LECTURE 2:
DATA COLLECTION
(WHAT TO MEASURE, AND HOW TO MEASURE IT)

Experimental Methods 1, E2019
BSc in Cognitive Science, Aarhus University
Wednesday 11/09/2019
Fabio Trecca

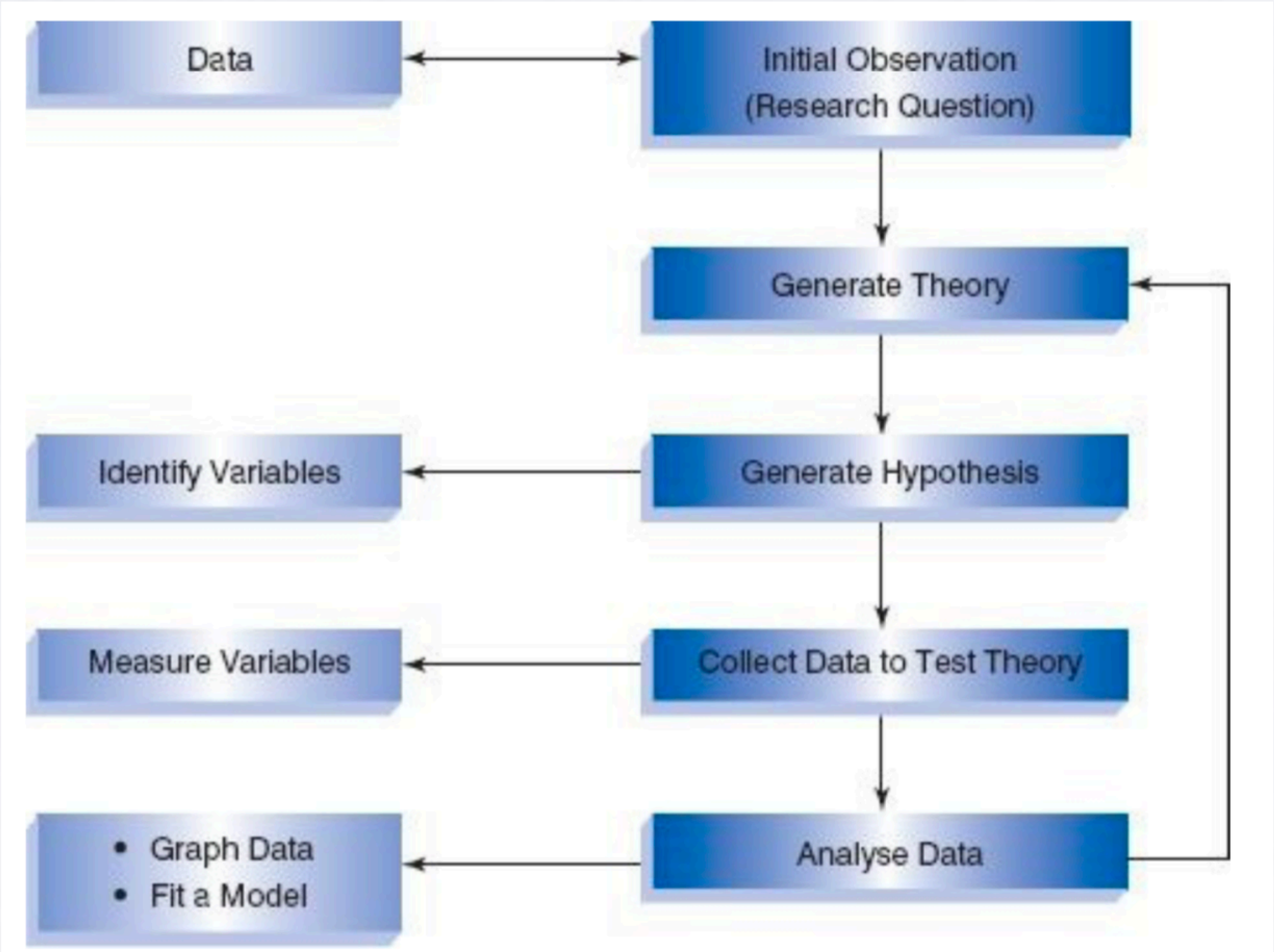
WHAT DO I NEED TO ANSWER A RESEARCH QUESTION?

- Data (1 data point = 1 individual observation)
- Explanation of the data

RECAP

- The study of human cognition is interdisciplinary
- It must rely on insights from many disciplines
- It combines 1st, 2nd, and 3rd person methods
- However, the word "Cognitive Science" reflects specifically the use of quantitative/experimental methods that characterize much of the discipline

THE RESEARCH PROCESS



Correlational vs. Experimental research

Between- vs. Within-subject design

Measurement error

Unsystematic vs. systematic variation

HYPOTHESIS

- H_0 (null hypothesis) = No difference between the means
- H_1 (alternative hypothesis) = Difference between the means
- Null hypothesis significance testing (NHST): we can't prove the H_1, but we can reject the H_0

WHY DO WE NEED STATS?

