Correlation

Ryan Miller

Grinnell College
Statistics

Introduction

So far, we've discussed *one-sample* and *two-sample* hypothesis tests, with both *categorical* and *quantitative* outcomes

- Two-sample tests always involve a binary categorical variable that divides our data into two groups (samples)
  - One-sample tests only involve the outcome variable
- We default to the Z-test when the outcome variable is categorical (expressed using proportions)
- We default to the T-test when the outcome variable is quantitative (expressed using means)

You might notice these tests span every combination of two variables with the exception of *two quantitative* variables

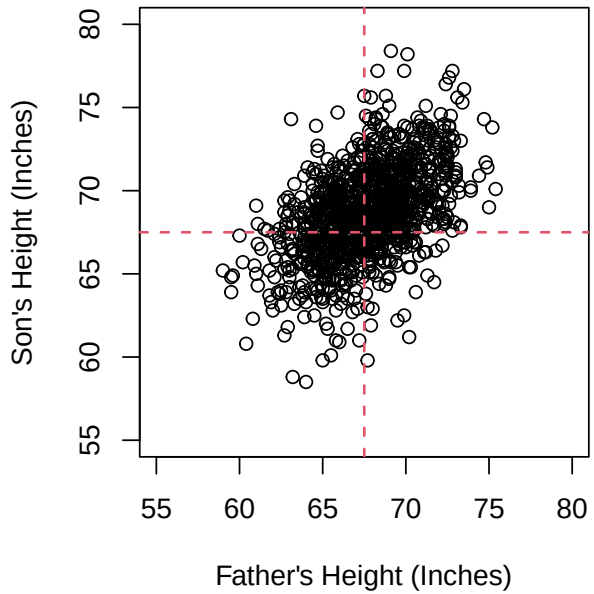
Pearson's Height Data

- 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 Height Data (cont.)

A **scatter plot** is used to graph two quantitative variables. Here the red lines are each variable's mean.



Does height appear to be heritable?

Pearson's Correlation Coefficient

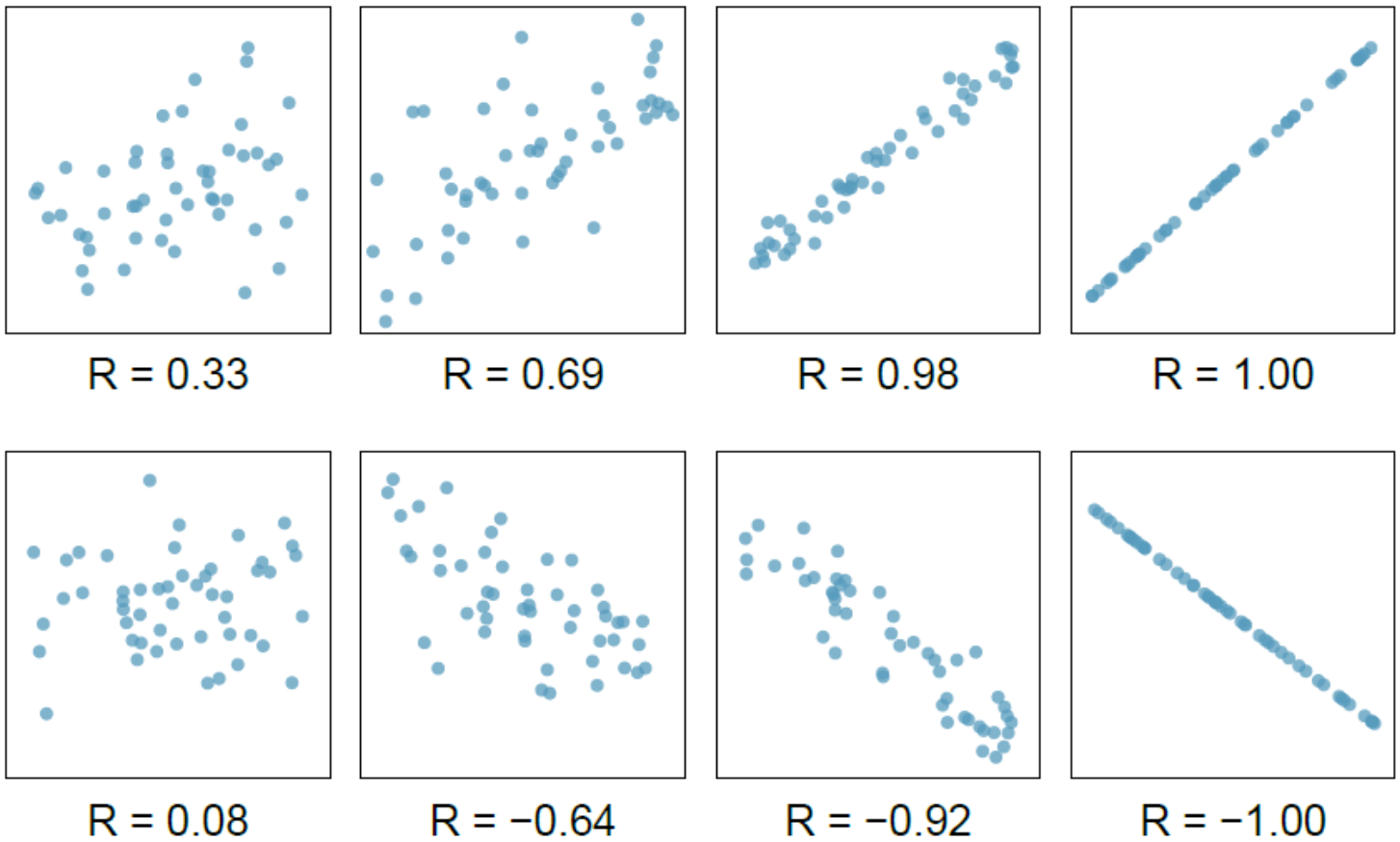
- The scatter plot clearly shows that father and son heights are related, but Pearson wanted to *quantify* the *strength of the relationship*
  - Building upon an idea from the French scientist Francis Galton, he developed **Pearson's correlation coefficient**:

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

- Here, $\overline{x}$ and $\overline{y}$ are the means of each variable, X and Y
  - s_x and s_y are the standard deviations of these variables

Correlation and Strength of Linear Association

Pearson's correlation, r, is a standardized measure of the *strength of linear association* between two quantitative variables:



Correlation and Strength of Linear Association (cont.)

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/

Correlation and Z-Scores

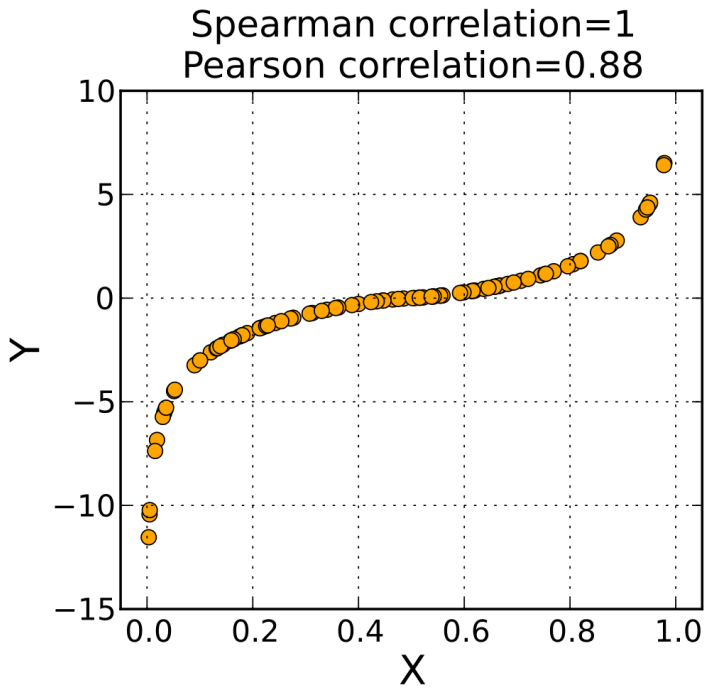
You might notice the Z-score transformation being used in Pearson's correlation (ie: $z_i = \frac{x_i - \overline{x}}{s_x}$:

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

- Each Z-score reflecst the standardized difference between an observed value, x_i, and the mean of the corresponding variable, $\overline{x}$.
- Pearson's use of the Z-score transformation allows the correlation coefficient to be standardized and unitless.

