Sample Quiz 6A

In a study reported in the July 2007 issue of the *Journal of Epidemiology and Community Health*, researchers investigated whether veterans are more likely to commit suicide than nonveterans. They spent 12 years following 104,000 veterans who had served in the armed forces between 1917 and 1994, and compared them with 216,000 nonveterans. They found that 197 veterans and 311 nonveterans committed suicide.

1. Identify the explanatory and response variables in this study.
2. Is this an experiment or an observational study? Explain briefly.
3. Notice that more nonveterans than veterans committed suicide (311 vs. 197).
Would you conclude that veterans are *less* likely to commit suicide than nonveterans? Explain.
4. Calculate the relative risk of committing suicide, comparing veterans to nonveterans.
5. Describe and justify the conclusion that you would draw from this study.

Solution to Sample Quiz 6A

1. The explanatory variable is *whether a subject is a veteran or a nonveteran* (a binary categorical variable). The response variable is *whether the subject committed suicide* (also a binary categorical variable).
2. This is an observational study because the researchers did not determine who would/would not be a veteran.
3. No, you cannot conclude from this information that veterans are less likely to commit suicide than nonveterans. Because the sample sizes differed, you need to consider the suicide rates, not simply the counts. The suicide rate for veterans is 197/104,000 or .001894, whereas the suicide rate for nonveterans is 311/216,000 or .00144.
4. The relative risk is of committing suicide is .001894/.00144 or 1.315.
5. This study indicates that veterans are about 1.3 times more likely than nonveterans to commit suicide. However, because this was an observational study, you cannot conclude that there is a causal relationship between serving in the armed forces and committing suicide.

Sample Quiz 6B

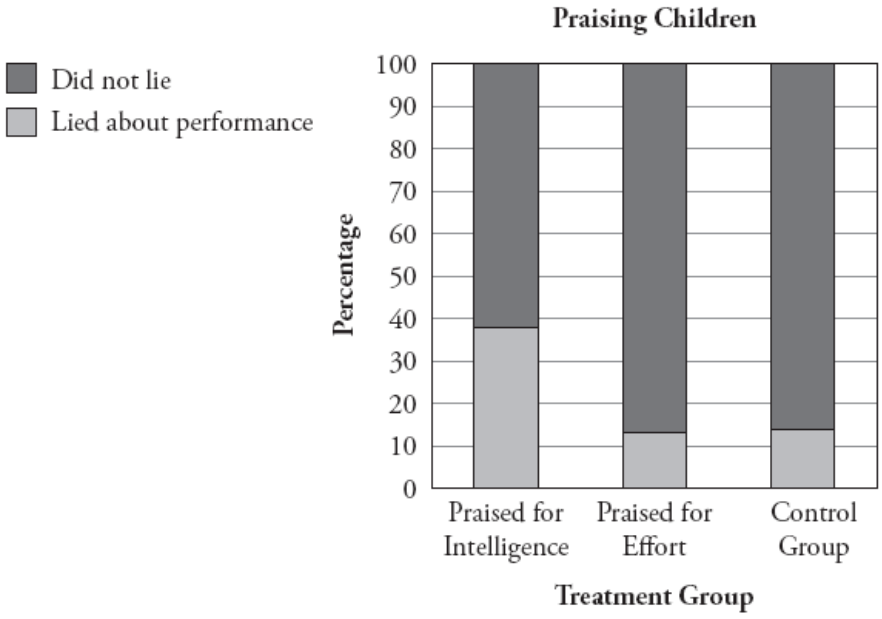
Is it better to praise a child's intelligence or effort? In a study reported in a 1998 issue of the *Journal of Personality and Social Psychology,* researchers investigated this question by randomly assigning fifth-graders into three groups. One group received praise for their intelligence after working on a set of problems; another group received praise for their effort; and a third (control) group received general praise that did not mention a specific attribute. Each child was then asked to write a description of the problems for a child in another state, including how many he or she solved correctly. One response variable was *whether the child lied about how many problems he/she had solved correctly.* The data are organized in the following table:

	Praised Intelligence	Praised Effort	Control Group	Total
Lied About Performance	11	4	4	19
Did Not Lie	18	26	25	69
Total	29	30	29	88

1–5. Analyze these data to investigate whether they suggest any relationship between the type of praise given to a child and the child's propensity for lying about performance. Write a paragraph describing your findings, and support your findings with a graphical display and calculations.

Solution to Sample Quiz 6B

1–5. These data indicate that there is virtually no difference between the control group and the group praised for their effort. In both of these groups, approximately 14% of the children lied about the number of problems they solved correctly. However, in the group who was praised for their intelligence, 38% of them lied. If you combine the "praised for effort" and control groups, you find the relative risk of lying in the "praised for intelligence group" versus the other two groups is 2.79. Thus, the children who are praised for their intelligence are almost three times more likely to lie about the number of problems they solved correctly than are the children in the other two groups. The following segmented bar graph displays these results:



Sample Quiz 7A

The following side-by-side stemplot displays the total number of points scored per Super Bowl football game for the first 41 Super Bowls (from 1967–2007), separated according to the first 20 games (1967–1986) and the next 21 games (1987–2007):

First 20 Games		Next 21 Games
97321	2	
87720	3	16799
777654	4	13456
640	5	23569
6	6	11599
	7	5

