Correlation

Ryan Miller

Grinnell College
Statistics

Introduction

- Our previous lecture introduced methods for numerically describing associations between *two categorical variables*
  - These included differences or ratios of conditional proportions, and odds ratios
- Today we'll introduce methods for numerically describing associations between *two quantitative variables*

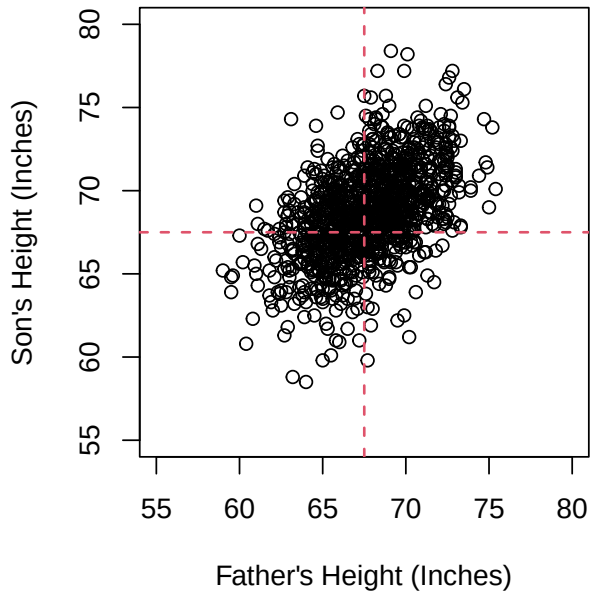
Pearson's Heights

- In the 1880s, the scientific community was fascinated by the idea of quantifying heritable traits
  - Karl Pearson, a now famous statistician, collected data on the heights (inches) of 1,078 fathers and their fully-grown first-born sons:

Father	Son
65	59.8
63.3	63.2
65	63.3
65.8	62.8
...	...

Pearson's Heights

Here are Pearson's height data on a scatter plot. The red lines are the mean of each variable.



Does height appear to be heritable?

Pearson's Correlation Coefficient

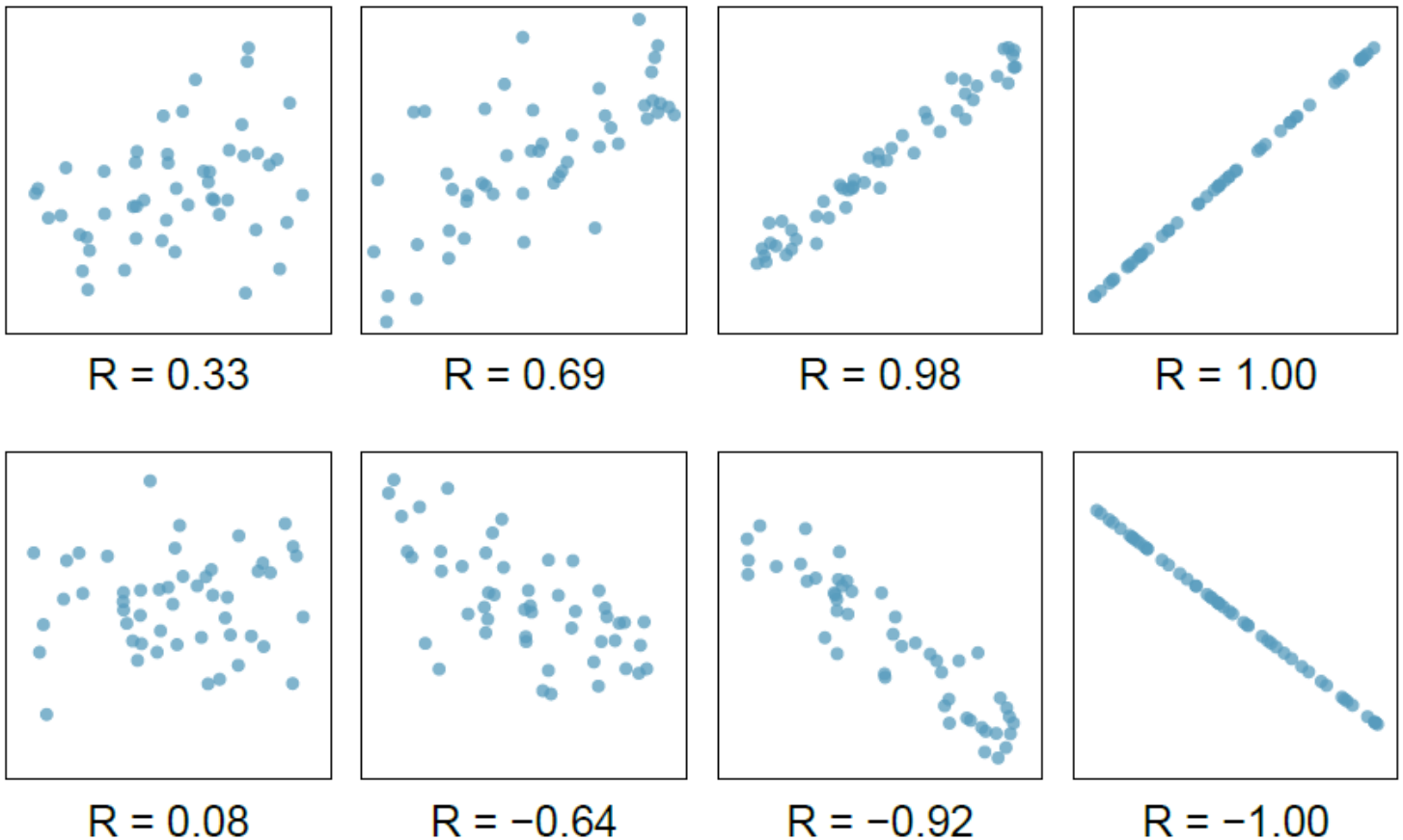
- Adult heights of fathers and sons are clearly associated, but Pearson wanted to *quantify* how strongly they are associated
  - Building upon an idea from the French scientist Francis Galton, he developed **Pearson's correlation coefficient**:

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

- Here, $\overline{x}$ and $\overline{y}$ are the mean values of two quantitative variables, X and Y
  - s_x and s_y are the standard deviations of these variables

Examples

Pearson's correlation, r, quantifies the *strength of linear association* between two quantitative variables



Correlation vs. Strength of Association

Whether a correlation is considered "strong" or "weak" depends upon the field:

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/

Z Scores

Pearson's correlation coefficient doesn't depend upon the units or context of the data due to its use of **z-scores**, or standardized representations of individual data-points:

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

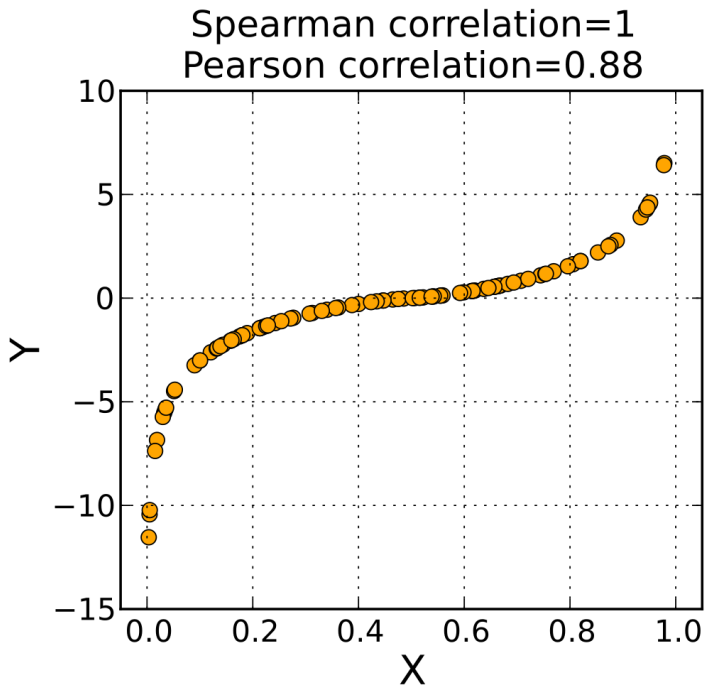
These Z-scores, $z_{x_i} = \frac{x_i - \overline{x}}{s_x}$, reflect the standardized difference between an observed value, x_i, and the mean of the corresponding variable, $\overline{x}$.

Z Scores

- In Pearson's data, sons had an average height of $\overline{x} = 63.3$ inches with a standard deviation of $s_x = 2.8$
  - So, we could describe a son who measured 68.7 inches as being 5.4 inches above average
  - Or we could use the z-score: $z = \frac{68.7 - 63.3}{2.8} = 1.9$, meaning they are 1.9 standard deviations above average
- Z-scores are most useful when data exist on different measurement scales or involve highly-specialized units

