Quantitative Variables – Part 1

Numerical Summaries and Visualization

Grinnell College

Goals for Class Today

We are going to learn how to do the following:

1. use graphs and plots to display data
2. describe what we see
    - establish a shared language for description

These are not always as easy as they sound

Motivation

Below is a data frame of 20 observations (out of a total 244) regarding the tips given to one waiter over the course of several months in one restaurant. This is a *few* years old

Total Bill	Tip	Sex	Smoker	Day	Time	Size
13.42	1.58	Male	Yes	Fri	Lunch	2
16.27	2.50	Female	Yes	Fri	Lunch	2
10.09	2.00	Female	Yes	Fri	Lunch	2
20.45	3.00	Male	No	Sat	Dinner	4
13.28	2.72	Male	No	Sat	Dinner	2
22.12	2.88	Female	Yes	Sat	Dinner	2
24.01	2.00	Male	Yes	Sat	Dinner	4
15.69	3.00	Male	Yes	Sat	Dinner	3
11.61	3.39	Male	No	Sat	Dinner	2
10.77	1.47	Male	No	Sat	Dinner	2
15.53	3.00	Male	Yes	Sat	Dinner	2
10.07	1.25	Male	No	Sat	Dinner	2
12.60	1.00	Male	Yes	Sat	Dinner	2
32.83	1.17	Male	Yes	Sat	Dinner	2
35.83	4.67	Female	No	Sat	Dinner	3
29.03	5.92	Male	No	Sat	Dinner	3
27.18	2.00	Female	Yes	Sat	Dinner	2
22.67	2.00	Male	Yes	Sat	Dinner	2
17.82	1.75	Male	No	Sat	Dinner	2
18.78	3.00	Female	No	Thur	Dinner	2

Do more customers come to the restaurant on certain days?

Hard to tell by looking at table

Motivation

Data collection has made remarkable progress in the last few decades, giving us a greater quantity of data than most could ever dream of. However, just looking at tables of data is not very useful.

Better approaches:

1. **Data Visualization** displaying data in ways that make patterns more noticeable
2. **Numerical Summaries** calculating numbers that tell us about certain aspects of the data

Collectively these are both sometimes called *descriptive statistics*

Data Visualization

Why do we graph data?

- It (hopefully) allows us to interpret the data...
  - quickly
  - easily

The type of graphs we use are determined by:

- type of data (quantitative vs. categorical)
- how many variables we are working with
- context: information we are trying to convey

Distribution

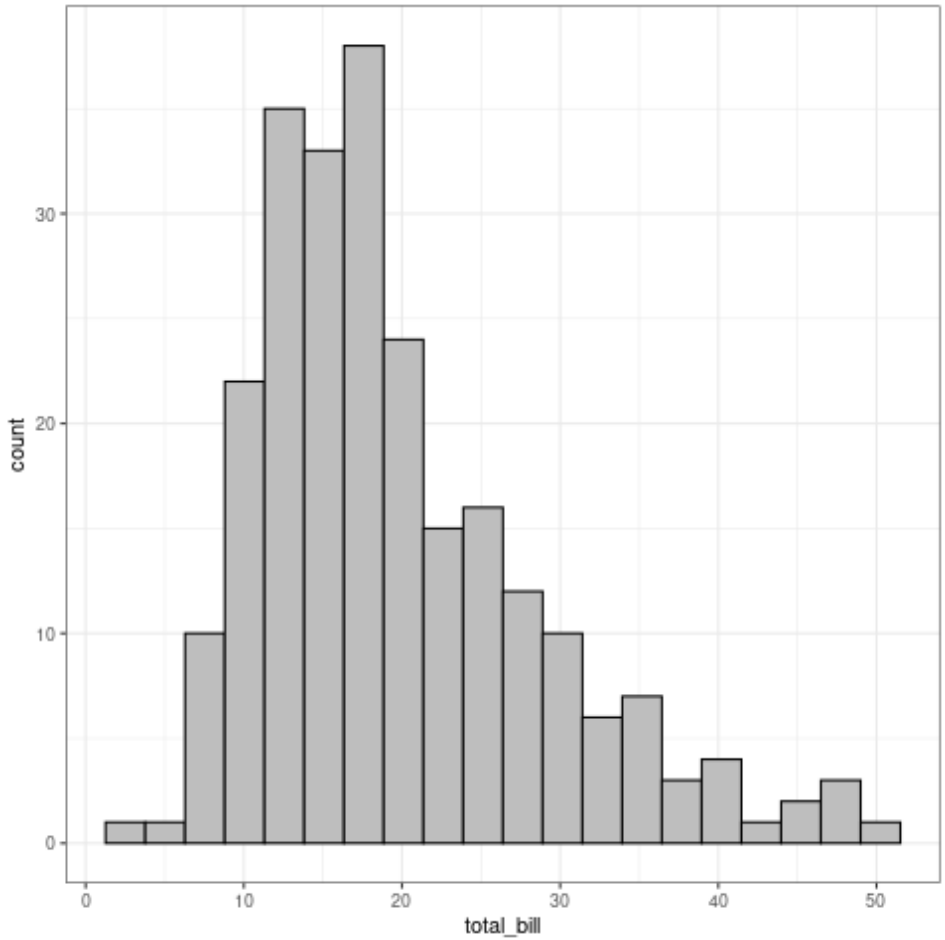
In order to better understand patterns in our data, we will often combine graphics with short descriptions of what we see. A term we will use often:

The **distribution** of a variable is a description of how frequently certain values of that variable show up in our data

- the values
- how often each value shows up

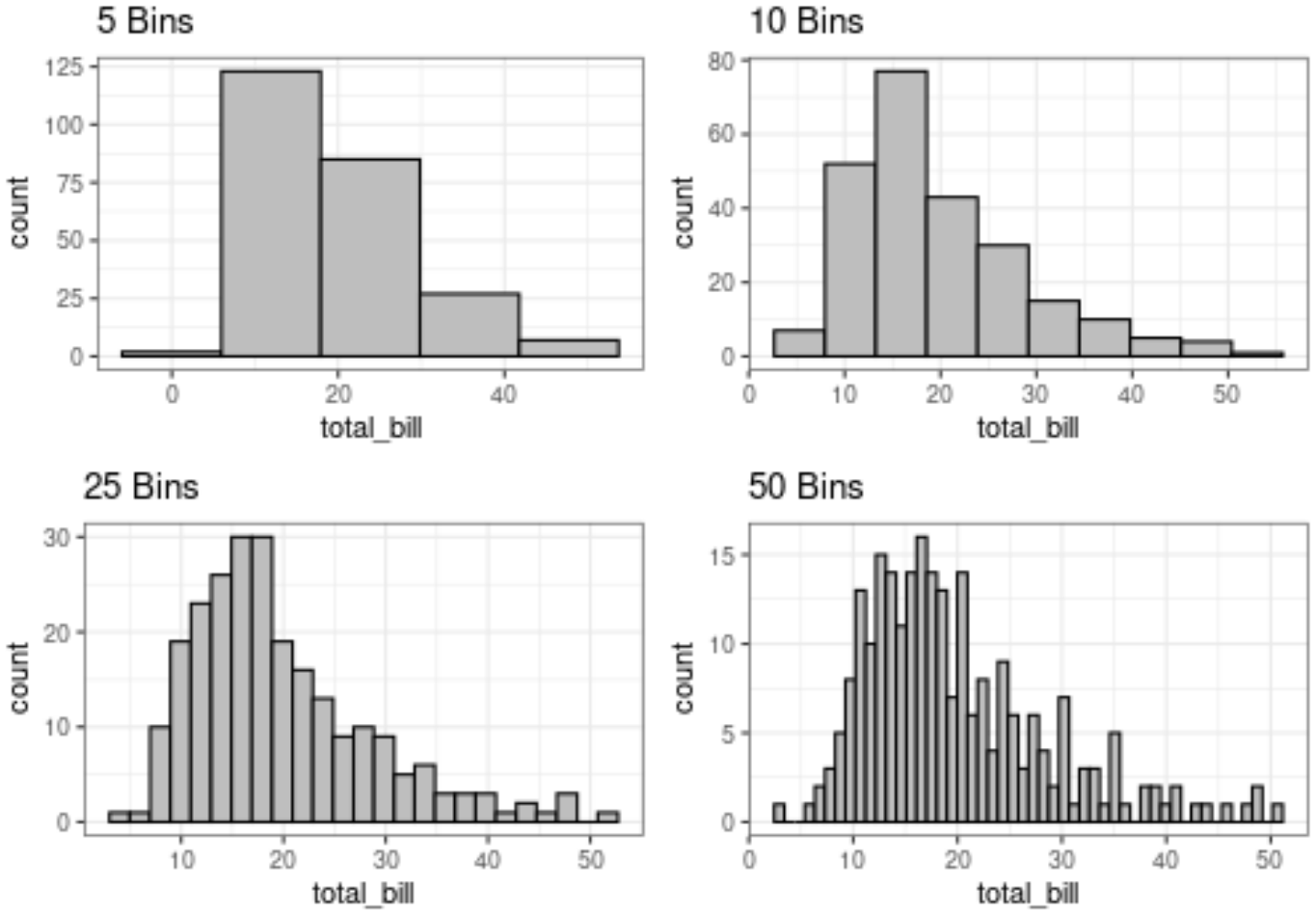
Histograms

A **histogram** is often used to show the distribution of values for a quantitative variable. Histograms group values into equally spaced intervals called bins, then make rectangles with height equal to the counts

