MLB 2018: Analysis of Batting Performance Report

Student id: 11647793

Overview:

Significant variations in batting performance were found between the outfield, infield, and catcher positions based on an ANOVA analysis of 2018 Major League Baseball player data, with a p-value of 0.048, rejecting the hypothesis that their batting performance is equal. Null hypothesis means, all the positions had the similar batting performance and Alternative hypothesis means, there are significant variations in batting performance of all the positions. This study aimed to reject the null hypothesis and to support alternative hypothesis suggesting differences in the batting performance.

Methodology:

1. Categorization of Positions: The positions of players are categorized into 3 groups: Infield (1B, 2B, 3B, SS), Outfield (LF, CF, RF), Catcher (C) and all other positions are excluded from analysis.**2. Identification of Outliers:** Using 1.5IQR rule, Outliers are identified within each category. Sample data points that lie outside of the upper boundary and lower boundary are considered as Outliers.

Below are the sample size, quartiles, and boundaries for each category:

The calculations included quartile values (Q1, Q3), interquartile range (IQR), and upper and lower boundaries for each position. **N**= Number of Observations, **Q1** = Quartile(range,1), **Q3**= Quartile(rangle,3), **IQR** = Q3-Q1, **Upper Boundary** = Q3+1.5*IQR, **Lower Boundary** = Q1-1.5*IQR, **Number of outliers below the lower limit** = COUNTIF (range,"<"&Lower Boundary), **Number of outliers above the upper limit**= COUNTIF (range,">"&Upper Boundary)

Outfield	
N	225
Q1	0.276
Q3	0.34
IQR	0.064
Upper Boundary	0.436
Lower Boundary	0.18

Infield	
N	288
Q1	0.27275
Q3	0.3375
IQR	0.06475
Upper Boundary	0.434625
Lower Boundary	0.175625

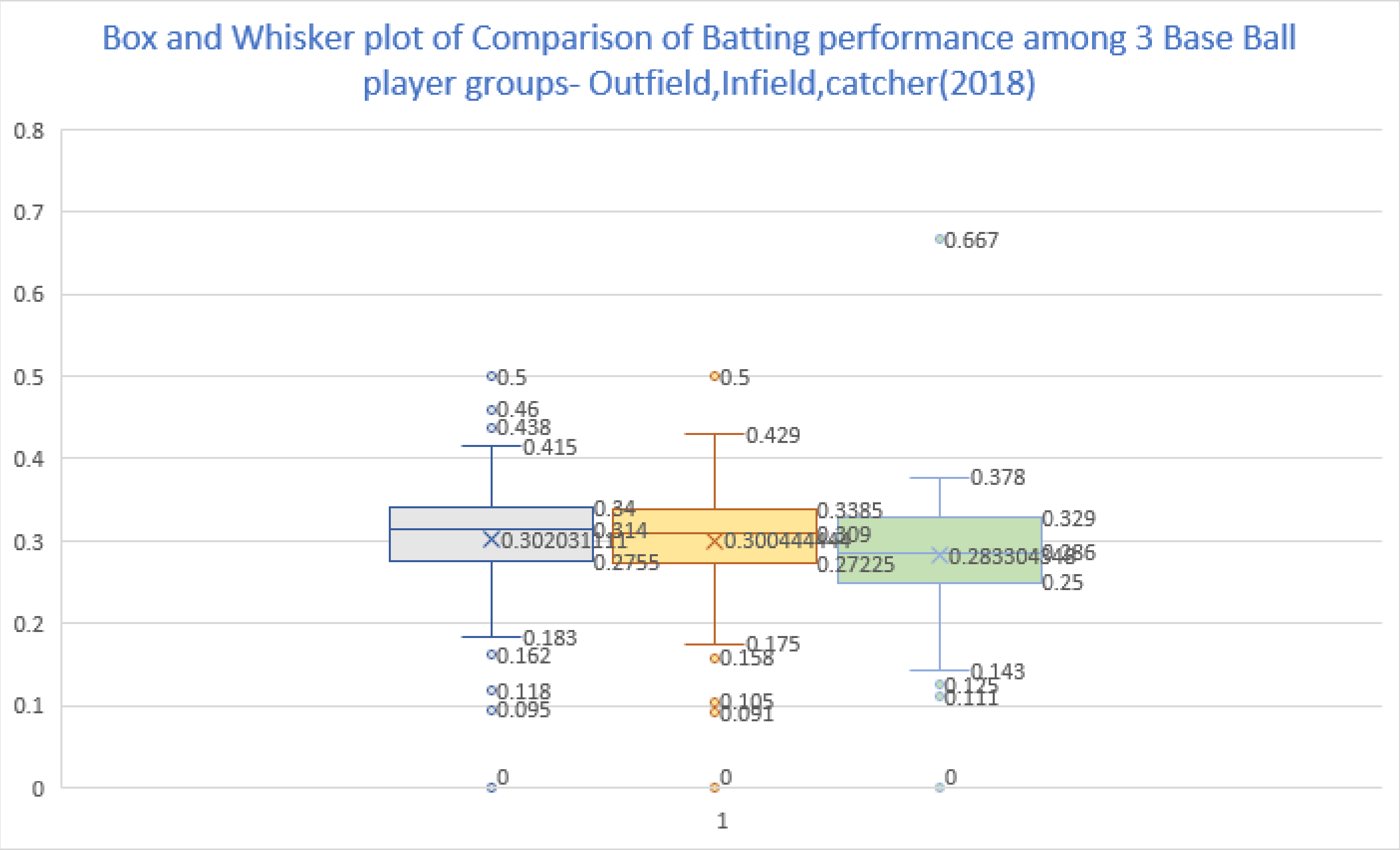
Catcher	
N	115
Q1	0.252
Q3	0.3285
IQR	0.0765
Upper Boundary	0.44325
Lower Boundary	0.13725

