Numerical Summaries

Grinnell College

September 4, 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 in:

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 approaches for identifying this information

1. Order Statistics
2. Moment Statistics

Order Statistics

Order statistics, perhaps unsurprisingly, are statistics based on the ordered values of a quantitative variable

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

1. Inference on these make few assumptions about how the data is distributed
2. Some summaries are robust to major fluctuations in the data
3. Readily interpretable

Percentiles

A **percentile** α is a number such that approximately $\alpha\%$ 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

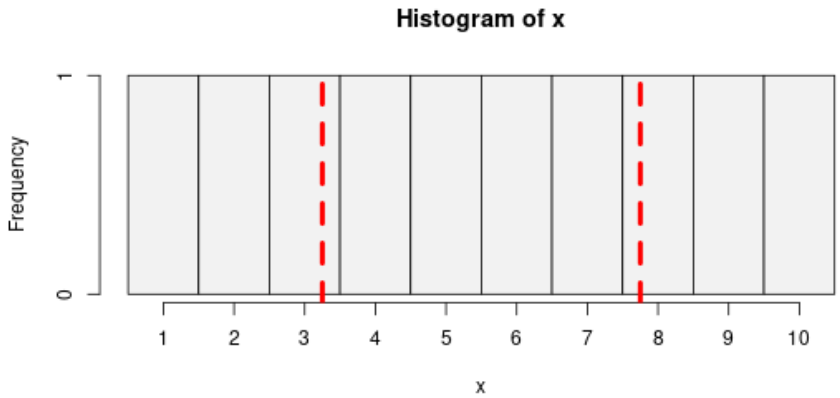
Along with the median, these numbers make up the *five-number summary* for describing data

IQR

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

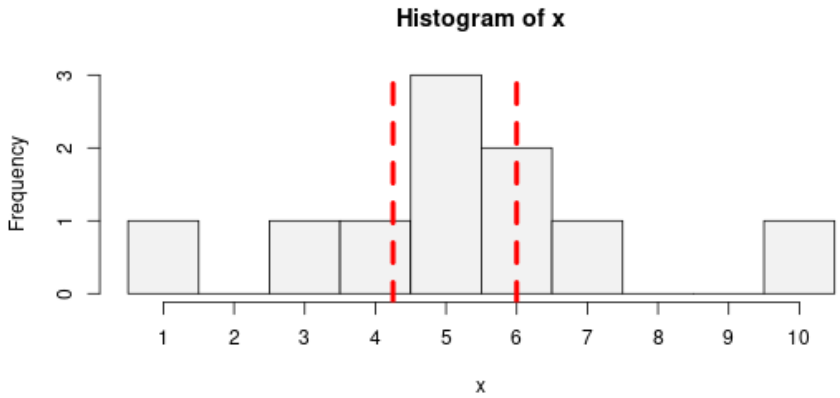
$x = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

