

*Topics covered on Exam 1: Chapters 1-10 in text.*

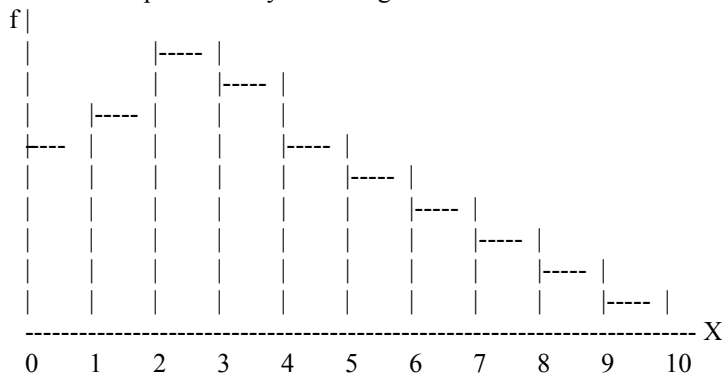
*This material is covered in webassign homework assignments 1 through 4 and worksheets 1-15.*

**Exam information: materials allowed:** calculator (no laptops), one 3 x 5 notecard (2 sided) with notes, definitions, formulas, etc. Standard normal table will be provided with the exam.

**WARNING!** The problems below may not cover all topics for which you are responsible on exam 1.

**Answers are at the end of the document.**

- The distribution of heights of American men aged 18 to 24 can be approximated by a normal model with mean 68 inches and standard deviation 2.5 inches. Half of all young men are shorter than
  - 65.5 inches
  - 68 inches
  - 70.5 inches
  - can't tell because the median height is not given
- Use the information in Problem 1 and the **68-95-99.7 rule** to determine the percentage of young men that are taller than 6' 1".
- The grade point averages (GPA) of 7 randomly chosen students in a statistics class are  
3.14 2.37 2.94 3.60 1.70 4.00 1.85  
Find these statistics:
  - mean
  - median and quartiles
  - range and IQR
- Refer to the information given in the previous problem. If  $\sum(y - \bar{y})^2 = 4.51$ , what is the standard deviation?
- A standardized test designed to measure math anxiety has a mean of 100 and a standard deviation of 10 in the population of first year college students. Which of the following observations would you suspect is an outlier?
  - 150
  - 100
  - 90
  - 125
  - none of the above
- A clerk entering salary data into a company spreadsheet accidentally put an extra "0" in the boss's salary, listing it as \$2,000,000 instead of \$200,000. Explain how this will affect these summary statistics for the company payroll:
  - measures of center: mean and median
  - measures of spread: range, IQR, and standard deviation
- The distribution represented by the histogram below is:



Choose one of the following:

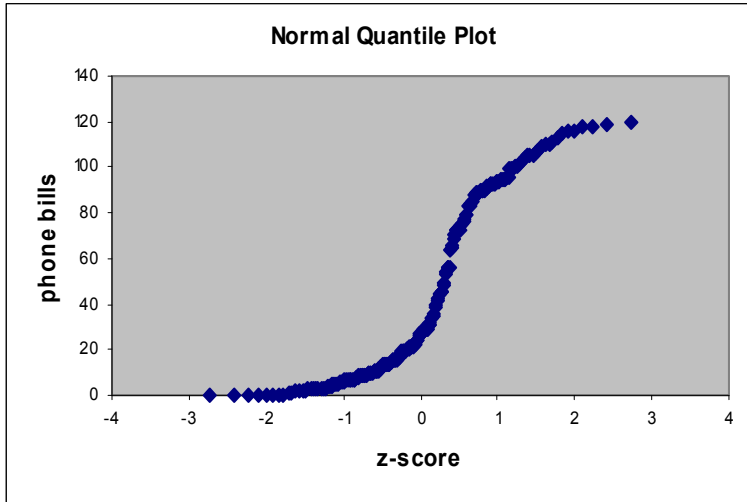
- skewed to the right.
- skewed to the left.
- symmetric.
- normal.

