Mt. Fuji, Japan

Data collection

What to measure, and how to measure it

Methods 1, E2021 - Lecture 2
Tuesday 7/9/2021
Fabio Trecca

Attendance registration

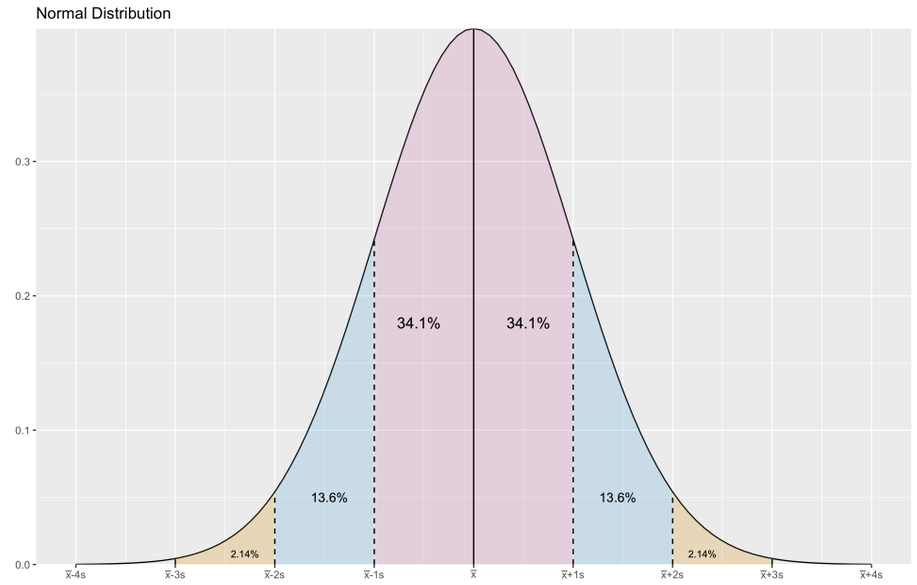
Check in using the PIN-code on the blackboard

Recap (1)

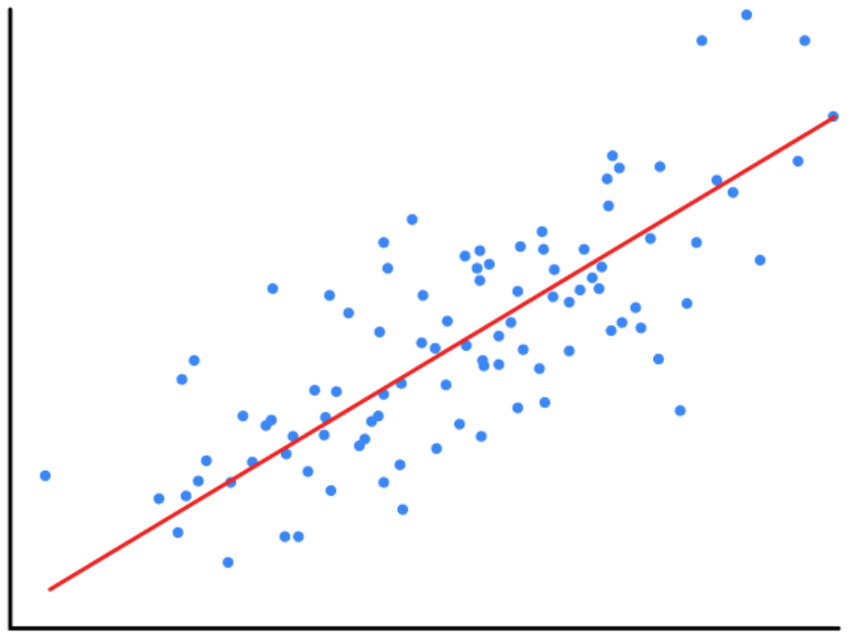
- The study of human cognition is interdisciplinary
- It must rely on insights from many disciplines
- It combines 1st, 2nd, and 3rd person methods
- However, the word “Cognitive Science” reflects specifically the use of quantitative/experimental methods that characterize much of the discipline

Recap (2)

- **Distribution**: the frequency/ probability with which certain values occur in a data set



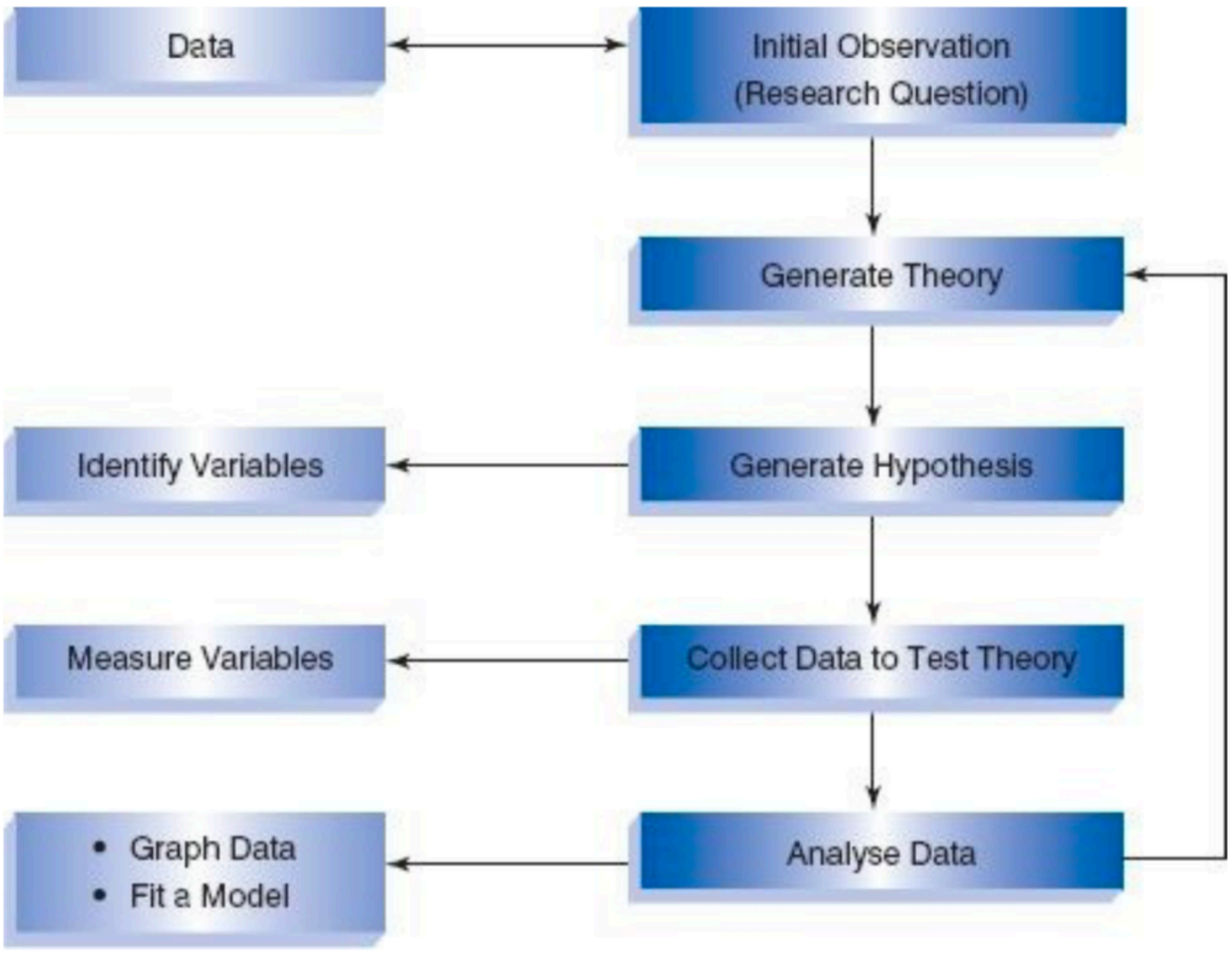
- **Model**: statistical summary of some data



What do i need to answer a research question?

