SIMULATING RARE BASEBALL EVENTS USING MONTE CARLO METHODS IN *EXCEL* AND *R*

J. Scott Billie
Department of Mathematical Sciences
U. S. Military Academy
West Point, NY 10996
john.billie@usma.edu

Michael Huber
Department of Mathematics and Computer Sciences
Muhlenberg College
Allentown, Pennsylvania 18104
huber@muhlenberg.edu

Scott Nestler
Department of Mathematical Sciences
U. S. Military Academy
West Point, NY 10996
scott.nestler@usma.edu

Gabriel Costa
Department of Mathematical Sciences
U. S. Military Academy
West Point, NY 10996

Abstract

Statistics and baseball have had a long and venerable relationship throughout the history of the sport. Practically anything and everything has been recorded and tracked by avid baseball fans. The same fans debate statistics and argue which of the hallowed records will ever be broken. Websites such as *Baseball-Reference.com* and *Baseball-statistics.com* now offer data sets that include hitting, pitching, fielding, and base running events for both individual players and collective teams. Simulating certain events, such as hitting streaks or number of wins in a season, have become effective approaches to answering "Will this record ever be broken?" One such seemingly-unbreakable record is Joe's DiMaggio's famous 56-game hitting streak of 1941. In this paper we apply Monte Carlo simulation techniques in both *Excel* and the statistical package *R* to determine the likelihood of such a rare event using actual data from the DiMaggio streak.

INTRODUCTION

Statistics and baseball have had a long and venerable relationship throughout the history of the sport. Practically anything and everything has been recorded and tracked by avid baseball fans. The same fans debate statistics and argue which of the hallowed records will ever be broken. Websites such as *Baseball-Reference.com* and *Baseball-statistics.com* now offer data sets that include hitting, pitching, fielding,

and base running events for both individual players and collective teams. In 1971, the Society for American Baseball Research (SABR) was formed, with the mission of fostering the study of baseball past and present, and to provide an outlet for educational, historical and research information about the game.[1] Based on increasing popularity among faculty and students, "Sabermetrics" was first offered as a visiting professor's course at the United States Military Academy (USMA) in 1996, and it has been offered as a mathematics elective every year since 2001. This course goes beyond traditional statistics to engage students in understanding both "old school" measures and "new school" statistics (such as Win Shares, OPS, Linear Weights Runs, Regression, etc.) in order to study and analyze the National Pastime. The USMA Sabermetrics course is designed to advance the student's understanding of statistical concepts and to develop skills in applying those statistics to baseball. Mathematical applications in the course consist of formulating and solving mathematical problems as well as interpreting and explaining results. In this paper we specifically discuss the course topic of "Streakiness" or "The Hot Hand" [1]. To visually demonstrate streakiness, a Monte Carlo simulation was built in *Excel* to simulate Joe DiMaggio's famous fifty-six game hitting streak in 1941. The model was later coded in the statistical programming language *R* code to both take advantage of the free-ware and provide additional analytical rigor.

BACKGROUND

Collectively our group grew up with either a love for the game or a fascination with statistics. We remember summer afternoons listening to the our favorite team and their iconic announcers (Bob Prince for the Pirates, Mel Allen for the Yankees, and Chuck Thompson for the Orioles). Baseball memories include the gravely voice of Bob Prince espousing the exploits of the Pittsburgh Pirates. Or Mel Allen rhetorically asking, "How about that?" after a particularly exceptional Bronx Bomber play. We played our own games using Topps baseball cards; naturally the beloved Buccos or Earl Weaver's Birds would always win (1971 and 1979 excluded). The statistics fed our interest in baseball, and like many fans, the love for the game soon spilled over into other baseball board games such as All Star Baseball (Cadaco-Ellis), Strat-O-Matic^{TM}, and APBA. As technology advanced, these games would migrate over to personal computers and eventually onto interactive gaming stations. As our interest in mathematics and operations research increased, it was natural to look to baseball for research opportunities. Dice games would turn into spreadsheet models and offering a college-level course on Sabermetrics was a welcome decision. Even now, as we teach these older simulation models to students (the first chapter of *Curve Ball* is entitled, "Simple Models from Tabletop Baseball Games"), we bring in the APBA set and let the students pick teams from decades ago, matching batter versus pitcher abilities. Somehow the 1927 Yankees seem to dominate the APBA game. We then discuss the new generation gaming system (X-Box or Play Station) versions that current students are familiar with and prefer. All lead to explorations of simulating this great game. As Albert and Bennett point out, "If you couldn't play baseball outdoors, the next best thing was baseball indoors [1]."