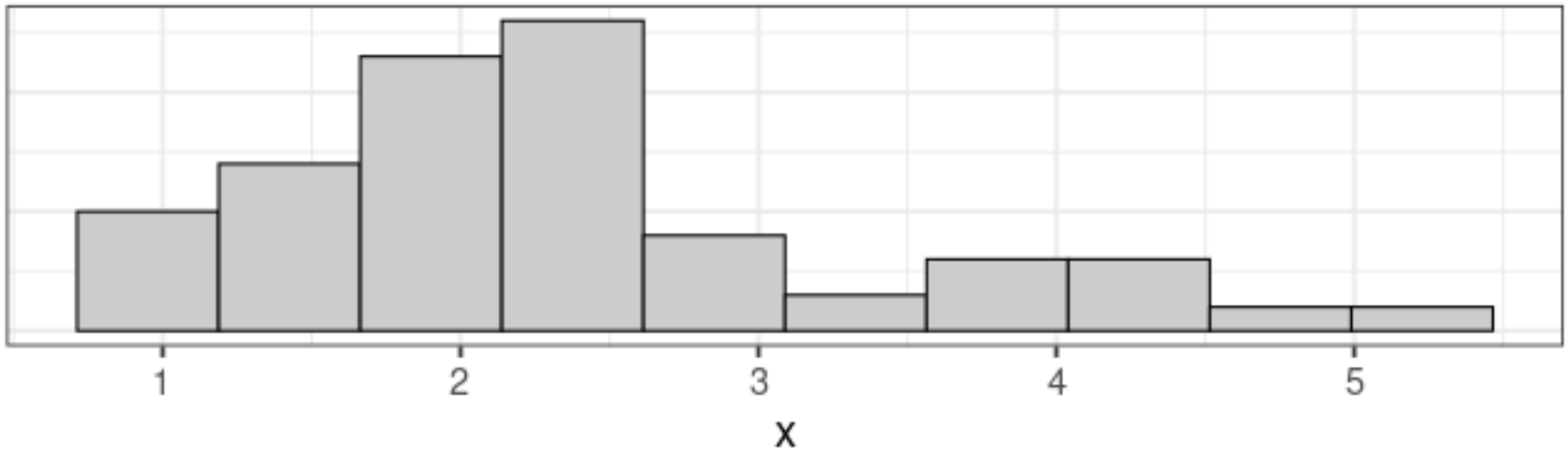
- $x_{\{25\}} = 3.25$, $x_{\{75\}} = 7.75$
- $IQR = 4.5$

