

# 1 Introduction

The Pythagorean Won-Loss formula has been around for decades. Postulated by Bill James in the early 1980s, the Pythagorean Won-Loss formula indicates the percentage of games (or winning percentage, WP) a baseball team should have won at a particular point in a season given as a function of average runs scored (RS) and average runs allowed (RA):

$$WP = \frac{RS^\gamma}{RS^\gamma + RA^\gamma}. \quad (1)$$

James initially postulated the exponent to be 2 (hence the name “Pythagorean” from a sum of squares). Empirical observation later suggested that  $\gamma \approx 1.8$  was more appropriate.

For decades, the Pythagorean Won-Loss formula gave a good indication of the percentage of games a baseball team should have won at a particular point in a season. Baseball statisticians could use the resulting percentage to predict a team’s won-loss record at the end of the season. Sabermetricians could also use the percentage to comment on a team’s level of over-performance/under-performance as well as the value of adding certain players to their lineup.

Until recently, however, the Pythagorean Won-Loss formula had no statistical verification. Miller (2007) addressed this issue by assuming that the runs scored and runs allowed follow separate independent continuous Weibull distributions. The key assumption is the independence of runs scored and allowed; the Weibull assumption was made to ensure the resulting integration led to a simple, closed-form expression. Upon making these assumptions, he was able to derive James’ formula in the form of the probability that the runs a particular team scores is greater than the runs it allows. He estimated this model via least squares and maximum likelihood estimation on 2004 American League data and determined that the appropriate value of  $\gamma$  was indeed around 1.8, consistent with empirical observation.

A few researchers have applied Bill James’ model to other sports. For example, Schatz (2003) applied the model to football and determined that an appropriate value of  $\gamma$  is around 2.37. Oliver (2004) did the same for basketball and determined that an appropriate value for  $\gamma$  is around 14. Rosenfeld et al. (2010) drew upon this research and used the Pythagorean Won-Loss formula to predict overtime wins in baseball, basketball, and football.

Cochran and Blackstock (2009) applied the Pythagorean Won-Loss formula to hockey. They used least squares estimation to estimate James’ model as well as several modifications. They found that James’ original Pythagorean Won-Loss

formula, with a value of around 1.927, is just as accurate as the results produced by more complex models.

No one to date, however, has provided a theoretical verification from first principles for applying the Pythagorean Won-Loss formula (or its generalizations) to any sport other than baseball. We do so for hockey. Specifically, we make the same assumptions that Miller (2007) made for baseball and find that the Pythagorean Won-Loss formula applies just as well to hockey than it does to baseball. Our results provide a theoretical explanation as to why the Pythagorean Won-Loss formula used in Cochran and Blackstock (2009), which is so parsimonious compared to other models used in that research, fits hockey data so well. Cochran and Blackstock's paper discusses several models generalizing and extending Pythagoras, though no theoretical justification for any model is given. These complex models have more degrees of freedom, arising from the inclusion of interaction terms. As their analysis shows that there is minimal gain in predictive power from these more complicated formulas, they recommend using the standard Pythagorean formula due to its simplicity.

In the following two sections, we provide theoretical justification from a statistical perspective for applying the simple and elegant Pythagorean Won-Loss formula to hockey. Afterwards, we examine a linear predictor that has been used in baseball (Jones and Tappin 2005). We apply this model to hockey, showing that this predictor is simply just the linearized version of the Pythagorean formula. Not only does our first-order approximation provide theoretical justification for the linear predictor, but also allows us to gain an interpretation of the best-fit slope parameter in terms of the average offense in the league and the Pythagorean exponent.

Our work is organized as follows. We first show that the distributions of goals scored and goals allowed can be treated as independent distributions. We then discuss our model and estimation results; in particular, we sketch the derivation of the Pythagorean Won-Loss formula from our assumptions about the independence and distribution of goals scored and allowed. Afterwards, we present tests that examine our model's statistical validity and prove that the linear predictor model is simply the linearization of the Pythagorean model. Finally, we conclude by summarizing our findings and discussing potential avenues of future research.

## **2 Statistical Independence of Goals Scored and Goals Allowed**

Naively, one would think that the distributions of goals scored and goals allowed are dependent. For example, if a team has a big lead, the coaching staff might

change which players they have on the ice or use up remaining time on the clock and not worry about scoring. On the other hand, if a team is trailing toward the end of a game the staff may pull out their goalie when they need a last minute goal to increase their probability of scoring.

Some of the above arguments also apply to other sports, including baseball. Recent research in sabermetrics (Ciccolella 2006; Miller 2007), however, suggests that the distributions of runs scored and runs allowed can be considered independent. One would do well to ask whether a similar argument is true about hockey, namely whether the distributions of goals scored and goals allowed are statistically independent, and if so, what distribution they follow. We did so by performing non-parametric statistical tests of Kendall's Tau and Spearman's Rho (Hogg et al 2005).

Given that each team plays 82 games over the course of a regular season and if  $GS_i$  is the number of goals scored in game  $i$  and  $GA_i$  is the number of goals allowed in game  $i$ , we say the pair  $i$  and  $j$  are concordant if

$$\text{sgn}((GS_i - GS_j)(GA_i - GA_j)) = 1$$

and discordant if

$$\text{sgn}((GS_i - GS_j)(GA_i - GA_j)) = -1,$$

with  $\text{sgn}(x) = 1$  if  $x > 0$ ,  $0$  if  $x = 0$  and  $-1$  otherwise. Based on these definitions, our Kendall's Tau test statistic for testing the null hypothesis that the distributions of goals scored and goals allowed are independent is

$$\tau = \frac{\#\{\text{concordant pairs}\} - \#\{\text{discordant pairs}\}}{\frac{1}{2} \cdot 82 \cdot 81}. \quad (2)$$

We also computed the classic Spearman's Rho coefficient to also test our null hypothesis. If we let  $R(GS_i)$  denote the rank of  $GS_i$  amongst  $GS_1, \dots, GS_{82}$  and  $R(GA_i)$  denote the rank of  $GA_i$  amongst  $GA_1, \dots, GA_{82}$ , then Spearman's rho is defined as follows:

$$\rho = \frac{\sum_{i=1}^{82} (R(GS_i) - \frac{n+1}{2})(R(GA_i) - \frac{n+1}{2})}{82(82^2 - 1)/12}. \quad (3)$$

We record the results of these tests in Tables 1 and 2.

