Chapter 8: Correlation

Context 2
Association 3
Dependent and independent variables 4
Summarizing data on two variables 5
Interpreting r 6
Values for r 7
SD line 8
Computing r 9

Context

- We looked at data collection:
  - controlled experiments (Ch 1)
  - observational studies (Ch 2)
  - sampling (Ch 19)
- Then we looked at summarizing data on one variable:
  - histogram (Ch 3)
  - average, median, SD (Ch 4)
  - normal approximation, percentiles, boxplot (Ch 5)
- Now we will look at summarizing the relationship between two variables. Example: height of fathers and sons. See overhead.

2 / 9

Association

- If there is a strong association between two variables, then knowing one helps a lot in predicting the other.
- If there is a weak association, then knowing one variable does not help much in guessing the other.
- Examples:
  - The length and weight of a 2 by 4 - strong association
  - Height of fathers and sons - medium association
  - The height of people and the size of their house - weak association

3 / 9

Dependent and independent variables

- Usually we label one variable as independent and one as dependent
- The independent variable is thought to influence the dependent variable
- Example:
  - A father's height influences the son's height
  - So the father's height is the independent variable and the son's height is the dependent variable.
- This is not a set and stone rule - depends on the research
- Which one to put on which axis?
  - The independent variable is put on the horizontal x-axis
  - The dependent variable is put on the vertical y-axis
- See overhead

4 / 9

Summarizing data on two variables

- Suppose you want to describe the relationship between two variables
- The scatter diagram is football shaped
- What information do we need? See overhead:
  - Center: point of averages
    - average of the x-values
    - average of the y-values
  - Spread:
    - SD of x-values
    - SD of y-values
  - Clustering around a line: correlation coefficient r
- So we need 5 numbers!

5 / 9

Interpreting r

- r measures clustering of the cloud of points around a line
- r can only have values between -1 and 1
- Direction of the linear relationship:
  - positive value $(+)$ means that the line/cloud slope up: as one variable increases, so does the other
  - negative correlation (-) means that the line/cloud slope down: as one variable increases, the other decreases
- Strength of the linear relationship:
  - A value close to 0 means that the linear relationship between the two variables is weak. The points are widely spread around the line.
  - A value close to -1 or 1 means that the linear relationship between the two variables is strong. The points are tightly clustered around the line.

6 / 9

Values for r

- See overhead for examples
- For social science studies: values of 0.3-0.7 are common
- Warnings:
  - $r = 0.8$ does not mean that 80% of the points are tightly clustered around the line
  - $r = 0.8$ does not indicate twice as much linearity as $r = 0.4$

7 / 9

SD line

- Points of scatter diagram tend to cluster around the SD line
- The SD line goes through the point of averages
- The SD line goes through all the points that are an equal number of SDs away from the average, for both variables. Example:
  - If a father is 1 SD above average in height, and his son is 1 SD above average in height, then the point falls on the SD line.
  - If a father is 1 SD above average in height, and his son is 0.5 SD above average in height, then the point does not fall on the SD line.
- The slope of the SD line is:
  - (SD of y)/(SD of x) if the correlation is positive
  - - (SD of y)/(SD of x) if the correlation is negative

8 / 9

Computing r

- Step 1: Convert the x-values to standard units
- Step 2: Convert the y-values to standard units
- Step 3: For each row in the table, work out the product:
  (x in standard units) $\times$ (y in standard units)
- Step 4: Take the average of the products. That is r.
- Example: see overhead
- Why does this work? See overhead
- What happens to r if we do a change of scale? (adding a number to all entries of the list, or multiplying all entries of the list by a number)

9 / 9