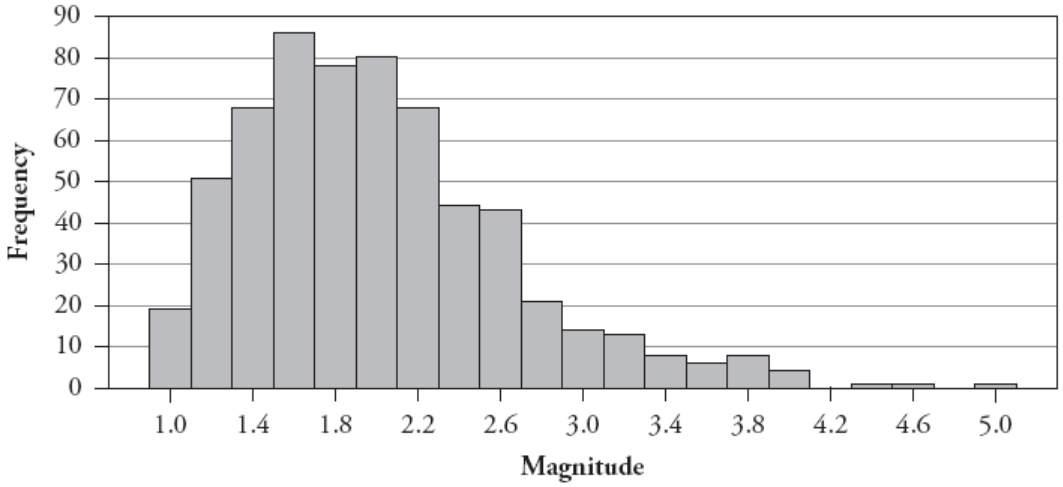
1. Does this stemplot enable you to determine how many points were scored in the first Super Bowl? If so, what is this number?
2. Does this stemplot enable you to determine how many of the first 41 Super Bowls had a total of 37 points? If so, what is this number?
3. Does this stemplot provide evidence that Super Bowl games have become more high-scoring over time, more low-scoring over time, or neither? Explain.
4. True or false? The five lowest-scoring Super Bowls were all played among the first 20 games.
5. True or false? The five highest-scoring Super Bowls were all played among the next 21 games.

Solution to Sample Quiz 7A

1. No, you cannot tell from this stemplot the order of the Super Bowl games.
2. Yes, you can tell that in 2 of the first 41 Super Bowls there were a total of 37 points.
3. This stemplot provides evidence that Super Bowls have become more high-scoring over time because the scores in the last 21 games tend to be slightly higher than the scores in the first 20 games.
4. True. The five lowest scores were 21, 22, 23, 27, and 29.
5. False. The five highest scores were 75, 69, 69, 66, and 65—and one of these was among the first 20 games.

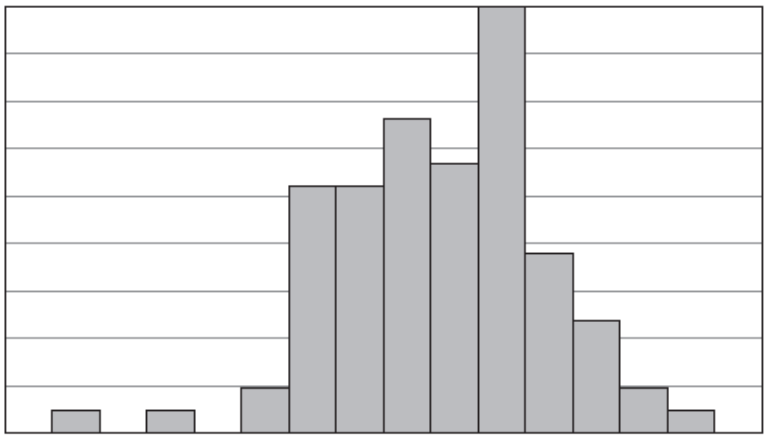
Sample Quiz 7B

The following histogram displays the magnitudes of the 614 earthquakes with Richter scale magnitude greater than 1.0 that occurred in the United States between March 25 and April 1, 2004:

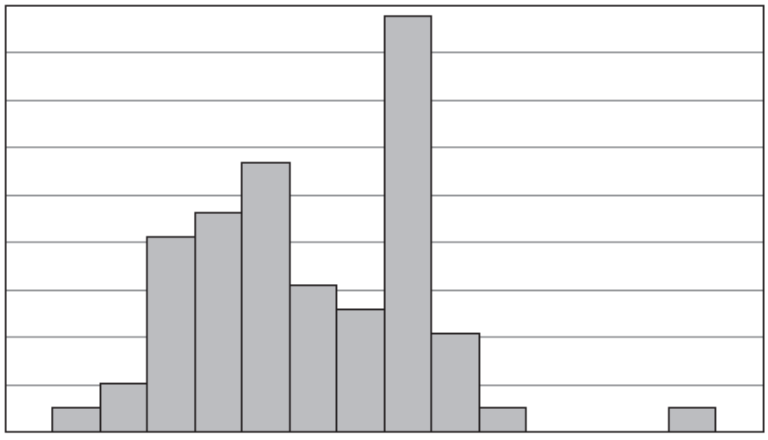


1. Describe the shape of this distribution.
2. Is the percentage of earthquakes of magnitude 3.0 or higher closest to 1%, 10%, or 25%?
The following histograms display student responses to several questions on a course survey taken during the first day of a statistics class.

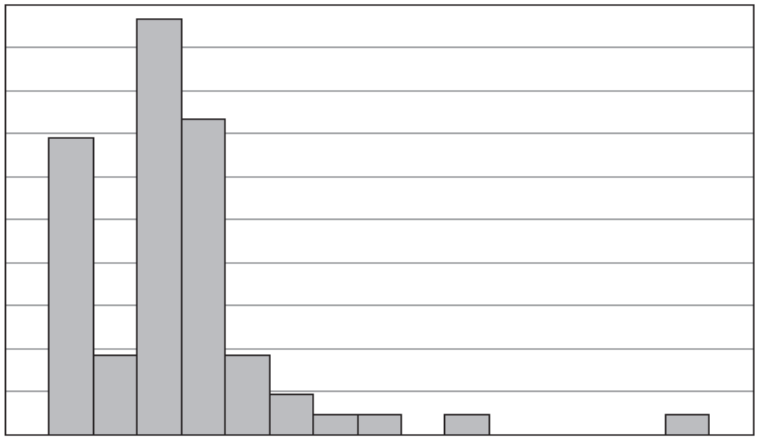
a.



