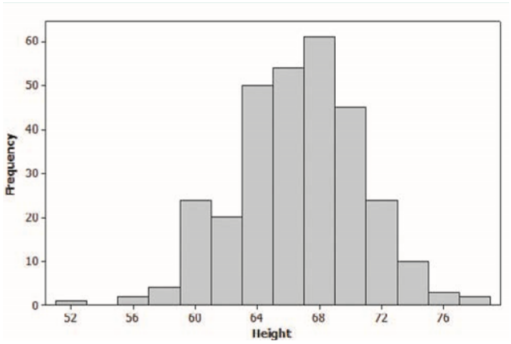
Chapter 1

One-variable data analysis

- **Multiple choice**

Section 1, Section 2, Section 3

1. The histogram below shows the heights of 300 randomly selected high school students. Which of the following is the best description of the shape of the distribution of heights?

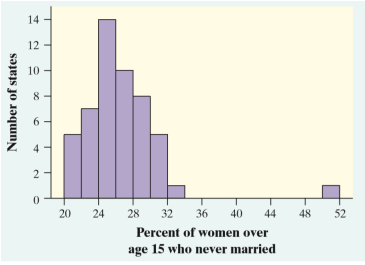


   (a) Roughly symmetric and unimodal
   (b) Roughly symmetric and bimodal
   (c) Skewed to the left
   (d) Skewed to the right

2. You look at real estate ads for houses in Naples, Florida. There are many houses ranging from $200,000 to $500,000 in price. The few houses on the water, however, have prices up to $15 million. The distribution of house prices will be

   (a) skewed to the left
   (b) roughly symmetric
   (c) skewed to the right
   (d) of two clusters

3. The following histogram shows the distribution of the percents of women aged 15 and over who have never married in each of the 50 states and the District of Columbia. Which of the following statements about the histogram is correct?



   (a) The center of the distribution is about 36%.
   (b) There are more states with percents above 32 than there are states with percents less than 24.

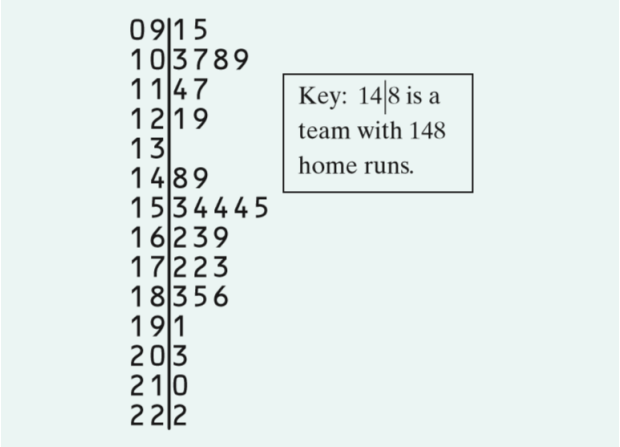
(c) It would be better if the values from 34 to 50 were deleted on the horizontal axis so there wouldnt be a large gap.
(d) There was one state with a value of exactly 33%.
(e) About half of the states had percents between 24% and 28%.

4. When comparing two distributions, it would be best to use relative frequency histograms rather than frequency histograms when
   (a) the distributions have different shapes.
   (b) the distributions have different spreads.
   (c) the distributions have different centers.
   (d) the distributions have different numbers of observations.
   (e) at least one of the distributions have outliers.

5. Which of the following is the best reason for choosing a stemplot rather than a histogram to display the distribution of a quantitative variable?
   (a) Stemplots allow you to split stems; histograms dont.
   (b) Stemplots allow you to see the values of individual observations
   (c) Stemplots are better for displaying very large sets of data.
   (d) Stemplots never require rounding of values.
   (e) Stemplots make it easier to determine the shape of a distribution.

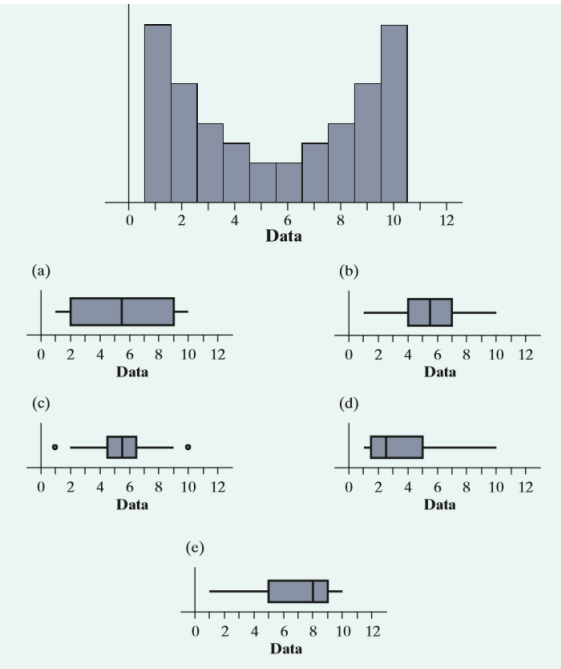
6. The scores on a statistics test had a mean of 81 and a standard deviation of 9. One student was absent on the test day, and his score wasnt included in the calculation. If his score of 84 was added to the distribution of scores, what would happen to the mean and standard deviation?
   (a) Mean will increase, and standard deviation will increase.
   (b) Mean will increase, and standard deviation will decrease.
   (c) Mean will increase, and standard deviation will stay the same.
   (d) Mean will decrease, and standard deviation will increase.
   (e) Mean will decrease, and standard deviation will decrease.

7. The stemplot shows the number of home runs hit by each of the 30 Major League Baseball teams in 2011. Home run totals above what value should be considered outliers?



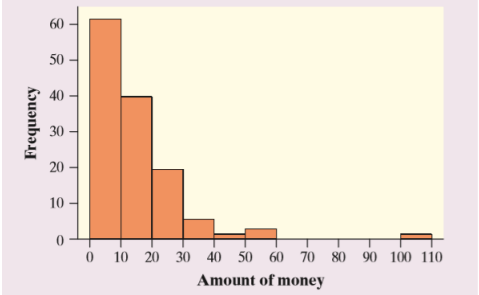
(a) 173 (b) 210 (c) 222, (d) 229 (e) 257

8. Which of the following boxplots best matches the distribution shown in the histogram?



9. You record the age, marital status, and earned income of a sample of 1463 women. The number and type of variables you have recorded is
   (a) 3 quantitative, 0 categorical.
   (b) 4 quantitative, 0 categorical.
   (c) 3 quantitative, 1 categorical.
   (d) 2 quantitative, 1 categorical.
   (e) 2 quantitative, 2 categorical.

10. In a statistics class with 136 students, the professor records how much money (in dollars) each student has in his or her possession during the first class of the semester. The histogram shows the data that were collected.