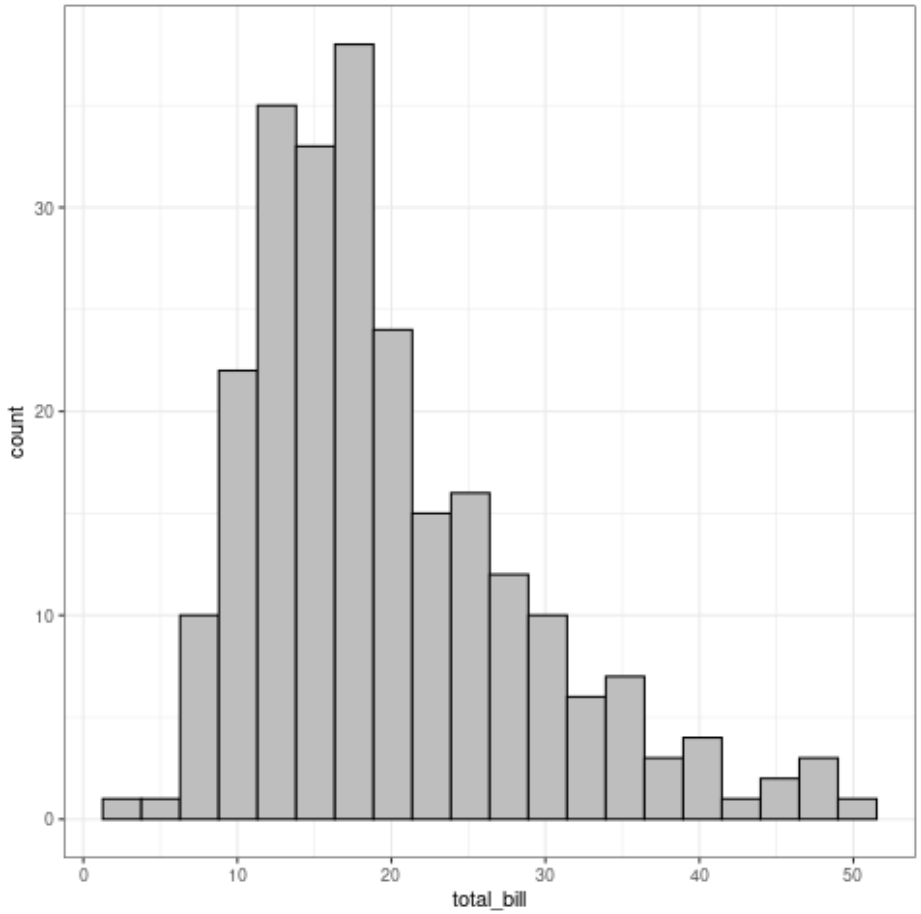


Histogram Bin Width

Using wider/narrower bin width can drastically change the histogram
- too wide: can't tell exactly where data points are
- too narrow: overly detailed and hard to read



Histograms – Description



There is quite a bit we can examine:

- What is a 'typical' value?
- How spread out is this data?
- Does it appear skewed (more data on one side?)

