Pythagoras at the Bat: An Introduction to Stats and Modeling

Steven J. Miller (sjm1@williams.edu)
Department of Math/Stats, Williams College
(faculty advisor to the baseball team)

Williams NorCal: San Francisco, January 4th, 2024

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

Goals of the talk:

- What are the right statistics to study?
- Discuss some fun applications of math/stats to baseball and life.
- Talk about opportunities for students at Williams.

Joint with many Williams students over the years, including:

John Bihn, Alex Cardonick, Mac Carso, George Carroll, Lucas Casso, Elliot Chester, Jakob Cohn, Thomas Coleman, Dan Costanza, Kevin Deptula, Jacob Eckerle, Bryan Eckleman, Cameron Edgar, Carson Eisenach, Alan Felix, Cole Futterman, William Gerson, Jen Gossels, Jay Habib, Charles Ide, Jake Jeffries, William Jeffries, Aviv Lipman, Victor Luo, Alex Matthews, Kyle McGrath, Cameron Miller, David Moon, James Murray, Sasha Palma, Jackson Parese, Francesca Paris, Bo Peponis, David Phillips, Chris Picardo, Jaclyn Porfilio, Ethan Prout, Tim Randolph, Evan Ruschil, John Sandifer, Jacob Siegel, Nick Skiera, Evan Skorpen, Michael Stone, Cole Whitehouse, Bryan Wooley.

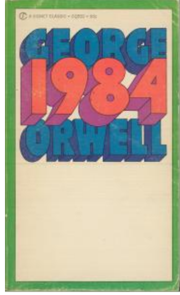
Goal is to find good statistics to describe real world.

Figure: Harvard Bridge, about 620.1 meters.

Goal is to find good statistics to describe real world.

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

Not “Who is on first” but “Who is in first”.



American League

EAST	W	L	PCT	GB
Boston Red Sox	41	19	.683	-
New York Yankees	37	17	.685	1
Tampa Bay Rays	28	30	.483	12
Toronto Blue Jays	26	33	.441	14.5
Baltimore Orioles	17	41	.293	23

American League

AL East	W	L	Pct	GB
Yankees	37	17	.685	1.0
Red Sox	41	19	.683	-
Rays	28	30	.483	12.0
Blue Jays	26	33	.441	14.5
Orioles	17	41	.293	23.0

Choosing the right statistic matters; depending on your choice, at the start of June 4, 2018 either the Red Sox or the NY Yankees are in first place in the American League East, and the other is in second! *(Left from ESPN, right from Google.)*

Standings

https://www.mlb.com/standings/2008-06-05

Regular Season | Wild Card | Postseason Picture | Spring Training

June 5 | 2008 | Division | Standard | Advanced

AL East	W	L	PCT	GB	WCGB	L10	STRK	RS	RA	DIFF	X-W/L	HOME	AWAY	>.500
Boston	38	25	.603	-	-	7-3	W3	321	267	+54	37-26	24-5	14-20	26-21
Tampa Bay	35	25	.583	1.5	-	5-5	L3	263	247	+16	32-28	24-10	11-15	34-23
Toronto	32	30	.516	5.5	4.0	5-5	L2	256	227	+29	34-28	15-11	17-19	24-23
Baltimore	30	30	.500	6.5	5.0	5-5	W1	247	263	-16	28-32	17-11	13-19	19-27
NY Yankees	30	30	.500	6.5	5.0	5-5	W2	272	281	-9	29-31	16-13	14-17	19-20

AL Central	W	L	PCT	GB	WCGB	L10	STRK	RS	RA	DIFF	X-W/L	HOME	AWAY	>.500
Chi White Sox	33	27	.550	-	-	6-4	W3	265	221	+44	35-25	16-10	17-17	14-22
Minnesota	31	29	.517	2.0	4.0	6-4	L1	279	287	-8	29-31	19-16	12-13	16-21
Cleveland	27	33	.450	6.0	8.0	4-6	L1	265	255	+10	31-29	16-16	11-17	20-24
Detroit	24	35	.407	8.5	10.5	3-7	L3	276	302	-26	27-32	12-14	12-21	18-27
Kansas City	23	37	.383	10.0	12.0	2-8	L3	218	281	-63	23-37	12-16	11-21	12-32

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

- $\mathrm{RS}_{\mathrm{obs}}$: average number of runs scored per game;
- $\mathrm{RA}_{\mathrm{obs}}$: average number of runs allowed per game;
- γ: some parameter, constant for a sport.