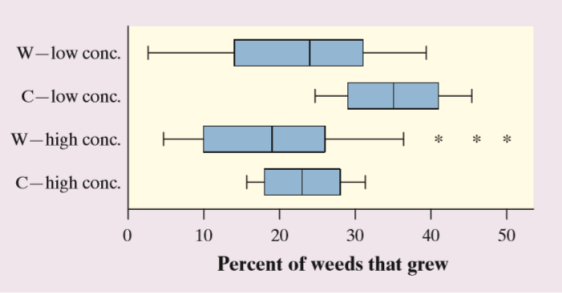


Which of the following statements about this distribution is not correct?
   (a) The histogram is right-skewed.
   (b) The median is less than $20.
   (c) The IQR is $35.

(d) The mean is greater than the median.

(e) The histogram is unimodal.

11. An experiment was conducted to investigate the effect of a new weed killer to prevent weed growth in onion crops. Two chemicals were used: the standard weed killer (C) and the new chemical (W). Both chemicals were tested at high and low concentrations on a total of 50 test plots. The percent of weeds that grew in each plot was recorded. Here are some boxplots of the results. Which of the following is not a correct statement about the results of this experiment



(a) At both high and low concentrations, the new chemical (W) gives better weed control than the standard weed killer (C).

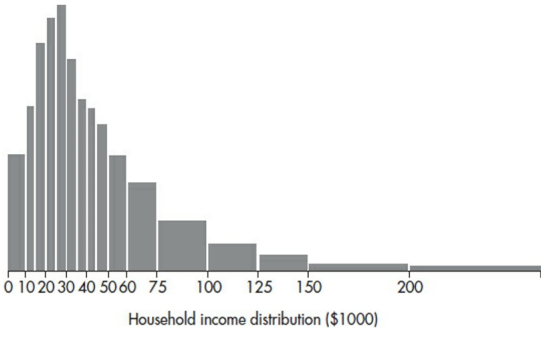
(b) Fewer weeds grew at higher concentrations of both chemicals.

(c) The results for the standard weed killer (C) are less variable than those for the new chemical (W).

(d) High and low concentrations of either chemical have approximately the same effects on weed growth.

(e) Some of the results for the low concentration of weed killer W show fewer weeds growing than some of the results for the high concentration of W.

12. The graph below shows household income in Laguna Woods, California.



What can be said about the ratio $\frac{\textbf{Mean family income}}{\textbf{Median family income}}$?

(a) Approximately zero

(b) Less than one but definitely above zero

(c) Approximately one

(d) Greater than one

(e) Cannot be answered without knowing the standard deviation

13. Which of the following are true statements?

I. The range of the sample data set is never greater than the range of the population.

II. The interquartile range is half the distance between the first quartile and the third quartile.

III. While the range is affected by outliers, the interquartile range is not.

(a) I only

(b) II only

(c) III only

(d) I and II

(e) I and III

14. Dieticians are concerned about sugar consumption in teenagers diets (a 12-ounce can of soft drink typically has 10 teaspoons of sugar). In a random sample of 55 students, the number of teaspoons of sugar consumed for each student on a randomly selected day is tabulated. Summary statistics are noted below:

Min = 10	Max = 60	$Q_1 = 25$	Median =31
$Q_3 = 38$	Mean = 31.4	n = 55	s = 11.6

Which of the follwing is a true statement?

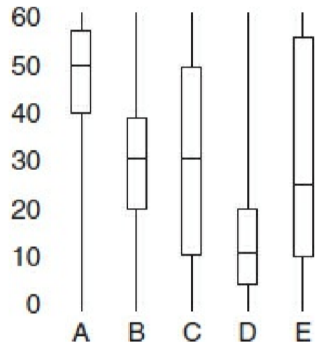
(a) None of the values are outliers.

(b) The value 10 is an outlier, and there can be no others.

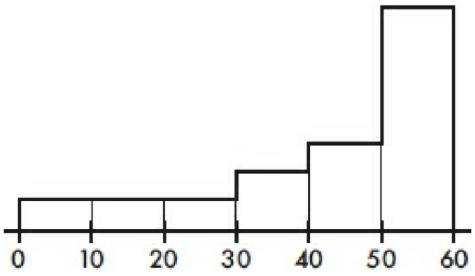
(c) The value 60 is an outlier, and there can be no others.

(d) Both 10 and 60 are outliers, and there may be others.

(e) The value 60 is an outlier, and there may be others at the high end of the data set.

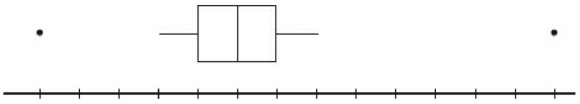


15. To which of the above boxplots does the following histogram correspond?



(a) A (b) B (c) C (d) D (e) E

16. Below is a boxplot of yearly tuition and fees of all four year colleges and universities in a Western state. The low outlier is from a private university that gives full scholarships to all accepted students, while the high outlier is from a private college catering to the very rich.



Removing both outliers will effect what changes, if any, on the mean and median costs for this states four year institutions of higher learning?

(a) Both the mean and the median will be unchanged.
(b) The median will be unchanged, but the mean will increase.
(c) The median will be unchanged, but the mean will decrease.
(d) The mean will be unchanged, but the median will increase.
(e) Both the mean and median will change.

17. If quartiles $Q_1 = 20$ and $Q_3 = 30$, which of the following must be true?

I. The median is 25.
II. The mean is between 20 and 30.
III. The standard deviation is at most 10.

(a) I only.
(b) II only.
(c) III only.
(d) All are true.
(e) None are true.

18. A 1995 poll by the Program for International Policy asked respondents what percentage of the U.S. budget they thought went to foreign aid. The mean response was 18%, and the median was 15%. (The actual amount is less than 1%.) What do these responses indicate about the likely shape of the distribution of all the responses?

(a) The distribution is skewed to the left.
(b) The distribution is skewed to the right.

(c) The distribution is symmetric around 16.5%.
(d) The distribution is bell-shaped with a standard deviation of 3%.
(e) The distribution is uniform between 15% and 18%.

19. Which of the following are true statements?

I. If the sample has variance zero, the variance of the population is also zero.
II. If the population has variance zero, the variance of the sample is also zero.
III. If the sample has variance zero, the sample mean and the sample median are equal.

(a) I and II
(b) II and III
(c) I and III
(d) I, II and III
(e) None of the above

20. When a set of data has suspect outliers, which of the following are preferred measures of central tendency and of variability?

(a) mean and standard deviation
(b) mean and variance
(c) mean and range
(d) median and range
(e) median and interquartile range

21. The 70 highest dams in the world have an average height of 206 meters with a standard deviation of 35 meters. The Hoover and Grand Coulee dams have heights of 221 and 168 meters, respectively. The Russian dams, the Nurek and Charvak, have heights with z-scores of +2.69 and 1.13, respectively. List the dams in order of ascending size.

