Unfair from the start? Expert witnesses use statistics to find out if a hiring test disproportionately affected African Americans.

Misuse of Bayesian Statistics in Court

Joseph B. Kadane

© Photographer: Coce Risteski | Agency: Dreamstime.com

Editor's Note: *This paper does not necessarily reflect the opinions of the U.S. Department of Justice.*

The Delaware State Police has set qualifications necessary for a person to be a state trooper. These include U.S. citizenship; residence in Delaware, Maryland, New Jersey, Pennsylvania, or Virginia; education to roughly the level of an associate's degree; a driver's license with no DUI convictions; and no felony or illegal drug use. They also require medical and physical tests and a test of reading and writing skills, called the ALERT test. The Civil Rights Division of the U.S. Department of Justice sued the State of Delaware, alleging that the Delaware State Police's use of the ALERT test on a pass-fail basis disproportionately eliminated African-American candidates for hire as state troopers.

The procedure used by the Delaware State Police was to examine applications to see if the applicant appeared to meet the qualifications and to check for convictions. The written ALERT test was then administered and further investigation and testing were performed on those who passed.

U.S. Department of Justice expert witness Bernard Siskin examined the data on all applicants permitted to take the ALERT examination, dividing them into those who passed and those who failed and those who were African-American and those who were not. This resulted in two-by-two tables, one for each year in question (i.e., 1991 to 1998). Using standard methods, he found African Americans were disproportionately failing the ALERT examination.

Delaware expert witness Elizabeth Becker, an economist, proposed that the relevant comparison should be limited to those whose qualifications were reported correctly on their applications and satisfied the Delaware requirements (i.e., those who would have survived the more in-depth investigation of their qualifications conducted after they passed the ALERT test). She posited that two African Americans and 22 whites met the minimum qualifications, while 30

African Americans and 73 whites did not. However, according to her, there were 258 African Americans and 1,310 whites listed as "questionable" or "unreviewable," whose status could not be determined. Hence, Becker's conception was difficult to implement.

The issue of which is the "right" reference group is a legal one, not a statistical one. While statisticians can observe that Siskin's reference group leads to a straight forward analysis and Becker's does not, the law must determine what the legally relevant question is.

Becker's Analysis

Becker first established she was unable to determine the qualification status of the vast preponderance of the applicants. As test-failures were not investigated further, and since the investigation done after the test required the cooperation of the applicant, this finding is not surprising. To substitute for the lack of such data, she proposed a Bayesian analysis.

The Bayesian calculations she used presupposed a Beta prior distribution with parameters α and β and binomial data with k successes in n trials. The posterior distribution is then a Beta distribution again, with parameters $\alpha+k$ and $\beta+(n\text{-}k)$. In turn, this posterior distribution is used recursively as the prior for the next group of binomial data. Such a recursive use is justified only if each set of binomial data is independent of the others and has the same probability of "success."

The heart of her analysis was to compare the proportion of minimally qualified African Americans among those who took the test to the proportion of minimally qualified African Americans among those who took the test and passed it. Lacking data on the minimal qualifications of those who failed the exam, however, her analysis required some heroic assumptions.

The test-takers were divided into 10 groups according to when they took the test. Below, I trace Becker's steps in analyzing the first group—those who took the exam before July 10, 1993—and, if successful, comprised class 61 in the Delaware Police Academy.

Becker began her analysis with a prior distribution on African-American availability with $\alpha = 8$ and $\beta = 45$, which she computed by rounding up the solutions to a set of equations using a mean of .1416 and a variance of 0.0023—which in turn came from the weighted mean of African-American availability among test-takers in the years under consideration—and a variance large enough to incorporate each of the observed African-American proportions in a 95% interval. Note the question of minimal qualifications is not addressed here.

She then took the numbers of African Americans and whites who took the exam to be in class 61, 41, and 236, respectively, and treated them as binomial data. This led to a new α=41+8=49 and β = 45+236=281. Note that this calculation uses the same data as in the prior (in attenuated form) and again ignores the issue of minimal qualifications.

Next, Becker drew from the 1990 census data relating to the proportion of African Americans who satisfied at least some of the minimal qualifications: residence and age, with an income constraint of $75,000/year. (The thought is that those with incomes greater than $75,000 would be unlikely to be interested in becoming Delaware state troopers.) These figures are 325 African Americans and 3,553 whites. This led to new values for α and β: α=49+325=374 and β=281+3553=3834. Thus, her estimate of the proportion of African Americans among the qualified pool is $\alpha/(\alpha+\beta) = \frac{374}{374+3834} = 8.89\%$.

Presumably, this is to be interpreted as her estimate of the proportion of African-American applicants among minimally qualified applicants.

