Baseball Statistics, Bootstrap Methods and the Boston Red Sox

Andrew R. Willis

11th September 2001

1 Introduction

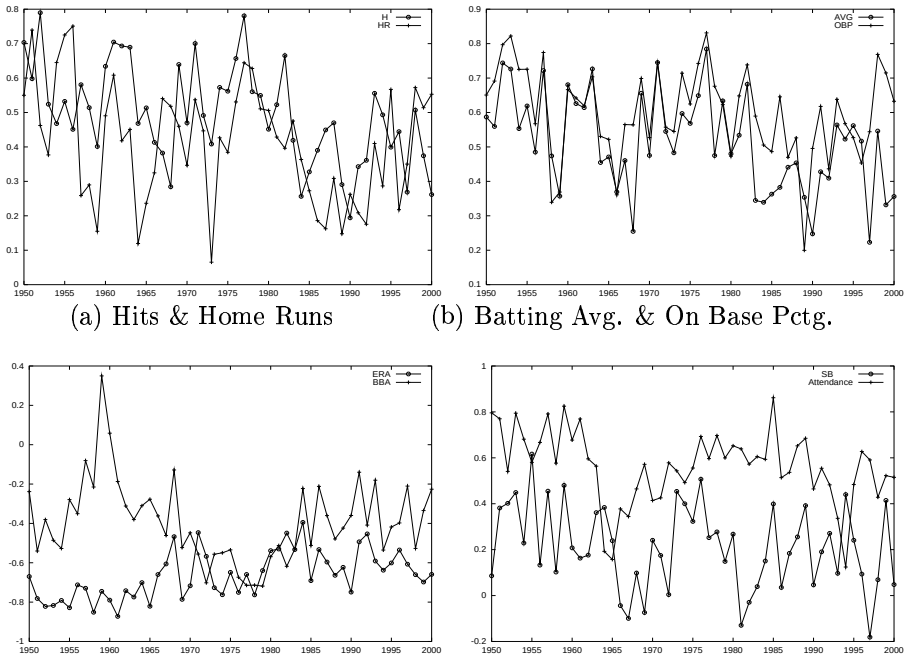
This report makes use of various methods of non-parametric statistics to invesigate some hypothesis relating to the sport baseball. Sports statistics are computed for almost all popular sports where they are utilized to estimate the success for athletes and entire teams. In sports, statistics are computed at the lowest level for each player. The team statistics are then formed as a function of the statistics of each member of the team. This report will concentrate solely on team statistics. Team sports statistics generally fall into one of 2 categories:

1. Offensive statistics.
2. Defensive statistics.

In baseball, offsensive statistics relate to batting where a team has the opportunity to score runs. Defensive statistics relate to pitching and fielding where a team is attempting to prevent runs from scoring. In this case, offensive statistics are well separated from defensive statistics, in the sense that for a given team there is no possibility of affecting offensive statistics such as runs scored while on defense and vice-versa. These statistics will be used to investigate 3 major conjectures.

1. What is the *best* single baseball statistic?
2. Is talent equally distributed in American and National Leagues?
3. How much variability is there in the outcome of a single season?

Two of these conjectures relate to baseball in general. The third uses bootstrapping methods to estimate the variability of the possible outcomes for a given season. Throughout the report we will concentrate on the win statistic for a team, w. The win statistic is the number of games won by a baseball team during regular season play (i.e. it does not include playoff games). Since teams play an equal number of games throughout each baseball season, the win statistic directly measures a teams success in winning baseball games. The win statistic will play a pivotal role in analysis of (1), (2) and (3).



(a) Hits & Home Runs (b) Batting Avg. & On Base Pctg.

(c) Earned Runs & Walks (d) Stolen Bases & Attendance

Figure 1: Correlation Coefficients of games won and various other statistics

2 *Best* Baseball Statistics and Correlation

Almost all statistics in sports relate either directly or indirectly to the teams success. Hence, it is clear that for a single team and season many of these statistics should correlate well with the win-loss record. A somewhat more subtle question is to determine those statistics which seem to correlate more strongly. To this end, correlation coefficients were computed between a set of offensive and defensive statistics and the number of team wins. The null hypothesis is that the *best* defensive statistic is more strongly correlated to winning than the *best* offensive statistic. The word *best* here is used to denote the statistic which has the highest mean correlation to the number of team wins. Initially work is done to identify the offensive and defensive statistic which has the highest mean correlation coefficient to the win statistic. Further tests are then applied to these two to determine which has the higher mean correlation to team wins. In addition, some miscellaneous statistics were correlated to team wins such as game attendance. Statistics such as this cannot be classified as either offensive or defensive.

