An Analysis of Baseball Statistics

Jack McKay
Period 1
June 2005
TJHSST

Abstract

For years, baseball theorists have pondered the most basic question of baseball statistics: which statistic most accurately predicts which team will win a baseball game. With this information, baseball teams can rely on technological, statistical-based scouting organizations. The book, Moneyball addresses the advent of sabermetric statistics in the 1980s and 1990s and shows how radical baseball thinkers instituted a new era of baseball scouting and player analyzation. This project analyzes which baseball statistic is the single most important. It has been found that new formulas, such as OBP, OPS, and Runs Created correlate better with the number of runs a team scores than traditional statistics such as batting average.

Findings: Season Simulation

	AB	H	RBI	BB	2B	3B	HR	TB	AVG	OBP	SLUG	OPS	RC
Orioles													
Roberts	748	253	108	102	46	15	40	449	0.338	0.418	0.600	1.018	0.251
Mora	712	244	102	123	48	5	34	404	0.343	0.440	0.567	1.007	0.249
Tejada	758	273	163	53	46	6	46	469	0.360	0.402	0.619	1.021	0.249
Sosa	760	194	145	36	34	1	51	383	0.255	0.289	0.504	0.793	0.146
Palmeiro	674	174	80	106	38	0	21	275	0.258	0.359	0.408	0.767	0.146
Lopez	712	224	125	44	47	3	40	397	0.315	0.354	0.558	0.912	0.198
Bigbie	709	193	116	33	33	3	45	367	0.272	0.305	0.518	0.822	0.158
Gibbons	689	196	65	38	35	5	25	316	0.284	0.322	0.459	0.781	0.148
Matos	671	185	88	35	28	3	26	297	0.276	0.312	0.443	0.754	0.138
Totals	6433	1936	992	570	355	41	328	3357	0.301	0.358	0.522	0.880	0.187
Yankees													
Jeter	744	282	131	104	43	12	40	469	0.379	0.455	0.630	1.086	0.287
Matsui	740	241	147	96	52	3	49	446	0.326	0.403	0.603	1.006	0.243
Rodriguez	743	254	178	70	59	5	61	506	0.342	0.399	0.681	1.080	0.271
Sheffield	750	242	154	44	37	3	60	465	0.323	0.360	0.620	0.980	0.223
Posada	718	159	89	60	25	1	34	288	0.221	0.281	0.401	0.683	0.113
Giambi	671	158	99	93	33	1	34	295	0.235	0.329	0.440	0.768	0.144
Williams	679	173	80	66	28	1	22	269	0.255	0.321	0.396	0.717	0.127
Martinez	653	171	68	70	35	1	26	286	0.262	0.333	0.438	0.771	0.146
Womack	669	199	58	35	29	11	15	295	0.297	0.332	0.441	0.773	0.147
Totals	6367	1879	1004	638	341	38	341	3319	0.295	0.359	0.521	0.881	0.187
Jays													
Rios	797	217	94	38	29	9	26	342	0.272	0.305	0.429	0.734	0.131
Catalanot	714	208	79	97	43	2	25	330	0.291	0.376	0.462	0.838	0.174
Hillenbra	762	255	142	36	50	6	48	461	0.335	0.365	0.605	0.970	0.221
Koskie	722	200	119	62	50	3	42	382	0.277	0.334	0.529	0.863	0.177
Wells	704	190	92	58	36	1	32	324	0.270	0.325	0.460	0.786	0.150
Hinske	682	198	94	59	37	4	35	348	0.290	0.347	0.510	0.857	0.177
Hudson	659	200	78	59	38	7	24	324	0.303	0.361	0.492	0.852	0.177
Adams	663	158	70	44	27	5	17	246	0.238	0.286	0.371	0.657	0.106
Zaun	603	183	88	85	39	1	26	302	0.303	0.390	0.501	0.890	0.195
Totals	6306	1809	856	538	349	38	275	3059	0.287	0.343	0.485	0.828	0.166

	Wins	Losses
Orioles	77	85
Yankees	95	67
Jays	71	91

Theory

The statistics which place an added emphasis on (1) walks taken or (2) hitting for power will prove to be far superior than statistics that do not have such an emphasis.

1. The importance of the walk: If you have ever played baseball, you may remember the frustration of having the opposition have a home run after your pitcher walked several batters. You might wonder why the pitcher couldn't just walk the batters after the home run, so that they wouldn't count as runs. Indeed, in the previous situation, a walk is as good as a home run. In fact, most of the time a player walks and then goes on to score, the player might as well have hit a home run – the end result would be the same. The difference between a batter getting out and not getting out is far more important than the difference between a player getting on base with a walk and a player getting on base with a hit.

