Journal of Wine Economics, Volume 12, Number 4, 2017, Pages 363–369
doi:10.1017/jwe.2017.16

The Distribution of Ratings Assigned to Blind Replicates*

Jeffrey C. Bodington [a]

Abstract

The inability of many wine judges to achieve perfect consistency by assigning the same rating to the same wine in a blind tasting is well established. Results for four wine tastings that include blind replicates are examined in this article. Although perfection is rare, the probability distributions of those results show that wine judges do tend to assign closer ratings to replicates than is likely due to chance alone. Approximately one-third of judges assign ratings that are within one rank of perfect consistency, and two-thirds assign ratings within two ranks of perfect consistency. This finding is sensitive to judges' capabilities, the mechanics of the tasting protocol, and the extent to which the replicate is different from other wines in the tasting. Much wine-related research to date takes judges' individual ratings as deterministic, yet these results show that those ratings are stochastic. These results yield a probability distribution that may guide future research concerning the uses and economic implications of wine ratings. (JEL Classifications: A10, C10, C00, C12, D12)

Keywords: blind, random, replicates, statistics, wine tasting.

I. Introduction

Blind replicates are unidentified double or triplicate samples of the same wine that are among the wines tasted by judges during a wine competition or assessment. They are used to test the consistency of scores or ranks assigned by judges, and they eliminate the complexity of differences between wines, because every wine in the replicate is by definition the same. Blind replicates also erase the complexity of differences between judges, because results can be evaluated intra- rather than inter-judge. Hodgson (2008) offers a famous analysis of wine judges' assessments of blind

*The author thanks Robert Hodgson for providing California State Fair Commercial Wine Competition data and essential explanations. The author also thanks Deborah Parker Wong and an anonymous reviewer for their perceptive and constructive comments. All remaining errors are the responsibility of the author alone.

[a] Bodington & Company, 50 California St., San Francisco, CA 94111; e-mail: jcb@bodingtonandcompany.com.

replicates, Ashton (2012) provides a review of three tastings with replicates, and Hodgson and Cao (2014) and Cicchetti (2014) also evaluate replicates.

The works referenced above show that a wine judge with near-perfect consistency, one who assigns the same score or rank to the same wine every time, is at best rare. This short article accepts and builds on that finding to examine the probability distribution of blind replicates. What is the shape of the distribution? How does the shape compare to that of a random distribution? What does the distribution imply about the information content of wine-assessment results and the ratings published by Robinson, Parker, Suckling, and others? Four sets of wine-tasting results that include blind replicates are described in Section II. The probability distributions of those results are evaluated in Section III, and conclusions and discussion follow in Section IV.

II. Wine-Tasting Data with Replicates

Four sets of wine-tasting results are described below. These four are employed because the data are available to ease replication and because the tastings were conducted according to published and blind wine-tasting protocols. Together, the sets include a large sample, small samples, quality ranks, and preference ranks. The data and MATLAB code written by the author to evaluate the distributions are available on request.

First, Hodgson (2008) describes his initiative to include blind triplicates among the wines entered in the California State Fair Commercial Wine Competition (CSF). The protocol for the CSF is described in Hodgson and Cao (2014, 64, 67–68), and I thank Hodgson for providing CSF data for 2005 through 2010. In flights of approximately 15 wines, panels composed of four or five judges tasted and awarded a medal to each wine before discussing them; each judge then had the option to revise the medal awarded to each wine. The medal awarded, or assigned, to each wine by each judge was a conclusion about absolute quality. The raw first-of-two-assignments data for 2010 evaluated herein contain medals at 10 levels of quality: No Award, Bronze–, Bronze, Bronze+, Silver–, Sliver, Silver+, Gold–, Gold, and Gold + . There were 72 triplicates among 2,929 wines. Gold+ down through No Award is an ordered set of 10 award-level ranks. Judges could, and often did, assign the same medal to several wines that had the same quality; thus, the judges' assignments of awards were with replacement. For graphical display in Section III, I converted the award medal levels to ordinal ranks, from the highest rank of 1 to the lowest rank of 10.