The input data set consisted of many team statistics for all baseball teams in the major leagues for the years 1950-2000. For each year the correlation coefficient between all offensive and defensive statistics and team wins was computed.

Generally, when computing offensive statistics, one expects to have a positive correlation, (i.e. when batting average for team A is high, the number of team A wins is high). Conversely, for most defensive statistics, one expects to find a negative correlation, (i.e. when the average earned runs against a team A is low, the number of team A wins is high). Of course, this is not true for all offensive or defensive statistics. To adjust for this, all defensive statistics were made to have a negative mean correlation and vice-versa for offensive statistics.

Correlated Statistics	Null Hyp	Significance	Fig. 1 Index	Conclusion
H/HR	Accept	NA	(a)	$E(H) \leq E(HR)$
OBP/AVG	Reject	0.05	(b)	$E(OBP) > E(AVG)$
BBA/ERA	Reject	0.05	(c)	$E(BBA) > E(ERA)$
Attendance/SB	Reject	0.05	(d)	$E(Attendance) > E(SB)$

Table 1: Sign Test on statistics correlation

From the resulting values, the 2 statistics which had the highest and lowest mean correlation value were extracted. The highest average value corresponds to the offsensive statistic with the strongest correlation and the lowest average value corresponds to the defensive statistic with the strongest correlation. Figure 1 shows the correlation coefficients found for some of the strongest correlated statistics over the years 1950-2000.

Having identified those statistics which are most correlated with team wins, the sign test was applied to see if there are significant differences in the mean correlation values for these statistics. The corresponding hypotheses for this test are given below:

Null Hypothesis: $E(Correlation(statistic\,A)) \leq E(Correlation(statistic\,B))$

Alternate Hypothesis: $E(Correlation(statistic\,A)) > E(Correlation(statistic\,B))$

Table 1 summarizes the results obtained from the test.

From Table 1, we may conclude that

1. For offensive statistics, the on base percentage (OBP) is more highly correlated with the number of game wins than the batting average (AVG) .
2. For defensive statistics, the earned runs against (ERA) a team is more highly correlated with the number of game wins than the number of walks allowed (BBA) by the team.
3. Attendance was more highly correlated with game wins than the number of stolen bases (SB).
4. The mean correlation for the number of hits (H) and the number of home runs (HR) are indistinguishable using this test. This implies that over a year, the number of hits and the number of home runs are a similar measure of the number of team wins.

As shown in the figure 1, the year to year values of each of these statistics tends to have a high degree of variability. Table 2 shows the confidence intervals for some of the statistics considered assuming a normal distribution of the correlation coefficient. From the table we can see that many of the intervals are large and overlapping, which makes it difficult to draw conclusions by comparison of the global data. However the plots from Figure 1 show that for a given year the values of one statistic is consistently above or below the value of the other statistic. Consequently, even though the confidence intervals for the true value of a two different correlation coefficients may overlap, we may still apply tests to determine whether there is a significant difference between the means of two statistics.

Correlated Statistic	95% Confidence Interval
OBP	(0.337147, 0.862854)
AVG	(0.233844, 0.792644)
ERA	(-0.902997, -0.426032)
BBA	(-0.804959, 0.021598)
Attendance	(0.247077, 0.881218)

Table 2: Confidence Intervals for some correlated statistics

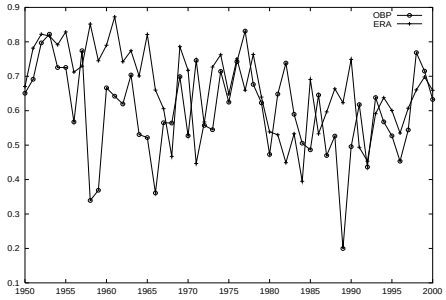


Figure 2: $|Correlation(ERA)|$ and $Correlation(OBP)$

Further tests were computed to compare the OBP and ERA statistics to determine which has the strongest mean correlation to team wins. Since, as mentioned before, ERA is negatively correlated, we compare $|ERA|$ to OBP. A one-tailed sign test was applied on these statistics in order to determine which statistic has a higher mean correlation value.