James' Won–Loss Formula (NUMERICAL Observation)

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

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

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.

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

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

Reduced to calculating

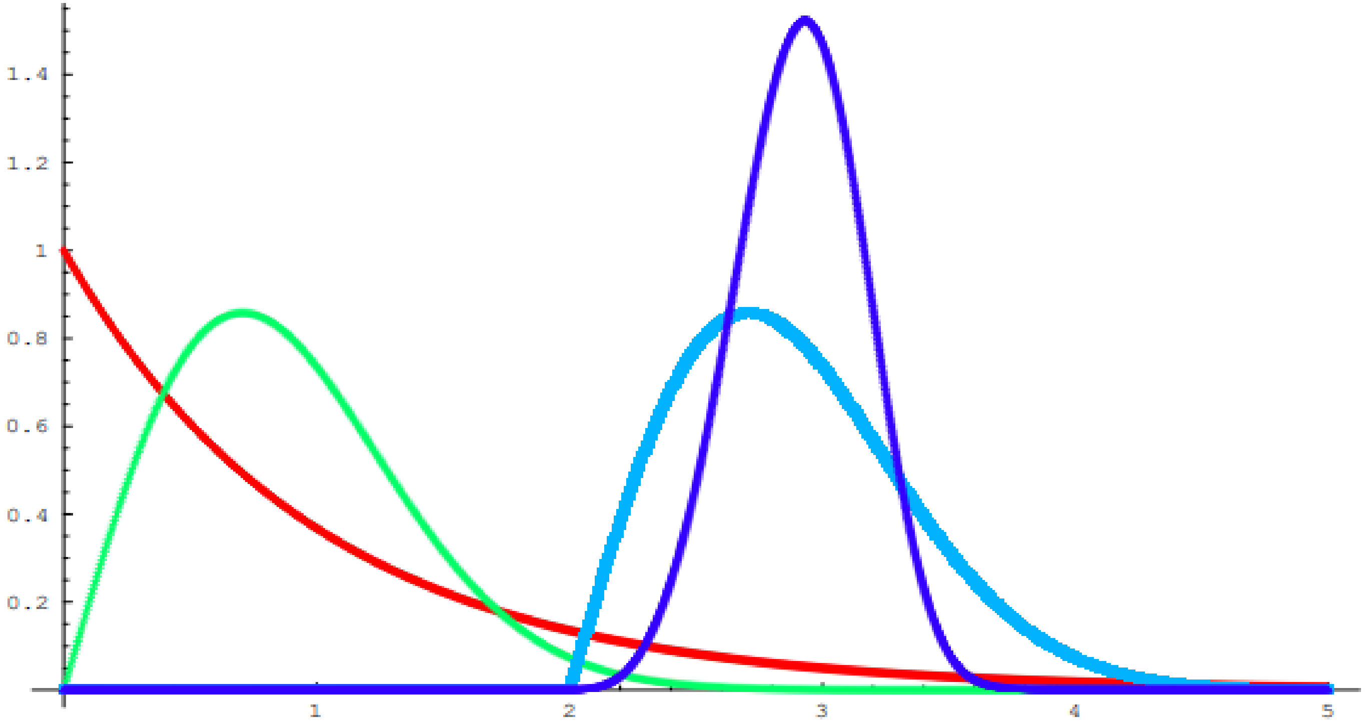
$$\int_{x=0}^{\infty}\left[\int_{y\le x} f_{\mathrm{RS}}(x)g_{\mathrm{RA}}(y)\mathrm{d}y\right]\mathrm{d}x$$

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?