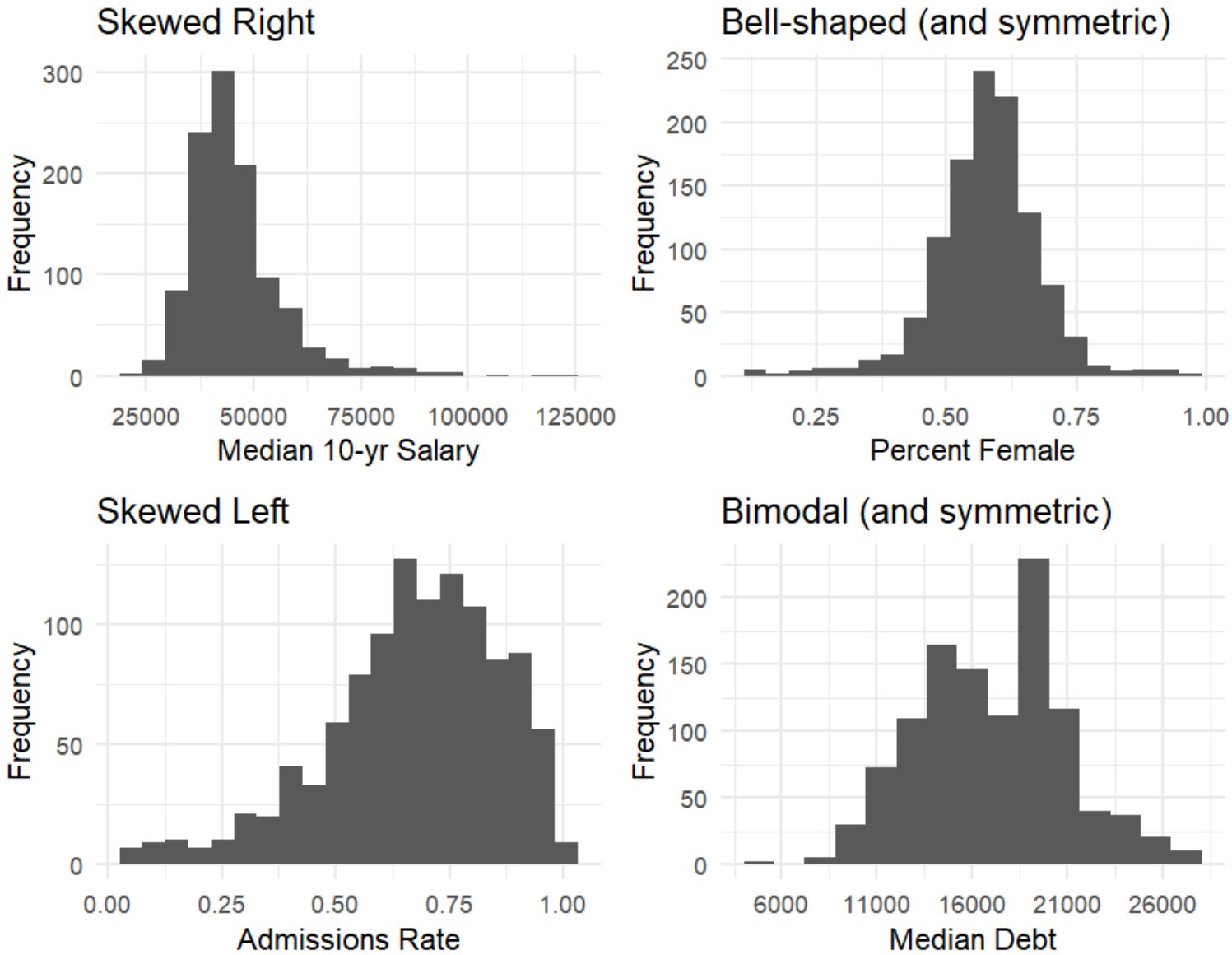
Quantitative Variable - Distribution

When describing the **distribution** of a quantitative variable we need to mention all of these things so we understand how it looks:

- **Shape** - is the distribution symmetric, skewed, number of 'modes'?
- **Center** - where does the data bunch up
- **Spread** - how spread out is the data (ie: range of values)
- **Outliers** - are there values that are much smaller/larger?
- **Context!!** - information related to what we are working with

Distribution - Shape Examples

The following is info from 1059 colleges all over the US

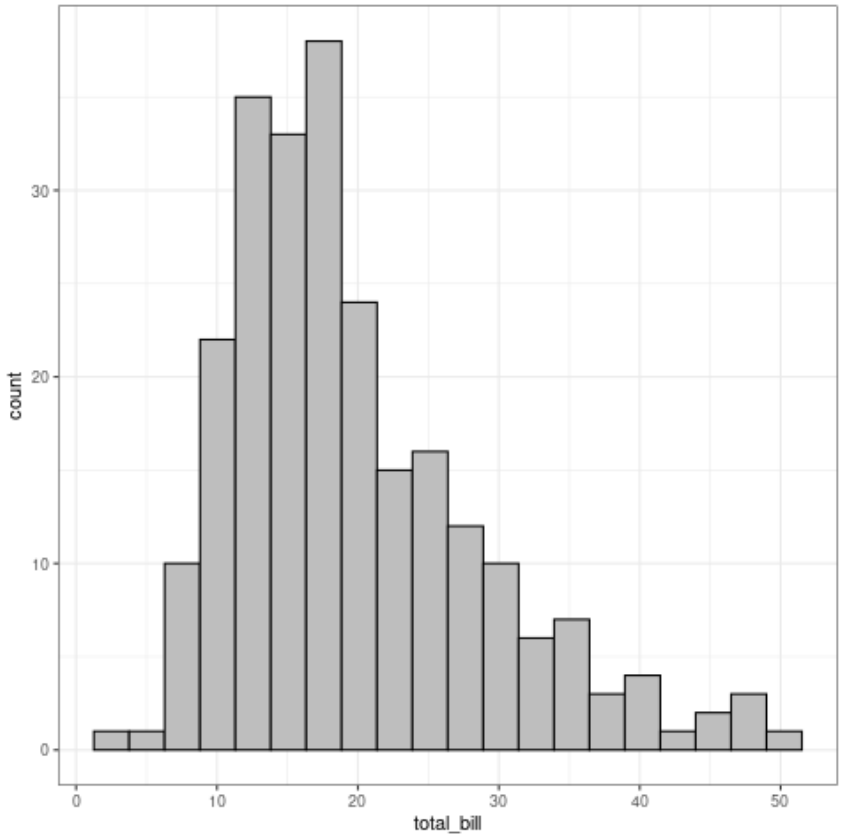


Outliers

Outliers are data points that look *unusual* in that they either don't follow a pattern that we see in the data or are far away from other points

In a histogram, we identify outliers by looking for data separated from the rest by gaps in the bins.

Practice – Describing a Distribution



The distribution of total bill amount (in dollars) for the waiter at this restaurant is...

