Correlation

Grinnell College

February 11, 2024

Review

- Measures of centrality
- Measures of spread
- Robust statistics
- Conditional Tables
- Standardization

Pearson's Height Data

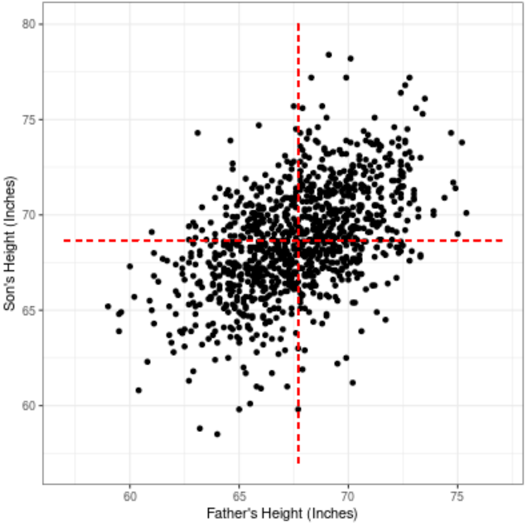
In the 1880's the scientific community was enthralled with the idea of quantifying heritable traits

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

Does height appear to be heritable?



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**:

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

As before, $\overline{x}$ and $\overline{y}$ are the mean values of the quantiative variables X and Y. Similarly, s_x and s_y are their standard deviations

z-scores and corrleation

Recall our previous discussion of z-scores and standardization

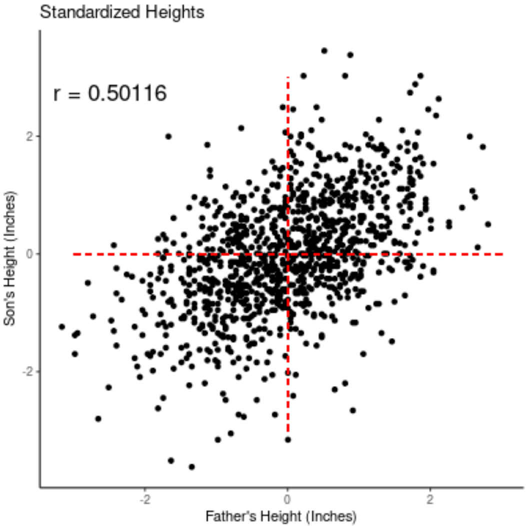
$$z_i = \frac{x_i - \overline{x}}{s_x}$$

And observe the relationship with the correlation coefficient:

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

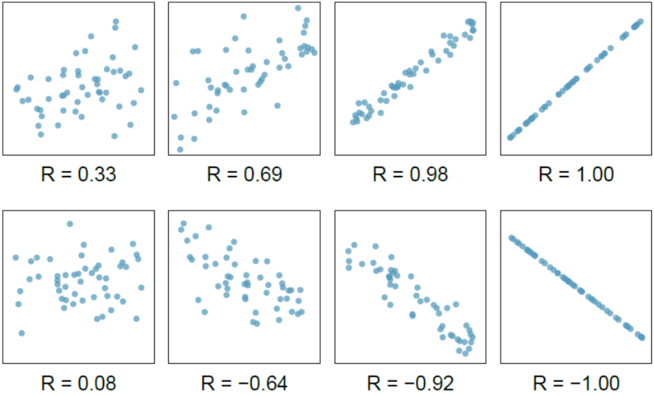
If above-average values of X are common among cases with above-average values of Y (or vice-versa), we should expect r to be positive

Height Data

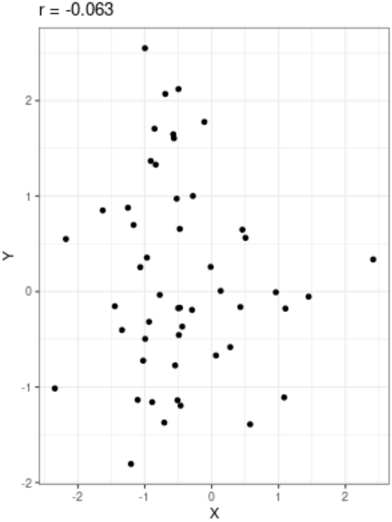

Standardized Heights
r = 0.50116
2
0
-2
Son's Height (Inches)
-2
0
2
Father's Height (Inches)