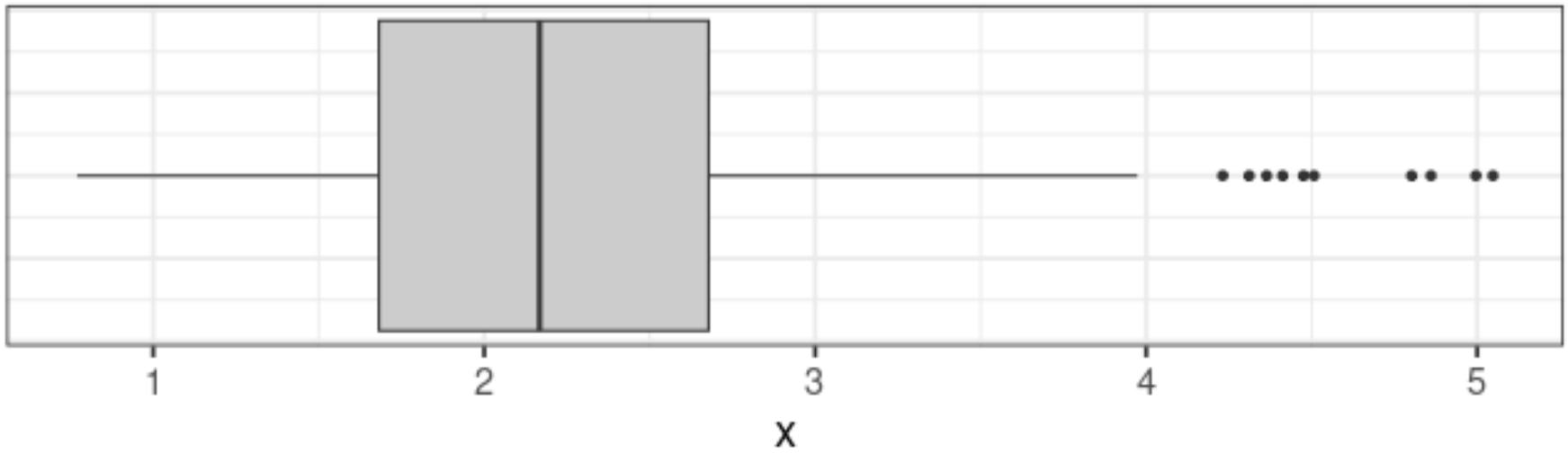


$x = \{1, 3, 4, 5, 5, 5, 6, 6, 7, 10\}$

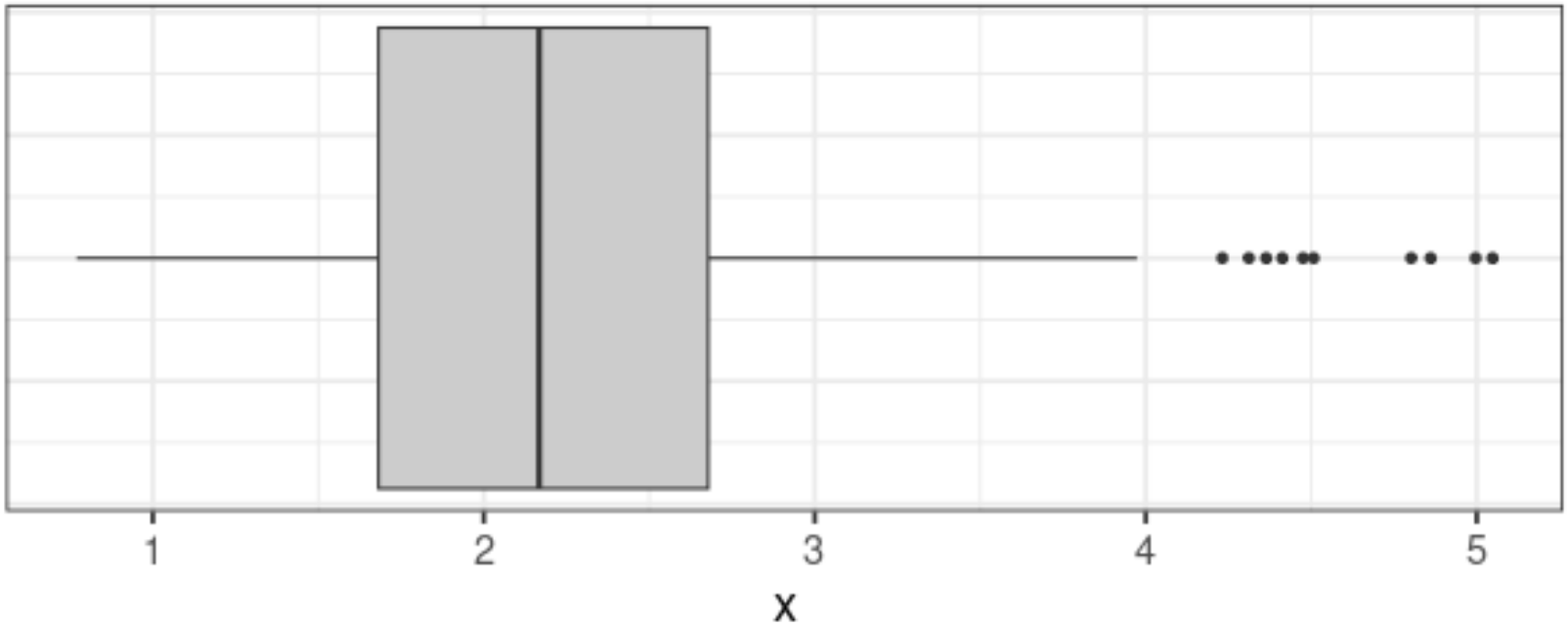
- $x_{\{25\}} = 4.25$, $x_{\{75\}} = 6$
- $IQR = 1.75$




1
2
3
4
5
x


1
2
3
4
5
x

Five Number Summary



- Median
- 25th Percentile (Q_1)
- 75th Percentile (Q_3)
- At most Q_1 - 1.5 $\times$ IQR
- At most Q_3 + 1.5 $\times$ IQR
- Outliers

Moment Statistics

Moment statistics are statistics that are based on the numerical magnitudes of all our observations

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

However, because they are associated with the magnitudes of *all* observations they can be sensitive to extreme values or 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*

