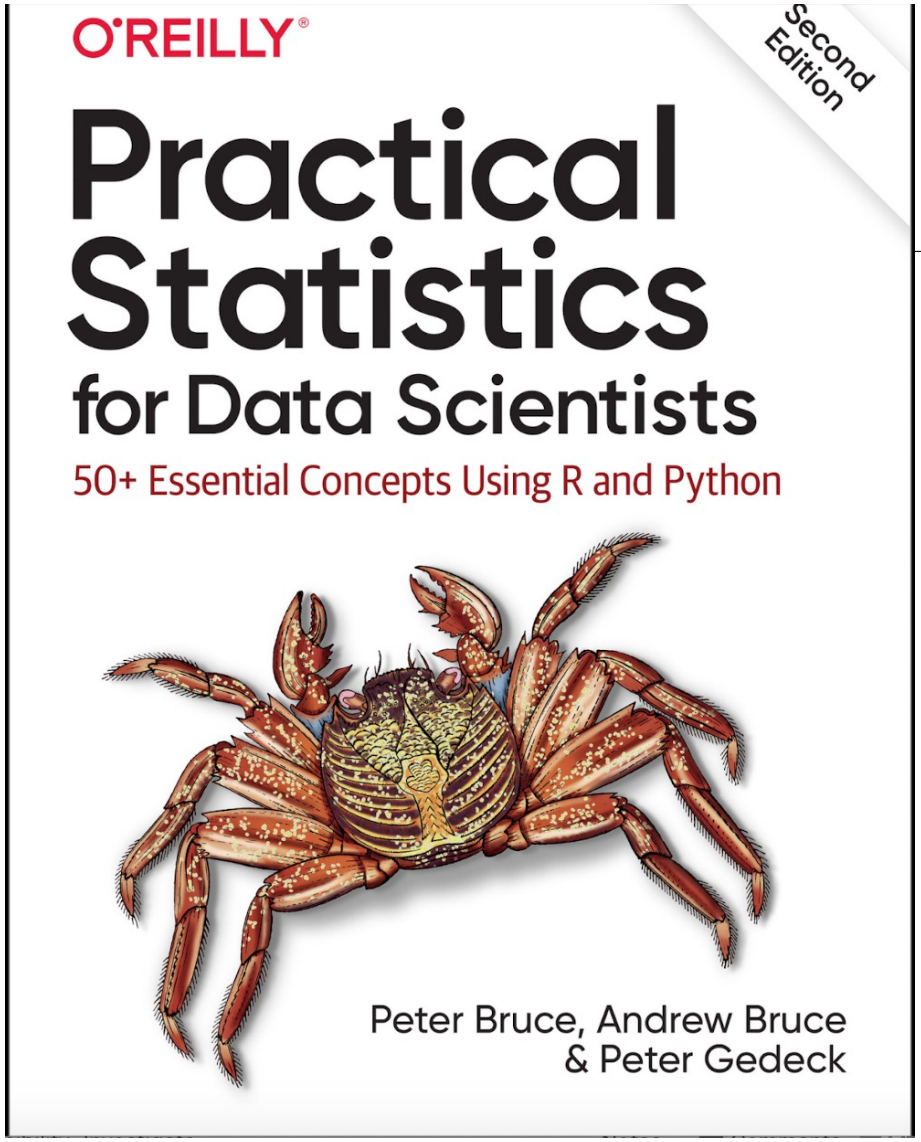
The project is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience



LAST LECTURE

- Introduced some definitions.



LAST LECTURE

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

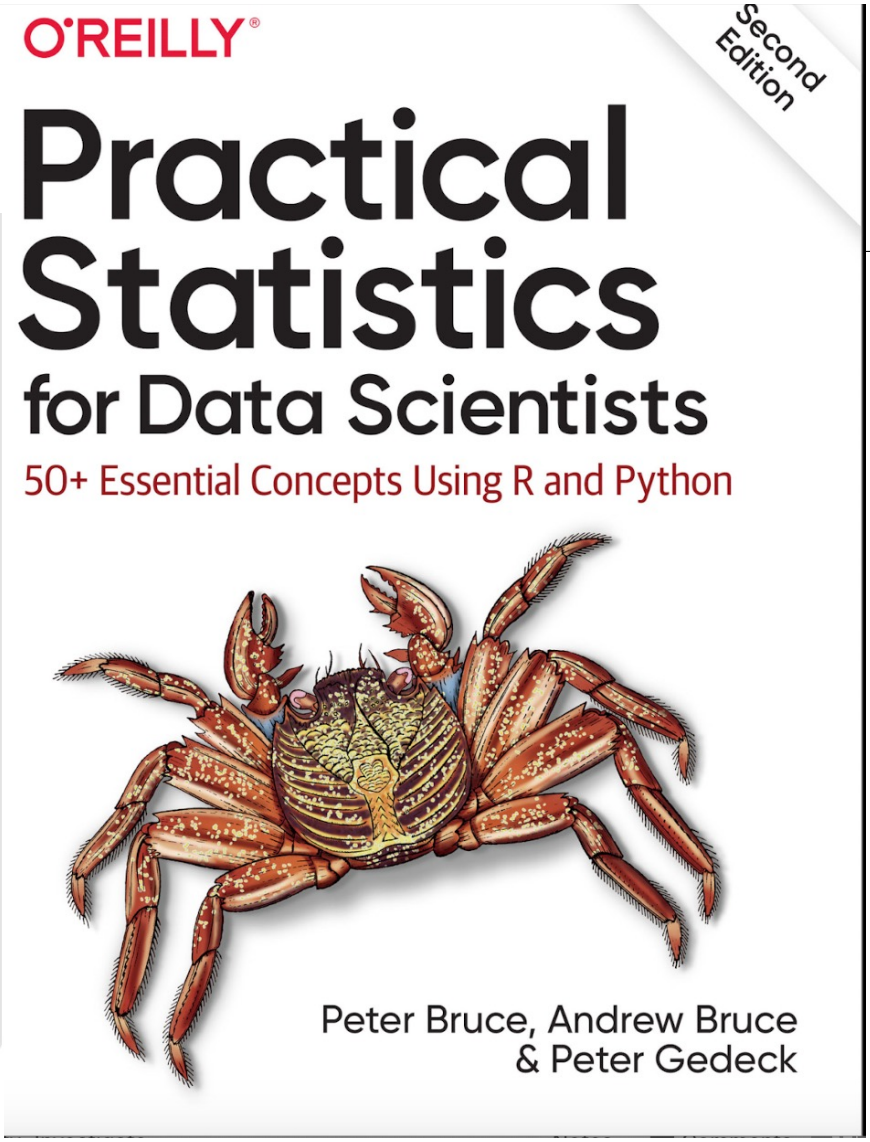
Records

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



LAST LECTURE

Robust

Not sensitive to extreme values.

