Correlation

Ryan Miller

Outline

1. Scatterplots
   - Describing form, strength, and direction
2. Quantifying strength of association
   - Pearson's correlation coefficient, alternatives
3. Common pitfalls
   - Outliers and non-linear data, ecological correlations

Pearson's height data

- Francis Galton and Karl Pearson, two pioneers of modern statistics, lived in Victorian England at a time when the scientific community was fascinated by the idea of quantifying heritable traits

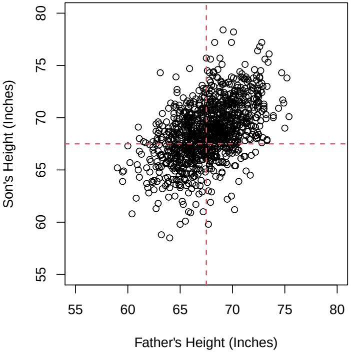
Pearson's height data

- Francis Galton and Karl Pearson, two pioneers of modern statistics, lived in Victorian England at a time when the scientific community was fascinated by the idea of quantifying heritable traits
- Wondering if height is heritable, they measured the heights 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
...	...

Scatterplots

A scatterplot can be used to visually identify whether these variables are related:



So, do you think there's an association between the height of a father and son?

Associations between quantitative variables

Using a scatterplot, we can qualitatively describe an association in terms of the following factors:

1) **Form** - what type of trend or pattern do the data seem to follow (ie: linear, logarithmic, exponential, etc.)
2) **Strength** - how closely or tightly do the individual data-points follow that trend or pattern
3) **Direction** - do larger values of the "X" variable tend to correspond with larger values of the "Y" variable (positive) or do they correspond with smaller values (negative)

Associations between quantitative variables

Using a scatterplot, we can qualitatively describe an association in terms of the following factors:

1) **Form** - what type of trend or pattern do the data seem to follow (ie: linear, logarithmic, exponential, etc.)
2) **Strength** - how closely or tightly do the individual data-points follow that trend or pattern
3) **Direction** - do larger values of the "X" variable tend to correspond with larger values of the "Y" variable (positive) or do they correspond with smaller values (negative)

Note: For some non-linear forms, it doesn't make sense to use "positive" and "negative" to describe direction.

Quantifying strength (linear associations)

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- Pearson's correlation coefficient, r, measures the strength of a *linear association* between X and Y

$$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Quantifying strength (linear associations)

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- Pearson's correlation coefficient, r, measures the strength of a *linear association* between X and Y

$$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- As you can see, when *above average* values in X are accompanied by *above average* values in Y there is a *positive contribution* to the correlation between X and Y

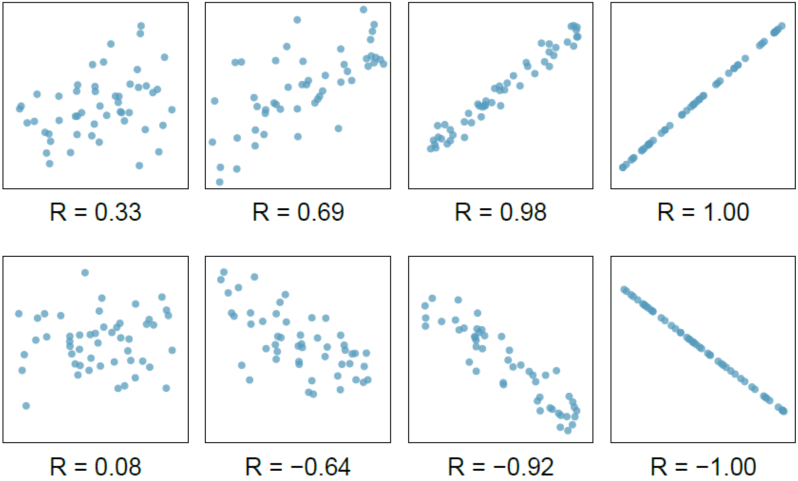
Quantifying strength (linear associations)

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- Pearson's correlation coefficient, r, measures the strength of a *linear association* between X and Y

$$r_{xy} = \frac{1}{n-1} \sum_i \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- As you can see, when *above average* values in X are accompanied by *above average* values in Y there is a *positive contribution* to the correlation between X and Y
- When *above average* values in X are accompanied by *below average* values in Y there is a *negative contribution* to the correlation between X and Y

Correlation examples



What is a “strong” correlation?

