Tutorial Worksheet 1 - Reviews of Descriptive Statistics (with Solutions)

Objective Questions

1. The _____ is the value you calculate when you want the arithmetic average.

   (a) Mean
   (b) Median
   (c) Mode
   (d) All of the above

   Solution: Mean.

2. The process of arranging data into rows and columns is called

   (a) Classification
   (b) Frequency distribution
   (c) Tabulation
   (d) Array

   Solution: Tabulation.

3. Find the median of the following data: 160, 180, 200, 280, 300, 320, 400

   (a) 140
   (b) 300
   (c) 180
   (d) 280

   Solution: 280.

4. The "average" type of grass used in UAE lawns is best described by

   (a) the mean
   (b) the median
   (c) the mode
   (d) the standard deviation

   Solution: the mode.

5. The median is a better measure of central tendency than the mean if

   (a) the variable is discrete
   (b) the distribution is skewed
   (c) the variable is continuous
   (d) the distribution is symmetric

   Solution: the distribution is skewed.

6. A set of data points follow a simple linear relation $y = 3x + 2$, where x is any integer number. The mean of the values of y for all values of x in the range $[1 \dots 100]$ (equally probable) is

   (a) 50
   (b) 50.5

(c) 152

(d) 153.5

Solution: 153.5.

7. The GM of the following data will be calculated as $X = [50, 125, 70, 56, 49, 98]$

   (a) 70

   (b) 74

   (c) 100

   (d) 101

   Solution: 70.

8. What is the primary characteristic of a set of data for which the standard deviation is zero?

   (a) All values of the variable appear with equal frequency.

   (b) All values of the variable have the same value.

   (c) The mean of the values is also zero.

   (d) None of the above is correct.

   Solution: All values of the variable have the same value.

9. If the standard deviation of x, y, z is p then the standard deviation of $3x + 5, 3y + 5, 3z + 5$ is?

   (a) $3p + 5$

   (b) $3p$

   (c) $p + 5$

   (d) $9p + 15$

   Solution: $3p$.

Subjective Questions

Problem 1. The wickets taken by a bowler in 10 cricket matches are as follows: 2 6 4 5 0 2 1 3 2 3 Find the mode of the data.

Solution:

No. of wickets taken by bowler in 10 cricket matches 2 6 4 5 0 2 1 3 2 3 . Since 2 wickets are taken by the bowler in maximum no. of matches. Hence the mode of the given data is 2.

Problem 2. If the mean of a frequency distribution is 100 and the coefficient of variation is 45%, then what is the value of variance.

Solution:

Coefficient of Variation = Standard Deviation/ Mean. Here, Coefficient of Variation = 0.45, Mean =100 $\Rightarrow$ Standard Deviation = $0.45 \times 100 = 45 \Rightarrow$ Variance = (Standard Deviation)$^2 = 45^2 = 2025$

Problem 3. For a given sample, the observation is as follows.

x	1	2	3	4	5	6
$f(x)$	25	50	10	30	40	20

x denotes a sample value and $f(x)$ denotes the frequency of occurrence of x. Find the five-point summary of the above data.

Solution:

Min $= 1, 1^{\text{st}}$ Quartile $(Q_1) = 2$, Max $= 6, 3^{\text{rd}}$ Quartile $(Q_3) = 5$, and median $= 4$

Problem 4. Calculate the mean, median and mode of the following data: $5, 10, 10, 12, 13$. Are these three equal?

Solution:

Sum of all observations $= 5+10+10+12+13 = 50$, and number of observations $= 5$. $\Rightarrow$ mean $= \frac{\text{sum of all observations}}{\text{Total observations}} = 50/5 = 10$
Here, $n = 5$ (odd). Hence, median $= [(5+1)/2]^{\text{th}}$ position $= 3^{rd}$ position $= 10$. Mode = Most frequent data $= 10$.
Hence, Mean = Median = Mode.

Problem 5. A frequency distribution of a set of 10 data is given below. Calculate the coefficient of variance of the data.