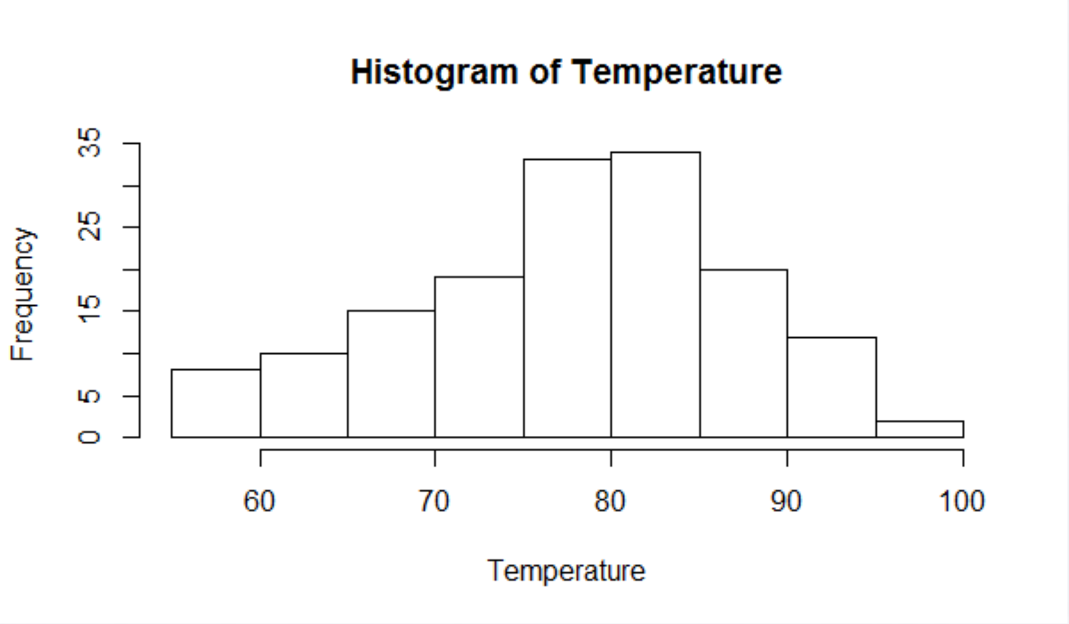
- To discern systematic variation from unsystematic variation

VARIABLES

- Categorical
  - Binary/Logical (frequency)
  - Nominal (frequency)
  - Ordinal (frequency + order)
- Continuous
  - Interval (full arithmetic)
  - Ratio (full arithmetic)

DATA ANALYSIS

- Plot the data (= frequency distribution, e.g., histograms)
  - what is the frequency with which certain values of my variables occur in relation to others?

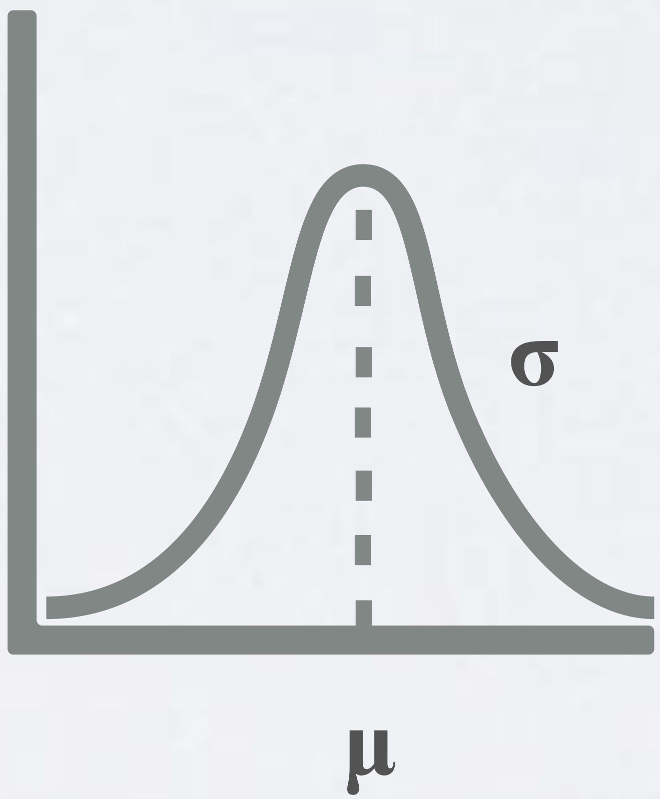


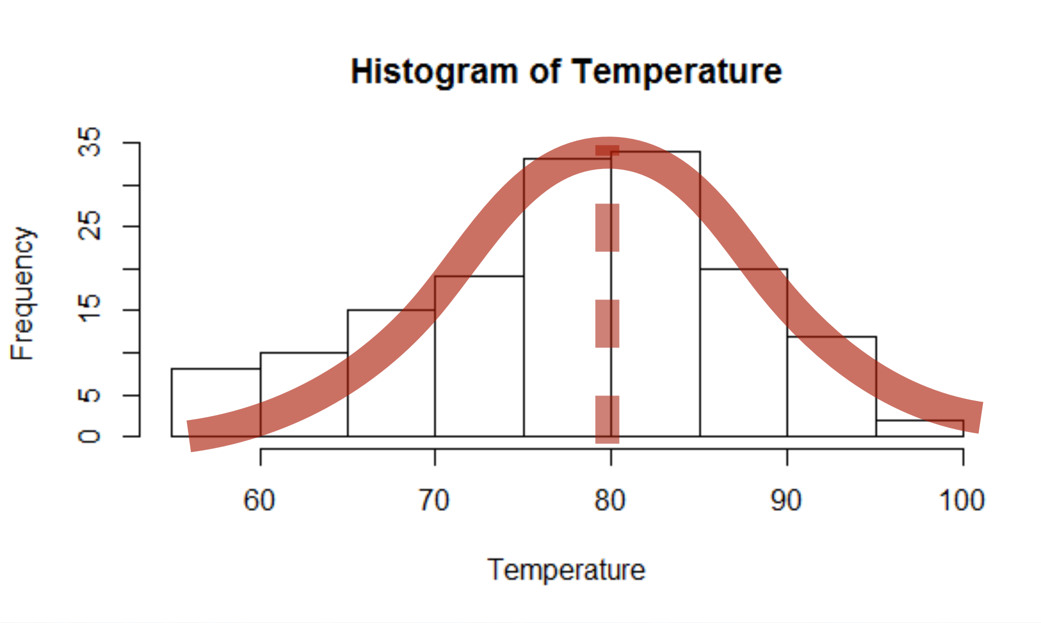
- Fit models (e.g., mean, correlation, linear regression)
  - what is the best way to summarise the raw data?

