Pythagoras on the Ice

Steven J. Miller (sjm1@williams.edu, Williams College)
(with Kevin Dayaratna, George Washington University)

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

Babson College, October 1, 2016

Introduction to the Pythagorean Won–Loss Theorem

Numerical Observation: Pythagorean Won–Loss Formula

Parameters

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

Numerical Observation: Pythagorean Won–Loss Formula

Parameters

- $\mathrm{PS_{obs}}$: average number of points scored per game;
- $\mathrm{PA_{obs}}$: average number of points allowed per game;
- γ: some parameter, constant for a sport.

James' Won–Loss Formula (NUMERICAL Observation)

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

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

Pythagorean Formula in Various Sports

James' Won–Loss Formula (NUMERICAL Observation)

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

Football: 2.37 (Schatz 2003).

Basketball: $\approx$14 (Oliver 2004).

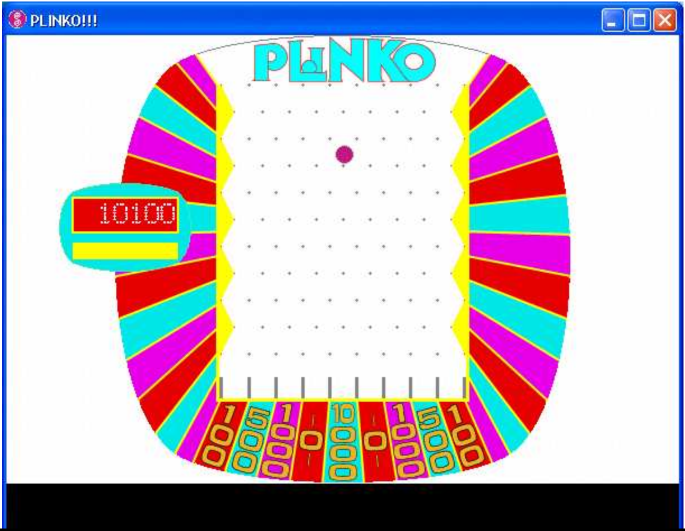
Hockey: 1.927 (Cochran and Blackstock 2009).

Goal is to provide a theoretical justification for hockey.

Applications of the Pythagorean Won–Loss Formula

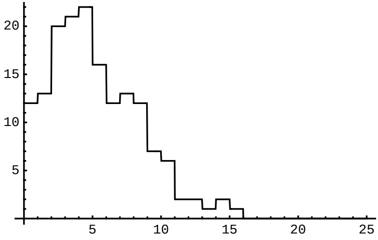
- Extrapolation: use half-way through season to predict a team's performance for rest of season.
- Evaluation: see if consistently over-perform or under-perform.
- Advantage: Other statistics / formulas (run-differential per game); this is easy to use, depends only on two simple numbers for a team.

Probability and Modeling

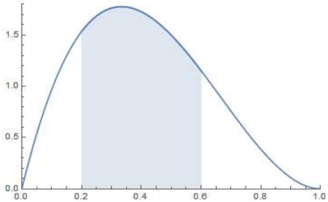


Observed scoring distributions

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



Probability Review

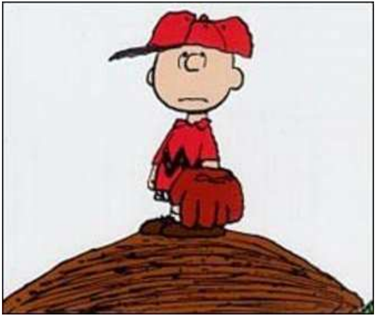


- Let X be random variable with density $p(x)$:
  - ◇ $p(x) \geq 0$;
  - ◇ $\int_{-\infty}^{\infty} p(x) \mathrm{d}x = 1$;
  - ◇ $\text{Prob}\,(a \leq X \leq b) = \int_a^b p(x) \mathrm{d}x$.
- Mean $\mu = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x$.
- Variance $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2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

- Goals Scored and Goals Allowed independent random variables;
- $f_{\mathrm{GS}}(x)$, $g_{\mathrm{GA}}(y)$: probability density functions for goals scored (allowed).

Modeling the Real World (cont)

Possible Model:

- Goals Scored and Goals Allowed independent random variables;
- $f_{\mathrm{GS}}(x)$, $g_{\mathrm{GA}}(y)$: probability density functions for goals scored (allowed).

