Journal of Wine Economics, Volume 12, Number 4, 2017, Pages 354–362
doi:10.1017/jwe.2017.14

https://doi.org/10.1017/jwe.2017.14 Published online by Cambridge University Press

Opinions versus Facts: A Bio-statistical Paradigm Shift in Oenological Research*

Dom Cicchetti [a]

Abstract

A substantial oenological literature exists on opinions of experts and neophytes as they relate to opinions about the quality of wines (Ashenfelter and Quandt, 1998; Cicchetti, 2004; Lindley, 2006). These opinions can be contrasted with factual binary questions about wine: Is it oaked? Does it contain sulfites? Is it filtered? Is the grape varietal Cabernet Sauvignon or Cabernet Franc? Syrah or Grenache? Pinot Noir or Gamay? Such factual binary issues are examined within the broader context of the various measures of factual judgment: Overall Accuracy (OA), Sensitivity (Se), Specificity (Sp), Predicted Positive Accuracy (PPA), and Predicted Negative Accuracy (PNA). The resulting biostatistical methodology derives from bio-behavioral diagnostic research investigations. The purpose of this report is to apply this methodology to the discipline of oenology to compare wine judgments with wine facts. Using hypothetical examples, wine judges' classifications of wines as oaked or unoaked were analyzed for their degree of accuracy. The results show that OA is a poor measure of the accuracy of binary judgments relative to Se, Sp, PPA, or PNA. The biostatistics of the problem could have wide-ranging applications in the design of future oenological research investigations, and in scientific research more broadly. (JEL Classifications: C1, L15, Q13)

Keywords: acccuracy, Binary Tasting Judgments, oenology.

I. Introduction

A substantial oenological literature exists on the opinions of experts and neophytes as they rate the quality of wines (Ashenfelter, 1998; Cicchetti, 2004; Lindley, 2006). In distinct contrast, the purpose of this article is to present a new methodology for examining the extent of agreement between wine tasters in their binary judgments of wine characteristics—that is, in providing answers to the following types of questions: Is a wine oaked? Does it contain sulfites? What is the grape varietal? Cabernet Sauvignon

* A brief summary of this research was presented by the author at the 2016 meeting of the American Association of Wine Economists (AAWE) in Bordeaux, France.

[a] Yale University School of Medicine, Department of Biometry, Yale University Home Office, Box 317, North Branford, CT 06471; e-mail: dom.cicchetti@yale.edu.

or Cabernet Franc? Grenache or Syrah? Pinot Noir or Gamay? When comparing this hypothetical example to an actual oenological research investigation, a number of other factors would, per force, need to be experimentally controlled, such as the type of oak, previous barrel use, the duration of exposure to oak, and other relevant factors. Similar reasoning would also apply to any binary variable that is of oenological interest. The reviewer's recognition of this caveat is appreciated.

The motivation for this research derives from several sources: participation in wine-tasting studies focusing on differentiating various wine varietals (e.g., at annual meetings of the New York and New World Wine Experiences, empirical studies in the science of wine (Goode, 2014), and some of the recent research undertaken by the author (Cicchetti and Cicchetti 2009; 2014).

An anonymous reviewer notes that some oenological research with a focus upon factual variables has been published, citing several relevant investigations. For example, Frøst and Noble (2002) study the influence of a taster's factual knowledge and level of expertise on preferences for certain characteristics of red wines, such as typical aromas and tastes. A second oenological investigation compares support vector machines (SVMs)—"supervised learning methods used for classification"— and three types of neural networks (NNs) for predicting the quality of wine based upon physiochemical data that include such variables as levels of alcohol, sulfates, citric acid, and fixed acidity (Nachev and Hogan, 2013, p. 310). The authors find that SVMs outperform each of three versions of NNs. In a third oenological investigation, Cortez et al. (2009) compare SVM to NN methods using a multiple-regression methodology to predict wine-taster preferences, including such variables as measures of acidity, residual sugar, chlorides, sulfates, and other physicochemical properties in red and white wines. Once again, SVMs outperform their methodological competitors.