The Sabermetrics course at West Point uses Albert and Bennett's *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game* as the primary textbook, but

[1] Mission statement found on the SABR home page, *www.sabr.org*.

also incorporates topical articles, papers and guest lecturers. This paper specifically looks at *Curve Ball*'s Chapter 5, which addresses the concept of streakiness. In the book, historical data is used to discuss this concept and provide two specific examples. The first example is the consistent versus streaky hitter (Todd Zeile in 1999) and the second example is consistent versus streaky team play (both NL and AL in 1998). We decided constructing a model to represent hitting streaks would be both interesting and beneficial for students.[2] During class we presented the *Excel* simulation and then discussed model development based on the Leemis and Park six-step process [11].

Students typically like the idea of having analytical solutions to all of their problems. It is reassuring for them to input numbers and "crank out" solutions. Unfortunately in the real world problems quickly become complex. In our case, we simply want to determine the probability of Joe DiMaggio getting a base hit in 56 consecutive games. We can use axiomatic probability (set theoretic approach) to determine the probability of getting a hit in a single at-bat (AB). We build a sample space of all previous events, in this case each AB is an experiment. We next determine the probability of getting the specific outcome of a hit (See Equation 1). The calculation is only good for one specific AB. The fraction will need to be adjusted every time the player comes to bat in order to account for the ever-increasing sample space.

$$P(\text{Getting a Hit}) = \frac{\text{Any Kind of Hit}}{\text{Total Number of AB}} \tag{1}$$

Because the probability of successfully getting a hit changes as the number of at-bats increases, we decided against an analytic approach. A dynamic system lent itself very well to a simulation model. Our fundamental approach was to first determine the number of AB for an individual game. Then for each AB, we used a Bernoulli trial to determine the outcome of the event to be either a hit or a miss. We then replicated the one game process over an entire 162 game season.[3] We used some routine "bookkeeping" to determine the longest hitting streak for that season. Since this will only provide us with one possible outcome (a point estimate), we elected to run several iterations or replications of multiple seasons to get a more accurate depiction of likely outcomes. Finally to reduce trial to trial variance, we ran three of the multiple seasons simulations. This repeated random sampling technique is the foundation of Monte Carlo simulation methods, that will be discussed in the next section.

In 1941, Joe DiMaggio of the New York Yankees had a magical season. The Hall of Famer hit .357, led the league in both runs batted in (RBI) and total base hits (H), and most astoundingly collected a base hit in 56 consecutive games. These accomplishments earned him the Most Valuable Player (MVP) Award in the American League.[4] Michael Seidel wrote that, "No other sustained performance in the history of baseball builds with the drama and explodes with the energy of Joe DiMaggio's

[2]Students take four core Mathematics courses at USMA. Specifically in the Introduction to Probability and Statistics course, students are introduced to building Monte Carlo simulations to provide insight and assist with their analysis.

[3]Using today's schedule of 162 games compared to the pre-expansion 154 games will only slightly increase the probability of hitting in 56 consecutive games. Further comparisons could determine by how much.

[4]Despite hitting .406 (the last player ever to bat above .400 in a season), Boston Red Sox outfielder Ted Williams would finish a distant second in the MVP voting.

