

UNIVARIATE STATISTICAL ANALYSIS

TUTORIAL SESSION 1

NUMERICAL DESCRIPTIVE MEASURES

Chapter 1

Question 1.13 (p.16)

Question 1.21 (p.18) – Going back to school

- Use this information to construct a bar chart.
- Use this information to construct a pie chart

Chapter 3

Question 3.24 (p.112-113) – age 25 to 34 with 4-year degree

Part a. only

Chapter 4

Question 4.1 (p.156) – Truck prices

Part a. only

Question 4.9 (p.158) – Speed-related crash fatalities

Question 4.17 (p.165) – Chess costs

Assuming the data is **sample**.

- Calculate the sample mean.
- Calculate the sample variance.
- Calculate the sample standard deviation.

Question 4.19 (p.165) – Cereals cost per serving

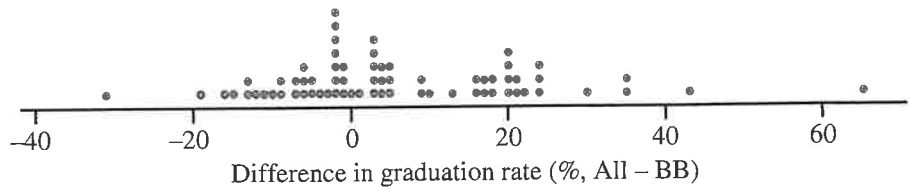
Assuming the data is the **population**.

- Calculate the population mean.
- Calculate the population variance.
- Calculate the population standard deviation.

The most interesting feature of the difference dotplot is the variability in the positive differences. Positive differences correspond to schools that have a lower graduation rate for basketball players. The positive differences range from 1% all the way up to 65%. There were five schools that had a graduation rate for all athletes that was 30 percentage points or more greater than the graduation rate for basketball players. (In case you were wondering, these schools were University of Oregon with a difference of 30%, Syracuse University and University of North Carolina Wilmington with differences of 35%, University of Cincinnati with a difference of 43%, and University of Connecticut with a difference of 65%.)

FIGURE 1.8

Dotplot of graduation rate differences (all athletes – basketball players).



EXERCISES 1.12 - 1.32

● Data set available online

- 1.12 Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.
- Number of students in a class of 35 who turn in a term paper before the due date
 - Gender of the next baby born at a particular hospital
 - Amount of fluid (in ounces) dispensed by a machine used to fill bottles with soda pop
 - Thickness of the gelatin coating of a vitamin E capsule
 - Birth order classification (only child, firstborn, middle child, lastborn) of a math major
 - The number of kernels in a bag of microwave popcorn that have not popped after 3 minutes of cooking
 - The number of students in a class of 35 who have purchased a used copy of the textbook
- 1.13 Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.
- Brand of computer purchased by a customer C
 - State of birth for someone born in the United States C
 - Price of a textbook NC
 - Concentration of a contaminant (micrograms per cubic centimeter) in a water sample NC
 - Zip code (Think carefully about this one.) C
 - Actual weight of coffee in a 1-pound can, labeled as containing 1 pound
- 1.14 For the following numerical variables, state whether each is discrete or continuous.
- The number of insufficient-funds checks received by a grocery store during a given month
 - The amount by which a 1-pound package of ground beef decreases in weight (because of moisture loss) before purchase
 - The length of a 1-year-old rattlesnake
 - The altitude of a location in California selected randomly by throwing a dart at a map of the state
 - The distance from the left edge at which a 12-inch plastic ruler snaps when bent sufficiently to break
 - The price per gallon paid by the next customer to buy gas at a particular station
- 1.15 For the following numerical variables, state whether each is discrete or continuous.
- The length of a 1-year-old rattlesnake
 - The altitude of a location in California selected randomly by throwing a dart at a map of the state
 - The distance from the left edge at which a 12-inch plastic ruler snaps when bent sufficiently to break
 - The price per gallon paid by the next customer to buy gas at a particular station
- 1.16 For each of the following situations, give a set of possible data values that might arise from making the observations described.
- The manufacturer for each of the next 10 automobiles to pass through a given intersection is noted.
 - The grade point average for each of the 15 seniors in a statistics class is determined.
 - The number of gas pumps in use at each of 20 gas stations at a particular time is determined.
 - The actual net weight of each of 12 bags of fertilizer having a labeled weight of 50 pounds is determined.
 - Fifteen different radio stations are monitored during a 1-hour period, and the amount of time devoted to commercials is determined for each.

