

Pythagoras at the Bat: An Introduction to Stats and Modeling

Cameron, Kayla and Steven Miller
(sjm1@williams.edu, Williams College)

<http://www.williams.edu/Mathematics/sjmiller>



Acknowledgments



Sal Baxamusa, Phil Birnbaum, Chris Chiang, Ray Ciccolella, Steve Johnston, Michelle Manes, Russ Mann, students of Math 162 and Math 197 at Brown, Math 105 and 399 at Williams.



Dedicated to my great uncle Newt Bromberg (a lifetime Red Sox fan who promised me that I would live to see a World Series Championship in Boston).



Chris Long and the San Diego Padres.

Introduction to the Pythagorean Won–Loss Theorem



Goals of the Talk

- Give derivation Pythagorean Won–Loss formula.
- Observe ideas / techniques of modeling.
- See how advanced theory enters in simple problems.
- Opportunities from inefficiencies.
- Xtra: further avenues for research for students.

Goals of the Talk

- Give derivation Pythagorean Won–Loss formula.
- Observe ideas / techniques of modeling.
- See how advanced theory enters in simple problems.
- Opportunities from inefficiencies.
- Xtra: further avenues for research for students.

GO SOX!

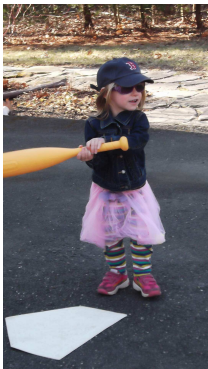
Baseball Review

Goal is to go from



Baseball Review

to



Baseball Review

to



Baseball Review



Baseball Review



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;
- γ : some parameter, constant for a sport.



86 Years & Worth the Wait

October 24, 2004

© 2004 Major League Baseball. All rights reserved.

MLB.com

© 2004 Major League Baseball. All rights reserved.

© 2004 Major League Baseball. All rights reserved.

© 2004 Major League Baseball. All rights reserved.

© 2004 Major League Baseball. All rights reserved.

© 2004 Major League Baseball. All rights reserved.

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;
- γ : some parameter, constant for a sport.

James' Won–Loss Formula (NUMERICAL Observation)

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

γ originally taken as 2, numerical studies show best γ for baseball is about 1.82.

Pythagorean Won–Loss Formula: Example

James' Won–Loss Formula

$$\text{Won – Loss Percentage} = \frac{\# \text{Wins}}{\# \text{Games}} = \frac{RS_{\text{obs}}^{\gamma}}{RS_{\text{obs}}^{\gamma} + RA_{\text{obs}}^{\gamma}}$$

Example ($\gamma = 1.82$): In 2009 the Red Sox were **95–67**. They scored 872 runs and allowed 736, for a Pythagorean prediction record of **93.4** wins and **68.6** losses; the Yankees were 103–59 but predicted to be **95.2–66.8** (they scored 915 runs and allowed 753).

Pythagorean Won–Loss Formula: Example

James' Won–Loss Formula

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

Example ($\gamma = 1.82$): In 2009 the Red Sox were **95–67**. They scored 872 runs and allowed 736, for a Pythagorean prediction record of **93.4** wins and **68.6** losses; the Yankees were 103–59 but predicted to be **95.2–66.8** (they scored 915 runs and allowed 753).

2011: Red Sox 'should' be 95-67, Tampa 'should' be 92-70....

Applications of the Pythagorean Won–Loss Formula

- **Extrapolation:** use half-way through season to predict a team's performance for rest of season.
- **Evaluation:** see if consistently over-perform or under-perform.
- **Advantage:** Other statistics / formulas (run-differential per game); this is easy to use, depends only on two simple numbers for a team.

Applications of the Pythagorean Won–Loss Formula

- **Extrapolation:** use half-way through season to predict a team's performance for rest of season.
- **Evaluation:** see if consistently over-perform or under-perform.
- **Advantage:** Other statistics / formulas (run-differential per game); this is easy to use, depends only on two simple numbers for a team.

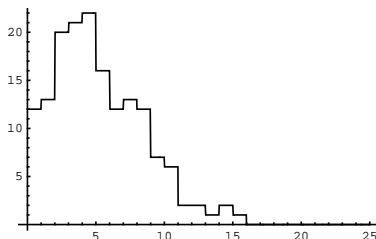
Red Sox: 2004 Predictions: May 1: 99 wins; June 1: 93 wins; July 1: 90 wins; August 1: 92 wins.
Finished season with 98 wins.

Probability and Modeling

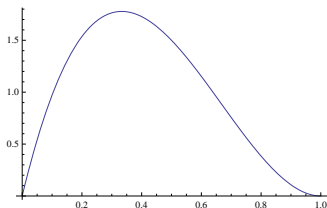


Observed scoring distributions

Goal is to model observed scoring distributions; for example, consider

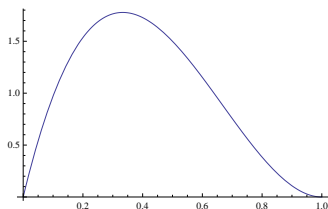


Probability Review



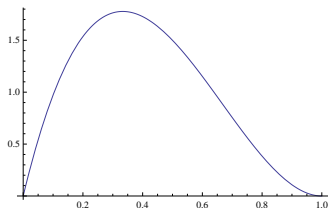
- Let X be random variable with density $p(x)$:
 - ◇ $p(x) \geq 0$;
 - ◇ $\int_{-\infty}^{\infty} p(x)dx = 1$;
 - ◇ $\text{Prob}(a \leq X \leq b) = \int_a^b p(x)dx$.

