Evaluation of the fairness for schools with different size in Cross Country Running competition

Yubo Wang
Department of Mathematics and Computer Science
Denison University
Granville, OH 43023
Email: wang_y2@denison.edu

***Abstract*—High school cross country racing is an outdoor running competition of American high schools. Like many other states, Ohio divides all the schools into different divisions based on the school size to maintain fair competition between the schools. In the Ohio cross country running, all schools are divided into three divisions based on the population size[3]. Our research focuses on the fairness of cross country running competition of girls in Division I of Ohio.**

I. INTRODUCTION

In the previous research[3], the author built a probability distribution of the per-mile pace of high school-aged female runners. Then the technique of Monte Carlo simulation is used to generate the result of state tournaments in many years. The result of the simulation is evaluated by a metric called normalized win rate.

$$\begin{aligned}\text{NWR} &= \frac{\text{actual win rate}}{\text{expected win rate}} \\ &= \frac{\frac{\text{championships}}{\text{seasons}}}{\frac{\text{school population}}{\text{total Ohio student population in the division}}}\end{aligned}$$

The normalized win rate shows the ratio between the actual win rate and the ideal win rate (the win rate when the competition is fair). According to the results, the schools with relatively small population size win less than they should and the relatively large schools win more than they should. Thus, the previous research found that the competition could be unfair for small schools.[3]

II. APPROACHES

Our goal is to verify the results in the previous research[3]. The result of previous research is obtained from simulation, so we use the historical records and analytical way to verify the result.

Our research is divided into two parts. In the first part, we look at the historical records of the top 16 schools in the past 21 years and to verify the simulation results from previous research[3]. In the second part, we apply the technique of order statistics to verify the previous results analytically.

We define two probability distributions of a single school being into the top 16 in the state tournament in a year.

- ideal distribution:
  In the ideal distribution, the probability of a single school being into the top 16 in the state tournament in a year is proportional to the ratio between its population and the total population in the division. According to the author of previous research, this is the fair way of competition[3]. (details of generating the distribution can be found in the appendix B1) Figure 1 shows the probability of being into the top 16 in the state tournament for each school with different population of girls that generated by the ideal distribution. Take Granville, Olentangy Orange and Mason as example, their probabilities of being into the top 16 according to the ideal distribution are shown below.

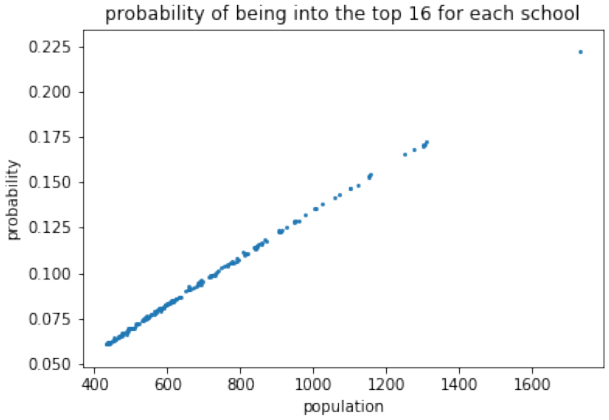

Fig. 1. probability distribution of being into the top 16 in the state tournament in a year from ideal distribution

pop	probability
433	0.068
849	0.12
1728	0.23

- empirical distribution:
  In the empirical distribution, the probability is no longer related to population size but related to the beta model from previous research. In this distribution, the smaller schools have a lower probability of being into the top 16 in the state tournament and the larger schools have a higher probability compared to the ideal distribution. (details of generating the distribution can be found in the

appendix B2) Figure 2 shows the probability of being

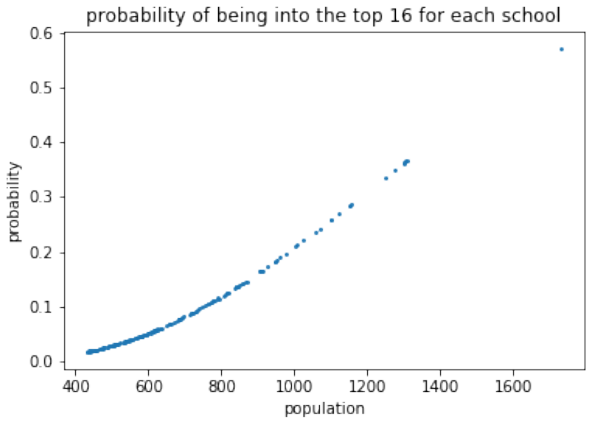

Fig. 2. probability distribution of being into the top 16 in the state tournament in a year from empirical distribution

into the top 16 in the state tournament for each school with different population of girls that generated by the empirical distribution. Take Granville, Olentangy Orange and Mason as example, their probabilities of being into the top 16 according to the empirical distribution are shown below.

pop	probability
433	0.012
845	0.14
1728	0.81

In the first part, we use the two samples proportion test on those data to do a hypothesis test. By applying two samples proportional test, we calculate, the t-score to evaluate if the probability of being into the top 16 for a school in the state tournament historical records is different significantly compare to the ideal distribution and empirical distribution. Also, we use the Bayesian formula to calculate which distribution is more likely to generate the results in the historical records in each year.

In the second part, we apply the technique of order statistics. As order statistic allow us to calculate the probability of each runner running at a specific place, we can calculate the win rate of each school in the competition.

III. PART I

In this part, we focus on the historical records and use the technique of hypothesis testing to check if the probability of being into the top 16 in the state tournament generated by two distributions are different significantly compare to the historical records. Also, we use the Bayesian formula to calculate which distribution is the historical records closest to.

A. Historical Records

We collect the historical records of XC Running competition in Ohio from 1999 to 2019[1][2]. [1] [2] In the historical records from each year, we extract the data of the top 16 schools in Division I.

For example, the top three schools in 2019 is shown below:

place	school	pop
1	Beavercreek	1154
2	Lakota East	1309
3	Hilliard Davidson	816

B. Hypothesis Testing

In this part, we use the technique of hypothesis testing to see if the historical records differ from the ideal distribution significantly.

Since we can approximate and verify that each school being into top 16 in the state tournaments each year is an independent event which means that a school being into the top 16 will not affect any other school being into top 16 at that year significantly.

