

# Correlation

Association between 2 Quantitative Variables

Grinnell College

## Review + Outline

We've talked about confidence intervals and hypothesis testing the last few weeks, which are incredibly useful tools. But there are other aspects to data analysis and other types of questions we may want to answer

Study design → lets us talk about those patterns and see how they apply

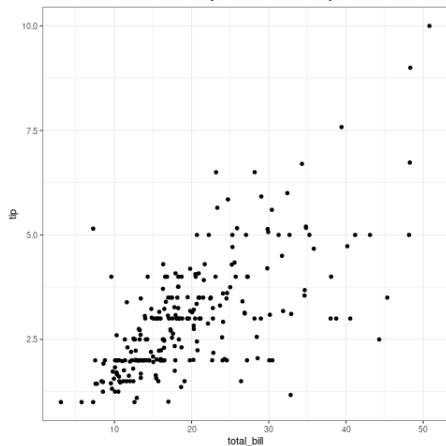
- ▶ representative sample → generalize from sample to populations
- ▶ random assignment → causal relationship

What we will be covering today:

- ▶ Revisiting scatterplots
- ▶ Measure the strength of the relationship between quantitative variables
- ▶ Working towards being able to make predictions using scatterplot

# Review – Scatterplots

When we want to plot two quantitative variables → scatterplots



Scatterplots let us see if there are *associations* between quantitative variables

# Review – Scatterplots

Describing associations in scatterplots:

- ▶ **Form:** pattern? (linear / non-linear / curved / none)
- ▶ **Strength:** weak / moderate / strong
- ▶ **Direction:** positive / negative
- ▶ **Outliers**

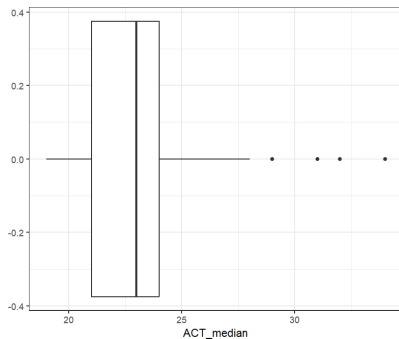
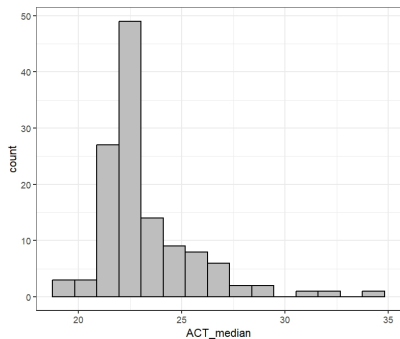
**Note:** Self-reminder to go back and look at Scatterplot examples in the Quantitative Variables 2 lecture

## Extra on Outliers

When we look for outliers in histograms and boxplots, it is fairly simple

- ▶ boxplot → points outside the 'whiskers'
- ▶ histogram → gaps inbetween the bins

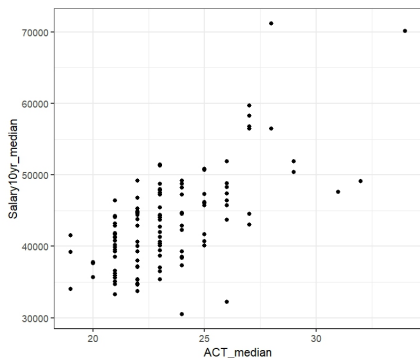
But! These do not always agree. We need to mention which we used to classify (i.e.: "...according to the histogram")



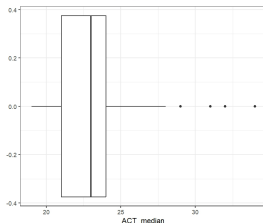
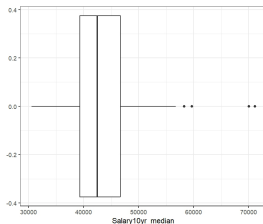
# Extra on Outliers

## Outlier in a scatterplot

- ▶ very small or large values for one of the variables (or both!)
- ▶ does not follow the overall pattern



```
> which(colleges$ACT_median == 35)
[1] 456
> colleges$Name[456]
[1] "Massachusetts Institute of Technology"
```



