Numerical Summaries

Grinnell College

September 16, 2024

Review

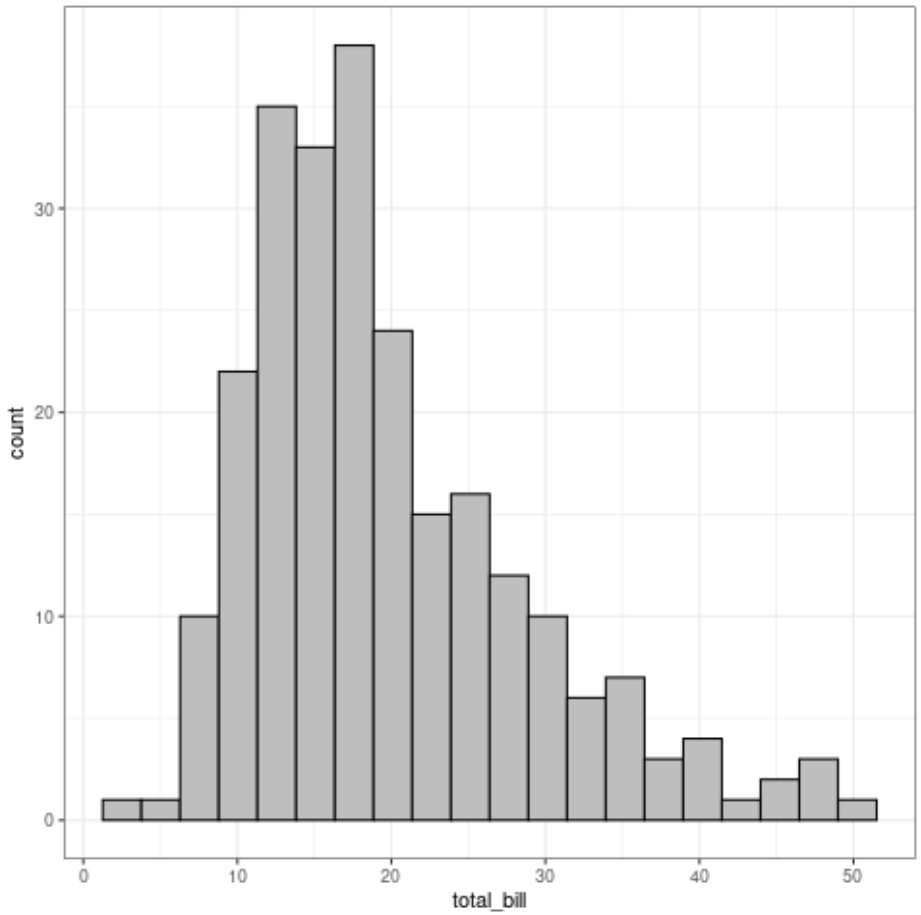
Graphical Summaries:

- Why do we create graphs?
- Types of plots
- Describing plots

We will now spend some more time on describing quantitative variables

Review – One Quantitative Variable

We've seen a few examples of histograms at this point.



Some of the things we've thought about:

- What are the most common values?
- How spread out is this data?
- What about the range of this data?
- Does it appear skewed or symmetric?

Review

We spent some class time over the last 2 weeks learning how to talk about our data. Largely our time has been spent on both of these:

1. **Data Visualization** displaying data in ways that make patterns more noticeable
2. **Numerical Summaries** calculating numbers that tell us about certain aspects of the data

Often these concepts ended up getting combined in that we would use numerical summaries to help us describe our graphs and plots. We are going to spend a bit more time today learning about numerical summaries.

Review – Quantitative Distribution

We need to mention all of the following things when we describe the **distribution** of a quantitative variable:

- **Shape** - how does the distribution look?
  - - symmetric?
  - - skewed?
  - - # of modes
- **Center** - where does the data bunch up
- **Spread** - how spread out is the data
- **Outliers** - are there values that are much smaller/larger than the rest?

Numerical Summaries

The **center** of a quantitative variable is meant to help us answer

- What are the most common values?
- Where does the data bunch up?
- What is the 'typical' value that most observations had?

The **spread** of a distribution is meant to tell us, literally, how spread out the values are. **Spread** also tells us how much *variability* there is

There are two separate approaches for describing center and spread

1. Order Statistics
2. Moment Statistics

Order Statistics

Order statistics are numerical summaries based on the ordered ranking of a quantitative variable (smallest to largest)

There are a few properties in particular that make order statistics useful:

1. They make no assumptions about how the data is distributed
2. Are generally robust to (unaffected by) major fluctuations in the data (i.e., outliers)
3. Easier to interpret

Review – Percentiles

A **percentile** α is a number such that $\alpha\%$ of our (quantitative) observations fall at or below this number when ranked from smallest to largest

Some percentiles have special names. The *median*, for example, is the 50th percentile.

Other notable percentiles include:

1. Minimum
2. 25th percentile or **first quartile** (Q_1)
3. 75th percentile or **third quartile** (Q_3)
4. Maximum

Median (Center)

The **median** is another name for the 50th percentile. It is frequently used as a measure of center.

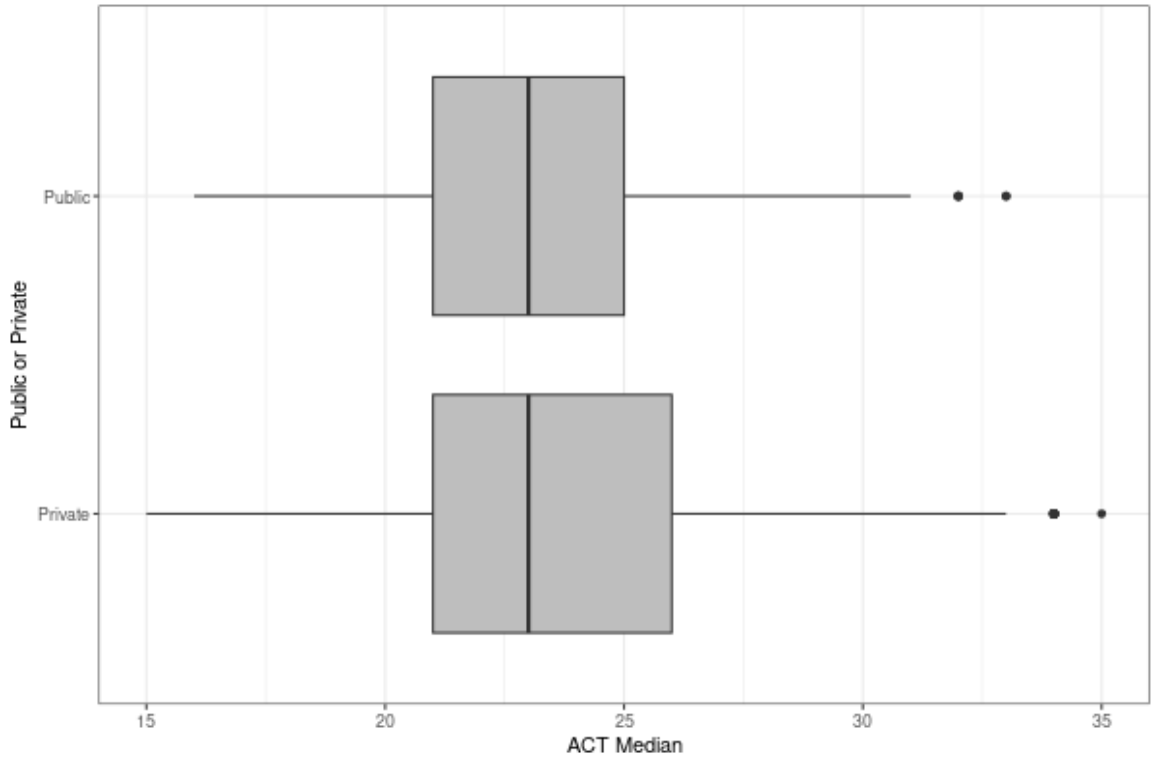
These are some other ways to think about the **median**.

