AMERICAN ASSOCIATION OF WINE ECONOMISTS

AAWE WORKING PAPER
No. 269
Economics

A MAXIMUM ENTROPY ESTIMATE OF UNCERTAINTY ABOUT A WINE RATING

Jeff Bodington

Dec 2021

www.wine-economics.org

A Maximum Entropy Estimate of Uncertainty about a Wine Rating

What can be deduced about the shape of a latent distribution from one observation?

Jeff Bodington[a]

Abstract

Much research shows that the ratings that judges assign to wines are uncertain and an acute difficulty in ratings-related research, and in calculating consensus among judges, is that each rating is one observation drawn from a unique and latent distribution that is wine- and judge-specific. A simple maximum entropy estimator is proposed that yields a maximum-entropy probability distribution for sample sizes of none, one, and more. A test of that estimator yields results that are consistent with the results of experiments in which blind replicates are embedded within flights of wines evaluated by trained and tested judges.

(JEL Classifications: A10, C10, C00, C12, D12)
Keywords: wine, judge, ratings, statistics, random

I. Introduction

Diverse research relies on the ratings that critics, judges, and consumers assign to wines. Recent examples include Gergaud *et al.* (2021) method of aggregating judges' ratings, Hölle *et al*. (2020) finding that customers' ratings of wines online can vary due to screen position alone, Corsi and Ashenfelter (2019) analysis of the correlation between weather data and ratings, Capehart (2019) analysis of whether or not training improves the accuracy of ratings, Lam *et al*. (2019) analysis of how ratings affect written reviews, and Malfeito-Ferreira *et al.* (2019) identification of the sensory and chemical differences between grand-gold and gold rated wines. In addition to such research, numerous critics, competitions, clubs, and vendors use ratings to compare wines, convey information, and sell wine.

This short article focuses on a difficulty with wine ratings for the uses above that is due the finding that each rating observed is one draw from a latent distribution that is wine- and judge-specific.

[a] Contact author at jcb@bodingtonandcompany.com.

What can be deduced about the shape of a latent distribution from one observation? Section II summarizes research showing that the ratings that critics and judges assign are stochastic, heteroscedastic, and may be affected by anchoring, expectations and serial position biases. Section III shows that those conditions lead to a problem that is inverse, ill posed with a sample size of one, and partially or wholly categorical rather than cardinal. A maximum entropy solution to that problem is posed in Section IV for sample sizes of none, one, and more. An example is presented in Section V and the results are consistent with the actual distributions observed for ratings assigned to blind replicates. Conclusions follow in Section VI.

II. The Maelstrom About a Rating Observed

Judges assign ratings to wines that are within a bounded set of scores, or an ordered set of categories, or ranks. Examples of scores include the 50- to 100-point scales used by *Wine Advocate* and *Wine Spectator*, U.C. Davis' zero to 20-point scale, Jacis Robinson's 12- to 20-point scale, and the zero- to 100-point scale prescribed by the International Organization of Wine and Vine's (OIV). Examples of categories include the Wine & Spirit Education Trust (WSET) six categories of quality (faulty, poor, acceptable, good, very good, and outstanding) and the California State Fair Commercial Wine Competition's (CSF) six or ten medals.[1] Further, some systems are forced rankings. If there are six wines in a flight, a judge must rank all six in order of relative preference. *Liquid Assets* and *San Francisco FOG* are examples of tasting groups who employ that approach. See reviews and comparisons of rating systems in Cicchetti & Cicchetti (2014), Kliparchuck (2013), and Veseth (2008).

Although tasters focus on the wine in the glass, much research shows that the ratings that they assign are affected by other factors. Some of the factors are supported by literature that is cited below. Other factors described below are reported as anecdotes and, when no literature is cited, those other factors are intended as hypotheses that remain to be tested.

[1] Depending on the year, the CSF has awarded six (No Award, Bronze, Silver, Gold-, Gold, and Gold+) or ten (No Award, Bronze–, Bronze, Bronze+, Silver-, Silver, Silver+, Gold–, Gold, and Gold+) ordered medals. The author holds a WSET Level III certification.

Stochastic Ratings

First, although wine ratings are not merely random, evidence that ratings are stochastic is abundant in the wine-related academic literature and trade press. Bodington (2017, 2020) summarizes and cites four experiments with blind replicates, more than 20 other evaluations that find uncertainty in ratings, and two texts that explain the neurological, physiological, and psychological reasons for variance in the ratings that the same judge assigns to the same wine. The stochastic nature of wine ratings is not unique. Kahneman *et al.* (2021, p. 80-86, 215-258) describe variance in wine ratings and in many other areas of human judgement including physicians' diagnoses, forensic experts' fingerprint identifications, and judges' sentencings of criminals.

