Journal of Forensic Economics 21(1), 2009, pp. 99-103

Comment: "A Closer Look at Correction for False Discovery Bias When Making Multiple Comparisons"

A. E. Rodriguez*

The John D. Finnerty paper in this issue discusses the advantages and explains the usage of the Benjamini-Hochberg (BH) correction for the error-rate, in instances of multiple tests conducted in a statistical analysis. The paper recommends the BH procedure over the Bonferroni or Šidák correction. Bonferroni and Šidák are considered more "conservative." In other words, Bonferroni and Šidák set a higher threshold for a plaintiff's statistical test to overcome. The Šidák adjustment procedure was proposed by David I. Tabak in an earlier volume of this *Journal* (Tabak, 2006).

There are two issues of relevance raised originally by Tabak's paper and now by Finnerty's paper but only one in contention. First: Bonferroni, Šidák, or Benjamini-Hochberg? Second: do we need to adjust error rates in the first place?

The first question is uncontroversial once it is determined one needs an adjustment. Which method to use (and there are many) is a matter of preference and a decision whose merits cannot be statistically established ex ante. A popular text puts it as follows:

> Several other procedures, such as Tukey's procedure and Duncan's multiple-range test have been developed to help in such situations. However, there is considerable controversy in the statistical community as to which procedure is "best." (Anderson, Sweeney and Williams, 2008 at 512)

The second question is more difficult, and one to which neither Finnerty nor Tabak provides good guidance.

The publication of Tabak's paper in this *Journal* (2006) left the impression that an error-rate adjustment was obligatory in *all* instances of multiple testing. Finnerty's paper only enhances this impression. Admittedly, both Tabak and Finnerty recognize exceptions to the rule but focus more on illustrating their preferred method (and with pummeling the reader with the spectre of Daubert) than on helping the reader understand when error rates should be adjusted and when they should not.

Why Should We Care?

As all NAFE practitioners know, a statistical method published in a refereed publication enters the realm of "theory and technique generally accepted by a scientific community," to use the language of Daubert. Thus, one can envi-

*Associate Professor, Department of Economics & Finance, University of New Haven. The author thanks Dwight Steward for insightful comments on an earlier draft.

sion a situation whereby opposing counsel, citing Tabak (and now Finnerty) as authority, moves to impugn an expert's opinion for failing to adjust error rates, with little regard for whether the adjustment is warranted. To be sure, the contrary argument can be clarified in redirect or in responsive pleadings. Still, this is a headache no one needs, especially when it is not entirely clear if and when an adjustment is required.

My apprehension is not entirely speculative. A challenge over the issue of omitting error rate adjustments was raised in a recent race and gender discrimination matter (*EEOC v. Autozone Inc.*, 2006).[1] The opinion concerns a motion for summary judgment filed by defendant Autozone alleging (among other things) that plaintiff's expert failed to employ a Bonferroni adjustment for multiple testing, in effect impugning plaintiff's expert statistical results. Plaintiff responded by claiming that multiple-testing is applicable in epidemiology and medical testing but not in employment litigation and therefore was not obliged to employ adjustments. The motion was granted in part and denied in part.

Joseph Gastwirth recently published an insightful comment on this particular case (Gastwirth, 2008). In his commentary Gastwirth appears to celebrate the judge's decision to dismiss the matter on the basis of the multiple testing criticism raised by the defense. To wit: "The decision is noteworthy because the judge realized that the principles of statistical inference remain the same regardless of the origin of the data" (Gastwirth, 2008 at 1). The Court however did no such thing. It effectively punted in frustration, unable to draw any conclusive assistance from either expert.

> Given the contradictory views on the use of statistical adjustments, particularly the Bonferroni adjustment, the court does not have a sufficient basis to find statistical adjustment was required in this case or that the non-utilization of any statistical adjustment makes Dr. Barnow's results unrealiable. Therefore, the court will not grant summary judgment on the EEOC's pattern or practice claims on this basis. (*EEOC v. Autozone Inc*. 2006, at 9)

Yet, *in this particular instance* Gastwirth is right. And, at the same time, the trial court judge is entirely correct in noting the confusion in the matter.

