Experiments and Numerical Data

July 30, 2019

Issues With Labs

rstudio.cloud/projects

Experiments

Studies where researchers assign treatments to cases are **experiments**.

Whenever an experiment utilizes randomly assigned treatments, we say that it is a **randomized experiment**.

- Note: ”treatment” refers to whatever explanatory variable we are most interested in.

Principles of Experimental Design

Four key principles:

1. Controlling
2. Randomization
3. Replication
4. Blocking

Principles of Experimental Design: Controlling

When treatments are assigned to cases, researchers do their best to **control** any other differences in the treatment groups.

- For example, if both groups are given a pill to take with water, we might instruct everyone to drink the full 8 oz of water in order to control for any impact of water consumption.

Principles of Experimental Design: Randomization

Researchers **randomize** cases into treatment groups.

- This helps account for any unmeasured variables.
- For example, if we're studying a new cancer therapy and dog ownership has a positive impact on cancer outlook, randomization helps ensure that we have similar numbers of dog owners in each treatment group.
- This helps minimize bias in our data.

Principles of Experimental Design: Replication

The more information we have, the more confident we can be in our results! We gather more information through **replication**.

- Suppose we have 3 treatment groups. Replication is just testing each treatment multiple times (multiple cases are assigned to each treatment group).

Principles of Experimental Design: Blocking

If we suspect (or know) that other variables are important in influencing a response, we can group cases into **blocks**.

- Cases within each block are then randomly assigned to each treatment.
- For example, if we are looking at a new asthma medication, we might block individuals by high, medium, and low severity of asthma. Then half of the individuals in each block would be assigned to the new medication.
- This helps ensure that each treatment group has similar numbers of patients from each severity level.

Principles of Experimental Design

- All experiments will use some form of controlling, randomization, and replication.
- Blocking is a slightly more advanced technique (in that it requires slightly more advanced methods to analyze).
- You will learn more about blocking if you take STAT 100B.

Bias in Human Experiments

Randomized experiments are the gold standard, but even they have their limitations!

Experiments involving people are especially prone to bias.

Example: Heart Attack Drugs

Suppose we are interested in whether a new drug helps to prevent repeated cardiac events in patients who have already had at least one heart attack.

- We get a random sample of 100 people who have had a heart attack in the past.
- 50 of them are randomly assigned to the treatment (our new drug). This is our **treatment group**.
- The other 50 do not receive the drug. This is our **control group**.

Can you think of anything that could bias our results?

Sources of Bias

- People who get the new drug expect it to work.
- E.g., people who did not get the drug may wonder if their study participation was worth the risk.
- Doctors may inadvertently affect the results through their optimism (or lack thereof) when administering the drug.

Reducing Bias in Human Experiments

We can reduce bias by

- Keeping patients uninformed about their treatment group.
  - We call these studies **blind**.
  - One way to keep studies blind is to give the control group a **placebo**.
- Keeping doctors uninformed about which treatment groups their patients are in.
  - We call these studies, where neither patient nor medical provider know the treatment group, **double-blind**.

Can you think of some ethical issues that may arise in randomized, double-blind, placebo-controlled studies?

Summarizing Data

Chapter 2 is all about summarizing data through summary statistics and graphs. We can get a lot of information out of these things!

These concepts are also important foundations for the rest of the course.

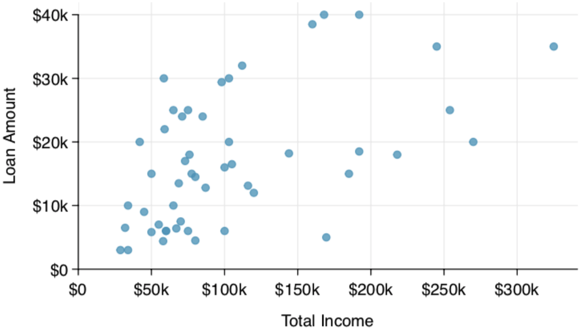
Numerical Data

Let's start by thinking of a simple numeric variable: the ages of everyone in this room.

Can you think of any ways to summarize all of our ages in only one or two numbers?

Scatterplots

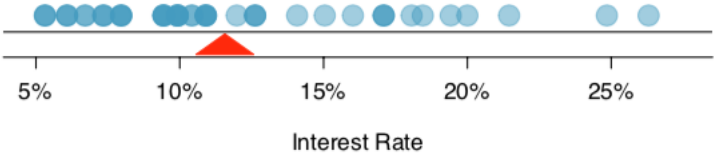
A **scatterplot** shows a case-by-case view of two numerical variables.