- Data, where 1 data point = 1 individual observation
- Explanation of the data

The research process



Correlational vs. Experimental research

Between- vs. Within-subject design

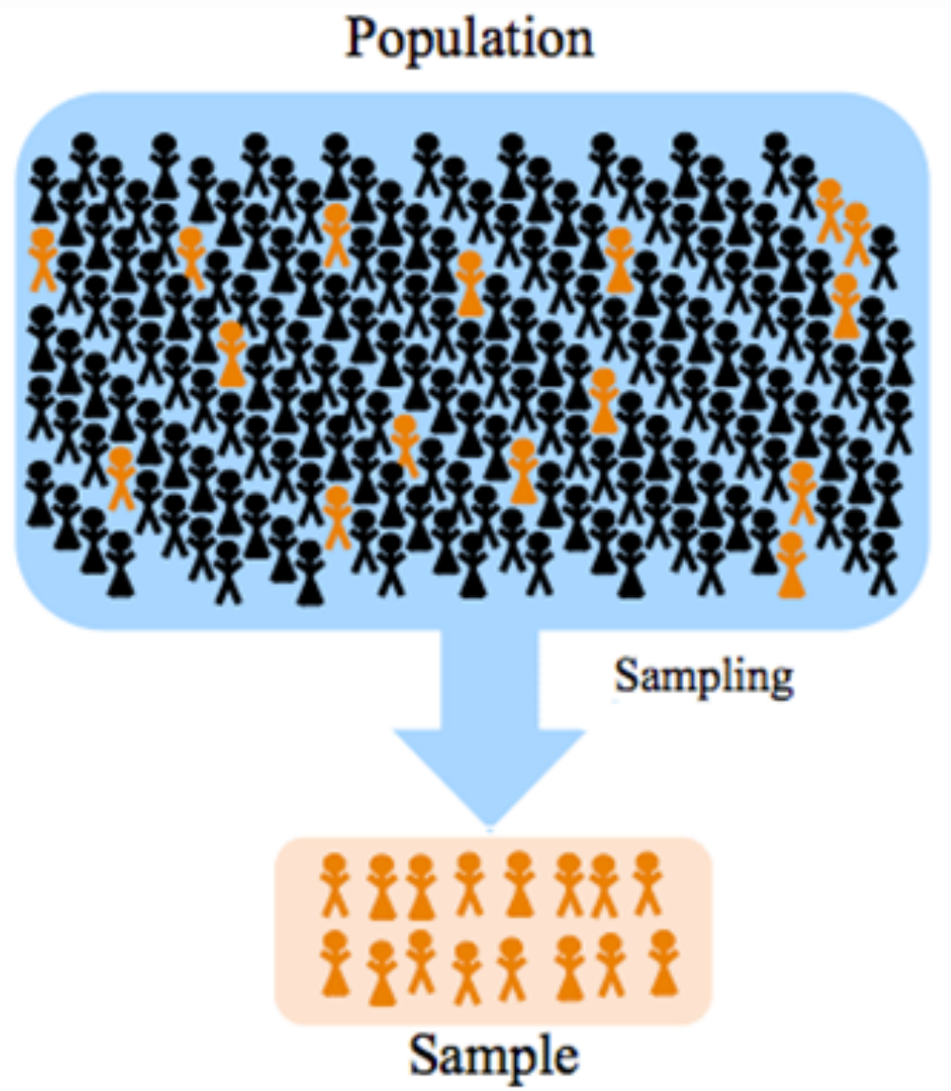
Measurement error

Unsystematic vs. systematic variation

etc.

Population vs. samples

- **Population** = The exhaustive collection of units to which we want to generalize our findings
  - N, μ, σ
- **Sample** = A smaller collection of units extracted from the population on which we can practically test our hypotheses
  - n, $\bar{x}$, s

