Z-scores and Correlation

Grinnell College

September 18, 2024

Today

Big things today:

- Standardization and z-scores
- Correlation
  - Pearson
  - Spearman (rank)
  - Ecological fallacies

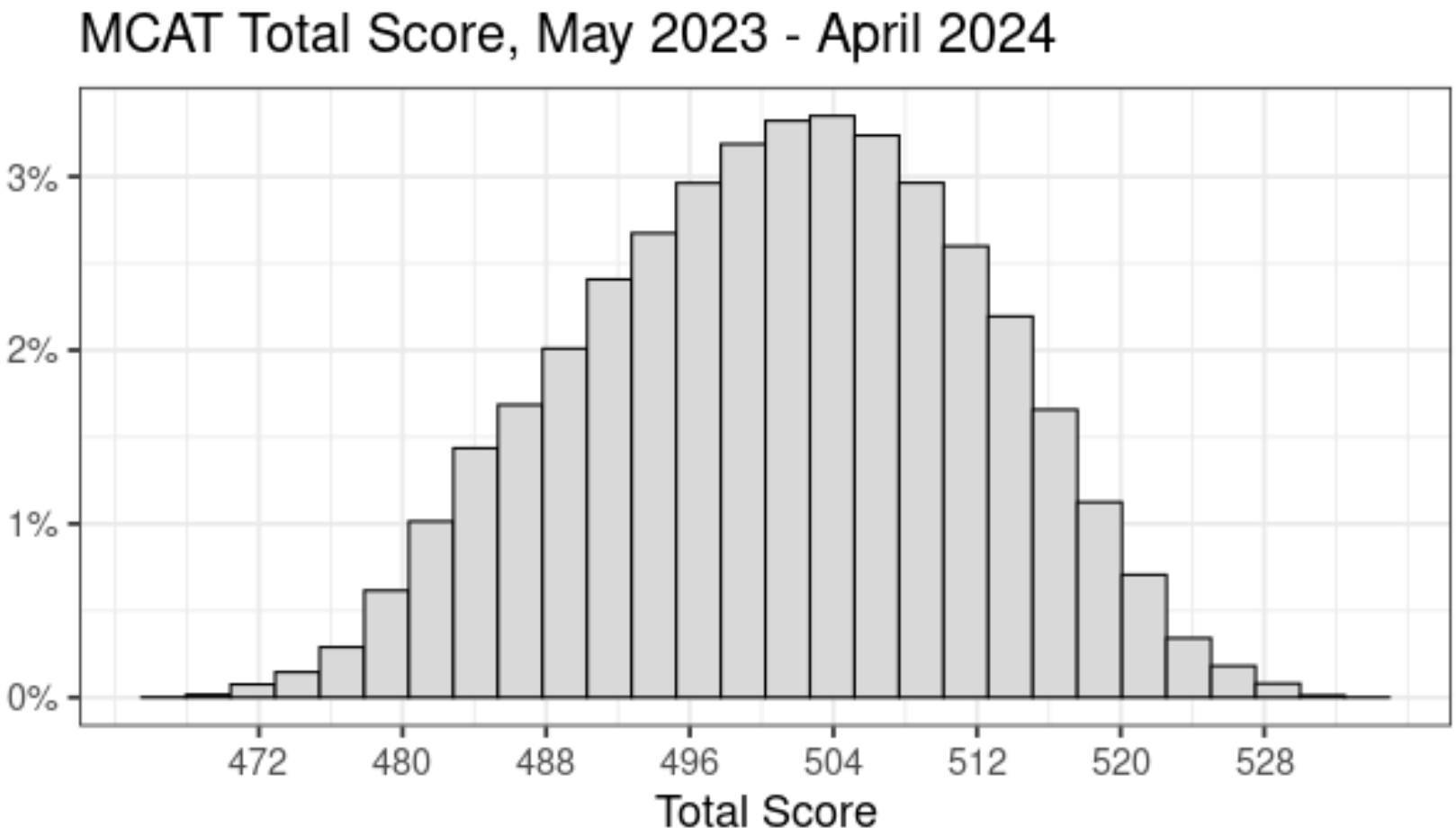
Z-scores

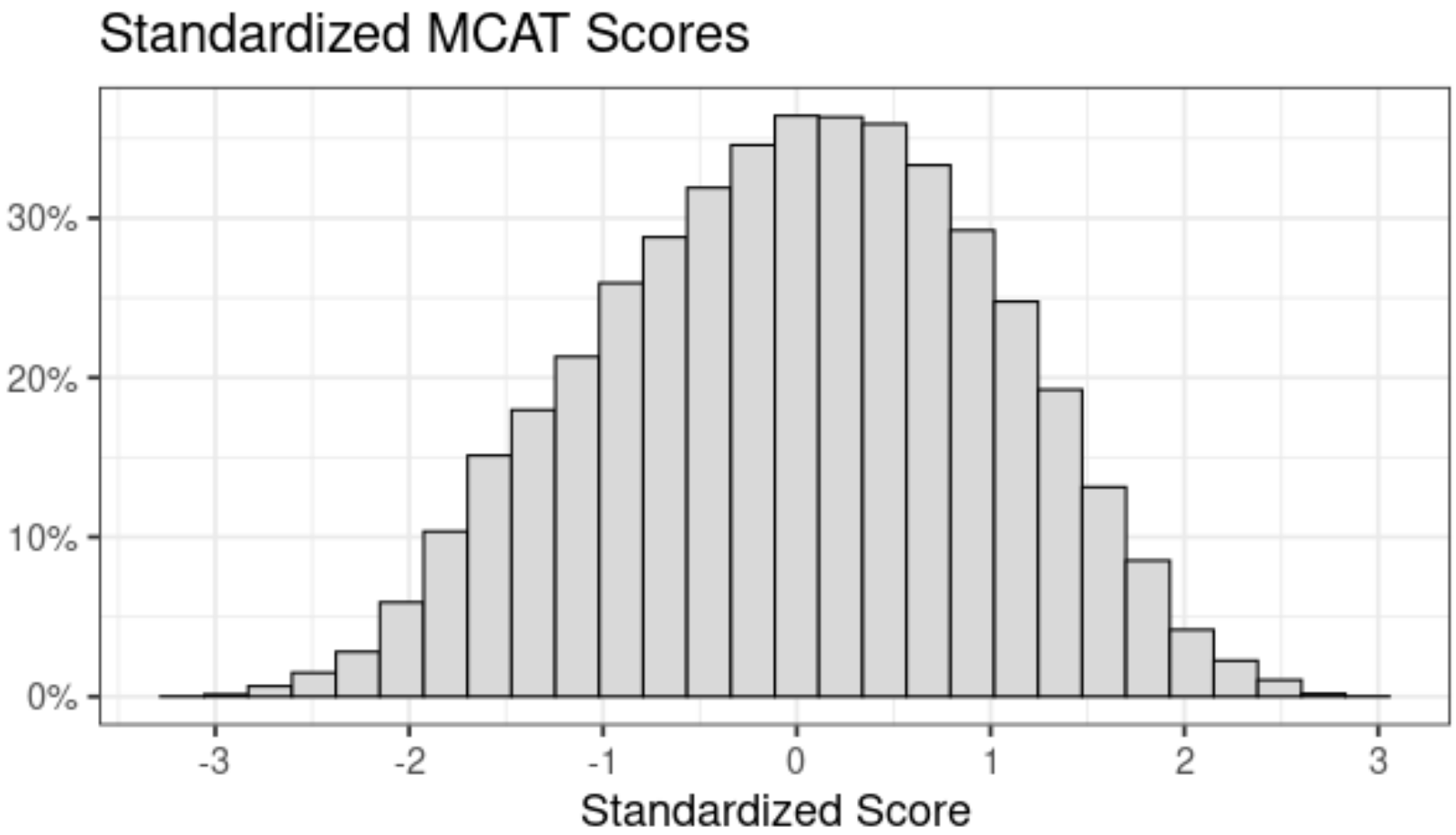
A **z-score** or **standardized score** is a measurement that describes an observations *value* relative to the mean and standard deviation of a group

