Lab 4a: Correlation

Stat 131A, Fall 2018

Learning Objectives:

- Use a correlation coefficient to describe the direction and strength of a linear relationship.
- Distinguish between association and causation.
- Identify lurking variables that may explain an observed relationship.

General Instructions

- Write your solutions in an `Rmd` (R markdown) file.
- Name this file as `lab04a-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `lab04a-gaston-sanchez.Rmd`).
- Knit your `Rmd` file as an html document (default option).
- Submit your `Rmd` and `html` files to bCourses, in the corresponding lab assignment.

Problem 1

A small data set is shown below.

```
x = c(1, 2, 3, 4, 5)
y = c(2, 3, 1, 5, 6)
```

a. Use `cor()` to compute the correlation coefficient r_{xy}.

b. If you compute r_{yx}, does this change the correlation between y and x?

c. Add 3 to each value of `y`, and compute again the correlation. Does this change the correlation?

d. Double each value of `x`, and compute again the correlation. Does this change the correlation?

e. Use `x` as in (d) and `y` as in (c) to compute the correlation. Does this change the correlation?

f. Use `x` and `y` as in (a) but you interchange the last two values (5 and 6) for `y`. Does this change the correlation?

Problem 2

Six data sets are shown below. In set 1, the correlation is 0.8571, and in set 2 the correlation is 0.7857. Find the correlation for the remaining data sets. You don't really need to do any arithmetic, but in case you do then do your calculations in R.

Set 1		*Set 2*		*Set 3*		*Set 4*		*Set 5*		*Set 6*	
x	y	x	y	x	y	x	y	x	y	x	y
1	2	1	2	2	1	2	2	1	4	0	6
2	3	2	3	3	2	3	3	2	6	1	9
3	1	3	1	1	3	4	1	3	2	2	3
4	4	4	4	4	4	5	4	4	8	3	12
5	6	5	6	6	5	6	6	5	12	4	18
6	5	6	7	7	6	7	5	6	10	5	21
7	7	7	5	5	7	8	7	7	14	6	15

Problem 3

Two wheathermen compute the correlation between daily maximum temperatures for Washington and Boston. One does it for June; the other does it for the whole year. Who gets the bigger correlation? ("Washington" is the city, not the state).

Problem 4

Correlation coefficient: If the correlation coefficient for a given scatterplot is r = -0.81, which of the following must be true about the relationship between the explanatory and response variable?

a. The association has a linear form.

b. The association must be positive.

c. The association must be negative.

d. The association is weakened by outliers.

Problem 5

Wedding expenses and marriage length: With the headline, "Want a happy marriage? Have a big, cheap wedding" CNN reported on a study that examined the correlation between wedding expenses and the length of marriages. The news article states, "A new study found

that couples who spend less on their wedding tend to have longer-lasting marriages than those who splurge."

What would be a reasonable explanation for the observed correlation?

a. Having an inexpensive wedding helps young couples avoid financial burdens that may strain their marriage.

b. Having an inexpensive wedding guarantees a couple will have a long-term marriage because the study shows a strong correlation between the two variables.

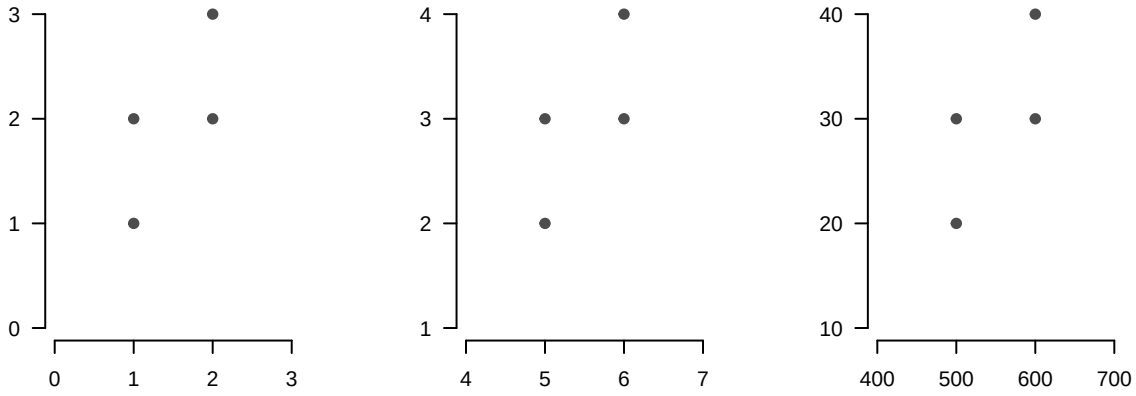
c. Having an inexpensive wedding has no impact on the length of marriage because the cost of a wedding is a confounding variable that explains the correlation.

Problem 6

Many studies have found an association between gas prices and car accidents (high gas prices lead to fewer auto accidents). One study found an association between gas prices and traffic congestion. Should you conclude that the decreasing cost of gas causes more traffic jams? Or can you explain a rising of traffic congestion in some other way?

