MODULE 2 EXAMPLES

TAs: Nihal Reddy
Email: nireddy@ucsd.edu
OH: Thursdays 6-7pm

Slide Credits: Kira Fleischer

PROBLEM #: KEY TOPICS FROM PROBLEM

Problem setup and description.

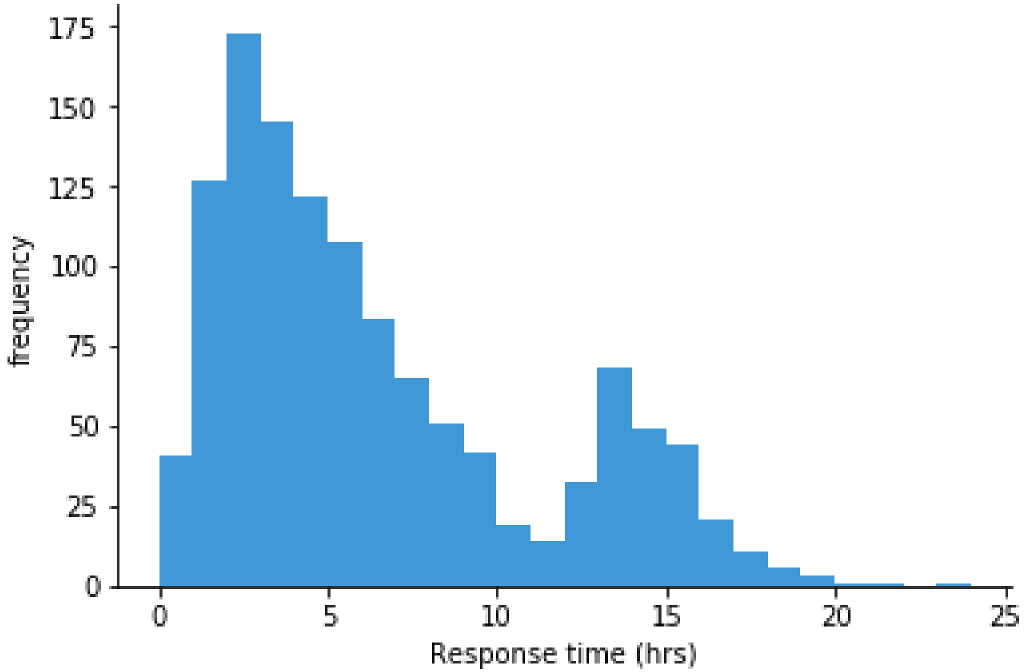
Question

Key notes from readings/lectures needed to answer the question

Solution: written with as much detail as we expect you to give on your homework sets

PROBLEM 1: HISTOGRAM

A phone company is interested in the response time of its employees and plotted a histogram as shown.



(a) Describe whether this distribution is skewed to the left, skewed to the right, or symmetric. Also discuss its possible modes with an explanation.