Nonlinear Relationships

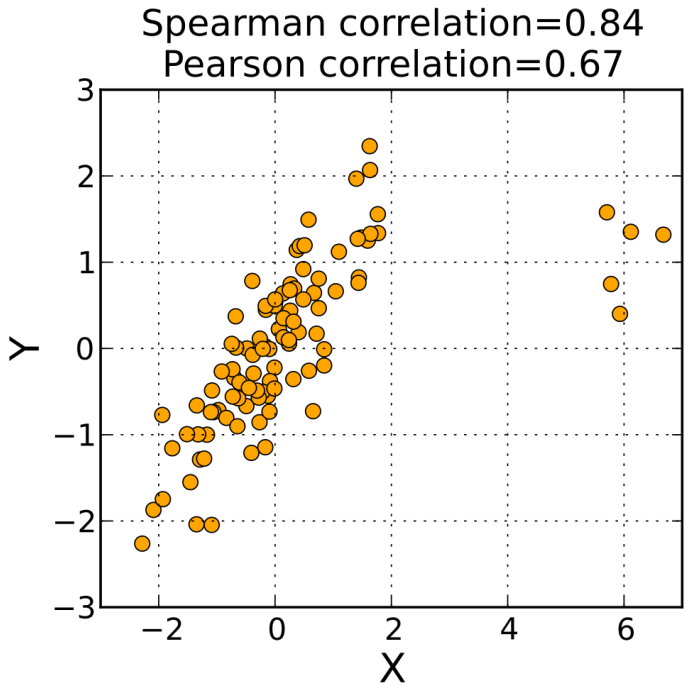
Spearman's rank correlation is an alternative that is suitable for quantifying the strength of non-linear associations:



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*:

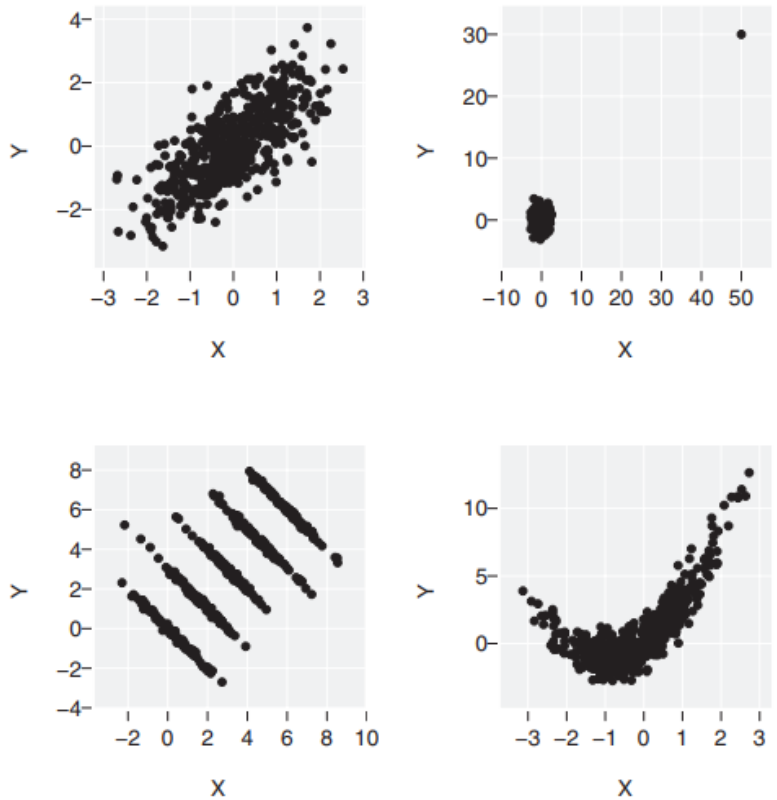
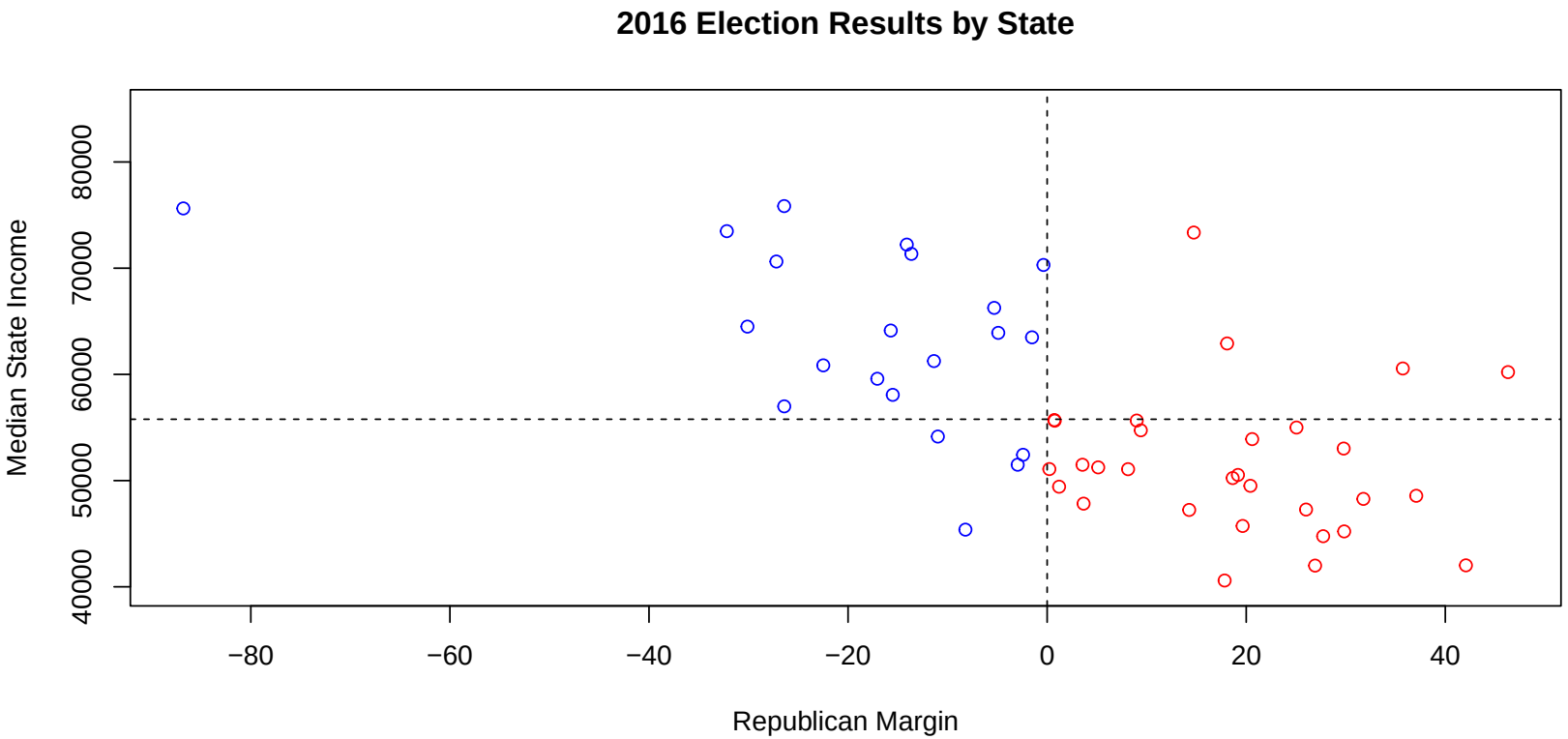