Team	$\tau$ (08-09)	$p$ (08-09)	$\tau$ (09-10)	$p$ (09-10)	$\tau$ (10-11)	$p$ (10-11)
Anaheim Ducks	0.075	0.313	-0.105	0.156	0.008	0.900
Atlanta Thrashers	-0.023	0.761	0.027	0.712	-0.061	0.411
Boston Bruins	0.126	0.089	-0.047	0.527	-0.108	0.145
Buffalo Sabres	-0.123	0.097	-0.063	0.395	0.083	0.262
Calgary Flames	-0.056	0.453	-0.055	0.456	-0.031	0.679
Carolina Hurricanes	-0.129	0.084	-0.165	0.027	-0.112	0.132
Chicago Blackhawks	-0.048	0.522	0.060	0.422	-0.056	0.453
Colorado Avalanche	0.036	0.627	-0.042	0.573	-0.047	0.527
Columbus Blue Jackets	0.042	0.570	0.063	0.399	-0.090	0.226
Dallas Stars	0.049	0.506	-0.089	0.232	-0.126	0.090
Detroit Red Wings	0.006	0.932	-0.009	0.906	-0.003	0.968
Edmonton Oilers	-0.042	0.576	0.017	0.824	-0.217	0.004
Florida Panthers	-0.105	0.156	0.003	0.971	-0.082	0.270
LA Kings	0.073	0.329	-0.017	0.817	-0.008	0.910
Minnesota Wild	-0.046	0.533	-0.025	0.737	-0.207	0.005
Montreal Canadiens	-0.079	0.290	-0.006	0.935	-0.171	0.021
Nashville Predators	0.109	0.143	0.078	0.296	-0.132	0.075
New York Islanders	-0.019	0.798	-0.056	0.449	0.017	0.817
New York Rangers	0.015	0.836	-0.097	0.191	-0.007	0.923
NJ Devils	-0.089	0.229	-0.096	0.197	-0.125	0.092
Ottawa Senators	0.034	0.647	-0.126	0.090	-0.088	0.235
Philadelphia Flyers	-0.038	0.607	-0.023	0.752	-0.097	0.191
Phoenix Coyotes	-0.008	0.910	-0.072	0.335	-0.006	0.932
Pittsburgh Penguins	-0.014	0.846	-0.041	0.584	-0.059	0.425
San Jose Sharks	0.083	0.262	-0.125	0.093	-0.047	0.525
St Louis Blues	0.032	0.665	-0.032	0.665	0.030	0.691
Tampa Bay Lightning	0.100	0.178	-0.065	0.379	-0.026	0.727
Toronto Maple Leafs	0.031	0.673	-0.043	0.559	-0.037	0.615
Vancouver Canucks	0.047	0.527	-0.130	0.081	-0.088	0.237
Washington Capitals	0.025	0.737	0.023	0.761	-0.036	0.632
Max	0.126		0.078		0.083	
Min	-0.129		-0.165		-0.217	

Table 1: Kendall's Tau and  $p$ -values for the National Hockey League, Seasons 2008-2009, 2009-2010 and 2010-2011.

Team	$\rho$ (08-09)	$p$ (08-09)	$\rho$ (09-10)	$p$ (09-10)	$\rho$ (10-11)	$p$ (10-11)
Anaheim Ducks	0.145	0.193	-0.123	0.272	0.030	0.789
Atlanta Thrashers	-0.007	0.950	0.084	0.452	-0.052	0.644
Boston Bruins	0.214	0.054	-0.032	0.772	-0.124	0.266
Buffalo Sabres	-0.164	0.141	-0.051	0.651	0.163	0.144
Calgary Flames	-0.057	0.614	-0.034	0.763	-0.015	0.896
Carolina Hurricanes	-0.152	0.172	-0.217	0.050	-0.116	0.300
Chicago Blackhawks	-0.041	0.717	0.117	0.296	-0.048	0.671
Colorado Avalanche	0.087	0.436	-0.026	0.819	-0.018	0.875
Columbus Blue Jackets	0.090	0.420	0.131	0.240	-0.086	0.442
Dallas Stars	0.101	0.367	-0.102	0.363	-0.153	0.169
Detroit Red Wings	0.047	0.673	0.028	0.805	0.025	0.820
Edmonton Oilers	-0.024	0.833	0.058	0.607	-0.290	0.008
Florida Panthers	-0.126	0.259	0.036	0.748	-0.071	0.526
LA Kings	0.139	0.212	0.016	0.889	0.021	0.853
Minnesota Wild	-0.037	0.738	0.001	0.992	-0.276	0.012
Montreal Canadiens	-0.085	0.447	0.030	0.791	-0.223	0.044
Nashville Predators	0.186	0.094	0.140	0.210	-0.166	0.136
New York Islanders	0.006	0.959	-0.049	0.663	0.051	0.652
New York Rangers	0.062	0.579	-0.120	0.283	0.019	0.867
NJ Devils	-0.104	0.353	-0.117	0.297	-0.166	0.136
Ottawa Senators	0.079	0.481	-0.177	0.111	-0.100	0.372
Philadelphia Flyers	-0.026	0.820	-0.012	0.913	-0.090	0.424
Phoenix Coyotes	0.012	0.918	-0.064	0.569	0.033	0.765
Pittsburgh Penguins	0.001	0.996	-0.030	0.791	-0.048	0.666
San Jose Sharks	0.153	0.169	-0.146	0.189	-0.032	0.774
St Louis Blues	0.083	0.458	-0.015	0.895	0.074	0.511
Tampa Bay Lightning	0.182	0.102	-0.063	0.573	0.001	0.996
Toronto Maple Leafs	0.075	0.505	-0.027	0.812	-0.034	0.765
Vancouver Canucks	0.093	0.404	-0.175	0.116	-0.092	0.413
Washington Capitals	0.062	0.583	0.071	0.528	-0.010	0.928
Max	0.214		0.140		0.163	
Min	-0.164		-0.217		-0.290	

Table 2: Spearman's Rho and  $p$ -values for the National Hockey League, Seasons 2008-2009, 2009-2010 and 2010-2011.

