PSY515: Module 1 Lecture

Welcome to Quant 1
(aka Statistical Methods)

Goals of this course

- Understand the basic principles that underlie the type of statistical models that are widely used in Cognitive Science
- Learn how to select, implement, and interpret these statistical models
- Communicate the results of statistical analyses (through text and visualization)
- Incorporate best practices for open and reproducible research

Format

- **Learn**
  - Lectures
  - Labs
  - Readings
- **Practice**
  - Weekly Quizzes
  - Homework
- **Demonstrate Mastery**
  - Final Project
- **Get Support**
  - Journal Entries

My Goals

- Prepare you to both understand statistical analyses in the research you read and produce statistical analyses yourselves
- Create situations in which you can practice these skills that will help you throughout your career
- Challenge you to learn new things in a supportive environment
- Work together with you to make this a high-quality learning experience
- Give everyone an A

When are office hours?

You tell me!

Factors to consider:

- Lab sessions (that will prepare you for the homework) are on Thursdays.
- Homework is due on Tuesdays.
- It seems like Friday or Monday would be optimal, but I know that folks might not be on campus.
- I can hold office hours in person and/or virtually on Mondays or virtually (only) on Fridays.

A Quick Note about AI

- The use of AI is **not permitted** in this course.
- Why?
- I would like you to develop a deep understanding of how to run and interpret statistical analyses. Using AI to automate coding and interpretation will not help you learn and understand these concepts. It would be a disservice to you.

Questions?

- Please read the rest of the syllabus on your own.
- For the rest of today:
  - Descriptive Statistics, Models, and Distributions

Why do we describe data?

- Find errors in data entry or collection
- Understand your data
- Explore descriptive research questions
- Overall, there's a lot to learn from descriptive statistics.

Distributions

A **distribution** is a description of the [relative] number of times a variable X will take each of its unique values.

▶ Code

Question:

If I know nothing about someone, for example a participant in the survey, but I had to guess their Happiness rating, what would be the *best* number to guess?

The Mean!

If I don't have any other information, then the best "model" of my dataset would be the mean or average observation.

Mean, μ

- The **mean** is the average. The population mean is represented by the Greek symbol μ.
- Example: a set of numbers is: `7, 5, 8, 4, 9, 3`.

For a vector x with length N, the mean (μ) of x is:

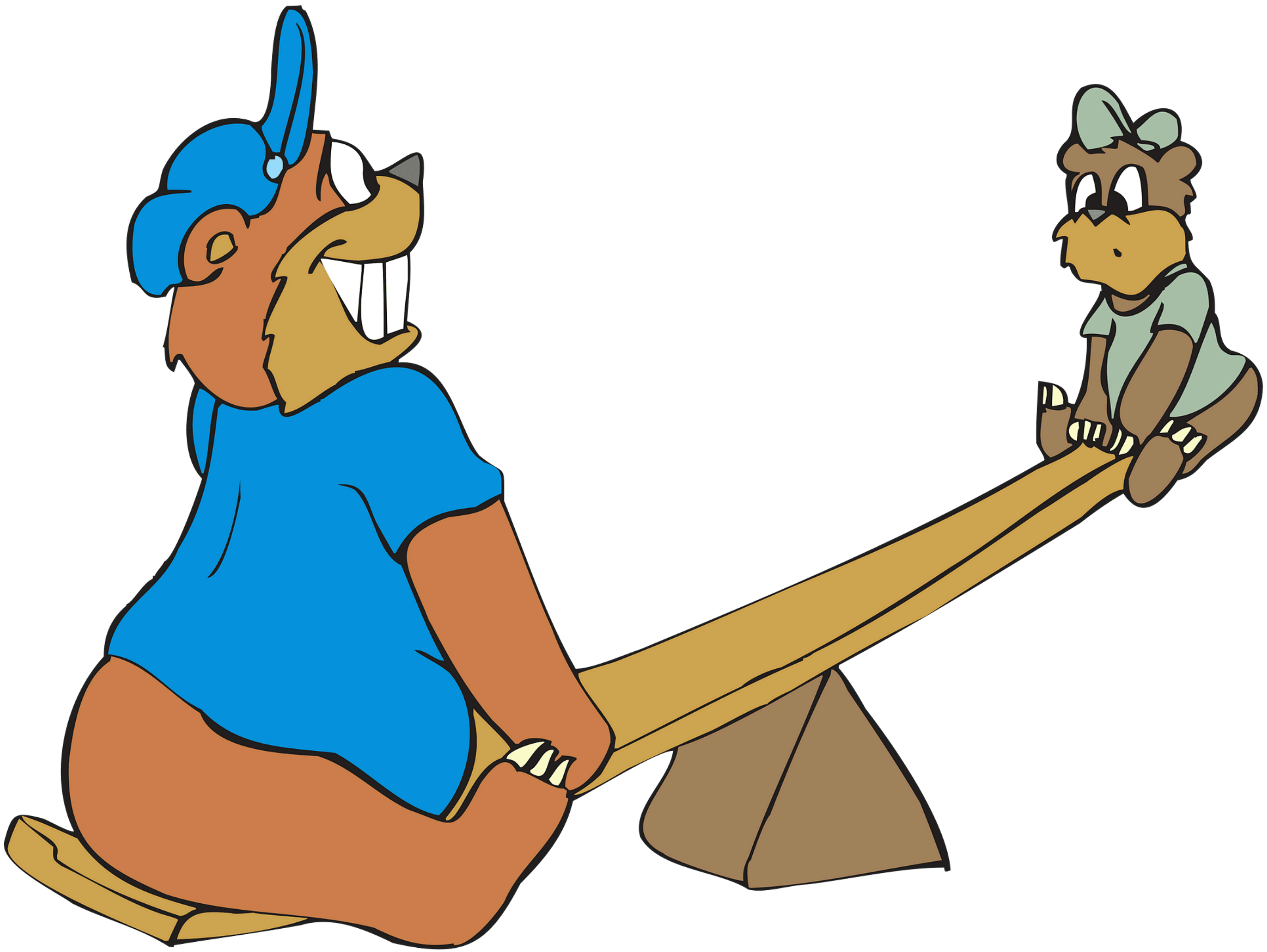
$$\mu = \frac{\Sigma(x_i)}{N} = \frac{7+5+8+4+9+3}{6} = \frac{36}{6} = 6$$

Properties of the mean

Example: a set of numbers is: `7, 5, 8, 4, 9, 3`. The mean of these numbers is 6.

- The mean can take a value not found in the dataset.
- Fulcrum of the data

The mean is the fulcrum of the data



Properties of the mean

Example: a set of numbers is: `7, 5, 8, 4, 9, 3`. The mean of these numbers is 6.

- The mean can take a value not found in the dataset.
- Fulcrum of the data

- The mean is strongly influenced by outliers.
- Deviations from the mean sum to 0

It’s important to remember that the mean of a population (or group) may not represent well some (or any) members of the population.

Example: André-François Raffray and the French apartment

Other measures of central tendency

- The **Mean** only one measure of *central tendency*
- **Median** – the middle point of the data
  - e.g., in the set of numbers `7, 10, 8, 3, 9, 3, 12`, the median number is 8.
  - You can see this if you write them in order: 3 3 7 **8** 9 10 12
- **Mode** – the number that most commonly occurs in the distribution.
  - e.g., in the set of numbers above, the mode is 3 because it occurs twice.

Center and spread

- Distributions are most often described by their **center** (mean/median) and **spread** (variance/standard deviation).
- Typically, these two parameters are used in common inferential techniques.
- The mean represents the average score in a distribution. A good measure of spread will tell us something about how the typical score deviates from the mean.
- Why can't we use the average deviation?