- the **median** divides the data into an upper and lower half
- the middle value of the data if arranged from smallest to largest
- about half the data is larger than the median and about half the data is smaller

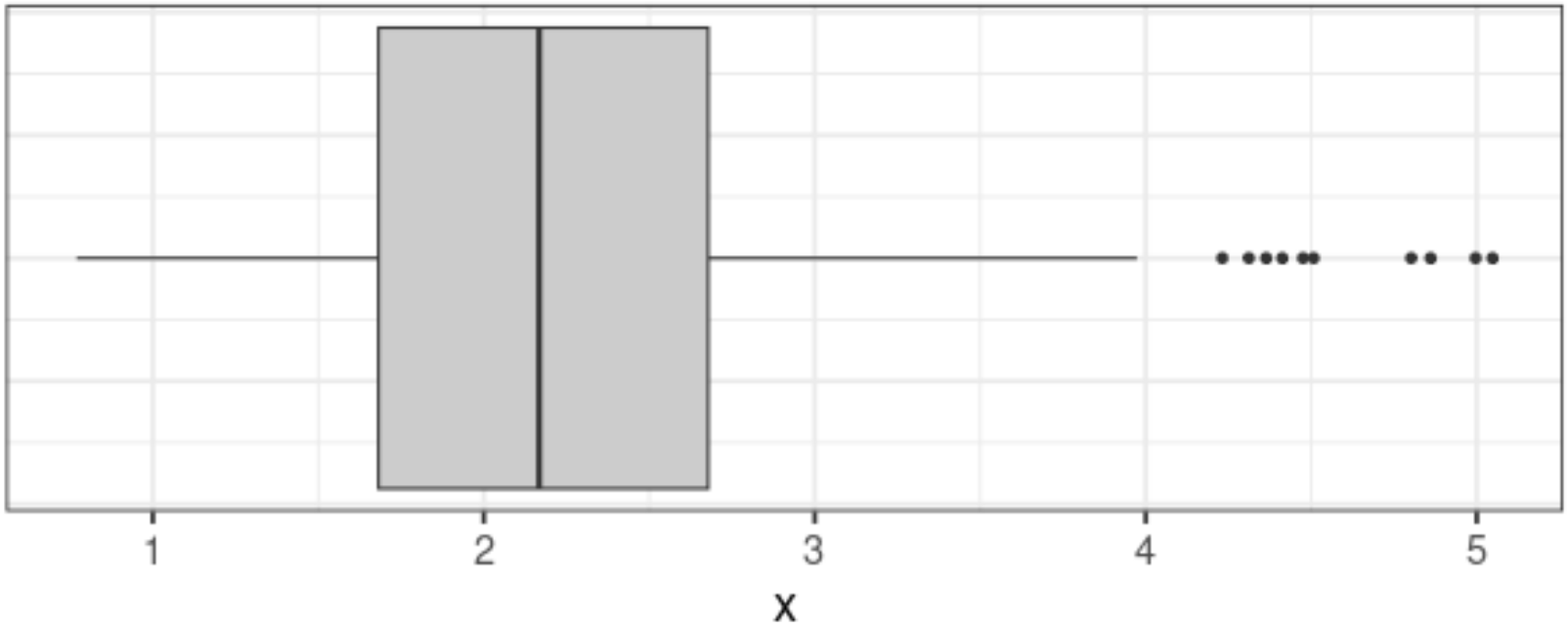
IQR (Spread)

The IQR is a measure of the spread of our data.

- calculation is Q3 - Q1
- measures the range of the middle 50% of the data



Five Number Summary



- Median
- 25th Percentile (Q_1)
- 75th Percentile (Q_3)
- Minimum or 1.5 $\times$ IQR
- Maximum or 1.5 $\times$ IQR
- Outliers

Moment Statistics

Moment statistics are statistics that are based on specific mathematical properties of our data. There are some very powerful math tools that can make use of these (we will see some shortly)

Unlike order statistics, moment statistics (largely) *do* make assumptions about how the data is distributed: as such, they can be very sensitive to unexpected fluctuations such as outliers

In this sense, we say that moment statistics *are not* robust

Some Notation

Before we continue, it is helpful to introduce some notation that can make doing calculations with data a little easier.

n is often used to denote the sample size (# of observations in our sample)

$\mathbf{x}_i$ is used to denote the values for a variable in our data set
i.e.: $\mathbf{x}_3$ is the 3rd value of the variable in the data set

$\sum$ is a symbol frequently used to denote adding a whole bunch of things together.

Mean (Center)

The **mean** is the same thing as the **average** value of the variable.

To find the value of the mean, we add up all the values of the variable and divide by the number of observations.

Often the *sample* **mean** of a variable is denoted as $\overline{x}$

Using the notation from the previous slide, the equation for the **mean** is

$$\overline{x} = \frac{\sum x_i}{n}$$

Spread – Standard Deviation

Standard Deviation – a way of measuring the typical deviation (distance) of each observation from the *mean*

- the symbol **s** is often used to denote the standard deviation of our sample (sample standard deviation)

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

Why do we take the square root? Taking the square root ensures that the standard deviation has the same units as the original variable

Why do we use n-1 and not n? It's complicated. Using n-1 gives us better estimates and predictions

Standard Deviation

Some properties of the **standard deviation**:

- measures spread (variability) from the mean
  - values close to the mean = smaller contribution to s
  - values far away from the mean = larger contribution to s
- cannot be negative ($s \geq 0$)
- has the same units as the original variable

You may hear the word **variance**.

- variance = s^2
- harder to interpret
- certain math scenarios where it is easier to work with than s

Which measures to use?

The *shape* of the distribution, as well as whether we have *outliers* will determine whether we use **order statistics** (median and IQR) or **moment statistics** (mean and standard deviation) to describe the center and spread

In general we prefer to use **moment statistics** (mean and standard deviation) if we can, but there are certain situations where the mean and standard deviation are not good measures of center and spread

Which measures to use?

