MATHEMATICA APPLICANDA
Vol. 40(2) 2012, p. 13–31

Zachary Hass (West Lafayette)
Wojbor A. Woyczynski (Cleveland)
Chris Yanosko (Cleveland)
Eric Becker (Cleveland)

A Lost Opportunity: Recovering the End of Major League Baseball's 1994 Strike Shortened Season

Abstract The 1994 Major League Baseball (MLB) Season in the United States ended prematurely when the players went on strike on August 12th, due to a labor disagreement with team owners. This paper describes the model estimation for predicting the runs scored in each of the unplayed games and gives the results of 1,000 simulations. Of particular interest are the Cleveland Indians and the Montreal Expos. The Expos were on pace to have the best season in franchise history (and the best record in the league), while the Indians were poised to begin a very successful run that could have ended the city's World Championship drought dating from 1948.

2010 Mathematics Subject Classification: BBB Primary 62J12; Secondary 62P99, 62F10.

Key words and phrases: Poisson mixed effects regression, simulations, baseball statistics, discrete Weibull distribution, Markov chains.

1. Introduction. Tensions ran high between players and owners heading into the 1994 Major League Baseball Season. The collective bargaining agreement had expired in December of 1993 as team owners pressed for a salary cap to fix the disparity in the league, hoping to make small market teams more competitive. Player trust of ownership, however, was suffering from the lingering effects of the recent free agency collusion charges that resulted in a $280 million settlement. The situation was exacerbated when ownership opted in June not to pay $7.8 million in pension and benefits that had previously been agreed upon. This was followed by the U.S. Senate's failure to approve anti-trust legislation which would have given the players leverage

in the negotiations. As a result, the players went on strike effective August 12, 1994. The strike lasted 232 days and was the 4th work stoppage for the league in the last 22 years. As many as 710 games went unplayed, depending on how the playoffs would have played out. It was the first time the World Series had been canceled since 1904 [30].

The strike cut short what was shaping up to be a very interesting season. Before the work stoppage, Matt Williams was on pace to match Roger Maris' single season home run record, and Tony Gwynn had the best chance of breaking .400 for a season batting average since Ted Williams. The Montreal Expos had a 74-40 record, on pace to have the best season in club history and were widely considered to be major World Series contenders, even though they had the 2nd lowest payroll in the league. Following the strike the attendance in MLB dropped across the board, and Montreal never recovered, eventually losing the franchise to Washington D.C., where they became the Nationals [30].

The Cleveland Indians also lost a great opportunity. Given its well-publicized championship drought in major sports, fans bemoan how often they have been on the wrong end of spectacular games or plays that decide championship runs. Not since 1964 has the city won a championship (NFL title), and not since 1948 had the Indians brought home the World Series. 1994 marked the first year in what would become a string of successful seasons as players such as Carlos Baerga, Kenny Lofton, and Sandy Alomar Jr. began to perform at a high level. At the time of the strike, Cleveland sat one game behind the White Sox in the race for the AL Central, and one game ahead of the Kansas City Royals for the Wild Card spot (the playoff spot reserved for the best record outside of a divisional champion in each league). The Indians would go on to make the playoffs in each of the following five seasons. Losing in the World Series in 1995 to the Braves, and in 1997 to the Marlins, they couldn't quite close the deal, coming as close as a single out from being crowned the champion [7]. Fans may always wonder what could have been, had the fateful 1994 season been played out? Could the Expos have parlayed a magical season into a sustainable, competitive franchise? Could the Indians have acquired the playoff experience necessary to close the deal in the subsequent World Series? This paper fits a statistical model to the data of the played season and then uses it to simulate the unplayed portion of the regular season 1,000 times to see what probably would have transpired.

The paper is written in the style that would appeal to a quantitative baseball expert. We assume that the reader is familiar with the fundamentals of the baseball game but do not include all the mathematical and statistical subtleties involved. Instead, we provide detailed citation for a mathematically inclined person.

2. Related Work. Professor Bolesław Kopociński of the University of Wrocław, Poland, addressed a similar problem in his two papers, *Components*

of the Game Result in a Football League [20] and *Unfinished League Season of Football* [21]. In the former, Professor Kopociński shows how Poisson regression is useful for estimating the different components that affect goals scored in a soccer match. He finds home field advantage, relative strength of teams (as measured by position in final standings), and a random component to be significant in predicting goals scored [20]. In the latter paper, Kopociński uses his method for estimating components to fit a model to the results of the 1939 Polish Premiere Soccer League season. That season ended about two-thirds of the way through as it was interrupted by the invasion of Poland by the German and Soviet armies and the resulting outbreak of World War II [21]. Although baseball is certainly a different sport from soccer, the parallels between the two situations are strong enough to extend Kopociński's work to our situation. All but 14 games of the 1994 season through August 11th were played, and the results are readily available for fitting a model to predict the latter part of the season (the other 14 were postponed, presumably due to inclement weather). Additionally, runs scored in baseball, like goals in soccer, is also a type of *count data* making Poisson Regression the logical first choice.

Relevant approaches to modeling scores in sports leagues may also be found in papers by Glickman and Stern [12], Keller [19], and Lee [23]. Also, in the baseball context, the Bradley-Terry model with random effects is described in James, Albert, and Stern [16], and a two-stage Bayesian model for predicting winners in Major League Baseball is proposed in Yang and Swartz [29]. However, to the best of our knowledge, no one made an effort to apply those techniques in the dramatic context of the unfinished 1994 season.

3. Data set. The amount of data about any single baseball season, readily available to the public, has exploded in recent years. The rising popularity of *sabermetrics*, a field of baseball research based on statistics and other objective evidence, has fueled much of the growth in this data[1].

