2002

Comment on the Age Discrimination Example

Dale A. Nance
Case Western University School of Law, dale.nance@case.edu

COMMENT

Dale A. Nance*

CITATION: Dale A. Nance, Comment on the Age Discrimination Example, 42 Jurimetrics J. 341–346 (2002).

The Federal Judicial Center's Research Division has developed a prototype for the computer assisted education of judges with regard to the statistical analysis of evidence.[1] In one part, the prototype presents an analysis in the context of a hypothetical discrimination case in which patterns of employment termination are used to infer whether disparate treatment (or disparate impact) on account of age occurred in the company's employment decisions. The materials develop the plaintiff's statistical case and the defendant's statistical reply. Along the way, the user is introduced to a number of mathematical concepts as well as several specific "significance tests," tests used (in this context) to assess whether the apparent difference in the treatment of employees based on age is a real difference or just an artifact of sampling.

Since I have been asked to comment on the quality of these learning materials, I will say that, on the whole, they seem to work quite well. The user is given just what is needed to understand the analysis without being overwhelmed by the mathematics. In fact, the presentation effectively concedes that the user may not understand the mathematics in detail and focuses on conveying an understanding of the structure of the arguments made and supported by standard statistical techniques. Presenting the materials in the context of proofs and counter-proofs in a hypothetical case assists in understanding not just the

*Dale A. Nance is Professor of Law, Chicago-Kent College of Law, Illinois Institute of Technology.

1. Robert Timothy Reagan, *Federal Judicial Center Statistical Examples Software Prototype: Age Discrimination Example*, 42 JURIMETRICS J. 281 (2002).

statistical tools, but how they are encountered in litigation. I should, perhaps, qualify this praise by confessing to a mathematics background that is probably more extensive than the average judge. Consequently, perhaps there are significant difficulties in the presentation that I failed to spot. Still, I think the prototype is a very useful tool, especially with the incorporation of a few changes.

I have three criticisms, two minor, the third more fundamental. First, the prototype does not indicate the extent of professional disagreement about the appropriateness of certain statistical calculations. For example, in presenting the plaintiff's case, the prototype illustrates the use of the "chi-squared" test in assessing whether an employee's risk of termination was affected by whether the employee was over or under 40 years old. The prototype then simply states a rule of thumb taken from a standard statistical text, namely, that "[u]se of the chi-squared test... traditionally was considered permissible when (i) the total number of observations, N, is greater than 40, or (ii) the total number of observations, N, is from 20 to 40, and all expected cell values are greater than or equal to 5."[2] (In the hypothetical context, N is the number of employment decisions, positive or negative.) This is one rule of thumb encountered, but it is not the only one,[3] and there is considerable theoretical disagreement on the proper rule, disagreement that ultimately relates to the nature of random sampling.[4] I would not suggest that the prototype go into such debates, but it would be worthwhile to reveal that the quoted rule of thumb is not carved in stone.[5]

My second minor complaint concerns the prototype's otherwise helpful explanation of the difference between "one-tailed" and "two-tailed" significance tests.[6] Roughly speaking, the "*p*-value" here is the probability of getting the observed data on employee discharges if the "null hypothesis" that there is no disparate impact based on age for all termination decisions is true.[7] More

---

2. *Id.* at 288 (Screen 4.3.1.2) (citing GEORGE W. SNEDECOR & WILLIAM G. COCHRAN, STATISTICAL METHODS 127 (8th ed. 1989)).

3. One standard text recommends the following "conservative" rule of thumb: "For tables with more than a single degree of freedom, a minimum expected frequency of 5 can be regarded as adequate for carrying out the Pearson chi-square test of association. However, when there is only a single degree of freedom, a minimum expected frequency of 10 is much safer." WILLIAM L. HAYS, STATISTICS 862 (5th ed. 1994).

