**Assignment: Summarizing Ratio Variables**

|  |  |
| --- | --- |
|  | **Unit 10** - Comprehending the meaning and implications of basic summary measures for ratio variables. Mean, median, quartiles, interquartile range, |
|  | Central tendency: mode, median and meanIQR and box plotVariance and standard deviation |  | (Mode (bimodal, multi-model)medianmeanquartileinterquartile rangebox plotoutliers |
|  | De Veaux, ch. 4, p. 52-66 (15 p.)And/Or:Babbie, Ch. 14, p. 416-425 |  | No groups are formed. This is an individual assignment. |  | 90 minutes |
|  | Part A is to be completed individually without SPSS or any other program. Part B is to be completed individually by using SPSS. |

Part I:

*Given:*

A researcher is interested in whether the use of social media in terms of time spent on platforms like twitter, facebook etc…. differs across neighborhoods in a city. S/he collects data in the form of exact number of minutes a specific computer in a specific neighborhood remained active in one of the predetermined social media platforms. S/he collects small and unequal numbers of observations in order to have a preliminary view of the situation. The following picture emerges. Here each figure represents the number of minutes spent on social media by a distinct person living in designated area.

Table 1: minutes spend on social media platforms: observations from 3 neighborhoods



1. Take a pencil, an eraser, several sheets of empty paper, and a hand-calculator. DO NOT use SPSS or any other program or programming language for this part. Rearrange the observations for each neighborhood in ascending order in the table below.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Neighborhood |  |  |  |  |  |  |  |  |  |  |  |
| A | 5 | 32 | 48 | 67 | 69 | **70** | 80 | 83 | 93 | 100 | 117 |
| B | 5 | 15 | 20 | 25 | **30** | **75** | 103 | 128 | 129 | 165 |  |
| C | 12 | 100 | 120 | 121 | **147** | 180 | 215 | 264 | 411 |  |  |

2. Calculate the mean and the standard deviation to the mean of time spent on social media for each neighborhood. Further, calculate the median, the 1st quartile, 3rd quartile and the Interquartile range. Put all of the values in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Neighborhood | Mean | Standard deviation | 1st quartile | Median | 3rd quartile | Interquartile range |
| A | 69.4 | 31.7 | 57.5 | 70 | 88 | 30.5 |
| B | 69.5 | 57.8 | 20 | 52.5 | 128 | 108 |
| C | 174.4 | 112.7 | 120 | 147 | 215 | 95 |

3. What can you say about the differences or similarities between neighborhood A and B, based on the mean and standard deviation?

A and B differ in terms of their standard deviations. Computers (people?) in A are more similar to each other compared to people in B

4. Why is the standard deviation important in this comparison?

Standard deviation informs us the extent to which observations are dispersed around the mean value. In other words, it is a measure of the ability of the mean value to represent the entire dataset

5. Why do the mean and median values differ so much between neighborhood B and C?

Because Neighborhood B and C contain some extreme values

6. Why do you think the median may be more useful in summarizing a dataset than the mean?

Because it is not affected by the extreme values

7. What does median tell you in terms of the distribution of observations across the range of data?

50% of observations fall below and another 50% fall above the median

8. What do the 1st and the 3rd quartiles inform you about?

they divide the first and the second halves into two more pieces; in this way the entire dataset divided into four equal pieces

9. Where do the middle 50% of the observations lie?

they lie in between 1st and 3rd quartiles

10. What does interquartile range tell you in this sense?

it informs us about the dispersion (min and max value) of the middle 50% of the data

11. Create a box plot. For this, find whisker ‘points’. Please fill-in the following table:

|  |  |  |
| --- | --- | --- |
|  |  | Neighborhoods |
|  |  | A | B | C |
|  | IQR (see table 1) | 30,5 | 108 | 195 |
| Lower whisker | 1st quartile minus 1.5 \* IQR | 11.75 | -142 | -22.5 |
|  | Smallest value in the dataset, which is within the 1st quartile minus 1.5 \* IQR | 32 | 5 | 12 |
|  | Outliers? (if so, which one) | 5 | no | no |
| Upper whisker | 3rd quartile plus 1.5\*IQR | 133.75 | 290 | 357.5 |
|  | Highest value in the dataset within 3rd quartile plus 1.5\*IQR | 117 | 165 | 264 |
|  | Outliers? | No | no | 411 |

12. Draw a boxplot for each of the three neighborhoods and identify the (possible) outliers.

No. of Observations

400

300

200

100

0 Neighborhoods

A B C



Only A and C have outliers

13. Examine the effect of outliers. Eliminate outliers from those datasets which have them, and then

1. re-calculate the mean and standard deviation for each neighborhood for which you have removed outliers.
2. Use these new estimations and those that you have done above to fill in the following table:

|  |  |  |
| --- | --- | --- |
| Neigborhood | Based on original data | After removal of outliers |
| Mean | SD | Mean | SD |
| A | 69.4 | 30.2 | 75.9 | 23.4 |
| B | 69.5 | 55.1 |  |  |
| C | 174.4 | 107.5 | 144.8 | 71.7 |

1. What is the impact of eliminating outliers on mean and standard deviation estimates? eliminating outliers decrease the standard deviation while changing the magnitude of the mean in opposite direction from the that of the outlier
2. How would this effect your interpretations in terms of similarities and differences between different neighborhoods?

the similarity between A and B appears to be the result of the outlier in B, eliminating this outlier makes the similarity disappear

Part II.