56-game consecutive hitting streak launched on a hazy Thursday afternoon in New York on May 15, 1941, and grounded on a damp summer night in Cleveland on July 17 [14]." A natural heroic legend was created. During the streak, DiMaggio sported a .408 batting average. Four of the pitchers he faced during the streak would eventually be elected to the Baseball Hall of Fame. The next closest consecutive games hitting streak is 45 games attained by Willie Keeler in 1897.[5] Pete Rose compiled a 44-game streak in 1978. The DiMaggio streak is arguably one of the unattainable feats in modern day baseball. Constructing a simulation of this event helps to demonstrate exactly how rare this event is, and whether or not we might expect to see the record broken.[6]

There have been several treatises on simulations of hitting streaks in recent years. Arbesman and Strogatz [5] use Monte Carlo simulation to analyze Joe DiMaggio and baseball streaks in general. Their approach focuses on aggregating AB for each game played, treating a player's AB per game as a constant for all games in a season. Rockoff and Yates [13] treat AB per game as a binomial random variable and equate the probability of success to a player's seasonal batting average. Albert and Williamson [2] investigate the hot hand with a Bayesian model, relating a parameter and statistic in a linear fashion. They simulate binomial data based on the set of hitting probabilities obtained via a Markov switching model (flip a coin to decide the state of the first game and simulate the Markov chain for subsequent states). Albright [3] explores streakiness via a Markov dependence of individual at-bats; in addition, he proposes a logistic regression probabilistic model to incorporate successes and failures in individual at-bats. Thomas [17] simulates the performance of hitters over their entire careers, estimating a distribution of hitting streaks. Our model is again different, using actual data from DiMaggio's streak (number of at-bats per game and number of hits per at-bats in the streak) to update the Monte Carlo simulation.

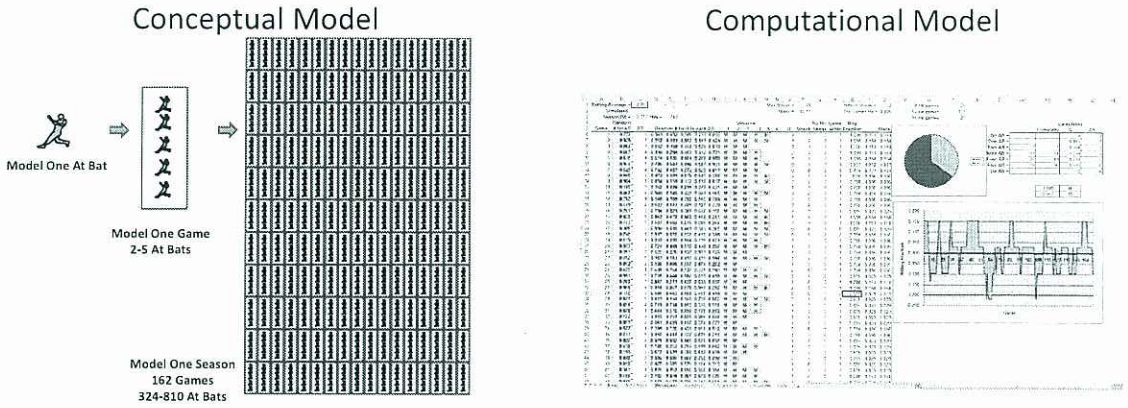
Monte Carlo Methods and Building the Conceptual Model

Most definitions of either a model or simulation allude to the fact that it is merely an abstraction of reality. The modeler must decide what aspects of the real world phenomena need to be replicated. For our model we looked at historical data for Joe DiMaggio's season batting average (the percentage of times he got a base hit during an at bat over the course of the season) and the number of at bats per game during his hitting streak. We employed a Monte Carlo simulation to generate the number of at bats per game and then use the player's season batting average to determine if each individual at bat resulted in either a "hit" or "miss" (not a hit). Each Monte Carlo iteration provides a possible outcome of the longest consecutive games hitting streak. Increasing the number of iterations assists in providing a more accurate depiction of the length of hitting streaks and the likelihood of a particular length based on our model. We will include software dependent techniques to increase the number of seasons replicated.

The first step is to make some key assumptions about the model we will build to represent the act of Joe DiMaggio hitting. We limited our data to 1941, the year of

[5]Keeler hit successfully in the last game of the 1896 season and then in the first 44 games of the 1897 season.

[6]Of note, Joe DiMaggio also holds the minor league baseball record of consecutive games with a hitting streak of 61, set in 1933 while playing for the Pacific Coast League's San Francisco Seals.