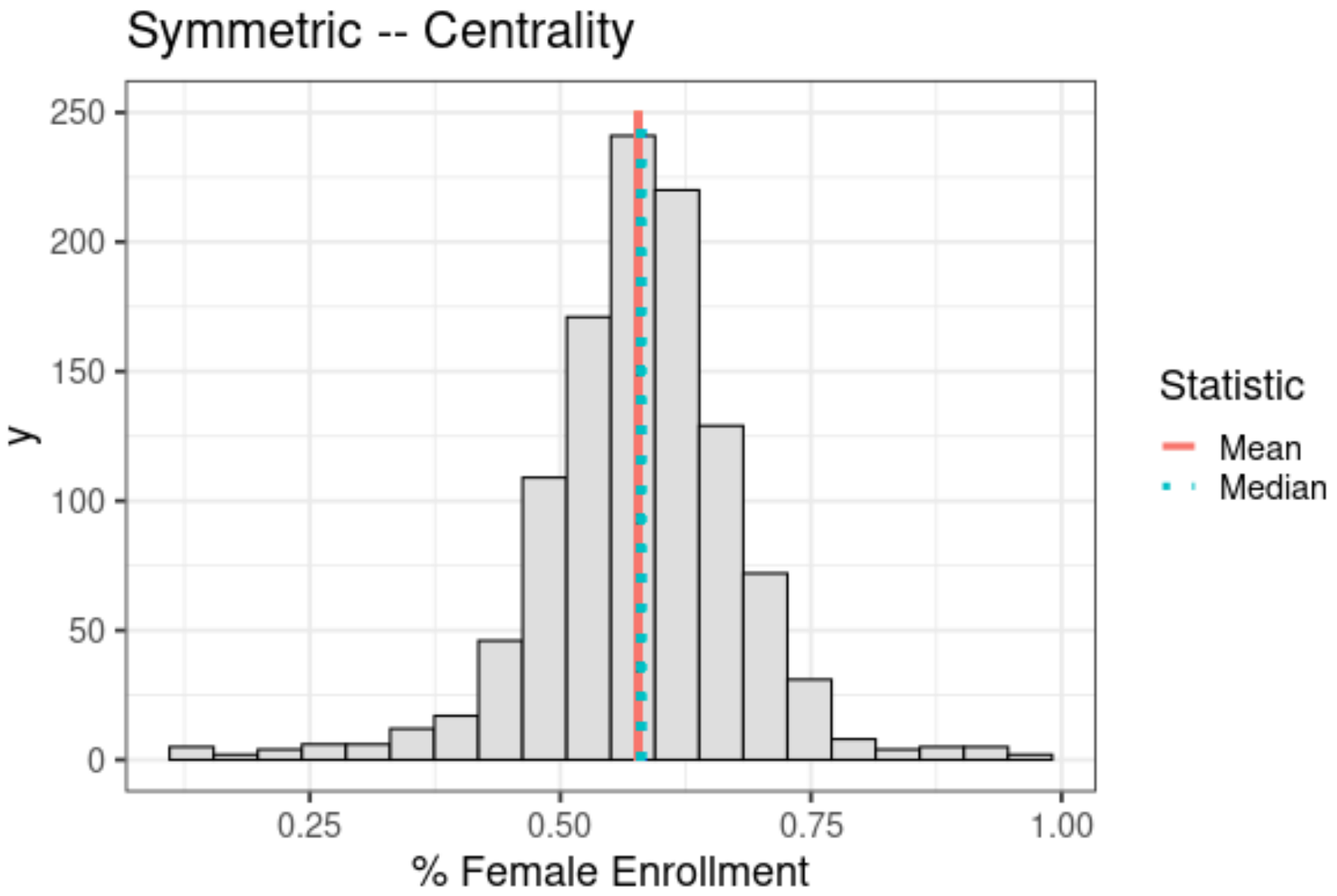
$$\overline{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$$