$\mu = ?, \sigma = ?$

THE NORMAL (FREQUENCY) DISTRIBUTION

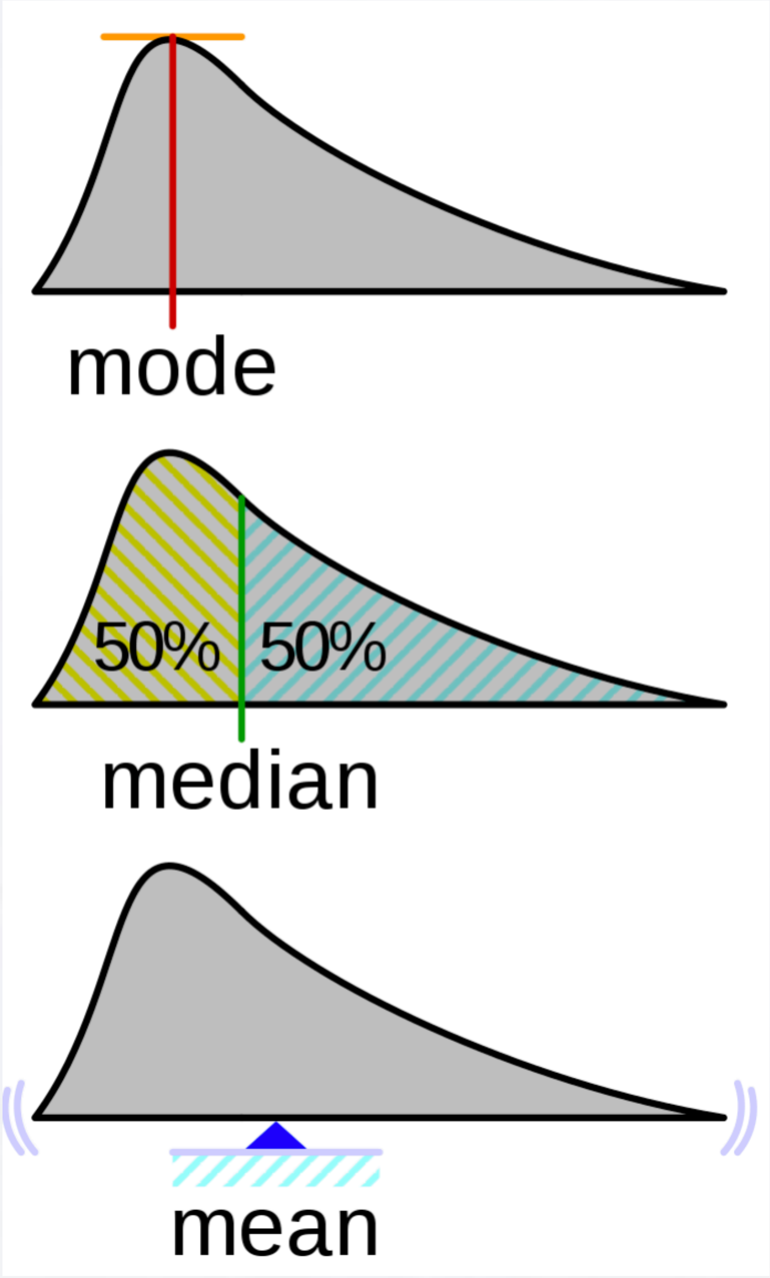
- A.k.a. Gaussian distribution, bell curve
- Symmetrical gravitation toward the mean with decreasing N of data points as we approach the tails
- Many cognitive and behavioural processes are normally distributed
- Defined by two parameters: mean (μ) and standard deviation (σ)
- Results from sum of independent events/ factors





