DESCRIPTIVE STATISTICS: HOMEWORK

EXERCISE 1

The following are real data from Santa Clara County, CA. As of March 31, 2000, there was a total of 3059 documented cases of AIDS in the county. They were grouped into the following categories (Source: Santa Clara County Public H.D.) Research question: Is there a difference between males and females with respect to engaging in one of the following activities (homosexual/Bisexual Contact, IV Drug User, Heterosexual Contact, Other) and then developing AIDS? Based on the above research question determine if row percentages or column percentages would be most appropriate for determining a relationship between variables. Next use your percentages to determine if there is a relationship or if the variables are independent.

Risk Factors

Gender	**Homosexual/ Bisexual**	**IV Drug User ***	**Heterosexual Contact**	**Other**
female	0	70	136	49
male	2146	463	60	135

* includes homosexual/bisexual IV drug users

EXERCISE 2

The following table identifies a group of children by one of four hair colors, and by type of hair. Based on the following research question determine if row or column percentages would be most appropriate for determining a relationship between variables. Next use your percentages to determine if there is a relationship or if the variables are independent. Is there a difference between different hair colors with respect to whether hair is wavy or straight?

Hair color

Hair Type	**Brown**	**Blond**	**Black**	**Red**	**Totals**
Wavy	20		15	3	43
Straight	80	15		12	
Totals		20			215

EXERCISE 3

A previous year, the weights of the members of the San Francisco *49ers* and the Dallas *Cowboys* were published in the *San Jose Mercury News*. The factual data are compiled into the following table. Based on the following research question determine if row or column percentages would be most appropriate for determining a relationship between variables. Next use your percentages to determine if there is a relationship or if the variables are independent. Is there a relationship between weight brackets of players and their shirt number?

Weight (in pounds)

Shirt #	**≤ 210**	**211 - 250**	**251 - 290**	**291 ≤**
1 - 33	21	5	0	0
34 - 66	6	18	7	4
66 - 99	6	12	22	5

EXERCISE 4

The chart below gives the number of suicides comparing blacks and whites estimated in the U.S. for a recent year by age, race and sex. We are interested possible relationships between age, race, and sex. We will let suicide victims be our population. (Source: The National Center for Health Statistics, U.S. Dept. of Health and Human Services). Based on the following research question determine if row or column percentages would be most appropriate for determining a relationship between variables. Next use your percentages to determine if there is a relationship or if the variables are independent. Is there a difference between race and sex with respect to the age of suicide?

Age

Race and Sex	**1 - 14**	**15 - 24**	**25 - 64**	**over 64**	**TOTALS**
white, male	210	3360	13,610		22,050
white, female	80	580	3380		4930
black, male	10	460	1060		1670
black, female	0	40	270		330
all others					
TOTALS	310	4650	18,780		29,760

EXERCISE 5

The data below was obtained from **www.baseball-almanac.com** showing hit information for 4 well known baseball players.Research Questions: Is there a difference between baseball players with respect to the type of hit? Based on the above research question determine if row percentages or column percentages would be most appropriate for determining a relationship

between variables. Next use your percentages to determine if there is a relationship or if the variables are independent.

Type of Hit

NAME	Single	Double	Triple	Home Run	TOTAL HITS
Babe Ruth	1517	506	136	714	2873
Jackie Robinson	1054	273	54	137	1518
Ty Cobb	3603	174	295	114	4189
Hank Aaron	2294	624	98	755	3771
TOTAL	8471	1577	583	1720	12351

