Lab 3b: Scatterplots and Association

Stat 131A, Spring 2019

Learning Objectives:

- Use a scatterplot to display the relationship between two quantitative variables.
- Describe the overall pattern (form, direction, and strength) and striking deviations from the pattern.
- 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.

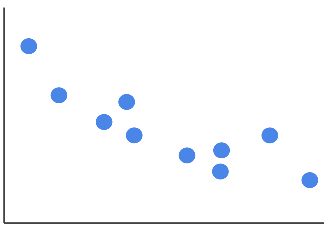
General Instructions

- Write your solutions in an `Rmd` (R markdown) file.
- Name this file as `lab03b-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `lab03b-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.

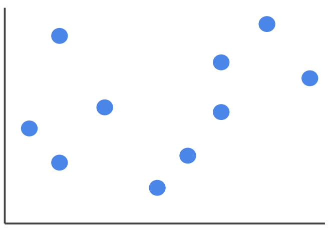
Problem 1

The scatterplots below differ in the direction of the association. The direction is described as positive association or negative association (or neither).

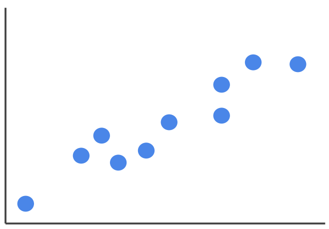
Scatterplot 1



Scatterplot 2



Scatterplot 3



a) Match descriptions to scatterplots.

i. x = city miles per gallons and y = highway miles per gallon for 10 cars.

ii. x = sodium (mg/serving) and y = Consumer Report rating for 10 brands of tomato soup.

iii. x = price ($) and y = sodium (mg/serving) for 10 brands of vegetable soup.

b) Based on the scatterplots, label each association as a positive association, a negative association, or neither.

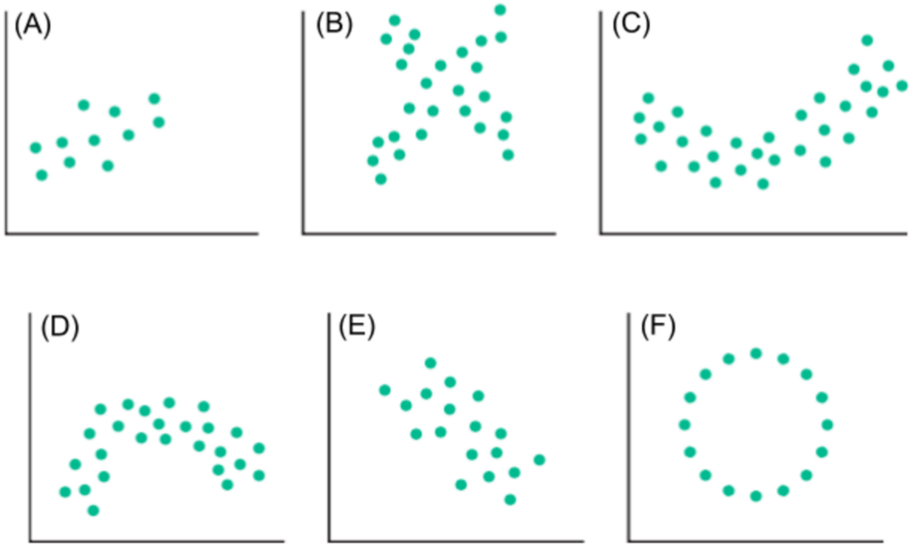
i. The association between city mpg and highway mpg for a sample of cars.

ii. The association between sodium in a serving and Consumer Report rating for different brands of tomato soup.

iii. The association between price and sodium in a serving for different brands of vegetable soup.

