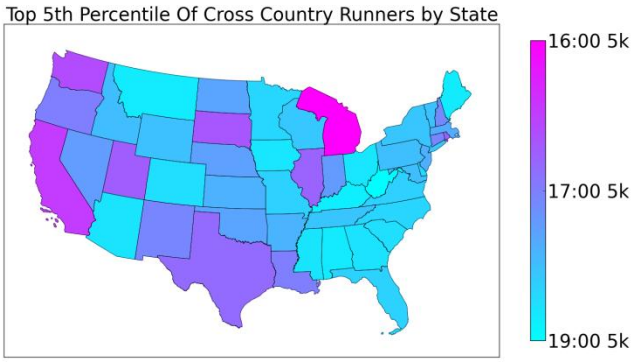
Cross Country: A Country Club Sport?

Dan Kearney

December 14, 2011

Abstract

Some states are consistently faster than others at high school cross country running. To use a recent example, in 2011, a New Jersey boy won the individual title at nationals in San Diego and a team from the same state won the team competition in Oregon. People often speculate that reasons ranging from climate to tradition affect running speeds above all else. Moreover, some suspect that the states that excel at running are simply the wealthiest ones. This paper aims to find if wealthier states are, in fact, faster than poor ones.



1: The states cover a wide distribution of running speeds.

Data

The cross country times used in this paper came from a website called milesplit.us. This website collects data from almost every cross country race in the country and uses that information to display individual rankings, define 'elites', and compare teams. Milesplit also produces articles, videos, interviews, and news about cross country running. Milesplit has 50 sister sites, one for each state. I chose use data from the 5000 meter race because milesplit has data from almost 540,000 athletes participating in the 5k, second to the 3-mile race which has around 84,000 athletes. Since there are probably only a small percentage of runners who have run a 3-mile race without running a 5k, I consider the 5k times race to be a good snapshot of American distance running. The data comes from the boys' 2011 season, which is completely concluded at the time of writing.

Data for sports enrollment comes from the 2010 release of the data of the National Federation of State High School Associations. State wealth data comes from the US Census Bureau for 2010.

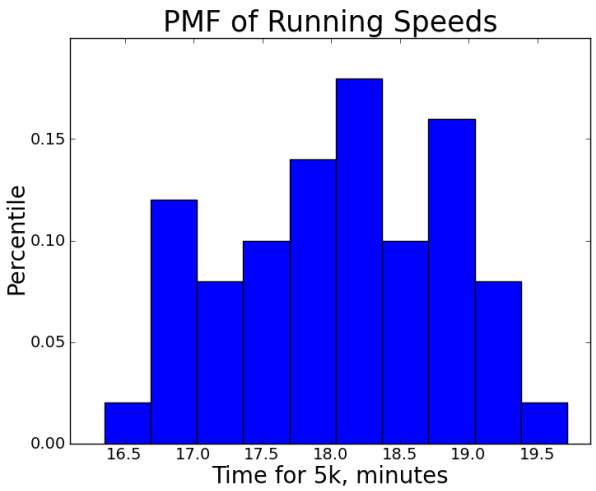
Analysis

In order to compare two states, I developed a metric that would not be skewed by larger states simply having more runners than small states. For example, ranking a state by its top ten runners' times would bias towards states with more runners. One way to compare states is to use percentiles, which are independent of the number of samples in a distribution. In running, it is intuitive to compare times: saying someone has run a five minute mile is much easier to comprehend than someone running 7 miles per hour. A drawback is that the percentiles are backward: typically, someone in the 95th percentile is elite and someone in the 5th is behind, but by comparing times the opposite is true. Regardless, the rest of this paper uses the convention of time, rather than speed, for computing percentiles.

Because I am interested in the truly elite of a state, I compare each state's 5th percentile. The 5th percentile for any given state typically consists of over 100 runners, which is large enough to draw summary statistics from without fear of a few outliers distorting those statistics. I compiled a database consisting of just enough runners' times to calculate the fifth percentile. As a result I only have a slice of the full distribution, and I developed a way to calculate the fifth percentile of that whole distribution given that I have a small slice of the data.