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

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


MCAT Total Score, May 2023 - April 2024
3%
2%
1%
0%
472
480
488
496
504
512
520
528
Total Score


Standardized MCAT Scores
30%
20%
10%
0%
-3
-2
-1
0
1
2
3
Standardized Score

MCAT Scores

Based on data from March 2023-April 2024, the average total MCAT score was 501.03, with a standard deviation of 10.961, giving us the following summary statistics:

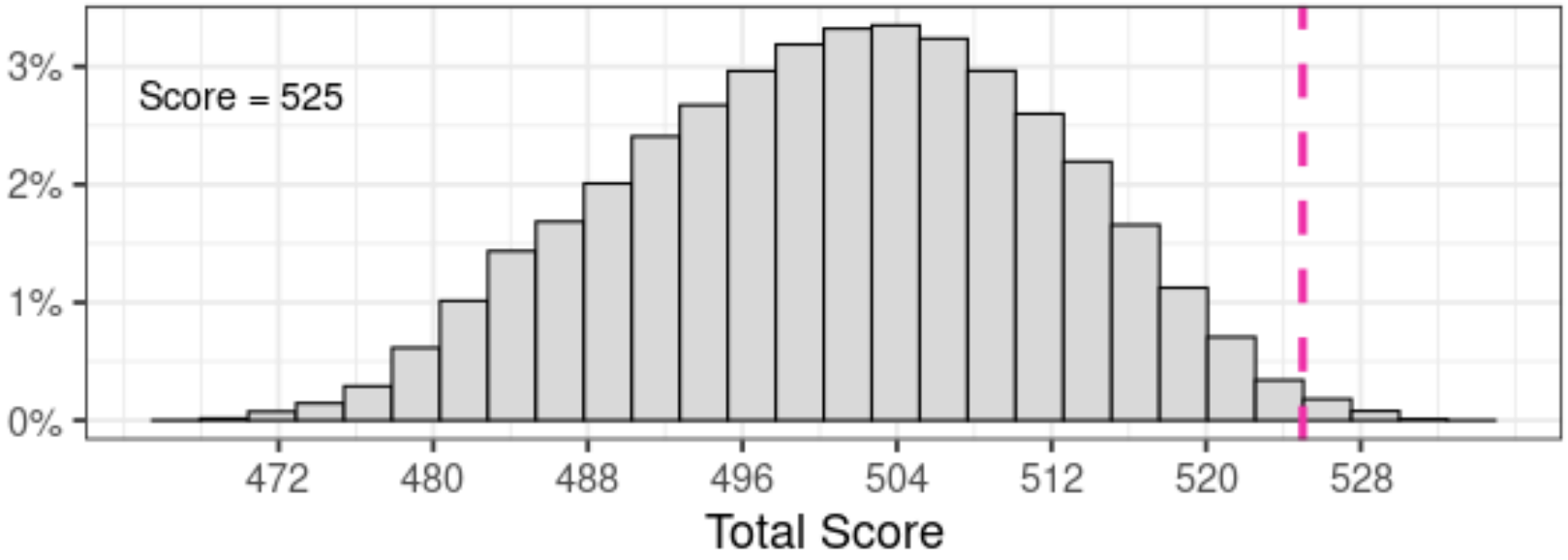
$$\mu = 501.03, \quad \sigma = 10.96$$

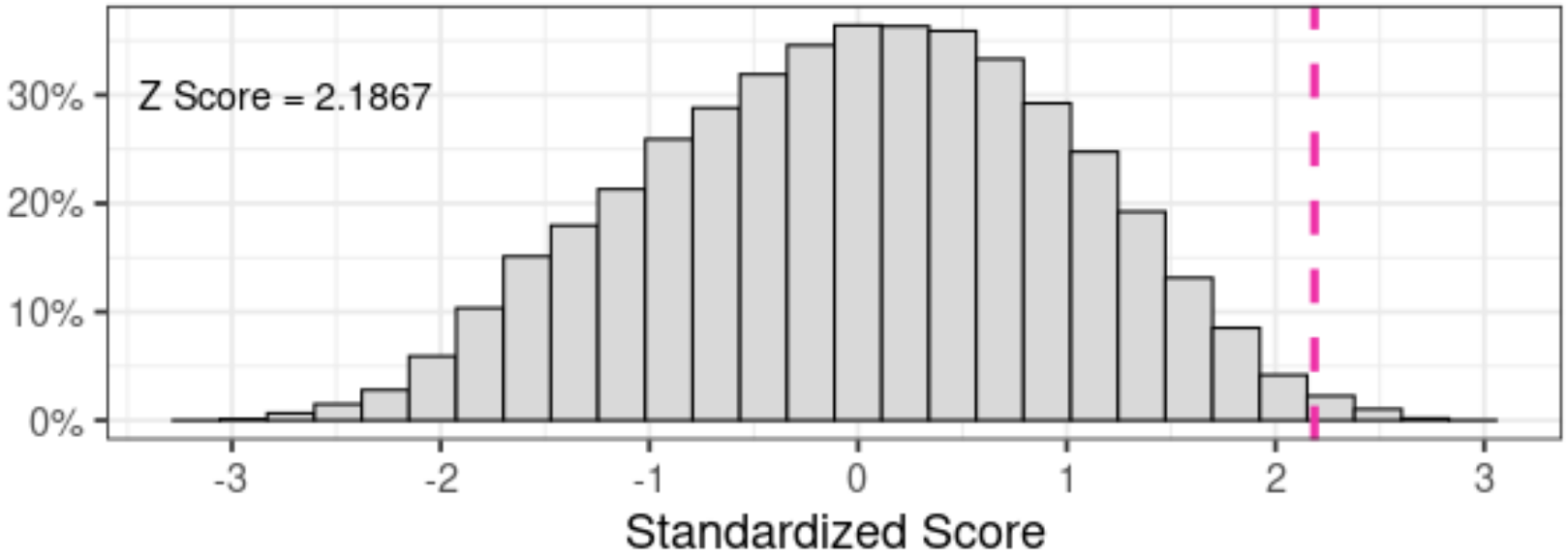
For an MCAT score of 525, we find a standardized value of

$$z = \frac{525 - 501.03}{10.96} = 2.18$$

This tells us that:

1. The observation is greater than the mean, as it is positive
2. The observation is 2.18 standard deviations greater than the mean