The debate over error rates has been raging in other fields, especially epidemiology (Goodman, 1998) since at least the early 1990s. In fact, the debate started practically in the early days of statistics amidst the debates between Fisher on the one hand and Neyman and Pearson on the other (Goodman, 1998).

Anticipating my point below, there are instances where adjustments may be required, such as instances of data mining as Frank Denton warned some time ago (Denton, 1985). From what one can infer based solely on reading publicly available material, the Autozone case proved to be a "fishing expedition" on the part of plaintiff's expert, perhaps warranting an error-rate adjustment. Incidentally, plaintiff's is a peculiar defense. Follett and Welch addressed er-

[1]I don't know if Tabak's paper was used as basis; I suspect not based on the relevant dates.

ror-rate adjustments in employment litigation over 25 years ago (Follett and Welch, 1983).

Why Do We Need Any Adjustments to an Error Rate or p-value?

The error rate sets a threshold that determines whether an observed result is so unlikely to have occurred by chance alone that it enables us to conclude as follows: the data appear to be consistent with an inference of discrimination. Implicit in this statement is the recognition that a significant difference may be observed by chance if a null hypothesis is true. Incorrectly declaring a null hypothesis rejected solely because of random error is called a Type-I error, or a false positive.

Standard scientific practice commonly establishes a cutoff point to distinguish statistical significance at the 0.05 level, a threshold well ensconced in law. This level implicitly assumes that we have accepted the possibility of drawing a significant outcome by chance 1 out of 20 times. We are after all, looking at one sample, one draw, from the whole population, and may have had the misfortune, as the case may be, of drawing the long straw. When more than one test is conducted, the chance of finding at least one test statistically significant due to chance increases (Tabak, 2006 and Finnerty, 2009).

Where Is the Confusion?

Multiple-comparison adjustments did not originate with Bonferroni. Rather, they go back to Neyman and Pearson (1928). Neyman and Pearson called for the use of their statistical test adjustment for *repeated testing of samples drawn from the same data set*. Neyman and Pearson were examining rates of defective materials and rejection of lots where there were multiple samples drawn from each lot. For a more relevant example, suppose a supermarket chain wonders whether sales of cereal vary depending on shelf height.[2] This inquiry is easily handled via a straightforward stratification and combination method, Gastwirth's recommended approach. If, in addition, the supermarket wants to know whether there are *individual differences* in cereal sales *between* the shelves, then an adjustment is necessary. Hence one goes to Bonferroni or BH or any other. In other words the adjustment applies when we are testing whether the null holds for *multiple* unknown parameters drawn from the same data set.

In our NAFE world, such an application would emerge if one were to test substrata of a given sample. Thus, to use the discrimination example in the Finnerty paper, one would have to adjust if the data in a racial discrimination case were subsequently sliced by gender, age, income groups, ad infinitum. Again, reading Pearson and Neyman correctly, one would also be obliged to adjust if the procedure is such a "fishing" expedition (Denton, 1985).

It is *not* clear to me that one would have to adjust when the testing is conducted on the various constituent departments or division of the hypothetical

[2]This example is from Albright, Winston and Zappe 2004, at 782.

defendant firm *unless* the narrative recognizes that it is either a fishing expedition or a substrata test—as was clearly the case seeking to determine the influence of shelf level on cereal sales, or, as it appears to be the case, in *Autozone*. Note that we were able to arrive at this conclusion only because we were made aware that the expert had tallied *and reported* the number of regressions and manifold permutations conducted.

As for the two papers, the confusion arises because the examples proffered by Finnerty and Tabak could be construed to be instances of substrata testing or fishing—but the context provided in the paper by the authors is not sufficient to recognize whether this is in fact the case. It the narrative reveals the exercise to be data fishing then we must adjust. Otherwise, there is no need. In the manner in which the examples are presented in the Finnerty paper, the statistical tests don't appear to warrant adjustment. They appear to be just separate tests.

