MOD 26 (PART 1) SCATTERPLOTS

Learning Goals

- Use a scatterplot to display the relationship between two quantitative variables.
- Describe the overall pattern and striking deviations from the pattern.

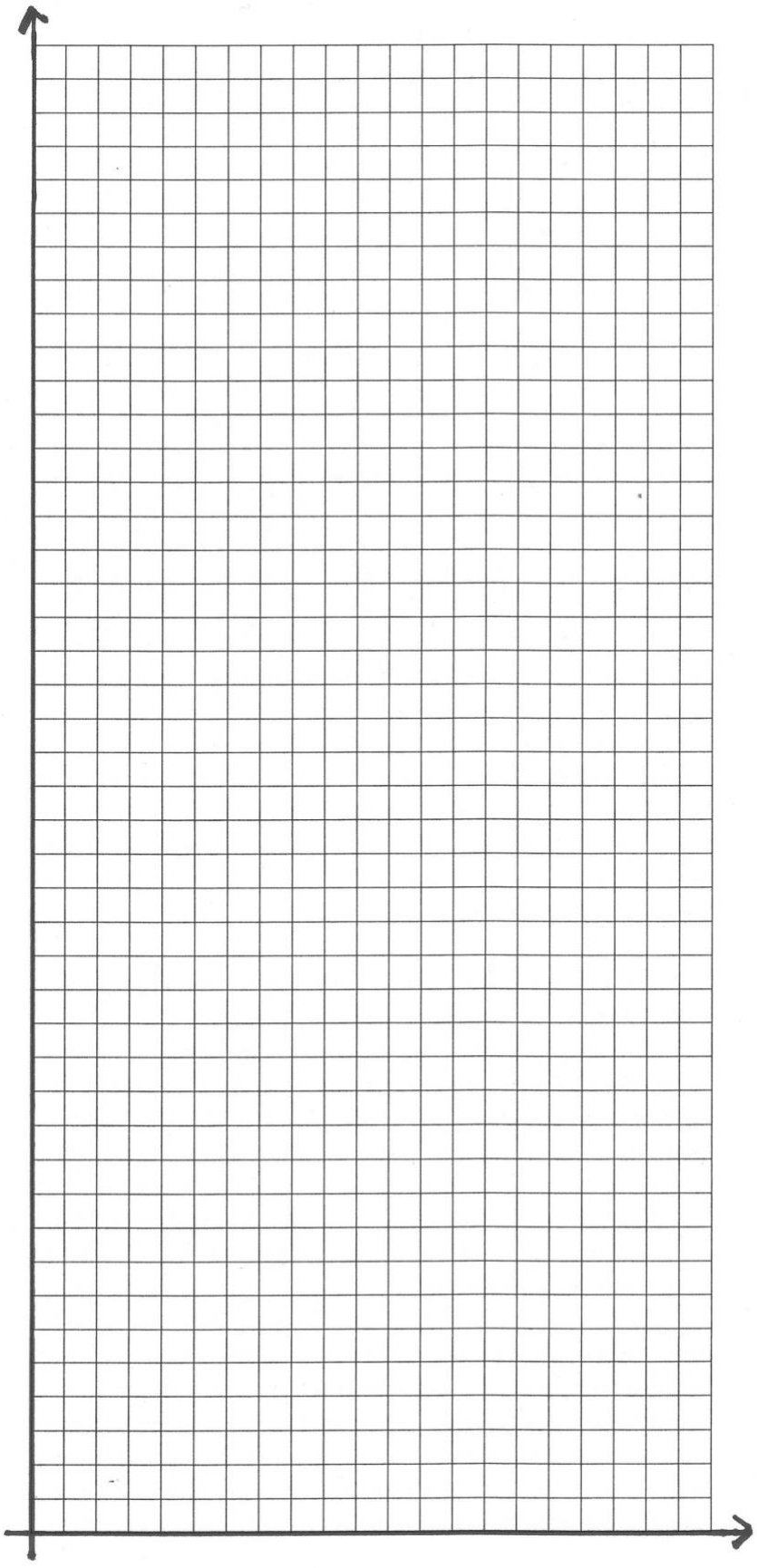
The following table describes the number of AIDS related deaths in New York City. Let n represent the number of AIDS related deaths in hundreds at time t (in years) since 1980 (in other words $t = 6$ would represent 1986 and $n = 2$ would be 200 deaths).

