Math 3007 (Spring 2021) Project Paper

Sean Hannaford, Navpreet Kaur, Danielle Moynihan

Title: #baseballgate: Did the MLB Change the Regulation Baseballs?

Abstract: Our project tests a popular baseball rumor: did the MLB secretly change the design of their baseballs midway through the 2015 season? This rumor spread as a way of explaining the sudden rise of home runs since these alleged new balls made pitching harder and hitting easier. The theory has been additionally bolstered with the confirmed news that the MLB plans to deaden the ball, likely as a result of the new and easier balls. Our hypothesis states that the balls were changed. In order to test this hypothesis, we analyzed baseball statistical data from all MLB teams before and after 2015, during which season the balls were supposedly switched, particularly emphasizing average home runs and strikeout statistics. Focusing particularly on the 2014 and 2016 seasons, we have consulted multiple tests to test the data and find that there is a distinct change between the two years, lending evidence to the rumor that the balls were switched midway through the 2015 season.

I. MOTIVATION

Midway through the 2015 season, the number of home runs increased noticeably between the first and second half of the season. Rumors began to circulate that the MLB changed the regulation balls for balls that made scoring easier for hitters. Since then, the amount of home runs has increased, and the MLB has stated they intend to make the balls harder to hit. For this reason, we decided to investigate for ourselves if, statistically, the balls are now easier to hit.

II. INTRODUCTION

The rumor of the MLB changing the design of their baseballs midway through the 2015 season attempted to explain why hitters seemed to have an easier time hitting home runs. These alleged new balls supposedly made pitching harder and hitting easier, which explained the sudden change in games. When the MLB recently announced their plans to deaden the ball, thus making it harder to hit home runs, theoretically switching the balls back, the changed ball rumor gained even more credibility. With this in mind, we decided to test the hypothesis that the balls were changed midway through the 2015 season. The best way to test this hypothesis would be to analyze data from the seasons before and after the 2015 season, since 2015 data could be unreliable if our hypothesis proved correct and the balls changed.

III. DATA COLLECTION

Luckily, baseball statistics are very popular and we did not have to collect the data of multiple seasons of baseball ourselves. We found our data from an MLB statistics website, baseball-statistics, which gave us every pitching and hitting statistic from the league across every season we could possibly want to analyze. The project data focuses particularly on average home

runs and strikeout statistics from the 2014 and 2016 seasons, to avoid any skewed data from potential a change in balls during the 2015 season.

In order to compare data further to strengthen our results, we tested 2014 and 2016 season data from the Nippon Professional Baseball League (NPB) league, using the website npb.jp. This website gave us a little more work in data collection, since it didn't provide a total count of home runs in the league in each season. Therefore, we had to count how many home runs each player on a team hit and then sum the runs to get the total number of home runs for each team, then again for the entire league.

IV. TESTING THE ASSUMPTIONS

When trying to determine which tests to use and whether the tests would be applicable, it was important to check to see if the normality and independence assumptions made for the tests, such as the two-sample procedure, are reasonable assumptions.

A. Checking for Normality

The first thing we wanted to do was check for normality so that we would be able to apply the two-sample procedure and regression model. Normality was tested using the chi-square goodness-of-fit test. Since we used a variety of data, we tested normality for home runs, strikeouts, batting averages, and on-base percentages for the years 2008-2019 at the $\alpha = 0.05$ level of significance. For all of our tests, we use the following null and alternative hypothesis: H_0:the data follows a normal distribution vs. H_A:the data does not follow a normal distribution. "The data" for all the tests depended on the year and which category we were testing for. For all tests, we used the sample mean,$\bar{x}$, as an estimate for μ and the sample standard deviation, s, as an

estimate for σ.Due to these 2 parameters, 2 degrees of freedom were subtracted. Then, for each model, assuming that H_0 is true, the appropriate computations were done.

There were a total of 12 tests done for each of the categories, so a total of 48 tests of normality. The data can be divided by the type of data we encountered as the following: symmetrical data, skewed data, data that does not work with our normal condition that the expected count is greater than or equal to 5, and rejection data. We will show examples of all 4 of these cases. However, it is important to note that for the case in which the data did not work under the normal condition that the expected count is greater than or equal to 5, a similar condition of at least 75% of the data is at least 5 was used so that the chi-square goodness of fit test was applicable.

I. For the data that was symmetrical, the chi-square goodness-of-fit test worked very well. All the estimated expected values were shown to be at least 5, so the test was applicable and we failed to reject the null hypothesis. The data for this is from the 2016 Home Runs dataset. The intervals were set up without looking at the data, and then we calculated the estimated probability and estimated expected count. The test was done as the following:

Assuming $H_0$is true, with $\bar{x}$ = 187 (estimate for μ) and $s = 31.9364$(estimate for σ), we calculated the probability and expected count for each interval.

$$(p_1)_e = P(N(187, (31.9364)^2) \leq 100) = P(Z \leq 100) \approx 0.0094$$

$$\rightarrow n(p_1)_e = 30(0.0094) = 0.282$$

$$(p_2)_e = P(100 \leq N(187, (31.9364)^2) \leq 120) = P(100 \leq Z \leq 120) \approx 0.0265$$

$$\rightarrow n(p_2)_e = 30(0.0094) = 0.795$$

$$(p_3)_e = P(120 \le N(187, (31.9364)^2) \le 140) = P(120 \le Z \le 140) \approx 0.0697$$

$$\rightarrow n(p_3)_e = 30(0.0697) = 2.091$$

$$(p_4)_e = P(140 \le N(187, (31.9364)^2) \le 160) = P(140 \le Z \le 160) \approx 0.1364$$

$$\rightarrow n(p_4)_e = 30(0.1364) = 4.092$$

$$(p_5)_e = P(160 \le N(187, (31.9364)^2) \le 180) = P(160 \le Z \le 180) \approx 0.1984$$

$$\rightarrow n(p_5)_e = 30(0.1984) = 4.952$$

$$(p_6)_e = P(180 \le N(187, (31.9364)^2) \le 200) = P(180 \le Z \le 200) \approx 0.215$$

$$\rightarrow n(p_6)_e = 30(0.215) = 6.45$$

$$(p_7)_e = P(200 \le N(187, (31.9364)^2) \le 220) = P(200 \le Z \le 220) \approx 0.1735$$

$$\rightarrow n(p_7)_e = 30(0.1735) = 5.205$$

$$(p_8)_e = P(220 \le N(187, (31.9364)^2) \le 240) = P(220 \le Z \le 240) \approx 0.1043$$

$$\rightarrow n(p_8)_e = 30(0.1043) = 3.129$$

$$(p_9)_e = P(240 \le N(187, (31.9364)^2)) = P(240 \le Z) \approx 0.0668$$

$$\rightarrow n(p_9)_e = 30(0.0668) = 2.004$$

All the expected counts are not at least 5, so the test is not applicable. Thus, we have to modify it as done in the second table shown below. Now, the test is applicable.