Probability Review



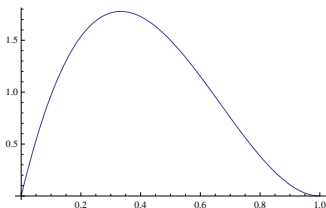
- Let X be random variable with density $p(x)$:
 - ◇ $p(x) \geq 0$;
 - ◇ $\int_{-\infty}^{\infty} p(x)dx = 1$;
 - ◇ $\text{Prob}(a \leq X \leq b) = \int_a^b p(x)dx$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)dx$.

Probability Review



- Let X be random variable with density $p(x)$:
 - ◇ $p(x) \geq 0$;
 - ◇ $\int_{-\infty}^{\infty} p(x)dx = 1$;
 - ◇ $\text{Prob}(a \leq X \leq b) = \int_a^b p(x)dx$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)dx$.
- Variance $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx$.

Probability Review



- Let X be random variable with density $p(x)$:
 - ◇ $p(x) \geq 0$;
 - ◇ $\int_{-\infty}^{\infty} p(x)dx = 1$;
 - ◇ $\text{Prob}(a \leq X \leq b) = \int_a^b p(x)dx$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)dx$.
- Variance $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx$.
- Independence: knowledge of one random variable gives no knowledge of the other.

Modeling the Real World

Guidelines for Modeling:

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



Modeling the Real World (cont)

Possible Model:

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

Modeling the Real World (cont)

Possible Model:

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

Won–Loss formula follows from computing

$$\int_{x=0}^{\infty} \left[\int_{y \leq x} f_{RS}(x) g_{RA}(y) dy \right] dx \quad \text{or} \quad \sum_{i=0}^{\infty} \left[\sum_{j < i} f_{RS}(i) g_{RA}(j) \right].$$

Problems with the Model

Reduced to calculating

$$\int_{x=0}^{\infty} \left[\int_{y \leq x} f_{RS}(x) g_{RA}(y) dy \right] dx \quad \text{or} \quad \sum_{i=0}^{\infty} \left[\sum_{j < i} f_{RS}(i) g_{RA}(j) \right].$$

Problems with the Model

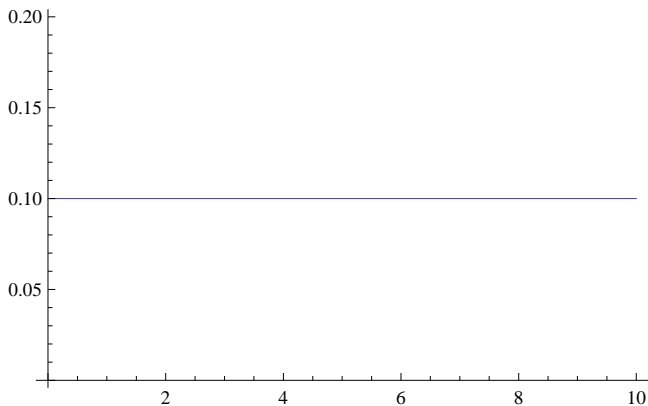
Reduced to calculating

$$\int_{x=0}^{\infty} \left[\int_{y \leq x} f_{RS}(x) g_{RA}(y) dy \right] dx \quad \text{or} \quad \sum_{i=0}^{\infty} \left[\sum_{j < i} f_{RS}(i) g_{RA}(j) \right].$$

Problems with the model:

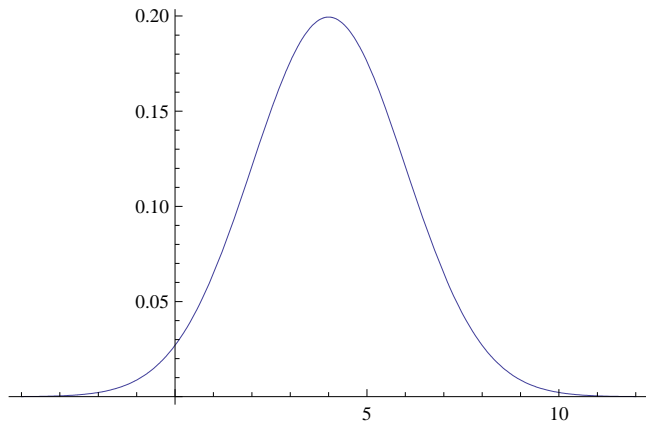
- What are explicit formulas for f_{RS} and g_{RA} ?
- Are the runs scored and allowed independent random variables?
- Can the integral (or sum) be computed in closed form?

