**Hudson River Undergraduate Mathematics Conference**

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

**Steven J. Miller**

**Brown University**

Westfield, MA, April $8^{\text{th}}$, $2006$

http://www.math.brown.edu/∼sjmiller

# Probability Review

## Probability density:

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

# Probability Review

Probability density:

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

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

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

# Probability Review

Probability density:

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

Mean (average value) $\mu = \int_{-\infty}^{\infty} x p(x)\mathrm{d}x$.

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

Independence: two random variables are independent if knowledge of one does not give knowledge of the other.

## Numerical Observation: Pythagorean Won-Loss Formula

Parameters:

- $\mathrm{RS}_{\mathrm{obs}}$: average number of runs scored per game;

- $\mathrm{RA}_{\mathrm{obs}}$: average number of runs allowed per game;

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

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

$$\mathrm{Won-Loss\ Percentage} \ = \ \frac{\mathrm{RS}_{\mathrm{obs}}{}^{\gamma}}{\mathrm{RS}_{\mathrm{obs}}{}^{\gamma} + \mathrm{RA}_{\mathrm{obs}}{}^{\gamma}}$$

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

## 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?

# 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;
- $f_{\mathrm{RS}}(x)$, $g_{\mathrm{RA}}(y)$: probability density functions for runs scored (allowed).
- Reduced to calculating

$$\int_x \left[\int_{y \le x} f_{\mathrm{RS}}(x) g_{\mathrm{RA}}(y) \mathrm{d}y\right] \mathrm{d}x \quad \text{or} \quad \sum_i \left[\sum_{j<i} f_{\mathrm{RS}}(i) g_{\mathrm{RA}}(j)\right].$$

# Problems with the Model

Reduced to calculating

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

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 $f_{\mathrm{RS}}$ and $g_{\mathrm{RA}}$?

# 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:

$$\Gamma(s) \;=\; \int_0^\infty e^{-u} u^{s-1} \mathrm{d}u \;=\; \int_0^\infty e^{-u} u^{s} \frac{\mathrm{d}u}{u}.$$

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

# Weibull Integrations (Continued)

$$\mu_{\alpha,\beta,\gamma} = \int_{\beta}^{\infty} x \cdot \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} \mathrm{d}x$$

# Weibull Integrations (Continued)

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

# Weibull Integrations (Continued)

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

Change variables: $u = \left(\frac{x-\beta}{\alpha}\right)^\gamma$. Then $\mathrm{d}u = \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} \mathrm{d}x$

# Weibull Integrations (Continued)

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

Change variables: $u = \left(\frac{x-\beta}{\alpha}\right)^{\gamma}$. Then $\mathrm{d}u = \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} \mathrm{d}x$ and

$$\begin{aligned}
\mu_{\alpha,\beta,\gamma} &= \int_{0}^{\infty} \alpha u^{\gamma^{-1}} \cdot e^{-u} \mathrm{d}u \ + \ \beta \\
&= \alpha \int_{0}^{\infty} e^{-u} u^{1+\gamma^{-1}} \frac{\mathrm{d}u}{u} \ + \ \beta \\
&= \alpha\Gamma(1+\gamma^{-1}) \ + \ \beta.
\end{aligned}$$

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* $(\alpha_{\rm RS}, \beta, \gamma)$ *and* $(\alpha_{\rm RA}, \beta, \gamma)$*;* $\alpha_{\rm RS}$ *and* $\alpha_{\rm RA}$ *are chosen so that the means are* $\rm RS$ *and* $\rm RA$*. If* $\gamma > 0$ *then*

