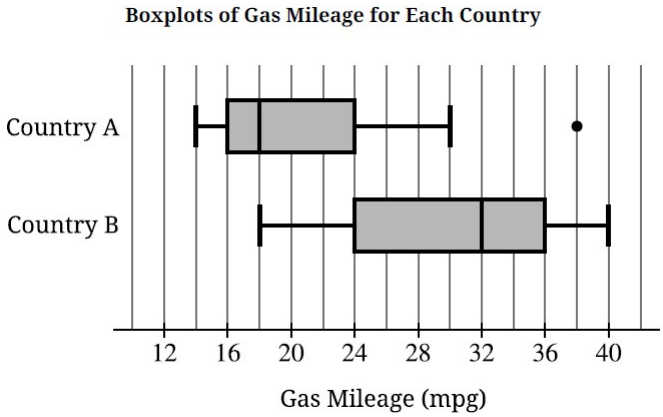
• The manager of an automotive company is interested in comparing the gas mileages for cars manufactured in Country A and cars manufactured in Country B. The manager selected a random sample of 100 cars manufactured in Country A and a random sample of 100 cars manufactured in Country B. The gas mileages for each sample, in miles per gallon (mpg), are summarized in the boxplots.



(A) Compare the distributions of gas mileage for the sample of cars manufactured in Country A and the sample of cars manufactured in Country B.

(B) For the distribution of gas mileage for the sample of cars manufactured in Country A, would you expect the mean to be greater than 18 mpg, less than 18 mpg, or equal to 18 mpg? Justify your answer.

(C) The manager will create a new boxplot with the combined data from the sample of cars manufactured in Country A and the sample of cars manufactured in Country B.

i. What is the range of the combined data set? Justify your answer.

ii. What is a possible value of the median of the combined data set? Justify your answer by referencing the boxplots shown.

Solutions:
(A). The distribution of gas mileage for the sample of cars manufactured in Country A has a lower center than the distribution of gas mileage for the sample of cars manufactured in Country B. The median gas mileage for the sample of cars manufactured in Country A (18 mpg) is less than the median gas mileage for the sample of cars manufactured in Country B (32 mpg).

The range of the gas mileages for the sample of cars manufactured in Country A (24 mpg) is slightly greater than the range of the gas mileages for the sample of cars manufactured in Country B (22 mpg). However, the IQR of the gas mileages for the sample of cars manufactured in Country A (8 mpg) is less than the IQR of the gas mileages for the sample of the cars manufactured in Country B (12 mpg).

The car manufactured in Country A with 38 mpg (the maximum of the sample of cars manufactured in Country A) is an outlier, while the distribution of gas mileage for the sample of cars manufactured

in Country B has no outliers

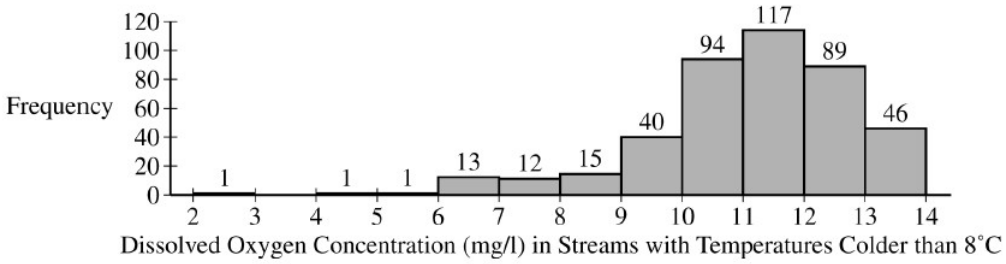
(B). The mean of the distribution of gas mileage for the sample of cars manufactured in Country A is expected to be greater than 18 mpg, the median of the distribution. Because the distribution of gas mileage for the sample of cars manufactured in Country A has an outlier to the right (or is skewed to the right), the mean of the distribution (which is not resistant) is expected to be pulled above the median (which is resistant) toward the higher values of gas mileage.

