Correlation

Association between 2 Quantitative Variables

Grinnell College

September 23, 2024

Outline

On Friday we covered study design and how that affects the conclusions we make. These ideas will stick with us for the rest of the semester.

What we will be covering today:

How to measure the strength of the relationship between quantitative variables.

Review – Z-scores

A **z-score** or **standardized score** is a measurement that describes an observations *value* relative to the mean and standard deviation of a group

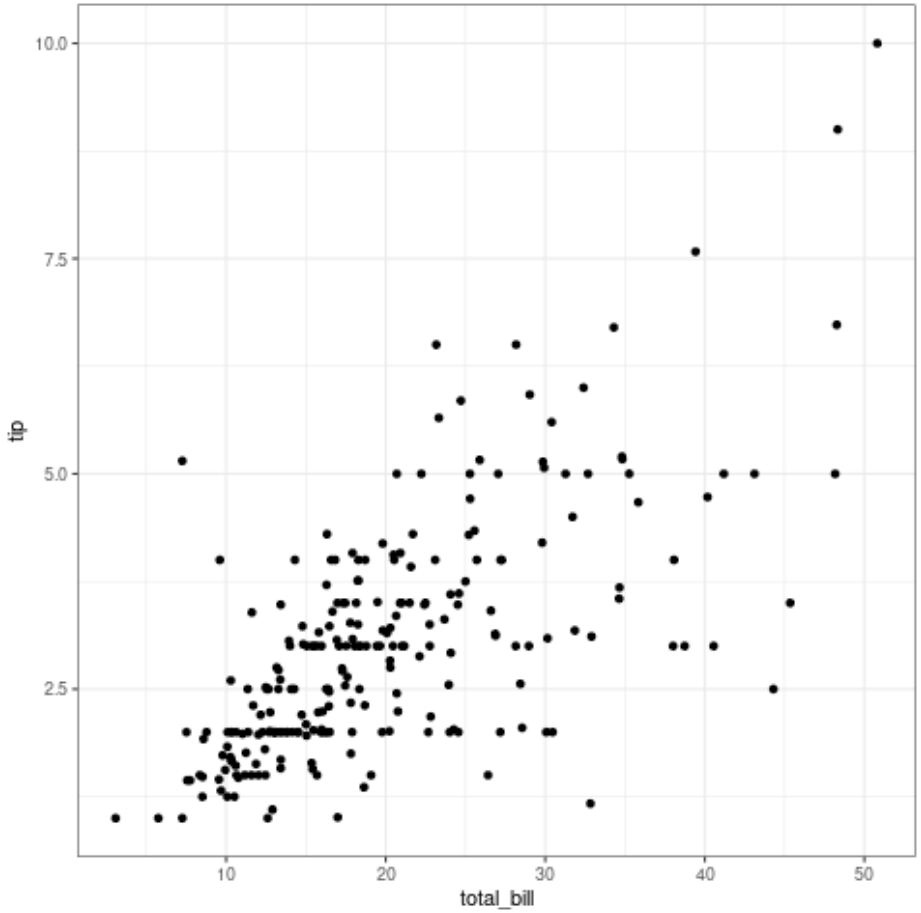
$$z_i = \frac{x_i - \mu}{\sigma}$$

In particular, there are two informative attributes related to a z-score:

1. The *sign* of the z-score tells us if the observation is above or below the group mean
2. The *magnitude* of the z-scores tells us how many standard deviations away from the mean an observation is

Review – Scatterplots

When we want to plot two quantitative variables $\rightarrow$ scatterplots



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

Review – Scatterplots

Describing associations in scatterplots:

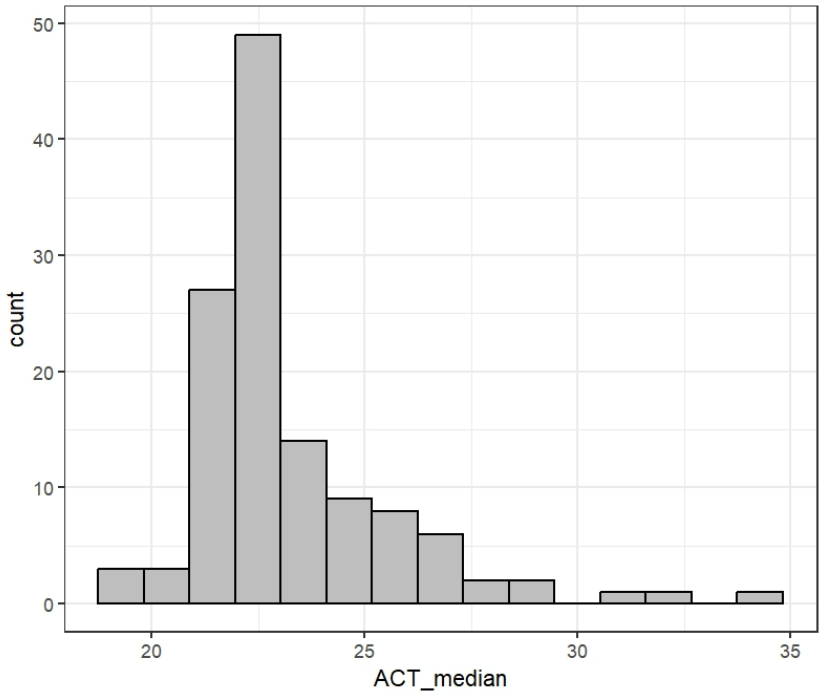
- **Form**: pattern? (linear / non-linear / curved / none)
- **Strength**: weak / moderate / strong
- **Direction**: positive / negative
- **Outliers**

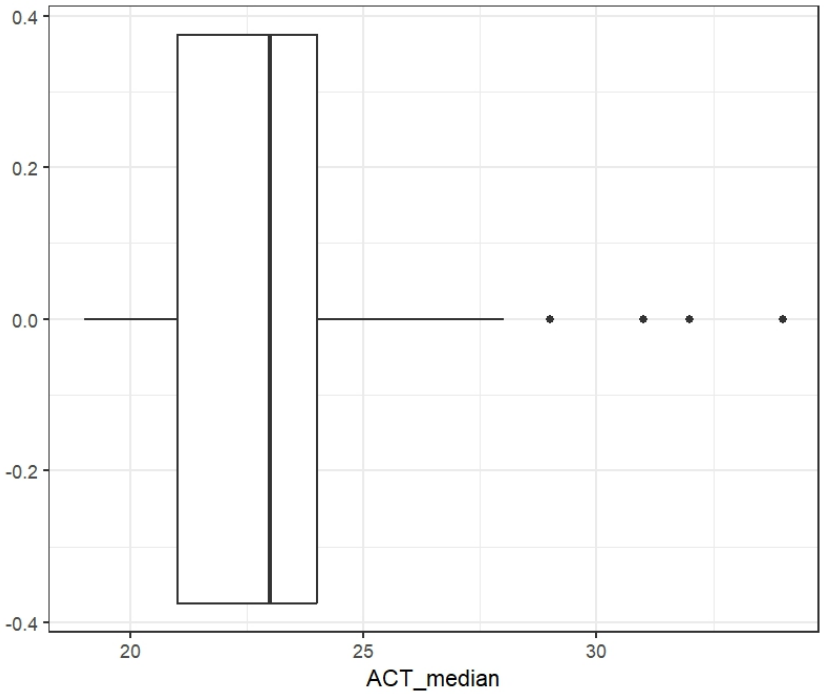
Extra on Outliers

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

- boxplot $\rightarrow$ points outside the 'whiskers'
- histogram $\rightarrow$ 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")