(a) Charvak, Grand Coulee, Hoover, Nurek
(b) Charvak, Grand Coulee, Nurek, Hoover
(c) Grand Coulee, Charvak, Hoover, Nurek
(d) Grand Coulee, Charvak, Nurek, Hoover
(e) Grand Coulee, Hoover, Charvak, Nurek

22. Jorges score on Exam 1 in his statistics class was at the 64th percentile of the scores for all students. His score falls

(a) between the minimum and the first quartile.
(b) between the first quartile and the median.
(c) between the median and the third quartile.
(d) between the third quartile and the maximum.
(e) at the mean score for all students.

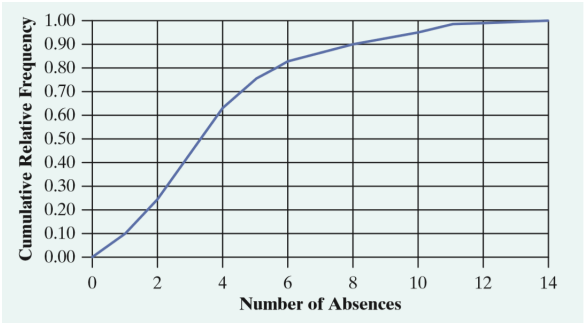
23. When Sam goes to a restaurant, he always tips the server \$2 plus 10% of the cost of the meal. If Sams distribution of meal costs has a mean of \$9 and a standard deviation of \$3, what are the mean and standard deviation of the distribution of his tips?

    (a) \$2.90, \$0.30
    (b) \$2.90, \$2.30
    (c) \$9.00, \$3.00
    (d) \$11.00, \$2.00
    (e) \$2.00, \$9.00

24. George has an average bowling score of 180 and bowls in a league where the average for all bowlers is 150 and the standard deviation is 20. Bill has an average bowling score of 190 and bowls in a league where the average is 160 and the standard deviation is 15. Who ranks higher in his own league, George or Bill?

    (a) Bill, because his 190 is higher than Georges 180.
    (b) Bill, because his standardized score is higher than Georges.
    (c) Bill and George have the same rank in their leagues, because both are 30 pins above the mean.
    (d) George, because his standardized score is higher than Bills.
    (e) George, because the standard deviation of bowling scores is higher in his league.

The following two problems refer to the following setting. The number of absences during the fall semester was recorded for each student in a large elementary school. The distribution of absences is displayed in the following cumulative relative frequency graph.



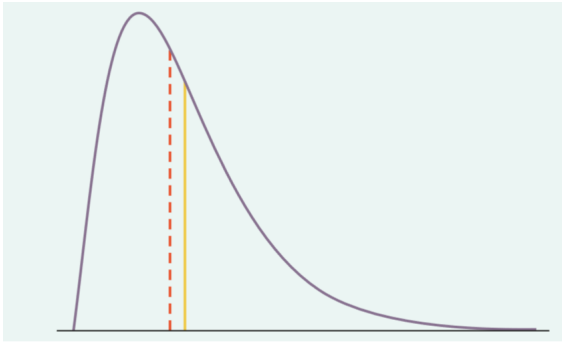
25. What is the interquartile range (IQR) for the distribution of absences?

    (a) 1 (b) 2 (c) 3 (d) 5 (e) 14

26. If the distribution of absences was displayed in a histogram, what would be the best description of the histograms shape?

    (a) Symmetric
    (b) Uniform
    (c) left-skewed
    (d) right-skewed
    (e) can not be determined

27. Two measures of center are marked on the density curve shown. Which of the following is correct?

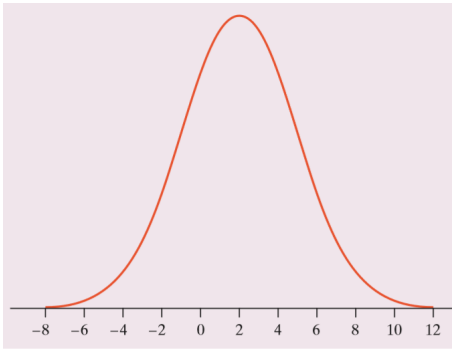


(a) The median is at the solid line and the mean is at the dashed line.
(b) The median is at the dashed line and the mean is at the solid line.
(c) The mode is at the dashed line and the median is at the solid line.
(d) The mode is at the solid line and the median is at the dashed line.
(e) The mode is at the dashed line and the mean is at the solid line.

28. Many professional schools require applicants to take a standardized test. Suppose that 1000 students take such a test. Several weeks after the test, Pete receives his score report: he got a 63, which placed him at the 73rd percentile. This means that

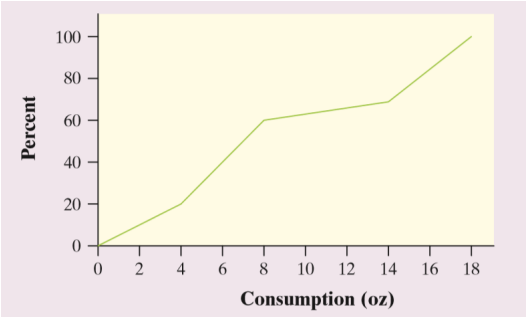
(a) Petes score was below the median.
(b) Pete did worse than 63% of the test takers.
(c) Pete did worse than 73% of the test takers.
(d) Pete did better than 63% of the test takers.
(e) Pete did better than 73% of the test takers.

29. For the Normal distribution shown, the standard deviation is closest to



(a) 0 (b) 1 (c) 2 (d) 3 (e) 5

30. The figure shows a cumulative relative frequency graph of the number of ounces of alcohol consumed per week in a sample of 150 adults who report drinking alcohol occasionally. About what percent of these adults consume between 4 and 8 ounces per week?



(a) 20% (b) 40% (c) 50% (d) 60% (e) 80%

31. Which of the following is not correct about a standard Normal distribution

    (a) The proportion of scores that satisfy $0 < z < 1.5$ is 0.4332.
    (b) The proportion of scores that satisfy $z < 1.0$ is 0.1587.
    (c) The proportion of scores that satisfy $z > 2.0$ is 0.0228.
    (d) The proportion of scores that satisfy $z < 1.5$ is 0.9332.
    (e) The proportion of scores that satisfy $z > 3.0$ is 0.9938.

The next two problems refer to the following setting. Until the scale was changed in 1995, SAT scores were based on a scale set many years ago. For Math scores, the mean under the old scale in the 1990s was 470 and the standard deviation was 110. In 2009, the mean was 515 and the standard deviation was 116.

32. What is the standardized score (z-score) for a student who scored 500 on the old SAT scale?

    (a) -30 (b) -0.27 (c) -0.13 (d) 0.13 (e) 0.27

33. Gina took the SAT in 1994 and scored 500. Her cousin Colleen took the SAT in 2013 and scored 530. Who did better on the exam, and how can you tell?