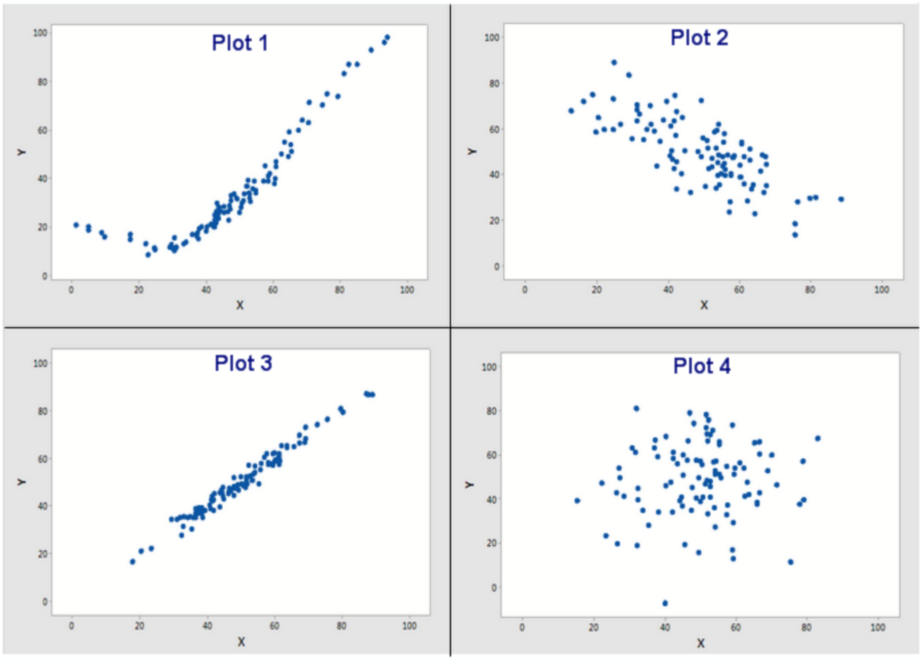
Problem 2

Describe each of the following scatter diagrams in terms of *strength of association*, *form*, and *direction*.



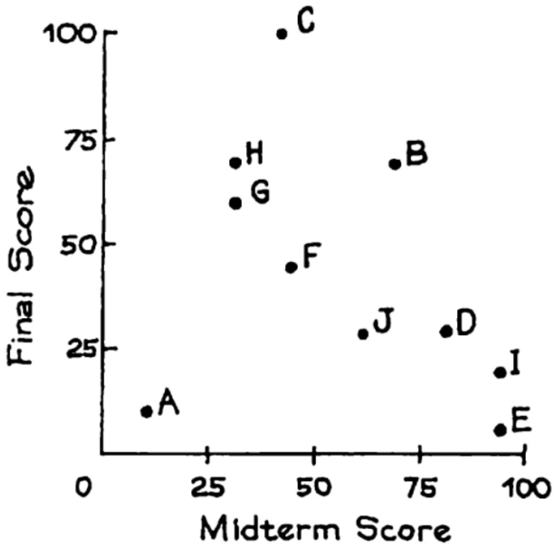
Problem 3

Provide descriptions of the overall pattern in each of the following scatterplots; in particular describe: direction, form, and strength, as well as striking deviations from the pattern.



Problem 4

Students named A, B, C, D, E, F, G, H, I, and J took a midterm and a final in a certain course. A scatter diagram for the scores is shown below.



a. Which students scored the same on the midterm as on the final?

b. Which students scored higher on the final?

c. Was the mean score on the final around 25, 50, or 75?

d. Was the SD of the scores on the final around 10, 25, or 50?

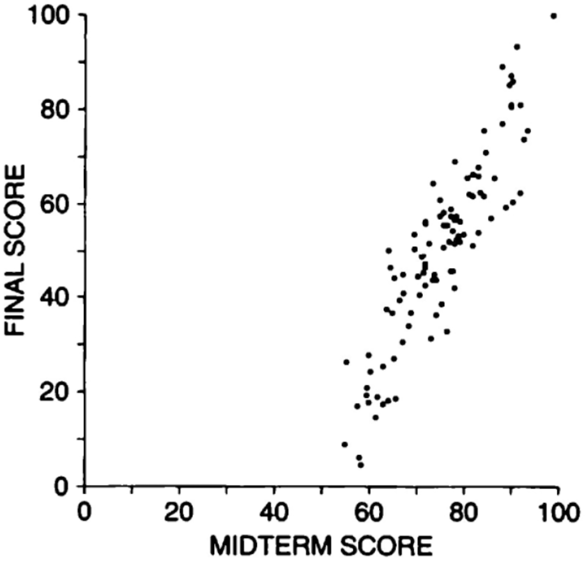
e. For the students who scored over 50 on the midterm, was the mean score on the final around 30, 50, or 70?

f. True or False: on the whole, students who did well on the midterm also did well on the final.

g. True or False: there is strong positive association between midterm scores and final scores.

