P-hacking’s day in court

A.E. Rodriguez
University of New Haven

Glenn McGee
University of New Haven

ABSTRACT

Concerns over potential error rates resulting from multiple testing have led researchers and scientists to call for adjustments in research protocols and in published work. *p*-value adjustments foil or mitigate the likelihood of false positives by raising the burden of statistical significance. Similar demands can be found in forensic settings where experts routinely offer statistical testimony in support of litigation. Calls for the adjustment of error rates in the reports and testimony of statistical experts in legal proceedings raise special ethical considerations that are distinct from those in medical and health-related fields.

In employment discrimination litigation, for example, statistical testing is often dispositive. Raising the bar associated with a particular statistical test to allegedly reduce the likelihood of a mistaken finding of discrimination when none exists, fails to recognize that the adjustment simultaneously diminishes the ability to detect the presence of a legitimate violation when one actually exists. If, in fact, there did exist discrimination the cost may be justice denied.

A sole focus on Type I error rate adjustments aimed at curtailing false positives, driven by fears of frivolous lawsuits, is to the detriment of Type II error and the increased likelihood of denying a worthy plaintiff his day in court.

Forensic statistical experts cannot accept or support “standalone” error-rate adjustments because of the duty of impartiality to both plaintiffs and defendants. Experts should acknowledge the potential need for adjustment in situations where they may be warranted but should advance these arguments in an expository manner disclosing the likelihood of all (Type I and Type II) error probabilities. An illustrative example is provided to demonstrate the relative simplicity of calculating the power of the utilized test. An enhanced understanding of the relative tradeoffs of error rate adjustments enhances the chances of a correct decision by the trier-of-fact.

Keywords: Daubert, expert witness testimony, error-rates, Bonferroni, moral hazard, statistical power, multiple-adjustments.

In a nutshell, an ethical dilemma exists when the entity conducting the significance test has a vested interest in the outcome of the test. (Ziliak & McCloskey, 2010)[1]

Those who put faith in the labors, expertise and testimony of experts have a right to expect that experts, themselves, will feel bound to an ethical obligation to research, analyze and report findings in an impartial, reliable manner (Faigman 1999, 2000). Outside the courtroom, questions are being raised anew about the credibility of researchers in the sciences, especially the social and health sciences (Carey, 2015; THE ECONOMIST, 2013). Although apprehensions with null hypothesis statistical testing are hardly novel (Cohen 1994; Jeffreys , 2003), the more recent criticism about error rates in science has re-emerged with a focus on selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results (Ioannidis, 2005; Nuzzo, 2014; Open Sciences Collaboration, 2015).

In the courtroom, experts called to provide statistical findings are increasingly asked to defend their conclusions not only against other experts, competing hypotheses, spurious allegations, and alternative data, but also against the possibility of deriving significant findings from possibly incidental inquiries – a criticism akin to those expressed across the sciences. For instance, Hersch & Bullock rue the obfuscation characterizing challenges to the use of regression analysis as an analytical tool in employment discrimination. “All too often, when a party presents regression analysis to assist its case, the opposing party launches spurious critiques challenging the validity of the analysis (Hersch & Bullock, 2014).” McCloskey & Ziliak offer testimony against unquestioned reliance on significance testing at the expense of practical importance (McCloskey & Ziliak, 2011). Ziliak & McCloskey, have alleged that the researchers responsible for Merck’s Vioxx clinical trials “fudged the data” to favor its case. “But an outsider could be forgiven for inferring that they dropped the three observations in order to get an amount of ststistical significance low enough to claim – illogically, but this is the usual procedure – a zero effect (Ziliak & McCloskey, 2009).” And with characteristic directness, Kaye notes “statistically significant results are nice to have. Scientists like them, and now litigants who rely on statistical evidence also want them. But the mere fact that an expert states that data are "significant" does not necessarily mean that the evidence satisfies the applicable burden of persuasion (Kaye, 1983).”

One of the areas singled out for criticism is the repeated application of significance testing – which complicates the interpretation of error rates. Significant findings will emerge if enough tests are applied, even when there is no real effect – an artifact of sampling known as a Type I error and popularly as p-hacking.