$Thus, decidetouseproportiontestingonthedataandtocomparetheproba$

1) Proportion Testing: As we can calculate the proportion of schools being into the top 16 in each year, proportion testing is a good way to test the data as we can compare two proportion with it.

We call a school to get into the top 16 to be a success and the probability of success can be calculated as the actual times that a school to get into the top 16 divided by the total number of chances for that school to get into the top 16 which is the number of years that a school get into the state tournament.

$$p_{\text{success}} = \frac{\text{a school to get into the top 16}}{\text{number of years}}$$

Then, we define the following hypothesis:

- H_0: the proportion of being into the top 16 in the state tournament from historical records and one of our distribution are the same
- H_a: the proportion of being into the top 16 in the state tournament from historical records and each distribution we mentioned above are not the same

However, because we only have the data for 21 years, the sample size from a single school is too small for us to do a valid hypothesis test. To solve this problem, we combine some schools with similar population size as one school and use that school to do the hypothesis testing.

We combine three schools with approximately the same size as one school so the sample size increase from 21 to 63. To use this technique, we have to verify that the proportion of

[1]The data from 2007 to 2019 are from https://ohio-cross-country-state-championships.runnerspace.com/

[2]The data from 1999 to 2006 are from https://ohsaa.org/sports/cc/pastresults.htm

being into the top 16 in the state tournament for the three schools can be approximated as independent from each other.

In this case, the total chance should be how many years we have times how many schools we combine as one. So, we can use these data to do the hypothesis testing. Then we can rewrite the formula of two samples proportion test.

$$t - score = \frac{p_1 - p_0}{\sqrt{p(1-p)\frac{2}{n}}}$$

$$p_0 = \text{proportion of being into the top 16 according to the distribution}$$

$$p_1 = \frac{\text{actual times that those schools to get into the top 16}}{\text{total number of chances for those schools to get into the top 16}}$$

$$p = \frac{p_1 + p_0}{2}$$

2) Independent Test of Samples: In order to combine some schools as one, we need to verify that their proportion of being into the top 16 in the state tournament can be approximated as independent which one school being into the top 16 in will not affect the proportion of other two schools being into the top 16.

However, the probabilities for the three schools are technically not independent as one school being into the top 16 will make other schools' probabilities decrease a little bit, but what we want to do is to verify if they are close enough to the independent so we can apply the technique of combining three schools.

In this process, we use the beta model from the previous paper to generate the top 16 schools in the state tournament for 24,000 times. Then we choose three schools with a similar population size of girls to see all the count of possible combinations of outcomes because we will combine two or three schools as one school in the hypothesis testing.

The details can be found in the appendix A.

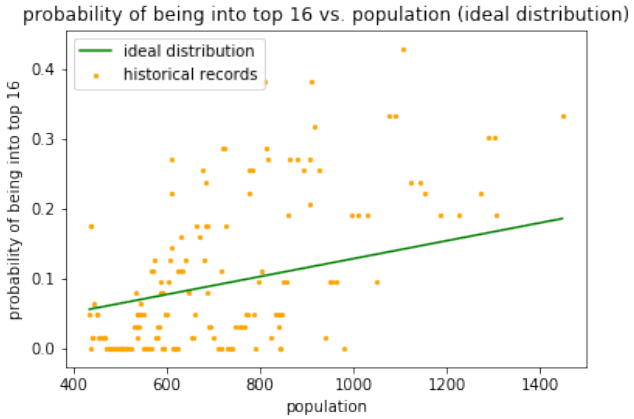

Fig. 3. proportion of being into top 16 vs. population

3) Hypothesis Testing With the Ideal Distribution and Historical Records: Figure 3 provides a general view of the proportion, the x-axis is the average population of the three schools that we combined and the y-axis is the proportion of being into the top 16 in the state tournament from historical records or from ideal distribution. According to the graph, we can conclude that the historical records are different from the ideal distribution. The larger schools have greater proportion and small schools have a lesser proportion of being into the top 16.

We set up a range of three schools which means we combine three schools with similar population size as one school. Then we traverse through all the schools and each time we do the hypothesis testing for three schools. We have the hypothesis that:

- H_0: the proportion of being into the top 16 in the state tournament from historical records and ideal distribution are the same
- H_a: the proportion of being into the top 16 in the state tournament from historical records and ideal distribution are not the same

Then we calculate the hypothesis testing by using the hypothesis above. For example, when we have schools of Mentor, Canton McKinley Senior and Lakota East:

We combine these three schools as one, we can find the average proportion of being into the top 16 from the ideal distribution which is 0.167. Then we can calculate the proportion of being into the top 16 in the historical records which is 0.190. Thus, we can calculate the t-score by applying the formula.

$$\begin{aligned} p_0 &= 0.167 \\ p_1 &= 0.190 \\ p &= \frac{0.167 + 0.190}{2} = 0.1785 \\ ts &= \frac{p_1 - p_0}{\sqrt{p(1-p)\frac{2}{n}}} \\ &= \frac{0.190 - 0.167}{\sqrt{0.1785(1-0.1785)\frac{2}{21*3}}} \\ &= 0.337 \end{aligned}$$

Then we can calculate the p-value by using the t-score and we need to notice that we use two tails test here when we doing the calculation.

Then we mode the slice of school and the new slice includes Canton McKinley Senior and Lakota East and Mason. By using the similar process, we can get the t-score for the hypothesis testing for combining these three schools is 0.060.

By doing the calculation above repeatedly, we have the following results.

Figure 4 shows the relation between t-score from the hypothesis testing and the population. The x-axis is the average population of the three schools that we combined and the

Fig. 4. t-score vs. population

y-axis is the t-score from the hypothesis testing. The two orange lines shows the range of t-scores with p-value of 0.1 (two tails) which means the points that are out of the range between the orange lines are the samples that reject the null hypothesis which shows the significant difference between ideal distribution and historical records.

Also, this graph makes the difference clear that we can see smaller schools get lower proportion of being into the top 16 and larger schools get larger proportion compare to the ideal distribution and the difference of proportion increase when the population size increase for the larger schools.