Second, Cicchetti (2014) reports results for a wine tasting conducted at the 2013 meeting of the American Association of Wine Economists (AAWE) in Stellenbosch, South Africa; the protocol is also described (3–4). Fifteen judges undertook a blind tasting of eight Sauvignon Blanc wines followed by eight Pinotage wines. Each flight contained a triplicate. Judges were instructed to score

each wine according to the absolute *Wine Spectator* scale: Poor/Unacceptable (for score range 50–69), Fair/Mediocre (70–79), Good/Above Average (80–89), and Excellent/Superior (90–100). Based on that ordered set of categories and the frequency distribution of scores shown in Appendix A, I converted the AAWE judges' scores for graphic display to ranks 1 (for score range 95–100), 2 (90–94), 3 (85–89), 4 (80–84), 5 (75–79), 6 (70–74), 7 (65–69), 8 (60–64), 9 (55–59), and 10 (50–54).

CSF and AAWE instructed judges to assign medal levels or scores according an assessment of quality. If a judge concluded that two wines had the same quality, he or she assigned the same medal or score to both wines. Statistically, assignment was then with replacement. In contrast, two other tasting protocols instructed judges to assign ranks according to relative preference, with no ties allowed. According to that protocol, assignment was without replacement. Princeton-based Liquid Assets conducted a blind tasting of eight 1970 Bordeaux wines (http://liquidasset.com/, Report #52). The flight contained one double, and eight judges each assigned a preference rank to each wine, with no ties allowed. San Francisco–based FOG conducted a blind tasting of six Syrah wines during 2017. The flight contained one double, and nine judges each assigned a preference rank to each wine, again with no ties allowed. The protocol for FOG is described in Bodington (2012, 182; 2015, 35).

In sum, four sets of blind replicates are evaluated in Section III. Two involve assignment with replacement, two involve assignment without replacement, and results for all four are expressed as ranks.

