EDP308: STATISTICAL LITERACY

The University of Texas at Austin, Fall 2020
RAZ: Rebecca A. Zárate, MA

Overview

- Types of Statistics
  - Descriptive vs. Inferential
- Central Tendency
  - Mean, Median, Mode
- Skewed Data
  - Left (negative skew)
  - Right (positive skew)
- Probability Distributions
  - Normal, Bimodal, Uniform
- Notation and Differentiations
  - Greek is for Populations, Roman is for Samples
- Calculating a Mean and Median in R

Describing the Data

What is the purpose?

- Statistics serve one of two purposes.
  - Used to DESCRIBE a sample data set
    - Summaries (mean, variance)
    - Visual representations (graphs, charts)
  - Used to INFER and draw conclusions about the population as a whole from a data set
    - Hypothesis testing
    - Variance comparisons
    - Regression analysis

Two Types of Statistics

<u>Descriptive Statistics</u>

- Summarizing sample data sets
  - Distribution
    - Frequency, %
  - Central Tendencies
    - Mean, median, modes
  - Measures of Spread
    - Standard deviation, variance
  - Measures of Association
    - Correlation

<u>Inferential Statistics</u>

- Inferring things about a population from sample
  - Hypothesis Testing
  - Determining Association
    - Regression Analysis
  - Comparing Means
    - T-tests
  - Comparing Variance
    - Chi-Squared
    - ANOVA

We'll focus on Descriptive Statistics for now.

Central Tendency

Central Tendency

Mean

Median

Mode

What are they?

What do they tell us?

Why use one over the other?

Central Tendencies

- Mean ($\bar{x}$):
  - Average of set
    - Ex. 1, 2, 3, 4, 5 = 15 (total)/5 (number of #s) = 3
    - 3 is the average
- Median:
  - Middle-ranked item of set, splits set 50%
  - Good for skewed data
    - Ex. 2, 2, 2, 5, 6, 7, 7
    - 5 is the median
- Mode:
  - Most recurrent item
  - Good for categorical data
    - Ex. Ex. 2, 2, 2, 5, 6, 7, 7
    - 2 is the most recurrent value

Money.

What is the average income in the USA?

(How could I ask this in a better way?)

Money.

- The mean income in the USA is around: $48-69k
  - How does this strike you? Sound right?

Income

Imagine a bar filled with your every day, average American…

With their average income…

Then…

Bill Gates walks into a bar…

…*EVERYONE INSIDE BECOMES A MILLIONAIRE!*…

…on average…

Those are just back statistics…

Mean vs. Median

(1)

Name	Annual Income
Tom	$32,000
Larry	$36,000
Susan	$39,000
Paul	$41,000
Marcus	$45,000
Randy	$50,000
Sandy	$57,000
Tim	$60,000
Pam	$65,000
Jack	$80,000

(2)

Name	Annual Income
Tom	$32,000
Larry	$36,000
Susan	$39,000
Paul	$41,000
Marcus	$45,000
Randy	$50,000
Sandy	$57,000
Tim	$60,000
Pam	$65,000
Bill Gates	$1,000,000,000