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

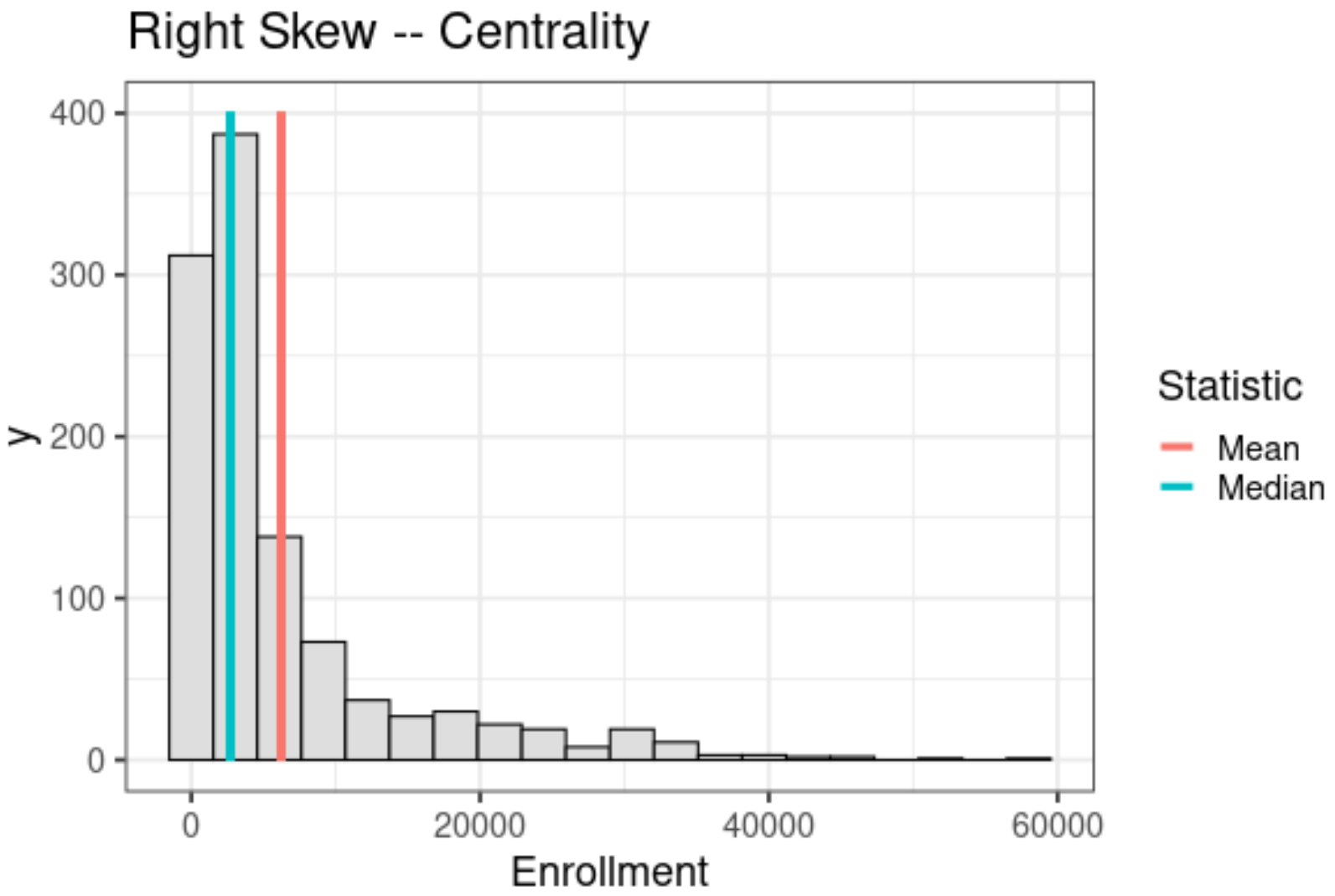
Unlike the median, which only uses the rank values of the observation (and the magnitude of a single observation), the mean uses the magnitude of all of the observed values