(C)

i. The maximum value in the combined data is 40 mpg because 40 mpg is the maximum gas mileage for the sample of cars manufactured in Country B, and as shown in the boxplot, all the gas mileages for the sample of cars manufactured in Country A are less than 40 mpg. The minimum value in the combined data is 14 mpg, because 14 mpg is the minimum mpg for the sample of cars manufactured in Country A, and as shown in the boxplot, all the gas mileages for the sample of cars manufactured in Country B are greater than 14. Thus, the range of the combined data set is 40-14=26 mpg.

ii. In the combined data, there are 200 gas mileages. The median is a value where at least half, or 100, of the gas mileages in the combined data are less than or equal to the median value and at least half, or 100, of the gas mileages in the combined data are greater than or equal to the median value. From the boxplot for the sample of cars manufactured in Country A, the third quartile, Q3, is 24 mpg indicating there are at least 75 gas mileages less than or equal to 24 mpg and at least 25 gas mileages greater than or equal to 24 mpg. From the boxplot for the sample of cars manufactured in Country B, the first quartile, Q1, is 24 mpg indicating there are at least 25 gas mileages less than or equal to 24 mpg and at least 75 gas mileages greater than or equal to 24 mpg. Thus, in the combined data set, there are at least 100 gas mileages less than or equal to 24 mpg and at least 100 gas mileages greater than or equal to 24 mpg, which implies 24 is the value of the median of the combined data set.

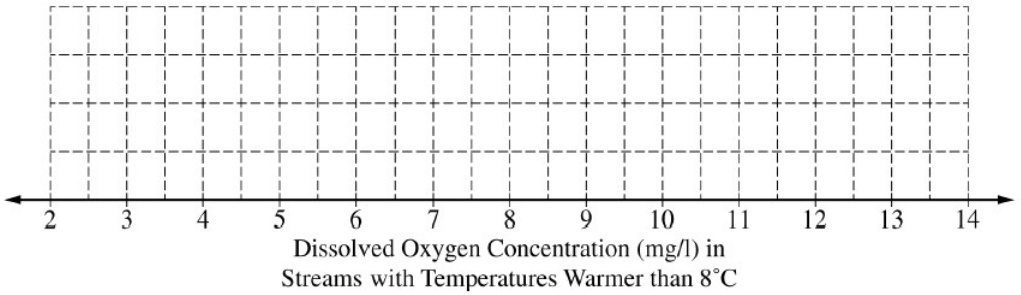
• As part of a study on the chemistry of Alaskan streams, researchers took water samples from many streams with temperatures colder than 8°C and from many streams with temperatures warmer than 8°C. For each sample, the researchers measured the dissolved oxygen concentration, in milligrams per liter (mg/l).



(A) The researchers constructed the histogram shown for the dissolved oxygen concentration in streams from the sample with water temperatures colder than 8°C. Based on the histogram, describe the distribution of dissolved oxygen concentration in streams with water temperatures colder than 8°C.

Min	Q1	Median	Q3	Max	Mean	Std. Dev.
2.10	4.39	5.43	6.12	13.45	5.54	1.64

(B) The researchers computed the summary statistics shown in the table for the dissolved oxygen concentration in streams from the sample with water temperatures warmer than 8°C. Use the summary statistics to construct a box plot for the dissolved oxygen concentration in streams with water temperatures warmer than 8°C. Do not indicate outliers.



(C) The researchers believe that streams with higher dissolved oxygen concentration are generally healthier for wildlife. Which streams are generally healthier for wildlife, those with water temperature colder than 8°C or those with water temperature warmer than 8°C? Using characteristics of the distribution of dissolved oxygen concentration for temperatures colder than 8°C and characteristics of the distribution of dissolved oxygen concentration for temperatures warmer than 8°C, justify your answer.