mean income of $50,500

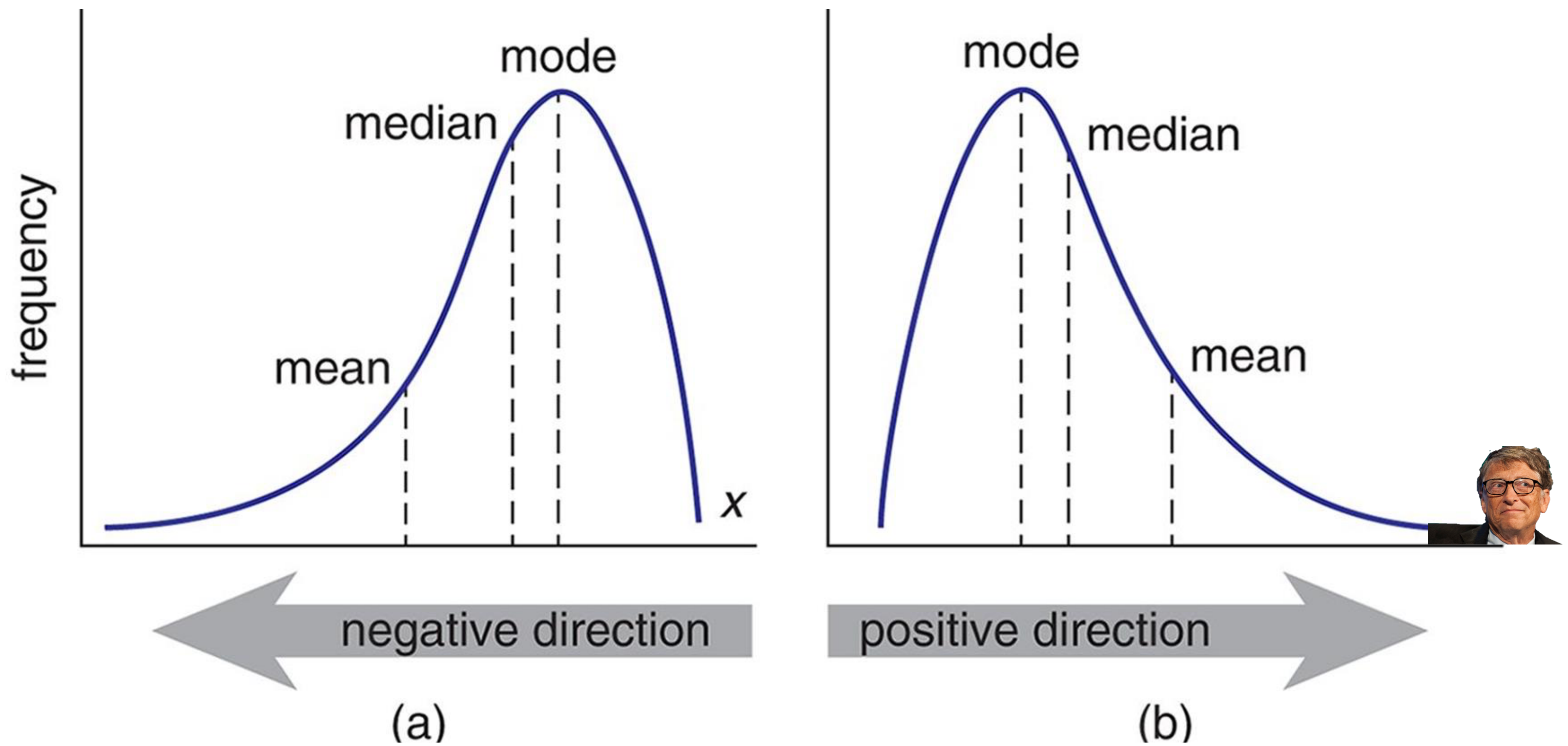
median income of $47,500

mean income $100,042,500

median income of $47,500

Sensitivity and Outliers

- Mean is sensitive to outlier, I'm look at you Bill…
  - Medians can be a more accurate representation.

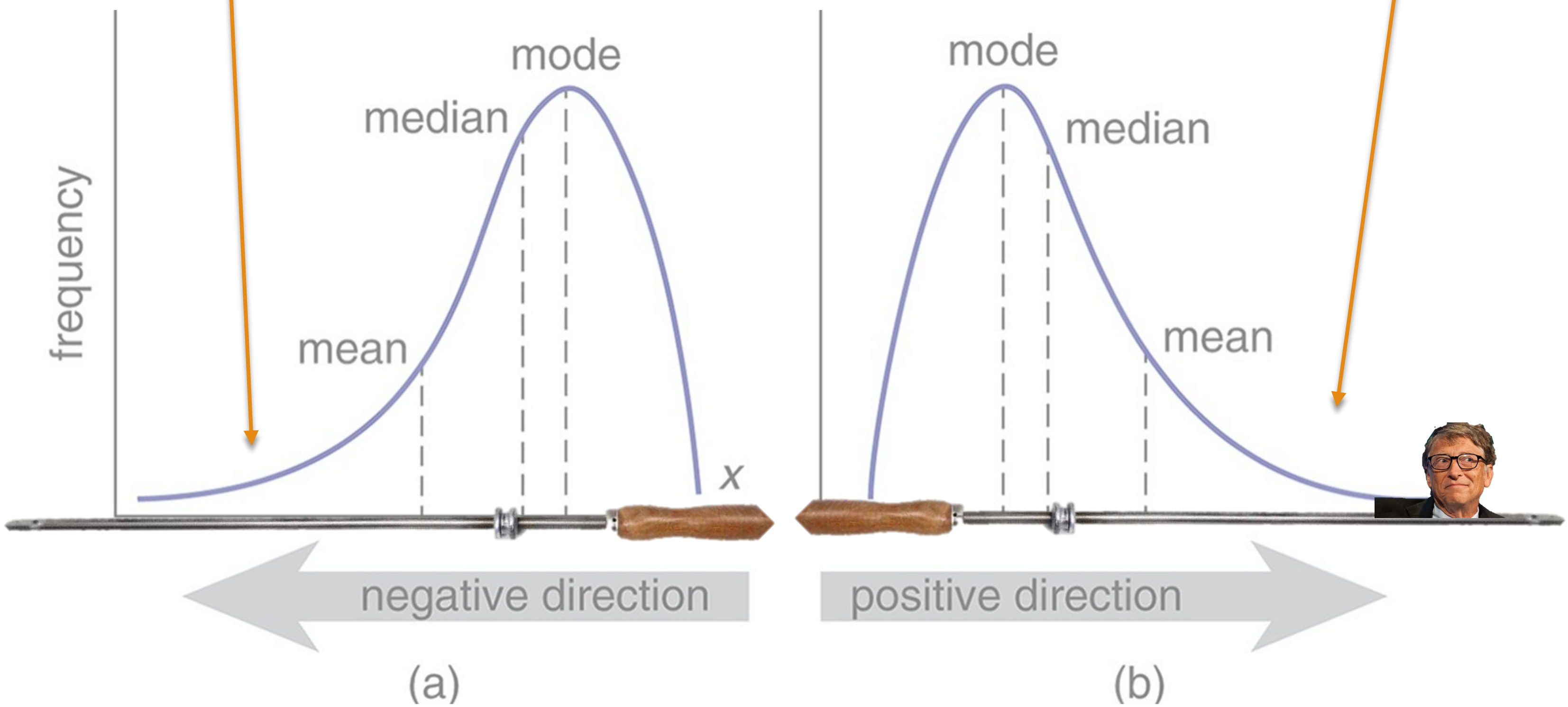


Skews

This is a LEFT skew,
Or negative skew.

This is a RIGHT SKEW
Or positive skew.

