UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECI 2143-06
PROBABILITY & STATISTICAL DATA ANALYSIS

Project 2 Report

Title : Case Study On The Factors That Affects Baseball Attendance

Prepared for : Dr. Chan Weng Howe

Prepared by : Amir Hakim bin Ahmed Mahir

Matric Number : A19EC0018

TABLE OF CONTENT

INTRODUCTION

a. Background of study

Major League Baseball

The Major League Baseball (MLB) is a North American professional baseball organization that was founded in 1903 by merging two United States professional baseball League, which were the National League (NL) and the American League (AL). in 1876, the National League was formed while the American League was established a little late in 1901. At first, the NL and AL acted as independent organizations and both of the leagues were involved in what was known as "baseball war" until they were merged. The leagues formed a truce in 1903 that caused the World Series to be created, where the winners from each league were matched to determine the national champion. In 1997, MLB introduced inter-league play, which mean that each NL team played a series of regular season games against AL teams with the same division. However, there is one major difference between the two leagues, which AL has adopted the designated hitter rule start form 1973 and has caused teams form the AL to score more runs than teams from NL.

Designated Hitter Rule

The rule is basically allowing teams to use another player to bat in place of the pitcher. The designated hitter does not take the field on defense as the pitcher is still part of the team's nine defensive players. The designated hitter must be selected before the game and he must come to bat once unless the opposing team changes their pitcher prior to that point. When a team chooses to not pick a designated hitter before the game, the team is forbidden to use a designated hitter for the entire match.

b. Objectives of study.

This study is held to find out whether the run scores in each match by the teams are related to their supporters' attendance. This is because baseball team owners believe that the designated hitter rule means more runs scored, which in turn means higher attendance. In addition to that, to find out that does the attendance affected the number of matches won by the team. Other than that, to determine whether is there any relationship between the base hit and attendance. Next, this study was held to find any proof that the mean runs scored on 2016 season is different from the mean runs scored of previous seasons.

c. Methodology.

The type of data used for this study is secondary data. The dataset for this study was acquired from two reliable website for sport's data, especially related to baseball in this case, which were the ESPN website and the Baseball Reference website. The purpose of both websites collecting statistical data for MLB is mainly to provide baseball fans with all stats related to both baseball players and teams throughout each season since the MLB was formed and also to determine the best team, manager and players for the season.

The method of this study is quantitative as this type of research allow me to analyses the data from the dataset that has been chosen. I have chosen four different types of statistical

analysis for this study which were 1 sample testing, correlation, regression and Analysis of Variance (ANOVA). From all of the variables provided in the dataset, only a few variables were chosen. The Table 1 below shows the different types of analysis and the variables for each of them.

Objective	Statistical Analysis	Variables Chosen
To find any proof that the mean runs scored on 2016 season is different from the mean runs scored of previous seasons.	1 sample hypothesis testing	Runs Made (Ratio)
To find out whether the run scores in each match by the teams are related to their supporters' attendance	Correlation	Runs Per Game (Ratio) Overall Average Attendance (Ratio)
To find out that does the attendance affected the number of matches won by the team.	Correlation	Overall Average Attendance (Ratio) Matches Won (Ratio)
To determine whether is there any relationship between the base hit and attendance	Regression	Base Hit (Ratio) Overall Average Attendance (Ratio)
To determine whether all the subcategories of Runs Made has the same mean	ANOVA	Runs Made (Ratio)

Table 1

All of the statistical analysis above were made inside the RStudio by importing the original dataset in the form of .xlsx file prior to the coding in RStudio. The significance level used for all the statistical analysis are the same, which was $\alpha = 0.05$.

DATA ANALSYSIS

a. 1 SAMPLE HYPOTHESIS TEST

The variable that has been chosen for this 1 sample test is Runs Made. Based on the website, https://www.baseball-reference.com/leagues/MLB/bat.shtml#all_teams_standard_batting_totals, I decided to take mean Runs Scored for each team from 2000 to 2016 as the population mean, which is $\mu = 738.67$. The 1 sample test was used to determine whether there is evidence to prove that the mean Runs Scored in 2016 exceeds 738.67 at 0.05 significance level.

