**Descriptive Statistics Using R**

The use of descriptive statistics is good way to describe, show, or summarize data in a meaningful way. Methods for descriptive statistics often include central tendency measures, variability measures, and graphs for visual representation. For this project, descriptive statistics was used for six data sets: Binominal, BN1, BN2, In, N1, N2.

**Results**

* 1. **Binomial**

For the Binomial dataset, the central tendency was measured using the mean. The arithmetic mean is computed by taking the sum of all the data values, and then divides the sum by the total number of values in the set. Using R, the arithmetic mean came out to 70.17. This means on average of the set of numerical values was 70.17.

For variability measures, the range, variance, and standard deviation were computed. The range measures how spread out the data is. The range is the difference between the maximum and minimum values in the data set. For the Binomial dataset, the maximum value was 84, the minimum value was 57, and the range came out to 27.

Next, the variance and standard deviation were measured. The variance measures how spread out the data is from the mean, while the standard deviation is the square root of the of the variance. Using R, the variance came out to be 21.01 and the standard deviation came out to be 4.69. This means on average the frequency is 70.17, plus or minus 4.69.

Below the data is presented using a histogram. The distribution of the histogram below shows a fairly normal distribution. Based on the histogram, it does not appear that there are any outliers since there aren't any bars that are separated from the other observations.

A screenshot of a cell phone

Description automatically generated

Another way to present the data is using a box plot. A box plot displays the five-number summary of a set of data, which is the minimum, first quartile, median, third quartile, and maximum. In R's box plot, an observation can indicate potential outlier if it falls more than 1.5 times the IQR (Inter-Quartile Range) below the first quartile (Q1) or above the third quartile (Q3). According to Greenwood and Banner (n.d.), potential outliers are plotted with circles and the whiskers of the box plot are reduced to only go as far as observations that are within 1.5\*IQR of the upper quartile and lower quartile. The box part of the boxplot goes from Q1 to Q3, and the median is represented as a line inside the box (Greenwood & Banner, n.d.). The five-number summary for the box plot is the following:

Min: 57.00

1st Quartile: 67

Median: 70.00

3rd Quartile: 73.00

Max: 84.00

Looking into the summary statistics the IQR is 14, which is computed by subtracting the 73, the third quartile with 67, the first quartile. The box plot below shows that there are two potential outliers. One potential outlier falls more than 1.5\***IQR**below Q1, and the other potential outlier falls more than 1.5\*IQR above Q3. The distribution is relatively symmetrical with the distance from the minimum to the median being the same as the distance from the median to the maximum. Also, the distance from Q1 to the median is relatively the same as the median to Q3.

A screenshot of a cell phone

Description automatically generated

* 1. **BN1**

For the dataset BN1, the central tendency was indicated using the mean. Using R, the mean came out to 9.99. This means the average of the set of numerical values was 9.99.

For variability measures, the range, variance, and standard deviation were computed. The range measures the dispersion of the data. For the range, the maximum value was 18.612 and the minimum values was 1.781. The difference between these two values makes the range to be 16.831. Next, the variance and standard deviation were computed. The variance measures how spread out the data is from the mean, while the standard deviation is the square root of the of the variance. Using R, the variance computed to 4.00 and the standard deviation computed to 2.00. This means on average the frequency is 9.99, plus or minus 2.00.

There are several ways the data can presented. One way is using a histogram. For the data set, the histogram shows a normal distribution and no outliers.

