# Pythagoras at the Bat:
# An Introduction to Stats and Modeling

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

http://web.williams.edu/Mathematics/sjmiller/public_html/

**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 150 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.

**Acknowledgments**

## 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**

- **G**ive derivation Pythagorean Won–Loss formula.

- **O**bserve ideas / techniques of modeling.

- **S**ee how advanced theory enters in simple problems.

- **O**pportunities from inefficiencies.

- **X**tra: further avenues for research for students.

GO SOX!

**Statistics**

Goal is to find good statistics to describe real world.

## Statistics

Goal is to find good statistics to describe real world.



**Figure:** Harvard Bridge, about 620.1 meters.

**Statistics**

Goal is to find good statistics to describe real world.



**Figure:** Harvard Bridge, 364.1 Smoots ($\pm$ one ear).

**Baseball Review**

Goal is to go from

## Baseball Review

**Baseball Review**

to

## Baseball Review

## Baseball Review

to

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



86 Years & Worth the Wait

October 24, 2004

**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.

### James' Won–Loss Formula (NUMERICAL Observation)

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

$\gamma$ originally taken as 2, numerical studies show best $\gamma$ for baseball is about 1.82.

**Pythagorean Won–Loss Formula: Example**

### James' Won–Loss Formula

$$\text{Won} - \text{Loss Percentage} \ = \ \frac{\#\text{Wins}}{\#\text{Games}} \ = \ \frac{\text{RS}_{\text{obs}}^{\gamma}}{\text{RS}_{\text{obs}}^{\gamma} + \text{RA}_{\text{obs}}^{\gamma}}$$

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

$$\text{Won} - \text{Loss Percentage} \ = \ \frac{\#\text{Wins}}{\#\text{Games}} \ = \ \frac{\text{RS}_{\text{obs}}^{\gamma}}{\text{RS}_{\text{obs}}^{\gamma} + \text{RA}_{\text{obs}}^{\gamma}}$$

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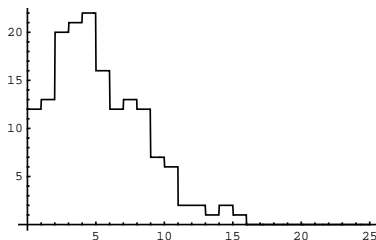
Red Sox: 2004 Predictions: May 1: 99 wins; June 1: 93 wins; July 1: 90 wins; August 1: 92 wins.
Finished season with 98 wins.
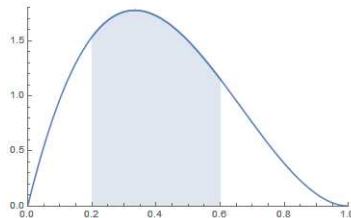
## Probability and Modeling

**Observed scoring distributions**

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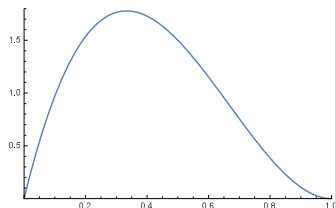
**Probability Review**



- Let $X$ be random variable with density $p(x)$:
  - $\diamond\ p(x) \geq 0$;
  - $\diamond\ \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  - $\diamond\ \mathrm{Prob}\,(a \leq X \leq b) = \int_{a}^{b} p(x)\mathrm{d}x$.

## Probability Review



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

## Probability Review
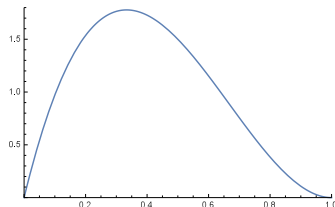


- Let $X$ be random variable with density $p(x)$:
  - $\diamond$ $p(x) \geq 0$;
  - $\diamond$ $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  - $\diamond$ $\mathrm{Prob}\,(a \leq X \leq b) = \int_{a}^{b} p(x)\mathrm{d}x$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$.
- Variance $\sigma^2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

