

A DERIVATION OF THE PYTHAGOREAN WON-LOSS FORMULA IN BASEBALL

STEVEN J. MILLER

ABSTRACT. It has been noted that in many professional sports leagues a good predictor of a team's end of season won-loss percentage is Bill James' Pythagorean Formula $\frac{RS_{\text{obs}}^\gamma}{RS_{\text{obs}}^\gamma + RA_{\text{obs}}^\gamma}$, where RS_{obs} (resp. RA_{obs}) is the observed average number of runs scored (allowed) per game and γ is a constant for the league; for baseball the best agreement is when γ is about 1.82. This formula is often used in the middle of a season to determine if a team is performing above or below expectations, and estimate their future standings.

We provide a theoretical justification for this formula and value of γ by modeling the number of runs scored and allowed in baseball games as independent random variables drawn from Weibull distributions with the same β and γ but different α ; the probability density is

$$f(x; \alpha, \beta, \gamma) = \begin{cases} \frac{\gamma}{\alpha} ((x - \beta)/\alpha)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$

This model leads to a predicted won-loss percentage of $\frac{(RS-\beta)^\gamma}{(RS-\beta)^\gamma + (RA-\beta)^\gamma}$; here RS (resp. RA) is the mean of the Weibull random variable corresponding to runs scored (allowed), and $RS - \beta$ (resp. $RA - \beta$) is an estimator of RS_{obs} (resp. RA_{obs}). An analysis of the 14 American League teams from the 2004 baseball season shows that (1) given that the runs scored and allowed in a game cannot be equal, the runs scored and allowed are statistically independent; (2) the best fit Weibull parameters attained from a least squares analysis and the method of maximum likelihood give good fits. Specifically, least squares yields a mean value of γ of 1.79 (with a standard deviation of .09) and maximum likelihood yields a mean value of γ of 1.74 (with a standard deviation of .06), which agree beautifully with the observed best value of 1.82 attained by fitting $\frac{RS_{\text{obs}}^\gamma}{RS_{\text{obs}}^\gamma + RA_{\text{obs}}^\gamma}$ to the observed winning percentages.

1. INTRODUCTION

The goal of this paper is to derive Bill James' Pythagorean Formula (see [Ja], as well as [An, Ol]) from reasonable assumptions about the distribution of scores. Given a sports league, if the observed average number of runs a team scores and allows are RS_{obs} and RA_{obs} , then the Pythagorean Formula predicts the team's won-loss percentage should be $\frac{RS_{\text{obs}}^\gamma}{RS_{\text{obs}}^\gamma + RA_{\text{obs}}^\gamma}$ for some γ which is constant for the league. Initially in baseball the exponent γ was taken to be 2 (which led to the name), though fitting γ to the observed records from many seasons lead to the best γ being about 1.82. Often this formula is applied part way through a season to estimate a team's end of season standings. For example, if halfway through a season a team has far more wins than this formula predicts, analysts often claim the team is playing over their heads and predict they will have a worse second-half.

Rather than trying to find the best γ by looking at many teams' won-loss percentages, we take a different approach and derive the formula and optimal value of γ by modeling the runs scored and allowed each game for a team as independent random variables drawn from Weibull distributions with the same β and γ but different α (see §3 for an analysis of the 2004 season which shows that, subject to the condition that the runs scored and allowed in a game must be distinct integers, the runs scored and allowed are statistically independent, and §4 for additional comments on the independence). Recall the three-parameter Weibull distribution (see also [Fe2])

Date: April 2, 2006.

2000 Mathematics Subject Classification. 46N30 (primary), 62F03, 62P99 (secondary).

Key words and phrases. Pythagorean Won-Loss Formula, Weibull Distribution, Hypothesis Testing.

is

$$f(x; \alpha, \beta, \gamma) = \begin{cases} \frac{\gamma}{\alpha} \left(\frac{x-\beta}{\alpha} \right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases} \quad (1.1)$$

We denote the means by RS and RA, and we show below that $RS - \beta$ (resp. $RA - \beta$) is an estimator of the observed average number of runs scored (resp. allowed) per game. The reason $RS - \beta$ and not RS is the estimator of the observed average runs scored per game is due to the discreteness of the runs scored data; this is described in greater detail below. Our main theoretical result is proving that this model leads to a predicted won-loss percentage of

$$\text{Won-Loss Percentage}(RS, RA, \beta, \gamma) = \frac{(RS - \beta)^\gamma}{(RS - \beta)^\gamma + (RA - \beta)^\gamma}; \quad (1.2)$$

note for all γ that if $RS = RA$ in (2.6) then as we would expect the won-loss percentage is 50%.

In §3 we analyze in great detail the 2004 baseball season for the 14 teams of the American League. Complete results of each game are readily available (see for example [Al]), which greatly facilitates curve fitting and error analysis. For each of these teams we used the method of least squares and the method of maximum likelihood to find the best fit Weibulls to the runs scored and allowed per game (with each having the same γ and both having $\beta = -.5$; we explain why this is the right choice for β below). Standard χ^2 tests (see for example [CaBe]) show our fits are adequate. For continuous random variables representing runs scored and runs allowed, there is zero probability of both having the same value; the situation is markedly different in the discrete case. In a baseball game runs scored and allowed *cannot* be entirely independent, as games do not end in ties; however, modulo this condition, modified χ^2 tests (see [BF, SD]) do show that, given that runs scored and allowed per game must be distinct integers, the runs scored and allowed per game are statistically independent. See [Ci] for more on the independence of runs scored and allowed.

Thus the assumptions of our theoretical model are met, and the Pythagorean Formula should hold for some exponent γ . Our main experimental result is that, averaging over the 14 teams, the method of least squares yields a mean of γ of 1.79 with a standard deviation of .09 (the median is 1.79 as well); the method of maximum likelihood yields a mean of γ of 1.74 with a standard deviation of .06 (the median is 1.76). This is in line with the numerical observation that $\gamma = 1.82$ is the best exponent.

