1. (Inference). A large exercise center has several thousand members from age 18 to 55 years and several thousand members age 56 and older. The manager of the center is considering offering online fitness classes. The manager is investigating whether members' opinions of taking online fitness classes differ by age. The manager selected a random sample of 170 exercise center members ages 18 to 55 years and a second random sample of 230 exercise center members ages 56 years and older. Each sampled member was asked whether they would be interested in taking online fitness classes.

The manager found that 51 of the 170 sampled members ages 18 to 55 years and that 79 of the 230 sampled members ages 56 years and older said they would be interested in taking online fitness classes.

At a significance level of $\alpha = 0.05$, do the data provide convincing statistical evidence of a difference in the proportion of all exercise center members ages 18 to 55 years who would be interested in taking online fitness classes and the proportion of all exercise center members ages 56 years and older who would be interested in taking online fitness classes? Complete the appropriate inference procedure to justify your response.

Solution:

Section 1
Let p_{younger} represent the proportion of all exercise center members from 18 to 55 years of age who would be interested in taking online fitness classes, and p_{older} represent the proportion of all exercise center members 56 years or older who would be interested in taking online fitness classes. The null hypothesis is $\text{H}_0 : p_{\text{younger}} = p_{\text{older}}$ and the alternative hypothesis is $\text{H}_a : p_{\text{younger}} \neq p_{\text{older}}$.

An appropriate inference procedure is a two-sample z-test for a difference of population proportions.

Section 2 The independent observations condition for performing the two-sample z-test for a difference in population proportions is satisfied because the data were obtained from a random sample of 170 exercise center members ages 18 to 55 years and a second random sample of 230 exercise center members ages 56 years and older.

The 10% condition must be met by both samples because sampling of exercise center members is done without replacement. There are more than $10(170) = 1,700$ adults from 18 to 55 years of age who are members of the exercise center and more than $10(230) = 2,300$ adults ages 56 years and older who are members of the exercise center.

The value of the sample proportions are $\hat{p}_{\text{younger}} = \frac{51}{170} = 0.3$ and $\hat{p}_{\text{older}} = \frac{79}{230} \approx 0.3435$.

The combined proportion is $\hat{p}_c = \frac{170(0.3)+230(0.3435)}{170+230} \approx 0.325$.

The sample size is large enough to support an assumption that the sampling distribution of $\hat{p}_{\text{younger}} - \hat{p}_{\text{older}}$ is approximately normal because $170(0.325) = 55.25$, $170(1 - 0.325) = 114.75$, $230(0.325) = 74.75$, and $230(1 - 0.325) = 155.25$ are all at least 10.

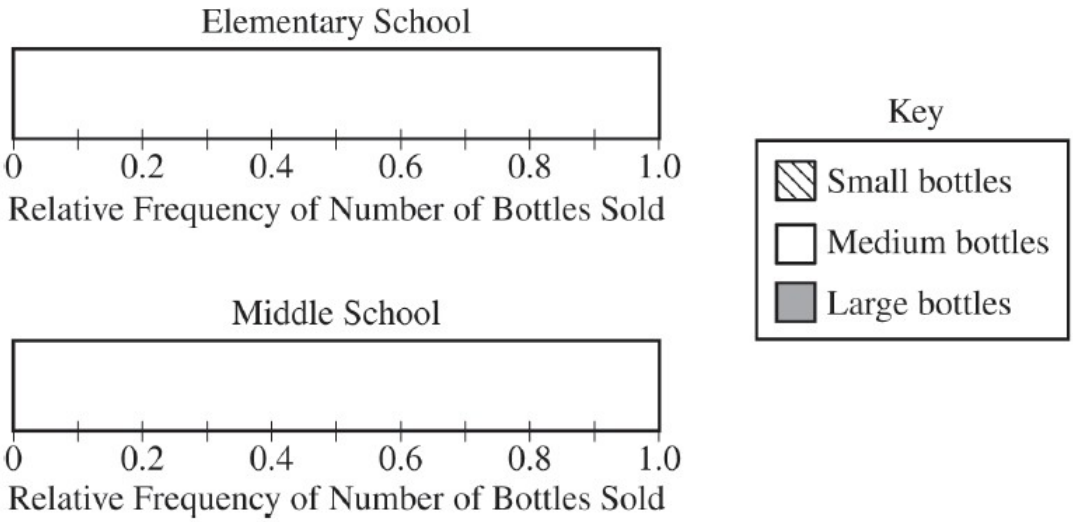
The value of the test statistic is $z = \frac{0.3-0.3435}{\sqrt{0.3250(1-0.3250)\left(\frac{1}{170}+\frac{1}{230}\right)}} \approx -0.918$.