Skews are always referring to the skewer…
Like Bill Gates who is a skew-er of mean income

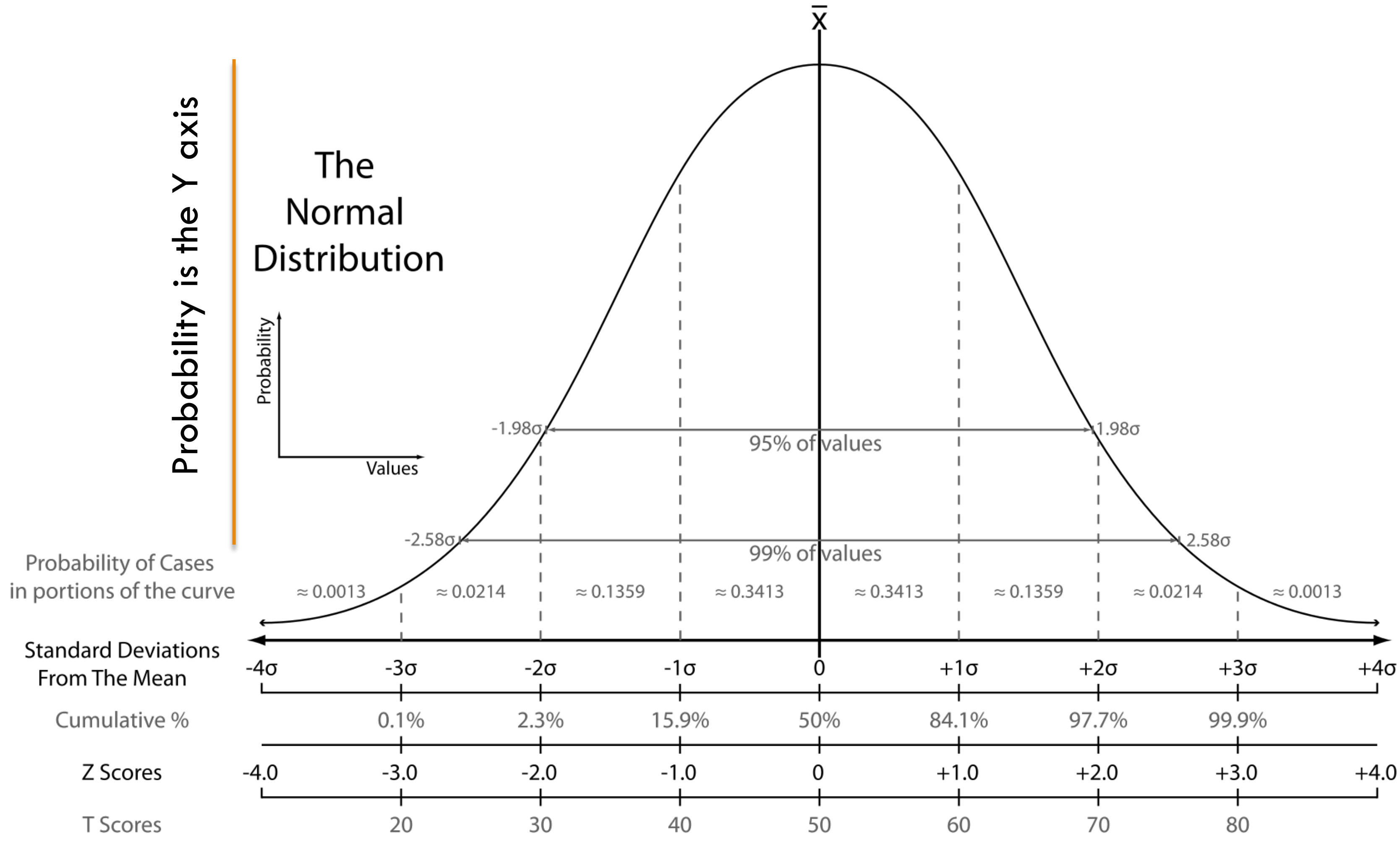


Skews vs Normality…

So if the income in America is skewed because of that top 1%, what does "normal" data look like?