The wine-related literature cited above supports several findings. A rating observed is one draw from a latent distribution, it is one instance of in some cases many potential instances. Ratings are heteroscedastic, so the distribution of ratings on a wine is wine- and judge-specific and different judges' ratings on the same wine are not identically distributed (ID). Some judges assign ratings more consistently than others, and some wines are more difficult to rate consistently than others. Research attempts to predict ratings from physiochemical properties have struggled to obtain statistical significance. Experiments with blind replicates show that, on average, the standard deviation of the rating that the same judge assigns to the same wine within a flight is approximately 1.3 out of 10 potential rating categories. And while some judges independently assign ratings that correlate well with each other, about 10% of CSF judges assign ratings that are indistinguishable from random assignments.

Although most ratings are assigned by judges prior to any discussion of the subject wines, pre-rating discussion sometimes occurs among panelists and some competitions require an initial rating, then discussion and then a post-discussion rating. According to Taber (2005, p. 300-301), discussion of the wines took place during the tasting at the 1976 Judgement of Paris. The CSF is an example of a competition in which judges assign an initial rating, discuss the wines with other judges and then assign another post-discussion rating. Both sets of ratings are reported to CSF officials and the author is not aware of any correlation or other comparisons made by the CSF.

Judges can influence each other so post-discussion ratings may not be statistically independent (I) and such ratings may be more highly correlated. When combined with the heteroscedasticity above, post-discussion ratings may not be statistically IID.

Anchoring, Expectation & Serial Position Biases

The score-based rating systems noted above also assign categories of quality or award to score thresholds and ranges. For example, the OIV system sets score thresholds for Bronze, Silver, and Gold medals at scores of 80, 85 and 90 respectively.[2] Bodington and Malfeito (2018) showed spikes in the frequencies of scores assigned just below those thresholds. Thus, some judges appear to anchor at OIV's category thresholds.

In addition to anchoring scores to category thresholds, there is anecdotal evidence of sequential anchoring. In a taste-and-score sequential protocol, a judge may assign a rating to the first wine and then rate the remaining wines "around" that anchor. A lag structure may also exist in which a judge rates around some composite of the most recent wines. The upper and lower bounds on ratings, whether numerical or categorical, may then merely bound a judge's assessments of relative preference.

Much research shows that judges' expectations affect the ratings they assign. Ashton (2014) found that judges assigned higher ratings to wines from New Jersey when told the wines were from California and lower ratings to wines from California when told the wines were from New Jersey. On that evidence, regardless of actual quality, an expectation of good quality may lead to a central tendency in ratings within whatever range of scores or categories indicates good quality. In addition, information provided about wines may alter expectations and ratings. For example, the pre-printed forms provided to CSF judges list the grape variety, vintage, alcohol by volume and

[2] The complete OIV award system is Bronze to wines with a mean score of at least 80 points (up to a maximum of 25% of all prized wines including Gold and Silver), Silver to wines with a mean score more than 84 points (up to a maximum of 12% of all wines entered), and Gold to wines with mean scores over 90 points (up to a maximum of 6% of all wines entered). A fourth medal, Great Gold, is awarded by a Grand Jury to the best wine in each of several categories (up to a maximum of 25% of the number of Gold medals).

residual sugar of a wine next to spaces where the judge writes in a comment and then a rating.[3] Whether or not such judgments should be represented as "blind" is open to debate. Whether or not having that information affects judges' ratings remains to be tested.

Some critics and competitions employ a sequential, step-by-step or taste-and-rate protocol. A critic or judge tastes a wine and assigns a rating, then tastes the next wine and assigns a rating to that wine, and so on. The Judgement of Paris, the CSF, and many publishing critics employ a sequential protocol.[4] The possibility that ratings assigned during taste-and-rate protocols are affected by serial position, rather than the intrinsic qualities of the wines and judges, is difficult assess and rule out. Serial position bias may occur in wine competitions due to palate fatigue, rest breaks, meal breaks, physiological and psychological factors.[5] There are anecdotal reports from judges who say there is temptation to assign a high rating to a dry and high-acid wine because it is refreshing in a sequence just after several off-dry and alcoholic wines. U.C. Davis' class for potential wine judges warns of position bias affecting differences in ratings due to the sequence of wines, breaks and lunch.[6] Filipello (1955, 1956, 1957) and Filipello and Berg (1958) conducted various tests using sequential protocols and found evidence of primacy bias. Mantonakis et al. (2009, p. 1311) found that "high knowledge" wine tasters are more prone than "low knowledge" tasters to primacy and recency bias. The sequence of wines tasted at the 1976 Judgment of Paris has never been disclosed so what effect position bias may have had on the results remains unknown.[7]

---

[3] Form provided to the author by the CSF on July 16, 2019.

