EXTENDING JAMES' PYTHAGOREAN FORMULA FOR HEAD-TO-HEAD MATCHUPS

RICHARD CLEARY, JAKE JEFFRIES, CAMERON MILLER, STEVEN J. MILLER, JAMES MURRAY, AND NICK SKIERA

ABSTRACT. Bill James' Pythagorean formula has a long history of estimating a team's winning percentage to within a few games each season, using just the total runs scored and allowed. Its importance is highlighted as it is one of the few non-historical statistics seen on expanded standings pages. For long seasons, small effects often average out, but this is not the case in the playoffs. We suggest a simple, easily implemented generalization that incorporates the average runs scored and allowed of the home and away teams to assign a probability for each to win: rescale the quantities relative to the league average. We demonstrate the value of this approach by looking at the results of all playoff series from 2001 through 2019. Taking into account the strengths of each team in each match-up, we predict the higher seeded team should win 80.18 series and lose 68.82, agreeing phenomenally well with the actual results of 80 and 69.

CONTENTS

1. INTRODUCTION

Bill James' Pythagorean formula has for years been a staple of sabermetric analysis. Using just the total runs-scored and runs-allowed of a team[1], it predicts their winning percentage for the season to be

$$\frac{\text{RunsScored}^{\gamma}}{\text{RunsScored}^{\gamma} + \text{RunsAllowed}^{\gamma}},$$

where γ is currently taken to be 1.87 by Baseball Reference[2]. Miller [Mi] provided a theoretical basis for this by showing it is a consequence of assuming runs scored and

2020 *Mathematics Subject Classification.* 62P99.

Key words and phrases. James' Pythagorean Won-Lost Formula, Weibull Distribution, Head-to-Head Matchups.

[1]Using average runs scored and allowed instead would give the same prediction.

[2]See https://www.baseball-reference.com/about/war_explained_runs_to_wins.shtml, though 1.83 is also often used; the value is different for different eras, from the deadball era to the steroid years

allowed are drawn from the three-parameter Weibull distribution:

$$\text{Prob}(X \in [a,b]) \;=\; \int_a^b f(x;\alpha,\beta,\gamma)dx, \tag{1.1}$$

where

$$f(x;\alpha,\beta,\gamma) \;=\; \frac{\gamma}{\alpha}\left(\frac{x-\beta}{\alpha}\right)^{\gamma-1} \exp\left(\left(\frac{x-\beta}{\alpha}\right)^{\gamma}\right). \tag{1.2}$$

One of its many uses is to assess the value of adding players to a roster, as we can use runs created statistics [A, C] to see how different signings would affect run production and defense (see also [Ad, BZ, Br, Go, De, Ke, Min, Pe, Vo] for work on how teams estimate performance). There is now an extensive literature on the formula, in baseball and in other sports; see for example [ADMY, CGLMP, DM, HJM, Hu, Ja, LM, Mi]. In addition to providing theoretical justifications, many of the above papers explore potential improvements to the formula, such as using linear combinations of Weibulls, and adjusting the run production to not count late inning runs in blowout games.

As none of these adjustments noticeably improves on the predictive power of the simpler formula, we pursued a new approach: incorporating the data of the two teams playing to extend the formula to predict head-to-head match-ups, and not just estimate a season's performance. This appears to be a promising area of research. We are not aware of many attempts to solve this problem. One is James' $\log 5$ formula (see [Bi] for a discussion of when it works and when it doesn't): if team A wins $p\%$ of their games and team B wins $q\%$, then

$$\text{Prob}(A \text{ beats } B) \;=\; \frac{p(1-q)}{p(1-q)+(1-p)q} \;=\; \frac{p-pq}{p+q-2pq}. \tag{1.3}$$

We can justify it through a simple conditional probability model: we assume A (resp. B) has a 'good' day with probability p (resp. q), and a 'bad' day with probability $1-p$ (resp. $1-q$). We keep playing until the two teams have different outcomes, leading to the conditional formula above.[3] Another is Heumann's [He] improvement to Pythagorean Wins. He notes that the standard Pythagorean formula cannot be the full story, as the sum of the expected number of wins across all teams does not equal games played in general for leagues with more than two teams. His adjustment takes into account head-to-head matchups, leading to the correct number of wins for the league; unfortunately this adjustment cannot be used for two teams that have never previously met.