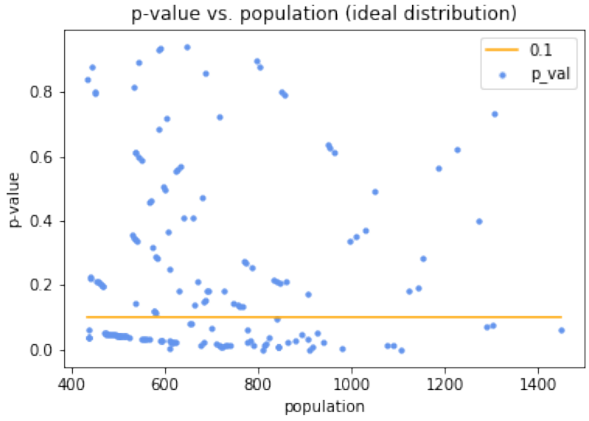

Fig. 5. p-value vs. population

Figure 5 shows the relationship between p-value and population. The x-axis is the average population of the three schools that we combined and the y-axis is the p-value that we calculate from t-score. The orange line shows the 0.05 (one tail on each side) of p-value and the samples below the line reject the null hypothesis which shows the significant difference between ideal distribution and historical records. There are about 47.9% percent of samples get a p-value smaller than 0.05. Thus, there are about half of the samples that reject the null hypothesis and we can conclude that the ideal distribution doesn't fit the historical records very well.

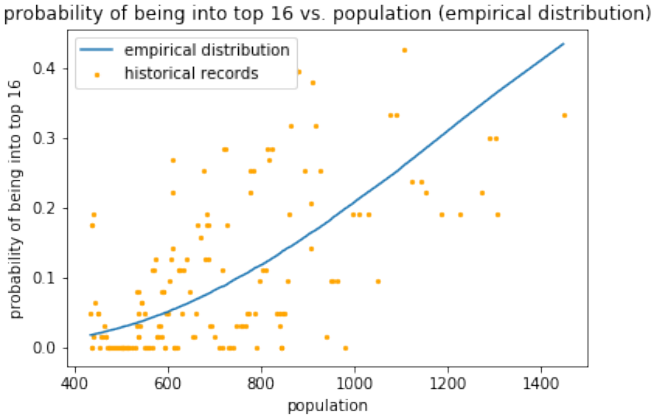

Fig. 6. proportion of being into top 16 vs. population

4) Hypothesis Testing Over the Empirical Distribution and Historical Records: Figure 6 provides a general view of the proportion, the x-axis is the average population of the three schools that we combined and the y-axis is the proportion of being into the top 16 in the state tournament. According to the graph, we can see that the empirical distribution seems to fit the historical records better than the ideal distribution but it doesn't work very well when the school population gets large.

Similarly, we set the hypothesis:

- H_0: the proportion of being into the top 16 in the state tournament from historical records and empirical distribution are the same
- H_a: the proportion of being into the top 16 in the state tournament from historical records and empirical distribution are not the same

Then, we use same formula to do the calculation. For example with the same samples with Mentor, Canton McKinley Senior and Lakota East:

We combine these three schools as one, we can find the average proportion of being into the top 16 from the empirical distribution which is 0.486. Then we can calculate the proportion of being into the top 16 in the historical records which is 0.190. Thus, we can calculate the t-score by applying the formula.

$$
\begin{aligned}
p_0 &= 0.486 \\
p_1 &= 0.190 \\
p &= \frac{0.486 + 0.190}{2} = 0.338 \\
ts &= \frac{p_1 - p_0}{\sqrt{p(1-p)\frac{2}{n}}} \\
&= \frac{0.190 - 0.486}{\sqrt{0.338(1-0.338)\frac{2}{21*3}}} \\
&= -3.51
\end{aligned}
$$

Then we can calculate the p-value by using the t-score and we need to notice that we use two tails test here when we doing the calculation.

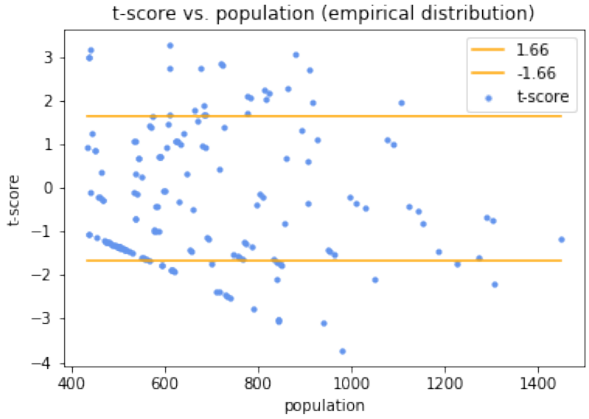

Fig. 7. t-score vs. population

Figure 7 shows the relation between t-score and population. The x-axis is the average population of the three schools that we combined and the y-axis is the t-score from the hypothesis testing. The two orange lines shows the range of t-scores with p-value of 0.1 (two tails) which means the points that are out of the range between the orange lines are the samples that reject the null hypothesis which shows the significant difference between empirical distribution and historical records.

Also, according the graph, we can see that the model fit the proportion of being into the top 16 for small and medium schools well, but most of the large schools get the t-score smaller than 0 which means they have the proportion less than we expect so they being into the top 16 less frequent then we expect.

Figure 8 shows the relationship between p-value and population. The x-axis is the average population of the three schools that we combined and the y-axis is the p-value that we calculate from t-score. The orange line shows the 0.05 (one tail on each side) of p-value and the samples below the line reject the null hypothesis which shows the significant difference between ideal distribution and historical records. There are

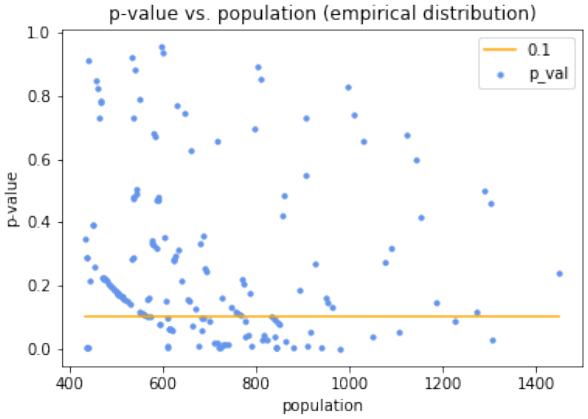

Fig. 8. p-value vs. population

about 30.3% percent of samples get a p-value smaller than 0.05. Thus, it fits the historical records better than the ideal distribution.