(a) Conceptual Model (b) Computational Model

Figure 1: Visualization of Conceptual Model and the *Excel* Computational Model.

the streak. For the season, DiMaggio played in 139 games and collected 193 hits in 541 at-bats. A simple ratio (similar to Equation 1) tells us that the Yankee Clipper would get a base hit in 35.7% of his at-bats. The batting average (AVG) is

$$AVG = \frac{H}{AB} = \frac{193}{541} = .357$$

We also assume that each AB is an independent event; otherwise, we would need to take into account such things as performance against individual pitchers and other situational effects.

Another important aspect of this model is determining the number of ABs for each game played. We assumed the number of ABs per game during the actual hitting streak provided a sound representational distribution. During the streak, Mr. DiMaggio had 223 official ABs. That number can be further broken down into three games with 2 ABs, eleven games with 3 ABs, twenty-six games with 4 ABs and sixteen games with 5 ABs. Based on these numbers, an empirical distribution is created and then converted into a cumulative distribution table, called a `VLOOKUP` table in *Excel*. See Table 1 for the cumulative at-bats. Typically the Yankee Clipper batted fourth in the line up and finished the season with the third most ABs for the Bronx Bombers. Of note, the possibility of getting no ABs in a game is not an option as this would effectively end the hitting streak.

A final assumption is that the random number function in *Excel*, `RAND`, provides sufficient rigor for our model.[7] The function returns a random number greater than or equal to 0 and less than 1, evenly distributed. The resulting number is a *Uniform*(0, 1) random variate. We then apply the Inverse Transform Method and use the random variate as an input to produce an outcome from our constructed empirical

[7] A good discussion on previous shortcomings of this function and the eventual solutions may be found at the Microsoft support site, *http://support.microsoft.com/kb/828795*.

distribution.

	Obs	Cumulative	%	AB
< Two AB	0	0	0.000	n/a
Two AB	3	3	0.054	2
Three AB	11	14	0.250	3
Four AB	26	40	0.714	4
Five AB	16	56	1.000	5

Table 1: Cumulative At Bats

Our initial decision was to build the specification model using *Excel*. The primary reason was tied to student familiarity with that particular software in core mathematics courses. Additionally, students have access to *Excel* after graduation.[8] However, due to an increased number of *R* users in the department, we decided to build a second model using that software. The following two sections describe the individual models.

The *Excel* Model

There are several positive aspects associated with modeling in *Excel*. The software program is ubiquitous, residing on laptops and personal computers throughout the world. Very little coding expertise is required, since the package is GUI-based. As a result, learning and mastering a few key functions can allow students to become simulation writers. Additional functionality may also be added through the use of *Microsoft Visual Basic*. The program is more than adequate for rough order of magnitude or "back of the envelope" calculations and finally, as mentioned previously, students get a healthy dose of spreadsheet modeling in their four core mathematics courses at West Point.

After determining limiting assumptions, we identify necessary input variables associated with a problem. As part of the spreadsheet modeling process, we require input variables to be entered as separate cells and not embedded into model equations. This technique allows for easy identification of input values, assists in code debugging and with quick value replacement aids in the conduct of sensitivity analysis. For this simulation, the batter's AVG (.357 for DiMaggio) and the two empirically based distributions (one for AB per game and other based on the AVG for hitting) are created and placed in the spreadsheet. At this point, a random number is generated for each game. The resulting random variate is used to determine the number of ABs per game. In *Excel*, this is accomplished with the `RAND` command and another *Excel* built-in function, the `VLOOKUP` command. The `VLOOKUP` function searches for an input value in the leftmost column of a table and then returns an output value in the same row from a column the user specifies. We use the last column in Table 1 to calculate the number of at-bats. For example, if the random number generated is 0.5431, we look at the column with "%" header, move down until we find the row that contains the lower value for the range (since it is a cumulative table). We then move along

[8]The US Army uses the Office suite in the work place and makes allowances to purchase home use software at a reduced cost.

the same row over to the *AB* column and determine that the number of at-bats for this event is 3. We now generate a random number for each game of the season (154 games prior to 1961 and 162 after that year).