Assuming commonly-accepted critical thresholds of 0.05 and 0.10, instituting Bonferroni corrections reduce these thresholds to 0.00167 and 0.00333. As values of  $\tau$  and  $\rho$  are almost always below 0.2 in magnitude and never exceed 0.29, even without making use of Bonferroni corrections, our results indicate that dependence between goals scored and goals allowed, if present in any manner, is quite weak. Thus, for all intents and purposes, our results suggest that it is safe to treat the distributions of goals scored and goals allowed as independent distributions. Intuitively, the effects we described at the beginning of the section probably contribute to the slight dependence in goals scored and goals allowed. These effects, however, essentially wash out, enabling one to be able to treat goals scored and goals allowed as if they come from independent distributions, similar to the findings in Ciccolella (2006) and Miller (2007) for baseball.

### 3 Model Development

In this section, we prove that if goals scored and allowed are drawn from independent translated Weibull distributions then the Pythagorean Won-Loss formula holds. The use of a translated Weibull density makes estimation, goodness of fit tests, and statistical independence tests easier. This is a rich family of distributions, and is able to model many one-hump phenomena (i.e., processes whose probability is strictly increasing to a maximum and then rapidly decreasing to zero afterwards).

Thus, we assume that the distribution of the number of goals a hockey team scores (GS) and the number of goals it allows (GA) each follow independent translated two-parameter Weibull distributions with the following probability density functions:

$$\begin{aligned} f(x; \alpha_{GS}, \gamma) &= \frac{\gamma}{\alpha_{GS}} \left( \frac{x+.5}{\alpha_{GS}} \right)^{\gamma-1} \exp \left( - \left( \frac{x+.5}{\alpha_{GS}} \right)^\gamma \right) I(x > -.5) \\ f(y; \alpha_{GA}, \gamma) &= \frac{\gamma}{\alpha_{GA}} \left( \frac{y+.5}{\alpha_{GA}} \right)^{\gamma-1} \exp \left( - \left( \frac{y+.5}{\alpha_{GA}} \right)^\gamma \right) I(y > -.5), \end{aligned} \quad (4)$$

where  $I(u > -.5)$  is the indicator random variable which is 1 if its argument  $u$  exceeds  $-.5$  and 0 otherwise. We specifically translated the Weibull densities by a factor of 0.5 to ensure that our data (the integer representing the score) is at the center of the bins for our chi-squared goodness of fit tests. Continuous distributions are used to facilitate computation by transforming sums into integrals, and facilitate getting a simple, closed-form expression such as the Pythagorean formula. Of course, continuous distributions do not truly represent reality as baseball and hockey teams only score integral values of points; however, the Weibull is a flexible distribution and by appropriately choosing its parameters, it can fit many data sets.

Miller (2007) showed the Pythagorean Won-Loss formula can be derived by computing the probability that the number of goals a team scores is greater than the number of goals it allows. We sketch the argument below:

$$\begin{aligned}
\text{Pythag}_{\text{WL}} &= \text{Prob}(X > Y) \\
&= \int_{-.5}^{\infty} \int_{-.5}^x f(x; \alpha_{\text{GS}}, \gamma) f(y; \alpha_{\text{GA}}, \gamma) dy dx \\
&= \int_{-.5}^{\infty} \int_{-.5}^x \frac{\gamma}{\alpha_{\text{GS}}} \left( \frac{x+.5}{\alpha_{\text{GS}}} \right)^{\gamma-1} \exp\left(-\left(\frac{x+.5}{\alpha_{\text{GS}}}\right)^{\gamma}\right) \\
&\quad \cdot \frac{\gamma}{\alpha_{\text{GA}}} \left( \frac{y+.5}{\alpha_{\text{GA}}} \right)^{\gamma-1} \exp\left(-\left(\frac{y+.5}{\alpha_{\text{GA}}}\right)^{\gamma}\right) dy dx \\
&= \int_0^{\infty} \frac{\gamma}{\alpha_{\text{GS}}} \left( \frac{x}{\alpha_{\text{GS}}} \right)^{\gamma-1} \exp\left(-\left(\frac{x}{\alpha_{\text{GS}}}\right)^{\gamma}\right) \\
&\quad \cdot \left[ \int_0^x \frac{\gamma}{\alpha_{\text{GA}}} \left( \frac{y}{\alpha_{\text{GA}}} \right)^{\gamma-1} \exp\left(-\left(\frac{y}{\alpha_{\text{GA}}}\right)^{\gamma}\right) dy \right] dx \\
&= \int_0^{\infty} \frac{\gamma}{\alpha_{\text{GS}}} \left( \frac{x}{\alpha_{\text{GS}}} \right)^{\gamma-1} \exp\left(-\left(\frac{x}{\alpha_{\text{GS}}}\right)^{\gamma}\right) \left[ 1 - \exp\left(-\left(\frac{x}{\alpha_{\text{GA}}}\right)^{\gamma}\right) \right] dx \\
&= 1 - \int_0^{\infty} \frac{\gamma}{\alpha_{\text{GS}}} \left( \frac{x}{\alpha_{\text{GS}}} \right)^{\gamma-1} \exp\left(-\left(\frac{x}{\alpha_{\text{GS}}}\right)^{\gamma} - \left(\frac{x}{\alpha_{\text{GA}}}\right)^{\gamma}\right), \quad (5)
\end{aligned}$$

as the first integral is one as it is a probability density. The remaining integral is almost a Weibull density with parameters  $\gamma$  and  $\alpha$ , with

$$\frac{1}{\alpha^{\gamma}} = \frac{1}{\alpha_{\text{GS}}^{\gamma}} + \frac{1}{\alpha_{\text{GA}}^{\gamma}}, \quad (6)$$

differing only by a constant factor. (The combination of  $\alpha_{\text{GS}}$  and  $\alpha_{\text{GA}}$  to get  $\alpha$  occurs in a variety of other problems, such as resistors in parallel or the reduced mass in the two-body problem in mechanics.) This leads to

$$\begin{aligned}
\text{Pythag}_{\text{WL}} &= 1 - \int_{x=0}^{\infty} \frac{\gamma}{\alpha_{\text{GS}}} \left( \frac{x}{\alpha_{\text{GS}}} \right)^{\gamma-1} \exp\left(-\frac{x}{\alpha}\right)^{\gamma} dx \\
&= 1 - \frac{\alpha^{\gamma}}{\alpha_{\text{GS}}^{\gamma}} \int_0^{\infty} \frac{\gamma}{\alpha} \left( \frac{x}{\alpha} \right)^{\gamma-1} \exp\left(-\frac{x}{\alpha}\right)^{\gamma} dx \\
&= 1 - \frac{\alpha^{\gamma}}{\alpha_{\text{GS}}^{\gamma}}
\end{aligned}$$