Numerical Summaries

Numerical Summaries are numbers that condense info about a variable into something easier to understand

Example:

Total Bill	Tip	Sex	Smoker	Day	Time	Size
13.42	1.58	Male	Yes	Fri	Lunch	2
16.27	2.50	Female	Yes	Fri	Lunch	2
10.09	2.00	Female	Yes	Fri	Lunch	2
20.45	3.00	Male	No	Sat	Dinner	4
13.28	2.72	Male	No	Sat	Dinner	2
22.12	2.88	Female	Yes	Sat	Dinner	2
24.01	2.00	Male	Yes	Sat	Dinner	4
15.69	3.00	Male	Yes	Sat	Dinner	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮

- What's the typical table size?
- Roughly how much do customers spend on a meal?
- How much disparity is there among bill amounts?

Note: Many times these are quite similar to center/spread which we just talked about!

Numerical Summaries

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

- What are the most common values?
- Where does the data bunch up?
- What is the 'typical' value that most observations had?

The **spread** of a distribution is meant to tell us, literally, how spread out the values are. **Spread** also tells us how much *variability* there is

There are two separate approaches for describing center and spread

1. Order Statistics
2. Moment Statistics

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. Work well regardless of what data looks like
2. Are generally robust to (unaffected by) skews/outliers

Percentiles

A **percentile** is a number such that some percent of our (quantitative) observations are equal to or smaller than this number

- helpful to have values arranged smallest to largest

Data = {1, 2, 4, 4, 4, 5, 5, 7} (8 total observations)

Example: 25% percentile (25% of values are equal to or smaller than this)

25% $\times$ 8 observations = 2 observations
We will find the 2nd observation in the data (smallest to largest) and this is the 25th percentile
{ 1, 2, 4, 4, 4, 5, 5, 7}

So... 25th percentile of this dataset is 2

Percentiles

A **percentile** is a number such that some percent of our (quantitative) observations are equal to or smaller than this number

- helpful to have values arranged smallest to largest

Note: Finding most Percentiles are not too bad with an even number of observations. Sometimes it is harder with odd number of observations.

Data $= \{2, 4, 10, 11, 13\}$

Finding 25th percentile:
$\frac{1}{5} = .2 = 20\%, \quad \frac{2}{5} = .4 = 40\%$

4 is the 25th percentile. It is also *every* percentile between and including 40th percentile too

Percentiles

A **percentile** is a number such that some percent of our (quantitative) observations are equal to or smaller than this number

- helpful to have values arranged smallest to largest

Another trick for finding percentiles: multiply percentage by # of observations to see how far over you have to count (always round up!)

Example: 25th percentile again
Data $= \{2, 4, 10, 11, 13\}$

There are 5 observations total
$25\% \times 5 = 0.25 \times 5 = 1.25$

round up to next number (2)
$\rightarrow$ the 2nd observation is the 25th percentile
$\rightarrow$ 25th percentile $= 4$

Percentiles

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

The **median** is another name for the 50th percentile. It cuts our data set in half (50% of observations are less than median, 50% more than)

Example 1 (odd # of observations), find value with equal amount of observations on either side

Data $= \{ 1, 3, 3, \boxed{6}, 7, 8, 9\} \rightarrow$ median $= 6$

Example 2 (even # of observations), find two innermost values and average them

Data $= \{ 3, 3, \boxed{6, 7}, 8, 9\} \rightarrow \frac{6+7}{2} = 6.5 \rightarrow$ median $= 6.5$

IQR

Q1 = first quartile = 25th percentile (a quarter of values are less than Q1)

Q2 = third quartile = 75th percentile (3 quarters of values less than Q3)

The **interquartile range** or **IQR** is the value of $Q_3 - Q_1$, and gives us the range of the middle 50% of our data

IQR

The **interquartile range** or **IQR** is the value of $Q_3 - Q_1$, and gives us the range of the middle 50% of our data

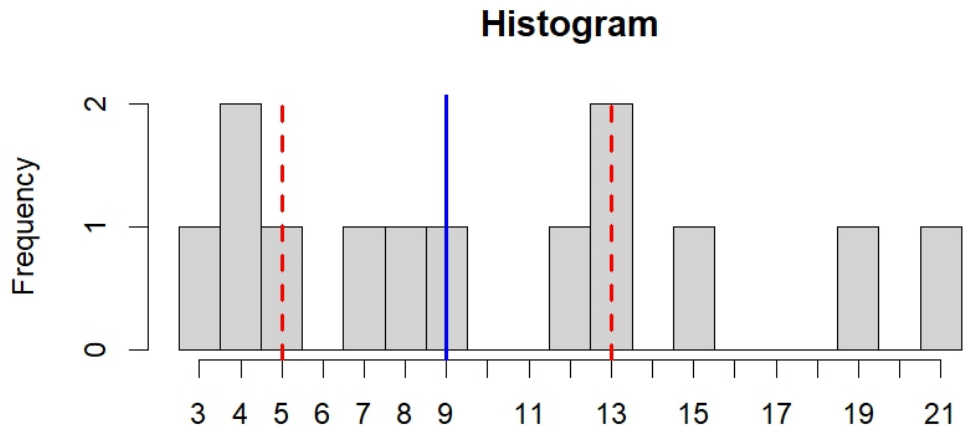
Data: $\{3, 4, 4, 5, 7, 8, 9, 12, 13, 13, 15, 19, 21\}$

