The Pythagorean Won-Loss Formula in Baseball: 
An Introduction to Statistics and Modeling

Steven J. Miller
Brown University

Westfield, MA, April 8th, 2006
http://www.math.brown.edu/~sjmiller
Probability Review

Probability density:

• $p(x) \geq 0$;
• $\int_{-\infty}^{\infty} p(x) \, dx = 1$;
• $X$ random variable with density $p(x)$: \( \text{Prob} \ (X \in [a, b]) = \int_{a}^{b} p(x) \, dx \).
Probability Review

Probability density:

• \( p(x) \geq 0; \)

• \( \int_{-\infty}^{\infty} p(x)\,dx = 1; \)

• \( X \) random variable with density \( p(x) \): \( \text{Prob} ( X \in [a, b]) = \int_{a}^{b} p(x)\,dx. \)

Mean (average value) \( \mu = \int_{-\infty}^{\infty} xp(x)\,dx. \)

Variance (how spread out) \( \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)\,dx. \)
Probability Review

Probability density:

- \( p(x) \geq 0 \);
- \( \int_{-\infty}^{\infty} p(x) dx = 1 \);
- \( X \) random variable with density \( p(x) \): \( \text{Prob} \ (X \in [a, b]) = \int_a^b p(x) dx \).

Mean (average value) \( \mu = \int_{-\infty}^{\infty} xp(x) dx \).

Variance (how spread out) \( \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx \).

Independence: two random variables are independent if knowledge of one does not give knowledge of the other.
Numerical Observation: Pythagorean Won-Loss Formula

Parameters:

- $RS_{obs}$: average number of runs scored per game;

- $RA_{obs}$: average number of runs allowed per game;

- $\gamma$: some parameter, constant for a sport.

Bill James’ Won-Loss Formula (NUMERICAL Observation):

$$\text{Won} - \text{Loss Percentage} = \frac{RS_{obs}\gamma}{RS_{obs}\gamma + RA_{obs}\gamma}$$

For baseball: $\gamma$ originally taken as 2.
Numerical studies show best $\gamma$ is about 1.82.
Guidelines for Modeling:

- Model should capture key features of the system;
- Model should be mathematically tractable (solvable).

In general these are conflicting goals. How should we try and model baseball games?
Modeling the Real World

Guidelines for Modeling:

- Model should capture key features of the system;
- Model should be mathematically tractable (solvable).

In general these are conflicting goals. How should we try and model baseball games?

Possible Model:

- Runs Scored and Runs Allowed independent random variables;
- \( f_{RS}(x) \), \( g_{RA}(y) \): probability density functions for runs scored (allowed).
- Reduced to calculating

\[
\int \left[ \int_{y \leq x} f_{RS}(x) g_{RA}(y) \, dy \right] \, dx \quad \text{or} \quad \sum_{i} \left[ \sum_{j < i} f_{RS}(i) g_{RA}(j) \right].
\]
Problems with the Model

Reduced to calculating

$$\int_x \left[ \int_{y \leq x} f_{RS}(x) g_{RA}(y) \, dy \right] \, dx \quad \text{or} \quad \sum_i \left[ \sum_{j<i} f_{RS}(i) g_{RA}(j) \right].$$

Problems with the model:

- Can the integral (or sum) be completed in closed form?
- Are the runs scored and allowed independent random variables?
- What are $f_{RS}$ and $g_{RA}$?
Three Parameter Weibull

Weibull distribution:

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

- $\alpha$: scale (meters versus centimeters);
- $\beta$: origin (translation, zero point);
- $\gamma$: shape (behavior near $\beta$ and at infinity).

Various values give different shapes, but can we find $\alpha$, $\beta$, $\gamma$ such that it fits observed data? Is the Weibull theoretically tractable?
Weibull Integrations

Let \( f(x; \alpha, \beta, \gamma) \) be the probability density of a Weibull(\( \alpha, \beta, \gamma \)):
\[
f(x; \alpha, \beta, \gamma) = \begin{cases} 
\frac{\gamma}{\alpha} \left( \frac{x-\beta}{\alpha} \right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\
0 & \text{otherwise.}
\end{cases}
\]

For \( s \in \mathbb{C} \) with the real part of \( s \) greater than 0, recall the \( \Gamma \)-function:
\[
\Gamma(s) = \int_{0}^{\infty} e^{-u} u^{s-1} \, du = \int_{0}^{\infty} e^{-u} u^{s-1} \, du.
\]

