

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

STAT 140: Week 2

My hope is that after Week 2 (09/07-09/10), you will be able to:

- Summarize numerical variables, in graphs and with statistics.
- Summarize categorical variables, in graphs and with statistics.
- Create histograms, density plots, box plots, and bar plots with `ggplot2`.

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Summarizing Data

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Now that we have discussed how to collect data (anecdotal evidence, observational studies, and experiments), as well as different types of variables (numerical, categorical, . . .), we can begin to learn how to summarize data.

- A **distribution** is a summary of what values a variable can take and how often those values occur.
- It is often useful to show a visual display of the distribution when discussing a variable.
- Visual displays for numerical and categorical variables can look quite different.

state.x77

STAT 140:
Week 2

Last week, we discussed loading data into R from a csv file. There are also several datasets included in R that you can use for practice.

- `state.area`: Numeric vector of state areas in square miles
- `state.division`: Factor giving state divisions (New England, Middle Atlantic...)
- `state.region`: Factor giving the region (Northeast, South, North Central, West) that each state belongs to
- `state.x77`: Matrix with 50 rows and 8 columns giving statistics on Population, Income, Illiteracy, Life Expectancy, Murder Rates, Percent HS Graduates, and Area

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

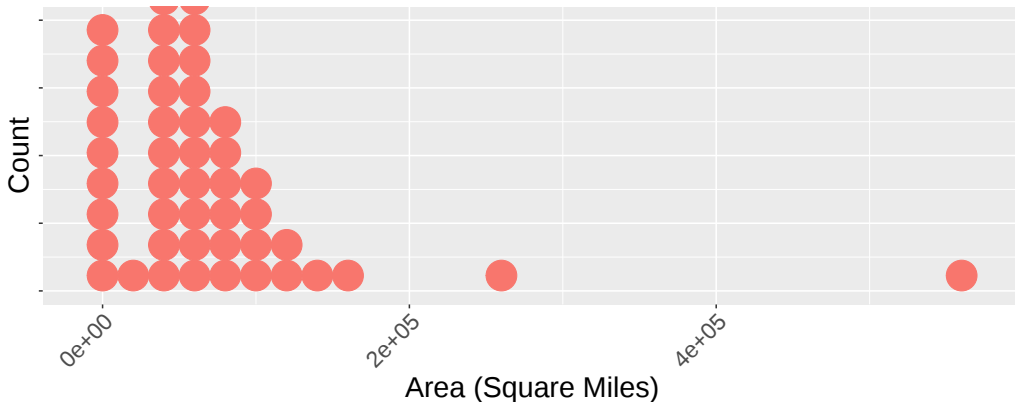
Review

Notation
Review

Stacked Dot Plots

STAT 140:
Week 2

- A **stacked dot plot** is a case-by-case view of data for a single numerical variable.



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

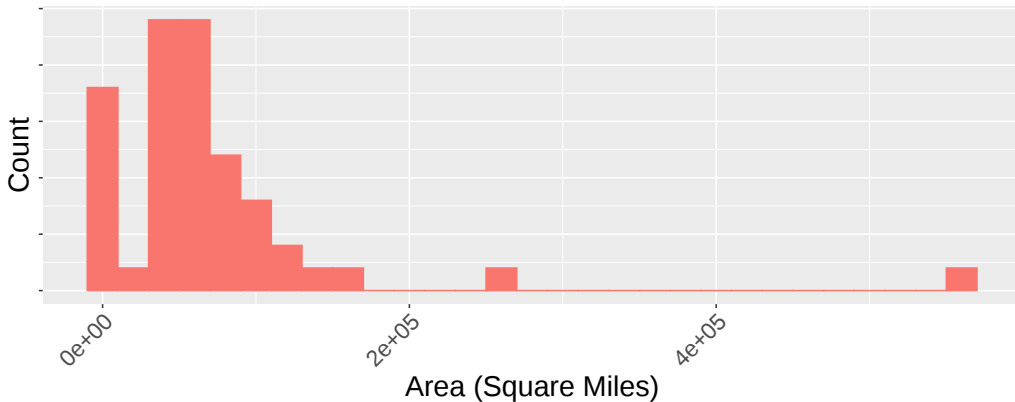
Review

Notation
Review

Histogram

STAT 140:
Week 2

- A **histogram** is a 'binned' version of a stacked dotplot, where 'bins' are ranges to which the data points belong.