Sums of squares

Our solution is to square deviations.

```
x = c(7, 5, 8, 4, 9, 3)
mean(x)
```

```
[1] 6
```

```
(deviation = x - mean(x))
```

```
[1]  1 -1  2 -2  3 -3
```

```
deviation^2
```

```
[1] 1 1 4 4 9 9
```

```
sum(deviation^2)
```

```
[1] 28
```

The sum of squared deviations is referred to as **the Sum of Squares (SS).**

Variance

We calculate the average squared deviation: this is our variance, σ^2:

```r
sum((x - mean(x))^2)/length(x)
```

```
[1] 4.666667
```

Standard Deviation

Standard deviation σ is the square root of the variance.

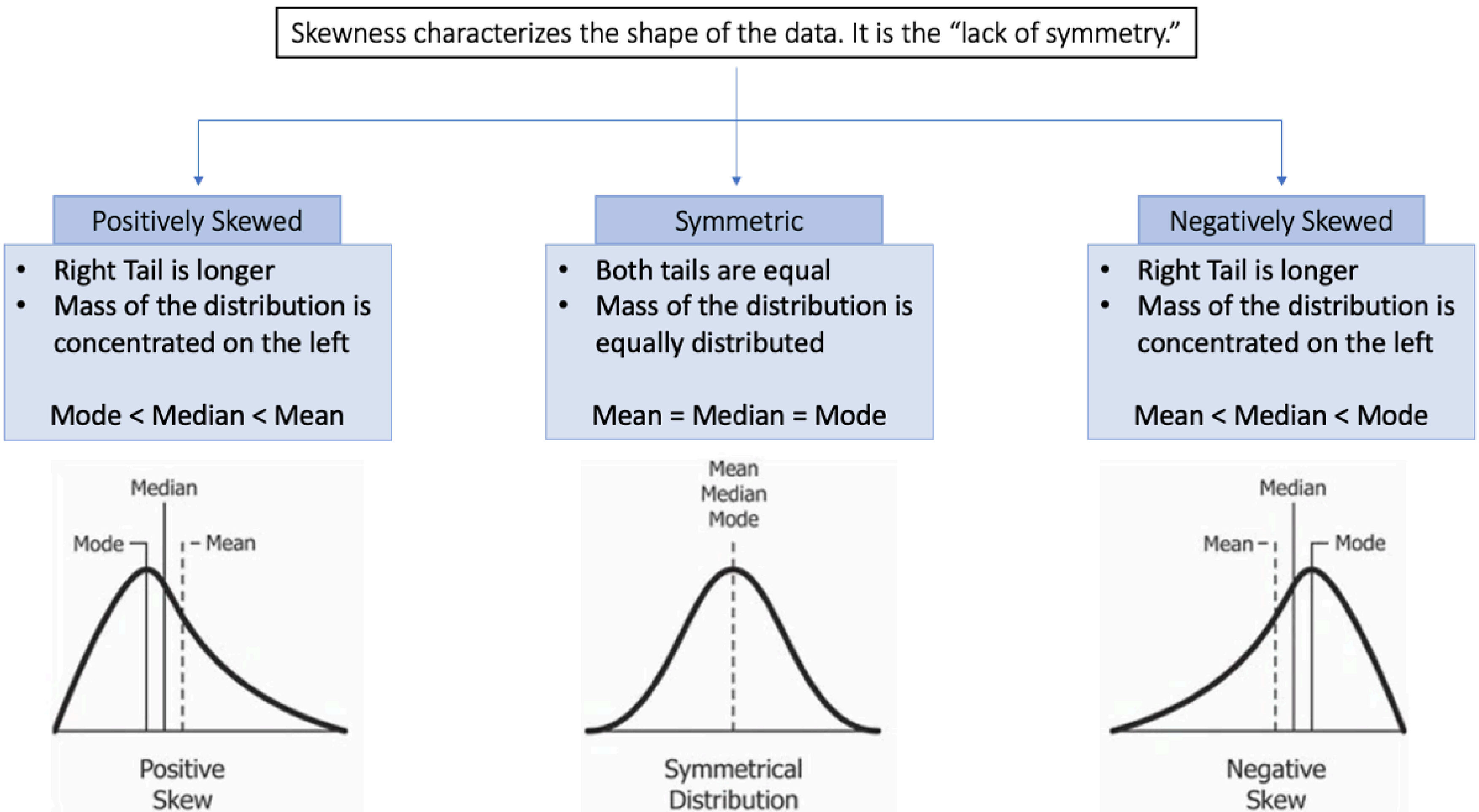
```
sqrt(sum(deviation^2)/length(deviation))
```

```
[1] 2.160247
```