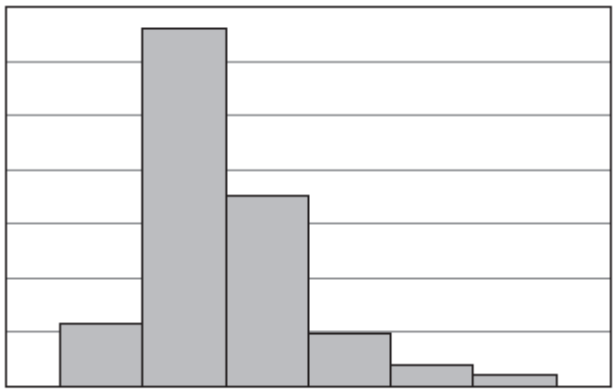
b.



c.



d.



3. Which histogram do you think displays the variable *number of siblings?* Justify your answer.
4. Which histogram do you think displays the variable *price paid for most recent haircut?* Justify your answer.
5. Which histogram do you think displays the variable *height?* Justify your answer.

Solution to Sample Quiz 7B

1. This histogram is strongly skewed to the right.
2. This percentage is closer to 10%.
3. Histogram d displays the variable *number of siblings.* It would be rare for students to have a large number of siblings and very common for students to have one or two siblings.
4. Histogram c displays the variable *price paid for most recent haircut.* Some students will get their hair cut for free by roommates or friends. College students (males particularly) will not be willing to pay a large amount for a haircut. A few females will pay very large amounts. It makes sense to have a large outlier for a particular visit.
5. Histogram a displays the variable *height.* You might expect a more symmetric distribution or one with two peaks (males and females). It's feasible that there were one or two shorter people in the class. Histogram b is also a possibility; perhaps the large peak comes at a value, such as six feet, that many students might round to. It is difficult to explain the extreme outlier in histogram b, unless this is a data entry error or a professional basketball player.

Sample Quiz 8A

The following table reports counts of the number of "close friends" reported by a sample of men and a sample of women:

Number of Close Friends	0	1	2	3	4	5	6	Total
Number of Respondents (male)	196	135	108	100	42	40	33	654
Number of Respondents (female)	201	146	155	132	86	56	37	813

1. Is *number of close friends* a quantitative or categorical variable?
2. Are these distributions roughly symmetric, skewed to the left, or skewed to the right? Explain briefly. (You do not need to construct any graphs.)
3. Calculate the median number of close friends for each gender.
4. Based on the shape of the distributions, do you expect the means to be greater than the medians, less than the medians, or very close to the medians? (Do not calculate either mean.)
5. For each gender, calculate the proportion who say that they have no close friends. Comment on how these proportions compare between men and women.

Solution to Sample Quiz 8A

1. This is a quantitative variable.
2. These distributions are skewed to the right. The frequencies are greatest at 0 and then roughly decrease as the number of close friends increases.
3. For males, the median is 1 friend (between 327th and 328th observations, 196 _ 135 _ 327).
For females, the median is 2 friends (407th observation; 201 _ 146 _ 407, 201 _ 146 _ 155 _ 407).
4. Because these distributions are skewed to the right, the means should be greater than the medians.
5. For males, the proportion who say they have no close friends is 196/654 or .2997.

For females, the proportion who say they have no close friends is 201/813 or .2472.

Thirty percent of the men say they have no close friends, whereas just under 25 percent of the women report no close friends. With samples this large, there is likely to be a significant difference in the proportions, with men 1.21 times more likely than women to report no close friends.

Sample Quiz 8B

1. Create an example of five hypothetical exam scores (between 0 and 100, with repeats allowed) with the property that 80% of the scores are less than the mean.
2. Create an example of five hypothetical exam scores (between 0 and 100, with repeats allowed) with the property that the mean is greater than three times the median.
3. What effect does adding 10 points to every exam score in a class have upon the mean score? In other words, how would the new mean compare to the old mean?
Be as specific as possible.
4. Is it possible for the mean value of a variable to be greater than all of the data values? Explain.
5. A trimmed mean is another measure of center. For example, one trimmed mean deletes the highest 5% of the values and the lowest 5% of the values, and then takes the mean of the remaining 90% of the values. Is this trimmed mean more or less resistant to outliers than the mean? Explain briefly.

Solution to Sample Quiz 8B

1. Many answers are possible. There should be one extremely high score (to inflate the mean above the other observations), and the other four scores should be very low. For example, with the exam scores {20, 20, 20, 20, 100}, the mean is 36.
2. The median will be the third ordered score, so the first three scores need to be low, whereas the last two scores need to be high in order to create a large mean.
For example, with the exam scores {10, 10, 10, 100, 100}, the median is 10 and the mean is 46.
3. Adding 10 points to every exam score will raise the mean score by exactly 10 points.
4. No, it is not possible for the mean value of a variable to be greater than all of the data values. The mean is an average of all the values and must have at least one data value on each side.
5. A trimmed mean is more resistant to outliers than the mean because the outliers have been trimmed away. Outliers will have virtually no effect on a trimmed mean.

Sample Quiz 9A

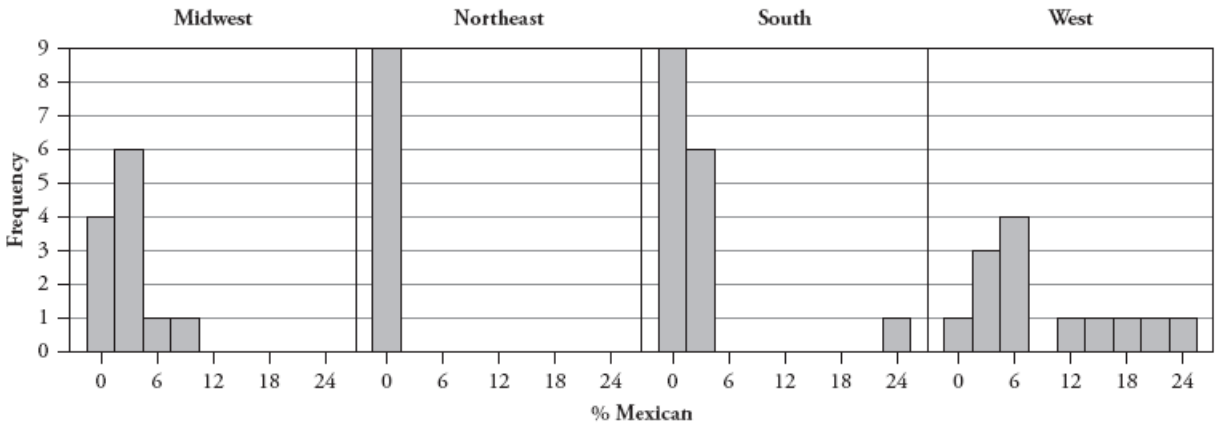
1. Create an example of five hypothetical exam scores (between 0 and 100, inclusive, with repeats allowed) with the property that the standard deviation is as small as possible.
2. What effect does adding ten points to every exam score in a class have on the interquartile range of the scores? Be specific and explain briefly.
3. What effect does adding ten points to every exam score in a class have on the standard deviation of the scores? Be specific and explain briefly.
4. Is the mean absolute deviation more or less resistant to outliers than the standard deviation? Explain briefly.
5. Would a manufacturer of candy bars want to have a larger standard deviation of the weights of candy bars coming off the assembly line or a small standard deviation of those weights? Explain briefly.

