Tutorial Topic 3 (Solution)

Numerical Descriptive Measures

Introduction

In this topic we will be looking at Numerical Descriptive Measures or, specifically, Measures of Central Tendency and Measures of Variation.

In the element Measures of Central Tendency one of the items we will be investigating is a pretty common term, an average. An average is a typical value but there are different types of averages. The most common average we use in everyday life is actually called the arithmetic mean. But there are two other types of averages. One is the Median, or the middle value (and this is the average we use when quoting an average house price in Real Estate) and the other is the Mode or the most occurring. And if I'm after a typical value, then I need to decide whether to use the Mean, the Median or, in some instances, the Mode.

Therefore, the aims of this tutorial are to:

- calculate and interpret numerical descriptive measures of central tendency, variation and shape for numerical data
- calculate and interpret descriptive summary measures for a population
- construct and interpret a box-and-whisker plot
- calculate and interpret the covariance and the coefficient of correlation for bivariate data

Textbook Questions/Answers/Readings

3.13 A manufacturer of torch batteries took a sample of 13 batteries from a day's production and used them continuously until they were drained. The numbers of hours they were used until failure were: [**Dataset:** BATTERIES.XLS]

342	426	317	545	264	451	
1049	631	512	266	492	562	298

a. Calculate the mean, median and mode. Looking at the distribution of times to failure, which measures of central tendency do you think are most appropriate and which least appropriate to use for these data? Why?

Hours: Original	
Mean	473.46
Standard Error	58.46
Median	451.00
Mode	#N/A
Standard Deviation	210.77
Sample Variance	44422.44
Kurtosis	4.18
Skewness	1.75
Range	785.00
Minimum	264.00
Maximum	1049.00
Sum	6155.00
Count	13.00

b. Calculate the range, variance and standard deviation.

As shown in Hours: original table above, Range = 785; Variance = 44422.44; Standard deviation = 210.77.

c. What would you advise if the manufacturer wanted to say in advertisements that these batteries 'should last 400 hours'? (Note: There is no right answer to this question; the point is to consider how to make such a statement precise.)

From the manufacturer's viewpoint, the worst measure would be to compute the percentage of batteries that last over 400 hours (8/13 = .61). The median (451) and the mean (473.5) are both over 400, and would be a better measure for the manufacturer to use in advertisements.

d. Suppose that the first value was 1,342 instead of 342. Repeat (a) to (c), using this value. Comment on the difference in the results.

Hours: Altered	
Mean	550.38
Standard Error	87.46
Median	492.00
Mode	#N/A
Standard Deviation	315.33
Sample Variance	99435.26
Kurtosis	2.70
Skewness	1.70
Range	1078.00
Minimum	264.00
Maximum	1342.00
Sum	7155.00
Count	13.00

The median seems to be better a descriptive measure of the data as the two outliers affect the mean.

From the manufacturer's viewpoint, the worst measure remains the percentage of batteries that last over 400 hours (9/13 = .69). The median (492) and the mean (550.38) are both well over 400, and would be a better measure for the manufacturer to use in advertisements.

The average life of batteries is higher in the second sample but the second sample appears to have a considerably higher deviation from the mean. The shape of the distribution of both is likely to be right-skewed as the mean is higher than the median.

Reading: Berenson Ch. 3, Section 3.1

3.27 A company wished to study its accounts receivable for two successive months. An independent sample of 50 accounts was selected for each month. The results are in the table below.

Frequency Distributions for Accounts Receivable		
Amount	March Frequency	April Frequency
$0 to under $2,000	6	10
$2,000 to under $4,000	13	14
$4,000 to under $6,000	17	13
$6,000 to under $8,000	10	10
$8,000 to under $10,000	4	0
$10,000 to under $12,000	0	3
Total	**50**	**50**

March			
Midpoint	Frequency	$f_j m_j$	$f_j m_j^2$
1,000	6	6,000	6,000,000
3,000	13	39,000	117,000,000
5,000	17	85,000	425,000,000
7,000	10	70,000	490,000,000
9,000	4	36,000	324,000,000
11,000	-	-	-
Totals	50	236,000	1,362,000,000
April			
Midpoint	Frequency	$f_j m_j$	$f_j m_j^2$
1,000	10	10,000	10,000,000
3,000	14	42,000	126,000,000
5,000	13	65,000	325,000,000
7,000	10	70,000	490,000,000
9,000	-	-	-
11,000	3	33,000	363,000,000
Totals	50	220,000	1,314,000,000

a. For each month, approximate the:

I. Mean

March: Mean = 236,000/50 = 4,720

April: Mean = 220,000/50 = 4,400

II. Standard Deviation

March: Standard deviation = $\sqrt{\frac{1362000000 - 50 \times 4720^2}{49}} = 2250.079...$

April: Standard deviation = $\sqrt{\frac{1314000000 - 50 \times 4400^2}{49}} = 2657.296...$

b. On the basis of your answers in (a), do you think the mean and the standard deviation of the accounts receivable have changed substantially from March to April? Explain.

The arithmetic mean has declined by $320 while the standard deviation has increased by $407.21.