Won–Loss formula follows from computing

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

Problems with the Model

Reduced to calculating

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

Problems with the Model

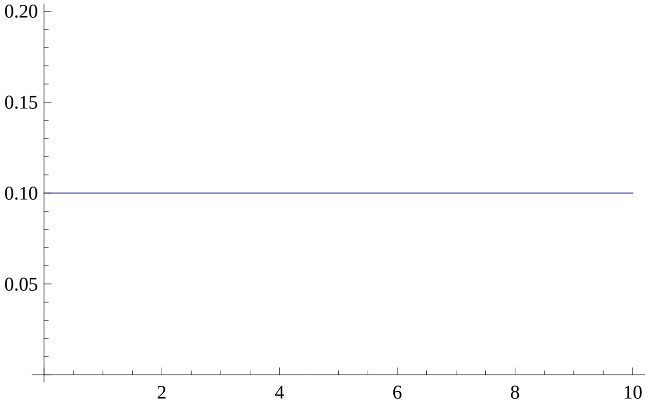
Reduced to calculating

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

Problems with the model:

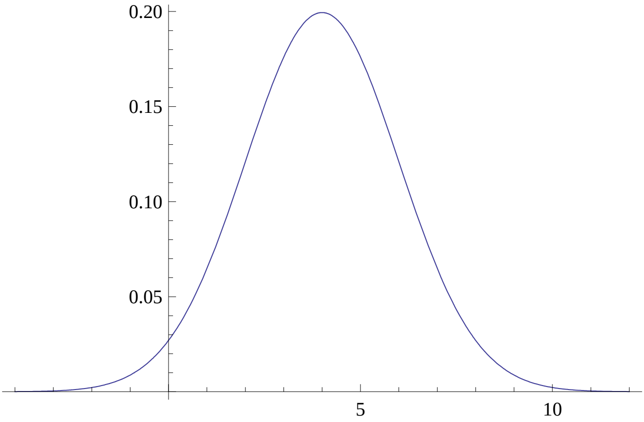
- What are explicit formulas for $f_{\rm GS}$ and $g_{\rm GA}$?
- Are the goals scored and allowed independent random variables?
- Can the integral (or sum) be computed in closed form?

Choices for $f_{\rm GS}$ and $g_{\rm GA}$



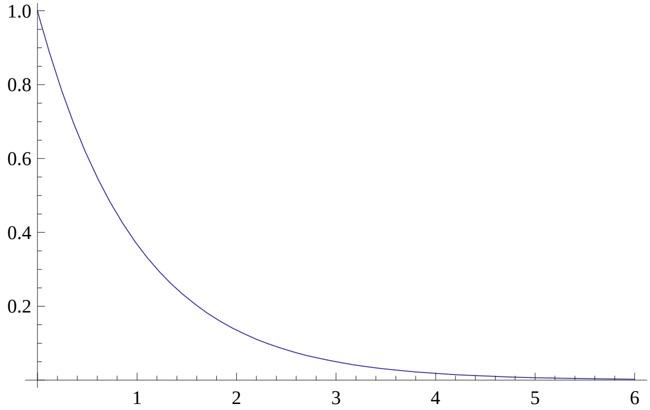
Uniform Distribution on $[0, 10]$.

Choices for f_{GS} and g_{GA}



Normal Distribution: mean 4, standard deviation 2.

Choices for f_{GS} and g_{GA}



Exponential Distribution: e^{-x}.

Three Parameter Weibull

Weibull distribution:

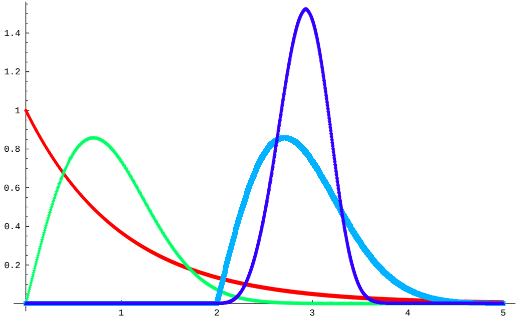
$$f(x;\alpha,\beta,\gamma) = \begin{cases} \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} e^{-((x-\beta)/\alpha)^{\gamma}} & \text{if } x \geq \beta \\ 0 & \text{otherwise.} \end{cases}$$

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

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

Weibull Plots: Parameters (α, β, γ):

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



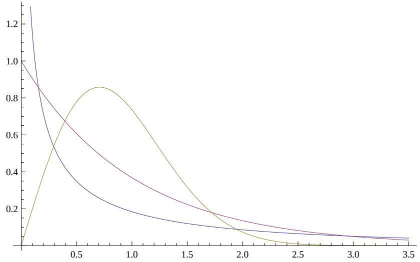
Red:(1, 0, 1) (exponential); Green:(1, 0, 2); Cyan:(1, 2, 2); Blue:(1, 2, 4)

Three Parameter Weibull: Applications

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

Arises in many places, such as survival analysis.