Solution to Sample Quiz 9A

1. Many answers are possible, but all five exam scores must be identical and result in a standard deviation of zero, which is as small as possible.
2. Adding ten points to each score will have no effect on the interquartile range of the scores. Both the first and third quartile will have been raised by ten points, so the difference between them will be unchanged.
3. Adding ten points to each score will have no effect on the standard deviation because the scores will still be the same distance from the mean (which will have also increased by 10 points). The spread of the distribution will not have changed—the entire distribution will simply have shifted up on a number line.
4. Neither is resistant, but the mean absolute deviation will be slightly more resistant to outliers because it does not square the distance between the outlier and the mean before including it in the sum.
5. A manufacturer of candy bars would want to have a small standard deviation of weights. If the standard deviation is large, then the manufacturer could be frequently giving away extra candy for free, which would be costly; or just as frequently, the manufacturer could be cheating customers out of candy. If the standard deviation is large, customers would most likely notice that many of their candy bars are underweight and would complain. With a small standard deviation, most candy bars would be near the advertised weight and customers would not feel cheated, nor would the manufacturer be wasting candy.

Sample Quiz 9B

The following histograms display the distributions of percentage of a state's residents who are Mexican, for each region of the country (West, South, Northeast, Midwest).



1. Which region do you suspect has the *smallest* standard deviation in these percentages? Explain.
2. Which region do you suspect has the *largest* standard deviation in percentages? Explain.
3. For the south region, would you recommend reporting the standard deviation or the interquartile range as your measure of the spread of the distribution? Explain.
4. Which of the following sets of five numbers has the *smallest* standard deviation? Explain your reasoning, but do not do any calculations.
(a) 0, 0, 0, 10, 10 (b) 5, 5, 6, 7, 7 (c) 0, 5, 5, 5, 10
5. Which of these three sets of numbers (above) has the *largest* standard deviation? Explain your reasoning, but do not do any calculations.

Solution to Sample Quiz 9B

1. The Northeast has the smallest standard deviation because the data are so tightly grouped around the mean.
2. The West will have the largest standard deviation because it has the largest "spread" from the mean. The South would also be a reasonable answer because it is difficult to know just how much that outlier will inflate the standard deviation value.
3. The interquartile range would be more appropriate because the outlier will inflate the standard deviation (so the measure will not describe the bulk of the data), but that value will not affect the interquartile range.
4. Set (b) has the smallest standard deviation because all of the numbers are so close to the mean (6).
5. Set (a) has the largest standard deviation because all of the numbers are at extreme distances from the mean (4).

Sample Quiz 10A

The following data are monthly rents (in dollars) of studio and one-bedroom apartments in Harrisburg and Philadelphia, Pennsylvania.

Harrisburg (n = 10): 500, 549, 569, 575, 585, 600, 630, 680, 705, 790

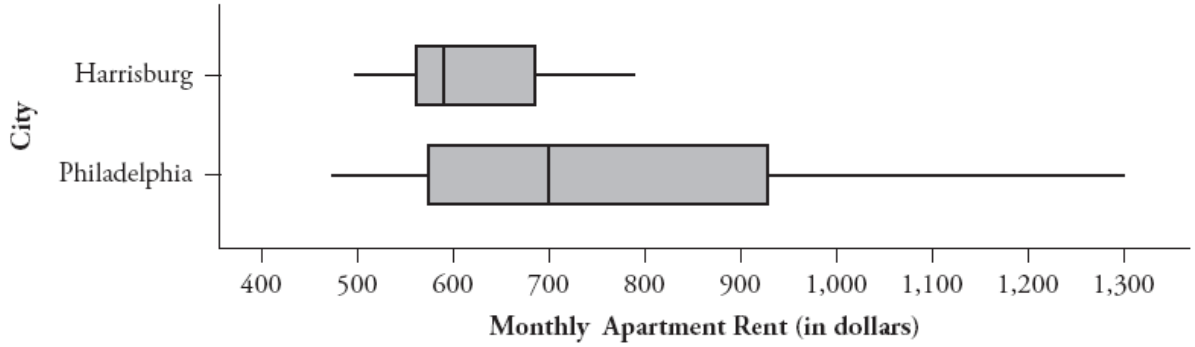
Philadelphia (n = 15): 475, 525, 540, 575, 600, 600, 645, 700, 725, 755, 885, 930, 965, 1180, 1300

1 and **2.** For each city, determine and report the five-number summary of these monthly rents.
3. Construct boxplots of the distributions of rent amounts in these two cities, using the same axis and scale. (Do not bother to check for outliers; there are no outliers in either distribution.)

4 and **5.** Compare and contrast the distributions of monthly apartment rents in these two cities. Refer to appropriate calculations and displays to support your comments.

Solution to Sample Quiz 10A

1 and **2.** Harrisburg: min = \$500, Q_L = \$569, median = \$592.5, Q_U = \$ 680, max = \$790
Philadelphia: min = \$475, Q_L = \$575, median = \$700, Q_U = \$930, max = \$1300
3.