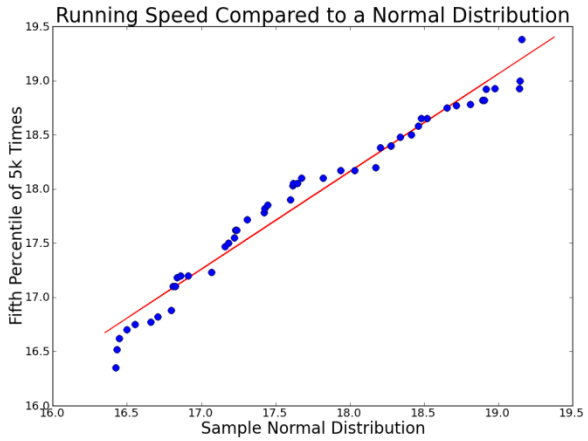
If I have the bottom 10% of the distribution about a state, I know that my slice's 100th percentile corresponds to the 10th percentile of the full distribution. Generally, my slice's top percentile is actually the $\frac{number\ of\ runners\ in\ slice}{total\ number\ of\ runners}$ percentile. Using that ratio, I can exactly calculate the value at any percentile of the full distribution, so long as the ratio is below 1. For example, the fifth percentile of the full distribution is that ratio, 10, times the desired percentile, .05, which equals .5. I can easily check that this is right: the 5th percentile of the full distribution is right smack in the middle of my slice, which is certainly the slice's 50th percentile. Applying this method to every state allows me to grab every state's fifth percentile, which means that I can compare each state by a metric that is not skewed by state size.

The distribution of the collective fifth percentiles from each state is in the PMF below. The 5th percentile varies hugely: the slowest state is about three minutes faster than the slowest.



2: PMF of the nation's 5th percentiles for time.

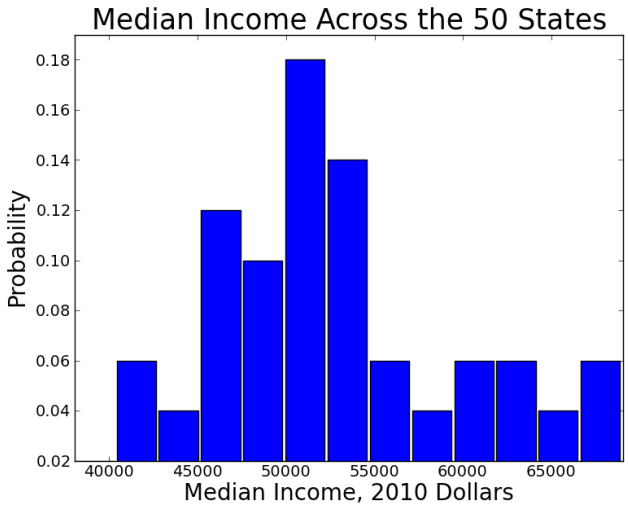
This looks somewhat like a normal distribution because the central peak makes it resemble the familiar 'bell curve' and the mean and median are almost the same, 18.04 minutes and 17.9 minutes. To see if the distribution of running times is normal, I made a simulation of a normal distribution with the same summary statistics as the distribution and plotted it against the actual data below.



3: Running speed fits well to a simulated normal distribution.

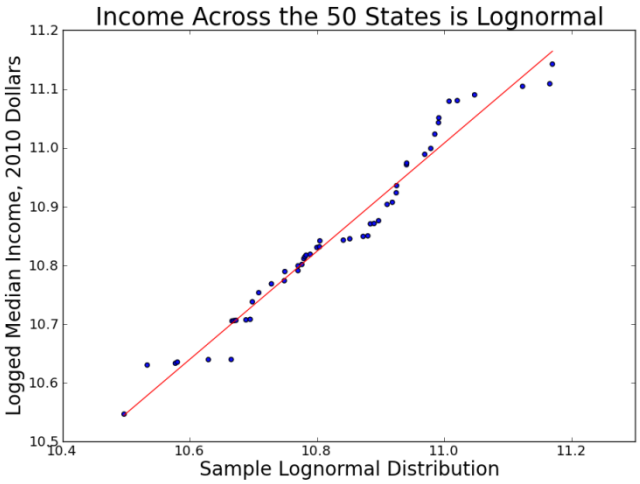
The simulated normal distribution forms a roughly linear relationship with the actual data, which leads me to conjecture that a normal distribution is a reasonable model for this data. The Pearson correlation for this data and a normal distribution is .98. I checked lognormal and uniform distributions as well. The correlation between this data and a lognormal distribution is significantly lower, around .8, and the correlation between this data and a uniform distribution is slightly only lower than that of the normal distribution, around .97. This simply means that the distribution is not well-behaved: it acts both like a uniform distribution and a normal distribution. As such, a normal distribution is a reasonable, though not perfect, model.