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



The Gamma Distribution and Weibulls

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

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

- 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.

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**

Pythagorean Won–Loss Formula: $\frac{\text{GS}_{\text{obs}}^{\gamma}}{\text{GS}_{\text{obs}}^{\gamma}+\text{GA}_{\text{obs}}^{\gamma}}$

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

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

Pythagorean Won–Loss Formula: $\frac{\mathrm{GS}_{\mathrm{obs}}^{\gamma}}{\mathrm{GS}_{\mathrm{obs}}^{\gamma}+\mathrm{GA}_{\mathrm{obs}}^{\gamma}}$

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

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

Take $\beta = -1/2$ (since goals integers).
$\mathrm{GS} - \beta$ estimates average goals scored, $\mathrm{GA} - \beta$ average goals allowed.
Weibull(α, β, γ) 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 $(\alpha_{\rm GS}, \beta, \gamma)$ and $(\alpha_{\rm GA}, \beta, \gamma)$ respectively. To have means of $\rm GS - \beta$ and $\rm GA - \beta$ our calculations for the means imply

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

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)

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

where we have set

$$
\frac{1}{\alpha^{\gamma}} = \frac{1}{\alpha_{\text{GS}}^{\gamma}} + \frac{1}{\alpha_{\text{GA}}^{\gamma}} = \frac{\alpha_{\text{GS}}^{\gamma} + \alpha_{\text{GA}}^{\gamma}}{\alpha_{\text{GS}}^{\gamma} \alpha_{\text{GA}}^{\gamma}}.
$$

Proof of the Pythagorean Won–Loss Formula (cont)

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

We substitute the relations for $\alpha_{\rm GS}$ and $\alpha_{\rm GA}$ and find that

$$\text{Prob}(X > Y) = \frac{({\rm GS} - \beta)^\gamma}{({\rm GS} - \beta)^\gamma + ({\rm GA} - \beta)^\gamma}.$$

Note ${\rm GS} - \beta$ estimates ${\rm GS}_{\rm obs}$, ${\rm GA} - \beta$ estimates ${\rm GA}_{\rm obs}$.

Analysis

Best Fit Weibulls to Data (Method of Maximum Likelihood)

Data for each teams for the 2008-2009, 2009-2010, and 2010-2011 regular seasons.

Estimated parameters simultaneously via maximum likelihood (MLE).

Performed goodness of fit tests as well as tests of statistical independence.

Will give some representative plots of observed and best fit Weibulls for 2011: Stanley Cup champions (Boston Bruins), their opponent (Vancouver Canucks), the New Jersey Devils (average team: 38 wins, 39 losses and 5 overtime losses), and Edmonton Oilers (worst record in 2011).

Best Fit Weibulls to Data (Method of Maximum Likelihood)

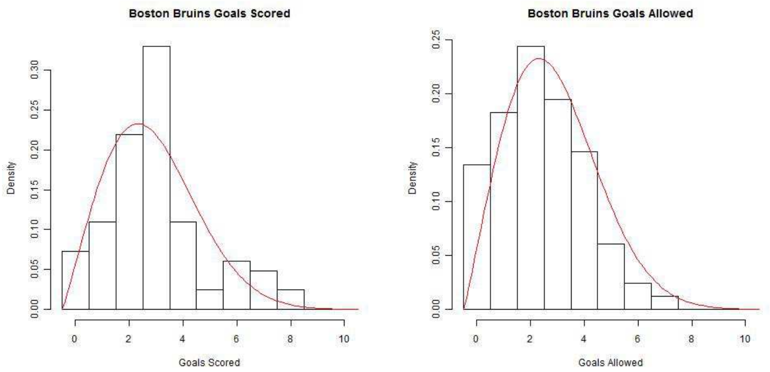


Figure: Boston Bruins: Goals scored and allowed, 2011.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