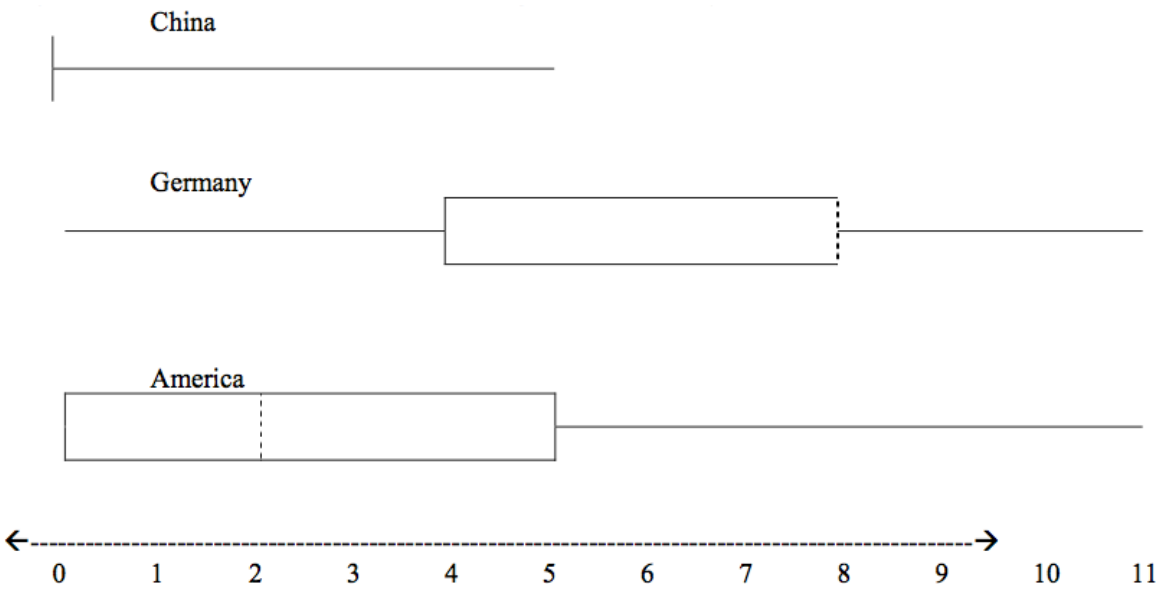
EXERCISE 6

An elementary school class ran 1 mile in an average of 11 minutes with a standard deviation of 3 minutes. Rachel, a student in the class, ran 1 mile in 8 minutes. A junior high school class ran 1 mile in an average of 9 minutes, with a standard deviation of 2 minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran 1 mile in an average of 7 minutes with a standard deviation of 4 minutes. Nedda, a student in the class, ran 1 mile in 8 minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

EXERCISE 7

In a survey of 20 year olds in China, Germany and America, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.

b. Explain how it is possible that more Americans than Germans surveyed have been to over eight foreign countries.
c. Compare the three box plots. What do they imply about the foreign travel of twenty year old residents of the three countries when compared to each other?

EXERCISE 8

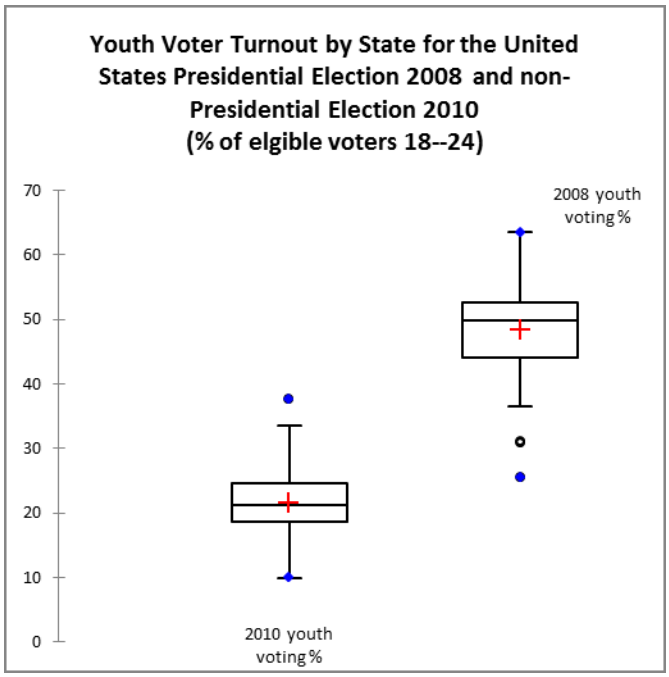
Below are the scores from two different math classes on the same exam.
Construct side by side outlier boxplots for each of the data sets. Include the five number summaries.

class 1	class 2
70	83
71	75
72	76
73	72
74	84
75	90
76	92
77	39
78	91
79	61
80	63
81	74
82	76
83	82
84	92
85	78
86	73
87	68
88	82
89	89
90	86
91	63
40	68
100	

EXERCISE 9

The graph below contains the data for youth voter turnout for the 2008 (Presidential Election) and 2010 (no Presidential Election). The 6 lowest and 6 highest youth turnout states for 2008 were: AR (31.0), GA (25.5), IA (63.5), ME (54.7), MN (62.9), NH (57.7), OH (57), OK (41.5), TN (41.4), TX (36.6), UT (30.9), WI (57.5). The 6 lowest and 6 highest youth turnout states for 2010 were: AR (15.2.0), DC (30.1), IN (9.9), KS (11.7), ME (31.4), NE (10.7), NM (13.5), ND (37.6), OR (32.9), SC (33.5), UT (11.8), WI (31.0). Based on this data and the graph and chart given answer the following questions.

1. What states are more than 1.5 IQR's from the 2008 and 2010 first and third quartiles?
2. Which quartile contain the Minnesota youth turnout (27.9) data for 2010?
3. Which of the following statement can be said about the difference in the IQR of the 2008 and 2010 data?
    a. There is more data in the 2008 than the 2010 IQR since the area is larger.
    b. The range of percents in youth voting in 2010 was about the same as the rage in 2008.
    c. The median youth voter turnout in 2008 was higher than then state with the highest percentage of youth voter turnout.
    d. 50% of the states in 2010 were below the turnout of the state with the lowest youth voter turnout in 2008.