My other significant dataset is the median income of each state. I chose to use the median income because medians are better for skewed distributions than means, and medians are well-published. I obtained the median income of each state through the US Census Bureau. I took each state's information and plotted it as a PMF.



4: This PMF shows median wealth across the 50 states is asymmetric.

The shape of this particular graph suggests that the relationship is lognormal because the data has a tail that is skewed to the right. The mean income, $52,156 is significantly larger than the median, $50,846, further suggesting that the distribution is lognormal. To put this postulate to the test, I made a simulation of a lognormal distribution with the same mean and standard deviation and plotted the simulation against the data.

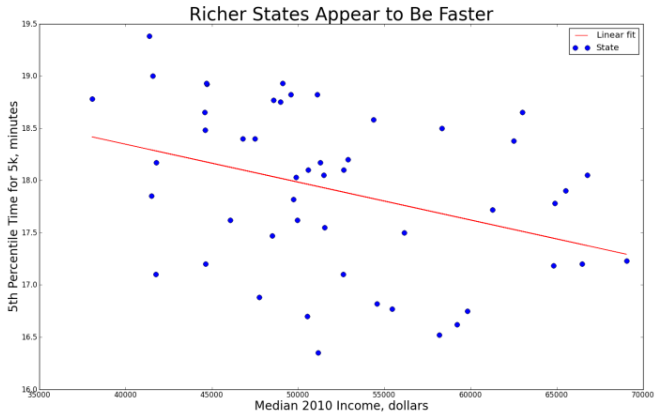


5: The median wealth fits nicely to a lognormal simulation.

The relationship is so close to linear that I can conclude that this is, in fact, a lognormal distribution. No other fit gave such a clean result. This is familiar as it is well-known that wealth distributions are often lognormal, exponential, Pareto, or otherwise skewed.

Wealth and Speed

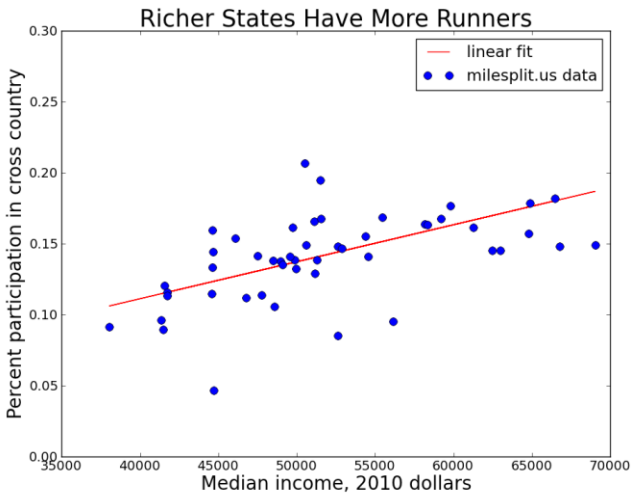
Using data from the US Census Bureau, I made a scatterplot of each state's 5th percentile 5k time against its median wealth. I also calculated the line that minimizes the squared error, which takes the equation $y = 3.64 * 10^{-5}x + 19.79$. This means that the best-fit fifth percentile time drops 21 seconds per $10,000 increase in median income. Given that the range of the data is $30,000, the difference between living in Alabama and New Hampshire might mean your state's best are a minute faster or slower. This is a surprising finding, and has implications on who decides to be a runner. It might mean that athletes with more money are more capable of running; this could mean buying more expensive gear or having more and safer space to run around.