[4] In contrast, some competitions employ an "open" protocol in which a flight is poured and judges can taste and re-taste the wines in any order and frequency. *Liquid Assets* and *San Francisco FOG* follow that open protocol.

[5] Serial position bias is common in many fields of judging. de Bruin (2005) examined singing and figure skating competition results and found position bias in both step-by-step and end-of-sequence sequential judging protocols.

[6] The author took the class and test for potential CSF judges at UC Davis.

[7] The Judgment's tasting protocol was sequential taste-and-score. The author confirmed, in email communications with both Mr. Taber and Mr. Spurrier, that the sequence of pour has never been disclosed.

Standing back, within the maelstrom described above, what can be deduced about a rating observed? A rating is a discrete score, ordered category or rank drawn from a bounded set. A rating is stochastic, and a rating observed is one draw from a wine- and judge-specific distribution. And that rating observed may have been affected by anchoring, expectation and/or serial position biases.

III. The Problem with One Observation

Except for rare experiments with blind replicates in a flight, critics and judges examine one wine (or each wine in a flight) and then assign one rating to that wine (or one rating to each wine in a flight). That is an acute problem. Literature cited above shows that ratings are heteroscedastic, ratings are not ID, so neither the collection of all judges' ratings on a wine nor the collection of one judge's ratings on other wines, can be employed to estimate the distribution of potential ratings by one judge on one wine. It is a unique distribution and only one sample drawn from that distribution is observed.

Score increments are usually whole numbers, some competitions allow half-points, so even score assignments are discrete. The scores, category choices and ranks used by critics, and allowed by competitions, are bounded sets. Although a rating is observed after it is assigned by a judge, aside from being discrete and bounded, no other information about the distribution of that rating is observed. Estimating the shape of the discrete and bounded distribution is thus an inverse problem. The shape, and any parameters describing that shape, must be inferred from the observation. And, unless the shape of the distribution can be defined by one parameter, the problem is ill posed. If ill posed, there are more unknown parameters than observations so a unique distribution can't be defined.

There is another difficulty. Although scores appear to be cardinal, the anchoring behaviors described above indicate that some judges mix the notions of cardinal scoring and ordered categories. The anchoring behaviors also indicate that some judges mix the notions of cardinal scoring and ranking. On that basis, treating scored ratings as cardinal appears perilous. And trying

to construct a model of a judge's scoring behavior, at minimum, exacerbates the inverse and ill-posed aspects of the problem.

The problem of what can be deduced about the shape of a latent distribution from one observation is not new. Several authors have examined what can be inferred about continuous and symmetric distributions; see Casella & Strawderman (1981), Rodriguez (1996), Golan, Judge & Miller (1996), Leaf, Hui & Lui (2009), and Cook & Harslett (2015). Their methods are summarized in Appendix A, and the discussion below focuses on what can be inferred from one observation about a bounded, discrete, ordinal scale, and probably asymmetric distribution.

IV. A Maximum Entropy Solution to An Inverse and Ill-Posed Problem

Hartley (1928) and Shannon (1948) posed the notion that uncertainty about something, in this case a judge's rating on a wine, can be expressed as information entropy. The higher the entropy, the higher the uncertainty. Jaynes (1957a, 1957b) then proposed the idea that a distribution that maximizes entropy assumes the least precision in what is known about the latent distribution of the data. When there aren't enough data to estimate a unique parameter, set that parameter to maximize entropy so that you don't pretend to know more than you actually know. This maximum entropy approach is often employed to solve inverse and ill-posed problems. See for example Golan, Judge & Miller (1997, p. 7-10 in particular).

Shannon (1948) defined the amount of information (i) in a random variable (x), in Equation (1), as the logarithm of the inverse of its probability (p). Building on that, he expressed information entropy (H), in Equation (2), as the expectation of information about a random variable. While the precise meaning of information entropy is controversial, H is highest for a uniform random distribution and lowest for a single-point, degenerate distribution. The informational distance (I) between two distributions, p and q, known as relative or cross-entropy, appears in Equation (3). See Rioul (2008) and Lombardi *et al.* (2015) for histories and discussions of these famous results.

$$i\big(p(x)\big) = ln\left(\frac{1}{p(x)}\right) = -\,ln\big(p(x)\big) \quad (1)$$

$$H = -\sum_{min}^{max} p(x)\, ln\left(p(x)\right) \quad (2)$$

$$I(p,q) = \sum_{min}^{max} q(x)\, ln\left({}^{q(x)}\!/_{p(x)}\right) \quad (3)$$