Probability Distributions

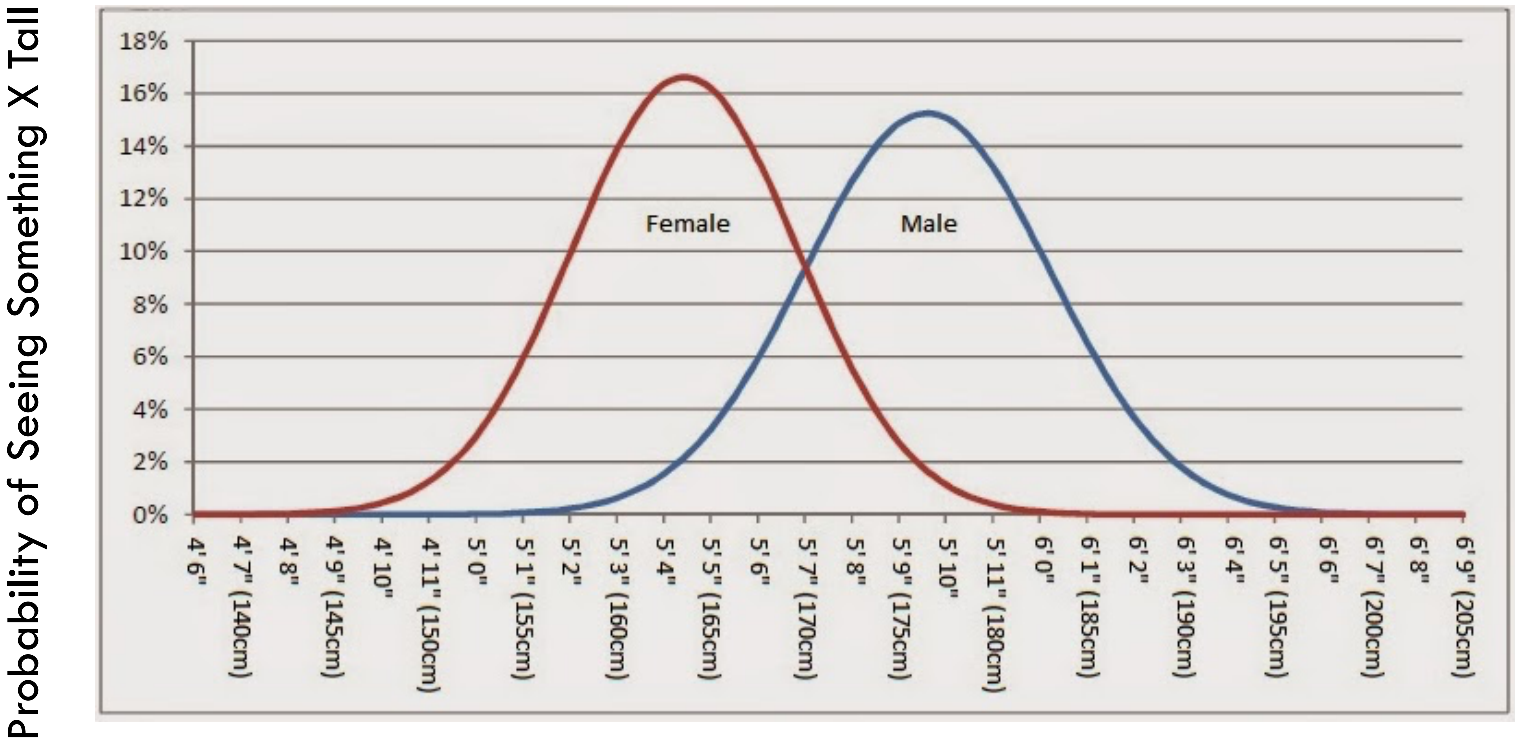
Behold! The Normal Distribution



What is “Normal?”

- Things that distribute “normally” are symmetrical, same amount below the mean as above the mean and is unimodal, meaning there is one big hump (the mode)
  - Natural Examples:
    - Human height, temperature, heart rate, blood-pressure
    - Delivery time, grades, guesses(?)
  - The typical value of something usually lingers (or clumps) around the mean and are more frequent.
    - Ex. The majority of females are around 5’4-ish with a few extremely tall or extremely short

Height by Gender



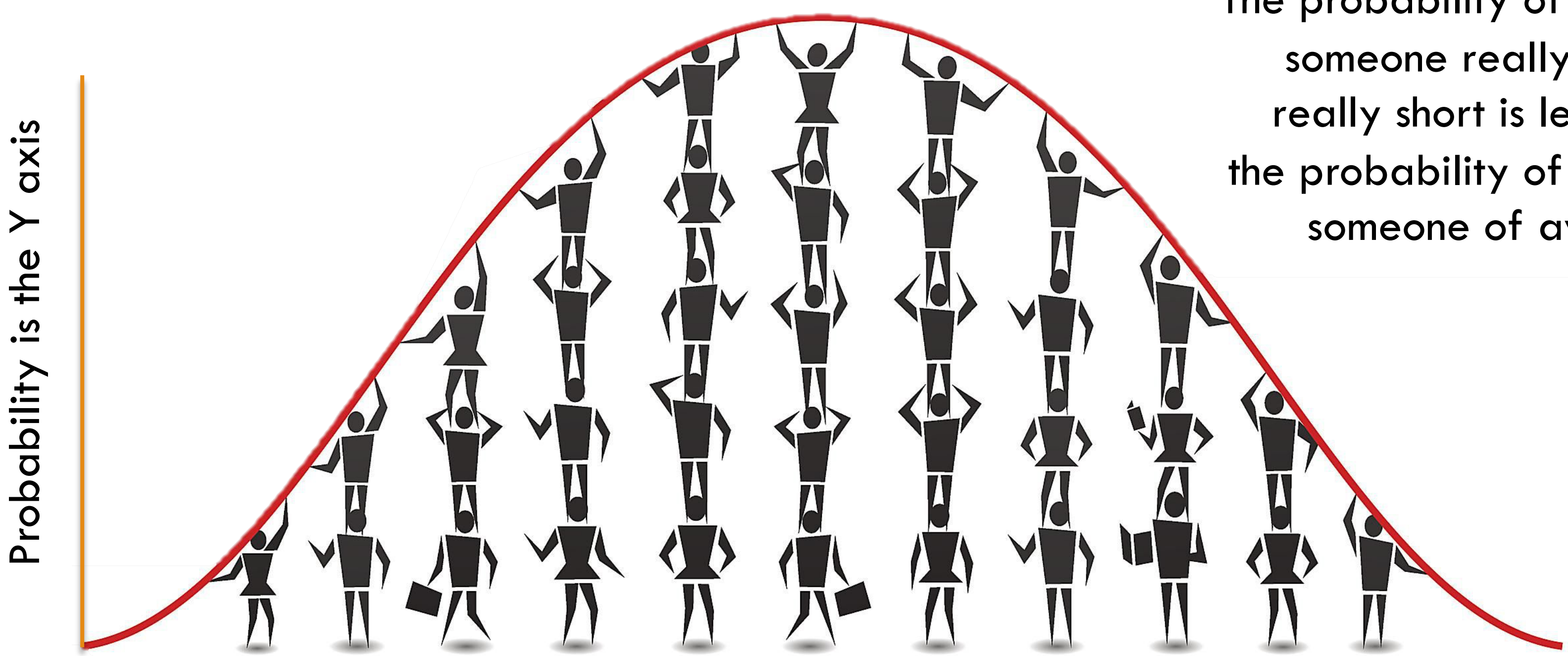
What do you notice in this graph? Variables?