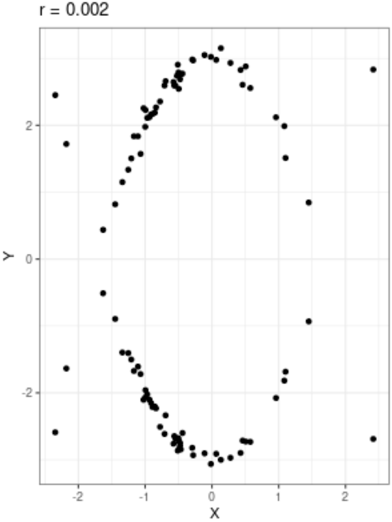
Correlation Examples

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



Correlation Examples





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/

Non-linear Assocation

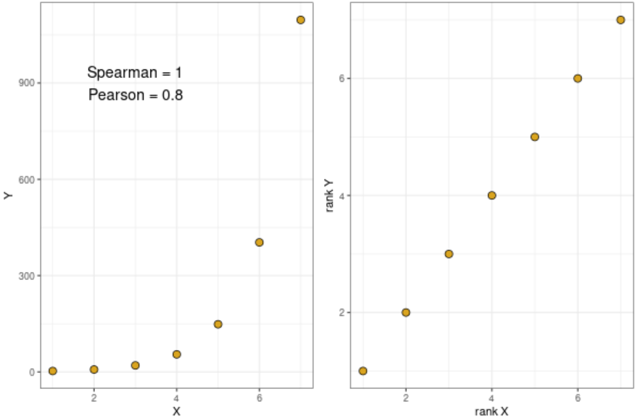
In addition to Pearson, we have **Spearman's rank correlation** (denoted ρ) where the values of X and Y are replaced with their rank order from smallest to largeset:

$$\begin{matrix} X = \{2, 4, 6, 10, 8\} \\ Y = \{7, 4, 1, 5, 3\} \end{matrix} \quad \Longrightarrow \quad \begin{matrix} X_{rank} = \{1, 2, 3, 5, 4\} \\ Y_{rank} = \{5, 3, 1, 4, 2\} \end{matrix}$$

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

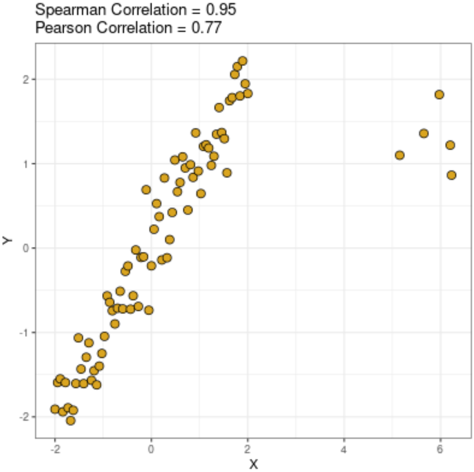
Non-linear Assocation

$$y = e^x$$

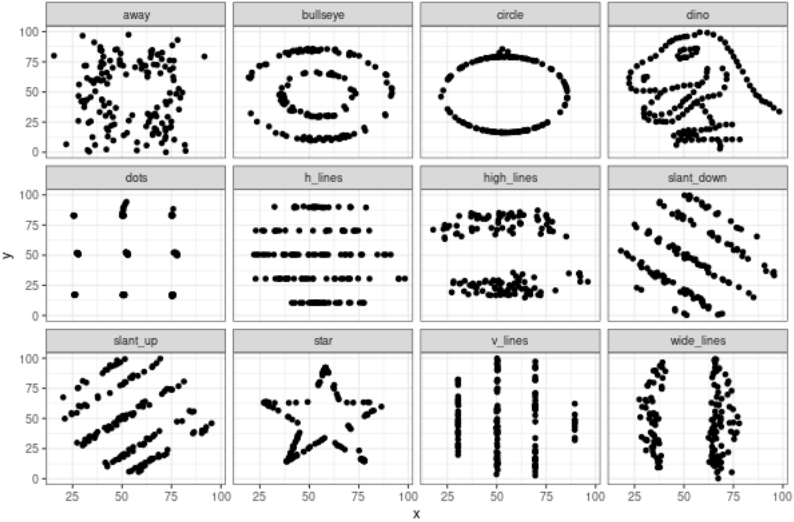


Spearman Correlation

Spearman's correlation is more robust to outliers



“Datasauraus Dozen”



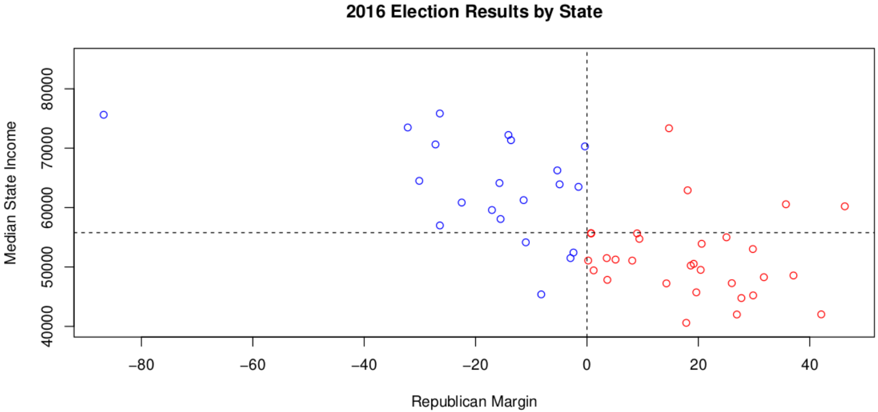
Ecological Correlation

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

- Countries
- States
- Schools

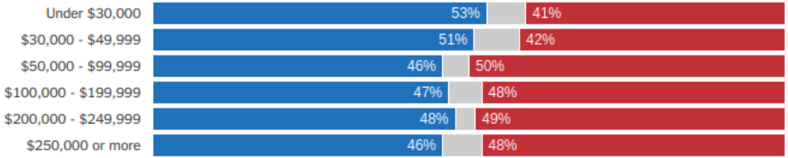
Ecological Correlations

Looking at the relationship between median state income and 2016 election results gives a correlation coefficient of $r = -0.63$