One of the key concepts of sabermetrics is the *run creation*. There are various formulas for run creation, but they all contain three basic components: the ability to get on-base, the ability to advance bases, and the number of opportunities to do both. On the other hand, the *Runs Above Replacement* (RAR) measure quantifies how many runs a player produces beyond what a replacement would produce over the course of a full season.[2] This important concept can be applied directly to pitching, fielding, and batting, and the results allow for relative comparisons amongst different players utilizing a single number. Pitching RAR, for instance, measures a pitchers ability to prevent runs from scoring [8].

[1] The name *sabermetrics* is derived from the name of the organization, Society for American Baseball Research (SABR).

[2] Here, by a replacement we mean an average player from a .320 winning percentage club.

Term (Abbreviation)	Formula
Hits (H)	
At-Bats (AB)	
Base-on-Balls (BB)	
Hit-by-Pitch (HBP)	
Sacrifice Fly (SF)	
Touched Bases (TB)	
Innings Pitched (IP)	
Earned Runs (ER)	
Batting Average (BA)	H / AB
On-Base Percentage (OBP)	(H +BB +HBP) / (AB +BB +HBP +SF)
Slugging Percentage (SLG)	TB/AB
On-Base Plus Slugging (OPS)	OBP + SLG
Earned Run Average (ERA)	(ER * 9) / IP
Walks and Hits Per Inning Pitched (WHIP)	(BB +H) / IP

Table 1: Selected variables and baseball statistics abbreviations.

The data set used here for our model fitting contains 3200 "observations" from 1600 games of all 28 MLB teams (Arizona and Tampa Bay did not enter the league until 1998). All statistics were taken from `Baseball-Reference.com` [27]. A number of different variables were tested as potential predictors; the basic baseball abbreviations and the formulas for those predictor variables are given in Table 1. The dependent variable, of course, is the number of *Runs Scored* (RS) by each team.

Our work shows that the most useful predictor turns out to be a *relative strength index* (RSI) based on the RAR measures which provides a much better differentiation across the aspects of the game than one based on the standings alone. In this paper RSI is defined as a sum of the offensive, fielding, and pitching RARs.

For each team, *offensive RAR* is a sum of the RARs for the 8 position players on each roster in the National League (NL), and the same sum plus the designated hitter's RAR in the American League (AL). Likewise, the *fielding RAR* is a sum of the 8 position players' fielding RARs. *Pitching RAR* was calculated as a sum of the RARs of the top 5 starting pitchers, the top closer, and the top 4 relievers for each ball club.

The *game relative strength* (GRS) for each team in a given pairing is the maximum of 0 and the difference between the RSIs of the team and its opponent. This choice of the definition of the game relative strength assures that this quantity remains nonnegative, an important property in our modeling.

4. Empirical Model. Baseball runs are count data, with the event counted being a run scoring play. This makes the Poisson distribution a natural first, and simplest (one parameter) candidate for our modeling. Our data

have the mean of about 4.9 runs per team per game which yields the model probability distribution

$$\textbf{Prob}\,[\#\,\text{of runs} = k] \approx e^{-4.9}\frac{4.9^k}{k!}, \qquad k = 0,\ 1,\ 2,\ \ldots \tag{4.1}$$

We would like to emphasize here that this choice (obviously influenced by Kopociński's papers) was the "rough first cut" and was made to simplify our calculations and permit access to standardized regression software. A discussion of the limitations of this choice – obvious from Figure 1 where the QQ plot, (see, e.g., Denker and Woyczynski [9], p. 69) showing the quantiles of the scored runs data from the completed part of the 1994 MLB season versus the quantiles of the Poisson distribution with the same mean $\lambda = 4.9$ — and the distributional issues involved can be found in Section 7.

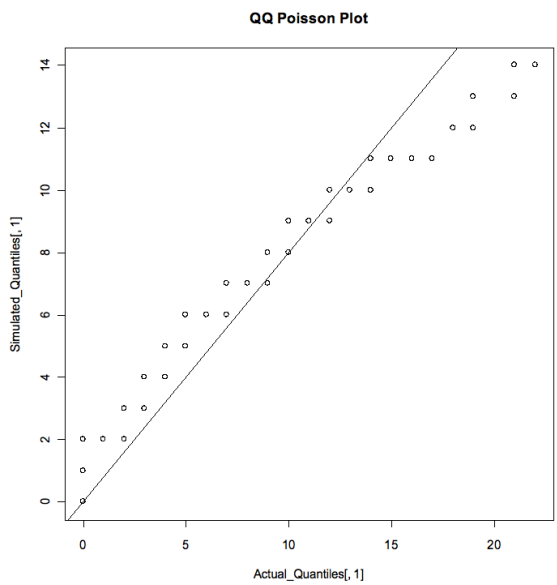


Figure 1: The QQ plot for the quantiles of the scored runs data from the completed part of the 1994 MLB season, versus the quantiles of the Poisson distribution with the same mean $\lambda = 4.9$.

As a consequence of the above choice the Poisson (rather than the standard Gaussian)) mixed effects predictive regression was employed. Mixed effects Poisson regression model is the special case of generalized linear model with the logarithmic link function and is appropriate for modeling of count data (see, e.g., Jiang [18], Cameron and Trivedi [5], or Christensen [6], for mathematical details). The formal assumption is that the response variable Y has a Poisson distribution. Rather than fitting a very high-dimensional and

computationally intensive Poisson regression model for the number of runs R_{ij} scored by team i against team j involving simultaneously all possible predictor variables,

$$R_{ij} = \mathcal{P}(\lambda_{ij}) \tag{4.2}$$

with the mean

$$\lambda_{ij} = c_1 \cdot \mathrm{HOME}_{ij} + c_2 \cdot \mathrm{BA}_i + c_3 \cdot \mathrm{OBP}_{ij} + c_4 \cdot \mathrm{OPS}_i + c_5 \cdot \mathrm{FLD}_i$$

$$+c_6 \cdot \mathrm{SP}_i + c_7 \cdot \mathrm{BP}_i + c_8 \cdot \mathrm{GRS}_{ij} + c_9 + \mathrm{RND}, \tag{4.3}$$

we opted for a collection of eight, lower dimensional Poisson regression models involving fewer regressor (predictor) variables which made "baseball sense". We employed the popular STATA computing platform using the "ready-to-use" `xtmepoisson` command for the mixed effects Poisson regression model. A comparison of the coefficients for these Poisson regression fittings, as well as their AIC scores, is shown in Table 2. Sizes of the 95 % confidence intervals for each coefficient are also displayed in parentheses.

