FEDERAL JUDICIAL CENTER STATISTICAL EXAMPLES SOFTWARE PROTOTYPE: AGE DISCRIMINATION EXAMPLE

Robert Timothy Reagan*

ABSTRACT: The Federal Judicial Center prepared prototype materials for a software product designed to teach statistics to judges with legal examples. The prototype presents Fisher's exact, chi-squared, and bivariate and multivariate logistic regression analyses in a hypothetical example of age discrimination litigation. The prototype also discusses odds ratios, expected values, and *p*-values. Eleven distinguished experts provide comments and reviews.

CITATION: Robert Timothy Reagan, Federal Judicial Center Statistical Examples Software Prototype: Age Discrimination Example, 42 Jurimetrics J. 281–295 (2002).

As part of the Federal Judicial Center's efforts to provide federal judges with reference materials on technical matters,[1] we have begun to develop a software product that teaches judges about statistical analysis using legal examples. What follows is a prototype based on a hypothetical age discrimination case. We present this example here, along with invited comments, to highlight challenges in teaching judges about statistics.

This prototype is not software, but rather, presentation text for inclusion in a software product. The example is presented in seven parts with various subparts. The sections are organized as "screens" that are presented here to show how the

* Robert Timothy Reagan is Senior Research Associate, The Federal Judicial Center, Thurgood Marshall Federal Judiciary Building, One Columbus Circle NE, Washington, DC 20002-8003. This example was prepared with considerable advice and suggestions from Joseph Gastwirth, Marc Rosenblum, and Joe Cecil, who bear no responsibility for its flaws. This work was supported in part by a grant from the Carnegie Corporation of New York. It is a work in progress and does not necessarily represent the views of the Federal Judicial Center.

1. Established in 1967 by 28 U.S.C. § 620 (2000), the Federal Judicial Center is the federal judiciary's research and education support agency.

information might be organized on the web.[2] Following screen 4 are three subparts—screens 4.1 through 4.3. Screen 4.1, in turn, is followed by two subparts—screens 4.1.1 and 4.1.2. The remaining parts and subparts are presented in the same fashion.

The prototype is designed to be part of an interconnected product that uses hyperlinks to help users navigate through instructional sections according to their own needs. For example, each figure would include a hyperlink to material explaining in detail how to read the figure; that detail, however, is not presented here. Also, technical terms, such as "natural logarithm" or "independent variable," would have hyperlinks for further explanation.[3]

The example that follows is purely hypothetical. It is meant to convey how an age discrimination case *might* be litigated, but not how an age discrimination case *should* be litigated.

Screen 1. Brief Description of Example

This example demonstrates both how statistical evidence can support a plaintiff's prima facie case of employment discrimination and how additional statistical evidence can support the defendant's nondiscriminatory explanation for the apparently discriminatory pattern.

The example arises under the federal Age Discrimination in Employment Act, 29 U.S.C. §§ 621-634 (1994 & Supp. V 1999) (ADEA). The ADEA creates a private cause of action based on disparate treatment. Whether the Act also provides a cause of action for disparate impact is an open question.[4] Although the data in this example do not support a claim of disparate treatment, they do support a claim of disparate impact.

Screen 2. Plaintiffs' Facts

Plaintiffs filed suit in federal court alleging that their former employer— Premium Investments—terminated their employment in violation of the Age Discrimination in Employment Act. The plaintiffs were terminated after a corporate reorganization following a merger. All were over 40 years old at the time of their terminations.

Plaintiffs presented a statistical report containing data on the termination of employees during the reorganization. Plaintiffs produced evidence that 10.1% of the employees under 40 years of age were terminated (23 out of 228), while 22.3% of the employees 40 years of age and older (56 out of 251) were terminated. Statistical analysis shows that a disparity as large as this (or larger) would have a prob-

2. For this reason, the presentation departs from usual law journal style of citing authorities in footnotes. Footnotes in screen text are notes to journal readers rather than parts of the prototype.

3. The product also could include hyperlinks to other products, such as the REFERENCE MANUAL ON SCIENTIFIC EVIDENCE (Federal Judicial Center ed., 2d ed. 1999).

4. The Supreme Court heard oral argument on this question on March 20, 2002, but dismissed the writ of certiorari as improvidently granted on April 1, 2002. *See* Adams v. Florida Power Corp., 255 F.3d 1322 (11th Cir.), *cert. granted*, 122 S. Ct. 643 (2001), *cert. dismissed*, 122 S. Ct. 1290 (2002).

ability of occurring by chance alone approximately equal to three one-hundredths of one percent.

Screen 3. The Law

The Age Discrimination in Employment Act forbids employment discrimination on the basis of age in hiring, discharging, or setting compensation, terms, conditions, or privileges of employment. 29 U.S.C. § 623(a)(1). The ADEA's protection extends only to those individuals who are 40 years of age or older. *Id.* § 631(a).

Although the Supreme Court has never expressly decided whether the burden-shifting method of proof developed under Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e *et seq.*, applies to claims under the ADEA, the Court has assumed that such a method does apply. *Reeves v. Sanderson Plumbing Prods.*, 530 U.S. 133, 142 (2000). With this method, if a plaintiff establishes prima facie evidence of discrimination, the defendant has the burden of producing rebuttal evidence of a legitimate, nondiscriminatory reason for the employment action against the plaintiff. *E.g.*, *O'Connor v. Consolidated Coin Caterers Corp.*, 517 U.S. 308, 310 (1996). If the defendant is able to produce such rebuttal evidence, then the plaintiff has the burden of proving that the employment action adverse to plaintiff nevertheless was because of plaintiff's age. *E.g.*, *St. Mary's Honor Center v. Hicks*, 509 U.S. 502, 507-08 (1993).

Screen 4. Statistical Analysis Supporting Plaintiffs' Prima Facie Case

Plaintiffs showed that the proportion of employees 40 years of age and older who were terminated was greater than the proportion of employees under 40 who were terminated. Plaintiffs showed that this result would be unlikely to result from chance by performing *Fisher's exact test* on the data. For purposes of illustration, however, other possible analyses of plaintiffs' data also are presented.[5]

Screen 4.1. The Data

At the time of plaintiffs' termination, defendant's workforce was 479 employees, of whom 251 (52.4%) were 40 years of age or older. After the merger, Premium terminated 79 of its employees, of whom 56 (70.9%) were 40 years of age or older.

Another way to describe these data is to observe that of the 228 employees under 40 years of age, 23 (10.1%) were terminated, but of the 251 employees 40 years of age or older, 56 (22.3%) were terminated.

5. I am grateful to Joseph Gastwirth for emphasizing the importance of using more than one method of analysis.

Screen 4.1.1. A Two-by-Two Table

The data can be displayed in a *two-way classification table* in which the data are classified according to (i) whether the employee was terminated or retained, and (ii) whether the employee was 40 years of age or older, or not. Because the two classifications in this example have two classes, or categories, each, the table would often be called a *two-by-two table*. The data are displayed in Table 1.

Table 1. Defendant's Terminations by Age Group

	Under 40 Years Old	40 Years Old Or Older	Total
Terminated	23	56	79
% of Age Group	(10.1%)	(22.3%)	(16.5%)
Expected Number	(37.6)	(41.4)	
Retained	205	195	400
% of Age Group	(89.9%)	(77.7%)	(83.5%)
Expected Number	(190.4)	(209.6)	
Total	228	251	479
% of Age Group	(100.0%)	(100.0%)	(100.0%)

Note. This two-way classification has two rows representing employment outcome—terminated and retained—and two columns representing age group—under 40 years old and 40 years old or older. For that reason, it is called a "two-by-two table." There is also a third row and a third column for totals. The spaces where the rows and columns intersect are called *cells*, and they contain the data. For example, the table states in the upper left-hand cell that 23 of the employees under 40 years old were terminated. The numbers in parentheses underneath, in this table, are percent of age group in that cell (a *column percentage*) and the "expected number" of observations for that cell. For example, in the upper left-hand cell, the 23 persons under 40 years old who were terminated were 10.1% of the employees under 40 years old. Because 16.5% of all the employees were terminated, if age were not a factor in terminations we would expect 16.5% of the employees under 40 years old to have been terminated and that would be 37.6 employees.

Screen 4.1.2. The Odds Ratio

One useful way to summarize the data in a two-by-two table is with an *odds ratio*. The odds on being terminated for employees under 40 years old are equal to the number of employees under 40 years old who were terminated divided by the number who were retained: 23/205 = 0.112. Analogously, the odds on being terminated for employees 40 years of age or older are 56/195 = 0.287. It is clear that the odds on being terminated are considerably greater for the older employees than for the younger employees. The odds ratio is 0.287/0.112 = 2.56.

The odds ratio is useful, because its value is the same regardless of whether the rows' odds are divided by each other or the columns' odds are divided by each

other. The columns' odds ratio for being terminated comparing older and younger employees is equal to the odds on being terminated for older employees divided by the odds on being terminated for younger employees, which has already been computed to be:

$$\text{Odds Ratio} = \frac{56 / 195}{23 / 205} = 2.56$$

The odds ratio for being older comparing terminated and retained employees is equal to the odds on being older for terminated employees divided by the odds on being older for retained employees, which is:

$$\text{Odds Ratio} = \frac{56 / 23}{195 / 205} = 2.56$$

Screen 4.2. Statistical Analysis: Fisher's Exact Test

A statistical test called *Fisher's exact test* shows that if 79 employees out of 479 employees were selected at random for termination, where 228 employees were under 40 years old and 251 were 40 years of age or older, the probability that the rates of termination for older and younger employees would be at least as different as observed in these data would be approximately three in ten thousand.

Screen 4.2.1. Expected Values

The statistical analysis of Table 1 compares the data observed in each *cell* (combination of employment action and age group) with what would be expected from knowing just the row and column totals. For example, if 79 out of 479 employees (16.5%) are terminated, we would expect approximately 16.5% of the employees under 40 years of age to be terminated, which equals 37.6 employees, if age were not a factor in termination. Because only 23 of the employees under 40 years old were terminated, they were terminated at a rate below expectation. Similarly, employees 40 years of age and older were terminated at a rate above expectation (56 actually terminated compared to 41.4 expected).

Screen 4.2.2. A Hypergeometric Random Variable

A statistical analysis can determine whether the differences between expected and observed values in Table 1 are greater than would be likely to result by chance, if age were not a factor. A test called *Fisher's exact test* computes the probability of all observations at least as far away from the expected values as the values actually observed. It assumes that the row and column totals, which often are called *marginal totals* or *marginals*, are given, or *fixed*. This means that, for this example, we assume that of the 479 employees, 79 had to be selected for termination, and the question is whether employees 40 years of age or older were selected at a rate more different from the rate for employees under 40 than would be expected by chance.

In a *two-by-two table* such as Table 1, if the marginals are fixed, then determining the value of any cell in the table necessarily determines the value of the three remaining cells. Otherwise the rows and columns would not add up to their fixed marginals. If the number of employees falling into any one cell is determined by chance alone, subject only to the fixed marginals, then the number of employees falling into that cell is a *hypergeometric random variable*.

Screen 4.2.3. *p*-Values

Fisher's exact test can be either *one-* or *two-tailed*. A two-tailed Fisher's exact test of Table 1 computes the probability, if terminations were determined at random without regard to age, of observing 56 or more employees 40 years of age or older being terminated, given that 16.5% of the workforce is terminated, as well as the probability of observing 26 or fewer employees 40 years of age or older being terminated (because these observations would also be at least as far from the expected value of 41.4 as what was observed). A one-tailed Fisher's exact test of Table 1 computes only the probability of observing 56 or more employees 40 years of age or older being terminated. Note that because the marginals are fixed, this is the same as computing the probability of 23 or fewer employees under 40 years old being terminated, or the probability of 205 or more employees under 40 years old being retained, or the probability of 195 or fewer employees 40 years of age or older being retained.

The probabilities computed by Fisher's exact test are .0003, two-tailed, and .0002, one-tailed. These probabilities would generally be reported by a statistician as $p = .0003$, or $p = .0002$, respectively, and they generally are referred to as "*p*-values." Interpreting the two-tailed test, a disparity in termination rates between the two age groups in Table 1 at least as large as that observed would result by chance approximately three in every ten thousand times. Because such a result would be so unlikely, this statistical analysis supports plaintiffs' prima facie case of their age being a factor in Premium's decision to terminate them.

Screen 4.3. Other Statistical Analyses

Seldom is there only one correct way to analyze data. In this example, plaintiff presented results from Fisher's exact test comparing the termination rates of employees 40 years of age and older to the termination rates of employees under 40 years of age. Plaintiff, instead, might have compared these rates with a *chi-squared test*. Alternatively, plaintiff might not have looked at just two age groups, but tested for a trend of increasing termination rates with age across all the age levels.[6] Plaintiff might have used *bivariate logistic regression* to do this.

6. Trend tests are illustrated in J.L. Gastwirth, *Statistical Evidence in Discrimination Cases*, 160(A) J. ROYAL STAT. SOC'Y 289 (1997).

Screen 4.3.1. Chi-Squared Test

Before powerful computers and ingenious algorithms made the computations involved in Fisher's exact test feasible for data such as those in Table 1, statisticians would usually have analyzed the data using a chi-squared test. The chi-squared test is so-named because the probabilities associated with the statistic computed in a chi-squared test are closely approximated by a theoretical probability distribution called the *chi-squared distribution*. According to the chi-squared test, the probability of observing by chance termination rates for older and younger employees at least as disparate as those observed in these data is approximately one in a thousand.

Screen 4.3.1.1. The Chi-Squared Formula

The general formula for chi-squared (χ^2) compares the number of observa tions in each unique category (O_i) to the number expected for each unique category (E_i), squares the difference, divides by the number expected, and sums the results for all unique categories.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

A handy formula for computing chi-squared for a two-by-two table appears in Table 2. The formula refers to the four cell values in the two-by-two table by the letters *a* through *d*, where *a* refers to first row, first column, *b* refers to first row, second column, *c* refers to second row, first column, and *d* refers to second row, second column (see Table 2). The formula in Table 2, unlike the general formula above, includes what is known as *Yates' correction for continuity*, which is considered appropriate when the marginals are fixed. *See*, *e.g.*, DAVID C. HOWELL, STATISTICAL METHODS FOR PSYCHOLOGY 146 (4th ed. 1997).

Table 2. Computation of Chi-Squared with Yates' Correction for Continuity for a Two-By-Two Classification Table

	Column 1	Column 2	Total
Row 1	*a*	*b*	*a+b*
Row 2	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*N*

$$\chi^2 = \frac{N(|ad - bc| - N/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

See WILLIAM L. HAYS, STATISTICS 861 (5th ed. 1994).

Screen 4.3.1.2. The *p*-Value

For the data in Table 1, chi-squared is equal to 12.09.

$$\chi^2 = \frac{479(|23 \cdot 195 - 56 \cdot 205| - 479 / 2)^2}{79 \cdot 400 \cdot 228 \cdot 251} = 12.09$$

The two-tailed probability of observing a chi-squared value at least this large, with one *degree of freedom* as exists here, is .001.

Because the chi-squared test only approximates the results of a Fisher's exact test, the Fisher's test usually is preferable if computer software for statistical analysis is available. Use of the chi-squared test instead 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. GEORGE W. SNEDECOR & WILLIAM G. COCHRAN, STATISTICAL METHODS 127 (8th ed. 1989).

Screen 4.3.2. Bivariate Logistic Regression

Logistic regression is a statistical technique in which a computer is used to derive an equation that expresses the probability of being in one of two categories as a function of one or more *independent variables*. Category membership is regarded as the *dependent variable*. Because the dependent variable in this case can take one of two values, it is a *dichotomous* variable. In this case, the two possibilities are "retained" and "terminated." With *bivariate* logistic regression there are only two variables analyzed—the dependent variable and one independent variable. In this case, the independent variable is age.

According to a logistic regression analysis, the probability of observing by chance a relation between age and probability of being terminated at least as large as that observed in these data is approximately one in ten thousand.

Screen 4.3.2.1. The Logistic Regression Equation

The type of equation derived with logistic regression is one that expresses the *natural logarithm* of the *odds* on being in one group as opposed to the other as some number (the constant) plus or minus some other number (a *coefficient*) times the value for the independent variable. (Other types of equations could be derived from the data with other types of statistical techniques.) The constant and the coefficient are selected by the computer, using a process called maximum likelihood estimation, so that the results of the equation are as compatible with the actually observed data as possible.

In this case, a logistic regression on plaintiffs' data yields the following equation:

$$\ln\left(\frac{\text{P(Terminated)}}{\text{P(Retained)}}\right) = -3.546 + 0.044 \cdot (\text{age})$$

Here, P(Terminated) refers to the probability of being terminated and P(Retained) refers to the probability of being retained. Their ratio is the odds on being terminated. Logistic regression assumes that the natural logarithm of the odds on being terminated is a linear function of age.

For example, for a 55-year-old employee, the probability of being terminated would be predicted to be equal to 24.5%, based on these employment data and the logistic regression analysis. This is how the prediction can be computed:

$$\ln\left(\frac{\text{P(Terminated)}}{\text{P(Retained)}}\right) = -3.546 + 0.044 \cdot (55) = -1.126$$

$$\Rightarrow \text{Odds(Terminated)} = \left(\frac{\text{P(Terminated)}}{\text{P(Retained)}}\right) = e^{-1.126} = 0.324$$

$$\Rightarrow \text{P(Terminated)} = \frac{\text{Odds(Terminated)}}{1 + \text{Odds(Terminated)}} = \frac{0.324}{1.324} = 0.245$$

The natural logarithm of the odds on being terminated would be predicted to be -1.126, which is -3.546 plus 0.044 times the age of 55. This implies that the odds on being terminated is the natural number *e*, which is approximately equal to 2.718, raised to the power -1.126, and this is equal to 0.324. The odds on being terminated for a 55-year-old, therefore, are approximately equal to one to three. This corresponds to a probability of approximately ¼. By actual computation, the predicted probability of being terminated is equal to 24.5%.

For a 32-year-old employee, on the other hand, the predicted probability of being terminated is only 10.5%.

$$\ln\left(\frac{\text{P(Terminated)}}{\text{P(Retained)}}\right) = -3.546 + 0.044 \cdot (32) = -2.138$$

$$\Rightarrow \text{Odds(Terminated)} = \left(\frac{\text{P(Terminated)}}{\text{P(Retained)}}\right) = e^{-2.138} = 0.118$$

$$\Rightarrow \text{P(Terminated)} = \frac{\text{Odds(Terminated)}}{1 + \text{Odds(Terminated)}} = \frac{0.118}{1.118} = 0.105$$

It is possible to express the relation between the predicted probability of being terminated and age as a single, albeit intimidating, equation.

$$\text{P(Terminated)} = \frac{e^{-3.546 + 0.044 \cdot (\text{age})}}{1 + e^{-3.546 + 0.044 \cdot (\text{age})}}$$

The predicted probability of being terminated, given that age is equal to 55, can be expressed as P(Terminated | age = 55), because in probability notation a vertical bar in a probability expression customarily indicates a specific condition for which the probability is computed. Similarly, the predicted probability of being termi-

nated, given that age is equal to 32, can be expressed as P(Terminated | age = 32). Using the alternative equation, these probabilities are the same as those already computed.

$$P(\text{Terminated}|\text{age} = 55) = \frac{e^{-3.546 + 0.044 \cdot (55)}}{1 + e^{-3.546 + 0.044 \cdot (55)}} = 0.245$$

$$P(\text{Terminated}|\text{age} = 32) = \frac{e^{-3.546 + 0.044 \cdot (32)}}{1 + e^{-3.546 + 0.044 \cdot (32)}} = 0.105$$

These, of course, are only two examples. Figure 1 displays the relation between predicted probability of termination and age for all ages in plaintiffs' data set. These ages range from 17 to 77.

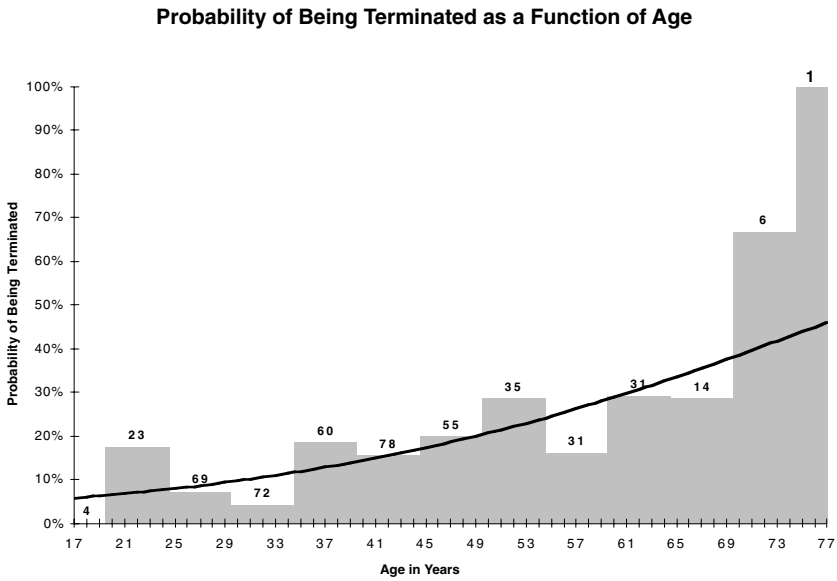

Figure 1. The solid line in this figure, which curves slightly and slopes up from left to right, represents predicted probability of termination as a function of age as computed by logistic regression from plaintiffs' data. The vertical bars in the figure with numbers atop them are included for comparison. Each bar represents the actual proportion of employees terminated for a five-year age span. For example, among employees 17 to 19 years of age, none was terminated, so there is no bar. Among employees 20 to 24 years of age, approximately 17% were terminated, so that is how high the bar is for that age group. The number atop each bar represents how many employees are in that age group. The termination rates in the age groups with more employees have a larger impact on the logistic regression results than the termination rates in the groups with fewer employees.