MEASURES OF CENTRAL TENDENCY

- mode
- median
- mean
- In normal distribution: mode=median=mean



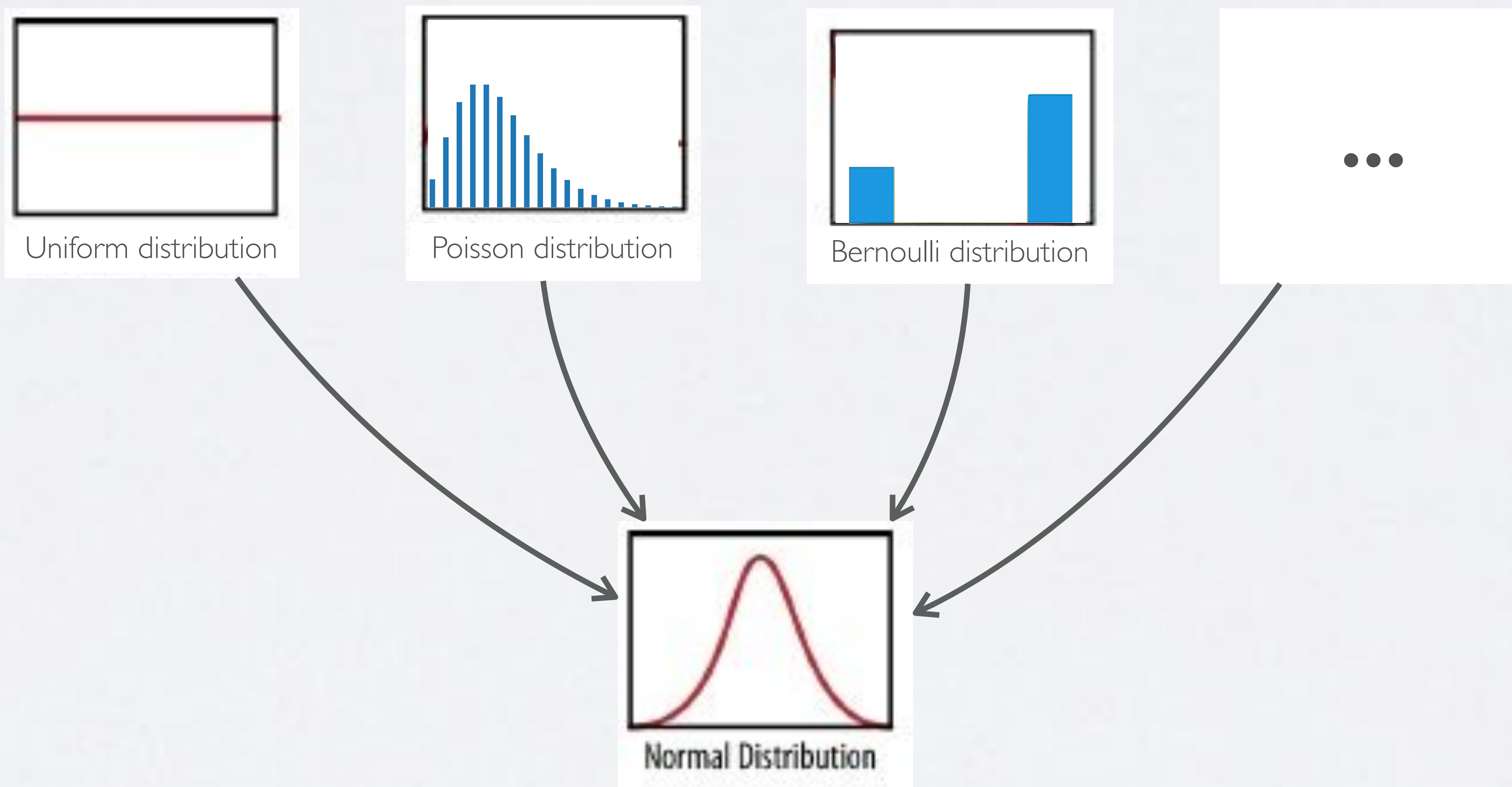
FREQUENCY DISTRIBUTION VS. PROBABILITY DISTRIBUTION

- Two ways of thinking about the same thing:
  - frequency distribution tells me something about the data I have
  - probability distribution allows me to use the data I have to predict the distribution of new data points

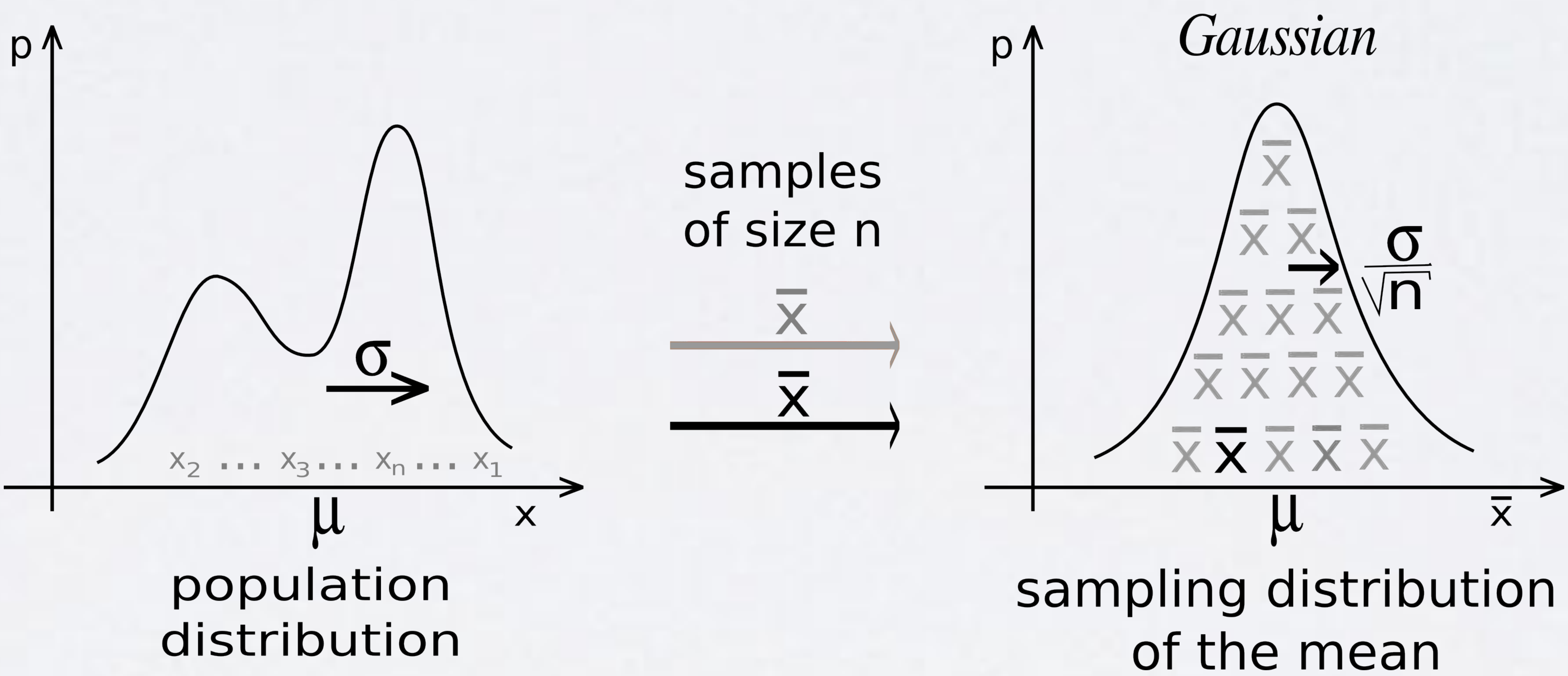
CENTRAL LIMIT THEOREM (1)