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

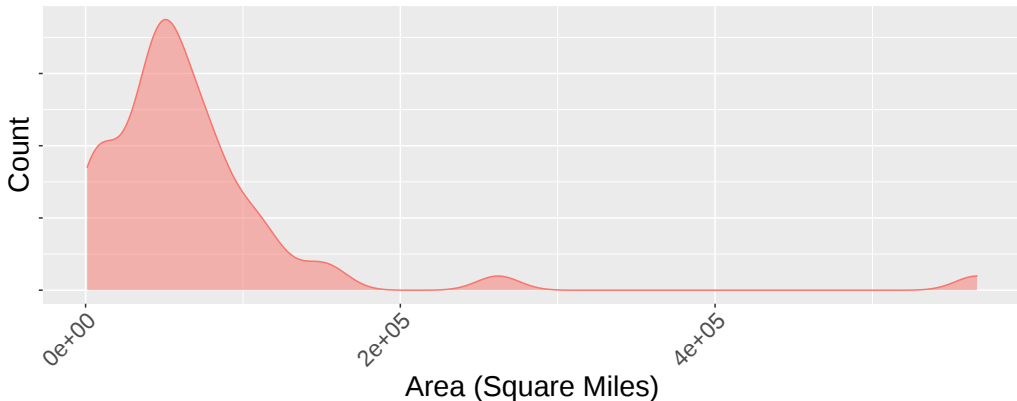
Review

Notation
Review

Density Plot

STAT 140:
Week 2

- A **density plot** can be thought of as a histogram with the binwidth approaching 0 (or the number of bins approaching infinity)—see this [animation](#).



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Displays for Numerical Variables 1

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

- We have already defined a **distribution** is a summary of what values a variable can take and how often those values occur.
- For density plots and histograms,
 - The x-axis is a summary of what values the variable takes.
 - The y-axis dictates how often those values occur.
- Density plots can be taken one step further—there is often a specific function that characterizes the theoretical distribution of a variable in addition to the curve we see.
- These special functions are known as **probability density functions**. We will discuss them more in Chapter 3.

Displays for Numerical Variables 2

STAT 140:
Week 2

To describe a distribution for a numerical variable, you should discuss

- Center – Where is the middle of the data?
- Spread – Are the data close together, or far apart?
- Unusual Values – Are there any values unlike the others?
- Shape – What does the distribution look like?

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

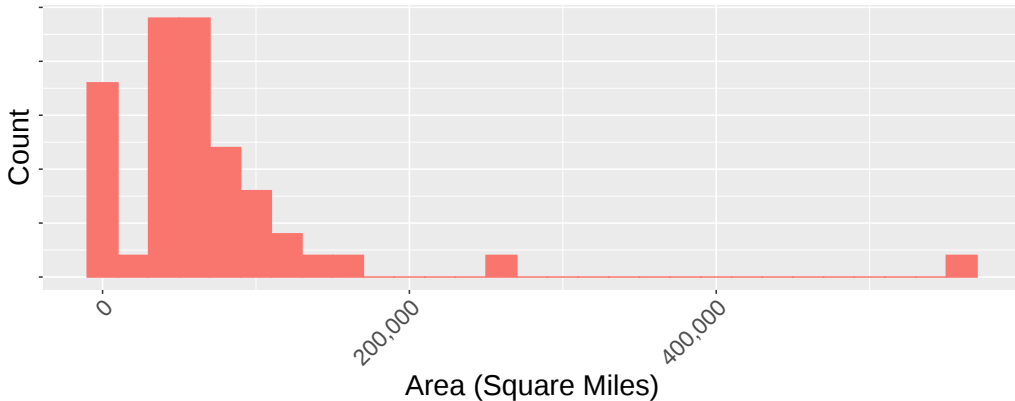
Review

Notation
Review

Example

STAT 140:
Week 2

Talk about the center, spread, shape, and unusual values for this histogram.



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Statistics for Center 1

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

- To formally discuss different aspects (center, shape, spread, and unusual values) of a distribution, we can use statistics in conjunction with a graph.
- What is a statistic?
 - A summary of an aspect of a sample, often a number.
- The **mean** and **median** are summaries of the center of a distribution/graph.
- When discussing the center of a distribution, you should report the mean or median (if available). Otherwise, you should eyeball the center.