8. Twenty-seven applicants interested in working for the Food Stamp program took an examination designed to measure their aptitude for social work. The following test scores were obtained:  
79, 93, 84, 86, 77, 63, 46, 97, 87, 88, 87, 92, 68, 72, 86, 98, 81, 70, 66, 98, 59, 76, 68, 91, 94, 85, 88.
- Find  $Q_1$ .
  - Construct a boxplot for these observations. Do you observe any outliers?
9. A manufacturer of television sets has found that for the sets they produce, the distribution of lengths of time until the first repair can be approximated by a normal model with a mean of 4.5 years and a standard deviation of 1.5 years. If the manufacturer sets the warrantee so that only 10.2% of the 1st repairs are covered by the warrantee, how long should the warrantee last?
10. Suppose the distribution of the amount of tar in cigarettes can be approximated by a normal model with a mean of 3.5 mg and a standard deviation of 0.5 mg.
- What proportion of cigarettes have a tar content that exceeds 4.25 mg?
  - In order to advertise as a low tar brand, a manufacturer must prove that their tar content is below the 25th percentile of the tar content distribution. Find the 25th percentile of the distribution of tar amounts.
11. Has the percentage of young girls drinking milk changed over time? The following table is consistent with the results from "Beverage Choices of Young Females: Changes and Impact on Nutrient Intakes" (Shanthy A. Bowman, Journal of the American Dietetic Association, 102(9), pp. 1234-1239):

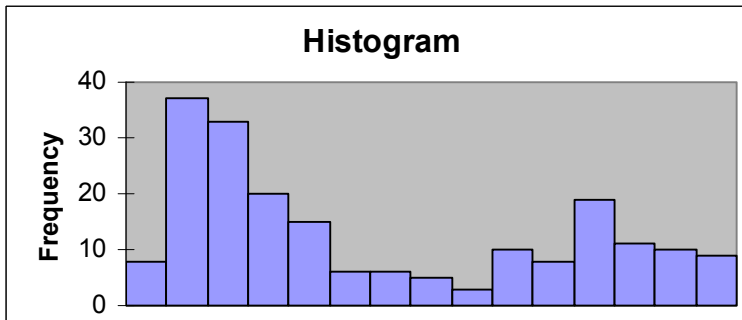
		Nationwide Food Survey Years			Total
		1987-1988	1989-1991	1994-1996	
Drinks Fluid Milk	Yes	354	502	366	1222
	No	226	335	366	927
	<b>Total</b>	<b>580</b>	<b>837</b>	<b>732</b>	<b>2149</b>

- Find the following:
    - What percent of the young girls reported that they drink milk?
    - What percent of the young girls were in the 1989-1991 survey?
    - What percent of the young girls who reported that they drink milk were in the 1989-1991 survey?
    - What percent of the young girls in 1989-1991 reported that they drink milk?
  - What is the marginal distribution of milk consumption?
12. It's the last inning of an important baseball game. The home team is losing by a run, the bases are loaded and the manager needs a pinch hitter. Two batters are available to pinch hit. Here are their statistics:
- | Player | Overall    | vs Left-handed pitching | vs Right-handed pitching |
|--------|------------|-------------------------|--------------------------|
| A      | 33 for 103 | 28 for 81               | 5 for 22                 |
| B      | 45 for 151 | 12 for 32               | 33 for 119               |
- Based on their overall batting averages and their batting averages against right-handed and left-handed pitchers, who would you select as the pinch hitter? What is this phenomenon called?
13. The mean SAT verbal score of next year's freshmen entering the local university is 600. It is also known that 69.5% of these freshmen have scores that are less than 625. If the distribution of scores can be approximated by a normal model, what is the standard deviation of the scores?
14. Two students are enrolled in an introductory statistics course at the University of Florida. The first student is in a morning section and the second student is in an afternoon section. If the student in the morning section takes a midterm and earns a score of 76, while the student in the afternoon section takes a midterm with a score of 72, which student has performed better compared to the rest of the students in his respective class? Assume that both distributions of test scores can be approximated by a normal model. For the morning class, the class mean was 64 with a standard deviation of 8. For the afternoon class, the class mean was 60 with a standard deviation of 7.5.
15. Suppose that a normal model describes the acidity (pH) of rainwater, and the water tested after last week's storm had a z-score of 1.8. This means that the acidity of that rain