4. *See, e.g.*, Neal E.A. Kroll, *Testing Independence in 2x2 Contingency Tables*, 14 J. EDUC. STAT. 47 (1989); Graham J.G. Upton, *A Comparison of Alternative Tests for 2x2 Comparative Trials*, 145 J. ROYAL STAT. SOC'Y (SERIES A) 86 (1982).

5. HAYS, *supra* note 3, qualifies the rule of thumb with the following statement:

> This rule of thumb is ordinarily conservative, and circumstances may arise in which smaller expected frequencies can be tolerated. In particular, if the number of degrees of freedom is large, it is fairly safe to use the Pearson chi-square test for association even if the minimum expected frequency is as small as one, provided that there are only a few cells with small expected frequencies (such as one of five or fewer).

Id. at 863.

6. Reagan, *supra* note 1, at 286 (Screen 4.2.3).

7. The *p*-value of a test statistic may be thought of as the "conditional false positive" rate for "rejecting the null hypothesis." In this context, if one were to infer disparate impact based on age in a large number M of cases with the same value of the test statistic, then, assuming no disparity, one would incorrectly infer disparity in about p×M of such cases.

precisely, for a one-tailed test in this context, the relevant "*p*-value" to be considered is the probability that the observed termination rate among the over-40 group would be *at least as much higher* than that for the under-40 group as was in fact observed *if* the probability of discharge was no different for the two groups. For a two-tailed test, the *p*-value is the probability that the observed termination rate among the over-40 group would be *at least as much higher or lower* (than that for the under-40 group) as the magnitude of the difference that was in fact observed *if* the probability of discharge was no different for the two groups.

This is well explained in the prototype's discussion of the use of Fisher's exact test, but that discussion omits an answer to a potentially important question: Which is the proper test (and associated *p*-value) to use, a one-tailed test or a two-tailed test? The discussion in this screen of the prototype simply concludes by relying on the two-tailed test *p*-value (0.0003) to infer that the pattern of employee discharge was unlikely to be the result of chance under the indicated assumption. But the prototype does not explain this choice of the two-tailed test *p*-value, perhaps because the one-tailed *p*-value is also very small (0.0002). This question is most troublesome if the size of the *p*-value becomes determinative, because the one-tailed *p*-value will always be smaller than the two-tailed *p*-value.[8] Whether the prototype's discussion is intended as suggesting that the two-tailed test should always be used should be clarified.[9]

This point brings us to my major concern. How large can the *p*-value be before the data are considered unrevealing? A typical convention in social science refuses to infer a difference in the underlying populations if the *p*-value exceeds 5% (0.05), that is, if the probability that the observed difference (or a greater one) would arise from mere chance (by sampling from populations that are equivalent using the parameters of interest) is greater than 0.05.[10] The idea is that we then should be content to allow the issue to remain in limbo pending further study. Generally speaking, however, the law has no such leisure. A finding that discrimination has not been proved is, for most practical purposes, equivalent to a finding that discrimination did not occur; principles of res judicata generally prevent readjudication of the claim at a later time, even if further evidence is obtainable. Yet trials cannot be indefinitely postponed pending further study, and courts should proceed on the best evidence then available.[11] It is, therefore,

---

8. *See* Daniel L. Rubinfeld, *Guide to Multiple Regression*, *in* 1 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY § 4-3.1.3 (David L. Faigman et al. eds., 2d ed. 2002).

9. *Cf.* Richard Goldstein, *Two Types of Statistical Errors in Employment Discrimination Cases*, 26 JURIMETRICS J. 32, 42–47 (1985) (endorsing judicial receptiveness to one-tailed tests in discrimination cases).

10. *See* David H. Kaye & David A. Freedman, *Statistical Proof*, *in* 1 MODERN SCIENTIFIC EVIDENCE, *supra* note 8, at §§ 3-4.2.1, 3-4.2.2.