What can we learn from the scatterplot?

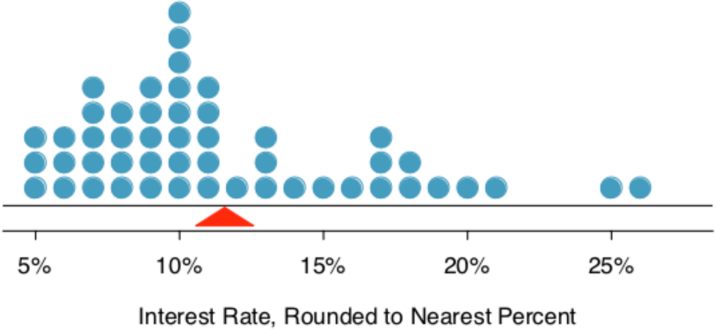
Dot Plots

A **dot plot** is like a scatterplot with only one variable. It shows how a single, *continuous* numerical variable falls on a number line.



Dot Plots

A **stacked dot plot** shows the same information for a *discrete* numerical variable.

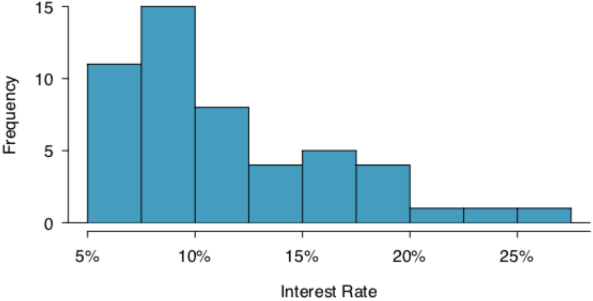


Histograms

A **histogram** is similar to a dot plot, but instead of showing the exact value for each observation, values are put into **bins**.

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	···	25.0% - 27.5%
Count	11	15	8	4	···	1

Figure 2.5: Counts for the binned `interest_rate` data.



The Mean

Both of the dot plots had a red arrow pointing to the **mean** (or **average**) of the variable.

You've probably calculated an average before, but if you haven't (or if you need a refresher), to find the mean you add all of the values and then divide by the number of values.

The Mean

For example, if we had a variable called `ages` with the values 21, 22, 26, 18, 19, and 21, the mean would be

$$\frac{\text{sum of values}}{\text{total \# of observations}} = \frac{21 + 22 + 26 + 18 + 19 + 21}{6}.$$

We denote the mean by $\boldsymbol{\bar{x}}$. In this case, $\bar{x} = 21.167$

The Mean

In math notation, the formula for the mean looks like this:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

In our example, $n = 6$ observations and each x_i is one of our ages.

Measures of Center

The mean is a common way to measure the center (middle) of the **distribution** of the data.

You can think of the distribution as the way that the data is *distributed* from left to right on a histogram.

Measures of Center

The mean of a variable is denoted by $\bar{x}$. This is what we refer to as the **sample mean**.

The mean of the entire population is typically something that we don't have exact data on (we usually don't have data for every single member of a population). Instead, we estimate the population mean using a sample mean.

The **population mean** is denoted by $\boldsymbol{\mu}$. This is the Greek letter *mu*.

That's a lot of symbols to remember?

Let's put them all in one place. We will add to this list as we go.

- n: number of observations/cases
- $\bar{x}$: sample mean
- μ: population mean

Data Density

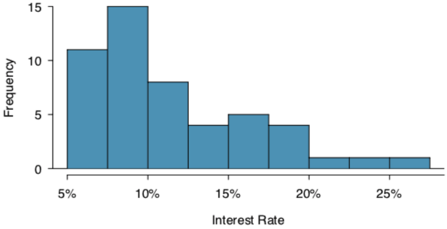
Now that we've brought up the distribution of the data, we can start to think about the density of the data.

Data density refers to the amount of data in any bin. (Taller bins mean more data density, or more data in the bin.)

From here, we can start to consider the *shape* of a distribution.

Shape

Remember our histogram?



- The sides of the distribution (on either side of the mean) are referred to as the **tails**.
- Here the data have a long, thin right tail, so we say that the shape is **right skewed**.

