

Running Head: ADVERSE IMPACT

Significance Tests and Confidence Intervals for the Adverse Impact Ratio

Scott B. Morris

Illinois Institute of Technology

Russell Lobsenz

Federal Bureau of Investigation

Abstract

Since the introduction of the Civil Rights Act of 1964, courts have increasingly relied on statistical evidence in adverse impact cases. There are several shortcomings associated with current statistical methods for determining adverse impact. An alternative framework for evaluating adverse impact based on confidence intervals around the impact ratio is proposed.

Correspondence concerning this article should be addressed to Scott B. Morris, Illinois Institute of Technology, Chicago, IL 60616 or morris@charlie.acc.iit.edu (email). Poster presented at the 13th annual Society for Industrial and Organizational Psychology Conference (April 1998, Dallas)

Significance Tests and Confidence Intervals for the Adverse Impact Ratio

Since the introduction of the Civil Rights Act of 1964, the courts have increasingly relied on statistical evidence in employment discrimination cases. Over the past thirty years the nature of statistical evidence has gradually evolved. The two most common approaches are the four-fifths rule outlined in the Uniform Guidelines (1978), and the Z-test for differences in selection rates.

These methods differ in two ways. First, the four-fifths rule focuses on an effect size (i.e., impact ratio), while the Z-test focuses on statistical significance. Second, they are based on different operational definitions of adverse impact. The impact ratio is a ratio of selection rates, while the Z-test is based on the difference between selection rates. As a result of these differences, the two methods will not necessarily reach the same conclusions.

In this paper, we propose an alternative framework for evaluating adverse impact based on confidence intervals around the impact ratio. Using this framework, practitioners can evaluate the practical significance of results by examining the estimate of effect size, and can evaluate statistical significance through the confidence interval, or through a test of significance on the impact ratio.

Defining Adverse Impact

The concept of adverse impact was first delineated by the Supreme Court decision in Griggs v. Duke Power Company (1971). Under the adverse impact theory, discrimination exists when there is evidence of a statistical disparity in selection or promotion rates (i.e., the proportion of applicants in each group who are selected). In other words, adverse impact exists

when a protected group (e.g., Blacks) is selected at a substantially lower rate than the majority group. Adverse impact can exist even if the practice appears neutral on the surface and when there is no evidence of an intent to discriminate.

The presence of adverse impact does not, in and of itself, make a selection system illegal. Rather, a finding of adverse impact gives rise to a prima facie case, or a presumption of discrimination. The employer must then defend the selection procedure by demonstrating that it is job related. When it can be shown that the procedure validly measures attributes that are essential for successful job performance, the procedure is permissible under Title VII.

Although Griggs defined the concept of adverse impact, the Supreme Court did not specify exactly how large the difference in selection rates must be in order to constitute a substantial discrepancy. Clearer guidance was provided by the EEOC with the adoption of the Uniform Guidelines on Employee Selection Procedures (1978). The Uniform Guidelines suggested a rule of thumb, commonly referred to as the four-fifths rule, which has become the standard methodology for determining adverse impact. This four-fifths rule is based on the Impact Ratio, which is the ratio of the selection rate for the minority group (SR_{\min}) to the selection rate for the majority group (SR_{maj}), or

$$IR = \frac{SR_{\min}}{SR_{\text{maj}}}. \quad (1)$$

According to the Uniform Guidelines (Section 4D, p. 38297),

A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5)(or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact...

The fact that the four-fifths rule is relatively straightforward and easy to implement has resulted in its becoming the most favored application for determining adverse impact in employment discrimination cases. However, not all commentators have embraced the four-fifths rule. For example, Boardman (1979) and Greenberg (1979) both demonstrated that the four-fifths rule has excessive Type I and Type II error rates. Type I errors indicate that the rule would falsely indicate many cases of adverse impact where none exists in the population (i.e., if the selection system were applied to all potential applicants). Type II errors indicates that the rule would often fail to identify adverse impact when it does exist in the population. Both results are based on the fact that the Impact Ratio is subject to considerable sampling error, particularly when sample sizes are small. More recently, Lawshe (1987) demonstrated the effect of sampling error on actual data. The Impact Ratio was computed for selection decisions for 11 jobs across two consecutive years. Although the same selection systems were used in both years, the Impact Ratios changed dramatically. This instability, Lawshe (1987) argued, seriously undermines the usefulness of the Impact Ratio in the determination of discrimination.