In order to obtain simple closed form expressions for the probability of scoring more runs than allowing in a game, we assume that the runs scored and allowed are drawn from continuous and not discrete distributions. This allows us to replace discrete sums with continuous integrals, and in general integration leads to more tractable calculations than summations. Of course assumptions of continuous run distribution cannot be correct in baseball, but the hope is that such a computationally useful assumption is a reasonable approximation to reality; it may be more reasonable in a sport such as basketball, and this would make an additional, interesting project. Closed form expressions for the mean, variance and probability that one random variable exceeds another are difficult for general probability distributions; however, the integrations that arise from a Weibull distribution with parameters (α, β, γ) are very tractable. Further, as the three parameter Weibull is a very flexible family and takes on a variety of different shapes, it is not surprising that for an appropriate choice of parameters it is a good fit to the runs scored (or allowed) per game. What is fortunate is that we can get good fits to both runs scored and allowed simultaneously, using the same γ for each; see [BFAM] for additional problems modeled with Weibull distributions. For example, $\gamma = 1$ is the exponential and $\gamma = 2$ is the Rayleigh distribution. Note the great difference in behavior between these two distributions. The exponential's maximum probability is at $x = \beta$, whereas the Rayleigh is zero at $x = \beta$. Additionally, for any $M > \beta$ any Weibull has a non-zero probability of a team scoring (or allowing) more than M runs, which is absurd of course in the real world. The tail probabilities of the exponential are significantly greater than those of the Rayleigh, which indicates that perhaps something closer to the Rayleigh than the exponential is the truth for the distribution of runs.

We have incorporated a translation parameter β for several reasons. First, to facilitate applying this model to sports other than baseball. For example, in basketball no team scores fewer than 20 points in a game, and it is not unreasonable to look at the distribution of scores above a baseline. A second consequence of β is that adding P points to both the runs scored and runs allowed each game does not change the won-loss percentage; this is reflected beautifully in (1.2), and indicates that it is more natural to measure scores above a baseline (which may be zero). Finally, and most importantly, as remarked there are issues in the discreteness of the data and the continuity of the model. In the least squares and maximum likelihood curve fitting we bin the runs scored and allowed data into bins of length 1; for example, a natural choice of bins is

$$[0, 1) \cup [1, 2) \cup \dots \cup [9, 10) \cup [10, 12) \cup [12, \infty). \quad (1.3)$$

As baseball scores are non-negative integers, all of the mass in each bin is at the left endpoint. If we use untranslated Weibulls (i.e., $\beta = 0$) there would be a discrepancy in matching up the means.

For example, consider a simple case when in half the games the team scores 0 runs and in the other half they score 1. Let us take as our bins $[0, 1)$ and $[1, 2)$, and for ease of exposition we shall find the best fit function constant on each bin. Obviously we take our function to be identically $\frac{1}{2}$ on $[0, 2)$; however, the observed mean is $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$ whereas the mean of our piecewise constant approximant is 1. If instead we chose $[-.5, .5)$ and $[.5, 1.5)$ as our bins then the approximant would also have a mean of $\frac{1}{2}$. Returning to our model, we see a better choice of bins is

$$[-.5, .5) \cup [.5, 1.5) \cup \dots \cup [7.5, 8.5) \cup [8.5, 9.5) \cup [9.5, 11.5) \cup [11.5, \infty). \quad (1.4)$$

An additional advantage of the bins of (1.4) is that we may consider either open or closed endpoints, as there are no baseball scores that are half-integral. Thus, in order to have the baseball scores in the *center* of their bins, we take $\beta = -.5$ and use the bins in (1.4). In particular, if the mean of the Weibull approximating the runs scored (resp. allowed) per game is RS (resp. RA) then $RS - \beta$ (resp. $RA - \beta$) is an estimator of the observed average number of runs scored (resp. allowed) per game.

2. THEORETICAL MODEL AND PREDICTIONS

We determine the mean of a Weibull distribution with parameters (α, β, γ) , and then use this to prove our main result, the Pythagorean Formula (Theorem 2.2). Let $f(x; \alpha, \beta, \gamma)$ be the probability density of a Weibull with parameters (α, β, γ) :

$$f(x; \alpha, \beta, \gamma) = \begin{cases} \frac{\gamma}{\alpha} \left(\frac{x-\beta}{\alpha} \right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases} \quad (2.1)$$

For $s \in \mathbb{C}$ with the real part of s greater than 0, recall the Γ -function (see [Fel]) is defined by

$$\Gamma(s) = \int_0^\infty e^{-u} u^{s-1} du = \int_0^\infty e^{-u} u^s \frac{du}{u}. \quad (2.2)$$

Letting $\mu_{\alpha, \beta, \gamma}$ denote the mean of $f(x; \alpha, \beta, \gamma)$, we have

$$\begin{aligned} \mu_{\alpha, \beta, \gamma} &= \int_\beta^\infty x \cdot \frac{\gamma}{\alpha} \left(\frac{x-\beta}{\alpha} \right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma} dx \\ &= \int_\beta^\infty \alpha \frac{x-\beta}{\alpha} \cdot \frac{\gamma}{\alpha} \left(\frac{x-\beta}{\alpha} \right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma} dx + \beta. \end{aligned} \quad (2.3)$$

We change variables by setting $u = \left(\frac{x-\beta}{\alpha}\right)^\gamma$. Then $du = \frac{\gamma}{\alpha} \left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} dx$ and we have

$$\begin{aligned}\mu_{\alpha,\beta,\gamma} &= \int_0^\infty \alpha u^{\gamma-1} \cdot e^{-u} du + \beta \\ &= \alpha \int_0^\infty e^{-u} u^{1+\gamma-1} \frac{du}{u} + \beta \\ &= \alpha \Gamma(1 + \gamma^{-1}) + \beta.\end{aligned}\tag{2.4}$$

A similar calculation determines the variance. We record these results:

Lemma 2.1. *The mean $\mu_{\alpha,\beta,\gamma}$ and variance $\sigma_{\alpha,\beta,\gamma}^2$ of a Weibull with parameters (α, β, γ) are*

$$\begin{aligned}\mu_{\alpha,\beta,\gamma} &= \alpha \Gamma(1 + \gamma^{-1}) + \beta \\ \sigma_{\alpha,\beta,\gamma}^2 &= \alpha^2 \Gamma(1 + 2\gamma^{-1}) - \alpha^2 \Gamma(1 + \gamma^{-1})^2.\end{aligned}\tag{2.5}$$

We can now prove our main result:

Theorem 2.2 (Pythagorean Won-Loss Formula). *Let the runs scored and runs allowed per game be two independent random variables drawn from Weibull distributions with parameters $(\alpha_{\text{RS}}, \beta, \gamma)$ and $(\alpha_{\text{RA}}, \beta, \gamma)$ respectively, where α_{RS} and α_{RA} are chosen so that the means are RS and RA. If $\gamma > 0$ then*

$$\text{Won-Loss Percentage}(\text{RS}, \text{RA}, \beta, \gamma) = \frac{(\text{RS} - \beta)^\gamma}{(\text{RS} - \beta)^\gamma + (\text{RA} - \beta)^\gamma}.\tag{2.6}$$

