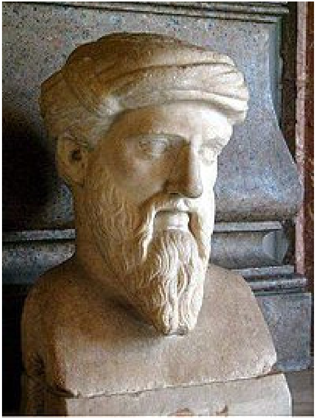
Pythagoras at the Bat: An Introduction to Stats and Modeling

Richard Cleary (rcleary@babson.edu) and
Steven J. Miller (sjm1@williams.edu)

SIGMAA Math & Sports Session: JMM, January 7, 2026

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

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!

Statistics

Goal is to find good statistics to describe real world.

Statistics

Goal is to find good statistics to describe real world.

Figure: Mass Ave Bridge, about 620.1 meters.

Statistics

Goal is to find good statistics to describe real world.

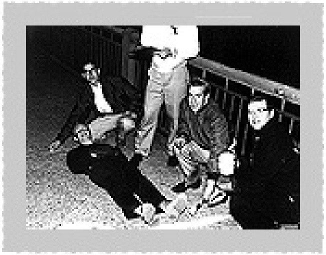


Figure: Harvard Bridge, 364.1 Smoots ($\pm$ one ear).

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.

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)

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

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

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

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).

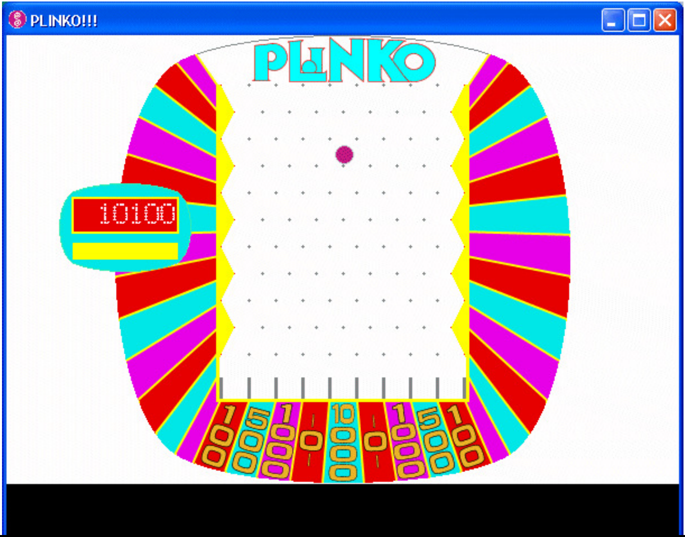
2011: Red Sox 'should' be 95-67, Tampa 'should' be 92-70....

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.

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



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

Won–Loss formula follows from computing

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

Problems with the Model

Reduced to calculating

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

Problems with the model:

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

Three Parameter Weibull

Weibull distribution:

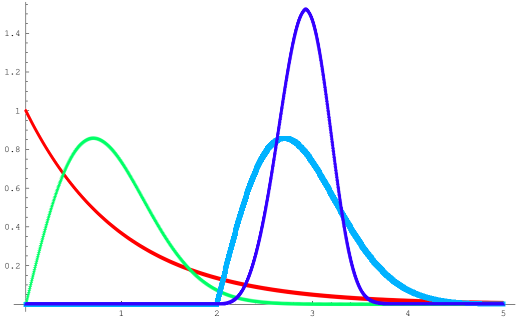
$$f(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$

- 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):

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



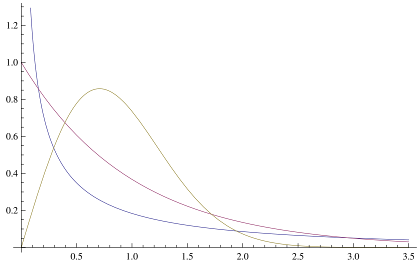
Red:(1, 0, 1) (exponential); Green:(1, 0, 2); Cyan:(1, 2, 2); Blue:(1, 2, 4)