AIDS and HIV Deaths in New York City (1981 – 1994)	
Year	Number of Deaths (in hundreds)
1981	1
1982	2
1983	6
1984	11
1985	18
1986	27
1987	34
1988	43
1989	54
1990	57
1991	65
1992	70
1993	74
1994	84

1) Why would we say that each of these variables is a quantitative variable as opposed to a categorical variable?

2) Which is the explanatory variable and which is the response variable?

3) Use the grid on the back of this page to make a scatterplot of the data (i.e. plot the data points). Be sure to place the explanatory variable on the horizontal axis (a.k.a. the x-axis) and the response variable on the vertical axis (a.k.a. the y-axis), and label your axes. Also, select the scales on each axis so that you use as much of the grid as possible. Which basic shape do these data make in the scatterplot?



4) Now we'll describe the **overall pattern** of the relationship between these two variables (using **direction**, **form**, and **strength**) and striking deviations (if any) from the overall pattern.

a) What is the **direction** of your scatterplot – positive (increasing), negative (decreasing), or neither?

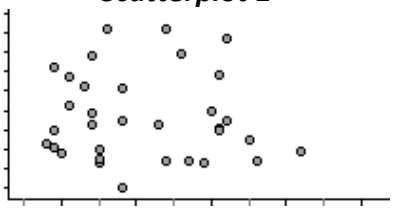
b) What overall **form** does the data in your scatterplot have (linear or curvilinear – if you don't know what these are – ask the teacher)?

c) Guesstimate the **strength** of the relationship of the data in your scatterplot (for now … weak, fairly strong, strong, or very strong). Explain your choice.

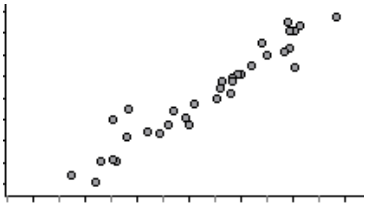
d) Are there any **striking deviations** from the overall pattern (a.k.a. **outliers**)?

5) These scatterplots show various body measurements for 34 adults who exercise several times each week. Describe the overall pattern with regard to direction, form, and strength. Then describe any striking deviations from the pattern.

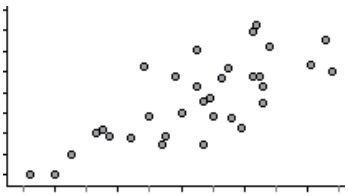
Scatterplot 1



Scatterplot 2

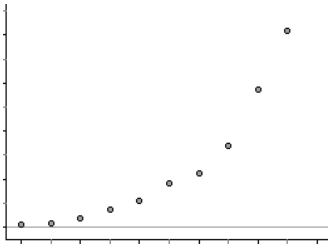


Scatterplot 3

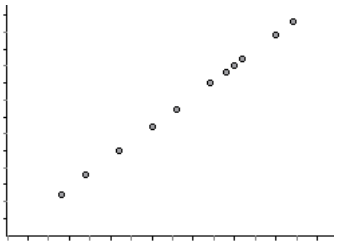


6) For each scatterplot, describe the overall pattern of the relationship between the two variables (using direction, form, and strength) and striking deviations (outliers) from the overall pattern.