What are the axes?

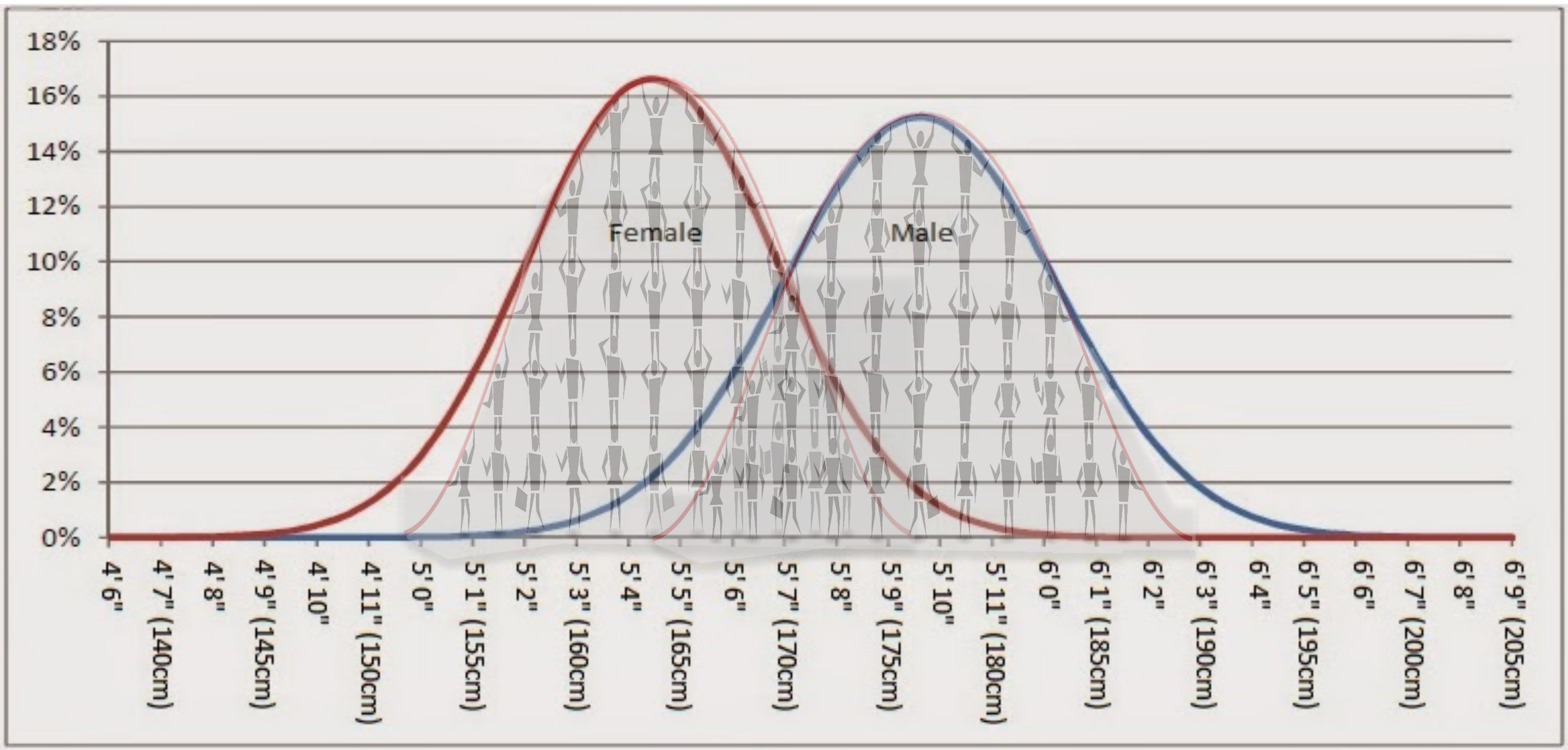
Piles of People

- Think of the curve as a pile of people…
  - Most people are piled up on top of each other in the middle while a few extremely low or extremely high cases are at the ends of the curve.