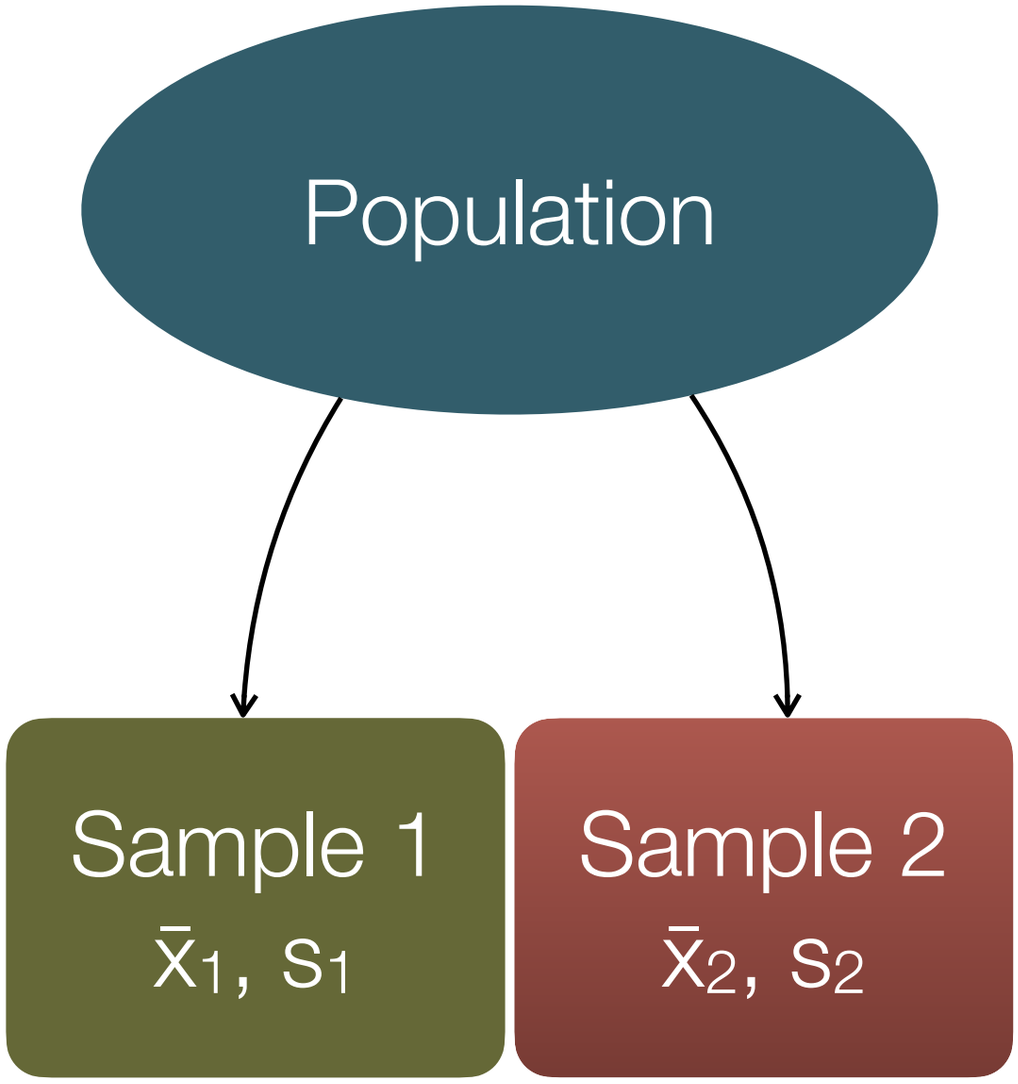


Hypothesis

- H_0 (null hypothesis) = no difference between the means

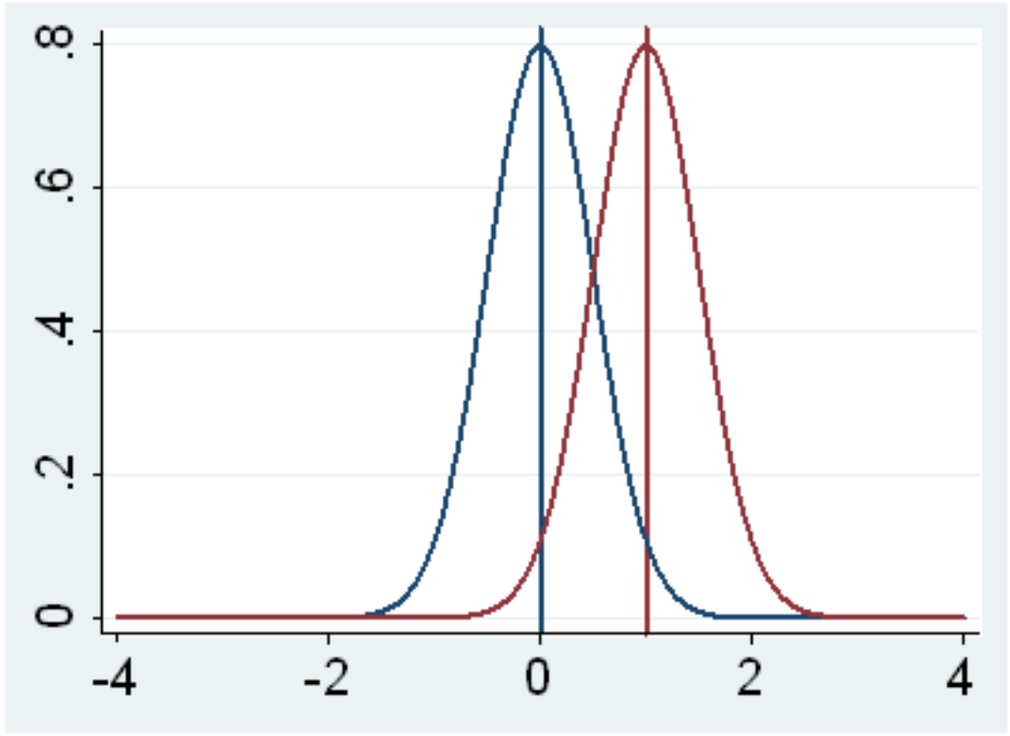
- H_1 (alternative hypothesis) = difference between the means

- Null hypothesis significance testing (NHST): we can't prove the H_1, but we can reject the H_0



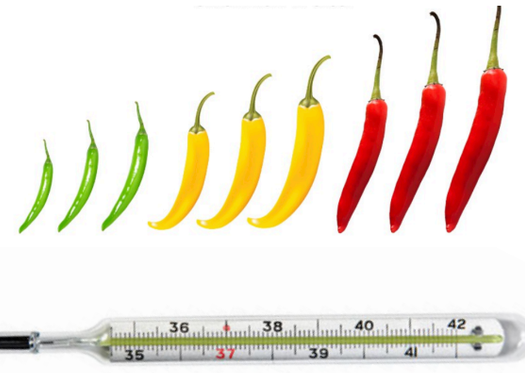
Why do we need stats?

- To discern systematic variation from unsystematic variance
  - **Systematic variance**: introduced by the experimental manipulation (good)
  - **Unsystematic variance**: inherent to the system and/or the measurement (bad)
    - aka error variance, residual variance, unexplained variance



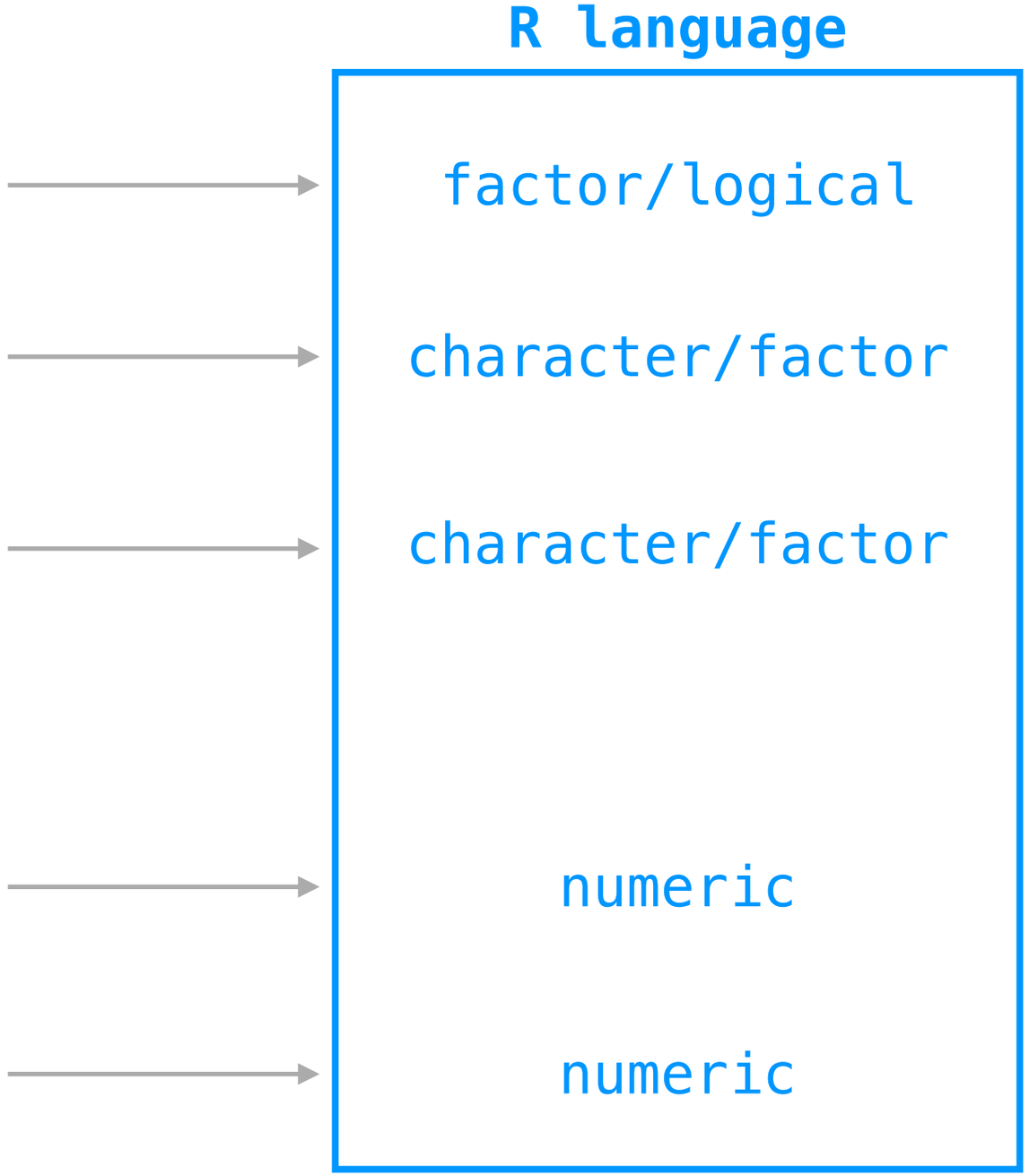
Variables

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



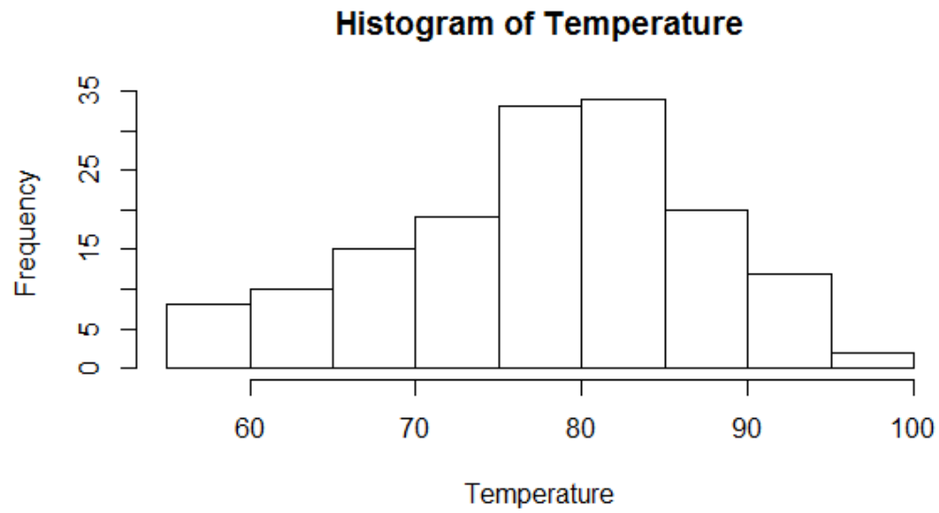
Variables

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



From data collection to 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?