Synonym

resistant

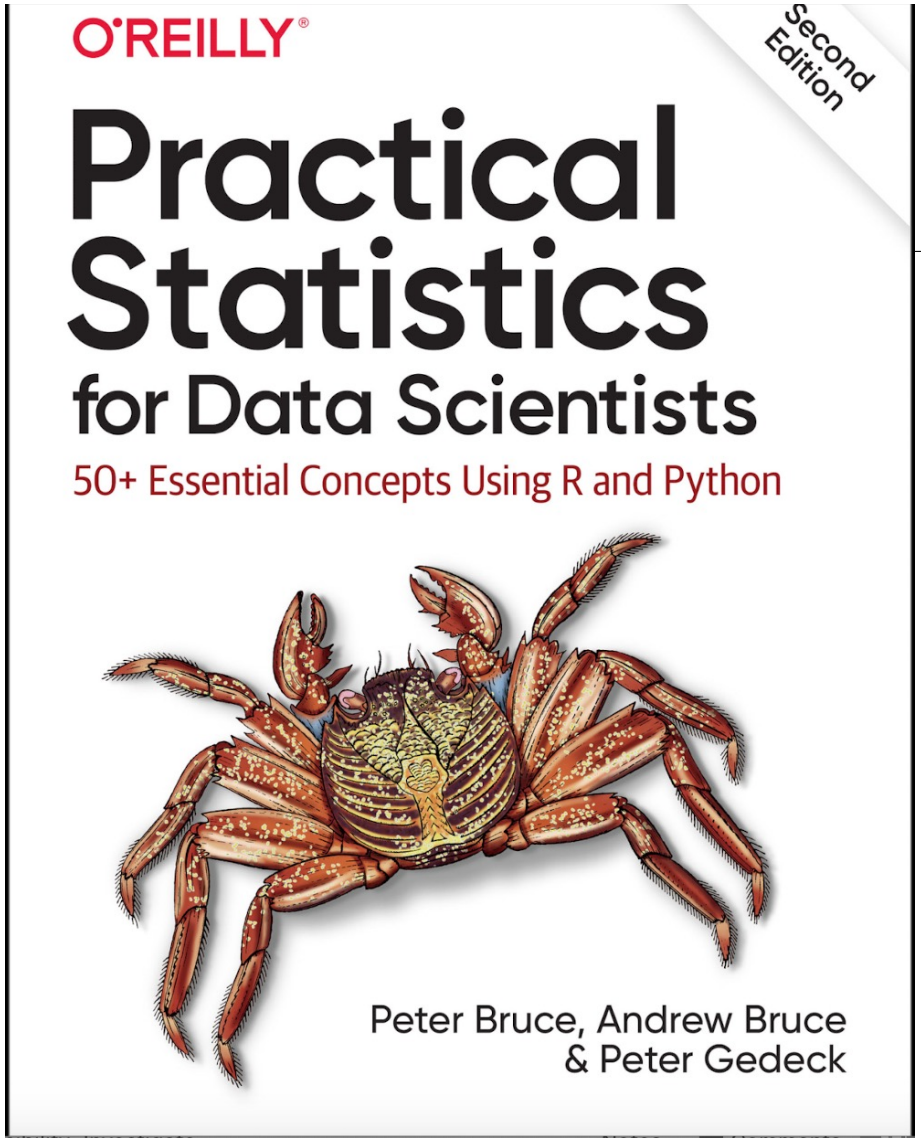
Outlier

A data value that is very different from most of the data.

Synonym

extreme value

Talked about robustness in the context of effects of extreme outliers on mean and median, but I think I neglected to say the word “robust.”



LAST LECTURE

Mean

The sum of all values divided by the number of values.

Synonym

average

Weighted mean

The sum of all values times a weight divided by the sum of the weights.

Synonym

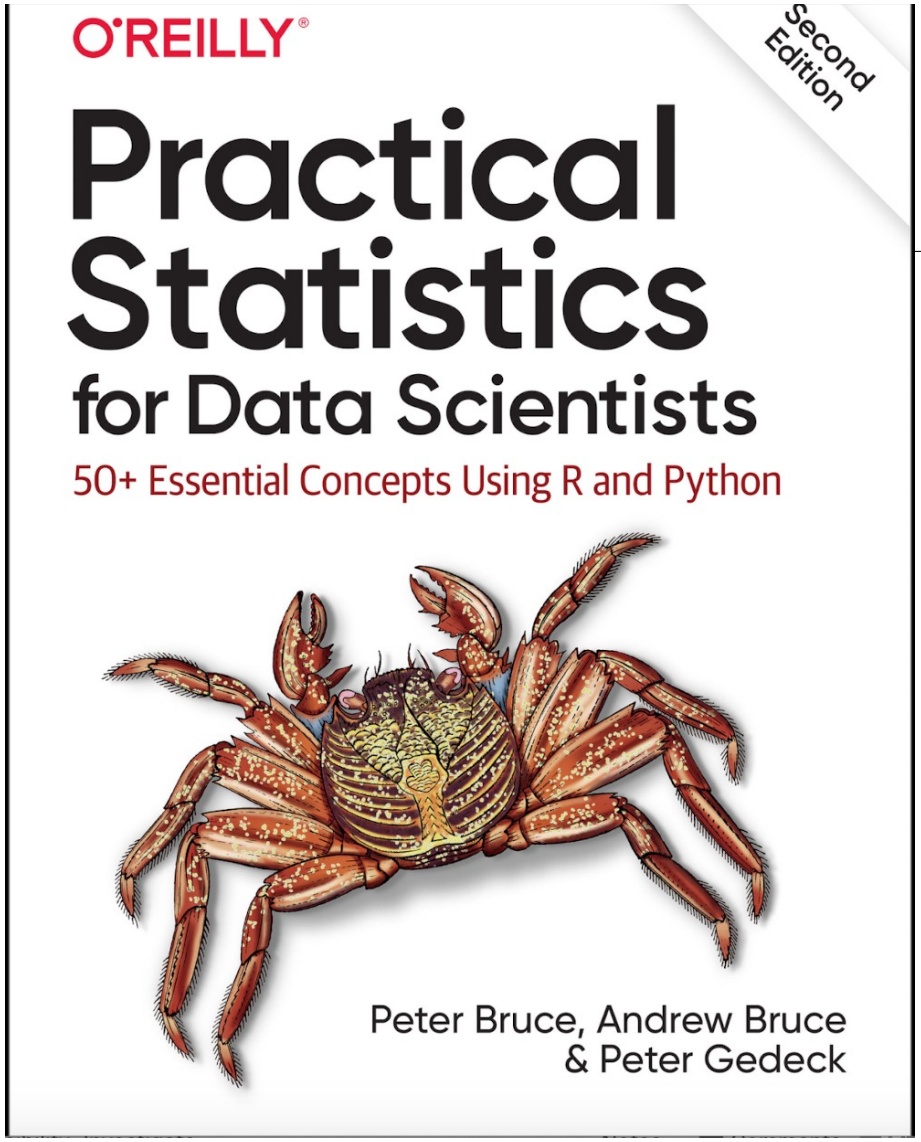
weighted average

Median

The value such that one-half of the data lies above and below.