5) Conclusion: Compare to the ideal distribution, empirical distribution better predicts small and medium schools, but neither of them predicts the proportion of being into the top 16 in the state tournament of large schools very well. But we can see the empirical distribution has less samples with p-value smaller than 0.05 so I believe that the empirical distribution better predict the historical records than the ideal distribution.

We can conclude that the majority of schools face the the situation that they have the average proportion of being into the top 16 in the state tournament different with the ideal proportion. There are lots of relatively small schools compare to others in Division I, and we have evidence that it is unfair for them as they have an average proportion of being into the top 16 in the state tournament lower than it should be.

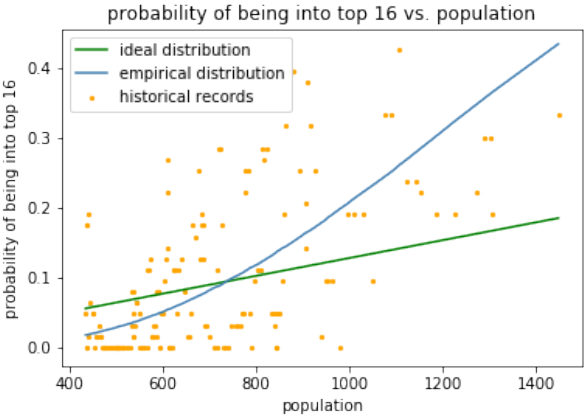

Fig. 9. proportion of being into top 16 vs. population

C. Bayesian Calculation

Bayesian formula allows us to calculation conditional probability. As we already using the simulation technique and hypothesis testing to figure out if the historical records had a significant difference compared to the hypothesis distribution. According to the formula of Bayes' theorem, we have:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

In this part, we are going to calculate which distribution is more likely to generate historical records by using Bayes' theorem. Our calculation has two parts. In the first part, we will calculate the probability of two distributions when the historical records are given. In the second part, we calculate the probability of generating historical records when each of the distribution is given.

Also, we define the probability of having each distribution is 0.5 initially as we have no evidence to support each of them.

1) Probability of Generating the Historical Records by Each Distribution: In this part, A in the formula is the event that a list of schools being into top 16 and B is the event that a distribution is generating the result. Thus, we can define the variables in the formula as follow:

- A: a list of schools being into the top 16 (from historical records)
- B: having a specific distribution (we use B_i and B_e to denote the ideal distribution and empirical distribution)
- A|B: generating a specific list of schools in the top 16 when having a distribution
- B|A: a distribution is used when we see a list of schools being into the top 16 in the state tournament

Notice that we calculate the probability of generating a specific list of schools in the top 16 in the state tournament when we using the empirical distribution (P(A—B_e)) from the simulation result of empirical distribution. Remind that we only count the frequency of having first place for each school rather than the frequency of getting into the top 16. Then we can calculate the probability of having the fist place for each school in the state tournament.

Then, we can apply the similar process when we calculate the ideal distribution in the appendix B section B1. Remind that we use the ratio between schools and total population as the probability of having the first place in the schools remain, but now we use the frequency of having the first place to calculate directly.

For example, if we see Beavercreek and Lakota East in the top 2 in a year (this is the result of 2019[1]), then the probability of this event when we use the empirical distribution is:

$$\frac{8446}{360000} \times \frac{12573}{360000 - 8446}$$

Which 8846 is the frequency for Beavercreek to get first place from the simulation and 360000 is the total number of simulation and $\frac{8446}{360000}$ is the probability of Beavercreek to get first place. Then consider the rest of schools, 12573 is the frequency for Lakota East to get first place from the simulation and 360000-8446 is the rest total number of simulation and $\frac{12573}{360000-8446}$ is the probability of Lakota East to get first place in the rest of schools which is the probability of being in the top 2 in the total schools.

Then we can calculate the probability of generating a specific list of schools in the top 16 in the state tournament when we using the ideal distribution ($P(A|B_i)$) using the similar process in the section 3.1.1. But hear, we only calculate the probability of getting the first place for each school rather than calculate the probability of being into the top 16.

For example, if we see Beavercreek and Lakota East in the top 2 in a year (this is the result of 2019[1]), then the probability of this event when we use the empirical distribution is:

$$\frac{1154}{117216} \times \frac{1309}{117216 - 1154}$$

Which 1154 is the population for Beavercreek and 117216 is the total population of division 1. Then 1309 is the population of Lakota East and 117216-1154 is the rest of population in the division without Beavercreek. Thus, $\frac{1154}{117216}$ represents the probability of getting the first place for Beavercreek and $\frac{1309}{117216-1154}$ represents the probability of Lakota East to get first place.

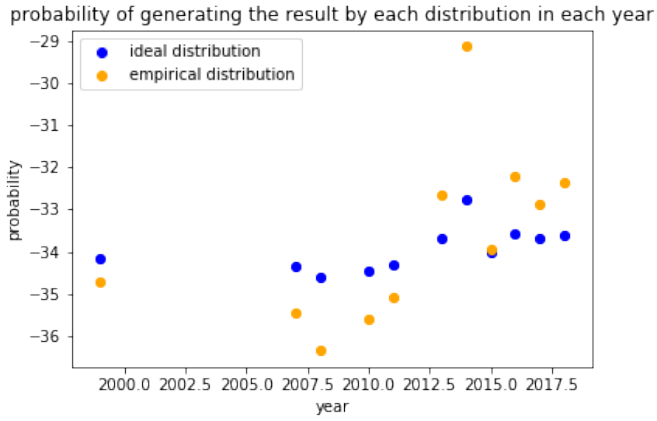

Fig. 10. probability of generating the result by each distribution in each year

Notice the y-axis is rescale by log, so the figures show there is significantly difference between the probability of generating the outcomes in each year by two distribution. In this case, we get similar outcomes than the last, part. The empirical distribution is more likely to generated the outcomes in the recent years, also it is more likely to generate the outcomes with average population larger than about 850.