A close up of a logo

Description automatically generated

Another way to present the data is using a box plot. The five-number summary for the box plot is the following:

Minimum: 1.781

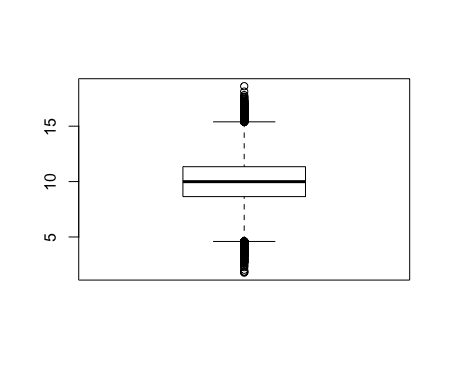
1st Quartile: 8.643

Median: 9.993

3rd Quartile: 11.343

Max :18.612

The IQR computes to 2.7. Below the boxplot shows a good number of potential outliers that fall more than 1.5\*IQR below Q1 and above Q3. The distribution is relatively symmetrical with the distance from the minimum to the median being the same as the distance from the median to the maximum. Also, the distance from Q1 to the median is relatively the same as the median to Q3.



* 1. **BN2**

For the BN2 dataset, the central tendency was identified using the mean. Using R, the mean came out to be 10.99. This means the average of the set of numerical values was 10.99.

For variability measures, the range, variance and standard deviation were computed. For the range, the maximum value was 15.161 and the minimum value was 6.638. The difference between these two values makes range to be 8.523. Next, using R the variance and standard deviation were measured. The variance of the data set is approximately 1 and the standard deviation is approximately 1. Both standard deviation and variance measure how spread out the data is from the mean. The standard deviation of 1 means on average the frequency was 10.99, give or take 1.

Below the data is presented using a histogram. The histogram shows a normal distribution with no outlier.

A screenshot of a cell phone

Description automatically generated

Along the histogram, the data can be presented with a box plot. The five-number summary for the box plot is the following:

Min: 6.638

1st Quartile: 10.321

Median: 10.998

3rd Quartile: 11.667

Max: 15.161

Looking at the five-number summary, the IQR computes to 1.346. Below the box plot shows that there are a number of potential outliers that fall more than 1.5\***IQR** below Q1 and above Q3. The distribution is roughly symmetrical with the distance from the minimum to the median being the same as the distance from the median to the maximum. Additionally, the distance from Q1 to the median is about the same as the median to Q3.



* 1. **In**

For the In dataset, the central tendency was identified using the mean. Using R, the mean came out to be 18.99. This means the average of the set of numerical values was 18.99.

For variability measures, the range, variance, and standard deviation were computed. For the range, the maximum value was 43.00, and the minimum amount was 3.00. The difference between these two values makes the range to be 40.00. Next, using R the variance and standard deviation were measured. The variance of the data set is 19.03, and the standard deviation is approximately 4.36. This means the average frequency was 18.99, plus or minus 4.36.

Below the data is presented using a histogram. The histogram for the data set shows a roughly normal distribution with no outliers.

A close up of a logo

Description automatically generated

In addition to a histogram, a box plot was also used to present the data. The five-summary for the box plot is the following:

Minimum: 3.00

1st Quartile: 16.00

Median: 19.00

3rd Qurtile:22.00

Max: 43.00

By subtracting the Q1 form Q3, IQR computes to 6. Below the box plot shows that there is a greater number of potential outliers that fall more than1.5\*IQR above Q3 than 1.5\*IQR below Q1. The box plot displays a slight right skew with the distance from the minimum to the median being smaller than the distance from the median to the maximum.

A screenshot of a cell phone

Description automatically generated

* 1. **N1**

For the N1.cvs dataset, the central tendency was identified using the mean. Using R, the mean came out to be 10.10. This means the average of the set of numerical values was 10.10.

For variability measures, the range, variance and standard deviation were computed. For the range, the maximum value was 15.77 and the minimum value was 5.06. The difference between these two values makes range to be 10.71. Next, using R the variance and standard deviation were measured. The variance of the data set is 3.90 and the standard deviation is approximately 1.97. This means on average the frequency was 10.10, plus or minus 1.97.

To present the data, a histogram was used. The histogram below shows a roughly normal distribution with an outlier.