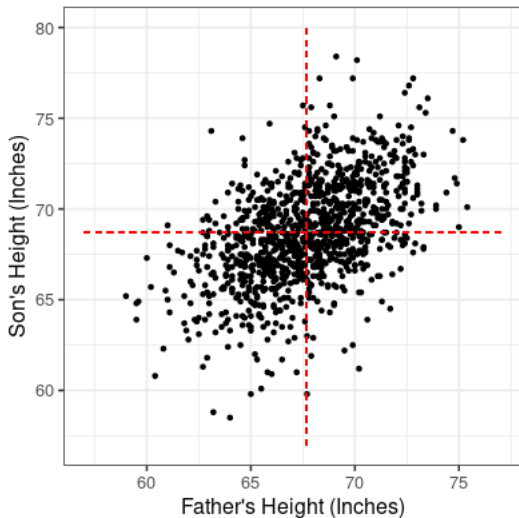
# Pearson's Height Data

In the 1880's the Western scientific community was enthralled with the idea of quantifying heritable traits

Karl Pearson collected data on the heights of 1,087 father's and their fully grown first born sons

Father	Son
65.0	59.8
63.3	63.2
65.0	63.3
65.8	62.8
61.1	64.3
63.0	64.2
⋮	⋮

# Height Data



# Pearson's Correlation Coefficient

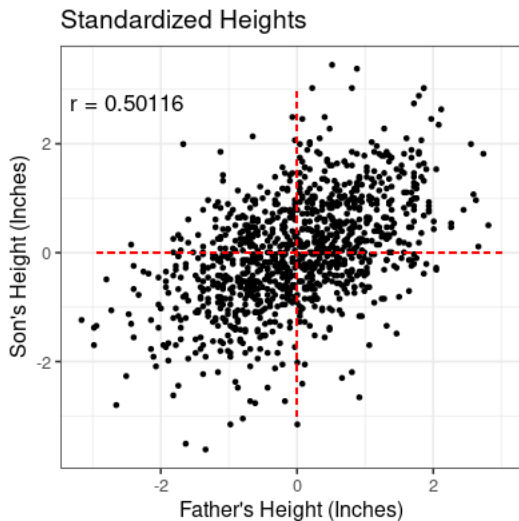
Heights clearly associated, but how to quantify?

Building upon the work from French scientist Francis Galton, Pearson developed the **Pearson's correlation coefficient (r)**:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

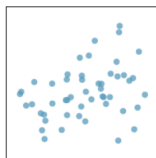
If above-average values of  $X$  are common among cases with above-average values of  $Y$  (or vice-versa), we should expect  $r$  to be positive

# Height Data – Standardized

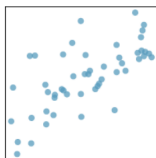


# Correlation Examples

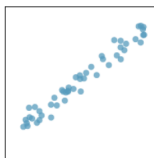
Pearson's correlation coefficient tells us the strength of *linear* association between two quantitative variables (and direction!)



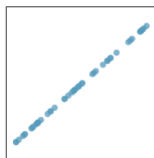
$R = 0.33$



$R = 0.69$



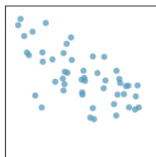
$R = 0.98$



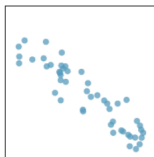
$R = 1.00$



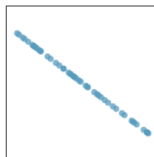
$R = 0.08$



$R = -0.64$



$R = -0.92$



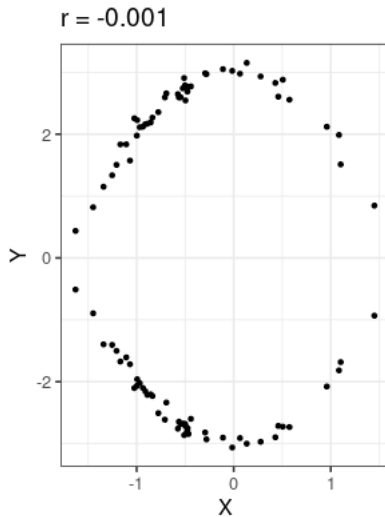
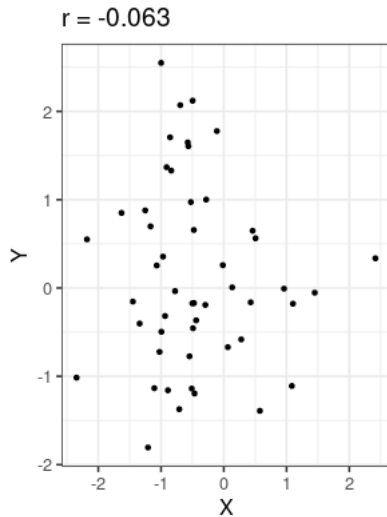
$R = -1.00$