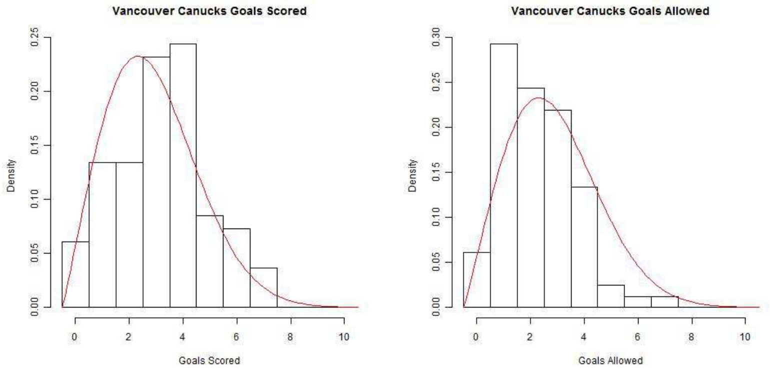


Figure: Vancouver Canucks: Goals scored and allowed, 2011.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

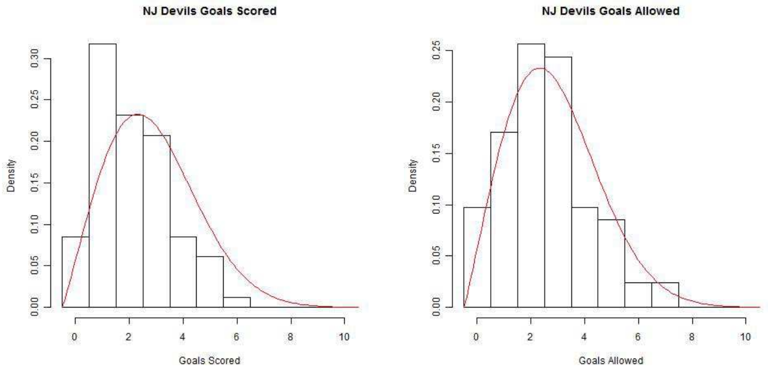


Figure: New Jersey Devils: Goals scored and allowed, 2011.

Best Fit Weibulls to Data (Method of Maximum Likelihood)

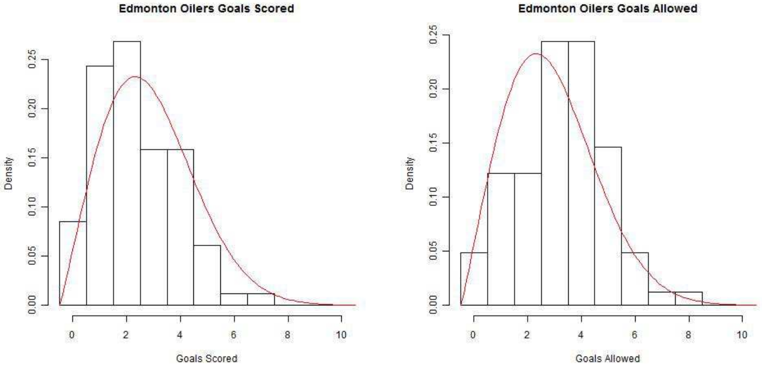


Figure: Edmonton Oilers: Goals scored and allowed, 2011.

Accuracy of Prediction

Team	Won	Lost	$\text{Actual}_{\text{WL}}$	$\text{Pythag}_{\text{WL}}$	Diff	γ	α_{GS}	α_{GA}
Boston Bruins	53	29	0.646	0.639	0.57	2.11	4.31	3.28
NJ Devils	51	31	0.622	0.565	4.71	1.99	3.91	3.43
Washington Capitals	50	32	0.610	0.534	6.25	2.31	4.24	4.00
Carolina Hurricanes	45	37	0.549	0.534	1.22	2.12	3.89	3.65
Pittsburgh Penguins	45	37	0.549	0.551	-0.16	2.24	4.21	3.84
Philadelphia Flyers	44	38	0.537	0.567	-2.46	2.37	4.25	3.79
New York Rangers	43	39	0.524	0.466	4.79	2.02	3.39	3.63
Buffalo Sabres	41	41	0.500	0.531	-2.55	2.17	4.00	3.78
Florida Panthers	41	41	0.500	0.506	-0.46	2.12	3.78	3.74
Montreal Canadiens	41	41	0.500	0.511	-0.86	2.45	4.01	3.94
Ottawa Senators	36	46	0.439	0.454	-1.27	2.27	3.54	3.84
Atlanta Thrashers	35	47	0.427	0.469	-3.46	2.31	4.13	4.36
Toronto Maple Leafs	34	48	0.415	0.442	-2.24	2.27	4.08	4.53
New York Islanders	26	56	0.317	0.339	-1.81	2.25	3.30	4.44
Tampa Bay Lightning	24	58	0.293	0.378	-6.96	2.31	3.50	4.34

Table: 2008-2009 National Hockey League Eastern Conference.