Problem 5

The scatter diagram below shows scores on the midterm and final in a certain course.



a. Was the average midterm score around 25, 50, or 75?

b. Was the SD of the midterm scores around 5, 10, or 20?

c. Was the SD of the final scores around 5, 10, or 20?

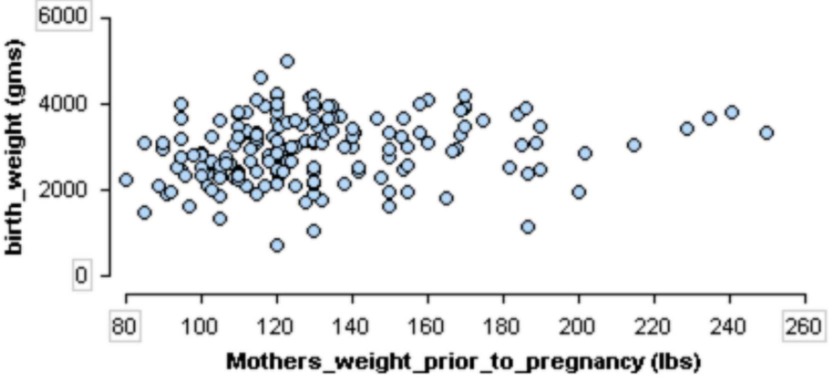
d. Which exam was harder—the midterm or the final?

e. Was there more spread in the midterm scores, of the final scores?

f. True or False: there was a strong positive association between midterm scores and final scores.

Problem 6

The data associated to the following scatterplot comes from a research study on new mothers to identify variables connected to low birth weights. This scatterplot investigates the relationship between two quantitative variables in the study: *mother's weight prior to pregnancy* and *baby's birth weight*.



a) What does a point represent in this scatterplot?

i. a mother's weight.

ii. a new mother.

iii. a baby's birth weight.

b) Approximately what is the pre-pregnancy weight of the woman with the heaviest baby?

i. 125 pounds.

ii. 250 pounds.

iii. 6,000 grams.

c) Approximately what is the range of birth weights for babies in this study?

i. 80 to 250 lb.

ii. 700 to 5,000 gm.

iii. 0 to 6,000 gm.

d) This study attempted to relate different variables to low birth weight. Which of the following variables in this data set could not be used with birth weight to create a scatterplot?

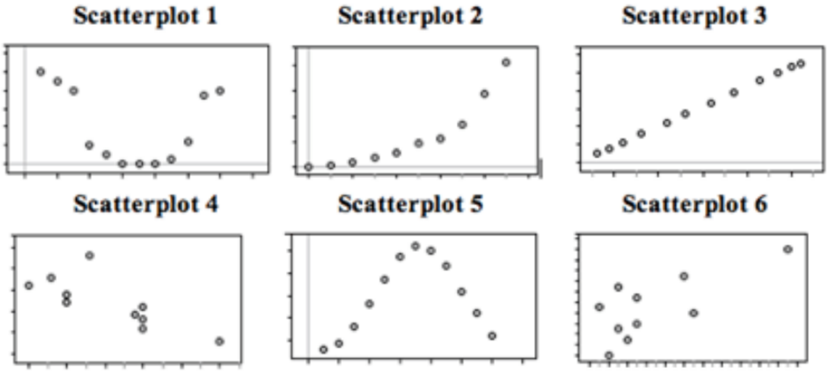
i. mother's age at delivery (years).

ii. number of doctor visits during 1st trimester of pregnancy.

iii. whether mother smoked during pregnancy (yes, no)

Problem 7

Indicate which scatterplot matches the given descriptions.



A: X = month (January = 1), Y = rainfall (inches) in Napa, CA in 2010 (Note: Napa has rain in the winter months and months with little to no rainfall in summer.)

B: X = month (January = 1), Y = average temperature in Boston MA in 2010 (Note: Boston has cold winters and hot summers.)

C: X = year (in five-year increments from 1970), Y = Medicare costs (in $) (Note: the yearly increase in Medicare costs has gotten bigger and bigger over time.)

