The Effect of School Size on Cross Country Performance

Matt Kretchmar
Department of Mathematics and Computer Science
Denison University
Granville, OH 43023
Email: kretchmar@denison.edu

***Abstract*—As in many states, Ohio divides its school districts into different athletic divisions based on school enrollment in order to maintain a fair and competitive environment between schools of different sizes. In the sport of cross country racing, all schools who field a team are grouped into one of three divisions. We examine the ability of teams to compete fairly within each division. Our approach is to create statistical models of high school runners and then use Monte Carlo techniques to simulate the competition between schools. Our model isolates differences in school enrollment to separate those effects from other factors impacting a school's cross country performance. Our analysis reveals a significant disparity in competitive equity between schools within divisions arising from large differences in school enrollment.**

I. Introduction

The primary approach in this paper is to use statistical modeling and Monte Carlo simulation techniques to study the effect of high school enrollment size on performance in the sport of cross country racing. Both statistical modeling and Monte Carlo simulation have an extensive and long history of application to sports analytics. Even as early as 1974, researchers were using Monte Carlo simulation to study the effects of batting order choice in baseball [10]. Other researchers have applied similar simulations to measure outcomes [8] and duration [9] in tennis matches. Certainly the explosion of Sabermetrics following the popularity of *Money Ball* shows just how wide sports analytics has spread and also entered the lexicon of popular culture [11]. We follow in this same tradition to apply similar techniques to distance running performance.

A. Background

The Ohio High School Athletic Association (OHSAA) governs and oversees Ohio's high school interscholastic athletic competitions[5]. Part of their responsibility is to sort Ohio's 735 recognized high school programs into different divisions in order to create fair competition for the teams involved. In the sport of cross country (XC), OHSAA divides the schools into three different, equally-sized athletic divisions[6][1]. The intention of the division process is to create an equal playing field of three divisions with the same number of teams, thus ensuring no team is advantaged or disadvantaged by competing against a greater or lesser number of teams.

Figure 1 shows the distribution of school population (enrollment numbers) for the 735 recognized high schools in the state of Ohio [2][7]. They vary from the smallest school, The Choffin Career Center, with 47 students to the largest, William Mason High School, with 3531 students. Notably, the distribution is not uniform; there are many more small schools and very few large schools. This distribution of Ohio high schools closely follows an exponential distribution.

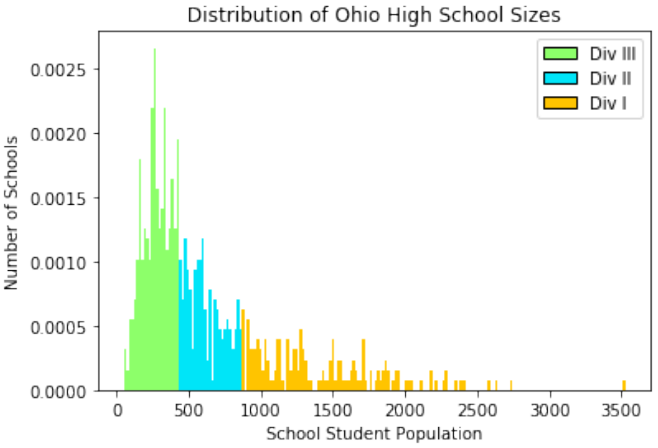

Fig. 1. School Enrollment Distribution

In this same figure, we see the three athletic divisions created for cross country based on equally-sized groups [3]. Division I, composed of the schools with the largest enrollment, is indicated with orange. Division II, medium sized schools, is in blue, while the smallest third of Ohio schools are placed in Division III shown in green. One of the tensions inherent in the division process becomes immediately evident. In order to maintain equal sized divisions, the difference in enrollment sizes among Division I schools is much larger than it is for the other two divisions. This disparity introduces competitive inequity in the sport counteracting the equity

[1]The actual division of teams is computed separately for boys and girls, and is based on a slightly more complicated enrollment computation.

[2]Only 501 of these schools fielded a cross country team in 2018.

[3]The actual divisional alignment is slightly different because many of these schools, especially the smaller ones, did not field an XC team.

sought by forming equal sized divisions. We explore this tension in this paper.