Homerun Class Intervals	Observed Count	Estimated Probability	Estimated Expected Count		Homerun Class Intervals	Observed Count	Estimated Probability	Estimated Expected Count
(0, 100]	0	0.0094	0.282	→	(0, 160]	7	0.242	7.26
(100, 120]	1	0.0265	0.795	**MODIFIED**	(160, 180]	5	0.1984	5.952
(120, 140]	3	0.0697	2.091	**VERSION**	(180, 200]	7	0.215	6.45

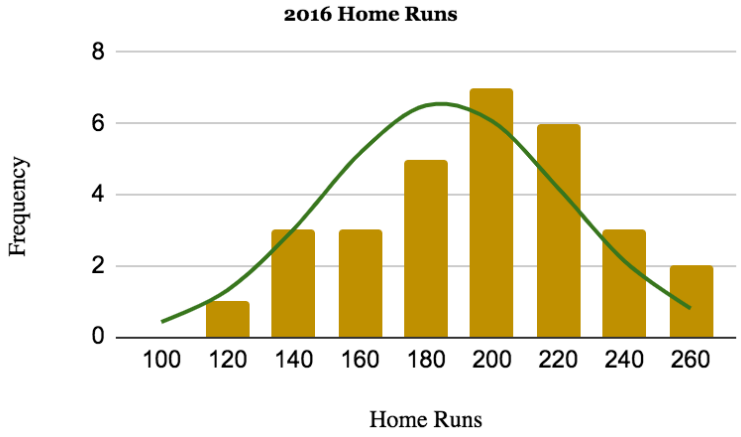
(140, 160]	3	0.1364	4.092		(200, 220]	6	0.1735	5.205
(160, 180]	5	0.1984	5.952		(220, ∞)	5	0.1711	5.133
(180, 200]	7	0.215	6.45			30	1	30
(200, 220]	6	0.1735	5.205					
(220, 240]	3	0.1043	3.129					
(240, ∞)	2	0.0668	2.004					
	30	1	30					

df = 5 - 1 - 2 = 2

$$d = \frac{(7-7.26)^2}{7.26} + \frac{(5-5.952)^2}{5.952} + \frac{(7-6.45)^2}{6.45} + \frac{(6-5.205)^2}{5.205} + \frac{(5-5.133)^2}{5.133} = 0.334$$

$$d^* = 5.9915 > d$$

Thus, we fail to reject the null hypothesis at the $\alpha = 0.05$ level of significance. So, we have convincing evidence that the home runs follow a normal distribution.



[Note: We can see from the histogram and normal approximation curve that the data fits the normal distribution curve pretty well.]

II. The next type of data we encountered was skewed data. For this type of data, when plotted on the histogram you could see that it was not okay to just assume normality and that a test was

needed. An example of this type of data was the 2013 Batting Averages. Our null hypothesis and alternative hypothesis are the same as above. Similar to the symmetric case, the intervals were chosen without looking at the data and then later modified. The test was done as the following: Assuming H_0 is true, with $\bar{x}$ = 0.2534 (estimate for μ) and $s = 0.0123$ (estimate for σ), we calculated the probability and expected count for each interval.

$$(p_1)_e = P(N(0.2534, (0.0123)^2) \leq 100) = P(Z \leq 0.24) \approx 0.1385$$

$$\rightarrow n(p_1)_e = 30(0.1385) = 4.1538$$

$$(p_2)_e = P(0.24 \leq N(0.2534, (0.0123)^2) \leq 0.245) = P(0.24 \leq Z \leq 0.245) \approx 0.1096$$

$$\rightarrow n(p_2)_e = 30(0.1096) = 3.2885$$

$$(p_3)_e = P(0.245 \leq N(0.2534, (0.0123)^2) \leq 0.25) = P(0.245 \leq Z \leq 0.25) \approx 0.1440$$

$$\rightarrow n(p_3)_e = 30(0.1440) = 4.3207$$

$$(p_4)_e = P(0.25 \leq N(0.2534, (0.0123)^2) \leq 0.255) = P(0.25 \leq Z \leq 0.255) \approx 0.1607$$

$$\rightarrow n(p_4)_e = 30(0.1607) = 4.8224$$

$$(p_5)_e = P(0.255 \leq N(0.2534, (0.0123)^2) \leq 0.26) = P(0.255 \leq Z \leq 0.26) \approx 0.1524$$

$$\rightarrow n(p_5)_e = 30(0.1524) = 4.5722$$

$$(p_6)_e = P(0.26 \leq N(0.2534, (0.0123)^2) \leq 0.265) = P(0.26 \leq Z \leq 0.265) \approx 0.1227$$

$$\rightarrow n(p_6)_e = 30(0.1227) = 3.6823$$

$$(p_7)_e = P(0.265 \leq N(0.2534, (0.0123)^2) \leq 0.27) = P(0.265 \leq Z \leq 0.27) \approx 0.0840$$

$$\rightarrow n(p_7)_e = 30(0.0840) = 2.5192$$

$(p_8)_e = P(0.27 \le N(0.2534, (0.0123)^2) \le 0.275) = P(0.27 \le Z \le 0.275) \approx 0.0488$

$\rightarrow n(p_8)_e = 30(0.0488) = 1.4640$

$(p_9)_e = P(0.275 \le N(0.2534, (0.0123)^2)) = P(0.275 \le Z) \approx 0.0392$

$\rightarrow n(p_9)_e = 30(0.0392) = 1.1769$

All the expected counts are not at least 5, so the test is not applicable. Thus, we have to modify it as done in the second table shown below. Now, the test is applicable.

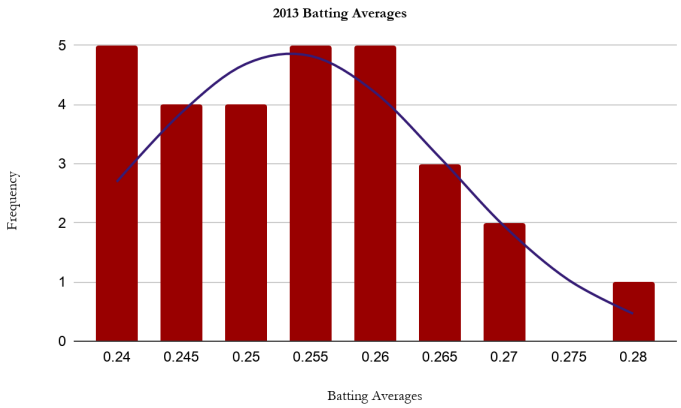
Intervals	Observed Counts	Estimated Probability	Estimated Count		Intervals	Observed Counts	Estimated Probability	Estimated Count
(0,0.24]	5	0.1385	4.1538		(0,0.245]	9	0.2481	7.4422
(0.24,0.245]	4	0.1096	3.2885	→	(0.245,0.255]	9	0.3048	9.1432
(0.245,0.25]	4	0.1440	4.3207	**MODIFIED**	(0.255,0.265]	8	0.2751	8.2545
(0.25,0.255]	5	0.1607	4.8224	**VERSION**	(0.265,∞)	4	0.1720	5.1601
(0.255,0.26]	5	0.1524	4.5722			30	1	30
(0.26,0.265]	3	0.1227	3.6823					
(0.265,0.27]	2	0.0840	2.5192					
(0.27,0.275]	0	0.0488	1.4640					
(0.275,∞)	2	0.0392	1.1769					
	30	1	30					

df = 4 - 1 - 2 = 1

$$d = \frac{(9-7.4422)^2}{7.4422} + \frac{(9-9.1432)^2}{9.1432} + \frac{(8-8.2545)^2}{8.2545} + \frac{(4-5.1601)^2}{5.1601} = 0.5790$$

$d^* = 3.8415 > d$

→ Thus, we fail to reject the null hypothesis at the α =0.05 level of significance. So, we have convincing evidence that the home runs follow a normal distribution.



[Note: We can see from the histogram and normal approximation curve that the data fits the normal distribution curve pretty well even with the left-skewness.]