Three Parameter Weibull: Applications

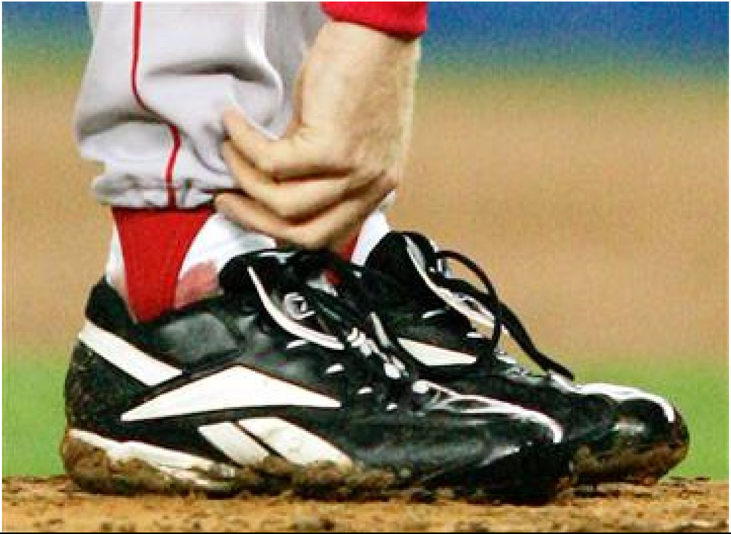
$$f(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$

Arises in many places, such as survival analysis.

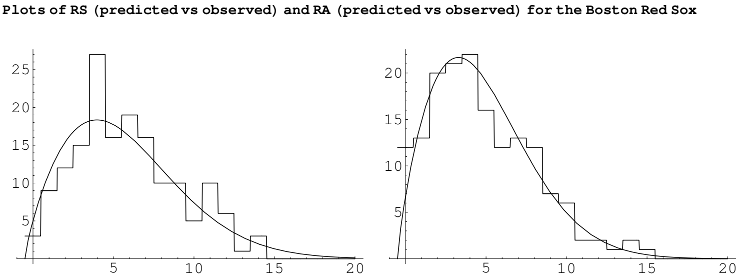
- $\gamma < 1$: high infant mortality;
- $\gamma = 1$: constant failure rate;
- $\gamma > 1$: aging process.



Analysis of 2004



Best Fit Weibulls to Data (Method of Maximum Likelihood)



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

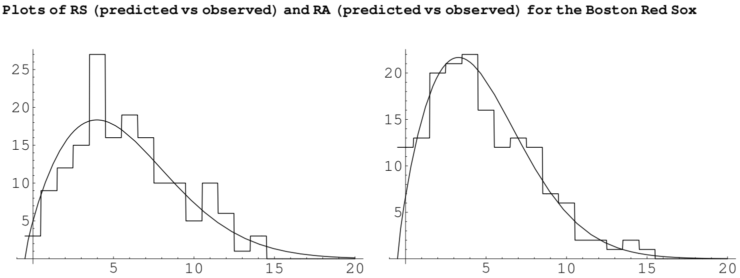
Best Fit Weibulls to Data: Method of Least Squares

- $\mathrm{Bin}(k)$ is the k^{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

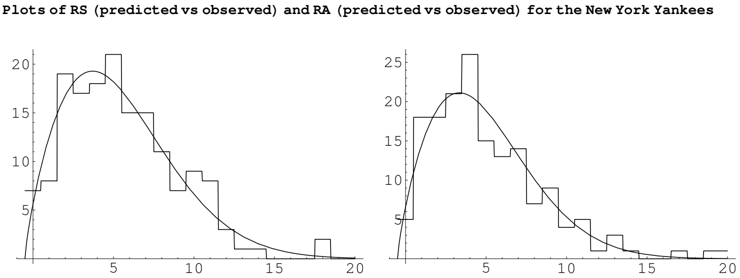
$$\sum_{k=1}^{\#\mathrm{Bins}} \left(\mathrm{RS}_{\mathrm{obs}}(k) - \#\mathrm{Games} \cdot A(\alpha_{\mathrm{RS}}, \gamma, k)\right)^2 + \sum_{k=1}^{\#\mathrm{Bins}} \left(\mathrm{RA}_{\mathrm{obs}}(k) - \#\mathrm{Games} \cdot A(\alpha_{\mathrm{RA}}, \gamma, k)\right)^2 .$$