- a. had a pH of 1.8
  - b. varied with a standard deviation of 1.8
  - c. had a pH 1.8 higher than the average rainfall
  - d. had a pH 1.8 times that of average rainwater
  - e. had a pH 1.8 standard deviations higher than that of average rainwater
16. The highway gas mileage  $x$ , measured in miles per gallon (mpg), of 26 models of midsize cars, have the following summary statistics:  $\bar{x} = 26.54$  mpg, median = 26 mpg,  $s = 3.04$  mpg,  $IQR = 3$  mpg. If you convert gas mileage  $x$  from miles per gallon to  $x_{new}$  which is measured in miles per liter, what are the new values of the summary statistics? (3.785 liters = 1 gallon).
17. Shown below is the normal probability plot for 200 monthly telephone bills.



Shown below is a histogram. Is this a histogram of the same data that was used to construct the normal probability plot?



18. A local plumber makes house calls. She charges \$30 to come out to the house and \$40 per hour for her services. For example, a 4-hour service call costs  $\$30 + \$40(4) = \$190$ .
- a. The table shows summary statistics for the past month. Fill in the table to find out the cost of the service calls.

Statistic	Hours of Service Call	Cost of Service Call
Mean	4.5	
Median	3.5	
Stan Dev	1.2	
IQR	2.0	
Minimum	0.5	

- b. This past month, the time the plumber spent on a particular service call had a z-score of  $-1.50$ . What is the z-score for the cost of the service call?
19. In 2013 the Department of Education published the Digest for Education Statistics, a collection of information about education in the United States. They reported the average amount (dollars per student) spent by public schools in each state and Washington, D.C. during the school year 2011-2012. The data was recorded according to whether the state lies east or west of the Mississippi River. A back-to-back stem and leaf display of the data is shown below. 6|7 denotes \$6,700.
- a. Which states, Eastern or Western, tend to spend more?
- b. Western states median = ? Eastern states  $Q_1 =$  ?

**Dollars Spent per Student**

Western States	Eastern States
2	4
99	4 5
33210	5 12
9776	5 6699
444200	6 114
7755	6 7788
3	7 2
	7 567789
	8 2
	8 6
0	9 22
	9 9
	10 2
	10

20. A copy machine dealer has data on the number  $x$  of copy machines at each of 89 customer locations and the number  $y$  of service calls in a month at each location. Summary calculations give  $\bar{x} = 8.4$ ,  $s_x = 2.1$ ,  $\bar{y} = 14.2$ ,  $s_y = 3.8$ , and  $r = .86$ . What is the slope of the least squares regression line of number of service calls on number of copiers?
21. In the setting of the previous problem, about what percent of the variation in number of service calls is explained by the linear relation between number of service calls and number of machines?
22. Outdoor temperature influences natural gas consumption for the purpose of heating a house. The usual measure of the need for heating is *heating degree days*. The number of heating degree days for a particular day is the number of degrees the average temperature for that day is below  $65^\circ\text{F}$ , where the average temperature for a day is the mean of the high and low temperatures for that day. An average temperature of  $20^\circ\text{F}$ , for example, corresponds to 45 heating degree days. A homeowner interested in switching to solar

heating panels collects the following data on her natural gas use for the months October through June, where  $x$  is heating degree days per day for the month and  $y$  is gas consumption per day in hundreds of cubic feet.

Month	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June
$x$	15.6	26.8	37.8	36.4	35.5	18.6	15.3	7.9	0
$y$	5.2	6.1	8.7	8.5	8.8	4.9	4.5	2.5	1.1