III. The third case was when we were not able to apply the chi-square goodness-of-fit test since we could not modify the data to have at least 4 class intervals all with expected counts greater than or equal to 5. For this situation, we instead used the condition that at least 75% of our data was at least 5, and thus the test became more appropriate for the situation. An example of this data was from the 2017 Strikeouts data. In general, the most modification and the use of this condition was used for the data for strikeouts. The test was done as the following:

Assuming H_0 is true, with $\bar{x}$ = 1336.8 (estimate for μ) and $s = 122.8698$ (estimate for σ), we calculated the probability and expected count for each interval.

$$(p_1)_e = P(N(1336.8, (122.8698)^2) \leq 1100) = P(Z \leq 1100) \approx 0.0270$$

$$\rightarrow n(p_1)_e = 30(0.0270) = 0.8092$$

$$(p_2)_e = P(1100 \leq N(1336.8, (122.8698)^2) \leq 1200) = P(1100 \leq Z \leq 1200) \approx 0.1058$$

$$\rightarrow n(p_2)_e = 30(0.1058) = 3.1740$$

$$(p_3)_e = P(1200 \le N(1336.8, (122.8698)^2) \le 1300) = P(1200 \le Z \le 1300) \approx 0.2495$$

$$\rightarrow n(p_3)_e = 30(0.2495) = 7.4851$$

$$(p_4)_e = P(1300 \le N(1336.8, (122.8698)^2) \le 1400) = P(1300 \le Z \le 1400) \approx 0.3142$$

$$\rightarrow n(p_4)_e = 30(0.3142) = 9.4267$$

$$(p_5)_e = P(1400 \le N(1336.8, (122.8698)^2) \le 1500) = P(1400 \le Z \le 1500) \approx 0.2114$$

$$\rightarrow n(p_5)_e = 30(0.2114) = 6.3434$$

$$(p_6)_e = P(1500 \le N(1336.8, (122.8698)^2) \le 1600) = P(1500 \le Z \le 1600) \approx 0.0760$$

$$\rightarrow n(p_6)_e = 30(0.0760) = 2.2787$$

$$(p_7)_e = P(1600 \le N(1336.8, (122.8698)^2)) = P(1600 \le Z) \approx 0.0161$$

$$\rightarrow n(p_7)_e = 30(0.0161) = 0.4828$$

All the expected counts are not at least 5, so the test is not applicable. Thus, we have to modify it as done in the second table shown below. Note that even after the modification we are not able to get all expected counts to be at least 5. Thus, we use the at least 75% is greater than or equal to 5 condition to make our test work.

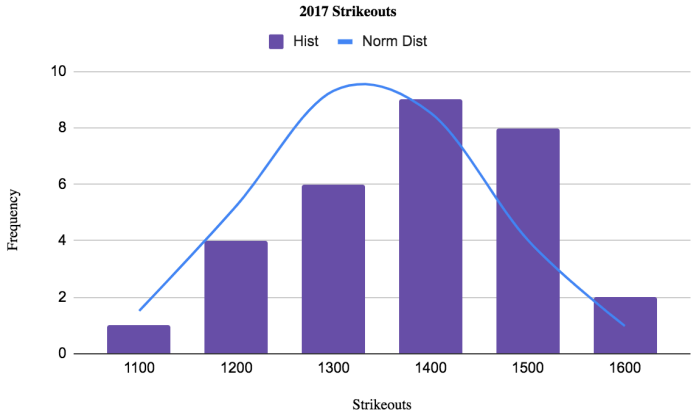
Intervals	Observed Counts	Estimated Probability	Estimated Count		Intervals	Observed Counts	Estimated Probability	Estimated Count
(0,1100]	1	0.0270	0.8092		(0,1200]	5	0.1328	3.9832
(1100,1200]	4	0.1058	3.1740	→	(1200,1300]	6	0.2495	7.4851
(1200,1300]	6	0.2495	7.4851	**MODIFIED**	(1300,1400]	9	0.3142	9.4267
(1300,1400]	9	0.3142	9.4267	**VERSION**	(1400,∞)	10	0.3035	9.1049
(1400,1500]	8	0.2114	6.3434			30	1	30
(1500,1600]	2	0.0760	2.2787					
(1600,∞)	0	0.0161	0.4828					
	30	1	30					

df = 4 - 1 - 2 = 1

$$d = \frac{(5-3.9832)^2}{3.9832} + \frac{(6-7.4851)^2}{7.4851} + \frac{(9-9.4267)^2}{9.4267} + \frac{(10-9.1049)^2}{9.1049} = 0.6615$$

$$d^* = 3.8415 > d$$

→ Thus, we fail to reject the null hypothesis at the α =0.05 level of significance. So, we have convincing evidence that the home runs follow a normal distribution.



[Note: We can see from the histogram and normal approximation curve that the data fits the normal distribution curve pretty well.]

IV. Despite the challenges we faced above regarding skewness or using a different condition, we were still able to prove normality. However, out of the 44 tests we conducted, there was 1 dataset for which we could not prove normality. This data was from 2008 Strikeouts. Thus, this year was removed from the regression model, which we will see later. The test was done as the following:

Assuming H_0 is true, with $\bar{x}$ = 1096.1 (estimate for μ) and $s = 116.4605$(estimate for σ), we calculated the probability and expected count for each interval.

$$(p_1)_e = P(N(1096.1, (116.4605)^2) \leq 900) = P(Z \leq 900) \approx 0.0461$$

$$\rightarrow n(p_1)_e = 30(0.0461) = 1.3832$$

$$(p_2)_e = P(900 \le N(1096.1, (116.4605)^2) \le 1000) = P(900 \le Z \le 1000) \approx 0.1585$$

$$\rightarrow n(p_2)_e = 30(0.1585) = 4.7559$$

$$(p_3)_e = P(1000 \le N(1096.1, (116.4605)^2) \le 1100) = P(1000 \le Z \le 1100) \approx 0.3087$$

$$\rightarrow n(p_3)_e = 30(0.3087) = 9.2616$$

$$(p_4)_e = P(1100 \le N(1096.1, (116.4605)^2) \le 1200) = P(1100 \le Z \le 1200) \approx 0.3005$$

$$\rightarrow n(p_4)_e = 30(0.3005) = 9.0146$$

$$(p_5)_e = P(1200 \le N(1096.1, (116.4605)^2) \le 1300) = P(1200 \le Z \le 1300) \approx 0.1462$$

$$\rightarrow n(p_5)_e = 30(0.1462) = 4.3850$$

$$(p_6)_e = P(1300 \le N(1096.1, (116.4605)^2) \le 1400) = P(1300 \le Z \le 1400) \approx 0.0355$$

$$\rightarrow n(p_6)_e = 30(0.0355) = 1.0637$$

$$(p_7)_e = P(1400 \le N(1096.1, (116.4605)^2) = P(1400 \le Z) \approx 0.0045$$

$$\rightarrow n(p_7)_e = 30(0.0045) = 0.1360$$

All the expected counts are not at least 5, so the test is not applicable. Thus, we have to modify it as done in the second table shown below. Now, the test is applicable.

Intervals	Observed Counts	Estimated Probability	Estimated Count
(0,900]	1	0.0461	1.3832
(900,1000]	5	0.1585	4.7559
(1000,1100]	12	0.3087	9.2616
(1100,1200]	3	0.3005	9.0146
(1200,1300]	8	0.1462	4.3850
(1300,1400]	1	0.0355	1.0637
(1400,∞)	0	0.0045	0.1360
	30	1	30

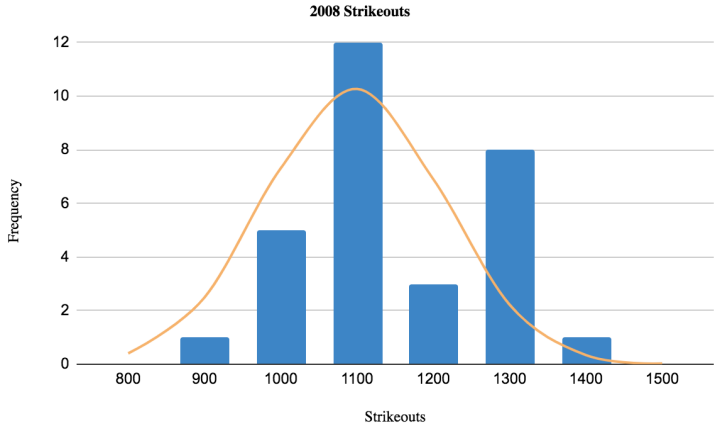
	Intervals	Observed Counts	Estimated Probability	Estimated Count
	(0,1000]	6	0.2046	6.1391
→	(1000,1100]	12	0.3087	9.2616
MODIFIED	(1100,1200]	3	0.3005	9.0146
VERSION	(1200,∞)	9	0.1862	5.5847
		30	1	30

df = 4 - 1 - 2 = 1

$$d = \frac{(6-6.1391)^2}{6.1391} + \frac{(12-9.2616)^2}{9.2616} + \frac{(3-9.0146)^2}{9.0146} + \frac{(9-5.5847)^2}{5.5847} = 6.9144$$

$$d^* = 3.8415 < d$$

→ Thus, we reject the null hypothesis at the α =0.05 level of significance. So, we have convincing evidence that the home runs do NOT follow a normal distribution.