In forensic statistical testing, an expert has considerable leeway and incentive to explore all relevant permutations of the data to draw the best outcome possible. Most of this work is often necessary, and professional judiciousness and reputational concern ward off most prospective abuses. But a moral hazard problem arises, nonetheless, because it is difficult to distinguish if a white paper or forensic report proffered in support of a legal claim has been unduly manipulated in support of a frivolous claim. In response, and rather than challenging the substance of the empirical work, it is often easier for the defense simply to advocate for an adjustment, to allegedly account for the statistical prejudicial nature of

[1] Brief for Respondents as Amicus Curiae *Matrixx Initiatives, Inc. et al v. Siracusano et al*, 563 U.S. (2011), (Docket No. 09-1156.), at 5.

plaintiff's actions. Error rate adjustments are invoked, warranted to ward off the possibility of spurious false positives.[2]

However, defense's incentive to challenge plaintiff's expert report is just as appealing in instances where the analysis may be procedurally and methodologically correct. Indeed, because of the asymmetry of information and the nature of model-building and statistical testing, any report can be impugned by this reasoning. Whether plaintiff expert's report is statistically "by the book," as to the numerous and necessary judgments undertaken, cannot be verified. Before its current incantation, the term data mining was considered a disparaging term. Data mining was considered a vile act akin to "massaging the data." To massage the data is of course, still considered unprofessional – but the term "data mining" itself appears to have lost its edge. One would have been "data mining," for example, if one tested numerous variables onto a multiple regression model searching for best fit or for statistical significance; or selecting a particular time span of data after testing time spans of various lengths in an analysis of time series. In principle, one should account for the impact of this selectivity by removing a "degree of freedom" or by using higher (than the conventional) nominal significance levels. Clearly, whether an expert massaged the data or the model or even if she in fact, adjusts the degrees of freedom accordingly, is unknown. Denton, (1985) At stake is the soundness of the judicial system's ability to adjudicate correctly, its ability to minimize its own error rates.

Clearly, understanding whether to adjust for error rates should be of significant interest to forensic experts and jurists just as it is a matter of great concern for medical and psychological researchers.[3] If the trier of fact remains without an understanding of the reasons underscoring the calls for, and opposition to, adjusting error rates, in practice, allegations of "flawed statistical methodology" surrounding the absence or presence of a multiple comparison adjustment in a proffered expert report, stands a considerable chance of simply contributing to the fog that envelops dueling expert witnesses in court proceedings (Faigman 1999, 2000; Hersch & Bullock, 2014; Solow 2006; Posner, 1999).

The concern is not merely theoretical. Consider a challenge over the issue of omitting error rate adjustments raised in a recent race and gender discrimination matter.[4] The pertinent 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. The Bonferroni adjustment sets the *p*-value at which the test is evaluated for significance based on the total number of tests being performed. Specifically, the *p*-value ultimately utilized in a Bonferroni

[2] The influential paper by Holm (1979) appears to have "popularized" the relevance of sequential Bonferroni. Yet, it appears that the debate over whether to adjust first emerges among *forensic statistical experts* with Tabak (2006). Finnerty (2009) subsequently recommended adjustments to the proper error-rate adjustment methodology while acknowledging the troubling Type I-Type II error rate tradeoff that arises with adjustment.

[3] Although our main point is a general one, our comments are directed largely to statistical experts who offer statistical testimony in support of litigation. Formally, "Relating to or dealing with the application of forensic knowledge to legal problems." Merriam-Webster, http://www.merriam-webster.com/dictionary/forensic [viewed April 2016]. Forensic statistical experts include accountants, economists, attorneys, organizational behavior specialists, psychologists, statisticians, among others. Relatedly, Zitzewitz (2012) writes about the field of "forensic economics" expansively, equating "forensic" with "exploratory," encompassing research that empirically investigates policy, regulatory, and political issues relying methodologically on microeconomic theory.

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

adjustment is obtained by dividing the original *p*-value chosen by the number of tests being performed. The plaintiff in *Autozone* 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.