X	1	2	3	4	5	6	7	8	9	10
F(x)	1	3	5	7	9	2	4	6	1	0

Solution:

$$\text{Here, } \mu = \frac{1+6+15+28+45+12+28+48+9+0}{1+3+5+7+9+2+4+6+1+0} = 5.052632$$

$$\text{and } \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 f_i = 4.267, \text{ that is } \sigma = 2.065$$

Hence, for the given data, CV $= \frac{\sigma}{\bar{\mu}} * 100 = 40.885\%$

Problem 6. The accompanying data on the number of minutes used for cell phone calls in one month was generated to be consistent with summary statistics published in a report of a marketing study of San Diego residents **(Tele-Truth, March 2009)**:

189	0	189	177	106	201	0	212	0
306	0	0	59	224	0	189	142	3
71	165	236	0	142	236	130		

(a) Would you recommend the mean or the median as a measure of center for this data set? Give a brief explanation of your choice.

(b) Compute a trimmed mean by deleting the three smallest observations and the three largest observations in the data set and then averaging the remaining 19 observations. What is the trimming percentage for this trimmed mean?

(c) What trimming percentage would you need to use in order to delete all of the 0 minute values from the data set? Would you recommend a trimmed mean with this trimming percentage? Explain why or why not.

Solution:

(a) Calculate the mean

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{25}(0+0+0+\ldots.+236+306) = 119.08$$

Calculating the median:

The median is the middle number of ascending order of the data for odd number of observations. The given number of observations are 25, the median is the middle number and it is 142.

From the mean and median, observe that the median is the best measure of central tendency because of the outliers presence.

(b) The trimmed mean is,

$$\bar{x}_{tr} = \frac{\sum x_i}{n} = 115.736$$

The trimmed percentage is,

$$\text{Percentage} = \frac{\text{Trimmed observations}}{\text{Total observations}} \times 100\% = \frac{3}{25} \times 100\% = 0.12 \times 100\% \approx 12\%$$

(c) The trimmed percentage is,

$$\begin{aligned}\text{Percentage} &= \frac{\text{Trimmed observations}}{\text{Total observations}} \times 100\% \\ &= \frac{7}{25} \times 100\% = 0.28 \times 100\% \approx 28\%\end{aligned}$$

No, we don't recommend this percentage of trimmed mean because it is quite high.

Problem 7. 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 $\hat{p} = .80$ for the entire sample of 25 patients?

Solution:

The given sequence of success and failure is S S F S S S F F S S .

(a) Compute the value of the sample proportion of success.

From the given sequence there are 10 observations. Of those, there are 7 successes (S). So, the sample proportion of success is given by,

$$\hat{p} = \frac{\text{Favourable cases}}{\text{Total number of observations}} = \frac{7}{10} = 0.7$$

Therefore, the sample proportion of success is 0.7.

(b) Replace each S with 1 and F with a 0. The resultant sequence is 1 1 0 1 1 1 0 0 1 1

Compute the sample for the coded data.

$$\bar{x} = \frac{\sum x}{n} = \frac{1 + 1 + 0 + \ldots + 1}{10} = \frac{7}{10} = 0.7$$

Therefore, the sample mean of the coded data is 0.7.

Observe sample proportion of successes and the coded sample mean are equal. That is

$$\bar{x} = \hat{p} = 0.7$$

(c) The researcher added 15 patients to the study. So, now the total number of observations is 25. Estimate the number of successes required, in order to get the sample proportion of 0.80.

$$\text{i.e., } \hat{p} = 0.80 \Rightarrow \frac{X}{n} = 0.80 \Rightarrow \frac{X}{25} = 0.80 \Rightarrow X = 0.80 \times 25 \Rightarrow X = 20$$

Therefore, the researcher would require 20 successes out of 25 observations i.e., 13 successes out of 15 new observations to get the sample proportion of 0.80.

