ATTENDANCE PREDICTION OF RED SOX HOME GAMES IN COMING YEARS

By:

Nikhil Dharane

IE 6200: Engineering Probability & Statistics

Guide: Prof. Rajesh Jugulum

Semester: Fall 2018

Northeastern

NORTHEASTERN UNIVERSITY
BOSTON

ABSTRACT

Our project aims at the prediction of attendance of the people attending Red Sox home games in upcoming years by identifying the factors on which the attendance depends and formulating a trend that can be used to predict the same in the future.

I got the data of wins, losses, position finished, attendance per game and overall attendance for the Red Sox game from the year 1998 to 2018 as a sample from the BaseballReference.com. A computational analysis such as mean, standard deviation and few distributions will be performed to predict the Red Sox's audience attendance per game over the next few years.

By using the various factors of each match and carrying out the statistical analysis, will be trying to observe and will conclude that how many people are likely to attend the match in coming years. It could be based on the number on matches Red Sox is winning, losing or the position they are holding. The attendance could be based on a single or multiple factor.

INTRODUCTION

The Boston Red Sox is a professional baseball team belonging to the Eastern Division of the American League of Major League Baseball. Its home field has been Fenway Park in Boston, Massachusetts since 1912. The Red Sox boast one of largest fan bases the League, but the smaller capacity of Fenway limits the number of fans who attend games in-person. Consequently, while The Red Sox average attendance percentage is among the league's highest, overall attendance is not among league leaders. Fenway has the longest consecutive sell-out record in the Major League, 794 sold out games from May 15, 2003 to April 10, 2013, including 820 playoff contests.

Our project mainly investigates the annual total audience attendance of the Red Sox from 1998 to 2018 at Fenway Park. Our project aims to predict attendance of Red Sox home games in upcoming years by collecting the previous years' data. The data population is the annual attendance rate of the Red Sox game hosted at Fenway Park. The sample is the annual attendance rate from 1998 to 2018. The annual survey data in the sample includes not only audience attendance but also the number of wins and losses of the Red Sox each year, the attendance per game, as well as the final ranking of the Red Sox in the League.

I performed the analysis by using Forecasting in Excel and hypothesis in Minitab. Calculating the confidence interval of the average attendance in the previous year, I will conclude the range of attendance per year for 2019, and so on. Then state the hypotheses to decide if there should be an exact amount of attendance in the upcoming years and find out if I can reject the null hypothesis. Finally, I can get a conclusion about the approximation of audience in 2019.

I used Forecasting part of Excel to estimate the attendance of the Red Sox in upcoming years after 2018. I also used Minitab and Tableau to predict the confidence interval of the Red Sox's audience attendance at Fenway Park in 2019. Since I used Excel to draw the conclusion that audience attendance will increase in the next year, I set a slightly higher benchmark than 2018, and then used Decision Rule to determine whether the setting is reasonable. After several calculations, a more accurate prediction value is finally obtained. Along with that I have also used SPSS to analyze whether attendance is related to win, lose and position of Red Sox in every game.