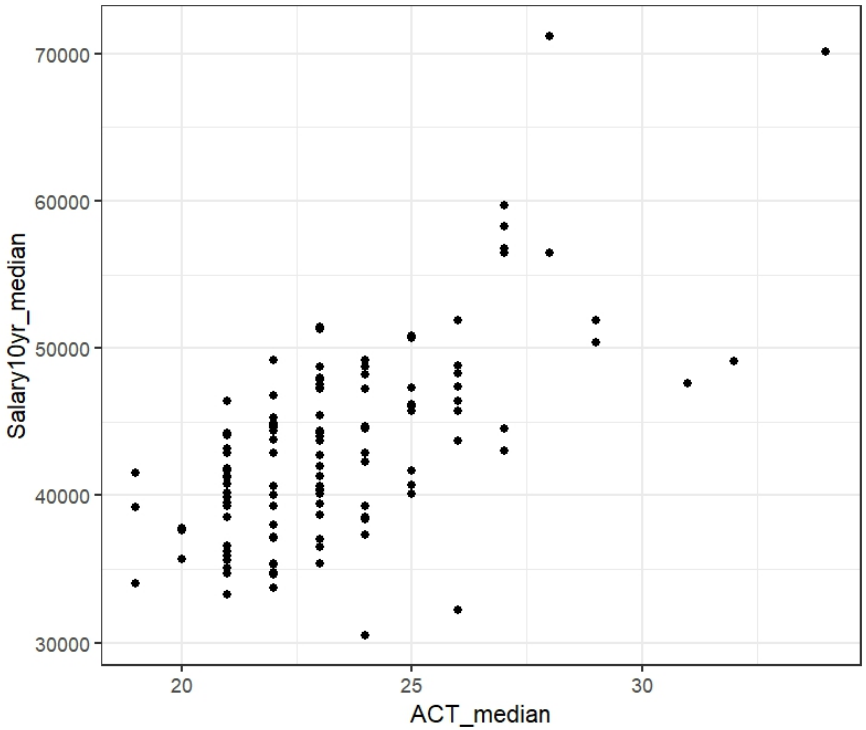




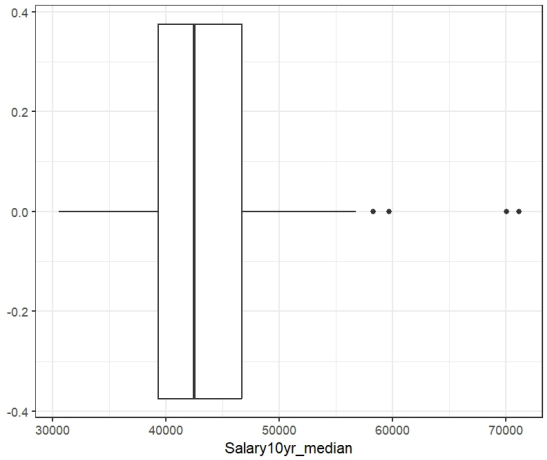
Extra on Outliers

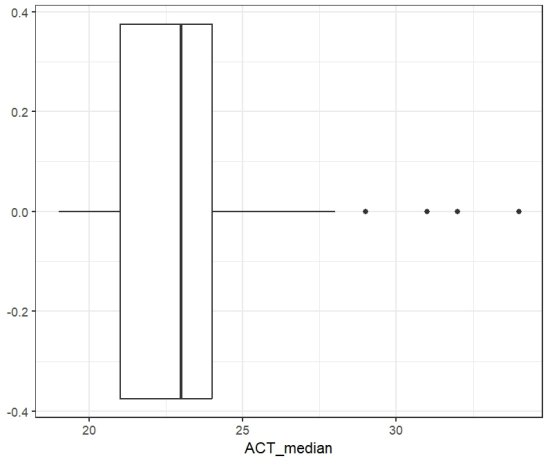
Outlier in a scatterplot

- 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"
```





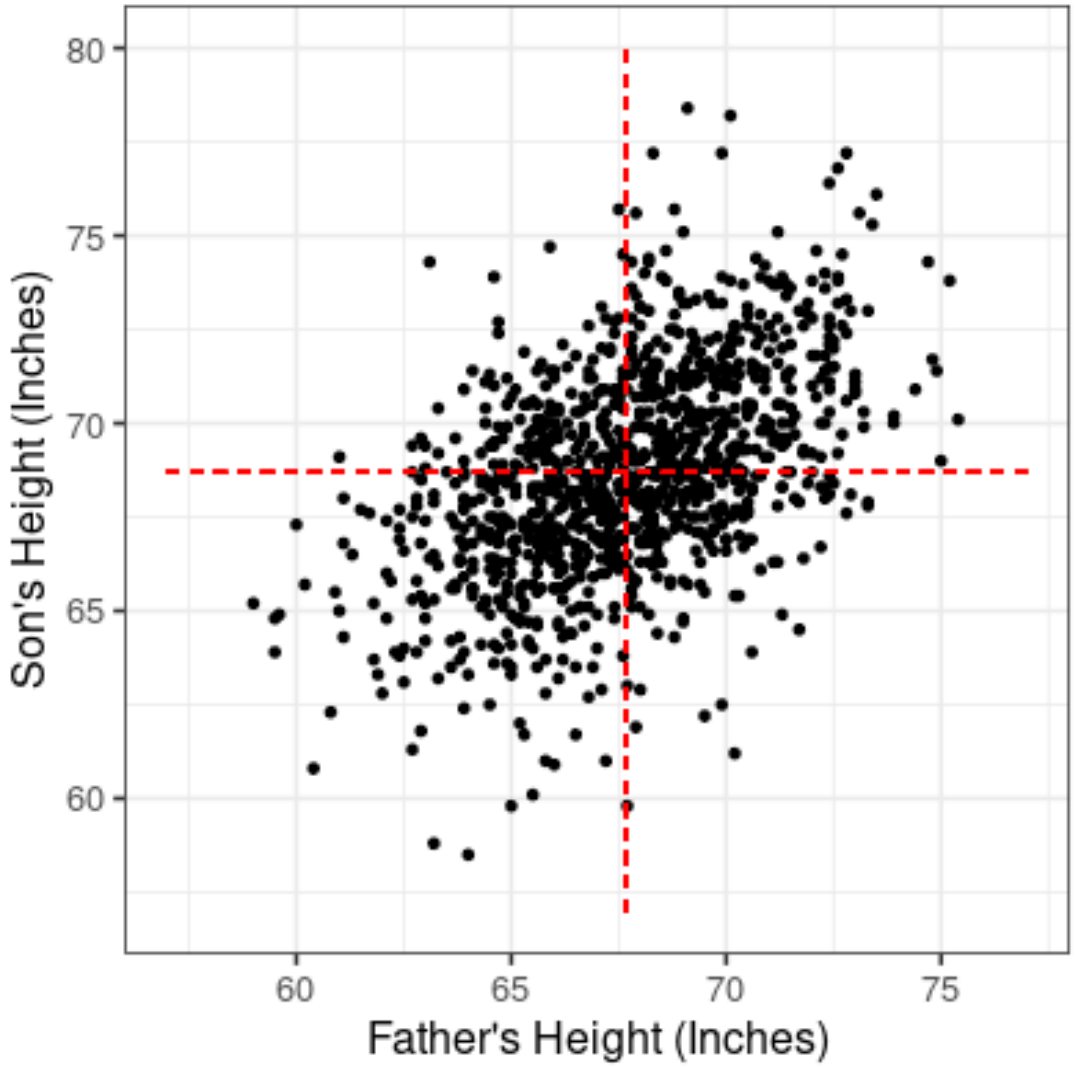
Pearson's Height Data

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


Son's Height (Inches)
80
75
70