Statistics for Center 2

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ Mean

- To compute the mean of a variable, we add up all of the values and divide by the total number of observations.
- The mean of a dataset is also called the average.
- Formally,

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- The R command for calculating the mean is `mean()`.

Statistics for Center 3

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ Median

- The median is the point at which the dataset is split in half.
- To compute the median of a variable, we order the values from smallest to largest.
 - If there are an odd number of observations, the median is the observation in the exact middle.
 - If there are an even number of observations, the median is the average of the two middle observations.
- The R command for calculating the median is `median()`.

Statistics for Center 4

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

```
state.x77.df <- as.data.frame(state.x77)
## To store the data as a data frame, we make a new object.
dim(state.x77.df)
```

```
## [1] 50 8
```

```
colnames(state.x77.df)
```

```
## [1] "Population" "Income"      "Illiteracy" "Life Exp"    "Murder"
## [6] "HS Grad"    "Frost"        "Area"
```

Statistics for Center 5

STAT 140:
Week 2

```
state.x77.df %>%  
  summarize(mean(Area), median(Area))
```

```
##    mean(Area) median(Area)  
## 1    70735.88     54277
```

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Statistics for Center 6

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

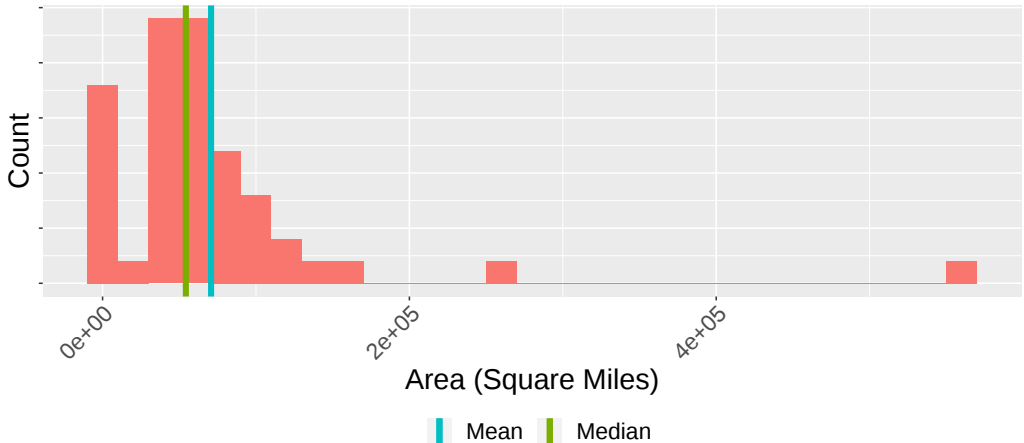
Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review



Statistics for Spread 1

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

- The **standard deviation** and **IQR** are summaries of the spread of a distribution/graph.
- When discussing the center of a graph, you should report the standard deviation or IQR (if available). Otherwise, you should eyeball the spread or report the range.

Statistics for Spread 2

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ Standard Deviation

- The standard deviation represents the typical distance of an observation in a dataset to the mean of the dataset.
- To compute the standard deviation...
 - We find the **deviation** of every observation in the dataset by subtracting the mean from each value.
 - We square each of the deviations and take the average. This quantity is known as the **variance**.
 - We take the square root of the variance.
- Formally,

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}}$$

- The R command for calculating the standard deviation is `sd()`.

Statistics for Spread 3

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ IQR

- The IQR, or InterQuartile Range, is the distance between the first and third quartiles of a dataset.
- To calculate the IQR, we subtract $Q1$ from $Q3$.

■ First Quartile

- The first quartile, or $Q1$, is the median of the bottom half of a dataset.

■ Third Quartile

- The third quartile, or $Q3$, is the median of the upper half of a dataset.

■ Range

- The range is the distance between the minimum and maximum values of a dataset.

