Data Summaries

Nate Wells

Math 141, 2/5/21

Outline

In this lecture, we will. . .

Outline

In this lecture, we will...

- Use `ggplot2` to create Barplots
- Investigate some options for further customizing graphs
- Discuss measurements of center and spread for quantitative data

Section 1

Common Graphs using ggplot2

The Five Named Graphs

- We focus on just 5 graphs fundamental to statistics (although other types exist)

The Five Named Graphs

- We focus on just 5 graphs fundamental to statistics (although other types exist)
    1. Scatterplots
    2. Linegraphs
    3. Histograms
    4. Boxplots
    5. **Barplots**

The Five Named Graphs

- We focus on just 5 graphs fundamental to statistics (although other types exist)
  1. Scatterplots
  2. Linegraphs
  3. Histograms
  4. Boxplots
  5. **Barplots**
- We'll use a common data set to investigate each graph: the Portland Biketown data

```
biketown <-
  read_csv("biketown.csv")
```

Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.

Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.
- We use Bar Charts to visualize the distribution of *categorical* variables, whose values are broken down into distinct levels.

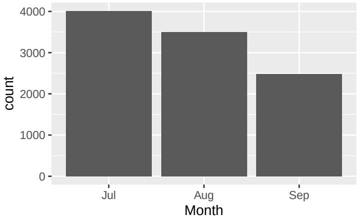
Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.
- We use Bar Charts to visualize the distribution of *categorical* variables, whose values are broken down into distinct levels.
- Investigate the distribution of bike use by `month`

Bar Charts

- Both Boxplots and Histograms show the distribution of *quantitative* variables.
- We use Bar Charts to visualize the distribution of *categorical* variables, whose values are broken down into distinct levels.
- Investigate the distribution of bike use by `month`

```
ggplot(data = biketown, mapping = aes(x = Month)) +
  geom_bar()
```



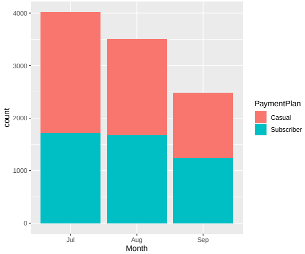
Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

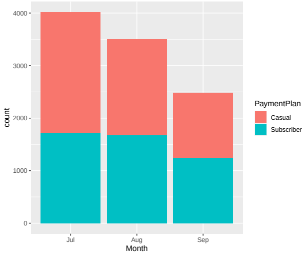
```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan)) +
  geom_bar()
```



Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan)) +
  geom_bar()
```

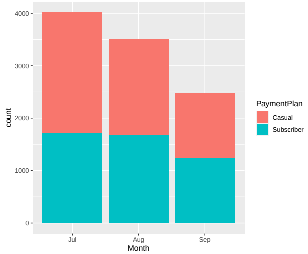


- Each bar divided into count by `fill` variable.

Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan)) +
  geom_bar()
```

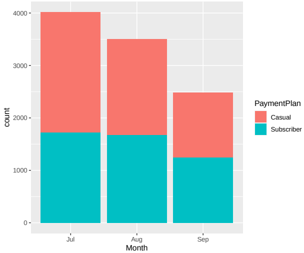


- Each bar divided into count by `fill` variable.
- Hard to make direct comparisons

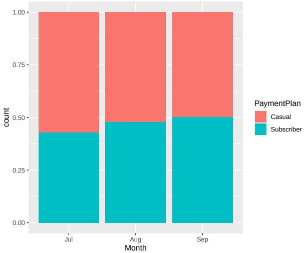
Segmented / Stacked Bar Charts

- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar()
```




```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar(position = "fill")
```

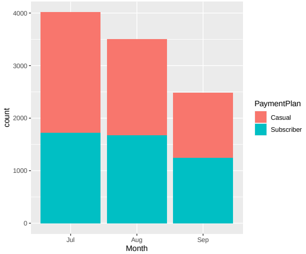


- Each bar divided into count by `fill` variable.
- Hard to make direct comparisons

Segmented / Stacked Bar Charts

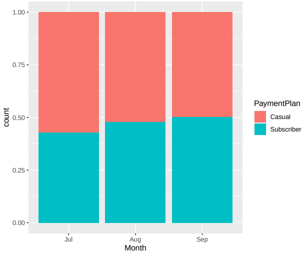
- Bar charts used to visualize the *joint distribution* of 2 categorical variables.

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar()
```



- Each bar divided into count by `fill` variable.
- Hard to make direct comparisons

```
ggplot(data = biketown,
       mapping = aes(x = Month,
                     fill = PaymentPlan))
  geom_bar(position = "fill")
```



- Each bar divided into proportion by `fill` variable.