Shape

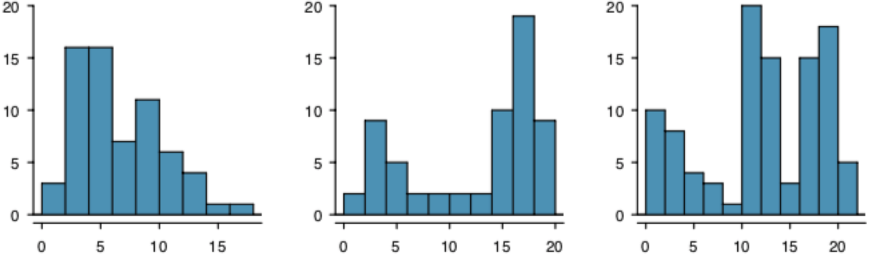
- If the data have a long, thin tail on the left, we say that the shape is **left skewed**.
- If the data have roughly equal tails, we say the distribution is **symmetric**.

Shape

We can also talk about the modes of a distribution. In a distribution, a **mode** is any prominent peak in the distribution. These can be found in a histogram!

- A distribution with one prominent peak is called **unimodal**.
- Distributions with two prominent peaks are **bimodal**.
- Distributions with three or more promiment peaks are **multimodal**.

Modes



How many modes are there in each distribution?
Remember that we only count *prominent* peaks.

Modes

Bin widths, our particular sample, and differing opinions can all impact where we see a ”prominent” mode.

...but that is okay! The goal of examining the shape of our data is simply to better understand the nature of our data. This allows us to make more informed technical decisions down the line.

Variability

We talked about the mean as a way to measure the center of the data, but the variability of data is also an important consideration.

Why might the variability be important?

Why Variability?

Suppose we want to know the average age in this class and take two random samples of size 10 each.

Sample 1:	22, 19, 20, 18, 20, 21, 20, 22, 20, 18
Sample 2:	12, 18, 32, 21, 19, 19, 17, 21, 22, 19

In both cases, we get a sample average of $\bar{x} = 20$.

How confident are you about our estimate of the average age in this class using Sample 1? What about Sample 2?

Variability

We can think about variability as how far away the observations are from the mean.

The distance between an observation and its mean is called the **deviation**. From Sample 1 (22, 19, 20, 18, 20, 21, 20, 22, 20, 18), the deviations for the first, second, and tenth observations are

$$
\begin{aligned}
x_1 - \bar{x} &= 22 - 20 = 2 \\
x_2 - \bar{x} &= 19 - 20 = -1 \\
x_{10} - \bar{x} &= 18 - 20 = -2
\end{aligned}
$$

Variability

We're interested in how far a typical observation is from the mean, but if we add up all of the deviations for a sample, we always get zero! Let's try it on Sample 1:

$$\begin{aligned}
&(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_9 - \bar{x}) + (x_{10} - \bar{x}) \\
&= (22 - 20) + (19 - 20) + (20 - 20) + (18 - 20) + (20 - 20) \\
&\quad + (21 - 20) + (20 - 20) + (22 - 20) + (20 - 20) + (18 - 20) \\
&= 2 + (-1) + 0 + (-2) + 0 + 1 + 0 + 2 + 0 + (-2) \\
&= 2 - 1 - 2 + 1 + 2 - 2 \\
&= 0
\end{aligned}$$

Note: A short proof of this will be posted on the course website.

Variability

So the average deviance doesn't work...

- This is because all of those positives and negatives end up balancing each other out.
- When we talk about variability, we aren't that interested in whether any particular point is above or below the mean.
- We really just want to know how far away it is.

Variability

There are two simple ways to get rid of the signs to focus on distance (without direction).

1. Take the absolute value of the number.
2. Square the number.

It turns out that there are a whole lot of mathematical reasons why it's easier to work with squares than with absolute values!

Variance

And so we come to the **variance**. The variance can be inconvenient to calculate by hand, but it goes something like this:

1. We square all of those deviations we calculated previously.
2. Add them up.
3. Take the average.

We denote our **sample variance** by $\boldsymbol{s^2}$.

Variance

Note: Technically, we divide by $n - 1$ instead of by n when we take our average. We may talk more about this later, but in the meantime just know that there's some mathematical nuance that makes the variance formula a little bit more complicated.