Best Fit Weibulls to Data (Method of Maximum Likelihood)



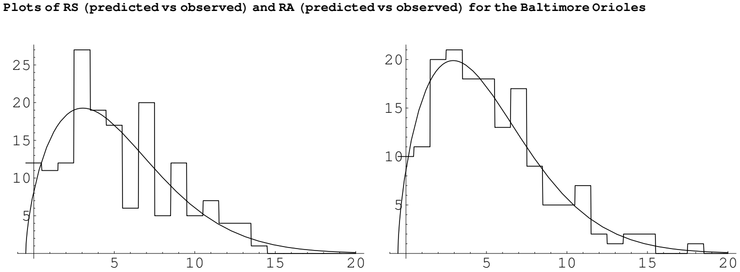
Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Best Fit Weibulls to Data (Method of Maximum Likelihood)



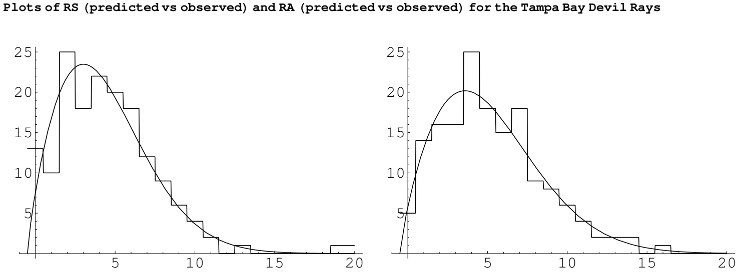
Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Best Fit Weibulls to Data (Method of Maximum Likelihood)



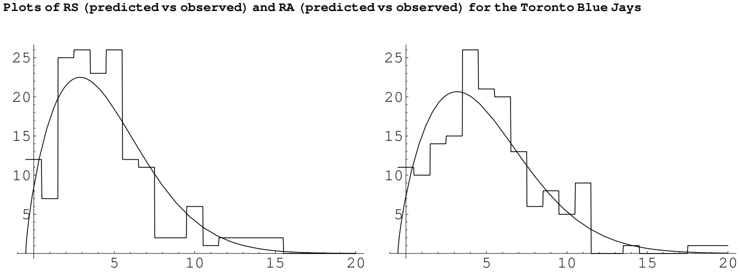
Using as bins $[-.5, .5] \cup [.5, 1.5] \cup \cdots \cup [7.5, 8.5] \cup [8.5, 9.5] \cup [9.5, 11.5] \cup [11.5, \infty)$.

Best Fit Weibulls to Data (Method of Maximum Likelihood)



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)



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

Head-to-Head

Issues with Pythagorean Head-to-Head

Does not ensure league averages to 500.

In 2025 predicts teams win on average 81.30 and lose 80.70 games.

In a 7 game series predicts Blue Jays win 3.82 out of 7, Dodgers win 4.10.

Issue: Does not take into account data from both teams. How to fix?

New Application: Head-to-Head

James Log-5 Method estimates the probability A beats B if A wins p and B wins q percent of the time:

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

New Application: Head-to-Head

James Log-5 Method estimates the probability A beats B if A wins p and B wins q percent of the time:

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

How to generalize with Pythagorean formula?

Joint with: Jake Jeffries, Cam Miller, James Murray, Sasha Palma and Nick Skiera.
Preprint: https://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/PythagBothTeams10.pdf.