- Calculate the correlation coefficient  $r$  and interpret its value; draw a scatterplot of the data.
  - Calculate the least squares regression line  $y = b_0 + b_1x$  of gas consumption  $y$  on heating degree days  $x$ . Draw the regression line on the scatterplot.
23. Each of the following statements contains a blunder. In each case explain what is wrong.
- "There is a high correlation between the sex of American workers and their income."
  - "We found a high correlation ( $r = 1.09$ ) between students' ratings of faculty teaching and ratings made by other faculty members."
  - "The correlation between planting rate and yield of corn was found to be  $r = .23$  bushel."
24. A study of 1,000 families gave the following results:  
 average height of husband =  $\bar{x} = 68$  inches; average height of wife =  $\bar{y} = 63$  inches;  
 $\frac{s_y}{s_x} = .925$  wife inches per husband inch;  $r = .25$ .
- Estimate the height of a wife when her husband is 72 inches tall.
    - 63 inches
    - 72 inches
    - 64 inches
    - none of these
    - need more information
  - A sociologist wants to reverse the roles of the variables and use the height of the wife to predict the height of the husband. So wife height is now the  $x$ -variable and husband height is the  $y$ -variable. What is the slope of the new least squares line?

**The information below is needed for questions 25 and 26.**

In finance, million dollar investments are made with the assistance of the Capital Asset Pricing Model (CAPM). The CAPM uses a least squares line to predict the annual rate of return ( $y$ ) of a stock based on the rate of return for the overall stock market ( $x$ ). The slope of the line is used to evaluate the risk of investing in the stock:

- Slope = 1: Average risk (neutral stock)
- Slope > 1: High risk (aggressive stock)
- Slope < 1: Low risk (conservative stock)

The data in the accompanying table are the annual rates of return for Disney stock ( $y$ ) and the rate of return for the overall stock market ( $x$ ) for an 8 year period (where rates of return are measured as a percent). The least squares line for the CAPM model is shown below the table

Year:	83	84	85	86	87	88	89	90
Disney rate of return ( $y$ )	2	-10	18	9	12	-1	-12	2
Overall market rate of return ( $x$ )	1	-5	12	5	7	0	-6	2

Results:  $\hat{y} = -.94 + 1.72x$ , sum of squares of residuals = 7.996

25. Give an interpretation of the slope of the least squares line:
- For a year with an overall market rate of return of 1%, we estimate Disney's rate of return to be 1.72%
  - For every 1.72% increase in overall market rate of return, we estimate Disney's annual rate of return to increase by 1%.
  - For every 1% increase in overall market rate of return, we estimate Disney's annual rate of return to increase by 1.72%.
  - For a year with a market rate of return of 0%, we estimate Disney's rate of return to be -.94%

26. Which of the following is the best interpretation of the sum of the squares of the residuals?
- No other line will produce a sum of squares of the residuals greater than 7.996.
  - The least squares line obtained from these data should yield predictions of Disney's rate of return that are accurate to within  $\pm 2\sqrt{7.996}$ .
  - The small value of 7.996 for the sum of squares of residuals indicates that the straight-line CAPM model is not useful for predicting Disney's rate of return ( $y$ ).
  - No other straight line fit to these data will produce a sum of squares of residuals smaller than 7.996

**Questions 27 and 28 refer to the following:**

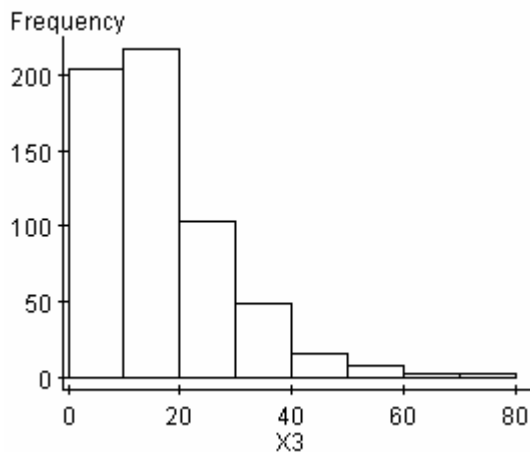
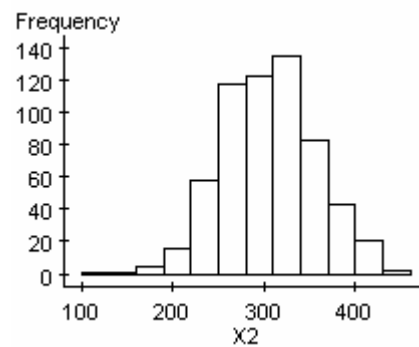
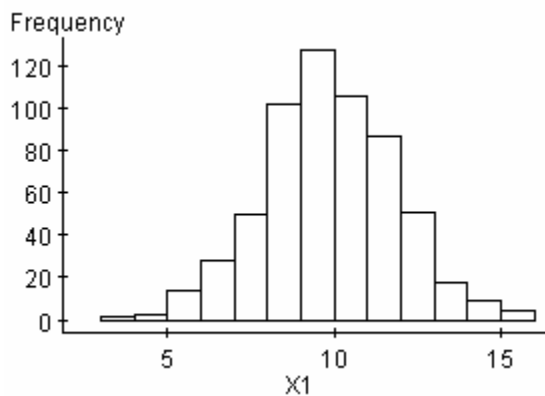
Data were collected to find the relationship between the labor ( $y$  in hours) required to produce lots of custom wood products and the size  $x$  of the lot. The following least squares regression equation was calculated from the data:

$$\hat{y} = 13.7 + 1.7x.$$

27. What is the predicted hours of labor for a lot size of 55?
28. One of the original data points is (20, 50.3). What is the residual when the lot size is 20?

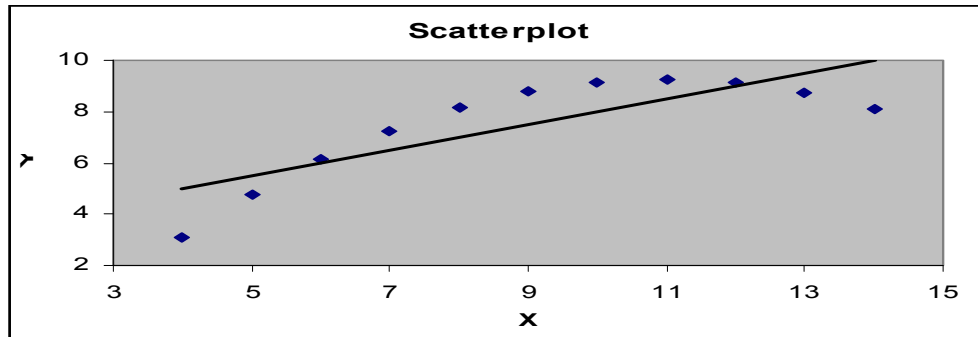
**Questions 29, 30, 31, 32 and 33 refer to the following:**

**Consider the following histograms of variables labeled X1, X2 and X3:**



29. The median for variable X2 would be around  
 a. 10   b. 305   c. 250   d. impossible to tell
30. The third quartile for variable X1 would be around  
 a. 12   b. 8   c. 5   d. 15

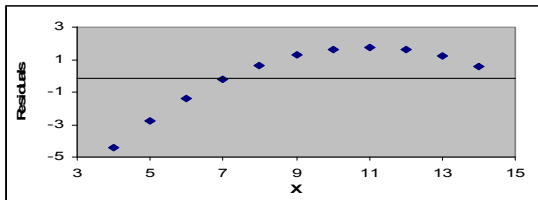
31. The distribution in which the mean and median are most different would be
  - a. X1
  - b. X2
  - c. X3
  - d. It is impossible to tell.
  
32. The standard deviation for variable X1 would be
  - a. About the same as the standard deviation for variable X2.
  - b. Smaller than the standard deviation for variable X2.
  - c. Larger than the standard deviation for variable X2.
  - d. It is impossible to tell.
  
33. The histograms above are the results of questions asked of a group of undergraduate students. Match the histogram (X1, X2, or X3) above to the appropriate question below.
  - a. How many hours did you work at a job last week? \_\_\_\_\_
  - b. What is your shoe size? \_\_\_\_\_
  - c. How much did you spend on textbooks (in dollars) this semester? \_\_\_\_\_
  
34. Shown below is a scatterplot with the corresponding least squares line.



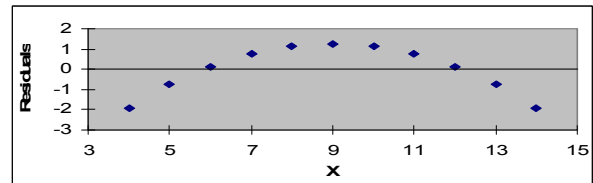
Choose the residual plot that corresponds to this scatterplot and least squares line.