Choices for f_{RS} and g_{RA}



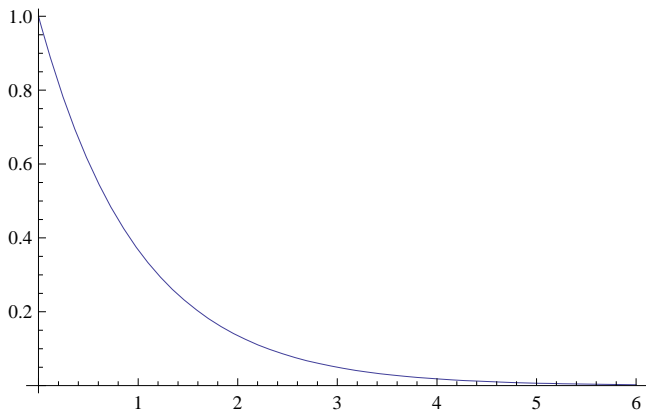
Uniform Distribution on $[0, 10]$.

Choices for f_{RS} and g_{RA}



Normal Distribution: mean 4, standard deviation 2.

Choices for f_{RS} and g_{RA}



Exponential Distribution: e^{-x} .

Three Parameter Weibull

Weibull distribution:

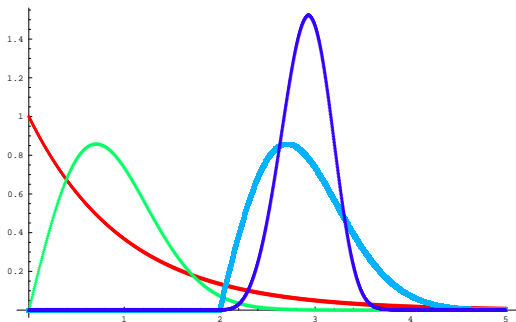
$$f(\mathbf{x}; \alpha, \beta, \gamma) = \begin{cases} \frac{\gamma}{\alpha} \left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^\gamma} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$

- α : scale (variance: meters versus centimeters);
- β : origin (mean: translation, zero point);
- γ : shape (behavior near β and at infinity).

Various values give different shapes, but can we find α, β, γ such that it fits observed data? Is the Weibull justifiable by some reasonable hypotheses?

Weibull Plots: Parameters (α, β, γ) :

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



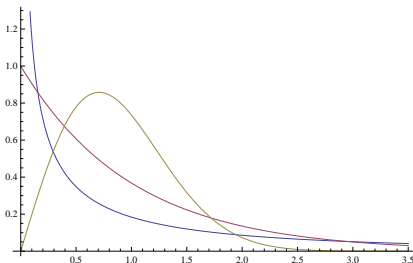
Red:(1, 0, 1) (exponential); Green:(1, 0, 2); Cyan:(1, 2, 2);
Blue:(1, 2, 4)

Three Parameter Weibull: Applications

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

Arises in many places, such as survival analysis.

- $\gamma < 1$: high infant mortality;
- $\gamma = 1$: constant failure rate;
- $\gamma > 1$: aging process.



The Gamma Distribution and Weibulls

- For $s > 0$, define the Γ -function by

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

- Generalizes factorial function: $\Gamma(n) = (n-1)!$ for $n \geq 1$ an integer.

A Weibull distribution with parameters α, β, γ has:

- Mean: $\alpha \Gamma(1 + 1/\gamma) + \beta$.
- Variance: $\alpha^2 \Gamma(1 + 2/\gamma) - \alpha^2 \Gamma(1 + 1/\gamma)^2$.

Weibull Integrations

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

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

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

A similar calculation determines the variance.