Applying Jaynes' notion of maximum entropy to none, one or more observations (x^o) drawn from a discrete and bounded but unknown distribution ($\hat{p}$) yields the solution proposed here in Equation (4) below. The first term $I(u,\hat{p})$ is the cross entropy between $\hat{p}$ and a uniform random distribution (u). The second term $n \cdot I(q|x^o,\hat{p})$ is the sample size (n) times the cross entropy between $\hat{p}$ and the actually-observed distribution ($q|x^o$). If $n = 0$ then the minimization in Equation (4) is a dual of the maximum entropy solution $u(all\ x) = 1/(max - min + 1)$.[8] If $n = 1$ then q is a one-hot vector where $q(x^o) = 1$. For $n = 3$, q could be the distribution observed for blind triplicates. As $n \to \infty$ then $\hat{p}$ tends to the distribution implied by a large sample and the influence of the random distribution, u, tends to zero.[9, 10] That minimization is the dual of a maximum likelihood solution for a large sample. See a derivation of Equation (4) in Appendix B and Golan, Judge & Miller (1996, p. 25, 41), Shashua (2008, p. 3-3) and Schapire (2014, p. 4) regarding the duality of maximum likelihood and minimum entropy.

$$\arg[\hat{p}] = argmin[\, I(u,\hat{p}) + n \cdot I(q|x^o,\hat{p})\,] \quad (4)$$

Among many potential PMFs for $\hat{p}$, a simple one is a categorical distribution that is discrete, bounded (*min, max*), and has a probability for every potential rating (r and a total of R ratings). $p(x = r) = p_{x=r}$ and $1 = \sum_{r=min}^{r=max}(p_{x=r})$. That PMF has $max - min + 1 = R$ unknown

[8] If $n = 0$ the minimization in Equation (4) solves to $\hat{p} = u$.

[9] As $n \to \infty$ the minimization in Equation (4) solves to a PMF where $\hat{p} \approx q|x^o$. For a categorical PMF with a probability parameter for every category, that minimization solves to $\hat{p} = q|x^o$.

[10] Substituting Equation (3) in to Equation (4) yields the more detailed expression that is implemented in the author's MATLAB code:

$$\arg[\hat{p}] = argmin\left[\sum_{min}^{max} u(x)\, ln\left({}^{u(x)}\!/_{\hat{p}(x)}\right) + n \cdot \sum_{min}^{max} q|x^o(x)\, ln\left({}^{q|x^o(x)}\!/_{\hat{p}(x)}\right)\right]$$

probabilities to estimate. For just one observation x^o, the maximum entropy solution to Equation (4) for R = 10 is $p(x = x^o) = 0.55$ and $p(all\ x \neq x^o) = 0.05$. That spikey solution is not very credible. It ignores the notion of central tendency. It ignores the evidence cited above in Section II that while most judges may not assign the exact same ratings to replicates in a flight, the ratings they do assign tend to cluster about nearby scores, categories or ranks.

Central tendency and clustering are notions about frequency and distance, near and far. Distance is easily expressed on a cardinal scale of scores, but the evidence cited in Section II shows that score assignments may be influenced by ordinal considerations and many rating systems are ordered categories alone. The difficulty with mapping ordinal categories to a cardinal scale is that, although categories may be adjacent, there is no information about the widths of categories or the transitions between categories. And if scores are interpreted as expressions of economic utility, the economics literature summarized in Barnett (2003, p. 41) is rich in what he calls the modern view that utility is ordinal, cannot employed to calculate indifference points between goods, and cannot be employed to compare one person to another. While using ratings to make inferences about utility and indifference points may violate the logic of utility theory, statistical analysis of ordinal ratings for other purposes is common. According to Chen & Wang (2014), ordinal data are the most frequently encountered type of data in social science.[11] Chen & Wang review methods of mapping categorical responses to numerical values, and those methods include merely ranking the categories. In their texts concerning analysis of categorical data, Agresti (2007, p. 43-44, 119, 195, 230, 299) and Lynch (2007, p. 219) discuss methods of mapping, that again include ranking, and they also present PMFs with dispersion parameters to describe central tendency and clustering in the distributions of categorical data.