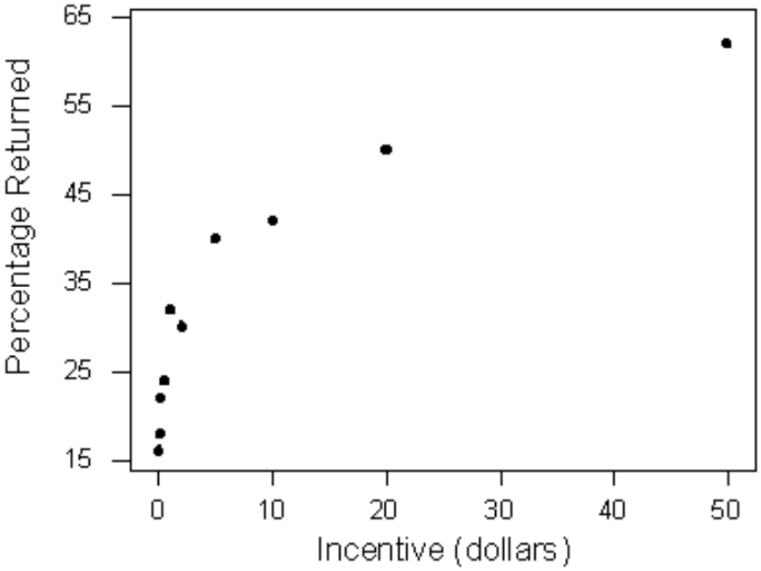
D: X = average temperature in Boston MA (F-degrees), Y = average temperature in Boston MA (C-degrees) each month in 2010.

E: X = chest girth (cm), Y = shoulder girth (cm) for a sample of men

F: X = engine displacement (liters), Y = city miles per gallon for a sample of cars (Note: engine displacement is roughly a measure of engine size. Large engines use more gas.)

Problem 8

Surveys can give useful information about a population if a random sample of people complete the survey. But the information may be biased if only a small percentage of the sample actually return the survey. Will more people complete a survey if they are paid? In this study researchers examine the relationship between monetary incentive and the percentage of the sample who complete the survey.



a) The direction of this relationship is:

i. positive.

ii. negative.

iii. neither positive nor negative.

b) In the context of this example, this means that when researchers promised higher payments, the percentage of participants who completed the survey:

i. increased.

ii. remained the same.

iii. decreased.

c) The form of the relationship is:

i. linear.

ii. curvilinear.

iii. neither linear nor curvilinear.

d) Based on the form of the relationship as it is illustrated above, the relationship is quite:

i. strong.

ii. weak.

e) The point (50, 64) is NOT an outlier with respect to:

i. Incentive.

ii. Percentage returned.

iii. Form and direction of the relationship.

Problem 9

In a study of blood pressure and number of children, it turns out that there is a strong positive correlation between blood pressure and how many children they have. TRUE or FALSE:

a. People with high blood pressure tend to have more children.

b. People with low blood pressure tend to have more children.

c. The scatterplot between blood pressure and number of children must be linear.

d. The person with the largest amount of children, is also the person with highest blood pressure.

e. People with more children tend to have higher blood pressure.

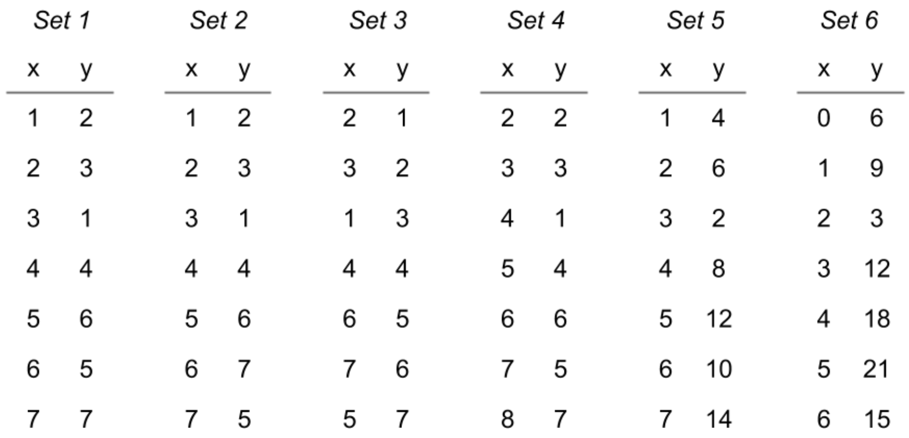
f. Having high blood pressure will cause having more children.

g. Having more children will cause high blood pressure.

h. The person with the lowest blood pressure must have no children.