Interestingly, defendants also claimed that plaintiff's expert used an arbitrary significance level that does not conform to the requirements of *Casteneda v. Partida*. The motion was granted in part and denied in part. Yet, an expert's choice to use Bonferroni, or any other related method like Sidak or Benjamini-Hochberg to adjust error rates, is an entirely arbitrary decision.[5] In fact, the 5-percent level common in null hypothesis significance tests and embodied into case law by Castaneda is itself entirely arbitrary and a matter of historical convention and forceful personalities (Curran-Everett, 2009). While arbitrary and often customary, the willful selection of operational error-rates is hardly value-free (Pittenger, 2001; Rudner, 1953).

The stakes in error rate adjustments are high in part because the deployment of arguments for adjustment is almost always made by defense counsel, and the resulting adjustments of error inevitably favor the defense. In an employment discrimination matter, for example, the burden is on plaintiffs to demonstrate that a workplace event has had disparate impact on members of a protected class and is therefore discriminatory and illegal. This burden can be met with a favorable showing of the 80 percent rule or by tests of statistical significance. A positive outcome on the statistical tests, showing that the members of a protected class were impacted at a higher rate, is necessary to establish a rebuttable presumption.

Intent on winning, an expert retained by plaintiff may be unable to avoid the temptation of aligning her testimony with the interests of her client – and deploy a barrage of tests with the expectation of obtaining at least one significant result (Solow, 2014). Culling and testing the data, per se, is not intrinsically flawed, and many practitioners view it as unavoidable and desirable, an important step in inductive reasoning.[6] Yet, it is precisely this multiple testing that raises the chances of a favorable results emerging by chance– the methodological pitfall behind p-hacking accusations. Notwithstanding methodological soundness, the presence of a statistically significant outcome seemingly underscoring the cause of action tends to easily be impugned as biased by opposing counsel.

Blindly embracing error rate corrections is a two-edged sword that can jeopardize the credibility of forensic statistical experts generally and individually. Contrary to the testing in medical fields where there may exist an asymmetry in the costs of errors, forensic experts must recognize that favoring a Type-I error-rate adjustment like Bonferroni unduly penalizes the ability of plaintiffs to seek recourse.

True, the moral hazard confronting plaintiff's expert remains. Transparency, full disclosure, and clarity are recommended; in instances where it may be appropriate to resort to a Type-I error adjustment – the expert should carefully explain the reasoning behind the decision to adjust error-rates, the prospective tradeoffs incurred and the impact on the conclusions. A statistical power analysis is also important. A power analysis is a statistical procedure used to detect a meaningful difference assuming that plaintiff's claim is true (Gastwirth & Bura, 2011). This power analysis should routinely accompany each test utilized in the analysis. These additional elements, associated with clear and cogent

[5] Sidak and Benjamini-Hochberg are alternative procedures that retain the broader objective of the Bonferroni procedure of reducing the error-rate without sacrificing power (MacDonald, 2014).

[6] But see, Denton (1985).

explanations, would enable the Court to properly integrate a finding of significance or non-significance when the situation calls for one.

CONTEXT

It is difficult to disentangle the calls for the error-rate adjustments that concern us here from the recent complaints against the blind reliance on null hypothesis significance testing (NHST), as they are related. In raising this issue, commentators are conflating their apprehension with calls for greater statistical fidelity in the reporting of results across numerous fields. These concerns have raised *inter alia*, calls for programs of formal replications of research results, for pre-publication of anticipated hypothesis and models, and even for abandonment of conventional hypothesis testing. The Journal of Basic and Applied Social Psychology, for instance, has banned the use of Null Hypothesis Statistical Tests and related statistical procedures (Trafimow & Marks, 2015). Similarly, the American Statistical Association has recently released a statement against *p*-values (Wasserstein & Lazar, 2016). Indeed, even in biomedical research, where a need for error-adjustments for multiple testing is generally accepted, there are several eloquent critics (Feise, 2002; Perneger, 1998; Rothman,1990).