Null Hypothesis: $E(|Correlation(ERA)|) \leq E(Correlation(OBP))$

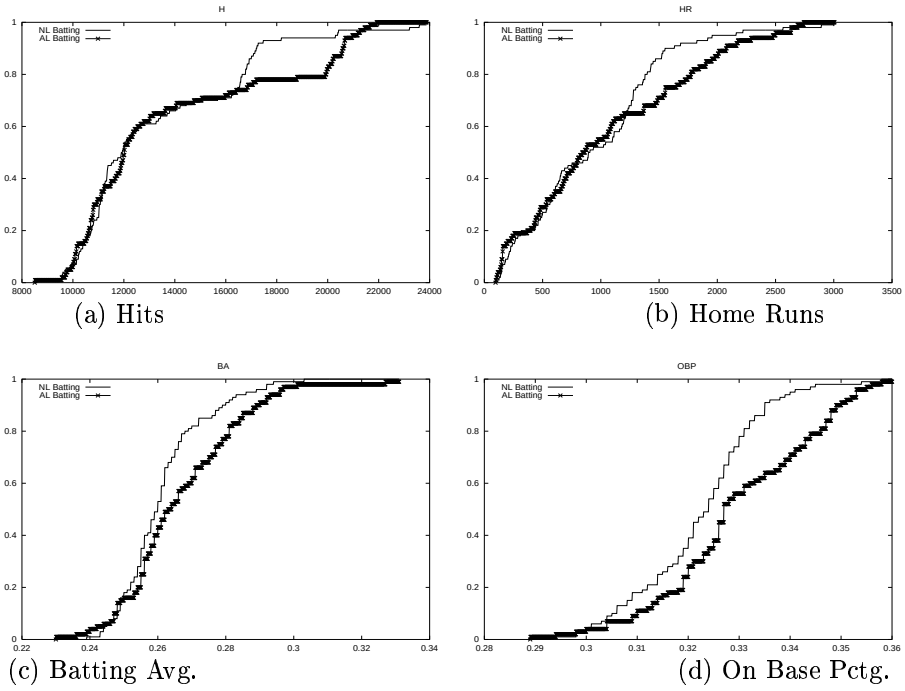
Alternate Hypothesis: $E(Correlation(|ERA|)) > E(Correlation(OBP))$

Figure 2 plots the correlation coefficients for the two statistics for the years 1950-2000. The null hypothesis in this case was rejected with significance 0.05, showing that the mean value of the ERA statistic is greater then the mean value of the OBP statistic. Hence, we may conclude that the ERA statistic is the *most* effective single statistic for measuring the amount of team wins. This confirms the initial hypothesis that the best defensive statistic is more strongly correlated with team wins than the best offensive statistic.

3 American and National Leagues

The goal of this section is to analyze the distribution of talent between the two leagues in major league baseball. This is accomplished via a test for equality of distribution for various statistics in each of the leagues. The assumption here is that since team statistics are a function of individual player statistics, for equal distributions of player talent in both leagues, the resulting distributions of team statistics in each league should be identical. The Komolgorov-Smirnov test was applied here instead of the chi-squared test since the data is continuous and the binning methods involved in the chi-squared test is more well suited to situations where the data takes on a discrete set of possible values.

The Komolgorov-Smirnov test makes use of the empirical cumulative distribution function (CDF) for the observed values



(a) Hits (b) Home Runs

(c) Batting Avg. (d) On Base Pctg.

Figure 3: Empirical distributions for offensive statistics

of a random variable X which we will denote as $F_X(x)$. Given 2 random variables A and B and a set of observed values for each, the two-tailed Kolmolgorov-Smirnov test has the following hypotheses:

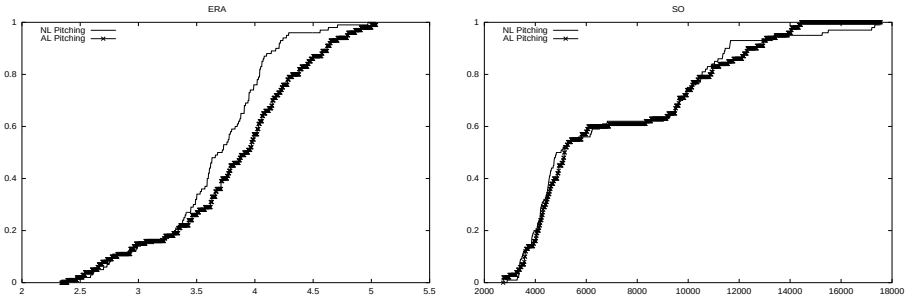
Null Hypothesis: $F_A(x) = F_B(x)$

Alternate Hypothesis: $F_A(x) \neq F_B(x)$

The empirical distributions of the offensive statistics were computed for the years 1901-2000, they are shown in Figure 3. In all but 2 of the cases examined, the Komolgorov-Smirnov test accepted the null hypothesis, indicating that the distributions in each league were close. However, in cases (c) and (d) the null hypothesis was rejected with significance 0.05. This corresponds to the statistics BA (referred to previously as AVG) and OBP respectively. In section 2, we noted that OBP was the *best* offensive statistic for predicting the number of team wins for a given season. In addition, we also found BA (AVG) to be one of the strongest correlated offensive statistics. Hence, this difference in OBP may indicate a difference between talent in the two leagues. The same trend is apparent in both of the statistics: for higher values of each statistic, the American League has more probability mass. Histograms of these statistics supporting this statement are avaiable in the Appendix, Figure 6. This is observable in the CDF plots since for the National League, the CDF is closer to 1 at higher values of these statistics than the American League. This leads us to believe that the American League has a higher mean statistic value for BA and OBP than the National League. One may conclude that the American League tends to have better offensive players (i.e. more offensive talent) than the National League.

To confirm this conjecture, the sign test was applied for BB, BA, and OBP statistics.

Null Hypothesis: $E(American\,League(X)) \leq E(National\,League(X))$



(a) Earned Run Average (b) Strike Outs

Figure 4: Empirical distributions for defensive statistics

Alternate Hypothesis: $E(American\,League(X) > E(National\,League(X))$

In all cases, the null hypothesis was rejected with significance 0.05, confirming that the American League has a higher mean value for each of these offsensive statistics.

Identical procedures were applied on the defensive statistics ERA and SO (strike outs). Figure 4 shows the empirical distributions of these statistics for the years 1901-2000. In this case, the null hypothesis is rejected for the ERA statistic.

This statistic reflects the results found for the offensive statistics. Since the American League was found to have a higher on-base percentage and mean batting average, one would expect American League defensive statistics such as the earned runs against (ERA) each team to be higher. Figure 4 clearly shows that the National League has more probability mass for smaller values of the ERA statistic. Histograms of these statistics supporting this statement are avaiable in the Appendix, Figure 6. This indicates that the mean ERA statistic value for the American League may be larger than the mean ERA statistic for the National League. This was validated with significance 0.05 as done for the offensive statistics. Since a high ERA indicates less wins, we must conclude that the National League has tends to have more defensive talent than the American League.

Hence, from the tests stated above we can conclude the following:

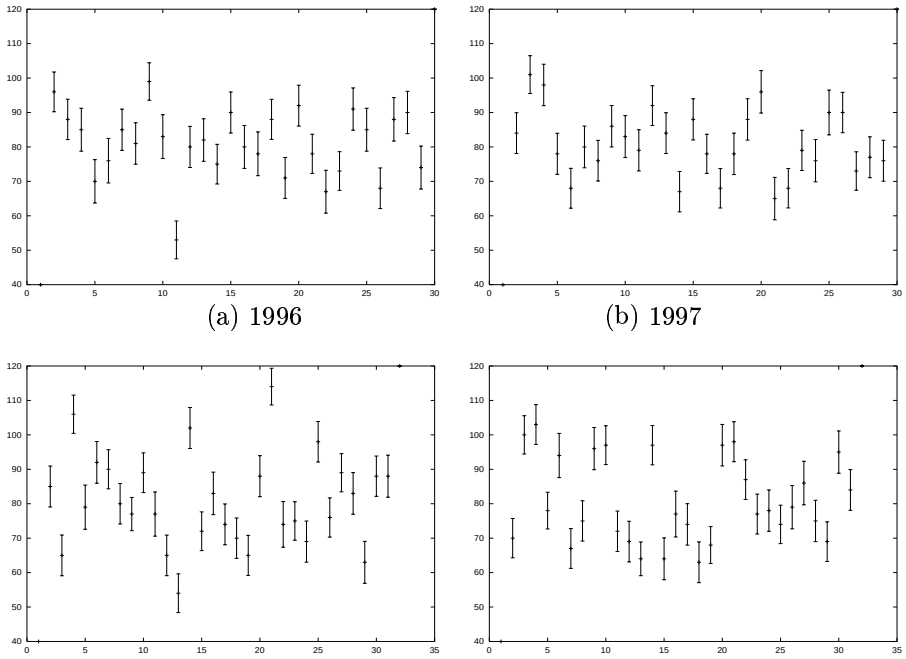
1. Offensive statistics indicate that more offensive talent is present in the American League than the National League