Screen 4.3.2.2. The *p*-Value

As with Fisher's exact test and the chi-squared test, the statistician can test whether the results of the logistic regression are unlikely to result by chance. To do this, the statistician tests the *statistical significance* of the coefficient for the independent variable. For these data, the probability of computing a logistic regression coefficient for age as far away from zero as the computed value of 0.044 is approximately 0.0001, or one chance in ten thousand.

The logic of statistical significance in a regression context is more complex than the logic of statistical significance in some other contexts. Here we assume that a logistic regression equation describes the relation between the independent variable (age) and the dependent variable (retained or terminated). What we do not know is the regression coefficient for the independent variable. So we estimate it from the data. Our estimate is 0.044. If the real value were zero, then our estimate would be as far away from zero as 0.044 approximately 0.01% of the time. (Note that if the coefficient for age were zero, then the probability of being terminated would be constant regardless of age.)

Screen 5. Defendant's Facts

Defendant produced evidence to rebut plaintiffs' prima facie case. Defendant contended that employee age was not a factor in its termination decisions—employees were terminated on the basis of defendant's need for the services provided by the employees and the employees' records of performance evaluations. In addition, managers charged with making the termination decisions were instructed to consider employees' loyalty to the new management. Defendant argued that because its merger and reorganization inspired significant employee dissent, it considered loyalty to new management, as opposed to nostalgia for old management, in assessing future productivity of its employees. Premium argued further that the reason termination rates appeared to be related to age was simply that employees with more years of service tended to be older.

Screen 6.1. The Data

Defendant offered as evidence data on age and years of employment with defendant for all 479 employees. These data are plotted in Figure 2. This type of plot is known as a *scatter plot* or *scattergram*. In this case, the horizontal axis (often referred to as the *x*-axis or *abscissa*) represents age of employee and the vertical axis (often referred to as the *y*-axis or *ordinate*) represents years of service, or years of employment with defendant. Each dot in the plot represents one employee, and the dot's position relative to the horizontal axis represents the employee's age, and the dot's position relative to the vertical axis represents the employee's years of service.

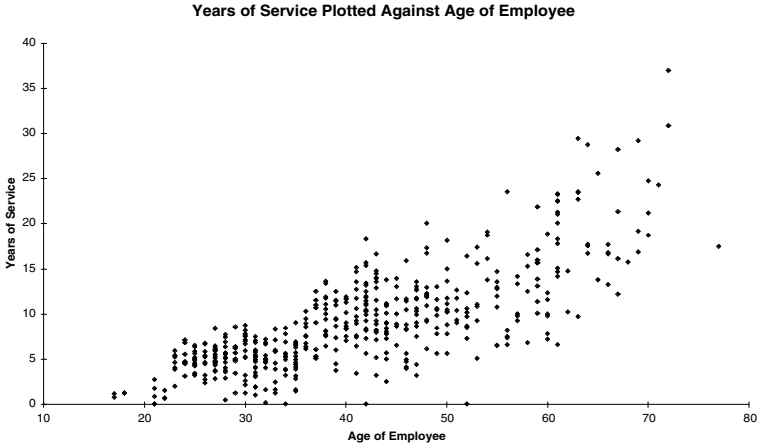

Figure 2. Years of service plotted against age of employee.

Defendant presented a second plot with the same data, but this time the employees who were terminated and the employees who were retained were plotted with different symbols. This plot is presented in Figure 3.

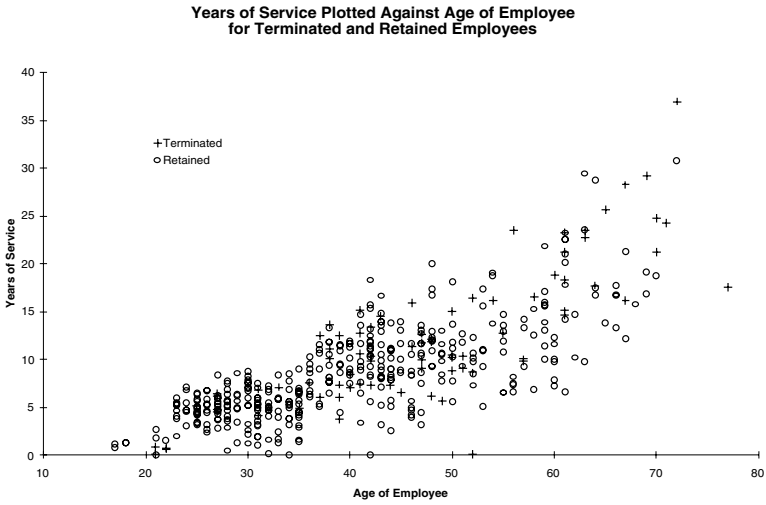

Figure 3. Years of service plotted against age of employee with a "+" representing a terminated employee and a "**o**" representing a retained employee. Of the 479 employees total, 79 (16.5%) were terminated and 400 (83.5%) were retained.

Screen 6.2. Statistical Analysis: Logistic Regression

Defendant argued that some employees who were terminated were selected for termination in part because of their years of service, but no employees were terminated because of their age. That older employees were more likely to be terminated resulted from their greater likelihood of having more years of service, defendant argued. To prove this explanation, defendant performed a *multivariate logistic regression analysis*. Multivariate logistic regression is similar to bivariate logistic regression, except that with multivariate logistic regression there is more than one independent variable.

Defendant argued that its statistical analysis showed that employees who had worked for the company for a long time were more likely to be terminated than recently hired employees, because the more senior employees were less likely to be loyal to new management. Although the more senior employees tended to be older than new hires, when length of employment is taken into account, age of employee is not statistically associated with likelihood of termination.

Screen 6.2.1. The Logistic Regression Equation

Here, the dependent variable is whether the employee was terminated or retained. The independent variables are (1) age and (2) years of service. According to defendant, a logistic regression analysis of the data depicted in Figure 3 resulted in the following equation.

$$\ln\left(\frac{\text{P(Terminated)}}{\text{P(Retained)}}\right) = -3.096 + 0.077 \cdot (\text{years of service}) + 0.016 \cdot (\text{age})$$

The following equation is merely another way to express the same relations.

$$\text{P(Terminated)} = \frac{e^{-3.096 + 0.077 \cdot (\text{years of service}) + 0.016 \cdot (\text{age})}}{1 + e^{-3.096 + 0.077 \cdot (\text{years of service}) + 0.016 \cdot (\text{age})}}$$

For example, for a 45-year-old employee who had been with defendant for thirteen years, the probability of being terminated would be predicted to be equal to 20.2%, based on the employment data and the logistic regression analysis.

$$\ln\left(\frac{\text{P(Terminated)}}{\text{P(Retained)}}\right) = -3.096 + 0.077 \cdot (13) + 0.016 \cdot (45) = -1.375$$

$$\text{P(Terminated)} = \frac{e^{-3.096 + 0.077 \cdot (13) + 0.016 \cdot (45)}}{1 + e^{-3.096 + 0.077 \cdot (13) + 0.016 \cdot (45)}} = 0.202$$

Figure 4 illustrates the estimated probability of termination as a function of years of service and age of employee. Estimated probabilities are given by the

logistic regression equation. The figure shows that probability of termination is more substantially related to years of service than age of employee.

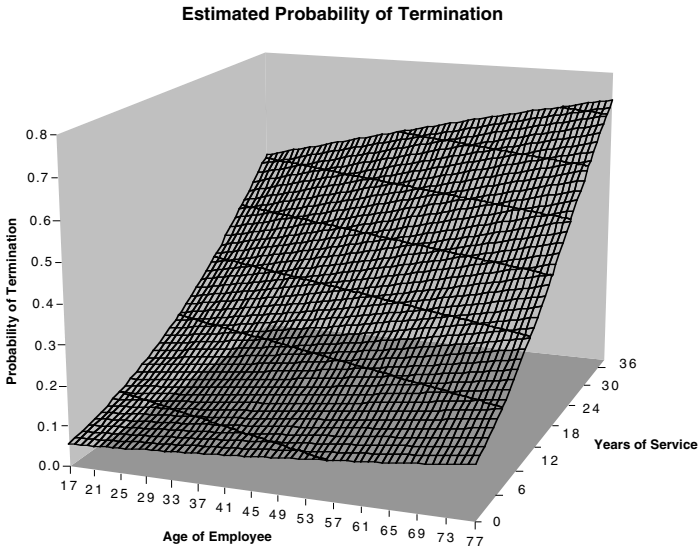

Figure 4. This three-dimensional chart plots estimated probability of termina tion, according to the logistic regression analysis, as a function of years of service and age of employee.

Screen 6.2.2. The *p*-Values

Although this analysis suggests that the probability of being retained is a function both of age and years of service, defendant performed *statistical significance* tests on the *coefficients* for age and years of service in the equation. According to these tests, the probability of obtaining a number as far away or farther from zero as the coefficient for years of service—the number 0.077—is approximately 0.019, or less than two in a hundred, if there were no true statistical relation between being terminated and years of service. On the other hand, the probability of observing a number as far away or farther from zero as the coefficient for age—the number 0.016—is approximately 0.307, a rather high probability. Defendant argued that this analysis shows that years of service was a real factor in determining whether its employees were terminated, but age was not.

Screen 7. Legal Analysis

The statistical evidence supports defendant's argument that its termination decisions were based on factors that included length of employment, but did not

nclude age as a separate factor. The data, therefore, do not support an action for ge discrimination on a *disparate treatment* theory. The Supreme Court has not yet etermined whether *disparate impact* is actionable under the ADEA. If it is, it is a uestion of fact whether defendant's reliance on perceived loyalty to new managem-ent, or length of service as a proxy for loyalty, was sufficiently necessary to the peration of its business to save defendant from liability arising from the disparate mpact such a policy would have on older employees.

this page is blank in original

COMMENT

Orley Ashenfelter*

CITATION: Orley Ashenfelter, Comment on the Age Discrimination Example, 42 Jurimetrics J. 297–300 (2002).

For two reasons, the use of statistical reasoning has become a common theme of the legal process in many areas. First, statistical reasoning and its implementation have become increasingly accessible with the spread of modern computers and readily available machine-readable data. In short, it is easier and cheaper to engage in statistical analyses than it has ever been. Second, the kind of complex, technical issues that confront the legal system are increasingly amenable to helpful explication by statistical methods.

Tim Reagan's software prototype[1] is motivated by the first of these factors, accessible data and software. Ironically, the legal issue it explicates, litigation over age discrimination,[2] is a standard problem in legal factfinding. I think this is both the strength and the weakness of this prototype example for the purpose of teaching statistical methods to those active in the legal system. On the one hand, because the empirical issue is clear (is there evidence the employer based termination decisions on age?), the kind of statistical methods to be employed are routine. This makes it much easier to provide the routine teaching materials that will be helpful in the analysis of the empirical issues. On the other hand, the nature of this routine material also makes it difficult to keep the reader's interest—in my experi-

* Orley Ashenfelter is Joseph Douglas Green 1895 Professor of Economics, Princeton University. The author is indebted to the literally hundreds of federal judges who have participated in his classroom teaching of statistical methods in the last two decades through various programs sponsored by the Law and Economics Centers at the University of Miami, Emory University, and George Mason University and by the Federal Judicial Center.

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

2. *Id.* at 282.

ence, this is the key to the successful teaching of statistical methods to those (lawyers!) who have selected a way of life deliberately designed to avoid them.
In what follows I provide constructive comments on several aspects of the prototype that I hope might make this kind of statistical material both more interesting and more credible.

I. SIGNIFICANCE: STATISTICAL AND SUBSTANTIVE

One of the most difficult issues in the teaching of statistical reasoning, but one that intrigues virtually all who encounter it, is the difficulty in determining what is a significant finding. In this prototype example, Reagan avoids the usual discussion of significance levels by reporting on *p*-values, or the probability of rejecting the null hypothesis of equal treatment, given the actual data configuration.[3] Instead of discussing the rejection of a null hypothesis, particular *p*-values are said to be "unlikely"[4] or "rather high,"[5] indicating rejection and the failure to reject, respectively.

Although I am sympathetic to the purpose of avoiding hard criteria for "statistical significance,"[6] I am not sure that this is the best way to do so. "What," I can hear the student ask, "makes 0.3 (the *p*-value of the coefficient on the age variable in the logit regression) a 'rather high' probability?" It would be better to provide an explicit but brief discussion of the criteria that lead to the conclusion that the particular result is "likely" or "unlikely" to be due to chance. In principle, different factfinders may use different standards—all that we can provide are the key criteria that should enter into the reasoning. In short, I think the example would benefit from a standard, old-fashioned discussion of Type I errors (concluding the firm discriminated when it did not) and Type II errors (concluding the firm did not discriminate when it did).

A separate but related issue is how important a statistical finding is to the substantive context in question. Minuscule differences from the null hypothesis will always be detected with large enough samples. Is a difference in the termination rates of 0.14 between those 32 and 55 years of age a *substantive* difference? Or, if both age and seniority have estimated positive effects on terminations, is it significant that the effect of seniority is 5 times as large as the effect of age?

The implied answers to both these questions given in the prototype are "yes," but the prototype is silent on the criteria that would be used by a scientist to determine substantive significance. Although it is difficult to provide even a rudimentary guide to the notion of substantive significance, doing so would increase the salience of the example and increase the reader's interest in it.

---

3. *Id*. at 286.
4. *Id*.
5. *Id*. at 294.
6. *Id*. at 291.

II. CAUSALITY

One of the most difficult aspects of the use of statistical reasoning is establishing the explanation for an empirical finding. It is well known that statistical reasoning alone rarely provides compelling evidence for causality. For most observers, however, causality is at the heart of what they wish to establish.

In the prototype, for example, the defendant attempts to rebut the claim that age is the cause of terminations with the argument that seniority, not age, is the cause of terminations.[7] Since seniority and age are only imperfectly correlated, it is, in principle, possible to assess this claim with data. Indeed, the defendant provides evidence that seniority is related to termination rates.[8]

What are the claims for causality being made by the plaintiff and defendant? The plaintiffs' causal claim is presumably that the employer's differential treatment resulted directly from an employee's age—and this claim leads immediately to the plaintiff's statistical analysis.[9] The defendant's causal claim appears to be that more senior workers were inferior in satisfying the defendant's need for services, had lower performance evaluations, and were less loyal to new management.[10] To implement these claims, the defendant must make the *additional* claim that seniority is an empirical proxy for these factors.

When laid bare in this way, it becomes apparent that the defendant's claims might be tested in several other ways, by using data on performance evaluations, salaries, and perhaps other variables. Such discussions, although they lead to ambiguity, stimulate interest and make it clear that correlation does not establish causality.

III. SELECTION EFFECTS IN THE EVIDENCE

A disturbing feature of statistical evidence in litigation is the possibility that the evidence is provided only because it is favorable to a particular party. If evidence is provided only because it is favorable, then it is very difficult to interpret conventional *p*-values.

There are many ways to demonstrate this point with concrete examples. Suppose, for example, that an unscrupulous firm wishes to advertise the effectiveness of a particular product that it knows to be ineffective based on evidence from randomized trials. By selecting to test the product in many different labs independently, this firm can find at least one result that will establish effectiveness at any *p*-value it selects. Then, to sell its product, all the firm need do is report the results from the lab that shows the product is effective and ignore all the others. (For example, 3 in 100 of the labs will discover that the performance they observe for the product would happen in only 0.03 of cases. Naturally, the firm will report only these three results, ignoring what was found by the other ninety-seven labs.)

7. *Id.*
8. *Id.* at 291–95.
9. *Id.* at 282–83.
10. *Id.* at 291.

The importance of being alert to the possibility of manipulative selection in the reporting of results is that it provides a strong incentive for the factfinder to (1) insist on reports of all the evidence, not just a part, and (2) consider alternative types of evidence relevant to the same issue. Although it provides no guaranteed resolution for the potential problem of the misuse of *p*-values, it does at least provide a mechanism that may help clarify the role of the statistical evidence.

This prototype statistical analysis should be a useful device for showing how the tools of statistical reasoning can help in the appraisal of the facts in a relatively routine type of litigation. The difficult question is, will the prototype be helpful in developing an appreciation for statistical reasoning in other situations?

COMMENT

David W. Barnes*

CITATION: David W. Barnes, Comment on the Age Discrimination Example, 42 Jurimetrics J. 301–307 (2002).

This comment addresses two questions: How useful is the "example and hyperlink method" as a teaching as opposed to reference tool? And, what do judges need to know about statistical analysis? The discussion of the example-and-hyperlink method includes observations about the accessibility of this teaching method to judges, whether it enables judges to generalize from the specific examples, and how the prototype might be improved by comparing the utility of statistical tests. The discussion of what judges need to know draws from years of teaching statistics to judges but still rests only on my intuition about what information about statistics they need to do their work as judges intelligently. Judges need to know less about formulas and more about when particular statistical approaches are appropriate, as well as the relative merits of statistical methods for different contexts, types of data, and questions.

I. THE EXAMPLE AND HYPERLINK METHOD

A. Accessibility

Robert Timothy Reagan has done a splendid job of illustrating the challenges in teaching judges statistics. It is apparent that every sentence uttered by a statistician needs explanation. The logistic regression example brilliantly illustrates the challenge. The paragraph introducing the topic requires cross-reference, presum-

* David W. Barnes is Distinguished Research Professor of Law, Seton Hall University. He is author of *Statistics as Proof: Fundamentals of Quantitative Evidence* (1983), the first casebook on statistical evidence.

ably by hyperlink, to *independent variables*, *dependent variables*, *dichotomous variables*, *regression*, *bivariate*, and both *logarithm* and *logistic*.[1] And, what does it mean to *derive an equation*? Presumably the hyperlink to *independent variable* would have its own hyperlink to *variable*, which in turn would have a hyperlink to *dichotomous*, and an explanation of what *independent* means. To understand the single tree that is logistic regression, many judges will have to descend through the trunk using hyperlinks to the root system that is the foundation of inferential statistics.

Statistical methods, like aspen trees, share a common root system. The foundational knowledge is not only the types of variables (e.g., dichotomous, independent) or analysis (e.g., bivariate). It also includes the rationale, process, and language of statistical inference. The aspen tree that starts a stand or forest of aspen sends out lateral roots for many yards that send up woody shoots that look like individual aspen trees. The entire forest shares a common root system. To understand an aspen shoot, one must understand the root system.

Statistical science has expanded in the same way. The rationale, process, and language of statistical inference are part of the root system. This root system is more than a collection of terms that can be explained in hyperlinks. Like the roots, they are interwoven. Hypothesis testing, for instance, can be illustrated by examples, but the logic of "failing to reject null hypotheses" rather than "accepting null hypotheses" using hyperlinks seems disjointed.

I am somewhat at a loss to explain why the example-and-hyperlink method may not work for foundational material, but two possibilities occur to me. One is that the software user may not know the importance of understanding a hyperlinked term and fail to appreciate a critical underlying term (such as the requirement of *independent* variables). The software user may not be able to tell whether he or she fully understands the content of a screen. That is why texts usually give all the basics before going on to the more complicated material. I am sure that Dr. Reagan envisions the judge starting at the beginning of the software and working all the way through, but a judge is equally likely to click on *logistic regression* and fail to understand the explanation. Hyperlinks strike me as more useful for supplying definitions than explaining logic.

The other reason the example-and-hyperlink approach may not be the best method for presenting foundational material is that the logic requires a more lengthy discussion than is easily presented on "screens." The logic of inferential reasoning is not a series of PowerPoint bullet points. Screen 4.3.2 illustrates the difficulty of explaining something complex in a series of slides. It appears to go on for pages. I would have to see the software presentation in operation to see whether this "slide" is more like an online text, which might be more effective for some teaching.

The designers of the prototype recognized that an individual statistical method cannot be appreciated without understanding the root system. Sophisticated con-

1. Robert Timothy Reagan, *Federal Judicial Center Statistical Examples Software Prototype: Age Discrimination Example*, 42 JURIMETRICS J. 281, 282 (2002).

sumers of statistics who understand the clever hyperlink system supporting the trees can devote their attention to the individual trees, as this particular excerpt from the prototype does. The prototype may be a great method for teaching about individual statistical techniques.

B. Generalizability

To be a useful teaching device, the example-and-hyperlink method must leave the judge able to apply the tool in another context. Will the software be organized around substantive law areas or around statistical tools? If it is the former, does a substantive law example lead judges to believe that the methods described are the only acceptable ones for that substantive legal question? Is the FJC implicitly giving its imprimatur to those tools? More important, how does the judge translate from the application in one substantive area to another? Organizing around statistical tools may not be as appealing to the legal user, but it might have the virtue of describing for each tool or statistical test the circumstances for which it is appropriate. It would be nice to see an outline of the entire project to see whether a user can learn enough from the application of logistic regression in the age discrimination context to see how it would be used to project lost earning capacity, for instance.

C. Applicability

A teaching method for judges must leave users with the ability to determine whether the tool was appropriately applied in another context. I am thinking not of another substantive legal context, but of another analytical context, with different sorts of data (continuous rather than categorical, for instance) or variables with different distributions. Again, the prototype might be improved by a general description of when each tool or method is appropriate. For instance, the discussion of the chi-squared test only describes its utility as something statisticians used "[b]efore powerful computers and ingenious algorithms made the computations involved in Fisher's exact test feasible."[2] It might be helpful, in the introductory material, to explain when a statistician's use of the chi-squared test is acceptable. First, what is the general applicability of the chi-squared test? Second, Screen 4.3.1 hints that the chi-squared test should not be used in this century. Is that the intended message? Should a judge refuse to admit evidence based on a chi-squared test?

