Methods For The Analysis Of Jury Panel Selections: Testing For Discrimination In A Series Of Panels

By PETER W. SPERLICH*
AND MARTIN L. JASPOVICE**

Introduction

In two previous articles, the authors addressed the constitutional requirements for the selection of grand juries[1] and the problem of discovery and proof of the presence of discrimination in the selection of a single jury panel.[2] The present article focuses on testing for discriminatory selection practices in a series of panels.[3] The development of

* Professor, Department of Political Science, the University of California, Berkeley.

** J.D., 1972, University of California; Member, California Bar.

The authors gratefully acknowledge the very helpful comments of Christopher Achen and William Zinn, and the many types of assistance received from the Institute of Governmental Studies and the Department of Political Science of the University of California at Berkeley. Roger Johnson wrote the computer program for the probability computations. Coronet Galloway, Katy Kiggins, Geoffrey Mires, and Titus Toyama assisted our work in various ways. Their help is much appreciated.

1. Sperlich & Jaspovice, *Grand Juries, Grand Jurors and the Constitution*, 1 HASTINGS CONST. L. Q. 63 (1974) [hereinafter cited as *Grand Juries*].

2. Sperlich & Jaspovice, *Statistical Decision Theory and the Selection of Grand Jurors: Testing for Discrimination in a Single Panel*, 2 HASTINGS CONST. L. Q. 75 (1975) [hereinafter cited as *Statistical Decision Theory*].

3. The word "panel" refers to the product of the initial selection of prospective jurors from the eligible population. It corresponds to such terms as "venire," "wheel," and "draw," which are also used by the courts.

The primary focus of the two previous articles was on the grand jury; this is also true here. However, nearly everything that is said about the selection and testing of grand jury panels can also be applied to the selection and testing of trial jury panels. It is only in the step(s) *after* the selection of the panels that grand jury and trial jury selection procedures differ significantly. The selection of the actual grand jurors from among the grand jury panelists tends to proceed in a random fashion, whereas the selection of the actual trial jurors from among the trial jury panelists is subject to the entirely non-random voir dire proceedings.

multi-panel testing methods is of considerable importance because several courts have held that discrimination can be proved only by evidence pertaining to a sequence of panels.

In *Strauder v. West Virginia*,[4] the Supreme Court first articulated the rule that the due process and equal protection clauses of the Fourteenth Amendment[5] proscribe the statutory exclusion of persons from jury service solely because of their race. Thereafter, in *Neal v. Delaware*,[6] the Court ruled that the Fourteenth Amendment also prohibits the racially discriminatory application of facially neutral jury selection statutes. Following *Strauder* and *Neal*, the Supreme Court has held that all jury panels must be representative of a fair cross section of the eligible community from which they are drawn.[7] The principle of cross-sectional representation requires that all eligible persons residing in the community be afforded an equal opportunity to be selected for jury service.[8]

A challenge to the selection procedure employed in choosing grand or petit jury panels can be presented in either or both of two basic forms: (1) the challenge can assail the constitutional validity of the statute or plan which prescribes the manner of selection, asserting that it is incapable of yielding the results demanded by law;[9] or (2) the challenge may concede the constitutionality of the governing statute but attack the manner in which it is applied. Because there are no longer any facially discriminatory jury selection statutes, the second form of challenge prevails today. The most widely utilized method of attacking the discriminatory application of facially neutral selection statutes entails a demonstration of a discriminatory outcome of the selection procedure.[10]

4. 100 U.S. 303 (1879).

5. U.S. CONST. amend. XIV.

6. 103 U.S. 370 (1880).

7. Duren v. Missouri, 99 S. Ct. 664 (1979); Taylor v. Louisiana, 419 U.S. 522, 526 (1975); Glasser v. United States, 315 U.S. 60, 85 (1942); Smith v. Texas, 311 U.S. 128, 130 (1940).

8. As stated by the Court in Smith v. Texas, 311 U.S. 128 (1940): "It is part of the established tradition in the use of juries as instruments of public justice that the jury be a body truly representative of the community. For racial discrimination to result in the exclusion from jury service of otherwise qualified groups . . . violates our Constitution and . . . is at war with our basic concepts of a democratic society and a representative government." *Id.* at 130.

9. For an example of a jury selection statute which is discriminatory on its face, see Strauder v. West Virginia, 100 U.S. 303 (1879).

10. The Supreme Court has consistently drawn a distinction between nonrepresentative selection outcomes which the evidence demonstrates are attributable to chance factors, and are therefore nondiscriminatory, and those attributable to determinative factors, which indi-

The initial burden is on the challenger to produce evidence sufficient to demonstrate what has come to be known as a prima facie case of discrimination.[11] This may be accomplished by presenting a statistical analysis which reflects a significant disparity over a period of time between the percentage of the allegedly excluded class in the community and the percentage of that same class included in the grand jury panel.[12] Once a prima facie case has been presented, the burden shifts to the state to establish that the proven disparity was not the product of discriminatory selection procedures.[13] One commentator explained the concept of the prima facie case as follows:

> Suppose, however, that Negroes have appeared on juries, though in a lower proportion than they occur among the eligible population. Obviously, not all eligible whites *or* Negroes are called for jury duty, and a number of legitimate factors, including chance, enter into the determination of who is called. The fact that a higher proportion of whites than of Negroes is called, therefore, does not itself give rise to an inference of discrimination. To suggest that discrimination is at work, and accordingly to call upon the state to show that it is not, the discrepancy must be great

cate discrimination. Thus, in Smith v. Texas, 311 U.S. 128 (1940), the Court declared that "[c]hance and accident alone could hardly have brought about the listing for grand jury service of so few Negroes from among the thousands shown by the undisputed evidence to possess the legal qualifications for jury service." *Id.* at 131. *See also* Eubanks v. Louisiana, 356 U.S. 584, 587 (1958); Hill v. Texas, 316 U.S. 400, 404 (1942).

11. The initial assertion of the prima facie case principle in the context of jury selection is found in Neal v. Delaware, 103 U.S. 370 (1880): "The showing thus made, including, as it did, the fact (so generally known that the court felt obliged to take judicial notice of it) that no colored citizen had ever been summoned as a juror in the courts of the State, —although its colored population exceeded twenty thousand in 1870, and in 1880 exceeded twenty-six thousand, in a total population of less than one hundred and fifty thousand, —presented a *prima facie* case of denial, by the officers charged with the selection of grand and petit jurors, of that equality of protection which has been secured by the Constitution and laws of the United States." *Id.* at 397. *See also* Duren v. Missouri, 99 S. Ct. 664 (1979) and Castaneda v. Partida, 430 U.S. 482 (1977), the most recent Supreme Court decisions affirming that the demonstration of a significant underrepresentation establishes a prima facie case of discrimination.

12. As stated by the court in People v. Newton, 8 Cal. App. 3d 359, 87 Cal. Rptr. 394 (1970): "While each jury roll or venire need not be a perfect mirror of the community . . . , any substantial disparity, over a period of time, between a group's percentage thereon and its percentage in the eligible population is prima facie evidence of discrimination, regardless of the source of jurors, and shifts the burden to the prosecution to justify the discrepancy." 8 Cal. App. 3d at 390, 87 Cal. Rptr. at 414 (citations omitted).

13. A prima facie case cannot be successfully rebutted by evidence consisting solely of general assertions by the jury selectors that they performed their duty in good faith without regard to racial criteria. Whitus v. Georgia, 385 U.S. 545, 551 (1967); Eubanks v. Louisiana, 356 U.S. 584, 587 (1958); Reece v. Georgia, 350 U.S. 85, 88 (1955); Norris v. Alabama, 294 U.S. 587, 598 (1935).

enough to eliminate chance as a reasonable explanation.[14]

It is thus not generally sufficient for a challenger merely to demonstrate a disparity between a particular group's representation in the population and that group's representation on a single jury. As stated by the Court in *Akins v. Texas*:[15]

> Purposeful discrimination is not sustained by a showing that on a single grand jury the number of members of one race is less than that race's proportion of the eligible individuals. . . . Defendants under our criminal statutes are not entitled to demand representatives of their racial inheritance upon juries before whom they are tried. . . . The mere fact of inequality in the number selected does not in itself show discrimination.[16]

To succeed with a statistical challenge, it is usually necessary to produce evidence of discrimination drawn from the general system of selection: a series of jury panels must be taken into account. Thus, in *People v. Pinell*,[17] where the defendants attacked the composition of the Marin County, California, grand jury panel from which the indicting grand jury was drawn, the appellate court overturned the trial court's finding of discrimination, stating:

> [T]he element of purpose may be deduced from a substantial history of gross inadequacy of representation of the class. . . . When a defendant shows such a history, he has made a prima facie case of discrimination, and the burden shifts to the prosecution to explain and justify the discrepancy. Since this grand jury is the first chosen in Marin County under the present method, there is no such history.[18]

14. Kuhn, *Jury Discrimination: The Next Phase*, 41 S. CAL. L. REV. 235, 252 (1968).

15. 325 U.S. 398 (1945).

16. *Id.* at 403. *See also* Frazier v. United States, 335 U.S. 497, 507 (1948).

17. 43 Cal. App. 3d 627, 117 Cal. Rptr. 913 (1974).

18. *Id.* at 633, 117 Cal. Rptr. at 916-17, citing Smith v. Texas, 311 U.S. 128, 131 (1940). *See* White v. State, 230 Ga. 327, 196 S.E.2d 849 (1973), *cert. denied*, 414 U.S. 886 (1973), where the appellant, who had produced evidence comparing the percentage of blacks in the general population with the percentage of blacks on a single jury panel, was held to have failed to carry his burden of proof. *Compare* Gould v. State, 131 Ga. App. 811, 207 S.E.2d 519 (1974) (where a prima facie case was found to have been established, and where *White* and Estep v. State, 129 Ga. App. 909, 201 S.E.2d 809 (1973) were distinguished in the following pertinent language: "In *White* and *Estep*, in addition to the evidence relative to the grand jury and traverse jury under challenge, '[the] appellant did introduce in evidence one previous grand jury and petit jury list, [but] he nowhere produced any evidence showing the number of negroes, women and young adults thereon.' . . . In the case at bar, the appellant introduced evidence that the percentages previously referred to with reference to the grand jury and traverse jury pools for 1971, are also applicable to the grand jury and traverse jury pools selected in 1960, 1962, 1963, 1965, 1967 and 1969. It was stipulated that there had been no significant change in any degree in the designated groups over the period between the 1960 census and 1970 census." 131 Ga. App. at 818-19, 207 S.E.2d at 525, quoting White v. State, 230 Ga. 327, 331, 196 S.E.2d 849, 853 (1973).

As applied to the particular case, the foregoing reasoning has the highly questionable implication that there can be no remedy for discrimination in the first (and second?) year of a particular jury selection plan, regardless of how severe that discrimination is. *Pinell* would seem to allow a county to perpetuate discriminatory jury selection methods simply by adopting new selection plans with some degree of frequency. However, these arguments aside, there can be no doubt that it is generally desirable and often essential to present serial evidence in challenges to jury selection procedures.

The major drawback for both courts and attorneys applying the prima facie principle is the lack of clear definitions and standards. The courts have not developed a benchmark against which statistical evidence can be measured to determine whether a prima facie case has been established,[19] and the result has been contradictory decisions.[20] Further, no discernable standards have been delineated governing the degree of proof necessary to rebut a prima facie case.[21] Finally, although it is clear that in order to comply with the requirement of demonstrating underrepresentation over time, statistical evidence of discrimination very frequently must be produced for more than one jury panel, whether "over time" contemplates two, three, four or more jury panels has never been made clear. It is the purpose of this article to contribute to a clarification of these issues as well as to present appropriate methods of testing for discrimination in a sequence of jury panels.

The present article is organized as follows. Part I provides an overview of the issues and techniques relating to single-panel discrimination testing, laying the foundation for the more complex concepts and methods required for multi-panel analysis. Part II is the core of this article, addressing the problems of multi-panel analysis and developing the methods for this type of testing. Section A delineates the three basic types of panel sequences, based on the nature and amount of discrimination against the "test group." Sections B, C, and D examine the applicability of the two key testing methods to the three types of panel sequences. The two testing methods are referred to as

19. In Whitus v. Georgia, 385 U.S. 545, 552 n.2 (1967) and Alexander v. Louisiana, 405 U.S. 625, 630 n.9 (1972), the Supreme Court noted the potential applicability of statistical probability analysis to jury selection procedures. However, as yet no decision has been premised upon such an analysis.

20. *See Grand Juries, supra* note 1, at 80-82.

21. While the Supreme Court has held that certain attempts to overcome a prima facie case will be insufficient, *see* note 13 *supra*, no standard delineating the essential elements of the proof necessary to rebut a prima facie case has been established.

the "aggregation method" and the "serial method." The former method is unitary; the latter method has three subtypes. Part II concludes with a section in which the applicability rules are restated in a single array.

Part III of this article applies the methods developed in Part II to large panels. Small panels were utilized in developing the testing methods to allow detailed explication of the computational procedures and the derivation of the various probability figures. Actual jury panels are larger than the examples used in Part II, and Part III demonstrates the application of the testing methods to realistic situations. Whereas Parts I-III focus on the discovery of discrimination in the selection of jury panels, Part IV deals briefly with the legal remedies which are available if discrimination is found. Finally, the Appendix brings a shift of orientation. It shows how *not* to test for discrimination in a series of panels. A previous contribution to the jury selection literature[22] recommended the use of the "product rule" for multiple-panel tests. It will be shown that the use of this rule leads to erroneous results, and that the aggregation and/or serial method should be employed instead.

I. Single-Panel Testing

There are two basic approaches to testing for discrimination in a sequence of jury panels. Before these methods can be considered, however, it is necessary to identify the correct technique for testing any single panel. The single-panel test, which was the subject of a previous article,[23] requires the use of *cumulative*, rather than specific, probabilities. It will be useful to restate briefly the single-panel criteria.