Synonym

50th percentile

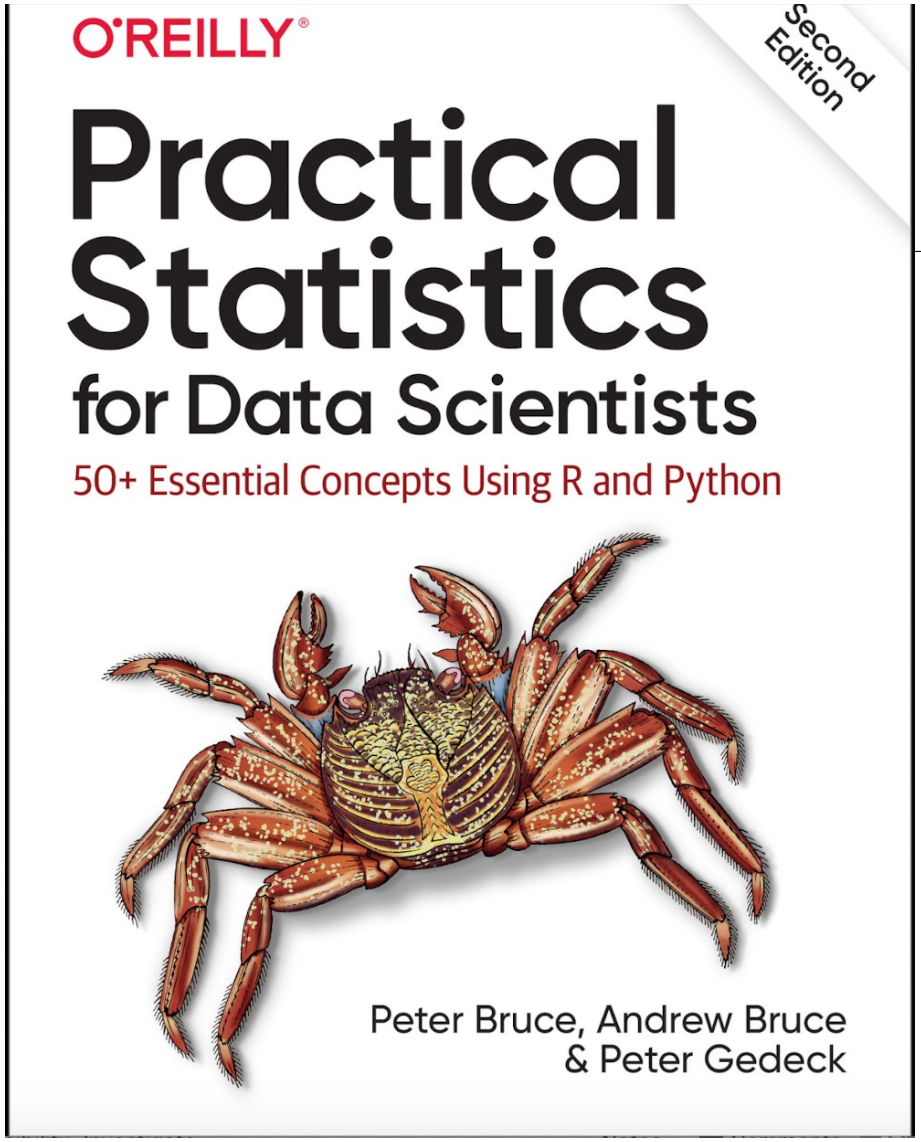


LAST LECTURE

- Types of Error
  - Systematic vs. Random
- Systematic error (Bias)
  - Observer bias
  - Selection bias
  - Measurement bias
  - Confounding factors
- Random error (Noise)
  - Measurement error
  - Heisenberg uncertainty

