The Rules of Victory: How to Win a Baseball Game

Beomsoo Kim, and Youngmin Kim

Abstract—There have been a lot of data visualization for sports, especially in baseball. However, most studies focused relationships between payment and performance, or relationships among baseball statistics. There is no research about relationships between win rate and baseball statistics, even today. Therefore, we want to figure out the most important baseball statistics to win a baseball game in KBO league (Korean pro-baseball league) using data visualization. In this paper, we extract recent several years baseball statistics such as win rate, batting average and earned run average. Then we analyze the correlations between win rate and some of baseball statistics using Pearson correlation analysis method. We visualize those in radar chart. Finally, we discuss about the result of data visualization and conclude our goal, the most important baseball statistics to win a baseball game.

Index Terms—Baseball, Korean baseball league, win rate

1 Introduction

Recently, in Nov. 2nd, 2016, Doosan Bears won the Korean Series (Championship Series of Korean pro-baseball league, KBO). That was their two-year consecutive victory and their 5th Korean champion title so far. They showed the overwhelming power in not only regular season, but also Korean Series. As a result, people say that the Doosan dynasty finally arrived. Actually, there were the teams that ruled the league and were called dynasty in KBO league. Haitai Tigers in the 80s and 90s, Hyundai Unicorns in early 2000s, SK wyverns in late 2000s, and Samsung Lions in early 2010s are them. From Haitai to Doosan, however, they each had different features. For example, SK Wyverns had powerful pitching strength but also had relatively weak batting strength. And Doosan Bears have fantastic starters and hitters but also have poor bullpen pitchers.

Here, we have a question. How to play baseball well? That is, how to win a baseball game? Even if we do not go far, the question is amplified in this year. Samsung Lions has batting average of 0.293 that is the 3rd place in the league, and Lotte Giants records 1009 strike-outs (for pitchers) that is the 2nd place in the league. However, they actually take 9th and 8th places, respectively in the league[1]. Therefore, we carry out research to find the answer of this question.

There have been many studies about baseball analysis. Scully studied the relationship between pay and performance of some baseball players in major league baseball [8]. Sommers and Quinton also did a research about pay and performance of players in major league baseball in case of the first family of Free Agents [10]. Few decades later, Hall and Szymanski studied the causality between team performance and payroll in the case of major league baseball and English soccer [6]. Finally, Averbukh and Brown did a research about baseball pay and performance again recently [1].

Lets focus on visualization rather than analysis. First, Hakes and Turner studied about pay, productivity, and aging in major league baseball [5]. That is, the visualization of the causality between the payment and performance of some baseball players, as aging. They represented data as lots of dots in x-y plane and the straight line which express the trends of dots. Second, Yagi and Funayama studied about the digitizing of the characteristics and visualization of the condition of the Japan pro-baseball hitter [12]. In this paper, they analyzed the condition of batters over one season. They also express data in typical method such as bar chart and broken line chart. Finally, Kono and Yamamoto did a research about visualization of baseball players defensive range using a probability ellipse [7]. They express the data as a lot of dots in the baseball park which is drawn as computer graphic.

Actually, our question - How to win a baseball game? - is about win which is the most basis of the sport. As you can see in above, however, we cannot find any paper which covers it. Most of the papers dealt with the relationship between the salary and the individual statistics, or the specific internal data of a baseball game. In other words, we cannot find that how one team can make a win in baseball game, rather than one player. As a result, we decide to get the answer by ourselves.

2 Data extraction and processing

2.1 Data Extraction

First of all, we visit the official KBO league site and the verified KBO league record site, statiz[2] that is usually quoted in news articles, to get the data of baseball. There are too much data and we cannot extract all. Because this work is not only inefficient, but also needless. So we have to decide the criterion to select which of baseball statistics. Our conclusion is simplicity. We want to make the visualization which is easy to understand for not only baseball lovers, but also the people who do not know anything of baseball. Therefore, we select the most basic and important statistics such as batting average (AVG) and earned run average (ERA). And we do not select too difficult statistics such as batting average on balls in play (BABIP) or defense-independent pitching statistics (DIPS). The baseball statistics we choose are as follows:

ERA Earned Run Average

WHIP Walks plus Hits divided by Innings Pitched

K/9 Strike-outs per 9 innings of pitchers

BB/9 Base on Balls per 9 innings of pitchers

HR/9 Home Runs per 9 innings of pitchers

K/BB Strike-outs divided by Base on Balls of pitchers

R_P Runs of pitchers

WAR_P total Wins Above Replacement player of pitchers (from statiz)

LOB% Left On Base Percentage of pitchers

AVG Batting Average of hitters

R_H Runs of hitters

OBP On Base Percentage of hitters

SLG Slugging percentage of hitters

OPS OBP + SLG

- *Beomsoo Kim, ECE of UNIST. E-mail: kbs9409@unist.ac.kr*
- *Youngmin Kim, ECE of UNIST. E-mail: bbonobono@unist.ac.kr*

Publication 05 Dec. 2016.
For information on obtaining reprints of this article, please send e-mail to: kbs9409@unist.ac.kr.

[1] Data source from KBO records (http://www.koreabaseball.com)

[2] Statiz.com (http://www.statiz.co.kr)

HR% Home Runs per plate of hitters

BB% Base on Balls per plate of hitters

K% Strike-outs per plate of hitters

BB/K Base on Balls divided by Strike-outs of hitters

RISP Batting Average when Runner In Scoring Position (2B or 3B) of hitters

WAR_H total Wins Above Replacement player of hitters (from statiz)

SB% Stolen Base Percentage

E/G Errors per games

Now, we have to consider the scope of years and teams. There are many choices: Collect the whole years data of KBO league (from 1982 to 2016) because KBO league has so short history. Include the disappeared teams data such as MBC Blue dragons and Ssangbangwool Raiders. However, there are not many remaining records and it is difficult to implement the visualization to handle disappeared teams. In conclusion, we choose recent 10 years, from 2007 to 2016. And the chosen 10 teams are Doosan Bears, Hanwha Eagles, KIA Tigers, KT Wiz, LG Twins, Lotter Giants, NC Dinos, Nexen Heroes, Samsung Lions, and SK Wyverns that exist now in KBO league.

2.1.1 Remarks

Since Nexen Heros founded in 2008, the data of 2007s Nexen Heroes is determined as Hyundai Unicorns which disbanded after 2007 season. Also, NC Dinos and KT Wiz entered the KBO league in 2013 and 2015, respectively. Therefore, those teams data are included only after 2013 and 2015, respectively.

2.2 Pearson Correlation Analysis

There are a lot of methods to analyze the data. Among those, we choose Pearson correlation analysis which is widely used in sciences. It was developed by Karl Pearson from a related idea introduced by Francis Galton [2, 3, 4, 9, 11]. The alternative method is regression analysis. As we will talk later, we draw a radar chart as a visualization. The radar chart shows many number of relationships, not the only one number of relationship. So we prefer just a number for each relationship rather than function, as a result of analysis. Since regression analysis results regression function, not some specific number, we choose Pearson correlation analysis rather than regression analysis. From now on, we will talk about what is PCC (Pearson Correlation Coefficient, ρ) and how to calculate it.

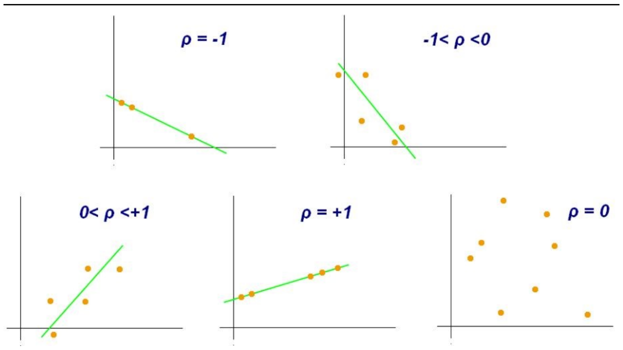


Fig. 1. Example of Scatter Diagrams with Different Values of PCC, ρ

PCC is a measure of the linear dependence between two variables, giving a value between -1 and +1 inclusive. +1 means that there is total positive linear correlation, 0 means that there is no linear correlation, and -1 means that total negative linear correlation (Figure 1). And the calculation of PCC is like in below[3]:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

where

- **cov** is the covariance
- σ_X is standard deviation X

The formula for ρ can be expressed in terms of mean and expectation. Since

- **cov(X,Y) = E[(X-μ_X)(Y-μ_y)]**

Then the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{E[(X-\mu_x)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{2}$$

where

- **cov** and σ_X are defined as above
- μ_X is the mean of **X**
- **E** is the expectation.