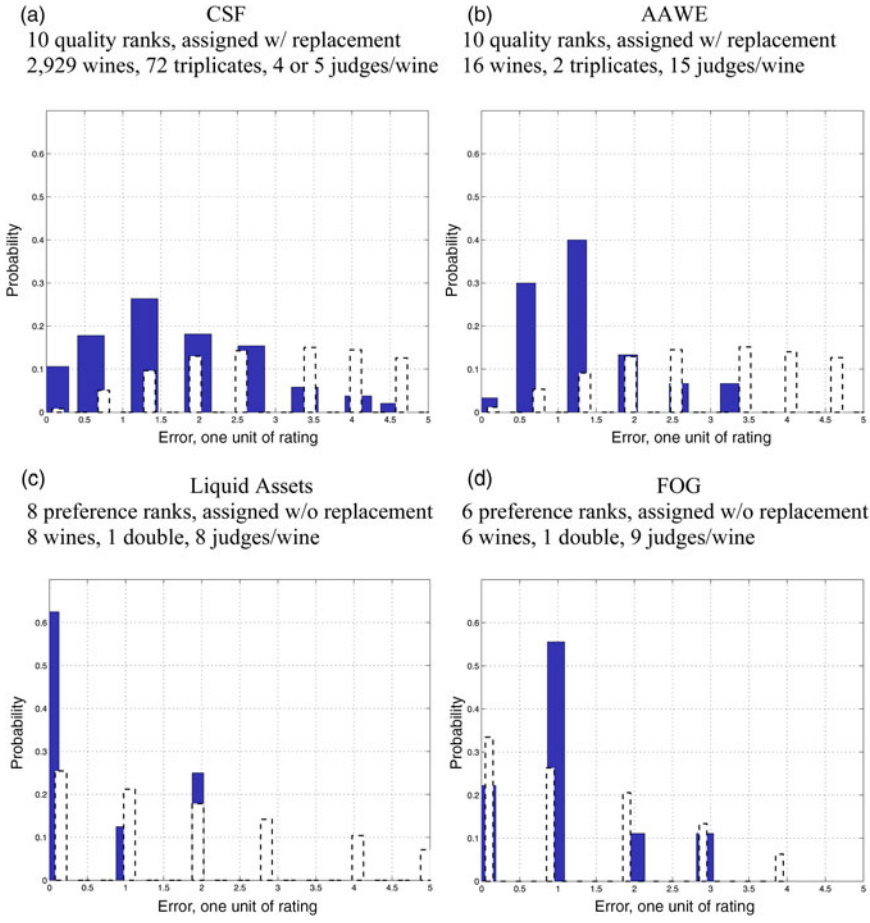
III. Distributions of Blind Replicates

The probability distributions of judges' ranks for the replicates described above are shown as solid vertical bars in Figure 1. For CSF in Figure 1(a) and AAWE in Figure 1(b), the horizontal axis is the average absolute difference in each judge's ranks. Perfect consistency is thus an average error of zero. For Liquid Assets in Figure 1(c) and FOG in Figure 1(d), judges assigned ranks that indicated relative preference, and again, as described above, assignment was without replacement. In those cases, due to assignment without replacement, judges could not assign the same rank to two wines. Perfect consistency in those cases is thus adjacent ranks. The horizontal axes in Figures 1(c) and 1(d) are then the difference in ranks minus unity so that, in all four figures, zero error on the horizontal axis indicates perfect consistency. Finally, the dashed line bars in each figure show the distribution of error, as though judges had assigned ranks randomly.

Figure 1 shows that, in aggregate, the wine judges did not assign ranks randomly. None of the four sold-bar distributions of observed ranks is a close match to the dashed-line-bar distributions of random ranks. Chi-square tests show that the p-values are < 0.01 for the null hypothesis that the distributions are the same. Some

Figure 1

Probability Distributions of Ranks Assigned to Blind Replicates



of the judges in each tasting did have perfect consistency, and most judges assigned ranks with lower error than is likely due to chance alone.

Although this finding is material and statistically significant, related findings illustrate difficulties in using replicates to assess judges' consistency. The observed distributions in Figures 1(a), 1(b), and 1(d) have a log-normal shape, but the one in Figure 1(c) does not. The methodologies for Figures 1(a) and 1(b) are similar, as are the shapes and ranges of the distributions. The methodologies for Figures 1(c) and 1(d) are similar, but the shapes and ranges of the distributions are not. That may be due, in part, to smaller sample sizes. In addition, the results in Figure 1(d) concern a tasting of six Syrah wines from one AVA that were made in a similar

style. The judges' notes show that most of the wines were difficult to differentiate from each other. One judge scored and ranked the wines; four of the six wines had tie scores, and she explained that the relative rank of those four was random. In that context, a low probability of perfect consistency in Figure 1(d) is not surprising. A replicate within a flight of otherwise very different wines, or a tasting protocol that allowed assignment with replacement, may have led to less error.

The statistics of experimental design also affect the distributions in Figure 1. The dashed-line bars show the distribution of error for random draws from a uniform distribution. Those shown are exact distributions calculated from 10,000 random draws according to each tasting protocol. In Figures 1(a) and 1(b), the shape of the random distribution is concave due to the increasing number of combinations that yield larger-than-small errors and the decreasing number of combinations that yield the largest average errors. In Figures 1(c) and 1(d), the shape of the random distribution is decreasing. Assignment of ranks without replacement shrinks the sample space and thus increases the probability of illusory consistency. In sum, the details of the tasting protocol, including whether assignment is with or without replacement, can affect the distribution of observed error.

IV. Conclusion and Discussion

The probability distributions of ranks assigned by judges to blind replicates in four wine tastings show that approximately a one-third probability of error less than ± 1.0 rank and a two-thirds probability of error less than ± 2.0 ranks. Evidence that a judge is more consistent than average may support lower error. Evidence that other wines among the replicates are similar to the replicates may be another partial explanation for error. Assessments of statistical significance must reflect the mechanics of the wine-tasting protocol. Obtaining more precision than that without further inquiry appears perilous.

Much research is devoted to the wine ratings, and their economic implications, published by Robinson, Parker, Suckling, and others. For example, Ashton (2013) examines the pairwise correlations of ratings assigned to Bordeaux wines by six widely published wine critics. Stuen et al. (2015) evaluate the correlations of the ratings assigned to California and Washington wines in five well-known publications. Marks (2015) assesses the value to consumers of the information conveyed by wine judges' qualifications, scores, score-component weights, tasting notes, and other factors. Ashenfelter and Jones (2013) examine the demand for expert opinions about wines. Cardebat et al. (2014), Masset et al. (2015), Ashton (2016), and Oczkowski (2016) employ judges' ratings to examine various aspects of Bordeaux wine prices. Judges' ratings are also employed to confer awards on the wines entered in dozens of state, county, magazine, and other wine competitions each year.