2) Probability of Having Two Distributions When We Have the Historical Records: In this part, A in the formula is the event that a distribution is generating the result and B is the

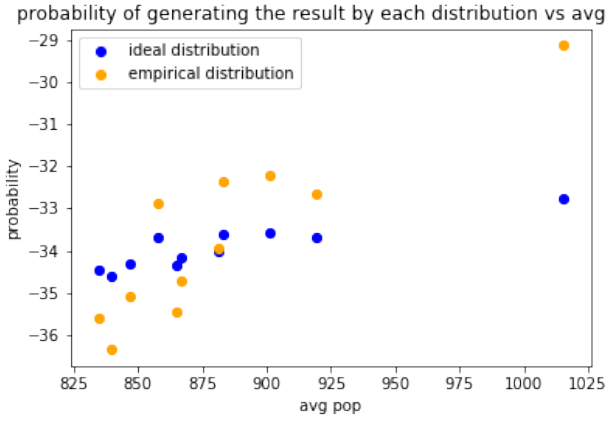

Fig. 11. probability of generating the result by each distribution vs avg pop

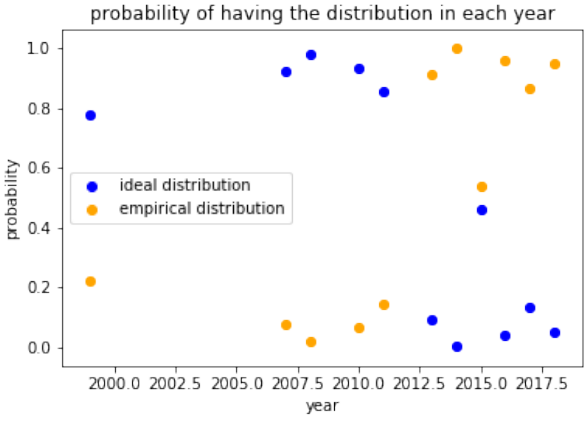

Fig. 12. probability of having the distribution in each year

event that a list of schools being into top 16. Thus, we can define the variables in the formula as follow:

- A: having a specific distribution (we use A_i and A_e to denote the ideal distribution and empirical distribution)
- B: a list of schools being into the top 16 (from historical records)
- A|B: a distribution is used when we see a list of schools being into the top 16 in the state tournament
- B|A: generating a specific list of schools in the top 16 when having a distribution

Notice the process of calculating $P(B|A)$ is shown in the previous section. Also notice that $P(B|A)$ in this section is denoted as $P(A|B)$ in the previous section due to the different definition to the events.

Thus, we can calculate the probability of having empirical distribution based on the information we have which we know a list of schools being into the top 16 in in the state tournament by using the following formula.

$$P(A|B) = \frac{P(B|A_e) \times P(A_e)}{P(B|A_i) + P(B|A_e)} = \frac{P(B|A_e) \times 0.5}{P(B|A_i) + P(B|A_e)}$$

Thus, we know the probability of having empirical distribution is 0.26 when we know Beavercreek and Lakota East are the top 2 schools in a state meet.

Then, we use the same process to calculate the outcomes for top 16 schools in each year.

Figure 12 shows the probability of having two distributions when the historical records are given in each year. The x-axis is the year and the y-axis is the probability of having each distribution. According to the graph, the results from 1999 to 2011 are likely to be generated by the ideal distribution and the results from 2012 to 2019 are likely to be generated by the empirical distribution. Notice that we use the population size of 2019 to do the calculation so it makes sense that the results in recent years are likely to be generated by the empirical distribution.

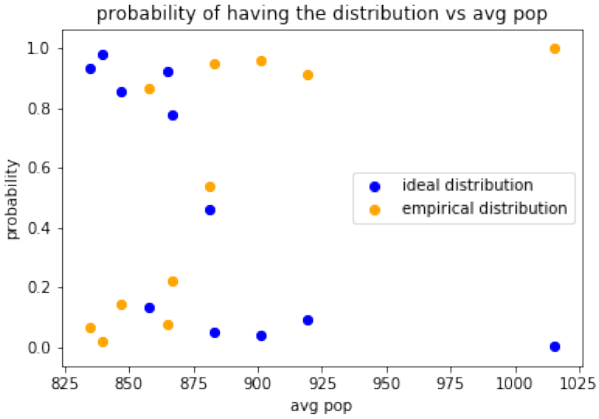

Fig. 13. probability of having the distribution vs avg pop

Figure 13 also shows the probability of having two distributions when the historical records are given in each year. Different with figure 12, we calculate the average population of girls for the schools being into the top 16 in each year and then use it to be the x-axis. So in this figure, x-axis is the average population of girls of the top 16 schools in each year and the y-axis is the probability of having each distribution.

3) Conclusion: Bayesian calculation provide us the analytical evidence that the results in recent years is likely to be generated by the empirical distribution, also the empirical distribution is likely to generate the outcomes in the recent years.

We can see that the average population increased in the recent years, so the empirical is more likely to generate the historical records.

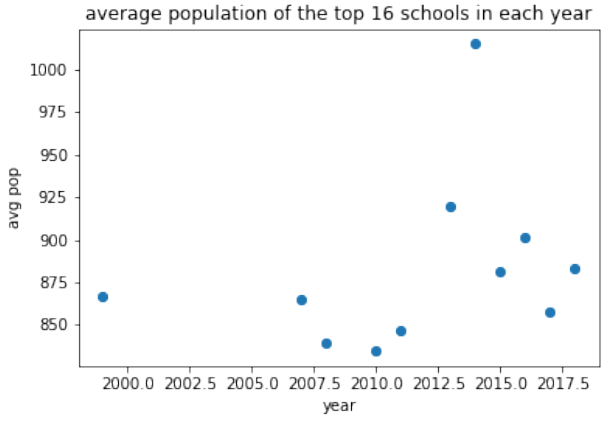

Fig. 14. average population of the top 16 schools in each year

IV. PART II

In this part, we are going to use the technique of order statistics to calculate the probability of having each results in the competition between two or three schools.