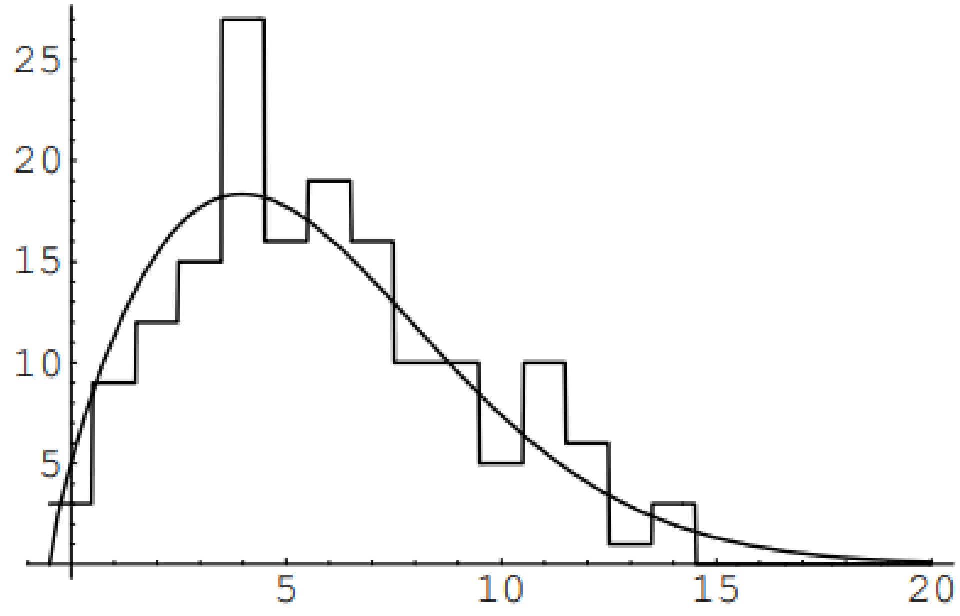
Weibull Plots: Parameters (α, β, γ):

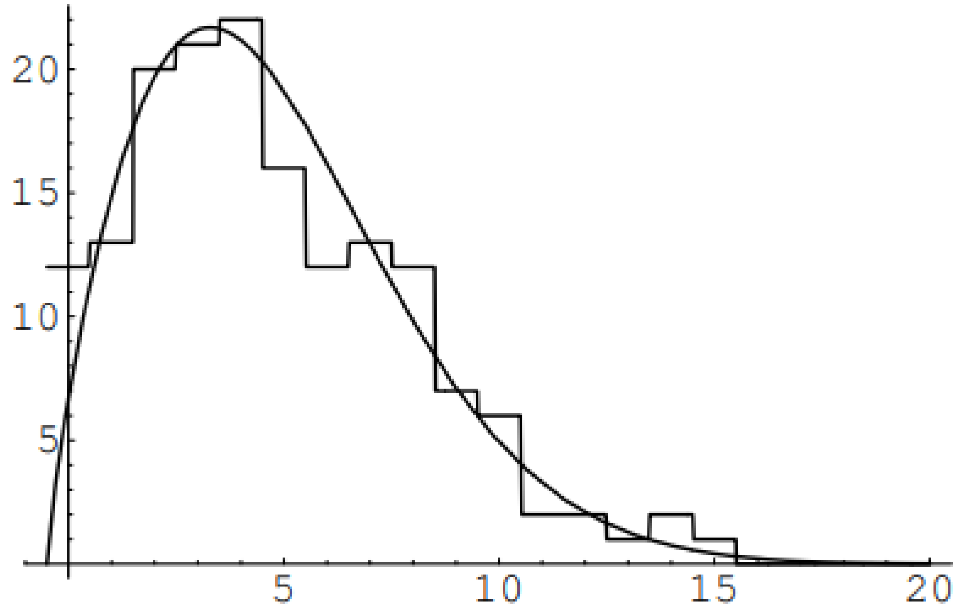
$$f(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$



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

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





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

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 $(\alpha_{\mathrm{RA}}, \beta, \gamma)$; 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 $\frac{(\mathrm{RS}-\beta)^{\gamma}}{(\mathrm{RS}-\beta)^{\gamma}+(\mathrm{RA}-\beta)^{\gamma}}$.

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

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.

Work with Williams Students:

- Calling games before done.
- Ballpark effects.
- Other approximations.
- Applying to individual matchups (Winter Study ‘24).