Accuracy of Prediction

Team	Won	Lost	$\text{Actual}_{\text{WL}}$	$\text{Pythag}_{\text{WL}}$	Diff	γ	α_{GS}	α_{GA}
San Jose Sharks	53	29	0.646	0.58	5.45	2.07	4.02	3.44
Detroit Red Wings	51	31	0.622	0.558	5.22	2.29	4.46	4.03
Calgary Flames	46	36	0.561	0.508	4.36	2.11	4.05	3.99
Chicago Blackhawks	46	36	0.561	0.572	-0.87	2.09	4.12	3.59
Vancouver Canucks	45	37	0.549	0.536	1.03	2.08	3.89	3.63
Anaheim Ducks	42	40	0.512	0.51	0.17	2.25	3.91	3.84
Columbus Blue Jackets	41	41	0.500	0.484	1.31	1.99	3.63	3.75
St Louis Blues	41	41	0.500	0.492	0.62	2.16	3.74	3.79
Minnesota Wild	40	42	0.488	0.555	-5.50	2.12	3.62	3.27
Nashville Predators	40	42	0.488	0.462	2.12	1.94	3.48	3.77
Edmonton Oilers	38	44	0.463	0.474	-0.83	2.09	3.79	3.98
Dallas Stars	36	46	0.439	0.474	-2.83	2.09	3.82	4.02
Phoenix Coyotes	36	46	0.439	0.423	1.31	2.00	3.44	4.01
LA Kings	34	48	0.415	0.469	-4.45	1.97	3.47	3.70
Colorado Avalanche	32	50	0.39	0.418	-2.26	2.00	3.39	4.00

Table: 2008-2009 National Hockey League Western Conference.

Accuracy of Prediction

Team	Won	Lost	$\text{Actual}_{\text{WL}}$	$\text{Pythag}_{\text{WL}}$	Diff	γ	α_{GS}	α_{GA}
Washington Capitals	54	28	0.659	0.635	1.93	2.57	4.8	3.87
NJ Devils	48	34	0.585	0.56	2.08	2.1	3.6	3.21
Buffalo Sabres	45	37	0.549	0.571	-1.81	2.21	3.84	3.37
Pittsburgh Penguins	47	35	0.573	0.548	2.08	2.18	4.14	3.79
Ottawa Senators	44	38	0.537	0.471	5.40	2.14	3.65	3.85
Boston Bruins	39	43	0.476	0.515	-3.28	1.99	3.41	3.30
Philadelphia Flyers	41	41	0.50	0.522	-1.82	1.94	3.82	3.65
Montreal Canadiens	39	43	0.476	0.489	-1.13	2.18	3.55	3.62
New York Rangers	38	44	0.463	0.512	-3.96	1.95	3.64	3.55
Atlanta Thrashers	35	47	0.427	0.468	-3.40	2.25	3.82	4.04
Carolina Hurricanes	35	47	0.427	0.471	-3.65	2.29	3.81	4.00
Tampa Bay Lightning	34	48	0.415	0.414	0.04	2.13	3.54	4.16
New York Islanders	34	48	0.415	0.424	-0.75	2.21	3.63	4.18
Florida Panthers	32	50	0.39	0.449	-4.81	1.97	3.47	3.85
Toronto Maple Leafs	30	52	0.366	0.407	-3.41	2.30	3.55	4.18

Table: 2009 - 2010 National Hockey League Eastern Conference.

Accuracy of Prediction

Team	Won	Lost	$\text{Actual}_{\text{WL}}$	$\text{Pythag}_{\text{WL}}$	Diff	γ	α_{GS}	α_{GA}
San Jose Sharks	51	31	0.622	0.579	3.51	2.23	4.14	3.59
Chicago Blackhawks	52	30	0.634	0.587	3.86	2.15	4.16	3.53
Vancouver Canucks	49	33	0.598	0.573	1.97	2.22	4.22	3.69
Phoenix Coyotes	50	32	0.61	0.545	5.33	2.17	3.64	3.35
Detroit Red Wings	44	38	0.537	0.532	0.37	2.15	3.73	3.51
LA Kings	46	36	0.561	0.56	0.12	2.24	3.93	3.54
Nashville Predators	47	35	0.573	0.501	5.95	2.14	3.65	3.65
Colorado Avalanche	43	39	0.524	0.498	2.19	2.25	3.82	3.84
St Louis Blues	40	42	0.488	0.498	-0.84	2.18	3.64	3.65
Calgary Flames	40	42	0.488	0.484	0.30	2.01	3.36	3.47
Anaheim Ducks	39	43	0.476	0.484	-0.66	2.35	3.86	3.97
Dallas Stars	37	45	0.451	0.476	-2.03	2.42	3.85	4.01
Minnesota Wild	38	44	0.463	0.45	1.12	2.50	3.60	3.91
Columbus Blue Jackets	32	50	0.39	0.408	-1.48	2.12	3.50	4.17
Edmonton Oilers	27	55	0.329	0.377	-3.87	2.35	3.55	4.40