[Note: When we draw the histograms we can see that the data is very roughly symmetric, so we could use the data, but we will be leaving it out since we tested for it.]

Note, that we do not include all our tests for normality in this paper due to the large number of tests that were conducted. However, all the tests could be characterized as having used a method similar to one of the 4 examples that were demonstrated.

B. The Independence Assumption

In our presentation, we attempted to prove independence between two different seasons by testing the league average number of home runs in relation to the year. Though we did this in an attempt to show that we can treat two separate seasons as independent from one another, we believe we inadvertently tested whether the league average number of home runs is independent of the season, which was not our intention. After considering this problem of determining

whether we could treat two different seasons as independent, we determined that we must settle with assuming this to be true. By assuming that subsequent seasons are independent, we can perform the tests we carried out in this project. However, if the seasons are not truly independent, we can trust that the tests are robust enough that any deviation from our assumption is negligible enough such that our results are still reliable.

Along with independence, we checked to see if our data for 2014 and 2016 were identically distributed since that is also an important assumption when using the Central Limit Theorem. Our null hypothesis was that the 2014 and 2016 followed the same distribution and our alternative hypothesis was they do not. We checked at the $\alpha = 0.05$ level of significance. First, we took the data from 2014 and found the observed probabilities.

Homerun Class Intervals	Observed Count	Observed Probability
(0, 100]	1	0.0333
(100, 120]	5	0.1667
(120, 140]	10	0.3333
(140, 160]	10	0.3333
(160, 180]	2	0.0667
(180, ∞)	2	0.0667

Then we created a table for the data we saw in 2016 as shown below.

Homerun Class Intervals	Observed Count
(0, 100]	0
(100, 120]	1
(120, 140]	3
(140, 160]	3
(160, 180]	5
(180, ∞)	18

So, now we wanted to check if the distribution found in the data for home runs in 2014 was similar to the distribution for home runs in 2016. So, we take the probability observed for each class interval for 2014, and assuming that the null hypothesis is true, use those probabilities to calculate the expected probabilities for 2016. (We are simply multiplying each probability from the 2014 table above by the total number of teams, 30). Doing this we get the following table,

Homerun Class Intervals	Observed Count	Estimated Expected Count USING 2014
(0, 100]	0	1
(100, 120]	1	5
(120, 140]	3	10
(140, 160]	3	10
(160, 180]	5	2
(180, ∞)	18	2

We see that not all of our estimated expected counts are greater than or equal to 5, so we need to modify our tables.

2014 ***2016***

Homerun Class Intervals	Observed Count	Estimated Probability	Estimated Expected Count
(0, 120]	6	0.2000	6
(120, 140]	10	0.3333	10
(140, ∞)	14	0.4667	14

Homerun Class Intervals	Observed Count	Estimated Expected Count USING 2014
(0, 120]	1	6
(120, 140]	3	10
(140, ∞)	26	14

Now, all of the estimated expected counts for 2016 are at least 5, so we can continue to check.

df = 3 - 1 = 2

$$d = \frac{(1-6)^2}{6} + \frac{(3-10)^2}{10} + \frac{(26-14)^2}{14} = 19.3524$$

$$d^* = 5.9915$$

$d > 5.9915$

→ Since the d-value is significantly greater than d^*, we reject H_0at the $\alpha = 0.05$ level of significance and thus, we have very convincing evidence that the distributions for 2014 and 2016 are different.

We apply this conclusion to all the years before 2015 and after 2015, to show that the distribution for the data before and after 2015 is not identically distributed.

V. TESTS

A. Two-Sample Procedure

In order to test our theory about a rumored ball change in 2015, we decided to test a variety of statistics in order to find if there was any distinct change after the 2015 season. We took two seasons—2014 and 2016— and compared many league average statistics to determine if there had been a distinct change in play. Particularly, we looked at home runs, strikeouts, OPS, and ERA. Home runs and strikeout are standard, basic hitting statistics which are good indicators of how well players are batting. Though not the most complex values, they are a good basic test for player performance. A high number of home runs means that hitters are hitting pitches often and well, while a high number of strikeouts means that hitters are failing more often. OPS and ERA are slightly more complex. OPS (On-base Plus Slugging), is a combined measure of the percentage of the time a batter gets on base, plus their slugging percentage—a rough measure of how many bases a player is hitting for (singles, doubles, triples, or home runs). ERA (Earned Run Average) is a pitching statistic that counts the average number of runs a pitcher allows per inning. Similarly to home runs, a high OPS indicates good hitter performance, while a high ERA

indicates poor pitcher performance. These statistics were compared using a two-sample T procedure.

The two-sample T procedure was used to test the league average values for these four statistics between 2014 and 2016. For each test, the null hypothesis was that the mean values of each league average statistic did not change between seasons. Our alternative hypothesis were determined based on our theory that hitter performance should increase in 2016, and pitcher performance should decrease. These tests were performed at the $\boldsymbol{\alpha} = 0.05$ level of confidence. Since there are 30 teams in the MLB, this yields 58 degrees of freedom, and t-value of $\pm\ 1.677$. It should be noted that these tests were first performed under the assumption that the variances between seasons were different, though this was later found to be a faulty assumption. The method below represents the tests that were performed under the assumption that the variances are equal, and used a pooled variance. A more in-depth discussion of the variances can be found in the next section, where we cover the F test for variance.

I. Home runs Ho : $\boldsymbol{\mu}_x = \boldsymbol{\mu}_y$ Ha: $\boldsymbol{\mu}_x < \boldsymbol{\mu}_y$

$$S_p^{\ 2} = \frac{(30-1)(633.085)+(30-1)(1019.931)}{30+30-2} = 826.508$$

$$t = \frac{(140-187)}{S_p\sqrt{1/30+1/30}} = -6.332$$

t < -1.667

Therefore we reject the null hypothesis

II. Strikeouts Ho : $\boldsymbol{\mu}_x = \boldsymbol{\mu}_y$ Ha: $\boldsymbol{\mu}_x > \boldsymbol{\mu}_y$

$$S_p^{\ 2} = \frac{(30-1)(11921.206)+(30-1)(14484.466)}{30+30-2} = 13,202.8$$

$$t = \frac{(1070.027-1427)}{S_p\sqrt{1/30+1/30}} = -1.735$$

t < +1.667

Therefore we fail to reject the null hypothesis.

III. OPS Ho : $\boldsymbol{\mu}_x = \boldsymbol{\mu}_y$ Ha: $\boldsymbol{\mu}_x < \boldsymbol{\mu}_y$

$$S_p^{\ 2} = \frac{(30-1)(0.000907)+(30-1)(0.000813)}{30+30-2} = 0.00086$$

$$t = \frac{(0.700-0.739)}{S_p\sqrt{1/30+1/30}} = -5.147$$

t < -1.667

Therefore we reject the null hypothesis.