CALCULATION AND ANALYSIS

Prediction of attendance per game of Boston Red Sox in the 2019

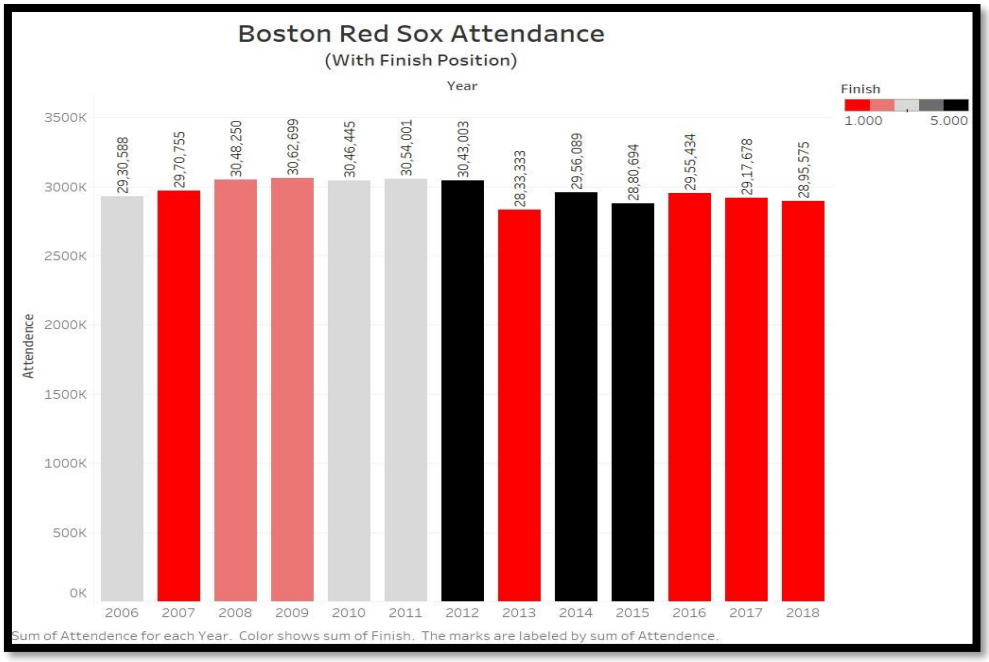

Figure 2.1 Boston Red Sox Attendance with finish position

Although we can see an increase of 2.6 percent in the 2015-2016 attendance, it slipped from 7th to 8th in attendance among all MLB ballparks. So, we can see that whenever the Boston Red Sox team secures a position less than the 3rd position, its attendance decreases in the coming year. Also, when the team secures 1st or 2nd position, the attendance increases in the coming year. We can see this trend from the graph above. The rise in attendance from the year 2003 to year 2005 is because of the increase of the number of seats in the stadium (Fenway Park). There are several other factors affecting this change in behaviour like marketing strategies, renovation of the stadium, etc, which are beyond the scope of this project.

Since Boston Red Sox secures 1st position for three years consecutively, we expect the attendance of the year 2019 to increase. Taking this assumption in consideration, we are going to perform a hypothesis to predict the attendance of Boston Red Sox games for the year 2019.

The below graph shows the average attendance/game of Boston Red Sox.