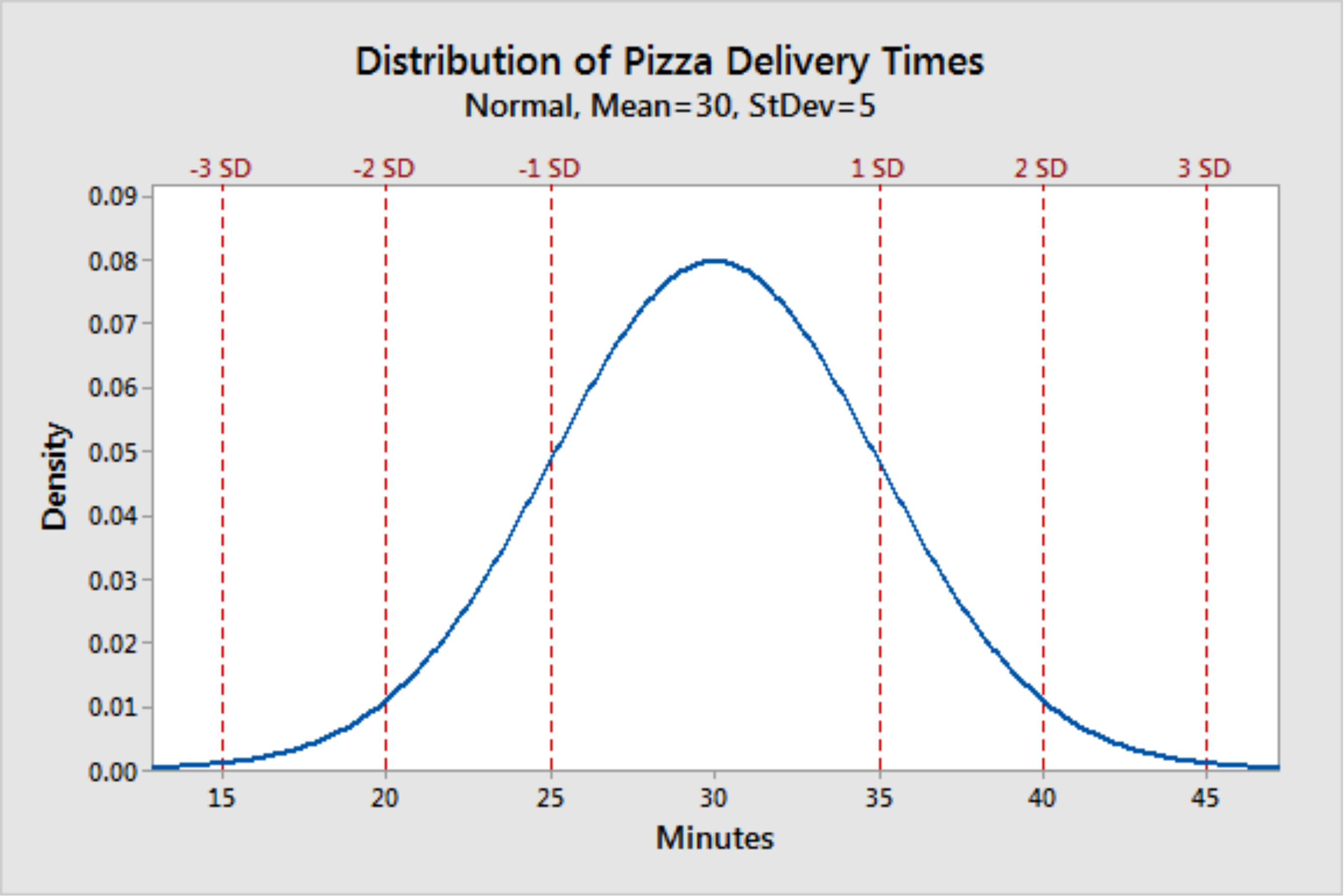
The probability of seeing someone really tall or really short is less than the probability of seeing someone of average height.