The corresponding p-value is $2 * P(z < -0.918) = 0.359$.

Section 3 Because the p-value of approximately 0.359 is greater than $\alpha = 0.05$, the null hypothesis should not be rejected. The results from this study do not provide convincing statistical evidence that the population proportion of exercise center members from 18 to 55 years of age who would be interested in taking online fitness classes is different from the population proportion of adults ages 56 years and older who would be interested in taking online fitness classes.

2. (Exploring Data). A local elementary school decided to sell bottles printed with the school district's logo as a fund-raiser. The students in the elementary school were asked to sell bottles in three different sizes (small, medium, and large). The relative frequencies of the number of bottles sold for each size by the elementary school were 0.5 for small bottles, 0.3 for medium bottles, and 0.2 for large bottles. A local middle school also decided to sell bottles as a fund-raiser, using the same three sizes (small, medium, and large). The middle school students sold three times the number of bottles that the elementary school students sold. For the middle school students, the proportion of bottles sold was equal for all three sizes.

(A) Complete the segmented bar graphs representing the relative frequencies of the number of bottles sold for each size by students at each school.

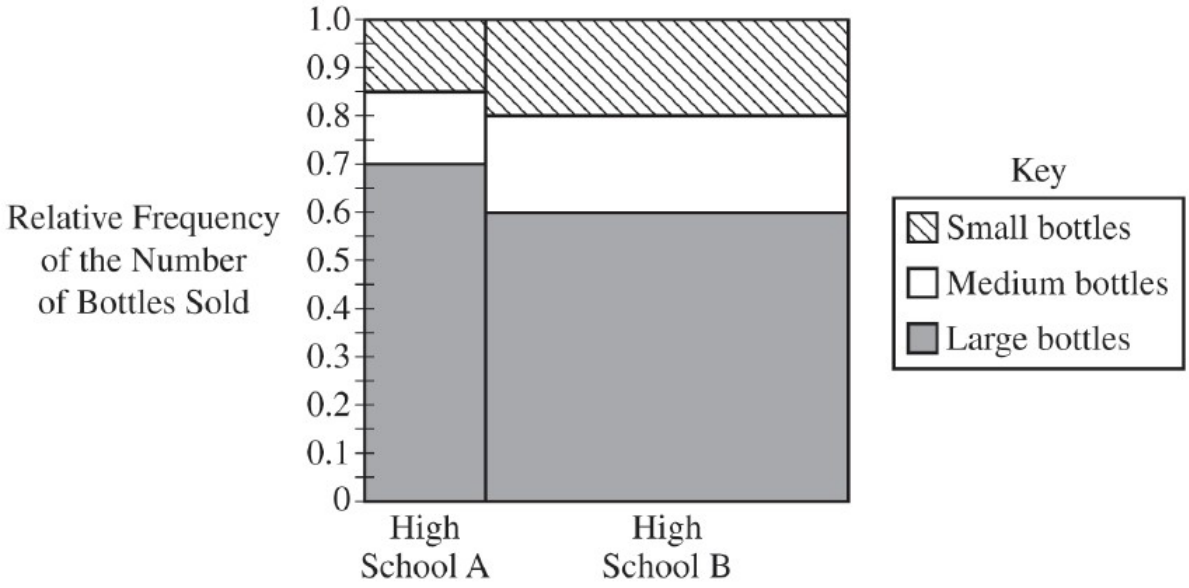


(B) An administrator at the elementary school concluded that the elementary school students sold more small bottles than the middle school students did. Is the elementary school administrator's conclusion correct? Explain your response.