Reading: Berenson Ch. 3, Section 3.3

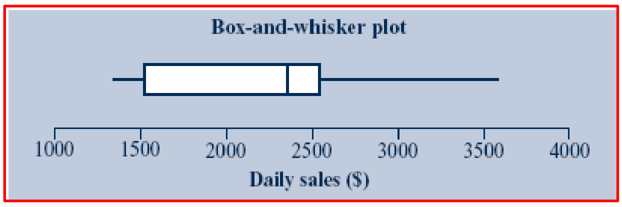
3.33 The sales per day, in dollars, at a certain store are: (from question 3.8) [Dataset: SALES.XLSX]

1520	2620	3360	3550	1350	2545	1430	2400	3580
2390	1525	2400	1420	1550	2390	1560	1680	2330

a. List the five-number summary.

Five-number summary 1350; 1525; 2360; 2545; 3580

b. Construct the box-and-whisker plot and discuss the daily sales distribution for the stores.



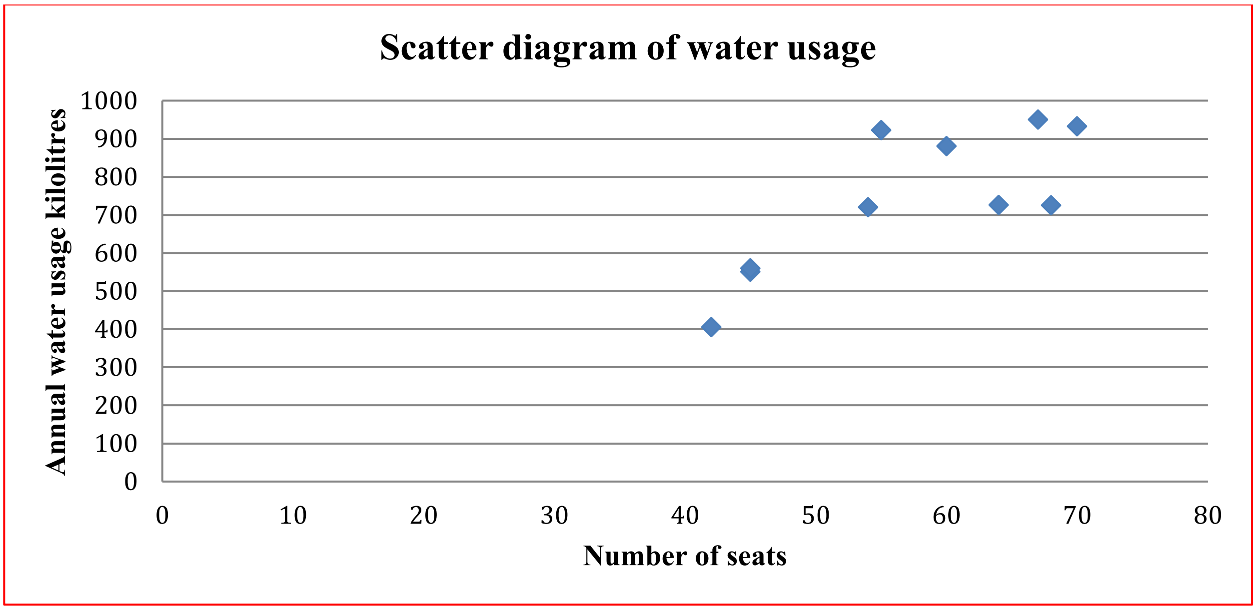
The results are inconsistent. The right-hand whisker is far longer than the left-hand and the distance from the median to $X_{largest}$ is greater than the distance from the median to $X_{smallest}$ indicating right-skewness. However, the left-hand side of the box is far longer than the right-hand side indicating left-skewness.

Reading: Berenson Ch. 3, Section 3.4

3.39 A local council is interested in the relationship between the size of local restaurants, measured as number of seats, and their annual water usage, in kilolitres. From a random sample of 10 local restaurants the following information was obtained. [Dataset: WATER2.XLS]

Number of seats X	Annual water usage Y (kilolitres)
60	880
45	550
54	720
68	725
70	932
55	922
67	950
45	560
64	726
42	405

a. Construct a scatter diagram for the data and comment on any apparent relationship between restaurant size and annual water usage.



The scatter diagram shows a positive relationship between size of local restaurants (as measured in number of seats) and annual water usage.

b. Calculate the sample covariance and coefficient of correlation. Are these values what you expected from the scatter diagram?

cov(seats, water) = 1546, r = 0.7941. A positive relationship was expected

c. What conclusions can you reach about the relationship between restaurant size and annual water usage?

Based on correlation coefficient, there is a fairly strong positive linear relationship between the size of local restaurants (as measured in number of seats) and annual water usage

Reading: Berenson Ch. 3, Section 3.5

TEXTBOOK REFERENCE:

Basic Business Statistics: Concepts and Applications. *Berenson, M.L. Levine, D.M. Szabat, K.A. O'Brien, M. Jayne, N. Watson, J.* 5th edition. 2019. Pearson Australia Group Pty Ltd. ISBN 9781488617249. Chapter 3, sections 3 to 3.6