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

## 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.

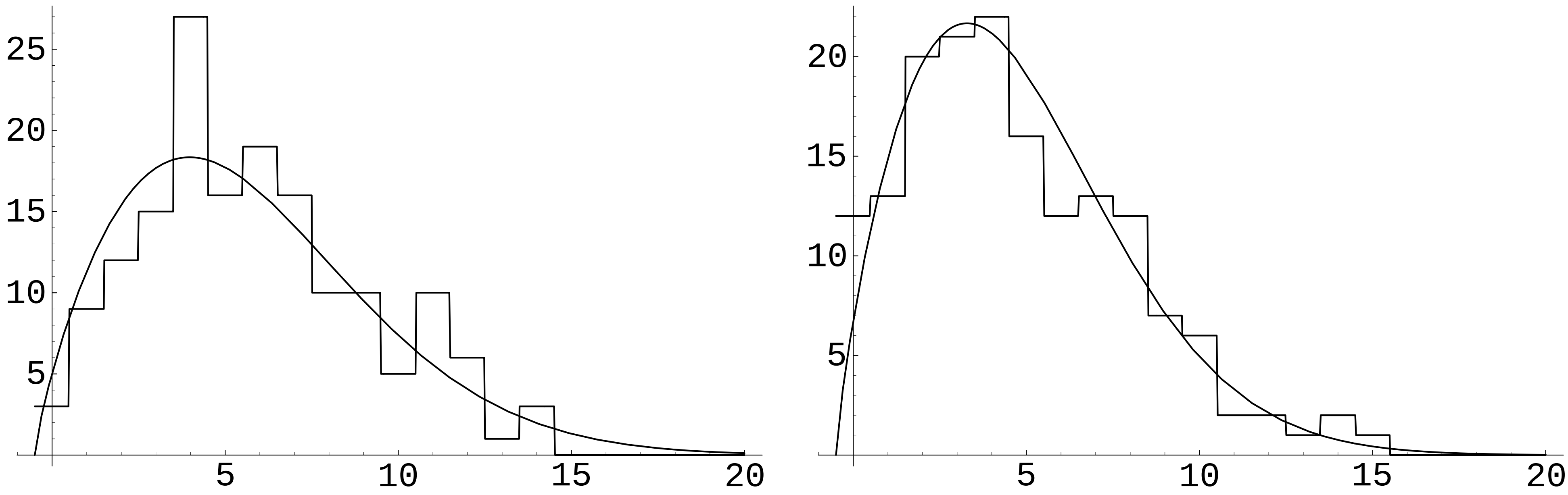
- $\mathrm{Bin}(k)$ is the $k^{\mathrm{th}}$ bin;
- $\mathrm{RS}_{\mathrm{obs}}(k)$ (resp. $\mathrm{RA}_{\mathrm{obs}}(k)$) the observed number of games with the number of runs scored (allowed) in $\mathrm{Bin}(k)$;
- $A(\alpha, \beta, \gamma, k)$ the area under the Weibull with parameters $(\alpha, \beta, \gamma)$ in $\mathrm{Bin}(k)$.

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

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

# Best Fit Weibulls to Data (Method of Maximum Likelihood)

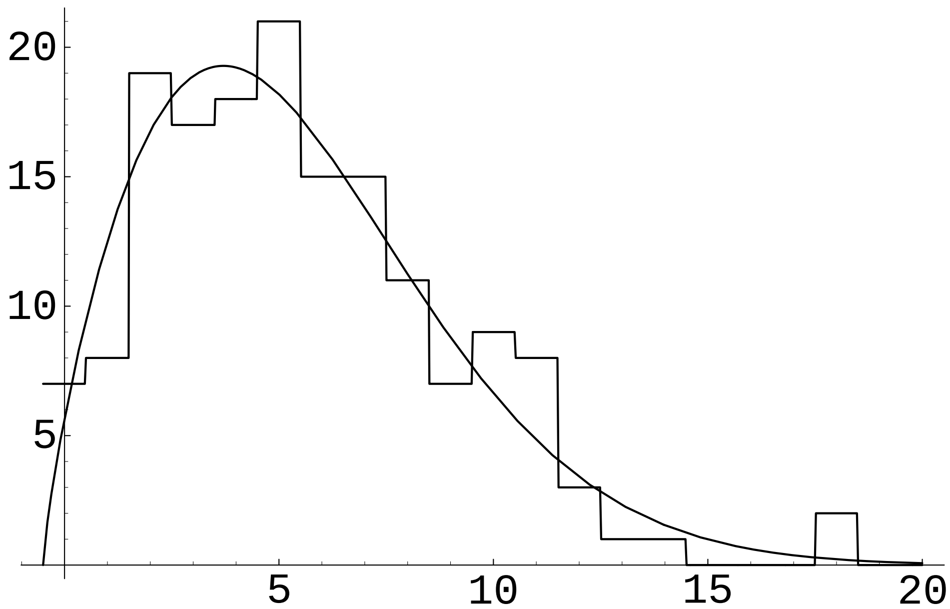
**Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Boston Red Sox**