4 and **5.** In general, the monthly apartment rents in Philadelphia are higher and more varied than in Harrisburg. Although the minimum and first quartile rents are similar (\$475 and \$575 in Philadelphia and \$500 and \$569 in Harrisburg), the rents in Philadelphia rapidly outpace Harrisburg thereafter. The median Philadelphia rent is more than \$100 more than in Harrisburg (\$700 vs. \$592.50), and the upper quartiles differ by \$250. The maximum Harrisburg rent is only \$790, whereas the strongly right-skewed Philadelphia rents have a maximum of \$1300.

Sample Quiz 10B

The following side-by-side stemplot displays the total number of points scored per Super Bowl football game for the first 41 Super Bowls (from 1967–2007), separated according to the first 20 games (1967–1986) and the next 21 games (1987–2007):

First 20 Games		Next 21 Games
97321	2	
87720	3	16799
777654	4	13456
640	5	23569
6	6	11599
	7	5

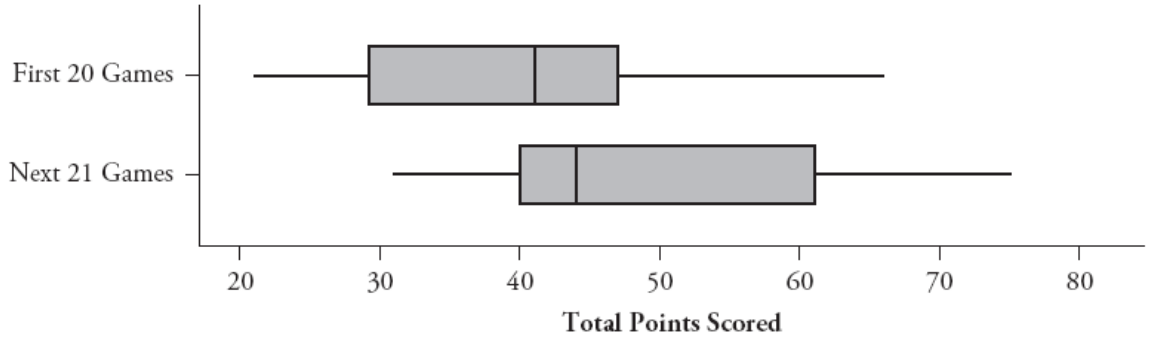
1 and **2.** For each group, determine the five-number summary of these total points.
3. Construct boxplots of the distributions of total points in these two groups, using the same axes and scale. (Do not bother to check for outliers; there are no outliers in either distribution.)

4 and **5.** Compare and contrast the distributions of total points between these two groups. Refer to appropriate calculations and displays to support your comments.

Solution to Sample Quiz 10B

1 and **2.** First 20 games: min = 21, Q_L = 29.5, median = 41, Q_U = 47, max = 66 points
Next 21 games: min = 31, Q_L = 40, median = 52, Q_U = 61, max = 75 points

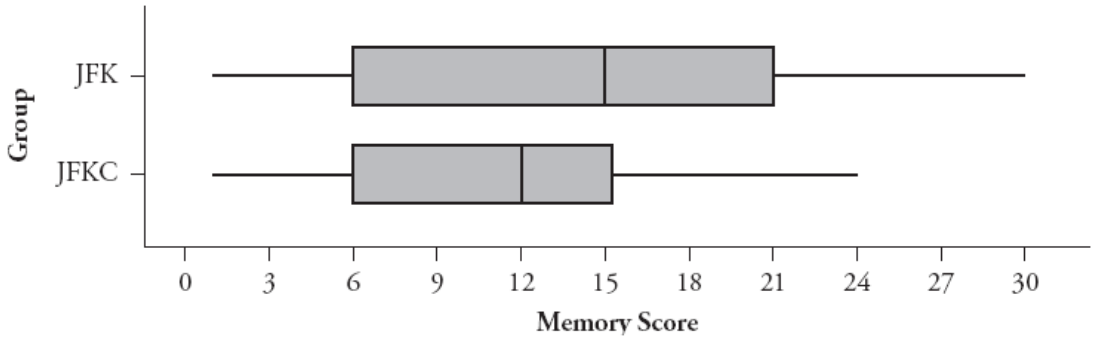
3.



4 and **5.** The total number of points scored in a Super Bowl has increased rather dramatically in the last 21 games. In at least 75% of the first 20 games, a total of less than 47 points were scored, whereas in at least half of the more recent 21 games, more than 52 points were scored. In over 25% of the last 21 games, at least 61 points have been scored, and in one game, a total of 75 points were scored. The minimum number of points scored in a game since 1987 is 31, whereas prior to this, 25% of the games saw a total score of less than 30 points, with a minimum score of 21 points.

Sample Exam 2A

1. An instructor conducted an in-class experiment where students memorized as many letters as possible from a strip of paper. Everyone received the same letters in the same order, but some students saw the letters in convenient three-letter groupings (such as JFK-CIA) whereas others saw less convenient groupings (such as JFKCIAF). Boxplots of the number of letters memorized correctly appear below:

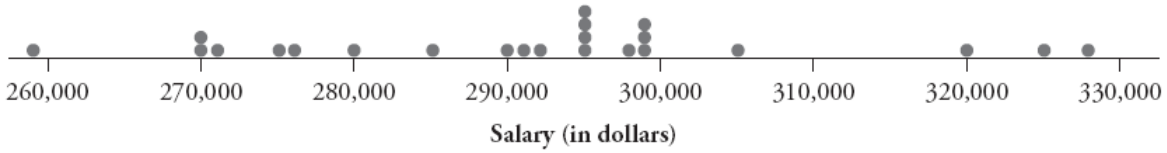


a. Report (as accurately as possible from the graph) the median score for each group:
JFK: JFKC:
b. Report (as accurately as possible from the graph) the interquartile range for each group:
JFK: JFKC:

c. Do the boxplots provide any evidence that students who received the letters in convenient three-letter groupings were able to memorize more letters than the other group? Explain briefly.