$$= \frac{\alpha_{\text{GS}}^\gamma}{\alpha_{\text{GS}}^\gamma + \alpha_{\text{GA}}^\gamma}, \quad (7)$$

where the last equality follows from simple algebra. The mean goals scored (GS) and mean goals allowed (GA) for our translated Weibull density are

$$\text{GS} = \alpha_{\text{GS}}\Gamma(1 + \gamma^{-1}) - .5, \quad \text{GA} = \alpha_{\text{GA}}\Gamma(1 + \gamma^{-1}) - .5, \quad (8)$$

where

$$\Gamma(s) = \int_0^\infty e^{-x} x^{s-1} dx \quad (\text{Re}(s) > 0) \quad (9)$$

is the Gamma function (see Miller (2007) for the calculation). Therefore, after a bit more algebra, we find

$$\text{Pythag}_{\text{WL}} = \frac{(\text{GS} + .5)^\gamma}{(\text{GS} + .5)^\gamma + (\text{GA} + .5)^\gamma}. \quad (10)$$

Estimation of the parameters of our Weibull densities enables us to compute the corresponding Pythagorean expectations.

## 4 Data and Results

We compiled data (goals scored and goals allowed) from ESPN.com for each of the 30 NHL teams over the course of the 2008-2009, 2009-2010, and 2010-2011 regular seasons. We estimated our parameters simultaneously via maximum likelihood (MLE). We also performed goodness of fit tests as well as tests of statistical independence. Figures 1 through 4 are some representative plots of the observed data and the best fit Weibulls for 2011. The complete plots are available from the authors. We have chosen the Stanley Cup champions, the Boston Bruins, their opponent, the Vancouver Canucks, the New Jersey Devils (whose 38 wins, 39 losses and 5 overtime losses makes them close to an average team), and the Edmonton Oilers, who had the worst record in 2011.

Our results from our maximum likelihood estimation, our computation of each of the 30 NHL team's Pythagorean won loss formula ( $\text{Pythag}_{\text{WL}}$ ), and our computed difference between the observed number of games won and the expected number of games won (Diff), are below in Tables 3 through 8:

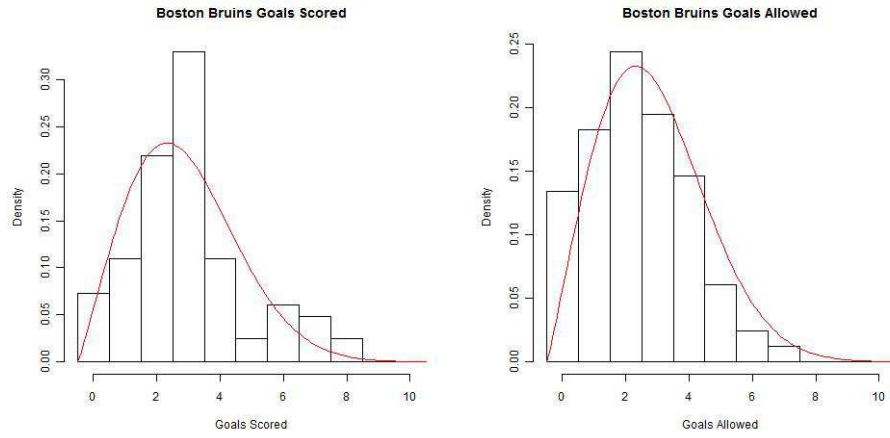


Figure 1: Boston Bruins: Goals scored and allowed, 2011.

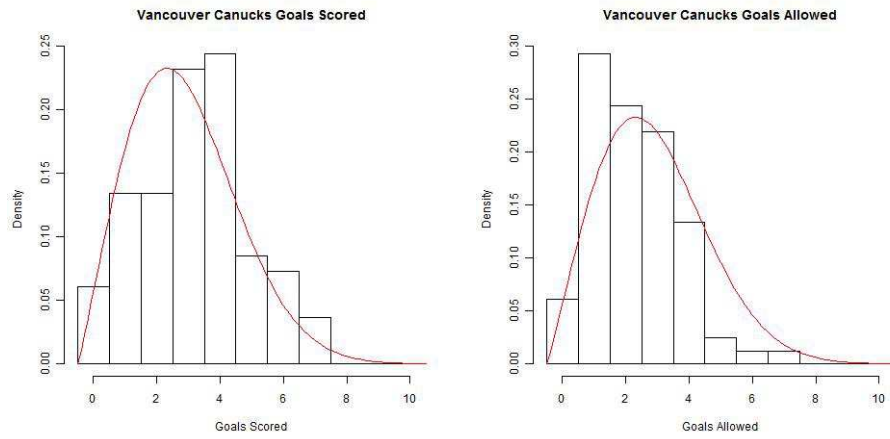


Figure 2: Vancouver Canucks: Goals scored and allowed, 2011.

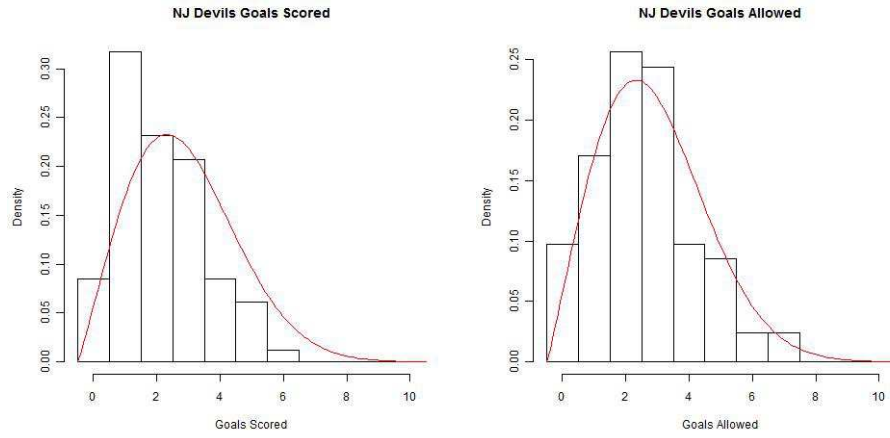


Figure 3: New Jersey Devils: Goals scored and allowed, 2011.