However, setting aside the objections and considering that costs of potential false positives are considerable in the medico-biological fields, there is perhaps good reason to enhance the burden of proof and consider adjusting conventional error rates. This is not the case for forensic statistical experts where there is no clear and marked difference between the costs of both false positives and false negatives in conventional civil litigation. Forensic statistical experts should not ignore proper statistical procedure; however, it should be acknowledged that the practice of forensics is different from other scientific fields in important ways. While concerns remain about type I error, one cannot disregard errors of the second type. Neither should decision-making be limited solely to sampling error – as happens when the focus is solely on *p*-values. Further, it is critical to recognize that for legal disputes where objectivity and impartiality are key, the manner and method by which forensic experts analyze and present conclusions to the court is distinct, relevant, and important.

STATISTICAL FORENSIC TESTIMONY AND MORAL HAZARD

There is considerable subjectivity in the assembling of a statistical report in a lawsuit that enables an expert to present a client in the best statistical light possible. For instance, in employment discrimination, there is disagreement as to who among the workforce should be considered the baseline against which the reduction in force is compared, *e.g.*, should the comparison set be those who applied for a job, or a larger set which also includes those who may have applied but for the knowledge of a discriminatory employer? There is disagreement about the unit of analysis, *e.g.,* should the examination occur at the store level, a regional level, or company-wide? There is disagreement on the possible job categories; *e.g.,* are "senior" managers in the sales group a classification distinct from "departmental" managers in the operations group?

The opportunity for manipulation is considerable. Specious reasoning can be found both at the broad methodological level and at discipline specific instances. As an illustration of the latter, consider Simmons *et al* (2011) detailing several ways in which scholars are able to selectively collect and analyze data so as to dramatically inflate the odds of obtaining a statistically significant result. At the field level, Solow (2006) lists numerous ways in which dubious testimony emerges in cases: "…ignoring facts that are inconvenient to the desired conclusion; making unsupportable assumptions, manipulating data or statistical results to support the desired result; and reaching conclusions based on unsupported or unsupportable theories." DeMartino (2013) documents, "the widespread expectation to supply what might be called pseudo-research to sustain decisions already taken." And more specifically, DeMartino reports that "several economic consultants identified the market pressure that sometimes push the economist in the direction of providing the client with the result that best serves its interest rather than that which is best supported by the evidence." He also notes that "pressure to sustain the position of the client has been particularly intensive in civil litigation where the parties often contest vast sums."[7]

And as an example of methodological obfuscation, consider *Carpenter v. Boeing,* an employment discrimination case. In *Carpenter,* defense's forensic expert successfully impugned plaintiff's statistical analysis alleging that variables were missing from the study. Boeing's claim was that variables other than those controlled for in plaintiff's expert report *could* be responsible for the observed disparities. The district court agreed with Boeing that a statistical study could not establish a claim without considering such variables and granted Boeing's motion for summary judgment on that basis. The 10th Circuit Court upheld.

How could tendentious arguments enter the determination of the proper error rate? To illustrate we resort to an example offered by Tabak to illustrate this point: "Consider an expert examining the share of women being laid off at a company undergoing a reduction in force" (Tabak, 2006). The company is one with a nation-wide presence defending a discrimination lawsuit. Suppose the null hypothesis of no discrimination is true. The company did not discriminate between men and women when instituting the reduction-in-force. The realized data is tested to establish whether there is any difference in treatment based on gender. If this test is repeated 20 times – drawing a different separable component every time (e.g. gender by individual region, gender by store, by department, by age, by level of job seniority, by job category, etc.), there is a 64 percent chance of obtaining at least one "significant" result.[8] These outcomes – a false positive - are entirely by chance because it is known – by design - that there is no discrimination. Yet, the finding of significance suffices to establish a rebuttable presumption of discrimination. The claim would proceed with an significant likelihood of prevailing. The prospective cost in this instance is the cost of settling the matter, or of a judgment favoring the plaintiff, in addition to the litigation costs. Returning to the example, consider the alternative: suppose the null hypothesis is not true. That is to say, assume that there is discrimination present. Then, an adjustment to the Type I error rate necessarily leads to an increase in Type II errors. And a conventional statistical test would be unlikely to establish that there is discrimination present when examining the proffered data. The cost in this instance is justice denied.

