Statistical Significance Standards for Basic Adverse Impact Analysis

Author:
David Morgan

Published July 2010

DCI CONSULTING

Statistical Significance Standards for Basic Adverse Impact Analysis

David Morgan

July 2010

This paper reviews basic statistical significance tests for adverse impact (AI) analyses of 2x2 tables. In this context, analysts are interested in whether employment decision (e.g., hiring, promotion, termination, etc.) rates between two groups are meaningfully different. For the purposes of this paper, meaningfully different refers to the confidence an analyst has that results are not due to chance, luck, or any other random factor. In other words, if a difference in rates exists, how likely is it due to chance? The less likely a disparity is due to chance, the more confident the analyst is that a difference in selection rates is real.

AI analyses are particularly relevant to employers covered by laws and regulations administered by the Equal Employment Opportunity Commission (EEOC) and the Office of Federal Contract Compliance Programs (OFCCP)—Federal agencies which primarily enforce Title VII of the Civil Rights Act (CRA) of 1964 and Executive Order (EO) 11246, respectively. These entities apply both reactive and proactive approaches to investigate claims and enforce non-discrimination in the workplace. Although neither statute referenced above explicitly mentions AI as we know it today, the idea did in fact grow out of the verbiage of Title VII:

It shall be an unlawful employment practice for an employer…[to] deprive or tend to deprive any individual of employment opportunities or otherwise <u>adversely affect</u> [emphasis added] *his status as an employee, because of such individual's race, color, religion, sex, or national origin.*

The landmark U.S. Supreme Court ruling in *Griggs v. Duke Power Co.* (1971) provided the first working definition of AI, in that specific discriminatory intent is not required for claims of discrimination; therefore, tests that have AI—an unintentionally disproportionate effect against a protected class—must be both job-related and consistent with business necessity. A document published by the U.S. Department of Labor, in conjunction with the U.S. Department

of Justice and EEOC, entitled the Uniform Guidelines on Employee Selection Procedures (UGESP), describes methodologies by which AI analyses are to be conducted by Federal enforcement agencies. See the first paper in this series (Colosimo, 2010) for a more thorough historical account of the origins of AI and its present-day implications in the EEO context.

Statistical Methods for Adverse Impact Analyses

Two statistical significance tests are most commonly used to analyze data for the purpose of identifying AI. They are: the 2 standard deviation (SD) test, also called the *Z* test, and Fisher's exact test (FET). Both approaches examine the relationship between two variables to determine whether a difference in employment decision rates is likely due to chance. The vast majority of OFCCP settlements have used *Z* tests as the primary measure of adverse impact. A typically accepted probability value linked to acceptable confidence that a disparity is not due to chance is less than .05. That is, there is a 95% probability that such a finding did not occur by chance alone, or that we are only willing to accept a false positive finding of AI less than 5 times out of 100[1].

With the *Z* test, *expected* selection rates for males and females, for example, are compared against *actual* selection rates for both groups. When a large discrepancy between expected and observed totals is statistically significant (e.g., more than 2 SDs apart—the threshold endorsed by OFCCP), one may assume that the result is probably not due to chance. Take Table 1 below, for example, which represents male and female applicants to a Customer Service Representative position for company *x*. As shown, out of 100 male applicants, 35 were selected for the position. Out of 100 female applicants, only 20 were selected for the position. Absent any bias in decisions, we would *expect* males and females to be selected at nearly the same rate (i.e., 28%, which is the approximate overall selection rate). Thus we would expect about 28 hires from each group. However, there were clearly more males selected for the position than females, but is this disparity statistically significant? When using the *Z* test to compare expected and actual selection rates for males (about 35%) and females (20%), there is in fact a statistically significant disparity slightly above 2 SDs (2.38). Additionally, the result is significant at less than a .05 probability value, which means the disparity between the groups is likely due to something more than just chance.

[1] From the statistical standpoint, a one-sided test—a test designed to identify potential discrimination against a focal group—has more statistical power to detect AI, if it exists; however, although this may be true, such an analysis is clearly inconsistent with the law under Title VII. Enforcement agencies such as the OFCCP will certainly have interest in pursuing situations where a minority group has a significantly higher selection rate than a non-minority group. The decision to use a two-sided test, however, should be made before conducting an AI analysis. Such a decision forces consideration of the consequences in both directions, maintaining consistency with Title VII and allowing for appropriate follow-up analyses and decisions.