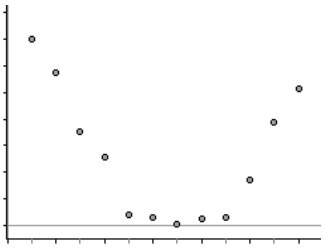
Scatterplot 4



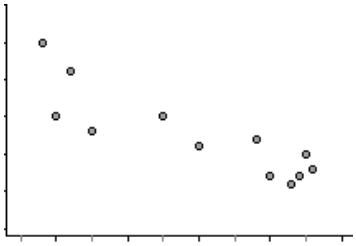
Scatterplot 5



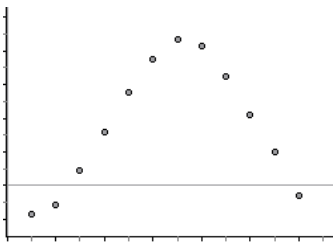
Scatterplot 6



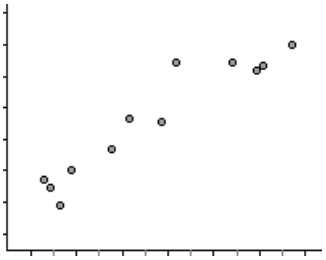
Scatterplot 7



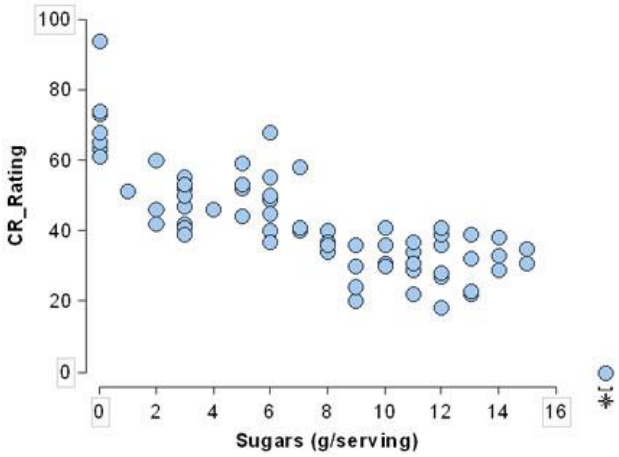
Scatterplot 8

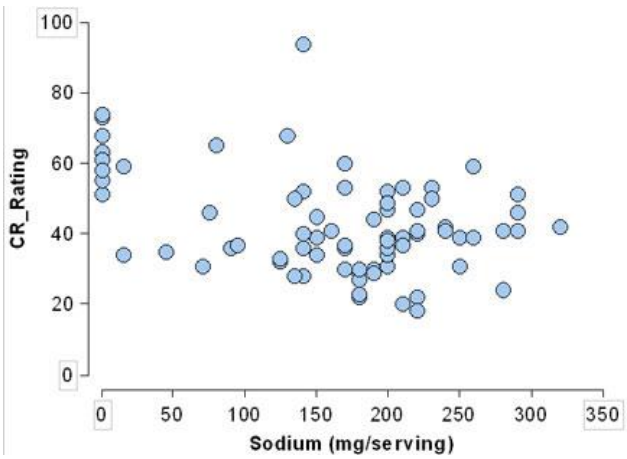


Scatterplot 9



7) Each scatterplot below relates one ingredient to the Consumer Report rating for 77 breakfast cereals. For each scatterplot, describe the overall pattern of the relationship between the two variables (using direction, form, and strength) and striking deviations (outliers) from the overall pattern. Also, for each scatterplot, indicate the explanatory and response variables and then describe what each dot represents





Overall pattern:

Explanatory variable:

Response variable:

Each dot represents:

Overall pattern:

Explanatory variable:

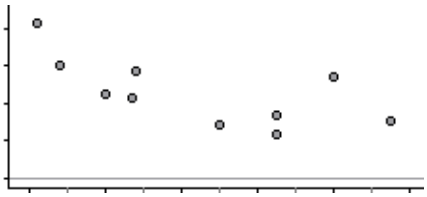
Response variable

Each dot represents:

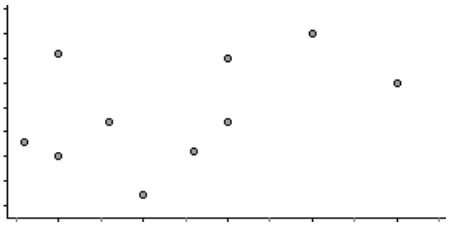
8) Match each pair of variables to a scatterplot, and label the variables on the appropriate axes. Briefly explain your reasoning.