Statistics for Spread 4

STAT 140:
Week 2

```
state.x77.df %>%  
  summarize(sd(Area))
```

```
##      sd(Area)
```

```
## 1  85327.3
```

```
## To find Q1 and Q3, we use the quantile function.
```

```
## The quantile function has multiple arguments.
```

```
state.x77.df %>%  
  summarize(quantile(Area, 0.25), quantile(Area, 0.75))
```

```
##      quantile(Area, 0.25) quantile(Area, 0.75)
```

```
## 1              36985.25              81162.5
```

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Five-Number Summary 1

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

- The **five-number summary** consists of the minimum, first quartile, median, third quartile, and maximum.
- In exploratory data analysis, it is common to report some sort of table with the five-number summary for each numerical variable in your dataset.

```
state.x77.df %>%  
  summarize(Minimum = min(Area),  
            First.Quartile = quantile(Area, 0.25),  
            Median = median(Area),  
            Third.Quartile = quantile(Area, 0.75),  
            Maximum = max(Area))
```

##	Minimum	First.Quartile	Median	Third.Quartile	Maximum
## 1	1049	36985.25	54277	81162.5	566432

Five-Number Summary 2

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ Boxplot

- A boxplot is a graphical display for the 5 number summary.
- Boxplots consist of two components:
 - The box, which spans from Q1 to Q3 and is split by the median.
 - The whiskers, which connect the box to the min. and max. values.
 - Outliers are sometimes specially marked.

Boxplots 1

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review



Boxplots 2

STAT 140:
Week 2

Boxplots are particularly useful when comparing one continuous and one categorical variable.

```
state.x77.df %>%  
  group_by(Region) %>%  
  summarize(Minimum = min(Area),  
            First.Quartile = quantile(Area, 0.25),  
            Median = median(Area),  
            Third.Quartile = quantile(Area, 0.75),  
            Maximum = max(Area))
```

```
## # A tibble: 4 x 6  
##   Region      Minimum First.Quartile Median Third.Quartile Maxim  
## * <fct>      <dbl>      <dbl>    <dbl>      <dbl>    <db  
## 1 Northeast    1049        7521     9027      30920    478  
## 2 South        1982       37294    46113     52481    2621
```

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Boxplots 3

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

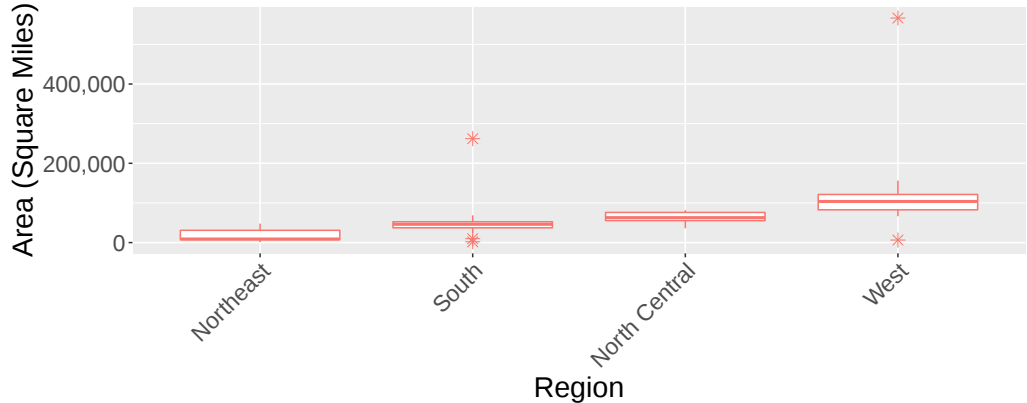
Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review



Unusual Values

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ Outliers

- Values that are unusually distant from the rest of the dataset are known as outliers.

■ Robust Statistics

- Robust statistics are statistics that are minimally influenced by outliers.
- You can also say these statistics are **resistant** to outliers.
- The median and the IQR are robust statistics.
- The mean and standard deviation are heavily influenced by outliers, or not resistant, to outliers.

Shape

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ Right Skewed

- A dataset is said to be right skewed if the distribution has a long right tail, or a handful of values at the right of the graph.
- Other ways to describe data that are right skewed are **skewed to the right** or **positively skewed**.

■ Left Skewed

- A dataset is said to be left skewed if the distribution has a long left tail, or a handful of values at the left of the graph.

