Duquesne Law Review

Volume 17 | Number 3 Article 10

1978

Additional Comments on the Application of Statistical Analysis to Differential Pass-Fail Rates in Employment Testing

B. C. Spradlin

J. W. Drane

Additional Comments on the Application of Statistical Analysis to Differential Pass-Fail Rates in Employment Testing

*B. C. Spradlin**
and
*J. W. Drane***

An article[1] in a recent issue of *The Harvard Law Review* compared a statistical technique for assessing the significance of differences in pass rates of applicant groups on employment tests with the "80% rule" adopted as part of the Federal Executive Agency Guidelines. The author demonstrated that the two methods may lead to conflicting results when applied to the test pass rates of applicant groups to determine if the tests have a disproportionate impact on one group.

The advantages of the statistical technique selected—a test of the difference between independent proportions[2]—when compared to the procedure adopted by Federal Guidelines were appropriately noted. The principal advantages were that the statistical test takes into account both sample size and the magnitude of the difference in the sample proportions. These are important considerations, and Professor Shoben presents a well-written advocacy for using statistical analysis for a more reliable inference concerning whether an employment test has an adverse impact on a protected class employee under Title VII[3] statutes.

However, the opinion expressed in this note is that the comment leaves two points that need clarification. The first is that statistical tests are, by and large, general in applicability and are not unique to any one setting. The second is that conflicting inferences about

* Ph. D; Vice President, Criterion Analysis, Inc., Dallas, Texas.

** P.E., Ph. D; Associate Professor, Department of Statistics, Southern Methodist University, Dallas, Texas.

1. Shoben, *Differential Pass-Fail Rates in Employment Testing: Statistical Proof Under Title VII*, 91 HARV. L. REV. 793 (1978).

2. Eberhardt & Fligner, *A Comparison of Two Tests for Equality of Two Proportions*, 31 AM. STATISTICIAN 151 (1977); P. HOEL, INTRODUCTION TO MATHEMATICAL STATISTICS § 9.3 (1971); W. MENDENHALL, INTRODUCTION TO PROBABILITY AND STATISTICS 202 (1975).

3. 42 U.S.C. §§ 2000e(1)-(17) (1970 & Supp. V 1975).

adverse impact on a given employee group may result when applying different methods of statistical analysis, in which case no inference should be made.

I. General Application of Supreme Court Criterion for Statistical Significance

The first point concerns the applicability of the statistical technique employed by the Supreme Court in the cases of *Castaneda v. Partida*[4] and *Hazelwood School District v. United States*[5] to cases concerning the significance of differences in pass-fail rates in employment tests. It is suggested here that the opinion expressed by the Court in *Castaneda* and *Hazelwood* implies a much broader concept of statistical analysis than the particular technique used in those cases. Although the sample statistics calculated in the examples of Professor Shoben's comment[6] are differences in proportions rather than proportions, the wording of the statistical inferences cited by the Supreme Court in *Castaneda* and *Hazelwood* apply directly to those examples.[7] A sample difference statistic as advocated by Professor Shoben is different from the sample statistics cited by the Court in the cases mentioned above in form only. The process of forming statistical inferences used by the Court may be

4. Castaneda v. Partida, 430 U.S. 482 (1977). The Court states: "As a general rule for such large samples [n=870 in this case], if the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that the jury drawing was random would be suspect to a social scientist." *Id.* at 496 n.17.

5. Hazelwood School District v. United States, 433 U.S. 299 (1977). The Court states: "The Court in *Casteneda* noted that 'as a general rule for such large samples, if the difference between the expected value and the observed number is greater than two or three standard deviations' then the hypothesis that teachers were hired without regard to race would be suspect." *Id.* at 309 n.14.

6. Shoben, note 1 *supra*.

7. *Id.* at 807. In *id.* at 807 n.53, under the null hypothesis, the expected value of the sample difference statistic P_w(sam)-P_b (sam) is 0. For this example, the calculated z is the number of standard deviations that the sample statistic P_w (sam)-P_b (sam) lies from its expected value under the null hypothesis. Therefore, in the wording of the Court in *Castaneda* and *Hazelwood*, "Since the difference between the expected value (0) and the observed value (.1) is less than two or three standard deviations the hypothesis of no difference in impact of the employee tests could not be rejected." *See* notes 4 & 5 *supra*. In Shoben, *supra* note 1, at 807 n.55, according to the Supreme Court criterion for significance, since the difference between the expected value of the statistic and the observed or sample value of the statistic is 3.13 standard deviations, the hypothesis of no difference in the impact of the employment tests on different employee groups is suspect.

applied generally and so may the criteria for rejection[8] of the null hypothesis. In fact, reduction in sample size does not restrict the generality of the process of developing inferences. In cases when the sample size does not comply with the Court's definition of "large",[9] a translation can be made so that the number of standard deviations between the observed statistic and its expected value can be stated in terms that reflect a large sample equivalent.