To continue her analysis, Becker then incorporated the test passers: 21 African Americans and 204 whites, leading to α=374+21=395 and β=3834+204=4038, yielding a mean of $\alpha/(\alpha+\beta)$=395/(395+4038)=8.91% with a standard deviation of

$$\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} =$$

$$\sqrt{\frac{(395)(4038)}{(395+4038)^2(395+4038+1)}} =$$

$$= .0043.$$

She took this, presumably, as the distribution of the proportion of African Americans among minimally qualified test-passers.

Finally, she performed a test of significance, asking whether the number 8.89% is unusual in a Beta distribution with mean 8.91% and standard deviation .0043. Using a normal approximation, she calculated the interval .0889±2(.0043)=(.0803, .0975), which includes .0891. Hence, she concluded the proportion of African Americans among minimally qualified test-passers is not significantly different than the proportion of African Americans among minimally qualified test-takers, so the ALERT test does not adversely impact African Americans.

Becker did several robustness studies to lend support to her conclusions. If the income constraint is relaxed, she found 334 African Americans and 3,943 whites and showed the results are similar. She changed the prior to be uninformative, varied the labor market data to include those with any college credits, and reduced the upper income constraint to $50,000/year. Again, she found similar results. Finally, she reduced the weight given to the labor market data to be equal to the average (over periods) of the applicant flow data and found her conclusion unchanged.

She did analogous calculations for each of the time periods involved to the same effect: no significant differences found. Hence, she concluded that African Americans were not adversely impacted by the use of the ALERT test by the Delaware State Police.

Critique

I was hired by the U.S. Department of Justice to examine Becker's analysis. While I found many matters to comment upon, it is most useful to start with the item that matters most to the results she found.

In Becker's analysis, the Beta distribution with α = 374 and β = 3834 plays two roles. She takes it to be the distribution of the proportion of African Americans among those minimally qualified and as the prior distribution for her distribution of the proportion of African Americans among those minimally qualified and who passed the test. This assumes implicitly that these proportions are the same, which is exactly what she purports to test.

To see how strongly this assumption matters in her analysis, imagine that among the 41 African Americans who took the test, none passed, while among the 236 whites, all passed. This leads to $\alpha = 374 + 0 = 374$ and $\beta = 3834 + 236 = 4070$, yielding a recomputed mean of $\alpha/(\alpha+\beta)=374/(374+4070) = 8.42\%$ and a standard deviation of

$$\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} =$$

$$\sqrt{\frac{(374)(4070)}{(374+4070)^2(374+4070+1)}} =$$

$= .0042.$

Thus, the 95% interval would be $.0889 \pm 2(.0042) = (.0815, .0973)$. As .0842 falls in this interval, Becker would conclude there is no evidence of adverse impact, even though there can be no greater disparity in test results than a 0% passing rate for African Americans and a 100% passing rate for whites. Becker's "test" is no test at all, and provides no information as to whether the exam at issue has an adverse impact on African Americans. (Again, similar results can be obtained for each of the other time periods).

There are, of course, other issues with Becker's analysis:

1. The race data on persons taking the test is used twice: once in the prior and again in the first updating. The race data on persons who passed the test is used thrice: in both of the above and in the last step. This violates the independence assumption that underlies use of the binomial likelihood.

2. The binomial data used in her recursions do not necessarily have the same probability of success; indeed, the very thing at issue is whether they do.

3. The test of significance used at the end is distinctly non-Bayesian. If θ_1 were the proportion of African Americans among minimally qualified test-takers and θ_2 the same among minimally qualified test-passers—and a joint distribution for $(\theta_{1,}\ \theta_2)$ were available—the posterior probability that $\theta_1 > \theta_2$ would be of interest. Equivalently, one could look at the ratio between the conditional probability of passing the test if one were minimally qualified and African American to the same if one were non-African American. Ratios less than one would be indicative of adverse impact.

4. There is no data on the proportions of African Americans and whites among test-failures who meet the minimal qualifications.

Conclusion

As an advocate for Bayesian analysis in court, it is appropriate for me to discuss a case in which I feel Bayesian analysis went awry. The emphasis on subjectivity (everyone has a right to his or her own opinion) puts one in a difficult position to critique truly bizarre assumptions. Yet that certainly will need to happen as more Bayesian analyses are presented to courts.

The court found that the legal theory upon which Becker relied was incorrect (i.e., that the relevant population legally is the population of test-takers, not restricted to those later found to be minimally qualified). Becker's testimony on the basis of her analysis was therefore not presented in court. After a trial centered on the construction of the ALERT test, the court found it does unlawfully disadvantage African Americans.

