

Chapter 2: Organizing, Displaying, and Describing Data

Qualitative Data

Qualitative data can be organized and displayed using tables and graphs.

Tables: showing counts (frequencies) and percentages or proportions (relative frequencies)

Bar Graphs: The length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

Pie Charts: categories of data are represented by wedges in the circle proportional in size to the percent of individuals in each category. A pie chart can be used when you are trying to display the relationship between parts and the whole and can be effective if you have qualitative data that fall into only a few categories.

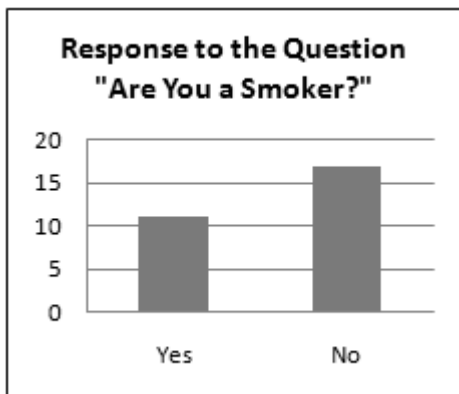
Example: A group of 28 students were asked whether or not they smoke. The following data set was obtained:

Y Y Y Y Y Y Y Y Y Y Y N N
N N N N N N N N N N N N N
N N

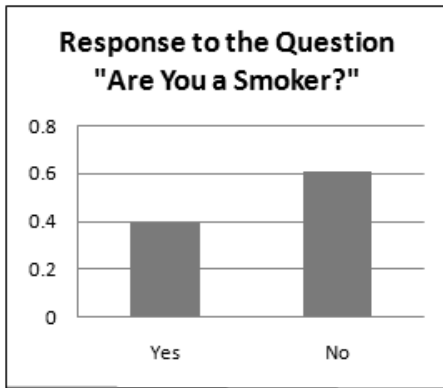
This data set can be organized into a frequency table by counting the number of yes and no responses.

Smoker	Frequency	Relative Frequency
Yes	11	0.393
No	17	0.607

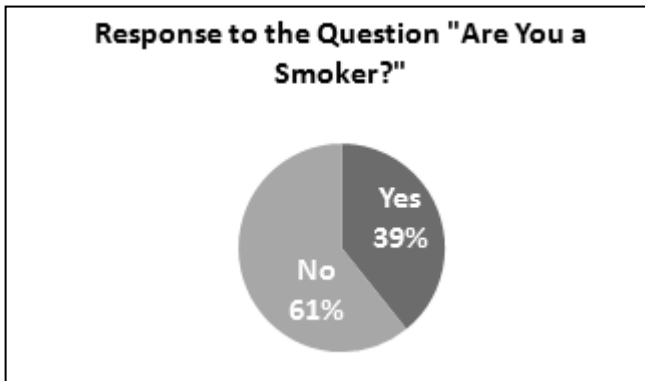
Here is a bar chart with frequency on the vertical axis.



Here is a bar chart with relative frequency on the vertical axis.



Here is a pie chart.



This data can be broken out further by gender. For example, suppose that of the 11 smokers, 4 are male and 7 are female. Suppose that of the 17 nonsmokers, 10 are male and 7 are female. We can organize this information into a contingency table.

	Male	Female	Total
Smoker	4	7	11
Nonsmoker	10	7	17
Total	14	14	28

We will talk about contingency tables later in the course.

Quantitative Data

Quantitative data can also be organized and displayed using tables and graphs. Quantitative data may be presented individually (ungrouped) or grouped into intervals.