Problem 10

Six data sets are shown below.

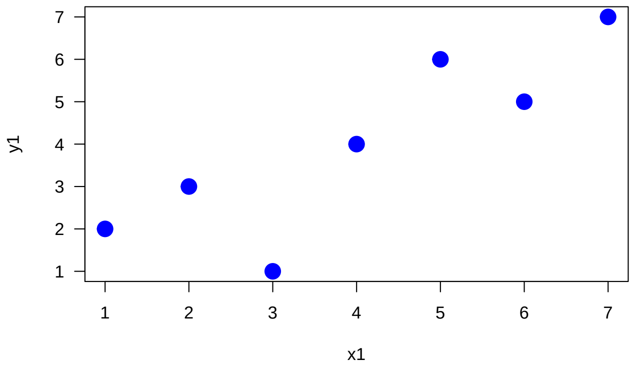


Set 1		*Set 2*		*Set 3*		*Set 4*		*Set 5*		*Set 6*	
x	y	x	y	x	y	x	y	x	y	x	y
1	2	1	2	2	1	2	2	1	4	0	6
2	3	2	3	3	2	3	3	2	6	1	9
3	1	3	1	1	3	4	1	3	2	2	3
4	4	4	4	4	4	5	4	4	8	3	12
5	6	5	6	6	5	6	6	5	12	4	18
6	5	6	7	7	6	7	5	6	10	5	21
7	7	7	5	5	7	8	7	7	14	6	15

Use R to create vectors for each pair of x and y values, and make scatter diagrams: e.g. for instance with `plot()`. Here's an example with the first set:

```
# set 1
x1 = c(1, 2, 3, 4, 5, 6, 7)
y1 = c(2, 3, 1, 4, 6, 5, 7)

plot(x1, y1,
     col = 'blue',  # color
     pch = 19,      # type of symbol (dot)
     cex = 2,       # size of dots
     las = 1        # y-tick marks perpendicular to y-axis
     )
```



Optional: You can try plotting with the package `"plotly"` which produces interactive graphics—nicer and more interesting than the static graphs obtained with `plot()`. To do this, you have to install the package:

```
# Run the following command on the console
# do NOT include this command in your Rmd file
install.packages("plotly")
```

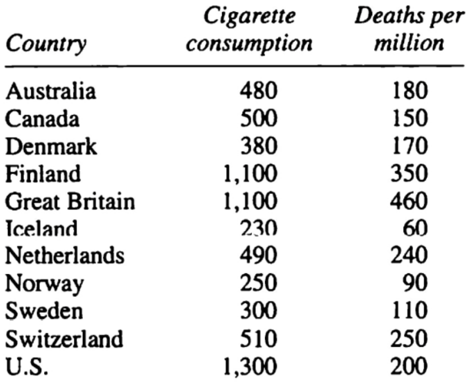
After `"plotly"` has been installed, then load it (include this in your Rmd file). By the way, the output of `plot_ly()` will only work when you knit an html file. If you try to knit using a different format, then `plot_ly()` won't work.

```
# (this is optional)
library(plotly)

plot_ly(x = x1, y = y1, type = "scatter", mode = "markers")
```

Problem 11

The following table shows per capita consumption of cigarettes in various countries in 1930, and the death rates from lung cancer for men in 1950. (In 1930, hardly any women smoked; and a long period of time is needed for the effects of smoking to show up.)



Country	*Cigarette consumption*	*Deaths per million*
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1,100	350
Great Britain	1,100	460
Iceland	230	60
Netherlands	490	240
Norway	250	90
Sweden	300	110
Switzerland	510	250
U.S.	1,300	200

a. Create a data frame in R for the data table. Your data frame should contain three columns: `country`, `cigarette`, and `deaths`. You will use this data frame in part (b), via `ggplot()`.

b. Use the package `ggplot2` to plot a scatter diagram for these data. Hint: `geom_point()` is your friend.

c. True or False: the higher cigarette consumption was in 1930 in one of these countries, on the whole the higher the death rate from lung cancer in 1950. Or can this be determined from the data?

d. True or False. Death rates from lung cancer tend to be higher among those persons who smoke more. Or can this be determined from the data?