ECON 3818

Chapter 2

Kyle Butts

27 September 2021

Chapter 2: Describing Distribution with Numbers

Chapter Overview

Population vs. Sample

Measures of Central Tendency

- Mean
- Median

Measures of Variability

- Quartiles
- Variance and Standard Deviation

Population vs Sample

Population: the entire entities under the study

- Examples: all men, all NBA players, all children under 5

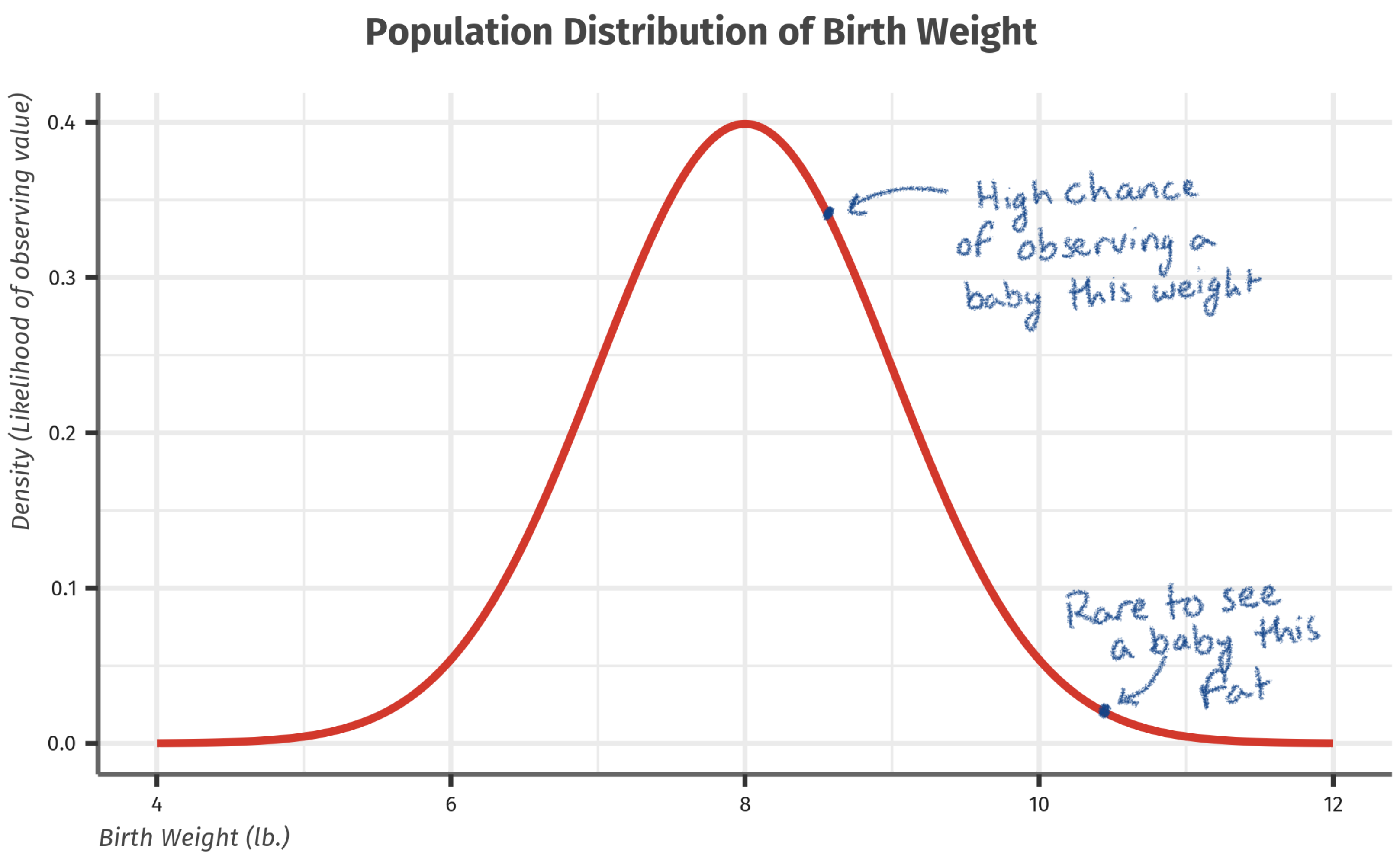
Sample: subset of the population

- Can be used to draw inferences about the population
- Examples: our class, Denver Nuggets players, daycares in Colorado
- Interested in parameters of the **population** distribution, we can estimate these parameters using data from **samples** since finding population parameters is infeasible

Population Distribution

Distribution of a variable: tells us *what values* it takes and *how often* it takes these values