These results are comparable to the findings of non-oenological biobehavioral diagnostic studies that have been reported earlier by the author and colleagues (Cicchetti et al., 1995). The results of this investigation show that NN is outperformed by Logistic Regression (LR), Linear Discriminant Function Analysis (LDFA), and Quadratic Discriminant Function Analysis (QDFA) in the diagnosis of the presence or absence of autism by 15 International Classification of Diseases (ICD-10) criteria. This result is consistent with the author's initial critique of NN research (Cicchetti, 1992) and with the results of an earlier study by Fletcher et al. (1978). These latter investigations reveal that a critical factor in applying NN and competing methodologies is not the sample size, per se, but rather the ratio of the number of *subjects* to the number of predictor *variables*, also known as the S/V ratio. Additionally, *shrinkage* is defined as the percentage loss in classification accuracy between the training set and the cross-validation results. If the training-set level of accuracy is 90% and the cross-validation level is 80%, this would indicate a 10% level of shrinkage. Thus, Fletcher and colleagues show that when LDFA is applied to classification studies in neuropsychology research, the S/V ratio is of critical importance: With an S/V of 1:1, the shrinkage estimates from test to cross-validation fell within a very narrow range of 34% to 36%, whether the sample sizes were 10, 25, or 50. For an S/V of 2:1, with sample

https://doi.org/10.1017/jwe.2017.14 Published online by Cambridge University Press

https://doi.org/10.1017/jwe.2017.14 Published online by Cambridge University Press

sizes of 20, 50, and 100, shrinkage estimates vary from 21% to 23%; for a 3:1 ratio, with sample sizes of 30, 75, and 150, the shrinkages were all 17%; when the S/V was 4:1, with sample sizes of 40, 100, and 200, the corresponding levels of shrinkage ranged between 13% and 15%; and, finally, for S/V ratios of 5:1, with sample sizes of 50, 125, and 250, the cross-validation shrinkages ranged between 9% and 12%. These data clearly indicate that sample sizes are of negligible importance in comparison to the size of the S/V in the design of classification research, whether in the arena of oenology research or more broadly. It should be noted that the S/V ratios in the Cicchetti et al. (1995) investigation are 15:1 for the autistic subjects and 17:1 for the non-autistic subjects. The S/V ratios in the Nachev and Hogan (2013) and Cortez et al. (2009) investigations are in excess of 50:1; one therefore assumes that the resulting shrinkage levels from the training to the cross-validation sets are very low.

In addition to the relative dearth of research on oenophiles' judgments in the factual realm, there is a second serious problem in the literature: the absence of a defensible biostatistical strategy to apply to this neglected realm of important oenological research. This second area of focus defines the main thrust of this report, with the resulting approach best understood as it applies to the accuracy of assessing oenological facts. The next section focuses on recommended biostatistical methods for providing answers in the study of oenological binary variables.

II. Hypothetical Oenophiles Distinguish between Oaked and Unoaked Wines

Consider that 10 apocryphal wine lovers consent to participate in a tasting with the objective of differentiating between five oaked and five unoaked wines of the same vintage year from around the globe. How accurate are their judgments? Here, we have a binary judgment to answer the research question: Are the wines classified as oaked, yes or no? Borrowing from the broader fields of biomedical science, physics, and chemistry, the components of wine judgment would be Overall Accuracy (OA), Sensitivity (Se), Specificity (Sp), Predicted Positive Accuracy (PPA), and Predicted Negative Accuracy (PNA).

III. Defining the Five Components of Judgmental Accuracy

Overall Accuracy (OA) refers to the total percentage of correct judgments. Applying the criteria of Cicchetti et al. (1995) to judgmental accuracy, less than 70% rates as poor, 70% to 79% is fair, 80% to 89% is good, and 90% to 100% is excellent. (These criteria also apply to each of the remaining four measures of judgmental accuracy.) *Sensitivity* (Se) refers to the percentage of oaked wines that are correctly judged as such. *Specificity* (Sp) refers to the percentage of unoaked wines that are correctly judged as such. *Predicted Positive Accuracy* (PPA) refers to the percentage of wines judged as oaked that are actually oaked. Finally, *Predicted Negative Accuracy* (PNA) refers to the percentage of wines judged as unoaked that are actually unoaked.

https://doi.org/10.1017/jwe.2017.14 Published online by Cambridge University Press

Table 1

10 Tasters Classify 100 Wines as Oaked or Unoaked—Summary Data*

Correct Classification *Taster:*	*Oaked (+)*	*Unoaked (−)*	*Totals:*
Oaked (+)	40 (22.5)	5 (22.5)	**45**
Unoaked (–)	10 (27.5)	45 (27.5)	**55**
Totals:	**50**	**50**	**100**

OA = (40 + 45)/100 = 85%; Se = 40/50 = 80%; Sp **=** 45/50 = 90%; PPA **=** 40/45 = 89%; and PNA **=** 45/55 = 82%.