Two high schools are also selling the bottles and are competing to see which one sold more large bottles.

(C) A mosaic plot for the distribution of the number of bottles sold by each of the high schools is shown here.

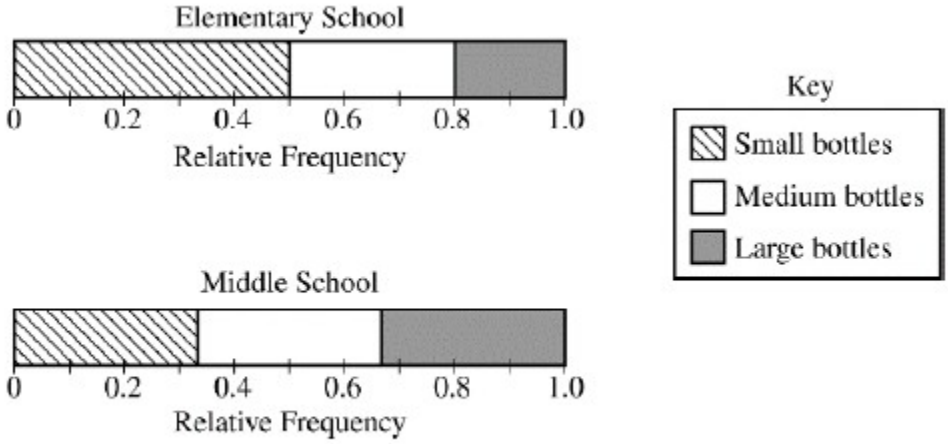
Distribution of the Number of Bottles Sold by High School



(i) Which of the two high schools sold a greater proportion of large bottles? Justify your answer.

(ii) Which of the two high schools sold a greater number of large bottles? Justify your answer.

Solution:
(A)



(B) No, the segment for small bottles for the elementary school is wider than the segment for small bottles for the middle school; however, the middle school students sold three times as many bottles as the elementary school students. So, if the elementary school sold x number of bottles, the middle

school sold 3x number of bottles.

For example, if the elementary students sold 100 bottles total, then they sold 0.5(100) or 50 small bottles. However, because the middle school students sold three times the total number of bottles as the elementary students, they would have sold 300 bottles total and $(0.\overline{3})(300) = 100$ small bottles. Because $100 > 50$, the middle school sold more small bottles, and the elementary school's administrator is not correct.

(C)

(i) The mosaic plot shows that the proportion of large bottles sold by High School A was 0.7 and the proportion of large bottles sold by High School B was 0.6. High School A sold a greater proportion of large bottles because $0.7 > 0.6$.

(ii) The number of bottles sold is represented by the area of the shaded region. The area of the rectangle representing large bottles sold by High School B is clearly larger than the area of the rectangle representing large bottles sold by High School A. Therefore, High School B sold more large bottles than High School A.

3.(Sampling and Experimental Design). A car maker produces four different models of cars: A, B, C, and D. A group of researchers is investigating which model of car has the longest distance traveled per gallon of gas (mileage). Higher mileage is considered better than lower mileage. The researchers will conduct a study in which they contact several owners of each model of car and ask them to estimate their mileage.

(A) Is this an observational study or an experiment? Justify your answer in context.

Model D has an autopilot feature, in which the car controls its own motion with human supervision. James owns a Model D car and will investigate whether using the autopilot feature results in higher mileage than not using the autopilot. James will drive his car on 70 different days to and from work, using the same route at the same time each day. James will record the mileage each day.

(B) James will use a completely randomized design to conduct his investigation. Describe an appropriate method James could use to randomly assign the two treatments, driving using the autopilot feature and driving with out using the autopilot feature, to 35 days each.

(C) After the investigation was completed, James verified that the conditions for inference were met and conducted a hypothesis test. He discovered the mean mileage when using the autopilot feature was significantly higher than the mean mileage when not using the autopilot feature.