B. Map of Paper

This section provides an overview of the division-creation problem in the state of Ohio. In Section II we first build a statistical model of high school running performance. We then use this model to simulate the formation of different high school cross country teams. This section culminates with a Monte Carlo algorithm to simulate the competition of those teams in XC meets. In Section III we analyze the results of those simulations. We show the statistical disparity in runners' abilities between a large school and a small school. The Monte Carlo simulations reveal how that disparity translates into a significant competitive advantage for larger schools. Section IV summarizes our main findings while Section V introduces possible future avenues of study.

II. Overview of the Model

The goal of this work is to isolate and investigate the effect of school size on a team's ability to compete in the sport of XC. We do so by first building a statistical model of runners' abilities and use that model to form XC teams. We then use Monte Carlo techniques to simulate these teams competing in a XC meet[4]. Without loss of generality we concentrate on female teams, though the analysis is presumed to be similar for male runners.

A. Distribution of Per-Mile Performance of High School Runners

Our first step is to model the running performance of 14 to 18 year-old females. We use data collected by HealthLine of 10,000 female runners who competed in 5k races in 2010 [3]. We combine that with the data of 90 million female recorded runs on strava in 2017 to build a model of running performance distribution [13].

Figure 2 shows the variation in running paces collected from our data for a general population of female, high school aged students. The pace varies from the fastest runners who are able to maintain a sub 6-minute per mile 5k pace, to the slower runners who average over 20 minutes per mile. We fit a beta probability distribution to this data as shown in the orange line.

The beta model is given by

$$\text{pdf} = f(\hat{x}, a, b) = \frac{\Gamma(a+b) \cdot \hat{x}^{a-1} \cdot (1-\hat{x})^{b-1}}{\Gamma(a)\Gamma(b)}$$

where $\hat{x}$ is the normalized (shifted and scaled) value of x and Γ is the generalized factorial function give by

$$\hat{x} = (x - \text{shift})/\text{scale}$$

$$\Gamma(z) = \int_0^\infty x^{z-1} \cdot e^{-x} dx = (z-1)!$$

Fitting the beta distribution model to our data yields the following parameters:

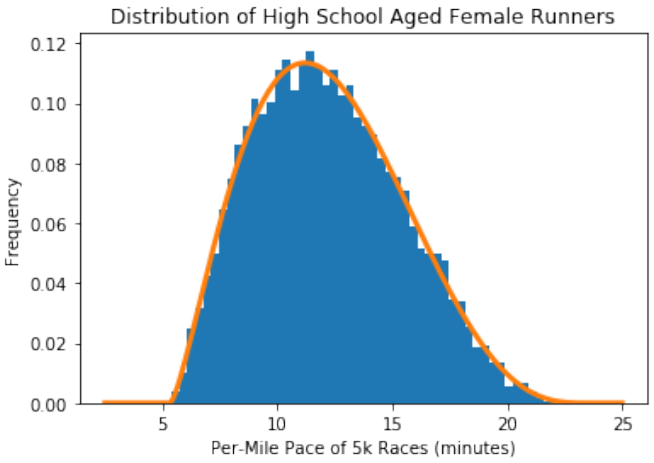

Fig. 2. Distribution of Per-Mile Pace of Female High School Students

Parameter	Value
a	2.464
b	3.856
shift	5.316
scale	17.608

For a school with N female students, we randomly draw N samples from this distribution. This gives us N virtual female students/athletes to form a school's entire female student population, and from which we can form an XC team.

B. Selection of a School Team

Our next modeling step is to create a virtual team for each school. High school varsity XC teams feature seven runners on their team[4]. From the N female sampled students, we select the fastest seven to form our team. This selection process ignores some of the complicating factors in real life. In the real world, for example, one of the top seven students may choose to participate in a different sport or no sport at all. Or one of the top runners might drop out due to illness, injury, or ineligibility. Our model makes the simplifying assumption that all top seven runners will choose to and be able to participate in XC for the school's team.

C. Simulation of an XC Meet

A meet is a 5k race between two or more teams. We simulate the competition in a meet in the following way. The 5k paces of each runner on each team are "average" paces. On any given day, an individual runner may run faster or slower than their average pace; it is not uncommon for a runner's pace to vary by 6% or more depending on the course, the weather conditions and other factors. An individual's running pace on a given day is again modeled by a beta distribution as shown in Figure 3. In this figure, we simulate a particular runner with an average 5k per-mile pace of 6:35 who may run as fast as 6:20 per mile or have a bad day and run closer to 7:00 pace. We

[4]A minimum of five runners are required to field a complete team.

sample randomly from this distribution to obtain that runner's pace on a specific day of competition.

