

# Numerical Summaries

Grinnell College

September 17, 2025

1. Brief review of histograms to jump into...
2. Numeric summaries of 1 variable (Part 1)
  - ▶ {Median, IQR, Range}  $\approx$  {Five Number Summary} vs
  - ▶ {Mean and Standard Deviation}
3. Numeric summaries for 2 variables (Part 2)
  - ▶ Pearson's Correlation
  - ▶ Spearman's Correlation

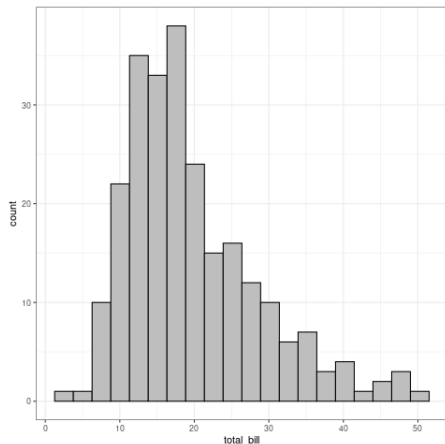
The main concepts in the class have been...

1. **Data Visualization** displaying data to allow quick and easy interpretations
  - ▶ The best graphics tell a story
2. **Numerical Summaries** that tell us about the data
  - ▶ Summaries necessarily lose information
  - ▶ Have already touched on many of them

Today we will fill out the numeric summaries more fully.

# Review – One Quantitative Variable

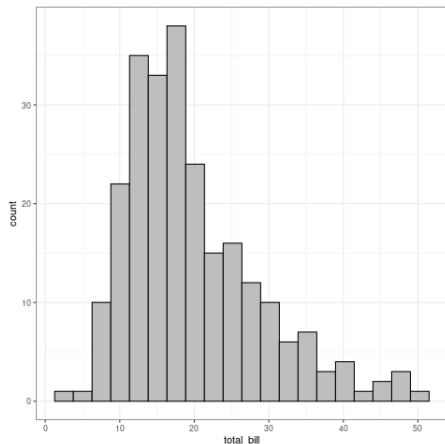
We've seen a few examples of histograms at this point.



Things we discuss when talking about the shape of a distribution?

# Review – One Quantitative Variable

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



Things we've discussed:

- How many modes does the data have?
- Are there outliers?
- Does it appear skewed or symmetric?

# 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
- **Outliers** - are there values that are much smaller/larger than the rest?
- **Center** - where does the data bunch up
- **Spread** - how spread out is the data

Today will tackle the last two points.

# Numerical Summaries

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

- What is the “middle” of our data?
- What is the central behavior of a variable?
  - ▶ As compared to estimating the maximum or minimums
- What is the 'standard' response to some treatment we apply in an experiment?

The **spread**, literally, is how spread out the values are.

- Statistics is the study of variability
- Without this things are deterministic

# Two Approaches for Center and Spread

## 1. Order Statistics

- ▶ Based on percentiles
  - ★ Median, IQR, Range
  - ★ Min,  $Q_1$ , Median,  $Q_3$ , Max
- ▶ Called “order” because of ordering the values smallest to largest
- ▶ Non-parametric statistics/applied to any shape

## 2. Moment Statistics

- ▶ Based on mathematical properties of our data
- ▶ Eg Mean and Standard Deviation
  - ★ Less common eg: skew, kurtosis, rate parameters
- ▶ Makes assumptions on the shape of the distribution
  - ★ Only as valid as those assumptions!!!
- ▶ Have nice properties usually
  - ★ Eg can use calculus



# Why?

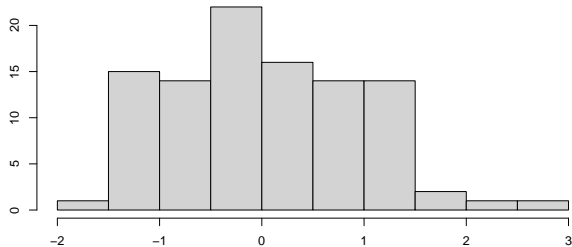
Differently shaped distributions require different approaches. Broadly...

