Biddle Consulting Group, Inc.

Important Developments in the EEO Analysis Field: Practical Significance and Fisher Exact Tests

One of the benefits of supporting our clients in numerous Title VII cases and audits is that we stay current with new developments relevant to EEO analyses. Through these experiences, we learn of new techniques that should be adopted, as well as previously-applied techniques that should be modified or abandoned altogether. Two such developments have emerged over the last few years that have led our firm to make changes to the latest version of the Adverse Impact Toolkit (version 4.1) as well as our AutoAAP® program and related tools.

The first change we have made focuses on applying "practical significance" tests to adverse impact analyses. The second change deals with the Fisher Exact Test, which is a commonly-applied statistical analysis used when conducting the "Selection Rate Comparison" adverse impact analyses. Both changes are discussed in more detail below.

Practical Significance

Since the publication of our book, *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*, 2nd Ed. (2006), there are new cases that re-interpret practical significance tests that had been applied in previous cases,[1] as well as cases that take more narrow interpretations of practical significance that have been used by other courts[2] or that do *not* support practical significance evaluations[3] at all. While the Uniform Guidelines on Employee Selection Procedures (Guidelines) are clear that practical significance evaluations are *conceptually relevant* to adverse impact analyses, we suggest employers "tread carefully" when evaluating the practical significance of adverse impact analysis results in light of these "mixed" recent findings. We do not believe that hard-and-fast practical significance rules should be applied when analyzing adverse impact.

In enforcement and litigation settings, we have found that the determination of adverse impact is a process whereby the "collective evidence" is weighed, *starting with* a finding of statistical significance that exceeds 2 standard deviations. If this first threshold is met, additional evaluations, such as the 80% test (or the "impact ratio") and practical significance evaluations can be used to evaluate the sum of adverse impact evidence in the situation. This is because the true definition (per the Guidelines) of adverse impact is essentially a "substantially different rate of selection."

[1] *Delgado v. Ashcroft*, **No. Civ.A. 99-2311(JR),** May 29, 2003.
[2] *OFCCP v. TNT Crust* (US DOL, Case No. 2004-OFC-3).
[3] *Dixon v. Margolis* (765 F.Supp. 454, N.D.Ill.,1991), *Washington v. Electrical Joint Apprenticeship & Training Committee of Northern Indiana*, 845 F.2d 710, 713 (7th Cir.), cert. denied, 488 U.S. 944, 109 S.Ct. 371, 102 L.Ed.2d 360 (1988). *Stagi v. National Railroad Passenger Corporation*, No. 09-3512 (3d Cir. Aug. 16, 2010).

For example, in the most recent circuit-level case dealing with practical significance, the court stated:

> Similarly, this Court has never established “practical significance” as an independent requirement for a plaintiff’s prima facie disparate impact case, and we decline to do so here. The EEOC Guidelines themselves do not set out “practical” significance as an independent requirement, and we find that in a case in which the statistical significance of some set of results is clear, there is no need to probe for additional “practical” significance. Statistical significance is relevant because it allows a fact-finder to be confident that the relationship between some rule or policy and some set of disparate impact results was not the product of chance. This goes to the plaintiff’s burden of introducing statistical evidence that is “sufficiently substantial” to raise “an inference of causation.” *Watson*, 487 U.S. at 994-95. There is no additional requirement that the disparate impact caused be above some threshold level of practical significance. Accordingly, the District Court erred in ruling “in the alternative” that the absence of practical significance was fatal to Plaintiffs’ case [*Stagi v. National Railroad Passenger Corporation*, No. 09-3512 (3d Cir. Aug. 16, 2010)].

For these reasons, we recommend that our clients rely primarily on statistical significance for determining adverse impact, and consider the 80% test and shortfall calculations (i.e., the number needed to eliminate a statistically significant finding) as two methods for sensibly evaluating practical significance.

The Fisher Exact Test

Partially due to its name (“exact”), the Fisher Exact Test (“FET” hereafter) has long been used for analyzing adverse impact in selection and hiring decisions in the EEO field. The use of FETs in the Title VII enforcement settings (litigation and government audits) has typically been limited to situations where sample sizes are small (e.g., less than 30, or whenever one group’s passing numbers fall below 5), although it has often been calculated using a variety of approaches (one-tail significance and two-tail significance levels using a wide variety of methodologies) when evaluating 2X2 tables (where two groups are compared with respect to two outcomes—e.g., passing and failing).