James is a member of a Model D club with thousands of members who all drive Model D cars. He will give a presentation at a Model D club members' meeting later this year and would like to state that the results of his hypothesis test apply to all Model D cars in his club. Another member of the club who is a statistician tells James his findings do not apply to all Model D cars in the club. What change would James need to make to his original study to be able to generalize to all Model D cars in the club?

Solution:
(A.) This is an observational study. The researchers had the car owners estimate their mileage. The car owners were not randomly assigned a car model, so no treatment was imposed.

(B.) Number the days in the experiment from 1 to 70. Using a random number generator, generate 35 unique integers from 1 to 70, inclusive. Assign the days with those 35 unique integers for James to drive the car with autopilot and assign the remaining 35 days for James to drive the car without autopilot.

(Alternative solution) Using 70 equally sized slips of paper, label 35 "with autopilot" and 35 "without autopilot." Mix the slips of paper in a bag. Each day for the 70 days, select a slip of paper (without replacement) to determine the driving method for that day.

(C) In order to generalize his findings to all Model D cars in his club, he would need to randomly select Model D cars from the club. He would then need to carry out a similar study using the Model D cars that were randomly sampled from the club.

4. (Probability and Sampling Distributions). In an online game, players move through a virtual world collecting geodes, a type of hollow rock. When broken open, these geodes contain crystals of different colors that are useful in the game. A red crystal is the most useful crystal in the game. The color of the crystal in each geode is independent and the probability that a geode contains a red crystal is 0.08.

(A) Sarah, a player, will collect and open geodes until a red crystal is found.

i. Calculate the mean of the distribution of the number of geodes Sarah will open until a red crystal is found. Show your work.

ii. Calculate the standard deviation of the distribution of the number of geodes Sarah will open until a red crystal is found. Show your work.

(B) Another player, Conrad, decides to play the game and will stop opening geodes after finding a red crystal or when 4 geodes have been opened, whichever comes first. Let $Y =$ the number of geodes Conrad will open. The table shows the partially completed probability distribution for the random variable Y.

Number of geodes Conrad will open, y	1	2	3	4
Probability, $P(Y = y)$	0.08	0.0736		

i. Calculate $P(Y = 3)$. Show your work.

ii. Calculate $P(Y = 4)$. Show your work.

(C) Consider the table and your results from part (b).

i. Calculate the mean of the distribution of the number of geodes Conrad will open. Show your work.

ii. Interpret the mean of the distribution of the number of geodes Conrad will open, which was calculated in part (c-i).

Solution:

(A) Let G represent the number of geodes a player opens until finding a red crystal. G follows a geometric distribution with $p = 0.08$.

(i) $\mu = E(G) = \dfrac{1}{0.08} \approx 12.5$ geodes

(ii) $\sigma_G = \dfrac{\sqrt{1 - 0.08}}{0.08} \approx 11.99$ geodes

(B) (i) $P(Y = 3) = (0.92)^2(0.08) \approx 0.067712$

(ii) $P(Y = 4) = 1 - P(Y = 1 \text{ or } 2 \text{ or } 3)$
$\approx 1 - (0.08 + 0.0736 + 0.067712)$
≈ 0.778688

OR

If Conrad opens 4 geodes, then he either finds no red geodes or he finds a red geode on the fourth one he opens; therefore, $P(Y = 4) = (0.92)^4 + (0.92)^3(0.08)$
≈ 0.778688.

Additional Notes:

- A response may satisfy component 2 or component 4 by the following or a combination of the following:
  - **Probability formula:** Displaying a correct formula for computing the geometric probability, such as:
    - $(1 - 0.08)^2(0.08)$ or $(0.92)^2(0.08)$ for part (b-i)
  - **Calculator function syntax:** Using calculator function notation with the correct value of the parameter identified, such as:
    - geompdf$(p = 0.08, x = 3)$ for part (b-i)
    - $1 -$ geomcdf$(p = 0.08, x = 3)$ for part (b-ii)
    - binompdf$(n = 4, p = 0.08, x = 0)$ + geompdf$(p = 0.08, x = 4)$ for part (b-ii)
    - binompdf$(n = 4, p = 0.92, x = 4)$ + geompdf$(p = 0.08, x = 4)$ for part (b-ii)
- An arithmetic or transcription error in a response can be ignored if correct work is shown.