Based on the problems with the Impact Ratio, Wing (1982) suggested combining this information with a test of statistical significance. Reliance on a significance test would control for Type I errors, although it would not assist in avoiding Type II errors (Schmidt, 1996). A preference for statistical significance has also been expressed by some courts (e.g., Hazelwood School District v. United States, 1977). For example, in Rich v. Martin-Marietta (1979), the court dismissed the four-fifths rule in favor of a test of statistical significance. Similarly, the OFCCP Compliance Manual (Office of Federal Contract Compliance Programs, 1993) recommends that a significance test be computed in addition to the four-fifths rule.

Several different statistical tests have been recommended for assessing adverse impact (e.g., chi-square, Fisher exact test, Z-test). The two most common approaches are the Z-test for the difference between two proportions, and the chi-square test for association in a 2x2 contingency table. Both tests are based on the normal approximation to the binomial distribution, and are mathematically equivalent. The Z-test for the difference between proportions (Z_D) is computed as follows (OFCCP, 1993),

$$Z_D = \frac{SR_{\min} - SR_{\max}}{\sqrt{SR_T(1 - SR_T)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \quad (2)$$

where SR_T is the total selection rate, and N_1 and N_2 are the number of applicants in each group. This approach is often labeled the 2-SD test, because the difference is considered significant if it is more than two standard deviations above or below zero (or more precisely, $|Z| > 1.96$, corresponding to the 95% confidence interval).

Although some courts have recognized a precise threshold for concluding a statistically significant result (e.g., two standard deviations, 95% confidence level), several have cautioned that significance criteria "must not be interpreted or applied so rigidly as to cease functioning as a guide and become an absolute mandate or proscription." (Albemarle Paper Company v. Moody, 1975). Additionally, in United States v. Test (1976), the court ruled that "the mathematical conclusion that the disparity between these two figures is 'statistically significant' does not, however, require an a priori finding that these deviations are 'legally significant'..." The court's skepticism of statistical significance is partially attributable to their appreciation of the effects of small sample sizes on outcomes of significance tests. For example, when large samples are

available, even small disparities will be statistically significant. Contrarily, when samples are very small, only large differences will produce statistically significant results. Thus, even if an employer is discriminating, the statistical test may have little power to detect it.

Understanding power is very important for interpreting non-significant results. If the test result is non-significant, it may seem appropriate to maintain the null hypothesis (i.e., no discrimination). However, when there is insufficient power to detect notable disparities, then non-significant results could be due to chance and may not imply that no discrimination exists.

In addition to having low power, the Z_D -test is also limited by the fact that it is based on an effect size (i.e., the difference between selection rates) that is conceptually different from the four-fifths rule (i.e., the ratio of selection rates). Given the widespread acceptance of the four-fifths rule, a significance test that is based on a different effect size is potentially confusing, because the two approaches are based on different operational definitions of adverse impact.

Statistical Significance of the Impact Ratio

A potential alternative that has received little attention is the significance test on the impact ratio directly. This would provide a significance test that is based on the same effect size as the four-fifths rule, thereby providing a consistent framework for assessing adverse impact. In addition, this approach can be used to build confidence intervals around the impact ratio.

Testing significance of the impact ratio is slightly more complex than existing procedures. Because the effect size is a ratio, the sampling distribution is non-symmetric. Values less than one are constrained to the interval zero to one, while values larger than one can be infinitely large. The typical solution is to compute the natural log of the ratio, which takes on positive values if the minority group has the larger selection rate, and negative values when the

selection rate is lower for minorities.

According to Fleiss (1994), the natural log of the impact ratio is approximately normally distributed with standard error,

$$SE_{IR} = \sqrt{\frac{1 - SR_{\min}}{N_{\min} SR_{\min}} + \frac{1 - SR_{maj}}{N_{maj} SR_{maj}}}. \quad (3)$$

Under the null hypothesis, a pooled estimate of the selection rate can be used, producing the test statistic,

$$Z_{IR} = \frac{\ln\left(\frac{SR_{\min}}{SR_{maj}}\right)}{\sqrt{\frac{1 - SR_T}{SR_T} \left(\frac{1}{N_{\min}} + \frac{1}{N_{maj}}\right)}}. \quad (4)$$

Power Analysis

In order to evaluate the utility of the new test (Z_{IR}), the power was compared to the Z_D test for the difference between proportions. Because both sampling distributions are approximately normal, the power can be determined by: (1) defining the minimum significant effect size (ES_{\min}), (2) converting ES_{\min} into a Z-score indicating its position in the sampling distribution relative to the population effect size, and (3) computing the probability of scores exceeding this value in a normal distribution.