II. WHAT JUDGES NEED TO KNOW

A. Reference Works and Teaching Materials

I have tried to be careful in the preceding comments to emphasize the teaching rather than reference applications of the prototype, since this product is being developed to teach judges. While the specific tools (logistic regression, chi- squared

2. *Id.* at 287.

analysis, and Fisher exact test) may be taught by the example-and-hyperlink method, I am not sure it works for the foundational materials. Although I have reservations about the hyperlink approach for teaching foundational statistical materials, it is a terrific structure for a reference source. For reference rather than teaching, the prototype is likely to be extremely useful to judges, who could use it as a dictionary or encyclopedia.

Use of a technical dictionary naturally requires some foundational knowledge. It would be difficult to learn the law from *Black's Law Dictionary*. The Federal Judicial Center already has such an encyclopedia, of course, in its *Reference Manual on Scientific Evidence*.[3]

The various chapters of the *Reference Manual on Scientific Evidence* are organized in a similar way. Each chapter poses a series of questions judges might ask about a particular courtroom presentation. In the section on data collection, for instance, the subsections are: "*Is the Measurement Process Reliable?,*" "*Is the Measurement Process Valid?*" and "*Are The Measurements Recorded Correctly?*" This organization focuses on what judges need to know about the statistical evidence. Perhaps the FJC's prototype would benefit from consideration of the context in which a judge would use this product.

A judge may be called on to understand statistical evidence in the roles of gatekeeper and factfinder. What would a judge in either of these contexts want to know about the proffered evidence? Judges would probably want to know many of the same things statisticians would when evaluating someone else's work. A statistician hearing the evidence is unlikely to recalculate the summary statistics; a judge never would. Do judges need to know the formulas for calculating chi- squared, an odds ratio, or Fisher's exact test? I think they need to know it only if it helps them determine whether the test is appropriate. As a *Daubert* gatekeeper excluding unreliable evidence, a judge might consider whether and when statisticians agree that using the chi-squared test is appropriate. As a factfinder, a judge might consider how much weight statistical evidence based on a chi-squared test rather than the Fisher Exact formula should be given. In its present form, the prototype does not help with those decisions.

B. Choices About Including Formulas

Deciding which formulas to include and explain is not easy. Revealing some formulas may help the user understand a statistic. I have seen some lights go on when judges recognize that the formula for the standard deviation resembles the formula for the average of a list of numbers (in that case, a list of differences from the mean). Otherwise, I am unsure what utility the formula has to someone who is not calculating the statistic.

---

3. The chapter on statistics is, fortunately for judges who master it before using the statistical software, an excellent set of foundational teaching materials. *See* David Kaye & David Freedman, *Reference Guide on Statistics*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 83 (Federal Judicial Center ed., 2d ed. 2000).

Judges do need to know something about the data requirements for a test and that might be illustrated by discussing the formula for a statistic. Assuming, for instance, that there is some legitimate excuse for using a chi-squared test now that we have "computers and ingenious algorithms," judges probably need to know the difference between observed and expected values to understand when the Yates correction for continuity is appropriate. Discussing how the chi-squared statistic is calculated might be a useful way of explaining what observed and expected values are and why they are important. Otherwise, I am not sure why a judge would want to know the formula. It is impossible to make any wise general statements about how much information is enough. The general guideline, however, should be whether the information helps the judge determine that the test was appropriate to the application at hand.

Choices about including formulas are complicated by the nature of some formulas. Showing the calculation of the Fisher exact test for the age discrimination example would be excruciating. That is undoubtedly why it is not shown. It would also be unproductive to show it, even if it were a short, straightforward bit of arithmetic. But why provide a detailed calculation of the chi-squared statistic and none for the Fisher exact test, which is, according to Screen 4.3.1.2, preferred? A judge might also wonder why there is any discussion of the computations for the chi-squared test at all, knowing that there are always computers available. Should a judge always reject chi-squared evidence? If not, when? Is it still appropriate to apply the "traditional" rule described in Screen 4.3.1? I imagine a statistician would say "Judge, it doesn't matter which test you use in this case." Could the software supply some reason for the judge to evaluate the statistician's assertion? Perhaps the software could be designed to give more explicit guidance.

Finally, with respect to including equations, Screen 4.3.2.1 is a tour-de-force explanation of the logistic regression equation. A judge would not *have* to go to that screen, but what general lesson would a judge take from those calculations? Could a judge apply this in a different context? Is there any reason for a judge to want to calculate these odds, unless he or she is attempting to replicate the statistician's work?

Aside from the lack of utility of that formula, presenting the formula raises its own questions, unanswered in the prototype. What, for instance, is a *natural log*? Are there other kinds of logs? Perhaps *unnatural*? Why do we use *natural logs* at all? And, what is this number *e*? If I did not already know the answers to these questions, I would feel the way I do reading a particularly difficult text written in French. When I must look up every word, I stop caring. And that is before being told by Screen 4.3.2.1 that odds of one to three corresponds to a ratio of 1 to 4. I know the hyperlinks are available on many of these concepts, but . . . .

Dr. Reagan chose wisely in developing the difficult topic of logistic regression for the prototype. Logistic regression is a nice example of the relative importance of equations and explanations. Explanations of the reasons for using logarithms and ratios, the implications of transforming variables, and the appropriateness of analysis without transformations are more important to judges in their roles

as gatekeepers or fact-finders than are the equations. Nothing except the daunting nature of the task, however, prevents the courageous inclusion of both. The hyperlink method, by including a series of "When-Is-This-Test-Appropriate?" and "Compare-This-Test-To-Another" hyperlinks as well as "Check-Out-The-Equation" hyperlinks, can offer whatever information the user desires.

C. Comparing Statistical Methods

Judges and other laypeople may reasonably wonder why different statistical tests give different *p*-values and what to do when that happens. Statisticians know that some tests have greater power, but rarely integrate that into the general discussion of the variety of available tests. This question must be on the minds of judges who are told on Screen 4.2.3 that the Fisher exact test gives a *p*-value of .0003, two-tailed, and .0002, one-tailed (without being told which to prefer, by the way, although the hyperlink to *One- and Two-tailed Tests* will take care of that); on Screen 4.3.1.2 that the chi-squared *p*-value is .001 (two-tailed, without explanation of why there is no one-tailed calculation); and on Screen 4.3.2.2 that the logistic regression *p*-value is .0001.

The Federal Rules of Evidence require federal judges to exclude unreliable expert testimony. Testimony based on a statistical test with insufficient power to answer the factual questions presented should be excluded. The appropriateness of a particular statistical method requires knowledge of more than which tests are appropriate for categorical as opposed to continuous variables. Statistical power is part of the analysis of whether a statistical test is appropriate in a particular context. At least, it seems part of the explanation for the different *p*-values in the age discrimination example. I would think judges determining the reliability or weight of statistical evidence would want to be able to evaluate this. Upon learning that different tests give different results, a judge is certain to be concerned about whether the appropriate test was used. The judge as gatekeeper probably has an obligation to be concerned.

Using the logistic regression analysis of the age discrimination example, I offer one more example of what judges may want to know. After this tutorial, judges can see and interpret the plots of data. But why it is useful to plot data? What can looking at the plots do for the judges' appraisal of the appropriateness of the statistical technique? Perhaps the software could explain that the plots show that the relationship between the variables is nonlinear rather than linear and therefore the data should be transformed. Otherwise, how is the judge to decide whether a logistic regression equation is superior to another model? Again, providing some way to compare methods is necessary. Perhaps the "Linear-Versus-Non-linear-Regression" hyperlink, one of the many ("Compare-These-Tests" hyperlinks) would address that problem.

The designers of the software are quite likely to already have thought of the difficulties I have raised. I apologize for any hint that they negligently overlooked something. It is hard to tell the elephant's appearance from a snapshot of its tail. Obliged to persist, however, I do recommend that the designers include a module on when to use what method of analysis. Since the choice of statistical tool depends on such things as the variables' distributions and relationships to one another and what one wants to know, this material needs to be presented in some integrated fashion. This can be done quite nicely using the example-and-hyperlink approach. I question whether the foundational material describing the methodology of statistical inference can be presented effectively by that approach. Lastly, a series of hyperlinks describing the utility and limitations of particular statistical tests and comparing tests would help judges evaluate the reliability and weight of statistical expert testimony.

this page is blank in original

COMMENT

John M. Conley*

CITATION: John M. Conley, Comment on the Age Discrimination Example, 42 Jurimetrics J. 309–313 (2002).

When trying to teach statistics to judges, one is inevitably confronted with the *Goldilocks and The Three Bears* conundrum: it is easy to make things too simple, even easier to make things too complicated, but almost impossible to get it just right. I fear that this is the fate of the Federal Judicial Center's model software problem.[1] It represents an admirable effort. The design is original, the problem itself is interesting, and the screens are full of useful information. I see it as an excellent tool for teaching statistical evidence to law students who have indicated their interest and aptitude by volunteering for such an elective course. It will also prove useful to judges who already have considerable experience with statistics. But I think that the vast majority of judges—beginning, captive consumers of statistics—will find that it gives them too little in the way of basics, but too much in the way of dense and intimidating detail.

I should begin with a brief note about my own experience with statistical evidence, which will also reveal something about my biases. I am not a professional statistician.[2] I am, rather, a lawyer who makes frequent use of statistics, both in academia and in practice. When I write on statistical topics, I always place

* John M. Conley is William Rand Keenan, Jr. Professor, University of North Carolina School of Law. He is the coauthor (with David Barnes) of *Statistical Evidence in Litigation: Methodology, Procedure and Practice* (1986), one of the first legal treatises on the topic. He has taught courses for judges in law and social science and in statistics.

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

2. I have a social science Ph.D., but it is in anthropology, an unapologetically qualitative field.

myself under professional tutelage, usually in the form of a co-author.[3] Notwithstanding this meager background, I have spent a great deal of time teaching basic statistics and other scientific methods to judges, often because I have been the best thing available on short notice. Since 1991, I have taught the law and social science course in the University of Virginia School of Law's Graduate Program for Judges (an LL.M. program). Over the same period, I have been the director of, and a statistics teacher in, Duke University School of Law's annual week-long "Judging Science" course. In addition to the several hundred judges I have taught in these and other venues, I have written and argued on statistical evidence before a number of courts, as an advocate rather than an expert. The point of all this is that, while I do not claim a professional knowledge of statistics, I am confident that I know as much as anyone about judges' capacities to learn it. I also believe that the run of judges with whom I have dealt is fairly representative, going well beyond the Posnerian super-judges who are customarily invited into law schools. Finally, I lack an intuitive sense of mathematics myself, so I feel their pain.

I realize, of course, that hands-on, experiential learning is now in vogue, at the expense of more abstract, didactic approaches. But my experience with teaching judges is that they require constant back-and-forth between theory and practice, with principles and applications continually reinforcing each other. It is unreasonable to expect them (or other non-specialist adult students) to deduce principles from relatively unadorned examples, or to understand and remember principles without the aid of vivid examples.

The software prototype, while providing a vivid legal example, offers insufficient general background to be useful to the beginner. The treatment of *p*-values illustrates the point. Screen 2 introduces the plaintiff's raw ratio data, which are straightforward enough.[4] The screen concludes with the remark that a disparity as great or greater than the one found would have a chance occurrence probability of about 0.03%.[5] This is the only point attributed to "statistical analysis" in the very first statistical screen,[6] so the reader will probably presume that it is important. But why? It is rarely self-evident to non-statisticians that the meaning of a disparity can be evaluated by calculating the probability that it would occur by chance. This question is finally addressed in Screen 4.2.3 (and again in Screens 4.3.1.2 and 4.3.2.2).[7] This, however, is deep within the program, beyond such concepts as odds ratios, Fisher's hypergeometric variables, and Fisher's exact test—concepts that may already have scared away the very people who need to understand the rationale for a *p*-value.

This is a long-winded way of saying that the program needs more introduc-

3. *See, e.g.,* DAVID W. BARNES & JOHN M. CONLEY, STATISTICAL EVIDENCE IN LITIGATION: METHODOLOGY, PROCEDURE AND PRACTICE (1986); David W. Peterson & John M. Conley, *Of Cherries, Fudge, and Onions: Science and Its Courtroom Perversion,* 64 L. & CONTEMP. PROBS. 213 (2001).

4. *See* Reagan, *supra* note 1, at 282–83.

5. *Id.* at 283.

6. *Id.*

7. *Id.* at 286, 288, 291. I understand that the plan is to have hyperlinks to definitions. Nevertheless, I think that an overarching conceptual issue like this is too important to handle in an aside.

tory material. It needs, for example, a general explanation of how statisticians evaluate the significance of a disparity, and how their evaluation can contribute to a judge's evaluation of legal significance. The example would immediately become an illustration of this larger point; the larger point could be continually reinforced throughout the evolution of the example. It is unrealistic to expect a beginner either to make sense of the cryptic introduction to *p*-values in Screen 2 or to fight through the ever-denser intermediate materials to reach the fuller explanations.

Admittedly, the program has to be kept short. I am usually all for brevity. But here, I think that what has been omitted is so important that it is worth the cost of another computer screen or two. However, if something must come out to put this in, I have a nomination: the logistic regression screens. I have done regression analyses, offered and attacked them in court, and taught the subject at an elementary level. These screens were hard. I read them. I concentrated. I reread them. I struggled. I cheated (I had to ask my math-teacher daughter to refresh me on natural logarithms; my failure of memory frustrated me). Finally, I got it. I was proud of myself. Then I wondered how many judges I know who would have stayed the course.

A related problem is the failure to anticipate questions that judges recurrently ask. For example, when an audience of judges has grasped the concept of a *p*-value (at, for instance, the 0.05 level), someone invariably asks, "So that means there's a 95% chance that the defendant is guilty or liable, right?" Wrong, of course, but that requires a fairly extended explanation that often runs contrary to judicial common sense. The second question is usually, "So how does this relate to the 51% preponderance-of-the-evidence standard?" It is not easy to explain why, if it relates at all, it does so in a very indirect way. Yet without such explanations, a little bit of statistical learning can be worse than useless—it can be dangerous.

I have two other general concerns. One relates to the equitable balance of the problem. My initial reaction was that it was too pro-plaintiff. The plaintiffs offer several statistical analyses, all tending to show age disparities that are very difficult to explain on the basis of chance alone.[8] The defendant's argument (as illustrated by Screen 6.2, for example) seems to boil down to this: "We didn't fire people because they were old per se. We fired them because they had been around a long time and thus might be loyal to their friends and colleagues; people who have been around a long time tend to be old."[9] If I were the plaintiffs' lawyer, I might stipulate to this argument. Most plaintiffs' lawyers that I know view their gut-level task as proving that the defendants are rotten people. Here, the defendant proves that point itself. I understand that the defendant's argument would create a legal defense, and that judges are not supposed to operate at a gut level. However, a teaching hypothetical should be balanced at every level.

The equitable and emotional imbalance of the case also may feed into the

8. *Id.* at 282–83, 283–91.
9. *See id.* at 291.

popular stereotype that social science evidence, including statistics, tends to favor the "progressive" side.[10] To some, this argues for lenient admissibility of such evidence; to others, it is grounds for rigorous scrutiny. I would prefer to dissolve these ideological associations. Statistical evidence is what it is, neither progressive nor reactionary. It should therefore be presented to judges in a context where the equities are more nearly balanced.

My final comments concern the discourse of statistics. Like any other science, statistics has its jargon. This cannot be avoided; even in an elementary exposition, terms of art must be used and explained. But, again like any other science, statistics also has a style, a way of rendering explanations. Much of this is a matter of preference and predilection, a reflection of the backgrounds and modal intellectual personalities of the people who practice statistics. But some of it is also the product of negotiation, of tacit agreements among professionals that certain ways of saying things are not wrong and thus acceptable.

It is not my intent to make a value judgment about statistical discourse. Indeed, I personally enjoy the precision, concision, and understatement that typify this discourse. It is, however, a distinctive style that is foreign to many literate readers, including those whose native language is law. Consider the example of Screen 4.2.2, which introduces the idea of a hypergeometric random variable.[11] I have always understood the point to be that it is a zero-sum game; if a person from one group gets a favorable outcome, then someone from the other group cannot. Each choice changes the odds for the next one.[12] I have often heard teachers use the analogy of drawing from a cookie jar which contains two kinds of cookies. This is different from, say, a bar exam with a fixed passing score that, in theory, everyone can pass. Rather than using such accessible language, however, Screen 4.2.2 turns to tight statistical prose and unapologetic statistical diction. The reader must instantly adjust, for example, to the idiosyncratic use of "marginal" as a noun.[13] Consequently, the reader's understanding of the definition of "hypergeometric random variable"[14] will depend on an appreciation of the terse and cryptic phrase, "subject only to the fixed marginals."[15] Similar problems permeate all of the screens that discuss specific statistical techniques.

Now, I understand that these phrasings are correct. Statisticians who use them

---

10. This stereotype goes back at least to the use of social science in *Brown v. Board of Education of Topeka*, 347 U.S. 483, 494–95, 495 n.11 (1954), and the irrationally hostile reaction to that evidence. *See* John M. Conley, *"The First Principle of Real Reform": The Role of Science in Constitutional Jurisprudence,* 65 N.C.L. REV. 935, 940 (1987). The stereotype has probably been reinforced by the use of statistical evidence by employment discrimination plaintiffs. *Id.* at 941–42.

11. *See* Reagan, *supra* note 1, at 285–86.

12. *See* BARNES & CONLEY, *supra* note 3, at § 4.17.

13 *See* Reagan, *supra* note 1, at 285.

14. *See id.* at 286.

15. *Id.* As evidence of just how odd this usage is outside of the circle of professional statisticians, I should point out that my WordPerfect spell checker has just rejected "marginals." There, it did it again.

will be immunized against charges of heterodoxy, let alone heresy. They are also grammatical; an intelligent and diligent reader should be able to work through them. Finally, they are short, which is usually a virtue. But why does the process have to be so hard? What's the rush? Why can't the explanation be given in looser, airier prose, using everyday words? Why not digress to a simple example or two? In other words, why not shape the writing to meet the needs of the audience?

Almost all of the judges to whom I have taught statistics and other scientific methods have come to the course exhibiting a certain fear and loathing. In my experience, the only way to deal with that predisposition is to begin with a demonstration, in everyday English, that statistical inference depends on relatively few concepts that are entirely logical. This requires a disproportionate amount of time, but the effort is essential. Thereafter, the adept and intrepid few can be led through the details. Despite my admiration for the project, I fear that the model problem has too little overview and too much detail. The Goldilocksian mean remains elusive.

this page is blank in original

COMMENT

Shari Seidman Diamond*

CITATION: Shari Seidman Diamond, Comment on the Age Discrimination Example, 42 Jurimetrics J. 315–320 (2002).

Federal judges face increasing demands to assess complex expert testimony that frequently includes statistical evidence. As a result, suitable teaching materials to help judges understand and feel comfortable with such technical evidence should be welcome indeed.[1] The software product foreshadowed in the age discrimination prototype by Robert Timothy Reagan[2] promises to make appropriate expertise more readily accessible. Judges can consult it at their leisure or when the demands of a particular case make it immediately relevant.

In commenting on the prototype, I offer two kinds of suggestions designed to build on a fine educational effort. The suggestions reflect a tension between coverage and concise presentation, and thus may appear inconsistent. My proposal is to add additional information that will further guide users and help them put the material in context, while keeping the presentation simple enough so that potential users are not discouraged from turning to it for assistance. The ingenious idea to present the material in the form of a software product makes both goals eminently achievable.

* Shari Seidman Diamond is Professor of Law and Psychology, Northwestern University Law School, and Senior Research Fellow, American Bar Foundation. She is the author of the guide to survey research in the FJC's *Reference Manual on Scientific Evidence*.

1. The Federal Judicial Center's REFERENCE MANUAL ON SCIENTIFIC EVIDENCE (2d ed. 2000) has already made a strong contribution in this area.

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

I. PROVIDING ADDITIONAL SIGNALS FOR THE USER

Reagan's example describing the statistical analyses that may flow from a case of alleged age discrimination contains a great deal of information. That coverage has both good and bad implications—good because it addresses the variety of statistical tests that the court may encounter,[3] and bad because, as a teaching device, it may discourage a learner who can easily handle some, but not all, of the information in a single dose.

One way to improve the usefulness of the presentation is to state explicitly that the statistical tests described (e.g., Fisher's exact test, chi-squared test, bivariate logistic regression) are alternatives that will generally produce the same outcome, but also to acknowledge that the tests can produce different results and to explain when and why they do. The contingencies are not obvious. The results of a Fisher's exact test and a chi-squared test will rarely differ. In contrast, a bivariate logistic regression may produce a result that differs from the one obtained by the other two tests when the relationship between age and probability of termination is not linear. The difference results from the fact that the regression can consider the full distribution of age values, whereas the other two tests analyze age in fewer categories (e.g., as 40-or-above versus below-40). The lesson that the user should take away is not to be daunted by the variety of available tests, but to be aware of their similarities and differences.

A second suggestion is simpler to implement. A great attraction of the software product is that it can provide an accessible glossary to the user by making words into hypertext so that the user can simply click to get a definition. In some situations, however, words that are terms of art may also have everyday meanings that will make users believe that they recognize them. In the age discrimination prototype, judges may not automatically take advantage of the hypertext option to seek a definition of, for example, "fixed" or "marginal."[4] This problem may be easily remedied by inserting a warning and an example at the beginning of the presentation, alerting users to the fact that some familiar words can have special meanings and that the hypertext signals those instances.

II. INTRODUCING ALTERNATIVE AND RESPONSIVE EXAMPLES

For the past seven years I have taught judges and attorneys about survey research and design at Duke's "Judging Science" program and at sessions on expert testimony sponsored by the American Law Institute-American Bar Association. I have also watched colleagues teach the same audiences about other forms of scientific evidence. In general, I have found that these sophisticated students are ea-

---

3. *See id*. at 284–91.

4. *See id*. at 285. A similar problem arises in jury instructions when the word used in the instruction such as "aggravating" has a variety of meanings, only one of which is applicable to the case at hand. *See* Shari S. Diamond & Judith N. Levi, *Improving Decisions on Death by Revising and Testing Jury Instructions*, 79 JUDICATURE 224, 232 (1996).

ger to learn but are sometimes mystified by the unfamiliar terrain of statistical reasoning. Although judges routinely make judgments based on probabilistic evidence, the evidence is not usually presented in numerical form. Thus, statistical evidence is particularly challenging.

