Chapter 9: More correlation

Features of r

r has no units

- The first step in computing r is a conversion to standard units
- As a result, original units like inch, pound, dollar, etc disappear. So r has no units.

3 / 19

Change of scale

- The first step in computing r is a conversion to standard units
- Remember that standard units do not change if we change the scale, i.e., if:
  - we add a constant to all entries of a list
  - we multiply all entries of a list by a positive number
- As a result, r is not affected by changes of scale. So r does not change if:
  - we add a constant to all entries of a list
  - we multiply all entries of a list by a positive number
- That is nice! So the correlation between height and weight of people is the same, whether we measure it in inches and pounds, or in centimeters and kilograms.

4 / 19

Changing the order of x and y

- Remember that r is the average of the products of x and y, after conversion to standard units
- Products do not depend on the order of the factors
  (for example: $3 \times 8 = 8 \times 3$)
- As a result, the correlation between x and y is the same as the correlation between y and x

5 / 19

Changing SDs

- The appearance of a scatter diagram depends on the SDs. See overhead (Figure 3 on page 145). These two plots have the same correlation coefficient!
- The first step in computing r is a conversion to standard units. Standard units measure how many SDs a value is away from the average.
- So r measures clustering not in absolute terms, but in relative terms – relative to the SD
- How to compare scatter diagrams?
  - In your mind's eye, draw them as Figures 6 and 7 on pages 127 and 129, so that the vertical and horizontal spread cover the same range as in these figures.
  - Then pick the panel it looks like most. That gives a good guess of r.

6 / 19

When to use r?

- r should only be used for football shaped scatter diagrams
- If the scatter diagram is not football shaped, r can be misleading. See overhead. Be aware of:
  - Outliers:
    - Outliers can throw off the value of r.
    - But don't automatically delete outliers. Investigate what is going on with these values. Example: ozone layer
  - Nonlinear association:
    - r measures linear association
    - If the scatter diagram looks curved, don't use r!

7 / 19

Ecological correlation

8 / 19

Gubernatorial elections

- Last Washington State gubernatorial elections
- Gregoire (Democrat) vs Rossi (Republican)
- Very tight race:
  - First count: Rossi won by a few hundred votes
  - Machine recount: Rossi won by a few hundred votes
  - Hand recount: Gregoire won by a few hundred votes

9 / 19

Illegal votes

- It turned out that there were illegal votes cast by felons
- Republican party went to court, and argued that Rossi would have won if one accounts for the illegal votes
- UW Stat professors were expert witness in the court case
- The Democrats won the case, and Gregoire is the current governor

10 / 19

The court case

- Case presented by the Republicans:
  - They identified precincts with a high percentage of Democratic voters
  - In those precincts, they identified illegal voters
  - It is impossible to find the ballots of these people, so we don't know what they voted. But the Republicans wanted to guess their vote. Example:
    - Suppose 60% voted Democratic in the precinct
    - Suppose they found 100 illegal voters
    - Then they estimated that 60 of these were for the Democrats, and 40 for the Republicans
    - They wanted to deduct these votes from the total
- What problems do you see?

Problems

- Main statistical problems:
  - The sample of precincts was a convenience sample. It is not fair to only look at a number of special hand-picked precincts.
  - The method for deducting the votes is not correct
- Moreover:
  - The democrats also looked for illegal voters:
    - They identified precincts with a high percentage of Republican voters
    - They found illegal voters in these precincts
  - If they followed the method proposed by the Republicans on their combined sample, then Gregoire would still win.

What is wrong with deduction method?

- We only have information on the percentage of Democratic/Republican votes for the precinct as a group. We do not know how each individual voted.
- The illegal voters are a special subgroup
- The Republicans assumed that this subgroup voted like the people did on average, but that does not have to be true!
- Example:
  - Overall, 55% of the votes were cast by women
  - Using the method of the Republicans, we would guess that 55% of the illegal voters were women
  - But that is not true! Illegal voters are mostly felons, and felons are mostly men. They are a special subgroup!

13 / 19

Main lesson

- Data on group level is called ecological data
- When you have data on group level (like proportion of Republican voters), then you cannot say much about subgroups or individuals
- If you want to conclude something about individual people, then you want to have data about individual people.

14 / 19

Effect on r

- What is the effect on the correlation coefficient?
  - Ecological correlations (=correlations based on rates or averages of a group) usually overstate the strength of the association.
  - Why is that?
    - There is a lot of variation between individuals
    - Taking rate or averages of groups removes some of the variation
    - This makes it seem as if there is more clustering
  - So watch out for this!

15 / 19

Association is not causation

Example 1

- The shoe size of children is strongly correlated with reading skills
- Is there an association between shoe size and reading skills? Yes, knowing the reading skills gives you information to guess the shoe size of a child.
- Is there causation? Does learning new words make your feet grow? No!
- There is a confounding factor:
  - As you grow older, your reading ability increases
  - As you grow older, your feet grow
  - So age is a difference between the two groups (the kids who read well, and the kids who read poorly), that affects the outcome (shoe size)
- See diagram on overhead

17 / 19

Example 2

- In the Great Depression (1929-1933), better educated people tended to be unemployed for shorter periods of time.
- Is there an association between education and unemployment periods? Yes, knowing somebody was highly educated, makes you guess that his/her unemployment periods were short
- Does education protect against unemployment? Perhaps...
- There is a confounding factor:
  - Highly educated people tended to be younger, because education goes up over time
  - Employers tended to prefer to hire younger people
  - So age is a difference between the two groups (highly educated people, and lowly educated people), that affects the outcome (unemployment period)
- See diagram on overhead

18 / 19

Examples from the news

- Recent examples from the news (see the tab 'Links' on the class website):
  - Dust may quell hurricanes
  - Study of psychiatric drug finds new is not better
  - Cola raises women's osteoporosis risks

19 / 19