a) x = city miles per gallons and y = highway miles per gallon for 10 cars
b) x = sodium (mg/serving) and y = Consumer Report quality rating for 10 salted peanut butters
c) x = price ($) and y = Consumer Report quality rating for 10 bicycle helmets

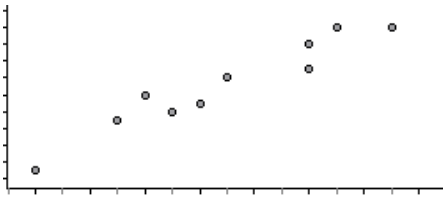
Scatterplot 10



Scatterplot 11



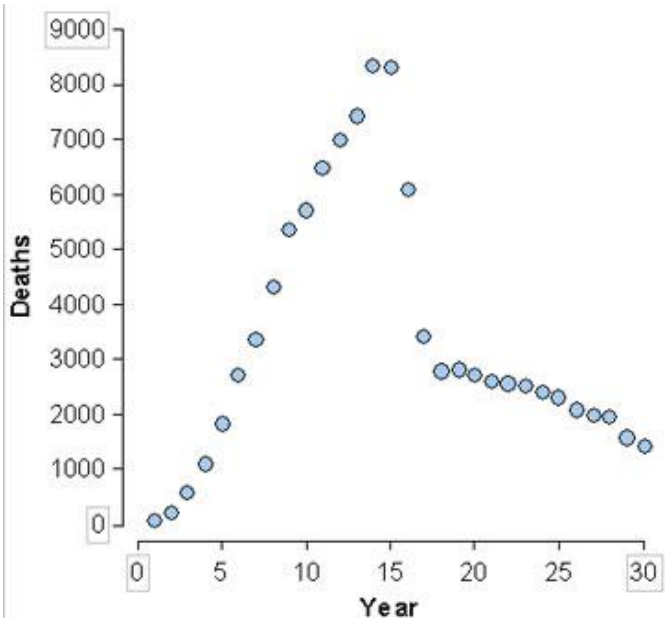
Scatterplot 12



9) For the scatterplots in the previous problem, describe what each dot represents.

10) Oops … we did not quite have all the AIDS related death data for New York City.

AIDS and HIV Deaths in New York City (1981 – 2010)	
Year	Number of Deaths
1981	59
1982	201
1983	593
1984	1107
1985	1828
1986	2720
1987	3350
1988	4300
1989	5358
1990	5724
1991	6475
1992	6985
1993	7429
1994	8355
1995	8322
1996	6078
1997	3428
1998	2795
1999	2805
2000	2710
2001	2577
2002	2554
2003	2520
2004	2387
2005	2318
2006	2090
2007	1988
2008	1940
2009	1589
2010	1413



a) Could we say that the form of this data is linear? Curvilinear? Explain.

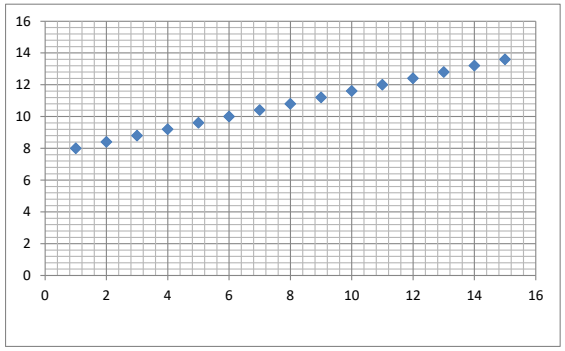
b) Is it possible to force this data set to have a linear form?