Once the number of *AB*s is established for each game, the next step is to determine whether the event results in a "hit" or a "miss" (not a hit). Because the *AB*s are considered to be independent events, a new random number is generated for each one. Again a similar `VLOOKUP` command is used. This time the cumulative distribution is broken into two categories with the lower range of the second category equal to the season batting average (.357 in this DiMaggio example). To assist in sensitivity analysis, a slider bar tool is used to rapidly change the input value for the batting average.

The final step of the simulation process is to collect the descriptive statistics of all the individual events. Both `IF` and `COUNT` functions are used to accomplish this in the spreadsheet. Hit totals are collected for each game. Consecutive games with hits are then determined by looking at the hit total for the previous game. The streak total is incrementally increased by one if there is at least one hit in the previous game. If there are no hits in the current game, the streak total resets to zero. In the Sabermetrics classroom, we encourage students to calculate additional values found in *Curve Ball's* Chapter 5, to include Total Games with Hits, Percentage of Games with Hits, Moving Average of the Fraction of Games with Hits and the Black.

The original *Excel* spreadsheet model allows for a single replication of one the *AB*s for one game. We were able to replicate an entire 162 game season by simply copying and pasting the desired number of rows. The simulation could then be ran several times by using the *F9* key to recalculate all the random numbers. Figure 1 provides visual representations of the conceptual and computational models.

Between instructors and students in the class, we have run the simulation several hundreds, if not thousands, of iterations. The majority of hitting streak lengths were within the 10 to 25 games range. Hitting steaks of length 26 to 40 did occur, but at a greatly reduced rate. Only rarely did a hitting steak of greater than 40 occur. These results seemed to make sense, but to better understand the distribution of possible outcomes, we needed to generate more than simple point estimates. A more analytically rigorous approach was to increase the number of replications of the simulation and then run multiple trials. We acomplished this using seven lines of Visual Basic code. Other add-in software packages specific to *Excel*, such as *Risk Solver*, *Crystal Ball*, and *@Risk* exist, however we sought out open source programs that students could take with them and use after graduation.

The *R* Model

Michael Crawley describes *R* as a high-level language and environment for data analysis and graphics. More specifically, *R* is an open source version of the popular statistics program *S-PLUS*. There are several benefits associated with this software. The most prominent feature is that the software may be downloaded from the Comprehensive *R* Archive Network website (*http://cran.r-project.org/*) and run on any platform free of charge.

In general, the structure of the R model is very similar to the *Excel* model. We start by again providing the input parameters for `Games` and `BatAvg`, but additionally provide `Paths`, which determines the number of iterations R will execute the model. A cumulative table representing the probability of player ABs in a game is again used, but coded with `else if` statements. A random number is again generated for the Monte Carlo simulation portion, but now using R's `runif(n, min=0, max=1)` command. This will generate n-*Uniform*(0,1) random numbers, which are then compared to the cumulative table's thresholds and produce the number of AB for each game. Some minor housekeeping is required to construct arrays to track the statistics of interest; for this model only the longest hitting streak (in games) and the number of games in a season with a hit are included, but others could easily be added using a similar technique. A comparable method is used to determine whether the outcome of each AB in each game is a "hit" or a "miss". The statistics of interest for that "season" are then stored and the large, outer loop repeats the process a specified number of times. Then, using the graphical capabilities of R, a histogram is generated representing the longest hitting streak in each of the simulated seasons. Similarly, we construct an Empirical Cumulative Distribution Function (ECDF) (see Figure 2). The location of a 56-game hitting streak is annotated on each plot by a dashed, red line.

Results and Analysis

We ran both the *Excel* and R Monte Carlo simulations for 32,500 iterations with five replications.[9] A summary of the descriptive statistics for the maximum consecutive games hitting streak is provided in Table 2. A quick visual inspection finds little difference between the measures of the mean, standard deviation and minimum for *Excel* and R. There does appear to be a difference between the maximum value for the two approaches. The wide spread found in the data may be explained due to this parameter residing in the right-hand tail of the distribution. Creating empirical distributions from the data will allow us to examine the probability of a batter achieving a specific number of consecutive games with a hit.