Both tests can be viewed as an effect size divided by a standard error. In both cases, the test will be significant if the absolute value of Z exceeds a critical value (here, 1.96 was used, corresponding to $\alpha=.05$). Thus, the statistic will be significant when the absolute value of the effect size is greater than 1.96 times the standard error for the test statistic (i.e., the denominator

of Equation 2 or 4). The relative position of this minimum significant effect size (ES_{\min}) in the sampling distribution was determined by a Z-score,

$$Z_{\min} = \frac{ES_{\min} - ES_{pop}}{SE}, \quad (5)$$

where ES_{pop} is the population effect size, and SE is the standard error of the sampling distribution. In computing power, the null hypothesis is assumed to be false; therefore, it is inappropriate to use a pooled estimate of the selection rate when computing the standard error for Equation 4. For the Z_{IR} test, the standard error was computed using Equation 3, and for the Z_D test, the standard error is (Fleiss, 1994)

$$SE_D = \sqrt{\frac{SR_{\min}(1-SR_{\min})}{N_{\min}} + \frac{SR_{maj}(1-SR_{maj})}{N_{maj}}}. \quad (6)$$

Power was computed as the probability of scores falling below Z_{\min} in a cumulative normal distribution. Because we are focusing on the situation where the minority group has a smaller selection rate, Z_{\min} will always be negative. Significant values will be those that are less than Z_{\min} (i.e., larger negative values). As a result, the probability of scores falling below Z_{\min} is equal to the power of the test.

The power of the test is a function of several parameters: the population effect size, the sample size, the selection rate, and the proportion of minority applicants. Power was computed under a variety of levels of each of these parameters. Two levels of the population effect size were chosen based on the practice of adverse impact analysis. Because, according to the four-fifths rule, an impact ratio greater than .8 does not indicate adverse impact, this value reflects the smallest meaningful effect size. Following the same logic, 3/5ths or .6 should represent a

moderate level of adverse impact. Thus, the population effect size was defined as the natural log of .8 and .6 for the impact ratio test. For the Z_D test, these values were transformed into a difference in selection rates as follows,

$$SR_{\min} - SR_{maj} = \frac{SR_T(1-IR)}{1-PCT_{\min}(1-IR)}, \quad (7)$$

where IR is the population impact ratio, and PCT_{\min} is the percentage of minority applicants. For the selection rate and percent minority applicants, values were chosen to reflect the ideal situation (50%), and a value more realistic for selection situations (10%). Sample size ranged from 100 to 1000. Power was assessed for $\alpha=.05$.

For both tests, power is greatest when the selection rate is high (50%) and the minority makes up a large percentage (50%) of the applicant pool (see Figure 1). Under these conditions, both tests have adequate power to detect moderate levels of adverse impact (IR=.6), even with fairly small sample sizes (i.e., N=100). However, in order to detect minimal adverse impact (IR=.8), a sample size of 600 would be required to achieve 80% power.

Power drops when either of these parameters is low (e.g., Figure 2 presents power when the proportion of minority applicants is 10%). When the adverse impact is moderate (IR=.6), the tests have adequate power only for very large samples (i.e., N>500). On the other hand, even very large samples (N=1000) do not have adequate power to detect small levels of adverse impact. Figure 2 also illustrates a slight power advantage for the test on the impact ratio over the test of the difference in proportions.

When both the selection rate and the percentage of minority applicants is small, power drops off dramatically (see Figure 3). Although there is a slight advantage for the impact ratio

test, neither has adequate power, even for sample sizes as large as 1000.

Confidence Intervals

For many years, researchers have argued against heavy reliance on small-sample studies with low power (most recently, Cohen, 1994; Schmidt, 1992, 1996). This is clearly the case with adverse impact statistics, where huge samples are necessary to achieve reasonable power under typical conditions. A common recommendation is to replace reports of significance and p-values with effect sizes and confidence intervals. While this does not eliminate the problem of low power, it provides a more comprehensive picture of the research results. Specifically, the effect size provides the best estimate of the parameter of interest (in this case, the impact ratio); while the confidence interval communicates the degree of sampling error in that estimate.