[7] The entire issue of the *Journal of Forensic Economics* (Volume 24, Issue 1, 2013) is devoted to forensic ethics. Interestingly, the matter raised here, the issue of multiple comparisons, is not mentioned.

[8] When α is the observed statistical threshold, if follows that the probability of observing at least one significant result is obtained from the following property of binomials:

Prob(at least one significant result)

= 1 - Prob(no significant result)

= 1 - Prob$(1- \alpha)^{20}$

= $1 – (1-0.05)^{20}$

= 0.6415;

The example is from Follett & Welch (1983); Barnes (1984) (The probabilistic result means simply that as the number of separable component parts of an examination or interview or work force increases, the probability of finding a part that violates the two standard deviation test for disparate impact increases even if the test or interive or employment practice is neutral.)

Reputational concerns and the adversarial nature of the American legal process offer some protections against obvert abuse. There are considerable costs to impugning an expert's reputation by taking ill-considered actions. These costs have to be weighed against the benefits of future business, which is sure to follow a winning record for any forensic expert. The legal process itself ensures that defendant will have an expert of their own; one capable of challenging any tendentious practices or methods plaintiff's expert may have used.
And then there is Daubert.[9] *Daubert v. Merrill-Dow Pharmaceuticals* set forth the present-day standard guiding the admission of expert testimony into legal proceedings.[10] Among the factors set forth in Daubert to limit or exclude absurd, unconventional, irrational, subjective, or pseudoscientific assertions is one that states "Additionally, in the case of a particular scientific technique, the court ordinarily should consider the known or potential rate of error." (Daubert, 1993)

DAUBERT WARNS OF ERROR RATES, NOT MERELY TYPE I ERRORS

The gatekeeping role bestowed by Daubert is enhanced if the trier-of-fact is provided with the likely consequences of adopting error rate adjustments. Specifically, an estimation of the likelihood of Type II error and the associated power of the test should accompany any demands for error-rate adjustments. An example provided below demonstrates the considerable advantages a statement specifying the power of the statistical test used offers in clarifying the issue with modest incremental demands asked of a statistical expert.

There are two possible outcomes from a statistical finding providing no support for an inference of discrimination. Either the defendant was not culpable—or they were culpable, and the statistical test utilized was not capable of detecting the realized disparity between group selection rates. The latter is known as a Type-II error – a false negative. It is an error because the test deployed was incapable of rejecting the null hypothesis of no discrimination. This error results from what is understood to be the low power of the utilized test. Although real and potentially damaging to one's case, the power tradeoff is often unobserved Colquhoun (2006). Consider the following illustration: an instance of a matter in which you are conducting 1000 tests. Suppose the first 900 tests are random numbers from a standard normal distribution. The last 100 tests are random numbers from a normal distribution with a mean of three and a standard deviation of one.

The hypothesis that the value of x is not different from 0 is tested. The alternative hypothesis is a one-sided test stating the value is larger than 0. The first 900 tests should fail to reject the null; any difference between the observed value and 0 is due to chance. The last 100 tests should reject the null: for the difference between the null and the actual realization of the data – of 3 – is not due to chance alone (in fact, it was deliberately designed that way). The value chosen to represent the actual outcome, 3, is arbitrary in this instance; but it is typically presented as a range. The result is to show that the power of the test used in the examination of discrimination is itself a range. In fact, providing a visual display of the sensitivity of the power of the test to the assumed counterfactual, enhances the impression provided.

These tests are repeated 1000 times and the results examined; the mean of the 1000 repetitions is determined to establish the false positive and the false negative rates.[11] In Table

[9] And, of course, its various state incantations.
[10] 509 U.S. 579 (1993).
[11] The R script is available upon request.

1 (Appendix) the Type-I error rate (false positives) is 0.041, close to the expected value of 0.05. The type-II error rate (false negatives) is 0.168.

Now consider adjusting the false positive error rate to increase the likelihood that the null is not unintentionally rejected. The Bonferroni adjustment – where the new false positive rate is a fraction of the initial one - accomplishes this. The Bonferroni adjustment is equal to α/1000 or approximately 0.00005. The false positive rate ("α") is divided by the number of tests: in this case 1000 tests.

