Did the Michigan Supreme Court Appreciate the Implications of Adopting the “Disparity of the Risk” Measure of Minority Representation in Jury Pools in People v. Bryant?

Joseph L. GASTWIRTH, Wenjing XU, and Qing PAN

Due to an error in the computer program used by Kent County, Michigan, for about 15 months, the African-American proportion of jury pools was about one-half their proportion of the eligible population. Subsequently, a number of defendants appealed their convictions because the jury pool did not fairly represent the community. Although a Federal Court of Appeals found that the statistical evidence helped the defendant establish that African-Americans were under-represented; in a different case, People v. Bryant, the Michigan Supreme Court found the statistics insufficient. The different conclusions arose because the Michigan Court adopted a new criterion that the “disparity of the risk” measure should be at least 0.50. The statistical properties of the measure will be described and it will be seen that it is equivalent to requiring that the Kolmogorov–Smirnov distance between the two relevant distributions is at least 0.50. If one is comparing two normal distributions, with different means and the same variance, the requirement implies that the effect size would need to be at least 1.35 before one could conclude that they differed. Since effect sizes of 0.8 are considered “large,” the criteria used by the Michigan Court are far too stringent and if adopted nation-wide would allow individuals to have trials in front of juries where minorities are substantially under-represented.

KEY WORDS: Binomial model; Disparity of the risk; Effect size; Fair representation; Jury discrimination; Kolmogorov–Smirnov distance.

INTRODUCTION

For a period of 15 months, an error in a computer program led to about half as many African-Americans being summoned for jury service as what would be expected given their percentage of age-eligible residents of Kent County. Several minority defendants whose trials were held during this time appealed their convictions on the basis that their juries were not a “fair representation” of the community. Although the Federal Court of Appeals (6th Circuit), in Ambrose v. Booker, found that the difference between the minority share (4.17%) of the jury pool during the period and their share (8.25%) of the community was sufficient to support the establishment of a prima facie case of under-representation, viewing the same data in People v. Bryant the Michigan Supreme Court said the statistical evidence was insufficient. The Michigan Court adopted a relatively new criterion that the statistical evidence a defendant challenging the representativeness of the jury pool should satisfy: the disparity of the risk must be at least 0.50.

The disparity of the risk (DR) measure is defined for the comparison of two binomial distributions in Section 2 and its use and statistical properties are illustrated. The concept is used to compare two normal distributions in Section 3. It will be seen that a DR of at least 0.50 implies that the effect size or ratio of the difference between the means of two normal to the standard deviation should be at least 1.35, noticeably greater than the value, 0.8, normally considered “large” in the “effect size” literature. In Section 4, the conclusions obtained from the DR criteria adopted by the Michigan Court are compared to inferences drawn from statistically sounder approaches on the data from Kent County and a more recent case, U.S. v. Hernandez-Estrada. It will be seen to be far more stringent than other measures of minority under-representation.

Before discussing the legal cases, it is important to recall that in Duren, the main “fair representation” case, the Supreme Court stated the criteria a defendant needs to satisfy: (1) the group alleged to be under-represented is a distinctive one; (2) the representation of this group on venires from which juries are selected is not fair and reasonable in relation to their number in the community; and (3) this under-representation is due to a systematic exclusion of the group in the jury-selection process, that is, the defendant needs to identify the aspect of the system that causes the under-representation. Statistical evidence is essential to satisfy the second criteria and is often useful for meeting the third. The disparity of the risk measure was proposed as a method for deciding whether the statistical evidence satisfies the second Duren criterion.