Table: 2009 - 2010 National Hockey League Western Conference.

Accuracy of Prediction

Team	Won	Lost	$\text{Actual}_{\text{WL}}$	$\text{Pythag}_{\text{WL}}$	Diff	γ	α_{GS}	α_{GA}
Pittsburgh Penguins	49	33	0.598	0.569	2.34	2.00	3.82	3.32
Washington Capitals	48	34	0.585	0.560	2.09	1.91	3.67	3.23
Philadelphia Flyers	47	35	0.573	0.572	0.12	2.14	4.15	3.62
Boston Bruins	46	36	0.561	0.586	-2.05	1.89	3.91	3.26
Tampa Bay Lightning	46	36	0.561	0.493	5.55	2.00	3.89	3.94
Montreal Canadiens	44	38	0.537	0.504	2.64	1.93	3.49	3.46
New York Rangers	44	38	0.537	0.571	-2.83	1.88	3.79	3.25
Buffalo Sabres	43	39	0.524	0.531	-0.57	2.14	3.93	3.71
Carolina Hurricanes	40	42	0.488	0.503	-1.26	2.17	3.84	3.82
NJ Devils	38	44	0.463	0.426	3.09	1.96	2.95	3.44
Toronto Maple Leafs	37	45	0.451	0.464	-1.04	2.09	3.65	3.91
Atlanta Thrashers	34	48	0.415	0.404	0.90	2.32	3.62	4.28
Ottawa Senators	32	50	0.390	0.386	0.36	2.07	3.20	4.01
Florida Panthers	30	52	0.366	0.442	-6.21	2.31	3.29	3.64
New York Islanders	30	52	0.366	0.455	-7.32	2.14	3.79	4.12

Table: 2010 - 2011 National Hockey League Eastern Conference.

Accuracy of Prediction

Team	Won	Lost	$\text{Actual}_{\text{WL}}$	$\text{Pythag}_{\text{WL}}$	Diff	γ	α_{GS}	α_{GA}
Vancouver Canucks	54	28	0.659	0.644	1.20	2.15	4.13	3.14
San Jose Sharks	48	34	0.585	0.562	1.88	2.21	3.94	3.51
Detroit Red Wings	47	35	0.573	0.541	2.61	2.24	4.16	3.86
Anaheim Ducks	47	35	0.573	0.500	5.96	2.11	3.82	3.82
LA Kings	46	36	0.561	0.526	2.91	1.98	3.52	3.34
Chicago Blackhawks	44	38	0.537	0.558	-1.77	2.29	4.08	3.68
Nashville Predators	44	38	0.537	0.549	-0.98	2.15	3.55	3.24
Phoenix Coyotes	43	39	0.524	0.495	2.44	2.16	3.68	3.71
Dallas Stars	42	40	0.512	0.464	3.94	2.23	3.61	3.85
Calgary Flames	41	41	0.500	0.524	-1.96	2.10	4.00	3.82
Minnesota Wild	39	43	0.476	0.450	2.13	2.03	3.40	3.76
St Louis Blues	38	44	0.463	0.497	-2.78	1.94	3.81	3.83
Columbus Blue Jackets	34	48	0.415	0.408	0.50	2.25	3.49	4.12
Colorado Avalanche	30	52	0.366	0.423	-4.70	2.42	3.83	4.35
Edmonton Oilers	25	57	0.305	0.374	-5.64	2.16	3.29	4.17

Table: 2010 - 2011 National Hockey League Western Conference.