While the FET has continued to be used in Title VII situations, the statistical and medical research fields have increasingly used alternative tests for 2X2 tables that more accurately address the statistical assumptions tied to the analysis, as well as increase statistical power to determine true differences between the two groups under study (e.g., Collins & Morris, 2008; Crans & Shuster, 2008; Lin & Yang, 2009). While several issues pertaining to the analysis of 2X2 tables have been debated since Fisher originally proposed his test (in 1934), two issues have drawn attention in recent years: the *conditional assumptions* relevant to the FET and its *conservative nature*. Each is discussed in more detail below.

Conditional Assumptions Relevant to the FET

The first issue deals with the *conditional assumptions* tied to the FET which are frequently not met in practice. There are three situations (or "models") in which 2X2 tables can be evaluated (see Collins & Morris, 2008):

> ***Model 1: Independence Trial*****—**The marginal proportions are assumed to be *fixed in advance* (i.e., proportion of each group and selection totals are fixed). The data are not viewed as a random sample from a larger population.
>
> ***Model 2: Comparative Trial*****—**Either the row or column totals are fixed. The applicants are viewed as random samples from two distinct populations (e.g., minority and majority). The proportion from each population is fixed (i.e., the marginal proportion on one variable is assumed to be constant across replications). The second marginal proportion (e.g., the marginal proportion of applicants who pass the selection test) is estimated from the sample data.
>
> ***Model 3: Double Dichotomy*****—**Neither row/column are assumed to be fixed. Applicants are viewed as a random sample from a population that is characterized by two dichotomous characteristics. No purposive sampling or assignment to groups is used, and the proportion in each group can vary across samples.

These three models can be summarized as "fixed margins," "mixed margins," and "free margins." The statistical field has more recently arrived at a consensus that the FET can *only be accurately applied in the first model*—the Independence Trial model (fixed margins). Because this model does not represent typical personnel selection data, "there is reason to question the appropriateness of the FET for adverse impact analysis" (Collins & Morris, 2008). The stringent requirements of the Independence Trial model requires that "both of the margins in a 2X2 table are fixed *by construction—i.e.*, both the treatment and outcome margins are *fixed a priori*" (Sekhon, 2005). In other words, for the conditional assumptions of the Independence Trial model to be met, the investigator needs to call out, in advance, the marginal totals of both the rows and columns *prior to conducting the experiment that will produce the numbers within each of the rows and columns*. This requirement is only rarely met when conducting adverse impact analyses. As Gimpel (2007) states, "Over decades there has been a lively debate among statisticians on the applicability of the conditional FET. The argumentation against the test mainly is that it conditions inference on both margins where only one margin is fixed by most experimental designs and the test is inherently conservative…the row and column marginal totals are fixed by the researcher prior to data collection" (p. 171).

Even in promotional analyses, where sometimes the total number of promotions may in fact be declared in advance of the promotional process, the conditional assumptions of the FET may be violated because the justification of conditional tests (those for "fixed" margins) "depends on the assumption that the process determining the fixed marginal counts *is not dependent on the process under study*…" (Gastwirth, 1997). In other words, the number of minority members hired out of a labor pool *should not provide information* about the odds ratio of the promotion rates, the parameter of interest. Gastwirth advises checking this assumption before calculating conditional tests in situations where the available sample results from a previous selection process that *may be affected by the same factors involved in the process being examined* (because the odds ratio of the hiring rates and promotion rates would be related).

In layoff (RIF) situations, where it appears that the row and column totals may be fixed in advance, sometimes the decision making process is *dynamic and continually flexible* until the final decisions have been made, thus weakening the strict conditional assumptions tied to the FET. In addition, the numbers in the process do in fact depend on the employer's previous selection and promotional practices.

When applying the three models above to typical adverse impact analyses (e.g., hiring, promotion, terminations, layoffs, etc.), it becomes clear that the conditional assumptions of the FET will only rarely be met. Furthermore, when the conditions are not met, the researcher is required to use either an unconditional test (such as Barnard's Exact Unconditional Test adopted in StatXact or the Boschloo Exact Unconditional Test[4] widely advocated in the statistical and medical fields) or a conditional test that adjusts for the conditional assumptions tied to the FET (e.g., the Lancaster mid-p, adopted in BCG's tools).