Score = 525
3%
2%
1%
0%
472
480
488
496
504
512
520
528
Total Score


Z Score = 2.1867
30%
20%
10%
0%
-3
-2
-1
0
1
2
3
Standardized Score

Slide

Clever transition slide to next topic

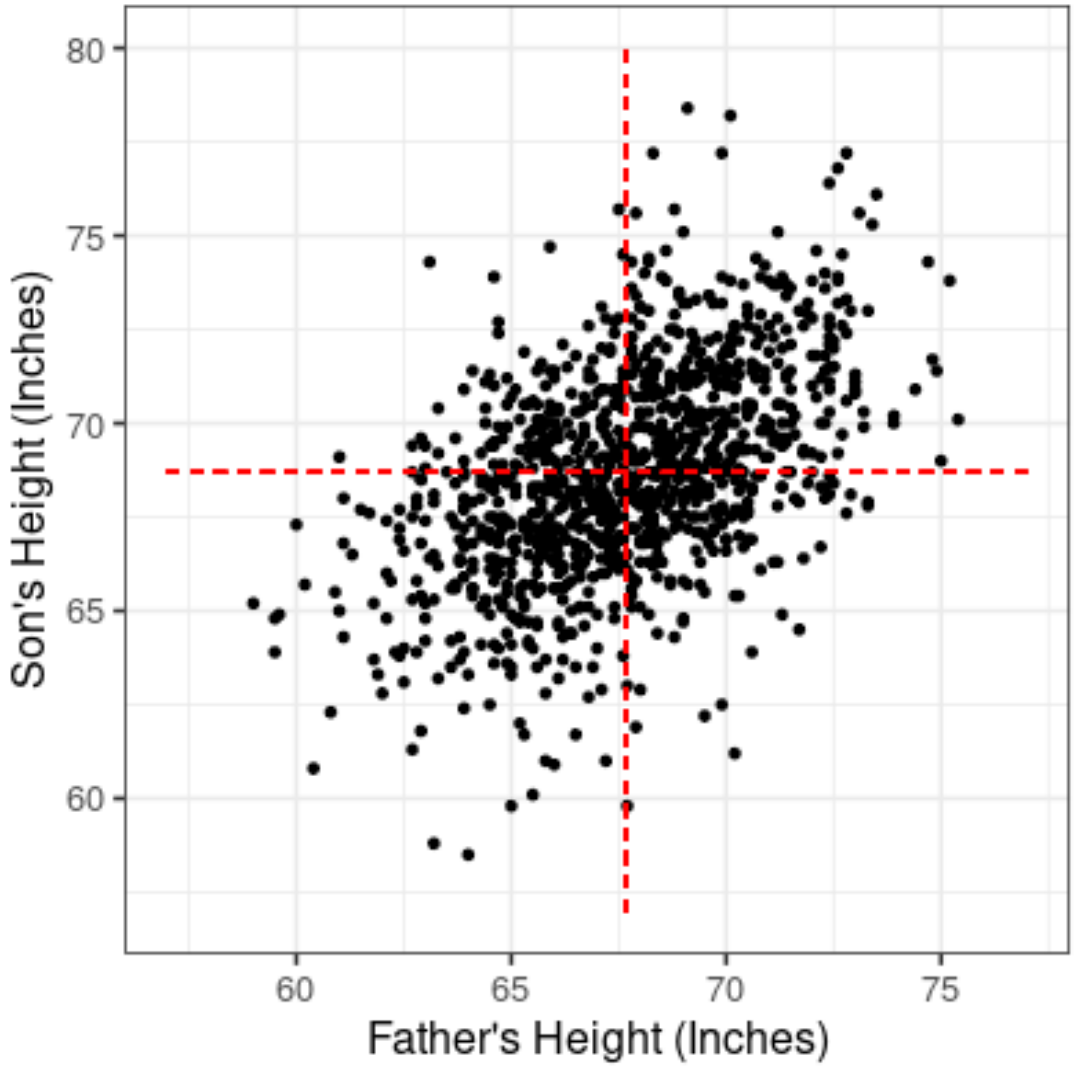
Pearson's Height Data

In the 1880's the 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


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

Pearson's Correlation Coefficient

Heights clearly associated, but how to quantify?

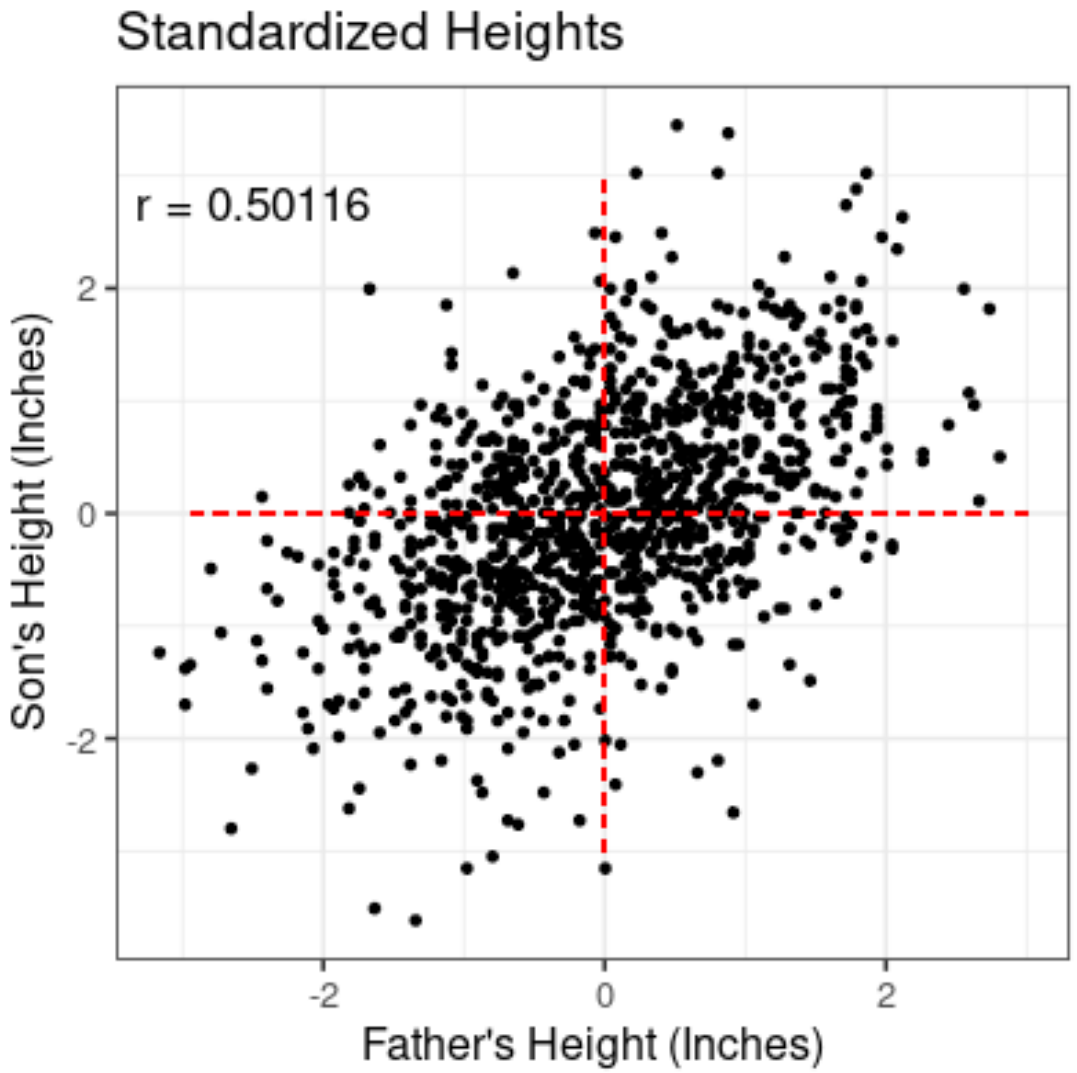
Building upon the work from French scientist Francis Galton, Pearson developed the **Pearson's correlation coefficient (r)**:

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

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

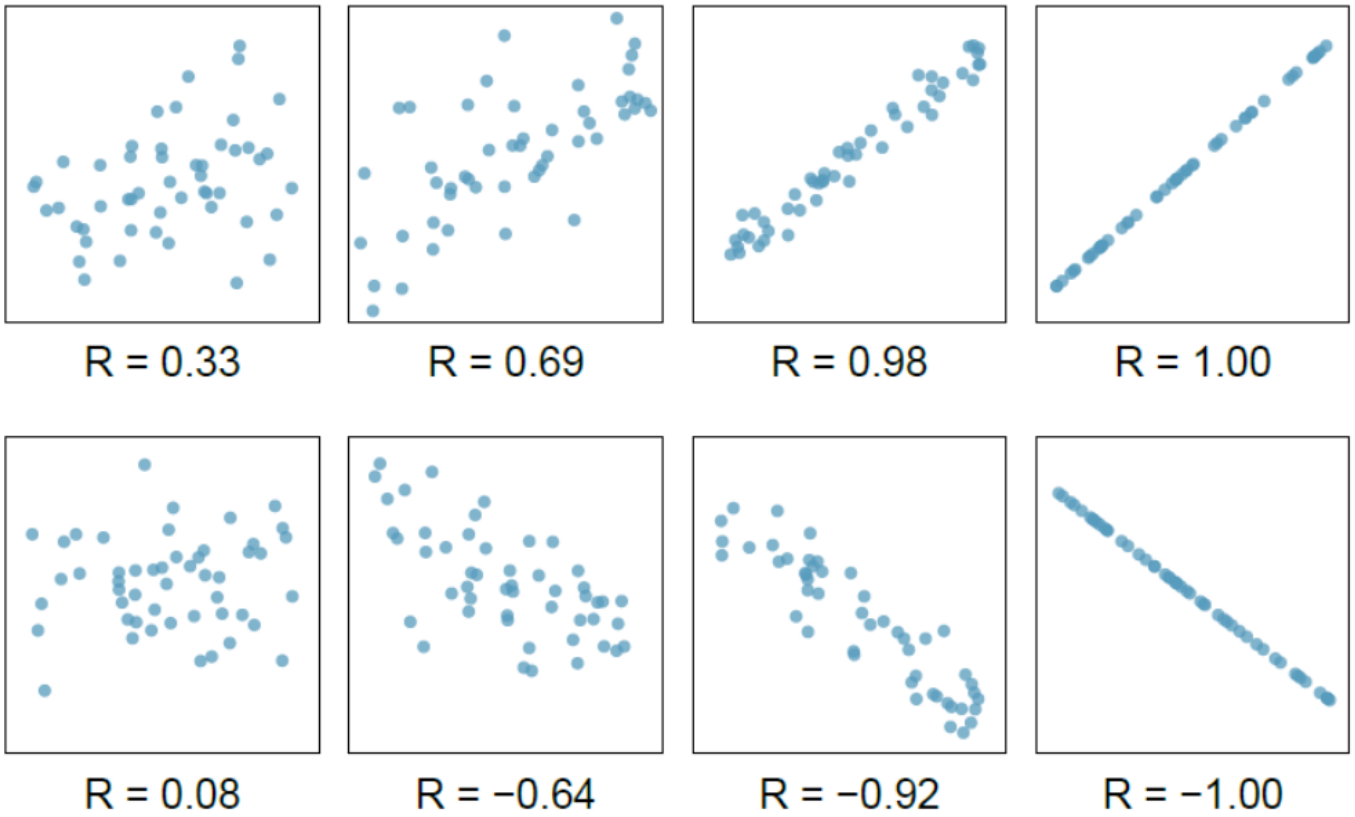
If above-average values of X are common among cases with above-average values of Y (or vice-versa), we should expect r to be positive

Height Data – Standardized



Correlation Examples

Pearson's correlation coefficient tells us the strength of *linear* association between two quantitative variables

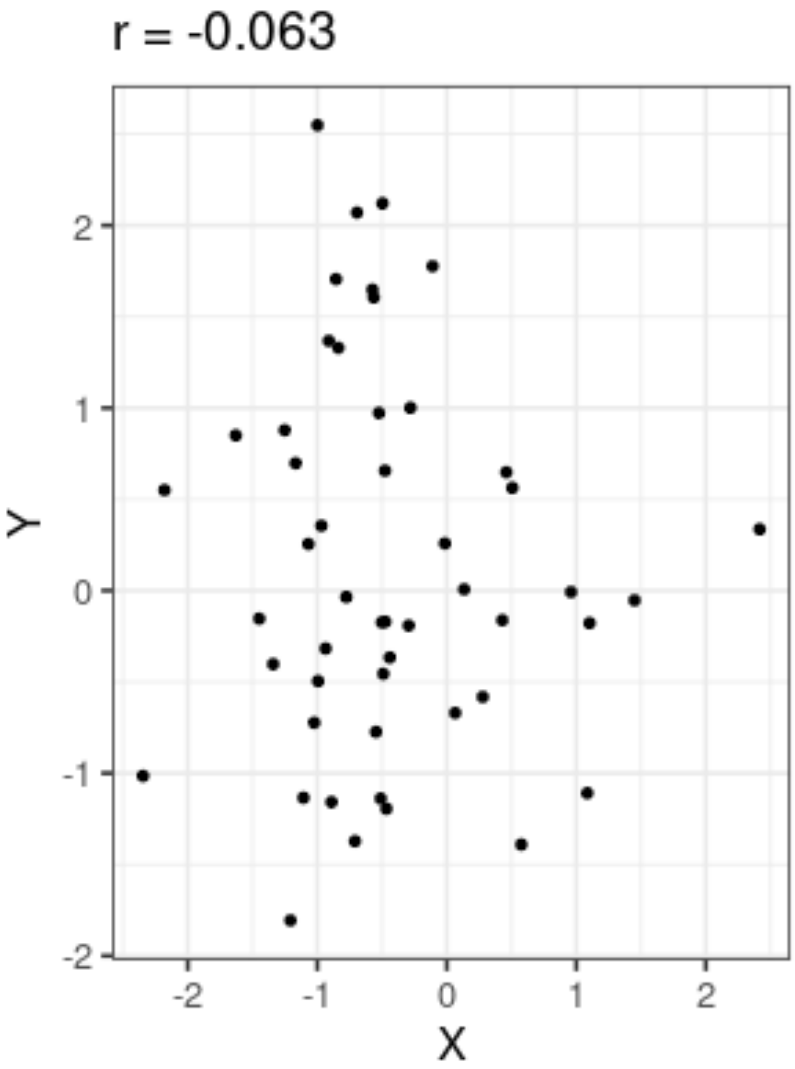


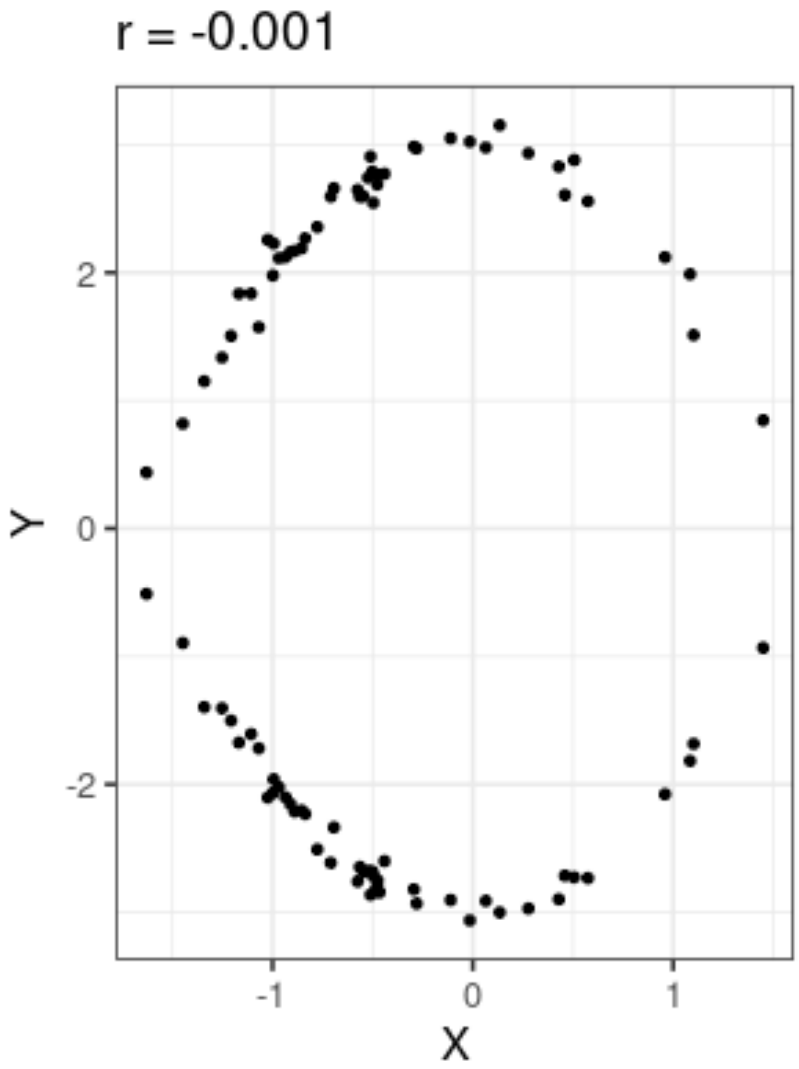
What is considered "strong"?

