Statistics

J. Alexander Branham

Fall 2016

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Statistics

Statistics

Statistics allow us to learn from data.

Data

Observations of a *variable*:

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Univariate statistics

Univariate statistics

A **statistic** summarizes data. You're already familiar with some common statistics, like averages.

We oftentimes want to find the "center" of the data — this describes typical values

$$x = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix}$$

The *mean* ($\bar{x}$) is calculated by summing the data, then dividing by the number of observations:

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

$$x = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix}$$

The *mean* ($\bar{x}$) is calculated by summing the data, then dividing by the number of observations:

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

The *median* is found by ordering the observations from highest to lowest and finding the one in the middle:

$$x = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{bmatrix}$$

The *mean* ($\bar{x}$) is calculated by summing the data, then dividing by the number of observations:

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

The *median* is found by ordering the observations from highest to lowest and finding the one in the middle: The *mode* is the most common number

Central tendency

- The mean balances the value on either side

Central tendency

- The mean balances the value on either side
- The median balances the number of observations on either side

Central tendency

- The mean balances the value on either side
- The median balances the number of observations on either side
- Which is a better measure?

Mean vs median

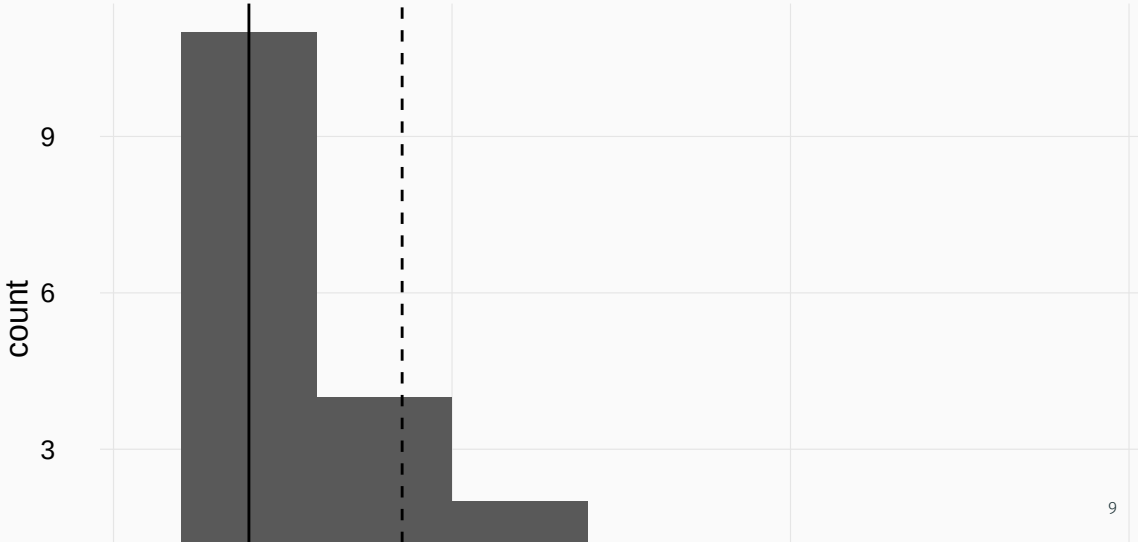
```
x <- c(1, 1, 2, 1, 1, 3, 2, 4, 2, 1,
       1, 1, 5, 7, 9, 4, 5, 6, 25)
mean(x)
```

```
## [1] 4.263158
```

```
median(x)
```

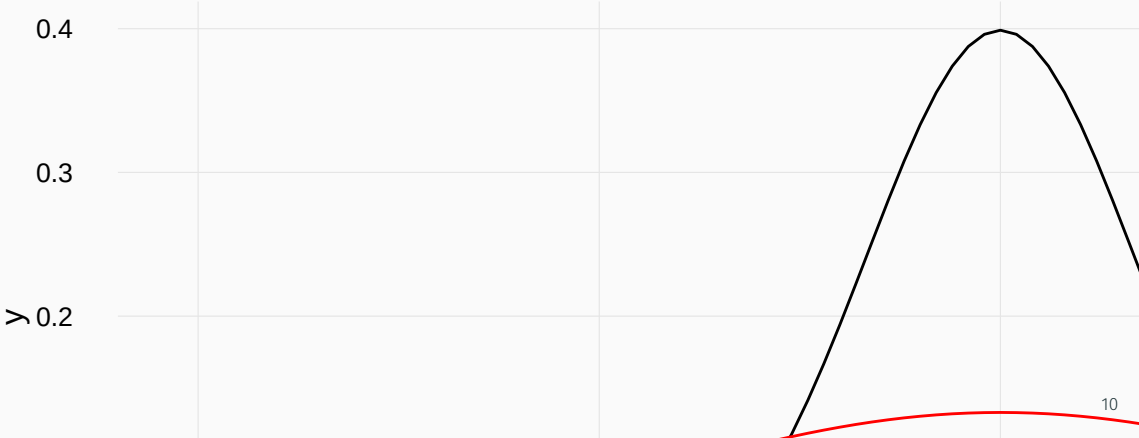
```
## [1] 2
```

MEAN VS MEDIAN



Variance

Finding central tendency is good, but we might go a step further. Consider these two distributions:



You try!

Here are the average high's from a previous year's math camp:

Day	M	Tu	W	Th	F	M	Tu
High	95	103	100	97	39	108	112

Find the mean, median, and mode.

You try!