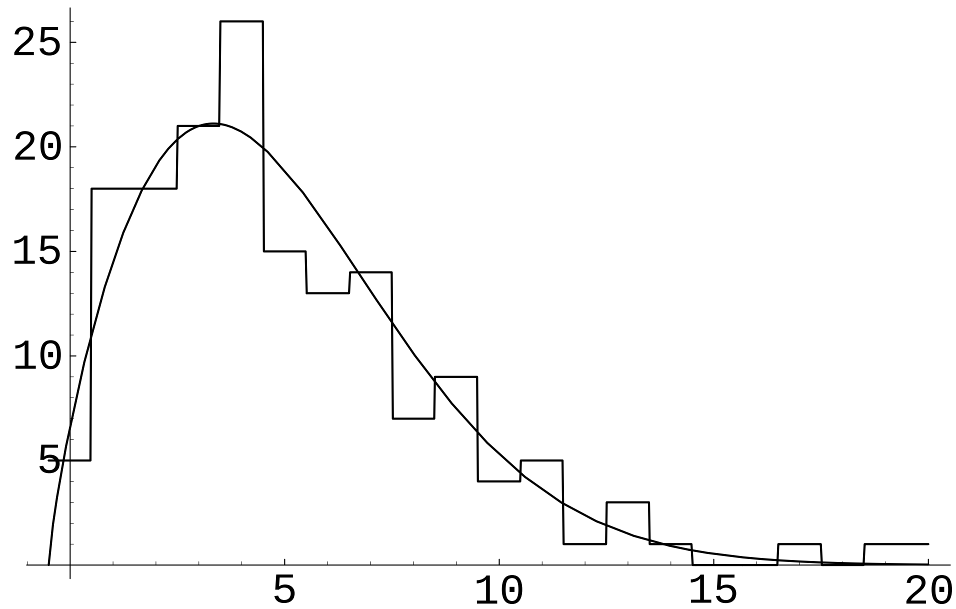



Using as bins

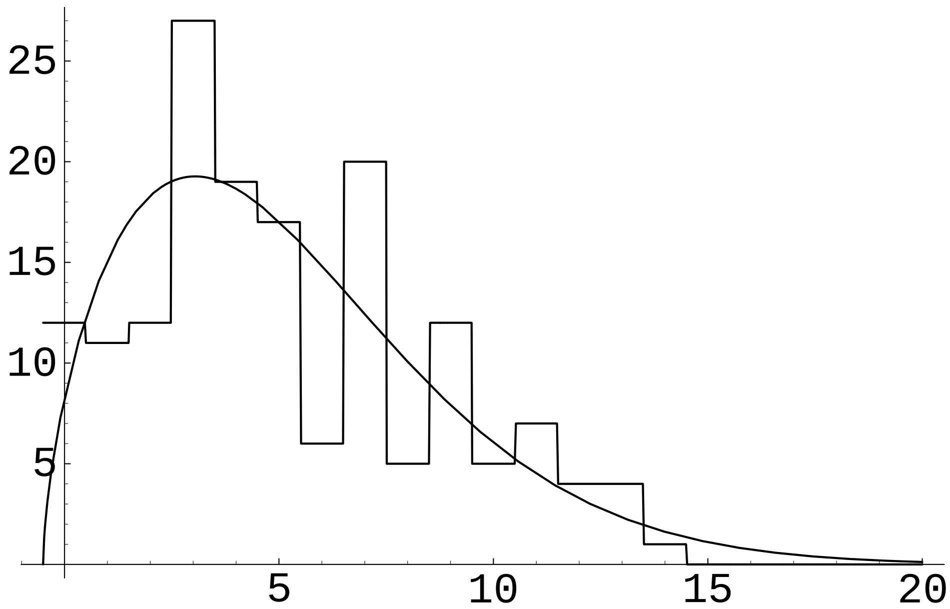
$$[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty).$$