- a. I    b. II    c. III    d. IV    e. none

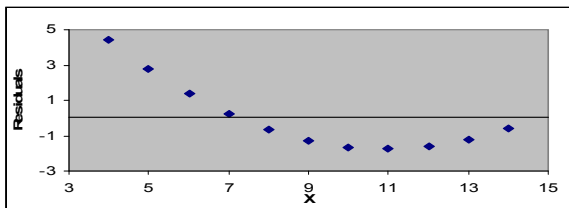
I.



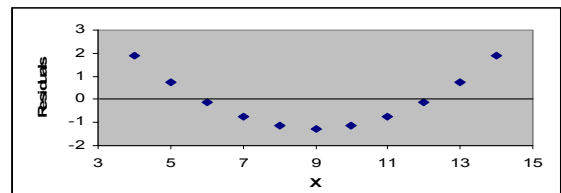
II



III



IV

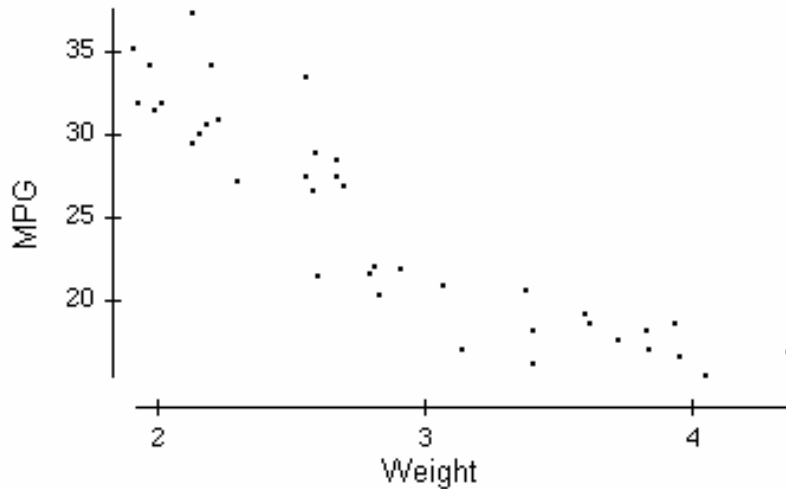


35. When every possible sample with the same number of observations is equally likely to be chosen, the selected sample is called a:
  - a. simple random sample.
  - b. stratified random sample
  - c. cluster random sample
  - d. systematic random sample

36. When the population is divided into mutually exclusive sets, and then a simple random sample is drawn from each set, this is called:
- simple random sampling.
  - stratified random sampling.
  - cluster random sampling.
  - systematic random sampling.
37. A marketing research firm divides the population of a state into geographic areas, and randomly selects some of the areas and takes a simple random sample of each selected area. This is an example of a
- cluster random sample
  - systematic random sample
  - simple random sample
  - stratified random sample
38. A simple random sample of 20 undergraduates at Johns Hopkins University found that 60% of those sampled felt that drinking was a problem among college students. A simple random sample of 20 undergraduates at Ohio State University found that 70% felt that drinking was a problem among college students. The number of undergraduates at Johns Hopkins University is approximately 2,000; the number at Ohio State is approximately 50,000. We conclude
- the sample from Johns Hopkins is much less representative of its population than the sample from Ohio State is of its population.
  - the sample from Johns Hopkins is much more representative of its population than the sample from Ohio State is of its population.
  - the samples from Johns Hopkins and Ohio State are equally representative of their respective populations.
  - it is impossible to make any statements about which sample is more representative of its corresponding population since the students surveyed attended different schools.
39. A simple random sample of 2% of the undergraduates at Johns Hopkins University found that 60% of those sampled felt that drinking was a problem among college students. A simple random sample of 2% of the undergraduates at Ohio State University found that 70% felt that drinking was a problem among college students. The number of undergraduates at Johns Hopkins University is approximately 2,000; the number at Ohio State is approximately 50,000. We conclude
- the sample from Johns Hopkins is much less representative of its population than the sample from Ohio State is of its population.
  - the sample from Johns Hopkins is much more representative of its population than the sample from Ohio State is of its population.
  - the samples from Johns Hopkins and Ohio State are equally representative of their respective populations.
  - it is impossible to make any statements about which sample is more representative of its corresponding population since the students surveyed attended different schools.