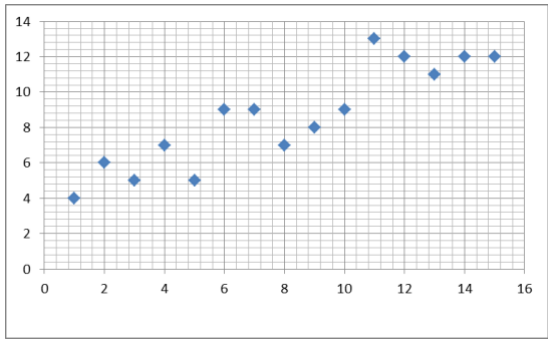
MOD 26 (PART 2) HOMEWORK: AN INTRODUCTION TO THE CORRELATION COEFFICIENT

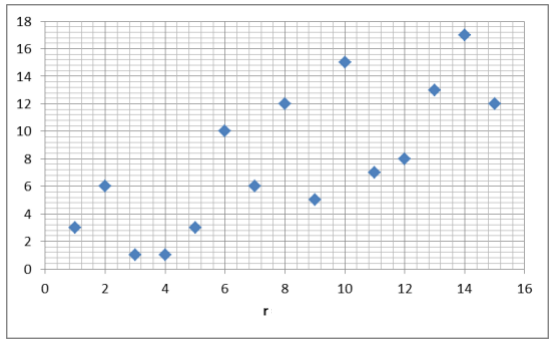
Learning Goals

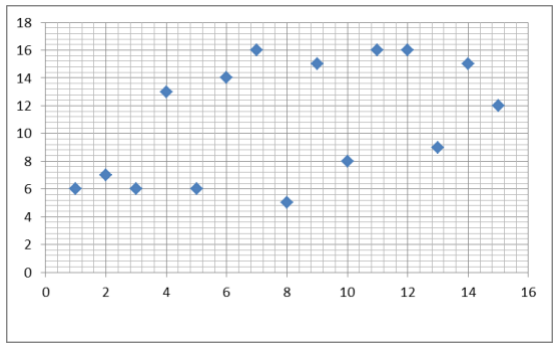
- Interpret the value of the correlation coefficient.

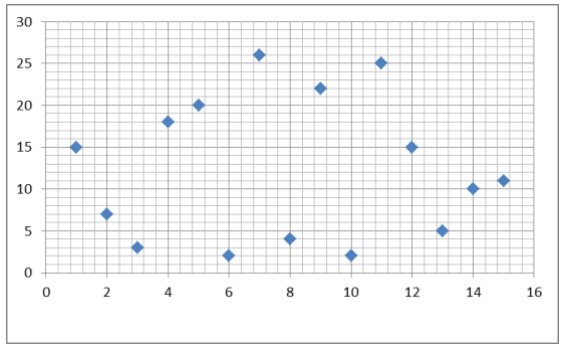
1) For each of the scatterplots rank the strength of relationship (how well it fits the form – which is linear for each of these scatterplots). Use only rankings from 0 to 1 where 0 indicates that the data does not fit the form and a 1 indicates that the data is perfectly linear. You may use rankings such as 0.7 etc. However, if the direction is negative, slap a negative sign on your ranking.