Other Projects with Williams Students:

- Lawsuit on [redacted].
- MLB lawsuit on [redacted].
- MLB lawsuit on [redacted].

Math 344: Spring 2023: The Mathematics of Sports

(with Alex Cardonick, Mac Carso, Jacob Eckerle, Ethan Prout, Evan Ruschil)

Value of a single vs double vs triple vs home run: is it really 1 – 2 – 3 – 4?

Run Expectancy

	Outs 0	Outs 1	Outs 2
000	0.481	0.254	0.098
100	0.859	0.509	0.224
020	1.100	0.664	0.319
120	1.437	0.884	0.429
003	1.350	0.950	0.353
103	1.784	1.130	0.478
023	1.964	1.376	0.580
123	2.292	1.541	0.752

Game State Probability

	Outs 0	Outs 1	Outs 2
000	0.244	0.175	0.139
100	0.059	0.070	0.071
020	0.015	0.026	0.033
120	0.014	0.025	0.031
003	0.002	0.009	0.014
103	0.005	0.011	0.016
023	0.003	0.007	0.008
123	0.004	0.009	0.011

Triple Gain

	Outs		
	0	**1**	**2**
000	0.869	0.696	0.255
100	1.491	1.441	1.129
020	1.250	1.286	1.034
120	1.913	2.066	1.924
003	1.000	1.000	1.000
103	1.566	1.820	1.875
023	1.386	1.574	1.773
123	2.058	2.409	2.601

HR Gain

	Outs		
	0	**1**	**2**
000	1.000	1.000	1.000
100	1.622	1.745	1.874
020	1.381	1.590	1.779
120	2.044	2.370	2.669
003	1.131	1.304	1.745
103	1.697	2.124	2.620
023	1.517	1.878	2.518
123	2.189	2.713	3.346

Triple Gain - Normalized

	Outs		
	0	**1**	**2**
000	0.212	0.122	0.035
100	0.088	0.101	0.080
020	0.019	0.033	0.034
120	0.027	0.052	0.060
003	0.002	0.009	0.014
103	0.008	0.020	0.030
023	0.004	0.011	0.014
123	0.008	0.022	0.029

Expected Triple Gain 1.033

HR Gain - Normalized

	Outs		
	0	**1**	**2**
000	0.244	0.175	0.139
100	0.096	0.122	0.133
020	0.021	0.041	0.059
120	0.029	0.059	0.083
003	0.002	0.012	0.024
103	0.008	0.023	0.042
023	0.005	0.013	0.020
123	0.009	0.024	0.037

Expected HR Gain 1.420

Model Triple/HR Value	SLG Triple/HR Value
0.728	0.750

Model HR/Triple Value	SLG HR/Triple Value
1.374	1.333

% Difference
3.08%

The ratios above for the model comparisons evaluate the normalized "Expected Triple Gain" divided by the normalized "Expected HR Gain" and compare it to the more well known ratio of 3:4, or 0.750, in the SLG equation. Based on these results, we find that SLG overvalues the triple relative to the home run since our model ratio is smaller than 0.750, representing a 3.08% overvaluation of the triple relative to SLG's value. Clearly a triple is less valuable than a home run; the question is how much less? Basing critical baseball decisions solely on a statistic like SLG may not represent the best decision. Nonetheless, these findings more importantly lay the basic idea and groundwork that lead us to carrying out other, more complex assumptions.

To simplify the analysis, we make the assumption that a single moves a baserunner one base, while a double moves a baserunner two bases.