A. Introduction of Order Statistics

Order statistics aims to calculate the probability distribution of a variable with order.

1) General Ideas: Suppose we have independent variables $X_1, ..., X_n$ such that $X_1 < X_2 < ... < X_n$. Then, we call X_r the rth order statistic[4]. In our research, the order statistic is defined by the per mile pace of runners, the runner with fastest per mile pace is the first order statistic and the runner with slowest per mile pace is the last order statistic.

For example, suppose A_1, A_2 are the two runners from a school A and A_1 has a per mile pace of 5.8 and A_2 has a per mile pace of 6.0. Then we call A_1 is the first order statistic and A_2 is the second order statistic.

2) Concepts: According to the definition, we are able to calculate the probability density function of a given order statistic.

PDF: The probability density function is used to calculate the probability of a variable having a exact value. Thus, we can express the PDF of rth order statistic (X_r) with population size of n take the value of x can be expressed as the probability of X_{r-1} smaller or equals to x and X_r equals to x and X_{r+1} is grater or equals to x.

$$\begin{aligned} f_r(x) &= Pr\{X_{r-1} \le x \cup X_r = x \cup X_{r+1} \ge x\} \\ &= Pr\{X_{r-1} \le x\} Pr\{X_r = x\} Pr\{X_{r+1} \ge x\} \\ &= nf(x)\binom{n-1}{r-1} F(x)^{r-1}(1-F(X))^{n-r} \\ &= \frac{n!}{(r-1)!(n-r)!} F(x)^{r-1} f(x)(1-F(X))^{n-r} \end{aligned}$$

By using the language in our problem, $F(x)^{r-1}$ calculate the probability of $r-1$ runners run have the per mile pace faster than the current r^{th} runner and $\frac{n!}{(r-1)!(n-r)!}$ calculate the combination of the those $r-1$ runners. Then $(1-F(X))^{n-r}$ calculate the probability of $n-r$ runners run slower than the current runner and then $f(x)$ calculate the probability' of current runner has the per mile pace of x.

B. Sample Calculation

In this section, we use the competition of two schools as an example. Mason is the largest school in Division I and Granville is the smallest and we use these schools as example. We use M_1 and M_2 to denote the runner has the fastest and second fastest average per mile pace in Mason and G_1 and G_2 for Granville.

To make the order statistic concept clear, we plot a probability density curve for top runners from both schools.

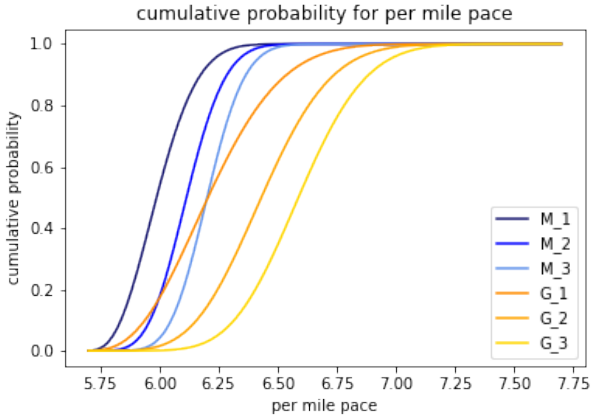

Fig. 15. cumulative probability for per mile pace

Figure 15 shows the cumulative probability for per mile pace for the top three runners from both school. The x-axis is the per mile pace and the y-axis is the cumulative probability of one of the top three runner run at a specific per mile pace.

In this situation, we have 6 possible outcomes totally and we use one of them as an example. Suppose we want to calculate one of the combination $M_1G_1M_2G_3$ which Mason gets first and third place and Granville get the second and the last from the four runners competition between Mason and Granville.

Then we can express it in the way of order statistics which is $M_1 < G_1 < M_2 < G_2$. Then we can simplify again to $M_1 < G_1, M_2, G_2$ and $M_1 < G_1 < (M_2, G_2)$ and $(M_1, G_1, M_2) < G_2$. Thus, we can calculate the PDF of these terms and integrate to get the probability of having this outcome.

$$
\begin{aligned}
&P\{MGMG\} \\
&= \int_0^\infty n_M P(M_1 = x_1) \int_{x_1}^\infty n_G P(G_1 = x_2) \\
&\int_{x_2}^\infty (n_M - 1) P(M_2 = x_3) \int_{x_3}^\infty (n_G - 1) P(G_2 = x_4) \\
&dx_4 dx_3 dx_2 dx_1 \\
&= \int_0^\infty n_M pdf_{M_1}(x_1) \int_{x_1}^\infty n_G pdf_{G_1}(x_2) \\
&\int_{x_2}^\infty (n_M - 1) pdf_{M_2}(x_3)(1 - cdf_{M_3}(x_3))^{n_M - 2} \\
&\int_{x_3}^\infty (n_G - 1) pdf_{G_2}(x_4)(1 - cdf_{G_3}(x_4))^{n_G - 2} \\
&dx_4 dx_3 dx_2 dx_1
\end{aligned}
$$

By calculating the possibilities of all six outcomes, we have the result like this:

outcome	probability
$M_1M_2G_1G_2$	0.64
$M_1G_1M_2G_2$	0.13
$M_1G_1G_2M_2$	0.03
$G_1G_2M_1M_2$	0.04
$G_1M_1M_2G_2$	0.13
$G_1M_1G_2M_2$	0.03

Then, we can calculate the probability of winning for each school. According to the rules of cross country running, we count the score for each team and the team with lowest score will win the competition. The score is counted as follows: the score a school get is equal to the sum of the places that each runner get in the competition. For example, when we have the result $M_1M_2G_1G_2$, then Mason gets $1 + 2$ scores and Granville gets $3 + 4$ scores so Mason win the competition. Then, as there are some outcomes that two school gets the same score, we have another outcome which is tie. But notice that we only have the tie situation when we have even number of runners from each school.