Note: $.25 \times 13 = 3.25 \rightarrow 4, \qquad .75 \times 13 = 9.75 \rightarrow 10$

$\{3, 4, 4, 5, 7, 8, 9, 12, 13, 13, 15, 19, 21\} \rightarrow \text{Q1} = 5$

$\{3, 4, 4, 5, 7, 8, 9, 12, 13, 13, 15, 19, 21\} \rightarrow \text{Q3} = 13$

IQR = Q3 - Q1 = 13 - 5 = 8



5-Number Summary

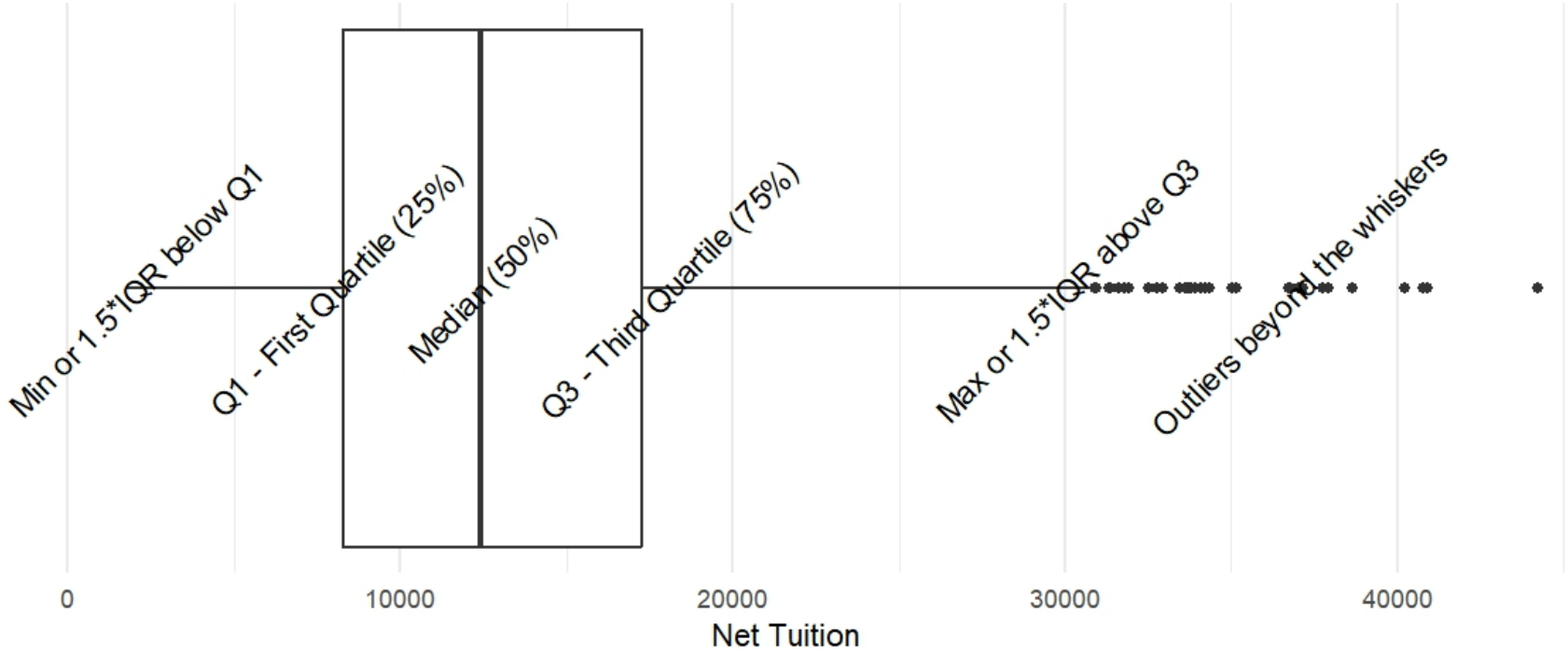
The **5-Number summary** is a collection of 5 numbers that together tell us quite a bit about the data we are looking at.

- Minimum
- Q1 (first quartile, 25th percentile)
- Median
- Q3 (third quartile, 75th percentile)
- Maximum