New Application: Head-to-Head (cont)

Adjust Pythagorean Formula, use both teams:

- home team $\mathrm{RS_h}, \mathrm{RA_h}$,
- away team $\mathrm{RS_a}, \mathrm{RA_a}$,
- league average runs scored per game is R,

New Application: Head-to-Head (cont)

Adjust Pythagorean Formula, use both teams:

- home team $\mathrm{RS_h}, \mathrm{RA_h}$,
- away team $\mathrm{RS_a}, \mathrm{RA_a}$,
- league average runs scored per game is R,
- adjusted home numbers:
  $\mathrm{RS_{h,adj}} = \mathrm{RS_h}(\mathrm{RA_a}/\mathrm{R})$,
  $\mathrm{RA_{h,adj}} = \mathrm{RA_h}(\mathrm{RS_a}/\mathrm{R})$:

New Application: Head-to-Head (cont)

Adjust Pythagorean Formula, use both teams:

- home team $\mathrm{RS_h}, \mathrm{RA_h}$,
- away team $\mathrm{RS_a}, \mathrm{RA_a}$,
- league average runs scored per game is R,
- adjusted home numbers:
  $\mathrm{RS_{h,adj}} = \mathrm{RS_h}(\mathrm{RA_a}/\mathrm{R})$,
  $\mathrm{RA_{h,adj}} = \mathrm{RA_h}(\mathrm{RS_a}/\mathrm{R})$:

$$\begin{aligned}&\mathrm{Prob(Home\ Team\ Wins)}\\ &= \frac{\mathrm{RS}_{\mathrm{h,adj}}^{\gamma}}{\mathrm{RS}_{\mathrm{h,adj}}^{\gamma} + \mathrm{RA}_{\mathrm{h,adj}}^{\gamma}} = \frac{(\mathrm{RS_h RA_a})^{\gamma}}{(\mathrm{RS_h RA_a})^{\gamma} + (\mathrm{RA_h RS_a})^{\gamma}}.\end{aligned}$$

New Application: Head-to-Head: Data

Looked at playoffs from 2001 – 2019.

Compared observed series won by home team to predicted (if predict home team wins with probability .72, count that as .72 of a win for home and .28 of a win for away).

Log-5: home wins 83.19 and loses 65.81.
Observed: home wins 80.00 and loses 69.00.

New Application: Head-to-Head: Data

Looked at playoffs from 2001 – 2019.

Compared observed series won by home team to predicted (if predict home team wins with probability .72, count that as .72 of a win for home and .28 of a win for away).

Log-5: home wins 83.19 and loses 65.81.
Observed: home wins 80.00 and loses 69.00.

Predicted: home wins 80.18 and loses 68.82!

New Application: Head-to-Head: Exponent

New adjusted numbers: What exponent b is best?

- $\mathrm{RS}_{\mathrm{h,adj}} = \mathrm{RS}_{\mathrm{h}}(\mathrm{RA}_{\mathrm{a}}/\mathrm{R})^b$.
- $\mathrm{RA}_{\mathrm{h,adj}} = \mathrm{RA}_{\mathrm{h}}(\mathrm{RS}_{\mathrm{a}}/\mathrm{R})^b$.
- $b = 0$ no adjustment; none if league average.
- $b \to \infty$: tremendous impact to small changes.

If symmetric (so average to .500) only possibility is $b = 1$.

