Correlation (part 1)

Ryan Miller

Introduction

At this point we've discussed associations in the follow contexts:

- Relating two categorical variables - contingency tables and differences in row/column proportions
- Relative one categorical and one quantitative variable - side-by-side graphs and differences in means

Introduction

At this point we've discussed associations in the follow contexts:

- Relating two categorical variables - contingency tables and differences in row/column proportions
- Relative one categorical and one quantitative variable - side-by-side graphs and differences in means

This presentation will cover the remaining scenario, relating two quantitative variables

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 hereditable traits

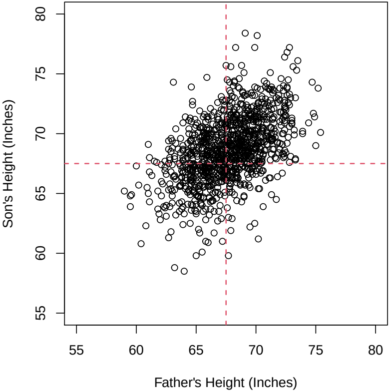
Pearson's Height Data

- 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 hereditable traits
- Wondering if height is hereditable, 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
...	...

Pearson's Height Data

Using a scatterplot an association is obvious:



But how do we summarize it?

Pearson's Correlation Coefficient

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

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

Pearson's Correlation Coefficient

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- The 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

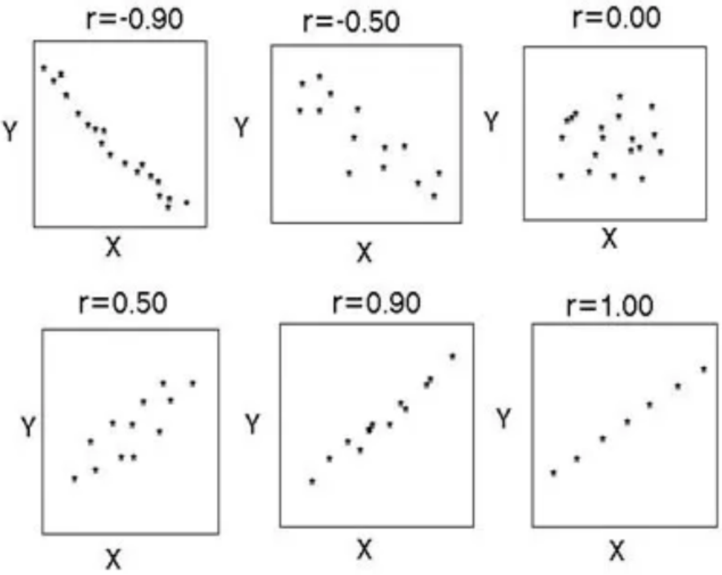
Pearson's Correlation Coefficient

- Consider two variables, X and Y, and their average values, $\bar{x}$ and $\bar{y}$
- The 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

Examples



Strength of Association

Whether a correlation is considered "strong" or "weak" depends on the 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

1. Open the "Tips" dataset in the "data explorer" app
2. Create a scatterplot using "tip" as the X variable and "tot_bill" as the Y variable
3. Describe whether you see an association
4. Describe the correlation coefficient between the two variables using the "Summarize the Data" tab