IV. ERA Ho : $\boldsymbol{\mu}_x = \boldsymbol{\mu}_y$ Ha: $\boldsymbol{\mu}_x < \boldsymbol{\mu}_y$

$$S_p^{\ 2} = \frac{(30-1)(0.190)+(30-1)(0.208)}{30+30-2} = 0.199$$

$$t = \frac{(3.738-4.184)}{S_p\sqrt{1/30+1/30}} = -3.876$$

t < -1.667

Therefore we reject the null hypothesis.

The results of these hypothesis tests showed that home runs, OPS, and ERA all increased between 2014 and 2016, indicating that hitters were playing better and pitchers were playing worse. One will notice, however, that we failed to reject the null hypothesis for strikeouts. In fact, performing an opposite hypothesis test that strikeouts increased after 2015, we found that we were actually able to reject the null hypothesis:

V. Strikeouts Ho : $\boldsymbol{\mu}_x = \boldsymbol{\mu}_y$ Ha: $\boldsymbol{\mu}_x < \boldsymbol{\mu}_y$

$$S_p^{\ 2} = \frac{(30-1)(11921.206)+(30-1)(14484.466)}{30+30-2} = 13,202.8$$

$$t = \frac{(1070.027-1427)}{S_p\sqrt{1/30+1/30}} = -1.735$$

$t < -1.667$

Therefore we reject the null hypothesis.

This means that, despite the fact that all our other tested values indicate that offense had clearly increased after 2015, there is clear proof that players were striking out even more than before. This goes against our initial assumption, but there is a reasonable explanation as to why this is happening. It is possible that the increase in strikeouts comes as a result of an increased trend in players attempting to hit multiple-base hits (doubles, triples, home runs), at the risk of being more likely to strikeout. According to our theory, it is possible that as it became easier to hit home runs, players attempted a more aggressive style of play which resulted in more strikeouts. Overall, however, the increased offensive output outweighed the higher rate of failure.

In summary, the T tests on home runs, strikeouts, OPS, and ERA showed a distinct change in play between 2014 and 2016. Though the increase in strikeouts was unexpected, we believe that the results support our theory that there might have been a change in the design in the ball in 2015 that changed gameplay in a way that benefited hitters.

B. F-Value Test

As stated in the previous section, the two-sample T tests were first performed under the assumption that the variances between samples were different, though this was found to not be the case. Initially, assuming that variances are different, we run into the Behrens-Fisher problem This problem is concerned with two=sample T tests which variances cannot be treated as equal, so it is inappropriate to use a pooled variance. To solve this problem, we approximated the number of degrees of freedom using:

$$df = \frac{(\theta + n/m)^2}{\theta^2/(n-1) + (n/m)^2/(m-1)}$$

Where n and m are the number of elements of each sample, and theta is the ratio of the sample variances of the two samples. Using this approximation for the degrees of freedom, one can perform T tests using the sample variances from the data.

We performed our first set of T tests using this process, but an analysis of the variances later on showed that these tests were invalid. To test the variances we used a F test for variance on each statistic: home runs, strikeouts, OPS, and ERA. For each, the null hypothesis was that the variances are equal, while the alternative hypothesis was that they are not. These tests were done at the $\boldsymbol{\alpha} = 0.05$ level of confidence. For an F random variable with 29,29 degrees of freedom, the bounds for this confidence level are (0.475, 2.101). The results are as follows.

I. Home runs Ho : $\boldsymbol{\sigma}_x = \boldsymbol{\sigma}_y$ Ha: $\boldsymbol{\sigma}_x \neq \boldsymbol{\sigma}_y$

$f = 633.085/1019.931 = 0.621$

$0.475 < f < 2.101$

Therefore we reject the null hypothesis

II. Strikeouts Ho : $\boldsymbol{\sigma}_x = \boldsymbol{\sigma}_y$ Ha: $\boldsymbol{\sigma}_x \neq \boldsymbol{\sigma}_y$

$f = 11921.206/14484.466 = 0.823$

$0.475 < f < 2.101$

Therefore we reject the null hypothesis

III. OPS Ho : $\boldsymbol{\sigma}_x = \boldsymbol{\sigma}_y$ Ha: $\boldsymbol{\sigma}_x \neq \boldsymbol{\sigma}_y$

$f = 0.000907/0.000813 = 1.116$

$0.475 < f < 2.101$

Therefore we reject the null hypothesis

IV. ERA Ho : $\boldsymbol{\sigma}_x = \boldsymbol{\sigma}_y$ Ha: $\boldsymbol{\sigma}_x \neq \boldsymbol{\sigma}_y$

$f = 0.190/0.208 = 0.913$

$$0.475 < f < 2.101$$

Therefore we reject the null hypothesis

Furthermore, the 95% confidence intervals for the ratios of the variances are

I. Home runs: (0.295, 1.304)

II. Strikeouts: (0.392, 1.729)

III. OPS: (0.531, 2.344)

IV. ERA: (0.438, 1.891)

What these tests and the confidence intervals show is that we cannot assume, for any of these tests, that the variances are different. After learning this, we threw out the results from our first round of tests, and reperformed the tests using a pooled variance, as can be seen in the previous section. Though the results of the two rounds of tests did not differ by much, we can be more confident in the validity of the subsequent tests.

C. Regression Models

In order to understand whether the trends were similar from before 2015 and after 2015, we created simple linear models using the least-squares regression line. We conducted three different tests, one testing home runs vs. strikeouts, one for batting averages vs. strikeouts, and one for on-base percentage vs. strikeouts. We used different *x* values for each of these models while keeping the *y* as strikeouts because of the conclusion we obtained earlier regarding there being evidence supporting that the number of home runs and the number of strikeouts is related. Note, that the batting average is that the average performance for the batters, so since we are looking across teams per year, we are looking at the average batting average for all 30 teams in the given year. Also, on-base percentage is a formula used to calculate how often a player

reaches base, so we will be taking the averages across all 30 teams for the given year. We wanted to use a variety of models to investigate whether or not there is actually a difference.

Our first model will be testing home runs and strikeouts. The questions we are asking through these models are: 1- Can we predict the number of strikeouts given the number of home runs in a year? 2- Does the data show significantly different data for the years 2009-2014 and 2016-2019? Note that we do not use 2008 and 2020 in this model since we were not able to prove normality for the strikeout data in 2008 and 2020 was an outlier in this case due to the pandemic, and we wanted to make sure that we removed any outliers. We should also take into consideration that the data set we use for the years after 2015 is very small, we are only looking at 4 years, thus the models not be entirely accurate to either indicate a difference or similarity.

For each model, we used the following assumptions:

1- $Y_i's$ are normally distributed, as was proven above.

2- $Y_i's$ are independent. This is an assumption we make since we consider each year to be independent.

3- The variance is constant. This means that we do not see any outliers or patterns in our residual plots (which will be shown ahead).

With these assumptions, we moved ahead to create our models. We proceeded by finding equations for the least-squares regression line/

<u>Home Runs vs Strikeouts</u>

The first table we created showed the average number of home runs and strikeouts across the 30 teams in the regular season for each year.

YEARS	2009	2010	2011	2012	2013	2014
x_i	168.0667	153.7667	151.7333	164.4667	155.3667	139.5333

y_i	1119.7000	1143.5333	1149.6000	1214.2000	1223.6667	1248.0333

Using the information from the table above, we were able to find the sample mean for both *x* and *y*, the summation of $x_i's$ and $y_i's$, and the sum of squares of $x_i's$ and $y_i'sso$ that we can find the appropriate *a* and *b* values for the least-square regression line.