STATISTICAL PROPERTIES OF THE DISPARITY OF THE RISK MEASURE

The disparity of the risk (DR) measure was introduced by Detre (1994) to compare two binomial random variables (X and Y) with cumulative distribution functions (cdfs), F(k) and G(k), with different success probabilities, p and π, where p < π, and the same number, n, of trials. In the context of evaluating the representativeness of juries, π is the minority fraction of the age-eligible population in the jurisdiction and is obtained from Census data; p is the minority proportion of individuals.
summoned for jury service during the period under study. For each \( k = 0, \ldots, n \), let \( D(k) = F(k) - G(k) \). The disparity of the risk (DR) measure is the maximum value of \( D(k) \) for all values \( k \leq n \pi r \) and the value of \( k \) at which the maximum difference occurs is denoted by \( k^* \). Notice that the DR measure is related to the Kolmogorov–Smirnov (KS) distance between \( F(x) \) and \( G(x) \), that is,

\[
\sup_x |F(x) - G(x)|.
\]

The only difference is that the DR measure restricts the calculation of the sup to values of \( x = \mu \), where \( \mu = E(Y) \). Because the question of minority under-representation only arises when \( p < \pi \), \( Y \) is stochastically larger than \( X \) (Klenke and Mattner 2010), that is, \( F(k) > G(k) \) for all \( k \) and it is shown in the Appendix that the DR measure is equivalent to the KS distance between two cdfs.

To illustrate the use of the DR measure, assume that a minority group form 20% of the age-eligible community, that is, \( \pi = 0.2 \), and that they form the proportion, \( p = 0.15 \), of individuals called for jury service over a period of time. Because the legal issue concerns the fairness of the system of obtaining jurors, the value of \( p \) should be calculated from data over several months or even a few years, so the sample size, \( n \), is large enough for statistical analysis. After studying grand juries of 23 individuals, Detre (1994) recommended that courts require a DR of at least 0.37, while Re (2012) recommended 0.50 as the critical value from considering 12 member juries. For the situation where \( p = 0.15 \) and \( \pi = 0.20 \), DR = 0.177 when \( n = 12 \) and equals 0.243, when \( n = 23 \). Although minorities are under-represented by 25% in the jury pool, the DR measure is noticeably lower than either the 0.37 or 0.50 thresholds.

The use of statistical hypothesis tests in discrimination cases was criticized by Detre (1994) and King (2007) because small differences can be classified as statistically significant in large samples. The DR measure, however, is also sensitive to the choice of \( n \). Returning to the case, \( p = 0.15 \) and \( \pi = 0.20 \). If \( n = 45 \), the typical size of a venire in Kent County, DR = 0.342 and if \( n = 132 \), the number of individuals called for jury service on the day of Bryant’s trial, DR = 0.55. Indeed, the following lemma shows that for any \( p_0 < \pi \), the value of DR approaches 1 as the sample size goes to infinity.

**Lemma 1.** When testing whether the parameter \( p_0 \) of a binomial random variable equals \( \pi \) versus the alternative \( p < \pi \), DR \( \to 1 \) as \( n \to \infty \) when the alternative hypothesis is true.

**Proof.** Let \( S_n' \) and \( S_n \) be two binomial random variables following \( B(n,\pi) \) or \( B(n,p) \) distributions, respectively, where \( \pi \) and \( p \) are the minority proportion under the null or alternative, respectively. From the jury sample, \( S_n = \sum_{i=1}^{n} X_i \). Applying Hoeffding’s inequality (1963) yields that \( P(X_n - p \geq \varepsilon) = e^{-2n\varepsilon^2} \) for any \( 0 < \varepsilon < 1 - p \), that is, \( P(S_n - np \geq ne) \leq e^{-2n\varepsilon^2} \). Consider \( S_n' \sim Bin(n,\pi) \), where \( 0 < p < \pi < 1 \). For any \( 0 < \varepsilon < p - \frac{\pi + p}{2} \) and \( \delta > 0 \), there exists an \( n_1 \) such that \( e^{-2n_1\varepsilon^2} < \delta \), and \( P(S_n - np \geq ne) < \delta \), for any \( n > n_1 \). Similarly, there exists an \( n_2 \) such that \( e^{-2n_2\varepsilon^2} < \delta \), and \( P(S_n' - np \geq ne) < \delta \) for any \( n > n_2 \). Thus, for any \( \varepsilon < \pi - \frac{\pi + p}{2} \), \( P(S_n' \geq n\varepsilon + \frac{\pi + p}{2}) = 1 - \delta \) and \( P(S_n' \geq n\varepsilon + \frac{\pi + p}{2}) < \delta \), for any \( n > \max(n_1,n_2) \).