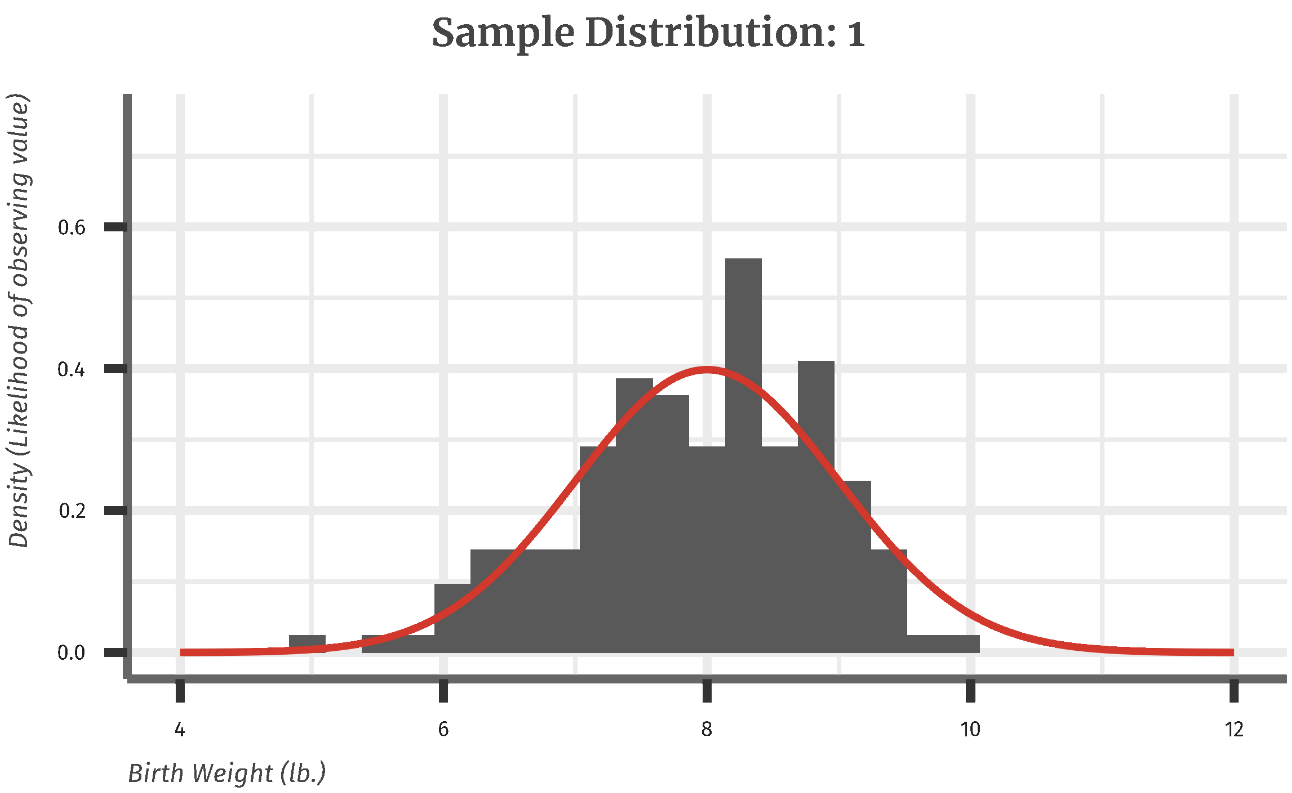
- We are interested in the underlying population distribution of some variable
- Fundamental problem of statistics is we can't collect data on every single observation



Population Distribution of Birth Weight
Density (Likelihood of observing value)
0.4
0.3
0.2
0.1
0.0
4
6
8
10
12
Birth Weight (lb.)
High chance of observing a baby this weight
Rare to see a baby this fat

Population Inference

What we do instead is use a sample of the population and use that sample distribution to determine parameters of interest



Parameters of Interest

Two primary **population** parameters of interest:

- Measures of central tendency:
  - Population mean, μ
  - Population median
- Measures of variability:
  - Population variance, σ^2

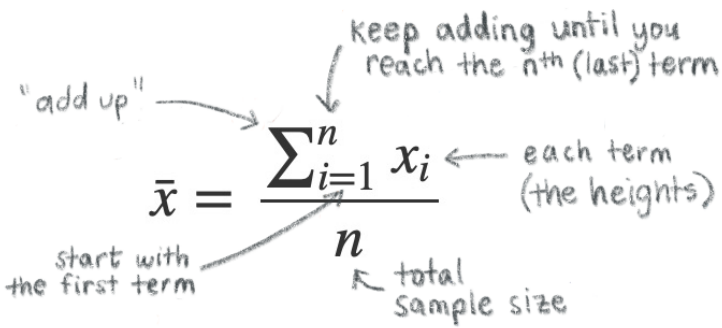
We will *estimate* these using the **sample** distribution