## Probability Review
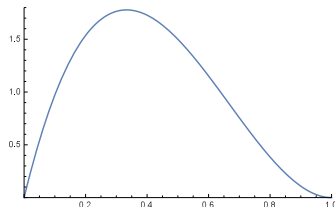


- Let $X$ be random variable with density $p(x)$:
  - $\diamond\ p(x) \geq 0$;
  - $\diamond\ \int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  - $\diamond\ \mathrm{Prob}\,(a \leq X \leq b) = \int_{a}^{b} p(x)\mathrm{d}x$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$.
- Variance $\sigma^2 =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

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

**Problems with the Model**

Reduced to calculating

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

**Problems with the Model**

Reduced to calculating

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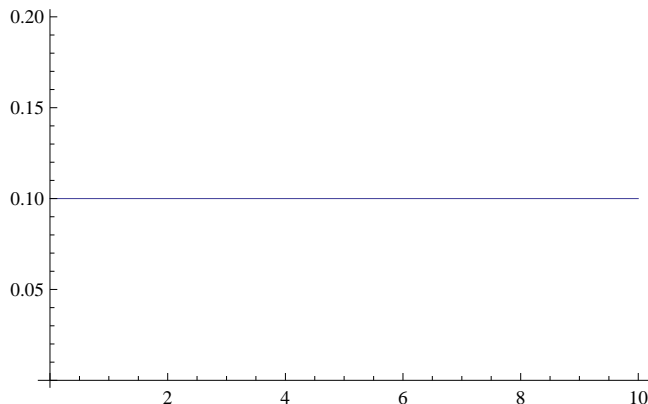$$

Problems with the model:

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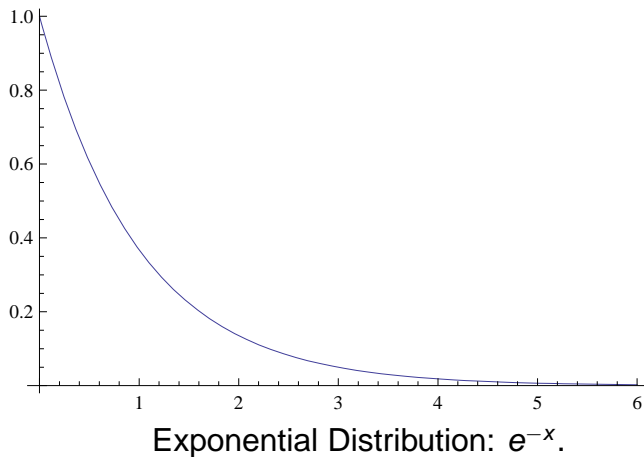
## Choices for $f_{RS}$ and $g_{RA}$



Uniform Distribution on $[0, 10]$.

## Choices for $f_{RS}$ and $g_{RA}$
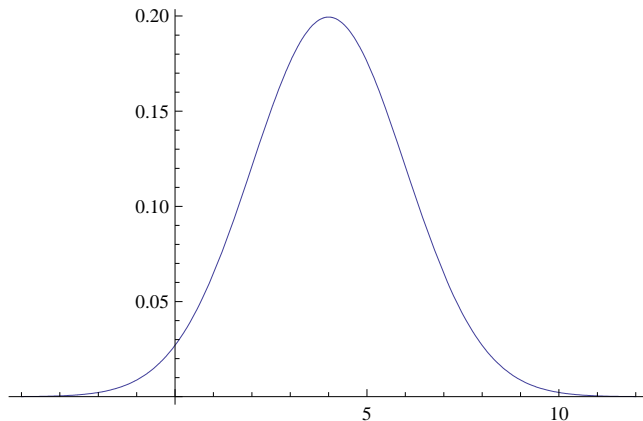


Normal Distribution: mean 4, standard deviation 2.

## Choices for $f_{RS}$ and $g_{RA}$



Exponential Distribution: $e^{-x}$.

**Three Parameter Weibull**

Weibull distribution:

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

- $\alpha$: scale (variance: meters versus centimeters);
- $\beta$: origin (mean: 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 justifiable by some reasonable hypotheses?