Statistic	2010 youth voting %	2008 youth voting %
No. of observations	51	51
No. of missing values	0	10
Minimum	9.9000	25.5000
Maximum	37.6000	63.5000
1st Quartile	18.7000	44.1000
Median	21.2000	49.9000
3rd Quartile	24.6500	52.5000
Mean	21.5961	48.3488
Variance (n-1)	34.7520	61.6996
Standard deviation (n-1)	5.8951	7.8549

EXERCISE 10

Interested in student athletes study habits, Abby conducts a survey of baseball and track and field athletes asking them how many hours a week they spend studying. Her data is below. Construct outlier boxplots for each group and compare and contrast the two groups.

Baseball	3	4	5	5	7	7	8	9	10	11					
Track and Field	0	4	6	9	9	10	11	11	12	13	14	14	15	15	17

EXERCISE 11

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best G.P.A. when compared to his school? Explain how you determined your answer.

Student	G.P.A.	School Ave. G.P.A.	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

EXERCISE 12

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, each asked adult consumers the number of fiction paperbacks they had purchased the previous month. The results are below.

Publisher A

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher B

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher C

# of books	Freq.	Rel. Freq.
0-1	20	
2-3	35	
4-5	12	
6-7	2	
8-9	1	

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of 1. For Publisher C, make bar widths of 2.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of 2.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

EXERCISE 13

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all on-board transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Below is a summary of the bills for each group.

Singles

Amount($)	Frequency	Rel. Frequency
51-100	5	
101-150	10	
151-200	15	
201-250	15	
251-300	10	
301-350	5	

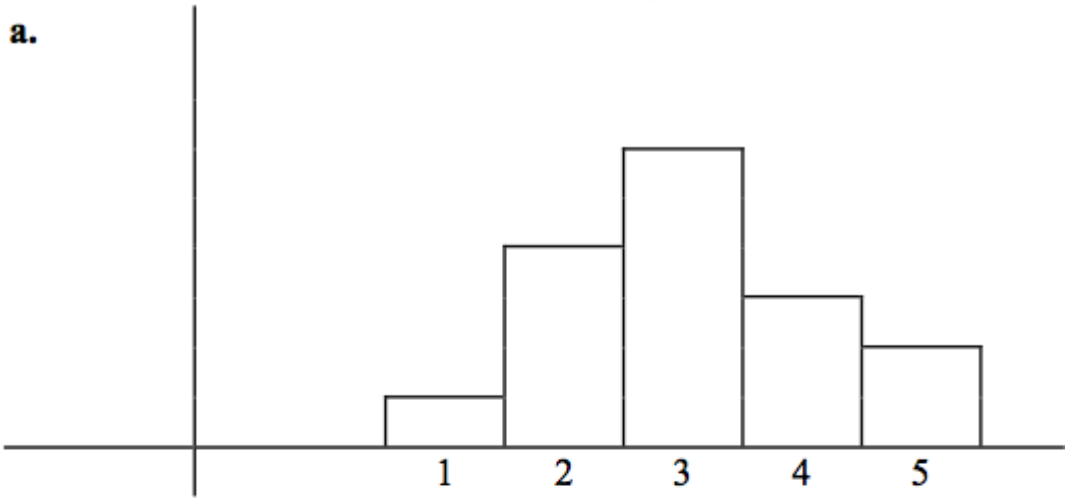
Couples

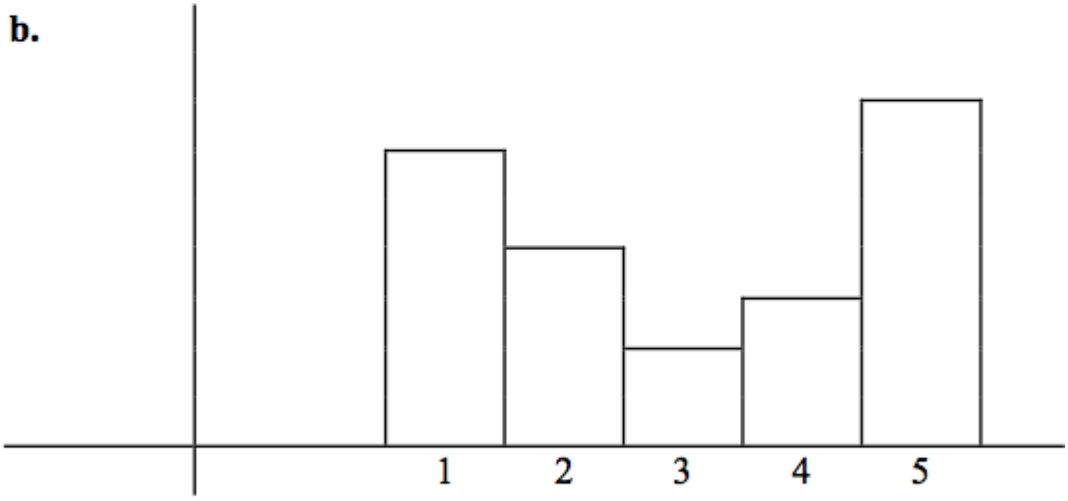
Amount($)	Frequency	Rel. Frequency
100-150	5	
201-250	5	
251-300	5	
301-350	5	
351-400	10	
401-450	10	
451-500	10	
501-550	10	
551-600	5	
601-650	5	