Organizing Quantitative Data Using a Frequency Table (Ungrouped Data)

Example: A group of 28 high school students were surveyed and asked the number of miles they live from their high school. The following data was obtained:

1 1 1 2 2 2 2 3 3 3 3 3 3
3 4 4 4 4 4 4 4 4 5 5 5 5
5 5

Number of miles (X)	Frequency	Relative Frequency
1	3	0.107
2	4	0.143
3	7	0.250
4	8	0.286
5	6	0.214

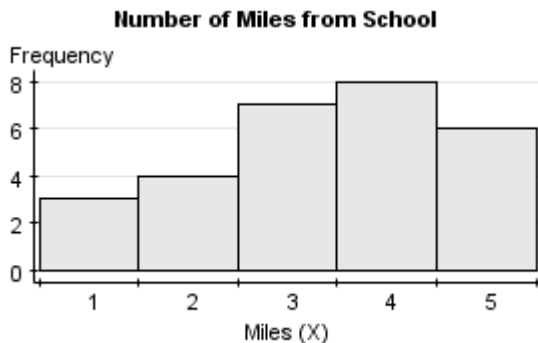
This is called an ungrouped frequency table because there are individual values in the first column.

There are five classes in this table since there are five rows (not including the column heading).

Displaying Ungrouped Quantitative Data Using a Histogram

Histogram – a graph that consists of contiguous boxes (i.e. the edges of the boxes touch). The horizontal axis is labeled with what the data represents. The vertical axis is labeled with either “frequency” or “relative frequency.”

Here is a histogram for this frequency table above. Notice that the horizontal axis is labeled “miles” and vertical axis is labeled “frequency.” The bars touch.



Example: A bank manager reported the following waiting times in minutes for a sample of customers.

1 0 0 3 4 2 2 3 1 2 3 1 0 0
2 3 0 5 3

Create an ungrouped frequency table and a histogram.

Solution:

Time (X)	Frequency	Relative Frequency
0	5	0.263
1	3	0.158
2	4	0.211
3	5	0.263
4	1	0.053
5	1	0.053

Here is the histogram associated with this frequency table.



Organizing Quantitative Data Using a Frequency Table (Grouped Data)

The first column of a grouped frequency table consists of intervals of numbers (instead of individual numbers). We can also call these intervals “classes” or “bins.” It is very important that each observation be counted only in one class.

Example: A group of 28 students were asked about the amount of sleep they received the previous night. The following data were obtained:

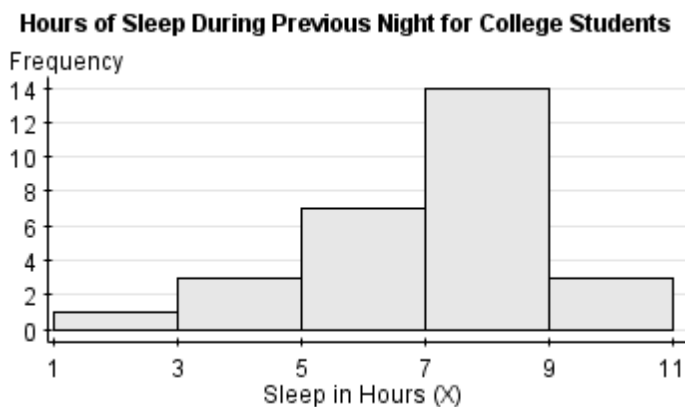
1 3 4 4 5 5 5 6 6 6 6 7 7
 7 7 7 7 7 8 8 8 8 8 8 8 9
 9 10

Amount of sleep in hours (X)	Frequency	Relative Frequency
1 - 3	1	0.036
3 - 5	3	0.107
5 - 7	7	0.250
7 - 9	14	0.500
9 - 11	3	0.107

It is very important that each observation be counted only in one interval. In this case, we will consider the interval from 1 - 3 to include all numbers from 1 up to but not including 3. The interval from 3 - 5 includes all numbers from 3 - 5 but not including 5.

There are five classes in this table since there are five rows (not including the column heading).

Below is a histogram for this frequency table.



