### Comparison of Chicago Bulls vs Miami Heat

Southern New Hampshire University

MAT 243 Project One Summary Report

Professor In Hak Moon

February 6, 2023

The problem that is being solved is analyzing a dataset of basketball games to understand the factors contributing to a team winning or losing. The data set being used is a collection of 246 games from the 1996 and 2013 NBA seasons, which includes information such as the teams playing, the scores, the game's location, and the game's outcome. To perform the analysis, statistical methods such as linear regression and logistic regression will be used to examine the relationship between different variables and the game's outcome. Additionally, data visualization techniques will be used to explore and understand the data provided.

The table shows that the team that was picked for the analysis is the Heat, and the years selected for the analysis are 2013-2015, while the team assigned for the comparative study is the Bulls and the range of years is 1996 - 1998.

This information can be used to compare the performance of the Heat during the years 2013-2015 and that of the Bulls during the years 1996-1998. This can be done by comparing various statistics such as points scored, points allowed, win-loss ratio, etc. Data visualization techniques can also be used to explore and understand the data.

|  | **Name of Team** | **Assigned Years** |
| --- | --- | --- |
| Personal Team | Miami Heat | 2013 – 2015 |
| Assigned Team | Chicago Bulls | 1996 – 1998 |

Data visualization is a powerful tool for studying data distributions and trends. We can quickly identify patterns, outliers, and other essential data features by creating visual representations of data. In this activity, I picked a histogram as the best plot to describe the data distribution of the variable. I chose this plot because it is a simple and effective way to show the issuance of a variable and how often specific values occur. By visually inspecting the histogram, we can see that the distribution of the variable is roughly symmetric, with most of the data points concentrated in the middle of the range. This signifies that there is no clear bias in the data and that it is likely to be a representative sample. Additionally, the histogram allows us to see any outliers, skewness, and kurtosis in the data.

In this activity, I picked a boxplot as the best plot to describe the data distribution of the variable for the assigned team. I chose this plot because it is a compact and informative way to show the distribution of a variable and its basic statistical properties, such as the median, quartiles, and outliers.

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generatedBy visually inspecting the boxplot, we can see that the distribution of the variable is roughly symmetric and has no outliers. The median, represented by the line inside the box, is near the center of the distribution and the quartiles, represented by the top and bottom of the box, show that the data is tightly distributed. This signifies that the variable is likely to be a representative sample with no extreme values. Additionally, we can see that the distribution of the variable is not skewed to the left or right, showing a normal distribution.

In general, data visualization is used to compare two different data distributions by creating visual representations of the data, such as bar charts, line plots, and scatter plots. These plots allow us to quickly show patterns, outliers, and other essential data features and compare them with different data sets.

In this activity, I picked a side-by-side boxplot as the best plot to compare the data distributions of my team with the assigned team. I chose this plot because it is a compact and informative way to show the distribution of a variable and its basic statistical properties, such as the median, quartiles, and outliers, for two different data sets on a single graph.

*Chart, box and whisker chart

Description automatically generatedChart, histogram

Description automatically generated*By visually inspecting the side-by-side boxplots, we can see that the two distributions are similar in shape and spread, but the assigned team has a slightly higher median and more outliers. This signifies that the two groups are identical in performance, but the given team performs marginally better with some extreme values.

Central tendency and variability measures are generally used to analyze a data distribution. Measures of central tendency, such as mean, median, and mode, provide a single value representing the distribution's center. Measures of variability, such as range, variance, and standard deviation, give information on the spread of the data.

The mean is the data set's average; in this scenario, it is 102.56. It is the center of the distribution by considering all the values of the data set.

The median is the middle value of the data set, and in this scenario, it is 103.0. It is the center of the distribution by dividing the data set into two equal parts.

The variance is a measure of the spread of the data set, being how far each number in the set is from the mean and is calculated as 126.05.

The Standard deviation is the square root of the variance, a measure of the spread of the data set, and it is calculated as 11.23.

The mean and the median are both measures of central tendency and are used to describe the distribution of points scored by your team in home games. For example, the mean of 102.56 and median of 103.0 are close to each other, showing that the data is symmetric and bell-shaped. This shows that the data is evenly distributed around the center.

The median is the best measure of the central tendency to use to represent the center of the distribution based on its skew. Since the data is symmetric and bell-shaped, the median is the best measure of central tendency as it is not affected by outliers or extreme values; it is more robust and provides a better representation of the central movement of the data.

|  |  |
| --- | --- |
| Statistic | Value |
| Mean | 102.56 |
| Median | 103.0 |
| Variance | 126.05 |
| Standard Deviation | 11.23 |

Confidence intervals are generally used in estimating a population's central tendency measures. They provide a range of values likely to contain the valid population means or median with a certain confidence level. The confidence level is usually set at 95% or 99%, meaning that if the process of sampling and estimation is repeated many times, the true population means or median, would be included in the calculated interval in 95% or 99% of the cases.