	Excel					*R*				
Trial #	1	2	3	4	5	1	2	3	4	5
Mean	18.93	18.97	19.00	18.90	18.94	19.00	18.93	18.93	18.94	18.94
Std Dev	5.74	5.75	5.81	5.76	5.79	5.81	5.75	5.79	5.78	5.74
Min	6	7	7	7	7	8	7	7	7	7
Max	64	68	74	71	62	66	83	74	89	60

Table 2: Descriptive Statistics for *Excel* and R Iterations

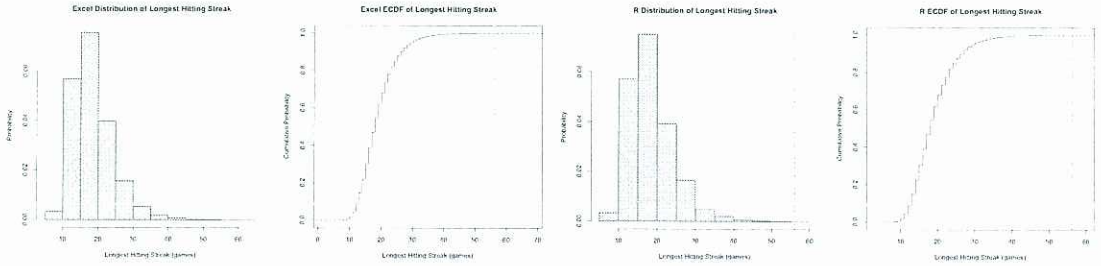
One method of capturing the probability distribution is to use an Empirical Distribution Function(EDF) or Empirical Cumulative Distribution Function (ECDF). Devore [7] provides the following definition of an EDF: Let $x_1, x_2, ..., x_n$ be a random sample. The *empirical distribution function* $F_n(x)$ is a function of x which equals the fraction

[9] The 32,500 number was based on a size limitation encountered when using the VBA code in the *Excel* spreadsheet.

of X_is that are less than or equal to x for each x, $-\infty \leq x \leq \infty$. Simply put, the EDF or ECDF represents the accumulation of actual sample data.

$$F_n(x) = (\frac{\text{observations} \leq x}{n}) \tag{2}$$

For each simulation replication we used the ECDF function in R to calculate the predicted probability of a 56 game consecutive hitting streak, $\hat{F_{56}}$. In our case we find the average $\hat{F_{56}}$ for both models to be between 0.9998277 and 0.9998954. It is apparent that this is a rare event, but to quantify exactly how rare it truly is requires two things— simulating a sufficient number of seasons (32,500 in this case) and then repeating it some number of times (5 in this case) to get a good representation of the true population distribution. As shown in Table 3, a 56-game hitting streak by a .357 hitter only occurs about 1 out of 5,000 (since $1 - 0.9998277 = 0.0001723$, this equates to approximately 2 out of 10,000 or roughly 1 in 5000) seasons. Note that simulating only 1000 seasons would likely result in an estimate of 1 for the ECDF (i.e., no occasions in 1000 seasons), since it is more rare than 1 in 1000. By modifying our simulation, we could easily perform sensitivity analysis if our batter was instead a .400 hitter (a more unlikely event) or a more average .250 hitter. The resulting histograms and ECDFs are provided for the first 32,500 iteration of both models. On all the graphs, the likelihood of a 56 game consecutive hitting streak is indicated with a dashed, red line. Both sets of graphs provide very similar results.



(a) *Excel* Histogram (b) *Excel* ECDF (c) *R* Histogram (d) *R* ECDF

Figure 2: Histograms and ECDFs

	Trials					
	1	**2**	**3**	**4**	**5**	$\hat{F_{56}}$
Excel	.999815	.999815	.999785	.999908	.999815	.9998277
R	.999877	.999785	.999908	.999385	.999969	.9998954

Table 3: Results of R Monte Carlo Simulation for 32,500 Replications/Trial

Conclusions and Recommendations

We were able to demonstrate the likelihood of a .357 batter (a very respectable average) to safely hit in 56 consecutive major league baseball games had a probability of 0.0001723 percent. A logical next step for this project would be to conduct sensitivity analysis. The .357 posted by DiMaggio is somewhat lofty, even using today's inflated hitting standards. By continuing to simulate individual ABs, we would have to determine the appropriate metrics for the previous inputs. In lieu of the .357 value, the major league average for each batting position (1 through 9) should be calculated. In addition, a new cumulative ABs table should also be constructed for each new batting average. A new simulation could then be run to determine the likelihood of the consecutive games hitting streak being broken.