2015	
Movie	2015 Sales (millions of dollars)
Star Wars: The Force Awakens	936.7
Jurassic World	652.3
Avengers: Age of Ultron	459.0
Inside Out	356.5
Furious 7	353.0
Minions	336.0
The Hunger Games: Mockingjay—Part 2	281.7
The Martian	228.4
Cinderella	201.2
Spectre	200.1
Mission Impossible—Rogue Nation	195.0
Pitch Perfect 2	184.3
The Revenant	183.6
Ant-Man	180.2
Home	177.4
Hotel Transylvania 2	169.7
Fifty Shades of Grey	166.2
The SpongeBob Movie: Sponge Out of Water	163.0
Straight Outta Compton	161.2
San Andreas	155.2

- Construct a dotplot of the 2014 ticket sales data. Comment on any interesting features of the dotplot. (Hint: See “What to Look For” in the Dotplots box on page 14.)
- Construct a dotplot of the 2015 ticket sales data. Comment on any interesting features of the dotplot.
- In what ways are the distributions of the 2014 and 2015 ticket sales observations similar? In what ways are they different?

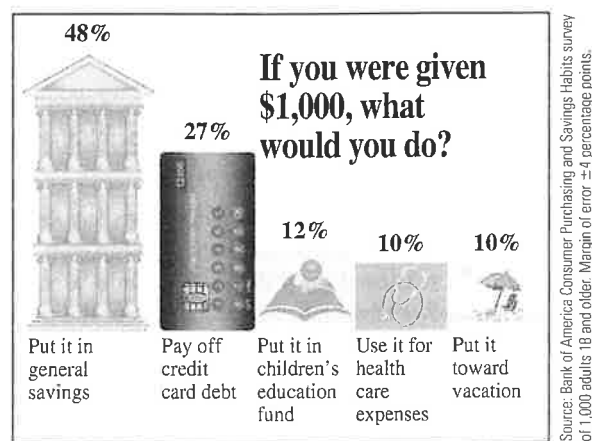
- 1.21 The report “With Their Whole Lives Ahead of Them” (publicagenda.org/files/theirwholelivesaheadofthem.pdf, retrieved February 6, 2018) includes data from a survey of 200 students who started college but did not complete a degree. Each of these students was asked, “How much have you thought about going back to school?” The accompanying frequency distribution summarizes the responses to this question.

Response	Frequency
A lot of thought	130
Some thought	48
No thought	18
Don't know	4

- Summarize the response data using a bar chart.
- Write a few sentences commenting on the distribution of the responses.

- 1.22 The following display is a graph similar to one that appeared in *USA TODAY* (June 29, 2009). This graph is meant to be a bar graph of responses to the question shown in the graph.

- Is response to the question a categorical or numerical variable?
- Explain why a bar chart rather than a dotplot was used to display the response data.
- There must have been an error made in constructing this graph. How can you tell that the graph is not a correct representation of the response data?



- 1.23 ● The accompanying table gives the total number visits and the number of unique visitors for some popular social networking sites in the United States for the month of July 2017. The number of unique visitors data are taken from the online article “Top 15 Most Popular Social Networking Sites” (ism.org/images/files/Social-media-platforms-from-Engage-to-Succeed-webinar.pdf, retrieved February 7, 2018). The total number of visits were estimated using data from semrush.com (retrieved February 7, 2018). The data on total visits and unique visitors were used to compute the values in the final column of the data table, in which

$$\text{visits per unique visitor} = \frac{\text{total visits}}{\text{number of unique visitors}}$$

Site	Unique Visitors (in millions)	Total Visits (in millions)	Visits per Unique Visitor
Facebook	1,500	20,730	13.8
YouTube	1,499	39,000	26.0
Twitter	400	13,860	34.7
Instagram	275	4,320	15.7
LinkedIn	250	1,398	5.6
reddit	125	2,445	19.6
Pinterest	105	5,970	56.9
flickr	80	84	1.1
meetup	42	45	1.1

The cumulative relative frequency plot can be used to obtain approximate answers to questions such as

What proportion of the observations is smaller than a particular value?

and

What value separates the smallest p percent from the larger values?

