Understanding Replacement and Incineration in Blaseball via Permutation Tests

Cuttlefishman (@Cuttlefishman#6487) (he/him)*
*Society for Internet Blaseball Research

ABSTRACT *Comparison via permutation tests of incinerated and replacement players to the non-incinerated and non-replacement populations indicates that player quality does not significantly different from each other for Season 3, when calculated either for player's star ratings or for various performance based statistics.*

KEYWORDS
Incineration
Replacement
Permutation Test

Dare pondus idonea fumo
Fit to give weight to smoke.

Persius
Satires, V, 20.

INTRODUCTION

Blaseball, even with the beginning of reliable statistics collection by the Society for Internet Blaseball Research (SIBR), is still a murky sport to understand. While attention has been played to many facets of the game in prior research, perhaps one of the most obscured part of the game is the nature of incineration and replacement in the Internet Blaseball League (IBL), the premier professional level baseball league, especially as the happenings of the both the office of the commissioner[1] and the underleagues remain highly mysterious.

This paper, using season 3 data, begins to try to measure the nature of incineration and replacement in the IBL as the Uncertainty Era brought about a substantial increase in player incinerations (and subsequent replacements). While the process of determining "replacement level" as done in reality league baseball sabermetrics relies on data not available to us at present[2], comparisons can be made with two main sources of data - the star ratings of players, and the value of their performance based stats.

DATA

This study makes use of a Season 3 player data, as captured, interpreted, and broken down by the the datawitches of SIBR (tehstone 2020), (no space 2020). This data set of 40 incinerated and 40 replacement players records most, although not all, players who have been incinerated or replaced so far in IBL's history[3].

This set is based off of hourly scrapes of player data from the Blaseball API, so star ratings should reflect the stlats just before incineration or after replacement occurs. This is then compared against non-replaced, non-incinerated players as a baseline measurement. All performance-based statistics are based off of the entirety of a player's season 3 performance (extended or reduced based off of the portion of the season that they played), where stats which are not plate appearance neutral are ignored (to avoid the obvious bias present where incinerated or replacement players who by nature missed some amount of games would obviously have lower values).

METHODOLOGY

The specific circumstances that this data set presents, where we have multiple sets from the same population, but with an unknown distribution and with differing samples presents a challenge to most standard probability tests. Varying measurements of the average can be compared directly; Table 1 for the star ratings of the population and the sampled groups, Table 6 for various plate appearance neutral performance-based batting statistics, and Table 7 for pitching appearance neutral performance-based pitching statistics.

On the surface this would indicate that these groups are identical in nature, but this is by itself is not sufficient evidence as such. However, permutation based probability tests provide the ability to evaluate these more completely, by comparing the current distribution of batting stars with other potential distributions to

Manuscript compiled: Tuesday 15th September, 2020

[1] Who is doing a great job.

[2] For more information on how the performance and value of a reality baseball league "replacement-level" player is measured and identified, sabermetrics site Fan-Graphs has provided a number of detailed explanations of their methodology. See (Slowibski 2010) and (Cameron 2013)

[3] Some issues with connecting with the IBL's API has meant that at parts, data is either missing or incomplete in nature. In the data set in used for this paper, this is most apparent with the fact that the pitching performance data set records 8 replacement players, but only 6 incinerated players.(no space 2020). While not ideal, this shouldn't make or break the research conducted in this paper.

evaluate if this understanding that the two groups are identical is true[4].

Table 1 Various Sample and Population Star Ratings

Sample	Measure of Average	Batting Stars	Pitching Stars
Season 3 Adjusted Player Population (n=321)	Mean	2.258567	1.788162
	Median	2	1.5
	Mode	2	1.5
Non-Replacement, Non-Incinerated Players (n = 243)	Mean	2.273663	1.829218
	Median	2	1.5
	Mode	2	1.5
Incinerated Players (n = 40)	Mean	2.3375	1.525
	Median	2	1.5
	Mode	2	1
Replacement Players (n=40)	Mean	2.1	1.8
	Median	2	1.75
	Mode	1	1.5

To facilitate measuring if these incinerated and replaced players are different from the population at large, a series of Asymptotic Two-Sample Fisher-Pitman Permutation Tests were run between Non-Replacement and Replacement and Non-Incinerated and Incinerated players respectively to test if these two samples are statistically distinguishable from each other[5], using the coin package[6] in the statistical software R. These tests were repeated for both batting and pitching stars, out of stlat based information, as well as for performance-based batting statistics of SLG, BA, BABIP, K%, BB%, AB/HR, OBP, and BlwOBA, and the pitching-based performance statistics of Avg RD / Loss, ERA, FIP, and WHIP.[7].