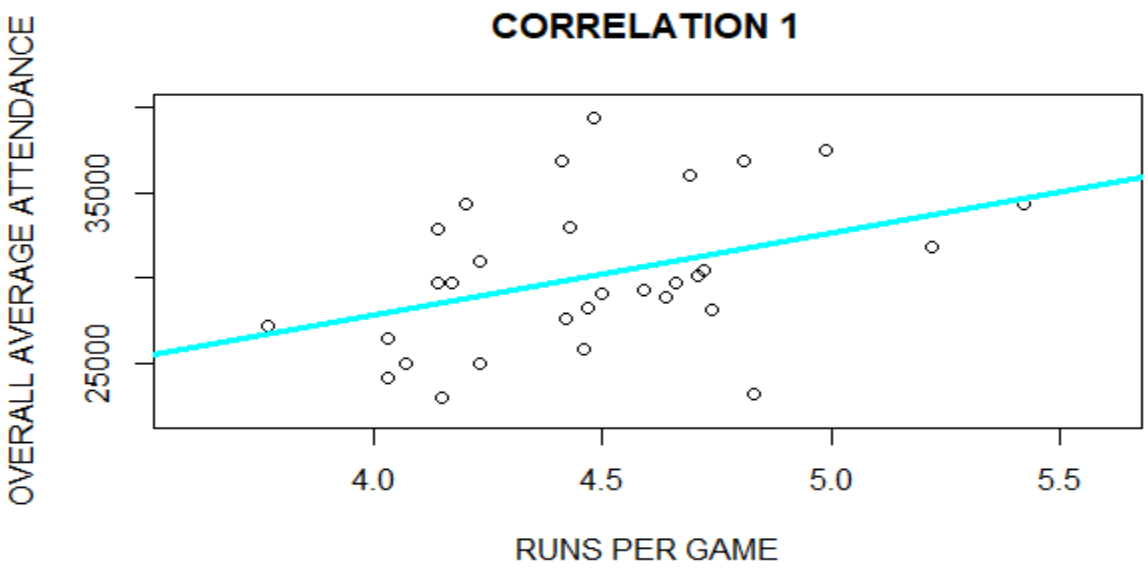
H0: $\mu = 738.67$

H1: $\mu \neq 738.67$

```
> alpha = 0.05
> t.test(x, mu=738.6706, alternative = "two.sided", conf.level = 0.95)

        One Sample t-test

data:  x
t = -1.2708, df = 29, p-value = 0.2139
alternative hypothesis: true mean is not equal to 738.6706
95 percent confidence interval:
 702.4768 747.1232
sample estimates:
mean of x
    724.8

> abs(qt(0.025,29))
[1] 2.04523
```

Based on the results of R-coding, the value of test statistics that I managed to get is t = -1.2708 and the p value = .2139. When compared the test statistics with the critical region, t = 2.0452 and t = -2.0452 at 29 degree of freedom, it is clear that test statistics does not fall within the critical region. The p-value is also not statistically significant as it is larger than $\alpha = 0.05$. So, the null hypothesis failed to be rejected as there is enough evidence to show that the mean Runs Scored in 2016 does not differ average runs made from 2000 to 2016.

b. <u>CORRELATION</u>



H0 : No linear correlation between Runs Made and Overall Average Attendance

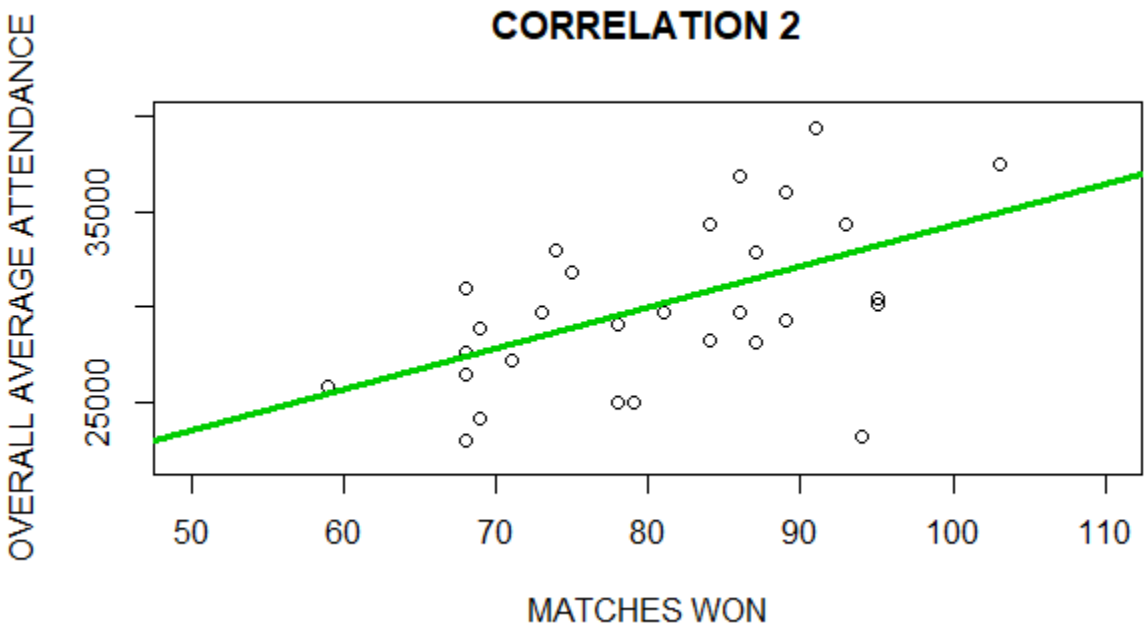
H1 : Linear correlation exists between Runs Made and Overall Average Attendance

```
> #hypothesis testing for correlation
> cor.test(x,y,conf.level = .95)

        Pearson's product-moment correlation

data:  x and y
t = 2.3236, df = 28, p-value = 0.02763
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04887083 0.66587777
sample estimates:
      cor
0.4020611

> #two tailed test 0.5/2,28
> abs(qt(0.025,28))
[1] 2.048407
```