The standard deviation is more interpretable than the variance. It can be thought of as the average distance of scores from the mean.

Skew

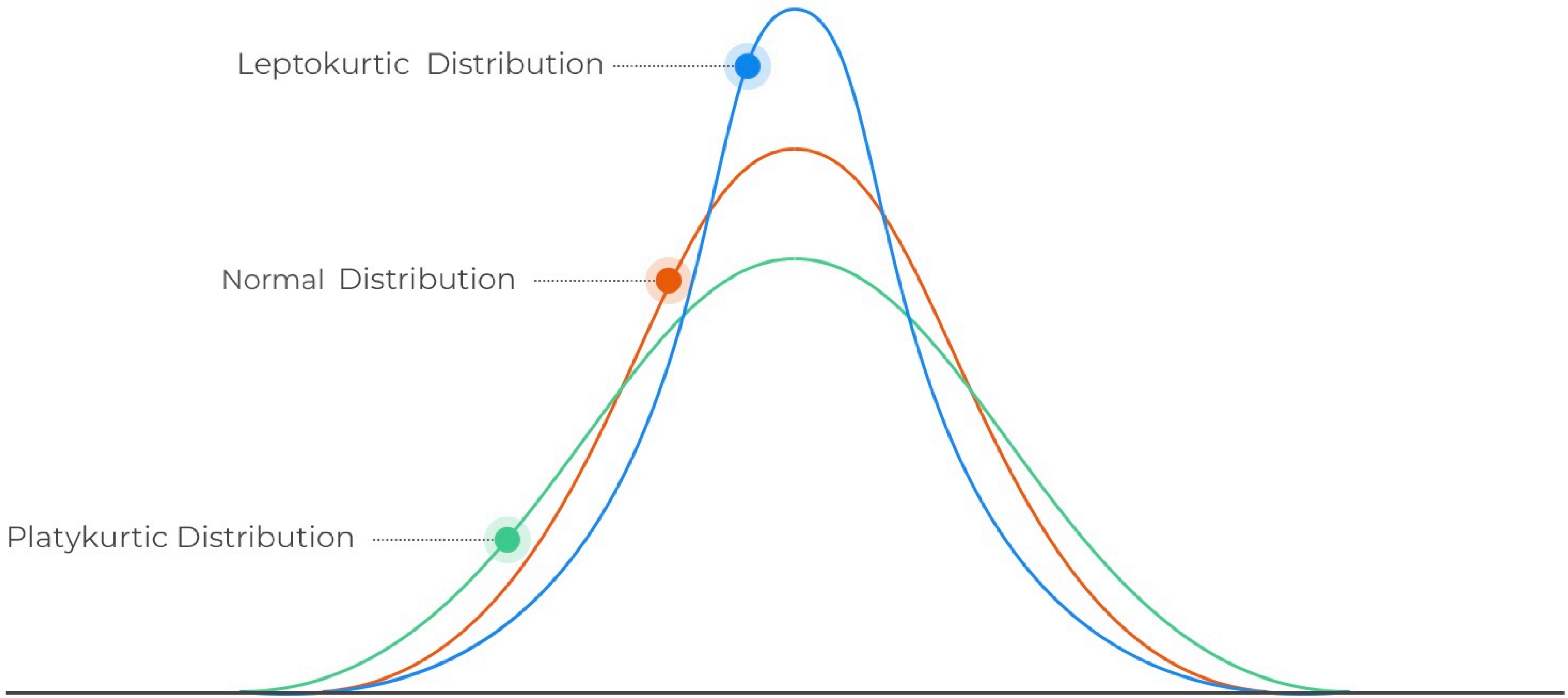
Skewness characterizes **symmetry** of a distribution.