One strategy for clear communication is to provide concrete examples of how the conclusion would change if the data were slightly different. The software delivery of the information proposed for the Federal Judicial Center's prototype makes such an approach quite feasible. Moreover, a software delivery system could even permit the user to modify the data and see how the modification affects the conclusions. In the example used by Reagan, 23 out of 228 employees, or 10.1%, under 40 years of age were terminated, while 56 out of 251, or 22.3%, of those employees 40 years and older were terminated.[5] An alternative analysis could show that if the company had been much smaller, the percentages of those terminated in each group could have remained the same (i.e., 10% and 22%) and yet the difference would not have been statistically significant. Such a modification would reveal the importance of sample size for producing statistically reliable results.[6] Other examples could show that a smaller percentage difference in a company with the same number of employees as in the original example would still provide a stable inference of difference. Comparing the two modifications would give some sense of the relative impact of sample size and effect size. A fully interactive module could permit the user to modify both the number of employees and the percent difference in termination rate.

Similarly, a modified version of the defendant's response to the plaintiff's prima facie case could show how to interpret the results if the age of the employee remained a significant predictor of termination after statistically controlling for years of service. Thus, the defendant's response might partially, but not fully, account for the plaintiff's showing of an increased likelihood of termination associated with employee age. The fact that these prototypes are to be presented in a software product opens the door to a rich set of possibilities without overburdening the user, who would only see the alternative version in response to a probe, such as: "Would you like to see what would happen if . . . ?" In this way, both broad coverage and a concise, uncluttered presentation can be achieved.

Presenting related versions of an example can help convey more clearly the attributes that matter in any example. Assume that a plaintiff brings a case alleging trademark infringement under the Lanham Act.[7] The relevant legal question is whether particular attributes of the alleged infringer's product (e.g., name or packaging) are likely to cause confusion about who manufactures the product. The plaintiff claims that the name and packaging of the defendant's car polish, FINISH 2001, will lead consumers to be confused, causing them to think that it is produced

5. Reagan, *supra* note 2, at 282–83.

6. Of course, in each of these instances, the analysis is being done on the entire population, raising some question about the appropriateness of *any* significance testing.

7. Lanham Act § 43(a), 15 U.S.C. § 1125(a) (1994 & Supp. 1999).

by the same company that manufactures the plaintiff's product, NU FINISH.[8] Each consumer in a survey is shown a bottle of NU FINISH. The consumer then views a display of products as they might appear in the marketplace and is asked to indicate which, if any, is produced by the same company that manufactures the first product. The display includes FINISH 2001. The rate of confusion is measured by the percentage of respondents who indicate that they think that FINISH 2001 is produced by the same company that produces the first product, NU FINISH. Assume that 25% of the survey respondents indicate that FINISH 2001 is manufactured by the same company that puts out NU FINISH. A one-in-four rate of confusion is generally enough to lead a federal court to find a likelihood of confusion.[9] However, the 25% may have included some amount of guessing or other indicator of confusion that was not caused specifically by the characteristics (i.e., name and packaging) of FINISH 2001.

The solution is found in the addition of a control group—and both parties in the case on which this example is based did conduct surveys that included control groups.[10] The difference in the controls they selected shows how alternative methods can affect the results. The defendant's expert chose another new competing product, PRISM.[11] Recall that the consumers in the survey were asked to indicate which products, if any, were put out by the same company that produces the product that they had just been shown (i.e., NU FINISH, the plaintiff's product). For respondents in the test group, the display included the defendant's product, FINISH 2001; for respondents in the control group, the control product, PRISM, replaced the FINISH 2001.[12] By comparing the selection rate in the control group for PRISM with the selection rate in the test group for FINISH 2001, the survey could test how much of the apparent confusion was attributable to the particular characteristics of the FINISH 2001 name and appearance, and how much was due to guessing or other forms of noise or confusion in the market generally.[13] In fact, the results of the survey showed that the selection rate for the PRISM product was no different from the selection rate for the FINISH 2001 product—indeed, it slightly exceeded it—and the court found that no likelihood of confusion had been shown.[14]

The value of a control group for permitting a clear test of a particular potential

---

8. This example is based, with some modifications, on *Reed-Union Corp. v. Turtle Wax, Inc.*, 869 F. Supp. 1304 (N.D. Ill. 1994), *aff'd*, 77 F.3d 909 (7th Cir. 1996).

9. *See, e.g.,* Exxon Corp. v. Texas Motor Exch. of Houston, Inc., 628 F.2d 500, 507 (5th Cir. 1980) (15%, 23% "strong evidence of likelihood of confusion"); James Burrough Ltd. v. Sign of Beefeater, Inc., 540 F.2d 266, 279 (7th Cir. 1976) (15% sufficiently "evidences a likelihood of confusion, deception, or mistake"); RJR Foods, Inc. v. White Rock Corp., No. 77 CIV 2329, 1978 WL 21389 (S.D.N.Y. Sept. 29, 1978), *aff'd*, 603 F.2d 1058, 1061 (2d Cir. 1979) (15-20% "reliable").

10. *See* 869 F. Supp at 1311–12.

11. *See id.* at 1312.

12. *See id.* at 1311.

13. *See id.* "Noise" is defined in *Reed-Union Corp.* as the "confusion by the consumer not caused by NU FINISH/FINISH 2001 similarity." *Id.* at 1311.

14. *See id.* at 1311–12.

cause depends on two important factors. First, the respondents must be randomly assigned to the test and control groups so that they are equivalent apart from, in this example, whether they are shown the display with the FINISH 2001 product or the one with the PRISM product. Second, the test and control stimuli must be as similar as possible in all ways except the characteristic that is being evaluated. The plaintiff's survey in the actual case shows how the intrusion of an extraneous difference can threaten the value of the control.[15] The plaintiff's expert used a product as a control that had the maker's name prominently displayed on the package (ARMOR ALL).[16] Neither FINISH 2001 nor NU FINISH (nor PRISM) display the manufacturer's name except in small print on the back of the bottle.[17] Displaying a major manufacturer's name on the bottle of the control product cued respondents in the control group as to the manufacturer's name.[18] Since all of the other ARMOR ALL products from the display also had that name prominently displayed on the front of the container, respondents did not link the NU FINISH with the ARMOR ALL product. The control group thus produced a *de minimis* noise level.[19] The court properly rejected this alternative control group design and found no evidence of likelihood of confusion.[20]

The value of discussing these alternative approaches is that they reveal: (1) the value of a control group, and (2) a comparison of approaches that can help the user understand how and why not all control groups are equal. The age discrimination problem described in the prototype offers possibilities for the same kind of variation. For example, the choice of years of service is a well-tailored instance of the kind of variable a defendant might offer to successfully rebut the plaintiff's prima facie case of age discrimination. But it is worth recognizing other possibilities. Suppose, for example, that the defendant can show that performance evaluations, which are both less reliable and potentially less valid measures, may account, but only partially, for the statistical relationship between age and termination rates. This example would muddy the waters considerably because the plaintiff could argue that the performance evaluations themselves reflect employer discrimination, and the defendant could argue that the evaluations are valid indicators of performance but are not sensitive enough to capture all of the variations in employee performance that led to the termination decisions. It is tempting to offer only an uncomplicated example in a prototype. Nonetheless, it may be unwise to leave the impression that judges can expect data in employment discrimination cases to be quite as neat as the current version of the hypothetical prototype may suggest. Again, the software user can control how much complication is intro-

15. *Id.* at 1311.

16. *Id.*

17. *Id.*

18. *Id.*

19. *Id.* That is, few respondents concluded that the ARMOR ALL product was made by the maker of NU FINISH.

20. *Id.* at 1311–12.

duced into the lesson. For example, a probe might enable the user to select among alternative variables that a defendant might attempt to use, and each of those variables could have different characteristics.

As the prototype stands, it will make a strong contribution to education on a complex topic. However, it can do more. While there are always opportunities to improve any product, the ones suggested here invite those writing the prototype to take advantage of the technology they propose to use in order to create an instrument that neither demands too much of the user nor oversimplifies the issues.

COMMENT

William B. Fairley*

CITATION: William B. Fairley, Comment on the Age Discrimination Example, 42 Jurimetrics J. 321–326 (2002).

The Federal Judicial Center's teaching example[1] on age discrimination lucidly exposits an oft-seen sequence of argument in a discrimination case. Call it the "standard paradigm." The standard paradigm has, however, to my mind, flaws in the logic of inference. At the least, the standard paradigm exemplifies a particular point of view on inference. For this reason, I would like to see it presented to judges accompanied by substantial discussion. The existence of different points of view on inference poses a pedagogical problem. The problem is acute for judges, for they, of all people, should not be prejudiced in advance as to proper forms of argument.

The first problem with the standard paradigm is that the probability model underlying the plaintiff's prima facie case is not remotely plausible. According to the FJC example,

> A statistical test called *Fisher's exact test* shows that if 79 employees out of 479 *employees were selected at random for termination* . . . the probability that the rates of termination for older and younger employees would be at least as different as observed in these data would be approximately three in ten thousand.[2]

Employees are not terminated "at random." A deliberative process characterizes all but the bizarre situation. The usefulness of a random selection model would have to be justified on grounds other than realism. Yet, realism is important in the discrimination case. Assume the usual random selection process with fixed mar-

---

* William B. Fairley is President of Analysis and Inference, Inc., a research and consulting firm with professional skills in statistics, economics, and finance. He is coeditor of *Statistics and Public Policy* (1977).

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

2. Reagan, *supra* note 1, at 285 (emphasis added).

ginal totals in the two-by-two table of counts of termination versus age, as in the example. The inference to be drawn, then, from a finding of a statistically significant difference in termination rates between older and younger employees according to Fisher's exact test is that either (1) there was a difference in the rates at which the two groups were selected, or (2) the selection did not proceed in the manner assumed. Without additional information or assumptions, it is impossible to tell which of the two explanations to credit.

The second problem with the standard paradigm is that, even if a random selection process is assumed, the particular model of probability that is assumed to govern the random selection is in practice either never stated or not validated.[3] This problem is illustrated by the concluding sentence in "Screen 2: Plaintiffs' Facts": "Statistical analysis shows that a disparity as large as this (or larger) would have *a probability of occurring by chance* alone approximately equal to three one-hundredths of one percent."[4] Of course, this is a conclusory statement that might be made in an opening statement in court. Nonetheless, the random sampling without replacement model implicitly assumed is not discussed in Section 2 and commonly would not be discussed in an actual case.

It is typical for "the probability" or "the chance" to be advanced—or understood by the reader—as "the" answer. That is, the calculated *p*-value is taken to show how likely it would be to see the disparity, or a greater one, if selection rates were the same for the older and the younger groups. But, of course, "the" probability or chance depends totally on the probability model assumed and can be arbitrarily different depending on the chosen model.

An example from my consulting experience, not with employment discrimination, but with racial discrimination in the provision of municipal hospital services, is *Bryan v. Koch*.[5] Plaintiff calculated infinitesimal chi-square probabilities from differences in the numbers of hospital beds closed down by the New York City Health and Hospitals Corporation in hospitals located in predominately minority areas versus other areas of New York. The probabilities were not applicable, however, because the decision to close "beds" was made by closing only a few "hospitals." The unit of decision was the hospital, not the bed. Once a hospital was closed, all the beds in it were closed. Instead of a sample of hundreds of beds, there was a sample on the order of 10 hospitals. Using a sample of hospitals, instead of a sample of beds, the difference in the rates at which hospitals were closed in predominately minority areas versus other areas did not approach significance.[6]

Concern with the validity of the model is critical. For example, contrary to the implicit or explicit model of independent selection for all individuals, it is com-

3. The fifth problem, discussed below, is that, whether or not the model is validated, it is not compared to equally or more plausible alternative models.

4. Reagan, *supra* note 1, at 283 (emphasis added).

5. 492 F. Supp. 212 (S.D.N.Y. 1980). For a detailed discussion of *Bryan*, see Thomas J. Sugrue & William B. Fairley, *A Case of Unexamined Assumptions: The Use and Misuse of the Statistical Analysis of Castaneda/Hazelwood in Discrimination Litigation*, 24 B.C. L. REV. 925, 953–55 (1983) (discussing *Bryan*).

6. Sugrue & Fairley, *supra* note 5, at 953–55.

mon for decisions to be made about entire groups of employees, such as those in certain departments, or those with certain skills. Thus, the independence of decisions between one employee and the other is not valid, and probability calculations made under the assumption that they are can be off by wide margins.

The third problem with the standard paradigm in practice is that the error created by ignoring the validity of the model almost always favors plaintiffs over defendants. The model almost universally assumed is one of independent or near-independent selections. Such a model is the one most likely to generate a statistically significant difference between two groups. It is anything but a "conservative" choice.

It might be argued that this direction of "error" is a feature of the weak burden of proof required for a prima facie case of discrimination. Perhaps the requirement of a statistically significant finding using *pro forma*, unrealistic probability assumptions works out to the low threshold of proof that courts would settle on in any case. However, unsupported models should not be advanced or swallowed uncritically. Different model failures in different situations can affect calculated significance levels differently. Unless the model is realistically appraised, there is no way of measuring how significant it is.

I also set aside the intrinsic difficulties in equating a *p*-value to the strength of the evidence.[7] There are a number of other difficulties intrinsic to the classical significance testing framework.[8]

The fourth problem with the standard paradigm in practice is that it invites confusion of causation and association. The text in the example states that:

> [I]f 79 out of 479 employees (16.5%) are terminated, we would *expect* approximately 16.5% of the employees under 40 years of age to be terminated, which equals 37.6 employees, *if age were not a factor in termination*. Because only 23 of the employees under 40 years old were terminated, they were terminated at a rate below expectation.[9]

The implication is that, because the observed number of younger employees terminated (23) is less than the "expected" number (37.6), age *is* a factor in termination. But this conclusion clearly goes beyond the statistical analysis. While an association between age and termination is plain, no direct evidence on the causal relationship of the two variables has been provided. The only conclusion supported by the statistical analysis is that if the observed rate of selection of younger employees is not equal to the expected number, then either: (1) there is a difference in the underlying true rates of selection of younger and older, or (2) the model is wrong. Without additional information or assumptions, there is no way to make a choice between the two possibilities.

At least part of the problem is semantic. "Expect," as used in the quotation, is a loaded way to describe a model-implied average number. The term leads the

7. *See* James O. Berger & Thomas Selke, *Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence*, 82 J. AM. STAT. ASS'N 112 (1987).

8. Kadane and Mitchell provide a list of these from a Bayesian perspective. *See* Joseph B. Kadane & Caroline Mitchell, *Statistics in Proof of Employment Discrimination Cases*, *in* LEGACIES OF THE 1964 CIVIL RIGHTS ACT 241 (Bernard Grofman ed., 2000).

9. Reagan, *supra* note 1, at 285 (emphasis added).

reader to hear "expected value" as synonymous with "a value that should be seen, and if it is not, then age must be a factor." The phrase, "if age were not a factor in termination," as used in the quotation, is neither a neutral nor technically precise substitute for, "if the rates of selection from the two groups were equal."[10] The former phrase strongly suggests discrimination is occurring, while the latter limits itself to a mathematical statement about model parameters. Thus, the reader is encouraged to infer a causal conclusion (discrimination caused the disparity in termination rates), even though the statistical test or *p*-value says nothing about causality.

Other terms in statistics, most notably the term "significant," also have different meanings or connotations in everyday language. When a statistician uses these terms for a lay audience, alarms should go off. Is the term being used in a purely technical sense, or to make an assertion that does not follow from the statistical analysis? In scientific practice, causal claims must be supported with an accepted theory or additional information that go directly to the existence of a causal, as opposed to a mere associational, relationship. For example, Cochran and Snedecor write, in one of the most widely cited texts in the scientific literature that "[b]efore claiming that a significant difference is caused by the variable under study, it is the investigator's responsibility to produce evidence that disturbing variables of this type [such as length of service, relevant skills, etc.] could not have produced the difference."[11] Neither in science nor in law must every possible disturbing variable be ruled out.[12] Nevertheless, a prima facie case of discrimination often can be established without any evidence that disturbing variables are not responsible for the difference. In science, you can't get to first base with a completely naïve statistic. In law, you can.

Again, what the law and scientific practice should require are not necessarily the same. However, much is lost by not at least acknowledging the protection that scientific practice provides. The basic distinction between causality and association constantly trips up the lay person—the judge and almost all other participants in a trial. Requiring explicit attention to the distinction from the beginning, though possibly through relaxing the extent of proof required, would seem to be at least as important in the legal as in the scientific arena.

There is another reason to avoid any but the most literal and carefully qualified use of the term "expected." The quarrel with the use of "expected" in the age discrimination example was not that the actual numerical answer for the expected value was wrong, but rather that the term leads lay readers down the garden path of causal conclusions unsupported by the analysis. There are also cases where what is called "expected" is the wrong answer. In *Sanderson v. City of New York*,[13] the "expected" number of promotions of older officers to the highest police posts was calculated by plaintiff's expert statistician. She noted that it was greater than the

10. *Id.*

11. GEORGE W. SNEDECOR & WILLIAM G. COCHRAN, STATISTICAL METHODS 127 (7th ed. 1980).

12. *See* Bazemore v. Friday, 478 U.S. 385 (1986).

13. No. 96-Civ. 3368 (LLS), 1998 WL 187834 (S.D.N.Y. Apr. 21, 1998).

actual number of such promotions. This "fact" was the basis for a claim that "age was a factor in promotions."[14] However, under a simple model positing two different levels of abilities among officers, if more able officers are more likely to be promoted at each stage of their careers than less able officers, then over time, the younger officers in the higher ranks are more likely be more able than older officers. This, then, could explain why older officers were not being promoted in proportion to their numbers (the "expected" proportion). Thus, the "expected" (in the technical sense) numbers were only "expected" (in an everyday sense) under an unrealistically simple model that did not take into account the basic factor of ability in making promotions. The expected values were only "correct" for a clearly erroneous model.

The fifth problem with the standard paradigm in practice is that it rarely countenances more than one model, and if it does, the alternative models are typically not compared. If the model assumed is arbitrary, so are any conclusions drawn from it. In the prototype, a plaintiff's logistic model relating rate of termination to age is examined. If this model is true, the simpler model of random sampling without replacement from a list of employees is true only as an approximation—and potentially a poor one. The more usual situation is where the plaintiff has only the random sampling model. The plaintiff has either not examined the more detailed model that relates rate of termination to age continuously, or having examined it, has not reported it. But, if rates of termination do not increase with age, can a claim of age discrimination be supported? Perhaps there are such cases, but they must be the exceptions.

The simple model of random sampling assumes that there are just two rates of termination. This assumption could be critical, but it may not be true. For example, suppose (unlike the FJC example) the average rate of termination for 40-and-over is higher than the average rate for under 40. Suppose further that the rate in the age range 40 to 59 is greater than that in the range 60-and-over. How could animus related to age produce such a pattern?

The sixth problem with the standard paradigm in practice is that it promotes a false sense of precision, especially with regard to the almost always false notion that very small probabilities can be supported by evidence. Values like 0.0001 or 0.0003 in the FJC example are calculated—often to even more decimal places—that suggest to the lay person that fine differences in probabilities can be detected by statistical analysis. As noted earlier, however, differences in plausible assumptions can swing "the probability" of a disparity by several orders of magnitude, so that the precision suggested by four decimal answers is spurious.

Tiny *p*-values are inherently suspect as answers to any real world probability question, if only because of the problem of "outrageous events." Mosteller and Wallace discuss the existence of "roguish" events that are not included in the sample probability model that yields a tiny probability.[15] The universal existence of what

14. Leona Lee, Report (Sept. 9, 1997) (on file with author).

15. FREDERICK MOSTELLER & DAVID L. WALLACE, APPLIED BAYESIAN AND CLASSICAL INFERENCE: THE CASE OF THE FEDERALIST PAPERS 90-91 (2d ed. 1984).

they call "outrageous events"—atypical events having small but not tiny probabilities—means that realistic answers are never tiny. Mosteller has given a tongue-in-cheek example of how he, Mosteller, could have defeated Muhammed Ali in the ring. How? Ali rushes out at the bell, steps on a shoelace, falls awkwardly to the mat, and knocks himself out cold.[16] The general proposition is that the probability of an event can never be less than the probability of some "outrageous" event that could produce it.

Tiny *p*-values exist only as logical deductions from very simple probability models that almost always fail to incorporate real world phenomena that would imply greater probabilities. However, tiny *p*-values deduced from simplified models are commonly produced and impress the unwary.

These problems with the FJC example are not with the careful exposition given but with standard practice in the field. Judges need to be exposed to critical commentary on common practice and to some of the debates on inference. Since time for in-service education is short, I question how much technique they can be taught beyond what they already know. More valuable than technique is exposure to the aims of data analysis and modeling. The use of examples may be an excellent vehicle for this exposure, but then the examples should inspire a critical examination of practice, including the logic of inference.

Another area for judicial education is in the subtleties of the relation of statistics to law. To provide just one example, Richard Lempert has written about the difference between judicial and statistical precedent:

> When a court is confused about competing statistical claims, the lure of apparently on-point precedent is attractive. But statistics are almost always case-specific, and what may make sense in the context of one case's statistics may not make sense in another statistical context . . . . If precedents were not statistical, courts would be well equipped—it is their stock in trade—to see how cases might be distinguished. But statistically untrained judges are poorly equipped to make distinctions regarding statistical precedent. Commentators need to help courts understand what statistical tests and conclusions mean in precedential cases and what this implies for the respect accorded such precedent in different future actions.[17]

To accomplish all of the aims of judicial education about statistics and to avoid the "a little knowledge is a dangerous thing" trap, judges, their clerks, and juries should increase their use of consultants. The role of consultants would *not* be to provide another opinion on which statistical analysis is right, but rather (among other contributions), to help the trier of fact understand what is being said. The plural in "consultants" is important. Different angles and points of view would seem to be an important safeguard against the undue influence of any one "expert" who has the ear of the magistrate.[18]

---

16. William B. Fairley, *Review of* Data for Decisions, *in* A STATISTICAL MODEL: FREDERICK MOSTELLER'S CONTRIBUTIONS TO STATISTICS, SCIENCE, AND PUBLIC POLICY 257 (Stephen E. Fienberg et al. eds., 1990).