We adjust the Pythagorean Formula to use data from both teams:

- home team $\text{RS}_\text{h}, \text{RA}_\text{h}$,
- away team $\text{RS}_\text{a}, \text{RA}_\text{a}$,
- league average runs scored per game is R,

[3]There are four outcomes, but the good-good and bad-bad, with probabilities pq and $(1-p)(1-q)$ are not accessible, and thus the probability A wins is the probability it has a good day and B has a bad day, $p(1-q)$, divided by the probability that exactly one has a good day, $p(1-q)+(1-p)q$.

- adjusted home numbers:
  $\mathrm{RS}_{\mathrm{h,adj}} = \mathrm{RS}_{\mathrm{h}}(\mathrm{RA}_{\mathrm{a}}/\mathrm{R})$,
  $\mathrm{RA}_{\mathrm{h,adj}} = \mathrm{RA}_{\mathrm{h}}(\mathrm{RS}_{\mathrm{a}}/\mathrm{R})$:

We introduce a new statistic, the **extended Pythagorean probability that the home team wins**:

$$\mathrm{Prob}(\text{Home Team Wins}) \;=\; \frac{\mathrm{RS}_{\mathrm{h,adj}}^{\gamma}}{\mathrm{RS}_{\mathrm{h,adj}}^{\gamma} + \mathrm{RA}_{\mathrm{h,adj}}^{\gamma}}. \tag{1.4}$$

In the next section we discuss our reasoning for proposing this model, and report on comparing its predictions to actual seasons. It is worth remarking that instead of (1.4) we can use an equivalent formula that does not involve the league average, as those factors are removed by multiplying by R^{γ}/R^{γ}:

$$\mathrm{Prob}(\text{Home Team Wins}) \;=\; \frac{(\mathrm{RS}_{\mathrm{h}}\mathrm{RA}_{\mathrm{a}})^{\gamma}}{(\mathrm{RS}_{\mathrm{h}}\mathrm{RA}_{\mathrm{a}})^{\gamma} + (\mathrm{RA}_{\mathrm{h}}\mathrm{RS}_{\mathrm{a}})^{\gamma}}. \tag{1.5}$$

We prefer (1.4) as the quantities have a natural interpretation, as we explain in the next section.

2. Extended Pythagorean Formula

Billy Beane has perhaps the most famous quote about the value of sabermetrics over short periods:

> *My shit doesn't work in the playoffs. My job is to get us to the playoffs. What happens after that is fucking luck.*

Over the course of a season, effects often average out and estimators often do a very good job; however, it is precisely the short playoff window where the games matter most. For example, the Pythagorean Formula is a good predictor against an average opponent, but there are no average opponents in the playoffs.[4] It makes no sense to try to predict a team's probability of winning looking only at their data; we need to incorporate whom they are playing.

We were inspired to rescale the runs-scored and runs-allowed by other analyses which took into account park effects, see for example [Sch1, Sch2]. Note that there is no change if the opponent is league average. If for example the away team allows 10% more runs than an average defense, we adjust the home team's runs-scored by increasing it by 10%. Of course, instead of a multiplicative rescaling we could do an additive shift. We prefer a multiplicative rescaling rather than an additive shift in runs scored because of the 'game to game' nature of our modeling. The approaches would be nearly equivalent in a regular season implementation, as can be seen by a Taylor series expansions, similar to [DM], where the authors proved that the linear Pythagorean formula follows from a multivariable Taylor series expansion of the common formula about the league average.