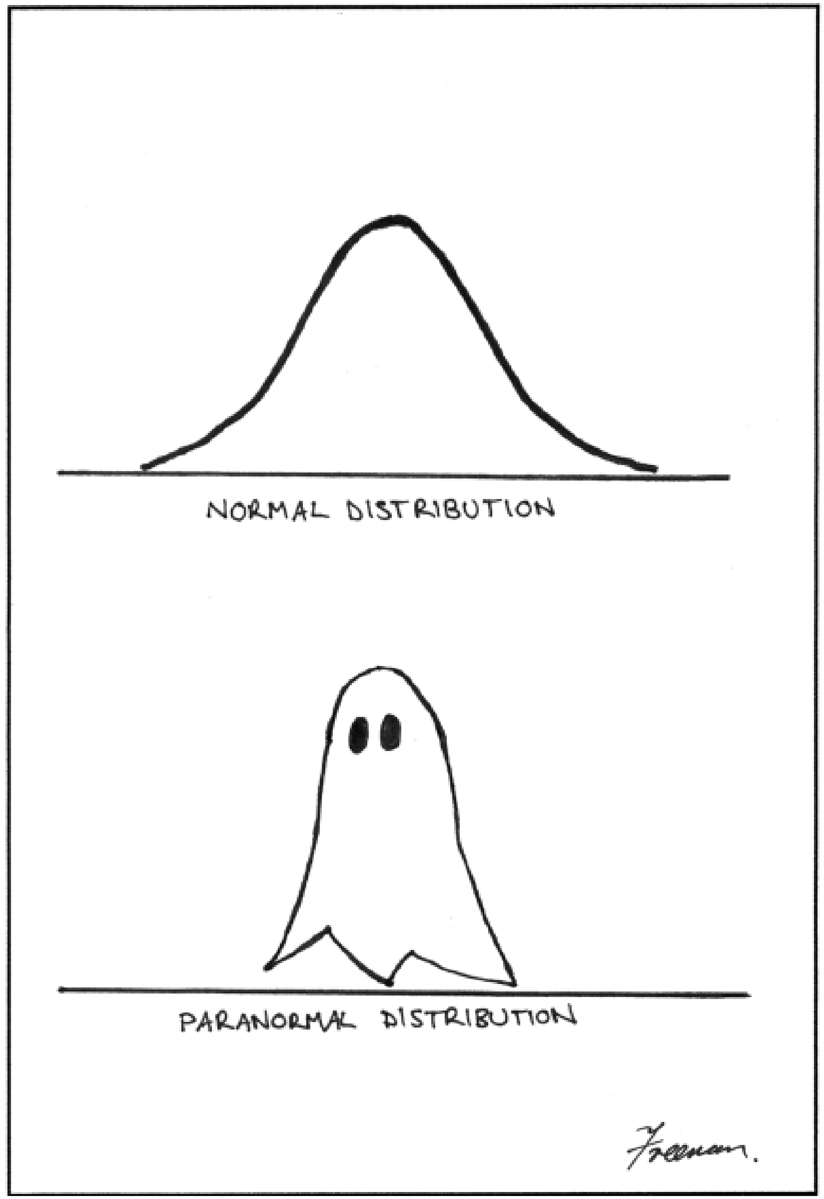
In general, when building statistical models, we must not forget that the aim is to understand something about the real world. Or predict, choose an action, make a decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world: our models are not the reality. Hand (2014)

Kurtosis

Kurtosis characterizes **tail-heaviness** of a distribution.