The Conservative Nature of the FET

The second issue with the FET pertains to its *unnecessary conservative* nature. This limitation occurs because the FET has "less power than conditional mid-p tests and unconditional tests" while these other tests "generally have higher power yet still preserve test size" (Lydersen, et. al, 2009). For this limitation alone, several statisticians have recommended that the "traditional FET should practically never be used" (Lydersen, et. al, 2009) because the "actual significance level (or size) being used is much less than the nominal level" (Lin & Yang, 2009). Agresti (2007, p. 48) recommends using the mid-p adjustment even in situations where the fixed marginal assumptions can be met "because the actual error rates [of the FET] is *smaller than the intended one*" (p. 48). Agresti states:

> For small samples, the exact distribution has relatively few possible values. The P-value also has relatively few possible values … discreteness affects error rates. Suppose like many methodologists, you will reject the null hypothesis if the P-value is less than or equal to 0.05. Because of the test's discreteness, the actual probability of type I error may be much less than 0.05. The [Fisher Exact Test] is conservative because the actual error rate is smaller than the intended one. To diminish the conservativeness, we recommend using the mid-p value. (p. 48)

Other categorical statistical books also recommend the mid-p as an effective way to meet the challenges offered by a discrete distribution when trying to apply a statistical test to correctly answer the .05 question (e.g., Simonoff, 2003; Rothman, 1986; Hirji, 2006).

While these issues have been percolating in the statistical journals for decades, the more recent application of Monte Carlo simulations conducted on the FET and similar tests now provide insight into just how conservative the FET can be, with estimates ranging between 50% below the desired .05 significance level (with small sample sizes around 50 or less) and 10-30%

[4] See, for example, http://www.stat.ncsu.edu/exact/

under the .05 level (with larger sample sizes ranging between 75 and 200). The net effect of these limitations relevant to the FET is that it sets a 2.4 standard deviation threshold (in small samples of 50 and smaller), rather than the desired 2.0 threshold, and a 2.1 – 2.2 standard deviation threshold in larger samples between 75 and 200. This can result in practitioners thinking they are using a 2 standard deviation threshold test to analyze adverse impact, while the conservative nature of the FET has the net effect of actually setting the bar higher for finding adverse impact.

Lancaster's Mid-p as a Viable Solution

For the reasons discussed above, we join with several other statisticians (see short list of journal articles and textbooks) advocating Lancaster's mid-p correction to the FET, which effectively corrects the FET to more accurately reflect the probability values of the adverse impact case analyzed in any one of the three models discussed above. While exact unconditional tests (like the widely-used Bernard or Boschloo Exact Unconditional Tests[5]) can be correctly used for mixed and free marginal situations, they were not tailored for fixed marginal situations. The mid-p, however, can be used for *all three* situations. While the mid-p is simply the FET with a statistical adjustment,[6] outputs from the mid-p typically correlate exceptionally high (r >.98) with the outputs from most unconditional exact tests.

There are several reasons that the mid-p is a "best choice" option as explained by Hirji (2006, pp. 218-219). In the Section titled, Why the Mid-P?, Hirji provides the basis for endorsing the mid-p as the preferred exact method (for either conditional or unconditional situations). These are:

1. Statisticians who hold very divergent views on statistical inference have either recommended or given justification for the mid-p method.
2. A mid-p version has been or can be devised for most of the statistics used in exact conditional and unconditional analysis of discrete data.
3. Mid-p versions for multi-parameter discrete data tests have also been devised.
4. The evidence function of the TST (twice the smaller tail) mid-p method is a continuous bimonotonic function (a function that, starting from a value at a central point, decreases evenly in either direction), giving a coherent set of p-values and nested, connected confidence intervals.
5. The shape and power function of the TST mid-p tests is generally close to the shape of the ideal power function. (This is an important distinction because it demonstrates that the power of the test is uniform, able to detect adverse impact when it exists across a variety of data sets—both balanced and unbalanced).
6. In a wide variety of designs and models, the mid-p rectifies the extreme conservativeness of the traditional exact conditional method without compromising the type I error in a serious manner.
7. The expected value of the one-sided mid-p is equal to .5. The TST mid-p has an empirical distribution that is close to the uniform (0, 1) distribution with expected values of around 0.5.

[5] See, for example, http://www.stat.ncsu.edu/exact/.

[6] The mid-p is simply calculated by first computing the natural FET then subtracting the hypergeometric probability value of the first observed set from the total cumulative probability from both tails.

8. Empirical studies show that the performance of the mid-p method resembles that of the exact unconditional methods and the conditional randomized methods.
9. With the exception of a few studies, most studies indicate that, in comparison with a wide variety of exact and asymptotic methods, the mid-p methods are among the preferred, if not the preferred ones.
10. The median unbiased estimate, derived from the TST mid-p evidence function, has good comparative small and large sample properties.
11. The progress in computing power and efficient algorithms has made computation of exact distributions, and thus the mid-p based indices, a practically feasible option for an extensive array of complex discrete models and designs.