11. I do not mean to suggest that courts should be entirely passive in accepting parties' claims that the evidence they present is in fact the best reasonably available. Active steps may be necessary to assure that the evidence considered is the best reasonably available at that time. *See, e.g.*, Dale A. Nance, *Evidential Completeness and the Burden of Proof*, 49 HASTINGS L.J. 621, 625 (1998)

important to recognize that when science is put to use in the service of the law, the legal system may need to employ somewhat different standards than science might otherwise employ.[12]

The prototype's discussion of the implications of the defendant's evidence in the hypothetical case illustrates the importance of these points. The defense uses a multivariate logistic regression analysis to support its claim that the company chose to consider, as one factor in discharge decisions, the length of employment (as distinguished from employee age), because the longer the employee had been with the company, the more loyalty the employee might be expected to have toward the old management. Thus, the defendant asserted, employee seniority rather than age was the factor cutting against employee retention. But because those with greater seniority also tend to be older, an *apparent* discrimination (disparate impact) against older employees appears. Now, the defendant's regression analysis actually suggests that *both* seniority *and* age are factors explaining termination, because when one restricts attention to those of comparable seniority, the regression reveals that older employees were still discharged at a higher rate. To be sure, differences in seniority account for more of the variation in discharge rates than do differences in age, but age discrimination may still have occurred.

To resolve this remaining issue, the prototype draws on the *p*-values for the regression. The *p*-value for the association with seniority is 0.019, while the *p*-value for the association with age is 0.307.[13] Based on these *p*-values, the prototype concludes that "[t]he statistical evidence supports defendant's argument that its termination decisions were based on factors that included length of employment, but did not include age as a separate factor."[14] The basis for this conclusion, however, is not explained. Although the prototype does not state that when a *p*-value is smaller than some conventional figure, like 0.05, then the association must be considered unproved in litigation, something like this appears to be entailed in the conclusion just quoted. Apparently, the *p*-value of 0.019

(recommending that the burden of production be understood as requiring evidence that is reasonably complete and explaining how this might be implemented).

12. *See generally* Peter Donnelly & Richard D. Friedman, *DNA Database Searches and the Legal Consumption of Scientific Evidence*, 97 MICH. L. REV. 931, 969–78 (1999). Beyond that, there is considerable skepticism about the selection of a particular *p*-value as marking the difference between usable results and unusable results, even in the context of relatively pure science. *See generally* THE SIGNIFICANCE TEST CONTROVERSY: A READER (Denton E. Morrison & Ramon E. Henkel eds., 1970).

13. Reagan, *supra* note 1, at 293. Without going into the details, these *p*-values relate to the coefficients for seniority and age in the regression equation. *See* Reagan, *supra* note 1, at 293–94 (Screens 6.2– 6.2.2). In this context one must distinguish between the strength of an association between two variables and the strength of the inference that such an association exists. The coefficients for seniority and for age in the regression equation estimate the strength of the association between those factors and the termination decision, while the *p*-values for those coefficients give the probability of observing an association of at least that magnitude if the actual coefficient is zero. In a given case, one might be very confident that an association exists, but the magnitude of that association might be practically unimportant. *See* Rubinfeld, *supra* note 8, § 4-3.1.

14. Reagan, *supra* note 1, at 294–95 (Screen 7).

authorizes the inference of an association between seniority and discharge because 0.019 is less than 0.05. At the same time, the *p*-value of 0.307 is greater than 0.05, and the prototype's conclusion, that an association between age and termination is not supported, might seem to follow.

If this is, indeed, the basis for the prototype's conclusion, the prototype inappropriately applies a convention from social science to the resolution of a legal dispute. There is, to be sure, judicial authority for this methodological transfer.[15] Nonetheless, the 0.307 probability of a coefficient as large or larger than that found for the age variable under the null hypothesis does not necessarily mean that the data are not probative of disparate treatment by age. Nor does it mean that we should consider the hypothesis of such disparate treatment unproved by legal standards. Whether the data are probative of disparate treatment (i.e., of the existence of a real association between age and termination, regardless of the strength thereof) depends as well on the probability that one would obtain at least such a disparity *if* there were disparate treatment by age.[16] As long as this probability is larger than the *p*-value, as it may well be in the context of the hypothetical, then there is reason to think the evidence is probative and favors the plaintiff, assuming that the expert is not considered wholly incredible. To be precise, the data are probative of a specific association between age and termination if it is more likely that one would observe these data when there is that association than when age and termination are not associated.[17] Whether, in turn, disparate treatment is proved to the applicable legal standard depends not only on this likelihood ratio, but also on the other, nonstatistical evidence in the case.[18]