Measuring Center: the Mean

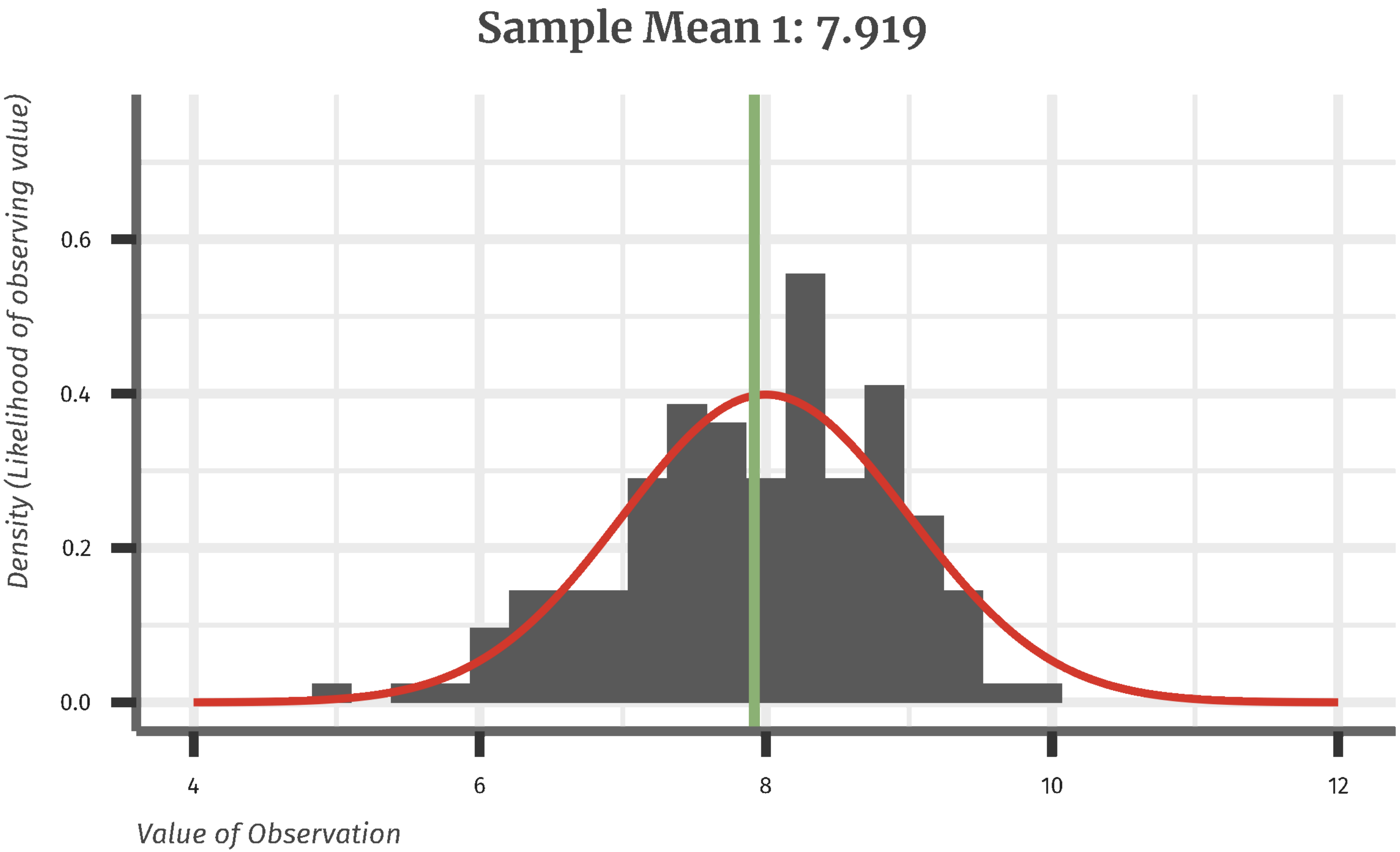
The most common measure of center is the arithmetic average, or **mean**