Ecological Correlations

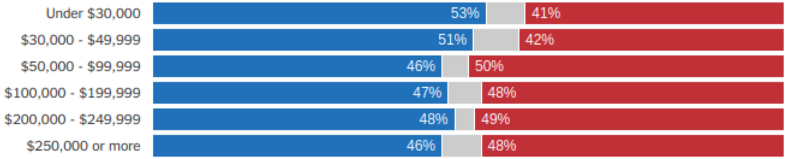
Using 2016 exit polls conducted by the NY Times, we can get a sense of party vote and income *at the individual level*



- Looking at individuals as cases *instead* of states, we see the opposite relationship

Ecological Correlations

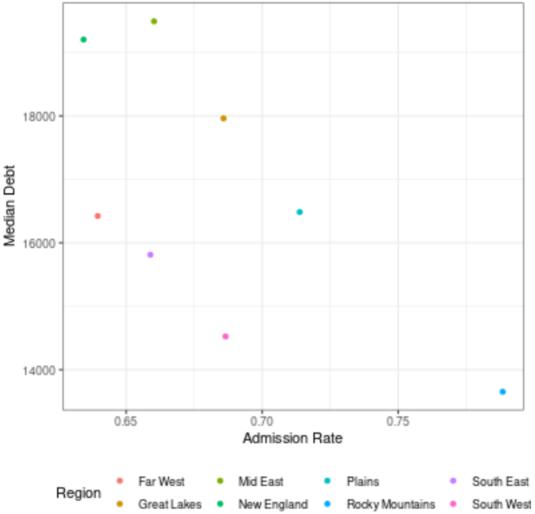
Using 2016 exit polls conducted by the NY Times, we can get a sense of party vote and income *at the individual level*



- Looking at individuals as cases *instead* of states, we see the opposite relationship
- This "reversal" is an example of the **ecological fallacy**
  - Inferences about individuals cannot *necessarily* be deduced from inferences about the groups they belong to

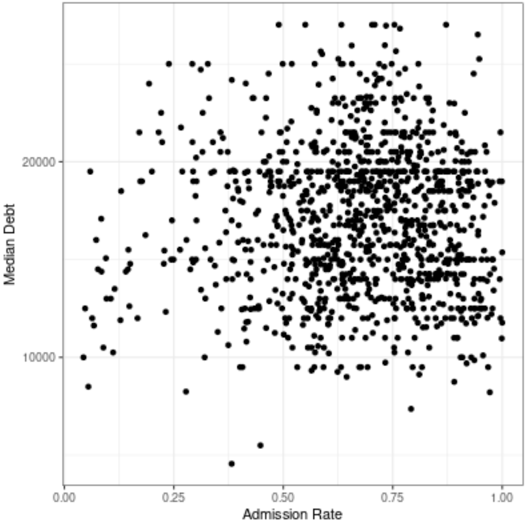
College Ecological Fallacy

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

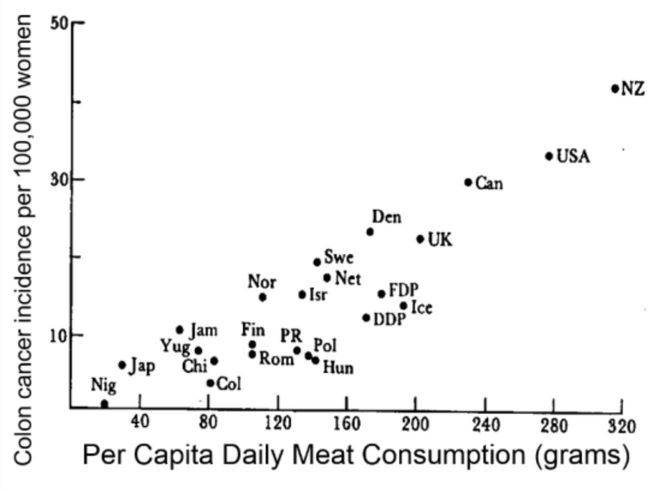


College Ecological Fallacy

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



Meat Consumption



Illiteracy (1930s Census data)

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

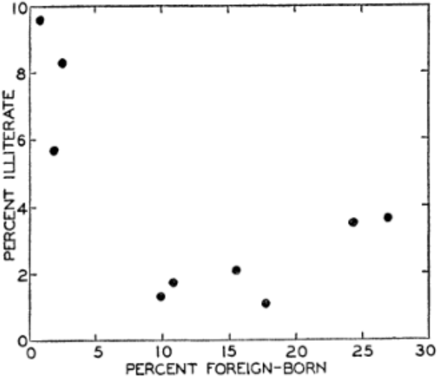


FIG. 3

Review

- **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