65
60
60
65
70
75
Father's Height (Inches)

Pearson's Correlation Coefficient

Heights clearly associated, but how to quantify?

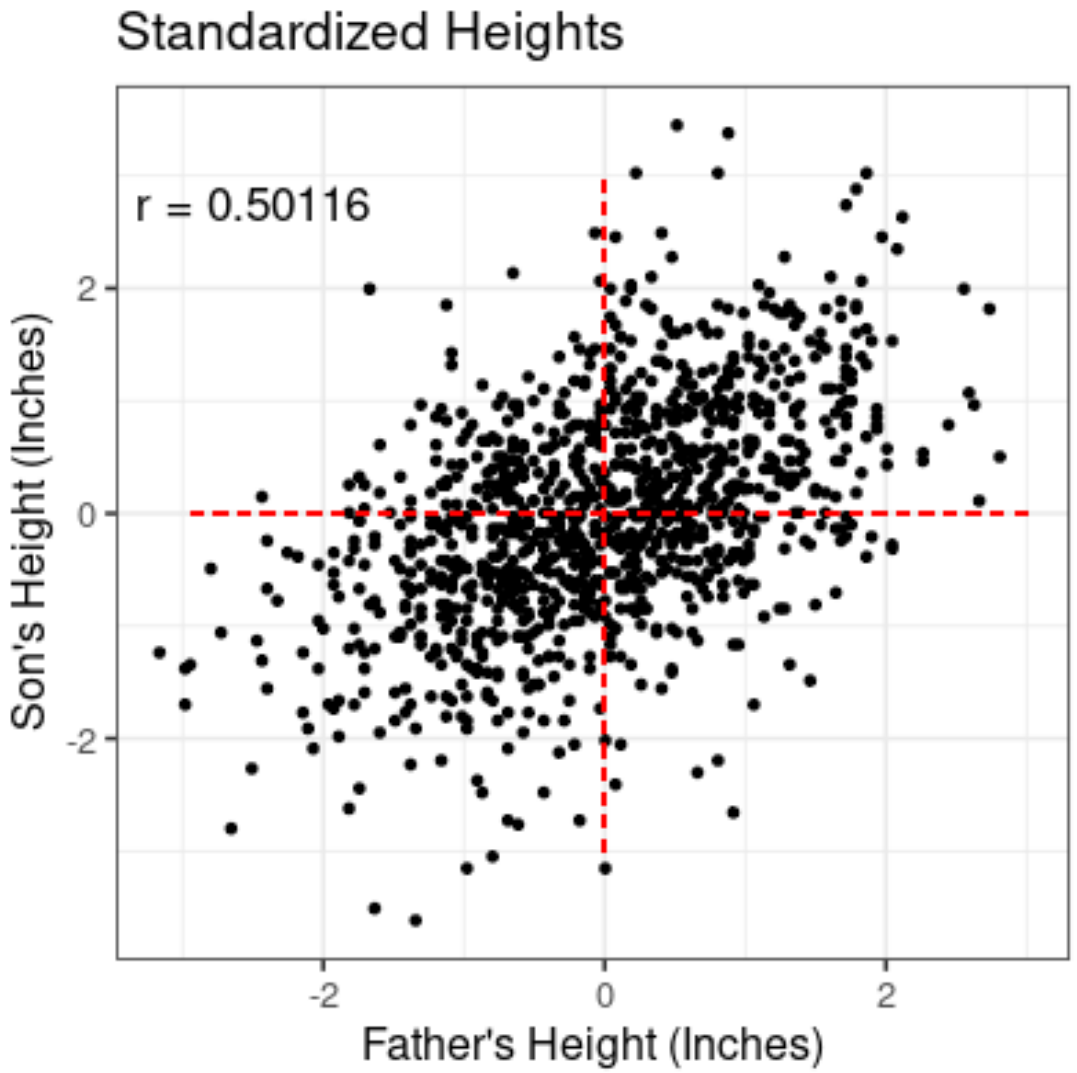
Building upon the work from French scientist Francis Galton, Pearson developed the **Pearson's correlation coefficient (r)**:

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

$$= \frac{1}{n-1}\sum_{i=1}^{n}(z_{x_i})(z_{y_i})$$