Skewed left: tail extends towards left

Skewed right: tail extends towards right

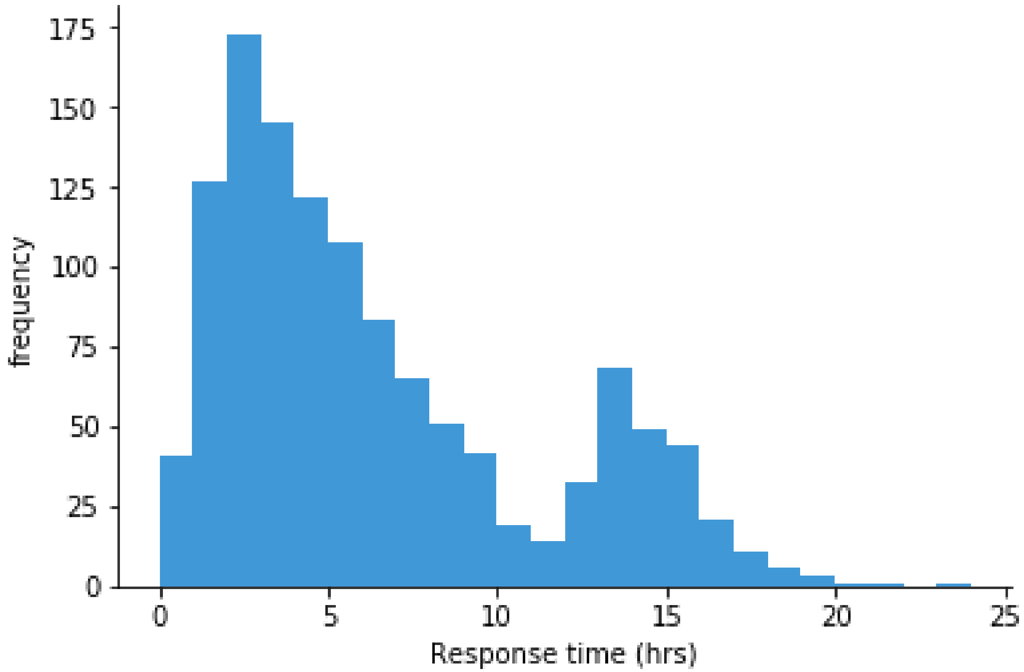
Symmetric: approximately no tails

Mode: peak in histogram

Solution: This histogram is skewed to the right, with one mode approximately in between 2-3 hours and another between 13-14 hours.

PROBLEM 1: HISTOGRAM

A phone company is interested in the response time of its employees and plotted a histogram as shown.



(b) Do you expect the mean or the median to be larger? Explain.

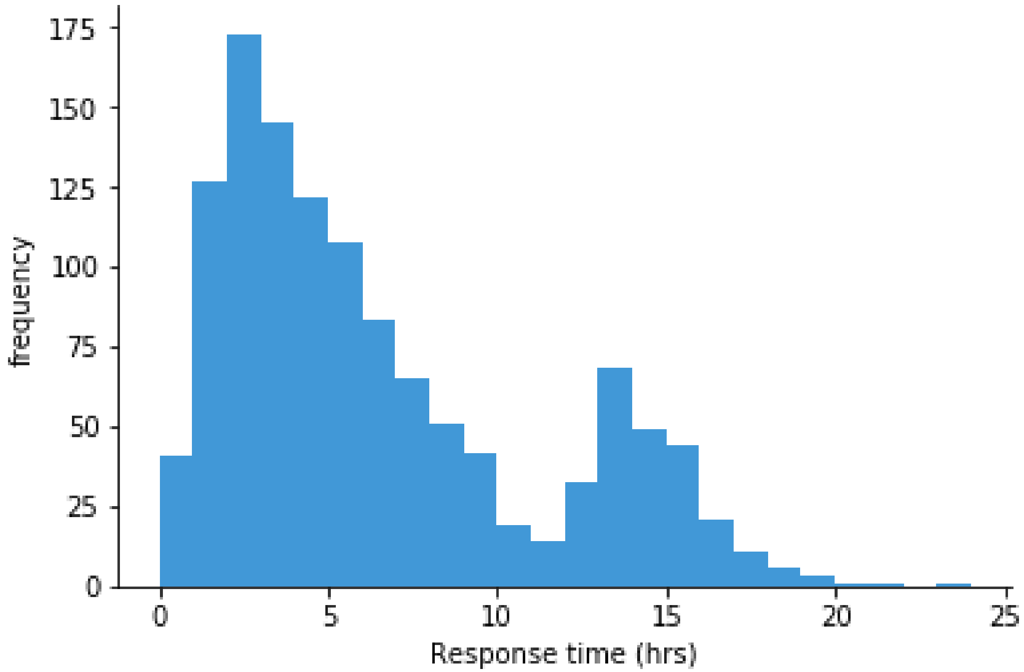
Skewed right: mean > median

Skewed left: mean < median

Solution: The mean is expected to be larger since this histogram is right skewed - the mean would be pulled more towards the upper tail.

PROBLEM 1: HISTOGRAM

A phone company is interested in the response time of its employees and plotted a histogram as shown.