Statutes and Constitutional decisions require that jury panels be a fair cross section of the community from which they are drawn.[24] The Constitution, however, does not require proportional representation of the various identifiable "distinct groups" of the community.[25] It is well that proportionality is not required, for even completely unbiased selection methods do not at all times result in proportionate representation. True random sampling—a completely unbiased selection method which gives everyone in the population the same chance of being se-

22. Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Cases*, 80 HARV. L. REV. 338 (1966).

23. *See* note 2 *supra*.

24. Thiel v. Southern Pacific Co., 328 U.S. 217, 220 (1946); Smith v. Texas, 311 U.S. 128, 130 (1940).

25. Cassell v. Texas, 339 U.S. 282, 286 (1950). For a detailed discussion of these points, see *Grand Juries, supra* note 1, at 63, 68-87.

lected—produces selection outcomes, some of which reflect the identical proportions found in the population and some of which do not.[26]

With the use of random selection procedures, however, the magnitude and likelihood of divergences from true proportionality are known. Basically, small divergences have a greater probability of occurring than large ones. Some large divergences, indeed, are altogether unlikely if the selection process is unbiased. The task is to determine whether the divergence from proportionality of a selection outcome (the jury panel) is small enough to be attributable to the normal fluctuations of random sampling and is thus compatible with the requirement of an unbiased selection. It must therefore be determined how large a divergence is "too large." It would be convenient if a single percentage difference could be specified, so that any difference of, for example, 20 percentage points or larger would constitute evidence that the selection had been biased. This, however, is not possible because the same percentage point difference varies in importance and meaning depending upon context. This will become evident when the following example is considered.

Percentage of Test Group	Case A	Case B
. . . in population	50%	20%
. . . in panel	40%	10%
Percentage point difference	—10 points	—10 points

While the difference is 10 percentage points in either case, it is clear that the problem of underrepresentation is considerably more severe in Case B. In Case B, the test group is underrepresented by 1/2 (50%), whereas in Case A the underrepresentation is only 1/5 (20%).

Similarly, any uniform specification of "excessive divergence" in terms of differences of proportion or percent, rather than of percentage points, is also misleading. As with percentage points, a given difference in proportion or percent may vary in importance and meaning from one context to another. Two contrasting cases will illustrate this point. With fair sampling methods, the underrepresentation by 1/2 in Case A is much less likely to occur than the underrepresentation by 1/2 in Case

26. H. Blalock, Social Statistics 149-54, 179-86, 195-99 (1979); W. Cochran, Sampling Techniques 11-15, 49-54 (1963); L. Kish, Survey Sampling 10-17, 45-47 (1965).

Percentage of Test Group	Case A	Case B
. . . in population	50%	2%
. . . in panel	25%	1%
Difference: proportion	—1/2	—1/2
Difference: percent	—50%	—50%

B. While the divergences are identical in magnitude, Case A is much less likely to be the result of normal sampling fluctuations than is Case B.

Courts have not always been cognizant of the varying importance of identical differences and have tended to give greater weight to larger disparities.[27] This can lead to erroneous interpretations and decisions. As the above illustrations suggest, a difference of 10% in one context can reveal much more discrimination than a difference of 50% in some other context. The answer to this problem is found in the computation of *probabilities*. Probabilities incorporate adjustments for the size of the test group in the population, which is what simple differences fail to do. Probabilities of the same magnitude thus have the same meaning regardless of the population context.

Computed probabilities fall into two classes: significant and non-significant. When testing for discrimination in a jury panel reveals a "significant probability" with respect to a particular group, one must conclude that the disparity between the proportion of that group in the population and its representation in the panel is too large to be attributable to normal sampling fluctuations. A significant probability therefore means that the selection of the panel was biased as to that group. A non-significant probability, by contrast, means that the divergence is likely to be the result of normal fluctuations. It should be noted, however, that non-significance is not proof of a fair selection. Non-significance merely means that in addition to bias as an explanation of the disparity, there exists a possible alternative explanation: chance fluctuation. Bias therefore cannot be positively asserted—at least not on the basis of a statistical challenge.[28]

The most common dividing line between significant and non-sig-

27. *See Grand Juries, supra* note 1, at 63, 80-82.

28. Examination of the jury panel selectors may demonstrate that a relatively small divergence was due to deliberate discrimination. *See id.* at 63, 77.

nificant outcomes is the probability of .05.[29] Occasionally, the less stringent .1 and the more stringent .01 are employed. The use of .05, however, generally is most advisable. Using .05, a probability of .05 or smaller (*e.g.*, .02, .001, .0008) is identified as significant, while a probability larger than .05 (*e.g.*, .057, .08, .2) is identified as non-significant. Specifically, a probability of .05 means that the observed outcome (test group membership in the jury panel) would occur by chance (*i.e.*, by normal sampling fluctuations) only once in twenty such selections of jury panels, or five times in one hundred such selections. This is too small a likelihood to be probable. Chance must be rejected as the explanation, and the existence of bias in the selection of the particular panel must be acknowledged.

The formula for the computation of specific probabilities has been produced and explained in an earlier article.[30] It will be repeated below. First, however, a complication must be considered: when working with large panels, *all* specific probabilities may be significant, including the probability attached to true proportional representation. There exists a large number of possible outcomes with respect to the composition of a large panel. Thus, considering only gender, a panel of 1,000 persons can consist of 1,000 men and zero women, 999 men and one woman, 998 men and two women, down to zero men and 1,000 women. Altogether, such a panel has 1,001 different possible outcomes. With this many potential outcomes, the probability of any one of them occurring becomes quite small, and even the most likely outcome (true proportionality) has only a small likelihood of actually occurring.[31]

The solution to this problem is found in the use of *cumulative* rather than specific probabilities.[32] The use of cumulative probabilities reorients the analysis from single selection outcomes to a *range* of selection outcomes. It identifies the significant part and the non-significant part of that range. Finally, the cumulative approach determines whether the particular panel being tested falls into the significant or the non-significant range. The meaning of a significant result is the same as that discussed above: chance fluctuations cannot serve as an expla-

29. H. BLALOCK, SOCIAL STATISTICS 157-61 (1979); P. JACOBSON, INTRODUCTION TO STATISTICAL MEASURES FOR THE SOCIAL AND BEHAVIORAL SCIENCES 129 (1976); R. RUNYON & A. HABER, FUNDAMENTALS OF BEHAVIORAL STATISTICS 167-69 (1971).

30. *Statistical Decision Theory, supra* note 2, at 75, 83-84.

31. This point emerges clearly in the tables appended to *Statistical Decision Theory, supra* note 2, at 75, 95-112.

32. In fact, it is never correct to use specific probabilities in this type of testing, even though for small panels the results of such usage do not lead to the same proliferation of significant findings.

nation for the divergence between population and jury panel, and bias in the selection must be acknowledged.

Cumulative probabilities are obtained by a summation of the specific probabilities. Beginning with the probability associated with the most discriminatory outcome, the process moves by step-wise additions from the tail to the center of the probability distribution until the specified significance level of, for example, .05 has been reached.[33] The specific probabilities included in the sum of .05 belong to the critical region.[34] The critical region—also known as the region of rejection and as the range of significant outcomes—is always much smaller than the total range of outcomes. Outcomes of relatively high probability, such as proportionality, will not be significant when cumulative probabilities are employed in the test for discrimination.

II. Multiple-Panel Testing

A. Orientation

All testing of sequences of jury panels must begin with the inspection of the specific panels included in the series. The first examination compares the population and panel percentages of the social groups upon which attention is being focused. If discrimination on the basis of gender is suspected, the percentages for women (or for men) will be compared. If racial discrimination has been alleged, the percentages for blacks (or for Chicanos, Asians or whites) will be compared. Similarly, there may be an examination of age, occupational or educational groups.[35]

In principle, the initial comparison of percentages will reveal one of three things: (1) the percentages of the test group in the population and in the panels are identical, in which case no further testing for discrimination is necessary;[36] (2) some or all of the panel percentages of the test group are smaller than the population percentage and those that are not smaller are the same as the population percentage, in which case testing for discrimination becomes necessary; or (3) some of the panel percentages of the test group are smaller and some are larger

33. The present discussion assumes a one-tailed test of significance.

34. For a fuller explanation of the use of cumulative probabilities, see *Statistical Decision Theory*, *supra* note 2, at 74, 85-94.

35. For a discussion and comprehensive listing of cognizable groups, see *Grand Juries*, *supra* note 1, at 73-74.

36. As seen in Part I, the testing is binary. Test group and residual group stand in a reciprocal relationship. If the percentages are identical for the test group, then they also must be for the residual group.

than the population percentage, in which case discrimination tests are also necessary.

The authors have never yet found the first condition to occur, nor is this likely to happen. Even the fairest selection system, a true random method, will not produce proportional representation for each single panel. This condition can thus be ignored. What is likely to occur instead is one of the following three situations:

1. All existing disparities between population and panel percentages are underrepresentations of the test group. This situation will be called *underrepresentations only*.
2. The existing disparities between population and panel percentages include underrepresentations and overrepresentations of the test group. The underrepresentations are more numerous and/or stronger than the overrepresentations. This situation will be called *underrepresentations predominant*.
3. The existing disparities between population and panel percentages include underrepresentations and overrepresentations of the test group. The underrepresentations and overrepresentations are balanced in number and/or strength. This situation will be called *balanced disparities*.

The next three sections will discuss testing for discrimination in each of these situations.

B. Underrepresentations Only

Two basic testing methods are available: serial testing and aggregation testing. The serial method, which appears in three different forms, regards each of the panels as a *separate event* for the purpose of computing probabilities. The aggregation method, in contrast, treats all the panels of the sequence as a *single event* when the probabilities are computed. Both methods have their areas of application.

1. The Ordinary Serial Test

For the purposes of this subsection, it is assumed that the initial comparison of population and panel percentages has shown that all disparities consist of underrepresentations of the test group. The most obvious method that can be applied in this situation is the "ordinary" serial test. It is distinguished from the aggregation test in that it treats the several panels as separate events, and it is distinguished from the other versions of the serial method[37] in that it takes into account disparities of only one type: underrepresentations of the test group.

The ordinary serial test requires the computation of the cumulative probabilities associated with the selection outcome (group composition) of each panel. Each panel is subjected to a single-panel test, as described in Part I. When the separate cumulative probabilities have been obtained, they can be inspected for patterns of discrimination. The computation proceeds in two steps. First, the specific probabilities for the selection outcomes at the "underrepresentation tail" of the distribution are computed. Second, these probabilities are summed, beginning with the most extreme outcome and ending when the sum reaches the chosen significance level (*e.g.*, .05). The outcomes included in this sum constitute the critical region.

The following is the formula for the computation of the specific probabilities:

$$Pr = \binom{N}{r} \left[\left(p^r\right) \left(q^{n-r}\right) \right]$$

The first component of the formula translates into $-\left[\frac{N!}{(r!) \quad (n-r!)}\right]$

and gives the *number of ways* in which a given selection outcome can occur.[38] N stands for the number of persons on the panel, r stands for the number of test group members on the panel, and n-r stands for the number of remaining panel members. The second component—

$$\left[\left(p^r\right) \left(q^{n-r}\right) \right]$$

—gives the *likelihood* of any one of these outcomes occurring.[39] Here *p* stands for the proportion of test group members in the eligible population, *q* stands for the proportion of the rest of the eligible population, and *r* and *n-r* have the same meaning as above. To obtain *Pr*, the specific probability of the selection outcome, the two components of the

37. They are the "differential" and the "two-tailed" serial tests.

38. For example, there are *two* ways to obtain a "3" by throwing a pair of dice: a "1" on the first die and a "2" on the second, and a "2" on the first die and a "1" on the second.

39. For example, either one of the two ways of obtaining a "3" by throwing a pair of dice has a likelihood of 1/36, or .0278.

formula are multiplied by each other.[40]

An example will illustrate the use of the formula. These are the assumed facts:

Blacks on jury panel:	5
Whites on jury panel:	95
Total number on panel:	100
Proportion of black in population:	0.2
Proportion of whites in population:	0.8

The formula thus takes this form:

$$\Pr = \left[\frac{100!}{(5!)\quad(95!)}\right]\left[\left(.2^{5}\right)\left(.8^{95}\right)\right]$$

It is obviously not practical to compute expressions such as "100 factorial" and ".8 to the 95th power" with paper and pen. Nor can this be done with most pocket calculators. A fairly powerful computer is needed, which gives this result: Pr=.000015. This means that the probability of obtaining a panel composed of five blacks and ninety-five whites under the above stated assumptions is fifteen in one million.

To execute the second step and obtain the cumulative probability for the present selection outcome, it is also necessary to obtain the specific probabilities for the more extreme values on the underrepresentation side. These are the results:

Blacks on jury panel: 0	Pr=	.0000000002
1		.0000000051
2		.000000063
3		.00000051
4		.0000031

When these probabilities are summed, including the previously ob-

40. Thus, the probability of obtaining a "3" in one throw of a pair of dice equals (2)(.0278) = .0556.

tained probability of .000015, the cumulative probability of .00001868 results. Rounded, this figure is .00002, or two in one hundred thousand.

It is clear that the summation of the six relevant specific probabilities has produced a sum less than the chosen significance level of .05.[41] The outcome of five black persons on a jury panel of 100 (5%) when blacks constitute 20% of the eligible population is thus *significant*. As explained above, this means that the discrepancy of fifteen percentage points cannot be explained by chance, but must instead be attributed to discrimination.[42] Turning to the testing of a *sequence* of jury panels, it will be remembered that the ordinary serial method first requires the computation of the cumulative probability associated with each of the panels. The determination of whether a single panel is discriminatory is made by the test described in the preceding paragraphs. The determination of whether a sequence of panels is discriminatory is made by examining the pattern of cumulative probabilities, looking particularly for the number of significant results among them.