The Pythagorean Theorem

American League



| Select favorite team ▼

| Standings as of Jun ▼ | 5 ▼ | 2008 ▼

| GO

| East | W | L | PCT | GB | L10 | STRK | INT | HOME | ROAD | X W-L | LAST GAME | NEXT GAME |
|-------------|----|----|------|------|-----|------|-----|-------|-------|-------|-------------------|-------------------|
| Boston | 37 | 25 | .597 | - | 6-4 | W2 | 3-0 | 23-5 | 14-20 | 36-26 | 6/4 v TB, W 5-1 | 6/5 v TB, 6:05P |
| Tampa Bay | 35 | 24 | .593 | 0.5 | 6-4 | L2 | 1-2 | 24-10 | 11-14 | 32-27 | 6/4 @ BOS, L 1-5 | 6/5 @ BOS, 8:05P |
| Toronto | 32 | 29 | .525 | 4.5 | 6-4 | L1 | 2-1 | 15-11 | 17-18 | 34-27 | 6/4 @ NYY, L 1-5 | 6/5 @ NYY, 1:05P |
| New York | 29 | 30 | .492 | 6.5 | 5-5 | W1 | 0-2 | 15-13 | 14-17 | 28-31 | 6/4 v TOR, W 5-1 | 6/5 v TOR, 1:05P |
| Baltimore | 28 | 30 | .483 | 7.0 | 4-6 | L1 | 2-1 | 17-11 | 11-19 | 27-31 | 6/4 @ MIN, L 5-7 | 6/5 @ MIN, 1:10P |
| Central | W | L | PCT | GB | L10 | STRK | INT | HOME | ROAD | X W-L | LAST GAME | NEXT GAME |
| Chicago | 32 | 26 | .552 | - | 6-4 | W2 | 3-0 | 15-9 | 17-17 | 34-24 | 6/4 v KC, W 6-4 | 6/5 v KC, 8:11P |
| Minnesota | 31 | 28 | .525 | 1.5 | 7-3 | W1 | 1-2 | 19-15 | 12-13 | 29-30 | 6/4 v BAL, W 7-5 | 6/5 v BAL, 1:10P |
| Cleveland | 27 | 32 | .458 | 5.5 | 4-6 | W1 | 0-3 | 16-16 | 11-16 | 31-28 | 6/4 @ TEX, W 15-9 | 6/5 @ TEX, 8:05P |
| Detroit | 24 | 35 | .407 | 8.5 | 3-7 | L3 | 1-2 | 12-14 | 12-21 | 27-32 | 6/4 @ OAK, L 2-10 | 6/6 v CLE, 7:05P |
| Kansas City | 23 | 36 | .390 | 9.5 | 2-8 | L2 | 2-1 | 12-16 | 11-20 | 23-36 | 6/4 @ CWS, L 4-6 | 6/5 @ CWS, 8:11P |
| West | W | L | PCT | GB | L10 | STRK | INT | HOME | ROAD | X W-L | LAST GAME | NEXT GAME |
| Los Angeles | 37 | 24 | .607 | - | 7-3 | W5 | 2-1 | 18-13 | 19-11 | 31-30 | 6/4 @ SEA, W 5-4 | 6/6 @ OAK, 10:05P |
| Oakland | 33 | 27 | .550 | 3.5 | 6-4 | W4 | 1-2 | 20-13 | 13-14 | 35-25 | 6/4 v DET, W 10-2 | 6/6 v LAA, 10:05P |
| Texas | 30 | 31 | .492 | 7.0 | 5-5 | L1 | 2-1 | 15-14 | 15-17 | 29-32 | 6/4 v CLE, L 9-15 | 6/5 v CLE, 8:05P |
| Seattle | 21 | 39 | .350 | 15.5 | 3-7 | L4 | 2-1 | 14-19 | 7-20 | 24-36 | 6/4 v LAA, L 4-5 | 6/6 @ BOS, 7:05P |

National League



| East | W | L | PCT | GB | L10 | STRK | INT | HOME | ROAD | X W-L | LAST GAME | NEXT GAME |
|--------------|----|----|------|------|-----|------|-----|-------|-------|-------|------------------|------------------|
| Philadelphia | 35 | 26 | .574 | - | 8-2 | L1 | 1-2 | 20-13 | 15-13 | 36-25 | 6/4 v CIN, L 0-2 | 6/5 v CIN, 1:05P |
| Florida | 32 | 26 | .552 | 1.5 | 4-6 | W1 | 1-2 | 18-12 | 14-14 | 29-29 | 6/4 @ ATL, W 6-4 | 6/5 @ ATL, 7:00P |
| New York | 30 | 28 | .517 | 3.5 | 7-3 | W2 | 2-0 | 17-11 | 13-17 | 30-28 | 6/4 @ SF, W 5-3 | 6/5 @ SD, 10:05P |
| Atlanta | 31 | 29 | .517 | 3.5 | 4-6 | L1 | 2-1 | 24-8 | 7-21 | 35-25 | 6/4 v FLA, L 4-6 | 6/5 v FLA, 7:00P |
| Washington | 24 | 35 | .407 | 10.0 | 3-7 | L3 | 1-2 | 13-16 | 11-19 | 23-36 | 6/4 v STL, PPD | 6/5 v STL, 7:10P |
| Central | W | L | PCT | GB | L10 | STRK | INT | HOME | ROAD | X W-L | LAST GAME | NEXT GAME |

Building Intuition: The log 5 Method

Assume team A wins p percent of their games, and team B wins q percent of their games. Which formula do you think does a good job of predicting the probability that team A beats team B ?

$$\frac{p + pq}{p + q + 2pq},$$

$$\frac{p + pq}{p + q - 2pq}$$

$$\frac{p - pq}{p + q + 2pq},$$

$$\frac{p - pq}{p + q - 2pq}$$

Building intuition: A wins p percent, B wins q percent

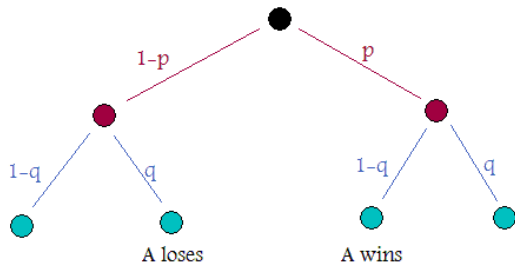
$$\frac{p + pq}{p + q + 2pq}, \quad \frac{p + pq}{p + q - 2pq}$$

$$\frac{p - pq}{p + q + 2pq}, \quad \frac{p - pq}{p + q - 2pq}$$

Consider special cases:

- 1 Prob(A beats B) + Prob(B beats A) = 1.
- 2 If $p = q$ then the probability A beats B is 50%.
- 3 If $p = 1$ and $q \neq 0, 1$ then A always beats B .
- 4 If $p = 0$ and $q \neq 0, 1$ then A always loses to B .
- 5 If $p > 1/2$ and $q < 1/2$ then Prob(A beats B) $> p$.
- 6 If $q = 1/2$ prob A wins is p ($p = 1/2$ the prob B wins is q).