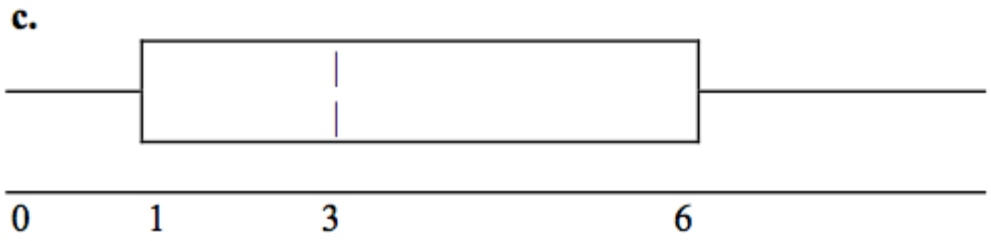
a. Fill in the relative frequency for each group.
b. Construct a histogram for the Singles group. Scale the x-axis by $50. widths. Use relative frequency on the y-axis.
c. Construct a histogram for the Couples group. Scale the x-axis by $50. Use relative frequency on the y-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the Couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by $50, scale it by $100. Use relative frequency on the y-axis.
f. Compare the graph for the Singles with the new graph for the Couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. By scaling the Couples graph differently, how did it change the way you compared it to the Singles?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person in a couple? Explain why in one or two complete sentences.

EXERCISE 14

Refer to the following histograms and box plot. Determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



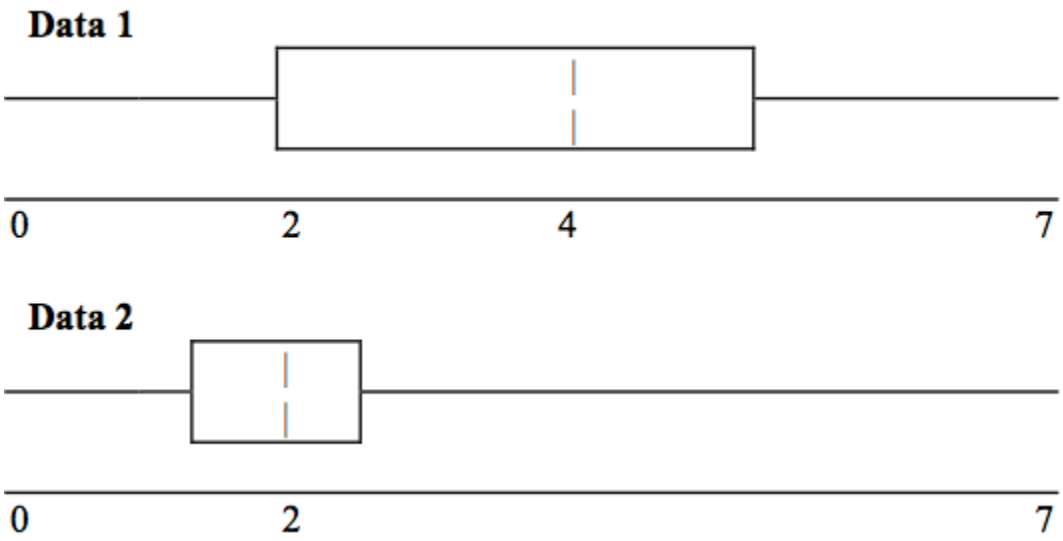




a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for (b) is larger than the standard deviation for (a).
d. We cannot determine if any of the third quartiles for the three graphs is different.