Proof. Let X and Y be independent random variables with Weibull distributions $(\alpha_{\text{RS}}, \beta, \gamma)$ and $(\alpha_{\text{RA}}, \beta, \gamma)$ respectively, where X is the number of runs scored and Y the number of runs allowed per game. As the means are RS and RA, by Lemma 2.1 we have

$$\begin{aligned}\text{RS} &= \alpha_{\text{RS}} \Gamma(1 + \gamma^{-1}) + \beta \\ \text{RA} &= \alpha_{\text{RA}} \Gamma(1 + \gamma^{-1}) + \beta.\end{aligned}\tag{2.7}$$

Equivalently, we have

$$\begin{aligned}\alpha_{\text{RS}} &= \frac{\text{RS} - \beta}{\Gamma(1 + \gamma^{-1})} \\ \alpha_{\text{RA}} &= \frac{\text{RA} - \beta}{\Gamma(1 + \gamma^{-1})}.\end{aligned}\tag{2.8}$$

We need only calculate the probability that X exceeds Y . Below we constantly use the integral of a probability density is 1. We have

$$\begin{aligned}\text{Prob}(X > Y) &= \int_{x=\beta}^\infty \int_{y=\beta}^x f(x; \alpha_{\text{RS}}, \beta, \gamma) f(y; \alpha_{\text{RA}}, \beta, \gamma) dy dx \\ &= \int_{x=\beta}^\infty \int_{y=\beta}^x \frac{\gamma}{\alpha_{\text{RS}}} \left(\frac{x-\beta}{\alpha_{\text{RS}}}\right)^{\gamma-1} e^{-((x-\beta)/\alpha_{\text{RS}})^\gamma} \frac{\gamma}{\alpha_{\text{RA}}} \left(\frac{y-\beta}{\alpha_{\text{RA}}}\right)^{\gamma-1} e^{-((y-\beta)/\alpha_{\text{RA}})^\gamma} dy dx \\ &= \int_{x=0}^\infty \frac{\gamma}{\alpha_{\text{RS}}} \left(\frac{x}{\alpha_{\text{RS}}}\right)^{\gamma-1} e^{-(x/\alpha_{\text{RS}})^\gamma} \left[\int_{y=0}^x \frac{\gamma}{\alpha_{\text{RA}}} \left(\frac{y}{\alpha_{\text{RA}}}\right)^{\gamma-1} e^{-(y/\alpha_{\text{RA}})^\gamma} dy \right] dx \\ &= \int_{x=0}^\infty \frac{\gamma}{\alpha_{\text{RS}}} \left(\frac{x}{\alpha_{\text{RS}}}\right)^{\gamma-1} e^{-(x/\alpha_{\text{RS}})^\gamma} \left[1 - e^{-(x/\alpha_{\text{RA}})^\gamma} \right] dx \\ &= 1 - \int_{x=0}^\infty \frac{\gamma}{\alpha_{\text{RS}}} \left(\frac{x}{\alpha_{\text{RS}}}\right)^{\gamma-1} e^{-(x/\alpha)^\gamma} dx,\end{aligned}\tag{2.9}$$

where we have set

$$\frac{1}{\alpha^\gamma} = \frac{1}{\alpha_{\text{RS}}^\gamma} + \frac{1}{\alpha_{\text{RA}}^\gamma} = \frac{\alpha_{\text{RS}}^\gamma + \alpha_{\text{RA}}^\gamma}{\alpha_{\text{RS}}^\gamma \alpha_{\text{RA}}^\gamma}.\tag{2.10}$$

Therefore

$$\begin{aligned}
\text{Prob}(X > Y) &= 1 - \frac{\alpha^\gamma}{\alpha_{\text{RS}}^\gamma} \int_0^\infty \frac{\gamma}{\alpha} \left(\frac{x}{\alpha}\right)^{\gamma-1} e^{-(x/\alpha)^\gamma} dx \\
&= 1 - \frac{\alpha^\gamma}{\alpha_{\text{RS}}^\gamma} \\
&= 1 - \frac{1}{\alpha_{\text{RS}}^\gamma} \frac{\alpha_{\text{RS}}^\gamma \alpha_{\text{RA}}^\gamma}{\alpha_{\text{RS}}^\gamma + \alpha_{\text{RA}}^\gamma} \\
&= \frac{\alpha_{\text{RS}}^\gamma}{\alpha_{\text{RS}}^\gamma + \alpha_{\text{RA}}^\gamma}. \tag{2.11}
\end{aligned}$$

Substituting the relations for α_{RS} and α_{RA} of (2.8) into (2.11) yields

$$\text{Prob}(X > Y) = \frac{(\text{RS} - \beta)^\gamma}{(\text{RS} - \beta)^\gamma + (\text{RA} - \beta)^\gamma}, \tag{2.12}$$

which completes the proof of Theorem 2.2. \square

Remark 2.3. The reason the integrations can be so easily performed (determining the normalization constants, the mean and variance, as well as calculating the probability that X exceeds Y) is that we have terms such as $e^{-u^\gamma} u^{\gamma-1}$; these are very easy to integrate. It is essential, however, that we also have a tractable expression for the mean in terms of the parameters. Fortunately this is possible as the mean is a simple combination of the Γ -function and the parameters. As we fix γ and then choose α_{RS} or α_{RA} , it is important that the argument of the Γ -function only involve γ and not α_{RS} or α_{RA} . If the argument of the Γ -function involved α_{RS} or α_{RA} , then we would have to solve equations of the form $\text{RS} = g(\alpha_{\text{RS}}, \gamma) \Gamma(h(\alpha_{\text{RS}}, \gamma))$ for some functions g and h . Inverting this to solve for α_{RS} as a function of γ and RS would be difficult in general. Finally we remark that the essential aspect of (2.8) is that α_{RS} is proportional to $\text{RS} - \beta$. It does not matter that the proportionality constant involves γ . While it is difficult to solve $\Gamma(1 + \gamma^{-1}) = z$ for γ , we do not need to; these factors cancel.

Remark 2.4. We take $\gamma > 0$ as if $\gamma < 0$ then (2.6) (while still true) is absurd. For example, if $\gamma = -.5$, $\beta = 0$, $\text{RS} = 25$ and $\text{RA} = 16$, then (2.6) predicts a winning percentage of

$$\frac{25^{-1/2}}{25^{-1/2} + 16^{-1/2}} = \frac{4}{9} < \frac{1}{2}; \tag{2.13}$$

thus a team that scores more runs than it allows is predicted to have a losing season! Of course, when $\gamma \leq 0$ we have a very strange probability distribution. Not only is the behavior near $x = 0$ interesting but we no longer have rapid decay at infinity (the probability now falls off as $x^{\gamma-1}$), and this is unlikely to be a realistic model.