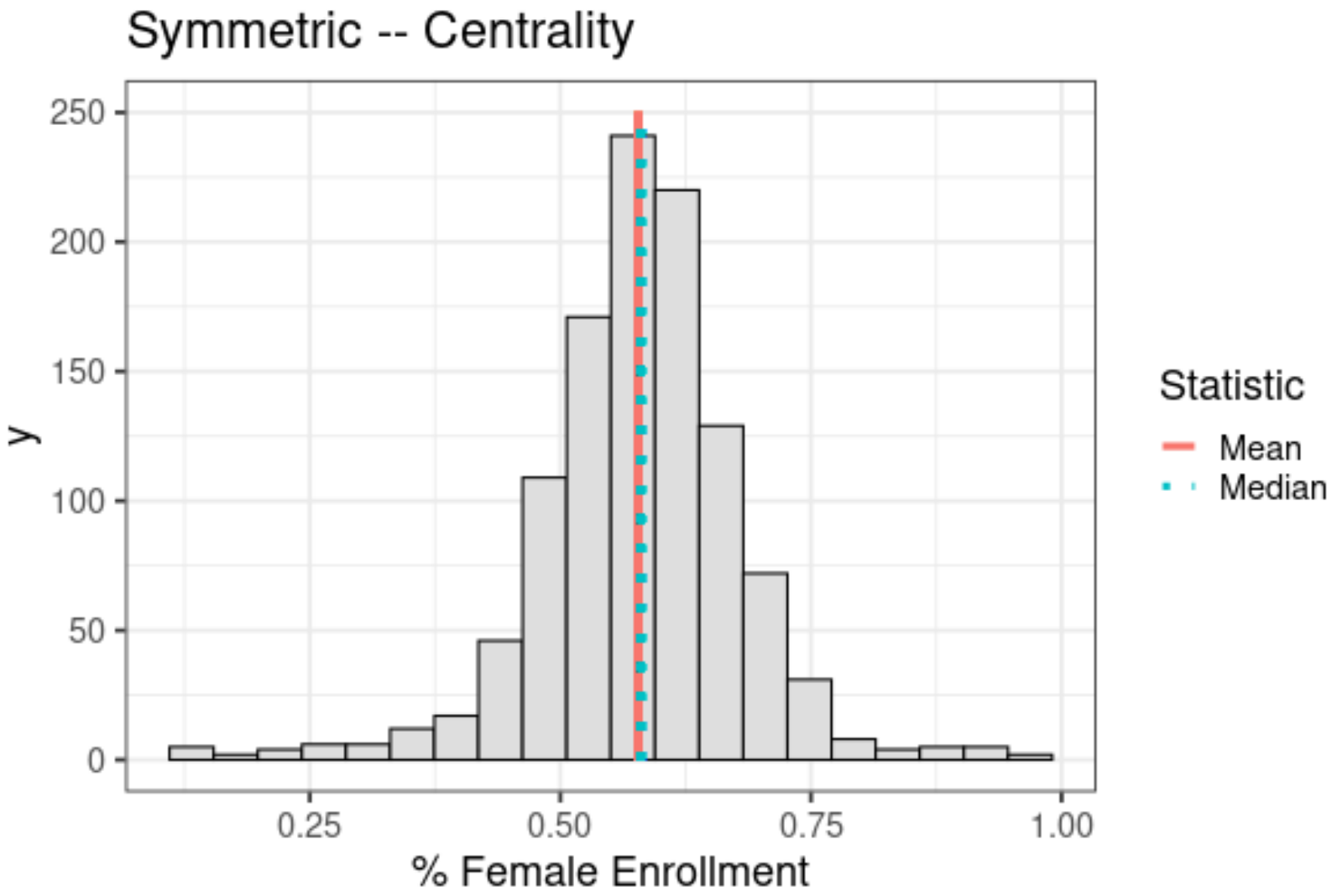
Order statistics are robust, moment statistics are not robust.

- A skewed distribution can affect the mean and std. dev. a lot
  - skew $\rightarrow$ mean & std. dev. not good measures of center & spread
- Outliers can affect the mean and std. dev. a lot
  - outliers $\rightarrow$ mean & std. dev. not good measures of center & spread

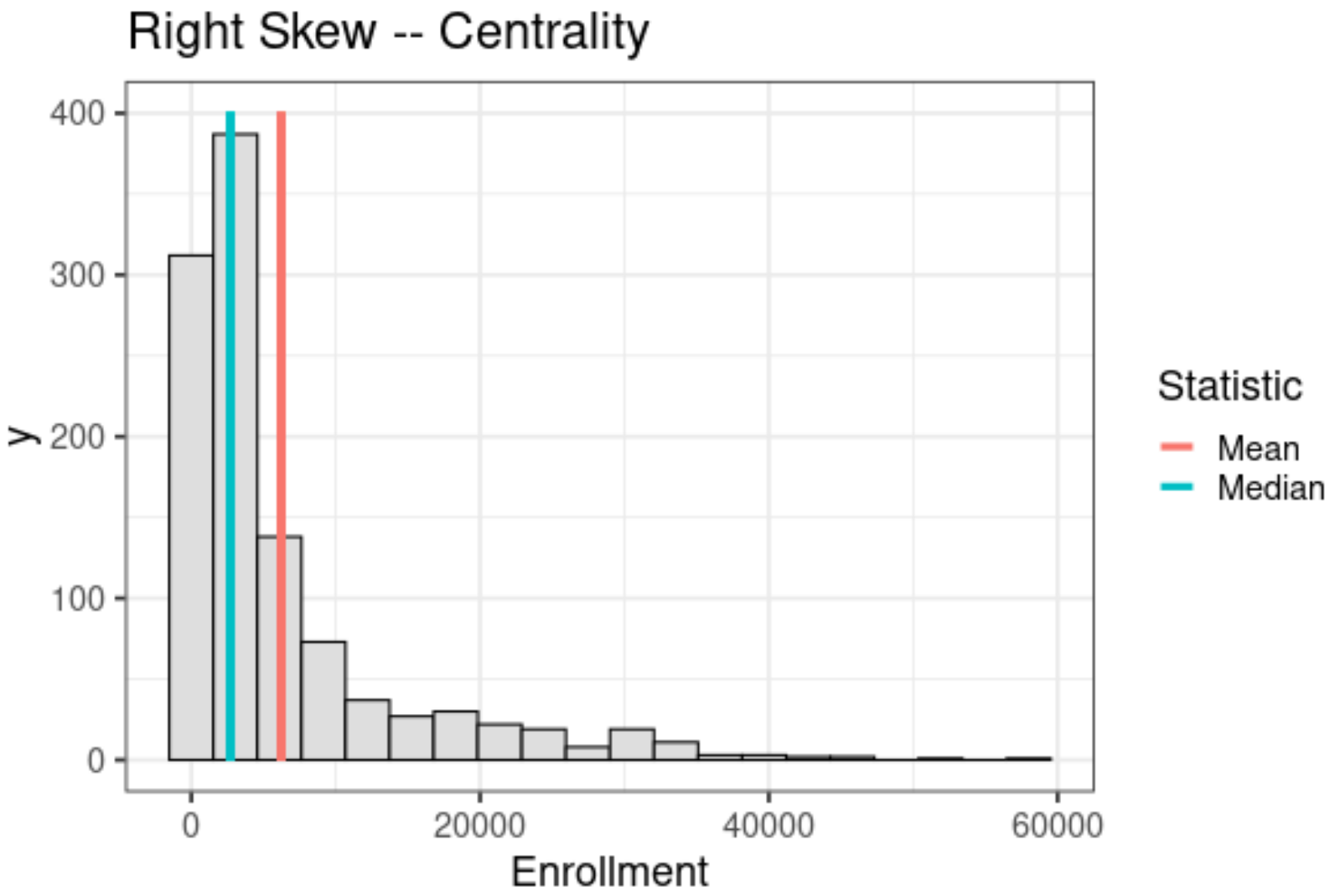
Summary:

Symmetric shape with no 'extreme' outliers $\rightarrow$ mean and std. dev.
Skewed shape or outliers (or both) $\rightarrow$ median and IQR