The formula for ρ can be expressed in terms of uncentered moments. Since

- $\mu_X = E[X]$
- $\mu_Y = E[Y]$
- $\sigma_X^2 = E[(X-E[X])^2] = E[X^2] - [E[X]]^2$
- $\sigma_Y^2 = E[(Y-E[Y])^2] = E[Y^2] - [E[Y]]^2$
- $E[(X-\mu_x)(Y-\mu_Y)] = E[(X-E[X])(X-E[X])] = E[XY] - E[X]E[Y]$,

The formula for ρ can also be written as

$$\rho_{X,Y} = \frac{E[XY]-E[X]E[Y]}{\sqrt{E[X^2]-[E[X]]^2}\sqrt{E[Y^2]-[E[Y]]^2}} \tag{3}$$

We find PCCs between win rate and each statistics for each years and teams using the method shown in above. And we visualize the result into radar chart described in below.

3 VISUALIZATION (DESIGN)

We need to visualize the analyzed data to understand easily. So we consider how to represent that effectively. We want to avoid the most general ways like line chart or bar chart because they are too boring and trite. Therefore, we want some fresh way of visualization which is colorful and has high readability.

Our choice is radar chart (Figure 2). It is more effective than line chart or bar chart in terms of space efficiency. Since we have 22 dimensions (statistics), line chart or bar chart should have long x-axis. However, radar chart occupies less space than those, because it can be drawn in limited space (in a circle). It is also efficient when comparing several graphs. Because they have different colors and are together in the same dimension (axis). However, it has limitations. It is hard to compare with other dimensions in opposite side in a graph. This is because radar chart is drawn in a circle. The other limitation is scalability. If we compare many number of graphs, it is hard to distinguish each other. Therefore, we cannot compare a large number of graphs. These are our remained future works.

[3] http://www.real-statistics.com/correlation/basic-concepts-correlation/

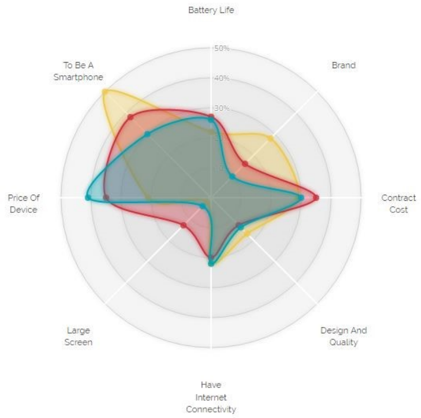


Fig. 2. Radar Chart in d3.js

We use already implemented version of radar chart in d3.js[4] (Figure 2). However, it is not suitable for our visualization and therefore we modify some codes. First, we change the axis grid format from percentage format to real number format. And remove other graphs (say, yellow and blue ones) because we need only one graph (say, red one) in combination mode. Then we modify radar lines interpolation type from cardinal-closed to linear-closed shape. Finally, we changed the number and name of axes to fit our data. As a result, our version of Radar chart is completed (Figure 3).

The actual implement codes consist of total 4 files:

index.html Main web page containing interactive tools, radar chart, and explanation of each baseball statistics,

RadarChart.js Implements the modified version of radar chart code,

script.js Bring data selectively according to web pages setting, calculate PCC, and do setting and function call of radar chart,

data.csv Raw data file and uploaded in web.