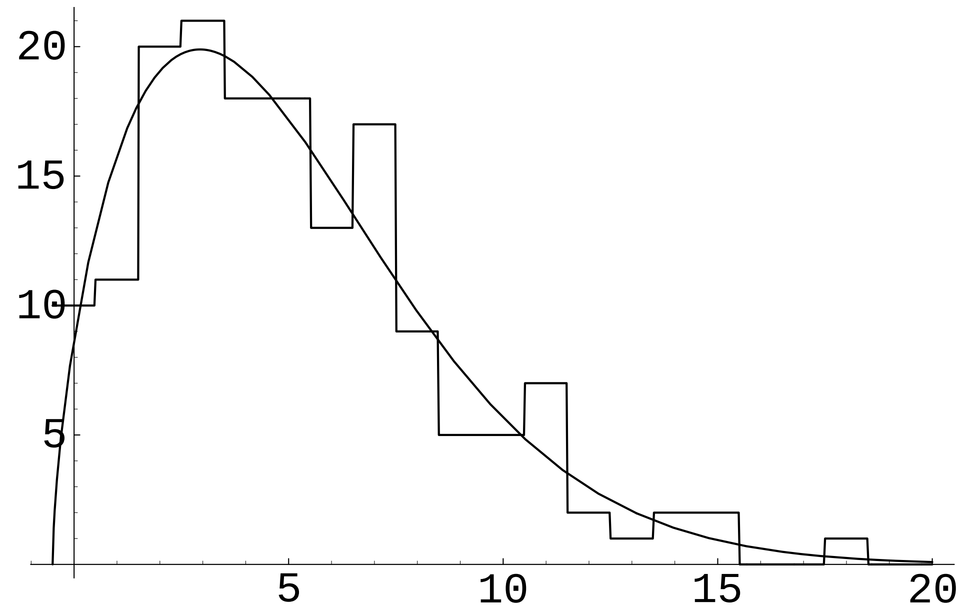
**Plots of RS (predicted vs observed) and RA (predicted vs observed) for the New York Yankees**



