Numerical Summaries

Grinnell College

January 28, 2026

Review

Graphical Summaries:

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

John Tukey quote:

"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** α 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

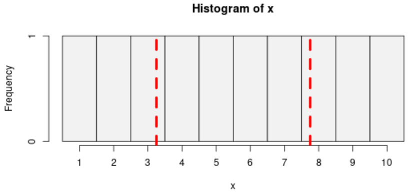
Along with the median, these numbers make up the *five-number summary* for describing data

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

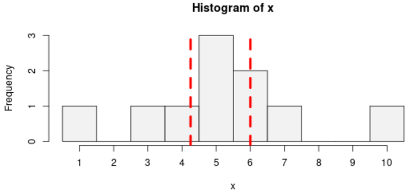
$x = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

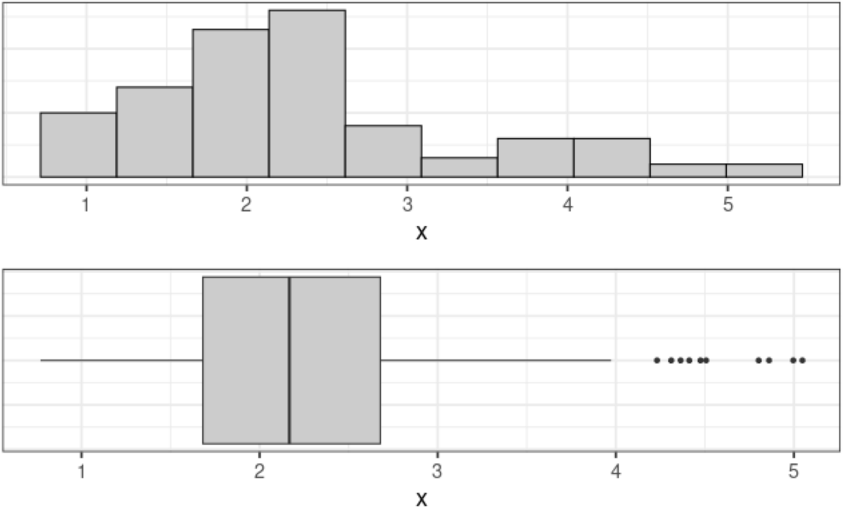
- $x_{\{25\}} = 3.25$, $x_{\{75\}} = 7.75$
- $IQR = 4.5$



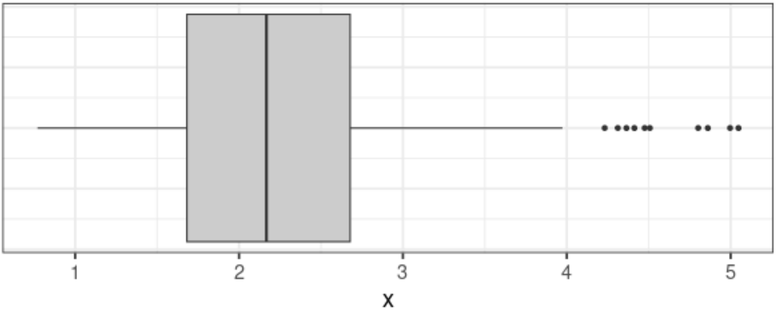
$x = \{1, 3, 4, 5, 5, 5, 6, 6, 7, 10\}$

- $x_{\{25\}} = 4.25$, $x_{\{75\}} = 6$
- $IQR = 1.75$




1
2
3
4
5
x
1
2
3
4
5
x

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

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 μ (mu or "myu") for *parameter*, $\overline{x}$ for *statistic*

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

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

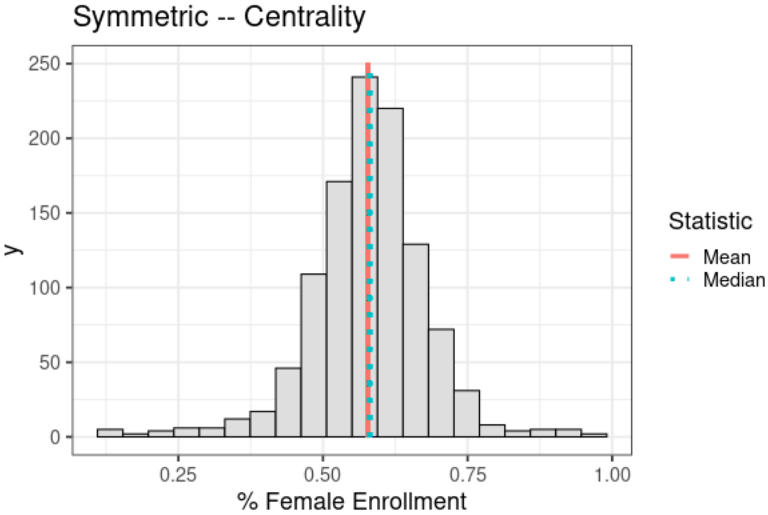
This gives us a sense of an *expected value*