A set of examples will clarify the decision process. The chosen test period is five years.[43] One jury panel was selected each year. Table 1 presents the other facts assumed for the examples.

TABLE 1: BASIC ASSUMPTIONS FOR EXAMPLES 1-18

For each year of the test period:	
Proportion of women in the population:	.5
Proportion of men in the population:	.5
Number of persons on each panel:	8

Given the assumptions of Table 1, there are nine possible compositions for each panel, ranging from zero women to eight women.[44] The specific and cumulative probabilities associated with each possible composition are given in Table 2.

41. In the present example, the range of significant outcomes actually extends from zero blacks to 13 blacks.

42. For a more detailed account of computational and decision processes, see *Statistical Decision Theory*, *supra* note 2, at 75, 82-94.

43. The question of how many panels should be included in the test period will be considered below.

44. The small number of eight panelists was chosen to permit the display of the probability figures for all possible panel compositions. While there are no jury panels of eight persons, the logic of the analysis is the same when applied to larger panels.

TABLE 2: PROBABILITIES FOR EXAMPLES 1-18

Number of Women	Probabilities: Specific	Probabilities: Cumulative	Interpretation (Significance at .05, one-tailed)
0	.00391	.00391	significant underrepresentation
1	.0313	.03521	
2	.1094	.14461	
3	.2187	.36331	
4	.2734	.63671	
5	.2187	.85541	
6	.1094	.96481	
7	.0313	.99611	
8	.00391	1.00002*	

* The specific probabilities of all possible compositions (selection outcomes) necessarily add to 1.00. The slight deviation from this figure is due to rounding.

Table 3 presents the composition by gender for each of the five panels of the first example.

TABLE 3: PANEL COMPOSITIONS FOR EXAMPLE 1

Panel	Women: Number	Women: Proportion	Men: Number	Men: Proportion
1	1	.125	7	.875
2	1	.125	7	.875
3	0	.000	8	1.000
4	0	.000	8	1.000
5	1	.125	7	.875

Table 4 presents the cumulative probabilities for the gender composition of each of the five panels. It will be remembered that it is the *pattern* of these probabilities which is used in the serial method to determine whether or not a sequence of panels is discriminatory.

TABLE 4: COMPUTATIONAL RESULTS FOR EXAMPLE 1

Panel	Cumulative Probability	Significance (0.05, one-tailed)
1	.03521	$<$ .05
2	.03521	$<$ .05
3	.00391	$<$ .05
4	.00391	$<$ .05
5	.03521	$<$ .05

Each of the results shown in Table 4 has a cumulative probability smaller than the chosen significance level of .05. Each of the results is therefore significant. This means that each of the five differences between the proportion of women in the population (Table 1) and the proportion of women on the panels (Table 3) must be attributed to discrimination.[45] The data show that the selection methods employed during the five years in question discriminated against women. The general determination is this: when each of the separate panels shows discrimination, the sequence taken as a whole is discriminatory.

The determination for Table 4 is easy to understand, but a problem arises if the separate results are not uniformly significant. Example 2 is a case in point:

TABLE 5: PANEL COMPOSITIONS FOR EXAMPLE 2

	Women		Men	
Panel	Number	Proportion	Number	Proportion
1	1	.125	7	.875
2	3	.375	5	.625
3	0	.000	8	1.000
4	2	.250	6	.750
5	1	.125	7	.875

Referring to Table 2, the following results can be found for the second example:

TABLE 6: COMPUTATIONAL RESULTS FOR EXAMPLE 2

Panel	Cumulative Probability	Significance (0.05, one-tailed)
1	.03521	< .05
2	.36331	n.s.
3	.00391	< .05
4	.14461	n.s.
5	.03521	< .05

n.s. = not significant at .05.

Three of the five results demonstrate discrimination against women. The question arises as to whether the sequence taken as a whole demonstrates discrimination. An intuitive answer might be that it does, because a majority of the panels shows discrimination. Although the answer happens to be correct, the decision rule is incorrect. A "majority pattern" is not required; in fact, most "minority patterns" demonstrate discrimination.

The key question is how often a significant cumulative probability would be obtained just by chance, as part of the usual sampling fluctuations. Using the customary decision criterion of a one-tailed significance test at .05, the answer is *one-in-twenty*. A significant result, on the

45. For a discussion of the relationship between differences in proportions and significance, see notes 24-27 and accompanying text *supra*.

average, will occur by chance once in twenty separate panel selections.[46] Thus, obtaining two significant results in twenty selections—or in five selections—demonstrates that the selection system produces divergences between population and panel characteristics in excess of that which can be attributed to normal, chance fluctuations. It follows that the selection method is discriminatory.

2. *Three Issues Related to Length of Test Sequence*

a. **Optimum and Feasibility**

The fact that at a .05 level of significance, an unbiased selection system on the average will produce by chance only one significant outcome in twenty separate selections defines the optimal number of panels for a serial test: *twenty*. A sequence of twenty panels containing no significant outcome or only one significant outcome cannot be interpreted as discriminatory. A sequence of twenty panels must be interpreted as discriminatory if it contains two or more significant outcomes. The one "permissible" significant outcome has an equal likelihood of occurring in any one of the twenty separate selections: the significant result is as likely to occur on the first as on the last selection. This fact has important consequences for the testing of sequences of less than twenty panels.

It is not usually practical, or even possible, to test a sequence of twenty jury panel selections since this normally means a time span of twenty years. Records are not often kept for such a long period of time, so that the characteristics of the panelists may no longer be obtainable. Nor is it possible to collect the required information by interviewing panelists who were called twenty years ago. Mortality and mobility make it impossible to reach enough former panelists to generate an adequate and valid data base. In addition, the laws governing jury selection will probably have changed several times during any period of twenty years. Accordingly, information concerning panelists chosen pursuant to former statutory provisions may not be appropriate for tests to determine the fair or discriminatory nature of the present selection system. Bowing to necessity and relevance, all challenges to jury selection systems will offer evidence to establish a prima facie case that is based on data for a number of jury panels much smaller than twenty. Experience indicates that the maximally feasible number does not nor-

46. It is assumed that the selections are made from the same population and by the same selection methods.

mally exceed five.[47]

b. The Interpretation of One Significant Panel in Five

A problem arises when a sequence of five panels or some other relatively short sequence is tested: the status of one singificant outcome becomes ambiguous. Since fair selection methods produce only one significant outcome in twenty selections, and since this outcome has an equal chance of occurring in any one of the twenty selections, the one "permissible" significant result will occur among the first five (or any five) selections only one time in four. This means that of four sequences of five panels, each containing one significant outcome, *only one* sequence will truly be non-discriminatory (hereinafter referred to as Type A); carrying the sequence to a full complement of twenty selections would produce no additional discriminatory outcomes. The other three sequences, when extended to twenty selections, would produce additional significant results and would thus be discriminatory (hereinafter referred to as Type B).

The problem, of course, is that it cannot be known whether a given sequence of five panels containing one significant result is an instance of Type A or Type B. It could well be argued that any such sequence should be regarded as belonging to Type B since the occurrence of Type B is three times as likely as the occurrence of Type A. Such an interpretation, however, would result in one error, on the average, for each set of four decisions. This error rate would correspond to a significance level of .25, which is too high to be acceptable. The scientifically and statistically proper rule in the present circumstances is to allow one significant result in any sequence of five panels without regarding the sequence as discriminatory.[48] It is clear that this rule increases the bur-

47. This judgment of a five-year maximum is based on the authors' personal experience with data collection efforts for the analysis of jury panel selections.

48. The question of the Type A versus Type B interpretation of one significant result in the sequence requires consideration with any test sequence of less than twenty panels. The relative probabilities of Type A and Type B are different, of course, for each such sequence.

In a sequence of 10 panels, for example, the probability is .5 that the one "non-discriminatory" result will occur in that sequence. Thus, the likelihood that a one-significant-result 10-year sequence belongs to Type A is .5; it is also .5 that the sequence belongs to Type B. An interpretation of the sequence as discriminatory (Type B) would result in one error, on the average, for each set of two decisions. This corresponds to a significance level of .5 and is much too high to be acceptable. The sequence thus must be interpreted as an instance of Type A, allowing one significant result in any sequence of 10 panels without regarding the sequence as discriminatory. The longer the sequence, the greater the error ratio associated with interpreting the sequence as Type B when there is one significant result. The shortest sequence—two panels—is the only one for which a Type B interpretation can be sustained. The probability is .1 that one "non-discriminatory" result will occur in this sequence. Type den of the challenger and gives an advantage to the other side. Fortunately, however, the practical consequences are not particularly severe. When there is discrimination or incompetence in a selection system, it would be quite unusual to find only one significant result in a series of five selections. More typically, all of the results or a great majority of them are significant.

c. Deviations from Sequence Normality

At a .05 level of significance, a sequence of twenty panels will contain one significant result *on the average*. This restrictive phrase has been used before; now its implications must be considered.

At the .05 level of significance, more than one significant difference between population and panel percentages is above chance expectations in a sequence of twenty panels. However, deviations from the expected result can also occur by chance. Any panel sequence has a binomial distribution of significant outcomes; for a panel of twenty, the binomial distribution has an $n = 20$ and $p = .05$. The number of significant outcomes can vary from zero to twenty. It must be determined how many significant outcomes in this sequence are too many: how many more significant panels than one (the "average expectation") can occur just by chance in a sequence of twenty panels. At .05, the answer to *two*.[49] Zero, one, and two significant panels are within the acceptable range of the distribution, using a one-tailed test of significance and a significance level of .05. Three to twenty significant panels fall into the critical region.

The situation is one step removed from, but exactly parallel to, the ordinary testing for significance when percentage differences are subjected to analysis. Both sets of outcomes are subject to normal sampling fluctuations. Both are subject to the same question of how many significant results are too many. In both cases the solution is found by computing and aggregating the probabilities of the specific outcomes in the test tail of the distribution. In both cases outcomes in the extreme end of the tail are too unlikely to have occurred by chance.

B interpretation of such a sequence would result in one error, on the average, for each set of 10 decisions. This corresponds to a significance level of .1, which is quite acceptable under common statistical standards.

A sequence of two panels containing one significant result may therefore be interpreted as Type B (discriminatory). All other sequences—from three to 20 panels—containing one significant result should be interpreted as Type A (discrimination not proven).

49. The computational process for this second-order problem is identical to that used for the determination of the probabilities of selection outcomes. *See* notes 37-42 and accompanying text *supra*.

A comparative display of terms will help to understand the new test situation in light of the more familiar one.

COMPARATIVE DISPLAY

Formula[50] Term	Testing: Panel Composition	Testing: Sequence Composition
N	number of persons on panel	number of panels in sequence
r	number of test group members on panel	number of significant panels in sequence
n-r	number of residual group members on panel	number of non-significant panels in sequence
p	proportion of test group in population	selected significance level (.05)
q	proportion of residual group in population	obverse of selected significance level (.95)

It then can be seen that:

(1) in a sequence of twenty panels, at a .05 level of significance, one significant panel will occur by chance on the average;

(2) in a sequence of twenty panels, at .05, as many as two significant panels may occur by chance.

The previous statement about panels of twenty must therefore be revised. Discrimination can be asserted not when there are two or more significant outcomes, but only when there are three or more.

As was noted earlier,[51] there is little possibility of testing panel sequences of twenty. The important question relates to the implications of the "sequence deviations" for realistic panel sizes. The answer is unexpectedly fortuitous: no adjustments are necessary for sequences of two to seven panels. Table 7 shows the cumulative probabilities associated with the various numbers of significant panels for a variety of panel sequences. It can be seen that for sequences of two to seven panels only the outcome of zero significant panels is non-significant. Outcomes of one or more significant panels are significant. The range of non-significant outcomes changes to zero-and-one for sequences of eight, nine, and ten panels.

50. *See* notes 38-40 and accompanying text *supra*.

51. *See* note 47 and accompanying text *supra*.

TABLE 7: CUMULATIVE PROBABILITIES OF CHANCE OCCURRENCES OF SIGNIFICANT PANELS IN VARIOUS PANEL SEQUENCES*

NUMBER OF SIGNIFICANT PANELS	NUMBER OF PANELS IN TEST SEQUENCE 2	3	4	5	6	7	8	9	10	20
0	.9025	.8574	.8145	.7738	.7351	.6983	.6634	.6302	.5987	.3585
1	.9975	.9927	.9860	.9774	.9672	.9556	.9428	.9288	.9139	.7358
2	1.0000	.9999	.9995	.9988	.9978	.9962	.9942	.9916	.9885	.9245
3		1.0000	.9999	.9999	.9999	.9998	.9996	.9994	.9990	.9841
4			1.0000	.9999	.9999	.9999	.9999	.9999	.9999	.9974
5				1.0000	.9999	.9999	.9999	.9999	.9999	.9997
6					1.0000	.9999	.9999	.9999	.9999	.9999
7						1.0000	.9999	.9999	.9999	.9999
8							1.0000	.9999	.9999	.9999
9								1.0000	.9999	.9999
10									1.0000	.9999
20										1.0000

*One-tailed test; .05 level of significance.

The problem of the correct interpretation of one significant outcome in shorter panels remains since the results for the remaining panels up to twenty are not known. The current issue, however, requires no further adjustment of the decision rules for the kind of panel sequences (two to seven panels) which are likely to be available for testing.