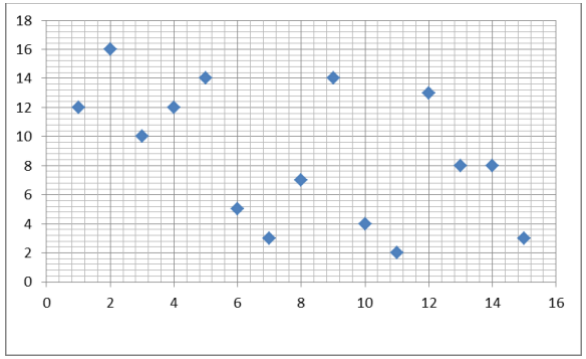




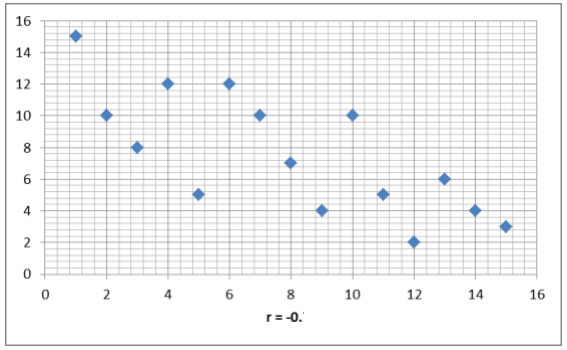


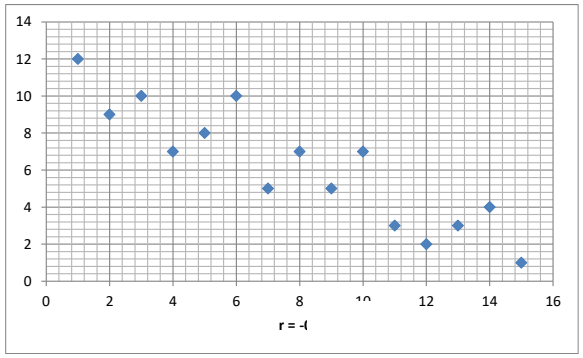


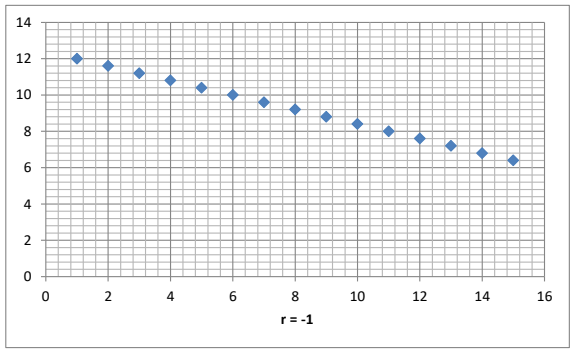




Continued on the back …

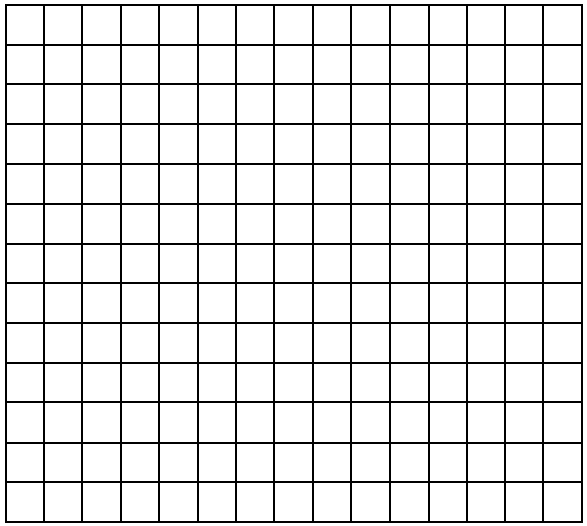


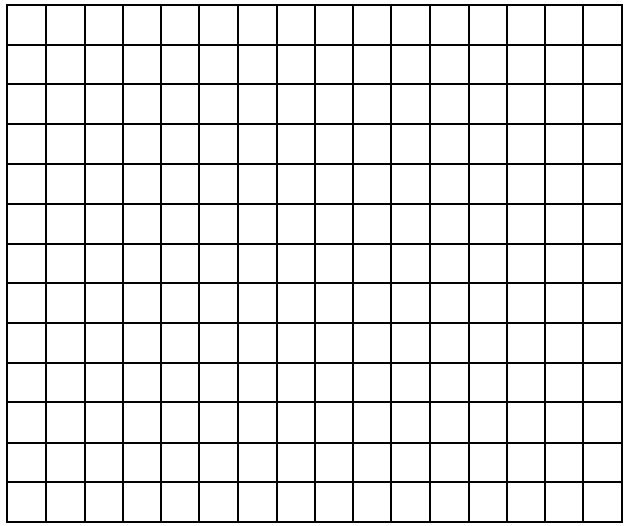




2) If we let r be each of the rankings you guesstimated for each of the scatterplots, then you guesstimated the correlation coefficient for each scatterplot. Below are the actual correlation coefficients for the scatterplots. Go back and label each with the correlation coefficient that is the best indicator of the strength of the relationship for the data (be sure to use the notation "r = " with the number). r = 0, r = 0.5, r = -0.5, r = 0.75, r = -0.75, r = 0.9, r = -0.9, r = 1, r = -1

3) Create two scatterplots each with 10 data points (one with a correlation coefficient of r = -0.8 and one with a correlation coefficient of 0.8).