The variables that have been chosen for the first correlation analysis is Runs Per Game and Overall Average Attendance. For the significance test for correlation, based on R coding, the null hypothesis which is there is no linear correlation between the Overall Average Attendance and Run Per Game is rejected. This is due to the test statistics gained from the correlation analysis in

the R-coding, t = 2.3236 falls within the two tailed critical regions at 0.05 significance level with degree of freedom of 28, t = 2.0484. Not to mention, p-value = .02763 are also statistically significant as it is less than $\alpha = 0.05$. Thus, there is enough evidence to conclude that linear correlation exists between the Overall Average Attendance and Runs Per Game.

Next, a Pearson's correlation was computed to determine the relationship between those two variables as both of them are ratio data types. There was a positive correlation between the two variables with values of r = 0.4021, n = 30 and p = .0276. The scatterplot above shows the result of the correlation analysis. We can clearly see the straight line on the graph that indicates the positive correlation between the two variables. The strength of the correlation is fairly weak as the *r* value, which is 0.4021 falls within the weak region. We can also notice that there are a few dots that are far from the correlation line projected inside the graph. Hence, why the correlation's strength is weak. So, it is safe to say that there was a weak, positive correlation between Runs Per Game and Overall Average Attendance. The Overall Average Attendance increase slightly with the increment of Runs Per Game.



H0 : No linear correlation between Matches Won and Overall Average Attendance

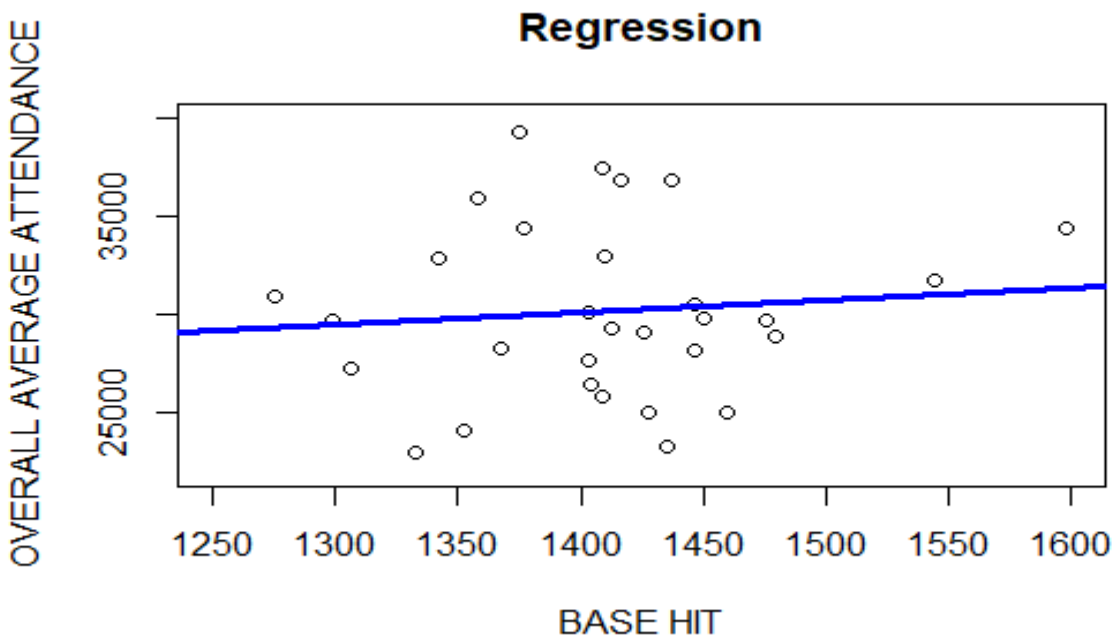
H1 : Linear correlation exists between Matches Won and Overall Average Attendance

```
        Pearson's product-moment correlation

data:  x and y
t = 3.2472, df = 28, p-value = 0.003021
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.200547 0.743246
sample estimates:
      cor
0.5230271

> #two tailed test 0.5/2,28
> abs(qt(0.025,28))
[1] 2.048407
```