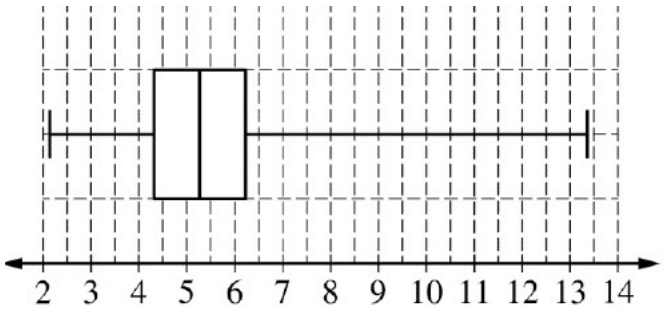
Solutions:

(A) The histogram of dissolved oxygen concentration in Alaskan streams with water temperatures colder than 8°C is unimodal and skewed left with a median between 11 and 12 mg/l.

The first quartile is in the bin from 10-11 mg/l and the third quartile is in the bin from 12-13 mg/l, so the IQR is approximately 2 mg/l.

There do not appear to be any high outliers, but there are several potential low outliers because the values in the 2-3, 4-5, and 5-6 bins are all certainly more than 1.5 IQR below the first quartile.

(B)



Dissolved Oxygen Concentration (mg/l) in
Streams with Temperatures Warmer than 8°C

(C) If the researchers' belief is correct, then streams with water temperature colder than 8°C are healthier for wildlife.

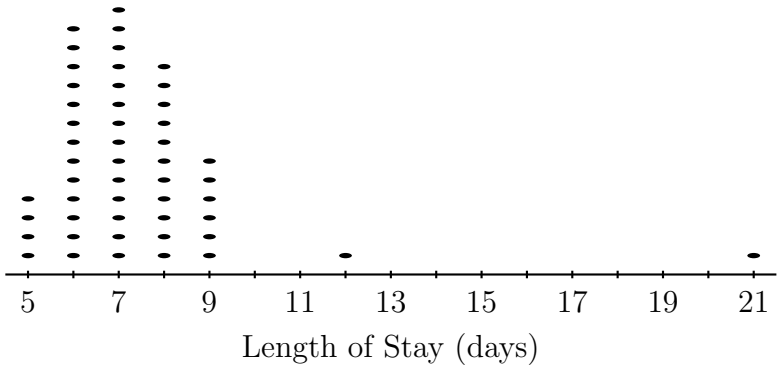
The distribution of dissolved oxygen concentration for colder streams has a higher center because its median (between 11 mg/l and 12 mg/l) is larger than the median for warmer streams (5.43 mg/l).

The shape of the distribution of dissolved oxygen concentration for colder streams is different from the shape of the distribution for warmer streams. The distribution of values of dissolved oxygen concentration for colder streams is skewed to the left but the distribution of values for warmer streams is skewed to the right.

Both distributions have a similar spread because they both have similar IQR values — approximately 2 mg/l for the colder streams and 1.73 mg/l for the warmer streams.

• The length of stay in a hospital after receiving a particular treatment is of interest to the patient, the hospital, and insurance providers. Of particular interest are unusually short or long lengths of stay.
A random sample of 50 patients who received the treatment was selected, and the length of stay, in number of days, was recorded for each patient. The results are summarized in the following table and are shown in the dotplot.

Length of stay (days)	5	6	7	8	9	12	21
Number of patients	4	13	14	11	6	1	1



(A) Determine the five-number summary of the distribution of length of stay.

(B) Consider two rules for identifying outliers, method A and method B. Let method A represent the $1.5 \times$ IQR rule, and let method B represent the 2 standard deviations rule.

(i) Using method A, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

(ii) The mean length of stay for the sample is 7.42 days with a standard deviation of 2.37 days. Using method B, determine any data points that are potential outliers in the distribution of length of stay. Justify your answer.