2. Defensive statistics indicate that more defensive talent is present in the National League than the American League.

Since in section 2 we decided that the best defensive statistic was a better indicator of team wins than the best offensive statistic, one could conjecture that the National League may have an edge in terms of winning baseball games. However, detailed treatment of this conjecture is not covered in this report.

4 Bootstrapping new seasons

In this section bootstrap methods were applied to simulate alternate possible outcomes to a single season. The chosen test statistic is the total wins for a given team throughout the season. The bootstrapped sample sets are drawn from a data set



(a) 1996 (b) 1997

(c) 1998 (d) 1999

Figure 5: Bootstrap results for 300 bootstrap sample sets

which consists of the win-loss record for every game of team A vs. team B throughout a single season. Datasets were available for the years 1996-1999.

In this case, the bootstrap method needed a small adjustment due to the nature of the data. For each N game series between team A and team B a bootstrap sample of X wins by team A precipitates N-X wins for team B. Hence, bootstrapping team A alone also determines part of the bootstrap sample for team B, which does not fit into the typical bootstrapping situation.

In Figure 5 the observed statistic value is plotted for the 1996-1999 baseball seasons for each team in the major leagues along with the associated standard error, $\widehat{se}_{boot}(w)$, estimated via the bootstrap method. For almost all situations, $\widehat{se}_{boot}(w)$, had a value between 5 and 6.

Figure 5 also shows that even though the variation of the win statistic is roughly the same for each team, the variance of the win statistic over all teams can change drastically from year to year. In years 1996 and 1997, we see that most teams have approximately 81 wins. However, for the year 1998 there is a large amount of variability of wins for each team with one team having an observed value of 114 wins. In 1999, there seems to be less variability than the previous year, yet, it seems like there are two strata of teams: those with approximately 93 wins and those with around 77 wins.

The Boston Red Sox win statistics are plotted as $x = \{3, 1, 4, 5\}$ for the $years = \{1996, 1997, 1998, 1999\}$. Since teams are listed alphabetically from left to right, the change in position was caused by team additions and name/location changes during the years analyzed.

From Table 3, we can see that in all years of consideration the Red Sox are consistently above the mean, indicating that

Year	Red Sox Wins	Division Position	League Avg. Wins	Avg. Wins of League Division Ldrs
1996	85	3	80.929	92.667
1997	84	4	80.929	91.500
1998	92	2	81	99.500
1999	94	2	80.900	98.333
2000	85	2	80.933	93.333

Table 3: The win statistic and the Boston Red Sox

recently they have consistently fielded better than average baseball teams. However, for the years considered they never win the first position in the American League East Division. Four of these years (all but 1997), the New York Yankees took first place in the division. It would be interesting to make a detailed comparison of the statistics for these two teams to detect possible reasons for this trend. However, the computed variability in the win statistic suggest that the Red Sox are consistently 1 standard deviation away from being classified as a Division Leader. Hence, to make a bid for the playoffs and possibly the World Series, the Red Sox need to improve their team statistics by improving their player roster.

5 Conclusions

Section 2 applied data from 1950-2000, to conclude that the offensive statistic which is most correlated with team wins is the on base percentage (OBP) for the team. We also found that the defensive statistic which is most correlated to team wins is the earned run average (ERA). These two statistics were then compared against each other and it was found that the ERA statistic had the highest correlation with team wins from all offensive and defensive statistics considered.

Section 3 used data from years 1901-2000, to conclude that the distribution of the offensive statistics OBP and batting average (BA) are different between the American and National Leagues. Further, the American League teams tend to have higher statistic values than National League teams. On the other hand, the distribution of defensive statistic ERA was found to be different between the two leagues. Here it was found that the National League teams tend to have lower values of this statistic than American League teams. This suggests that the American League has more offensive talent than the National League and vice-versa for defensive talent. Since low values of the defensive statistic ERA was found in section 2 to have the highest correlation to team wins we made the conjecture that perhaps the National League has a slight edge in terms of having more team wins.