In Table 2 (Appendix) the Type-I error rate has been reduced to 0.000023. However, the Type-II error rate has increased considerably to 0.896; the Type-I error rate was reduced at the expense of Type-II errors. In effect, the power of this test has collapsed. Power refers to the test's ability to correctly reject the null hypothesis when the alternative is true – as in table 2. The power of the test drops from approximately 90 percent to 10 percent.
Thus, this Bonferroni adjusted test is – for all practical purposes - incapable of detecting the presence of truly aggrieved plaintiffs. A consequence of Type-II error inflation in the appraisal is that legitimate plaintiffs are unfairly dismissed and thereby unable to obtain redress. This pitfall, of course, is especially vexing in small samples, which are generally characterized by a low ability to avoid Type-II errors. Put differently, tests on small samples are beset by low power (Colquhoun, 2006).

When choosing a threshold value, whether 5 percent, Bonferroni adjusted, or any other value for that matter, it is important to select one based on the best balance among the error rates. Daubert is silent on this decision, and for good reason. The balance between metrics cannot be established *ex ante* – and must be context driven: medical research is not comparable to statistical forensic testimony in a Title VII lawsuit. If there is little downside to concluding that the data is not consistent with discrimination, then a low false positive rate may be an adequate trade-off for a high false positive rate. For instance, if examining the accuracy of a trial medication, one may want to avoid spurious false positives. On the other hand, if examining the likelihood of default for loan applications, one may want the false positives to be a bit higher.

The value of forensic economic testimony lies in its impartiality. Calls for multiple adjustments may be necessary. If this is the case, then a disclosure of the power of the test should accompany any proffered outcomes. More specifically, a power of the test keyed on various plausible counterfactuals designed to accommodate legitimate instances of discrimination.

CONCLUDING COMMENTS

The adjustment of error rates is not in itself concerning in those settings where the correction of error is likely to be both methodologically and topically appropriate. In many sets of data bearing on medicine, such adjustments can be well-nigh essential. However, the adjustment of error rates in the adjudication of presentations of data from experts in statistical forensics amounts to an intractable and important value judgment pretextually cast in the language of statistical fidelity.

Opting for an adjustment advocates the importance and relevance of Type-I error-rates over Type-II error rates. Or it may simply hide the contribution of the dueling forensic experts behind the "impenetrable wall of esoteric knowledge" in a manner not susceptible to cross-examination (Posner, 1999); after all, the statistical properties of error-rate adjustments are well documented in the literature.

The discussion and debate about how to compute and give a courtroom valence to error rate adjustments in forensic economics is a matter for the profession to discuss and resolve in the peer-reviewed literature, not in the courtroom, and the correct answer in the matter is not one jurists are either prepared to make or reverse with any appropriate rigor. However, the primary concern is normative: one of ethics and justice, because the defense is inevitably the beneficiary of any evaluation of forensic economic data if error rate adjustment is requested and debated.

The discussed *Autozone* case is a relatively recent case in which the matter of error-rate adjustment figured prominently in the battle of experts. Professor Joseph Gastwirth published an insightful comment on this particular case (Gastwirth, 2008). In his commentary, Professor Gastwirth appears to celebrate the judge's decision to dismiss the matter on the basis of the multiple testing criticism raised by the defense. But a close reading of the decision reveals a more sobering explanation than the one offered by Professor Gastwirth. The Court effectively punted in frustration, unable to draw any conclusive assistance from either expert. The Courts stated as follows:

> "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 unreliable. Therefore, the court will not grant summary judgment on the EEOC's pattern or practice claims on this basis." *(EEOC v. Autozone Inc., 2006)*

The trial court judge is correct in noting the confusion in the matter, an inescapable outcome that emerges because the question of whether to adjust the error rates cannot be settled a priori.