Hirji concludes by stating: “The mid-p method is thus a widely-accepted, conceptually sound, practical and among the better of the tools of data analysis. Especially for sparse and not that large a sample size discrete data, we thereby echo the words of Cohen and Yang (1994) that it is among the ‘sensible tools for the applied statistician.’”

We are pleased to continue our research in these (and other related) areas of EEO analyses, and welcome feedback from other practitioners who can offer insights on the same.

Please feel free to contact us at (800) 999-0438 or staff@biddle.com for additional information on either of these matters.

Sincerely,

Dan A. Biddle, Ph.D.
CEO, Biddle Consulting Group, Inc.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley.

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

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* (8), 3598-3611.

Gimpel, H. (2007). Preferences in Negotiations: The Attachment Effect (Lecture Notes in Economics and Mathematical Systems): Author.

Hirji, K. F. (2006). *Exact analysis of discrete data*. CRC Press.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56, 223-234.

Lin, C.Y, Yang, M.C. (2009). Improved p-value tests for comparing two independent binomial proportions. *Communications in Statistics - Simulation and Computation, 38* (1), 78 – 91.

Lydersen, S. Fagerland, M.W., Laake, P. (2009). Recommended tests for association in 2X2 tables. *Statistics in Medicine, 28,* 1159–1175.

Rothman K, J. (1986). *Modern epidemiology*. Boston: Little, Brown.

Sekhon, J.S. (2005). Making inferences from 2 × 2 tables: the inadequacy of the Fisher Exact Test for observational data and a principled Bayesian alternative. Travers Department of Political Science, Survey Research Center, UC Berkeley: Author.

Simonoff, J.S. (2003). *Analyzing categorical data*. Springer-Verlag: New York, NY.

References Supporting the Use of Lancaster's Mid-p Adjustment to the FET

The FET assumes that the row and column totals are known in advance. In cases where this assumption is not met, the FET is very conservative, resulting in Type I error which is *below the nominal significance level* (typically .05). The strict "fixed margins known in advance" requirement of the FET is not met in many experimental designs and almost all non-experimental ones. A natural way to apply a continuity correction to adjust for this limitation is through the notion of a mid-p value, which is obtained by subtracting half the probability of the observed statistic from the cumulative, two-tail exact p-value. While this proposal was first made by Lancaster in 1961, it has since been supported by many other distinguished professionals (some of these are provided below).

Journal References

Barnard, G.A. (1989). On alleged gains in power from lower p-values. *Statistics in Medicine, 8*:1469-1477.

Franck, W.E. (2007). P-values for discrete test statistics. *Biometrical Journal, 28* (4), 403-406.

Hirji, K. F. (1991). A comparison of exact, mid-P, and score tests for matched case-control studies. *Biometrics, 47*: 487-496.

Hirji, K., Tan, S.J., Elasho, R.M. (1991). A quasi-exact test for comparing two binomial proportions, *Statistics in Medicine, 10,* 1137-1153.

Hirji, K., Tang, M.L., Vollset, S.E., Elasho, R.M. (1994). Efficient power computation for exact and mid-p tests for the common odds ratio in several 2X2 tables, *Statistics in Medicine, 13*, 1539-1549.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association,* 56, 223-234.

Miettinen, O.S., & Nurminen, M (1985). Comparative analysis of two rates. *Statistics in Medicine*, 4, 213-226.

Plackett, R. L. (1984). Discussion of Yates' 'Tests of significance for 2×2 contingency tables'. *Journal of Royal Statistical Society*, Series A, 147, 426-463.

Stone, M. (1969). The role of significance testing: some data with a message. *Biometrika, 56,* 485493.

Upton G. (1992). Fisher's exact test. *Journal of the Royal Statistical Society, Series A,* 155: 395–402.

Williams, D. A. (1988). Tests for differences between several small proportions, *Applied Statistics, 37*, 421-434.

Statistical Textbooks

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd ed., Wiley.

Anscombe, F. J. (1981). *Computing in Statistical Science through APL.* Springer-Verlag, New York.

Hirji, K. F. (2006). *Exact analysis of discrete data*. CRC Press.

Pratt J.W., Gibbons J.D. (1981). Concepts of Nonparametric Theory. Springer-Verlag, New York.

Rothman K, J. (1986). *Modern epidemiology*. Boston: Little, Brown.

Simonoff, J.S. (2003). *Analyzing categorical data*. Springer-Verlag: New York, NY.