Without performing such assessment, going beyond the calculation of the *p*-value, all one can say about the relationship between the *p*-value for the age-termination association and the proposition that disparate treatment has occurred is that, *ceteris paribus*, the smaller the *p*-value, the more the data support an inference of disparate treatment.[19] As Professor David Kaye has nicely put it:

---

15. *See, e.g.*, Segar v. Smith, 738 F.2d 1249 (D.C. Cir. 1984).

16. If the *p*-value is thought of as a conditional false positive rate (*see supra* note 7), then the probability mentioned here should be thought of as a "conditional *true* positive," the rate at which one would correctly infer disparity if one consistently inferred it from the test statistic computed for these data and the hypothesized disparity were true. This is what statisticians refer to as the "power" of a test. *See* Kaye & Freedman, *supra* note 10, § 3-4.3.1; Goldstein, *supra* note 9, at 34–42.

17. This is well understood in the context of forensic identification evidence. Conditional false positive rates quite analogous to *p*-values are accepted as probative even though they are much higher than 0.05. *See, e.g.*, People v. Mountain, 486 N.E.2d 802, 805 (N.Y. 1985) (holding that evidence of a match between the defendant's blood type and the blood type of the perpetrator is *relevant*—not to say dispositive beyond reasonable doubt—to prove identity, even though the probability of getting such a match *if* the defendant were innocent was as high as 0.40); 1 MCCORMICK ON EVIDENCE § 205, at 753 (John W. Strong ed., 5th ed. 1999).

18. *See* David H. Kaye, *Statistical Significance and the Burden of Persuasion*, 46 LAW & CONTEMP. PROBS. 13, 22–23 (1983).

19. The problem surfaces in a more subtle way in the prototype's discussion of the two-tailed *p*-value for the plaintiff's use of the Fisher's exact test (0.0003). The prototype states, "a disparity in termination rates between the two age groups at least as large as that observed would result by chance

> [T]here is no sharp border between "significant," and "insignificant." Although a few commentators and courts have inadvertently suggested otherwise, as the *P*-value decreases, evidence gradually becomes stronger. As a result, most modern statistics texts and journals discourage the reporting of results as "significant" or "insignificant" in favor of explicit statements of *P*-values. Courts should do likewise. There is no strictly objective basis, in science or in anything else, for believing that a proposition is true simply because the evidence for it is "statistically significant" at the .05 level.[20]

The net effect of using the relatively conservative (because larger) two-tailed probability that is compared to a conservative (because small) significance level (0.05), concluding that the hypothesis of discriminatory treatment is unproven if the former is larger than the latter, is to reduce the number of cases in which discrimination is found. That is not a reason to reject the conservative conventions, but it emphasizes what is at stake in choosing appropriate conventions and suggests that we need to examine the reasons carefully. It would be very helpful if the prototype explicitly acknowledged these issues.

---

approximately three in every ten thousand times. *Because such a result would be so unlikely, this statistical analysis supports plaintiff's prima facie case . . . .*" Reagan, *supra* note 1, at 286 (Screen 4.2.3) (emphasis added). The italicized conclusion is misleading, if not incorrect, because all the statistician can infer from the small *p*-value is that the statistical analysis supports the plaintiff's prima facie case *more than it would* if the *p*-value were larger, *ceteris paribus*; even a substantially larger *p*-value would not, by itself, imply that the plaintiff's case is *not* supported. To draw the italicized conclusion, one must have assessed the conditional true positive rate as well (*see supra* note 18), about which the prototype is silent.

20. David H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1344–45 (1986).