$\bar{x}$	155.4889	Σx_i=	932.9333	Σx_i^2=	145571.2333
$\bar{y}$	1183.1222	Σy_i=	7098.7333	Σy_i^2=	8412205.6867

Now, that we have these values we can calculate the *a* and *b*. Remember that *b* represents the slope and using this data we calculate it as the following:

$$b = \frac{6\Sigma x_i y_i - (932.9333)(7098.7333)}{6(145571.2333) - (932.9333)^2} = \frac{6614453.307 - 6622644.683}{873427.398 - 870364.542} = \frac{-8191.38}{3062.856} = -2.6746$$

$$a = 1183.1222 - (-2.6746 * 155.4889) = 1183.1222 + 415.8706 = 1598.9873$$

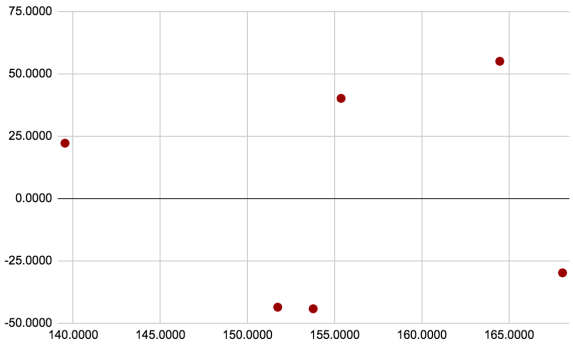
$$\Rightarrow \hat{y} = 1598.9873 - 2.6746x$$

Using the least-square regression line we found, we want to find the residuals by doing $y_i - \hat{y}(x_i)$.

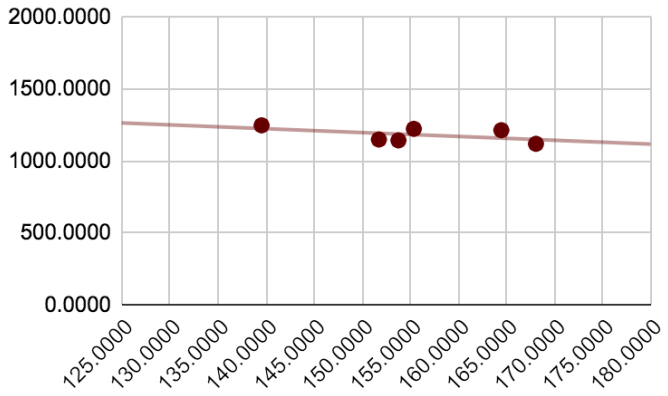
x_i	y_i	$\hat{y}(x_i)$	e_i
168.0667	1119.7000	1149.4762	-29.7762
153.7667	1143.5333	1187.7230	-44.1896
151.7333	1149.6000	1193.1613	-43.5613
164.4667	1214.2000	1159.1048	55.0952
155.3667	1223.6667	1183.4436	40.2231
139.5333	1248.0333	1225.7914	22.2419

After finding the residuals, we want to take the sum of the residuals to see how accurate our linear regression line is. The sum of the residuals is 0.0330. This shows that our linear regression line is somewhat close to 0, so it is a pretty close line for the model. First, we will graph the

residual plots to make sure that there are no outliers or patterns so that our third assumption is satisfied.



From this residual plot, we see an equal number of residuals above and below the line and no significant operators, so we can assume that the variance is constant and proceed to plot the least-squares regression line.



This linear regression model shows a negative trend between the average number of home runs and the average number of strikeouts for the years before 2015. Now, we will compare it for the years after 2015.

YEARS	2016	2017	2018	2019
x_i	187.0000	203.5000	186.1667	225.8667
y_i	1299.5000	1336.8000	1373.5667	1427.4333

Using a similar technique from above, we will be finding all the values needed to calculate for the slope *b* and y-intercept *a*.

$\bar{x}$	200.6333	Σx_i =	802.5333	Σx_i^2 =	162055.0289
$\bar{y}$	1359.3250	Σy_i =	5437.3000	Σy_i^2 =	7399985.7989

$$b = \frac{4\Sigma x_i y_i - (802.5333)(5347.3000)}{4(1162055.0289) - (802.5333)^2} = 2.1764$$

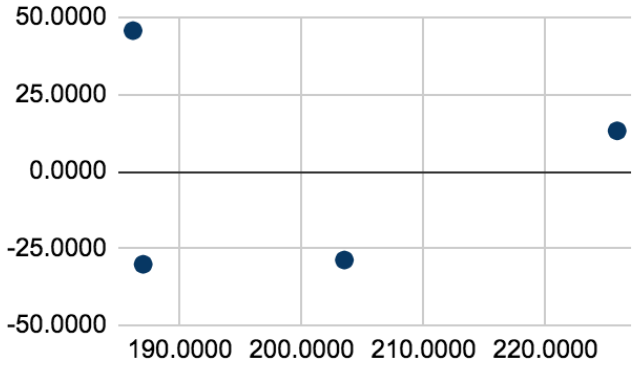
$$a = 1359.3250 - (2.1764 * 200.6333) = 922.6745$$

$$\Rightarrow \hat{y} = 922.6745 + 2.1764x$$

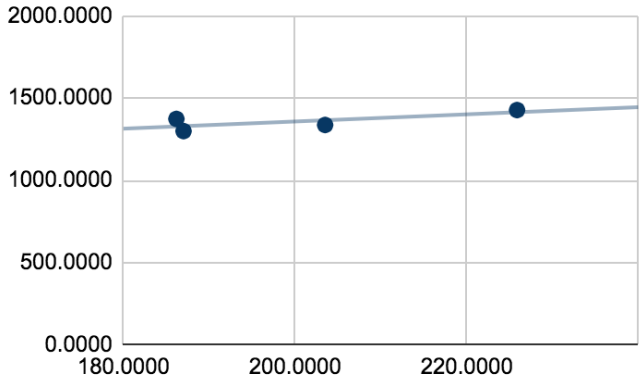
Using the least-square regression line we found, we want to find the residuals by doing $y_i - \hat{y}(x_i)$.

x_i	y_i	$\hat{y}(x_i)$	e_i
187.0000	1299.5000	1329.6613	-30.1613
203.5000	1336.8000	1365.5719	-28.7719
186.1667	1373.5667	1327.8476	45.7190
225.8667	1427.4333	1414.2507	13.1826

After finding the residuals, we want to take the sum of the residuals to see how accurate our linear regression line is. The sum of the residuals is -0.0315. This shows that our linear regression line is somewhat close to 0, so it is a pretty close line for the model. First, we will graph the residual plots to make sure that there are no outliers or patterns so that our third assumption is satisfied.



From this residual plot, we see an equal number of residuals above and below the line and no significant operators, so we can assume that the variance is constant and proceed to plot the least-squares regression line.



This linear regression model shows a positive trend between the average number of home runs and the average number of strikeouts for the years after 2015. Comparing this to the results from before 2015, we see a clear difference between before and after 2015 in the relationship between home runs and strikeouts. It is also important to remember that we had a very small number of years after 2015 to look at. We were able to use the average number of home runs to create a pretty close linear regression line and to predict the number of strikeouts and we could see that the linear regression models were different.

Batting Averages vs Strikeouts

The first table we created showed the average number of batting averages and strikeouts across the 30 teams in the regular season for each year.