2. In a recent study, researchers followed 104,000 U.S. veterans who had served in the armed forces and a comparison group of 216,000 nonveterans. Over a period of 12 years, they found that 197 veterans and 311 nonveterans committed suicide.
a. Calculate the relative risk of suicide, comparing veterans to nonveterans.
b. Write a sentence interpreting this relative risk value.
c. A segmented bar graph would not provide much helpful information for these data. Explain why not. (Do not bother to produce this graph.)

3. The following dotplot displays the salaries of the 23 presidents of California State University campuses for the 2007_08 academic year:



These salaries, arranged in order, are listed here:

258,680	270,000	270,315	270,568	275,000	276,055	279,500	285,000
290,000	291,179	292,000	295,000	295,000	295,000	295,000	297,870
298,749	299,000	299,435	305,008	320,329	325,000	328,209	

a. Determine the five-number summary of these salaries. Consider the following computer output:

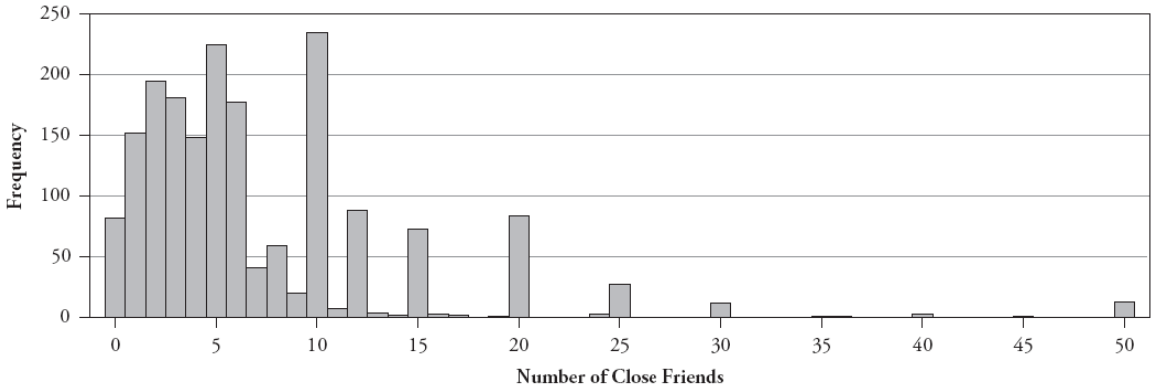
Variable	N	Mean	StDev
salary	23	291,822	17,669

b. The salary for Cal Poly's President Baker is $328,209. Calculate the *z*-score for his salary, and decide whether his salary is more than two standard deviations above the mean.
c. Check whether President Baker's salary is an outlier, according to the 1.5 _ IQR rule.

4. The 2001_2002 National Health and Nutrition Examination Survey asked people over the age of 60 how many close friends they had. The answers for the 1840 respondents appear in the following table. (For example, 82 people answered that they have 0 close friends, and 13 people answered that they have 50 close friends, and nobody answered that they have 18 close friends.)

Answer	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Count	82	152	195	181	148	225	177	41	59	20	235	7	88	4
Answer	14	15	16	17	19	20	24	25	30	35	36	40	45	50
Count	2	73	3	2	1	84	3	27	12	1	1	3	1	13

Consider the following histogram of these data:

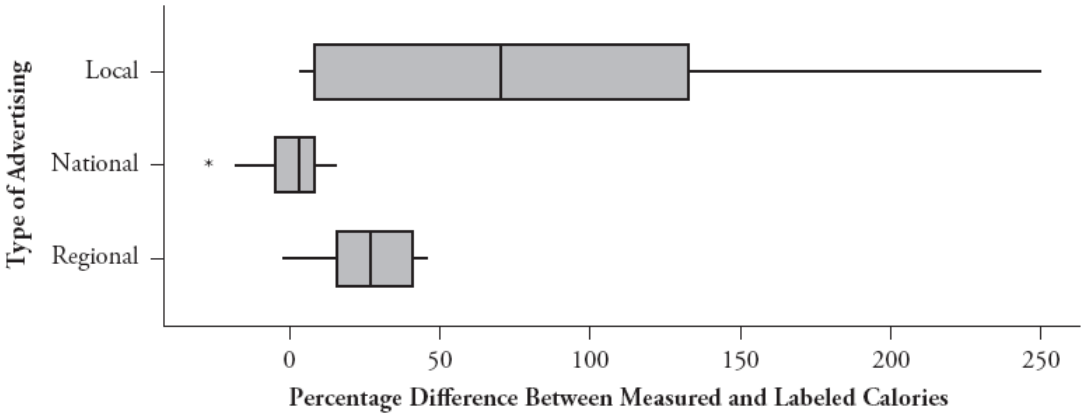


a. Write a paragraph describing key features of this distribution.
b. Determine the median of these data.
c. Would you expect the mean to be greater than the median, less than the median, or about the same as the median? Explain briefly without calculating the mean.

5. Suppose Mary records the ages of people entering a McDonald's restaurant in a suburban shopping plaza tomorrow, while Abby records the ages of people entering a popular food court on a nearby college campus. Who would you expect to have the higher standard deviation of these ages: Mary (suburban McDonald's) or Abby (campus food court)? Explain briefly.

6. Construct a hypothetical example of ten exam scores (between 0 and 100, inclusive, possibly including repeats) so that the interquartile range equals zero and the mean is greater than the median. (Both of these properties should hold for the example that you create.) Also report the values of the mean, median, and IQR for your example. (If you are unable to create an example for which both properties hold, then for partial credit try to create an example in which one of the properties holds.)

7. In a recent study, researchers purchased 40 food items in New York City and determined the actual calorie content of each through a laboratory analysis. They then calculated the percentage difference between the actual calorie content and the calorie count listed on the item's label. (A positive percentage difference corresponds to a food item whose actual calorie content was higher than what the label claimed.) Each food item was also classified according to whether it was marketed locally, nationally, or regionally. The boxplots below were constructed to compare the distributions:



Write a paragraph summarizing what these boxplots reveal about the percentage differences between measured and labeled calorie content among the three groups of food items.