3. NUMERICAL RESULTS: AMERICAN LEAGUE 2004

We analyzed the 14 teams¹ of the American League from the 2004 season in order to determine the reasonableness of the assumptions in our model; we leave the National League teams as an exercise to the reader. We used the method of least squares² and the method of maximum

¹The teams are ordered by division (AL East, AL Central, AL West) and then by number of regular season wins, with the exception of the Boston Red Sox who as World Series champions are listed first.

²We minimized the sum of squares of the error from the runs scored data plus the sum of squares of the error from the runs allowed data; as $\beta = -.5$ there were three free parameters: α_{RS} , α_{RA} and γ . Specifically, let $\text{Bin}(k)$ be the k^{th} bin from (1.4). If $\text{RS}_{\text{obs}}(k)$ (resp. $\text{RA}_{\text{obs}}(k)$) denotes the observed number of games with the number of runs scored (allowed) in $\text{Bin}(k)$, and $A(\alpha, \beta, \gamma, k)$ denotes the area under the Weibull with parameters (α, β, γ) in $\text{Bin}(k)$, then for each team we found the values of $(\alpha_{\text{RS}}, \alpha_{\text{RA}}, \gamma)$ that minimized

$$\sum_{k=1}^{\#\text{Bins}} (\text{RS}_{\text{obs}}(k) - \#\text{Games} \cdot A(\alpha_{\text{RS}}, -.5, \gamma, k))^2 + \sum_{k=1}^{\#\text{Bins}} (\text{RA}_{\text{obs}}(k) - \#\text{Games} \cdot A(\alpha_{\text{RA}}, -.5, \gamma, k))^2. \tag{3.1}$$

likelihood³ with the bins of (1.4). For each team we simultaneously found the best fit Weibulls of the form $(\alpha_{RS}, -.5, \gamma)$ and $(\alpha_{RA}, -.5, \gamma)$. We then compared the predicted number of wins, losses, and won-loss percentage with the actual data:

Results from the Method of Least Squares						
Team	Obs Wins	Pred Wins	ObsPerc	PredPerc	GamesDiff	γ
Boston Red Sox	98	94.4	0.605	0.582	3.6	1.80
New York Yankees	101	92.2	0.623	0.569	8.8	1.77
Baltimore Orioles	78	84.0	0.481	0.518	-6.0	1.63
Tampa Bay Devil Rays	70	71.9	0.435	0.446	-1.9	1.82
Toronto Blue Jays	67	64.2	0.416	0.399	2.8	2.01
Minnesota Twins	92	91.6	0.568	0.566	0.4	1.80
Chicago White Sox	83	81.6	0.512	0.503	1.4	1.71
Cleveland Indians	80	80.2	0.494	0.495	-0.2	1.81
Detroit Tigers	72	78.1	0.444	0.482	-6.1	1.76
Kansas City Royals	58	61.8	0.358	0.381	-3.8	1.80
Los Angeles Angels	92	92.3	0.568	0.570	-0.3	1.68
Oakland Athletics	91	79.6	0.562	0.491	11.4	1.79
Texas Rangers	89	86.7	0.549	0.535	2.3	1.88
Seattle Mariners	63	72.8	0.389	0.449	-9.8	1.76

Results from the Method of Maximum Likelihood						
Team	Obs Wins	Pred Wins	ObsPerc	PredPerc	GamesDiff	γ
Boston Red Sox	98	93.0	0.605	0.574	5.03	1.82
New York Yankees	101	87.5	0.623	0.540	13.49	1.78
Baltimore Orioles	78	83.1	0.481	0.513	-5.08	1.66
Tampa Bay Devil Rays	70	69.6	0.435	0.432	0.38	1.83
Toronto Blue Jays	67	74.6	0.416	0.464	-7.65	1.97
Minnesota Twins	92	84.7	0.568	0.523	7.31	1.79
Chicago White Sox	83	85.3	0.512	0.527	-2.33	1.73
Cleveland Indians	80	80.0	0.494	0.494	0.	1.79
Detroit Tigers	72	80.0	0.444	0.494	-8.02	1.78
Kansas City Royals	58	68.7	0.358	0.424	-10.65	1.76
Los Angeles Angels	92	87.5	0.568	0.540	4.53	1.71
Oakland Athletics	91	84.0	0.562	0.519	6.99	1.76
Texas Rangers	89	87.3	0.549	0.539	1.71	1.90
Seattle Mariners	63	70.7	0.389	0.436	-7.66	1.78

Using the method of least squares, the mean of γ over the 14 teams is 1.79 with a standard deviation is .09 (the median is 1.79); using the method of maximum likelihood the mean of γ over the 14 teams is 1.74 with a standard deviation of .06 (the median is 1.76). Note that the numerically observed best exponent of 1.82 is well within this region for both approaches.

We now consider how close the estimates of team performance are to the observed season records. For the method of least squares, over the 14 teams the mean number of the difference between observed and predicted wins was 0.19 with a standard deviation of 5.69 (and a median of 0.07); if we consider just the absolute value of the difference then we have a mean of 4.19 with a standard deviation of 3.68 (and a median of 3.22). For the method of maximum likelihood, over the 14 teams the mean number of the difference between observed and predicted wins was -0.13 with a standard deviation of 7.11 (and a median of 0.19); if we consider just the absolute value of the difference then we have a mean of 5.77 with a standard deviation of 3.85 (and a median of 6.04). This is consistent with the observation that the Pythagorean Formula is usually accurate to about four games in a 162 game season.

For the remainder of the paper, we analyze the fits from the method of maximum likelihood; these fits were slightly better than those from the method of least squares. The estimates from the

³Notation as in Footnote 2, the likelihood function of the sample is

$$L(\alpha_{RS}, \alpha_{RA}, -.5, \gamma) = \frac{\#Games}{RS_{obs}(1), \dots, RS_{obs}(\#Bins)} \prod_{k=1}^{\#Bins} A(\alpha_{RS}, -.5, \gamma, k)^{RS_{obs}(k)} \cdot \frac{\#Games}{RA_{obs}(1), \dots, RA_{obs}(\#Bins)} \prod_{k=1}^{\#Bins} A(\alpha_{RA}, -.5, \gamma, k)^{RA_{obs}(k)}. \quad (3.2)$$

For each team we find the values of the parameters α_{RS} , α_{RA} and γ that maximize the likelihood. Computationally, it is equivalent to maximize the logarithm of the likelihood, and we may ignore the multinomial coefficients as they are independent of the parameters.

method of maximum likelihood enjoy many desirable properties, including being asymptotically minimum variance unbiased estimators and yielding sufficient estimators (whenever they exist).