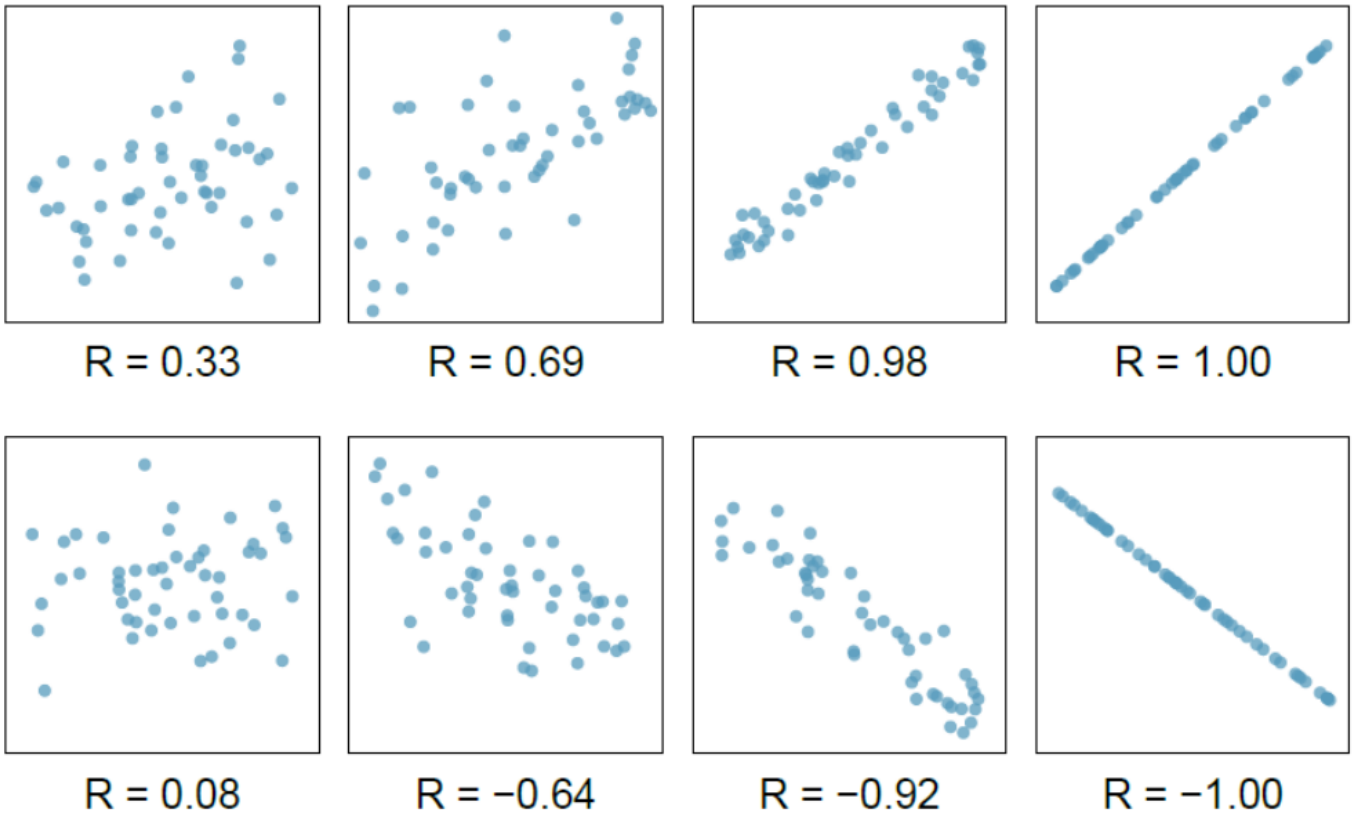
If above-average values of X are common among cases with above-average values of Y (or vice-versa), we should expect r to be positive

Height Data – Standardized



Correlation Examples

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

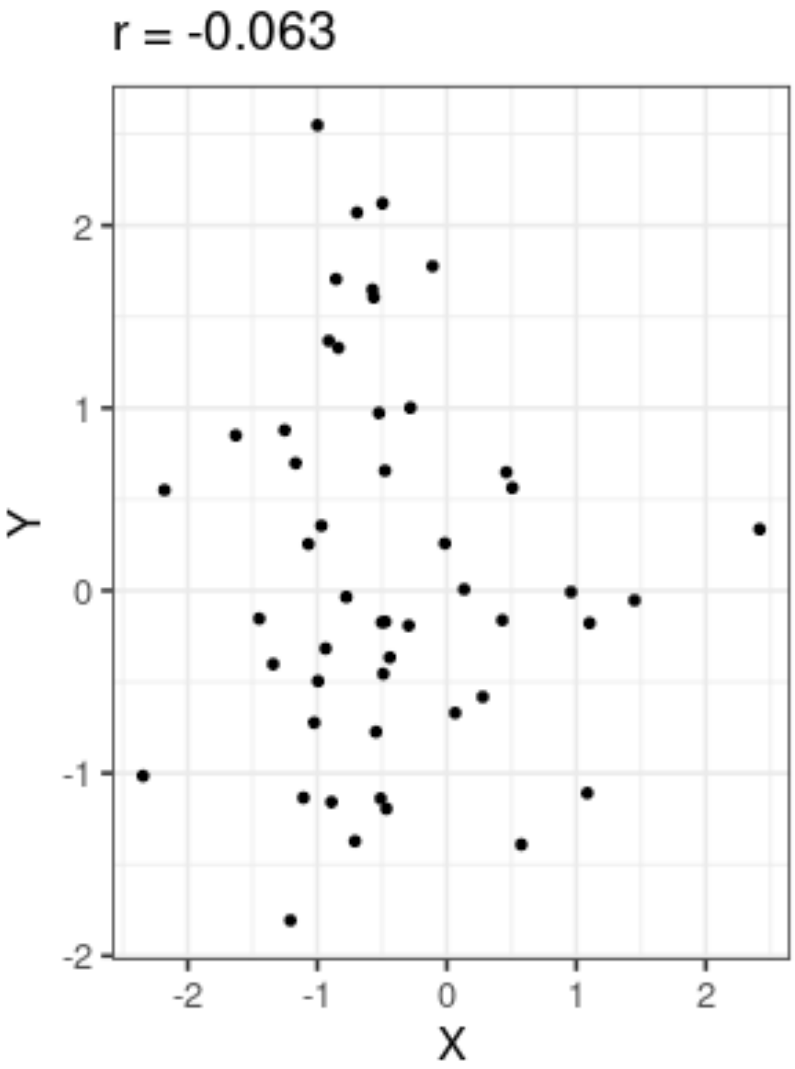


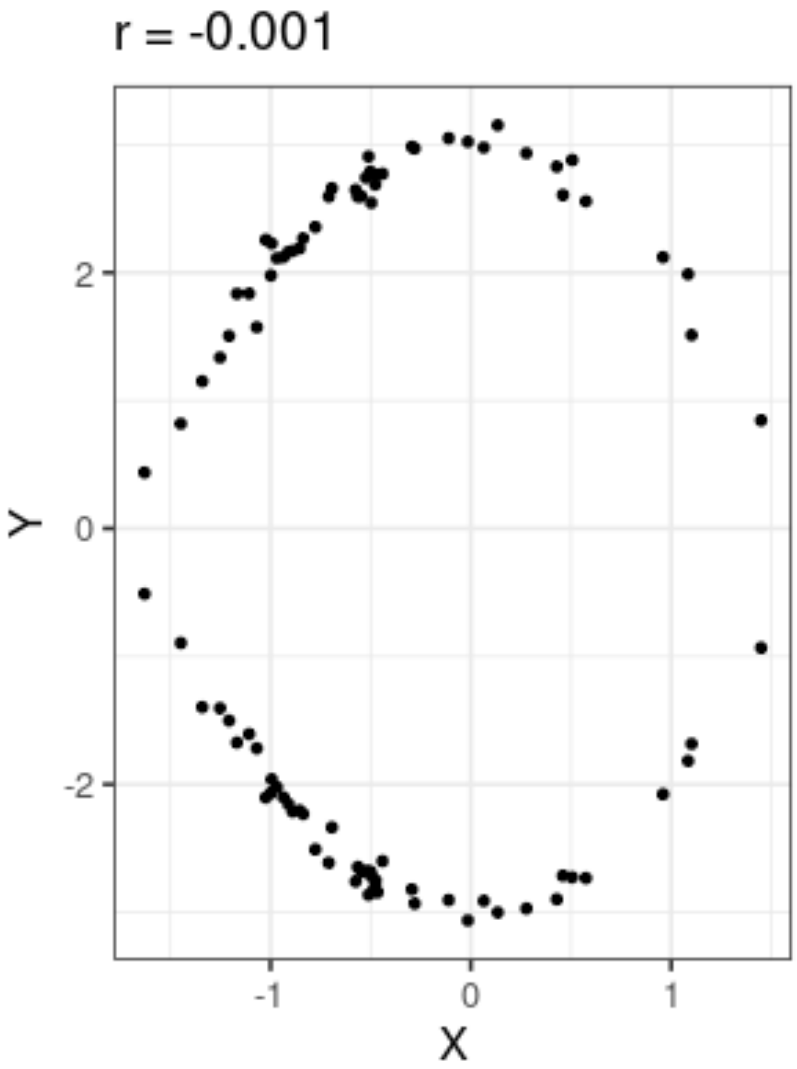