Questions:

How many students got 7 hours of sleep or more? 17

What proportion of students got less than 3 hours of sleep? 0.036

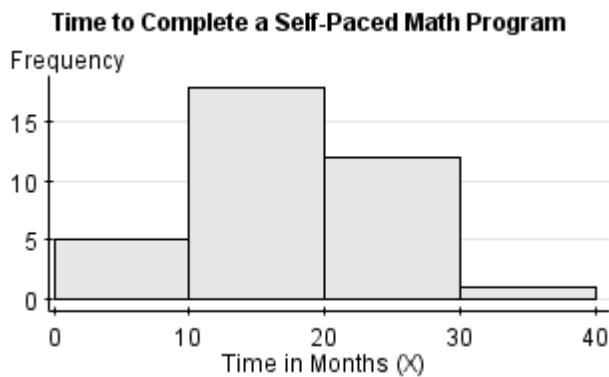
Example: The time (in months) for students to complete a self-paced math program were recorded, and the data are shown below.

24 20 16 32 14 22 2 12 24 6 10 20
 8 16 12 24 14 20 18 14 16 18 20 22
 24 26 28 18 14 10 12 24 6 12 18 16

Create a grouped histogram. Use a class width of 10 and a starting value of 0.

Solution: Create the classes for the frequency table. Then count the number of values falling into each interval.

Time (X)	Frequency	Relative Frequency
0 -- 10	4	0.111
10 - 20	18	0.500
20 - 30	13	0.361
30 - 40	1	0.028



Questions:

How many people took between 10 and 20 months to complete the program? *18*

What proportion of people took fewer than 10 months to complete the program? *0.111*

Stemplot (or Stem-and-Leaf Plot)

Stemplots-- A graph for which every number is split into two parts—a stem and a leaf. It allows us to reconstruct the original data if desired. Stemplots work well for small data sets, but not as well for large data sets.

Example: Create a stem and leaf plot for the following ages:

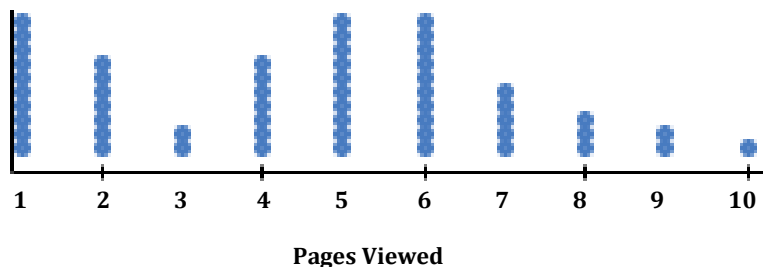
31, 32, 32, 32, 33, 35, 36, 36, 37, 37, 38, 39, 39, 40, 40, 40, 41, 42, 42, 43, 43, 44, 45, 45, 46, 46, 47, 48, 48, 51, 53, 55, 56, 56, 60, 60, 61, 62, 76, 78, 82, 97

Stem (tens)	Leaves (units)
3	1 2 2 2 3 5 6 6 7 7 8 9 9
4	0 0 0 1 2 2 3 3 4 5 5 6 6 7 8 8
5	1 3 5 6 6
6	0 0 1 2
7	6 8
8	2
9	7

Dotplot

A dotplot is a graph in which each data value is plotted as a point (or dot) along a scale of values.

Example: The following dotplot shows the number of pages viewed on a website by an individual visitor.



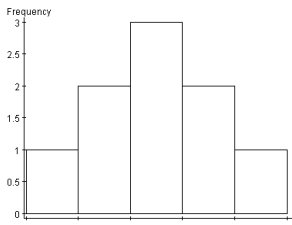
Characteristics of Graphical Displays

When looking at a graphical display of data, it's a good idea to examine the following features:

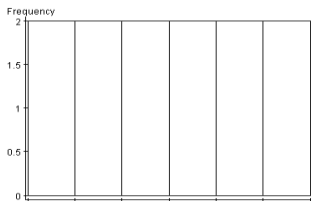
- Center—What is a typical value in the data set?
- Spread—Are the values clustered tightly together or are they spread out?
- Gaps—Are there areas in the graph where there are no values?
- Outliers—Are there extreme values?
- Shape—Is there a pattern to the data? See below.