# What is considered “strong”?

<b>Correlation Coefficient</b>		<b>Dancey &amp; Reidy (Psychology)</b>	<b>Quinnipiac University (Politics)</b>	<b>Chan YH (Medicine)</b>
+1	-1	Perfect	Perfect	Perfect
+0.9	-0.9	Strong	Very Strong	Very Strong
+0.8	-0.8	Strong	Very Strong	Very Strong
+0.7	-0.7	Strong	Very Strong	Moderate
+0.6	-0.6	Moderate	Strong	Moderate
+0.5	-0.5	Moderate	Strong	Fair
+0.4	-0.4	Moderate	Strong	Fair
+0.3	-0.3	Weak	Moderate	Fair
+0.2	-0.2	Weak	Weak	Poor
+0.1	-0.1	Weak	Negligible	Poor
0	0	Zero	None	None

Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6107969/>

# Correlation Examples



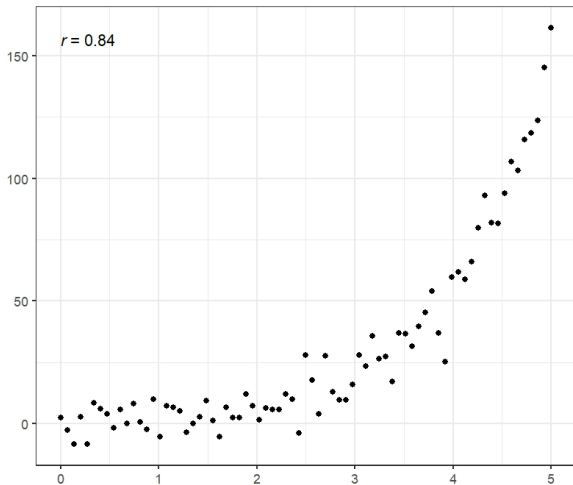
# Correlation Properties

## Properties:

- ▶  $r$  has no units
- ▶  $r$  measures the strength of a linear relationship
- ▶  $r$  is between  $-1$  and  $1$
- ▶ The closer  $r$  is to  $0$   $\rightarrow$  weaker relationship
- ▶ The further  $r$  is from  $0$   $\rightarrow$  stronger relationship
- ▶  $r=0$   $\rightarrow$  no linear relationship
- ▶ changing scale of either variable doesn't affect  $r$  value

# Pitfalls

If we get a value for  $r$  close to  $+1$  or  $-1$ , it does **not** mean the relationship actually is linear (double-check the scatterplot!)



## Non-linear Association

In addition to Pearson, we have **Spearman's rank correlation** (denoted  $\rho$ ) where the values of  $X$  and  $Y$  are replaced with their rank order from smallest to largest before correlating:

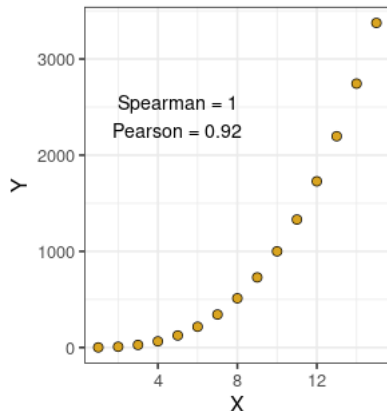
$$\begin{array}{l} X = \{2, 4, 6, 9, 8\} \\ Y = \{7, 4, 1, 5, 3\} \end{array} \implies \begin{array}{l} X_{rank} = \{1, 2, 3, 5, 4\} \\ Y_{rank} = \{5, 3, 1, 4, 2\} \end{array}$$

Whereas Pearson's  $r$  measures *linear association*, Spearman's  $\rho$  measures the *monotonic association* (increasing or decreasing)