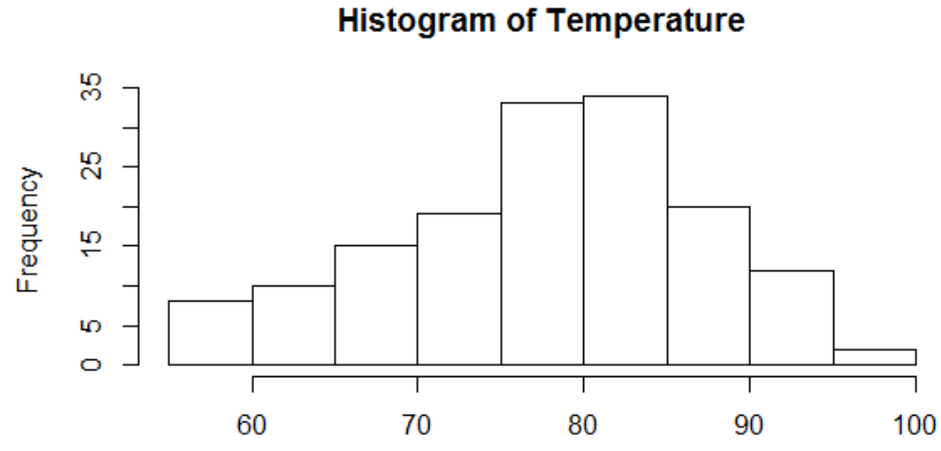
- Fit models (e.g., mean, correlation, linear regression)
  - what is the best way to summarise the raw data?

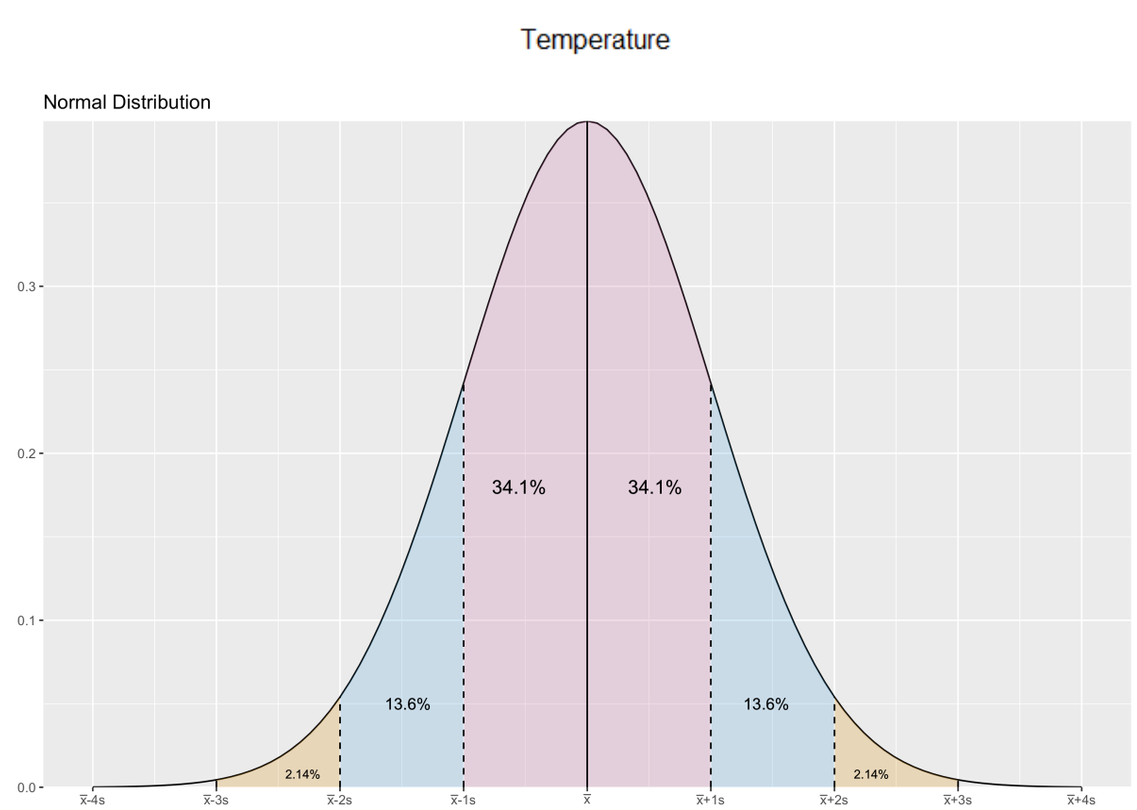
$$\mu = ?, \sigma = ?$$

$$\bar{x} = ?, s = ?$$

Frequency vs. probability distributions

- 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





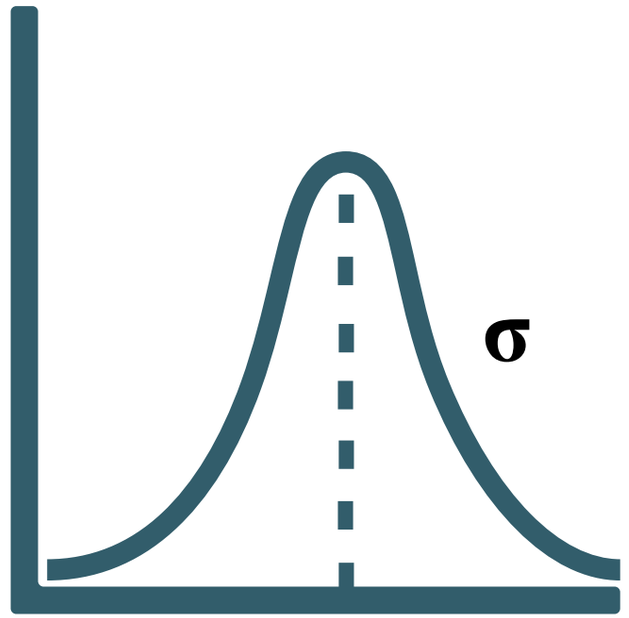
The normal (frequency) distribution

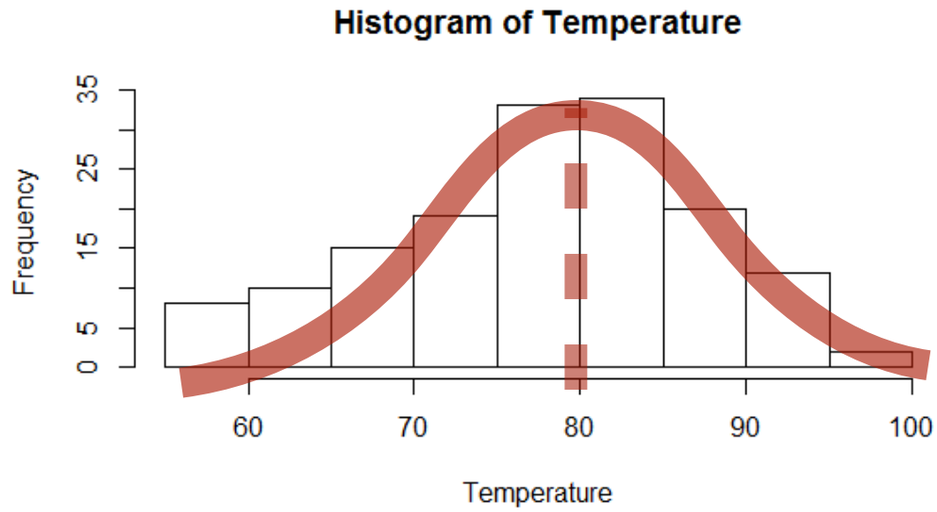
- Symmetrical gravitation toward the mean with decreasing N of data points as we approach the tails

- Many cognitive and behavioural processes are normally distributed

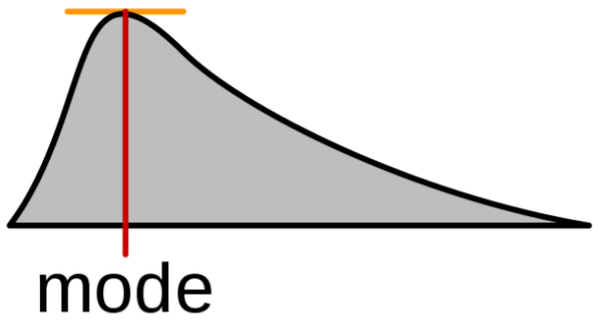
- Defined by two parameters: mean (μ) and standard deviation (σ)