Distribution Shapes

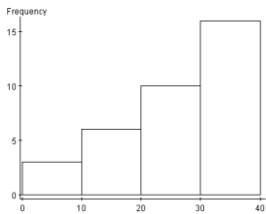
Symmetric—One side is (approximately) a mirror image of the other. The example below is a bell-shaped, symmetric distribution.



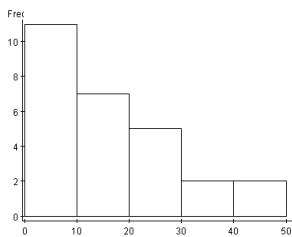
Uniform—Every class has equal frequency.



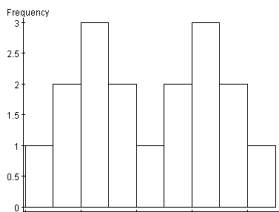
Skewed Left—The distribution has a long left tail.



Skewed Right—The distribution has a long right tail.



Bimodal—The distribution has two peaks.



Examples: What type of distribution do you think that the following variables would have? (*answers will vary*)

1. Hours spent by U.S. adults watching football on Thanksgiving Day.
2. Amount won by all people playing a particular state's lottery last week.
3. Ages of the faculty members at your school.
4. Last digit of phone numbers on your campus.
5. Heights of 30-year-old men and women.

Numerical Descriptions of Data

Measures of Center

Measure of Center—a value at the center or middle of data set, or a “typical” value.

Two Measures of Center

- Mean—The average of a set of values.
- Median—The middle value of an ordered data set.

Measuring Center: The Mean

Notation:

n is the number of data values in a sample

N is the number of data values in a population

X is the variable used to represent the individual data values

Σ is the sum of a set of values

Sample mean: $\bar{x} = \frac{\Sigma x}{n}$

Population mean: $\mu = \frac{\Sigma x}{N}$

Example: Suppose there are five people with the following incomes:

18,000 20,000 22,000 24,000 30,000

The mean income is \$22,800.

Measuring Center: The Median

The sample median (M) is the middle value of an ordered data set. To find the sample median,

1. Arrange the numbers in increasing order.
2. The median is the middle number. If there are two numbers in the middle, the median is the average of the two numbers.

Example: Suppose there are five people with the following incomes:

18,000 20,000 24,000 22,000 30,000

The median income is \$22,000.

Suppose that for the previous example, the highest-paid person makes \$30,000,000 instead of \$30,000 . Here is the new salary data.

18,000 20,000 24,000 22,000 30,000,000

What are the mean and median?

The mean is \$6,016,800.

The median is still \$22,000.

Comments

1. Note that the mean is affected by extreme data values. The median is resistant to extreme values.
2. If a distribution is symmetric, the mean and median are the same.
3. If a distribution is left-skewed (long tail on the left), the mean is less than the median.
4. If a distribution is right-skewed, the mean is greater than the median.

Measures of Spread

Measure of spread tell us if the values in a data set are clustered together or spread out. We will look at three measures of spread:

- Range
- Standard Deviation
- Interquartile Range (IQR)

We will look at the first two measures of spread now, and the IQR later.

Measuring Spread: The Range

The range is the difference between the largest and smallest values in a data set.

Example: Suppose there are five people with the following incomes:

18,000 20,000 22,000 24,000 30,000

What is the range?

Range = $30,000 - 18,000 = 12,000$

Note that the range is not resistant to extreme values since it depends on only the maximum and minimum values.

Measuring Spread: The Standard Deviation

The standard deviation provides a measure of the spread of data values around the mean of a data set. Values close together will yield a small standard deviation, whereas values spread farther apart will yield a larger standard deviation.