Comment

Elizabeth Becker

I submit these comments in reply to Joseph B. Kadane's critique of a statistical analysis I submitted for an employment litigation in which we were recently retained by opposing counsel. The criticisms he mounts in his discussion of a small part of my analysis, even if true, would not have altered my conclusions. I tested the sensitivity of my results to the particular criticisms raised by Kadane at the time I submitted my analysis to the court. Although he does not report the specifics of those additional analyses, my results were robust to these criticisms. Moreover, Kadane offered testimony to the court supporting a deeply flawed statistical analysis conducted by another of the plaintiffs' experts, Bernard Siskin. The reliance by the court on that analysis resulted in a disturbing finding by the court.

I was retained by counsel for the defendant in this litigation by the U.S. Department of Justice against the Delaware State Police (DSP). The DoJ alleged the ALERT test of reading and writing proficiency used by DSP to screen applicants had an adverse impact on African-American applicants relative to white applicants (i.e., white applicants passed the test at higher rates than African-American applicants).

The ALERT test was part of a lengthy review process for selecting applicants for entry-level positions. The first step was the completion of an application verifying that each applicant met certain minimum qualifications to be an applicant. These qualifications were education, age, citizenship, and residency, as well as specific standards relating to driving records, drug use, and criminal activity. Applicants who attested to having met these minimum qualifications were screened in a rudimentary fashion to determine whether their representations were true. Those who passed this preliminary screen were permitted to sit for the ALERT exam.

Data regarding test-takers and scores on the test were stipulated by the parties in the litigation. Thus, there was no disagreement as to who sat for the exam and who passed. Test-takers who scored sufficiently high were permitted to continue in the review process to determine whether they possessed additional qualifications—such as physical and medical fitness and satisfactory performance on an oral exam—to be hired. These were not the qualifications necessary to be applicants, but rather additional qualifications necessary to be hired as a state trooper.

In the process of this more lengthy review, it became evident for a large proportion of test-takers that, although they had attested to meeting the specified application standards, they actually lacked the required minimum qualifications to be applicants. Based on analyses of these test-passers, I determined that as many as 40% of the test-takers may have lacked the minimum qualifications to be applicants. Had this been known prior to the administration of the ALERT exam, these persons would not have been considered qualified applicants and would not have been permitted to sit for the exam.

Counsel for the DoJ presented a legal argument that the test caused adverse impact among all African-American applicants allowed to sit for the exam, relative to white applicants who took the exam. They retained Bernard Siskin to prepare a statistical analysis of pass rates for the two groups of test-takers. He found, using chi-square analyses of two-by-two tables, statistically significant differences in pass rates across eight pools of test-takers. The chi-square statistic is known to be sensitive to sample size. The presence of unqualified applicants in the population, thus, clearly has statistical implications.

Yet, Siskin asserted, and was supported in this position by Kadane, that there was no consequence to the inclusion of the unqualified applicants in his analysis. They argued that due to the data stipulation regarding test-takers and test outcomes, those data were adequate for analyzing adverse impact. They also defended the position that, if one were to assume the number of unqualified applicants—whatever that number may have been—was distributed among all applicants proportionately by race, then their presence in the applicant pool would have no impact on the statistical conclusions of adverse impact. Essentially, they adopted a position that sample size has no impact on analysis of two-by-two tables. Siskin made no attempt to correct his analyses for the presence of unqualified applicants.

Counsel for the defendant, DSP, presented a legal argument that an employment practice can have an adverse impact only if it deprives qualified applicants of employment. They argued that a simple comparison of differences in outcomes among a pool of test-takers that included both qualified and unqualified applicants, therefore, could not be the basis for a finding of adverse impact. I was asked by counsel for DSP to rely upon this legal argument and to prepare a statistical assessment of the potential adverse impact of the ALERT exam on qualified applicants. I proposed that the relevant analysis be limited to those who would have been allowed to sit for the exam, had they been qualified applicants. As the qualifications of those who failed the test were not observable, I estimated what the availability of African Americans in the pool of test-takers would have been, had only qualified applicants been allowed to take the test. To develop this estimate of availability, I relied in part on data relating to test-takers. I supplemented that information with data from an assessment of labor market availability of persons meeting the minimum qualifications to be applicants.

Despite the fact that Kadane, himself, has advocated reliance on labor market data to help understand the racial composition of persons available for employment in circumstances where applicant flow data is imperfect, he asserted and continues to assert that such data were not relevant to the question in this matter. The representation of African Americans in the labor market was significantly lower than the representation among test-takers. The labor market availability ranged between about 8% and 9%, depending on particular assumptions. The representation among test-takers ranged from 10.7% to 25.6% across the various pools of test-takers.

Kadane and Siskin explained this large disparity in the availability of African Americans between the labor market and the test-takers as resulting from a higher level of interest in law enforcement among African Americans. Neither offered corroboration for this bizarre assertion. Despite widespread understanding that African Americans have lower levels of educational attainment and higher rates of arrest for driving infractions, drug use, and criminal activity, neither reached the obvious explanation that the difference was attributable to a larger presence of unqualified African Americans among test-takers. Thus, both were comfortable with an assumption that number of unqualified applicants could be distributed among all applicants proportionately by race.