Let \( \mu_{\alpha, \beta, \gamma} \) denote the mean of \( f(x; \alpha, \beta, \gamma) \).
Weibull Integrations (Continued)

\[
\mu_{\alpha, \beta, \gamma} = \int_{\beta}^{\infty} x \cdot \frac{\gamma}{\alpha} \left( \frac{x - \beta}{\alpha} \right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma} \, dx
\]
Weibull Integrations (Continued)

\[
\mu_{\alpha, \beta, \gamma} = \int_{\beta}^{\infty} x \cdot \frac{\gamma}{\alpha} \left( \frac{x - \beta}{\alpha} \right)^{\gamma-1} e^{-\left(\frac{x - \beta}{\alpha}\right)\gamma} \, dx
\]

\[
= \int_{\beta}^{\infty} \frac{x - \beta}{\alpha} \cdot \frac{\gamma}{\alpha} \left( \frac{x - \beta}{\alpha} \right)^{\gamma-1} e^{-\left(\frac{x - \beta}{\alpha}\right)\gamma} \, dx + \beta.
\]
Weibull Integrations (Continued)

\[ \mu_{\alpha,\beta,\gamma} = \int_{\beta}^{\infty} x \cdot \frac{\gamma}{\alpha} \left( \frac{x - \beta}{\alpha} \right)^{\gamma-1} e^{-\left((x-\beta)/\alpha\right)^\gamma} \, dx \]

\[ = \int_{\beta}^{\infty} \frac{x - \beta}{\alpha} \cdot \frac{\gamma}{\alpha} \left( \frac{x - \beta}{\alpha} \right)^{\gamma-1} e^{-\left((x-\beta)/\alpha\right)^\gamma} \, dx + \beta. \]

Change variables: \( u = \left( \frac{x-\beta}{\alpha} \right)^\gamma \). Then \( du = \frac{\gamma}{\alpha} \left( \frac{x-\beta}{\alpha} \right)^{\gamma-1} \, dx \)
Weibull Integrations (Continued)

\[ \mu_{\alpha, \beta, \gamma} = \int_{\beta}^{\infty} x \cdot \frac{\gamma}{\alpha} \left( \frac{x - \beta}{\alpha} \right)^{\gamma - 1} e^{-((x-\beta)/\alpha)^{\gamma}} \, dx \]

\[ = \int_{\beta}^{\infty} \frac{x - \beta}{\alpha} \cdot \frac{\gamma}{\alpha} \left( \frac{x - \beta}{\alpha} \right)^{\gamma - 1} e^{-((x-\beta)/\alpha)^{\gamma}} \, dx + \beta. \]

Change variables: \( u = \left( \frac{x-\beta}{\alpha} \right)^{\gamma} \). Then \( du = \frac{\gamma}{\alpha} \left( \frac{x-\beta}{\alpha} \right)^{\gamma - 1} \, dx \) and

\[ \mu_{\alpha, \beta, \gamma} = \int_{0}^{\infty} \alpha u^{\gamma - 1} \cdot e^{-u} \, du + \beta \]

\[ = \alpha \int_{0}^{\infty} e^{-u} u^{1+\gamma - 1} \frac{du}{u} + \beta \]

\[ = \alpha \Gamma(1 + \gamma^{-1}) + \beta. \]

A similar calculation determines the variance.
Derivation of the Pythagorean Won-Loss Formula

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

\[
\text{Won-Loss Percentage}(RS, RA, \beta, \gamma) = \frac{(RS - \beta)\gamma}{(RS - \beta)\gamma + (RA - \beta)\gamma}.
\]
Best Fit Weibulls to Data: Method of Least Squares

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.

- Bin\((k)\) is the \(k\)th bin;
- \(RS_{\text{obs}}(k)\) (resp. \(RA_{\text{obs}}(k)\)) the observed number of games with the number of runs scored (allowed) in Bin\((k)\);
- \(A(\alpha, \beta, \gamma, k)\) the area under the Weibull with parameters \((\alpha, \beta, \gamma)\) in Bin\((k)\).

Find the values of \((\alpha_{RS}, \alpha_{RA}, \gamma)\) that minimize

\[
\sum_{k=1}^{\text{Bins}} (RS_{\text{obs}}(k) - \text{Games} \cdot A(\alpha_{RS}, -.5, \gamma, k))^2 \\
+ \sum_{k=1}^{\text{Bins}} (RA_{\text{obs}}(k) - \text{Games} \cdot A(\alpha_{RA}, -.5, \gamma, k))^2.
\]
Best Fit Weibulls to Data (Method of Maximum Likelihood)

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

Using as bins

\([-0.5, 0.5] \cup [0.5, 1.5] \cup \cdots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)\).
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
<table>
<thead>
<tr>
<th>Team</th>
<th>RS+RA $\chi^2$: 20 d.f.</th>
<th>Indep $\chi^2$: 109 d.f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston Red Sox</td>
<td>15.63</td>
<td>83.19</td>
</tr>
<tr>
<td>New York Yankees</td>
<td>12.60</td>
<td>129.13</td>
</tr>
<tr>
<td>Baltimore Orioles</td>
<td>29.11</td>
<td>116.88</td>
</tr>
<tr>
<td>Tampa Bay Devil Rays</td>
<td>13.67</td>
<td>111.08</td>
</tr>
<tr>
<td>Toronto Blue Jays</td>
<td>41.18</td>
<td>100.11</td>
</tr>
<tr>
<td>Minnesota Twins</td>
<td>17.46</td>
<td>97.93</td>
</tr>
<tr>
<td>Chicago White Sox</td>
<td>22.51</td>
<td>153.07</td>
</tr>
<tr>
<td>Cleveland Indians</td>
<td>17.88</td>
<td>107.14</td>
</tr>
<tr>
<td>Detroit Tigers</td>
<td>12.50</td>
<td>131.27</td>
</tr>
<tr>
<td>Kansas City Royals</td>
<td>28.18</td>
<td>111.45</td>
</tr>
<tr>
<td>Los Angeles Angels</td>
<td>23.19</td>
<td>125.13</td>
</tr>
<tr>
<td>Oakland Athletics</td>
<td>30.22</td>
<td>133.72</td>
</tr>
<tr>
<td>Texas Rangers</td>
<td>16.57</td>
<td>111.96</td>
</tr>
<tr>
<td>Seattle Mariners</td>
<td>21.57</td>
<td>141.00</td>
</tr>
</tbody>
</table>