2. The importance of hitting for power: After getting on base, hitting for power is most important. Hitting for power is necessary to convert the runners on base into runs. Therefore, the importance of the walk and power hitting are somewhat dependent – their importances are derived from each other. Walks are so important because of the possibility of the baserunner coming around to score, most likely on an extra base hit. Likewise, power hitting is so important not because it helps the hitter himself score (the hitters afterward are responsible for that), but because it helps the runners already on base score.

Method

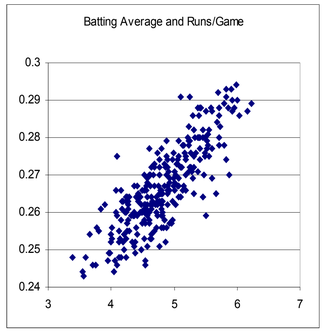
My analysis had two parts. First, I obtained statistical data about baseball teams in the past ten years and entered it into a Microsoft Excel spreadsheet. I calculated the correlation between certain statistics of teams and the number of runs they scored that season.

The second part of my research consists of a computer program in C++. The user can pick the number of games he wishes to simulate. The program displays the percent of games in which a certain statistic accurately predicts the winner of the game.

Findings: 100 Games Simulation

The winning team in each team had a higher OPS 87 times and a lower OPS 13 times.
The winning team in each team had a higher SLUG 84 times and a lower SLUG 16 times.
The winning team in each team had a higher AVG 81 times and a lower AVG 19 times.
The winning team in each team had a higher OBP 80 times and a lower OBP 20 times.
The winning team in each team had a higher Runs Created 82 times and a lower Runs Created 18 times.

Correlation Charts Based on Real Data



Conclusion

The best statistic is the one with the most criterion validity. It is possible for a statistic to calculate the percentage of the time a baseball players hits a home run on a Tuesday. But, no statistician would pay attention to the statistic. It does not measure anything important, and therefore has limited predictive ability. Furthermore, a player who excels in this statistic may not help his team win games – the statistic is too obscure.

When dealing with common statistics, statisticians should be similarly concerned. Batting average has been the statistic of choice for the past century. However, my analysis of statistics from a decade of baseball seasons and from a computer baseball simulation illustrate the problem with using batting average to evaluate a hitter. From the correlation charts, it's evident that one team hit .275 for an entire season and only averaged four runs per game. Meanwhile, another team hit .260 for an entire season and averaged five-and-a-half runs per game. This fluctuation demonstrates the misleading nature of the batting average statistic.

Introduction to Sabermetrics

For some time, a baseball debate has been brewing. Newcomers and sabermetricians (the "Statistics Community") feel that baseball can be analyzed as a scientific entity. The Sabermetric Manifesto by Bill James serves as the Constitution for these numbers-oriented people. Also, Moneyball by Michael Lewis serves as the successful model of practical application of their theories. Traditional scouts (the "Scouting Community") contend that baseball statistics should not over-analyzed and stress the importance of intangibles and the need for scouts. The debate can also be interpreted in terms of statistics. Baseball lifers feel that stats such as batting average are the most important. Meanwhile, the Statistics Community feels that complex, formulaic stats can better predict a player's contributions to a team. The discussion continues in the offices of baseball teams around the country: are computer algorithms better than human senses?

From a statistical sense, baseball is an ideal sport. Plate appearances are discrete events with few, distinct results. In fact, results can be limited to a few distinct outcomes: hit, walk, or out. Outcomes can also be expressed more specifically: single, double, triple, home run, walk, strike-out, fly-out… etc. Most importantly, the outcomes of past plate appearances can accurately predict the outcomes of future plate appearances. Baseball statisticians continue to desire more information in their field in order to become better at analyzing the past and predicting the future.

This project concerns itself with testing the sabermetric statistical subtheory. I will identify the baseball statistic that best measures and correlates with run production.

Overall, the correlation charts and the bar chart illustrate that OPS and Runs Created are both better statistics than batting average and should be used more prominently.

Definition of Terms

BA – Batting Average
OBP – On Base Percentage
OPS – On Base Percentage Plus Slugging
OPS Adjusted – On Base Percentage * 1.2 Plus Slugging Percentage
Runs Created – On Base Percentage * Slugging Percentage