Whether a correlation is considered “strong” or “weak” can depend on your discipline:

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/

Practice

Load the College19 Complete Dataset (available on our website) into StatKey, then describe the *form*, *strength*, and *direction* in the following scatterplots:

1) X = Adm_rate, Y = Net_Tuition
2) X = Enrollment, Y = Avg_Fac_Salary

Practice (solutions)

1) Roughly linear, weak-to-moderate ($r = -0.366$), negative
2) Non-linear (logarithmic), moderate, positive

Quantifying strength (non-linear associations)

Methods for quantifying strength of non-linear association are *beyond the scope of this course*, nevertheless few a listed below (along with brief descriptions) for your reference:

- Spearman's rank correlation - correlates the ordered ranks of each variable (assumes a monotone form)
- Kendall's rank correlation - measures concordance (ie: $+,+$ or -,- pairs, relative to the average, across variables)
- R^2 (coefficient of variation) - a model-based measure of how much variability in an outcome variable can be explained by a function of the explanatory variable

Common mistakes and misconceptions

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

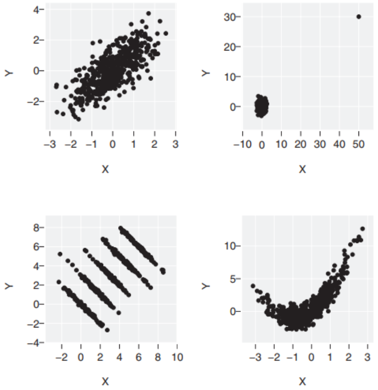
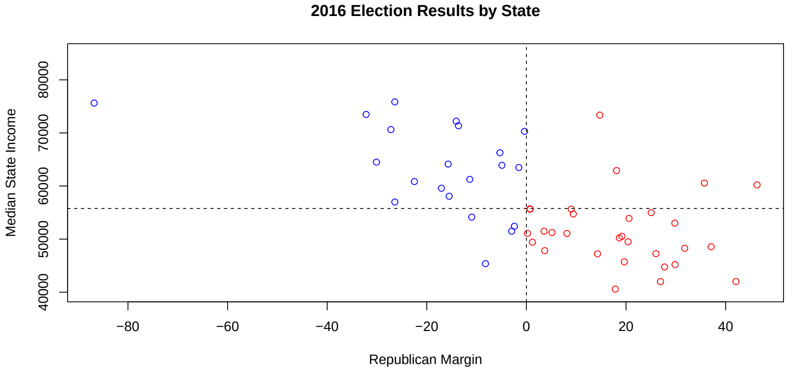


Fig. 6.1. Studying dependence between X and Y. All four pairs of variables have correlation approximately equal to 0.7, but they all have very different patterns. Only the top left plot shows two variables matching a dependence modeled by correlation.

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

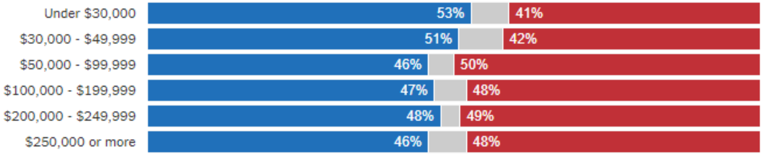
Ecological correlations



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

The ecological fallacy

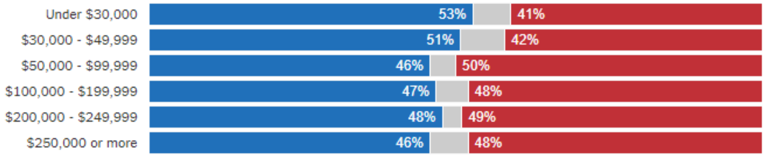
Using 2016 exit polls, conducted by the NY Times (Link), we can get a sense of how party vote and income are related *for individuals*:



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

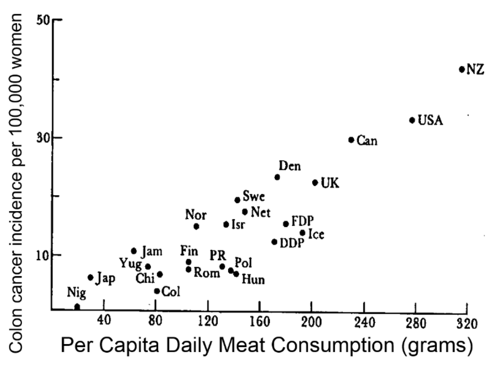
The ecological fallacy

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
  - The lesson here is we should use data where the cases align with who/what we're aiming to describe

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

- **Scatterplots** are used to describe the *form*, *strength*, and *direction* of an association between two quantitative variables
- **Pearson's correlation coefficient** is common way to measure the strength of linear association
  - Avoid relying too heavily on the correlation coefficient when the data contain outliers and non-linear relationships
- Be careful when interpreting **ecological correlations**, you should never infer beyond the cases that the data are describing