EXERCISE 15

Refer to the following box plots.



a. In complete sentences, explain why each statement is false.
    i. Data 1 has more data values above 2 than Data 2 has above 2.
    ii. The data sets cannot have the same mode.
    iii. For Data 1, there are more data values below 4 than there are above 4.
b. For which group, Data 1 or Data 2, is the value of "7" more likely to be an outlier? Explain why in complete sentences

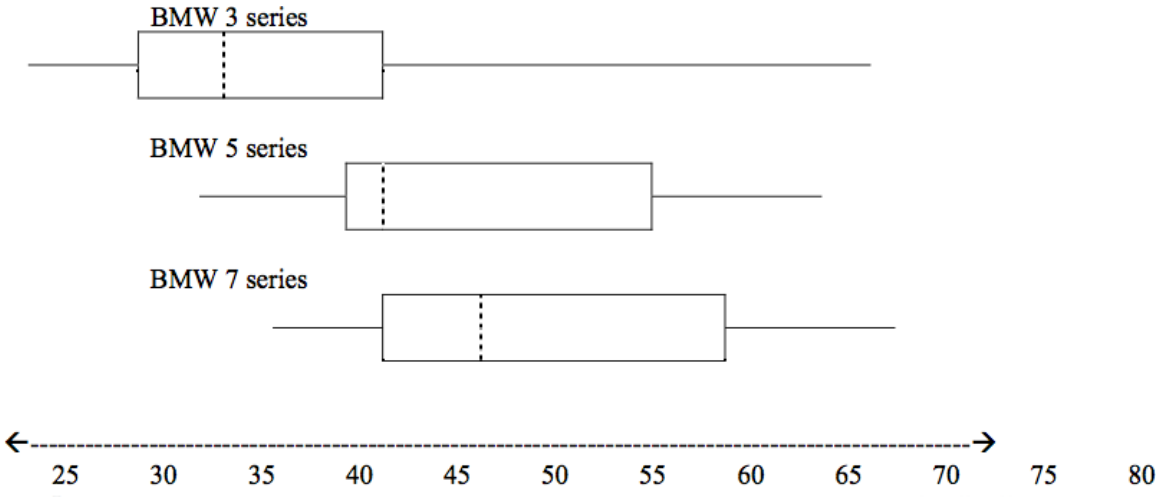
EXERCISE 16

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years. (Source: Bureau of the Census)

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

EXERCISE 17

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
b. Which group is most likely to have an outlier? Explain how you determined that.
c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
d. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is that spread?
e. Look at the BMW 5 series. Which quarter has the largest spread of data? What is that spread?
f. Look at the BMW 5 series. Find the Inter Quartile Range (IQR).
g. Look at the BMW 5 series. Are there more data in the interval 31-38 or in the interval 45-55? How do you know this?
h. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

    i. 31-35
    ii. 38-41
    iii. 41-64

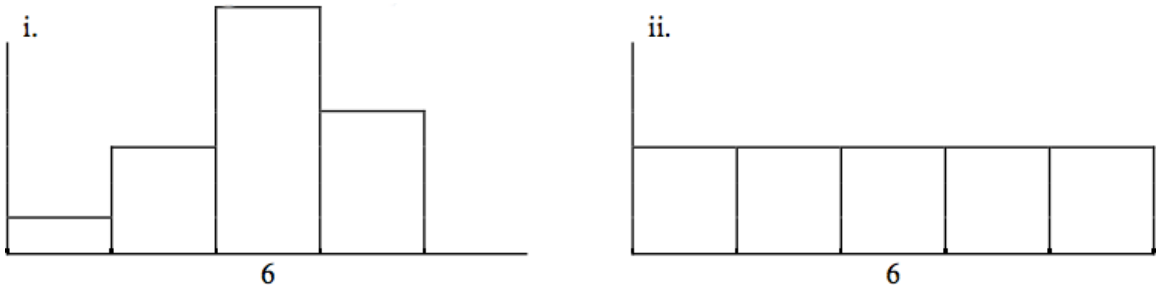
EXERCISE 18

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information:

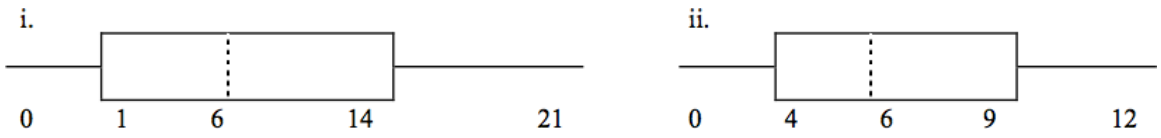
	Javier	Ercilla
$\bar{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct?
b. Explain what the difference in the results of the surveys implies about the data.

c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



EXERCISE 19

Below are the 2008 obesity rates by U.S. states and Washington, DC. (Source: http://www.cdc.gov/obesity/data/trends.html#State)

State	Percent	State	Percent
Alabama	31.4	Montana	23.9
Alaska	26.1	Nebraska	26.6
Arizona	24.8	Nevada	25
Arkansas	28.7	New Hampshire	24
California	23.7	New Jersey	22.9
Colorado	18.5	New Mexico	25.2
Connecticut	21	New York	24.4
Delaware	27	North Carolina	29
Washington DC	21.8	North Dakota	27.1
Florida	24.4	Ohio	28.7
Georgia	27.3	Oklahoma	30.3
Hawaii	22.6	Oregon	24.2
Idaho	24.5	Pennsylvania	27.7
Illinois	26.4	Rhode Island	21.5
Indiana	26.3	South Carolina	30.1
Iowa	26	South Dakota	27.5
Kansas	27.4	Tennessee	30.6
Kentucky	29.8	Texas	28.3