For example, to determine the approximate proportion of years with annual rainfall less than 9.5 inches, we would follow a vertical line up from 9.5 on the x -axis and then read across to the y -axis to obtain the corresponding relative frequency, as illustrated in Figure 3.31(a). Approximately 0.55, or 55%, of the years had annual rainfall less than 9.5 inches.

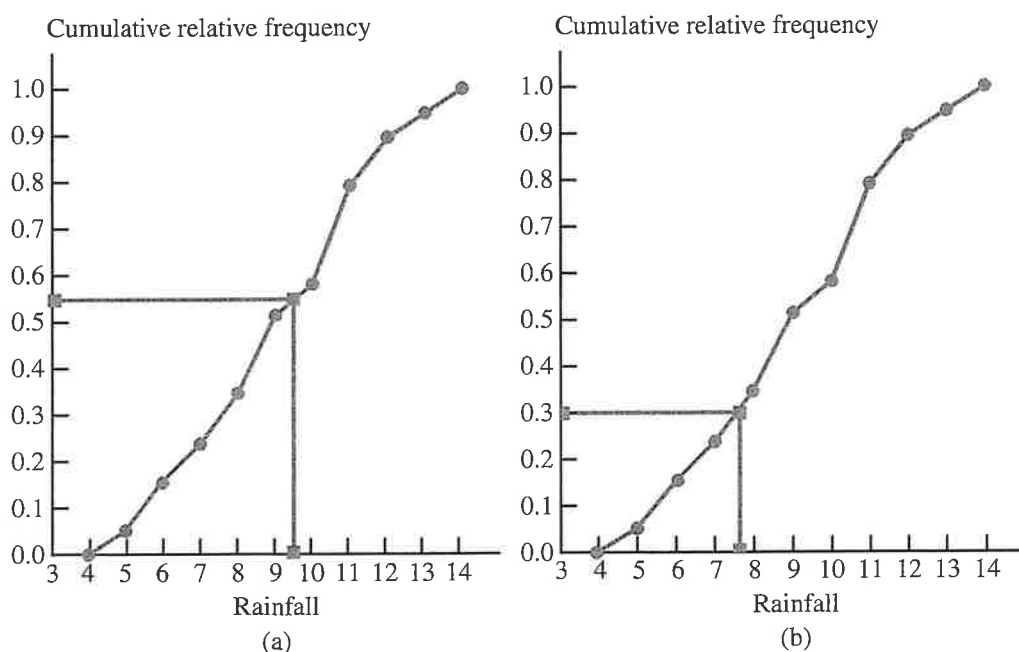


FIGURE 3.31

Using the cumulative relative frequency plot.

(a) Determining the approximate proportion of years with annual rainfall less than 9.5 inches.

(b) Finding the amount of rainfall that separates the 30% of years with the lowest rainfall from the 70% with higher rainfall.

Similarly, to find the approximate amount of rainfall that separates the 30% of years with the smallest annual rainfall from years with higher rainfall, start at 0.30 on the cumulative relative frequency axis and move across and then down to find the corresponding rainfall amount, as shown in Figure 3.31(b). Approximately 30% of the years had annual rainfall of less than 7.6 inches.

EXERCISES 3.24 - 3.39

● Data set available online

3.24 The data in the accompanying table are from the Organization for Economic Co-operation and Development (data.oecd.org/eduatt/population-with-tertiary-education.htm, retrieved February 18,

2018). Entries in the table are the percentage of 25- to 34-year-old people who have completed a 4-year college degree for 27 countries in 2016.

Country	Percentage of people age 25 to 34 with 4-year degree	Country	Percentage of people age 25 to 34 with 4-year degree
Australia	49.3	Italy	25.6
Austria	39.7	Japan	60.1
Canada	60.6	Mexico	21.8
Colombia	28.1	New Zealand	43.4
Costa Rica	28.9	Norway	48.6
Czech Republic	32.6	Poland	43.5
Denmark	45.9	Portugal	35.0
Finland	41.1	Spain	41.0
France	44.0	Sweden	47.2
Germany	30.5	Switzerland	48.8
Greece	41.0	Turkey	30.5
Hungary	30.4	United Kingdom	52.0
Iceland	43.3		
Israel	47.4	United States	47.5

- c. Based on the histograms from Parts (a) and (b), write a few sentences commenting on the similarities and differences in the distribution of SAT critical reading scores for males and females.