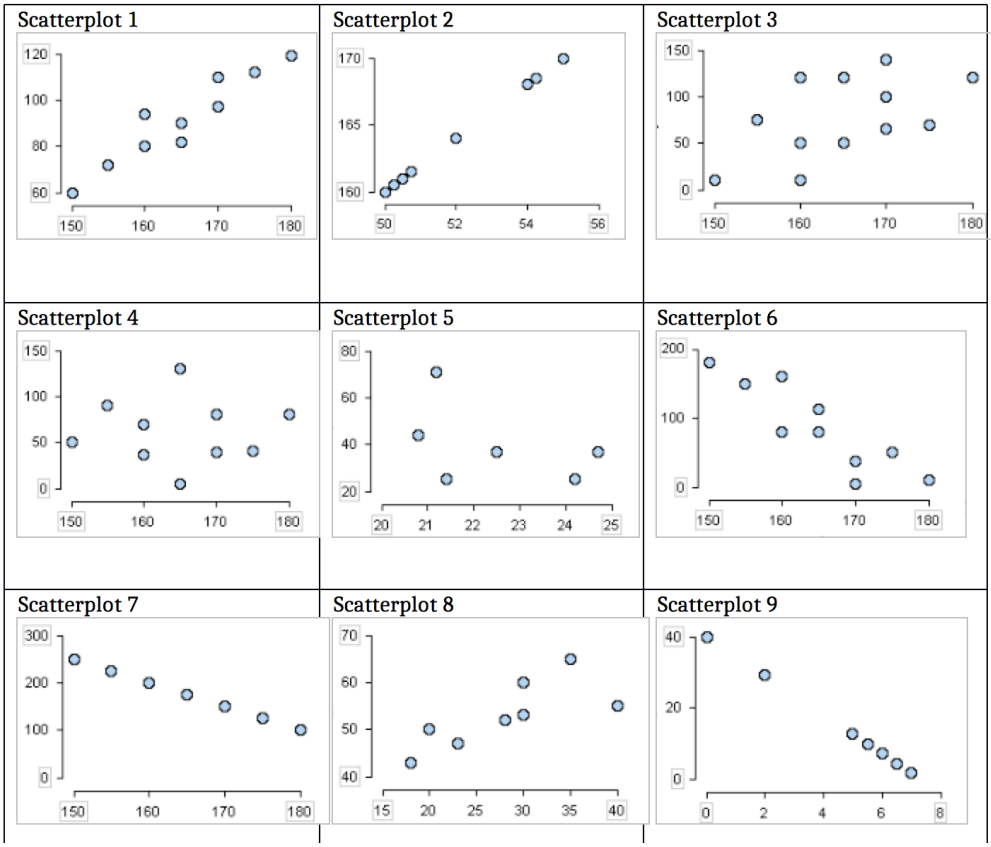
Problem 7

Below are three scatter diagrams. Do they have the same correlation? Yes or No, and why?



Problem 8

Label each scatterplot with its correlation coefficient given below.



a. Correlation $r = 1$

b. Correlation $r = -1$

c. Correation $r = -0.44$

d. Correlation $r = 0.94$

e. Correlation $r = -0.88$

f. Correlation $r = 0.55$

g. Correlation $r = 0.75$

h. Correlation $r = 0.01$

Problem 9

A class of 15 students happens to include 5 basketball players. True or False, and explain: the relationship between heights and weights for this class should be summarized using r.

Problem 10

The correlation between height and weight among men age 18-74 in the U.S. is about 0.40. Say whether each conclusion below follows from the data; explain your answer.

a. Taller men tend to be heavier.

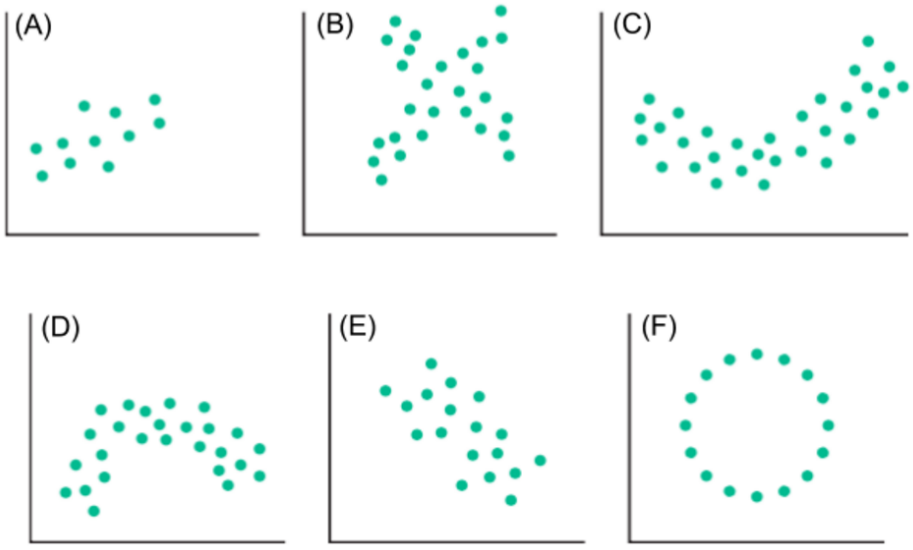
b. The correlation between weight and height for men age 18-74 is about 0.40.

c. Heavier men tend to be taller.

d. If someone eats more and puts in 10 pounds, he is likely to get somewhat taller.

Problem 11

Which of the following six scatter diagrams should be summarized by r? Explain.



Problem 12

On a multiple-choice exam, there are 100 problems. Let X be the number of problems a student got right, and Y the number a student got wrong. If the average and SD of X is 60 and 10, respectively, find:

a. The average and SD of Y.

b. What is the correlation between X and Y?

Problem 13

Consider the following options for a correlation coefficient r:

a. exactly -1
b. close to -0.5
c. close to 0
d. close to 0.5
e. exactly 1

If women always married men who were five years older, the correlation coefficient between the ages of husbands and wives would be _____. Choose one of the options above, and explain your reasoning.

Problem 14

A study of tutoring services at a community college shows a strong positive association between time spent in the math lab and math exam average.

Decide whether or not each statement below is valid or invalid, and explain your reasoning.

a. Students who spend more time in the math lab tend to have higher exam averages.

b. The tutoring services in the math lab are improving student performance on math exams.

c. To improve student performance in math classes, the college should require students to attend the math lab.