The standard deviation as “roughly, the average distance of the observations from the mean.”

$$\text{Sample standard deviation: } s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$\text{Population standard deviation : } \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Example: Find the standard deviation for the following sample data.

1 3 5 7 9

Solution:

The mean of this data set is 5.

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	$1 - 5 = -4$	16
3	$3 - 5 = -2$	4
5	$5 - 5 = 0$	0
7	$7 - 5 = 2$	4
9	$9 - 5 = 4$	16

$$\sum(x - \bar{x})^2 = 40$$

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{40}{5 - 1}} = \sqrt{10} = 3.2$$

Interpretation: These values are roughly 3.2 units away from their mean of 5 on average.

Characteristics of standard deviation

1. It is a measure of variation from all values from the mean
2. It is usually positive. It is zero only when all data values are the same.
3. The presence of outliers can cause a big increase in the standard deviation.
4. The units of standard deviation are the same as the units of the original data values.

Sample Variance

Sample variance: $s^2 =$ square of the sample standard deviation s

Population variance: $\sigma^2 =$ square of the population standard deviation σ

Using the Standard Deviation

Unusual Values

In general, we can consider values that are more than two standard deviations away from the mean as “unusual.”

Example: IQ scores have a mean of 100 and a standard deviation of 15.

a. Find the IQ scores that are two standard deviations above and below the mean.

Minimum usual value: $100 - 2(15) = 70$

Maximum usual value: $100 + 2(15) = 130$

b. Using the definition, would an IQ score of 140 be considered unusual?

Yes, because it higher than the maximum usual value.

Range Rule of Thumb

The range of a data set is approximately four times the standard deviation. There are two ways to use this:

1. You are given a data set and asked to estimate the standard deviation, or
2. You are given only the mean and the standard deviation of a data set (but not the original data) and asked to estimate the range.

Example of case 1: Estimate the standard deviation for the following data set:

17 38 27 14 18 34 16 42 28 24 40 20 23 31 37 21 30 25

Solution: Range = max – min = 42 – 14 = 28

Estimated sample standard deviation = 7

Actual sample standard deviation (s) = 8.66

Example of case 2: Heights of women have a mean of 63.6 inches and a standard deviation of 2.5 inches. Estimate the range.

Solution:

$$63.6 - 2(2.5) = 58.6 = \text{minimum "usual" value}$$

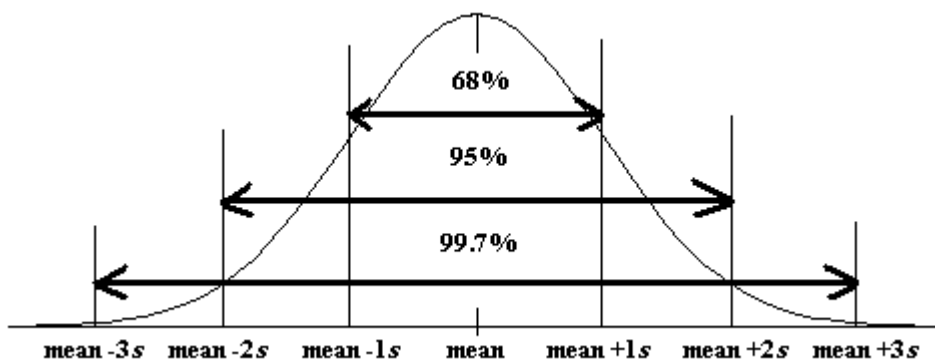
$$63.6 + 2(2.5) = 68.6 = \text{maximum "usual" value}$$

The range is approximately $68.6 - 58.6 = 10$

The Empirical Rule

For data sets having a distribution that is approximately bell-shaped,

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.



Example: The average recovery time for a certain type of surgery is 4 months with a standard deviation of 0.5 month. The distribution of recovery times is bell-shaped.

a. What percentage of people have recovery times between 3 and 5 months?

b. Approximately 68% of people recover in what time period?