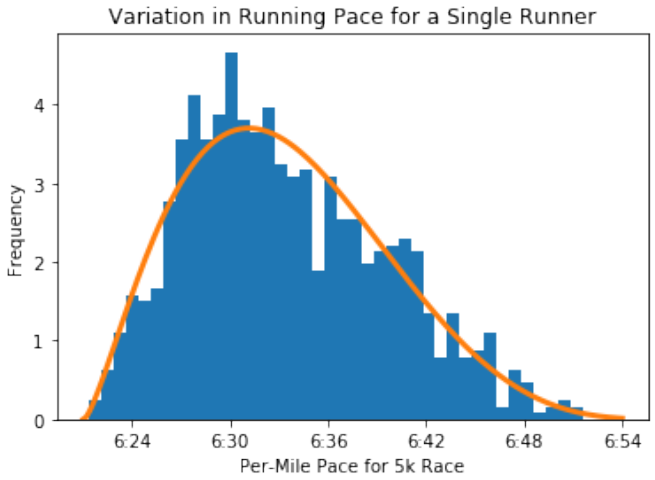

Fig. 3. Variation in Per-Mile Pace of an Individual Runner

The parameters for this beta model are:

Parameter	Value
a	2.396
b	4.638
shift	6.531
scale	0.607

Note that this model is specific to one female runner. Each runner will have their own model, particularly their own shift and scale parameters based on their unique average running pace.

After each individual runner's pace is determined for that particular event, the 5k running times for each runner are computed and the racers are ranked accordingly. Teams score by adding the sum of the finishing places for the first five runners from each team [5]. Teams are ranked by lowest team score first to highest team score last[2]. We simulate many races over many possible seasons to determine how often schools of different sizes are able to finish in the top of a division, or to win a particular meet (i.e. the state title).

D. Meets and Seasons

We build our model to capture the same key competitions that lead to a state championship in Ohio. Namely, schools within each division are grouped into four regions, each region into several districts, and each district comprised of various leagues/teams. A team must first place high enough in their local league to obtain entry in their district race. The top teams from each district race then go on to compete at their regional race. The top teams from each of the four regions then compete in the state meet to determine the state champion.

$$\text{League} \rightarrow \text{District} \rightarrow \text{Region} \rightarrow \text{State Meet}$$

[5]There can be complicating factors where this is not exactly the case.

Our model features 168 teams in each Division [6]. We group them into twelve districts comprised of 14 teams per district. The top 5 teams from each district race then go on to one of four regions. There are three districts funneling into each region, so the top 5 from each district produces regions of 15 teams. The top 5 teams from each region then go on to states, which features the top 20 teams from the division. This process is independent and parallel in each division, so the state meet will features three different girl's races, each with 20 teams (Div I, Div II, and Div III).

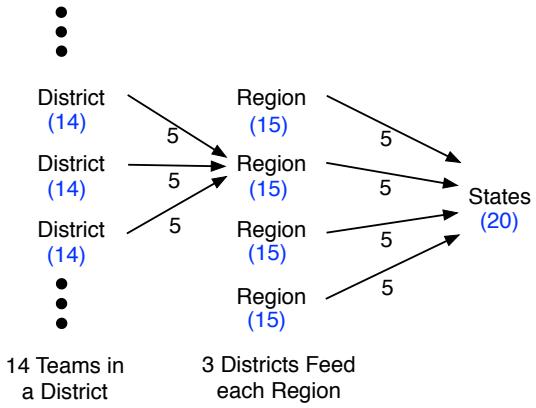

Fig. 4. Districts, Regions, States

Our model randomly assigns teams to different district groups for each season. This is unlike the real situation in which teams are grouped geographically and generally compete against the same teams in each district annually. We do this so as not to bias the model against one particular geographic alignment of teams within districts and regions.