We performed χ^2 tests to determine the goodness of the fit from the best fit Weibulls from the method of maximum likelihood⁴. For the Weibulls approximating the runs scored and allowed per game we used the bins of (1.4):

$$[-.5, .5] \cup [.5, 1.5] \cup \dots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty). \quad (3.4)$$

There are 20 degrees of freedom for these tests. For 20 degrees of freedom the critical thresholds are 31.41 (at the 95% level) and 37.57 (at the 99% level).

We also tested the independence of the runs scored and runs allowed per game (a crucial input for our model). As this test requires each row and column to have at least one non-zero entry, here we broke the runs scored and allowed into bins

$$[0, 1) \cup [1, 2) \cup [2, 3) \cup \dots \cup [8, 9) \cup [9, 10) \cup [10, 11) \cup [11, \infty). \quad (3.5)$$

This gives us an $r \times c$ contingency table (with $r = c = 12$); however, as the runs scored and allowed per game can never be equal, we actually have an incomplete two-dimensional contingency table with $(12 - 1)^2 - 12 = 109$ degrees of freedom; see [BF, SD]. This complication is not present in the theoretical model, as if the runs scored and allowed are drawn from continuous distributions (in this case, Weibulls), there is zero probability of both values being equal. This difficulty is due to the fact that the runs scored and allowed in a game must be distinct integers. We describe the modified χ^2 test for an incomplete two-dimensional contingency table with diagonal entries forced to be zero (these are called structural or fixed zeros).

Let $\text{Bin}(k)$ denote the k^{th} bin in (3.5). For our 12×12 incomplete contingency table with these bins for both runs scored and allowed, the entry $O_{r,c}$ corresponds to the observed number of games where the team's runs scored is in $\text{Bin}(r)$ and the runs allowed are in $\text{Bin}(c)$; note⁵ $O_{r,r} = 0$ for all r . We use the iterative fitting procedure given in the appendix to [BF] to obtain maximum likelihood estimators for the $E_{r,c}$, the expected frequency of cell (r, c) under the assumption that, given that the runs scored and allowed are distinct, the runs scored and allowed are independent. For $1 \leq r, c \leq 12$, let $E_{r,c}^{(0)} = 1$ if $r \neq c$ and 0 if $r = c$. Set

$$X_{r,+} = \sum_{c=1}^{12} O_{r,c}, \quad X_{+,c} = \sum_{r=1}^{12} O_{r,c}. \quad (3.6)$$

Then

$$E_{r,c}^{(\ell)} = \begin{cases} E_{r,c}^{(\ell-1)} X_{r,+} / \sum_{c=1}^{12} E_{r,c}^{(\ell-1)} & \text{if } \ell \text{ is odd} \\ E_{r,c}^{(\ell-1)} X_{+,c} / \sum_{r=1}^{12} E_{r,c}^{(\ell-1)} & \text{if } \ell \text{ is even,} \end{cases} \quad (3.7)$$

and

$$E_{r,c} = \lim_{\ell \rightarrow \infty} E_{r,c}^{(\ell)}; \quad (3.8)$$

⁴Using the bins from (1.4) (and the rest of the notation as in Footnote 2), we studied

$$\sum_{k=1}^{\#\text{Bins}} \frac{(\text{RS}_{\text{obs}}(k) - \#\text{Games} \cdot A(\alpha_{\text{RS}}, -.5, \gamma, k))^2}{\#\text{Games} \cdot A(\alpha_{\text{RS}}, -.5, \gamma, k)} + \sum_{k=1}^{\#\text{Bins}} \frac{(\text{RA}_{\text{obs}}(k) - \#\text{Games} \cdot A(\alpha_{\text{RA}}, -.5, \gamma, k))^2}{\#\text{Games} \cdot A(\alpha_{\text{RA}}, -.5, \gamma, k)}. \quad (3.3)$$

This has a χ^2 distribution with $2(\#\text{Bins} - 1) - 1 - 3 = 20$ degrees of freedom (the factor of 3 which we subtract arises from estimating three parameters, α_{RS} , α_{RA} and γ ; β was not estimated, as it was taken to be $-.5$).

⁵The reason $O_{r,r}$ should equal zero is that a team cannot score and allow the same number of runs in a game, as baseball does not allow ties (except for an occasional All-star game). The first 11 bins each contain exactly one score, so for $r \leq 11$, $O_{r,r} = 0$. The final bin, however, contains all scores from 11 to ∞ , and thus it is possible for the runs scored and allowed to be unequal and both in this bin; however, the probability is so small here that we may simply replace all runs scored or allowed exceeding 11 with 11. Of the 14 teams, 7 have $O_{12,12} = 0$, 5 (teams 3, 4, 5, 9 and 12) have $O_{12,12} = 1$, 1 (team 8) has $O_{12,12} = 2$ and 1 (team 7) has $O_{12,12} = 3$.

the iterations converge very quickly in practice⁶. Then

$$\sum_{r=1}^{12} \sum_{\substack{c=1 \\ c \neq r}}^{12} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}} \quad (3.9)$$

is approximately a χ^2 distribution with $(12 - 1)^2 - 12 = 109$ degrees of freedom. The corresponding critical thresholds are 134.4 (at the 95% level) and 146.3 (at the 99% level).

We summarize our results below; the first column is the χ^2 tests for the goodness of fit from the best fit Weibulls, and the second column is the χ^2 tests for the independence of the runs scored and runs allowed.

Results from the Method of Maximum Likelihood		
Team	RS+RA χ^2 : 20 d.f.	Indep χ^2 : 109 d.f.
Boston Red Sox	15.63	83.19
New York Yankees	12.60	129.13
Baltimore Orioles	29.11	116.88
Tampa Bay Devil Rays	13.67	111.08
Toronto Blue Jays	41.18	100.11
Minnesota Twins	17.46	97.93
Chicago White Sox	22.51	153.07
Cleveland Indians	17.88	107.14
Detroit Tigers	12.50	131.27
Kansas City Royals	28.18	111.45
Los Angeles Angels	23.19	125.13
Oakland Athletics	30.22	133.72
Texas Rangers	16.57	111.96
Seattle Mariners	21.57	141.00

Except for the Weibulls for the runs scored and allowed for the Toronto Blue Jays, and the independence of runs scored and runs allowed for the Chicago White Sox⁷, all test statistics are well below the 95% critical threshold (31.41 as there are 20 degrees of freedom). As we are performing multiple comparisons, chance fluctuations should make some differences appear significant (for example, if the null hypothesis is true and 10 independent tests are performed, there is about a 40% chance of observing at least one statistically significant difference at the 95% confidence level). We must therefore adjust the confidence levels. Using the common, albeit conservative, Bonferroni⁸ adjustment method for multiple comparisons, at the 95% confidence level we find significant fits for all but the Toronto Blue Jays' runs scored and allowed and the independence of runs scored and allowed for the Chicago White Sox; however, both just barely miss at the 95% confidence level (41.18 versus 41.14 for the Blue Jays, and 153.07 versus 152.9 for the White Sox). Thus the data validates our assumption that, given that runs scored and allowed cannot be equal, the runs scored and allowed per game are statistically independent events, and that the parameters from the method of maximum likelihood give good fits to the observed distribution of scores. In Appendix A we provide plots comparing the observed distribution of runs scored and allowed versus the best fit predictions, where even a visual inspection shows the agreement between our theory and the data.