RESULTS

The results of the permutation tests for batting and pitching stars and for performance-based statistics can be seen in Tables 2, 3, and 4 respectively.

ANALYSIS

Both incinerated and replacement players fall well into the centre of any reasonable confidence interval for batting, pitching stars, and the majority of performance-based statistics observed. Assuming a 95% confidence interval, only the ERA and FIP for incinerated pitchers fall outside. Looking more closely at the statistics lines, Paul Barnes stands out as a pitcher from this period, with a recorded ERA of 45 and FIP of 0, as he only pitched for a single inning prior to incineration. This seems to explain a large part of why ERA and FIP are outside of the confidence interval, as when Barnes is removed and ERA and FIP tests are recalculated, ERA falls within the 95% confidence interval and FIP is significantly closer to it, as shown in Table 5.

This points to the potential that FIP in blaseball might be calculated in a manner which is erroneous potentially, although further evidence from a broader span of seasons and a larger dataset would be needed before any persuasive argument to this effect could be made.

Reviewing the data as a whole, the fact that incinerated and replacement players fall well within the centre of their non-replacement, non-incinerated siblings in both star ratings and performance-based statistics is significant. This is the first strong evidence that incineration as an effect is not targeting any specific

Table 2 Results of Star Rating Asymptotic Two-Sample Fisher-Pitman Permutation Tests

Sample Pair	Test	Z-Score	p-value
Incinerated, Not Incinerated Samples	Batting Stars	0.54694	0.5844
	Pitching Stars	-1.7622	0.07804
Replacement, Not Replacement Samples	Batting Stars	1.0987	0.2719
	Pitching Stars	-0.07927	0.9368

Table 3 Results of Plate Appearance Neutral Performance-based Batting Statistics Asymptotic Two-Sample Fisher-Pitman Permutation Tests

Sample Pair	Test	Z-Score	p-value
Incinerated, Not Incinerated Samples	SLG	-0.39456	0.6932
	BA	-0.54468	0.586
	BABIP	-0.59221	0.5537
	K%	-1.4347	0.1514
	BB%	0.59748	0.5502
	AB/HR	-0.42868	0.6682
	OBP	0.077131	0.9385
	BlwOBA	-0.12448	0.9009
Replacement, Not Replacement Samples	SLG	1.5312	0.1257
	BA	1.8925	0.05843
	BABIP	1.9313	0.05345
	K%	-0.1875	0.8513
	BB%	0.97569	0.3292
	AB/HR	0.64164	0.5211
	OBP	1.8035	0.07131
	BlwOBA	1.8944	0.05817

[4] Please see (Wilber 2019) for a more detailed explanation of permutation tests
[5] That is, the alternative hypothesis where $\mu \neq 0$
[6] https://cran.r-project.org/web/packages/coin/index.html
[7] For details to how these are calculated, see (no space 2020)

Table 4 Results of Pitching Appearance Neutral Performance-based Pitching Statistics Asymptotic Two-Sample Fisher-Pitman Permutation Tests

Sample Pair	Test	Z-Score	p-value
Incinerated, Not Incinerated Samples	Avg RD / Loss	-0.69731	0.4856
	ERA	3.2735	0.001062
	FIP	-2.4649	0.01371
	WHIP	0.22414	0.8227
Replacement, Not Replacement Samples	Avg RD / Loss	0.81	0.4179
	ERA	-0.46474	0.6421
	FIP	-0.67268	0.5012
	WHIP	-1.5425	0.123

Table 5 Results of Barnes Adjusted Pitching Appearance Neutral Performance-based Pitching Statistics Asymptotic Two-Sample Fisher-Pitman Permutation Tests

Sample Pair	Test	Z-Score	p-value
Incinerated, Not Incinerated Samples	ERA	-1.4754	0.1401
	FIP	-2.1465	0.03183

statistical subcategory of players and that for blaseball, replacement players are statistically indistinguishable from the current rosters. As much as the actions of a rogue umpire when one's favorite player is incinerated saddens us, one should take some consolation in the fact that umpires are incinerating indiscriminately.