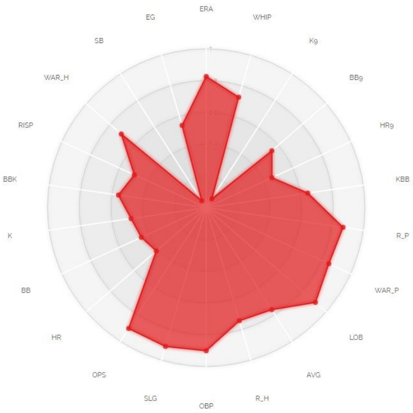
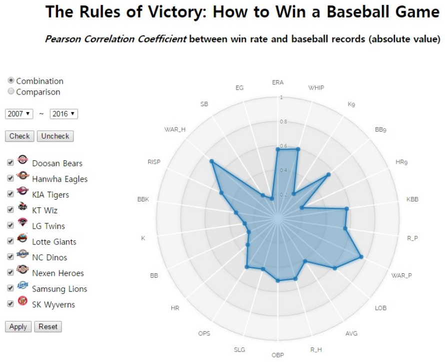


Fig. 3. Our Version of Radar Chart

The interactive setting tool is on left side of radar chart (Figure 4). First, there are two modes: Combination mode and comparison mode. In combination mode, the dataset of selected years and teams are drawn as only one graph of radar chart. Therefore, if we apply new dataset, the radar chart is renewed. In other words, the existing graph disappears (Figure 5). In comparison mode, on the other hand, the dataset of selected years and teams are appended to radar chart as an independent graph. Therefore, we can compare some graphs consist of datasets we want (Figure 6). The dataset setting part is under the mode selection part. We can choose start year and end year from 2007 to 2016. And we can select the teams from Doosan Bears to SK Wyverns. Check and uncheck buttons do check or uncheck all the boxes of teams. Finally, apply and reset buttons do apply this setting or reset all of the settings, literally. The complete version of radar chart is shown in (Figure 4, 5, 6):


The Rules of Victory: How to Win a Baseball Game

Pearson Correlation Coefficient between win rate and baseball records (absolute value)

● Combination
○ Comparison

2007 ▾ ~ 2016 ▾

☑ Doosan Bears
☑ Hanwha Eagles
☑ KIA Tigers
☑ KT Wiz
☑ LG Twins
☑ Lotte Giants
☑ NC Dinos
☑ Nexen Heroes
☑ Samsung Lions
☑ SK Wyverns

Fig. 4. Our Visualization (Final release)

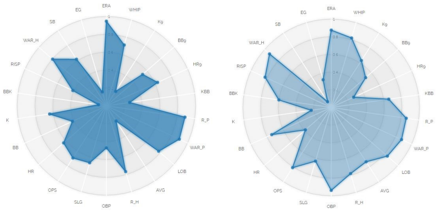


Fig. 5. Combination mode: the left one is graph of (2007-2007, 10 teams) and the right one is graph of (2008-2008, 10 teams)

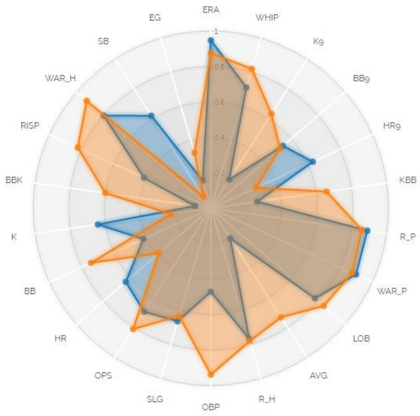


Fig. 6. Comparison mode: blue one is graph of (2007-2007, 10 teams) and orange one is graph of (2008-2008, 10 teams)

[4] From d3.js version of radar chart creator, Nadieh Bremer (http://bl.ocks.org/nbremer/)

4 Result & Discussion

Using this application, we can see not only the answer of our question, but also other additional phenomena. First, overall, the PCC values of pitchers are higher than those of hitters. Lets look at the Figure 7. It describes the largest dataset we can choose (2007-2016, 10 teams). The PPC values of pitchers like ERA, WHIP, and BB/9 are higher than those of hitters like AVG, OPS, and RISP. It is very interesting because it shows the answer of our question clearly. In other words, the pitcher statistics are more important to win a baseball game rather than the hitter statistics.

Second, the PPC values of WAR_P and WAR_H are significantly higher than others. Actually, it is not much surprising. In definition, WAR is an attempt by the sabermetrics baseball community to summarize a players total contributions to their team in one statistic[5]. Therefore, we already expected this phenomenon before the analysis as we can see in definition.