Here are the average high's from a previous year's math camp:

Day	M	Tu	W	Th	F	M	Tu
High	95	103	100	97	39	108	112

Find the mean, median, and mode.

What's weird with this data?

You try (answers)

```
x <- c(95, 103, 100, 97, 39, 108, 112)
mean(x)
```

```
## [1] 93.42857
```

```
median(x)
```

```
## [1] 100
```

Variance

Variance measures how spread out a distribution is. One way to calculate it is like so:

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

That measures the sum of the squared average deviation from the mean

"Squared average deviation from the mean" is a bit weird, though, so oftentimes we use standard deviations instead, which is just the squared root of the variance:

$$s_x = \sqrt{s_x^2}$$

YOU TRY!

I flipped a coin 4 times, one of which was a heads. What's the mean of the data? What's the variance? The standard deviation?

You try (answers)

```
x <- c(1, 0, 0, 0)
mean(x)
```

```
## [1] 0.25
```

```
var(x)
```

```
## [1] 0.25
```

```
sd(x)
```

```
## [1] 0.5
```

Bivariate statistics

Bivariate statistics

Thus far, we've focused on statistics that summarize just one variable. But we're oftentimes interested in relationships between different variables. This can be hard to see with the raw data, though:

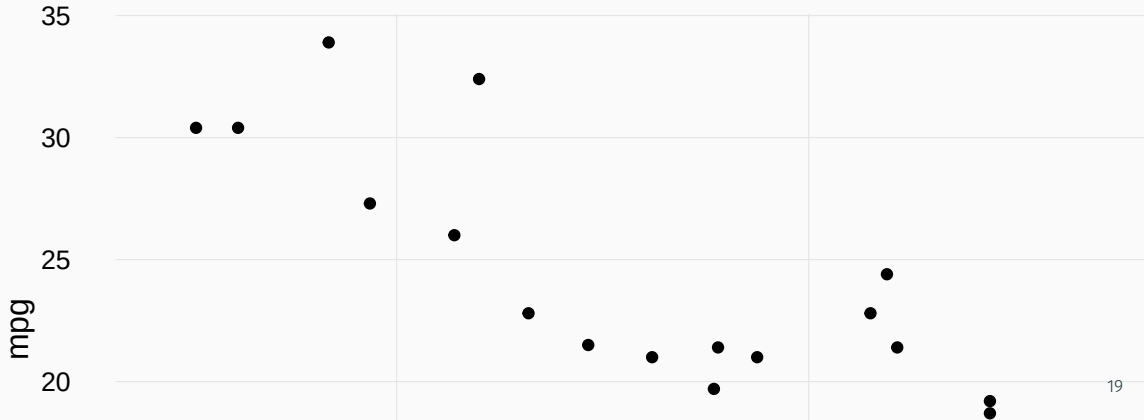
What's the relationship between `mpg` and `wt`?

```
head(mtcars, 9)
```

```
##                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
```

Always plot your data!!!

```
ggplot(mtcars) +
  geom_point(aes(wt, mpg))
```



Covariance

Covariance measures the direction of a relationship:

```
with(mtcars, cov(wt, mpg))
```

```
## [1] -5.116685
```

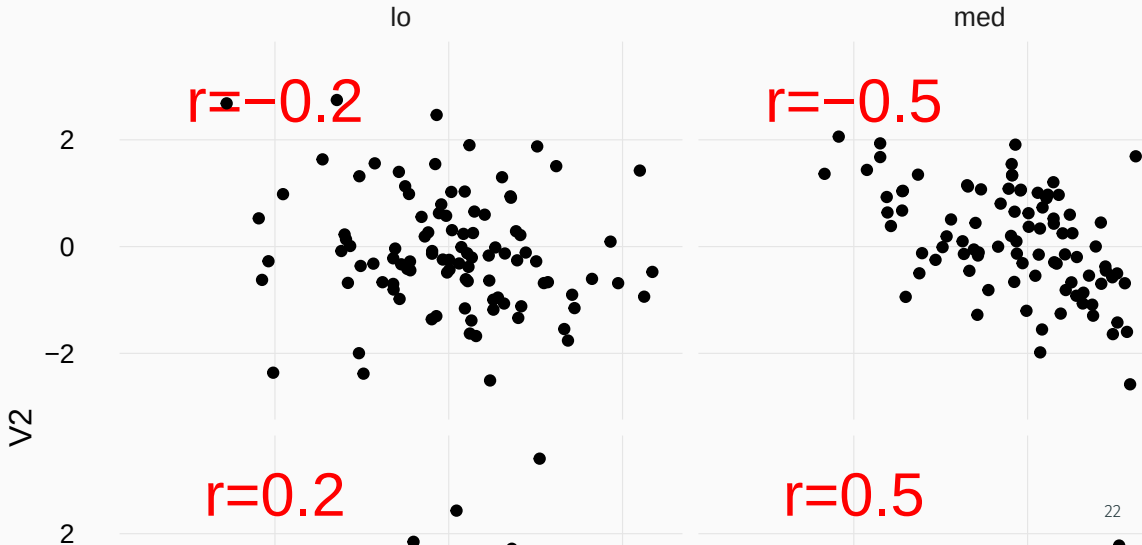
Correlation

Correlation (pearson's *r*) captures the direction and strength of a linear relationship between two variables.

Ranges from —1 to 1

```
with(mtcars, cor(wt, mpg))
```

```
## [1] -0.8676594
```


lo
med
r=−0.2
r=−0.5
r=0.2
r=0.5
V2
2
0
−2
2

Always plot your data