# MAT 243 Project One Summary Report

Benjamin Leanna

Benjamin.leanna@snhu.edu

Southern New Hampshire University

## Introduction: Problem Statement

The problem that I’m going to solve is that as a data analyst for a basketball team, I have access to a large set of historical data that I can use to analyze performance patterns. The difference patterns that they are looking for are for the Chicago Bulls from the years 1996-1998, the Miami Heat from years 2013-2015, and not just the comparison of the two teams, but also the relative skills of both teams and relative skills of all teams in those years.

The different data sets that I will be using are points scored by the team in a game to measure the relative skill level of the team in the league from the years when the team played the games.

Different statistical methods I’ll be using for this analysis will include calculating the mean, median, variance, and standard deviation along side the confidence intervals.

## Introduction: Your Team and the Assigned Team

The team that was assigned to me was the Chicago Bulls between the years of 1996 and 1998. The team that I chose was the Miami Heat between the years of 2013 and 2015.

Table 1. Information on the Teams

|  | **Name of Team** | **Assigned Years** |
| --- | --- | --- |
| 1. Yours | Miami Heat | 2013 - 2015 |
| 2. Assigned | Chicago Bulls | 1996-1998 |

## Data Visualization: Points Scored by Your Team

Data visualization is basically another form of visual art that grabs our interest and keeps our eyes on what is being delivered. Its tools are essential in analyzing tremendous amounts of information and make data-driven decisions in a way that is easy to look at, see the trends, and notice the outliers.

**![Chart

Description automatically generated]()**

For the data visualization of my team of points scored over the years, I decided to choose the scatterplot instead of the histogram for specific reason. The reason being that even though the histogram shows the frequency of games the points were scored in, the scatterplot, to me, showed a better visual on the average scoring over the years.

With this representation it not only shows a nice curve of the average but the minimal outliers that also are frequented.

## Data Visualization: Points Scored by the Assigned Team

*![Chart

Description automatically generated]()*

Again, for the data visualization of the assigned team to me, I also chose the scatterplot; And for the same reasons. You can, again, see the average over the years and this time see the minimum average is a bit higher than it was with the team I chose.

You can also see here a lot more outliers than my team in both the higher and lower ends of the scorring. This type of data can be used in other areas as such as consistancy, or reliability in those years.

## Data Visualization: Comparing the Two Teams

**![Chart, histogram

Description automatically generated]()**

In this project I was asked between two different charts to represent the points scored in a distribution; They were a boxplot and a histogram. From the chart above you see that I chose the histogram, and for a few reasons. First off, if you were to look at the boxplot in the comparison, the two teams looked very very closely the same. Next, I noticed the way they were laid out was a bit different as well. The boxplot had ‘Y’ assigned with the points and ‘X’ just being the two teams. This visualization wasn’t compelling when trying to see the difference in the two. The histogram using a translucent layover of the two teams was much easier, and appealing, to look at for the data.

When looking at this distrubution on the histogram chart, you can easily see that the assigned team in regards to the team that I chose has less games with scores between 85 and 100 points, but more games with scores between 100 and 130 points. This would most likely signify that the assigned team in their years had more overall points and higher team score than the team that I had picked.

## Descriptive Statistics: Relative score of Your Team.

| **Statistic** | **Value** |
| --- | --- |
| Mean | 1617.48 |
| Median | 1652.57 |
| Variance | 9987.20 |
| Standard Deviation | 99.94 |

Central tendency measurements help you to find the middle, or the average, of a dataset. By using the mode (the most frequent value), the median (the middle number in an ordered dataset), the mean (the sum of all values divided by the total number of values), the variance (the average of the squared differences of a random variable from its mean), and the standard deviation (the average amount of variability in your dataset), you can visually see the distribution of the dataset in different ways.

For a normal distribution the data is symmetrically distributed with no skew, while a distribution positively will be to the left, and to the right negatively. Our skew, for the most part, is bell-shaped and leaning more towards a normal distribution.