Using these two data sources and Bayesian techniques, I formed an estimate of the availability of qualified African-American applicants, relative to whites. When I compared that availability to representation among test-passers, I found an absence of statistical evidence of adverse impact among qualified applicants. The specific statistical analysis critiqued by Kadane is one among many I prepared to assess potential adverse impact among qualified applicants. His two concerns with this specific analysis are (1) the recursive development of an estimate of the availability of African Americans among qualified test-takers for a particular applicant class violates an assumption of independence because I based my prior for that specific class roughly on the availability across all classes and (2) my comparison of availability among test-takers to representation among test-passers is statistically inappropriate. Kadane fails to report that when I tested the sensitivity of these results (even in my original report) to the specific assumptions in this particular analysis, I found them to be robust.

Consider the calculations for the particular class reviewed by Kadane. When I use a prior roughly informed by availability among test-takers across all classes, combine that with information from the labor market, and then with specific information relating to availability within the particular applicant class, I derive an availability estimate among the qualified applicant pool of 8.89%. Replacing the informed prior with an uninformed one yields an estimate of 8.83%. This is a negligible difference, but actually yields a slightly lower estimate of African-American availability. It is not surprising that moving from a prior based on a wildly variant availability among test-takers with an uninformed prior had little impact on my analyses.

Kadane's concerns regarding this recursive derivation of the availability of African Americans should be allayed with the use of an uninformed prior. The uninformed prior, the labor market statistics, and the test taker data are now completely independent. Now compare this availability with representation of African Americans among test-passers. As reported by Kadane, African Americans are 21 of 225, or 9.33%, of test-passers for this particular class. This is higher than the availability. Kadane decries the method I use to make a statistical comparison of these numbers. Yet, he has not presented—and I cannot imagine—a statistical technique that would support a finding of adverse impact in a circumstance in which there is a higher representation of African Americans among test-passers than among those available to take the test. A higher-than-expected representation of African-American test-passers was observed in five of the eight applicant classes studied when I relied on an uninformed prior. It would take some truly bizarre assumptions to infer from this an adverse impact on African Americans.

His computation that I would have found no adverse impact, even had all the whites passed and all the African Americans failed, is a result for the particular set of assumptions in this single computation that he chose to analyze and does not hold for other pools of test-takers or other analyses I conducted. It is a result of the large weight given the labor market availability and the enormous variance that emerges when the labor market availability is combined with a measure of availability among the test-takers that is so much larger than the labor market availability. Essentially, these two numbers are so at odds with one another that we are left with an extremely imprecise understanding as to what the real availability of qualified African-American applicants may have been. Therefore, it makes sense that almost any observed availability among test-passers would be unsurprising.

Nonetheless, the extreme nature of this particular outcome is disturbing. That is precisely why I prepared alternative analyses in which I significantly reduced the weight of the labor market data in my computations. Despite this change, which eliminates the one strange result that Kadane has winnowed out of context from my entire report, I found no statistical evidence of adverse impact.

I also raised a similar complaint about the use of the chi-square statistic by Siskin. Due to its extreme sensitivity to sample size, I computed that African Americans could have had a pass rate as close as 95% of the rate for whites, and Siskin could still show a statistically significant adverse impact. Yet, Kadane stated that if the assumption that African-American and white test-takers are equally likely to be minimal is true, then I have "no grounds to complain about Siskin's analysis."

In the end, this dispute about the appropriate interpretation of the statistics I presented was moot. The court accepted the legal arguments presented by counsel for the DoJ and granted their motion for summary judgment on the issue of adverse impact. The court's reasoning was that because the DSP allowed applicants to sit for the exam, they were de facto "qualified to sit" for the exam, and thus subject to adverse impact from the exam. It found that the fact that many of them failed to have the minimum qualifications required in order to take the test became irrelevant when the DSP allowed them to take the test. Thus, no data other than the stipulated information regarding pass rates was considered relevant. My analysis, or any other analysis focused on qualifications for that matter, was at that point irrelevant to the court's decision.

In support of this reasoning, the court specifically adopted the argument presented by Siskin and Kadane that the existence of nonqualified test-takers in the pool, however large that number may be, will have no impact on the statistical conclusions about adverse impact. So, we now have a legal ruling that applicants failing to meet the requirements set out by an employer to participate in a hiring process can be adversely affected by a specific step in that process. This is paired with a legal finding that the presence of unqualified persons in an applicant pool is statistically irrelevant to a study of adverse impact, however large their presence may be. In fact, remedial actions requested by the DoJ were actually premised on the inflated measures of hiring shortfalls that emerged from this analysis. I am surprised that Kadane criticizes my contribution in this matter without recognizing the damaging influence of his own.