Model	Home	BA	OBP	OPS	Fielding	Starting Pitching	Bull Pen	Strength	Constant	Random	AIC
1	-0.007 (0.686)	-4.832** (0.021)	0.905 (0.309)	3.202*** (0.000)					0.070 (0.210)	0.183+	17071
2	-0.007 (0.690)				-0.002*** (0.004)	-0.001*** (0.002)	-0.002** (0.024)		1.710*** (0.000)	0.190+	17109
3	-0.006 (0.690)							0.002*** (0.000)	1.511*** (0.000)	0.173+	17012
4	-0.006 (0.719)	-4.078** (0.033)		2.160*** (0.001)				0.002*** (0.000)	0.951*** (0.007)	0.169+	16994
5	-0.005 (0.693)				-0.001 (0.158)	-0.000 (0.987)	-0.000 (0.839)	0.002*** (0.000)	1.518*** (0.000)		17013
6					-0.001** (0.027)	-0.000 (0.142)	-0.001 (0.328)	-4.467** (0.020)		0.167+	16989
7		-4.079** (0.033)		2.181*** (0.001)	-0.001 (0.112)			0.002*** (0.000)	0.930*** (0.008)	0.168+	16990
8		-4.081** (0.033)		2.159*** (0.001)				0.002*** (0.000)	0.946*** (0.0007)	0.169+	16992

Table 2: Selected Poisson Mixed Effects regressions; *Runs Scored* is the dependent variable here. (****Significant at the 1% level, **5% level, +10% level.*

Recall that AIC (see, e.g., [1], or [4], for more details), which rewards goodness of fit of the model but also penalizes overfitting by including too many parameters, measures the loss of information (increases of entropy) caused by using the particular model under consideration instead of the "true" model. Utilizing the concept to compare two different models, with AICs, say, 16992 and 16990, respectively, produces the information that the first model is $\exp((16990 - 16992)/2)) = .37$ as probable as the second to minimize the information loss.

The random component was used to capture such things as weather conditions, player illness, or even a swarm of midges converging on a pitcher (as

was the case in the famous game 2 of the 2007 American League Division Series between the Indians and Yankees) [7]. One observation is that the home field advantage was not a useful predictor of the number of runs scored; note the large confidence intervals associated with the corresponding coefficients. This is not inconsistent with Levernier and Barilla's work [24] analyzing the home-field advantage in Major League Baseball using logit models.

Observe that all the fittings in Table 2 resulted in the negative coefficient for the BA term in the expression for the expected number of runs λ. This, obviously, is counterintuitive as the baseball "common sense" would indicate that the increased batting average would result in a larger number of runs scored. But we have to remember that OPS is also in the model and it is a broader offensive statistics and does have a positive coefficient as one would expect. So, in our model the BA plays the role of an adjustment variable for the effect of OPS. OPS measure how often someone gets on base (regardless of the method) and how far they get around the bases on average. BA measures only how often a player gets on base through a hit out of those at bats that qualify (it doesn't count sacrifice flies or bunts for example). Although the phenomenon may seem controversial it turned out to be significant and we decided to include it to make the model more flexible. The '94 average National League OPS was 0.794, and average BA was 0.267. Multiplying these numbers by the coefficients gives an average positive effect on the number of runs scores as one would expect (around 0.63, or so).

Of all the candidate models fit to the data, we selected the one with the lowest AIC (Akaike Information Criterion) but balanced the choice by taking into account the appearance of the most statistically significant coefficients. As a result of the above considerations we settled on the last model (number 8) in Table 2 to model the number of runs R_{ij} scored by team i against team j;

$$R_{ij} = \mathcal{P}[1.115 - 4.081 \cdot \mathrm{BA}_i + 2.159 \cdot \mathrm{OPS}_i + 0.002 \cdot \mathrm{GRS}_{ij}], \tag{4.4}$$

where $\mathcal{P}[\lambda]$ denotes a Poisson random variable with mean λ. The constant term combines both the constant and the random term (0.946, and 0.169, respectively). To predict the runs scored for a game, for any combination of teams, one needs the teams' Batting Averages, On-Base Plus Slugging, and the Relative Strength measure between the two team's Runs Above Replacement measure. Inserting this information into the model provides the Poisson random variable for a given team against a given opponent, which is then plugged into a Poisson random number generator, providing the score of that team for a particular game and simulation.

5. Simulations. The schedule of regular 1994 season games taken from `RetroSheet.org` [28] shows that 655 games originally scheduled after August 11, 1994 were not played. Furthermore, it can be deduced upon careful examination that an additional 14 games were postponed and not made up

elsewhere in the schedule leaving 669 games to be simulated to finish the regular season.

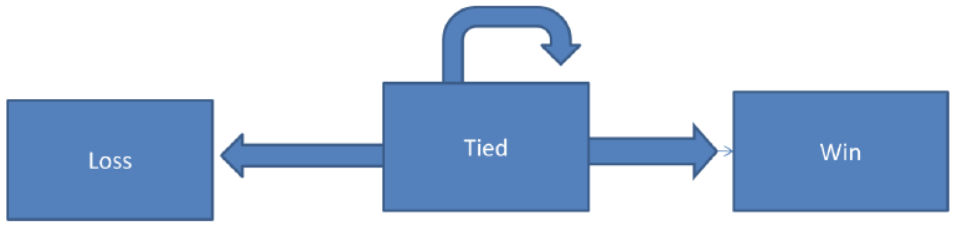


Figure 2: Transition graph of the Markov Chain used to resolve the outcome of the extra innings

Major League Baseball games do not end in ties. If the score is tied after nine innings of play, an additional inning is played. If the score is again tied after this additional inning, they will play an eleventh inning, and so on, until there comes an inning where one team emerges victorious. To resolve this, we used *Markov Chain modeling* (see, e.g., [3]). This is appropriate since a Markov Chain is a random process where the probabilities of advancing to any particular state depend only on the current state. Figure 2 shows an illustration of the state space and its transition graph for the relevant Markov Chain.