What is considered "strong"?

Correlation Coefficient		Dancey & Reidy (Psychology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+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





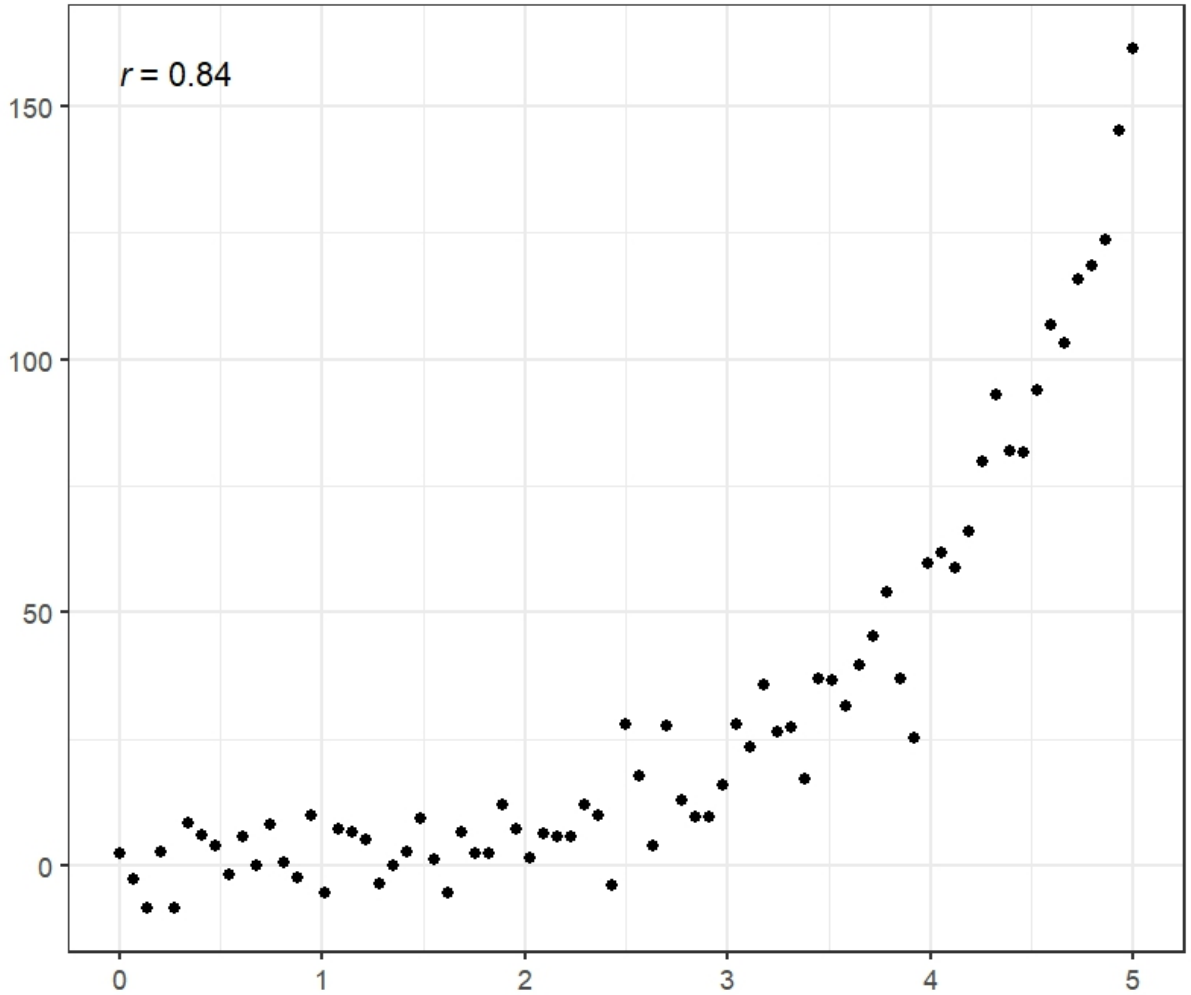
Correlation Properties

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

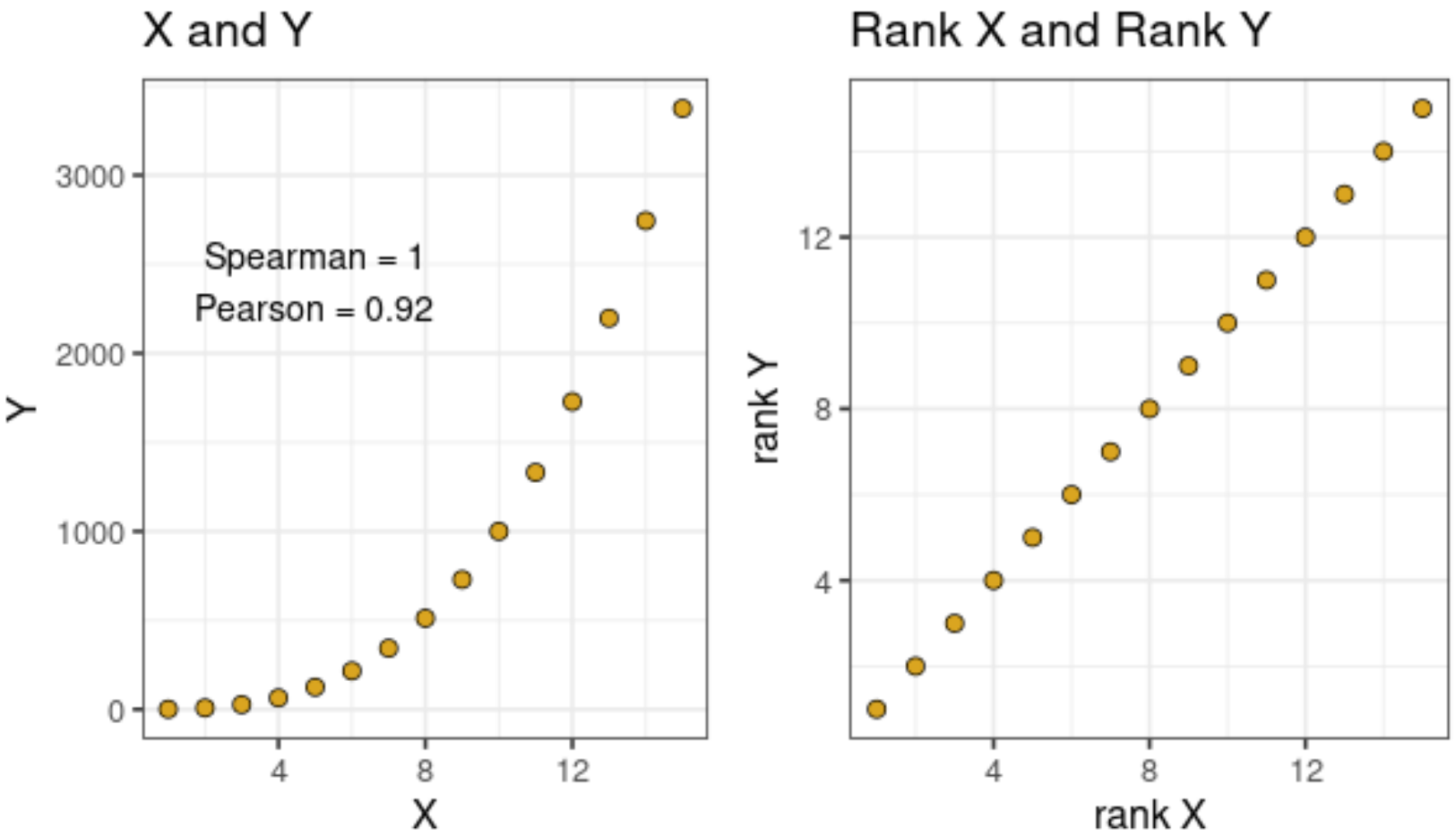