In this scenario, the 95% confidence interval for teams' average relative skill (ELO) from 2013 to 2015 is (1502.02, 1507.18). This means that we are 95% confident that the actual average relative skill of teams in the league lies between 1502.02 and 1507.18 in the years 2013 to 2015.

The interval would have been different if a different confidence level had been used. For example, if a 99% confidence level was used, the gap would be more expansive, resulting in more uncertainty around the valid population means.

The probability that a given team in the league has a relative skill level less than that of the team you picked is 0.841. Therefore, it is not unusual that a team has a skill level less than your team as the probability is relatively high.

|  |  |
| --- | --- |
| Confidence Level (%) | Confidence Interval |
| 95% | (1502.02, 1507.18) |

The 95% confidence interval for the average relative skill (ELO) of teams in the years 1996 to 1998 is (1487.71, 1493.6). This means that we are 95% confident that the true average relative skill of teams in the league lies between 1487.71 and 1493.6 in the years 1996 to 1998.

If a different confidence level had been used, the interval would have been different. For example, if a 99% confidence level was used, the interval would be wider, resulting in more uncertainty around the true population means.

When compared to the previous confidence interval, this interval is narrower. This signifies that the relative skill of teams in the range of years 1996 to 1998 is more consistent than the average relative skill of teams in the range of years 2013 to 2015. This could be due to several factors, such as differences in the overall competitiveness of the league, changes in the rules or coaching strategies, or other external factors.

|  |  |
| --- | --- |
| Confidence Level (%) | Confidence Interval |
| 95% | (1487.71, 1493.6) |

The statistical analyses that were performed in this scenario provided information about the distribution of points scored by a team in home games, the comparison of two different data distributions, and the estimation of the measures of central tendency for a population.

The practical importance of these analyses is that they allow us to understand a team's performance better and identify patterns, outliers, and other essential data features. For example, understanding the distribution of points scored by a team in home games can identify trends and predict future performance. We can identify similarities and differences between groups by comparing data distributions and making inferences about their relative performance. By estimating the measures of central tendency for a population, we can make inferences about the overall performance of teams in a league over a certain period.

In terms of the scenario, these results mean that we can identify patterns in the team's performance in home games and make predictions about future performance. They also indicate that the team is similar in performance to the assigned team, but the assigned team has a slightly better performance with some extreme values. Furthermore, the results indicate that the relative skill of teams in the league in the range of 1996 to 1998 is more consistent than the average relative skill of teams in the years 2013 to 2015. This could be due to several factors, such as differences in the overall competitiveness of the league, changes in the rules or coaching strategies, or other external factors. Overall, these results provide a better understanding of the performance of teams in the league and allow for more informed decision-making and predictions about future performance.

|  | **game\_id** | **year\_id** | **fran\_id** | **pts** | **opp\_pts** | **elo\_n** | **opp\_elo\_n** | **game\_location** | **game\_result** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 199511030CHI | 1996 | Bulls | 105 | 91 | 1598.2924 | 1531.7449 | H | W |
| **1** | 199511040CHI | 1996 | Bulls | 107 | 85 | 1604.3940 | 1458.6415 | H | W |
| **2** | 199511070CHI | 1996 | Bulls | 117 | 108 | 1605.7983 | 1310.9349 | H | W |
| **3** | 199511090CLE | 1996 | Bulls | 106 | 88 | 1618.8701 | 1452.8268 | A | W |
| **4** | 199511110CHI | 1996 | Bulls | 110 | 106 | 1621.1591 | 1490.2861 | H | W |

|  | **game\_id** | **year\_id** | **fran\_id** | **pts** | **opp\_pts** | **elo\_n** | **opp\_elo\_n** | **game\_location** | **game\_result** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 201210300MIA | 2013 | Heat | 120 | 107 | 1666.3193 | 1586.1121 | H | W |
| **1** | 201211020NYK | 2013 | Heat | 84 | 104 | 1647.6675 | 1548.2699 | A | L |
| **2** | 201211030MIA | 2013 | Heat | 119 | 116 | 1650.0934 | 1554.4674 | H | W |
| **3** | 201211050MIA | 2013 | Heat | 124 | 99 | 1656.5652 | 1504.0280 | H | W |
| **4** | 201211070MIA | 2013 | Heat | 103 | 73 | 1659.7239 | 1361.5804 | H | W |

\*Only showing the first five entries.