While the applications above are clear that ratings often differ from judge to judge, they are silent or less formal about the possibility that each judge's own ratings could differ with repeated samples. The results in Section III concerning blind replicates

show that most judges' ratings are not deterministic – they are stochastic. If a judge were to resample a wine, then that judge may render a different rating; the distribution of those ratings would look something like the distributions in Figure 1. Unfortunately, the literature on stochastic ratings appears to be sparse. Neither Marden (1995) nor Alvo and Yu (2014) mention nondeterministic ranks in their widely referenced texts. Marley (1993) proposes evaluating the expectations of random ranks, but he enumerates limitations to that approach. Niu et al. (2013) employ the expectations of random ranks in large-sample search applications and, like Marley, find limitations. Examining the implications of judge-level stochastic ratings with regard to the economics of wine seems to be an evidence-based and worthwhile next step to take.

References

Alvo, M., and Yu, P. L. H. (2014). *Statistical Methods for Ranking Data*. New York: Springer.

Ashenfelter, O., and Jones, G. V. (2013). The demand for expert opinion: Bordeaux wine. *Journal of Wine Economics*, 8(3), 285–293.

Ashton, R. H. (2012). Reliability and consensus of experienced wine judges: Expertise within and between? *Journal of Wine Economics*, 7(1), 70–87.

Ashton, R. H. (2013). Is there consensus among wine quality ratings of prominent critics? An empirical analysis of red Bordeaux, 2004–2010. *Journal of Wine Economics*, 8(2), 225–234.

Ashton, R. H. (2016). The value of expert opinion in the pricing of Bordeaux wine futures. *Journal of Wine Economics*, 11(2), 261–288.

Bodington, J. C. (2012). 804 tastes: Evidence on preferences, randomness, and value from double-blind wine tastings. *Journal of Wine Economics*, 7(2), 181–191.

Bodington, J. C. (2015). Evaluating wine-tasting results and randomness with a mixture of rank preference models. *Journal of Wine Economics*, 10(1), 31–46.

Cardebat, J. M., Figuet, J. M., and Paroissien, E. (2014). Expert opinion and Bordeaux wine prices: An attempt to correct biases in subjective judgments. *Journal of Wine Economics*, 9 (3), 282–303.

Cicchetti, D. (2014). Blind tasting of South African wines: A tale of two methodologies. American Association of Wine Economists, Working Paper No. 164, August. Available at www.wine-economics.org/aawe-working-paper-no-164-economics/.

Hodgson, R. T. (2008). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2), 105–113.

Hodgson, R. T., and Cao, J. (2014). Criteria for accrediting expert wine judges. *Journal of Wine Economics*, 9(1), 62–74.

Liquid Assets (2002). http://liquidasset.com/.

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. London: Chapman & Hall.

Marks, D. (2015). Seeking the veritas about the vino: Fine wine ratings as wine knowledge. *Journal of Wine Research*, 26(4), 319–335.

Marley, A. A. J. (1993). Aggregation theorems and the combination of probabilistic rank orders. In: Critchlow, D. E., Fligner, M. A., and Verducci, J. S. (eds.), *Probability Models and Statistical Analyses for Ranking Data*. New York: Springer-Verlag. Chapter 12.

Masset, P., Weisskopf, J.-P., and Cossutta, M. (2015). Wine tasters, ratings, and *en primeur* prices. *Journal of Wine Economics*, 10(1), 75–107.

Niu, S., Lan, Y., Guo, J., and Cheng, X. (2013). *Stochastic rank aggregation. Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, Bellevue, WA, August 11–15.

Oczkowski, E. (2016). Identifying the effects of objective and subjective quality on wine prices. *Journal of Wine Economics*, 11(2), 249–260.

Stuen, E. T., Miller, J. R., and Stone, R. W. (2015). An analysis of wine critic consensus: A study of Washington and California wines. *Journal of Wine Economics*, 10(1), 47–61.

Appendix: AAWE, Frequency Distribution of Scores

Results show local peaks at 50, 55, 60, and so on through 95.