Our model simulates 10,000 different seasons. Per each season, we re-select each school's N female students (using different values of N for each school); re-form each school's XC team, randomly re-align the districts and regions, and then simulate the district $\rightarrow$ region $\rightarrow$ state meet sequence to determine how each school fares. This process simplifies and deviates from real life where three quarters of the students from one season typically return for the next season (only graduating seniors leave). We do not simulate this carry-over effect between seasons for a school. Nor do we capture the effect of good coaching; namely schools with top programs and good coaching "grow" their talent and are likely to see high performance carry-over effects from year to year compared to schools who do not have such programs. These simplifications are intentional since our goal is to isolate the effects of school enrollment size from the other factors that influence XC performance.

III. Analysis

In this section we use our model to examine the relationship between school size (female enrollment) and competitive

[6]In the 2018 Ohio XC season, each division had 167 teams.

results in the state meet. We start our analysis by creating two imaginary teams that reflect actual teams in Ohio. Team1 is simulated using the population of the largest school in Division I (1728 female students) while Team2 simulates the smallest school in Division I with 433 female students. Therefore, $N_1 = 1728$ and $N_2 = 433$.

A. Fastest Runner

We start our analysis by first examining the per-mile pace distribution of the fastest runner from each team. Because we are only interested in learning about distribution of runners, we simulate only 1000 seasons instead of the 10,000 seasons for the full competitive model. For each season we randomly sample the beta distribution model to form the female population for each school. We select the fastest female runner from each school's population and examine their distribution across the 1000 different seasons.

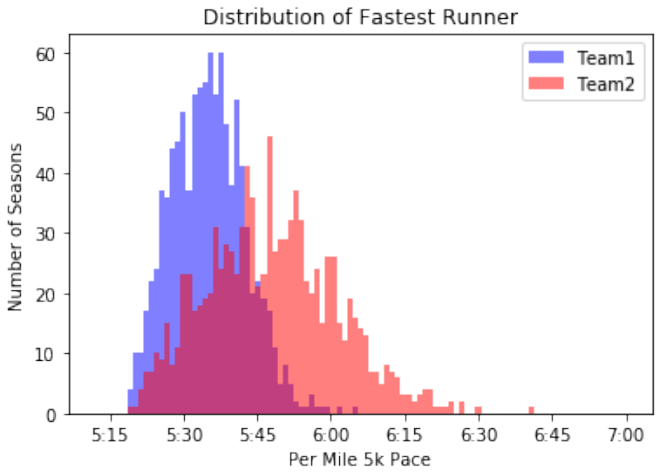

Fig. 5. Distribution of Per-Mile Pace of Fastest Female Runner

Team	Female Population	Avg Pace	Stdev of Pace
Team1	1728	5:35	.123
Team2	433	5:48	.219

Fig. 6. Fastest Runner Data

In Figure 5 we see the distribution of the fastest female runner from each school. On average, the fastest runner from Team1 is about 13 seconds per mile (about 41 seconds in a 5k) faster than the fastest runner from Team2. However there is significant overlap in the distributions of 1000 seasons. Team2, the smaller team, fields a faster runner than Team1 in about 20% of the seasons. Note also that the standard deviation of paces for the fastest runner from Team1 is much smaller. Though Team1 fields a faster runner on a consistent basis, the real advantage for Team1 is less variation in their top runner than in the smaller Team2. This concept will be pivotal as we expand the team to a full seven-person roster.

B. Team Selection

We now simulate the selection of a seven-member team roster over the same 1000 seasons. Again, for each season we sample N times to create a school's female population; now instead of selecting just the fastest runner, we draw the top seven female runners from each school and determine their average 5k paces over those 1000 seasons.

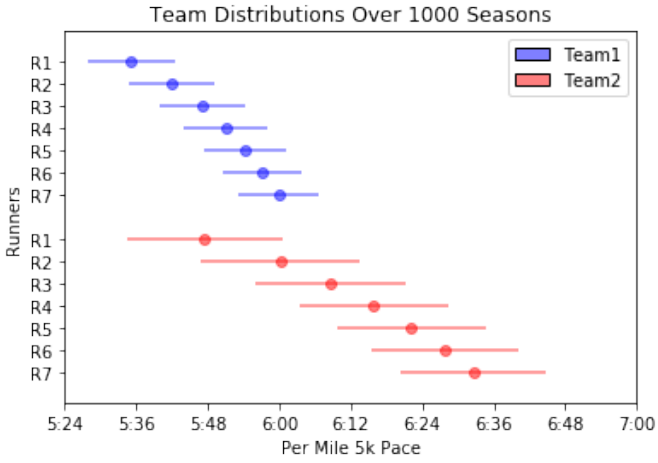

Fig. 7. Distribution of Per-Mile Pace of Teams

Figure 7 reveals a much greater disparity of runner performances for whole teams than we saw for individuals. Team1 (from the large school) is in blue at the top; Team2 (from the small school) is in red at the bottom. We plot the average pace of each placed runner over 1000 seasons as a dot. The error bars indicate the standard deviation of that placed runner over the 1000 seasons.

