

# Numerical Summaries

Grinnell College

September 9, 2024

## Graphical Summaries:

- Why create graphs?
- Types of plots?
- Notable aspects?

## John Tukey quote (again):

*“Numerical summaries focus on expected values, graphical summaries focus on unexpected values”*

Today we focus on univariate quantitative summaries

# Numerical Summaries

As with graphical summaries, there are typically a few attributes that we are interested:

1. Where is our data centered?
2. How spread out is it from the center?

To this end, we will mostly concern ourselves with two orders of thought here for identifying this information

1. Order Statistics
2. Moment Statistics

# Order Statistics

**Order statistics**, perhaps unsurprisingly, are statistics based on the ordinal ranking of a quantitative variable

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 (i.e., outliers)
3. Readily interpretable

# Percentiles

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

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

Along with the median, these numbers make up the *five-number summary* for describing data

# IQR

The **interquartile range** or **IQR** is the value of  $Q_3 - Q_1$ , giving the breadth of the middle 50% of the observed data

$$x = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

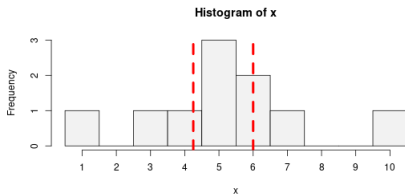
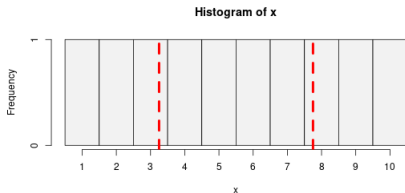
- $x_{\{25\}} = 3.25, x_{\{75\}} = 7.75$

- $IQR = 4.5$

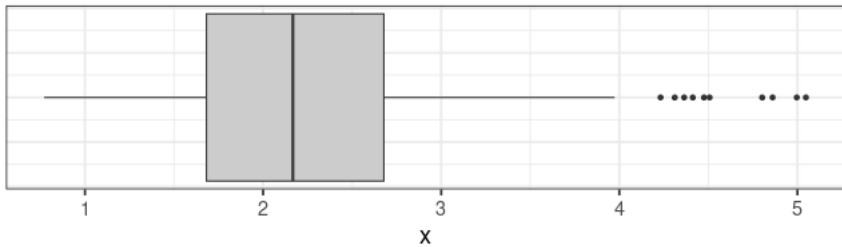
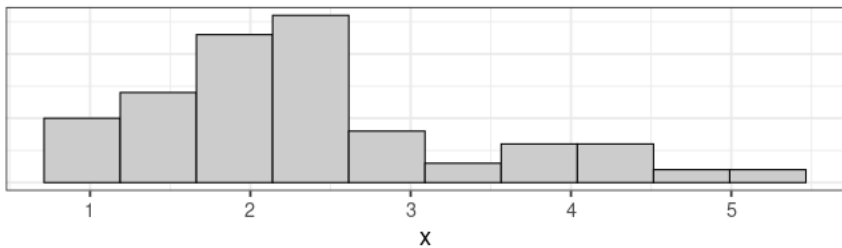
$$x = \{1, 3, 4, 5, 5, 5, 6, 6, 7, 10\}$$

- $x_{\{25\}} = 4.25, x_{\{75\}} = 6$

- $IQR = 1.75$

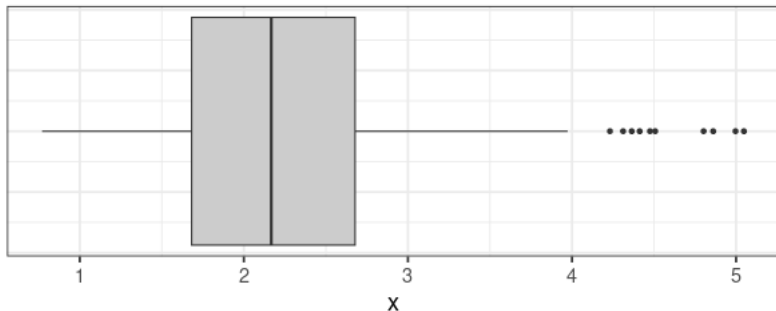


Forgotten slide on median (maybe include plots before/after outliers)