*Applying the criteria of Cicchetti et al. (1995) to levels of judged accuracy: < 70% = Poor; 70%–79% = Fair; 80%–89% = Good; and 90%–100% = Excellent; the figures in parentheses refer to the proportion of accurately judged wines expected on the basis of chance alone (PC); the method used to obtain the four figures in parentheses is the same one used in the chi-square (d) test. To obtain the overall taster agreement with the correct classification, one simply adds the two figures in parentheses that appear on the main diagonal and then divides the result by 100; PC then becomes (22.5 + 27.5)/100 = .50, or 50%.

The simulated summary data across the 10 wine judges appear in contingency table format, as shown in Table 1.

In binary classification data, such as in Table 1, the off-diagonal cases represent two types of errors, classified as Type I and Type II. The Type I errors can also be conceptualized as false positives, or the extent to which a true negative result (i.e., the wine is unoaked (–)) is misjudged as a positive one (i.e., the wine is oaked (+)). In distinct contrast, the Type II error occurs when the wine is incorrectly judged as oaked (+) (i.e., the wine is actually unoaked (– –)). This type can also be referred to as a false negative.

Applying this reasoning to the hypothetical summary data in Table 1, five unoaked wines (negative for oak) are misclassified as oaked (positive for oak). These results represent the tasters' false-positive judgments (Type I errors); correspondingly, 10 oaked (+) wines have been misclassified as unoaked (–). These results represent false-negative judgments (Type II errors).

These hypothetical wine data can also be expressed separately for each of the judges; see Tables 2 and 3. Again, less than 70% rates as poor, 70% to 79% is fair, 80% to 89% is good, and 90% to 100% is excellent in terms of wine-judging accuracy (Cicchetti et al., 1995).

The notations presented in Table 3 are defined as follows:

(++) means the taster judges the wine as oaked (+), and the correct judgment is oaked (+).

(− −) means that the taster judges the wine as unoaked (– –), and the correct judgment is unoaked (– –).

(+ −) means the taster judges the wine as oaked (+), but the correct judgment is unoaked (–).

(− +) means the taster judges the wine as unoaked (–), but the correct judgment is oaked (+).

Table 2
Each of 10 Tasters Classifies 10 Wines as Oaked or Unoaked

Taster:	*PO:*	*(+ +)*	*(− −)*	*(+ −)*	*(− +)*	*Marginals*		*Se*	*Sp*	*PPA*	*PNA*
Five	100	5	5	0	0	5–5;	5–5	100	100	100	100
Six	100	5	5	0	0	5–5;	5–5	100	100	100	100
Three	90	5	4	1	0	5–5;	6–4	100	80	83	100
One	90	4	5	0	1	5–5;	4–6	80	100	100	83
Seven	90	4	5	0	1	5–5;	4–6	80	100	100	83
Two	80	5	3	2	0	5–5;	7–3	40	100	29	100
Four	80	5	3	2	0	5–5;	7–3	100	60	71	100
Eight	80	3	5	0	2	5–5;	3–7	60	100	100	71
Nine	70	2	5	0	3	5–5;	2–8	40	100	100	63
Ten	70	2	5	0	3	5–5;	2–8	40	100	100	63
Totals:		**40**	**45**	**5**	**10**	**50–50;**	**45–55**				

Marginals occur in two sets, as they appear in Tables 2 and 3. The first set for each case refers to the numbers of oaked (+) and unoaked (–) wines. For Case 1, 48 of the wines are oaked (+), and 52 are unoaked (–). The second set for each case refers to the wine tasters' judgments as to whether a wine is oaked (+) or unoaked (–). For Case 1, 51 wines are judged to be oaked (+), and 49 are judged to be unoaked (–). The same design holds for Cases 2 through 8.

PC refers to the extent of agreement between a given taster and the correct classification concerning whether a wine is oaked. The calculation is the same as for the familiar and venerable chi-square (d) test; the PC calculation method is illustrated in Table 1.

IV. Why Overall Accuracy Is an Invalid Indicator of Judges' Evaluations

OA is not an adequate measure of wine judges' accuracy for several key reasons. Because it is an omnibus measure, a high level of OA (≥80%) can be associated with a wide range of levels of the remaining four components of judged accuracy. The simulated data in Table 2 strongly support this caveat. The OA level of 85% is considered good; the OA across the 10 apocryphal wine judges ranges between 70% (barely acceptable or fair) and 100% (perfect). For Wine Judges 5 and 6, Se varies between 40% and 100%, Sp between 60% and 100%, PPA between 29% and 100%, and PNA between 63% and 100%. The simulated data in Table 3 support the same argument: Depending on how the results distribute themselves to produce the same level of OA (here again, a value of 85%), Se values vary between 0% and 88%, and PPA values range similarly from 0% to 82%. The distribution across the four accuracy indices (Se, Sp, PPA, and PNA) is best for Case 1 (88%, 83%, 82%, and 88%, respectively) and worst for Case 8 (0%, 90%, 0%, and 93%, respectively).