The procedure for computing confidence intervals around the impact ratio requires a number of steps, which are outlined in Table 1. To illustrate, consider a selection situation with 50 minority and 450 majority applicants. Among those hired, five are minorities, and 72 are majority members, resulting in an impact ratio of .63 (the natural log of which is -.47). The standard error would be

$$SE_{IR} = \sqrt{\frac{1-0.10}{(50)(0.1)} + \frac{1-0.16}{(450)(0.16)}} = 0.44.$$

The resulting confidence interval for the log value would range from -1.33 to .39. In the original metric, the confidence interval ranges from .27 to 1.47. The confidence interval clearly indicates the problem with sampling error. While the impact ratio of .63 suggests adverse impact, it can be expected that substantially higher or lower ratios could occur simply due to sampling error.

Because this confidence interval includes one, the degree of adverse impact is not statistically

significant.

Conclusion

The two most common approaches to assessing adverse impact (i.e., the four-fifths rule and the Z_D test on the difference in proportions) are based on different operational definitions of adverse impact, and focus on different aspects of the data (effect size vs. statistical significance). An alternative statistical test, the impact ratio test (Z_{IR}), has two advantages over the Z_D test. First, the power of Z_{IR} is slightly better when the proportion of minorities in the applicant pool is small. As a result of low power under these conditions, employers who discourages minority applicants will have less chance of detecting adverse impact. This is clearly contrary to the intent of the *Uniform Guidelines*. A second advantage to Z_{IR} is that, in addition to the significance test, confidence intervals can be built around the impact ratio.

The proposed method, which includes estimates of the impact ratio along with confidence intervals, provides a more informative and more consistent framework for assessing adverse impact. While this procedure is more complex than current methods for assessing adverse impact, the complexity is offset by several advantages. First, it provides an indication of adverse impact in a common metric (i.e., the impact ratio), with a generally accepted criterion for practical significance ($IR < 0.8$). Thus, the recommended procedure does not require a fundamental change in how organizations and the courts define adverse impact. Second, it provides sufficient information so that both practical and statistical significance can be evaluated, through the inclusion of the standard error. Third, the width of the confidence interval should provide practitioners with a better sense of the general lack of power of these tests, and the potential for inaccurate decisions based on sampling error.

References

- Albemarle Paper Co. v. Moody, 422 US 405 (1975).
- Boardman, A. E. (1979). Another analysis of the EEOC 'four fifths' rule. Management Science, 25, 770-776.
- Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.
- Greenberg, I. (1979). An analysis of the EEOC "four-fifths rule," Management Science, 25, 762-769.
- Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), The Handbook of Research Synthesis (pp. 245-260). NY: Russell Sage Foundation.
- Griggs v. Duke Power Co., 401 U.S. 424 (1971).
- Hazelwood School District v. United States, 433 U.S. 299 (1977).
- Lawshe, C. H. (1987). Adverse impact: Is it a viable concept? Professional Psychology: Research and Practice, 18, 492-497.
- Office of Federal Contract Compliance Programs (1993). Federal contract compliance manual. Washington, D.C.: Department of Labor, Employment Standards Administration, Office of Federal Contract Compliance Programs (SUDOC# L 36.8: C 76/993).
- Rich v. Martin-Marietta, 467 F. Supp. 587 (D. Col. 1979).
- Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American Psychologist, 47, 1173-1181.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. Psychological Methods, 1, 115-129.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register, 43, 38290-38315.

United States v. Test, 550 F2d 577 (1976).

Wing, H. (1982). Statistical hazards in the determination of adverse impact with small samples. Personnel Psychology, 35, 153-162.

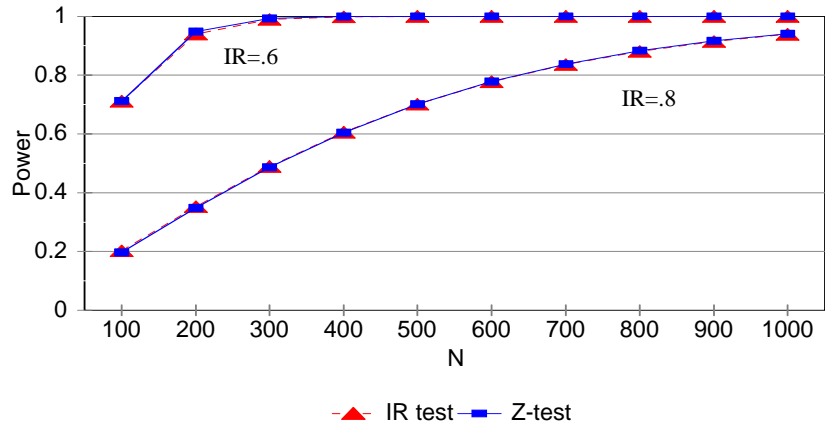
Table 1

Computing confidence intervals on the adverse impact ratio.