Head-to-Head: Exponent II (from paper)

It is important to note that the probabilities summing to 1 would not hold in general if instead of rescaling by quantities such as $\mathrm{RS_a/R}$ we instead rescaled by $(\mathrm{RS_a/R})^b$ for $b \neq 1$; doing so would magnify or diminish the adjustment (as $b \to 0$ it reduces to the original Pythagorean formula, while $b \to \infty$ gives tremendous impact to small changes): in obvious notation we now have

$$\begin{aligned} P_{h,a}(b) &= \frac{(\mathrm{RS_h RA_a^{\it b}})^\gamma}{(\mathrm{RS_h RA_a^{\it b}})^\gamma + (\mathrm{RA_h RS_a^{\it b}})^\gamma} + \frac{(\mathrm{RS_a RA_h^{\it b}})^\gamma}{(\mathrm{RS_a RA_h^{\it b}})^\gamma + (\mathrm{RA_a RS_h^{\it b}})^\gamma} \\ &= \frac{\sigma_h \alpha_a^b}{\sigma_h \alpha_a^b + \alpha_h \sigma_a^b} + \frac{\sigma_a \alpha_h^b}{\sigma_a \alpha_h^b + \alpha_a \sigma_h^b} \\ &= \frac{\sigma_h \alpha_a^b(\sigma_a \alpha_h^b + \alpha_a \sigma_h^b) + \sigma_a \alpha_h^b(\sigma_h \alpha_a^b + \alpha_h \sigma_a^b)}{(\sigma_h \alpha_a^b + \alpha_h \sigma_a^b)(\sigma_a \alpha_h^b + \alpha_a \sigma_h^b)} \\ &= \frac{\sigma_h \sigma_a \alpha_h^b \alpha_a^b + \sigma_h^{b+1} \alpha_a^{b+1} + \sigma_h \sigma_a \alpha_h^b \alpha_a^b + \sigma_a^{b+1} \alpha_h^{b+1}}{\sigma_h \sigma_a \alpha_h^b \alpha_a^b + \sigma_h^{b+1} \alpha_a^{b+1} + \sigma_h^b \sigma_a^b \alpha_h \alpha_a + \sigma_a^{b+1} \alpha_h^{b+1}}, \end{aligned}$$

and if $b \neq 1$ the third (after sorting) term in the numerator does not match the corresponding term in the denominator, though all the other terms do match. It is interesting that the only adjustment which is permissible under symmetry constraints (as the probability one team wins must equal the probability the other loses) is a simple multiplicative rescaling.

2025 World Series: Dodgers vs Rays

Predicts 46% chance of Toronto winning.

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.
- Adjusted Pythagorean formula for head-to-head match-ups.

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.

Some References

References

◇ S. Baxamusa, *Weibull worksheet*,
http://www.beyondtheboxscore.com/story/2006/4/30/114737/251.
◇ S. Baxamusa, *Run distribution plots for various teams*,
http://www.beyondtheboxscore.com/story/2006/2/23/164417/484.
◇ P. Birnbaum, *Sabermetric Research: Saturday, April 24, 2010*, see
http://blog.philbirnbaum.com/2010/04/marginal-value-of-win-in-baseball.html.
◇ R. Cleary, J. Jeffries, C. Miller, S. J. Miller, J. Murray, S. Palma and N. Skiera, *Adjusting James' Pythagorean Theorem for Head-to-Head Matchups*, preprint, https://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/PythagBothTeams10.pdf.
◇ R. Clearly and P. Staab, *Same-Score Streaks: A Case Study in Modeling*, Math Horizons **28** (2021), no. 4, pages 5–9. https://www.tandfonline.com/doi/citedby/10.1080/10724117.2021.1881355?scroll=top&needAccess=true.
◇ T. Corcoran, J. Gossels, V. Luo, S. J. Miller and J. Porfilio, *Pythagoras at the Bat*, in Social Networks and the Economics of Sports (edited by Panos M. Pardalos and Victor Zamaraev), Springer-Verlag, 2014, pages 89–114. http://arxiv.org/pdf/1406.0758.
◇ K. Dayaratna and S. J. Miller, *First Order Approximations of the Pythagorean Won-Loss Formula for Predicting MLB Teams Winning Percentages*, By The Numbers – The Newsletter of the SABR Statistical Analysis Committee **22** (2012), no 1, 15–19.
◇ H. Hundel, *Derivation of James' Pythagorean Formula*, 2003; see
https://groups.google.com/forum/#!topic/rec.puzzles/O-DmrUljHds.
◇ B. James, *1981 Baseball Abstract*, self-published, Lawrence, KS, 1981.
◇ M. Jones and L. Tappin, *The Pythagorean Theorem of Baseball and Alternative Models*, The UMAP Journal **26** (2005), no. 2, 12 pages. https://www.comap.com/membership/member-resources/item/the-pythagorean-theorem-of-baseball-and-alternative-models-umap.
◇ V. Luo and S. J. Miller, *Relieving and Readjusting Pythagoras*, By The Numbers – The Newsletter of the SABR Statistical Analysis Committee **25** (2015), no. 1, 5–14.
◇ S. J. Miller, *A derivation of the Pythagorean Won-Loss Formula in baseball*, Chance Magazine **20** (2007), no. 1, 40–48 (an abridged version appeared in The Newsletter of the SABR Statistical Analysis Committee **16** (February 2006), no. 1, 17–22, and an expanded version is online at http://arxiv.org/pdf/math/0509698).
◇ A. K. Yang and S. J. Miller, *Code and Data for Applications of Improvements to the Pythagorean Won-Lost Expectation in Optimizing Rosters*, https://web.williams.edu/Mathematics/sjmiller/public_html/