Single Gain - Normalized

	Outs		
	0	**1**	**2**
000	0.092	0.045	0.018
100	0.034	0.026	0.015
020	0.010	0.012	0.005
120	0.012	0.016	0.010
003	0.001	0.005	0.012
103	0.003	0.008	0.015
023	0.002	0.005	0.007
123	0.004	0.009	0.011

Expected Single Gain **0.379**

Double Gain - Normalized

	Outs		
	0	**1**	**2**
000	0.151	0.072	0.031
100	0.065	0.061	0.025
020	0.015	0.026	0.033
120	0.021	0.037	0.036
003	0.002	0.006	0.014
103	0.006	0.014	0.018
023	0.003	0.009	0.014
123	0.007	0.017	0.020

Expected Double Gain **0.701**

Summing these together, we find that the expected gain from a single and a double to be 0.379 and 0.701, respectively. As mentioned previously, we need to compare how these values relate to the SLG statistic. Within the slugging percentage formula used by the MLB, the relative value of a single over a double is 0.5. However, our analysis shows that relative value should be 0.541. Therefore, the SLG statistic is undervaluing the single over the double by 7.53%. These values, in the context of slugging percentage, are shown below:

Model Single/Double Value	SLG Single/Double Value
0.541	0.500

Model Double/Single Value	SLG Double/Single Value
1.849	2.000

% Difference
-7.53%

Assumptions for Runner Advancement (excluding the batter)

1 base on 1B	2 bases on 1B	3 bases on 1B	2 bases on 2B	3 bases on 2B
0.590	0.400	0.010	0.750	0.250

Resulting Game State - Single

	1 Base		2 Bases		3 Bases	
	Result	Runs Scored	**Result**	Runs Scored	**Result**	Runs Scored
000	100	0	100	0	100	0
100	120	0	103	0	100	1
020	103	0	100	1	100	1
120	123	0	103	1	100	2
003	100	1	100	1	100	1
103	120	1	103	1	100	2
023	103	1	100	2	100	2
123	123	1	103	2	100	3

Resulting Game State - Double

	2 Bases		3 Bases	
	Result	Runs Scored	**Result**	Runs Scored
000	020	0	020	0
100	023	0	020	1
020	020	1	020	1
120	023	1	020	2
003	020	1	020	1
103	023	1	020	2
023	020	2	020	2
123	023	2	020	3

Single Gain - Normalized

	Outs 0	1	2
000	0.092	0.045	0.018
100	0.043	0.034	0.017
020	0.011	0.016	0.015
120	0.015	0.023	0.019
003	0.001	0.005	0.012
103	0.004	0.009	0.016
023	0.003	0.006	0.010
123	0.005	0.011	0.014

Expected Single Gain	**0.442**

Double Gain - Normalized

	Outs 0	1	2
000	0.151	0.072	0.031
100	0.067	0.066	0.038
020	0.015	0.026	0.033
120	0.022	0.039	0.041
003	0.002	0.006	0.014
103	0.006	0.014	0.021
023	0.003	0.009	0.014
123	0.007	0.017	0.022

Expected Double Gain	**0.736**

By summing these all together, we again calculated the average relative contribution of a single and a double. We found in this analysis that on average, a single adds 0.442 expected runs to an inning, and a double adds 0.736 expected runs to an inning.

References

Talks: https://web.williams.edu/Mathematics/sjmiller/public_html/math/talks/talks.html

Baseball / Sabermetrics Talks

- The Pythagorean Won-Loss Formula in Baseball (An Introduction to Statistics and Modeling) (60 minute version) Brown University (9/28/05) and (12/7/05): slides(tab) slide(no tab) paper_pdf. Shorter 15 minute version: 2006 Hudson River Undergraduate Mathematics Conference (4/8/06): slides(tab) slides(no tab)
- Pythagoras at the Bat: An Introduction to Statistics and Modeling (20 minute version): Society For American Baseball Research, Boston Meeting, Boston, MA (5/20/06): slides(tab) slides(no tab) paper_pdf. Williams College (40 minute version, 1/15/08): pdf. Holy Cross (50 minute version, 2/7/08): pdf. Western New England College (40 minute version, 1/15/08): pdf. Connecticut Smoky Joe Wood SABR Chapter, Hamden, CT (30 minute version, 2/16/08) pdf. PROMYS Program, Boston University (40 minute version, 7/25/08): pdf. Bennington College (2/27/09): pdf Hampshire College (7/22/09): pdf Wellesley (9/21/09): pdf Awards Night, University of Connecticut (4/12/10): pdf Virginia Tech (3/28/11): pdf UMass Amherst (10/19/11) and Fitchburg State University (11/3/11): pdf Boston College (3/29/12) pdf Science Days talk to Prospective Williams Students (8/15/14) pdf and 8/12/16 pdf and 8/10/18 pdf University of Vermont (12/2/16, with an intro on undergraduate research) pdf Developer Thursdays at Cloud85 (6/8/17): pdf University of Michigan (11/29/17): pdf
- Pythagoras on the Ice: Hockey Conference, Babson (10/1/16): pdf
- See also talks on **Computers and Mathematics Education**.