- Symmetric with no (few?) outliers works well with moment statistics at this level
  - ▶ Mean and Standard Deviation
  - ▶ Want to lean towards this due to nice math properties
- Weird shaped distributions should be approached with order statistics
  - ▶ Five Number Summary
  - ▶ Median, IQR, Range

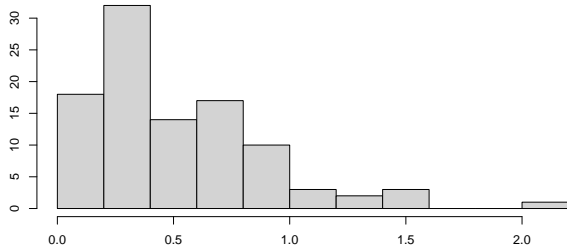
Failure to differentiate properly will lead to choosing the incorrect analysis

# What do I mean by “weird”?

**Normal**



**Gamma**



# 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 major fluctuations in the data
  - ▶ Eg Outliers do little to them
  - ▶ (Rhetorical) How many outliers do we need to make  $Q_3$  move?
3. Easier to interpret

# Review – Percentiles

A **percentile**  $\alpha$  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

# Quartiles and IQR and Range

**Quartiles** are the names for the 25% of data markers

- $Q_1 = 25\%$ 
  - ▶ = median of the lower half of the data
- $Q_3 = 75\%$ 
  - ▶ = median of the upper half of the data
- If the number of observations is odd, include the median when finding the median of the lower and upper half
  - ▶ Eg 1,1,2,3,5 - (1,1,2) and (2,3,5)

**IQR** is the width of the middle 50% of the data

- $Q_3 - Q_1$

**Range** is the total width of your data

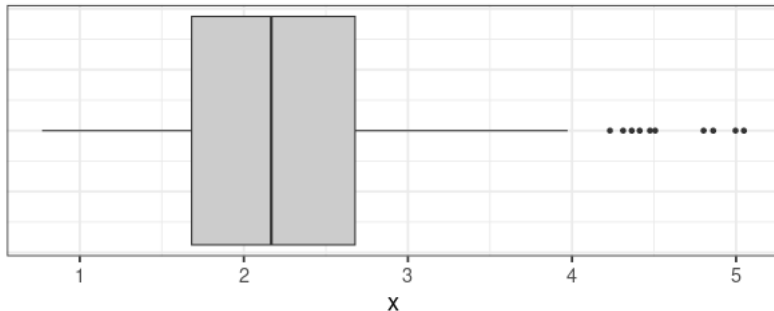
- Max - Min
- Largest variability seen in the data

# Question?

What's  $Q_2$ ?

# Five Number Summary

Min,  $Q_1$ , Median,  $Q_3$ , Max



Challenge: Visually guesstimate the five number summary of this boxplot

("Guesstimate" is a highly technical term.....)



# 5 Number Summary vs Median and IQR

I generally prefer the 5 number summary, why?

## 5 Number Summary vs Median and IQR?

I generally prefer the 5 number summary, why?

The data is less *processed*

- Information is lost when we summarize
- I can find IQR with the 5 number summary
- I can't find 5 number summary with median and IQR

# Moment Statistics

**Moment statistics** are statistics that are based on specific mathematical properties of our data (mean and st. dev.).

- Often have very nice properties
  - ▶ eg we can use calculus
- Make assumptions about the shape of the data
  - ▶ Assumptions bring information a priori but if violated leads to problems (bias)
- Can be moderately sensitive to fluctuations in the data
  - ▶ ie NOT robust
  - ▶ eg outliers wreck the mean