**Weibull Plots: Parameters $(\alpha, \beta, \gamma)$:**

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



Red:(1, 0, 1) (exponential); Green:(1, 0, 2); Cyan:(1, 2, 2); Blue:(1, 2, 4)
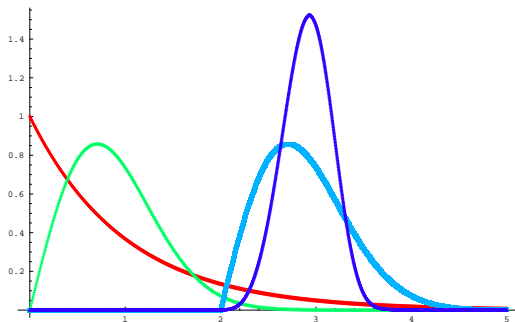
**Three Parameter Weibull: Applications**

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

Arises in many places, such as survival analysis.

- $\gamma < 1$: high infant mortality;
- $\gamma = 1$: constant failure rate;
- $\gamma > 1$: aging process.
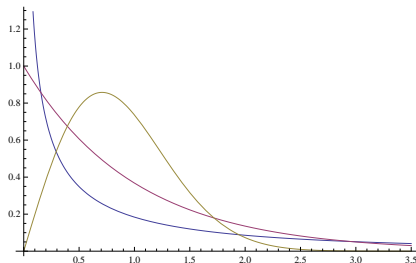
**The Gamma Distribution and Weibulls**

- For $s > 0$, define the $\Gamma$-function by

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

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

A Weibull distribution with parameters $\alpha, \beta, \gamma$ has:

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

**Weibull Integrations**

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

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

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

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 | 36-25 | 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, 6:05P |
| 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*? Why?

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

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

**Estimating Winning Percentages**

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

How can we test these candidates?

Can you think of answers for special choices of $p$ and $q$?

**Estimating Winning Percentages**

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

Homework: explore the following:
$\diamond$ $p = 1$, $q < 1$ (do not want the battle of the undefeated).

$\diamond$ $p = 0$, $q > 0$ (do not want the Toilet Bowl).

$\diamond$ $p = q$.

$\diamond$ $p > q$ (can do $q < 1/2$ and $q > 1/2$).

$\diamond$ Anything else where you 'know' the answer?

## Estimating Winning Percentages

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

**Estimating Winning Percentages**

$$\frac{p - pq}{p + q - 2pq} = \frac{p(1 - q)}{p(1 - q) + (1 - p)q}$$

Homework: explore the following:

$\diamond$ $p = 1$, $q < 1$ (do not want the battle of the undefeated).

$\diamond$ $p = 0$, $q > 0$ (do not want the Toilet Bowl).

$\diamond$ $p = q$.

$\diamond$ $p > q$ (can do $q < 1/2$ and $q > 1/2$).

$\diamond$ Anything else where you 'know' the answer?

## Estimating Winning Percentages: 'Proof'

Start

A has a good game with probability p

B has a good game with probability q

**Figure:** First see how A does, then B

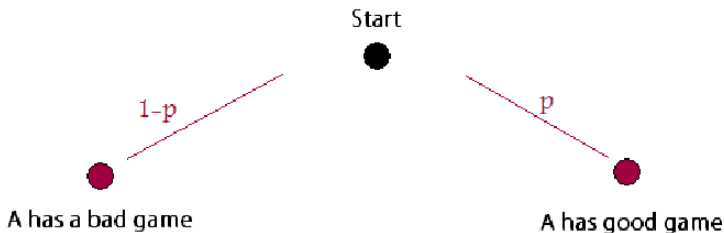**Estimating Winning Percentages: 'Proof'**



**Figure:** Two possibilities: *A* has a good day, or *A* doesn't.
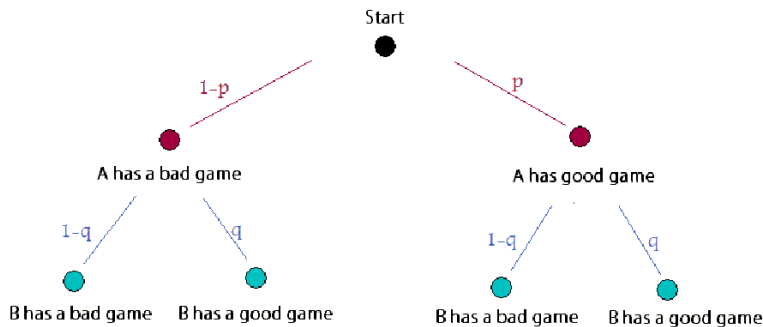
**Estimating Winning Percentages: 'Proof'**