So, for the second correlation analysis, Matches Won and Overall Average Attendance have been chosen, from the test of significance for correlation, I managed to get the test statistic value, t = 3.2472. When compared it to the two tailed critical value, which has 28 degree of freedom and at 0.05 significance level, t = 2.0484, the test statistic is greater than the critical value. The p = .0030 are also statistically significant as it is less than $\alpha = 0.05$. So, the null hypothesis, which there are no linear correlation occur between Matches Won and Overall Average Attendance is rejected as there is sufficient evidence to prove that linear correlation exists between the two variables that were being analyzed at 95% confidence interval.

Then, due to Matches Won and Overall Average Attendance both are ratio data types, Pearson's coefficient was used to determine the relationship between them. Positive correlation exists between Matches Won and Overall Average Attendance. From the R coding, I managed to get r = 0.5230, n = 30, and p = .0030. The scatterplot above was based on the results of the correlation analysis. We can see that the correlation is positive and based on the *r* value, which is 0.5230, the strength of the correlation is moderate because it is greater than 0.50 and less than 0.80 and also most of the dots were plotted very closed to the correlation line of best fit. Overall, there is a moderate, positive correlation exists between Matches Won and Overall Average Attendance meaning that the more the matches won by a team, the more the overall average attendance for the team.

c. **REGRESSION**



H0: β = 0 (no linear relationship)

H1: β ≠ 0 (linear relationship does exist)

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-7074.8 -2444.0  -740.6  2990.2  9385.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21465.873  17243.777   1.245    0.224
x               6.173     12.223   0.505    0.618

Residual standard error: 4463 on 28 degrees of freedom
Multiple R-squared:  0.009026,  Adjusted R-squared:  -0.02637
F-statistic: 0.255 on 1 and 28 DF,  p-value: 0.6175

> #two tailed test 0.5/2,28
> abs(qt(0.025,28))
[1] 2.048407
```

For the regression analysis, the variables that have been chosen were Base Hit for independent variable and Overall Average Attendance for the dependent variable. A simple linear regression method was used to calculated the Overall Average Attendance based on Base Hit. For the t-test for the population slope, based on R coding, the value of test statistic computed was 0.505 while the critical value for 28 degree of freedom and at 0.05 significance level, t=2.048. So, it is clear that the test statistic does not fall within any of the critical region which are t<-2.048 and t>2.048. The p-value = .618 and it is greater than 0.05 significance level making the result is not statistically significant. Thus, we failed to reject the null hypothesis as there is sufficient evidence that Base Hit does not influence the Overall Average Attendance at 0.05 significance level.

Next, based on the results on r-coding significant regression equation was found based on the R coding which is ($F(1,28) = 0.255$, $p = .618$), and the coefficient of determination, R^2 is 0.009026. Based on the regression graph, we can clearly see that the relationship between Base Hit and Overall Average Attendance is very weak and also very close to be interpret as no relationship occur between them. This is because value of R squared, which is 0.009026 is very nearly to the value of 0. The Overall Average Attendance that has been predicted, $\hat{y} = 21465.873 + 6.173$(Base Hit). The Overall Average Attendance increase 6.173 for each of Base Hit point scored.

There might be a Type II error occurred in this analysis as there is evidence that show that there is some kind of relationship between Base Hit and Overall Average Attendance.

d. ANOVA

For the one-way analysis of variance, I have chosen one variable only with ratio data type. This is due to the fact that the dataset that I used lack of nominal data type. the variable that have been chosen was Runs Made and I have split the variable into three separate categories, which are ≤ 700, 700-750 and >750 just like the table shown below.

≤ 700	**700-750**	**>750**
653	715	752
655	716	759
671	717	763
671	722	765
672	724	768
675	725	777
680	729	779
686	744	808
686	750	845

Notice that there are 9 observation for each category making the total observations is 27, when originally there are 30. 3 observations had to be removed due to one of the categories has 9 only observation while the other two have 10 and 11 respectively. I had to make the observation equal for each category because the analysis of variance that I have carried out was one-way ANOVA with equal sample sizes.

H0: The mean of runs made is the same for all categories.

H1: At least one mean is different