    (a) Colleen, she scored 30 points higher than Gina
    (b) Colleen, her standardized score is higher than Ginas.
    (c) Gina, her standardized score is higher than Colleens.
    (d) Gina, the standard deviation was bigger in 2013.
    (e) The two cousins did equally welltheir z-scores are the same.

34. Which of the following can be used to describe the center of a data set?

    (a) $\frac{Q_1 - Q_3}{2}$
    (b) $\frac{Q_1 + Q_3}{2}$
    (c) $\frac{\text{range}}{2}$
    (d) $\frac{\text{Standard deviation}}{2}$

35. A set of 5,000 scores on a college readiness exam are known to be approximately normally distributed with mean 72 and standard deviation 6. To the nearest integer value, how many scores are there between 63 and 75?

(a) 0.6247
(b) 4,115
(c) 3,650
(d) 3,123
(e) 3,227

36. Free-response questions on the AP Statistics Exam are graded on 4, 3, 2, 1 or 0 basis. Question # 2 on the exam was of moderate difficulty. The average score on question # 2 was 2.05, with a standard deviation of 1. To the nearest tenth, what score was achieved by a student who was at the 90th percentile of all students on the test? You may assume that the scores on the question were approximately normally distributed.

(a) 3.5
(b) 3.3
(c) 2.9
(d) 3.7
(e) 3.1

- **Free Response**

1. **Do pets or friends help reduce stress**

   If you are a dog lover, having your dog with you may reduce your stress level. Does having a friend with you reduce stress? To examine the effect of pets and friends in stressful situations, researchers recruited 45 women who said they were dog lovers. Fifteen women were assigned at random to each of three groups: to do a stressful task alone, with a good friend present, or with their dogs present. The stressful task was to count backward by 13s or 17s. The womans average heart rate during the task was one measure of the effect of stress. The table below shows the data.

Average heart rates during stress with a pet (P), with a friend (F), and for the control group (C)

GROUP	RATE	GROUP	RATE	GROUP	RATE	GROUP	RATE
P	69.169	P	68.862	C	84.738	C	75.477
F	99.692	C	87.231	C	84.877	C	62.646
P	70.169	P	64.169	P	58.692	P	70.077
C	80.369	C	91.754	P	79.662	F	88.015
C	87.446	C	87.785	P	69.231	F	81.600
P	75.985	F	91.354	C	73.277	F	86.985
F	83.400	F	100.877	C	84.523	F	92.492
F	102.154	C	77.800	C	70.877	P	72.262
P	86.446	P	97.538	F	89.815	P	65.446
F	80.277	P	85.000	F	98.200		
C	90.015	F	101.062	F	76.908		
C	99.046	F	97.046	P	69.538		

Use what you have learnt in this chapter to analyze the data.

(a) Construct an appropriate graph for comparing the heart rates of the women in the three groups.

(b) Calculate numerical summaries for each groups data. Which measures of center and spread would you choose to compare? Why?

(c) Determine if there are any outliers in each of the three groups. Show your work.

(d) Write a few sentences comparing the distributions of heart rates for the women in the three groups.

(e) Based on the data, does it appear that the presence of a pet or friend reduces heart rate during a stressful task? Justify your answer.

2. **I think I can**

   An important measure of the performance of a locomotive is its adhesion, which is the locomotives pulling force as a multiple of its weight. The adhesion of one 4400-horsepower diesel locomotive varies in actual use according to a Normal distribution with mean $\mu = 0.37$ and standard deviation $\sigma = 0.04$.

   (a) For a certain small trains daily route, the locomotive needs to have an adhesion of at least 0.30 for the train to arrive at its destination on time. On what proportion of days will this happen? Show your method.

   (b) An adhesion greater than 0.50 for the locomotive will result in a problem because the train will arrive too early at a switch point along the route. On what proportion of days will this happen? Show your method.

   The locomotives manufacturer is considering two changes that could reduce the percent of times that the train arrives late. One option is to increase the mean adhesion of the locomotive. The other possibility is to decrease the variability in adhesion from trip to trip, that is, to reduce the standard deviation.

   (c) If the standard deviation remains at $\sigma = 0.04$, to what value must the manufacturer change the mean adhesion of the locomotive to reduce its proportion of late arrivals to only 2% of days? Show your work.

   (d) If the mean adhesion stays at $\mu = 0.37$, how much must the standard deviation be decreased to ensure that the train will arrive late only 2% of the time? Show your work.

   (e) Which of the two options in parts (a) and (b) do you think is preferable? Justify your answer. (Be sure to consider the effect of these changes on the percent of days that the train arrives early to the switch point.)

3. The distribution of scores on a recent test closely followed a Normal distribution with a mean of 22 points and a standard deviation of 4 points.

   (a) What proportion of the students scored at least 25 points on this test?

   (b) What is the 31st percentile of the distribution of test scores?

   (c) The teacher wants to transform the test scores so that they have an approximately Normal distribution with a mean of 80 points and a standard deviation of 10 points. To do this, she will use a formula in the form:

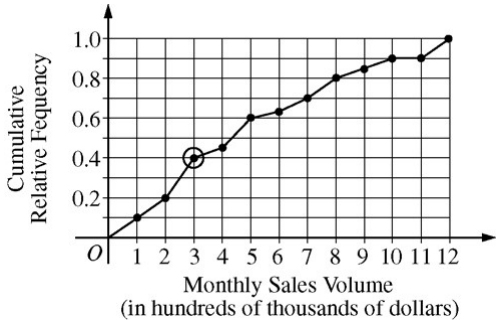
   $$newscore = a + b(oldscore)$$

   Find the values of a and b that the teacher should use to transform the distribution of test scores.

   (d) Before the test, the teacher gave a review assignment for homework. The maximum score on the assignment was 10 points. The distribution of scores on this assignment had a mean of 9.2 points and a standard deviation of 2.1 points. Would it be appropriate to use a Normal distribution to calculate the proportion of students who scored below 7 points on this assignment? Explain.

4. **2006FRB1**

A large regional real estate company keeps records of home sales for each of its sales agents. Each month, the company publishes the sales volume for each agent. Monthly sales volume is defined as the total sales price of all homes sold by the agent during a month. The figure below displays the cumulative relative frequency plot of the most recent monthly sales volume (in hundreds of thousands of dollars) for these agents.



(a) In the context of this question, explain what information is conveyed by the circled point.

(b) What proportion of sales agents achieved monthly sales volumes between \$700,000 and \$800,000?

(c) For values between 10 and 11 on the horizontal axis, the cumulative relative frequency plot is flat. In the context of this question, explain what this means.

(d) A bonus is to be given to 20 percent of the sales agents. Those who achieved the highest monthly sales volume during the preceding month will receive a bonus. What is the minimum monthly sales volume an agent must have achieved to qualify for the bonus?

Chapter 2

Two-variable data analysis

Multiple Choice

1. In the 2010–2011 season, the Dallas Mavericks won the NBA championship. The two-way table below displays the relationship between the outcome of each game in the regular season and whether the Mavericks scored at least 100 points.