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

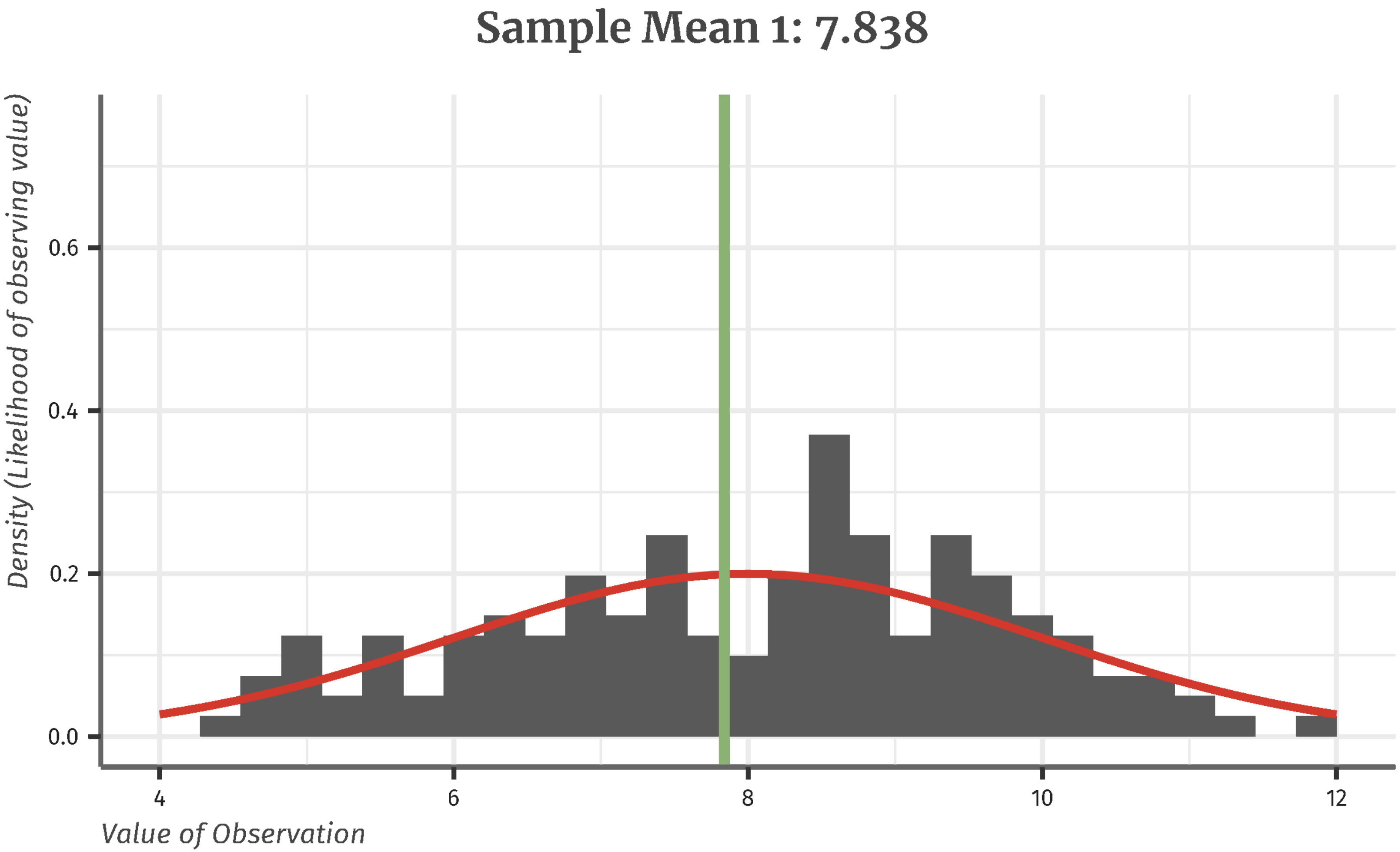
or more compactly:



Population Inference: Mean



Population Inference: Mean



Measuring Center: the Median

The median is the midpoint of a distribution

- Is more resistant to the influence of **extreme observations**

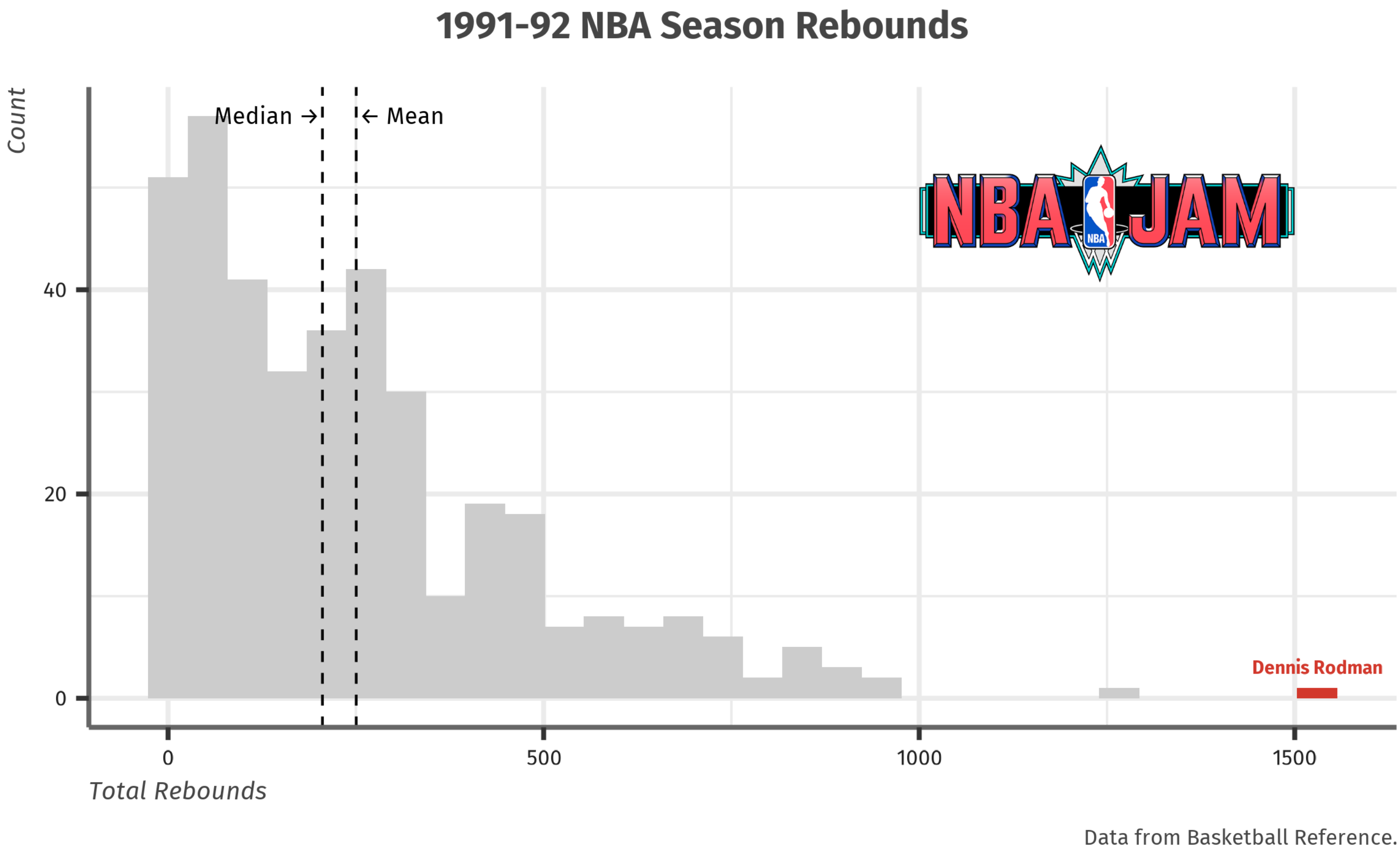
How to calculate median:

- Arrange observations from smallest to largest
- If there is odd number of observations, the median is the center observation. If there are even number of observations, the median is the average of two center observations

Mean vs. Median

- Although we will primarily be using the mean throughout the semester, the biggest drawback of the mean is that it is not resistant to **outliers**
- The median, however, is resistant to **outliers** so it can be important to calculate for smaller samples

Mean vs. Median Example



Median: 205.5 rebounds and **Mean**: 250.5 rebounds

Clicker Question

What is the sample mean of the participants's age?

Sample of individuals

AGE	SEX	BMI	DRINKS PER WEEK
59	male	32.26	3 drinks
62	male	25.09	2 drinks
60	female	32.58	1 drink
18	male	99.99	6 drinks
57	female	31.88	2 drinks
56	male	42.80	3 drinks

a. 58
b. 51.2
c. 52
d. 49.7

Clicker Question

Which measure of central tendency best describes the age of participants?

Sample of individuals

AGE	SEX	BMI	DRINKS PER WEEK
59	male	32.26	3 drinks
62	male	25.09	2 drinks
60	female	32.58	1 drink
18	male	99.99	6 drinks
57	female	31.88	2 drinks
56	male	42.80	3 drinks

a. Median

b. Mean

Measuring Variability

Measures of central tendency do not tell the whole story. To further characterize the distribution, we need to know how the data is spread out

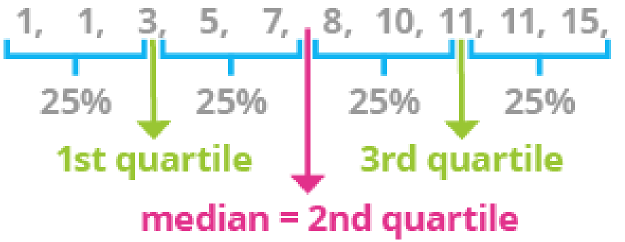
- Quartiles
- Variance

Variability: Quartiles

Measure of center alone can be misleading. One way to measure variability is to use quartiles.

How to calculate quartiles:

- Arrange observations in increasing order and locate median
- The first quartile is the median of the observations located to the left of the median
- The third quartile is the median of observations located to the right of the median



Boxplots

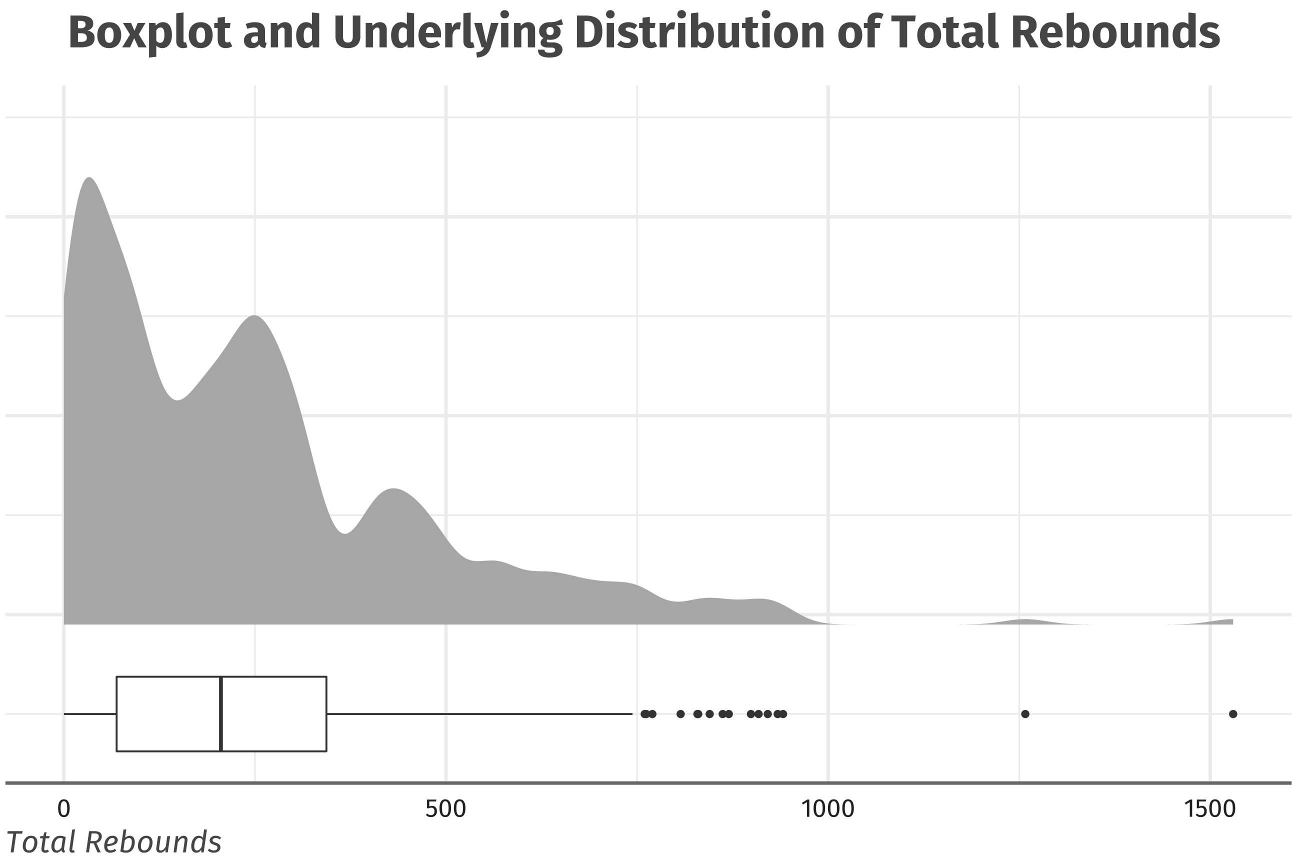
Five-number summary: smallest observation (minimum), the first quartile, the median, the third quartile, and the largest observation (maximum)