```
            Df Sum Sq Mean Sq F value   Pr(>F)
ind          2  51956   25978   67.46 1.41e-10 ***
Residuals   24   9242     385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, from the results of R-coding for ANOVA, the value of *F* test statistics is 67.46 and when comparing it to the critical value of *F* at $\alpha = 0.05$ significance level from *F*-distribution table, $F = 3.40$, the test statistics clearly lies within the critical region. The p-value of this test is <.000 and also less than $\alpha = 0.05$ making it statistically significant. The null hypothesis is rejected as there is sufficient evidence to show that all of the different categories does not have the same mean of Runs Made. There was a significant effect of mean of runs scored at 0.05 significant level for the three categories [$F(2, 24) = 67.46$, $p < 0.000$].

DISCUSSION AND CONCLUSION

First and foremost, the 1 sample hypothesis testing that was carried out have proven to us that the average runs made for each team in the 2016 season does not differ the average runs made from the 2000 season up until 2016 season. From this, I can conclude that the different of year does not affect the runs made. This is due to each year there will always be players who perform very well to increase the number of runs made and there are also those who have a bad spell for the entire season causing the number of runs made to not pass through the overall runs scored from the previous seasons. This has made the average runs made to be maintained at a certain range. Hence why the average run made in 2016 are basically the same as the average runs made from 2000 to 2016 season.

For the correlation part, I have done two analysis so that more information can be acquired. The first correlation analysis is between Runs Per Game and Overall Average Attendance. The test statistics for this analysis falls within the critical region, which prove that there is a linear correlation exists between the two variables. The scatterplot, which was plotted for this analysis has showed that there is a positive relationship between Runs Per Game and Overall Average Attendance, and the result from *cor.test()* function in R-coding has proven that the relationship between the variables is fairly weak. So, it is safe to say that Runs Per Game do affects the Overall Average Attendance in a positive direction but has only a small increment. Moving on to the second correlation analysis, which is between Matches Won and Overall Average Attendance. From the analysis, there has been a proof to show that linear correlation exists between the two variables and it is moderately positive. Overall, I conclude that Runs Per Game, Overall Average Attendance and Matches Won are all related to each other and moving together in the positive direction.

Next, the regression analysis was performed to find the linear relationship between Base Hit (independent variable) and Overall Average Attendance (dependent variable). Results of the R-coding, the test statistics for the t-test of population slope does not falls within the critical region which shows that the Base Hit does not has any effects to the change of Overall Average Attendance. The value of r squared gained from the analysis is very close to zero. From the regression line computed on the scatterplot, I can barely see the increment throughout the graph but because of I assumed that there is Type II error in here, I concluded that basically there is very weak relationship between Base Hit and Overall Average Attendance.

Lastly, the ANOVA test was done on only one variable only which is Runs Made. This is due to there is not much nominal variables in the dataset that I have chosen. I have separated the Runs Made variable into three different range and the result have shown that those three different range have different mean from each other. This is clearly different from the start as the I categorized the observations based on range. So, for example based on the analysis result, the first category (<=700) has the mean of 672.11 while the second category (>700 & <=750) has 726.89 and the third category (>750) has 779.56. From here, I already can conclude that the mean is different for each of the categories due to categorizing them based on the range of its value.

Overall, I can conclude that runs made per game do impact the attendance of baseball fans at the stadium whether home or away attendance while also affects the number of matches won. I also can conclude that the base hit scores cause very little changes on the overall attendance unlike runs scored, which has impact on it. In addition to that, I also found out that the mean runs scored on 2016 season is the same as the mean runs scored of previous seasons.

REFERENCES

- 2016 Major League Baseball Season Summary. (n.d.). Retrieved from http://www.baseball-reference.com/leagues/MLB/2016.shtml
- Augustyn, A. (2020, April 10). Major League Baseball. Retrieved from https://www.britannica.com/topic/Major-League-Baseball
- Major League Baseball Batting Year-by-Year Averages. (n.d.). Retrieved from https://www.baseball-reference.com/leagues/MLB/bat.shtml
- MLB Attendance Report - 2016. (n.d.). Retrieved from http://www.espn.com/mlb/attendance/_/year/2016
- National League of baseball is founded. (2009, November 16). Retrieved from https://www.history.com/this-day-in-history/national-league-of-baseball-is-founded
- What is a Designated Hitter Rule?: Glossary. (n.d.). Retrieved June from http://m.mlb.com/glossary/rules/designated-hitter-rule