In addition to Pearson, we have **Spearman's rank correlation** (denoted ρ) where the values of X and Y are replaced with their rank order from smallest to largest before correlating:

$$\begin{aligned} X &= \{2, 4, 6, 9, 8\} \\ Y &= \{7, 4, 1, 5, 3\} \end{aligned} \quad \Longrightarrow \quad \begin{aligned} X_{rank} &= \{1, 2, 3, 5, 4\} \\ Y_{rank} &= \{5, 3, 1, 4, 2\} \end{aligned}$$

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

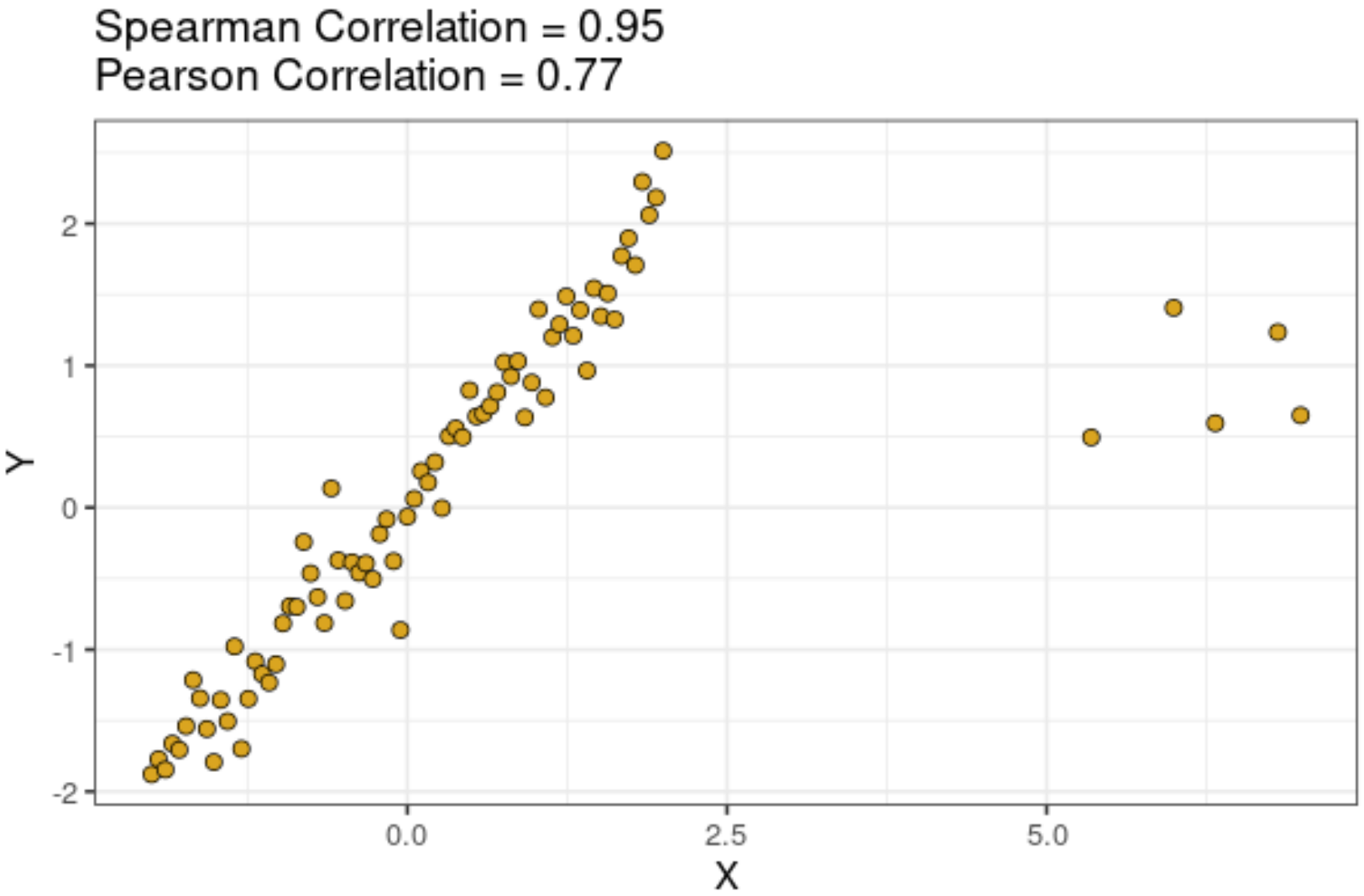
Non-linear Assocation

$$y = x^3$$

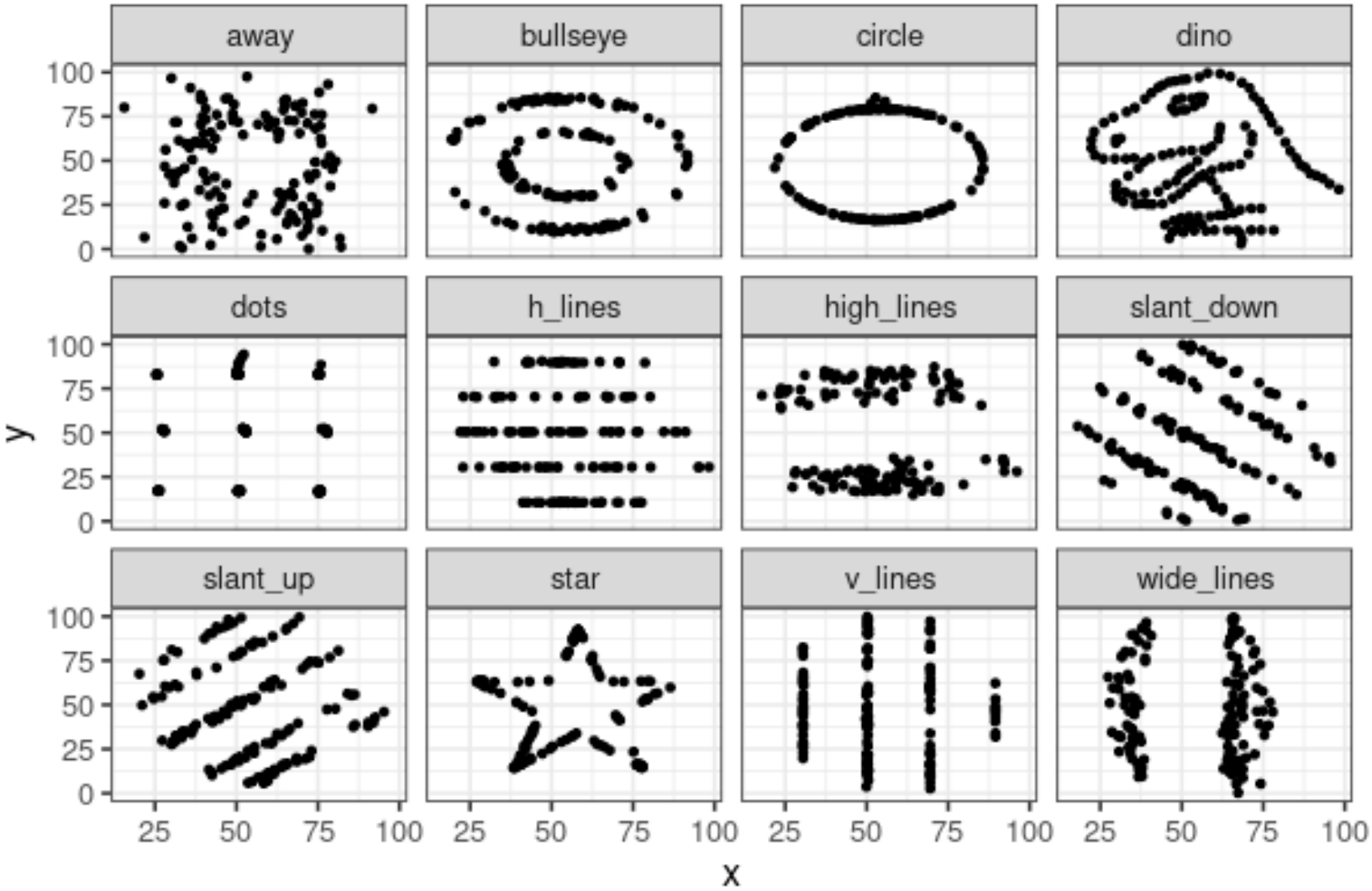


Spearman Correlation

Spearman's correlation is more robust to outliers



"Datasauraus Dozen"



Ecological Correlation

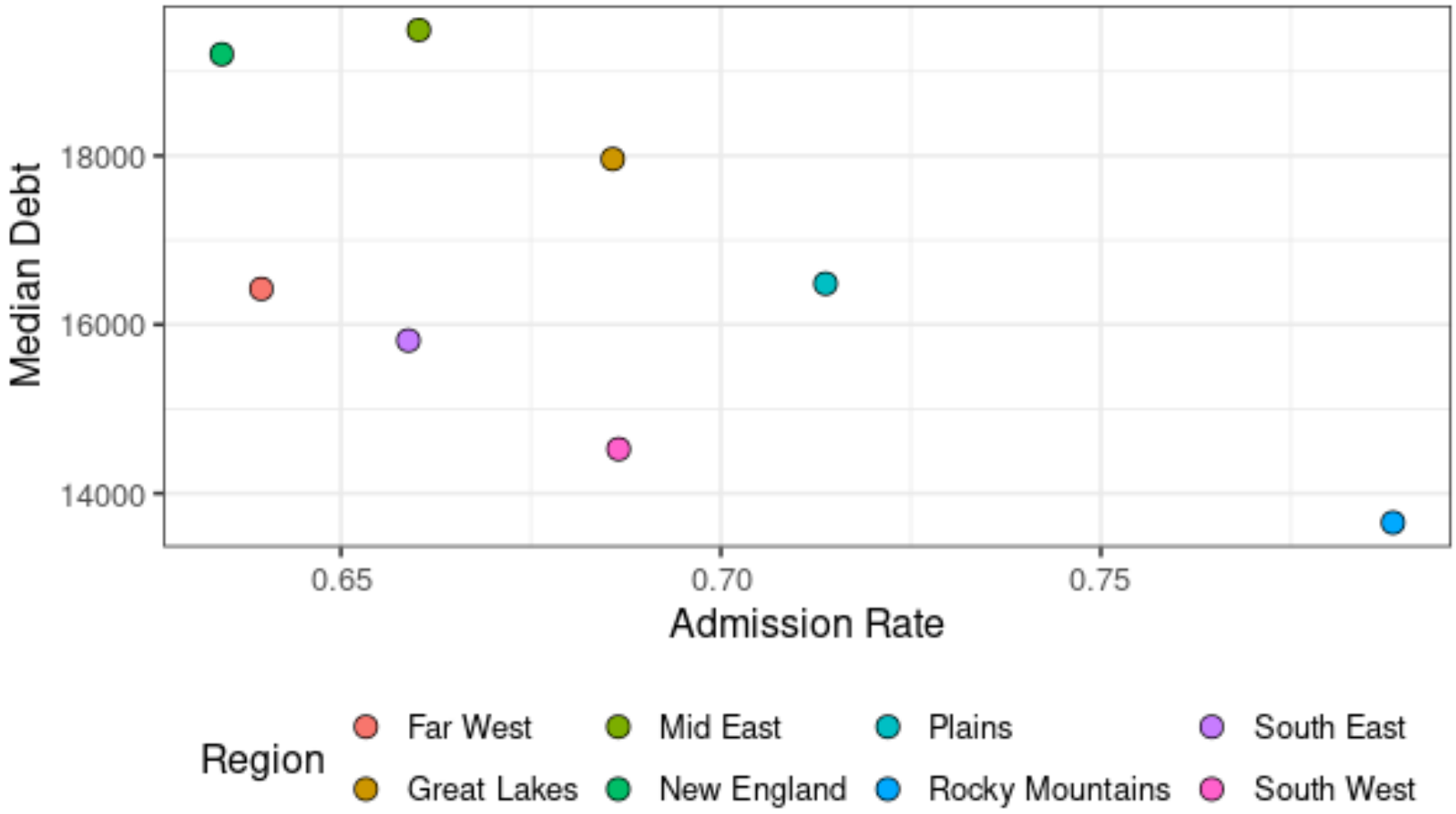
Ecological correlations compare variables for data that have been aggregated at an ecological level