For Team1, we see that their fastest placed runner averages about 5:35 per mile while their slowest, the seventh placed runner, averages about 6:00 per mile. For Team2, we see that the fastest runner averages 5:48 per mile while the slowest averages about 6:34 pace. This figure illustrates the powerful nature of order statistics by varying population size. The larger school is able to field seven much faster runners than the smaller school. Team 1's seventh fastest runner, is still faster on average than Team 2's second fastest runner!

C. Analysis of Meet Simulation

We use the techniques outlined in Section II to simulate a series of full XC seasons.

- We consider 168 schools from each of Ohio's three divisions. We use the actual enrollment numbers N for each school using OSHAA data.
- We simulate 10,000 different XC seasons. For each season, we randomly sample running distributions for each school's female student population and select the top seven fastest female runners for each school's team during that season. Each season starts with a new batch of seven different runners.

- For each season, we randomly sort the teams into different districts and regions.
- For each season, we simulate each district championship race, moving the top five teams into each region. We simulate each of the four regional championships, sending the top five teams on to states. We then simulate the state championship of the top 20 teams in Ohio.
- This gives us a baseline of 10,000 different seasons. We compute each team's number of wins (first place finishes or state championship titles), second place finishes, third place finishes and the number of appearances in the state championship meet.

Table 8 shows the results from Division I for the simulation. We see three of the larger teams which have historically done very well in state XC meets. We also see the smallest team in Div I. Mason, the largest Div I program claims 850 state titles (8.5% of them!!!) while also placing second 618 times and third 528 times; they appear in the state meet 6244 times (62% of the time!). At the other end of the spectrum, the smallest Div I school, Granville, wins two state titles among its 249 state championship appearances.

School	**Size**	**1st**	**2nd**	**3rd**	**Apps**
Mason	1728	850	618	528	6244
Centerville	1298	322	286	246	4155
BeaverCreek	1154	228	196	217	3415
...	...	...	...	...	...
Granville	433	2	2	3	249

Fig. 8. Div I Simulation Results

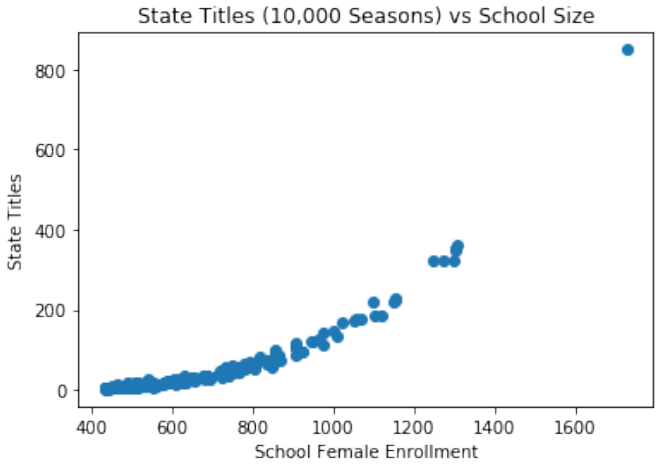

Fig. 9. State Title Count for Division I

Figure 9 shows a school's state championship count in these 10,000 simulated seasons. Clearly, larger teams win much more frequently as expected. Figure 10 shows the number of times each program makes it to the state meet (effectively a top-20 finish).