Erring on the side of practicality and transparency, statistical forensics experts should refuse not only to take sides but should abjure the obfuscation that characterizes exchanges between dueling experts. They should strive to set forth their analysis and methodology and establish 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 the proffered outcomes as a result of applying and not applying error rate adjustments should convey a more thorough impression of the likelihood of discrimination and the validity of the statistical study.

Experts should report p-values, confidence intervals, the power of the tests, and opinions as to the need for adjustments and their consequences. The expert should explain the tradeoff entailed in tweaking the Type-I error rate by providing estimates of the increases in Type-II errors. Simply put, in seeking increasingly stringent errors rates to accommodate the increased likelihood of false positives, the likelihood of false negatives increases. That is to say, it increases the chances of failing to detect instances where the null is rejected. This means that there is an increased chance that a deserving plaintiff will fail to obtain relief as a result of statistical artifact.

These are the elements of expert's own professional 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 they exclude someone from their day in court.

Ethics demands and justice requires that expert witnesses who bring statistical data to the courtroom be accountable for the reality that error rate adjustments introduce inevitable, scientifically demonstrable bias and that such adjustments rely more upon the confusion and fatigue of jurists and juries than upon valid presentations of scientifically rigorous expertise.

REFERENCES

Barnes, D. W. (1984). The Problem of Multiple Components in Title VII Litigation: A Comment. *Law and Contemporary Problems, 46*(4), pp. 185-188.

Carey, B. (2015, August 25). Many Psychology Findings Not As Strong as Claimed, Study Says. *The New York Times*.

Cohen, D. (1962). The Statistical Power of Abnormal Social Science Research - a Review. *Journal of Abnormal Social Psychology, 62*, pp. 145-153.

Cohen, J. (1994, December). The Earth is Round. *American Psychologist, 12*, pp. 997-1003.

Colquhoun, D. (2014). An Investigation of the False Discovery Rate and the Misinterpretation of p-Values. *Royal Society Open Science, 1*. doi:dx.doi.org/10.1098/rsos.140216

Curran-Everett, D. (2009, June). Explorations in statistics: hypothesis tests and P values. *Advances in Psychology Education, 33*(2), 81-86.

Daubert v. Merrell Dow Pharmaceutical, 509 (U.S. 1993).

Demartino, G. (2013). Professional Economic Ethics: The Posnerian and Naive Perspectives. *Journal of Forensic Economics, 24*(1), 3-18.

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

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

Faigman, D. L. (1999). *Legal Alchemy: The Use and Misuse of Science in the Law*.

Faigman, D. L. (2000). Lecture: The Use and Misuse of Science in the Law. *Yale Journal of Law and Technology, 2*. Retrieved April 18, 2016, from http://digitalcommons.law.yale.edu/yjolt/vol2/iss1/3

Federal Rules of Evidence. Rule 102.

Federal Rules of Evidence. Rule 702.

Feise, R. J. (2002, June 17). Do Multiple Outcome Measures Require p-value Adjustment? *BMC Medical Research Methodology, 2*(8), 1-4.

Finnerty, J. (2009). A Closer Look at Correcting for False Discovery Bias When Making Multiple Comparisons. *Journal of Forensic Economics, 21*(1), 55-62.

Follett, R., & Welch, F. (1983). Testing for Discrimination in Employment Practice. *Law & Contemporary Problems, 46*(4), 171-184.

Freedman, D. H. (2010, November). Lies, Damned Lies, and Medical Sciences. *The Atlantic*. Retrieved from http://www.theatlantic.com/magazine/archive/2010/11/lies-damned-lies-and-medical-science/308269/

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

Gastwirth, J. L., & Bura, E. (2011). Some Important Statistical Issues Courts Should Consider in their Assessment of Statistical Analysis Submitted in Class Certification Motions: Implications for Dukes v. Wal-mart. *Law, Probability and Risk, 10*, pp. 225-263.

Hersch, J., & Bullock, B. D. (2014). The Use and Misuse of Econometric Evidence in Employment Discrimination Cases. *Washington and Lee Law Review, 71*(4), pp. 2365-2429.