Solution to Sample Exam 2A

1. a. JFK: 15 JFKC: 12
b. JFK: 21 _ 6 _ 15 JFKC: 15 _ 6 _ 9

2. a. The conditional proportion of suicides among veterans is 197/104,000 or .00189. Among nonveterans, the conditional proportion of suicides is 311/216,000 or .00144. The relative risk is, therefore, .01189/.00144 or 1.32.
b. Veterans are 1.32 times more likely to commit suicide than nonveterans.
c. Because the conditional proportions of suicide are so small (less than .002) in both groups, the "suicide" segment of the bar graphs would be essentially invisible for both groups.

3. a. Minimum: $258,680 Maximum: $328,209
(n + 1)/2 = 12, so the median is the 12th ordered value: $295,000
The lower quartile is the median of the 11 values below the (overall) median, so it is the 6th value, which is $276,055. The upper quartile is similarly the 6th value from the top: $299,000.
b. The z-score for President Baker's salary is (328,209 _ 291,822)/17,669 or 2.06. This means that his salary is slightly more than two standard deviations above the average salary among CSU presidents.
c. The IQR is 299,000 - 276,055 or 22,945, so 1.5 × IQR = 34,417.5. To be a high outlier, a salary must be greater than 299,000 - 34,417.5 or 333,417.5. Baker's salary is less than this, so it is not an outlier.
4. a. The distribution of number of close friends is sharply skewed to the right. Most people say that they have 6 or fewer close friends, although the most common response is 10 close friends and some people answered as high as 50 close friends.
b. To find the median, remember that the sample size is 1840 people, so the median is in position (1840 + 1)/2 = 920.5, so we take the value of the median to be the average of the 920th and 921st values. Adding the counts reveals that these values are at the value 5, which is, therefore, the median.
c. The distribution is sharply skewed to the right, so the mean will be greater than the median.

5. Mary would have the greater standard deviation of ages, because McDonald's would see much more variability in its customers' ages than the campus food court would. McDonald's would have many children and elderly people as customers, as well as college-aged and middle-aged people. On the other hand, the campus food court would have mostly college-aged customers, with some older faculty and staff members and visitors.
6. One example that works is {50, 50, 50, 50, 50, 50, 50, 50, 100, 100}. The lower quartile (third value in order) is 50, and the upper quartile (third value from the top, in order) is 50, so the interquartile range is 0. The median is also 50, and the mean is 600/10 = 60, so the mean is greater than the median.

7. The most striking aspect of these data is that locally marketed food items tend to have many more calories than advertised. The median discrepancy in this group is over 50%. There is also tremendous variability in these discrepancy percentages for the local items, ranging from close to 0 to almost 250%. On the other extreme, the nationally marketed items tend to have calorie amounts very close to what is advertised, with very little variation. The median discrepancy percentage in this group is close to 0, and there is little variability except for an outlier that actually has fewer calories than advertised. The regionally marketed items fall in between local and nationally marketed, both in terms of center and spread. The regionally marketed items do tend to have more calories than stated, but nowhere near as many as the local ones, and the variability is less than with the local items as well.

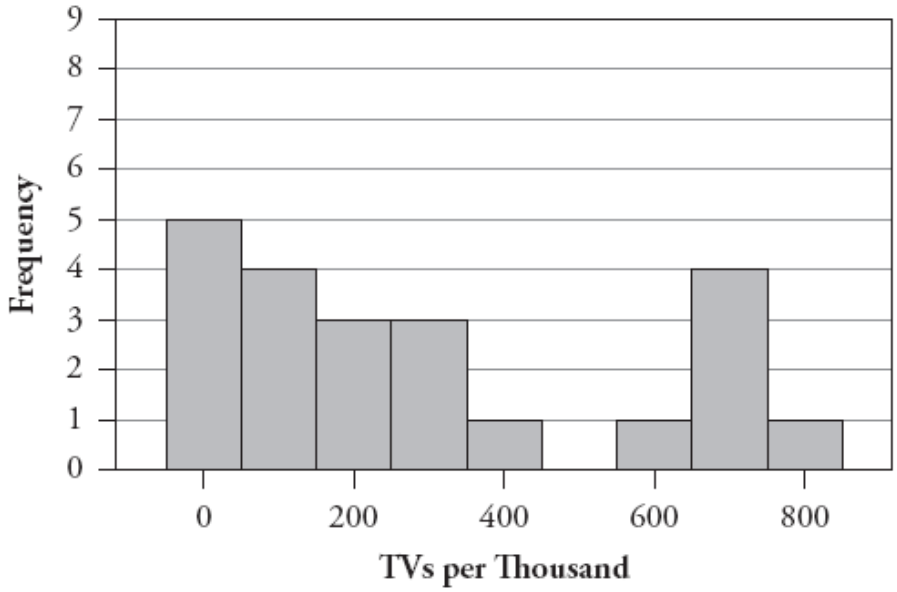
Sample Exam 2B

1. Suppose Ben records the noon temperature in New York City on every day in the month of June, and Frank records the noon temperature in New York City on every day in an entire year. Which one (Ben or Frank) would you expect to have the greater standard deviation of temperatures, or would you expect the standard deviations to be very similar? Explain briefly.

2. The following histogram displays the life expectancies (in years), as reported by *The World Almanac and Book of Facts 2006,* for a sample of 22 countries:

a. Describe (in five words or less) the shape of this distribution.
b. Two measures of center for this distribution are 67.0 and 70.7. One of these is the mean and the other is the median. Which is which? Explain your choice.

Now consider a histogram of the numbers of televisions per thousand people in these countries:



c. Write a paragraph describing key features of this distribution.

3. a. Construct a hypothetical example of ten exam scores (between 0 and 100, inclusive, possibly including repeats) so that the mean is greater than two times the median. Also report the values of the mean and median for your example.
b. Construct another (or different) hypothetical example of ten exam scores (between 0 and 100, inclusive, possibly including repeats) so that the interquartile range equals 50. Also report the values of the quartiles for your example.

4. Suppose a company that has recently fired many of its employees produces the following data on the gender (male/female) and job type (managerial/clerical) of the employees.

