A RE-ANALYSIS OF THE UNIVERSITY OF CHICAGO JUDGE-JURY AGREEMENT STUDY

Joseph L. Gastwirth, George Washington University and Michael D. Sinclair, Response Analysis
Joseph L. Gastwirth, GWU, 2201 G Street Rm. 315 Washington D.C. 20052[1].

Keywords: Judicial Assessment, Diagnostic Tests, Classification Errors.

I. Introduction

The Chicago Jury Study, titled The American Jury (1966) by Harry Kalven and Hans Zeisel examined the direction of disagreements between the judge and the jury and the explanatory factors that contributed to the disagreements under various situations. In this paper we have applied and extended the Hui and Walter (1980) diagnostic test paradigm to estimate the level of error in the verdicts. To best apply this technique, we needed to better understand the factors that contributed to disagreements and how these varied under different situations. As a result, we first conducted a reanalysis of the data using more recent categorical data methods to quantify the individual and combined effects of the factors contributing to the disagreements. While we will not present the detailed reanalysis, we discuss the results and use them in applying the Hui and Walter method.

In the assessment of diagnostic tests, the subjects are evaluated using both the new test (possible a screening procedure) and a confirmatory test with a known or presumably negligible level of error. If such a confirmatory test does not exist (i.e. all other available tests have non-negligible levels of error) an accurate assessment is not possible. The innovative aspect of the Hui and Walter (H&W) method for the evaluation of diagnostic test is that it allows the researcher to evaluate the error rates of both tests applied. To apply the method, one must select two populations or subpopulations with different prevalences. The error rates for each test are assumed to be equal across the two subpopulations, but the error rates for each test are allowed to differ. With this setup, the error rates for the two tests can be estimated along with an estimate of the true prevalence rate for each subpopulation.

The application of the H&W methodology has also proven effective in other applications. Sinclair and Gastwirth (1996) applied the H&W method procedure and modifications of the technique to the analysis of reinterview surveys. This application was primarily successful due to the availability of two or more demographic groups that one could reasonably assume meet the equal error rate requirement. Therefore, we attempted to apply these methods to the judge-jury disagreement data. Based on our reanalysis of the original data and a review of the sensitivity of the H&W method to the equal error rate requirement we found that basic H&W method was limited in application. By extending the model in section IV, we were able to obtain more accurate estimates of the verdict error rates.

Since there has been interest in conducting a similar study of judge-jury agreement in the future, we have provided suggestions for a new design in section V that are less dependent on the assumptions required by the analytical methods presented.

II. The Hui and Walter Method

The frequency of cases resulting in a verdict (guilty or acquit) from the jury and the judge can be expressed as a 2 X 2 table. We will index such a table for a specific crime by the letter g. We will denote the frequency of cases of crime/data type g that the jury gives a verdict of status i, i= 1 for guilty and i=2 for acquitted and the judge classifies as status j, j=1 for guilty and j=2 for acquitted by n_{gij}. Let π_g denote the true unknown prevalence rate of guilt in the gth subpopulation and let $\alpha_{r,g}$ and $ß_{r,g}$ denote the unknown false positive and false negative rates. These error rates are indexed by the letter r, such that r=1 corresponds to the jury's verdict and r=2 for the judge's verdict for subpopulation g. The false positive rate, $\alpha_{r,g}$ refers to the probability that evaluation procedure r will classify the person as guilty when in truth the person should have been acquitted. Similarly, the false negative rate, $ß_{r,g}$ is the probability that the evaluation procedure r will acquit in the case when in truth the party was guilty. Hence, the false negative rate corresponds to the level of leniency in the cases by evaluation procedure r. One (1) minus each of these parameters corresponds to the specificity and sensitivity of the classification procedures, respectively (Gastwirth 1987; Brookmeyer and Gail 1994).

Assuming conditional independence between the two evaluator's errors, (i.e. the verdicts from the judge and the jury are provided independently) the multinomial probabilities associated with the cell frequencies are as follows:

[1] This research was supported in part by a grant from the National Science Foundation.

$$\begin{aligned}
P(i=1, j=1) &= \pi_g (1-\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(\alpha_{1,g} \times \alpha_{2,g}) \\
P(i=2, j=1) &= \pi_g (\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(\alpha_{2,g}) \\
P(i=1, j=2) &= \pi_g (1-\beta_{1,g}) \beta_{2,g} + (1-\pi_g)(\alpha_{1,g})(1-\alpha_{2,g}) \\
P(i=2, j=2) &= \pi_g (\beta_{1,g} \times \beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(1-\alpha_{2,g})
\end{aligned}$$

In the above expression, we find that we have a total of five parameters, but only three independent cell entries (or degrees of freedom) from which to conduct the estimation. Therefore, the number of parameters must be reduced for estimation purposes.