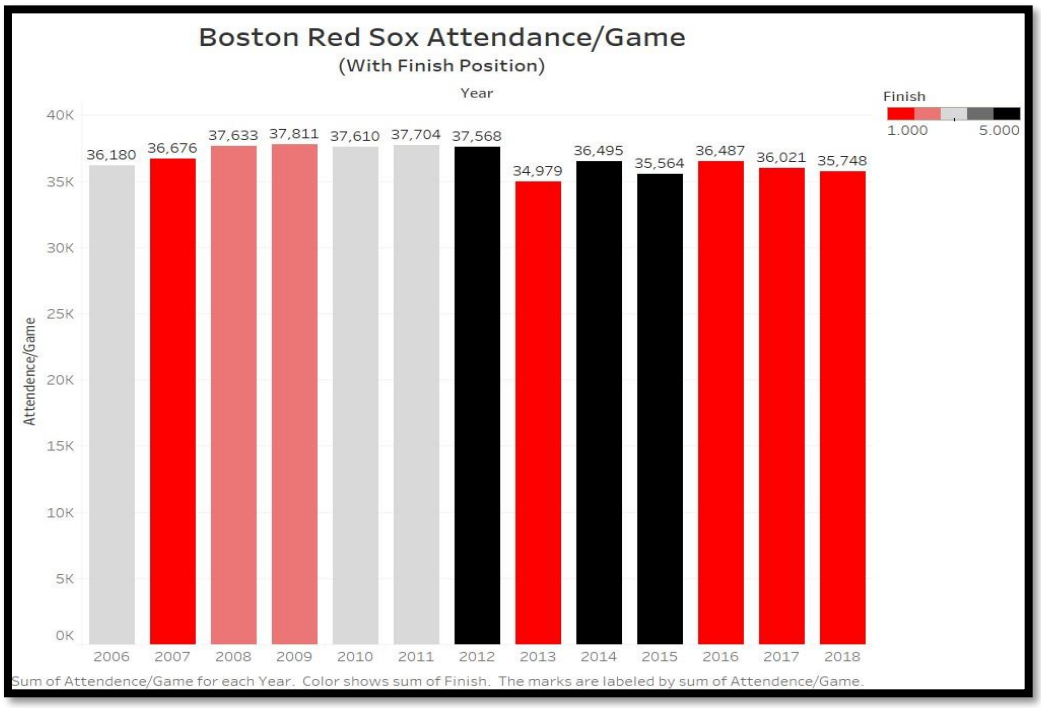

Figure 2.2 Boston Red Sox Attendance/Game with finish position

So, we are going to assume that the average attendance/game of Red Sox for the coming year (2019) will be greater than 35,800. Taking this value into consideration, we will perform hypothesis and find a range of values which will contain the attendance of year 2019.

Test	
Null hypothesis	H_0: $\mu <= 35800$
Alternative hypothesis	H_1: $\mu > 35800$

T-Value	P-Value
3.26	0.003

We get p-value (0.003) < α (0.05) ⟹ We reject the null hypothesis. Therefore, our attendance for next year is greater than 35,800.

Now, we will assume that our attendance for next year is greater than 36,200.

Test		
Null hypothesis	H_0: $\mu <= 36200$	
Alternative hypothesis	H_1: $\mu > 36200$	
T-Value	P-Value	
1.73	0.055	

We get p-value (0.055) > α (0.05) ⟶ We fail to reject the null hypothesis. Therefore, our attendance for next year is between 35,800 and 36200.

Now, we will assume that our attendance will be greater than 36,100.

Test		
Null hypothesis	H_0: $\mu <= 36100$	
Alternative hypothesis	H_1: $\mu > 36100$	
T-Value	P-Value	
2.11	0.028	

We get p-value (0.028) > α (0.05) ⟶ We reject the null hypothesis. Therefore, our attendance is between 36,100 and 36,200.

Now, we will assume that our attendance will be greater than 36,150.

Test		
Null hypothesis	H_0: $\mu <= 36150$	
Alternative hypothesis	H_1: $\mu > 36150$	
T-Value	P-Value	
1.92	0.039	

We get p-value (0.039) > α (0.05) ⟶ We reject the null hypothesis. Therefore, our attendance is between 36,150 and 36,200.

Now, we will assume that our attendance will be greater than 36,180.

Test	
Null hypothesis	H_0: $\mu <= 36180$
Alternative hypothesis	H_1: $\mu > 36180$
T-Value	P-Value
1.81	0.048

We get p-value (0.048) > α (0.05) ⟶ We reject the null hypothesis. Therefore, our attendance is between 36,180 and 36,200.

Through this, we conclude that the attendance of Boston Red Sox for the year 2019 will be between **36,180** and **36,200**.

Prediction of attendance using Forecast function in excel

Here we have used the Excel to forecast the attendance of the Red Sox Home games, we used the whole data set from 1998 to 2018 to predict the attendance for next 5 year; a longer duration can be considered as well. We considered a span of 5 years to show and analyse the importance of confidence Interval and how range changes when the intervals are changed.

Year	Attendence/Game	Forecast(Attendence/Game)	Lower Confidence Bound(Attendence/Game)	Upper Confidence Bound(Attendence/Game)
2018	35,748	35,748	35,748	35,748
2019		36,052	34,041	38,062
2020		36,355	33,650	39,061
2021		36,659	33,402	39,916
2022		36,963	33,234	40,691
2023		37,266	33,118	41,414

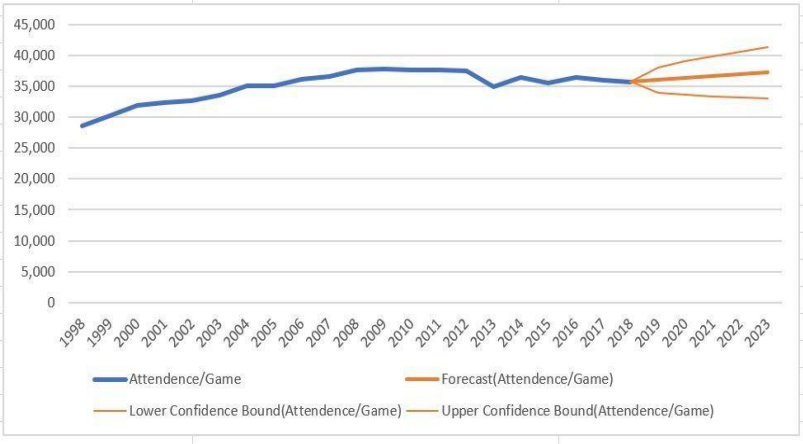

Figure 3.1 Forecast with 90% C.I.