Hence, \( \sup_x \{S_n - S'_n\} \geq P(S_n \leq np + \frac{\pi + p}{2}) - P(S'_n \leq np + \frac{\pi + p}{2}) > 1 - 2\delta \), for any \( n > \max(n_1,n_2) \), that is, \( 1 - 2\delta < DR(Bin(n,p), Bin(n,\pi)) \leq 1 \), for any \( \delta > e^{-2n_1\varepsilon^2} \) and \( \delta > e^{-2n_2\varepsilon^2} \). Therefore, as \( n \to \infty \), \( DR(Bin(n,p), Bin(n,\pi)) \to 1 \).

Consequently, the appropriate threshold for the DR measure also depends on the size of the sample studied. Gastwirth and Xu (2014) showed that statistical tests can be adapted for sample sizes typically occurring in discrimination cases by reducing the level required for statistical significance, while ensuring that the test has high power to detect a legally meaningful alternative, for example, by making the Type I and Type II errors equal each other.

**IMPLICATIONS OF THE DR ≥ 0.50 CRITERION FOR THE COMPARISON OF TWO NORMAL DISTRIBUTIONS**

Because the normal is the most frequently used distribution, the KS interpretation of the DR measure will be illustrated by comparing two random variables \( X_1 \) and \( X_2 \), which follow \( N(0,1) \) and \( N(\theta,1) \) (\( \theta \geq 0 \)) distributions, respectively. Thus, \( \theta \) is the difference between the means of the distributions or Cohen’s \( D \) measure of “effect size” as both distributions have variance 1. Denoting the cdf of an \( N(0,1) \) random variable by \( \Phi(x) \), the calculation of the DR measure is facilitated by

**Lemma 2.** The DR measure of the difference between \( X_1 \) and \( X_2 \) is \( 2\Phi(\theta/2) - 1 \), and occurs at \( x = \theta/2 \).

**Proof.** Let \( F(x) \) and \( G(x) \) be the cdf of \( X_1 \) and \( X_2 \), thus \( |F(x) - G(x)| = \Phi(x) - \Phi(x-\theta) = h(x) \). The derivative of \( h(x) = \Phi(x) - \Phi(x-\theta) \), which equals 0 when \( x = \theta/2 \). Thus, \( \sup_x |F(x) - G(x)| = \Phi\left(\frac{\theta}{2}\right) - \Phi\left(-\frac{\theta}{2}\right) = 2\Phi\left(\frac{\theta}{2}\right) - 1 \).

**Comment:** If one plots the density functions of the two normal variables, they cross at \( \theta/2 \), that is, for \( x < \theta/2 \), the density of \( X_1 \) is greater than the density of \( X_2 \) while the reverse holds for \( x > \theta/2 \).

Lemma 2 implies that to satisfy the DR ≥ 0.50 criteria, the difference \( \theta \) between the means of the two distributions is the root of \( 2\Phi(\theta/2) - 1.5 = 0 \), which is \( \theta = 1.349 \). To appreciate the magnitude of the DR ≥ 0.50 criteria, recall that Cohen’s original classifications considered effect sizes of 0.8, 0.5, and 0.2 as large, moderate, or small are used as a guideline in the social science literature (Ellis 2010, p. 41). In the context of social policy and health, McCartney and Rosenthal (2000) noted that researchers seldom see an intervention having a Cohen’s \( D \) of 0.80 and provide examples of practically important interventions with effect sizes less than 0.20. Similarly, Durlak (2009) reported that educational programs designed to improve academic performance are considered important if they have an effect size of around 0.20, so an effect size needs to be interpreted in the subject matter context. Thus, in most areas of social and medical science, many beneficial interventions would not be considered important if the DR ≥ 0.50 criteria adopted by the Michigan Supreme Court or its “effect size” equivalent were applied.