Variance

Let's return to our example and Sample 1. We already calculated our deviations, but this time we square them before adding them up.

$$\begin{aligned}
&(22-20)^2+(19-20)^2+\cdots+(20-20)^2+(18-20)^2 \\
&=2^2+(-1)^2+0^2+(-2)^2+0^2+1^2+0^2+2^2+0^2+(-2)^2 \\
&=4+1+4+1+4+4 \\
&=18
\end{aligned}$$

And then we divide by $n-1=9$

$$\frac{18}{9}=2.$$

Standard Deviation

The variance can be described as the average squared distance from the mean. That probably doesn't sound like a very intuitive way to measure variability.

However, the **standard deviation** is easier to conceptualize than the variance: it gets at our original goal of estimating how far a typical observation is from the mean.

Standard Deviation

Fortunately for us, the standard deviation doesn't require any additional mathematical nuance! In order to calculate the standard deviation, we simply take the square root of the variance.

Returning again to our example,

$$s = \sqrt{s^2} = \sqrt{2} \approx 1.414$$

Standard Deviation

In general,

- 70% of the data will fall within one standard deviation of the mean.
- 95% of the data will fall within two standard deviations of the mean.

...but these are not strict rules!

Population Variability

Like the mean, the **sample variance** and **sample standard deviation** also have population counterparts.

- The **population variance** is denoted $\boldsymbol{\sigma^2}$.
- The **population standard deviation** is denoted $\boldsymbol{\sigma}$.

σ is the Greek letter *sigma*. (We often use Greek letters to denote values from our population.)

Mean and Standard Deviation

Much of what we do in statistics is (1) estimate quantities and (2) determine how uncertain we are about those estimates.

- The mean is often a quantity of interest.
- The standard deviation helps us determine how uncertain we are about this quantity.

We will talk more about uncertainty in Chapter 5.

Symbols to Remember

Let's update our list with variance and standard deviation.

- n: number of observations/cases
- $\bar{x}$: sample mean
- μ: population mean
- s^2: sample variance
- s: sample standard deviation
- σ^2: population variance
- σ: population standard deviation

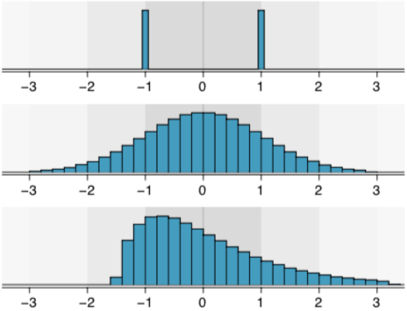
Mean, Standard Deviation, and Shape

Mean, standard deviation, and shape together give us a good description of our distribution.

- If any one of these is missing, we miss crucial information.
- Without the mean, we lack information about the center of the distribution.
- Without the standard deviation, we are unable to capture how spread out the data are.

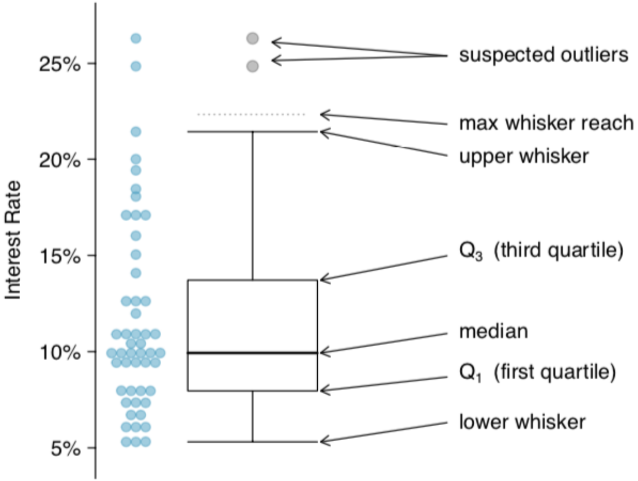
Why Shape?

These three distributions have the same mean ($\bar{x} = 0$) and standard deviation ($s = 1$)!



A good description of shape should include modality and skewness (or symmetry). To give an even clearer picture, we can report where the modes are and the sharpness of the peaks.