Solution:

a. The value 3 is two standard deviations below the mean; the value 5 is two standard deviations above the mean. The empirical rule tells us that 95% of values fall within 2 standard deviations of the mean. Therefore, approximately 95% of recovery times are between 3 and 5 months.

b. Approximately 68% of people will have a recovery time of between 3.5 and 4.5 months. These limits are one standard deviation above and below the mean.

Z-scores

A standard score, or z score, is the number of standard deviations that a given value x is located above or below the mean. It can be used to compare values from different data sets.

Example: IQ scores have a mean of 100 and a standard deviation of 15. Find the z-scores corresponding to the following IQ scores: 115, 130, 85.

Solution:

IQ score (X)	Z	
115	1	<i>(115 is one standard deviation above the mean.)</i>
130	2	<i>(130 is two standard deviations above the mean.)</i>
85	-1	<i>(85 is two standard deviations below the mean.)</i>

Whenever a value is less than the mean, its corresponding z score is negative.
Whenever a value is greater than the mean, its corresponding z score is positive.

A general formula for computing z-scores is:

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

Example: Find the z-score corresponding to a IQ score of 109.

Solution: $z = (109 - 100)/15 = 0.6$

We can also go in the opposite direction, converting z-scores to IQ scores.

Example: Find the IQ scores corresponding to the following z-scores: -2, 3, 0.

Solution:

IQ score (X)	Z	
70	-2	<i>(70 is two standard deviations below the mean.)</i>
145	3	<i>(145 is three standard deviations above the mean.)</i>
100	0	<i>(100 is the mean.)</i>

A general formula for finding a value when given a z-score is

$$X = \text{mean} + Z * \text{standard deviation}$$

Using Z-scores to Make Comparisons Between Different Groups

Example: Suppose for all graduating seniors, marketing majors earn an average of \$42,500 with a standard deviation of \$1000. Kate is a marketing major with an offer of \$43,000.

Accounting majors earn an average of \$46,000 with a standard deviation of \$1500. Tom is an accounting major with an offer of \$44,500.

Find the z-scores associated with their starting salaries.

Solution:

$$\text{Kate: } z = \frac{43,000 - 42,500}{1,000} = 0.5$$

$$\text{Tom: } z = \frac{44,500 - 46,000}{1,500} = -1.0$$

Although Tom's starting salary is higher, Kate's salary is higher relative to her peer group because her z-score is higher.

Measures of Relative Standing: Percentiles and Quartiles

A percentile is the value of a variable below which a certain percent of observations fall. For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found.

The 25th percentile is also known as the first quartile (Q_1), the 50th percentile as the median or second quartile (Q_2), and the 75th percentile as the third quartile (Q_3).

Finding Quartiles

To compute quartiles:

1. Put the data in order from smallest to largest.
2. Find the median; this is the second quartile.
3. The first quartile Q_1 is the median of the lower half of the data. If there are an odd number of values in the data set, do not include the middle number in the lower half.
4. The third quartile Q_3 is the median of the upper half of the data. If there are an odd number of values in the data set, do not include the middle number in the upper half.

Previously, it was mentioned that the interquartile range (IQR) is a measure of spread of a data set. The interquartile range (IQR) is the difference between the third and first quartiles, and is a measure of the spread of the middle 50% of the data.

$$\text{IQR} = Q_3 - Q_1$$

Example (*even number of values*): Find the quartiles and IQR for the following data set, which is the number of inches of snow reported in randomly selected U.S. cities for September 1 through January 10.

9 8 13 4 3 21 15 3 11 24 34 17

Solution:

Put the data in order.

3 3 4 8 9 11 13 15 17 21 24 34

The median is 12, which is the average of the two middle numbers 11 and 13.

The first quartile is 6, which is the median of the lowest six numbers (2, 2, 4, 6, 9, 11).