6: There appears to be a correlation between median income and 5th percentile running times.

Other Factors

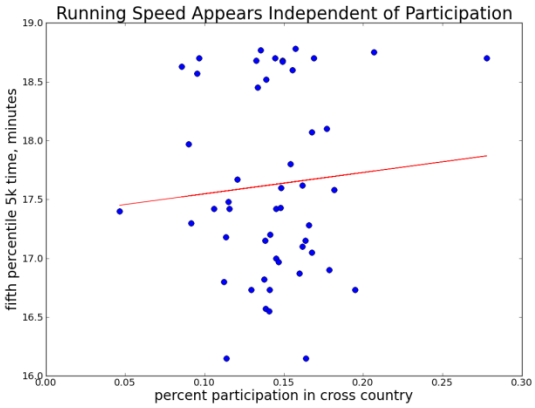
I explored other relationships to try to determine if the wealth trend might be caused by other factors. I found another trend: richer states have a higher percentage of their athletes run track than poor states. This suggests that richer states might simply value running more than other states, and that could be why the above correlation exists.



7: Richer states have a higher proportion of runners than poorer states.

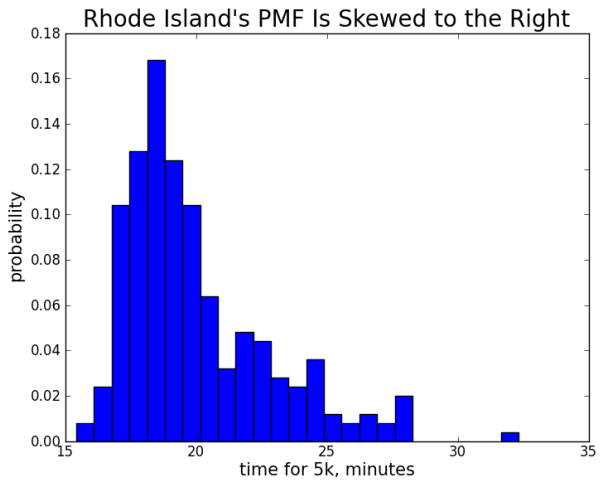
This relationship makes me question the cause of the median-speed relationship. It might be possible that states with higher proportions of runners are faster fundamentally, and income is unrelated to those differences in speed. On the surface, that might be reasonable: if more athletes choose to run cross country, there will be certainly more elite runners on every team.

I made a simulation in order to determine if a high proportion of runners causes a state's elite runners to be faster. Doing so involved re-sampling my milesplit data. I took all of my data from every state and combined them into one pool. Next, I removed any bias towards large states by assuming that each state has the same number of athletes, say 1000. For each state, I pulled the actual proportion of runners from the state's pool of 1000 athletes into a separate pool we call cross country runners; the only variable here is that some states will pull more of their 'athletes' into cross country than others. The plot below shows that in this world, where state size is not a factor and state income is ignored, the proportion of runners does not change the speed of the state.



8: Resampling the distribution entirely on proportion of runners.

I ran the simulation many times, always getting similar results. The fit tends to have a low R^2 value, with a Spearman correlation of around .1, which suggests that the two datasets are barely correlated. The slope of the data is usually low, generally around .05 minutes per percent increase in cross country participation. This simulation shows that the percent participation in cross country does not affect the speed of the runners: after all, a state with more runners is just as likely to draw additional fast ones as slow ones. To explore if that is true, I made a PMF of a state's runners to examine the distribution of runners. I chose a small state, Rhode Island, which I have the entire dataset for.



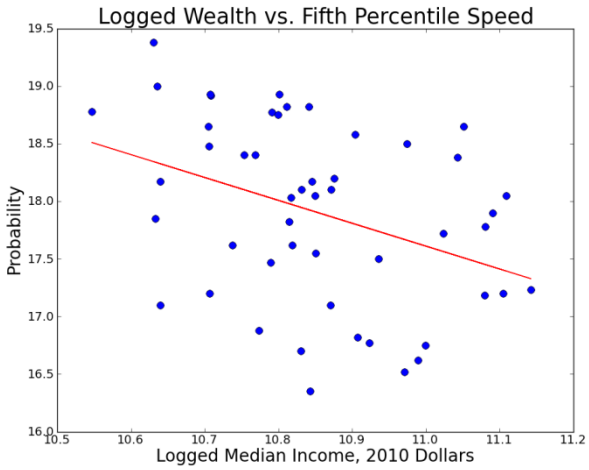
9: Rhode Island's PMF is asymmetric.