Goodness of Fit: 2008-2009 season

Team	$\chi^2_{\rm GS}$	d.f.	p-value	$\chi^2_{\rm GA}$	d.f.	p-value
Anaheim Ducks	3.46	8	0.902	5.94	9	0.746
Atlanta Thrashers	4.08	9	0.906	4.70	9	0.860
Boston Bruins	4.16	9	0.900	2.75	8	0.949
Buffalo Sabres	4.16	9	0.900	2.75	8	0.949
Calgary Flames	4.45	8	0.815	1.06	8	0.998
Carolina Hurricanes	12.33	9	0.195	4.51	7	0.720
Chicago Blackhawks	7.82	9	0.553	6.73	8	0.566
Colorado Avalanche	9.58	7	0.214	10.54	9	0.308
Columbus Blue Jackets	1.71	8	0.989	11.24	8	0.189
Dallas Stars	7.16	10	0.710	9.77	7	0.202
Detroit Red Wings	13.53	8	0.095	13.16	9	0.155
Edmonton Oilers	12.05	9	0.211	9.40	10	0.494
Florida Panthers	5.78	9	0.761	14.59	8	0.068
LA Kings	11.01	7	0.138	6.78	8	0.561
Minnesota Wild	10.59	8	0.226	8.36	7	0.302
Montreal Canadiens	9.73	7	0.204	4.20	8	0.839
Nashville Predators	8.10	8	0.423	7.52	9	0.583
New York Islanders	9.28	7	0.233	8.82	9	0.454
New York Rangers	9.75	7	0.203	8.64	9	0.471
NJ Devils	7.76	9	0.558	3.58	8	0.893
Ottawa Senators	7.12	7	0.417	4.57	8	0.803
Philadelphia Flyers	8.05	9	0.529	7.17	7	0.411
Phoenix Coyotes	6.87	7	0.442	5.18	8	0.739
Pittsburgh Penguins	7.27	9	0.609	8.80	8	0.359
San Jose Sharks	14.03	8	0.081	12.11	7	0.097
St Louis Blues	8.31	7	0.306	8.52	7	0.289
Tampa Bay Lightning	8.58	8	0.379	9.19	9	0.420
Toronto Maple Leafs	6.63	9	0.676	35.72	8	< 0.001
Vancouver Canucks	8.79	8	0.360	9.07	7	0.248

Goodness of Fit: 2009-2010 season

Team	$\chi^2_{\rm GS}$	d.f.	p-value	$\chi^2_{\rm GA}$	d.f.	p-value
Anaheim Ducks	13.05	8	0.110	1.11	8	0.997
Atlanta Thrashers	3.86	8	0.869	6.74	8	0.565
Boston Bruins	6.76	7	0.454	5.90	8	0.659
Buffalo Sabres	7.68	8	0.465	2.60	7	0.919
Calgary Flames	5.69	7	0.576	11.51	9	0.243
Carolina Hurricanes	8.61	9	0.474	7.06	8	0.531
Chicago Blackhawks	5.09	8	0.747	11.05	8	0.199
Colorado Avalanche	10.60	7	0.157	10.54	9	0.308
Columbus Blue Jackets	9.23	8	0.323	7.33	9	0.603
Dallas Stars	4.64	8	0.795	4.34	7	0.740
Detroit Red Wings	10.41	9	0.318	3.59	7	0.825
Edmonton Oilers	4.01	7	0.779	3.36	8	0.910
Florida Panthers	6.51	8	0.590	9.09	8	0.335
LA Kings	9.53	8	0.299	4.85	8	0.774
Minnesota Wild	1.69	7	0.975	2.61	7	0.918
Montreal Canadiens	8.03	7	0.330	5.90	8	0.659
Nashville Predators	9.01	8	0.342	5.67	8	0.684
New York Islanders	4.07	7	0.772	2.43	8	0.965
New York Rangers	4.44	9	0.880	6.24	9	0.716
NJ Devils	3.38	8	0.909	3.86	6	0.696
Ottawa Senators	3.98	8	0.859	4.49	8	0.811
Philadelphia Flyers	4.53	8	0.807	2.53	9	0.980
Phoenix Coyotes	5.16	7	0.640	9.15	7	0.242
Pittsburgh Penguins	9.16	9	0.423	4.97	8	0.761
San Jose Sharks	8.61	10	0.570	10.25	9	0.331
St Louis Blues	3.63	8	0.889	7.25	8	0.510
Tampa Bay Lightning	3.19	8	0.922	5.18	9	0.818
Toronto Maple Leafs	8.38	7	0.300	8.37	8	0.398
Vancouver Canucks	9.18	9	0.421	5.83	9	0.757