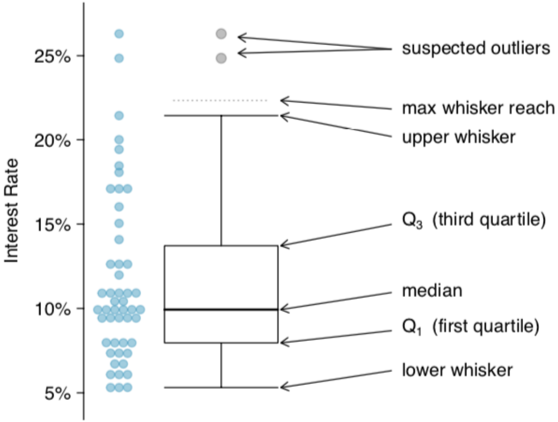
Box Plots



A stacked dot plot next to a vertical box plot.

The Median

The first step in constructing a box plot is to draw a line at the median.



The Median

- The **median** takes the data and splits it in half.
- The median is also called the **50th percentile** because 50% of the data is below this value.
- The median is another measure of center.
- To find the median, we sort our numerical variable and then find the halfway point.

The Median

If we have an odd number of observations, say,

$$1, 2, 3, 4, 5$$

we take the observation in the middle (the $\frac{n+1}{2}$th observation).

In this case,

$$1, 2, \mathbf{3}, 4, 5$$

3 is the median.

The Median

- If we have an even number of observations

  $$1, 2, 3, 4, 5, 6$$

  we cut the data exactly in half

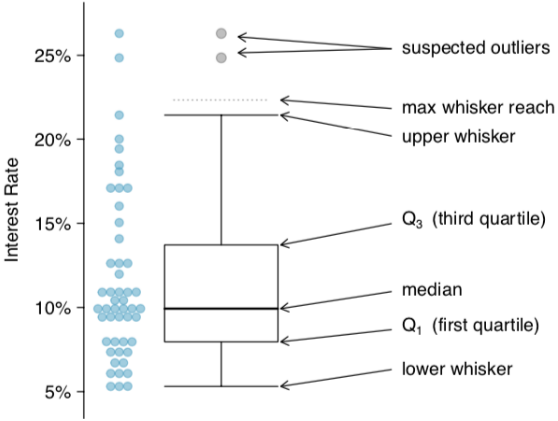
  $$1, 2, 3 \quad | \quad 4, 5, 6$$

  and the median is the average of the two observations closest to the halfway point

  $$\frac{3+4}{2} = 3.5$$

Quartiles

The next step in our box plot is to draw a box connecting the first and third quartiles.



Quartiles

Quartiles split our data into *quarters*.

- 25% of the data falls below the **first quartile** (Q1).
    - This is the 25th percentile.
- 50% of the data falls below the median.
- 75% of the data falls below the **third quartile** (Q3).
    - This is the 75th percentile.

What percent of the data falls between Q1 and the median? What percent between Q1 and Q3?

Finding Quartiles

1. Find the median.
2. Take all of the data that falls *below* the median and find the middle of that data using the same steps we used to find the median. This is the first quartile.
3. Repeat with the data that falls *above* the median. This is the third quartile.

Interquartile Range

- The distance between the first and third quartiles is referred to as the **interquartile range** (or IQR).
- This value is easy to calculate!

$$IQR = Q3 - Q1$$

- The IQR is another measure of variability.

Whiskers

Now we need to find the whiskers.

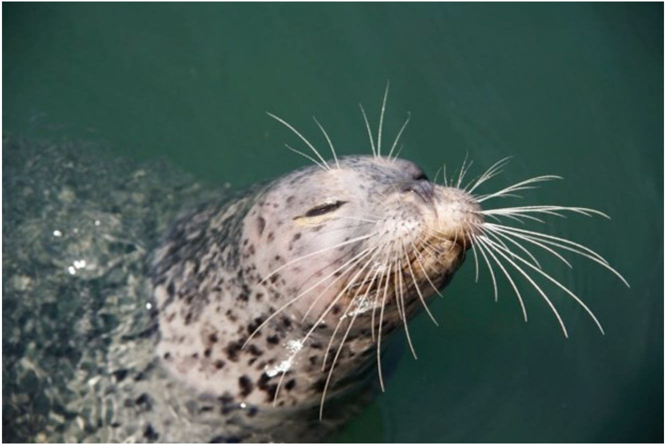
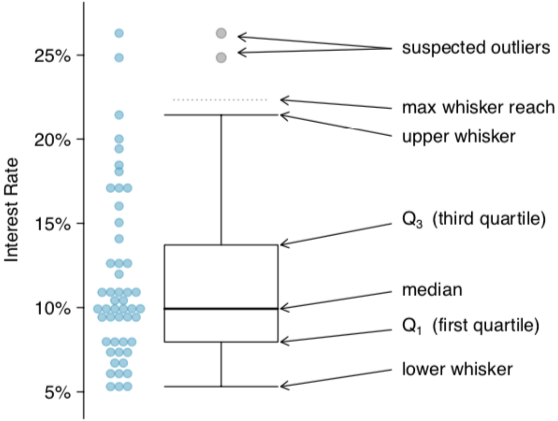