**Figure:** *B* has a good day, or doesn't.

Intro
0000000

Prob & Modeling
00000000000

Pythag Thm
0000000

Analysis of '04
0000000

Adv Theory
000

Summary
0000

Refs
0

Appendices
00000000000

## Estimating Winning Percentages: 'Proof'
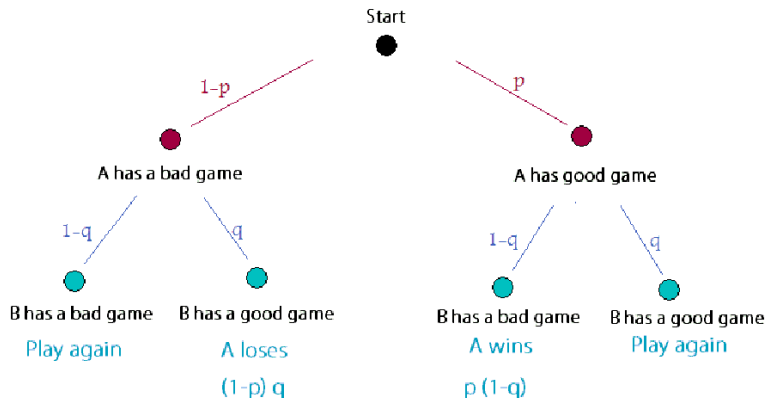


**Figure:** Two paths terminate, two start again.
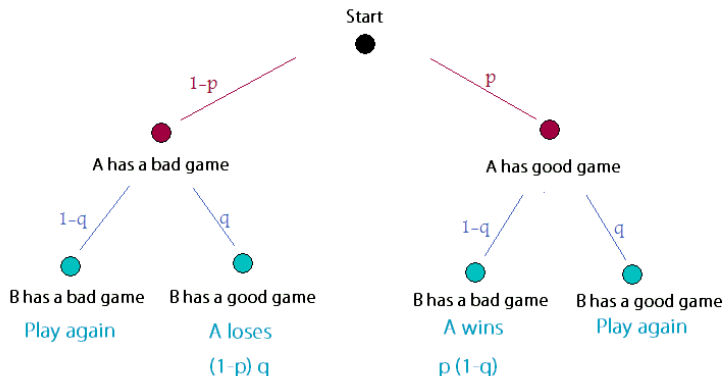
Intro
○○○○○○○
Prob & Modeling
○○○○○○○○○○○
Pythag Thm
○○○●○○○○
Analysis of '04
○○○○○○○
Adv Theory
○○○
Summary
○○○○
Refs
○
Appendices
○○○○○○○○○○○○

# Estimating Winning Percentages: 'Proof'



Figure: Probability $A$ beats $B$

**Pythagorean Won–Loss Formula:** $\frac{\mathrm{RS}_{\mathrm{obs}}^{\gamma}}{\mathrm{RS}_{\mathrm{obs}}^{\gamma} + \mathrm{RA}_{\mathrm{obs}}^{\gamma}}$

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

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

**Pythagorean Won–Loss Formula:** $\dfrac{\mathrm{RS}_{\mathrm{obs}}^{\gamma}}{\mathrm{RS}_{\mathrm{obs}}^{\gamma}+\mathrm{RA}_{\mathrm{obs}}^{\gamma}}$

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

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

Take $\beta = -1/2$ (since runs must be integers).
$\mathrm{RS} - \beta$ estimates average runs scored, $\mathrm{RA} - \beta$ estimates average runs allowed.
Weibull with parameters $(\alpha, \beta, \gamma)$ has mean
$\alpha\Gamma\left(1 + 1/\gamma\right) + \beta$.