Comparing Mean with Median



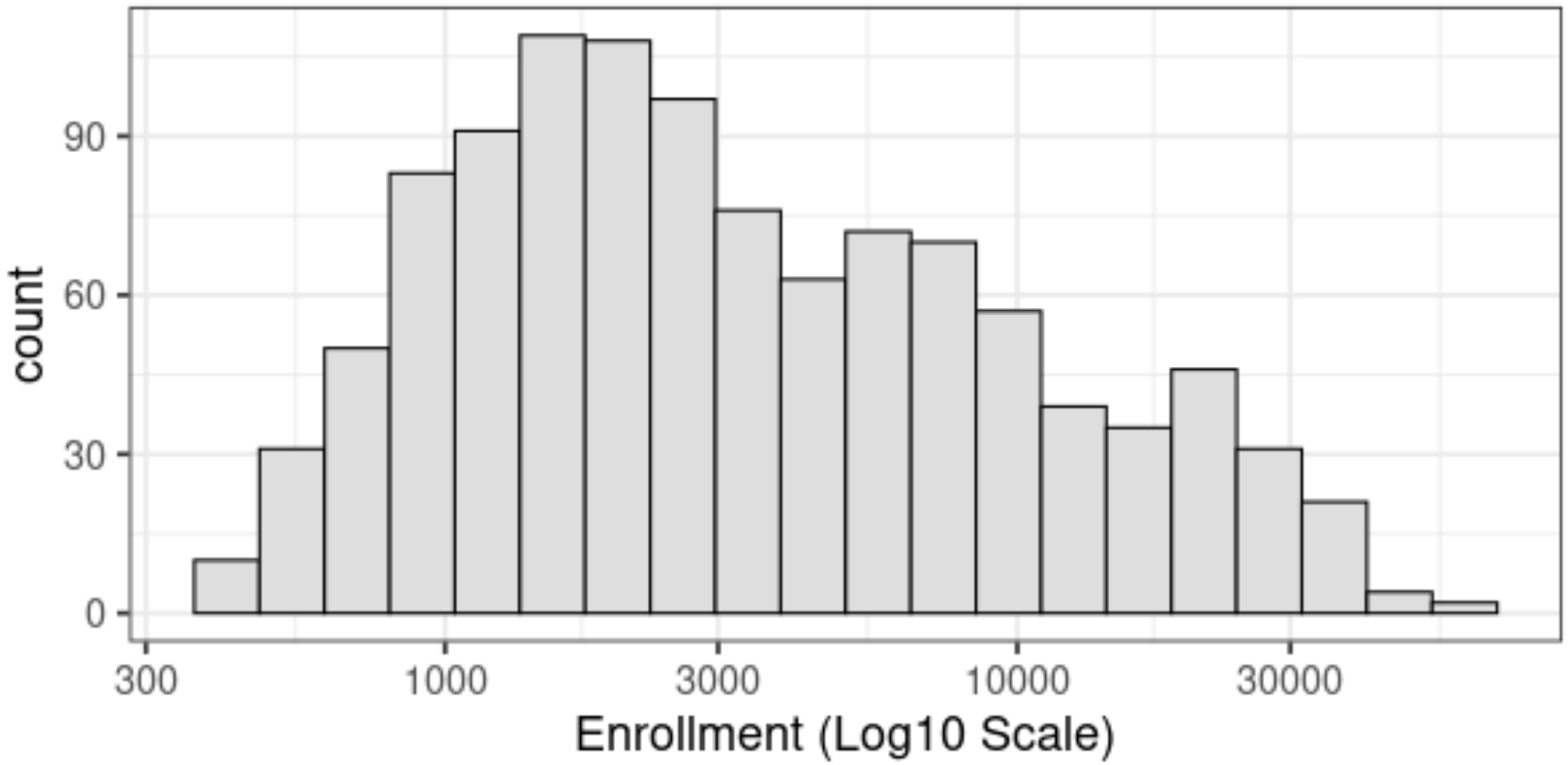
Comparing Mean with Median



Practice

For each of the following variables visualized below:

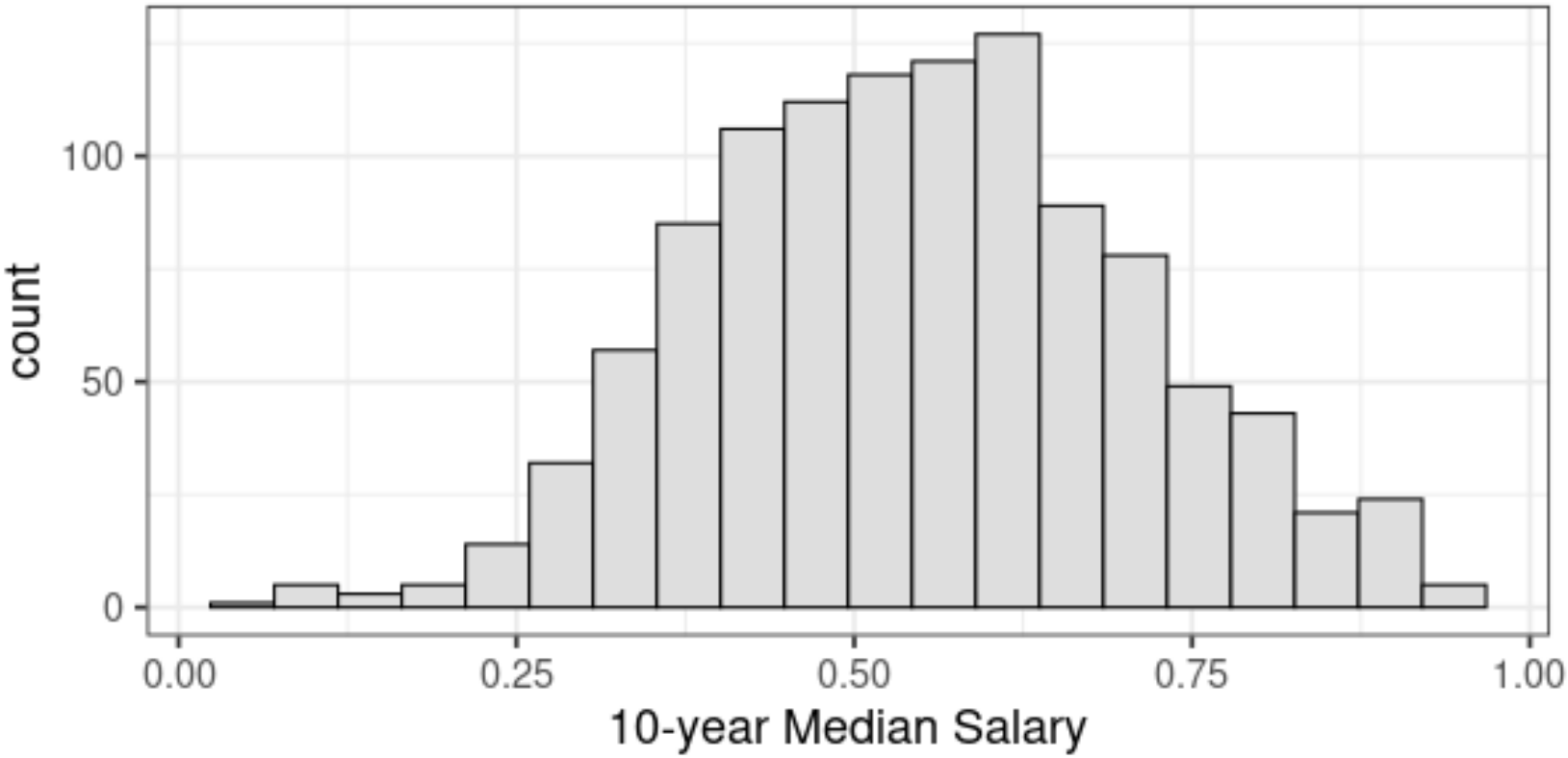
1. Determine approximate mean and median and which should be larger. How do you know?
2. Decide whether standard deviation or IQR is more appropriate for describing variability