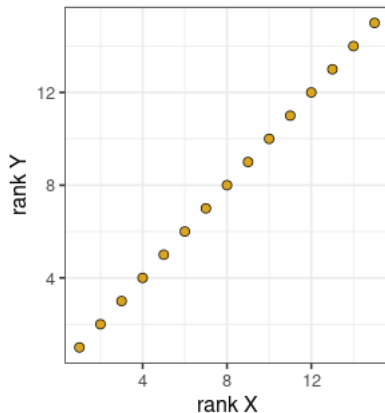
# Non-linear Association

$$y = x^3$$

X and Y



Rank X and Rank Y

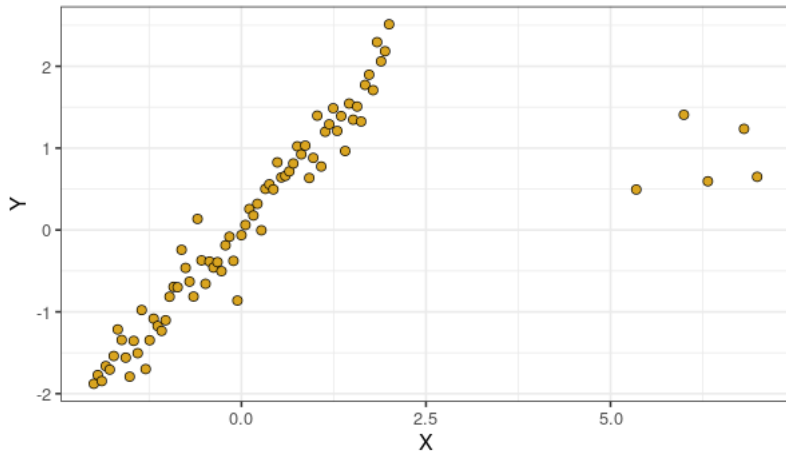


# Spearman Correlation

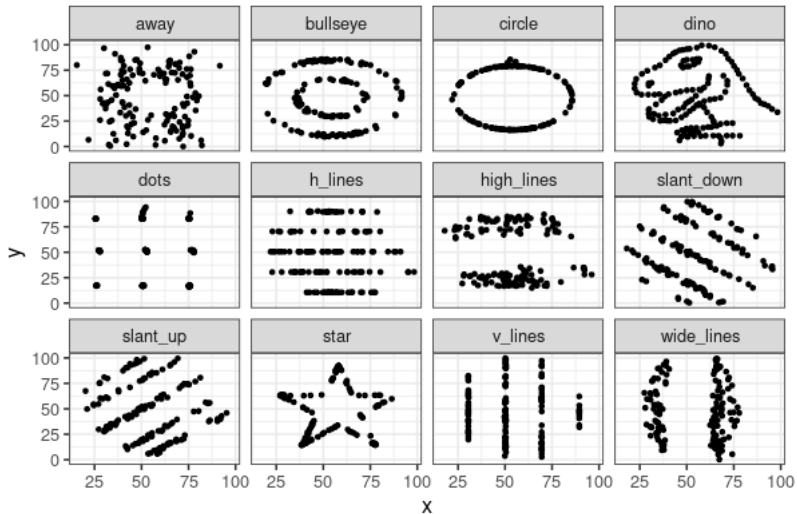
Spearman's correlation is more robust to outliers

Spearman Correlation = 0.95

Pearson Correlation = 0.77



# “Datasaurus Dozen”



# Ecological Correlation

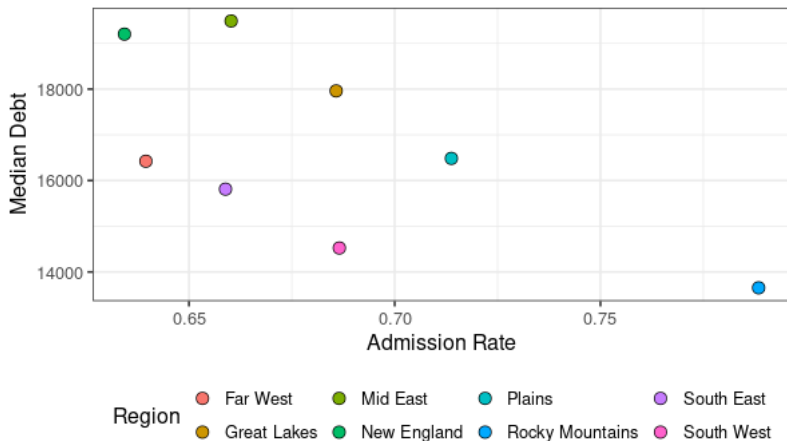
**Ecological correlations** compare variables for data that have been aggregated at an ecological level