Using the best fit parameters of the Weibulls, Lemma 2.1 provides an estimate for the mean number of runs scored and allowed per game. We are of course primarily interested in estimating γ and not the mean number of runs scored or allowed per game, because these are of course known from the season data; however, this provides an additional test to see how well our theory agrees with the data.

⁶If we had a complete two-dimensional contingency table, then the iteration reduces to the standard values, namely $E_{r,c} = \sum_{c'} O_{r,c'} \cdot \sum_{r'} O_{r',c} / \#Games$.

⁷The Chicago White Sox had the largest value of $O_{12,12}$ in the independence tests, namely 3. If we replace the last bin in (3.5) with two bins, [11, 12) and [12, ∞), then $r = c = 13$, $O_{12,12} = 0$ and $O_{13,13} = 1$. There are $(13 - 1)^2 - 13 = 131$ degrees of freedom. The corresponding critical thresholds are 158.7 (at the 95% level) and 171.6 (at the 99% level), and the observed value of the χ^2 statistic for the Chicago White Sox is 164.8.

⁸Using the Bonferroni adjustment for multiple comparisons divides the significance level α by the number of comparisons, which in our case is 14. Thus for the Weibull tests with 20 degrees of freedom the adjusted critical thresholds are 41.14 (at the 95% level) and 46.38 (at the 99% level); for the independence tests with 109 degrees of freedom the adjusted critical thresholds are 152.9 (at the 95% level) and 162.2 (at the 99% level).

As the number of games each team played is so large⁹, we use a z -test to compare the observed versus predicted means. The critical z -values are 1.96 (at the 95% confidence level) and 2.575 (at the 99% confidence level).

Results from the Method of Maximum Likelihood						
Team	Obs RS	Pred RS	z -stat	Obs RA	Pred RA	z -stat
Boston Red Sox	5.86	5.80	0.24	4.74	4.83	-0.35
New York Yankees	5.54	5.47	0.24	4.99	4.95	0.12
Baltimore Orioles	5.20	5.26	-0.22	5.12	5.08	0.16
Tampa Bay Devil Rays	4.43	4.41	0.12	5.23	5.21	0.09
Toronto Blue Jays	4.47	4.51	-0.18	5.11	4.96	0.59
Minnesota Twins	4.81	4.74	0.32	4.41	4.48	-0.28
Chicago White Sox	5.34	5.40	-0.22	5.13	5.05	0.34
Cleveland Indians	5.30	5.18	0.40	5.29	5.26	0.09
Detroit Tigers	5.10	5.06	0.18	5.21	5.13	0.27
Kansas City Royals	4.44	4.48	-0.13	5.59	5.46	0.48
Los Angeles Angels	5.16	5.10	0.22	4.53	4.59	-0.22
Oakland Athletics	4.90	4.85	0.18	4.58	4.63	-0.19
Texas Rangers	5.31	5.29	0.05	4.84	4.82	0.08
Seattle Mariners	4.31	4.29	0.10	5.08	5.03	0.18

We note excellent agreement between all the predicted average runs scored per game and the observed average runs scored per game, as well as between all the predicted average runs allowed per game and the observed average runs allowed per game. Performing a Bonferroni adjustment for multiple comparisons gives critical thresholds of 2.914 (at the 95% level) and 3.384 (at the 99% level). At the 95% level (resp. 99% level) all 14 teams have significant fits.

As a final experiment, instead of finding the best fit Weibulls team by team, we performed a similar analysis for each division in the American League in 2004. For example, in the AL East there are 5 teams (the World Champion Boston Red Sox, the New York Yankees, the Baltimore Orioles, the Tampa Bay Devil Rays and the Toronto Blue Jays), and we found the least squares fit to the data with the 11 free parameters

$$\alpha_{RS,BOS}, \alpha_{RA,BOS}, \dots, \alpha_{RS,TOR}, \alpha_{RA,TOR}, \gamma. \quad (3.10)$$

The five teams in the AL East (resp., the five teams of the AL Central and the four teams of the AL West) give a best fit value of γ of 1.793 (resp., 1.773 and 1.774), which again is very close to the numerically observed best value of γ of 1.82. Using the method of maximum likelihood gives best fit values of γ of 1.74 for the AL East, 1.75 for the AL Central and 1.73 for the AL West.

4. CONCLUSIONS AND FUTURE WORK

Bill James' Pythagorean Won-Loss Formula may be derived from very simple and reasonable assumptions (namely, that the runs scored and allowed per game are independent events drawn from Weibulls with the same β and γ). Using the method of least squares or the method of maximum likelihood, we can find the best fit values of these parameters from the observed game scores. Using the method of maximum likelihood, for the 2004 baseball season for each team in the American League the fits were always significant at the 95% confidence level (except for the Toronto Blue Jays, which just missed), the assumption that, given that the runs scored and allowed in a game are distinct integers, the runs scored and allowed per game are independent events was validated, and the best fit exponent γ was about 1.74 with a standard deviation of .06, in excellent agreement with the observation that 1.82 is the best exponent to use in the Pythagorean Formula (the method of least squares gives a best fit value for γ of 1.79 with standard deviation .09). Note that we obtain our value of the exponent γ not by fitting the Pythagorean Formula to the observed won-loss percentages of teams, but rather from an analysis of the distribution of scores from individual baseball games. Assuming teams behave similarly from year to year, there is now a theoretical justification for using the Pythagorean Formula to predict team performances in future seasons (with an exponent around 1.74 to 1.79 and using the observed average runs scored and allowed).

