EXHIBIT A

Reference Manual on Scientific Evidence

Second Edition

Federal Judicial Center 2000

An electronic version of the *Reference Manual* can be downloaded from the Federal Judicial Center's site on the World Wide Web. Go to

http://air.fjc.gov/public/fjcweb.nsf/pages/16

For the Center's overall homepage on the Web, go to

http://www.fjc.gov

Rights Act, if the proposed examination excludes a disproportionate number of women, the employer needs to show that the exam is job related.[110]

To see whether there is disparate impact, the employer administers the exam to a sample of 50 men and 50 women drawn at random from the population of job applicants. In the sample, 29 of the men but only 19 of the women pass; the sample pass rates are therefore 29/50 = 58% and 19/50 = 38%. The employer announces that it will use the exam anyway, and several applicants bring an action under Title VII. Disparate impact seems clear. The difference in sample pass rates is 20 percentage points: 58% − 38% = 20 percentage points. The employer argues, however, that the disparity could just reflect random error. After all, only a small number of people took the test, and the sample could have included disproportionate numbers of high-scoring men and low-scoring women. Clearly, even if there were no overall difference in pass rates for male and female applicants, in some samples the men will outscore the women. More generally, a sample is unlikely to be a perfect microcosm of the population; statisticians call differences between the sample and the population, just due to the luck of the draw in choosing the sample, "random error" or "sampling error."

When assessing the impact of random error, a statistician might consider the following topics:

- *Estimation.* Plaintiffs use the difference of 20 percentage points between the sample men and women to estimate the disparity between all male and female applicants. How good is this estimate? Precision can be expressed using the "standard error" or a "confidence interval."
- *Statistical significance.* Suppose the defendant is right, and there is no disparate impact: in the population of all 5,000 male and 5,000 female applicants, pass rates are equal. How likely is it that a random sample of 50 men and 50 women will produce a disparity of 20 percentage points or more? This chance is known as a *p*-value. Statistical significance is determined by reference to the *p*-value, and "hypothesis testing" is the technique for computing *p*-values or determining statistical significance.[111]
- *Posterior probability.* Given the observed disparity of 20 percentage points in the sample, what is the probability that—in the population as a whole—men and women have equal pass rates? This question is of direct interest to the courts. For a subjectivist statistician, posterior probabilities may be com-

110. The seminal case is *Griggs v. Duke Power Co.*, 401 U.S. 424, 431 (1971). The requirements and procedures for the validation of tests can go beyond a simple showing of job relatedness. *See, e.g.*, Richard R. Reilly, *Validating Employee Selection Procedures, in* Statistical Methods in Discrimination Litigation, *supra* note 11, at 133; Michael Rothschild & Gregory J. Werden, *Title VII and the Use of Employment Tests: An Illustration of the Limits of the Judicial Process*, 11 J. Legal Stud. 261 (1982).

111. "Hypothesis testing" is also called "significance testing." For details on the example, see *infra* Appendix, especially note 245.

the .05 level. If the threshold were set lower, say at .01, the result would not be significant.[137]

In practice, statistical analysts often use certain preset significance levels—typically .05 or .01.[138] The .05 level is the most common in social science, and an analyst who speaks of "significant" results without specifying the threshold probably is using this figure.[139] An unexplained reference to "highly significant" results probably means that p is less than .01.[140]

Since the term "significant" is merely a label for certain kinds of p-values, it is subject to the same limitations as are p-values themselves. Analysts may refer to a difference as "significant," meaning only that the p-value is below some threshold value. Significance depends not only on the magnitude of the effect, but also on the sample size (among other things). Thus, significant differences are evidence that something besides random error is at work, but they are not evidence that this "something" is legally or practically important. Statisticians distinguish between "statistical" and "practical" significance to make the point. When practical significance is lacking—when the size of a disparity or correlation is negligible—there is no reason to worry about statistical significance.[141]

As noted above, it is easy to mistake the p-value for the probability that there is no difference. Likewise, if results are significant at the .05 level, it is tempting to conclude that the null hypothesis has only a 5% chance of being correct.[142]

137. For another example of the relationship between a test statistic and significance, see *infra* § V.D.2.

138. The Supreme Court implicitly referred to this practice in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977), and *Hazelwood School District v. United States*, 433 U.S. 299, 311 n.17 (1977). In these footnotes, the Court described the null hypothesis as "suspect to a social scientist" when a statistic from "large samples" falls more than "two or three standard deviations" from its expected value under the null hypothesis. Although the Court did not say so, these differences produce p-values of about .05 and .01 when the statistic is normally distributed. The Court's "standard deviation" is our "standard error."