Height by Gender



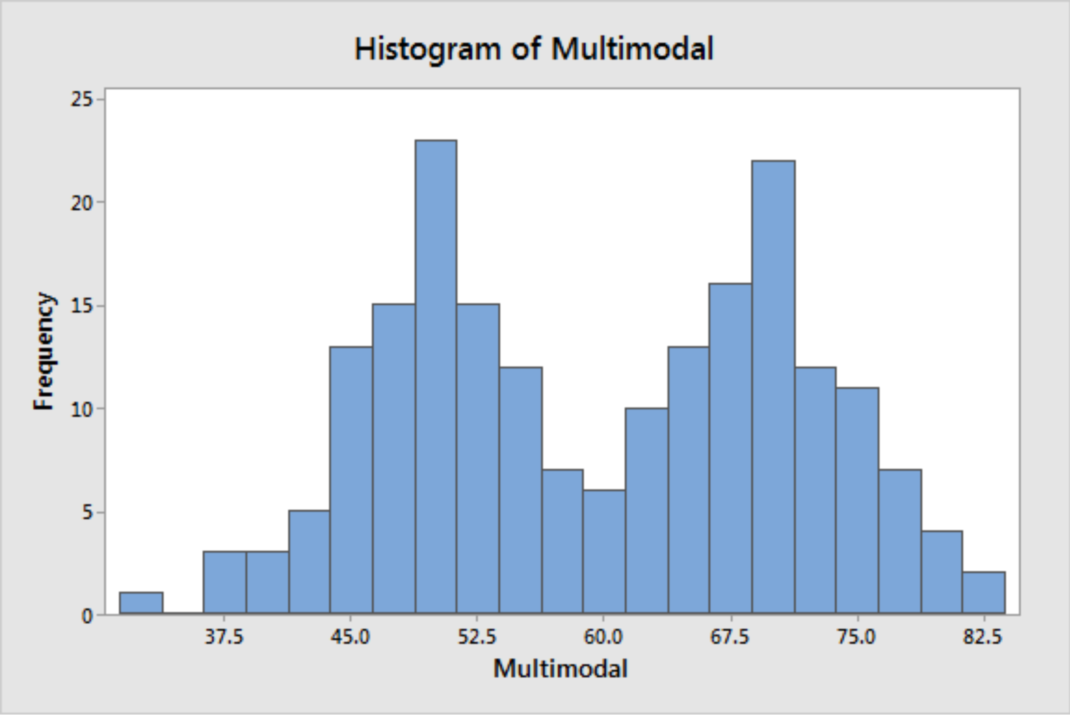
40 Minutes or Less or It's Free…



Bimodal Distributions

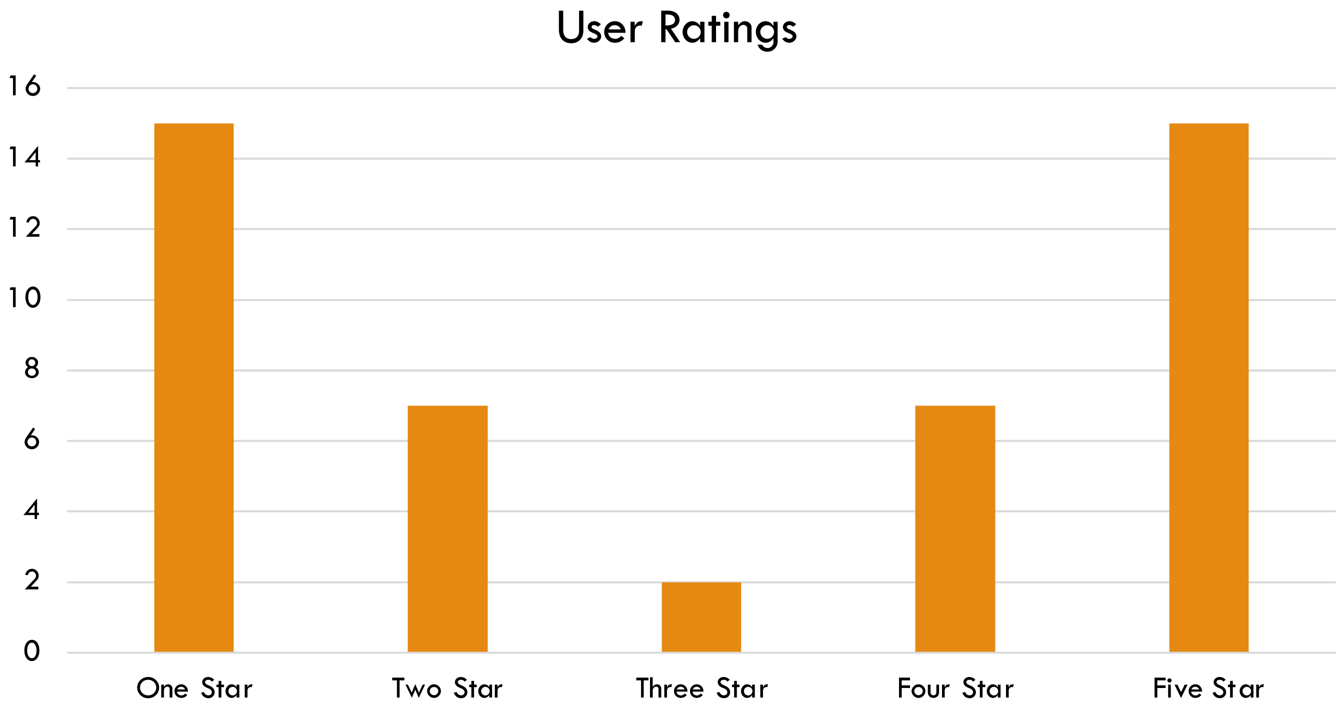
- Bimodal (or multimodal if more than 2)
  - Two distinct humps rather than one normal one
    - Two (or more) modes, the humps

What kind of data could produce this?



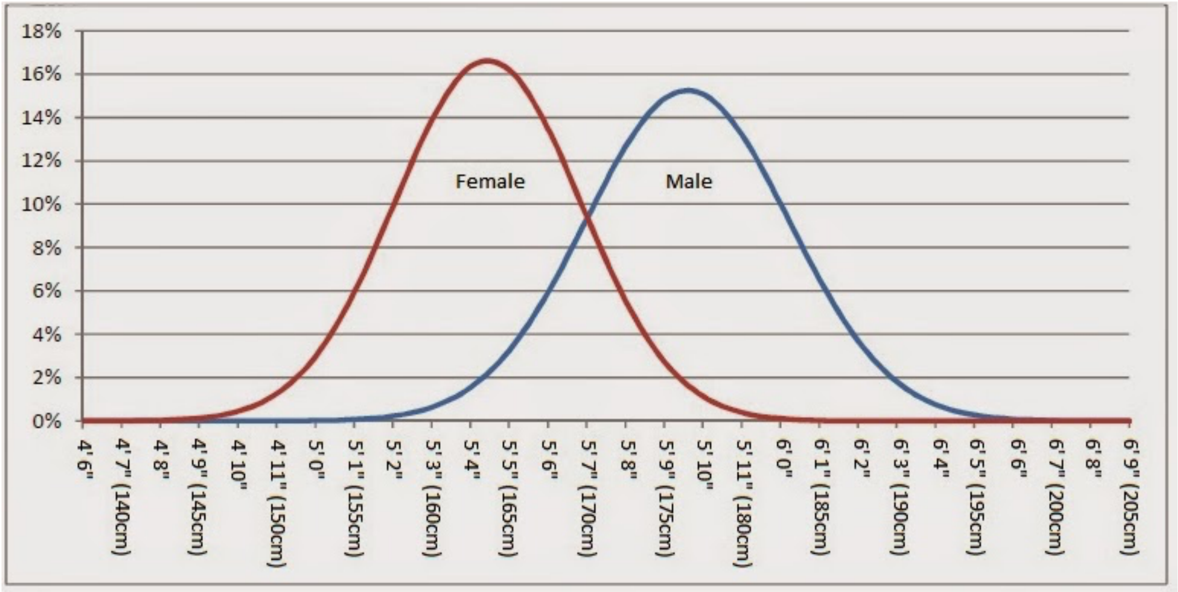
Amazon Reviews

- Bimodal
  - User rating can look like this usually because people who are extremely satisfied or extremely unsatisfied feel motivated to share their opinion.