Score on SAT Critical Reading Exam	Relative Frequency for Males	Relative Frequency for Females
200 to < 300	0.049	0.038
300 to < 400	0.151	0.156
400 to < 500	0.308	0.332
500 to < 600	0.285	0.286
600 to < 700	0.155	0.144
700 to < 800	0.052	0.044

- 3.27 ● The data in the accompanying table represents the percentage of workers who are members of a union for each U.S. state and the District of Columbia (AARP Bulletin, September 2009).

- a. Construct a histogram of these data using the class intervals 20 to < 30, 30 to < 40, ..., 60 to < 70. (Hint: See Example 3.16.)
- b. Write a few sentences describing the shape, center, and variability of the distribution.

- 3.25 ● The accompanying data on annual maximum wind speed (in meters per second) in Hong Kong for each year in a 45-year period were given in an article that appeared in the journal *Renewable Energy* (March, 2007).

30.3 39.0 33.9 38.6 44.6 31.4 26.7 51.9 31.9
 27.2 52.9 45.8 63.3 36.0 64.0 31.4 42.2 41.1
 37.0 34.4 35.5 62.2 30.3 40.0 36.0 39.4 34.4
 28.3 39.1 55.0 35.0 28.8 25.7 62.7 32.4 31.9
 37.5 31.5 32.0 35.5 37.5 41.0 37.5 48.6 28.1

- a. Use the annual maximum wind speed data to construct a histogram.
- b. Is the histogram approximately symmetric, positively skewed, or negatively skewed?
- c. Would you describe the histogram as unimodal, bimodal, or multimodal?

- 3.26 ● The accompanying relative frequency table is based on data from the 2016 College Bound Seniors Report (collegeboard.org, retrieved February 18, 2018).

- a. Construct a relative frequency histogram for SAT critical reading score for males.
- b. Construct a relative frequency histogram for SAT critical reading score for females.

State	% of Workers who Belong to a Union	State	% of Workers who Belong to a Union
Alabama	9.8	Nebraska	8.3
Alaska	23.5	Nevada	16.7
Arizona	8.8	New Hampshire	3.5
Arkansas	5.9	New Jersey	6.1
California	18.4	New Mexico	10.6
Colorado	8.0	New York	18.3
Connecticut	16.9	North Carolina	7.2
Delaware	12.2	North Dakota	24.9
District of Columbia	13.4	Ohio	14.2
Florida	6.4	Oklahoma	6.6
Georgia	3.7	Oregon	16.6
Hawaii	24.3	Pennsylvania	15.4
Idaho	7.1	Rhode Island	16.5
Illinois	16.6	South Carolina	3.9
Indiana	12.4	South Dakota	5.0
Iowa	10.6	Tennessee	5.5
Kansas	7.0	Texas	4.5
Kentucky	8.6	Utah	5.8
Louisiana	4.6	Vermont	4.1
Maine	12.3	Virginia	10.4
Maryland	15.7	Washington	19.8
Massachusetts	12.6	West Virginia	13.8
Michigan	18.8	Wisconsin	15.0
Minnesota	16.1	Wyoming	7.7
Mississippi	5.3		
Missouri	11.2		
Montana	12.2		

Example 4.6 Can You Hear Me Now?

Understand the context ►

It is not uncommon for a cell phone user to complain about the quality of his or her service provider. Suppose that each person in a sample of $n = 15$ cell phone users is asked if he or she is satisfied with the cell phone service. Each response is classified as S (satisfied) or F (not satisfied). The resulting data are



Getty Images

Consider the data ►

S F S S S F F S S F
S S S F F

Do the work ►

This sample contains nine S's and $n = 15$, so

$$\hat{p} = \frac{9}{15} = 0.60$$

Interpret the results ►

This means that 60% of the sample responses are S's. Of those surveyed, 60% are satisfied with their cell phone service.

The letter p is used to denote the **population proportion of S's**.* We will see later how the value of $\hat{p}$ from a particular sample can be used to draw conclusions about the population proportion p .

*Note that this is one situation in which we will not use a Greek letter to denote a population characteristic. Some statistics books use the symbol π for the population proportion and p for the sample proportion. We will not use π in this context so that there is no confusion with the mathematical constant $\pi = 3.14\dots$

EXERCISES 4.1 - 4.16

● Data set available online

- 4.1 The following are the prices (in dollars) of the six all-terrain truck tires rated most highly by *Consumer Reports* in 2018 (consumerreports.org, retrieved February 22, 2018):

159.00 199.00 157.00 127.65 123.99 126.00

- Calculate the values of the mean and median.
- Why are these values so different?
- Which of the two—mean or median—appears to be better as a description of a typical value for this data set? (Hint: See Example 4.5.)