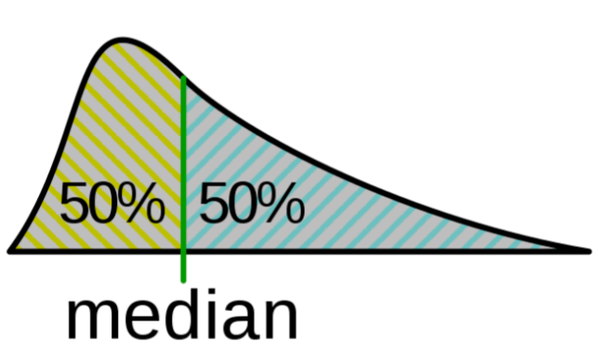
- Results from sum of independent events/ factors

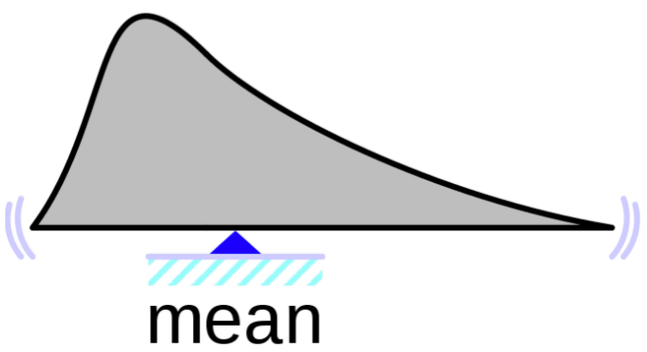


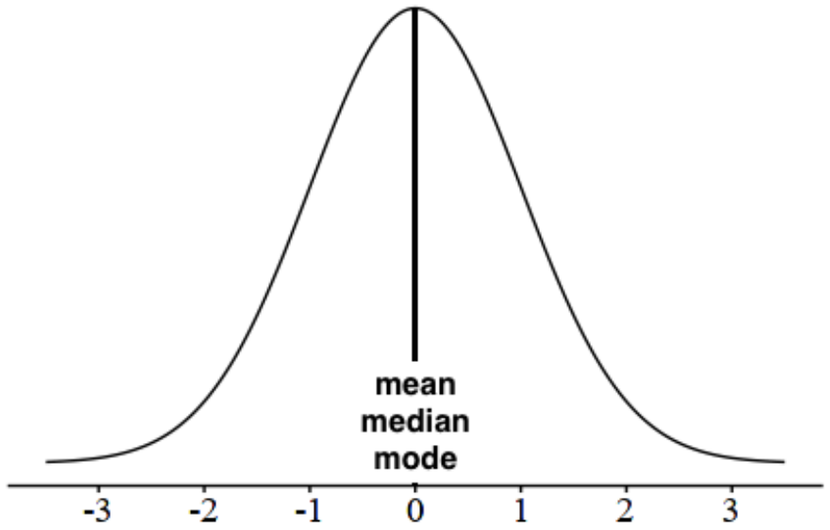


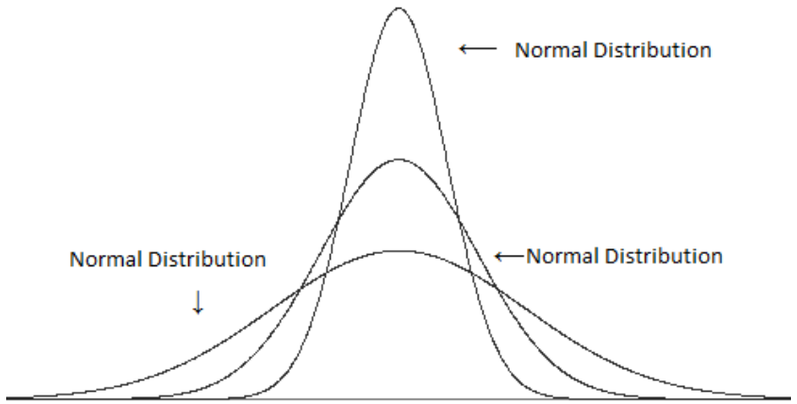
Measures of central tendency: mode, median, and mean

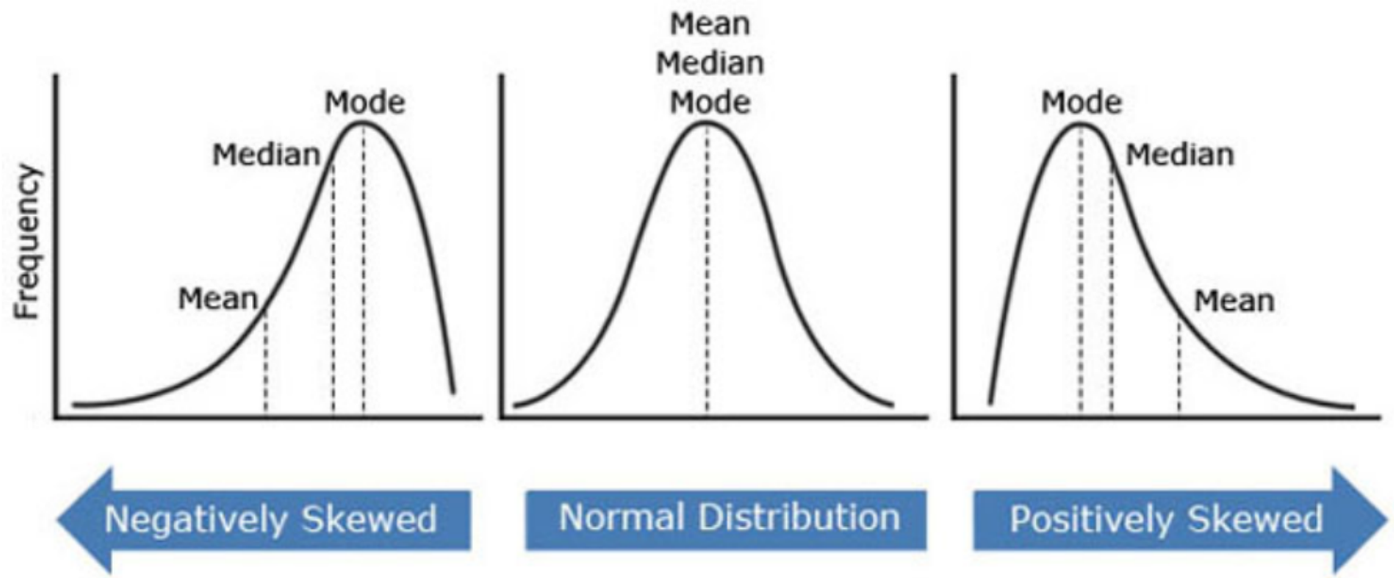






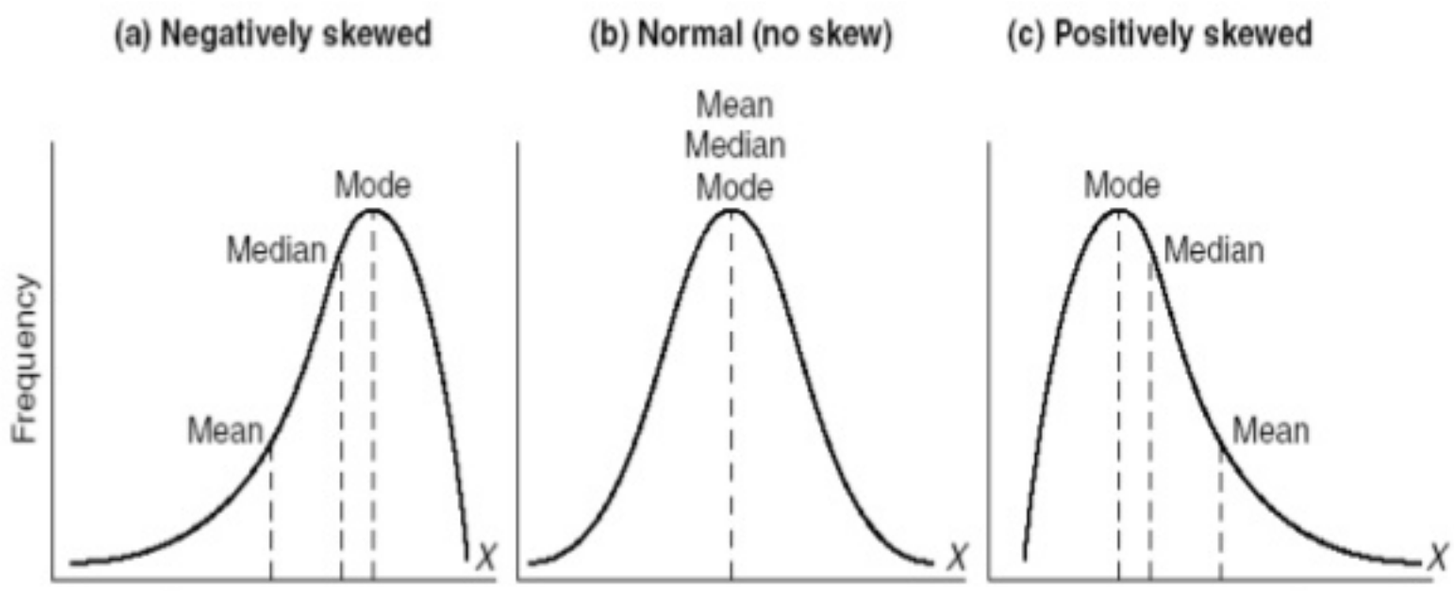






Deviation from normality (1)