	100 or more points	Fewer than 100 points	Total
Win	43	14	**57**
Loss	4	21	**25**
Total	**47**	**35**	**82**

Which of the following is the best evidence that there is an association between the outcome of a game and whether or not the Mavericks scored at least 100 points?

(a) The Mavericks won 57 games and lost only 25 games.

(b) The Mavericks scored at least 100 points in 47 games and fewer than 100 points in only 35 games.

(c) The Mavericks won 43 games when scoring at least 100 points and only 14 games when scoring fewer than 100 points.

(d) The Mavericks won a higher proportion of games when scoring at least 100 points (43/47) than when they scored fewer than 100 points (14/35).

(e) The combination of scoring 100 or more points and winning the game occurred more often (43 times) than any other combination of outcomes.

2. The following partially complete two-way table shows the marginal distributions of gender and handedness for a sample of 100 high school students.

	Male	Female	Total
Right	x		**90**
Left			**10**
Total	**40**	**60**	**100**

If there is no association between gender and handedness for the members of the sample, which of the following is the correct value of x?

(a) 20 (b) 30 (c) 36 (d) 45

(f) Impossible to determine without more information.

3. You have data for many years on the average price of a barrel of oil and the average retail price of a gallon of unleaded regular gasoline. If you want to see how well the price of oil predicts the price of gas, then which should be taken as the explanatory variable?

(a) the price of oil

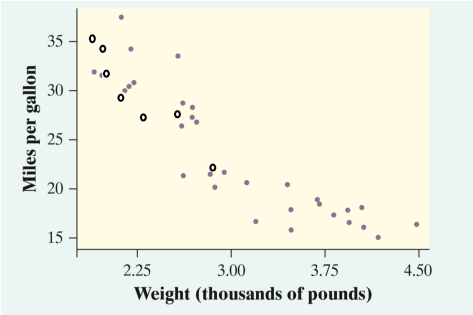
(b) the price of gas

(c) the year

(d) either oil or gas price

(e) time

4. The following graph plots the gas mileage (miles per gallon) of various cars from the same model year versus the weight of these cars in thousands of pounds. The circles correspond to cars made in Japan. From this plot, we may conclude that



(a) there is a positive association between weight and gas mileage for Japanese cars.

(b) the correlation between weight and gas mileage for all the cars is close to 1.

(c) there is little difference between Japanese cars and cars made in other countries.

(d) Japanese cars tend to be lighter in weight than other cars.

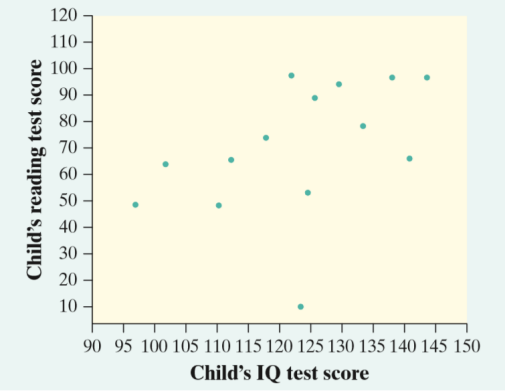
(e) Japanese cars tend to get worse gas mileage than other cars.

5. If women always married men who were 2 years older than themselves, what would the correlation between the ages of husband and wife be?

(a) 2 (b) 1 (c) 0.5 (d) 0

(e) Cant tell without seeing the data

6. The figure below is a scatterplot of reading test scores against IQ test scores for 14 fifth-grade children. There is one low outlier in the plot. What effect does this low outlier have on the correlation?



(a) It makes the correlation closer to 1.

(b) It makes the correlation closer to 0 but still positive.

(c) It makes the correlation equal to 0.

(d) It makes the correlation negative.

(e) It has no effect on the correlation.

7. Which of the following is not a characteristic of the least-squares regression line?

   (a) The slope of the least-squares regression line is always between -1 and 1.

   (b) The least-squares regression line always goes through the point $(\bar{x}, \bar{y})$.

   (c) The least-squares regression line minimizes the sum of squared residuals.

   (d) The slope of the least-squares regression line will always have the same sign as the correlation

   (e) The least-squares regression line is not resistant to outliers.

8. Each year, students in an elementary school take a standardized math test at the end of the school year. For a class of fourth-graders, the average score was 55.1 with a standard deviation of 12.3. In the third grade, these same students had an average score of 61.7 with a standard deviation of 14.0. The correlation between the two sets of scores is r = 0.95. Calculate the equation of the least-squares regression line for predicting a fourth-grade score from a third-grade score.

   (a) $\hat{y} = 3.60 + 0.835\,x$

   (b) $\hat{y} = 15.69 + 0.835\,x$

   (c) $\hat{y} = 2.19 + 1.08\,x$

   (d) $\hat{y} = -11.54 + 1.08\,x$

   (e) Can not be calculated without the data.

9. Using data from the 2009 LPGA tour, a regression analysis was performed using x = average driving distance and y = scoring average. Using the output from the regression analysis shown below, determine the equation of the least-squares regression line.

Predictor	Coef	SE Coef	T	P
Constant	87.974	2.391	36.78	0.000
Driving Distance	−0.060934	0.009536	−6.39	0.000

S = 1.01216 R-Sq = 22.1% R-Sq(adj) = 21.6%

   (a) $\hat{y} = 87.947 + 2.31\,9x$

   (b) $\hat{y} = 87.947 + 1.01216\,x$

(c) $\hat{y} = 87.947 - 0.060934\,x$

(d) $\hat{y} = -0.060934 + 1.01216\,x$

(e) $\hat{y} = -0.060934 + 87.947\,x$

The following 5 problems refer to the following setting.

Measurements on young children in Mumbai, India, found this least-squares line for predicting height y from arm span x:

$$\hat{y} = 6.4 + 0.93\,x$$

Measurements are in centimeters (cm).

10. By looking at the equation of the least-squares regression line, you can see that the correlation between height and arm span is

    (a) greater than zero.

    (b) less than zero.

    (c) 0.93

    (d) 6.4

    (e) Can not tell without seeing the data.

11. In addition to the regression line, the report on the Mumbai measurements says that $r^2 = 0.95$. This suggests that

    (a) although arm span and height are correlated, arm span does not predict height very accurately

    (b) height increases by $\sqrt{0.95} = 0.97$ cm for each additional centimeters of arm span.

    (c) 95% of the relationship between height and arm span is accounted for by the regression line.

    (d) 95% of the variation in height is accounted for by the regression line.

    (e) 95% of the height measurements are accounted for by the regression line.

12. One child in the Mumbai study had height 59 cm and arm span 60 cm. This childs residual is

    (a) -3.2 cm (b) -2.2 cm (c) -1.3 cm (d) 3.2 cm (e) 62.2 cm

13. Suppose that a tall child with arm span 120 cm and height 118 cm was added to the sample used in this study. What effect will adding this child have on the correlation and the slope of the least-squares regression line?