Since we were not interested in how many innings the game took to be resolved, we only needed the asymptotic probabilities of victory given a team's λ (the expectation of the Poisson random variable) and the opponent's λ. Such a calculation can be done by building a matrix of probabilities for each of the possible runs scored, with the columns referring to *home team's* runs, and rows referring to *away team's* runs, both beginning at 0. The sum of the lower triangular portion of the matrix is the home's probability of loss, while the upper triangular portion sums to the home's probability of victory, and the diagonal sums to the probability of a tie, that is, an additional extra inning of play. The win-loss probabilities were then rescaled to sum up to one (asymptotic probabilities). This was done for all potential match-ups.

It is worth noting that this method did not calculate additional runs scored in the extra innings, which reduced average runs scored for the simulated seasons. End-of-season standings' ties were also prominent. The pairs of teams which ended the regular season tied most often in 1,000 simulations are given in Table 3, together with the corresponding frequencies. Only end of season ties that affected which clubs would qualify for the playoffs were resolved. Interestingly, several three-way end-of-season ties were encountered. In the case of a three-way tie, each team was assigned a random slot, with

Team A and Team B playing each other, and Team C playing the winner for the right to advance to the playoffs. More unique situations exist based on whether the tie was for the wild card or for the division, but all were resolved according to the MLB policy [25].

Cincinnati Reds	Houston Astros	73
Cleveland Indians	Chicago White Sox	65
San Francisco Giants	Los Angeles Dodgers	63
Texas Rangers	Oakland Athletics	41
Texas Rangers	Seattle Mariners	36
Cleveland Indians	Kansas City Royals	35
Montreal Expos	Atlanta Braves	35
Chicago White Sox	Kansas City Royals	28
Seattle Mariners	Oakland Athletics	27
Cleveland Indians	Baltimore Orioles	23
Cincinnati Reds	Atlanta Braves	23

Table 3: Pairs of teams which ended the regular season tied most often in 1,000 simulations. The right-most column shows the corresponding frequencies

6. Results. In all, 1,000 regular season simulations were calculated and tabulated. The cumulative results for each MLB team are given in Table 4.

The consecutive columns show the final average (over 1000 simulated seasons) number of wins and losses, the percentage of simulations in (or, equivalently, probability with) which the team made the playoffs and earned the wild card. Also, in the final two columns, shown are the probabilities that the team won its division, and the frequency with which it participated in a play-in game.

Twelve of the teams did not make the playoffs a single time out of a thousand. This is not unexpected, as teams are often *out of the race* by August. Two of these, though, Toronto and Detroit ended the season in a tie for the wild card before subsequently losing the play-in game. The AL West stayed true to expectations, putting forth a sub .500 champion in the Seattle Mariners (46% of simulations) with an average mark of 75–87. The National League West did slightly better with an average champion in the Los Angeles Dodgers (76% of simulations) with a record of 84–78. In both leagues, the Central and Eastern Divisions competed for the wild card spot. In the National League, three of the playoff spots went to the Montreal Expos, Atlanta Braves, Cincinnati Reds, or the Houston Astros in every season. Out of the thousand simulations, in only 11 did the Expos fail to make the playoffs, and only 8.1% of the time were they the wild card.

AL EAST	Wins	Losses	Playoffs	Wild Card	Division	Play-in Ties
NYY	98.32	63.70	0.980	0.010	0.970	13
BAL	88.86	73.18	0.109	0.079	0.030	46
TOR	77.42	84.58	-	-	-	1
BOS	76.12	85.88	-	-	-	0
DET	79.01	82.99	-	-	-	1

AL CENTRAL	Wins	Losses	Playoffs	Wild Card	Division	Play-in Ties
CHW	95.98	66.12	0.911	0.661	0.250	103
CLE	99.06	63.02	0.982	0.232	0.750	77
KCR	86.25	75.76	0.018	0.018	-	9
MIL	74.22	87.78	-	-	-	0
MIN	71.21	90.79	-	-	-	0

AL WEST	Wins	Losses	Playoffs	Wild Card	Division	Play-in Ties
TEX	74.42	87.65	0.345	-	0.345	70
OAK	72.82	89.23	0.188	-	0.188	56
SEA	75.21	86.88	0.460	-	0.460	87
CAL	65.33	96.67	0.005	-	0.005	1

NL EAST	Wins	Losses	Playoffs	Wild Card	Division	Play-in Ties
MON	101.57	60.46	0.989	0.081	0.908	36
ATL	94.66	67.45	0.736	0.644	0.092	106
NYM	77.42	84.58	-	-	-	0
PHI	78.61	83.39	-	-	-	0
FLA	70.55	91.45	-	-	-	0

NL CENTRAL	Wins	Losses	Playoffs	Wild Card	Division	Play-in Ties
CIN	94.52	67.56	0.731	0.124	0.607	82
HOU	93.14	68.98	0.545	0.152	0.393	114
PIT	73.66	88.34	-	-	-	0
STL	72.30	89.70	-	-	-	0
CHC	71.33	90.67	-	-	-	0

NL WEST	Wins	Losses	Playoffs	Wild Card	Division	Play-in Ties
LAD	83.95	78.11	0.758	-	0.758	58
SFG	80.54	81.51	0.236	-	0.236	56
COL	73.41	88.60	0.006	-	0.006	2
SDP	68.57	93.43	-	-	-	0

Table 4: Results of the simulations

The most probable playoff picture in the National League would have pitted the Montreal Expos (1st Seed) vs. the Los Angeles Dodgers (3rd seed), with the winner taking on the winner of the Cincinnati Reds (2nd seed) vs. the Atlanta Braves (4th seed). The one and four seeds would not meet in divisional round as they are from the same division.