17. Richard Lempert, *Befuddled Judges: Statistical Evidence in Title VII Cases*, *in* LEGACIES OF THE 1964 CIVIL RIGHTS ACT 278, 278 (Bernard Grofman ed., 2000).

18. *See* William B. Fairley & Frederick Mosteller, *Trial of an Adversary Hearing: Public Policy in Weather Modification*, 3 INT'L J. MATH EDUC. IN SCIENCE & TECH. 375, 375–83 (1972) (detailing one such scheme).

COMMENT

David A. Freedman*

CITATION: David A. Freedman, Comment on the Age Discrimination Example, 42 Jurimetrics J. 327–331 (2002).

The Federal Judicial Center is an institution that I greatly admire. However, its prototype module does not seem to be a promising vehicle for teaching statistics to judges. The format and organization are likely to be confusing, the notation will be hard to follow, and—most important—the statistical content is misleading in a variety of ways. Creating teaching materials for the legal profession is an excellent idea, but other directions are more promising.

I. FORMAT

Browsing hypertext is different from browsing in a book. Although some people will prefer hypertext, a book is likely to be a better way to communicate statistical material to nonstatistical readers. Furthermore, judges and lawyers will often try to master technical material only when they need it—in the middle of a trial. In that environment, clicking on links may be especially frustrating (like trying to learn a foreign language in a hurry, by reading a dictionary where every definition refers to other definitions).

* David A. Freedman is Professor of Statistics, University of California, Berkeley. He is a coauthor of *Statistics* (3d ed. 1998), a leading college textbook on the topic, and of the guide to statistics in the FJC's *Reference Manual on Scientific Evidence*. Tim Reagan kindly provided the data used in the hypothetical and a full description of the computer simulation used to generate the data. Statistical results reported here were computed from these data. Don Ylvisaker made helpful comments.

II. ORGANIZATION

The module includes material on tables, scatter diagrams, odds ratios, expected values, *p*-values, chi-squared tests, Fisher's exact test, and logistic regression.[1] This is all loosely held together within the context of a hypothetical, but there is little sense of orderly progression from one topic to the next. Creating materials that help judges and lawyers to read tables and scatter diagrams would be an accomplishment, and a reasonable goal for a few pages of text-based materials, which can start at the beginning and develop the ideas in logical sequence.

III. NOTATION

The notation is quite complex.[2] How is the audience to know that "ln" stands for the natural logarithm, or that natural logarithms differ from common logarithms? In the first place, what are logarithms? Formulas involving exponentials may be equally obscure: what is e, let alone e^x? Will judges know that mathematicians use "⇒" as short-hand for "implies"? The materials suggest that readers will find answers by clicking on hyperlinks[3]—not a plausible claim. Even the scatter diagram[4] is problematic: thirty years of teaching statistics convince me that people need a lot of help when they start to read statistical charts and tables.

IV. THE SUBSTANCE OF THE CASE

The focus of the analysis in the hypothetical seems to be a logistic regression model applied to an age discrimination case.[5] However, a critical point is given short shrift. Statistical models make assumptions; inferences are conditional on these assumptions: therefore, assumptions need to be explored. The module barely acknowledges this fundamental question, and the analysis therefore seems naive. I will provide some additional detail, at the expense of introducing further technicalities.

The materials specify a logistic regression model, the explanatory variables being age and years of service.[6] Why these two variables rather than two other variables? Why linearity of log odds in the chosen variables? What about independence?[7] Even if the logistic model is accepted (rather than, say, the probit), and the

---

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

2. *See id.* at 288–90 (Screen 4.3.2.1).

3. *Id.* at 282.

4. *Id.* at 291–92 (Screen 6.1).

5. *Id.* at 293–94.

6. *Id.* at 293.

7. The model assumes that termination decisions are made separately for each employee, without paying attention to the outcomes for other employees. If, for example, the company has two computer programmers and decides to terminate one, the odds that it will terminate the other are unaffected—according to the model. In this respect among others, the model is unrealistic. The independence assumption is used to fit the model and to test the significance of coefficients. If this assumption is wrong, the statistical tests are unreliable.

two relevant variables are age and years of service, conclusions may be dramatically wrong if the linearity assumption is wrong. By way of illustration, suppose that in truth, the company fires employees based on age only: the log odds of an employee being fired is

$$\log \frac{P}{1-P} = -2.2 + 0.1 \times \sqrt{\text{age}} \qquad (1)$$

The data look very similar to the data in the hypothetical. Suppose the statistician, being unaware of the company's termination policy displayed in equation (1), fits a logistic model as in the hypothetical—wrongly assuming linearity. Both coefficients in the (misspecified) model are insignificant: although the company is guilty, it will be exonerated by the statistical analysis. In other words, if the model is wrong—i.e., does not describe the mechanism by which the data are generated—inferences are unreliable.

Believers in logistic regression may well ask, why would log odds of termination depend on the square root of age, as in (1)? This is a good question, which only invites another question: why would the log odds be linear in age? It may also be asked, if the true relationship involves a square root, why fit a straight line? (This is an example of "specification error.") Generally, specification error is only to be expected, because statisticians seldom know the true relationships. Perhaps in recognition of this point, the data in the hypothetical were constructed so that the logistic regression model (Screen 6.2.1) is incorrectly specified.[8] The uncertainties created by specification error need to be acknowledged.

The chief statistical finding in the hypothetical[9] can be summarized as follows. The *t*-statistics for age and years of service are about 1.0 and 2.3, respectively: the first is insignificant, the second significant. However, even if the logistic model is accepted—if the two relevant variables are age and years of service, and if linearity is granted—the statistical findings are still unreliable. More specifically, the difference between the *t*-statistics is 1.3 with a standard error of about 1.9. Thus, the difference between the two explanatory variables, in terms of significance of impact on the firing decision, is well within the range of chance variation. No reliable conclusion can be drawn from these data, because the two ex-

8. In essence, data for the module were generated by Dr. Reagan, using a model where the probability of termination was 0.02 times years of service but did not depend on age—except that employees with less than one year of service were terminated with probability 0.50. This is almost a linear probability model; it is not a logistic model. The logistic model is robust, in that it found no dependence on age, despite the deliberate misspecification. With other assumptions about the true data-generation process, as in equation (1) here, the model will not be so robust. I used Dr. Reagan's data on age and years of service, but generated the termination data from equation (1) rather than his model. Equation (1) uses natural logarithms to base $e = 2.71828...$; and P is the probability of termination.

9. Reagan, *supra* note 1, at 294 (Screen 6.2.2).

planatory variables are so highly correlated.[10] In short, even on its own terms, the chief statistical argument in the module is a failure.

V. OTHER COMMENTS

The materials state that the hypothetical is meant to convey how a case might be litigated, not how it should be litigated.[11] This disclaimer is all to the good. But the legitimacy of the statistical arguments remains a central—and unexplored—issue. Screen 4.3 suggests that there are multiple "correct" ways to analyze data, with an implication that the logistic regression is one such analysis.[12] That unfortunate implication is reinforced elsewhere in the materials.

Screen 4.3.2 says that the "computer is used to derive an equation," suggesting an objective basis for the logistic regression.[13] No. The computer is told to assume the equation. All that the computer does is to estimate parameters and compute *p*-values, based on assumptions determined by the analyst. Screen 4.3.2.2 finally states one of the assumptions behind logistic regression,[14] but that is late in the day, and the statement is rather cryptic. The independence assumption is not discussed at all. In sum, the materials fail to make a critical distinction. You can fit a model by a rigorous computer algorithm and make rigorous statistical tests of hypotheses, but if the crucial assumptions behind the model cannot be defended, the "rigor" is only cosmetic.

Screen 7 states that "[t]he statistical evidence supports defendant's argument . . . ."[15] Surely the legal implications of the statistical analysis depend on the status of the underlying assumptions. If the model is reasonable, that is one thing; if the model is unreasonable, that is another. "Business necessity" arguments about the variables are discussed, but that is somewhat different from the statistical issues. The validity of the statistical model needs to be assessed by the trier of fact, in the light of the evidence presented by the parties; the materials provide little help for a judge who might have to ponder such an issue.[16]

Moreover, many details are wrong. For instance, Screen 4.3.1.2 claims that the chi-squared approximation is good when there are 40 or more observations.[17] However, if there is a cell with a very small expected value, the chi-squared

---

10. *Id.* at 292 (Screen 6.1, Figure 3).
11. *Id.* at 282.
12. *Id.* at 286.
13. *Id.* at 288.
14. *Id.* at 291.
15. *Id.* at 294.
16. Some observers may urge that logistic regression is probative because it is standard. This seems to be a confusion. Any model for age discrimination is likely to be viewed as high science in some circles and low comedy in others. It may also be urged that nothing is perfect, and modeling has to be better than not modeling. The imperfections of the world are easily seen; the consequent advantages of logistic regression are less clear.
17. *Id.* at 288.

approximation is likely to be poor, with 40 observations—or 400.[18] One more example will suffice. Screen 4.3.1.1 recommends the continuity correction, citing Howell.[19] This book is not authoritative, and it is widely known that the continuity correction may introduce a substantial distortion in the *p*-values computed from larger tables; specific results depend on the model being tested.[20] These are minor details. Why do they need to be addressed in the materials? If they are to be discussed, greater care is necessary.

To sum up, FJC teaching materials should help judges sift technical evidence, including the critical assumptions made by party experts. The materials under discussion do not help very much in this regard, because they focus more on technique than substance. Even at the technical level, there are numerous errors, both stylistic and substantive.

---

18. For example, the chi-squared approximation should not be used on the table

1	1
1	397

The null distribution of the test statistic (defined as for Table 2, Screen 4.3.1.1) puts most of its mass near 24 with the rest near 224, both values inconceivably remote from what is probable under the chi-squared distribution with one degree of freedom.

19. *Id.* at 287 (citing DAVID C. HOWELL, STATISTICAL METHODS FOR PSYCHOLOGY (4th ed. 1997)).

20. A correction for continuity is appropriate with one degree of freedom. *Cf.* GEORGE W. SNEDECOR & WILLIAM G. COCHRAN, STATISTICAL METHODS 126–27, 198 (8th ed. 1989). For an example in the other direction, suppose a die is rolled 60 times; we make a chi-squared test of the null hypothesis that the die is fair. The uncorrected chi-squared statistic follows the chi-squared distribution (with five degrees of freedom) almost perfectly in the critical range from 10 to 15, but the continuity correction (*see* SNEDECOR & COCHRAN *supra*) makes *p* too small by a factor of two or three. *See also* DAVID FREEDMAN ET AL., STATISTICS 525–31, A–36 (3d ed. 1998). The best advice might be to consult a statistician before using the continuity correction.

this page is blank in original

COMMENT

Joseph L. Gastwirth*

CITATION: Joseph L. Gastwirth, Comment on the Age Discrimination Example, 42 Jurimetrics J. 333–340 (2002).

This is an interesting example[1] designed both to assist judges in understanding statistical evidence and evaluating its relevance or "fit" to the issues involved.[2] The concepts and techniques discussed (e.g., the odds ratio, Fisher's exact test, and logistic regression) are quite important.[3] I am unsure, however, of the "fit" of the defendant's logistic analysis in the legal context. This is especially true given the portion of *Texas Department of Community Affairs v. Burdine*[4] that states that

---

*Joseph L. Gastwirth is Professor of Statistics and Economics, George Washington University, and currently Visiting Scientist at the Division of Cancer Epidemiology and Genetics, National Cancer Institute. He is the author of the textbook, *Statistical Reasoning in Law and Public Policy* (1988), and the editor of *Statistical Science in the Courtroom* (2000). He thanks Drs. Jay Lubin, Marc Rosenblum, Binbing Yu, and Gang Zheng for helpful discussions about the legal and statistical issues and Dr. T. Reagan for providing the underlying data.

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

2. "Fit" plays an important role in the gatekeeping task the Court assigned trial judges in *Daubert v. Merrell Dow Pharms., Inc*., 509 U.S. 579 (1993). For further discussion and references, see Marc Rosenblum, *On the Evolution of Analytical Proof, Statistics, and the Use of Experts in EEO Litigation, in* STATISTICAL SCIENCE IN THE COURTROOM 161 (Joseph L. Gastwirth ed., 2000), and Symposium, *At the Daubert Gate: Managing and Measuring Expertise in an Age of Science, Specialization and Speculation*, 57 WASH. & LEE L. REV. 901 (2000). *But see* D.H. Kaye, *The Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies*, 87 VA. L. REV. 1933, 1961 (2001) (arguing that "[a]s a logical matter, however, the fit requirement is superfluous").

3. Reagan, *supra* note 1, at 284–85, 288.

4. 450 U.S. 248 (1981); *see also* Reeves v. Sanderson Plumbing Prods., 530 U.S. 133 (2000). I am grateful to Dr. Marc Rosenblum for pointing out the importance of *Reeves*.

the plaintiff should have a full and fair opportunity to show that the defendant's explanation is a pretext. No discussion of the "pretext" stage is given; Section I demonstrates why this is a serious flaw in the current version. Section II identifies statistical issues that deserve more discussion, such as why the odds ratio is appropriate for layoff cases but probably not for typical hiring cases.[5] Section III offers some suggestions on how to improve the teaching utility of the prototype.

I. THE AGE DISCRIMINATION EXAMPLE

The plaintiffs, whose specific ages are not given in a list, were terminated after a merger and reorganization.[6] The statistical data reported in Table 1[7] of the example are analyzed using Fisher's exact test, which considers all employees over 40-years-old as having the same probability of being terminated and tests whether this probability is the same as that of employees younger than 40. Because the test is a conditional one in that it uses the fact that 79 individuals were terminated, it is an appropriate procedure. However, it is less informative than a test that is directed at an increasing trend in firing due to age.[8]

Table 1: The Number and Percentage of Employees Fired by Age Category

Age Category	Under 40	40 to 50	50 to 60	60 and over
Retained	205	110	51	34
Fired	23	23	15	18
Total	228	133	66	52
% Fired	10.1	17.3	22.7	34.6

Table 1 groups the data by increasing age; the probability of being fired seems to rise with age. We test the null hypothesis of *no age effect* against a linearly increasing trend in the probability of being fired with age using the Cochran-Armitage (CA) test.[9] The resulting p-value of 0.0000074 is quite a bit less than the

5. Integrating the software with the Federal Judicial Center's *Manual on Scientific Evidence* and appropriate statistical texts would provide greater understanding.

6. Reagan, *supra* note 1, at 282.

7. *Id*. at 284.

8. Such a test is described in several statistics texts. *See* ALAN AGRESTI, CATEGORICAL DATA ANALYSIS 118–19 (1990); JOSEPH L. FLEISS, STATISTICAL METHODS FOR RATES AND PROPORTIONS 96–9 (1973); PETER SPRENT, DATA DRIVEN STATISTICAL METHODS 374–78 (1998). A trend test is suggested as an appropriate procedure for age cases in Joseph L. Gastwirth, *Statistical Evidence of Discrimination*, 160 J. ROYAL STAT. SOC'Y 289 (1997). The data from an age discrimination case that settled prior to trial are also discussed in SPRENT, *supra*, at 378.

9. *See* authorities cited *supra* note 8. The test essentially correlates the difference between the percentage of individuals in each group who were fired and the overall percentage with a trend of 1, 2, 3, and 4. These weights can be modified to account for other information. For instance, if there were written or oral comments that individuals should stop working at 50, then one could combine the 40 to 50 year olds with those under 40 using weights 1, 1, 2, and 3.

value of 0.0002 obtained from Fisher's test. Clearly, this analysis strengthens the plaintiff's prima facie case.

However, the employer contended that age was not a factor when it considered the need for the employees' services, their performance evaluations, and the manager's assessment of their potential loyalty to the new management.[10] The defendant also asserted that senior employees were less likely to be loyal to management, using seniority as a surrogate or proxy for loyalty.[11] The first part of defendant's statistical evidence consists of two plots that clearly show that seniority and age are strongly related and that, in each age group, retained employees usually had less seniority than terminated employees.[12]

The final analysis offered by the defendant is a logistic regression relating the probability of being fired to years of service and age.[13] Figure 4 is a useful plot showing that the coefficient on seniority is bigger than that for age, indicating that although increased age still appears to be related to one's probability of termination, seniority plays a more significant role.[14] Formal statistical tests show that seniority is statistically significant at the commonly accepted 0.05 level, while age is not significant at that level. The legal analysis provided indicates that the data do not support an action for age discrimination under the *disparate treatment* theory and leaves open the appropriateness of the use of seniority as a proxy for loyalty in *disparate impact* cases. Although the prototype's analysis stops here, it is instructive to consider arguments that a plaintiff might raise to show that the reasons offered by the employer are pretextual. While the defendant claimed to consider past job evaluations, the company's future need for specific skills, and loyalty, the "explanation" only incorporated the rather "subjective" factor of loyalty. The company did not conduct an interview or use a previously validated loyalty test.[15] Rather, the logistic regression *only* incorporates one factor that is assessed by a proxy variable that is highly correlated with age r = .77). The plaintiff demonstrates the effect of this high correlation between variables age and seniority by submitting a logistic equation that incorporates an interaction term[16] along with age and seniority. The results given in Table 2 below indicate that although the model as a whole is highly predictive,[17] *none* of the individual predictors is significant, even though the interaction variable has the smallest *p*-value. This implies that it will be difficult to distinguish the effect of age from that of seniority, especially if they have a

10. Reagan, *supra* note 1, at 291.

11. *Id.*

12. *Id.* at 291–92.

13. *Id.* at 293.

14. *Id.* at 294.

15. I do not know whether such a test exists. If one is not available, then the defendant should provide evidence of this fact to justify using the proxy of seniority, which is so clearly correlated with age.

16. This adds a new variable, age times seniority, that looks for a joint effect of age and seniority.

17. The overall test is at the 0.001 level.

joint effect. This indicates that the factfinder needs to make sure that the "proxy" variable is not a "cover" for the legally protected characteristic (age).

Table 2: Results of a Logistic Analysis with an Interaction Term

Variable	Estimate	Standard Error	*p*-value
Constant	-2.54	.8364	.002
Age	.0058	.0202	.772
Seniority	.0038	.0970	.965
Interaction	.0013	.0016	.424

Plaintiffs might well question the use of seniority as a "proxy" for loyalty. Older employees are likely to be more loyal, in part because they have acquired firm-specific knowledge and also because their alternative job prospects may not be as bright as those of younger employees.[18] In particular, older employees do not turn over at the rate of younger ones, enabling the employer to save the cost of training new employees.[19] Thus, seniority, a factor usually positively related to pay or productivity, is now considered a substitute for a subjective assessment of "loyalty" to new management. The hypothetical indicates that the defendant claimed that the merger and reorganization inspired significant employee dissent; however, it presented no evidence indicating that more senior employees expressed greater unhappiness than junior ones.

Assuming that an unusual amount of employee dissent arose from the merger, finding out when it began becomes important. For example, suppose the acquiring company had a reputation for laying off senior workers in previous mergers; suppose also that the president of the firm had circulated an e-mail to middle managers to proceed with the same strategy. Furthermore, the e-mail described senior employees as "old geezers." As often happens in the modern world, someone forwarded the e-mail to an employee in the firm being acquired, perhaps to warn the recipient of what the future might portend. If the dissension arose *after* these events, it seems that the distinct likelihood of age discrimination by the defendant *created* the dissent itself. Therefore, employee unhappiness is not an appropriate variable for the defendant to offer as an explanation, much less to use seniority as a "proxy" for it. Of course, there may well be alternative scenarios more favorable to the defendant. The main point is that the example[20] fails to discuss one of the three stages in *disparate treatment* cases and simply accepts the employer's claim about dissent, without showing that the degree of dissent was related to an employee's length of service. Indeed, 5 of the 11 employees with less than one year of senior-

18. RICHARD A. POSNER, AGING AND OLD AGE 75 (1995).

19. In fact, in *Reeves v. Sanderson Plumbing Products*, 530 U.S. 133 (2000), the defendant had replaced the plaintiff three times because two of the replacements, all in their thirties, had left.

20. Reagan, *supra* note 1.

ity were fired. This percentage of 44.5 exceeds that of any age group in our Table 1.

Plaintiffs might well point out that the defendant did not incorporate the job evaluations into the model. This is important because studies reviewed have shown that older employees perform similarly to younger ones.[21] This may be the result of selection; for example, employers may weed out poor performers so the abilities of remaining older workers are higher than average. For our purpose, the failure of a party to include a clearly relevant variable may render the analysis inadmissible under *Daubert*.[22] Indeed, in *Diehl v. Xerox Corp*.,[23] the trial judge did not credit a statistical analysis that failed to incorporate a major variable. In that case, plaintiff's expert did not include performance histories or a skills assessment study in a regression analysis whereas defendant's analysis using them along with the variable of seniority indicated that older workers were *favored*.[24] Thus, it seems that the logistic analysis may place too much weight on the variables age and seniority, even assuming that the defendant's explanation truthfully describes what its managers did. If in reality the employer *did not consider job evaluations* or favored more senior workers as in *Diehl*, then Reeves allows the fact-finder to conclude that the reasons offered were pretextual.[25]

The logistic model considers age as a continuous variable. This is true biologically, but the law treats all employees under the age of 40 as one group in that the Age Discrimination in Employment Act[26] (ADEA) only protects employees at least 40 years of age.[27] Thus, plaintiffs might argue that even if only one explanatory variable was appropriate, so the issue of omitted factors does not arise, the logistic model is so inappropriate that it should be deemed inadmissible. This may be too harsh a result, as the data clearly indicate that age or seniority is related to the firings. The defendant's regression should be accepted into evidence, but given less weight than a model reflecting the true legal status of employees under 40. In *Mangold v. California Public Utilities Commission*,[28] the defendant objected to the plaintiff's expert correlating the performance on a *subjective* promotion exam with age and asked for a comparison of test-takers over 40 with those under 40.[29] The opinion notes that the "favored" individuals need only be "substantially younger" than the plaintiff and so need not be under forty.[30] Thus, a logistic regression, incorporating the data on the relevant factors that were available to the defen-

21. POSNER, *supra* note 18, at 75 n.17.

22. *See* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993).