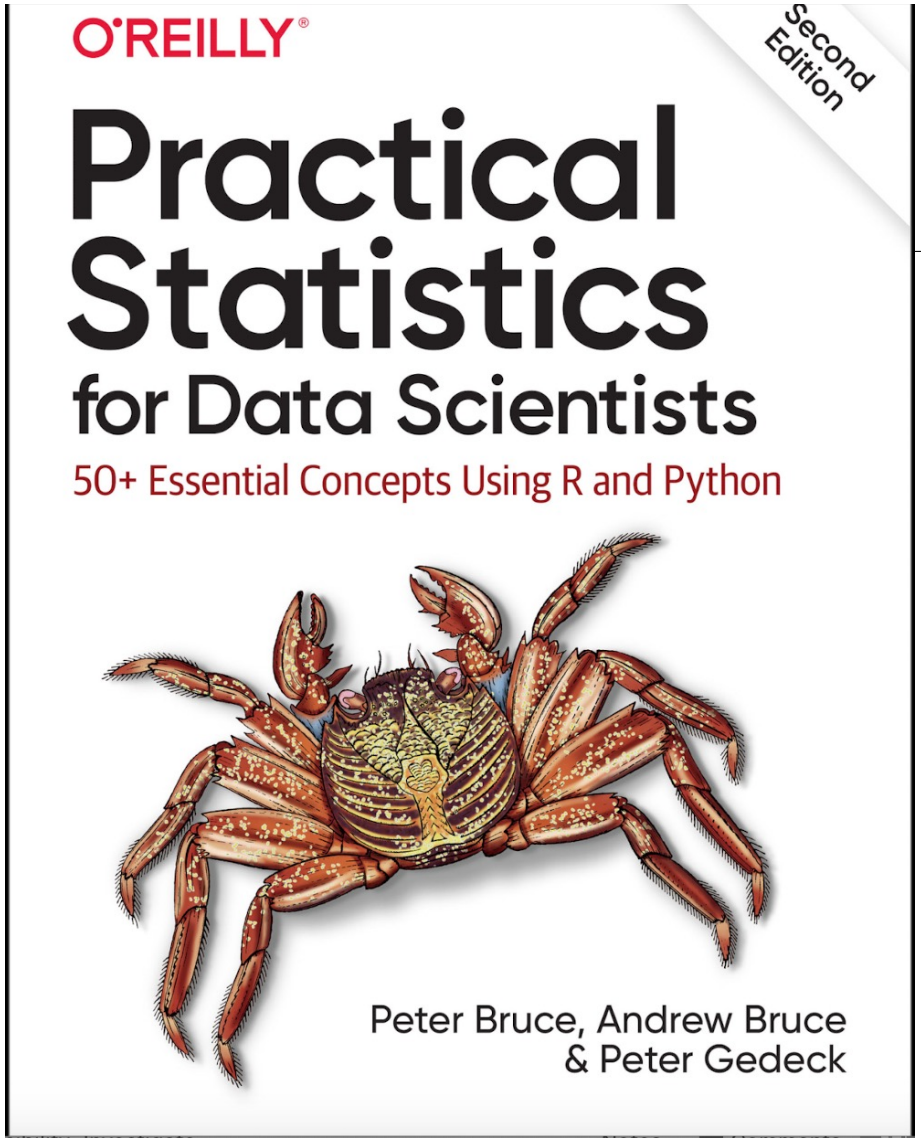
READ ABOUT

- Weighted mean
- Weighted median
- Trimmed mean



TODAY

- Types of Data (1st part of ch 1)
- Examples of
  - Measurement bias
  - Confounding factors
  - Measurement error
- Distributions (1st part of ch 2)
  - Bernoulli
  - Binomial
  - Gaussian
- Dispersion (chapter 1)



TYPES OF DATA

KEY TERMS FOR DATA TYPES

Numeric

Data that are expressed on a numeric scale.

Continuous

Data that can take on any value in an interval. (*Synonyms*: interval, float, numeric)

Discrete

Data that can take on only integer values, such as counts. (*Synonyms*: integer, count)

Categorical

Data that can take on only a specific set of values representing a set of possible categories. (*Synonyms*: enums, enumerated, factors, nominal)

Binary

A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (*Synonyms*: dichotomous, logical, indicator, boolean)

Ordinal

Categorical data that has an explicit ordering. (*Synonym*: ordered factor)

DISCRETE NUMERIC DATA

Discrete:

- Dice rolls
- Students in classroom
- Number of family members
- Number of cars on road

...

CONTINUOUS NUMERIC DATA

Continuous:

- Height
- Weight
- Temperature

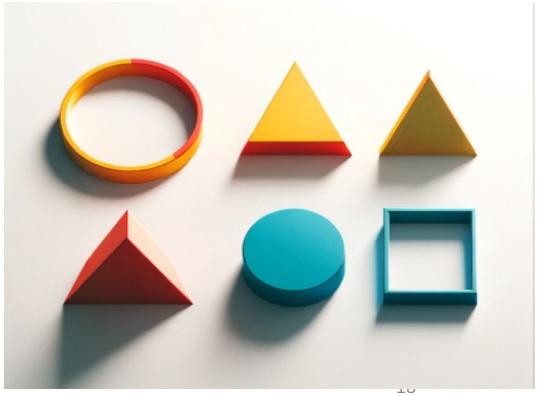
CATEGORICAL DATA

Ordinal

- Education level:
  - High school
  - Undergraduate
  - Graduate
- Rating (1-5 stars)
- Satisfaction
  - Very dissatisfied
  - Dissatisfied
  - Neutral
  - Satisfied
  - Very Satisfied

Non-ordinal

- Types of animals
- Shapes



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.

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.

DISTRIBUTIONS

Wikipedia says,

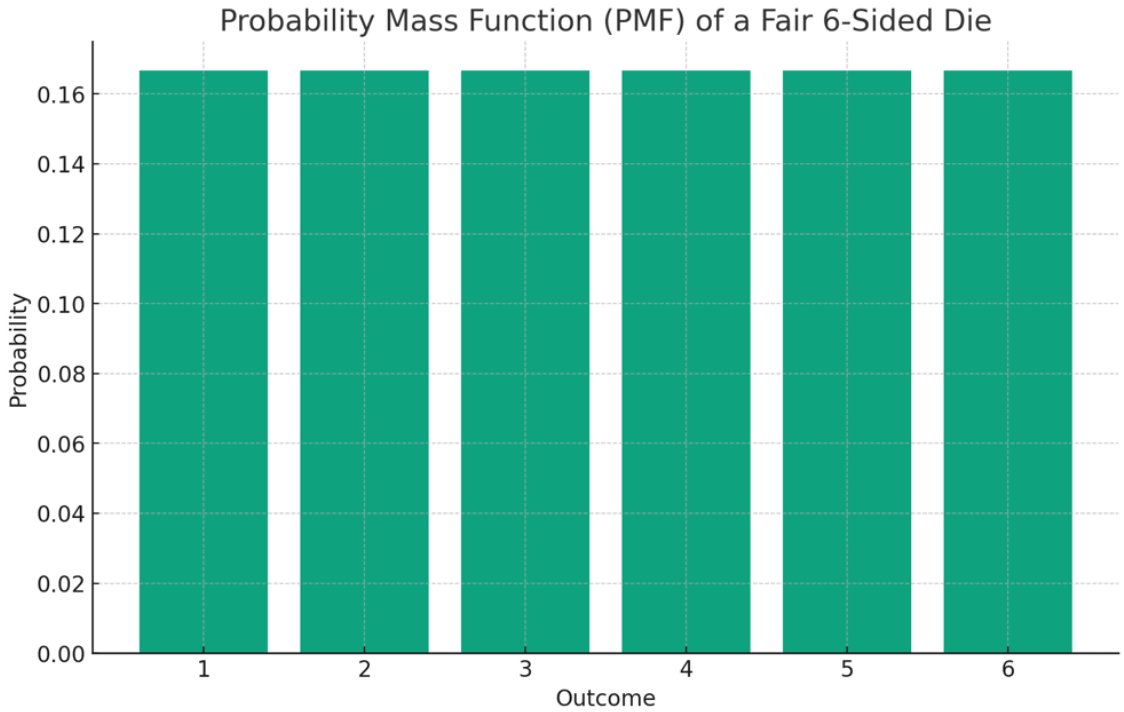
In probability theory and statistics, a **probability distribution** is the mathematical function that gives the probabilities of occurrence of different possible **outcomes** for an experiment.[1][2] It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities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

PROBABILITY MASS FUNCTION

Describes the probability of each discrete outcome.