Note that (1.4) has many desirable properties. In addition to reducing to the standard formula when the opponent's offense and defense are league average, we have the adjusted probability that the home team wins plus the adjusted probability that the away

[4]Unless of course a team has been plagued with several late season injuries, and is thus very different than their early season iteration.

team wins equals 1. To prove this, it is cleaner to use (1.5):

$$\begin{aligned} P_{h,a} &= \text{Prob(Home Team Wins)} + \text{Prob(Away Team Wins)} \\ &= \frac{(\mathrm{RS_h RA_a})^\gamma}{(\mathrm{RS_h RA_a})^\gamma + (\mathrm{RA_h RS_a})^\gamma} + \frac{(\mathrm{RS_a RA_h})^\gamma}{(\mathrm{RS_a RA_h})^\gamma + (\mathrm{RA_a RS_h})^\gamma}. \end{aligned} \tag{2.1}$$

To simplify displaying the algebra, set (σ for score, α for allow)

$$\sigma_t := \mathrm{RS_t}, \quad \alpha_t := \mathrm{RA_t}, \quad t \in \{h, a\}. \tag{2.2}$$

Then we have the sum of the probabilities of each winning is

$$\begin{aligned} P_{h,a} &= \frac{\sigma_h \alpha_a}{\sigma_h \alpha_a + \alpha_h \sigma_a} + \frac{\sigma_a \alpha_h}{\sigma_a \alpha_h + \alpha_a \sigma_h} \\ &= \frac{\sigma_h \alpha_a (\sigma_a \alpha_h + \alpha_a \sigma_h) + \sigma_a \alpha_h (\sigma_h \alpha_a + \alpha_h \sigma_a)}{(\sigma_h \alpha_a + \alpha_h \sigma_a)(\sigma_a \alpha_h + \alpha_a \sigma_h)} \\ &= \frac{\alpha_h^2 \sigma_a^2 + 2\alpha_a \alpha_h \sigma_a \sigma_h + \alpha_a^2 \sigma_h^2}{\alpha_h^2 \sigma_a^2 + 2\alpha_a \alpha_h \sigma_a \sigma_h + \alpha_a^2 \sigma_h^2} = 1. \end{aligned} \tag{2.3}$$

Thus, similar to Heumann's model, our approach also yields a league average winning percentage of .500; however, our approach is easier as we can make our prediction without having the data broken down team-by-team, and as remarked we can also handle the case of two teams which have never met, or only met a few times.

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

$$\begin{aligned} P_{h,a}(b) &= \frac{(\mathrm{RS_h RA_a^b})^\gamma}{(\mathrm{RS_h RA_a^b})^\gamma + (\mathrm{RA_h RS_a^b})^\gamma} + \frac{(\mathrm{RS_a RA_h^b})^\gamma}{(\mathrm{RS_a RA_h^b})^\gamma + (\mathrm{RA_a RS_h^b})^\gamma} \\ &= \frac{\sigma_h \alpha_a^b}{\sigma_h \alpha_a^b + \alpha_h \sigma_a^b} + \frac{\sigma_a \alpha_h^b}{\sigma_a \alpha_h^b + \alpha_a \sigma_h^b} \\ &= \frac{\sigma_h \alpha_a^b (\sigma_a \alpha_h^b + \alpha_a \sigma_h^b) + \sigma_a \alpha_h^b (\sigma_h \alpha_a^b + \alpha_h \sigma_a^b)}{(\sigma_h \alpha_a^b + \alpha_h \sigma_a^b)(\sigma_a \alpha_h^b + \alpha_a \sigma_h^b)} \\ &= \frac{\sigma_h \sigma_a \alpha_h^b \alpha_a^b + \sigma_h^{b+1} \alpha_a^{b+1} + \sigma_h \sigma_a \alpha_h^b \alpha_a^b + \sigma_a^{b+1} \alpha_h^{b+1}}{\sigma_h \sigma_a \alpha_h^b \alpha_a^b + \sigma_h^{b+1} \alpha_a^{b+1} + \sigma_h^b \sigma_a^b \alpha_h \alpha_a + \sigma_a^{b+1} \alpha_h^{b+1}}, \end{aligned} \tag{2.4}$$

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

3. Comparison with Postseasons

The true value of a model is not whether or not it is elegant or beautiful, but rather whether or not it does a good job explaining what has occurred and predicting what will follow. To that end, we looked at the results of all playoff series from 2001 through

2019. We chose this window as it provides a large number of head-to-head match-ups of strong teams, and ends before major rule changes which can affect the exponent which is known to be different in different eras. Thus we stop before the Covid season, before the new rules for extra innings, and so on.