Image from BBC Wildlife
www.discoverwildlife.com/animal-facts/mammals/how-do-whiskers-work/

Whiskers

Now we need to find the whiskers.



Whiskers

The **whiskers** capture (most of) the rest of the data.

- Each whisker is no longer than

  $$1.5 \times IQR.$$

  and stops at the point closest to, but still within, this range.

Whiskers

- The upper whisker goes no farther than

  $$Q3 + 1.5 \times IQR$$

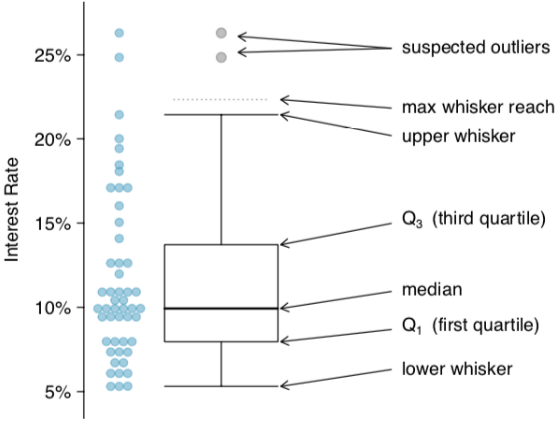
  and the lower whisker no farther than

  $$Q1 - 1.5 \times IQR$$

We may choose not to include the maximum upper reach and minimum lower reach on our box plot, but we always include the whiskers themselves.

Outliers

Finally, we add any outliers by labeling each one with a dot.



Outliers

Since we've already built the rest of our boxplot, we can start to think about outliers as whatever is left out.

- We label these observations specifically because they are *unusual* or *extreme*.
- Observations that are unusually far from the rest of the data are referred to as **outliers**.

Why Examine Outliers?

- Identify sources of strong skew.
- Provide insight into potentially interesting properties of the data.
- Identify possible data collection or data entry errors.

Robust Statistics

Suppose we have some data:

$$3, 6, 7, 4, 10, 8, 1, 5, 2, 9$$

and I replace the largest observation (10) with a significantly larger value (35).

$$3, 6, 7, 4, 35, 8, 1, 5, 2, 9$$

Robust Statistics

For our original data,

$$3, 6, 7, 4, 10, 8, 1, 5, 2, 9$$

we get the following:

median	IQR	$\bar{x}$	s
5.5	4.5	5.5	3.03

What do you think will happen to our sample statistics (mean, median, standard deviation, and IQR) when I replace 10 with 35?

Robust Statistics

Replacing 10 with 35, these numbers shift somewhat:

	median	IQR	$\bar{x}$	s
Original Data	5.5	4.5	5.5	3.03
Modified Data	5.5	4.5	8.0	9.83

The median and IQR are exactly the same, but the mean and standard deviation change quite a bit!

Robust Statistics

We say that the median and IQR are **robust statistics** or that they are *robust to* outliers, meaning that their values are minimally effected by these extreme observations.

	Robust		Not Robust	
	median	IQR	$\bar{x}$	s
Original Data	5.5	4.5	5.5	3.03
Modified Data	5.5	4.5	8.0	9.83

Why do you think the mean and standard deviation changed so much, but the median and IQR did not?

When Are Robust Statistics Important?

- Suppose you wanted to know about the typical home price in the United States in 2018.
- Recall that the mean and median are both measures of center.
- Would you look at the mean or the median? Why?

When Are Robust Statistics Important?

As long as you can defend your answer, there is value to each option!

- If we wanted to know what the typical homeowner is spending, the median would be more useful.
- If we wanted our estimate to scale, e.g., to estimate how much total money was spent on homes in 2018, the mean might be a better option.