Nonlinear Relationships

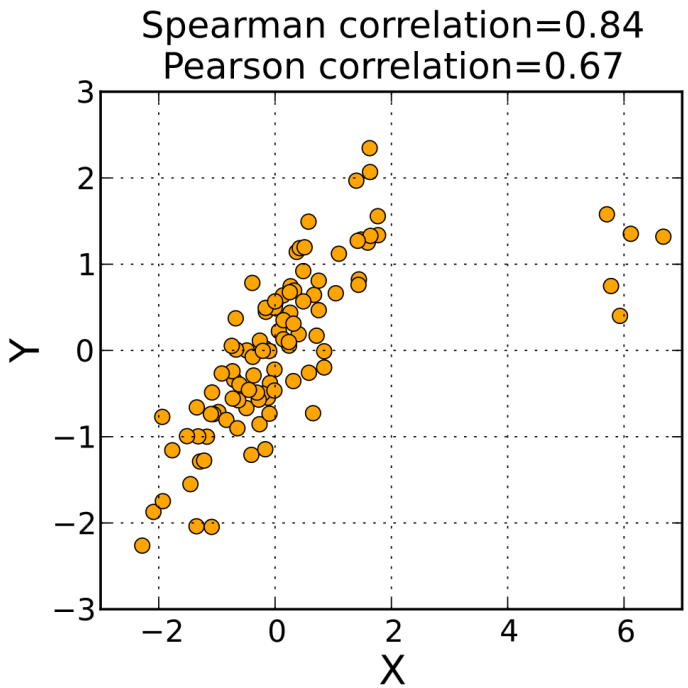
Spearman's rank correlation is an alternative that is suitable for quantifying the strength of non-linear associations:



The values of X and Y are separately ranked from 1 to n and these ranks are used as variables in the correlation calculation.

Spearman's Rank Correlation

Spearman's rank correlation is also more *robust* to outliers:



However, a downside of Spearman's correlation (and Pearson's correlation too) is that it only captures *monotonic* associations

Common Misconceptions

From Cook & Swayne's *Interactive and Dynamic Graphics for Data Analysis*:

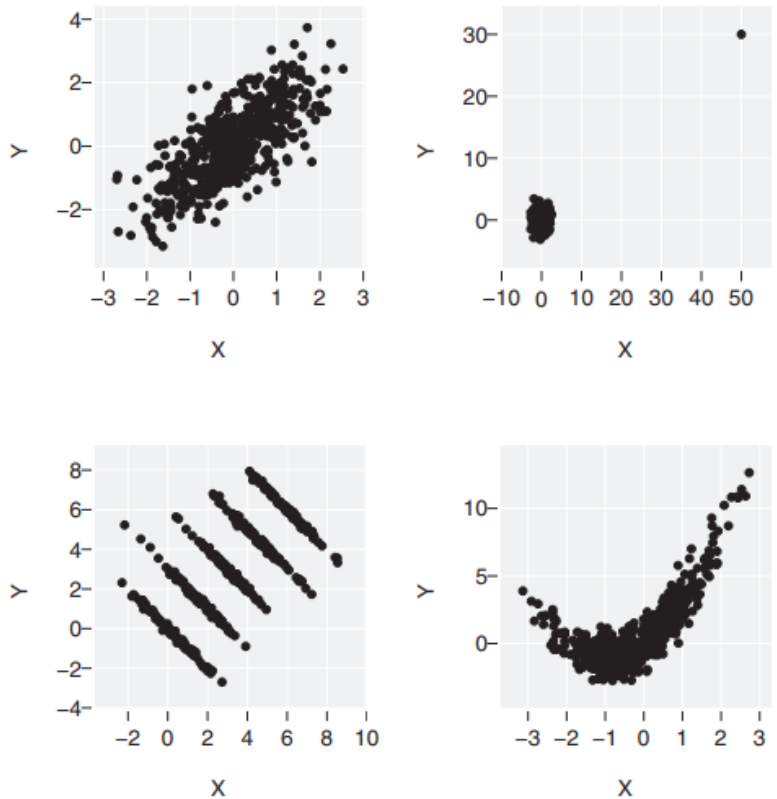
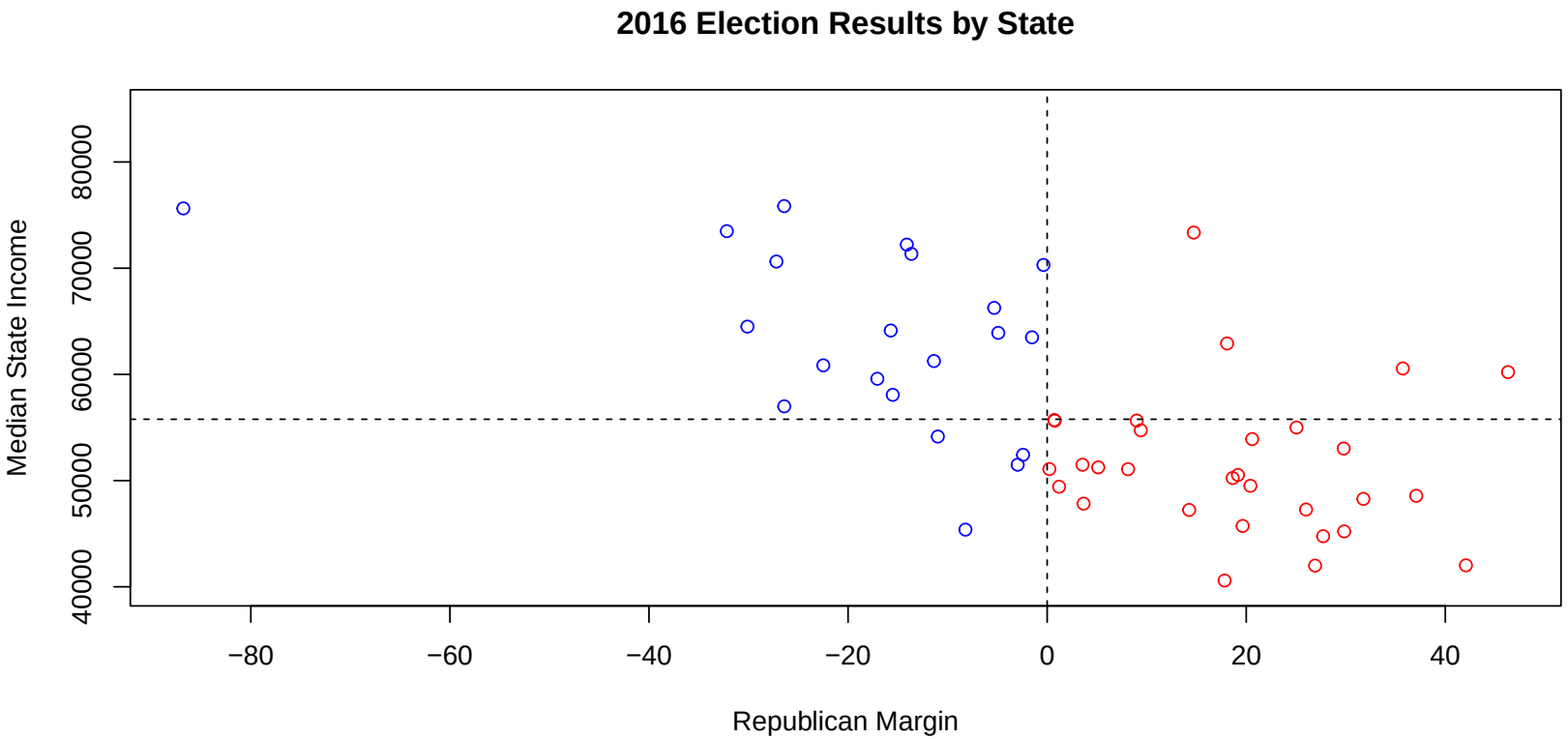