Louisiana	28.3	Utah	22.5
Maine	25.2	Vermont	22.7
Maryland	26	Virginia	25
Massachusetts	20.9	Washington	25.4
Michigan	28.9	West Virginia	31.2
Minnesota	24.3	Wisconsin	25.4
Mississippi	32.8	Wyoming	24.6
Missouri	28.5		

a. Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: The x-axis is labeled with the state names.
b. Use a random number generator to randomly pick 8 states. Construct a bar graph of the obesity rates of those 8 states.
c. Construct a bar graph for all the states beginning with the letter "A."
d. Construct a bar graph for all the states beginning with the letter "M."

EXERCISE 20

A music school has budgeted to purchase 3 musical instruments. They plan to purchase a piano costing $3000, a guitar costing $550, and a drum set costing $600. The average cost for a piano is $4,000 with a standard deviation of $2,500. The average cost for a guitar is $500 with a standard deviation of $200. The average cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer numerically.

EXERCISE 21

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the table below. (Note that this is the data presented for publisher B in homework exercise 13).

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the IQR to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?

c. Are any data values further than 2 standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts (a) and (c) of this problem give the same answer?
e. Examine the shape of the data. Which part, (a) or (c), of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

EXERCISE 22

For each situation below, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than all other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

EXERCISE 23

In 1990 the number of driver deaths per 100,000 for the different age groups was as follows (Source: The National Highway Traffic Safety Administration's National Center for Statistics and Analysis):

Age	Number of driver deaths per 100,000
15 - 24	28
25 - 39	15
40 - 69	10
70 - 79	15
80+	25

a. For each age group, pick the midpoint of the interval for the x value. (For the 80+ group, use 85.)
b. Using “ages” as the independent variable and “Number of driver deaths per 100,000” as the dependent variable, make a scatter plot of the data.
c. Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Pick two ages and find the estimated fatality rates.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between age of a driver and driver fatality rate?

EXERCISE 24

The average number of people in a family that received welfare for various years is given below. (Source: House Ways and Means Committee, Health and Human Services Department)

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

a. Using "year" as the independent variable and "welfare family size" as the dependent variable, make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is it significant?
d. Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
e. Use the two points in (d) to plot the least squares line on your graph from (b).
f. Based on the above data, is there a linear relationship between the year and the average number of people in a welfare family?
g. Using the least squares line, estimate the welfare family sizes for 1960 and 1995. Does the least squares line give an accurate estimate for those years? Explain why or why not.
h. Are there any outliers in the above data?
i. What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.

EXERCISE 25

Use the AIDS data from the practice for this section, but this time use the columns "year #" and "# new AIDS deaths in U.S." Answer all of the questions from the practice again, using the new columns.

EXERCISE 26

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). (Source: *Microsoft Bookshelf*)

Height (in feet)	Stories
1050	57
428	28
362	26
529	40
790	60
401	22

380	38
1454	110
1127	100
700	46

a. Using "stories" as the independent variable and "height" as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
h. Are there any outliers in the above data? If so, which point(s)?
i. What is the estimated height of a building with 6 stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
j. Based on the least squares line, adding an extra story adds about how many feet to a building?

EXERCISE 27

Below is the life expectancy for an individual born in the United States in certain years. (Source: National Center for Health Statistics)

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75.0
1992	75.7

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.

f. Why aren't the answers to part (e) the values on the above chart that correspond to those years?
g. Use the two points in (e) to plot the least squares line on your graph from (b).
h. Based on the above data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the above data?
j. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.

EXERCISE 28

The percent of female wage and salary workers who are paid hourly rates is given below for the years 1979 - 1992. (Source: Bureau of Labor Statistics, U.S. Dept. of Labor)

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

a. Using "year" as the independent variable and "percent" as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated percents for 1991 and 1988.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
h. Are there any outliers in the above data?
i. What is the estimated percent for the year 2050? Does the least squares line give an accurate estimate for that year? Explain why or why not?

EXERCISE 29

The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition 10, for various pages is given below.

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page 10 and on page 70.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
h. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
i. Is the least squares line valid for page 200? Why or why not?

The next two questions refer to the following data:

The cost of a leading liquid laundry detergent in different sizes is given below.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

EXERCISE 30

a. Using "size" as the independent variable and "cost" as the dependent variable, make a scatter plot.

b. Does it appear from inspection that there is a relationship between the variables? Why or why not?

c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$

d. Find the correlation coefficient. Is it significant?

e. If the laundry detergent were sold in a 40 ounce size, find the estimated cost.

f. If the laundry detergent were sold in a 90 ounce size, find the estimated cost.

g. Use the two points in (e) and (f) to plot the least squares line on your graph from (a).

h. Does it appear that a line is the best way to fit the data? Why or why not?

i. Are there any outliers in the above data?

j. Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost? Why or why not?