- 4.2 ● The article “Caffeine Content of Drinks” (caffeineinformer.com/the-caffeine-database, retrieved February 22, 2018) gave the following data on caffeine concentration (mg/ounce) for eight top-selling energy drinks:

Energy Drink	Caffeine Concentration (mg/oz)
Red Bull	9.5
Monster	10.0

(continued)

Energy Drink	Caffeine Concentration (mg/oz)
Rockstar	10.0
Full Throttle	10.0
No Fear	11.4
Amp	8.9
SoBe Adrenaline Rush	9.5
Tab Energy	9.1

- What is the value of the mean caffeine concentration for this set of top-selling energy drinks?
 - Coca-Cola has 2.9 mg/ounce of caffeine and Pepsi Cola has 3.2 mg/ounce of caffeine. Write a sentence explaining how the caffeine concentration of top-selling energy drinks compares to that of these colas.
- 4.3 ● Consumer Reports Health (consumerreports.org/health) reported the accompanying caffeine concentration (mg/cup) for 12 brands of coffee:

- b. When would it be reasonable to generalize from this sample to the population of all seniors at this university?

- 4.9 ● The U.S. Department of Transportation reported the number of speed-related crash fatalities for the 15 states that had the highest numbers of these fatalities in 2012 (*Traffic Safety Facts 2012 Data, Speeding*, May 2014).

State	Speed-Related Traffic Fatalities
Texas	1,247
California	916
Pennsylvania	614
North Carolina	440
Illinois	387
Florida	361
New York	360
Ohio	356
Missouri	326
South Carolina	316
Arizona	297
Alabama	272
Virginia	271
Michigan	250
Oklahoma	218

- a. Calculate the mean number of speed-related fatalities for these 15 states.
 b. Calculate the median number of speed-related fatalities for these 15 states.
 c. Explain why it is not reasonable to generalize from this sample of 15 states to the other 35 states.

- 4.10 The ministry of Health and Long-Term Care in Ontario, Canada, publishes information on its web site (health.gov.on.ca) on the time that patients must wait for various medical procedures. For two cardiac procedures completed in fall of 2005, the following information was provided:

	Number of Completed Procedures	Median Wait Time (days)	Mean Wait Time (days)	90% Completed Within (days)
Angioplasty	847	14	18	39
Bypass surgery	539	13	19	42

The median wait time for angioplasty is greater than the median wait time for bypass surgery but the mean wait time is shorter for angioplasty than for bypass surgery. What does this suggest about the distribution of wait times for these two procedures?

- 4.11 Houses in California are expensive, especially on the Central Coast where the air is clear, the ocean is blue, and the scenery is stunning. The median home price in San Luis Obispo County reached a new high in July 2004, soaring to \$452,272 from \$387,120 in March 2004 (*San Luis Obispo Tribune*, April 28, 2004).

The article included two quotes from people attempting to explain why the median price had increased. Richard Watkins, chairman of the Central Coast Regional Multiple Listing Services was quoted as saying, "There have been some fairly expensive houses selling, which pulls the median up."

Robert Kleinhenz, deputy chief economist for the California Association of Realtors explained the volatility of house prices by stating: "Fewer sales means a relatively small number of very high or very low home prices can more easily skew medians."

Are either of these statements correct? For each statement that is incorrect, explain why it is incorrect and propose a new wording that would correct any errors in the statement.

- 4.12 Consider the following statement: More than 65% of the residents of Los Angeles earn less than the average wage for that city. Could this statement be correct? If so, how? If not, why not?

- 4.13 A sample consisting of four pieces of luggage was selected from among the luggage checked at an airline counter, yielding the following data on x = weight (in pounds):

$$x_1 = 33.5, \quad x_2 = 27.3, \quad x_3 = 36.7, \quad x_4 = 30.5$$

Suppose that one more piece is selected and denote its weight by x_5 . Find a value of x_5 such that $\bar{x}$ = sample median.