To conduct estimation, we follow the Hui and Walter (1980) method developed for the evaluation of diagnostic tests. We select two populations or subpopulations with different prevalences, i.e., $\pi_1 \neq \pi_2$, such that, the error rates for each evaluator are equal across the two subpopulations. Note that the error rates associated with each evaluator are allowed to differ. This implies, $\beta_r = \beta_{r,1} = \beta_{r,2}$, and $\alpha_r = \alpha_{r,1} = \alpha_{r,2}$, with $\beta_1 \neq \beta_2$, $\alpha_1 \neq \alpha_2$. Under these conditions, the number of parameters reduces to six (two prevalence rates, one for each subpopulation and four common error rates). Given that the two 2X2 tables contain six degrees of freedom, estimation is possible. Closed form formulas for the estimates are given in the original paper. Estimated variances for the estimators are derived from the estimated asymptotic information matrix.

The H&W approach yields estimates of the prevalence rates and the misclassification rates. Sinclair and Gastwirth (1994) showed that these estimates may be biased when the test errors are not the same across the two populations (i.e. $\alpha_{r,1}$ not equal to $\alpha_{r,2}$ and/or $\beta_{r,1}$ not equal to $\beta_{r,2}$, for the first test ,r =1, or both tests, r= 1 and 2). In some situations, the method is quite sensitive to a violation to this assumption, but in other cases the method is very robust. In particluar, the sensitivety of the H&W method to a violation in the equal error rate assumption decreases as the difference in the prevelance rates in the two populations increases.

III. Application of the Basic Method

The original data consisted of several 2X2 tables for a variety of crimes. Treating the crimes as subpopulations, we set out to find two crimes that would meet the H&W requirements. Our reanalysis of the original data showed that the presence of a prior record played a key role in a disagreement between the judge and the jury. If two crimes had a similar high proportion of cases with a prior record, we suspected that occurrence of a prior record would probably be known by the jury, and as a result, the jury error rates should be similar for the two crimes. Futhermore, we attempted to find two such crimes that also had different prevelance rates to reduce the sensitivety of the H&W method to any differences in the error rates for the two crimes.

We present an example in Table 1 using auto theft and burglary cases from Table 19 (page 72 K&Z). Both of these crimes have a high rate of prior records/low percentage of first offenders (32% vs. 22% respectively). Both of these crimes also showed differences in the proportion of guilty verdicts by both the judge and the jury (.74/.91 jury/judge Burglary vs. .67/.80 jury/judge auto theft).

The results in Table 1 show that the juries tend to be substantially more lenient than the judges (β_1 =.192 vs β_2= .012). However, the probability of assigning a verdict of guilt incorrectly seems to be undetectable for both groups given the data collected. Given that the judge's error rates are lower than those from the jury, we see that the model estimate of the true guilt is closer to the judge's assessment.

IV. Extensions of the H&W Method

In table 146, K&Z present a comparison of judge and jury verdicts by the presence or absence of a jury request for more information on the legal definition of the crime, the law governing the evidence, sentencing etc. In this data, the jury verdict has three outcomes, acquitted, hung (unable to decide) and guilty. As in the prior models, the judge only has two outcomes, guilty or acquitted and case has only two true states, guilty or innocent. As a result, the data is structured in terms of two 3 x 2 tables (one for cases in which the jury requested information and one for those cases in which no request was made). We denote the frequency of cases of type g, g=1 for a request and g=2 for no request that the jury gives a verdict of status i, i= 1 for guilty and i=2 for hung and i= 3 for acquitted and the judge classifies as status j, j=1 for guilty and j=2 for acquitted by n_{gij} . Let π_g denote the prevalence rate of guilt in the gth subpopulation.

Given a three outcome structure, for the classification rates, define β_{grij} as the probability that evaluator, r, r=1 for jury and r=2 for judge, will classify a case from type g to be in category i, i=1,2 and 3 when the true status of the individual is category j. For example, β_{1131} denotes the probability that a request case (g=1) is classified by the jury (r=1) as acquitted (i=3) when the true status is guilty (j=1). Since the judge only has two possible verdict outcomes, β_{g221} and β_{g223} are equal to zero.

