Descriptive Statistics

Part 3 - Correlation

Ryan Miller

Grinnell College
Statistics

Introduction

So far we've discussed descriptive statistics for the following scenarios:

- Univariate (summarizing the distribution of a single variable)
  - *One categorical variable* - one-way tables of frequencies or proportions
  - *One quantitative variable* - mean and median (center), standard deviation, IQR, and range (spread)
- Bivariate (summarizing the association between two variables)
  - *Two categorical variables* - two-way tables, conditional proportions, risk difference, relative risk, and odds ratio
  - *One categorical and one quantitative variable* - differences in conditional means (or medians) across groups

Today we'll cover the final bivariate scenario - two quantitative variables

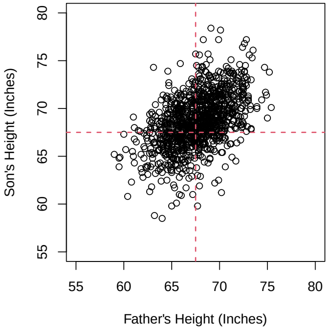
Pearson's height data

- 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

Here are Pearson's height data on a scatter plot:



Does height appear to be heritable?

Pearson's Correlation Coefficient

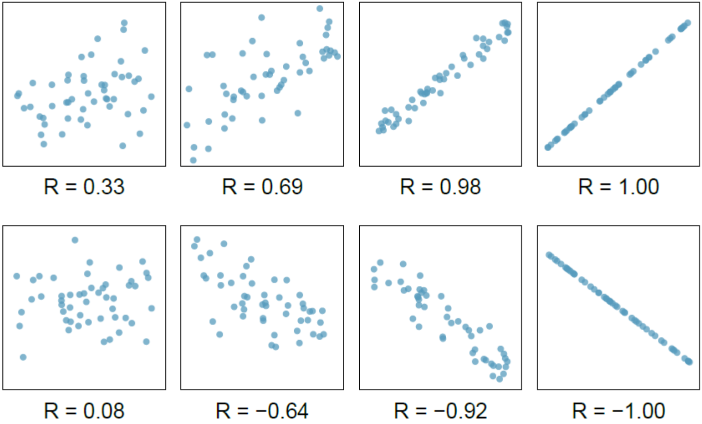
- The adult heights of fathers and their sons are clearly associated, but Pearson wanted to *quantify* how strongly they were associated
  - Building upon an idea from the French scientist Francis Galton, Person developed **Pearson's correlation coefficient**:

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

- 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

Correlation examples

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



Grinnell College
Statistics

What is a "strong" correlation?

Whether a correlation is considered "strong" or "weak" depends upon your 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/

Digression - Z-scores

Standardization using **z-scores** is a common approach statisticians use to analyze variables that are measured on very different scales:

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

- In Pearson's data, sons had an average height of $\overline{x} = 63.3$ inches with a standard deviation of $s = 2.8$
  - So, we could describe a son who measured 68.7 inches as 5.4 inches above average
  - We could also describe them with the z-score: $z = \frac{68.7 - 63.3}{2.8} = 1.9$, meaning they are 1.9 standard deviations above average

Digression - Z-scores

A practical advantage of standardization is that it makes variables more interpretable by non-experts

- If you were told that your blood urea concentration is 50 mg/dL you'd likely have no idea what to think
  - However, if you were told this is 4 standard deviations above average you'd quickly realize your blood urea is unusually high

Correlation and Z-scores

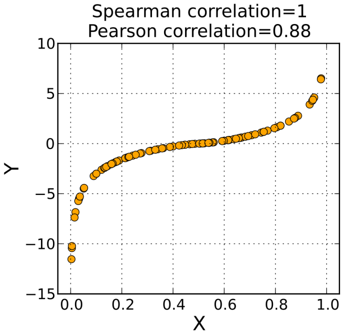
Now that we've defined z-scores, you should notice a connection with Pearson's correlation coefficient:

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

- Thus, correlation is just the average product of z-scores within a data set
  - So, if above-average values of X (positive z-scores) are common among cases with above-average values of Y we expect r to be positive

Non-linear correlation?