This gives us a sense of an *expected value*

Comparing Mean with Median



Comparing Mean with Median



Standard Deviation

Greek letter σ (sigma) for *parameter* and s for *statistic*

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

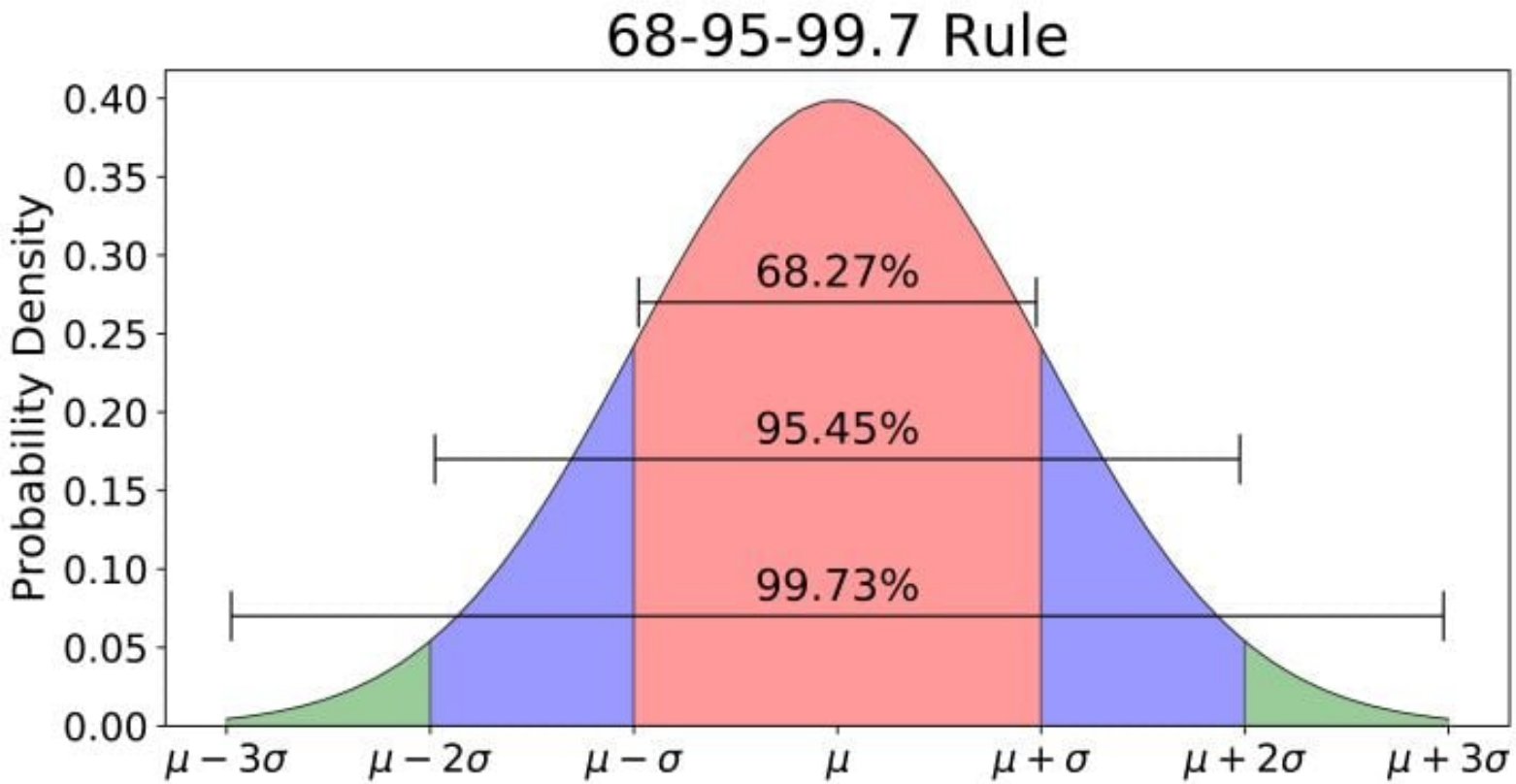
The **standard deviation** provides a measure (kind of) of the average expected distance of our observations from their mean

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

A standardized unit of distance can be used to determine which observations are outliers

68-95-99 Rule (Example)

An example of using moments statistics for bell-shaped distributions



Order and Moment Statistics

Order Statistics

- Based mainly on the ordered position
- Resistant to extreme values
- Useful for skewed distributions

Moment Statistics

- Uses the magnitude of each observation
- Sensitive to extreme values
- Strong algebraic and inferential properties

Report center and spread statistics as matched pairs: median/IQR or mean/SD. You should always inspect a plot of the data before deciding which summaries are useful