    (a) Correlation will increase, slope will increase.

    (b) Correlation will increase, slope will stay the same.

(c) Correlation will increase, slope will decrease.

(d) Correlation will stay the same, slope will stay the same.

(e) Correlation will stay the same, slope will increase.

14. Suppose that the measurements of arm span and height were converted from centimeters to meters by dividing each measurement by 100. How will this conversion affect the values of r^2 and s?

(a) r^2 will increase, s will increase.

(b) r^2 will increase, s will stay the same.

(c) r^2 will increase, s will decrease.

(d) r^2 will stay the same, s stay the same.

(e) r^2 will stay the same, s will decrease.

15. The fraction of the variation in the values of y that is explained by the least-squares regression of y on x is

(a) the correlation.

(b) the slope of the least-squares regression line.

(c) the square of the correlation coefficient.

(d) the intercept of the least-squares regression line.

(e) the residual.

16. An AP® Statistics student designs an experiment to see whether todays high school students are becoming too calculator-dependent. She prepares two quizzes, both of which contain 40 questions that are best done using paper-and-pencil methods. A random sample of 30 students participates in the experiment. Each student takes both quizzes —one with a calculator and one without— in a random order. To analyze the data, the student constructs a scatterplot that displays the number of correct answers with and without a calculator for each of the 30 students. A least-squares regression yields the equation

$$\widehat{\text{Calculator}} = -1.2 + 0.865\,(\text{Pencil}), \qquad r = 0.79$$

Which of the following statements is/are true?

I. If the student had used Calculator as the explanatory variable, the correlation would remain the same.

II. If the student had used Calculator as the explanatory variable, the slope of the least-squares line would remain the same.

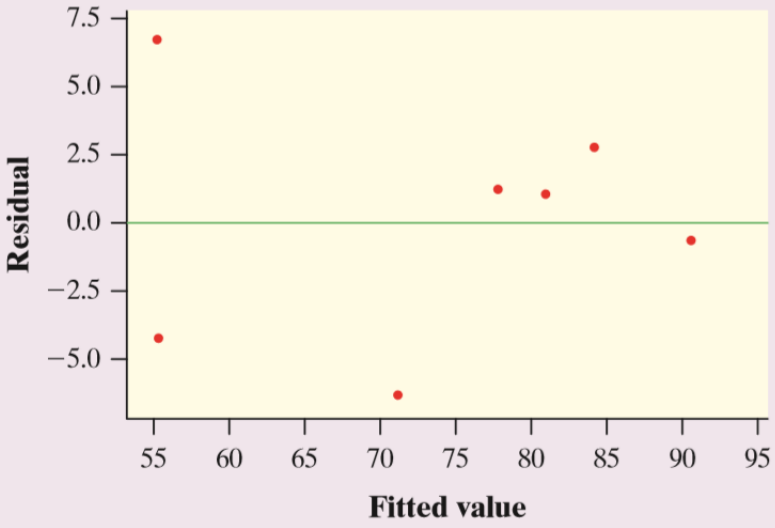
III. The standard deviation of the number of correct answers on the paper-and-pencil quizzes was larger than the standard deviation on the calculator quizzes.

(a) I only
(b) II only
(c) III only
(d) I and III only
(e) I, II and III

The following two problems refer to the following setting. Scientists examined the activity level of 7 fish at different temperatures. Fish activity was rated on a scale of 0 (no activity) to 100 (maximal activity). The temperature was measured in degrees Celsius. A computer regression printout and a residual plot are given below. Notice that the horizontal axis on the residual plot is labeled Fitted value.

Predictor	Coef	SE Coef	T	P
Constant	148.62	10.71	13.88	0.000
Temperature	-3.2167	0.4533	-7.10	0.001

S = 4.78505 R-Sq = 91.0% R-Sq(adj) = 89.2%



17. What was the activity level rating for the fish at a temperature of 20°C?

(a) 87 (b) 84 (c) 81 (d) 66 (e) 3

18. Which of the following gives a correct interpretation of s in the above setting?

(a) For every 1°C increase in temperature, fish activity is predicted to increase by 4.785 units.
(b) The typical distance of the temperature readings from their mean is about 4.785°C.
(c) The typical distance of the activity level ratings from the least-squares line is about 4.785 units.
(d) The typical distance of the activity level readings from their mean is about 4.785.
(e) At a temperature of 0°C, this model predicts an activity level of 4.785.

19. Which of the following statements is not true of the correlation r between the lengths in inches and weights in pounds of a sample of brook trout?

    (a) r must take a value between 1 and 1.
    (b) r is measured in inches.
    (c) If longer trout tend to also be heavier, then $r > 0$.
    (d) r would not change if we measured the lengths of the trout in centimeters instead of inches.
    (e) r would not change if we measured the weights of the trout in kilograms instead of pounds

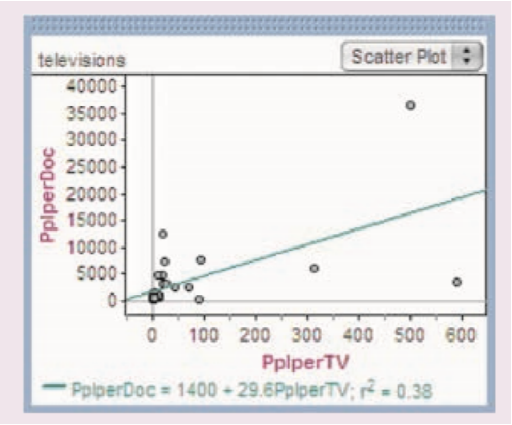
20. When we standardize the values of a variable, the distribution of standardized values has mean 0 and standard deviation 1. Suppose we measure two variables X and Y on each of several subjects. We standardize both variables and then compute the least-squares regression line. Suppose the slope of the least-squares regression line is -0.44. We may conclude that

    (a) the intercept will also be -0.44.
    (b) the intercept will be 1.0.
    (c) the correlation will be $1/-0.44$.
    (d) the correlation will be 1.0.
    (e) the correlation will also be -0.44.

21. There is a linear relationship between the number of chirps made by the striped ground cricket and the air temperature. A least-squares fit of some data collected by a biologist gives the model $\hat{y} = 25.2 + 3.3\,x$, where x is the number of chirps per minute and $\hat{y}$ is the estimated temperature in degrees Fahrenheit. What is the predicted increase in temperature for an increase of 5 chirps per minute?

    (a) 3.3 °F (b) 16.5 °F (c) 25.2 °F (d) 28.5 °F (e) 41.7 °F

22. A data set included the number of people per television set and the number of people per physician for 40 countries. The Fathom screen shot below displays a scatterplot of the data with the least-squares regression line added. In Ethiopia, there were 503 people per TV and 36,660 people per doctor. What effect would removing this point have on the regression line?