The researcher decides to collect more data in order to make a thorough comparison between neighborhoods. The resulting dataset (“neighborhood\_data.sav”) is available on blackboard. Please download this dataset and complete the following exercises by using SPSS.

1. Prepare a **descriptive table** (analyze-explore then check statistics) with mean, sd, 1st quartile, 3rd quartile, median, interquartile range, minimum and maximum values for each neighborhood. How do these numbers inform you as to the distribution of observations across the range?

|  |
| --- |
| **Descriptives** |
|  | Statistic | Std. Error |
| city A social media use | Mean | 69,43 | ,939 |
| 95% Confidence Interval for Mean | Lower Bound | 67,59 |  |
| Upper Bound | 71,27 |  |
| 5% Trimmed Mean | 70,38 |  |
| Median | 71,00 |  |
| Variance | 881,188 |  |
| Std. Deviation | 29,685 |  |
| Minimum | 1 |  |
| Maximum | 121 |  |
| Range | 120 |  |
| Interquartile Range | 37 |  |
| Skewness | -,605 | ,077 |
| Kurtosis | -,070 | ,155 |
| city B social media use | Mean | 72,15 | 1,765 |
| 95% Confidence Interval for Mean | Lower Bound | 68,69 |  |
| Upper Bound | 75,61 |  |
| 5% Trimmed Mean | 70,71 |  |
| Median | 72,00 |  |
| Variance | 3115,824 |  |
| Std. Deviation | 55,820 |  |
| Minimum | 1 |  |
| Maximum | 169 |  |
| Range | 168 |  |
| Interquartile Range | 106 |  |
| Skewness | ,286 | ,077 |
| Kurtosis | -1,466 | ,155 |
| city C social media use | Mean | 171,36 | 3,483 |
| 95% Confidence Interval for Mean | Lower Bound | 164,52 |  |
| Upper Bound | 178,19 |  |
| 5% Trimmed Mean | 167,04 |  |
| Median | 144,00 |  |
| Variance | 12129,846 |  |
| Std. Deviation | 110,136 |  |
| Minimum | 1 |  |
| Maximum | 422 |  |
| Range | 421 |  |
| Interquartile Range | 109 |  |
| Skewness | ,794 | ,077 |
| Kurtosis | ,215 | ,155 |

1. Create a **boxplot** in ‘graphs’ for each neighborhood. By using the boxplots compare three neighborhoods in terms of the way in which their residents are spread across the range. Can you interpret what you see in terms of the focus of the research?



1. Do you see any outliers in any one of the neighborhoods? If so, **eliminate the outliers**, and prepare the descriptives again for the mean, median and standard deviation scores and compare them with what you have obtained in (1). What is the effect of eliminating outliers? What might be the substantive meaning of outliers in a specific neighborhood?

**4 outliers. Doesn’t affect the descriptives very much. As there are only 4 outliers on 1000 cases, the influences is very low.**

|  |
| --- |
| **Descriptives** |
|  | Statistic | Std. Error |
| city A social media use | Mean | 69,54 | ,938 |
| 95% Confidence Interval for Mean | Lower Bound | 67,70 |  |
| Upper Bound | 71,38 |  |
| 5% Trimmed Mean | 70,51 |  |
| Median | 71,00 |  |
| Variance | 876,870 |  |
| Std. Deviation | 29,612 |  |
| Minimum | 1 |  |
| Maximum | 121 |  |
| Range | 120 |  |
| Interquartile Range | 37 |  |
| Skewness | -,605 | ,077 |
| Kurtosis | -,059 | ,155 |
| city B social media use | Mean | 72,09 | 1,770 |
| 95% Confidence Interval for Mean | Lower Bound | 68,62 |  |
| Upper Bound | 75,56 |  |
| 5% Trimmed Mean | 70,64 |  |
| Median | 72,00 |  |
| Variance | 3119,181 |  |
| Std. Deviation | 55,850 |  |
| Minimum | 1 |  |
| Maximum | 169 |  |
| Range | 168 |  |
| Interquartile Range | 106 |  |
| Skewness | ,289 | ,077 |
| Kurtosis | -1,464 | ,155 |
| city C social media use | Mean | 170,44 | 3,466 |
| 95% Confidence Interval for Mean | Lower Bound | 163,64 |  |
| Upper Bound | 177,24 |  |
| 5% Trimmed Mean | 166,02 |  |
| Median | 143,50 |  |
| Variance | 11967,608 |  |
| Std. Deviation | 109,397 |  |
| Minimum | 1 |  |
| Maximum | 422 |  |
| Range | 421 |  |
| Interquartile Range | 109 |  |
| Skewness | ,802 | ,077 |
| Kurtosis | ,266 | ,155 |

<<end of the assignment >>