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





Non-linear Assocation

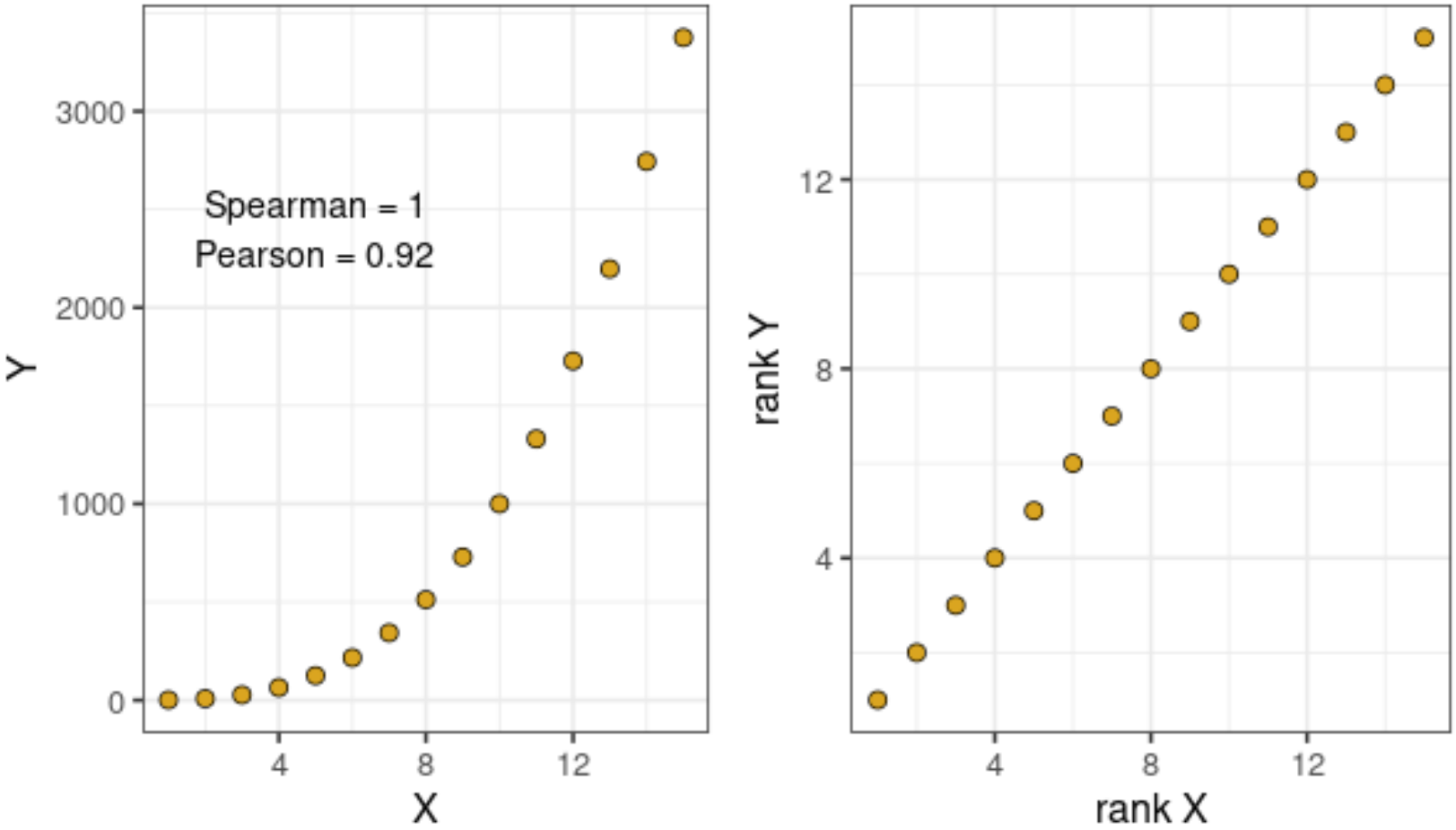
In addition to Pearson, we have **Spearman's rank correlation** (denoted ρ) where the values of X and Y are replaced with their rank order from smallest to largest before correlating:

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

Whereas Pearson's r measures *linear association*, Spearman's ρ measures the *monotonic association*