# Notation

Let's get on the same page for notation

- Sample size is denoted with **n** (or N)
- $x_i$  is used to denote the  $i^{th}$  value of  $x$  in our data set
  - ▶ eg  $x_3$  is the 3rd value of  $x$
- A bar over a variable indicates an estimated mean
  - ▶  $\bar{x}$  is the mean of  $x$  in our sample
- $\sum$  stands for “the summation of” (ie add these up)
  - ▶  $\sum_{i=1}^5 (i) = 1 + 2 + 3 + 4 + 5 = 15$
- Not needed but  $\prod$  is for the product of things

# R Mini Coding Camp Coming Up

- Not turned in
- Boot up R & make a new markdown/script
- Copy and paste the line to read in the “colleges” data from the data visualization lab
  - ▶ Hint: search for the word “csv”
- I'll post it after class

# Mean (Center)

In common parlance the (arithmetic) **mean** is the same thing as the **average**.

- 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 the variable  $x$  is denoted as  $\bar{x}$
- Using the notation from the previous slide, the equation for the mean is

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

# Spread – Standard Deviation

**Standard Deviation** is the measure of typical deviation (distance) of all the observations from the mean

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

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

**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 unbiased estimates and predictions. Vaguely it's the price paid for using the mean in our formula

# 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**.

- sample variance =  $s^2$
- harder to interpret
- Many distributions are defined with variance instead of  $s$ 
  - ▶ Eg [Wiki for Chi-squared Distribution](#) only has variance listed
  - ▶ Higher up in statistics “variance” is more prevalent



# Which measures to use?

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

- Generally default towards moment statistics
- Up to the histogram to convince me to use order statistics
  - ▶ For those who know the term, order statistics lack the power of moment statistics
  - ▶ But they (order stats) also don't have assumptions

# Which measures to use?

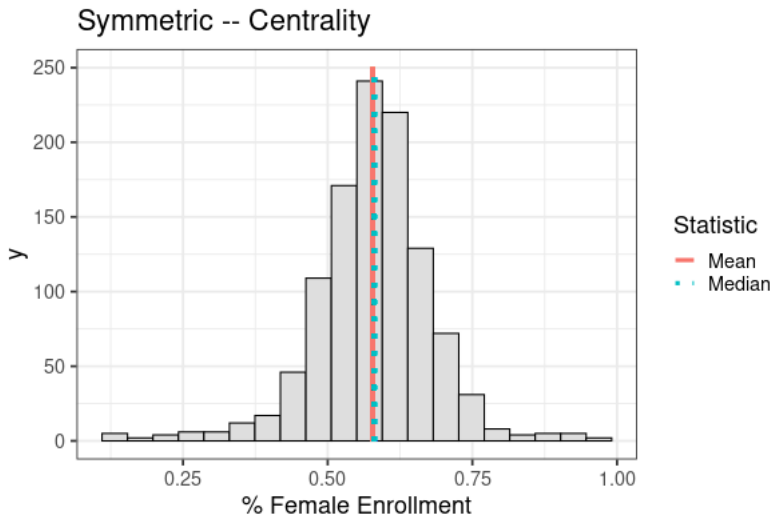
Order statistics are robust, moment statistics are not.

- 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

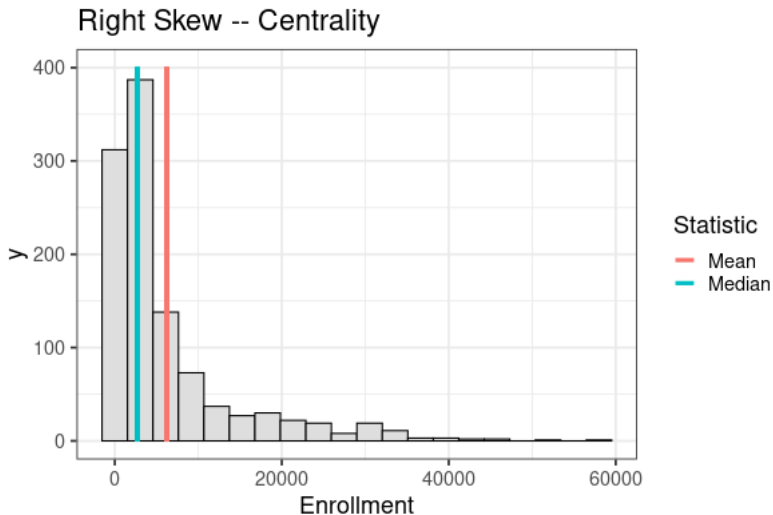
Summary:

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



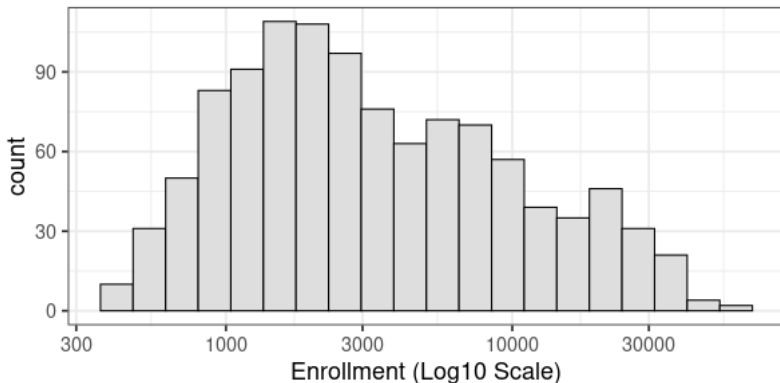
# Comparing Mean with Median



# Practice

For each of the following variables visualized below:

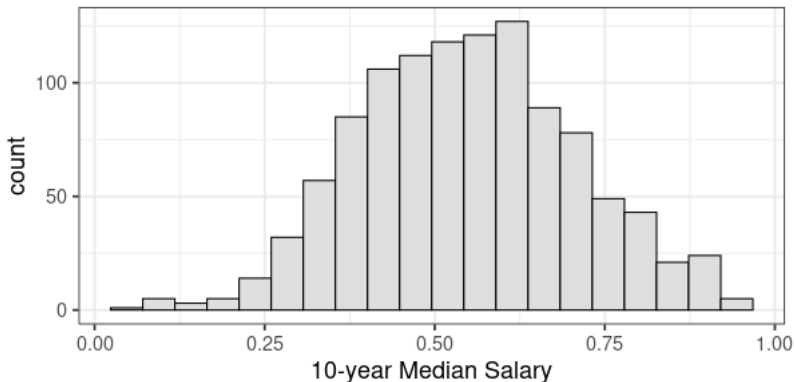
1. Determine approximate mean and median and which should be larger. How do you know?
2. Decide whether standard deviation or IQR is more appropriate for describing variability



# Practice

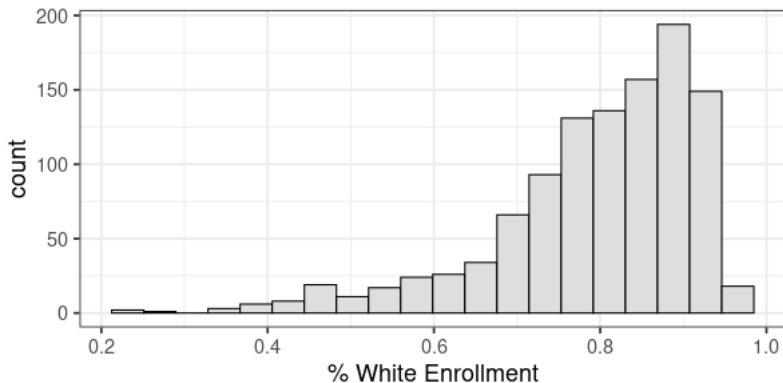
For each of the following variables visualized below:

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



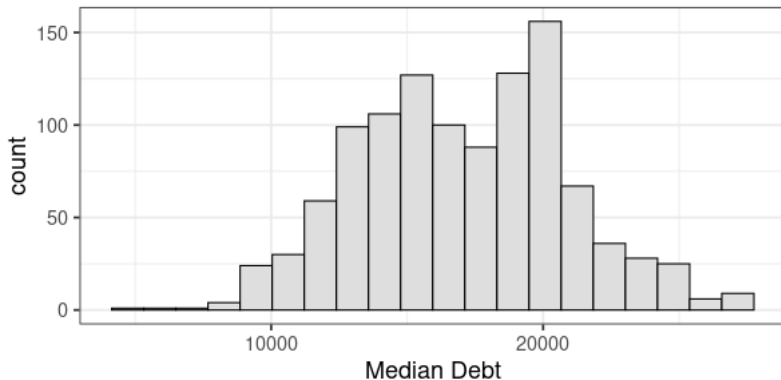
# Practice

Describe the distribution of '% White Enrollment.'



# Practice

Describe the distribution of Median Debt.





# Advantages and Disadvantages

## Order Statistics

### Advantages:

- Robust to outliers (besides range)
- A “better” center for skewed data

### Disadvantages:

- Discards most data
- Math properties that are...of limited use
- Statistically less powerful

## Moment Statistics

### Advantages:

- Very useful math properties
- Utilizes all of the data
- Higher power

### Disadvantages:

- Uses all of the data
- 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)

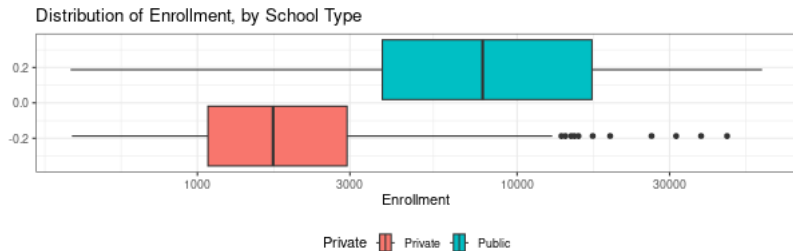
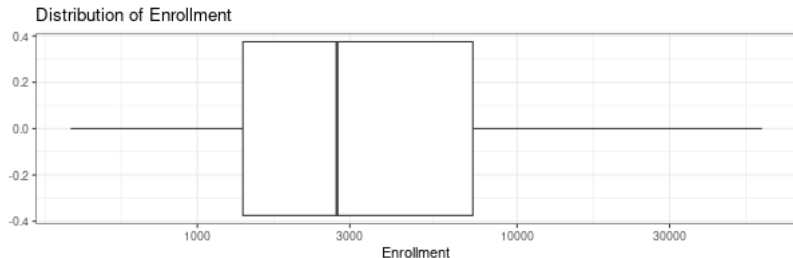
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
401	1388	2733	6241	7272	58392

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 (= Statistics for Each Category)

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



# Parameter vs Statistics

It's important to distinguish between when we are talking about numeric summaries related to the sample or the population

**Statistics** are numerical summaries calculated from the *sample*.

- typically statistics are denoted using lowercase Latin alphabet characters
  - ▶ sample mean =  $\bar{x}$
  - ▶ 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

# Parameter vs Statistics

Statistics are to Samples what Parameters are to Populations Example:

Suppose we are interested to see the effects of daily aspirin usage in younger aged people in increasing the years until the onset of the patients first cardiovascular event (eg heart attack). In particular, we want to increase the mean number of years. We randomly sample 500 people between the ages of 18 and 24 and gave them aspirin daily. We then recorded the time from taking the aspirin until their first cardiovascular event.

- Population
- Sample
- Parameter
- Statistic

# Parameter vs Statistics

Statistics are to Samples what Parameters are to Populations Example:

Suppose we are interested to see the effects of daily aspirin usage in younger aged people in increasing the years until the onset of the patients first cardiovascular event (eg heartattack). In particular, we want to increase the mean number of years. We randomly sample 500 people between the ages of 18 and 24 and gave them aspirin daily. We then recorded the time from taking the aspirin until their first cardiovascular event.

- Population: All “younger aged” people
- Sample: 500 people aged 18-24
- Parameter: mean years until the first cardiovascular event of all “younger aged” people
- Statistic: mean years until the first cardiovascular event of our sample