- 4.14 Suppose that 10 patients with meningitis received treatment with large doses of penicillin. Three days later, temperatures were recorded, and the treatment was considered successful if there had been a reduction in a patient's temperature. Denoting success by S and failure by F, the 10 observations are

S S F S S S F F S S

- a. What is the value of the sample proportion of successes?
 b. Replace each S with a 1 and each F with a 0. Then calculate $\bar{x}$ for this numerically coded sample. How does $\bar{x}$ compare to $\hat{p}$?
 c. Suppose that it is decided to include 15 more patients in the study. How many of these would have to be S's to give $\hat{p} = 0.80$ for the entire sample of 25 patients?

The sample mean and standard deviation for this data set are 4775 and 1867, respectively. If we were to change the two largest values from 6900 and 7700 to 10,900 and 11,700 (so that they still remain the two largest values), the median and interquartile range would not be affected, but the mean and the standard deviation would change to 7500 and 2390, respectively. The value of the interquartile range is not affected by a few extreme values in the data set.

EXERCISES 4.17 - 4.33

● Data set available online

- 4.17 ● The following data are costs (in cents) per ounce for nine different brands of sliced Swiss cheese (consumerreports.org):

29 62 37 41 70 82 47 52 49

- Calculate the variance and standard deviation for this data set. (Hint: See Example 4.8.)
- If a very expensive cheese with a cost per slice of \$1.50 (150 cents) was added to the data set, how would the values of the mean and standard deviation change?

- 4.18 ● Cost per serving (in cents) for six high-fiber cereals rated very good and for nine high-fiber cereals rated good by *Consumer Reports* are shown below. Write a few sentences describing how these two data sets differ with respect to center and variability. Use summary statistics to support your statements.

Cereals Rated Very Good

46 49 62 41 19 77

Cereals Rated Good

71 30 53 53 67 43 48 28 54

- 4.19 ● Combining the cost-per-serving data for high-fiber cereals rated very good and those rated good from the previous exercise gives the following data set:

46 49 62 41 19 77 71 30
53 53 67 43 48 28 54

- Calculate the quartiles and the interquartile range for this combined data set. (Hint: See Example 4.9.)
- Calculate the interquartile range for just the cereals rated good. Is this value greater than, less than, or about equal to the interquartile range computed in Part (a)?

- 4.20 ● The paper "Caffeinated Energy Drinks—A Growing Problem" (*Drug and Alcohol Dependence* [2009]: 1–10) gave the accompanying data on caffeine per ounce for eight top-selling energy drinks and for 11 high-caffeine energy drinks:

Top-Selling Energy Drinks

9.6 10.0 10.0 9.0 10.9 8.9 9.5 9.1

High-Caffeine Energy Drinks

21.0 25.0 15.0 21.5 35.7 15.0
33.3 11.9 16.3 31.3 30.0

The mean caffeine per ounce is clearly higher for the high-caffeine energy drinks, but which of the two groups of energy drinks (top-selling or high-caffeine) is the most variable with respect to caffeine per ounce? Justify your choice based on a measure of variability.

- 4.21 The accompanying data are consistent with summary statistics that appeared in the paper "Shape of Glass and Amount of Alcohol Poured: Comparative Study of Effect of Practice and Concentration" (*British Medical Journal* [2005]: 1512–1514). Data represent the actual amount poured (in ml) into a tall, slender glass for individuals who were asked to pour 44.3 ml (1.5 ounces). Calculate and interpret the values of the mean and standard deviation.

44.0 49.6 62.3 28.4 39.1 39.8 60.5 73.0 57.5
56.5 65.0 56.2 57.7 73.5 66.4 32.7 40.4 21.4

- 4.22 The paper referenced in the previous exercise also gave data on the actual amount poured (in ml) into a short, wide glass for individuals who were asked to pour 44.3 ml (1.5 ounces).

89.2 68.6 32.7 37.4 39.6 46.8 66.1 79.2 66.3
52.1 47.3 64.4 53.7 63.2 46.4 63.0 92.4 57.8

- Calculate and interpret the values of the mean and standard deviation.
- What do the values of the mean amount poured in the short, wide glass and the mean calculated in the previous exercise suggest about the shape of glasses used?

- 4.23 The price (in dollars) of the eight smart phones that were rated highest by *Consumer Reports* in 2018 (consumerreports.org, retrieved February 23, 2018) were

730 850 830 800 700 1000 950 520