Section 2

Extending ggplot2

Adding additional variables?

Scatterplots, side-by-side boxplots, and segmented barcharts all show relationships between 2 variables.

But what can we do to simultaneously explore 3 variables?

Adding additional variables?

Scatterplots, side-by-side boxplots, and segmented barcharts all show relationships between 2 variables.

But what can we do to simultaneously explore 3 variables?

1. 3D Scatterplots; possible, but challenging to code and interpret (still limited to 2d display)

Adding additional variables?

Scatterplots, side-by-side boxplots, and segmented barcharts all show relationships between 2 variables.

But what can we do to simultaneously explore 3 variables?

1. 3D Scatterplots; possible, but challenging to code and interpret (still limited to 2d display)
2. Map variables to additional aesthetics (beyond just `x` and `y`)

Adding additional variables?

Scatterplots, side-by-side boxplots, and segmented barcharts all show relationships between 2 variables.

But what can we do to simultaneously explore 3 variables?

1. 3D Scatterplots; possible, but challenging to code and interpret (still limited to 2d display)
2. Map variables to additional aesthetics (beyond just `x` and `y`)
3. Show several 2D plots side-by-side.

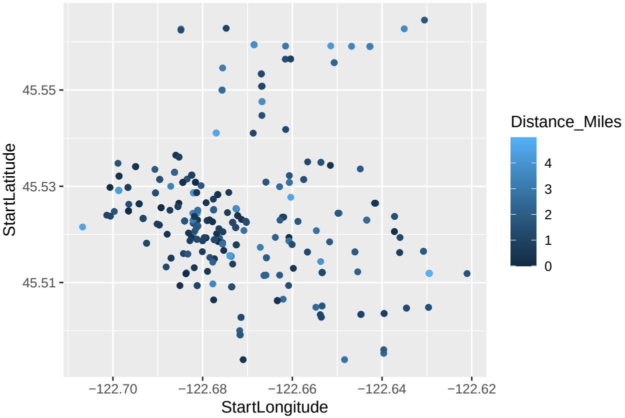
Multiple Variables on 2d Plots

Does ride distance depend on start location?

Multiple Variables on 2d Plots

Does ride distance depend on start location?

```
ggplot(data = biketown_sample,
       mapping = aes(x=StartLongitude, y=StartLatitude, color=Distance_Miles))+
  geom_point()
```



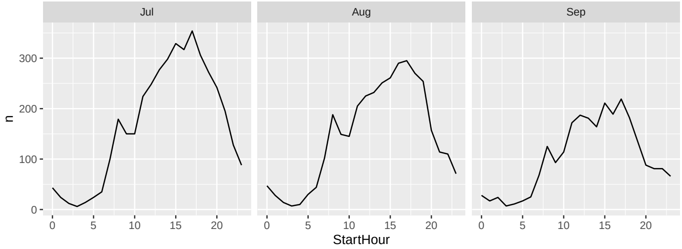
Facets

- Faceting is used to split one graphic into many smaller ones, based on the values of a categorical variable.

Facets

- Faceting is used to split one graphic into many smaller ones, based on the values of a categorical variable.

```
ggplot(data = biketown2, mapping = aes(x = StartHour, y = n)) +
  geom_line() +
  facet_wrap(~Month, ncol = 3)
```

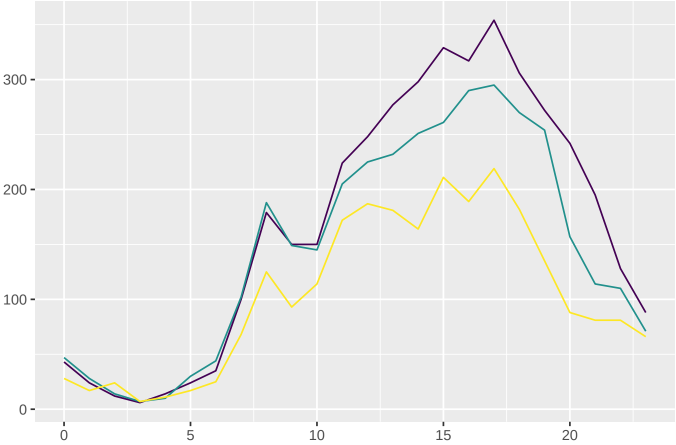


Adding Context

- Adding titles and axes labels to graphs greatly improves clarity.