Practice

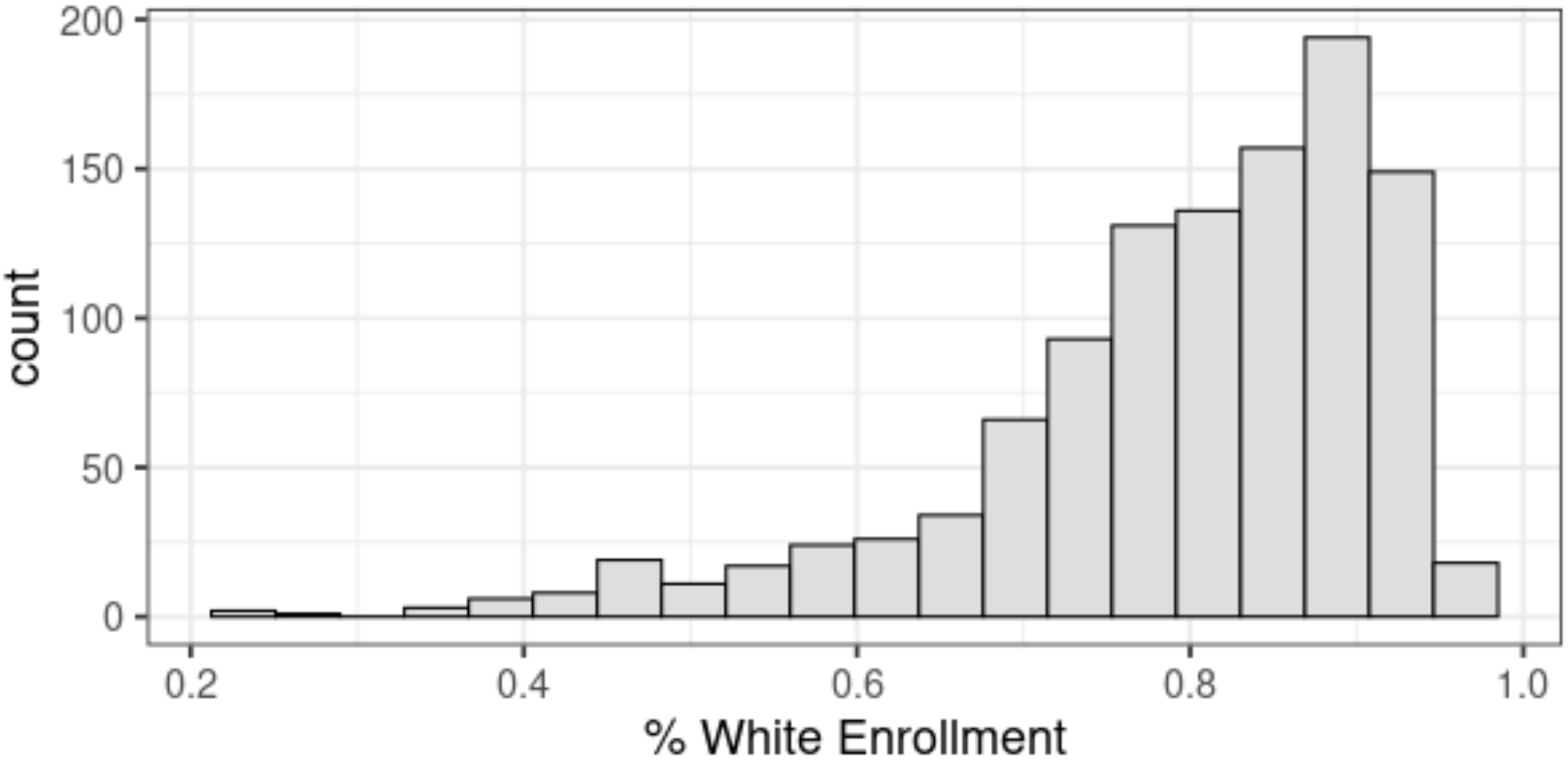
For each of the following variables visualized below:

1. Determine approximate mean and median. Are they very different?
2. Decide whether standard deviation or IQR is more appropriate for describing variability



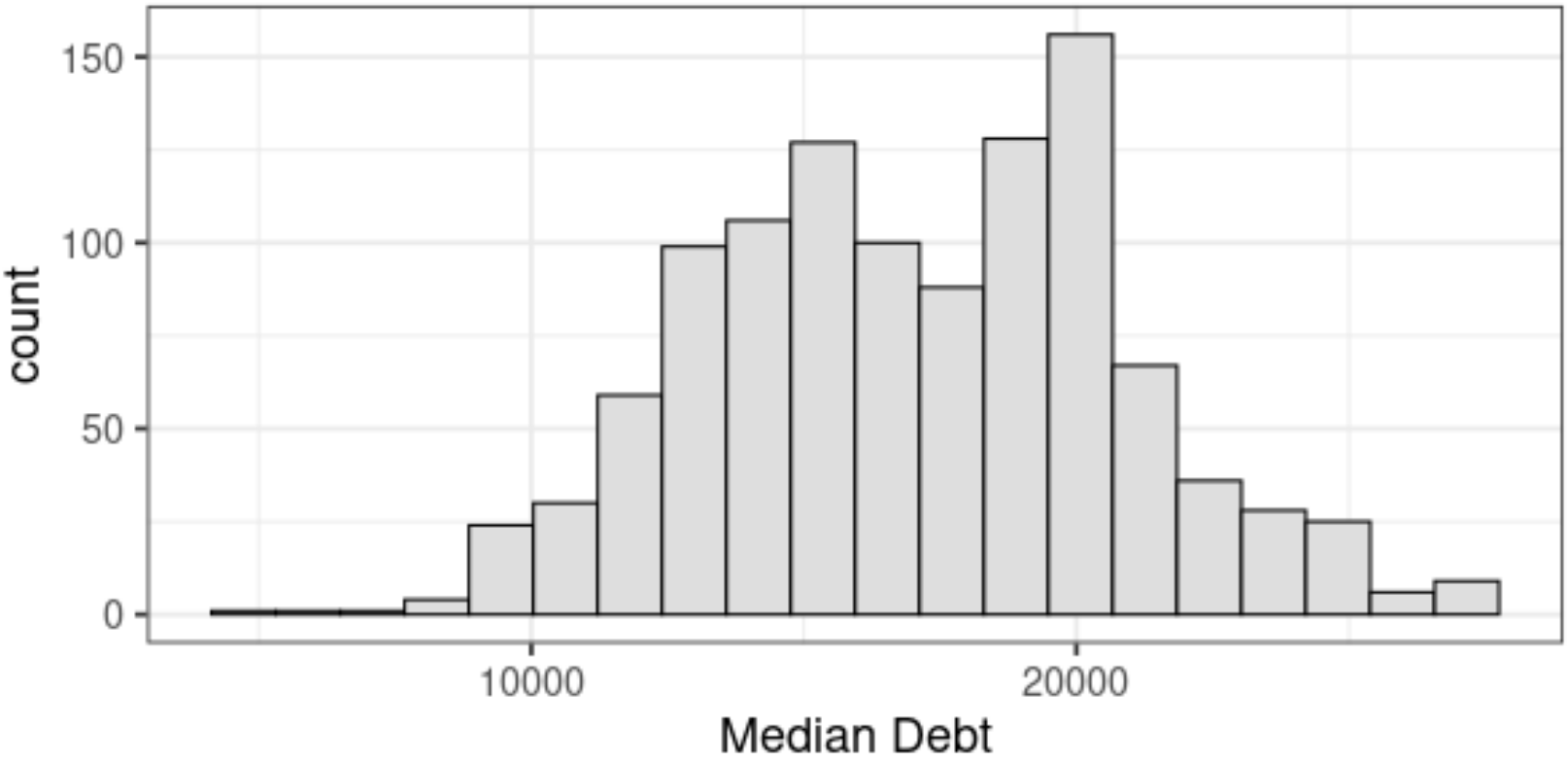
Practice

Describe the distribution of '% White Enrollment.'



Practice

Describe the distribution of Median Debt.



Advantages and Disadvantages

Order Statistics

Advantages:

- Robust to outliers
- More "correct" center for skew

Disadvantages:

- Discards most data
- No nice math properties

Moment Statistics

Advantages:

- Very useful math properties
- Utilizes all of the data

Disadvantages:

- Sensitive to outliers
- Sensitive to skew

Comparing Quantitative Variables

We can use **center** and **spread** to compare distributions. We typically refer to measures of centrality when discussing association.