Box plots

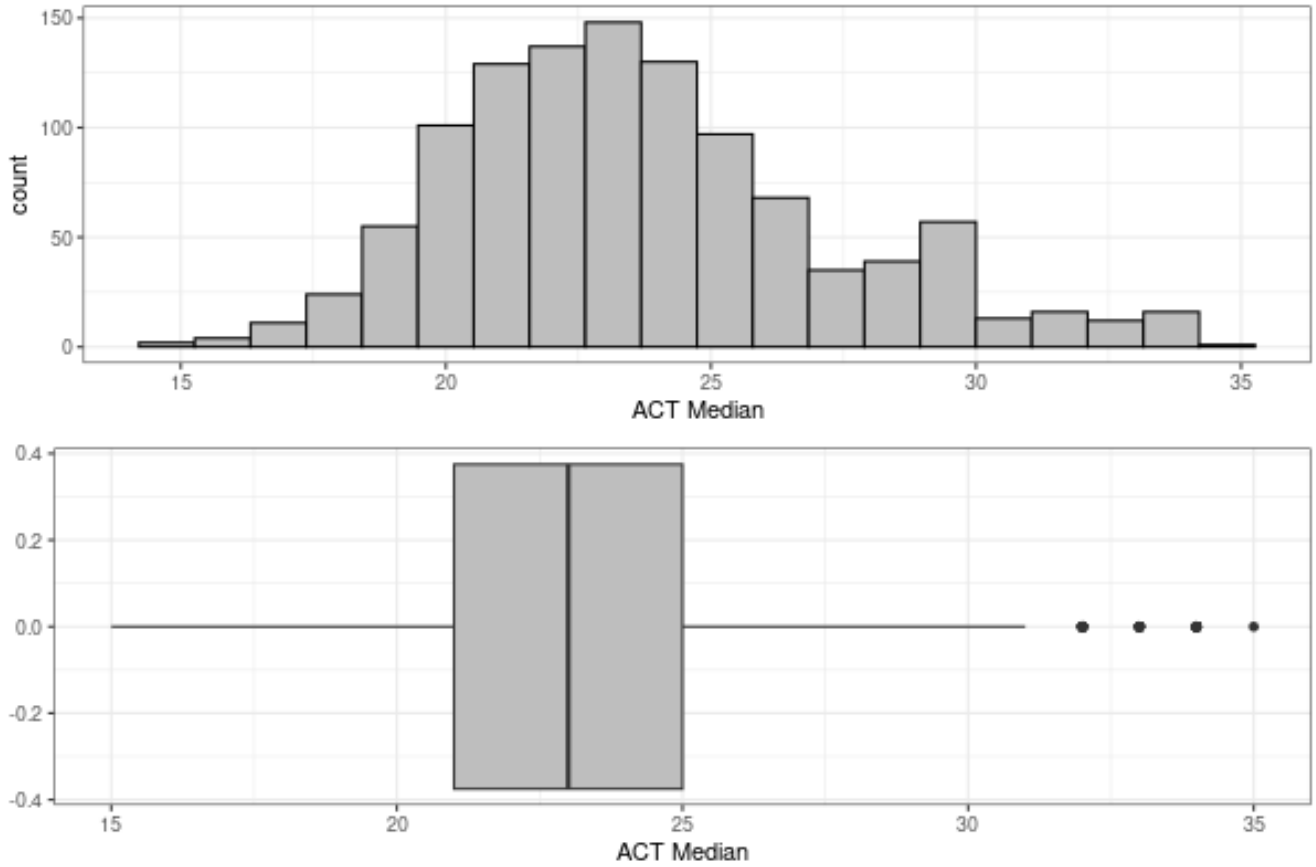
A **Box plot** is another way to display a quantitative variable, specifically it displays the 5-number-summary

ex) 2019 College data, tuition



Note: I am not ever going to make you calculate how far out we need to draw the outer lines

Histogram vs. Box plot



Using either will (generally) give us the same distribution description

- skew is sometimes harder to describe with boxplots
- outliers classification is different

Moment Statistics

Moment statistics are statistics that are based on specific mathematical properties of our data. There are some very powerful techniques that can make use of these (we will see some in a few weeks)

Unlike order statistics, moment statistics (largely) *do* care about how the data looks: they are very sensitive to skews and outliers

In this sense, we say that moment statistics *are not* robust

Some Notation

Before we continue, it is helpful to introduce some notation that can make doing calculations with data a little easier.

n often denotes the sample size (# of observations in our sample)

$\mathbf{x}_i$ is used to denote the values for a variable in our data set
i.e.: $\mathbf{x}_3$ is the 3rd value of the variable in the data set

$\sum$ is a symbol frequently used to denote adding a whole bunch of things together.

Mean (Center)

The **mean** is the same thing as the **average** value of the variable.

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 a variable is denoted as $\overline{x}$

Using the notation from the previous slide, the equation for the **mean** is

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

Spread – Standard Deviation

Standard Deviation – a way of measuring the typical deviation (distance) of each observation from the *mean*

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

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

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 better results for stuff we will do later

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**.

- variance = s^2
- harder to interpret
- in certain scenarios it is easier to work with than s

Example – Mean and SD

Data $= \{3, 5, 7\}$ (these are our x_i values)

Calculating Mean

$\bar{x} = \frac{\sum x_i}{n} = \frac{3+5+7}{3} = \frac{15}{3} = 5$

Calculating Standard Deviation

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

Easiest to calculate this in parts:

$\sum(x_i - \bar{x})^2 = (3-5)^2 + (5-5)^2 + (7-5)^2 = (-2)^2 + 0^2 + (2)^2 = 8$

$\frac{1}{n-1}\sum(x_i - \bar{x})^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} = \frac{8}{2} = 4$

$s = \sqrt{4} = 2$

Which measures to use?