Holm, S. (1979). A Simple Sequentially Rejective Test Procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Hotz, R. L. (2007, September 14). Most Science Studies Appear to Be Tainted by Sloppy Analysis. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/SB118972683557627104

Ioannidis, J. A. (2005, August). Why Most Published Research Findings Are False. *PLoS Medicine, 2*(8), pp. 0696-0701.

Jeffreys, H. (2003). *Theory of Probability* (3rd ed.).

Kaye, D. H. (1983). Statistical Significance and the Burden of Persuasion. *Law & Contemporary Problems, 46*, pp. 13-23.

MacDonald, J. H. (2014). Multiple Comparisons. In J. H. MacDonald, & J. H. MacDonald (Ed.), *Handbook of Biological Statistics* (pp. 254-260). Baltimore, Maryland: Sparky House Publishing. Retrieved from Handbook of Biological Statistics: http://www.biostathandbook.com/multiplecomparisons.html

Matrixx Initiatives, Inc. et al v. Siracusano, 563 (U.S. 2011).

Nuzzo, R. (2014, February 13). Statistical Errors. *Nature, 506*, pp. 150-152.

Open Sciences Collaboration. (2015, August 28). Estimating the Reproducibility of Psychological Science. *Science, 349*(6251), 943-951.

Perneger, T. J. (1998). What's Wrong With Bonferroni Adjustments. *BMJ*(316), 1236-8.

Peterson, D. W. (2005). On Forensic Decision Analysis. *Journal of Forensice Economics, 18*, pp. 11-62.

Pittenger, D. J. (2001). Hypothesis Testing as a Moral Choice. *Ethics & Behavior, 11*(2), pp. 151-162.

Posner, R. A. (1999). Law and Economics of the Expert Witness. *Journal of Economic Perspectives, 13*(2).

Rodriguez, A. E. (2009). A Closer Look at Correction for False Discovery Bias When Making Multiple Comparisons. *Journal of Forensic Economics, 21*(1), 99-103.

Rothman, K. J. (1990, January). No Adjustments Are Needed for Mulitple Comparisons. *Epidemiology*, 43-46.

Rudner, R. (1953). The Scienctist Qua Scientist Makes Value Judgments. *Philosophy of Science*, pp. 1-6.

Rule 702: Testimony by an Expert Witness. (n.d.). *Federal Rules of Evidence*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011, November). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Presenting Anything as Significant. *Psychological Science*, 1359-1366. Retrieved from http://www.researchgate.net/publication/51725721_False-Positive_Psychology_Undisclosed_Flexibility_in_Data_Collection_and_Analysis_Allows_Presenting_Anything_as_Significant?enrichId=rgreq-08f6db2e-afcc-4600-94f9-64aa26e38025&enrichSource=Y292ZXJQYWdlOzUxN

Simpson, M. S. (2005). Using Forensic Decision Analysis to Investigate Claims of Pay Discrimination. *Journal of Forensic Economics, 18*(1), pp. 63-81.

Solow , J. L. (2006, Winter). Doing Good Economics in the Courtroom: Thoughts on Daubert and Expert Testimony in Antitrust. *The Journal of Corporation Law*, 489-502.

Stanley, T. D. (2001, Summer). Wheat from Chaff: Meta-Analysis as Quantitative Literature Review. *Journal of Economic Perspectives, 15*(3), 131-150.

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

The Economist. (2013, October 19). Trouble at the Lab. Retrieved from http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(1), 1-2.

Wasserstein, R. L., & Lazar, N. (2016, March 7). The ASA's Statement on p-values: context, process, and purpose. *The American Statistician*, 129-133. Retrieved from http://dx.doi.org/10.1080/00031305.2016.1154108

Ziliak, S. T., & McCloskey, D. N. (2009). *The Cult of Statistical Significance.* Ann Arbor: The University of Michigan Press.

Zitzewitz, E. (2012). Forensic Economics. *Journal of Economics Literature, 50*(3), 731-769.

APPENDIX

Table 1

No Adjustment		
	Null is True	**Null is False**
False Positive	0.041	0.830
False Negative	0.958	0.168

Table 2

Bonferroni Adjusted		
	Null is True	**Null is False**
False Positive	0.000023	0.104
False Negative	0.99	0.896