The classification rates can be divided into two groups corresponding to those associated with a correct classification and those associated with an erroneous classification. Note that for each g and r, the probability evaluator r, classifies a truly guilty case from type g correctly as guilty, is equal to $\beta_{gr11} = (1-\beta_{gr21} - \beta_{gr31})$. The corresponding probability for acquitted is $\beta_{gr33} = (1-\beta_{gr13} - \beta_{yr23})$. Hence, the correct classification rates are simply determined by the error rates. The expected probabilities associated with each of the six cells, P(i,j) for a particular type, g are given below.

Table 1
Hui and Walter Analysis of Burglary and Auto Theft Cases

Pop	Sample Size	Hui and Walter Model Estimates					Observed Jury Prevalence	Observed Judge Prevalence
		Beta 1 Jury Class: NG True: Guilty	Beta 2 Judge Class: NG True: Guilty	Alpha 1 Jury Class: Guilty True: NG	Alpha 2 Judge Class: Guilty True: NG	Prevalence Guilty		
Burglary	298	.192 (.044)	.012 (.015)	.009 (.089)	.000 (.399)	.921 (.037)	.740	.910
Auto Theft	111					.810 (.087)	.670	.800

$$
\begin{aligned}
P(1,1) &= \pi_g \times (1-\beta_{g121}-\beta_{g131}) \times (1-\beta_{g231}) + (1-\pi_g) \times \beta_{g113} \times \beta_{g213} \\
P(1,3) &= \pi_g \times (1-\beta_{g121}-\beta_{g131}) \times \beta_{g231} + (1-\pi_g) \times \beta_{g113} \times (1-\beta_{g213}) \\
P(2,1) &= \pi_g \times \beta_{g121} \times (1-\beta_{g231}) + (1-\pi_g) \times \beta_{g123} \times \beta_{g213} \\
P(2,3) &= \pi_g \times \beta_{g121} \times \beta_{g231} + (1-\pi_g) \times \beta_{g123} \times (1-\beta_{g213}) \\
P(3,1) &= \pi_g \times \beta_{g131} \times (1-\beta_{g231}) + (1-\pi_g) \times (1-\beta_{g123}-\beta_{g113}) \times \beta_{g213} \\
P(3,3) &= \pi_g \times \beta_{g131} \times \beta_{g231} + (1-\pi_g) \times (1-\beta_{g123}-\beta_{g113}) \times (1-\beta_{g213}))
\end{aligned}
$$

The above data model contains seven parameters and the table for a given type g a total of five degrees of freedom.

Upon review of the data we found that the observed prevalence of guilt for the jury and the judge differed only slightly depending on whether or not the jury made a request for additional information (.62/.65 jury/request/no-request, .86/.84 judge/request/no-request). As a result, we assumed that the underlying prevalence rate was equal for the request and no-request cases (i.e. $\pi = \pi_1 = \pi_2$). Given that we expected the error rates for the jury to be different depending on whether or not a request was made, we assumed that for the request cases (g=1) that

$$\beta_{11ij} = C1 \times \beta_{21ij} \text{ for all i and j,}$$

where C1 is assumed to be greater than or equal to zero. This assumption implies that the error rates for the jury among those with a request (g=1) are a common multiple larger or smaller than those by the jury without a request (g=2). For future presentation, we will delete the g subscript from the no request (g=2) jury error rates (i.e., $\beta_{21ij} \Rightarrow \beta_{1ij}$).

Similarly, we assume that for the judge,

$$\beta_{g2ij} = C2 \times \beta_{21ij} \text{ for g = 1 and 2 and i=1,3 and j=1,3}$$

where C2 is assumed to be greater than or equal to zero.

This assumption implies that the error rates for the judge (r=2) are the same for the request (g=1) and no request (g=2) cases and that these errors are common multiple larger or smaller than those by the jury (r=1) without a request (g=2).

With these assumptions, the number of parameters is reduced to four error rates for the jury among the no request cases, β_{121}, β_{131}, β_{113} and β_{123}, a common prevalence rate, π, and two unknown fractions, C1 and C2 (seven parameters in total). The results of the data analysis are given in Table 2

The estimate of β_{113}, defined as the probability of the jury yielding a verdict of guilty when the case should have been acquitted was not significant from zero. Therefore, we evaluated a reduced model that assumed β_{113} was zero. The results for the reduced model are presented in Table 3. The ln likelihood for the initial/reduced model is equal to -4031.7/-4032.8. Hence, the data is consistent with the reduced model.