To compare predictions to outcomes we decided on the following metric: if our model predicts the home team to win the series against a given opponent $p\%$ of the time we credit the home team with p wins and the away team with $1-p$. We then sum all the credits given to the home teams and to the away teams, and compare those two numbers with the observed results, remembering an adage from Ernest Rutherford:

> *If your experiment needs statistics, you ought to have done a better experiment.*

Before reporting on the results, we remark on how we computed the probability the home team won; for the purpose of our analysis the home team is designated the team with the higher seed and thus home for the first game, which is not necessarily the stronger team. We use our adjusted formula to calculate the probability the home team beats the away team in a given game: $p = p(\mathrm{RS}_h, \mathrm{RA}_h, \mathrm{RS}_a, \mathrm{RA}_a)$; for simplicity we do not take into account home field advantage, which changes throughout most of the series. We then compute the probability that the home team reaches $n+1$ wins before the away team in a best of $2n+1$ series:

$$\text{Prob(Home Team Wins)} \;=\; \sum_{m=n+1}^{2n+1} \binom{m-1}{n-1} p^{n+1}(1-p)^{m-(n+1)}; \tag{3.1}$$

this is because if the home team wins in $m \in \{n+1, \ldots, 2n+1\}$ games then they win exactly $m-1$ of the first games and win the m^{th} game. Note the longer the series, the greater the likelihood the better team will win; see Figure 1 (as well as [CM] for applications of this analysis to compare success in sports such as basketball and football, which have very different playoff structures).

We used data from Baseball-Reference[5] for the playoff match-ups, results, and team statistics; we do not need to describe any detailed statistical test as to whether or not our prediction aligned well with reality!

- Observed: Higher seed wins 80.00 and loses 69.00.
- Log-5 Method: Higher seed wins 83.19 and loses 65.81.
- Extended Pythagorean: home wins 80.18 and loses 68.82.

Given that we are dealing with integer quantities, our predictions round perfectly to the results, and Rutherford would be proud!

4. Conclusion and Future Work

Our purpose was to adjust the Pythagorean Formula to be of use for head-to-head match-ups; while our analysis was focused on baseball, similar comparisons can be done for other sports. We obtain a very easy to implement formula using the four key pieces of information: each team's runs scored and allowed (while we use the league average in describing the method, it is not needed in the actual computations). With