This translation is called a normal equivalent deviate.[10] The procedure allows the results from any small sample statistical test of significance (for example a t test)[11] to be stated in terms of a large sample standard normal deviate (a z statistic).[12] For example, suppose it is of interest to determine whether the differential in average salaries for men and women is statistically significant for a given small group of employees. Further, suppose there are three male ($n=3$) and four female ($n=4$) employees in the group and a t statistic reflecting sample salary differences is calculated[13] to be equal to

8. *See* notes 4 & 5 *supra*.

9. *See* 430 U.S. at 496 n.17.

10. D. FINNEY, STATISTICAL METHOD IN BIOLOGICAL ASSAY 443, 452 (2d ed. 1964).

11. *See* authorities cited in note 2 *supra*.

12. *See* Shoben, *supra* note 1.

13. For this calculation, the hypothesis to be tested is $H_o: M_1 = M_2$ (i.e., there is no difference in the means of male and female salaries) or stated alternatively $H_o: M_1 - M_2 = 0$. Where M_1 = universe mean salary for men, and M_2 = universe mean salary for women. The t statistic is used to test the hypothesis. To calculate t, we need $\bar{X}_1$ = sample mean salary for men and $\bar{X}_2$ = sample mean salary for women and the standard error of the difference $\bar{X}_1 - \bar{X}_2$. The degrees of freedom are given by $n_1 + n_2 - 2 = 5$, since $n_1 = 3$ and $n_2 = 4$. The t statistic has the form:

$$t = (\bar{X} - \bar{X}) \div \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where S_1^2 and S_2^2 are the sample variances for men and women respectively. The denominator is the standard error of $\bar{X}_1 - \bar{X}_2$.

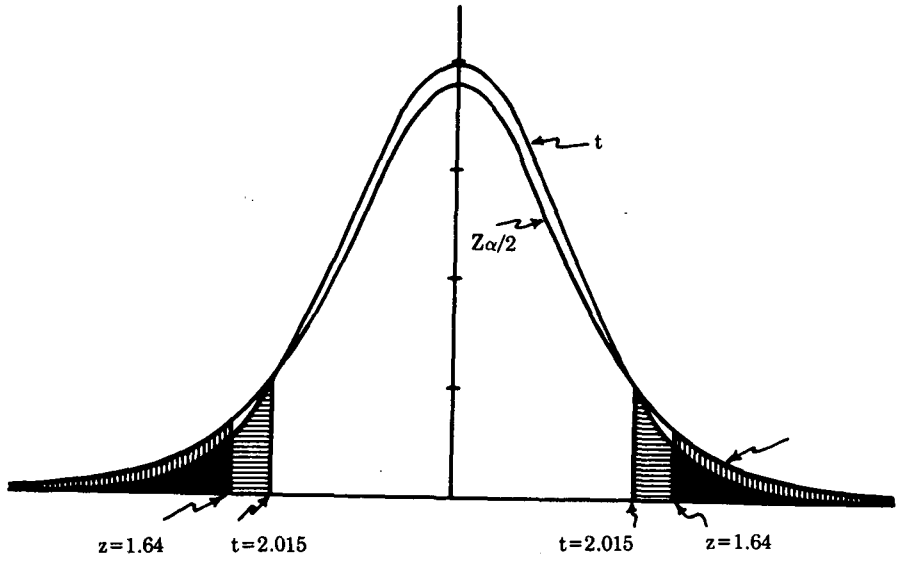
If we assume that a t calculated in this manner is equal to 2.015, we then find from Table IV of R. FISHER, STATISTICAL METHODS FOR RESEARCH WORKERS that the level of significance is = .1. In other words, we would expect to see a difference in men and women's salaries as large or larger than the one we observe about .1 of the time in repeated sampling where $n_1 = 3$, and $n_2 = 4$.

14. The translation from $t = 2.015$ to a standard normal deviate $z = 1.64$ can be shown by the following diagram:

2,015. This small t value can be translated[14] directly to a standard normal deviate (z statistic) equal to 1.64. The resulting value can then be analyzed in the context of the Court's definition of significance for large samples. In this example, the hypothesis of no difference in male and female salaries could not be rejected according to the Court's criterion. A similar procedure can be followed allowing the results of any statistical test of significance to be placed within the context of the Court's definition of significance in *Hazelwood* and *Castaneda.*

II. Conflicting Results from Application of Different Statistical Tests

The second point for clarification concerns the fact that conflicting inferences about adverse impact on a given employee group may occur when different methods of statistical analysis are used. This can happen using the statistical approach on the same data. This point can be illustrated with the following hypothetical set of data.



In other words $Z\alpha/2=1.64$ is the standard normal deviate that implies the same level of significance as the t value of 2.015 from the given small sample.

15. Shoben, *supra* note 1, at 804.

Suppose five hundred (500) white employees take an employment test and nineteen (19) are successful. One hundred fifty (150) black employees take the same test and one (1) is successful. These results can be displayed in a two-way table as shown below.