(a) Slope would increase; y intercept would increase.

(b) Slope would increase; y intercept would decrease.

(c) Slope would decrease; y intercept would increase.

(d) Slope would decrease; y intercept would decrease.

(e) Slope and y intercept would stay the same

23. Suppose a study finds that the correlation coefficient relating family income to SAT scores is $r = +1$. Which of the following are proper conclusions?

I. Poverty causes low SAT scores.

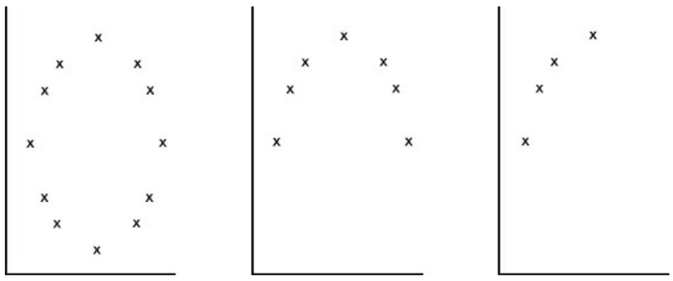
II. Wealth causes high SAT scores.

III. There is a very strong association between family income and SAT scores.

(a) I only (c) III only (e) I, II and III

(b) II only (d) I and II

24. Consider the following three scatterplots:



(a) None are 0.

(b) One is 0, one is negative, and one is positive.

(c) One is 0, and both of the others are positive.

(d) Two are 0, and the other is 1.

(e) Two are 0, and the other is close to 1.

Free Response

1. Using data from the 2010 census, a random sample of 348 U.S. residents aged 18 and older was selected. Among the variables recorded were gender (male or female), housing status (rent or own), and marital status (married or not married).

   The two-way table below summarizes the relationship between gender and housing status.

	Male	**Female**	**Total**
Own	132	122	**254**
Rent	50	44	**94**
Total	**182**	**166**	**348**

   (a) What percent of males in the sample own their home?

   (b) Make a graph to compare the distribution of housing status for males and females.

   (c) Using your graph from part (b), describe the relationship between gender and housing status.

   (d) The following two-way table below summarizes the relationship between marital status and housing status.

	Married	**Not married**	**Total**
Own	172	82	**254**
Rent	40	5554	**94**
Total	**212**	**136**	**348**

   For the members of the sample, is the relationship between marital status and housing status stronger or weaker than the relationship between gender and housing status that you described in part (c)? Justify your choice using the data provided in the two-way tables.

2. **2007FR6**

A study was designed to explore subjects ability to judge the distance between two objects placed in a dimly lit room. The researcher suspected that the subjects would generally overestimate the distance between the objects in the room and that this overestimation would increase the farther apart the objects were.

The two objects were placed at random locations in the room before a subject estimated the distance (in feet) between those two objects. After each subject estimated the distance, the locations of the objects were rerandomized before the next subject viewed the room.

After data were collected for 40 subjects, two linear models were fit in an attempt to describe the relationship between the subjects perceived distances (y) and the actual distance, in feet, between the two objects.

$$\text{Model 1: } \hat{y} = 0.238 + 1.080\times(\text{actual distance})$$

The standard errors of the estimated coefficients for Model 1 are 0.260 and 0.118, respectively.

$$\text{Model 2: } \hat{y} = 1.102\times(\text{actual distance})$$

The standard error of the estimated coefficient for Model 2 is 0.393.

(a) Provide an interpretation in context for the estimated slope in Model 1.

(b) Explain why the researcher might prefer Model 2 to Model 1 in this context.

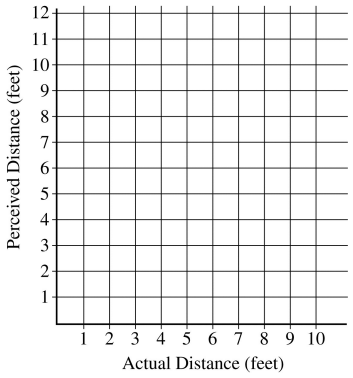
(c) Using Model 2, test the researchers hypothesis that in dim light participants overestimate the distance, with the overestimate increasing as the actual distance increases. (Assume appropriate conditions for inference are met.)

The researchers also wanted to explore whether the performance on this task differed between subjects who wear contact lenses and subjects who do not wear contact lenses. A new variable was created to indicate whether or not a subject wears contact lenses. The data for this variable were coded numerically (1 = contact wearer, 0 = noncontact wearer), and this new variable, named contact, was included in the following model.

$$\text{Model 3:} \hat{y} = 1.0\times(\text{actual distance}) + 0.12\times(\text{contact})\times(\text{actual distance})$$

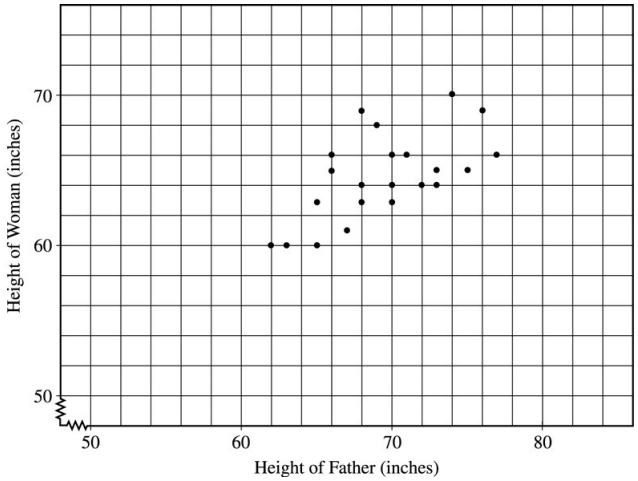
The standard errors of the estimated coefficients for Model 3 are 0.357 and 0.032, respectively.

(d) Using Model 3, sketch the estimated regression model for contact wearers and the estimated regression model for noncontact wearers on the grid below.



(e) In the context of this study, provide an interpretation of the estimated coefficients for Model 3.

3. **2007FRB4**
Each of 25 adult women was asked to provide her own height (y), in inches, and the height (x), in inches, of her father. The scatterplot below displays the results. Only 22 of the 25 pairs are distinguishable because some of the (x,y) pairs were the same. The equation of the least squares regression line is $\hat{y} = 35.1 + 0.427x$

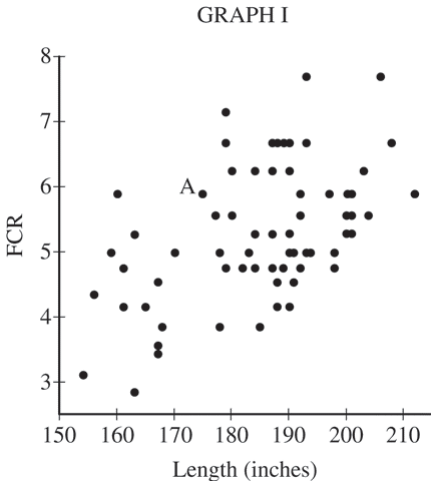


(a) Draw the least squares regression line on the scatterplot above.