**APPLICATION TO DATA FROM TWO JURY DISCRIMINATION CASES**

Courts use a variety of statistical measures in addition to formal statistical tests when examining the demographic composition of data on jury pools when deciding whether
minorities are fairly represented. When a minority group forms the fraction, \( p \), of the large jury pool and the fraction \( \pi \) of the eligible population, three commonly considered measures are: the absolute disparity (AD) \( p - \pi \), the comparative disparity (CD) \( (p - \pi)/\pi \), and absolute impact (AI) \( AD^* n \), where \( n \) is the size of a typical jury venire, that is, the AI is the average shortfall in the number of minority members in a jury venire. These procedures and the normal approximation to the standard one-sample test for the equality of \( p \) and \( \pi \), referred to as “standard deviation” analysis in many legal opinions, were discussed by Finkelstein (1966), Kaye (1985), Gastwirth (1988), and Finkelstein and Levin (2001). Recently, Gastwirth and Pan (2012) recommended the selection ratio, that is, the ratio of the probability a minority member is called for jury service to the corresponding probability of a majority member, over the use of CD as the CD is the ratio of the probability a minority member is called for jury service to the corresponding probability of all eligible members because the data on all members of the eligible community include the minority population. In the recent Berghuis v. Smith cases, the Court noted that all the measures are “imperfect,” so lower courts can apply the most appropriate one for each case.

At the time of the trials of Bryant and Ambrose, African-Americans formed 8.25% (\( \pi = 0.0825 \)) of the age-eligible residents of Kent County but during the April 2001 to August 2002 period when the erroneous computer was used, they formed only 4.17% (\( p = 0.0417 \)) of the 3898 individuals who served on jury venires. The standard statistical test yields a \( Z = -9.23 \), a highly significant result. The AD = 0.0408, the CD = 0.4945, and the SR = 0.4839, that is, African-Americans had just under one-half the chance of being on a jury venire during this period than Whites. As 45 individuals are on a jury venire, the AI = 1.83 reflecting the fact that a venire of 45 drawn from the community would be expected to have 3.71 African-Americans, nearly twice the average number, 1.88, who served on venires during the period.

In addition to the results of a statistical test, the basic measures of the shortfall in minority members of the jury pool should be considered. Here, the AD = −0.0408, the CD = −0.4945, and the SR = 0.4839. Given the substantial and statistically significant under-representation of African-Americans in Kent County jury pools during the period, the statistics seem sufficient to the Duren requirement. Furthermore, the defendants identified the faulty computer program as the flaw in the system that led to the under-representation and one of their experts, Prof. E. Rothman, used a Cusum chart and related statistical techniques that identified the time when the under-representation started, that is, the new computer program was introduced, and the increase in minority representation when it was replaced.

The recent case, U.S. v. Hernandez-Estrada, in an en banc decision the 9th Circuit realized that requiring an AD of 0.10 was inappropriate when a minority group forms a small percentage of the population and overruled its precedent of exclusively using that criterion. One aspect of the case concerned whether African-Americans were under-represented in jury pools. They formed 5.2% of the age-eligible residents of the Southern District of California, which includes Los Angeles; but only 3.5% of the jury pool. The Court realized that using the existing “an AD of at least 0.10” standard would not find under-representation if African-Americans were never called for jury service, that is, none of the 40,000 or so individuals called for jury service during the relevant period were African-American. In the first appellate opinion, Judge Kozinski calculated the formal statistical test; finding a highly significant disparity (\( Z < -14 \)). Although the AD is only 0.017, the CD is 0.327, and the SR = 0.661. Given the large sample, statistical significance is not surprising; however, the fact that an African-American citizen had just two-thirds the probability of being called for jury service as others should raise a question about the fairness of the system; especially as government guidelines (29 CFR Part 1607) call for employment practices in which the pass rate of minorities is less than 80% that of the majority to be validated as job-related.