Thus, we can calculate the win rate of each school.

$$
\begin{aligned}
P_{\text{Mason win}} &= 0.64 + 0.13 \\
P_{\text{Granville win}} &= 0.04 + 0.03 \\
P_{\text{tie}} &= 0.03 + 0.13
\end{aligned}
$$

Ideally, we should evaluate the integral from negative infinity to positive infinity. However, to save time, we integrate from 5.7 yo 7.7 as it is very unlikely for a runner to have pace outer the range according to out test.

1) Results: Then we repeat the above process of calculation to calculate the win rate for each school and the probability of tie situation for different number of runners from two and three schools. Then we plot the following graphs.

Figure 16 shows the trend of probability of each results in the two school competition. The x-axis if the number of runners from each school and the y-axis is the probability of each results.

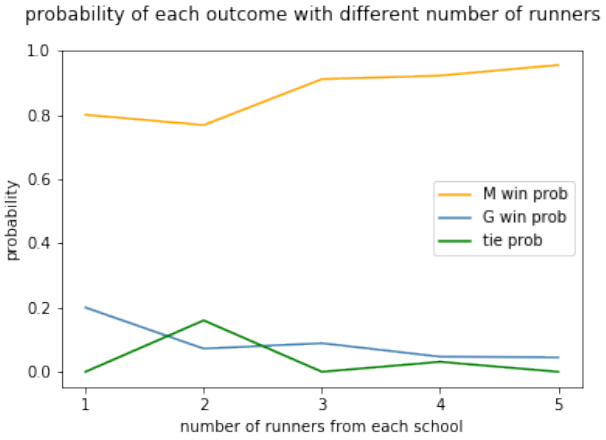

Fig. 16. probability of each outcome with different number of runners in each team

According to the graph, we can see that the larger school have the increased line of score with population in each team increase and it trends that Mason has probability of winning nine times than Granville and the difference will continue increase when the population in each team increase.

We also calculate the probability of each outcomes of three schools competition. Notice that in the situation of three schools competition, we only count the situation that two teams both get lowest score as tie situation. For example we have three schools A, B, C and A, B both get 10 score and C gets 11, then we count this result as a tie situation, but if A gets 10 and B, C get 11, then it will be counted as school A wins.

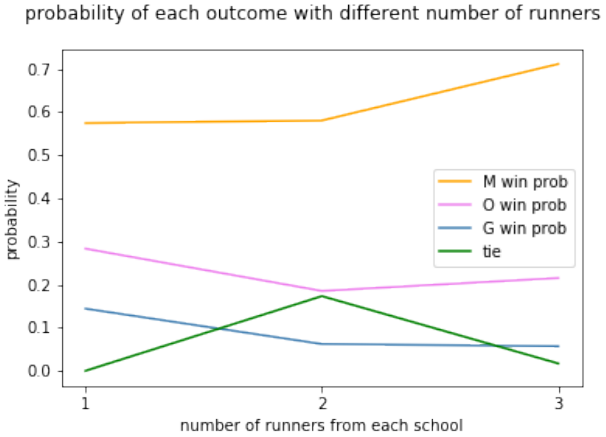

Fig. 17. probability of each outcome with different number of runners in each team

$definewhatisO$

Figure 17 shows the trend of probability of each results in the three school competition. The x-axis if the number of runners from each school and the y-axis is the probability of each results.

The three schools competition shows the similar pattern than the two schools competition result, the probability of being into the top 16 in the state tournament of Mason increase by having more runners in each school and the probability of other two small schools decrease.

2) Conclusion: According to the ideal distribution, the probability of being into the top 16 in the state tournament between two schools competition between Mason and Granville should be $\frac{1728}{1728+433} = 0.80$, but according to the graph, when we have more than two runners in each team, Mason has a probability of being into the top 16 in the state tournament much higher than 80% which lead to unfairness of competition between the two schools. Also we have similar result in the three schools competition. The win rate of Mason should be $\frac{1728}{1728+849+433} = 0.57$, but according to the graph we can see that when each team has more than one runner, Mason has a win rate much higher than we expect which lead to unfairness of competition between the two schools.

Thus, the two and three school competition shows unfairness of the competition to the small schools and this will also be a problem in the real cross country running competition.

V. CONCLUSION

In this research, we verify the result by two ways.

In the first part, we work on the historical records and did hypothesis testing and Bayesian calculation to verify the previous results.

In the section III-B, we did proportion testing on the probability of being into the top 16 in the state tournament from both distributions and the historical records. According to the result, empirical distribution better predicts the probability of being into the top 16 in the state tournament of small and medium schools, but neither of them predicts the probability of being into the top 16 in the state tournament of large schools very well. Also, we the empirical distribution seems closer to the historical records according to the p-value we have.

Also, we can conclude that the majority of schools face the the situation that they have the average probability of being into the top 16 in the state tournament different with the ideal probability. There are lots of relatively small schools compare to others in Division I, and we have evidence that it is unfair for them as they have an average probability of being into the top 16 in the state tournament lower than it should be.

In the section III-C, we calculate which distribution is more likely to generate the historical records and the probability of having each distribution when we have a list of schools being into the top 16 in the state tournament. We can conclude that the results from 1999 to 2011 are likely to be generated by the ideal distribution and the results from 2012 to 2019 are likely to be generated by the empirical distribution. Also, by looking at the average population of the top 16 schools in each year, we can conclude that the empirical distribution is more likely to generate the result of schools with larger population which is also match with the results in the section III-C.

In the second part, we calculate the probability of winning of two and three schools competition. The calculation is based on the empirical distribution as we believe it is more likely to generate the historical records. The result shows that the larger school win more frequent than the ideal win rate. Thus, we are able to conclude that it is not fair for schools with smaller population size of girls.

Thus, in this research, we provide several evidence to support the result from previous research which it is not fair for schools with smaller population size of girls.