(C) Explain why method A might identify more data points as potential outliers than method B for a distribution that is strongly skewed to the right.

Solutions:

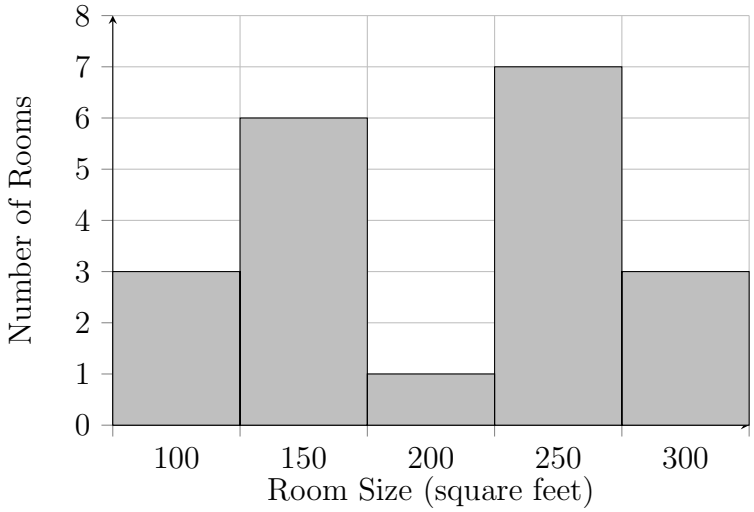
(A) The five-number summary of the distribution of length of stay is:
Minimum = 5 days
Lower quartile (Q_1) = 6 days
Median = 7 days
Upper quartile (Q_3) = 8 days
Maximum = 21 days

(B) (i) The patients who stayed for 12 days and 21 days are considered outliers using method A. An outlier using method A is a value greater than $1.5 \times$ IQR above the third quartile (Q_3) or more than $1.5 \times$ IQR below the first quartile (Q_1). Because
$Q_1 - 1.5 \times \text{IQR} = 6 - 1.5(8-6) = 3$, then any values below 3 are considered outliers. There are no such values. Because
$Q_3 + 1.5 \times \text{IQR} = 8 + 1.5(8-6) = 11$, then any values above 11 are considered outliers.

(ii) The patient who stayed for 21 days is the only outlier using method B. An outlier using method B is a value located 2 or more standard deviations above, or below, the mean. Because
$\text{Mean} \pm 2 \times \text{SD} = 7.42 \pm 2(2.37)$, then any value that is outside of the interval $(2.68, 12.16)$ is considered an outlier.

(C) Quartiles and the IQR are less sensitive to extreme values in strongly skewed distributions than the mean and standard deviation. Relative to the quartiles, the mean is pulled more toward the extreme values in the longer tail of a strongly skewed distribution.
For a distribution that is strongly skewed to the right, the sample mean will be pulled more toward the extreme values in the longer right tail of the distribution than the sample median, and the ratio of the standard deviation to the IQR will tend to be larger than that for more nearly symmetric distributions. As a result, this pulls the value of the outlier criterion for method B, Mean $+ 2 \times SD$, more toward the extreme values in the right tail of the distribution than the outlier criterion for method A, $Q_3 + 1.5 \times IQR$. This decreases the ability of method B to identify outliers relative to method A, which means that method A may identify more outliers than method B for a distribution that is strongly skewed to the right.

• The sizes, in square feet, of the 20 rooms in a student residence hall at a certain university are summarized in the following histogram.