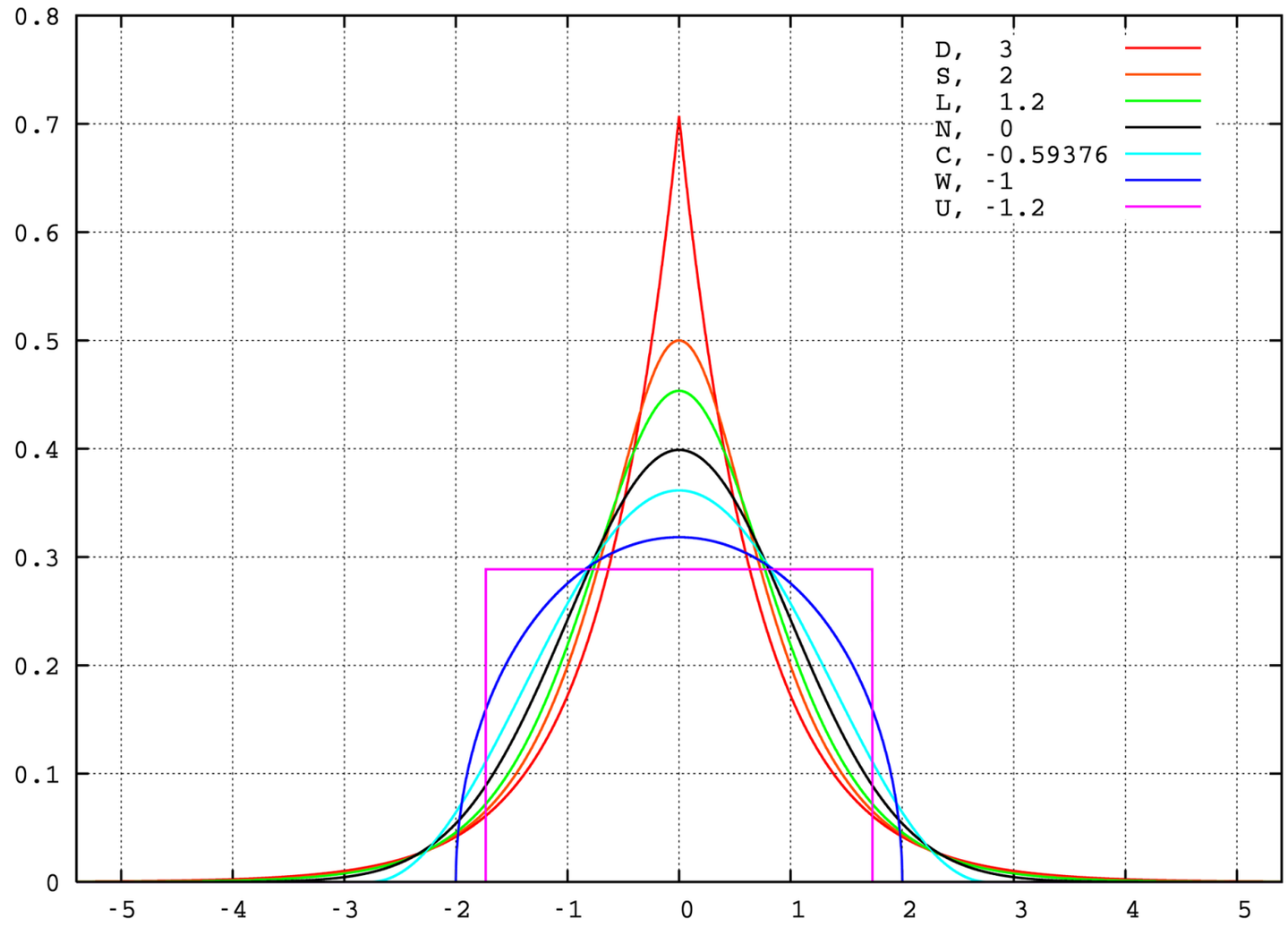
- Skewness
  - Most values on one side of the distribution, few on the other
  - Normal distribution has skewness = 0



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

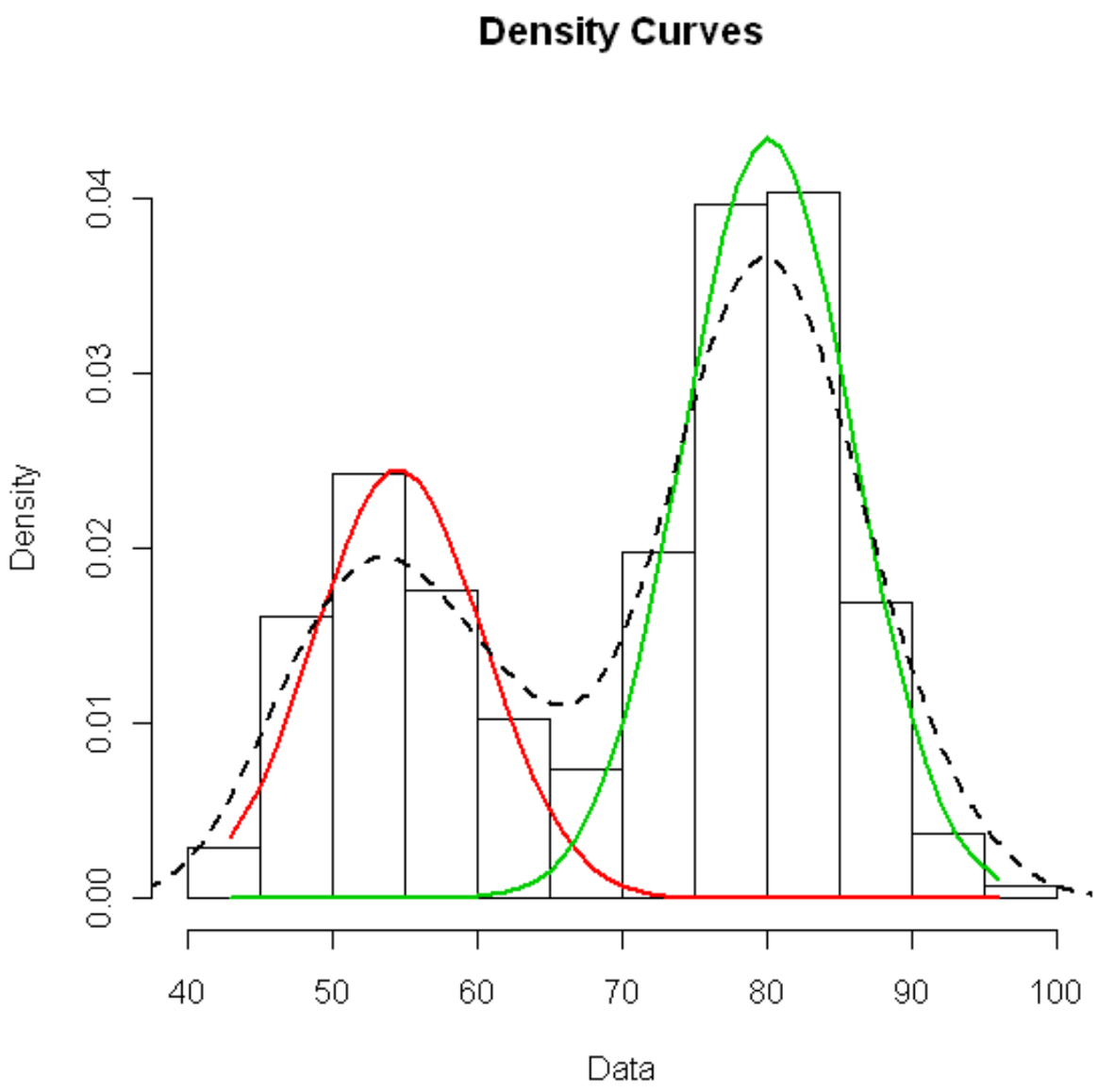
Deviation from normality (2)

- Kurtosis
  - 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
- Different generative processes?
- E.g., height in men vs. women



Why is the normal distribution important?

