3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11643

Sample Size Sensitivity in Descriptive Baseball Statistics

John Kulas[1], Marlee Wanamaker[1], Diuky Padron-Marrero[1], Hui Xu[2]
[1]Department of Psychology, Montclair State University, Montclair, NJ, USA, [2]US Bancorp, Minneapolis, MN.

Abstract

This paper presents one element of a larger project that probes for systematic and predictable patterns of variability/volatility in baseball's descriptive statistics. The larger project standardizes many baseball indices along an event metric and provides relative estimates of each index's point of inflection toward an empirical asymptote. Specifically these estimates reflect deviations in sensitivity to "sample size" (e.g., which descriptive statistics are more or less robust across events). The end purpose of this broader investigation is a qualifier to be associated with such statistics: sample size sensitivity (Triple S). Not because it's needed, but because, colloquially, discussions of baseball statistics are commonly qualified by the cautionary statement, "well, it's a small sample size". The current presentation highlights the process and results of estimating the logarithmic event function of one statistic, batting average, and we will provide real-time projections of accuracy (our estimated function versus in-coming baseball data that occurs during the CARMA conference). Results have implications for the integration of BigData applications into digestable summary statistics that appeal to a broad-reaching audience with practical implications and meaning.

Keywords: *logarithmic function estimation; baseball data; predictive model.*

1. Introduction

Job performance is dynamic – e.g., it changes over time and context (see, for example, Sturman, 2003). Although the dynamic nature of performance has been acknowledged for a very long time, Kane (1996) was perhaps the first to propose that researchers should conceptualize job performance as a *distribution of outcomes*. Within the social sciences, this perspective was originally cited as impactful, but as of February 2020, this unique conceptualization of worker performance had only realized 76 academic citations. Recently, however, the potential of BigData to inform all aspects of work (including performance) has been met with a proliferation of interest and investigations (e.g., Campion, Campion, & Campion, 2018; Gunasekaran, Papadopoulos, Dubey, Wamba, Childe, Hazen, & Akter, 2017; Tonidandel, King, & Cortina, 2016), including the potential to revisit Kane's (1996) proposal of defining job performance via dynamic functional distribution.

Parallel to the emergent interest of BigData applications to organizational phenomena, in October of 2017 the Journal of Business and Psychology published a special issue dedicated to the interdisciplinary relevance of athletics and organizations. In their initial call for paper and subsequent introduction, the special issue editors noted "how studies of sports can readily be compared to and applied to the study and practice of work in organizations" (Gentry, Hoffman, & Lyons, 2017, p. 509). The current CARMA presentation integrates these two relative "newcomers" into the organizational sciences: athletics and BigData. Specifically, we resurrect Kane's (1996) perspective on performance distributions, leveraging baseball data to inform the modeling of performance over time.

Dalal, Nolan, and Gannon (2017) posed similar questions, tracking performance (goals, assists, and positive/negative differential) based on the occurrence or absence of previous shared experience with teammates (their sample was Olympic hockey players, permitting an estimate of players who had and hadn't previously been "teammates"). They noted the particular relevance of their sample to the construction and utilization of temporary teams used by traditional corporate organizations.

Similarly oriented, Heazlewood (2006) attempted to predict performance of Olympic swimming athletes. He found that nonlinear models were better predictors of performance than linear models. Results were also more accurate for races that were shorter distances. The predictions were made using mathematical models that predicted performance in 1996 and 1998 and were evaluated based on how closely they predicted performance in events that had not happened yet at the time of the analysis. These nonlinear models were again noted as patterns also likely to occur within more traditional worker contexts – perhaps having implications on the duration of work tasks and suggesting qualitatively different approaches toward modeling performance across different task periods.

Hofmann, Jacobs, and Gerras (1992) applied a historical equivalent to our current pursuit: "mapping" performance across time in two samples of baseball players. Their interest was in relative rank orderings scross time and the stability of such orderings. Due to data capturing limitations of the age, these authors were reliant on annual summary data from 204 professional baseball players presented within *The Baseball Encyclopedia* (1990). Their findings suggest a common nonlinear inverted-U shaped trajectory for offensive performance (batting average), with pitching data (earned run averages) exhibiting more linearity over time (e.g., ERAs deteriorated fairly consistently across years played). They note possible implications regarding patterns of performance for traditional workers across seniority and tenure.

The current CARMA presentation utilizes similar information as Hofmann et al. (1992), but does so with the advantage of contemporary data-capturing capabilities. Specifically, we capture *event-level* data (e.g., each pitch of a baseball) in an attempt to model differences across descriptive statistic stability. For the current presentation, we focus on one index of offensive performance: batting average. Across players and years, we model the functional degradation and eventual stability of this statistic, and use this empirical function to predict player performance during the July 8-9 conference period.