Consider the five-number summary for our Enrollment variable (# students enrolled at a college)

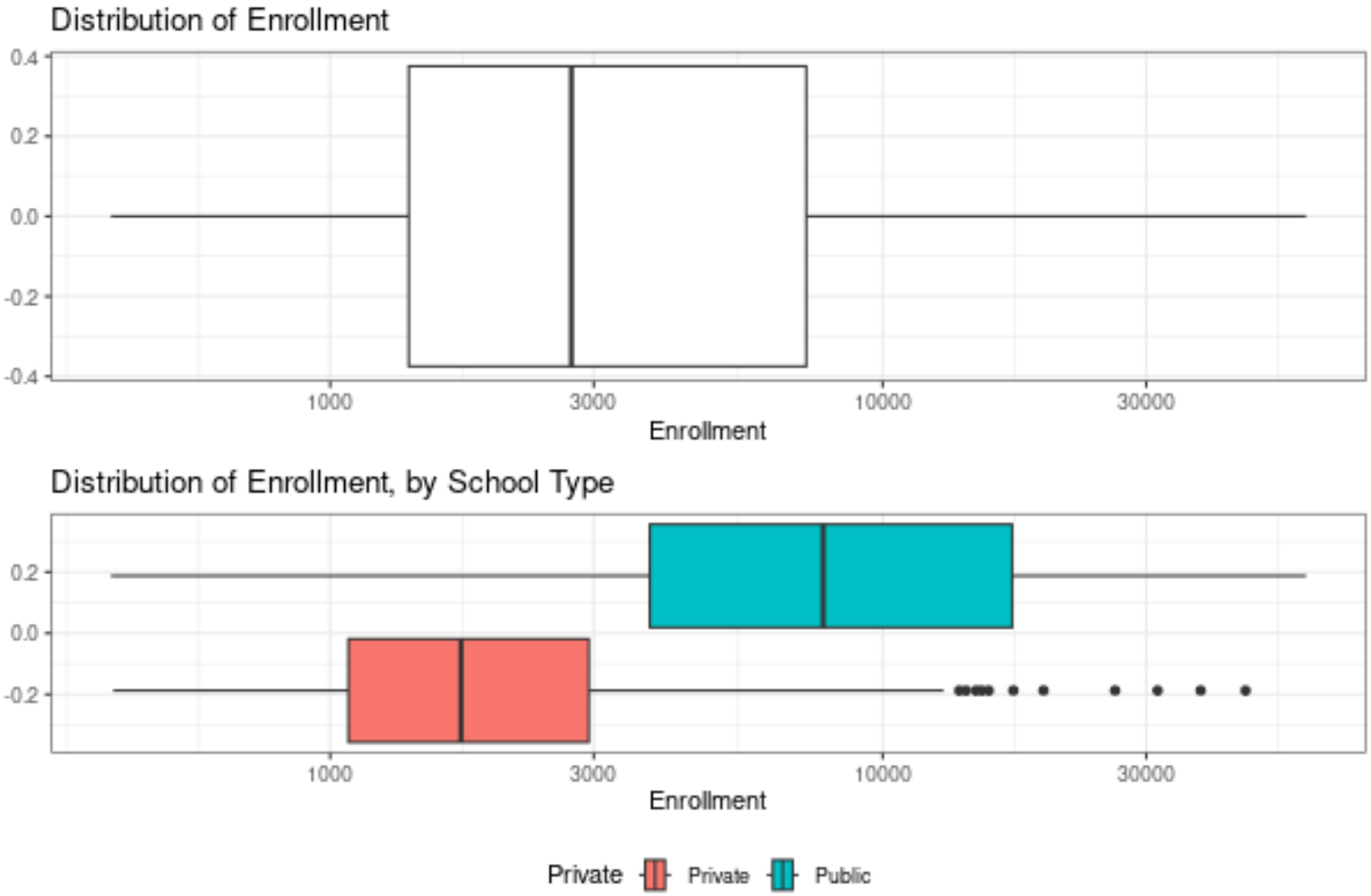
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
401	1388	2733	6241	7272	58392

Now lets look at the same variable, but add in info Public/Private

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Private	405	1079	1725	2720	2938	45370
Public	401	3788	7803	11325	17152	58392

Conditional Statistics

Which type of college tends to have more students enrolled? (center)



Parameter vs Statistics

Sometimes we will need to distinguish whether we are talking about things like mean/median/IQR/Std. dev. for a sample of data or for a population

Statistics are numerical summaries calculated from the *sample*.

- typically statistics are denoted using lowercase English alphabet characters
  - sample mean $= \overline{x}$
  - sample standard deviation $=$ s

Parameters are numerical summaries calculated from the population.

- typically parameters are denoted using Greek alphabet characters
  - population mean $= \mu$
  - population standard deviation $= \sigma$
- almost always the value of parameters are unknown to us (we want to learn about their values)

Z-scores

Sometimes we want to compare two variables that measure similar things but use different scales. (i.e.: ACT or SAT exam scores)

Standardizing our values is a way of adjusting them so that we can directly compare them. These adjusted values are called **z-scores**.

There are two formulas for calculating z-scores depending on if we are using sample or population info

$$z_i = \frac{x_i - \overline{x}}{s_x} \qquad z_i = \frac{x_i - \mu}{\sigma}$$

- in order to use μ and σ you need to actually be told these values, often a question prompt will tell you what to use

Z-score Properties

We can talk about standardizing a single value or an entire variable (standardizing *all* values of the variable)

If we standardize a variable, what we are doing is scaling each value so that the mean becomes 0 and the standard deviation becomes 1.

Properties of Z-Scores

- puts values on the same scale
- mean (center) $= 0$, std. dev (spread) $= 1$
- no units
- does not change the overall shape of the distribution

Z-Score Comparisons

Z-score interpretation: The value we get is the number of standard deviations the value is away from the mean

- positive z-scores are larger than the mean
- negative z-scores are smaller than the mean

Examples

- If z = 1.5, then the observation is 1.5 standard deviations larger than the mean
- If z = -1, then the observation is 1 standard deviation less than the mean

Example (Grinnell)

In our college dataset:

- the average ACT Median is $\overline{x} = 23.58$
- standard deviation of $s_x = 3.55$

Grinnell College has a median ACT of 32

- We can calculate its standardized value as:

$$z_{Grinnell} = \frac{32 - 23.58}{3.55} = 2.37$$

- This indicates that the median ACT at Grinnell College is 2.37 standard deviations above the average

Example (Iowa State University)

In our college dataset:

- the average ACT Median is $\overline{x} = 23.58$
- standard deviation of $s_x = 3.55$

Iowa State University has a median ACT of 25

- We can calculate its standardized value as:

$$z_{ISU} = \frac{25 - 23.58}{3.55} = 0.40$$

- This indicates that the median ACT at ISU is 0.4 standard deviations above the average

Example (Comparison)

We can compare Grinnell to ISU in terms of either mean 'ACT Median' directly, or we can compare their z-scores.

$$ACT_{Grinnell} = 32, \qquad ACT_{ISU} = 25$$
$$z_{Grinnell} = 2.37, \qquad z_{ISU} = 0.40$$

Which has the better score?
Grinnell, since the z-score is larger

Another example (ACT vs SAT)