We can use the **boxplot** using this five number summary to display quantitative data

How to make a boxplot:

- A central box spans the first and third quartiles
- A line in the box marks the median
- Line extends from the box out to the smallest and largest observations

Boxplots



Boxplot and Underlying Distribution of Total Rebounds
0
500
1000
1500
Total Rebounds

Interquartile Range

The **interquartile range**, IQR, is the distance between the first and third quartiles

- IQR = $Q_3 - Q_1$
- The IQR measures the spread of the data and it also helps to identify outliers

Rule for outliers:

- An observation is an outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first

Variability: Variance

Variance: denoted, s^2, measures how "spread out" the data are on average

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1},$$

or more compactly

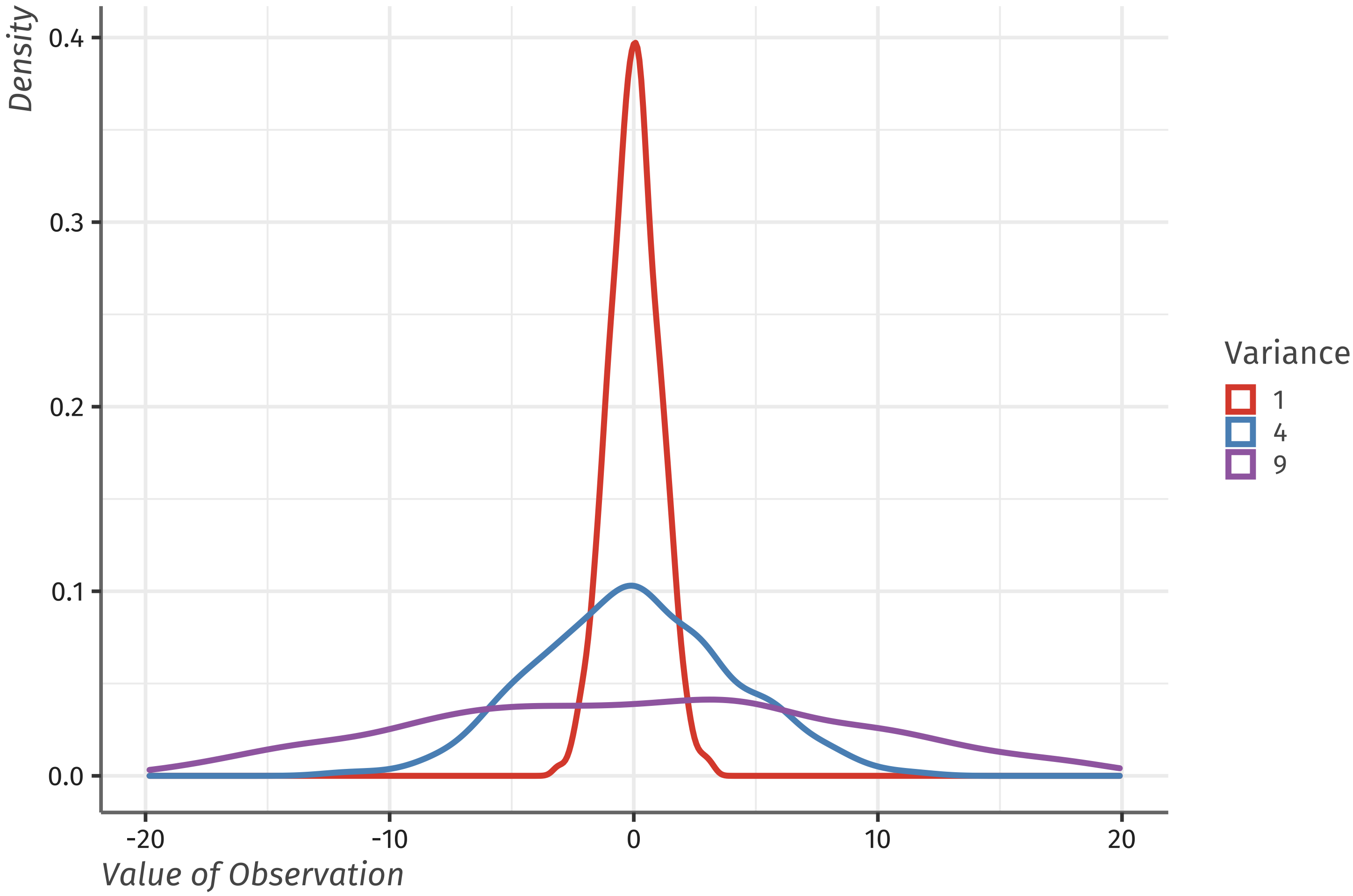
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