- Given a dataset with unknown underlying distribution, the sample means will approximate the normal distribution
- Samples should be of sufficient size

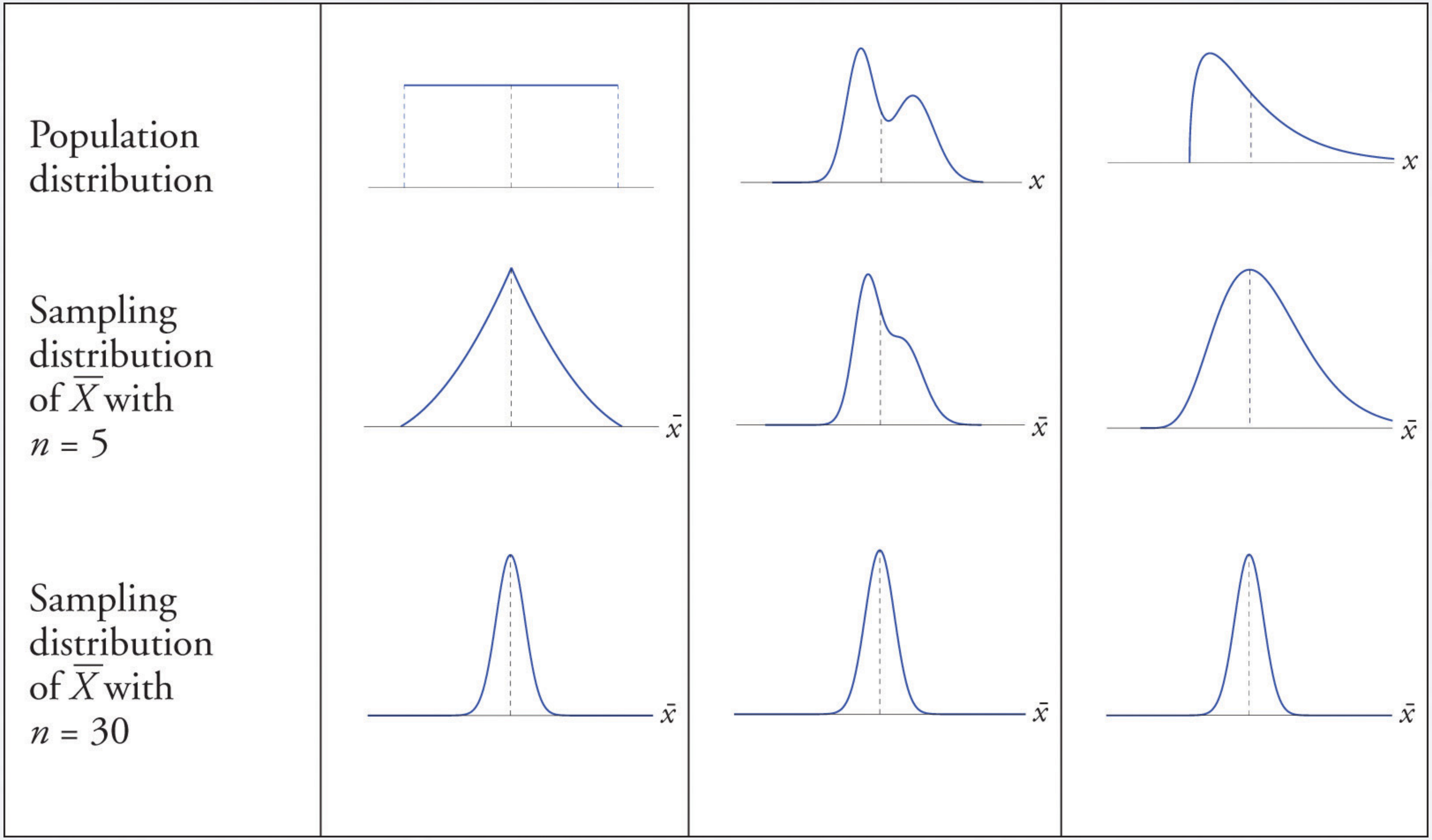
CENTRAL LIMIT THEOREM (2)



CENTRAL LIMIT THEOREM (3)