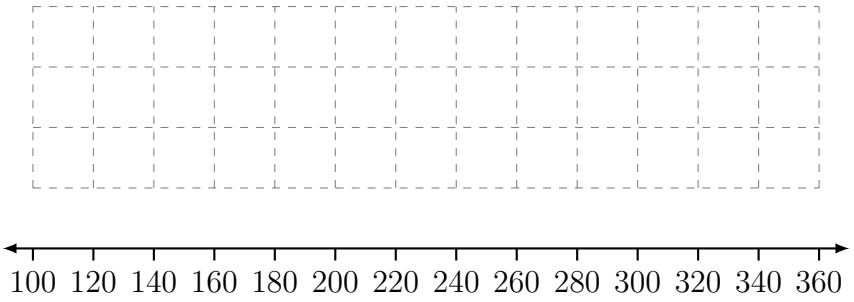


(A) Based on the histogram, write a few sentences describing the distribution of room size in the residence hall.

(B) Summary statistics for the sizes are given in the following table.

Mean	Standard Deviation	Min	Q1	Median	Q3	Max
231.4	68.12	134	174	253.5	292	315

Determine whether there are potential outliers in the data. Then use the following grid to sketch a boxplot of room size.

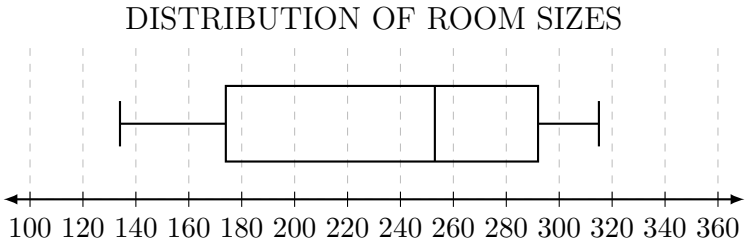


(C) What characteristic of the shape of the distribution of room size is apparent from the histogram but not from the boxplot?

Solutions:

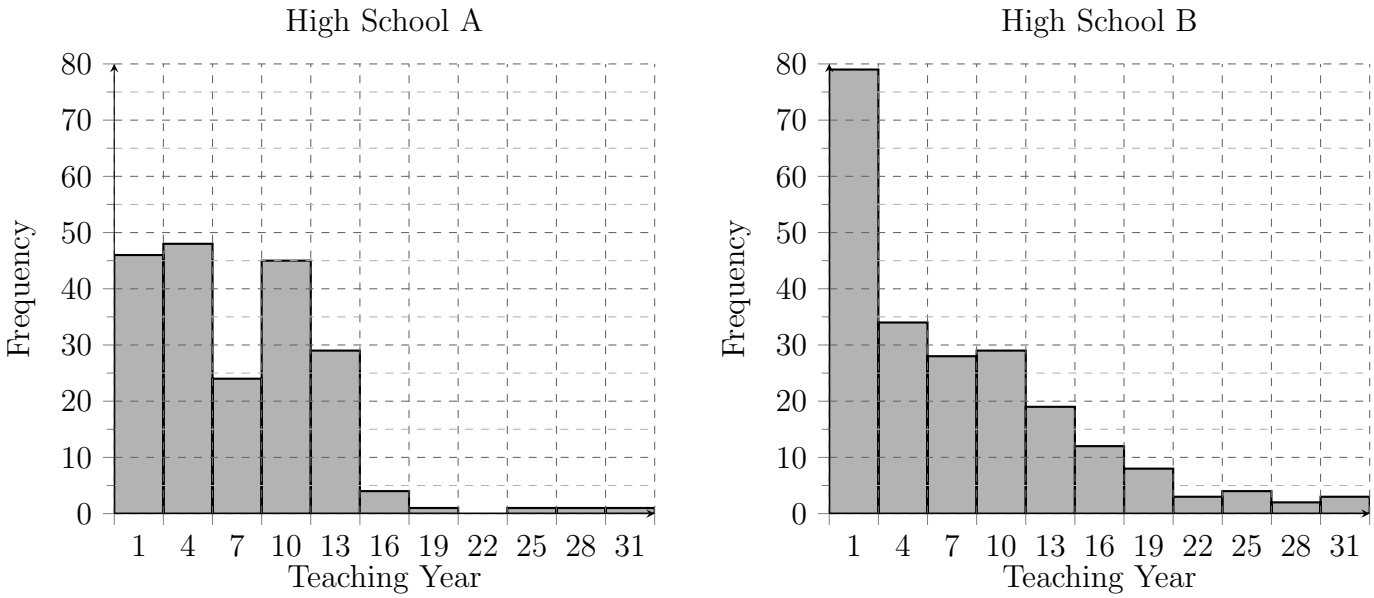
(A) The distribution of the sample of room sizes is bimodal and roughly symmetric with most room sizes falling into two clusters: 100 to 200 square feet and 250 to 350 square feet. The center of the distribution is between 200 and 300 square feet. The range of the distribution is between 150 and 250 square feet. There are no apparent outliers.