(C) (i) $\mu = E(Y)$
$\approx (1)(0.08) + (2)(0.0736) + (3)(0.0677) + (4)(0.778688)$
$\approx 0.08 + 0.1472 + 0.2031 + 3.1148$
≈ 3.545 geodes.

(ii) The mean of 3.545 geodes is the average number of geodes that result from a long run of many, many trials of opening randomly selected geodes and counting the number opened until either a red geode is found or the fourth geode is opened.

5. (Multi-Focus). Baseball cards are trading cards that feature data on a player's performance in baseball games. Michelle is at a national baseball card collector's convention with approximately 20,000 attendees. She notices that some collectors have both regular cards, which are easily obtained, and rare cards, which are harder to obtain. Michelle believes that there is a relationship between the number of months a collector has been collecting baseball cards and whether the majority of the cards in their collection are regular or rare. She obtains information from a random sample of 500 baseball card collectors at the convention and records how many full months they have been collecting baseball cards and whether the majority of the cards in their card collection are regular or rare. Her results are displayed in a two-way table.

Majority Type of Baseball Cards and Months of Collecting Baseball Cards

	Fewer Than 6 Months	6 - 10 Months	11 - 15 Months	16 - 20 Months	21 or More Months	Total
Has a Majority of Regular Baseball Cards	80	84	71	76	112	423
Has a Majority of Rare Baseball Cards	11	16	9	6	35	77
Total	91	100	80	82	147	500

(A) If one collector from the sample is selected at random, what is the probability that the collector has been collecting baseball cards for 11 or more months and has a majority of regular baseball cards? Show your work.

(B) Given that a randomly selected collector from the sample has been collecting baseball cards for fewer than 6 months, what is the probability the collector has a majority of regular baseball cards? Show your work.

(**C**) Michelle believes there is a relationship between the number of months spent collecting baseball cards and which type of card is the majority in the collection (regular or rare).

i. Name the hypothesis test Michelle should use to investigate her belief. Do not perform the hypothesis test.

ii. State the appropriate null and alternative hypotheses for the hypothesis test you identified in (c-i). Do not perform the hypothesis test.

(D) After completing the hypothesis test described in part (c), Michelle obtains a p-value of 0.0075. Assuming the conditions for inference are met, what conclusion should Michelle make about her belief? Justify your response.

Solution:

(A) $P(11+ \text{ months} \cap \text{majority regular cards})$
$= \dfrac{71 + 76 + 112}{500}$
$= \dfrac{259}{500}$
$= 0.518$

(B) $P(\text{majority regular cards} \mid \text{fewer than 6 months})$
$= \dfrac{P(\text{majority regular cards} \cap \text{fewer than 6 months})}{P(\text{fewer than 6 months})}$
$= \dfrac{\frac{80}{500}}{\frac{91}{500}}$
$= \dfrac{80}{91}$
≈ 0.879

(C) (i) Michelle should conduct a chi-square test for independence between months collecting baseball cards and majority card status for all baseball card collectors at the convention.

(ii) H_0: There is not an association between months collecting baseball cards and majority card status for all baseball card collectors at the convention.
H_a: There is an association between months collecting baseball cards and majority card status for all baseball card collectors at the convention.
OR
H_0: Months collecting cards and majority card status are independent for all baseball card collectors at the convention.
H_a: Months collecting cards and majority card status are not independent for all baseball card collectors at the convention.