SAMPLING DISTRIBUTION OF SAMPLE MEANS



https://saylordotorg.github.io/text_introductory-statistics/s10-02-the-sampling-distribution-of-t.html

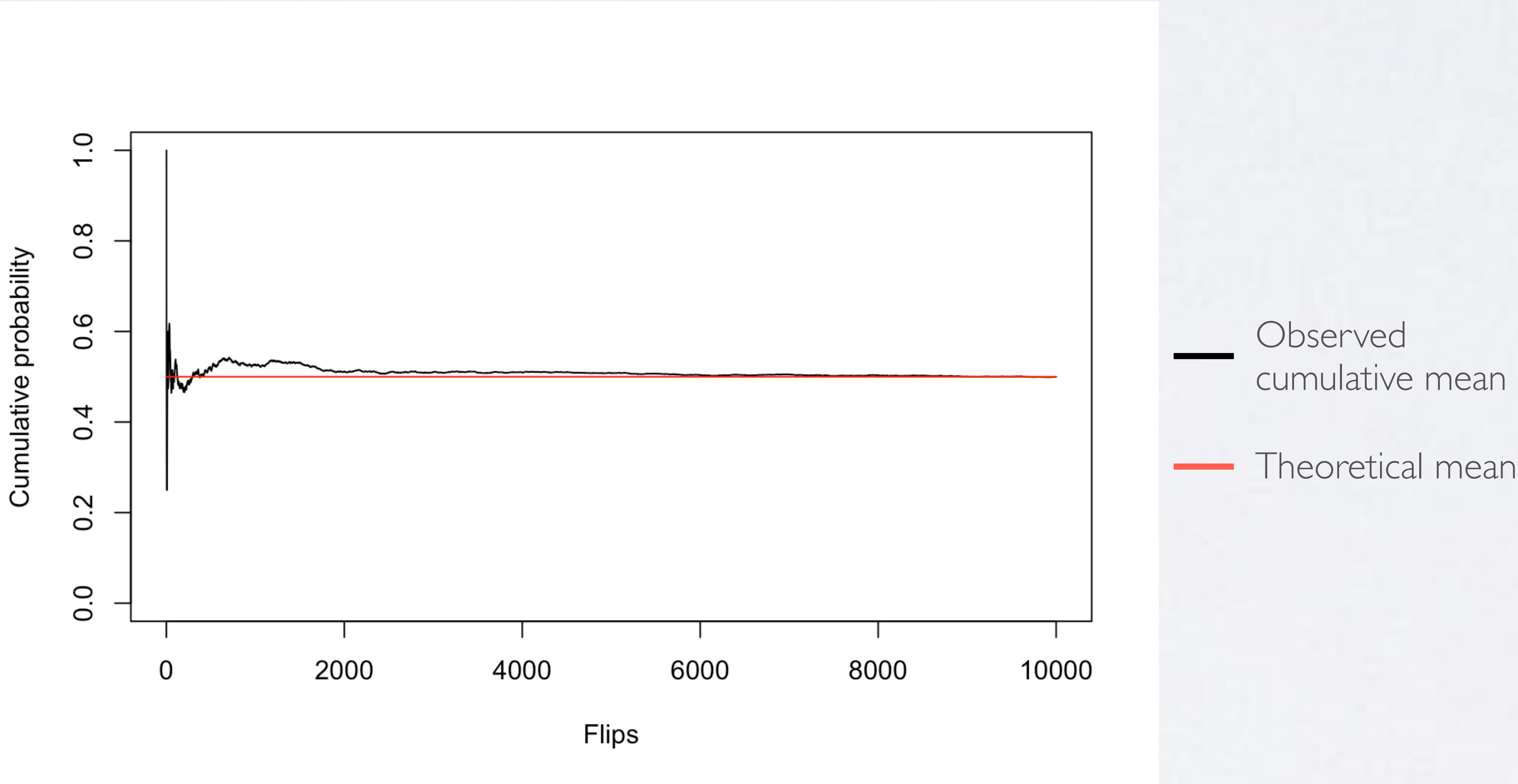
LAW OF LARGE NUMBERS (1)

- $\theta = 0.5$

- Average results obtained from a large number of trials will tend to become closer to the expected value as more trials are performed

- Observed probability approaching the theoretical probability

LAW OF LARGE NUMBERS (2)