3. The Ordinary Serial Test: Summary

The following are the decision rules for ordinary serial testing by which to determine whether a sequence of jury panels is discriminatory or not in respect to some cognizable group such as women, blacks, or blue collar workers:

DECISION RULES FOR ORDINARY SERIAL TESTS

RULE 1. Disparities between the population and panel proportions of a given group may be due to chance *or* to discrimination when in a sequence of seven panels or fewer, *no or only one* panel result is significant. In such a case, discrimination cannot be positively asserted on the basis of the ordinary serial test.

ADDENDUM TO RULE 1. When one result is significant in a sequence of two panels, discrimination can be asserted if the significance level of .1 is adopted.[52]

RULE 2. Disparities between the population and panel proportions of a given group are due to discrimination when in a sequence of seven panels or fewer, *two or more* of the panel results are significant. In such a case, discrimination in the selection of the panelists must be asserted.

In accordance with these rules, it can be seen that Example 3 (below) requires the conclusion that the selection system employed to choose the five panels is discriminatory.

TABLE 8: PANEL COMPOSITIONS FOR EXAMPLE 3

	Women		Men	
Panel	Number	Proportion	Number	Proportion
1	1	.125	7	.875
2	4	.500	4	.500
3	2	.250	6	.750
4	0	.000	8	1.000
5	2	.250	6	.750

TABLE 9: COMPUTATIONAL RESULTS FOR EXAMPLE 3

Panel	Cumulative Probability	Significance (0.05, one-tailed)
1	.03521	< .05
2	.63671	n.s.
3	.14461	n.s.
4	.00391	< .05
5	.14461	n.s.

n.s. = not significant at .05.

The sequence of panels found in Example 3—and thus the selection

52. *See* note 48 *supra*.

system in use for the five years—must be regarded as discriminatory since there are *two* significant results among the five panels.

The same rules lead to the conclusion that discrimination cannot be positively asserted in respect to the pattern of results found in Example 4:

TABLE 10: PANEL COMPOSITIONS FOR EXAMPLE 4

	Women		Men	
Panel	Number	Proportion	Number	Proportion
1	3	.375	5	.625
2	4	.500	4	.500
3	2	.250	6	.750
4	0	.000	8	1.000
5	4	.500	4	.500

TABLE 11: COMPUTATIONAL RESULTS FOR EXAMPLE 4

Panel	Cumulative Probability	Significance (.05, one-tailed)
1	.36331	n.s.
2	.63671	n.s.
3	.14461	n.s.
4	.00391	< .05
5	.63671	n.s.

n.s. = not significant at .05.

It cannot be asserted that this sequence—and thus the underlying selection system—must be regarded as discriminatory since there is only one significant result in five selections.

4. *The Aggregation Test*

The second basic way to test for discrimination in a sequence of jury panels is to *aggregate* the separate panels into a single set, treating them, in effect, as a single event. Testing for discrimination against women, for example, the method *first* adds the number of women and (separately) the number of men included in the separate panels, and *second* computes the cumulative probability on the basis of these sums.[53] The aggregation method clearly requires fewer probability computations than the serial method; only one cumulative probability is required for a sequence of any number of panels. The need for fewer computations, however, is not a sufficient reason for adopting a statisti-

53. *See* notes 32-34 and accompanying text *supra*.

cal method. The method's conceptual validity, computational results, and significance decisions must be examined.

Conceptually, the aggregation method has excellent standing. Aggregation increases the reliability of the test by increasing the size of the sample, *i.e.*, the number of persons on which the test is based. Two requirements must be satisfied for aggregations of this type to be valid: (1) the selection method must remain the same throughout; and (2) the several selections must be made from the same universe.[54] The first requirement presents no difficulty in the current context. The selection of jury panels is determined by statutes and court-approved jury selection plans. While statutes and plans can and do change, any challenge to a jury selection system will necessarily limit itself to a single framework and will not include panels selected under different statutes and plans.

The second requirement, in essence, means that there should be no significant changes in the sampling universe during the period in which the several samples (panels) are drawn. Two aspects of the requirement can be distinguished: (a) the sampling process should include full replacement; all cases selected in a prior sample should be returned to the universe before the drawing of a subsequent sample; and (b) the basic composition of the sampling universe should not change during the sampling period. The *replacement* condition is generally satisfied in the jury panel selection context. For example, if the selection is made from voter registration lists, panelists of one panel are not excluded from the selection of the next panel.[55] Where selections are made in a less systematic fashion *e.g.*, the "key-man" format[56], frequent repetitions of panelists in numerous court jurisdictions have been observed. Replacement thus appears to be common practice. The few cases of non-replacement that may occur occasionally have no significant effect on the outcome of the discrimination test.

54. The first requirement is implied in the task at hand: to test whether *the particular method employed* by the court at the time of the relevant litigation is discriminatory. The second requirement means that there must not have been a major change in the nature of the jurisdiction, *e.g.*, change of county or district borders, during the time-span of the panel sequence being tested.

55. In some courts, persons who have actually *served* as jurors may be granted an excuse from service for a specified period of time. The number of actual grand jurors relative to the population, however, is exceedingly small. Such excuses therefore do not affect the composition of the universe in any significant way.

56. The key-man selection system vests in judges or jury commissioners the power to choose prospective grand jurors. The so-called "key-men" most often are personal friends or acquaintances of the selectors and are nominated to serve as grand jurors or are asked to supply names of qualified candidates.

The *composition* condition, on first inspection, may appear to present some difficulties. It is obvious that the population of a county or of a federal district does not remain exactly the same for a period of several years. New population cohorts, reaching the minimum legal age for jury service, are added to the universe, while at the same time, mortality subtracts from it. Additionally, geographical mobility both adds to and subtracts members from the relevant population. The effects of mortality and legal maturity, however, are too small to be taken into account given the five-year timespan involved. For example, the proportion of men and women in the eligible population is highly stable, even though some specific men and women will die and some other specific men and women will reach the minimum legal age during that span of years. The important point to note is that the requirement of stable composition does not refer to the perpetuation of specific individuals, but rather to the maintenance of the relative proportions of the various types of persons.[57]

Mobility has the potential for causing important compositional changes. Such changes are not likely to occur in respect to gender. However, with respect to characteristics such as occupation, religion and race, mobility can produce important changes even within a five-year period. Fortunately, it is generally known (or, in any case, it can be ascertained) whether such changes have taken place during the test period.[58] If the changes are fairly small, averages can be employed as indicators of the population proportions. If the changes are very large, it will be advisable to employ the serial method, discussed in the previous section. Barring the highly unusual circumstance of drastic population changes during the test period, the aggregation method is a valid tool for the testing of discrimination in jury selection.

The computational results of the aggregation method and the resulting significance decisions are similar but not fully identical to those of the serial method. This can be shown by a recomputation of the previous examples together with a new example. The serial testing results for the new Example 5 must be given first.

57. To give a different illustration, in a repeated sampling from a standard deck of cards, the sampling results will not be affected if at some point during the sampling period the four aces of the original deck are removed and replaced by the four aces of another standard deck.

58. Not all changes need to be taken into account. Only those changes are relevant which regard population characteristics that are part of the test for and charge of discrimination. For example, if discrimination is charged in respect to gender, changes in religious composition are not relevant.

TABLE 12: PANEL COMPOSITIONS FOR EXAMPLE 5

	Women		Men	
Panel	Number	Proportion	Number	Proportion
1	4	.5	4	.5
2	4	.5	4	.5
3	4	.5	4	.5
4	4	.5	4	.5
5	4	.5	4	.5

These are the results, as obtained from Table 2:

TABLE 13: COMPUTATIONAL RESULTS FOR EXAMPLE 5

Panel	Cumulative Probability	Significance (.05, one-tailed)
1	.63671	n.s.
2	.63671	n.s.
3	.63671	n.s.
4	.63671	n.s.
5	.63671	n.s.

n.s. = not significant at .05.

Using the rules of the ordinary serial testing method, it is clear that the sequence of Example 5, and thus the underlying selection system, cannot be regarded as discriminatory since there is no significant result among the five panel computations.

The computational results, significance decisions, and discrimination determinations of the serial and aggregation methods can now be analyzed. Table 14 presents the comparisons.

59. *See* note 48 and accompanying text *supra*.

TABLE 14: COMPARISON OF ORDINARY SERIAL AND AGGREGATION METHODS

	Ordinary Serial Method		Aggregation Method*						
			Combined Panel Data				Results		
Example	Number of Significant Panels	Discrimination Asserted	N	Proportion	N	Proportion	Cumulative Probability	Significance (.05, one-tailed)	Discrimination Asserted
1	5	yes	3	.075	37	.925	.(8)9733**	< .05	yes
2	3	yes	7	.175	33	.825	.(4)2114	< .05	yes
3	2	yes	9	.225	31	.775	.(3)3398	< .05	yes
4	1	no***	13	.325	27	.675	.0192	< .05	yes
5	0	no	20	.500	20	.500	.5627	n.s.	no

* Common data for aggregation method: proportion of women in population = 0.5; proportion of men in population = .5; number of persons on aggregated panel = 40.

** The number in parentheses, following the decimal point, indicates the number of zeros to be inserted in this place; .(8)9733, consequently 0.(8) is equivalent to .000000009733. This style of presentation avoids long figures and the task of counting zeros.

*** Using the serial method, this case is ambiguous. It is not significant only under the very generous assumption of "no additional significant results in the next fifteen panels."[59]

There is agreement in four out of the five cases. Examples 1, 2 and 3 are clearly significant when the ordinary serial method is applied; they are also clearly significant when the aggregation method is used. Example 5 shows similar agreement on the non-significant side. For Example 4, however, the two methods produce different results.

Example 4 represents the case that is the most difficult to interpret under the rules of the ordinary serial method. There is one significant result among the five tests (panels) of Example 4. The serial method thus interprets the sequence as a whole as *not* significant. This interpretation requires the assumption, however, that fifteen additional panels, selected from the same universe and selected by the same method as the initial five panels, will contain *no* additional significant outcomes. Though the assumption is required under the conservative rules of scientific inference, it is, in fact, quite unrealistic.

The ordinary serial method is the cruder method of the two. It is fully precise only in its testing of the *separate* outcomes (panels). But in its evaluation of *sequences* of outcomes, it relies on somewhat imprecise judgments. The aggregation method, in contrast, enjoys full precision for sequences of outcomes. When the data are clear-cut (Examples 1, 2, 3 and 5) the serial method performs about as well as the aggregation method and produces the same conclusions regarding significance and discrimination. The serial method does not perform as well as the aggregation method when the data patterns are less clear-cut and when the probabilities, whether significant or not, are close to the pre-set significance level.

Since the aggregation method is at least as precise as the ordinary serial method in most circumstances, and more precise in others, it follows that the aggregation method should be used as the primary method of testing for discrimination in a series of jury panels. The ordinary serial method may be used to provide additional information—the probabilities of the single panel results—which may be of interest to the court. In cases of disagreement between the two methods, however, the results of the aggregation method must be used to determine significance and discrimination. The rules for aggregation testing resemble those of single-panel testing—not surprisingly, since the sequence of panels is, in fact, treated as a single event.

DECISION RULES FOR AGGREGATION TESTS

RULE 1. Disparities between the population and panel proportions of a given group may be due to chance *or* to discrimination when, adding across all panels, the total number of persons belonging to that group has a probability of being selected under random assumptions *greater* than the pre-set significance level (*i.e.*, .05). In such a case, discrimination cannot be positively asserted on the basis of the aggregation test.

RULE 2. Disparities between the population and panel proportions of a given group are due to discrimination when, adding across all panels, the total number of persons belonging to that group has a probability of being selected under random assumptions *equal to or smaller* than the pre-set significance level (*i.e.*, .05). In such a case, discrimination in the selection of panelists must be asserted.

5. Concealed Discrimination

A final cautionary note is in order about situations in which *all* of the existing disparities are underrepresentations of the test group. Such situations are highly suspect and most likely reflect discrimination even when not adjudged significant by either testing method. Fair sampling methods (*e.g.*, random sampling) produce symmetric distributions of outcomes.[60] Even the best sampling method does not produce series of outcomes in which all outcomes are identical to the population in respect to some characteristic or value. However, fair sampling methods produce a number of non-disparities as well as a *balanced* set of disparities.

Assuming 50% women in the population, the normal fluctuations of a fair selection method produce a number of outcomes with 50% women and a (larger) number of outcomes which differ from the population characteristic of 50% women. However, if the selection method is truly unbiased and if the number of samples is sufficiently large for testing, about one-half of the disparities will be underrepresentations of women and one-half will be overrepresentations.[61] In certain circum-

60. H. BLALOCK, SOCIAL STATISTICS 156-57, 179-86 (1979); L. KISH, SURVEY SAMPLING 11-14 (1965).

61. Relatively short sequences, such as four or five panels, may not show disparities *exactly* in balance. But even here, the disparities should not be all of one kind *if* a fair selection method was used.

stances neither the aggregation nor the one-tailed serial method will show significance, though there is persistent and exclusive underrepresentation of one group. In such cases, non-significance should not be the end of the investigation; more detailed inquiries into the nature of the court's selection method should be made.

One can imagine, for example, a biased but statistically sophisticated jury commissioner who persistently underselects women exactly to the borderline of non-significance. In the framework of the current examples, that would mean a sequence of panels in which each eight-person panel included three women. Table 2 shows that the outcome of "three women" is not significant for any particular panel. Thus, the one-tailed serial method will not lead to a significant interpretation for the sequence as a whole. For the total sequence there are fifteen women (5 x 3) among forty panelists (5 x 8). Table 15 (below) shows that the outcome of "fifteen women" is not significant when the five panels are treated as a single event. Thus, the aggregation method will not lead to a singificant interpretation for the sequence as a whole.

In certain circumstances, then, neither testing method will lead to the inference of discrimination, though the persistent albeit non-significant underrepresentations are not compatible with a fair selection system and must be the result of a discriminatory method of selecting the panels. Further investigations into the exact process of jury panel selection are clearly in order.