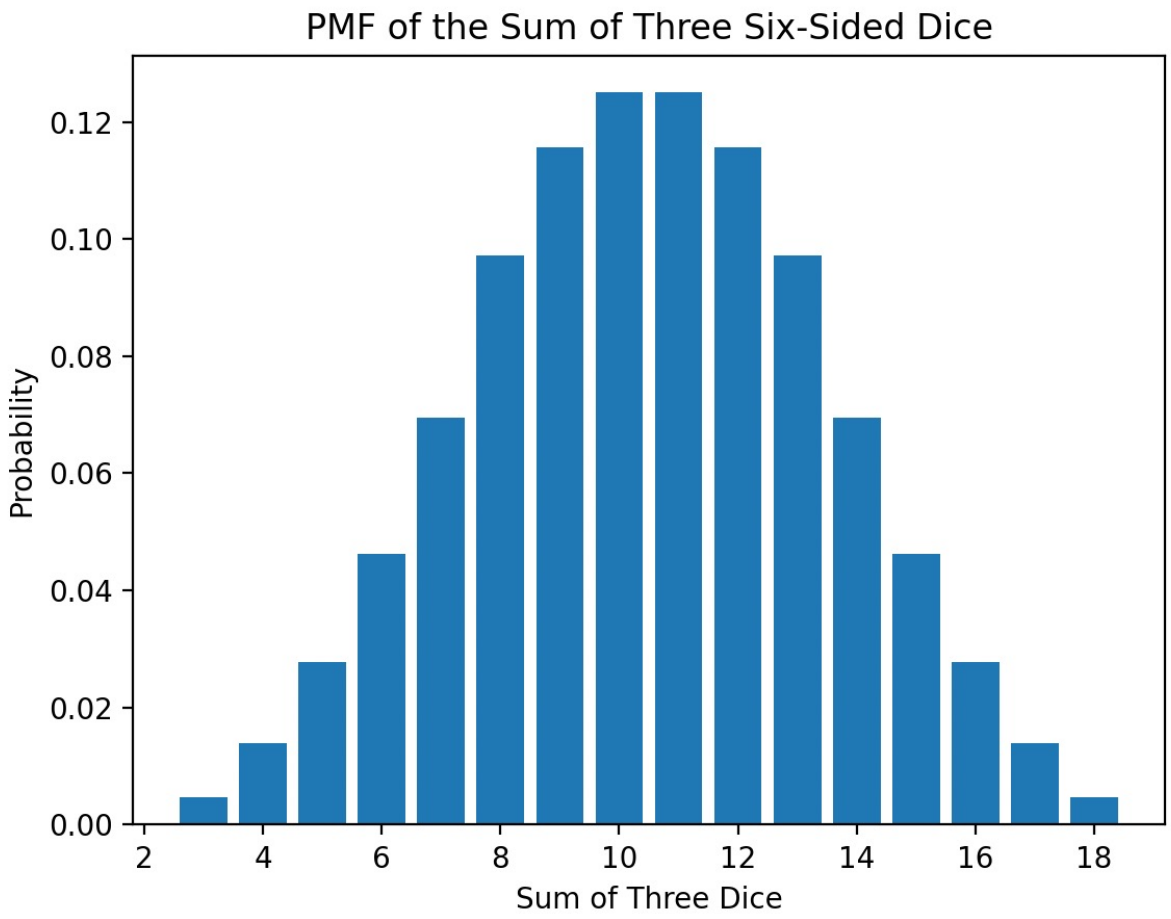
For discrete random variables, a PMF loloks like a histogram where each bin refers to a single outcome.

Sum of probabilities must be 1.