Building intuition: Sketch of proof: $\frac{p-pq}{p+q-2pq}$



- A beats B has probability $p(1 - q)$.
- A and B do not have the same outcome has probability $p(1 - q) + (1 - p)q$.
- $\text{Prob}(A \text{ beats } B) = \frac{p(1-q)}{p(1-q)+(1-p)q} = \frac{p-pq}{p+q-2pq}$.

Pythagorean Won–Loss Formula: $\frac{RS_{obs}^\gamma}{RS_{obs}^\gamma + RA_{obs}^\gamma}$

Theorem: Pythagorean Won–Loss Formula (Miller '06)

Let the runs scored and allowed per game be two independent random variables drawn from Weibull distributions $(\alpha_{RS}, \beta, \gamma)$ and $(\alpha_{RA}, \beta, \gamma)$; α_{RS} and α_{RA} are chosen so that the Weibull means are the observed sample values RS and RA . If $\gamma > 0$ then the Won–Loss Percentage is $\frac{(RS-\beta)^\gamma}{(RS-\beta)^\gamma + (RA-\beta)^\gamma}$.

Pythagorean Won–Loss Formula: $\frac{RS_{obs}^\gamma}{RS_{obs}^\gamma + RA_{obs}^\gamma}$

Theorem: Pythagorean Won–Loss Formula (Miller '06)

Let the runs scored and allowed per game be two independent random variables drawn from Weibull distributions $(\alpha_{RS}, \beta, \gamma)$ and $(\alpha_{RA}, \beta, \gamma)$; α_{RS} and α_{RA} are chosen so that the Weibull means are the observed sample values RS and RA . If $\gamma > 0$ then the Won–Loss Percentage is $\frac{(RS-\beta)^\gamma}{(RS-\beta)^\gamma + (RA-\beta)^\gamma}$.

Take $\beta = -1/2$ (since runs must be integers).

$RS - \beta$ estimates average runs scored, $RA - \beta$ estimates average runs allowed.

Weibull with parameters (α, β, γ) has mean $\alpha\Gamma(1 + 1/\gamma) + \beta$.

Proof of the Pythagorean Won–Loss Formula

Let X and Y be independent random variables with Weibull distributions $(\alpha_{RS}, \beta, \gamma)$ and $(\alpha_{RA}, \beta, \gamma)$ respectively. To have means of $RS - \beta$ and $RA - \beta$ our calculations for the means imply

$$\alpha_{RS} = \frac{RS - \beta}{\Gamma(1 + 1/\gamma)}, \quad \alpha_{RA} = \frac{RA - \beta}{\Gamma(1 + 1/\gamma)}.$$

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

Proof of the Pythagorean Won–Loss Formula (cont)

$$\text{Prob}(X > Y) = \int_{x=\beta}^{\infty} \int_{y=\beta}^x f(\mathbf{x}; \alpha_{RS}, \beta, \gamma) f(\mathbf{y}; \alpha_{RA}, \beta, \gamma) dy dx$$

Proof of the Pythagorean Won–Loss Formula (cont)

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

where we have set

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

Proof of the Pythagorean Won–Loss Formula (cont)

$$\text{Prob}(X > Y) = 1 - \frac{\alpha^\gamma}{\alpha_{RS}^\gamma} \int_0^\infty \frac{\gamma}{\alpha} \left(\frac{x}{\alpha}\right)^{\gamma-1} e^{(x/\alpha)^\gamma} dx$$

Proof of the Pythagorean Won–Loss Formula (cont)

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

We substitute the relations for α_{RS} and α_{RA} and find that

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

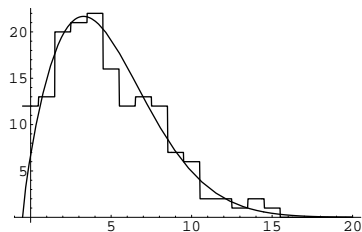
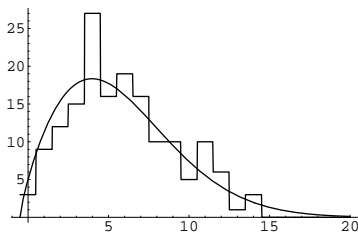
Note $\text{RS} - \beta$ estimates RS_{obs} , $\text{RA} - \beta$ estimates RA_{obs} .

Analysis of 2004



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

Best Fit Weibulls to Data: Method of Least Squares

- $\text{Bin}(k)$ is the k^{th} bin;
- $\text{RS}_{\text{obs}}(k)$ (resp. $\text{RA}_{\text{obs}}(k)$) the observed number of games with the number of runs scored (allowed) in $\text{Bin}(k)$;
- $A(\alpha, \gamma, k)$ the area under the Weibull with parameters $(\alpha, -1/2, \gamma)$ in $\text{Bin}(k)$.