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

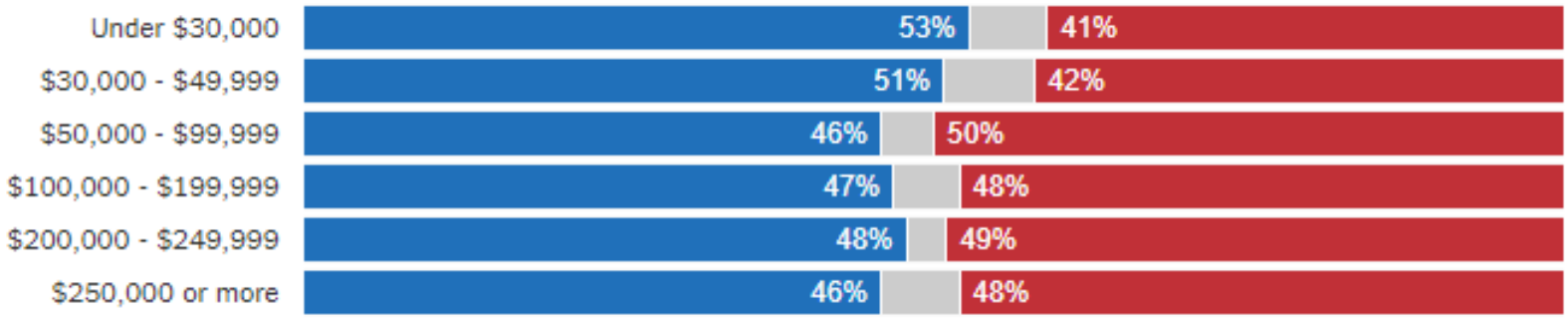
Common Misconceptions



- $r = -.63$, 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
- Today's lab will cover hypothesis tests involving correlation