In contrast with the inference drawn from the statistical significance of the data and the value of SR, indicating a practically meaningful difference, the DR measure is only 0.125. In fact, because the probability that a random sample of 12 from the eligible community would contain no African-Americans is 0.527, the maximum value of the DR measure is 1–0.527 = 0.473, that is, DR would not reach 0.50 even if no African-Americans had been among the thousands of individuals called for jury service during the period examined. The Hernandez-Estrada decision, however, stated that in addition to demonstrating a statistically significant under-representation, following the criteria the Court established in Duren, the defendant also needs to provide evidence showing how the under-representation resulted from the system used to recruit potential jurors, which was not done in the case.

APPENDIX

Theorem 1. When \( p < \pi \), the disparity of the risk is equivalent to the Kolmogorov–Smirnov distance between the two binomial distributions \( X_1 \sim \text{Bin}(n,p) \), \( X_2 \sim \text{Bin}(n,\pi) \), with distribution functions \( F(x) \) and \( G(x) \), respectively, that is, \( \sup_{x \leq \pi} |F(x) - G(x)| = \sup_x [F(x) - G(x)] = \sup_x |F(x) - G(x)| \).

Furthermore, the supreme is attained when

\[
X = k = \lfloor \frac{n \log \frac{1 - \pi}{1 - p}}{\log \frac{p(1 - p)}{\pi(1 - \pi)}} \rfloor,
\]

where \( \lfloor \cdot \rfloor \) is the floor function.

Proof. The proof is accomplished in two stages. First, the value \( k \) at Kolmogorov–Smirnov distance reaches its maximum will be obtained. Then, it will be shown that \( k \leq nt \).

As noted in Klenke and Mattner (2010), when \( p < \pi \), \( X_1 \) is stochastically smaller than \( X_2 \), that is, \( F(x) \geq G(x) \). Thus,

\[
\sup_x |F(x) - G(x)| = \sup_x [F(x) - G(x)] = \sup_x \left\{ \sum_{j=1}^{x} \binom{n}{j} [p^j(1 - p)^{n-j} - \pi^j(1 - \pi)^{n-j}] \right\}
\]

\[
= \sup_x \left\{ \sum_{j=1}^{x} \binom{n}{j} h(j) \right\} = S(x),
\]
where \( h(j) = p^j(1 - p)^{n-j} - \pi^j(1 - \pi)^{n-j} \). As long as \( h(x) \geq 0 \), \( S(x) \) increases and reaches its maximum at \( k \), where \( h(k) \geq 0, h(k+1) < 0 \). When \( h(j) \geq 0, \)

\[
\left( \frac{p}{1-p} \right)^j (1-p)^n \geq \left( \frac{\pi}{1-\pi} \right)^j (1-\pi)^n
\]

\[
\Rightarrow j \log \left( \frac{p/(1-p)}{\pi/(1-\pi)} \right) \geq n \log \left( \frac{1-\pi}{1-p} \right).
\]

As \( p < \pi, \frac{p}{1-p} < \frac{\pi}{1-\pi} \) or \( \log \frac{p/(1-p)}{\pi/(1-\pi)} < 0 \). Thus (1) is satisfied when

\[
j \leq \frac{n \log \left( \frac{1-\pi}{1-p} \right)}{\log \left( \frac{p/(1-p)}{\pi/(1-\pi)} \right)}.
\]

Denote \( \downarrow \left( \frac{n \log \frac{1-\pi}{1-p}}{\log \frac{p/(1-p)}{\pi/(1-\pi)}} \right) \) by \( \hat{j} \), it follows that \( h(j) \geq 0 \) if \( j \leq \hat{j} \) and \( h(j) \leq 0 \) otherwise. Thus, the maximum of \( S(x) = F(x) - G(x) = \sum_{j=1}^{n} h(j) \) is attained at the point \( k \), where \( k = \hat{j} = \downarrow \left( \frac{n \log \frac{1-\pi}{1-p}}{\log \frac{p/(1-p)}{\pi/(1-\pi)}} \right) \), that is, \( \sup \{ F(x) - G(x) \} = F(k) - G(k) \).