Table 1. Adverse impact table for male and female applicants.

	Males	Females	
Selected	35	20	55
Rejected	65	80	145
	100	100	

Unfortunately, the *Z* test is not only influenced by the size of an effect, but also the sample and cell sizes. A very large sample size may artificially inflate an effect, and a very small sample size may be inappropriate to test using this method. In situations where few employment decisions are made or there is a very small number of applicants from one group, the *Z* test may overestimate the standard deviation. For example, OFCCP uses FET when the sample size is less than 30 and there are less than 5 persons in each subgroup. The second method, FET, provides an exact probability value, of which the *Z* test only approximates, and has no minimal sample or cell size requirements. Although FET and *Z* outcomes tend to mirror each other as the sample size increases, FET is generally known for providing a more accurate probability value when the sample size is small. Other inherent dissimilarities between the two statistical methods exist as well, as is apparent when various selection models are being analyzed. These issues, recent developments, and implications and applications are discussed.

Some Debates Regarding Adverse Impact Analyses

Conditional Versus Unconditional Assumptions

Collins and Morris (2008) point out that statisticians and experts in the field of industrial and organizational (I/O) psychology have debated over the issue of assessing statistical significance for AI, for decades. Although this debate will probably not be resolved anytime soon, it is important to briefly cover the theoretical models on which arguments for or against particular sampling models are based. The way in which one frames an AI analysis, will affect how the analyses are conducted, and consequently the outcome of the results. Theoretical models that lay the foundation for basic AI analyses typically depend on whether or not the data are fixed in nature. "Fixed" in this sense refers to whether certain characteristics of the applicant pool such as demographics, and the number of employment decisions, were known beforehand.

Statistical tests that assume fixed data are sometimes referred to as conditioned[2] tests because they are used to examine predetermined, or conditioned, values. Tests that assume the data are not fixed are sometimes called unconditioned[3] tests because their marginal values are unknown. Importantly, there further exists statistical tests that assume just part of the data are fixed. Take, for example, the scenario in which we are going to hire *x* number of people, but the applicant pool has not yet been defined. We know how many individuals we intend to hire for vacant positions, but we do not yet know the characteristics of the applicant pool, since no one has applied. In many EEO situations it may be unclear exactly which model best mirrors reality.

Mid-p correction to Fisher's exact test

Agresti (1992) warns that a conditional (hypergeometric) test like FET can be somewhat conservative—mainly from the conditional distribution being "highly discrete." The probability of finding AI if it exists, therefore, may be considerably less than expected. Lydersen, Fagerland and Laake (2009) tend to agree. These researchers suggest that unconditional tests are generally more powerful than FET. However, when a conditional test like FET must be used given some underlying rationale, Lancaster's (1949) mid-*p* adjustment may be implemented to account for this general misalignment. Other adjustments to FET besides mid-*p* are available (Crans and Shuster, 2008; Seneta, Berry & Macaskill, 1999). For example, one approach referred to as "Adjusted FET" takes sample size and statistical power into account, and attempts to reduce the conservative nature of traditional FET by increasing the nominal significance level. Instead of using the typically accepted probability value of .05 (95% probability that a finding of AI did not occur just by chance), an alternatively less conservative probability value (e.g., .06) may be used. It should be noted that despite recent research in this area, traditional uncorrected FET will likely produce the fewest false positives (i.e., disparities that are erroneously flagged as statistically significant) in the vast majority of situations.

Concluding Points for Appropriate Adverse Impact Analyses

Understanding appropriate statistical significance standards used to enforce Federal law is important from both the enforcement and risk-management perspectives. Two common

[2] According to some experts, the hypergeometric model is the correct model to use when the values are fixed (Gastwirth, 1988). That is, we know exactly how many people will be selected. A double-binomial (unconditioned) approach for this type of analysis was at one time endorsed in the literature by Upton (1982), however, Upton (1992) has since revised that stance, endorsing a hypergeometric model instead.