2. Methods

Play-by-play data from all regular season major league baseball games played from April 2008 to October 2015 was retrieved from baseballsavant via Bill Petti's database building script (https://billpetti.github.io/2018-02-19-build-statcast-database-rstats/). Each datafile contains approximately 700,000 individual *plays* – the most common form of play is a *pitch* (that is, the pitcher throws the baseball to his catcher, while a batter either attempts to swing or not). For the purposes of batting average, we collapsed these individual pitching plays into offensive player *at bats*. An *at bat* is a plate appearance that results in our focal event – the presence (1) or absence (0) of a "hit".

Each year, every offensive player begins with a simple batting average of zero. After one *at bat* the player's batting average either stays at zero (he did not record a hit) or rises to 1.0 (he did record a hit of some sort – a single, double, triple, or home run). Upon subsequent *at bats*, the player's batting average reflects the cumulative number of hits divided by the cumulative number of opportunities (at-bats). Eventually, most batting averages tend to stabalize due to the sheer number of opportunities accumulated throughout the baseball season (the denominator of the batting average statistic becomes quite large, effectively neurtering the influence of the binary numerator event [hit (1) or miss (0)].

After computing sequential cumulative batting averages for each player across the course of his full season, we next calculated absolute batting average difference between each player's *at bat*. Figure 1 illustrates this information with a small subset of 2008 data - these are American League (AL) first basemen. The x-axis reflects the *number of at-bats* and is truncated at 160 for simplicity of visual presentation (e.g., our goal was to make the Figure 1 presentation as easily interpretable as possible). Here, the x-axis origin reflects the progression from a player's *first* at bat to their *second* (because each player's first at bat was the first meaningful recording of the possible hit event). The y-axis reflects the absolute deviation from the first to second event – as can be seen in Figure 1, the largest absolute deviation from the first to second at bat is .5 – this happens when the hitter alternates event outcomes across the two at bats (e.g., misses the first and hits the second or vice versa). This left-most event also represents the greatest opportunity for a large index due to the small denominator (2).

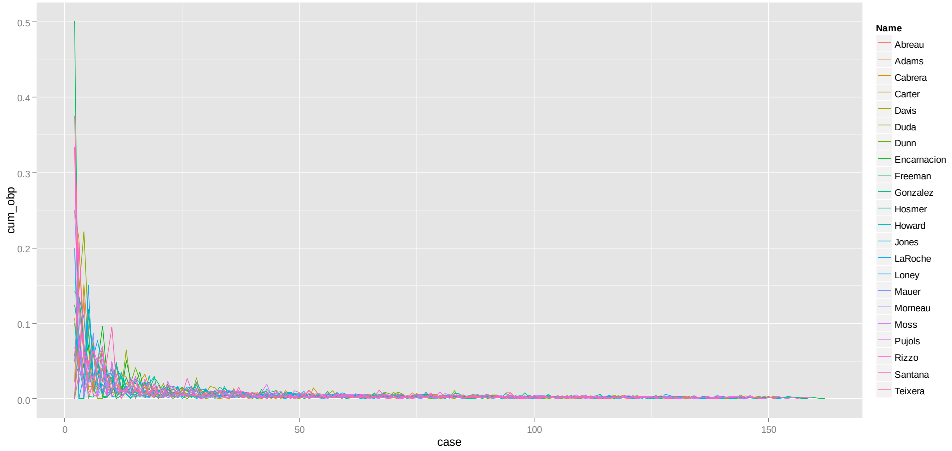

Figure 1. Small subset example of raw cumulative (absolute) average discrepancy.

3. Results

For purposes of function estimation, we were interested in the variance within vertical arrays. Figure 1 reflects this systematic pattern of heteroskedasticity, with greater variability in estimates near the origin (and less variability as plate appearances increase; e.g., to the "right" of Figure 1). The pattern is a bit more visually evident with the *standard deviation* of average absolute discrepancies, and these are presented for the first 100 at bats across all offensive players for the 2008 baseball season (see Figure 2).