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, 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
Chicago	38	22	.633	-	9-1	L1	0-0	26-8	12-14	39-21	6/4 @ SD, L 1-2	6/5 @ LAD, 10:10P

The Gamma Distribution and Weibulls

- For $s > 0$, define the Γ-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 α, β, γ has:

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

Weibull Integrations

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

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

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

$$\mu_{\alpha,\beta,\gamma} = \int_0^\infty \alpha u^{1/\gamma} \cdot e^{-u} \mathrm{d}u \; + \; \beta$$

$$= \alpha \int_0^\infty e^{-u} \, u^{1+1/\gamma} \, \frac{\mathrm{d}u}{u} \; + \; \beta$$

$$= \alpha\Gamma(1+1/\gamma) \; + \; \beta.$$

A similar calculation determines the variance.

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 $(\alpha_{\rm RS}, \beta, \gamma)$ and $(\alpha_{\rm RA}, \beta, \gamma)$ respectively. To have means of $\rm RS - \beta$ and $\rm RA - \beta$ our calculations for the means imply

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

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)

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

where we have set

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

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

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

Note $\mathrm{RS} - \beta$ estimates $\mathrm{RS}_{\rm obs}$, $\mathrm{RA} - \beta$ estimates $\mathrm{RA}_{\rm obs}$.

Appendices

Appendix I: Proof of the Pythagorean Won–Loss Formula

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

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

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)

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

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

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

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

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

where we have set

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

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

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

We substitute the relations for $\alpha_{\rm RS}$ and $\alpha_{\rm RA}$ 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 $\mathrm{Bin}(k)$ denote the k^{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 $\mathrm{Bin}(c)$.
- $E_{r,c} = \frac{\sum_{c'} O_{r,c'} \cdot \sum_{r'} O_{r',c}}{\#\mathrm{Games}}$ is the expected frequency of cell (r, c).
- Then

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

is a χ^2 distribution with $(\#\mathrm{Rows} - 1)(\#\mathrm{Columns} - 1)$ 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 $1 \le r, c \le 12$, 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: 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:

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

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

◇ $p = q$.

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

◇ Anything else where you 'know' the answer?

Estimating Winning Percentages

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

Estimating Winning Percentages

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

Homework: explore the following:

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

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

◇ $p = q$.

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

◇ Anything else where you 'know' the answer?

Estimating Winning Percentages: 'Proof'

Figure: First see how A does, then B.