FUTURE WORK

This work was by all means the first investigation of incineration and replacement published by SIBR. Further work will need to be done to continue from this start, confirming with larger datasets that incineration and replacement are not targeted or statistically discernible in future seasons, or by which the specific mechanisms of the frequency of incinerations is understood in. At the time of this paper's publication, the emergence of new weather conditions observed on the blaseball pitch, such as feedback may have influenced the frequency of the required solar eclipses need for incinerations from what they were observed to be in Season 3.

ACKNOWLEDGEMENTS

Special thanks to BaronBliss (@baronbliss#7135) (any), the BLaTeX (Blaron of Blaseball LateX), for creating the SIBR template and initial help in bringing this paper into the template.

Special thanks as well to my various statistics teachers who laid my basis of understanding for this research.

REFERENCES

Cameron, D., 2013 The recent examples of a replacement level player. https://blogs.fangraphs.com/the-recent-examples-of-a-replacement-level-player/.

no space, 2020 Season 3 derived statistics. Google Sheets document, https://docs.google.com/spreadsheets/d/1qfMtLywJzgxB8LNzJtFRc05EW73dsT3Dut8RyfB7NxM/edit?usp=sharing.

Slowibski, S., 2010 Replacement level. https://library.fangraphs.com/misc/war/replacement-level/.

tehstone, 2020 Aggregated statistics for blaseball season 3. Google Sheets document, https://docs.google.com/spreadsheets/d/1tGDP50yFYbYYrptcUB-735D75ZOTDxHHV-XC32wUFdc/edit?usp=sharing.

Wilber, J., 2019 The permutation test: A visual explanation of statistical testing. https://www.jwilber.me/permutationtest/.

Table 6 Various Sample and Population Plate Appearance Neutral Performance-based Batting Statistics

Sample	Statistic	Mean	Median
Season 3 Adjusted Batting Population (n = 200)	SLG	0.441102203970863	0.432002388627351
	BA	0.267134605986905	0.268306332842415
	BABIP	0.279449891045805	0.281599483839253
	K%	0.134325025405184	0.134572882588150
	BB%	0.634371200853981	0.482758620689655
	AB/HR	52.155692195926800	29.3888888888889
	OBP	0.309827327474848	0.308952676790919
	BlwOBA	0.185826458	0.185328047
Non-Replacement, Non-Incinerated Batters (n = 151)	SLG	0.447052806856611	0.432297965915338
	BA	0.269805906301752	0.269145765176300
	BABIP	0.282505294545485	0.282247284878864
	K%	0.136084281198298	0.135443668993021
	BB%	0.642419253535039	0.471217105263158
	AB/HR	53.9315483587587	29.477272727272700
	OBP	0.311984718386041	0.308952676790919
	BlwOBA	0.187536869	0.186321825
Incinerated Batters (n = 21)	SLG	0.4329688721859	0.426923076923077
	BA	0.262090471090517	0.279245283018868
	BABIP	0.273230521324193	0.272727272727273
	K%	0.11752363037374	0.1
	BB%	0.70798906569175	0.626577840112202
	AB/HR	45.8627604166667	25.775
	OBP	0.309010131298295	0.321875
	BlwOBA	0.185017148	0.203442308
Replacement Batters (n = 29)	SLG	0.414849487715728	0.424019607843137
	BA	0.252558006912412	0.25856697819314
	BABIP	0.262580733185137	0.272727272727273
	K%	0.136151330249567	0.142857142857143
	BB%	0.537076491780635	0.448275862068966
	AB/HR	45.1404903378571	29.3333333333333
	OBP	0.293934460432795	0.305882352941176
	BlwOBA	0.175582095	0.182136076

Table 7 Various Sample and Population Pitching Appearance Neutral Performance-based Pitching Statistics

Sample	Statistic	Mean	Median
Season 3 Adjusted Pitching Population (n = 106)	Avg RD / Loss	3.22	3.13
	ERA	5.60	5.12
	FIP	6.01	5.90
	WHIP	1.40	1.37
Non-Replacement, Non-Incinerated Pitchers (n = 92)	Avg RD / Loss	3.26	3.17
	ERA	5.19	4.99
	FIP	6.06	5.93
	WHIP	1.39	1.36
Incinerated Pitchers (n = 6)	Avg RD / Loss	2.91	3.00
	ERA	11.01	4.39
	FIP	4.89	4.75
	WHIP	1.43	155
Replacement Pitchers (n = 8)	Avg RD / Loss	2.92	3.08
	ERA	6.26	6.11
	FIP	6.27	6.22
	WHIP	1.57	1.55