TABLE 15: PROBABILITIES FOR PANEL N = 40*

Number of Women	Probabilities: Specific	Probabilities: Cumulative	Interpretation (Significance at .05, one-tailed)
10	.000771	.00112	
11	.002103	.00322	
12	.005081	.0083	significant underrepresentation
13	.0109	.0192	
14	.0211	.0403	
15	.0366	.0769	
16	.0572	.1341	
17	.0807	.2148	nonsignificant underrepresentation
18	.1031	.3179	
19	.1194	.4373	
20	.1254	.5627	proportionality

* To save space, the distribution is presented in abbreviated form. All outcomes from "no women" to "nine women" are significant.

C. Underrepresentations Predominant

The preceding section investigated sequences in which *all* population/panel disparities were underrepresentations of the test group. The present section will examine sequences which exhibit both types of disparities, but in which the underrepresentations of the test group are more numerous and/or stronger than the overrepresentations of the test group.[62]

1. Application of Ordinary Serial and Aggregation Tests

The first task is to determine how well the ordinary serial and aggregation methods perform in the present context. A set of nine examples (Numbers 6-14) provides the required information. The first three examples (6-8) introduce increasing degrees of numerical predominance of the underrepresentations (3:2, 4:2 and 5:1, respectively). The next three examples (9-11) introduce increasing degrees of "differential strength" of the underrepresentations. The underrepresentations are, in order, 1, 2 and 3 more steps distant from proportionality than the overrepresentations. The last group of three examples (12-14) combines the two forms of increasing predominance. Test sequences of six panels are employed for these examples in order to obtain greater flexibility in showing various degrees of underrepresentation predominance.

62. It is not necessary to extend the analysis to cases of disproportionate *over*-representations of the test group. In binary testing situations, such as that presented here, the overrepresentation of one group necessarily implies an equivalent underrepresentation of the other.

TABLE 16: PROBABILITIES FOR PANEL N = 48*

Number of Women	Probabilities: Specific	Probabilities: Cumulative	Interpretation (Significance at .05, one-tailed)
10	.0000232	.0000309	significant underrepresentation
11	.0000803	.000111	
12	.000248	.000359	
13	.000685	.001044	
14	.00171	.002754	
15	.00388	.006634	
16	.00801	.01464	
17	.0151	.02974	
18	.0260	.05574	nonsignificant underrepresentation
19	.0410	.09674	
20	.0595	.1562	
21	.0793	.2355	
22	.0973	.3328	
23	.1100	.4428	
24	.1146	.5574	proportionality

* To save space, the distribution is presented in abbreviated form. All outcomes from "no women" to "nine women" are significant.

Table 16 presents the probabilities associated with outcomes in panels of forty-eight persons. These figures are required for the aggregation test. Table 17 shows the gender compositions of the six panels included in each example. To save space, only the data for women are given. The data for men are, of course, implied in these figures.

Tables 18-20 present the computational results and the interpretations for Examples 6-14. It can be seen that there are considerable differences in the interpretations (determinations of discrimination). The ordinary serial test leads to the conclusion of discrimination for all but one of the examples. The aggregation test produces this conclusion for only four of the nine examples.

TABLE 17: PANEL COMPOSITIONS FOR EXAMPLES 6–14

Panel	Data for Women	Examples 6	7	8	9	10	11	12	13	14
1	Number	8	8	8	5	5	5	7	7	7
	Proportion	1.00	1.00	1.00	.625	.625	.625	.875	.875	.875
2	Number	6	6	3	5	5	5	5	5	2
	Proportion	.750	.750	.375	.625	.625	.625	.625	.625	.250
3	Number	4	3	2	5	5	5	4	2	1
	Proportion	.500	.375	.250	.625	.625	.625	.500	.250	.125
4	Number	2	2	1	2	1	0	2	1	1
	Proportion	.250	.250	.125	.250	.125	.000	.250	.125	.125
5	Number	1	1	1	2	1	0	1	1	0
	Proportion	.125	.125	.125	.250	.125	.000	.125	.125	.000
6	Number	0	0	0	2	1	0	0	0	0
	Proportion	.000	.000	.000	.250	.125	.000	.000	.000	.000
1–6	Number	21	20	15	21	18	15	19	16	11
	Proportion	.438	.417	.313	.438	.375	.313	.396	.333	.229

TABLE 18: COMPUTATIONAL RESULTS FOR EXAMPLES 6–8

Panel	Example 6 Cumulative Probability	Example 6 Significance (.05, one–tailed)	Example 7 Cumulative Probability	Example 7 Significance (.05, one–tailed)	Example 8 Cumulative Probability	Example 8 Significance (.05, one–tailed)
			ORDINARY SERIAL COMPUTATIONS			
1	1.00	n.s.	1.00	n.s.	1.00	n.s.
2	.96481	n.s.	.96481	n.s.	.36331	n.s.
3	.63671	n.s.	.36331	n.s.	.14461	n.s.
4	.14461	n.s.	.14461	n.s.	.03521	< .05
5	.03521	< .05	.03521	< .05	.03521	< .05
6	.00391	< .05	.00391	< .05	.00391	< .05
			ORDINARY SERIAL SEQUENCE INTERPRETATION			
1–6	significant		significant		significant	
			AGGREGATION COMPUTATIONS			
1–6	.2355	n.s.	.1562	n.s.	.006634	< .05
			AGGREGATION SEQUENCE INTERPRETATION			
1–6	not significant		not significant		significant	

TABLE 19: COMPUTATIONAL RESULTS FOR EXAMPLES 9–11

	Example 9		Example 10		Example 11	
Panel	Cumulative Probability	Significance (.05, one-tailed)	Cumulative Probability	Significance (.05, one-tailed)	Cumulative Probability	Significance (.05, one-tailed)
			ORDINARY SERIAL COMPUTATIONS			
1	.85541	n.s.	.85541	n.s.	.85541	n.s.
2	.85541	n.s.	.85541	n.s.	.85541	n.s.
3	.85541	n.s.	.85541	n.s.	.85541	n.s.
4	.14461	n.s.	.03521	$<$ .05	.00391	$<$ .05
5	.14461	n.s.	.03521	$<$ .05	.00391	$<$ .05
6	.14461	n.s.	.03521	$<$ .05	.00391	$<$ .05
			ORDINARY SERIAL SEQUENCE INTERPRETATION			
1–6	not significant		significant		significant	
			AGGREGATION COMPUTATIONS			
1–6	.2355	n.s.	.05574	n.s.	.006634	$<$ .05
			AGGREGATION SEQUENCE INTERPRETATION			
1–6	not significant		not significant		significant	

TABLE 20: COMPUTATIONAL RESULTS FOR EXAMPLES 12–14

	Example 12		Example 13		Example 14	
Panel	Cumulative Probability	Significance (.05, one-tailed)	Cumulative Probability	Significance (.05, one-tailed)	Cumulative Probability	Significance (.05, one-tailed)
	ORDINARY SERIAL COMPUTATIONS					
1	.99611	n.s.	.99611	n.s.	.99611	n.s.
2	.85541	n.s.	.85541	n.s.	.14461	n.s.
3	.63671	n.s.	.14461	n.s.	.03521	< .05
4	.14461	n.s.	.03521	< .05	.03521	< .05
5	.03521	< .05	.03521	< .05	.00391	< .05
6	.00391	< .05	.00391	< .05	.00391	< .05
	ORDINARY SERIAL SEQUENCE INTERPRETATION					
1–6	significant		significant		significant	
	AGGREGATION COMPUTATIONS					
1–6	.09674	n.s.	.01464	< .05	.000111	< .05
	AGGREGATION SEQUENCE INTERPRETATION					
1–6	not significant		significant		significant	

The two methods produce different results and interpretations in four out of the nine examples because the ordinary serial method does not take into account compensating disparities, *i.e.*, overrepresentations of the test group. Regarding Examples 6, 7, 10 and 12, the aggregation method finds the sum of the underrepresentations sufficiently compensated by the sum of the overrepresentations, and the result is non-significance. The ordinary serial method observes only that there are, respectively, two, two, three and two significant overrepresentations, which give significance to the sequence as a whole.

The aggregation method provides the better test in the current context.[63] It must now be determined whether another version of the serial method is more appropriate to the "underrepresentations predominant" situation, and whether it will produce inferences more closely resembling those of the aggregation test.

2. The Differential Serial Test

The serial method can be modified by shifting the focus of attention from the absolute to the relative number of underrepresentations, that is, to the difference between the number of underrepresentations and the number of overrepresentations of the test group in the sequence of panels under examination.[64] A fair selection method will produce the same number of negative and positive deviations from the true population value (*e.g.*, 50% women) in the sampling distribution.[65] If both types of deviations occur, but one more often than the other, it can be questioned whether the imbalance is compatible with fair selection procedures.[66] Unfortunately, the differential serial test requires a larger

63. It will be seen below, however, that the aggregation method becomes inappropriate in the very rare case of completely balanced disparities.

64. The differential serial test could be given a restrictive formulation by working only with the difference between the number of *significant* underrepresentations and overrepresentations. The consequence would be a further increase in the number of panels required for the test sequence. Taking into account *all* differences, the required minimum number is eight, as shown below. The reformulations would require eight panels *with significant deviations* among the test panels. A test series of eight panels will not always include eight panels with significant deviations. A sequence of more than eight panels may often be required. How many more panels beyond eight may be needed cannot be specified a priori.

65. *See* note 61 and accompanying text *supra*.

66. Each deviation, it should be noted, is given equal weight. A small deviation in one direction can thus cancel out a larger deviation in the other direction. It would be possible to refine the differential serial test so that the magnitudes of the deviations (their relative probabilities) are taken into account. At that point, however, the testing method gains unwarranted complexity for what is a fairly minor, auxiliary procedure. Restricting the test to significant differences, *see* note 64 *supra*, however, achieves a major part of the differential-magnitude adjustment: fairly large deviations can be canceled only by other fairly large ones.

number of panels in the test sequence than are normally available. Specifically, the minimum number of panels needed is *eight*; it is not possible to obtain significance at .05 for numerical differences between the two types of deviations in test sequences of seven or fewer panels.[67] A discussion of the required computations will demonstrate why this is the case.

The differential serial test employs the same formula as was used to establish the probabilities of the selection outcome.[68] A comparison of the meanings of the symbols in both contexts will clarify the similarities and differences.

COMPARATIVE DISPLAY

Formula: Term	Testing: Single-Panel Test	Testing: Panel-Sequence Test[69]
N	number of persons in panel	number of panels in sequence
r	number of test group members in panel	number of underrepresentations in sequence
n–r	number of remaining panel members	number of overrepresentations in sequence
p	proportion of test group members in eligible population	proportion of underrepresentations of all deviations in sampling distribution = .5
q	proportion of test group in eligible population	proportion of overrepresentations of all deviations in sampling distribution = .5

The most important point concerns the meaning of "N." Whereas in the previous context N referred to the number of persons in the jury panel, it now refers to the number of panels in the test sequence. The number of potential jurors in the jury panel is always larger than eight; indeed, this number is measured in the hundreds and even in the thousands.[70] The number of panels in a test sequence, however, is usually less than eight. Since the differential serial test requires a sequence of at least eight panels, it is clear that in most testing situations it will not be applicable even as an auxiliary method.

67. If all deviations are in one direction (*i.e.*, underrepresentation), significance can be obtained for a test sequence as small as five panels. This, however, would be a situation of "underrepresentations only" rather than of "underrepresentations predominant."

68. *See* notes 38-40 and accompanying text *supra*.

69. Panel outcomes which correspond exactly to the population value (*e.g.*, 50% women) are excluded from the computation. Therefore, the proportion of each deviation is set at .5 (50% of all the deviations in the sampling distribution).

70. This is the reason why it was not necessary to consider the current problem of having a small value for N in the previous context.

Examples 6-14, employed previously in this section, contain only six panels. Thus, they cannot be used to illustrate the differential serial test. A new example (designated "A") will be introduced solely for the purposes of this section. The example consists of ten panels of eight persons each. The test group is women, who are assumed to constitute 50% of the relevant population. Table 21 provides the details.

TABLE 21: THE PANELS OF EXAMPLE A

Panel	1	2	3	4	5	6	7	8	9	10
Number of Women	1	0	8	0	0	1	1	1	0	0
Cumulative Probability*	.035	.004	.004	.004	.004	.035	.035	.035	.004	.004
Significant Deviation**	U	U	O	U	U	U	U	U	U	U

* See Table 2 *supra* for the exact probabilities.
** U = significant underrepresentation.
O = significant overrepresentation.

The numerical values for the computation formula are: N = 10, r = 9, N-r = 1, p = .5, q = .5. Carrying out the computation gives a specific probability of .0098 and a cumulative probability of .0108. The difference between nine underrepresentations and one overrepresentation in a test sequence of ten deviating panels is therefore significant. The surplus of eight underrepresentations would be obtained by an unbiased selection method only once in a hundred selections of ten panels of the present type. Using the customary standard of a .05 significance level, the differential serial test reveals discrimination in this sequence of ten panels.

The test sequence of Example A is also found to be significant when the other two tests are employed. Under the rules of the ordinary serial test,[71] disparities between the population and panel proportions are due to discrimination when in a sequence of ten panels three or more of the panels show significant underrepresentations. Since the test sequence shows eight significant underrepresentations, discrimination must be inferred.

The aggregation test combines the several panels of a test sequence

71. *See* Table 7 *supra.*

into a single set, treating them as a single event and computing a single probability for it. If this probability is smaller than 0.05, the result is significant and discrimination must be asserted. Combining the panels gives the following numerical values for the computation formula: N = 80, r = 12, N-r = 68, p = .5, q = .5. Completing the computation gives a specific probability of .00000000005 and a cumulative probability of .00000000006. Once again, the test results reveal discrimination. For Example A, all three methods thus produce the same conclusion. As noted earlier,[72] however, the differential serial method can be employed only in sequences of eight panels or more. Thus, it will only rarely be usable as an auxiliary method to the basic aggregation test.