Adding Context

- Adding titles and axes labels to graphs greatly improves clarity.



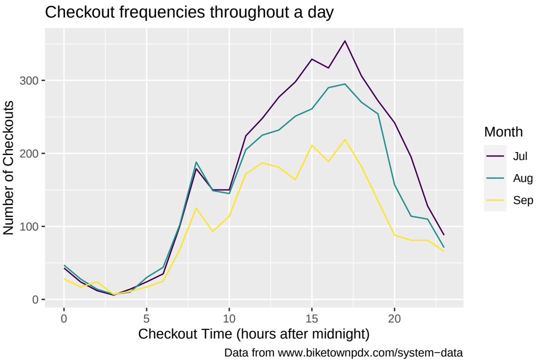
Adding Context

- Adding titles, captions, and axis labels greatly improves clarity.

Adding Context

- Adding titles, captions, and axis labels greatly improves clarity.

```
ggplot(data = biketown2, mapping = aes(x = StartHour, y = n, color = Month)) +
  geom_line( ) +
  labs(x = "Checkout Time (hours after midnight)", y = "Number of Checkouts",
       title  = "Checkout frequencies throughout a day",
       caption = "Data from www.biketownpdx.com/system-data")
```



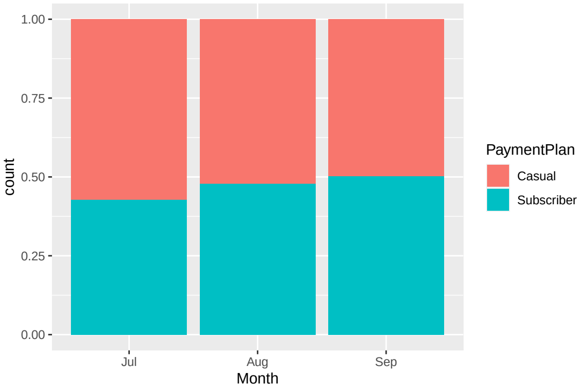
Change Graphic Colors

By default, R uses `Teal` and `Salmon` colors when plotting cat. variables with 2 levels

Change Graphic Colors

By default, R uses `Teal` and `Salmon` colors when plotting cat. variables with 2 levels

```
ggplot(data = biketown, mapping = aes(x = Month, fill = PaymentPlan)) +
  geom_bar(position = "fill")
```



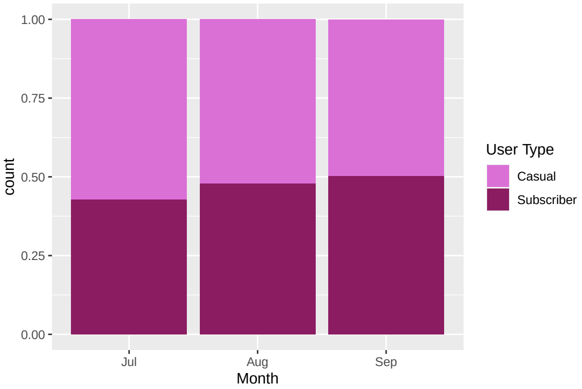
Change Graphic Colors

But it's possible to alter this

Change Graphic Colors

But it's possible to alter this

```
ggplot(data = biketown, mapping = aes(x = Month, fill = PaymentPlan)) +
  geom_bar(position = "fill")+
  scale_fill_manual(name = "User Type",
                    values = c("orchid", "maroon4"))
```