23. 933 F. Supp. 1157 (W.D.N.Y. 1996).

24. *Id*. at 1162, 1165.

25. *See* Shannen v. Fireman's Fund Ins. Co., 156 F. Supp. 2d 279, 291 (S.D.N.Y. 2001) (citing cases allowing plaintiffs to establish pretext by demonstrating inconsistencies and contradictions in the defendant's reasons or factual errors in its statements).

26. 29 U.S.C. §§ 621-34 (1994 & Supp. V 1999).

27. 29 U.S.C. § 631.

28. 67 F.3d 1470 (9th Cir. 1995).

29. *Id*. at 1476.

30. *Id*.

dant, is likely to be more informative than a simple comparison of firing rates between those under and over forty. Additionally, checking of the fit of the data to the assumptions underlying the model is desirable.

The Cochran-Armitage trend test, however, can be extended to this situation by modeling the log-odds of being terminated as a function of an employee's age group. One would then use a trend test of the odds-ratios after incorporating the other legitimate factors.[31] Because neither party submitted the most objective information, that of the performance ability of employees, we cannot conduct a complete analysis. For exploratory purposes, we ran two regression models. In the first, which did not use seniority, the odds ratios, relative to the under forty group, were 1.864 for the 40-to-50 age group, 2.621 for the 50-to-60, and 4.719 for the sixty-plus group. When seniority was included, those odds ratios were reduced to 1.27, 1.537, and 1.57, respectively. A test for an increasing trend of these odds ratios is not statistically significant; however, there still appears to be an increasing trend with age. Thus, the ultimate decision may depend on the other circumstances surrounding the lay-off, the appropriateness of using seniority as a "proxy" for loyalty, or what the job evaluation data indicate. If the job evaluations did not correlate with seniority, they would likely diminish its importance in a complete logistic model.

II. OTHER STATISTICAL ISSUES

While the odds ratio is the appropriate measure to examine termination data,[32] it may not be appropriate for hiring or promotion data where the ratio of the minority success rate to that of the majority group is often used. In *Bew v. City of Chicago*,[33] 98.595% of the blacks passed an exam compared to about 99.9952% of the whites.[34] The ratio of the success rates is 0.9864, far exceeding the simple "four-fifths" rule that has been used as a screening device.[35] Common sense suggests that with such a high percentage of minorities passing the exam and the ratio of the success rates being so high, the exam had a minimal impact on the prospects of a black applicant.[36] However, the odds-ratio, being symmetric in pass and fail rates, equals 28.87. Thus, blacks had only *one-twenty-fifth* the odds of passing as whites. But to conclude the data in *Bew* are stronger evidence of discrimination than the

31. This technique is commonly used in epidemiological studies. *See* Norman E. Breslow et al., *Multiplicative Models and Cohort Analysis*, 83 J. AM. STAT. ASS'N 1, 5 (1983).

32. *See* Joseph L. Gastwirth & Samuel W. Greenhouse, *Biostatistical Concepts and Methods in the Legal Setting*, 14 STAT. IN MED. 1641, 1642 (1995) (recalling that the odds ratio is the correct parameter specifying the distribution of the number of members of the protected class who are laid off).

33. 979 F. Supp. 693 (N.D. Ill. 1997).

34. *Id.* at 696 n.6.

35. *Id.*

36. The sample sizes were quite large and the usual test of significance yielded a difference of –5 standard deviations, well above the usual criteria of two to three. *Id.* at 696. Thus, at the summary judgment stage, it was reasonable for the judge to decide that there was a material issue of fact.

odds ratio of 2.56 in the FJC example is not sensible. The ratio of retention rates can be misleading in termination cases,[37] and judges should realize that several measures of the difference between two proportions can be used.[38]

It may be helpful to readers with a legal background to inform them of a basic difference in the assumptions underlying Fisher's exact test and the usual method for comparing two proportions: the exact test conditions on the number of employees fired. The numbers of fired employees in each of the age categories must add to the total. This implies that they are statistically dependent. In the typical case that concerns passing an entrance or promotion exam, the passing score is announced before the test so that the number of people passing in one group does not affect the number in another. Using Fisher's exact test in that situation typically entails a loss of power relative to recently developed methods.[39]

Athough the prototype provides a good discussion of the logistic model and interpretation of the coefficients, some mention of the "goodness of fit" as well as its explanatory power would be useful. Goodness of fit is concerned with how well the model fits the data. For example, in Figure 1, it appears that the model underpredicts the probability of being fired for employees in the upper age range. For assessing the explanatory power of ordinary regression, one uses the proportion of variance explained (R^2 or adjusted R^2); however, the most appropriate measure for logistic and similar binary regression models is still an active research area.[40]

The inclusion of the chi-squared approximation to Fisher's exact test should be useful to judges, as it provides a relatively simple method to obtain a reliable result under certain conditions. Indeed, it would be helpful to mention still other approaches to analyzing the data. We have already provided one way and will simply observe that one could also have stratified the data in our Table 1 into seniority categories (e.g. by quintiles). The stratified version of the trend test could

---

37. *See* Gastwirth & Greenhouse, *supra* note 32, at 1642.

38. The defendant in *Bew* prevailed by demonstrating that the test was job-related and the cut-off score was reasonable. 252 F.3d at 891. The measure of impact is important as it may play a role in evaluating the evidence validating a test. A test with a large impact will likely need a greater degree of predictive validity than one with a small impact.

39. See Roger L. Berger and Dennis D. Boos, *P Values Maximized Over a Confidence Set for the Nuisance Parameter*, 89 J. AM. STAT. ASS'N 1012 (1994) for the analysis of a two-by-two table, and Boris Freidlin and Joseph L. Gastwirth, Unconditional Versions of Several Tests Commonly Used in the Analysis of Contingency Tables, 55 BIOMETRICS 264 (1999) for the extension to combination and trend tests.

40. *See* Edward L. Korn & Richard Simon, *Explained Residual Variation, Explained Risk and Goodness of Fit*, 45 AM. STATISTICIAN 201 (1991); J.G. Pigeon & Joseph F. Heyse, *An Improved Goodness of Fit Statistic for Probability Prediction Models*, 41 BIOMETRICAL J. 71 (1999); Efstathia Bura & Joseph L. Gastwirth, *The Binary Regression Quantile Plot: Assessing the Importance of Predictors in Binary Regression Visually*, 43 BIOMETRICAL J. 1 (2001). Basic texts that could be cited in teaching materials include P.W. HOSMER & STANLEY LEMESHOW, APPLIED LOGISTICAL REGRESSION (1989) and DAVID G. KLEINBAUM, LOGISTIC REGRESSION: A SELF-LEARNING APPROACH (1994). For more technical discussions, see PETER MCCULLAGH & JOHN A. NELDER, GENERALIZED LINEAR MODELS (1989) and ALAN AGRESTI, CATEGORICAL DATA ANALYSIS 84–96 (1990).

then be used to assess the role of age, assuming that seniority was a proper substitute for loyalty. It is important for the judiciary to appreciate that often there are several reasonable approaches that typically yield similar, but not identical, results.

III. SUGGESTED CHANGES

Although this comment has raised several concerns about the appropriateness of the analysis in the context of the prototype, the problems can be remedied by modifying the explanation offered by the defendant. As we saw, seniority reduces the apparent age effect to a nonstatistically significant one using two alternative approaches based on logistic regression. The issues then became (1) the appropriateness of using seniority as a proxy for loyalty, and (2) the lack of the most objective data—job evaluations—in the model. If the FJC replaced seniority by the average of the last two years of job evaluations, then the logistic analyses would be proper. This is especially true here, as the former management made those evaluations and, consequently, no "bias" in them can be attributed to the new management. Then the "pretext" phase might concern whether the new management really used the evaluations or "adjusted" them in some manner. These are factual, rather than statistical, issues.

By modifying the scenario, the example could then introduce some of the basic assumptions underlying the various statistical procedures. This should be germane to other uses of two-by-two tables and logistic regression as legal evidence. Many epidemiological studies are submitted as evidence in toxic tort and environmental cases, so a basic knowledge of how the analysis relates to the way the data were collected should be useful to judges.

Finally, as the discussion of *Bew*[41] reminds us, statistical significance depends on the sample size as well as magnitude of the difference. Hopefully, other parts of the instructional materials will discuss this as well as the important issue of the power, or ability to detect a true difference, of a statistical test.[42]

---

41. *See supra* notes 33–38 and accompanying text.

42. *See* MICHAEL O. FINKELSTEIN & BRUCE LEVIN, STATISTICS FOR LAWYERS 186–88 (1990); 1 JOSEPH L. GASTWIRTH, STATISTICAL REASONING IN LAW AND PUBLIC POLICY 180–84, 257–58 (1988); David H. Kaye & David A. Freedman, *Reference Guide in Statistics*, *in* REFERENCE MANUAL ON SCIENTIFIC EVIDENCE 85–177 (2d ed. 2000). In this comment, we adopted the 0.05 level commonly used in the social sciences to determine when a result is statistically significant. In contrast to those studies where the researcher often can decide on the sample size in the discrimination setting, the numbers of employees in the various protected groups have been determined by the employer's reaction to economic circumstances. To be fair to both parties, the power of the test used to detect a meaningful disparity should be considered in setting the cut-off level for significance. *See* RICHARD A. POSNER, FRONTIERS OF LEGAL THEORY 373–74 (2001) (noting that because there is no special legal significance in the 0.05 level and because the 0.05 convention is rooted in considerations unrelated to litigation, statistical evidence not reaching it should not be excluded).

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 statistical tools, but how they are encountered in litigation. I should, perhaps, qualify this praise by confess-

*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).

ing 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 precisely, for a

---

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 $p \times M$ of such cases.

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, important to recog-

---

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)

nize 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 authorizes

---

(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).

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

> [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.

---

termination rates between the two age groups at least as large as that observed would result by chance 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 16), about which the prototype is silent.

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

COMMENT

Marc Rosenblum*

CITATION: Marc Rosenblum, Comment on the Age Discrimination Example, 42 Jurimetrics J. 347–350 (2002).

The Federal Judicial Center (FJC) tutorial[1] is a logical extension of the *Reference Manual on Scientific Evidence*.[2] By introducing a specific, detailed scenario, the FJC aims to hit two birds with one stone: when and how to do statistical comparisons in an age discrimination context and, by extension, when and how to do similar comparisons in other discrimination scenarios. Federal judges must evaluate statistical assumptions and methods to satisfy the standards in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[3] and the type of material presented here is an excellent first step toward raising the level of trial court statistical literacy generally.

But a trial court's most important task with respect to statistical evidence (in my judgment) is less the mastery of formulae and more the understanding of what each litigant's expert is measuring. Slight changes in underlying assumptions can lead to crucial differences in the way the statistics are calculated and in the way that legal interpretation and precedent apply. This comment identifies information missing from the screens in the FJC prototype that a trial court should know. It shows that some of the findings presented in the scenario do not inevitably follow.

* Marc Rosenblum is Chief Economist, Equal Employment Opportunity Commission, and former Adjunct Professor of Law, Georgetown University Law Center. These comments do not necessarily reflect the views of any federal agency.

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

2. REFERENCE MANUAL ON SCIENTIFIC EVIDENCE (Federal Judicial Center ed., 2d ed. 2000).

3. 509 U.S. 579 (1993).

I. SCREENS 1 AND 3—BRIEF DESCRIPTION AND LAW

While the Supreme Court has not expressly addressed the applicability of disparate impact theory to the Age Discrimination in Employment Act (ADEA),[4] the courts of appeals are roughly split on this issue, a fact that might be mentioned. At least for judges in circuits that recognize disparate impact theory in ADEA cases, the possibility that certain real fact patterns could implicate both theories might be noted.

II. SCREEN 2—FACTS

Although ADEA protection "extends only to those individuals who are 40 years of age or older,"[5] the scenario fails to distinguish between the lower limit for statutory protection and the specific ages at which particular employers may decide to treat age as a proxy for declining productivity or ability to perform job tasks. The majority of lower courts accept analytical comparisons that reflect actual employment practices.[6] In many instances, employers may not take age into account before age 50 or 55, so that inclusion of persons between ages 40 and 49 as older and retained would bias the statistical comparisons to favor employers.

Odds ratios and other measures of the likelihood of termination as a function of age are mentioned later in the prototype (see Figure 1) and can accommodate age-banding (usually into five-year cohorts). Because counsel and experts incorporate such concepts into their statistical presentations and arguments, they should be integrated into the tutorial.

III. SCREEN 4—STATISTICAL ANALYSIS

These screens illustrate the distinction between providing judges with the key overview of why statistical significance tests are invaluable evidentiary tools and the minutiae of calculation. In particular, Screen 4.3.1 and Table 2 referring to chi-squared tests are superfluous and should be omitted. As the tutorial acknowledges, this test "only approximates the results of a Fisher's exact test."[7]

Several different issues arise with respect to Screen 4.3.2.1, the log regression, which may not be entirely correct in either its assumption of a dichotomous dependent variable or of linearity. First, the assumption of linearity required to

---

4. 29 U.S.C. §§ 621–634 (2000).

5. Reagan, *supra* note 1, at 283.

6. *O'Connor v. Consolidated Coin Caterers*, 517 U.S. 308, 312 (1996), instructs trial courts to focus on the relative age spread between terminated employees and their replacements, rather than to emphasize the age-40 statutory coverage level.

7. Reagan, *supra* note 1, at 288. Pre-computer analysis often relied on approximations of exact tests that were burdensome to calculate. The chi-squared test's popularity stems from that time period, and there are numerous ancient precedents for using chi-squared tests. However, it is fully appropriate for the FJC to advise judges that methodologically outdated precedents may be disregarded in favor of reliable statistical methods.

estimate log odds may not always be true. It may be useful to suggest that judges probe that point and suggest to them what type of analysis applies when the actual function is nonlinear (and, presumably, increasing). Second, even assuming the function to be linear, the use of actual age values below the statutory limit of 40 is incorrect. The better view is that all persons below age 40 carry a constant age value, i.e., 39, because the legal implication of various plus-40 distributions does not carry back to sub-40 distributions. This is seen clearly in Figure 1 and the related discussion concerning termination as a function of age. Properly framed, the bars should show all persons below age 40 in one cohort, and then each successive five-year cohort, until all observations are accounted for.

Thus, the calculated example showing the predicted probability of a thirty-two year old employee's termination is neither statistically significant or legally relevant, whereas the predicted probability of a 55-year-old employee's termination, also nonsignificant, is relevant. The employee cannot use statistics to infer an age-related termination.

IV. SCREEN 5—DEFENDANT'S FACTS

Defendant sets up a fact pattern to rebut plaintiff's prima facie case and applies in Screen 6 the assumptions that are presented in Screen 5. But the reader cannot reach Screen 6 without recognizing and agreeing with several assumptions that do not necessarily flow as inevitably as they are presented. First, this defense assumes that the prospective loyalty of incumbent employees to new management can be objectively quantified and measured and that such data have been obtained as part of the normal performance-appraisal process unrelated to subsequent downsizing. Second, the use of this information as an independent variable in the analysis of age-based terminations further assumes that future productivity and future loyalty are not auto-correlated and that both variables are something more than subjective age proxies.

The positive correlation between age and years of service does not, itself, affect the relationship between termination and age, any more than it suggests an inverse relationship between length of service and disloyalty (as posited by the defense scenario). The software prototype would be better served by including one or more objective performance or productivity variables and regressing them on an outcome variable.[8] In this case, prospective disloyalty is explained by age, because incumbent employees are presumably less inclined to be loyal to new management. In short, future loyalty by age is simply too subjective and unknown to be taken into statistical account in explaining terminations by age.

8. In fact, the approach taken in Screen 5 is reminiscent of the circularity rejected 25 years ago in *James v. Stockham Valves & Fittings Co.,* 559 F.2d 310 (5th Cir. 1977), where the defense argued that employee skill was explained by grade, and that employee grade, in turn, was explained by skill rankings (by race, where all higher-skilled and graded workers were white and all lower-skilled and graded workers were black).

V. SCREEN 6—STATISTICAL REBUTTAL

Figures 2, 3 and 4 should, if they are presented at all, appear in an appendix rather than the text. And, as indicated above, figure 4, as shown, is incorrect. Instead, the text could explain in more detail the difference between a dichotomous dependent variable (requiring the result to be expressed as a log) and a continuous dependent variable, which does not.

VI. SCREEN 7—LEGAL ANALYSIS

For the reasons covered above, I disagree in two principal respects with the software prototype's legal analysis. First, defendant's explanation does not statistically rebut plaintiff argument that age was a factor in the employee terminations. Second, in many circuits, plaintiff is not constrained from arguing that the terminations had a disparate impact on older employees, aside from (or in conjunction with) the claimed disparate treatment against certain employees on the basis of their age.

For the software prototype to have maximum value, the judges using it should not have concerns that the underlying assumptions and fact pattern are problematic. Only then can the judges best focus on the statistical applications.

COMMENT

Michael J. Saks*

CITATION: Michael J. Saks, Comment on the Age Discrimination Example, 42 Jurimetrics J. 351–355 (2002).

I. ON LAW AND STATISTICS

Raising the statistical literacy of judges, very few of whom came to their positions from backgrounds involving quantitative or scientific skills, is an extremely difficult undertaking. Law has been a field of the humanities, not the sciences.[1] Law students, and therefore the lawyers and judges they become, are people of the word, not of the number, and not of the conceptualizations that numbers, data, and empirical research involve. I have no doubt that one of the hardest things judges or lawyers are asked to understand during their careers is statistical reasoning.[2]

That law and our society would be better off if statistical literacy could be added to lawyers' intellectual repertoire is something that has been recognized at least since Oliver Wendell Holmes's famous address, *The Path of the Law*.[3] But the

* Michael J. Saks is Professor, Arizona State University College of Law and Department of Psychology. A psychologist, he has taught statistics courses for judges as well as undergraduates.

1. "[T]he intellectual life of the whole of western society is increasingly being split into two polar groups. . . . Literary intellectuals at one pole—at the other scientists. . . . Between the two a gulf of mutual incomprehension." C.P. SNOW, THE TWO CULTURES AND THE SCIENTIFIC REVOLUTION 4 (1959).

2. I say this as a person who taught a course involving moderate amounts of statistics, every other summer for a decade, to judges in the University of Virginia's LL.M. degree program.

3. Oliver Wendell Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 469 (1897) ("For the rational study of the law the black-letter man may be the man of the present, but the man of the future is the man of statistics and the master of economics.").

legal profession is not a great deal closer to that literacy now than when Holmes made his observations more than a century ago.

Our modern, data-based society depends heavily on empirical data and the statistical analyses to make sense of those data. People from a vast array of fields —from biomedicine to engineering to management to the social sciences to zoology—share a set of concepts, a language, and an epistemology that lawyers and judges do not comprehend. Ironically, those assigned major responsibility to resolve important issues that emerge from those other domains are barely able to understand the information on which a factual understanding of the issues often depends. This is not a healthy situation. Thus, the goals implicit in the Federal Judicial Center's creation of computer assisted statistical lessons for judges[4] will be difficult to achieve, but could hardly be more important.

Having said that, is this effort aimed at the right level? I think judges would benefit from understanding statistical reasoning at a conceptual level—to understand where and why the data are relevant to solving the problem at hand. Indeed, an understanding of research methodology, of the logic of drawing inferences from empirical evidence (the scientific method) would be even more valuable to judges.[5] I think judges would also benefit from learning how to interpret the results of statistical analysis, so that they can have a proper appreciation of what the numbers tell us about the phenomena in question. But it is not as clear that moving closer to the computational level or the underlying mathematics offers skills that are beneficial to judges.[6]

II. ON STRATEGIZING ABOUT TEACHING STATISTICS

The traditional way to teach statistics is to begin with mathematical foundations—theorems, proofs, formulas, distributions, and the like. After all, there is no point in asking people to trust what you are doing if you do not lay out all of the premises. Then the students can begin to solve problems; learn formulas, cookbook procedures, or computer programs; calculate statistics on a database; and interpret the results of the analyses.

I taught undergraduate statistics for a number of years and came to believe that, at least for that "educated public," the traditional approach was backwards. Asking students to learn a lot of underlying premises before they come close to seeing what the tools are to be used for seems meaningless and boring. But if one

4. *See* Robert Timothy Reagan, *Federal Judicial Center Statistical Examples Software Prototype: Age Discrimination Example*, 42 JURIMETRICS J. 281, 287 (2002).

5. The entire enterprise of inferential statistics is a bloated—though essential—footnote to empirical research, which answers essentially one question among many that arise in interpreting empirical evidence, namely, "how likely is it that the observed differences are the result of chance?" Descriptive statistics summarize large amounts of data.

6. But assuming the prototype under comment is the correct level at which to try to educate judges, the job has been done exceptionally well and is presented remarkably clearly. I would love to have such computer assisted materials available for a course in law and statistics for law students.

begins at the end—with a practical problem, showing how a useful answer can be found by analyzing and interpreting data—the students immediately see the value of the whole enterprise. With their interest stimulated, one can peel away the layers and delve into the details.