https://doi.org/10.1017/jwe.2017.14 Published online by Cambridge University Press

Table 3
Eight Faces of Overall Accuracy in Judging Oaked and Unoaked Wines

Case	*PO*	*(++)*	*(– –)*	*(+ –)*	*(– +)*	*Marginals*	*PC*	*Se*	*Sp*	*PPA*	*PNA*
One	85	42	43	9	6	(.48-.52; .51-.49)	.50	88	83	82	88
Two	85	27	58	9	6	(.33-.67; .36-.64)	.55	82	87	75	91
Three	85	20	65	9	6	(.26-.74; .29-.71)	.60	77	88	69	92
Four	85	15	70	9	6	(.21-.79; .24-.76)	.65	71	89	63	92
Five	85	11	74	9	6	(.17-.83; .20-.80)	.70	65	89	55	93
Six	85	7	78	9	6	(.13-.87; .16-.84)	.75	54	90	44	93
Seven	85	4	81	9	6	(.10-.90; .13-.87)	.80	40	90	31	93
Eight	85	0	85	9	6	(.06-.94; .09-.91)	.86	0	90	0	93

*The clinical or practical significance of Se, Sp, PPA, and PNA values can be interpreted according to the criteria of Cicchetti et al. (1995) regarding levels of accuracy in differentiating between oaked and unoaked wine, wherein < 70% = Poor; 70%–79% = Fair, 80%–89% = Good; and 90%–100% = Excellent.

V. Implications for Future Research Investigations of Factual Binary Variables

The hypothetical data in Table 1 provide important information for future research studies designed to test wine judges' ability to correctly answer binary questions, such as whether a wine is oaked, whether one can distinguish Grenache from Syrah or filtered from unfiltered wines, and so forth. The simulated data in Table 3 indicate that the best design includes each binary variable with approximately 50% frequency. This advice also has implications across diverse fields of the behavioral and biomedical sciences and biostatistics—in fact, wherever a gold standard is available. The methodology and caveats also apply to diagnostic areas in which proxy gold standards are used, such as best clinical judgment in lieu of an unavailable gold standard (e.g., in judging the validity or accuracy of the binary diagnosis of autism (Cicchetti et al., 1995) and other biomedical disorders (Feinstein, 1987)). The term *gold standard* can be defined in relative or absolute terms. Perusing the literature, the more accurate definition, in this author's opinion, is more in accord with a relative concept. In this important regard, the definition offered by Versi (1992) in a letter to the editor of *British Medical Journal* makes eminently good sense. Versi regards the gold standard as not the perfect test but the best available at a given moment in time, which means it can and will be replaced by a better test once one becomes available.

If one examines the hypothetical data in Table 3, two phenomena become apparent: First, the driving force behind the varying levels of PC, Se, Sp, PPA, and PNA is the extent of maldistribution among the proportions of cases that are correctly judged as positive (oaked) or negative (unoaked); and second, as the level of the maldistribution increases, so does the dissimilarity in the values of PC, Se, Sp, PPA, and PN. It is also clear that the distribution of positive and negative cases produces the best results when their apportionment is as close to 50% as possible, as when the correct classification is 48% (+) and 52% (–).

Table 4

Correlating 85% Agreement Levels for Oaked/Unoaked Wines with: Chance Agreement Levels (PC) and Ranges of Se, Sp, PPA and PNA Values.

Correct Classification:				
(+)	*(–)*	*Difference (D):*	*PC:*	*RV (%):*
48	52	4	0.50	6
33	67	34	0.55	9
26	74	48	0.60	15
21	79	58	0.65	29
17	83	66	0.70	38
13	87	74	0.75	49
10	90	80	0.80	62
6	94	88	0.86	93

Note: The Pearson correlation between D and PC is + 0.96, and that between PC and RV is 0.97. These nearly perfect correlations represent very large Effect Sizes (ES) according to Cicchetti's (2008) extended ES categories of Cohen (1988).

As shown in Table 4, the Pearson correlation between the maldistribution of positive and negative cases and PC is nearly perfect, at 0.96; the Pearson correlation between PC and the range of percentages across Se, Sp, PPA, and PNA is also nearly perfect, at 0.97.