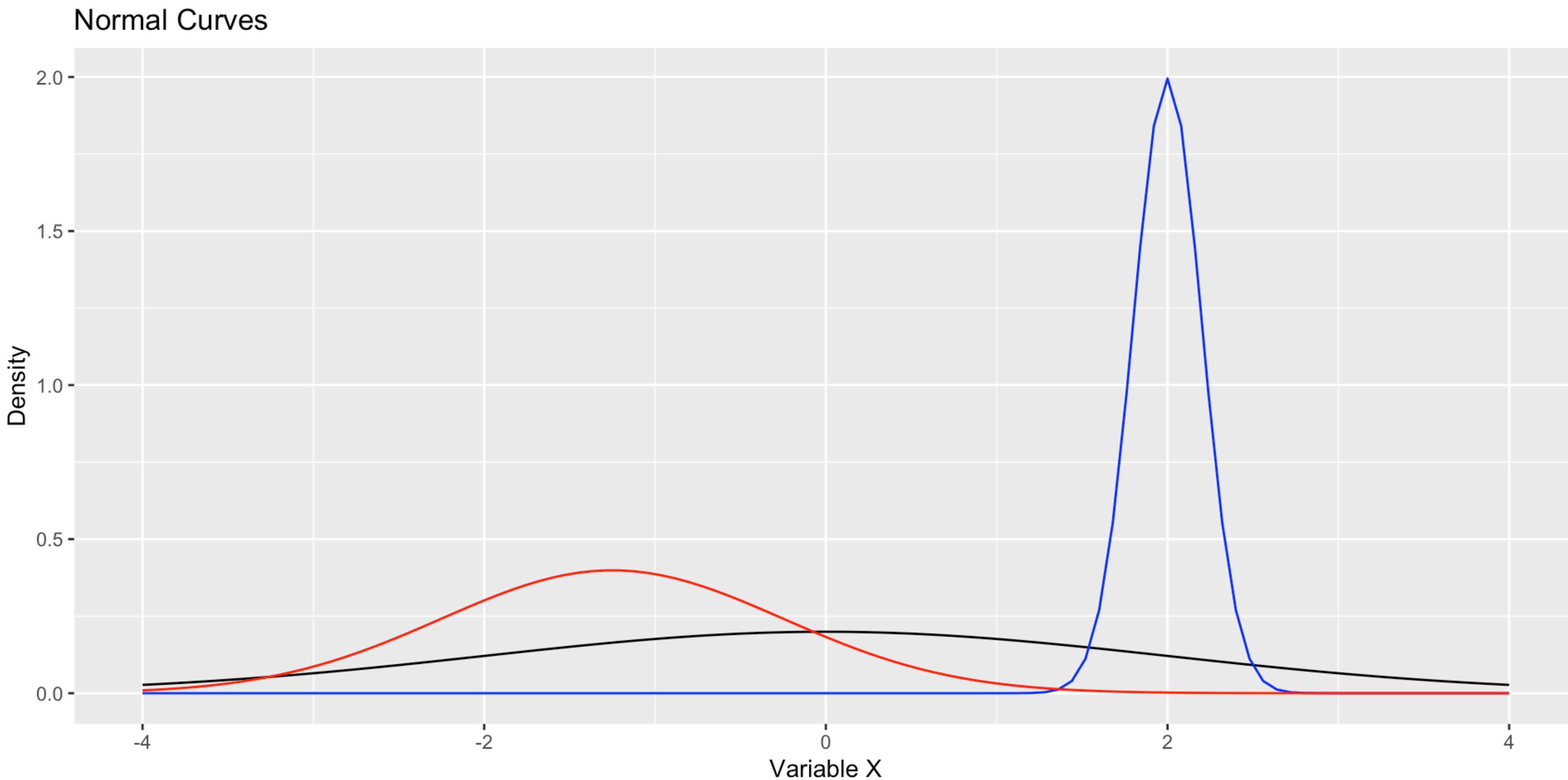
The Normal Distribution



Characteristics of the normal distribution

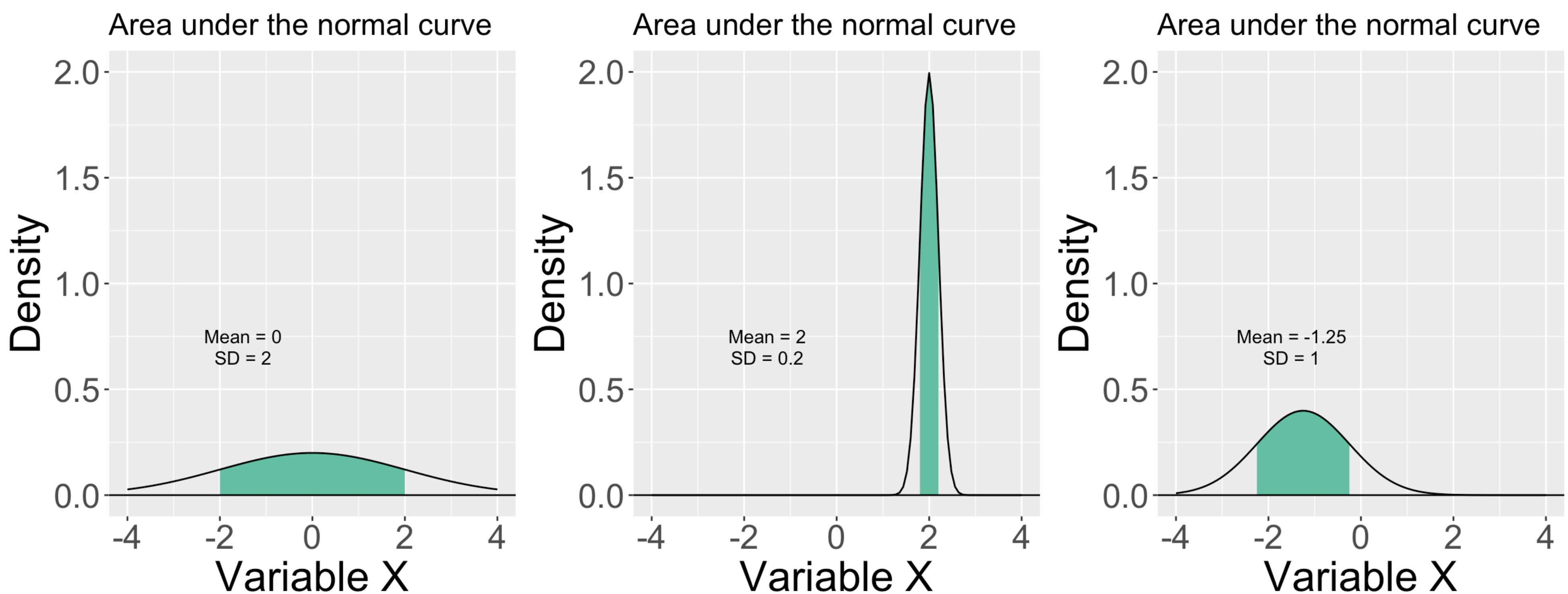
- The mean and standard deviation are independent.
- The distribution is unimodal and symmetric.
- The area of under the curve between corresponding locations, in standard deviation units, is the same regardless of μ and σ.
  - For example, in a normal distribution, approximately 68% of the area under the curve falls between 1 σ below the mean and 1 σ above mean—for every normal curve (regardless of the value of the mean and standard deviation).

▶ Code



All of these distributions are normal and have an equivalent area (proportion) that falls between one standard deviation below and one above their respective means.