[5] https://www.baseball-reference.com/

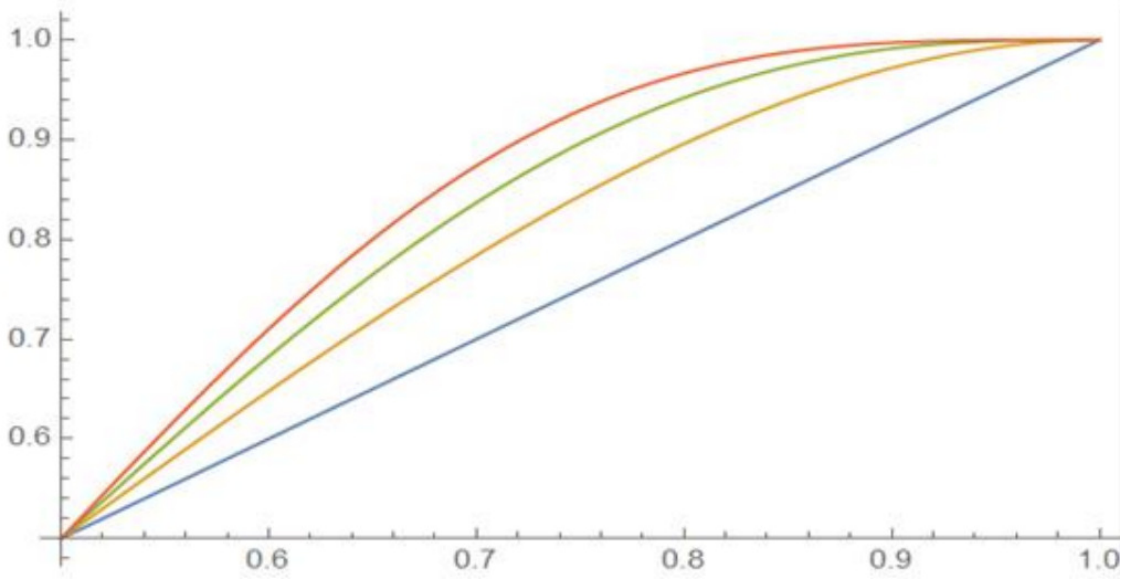

FIGURE 1. The probability a team with probability p of winning a given game (x-axis) wins a best of $2n+1$ series for $2n+1 \in \{1,3,5,7\}$ (the bottom curve is a best of one, and not surprisingly the probability of winning the series is the same as winning a game, and each curve higher is a series of length two more, note how quickly the better team's chances of winning a best of seven approach 1. Image from [CM].

over 150 playoff series as data, we have a large enough sample to do a reasonable comparison of our predictions and results, and found phenomenal agreement.

While there are other adjustments one can consider, if we wish to keep the symmetry of the probability one team wins equals the probability the other loses, the only possible scaling of the form $(R_{\text{opponent}}/R_{\text{league}})^b$ is to have $b = 1$. As an experiment we tried anyone varying b and computing the number of series win credits to assign to the home team; the exponent value which best fit the actual data happened to be $b = 1$.

A major advantage of our analysis is that we are incorporating data from both teams in an estimation of the probability of either winning. There are many related projects that can build on this. We are still treating each team as a monolithic quantity: it does not depend who you start, or what starter you face; all that matters is your average properties. This of course is not true, though the hope is that correcting for these changes will be lower order effects. For example, for platoon effects a team will often change some of their starters based on the handedness of the opposing pitcher, impacting their run production as well as their defense. It would be interesting to break teams into sub-teams, taking into account who they have on the mound and the handedness of their pitcher. Additionally, one could do a finer analysis of series and try to incorporate home field advantages game by game, though there are so many effects that can be of greater importance but are not easily incorporated. For example, in 2005 the White Sox played the Red Sox in the first round of the playoffs; the White Sox had clinched their spot with enough time to rest their starters, while the Red Sox were fighting till the last

game of the season to reach the playoffs. It is reasonable to posit that this difference helped lead to the White Sox sweep (14-2, 5-4, 5-3).

REFERENCES

[Ad] J. Adler, *Baseball Hacks: Tips & Tools for Analyzing and Winning with Statistics*, O'Reilly Media, 2006

[A] J. Albert, *A Breakdown of a Batter's Plate Appearance – Four Hitting Rates*, By the Numbers **16** (2006), no. 1, 23–29.

[ADMY] A. Almeida, K. Dayaratna, S. J. Miller and A. Yang, *Applications of Improvements to the Pythagorean Won-Loss Expectation in Optimizing Rosters*, in Artificial Intelligence, Optimization, and Data Sciences in Sports, Springer Optimization and its Applications series. To appear.

[Ba] S. Foreman, https://www.baseball-reference.com.

[Bi] P. Birnbaum, *When* $\log 5$ *does and doesn't work*, Sabermetric Research: January 7, 2016: http://blog.philbirnbaum.com/2016/01/when-log5-does-and-doesnt-work.html.

[Br] J. C. Bradbury, *The Baseball Economist: The Real Game Exposed*, Dutton, 2007.

[BZ] B. Baumer and A. ZImbalist, *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*, University of Pennsylvania Press, 2014.

[C] F. M. Chimkin, *Another Look at Runs Created*, Baseball Research Journal (2003), https://sabr.org/journal/article/another-look-at-runs-created.

[CM] R. Cleary and S. J. Miller, *GOATs and BOATs; or When Might 11/13 be Less Than 6/18?*, Mathematics in Sports **5** (2023), no. 1, 7 pages. https://libjournals.unca.edu/OJS/index.php/mas/article/view/30/16.