Year	Attendence/Game	Forecast(Attendence/Game)	Lower Confidence Bound(Attendence/Game)	Upper Confidence Bound(Attendence/Game)
2018	35,748	35,748	35,748	35,748
2019		36,052	31,297	40,806
2020		36,355	29,955	42,755
2021		36,659	28,955	44,363
2022		36,963	28,143	45,782
2023		37,266	27,455	47,077
2024		37,570	26,856	48,283

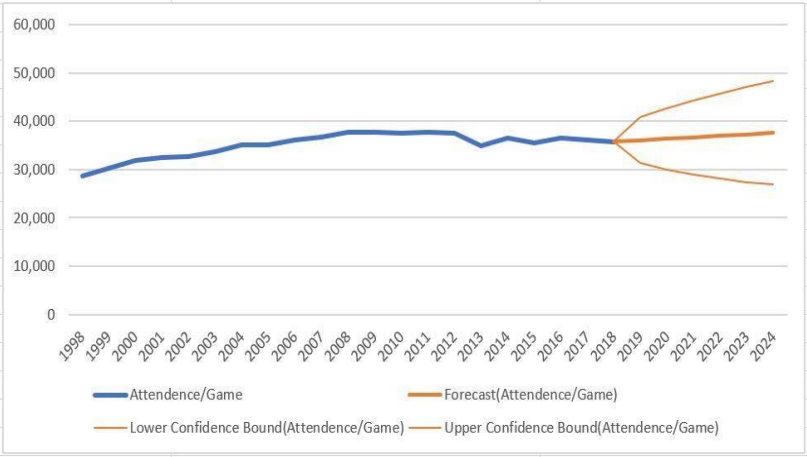

Figure 3.2 Forecast with 99.99% C.I.

The forecast sheet function helps us select the range of data to be used for predication, it requires at least 2 set of data to predict the next data set, we have options to select the desired confidence interval, the number of years to be predicted, starting data set from where we want to consider the data and other advanced options like filling mission point and aggregating duplicates using various methods.

I did the analysis by considering few confidence intervals as we observe the predicted overall attendance we can see that it is constantly increasing in each year this could be because of various reasons one of the reasons could be because of the three consecutive first position secured.

As we know in theory that as the confidence interval increases the range i.e. lower and upper bound increases as well which can be clearly observed from the above data. For example, when 90% confidence interval is considered we have the Lower bound as 34,041 and Upper bound as 38,062 giving us a range of 4,021 but when a 99.99% of confidence interval is considered we have the Lower bound as 31,297 and Upper bound as 40,806 giving us a range of 9,509. It can be clearly observed that the range has increased from 90% CI to 99.99% CI.

CONCLUSION

My aim was to predict the attendance of the Boston Red Sox home games for the upcoming years. From our findings, we drew a pattern from which we can derive the attendance for upcoming years. Neglecting the variables such as marketing strategies, commercials, etc., we forecasted the attendance of the year 2019 to be between 36,180 to 36,200 by using Test of Hypotheses. I used IBM's SPSS (Statistical Product and Service Solutions) to find the correlation between Wins, Loses, Finish Position and Attendance. Also, the pattern of increase in attendance was observed in the graphs created in Tableau. Lastly, I used Microsoft Excel forecast sheet function to predict the attendance for upcoming years and the values for same matched our hypotheses.

REFERENCES

Data Set:

The data set for our project was collected from the following websites.

https://www.baseball-reference.com/teams/BOS/attend.shtml

http://www.baseball-almanac.com/teams/rsoxatte.shtml

Book:

Probability & Statistics for Engineers & Scientists – Ronald Walpole, Raymond Myers, Sharon Myers, Keying Ye – Ninth Edition