(c) If you were to report either the mean or the median of the distribution, which one would you report, and why?

If there is any skew, report the median since this statistic is resistant to skew.

Solution: The median is resistant to the skewed shape of the distribution, so it is a better choice.

PROBLEM 2: WEIGHTED AVERAGES

To select a team of basketball players, the coach would like to start by measuring the heights of 40 students who are trying out for the team. One student was sick on that day and the coach first measured those that were present. He found that among the 39 students, the average height is 190cm with a standard deviation of 7 cm. Later on, he measured the absent student's height which turned out to be 193 cm.

(a) Will the new student's height increase or decrease the average height?

New height less than mean: decrease average height

New height greater than mean: increase average height

Solution: Increase: the new height is larger than the mean of the 39 other students' heights.

PROBLEM 2: WEIGHTED AVERAGES

To select a team of basketball players, the coach would like to start by measuring the heights of 40 students who are trying out for the team. One student was sick on that day and the coach first measured those that were present. He found that among the 39 students, the average height is 190cm with a standard deviation of 7 cm. Later on, he measured the absent student's height which turned out to be 193 cm.

(b) What is the new average?

$$\text{Weighted Average} = \frac{(N_1 \times \mu_1) + (N_2 \times \mu_2)}{N_1 + N_2}$$

Solution: Calculate a weighted mean. Use a weight of 39 for the old mean and 1 for the new mean:

$(39 \times 190 + 1 \times 193)/40 = 190.075$

PROBLEM 2: WEIGHTED AVERAGES

To select a team of basketball players, the coach would like to start by measuring the heights of 40 students who are trying out for the team. One student was sick on that day and the coach first measured those that were present. He found that among the 39 students, the average height is 190cm with a standard deviation of 7 cm. Later on, he measured the absent student's height which turned out to be 193 cm.

(c) Does the new student's height increase or decrease the standard deviation of the heights?

New height less than 1 std. dev. away from old mean: decrease std. dev.

New height greater than 1 std. dev. away from old mean: increase std. dev.

Solution: Since the new height is 3 cm away from the previous mean, and the standard deviation is 7 cm, the new height is less than 1 standard deviation away from the previous mean, so it decreases the standard deviation.

PROBLEM 3: MEAN, MEDIAN, QUARTILES, RANGE, IQR

A class has 20 students who receive the following grades on an exam with 200 points:

144 154 94 62 190 13 83 115 67 91 123 84 19 30 144 90 151 95 101 172

Please find these statistics:

(a) Mean

(b) Median and quartiles

(c) range and IQR

Mean: $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$

Median: 50th percentile. Order data smallest to largest, then take middle value if *n* is odd, or average the two middle numbers if *n* is even.

Quartiles:

- Q1: 25th percentile. Median of the first 50% of data
- Q3: 75th percentile. Median of the second 50% of data

Range = max – min

IQR = Q3 – Q1

Solution: (be sure to include formulas in your answers) (a) 101.1

(b) Q1: 75; Median: 94.5; Q3: 144.0

(c) Range: 177; IQR: 69

PROBLEM 4: STATISTICS WITHOUT RAW DATA

The following are the statistics for the payroll of a company:

lowest salary = $350; mean = $600; median = $700; range = $1000; IQR = $600; first quartile = $400; standard deviation = $300.

(a) Is this distribution skewed to the left, right, or symmetric?

Mean < median: skewed left

Mean > median: skewed right

Solution: Since the mean is smaller than the median, this would be skewed to the left.

PROBLEM 4: STATISTICS WITHOUT RAW DATA

The following are the statistics for the payroll of a company:

lowest salary = \$350; mean = \$600; median = \$700; range = \$1000; IQR = \$600; first quartile = \$400; standard deviation = \$300.

(b) Between what two values are the middle 50% found?

Middle 50% = 25% → 75% ⟺ Q1 → Q3

Solution: IQR = Q3 – Q1

$\Longrightarrow$ 600 = Q3 – 400

$\Longrightarrow$ Q3 = 1000

$\Longrightarrow$ Middle 50% of data between \$400 and \$1000

PROBLEM 4: STATISTICS WITHOUT RAW DATA

The following are the statistics for the payroll of a company:

lowest salary = \$350; mean = \$600; median = \$700; range = \$1000; IQR = \$600; first quartile = \$400; standard deviation = \$300.