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

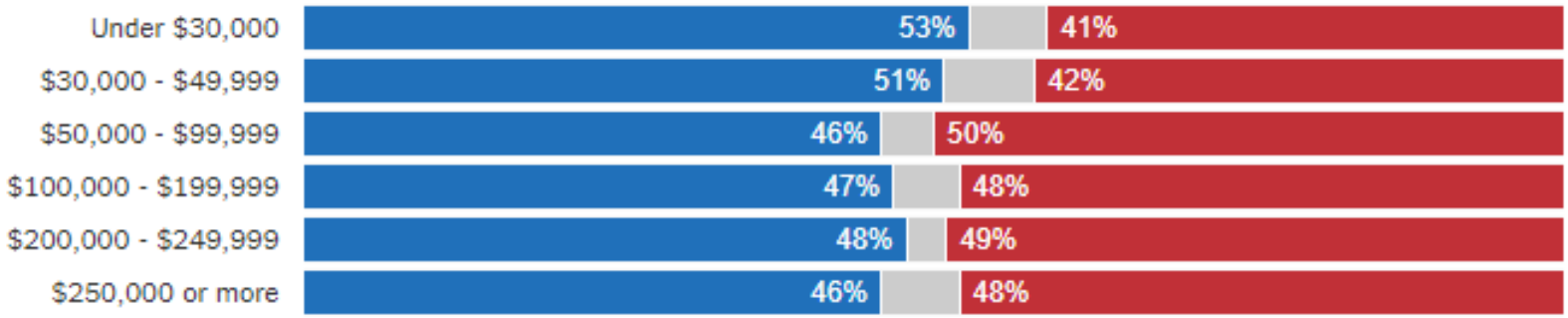
Common Misconceptions



- $r = -.63$, so do republicans earn lower incomes than democrats?

The Ecological Fallacy

Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- Looking at individuals as cases there is an opposite relationship between political party and income
- This "reversal" is an example of the **ecological fallacy**
  - Inferences about individuals cannot necessarily be deduced from inferences about the groups they belong to

Conclusion

- **Pearson's correlation coefficient** is a common way to measure the strength of linear association
  - Correlation is the *average product of z-scores*
- You may opt for **Spearman's rank correlation** if your data contain outliers or non-linear (but monotonic) relationships
- Be careful when interpreting **ecological correlations**, you need to carefully consider how a case is defined in your data, particularly when aggregation is involved