EXHIBIT K

Reference Manual on Scientific Evidence

Second Edition

Federal Judicial Center 2000

This Federal Judicial Center publication was undertaken in furtherance of the Center's statutory mission to develop and conduct education programs for judicial branch employees. The views expressed are those of the authors and not necessarily those of the Federal Judicial Center.

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.

puted using "Bayes' rule." Within the framework of classical statistical theory, however, such a posterior probability has no meaning.[112]

- *Applicability of statistical models.* Statistical inference—whether done with confidence intervals or significance probabilities, by objective methods or subjective—depends on the validity of statistical models for the data. If the data are collected on the basis of a probability sample or a randomized experiment, there will be statistical models that fit the situation very well, and inferences based on these models will be quite secure. Otherwise, calculations are generally based on analogy: this group of people is like a random sample, that observational study is like a randomized experiment. The fit between the statistical model and the data may then require examination: how good is the analogy?

A. Estimation

1. What Estimator Should Be Used?

An estimator is a statistic computed from sample data and used to estimate a numerical characteristic of the population. For example, we used the difference in pass rates for a sample of men and women to estimate the corresponding disparity in the population of all applicants. In our sample, the pass rates were 58% and 38%; the difference in pass rates for the whole population was estimated as 20 percentage points: 58% – 38% = 20 percentage points. In more complex problems, statisticians may have to choose among several estimators. Generally, estimators that tend to make smaller errors are preferred. However, this idea can be made precise in more than one way,[113] leaving room for judgment in selecting an estimator.

2. What Is the Standard Error? The Confidence Interval?

An estimate based on a sample is likely to be off the mark, at least by a little, due to random error. The standard error gives the likely magnitude of this random error.[114] Whenever possible, an estimate should be accompanied by its standard

112. This classical framework is also called "objectivist" or "frequentist," by contrast with the "subjectivist" or "Bayesian" framework. In brief, objectivist statisticians view probabilities as objective properties of the system being studied. Subjectivists view probabilities as measuring subjective degrees of belief. Section IV.B.1 explains why posterior probabilities are excluded from the classical calculus, and section IV.C briefly discusses the subjectivist position. The procedure for computing posterior probabilities is presented *infra* Appendix. For more discussion, see David Freedman, *Some Issues in the Foundation of Statistics*, 1 Found. Sci. 19 (1995), *reprinted in* Topics in the Foundation of Statistics 19 (Bas C. van Fraasen ed., 1997).

113. Furthermore, reducing error in one context may increase error in other contexts; there may also be a trade-off between accuracy and simplicity.

114. "Standard errors" are also called "standard deviations," and courts seem to prefer the latter term, as do many authors.

error.[115] In our example, the standard error is about 10 percentage points: the estimate of 20 percentage points is likely to be off by something like 10 percentage points or so, in either direction.[116] Since the pass rates for all 5,000 men and 5,000 women are unknown, we cannot say exactly how far off the estimate is going to be, but 10 percentage points gauges the likely magnitude of the error.

Confidence intervals make the idea more precise. Statisticians who say that population differences fall within plus-or-minus 1 standard error of the sample differences will be correct about 68% of the time. To write this more compactly, we can abbreviate "standard error" as "SE." A 68% confidence interval is the range

estimate – 1 SE to estimate + 1 SE.

In our example, the 68% confidence interval goes from 10 to 30 percentage points. If a higher confidence level is wanted, the interval must be widened. The 95% confidence interval is about

estimate – 2 SE to estimate + 2 SE.

This runs from 0 to 40 percentage points.[117] Although 95% confidence intervals are used commonly, there is nothing special about 95%. For example, a 99.7% confidence interval is about

estimate – 3 SE to estimate + 3 SE.

This stretches from –10 to 50 percentage points.

The main point is that an estimate based on a sample will differ from the exact population value, due to random error; the standard error measures the likely size of the random error. If the standard error is small, the estimate probably is close to the truth. If the standard error is large, the estimate may be seriously wrong. Confidence intervals are a technical refinement, and

115. The standard error can also be used to measure reproducibility of estimates from one random sample to another. *See infra* note 237.

116. The standard error depends on the pass rates of men and women in the sample, and the size of the sample. With larger samples, chance error will be smaller, so the standard error goes down as sample size goes up. ("Sample size" is the number of subjects in the sample.) The Appendix gives the formula for computing the standard error of a difference in rates based on random samples. Generally, the formula for the standard error must take into account the method used to draw the sample and the nature of the estimator. Statistical expertise is needed to choose the right formula.

117. Confidence levels are usually read off the normal curve (*see infra* Appendix). Technically, the area under the normal curve between –2 and +2 is closer to 95.4% than 95.0%; thus, statisticians often use ±1.96 SEs for a 95% confidence interval. However, the normal curve only gives an approximation to the relevant chances, and the error in that approximation will often be larger than the difference between 95.4% and 95.0%. For simplicity, we use ±2 SEs for 95% confidence. Likewise, we use ±1 SE for 68% confidence, although the area under the curve between –1 and +1 is closer to 68.3%. The normal curve gives good approximations when the sample size is reasonably large; for small samples, other techniques should be used.