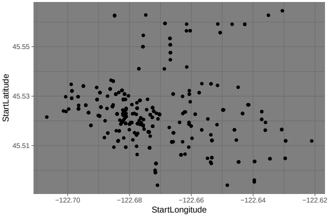
Change Theme

We can also control the styling of other plot elements via `theme`

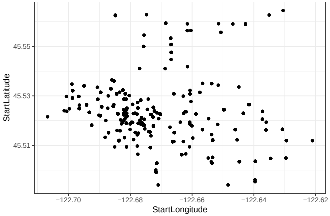
Change Theme

We can also control the styling of other plot elements via `theme`

```
ggplot(data = biketown_sample,
       mapping = aes(x = StartLongitude,
                     y = StartLatitude))+
  geom_point()+
  theme_dark()
```




```
ggplot(data = biketown_sample,
       mapping = aes(x = StartLongitude,
                     y = StartLatitude))+
  geom_point()+
  theme_bw()
```



Re-order bars

For categorical variables, values are often displayed in alphabetical order. We can change that by changing the way the data is stored:

Re-order bars

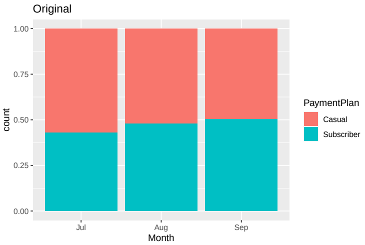
For categorical variables, values are often displayed in alphabetical order. We can change that by changing the way the data is stored:

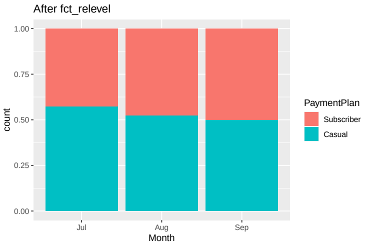
```
biketown <- mutate(biketown, PaymentPlan =
                     fct_relevel(PaymentPlan,
                                 "Subscriber", "Casual"))
```

Re-order bars

For categorical variables, values are often displayed in alphabetical order. We can change that by changing the way the data is stored:

```
biketown <- mutate(biketown, PaymentPlan =
                    fct_relevel(PaymentPlan,
                                "Subscriber", "Casual"))
```





Section 3

Data Summaries

Exam Statistics

Suppose you are an instructor trying to gauge class performance on an exam. You have exam scores for 200 intro stat students.

Exam Statistics

Suppose you are an instructor trying to gauge class performance on an exam. You have exam scores for 200 intro stat students.

What summarizing information would it be helpful to know in order to assess how well the class did?

Exam Statistics

Suppose you are an instructor trying to gauge class performance on an exam. You have exam scores for 200 intro stat students.

What summarizing information would it be helpful to know in order to assess how well the class did?

1. What was the typical value (maybe average or median)?
2. How much variation was there in scores?
3. What was the shape of the data?
4. Were there any outliers?

The Mean

The **mean** or average of a data set is one measure of *center*, obtained by adding all observed values and dividing by their number:

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

where n is the number of observations and x_i is the value of the ith observation.

The Mean

The **mean** or average of a data set is one measure of *center*, obtained by adding all observed values and dividing by their number:

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

where n is the number of observations and x_i is the value of the ith observation.

```
mean(biketown_short$Distance_Miles)
```

```
## [1] 1.677599
```

The Mean

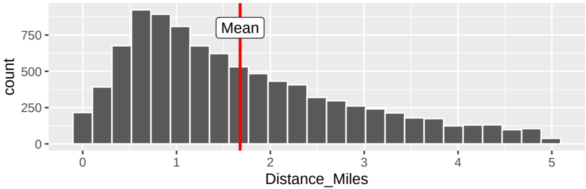
The **mean** or average of a data set is one measure of *center*, obtained by adding all observed values and dividing by their number:

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

where n is the number of observations and x_i is the value of the ith observation.

```
mean(biketown_short$Distance_Miles)
```

```
## [1] 1.677599
```



- If the histogram were made of solid material, the mean would be the point along the horizontal axis where the solid is perfectly balanced.

The Median

The **median** is another measure of *center* and separates data into two equally sized sets.

The Median

The **median** is another measure of *center* and separates data into two equally sized sets.

Suppose the n values are ordered from least to greatest. The median is the value in the middle of the list.

- If n is even, then there are two middle values, and the median is their average.

The Median

The **median** is another measure of *center* and separates data into two equally sized sets.

Suppose the n values are ordered from least to greatest. The median is the value in the middle of the list.