139. Some have suggested that data not "significant" at the .05 level should be disregarded. *E.g.*, Paul Meier et al., *What Happened in* Hazelwood: *Statistics, Employment Discrimination, and the 80% Rule*, 1984 Am. B. Found. Res. J. 139, 152, *reprinted in* Statistics and the Law, *supra* note 1, at 1, 13. This view is challenged in, e.g., Kaye, *supra* note 118, at 1344 & n.56, 1345.

140. Merely labeling results as "significant" or "not significant" without providing the underlying information that goes into this conclusion is of limited value. *See, e.g.*, John C. Bailar III & Frederick Mosteller, *Guidelines for Statistical Reporting in Articles for Medical Journals: Amplifications and Explanations*, *in* Medical Uses of Statistics, *supra* note 28, at 313, 316.

141. *E.g.*, Waisome v. Port Auth., 948 F.2d 1370, 1376 (2d Cir. 1991) ("though the disparity was found to be statistically significant, it was of limited magnitude"); *cf.* Thornburg v. Gingles, 478 U.S. 30, 53–54 (1986) (repeating the district court's explanation of why "the correlation between the race of the voter and the voter's choice of certain candidates was [not only] statistically significant," but also "so marked as to be substantively significant, in the sense that the results of the individual election would have been different depending upon whether it had been held among only the white voters or only the black voters").

142. *E.g.*, *Waisome*, 948 F.2d at 1376 ("Social scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for a deviation could be random . . . ."); Rivera v. City of Wichita Falls, 665 F.2d 531, 545 n.22 (5th Cir. 1982) ("A variation

the size of the sample. Discerning subtle differences in the population requires large samples; even so, small samples may detect truly substantial differences.[145]

When a study with low power fails to show a significant effect, the results are more fairly described as inconclusive than as negative: the proof is weak because power is low.[146] On the other hand, when studies have a good chance of detecting a meaningful association, failure to obtain significance can be persuasive evidence that there is no effect to be found.[147]

2. *One- or Two-tailed Tests?*

In many cases, a statistical test can be done either one-tailed or two-tailed. The second method will produce a *p*-value twice as big as the first method. Since

145. For simplicity, the numerical examples of statistical inference in this reference guide presuppose large samples. Some courts have expressed uneasiness about estimates or analyses based on small samples; indeed, a few courts have refused even to consider such studies or formal statistical procedures for handling small samples. *See, e.g.*, Bunch v. Bullard, 795 F.2d 384, 395 n.12 (5th Cir. 1986) (that 12 of 15 whites and only 3 of 13 blacks passed a police promotion test created a prima facie case of disparate impact; however, "[t]he district court did not perform, nor do we attempt, the application of probability theories to a sample size as small as this" because "[a]dvanced statistical analysis may be of little help in determining the significance of such disparities"); United States v. Lansdowne Swim Club, 713 F. Supp. 785, 809–10 (E.D. Pa. 1989) (collecting cases). Other courts have been more venturesome. *E.g.*, Bazemore v. Friday, 751 F.2d 662, 673 & n.9 (4th Cir. 1984) (court of appeals applied its own *t*-test rather than the normal curve to quartile rankings in an attempt to account for a sample size of nine), *rev'd on other grounds*, 478 U.S. 385 (1986).

Analyzing data from small samples may require more stringent assumptions, but there is no fundamental difference in the meaning of confidence intervals and *p*-values. If the assumptions underlying the statistical analysis are justified—and this can be more difficult to demonstrate with small samples—then confidence intervals and test statistics are no less trustworthy than those for large samples. Aside from the problem of choosing the correct analytical technique, the concern with small samples is not that they are beyond the ken of statistical theory, but that (1) the statistical tests involving small samples might lack power, and (2) the underlying assumptions may be hard to validate.

146. In our example, with $\alpha = .05$, power to detect a difference of 10 percentage points between the male and female job applicants is only about 1/6. *See infra* Appendix. Not seeing a "significant" difference therefore provides only weak proof that the difference between men and women is smaller than 10 percentage points. We prefer estimates accompanied by standard errors to tests because the former seem to make the state of the statistical evidence clearer: The estimated difference is 20 ± 10 percentage points, indicating that a difference of 10 percentage points is quite compatible with the data.