- ▶ Countries
- ▶ States
- ▶ Schools

The *ecological fallacy* is a fallacy in which a conclusion is drawn that, because a correlation exists at a group level, it must exist at the individual level as well

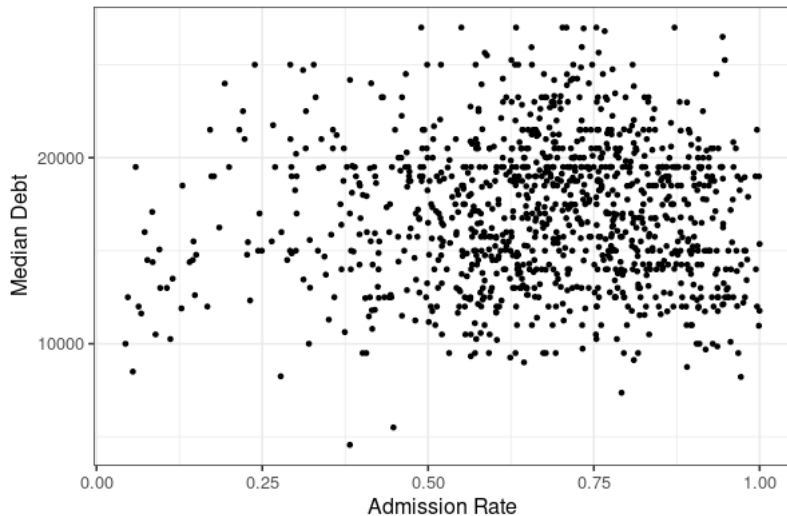
# College Ecological Fallacy

Grouping by region, the correlation between (mean) admission rate and (mean) median debt is  $r = -0.66$

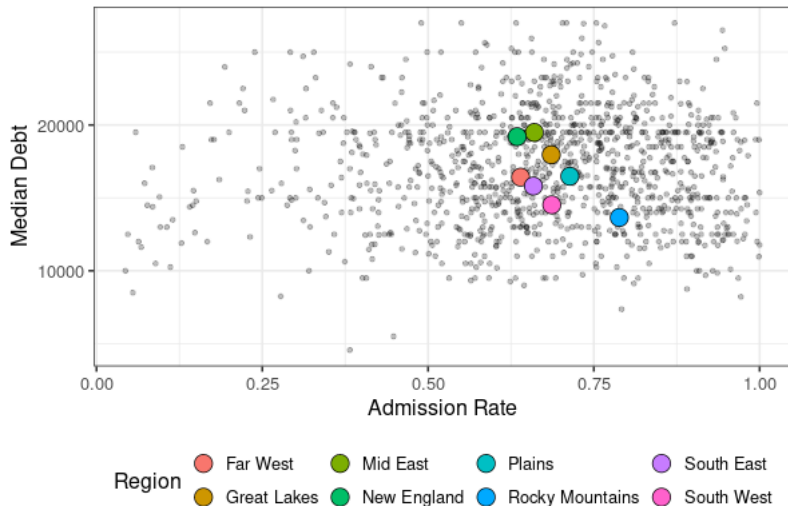


# College Ecological Fallacy

This completely disappears when we remove consideration of region, with  $r = 0.02$



# College Ecological Fallacy



# Correlation $\neq$ Causation

We can have a large correlation value between 2 variables. This does not mean the explanatory variable is *causing* a change in the response variable.

Examples with high correlation but where no causal claims can be made:

- ▶ Literacy Rate and Gross Domestic Product (GDP) in countries
- ▶ average number of TVs in a household and Life expectancy of countries
- ▶ ice-cream sales and shark attacks

**Lurking Variable:** a third variable that explains the relationship between two variables with high correlation

- ▶ **Pearson's correlation** strength of *linear association* (and direction)
  - ▶ Correlation is *average product of z-scores*
- ▶ **Spearman rank correlation** useful for data with outlier's or non-linear (but monotone) relationship
- ▶ Be careful with **ecological correlations** – inference for a group is not always valid for individuals
- ▶ Correlation  $\neq$  Causation