20 d.f.:  31.41 (at the 95% level) and  37.57 (at the 99% level).
109 d.f.: 134.4 (at the 95% level) and  146.3 (at the 99% level).

**Bonferroni Adjustment:**
20 d.f.:  41.14 (at the 95% level) and  46.38 (at the 99% level).
109 d.f.: 152.9  (at the 95% level) and  162.2  (at the 99% level).
### Testing the Model: Data from Method of Maximum Likelihood

<table>
<thead>
<tr>
<th>Team</th>
<th>Obs Wins</th>
<th>Pred Wins</th>
<th>ObsPerc</th>
<th>PredPerc</th>
<th>GamesDiff</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston Red Sox</td>
<td>98</td>
<td>93.0</td>
<td>0.605</td>
<td>0.574</td>
<td>5.03</td>
<td>1.82</td>
</tr>
<tr>
<td>New York Yankees</td>
<td>101</td>
<td>87.5</td>
<td>0.623</td>
<td>0.540</td>
<td>13.49</td>
<td>1.71</td>
</tr>
<tr>
<td>Baltimore Orioles</td>
<td>78</td>
<td>83.1</td>
<td>0.481</td>
<td>0.513</td>
<td>-5.08</td>
<td>1.60</td>
</tr>
<tr>
<td>Tampa Bay Devil Rays</td>
<td>70</td>
<td>69.6</td>
<td>0.435</td>
<td>0.432</td>
<td>0.38</td>
<td>1.80</td>
</tr>
<tr>
<td>Toronto Blue Jays</td>
<td>67</td>
<td>74.6</td>
<td>0.416</td>
<td>0.464</td>
<td>-7.65</td>
<td>1.90</td>
</tr>
<tr>
<td>Minnesota Twins</td>
<td>92</td>
<td>84.7</td>
<td>0.568</td>
<td>0.523</td>
<td>7.31</td>
<td>1.71</td>
</tr>
<tr>
<td>Chicago White Sox</td>
<td>83</td>
<td>85.3</td>
<td>0.512</td>
<td>0.494</td>
<td>-2.33</td>
<td>1.71</td>
</tr>
<tr>
<td>Cleveland Indians</td>
<td>80</td>
<td>80.0</td>
<td>0.494</td>
<td>0.494</td>
<td>0.0</td>
<td>1.71</td>
</tr>
<tr>
<td>Detroit Tigers</td>
<td>72</td>
<td>80.0</td>
<td>0.444</td>
<td>0.494</td>
<td>-8.02</td>
<td>1.71</td>
</tr>
<tr>
<td>Kansas City Royals</td>
<td>58</td>
<td>68.7</td>
<td>0.358</td>
<td>0.424</td>
<td>-10.65</td>
<td>1.71</td>
</tr>
<tr>
<td>Los Angeles Angels</td>
<td>92</td>
<td>87.5</td>
<td>0.568</td>
<td>0.540</td>
<td>4.53</td>
<td>1.71</td>
</tr>
<tr>
<td>Oakland Athletics</td>
<td>91</td>
<td>84.0</td>
<td>0.562</td>
<td>0.519</td>
<td>6.99</td>
<td>1.71</td>
</tr>
<tr>
<td>Texas Rangers</td>
<td>89</td>
<td>87.3</td>
<td>0.549</td>
<td>0.539</td>
<td>1.71</td>
<td>1.90</td>
</tr>
<tr>
<td>Seattle Mariners</td>
<td>63</td>
<td>70.7</td>
<td>0.389</td>
<td>0.436</td>
<td>-7.66</td>
<td>1.71</td>
</tr>
</tbody>
</table>

$\gamma$: mean $= 1.74$, standard deviation $= .06$, median $= 1.76$; close to numerically observed value of 1.82.

The mean number of the difference between observed and predicted wins was $-.13$ 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).
Conclusions

• Can find parameters such that the Weibulls are good fits to the data;

• The runs scored and allowed per game are statistically independent;

• The Pythagorean Won-Loss Formula is a consequence of our model;

• Our best value of $\gamma$ of about 1.74 is close to the observed best 1.82.
Future Work

• Micro-analysis: runs scored and allowed are not entirely independent (big lead, close game), run production smaller for inter-league games in NL parks, et cetera.

• What about other sports? Does the same model work? How does $\gamma$ depend on the sport?

• Are there other probability distributions that give integrals which can be determined in closed form?