Model	Smart Fortwo	Chevrolet Aveo	Mini Cooper	Toyota Yaris	Honda Fit	Hyundai Accent	Kia Rio
Repair Cost	$1,480	$1,071	$2,291	$1,688	$1,124	$3,476	$3,701

Problem 8. The Insurance Institute for Highway Safety, published data on repair costs for cars involved in different types of accidents on June 11, 2009. In one study, seven different 2009 models of mini-cars and micro-cars were driven at 6 mph straight into a fixed barrier. The following table gives the cost of repairing damage to the bumper for each of the seven models:

(a) Compute the values of the variance and standard deviation. The standard deviation is fairly large. What does this tell you about the repair costs?

(b) The Insurance Institute for Highway Safety also gave bumper repair costs in a study of six models of minivans (December 30, 2007). Write a few sentences describing how mini-cars, micro-cars, and minivans differ with respect to typical bumper repair cost and bumper repair cost variability.

Model	Honda Odyssey	Dodge Grand Caravan	Toyota Sienna	Chevrolet Uplander	Kia Sedona	Nissan Quest
Repair Cost	$1,538	$1,347	$840	$1,631	$1,176	$1,603

Solution:

(a) We find the sample mean, $\bar{x}$.

$$\bar{x} = \frac{\sum x}{n} = \frac{1480 + 1071 + \cdots + 3476 + 3701}{7} = \frac{14831}{7} = 2118.714$$

The following data shows the repair costs for cars involved in different types of accidents.

The calculation of the variance:

Model	x	$x - \bar{x}$	$(x - \bar{x})^2$
Smart Fortwo	1480	−638.714	407955.6
Chevrolet Aveo	1071	−1047.71	1097705
Mini Cooper	2291	172.286	29682.47
Toyota Yaris	1688	−430.714	185514.5
Honda Fit	1124	−994.714	989455.9
Hyundai Accent	3476	1357.286	1842225
Kia Rio	3701	1582.286	2503629
Sum	14831	0	7056167

The sample variance,

$$s^2 = \frac{1}{n-1}\sum(x - \bar{x})^2 = \frac{7056167}{7-1} = 1176028$$

The sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum(x - \bar{x})^2} = \sqrt{\frac{1176028}{7-1}} = 1084.448$$

The data set has much variability because it consisting of all 7 observation. And also it tells us that the there is considerable variation in the repair costs.

(b) The following data shows the repair costs for cars involved in different types of accidents. We find the sample mean, $\bar{x}$.

$$\bar{x} = \frac{\sum x}{n} = \frac{1538 + 1347 + \cdots + 1176 + 1603}{6} = \frac{6505.631}{6} = 1355.833$$

The sample variance,

$$s^2 = \frac{1}{n-1}\sum(x-\bar{x})^2 = \frac{468494.8}{6-1} = 93698.97$$

The sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum(x-\bar{x})^2} = \sqrt{\frac{468498.8}{6-1}} = 306.1029$$

The data set has much variability because it consisting of all 7 observation. And also it tells us that the there is considerable variation in the repair costs.

Comparing the results of part (a) with part (b), the mean and standard deviations are less than the results of part (a). It tells us that the mean repair cost is very low, whereas the standard deviation of minivans showing lower repair cost variability for the minivans.

Problem 9. For the data on the number of minutes used for cell phone calls published by San Diego residents used in Problem 6.

(a) Compute the values of the quartiles and the interquartile range for this data set.

(b) Explain why the lower quartile is equal to the minimum value for this data set. Will this be the case for every data set? Explain.

Solution:

(a) Number of observation = 25 $\Rightarrow$ Median = $[\frac{25+1}{2}]^{th}$ observation = 142, 1$^{\text{st}}$ Quartile (Q_1) = Median of first 13 observations (arranged in ascending order) = 0, 3$^{\text{rd}}$ Quartile (Q_3) = Median of last 13 observations (arranged in ascending order)= 189

IQR = Upper Quartile - Lower Quartile = $189 - 0 = 189$.

(b) In the given data, there are seven 0's, and the total number of data is 25. The lower quartile is the median for the first half of the observations (arranged in ascending order). Here, there are seven 0's. Thus, the lower quartile and the minimum value are the same.

No this will not be true for every dataset

Problem 10. Fiber content (in grams per serving) and sugar content (in grams per serving) for 18 high fiber cereals are shown below.