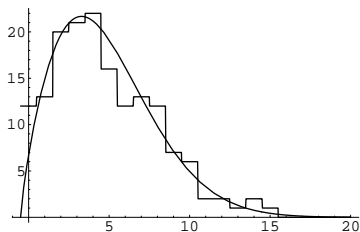
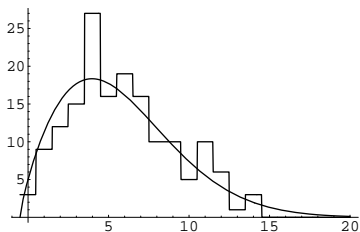
Find the values of $(\alpha_{\text{RS}}, \alpha_{\text{RA}}, \gamma)$ that minimize

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

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

Best Fit Weibulls to Data (Method of Maximum Likelihood)

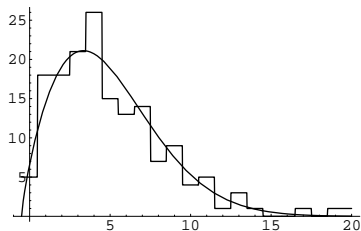
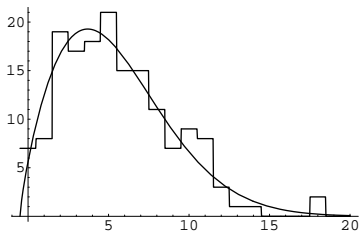
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Boston Red Sox



Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \dots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

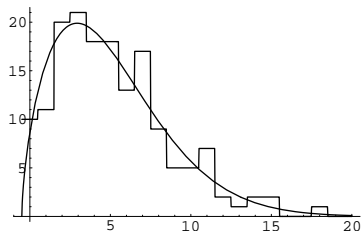
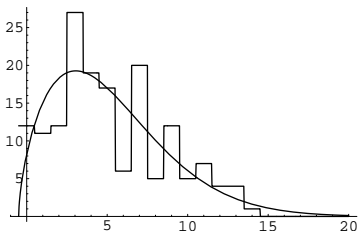
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the New York Yankees



Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \dots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

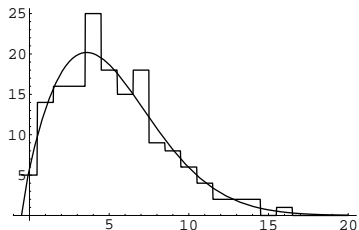
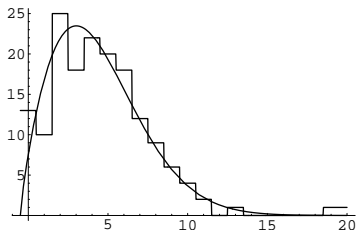
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Baltimore Orioles



Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \dots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

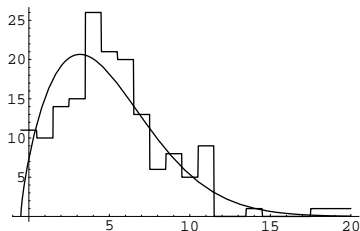
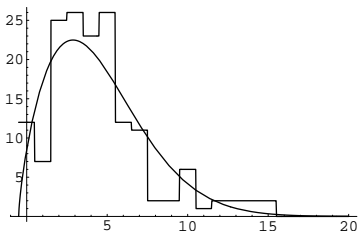
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Tampa Bay Devil Rays



Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \dots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

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



Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \dots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Advanced Theory

Bonferroni Adjustments

Fair coin: 1,000,000 flips, expect 500,000 heads.

Bonferroni Adjustments

Fair coin: 1,000,000 flips, expect 500,000 heads.
About 95% have $499,000 \leq \#Heads \leq 501,000$.

Bonferroni Adjustments

Fair coin: 1,000,000 flips, expect 500,000 heads.
About 95% have $499,000 \leq \#Heads \leq 501,000$.

Consider N independent experiments of flipping a fair coin 1,000,000 times. *What is the probability that at least one of set doesn't have $499,000 \leq \#Heads \leq 501,000$?*

| N | Probability |
|----------|--------------------|
| 5 | 22.62 |
| 14 | 51.23 |
| 50 | 92.31 |

See unlikely events happen as N increases!

Data Analysis: χ^2 Tests (20 and 109 degrees of freedom)

| 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 |

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).

Data Analysis: Structural Zeros

- For independence of runs scored and allowed, use bins $[0, 1) \cup [1, 2) \cup [2, 3) \cup \dots \cup [8, 9) \cup [9, 10) \cup [10, 11) \cup [11, \infty)$.
- Have an $r \times c$ contingency table with **structural zeros** (runs scored and allowed per game are never equal).
- (Essentially) $O_{r,r} = 0$ for all r , use an iterative fitting procedure to obtain maximum likelihood estimators for $E_{r,c}$ (expected frequency of cell (r, c) assuming that, given runs scored and allowed are distinct, the runs scored and allowed are independent).

Summary

Testing the Model: Data from 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 |

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

Conclusions

- Find parameters such that Weibulls are good fits;
- Runs scored and allowed per game are statistically independent;
- Pythagorean Won–Loss Formula is a consequence of our model;
- Best γ (both close to observed best 1.82):
 - ◇ Method of Least Squares: 1.79;
 - ◇ Method of Maximum Likelihood: 1.74.

