Evaluation of the Fairness for Schools with Different Size in Cross Country Running

Yubo Wang
email: wang_y2@denison.edu
Department of Mathematics and Computer Science
Research Advisor: Matt Kretchmar

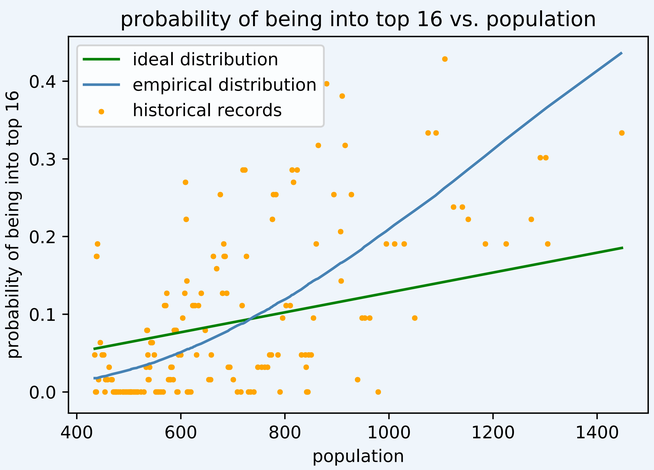
Introduction

High school cross country racing is an outdoor running competition of American high schools. In the Ohio cross country running, all schools are divided into three divisions based on the population size to maintain fair competition.

The previous research on this problem had shown the competition between school with different size is not fair. The author of previous research built a beta model of running abilities then the Monte Carlo was used to simulate the competition between schools with different sizes.

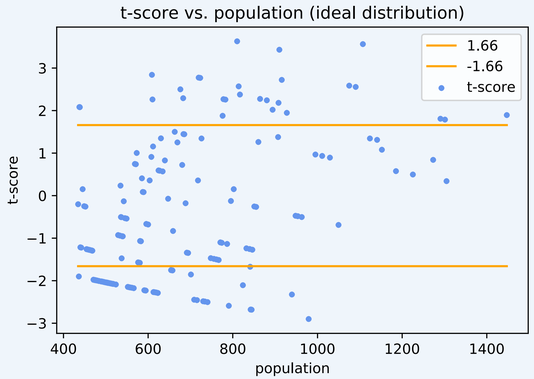
In this research, our goal is to verify the previous work. We will verify the result in two ways: 1. Statistical comparison between historical records against the simulation result, 2. Doing an analytical calculation based on order statistics. In our research, we introduce two probability distribution of the probability of a school being into the top 16 based on previous ideas:
1. The ideal distribution is based on the fair win rate which is proportional to the ratio between the school population and total population in the division. According to the author of previous research, this is the fair win rate of the competition.
2. The empirical distribution is based on the simulation results of the beta model that we simulate the per mile pace of each runner and see which schools being into the top 16.

The figure shows the probability of being into top 16 from both distributions and historical records for the schools with different population. The orange points are the historical records and the blue curve show the distribution based on the beta model and the green curve show the ideal distribution.

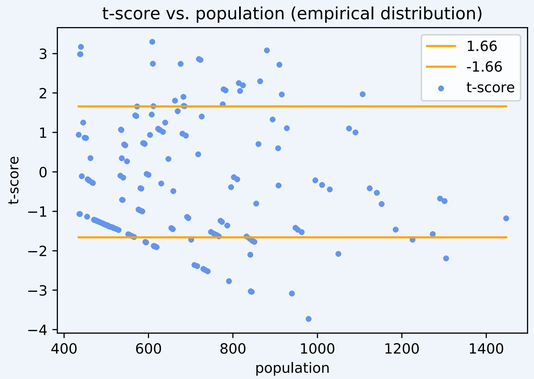


Hypothesis Testing

In this part, we focus on historical records. We use the two samples proportional testing to see if there is significant difference between the historical records and each distribution. The following figures shows the result of the hypothesis testing. The x-axis is the school population and y-axis is the t-score from the hypothesis testing. The blue points are the t-scores of samples with different population size and the orange line shows the range of passing the hypothesis testing. If the points are not between the range, it means that the distribution doesn't predict the historical record at that point very well.



This figure shows the result from hypothesis test between ideal distribution and historical records. We see that it predicts the small schools poorly because lots of small schools have the t-score below the orange line which means those small schools have the probability of being into the top 16 smaller than the prediction from the ideal distribution.



This figure shows the result from hypothesis test between empirical distribution and historical records. We see this distribution predicts the most of samples well except for the large schools as we see the most of large schools have a t-score smaller than zero which they have the probability of being into the top 16 smaller than the prediction from the empirical distribution.

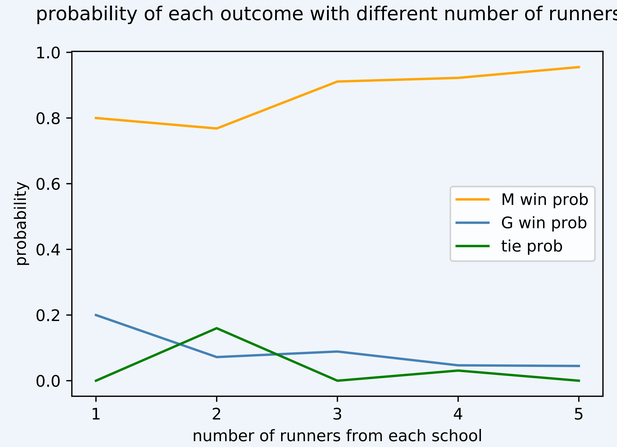
Thus, we can conclude that the empirical distribution better predicts the historical data as most of points are between the range of orange lines. But neither of them predicts the large schools very well.

Works Cited

1. Ohio Cross Country State Championships: https://ohio-cross-country-state-championships.runnerspace.com
2. Ohsaa: https://ohsaa.org/sports/cc/pastresults.htm
3. Matt Kretchmar. The Effect of School Size on Cross Country Performance
4. H. A. David, H. N. Nagaraja. Order Statistics. Wiley-Interscience. 3 edition. August 4, 2003

Order Statistics

In this part, we calculate the win rate for each school to verify the results from previous research analytically. However, due to the calculation complexity, we can only perform the calculation up to 10 runners. We decide to do some calculation between two schools' competition. We choose two schools from Division I: Mason (girl population: 1728) and Granville (girl population: 433).



The figure shows the win rate of the competition between Mason and Granville. The x-axis is the number of runners in each team and the y-axis is the probability winning the race. According to the ideal distribution, the fair win rate for Mason should be 80%. We see that the win rate of Mason is about 80% with one runner in each team, then it decreases a little bit when there are two runners in each team because of the tie situation. With a full team of 5 runners, Mason win much more frequent than the ideal distribution.

Conclusion

The results support the conclusion from the simulation results from the previous research that the way of dividing all the schools into three divisions is unfair to small schools. If you want to know more about our research, you can find a paper and some sample code in the repository: https://github.com/YuboW1/Evaluation_of_the_Fairness_for_Schools_with_Different_Size_in_Cross_Country_Running

Acknowledgements

This project was funded by William G. Bowen & Mary Ellen Bowen Research Endowment at Denison University. Thanks to Matt Kretchmar (the author of previous research) to help with the research.