This figure, which is consistent with other states in the US, shows that the distribution of runners is actually skewed to the right: a state has many more slow runners than fast. As a result, pulling more runners is more likely to decrease a state's speed than increase it. This test makes me confident that the median income causes a state's speed to change, not its proportion of runners.

Correlation Testing

Although the visual plot of the data is a good way to show a trend in a particular dataset, it does not discuss the quantitative likelihood of my hypothesis. The following section discusses testing my hypothesis against the null hypothesis, that median income does not affect the speed of a state.

The simplest way to quantify the strength of the relationship between two variables is the Pearson correlation, which works best when both fields are drawn from the same distribution. In this case, I determined earlier that the median income approximately forms a lognormal distribution, and the runners' speed can be roughly approximated by a normal distribution. To perform a legitimate Pearson correlation, I must cast the lognormal distribution to a normal distribution by taking the logarithm of every data point. The Pearson coefficient of that transformed data is .39. The graph of that distribution, shown below, visually has a tighter correlation than the lognormal graph.



10: Applying a log transform to the median wealth data makes the relationship clearer.

A ρ value of .39 implies a noisy but significant correlation. The Spearman coefficient of the two lists is .41. Just to compare, the Pearson coefficient of the raw (non-logged) data is .36.

I think that the Spearman coefficient is more accurate. The runners' speeds is somewhere in between a normal distribution and a uniform distribution, and so comparing it a true normal distribution as the Pearson correlation requires is likely to be an approximation. The Spearman coefficient is less affected by this problem of distribution matching.

This correlation, .41, is not particularly high. With a value like this, it is necessary to calculate the probability of a correlation with such a ρ value happening by chance.

Exploring the Null Hypothesis

A simple way to check the probability of the null hypothesis is by re-sampling and counting. If I randomly reassign all of the fifth-percentile times to each state randomly, we can count the probability of the random data having a ρ value greater than or equal to .41. I ran the test over 10000 iterations and found that the resample had such a ρ value only .34 % of the time. This means that the null hypothesis is true .34% of the time by chance; this is a very low probability, and suggests that the null hypothesis is not correct.

A related test searches for the probability that my slope, 21 seconds faster per $10,000 dollars income, would appear by chance. Modifying the above test gave a p-value of about 1%, which further suggests that the null hypothesis is incorrect.

Though these values suggest that the correlation is real, this distribution has a high covariance, and many states might actually be on the wrong side of the best-fit line. Using Bayesian logic, I can determine the probability that a given state will actually have the above behavior; however, it is easier to simply count. For example, simple counting determines that a state that is above the median for wealth has a 60% chance of being above the median for running speed. This means that in its simplest form, the conjecture that faster states are richer states is correct 60% of the time.

Conclusion

These findings have made me confident that the hypothesis suggested, that wealthier states are faster, is true. The significant Spearman's correlation, the inconclusive error testing, and the actually counted 60% probability for the median case all make a strong case that this hypothesis is real. It still leaves the question why. It might be state-of-the-art equipment, better places to run, or more money in the school's programs that causes this; it could be that wealthier people push their children away from contact sports; it could be that the leisure time associated with the upper-middle-class lifestyle simply allows more time for training. Regardless of this correlation, it does not preclude anyone from being successful in running. It is still anyone's game, and though some states might have a statistical edge, the underdog can always win in the end.

References

Data came from http://usa.milesplit.com/, http://www.census.gov/, and the National Federation of State High School Associations.

I learned probability and statistics and used modules from Allen Downey.

Downey, Allen. Think Stats. O'Reilly. 2011.

I used Python's Numpy, Matplotlib, MySQLdb, math, random, sys, csv, itertools, pickle, and Basemap modules.