▶ Code

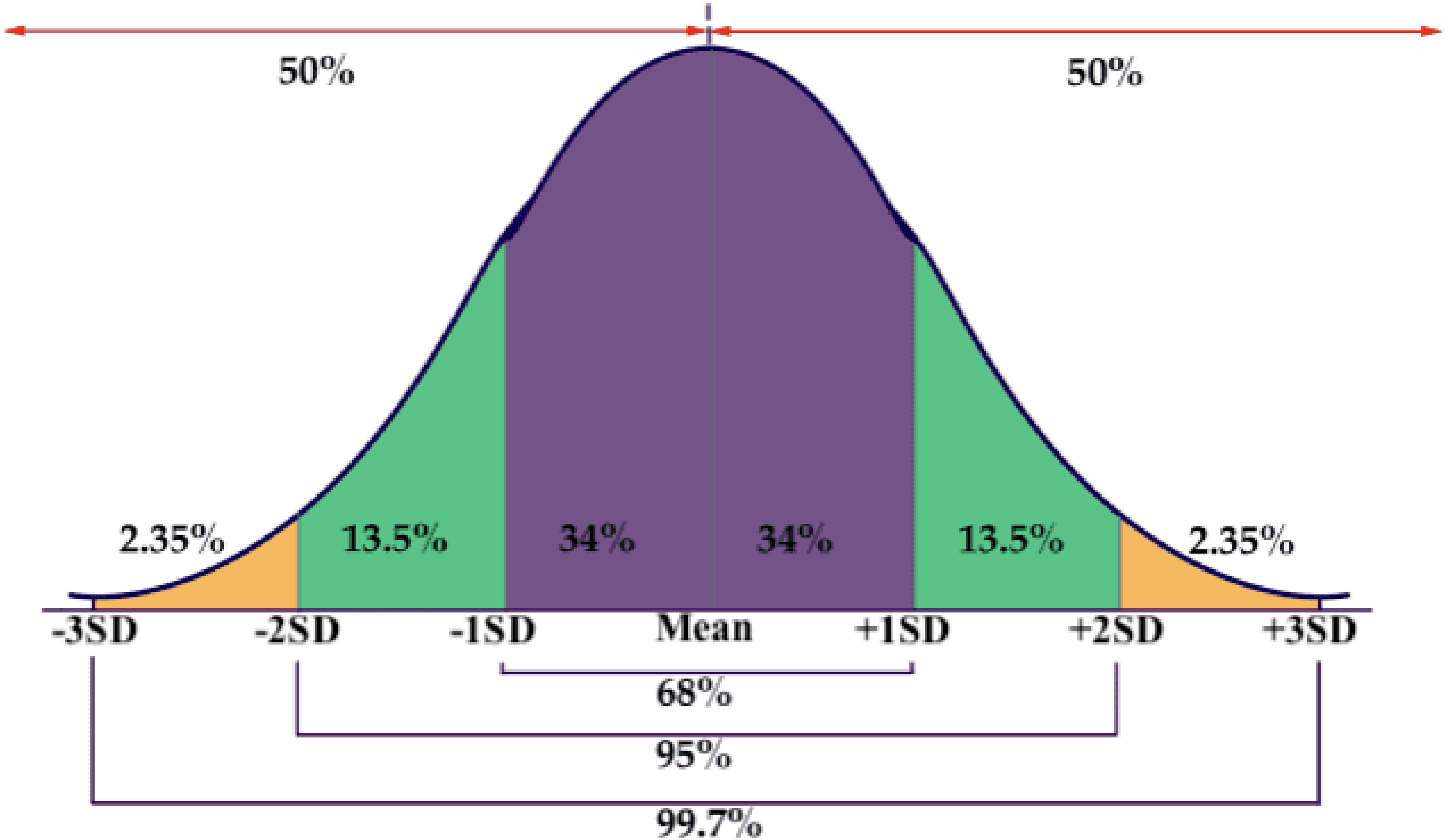


The Empirical Rule

In a normal distribution:

- Approximately 68% of the data falls within **one** standard deviation of the mean.
- Approximately 95% of the data falls within **two** standard deviations of the mean.
- Approximately 99.7% of the data falls within **three** standard deviations of the mean.

The Empirical Rule



Questions?