The third quartile is 19, which is the median of the highest six numbers (13, 15, 17, 21, 24, 34).

The interquartile range for the above data set is 13.

Example (*odd number of values*): Find the quartiles and IQR for the following data set, which contains quiz grades.

90 92 93 88 95 88 97 87 98 57 31

Solution:

Put the data in order.

31 57 87 88 88 90 92 93 94 97 98

The median is 90, the middle number.

The first quartile is 87, which is the median of the lowest five numbers (31, 57, 87, 88, 88).

The third quartile is 94, which is the median of the highest five numbers (92, 93, 94, 97, 98).

The interquartile range for the above data set is 7.

Outliers

Outlier – an unusually high or low value in a data set. Data values that fall beyond the following limits are considered outliers.

Lower limit: $Q_1 - 1.5 * (IQR)$

Upper limit: $Q_3 + 1.5 * (IQR)$

Example: Determine if there are any outliers for the following data set.

9 8 13 4 3 21 15 3 11 24 41 17

Solution:

For this data set, the median = 12, $Q_1 = 6$, $Q_3 = 19$.

Therefore, the IQR = $19 - 6 = 13$.

Lower Limit = $6 - 1.5 * 13 = -13.5$

Upper Limit = $19 + 1.5 * 13 = 38.5$

Are there any values in the data set that are less than -13.5 or greater than 38.5? Yes: 41. Therefore, 41 is an outlier.

Five-Number Summaries and Boxplots

The five-number summary consists of the following:

- minimum
- first quartile (Q_1)
- median
- third quartile (Q_3)
- maximum

Example (*odd number of values*): Given the following data set, find the five-number summary.

3 2 1 4 4 7 15 12 8 10 15

Solution: Put the numbers in order from smallest to largest.

1 2 3 4 4 7 8 10 12 15 15

minimum = 1

median = 7

$Q_1 = 3$

$Q_3 = 12$

maximum = 15

Example: (*even number of values*) Given the following data set, find the five-number summary.

2 4 7 8 1 3 4 10 12 15

Solution: Put the numbers in order from smallest to largest.

1 2 3 4 4 7 8 10 12 15

minimum = 1
median = 5.5
 $Q_1 = 3$
 $Q_3 = 10$
maximum 15

A boxplot is a graphical representation of the five-number summary.

To create a boxplot:

1. Find the five-number summary.
2. Make a box with ends at the quartiles Q_1 and Q_3
3. Draw a line in the box at the median.
4. Check for outliers using the $1.5 * IQR$ rule and if any, plot them individually.
5. Extend lines from end of box to smallest and largest observations that are not outliers.

Example: The following data set consists of the waiting times for a sample of calls to a customer service center. Find the five-number summary and create a boxplot.

2, 15, 7, 3, 1, 4, 8, 10, 4, 15, 20, 12

Solution-

Find the five-number summary. Put the data in order.

1, 2, 3, 4, 4, 7, 8, 10, 12, 15, 15, 20

Median = 7.5 (the average of the two numbers in the middle)
 $Q_1 = 3.5$
 $Q_3 = 13.5$
Minimum = 2
Maximum = 20