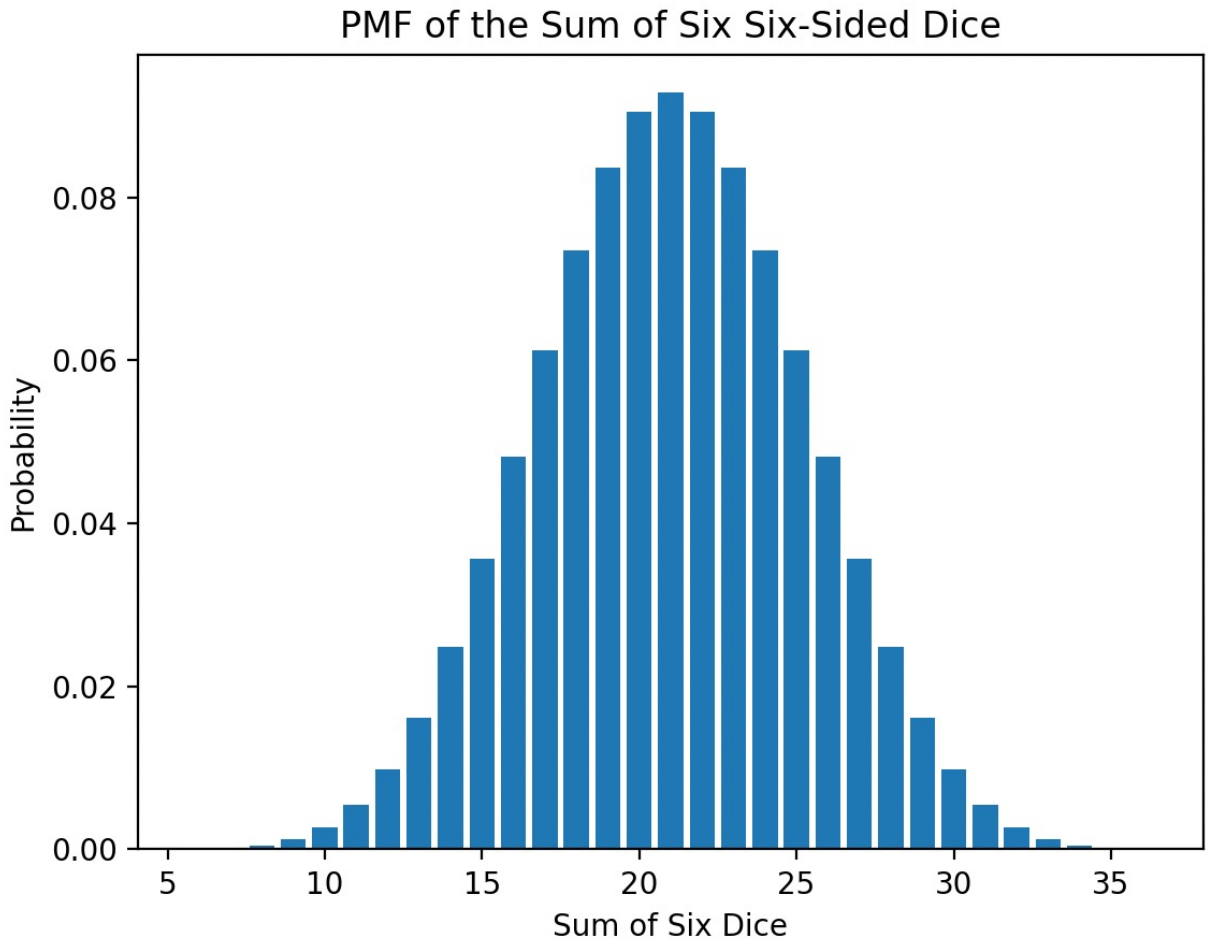
PROBABILITY MASS FUNCTION

Sum of three dice



PROBABILITY MASS FUNCTION

Sum of six six-sided dice

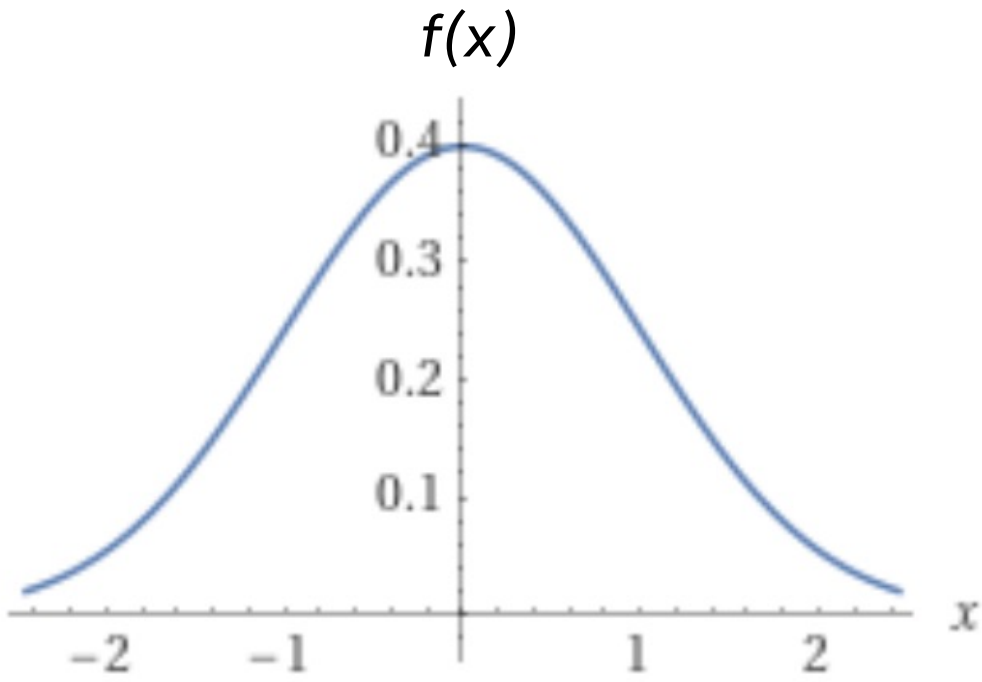


PROBABILITY DENSITY FUNCTION (PDF)

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

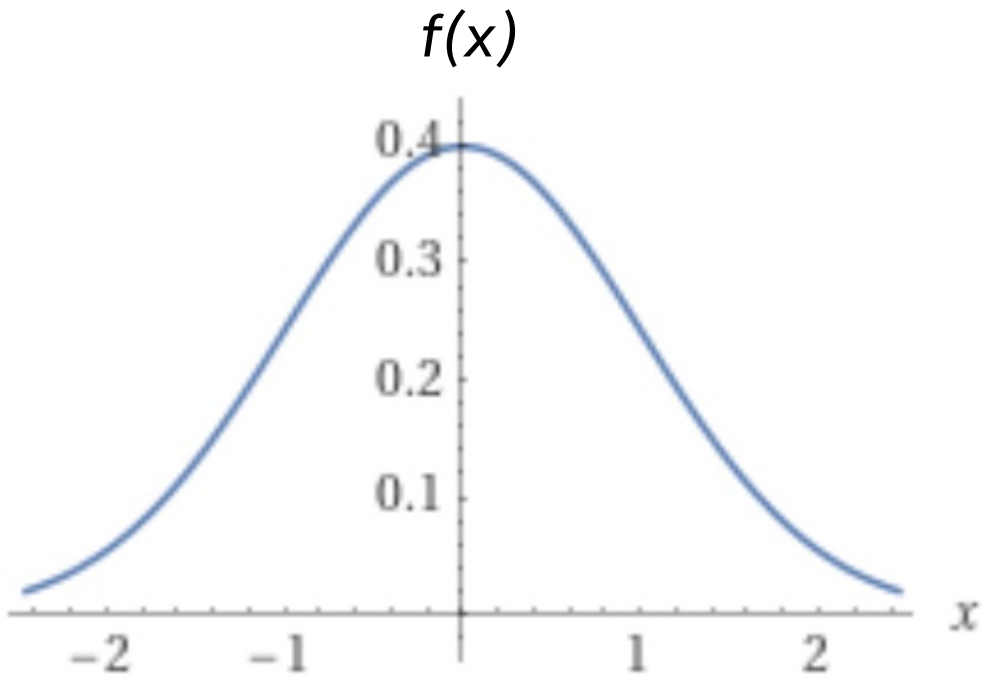
Sum of independent random variables with

- Finite mean
- Finite variance

Tend toward a Gaussian

Gaussian

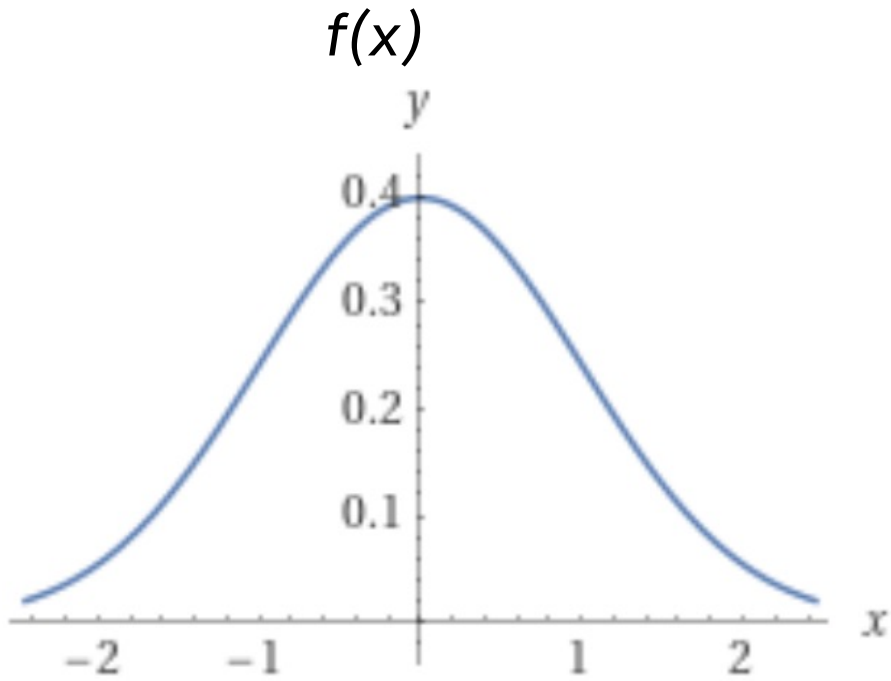
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