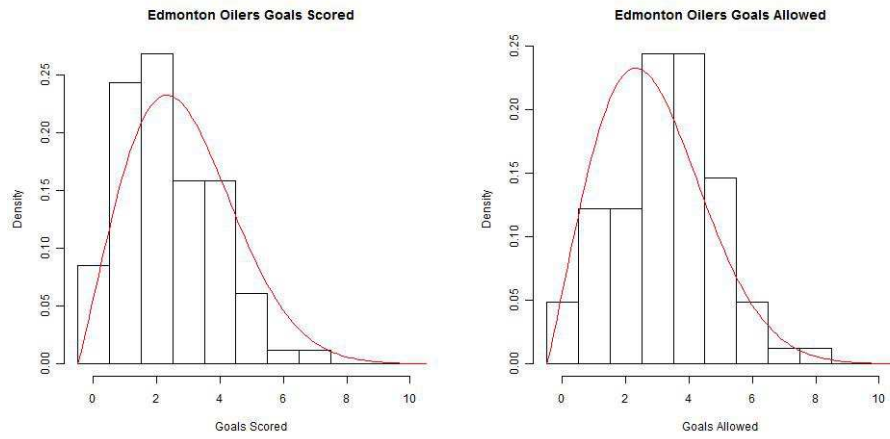


Figure 4: Edmonton Oilers: Goals scored and allowed, 2011.

Team	Won	Lost	Actual <sub>WL</sub>	Pythag <sub>WL</sub>	Diff	$\gamma$	$\alpha_{GS}$	$\alpha_{GA}$
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 3: 2008-2009 National Hockey League Eastern Conference.

Team	Won	Lost	Actual <sub>WL</sub>	Pythag <sub>WL</sub>	Diff	$\gamma$	$\alpha_{GS}$	$\alpha_{GA}$
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 4: 2008-2009 National Hockey League Western Conference.

Team	Won	Lost	Actual <sub>WL</sub>	Pythag <sub>WL</sub>	Diff	$\gamma$	$\alpha_{GS}$	$\alpha_{GA}$
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 5: 2009 - 2010 National Hockey League Eastern Conference.

Team	Won	Lost	Actual <sub>WL</sub>	Pythag <sub>WL</sub>	Diff	$\gamma$	$\alpha_{GS}$	$\alpha_{GA}$
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 6: 2009 - 2010 National Hockey League Western Conference.

Team	Won	Lost	Actual <sub>WL</sub>	Pythag <sub>WL</sub>	Diff	$\gamma$	$\alpha_{GS}$	$\alpha_{GA}$
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 7: 2010 - 2011 National Hockey League Eastern Conference.

The maximum likelihood estimated value of  $\gamma$  is almost always slightly above 2, averaging 2.15 for the 2008-2009 season (standard deviation 0.13), 2.20 for the 2009-2010 season (standard deviation 0.15), and 2.11 (standard deviation 0.14) for the 2010-2011 season, which is reasonably close to the estimates computed in Cochran and Blackstock (2009). Our results also indicate that many of the top teams, including the Washington Capitals, NJ Devils, San Jose Sharks, and the Chicago Blackhawks and Vancouver Canucks performed better than expected over the seasons examined.

## 5 Model Testing: Goodness of Fit

We performed chi-squared goodness of fit tests to determine how well the Weibull densities conform to the true distributions of goals scored and goals allowed. For

Team	Won	Lost	Actual <sub>WL</sub>	Pythag <sub>WL</sub>	Diff	$\gamma$	$\alpha_{GS}$	$\alpha_{GA}$
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 8: 2010 - 2011 National Hockey League Western Conference.

most teams, we tested the joint distributions by splitting our data based on the following bins:

$$[-.5, .5], [.5, 1.5], [1.5, 2.5], \dots, [8.5, 9.5], [9.5, \infty).$$

These bins are appropriate to ensure that our data occurs in the center of our bins (this is always true, as the goals scored and allowed must be non-negative integers). The number of bins was determined on a team by team basis according to each team's distribution of goals scored and goals allowed.

To perform our test, we computed the following statistics (Shao, 1999):

$$\begin{aligned} \chi_{GS}^2 &= \sum_{k=1}^{\#bins} \frac{\left( GS_{obs} - \#games \int_{a_k}^{a_{k+1}} f(x; \alpha_{GS}, \gamma) dx \right)^2}{\#games \int_{a_k}^{a_{k+1}} f(x; \alpha_{GS}, \gamma) dx} \\ \chi_{GA}^2 &= \sum_{k=1}^{\#bins} \frac{\left( GA_{obs} - \#games \int_{a_k}^{a_{k+1}} f(y; \alpha_{GA}, \gamma) dy \right)^2}{\#games \int_{a_k}^{a_{k+1}} f(y; \alpha_{GA}, \gamma) dy}, \end{aligned} \quad (11)$$

where  $GS_{obs}(k)$  and  $GA_{obs}(k)$  are the number of entries into the bin with left endpoint  $a_k$  and right endpoint  $a_{k+1}$  and

$$\frac{\#games \int_{a_k}^{a_{k+1}} f(x; \alpha_{GS}, \gamma) dx}{\#games \int_{a_k}^{a_{k+1}} f(y; \alpha_{GA}, \gamma) dy} \quad (12)$$

(with there being 82 games in a hockey season) is the expected proportion of the number of games a team should have in  $[a_k, a_{k+1})$  if the goals scored and allowed are drawn from independent Weibulls.

We tested the null hypothesis that the distributions of goals scored and goals allowed for each particular team follow Weibull distributions. If our data actually conforms to the Weibull distribution, then the chi-square statistics should follow a chi-squared distribution with  $\#bins - 1$  degrees of freedom. We have the capacity to reject this null hypothesis at significance level  $\alpha$  if the chi-square value is greater than or equal to the  $(1 - \alpha)^{\text{th}}$  quantile of a chi-squared distribution with  $\#bins - 1$  degrees of freedom (Shao 2009). See Tables 9 to 11 for the results.