DECISION RULES FOR DIFFERENTIAL SERIAL TESTS

RULE 1. Disparities between the population and panel proportions of a given group may be due to chance *or* to discrimination when in a series of eight panels or more the probability of obtaining a surplus of underrepresentations relative to overrepresentations is larger than .05. In such a case, discrimination cannot be positively asserted on the basis of the differential serial test.

RULE 2. Disparities between the population and panel proportions of a given group are due to discrimination when in a series of eight panels or more the probability of obtaining a surplus of underrepresentations relative to overrepresentations is .05 or smaller. In such a case, discrimination in the selection of the panelists must be asserted.

D. Balanced Disparities

This section focuses on the last of the three basic types of outcome sequences: *balanced disparities*. This type of situation is very rare. It is also the most complex situation, from a technical as well as from a legal perspective.

1. Application of Differential Serial and Aggregation Tests

The first task is to determine how the differential serial and aggregation methods work in the context of balanced under- and over-representations.[73] Examples 15-18 provide the test data. Each sequence is in

72. *See* note 67 and accompanying text *supra*.

73. It was seen earlier that the ordinary serial method is not appropriate for sequences that contain underrepresentations *and* overrepresentations.

exact balance. The sequences differ in the number and strength of the disparities. Table 22 presents the gender compositions of the panels. Table 23 gives the results of the tests.

TABLE 22: PANEL COMPOSITIONS FOR EXAMPLES 15–18

Panel	Data for Women	Examples 15	16	17	18
1	Number	8	8	8	8
	Proportion	1.00	1.00	1.00	1.00
2	Number	6	6	7	0
	Proportion	.750	.750	.875	.000
3	Number	4	6	6	8
	Proportion	.500	.750	.750	1.00
4	Number	4	2	2	0
	Proportion	.500	.250	.250	.000
5	Number	2	2	1	8
	Proportion	.250	.250	.125	1.00
6	Number	0	0	0	0
	Proportion	.000	.000	.000	.000
1–6	Number	24	24	24	24
	Proportion	.500	.500	.500	.500

TABLE 23: COMPUTATIONAL RESULTS FOR EXAMPLES 15–18

	Testing Method				
	Differential Serial Test			Aggregation Test	
	Number of Significant		Interpretation		Interpretation
Example	Under–Representations	Over–Representations	(Significant at .05, one–tailed)	Cumulative Probability	(Significant at .05, one–tailed)
15	1	1	not significant	.5574	not significant
16	1	1	not significant	.5574	not significant
17	2	2	not significant	.5574	not significant
18	3	3	not significant	.5574	not significant

Table 23 shows complete agreement between the two testing methods in a variety of balanced situations: no sequence is significant. Technically, the uniformly non-significant results are not surprising: each sequence *as a whole* exhibits proportional representation of men and women. But the conclusion to which both tests seem to point—that a fair selection method was employed in selecting the panels under inspection—is not entirely satisfactory. Example 18, for one, shows six significant disparities in six selections. Every single panel of this sequence shows extreme discrimination—against men or against women. It can therefore be questioned whether it is proper to conclude that the method of selection which produced these six panels was fair and unbiased.

2. Compensating Disparities and Constitutional Issues

Two principles are at stake in the testing of jury panels: the fairness of the selection method and the representativeness of the panels. The first relates to the due process and equal protection clauses of the Fourteenth Amendment and has been interpreted to mean that all citizens eligible for jury service should have an *equal chance* of being chosen.[74] The second principle relates to the cross-sectional requirement that the panels be *representative* of the community from which they are drawn.[75] The two principles are closely related. Most often the representativeness requirement is not satisfied because the selection method did not give to all eligible citizens the same chance of being chosen. Discriminatory selection methods produce unrepresentative panels. This is an iron law in respect to single panels. In the testing of sequences of panels, however, *unstable* discriminatory selection methods can produce overall results which look non-discriminatory. This, of course, is precisely the case with the "balanced significant disparities" under discussion.

Given a sequence of panels in which one-half included no women and the other half included no men, it could not be argued that men and women had the same chance of being chosen in the selection of the particular panels or that the particular panels were representative of the community. Assuming panels of equal size, an even number of panels and a population consisting of 50% men and 50% women, the aggregation method will produce a *non*-significant cumulative probability and thus *no* inference of discrimination, even if one-half of the panels in-

74. Peters v. Kiff, 407 U.S. 493, 503 (1972). The only method capable of producing this result is a true random selection method.

75. *See, e.g.*, Taylor v. Louisiana, 419 U.S. 522 (1975).

clude no men and one-half of them include no women. Indeed, in spite of the greatest possible discrimination in the selection process and the greatest possible lack of representativeness in each jury panel, some might argue that the sequence is an example of an unbiased selection process because the aggregate percentages correspond to those which would be expected under proportional representation.

In the situation of "balanced disparities," the aggregation method as well as the differential serial method fail to reveal the existence of discriminatory selection practices.[76] Another testing method must therefore be found.

3. *The Two-Tailed Serial Test*

Before a test appropriate for "balanced disparities" can be discussed, attention must be given to the issue of *two-tailed* testing.[77] The methods introduced in the preceding sections were aimed *only* at the discovery of underrepresentation of the test group (*e.g.*, women), even where information about overrepresentations was taken into account.[78] The residual group (*e.g.*, men) did not require separate and specific attention within that framework: any underrepresentation of the test group was necessarily reflected by a corresponding overrepresentation of the residual group. Testing for discrimination in the context of "balanced disparities" requires a shift in focus. The new method must be able to take into account underrepresentations of either group, test or residual. In fact, it must treat *both* groups as test groups.

Testing for the underrepresentation of one group involves only one side of the sampling distribution.[79] The serial and aggregation methods discussed above therefore employed a one-tailed test of significance. Testing for the underrepresentation of both groups involves both sides of the sampling distribution. It requires the use of two-tailed

76. If the ordinary serial test had been used, Examples 17 and 18 would have emerged as significant. These results, however, do not make the ordinary serial test applicable to "balanced disparities" situations.

77. *See generally* H. BLALOCK, SOCIAL STATISTICS 163-65 (1979); J. FLEISS, STATISTICAL METHODS FOR RATES AND PROPORTION 20-22 (1973); L. FREEMAN, ELEMENTARY APPLIED STATISTICS FOR STUDENTS IN BEHAVIORAL SCIENCE 153-54 (1965).

78. The emphasis on one-tailed testing is realistic. Actual sequences of jury panels tend to show consistent disparities of one type in respect to each relevant social characteristic. It is rare for both types of disparities to occur. It is even rarer for the two types of disparities to be balanced, should there be both underrepresentations and overrepresentations. And it is practically unheard of to find panel sequences that display significant disparities in balance. The present discussion of two-tailed testing is included to provide a complete analysis for all possibilities, and not because the reader will encounter it with great frequency.

79. *See* Table 2 *supra*.

tests of significance. The level of significance remains the same in two-tailed testing as it was in one-tailed testing (*e.g.*, .05). The critical region (region of rejection) therefore also remains the same. The difference is found in the fact that the critical region must be divided and allocated evenly to the two sides of the sampling distribution. It will be remembered that in the one-tailed testing situation the specific probabilities are summed from only one side of the sampling distribution,[80] beginning with the most extreme outcome and continuing until the sum corresponds to the chosen significance level (.05). In the two-tailed testing situation the specific probabilities are summed from both sides of the distribution, beginning with the extreme outcomes and continuing until the sum of each side corresponds to one-half of the chosen significance level (.025).[81] Table 24, using the same basic data as Table 2, gives an example of this process.

80. *Id.*
81. See note 77 *supra.*

TABLE 24: PROBABILITY CUMULATION FOR TWO-TAILED TESTING

Number of Women	Specific Probabilities	Cumulative Probabilities — Testing for the Underrepresentation of Women	Cumulative Probabilities — Testing for the Underrepresentation of Men	Interpretation (Significance at .05, two-tailed)
0	.00391	.00391		Significant underrepresentation: women
1	.0313	.03521		
2	.1094	.14461		non-significant underrepresentation: women
3	.2187	.36331		
4	.2734	.63671	.63671	proportional representation
5	.2187		.36331	
6	.1094		.14461	non-significant underrepresentation: men
7	.0313		.03521	
8	.00391		.00391	significant underrepresentation: men

Serial tests and aggregation tests are the two basic methods for determining significance and discrimination in a sequence of panels. It was seen earlier that the aggregation method is not appropriate for "balanced significant disparities."[82] A method appropriate for this situation must be found in the context of *serial* testing and in the logic of the "maximum number of allowable significant disparities."[83] The new test can be designated the "two-tailed serial method." The name implies the procedure. The two-tailed serial method gives equal attention to significant disparities of either type. The sequence of panels is examined to find the number of significant selection outcomes. The sequence as a whole is significant—and discrimination must be asserted[84]—if in a sequence of twenty panels or fewer, two or more significant outcomes are found. This follows the standard serial rules.[85] The only new procedure is that now a particular panel will be interpreted as significant if its cumulative probability is .025 or smaller.[86]

Examples 15-18 can be used once more. Table 25 gives the results for the two-tailed serial method. It will be remembered that the differential serial test and the aggregation test produced non-significant results for each one of the four examples.[87]

TABLE 25: EXPANDED RESULTS FOR EXAMPLES 15–18

	Number of Significant			Interpretation
Example	Under–Representations	Over–Representations	Total Number of Significant Disparities	(Significance at .05, two–tailed)
15	1	1	2	significant
16	1	1	2	significant
17	1	1	2	significant
18	3	3	6	significant

82. *See* notes 73-76 and accompanying text *supra.*

83. *See* notes 46-51 and accompanying text *supra.*

84. Discrimination (giving one person less of a chance to be chosen than another) can occur for at least two reasons: (1) a deliberate, purposive, motivated underselection of persons belonging to the test group; and (2) an erroneous, unintended underselection of persons belonging to the test group because of technical defects in the design and/or execution of the selection process. The first cause is likely to produce disparities tending predominantly in one direction (*i.e.* underrepresentation of the test group). The second cause may produce a variety of abnormalities, depending on the exact nature of the technical defect. It may produce one-sided disparities much like the first cause. It may also produce situations characterized in this paper as "balanced disparities."

85. *See* text accompanying note 52 *supra.*

86. Two-tailed testing divides the critical region of .05 into two regions of .025 each, one at each end of the distribution. *See* note 77 *supra.*

87. *See* notes 73-76 and accompanying text *supra.*

The two-tailed serial test produces significant results for all four of the examples. In situations of "balanced disparities," this method is preferable to the others not only statistically but also from the perspective of constitutional requirements. As noted earlier, when some panels consist entirely (or almost entirely) of members of one gender—or of one race, class, occupational or age group—*both* constitutional principles are violated: fairness (equal chance) in the selection process and representativeness of the panels as the selection outcome.[88] The rules for the test can be stated as follows:

DECISION RULES FOR TWO-TAILED SERIAL TEST

RULE 1. Disparities between the population and panel proportions of two complementary groups, both of which exhibit significant underrepresentations, may be due to chance *or* to discrimination when in a sequence of seven panels or fewer, *no or only one* panel result is significant. In such a case, discrimination cannot be positively asserted on the basis of the two-tailed serial test.

RULE 2. Disparities between the population and panel proportions of two complementary groups, both of which exhibit significant underrepresentation, are due to discrimination when in a sequence of seven panels or fewer, *two or more* of the panel results are significant. In such a case, discrimination in the selection of panelists must be asserted.

It must be pointed out that the authors are not aware of any case directly sanctioning the two-tailed method of testing for discrimination in the selection of jury panels. Yet, it seems that the use of this type of discrimination analysis can be supported on the basis of generally accepted principles and rulings of the courts. The very premise of the prima facie rule[89] is that chance and accident alone are insufficient to explain the sweeping exclusion of specific groups from jury service.[90] Similarly, a pattern of balanced disparities must be attributed to determinative factors and must therefore be condemned. Alternating total exclusion with total inclusion results in the removal of a particular set

88. *See* notes 74-75 and accompanying text *supra*.

89. *See* notes 11-21 and accompanying text *supra*. *See also Grand Juries, supra* note 1, at 63, 78-79.

90. Eubanks v. Louisiana, 356 U.S. 584 (1958); Smith v. Texas, 311 U.S. 128 (1940).

of attitudes and attributes from each jury ultimately selected. The Supreme Court has left no doubt that such results will not be tolerated:

> When any large and identifiable segment of the community is excluded from jury service, the effect is to remove from the jury room qualities of human nature and varieties of human experience, the range of which is unknown and perhaps unknowable. It is not necessary to assume that the excluded group will consistently vote as a class in order to conclude, as we do, that their exclusion deprives the jury of a perspective on human events that may have unsuspected importance in any case that may be presented.[91]

It is no more acceptable to exclude men from every other jury panel, and at the same time exclude women from each alternate panel, than it is to exclude men or women from all panels. In each instance all juries will be deprived of the full range of experience required to provide a complete perspective on human events. Similarly, near total exclusion found in alternating panels (*e.g.*, 99%/1%, 1%/99%, etc.) presents an equally discriminatory and unconstitutional result.

E. Summary of Rules

The preceding pages introduced four testing methods in the context of three types of panel sequences. The discussions included a variety of computational examples and references to basic statistical and legal principles. Attention was given not only to which testing methods were appropriate, but also to those which were not appropriate in a given context and for what reasons. Table 26 provides a schematic overview for the materials of Part II. It shows in which context a testing method is the primary, auxiliary or only test for discrimination in the selection of jury panels, and in which situations the method should not be employed (indicated by a dash).