(c) Suppose the company gives every employee a \$40 raise. Write down the new values of each statistics.

Addition shifts the distribution to the right (changes the center), but does not change the spread.

Solution: lowest salary = \$390; mean = \$640; median = \$740; range = \$1000; IQR = \$600; first quartile = \$440; standard deviation = \$300.

PROBLEM 4: STATISTICS WITHOUT RAW DATA

The following are the statistics for the payroll of a company:

lowest salary = \$350; mean = \$600; median = \$700; range = \$1000; IQR = \$600; first quartile = \$400; standard deviation = \$300.

(d) Now suppose that the company gives every employee a 20% raise. Write down the new values of each statistics.

Multiplication changes the center *and* changes the spread.

Solution: lowest salary = \$420; mean = \$720; median = \$840; range = \$1200; IQR = \$720; first quartile = \$480; standard deviation = \$360.

PROBLEM 5: CATEGORICAL VARIABLES

Suppose you consider a data set that has the variable eye color, where eye color is coded as follows, and suppose this data set yields the following statistics:

Blue = 0

Green = 1

Brown = 2

Hazel = 3

Count	20
Mean	1.45
Median	2
Std. Dev.	1.32
Q1	0
Q3	3

Discuss what these statistics can tell

Solution: Since eye colors are categorical, not quantitative, the statistics would not be useful nor appropriate.

PROBLEM 6: DETERMINING DISTRIBUTIONS FROM DATA

The following summary statistics give some information on the percentage of students that

Count	48
Mean	68.35
Median	69.90
Std. Dev.	10.20
Min	43.20
Max	87.40
Range	44.20
Q1	59.15
Q3	74.75

(a) Do you think this distribution is skewed or symmetric? Explain.

Mean < median: skewed left

Mean > median: skewed right

Mean ≈ median: symmetric

Solution: This would be slightly left skewed since the mean is slightly less than the median while the 25th percentile is farther away from the median than the 75th percentile.

PROBLEM 6: DETERMINING DISTRIBUTIONS FROM DATA

The following summary statistics give some information on the percentage of students that

Count	48
Mean	68.35
Median	69.90
Std. Dev.	10.20
Min	43.20
Max	87.40
Range	44.20
Q1	59.15
Q3	74.75

(b) Are there any outliers?

A data point $\boldsymbol{x}$ is an outlier if

- $\boldsymbol{x} < \text{Q1} - 1.5 \times \text{IQR}$
- $\boldsymbol{x} > \text{Q3} + 1.5 \times \text{IQR}$

Solution: $\text{Q1} - 1.5 \times \text{IQR} = 59.15 - 1.5 \times (74.75 - 59.15) = 35.75$;

$\text{Q3} + 1.5 \times \text{IQR} = 74.75 + 1.5 \times (74.75 - 59.15) = 98.15$;

There are no outliers since all data are within the fence.

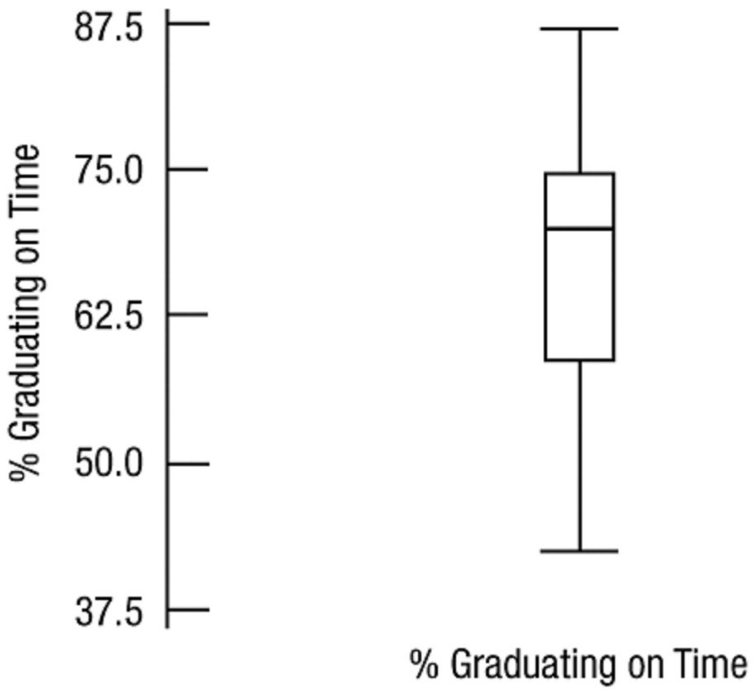
PROBLEM 6: DETERMINING DISTRIBUTIONS FROM DATA