With the exception of the Toronto Maple Leafs Goals allowed in 2008-2009, our  $p$  values for the 2008-2009 and 2009-2010 seasons are always well above commonly accepted critical thresholds (0.05 and 0.10), so we can say with reasonable certainty that the distributions of goals scored and goals allowed for our teams actually follow Weibull distributions. For the 2010-2011 season, our  $p$  values are almost always above these thresholds as well, with the exception of the 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. After instituting Bonferroni corrections, however, our critical thresholds drop to 0.00167 and 0.00333 respectively, and all our distributions except the Toronto Maple Leafs Goals Allowed in 2008-2009 and the Philadelphia Flyers Goals Allowed in 2010-2011 fall below our necessary critical thresholds. As a result, we can conclude with reasonable certainty that virtually all of our distributions of goals scored and goals allowed for each of our 30 teams adheres to a Weibull distribution.

## 6 Derivation of Linear Predictor

Jones and Tappin (2005) have used the following predictor for a team's winning percentage:

$$WP = .500 + \beta(PS - PA); \quad (13)$$

Team	$\chi^2_{GS}$	d.f.	<i>p</i> -value	$\chi^2_{GA}$	d.f.	<i>p</i> -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
Washington Capitals	11.13	7	0.133	11.51	7	0.118

Table 9: Results of chi-squared goodness of fit tests for the 2008-2009 season.

Team	$\chi^2_{GS}$	d.f.	<i>p</i> -value	$\chi^2_{GA}$	d.f.	<i>p</i> -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
Washington Capitals	8.49	8	0.387	5.85	7	0.558

Table 10: Results of chi-squared goodness of fit tests for the 2009-2010 season.

Team	$\chi^2_{GS}$	d.f.	<i>p</i> -value	$\chi^2_{GA}$	d.f.	<i>p</i> -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
Washington Capitals	10.29	7	0.173	6.93	7	0.436

Table 11: Results of chi-squared goodness of fit tests for the 2010-2011 season.

here WP is the team's winning percentage, PS is the average points scored (goals in hockey, runs in baseball, et cetera) and PA is the average points allowed. Notice that if  $PS = PA$  then the team is predicted to win half its games. Typically  $\beta$  is a small number. Thus for observed values of PS and PA we do not need to worry about the above expression exceeding 1.000 or falling below .000. For example, in baseball in 2010 runs scored ranged from 513 to 859 and runs allowed from 581 to 845; in hockey in 2010-2011 the ranges were 174 to 262 for goals scored and 185 to 288 for goals allowed. For these ranges, the winning percentages are all 'reasonable'.

We now show how the prediction in (13) follows from the Pythagorean formula. We assume there is some exponent  $\gamma$  such that

$$WP = \frac{PS^\gamma}{PS^\gamma + PA^\gamma}. \quad (14)$$

We provide two statistical 'proofs' of the linear formula. These require different backgrounds; our goal in including both is to make the arguments accessible to a wide audience. The first assumes multivariable calculus, and is quite fast and simple; the second uses only single variable calculus, but takes significantly longer and requires some clever algebraic manipulations and several approximations. After the proofs, we compare our predicted best fit values with observed data.

## 6.1 First Proof (Assuming Multivariable Calculus)

In this subsection we assume the reader is familiar with multivariable calculus. Recall the second order Taylor series expansion of a function  $f(x, y)$  about the point  $(a, b)$  is

$$\begin{aligned} f(x, y) = & f(a, b) + \frac{\partial f}{\partial x} \Big|_{(a, b)} (x - a) + \frac{\partial f}{\partial y} \Big|_{(a, b)} (y - b) \\ & + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \Big|_{(a, b)} (x - a)^2 + \frac{\partial^2 f}{\partial x \partial y} \Big|_{(a, b)} (x - a)(y - b) + \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \Big|_{(a, b)} (y - b)^2 \\ & + \text{higher order terms;} \end{aligned} \quad (15)$$

here the higher order terms involve products of  $(x - a)$  and  $(y - b)$  to the third and higher powers. The tangent plane approximation, which means keeping just the constant and linear terms, is

$$f(x, y) = f(a, b) + \frac{\partial f}{\partial x} \Big|_{(a, b)} (x - a) + \frac{\partial f}{\partial y} \Big|_{(a, b)} (y - b). \quad (16)$$

Let  $P_{\text{ave}}$  denote the average number of points scored in the league. We let

$$f(x, y) = \frac{x^\gamma}{x^\gamma + y^\gamma}. \quad (17)$$

We now expand about the point  $(a, b) = (P_{\text{ave}}, P_{\text{ave}})$ , with  $x = \text{PS}$  and  $y = \text{PA}$ , so

$$\begin{aligned} f(P_{\text{ave}}, P_{\text{ave}}) &= .500 \\ \frac{\partial f}{\partial x} &= \frac{\gamma x^{\gamma-1} y^\gamma}{(x^\gamma + y^\gamma)^2} \Rightarrow \frac{\partial f}{\partial x}(P_{\text{ave}}, P_{\text{ave}}) = \frac{\gamma}{4P_{\text{ave}}} \\ \frac{\partial f}{\partial y} &= -\frac{\gamma x^\gamma y^{\gamma-1}}{(x^\gamma + y^\gamma)^2} \Rightarrow \frac{\partial f}{\partial y}(P_{\text{ave}}, P_{\text{ave}}) = -\frac{\gamma}{4P_{\text{ave}}}. \end{aligned} \quad (18)$$

Noting that the predicted winning percentage is  $f(\text{PS}, \text{PA})$ , we see that the first order, multivariate Taylor series expansion gives

$$\text{WP} \approx .500 + \frac{\gamma}{4P_{\text{ave}}}(\text{PS} - P_{\text{ave}}) - \frac{\gamma}{4P_{\text{ave}}}(\text{PA} - P_{\text{ave}}) = .500 + \frac{\gamma}{4P_{\text{ave}}}(\text{PS} - \text{PA}). \quad (19)$$

## 6.2 Second Proof (Single Variable Calculus)

Remember that we assume there is some exponent  $\gamma$  such that

$$\text{WP} = \frac{\text{PS}^\gamma}{\text{PS}^\gamma + \text{PA}^\gamma}. \quad (20)$$

We multiply the right hand side by  $(1/\text{PS}^\gamma)/(1/\text{PS}^\gamma)$  and find

$$\text{WP} = \frac{1}{1 + \frac{\text{PA}^\gamma}{\text{PS}^\gamma}} = \left(1 + \left(\frac{\text{PA}}{\text{PS}}\right)^\gamma\right)^{-1}. \quad (21)$$