Below are the list of outliers from each category:

Outfield: 15 outliers (11 sample points below the lower limit and 4 sample points above the upper limit- 0,0,0.118,0.162,0.167,0.5,0,0.5,0.438,0,0.095,0.17,0.171,0.125,0.46) **, Infield**: 13 outliers (12 sample points below the lower limit and 1 sample point above the upper limit - 0.5,0.167,0.105,0,0.167,0.167,0,0.175,0.167,0,0.091,0,0.158)**, Catcher**: 6 outliers (4 sample

points below the lower limit and 2 sample points above the upper limit - 0.667,0.667,0.115,0.111,0,0.125).

Box and Whisker Plot- Comparison of Batting performance by player position:



Analysis:

The Baseball player league 2018 dataset consists of 1270 players, with 628 categorized as 3 groups (outfielders, Infielders or catchers) according to the positions specified. Infield (1B, 2B, 3B, SS), Outfield (LF, CF, RF), Catcher (C) and all other positions are excluded from analysis.

The variation of batting performance among these groups was measured by using ANOVA test based on On-Base Percentage (OBP) data provided in dataset. Below are the results of Anova: Single factor analysis.

Null Hypothesis: The null hypothesis (H0) states that the mean OBP does not significantly differ between the player positions.

Case:1 - ANOVA Result Without removing Outliers:

F-Statistic: 3.050, **P-value:** 0.048,

Degrees of Freedom: Between Groups = 2, Within Groups = 625

Anova: Single Factor

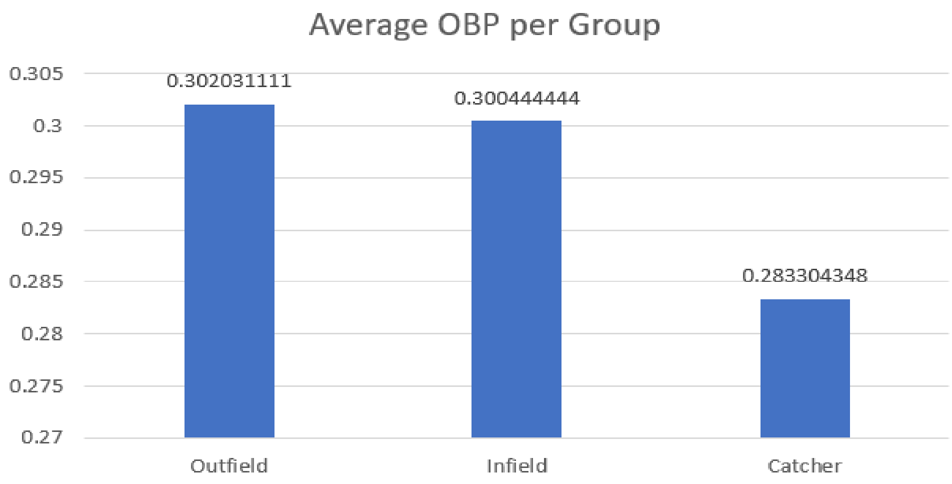
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Outfield	225	67.957	0.302031111	0.005224887
Infield	288	86.528	0.300444444	0.004078589
Catcher	115	32.58	0.283304348	0.00660795

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.030202821	2	0.01510141	3.05031058	0.048049684	3.010137326
Within Groups	3.094236241	625	0.004950778			
Total	3.124439062	627				