It remains to show \( k \leq n\pi \). Consider the function \( g(\theta) = \theta^n(1-\theta)^{1-\pi} \), where \( \theta \in (0, 1) \). Solving \( \frac{d \ln g(\theta)}{d \theta} = 0 \) yields \( \frac{\pi-\theta}{\theta(1-\theta)} = 0 \), that is, \( \hat{\theta} = \pi \). Furthermore, \( \frac{d^2 \ln g(\theta)}{d \theta^2}|_{\theta=\pi} = -\frac{1}{\theta(1-\theta)} < 0 \). Therefore, \( g(\theta) \) reaches maximum when \( \theta = \pi \), that is, for any \( \theta \in (0, 1) \), \( g(\theta) \leq g(\pi) \). Thus, \( g(p) \leq g(\pi) \), or \( \pi^n(1-p)^{1-\pi} \leq \pi^\pi(1-\pi)^{1-\pi} \) or \( \pi \log \frac{p/(1-p)}{\pi/(1-\pi)} \leq \log \frac{1-\pi}{1-p} \).

As \( \log \frac{p/(1-p)}{\pi/(1-\pi)} < 0 \), \( \pi \geq \frac{\log \frac{1-\pi}{1-p}}{\log \frac{p/(1-p)}{\pi/(1-\pi)}} \Rightarrow \frac{n \log \frac{1-\pi}{1-p}}{\log \frac{p/(1-p)}{\pi/(1-\pi)}} \leq n\pi \), that is, \( \hat{j} \leq n\pi \).

**REFERENCES**

29 CFR Part 1607. (1978), *Uniform Guidelines on Employee Selection Procedures*. [131]

Detre, P. (1994), “A Proposal for Measuring Underrepresentation in the Composition of the Jury Wheel,” *Yale Law Journal*, 103, 1913–1938. [129,130]

Durlak, J. A. (2009), “How to Select, Calculate and Interpret Effect Sizes,” *Journal of Pediatric Psychology*, 34, 917–928. [130]

Ellis, P. D. (2010), *The Essential Guide to Effect Sizes*, Cambridge: Cambridge University Press. [130]

Finkelstein, M. O. (1966), “The Application of Statistical Decision Theory to the Jury Selection Cases,” *Harvard Law Review*, 80, 338–376. [131]

Finkelstein, M. O., and Levin, B. (2001), *Statistics for Lawyers*, New York: Springer-Verlag. [131]

Gastwirth, J. L. (1988), *Statistical Reasoning in Law and Public Policy: Statistical Concepts and Issues of Fairness* (Vol. 1), Orlando, FL: Academic Press. [131]

Gastwirth, J. L., and Pan, Q. (2012), “Statistical Measures and Methods for Assessing the Representativeness of Juries: A Reanalysis of the Data in *Berghuis v. Smith*,” *Law, Probability and Risk*, 10, 17–57. [131]

Gastwirth, J. L., and Xu, W. (2014), “Statistical Tools for Evaluating the Adequacy of the Size of a Sample on Which Statistical Evidence is Based,” *Law, Probability and Risk*, 13, 277–306. [130]

Kaye, D. H. (1985), “Statistical Analysis in Jury Discrimination Cases,” *Jurimetrics*, 25, 274–289. [131]

King, A. G. (2007), “Gross Statistical Disparities as Evidence of a Pattern and Practice of Discrimination: Statistical Versus Legal Significance,” *The Labor Lawyer*, 22, 271–280. [130]

Klenke, A., and Mattner, L. (2010), “Stochastic Ordering of Classical Discrete Distributions,” *Advances in Applied Probability*, 42, 392–410. [130,131]

McCartney, K., and Rosenthal, R. (2000), “Effect Size, Practical Importance, and Social Policy for Children,” *Child Development*, 71, 173–180. [130]

Re, R. M. (2012), “Jury Poker: A Statistical Analysis of the Fair Cross-Section Requirement,” *Ohio State Journal of Criminal Law*, 8, 533–552. [130]