**Proof of the Pythagorean Won–Loss Formula**

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

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

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)**

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

$$
= \int_{\beta}^{\infty} \int_{\beta}^{x} \frac{\gamma}{\alpha_{\text{RS}}} \left( \frac{x - \beta}{\alpha_{RS}} \right)^{\gamma-1} e^{-\left( \frac{x-\beta}{\alpha_{\text{RS}}} \right)^{\gamma}} \frac{\gamma}{\alpha_{\text{RA}}} \left( \frac{y - \beta}{\alpha_{\text{RA}}} \right)^{\gamma-1} e^{-\left( \frac{y-\beta}{\alpha_{\text{RA}}} \right)^{\gamma}} \mathrm{d}y \mathrm{d}x
$$

$$
= \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\text{RS}}} \left( \frac{x}{\alpha_{RS}} \right)^{\gamma-1} e^{-\left( \frac{x}{\alpha_{\text{RS}}} \right)^{\gamma}} \left[ \int_{y=0}^{x} \frac{\gamma}{\alpha_{\text{RA}}} \left( \frac{y}{\alpha_{\text{RA}}} \right)^{\gamma-1} e^{-\left( \frac{y}{\alpha_{\text{RA}}} \right)^{\gamma}} \mathrm{d}y \right] \mathrm{d}x
$$

$$
= \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\text{RS}}} \left( \frac{x}{\alpha_{RS}} \right)^{\gamma-1} e^{-(x/\alpha_{\text{RS}})^{\gamma}} \left[ 1 - e^{-(x/\alpha_{\text{RA}})^{\gamma}} \right] \mathrm{d}x
$$

$$
= 1 - \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\text{RS}}} \left( \frac{x}{\alpha_{RS}} \right)^{\gamma-1} e^{-(x/\alpha)^{\gamma}} \mathrm{d}x,
$$

where we have set

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

Intro
0000000
Prob & Modeling
00000000000
Pythag Thm
0000000●
Analysis of '04
0000000
Adv Theory
000
Summary
0000
Refs
○
Appendices
00000000000

**Proof of the Pythagorean Won–Loss Formula (cont)**

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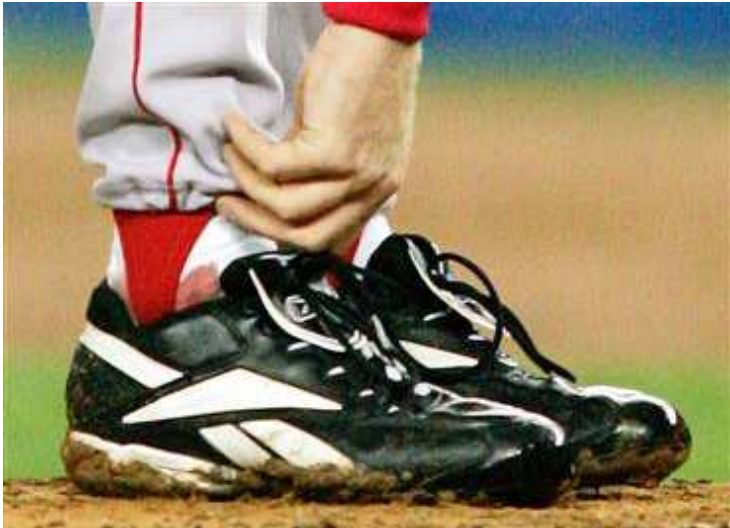$$

**Proof of the Pythagorean Won–Loss Formula (cont)**

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

We substitute the relations for $\alpha_{\mathrm{RS}}$ and $\alpha_{\mathrm{RA}}$ and find that

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

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

## 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
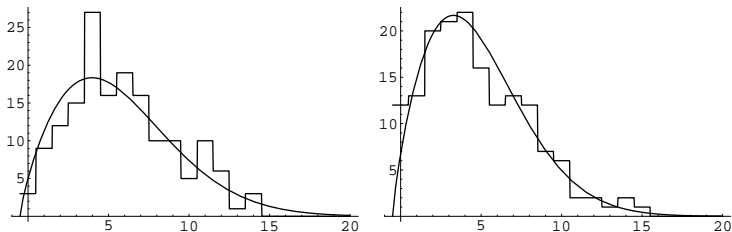


Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [7.5, 8.5]$
$\cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.
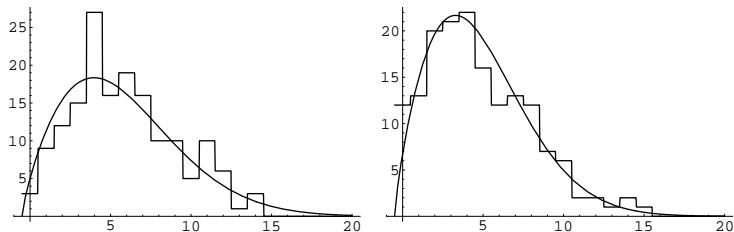
**Best Fit Weibulls to Data: Method of Least Squares**