TABLE 26: TESTS AND SITUATIONS

	Testing Method			
		Serial Tests		
Situation	Aggregation Test	Ordinary	Differential	Two-Tailed
Underrepresentations Only	PRIMARY	AUXILIARY	—	—
Underrepresentations Predominant	PRIMARY	—	AUXILIARY	—
Balanced Disparities	—	—	—	ONLY

91. Peter v. Kiff, 407 U.S. 493, 503-04 (1972).

III. Multiple-Panel Testing Methods Applied to Large Panels

Panels of eight persons were used in the preceding sections to give the reader a clear understanding of how the various probability figures were obtained, how they were used, and how they produced the results and interpretations for each of the panel sequences. The logic of test development and the results of the tests were not affected by the small size of the panels. All real jury panels are, of course, larger than eight persons; they range from about forty or fifty to several thousand members. It will be useful to apply the testing methods developed in this article to a set of panels of realistic size. This application will illustrate how the various tests are applied in "real life" cases.

The examples are organized in accordance with the three basic types of sequences: underrepresentations only, underrepresentations predominant and balanced disparities. Three examples are given for each type of situation. Each set of examples includes outcomes associated with a reasonably good jury panel selection method, a poor one and a very poor one. Only the appropriate testing methods will be applied.[92]

Table 27 presents the composition of the first three panel sequences. All existing disparities in these examples are *underrepresentations* of the test group (women). Two testing methods are therefore applicable: the aggregation test (of primary relevance) and the ordinary serial method (playing an auxiliary role). Table 28 gives the computational results and interpretations.

Example 19 shows panel compositions that are likely to emerge from a reasonably good selection method. All of the existing disparities are very small; none of them is significant. The only reason that some doubt remains about the validity of the selection method is that there are no overrepresentations. It will be remembered that a truly unbiased sampling will produce disparities of both types.[93] If possible—if there have been no changes in the selection method and if older data are still obtainable—more panels should be included in the test sequence to determine whether the disparities are also all of one kind in the longer sequence. If older panels are not available, the outcomes of future selections should be watched carefully. If the one-sideness of the disparities persists, an effort should be made to discover the underlying causes and a remedy should be developed. As applied to the cur-

92. *See* Table 26 *supra*. The following data apply to Examples 19-27: (a) proportion of women in the relevant population = .5; (b) proportion of men in the relevant population = .5; (c) size of each panel = 300 persons; (d) test sequence for each example = five panels.

93. *See* note 26 and accompanying text *supra*.

rent sequence, however, neither test leads to the assertion of discrimination.

Example 20 exhibits panel compositions that are likely to emerge from a poorly designed and/or biased selection system. All disparities are of one type. Some of them are quite large and several of them are significant. There is no doubt that the system is biased against women and should be changed. The results of both tests require the assertion of discrimination.

Example 21 shows panel compositions of the kind that emerge from grossly incompetent and/or extremely biased selection systems. Every outcome is an underrepresentation of the test group. All disparities are large and all of them are significant. The selection system is strongly biased against women. It does not fulfill constitutional requirements and its reform is mandatory. Both tests clearly indicate the presence of discrimination in the selection of the jury panelists.[94]

TABLE 27: PANEL COMPOSITIONS FOR EXAMPLES 19–21

		Example 19		Example 20		Example 21	
Panel	Data	Women	Men	Women	Men	Women	Men
1	Number	150	150	134	166	120	180
	Proportion	.500	.500	.447	.553	.400	.600
2	Number	148	152	145	155	130	170
	Proportion	.493	.507	.483	.517	.433	.567
3	Number	147	153	140	160	125	175
	Proportion	.490	.510	.467	.533	.417	.583
4	Number	150	150	130	170	115	185
	Proportion	.500	.500	.433	.567	.383	.617
5	Number	149	151	131	169	100	200
	Proportion	.497	.503	.437	.563	.333	.667
1–5	Number	744	756	680	820	590	910
	Proportion	.496	.504	.453	.547	.393	.607

94. In the experience of the authors, actual test sequences of panels most often tend to have the appearance of Examples 20 and 21. Sequences with partially or fully compensating disparities are rare when methods other than true random selection procedures are employed by the court. Whatever the existing bias might be, it tends to persist as long as the same selection method is in use.

TABLE 28: COMPUTATIONAL RESULTS FOR EXAMPLES 19-21

	Example 19		Example 20		Example 21	
Panel	Cumulative Probability	Significance (.05, one-t.)	Cumulative Probability	Significance (.05, one-t.)	Cumulative Probability	Significance (.05, one-t.)
			ORDINARY SERIAL COMPUTATIONS			
1	.5230	n.s.	.0367	< .05	.000317	< .05
2	.4313	n.s.	.3017	n.s.	.0121	< .05
3	.3864	n.s.	.1363	n.s.	.00229	< .05
4	.5230	n.s.	.0121	< .05	.0000313	< .05
5	.4770	n.s.	.0163	< .05	.(8)401*	< .05
			ORDINARY SERIAL SEQUENCE INTERPRETATION			
1–5	not significant		significant		significant	
			AGGREGATION COMPUTATIONS			
1–5	.3882	n.s.	.000165	< .05	.(16)683*	< .05
			AGGREGATION SEQUENCE INTERPRETATION			
1–5	not significant		significant		significant	

* The figure in parenthesis indicates the number of zeros to be inserted after the decimal point.

Table 29 presents the compositions of the second set of panel sequences. Both types of disparities are found in the examples, but the *underrepresentations are predominant.* The applicable testing methods are therefore the aggregation test (primary) and the differential serial test (auxiliary). Table 30 gives the computational results and interpretations.

Example 22 shows panel compositions characteristic of a fairly good selection system. The existing disparities are quite small—none is larger than 1.7 percentage points—and no disparity is significant. In the present sequence of five panels underrepresentations predominate over overrepresentations by a ratio of three to one. If the selection system is completely unbiased, larger sequences should reflect the two types of disparities in closer balance. If the imbalance persists, some remedy should be adopted. Neither testing method, in any case, demonstrates discrimination for the current sequence of panels.

Example 23 exhibits panel compositions that are likely to emerge from a poorly designed and/or biased selection system. Three significant disparities are found in only five panels, including a deviation of as large as eight percentage points. Neither test, however, leads to a conclusion of discrimination against women. The reason is that there is only one more significant underrepresentation than significant overrepresentations. It is likely, however, that a larger sequence would reveal a greater predominance of underrepresentations and thus significance and bias against women for the sequence as a whole. This is also indicated by the fact that the aggregation test just barely fails to produce a significant result (cumulative probability for the sequence = .0638). If additional past panels are not available, the selection system should be retested regularly in future years. Even though significance is not obtained for the present sequence of five panels, the selection system clearly should be reformed.

Example 24 shows compositions of the kind that emerge from grossly incompetent and/or extremely biased selection systems. In a sequence of five panels there are three significant underrepresentations of the test group and one significant underrepresentation of the residual group. All four disparities are substantial, including differences as large as 16.7 percentage points. There is no doubt that the selection system discriminates against women. Both tests show significance and the selection system clearly fails to meet constitutional requirements.

TABLE 29: PANEL COMPOSITIONS FOR EXAMPLES 22–24

		Example 22		Example 23		Example 24	
Panel	Data	Women	Men	Women	Men	Women	Men
1	Number	147	153	133	167	110	190
	Proportion	.490	.510	.443	.557	.367	.633
2	Number	146	154	142	158	100	200
	Proportion	.487	.513	.473	.527	.333	.667
3	Number	150	150	174	126	145	155
	Proportion	.500	.500	.580	.420	.483	.517
4	Number	155	145	131	169	175	125
	Proportion	.517	.483	.437	.563	.583	.417
5	Number	148	152	140	160	100	200
	Proportion	.493	.507	.467	.533	.333	.667
1–5	Number	746	754	720	780	630	870
	Proportion	.497	.503	.480	.520	.420	.580

TABLE 30: COMPUTATIONAL RESULTS FOR EXAMPLES 22-24

Panel	Example 22 Cumulative Probability	Example 22 Significance (.05, one-tailed)	Example 23 Cumulative Probability	Example 23 Significance (.05, one-tailed)	Example 24 Cumulative Probability	Example 24 Significance (.05, one-tailed)
			PROBABILITY COMPUTATIONS			
1	.3864	n.s.	.0283	< .05	.(5)195	< .05
2	.3431	n.s.	.1933	n.s.	.(8)401	< .05
3	.5230	n.s.	.9977	> .95	.3017	n.s.
4	.7373	n.s.	.0163	< .05	.9984	> .95
5	.4313	n.s.	.1363	n.s.	.(8)401	< .05
			DIFFERENTIAL SERIAL INSPECTION			
			Number of Significant Under- and Over-Representations			
	Under-R.	Over-R.	Under-R.	Over-R.	Under-R.	Over-R.
1-5	0	0	2	1	3	1
			DIFFERENTIAL SERIAL SEQUENCE INTERPRETATION			
1-5	not significant		not significant		significant	
			AGGREGATION COMPUTATIONS			
1-5	.4283	n.s.	.0638	n.s.	.(9)313	< .05
			AGGREGATION SEQUENCE INTERPRETATION			
1-5	not significant		not significant		significant	

Table 31 presents the compositions of the third set of panel sequences. Both types of disparities are found in the examples, but the disparities are *completely balanced* in number and magnitude. The only applicable testing method is therefore the two-tailed serial test.[95] Table 32 gives the test results.[96]

Example 25 reflects an excellent selection system. The existing disparities are small, non-significant and balanced. These are the kinds of panels that will be produced by a truly unbiased selection procedure. Courts should not be satisfied until their jury panels take on the appearance of the panels of this sequence. As would be expected, the two-tailed serial test shows a non-significant result.

Example 26 exhibits panel compositions that are likely to emerge from a poorly designed and/or biased selection system. The sequence as a whole is characterized by proportional representation, but two of the panels are unacceptably far removed from balanced representativeness. Men and women are each underrepresented by twenty-five percentage points, which is equivalent to an underrepresentation of 50%. The test shows a significant result. The system does not meet constitutional requirements and must be changed.

Example 27 shows compositions of the kind that can emerge only from grossly incompetent and/or extremely biased selection systems.[97] Out of five panels, two include no women and two include no men. These four panels are as far removed from representativeness as is possible. It is clear that the selection system does not fulfill constitutional requirements and must be changed. As would be expected, the two-tailed serial test produces a significant result.

95. As noted earlier, in the two-tailed testing situation a cumulative probability must be .025 or smaller in order to be significant—assuming maintenance of the customary .05 level of significance. *See* text accompanying note 81 *supra*.

96. The aggregation test is *not* appropriate for panel sequences showing balanced disparities. *See* text accompanying note 76 *supra*. It is of interest to note, however, that an aggregation computation would produce a cumulative probability of .5103, rendering all three examples equally non-significant.

97. The authors have never encountered a sequence of jury panels so drastically discriminatory *and balanced*. Many jury panels have shown the complete exclusion of a social group (especially blacks in the southern states). But these exclusions are not balanced by compensating exclusions of the corresponding group (*e.g.*, whites). If a sequence of panels as shown in Example 27 were actually found, incompetence rather than "deliberate" bias would be suspected. Discrimination, however, is equally unacceptable whether it results from the incompetence or the prejudices of court administrators.

TABLE 31: PANEL COMPOSITIONS FOR EXAMPLES 25–27

		Example 25		Example 26		Example 27	
Panel	Data	Women	Men	Women	Men	Women	Men
1	Number	155	145	135	165	0	300
	Proportion	.517	.483	.450	.550	.000	1.00
2	Number	140	160	150	150	300	0
	Proportion	.467	.533	.500	.500	1.00	.000
3	Number	150	150	75	225	150	150
	Proportion	.500	.500	.250	.750	.500	.500
4	Number	145	155	165	135	300	0
	Proportion	.483	.517	.550	.450	1.00	.000
5	Number	160	140	225	75	0	300
	Proportion	.533	.467	.750	.250	.000	1.00
1–5	Number	750	750	750	750	750	750
	Proportion	.500	.500	.500	.500	.500	.500

TABLE 32: COMPUTATIONAL RESULTS FOR EXAMPLES 25–27

Panel	Example 25 Cumulative Probability	Example 25 Significance (.05, one-tailed)	Example 26 Cumulative Probability	Example 26 Significance (.05, one-tailed)	Example 27 Cumulative Probability	Example 27 Significance (.05, one-tailed)
			PROBABILITY COMPUTATIONS			
1	.3017	n.s.	.04696	n.s.	.(90)49	< .025
2	.1363	n.s.	.5230	n.s.	.(90)49	< .025
3	.5230	n.s.	.(18)72	< .025	.5230	n.s.
4	.3017	n.s.	.04696	n.s.	.(90)49	< .025
5	.1363	n.s.	.(18)72	< .025	.(90)49	< .025
			TWO-TAILED SERIAL INSPECTION			
	Number of Significant Outcomes		Number of Significant Outcomes		Number of Significant Outcomes	
1–5	0		2		4	
			TWO-TAILED SERIAL SEQUENCE INTERPRETATION			
1–5	not significant		significant		significant	

IV. Remedies

When the testing of a sequence of jury panels produces significant results, the observed disparities between the population and panel compositions of the test group(s) must be attributed to discriminatory practices in the selection of the panelists.[98] Such a result requires two types of remedies, one focused on the future and one on the present and the past.