Mean and Median

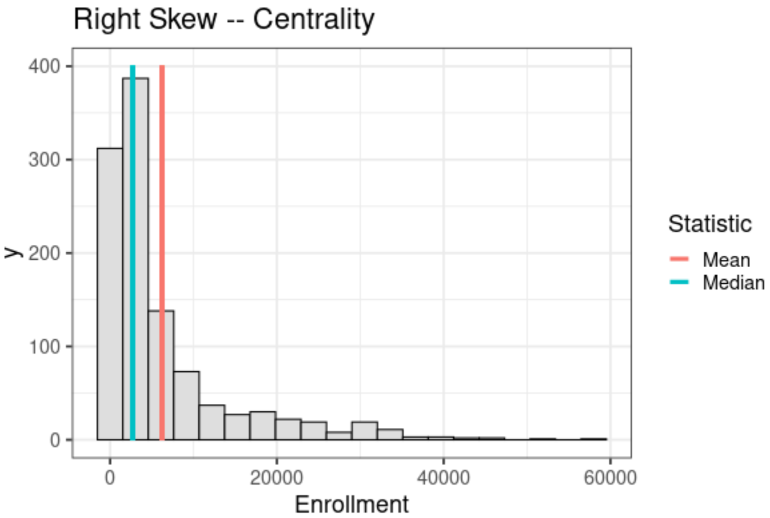
Suppose I invite you to play a game in which a standard six-sided dice is rolled. If the dice lands on a 6, you win \$10, otherwise you win nothing.

- What is the median amount of money you are expected to win?
- What is the mean?
- If it costs \$1.50 to play this game, would you play? Is the mean or median

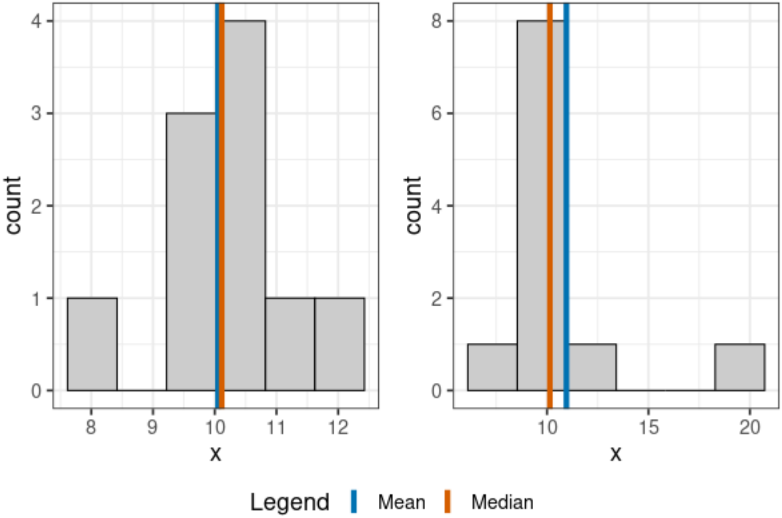
Comparing Mean with Median



Comparing Mean with Median



Outliers



count
4
3
2
1
0
8
9
10
11
12
x
count
8
6
4
2
0
10
15
20
x
Legend
Mean
Median

Standard Deviation

Greek letter σ (sigma) for *parameter* and $\hat{\sigma}$ for *statistic*

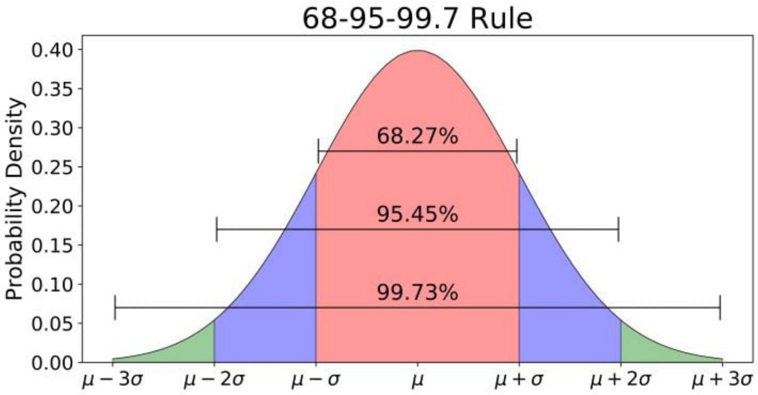
$$\hat{\sigma} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