# Five Number Summary



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

# Moment Statistics

**Moment statistics** are statistics that are based on specific mathematical properties of our data

Because they are oriented around known properties, they are associated with very powerful theoretical tools that provide context to their behavior

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

# Mean

Greek letter  $\mu$  (mu or “myu”) for *parameter*,  $\bar{x}$  for *statistic*

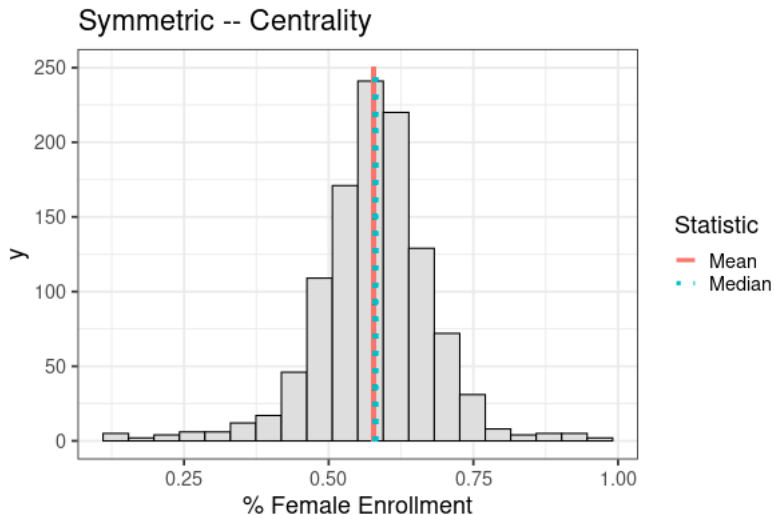
$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

The **mean**, or **arithmetic average**, describes the “center of mass” of a quantitative variable

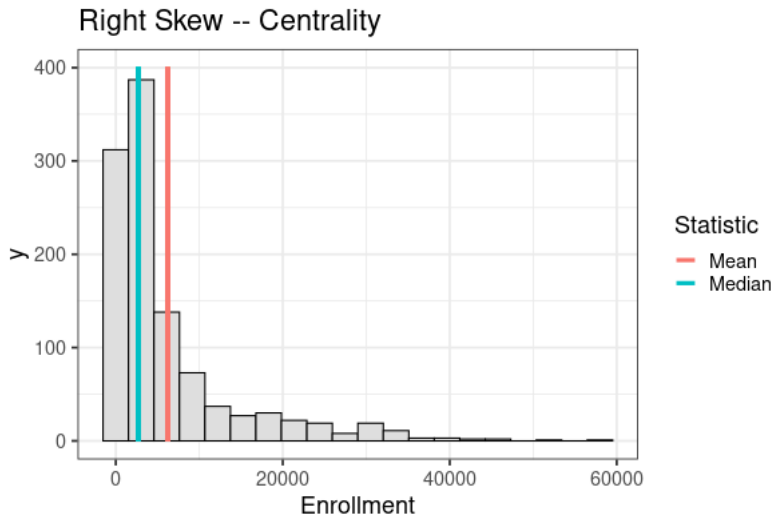
Unlike the median, which only uses the value of a single observation, the mean uses information from all of the observed values