An interesting future project would be to perform a more micro analysis to incorporate lower order effects, though our simple model is quite effective at fitting the data and predicting the

⁹All teams played 162 except for the Tampa Bay Devil Rays and the Toronto Blue Jays, who had a game rained out and only played 161 games in 2004.

best exponent γ (see for example [Sch1, Sch2], where such an analysis is performed to determine the all-time best hitters and sluggers). For example, one might break down runs scored and allowed per inning. If a team has a large lead it often pulls its good hitters to give them a rest, as well as bringing in weaker pitchers to give them some work; conversely, in late innings in close games managers often have pitch-runners for slow good hitters who get on base (to get a run now with a potential cost of runs later through the loss of the better hitter from the lineup), and star relievers (when available) are brought in to maintain the lead. Further there are slight differences because of inter-league play. For example, the American League teams lose their DH for games in National League parks, and thus we expect the run production to differ from that in American League parks. Further, using the analysis in [Sch1, Sch2] one can incorporate ballpark effects (some ballparks favor pitchers while others favor hitters). Such an analysis might lead to new statistics of adjusted runs scored and allowed per game. Additionally, teams out of the playoff race often play their last few games differently than when they are still in contention, and perhaps those games should be removed from the analysis.

One can also further examine the independence of runs scored and allowed. As baseball games cannot end in a tie, runs scored and allowed are never equal in a game; however, they can be equal after 9 innings. One avenue for research is to classify extra-inning games as ties (while recording which team eventually won). Also, if the home team is leading after the top of the ninth then it does not bat, and this will effect its run production. See [Ci] for an analysis of some of these issues.

Finally, it would be fascinating to see if this (or a similar) model is applicable to other sports with long seasons. While football has a relatively short season of 16 games, basketball and hockey have 82 games a season. The scores in basketball are more spread out than hockey, which is more compact than baseball; it would be interesting to see what affect these have on the analysis and whether or not the fits are as good as baseball.

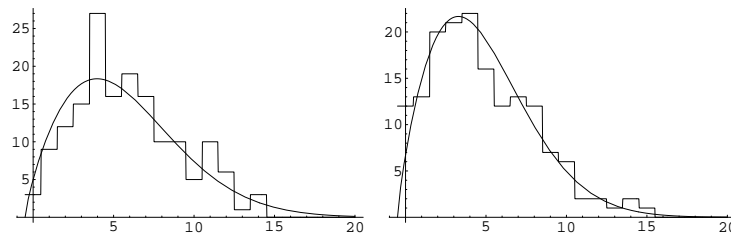
ACKNOWLEDGEMENTS

I would like to thank Russell Mann and Steven Johnson for introducing me to the Pythagorean Formula in baseball, Kevin Dayaratna for inputting much of the baseball data, Jeff Miller for writing a script to read in baseball data from the web to the analysis programs, Gerry Myerson for catching some typos, Eric T. Bradlow for helpful comments on an earlier draft, Ray Ciccolella for discussions on the independence of runs scored and allowed, and Stephen D. Miller for suggesting the National League exercise.

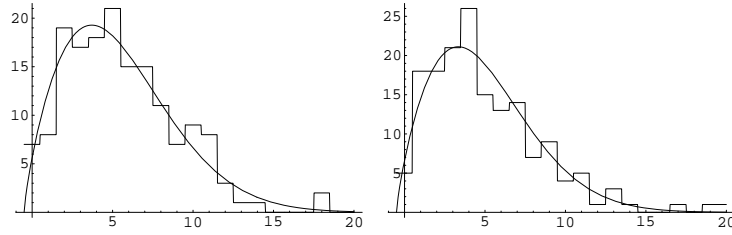
APPENDIX A. PLOTS OF BEST FIT WEIBULLS

Below we plot the best fit Weibulls against the observed histograms of runs scored and allowed. We use the bins of (1.4).

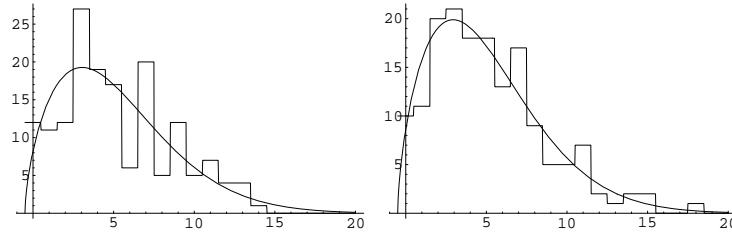
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Boston Red Sox



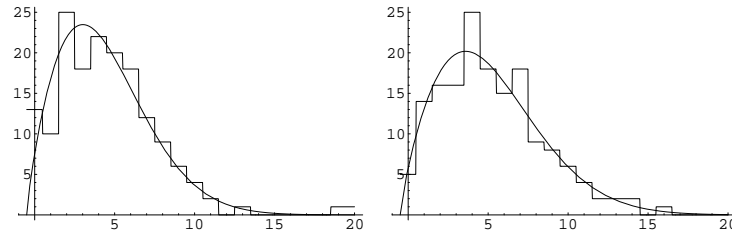
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the New York Yankees



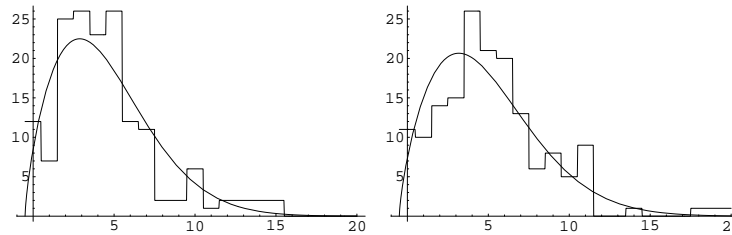
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Baltimore Orioles



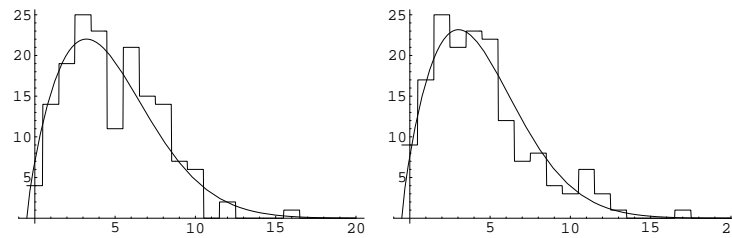
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Tampa Bay Devil Rays



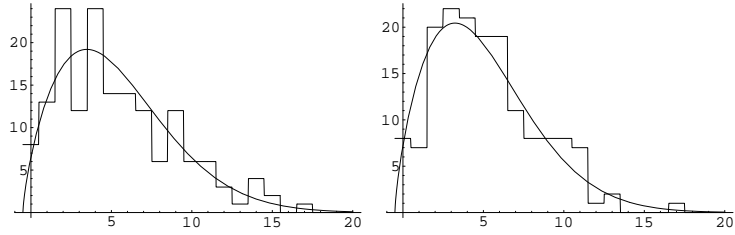
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Toronto Blue Jays