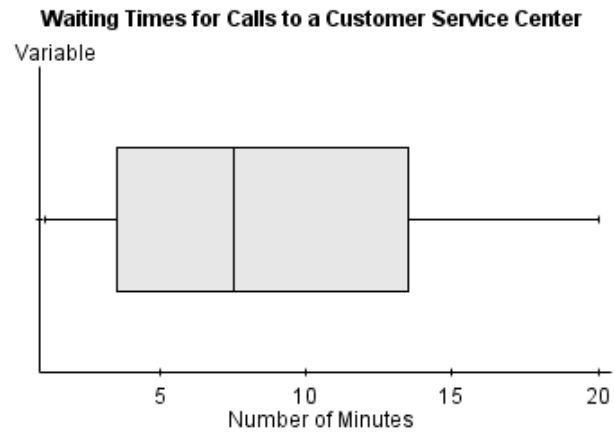
$$\text{IQR} = Q_3 - Q_1 = 10$$

$$1.5 * \text{IQR} = 15$$

$$\text{Lower limit} = Q_1 - 1.5 * \text{IQR} = 3.5 - 15 = -11.5$$

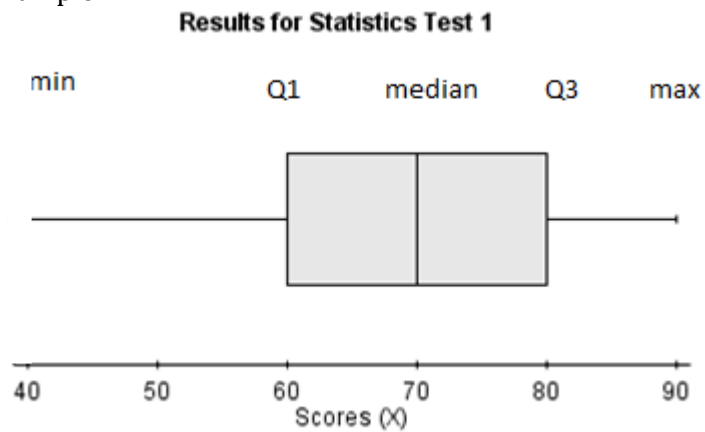
$$\text{Upper limit} = Q_3 + 1.5 * \text{IQR} = 13.5 + 15 = 28.5$$

There are no outliers since there are no values outside of the range (-11.5, 18.5).



Interpretation of Boxplots

Example:

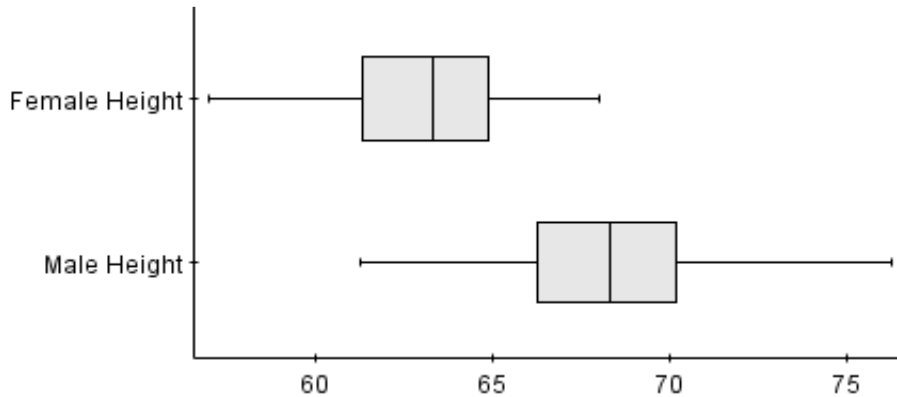


Here is some information that we could get from this boxplot:

- 50% of students scored between 60 and 80.
- The bottom 25% of students scored between 40 and 60.
- The top 25% of students scored between 80 and 90.
- There were no outliers.
- The minimum score on this test was 40.
- The maximum score on this test was 90.

Comparing Boxplots

Example: Heights in inches were recorded for a sample of men and a sample of women. The following boxplots were created.



Here are some observations from the graph:

- The minimum female height in the sample was about 57 inches; the maximum female height was about 68 inches.
- The minimum male height in the sample was about 61 inches; the maximum male height was about 76 inches.
- The median female height was about 63 inches; the median male height was about 68 inches.
- The middle 50% of female heights were between 61 to 64 inches (approximately).
- The middle 50% of male heights were between 66 to 70 inches (approximately).
- The IQR for female heights is approximately $64 - 61 = 3$ inches.
- The IQR for male heights is approximately $70 - 66 = 4$ inches.

Boxplots and Distribution Shapes