- If n is even, then there are two middle values, and the median is their average.

```
median(biketown_short$Distance_Miles)
```

```
## [1] 1.39
```

The Median

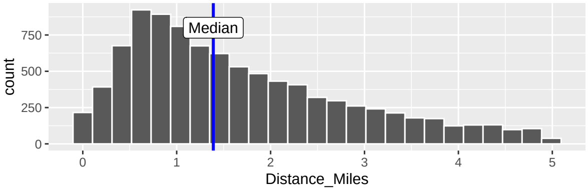
The **median** is another measure of *center* and separates data into two equally sized sets.

Suppose the n values are ordered from least to greatest. The median is the value in the middle of the list.

- If n is even, then there are two middle values, and the median is their average.

```
median(biketown_short$Distance_Miles)
```

```
## [1] 1.39
```



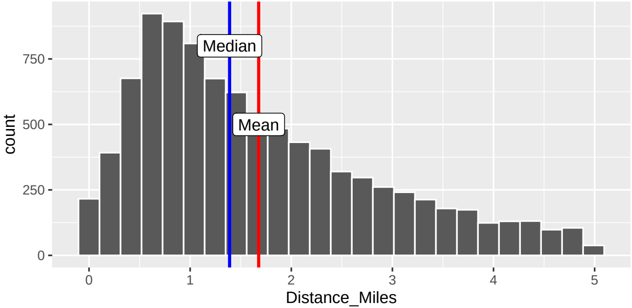
- The median corresponds to the line that divides a histogram into two equal area pieces.

Mean, Median, and Skew

Both mean and median represent *typical* values for a data set.

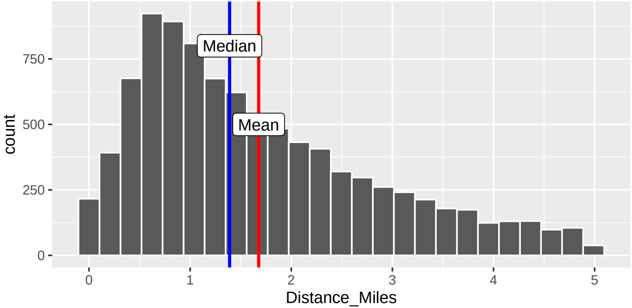
Mean, Median, and Skew

Both mean and median represent *typical* values for a data set.



Mean, Median, and Skew

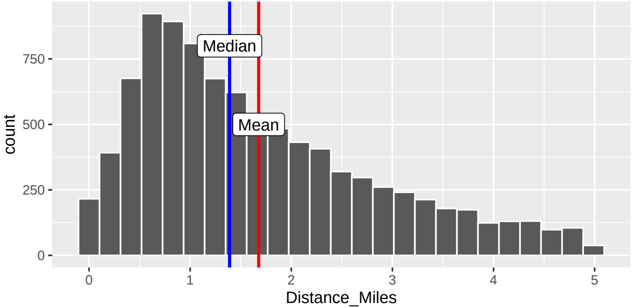
Both mean and median represent *typical* values for a data set.



- In non-symmetric distributions, the mean will further along the direction of skew than the median.

Mean, Median, and Skew

Both mean and median represent *typical* values for a data set.



- In non-symmetric distributions, the mean will further along the direction of skew than the median.
  - Why?

Robustness

Consider two data sets, one with a large outlier and one without:

```
my_data <- c(1, 2, 5, 7, 8, 10)
my_data_with_oulier <- c(1, 2, 5, 7, 8, 100)
```

Robustness

Consider two data sets, one with a large outlier and one without:

```
my_data <- c(1, 2, 5, 7, 8, 10)
my_data_with_oulier <- c(1, 2, 5, 7, 8, 100)
```

The mean value of a dataset is very sensitive to outliers.