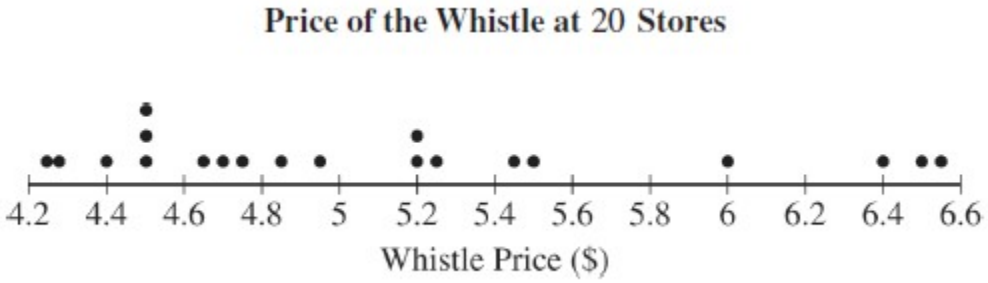
(D)

Because the p-value of 0.0075 is less than any reasonable α level such as 0.05 or 0.10, the null hypothesis should be rejected. The data provide convincing statistical evidence that there is an association between months collecting baseball cards and majority card status for all baseball card collectors at the convention.

6 (Investigative Task). A company sells a certain type of whistle. The price of the whistle varies from store to store. Julio, a statistician at the company, wants to estimate the mean price, in dollars ($), of this type of whistle at all stores that sell the whistle.

A
i. Identify the appropriate inference procedure for Julio to use.
ii. Describe the parameter for the inference procedure you identified in part (a-i) in context.

Julio called the managers of 20 randomly selected stores that sell the whistle and recorded the price of the whistle at each store. Following is a dotplot of Julio's data.



The summary statistics for Julio's data are shown in the following table.

Summary Statistics for Julio's Data

Sample Size	Mean	Standard Deviation	Minimum	Q_1	Median	Q_3	Maximum		
20	5.12	0.743	4.25	4.51	4.885	5.475	6.58		

(B) Julio wants to examine some characteristics of the distribution of the sample of whistle prices.

i. Describe the shape of the distribution of the sample of whistle prices. Justify your response using appropriate values from the summary statistics table.

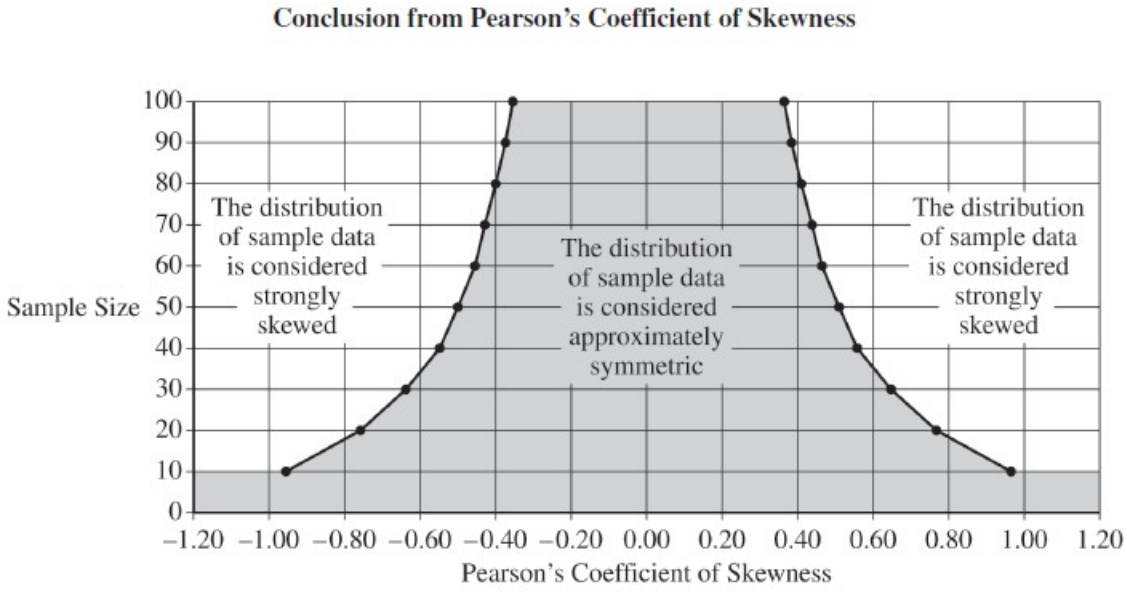
ii. Using the 1.5 × IQR rule, determine whether there are any outliers in the sample of whistle prices. Justify your response.