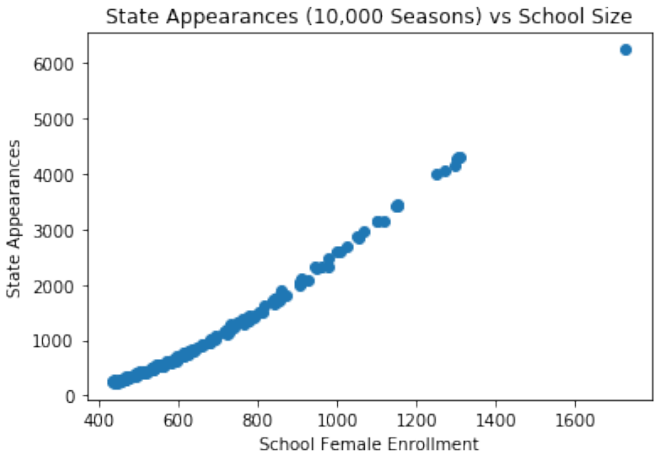

Fig. 10. Number of State Meet Appearances for Division I

D. Proportional Equality

Notably, the relationship between state championships and school size in Figure 9 is not linear (not a straight line) indicating that finish place is not proportional to school size. In his work *Nicomachean Ethics*, Aristotle establishes the *principle of proportionality* as a standard for measuring equity[1]. While we should expect large schools to win more often – that is the nature of large vs small – the winning rate of a school should be *in proportion* to its relative size. If Team1 hails from a school that is four times larger than a school fielding Team2, we should expect Team1 to win four times more often.

To evaluate winning percentage on the basis of proportionality we create a normalized win rate. A team's actual winning rate is computed as the number of championships divided by the total number of seasons. This win rate should be in proportion to the school's relative size. We compute the expected win rate by measuring the school's size as a fraction of all Ohio students (sum of all school sizes in that division).

$$\text{normalized win rate} = \frac{\text{actual win rate}}{\text{expected win rate}} = \frac{\frac{\text{championships}}{\text{seasons}}}{\frac{\text{school population}}{\text{total Ohio student population}}}$$

Figure 11 shows the expected (based on proportional population) and simulated wins for the largest and smallest teams in Division I.

In a fair system, the normalized win rate for schools should be approximately 1.0 (the expected and actual win rates should be nearly identical). The percentage of time a particular school wins a state title should be about equal to the percentage of students who attend that school from the whole state student population in that division. Schools that have a normalized win rate significantly above 1 are winning more often than they should on the basis of their school size. A normalized win rate of 2 indicates a school is winning twice as often as it should. Similarly, a normalized win rate of below 1 indicates a school winning much less frequently than it should.

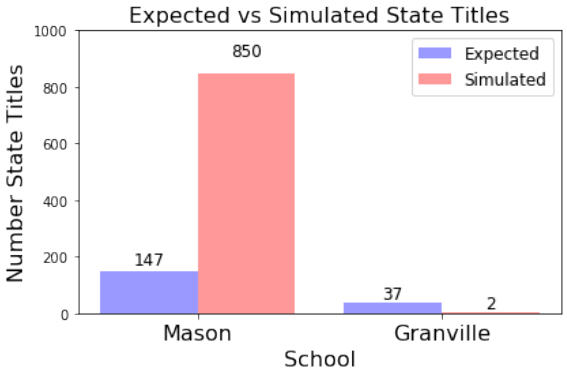

Fig. 11. Expected VS Actual Titles

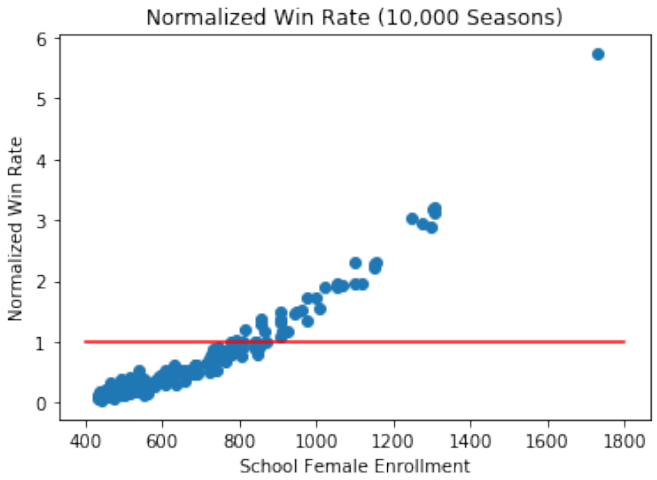

Fig. 12. Normalized Wins for Teams

Figure 12 shows the normalized win rate for Division I schools as a function of their school female population. Ideally, all schools (large and small) should be clustered on or near the red line (the 1.0 mark for normalized wins). As we can see, the larger schools have normalized win rates well above 1, with the largest school winning more than five times more frequently than it should. The smaller schools have normalized win rates well below 1, hovering near zero. Mason, the largest Div I program, has a normalize win rate of 5.75 indicating they win almost six times as many state titles as they should on the basis of their school size. Granville, with a normalized win rate of 0.0540, wins 18.5 times less frequently than it should. To put it another way, Mason is almost four times larger than Granville in population, but wins state titles 106 times more often than Granville.