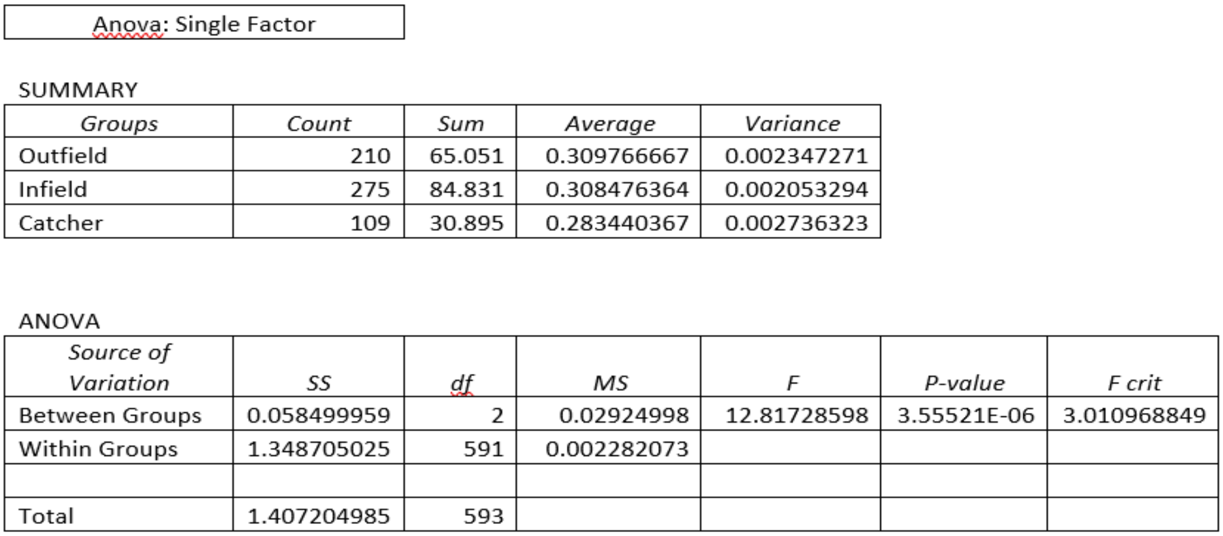
- From the above result it is observed that p-value (0.048) is less than significance level (0.05) which is the evidence to reject the null hypothesis(H0).
- With a total variance of 3.1244, with the observed between-groups variance was 0.0302, meaning that variations in positions might account for around 3% of the variance in batting performance.
- The variability among the group means is greater than what would be predicted by chance, according to the F-statistic of 3.050.



Case:2 - ANOVA Result After Removing Outliers

F-Statistic: 12.81729, **P-value:** 0.00000355521,

Degrees of Freedom: Between Groups = 2, Within Groups = 591



Anova: Single Factor

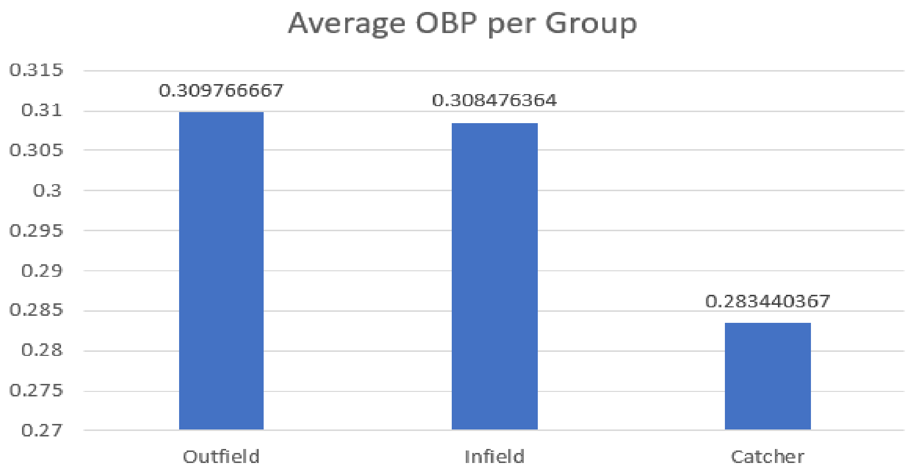
SUMMARY

Groups	Count	Sum	Average	Variance
Outfield	210	65.051	0.309766667	0.002347271
Infield	275	84.831	0.308476364	0.002053294
Catcher	109	30.895	0.283440367	0.002736323

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.058499959	2	0.02924998	12.81728598	3.55521E-06	3.010968849
Within Groups	1.348705025	591	0.002282073			
Total	1.407204985	593				

"After analyzing the ANOVA summary, it's clear that the p-value (0.00000356) surpasses the conventional significance threshold of 0.05. This indicates a strong indication to reject the null hypothesis (H0) and suggests that there are significant differences among the group means. The dataset's total sum of squares (SS) is 1.407205, with the between-groups sum of squares being 0.0585. This suggests that the three distinct groups- Outfield, Infield, and Catcher- account for approximately 4.16% of the total variance in the data. The F-statistic of 12.81729 further supports this conclusion, revealing that the variability among the group means is more significant than what would be expected due to random chance alone.



Therefore, from both the cases 1&2, we can conclude that there is statistically significant difference in mean of batting performance among three groups outfielders, Infielders, Catchers in Baseball player league during the analyzed period.

Name: Archana Goli
Student Id: 11647793

Mail Id: ArchanaGoli@my.unt.edu