APPENDIX

A. Independent test for three schools

Ideally, when the events of each school being into the top 16 is independent, the probability of having two specific schools being into the top 16 should be the multiplication of probability of each school being into the top 16.

$$P(A \text{ and } B) = P(A) \times P(B)$$

As we choose three schools with same population size, by applying the formula, we can calculate the expected frequency of each events and evaluate if they are independent or not by the chi-square test of fitting by using the formula: (we use n to represent the total times of simulation)

- H_0: observed and expected value follows the same distribution
- H_A: observed and expected value follows different distributions

$$X^2 = \sum_{i=0}^{n-1} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i}$$

- schools with population 433, 434 and 437
  We start from some small schools, we choose three smallest schools from Division I which is Granville, Lakeside and Hamilton Township. From the simulation of each single school, three of them have the very similar probability of being into the top 16 so we their average to do the calculation.
  The results shows that a single school being into the top 16 is 6.153%. Then we can use it to calculate the expected result of three schools competition.

$$P(\text{zero school}) = (1 - 0.06153)^3 \times n$$
$$P(\text{one school}) = 0.06153 \times (1 - 0.06153)^2 \times n$$
$$P(\text{two schools}) = 0.06153^2 \times (1 - 0.06153) \times n$$
$$P(\text{three schools}) = 0.06153^3 \times n$$

Thus, we get the expected value from the calculation and we can make the following table:

	G	L	H	GL
observed	1320	1339	1281	78
expected	1301	1301	1301	85
	GH	LH	GLH	None
observed	81	86	6	19809
expected	85	85	6	19837

$$c = \frac{(1320-1301)^2}{1301} + \ldots + \frac{(19809-19837)^2}{19837} = 2.6301$$

Thus, we get the following result

$$X^2(7, N = 24000) = 2.6301, p = 0.9170$$

- schools with population 841, 841 and 841
  Then we choose some schools with midian population size which are Lancaster, Solon and Westerville Central. Then by the same process, we get the following results: A single school in these three being into the top 16 has a probability of 0.1111.

	L	S	W	LS
observed	2135	2162	2095	267
expected	2117	2117	2117	266
	LW	SW	LSW	None
observed	231	253	34	16823
expected	266	266	33	16859

$$X^2(7, N = 24000) = 6.5981, p = 0.4719$$

- schools with population 1305, 1309 and 1728
  Then we choose some schools with midian population size which are Caton McKinley Senior, Lakota East and Mason. Then by the same process, we get the following results:
  Probability of Caton McKinley Senior and Lakota East being into the top 16 is 0.1726 and the probability of Mason is 0.2188.

	C	L	M	CL
observed	2725	2716	3652	558
expected	2678	2678	3594	559
	CM	LM	CLM	None
observed	751	716	131	12751
expected	750	750	156	12835

$$X^2(7, N = 24000) = 8.5273, p = 0.2884$$

All the three results shows the p-value larger than 0.05 which we have no evidence to reject the null hypothesis.

Thus we can conclude that three schools with similar population size of girls being into the top 16 for a school in a year are independent events so we can combine then as one school.

B. Generating two probability distributions

In this part, we are going to describe that how we compute the two distributions.

1) Ideal Distribution: According to the author of previous research, the fair win rate in the state tournament should equal to the ratio between the girl population of the school and the total girl population within the division. Notice that all the data of population of all schools are collected by the author of previous research in 2019[3].

$$\text{Win rate} = \frac{\text{school girl population}}{\text{total girl population in the division}}$$

Thus, we calculate the win rate of each school. We firstly simulate the top school, then we remove that school in the total population then calculate the second school. In the formula below, we use S_x to represent different schools and use a number and a equal sign to represent its place. Then we use x_i for different ids for schools.

$$\mathrm{S}_{id} = place$$

For example $S_0 = 1$ represents the school with ID 0 get the first place.

$$\begin{aligned} P_{\text{top 16}} =& P(S_{x_16} = 16 | S_{x_15} = 15, \ldots, S_{x_1} = 1) \times \\ & \ldots \times P(S_{x_2} = 2 | S_{x_1} = 1) \times P(S_{x_1} = 1) \end{aligned}$$

Then we use p_t to denote total girl population within the division and p_x to denote the girl population of the school x.

Then the probability of having the first place can be calculated.

$$P(S_{x_1} = 1) = \frac{p_{x_1}}{p_t}$$

When we have the probability of the first place, we can calculate the probability of the second place for a school.

$$P(S_{x_2} = 2 | S_{x_1} = 1) = \frac{p_{x_2}}{p_t - p_{x_1}}$$

This formula means that x_2 gets the first place in the rest of schools when we know x_1 already ge the first place.

Then, by repeat this process, we can calculate the probability of each school being into the top 16 in the state tournament. We firstly calculate the probability of a specific school s_1 get the first place then calculate in this situation, the probability of each school get the second place, then by repeat this process sixteen times, we can get the probability of each school being into the top 16.

By using the top 2 schools from 2019 as an example, we have Beavercreek and Lakota East are the top 2 in the state tournament.

The probability of each school gets the first place is

$$\begin{aligned} P(S_{\text{Beavercreek}} = 1) &= \frac{p_{\text{Beavercreek}}}{p_t} \\ &= \frac{1154}{117216} \end{aligned}$$

Then, the probability of Lakota East gets the second place when we know Beavercreek gets the first place its

$$\begin{aligned} P(S_{\text{Lakota East}} = 2 | S_{\text{Beavercreek}} = 1) \\ = \frac{p_{\text{Lakota East}}}{p_t} \\ = \frac{1309}{117216 - 1154} \end{aligned}$$

This is the process we calculate the probability of each school being into the top 2, by changing this two school to each school in the Division I, we can calculate the probability of each school being into the top 2. Then, by using the similar process, we are able to calculate the probability of each school being into the top 16.

2) Empirical Distribution: Rather than calculating the probability of each school being into each place, to generate the empirical distribution, we generate top 16 schools directly by generating the running pace of each runner for each school by using the beta model that introduced in the previous research.

For each school, we generating the same number of running pace with the population of each school and choose the top 5 runners. [3] Then we use the running pace of 5 runners from each school to simulate the result of each tournament and finally generate the result of the state tournament.

Then, by using the frequency of each school being into the top 16 in the state tournament, we can calculate their probability.

ACKNOWLEDGMENTS

The author would like to thank William G. Bowen & Mary Ellen Bowen Research Endowment for funding this work. Thanks Matt Kretchmar, the author of previous paper[3] providing help during the research.

REFERENCES

[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

[3] 5 is the minimum number of runners that is introduced in the previous research