The following summary statistics give some information on the percentage of students that

Count	48
Mean	68.35
Median	69.90
Std. Dev.	10.20
Min	43.20
Max	87.40
Range	44.20
Q1	59.15
Q3	74.75

(c) Draw a boxplot for this.

Solution:



PROBLEM 7: 5-NUMBER SUMMARIES

A class of students take a quiz and the 5-number summaries are given for the 18 freshmen and 15 sophomores.

Summary	Min	Q1	Median	Q3	Max
Freshmen	3	4.5	6.5	8.5	9.5
Sophomores	4	6	7.5	9	10

(a) Do the sophomores or freshmen have the highest score?

Look at max.

Solution: Sophomores: 10 > 9.5.

PROBLEM 7: 5-NUMBER SUMMARIES

A class of students take a quiz and the 5-number summaries are given for the 18 freshmen and 15 sophomores.

Summary	Min	Q1	Median	Q3	Max
Freshmen	3	4.5	6.5	8.5	9.5
Sophomores	4	6	7.5	9	10

(b) Do the sophomores or freshmen have a greater range?

Range = max – min.

Solution: Freshmen: 9.5 – 3 = 6.5 > 10 – 4 = 6.

PROBLEM 7: 5-NUMBER SUMMARIES

A class of students take a quiz and the 5-number summaries are given for the 18 freshmen and 15 sophomores.

Summary	Min	Q1	Median	Q3	Max
Freshmen	3	4.5	6.5	8.5	9.5
Sophomores	4	6	7.5	9	10

(c) Do the sophomores or freshmen have a greater IQR?

IQR = Q3 – Q1.

Solution: Freshmen: 8.5 – 4.5 = 4 > 9 – 6 = 3.

PROBLEM 7: 5-NUMBER SUMMARIES

A class of students take a quiz and the 5-number summaries are given for the 18 freshmen and 15 sophomores.

Summary	Min	Q1	Median	Q3	Max
Freshmen	3	4.5	6.5	8.5	9.5
Sophomores	4	6	7.5	9	10

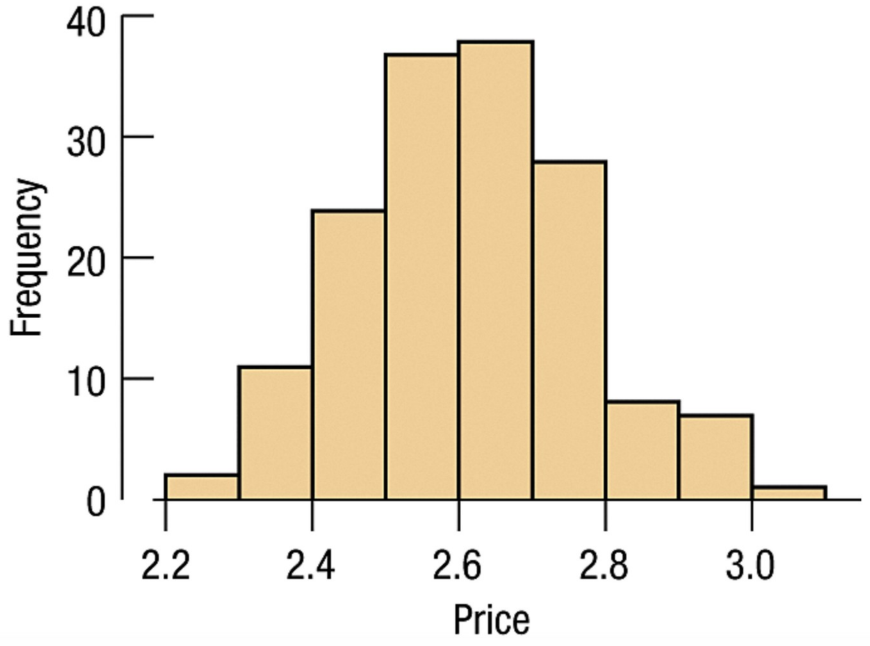
d) If the mean of the scores of the freshmen is 6.5 and that of the sophomores is 7, then what is the overall mean for the whole class?