The most likely substitutes to this scenario, according to our table, would be the San Francisco Giants replacing the Dodgers 23.6% of the time, and the Houston Astros in the playoffs 54.5% of the time for the Reds, Braves, and Expos in the order of most likely.

In the American League, five teams competed for the other three spots opposite the AL West champion, the New York Yankees, Baltimore Orioles, Chicago White Sox, Cleveland Indians, and Kansas City Royals. In the East, the Yankees took the division crown 97% of the time, and the Orioles were the third most likely to take the wild card at 7.9%.

In the Central, where only a game separated 1st, 2nd, and 3rd place before the simulation, it was the Indians who took the divisional crown most often at 75% of the time, with the White Sox taking it the other 25%of the time. Of the five, the Royals were the least likely to make the playoffs at only 1.8% of simulations. It should be noted that the Royals strength that year, in respect to RAR was their pitching, as they were ranked number

one overall. In batting and fielding they were ranked 19th and 20th respectively.

The most likely playoff picture for the American League would pit the Yankees (1st seed) vs. the Chicago White Sox (4th seed) with the winner playing the victor of the Cleveland Indians (2nd seed) vs. the Seattle Mariners (3rd seed). The Texas Rangers would be most likely to replace the Mariners (34.5% of the time), then the Oakland A's (18.8%), and California Angels (0.5%). The Orioles (10.9%) and Royals (1.8%) would replace the other three.

7. Refinements, Future Work, Conclusions. At this point a number of thoughts come to mind:

7.1. How good is the Poisson approximation? The answer is: not very good but, perhaps, good enough for our purposes. We believe that a more sophisticated, long-tail model mentioned below would not have changed the final predictions significantly; the winning team running up the score seldom changes the final outcome of the game. The QQ-plot in Figure 1 shows that the choice of the Poisson mixed effects regression model is not completely unreasonable. But a more subtle analysis of the data shows that the tail of the probability distribution of the number of runs scored is much heavier that that of the Poisson random variable $\mathcal{P}(\lambda)$ with the same mean. The probability of a Poisson random variable with $\lambda = 4.9$ being equal to, say, $k = 10$, is equal to 0.016, whereas, the empirical data of runs scored show the corresponding frequency to be frequency 0.034. Table 5 shows all the original relative frequencies f_k of the number of runs k scored in games actually played in the 1994 season.

k	f_k	k	f_k
0	0.04477	12	0.01502
1	0.08547	13	0.01095
2	0.11834	14	0.00563
3	0.12398	15	0.00250
4	0.13807	16	0.00219
5	0.11646	17	0.00125
6	0.09987	18	0.00031
7	0.08140	19	0.00093
8	0.05666	20	0,00000
9	0.04383	21	0.00062
10	0.03381	22	0.00031
11	0.01753		

Table 5: Relative frequencies f_k of the number of runs k scored in games actually played in the 1994 season.

More significantly, the variance of the the above Poisson distribution is 4.9, while the corresponding empirical variance for our data is 10.4. But with only one parameter in the Poisson distribution once one fits the mean, there is no flexibility to fit the variance as well.

Furthermore, Poisson distributions have, essentially, Gaussian tails, since (see, e.g., Glynn [13]) **Prob** $[\mathcal{P}(\lambda) \geqslant k]$ is bounded from above by (up to some constants) $1 - \Phi(k)$, where $\Phi(k) = \int_{-\infty}^{k} e^{-x^2/2}\, dx/\sqrt{2\pi}$ is the cumulative distribution function, CDF, of the standard Gaussian distribution. The former is a very rapidly decaying function of k. Indeed,

$$1 - \Phi(k) \leqslant \frac{1}{\sqrt{2\pi}k} e^{-\left(\frac{k}{\sqrt{2}}\right)^2}. \tag{7.1}$$

Thus, a slower-decaying probability distribution with, perhaps, two parameters may be in order. A natural choice here would be to replace the exponent k^2 in (7.1) by a slower growing k^α with $\alpha < 2$. Thus we settled on the CDF of an integer-valued random variable $\mathcal{W}$ of the form

$$\Phi_{\alpha,\beta}(k) = \textbf{Prob}\,[\mathcal{P}(\lambda) \leqslant k] = 1 - e^{-\left(\frac{k+1}{\beta}\right)^\alpha}, \qquad k = 0, 1, 2 \ldots \tag{7.2}$$

This distribution is known as the discrete Weibull distribution and has been considered in the context of reliability (see, Nakagawa and Osaki [26]), and microbial counts (see, Englehardt and Li [10]]. The parametric estimation issue for (7.2) was studied by Ali Khan et al. [2].

The result was a surprisingly good fit shown in Figure 3 (left). The dots indicate the values of $\Phi_{\alpha,\beta}(k)$, for $\alpha = 1.72$, and $\beta = 6.04$, and the continuous line shows the (interpolated) values of the empirical CDF obtained from the actual frequencies in Table 5. For comparison, in Figure 3 (right), we are showing the Poisonian CDF (dots), with the mean $\lambda = 4.9$, versus the same empirical CDF (continuous line). The results speak for themselves. The CDF $\Phi_{\alpha,\beta}(k)$ may be viewed as a discrete version of the Weilbull, or stretched exponential, distribution.