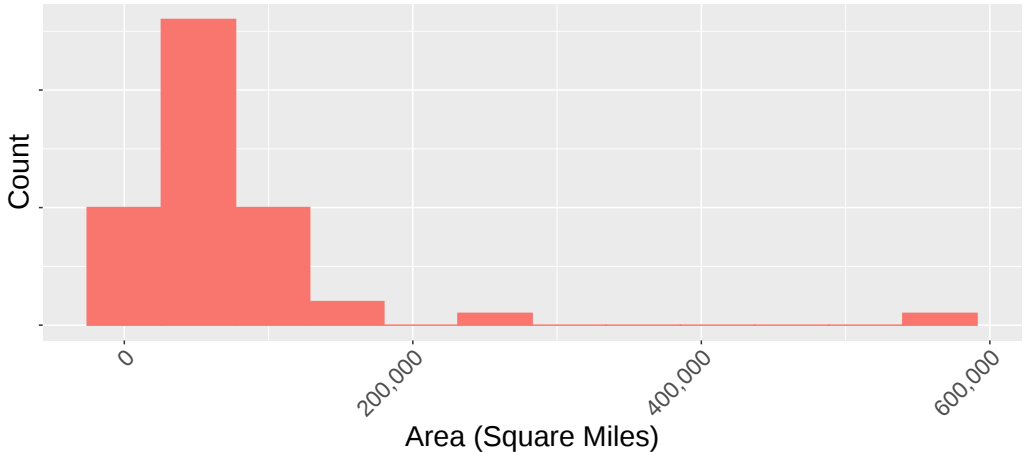
■ Symmetric

- A dataset is said to be symmetric if the tails are of equal size on both sides of the distribution.

Shape: State Areas

STAT 140:
Week 2

The distribution of state areas can be described as right skewed.



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

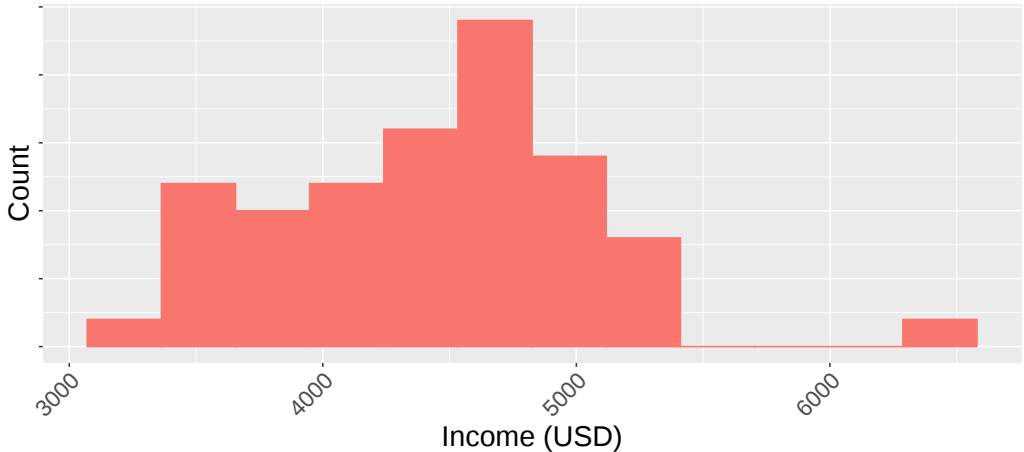
Review

Notation
Review

Shape: State Income

STAT 140:
Week 2

How would you describe the distribution of state per capita incomes?



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

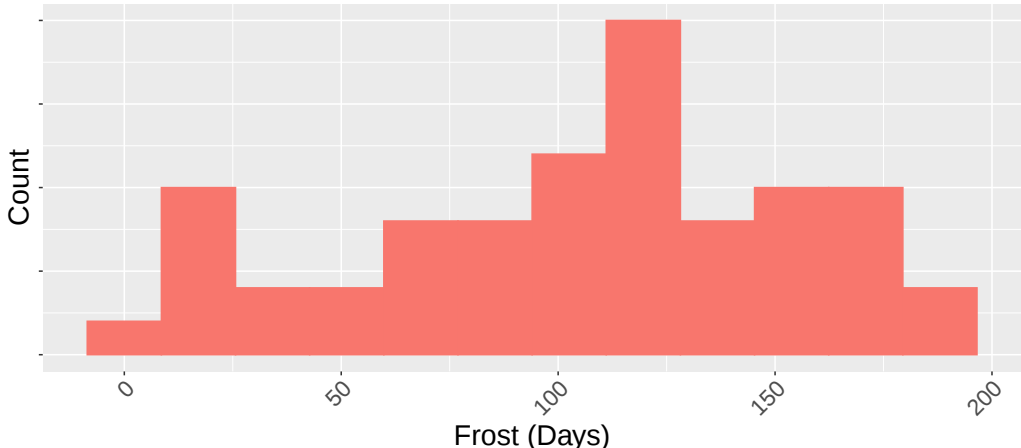
Review

Notation
Review

Shape: State Weather

STAT 140:
Week 2

How would you describe the distribution of the mean number of days with minimum temperatures below freezing per state?