40. Consider the following scatterplot.



Which of the following is a plausible value for the correlation coefficient between weight and MPG?

- a. -0.9   b. -1.0   c. +0.2   d. +0.9   e. +0.7

41. If a least squares line was fit to the data shown in the above scatterplot, would the slope of the least squares be positive or negative?

- a. positive                      b. negative

## ANSWERS

1. b    2. 2.5% (6' 1"=73" is how many standard deviations above the mean?)  
 3. a) 2.8   b) median 2.94; quartiles: since there are 7 (odd) observations, include the overall median in each half of the data)  $Q_1$  is the median of the smallest 4 observations so  $Q_1 = \frac{1.85+2.37}{2} = 2.11$ ;  $Q_3$  is the median of the largest 4 observations so  $Q_3 = \frac{3.14+3.6}{2} = 3.37$  c) range =  $4 - 1.7 = 2.3$ ; IQR =  $3.37 - 2.11 = 1.26$   
 4. .87    5. a  
 6. a) the median will probably be unaffected; the mean will be larger   b) the range and standard deviation will increase, the IQR will be unaffected    7. Skewed to the right.  
 8. a. The first step is to order the data. Then compute the overall median. Since there are 27 observations, the median is the observation in position 14: median = 85. Compute  $Q_1$ : we want the median of the lower half. Since we have an odd number of observations (27), **include the overall median in both halves of the data**. There are 14 observations in the lower half, including the overall median. The median of these lower 14 observations is the mean of the 2 middle observations in positions 7 and 8, so  $Q_1 = \frac{70+72}{2} = 71$ .  
 b. Note that  $Q_3$  is the median of the 14 observations in the upper half, including the overall median. So  $Q_3$  is the mean of the 2 middle observations in positions 20 and 21:  $Q_3 = \frac{88+91}{2} = 89.5$ .  
 IQR =  $Q_3 - Q_1 = 89.5 - 71 = 18.5$ ;  $1.5 * IQR = 27.75$ .  
Boundaries for outliers:  $Q_1 - 1.5 * IQR = 71 - 27.75 = \boxed{43.25}$ ;  $Q_3 + 1.5 * IQR = 89.5 + 27.75 = \boxed{117.25}$ .  
 Since the smallest observation is 46 and the largest observation is 98, there are no outliers. See boxplot below (the diamonds above the box represent the individual data values).



9.  $z = -1.27$ ;  $x = 2.595$  years. 10. a. .0668 b.  $z = -0.675$ ;  $x = 3.16$ . 11. a1. 56.9% a2. 38.9% a3. 41.1% a4. 60% b. Yes: 56.9%; No: 43.1%. 12. Player A overall batting avg. = .320; Player B overall batting avg. = .298. **Choose player A.** Player A vs right-handed pitchers = .227, Player B vs right-handed pitchers = .277; Player A vs left-handed pitchers = .346; Player B vs left-handed pitchers = .375. Player B has the higher batting average against both right-handed and left-handed pitchers; **choose Player B.** Simpson's paradox. 13.  $0.51 = (625 - 600) / \sigma \Rightarrow \sigma = 49.02$  14.  $z_1 = (76 - 64) / 8 = 1.5$ ;  $z_2 = (72 - 60) / 7.5 = 1.6$ . The student in the afternoon section performed better. 15. e. 16.  $\bar{x}_{new} = 7.01$  miles per liter;  $median_{new} = 6.87$  miles per liter;  $s_{new} = .803$  miles per liter;  $IQR_{new} = .793$  miles per liter. 17. Yes. Note upward curvature on left portion of the normal probability plot (bills are not less than \$0); plot very "steep" in central portion, which means there are not many observations in middle of data; downward curvature in right portion of plot indicates that the data has a shorter right tail than a normal distribution. 18. a.