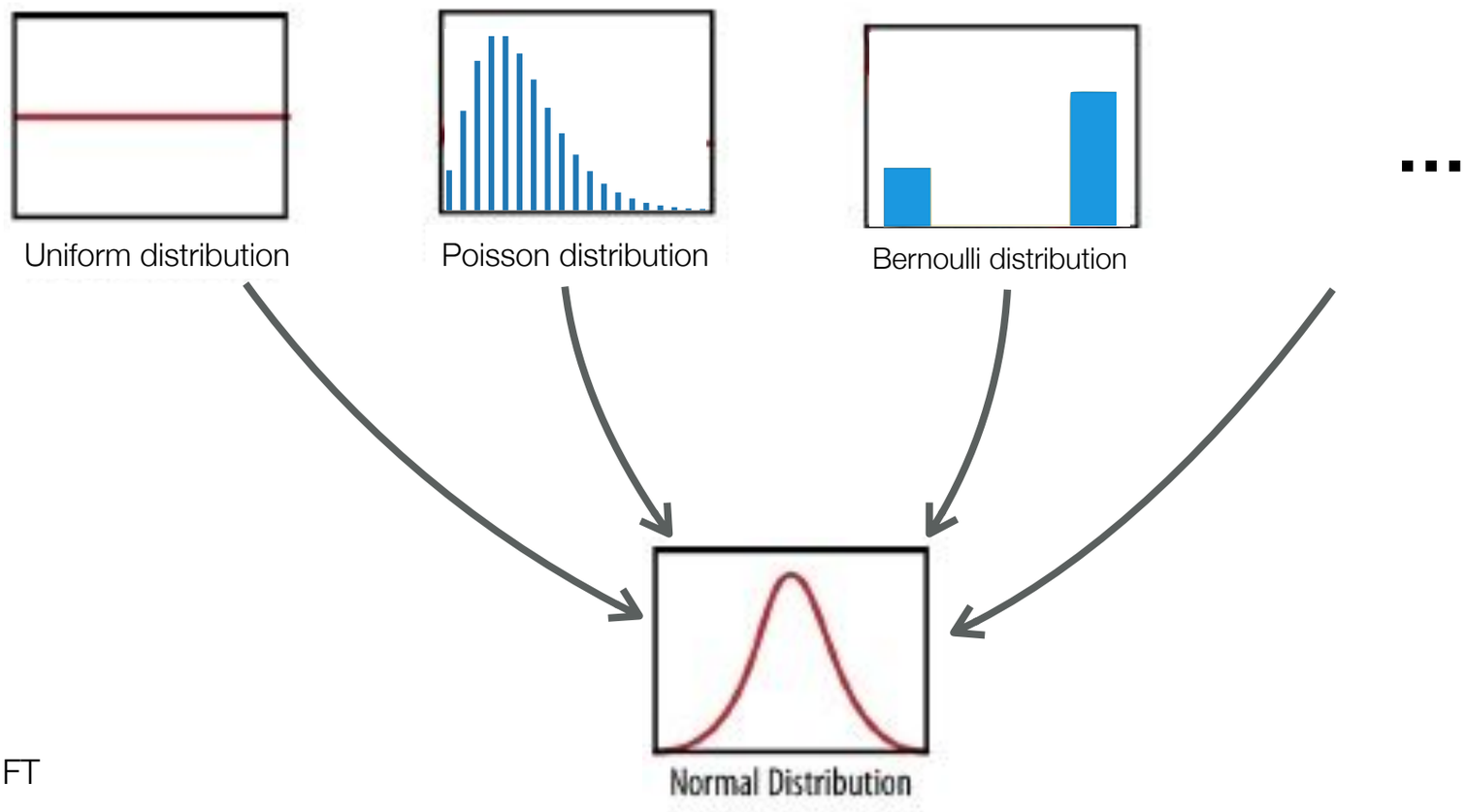
- Many statistical tests (e.g., t-test) will only work on data that are normally distributed

- Most linear models (e.g., regression) will only work on data whose residuals (=measurement error) is normally distributed

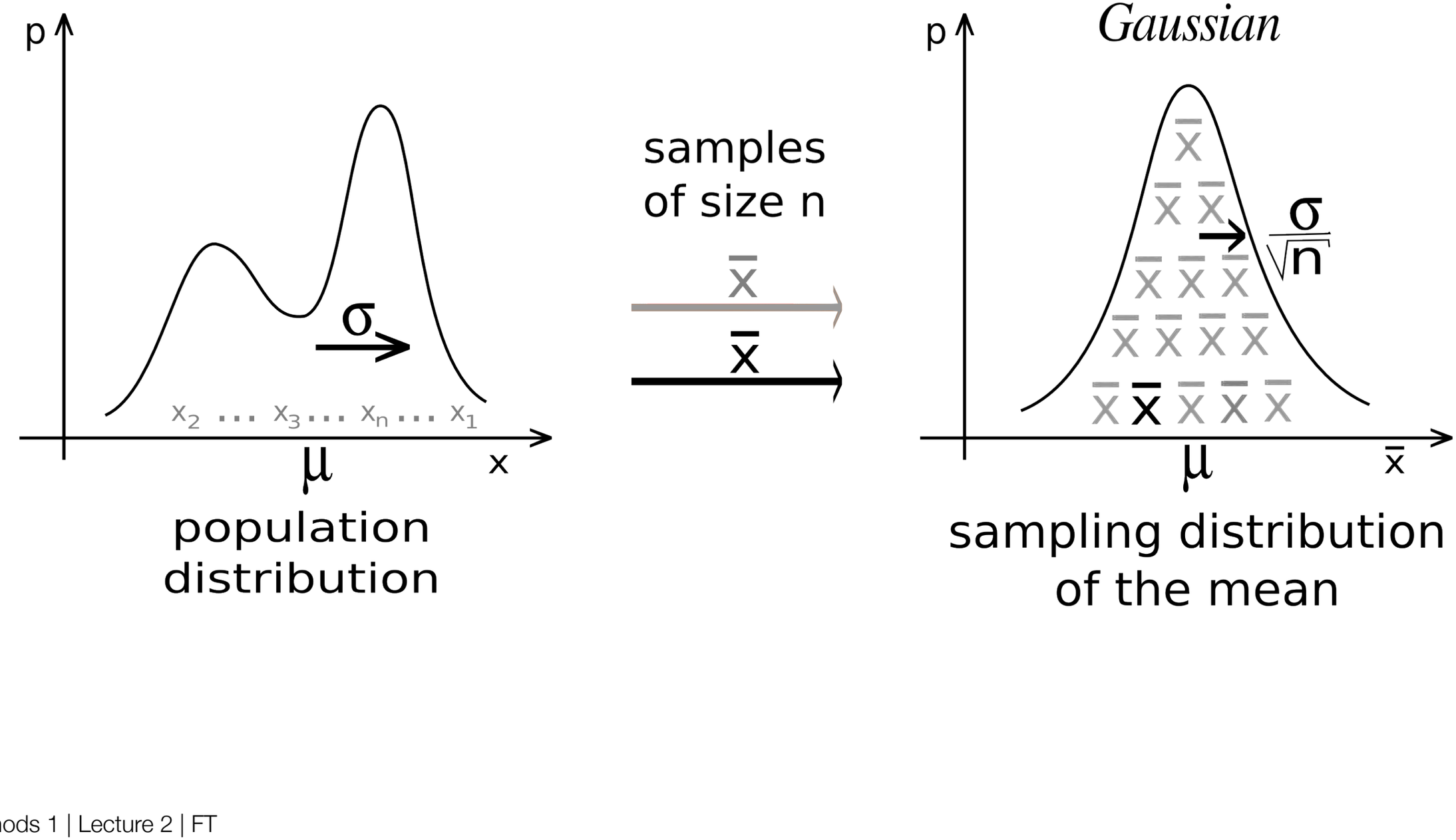
Central limit theorem (1)