YEARS	2009	2010	2011	2012	2013	2014
x_i	0.2623	0.2573	0.2549	0.2547	0.2534	0.2511
y_i	1119.7000	1143.5333	1149.6000	1214.2000	1223.6667	1248.0333

$\bar{x}$	0.2556	Σx_i=	1.5337	Σx_i^2=	0.3921
$\bar{y}$	1183.1222	Σy_i=	7100.2670	Σy_i^2=	8412205.6867

$$b = \frac{6\Sigma x_i y_i - (1.5337)(7100.2670)}{6(0.3921) - (1.5337)^2} = -\ 17067.4335$$

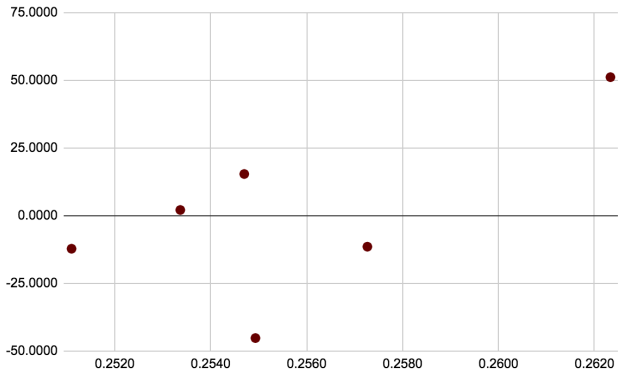
$$a = 1183.1222 - (-\ 17067.4335 * 0.2556) = 5545.8427$$

$$\Rightarrow \hat{y} = 5545.8427 - 17067.4335x$$

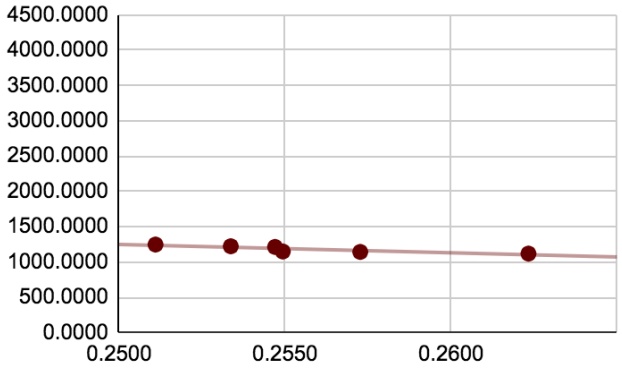
Using the least-square regression line we found, we want to find the residuals by doing $y_i - \hat{y}(x_i)$.

x_i	y_i	$\hat{y}(x_i)$	e_i
0.2623	1119.7000	1068.4860	51.2140
0.2573	1143.5333	1154.9610	-11.4276
0.2549	1149.6000	1194.7850	-45.1850
0.2547	1214.2000	1198.7674	15.4326
0.2534	1223.6667	1221.5240	2.1427
0.2511	1248.0333	1260.2101	-12.1768

The sum for the residuals is -0.0001 which is extremely close to 0, so we will suppose that our least-squares regression line is appropriate for the data. We can plot the residuals first to check if this line is appropriate for the data.



We do not see any significant outliers or patterns, so it is appropriate to use $\hat{y} = 5545.8427 - 17067.4335x$. Now, we can plot the regression line to see what type of trend we see with the data.



The line shows a negative relationship between the average of batting averages and the average number of strikeouts. Now, we want to see if this is significantly different from the after 2015 model.

YEARS	2016	2017	2018	2019
x_i	0.2553	0.2549	0.2478	0.2522
y_i	1299.5000	1336.8000	1373.5667	1427.4333

$\bar{x}$	0.2526	Σx_i=	1.0102	Σx_i^2=	0.2552

$\bar{y}$	1359.3250	Σy_i=	5437.3000	Σy_i^2=	7399985.7989

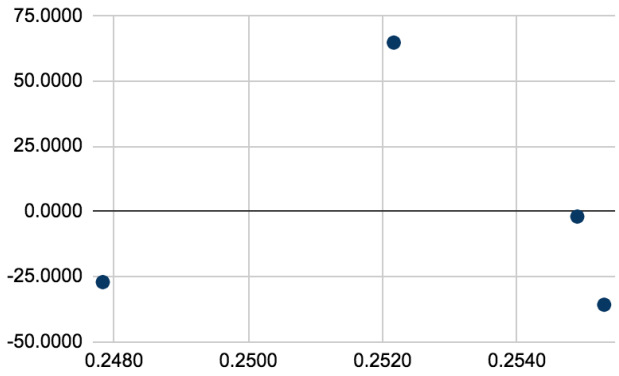
$$b = \frac{4\Sigma x_i y_i - (1.0102)(5437.3000)}{4(0.2552) - (1.0102)^2} = -\ 8758.2757$$

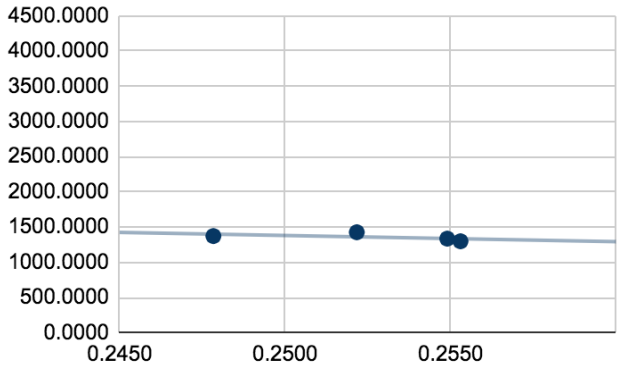
$$a = 1359.3250 - (-\ 8758.2757 * 0.2526) = 3571.2275$$

$$\Rightarrow \hat{y} = 3571.2275 - 8758.2757x$$

x_i	y_i	$\hat{y}(x_i)$	e_i
0.2553	1299.5000	1335.2397	.739-357
0.2549	1336.8000	1338.7430	-1.9430
0.2478	1373.5667	1400.6348	-27.0682
0.2522	1427.4333	1362.6823	64.7510

The sum of the residuals in 0.0001 so our regression line is appropriate for the data set. We can go ahead and plot the residuals to see if there are any outliers or patterns, and there are none so we can proceed ahead and draw the least-squares regression line.





Similar to before 2015, we see a negative relationship between batting averages and strikeouts. However, we can see that the equations for the two lines are different:

$\hat{y} = 5545.8427 - 17067.4335x$ and $\hat{y} = 3571.2275 - 8758.2757x$. The second one has a lower y-intercept and slope. There is a difference between the two models, however, it is important to remember that we only have data for 4 years after 2015 and that the least-squares regression line may have been different if more data was present.

On-Base Percentages vs, Strikeouts

The first table we created showed the average number of on-base percentages and strikeouts across the 30 teams in the regular season for each year.

YEARS	2009	2010	2011	2012	2013	2014
x_i	0.3328	0.3261	0.3205	0.3190	0.3175	0.3137
y_i	1119.7000	1143.5333	1149.6000	1214.2000	1223.6667	1248.0333

$\bar{x}$	0.3216	Σx_i =	1.9295	Σx_i^2 =	0.6207
$\bar{y}$	1183.1222	Σy_i =	7098.7333	Σy_i^2 =	8412205.6867

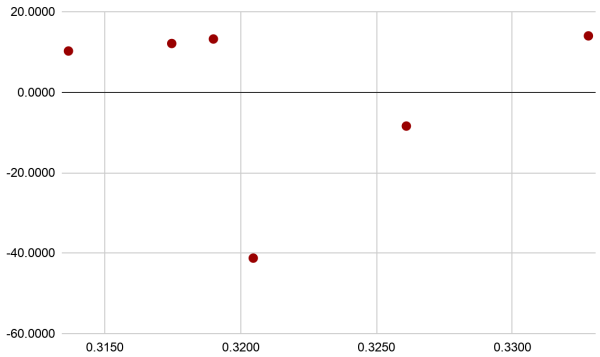
$$b = \frac{6\Sigma x_i y_i - (1.9295)(7098.7333)}{6(0.6207) - (1.9295)^2} = -\ 6903.9569$$

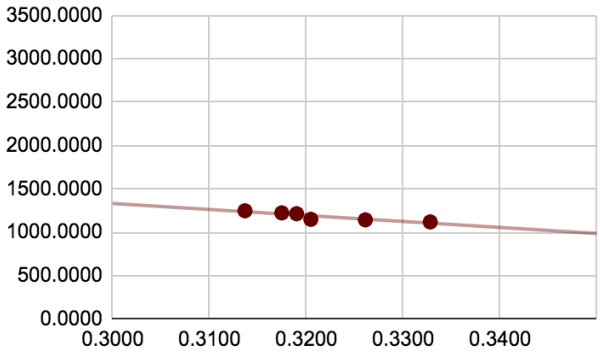
$$a = 1183.1222 - (-\ 6903.9569 * 0.3216) = 3403.3197$$

$$\Rightarrow \hat{y} = 3403.3197 - 6903.9569x$$

x_i	y_i	$\hat{y}(x_i)$	e_i
0.3328	1119.7000	1105.6828	14.0172
0.3261	1143.5333	1151.9394	-8.4060
0.3205	1149.6000	1190.8316	-41.2316
0.3190	1214.2000	1200.9574	13.2426
0.3175	1223.6667	1211.5435	12.1232
0.3137	1248.0333	1237.7786	10.2548

The sum of the residuals is 0.00 so the line is a good fit for the data, and we can go ahead and plot it.