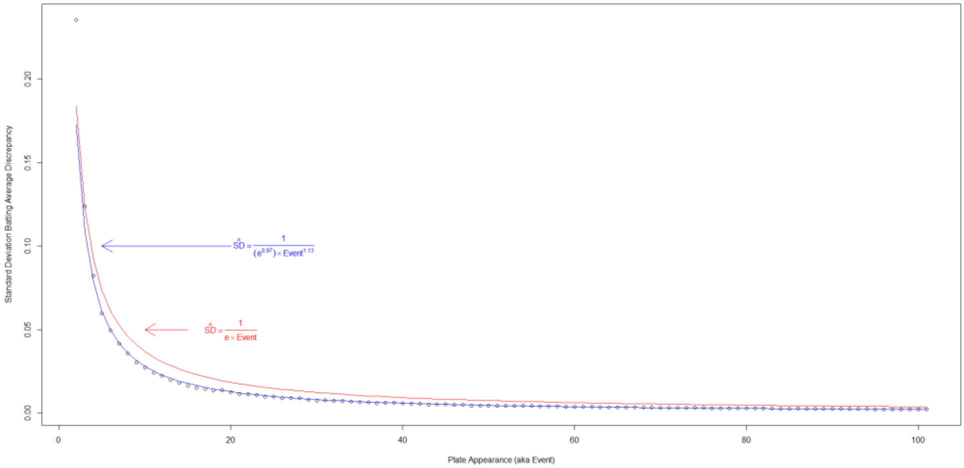

Figure 2. Predicted standard deviation as a function of event (in this case, "at bat" aka plate appearance).

Fitting a logarithmic regression to these standard deviations yields a reasonably predictive function: the predicted standard deviation (within vertical array) approximates $\frac{1}{e * Event}$. Via application of 8 years of baseball data, however, we were able to specify very slight modifyers to this general pattern: our empirical regression equation has slightly modified intercept and slope. For example, the 2008 data function was: $\log\widehat{\left(\frac{1}{sd}\right)} = .97 + 1.13 * \log(event)$. These functions explained the patterns of heteroskedasticity very well (R^2 = .9975, F = 39,690, p < .05 [again, only 2008 data]), and is presented visually via the blue function in Figure 2. Algebraically, our predictive model (solving for standard deviation in plate appearance batting averages instead of a logarithmic transformation of these) simplifies to: $\widehat{sd} = \frac{1}{e^{.97} * event^{1.13}}$. We also estimated similar functions for the other seven years of retrieved data. The CARMA presentation is dynamic, updating MLB player events and presenting as residual values to our aggregated (across 8 years) predictive function for batting average stability.

4. Discussion

For the purposes of this presentation, we focused on modeling the heteroskedasticity of a descriptive baseball statistic via standard deviation specification – by computing a simple standard deviation within each "array" (arrays are performance events – in Figure 1 the x-axis represents these events [e.g., an MLB "plate appearance"]). In broader applications, an "event" can be a "widget" (e.g., production), a work period (e.g., hour, shift, week, month), or service event (e.g., customer/consumer rating). Our ultimate interest, therefore, is twofold: 1) we intend to model similar functions across different baseball statistics, taking note of functional asymptotes and points of inflection, and 2) we hope to apply the general procedure of functional estimation across events to more common occurences of performance. Sturman (2003) notes in his meta-analysis that performance trends across time do tend to be different for different types of job, and so the estimate of functions across different baseball indicies may very well parallel different functions estimated across different jobs. Similar to the perspectives of both Kane (1996) and Hofmann, Jacobs, and Gerras (1992), our ultimate goal is to leverage insights taken from athletic performance in an attempt to conceptualize job performance in a new manner (here being operationalized as a predictive function across events).

References

Campion, M. C., Campion, M. A., & Campion, E. D. (2018). Big data techniques and talent management: Recommendations for organizations and a research agenda for IO Psychologists. *Industrial and Organizational Psychology*, *11*(2), 250-257.

Dalal, D. K., Nolan, K. P., & Gannon, L. E. (2017). Are pre-assembly shared work experiences useful for temporary-team assembly decisions? A study of Olympic ice hockey team composition. *Journal of Business and Psychology, 32,* 561-574.

Gentry, W. A., Hoffman, B. J., & Lyons, B. D. (2017). Box scores and bottom lines: Sports data can inform research and practice in organizations. *Journal of Business and Psychology, 32,* 509-512.

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research*, *70,* 308-317.

Heazlewood, T. (2006). Prediction versus reality: The use of mathematical models to predict elite performance in swimming and athletics at the Olympic games. *Journal of Sports Science and Medicine, 5,* 541-547.

Hofmann, D. A., Jacobs, R., & Gerras, S. J. (1992). Mapping individual performance over time. *Journal of Applied Psychology, 77,* 185-195.

Kane, J. S. (1996). The conceptualization and representation of total performance effectiveness. *Human Resource Management Review, 6,* 123-145.

Ployhart, R. E., & Hakel, M. D. (1998). The substantive nature of performance variability: Predicting interindividual differences in intraindividual performance. *Personnel Psychology, 51,* 859-901.

Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships. *Journal of Management, 29,* 609-640.

The baseball encyclopedia (8th edition). (1990). New York: MacMillan.

Tonidandel, S., King, E. B., & Cortina, J. M. (2016). Big Data methods: Leveraging modern data analytic techniques to build organizational science. *Organizatinoal Research Methods, 21,* 525-547.