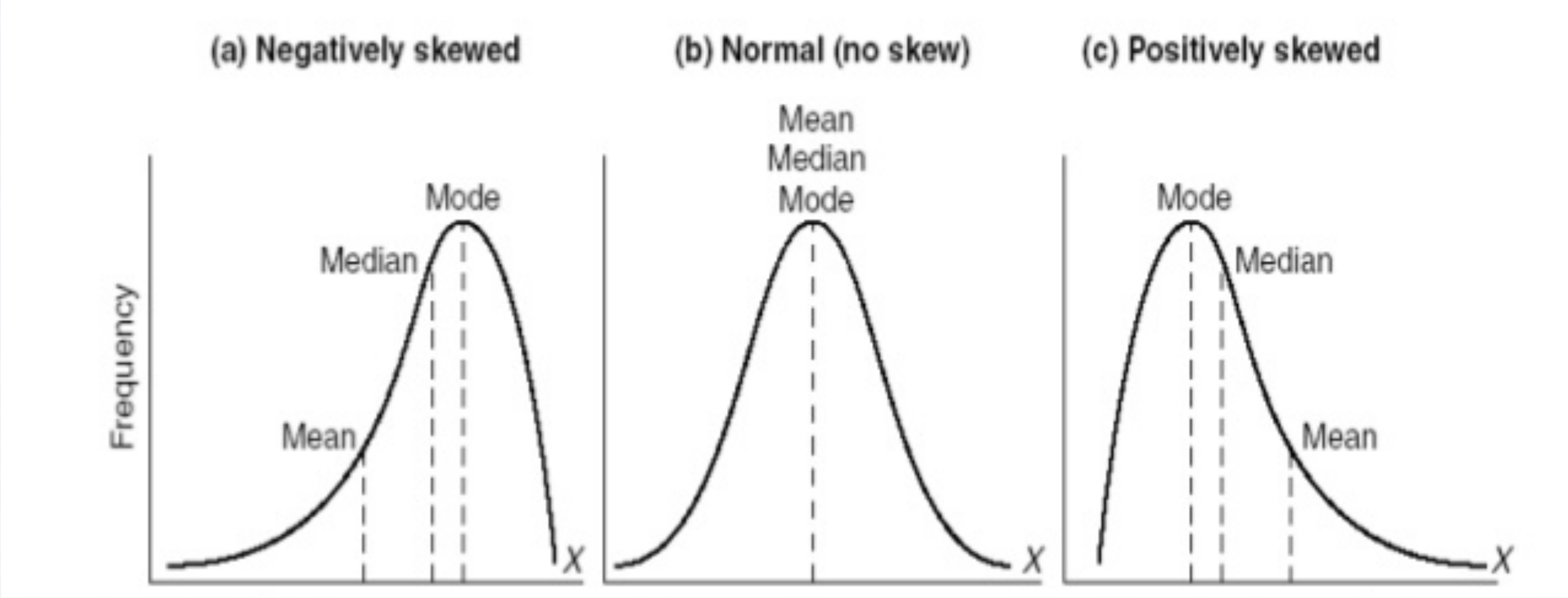


LAW OF LARGE NUMBERS (3)

- Important implication:
- The more you sample from a population (e.g., participants in an experiment), the closer the sample mean will be to the population mean
- Over time, independent event (generated by a random process) tend to approximate a normal distribution (the expected mean)

DEVIATION FROM NORMALITY (1)

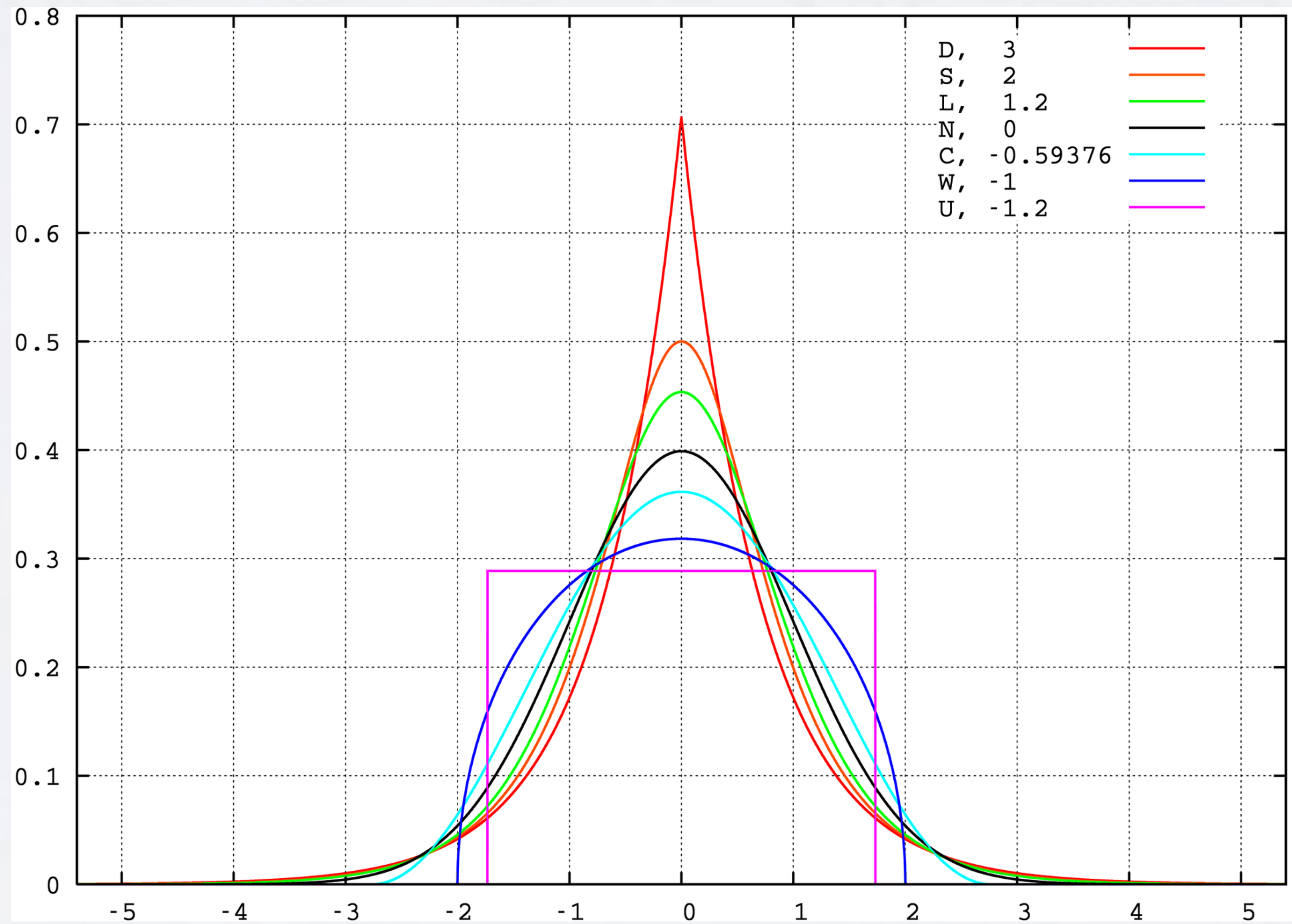
- Skewness
  - Most values on one side of the distribution, few on the other
  - Normal distribution has skewness = 0