Figure 13 shows the normalized win rate for all three Ohio divisions. The y-axis is now in logarithmic scale to more accurately reflect the low winning rate for smaller teams. Schools are plotted with a green dot if their win rate is above 1.0, and a red dot if it falls below 1.0. We see the same kind of disparity in Division II and Division III programs. The normalized win rate for the smallest programs is around 0.2 (winning about 1/5 the rate as they should), while the largest programs have a normalized win rate of 2 (winning twice as often as they should). While the inequity is present in all three divisions, it is significantly more pronounced in Division I. We hypothesize this is an effect of the larger discrepancy between school sizes in Division I. We also see that the majority of schools have a normalized win rate below 1.0 – most programs are at a disadvantage to their larger schools within the same division.

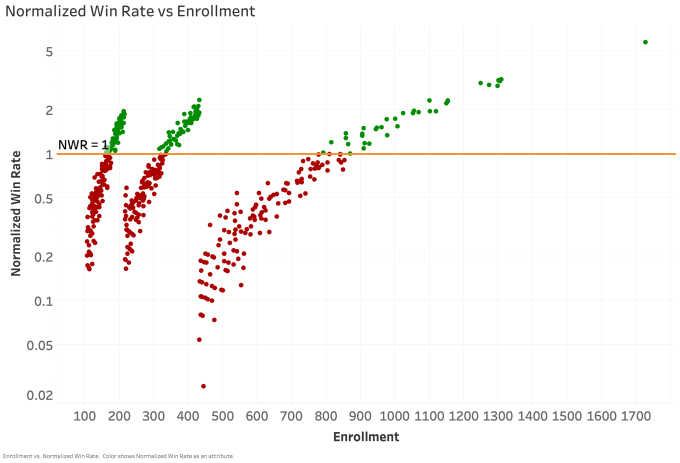

Fig. 13. Normalized Wins for All Teams

IV. Conclusion

The *ceteris paribus* assumption in this work is that all other things are considered equal. Of course that is not the case. There are strong programs and weak programs. Excellent coaching and less effective coaching. Legacy programs supported by the community and programs that struggle to field a team each year. These other factors do contribute significantly to the effectiveness of various XC programs and their ability to compete at high levels, and we do see some instances where smaller schools on occasion score well in a state meet. In this work, however, we wish to eliminate those other factors to isolate only the effects of school size on a program's fair ability to compete in the state meet.

Small communities take great pride in the success of their high school athletes. It is not uncommon to drive into a small town and see a sign greeting arrivers with something like "Home of the 1996 Panther State Champions"! Wins like this define communities for years. These kind of opportunities should be *possible*. They may be rare, but they should be attainable. If a team believed that its chances for this kind of success were zero, they would lose motivation to compete and even to participate. The OHSAA lists first and foremost among its commitments: *'Establishing and regulating regular season and tournament standards in order for competition to be fair and equitable.'*[5] The notion of equitability is fundamental to the nature of sports and the reason that governing organizations like OHSAA exist.

This work shows that the process by which Ohio schools competing in the sport of cross country does not create an equitable situation. The process is grossly inequitable. This is not an intentional act of the division process. In fact, the process was created in an attempt to promote greater equity. It is only through careful analysis that the true level of inequity becomes apparent. Among the significant conclusions of this work are:

- Individual state champions in Division I can come from small schools. Our analysis shows that it is possible that an individual runner from a small school can outperform runners from larger schools. Given there are significantly more smaller schools than larger, it should not be an uncommon occurrence at state championship meets.
- However, the power of order statistics disproportionately favors larger schools and creates a significant disparity in running talent beyond the top runner. Runners two through seven on teams from smaller schools are consistently and significantly slower than their counterparts on teams from larger schools.
- In simulated races, large teams have a disproportionate opportunity to win. The larger schools win state titles up to five times more often than they should based solely on their school size. The smallest schools in Division I effectively have zero chances of ever winning a state championship meet.
- In a fair system, a team should win titles *in proportion* to its relative size. When we normalize a school's win rate by their school population, we find that the largest schools win about five times more often than they should, while the smallest schools win about 18 times less frequently than they should. The effect of rank statistics greatly skews a team's ability to compete fairly. The result is a grossly unfair playing field that greatly favors the larger schools.