Weighted Average $= \dfrac{(N_1 \times \mu_1) + (N_2 \times \mu_2)}{N_1 + N_2}$

Solution: (6.5 × 18 + 7 × 15) / 33 = 6.73.

PROBLEM 8: DETERMINING STATISTICS FROM HISTOGRAMS

The following histogram shows the price of candies in a local store in La Jolla

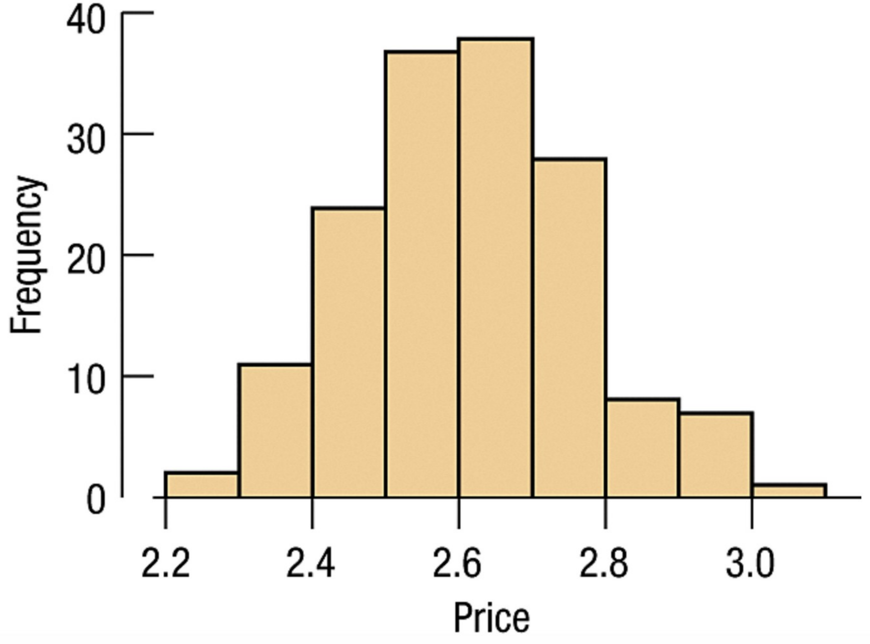


(a) Would it be appropriate to use the mean and standard deviation to summarize the distribution, and why?

Solution: Yes, because the distribution is symmetric and unimodal.