- 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

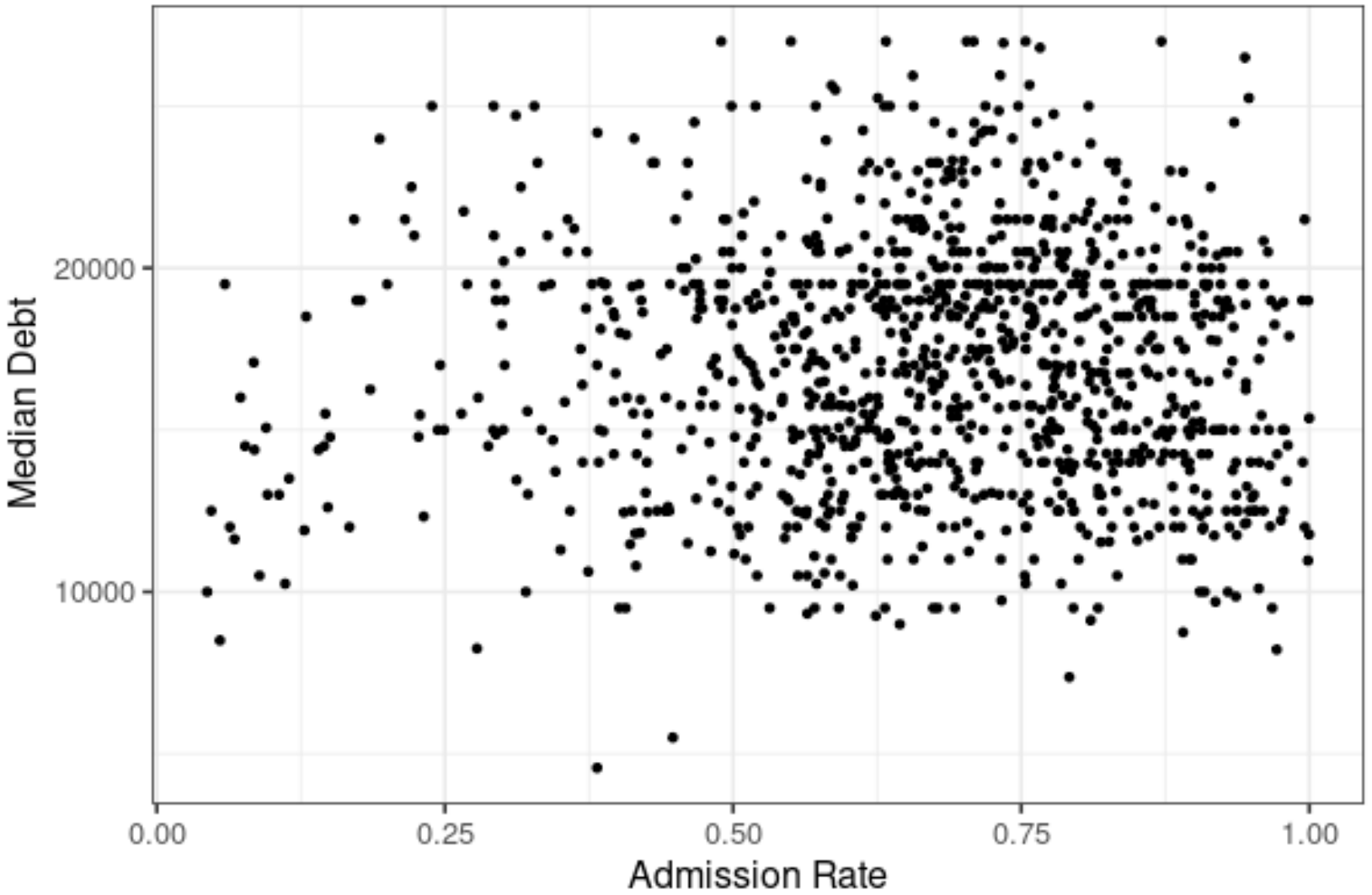
College Ecological Fallacy

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

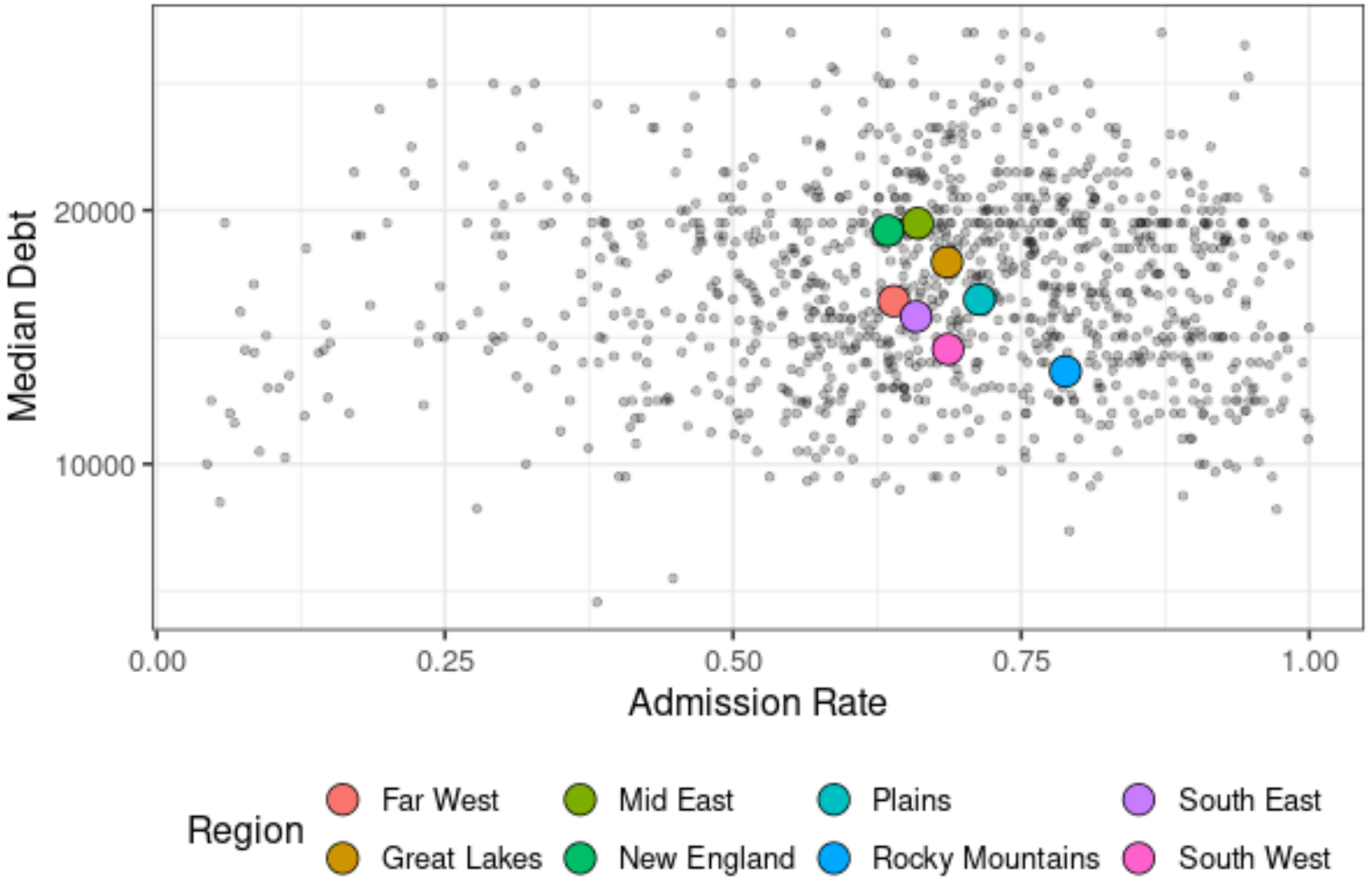


College Ecological Fallacy

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

Review

- **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