More problematic is the case where one conforms to statistical fidelity. Statistical purity appears, on its face, silly and nonsensical. It would require error-rate adjustments for all sorts of testing: by my research assistant, by third parties and so forth, every time the data is used. We would be obliged to adjust, adjust, and adjust some more, every time additional data were received, a report revised, the empirics updated at the request of counsel. The inevitability of receding into this miasma was, in fact, the main point in Denton's paper. Denton holds that distorting error rates is an inescapable fate even if one attempted to shun "data mining:" *"the use of the same data set by more than one investigator distorts the probabilities associated with reported hypothesis tests, even if no individual investigator engages in data mining"* (Denton, 1885, at 124).

Statistical fidelity, if one is true, would inevitably create an ethical-moral hazard problem, where we would limit the number of reported results, calculatingly shade our exploratory data efforts, deliberately muddle discussion on our methodology and generally embrace a lack of candor.

What to Do?

There are obvious situations where adjustments are required: arbitrarily slicing and dicing the same data set is one of them. In situations where substrata need testing, combination methods are applicable. I suspect that many NAFE practitioners are already relying on stratification and combination methods. These aggregations are required for handling small data sets, such as when the number of hires, or number of employees laid off, and so forth, are not enough to ensure adequate testing.

In all other situations, I err on the side of practicality and transparency. The best tonic is to set forth one's analysis and methodology and establishing their relevance within the overall context. After all, the inference of discrimination has to be based on admissible evidence and the statistical report is unlikely to be the sole probative element considered. Alluding to corroborating or contradictory evidence should convey a more thorough impression of the likelihood of discrimination and the validity of the statistical study.

The expert should report p-values and proffer an opinion as to why adjustments may or may not be necessary. Importantly, the expert should explain the tradeoff entailed in tweaking the Type-I error rate. In seeking increasingly stringent errors rates to accommodate the increased likelihood of false positives, we increase the likelihood of false negatives. That is to say, we increase the chances of failing to detect instances where the null is rejected. This means that there is an increasing chance that a deserving plaintiff will fail to obtain relief as a result of statistical artifact. These are the elements of this cost-benefit analysis. It means balancing the benefit of statistical fidelity on the one hand, a goal that may be inherently elusive against the increased possibility that we exclude someone from their day in court.

The second point that needs to be conveyed is Denton's argument that the nature of relying on an expert witness to opine on statistical analysis in a legal proceeding inexorably and inevitably leads to the distortion of the relevant probabilities, a matter beyond the control of any one person. As a result, the applicable error rate to emerge out of the serial compounding of these multiple instances of uncertainty is impractical and probably unknowable.

In this manner the basis and know-how for determining whether the results are "significant" is provided to the trier-of-fact, warts and all. It is they who should have the ultimate say.

References

Albright, S. Christian, Wayne L. Winston, and Christopher Zappe, *Data Analysis for Managers*. Belmont, CA: Brooks/Cole-Thomson, 2004.

Anderson, David R., Dennis J. Sweeney, and Thomas A. Williams, *Statistics for Business and Economics*, Mason, Ohio: Thomson South-Western, 10e, 2008.

Denton, Frank T., "Data Mining as an Industry," *The Review of Economics and Statistics*, 1985, 67(1), 124-127.

Finnerty, John, "A Closer Look at Correcting for False Discovery Bias When Making Multiple Comparisons," *Journal of Forensic Economics*, December 2009, 21(1).

Follett, Robert, and Finis Welch, "Testing for Discrimination in Employment Practices," *Law and Contemporary Problems*, 1983, 46(4), 171-184.

Gastwirth, Joseph L., "Case comment: An Expert's Report Criticizing Plaintiff's Failure to Account for Multiple Comparisons is Deemed Admissible in *EEOC v. Autozone*," *Law, Probability and Risk*, 2008, 7(1), 61-74.

Goodman, Steven N., "Multiple Comparisons, Explained," *American Journal of Epidemiology*, 1998, 147, 807-812.

Neyman, J., and E. S. Pearson, "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," *Biometrika*, 1928.

Tabak, David, "Multiple Comparisons and the Known and the Potential Error Rate," *Journal of Forensic Economics*, 2006, 19(2), 231-236.

Court Cases

EEOC v. Autozone Inc., 00-2923 Ma/A, US District Court for the Western District of Tennessee, August 29, 2006.