PROBLEM 8: DETERMINING STATISTICS FROM HISTOGRAMS

The following histogram shows the price of candies in a local store in La Jolla

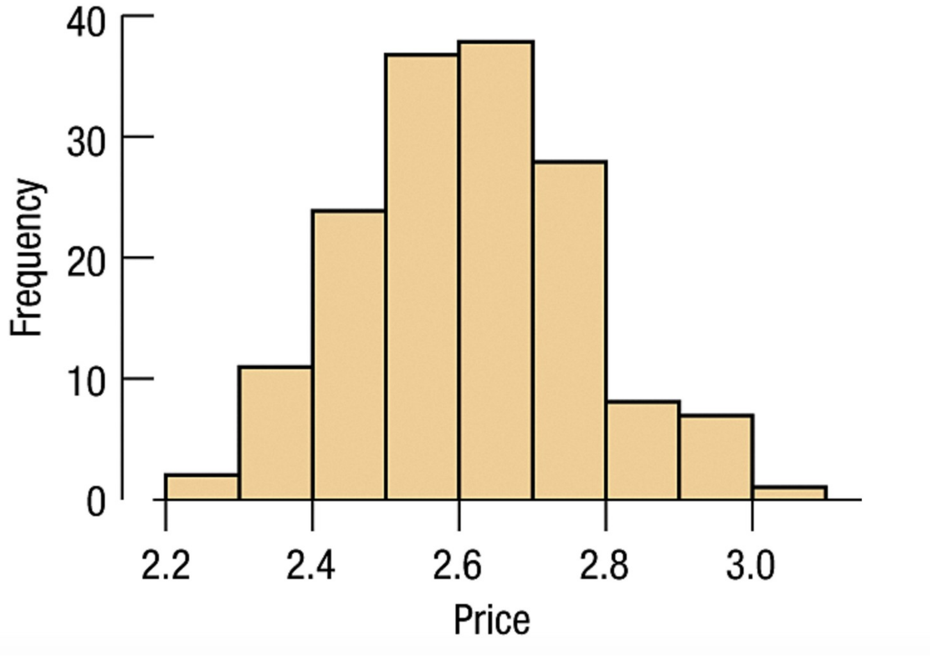


(b) Which is the mean closest to: $2.4, $2.6, or $2.8?

Solution: $2.6 because that's the balancing point of the histogram.

PROBLEM 8: DETERMINING STATISTICS FROM HISTOGRAMS

The following histogram shows the price of candies in a local store in La Jolla



(c) Which do you think is a better estimate for the standard deviation: $0.15, $0.5, or $1? Explain.

Solution: $0.15 because that's a typical distance from the mean. There aren't a lot of data points above $0.5 or $1 away from the mean.

PROBLEM 9: MEASURES OF CENTER AND SPREAD

Bob, Alice, and Charlie are three employees in a small company. Bob's salary is the lowest among the three. **If Bob now becomes a part-time worker and hence had a salary reduction, how would the following statistics of the salaries of these three employees be affected?**

(a) measures of center: Mean, Median

Changing the minimum of the data will not change the median, it will only change the mean.

Solution: The median will not be affected much while the mean would decrease.

PROBLEM 9: MEASURES OF CENTER AND SPREAD

Bob, Alice, and Charlie are three employees in a small company. Bob’s salary is the lowest among the three. **If Bob now becomes a part-time worker and hence had a salary reduction, how would the following statistics of the salaries of these three employees be affected?**

(b) measures of spread: standard deviation, range, and IQR. Please provide some arguments to why the standard deviation increases/decreases.

Solution: The range, standard deviation, and IQR would all increase.

The standard deviation measures the spread or variability of a dataset around its mean. It tells us how much the values in the dataset deviate from the mean on average. When the minimum value in a dataset decreases (while other values remain the same), the spread of the data increases because there is a larger gap between the minimum value and the mean. This causes individual data points to deviate more from the mean, which increases the average deviation and thus the standard deviation.

PROBLEM 10: COMPARING DISTRIBUTIONS BY MEAN AND STANDARD DEVIATION

Without doing any computations, compare the distributions by comparing their means and standard deviations.

(a) (i) 40 46 48 49 50 60 72

(ii) 40 46 48 49 50 60 80

Solution: The second distribution has a higher mean since the last number 80 > 72, and a higher standard deviation since 80 is farther away from the rest of the data than 72.

PROBLEM 10: COMPARING DISTRIBUTIONS BY MEAN AND STANDARD DEVIATION

Without doing any computations, compare the distributions by comparing their means and standard deviations.

(b) (i) 50 65 80 95 110

(ii) 30 40 80 120 130

Solution: Both distributions have the same mean since they are both centered around 80. However the second distribution has a greater standard deviation since the rest of the data are farther from the mean than the first.