Estimating Winning Percentages: 'Proof'

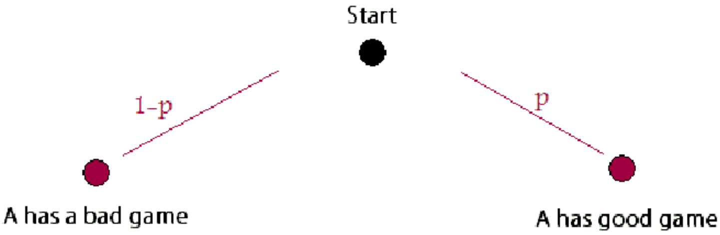


Figure: Two possibilities: *A* has a good day, or *A* doesn't.

Estimating Winning Percentages: 'Proof'

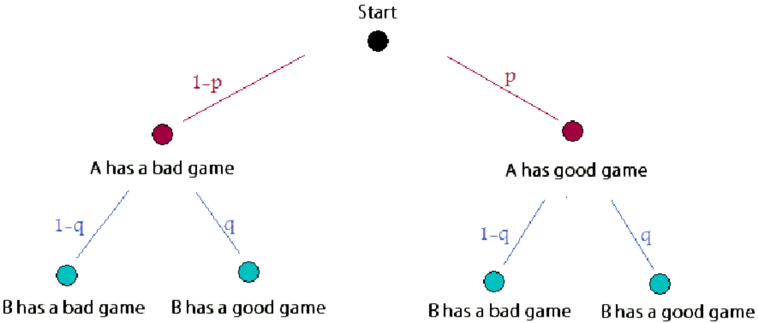


Figure: *B* has a good day, or doesn't.

Estimating Winning Percentages: 'Proof'

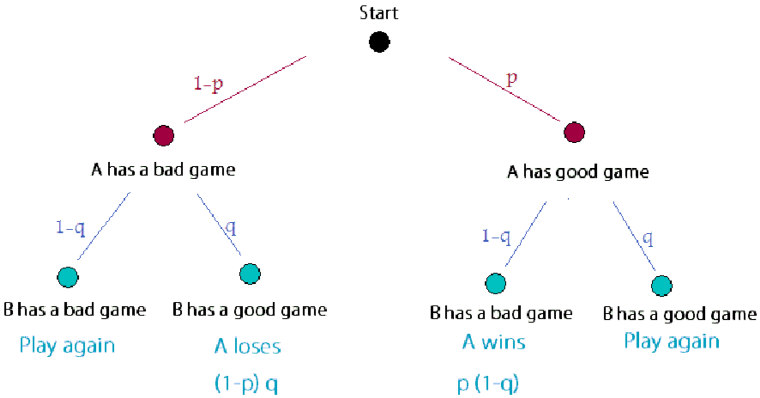


Figure: Two paths terminate, two start again.

Estimating Winning Percentages: 'Proof'

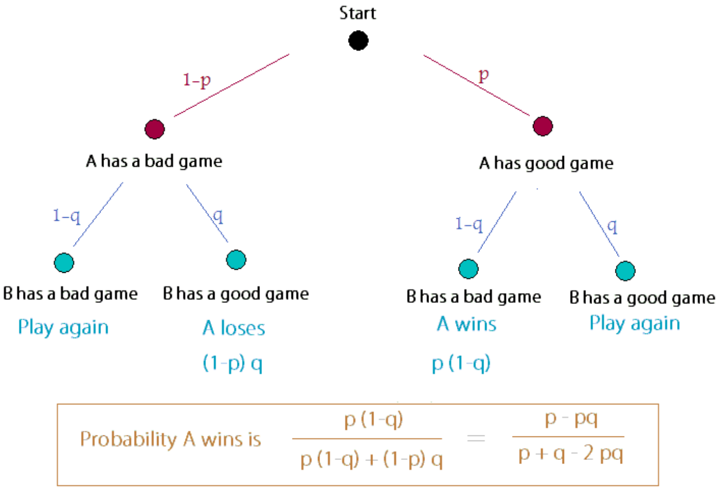


Figure: Probability *A* beats *B*

Appendix IV: Best Fit Weibulls from Method of Maximum Likelihood

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

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

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

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