The results from Table 3 indicate the jury will provide a verdict of acquitted when the case should have received a verdict of guilty about 19% (β_{131} =.1912) of the time (+/- 1.4%). The jury is unable to decide on a verdict, (i.e. hung) when the case is guilty about 4.7% (β_{121} =.0470), of the time (+/- .7%) and when the case should be acquitted at about 8.4% (β_{123} =.0843) of the time (+/- 2.5%). On the other hand, the judge only acquitted when the case was guilty at about 3% ($\beta_{231} = \beta_{131} \times C2$ =.1912×.1623=.0310) of the time (+/-.7% [2]). Hence, the jury is considerably more lenient than the judge in the cases studied. Furthermore, the need for further information increases the jury's error rates by about 32% (C1= 1.317) (+/- 18%).

It is comforting to note that the estimate of β_{131} =.1912 is also close to the estimate of the jury's false

[2] The estimated standard error of the product of C2 and β_{131} was computed using a first order taylor series approximation (Wolter, 1985)

negative rate (acquit when guilty) of β_1 = .192 given in Table 1. The judge's false negative rate in Table 1, β_2 = .012 is also relatively close to the estimate obtained above, β_{231} =.0310.

The above example illustrates how these models can be adapted to analyze data of this type and the results provide some fairly reliable estimates of the errors in verdicts. The above data model incorporates many assumptions about the error rates; however, each of these assumptions are plausible. Furthermore, sample sizes of 2,993 non-request and 584 with a request help to stabilize the estimates. The fact the underlying prevalence of guilty might differ slightly between the request and no request cases could distort the results, but we expect that this aspect only results in providing an estimate of the true prevalence rate that represents and average of the two underlying rates. The assumption that the judge has the same level of error among the request and no request cases may be questionable as well, but given the judge is substantially less lenient than the jury the impact of any of these differences should be minor.

V. Alternative Design Strategy

With dichotomous verdicts (acquitted vs. guilty) for s, s=1,...,S, subpopulations or crimes, and r evaluation techniques, r=1,..R, there are (2^R -1) ×S degrees of freedom from which to estimate (2R+1) × S parameters. To generalize this relationship further, define:

V_r As the number of possible response outcomes for evaluator r (e.g. for three outcomes, $V_r = 3$)

d As the number of possible true categories that the subject can belong to. Note that in general d = V_r, for all r = 1,..,R. In all the examples studied in this paper, d has been equal to two for guilty and acquitted.

Then, the degrees of freedom and the number of parameters associated with a $V_1 \times V_2 \times ... V_R$ table is as follows:

$$Df = \left[\left(\prod_{r=1}^{R} V_r\right) - 1 \right] \times S$$

$$Parameters = \left[\left(\sum_{r=1}^{R} (V_r - 1) \times d \right) + (d-1) \right] \times S$$

Note that in the above expression for the number of parameters, that the first summation term represents the number of unique classification/error rates defined for this structure and the (d-1) term represents the number of unique prevalence rates for a given subtables. For example, if the jury has three outcome categories (V_1 = 3) as presented in section IV, and the judge two outcome categories (V_2 = 2), with d=2 for dichotomous underlying classifications, we have (3 × 2) - 1 or five degrees of freedom for each subpopulation. In contrast, we have (3-1) × 2 + (2-1) × 2, equal to 6, plus (2-1) or 7 parameters for each subpopulation. Hence in these situations, with only two evaluators, the data model is over-parameterized, and one must make some assumptions about the parameters in order to conduct estimation. Walter and Irwig (1988) indicated that with the availability of outcomes from three testing procedures, one can overcome the types of assumptions required in the original H&W method and the alternative design studied here. With three test results for each respondent the model is saturated for only one population and parameter estimation is possible. For example, in the generalized equation for three outcomes for the first evaluator (V_1 = 3) and two outcomes for the second and third evaluators (V_2 = 2, V_3 = 2) we have {(3 × 2 × 2) minus 1}= 11 × s degrees of freedom, and {(2 × 2) + (1 × 2) + (1 × 2) +(2-1)} = 9 × s parameters. Hence, if the study is designed to have three evaluators rather than two, complete parameter estimation is possible for individual crimes or any aggregate level. Such a procedure could be accomplished by having the a secondary expert or a retired judge sit in on the trial. The potential to have more than two evaluators is a unique to studies of this type. In the survey environment, conducting more than two interviews with the same respondent is unfeasible.

The data models this paper are developed under the assumption that the error rates from each testing procedure are conditionally independent given the true status. Therefore, to apply the model with the availability of three evaluators, we must be sure that the evaluators provide their verdicts independently of the others. A correlation in the error rates results in a bias in the estimates produced from the H&W method (Vacek, 1985).