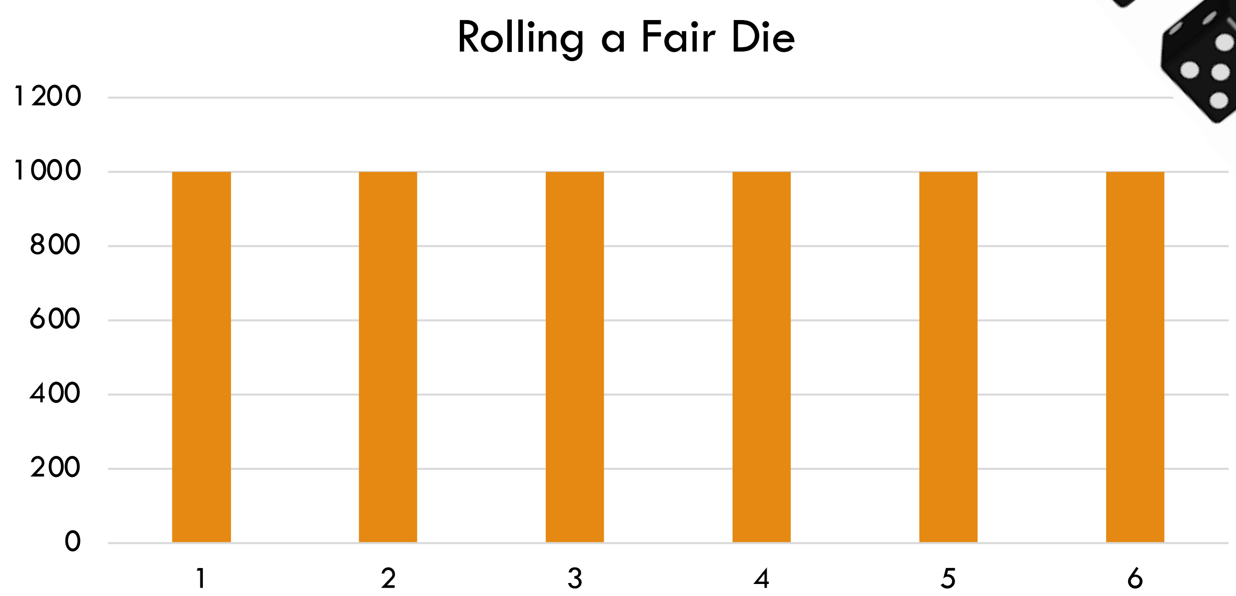
Bimodal Distributions

- Can also happen when you have two distinct groups answering the same questions.
  - Ex. Measuring height among both sexes
  - Ex. Clinical vs. Non-Clinical populations



Uniform Distributions

- Uniform distributions (also called rectangular) occur when all the possible values have equal likelihood of occurring
  - Like rolling a die



Probability Distributions

- There are many different types of distributions that are used for different types of data… These are just a few…
  - Normal Distribution ★
  - Binomial Distribution
  - Uniform Distribution
  - Poisson Distribution
  - Bernoulli Distribution

We'll focus on the Normal distribution for this course.

- All distributions help us to quantify and determine the *probability* of seeing a particular observations
  - Ex. The probability of seeing a woman that is 5'4 is about .18

Notation and Differentiations

Populations and Samples

- Now that we are starting to dive into numbers, we need to have a way to label them in such a way that we know if are talking about a sample or a population.
- When we design a study, we first define
  - The population of interest
    - Ex. What is the average level of stress for <u>all college students in America</u>?
- Reality Check: Can we ask every single college student in America their level of stress? No… Instead we must take a sample from the population
    - Ex. Sample 1,000 students from UT, St. Edward's, and ACC

Wording…

- We have different vocabulary for the numbers depending on if we are talking about a Population or a Sample

<u>P</u>arameters are for <u>P</u>opulations

<u>S</u>tatistics are for <u>S</u>amples

An average is example of a parameter for a population and statistic for a sample.

Who are we talking about?

- In statistics sometimes you will see common letters but sometimes you will see something that looks like Greek, which it is…
- These variable distinctions tell you whether you are talking about an entire population or just a small sample from the population.
- These distinction will become more important as we move through the course…
  - Equations change depending on whether you are working with an entire population or just a sample.

Who are we talking about?

Attribute	Population	Sample
Includes	Complete set	Subset of population
Mean	μ ("mu")	$\bar{x}$ ("x bar")
Sum of Squares	SS ("Sum of Squares")	SS ("Sum of Squares")
Variance	σ^2 ("sigma squared")	s^2 ("variance")
Standard Deviation	σ ("sigma")	s ("standard deviation")
Size	N	n
Numerical Descriptor	"Parameter"	"Statistic"

Up Next…

- We now know how to quantify the average value of a dataset, next we will quantify the average amount of *difference* in a dataset…

Variance

Calculating a Mean and Median in R

Calculating a Mean and Median in R

```r
############################
##### MEAN AND MEDIAN #######
############################
#### HISTOGRAM & BOXPLOT ####
############################

# Data from 20 women asking their height
height <- c(69, 63, 54, 61, 68, 61, 62, 56, 64, 66, 60, 61, 73, 63, 65, 72, 70, 59, 76, 59)

# Using the R function "mean()" we can quickly calculate the mean whic is 64.1 inches
mean(height)

# Using the R function "median()" we can quickly calculate the median whic is 63 inches
median(height)

# Here you can make a quick histogram
hist(height)

# And here a quick boxplot
boxplot(height)
```

Calculating a Mean and Median in R

```
##############################
##### MEAN AND MEDIAN #######
##############################
#### HISTOGRAM & BOXPLOT ####
##############################

# Data from 20 women asking their height
height <- c(69, 63, 54, 61, 68, 61, 62, 56, 64, 66, 60, 61, 73, 63, 65, 72, 70, 59, 76, 59)

# Using the R function "mean()" we c
mean(height)

# Using the R function "median()" we
median(height)

# Here you can make a quick histogra
hist(height)

# And here a quick boxplot
boxplot(height)
```