A close up of a logo

Description automatically generated

In addition to a histogram, a box plot was also used to present the data. The 5-number summary for the box plot is the following:

Minimum: 5.064

1st Quartile: 8.75

Median: 10.062

3rd Quartile:11.397

Maximum: 15.767

By subtracting the third quartile for the first quartile, the IQR for the box plot computes to 2.647. Below box plot shows that there is one potential outlier that falls more than 1.5\*IQR above Q3. The box plot displays a slight left skew with the distance from the minimum to the median being larger than the distance from the median to the maximum.

A screenshot of a cell phone

Description automatically generated

* 1. **N2**

For the N2 dataset, the central tendency was identified using the mean. The mean was computed with R. The value of the mean came out to be 11.68. This means the average of the set of numerical values was 11.68.

For variability measures, the range, variance and standard deviation were computed. For the range, the maximum value was 22.63 and the minimum value was 0.65. The difference between these two values makes range to be 21.98. Using R, the variance computed to 18.54 and the standard deviation computed 4.31. This means the average frequency was 11.68, plus or minus 4.31.

Below the data is presented using a histogram. The histogram shows a roughly normal distribution with no outliers.

A close up of a logo

Description automatically generated

In addition to the histogram, the data is also presented by a box plot. The 5-number summary for the box plot is the following:

Minimum: 0.6503

1st Quartile: 8.9959

Median: 11.8934

3rd Quartile: 14.0504

Max: 22.6275

The IQR for the box plot computes to 5.0545. The box plot shows that there is one potential outlier that falls more than 1.5\*IQR below Q1, and another potential outlier that falls more than 1.5\*IQR above Q3. The distribution is slightly skew left with the distance from the minimum to the median being larger than the distance from the median to the maximum. Also, the distance from Q1 to the median is larger than the median to Q3.

A screenshot of a social media post

Description automatically generated

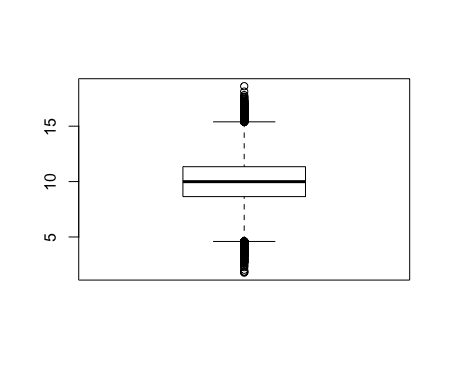
**Comparing Data Sets**

**2.1 BN1 vs BN2**

A close up of a logo

Description automatically generatedA screenshot of a cell phone

Description automatically generated



When measuring the central tendency, BN2 had a larger mean value of 10.99, than BN1 with a mean value of 9.99. This means on average BN1 had a higher frequency than BN2. Next, when comparing the variability of BN1 and BN2, BN1 had a larger standard deviation of 2.00, compared to BN2’s standard deviation of 0.99. Additionally, BN1 had a larger range of 16.831, compared to BN2’s range of 8.523. This means that the data in BN1 is more spread out than BN2. Shown on the histogram, both BN1 and BN2 are roughly symmetrical and have no outliers. However, shown on the box plot, both BN1 and BN2 show high levels of potential outliers that fall more than 1.5\*IQR below Q1 and above Q3.

**2.2 N1 vs N2**

A close up of a logo

Description automatically generatedA close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generated

When measuring the central tendency, N2 had a larger mean value of 11.68, than BN1 with a mean value of 10.10. This means on average N2 had a higher frequency than N1. Next, when comparing the variability of N1 and N2, N1 had a smaller standard deviation of 1.98, compared to N2’s standard deviation of 4.30. Additionally, N1 had a smaller range of 10.71, compared to N2’s range of 21.98. This means the values in N1 is less spread out than N2. On the histogram, both N1 and N2 have a relatively normal distribution. However, on the histogram it seems that N1 had an outlier and N2 did not. Additionally, when looking at the box plot it seems that BN1 had one potential outlier that falls more than 1.5\*IQR above Q3, and BN2 two potential outliers. BN2 had potential outliers that fall more than 1.5\*IQR below Q1 and above Q3.