Step	Equation
1. Compute Impact Ratio	$IR = \frac{SR_{\min}}{SR_{\max}}$
2. Compute natural log of IR	$\ln(IR)$
3. Compute SE_{IR}	$SE_{IR} = \sqrt{\frac{1 - SR_{\min}}{N_{\min} SR_{\min}} + \frac{1 - SR_{\max}}{N_{\max} SR_{\max}}}$
4. Compute confidence bounds	$\ln(CL) = \ln(IR) \pm 1.96 SE_{IR}$
5. Retranslate into original metric (antilog transformation)	$CL = e^{\ln(CL)}$

Figure 1

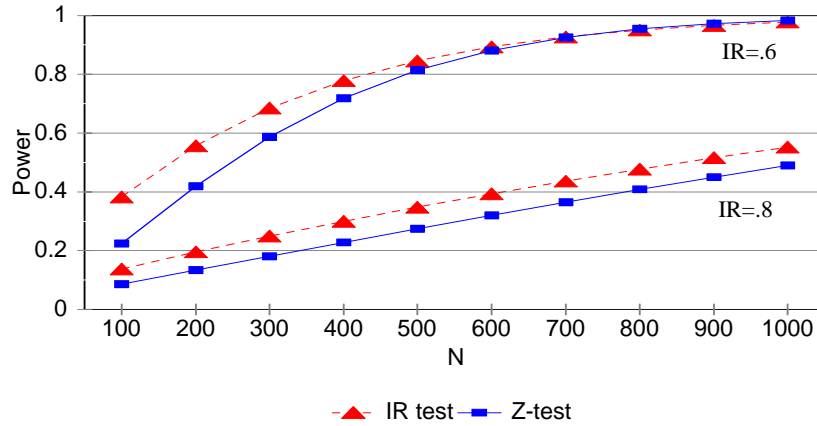
Power of Adverse Impact Statistics with 50% Selection Rate and 50% Minority Applicants.



Note: IR refers to the population impact ratio, IR test is the significance test on the impact ratio, Z-test is the significance test on the difference in proportions.

Figure 2

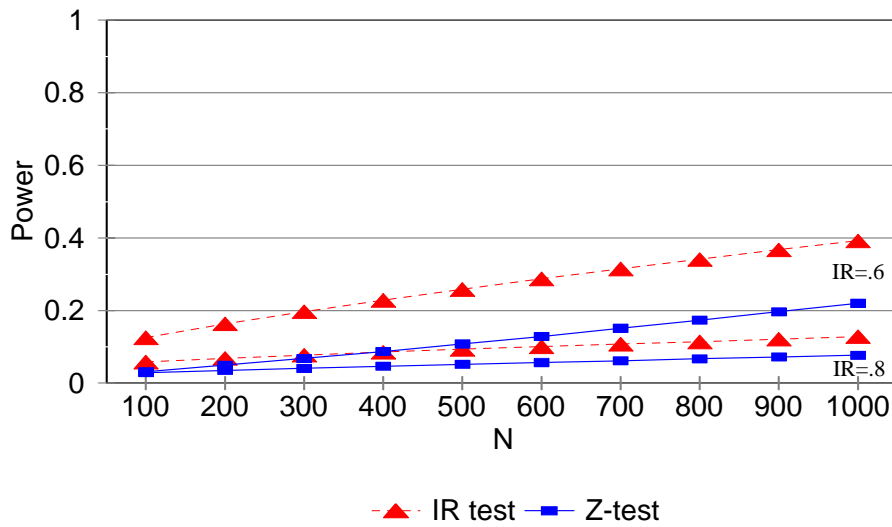
Power of Adverse Impact Statistics with 50% Selection Rate and 10% Minority Applicants.



Note: IR refers to the population impact ratio, IR test is the significance test on the impact ratio, Z-test is the significance test on the difference in proportions.

Figure 3

Power of Adverse Impact Statistics with 10% Selection Rate and 10% Minority Applicants.



Note: IR refers to the population impact ratio, IR test is the significance test on the impact ratio, Z-test is the significance test on the difference in proportions.