sum over sample

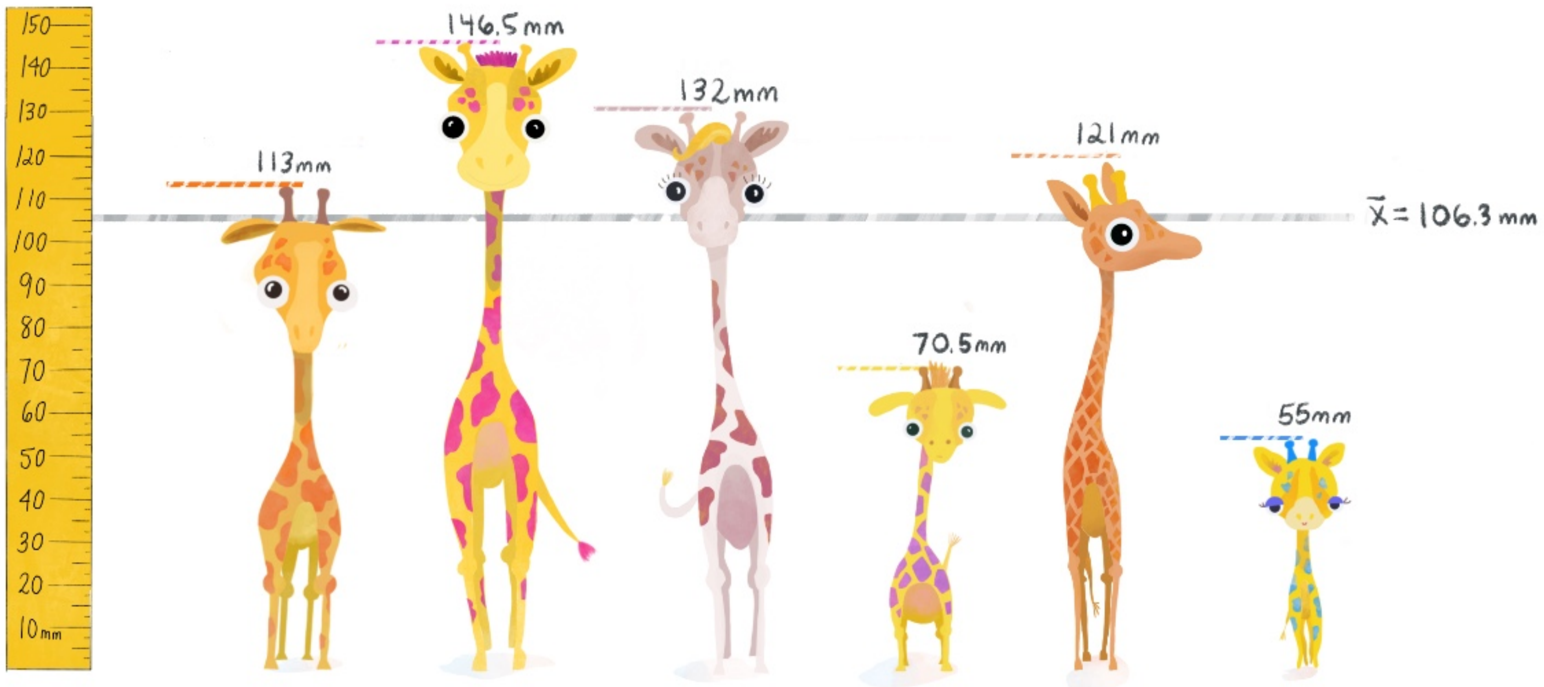
degrees of freedom

Deviations from mean

Visualizing Variance



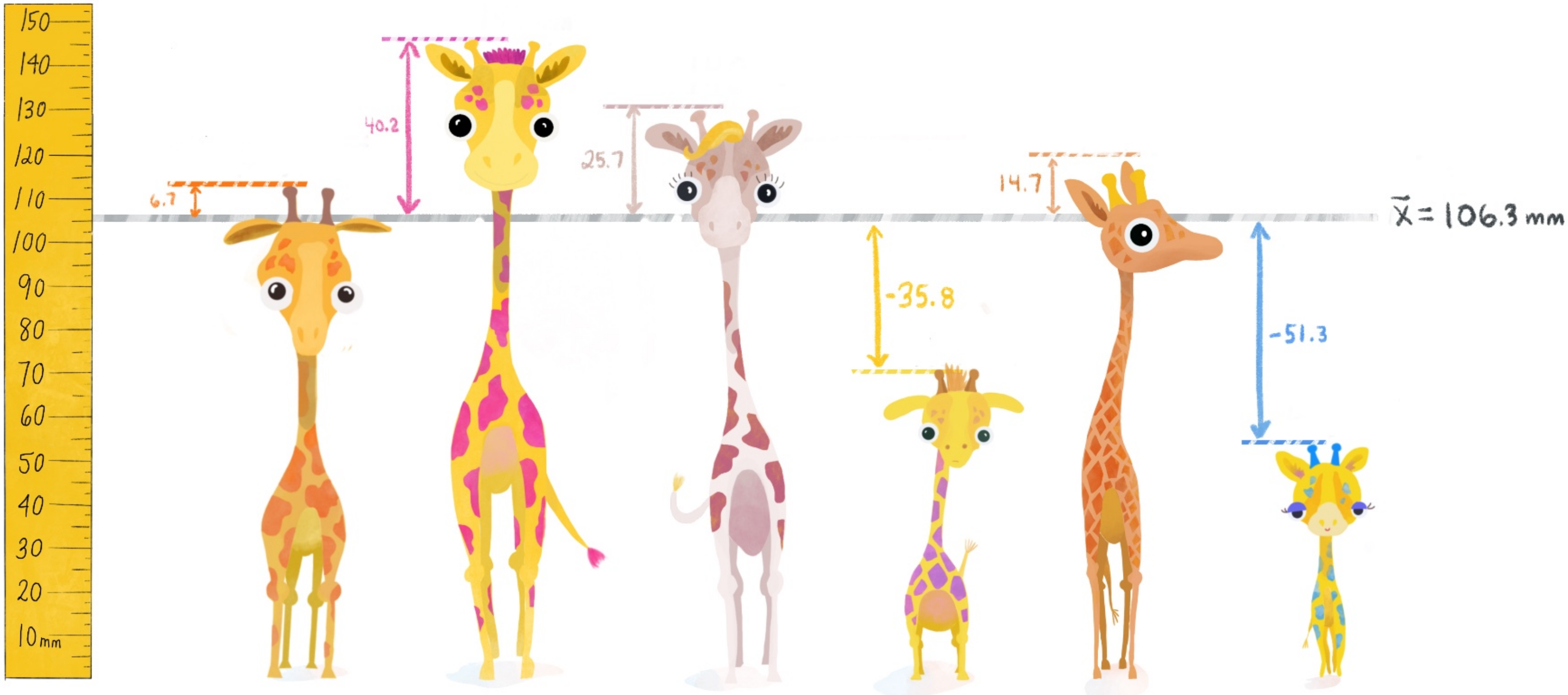
Example



1. Calculate the mean height in sample

Figure from Teacups, Giraffes, & Statistics

Example



2. Calculate deviations from mean
3. Square and sum

Figure from Teacups, Giraffes, & Statistics

Variability: Standard Deviation

Standard deviation: looks at how far each observation is from the mean; square root of the variance

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{s^2}$$

- $n-1$ is referred to as the degrees of freedom
- s measures variability about the mean
  - More variable $\implies$ larger s
- s is always greater than or equal to zero, but usually > 0
  - When would it be $= 0$?
- s is not resistant to outliers.

Practice Question

Calculate the standard deviation of age?

Sample of individuals

AGE	SEX	BMI	DRINKS PER WEEK
59	male	32.26	3 drinks
62	male	25.09	2 drinks
60	female	32.58	1 drink
18	male	99.99	6 drinks
57	female	31.88	2 drinks
56	male	42.80	3 drinks

Summary of Summary Statistics

Two basic ways to summarize the center and spread of a distribution

- Mean and standard deviation (or variance)
- The five-number summary

When to Use Which

Use $\bar{x}$ and s when the distribution is reasonably symmetric and free of outliers

Use five-number summary if distribution is skewed, or has outliers

Greek Letters and Statistics

Latin Letters

- Latin letters like $\bar{x}$ and s^2 are calculations that represent guesses (estimates) at the population values.

Greek Letters

- Greek letters like μ and σ^2 represent the truth about the population.

The goal for the class is for the latin letters to be good guesses for the greek letters:

$$\text{Data} \longrightarrow \text{Calculation} \longrightarrow \text{Estimates} \longrightarrow^{hopefully!} \text{Truth}$$

For example,

$$X \longrightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \bar{x} \longrightarrow^{hopefullly!} \mu$$