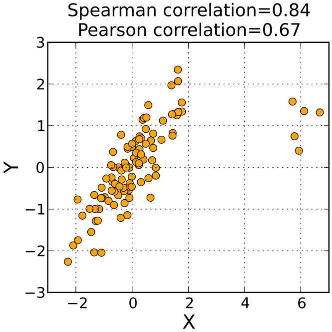
Spearman's rank correlation is an alternative that is suitable for quantifying the strength of non-linear associations



The values of X and Y are each ranked from 1 to n and these ranks are used to calculate correlation

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

Grinnell College
Statistics

Common mistakes and misconceptions

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

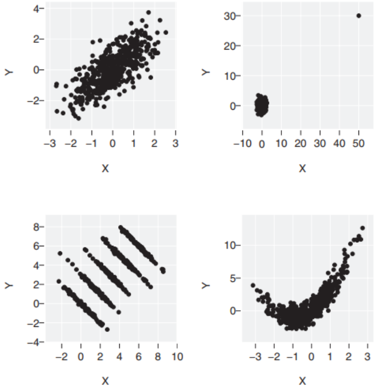
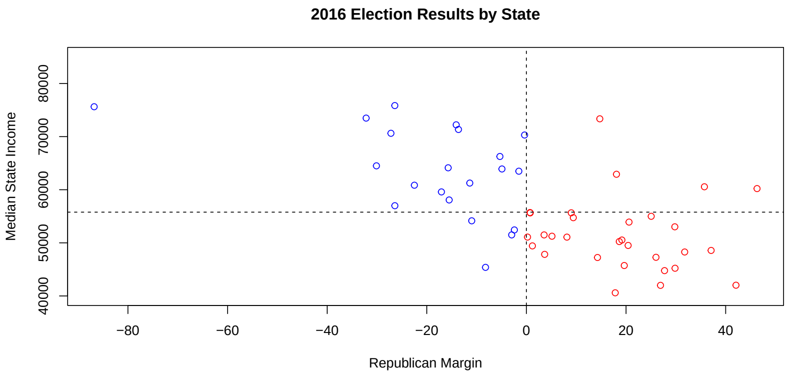


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

Ecological correlations

- **Ecological correlations** compare variables at an ecological level (ie: The cases are aggregated data - like countries or states)
  - There's nothing inherently bad about this type of analysis, but the results are often misconstrued
- Let's look at the correlation between a US state's median household income and how that state voted in the 2016 presidential election

Ecological correlations

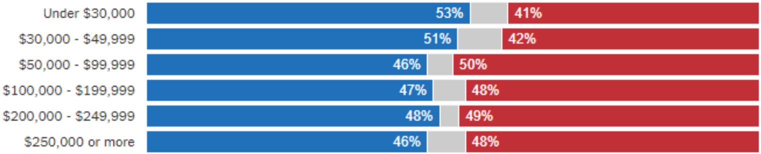


- $r = -.63$, so do republicans earn lower incomes than democrats?

Grinnell College
Statistics

The ecological fallacy

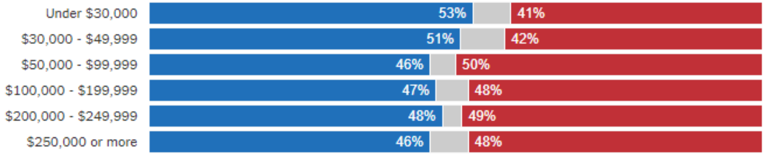
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- Looking at individuals as cases there is an opposite relationship between political party and income

The ecological fallacy

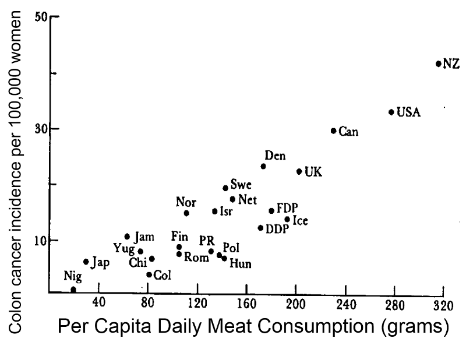
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



- 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

Practice

1) Describe the association (form, strength, and direction) and estimate the correlation coefficient
2) Explain how the ecological fallacy might impact the conclusion most people are tempted to draw from this graph



Pratice (solution)

1) There is a strong, positive, and approximately linear relationship between a country's meat consumption and its colon cancer incidence (among women). A reasonable estimate for the correlation might be around 0.8.
2) Most would interpret this graph as *individuals* who eat more meat being more likely to *individually* develop colon cancer. However, that conclusion is not justified by these data alone.

Conclusion

- **Pearson's correlation coefficient** is 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 should never infer beyond the cases that the data are describing