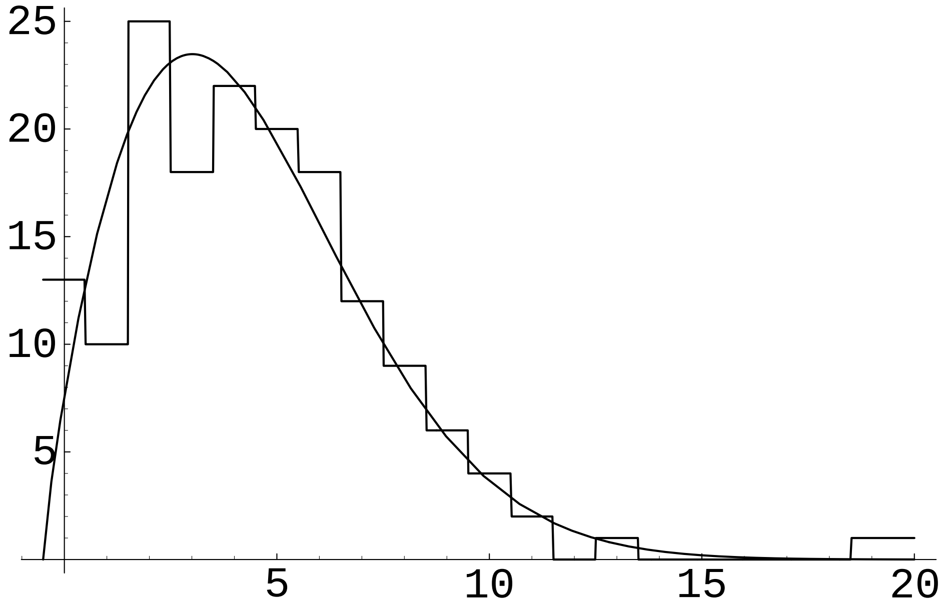



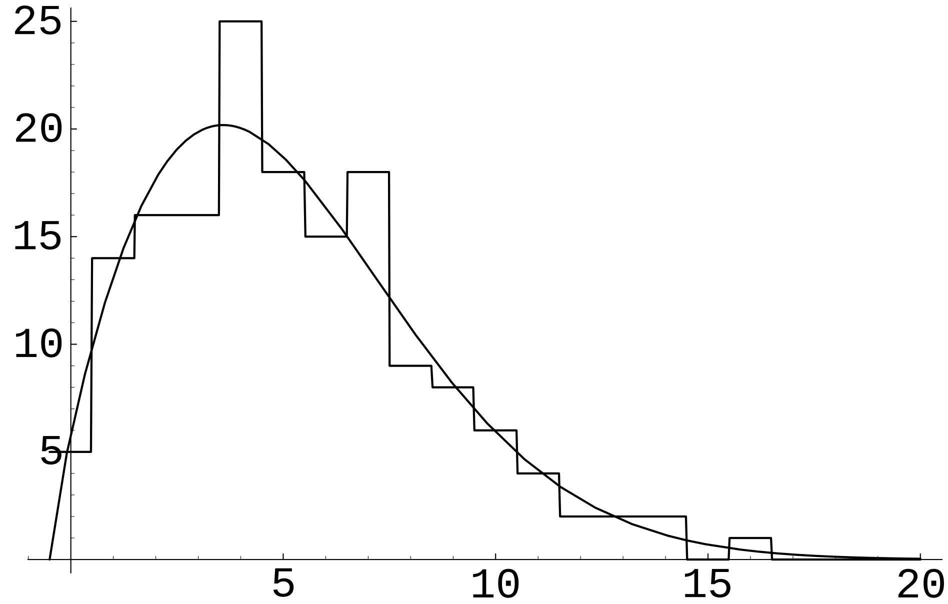
**Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Baltimore Orioles**