- $\mathrm{Bin}(k)$ is the $k^{\text{th}}$ 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 $\mathrm{Bin}(k)$;
- $A(\alpha, \gamma, k)$ the area under the Weibull with parameters $(\alpha, -1/2, \gamma)$ in $\mathrm{Bin}(k)$.

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

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

**Best Fit Weibulls to Data (Method of Maximum Likelihood)**
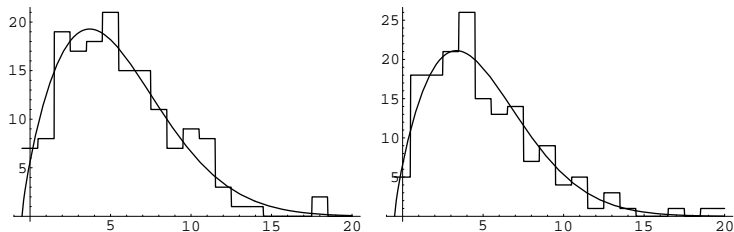
Plots of RS (predicted vs observed) and RA (predicted vs observed) for the Boston Red Sox



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

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

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

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

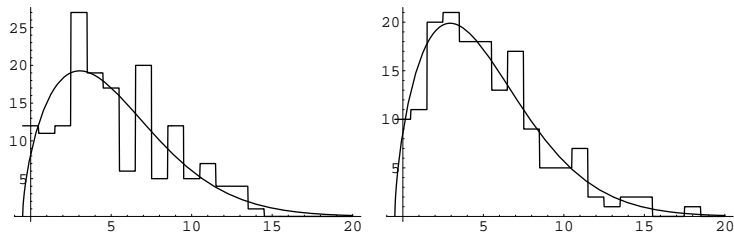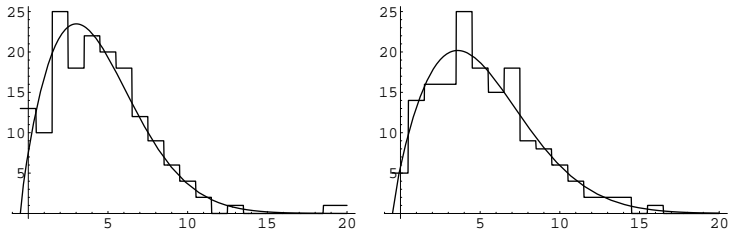**Best Fit Weibulls to Data (Method of Maximum Likelihood)**

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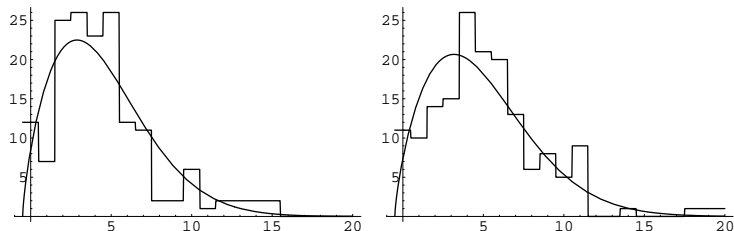


Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

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 $499,000 \leq \#\mathrm{Heads} \leq 501,000$.

**Bonferroni Adjustments**

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

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* $499,000 \leq \#\text{Heads} \leq 501,000$?

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

See unlikely events happen as *N* increases!

**Data Analysis: $\chi^2$ Tests (20 and 109 degrees of freedom)**

| 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).

**Data Analysis: Structural Zeros**

- For independence of runs scored and allowed, use bins $[0, 1) \cup [1, 2) \cup [2, 3) \cup \cdots \cup [8, 9) \cup [9, 10) \cup [10, 11) \cup [11, \infty)$.

- 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 | $\gamma$ |
|------|----------|-----------|---------|----------|-----------|----------|
| 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 |

$\gamma$: 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 $\gamma$ (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 $\gamma$ 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.