Robustness

Consider two data sets, one with a large outlier and one without:

```
my_data <- c(1, 2, 5, 7, 8, 10)
my_data_with_oulier <- c(1, 2, 5, 7, 8, 100)
```

The mean value of a dataset is very sensitive to outliers.

```
mean(my_data)
```

```
## [1] 5.5
```

```
mean(my_data_with_oulier)
```

```
## [1] 20.5
```

Robustness

Consider two data sets, one with a large outlier and one without:

```
my_data <- c(1, 2, 5, 7, 8, 10)
my_data_with_oulier <- c(1, 2, 5, 7, 8, 100)
```

The mean value of a dataset is very sensitive to outliers.

```
mean(my_data)
```

```
## [1] 5.5
```

```
mean(my_data_with_oulier)
```

```
## [1] 20.5
```

The median, however, is not.

```
median(my_data)
```

```
## [1] 6
```

```
median(my_data_with_oulier)
```

```
## [1] 6
```

Measures of Variability

We'd like to assess how variable the data set is.

- Are values usually close to the mean, or are they spread out?

Measures of Variability

We'd like to assess how variable the data set is.

- Are values usually close to the mean, or are they spread out?

How can we find the typical amount an observation differs from the mean observation?

Measures of Variability

We'd like to assess how variable the data set is.

- Are values usually close to the mean, or are they spread out?

How can we find the typical amount an observation differs from the mean observation?

Guess 1: Compute the average difference

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

Distance_Miles	Mean	Deviations
1.57	1.2	0.37
2.09	1.2	0.89
0.38	1.2	-0.82
0.86	1.2	-0.34
1.10	1.2	-0.10

Measures of Variability

We'd like to assess how variable the data set is.

- Are values usually close to the mean, or are they spread out?

How can we find the typical amount an observation differs from the mean observation?

Guess 1: Compute the average difference

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

Distance_Miles	Mean	Deviations
1.57	1.2	0.37
2.09	1.2	0.89
0.38	1.2	-0.82
0.86	1.2	-0.34
1.10	1.2	-0.10

- What's the problem?

Measures of Variability

We'd like to assess how variable the data set is.

- Are values usually close to the mean, or are they spread out?

How can we find the typical amount an observation differs from the mean observation?

Guess 1: Compute the average difference

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

Distance_Miles	Mean	Deviations
1.57	1.2	0.37
2.09	1.2	0.89
0.38	1.2	-0.82
0.86	1.2	-0.34
1.10	1.2	-0.10

- What's the problem?

Avg_Deviations
0

Measures of Variability

The fix?

Measures of Variability

The fix?

Guess 2: Compute the average *squared* difference

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

Distance_Miles	Mean	Sq_Deviation
1.57	1.2	0.1369
2.09	1.2	0.7921
0.38	1.2	0.6724
0.86	1.2	0.1156
1.10	1.2	0.0100

Measures of Variability

The fix?

Guess 2: Compute the average *squared* difference

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

Distance_Miles	Mean	Sq_Deviation
1.57	1.2	0.1369
2.09	1.2	0.7921
0.38	1.2	0.6724
0.86	1.2	0.1156
1.10	1.2	0.0100

- This is called the **Population Variance**

Measures of Variability

The fix?

Guess 2: Compute the average *squared* difference

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

Distance_Miles	Mean	Sq_Deviation
1.57	1.2	0.1369
2.09	1.2	0.7921
0.38	1.2	0.6724
0.86	1.2	0.1156
1.10	1.2	0.0100

- This is called the **Population Variance**

Pop_Variance
0.3454

Standard Deviation

The population variance does measure spread of data.

$$\text{Population Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard Deviation

The population variance does measure spread of data.

$$\text{Population Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

But it does have two small problems:

Standard Deviation

The population variance does measure spread of data.

$$\text{Population Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

But it does have two small problems:

1. When sampling, it tends to *underestimate* the variability in the population. So we increase it by dividing by something smaller:

Standard Deviation

The population variance does measure spread of data.

$$\text{Population Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

But it does have two small problems:

1. When sampling, it tends to *underestimate* the variability in the population. So we increase it by dividing by something smaller:

$$\text{Sample Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard Deviation

The population variance does measure spread of data.

$$\text{Population Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

But it does have two small problems:

1. When sampling, it tends to *underestimate* the variability in the population. So we increase it by dividing by something smaller:

$$\text{Sample Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2. Because observations are squared, it is no longer measured in same *units* as original data (i.e. if data is in miles, then variance is in sq. miles). So we take square roots:

Standard Deviation

The population variance does measure spread of data.

$$\text{Population Variance} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

But it does have two small problems:

1. When sampling, it tends to *underestimate* the variability in the population. So we increase it by dividing by something smaller:

$$\text{Sample Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

2. Because observations are squared, it is no longer measured in same *units* as original data (i.e. if data is in miles, then variance is in sq. miles). So we take square roots:

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

Visualizing Standard Deviation

The standard deviation measures the typical size of deviations of observations from the mean.