(B) The interquartile range is IQR $= 292 - 174 = 118$ square feet. There are no potential outliers because the minimum room size of 134 square feet does not fall below $Q_1 - 1.5(\text{IQR}) = -3$ square feet, and the maximum room size of 315 square feet does not exceed $Q_3 + 1.5(\text{IQR}) = 469$ square feet.



(C) The histogram clearly shows the bimodal nature of the distribution of room sizes, but this is not apparent in the boxplot.

• The following histograms summarize the teaching year for the teachers at the high schools A and B.



Teaching year is recorded as an integer, with first-year teachers recorded as 1, second-year teachers recorded as 2, and so on. Both sets of data have a mean teaching year of 8.2, with data recorded from 200 teachers at High School A and 221 teachers at High School B. On the histograms, each interval represents possible integer values from the left endpoint up to but not including the right endpoint.

(A) The median teaching year for one high school is 6, and the median teaching year for the other high school is 7. Identify which high school has each median and justify your answer.

(B) An additional 18 teachers were not included with the data recorded from the 200 teachers at High School A. The mean teaching year of the 18 teachers is 2.5. What is the mean teaching year for all 218 teachers at High School A?

(C) The standard deviation of the teaching year for the 221 teachers at High School B is 7.2. If one teacher is selected at random from High School B, what is the probability that the teaching year for the selected teacher will be within 1 standard deviation of the mean of 8.2 ? Justify your answer.

Solution

(A) The median teaching year for High School A is any value with 100 data values at or below it and 100 data values at or above it. The median teaching year for High School B is the 111th value in the ordered list of values. For High School A the median is in the interval that starts at 7 and ends just before 10, because there are only 94 data values below 7 and 106 data values of at least 7. Therefore the median cannot be less than 7. For High School B the median is in the interval that starts at 4 and ends just before 7 because there are more than half (113) of the data values less than 7. Therefore the median must be less than 7. So High School A must be the one with a median of 7, and High School B must be the one with a median of 6.

Another way to determine which school has the median of 7 is to notice that the distribution for High School B is highly skewed to the right, whereas the distribution for High School A is bimodal with a few possible outliers on the right. A distribution that is highly right-skewed is likely to have a substantially larger mean than median. The mean of both distributions is given as 8.2 years, so it makes sense that the highly right-skewed distribution (High School B) is the one with the bigger gap between the mean and median and, therefore, the one with the lower median of 6.

(B) The mean for the original 200 teachers was given as 8.2 years, and the mean for the additional 18 teachers is 2.5 years. Therefore the mean for the combined data set is:

$$\frac{(200)(8.2)+(18)(2.5)}{200+18} = \frac{1,640+45}{218} \approx 7.73 \text{ years.}$$

(C) The interval mean plus or minus 1 standard deviation on either side of the mean is 8.2 ± 7.2, or from 1.0 year to 15.4 years. Because teaching year is recorded as an integer, the interval includes teaching years 1 to 15. The number of teachers in that interval can be found by adding the heights of the five bars in the histogram for the intervals from 1 to 16, which includes $79+34+28+29+19 = 189$. Therefore the probability is $\frac{189}{221} \approx 0.8552$.

Problems adapted from the College Board tests.