Remark 7.1 The statistical environment **R** contains a parametric estimation package for the Discrete Inverse Weibull Distribution. Four methods are provided: the method of moments, the method of proportion, the heuristic algorithm, and the method of (inverse) Weibull probability paper plot in [17]. They all gave much worse results than the basic method used by us to produce the fit visualized in Fig. 3 (left). We just scanned numerically the 2-D parameter space and found the values of the parameters α, and β, that minimized the Kolmogorov-Smirnov (KS) uniform distance between the fit and the empirical CDF, which turned out to be equal to 0.016. A thorough asymptotic statistical analysis of the minimum KS distance can be found

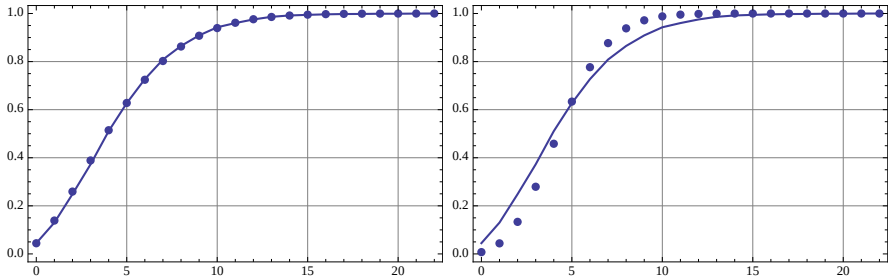


Figure 3: *Left:* The dots indicate the values of the CDF $\Phi_{\alpha,\beta}(k)$, , for $\alpha = 1.72$, and $\beta = 6.04$, and the continuous line shows the (interpolated) values of the empirical CDF obtained from the actual frequencies in Table 5 showing a surprisingly good fit. *Right:* For comparison, i the Poisonian CDF (dots), with the mean $\lambda = 4.9$, does not fit nearly as well the empirical CDF (continuous line).

in, .e.g., [22], and [14, 15]. We are not aware of any sufficiently deep mathematical analysis of the estimators proposed by [17] to recommend them. An implementation of **R** package to the problem considered in the present paper were not encouraging either.

7.2. Future Work. So, why didn't we work with $\Phi_{\alpha,\beta}(k)$ instead of the Poissonian model? The answer is that, for the Poissonian model we had tools of the mixed effects regression readily available in STATA software. Given the amazingly good fit shown in Figure 3 (left), it certainly would be worthwhile to develop similar analytic and software tools for the $\Phi_{\alpha,\beta}(k)$ model using the theory available in, e.g. Jiang's book [18]. It will take some serious mathematical effort but we intend to work on it in the future.

Team	Played Winning %	Simulated Winning %	Projected Winning %
New York Yankees	0.619	0.521	0.547
Detroit Tigers	0.461	0.521	0.498
Cleveland Indians	0.584	0.558	0.649
Kansas City Royals	0.557	0.495	0.535
California Angels	0.409	0.460	0.446
Montreal Expos	0.649	0.525	0.540
Chicago Cubs	0.434	0.483	0.348
San Diego Padres	0.402	0.491	0.421

Table 6: Checking Reasonability of Our Projections for Select Teams.

There may be some deeper reason for the appropriateness of the $\Phi_{\alpha,\beta}(k)$ model in baseball. In soccer games considered by Kopocinski [20, 21], the

Poisson model was totally appropriate because of the low typical scores. The corresponding cumulative score count process $N(t)$ is the Poisson process (as a function of time t). It is memoryless, and the number of goals scored in disjoint time intervals are assumed to be independent random variables. The "instantaneous" rate of change of the score is constant in time in this model. Also, one can score only one goal at a time. On the other hand, in baseball, the score can advance instantaneously by up to four runs and there may be a tendency for teams ahead in the game to pile up runs, which would make the instantaneous rate of change of the score time-dependent. The continuous-time version of the $\Phi_{\alpha,\beta}(k)$ model is often encountered in reliability studies, and the corresponding Weibull distribution is the limit distribution for the minimum of independent identically distributed random variables (also, see, Denker and Woyczynski [9], and Ferguson [11].

Obviously, the fact the baseball score can advance by up to four runs, makes the compound Poisson model also a good candidate for our future studies.

Subsequent work could take the simulation all the way through the playoffs, either by simulating playoff games based on the most probable playoff match ups or by finishing each individually simulated season off with the appropriate playoff simulation, or perhaps some third method. One thing that is clear, however, is that a great opportunity for baseball fans was lost that year, especially those that followed the Indians and the Expos.

Application of our approach to prediction of outcomes of any given season based on a partial season information is also likely to be of some use to the baseball insiders and outside bettors (wherever legal).

7.3. Conclusions. How reasonable are the results of this paper? Table 6

Team Top 5	**Simulated Winning %**
Cleveland Indians	0.674
Cincinnati Reds	0.593
Chicago White Sox	0.590
New York Yankees	0.578
Houston Astros	0.576
Team Bottom 5	
Pittsburgh Pirates	0.430
Florida Marlins	0.416
St. Louis Cardinals	0.402
California Angels	0.390
Minnesota Twins	0.387

Table 7: Simulated winning percentages of top five and bottom five performers.

gives select teams' winning percentages for the games already played, for the games simulated, and projected winning percentages. The projection is

based on a simple weighted average of each team remaining on the schedule multiplied by the winning percentage against that opponent in games actually played that year. The weight is given to the number of games remaining against that opponent.

To illustrate, at the time of the strike the Yankees had a winning percentage of 0.619. In the simulation, their average winning percentage was 0.578. Given who the Yankees had left on their schedule and how well they had previously done (for example they had not lost to the Indians all season and had 3 games remaining against them, which would be counted as $3*1.000 = 3$ wins) they were projected to finish with a winning percentage of 0.547. This serves as a sort of baseline for the results. In this case the drop off in wins in our simulation is reasonable given the toughness of the Yankees remaining schedule.