(b) One fathers height was x = 67 inches and his daughters height was y = 61 inches. Circle the point on the scatterplot above that represents this pair and draw the segment on the scatterplot that corresponds to the residual for it. Give a numerical value for the residual.

(c) Suppose the point x = 84, y = 71 is added to the data set. Would the slope of the least squares regression line increase, decrease, or remain about the same? Explain. (Note: No calculations are necessary to answer this question.)
Would the correlation increase, decrease, or remain about the same? Explain. (Note: No calculations are necessary to answer this question.)

4. **2014FR6**
Jamal is researching the characteristics of a car that might be useful in predicting the fuel consumption rate (FCR); that is, the number of gallons of gasoline that the car requires to travel 100 miles under conditions of typical city driving. The length of a car is one explanatory variable that can be used to predict FCR. Graph I is a scatterplot showing the lengths of 66 cars plotted with the corresponding FCR. One point on the graph is labeled A.



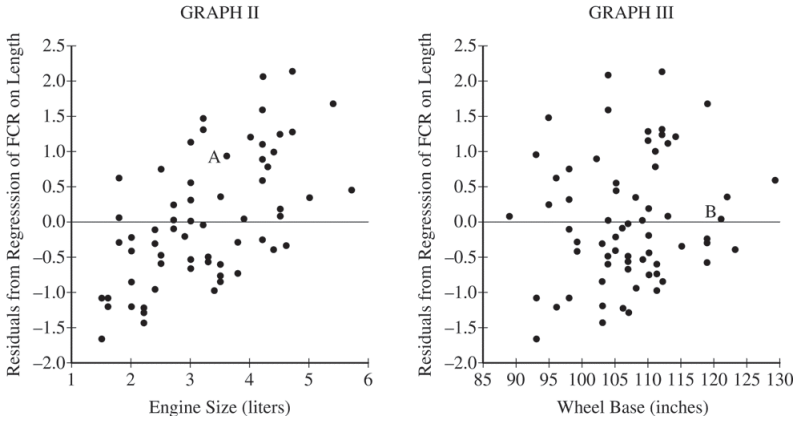
Jamal examined the scatterplot and determined that a linear model would be a reasonable way to express the relationship between FCR and length. A computer output from a linear regression is shown below.

Linear Fit
FCR= $-1.595789 + 0.0372614 \times$Length

Summary of Fit	
R-square	0.250401
Root Mean Square Error	0.902382
Observations	66

(a) The point on the graph labeled A represents one car of length 175 inches and an FCR of 5.88. Calculate and interpret the residual for the car relative to the least squares regression line.

(b) Jamal knows that it is possible to predict a response variable using more than one explanatory variable. He wants to see if he can improve the original model of predicting FCR from length by including a second explanatory variable in addition to length. He is considering including engine size, in liters, or wheel base (the length between axles), in inches. Graph II is a scatterplot showing the engine size of the 66 cars plotted with the corresponding residuals from the regression of FCR on length. Graph III is a scatterplot showing the wheel base of the 66 cars plotted with the corresponding residuals from the regression of FCR on length.



In graph II, the point labeled A corresponds to the same car whose point was labeled A in graph I. The measurements for the car represented by point A are given below.

FCR	Length(inches)	Engine Size (liters)	Wheel Base (inches)
5.88	175	3.6	93

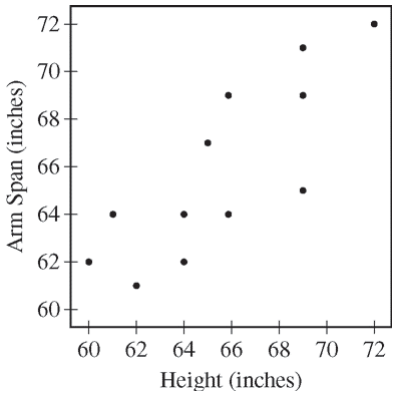
(i) Circle the point on graph III that corresponds to the car represented by point A on graphs I and II.

(ii) There is a point on graph III labeled B. It is very close to the horizontal line at 0. What does that indicate about the FCR of the car represented by point B?

(c) Write a few sentences to compare the association between the variables in graph II with the association between the variables in graph III.

(d) Jamal wants to predict FCR using length and one of the other variables, engine size or wheel base. Based on your response to part (c), which variable, engine size or wheel base, should Jamal use in addition to length if he wants to improve the prediction? Explain why you chose that variable.

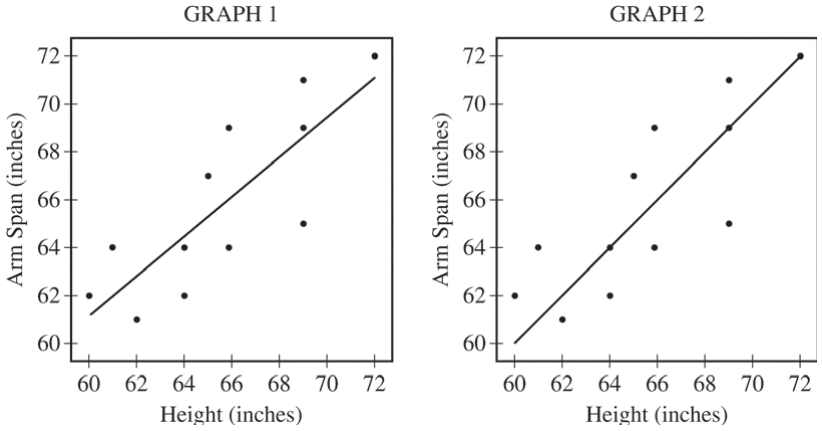
5. **2015FR5**
A student measured the heights and the arm spans, rounded to the nearest inch, of each person in a random sample of 12 seniors at a high school. A scatterplot of arm span versus height for the 12 seniors is shown.



(a) Based on the scatterplot, describe the relationship between arm span and height for the sample of 12 seniors. Let x represent height, in inches, and let y represent arm span, in inches. Two scatterplots of the same data are shown below. Graph 1 shows the data with the least squares regression line

$$\hat{y} = 11.74 + 0.8247x$$

And graph 2 shows the data with the line $y = x$.



(b) The criteria described in the table below can be used to classify people into one of three body shape categories: square, tall rectangle, or short rectangle.

Square	Tall Rectangle	Short Rectangle
Arm span is equal to height.	Arm span is less than height.	Arm span is greater than height.

(i) For which graph, 1 or 2, is the line helpful in classifying a students body shape as square, tall rectangle, or short rectangle? Explain.

(ii) Complete the table of classifications for the 12 seniors.

Classification	Square	Tall Rectangle	Short Rectangle
Frequency			

(c) Using the best model for prediction, calculate the predicted arm span for a senior with height 61 inches.