mode>median>mean   mode=median=mean   mode<median<mean

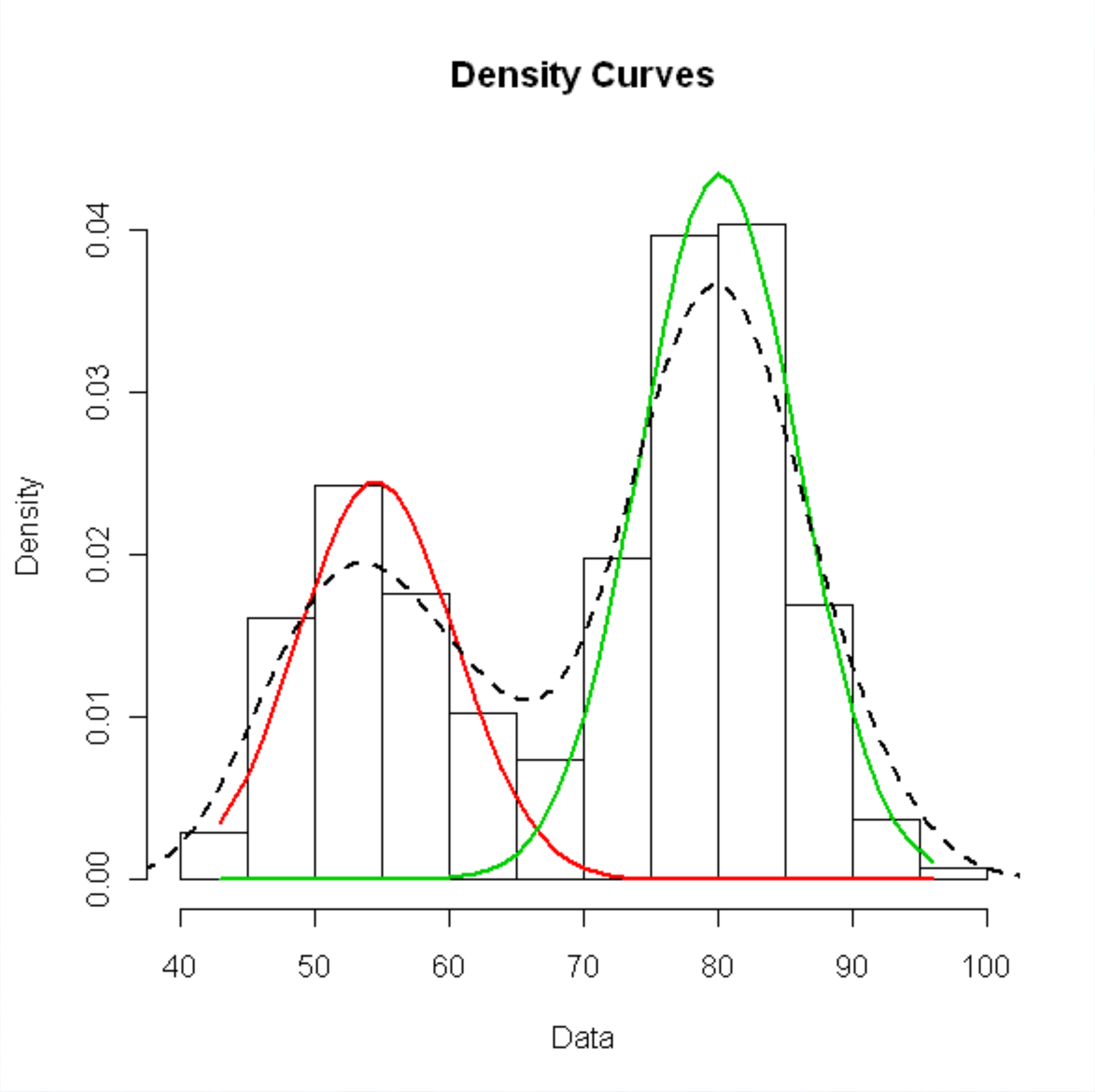
DEVIATION FROM NORMALITY (2)

- <u>Kurtosis</u>
  - How light- vs. heavy-tailed a distribution is
  - Leptokurtic (k > 0) or platykurtic (k < 0)
  - Normal distribution has kurtosis = 0



BIMODALITY

- Distributions with two modes
- May be a sign of two underlying unimodal distributions
- Different generative processes?
- E.g., height in men vs. women



WHY IS THE NORMAL DISTRIBUTION IMPORTANT?

- Many statistical tests (e.g., t-test) will only work on data that are normally distributed
- Most linear models (e.g., regression) will only work on data whose residuals (=measurement error) is normally distributed

TOMORROW

- Data mining = working with data sets with the purpose to discover patterns and insights
- Please make sure to download the CogSciPersonalityTest2019Data before this class.