Visualizing Standard Deviation

The standard deviation measures the typical size of deviations of observations from the mean.

For most data sets, almost all observations are within a distance of 2 standard deviations of the mean:

Visualizing Standard Deviation

The standard deviation measures the typical size of deviations of observations from the mean.

For most data sets, almost all observations are within a distance of 2 standard deviations of the mean:

```
sd(biketown_short$Distance_Miles)
```

```
## [1] 1.172257
```

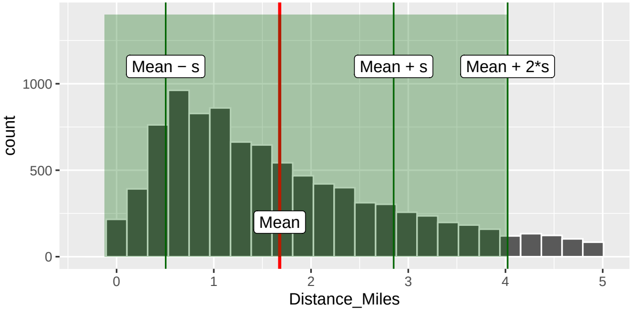
Visualizing Standard Deviation

The standard deviation measures the typical size of deviations of observations from the mean.

For most data sets, almost all observations are within a distance of 2 standard deviations of the mean:

```
sd(biketown_short$Distance_Miles)
```

```
## [1] 1.172257
```



Quartiles and IQR

Where the median divides data into equal halves, *quartiles* divide data into equal quarters

Quartiles and IQR

Where the median divides data into equal halves, *quartiles* divide data into equal quarters

- 25% of all observations are less than the *first quartile* $Q1$
- 25% of all observations are greater than the *third quartile* $Q3$

Quartiles and IQR

Where the median divides data into equal halves, *quartiles* divide data into equal quarters

- 25% of all observations are less than the *first quartile* $Q1$
- 25% of all observations are greater than the *third quartile* $Q3$

```
quantile(biketown_short$Distance_Miles, c(.25, .75))
```

```
##  25%  75%
## 0.75 2.38
```

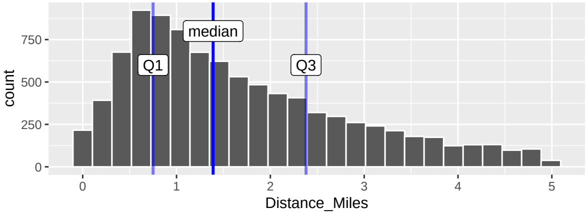
Quartiles and IQR

Where the median divides data into equal halves, *quartiles* divide data into equal quarters

- 25% of all observations are less than the *first quartile* $Q1$
- 25% of all observations are greater than the *third quartile* $Q3$

```
quantile(biketown_short$Distance_Miles, c(.25, .75))
```

```
##   25%  75%
## 0.75 2.38
```



- The *IQR* is the distance between the 1st and 3rd quartile: $\mathrm{IQR} = Q3 - Q1$

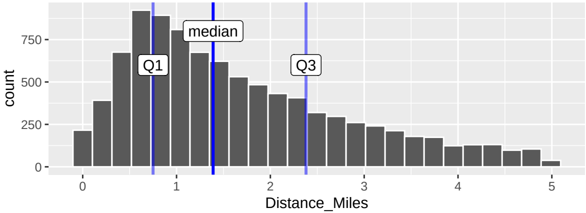
Quartiles and IQR

Where the median divides data into equal halves, *quartiles* divide data into equal quarters

- 25% of all observations are less than the *first quartile* $Q1$
- 25% of all observations are greater than the *third quartile* $Q3$

```
quantile(biketown_short$Distance_Miles, c(.25, .75))
```

```
##  25%  75%
## 0.75 2.38
```



- The *IQR* is the distance between the 1st and 3rd quartile: $\mathrm{IQR} = Q3 - Q1$
- Comparing $\mathrm{Median} - Q1$ and $Q3 - \mathrm{Median}$ can show shape of distribution.