There are many ways to attack the algebra. In the analysis below, we constantly replace complicated functions by their linear approximations (i.e., their first order Taylor series). Inspired by the logit model, let  $u_{\text{PS}} = \ln(\text{PS})$  and  $u_{\text{PA}} = \ln(\text{PA})$ , so  $\text{PS} = \exp(u_{\text{PS}})$  and  $\text{PA} = \exp(u_{\text{PA}})$ . Then  $(\text{PA}/\text{PS})^\gamma = (\exp(u_{\text{PA}})/\exp(u_{\text{PS}}))^\gamma$ , which is  $\exp(-\gamma(u_{\text{PS}} - u_{\text{PA}}))$ . We thus have

$$\text{WP} = (1 + \exp(-\gamma(u_{\text{PS}} - u_{\text{PA}})))^{-1}. \quad (22)$$

We now make some approximations. While there will obviously be some loss in predictive power from these choices, it will lead to a very simple, final expression. As we expect PS and PA to be of comparable size, the difference of their logarithms  $(u_{\text{PS}} - u_{\text{PA}})$  should be small; for example, if  $\text{PS} = 800$  and  $\text{PA} = 600$  (reasonable numbers in baseball), one finds  $u_{\text{PS}} - u_{\text{PA}} \approx .288$ ; we would get the same answer if we used 400 and 300 respectively, reasonable numbers for hockey. We Taylor expand the exponential function, noting

$$\exp(x) = 1 + x + \text{higher order terms}. \quad (23)$$

We drop these higher order terms as we have  $x = -\gamma(u_{\text{PS}} - u_{\text{PA}})$ . In other words, we are only keeping the constant and linear terms; note that if we only kept the

constant term, there would be no dependence on points scored or allowed! We thus find

$$\text{WP} \approx (1 + 1 - \gamma(u_{\text{PS}} - u_{\text{PA}}))^{-1} = \frac{1}{2} \left(1 - \frac{\gamma}{2}(u_{\text{PS}} - u_{\text{PA}})\right)^{-1}. \quad (24)$$

We now expand using the geometric series formula, which says

$$\frac{1}{1-r} = 1 + r + r^2 + r^3 + \dots \quad (25)$$

for  $|r| < 1$ . We take  $r = \frac{\gamma}{2}(u_{\text{PS}} - u_{\text{PA}})$ , and again only keep the constant and linear term,  $1 + r$ , yielding

$$\text{WP} \approx \frac{1}{2} \left(1 + \frac{\gamma}{2}(u_{\text{PS}} - u_{\text{PA}})\right) = \frac{1}{2} + \frac{\gamma}{4}(u_{\text{PS}} - u_{\text{PA}}). \quad (26)$$

We need to do a little more analysis to obtain a formula that is linear in  $\text{PS} - \text{PA}$ . Recalling that the  $u$ 's are the logarithms of the points, we have

$$u_{\text{PS}} - u_{\text{PA}} = \ln \text{PS} - \ln \text{PA} = \ln \frac{\text{PS}}{\text{PA}} = \ln \frac{\text{PA} + \text{PS} - \text{PA}}{\text{PA}} = \ln \left(1 + \frac{\text{PS} - \text{PA}}{\text{PA}}\right). \quad (27)$$

We now Taylor expand the logarithm. We have  $\log(1+x) = x$  plus higher order terms. For us,  $x = \frac{\text{PS} - \text{PA}}{\text{PA}}$  is much less than 1, and thus we again only keep up to the linear term. Substituting yields

$$\text{WP} \approx \frac{1}{2} + \frac{\gamma}{4} \frac{\text{PS} - \text{PA}}{\text{PA}}. \quad (28)$$

We make one last simplification. To first order, the  $\text{PA}$  in the denominator can be replaced by  $\text{P}_{\text{ave}}$ , the average number of points scored in the league. We have (finally) reached our linear approximation,

$$\text{WP} \approx .500 + \frac{\gamma}{4\text{P}_{\text{ave}}}(\text{PS} - \text{PA}). \quad (29)$$

Thus, in the simple linear approximation model, the 'interesting' coefficient should be approximately  $\gamma/4\text{P}_{\text{ave}}$ .

### 6.3 Comparison with Previous Seasons

Linear approximations such as this have been used for many years by a variety of people. For example, Michael Jones and Linda Tappin (see ScienceDaily, March 30, 2004, and Jones and Tappin 2005) used this linear model for baseball. They

wrote  $WP = .500 + \beta(PS - PA)$ , and by looking at the seasonal data from 1969 to 2003 found the best values of  $\beta$  ranged from .00053 to .00078, with an average value of .00065. Taking their average value of .00065 and using  $\gamma = 1.81$  leads to a predicted value of 696 runs scored per team per year, or about 4.3 runs per game. Conversely, using the average number of runs scored in 2010 by American League teams (721) and their average value of  $\beta$ , one gets a prediction of 1.88 for  $\gamma$ .

The above analysis provides some support for the linear model. In particular, the slope is no longer a mysterious quantity, but is naturally related to the exponent and average scoring in the league. Our analysis of NHL data provides further support. Using the method of least squares, we find that for the 2008-2009, 2009-2010, and 2010-2011 seasons,  $\beta$  is 0.183, 0.185, and 0.173 respectively. (Actually, we fit the more general model  $WP = \alpha + \beta(PS - PA)$ , and found the least-squares values of  $\alpha$  to be .500, .500 and .501, in excellent agreement with what we would expect.) As a result, the estimated value of  $\gamma$  for each of these seasons is 2.132, 2.105, 1.930, which is in beautiful accordance with our results from maximum likelihood estimation of our Weibull densities.

## 7 Conclusion and Future Research

Our results provide statistical justification for applying the simple and elegant Pythagorean Won-Loss formula to a sport other than baseball, namely hockey. We estimated  $\gamma$  to be between 2.1 and 2.3, which is reasonably close to Cochran and Blackstock's 2009 findings. Furthermore, our goodness of fit and statistical independence tests are stronger than Miller's findings for baseball (Miller 2007). When Miller performed similar chi-squared tests on 2004 American League data, Weibull densities were deemed to be inappropriate to describe one team's distribution of runs scored and runs allowed. Our results are more robust over a larger array of teams, illustrating that the Pythagorean Won-Loss formula is just as applicable, if not more applicable, to hockey than to baseball. Finally, we provided a theoretical justification for an existing linear model and gave an interpretation of the slope in terms of  $\gamma$  and the average scoring within the league.