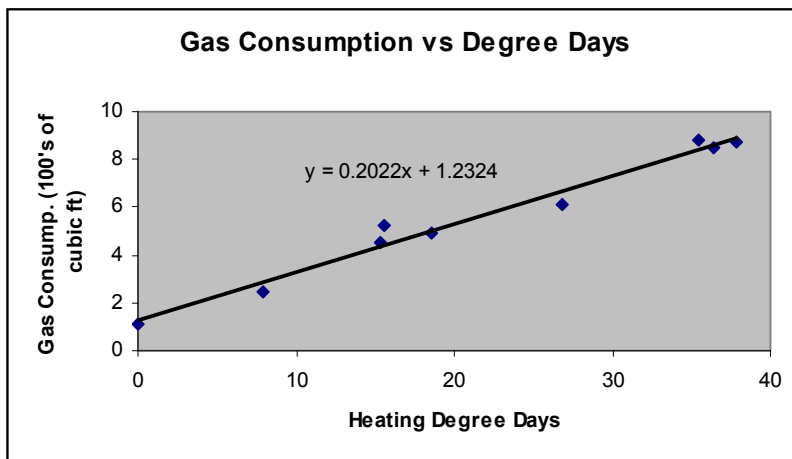
Statistic	Hours of Service Call	Cost of Service Call
Mean	4.5	\$210
Median	3.5	\$170
Stan Dev	1.2	\$48
IQR	2.0	\$80
Minimum	0.5	\$50

b. - 1.50

19. a. Eastern b. West median = \$5,950, East  $Q_1 = \$6,000$

20.  $b = r \left( \frac{s_y}{s_x} \right) = .86 \left( \frac{3.8}{2.1} \right) = 1.56$ . 21.  $r^2 = (.86)^2 = .74$

22. a)  $r = .989$ . There is a strong positive linear relationship between heating degree days and gas consumption. b).  $s_x = 13.419$ ;  $s_y = 2.74$ ; slope  $b_1 = .202$ , intercept  $a = \bar{y} - b\bar{x} = 1.23$



23. a. The correlation we are studying measures the linear relationship between 2 quantitative

variables; sex is a categorical variable.

- b.  $-1 \leq r \leq 1$  is violated.  
 c.  $r$  has no units.

**24A** c. **method 1:**  $b_1 = r * \left(\frac{s_y}{s_x}\right) = .25 * (0.925) = \boxed{.231}$ ;  $b_0 = \bar{y} - b_1 \bar{x} = 63 - .231 * 68 = \boxed{47.292}$ ;  
 $\hat{y} = 47.292 + .231 * 72 = 63.924 \approx 64$ .

**method 2:** The husband is 4 inches, or  $4/2.7 = 1.5$  standard deviations above the mean husband height. The wife's height is predicted to be above average by  $.25(1.5) = .4$  standard deviations, or  $.4 \times 2.5$  inches = 1 inch. Recall  $b_1 = r * \frac{s_y}{s_x}$ . So the wife's height is  $63 + 1 = 64$ .

**24B** Did you guess *new slope* =  $\frac{1}{\text{original slope}} = \frac{1}{.231} = 4.33$ ? **WRONG!** The roles of  $x$  and  $y$  are reversed so  
*new slope* =  $r * \left(\frac{s_x}{s_y}\right) = .25 * \left(\frac{1}{0.925}\right) = .25 * (1.081) = \boxed{.27}$ .

25. c. 26. d 27. 107.2 28. observed  $y$  – predicted  $y = 2.6$  hours. 29. b 30. a 31. c 32. b

33. a. X3 b. X1 c. X2 34. b 35. a 36. b 37. a 38. c 39. a 40. a 41. b. negative (from the formula  $b_1 = r * \frac{s_y}{s_x}$  for the slope of the least squares line, the correlation and the slope always have the same sign since the standard deviations  $s_y$  and  $s_x$  are positive).