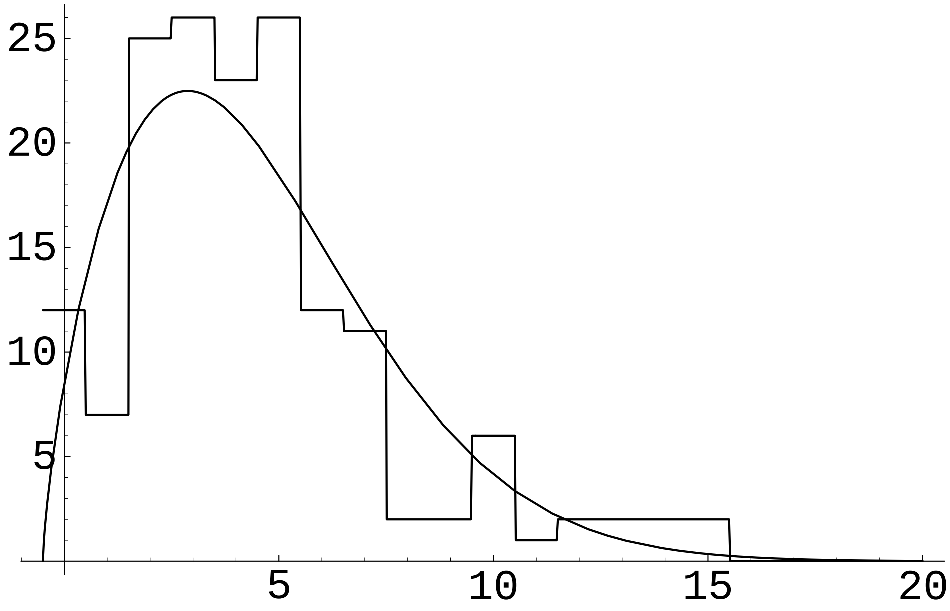



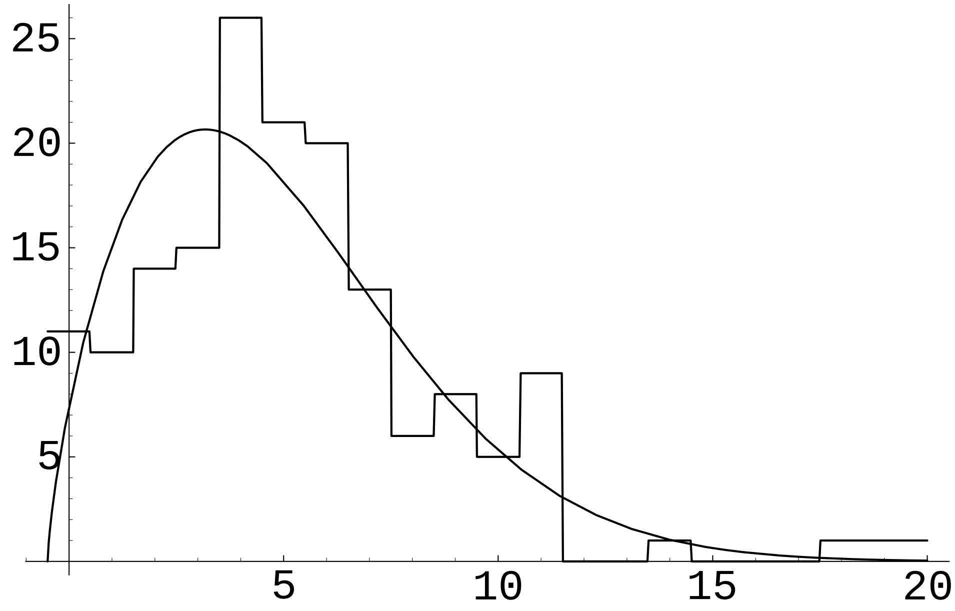
**Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Tampa Bay Devil Rays**



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



| Team | RS+RA $\chi 2$: 20 d.f. | Indep $\chi 2$: 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 |

$20$ 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

| Team | Obs Wins | Pred Wins | ObsPerc | PredPerc | GamesDiff | $\gamma$ |
|---|---|---|---|---|---|---|
| Boston Red Sox | 98 | 93.0 | 0.605 | 0.574 | 5.03 | 1.8 |
| New York Yankees | 101 | 87.5 | 0.623 | 0.540 | 13.49 | 1.7 |
| Baltimore Orioles | 78 | 83.1 | 0.481 | 0.513 | -5.08 | 1.6 |
| Tampa Bay Devil Rays | 70 | 69.6 | 0.435 | 0.432 | 0.38 | 1.8 |
| Toronto Blue Jays | 67 | 74.6 | 0.416 | 0.464 | -7.65 | 1.9 |
| Minnesota Twins | 92 | 84.7 | 0.568 | 0.523 | 7.31 | 1.7 |
| Chicago White Sox | 83 | 85.3 | 0.512 | 0.527 | -2.33 | 1.7 |
| Cleveland Indians | 80 | 80.0 | 0.494 | 0.494 | 0. | 1.7 |
| Detroit Tigers | 72 | 80.0 | 0.444 | 0.494 | -8.02 | 1.7 |
| Kansas City Royals | 58 | 68.7 | 0.358 | 0.424 | -10.65 | 1.7 |
| Los Angeles Angels | 92 | 87.5 | 0.568 | 0.540 | 4.53 | 1.7 |
| Oakland Athletics | 91 | 84.0 | 0.562 | 0.519 | 6.99 | 1.7 |
| Texas Rangers | 89 | 87.3 | 0.549 | 0.539 | 1.71 | 1.9 |
| Seattle Mariners | 63 | 70.7 | 0.389 | 0.436 | -7.66 | 1.7 |

$\gamma$: 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?