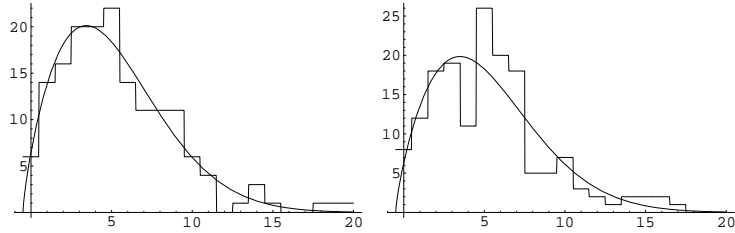
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Minnesota Twins



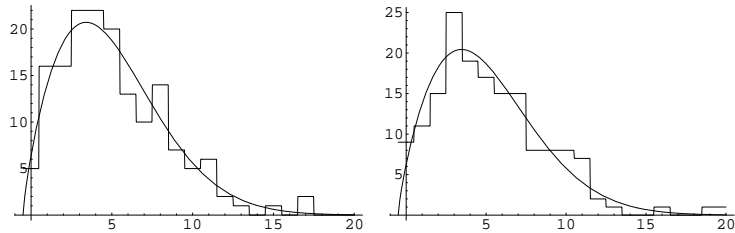
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Chicago White Sox



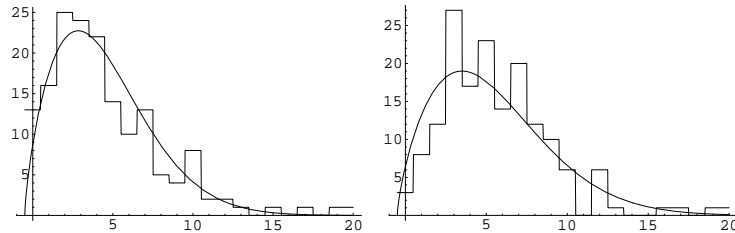
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Cleveland Indians



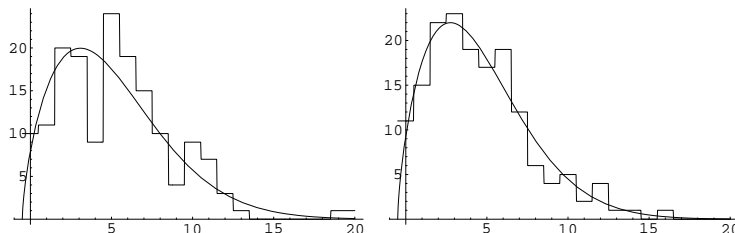
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Detroit Tigers



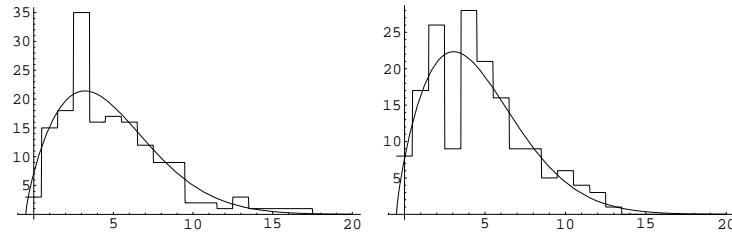
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Kansas City Royals



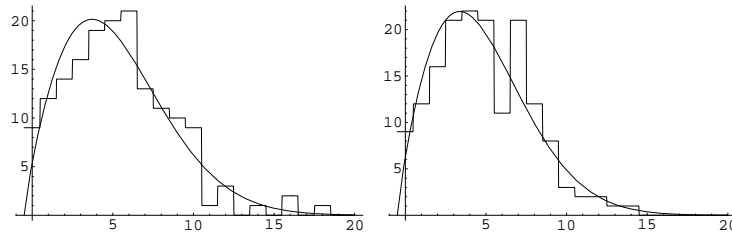
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Los Angeles Angels



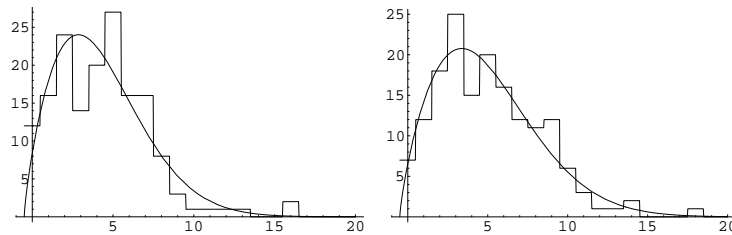
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Oakland Athletics



Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Texas Rangers



Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Seattle Mariners



REFERENCES

- [Al] Baseball Almanac, <http://baseball-almanac.com>.
- [An] J. Angus, <http://forums.mlbcenter.com/showthread.php?t=14121>.
- [BF] Y. M. M. Bishop and S. E. Fienberg, *Incomplete Two-Dimensional Contingency Tables*, *Biometrics* **25** (1969), no. 1, 119–128.
- [BFAM] E. T. Bradlow, P. S. Fader, M. Adrian and B. McShane, *Count Models Based on Weibull Interarrival Times*, preprint.
- [CaBe] G. Casella and R. Berger, *Statistical Inference*, Second Edition, Duxbury Advanced Series, 2002.
- [Ci] Ray Ciccolella, *Are Runs Scored and Runs Allowed Independent?*, preprint.
- [Fe1] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. I.*, third edition, Wiley, New York 1968.
- [Fe2] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. II.*, third edition, Wiley, New York 1971.
- [Ja] B. James, *Baseball Abstract 1983*, Ballantine, 238 pages.
- [Ol] L. Dean Oliver, <http://www.rawbw.com/~deano/helpscrn/pyth.html>.
- [SD] I. R. Savage and K. W. Deutsch, *A Statistical Model of the Gross Analysis of Transaction Flows* *Econometrica* **28** (1960), no. 3 551–572.
- [Sch1] M. J. Schell, *Baseball's All-Time Best Hitters*, Princeton University Press, Princeton, NJ, 1999.
- [Sch2] M. J. Schell, *Baseball's All-Time Best Sluggers*, Princeton University Press, Princeton, NJ, 2005.

DEPARTMENT OF MATHEMATICS, BROWN UNIVERSITY, 151 THAYER STREET, PROVIDENCE, RI 02912
 E-mail address: sjmill@math.brown.edu