WARNING: You will need to bring your graphing calculator to class every day.

MOD 26 (PART 3) CAUSATION AND LURKING VARIABLES

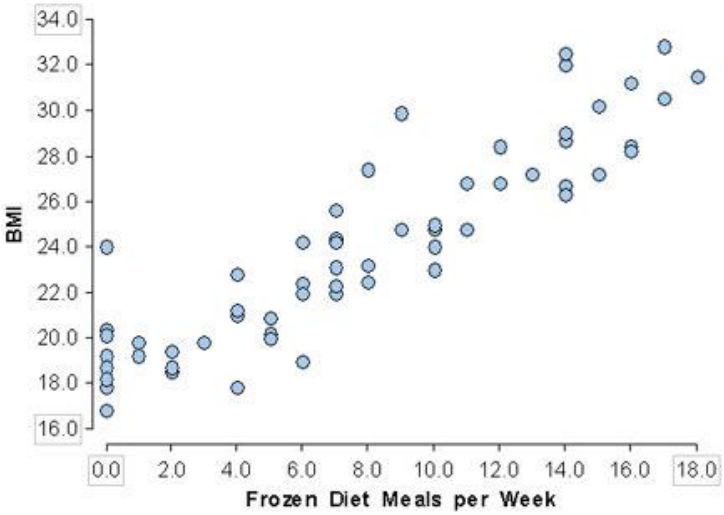
Learning Goals

- Distinguish between association and causation.
- Identify lurking variables that may explain an observed relationship.
- Recognize and explain the phenomenon of Simpson's Paradox as it relates to interpreting the relationship between two variables.

1) For each of the following scatterplots:

   a) identify the explanatory and response variables,
   b) guesstimate the correlation coefficient r (where the correlation coefficient indicates both the strength of the association and the direction of the relationship between the variables),
   c) if the correlation coefficient indicates a strong association, write a sentence indicating that the explanatory variable "causes" the response variable (even if it does not make sense),
   d) determine whether or not the sentence you wrote in part c) makes sense (explain why or why not), and
   e) if the causation sentence does not make sense, but there is a strong correlation, provide an explanation as to why this might happen.

   This scatterplot represents 60 female Cuyamaca College students between the ages of 18 and 35 who were selected at random. They were asked how many frozen diet meals (on average) they ate each week. Then their body mass index (BMI) was measured. A BMI under 18.5 is considered "underweight". A BMI between 18.5 and 24.9 is considered "normal weight". A BMI between 25 and 29.9 is considered overweight. And a BMI of 30 or more is considered "obese".



   a) Explanatory & response variables:

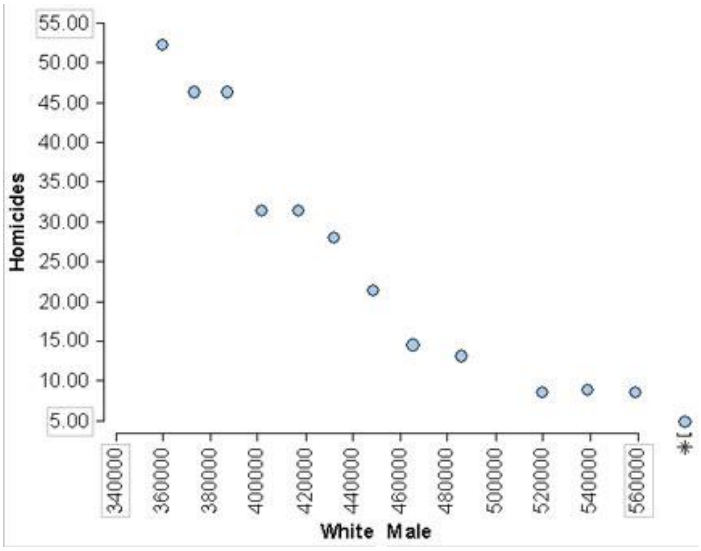
   b) Guesstimate correlation coef:

   c) Causation sentence:

   d) Does the causation sentence make sense?

   e) If the causation sentence does not make sense but the correlation is strong, explain why?