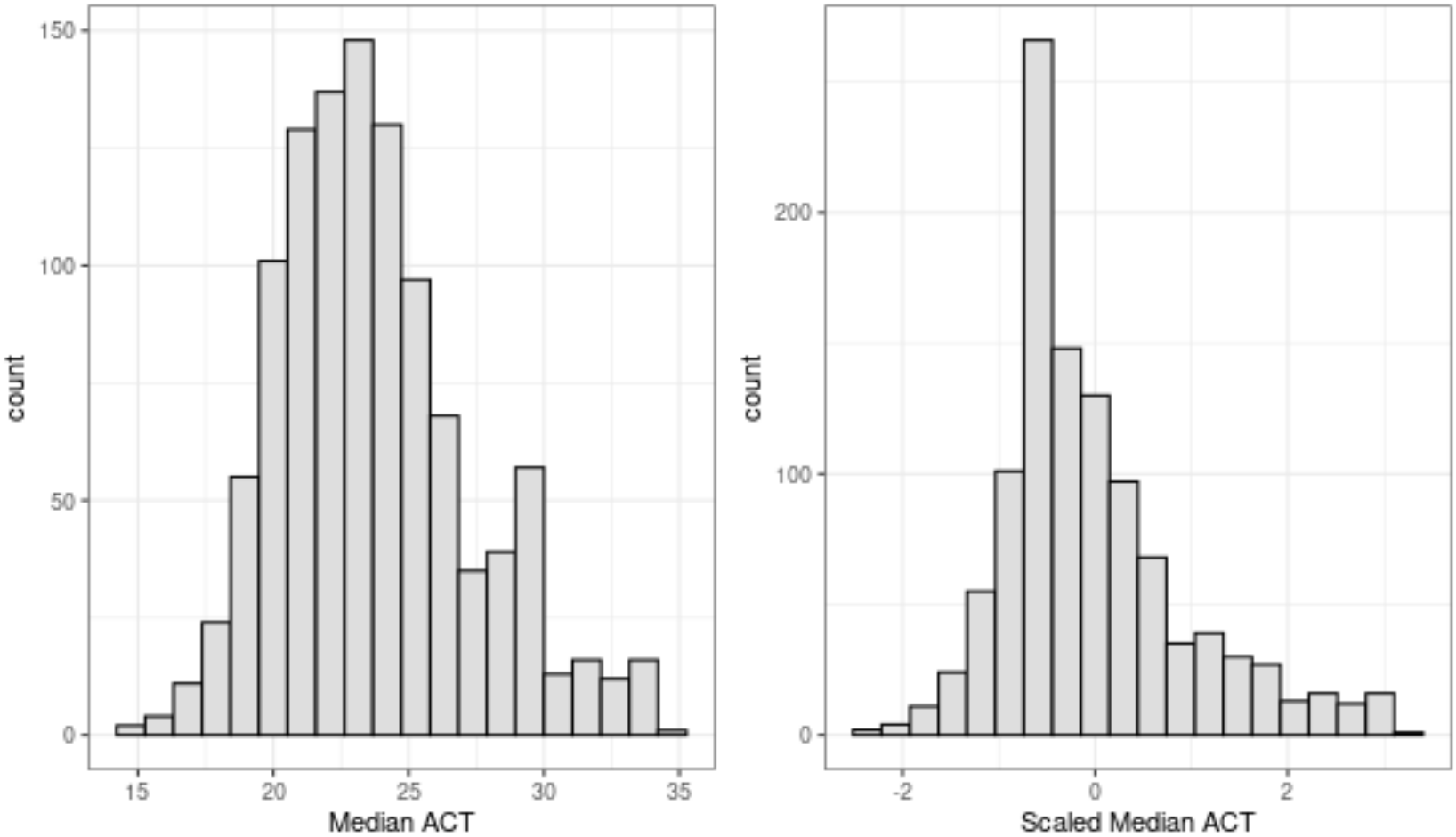
- The average score on the ACT English exam is 21.0 with a standard deviation of 4.0.
- The average score on the SAT Verbal exam is 520 with a standard deviation of 100.
- Kala scored a 27 on the ACT English exam.
- Nia scored a 770 on the SAT Verbal exam.

$$z_{Kala} = \frac{27 - 21}{4} = 1.5 \qquad z_{Nia} = \frac{770 - 520}{100} = 2.5$$

Who scored better? Nia, since Nia's z-score is larger

Transformation

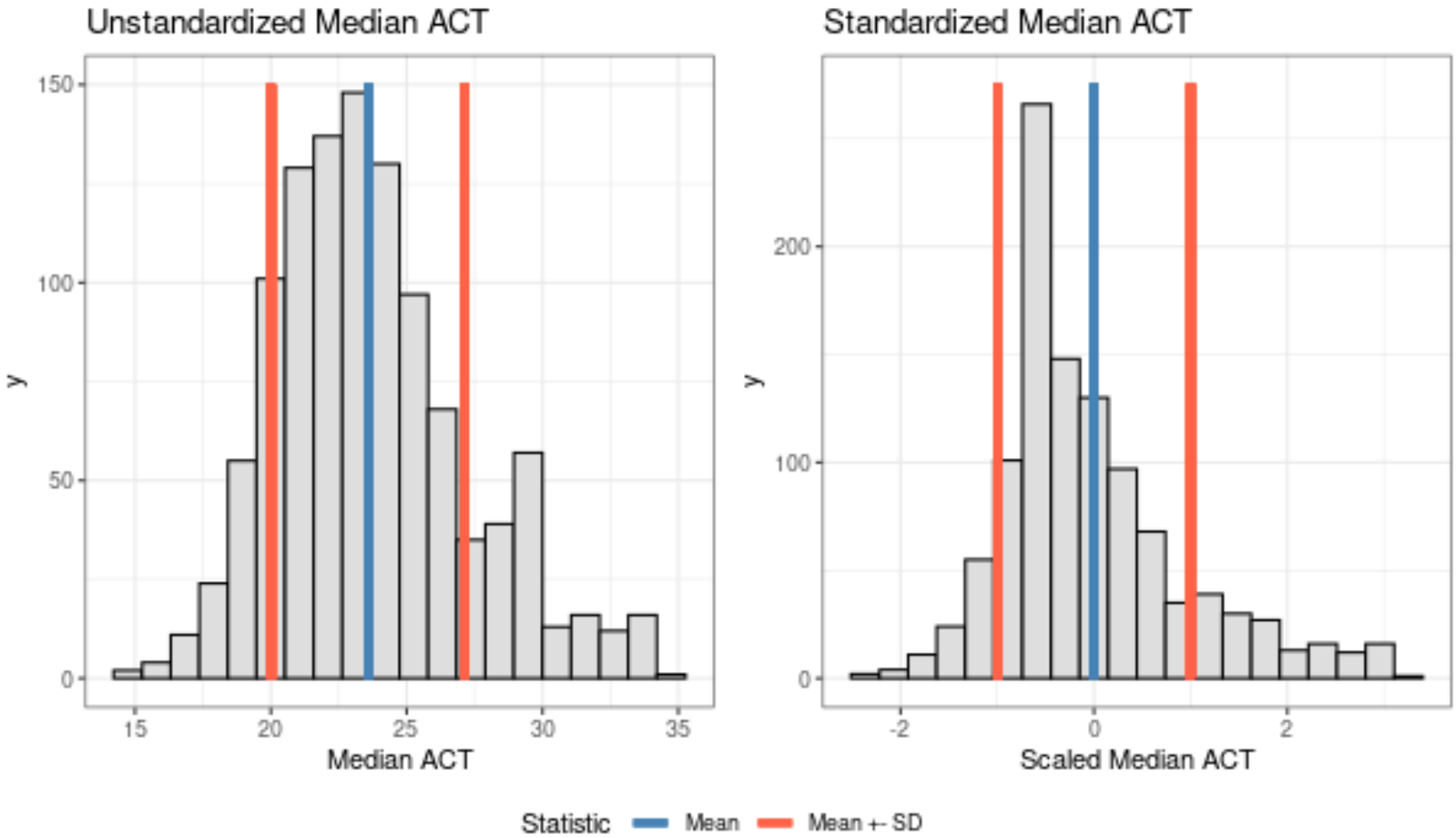
We can **standardize** all the values of a variable. The resulting standardized Values are either squeezed or stretched, but their relative positions stay the same