## Descriptive Statistics: Relative Score of Assigned Team

| **Statistic** | **Value** |
| --- | --- |
| Mean | 1739.80 |
| Median | 1751.23 |
| Variance | 2651.55 |
| Standard Deviation | 51.49 |

Using these stats just in a comparison to my team, I can quickly see that mean and median here is much higher; Meaning the middle and the average of all the scores is better than my team’s is. There’s also a much lower variance and standard deviation indicating a much closer central tendency to the mean.

## Confidence Intervals for the Average Relative Skill of All Teams in Your Team’s Years

*In the Python script, you calculated a 95% confidence interval for the average relative skill of all teams in the league during the years of your team. Additionally, you calculated the probability that a given team in the league has a relative skill level less than that of the team that you picked.*

*See Step 8 in the Python script to address the following items:*

* Report the confidence interval in a formatted table as shown below.

Table 4. Confidence Interval for Average Relative Skill of Teams in Your Team’s Years

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

Confidence levels, in layman’s terms, are just the percentage of confidence someone has in their estimate. Thus, in turn, resulting in how much caution is required when using the estimate. When using them in estimating the measures for a central tendency for a population, the government, according to The Census Bureau, routinely employs 90% confidence intervals.

For my team, I used a 95% confidence interval meaning the estimated average relative score of my team is about 1502.02 to 1507.18. This means I am 95% sure in the samplings of this to be correct.

By changing the confidence levels, the outcome would be different as well. For example, if you were to change the confidence level to 90%, the interval would be (1502.44, 1506.77) and on the other end with the level being 99%, the interval would be (1501.21, 1507.99). Looking at this, we could assume that the lower the confidence level, the tighter the estimation can be and vice versa when the confidence level is higher.

## Confidence Intervals for the Average Relative Skill of All Teams in the Assigned Team’s Years

Table 5. Confidence Interval for Average Relative Skill of Teams in Assigned Team’s Years

| **Confidence Level (%)** | **Confidence Interval** |
| --- | --- |
| 95 | (1487.66, 1493.65) |

Here we can see the assigned team’s confidence interval as (1487.66, 1493.65). While this is lower than the team I chose, that doesn’t mean that the teams actual skill level is lower. The estimated skill level for the team can still be higher than our simply because were are comparing different years.

Again, by changing the confidence levels here to 90% we get an interval of (1488.14, 1493.17) and an interval of (1486.72, 1494.59) at 95%. This is showing again that at a lower level the interval gets tighter and looser at higher levels.

## Conclusion

In conclusion here, these types of analysis reports are important in a few ways. Like I had said earlier in the report, looking at just scores for the teams and seeing their averages or medians, this doesn’t always provide a clear image of the skill of the two in comparison. By using data visualization in reference to each team in their specific time periods, we can think of it as looking at inflation. Just because something actually cost less 10 years ago, does that also mean it is more expensive nowadays if we take inflation into account.

The results in this case show that the actual estimated skill level of the Chicago bulls from 1996 to 1997 is indeed better than the estimated skill level of the Miami Heat from 2013 to 2015.

.

## Citations

*What Is Data Visualization? Definition, Examples, And Learning Resources*. (n.d.). Tableau. <https://www.tableau.com/learn/articles/data-visualization>

Bhandari, P. (2022, June 9). *Central Tendency | Understanding the Mean, Median & Mode*. Scribbr. <https://www.scribbr.com/statistics/central-tendency/>

Math.net. (n.d.). *Variance*. https://www.math.net/variance

Bhandari, P. (2022a, May 25). *How to Calculate Standard Deviation (Guide) | Formulas & Examples*. Scribbr. <https://www.scribbr.com/statistics/standard-deviation/>

US Census Bureau. (2021, October 8). *A Basic Explanation of Confidence Intervals*. Census.Gov. https://www.census.gov/programs-surveys/saipe/guidance/confidence-intervals.html