Future Work

- **Micro-analysis:** runs scored and allowed aren't independent (big lead, close game), run production smaller for inter-league games in NL parks,
- **Other sports:** Does the same model work? Basketball has γ between 14 and 16.5.
- **Closed forms:** Are there other probability distributions that give integrals which can be determined in closed form?
- **Valuing Runs:** Pythagorean formula used to value players (10 runs equals 1 win); better model leads to better team.

References

References

● Baxamusa, Sal:

◇ Weibull worksheet: <http://www.beyondtheboxscore.com/story/2006/4/30/114737/251>

◇ Run distribution plots for various teams:

<http://www.beyondtheboxscore.com/story/2006/2/23/164417/484>

● Miller, Steven J.:

◇ *A Derivation of James' Pythagorean projection*, By The Numbers – The Newsletter of the SABR Statistical Analysis Committee, vol. 16 (February 2006), no. 1, 17–22.

<http://www.philbirnbaum.com/btn2006-02.pdf>

◇ *A derivation of the Pythagorean Won–Loss Formula in baseball*, Chance Magazine **20** (2007), no. 1, 40–48. http://www.williams.edu/go/math/sjmillier/public_html/math/talks/talk.html

Appendices

Appendix I: Proof of the Pythagorean Won–Loss Formula

Let X and Y be independent random variables with Weibull distributions $(\alpha_{RS}, \beta, \gamma)$ and $(\alpha_{RA}, \beta, \gamma)$ respectively. To have means of $RS - \beta$ and $RA - \beta$ our calculations for the means imply

$$\alpha_{RS} = \frac{RS - \beta}{\Gamma(1 + 1/\gamma)}, \quad \alpha_{RA} = \frac{RA - \beta}{\Gamma(1 + 1/\gamma)}.$$

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

Appendix I: Proof of the Pythagorean Won-Loss Formula (cont)

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

where we have set

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

Appendix I: Proof of the Pythagorean Won–Loss Formula (cont)

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

We substitute the relations for α_{RS} and α_{RA} and find that

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

Note $\text{RS} - \beta$ estimates RS_{obs} , $\text{RA} - \beta$ estimates RA_{obs} .

Appendix II: Best Fit Weibulls and Structural Zeros

The fits *look* good, but are they? Do χ^2 -tests:

- Let $\text{Bin}(k)$ denote the k^{th} bin.
- $O_{r,c}$: the observed number of games where the team's runs scored is in $\text{Bin}(r)$ and the runs allowed are in $\text{Bin}(c)$.
- $E_{r,c} = \frac{\sum_{c'} O_{r,c'} \cdot \sum_{r'} O_{r',c}}{\# \text{Games}}$ is the expected frequency of cell (r, c) .
- Then

$$\sum_{r=1}^{\# \text{Rows}} \sum_{c=1}^{\# \text{Columns}} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}$$

is a χ^2 distribution with $(\# \text{Rows} - 1)(\# \text{Columns} - 1)$ degrees of freedom.

Appendix II: Best Fit Weibulls and Structural Zeros (cont)

For independence of runs scored and allowed, use bins

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

Have an $r \times c$ contingency table (with $r = c = 12$); however, there are *structural zeros* (runs scored and allowed per game can never be equal).

(Essentially) $O_{r,r} = 0$ for all r . We use the iterative fitting procedure 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

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

Then

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

and

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

the iterations converge very quickly. (If we had a complete two-dimensional contingency table, then the iteration reduces to the standard values, namely $E_{r,c} = \sum_{c'} O_{r,c'} \cdot \sum_{r'} O_{r',c} / \# \text{Games}$.) Note

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

Appendix III: Central Limit Theorem

Convolution of f and g :

$$h(y) = (f * g)(y) = \int_{\mathbb{R}} f(x)g(y-x)dx = \int_{\mathbb{R}} f(x-y)g(x)dx.$$

X_1 and X_2 independent random variables with probability density p .

$$\text{Prob}(X_i \in [x, x + \Delta x]) = \int_x^{x+\Delta x} p(t)dt \approx p(x)\Delta x.$$

$$\text{Prob}(X_1 + X_2 \in [x, x + \Delta x]) = \int_{x_1=-\infty}^{\infty} \int_{x_2=x-x_1}^{x+\Delta x-x_1} p(x_1)p(x_2)dx_2dx_1.$$

As $\Delta x \rightarrow 0$ we obtain the convolution of p with itself:

$$\text{Prob}(X_1 + X_2 \in [a, b]) = \int_a^b (p * p)(z)dz.$$

Exercise to show non-negative and integrates to 1.

Appendix III: Statement of Central Limit Theorem

- For simplicity, assume p has mean zero, variance one, finite third moment and is of sufficiently rapid decay so that all convolution integrals that arise converge: p an infinitely differentiable function satisfying

$$\int_{-\infty}^{\infty} xp(x)dx = 0, \quad \int_{-\infty}^{\infty} x^2 p(x)dx = 1, \quad \int_{-\infty}^{\infty} |x|^3 p(x)dx < \infty.$$

- Assume X_1, X_2, \dots are independent identically distributed random variables drawn from p .
- Define $S_N = \sum_{i=1}^N X_i$.
- Standard Gaussian (mean zero, variance one) is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Central Limit Theorem Let X_i, S_N be as above and assume the third moment of each X_i is finite. Then S_N/\sqrt{N} converges in probability to the standard Gaussian:

$$\lim_{N \rightarrow \infty} \text{Prob} \left(\frac{S_N}{\sqrt{N}} \in [a, b] \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