EXERCISE 31

a. Complete the above table for the cost per ounce of the different sizes.

b. Using "Size" as the independent variable and "Cost per ounce" as the dependent variable, make a scatter plot of the data.

c. Does it appear from inspection that there is a relationship between the variables? Why or why not?

d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$

e. Find the correlation coefficient. Is it significant?

f. If the laundry detergent were sold in a 40 ounce size, find the estimated cost per ounce.

g. If the laundry detergent were sold in a 90 ounce size, find the estimated cost per ounce.

h. Use the two points in (f) and (g) to plot the least squares line on your graph from (b).

i. Does it appear that a line is the best way to fit the data? Why or why not?

j. Are there any outliers in the above data?

k. Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost per ounce? Why or why not?

EXERCISE 32

According to flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000

3,000,000	1,350,000

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. Find the estimated total cost for a net taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Based on the above, what would be the probate fees and taxes for an estate that does not have any assets?

EXERCISE 33

The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.

g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?

EXERCISE 34

Below are the average heights for American boys. (Source: Physician's Handbook, 1990)

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. Find the estimated average height for a one year–old. Find the estimated average height for an eleven year–old.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Use the least squares line to estimate the average height for a sixty–two year–old man. Do you think that your answer is reasonable? Why or why not?

EXERCISE 35

The following chart gives the gold medal times for every other Summer Olympics for the women's 100 meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8

1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from (f) different from the chart values?
h. Use the two points in (f) to plot the least squares line on your graph from (b).
i. Does it appear that a line is the best way to fit the data? Why or why not?
j. Use the least squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

The next three questions use the following information.

Use the following state information for problems 15 – 17.

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado		1876	38	104,100
Hawaii		1959	50	10,932
Iowa		1846	29	56,276
Maryland		1788	7	12,407
Missouri		1821	24	69,709
New Jersey		1787	3	8,722
Ohio		1803	17	44,828
South Carolina	13	1788	8	32,008
Utah		1896	45	84,904
Wisconsin		1848	30	65,499

EXERCISE 36

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

EXERCISE 37

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

a. Let rank be the independent variable and area be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least squares line to estimate the area of a new state that enters the Union. Can the least squares line be used to predict it? Why or why not?
k. Delete “Hawaii” and substitute “Alaska” for it. Alaska is the fortieth state with an area of 656,424 square miles.

l. Calculate the new least squares line.
m. Find the estimated area for Alabama. Is it closer to the actual area with this new least squares line or with the previous one that included Hawaii? Why do you think that's the case?
n. Do you think that, in general, newer states are larger than the original states?

EXERCISE 38

We are interested in whether there is a relationship between the rank of a state and the year it entered the Union.

a. Let year be the independent variable and rank be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Why must the relationship be positive between the variables?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Let's say a fifty-first state entered the union. Based upon the least squares line, when should that have occurred?
g. Using the least squares line, how many states do we currently have?
h. Why isn't the least squares line a good estimator for this year?

EXERCISE 39

Below are the percents of the U.S. labor force (excluding self-employed and unemployed) that are members of a union. We are interested in whether the decrease is significant. (Source: Bureau of Labor Statistics, U.S. Dept. of Labor)

Year	Percent
1945	35.5
1950	31.5
1960	31.4
1970	27.3
1980	21.9
1986	17.5
1993	15.8

a. Let year be the independent variable and percent be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Why will the relationship between the variables be negative?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?

f. Based on your answer to (e), do you think that the relationship can be said to be decreasing?
g. If the trend continues, when will there no longer be any union members? Do you think that will happen?

The next two questions refer to the following: The data below reflects the 1991-92 Reunion Class Giving. (Source: SUNY Albany alumni magazine)

Class Year	Average Gift	Total Giving
1922	41.67	125
1927	60.75	1,215
1932	83.82	3,772
1937	87.84	5,710
1947	88.27	6,003
1952	76.14	5,254
1957	52.29	4,393
1962	57.80	4,451
1972	42.68	18,093
1976	49.39	22,473
1981	46.87	20,997
1986	37.03	12,590

EXERCISE 40

We will use the columns "class year" and "total giving" for all questions, unless otherwise stated.

a. What do you think the scatter plot will look like? Make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. What does it imply about the significance of the relationship?
d. For the class of 1930, predict the total class gift: _____
e. For the class of 1964, predict the total class gift: _____
f. For the class of 1850, predict the total class gift: _____ Why doesn't this value make any sense?