[CGLMP] T. Corcoran, J. Gossels, V. Luo, S. J. Miller and J. Porfilio, *Pythagoras at the Bat*, in Social Networks and the Economics of Sports (edited by Panos M. Pardalos and Victor Zamaraev), Springer-Verlag, 2014, pages 89–114.

[De] D. Decatur, *Behind-the-Scenes Baseball*, ACTA Sports, 2006.

[DM] K. Dayaratna and S. J. Miller, *First Order Approximations of the Pythagorean Won-Loss Formula for Predicting MLB Teams Winning Percentages*, By The Numbers – The Newsletter of the SABR Statistical Analysis Committee **22** (2012), no 1, 15–19.

[Fo] J. Tatsubana, https://footystats.org/england/premier-league/draws, https://footystats.org/england/premier-league/average-total-goals-table.

[Go] S. Goldman (editor), *Extra Innings: More Baseball Between the Numbers from the Team at Baseball Prospectus*, Basic Books, 2012.

[HJM] C. N. B. Hammond, W. P. Johnson and S. J. Miller, *The James Function*, Mathematics Magazine **88** (2015) 54–71.

[He] J. Heumann, *An improvement to the baseball statistic "Pythagorean Wins"*, Journal of Sports Analytics **2** (2016) 49–59.

[Hu] H. Hundel, *Derivation of James' Pythagorean Formula*, 2003; see https://groups.google.com/forum/#!topic/rec.puzzles/O-DmrUljHds.

[Ja] B. James, *1981 Baseball Abstract*, self-published, Lawrence, KS, 1981.

[JT] M. Jones and L. Tappin, *The Pythagorean Theorem of Baseball and Alternative Models*, The UMAP Journal **26** (2005), no. 2, 12 pages. https://www.comap.com/membership/member-resources/item/the-pythagorean-theorem-of-baseball-and-alternative-models-umap.

[Ke] J. Keri (editor), *Baseball Between the Numbers: Why Everything You Know About the Game Is Wrong*, Basic Books, 2006.

[LM] V. Luo and S. J. Miller, *Relieving and Readjusting Pythagoras*, By The Numbers – The Newsletter of the SABR Statistical Analysis Committee **25** (2015), no. 1, 5–14.

[MABF] B. McShane, M. Adrian, E. T. Bradlow, P. S. Fader, *Count Models Based on Weibull Interarrival Times*, Journal of Business & Economic Statistics **26** (2008), no 3, 369–378.

[Mi] S. J. Miller, *A derivation of the Pythagorean Won-Loss Formula in baseball*, Chance Magazine **20** (2007), no. 1, 40–48 (an abridged version appeared in The Newsletter of the SABR Statistical Analysis Committee **16** (February 2006), no. 1, 17–22, and an expanded version is online at http://arxiv.org/pdf/math/0509698).

[Min] R. B. Minton, *Sports Math: An Introductory Course in the Mathematics of Sports Science and Sports Analytics*, CRC Press, 2017

[Pe] D. Perry, *Winners: How Good Baseball Teams Become Great Ones*, Wiley, 2006.

[Sch1] M. J. Schell, *Baseball's All-Time Best Hitters*, Princeton University Press, Princeton, NJ, 1999.

[Sch2] M. J. Schell, *Baseball's All-Time Best Sluggers*, Princeton University Press, Princeton, NJ, 2005.

[Vo] R. Vollman (with T. Awad and I Fyffe), *Stat Shot: The Ultimate Guide to Hockey Analytics*, ECW Press, 2016.

[Wi] Wikipedia, *Pythagorean Expectation*, http://en.wikipedia.org/wiki/Pythagorean_expectation.

Email address: rcleary@babson.edu

DEPARTMENT OF MATHEMATICS, ANALYTICS, SCIENCE AND TECHNOLOGY, BABSON COLLEGE, WELLESLEY HILLS, MA 02481

Email address: jj21@williams.edu

WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267

Email address: cmill17sport@gmail.com

MT GREYLOCK REGIONAL HIGH SCHOOL, WILLIAMSTOWN, MA 01267

Email address: sjm1@williams.edu, Steven.Miller.MC.96@aya.yale.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267, USA

Email address: jdm10@williams.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267, USA

Email address: ns16@williams.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267, USA