Section 4 used data from years 1996-1999 to apply bootstrap methods to analyze the variability of the win statistic within a single season. It was found that the win statistic varies similarly for most all teams regardless of their position. This variation seemed to be about $\pm[5-6]$ games for all of the teams. This was followed up with a short discussion about the Red Sox team. In this case, we found that although the Red Sox consistently perform better than the league average by about 1 std. dev. they also perform below the first place team by about the same amount. Hence, it seems that they are consistently a good team, yet not quite good enough to capture first place. It was concluded that the Red Sox need to improve their team statistics in order to become a contender for a Division or World Series Championship.

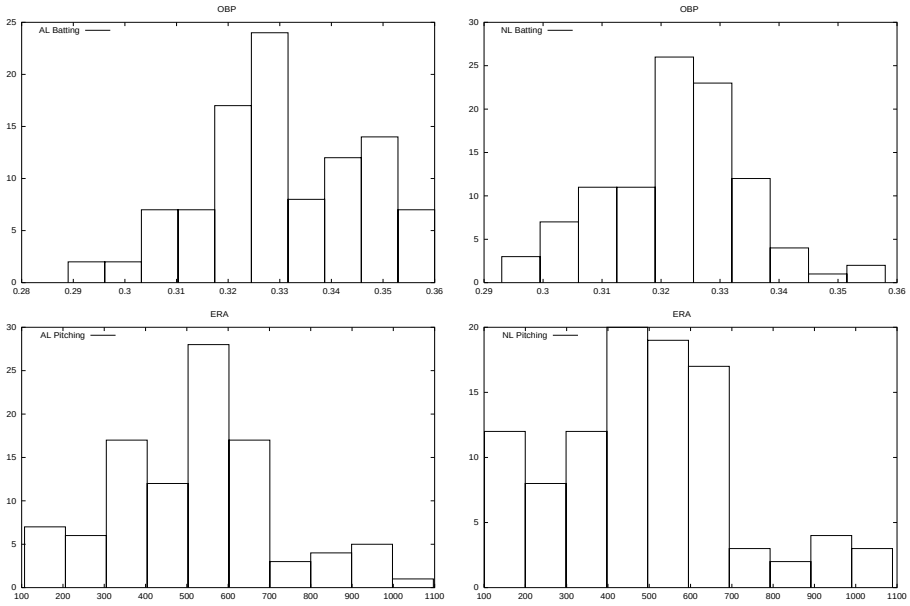


Figure 6: Histograms of American and National League OBP and ERA statistics.

6 Appendix

6.1 Data Sources

Archive containing team statistics used in Section 2

http://www.baseball1.com

Archive containing league statistics used in Section 3

http://www.baseball-reference.com

Archive containing game-by-game logs used in Section 4

http://www.baseball-reference.com

6.2 Statistic Definitions

Abbreviation	Definition	Statistic Classification
Team	Team Name	—
Lg	League (National or American)	—
Div	Division (3 Divisions in each League)	—
Pos	Teams Division Position in Final Standings	—
G	Regular Season Games Played	—
W	Games Won	—
L	Games Lost	—
GB	Games behind the division leading team	—
R	Runs scored	offensive
RA	Opponents Runs scored	defensive
AB	At bats (offensive)	offensive
H	Hits by batters (offensive)	offensive
2B	Doubles (offensive)	offensive
3B	Triples	offensive
HR	Home Runs	offensive
BB	Walks by batters	offensive
HBP	Batters hit by pitch	offensive
SF	Sacrifice flies	offensive
SO	Strike outs	offensive
AVG,BA	Batting average	offensive
OBP	On base percentage	offensive
SLG	Slugging percentage	offensive
SB	Stolen bases	offensive
CS	Players caught stealing	—
SHO	Shutouts	defensive
SV	Saves	defensive
IP	Innings Pitched	defensive
ER, ERA	Earned Runs allowed	defensive
HA	Hits allowed	defensive
HRA	Home runs allowed	defensive
BBA	Walks allowed	defensive
SOA	Strikeouts by pitchers	defensive
Attendance	Stadium Attendance	—