	Pass	Fail	Total
Whites	19	481	500 (Number of Whites N_W)
Blacks	1	149	150 (Number of Blacks N_B)
	20	630	650

The null hypothesis to be tested is that there is no difference in the pass rates for black and white employees (i.e., H :P (pop)—P (pop)=0). The statistical test[16] suggested in Professor Shoben's comment applies to these data as follows:

(a.) H_O :P_W (pop)–P_B (pop)=0

(b). P_W (sample)=19/500 = .0380; P_B (sample)=1/150 = .00667

(c.) Calculate the overall proportion of people in both groups who passed the test. For all 650,

$$\frac{19+1}{650} = .0308$$

(d.) Compute the overall proportion of people in both groups who did not pass the test. For all 650 employees,

$$1-.0308=.9692$$

(e.) Multiply the proportions from steps c and d.

$$\text{Prod.}=(.0308)(.9692)=.02985$$

(f.) Apply the following formula to calculate the standard error:

$$\text{Standard Error} = \sqrt{\frac{\text{Prod.}}{N_W} + \frac{\text{Prod.}}{N_B}}$$

where N_W and N_B are the number of whites and the number of blacks respectively in the sample. Therefore,

$$\text{Standard Error} = \sqrt{\frac{.02985}{500} + \frac{.02985}{150}} = .0161.$$

Calculate the z statistic according to the following fomula:

$$z = \frac{P_W \text{ (sample)} - P_B \text{ (sample)}}{\text{standard error}} = \frac{.038-.00667}{.0161} = \underline{\underline{1.95}}$$

16. *Id.*

Thus, if the level of significance chosen for this test were a=.05, the critical value for z is 1.96, and the null hypothesis of no difference in pass rates could not be rejected.

Another acceptable statistical procedure for testing the null hypothesis of no difference in pass rates between white and black employees is the Likelihood Ratio Chi-Square Test.[17] When this procedure is applied to the data in the contingency table shown above a statistic, G, is calculated.

$$\begin{aligned} G^2 &= 2[19\log(19) + 481\log(481) + 1\log(1) + 149\log(149) \\ &\quad - 500\log(500) - 150\log(150) - 20\log(20)\ 630\log(630) \\ &\quad + 650\log(650)] \\ &= 5.078 \end{aligned}$$

wherein log(X) is the natural logarithm of X. G^2 is to be compared to tabled values of chi-square. Assuming a significance level of .05 the critical value of chi-square with one degree of freedom is $X^2 = 3.841$[18] while the calculated value of G is 5.071. G^2 exceeds 3.841. Therefore, the null hypothesis of no difference in pass rates between

17. *See* Eberhardt & Fligner, supra note 2, at 151; S. Wilks, *The Large-Sample Distribution of The Likelihood Raio for Testing Composite Hypothesis*, 9 ANNALS OF MATHEMATICAL STATISTICS 60-62 (1938); S. KULLBACK, INFORMATION THEORY AND STATISTICS (1959); M. BISHOP, S. FIENBERG & P. HOLLAND, DISCRETE MULTIVARIATE ANALYSIS (1976).

The Likelihood Ration Chi-Square, G, for contingency or cross tabulated tables has the general form,

$$G_2 = \sum_j \sum_i n_{ij} \log[n_{ij}/(n_{i.}\, n_{.j}/n_{..})]$$

with degrees of freedom equal (r-1)(c-1), wherein r = number of rows and c = number of columns; and log (.) is the natural logarithm of whatever is in parentheses.

For a two by two contingency table,

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{..}$

where $n_{i.}$ represents a row total, $n_{.j}$ represents a column total and $n_{..}$ represents the overall total, the generalized form shown in (a.) *supra* may be reduced to.

$$\begin{aligned} G^2 &= 2\,[n_{11}\log(n_{11}) + n_{12}\log(n_{12}) + n_{21}\log(n_{21}) + n_{22}\log(n_{22}) \\ &\quad - n_{1.}\log(n_{1.}) - n_{2.}\log(n_{2.}) - n_{.1}\log(n_{.1}) - n_{.2}\log(n_{.2}) \\ &\quad + n_{..}\log(n_{..})] \end{aligned}$$

with degrees of freedom given by,

(2-1)(2-1) = 1.

18. F. ROHLF & R. SOKOL, STATISTICAL TABLES, Table R (1969).

white and black employees must be rejected. This result is in contrast to the conclusion reached when the same hypothesis about the same data was tested through the calculation of a z statistic. Further, it sometimes occurs that a hypothesis which cannot be rejected by the Likelihood Ratio Chi-Square may be rejected by the z test. The contradiction which may occur in test results does not always work in the same direction for these same two tests.

The point made by Professor Shoben[19] is that the application of Federal Executive Agency Guidelines for assessing significance in differences in pass rates of employee groups on employee tests is lacking as a reliable approach in determining adverse impact. The comment demonstrated that a statistical procedure for testing hypothesis is a more reliable inferential process. Further demonstrated was the fact that the statistical test (a z test) may produce results which contradict those produced by the Federal Guidelines when applied to the same data.

The point made in this note is that the same kind of contradiction may occur between the results of two equally acceptable statistical tests. The authors of this comment suggest that when this happens, no inference of adverse impact can be made.

19. Shoben, note 1 *supra*.