EXERCISE 41

We will use the columns "class year" and "average gift" for all questions, unless otherwise stated.

a. What do you think the scatter plot will look like? Make a scatter plot of the data.

b. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$

c. Find the correlation coefficient. What does it imply about the significance of the relationship?

d. For the class of 1930, predict the total class gift: _____

e. For the class of 1964, predict the total class gift: _____

f. For the class of 2010, predict the total class gift: _____ Why doesn't this value make any sense?

EXERCISE 42

We are interested in exploring the relationship between the weight of a vehicle and its fuel efficiency (gasoline mileage). The data in the table show the weights, in pounds, and fuel efficiency, measured in miles per gallon, for a sample of 12 vehicles.

Weight	Fuel Efficiency
2715	24
2570	28
2610	29
2750	38
3000	25
3410	22
3640	20
3700	26
3880	21
3900	18
4060	18
4710	15

Table 15

A. Graph a scatterplot of the data.

B. Find the correlation coefficient and determine if it is significant.

C. Find the equation of the best fit line.

D. Write the sentence that interprets the meaning of the slope of the line in the context of the data.

E. What percent of the variation in fuel efficiency is explained by the variation in the weight of the vehicles, using the regression line? (State your answer in a complete sentence in the context of the data).

F. Accurately graph the best fit line on your scatterplot.

G. For the vehicle that weighs 3000 pounds, find the residual (y-yhat). Does the value predicted by the line underestimate or overestimate the observed data value?

H. Identify any outliers, using either the graphical or numerical procedure demonstrated in the textbook.

I. The outlier is a hybrid car that runs on gasoline and electric technology, but all other vehicles in the sample have engines that use gasoline only. Explain why it would be appropriate to remove the outlier from the data in this situation. Remove the outlier from the sample data. Find the new correlation coefficient, coefficient of determination, and best fit line.

J. Compare the correlation coefficients and coefficients of determination before and after removing the outlier, and explain in complete sentences what these numbers indicate about how the model has changed.

EXERCISE 43

The four data sets below were created by statistician Francis Anscomb. They show why it is important to examine the scatterplots for your data, in addition to finding the correlation coefficient, in order to evaluate the appropriateness of fitting a linear model.

Set 1		Set 2		Set 3		Set 4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	9	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Table 16

A. For each data set, find the least squares regression line and the correlation coefficient. What did you discover about the lines and values of r?

For each data set, create a scatter plot and graph the least squares regression line. Use the graphs to answer the following questions:

B. For which data set does it appear that a curve would be a more appropriate model than a line?

C. Which data set has an **influential point** (point close to or on the line that greatly influences the best fit line)?

D. Which data set has an **outlier** (obviously visible on the scatter plot with best fit line graphed)?

E. Which data set appears to be the most appropriate to model using the least squares regression line?

Try these multiple choice questions.

EXERCISE 44

A correlation coefficient of -0.95 means there is a _____ between the two variables.

A. Strong positive correlation
B. Weak negative correlation
C. Strong negative correlation
D. No Correlation

EXERCISE 45

According to the data reported by the New York State Department of Health regarding West Nile Virus for the years 2000-2008 **(http://www.health.state.ny.us/nysdoh/westnile/update/update.htm)**, the least squares line equation for the number of reported dead birds (x) versus the number of human West Nile virus cases (y) is **y-hat = 19.2399 + 0.0257x**. If the number of dead birds reported in a year is 732, how many human cases of West Nile virus can be expected? r = 0.5490.

A. No prediction can be made
B. 19.6
C. 15
D. 38.1

The next two questions refer to the following data (showing the number of hurricanes by category to directly strike the mainland U.S. each decade) obtained from www.nhc.noaa.gov/gifs/table6.gif A major hurricane is one with a strength rating of 3, 4 or 5.

Decade	Total Number of Hurricanes	Number of Major Hurricanes
1941-1950	24	10
1951-1960	17	8
1961-1970	14	6
1971-1980	12	4
1981-1990	15	5
1991-2000	14	5
2001 – 2004	**9**	**3**

EXERCISE 46

Using only completed decades (1941 – 2000), calculate the least squares line for the number of major hurricanes expected based upon the total number of hurricanes.

A. y-hat = -1.67x + 0.5
B. y-hat = 0.5x – 1.67
C. y-hat = 0.94x – 1.67
D. y-hat = -2x + 1

EXERCISE 47

The data for 2001-2004 show 9 hurricanes have hit the mainland United States. The line of best fit predicts 2.83 major hurricanes to hit mainland U.S. Can the least squares line be used to make this prediction?

A. No, because 9 lies outside the independent variable values
B. Yes, because, in fact, there have been 3 major hurricanes this decade
C. No, because 2.83 lies outside the dependent variable values
D. Yes, because how else could we predict what is going to happen this decade.