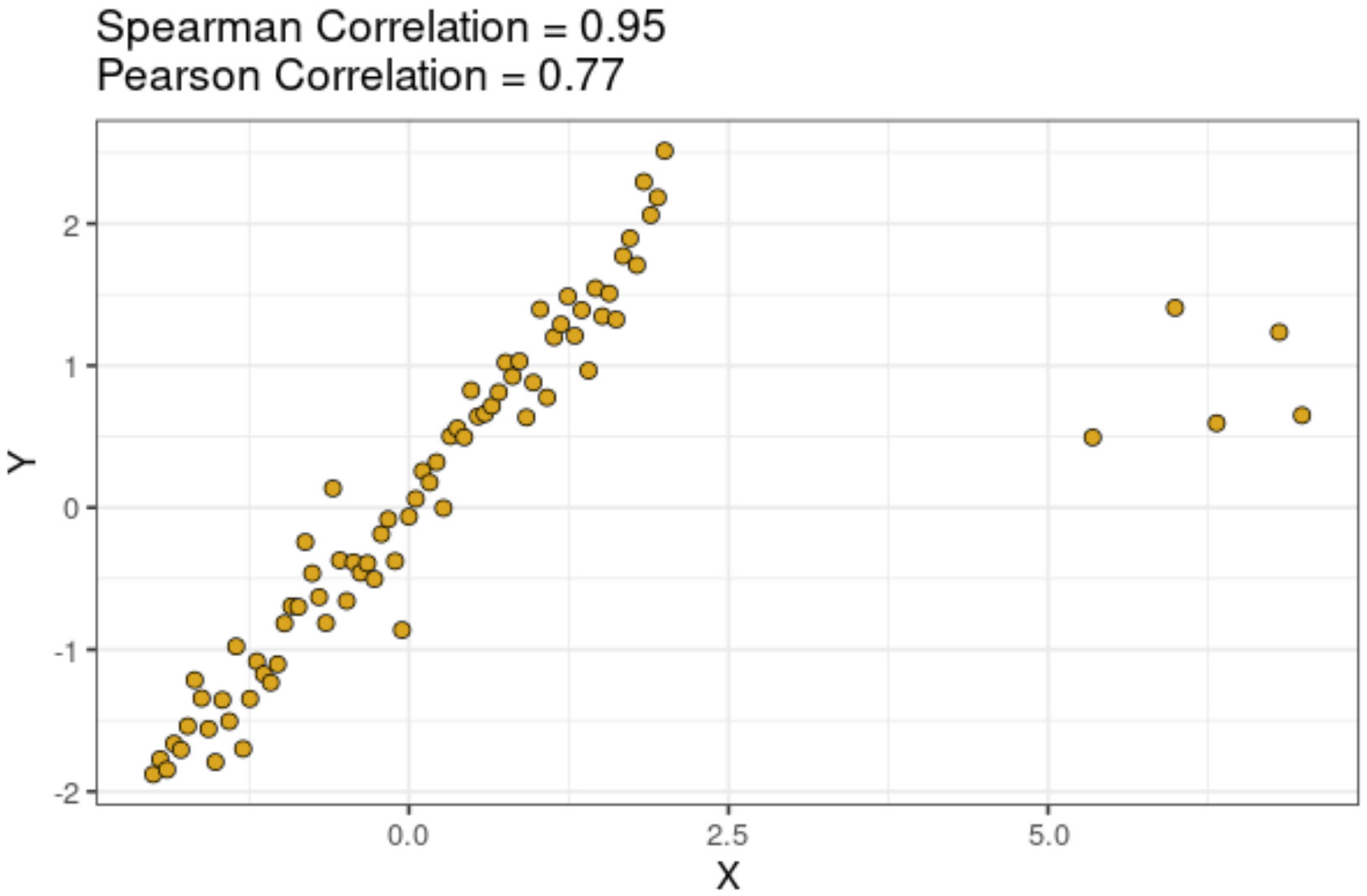
Non-linear Assocation

$$y = x^3$$

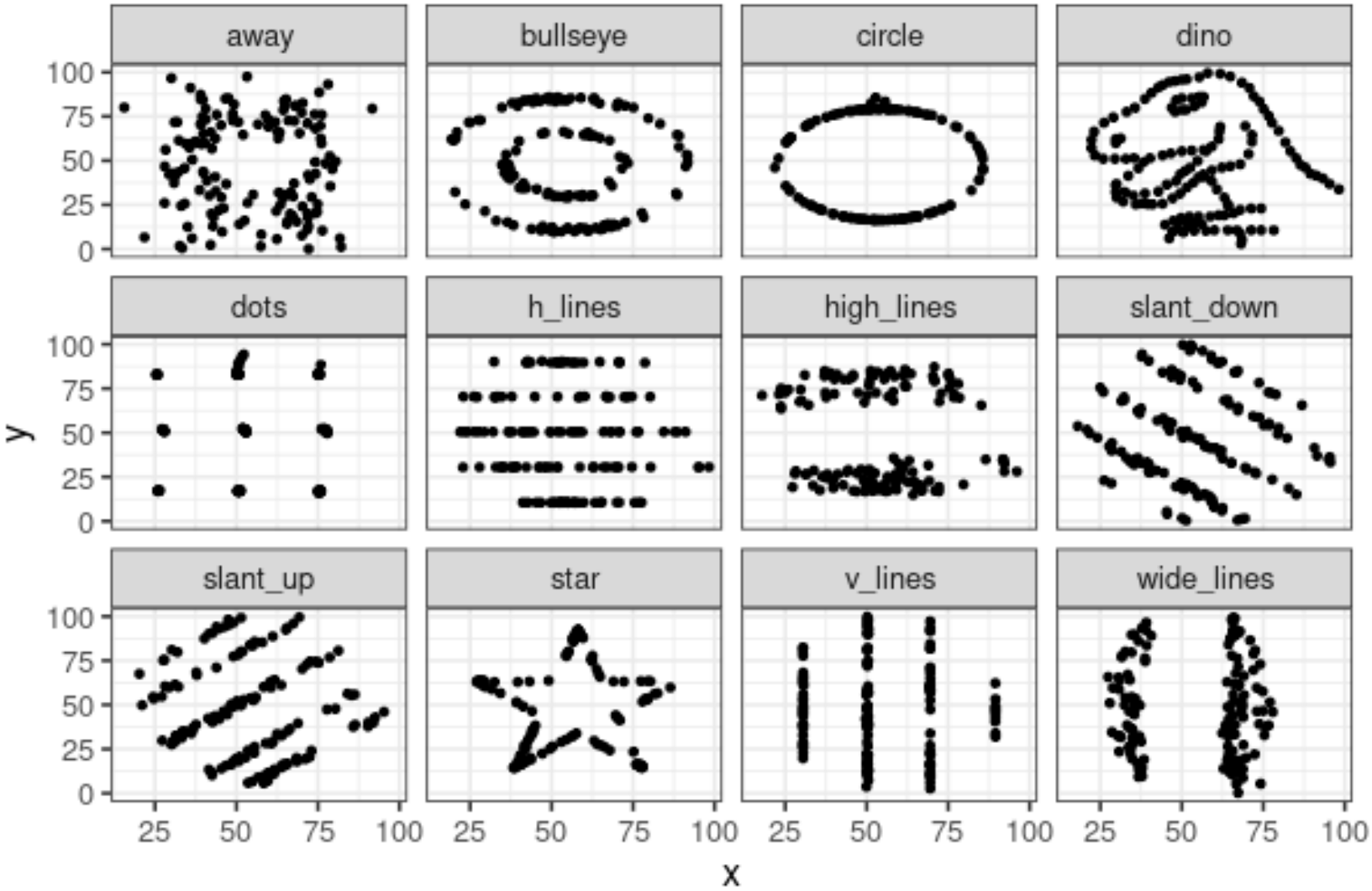


Spearman Correlation

Spearman's correlation is more robust to outliers



"Datasauraus Dozen"



Ecological Correlation

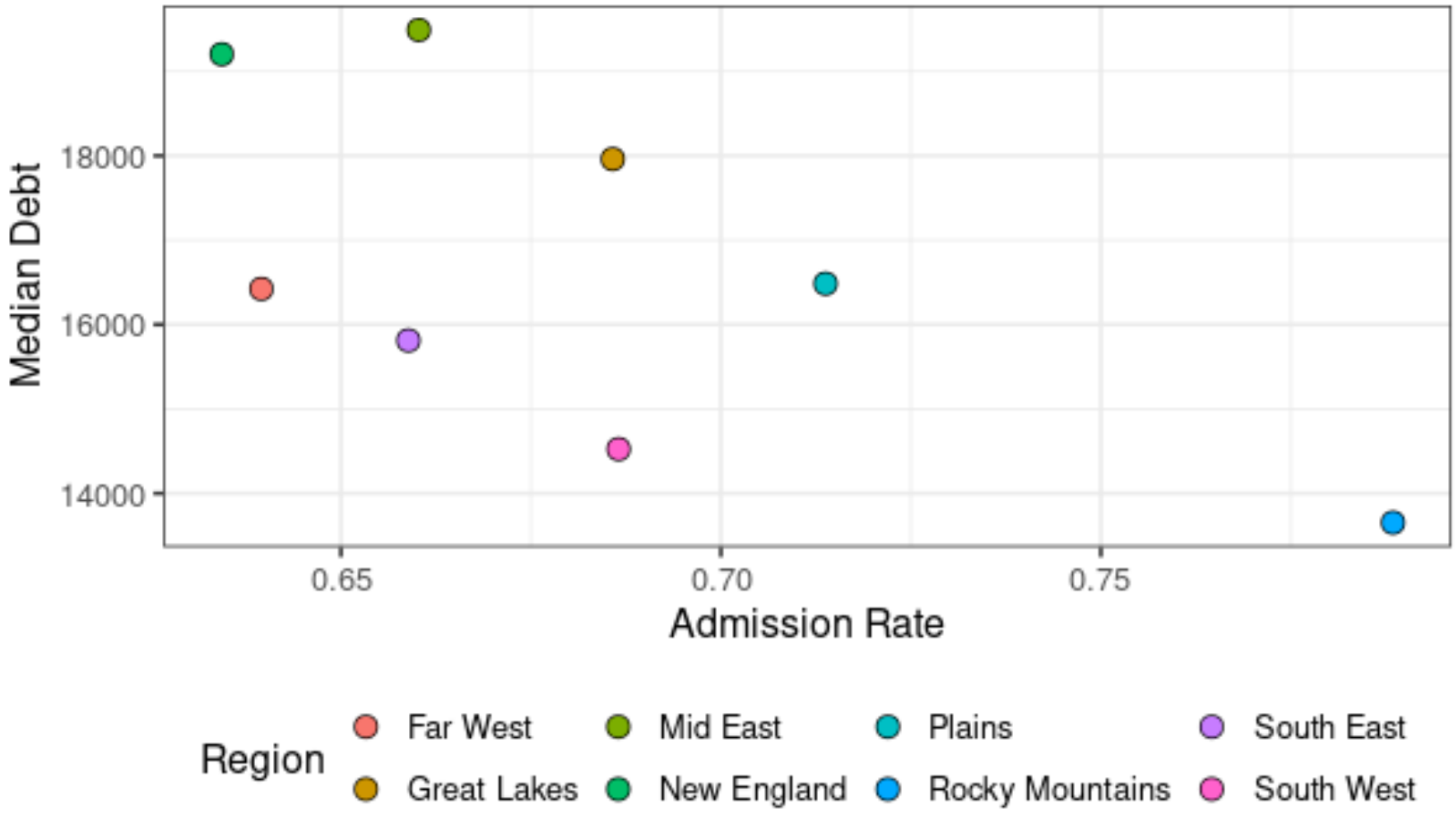
Ecological correlations compare variables for data that have been aggregated at an ecological level