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Shape, continued

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

■ Mode

- A prominent peak in the distribution.
- Can also be thought of as the most common value of the dataset for discrete data.

■ Unimodal

- A dataset with one prominent peak.

■ Bimodal

- A dataset with two prominent peaks.

■ Multimodal

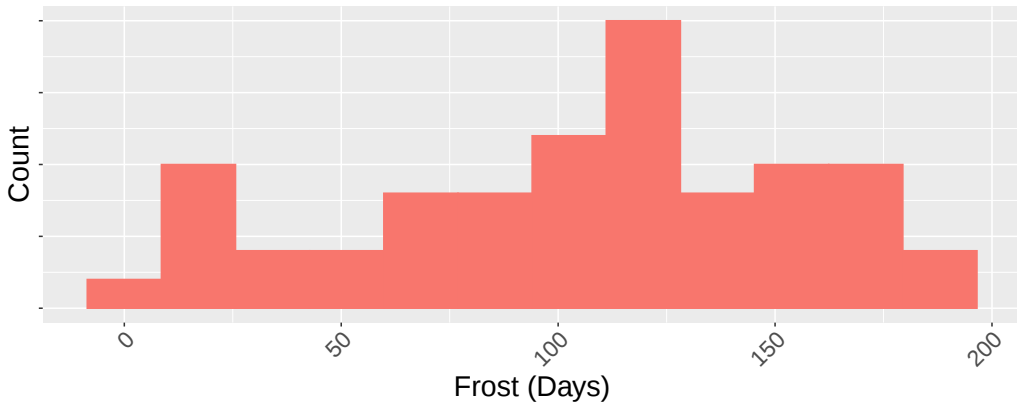
- A dataset with multiple prominent peaks.

- A dataset with no prominent peaks could be described as **uniformly distributed**.

Shape: State Weather

STAT 140:
Week 2

How would you describe the distribution of the mean number of days with minimum temperatures below freezing per state?



Shape: State Income

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

How would you describe the distribution of state per capita incomes?



Displays for Categorical Variables

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

- One last time! A **distribution** is a summary of what values a variable can take and how often those values occur.
- Bar plots (or bar charts) are one way to show the distribution of categorical variables.
- For any graphical display of a distribution,
 - The x-axis is a summary of what values the variable takes.
 - The y-axis dictates how often those values occur.
- In a bar plot, each level of the variable has a bar, and the height of the bar is proportional to the proportion of observations at that level.

Bar Plots

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

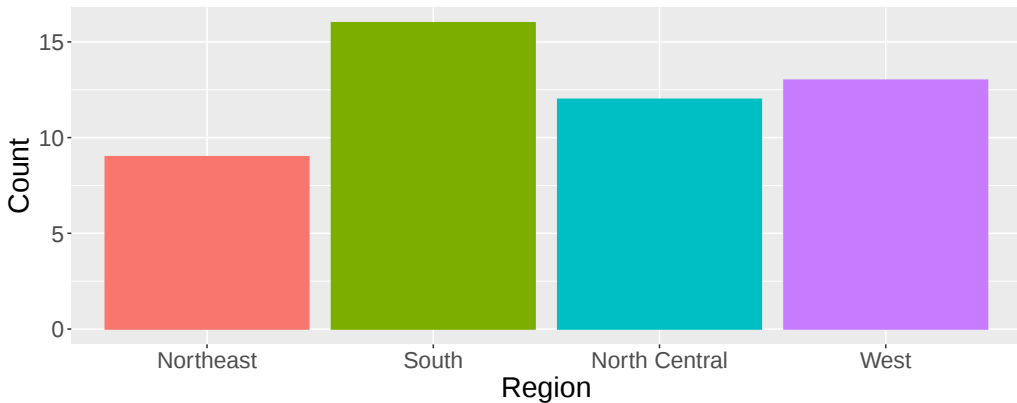
Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review