References

Papers: https://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/papers.html

- A derivation of James' Pythagorean projection, By The Numbers -- The Newsletter of the SABR Statistical Analysis Committee (**16** (February 2006), no. 1, 17--22). pdf (expanded version: pdf). Chance **(20** (Winter 2007), no. 1, 40-48**).**
- First Order Approximations of the Pythagorean Won-Loss Formula for Predicting MLB Teams Winning Percentages (with Kevin Dayaratna), By The Numbers -- The Newsletter of the SABR Statistical Analysis Committee, pdf (expanded version with appendix proving main result with just one variable calculus: pdf); entire issue of By The Numbers here **(**22 **(2012),** no 1, 15--19)**.**
- The Pythagorean Won-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey (with Kevin Dayaratna), The Hockey Research Journal: A Publication of the Society for International Hockey Research. ((2012/2013), pages 193--209) pdf
- Pythagoras at the Bat (with Taylor Corcoran, Jen Gossels, Victor Luo and Jaclyn Porfilio), book chapter in Social Networks and the Economics of Sports (edited by Panos M. Pardalos and Victor Zamaraev, Springer-Verlag, 2014, pages 89--114). Pdf
- Relieving and Readjusting Pythagoras (with Victor Luo), By The Numbers -- The Newsletter of the SABR Statistical Analysis Committee. **(25** (2015), no. 1, 5-14**)** pdf (older, expanded version: pdf)
- Applications of Improvements to the Pythagorean Won-Loss Expectation in Optimizing Rosters (Alexander Almeida, Kevin Dayaratna, Steven J. Miller and Andrew Yang), book chapter submitted to Artificial Intelligence, Optimization, and Data Sciences in Sports. pdf

Sieze opportunities: Never know where they will lead.

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.

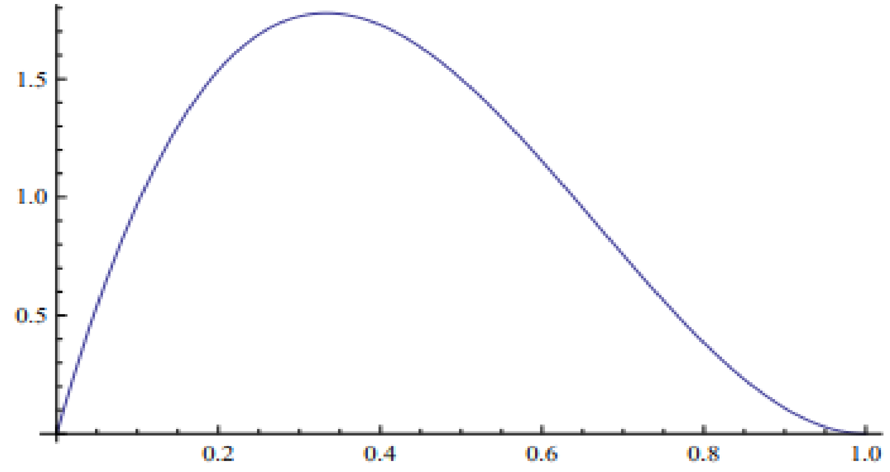
Thank you!

Longer, more technical versions of the talk:

- Pythagoras at the Bat: https://www.youtube.com/watch?v=gFDly_6qOn4 (slides here)
  - Version for 2014 prospective days at Williams here, with a little less calculus (2015 slides here, and 2015 video here).

Appendix with the integration details follows….

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.

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?

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

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

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.

Proof of the Pythagorean Won–Loss Formula

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

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

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

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

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

where we have set

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

Proof of the Pythagorean Won–Loss Formula (cont)

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

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}.

Data Analysis: χ^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).

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.