Who appears to have missed out the most from the premature end to the season? Table 7 gives gives an answer for the top five and bottom five performers in the simulations. After 1,000 simulations, the Cleveland Indians make the playoffs 82.8% of the time and the Montreal Expos breeze into postseason play a convincing 98.7% of the time. Although the model could be improved by considering a Weibull-type or Compound Poisson Process regression model before fans demand that record books show an additional playoff berth, it did work reasonably well in simulation, providing fairly convincing results.

8. Epilogue. We dedicate this paper to the endlessly suffering, but dedicated fans of the Cleveland Indians. During the past half-century the Indians teased us seven times by getting into the post season. Yet, since 1948 they brought no World Championship home. But, Leo Tolstoy begins his short story, also (not accidentally) titled *A Lost Opportunity*, with a citation from St. Matthew xviii., 21-35: "Then came Peter to Him, and said, Lord, how oft shall my brother sin against me, and I forgive him? till seven times? . . . " So, perhaps, eighth time's the charm.

Acknowledgment: The authors would like to thank Jim Albert of the Bowling Green State University for valuable bibliographical comments that helped us improve the paper which, however, probably still ended up not quite up to his usual high sports statistics standards. Krzysztof Szajowski of the Wroclaw University of Technology educated us about the history of the discrete Weibull distribution, and helped with the formal parametric estimation procedure. We are grateful for his interest in the paper.

References

[1] H. Akaike, *A new look at the statistical model identification.* IEEE Transactions on Automatic Control 19 (6) (1974) : 716–723. MR0423716; Zbl 0314.62039; doi: 10.1109/TAC.1974.1100705

[2] M.S. Ali Khan, A. Khalique, and A.M. Abouammoth, *On estimating parameters in a discrete Weibull distribution*. IEEE Transactions on Reliability 38 (3) (1989) : 348-350. Zbl 0709.62640; doi: 10.1109/24.44179

[3] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer-Verlag, pages: xviii+444, 1999. ISBN 0-387-98509-3; MR1689633

[4] K. P. Burnham, and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag, pages: xxvi+488, 2002. MR1919620

[5] A.C. Cameron, and P.K. Trivedi . *Regression analysis of count data*, Cambridge University Press, 1998. ISBN 0-521-63567-5; MR1648274

[6] R. Christensen, *Log-linear models and logistic regression*. Springer Texts in Statistics (Second ed.). New York: Springer-Verlag, pages: xvi+483, 1997, ISBN 0-387-98247-7. MR1633357;

[7] *Cleveland Indians*. Wikipedia. 19 April 2011. Wikimedia Foundation Inc. Last Accessed 21 April 2011. <http://en/wikipedia.org/wiki/Cleveland_Indians>

[8] G.B. Costa, M. R. Huber, and J.T. Saccoman. *Understanding Sabermetrics: an Introduction to the Science of Baseball Statistics*. : McFarland , Jefferson, NC, 2008.

[9] M. Denker and W.A. Woyczynski, *Introductory Statistics and Random Phenomena: Uncertainty, Complexity and Chaotic Behavior in Engineering and Science*. Birkhäuser Boston Inc., 1998. MR1643201

[10] J.D. Englehardt and R. Li, *The discrete Weibull distribution: An alternative for correlated counts with confirmation for microbial counts in water* . Risk Analysis 31 (3) (2011) : 370-381. doi: 10.1111/j.1539-6924.2010.01520.x

[11] N. Ferguson, *Large Sample Theory*. Chapman and Hall, London, 1996. MR1699953

[12] M.E. Glickman and H.S. Stern *A state-space model for National Football League scores*, J. Amer. Statist. Assoc. 93 (1998): 25–35. Zbl 0915.62078

[13] P.W. Glynn, *Upper bounds on Poisson tail probabilities*, Operations Research Letters 6 (1987): 9–14. Zbl 0616.60020; doi: 10.1016/0167-6377(87)90003-4

[14] L. Györfi, I. Vajda, and E.C. van der Meulen, *Minimum Kolmogorov distance estimates for multivariate parametrized families*, American Journal of Mathematical and Management Sciences 16(1-2) (1996): 167–191. Zbl 0883.62053

[15] L. Györfi, I. Vajda, and E.C. van der Meulen, *Minimum Kolmogorov distance estimates of parameters and parametrized distributions*, Metrika 43(3) (1998): 237-255. Zbl 0855.62016

[16] B. James, J. Albert and H.S. Stern, *Answering questions about baseball using statistics*, Chance 6(1993): 17–22. doi: 10.1137/1.9780898718386.ch15

[17] M.A. Jazi, C.-D. Lai and M.H. Alamatsaz, *A discrete inverseWeibull distribution and estimation of its parameters*, Statistical Methodology, 7(2010): 121-132.

[18] J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics. New York: Springer-Verlag, pages xiv+257, 2007. ISBN 978-0-387-47941-5; 0-387-47941-4. MR2308058

[19] J.B. Keller, *A characterization of the Poisson distribution and the probability of winning a game*, Amer. Statistician 48(1994): 294–298. MR1321895 doi: 10.2307/2684837

[20] B. Kopociński, *Components of the Game Result in a Football League*, Applicationes Mathematicae, 28 (2001): 55–72. MR1819529 doi: 10.4064/am28-1-4

[21] B. Kopociński, *Unfinished League Season of Football*, Demonstratio Mathematica, 34 (2001): 461–468. MR1833200

[22] A. Kozek, *On minimum distance estimation using Kolmogorov-Lévy type metrics*, Australian and New Zealand J. Statist. 40(3) (1998): 317-333. MR1649478 doi: 10.1111/1467-842X.00036

[23] A. Lee, *Modeling scores in the Premier League: Is Manchester United really the best?* Chance 10(1997): 15–19. doi: 10.1137/1.9780898718386.ch40

[24] W. Levernier and A.G. Barilla, *An Analysis of the Home-Field Advantage in Major League Baseball Using Logit Models: Evidence from the 2004 and 2005 Seasons,* Journal of Quantitative Analysis in Sports: 3(1) (2007): 1-22. doi: 10.2202/1559-0410.1045

[25] *Major League Baseball Tie-Breaking Procedures*, Wikipedia. 4 October 2010. Wikimedia Foundation Inc. Last Accessed 23 April 2011. <http://en.wikipedia.org/wiki/ >

[26] T. Nakagawa and S. Osaki, *The discrete Weibull distribution.* IEEE Transactions on Reliability R-24 (5) (1974) : 300-301.