Statistics for Categorical Variables

STAT 140:
Week 2

- Often, we use frequency tables to report a categorical variable's distribution, showing counts.

```
state.x77.df %>%  
  group_by(Region) %>%  
  summarize(n = n())
```

```
## # A tibble: 4 x 2  
##   Region          n  
## * <fct>         <int>  
## 1 Northeast          9  
## 2 South             16  
## 3 North Central     12  
## 4 West              13
```

Region	Count
Northeast	9
South	16
North Central	12
West	13
Total	50

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Statistics for Categorical Variables

STAT 140:
Week 2

- We might also use proportions to talk about the distribution of a categorical variable.

```
state.x77.df %>%  
  group_by(Region) %>%  
  summarize(n = n()/50)
```

```
## # A tibble: 4 x 2  
##   Region          n  
## * <fct>         <dbl>  
## 1 Northeast     0.18  
## 2 South         0.32  
## 3 North Central 0.24  
## 4 West          0.26
```

Region	Count
Northeast	0.18
South	0.32
North Central	0.24
West	0.26
Total	1.00

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Statistics for Categorical Variables

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

- Often, we represent a proportion with a $\hat{p}$ (p-hat) and a subscript.
- For example, we represent the proportion of states in the Northeast region of the United States with $\hat{p}_{NE}$.

Graph Types

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

- One Categorical Variable: Bar Chart/Bar Plot
- One Numerical Variable: Histogram/Density
- One Categorical, One Numerical Variable: Box PLots
- *Two Numerical Variables: Scatterplot*

Practice Questions 1

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

A survey of students at California Polytechnic State University (San Luis Obispo) collected data to investigate weights of college student backpacks. What type of graph is most appropriate for each of the following variables?

- Academic major
- The ratio of backpack weight to body weight
- Number of credits taken that quarter
- Backpack weights between undergraduate and graduate students
- Year in school

Practice Questions 2

STAT 140:
Week 2

Describe the histogram in terms of its center, spread, shape, and unusual values.



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Practice Questions 3

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

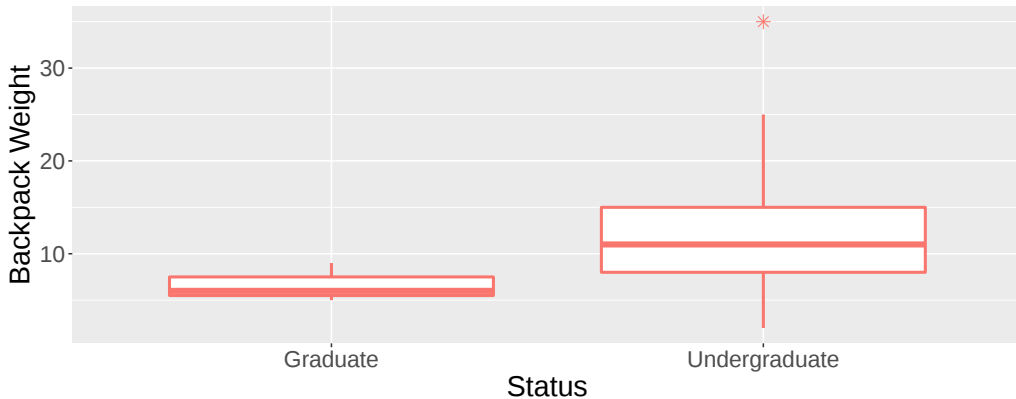
Notation
Review

- For the ratio of backpack weight to body weight, do you predict the mean will be:
 - Greater than the median,
 - Approximately equal to the median, or
 - Less than the median?
- Why?
- Would you recommend using the mean or the median for this variable? Why?

Practice Questions 4

STAT 140:
Week 2

What do these boxplots say about the distributions of backpack weight in undergraduate and graduate students?



Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Statistics Table

STAT 140:
Week 2

Goals for the
Week

Distributions

Center

Spread

Five-Number
Summary

Unusual
Values

Shape

Categorical
Variables

Review

Notation
Review

Characteristic	Statistic
Mean	$\bar{x}$
Standard Deviation	s
Proportion	$\hat{p}$