The prototype under review avoids the start-with-the-premises approach. It begins with a familiar problem in law for which empirical data and statistical analysis are relevant, and shows how the findings can be interpreted to help answer the factual question before the court. Nevertheless, it moves along unevenly—at times taking a very basic approach, other times using a rather mathematical and formula-based one, and occasionally offering some conceptual knowledge. It delves into details, and sometimes with a degree of particularity that I am uncertain will have much value for the judge-students.[7]

Perhaps the rationale is to mix levels of learning in an effort to keep the presentation more interesting. Perhaps some of these concerns would be diminished if this were one lesson among many, with suitable lessons preceding this one.

III. ON THE SPECIFIC PROTOTYPE

As I have noted, assuming that this is the correct level at which to approach the judge-students, the problem is an excellent one and is presented extraordinarily well. It provides a clear, complete, compact, comprehensible, and legally meaningful example. It also offers an opportunity to practice interpretation of 2×2 tables and graphs.

There are, however, some things that might bear clarifying and others that might benefit from elaboration or more emphasis. In explaining odds ratios and probabilities, it might help to make a more direct comparison between the two and to explain more about interpreting them. This could be done in a different lesson or made accessible in this one through a hyperlink. People who are not accustomed to working with both odds and probabilities can overlook some strange differences between them and be led to incorrect interpretations about what has been learned from them. Both probabilities and odds have a minimum of zero, but probabilities have a maximum of one while odds can be infinite. This potentially vast gap between probabilities and odds is not immediately apparent to most people. At the bottom of Table 1, the odds of being in one group versus another and the probability of being in the first group are close or identical. But at the top of the table, the odds and the probabilities diverge massively. Similarly, the distribution of probabilities is symmetrical (ranging from 0.02 at the bottom to 0.98 at the top), but the distribution of odds is extremely skewed (from 0.02 at the bottom to 50 at the top of the table).

7. I suspect that most judges will regard the level of detail as of little use, if my own efforts at computer-assisted statistical analysis demonstrations for my judge-students at Virginia are any guide.

Table 1. Comparison of Equivalent Probabilities and Odds

Size of membership in two groups	Probabilities	Odds
50 v. 1	50/51 = .98	50:1 = 50
20 v. 1	20/21 = .95	20:1 = 20
9 v. 1	9/10 = .90	9:1 = 9
5 v. 1	5/6 = .83	5:1 = 5
2 v. 1	2/3 = .67	2:1 = 2
1 v. 1	1/ 2 = .50	1:1 = 1
1 v. 2	1/3 = .33	1:2 = .50
1 v. 5	1/6 = .17	1:5 = .20
1 v. 9	1/10 = .10	1:9 = .11
1 v. 20	1/21 = .05	1:20 = .05
1 v. 50	1/51 = .02	1:50 = .02

Note: Probabilities range from 0 to 1. Odds range from 0 to ∞.

Some elaboration might also be useful in explaining how Fisher's exact probability is calculated, especially in light of the priority given that test.[8] That initial statistic is simply presented,[9] as a rabbit from a hat. However, the less important chi-squared test which follows it is explained by conceptual and computational formula.[10]

The first exposure to a conclusion from a statistical analysis might benefit from more emphasis, such as by stating the conclusion in different words and re-explaining it. The prototype states: "Because such a result would be so unlikely, this statistical analysis supports plaintiffs' prima facie case of their age being a factor in Premium's decision to terminate them."[11] In its context, this comes so quietly it may be too understated. This is the exclamation point of the analysis.

8. By priority, I mean both that it comes first in the presentation and that it is preferable to the chi-squared test whenever a computer is available.

9. *See* Reagan, *supra* note 4, at 285.

10. *See id*. at 286-88.

11. *Id*. at 286.

The judge-students need to have its importance underscored and to know what it is and why it is.

Occasionally, the prototype gives conclusory rules without the beginning of a reason for them.[12] In one way or another, explanation needs to be provided. Again, if doing so on the main screen seems too much of an aside, it could be done by hyperlink. And perhaps the rule-of-thumb as well as its explanation needs to be demoted to a hyperlink.

In an effort to familiarize judge-students with the way statisticians and users of statistical tests are likely to communicate—that is, in numbers—every time a probability is given I would provide the numbers along with the words. So, for example, in addition to saying, "the probability . . . is approximately one in ten thousand," I would add: $p = 0.0001$.

Things that are fundamental and yet opaque may need to be explained in the main screens, or in the main screens of earlier lessons. For example, the concepts of degrees of freedom and of one- versus two-tailed tests, are presented with no elaboration. Such things might be candidates for being more than hyperlinks.

A few sentences may not make any sense to the judge-students, no matter how many times they re-read them. A good example is: "The type of equation derived with logistic regression is one that expresses the *natural logarithm* of the *odds* on being in one group as opposed to the other as some number (the constant) plus or minus some other number (the *coefficient*) times the value for the independent variable."[13] The notion expressed in this sentence, like others, may be easily conveyed in a classroom, where an instructor can write an equation on the blackboard, explain its pieces, and answer students' questions. But transferring this process to an on-screen format presents a continuing challenge.[14]

12. An example is the rule for when the chi-squared test may be used as an approximation of the exact probabilities, which could be obtained from the Fisher's exact test. *Id.* at 286–87.

13. *Id.* at 288.

14. Such sentences could be hyperlinked for an explication of the whole sentence, with a video clip explaining the parts of the equation being described.

this page is blank in original

COMMENT

Sheldon L. Trubatch*

CITATION: Sheldon L. Trubatch, Comment on the Age Discrimination Example, 42 Jurimetrics J. 357–362 (2002).

Teaching statistics to judges is important for enhancing decision making in several legally important areas, and the general theme is important in this time of increasing disputes involving any number of technical issues. Therefore, although the teaching module developed by the Federal Judicial Center[1] is important in itself, it is even more important as an illustration of how to use the Internet to enable judges to educate themselves on technical matters. My comments address both aspects of this module but focus on the broader implications.

I. TRAINING FOR ADULTS

Lawyers and judges are trained to be lifelong learners. Not only must they keep abreast of changes in the law, but they also must acquire a basic understanding of scientific and technical developments. Nevertheless, there is generally an understandable difference between keeping up with legal developments and acquiring new information about scientific and technical matters. Law school established the foundation for adding new legal information to what is already known, but most lawyers and judges did not acquire a comparable foundation in scientific and technical matters during their education. Accordingly, most judges and law-

* Sheldon L. Trubatch is the President of HelperSoft, Inc. He also practices law in Washington D.C. He has been involved in conveying technical material to adults and students as a professor of physics, a lawyer for the government, a large corporation, and many clients in private practice.

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

yers are no different from any other educated layman when it comes to learning about scientific or technical matters.

Research has identified several aspects of effective adult education.[2] First, the teaching and learning processes should be interactive. In this case, the module is not interactive. However, interactivity could be realized by including a means to change some of the numerical assumptions to see how the results of the statistical analysis would vary. Second, the learner must be treated with dignity and respect. The tone of the FJC material meets this goal. Third, the material should address what the learner wants to know. To accomplish this, the material should be vetted with several judges, and their feedback should be incorporated. Fourth, the material should be made relevant to the learner by relating it to the learner's experience. In this case, relevance is established by relating the material to the issues presented for decision. Finally, the material must be valuable enough to the learner to support a change in the learner's attitude. In this case, the value of the material is clear if it will enable a judge to better support a legal decision.

II. TRAINING TAILORED FOR JUDGES

In discussing scientific and technical evidence in the courtroom, the Science, Technology and Law Panel of the National Research Council "noted that judges have little spare time for in-depth study of science or engineering."[3] Among the mechanisms discussed by the panel for helping courts to deal more effectively with scientific and technical evidence were short courses[4] and alternative delivery modes, including video, print, or online summaries.[5]

---

2. Diane Nowlan, *Principles of Adult Education as Related to Instructional Development*, *at* http://www.ucalgary.ca/UofC/faculties/EDUC/jdnowlan/adult/ html (2000); University of Nebraska Lincoln, *Principles of Adult Education E-2*, *at* http://4h.unl.edu/volun/arlen/principl.htm (n.d.).

3. NAT'L RESEARCH COUNCIL, SCIENCE, TECHNOLOGY, AND LAW PROGRAM, A CONVERGENCE OF SCIENCE AND LAW: A SUMMARY REPORT OF THE FIRST MEETING OF THE SCIENCE, TECHNOLOGY, AND LAW PANEL 7 (2001) [hereinafter cited as NRC PANEL].

4. *Id.* at 9. Short courses on science and technology are given at Duke University's Private Adjudication Center, the University of Virginia's Graduate Program for Judges, and the National Judicial College. Attendance is reported to be sparse "due to competing demands on judges' time" and limitations on court budgets. *Id.* at 9.

5. *Id.* As for alternative delivery modes, the Internet was recognized as providing opportunities for "learning on demand" or "just-in-time" learning. *Id.* Internet or e-learning, according to Jonathon Levy of Harvard Business School Publishing, supports a granular, dynamic learning model driven by the learner that is "just for you, just enough, and just in time." Donna Klinger, *A Guttenberg Moment*, BUS. OFFICER MAG., Jul. 2001, at 45. Practical limitations also were identified, including the resources needed to develop material and to ensure that the material is accurate, represents the consensus of the field, and is not biased. NRC Panel, *supra* note 3, at 9. These organizations include the New Mexico Judicial Education Center (*see* Program Descriptions, http://jec.unm.edu/ training/prog_conf.htm (last visited Apr. 8, 2002)), the Judicial Education Reference, Information, and Technical Transfer Project (a national clearinghouse for information on continuing judicial branch education; *see* JERITT, http://jeritt.msu.edu/default.asp (last visited Apr. 8, 2002)), and the National Association of State Judicial Educators (a clearinghouse for an emerging body of specialized judicial education materials and techniques, and a source of instruction; *see* NASJE Home, http://nasje.unm.edu (last visited Apr. 8, 2002)).

Several organizations have started to address the special needs of judicial education to make it effective. The National Center for State Courts[6] recommends a comprehensive judicial education process that incorporates "*appropriate adult education practices*,"[7] including needs assessments (systematic assessment and analysis of judges' learning needs), learning objectives (clear, concise written statement of objectives and skill levels), learning activities (promotion of learning by active participation in the process),[8] and conducive learning environment (appropriate instructional aids and support).[9]

Application of these adult education principles for judges to this prototype raises four issues. First, no indication is given as to whether a needs assessment was conducted. In this case, such an assessment would give judges an opportunity to provide important input about which specific aspects of applying statistics to employment discrimination cases give rise to the most difficulties. Such information would help the drafters to focus on specific topics that judges want to learn about.

Second, the learning objective for this prototype is implicit in the statement that "[t]his example demonstrates both how statistical evidence can support a plaintiff's prima facie case of employment discrimination and how additional statistical evidence can support the defendant's nondiscriminatory explanation for the apparently discriminatory pattern."[10] This description, however, does not do full justice to the scope of the learning objectives addressed by the material. These learning objectives include: (1) how facts about numbers can be treated by statistical methods; (2) the vocabulary of statistics; (3) the various alternative statistical methods available; (4) determination of the appropriateness of a particular statistical method for analyzing the facts; (5) the competing uses of statistics based on the same facts; and (6) the limited use of statistics as a tool (and not as the determinative basis) for a legal determination. It might be useful to state explicitly that all of these goals are addressed to some extent by the material and to indicate, without being overtly didactic, where each is being addressed.

Third, active participation is not encouraged in this prototype. Although many judges may not be numerate, they may be curious about how the statistical results would change with different numerical values for the facts. Changing the numeri-

---

6. The Center is developing self-paced interactive programs that merge advanced technological resources with proven adult learning principles. *See* National Center for State Courts, Institute for Court Management, *at* http://www.ncsconline.org/Education/index.html (Apr. 6, 2002).

7. National Association of State Judicial Educators, *Principles and Standards of Continuing Legal Education* 2 (1991), *available at* http://jeritt.msu.edu/pdf/Standardsforweb2.pdf (emphasis added).

8. *Id*. For example, in the training material on "How to Tell Good Science from Bad Science," described at National Center for State Courts, Institute for Court Management, *How to Tell Good Science from Bad Science*, http://www.ncsc.dni.us/ICM/distance/science.html (last visited Apr. 8, 2002), participants are invited to interact with the presenters.

9. National Center for State Courts, *supra* note 6.

10. Reagan, *supra* note 1, at 282.

cal facts should provide greater appreciation for the importance of numerical values to the statistical conclusions. Support for such active engagement with the numerical facts could be provided by incorporating an option to modify the numerical facts to see the changes in statistical results.[11]

Finally, although the physical aspects of the learning environment are not an issue for Internet-based training, the extent to which a participant can interact with an instructor or peers is important. Because peer interaction is often an important aspect of learning for adults, consideration should be given to providing for such interaction in the Internet presentation of the material. Such opportunities for interaction are routinely provided in online courses that satisfy requirements for continuing legal education.

III. TRAINING JUDGES ON QUANTITATIVE MATERIAL

There is a paucity of material for training judges, or even lawyers, on statistical concepts. By contrast, there is a substantial volume of material on statistics training for professionals who use statistics directly in their work.[12] It is likely that training material on statistics for judges has not been voluminously produced for two related reasons: most judges are lawyers, and most lawyers select the law as a career because they are neither mathematically nor technically inclined.

In the extreme, disinclination to mathematics may result in panic attacks called math anxiety.[13] Although most judges may not suffer from math anxiety, it may be assumed that a fair number of judges do not relish learning enough mathematics to understand and decide between competing mathematical explanations of job discrimination. Mathematical training material for judges may, therefore, be improved by incorporating some of the techniques that have been developed to overcome math anxiety. In particular, it may be useful to include in this material an introduction that urges judges to directly confront any negative emotional responses that may be triggered by previous experiences with learning mathematics, and to overcome those responses by focusing on the importance of learning enough about mathematics to decide a case competently.[14]

---

11. An example of interactivity that immediately reinforces the material learned is provided by some of the New Mexico Joint Education Center programs. For example, the "Introduction to Basic Hearsay Evidence" includes eight topics with twelve questions in the form of hypotheticals as exercises for the material presented. *See* New Mexico Judicial Education Center, *How to Use Hearsay Exercises*, http://jec.unm.edu/training/hearsay/about.htm (n.d.). This approach to demonstrating the ability to apply the material presented is no different from the criterion applied in teaching physics, which is, "if you cannot do the calculations, you have not learned the material."

12. *See* Internet Interface for Statistics Education, *at* http://acad.cgu.edu/wise/ (n.d.).

13. SHEILA TOBIAS, OVERCOMING MATH ANXIETY (1993).

14. *See* Sandra Manigault, *Coping with Math Anxiety*, *at* http://www.mathacademy.com/pr/minitext/anxiety (last visited Apr. 8, 2002).

IV. TRAINING MATERIAL MUST BE DESIGNED FOR THE INTERNET AS A MEDIUM

For training over the Internet to be effective, the training must be consistent with the special techniques of effective writing for the Internet. Studies show that Internet site visitors rarely read Internet pages word by word.[15] Visitors scan each page to pick out individual words and sentences. To be effective, Internet pages must employ concise text in a scannable format. Conciseness criteria include: inverted pyramid style, starting with the conclusion; half the word count (or less) of conventional writing; and one idea per paragraph. Text can be made scannable by liberally using structural elements such as: headings, bulleted lists, topic sentences, tables of contents, highlighted keywords (hypertext links, typeface variations and colors), large type, bold text, and informative subheadings.

Judges want early assurance that the training material meets their needs. A concise introduction on one screen provides that assurance. Moreover, material should be divided into small, self-contained units. A roadmap should show the relations among the units. This will enable judges to interactively personalize the material to meet their perceived needs. Related units should be hypertext-linked for ease of navigation, and graphics should be used only to support the text, not as decoration.

The training material is not structured to be read directly from the screen of a computer monitor. Separation of the written text into "bite-sized" chunks called "screens" is not sufficient. Conciseness and scannability require substantial changes to the written words.

Application of the web-design principles of Part IV suggests that Screen 3, "The Law," should be separated into screens on such subtopics as (1) Age Discrimination in Employment Act, (2) burden shifting method of proof, and (3) applicability of the burden shifting methodology.[16]

15. John Morkes & Jakob Nielsen, *Concise, SCANNABLE, and Objective, How to Write for the Web* (1997), *at* www.useit.com/papers/webwriting/writing.html.

16. *See* Reagan, *supra* note 1, at 283. Examples of these screens are given in the appendix.

APPENDIX

New Screen 1

Age Discrimination in Employment Act (ADEA), 29 U.S.C. § 621 et seq. (1994 & Supp. 1999).

Forbids Age-Based Discrimination

- Hiring
- Discharging
- Compensation
- Employment terms, conditions, privileges

Protects only individuals forty years of age or older

Plaintiffs must prove age-based discrimination
(underlining indicates hypertext link to New Screen 2)

New Screen 2

Three Step Burden Shifting Method of Proof

Plaintiff: establish prima facie evidence of discriminatory employment action by defendant
O'Connor v. Consolidated Coin Caterers, 517 U.S. 308, 311–12 (1996).

Defendant: burden to show legitimate, nondiscriminatory reason for employment action against plaintiff
Id. at 311

Plaintiff: burden to prove employment action really related to plaintiff's age
St. Mary's Honor Center v. Hicks, 509 U.S. 502, 507–08 (1993).

New Screen 3

Applicability of Burden Shifting Method of Proof

Developed under Title VII of Civil Rights Act of 1964, 42 U.S.C. §2000e *et seq.* (1994 & Supp 1999). *See* McDonnell Douglas v. Green, 411 U.S. 792 (1973).

Applicability to ADEA cases never explicitly decided by Supreme Court

Applicability to ADEA cases assumed by Supreme Court. *Reeves v. Sanderson*, 530 U.S. 133, 142 (2000); *cf. O'Connor*, 517 U.S. at 311.

REPLY

Robert Timothy Reagan *

CITATION: Robert Timothy Reagan, Reply to Comments on the Age Discrimination Example, 42 Jurimetrics J. 363–372 (2002).

Eleven distinguished experts on law and statistics have graciously provided valuable feedback on our computer software prototype to teach judges statistics using legal examples.[1] The commentators have offered an informative range of thoughtful reviews: "failure,"[2] "interesting example,"[3] "admirable effort,"[4] "on the whole, [the materials] seem to work quite well,"[5] "excellent first step,"[6] "fine educational effort,"[7] "lucid[ ] exposit[ion],"[8] "tour-de-force explanation of the logistic

---

* Robert Timothy Reagan is Senior Research Associate, Federal Judicial Center. The views expressed in this reply are those of the author and not necessarily of the Federal Judicial Center.

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

2. David A. Freedman, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 327, 329 (2002); *see also id*. at 328 ("naive").

3. Joseph L. Gastwirth, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 333. 333 (2002); *see also id*. at 334 ("serious flaw"); *id*. at 339 ("good discussion of the logistic model and interpretation of the coefficients").

4. John M. Conley, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 309, 309 (2002).

5. Dale A. Nance, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 341, 341 (2002).

6. Marc Rosenblum, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 347, 347 (2002).

7. Shari Seidman Diamond, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 315, 315 (2002); *see also id*. ("ingenious idea").

8. William B. Fairley, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 321, 321 (2002); *see also id*. at 326 ("careful exposition").

regression equation,"[9] and "the job has been done exceptionally well and is presented remarkably clearly."[10] The prototype and the comments highlight important general issues of how instructional materials should be presented. The comments also provide many useful lessons on specific content issues for such products, and, of course, offer very helpful remarks about our particular prototype.

I. GENERAL CONTENT ISSUES

The prototype is aimed at several different levels of quantitative sophistication, with the idea that its hyperlink format will allow users to explore the material according to their own levels of comfort. The product certainly could be more interactive than contemplated by the prototype, but additional interactivity presents additional challenges.

The goal is to teach statistics, not law, so the prototype avoids murky legal issues and uses the most authoritative legal citations. The example also was designed so as not to be especially favorable to either plaintiffs or defendants.

Quantitative Sophistication. Presenting difficult material to smart people requires a careful balancing of clarity and sophistication.[11] Teaching statistics to judges also requires the recognition that judges will vary in their preparation and interest. Consequently, information in the prototype is pitched at various levels of sophistication.[12]

"The world of law is a world of words,"[13] and it often is said that "most lawyers select the law as a career because they are neither mathematically nor technically inclined."[14] This does not mean, however, that judges cannot learn how to understand and evaluate statistical evidence. On the one hand, federal judges are among the most intelligent and best-educated members of society. Although they may not be drawn to the law because of their facility with mathematics, they often are attracted to law because of the importance of logic and analytical reasoning to

---

9. David W. Barnes, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 301, 305 (2002); *see also id.* at 303 ("great method"); *id.* at 301 ("splendid job"); *id.* ("brilliant[] illustrat[ion]").

10. Michael J. Saks, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 351, 352 n.6 (2002); *see also id.* at 353 ("the problem is an excellent one and is presented extraordinarily well").

11. *See, e.g.*, Diamond, *supra* note 7, at 315 (recognizing the "tension between coverage and concise presentation").

12. *See* Saks, *supra* note 10, at 353 ("The prototype . . . moves along unevenly—at times taking a very basic approach, other times using a rather mathematical and formula-based one, and occasionally offering some conceptual knowledge.").

13. Robert Timothy Reagan, *Supreme Court Decisions and Probability Theory: Getting the Analysis Right*, 77 U. DET. MERCY L. REV. 835, 835 (2000).

14. Sheldon L. Trubatch, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 357, 360 (2002); *see also* Saks, *supra* note 10, at 351 ("Law students, and therefore the lawyers and judges they become, are people of the word, not of the number . . . .").

the field. On the other hand, mastering statistical analysis usually is no walk in the park, even for persons drawn to more "technical" disciplines.