PROBABILITY DENSITY FUNCTION

For all probability density functions (PDFs):

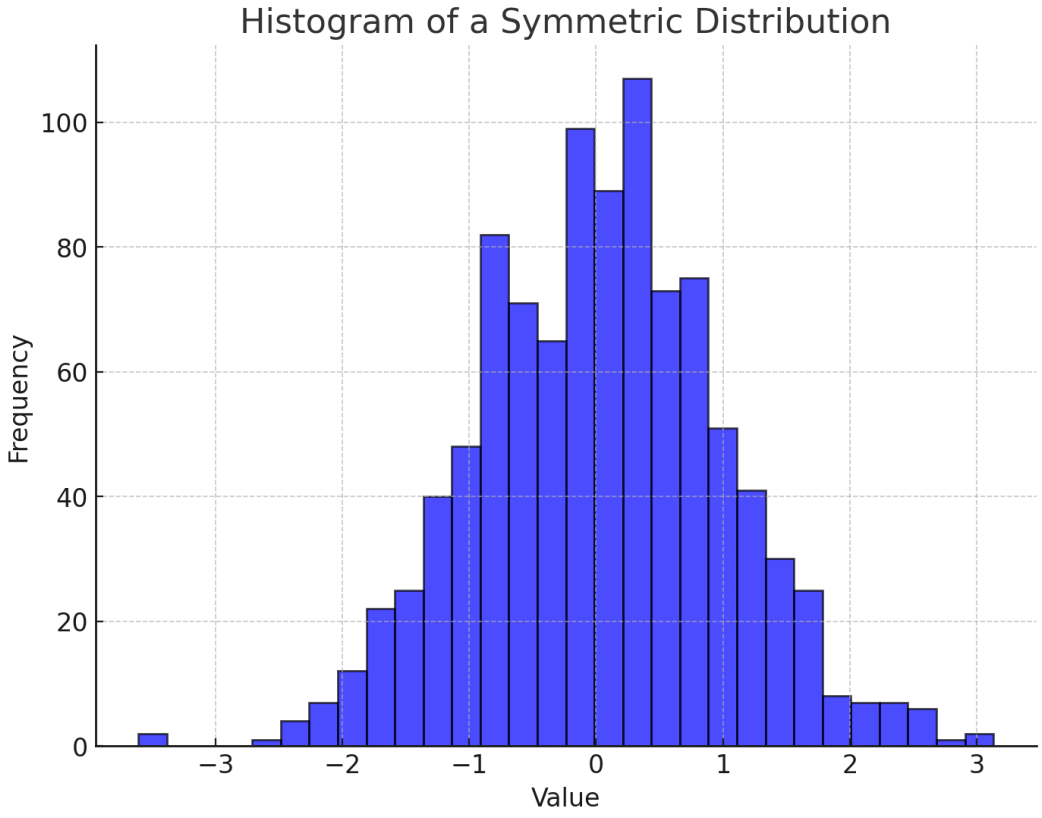
- Function is non-negative for all x.
- The integral over the entire range is 1.



UNDERLYING DISTRIBUTION

In Data Science, we usually do not know the underlying distribution. We instead deal with samples

We do not know the PDF, but we can visualize a continuous Random Variabe (R.V.) using a histogram.

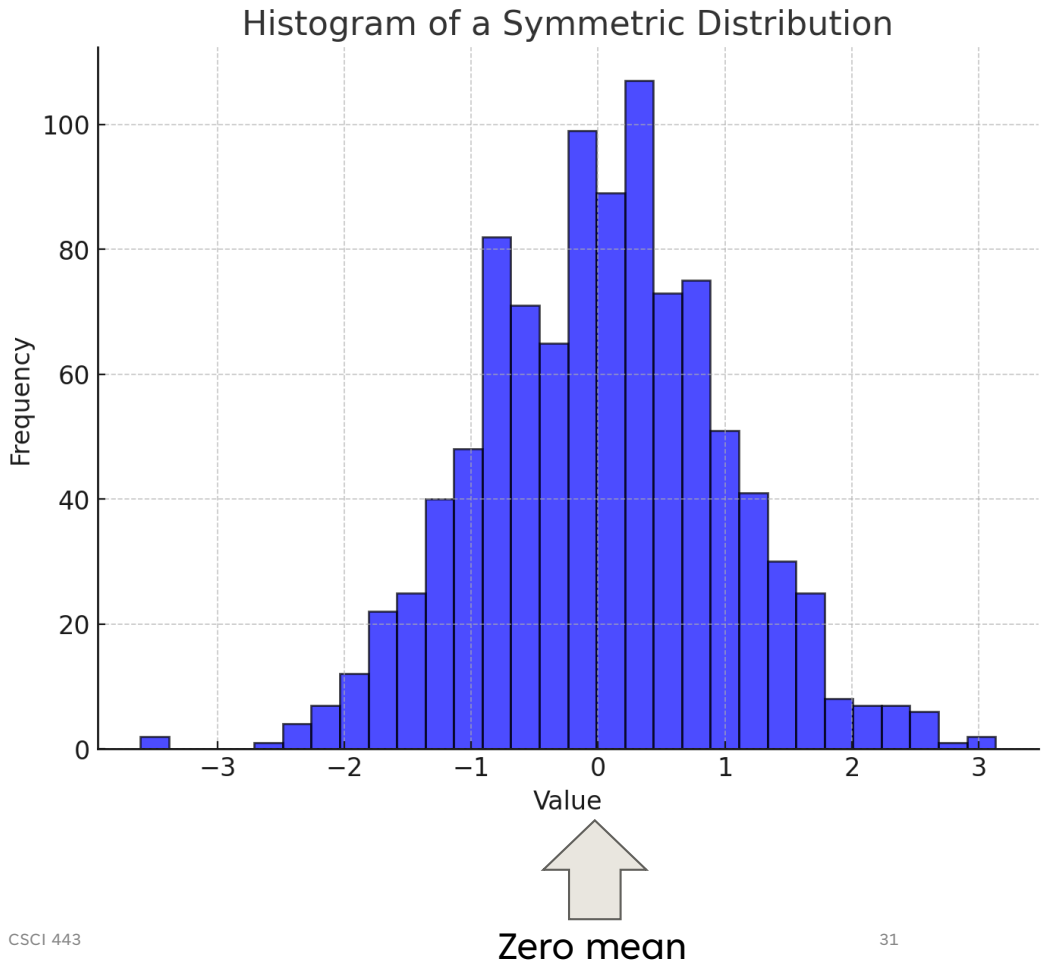


STATISTICS FROM SAMPLES

In Data Science, we estimate properties of the distribution using statistics.

For central tendency:

- mean
- median
- ...