Regarding the future, it is clear that the selection system must be changed so that properly composed panels will be chosen in subsequent years. Representative panels will be selected when two conditions are met: (1) the selections must be made from a *complete listing* of the eligible population; and (2) the selections of the particular panelists must be made by a *true random process*. The first requirement mandates the use of multiple listing sources. Placing sole reliance on lists of registered voters, for example, is not adequate. Methods are now available to integrate a variety of source lists at a relatively low cost in money and time, thereby producing a combined list that comes reasonably close to being a complete listing of the eligible population.[99] The second requirement mandates selection procedures such as the blind drawing of slips of paper from a well-mixed urn, or, much more efficiently, the use of a random number technique. It unequivocally prohibits the use of such procedures as "key-man" selection, in which judges and/or jury commissioners appoint friends and acquaintances to the jury panel.

Regarding the present and the past, whether a criminal defendant intends to challenge the selection of the grand jury that indicted him and/or the selection of the petit jury that will hear his case, the same procedures will be utilized. Initially, a pretrial motion must be made setting forth the basis of the challenge,[100] and an evidentiary hearing must be held to allow presentation of the necessary expert testimony.[101] A motion directed to the composition of the grand jury panel is one to

98. It must be remembered, however, that non-significant results do *not* prove that there was no discrimination. Non-significance only means that there is an alternative explanation for the observed disparities between population and panel compositions: random sampling fluctuations. Since discrimination does remain a possible explanation, selection systems that produce large and/or one-sided disparities should be reformed even when the tests do not show significance.

99. The Superior Court of Alameda County, California, for example, now uses a combined list based on voter registration and drivers license records. Integrated lists can be generated by hand sorts or by computer methods.

100. Agnew v. United States, 165 U.S. 36 (1897).

101. *See* Montez v. Superior Court, 10 Cal. App. 3d 343, 88 Cal. Rptr. 736 (1970).

quash the indictment; a motion assailing the makeup of the petit jury panel may be labeled a motion to quash the trial jury. If either motion is denied, the defendant may seek immediate relief from the appellate courts in an effort to halt the proceedings,[102] and, if unsuccessful, he may proceed to trial and if convicted appeal on the basis of an improperly constituted jury panel.[103] If a motion to quash an indictment is granted, the state may seek a new indictment from a properly constituted grand jury or may proceed by way of the "information" process.[104] In the event a motion to quash the trial jury is granted, the state may adopt new selection methods calculated to produce the necessary cross-sectional representation and then proceed to trial. In either instance, of course, the state may first seek appellate review of the trial court's ruling on the motion.[105]

Appendix: How *Not* to Test for Discrimination in Series of Panels

Readers familiar with basic probability theory may have wondered why the authors have not utilized the product rule to determine the probability of a sequence of panels, particularly since Professor Finkelstein employed this technique in his now classical article, *The Application of Statistical Decision Theory to the Jury Discrimination Cases*.[106] As will be shown, however, the product rule is *not* an appropriate technique for determining significance and discrimination in the context of jury panel selection.

When applied to independent events,[107] the product rule states that the probability of obtaining both events (outcomes) is the product of the probabilities associated with each.[108] The rule can be extended

102. A writ of prohibition is available for this purpose in California, pursuant to California Code of Civil Procedure section 1102 *et seq.* (West Supp. 1979). Ganz v. Justice Court, 273 Cal. App. 2d 612, 78 Cal. Rptr. 348 (1969).

103. The right to appeal on the basis of an improperly constituted jury panel following conviction is available only if the defendant has brought a timely challenge to the composition of the jury in the trial court. Michel v. Louisiana, 350 U.S. 91 (1955).

104. The dismissal of an indictment, information or complaint charging a felony is not a bar to another prosecution for the same offense. CAL. PENAL CODE § 999 (West 1970); People v. Combes, 56 Cal. 2d 135, 145, 14 Cal. Rptr. 4 (1961).

105. *See* Kronenberg, *Right of a State to Appeal in Criminal Cases*, 49 J. CRIM. L.C. & P.S. 473, 476-77 & n.24 (1959). Regarding federal cases, *see* Friedenthal, *Government Appeals in Federal Criminal Cases*, 12 STAN. L. REV. 71 (1959).

106. 80 HARV. L. REV. 338 (1966).

107. Periodic selections of jury panels, when based on a random selection method, are independent events. Having been selected or not having been selected for a prior panel does not change the likelihood of a person's being selected for a subsequent panel.

108. The product rule (also known as the multiplication rule) states that "[t]he

to any number of events. An example will clarify the rule. For an unbiased coin, the probability of getting "heads" is .5, as is the probability of getting "tails." It is desired to test whether a given coin is in fact unbiased. The hypothesis is that the coin is biased against tails, producing a disproportionately large number of heads when flipped.[109] For purposes of simplicity, a small number of tests and few flips will be used. In a real testing situation a larger number of both would be desirable. The logic of the exercise, however, would be no different.

Four tests will be made of the coin in question (each test corresponds to the selection of a jury panel). Each test will consist of four flips of the coin (each flip corresponds to the selection of one panelist). There are sixteen different decision chains for each test. However, there are only five different types of outcomes. Each decision chain has the same probability of occurring: 1/16. But the outcomes have different probabilities of occurring since some can result from only one and some from a number of decision chains. The following decision tree

DECISION TREE

Flip 1	Flip 2	Flip 3	Flip 4	Outcome Type
H	H	H	H	a
			T	b
		T	H	b
			T	c
	T	H	H	b
			T	c
		T	H	c
			T	d
T	H	H	H	b
			T	c
		T	H	c
			T	d
	T	H	H	c
			T	d
		T	H	d
			T	e

probability of two independent outcomes occurring together is the product of their respective probabilities." P. JACOBSON, INTRODUCTION TO STATISTICAL MEASURES FOR THE SOCIAL AND BEHAVIORAL SCIENCES 125 (1976). *See also* H. BLALOCK, SOCIAL STATISTICS 124-29 (1979).

109. The ability to predict the direction of the bias (*e.g.*, against "tails") permits the use of a one-tailed test of significance. Testing for bias without a prediction about its direction would require the use of a two-tailed test of significance.

illustrates these possibilities. Table 33 presents the specific probabilities associated with each test outcome.

TABLE 33: OUTCOMES AND PROBABILITIES

Outcome Type	Number of Heads	Number of Tails	Frequency of Occurrence	Probability Fraction	Probability Decimal
a	4	0	1	1/16	.0625
b	3	1	4	4/16	.25
c	2	2	6	6/16	.375
d	1	3	4	4/16	.25
e	0	4	1	1/16	.0625
a–e	—	—	16	16/16	1.0000

Table 33 shows that, for example, the specific probability of getting three heads and one tail in four flips of a fair coin is 0.25. The product rule indicates that the probability of obtaining this outcome in all four tests is found via the multiplication of the separate probabilities: .25 × .25 × .25 × .25 (or $.25^4$). The product of this multiplication is .00390625. The probability of getting three heads and one tail for each test in a sequence of four tests, each consisting of four flips of a coin, is .00390625, or roughly four times in one thousand. Only four times in a thousand four-test/four-flip trials would this sequence of outcomes occur with an unbiased coin. This is a fairly small (and significant) probability, and it must be concluded that the coin is biased.

The outcomes of each test do not need to be the same for the application of the product rule. Table 34 shows a different sequence of outcomes.

TABLE 34: SEQUENCE WITH DIFFERENT TEST RESULTS

Test	Number of Heads	Number of Tails	Specific Probability
1	3	1	.25
2	4	0	.0625
3	2	2	.375
4	3	1	.25

The product rule indicates that the probability of obtaining this sequence of outcomes is .25 × .0625 × .375 × .25, which equals .001465, or about one and one-half times in ten thousand. Only one and one-half times in ten thousand four-test/four-flip trials would this sequence of outcomes occur with an unbiased coin. This is a significant result (smaller than .05). The null hypothesis that the coin is unbiased must be rejected; bias (discrimination) must be asserted.

The probabilities of the *specific* outcomes have thus far been used in this Appendix. It will be remembered, however, that in the context of testing for discrimination in the selection of jury panels (and in many other contexts) *cumulative* probabilities must be used.[110] The proper question is *not* "what is the probability of getting two tails in four flips?" but "what is the probability of getting two tails *or fewer* in four flips?"[111] Testing for bias against "tails," the specific probabilities for the tail side of the sampling distribution are therefore cumulated. Table 35 presents the revised array for the last example.

TABLE 35: SEQUENCE WITH DIFFERENT TEST RESULTS: ADJUSTED

Test	Number of Heads	Number of Tails	Specific Probability	Cumulative Probability
1	3	1	.25	.3125
2	4	0	.0625	.0625
3	2	2	.375	.6875
4	3	1	.25	.3125

The product rule indicates that the correct (cumulatively-based) probability of this sequence of outcomes is $.3125 \times .0625 \times .6875 \times .3125$, which equals .00420, or approximately four in one thousand. Only four times in one thousand four-test/four-flip trials would an unbiased coin produce this sequence of outcomes. The result is significant (smaller than .05). The null hypothesis of an unbiased coin must be rejected; bias (discrimination) must be asserted.

The key question considered in this Appendix can now be raised: the propriety of using the product rule to determine the probability and significance of series of jury panel selections. Professor Finkelstein recommended this application of the product rule:

> In *Swain*, however, venires with five or fewer Negroes appeared in thirty consecutive cases. The probability of this occurrence, applying the product rule, is $.2^{30} = 4.63 \times 10^{-21}$. This means that, on the average, only one in more than one hundred

110. *See* notes 32-34 and accompanying text *supra*.

111. A commonly employed alternative formulation is: "What is the probability of getting no more than two tails in four flips?"

> million trillion groups each containing thirty venires would consist solely of venires which were not more than 15% Negro.[112]

He had previously determined that the probability of obtaining no more than five blacks in *one* thirty-person venire was .2026, given that eligible blacks constituted 25% of the population. To determine the probability of a sequence of thirty such venires, as the above quotation shows, Professor Finkelstein multiplied the probabilities of the thirty venires: $.2^{30}$. It is suggested that this application of the product rule is inappropriate. The method *under*-estimates the compound probabilities—the longer the panel sequence, the greater the underestimation. This point can be clarified by returning to some earlier examples.

Table 4 presents the cumulative probabilities for the first example of this article. The example consists of five panels (tests), each containing eight persons (flips). The cumulative probabilities are: .03521, .03521, .00391, .00391 and .03521. Multiplying these five figures gives the product of .000000000667, or approximately seven times in ten billion. Preserving the .05 level of significance, this is obviously a significant result. Discrimination must be asserted for the selection process that produced the panels of Example 1. The aggregation and ordinary serial methods reach the same conclusion.[113] Indeed, the cumulative sequence probability computed for the aggregation test is very similar to the figure produced by the product rule: .000000009733, or about one in one hundred million.

The basic problem with using the product rule—at least in the present context—is that it requires the multiplication of fractions by other fractions. The inescapable result is an ever smaller fraction (an ever smaller probability), regardless of the size of the component fractions. This means that *any* sequence of panels will be found to be significant, regardless of the composition of the panels, if the test sequence is long enough. Table 2 indicates that having three women on a panel of eight persons is not a significant outcome when women constitute 50% of the eligible population. The cumulative probability for the panel is .36331. Applying the product rule to sequences of such panels produces the following results: two panels = .132, three panels = .048. Only three such panels in sequence are needed to produce significance for the sequence as a whole.

This result may seem proper. It relates to an important problem that was considered earlier: persistent non-significant disparities can be

112. Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Cases*, 80 HARV. L. REV. 338, 357 (1966).

113. *See* Table 14 *supra*.

the result of a discriminatory selection system. The aggregation method was introduced precisely because the ordinary serial method could not always detect this type of bias. (The aggregation method does not work perfectly in these circumstances but is better than the serial method.) The product rule, however, lacks the "restraint" of the aggregation method. An additional illustration will make this clear. Example 5 consists of five panels of eight persons, each panel including four women. Since women were assumed to constitute 50% of the eligible population, each panel as well as the sequence as a whole shows the best possible selection result: proportional representation. Each panel probability is non-significant (.63671). The probability for the sequence as a whole is also non-significant, whether tested by the aggregation ($p = .5627$) or by the serial method. The application of the product rule, however, gives this result: $.63671^5 = .1046$. The addition of only two more panels to the sequence produces this probability: $.63671^7 = .0424$. The product rule method would therefore conclude that a sequence of seven panels which separately and jointly exhibit proportional representation of women (the test group) discriminates against women—obviously an unacceptable result. In contrast, the aggregation and serial methods produce non-significant results for the seven-panel sequence.

The product rule can be used to determine how likely it is that a given sequence of outcomes will occur. The likelihood of obtaining a sequence of seven panels of eight persons, each including four or fewer women is indeed .0424. The product rule cannot be used, however, to infer bias and discrimination in the context of jury panel selection. The technical reason for the inapplicability of the product rule is found in the fact that as more panels are added to a sequence, the number of decision chains increases and the likelihood that any particular chain of outcomes will occur becomes even smaller. Even the most likely chains acquire a low absolute probability. And as seen above, even with relatively few panels in the sequence, the likelihood of a particular chain quickly falls below the significance level.

This problem exactly parallels the difficulties which led to the conclusion that specific probabilities cannot be used in the testing of single panels.[114] If a panel is large enough—and it does not have to be very large—every outcome, even the most likely one (proportionality), acquires a very low probability. As panels get larger, there are ever more possible different outcomes. Since the sum of the probabilities of all

114. *See* notes 30-31 and accompanying text *supra*.

outcomes remains unity (1.0), the specific probability of each outcome necessarily becomes smaller with each increase in panel size. For a test group that constitutes 50% of the eligible population, there is no non-significant outcome in a panel of as few as 300 persons.[115]

Contrary to Professor Finkelstein's suggestion, the product rule therefore cannot be applied to determine discrimination in the selection of jury panel members. Depending on the nature of the disparities in the sequence of panels, the aggregation method and/or one version or the other of the serial method must be employed.

115. *Statistical Decision Theory, supra* note 2, at 85-94.