Appendix III: Proof of the Central Limit Theorem

- The Fourier transform of p is

$$\hat{p}(y) = \int_{-\infty}^{\infty} p(x)e^{-2\pi ixy} dx.$$

- Derivative of \hat{g} is the Fourier transform of $2\pi ixg(x)$; differentiation (hard) is converted to multiplication (easy).

$$\hat{g}'(y) = \int_{-\infty}^{\infty} 2\pi ix \cdot g(x)e^{-2\pi ixy} dx.$$

If g is a probability density, $\hat{g}'(0) = 2\pi i\mathbb{E}[x]$ and $\hat{g}''(0) = -4\pi^2\mathbb{E}[x^2]$.

- Natural to use the Fourier transform to analyze probability distributions. The mean and variance are simple multiples of the derivatives of \hat{p} at zero: $\hat{p}'(0) = 0$, $\hat{p}''(0) = -4\pi^2$.
- We Taylor expand \hat{p} (need technical conditions on p):

$$\hat{p}(y) = 1 + \frac{p''(0)}{2}y^2 + \dots = 1 - 2\pi^2y^2 + O(y^3).$$

Near the origin, the above shows \hat{p} looks like a concave down parabola.

Appendix III: Proof of the Central Limit Theorem (cont)

- $\text{Prob}(X_1 + \dots + X_N \in [a, b]) = \int_a^b (p * \dots * p)(z) dz.$
- The Fourier transform converts convolution to multiplication. If $\text{FT}[f](y)$ denotes the Fourier transform of f evaluated at y :

$$\text{FT}[p * \dots * p](y) = \widehat{p}(y) \cdot \dots \cdot \widehat{p}(y).$$

- Do not want the distribution of $X_1 + \dots + X_N = x$, but rather

$$S_N = \frac{X_1 + \dots + X_N}{\sqrt{N}} = x.$$

- If $B(x) = A(cx)$ for some fixed $c \neq 0$, then $\widehat{B}(y) = \frac{1}{c} \widehat{A}\left(\frac{y}{c}\right).$

- $\text{Prob}\left(\frac{X_1 + \dots + X_N}{\sqrt{N}} = x\right) = (\sqrt{N}p * \dots * \sqrt{N}p)(x\sqrt{N}).$

- $\text{FT}\left[(\sqrt{N}p * \dots * \sqrt{N}p)(x\sqrt{N})\right](y) = \left[\widehat{p}\left(\frac{y}{\sqrt{N}}\right)\right]^N.$

Appendix III: Proof of the Central Limit Theorem (cont)

- Can find the Fourier transform of the distribution of S_N :

$$\left[\widehat{p} \left(\frac{y}{\sqrt{N}} \right) \right]^N.$$

- Take the limit as $N \rightarrow \infty$ for **fixed** y .
- Know $\widehat{p}(y) = 1 - 2\pi^2 y^2 + O(y^3)$. Thus study

$$\left[1 - \frac{2\pi^2 y^2}{N} + O \left(\frac{y^3}{N^{3/2}} \right) \right]^N.$$

- For any **fixed** y ,

$$\lim_{N \rightarrow \infty} \left[1 - \frac{2\pi^2 y^2}{N} + O \left(\frac{y^3}{N^{3/2}} \right) \right]^N = e^{-2\pi y^2}.$$

- Fourier transform of $e^{-2\pi x^2}$ at x is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Appendix III: Proof of the Central Limit Theorem (cont)

We have shown:

- the Fourier transform of the distribution of S_N converges to $e^{-2\pi y^2}$;
- the Fourier transform of $e^{-2\pi y^2}$ is $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Therefore the distribution of S_N equalling x converges to $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

We need complex analysis to justify this conclusion. Must be careful: Consider

$$g(x) = \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

All the Taylor coefficients about $x = 0$ are zero, but the function is not identically zero in a neighborhood of $x = 0$.

Appendix IV: Best Fit Weibulls from Method of Maximum Likelihood

The likelihood function depends on: $\alpha_{RS}, \alpha_{RA}, \beta = -.5, \gamma$.

Let $A(\alpha, -.5, \gamma, k)$ denote the area in $\text{Bin}(k)$ of the Weibull with parameters $\alpha, -.5, \gamma$. The sample likelihood function

$L(\alpha_{RS}, \alpha_{RA}, -.5, \gamma)$ is

$$\begin{aligned} & \binom{\# \text{Games}}{\text{RS}_{\text{obs}}(1), \dots, \text{RS}_{\text{obs}}(\# \text{Bins})} \prod_{k=1}^{\# \text{Bins}} A(\alpha_{RS}, -.5, \gamma, k)^{\text{RS}_{\text{obs}}(k)} \\ & \cdot \binom{\# \text{Games}}{\text{RA}_{\text{obs}}(1), \dots, \text{RA}_{\text{obs}}(\# \text{Bins})} \prod_{k=1}^{\# \text{Bins}} A(\alpha_{RA}, -.5, \gamma, k)^{\text{RA}_{\text{obs}}(k)}. \end{aligned}$$

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.