Overall	Male	Female
Fired	160	20
Not fired	640	180
Total	800	200
% fired	20%	10%

Managerial	Male	Female
Fired	155	15
Not fired	445	35
Total	600	50
% fired	25.83%	30%

Clerical	Male	Female
Fired	5	5
Not fired	195	145
Total	200	150
% fired	2.50%	3.33%

a. Show that Simpson's paradox holds for these data. (Point out the relevant percentages and the relevant inequalities, i.e., which percentages are greater than which others.)
b. Write a sentence or two explaining, as if to an educated person with no knowledge of statistics, why the paradox happens in this context.

5. The following data are the speaking rates (in words per minute) of presidential candidates during televised debates in late September and early October of 2007:

Republicans:

Brownback	Giuliani	Huckabee	Hunter	McCain	Paul	Romney	Tancredo	Thompson
196	206	207	189	175	187	221	196	182

Democrats:

Biden	Clinton	Dodd	Edwards	Gravel	Kucinich	Obama	Richardson
190	183	222	201	174	189	187	161

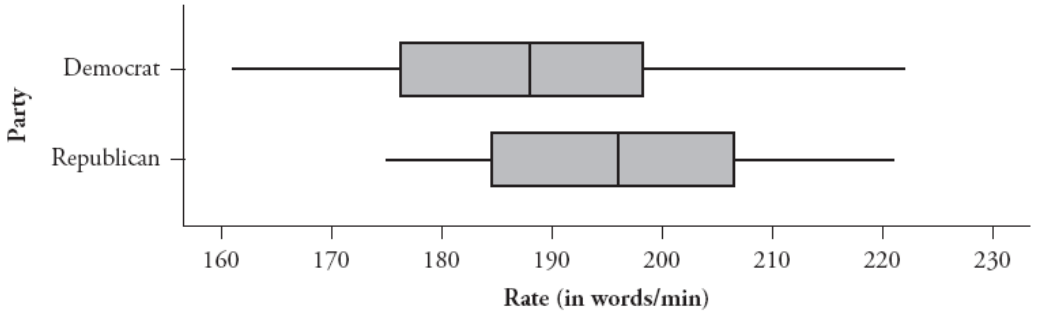
Consider the following computer output, for the speaking rates of candidates when the two parties are combined:

Variable	N	Mean	StDev	Minimum	Q1	Median
Rate (in words/min)	17	192.12	16.01	161.00	182.50	189.00

Variable	Q3	Maximum	Range	IQR
Rate (in words/min)	203.50	222.00	61.00	21.00

a. Use this output to determine whether there are any outliers in the data (again with the two parties combined).

The following boxplots display the speaking rates:



b. Report the median speaking rate for each party as accurately as you can from the graph.
c. Report the interquartile range of the speaking rates for each party as accurately as you can from the graph.
d. Summarize what these boxplots reveal about the speaking rates of the two parties' presidential candidates.

6. For each of the following, indicate whether it is resistant to outliers or not. Do not bother to explain your answers, except for the last one.
a. Mean
b. Median
c. Standard deviation
d. (Lower quartile + upper quartile)/2
e. (Maximum + minimum)/2
f. Mean absolute deviation (Explain your answer.)

Solution to Sample Exam 2B

1. Frank would have the greater standard deviation. Noontime temperatures in New York City vary more throughout the year, which would include very warm and very cold days, than in the month of June, which would have generally warm days.

2. a. The distribution of life expectancies is skewed to the left.
b. Because of the skew to the left, the mean is less than the median. So the mean is 67.0 and the median is 70.7.
c. The distribution of televisions per thousand people has an interesting bimodality. A large group of countries has relatively few (0-300 or so) televisions per thousand people. A smaller but noticeable group has many (600-800 or so) televisions per thousand people. There are a few, but not many, countries in between these extremes.

3. a. One example that works is {20, 20, 20, 20, 20, 20, 100, 100, 100, 100}. The median of these ten scores is 20, and the mean is 52. It's not necessary to have so many duplicate scores. The key is to have six fairly small scores, so the median (the average of the 5th and 6th values) will be fairly small.
b. One example that works is {50, 50, 50, 50, 50, 100, 100, 100, 100, 100}. The lower quartile is 50, and the upper quartile is 100, so the interquartile range is 100 - 50 or 50. It's not necessary to have so many duplicate scores. The key is for the third ordered score (the lower quartile) to be 50 points below the eighth ordered score (the upper quartile).

4. a. Simpson's paradox occurs here because a greater proportion of men than women were fired overall (20% of men were fired compared to 10% of women), but within both job classifications (managerial and clerical), women were fired at a greater rate than men (30% vs. 25.83% for managerial, 3.33% vs. 2.50% for clerical).
b. The explanation is that firings were more likely to occur in managerial positions, which were more likely to be filled by men, so men were fired at a higher rate overall, despite having a lower firing rate for both kinds of jobs, because they had the kinds of jobs (managerial) that were more likely to experience firings.

5. a. The IQR is 20.56, so $1.5 \times$ IQR = 31.5. Therefore, an outlier on the high end would have to be greater than 203.5 + 31.5 or 235 words per minute. None of the candidates spoke this quickly, so there are no high outliers. On the low end, an outlier would have to be less than 182.5 - 31.5 or 151 words per minute. None of the candidates spoke this slowly, so there are no low outliers.
b. Republican candidates tend to speak more quickly than Democrats, as seen by the higher median and quartiles for Republicans. The median speaking rate for Republicans appears to be about 196 words per minute, compared to about 188 words per minute for the Democrats' median. The variability of speaking rates within each party appears to be similar, as the interquartile range appears to be about 20 words per minute for each group. It is hard to judge shape from boxplots, but both boxplots look fairly symmetric. No candidate in either party is labeled as an outlier by the boxplots.

6. a. No, the mean is not resistant.
b. Yes, the median is resistant.
c. No, the standard deviation is not resistant.
d. Yes, this is resistant (because the quartiles are resistant).
e. No, this is not resistant (because the minimum and maximum are not resistant).
f. No, the mean absolute deviation is not resistant. An extreme value would have a large absolute deviation and would influence the mean enough that many values could have large absolute deviations. This, in turn, would affect the mean absolute deviation considerably because the mean is not resistant to outliers.