Other areas to explore include determining if there is any difference between the Monte Carlo simulation models run in the different software programs. The initial results are promising based on similar outcomes, however, additional statistical testing should be conducted to ensure there is no difference between the results. A final consideration for this paper is to determine the appropriate probability distribution associated with our parameter of interest. The following is a technique used in our Introduction into Probability and Statistics Course. We assume it would follow a Poisson distribution and be related to other rare events in baseball such as no-hitters and hitting for the cycle [9].

References

[1] Albert, J. and J. Bennett. 2003. *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game.* Springer - Verlag, Inc.: New York, New York.

[2] Albert, J. and P. Williamson. 2001. *Using Model/Data Simulations to Detect Streakiness.* The American Statistician. Vol. 55, No. 1, pp. 41-50.

[3] Albright, S. C. 1993. *A Statistical Analysis of Hitting Streaks in Baseball.* Journal of the American Statistical Association, Vol. 88, No. 424 (Dec., 1993), pp. 1175-1183.

[4] Albright, S. C. 1993. *A Statistical Analysis of Hitting Streaks in Baseball: Rejoinder.* Journal of the American Statistical Association, Vol. 88, No. 424 (Dec., 1993), pp. 1194-1196.

[5] Arbesman, S. and S. H. Strogatz. 2008. *A Monte Carlo Approach to Joe DiMaggio and Streaks in Baseball.* Unpublished. *http://arxiv.org/abs/0807.5082.*

[6] Crawley, M. J. 2007. *The R Book.* John Wiley & Sons, Ltd. West Sussex, England.

[7] Devore, J. L. 2008. *Probability and Statistics for Engineering and the Sciences, Revised Seventh Edition, United States Military Academy.* Cengage Learning: New York.

[8] Law, A. M. and Kelton, W. D. 2000. *Simulation Modeling and Analysis.* McGraw-Hill, Inc.: Boston.

[9] Huber, M. and A. Glen. 2007. *Modeling Rare Baseball Events - Are They Memoryless?.* Journal of Statistics Education, Volume 15, Number 1, 2007.

[10] Lawson, B., and L. M. Leemis. 2008. Monte Carlo and Discrete-Event Simulations in C and R. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Möch, T. Jefferson, J. W. Fowler.

[11] Leemis, L. M. and S. K. Park. 2006. *Discrete-Event Simulation: A First Course.* Pearson Prentice Hall: Upper Saddle River, New Jersey.

[12] Maindonald, J. and J. Braun. 2005. *Data Analysis and Graphics Using R, An Example-based Approach.* Cambridge University Press: New York, New York.

[13] Rockoff, D. M. and P. A. Yates. 2009. *Chasing DiMaggio: Streaks in Simulated Seasons Using Non-Constant At-Bats.* Journal of Quantitative Analysis in Sports. Vol. 5, Issue 2, Article 4.

[14] Seidel, Michael. 1988. *Streak: Joe DiMaggio and the Summer of '41.* McGraw-Hill Book Company: New York, New York.

[15] Siwoff, Seymour. 2004. *The Book of Baseball Records.* Elias Sports Bureau, Inc.: New York, New York.

[16] Stern, S. S. and C. N. Morris. 1993. *A Statistical Analysis of Hitting Streaks in Baseball: Comment.* Journal of the American Statistical Association. Vol. 99, No. 434, pp 1189-1194.

[17] Thomas, A. C. 2008. *Simulating Record Accomplishments in Baseball, Or, That's the Second-Biggest Hitting Streak I've Ever Seen!.* Unpublished. it http://www.people.fas.harvard.edu/ acthomas/papers/sims-writeup.pdf.

[18] Walkenbach, J. 2004. *Microsoft Office Excel 2003 Power Programming with VBA.* Wiley Publishing, Inc.: Indianapolis, Indiana.

[19] Excel Tip.com website. http://www.exceltip.com/st/Using_Loops_in_VBA in Microsoft Excel/628.html. Accessed 2 November 2009.