We see a negative relationship between on-base percentage and strikeouts. This is consistent

with the negative slope seen for the models before 2015 in the other 2 models above. Now, to check if this is similar to the linear regression model for after 2015.

YEARS	2016	2017	2018	2019
x_i	0.3214	0.3243	0.3180	0.3224
y_i	1299.5000	1336.8000	1373.5667	1427.4333

$\bar{x}$	0.3215	Σx_i=	1.2861	Σx_i^2=	0.4136
$\bar{y}$	1359.3250	Σy_i=	5437.3000	Σy_i^2=	7399985.7989

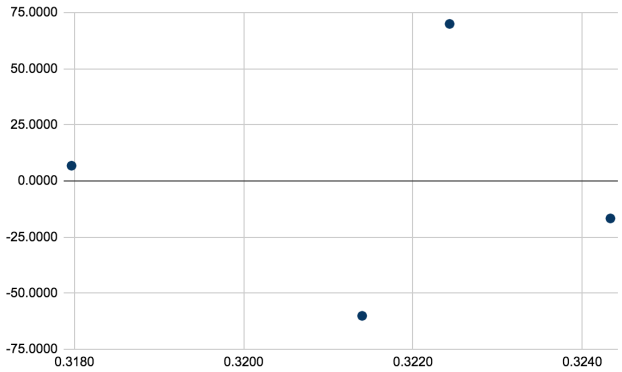
$$b = \frac{4\Sigma x_i y_i - (1.2861)(5437.3000)}{4(0.4136) - (1.2861)^2} = -2084.7792$$

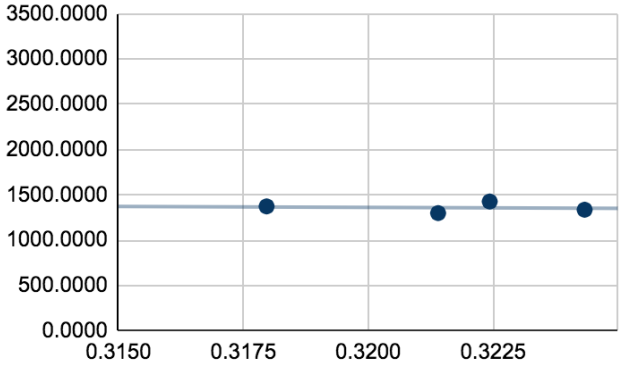
$$a = 1359.3250 - (-2084.7792 * 0.3215) = 2029.6510$$

$$\Rightarrow \hat{y} = 2029.6510 - 2084.7792x$$

x_i	y_i	$\hat{y}(x_i)$	e_i
0.3214	1299.5000	1359.6030	-60.1030
0.3243	1336.8000	1353.4876	-16.6876
0.3180	1373.5667	1366.7607	6.8060
0.3224	1427.4333	1357.4487	69.9846

The sum of the residuals here again is 0.00, so we know that the line is a good fit for the data.





When we plot it and after seeing the slope, we see that there is a negative slope, however the slope is much lower than the one we have for the before 2015 least-squares regression line. This shows us that there has been a change.

We were able to find least-square regression lines to predict strikeouts from home runs, batting averages, and on-base percentages, All three regression models have shown us changes in the before 2015 and after 2015 years, but the most significant one has been the home runs vs strikeouts, which has changed dramatically. It is also important to not that it is possible that if we had data from more years after 2015, we could have seen a greater difference in the batting average vs strikeouts and on-base percentage vs strikeouts models. However, even through the limited amount of data we see that there has been a change in the MLB.

VI. Comparison to the NPB

In the previous sections we found evidence that shows that there was a distinct change in play after the 2015 season. Though these tests could not test specifically if this change came as the results of a change in the ball, it is possible to further test our theory by analyzing baseball leagues separate from the MLB. If our theory that the change in play is due to a change in the MLB standard baseball is correct, then we would not necessarily expect that unrelated leagues

would see a similar pattern over that time period. If other leagues did see a similar pattern, then it is possible that the change in play is not unique to the MLB, and the changed-ball theory is weakened. If there is no such similar pattern, then the theory is strengthened, since it seems to imply that the change is specific to the MLB. To test this, we analyzed the Nippon Professional Baseball League (NPB). Specifically, we performed a test similar to that in section 5.A: a two-sample T procedure testing whether the league average home runs increased between 2014 and 2015 at the $\boldsymbol{\alpha} = 0.05$ level of confidence. As with the previous tests, we assumed that the variances are different. The result is as follows:

Home runs Ho : $\boldsymbol{\mu}_x = \boldsymbol{\mu}_y$ Ha: $\boldsymbol{\mu}_x < \boldsymbol{\mu}_y$

$$S_p^{\,2} = \frac{(30-1)(575.720)+(30-1)(545.670)}{30+30-2} = 560.691$$

$$t = \frac{(113.083-111.636)}{S_p\sqrt{1/30+1/30}} = 0.237$$

t > -1.667

Therefore we fail to reject the null hypothesis

This test showed that there is not substantial evidence that home runs increased between 2014 and 2016 in the NPB. This confirms our suspicions that the increase in offensive capability in the MLB may be unique to the MLB, since we could not find a similar increase in the NPB.

VII. CONCLUSION

Our results find that home runs in the MLB seasons after 2015 increased significantly, which seemed to corroborate with our hypothesis, but so do the number of strikeouts, which our hypothesis did not account for. While our results indicate that a major change in both home run and strikeout statistics, we were unable to determine if the only plausible explanation was a ball

change. There are plenty of other plausible explanations, particularly because of the increase in strikeouts, which could explain the increase in home runs; batters might have taken a more aggressive approach during the 2015 season, for instance, meaning they hit more home runs but also increased their odds of striking out as opposed to playing it safe and not striking out. Our data cannot definitively state whether or not the balls were switched, unfortunately, though we do find that baseball fans and conspiracy theorists were at least partially correct: there was a statistically significant increase in home runs starting in the middle of the 2015 season. In conclusion, we find strong evidence to support our hypothesis, that the MLB switched the balls, but we cannot ultimately prove this.

VIII. FURTHER WORK

If we had more time or continue to work on this project, it would be interesting to compare our findings with even more baseball leagues across the world. Another interesting thing we found while doing this project was that the MLB will be slightly modifying its baseballs sometime during this year. So it would be interesting to compare the data from this year with the results we found in our project. There are also other factors of data that we were not able to take into account when doing this presentation, such as looking at data for doubles, triples, and etc, which we had originally talked about.

IX. REFERENCES

"MLB Stats, Scores, History, & Records." *Baseball*, www.baseball-reference.com/.

MLB Stats.https://www.mlb.com/stats/

一般社団法人日本野球機構 . “2014 Regular Season.” *2014 Regular Season | NPB.jp* 日本野球機構, npb.jp/bis/eng/2014/stats/.

一般社団法人日本野球機構 . “2016 Regular Season.” *2016 Regular Season | NPB.jp* 日本野球機構, npb.jp/bis/eng/2016/stats/.