[3] Recent scholarly material suggests that an unconditional model may be appropriate when an AI analysis is conducted for a scenario in which a selection instrument with a passing score is going to be administered (Harpe, 2009; Collins & Morris, 2008), since those from the population of test-takers that will meet or exceed the passing score—and possibly their accompanying demographic representation—cannot be known in advance. A double-binomial statistical test may be the best statistical method to use in this circumstance.

approaches to analyzing data with respect to AI—*Z* and FET—were discussed. When sample sizes are large, results will be very similar across these two methods; however, when sample sizes are small the *Z* test may not be accurate. In the past FET has been regarded as the uniformly most powerful unbiased test (Lang and Steel, n.d.). However, there is renewed debate in the field as to when FET is appropriate. Some argue that FET is not appropriate for analyses of instruments such as cognitive ability or math tests when the pass rates are not known in advance, because passing rates better represent a sampling model called the double-binomial model, which is designed to examine data that is not fixed in nature.

As discussed, when FET is used, modifications such as mid-*p* or Adjusted FET may be necessary to correct for the conservative nature of conditional tests (such as traditional uncorrected FET) given the general misalignment between nominal significance levels associated with tests for continuous data, and tests for discrete data. However, this should be done knowing that Federal regulatory enforcement agencies such as the OFCCP will likely continue the practice of using traditional FET, especially for AI analyses where there are a small number of applicants in the pool ($n<30$). As such, Barnard (1990) cautions that the work of regulatory agencies (such as the OFCCP) must "proceed by rule" and "statistical issues raised in such agencies have more in common with sampling inspection problems than with those involved in scientific inference." Realization and applicability of this issue has been slow to reach the context of AI analyses; however, arguments made for these adjustments are certainly noteworthy, and understanding why they are made may assist EEO professionals in identifying the most appropriate analyses for detecting AI in their organizations.

References

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science, 7*, 131-177.

Barnard, G.A. (1990). Must clinical trials be large? The interpretation of *p*-values and the combination of test results. *Statistics in Medicine, 9*, 601-614.

Collins, M.W. & Morris, S.B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology*, *93*, 463-471.

Colosimo, J., Dunleavy, E.M. & Cohen, D. (2010). *Primer adverse impact analyses*. Unpublished manuscript. (Available from DCI Consulting Group, Inc., 1920 I Street NW, Washington, DC 20006).

Crans, G.G. & Shuster, J.J. (2008). How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. *Statistics in medicine, 27*, 3598-3611.

Gastwirth, J.L. (1988). *Statistical reasoning in law and public policy* (Vol. 1). San Diego, CA: Academic Press.

Griggs v. Duke Power Company. 401 U.S. 424 (1971).

Harpe, L.D. (2009). Steps analysis—successfully hiring talent one step at a time. Session presented at the 27th Annual Industry Liaison Group National Conference, Atlanta, Georgia.

Lancaster, H.O. (1949). Statistical control of counting experiments. *Biometrika, 39*, 419-422.

Lang, E.L. & Steel, L. (n.d.). Issues relative to the statistical methods employed. In Palmer, C. & Gilmartin, K. (Eds.), *Application of statistical methods to the analysis of employment data: A guide to the use of statistics in the adjudication of discrimination claims* Unpublished manuscript. (American Institutes for Research, 1000 Thomas Jefferson Street Washington, D.C. 20007).

Lydersen, S., Fagerland, M.W. & Laake, P. (2009). Recommended tests for association in 2×2 tables. *Statistics in Medicine, 28*, 1159-1175.

Seneta, E., Berry, G., & Macaskill, P. (1999). Adjustment to Lancaster's mid-*p*. *Methodology and Computing in Applied Probability*, *1*, 229-240.

Uniform guidelines on employee selection procedures. Fed. Reg., 43, 38,290-38,315 (1978).

Upton, G.J. (1982). A comparison of alternative tests for the 2x2 comparative trial. *Journal of the Royal Statistical Society*, 145, 86-105.

Upton, G.J. (1992). Fisher's exact test. *Journal of the Royal Statistical Society*, 155, 395-402.