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

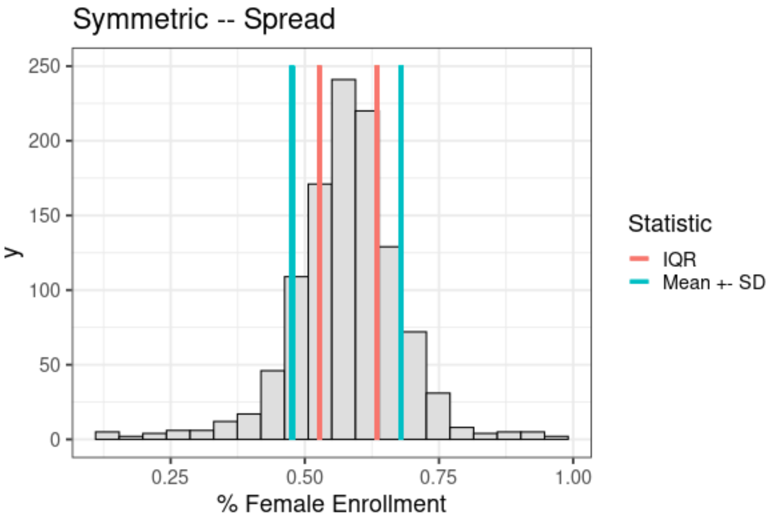
Because it is denoted in the same units as the variable in question, we can use it to construct ranges of value alongside the mean (e.g., $\mu \pm \sigma$)

A standardized unit of distance is used to determine what is an outlier

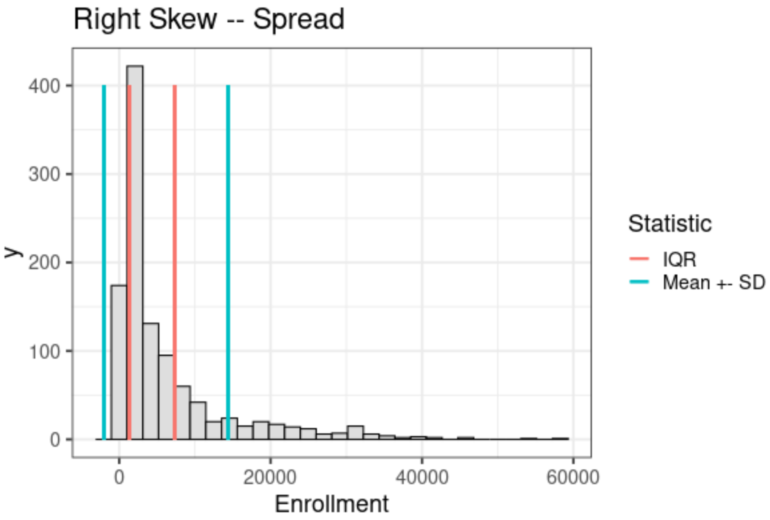
68-95-99 Rule (Example)



SD and IQR

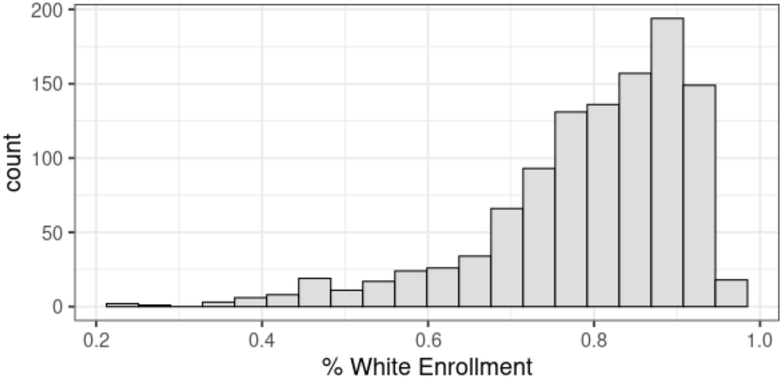

Symmetric -- Spread
250
200
150
100
50
0
y
0.25
0.50
0.75
1.00
% Female Enrollment
Statistic
IQR
Mean +- SD

SD and IQR


Right Skew -- Spread
400
300
200
100
0
y
0
20000
40000
60000
Enrollment
Statistic
IQR
Mean +- SD

Practice

1. Do you think that the mean or median should be larger in this case? 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