Order statistics are robust, moment statistics are not robust.

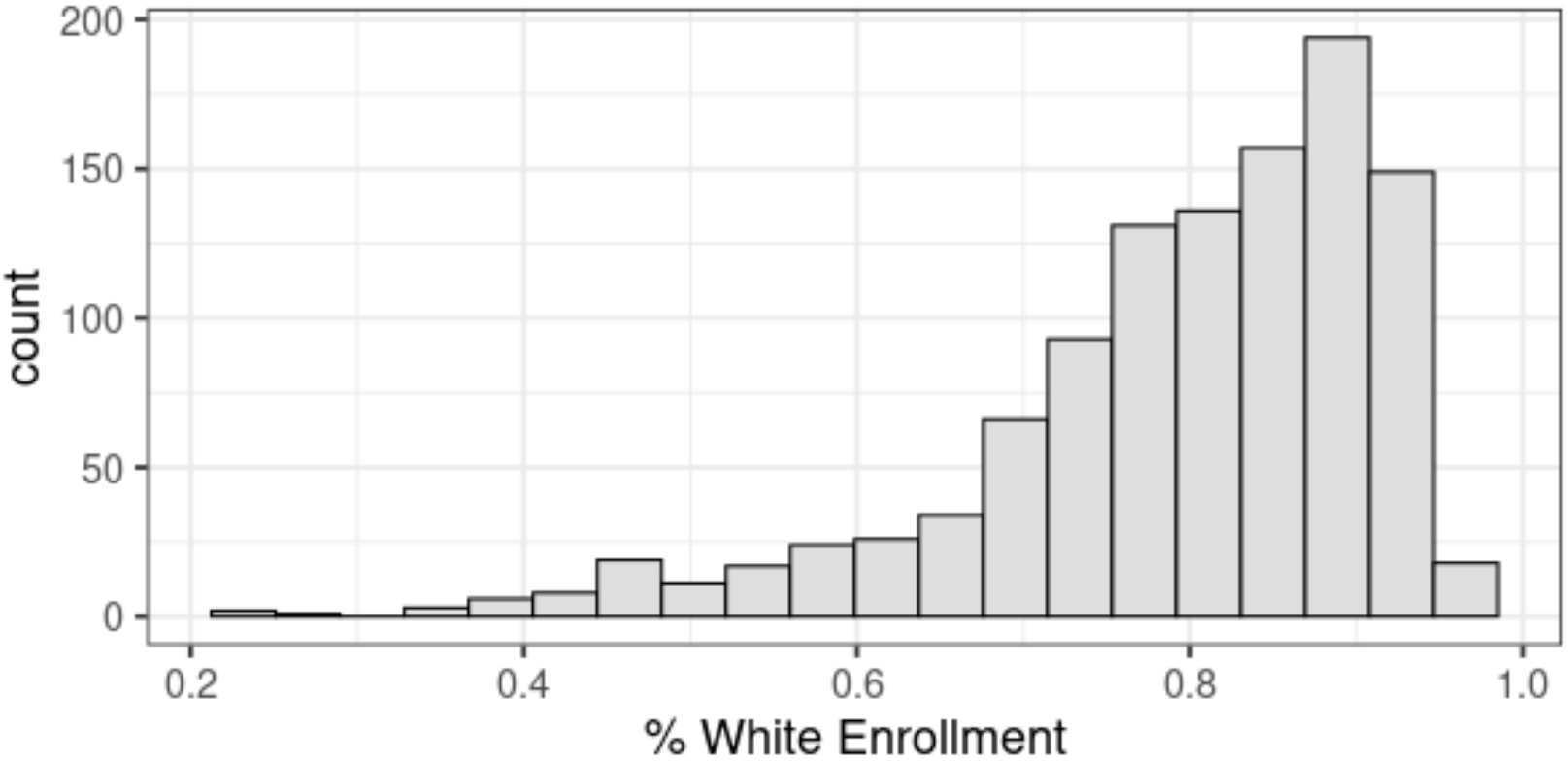
- A skewed distribution can affect the mean and std. dev. a lot
  - skew $\rightarrow$ mean & std. dev. not good measures of center & spread
- Outliers can affect the mean and std. dev. a lot
  - outliers $\rightarrow$ mean & std. dev. not good measures of center & spread

Summary:

Symmetric shape with no outliers $\rightarrow$ mean and std. dev.
Skewed shape or outliers (or both) $\rightarrow$ median and IQR
Multimodal $\rightarrow$ median and IQR

Example – Describing a Distribution (better)

Describe the distribution of '% White Enrollment' for colleges



The percent of white enrollment is skewed-left with a median of 82% and an IQR of 15% (89% - 74%). There are several low percentage outliers.

Note: The average is 79%. If I used average, I am underestimating the typical % white enrollment at colleges