These data were collected by J.C. Fisher and used in his paper: "Homicide in Detroit: The Role of Firearms", Criminology, vol.14, 387-400 (1976). The data are on the homicide rate in Detroit for the years 1961-1973. The number of “Homicides” are given per 100,000 people, and “White Males” indicates the white male population in Detroit.



a) Explanatory & response variables:

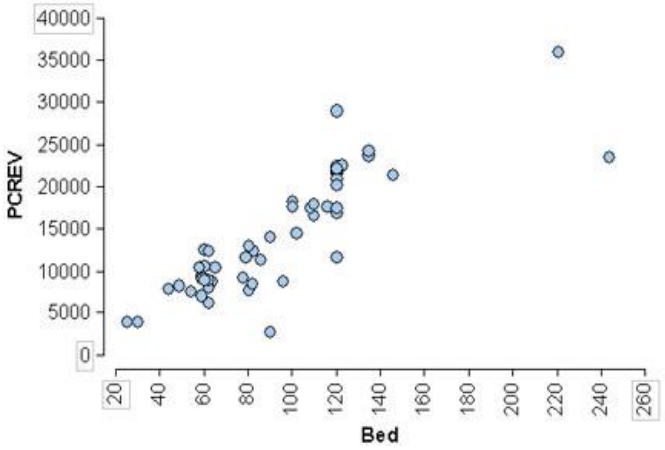
b) Guesstimate correlation coef:

c) Causation sentence:

d) Does the causation sentence make sense?

e) If the causation sentence does not make sense but the correlation is strong, explain why?

These data were collected by the Department of Health and Social Services of the State of New Mexico and cover 52 of the 60 licensed nursing facilities in New Mexico in 1988. “PCREV” is the annual total patient care revenue for the nursing home (in hundreds of dollars). “Bed” is the number of beds in the nursing home.



a) Explanatory & response variables:

b) Guesstimate correlation coef:

c) Causation sentence:

d) Does the causation sentence make sense?

e) If the causation sentence does not make sense but the correlation is strong, explain why?

2) A university offers only two degree programs, one in electrical engineering, and one in teacher education. Admission to these programs is competitive, and the women's caucus suspects discrimination against women in the admission process. The caucus obtains the following data from the university.

		Gender	
		Male	Female
Admission Decision	Admit	35	20
	Deny	45	40
	Total		

		Gender	
		Male	Female
Admission Decision	Admit		
	Deny		
	Total		

a) Complete the first table and then convert it to a table of percentages (enter the percentages in the second table). Round the percentages to one decimal place.

b) Is there an association between gender and admission decision? In other words, are men more likely to be admitted? If so, does it seem the University is indeed discriminating against women?

In its defense the university produced a three-way table that classifies students by gender, admission decision, **and** the program to which they applied. Does the association in part a) hold up, or has the direction of the association changed? Explain.

Electrical Engineering

Admission Decision	Gender: Male	Gender: Female
Admit	30	10
Deny	30	10
Total		

Teacher Education

Admission Decision	Gender: Male	Gender: Female
Admit	5	10
Deny	15	30
Total		

Electrical Engineering

Admission Decision	Gender: Male	Gender: Female
Admit		
Deny		
Total		

Teacher Education

Admission Decision	Gender: Male	Gender: Female
Admit		
Deny		
Total		

c) Complete the tables to tables of percentages. Round the percentages to one decimal place.

d) This has been an example of Simpson's Paradox. Try to write a definition for this paradox.