It can often be difficult to determine whether the distribution of sample data is skewed by looking at a graph of the data and the summary statistics, particularly when the sample size is small. Thus, statisticians sometimes measure how skewed a data set is. One such measure is Pearson's coefficient of skewness, which is calculated using the following formula.

$$\text{Pearson's Coefficient of Skewness} = \frac{3(\bar{x} - m)}{s}$$

In the formula, $\bar{x}$ is the sample mean, m is the sample median, and s is the sample standard deviation.
(C) (**i**) Calculate Pearson's coefficient of skewness for Julio's sample of 20 whistle prices. Show your work.

The following graph shows conclusions that can be made about the shape of the distribution of sample data based on Pearson's coefficient of skewness and sample size.



(ii) Indicate the value of the Pearson's coefficient of skewness you calculated in part (c-i) for the appropriate sample size by marking it with an "X" on the preceding graph.

(D) Consider your work in part (c).
(i) What should you conclude about the shape of the distribution of the sample of whistle prices? Justify your response.
Julio's inference procedure in part (a-i) needs one of the following requirements to be satisfied to verify the normality condition.

- The sample size is greater than or equal to 30.
- If the sample size is less than 30, the distribution of the sample data is not strongly skewed and does not have outliers.

(ii) Using your response to (d-i) and the preceding requirements, is the normality condition satisfied for Julio's data? Explain your response.

Solution

(A) (i) The inference procedure that should be used to estimate the mean price, in dollars ($), of this type of whistle at all stores that sell the whistle is a one-sample t-interval for a population mean.

(ii) The parameter of interest is the mean whistle price, in dollars ($), of this type of whistle at all stores that sell the whistle.

(B) (i) The distribution of the sample of whistle prices appears slightly skewed to the right, because the mean is slightly higher than the median.

(ii) Based on the $1.5 \times$ IQR rule, there are not whistle prices in this sample that would be considered outliers. A whistle price is an outlier using this method if it is more than $1.5 \times$ IQR below the first quartile (Q_1) or more than $1.5 \times$ IQR above the third quartile (Q_3). Because

$$\begin{aligned} Q_1 - 1.5 \times \text{IQR} &= 4.51 - 1.5(5.475 - 4.51) \\ &= 3.0625, \end{aligned}$$

and the minimum value (4.25) is greater than 3.0625, there are no outliers to the left. Because

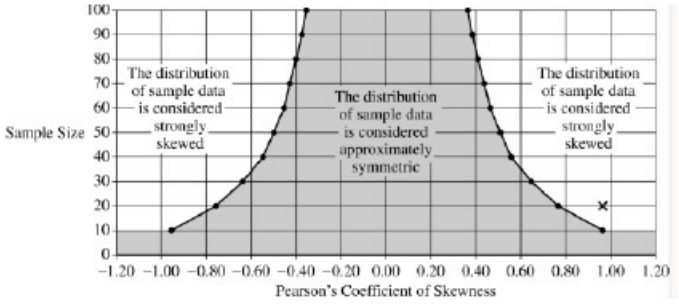
$$\begin{aligned} Q_3 + 1.5 \times \text{IQR} &= 5.475 + 1.5(5.475 - 4.51) \\ &= 6.9225, \end{aligned}$$

and the maximum value (6.58) is less than 6.9225, there are no outliers to the right.

(C) (i) Pearson's coefficient of skewness is

$$\frac{3(5.12 - 4.885)}{0.743} \approx 0.949.$$

(ii)



(D) (i) Looking at the graph in part (c), for a sample size of 20, and a skewness coefficient of 0.949, this point falls in "the distribution of sample data is considered strongly skewed" region. Therefore, we would consider the shape of the distribution of the sample of whistle prices to be strongly skewed.

(ii) No, based on the response to part (d-i). Julio's data would not satisfy the normality condition because neither of the criteria listed are met. Julio only has a sample size of 20, which is less than 30, and Pearson's coefficient of skewness indicates the distribution of sample data is strongly skewed.

Problems adapted from the College Board released tests.