147. Some formal procedures are available to aggregate results across studies. *See In re* Paoli R.R. Yard PCB Litig., 916 F.2d 829 (3d Cir. 1990). In principle, the power of the collective results will be greater than the power of each study. *See, e.g.*, The Handbook of Research Synthesis 226–27 (Harris Cooper & Larry V. Hedges eds., 1993); Larry V. Hedges & Ingram Olkin, Statistical Methods for Meta-Analysis (1985); Jerome P. Kassirer, *Clinical Trials and Meta-Analysis: What Do They Do for Us?*, 327 New Eng. J. Med. 273, 274 (1992) ("[C]umulative meta-analysis represents one promising approach."); National Reseach Council, Combining Information: Statistical Issues and Opportunities for Research (1992); Symposium, *Meta-Analysis of Observational Studies*, 140 Am. J. Epidemiology 771 (1994). Unfortunately, the procedures have their own limitations. *E.g.*, Diana B. Petitti, Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine (2d ed. 2000); Michael Oakes, Statistical Inference: A Commentary for the Social and Behavioural Sciences 157 (1986) ("a retrograde development"); John C. Bailar III, *The Promise and Problems of Meta-Analysis*, 337 New Eng. J. Med. 559 (1997) (editorial); Charles Mann, *Meta-Analysis in the Breech*, 249 Science 476 (1990).

small *p*-values are evidence against the null hypothesis, a one-tailed test seems to produce stronger evidence than a two-tailed test. However, this difference is largely illusory.[148]

Some courts have expressed a preference for two-tailed tests,[149] but a rigid rule is not required if *p*-values and significance levels are used as clues rather than as mechanical rules for statistical proof. One-tailed tests make it easier to reach a threshold like .05, but if .05 is not used as a magic line, then the choice between one tail and two is less important—as long as the choice and its effect on the *p*-value are made explicit.[150]

3. How Many Tests Have Been Performed?

Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield "significant" findings, even when there is no real effect. Consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce ten heads when tossed ten times is $(1/2)^{10} = 1/1{,}024$. Observing ten heads in the first ten tosses, therefore, would be strong evidence that the coin is biased. Nevertheless, if a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. The test—looking for a run of ten heads—can be repeated far too often.

148. In our pass rate example, the *p*-value of the test is approximated by a certain area under the normal curve. The one-tailed procedure uses the "tail area" under the curve to the right of 2, giving $p = .025$ (approximately). The two-tailed procedure contemplates the area to the left of -2, as well as the area to the right of 2. Now there are two tails, and $p = .05$. *See infra* Appendix (figure 13); Freedman et al., *supra* note 16, at 549–52.

According to formal statistical theory, the choice between one tail or two can sometimes be made by considering the exact form of the "alternative hypothesis." *See infra* § IV.C.5. In our example, the null hypothesis is that pass rates are equal for men and women in the whole population of applicants. The alternative hypothesis may exclude a priori the possibility that women have a higher pass rate, and hold that more men will pass than women. This asymmetric alternative suggests a one-tailed test. On the other hand, the alternative hypothesis may simply be that pass rates for men and women in the whole population are unequal. This symmetric alternative admits the possibility that women may score higher than men, and points to a two-tailed test. *See, e.g.*, Freedman et al., *supra* note 16, at 551. Some experts think that the choice between one-tailed and two-tailed tests can often be made by considering the exact form of the null and alternative hypothesis.

149. *See, e.g.*, Baldus & Cole, *supra* note 89, § 9.1, at 308 n.35a; The Evolving Role of Statistical Assessments as Evidence in the Courts, *supra* note 1, at 38–40 (citing EEOC v. Federal Reserve Bank, 698 F.2d 633 (4th Cir. 1983), *rev'd on other grounds sub nom.* Cooper v. Federal Reserve Bank, 467 U.S. 867 (1984)); Kaye, *supra* note 118, at 1358 n.113; David H. Kaye, *The Numbers Game: Statistical Inference in Discrimination Cases*, 80 Mich. L. Rev. 833 (1982) (citing Hazelwood Sch. Dist. v. United States, 433 U.S. 299 (1977)). Arguments for one-tailed tests are discussed in Finkelstein & Levin, *supra* note 1, at 125–26; Richard Goldstein, *Two Types of Statistical Errors in Employment Discrimination Cases*, 26 Jurimetrics J. 32 (1985); Kaye, *supra* at 841.

150. One-tailed tests at the .05 level are viewed as weak evidence—no weaker standard is commonly used in the technical literature.