Two things make statistics difficult—mathematics and logic. It is okay to expose judges to mathematics. It may be true that "[j]udges need to know less about formulas and more about when particular statistical approaches are appropriate,"[15] but surely some judges have the interest and ability to improve their familiarity with statistical computations.[16] These judges may find helpful a teaching tool that presents statistics in a legal context without shying away from mathematics.

The logic of statistical analysis may be difficult at first, but is attainable with sufficient exposure. "It [may be] rarely self-evident to non-statisticians that the meaning of a disparity can be evaluated by calculating the probability that it would occur by chance,"[17] but familiar expressions such as "what are the odds?" and "coincidence? I don't think so" suggest that statistical reasoning is occasionally intuitive.[18]

Linear and Hyperlink Presentations. Perhaps the best way to teach *mastery* of a discipline is to "start at the beginning and develop the ideas in logical sequence."[19] Although many judges want to learn more about statistics, few to none have the time or interest to master the discipline. So the prototype immerses its audience into statistical analyses with the fundamentals often presented through hyperlinks instead of earlier lessons.[20]

With the traditional textbook approach, the author decides the most effective order of presentation for the student. In contrast, with a hyperlink approach, the student has more control over order of presentation.[21] The advantage of the one-

---

15. Barnes, *supra* note 9, at 301; *see also* Saks, *supra* note 10, at 352; Rosenblum, *supra* note 6, at 347. *But see* Nance, *supra* note 5, at 341 ("The user is given just what is needed to understand the analysis without being overwhelmed by the mathematics.").

16. I am not sure I would go quite as far as Dr. Trubatch, who observes, "if you cannot do the calculations, you have not learned the material." Trubatch, *supra* note 14, at 360 n.11. I do think that familiarity with calculations promotes mastery of the material.

17. Conley, *supra* note 4, at 310.

18. Compare also the observation by Professor Ashenfelter that "statistical reasoning alone rarely provides compelling evidence for causality," Orley Ashenfelter, *Comment on the Age Discrimination Example*, 42 JURIMETRICS J. 297, 299 (2002), with a definition of *res ipsa loquitur*: "Rebuttable presumption or inference that defendant was negligent, which arises upon proof . . . that the accident was one which ordinarily does not happen in absence of negligence." BLACK'S LAW DICTIONARY 1305 (6th ed. 1990).

19. Freedman, *supra* note 2, at 328.

20. "The software prototype, while providing a vivid legal example, offers insufficient general background to be useful to the beginner." Conley, *supra* note 4, at 310. We might want to design the ultimate product so that users can choose between earlier lessons or hyperlinks to acquire the fundamentals.

21. "Asking students to learn a lot of underlying premises before they come close to seeing what the tools are to be used for seems meaningless and boring." Saks, *supra* note 10, at 352. Professor Barnes observes that statistical analyses rely on a complex network of foundation concepts, like the root system of an aspen forest. Barnes, *supra* note 9, at 302. Interestingly, instead of concluding that

size-fits-all approach is that the presenter has more experience with persons learning the material; the disadvantage is that one size does *not* fit all. The hyperlink approach provides the user with a greater variety of "sizes"; nevertheless, students might not pick the right size. In particular, students might not click on terms they think they understand, but do not.[22]

The adaptability of the hyperlink approach to a diverse body of students is likely to make it very useful as a way to present difficult material to judges with highly varied backgrounds, interests, needs, and time. But overcoming the disadvantages will require very careful design work.

Interactivity. Presenting information with a computer instead of a book allows for more interactivity, but our prototype has not pursued interactivity beyond clicking on hyperlinks.[23] "However, interactivity could be realized by including a means to change some of the numerical assumptions to see how the results of the statistical analysis would vary."[24]

Enhanced interactivity is a terrific idea, but presents substantial challenges. For example, statistical significance in the prototype was pretty clear. If users are free to change the data, statistical significance might approach more borderline levels and it may be difficult to ensure that a help window on interpreting borderline results "pops up" in such circumstances. Also, nonstatisticians often are skeptical of conclusions that would be substantially changed if one or a few data points were slightly altered,[25] possibly because they overestimate the prior probabilities of those slight changes.

Legal Authority. If the law is to serve only as a backdrop for the statistical presentation, the law probably should be clear. This implies a strong preference for citations to statutory text and Supreme Court opinions rather than lower-court opinions.[26]

The one area of the law discussed in the prototype that is unclear is whether the Age Discrimination in Employment Act[27] provides a cause of action for dispar-

---

this makes statistics particularly suitable to presentation by a hyperlink approach, Professor Barnes suggests that this approach is unlikely to work. *Id.* at 302–03.

22. Barnes, *supra* note 9, at 302 ("the software user may not know the importance of understanding a hyperlinked term"); Diamond, *supra* note 7, at 316 ("some familiar words can have special meanings").

23. Trubatch, *supra* note 14, at 358, 361.

24. *Id.* at 358; *see also* Diamond, *supra* note 7, at 317.

25. David H. Kaye, *Improving Legal Statistics*, 24 LAW & SOC'Y REV. 1255, 1267 (1990) ("[R]esults can be sensitive to slight changes in cells of tables that contain small numbers—an appealing argument, at least superficially."); Joseph L. Gastwirth, *Statistical Issues Arising in Equal Employment Litigation*, 36 JURIMETRICS J. 353, 367 (1996) ("It has become almost routine for defendants to challenge statistical results near the borderline of statistical significance for their sensitivity to changes of a few observations.").

26. This is one reason why we did not pursue an approach such as the one pursued in Professor Gastwirth's comment, where the professor ably developed a case for plaintiffs using district court and court of appeals cases. Gastwirth, *supra* note 3, at 336–38; *see also* Rosenblum, *supra* note 6, at 347, 349 n.8 (advocating citations to circuit law).

27. 29 U.S.C. §§ 621-634 (1994 & Supp. V 1999).

ate impact.[28] But at the time of this writing it is clear that this is unclear, because the Supreme Court has said so.[29]

Balance. It is very important that a product such as our prototype not intrude on the province of judges by explicitly or implicitly advocating answers to unsettled questions.[30] The prototype thus begins with evidence favorable to the plaintiffs and ends with evidence favorable to the defendant, but because of the unsettled question of whether disparate impact is actionable, no party clearly wins.

Even so, some users may find the materials too hard on either plaintiffs[31] or defendants.[32] Balance cannot necessarily be achieved with each individual example. It is more achievable in the product as a whole.

II. SPECIFIC CONTENT ISSUES

The final product will have to include lessons on several topics not well-developed in the prototype, including, among others, statistical significance, one- and two-tailed tests, size of effect and level of significance, type I and type II errors, power, and the fallacy of the transposed conditional.

Statistical Significance. The prototype does not include a lesson on criteria for statistical significance,[33] but the complete product must include one. This will not be an easy lesson to craft. It would not be proper, for example, to say or imply that the law should accept the conventional threshold of $p = 0.05$ as a firm criterion for admissibility or proof. "It is . . . 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."[34] So the prototype illus-

---

28. Reagan, *supra* note 1, at 295.

29. Hazen Paper Co. v. Biggins, 507 U.S. 604, 610 (1993) ("[W]e have never decided whether a disparate impact theory of liability is available under the ADEA"). On March 20, 2002, the Court heard oral argument in a case that was likely to settle the question, but dismissed the writ of certiorari as improvidently granted. *See* Adams v. Florida Power Corp., 255 F.3d 1322 (11th Cir.), *cert. granted*, 122 S. Ct. 643 (2001), *cert. dismissed*, 122 S. Ct. 1290 (2002).

30. Conley, *supra* note 4, at 312 ("a teaching hypothetical should be balanced at every level").

31. *See* Nance, *supra* note 5, at 345 ("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").

32. *See* Fairley, *supra* note 8, at 323 ("the error created by ignoring the validity of the model almost always favors plaintiffs over defendants").

33. Ashenfelter, *supra* note 18, at 298 ("Reagan avoids the usual discussion of significance levels by reporting on *p*-values").

34. Nance, *supra* note 5, at 343–44; *see also id.* at 343 (noting that the law cannot leisurely wait for further study); *id.* at 344 n.12 (observing that even in science the criterion for statistical significance is not always firm); Ashenfelter, *supra* note 18, at 298 (expressing sympathy with "the purpose of avoiding hard criteria for 'statistical significance'"). *But see* Gastwirth, *supra* note 3, at 335 ("statistically significant at the commonly accepted 0.05 level").

trates statistically significant results with very low *p*-values[35] and nonsignificant results with very high *p*-values.[36]

Other difficult issues the product will have to address include the following: (1) how the *p*-value relates to the standard of proof,[37] (2) what to conclude when different tests yield different *p*-values,[38] and (3) whether *p*-values make sense when one has all the data rather than a sample.[39] Perhaps even more challenging is the question of what impact on interpretation of *p*-values the parties' ability to select which evidence to present should have.[40]

One- and Two-Tailed Tests. The prototype acknowledges that statistical tests can be one- or two-tailed,[41] but it offers scant guidance on how to know which to prefer.[42] This is because the issue will be a difficult one to present properly.

One-tailed tests are used when prior theory justifies their extra power, but their logic can be slippery. A two-tailed test assesses whether something other than chance is at work. A one-tailed test assesses whether something more specific is at work, but it also treats the other tail as impossible or irrelevant. A one-tailed test to prove sex discrimination, for example, might be considered unfair to the defendant, because the defendant might be liable to other plaintiffs if the results were in the other tail.

I am sympathetic to relaxed criteria for statistical significance but strict criteria for one-tailed tests.[43] Development of our full product will require delicate crafting, because this view has not yet been embraced universally.

Size of Effect and Level of Significance. Persons learning about statistical analysis often confuse size of effect with level of significance.[44] There are two reasons for this. First, each has probative value—size of effect relates to magnitude of consequences, level of significance relates to credibility of evidence.[45] Sec-

---

35. Reagan, *supra* note 1, at 286 ($p = .0002$, $p = .0003$); *id.* at 288 ($p = .001$); *id.* at 291 ($p = .0001$); *id.* at 294 ($p = .019$). *But see* Fairley, *supra* note 8, at 326 ("Tiny *p*-values exist only as logical deductions from very simple probability models that almost always fail to incorporate real world phenomena that would imply greater probabilities.").

36. Reagan, *supra* note 1, at 294 ($p = .307$). *But see* Nance, *supra* note 5, at 345 (observing that even a *p*-value of .307 can be probative).

37. Conley, *supra* note 4, at 311.

38. Barnes, *supra* note 9, at 306; *see* Gastwirth, *supra* note 3, at 334 (suggesting that $p = .0000074$ and $p = .0002$ are likely to have different legal consequences).

39. Diamond, *supra* note 7, at 317 n.6. The logic of treating all the data as if they were a sample is to recognize that the data are produced by a process under study that could by chance have produced somewhat different data.

40. Ashenfelter, *supra* note 18, at 299 ("If evidence is provided only because it is favorable, then it is very difficult to interpret conventional *p*-values.").

41. Reagan, *supra* note 1, at 286, 288.

42. Barnes, *supra* note 9, at 305: Nance, *supra* note 5, at 342–43.

43. *See* Nance, *supra* note 5, at 345–46 (cautioning against strict criteria for statistical significance and raising the question of standards for the use of one-tailed tests).

44. Fairley, *supra* note 8, at 323–24.

45. *See* Gastwirth, *supra* note 3, at 334 (observing that a lower *p*-value strengthens the case).

ond, if power of the test is held constant, one generally serves as a proxy for the other.[46] The final product, however, should provide lessons clarifying this distinction.[47]

Type I and Type II Errors. The prototype does not discuss type I and type II error, but the final product will have to somewhere.[48] Such discussion is omitted from the prototype, because discussion of criteria for statistical significance was omitted. But type I and type II errors are concepts with very general application. Type I errors are simply false positives, or erroneous findings in favor of the party with the burden of proof. Type II errors are simply false negatives, or erroneous findings against the party with the burden of proof.

Power. Lessons on statistical power also will be very important components of the final product.[49]

Transposed Conditionals. The final product will have lessons steering users away from the fallacy of the transposed conditional.[50] The probability that a fact is true given certain evidence is not necessarily the same as the probability of obtaining the evidence if the fact were true.[51] *P*-values are probabilities of obtaining evidence assuming a statistical hypothesis is true. They are not probabilities that the null hypothesis—that the data resulted purely by chance—is true.[52]

III. MISCELLANEOUS PROTOTYPE ISSUES

Appropriateness of Analysis. The purpose of the product is to teach judges what statistical analyses mean, not to provide rules for which analyses to admit or rely on as evidence. One message the presentation clearly implies, however, is that "often there are several reasonable approaches that typically yield similar, but not identical, results."[53]

The presentation of Fisher's exact test, the chi-squared test, and bivariate logistic regression is meant to show that the same data can be analyzed legitimately

---

46. *See, e.g.*, ROBERT ROSENTHAL & RALPH L. ROSNOW, ESSENTIALS OF BEHAVIORAL RESEARCH xvi (1984) (observing that under many circumstances, level of significance = size of effect × size of study); *see also* Ashenfelter, *supra* note 18, at 298 ("Minuscule differences from the null hypothesis will always be detected with large enough samples.").

47. Diamond, *supra* note 7, at 317 (suggesting instruction on "the relative impact of sample size and effect size"); Nance, *supra* note 5, at 344 n.13 ("one must distinguish between the strength of an association between two variables and the strength of the inference that such an association exists").

48. Ashenfelter, *supra* note 18, at 298.

49. Barnes, *supra* note 9, at 306; Gastwirth, *supra* note 3, at 340; Nance, *supra* note 5, at 346 n.19.

50. Conley, *supra* note 4, at 310–11.

51. *See, e.g.*, Robert Timothy Reagan, Book Review, 52 OKLA. L. REV. 291, 299 (1999).

52. Nance, *supra* note 5, at 345 n.16.

53. Gastwirth, *supra* note 3, at 339; *see also* Diamond, *supra* note 7, at 316 ("The lesson that the user should take away is not to be daunted by the variety of available tests, but to be aware of their similarities and differences.").

in different ways. That does not mean all methods are equally legitimate,[54] but if a legitimate method is used to establish a prima facie case, then the burden shifts to the defendant to produce evidence that a different analysis would support a different conclusion.

The presentation is asymmetric in that plaintiffs' evidence is analyzed three ways, but defendant's evidence is analyzed only one way.[55] This probably is an artifact of the product's still being in development. Perhaps the age discrimination example should simply compare a bivariate logistic regression with a multivariate logistic regression[56] and present Fisher's exact test and the chi-squared test in other lessons, but because those other lessons are not yet developed, everything is presented in the same lesson.

Another form of asymmetry is the presentation of the computational formula for the chi-squared test, but not the computational formulas for Fisher's exact test or logistic regression. Obviously, this is because the formula for chi-squared is the easiest to master and the only one that does not realistically require a computer for computation. But perhaps the final product should include hyperlinks to the omitted formulas.[57]

Assumptions and Models. Statistical tests can be highly informative despite the fact that their underlying assumptions are never *precisely* true. For example, it cannot possibly be the case that defendant treated persons over 40 differently from persons under 40, but treated all persons over 40 the same as each other and all persons under 40 the same as each other.[58] A statistical analysis must use a reason-

---

54. *See* Reagan, *supra* note 1, at 288 ("Because the chi-squared test only approximates the results of a Fisher's exact test, the Fisher's test usually is preferable if computer software for statistical analysis is available."); *see also* Barnes, *supra* note 9, at 303 ("Screen 4.3.1 hints that the chi-square test should not be used in this century."); Saks, *supra* note 10, at 354, 354 n.8. Note that the clear advantage of a chi-squared test over Fisher's exact test is that a judge or attorney can do the test with a hand calculator. I disagree with Dr. Rosenblum's suggestion that chi-squared analyses be deemed inadmissible. *See* Rosenblum, *supra* note 6, at 348–49.

55. Ashenfelter, *supra* note 18, at 299 ("defendant's claims might be tested in several other ways"); Gastwirth, *supra* note 3, at 334 (suggesting a Cochran-Armitage trend test). In an effort to rebut defendant's evidence, Professor Gastwirth hypothesizes additional facts and advocates analyses demonstrating pretext. Gastwirth, *supra* note 3, at 333–34, 335, 337, 340.

56. *See* Freedman, *supra* note 2, at 328 ("The focus of the analysis in the hypothetical seems to be a logistic regression model applied to an age discrimination case.").

57. Barnes, *supra* note 9, at 305; Saks, *supra* note 10, at 355.

58. "Employees are not terminated 'at random.'" Fairley, *supra* note 8, at 321–22; *see also* Rosenblum, *supra* note 6, at 348 ("In many instances, employers may not take age into account before age 50 or 55, so that inclusion of persons between ages 40 and 49 as older and retained would bias the statistical comparisons to favor employers.") *But see* Gastwirth, *supra* note 3, at 334 (observing that Fisher's exact test "is an appropriate procedure" because it treats "all employees over 40-years-old as having the same probability of being terminated and tests whether this probability is the same as that of employees younger than 40"). It is intriguing to contemplate whether we should endorse peculiar statistical models arguable implied by the law—age does not matter until the employee's fortieth birthday, and then it matters equally every year, for example. *See* Rosenblum, *supra* note 6, at 349 ("the use of actual age values below the statutory limit of 40 is incorrect. The better view is that all persons below age 40 carry a constant age value, i.e., 39").

able model, but if we had certain knowledge of the objectively correct model, we might not need to do the statistical analysis.[59]

The prototype is designed to illustrate how erroneous results can arise from the failure to take into account an explanatory variable, and other lessons can illustrate other model difficulties, such as erroneous assumptions of independence.[60]

Statistical and Decision-Making Proxies. The prototype hypothetical somewhat facetiously proposes that defendant terminated employees excessively loyal to previous management, who would have tended to have more seniority, and therefore would have tended to be older.[61] The prototype does not make clear that length of service is a *statistical* proxy for loyalty in the analysis (how can loyalty be measured objectively?),[62] rather than a *decision-making* proxy.[63] This was an error.

Meaning of "Factor." The prototype says that we would expect older and younger employees to be terminated at similar rates "if age were not a factor in termination."[64] There is a danger that users will understand "factor" to mean "causal factor."[65] This illustrates the tension between clarity and precision.

Perhaps the phrase should have read, "if age were not a factor considered in terminations and not associated with any factors that were considered," or, "if age were not a factor in termination (direct or indirect, causal or otherwise)," but sometimes it is necessary to temporarily suspend precision for the sake of clarity.[66]

Chi-Squared Rules. The prototype states the conventional frequency rules for when a chi-squared test is likely to provide a sufficiently accurate *p*-value[67] and suggests that Yates' correction for continuity might be used on a two-by-two table when the marginals are fixed.[68] Although these rules are followed widely, they do

---

59. *But see* Freedman, *supra* note 2, at 329 (suggesting that you cannot learn from statistical analysis unless you know in advance what you are trying to discover: "if the model is wrong—i.e., does not describe the mechanism by which the data are generated—inferences are unreliable").

60. *See* Fairley, *supra* note 8, at 322–33 (observing that termination decisions might not be independent); Freedman, *supra* note 2, at 328 n.7 (same).

61. Reagan, *supra* note 1, at 291.

62. *Id.* at 293 ("the more senior employees were less likely to be loyal to new management"); *see* Rosenblum, *supra* note 6, at 349 ("future loyalty . . . is simply too subjective and unknown to be taken into statistical account").

63. Reagan, *supra* note 1, at 293 ("employees . . . were terminated . . . in part because of their years of service"); *id.* at 295 ("defendant's reliance on . . . length of service as a proxy for loyalty").

64. Reagan, *supra* note 1, at 284 tbl. 1, 285.

65. Fairley, *supra* note 8, at 323–24; *see also* Ashenfelter, *supra* note 18, at 299 ("correlation does not establish causality").

66. *See, e.g.*, Conley, *supra* note 4, at 312–13 (objecting to complex prose elsewhere in the prototype); Saks, *supra* note 10, at 355 (same).

67. Reagan, *supra* note 1, at 288 (citing GEORGE W. SNEDECOR & WILLIAM G. COCHRAN, STATISTICAL METHODS 127 (8th ed. 1989)).

68. *Id.* at 287 & tbl. 2 (citing WILLIAM L. HAYS, STATISTICS 861 (5th ed. 1994); DAVID C. HOWELL, STATISTICAL METHODS FOR PSYCHOLOGY 146 (4th ed. 1997)).

not always yield accurate results.[69] This does not matter for the example presented, but it is a challenging design issue to determine how aware to make users of arcane statistical issues that do not apply to the current example, but might apply to other situations they might encounter.

Correlated Independent Variables. The prototype example was designed so that the variable that plaintiffs argued caused their injuries was not causal, but merely correlated with causal variables. It is important to acknowledge that when two independent variables are both correlated with a dependent variable, but also highly correlated with each other, it is difficult accurately to assess the "real" statistical relationships among the variables.[70]

Teaching statistics to judges is no easy task, because teaching statistics to anyone is no easy task. It is, nevertheless, an important task, and this published symposium provides a treasure-trove of useful insights.

---

69. Freedman, *supra* note 2, at 330–31, 328 n.7; Nance, *supra* note 5, at 342, 342 n.3–5. Professor Freedman objects to reliance on the Howell text, Freedman, *supra* note 2, at 331, although Professor Howell supports Professor Freedman's objection to Yates' correction for continuity, HOWELL, *supra* note 68. I cited Howell in the prototype because I find it provides the best combination of clarity and sophistication among statistics texts I have seen and because it has a nice caution on the use of the Yates' correction.

70. Freedman, *supra* note 2, at 329 ("No reliable conclusion can be drawn from these data, because the two explanatory variables are so highly correlated."); Gastwirth, *supra* note 3, at 335 ("[I]t will be difficult to distinguish the effect of age from that of seniority . . . ."); Rosenblum, *supra* note 6, at 350 ("defendant's explanation does not statistically rebut plaintiff argument that age was a factor in the employee terminations.").