There are a number of potential avenues of future research that we hope this work will encourage:

1. Future research should go on to examine the statistical appropriateness of applying the Pythagorean Won-Loss formula to other sports, such as basketball and soccer. Researchers could then use the formula as a basis for comparing teams of different eras and understanding the effects of hiring well-known

coaches or superstars, as well as the expected gains resulting from mid-season signings.

2. One could also perform a more micro analysis as suggested in Miller (2007) to incorporate lower order effects. Baseball has several natural candidates, ranging from park effects to the presence or absence of a designated hitter depending on where the game is played. Similarly, there are natural candidates to investigate in hockey. The first is rink effects, ranging from having the home crowd to slight differences in the rinks (see Weiner 2009 for some of the differences between rinks, even though they all have the same dimensions for the ice). Other items include power plays (which means both how well a team does on power plays, as well as how likely they or the opponent is to provide an opportunity), 'meaningless' goals late in the game (such as goals scored by the leading team when the trailing team pulls its goalie), and overtime scoring (and its relation to classifying the game as a win or a loss). As our model already does a great job explaining the data, it is likely that these are lower order effects that mostly wash out, but it would still be interesting to see the size of their effects.
3. Almost surely professional sports players do not discuss how to ensure their scoring conforms to a Weibull distribution. Regardless, we used such a model here as doing so leads to a tractable double integral that can be solved in closed form. One of primary advantages of the Pythagorean formula is the simplicity of the resulting statistic; however, in an age of powerful and ever-present computing power, the need for a simple statistic is lessened. Consequently, there are several other approaches one may take:
  - (a) One possibility is to look at linear combinations of Weibull distributions. The resulting fit to the data cannot be worse, as our situation is just the special case of one Weibull distribution. One would have a sum of individually tractable integrals, all yielding closed-form expressions.
  - (b) Along these lines, one could replace a Weibull distribution with a linear combination of a Weibull distribution and a point mass at zero. Such a model allows one to accommodate for the probability of being shut out and have another density to model scoring. A similar idea is used via a quasi-geometric model in (Glass and Lowry, 2008) to model scoring in baseball games.
  - (c) The scoring data for both baseball and hockey is well-modeled by a one-hump distribution, namely the probability initially rises to a maximum and then continuously falls. Instead of using a Weibull distribution, one could

use a Beta distribution instead, where the density becomes

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} I(0 \leq x \leq 1), \quad (30)$$

with  $a, b > 0$  our shape parameters,  $\Gamma$  the Gamma function (which is a generalization of the factorial function, with  $\Gamma(n+1) = n!$  for  $n$  a non-negative integer) and  $I(0 \leq x \leq 1)$  is the indicator function which is 1 for  $x$  between 0 and 1 and 0 otherwise. For many choices of  $a$  and  $b$  we find that a Beta distribution captures the general shape of the observed scoring data; however, while closed-form expressions exist for the mean and the variance of the Beta distribution in terms of its parameters, for general choice of the parameters we do not have a nice closed form expression for the needed double integral. Thus, if Beta distributions were to be used, one would be reduced to numerical approximations to find the dependence of the winning percentage on the parameters of the teams.

## References

American Institute Of Physics, *Streamlining The 'Pythagorean Theorem Of Baseball'* (2004, March 30), ScienceDaily. Retrieved August 21, 2011, from <http://www.sciencedaily.com/releases/2004/03/040330090259.htm>.

Casella G. and Berger R., *Statistical Inference*, Second Edition, Duxbury Advanced Series, 2002.

Ciccolella, R. *Are Runs Scored and Runs Allowed Independent*, *By the Numbers* **16** (2006), no. 1, 11–15.

Cochran, J. and Blackstock, R. *Pythagoras and the National Hockey League*, *Journal of Quantitative Analysis in Sports* (2009), **5** (2), Article 11.

ESPN.com, [http://espn.go.com/nhl/standings/\\_/type/expanded/year/2010](http://espn.go.com/nhl/standings/_/type/expanded/year/2010).

Glass, D. and Lowry, P. J. *Quasigeometric Distributions and Extra Inning Baseball Games*, *Mathematics Magazine* (2008), **81** (2), 127-137.

Hogg, R.V.; Craig, A.T.; and McKean, J.W., *Introduction to Mathematical Statistics*, Sixth Edition, Prentice Hall Inc, 2004.

James, B. *The Bill James Abstract*, self-published, 1979.

James, B. *The Bill James Abstract*, self-published, 1980.

James, B. *The Bill James Abstract*, self-published, 1981.

James, B. *The Bill James Abstract*, Ballantine Books, 1982.

James, B. *The Bill James Abstract*, Ballantine Books, 1983.

Jones, M. A. and Tappin, L. A., *The Pythagorean Theorem of Baseball and Alternative Models*, *The UMAP Journal* 26.2 (2005), 12 pages.

Miller, S.J., *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 available at <http://arxiv.org/abs/math/0509698>.

Oliver, D. *Basketball On Paper*, Potomac Books, 2004.

Rosenfeld, J.W.; Fisher, J.I.; Adler, D; and Morris, C. *Predicting Overtime with the Pythagorean Formula*, *Journal of Quantitative Analysis in Sports* (2010), **6**, no. 2.

Ryder, A. *Win Probabilities: a tour through win probability models for hockey* (2004), *Hockey Analytics* [http://www.hockeyanalytics.com/Research\\_files/Win\\_Probabilities.pdf](http://www.hockeyanalytics.com/Research_files/Win_Probabilities.pdf).

Schatz, A. *Pythagoras on the Gridiron*. *Football Outsiders*, July 14, 2003. <http://www.footballoutsiders.com/stat-analysis/2003/pythagoras-gridiron>.

Shao, J. *Mathematical Statistics*, Springer, 1999.

Weiner, E. *Not every 200 foot by 85 foot NHL rink is the same*, *Off the Wall*, October 9, 2009 (5:00pm). <http://www.nhl.com/ice/news.htm?id=501626>.