This gives us a sense of an *expected value*

# Comparing Mean with Median



# Compare mean and median with stuff



# Standard Deviation

Greek letter  $\sigma$  (sigma) for *parameter* and  $s$  for *statistic*

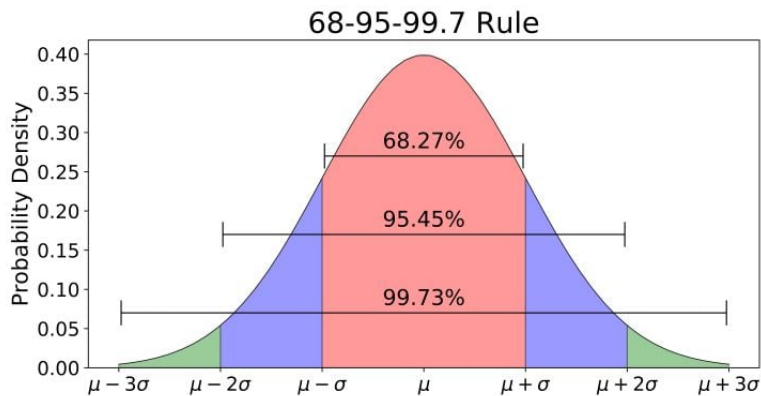
$$s = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

The **standard deviation** provides a measure of the average expected distance of our observations from their mean

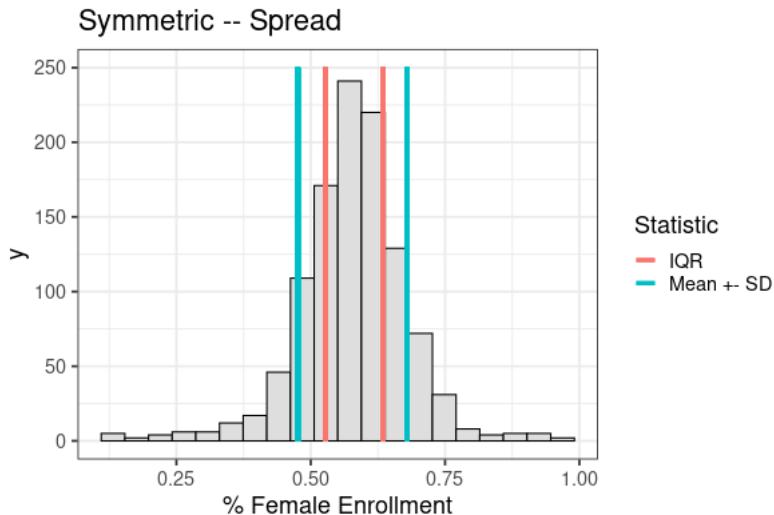
Because it is denoted in the same units as the variable in question, we can use it to construct ranges of values, i.e.,  $\mu \pm \sigma$

Often used to determine what is an outliers (i.e., how many standard deviations away)