**Smoots**

Sieze opportunities: Never know where they will lead.

**Smoots**

Sieze opportunities: Never know where they will lead.

**Smoots**

Sieze opportunities: Never know where they will lead.



Oliver Smoot: Chairman of the American National
Standards Institute (ANSI) from 2001 to 2002, President
of the International Organization for Standardization (ISO)
from 2003 to 2004.

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://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/PythagWonLoss_Pape

  ◇ *Pythagoras at the Bat* (with Taylor Corcoran, Jennifer Gossels, Victor Luo, Jaclyn Porfilio). Book chapter in Social Networks and the Economics of Sports (organized by Victor Zamaraev), to be published by Springer-Verlag. http://arxiv.org/pdf/1406.0758.

  ◇ *Relieving and Readjusting Pythagoras* (senior thesis of Victor Luo, 2014).
  http://arxiv.org/pdf/1406.3402.

Appendices

**Appendix I: Proof of the Pythagorean Won–Loss Formula**

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

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

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)**

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

where we have set

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

**Appendix I: 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 $\alpha_{\text{RS}}$ and $\alpha_{\text{RA}}$ 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 $\text{RS}_{\text{obs}}$, $\text{RA} - \beta$ estimates $\text{RA}_{\text{obs}}$.

**Appendix II: Best Fit Weibulls and Structural Zeros**

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

- Let $\mathrm{Bin}(k)$ denote the $k^{\text{th}}$ bin.
- $O_{r,c}$: the observed number of games where the team's runs scored is in $\mathrm{Bin}(r)$ and the runs allowed are in $\mathrm{Bin}(c)$.
- $E_{r,c} = \frac{\sum_{c'} O_{r,c'} \cdot \sum_{r'} O_{r',c}}{\#\text{Games}}$ 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 $\chi^2$ 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

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

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

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

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 $g$:

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

$X_1$ and $X_2$ independent random variables with probability density $p$.

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

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

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

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

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

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

● Assume $X_1, X_2, \ldots$ 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:

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

## Appendix III: Proof of the Central Limit Theorem

- The Fourier transform of $p$ is

$$\widehat{p}(y) \;=\; \int_{-\infty}^{\infty} p(x)e^{-2\pi ixy}\mathrm{d}x.$$

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

$$\widehat{g}'(y) \;=\; \int_{-\infty}^{\infty} 2\pi ix \cdot g(x)e^{-2\pi ixy}\mathrm{d}x.$$

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

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

- We Taylor expand $\widehat{p}$ (need technical conditions on $p$):

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

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

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

- $\text{Prob}(X_1 + \cdots + X_N \in [a, b]) = \int_a^b (p * \cdots * p)(z)\mathrm{d}z.$

- The Fourier transform converts convolution to multiplication. If $\text{FT}[f](y)$ denotes the Fourier transform of $f$ evaluated at $y$:

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

- Do not want the distribution of $X_1 + \cdots + X_N = x$, but rather $S_N = \frac{X_1 + \cdots + X_N}{\sqrt{N}} = x.$

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

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

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

## 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 $N \to \infty$ 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** $y$,

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

● Fourier transform of $e^{-2\pi y^2}$ 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 $\mathrm{Bin}(k)$ of the Weibull with
parameters $\alpha, -.5, \gamma$. The sample likelihood function
$L(\alpha_{RS}, \alpha_{RA}, -.5, \gamma)$ is

$$\binom{\#\mathrm{Games}}{\mathrm{RS}_{\mathrm{obs}}(1), \ldots, \mathrm{RS}_{\mathrm{obs}}(\#\mathrm{Bins})} \prod_{k=1}^{\#\mathrm{Bins}} A(\alpha_{RS}, -.5, \gamma, k)^{\mathrm{RS}_{\mathrm{obs}}(k)}$$

$$\cdot \binom{\#\mathrm{Games}}{\mathrm{RA}_{\mathrm{obs}}(1), \ldots, \mathrm{RA}_{\mathrm{obs}}(\#\mathrm{Bins})} \prod_{k=1}^{\#\mathrm{Bins}} A(\alpha_{RA}, -.5, \gamma, k)^{\mathrm{RA}_{\mathrm{obs}}(k)}.$$

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