DISPERSION

We estimate variability (a.k.a., dispersion) using a variety of metrics

(chapter 1)

Deviations

The difference between the observed values and the estimate of location.

Synonyms

errors, residuals

Variance

The sum of squared deviations from the mean divided by $n - 1$ where n is the number of data values.

Synonym

mean-squared-error

Standard deviation

The square root of the variance.

Mean absolute deviation

The mean of the absolute values of the deviations from the mean.

Synonyms

l1-norm, Manhattan norm

Median absolute deviation from the median

The median of the absolute values of the deviations from the median.

DISPERSION

We estimate variability (a.k.a., dispersion) using a variety of metrics

(chapter 1)

Range

The difference between the largest and the smallest value in a data set.

Order statistics

Metrics based on the data values sorted from smallest to biggest.

Synonym

ranks

Percentile

The value such that *P* percent of the values take on this value or less and (100–P) percent take on this value or more.

Synonym

quantile

Interquartile range

The difference between the 75th percentile and the 25th percentile.

MEAN DEVIATION

We estimate variability (a.k.a., dispersion) using a variety of metrics

(chapter 1)

Range

The difference between the largest and the smallest value in a data set.

Order statistics

Metrics based on the data values sorted from smallest to biggest.

Synonym

ranks

Percentile

The value such that *P* percent of the values take on this value or less and (100–P) percent take on this value or more.

Synonym

quantile

Interquartile range

The difference between the 75th percentile and the 25th percentile.

STANDARD DEVIATION AND GAUSSIAN DISTRIBUTION

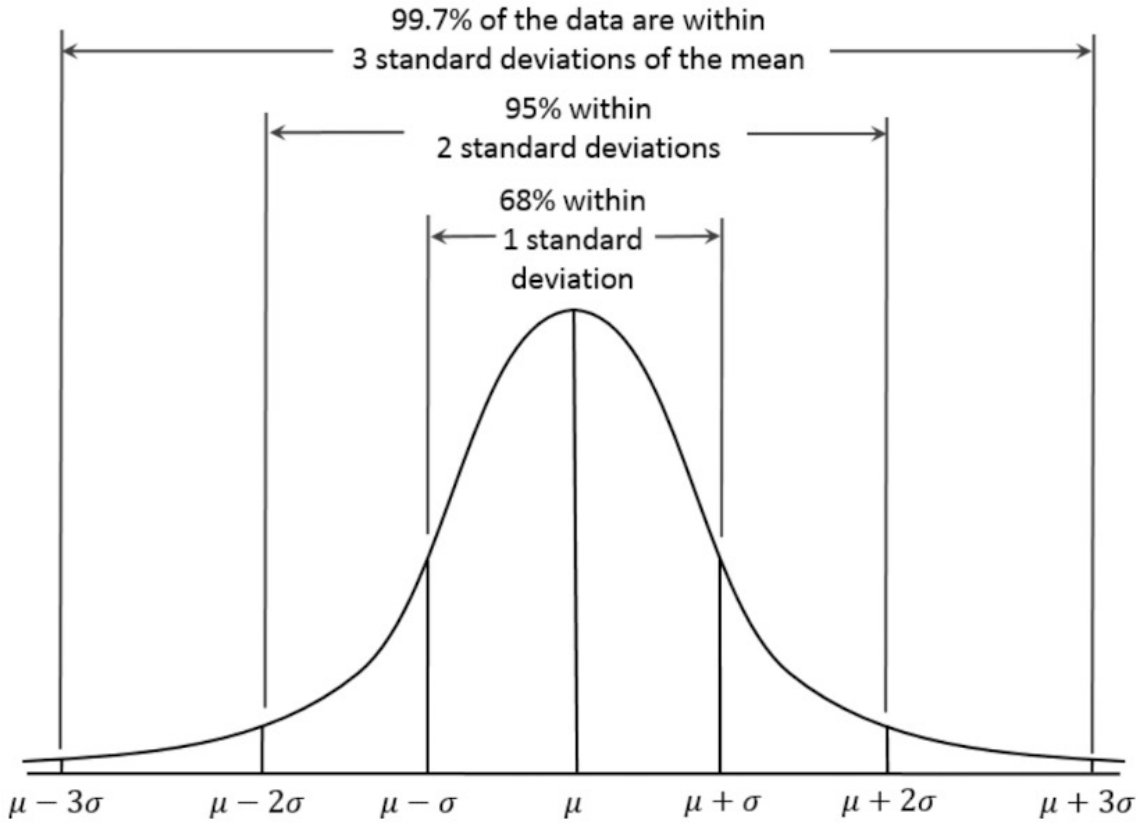
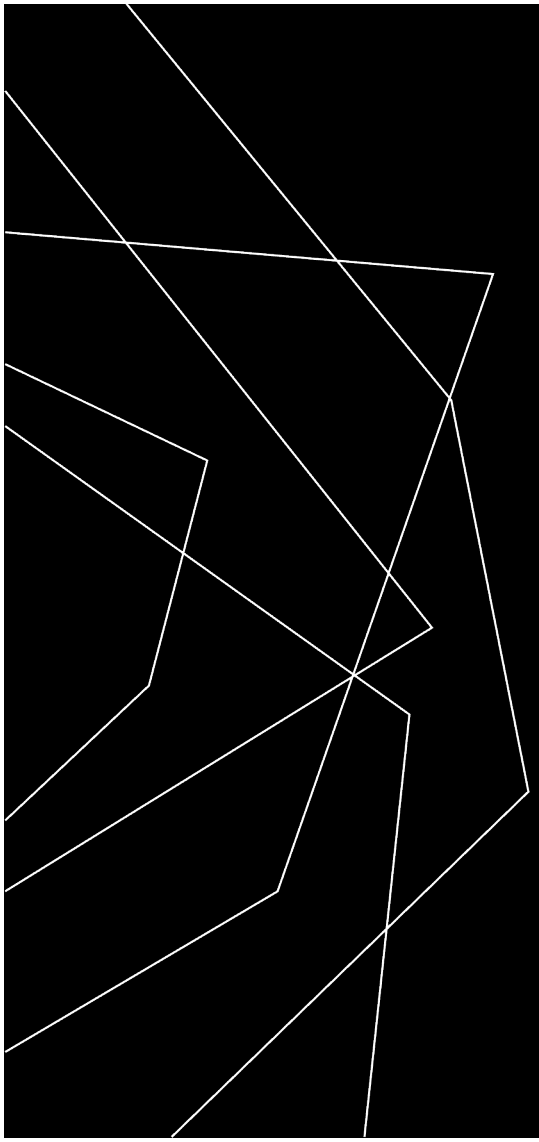


Figure 2-10. Normal curve



THANK YOU

David Harrison

Harrison@cs.olemiss.edu