- 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

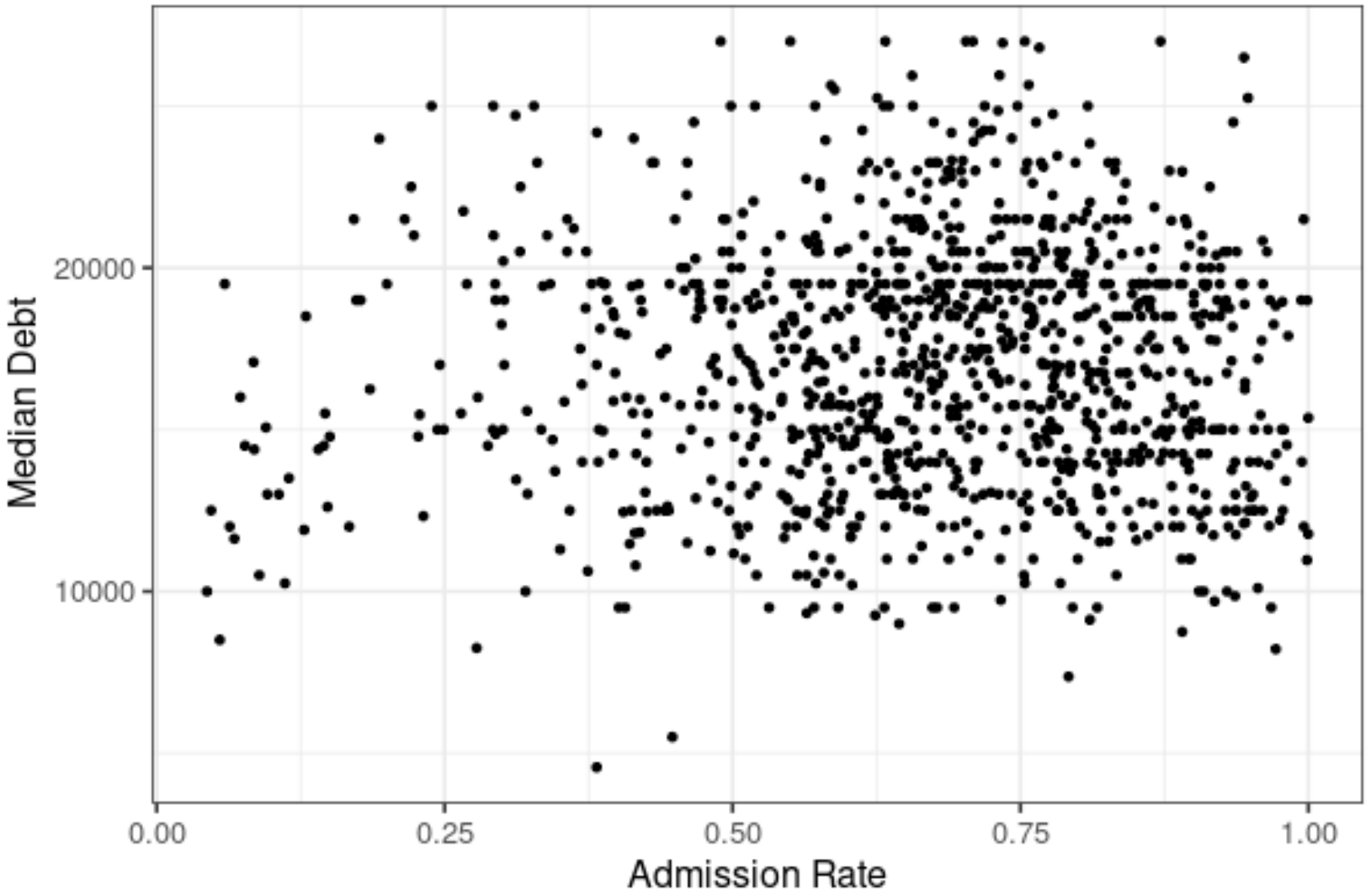
College Ecological Fallacy

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

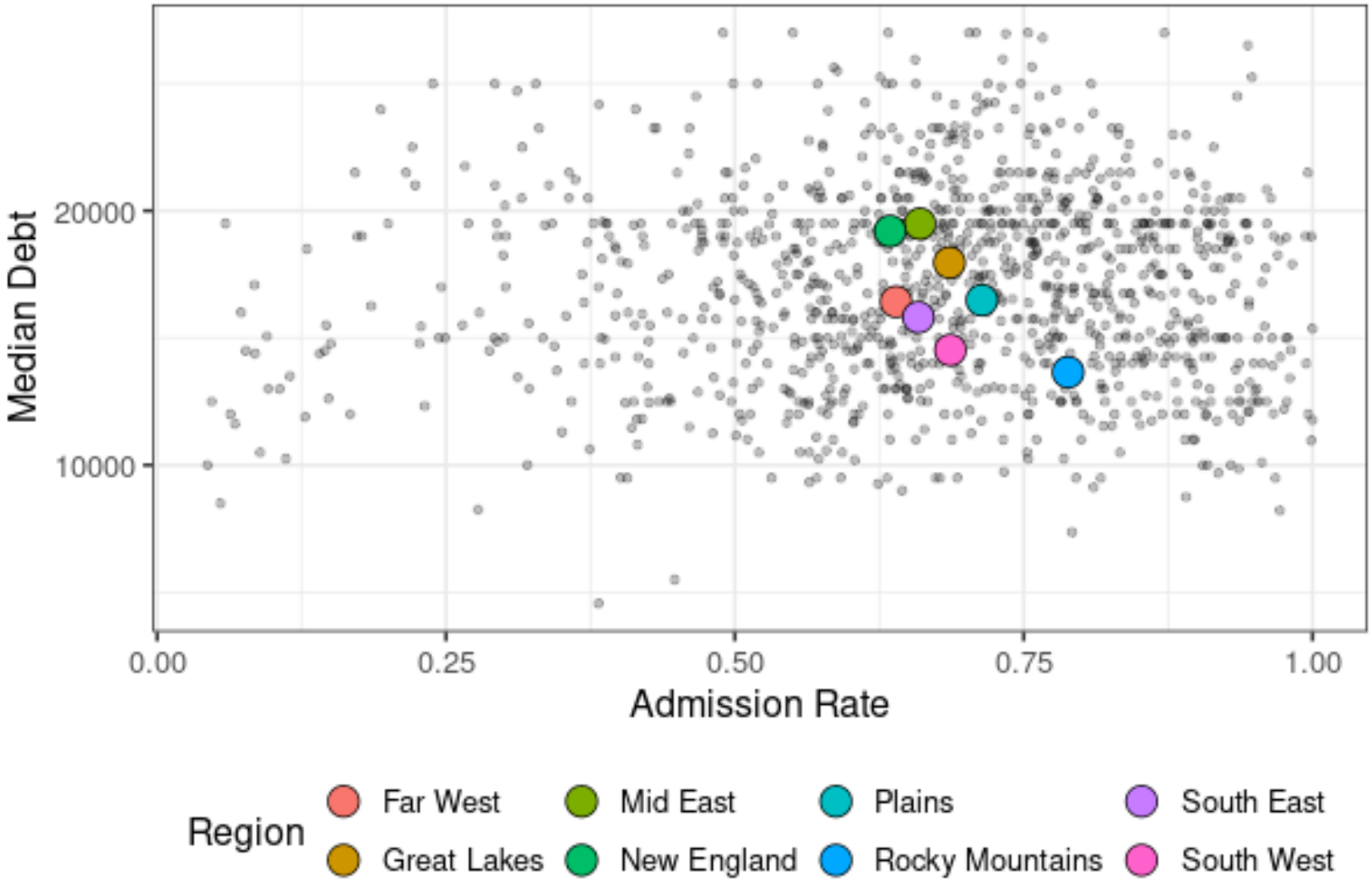


College Ecological Fallacy

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



College Ecological Fallacy



Illiteracy (1930s Census data)

Correlation between illiteracy and % foreign born is $r = -0.46$