These findings provide strong support for designing any oenological investigation based upon binary factual variables such that 50% of the wines represent each of the two sides of the binary/winery coin—namely, positive (+) and negative (–).

VI. Conclusions

The purpose of this report is to utilize methodology from the fields of biobehavioral medical and physical sciences to compare wine judgments to binary wine facts. In a hypothetical example, wine judges' classifications of wines as oaked or unoaked are analyzed for their degree of accuracy. The biostatistics of the problem have broad applications in future oenological research investigations as well as in other scientific disciplines; the results are in distinct contrast to the wider area of research in which two or more judges' opinions about the quality of wine are compared. As the anonymous reviewer notes, the sample size is import in the design of oenological research. Three variables that loom large are palate fatigue/dead palate, the scheduled time of day for the tastings, and the known taster variations in recognition threshold levels for the perceived levels of sugar in any given wine (Cicchetti and Cicchetti, 2008). Once these factors are controlled, a valid and simple test allows one to accurately estimate appropriate sample sizes (Kraemer and Thiemann, 1987). Finally, a more comprehensive report on the reliability and accuracy of binary diagnoses has recently been published by the author and colleagues (Cicchetti, Klin and Volkmar, 2017); these findings are currently under evaluation for their relevance to oenological research.

Acknowledgment

This is to acknowledge the valuable suggestions of an anonymous reviewer concerning the often unrecognized complexity of oenological binary variables, sample-size considerations, and references to additional oenological research for its relevance to this investigation. These comments are appreciated and have, along with other related issues, been woven into the fabric of the current version of the revised text.

References

Ashenfelter, O., and Quandt, R. (1998). Analyzing a wine tasting statistically. *Chance*, 12(3), 16–20.

Cicchetti, D. V. (1992). Neural networks and diagnosis in the clinical laboratory: State of the art. *Clinical Chemistry*, 38, 9–10.

Cicchetti, D. V. (2004). Who won the 1976 blind tasting of French Bordeaux and U.S. Cabernets? Parametrics to the rescue. *Journal of Wine Research*, 15, 211–220.

Cicchetti, D. V. (2008). From Bayes to the just noticeable difference to effect sizes: A note to understanding the clinical and statistical significance of oenological research findings. *Journal of Wine Economics*, 3, 185–193.

Cicchetti, A. F., and Cicchetti, D. V. (2008). The balancing act in consistent wine tasting and wine appreciation: A tale told by two brothers. Part I. Consistency in wine tasting and appreciation: A personal-experiential perspective. *Journal of Wine Research*, 19, 115–121.

Cicchetti, D. V., and Cicchetti, A. F. (2009). Wine rating scales: Assessing their utility for producers, consumers, and oenologic researchers. *International Journal of Wine Research*, 1, 73–83.

Cicchetti, D. V., and Cicchetti, A. F. (2014). Two oenological titans rate the 2009 Bordeaux wines. *Wine Economics and Policy*, 3(1), 28–36.

Cicchetti, D. V., Klin, A., and Volkmar, F. R. (2017). Assessing binary diagnoses of bio-behavioral disorders: The clinical relevance of Cohen's Kappa. *Journal of Nervous and Mental Disease*, 205(1), 58–65.

Cicchetti, D. V., Volkmar, F. R., Klin, A., and Showalter, D. (1995). Diagnosing autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. *Child Neuropsychology*, 1(1), 26–37.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547–553.

Feinstein, A. R. (1987). *Clinimetrics.* New Haven, CT: Yale University Press.

Fletcher, J. M., Rice, W. I., and Ray, R. M. (1978). Linear discriminant function analysis in neuropsychological research: Some uses and abuses. *Cortex*, 14, 564–577.

Frøst, M. B., and Noble, A. (2002). Preliminary study of the effect of knowledge and sensory expertise on liking for red wines. *American Journal of Enology and Viticulture*, 53(4), 275–284.

Goode, J. (2014). *The Science of Wine: From Vine to Glass.* 2nd ed. Berkeley: University of California Press.

Kraemer, H. C., and Thiemann, S. (1987). *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage.

Lindley, D. V. (2006). Analysis of a wine tasting. *Journal of Wine Economics*, 1(1), 33–41.

Nachev, A., and Hogan, M. (2013). Using data mining techniques to predict product quality from physicochemical data. *Proceedings of the International Conference on Artificial Intelligence*, 1, 308–314.

Versi, E. (1992). "Gold standard" is an appropriate term. *British Medical Journal*, 305, 187 (Letter to the Editor).

https://doi.org/10.1017/jwe.2017.14 Published online by Cambridge University Press