Fiber Content	7	10	10	7	8	7	12	12	8	13	10	8	12	7	14	7	8	8
Sugar Content	11	6	14	13	0	18	9	10	19	6	10	17	10	10	0	9	5	11

(a) Find the median, quartiles, and interquartile range for the fiber content data set.

(b) Find the median, quartiles, and interquartile range for the sugar content data set.

(c) Are there any outliers in the sugar content data set?

(d) Explain why the minimum value for the fiber content data set and the lower quartile for the fiber content data set are equal.

(e) Construct a comparative boxplot and use it to comment on the differences and similarities in the fiber and sugar distributions.

Solution:

(a) Arranging the fiber content dataset in ascending order we obtain,

7 7 7 7 7 8 8 8 8 8 10 10 10 12 12 12 13 14

Number of observation = 18 is even. Hence, Median = $\frac{1}{2}\left(\frac{18}{2}^{th} + \left(\frac{18}{2}+1\right)^{th}\right)$. Thus, median = $(8+8)/2 = 8$.

1^{st} Quartile (Q_1) = Median of first half of observations = Median of (7 7 7 7 7 8 8 8 8) = 7.

3^{rd} Quartile (Q_3) = Median of second half of observations = Median of (8 10 10 10 12 12 12 13 14) = 12.

IQR = Upper Quartile − Lower Quartile = 12 − 7 = 5.

(b) Arranging the sugar content dataset in ascending order we obtain,

0 0 5 6 6 9 9 10 10 10 10 11 11 13 14 17 18 19

Number of observation = 18 is even. Hence, Median = $\frac{1}{2}\left(\frac{18}{2}^{th} + \left(\frac{18}{2}+1\right)^{th}\right)$. Thus, median = $(10+10)/2 = 10$

1^{st} Quartile (Q_1) = Median of first half of observations = Median of (0 0 5 6 6 9 9 10 10) = 6.

3^{rd} Quartile (Q_3) = Median of second half of observations = Median of (10 10 11 11 13 14 17 18 19) = 13.

IQR = Upper Quartile − Lower Quartile = 13 − 6 = 7.

(c) We can use the IQR method of identifying outliers to set up a "fence" outside of Q_1 and Q_3. Any values that fall outside of this fence are considered outliers. To build this fence we take 1.5 times the IQR and then subtract this value from Q_1 and add this value to Q_3. This gives us the minimum and maximum fence posts that we compare each observation to. Any observations that are more than 1.5 IQR below Q_1 or more than 1.5 IQR above Q_3 are considered outliers.

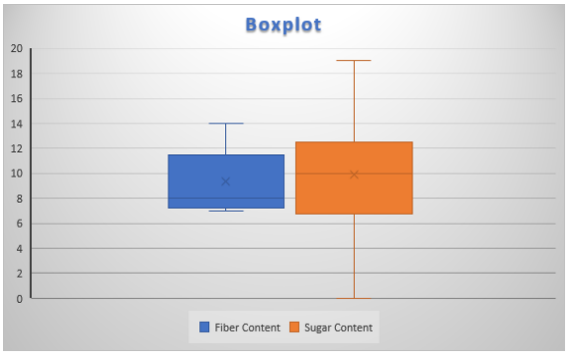
For sugar content dataset we calculated the following:

IQR $= 7, Q_1 = 6, Q_3 = 13 \Rightarrow$ Lower fence $= 6 - (1.5 \times 7) = -4.5$, and Upper fence $= 13 + (1.5 \times 7) = 23.5$.

Since, no observations lie outside the fence, so we conclude that there are no outliers in the dataset.

(d) In the fiber content data, there are five 7's and total number of data is 18. The lower quartile is the median for the lower half of the observations. That is, the 5^{th} observation is the lower quartile. Here, there are five 7's. Thus, the lower quartile and the minimum value is the same.

(e) The required boxplot is constructed and the following comments are made:



- The distribution of sugar content is roughly symmetrical, whereas the fiber content distribution seems to be positively skewed.
- Neither distribution contains outliers.
- There are more grams per serving of sugar content than of fiber content in these cereals (median of sugar content is higher).
- The sugar content is more variable.