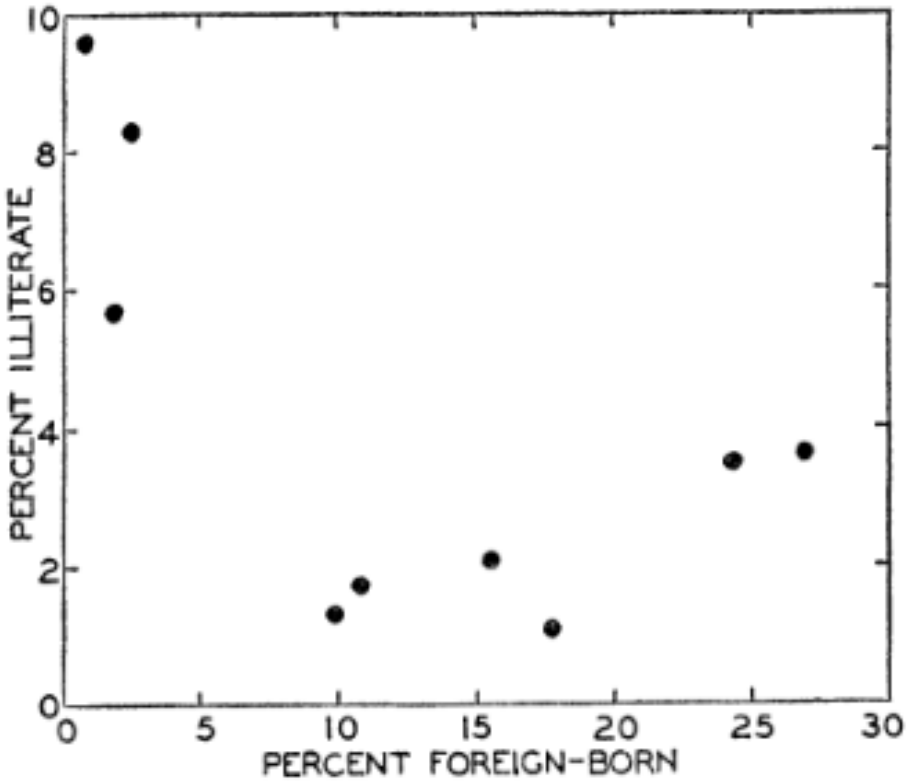
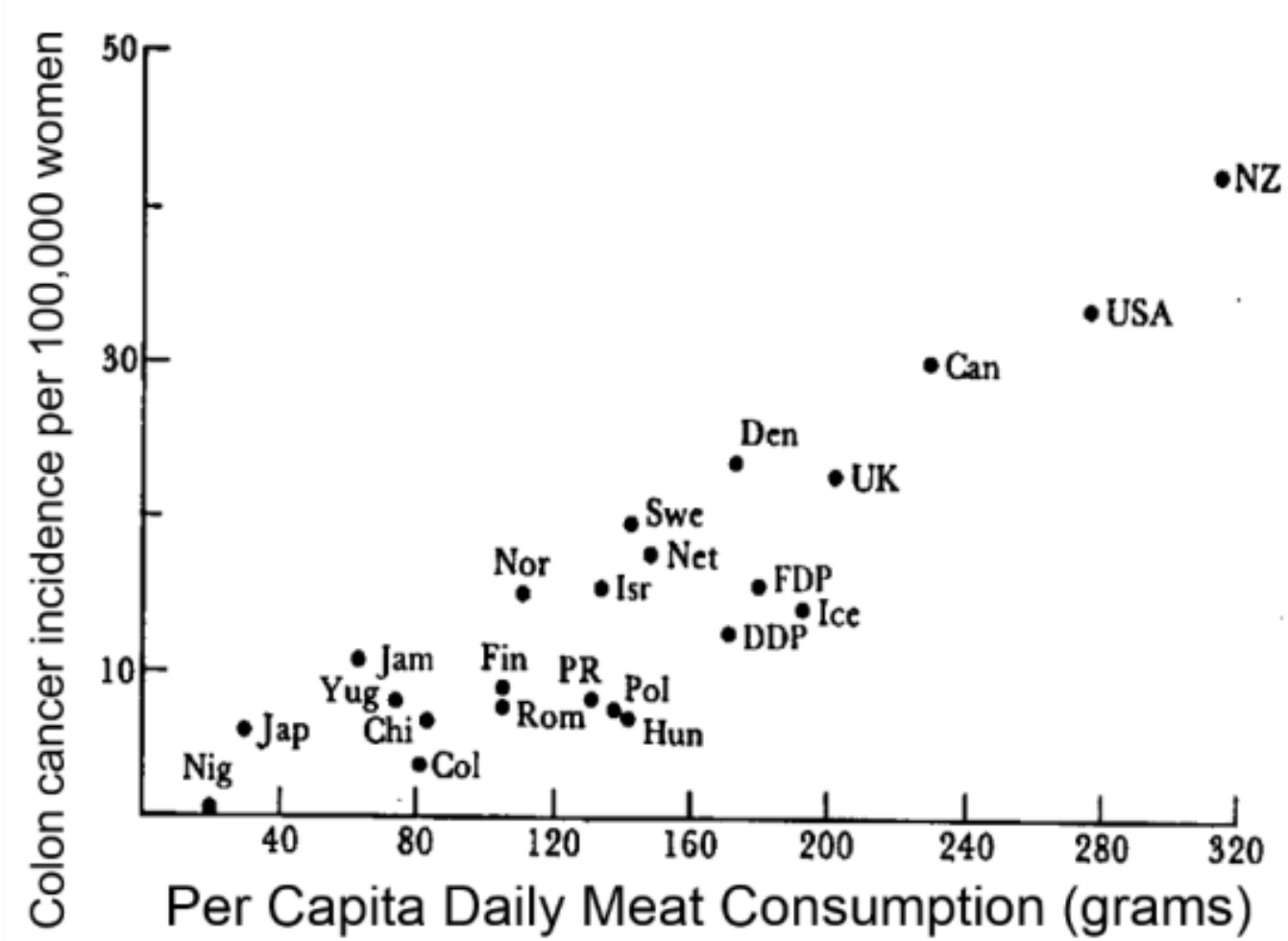


FIG. 3

Meat Consumption



Review

- **Z-scores** or **standardized scores** give us a way to relate the value of an obseration to the mean/standard deviation of a group
- **Pearson's correlation** strength of *linear association*
  - 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** – you should never infer beyond the specific data that you have at hand