Goodness of Fit: 2010-2011 season

Team	$\chi^2_{\rm GS}$	d.f.	p-value	$\chi^2_{\rm GA}$	d.f.	p-value
Anaheim Ducks	2.13	8	0.977	8.82	9	0.455
Atlanta Thrashers	3.80	8	0.875	9.08	10	0.524
Boston Bruins	17.08	9	0.047	4.43	8	0.816
Buffalo Sabres	3.85	9	0.921	4.68	8	0.791
Calgary Flames	3.84	9	0.921	8.75	8	0.364
Carolina Hurricanes	10.24	8	0.249	16.26	9	0.062
Chicago Blackhawks	3.42	8	0.905	6.86	7	0.444
Colorado Avalanche	6.99	8	0.537	15.46	8	0.051
Columbus Blue Jackets	7.35	7	0.393	8.38	8	0.397
Dallas Stars	7.54	7	0.375	6.80	8	0.559
Detroit Red Wings	4.92	8	0.766	6.88	8	0.550
Edmonton Oilers	3.54	8	0.896	9.96	9	0.354
Florida Panthers	6.98	8	0.539	15.39	6	0.017
LA Kings	10.22	7	0.176	8.34	8	0.401
Minnesota Wild	4.70	7	0.696	6.33	9	0.707
Montreal Canadiens	19.03	9	0.025	5.53	9	0.786
Nashville Predators	8.60	7	0.283	5.22	7	0.633
New York Islanders	3.66	9	0.932	5.54	8	0.699
New York Rangers	5.03	9	0.832	10.23	7	0.176
NJ Devils	4.91	7	0.671	6.94	8	0.544
Ottawa Senators	6.79	7	0.451	8.61	8	0.376
Philadelphia Flyers	4.60	9	0.867	57.94	8	< 0.001
Phoenix Coyotes	7.67	7	0.363	13.45	8	0.097
Pittsburgh Penguins	4.26	9	0.893	6.20	8	0.624
San Jose Sharks	10.26	7	0.174	7.81	7	0.350
St Louis Blues	5.98	9	0.742	8.64	9	0.471
Tampa Bay Lightning	6.55	9	0.684	6.02	9	0.738
Toronto Maple Leafs	12.82	8	0.118	6.66	8	0.573
Vancouver Canucks	7.74	8	0.459	9.18	8	0.327

Goodness of Fit:

Save for Toronto Maple Leafs goals allowed in 2008-2009, p values for 2008-2009 and 2009-2010 below critical thresholds (0.05 and 0.10); supports goals scored and allowed are Weibull.

2010-2011 similar save Boston Bruins goals scored, the Carolina Hurricanes goals allowed, the Colorado Avalanche goals allowed, the Montreal Canadiens goals scored, the Florida Panthers goals allowed, and the Philadelphia Flyers goals allowed.

Bonferroni corrections drop thresholds to 0.00167 and 0.00333, all save Toronto goals allowed in 2008-2009 and the Philadelphia goals allowed in 2010-2011 below.

Goals scored and allowed appear Weibull.

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

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

Team	RS+RA χ2: 20 d.f.	Indep χ2: 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

Conclusions / Future Work

- Find parameters such that Weibulls are good fits.
- Goals scored and allowed per game are statistically independent.
- Pythagorean Won–Loss Formula is a consequence of our model.
- Future Work: Adjust for big leads, overtime, penalty time....

Thank you!

References

References

Taylor Corcoran, Jennifer Gossels, Victor Luo, Steven J. Miller, and Jaclyn Porfilio, *Pythagoras at the Bat*. Book chapter in Social Networks and the Economics of Sports (organized by Victor Zamaraev), book chapter 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.

Kevin Dayaratna and Steven J. Miller, *The Pythagorean Won-Loss Formula and Hockey: A Statistical Justification for Using the Classic Baseball Formula as an Evaluative Tool in Hockey*, The Hockey Research Journal: A Publication of the Society for International Hockey Research. ((2012/2013), pages 193–209). http://web.williams.edu/Mathematics/sjmiller/public_html/math/papers/DayaratnaMiller_2012%20-%20SIHR%20Extended%20VersionArxivFebruary2013.pdf.

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

Steven J. Miller, *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

Steven J. Miller,*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_Paper.pdf

Appendices

Appendix I: 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 I: 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 II: 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 $\text{Bin}(k)$ of the Weibull with parameters $\alpha, -.5, \gamma$. The sample likelihood function $L(\alpha_{\rm RS}, \alpha_{\rm RA}, -.5, \gamma)$ is

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

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

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.