Third, the PPC values of AVG and E/G are quite low and that of BB/9 is quite high. It is surprising because we expected that AVG and E/G highly correlated with win rate before analysis. However, it was wrong actually. On the other hand, we expected that BB/9 less correlated with win rate compared to other statistics. It was also wrong.

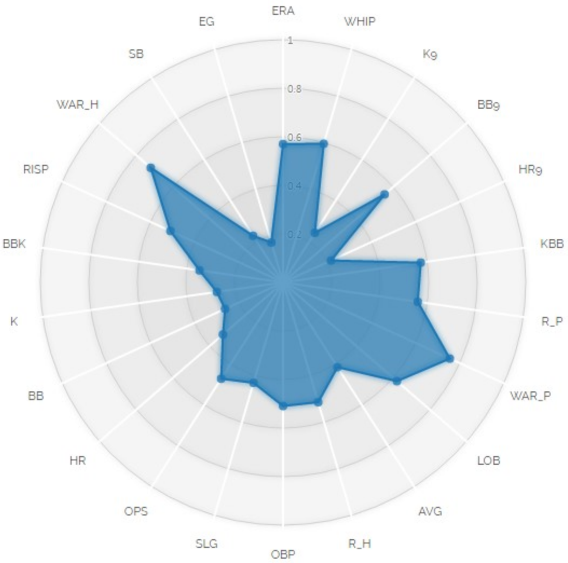


Fig. 7. radar chart (dataset: 2007-2016, 10 teams))

5 Conclusion

There are a lot of baseball statistics and they are relevant with win rate. According to our study, the statistics of pitchers are the most important to win a baseball game. There has been a dispute about that baseball is pitcher game or hitter game. In this paper, we figure out the answer of this dispute, baseball is pitcher game. And sabermetrics such as WAR highly connected to win rate. This is the projected result. Basically, they are made for that reason. However, AVG and E/G less correlate with win rate. On the other hand, BB/9 highly correlate with win rate. These two results are impressive.

In conclusion, we can say that if you want to win a baseball game, you should focus on defense rather than attack. Then, you can win a baseball game easily.

[5] Definition of ”WAR” from Fangraphs (http://www.fangraphs.com/library/war/)

Acknowledgments

KBO (Korea Baseball Organization) records (http://www.koreabaseball.com), statiz (http://www.statiz.co.kr), Charles Zaiontz / real-statistic (http://www.real-statistics.com), Nadieh Bremer (http://bl.ocks.org/nbremer/), and fangraphs (http://www.fangraphs.com).

References

[1] M. Averbukh, S. Brown, and B. Chase. Baseball pay and performance. 2015.

[2] F. Galton. Typical laws of heredity. *Nature*, 15(388, 389, 390):492–495; 512–514 532–533, April 1877.

[3] F. Galton. The british association: Section ii, anthropology: Opening address by francis galton, f.r.s., etc., president of the anthropological institute, president of the section. *Nature*, 32(830):507–510, Semtember 1885.

[4] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[5] J. K. Hakes and C. Turner. Pay, productivity and aging in major league baseball. *Journal of Productivity Analysis*, 35(1):61–74, 2011.

[6] S. Hall, S. Szymanski, and A. S. Zimbalist. Testing causality between team performance and payroll the cases of major league baseball and english soccer. *Journal of Sports Economics*, 3(2):149–168, May 2002.

[7] K. Kono and Y. Yamamoto. Visualization of baseball player's defensive range using a probability ellipse. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015), 2015 13th International Conference on*, pages 38–41. IEEE, 2015.

[8] G. W. Scully. Pay and performance in major league baseball. *The American Economic Review*, 64(6):915–930, December 1974.

[9] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, September 1996.

[10] P. M. Sommers and N. Quinton. Pay and performance in major league baseball: The case of the first family of free agents. *The Journal of Human Resources*, 17(3):426–436, 1982.

[11] S. M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2):73–79, 1989.

[12] K. Yagi, T. Funayama, and Y. Yamamoto. The digitizing of the characteristic and visualization of the wave of the condition of batter of the professional baseball. In *ICT and Knowledge Engineering (ICT and Knowledge Engineering), 2014 12th International Conference on*, pages 43–47. IEEE, 2014.