Transformation

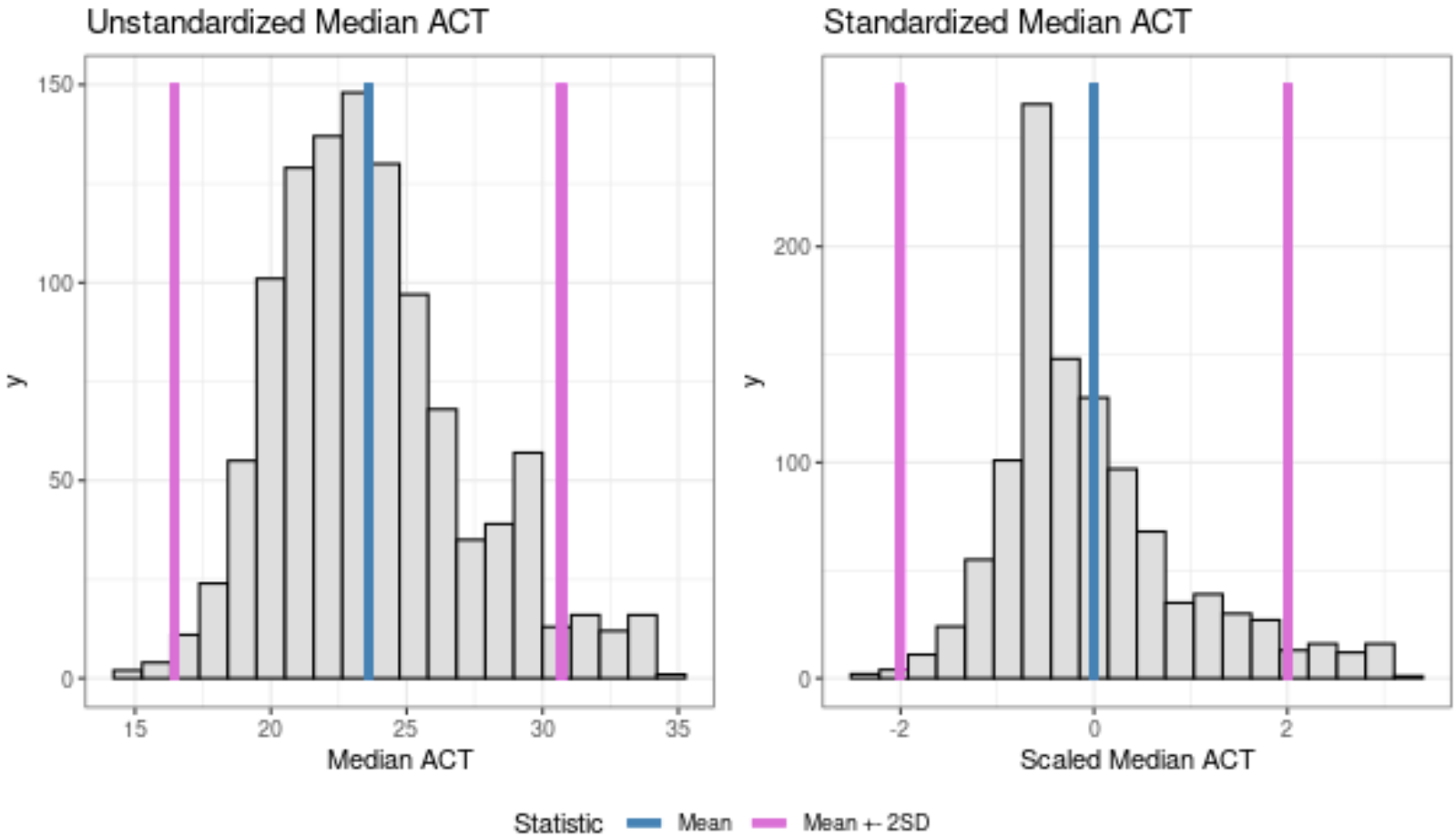
Similar to IQR giving us a range of common values, sometimes mean and standard deviation are used to create ranges of common values. Nearly the same amount of data still falls within this interval

Mean and one standard deviation:



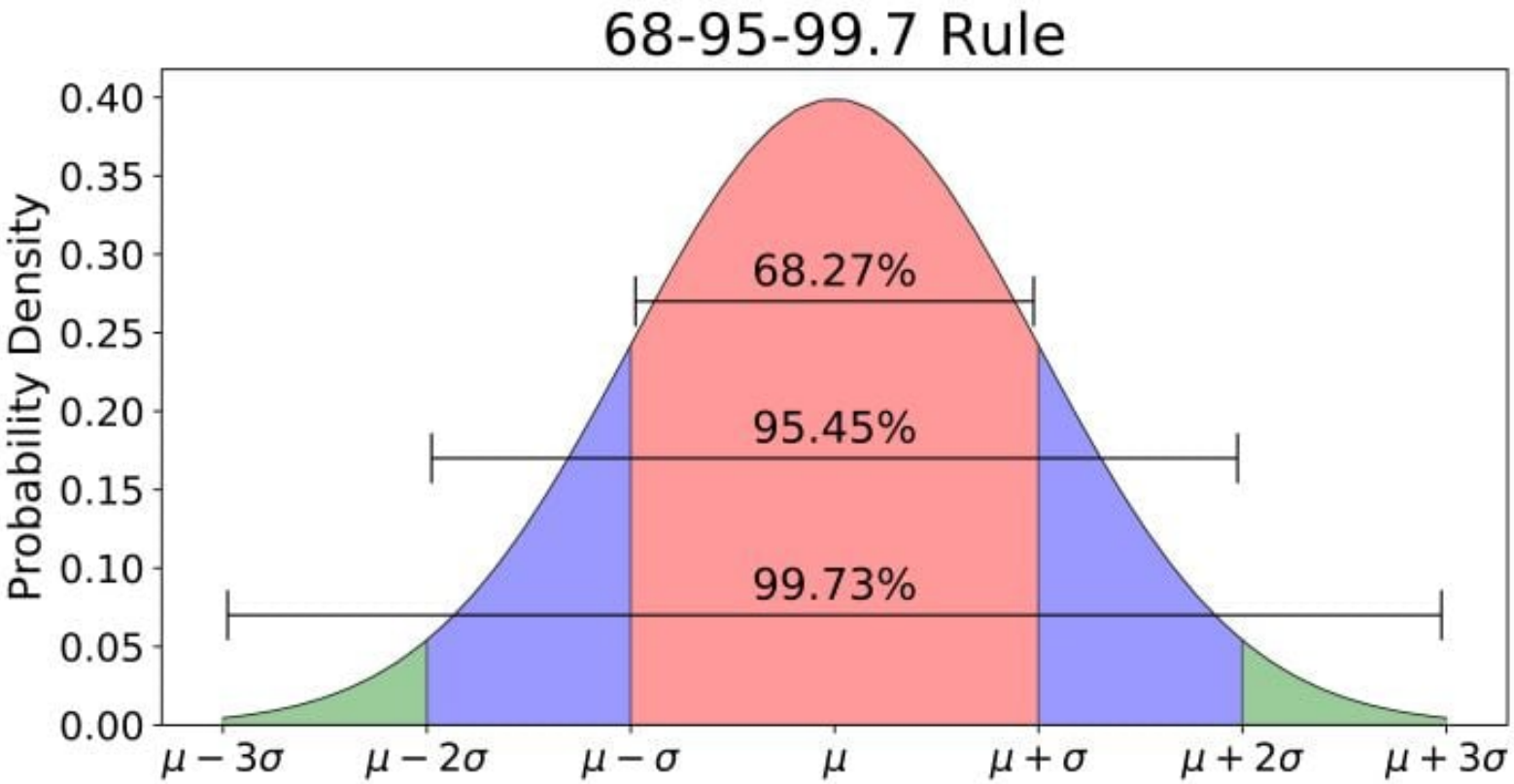
Transformation

Mean and two standard deviations



68-95-99 Rule

Useful property of the mean and standard deviation when we have a bell-shaped distribution (unimodal and symmetric).



Review

1. What is the difference between Order and Moment statistics?
2. What are the measures of center and spread we saw?
3. How do shape/outliers affect the measures of center/spread we use?
4. What is the purpose of standardizing a variable?

Additional question: Do you prefer examples where I write answers on the board, or do you prefer me putting them in the slides and explaining the answer (board / slides)