# 68-95-99 Rule (Example)

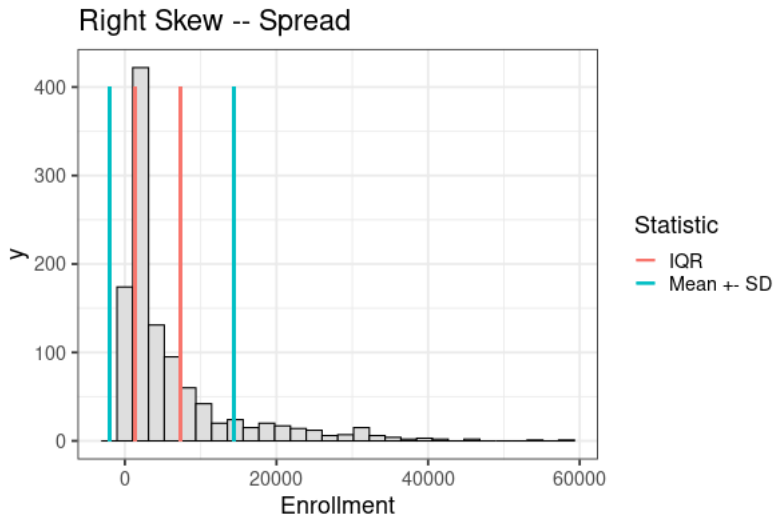


# Compare sd and IQR with stuff





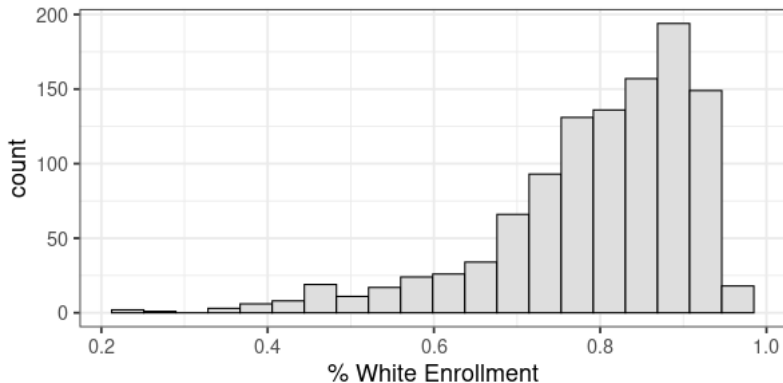
# Compare sd and IQR with stuff



# Practice

For each of the following variables visualized below:

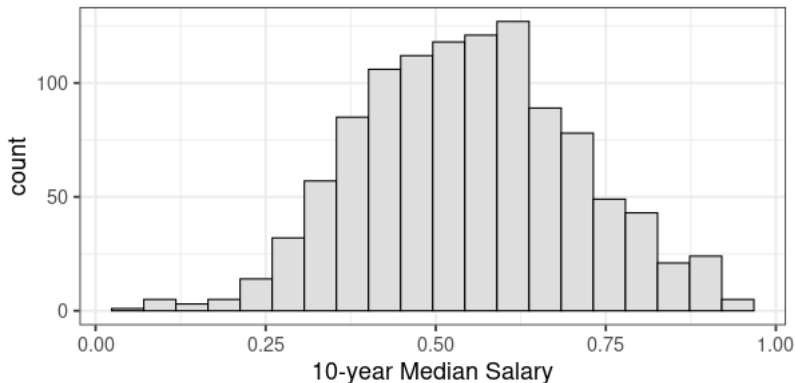
1. Determine approximate mean and median and which should be larger. How do you know?
2. Decide whether standard deviation or IQR is more appropriate for describing variability



## Practice

For each of the following variables visualized below:

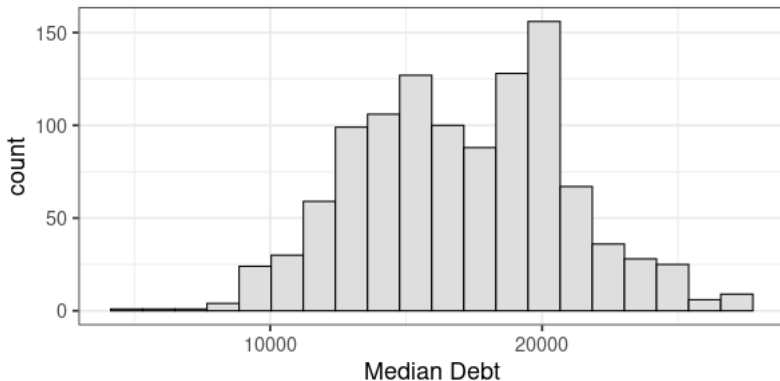
1. Determine approximate mean and median and which should be larger. How do you know?
2. Decide whether standard deviation or IQR is more appropriate for describing variability



## Practice

For each of the following variables visualized below:

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



# 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 for inference

- Utilizes all of the data

Disadvantages:

- Sensitive to outliers

- Sensitive to skew

# Things to Know

1. Center vs spread
2. Major quantiles (25, median, 75)
3. Identify components of boxplot
4. Effects of skew and outliers on various measures