The use of a third evaluator also allows the researcher to obtain a second assessment of the various factors as viewed by the judge in a study of this type, such as the clarity of evidence, the superiority of the lawyer, and the sympathetic nature of the defendant. Some limited data simulations have also shown that sample size of 1,000 cases should be sufficient to yield statistically significant estimates of the parameter values even for underlying error rates of .005 (α_3 = .005). However, further exploration on this issue is needed to determine the appropriate sample sizes required for a specified set of study goals.

VI. Closing Remarks

The application of the H&W method has shown to be an effective procedure for evaluating the error rates in diagnostic testing and in survey data collection procedures. In reanalyzing the K&Z data, we have found that the clarity of the evidence plays a substantial role in the level of disagreement as well as the seriousness of the crime, the presence of a prior record and the sympathetic nature of the defendant. These differences in the level of disagreements by crime hinder the pairing of crimes or data groups that is required in the application of the H&W method. However, by modifying the data structure suggested by the H&W method, we have shown that reliable estimates of the error rates can be developed. In particular, we have estimated the level of leniency in the jury verdicts for the sample of cases in the K&Z study and found the jury levels to be considerably higher than the judge. Given the apparent negligible level of false positives (an innocent person is found guilty) we were unsuccessful in estimating these rates with the sample sizes available.

Our paper is the first use of the H&W diagnostic test paradigm to analyze judge-jury data. Further exploration of the analytical procedures and their application to judge-jury agreement data is needed. In future studies of jury verdicts, we suggest a third evaluator and have provided a technique for analyzing the data.

References:

Brookmeyer, R. and Gail, M. (1994), *Aids Epidemiology, A Quantitative Approach*, New York: Oxford Press.

Gastwirth, J. L. (1987), "The statistical precision of medical screening procedures: Application to polygraph and AIDs antibodies test data," *Statistical Science*, (3), 213-238.

Gastwirth, J. L. (1988), *Statistical Reasoning in Law and Public Policy*, Orlando FL: Academic Press

Hui, Sui L. and Steven D. Walter, (1980) "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36,167-171.

Kalven, Harry Jr. and Zeisel, Hans, (1966) *The American Jury*, Boston Little Brown and Company

Sinclair, M. D. (1994), "Evaluating reinterview survey methods for measuring response errors," Ph.D. Dissertation, George Washington University.

Sinclair, M. D. and Gastwirth, J. L. (1994) "On Designs for Measuring Prevalence Rates and Error Rates from Diagnostic Testing Procedures," *Proceedings of the 1994 Summer Conference of the American Statistical Association, Bio-pharmaceutical Section*.

Sinclair Michael D. and Joseph L. Gastwirth (1996) "On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data," *Journal of the American Statistical Associati*on, Vol 91 pgs 961-969.

SAS Institute Inc. (1988), *SAS/STAT User's Guide* (Release 6.03), Cary, NC: Author.

Walter, S.D. and Irwig, L.M. (1988). Estimation of Test Error Rates, Disease Prevalence, and Relative Risk from Misclassified Data: A Review. *Journal of Clinical Epidemiology* 41, 923-937

Vacek, Pamela M., (1985) "The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests," *Biometrics* 41, 959-968.

Wolter, Kirk M., *(1985) Introduction to Variance Estimation* New York: Springer-Verlag.

Table 2
Modified Full Analysis of Table 146 Jury Request vs. No Request

Parameter	Description		Estimate	Estimated Standard Error
Jury Error Rates No Request	Classified as	True Status		
β_{121}	Hung	Guilty	.0469	.0039
β_{131}	Acquitted	Guilty	.1908	.0081
β_{113}	Guilty	Acquit	.1169	.0791
β_{123}	Hung	Acquit	.0780	.0114
C1	Ratio of Jury Error Rates with Request to no Request		1.334	.0996
C2	Ratio of Judge Error Rates to Jury Error Rates (no request)		0.000	.1361
π	Estimated Prevalence of Guilt		.8345	.0200

Table 3
Modified Reduced Model Analysis of Table 146 Jury Request vs. No Request

Parameter	Description		Estimate	Estimated Standard Error
Jury Error Rates No Request	Classified as	True Status		
β_{121}	Hung	Guilty	.0470	.0038
β_{131}	Acquitted	Guilty	.1912	.0076
β_{113}	Guilty	Acquit	Assumed to be negligible	
β_{123}	Hung	Acquit	.0843	.0128
C1	Ratio of Jury Error Rates with Request to no Request		1.317	.0973
C2	Ratio of Judge Error Rates to Jury Error Rates (no request)		0.1623	.0195
π	Estimated Prevalence of Guilt		.8612	.0060