- Given a population with unknown underlying distribution, the **sampling distribution of the sample means** will approximate the normal distribution

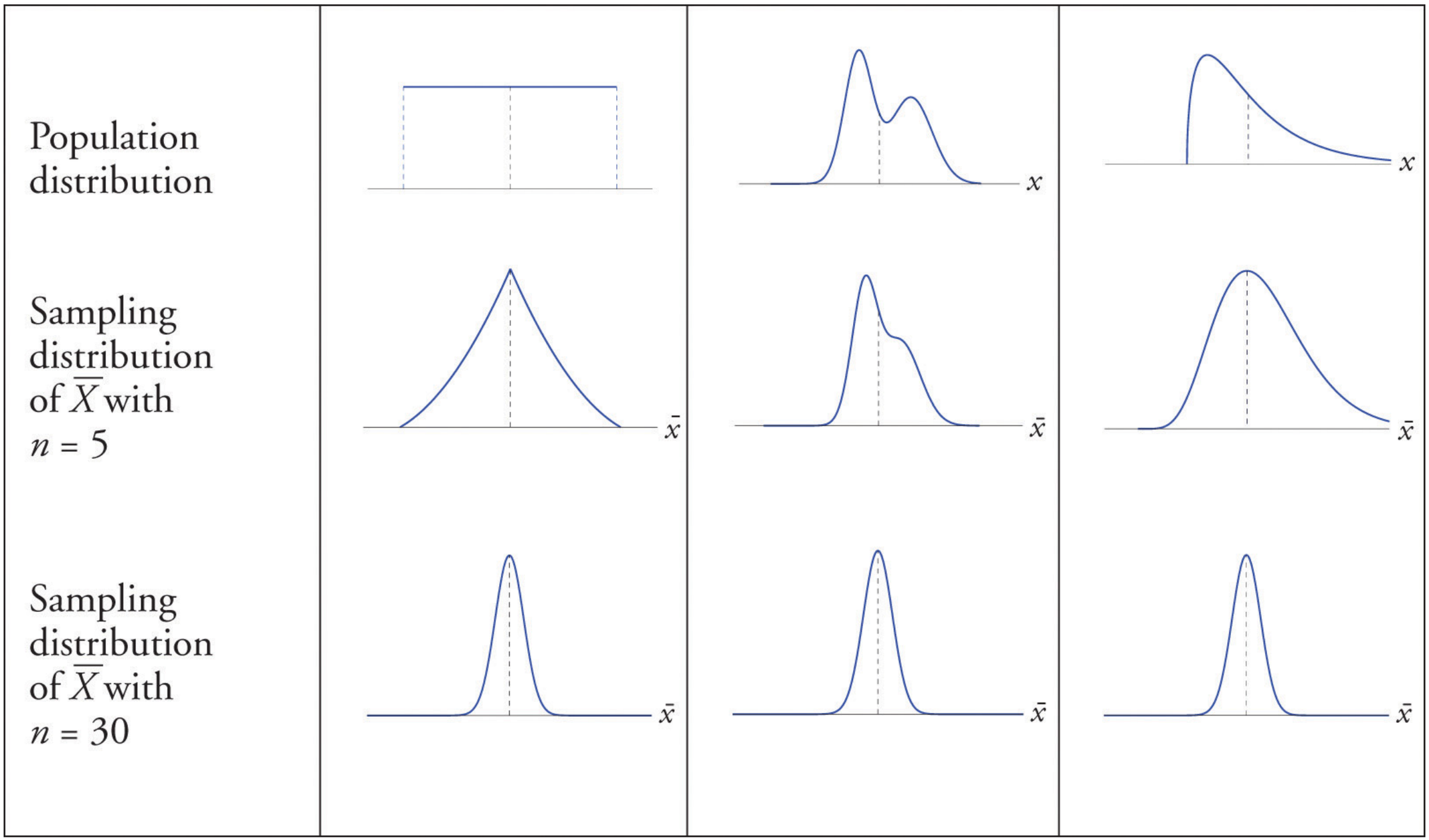
- Samples should be of sufficient size



Central limit theorem (2)



Sampling distribution of sample means



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

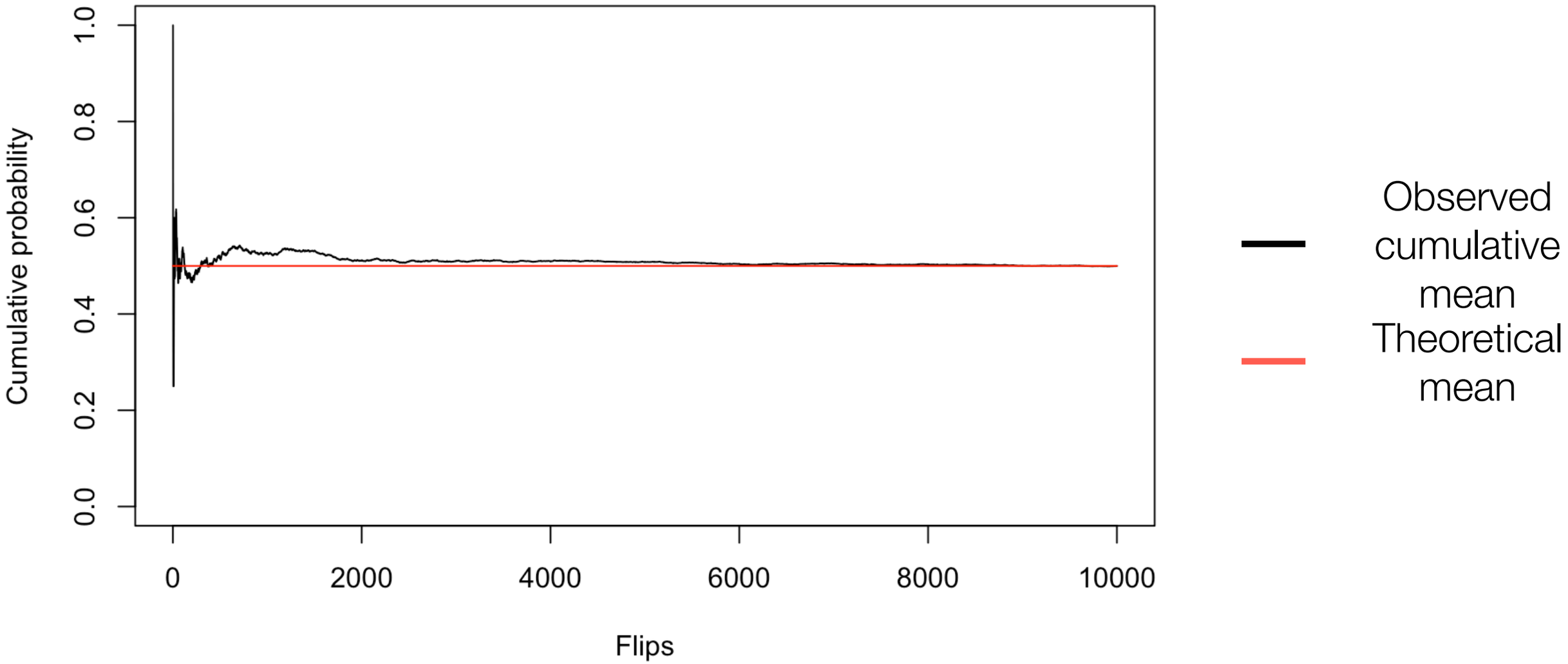
Law of large numbers (1)

- $\theta = 0.5$

- 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 implications:
  - **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)

Tomorrow

- **Exercise on data mining**: working with data sets with the purpose to discover patterns and insights

- Please make sure to download the CogSci Personality Test data before class