V. Future Inquiries

This paper analyzes the fairness of the current divisional process for Ohio's cross country programs. It does not propose alternatives nor evaluate them on their potential equality. An obvious next direction is to explore alternative methods of division and apply similar rigor in determining their ability to present a level playing field to all of Ohio's student athletes.

This paper simulates girls XC teams under the assumption that the analysis of boys teams would be very similar. Data suggests that boys and girls develop running ability differently; girls physically mature at a younger age and thus are closer to their top talent level when starting a high school career as a freshman than would a similar age boy. Consequently imbalances in depth of talent are possibly even more pronounced for boys teams where freshmen cannot contribute to a team's varsity roster in the same way as upperclass students. We would like to pursue this line of inquiry with actual data from boys and girls teams.

Another factor not considered in this paper is a team's training approach[7]. Smaller teams with fewer top-level athletes must necessarily take a more cautious approach in training, perhaps reducing the intensity and/or mileage of the training program. The risk of getting just one athlete injured is too high, as there is no one left to replace them at the same level. Larger teams might have up to two dozen athletes near the top level. These programs can afford to push their athletes harder and thus elevate the ones who do not get injured to a higher performance level. If one of the top athletes gets injured, there is another of almost equal ability ready to step in. The very largest programs can actually field two or more varsity level teams. They can afford to send their "B-team" to a district or regional final while optimizing the resting/training schedule of the A-team for the state meet. These other built-in advantages of size are not captured in this model.

During my conversations with various coaches and student athletes, more than one coach/athlete suggested that private and parochial schools win more frequently than they should (on the basis of student population) because they recruit good athletes from nearby school districts. While this effect is possibly more pronounced in other sports (football, basketball, volleyball), it may be worth investigating its presence in XC.

Acknowledgments

The authors would like to thank the Anderson Scholarship Fund and the William G. Bowen and Mary Ellen Bowen Endowed Fund for funding this work. Thanks to Granville Cross Country Coach Bart Smith and former coach Dave Agosta for providing data and insights as to Ohio's process of creating divisions.

References

[1] Aristotle, *Nicomachean Ethics*. New York, NY: Macmillan Publishing Company, 1962.
[2] FloTrack: http://www.wfltrack.org/cross/resources/HowToScoreXC.pdf
[3] Health Line: https://www.healthline.com/health/average-mile-time
[4] Liu, Jun S. *Monte Carlo Strategies in Scientific Computing*. New York, NY: Springer-Verlag. 2001.
[5] Ohio High School Athletic Association: https://ohsaa.org.
[6] OHSAA XC: https://www.ohsaa.org/sports/cc
[7] Ohio School Enrollment Data: http://education.ohio.gov/Topics/Data/Frequently-Requested-Data/Enrollment-Data
[8] Kovalchik, Stephanie Ann, and Martin Ingram. Estimating the duration of professional tennis matches for varying formats. *Journal of Quantitative Analysis in Sports*. Vol 14. No 1. 20018.
[9] Newton, Paul K., and Kamran Aslam. Monte Carlo Tennis: A Stochastic Markov Chain Model. *Journal of Quantitative Analysis in Sports*. Vol 5. No 3. 2009.
[10] Freeze, R. Allan. An Analysis of Baseball Batting Order by Monte Carlo Simulation. *Operations Research*. Vol 22. No. 4. Operations ResearchVol. 22, No. 4. 1974.
[11] Lewis, Michael M. *Moneyball: The Art of Winning an Unfair Game*. New York: W. W. Norton. 2003.
[12] Outdoor Industry: https://outdoorindustry.org/press-release/2015-strava-insights-show-cycling-and-running-landscape-in-the-u-s/.
[13] Runner's World: https://www.runnersworld.com/news/g25333911/strava-annual-report-running-statistics/.

[7]Thanks to Dave Agosta for this idea.