[27] *Major League Statistics and Information*, Sports Reference LLC. Baseball-Reference.com, Sean Forman, Accessed between 9/2010 - 12/2010. <http://www.baseball-reference.com/>

[28] D. Smith, et al. *Regular Season Games.* Retrosheet. Mark Pankin. 2010. Retrosheet Inc. Accessed 21 September 2010. <http://www.retrosheet.org/schedule/1994sked.txt>

[29] T.Y. Yang and T. Swartz, *A two-stage Bayesian model for predicting winners in Major League Baseball*, Journal of Data Science 2(2004): 61–73.

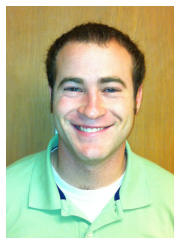
[30] *1994-95 Major League Baseball Strike.* Wikipedia. 9 April 2011. Wikimedia Foundation, Inc. Last Accessed 21 April 2011. <http://en.wikipedia.org/wiki/1994>

Stracone nadzieje: Symulacja dokończenia przerwanego przez strajk sezonu amerykańskiej ligi baseballowej

Streszczenie: Amerykański sezon baseballowy w roku 1994 zakończył się przedwcześnie 12 sierpnia, kiedy zawodnicy zadeklarowali strajk z powodu kontraktowych konfliktów z wlaścicielami klubów. Dla kibiców w Ameryce, gdzie baseball jest narodowym sportem granym od dziecka, było to dramatyczne wydarzenie. Niniejsza praca proponuje model estymacji prognozy, opartej na mieszanych modelach liniowych, wyników w każdej z nierozegranych z powodu strajku gier sezonu i przedstawia wyniki 1000 symulacji Monte Carlo i przewidywania ostatecznej klasyfikacji sezonu. Szczególnie intereresująca była sytuacja dwóch klubów: Cleveland Indians i Montreal Expos. Expos mieli szanse pobicia swojego własnego rekordu wygranych gier w jednym sezonie i osiągniecia najlepszego wyniku w całej lidze. Sezon Indians rownież się zapowiadał bardzo pomyślnie i otwierał realistyczną możliwość wygrania World Series, po raz pierwszy od 1948 roku.

Chociaż, począwszy od igrzysk w Los Angeles w 1984 roku, baseball jest regularnym sportem olimpijskim (bardzo rozpowszechnionym nie tylko w USA ale rownież w Ameryce Łacińskiej i Japonii) i po raz pierwszy był sportem demonstracyjnym już na Sztokholmskich igrzyskach w roku 1912, gra nie jest popularna w Polsce (choć istnieje) i jej zasady nie są powszechnie rozumiane. W tym celu (i w duchu kulturalnego zbliżenia między polskimi i amerykańskimi kibicami, w którym niniejsza praca została przedstawiona do międzynarodowego czasopisma redagowanego w Polsce) sugerujemy by czytelnik, przed przestudiowaniem naszej pracy, zajrzał do jednej z paru polskich stron internetowych wyjaśniających zasady baseballa. Doskonałym przykładem jest tutaj barwna strona http://www.baseball.pl/o-grach/zasady.html, w której wyjaśnienia są po polsku, ale z dokładnymi odnośnikami (w nawiasach) do klasycznej angielskiej terminologii. Oczywiście, w internecie amerykanskim, zagooglowanie terminu *baseball* oddaje błyskawicznie, w ciągu .21 sekund, listę 549 milionów związanych z tematem stron internetowych i terminologia baseballowa znalazła mocne odzwierciedlenie w codziennej idiomatyce języka angielskiego w Ameryce. Tutaj dobrym źrodłem dla początkującego fanatyka baseballa (i studenta Amerykańskiego angielskiego) jest http://en.wikipedia.org/wiki/Baseball_rules

Słowa kluczowe: Regresja poissonowska z mieszanymi efektami, symulacje, statystyki basebolowe, dyskretny rozklad Weibulla, łańcuchy Markowa.



Zachary Haas was born in Fennimore, WI, USA, in 1988. He received his BA from Case Western Reserve University in Statistics and Economics in 2010, where he also played for the football team. Currently he is working towards a MS in Applied Statistics at Purdue University. His email is zhass@purdue.edu. He was assisted in this project by his Case undergraduate colleagues Chris Yanosko i Eric Becker.

Wojbor A. Woyczyński was born in Częstochowa, Poland, in 1943. He received his M.Sc, in Electrical and Computer Engineering from the Wrocław University of Technology in 1966, his Ph.D in Mathematics from the University of Wrocław, in 1968, and his Habilitation in 1972. He is Professor in the Department of Statistics, and Director of the Center for Stochastic and Chaotic Processes in Science and Technology, at Case Western Reserve University, Cleveland, OH 44122, U.S.A. His email is: waw@case.edu, and the information about his work can be viewed at http:/sites.google.com/a/case.edu/waw.

Zachary Hass
Purdue University
Department of Statistics, 250 N. University Street, West Lafayette, IN 47907-2066, U.S.A.
E-mail: zhass@purdue.edu

Wojbor A. Woyczynski
Case Western Reserve University
Department of Statistics and Center for Stochastic and Chaotic Processes in Science and Technology, Cleveland, OH 44106, U.S.A.
E-mail: waw@case.edu
URL: http:/sites.google.com/a/case.edu/waw

Chris Yanosko
Case Western Reserve University

Eric Becker
Case Western Reserve University

(Received: 28th of October 2012)