Although the research cited above explains several methods of mapping categorical data to numerical values and several PMFs, this article does not pursue finding a "best" method of

[11] Chen & Wang (2014) give the example of "strongly agree," "agree," "have no opinion," "disagree," and "strongly disagree" as a common ordinal scale.

mapping and/or a best PMF. The best of those probably depends on the form of the rating data and the intended analysis or hypothesis test.[12]

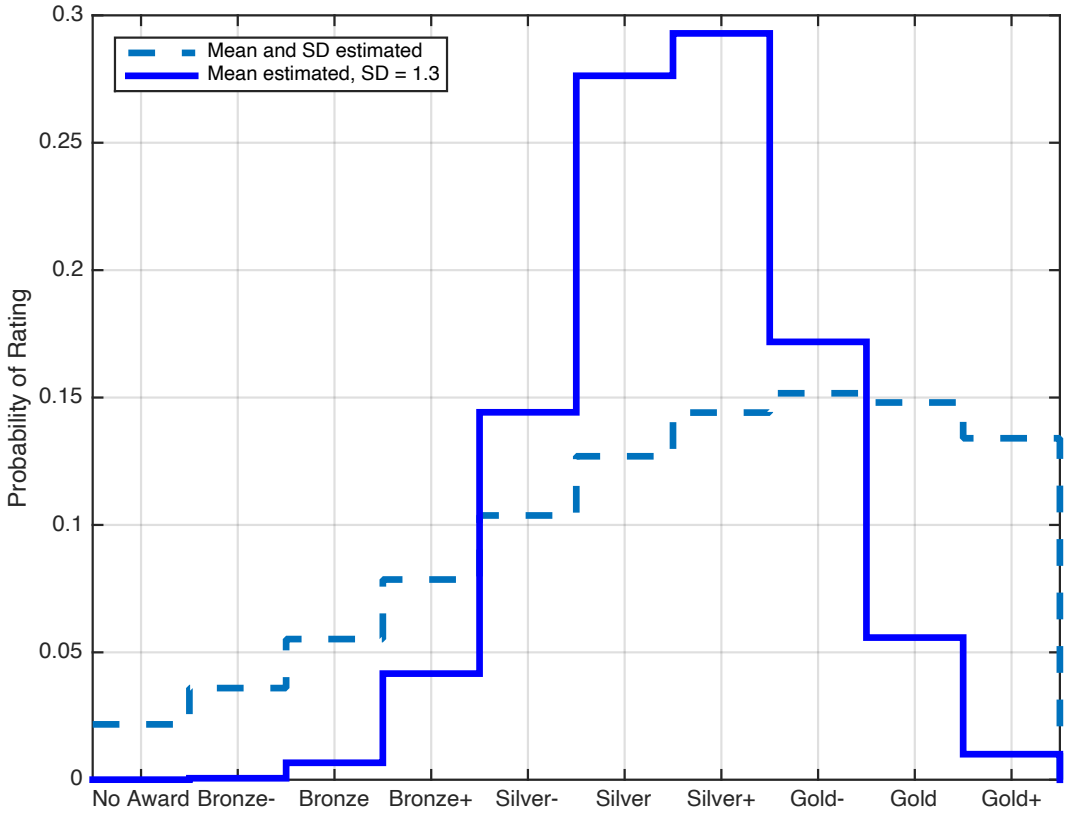
V. An Example

Suppose in this example that a judge assigns a rating of "*G old–*" to a wine entered in a competition with ten ordered categories of rating; No Award, Bronze–, Bronze, Bronze+, Silver-, Silver, Silver+, Gold–, Gold, and Gold+. In that case, and mapping from categories to ranks, $x^0 = 8 \in (1, 10)$. Suppose further that the latent distribution of that rating can be described by a discrete and bounded Gaussian function of mean (μ) and variance (σ^2). With emphasis, the intent is to describe the distribution of potential ordinal ratings assigned by one judge to one wine and there is no intent to make any inference about economic utility or to make a comparison with any other wine or judge. Knowing only that $x^0 = 8 \in (1, 10)$ and using Equation (4), the maximum-entropy estimate of the latent distribution of that rating observed appears as the dashed line in Figure 1. The stair-step shape of the distribution reflects the discrete categories. Next, evidence cited in Section II shows that the standard deviation (SD) in ratings assigned by trained and tested CSF wine judges to blind triplicates averages approximately 1.3 out of 10 ratings; $\sigma = 1.3$. Using that estimate of SD and Equation (4), the maximum-entropy estimate of μ yields the solid line in Figure 1. MATLAB code for Figure 1 is available on request.

[12] For example, the PMF will have one form for categories assigned with replacement and a different form for ranks assigned without replacement.

Figure 1

Estimated Probability Distributions of One Observed Rating of "Gold-"



Although $\sigma = 1.3$ was set exogenously for the solid line in Figure 1, it may be estimated using cross section data from competition results where panels of judges evaluate the same wines.[13] Like for choosing a "best" PMF, this article does not propose a best method of using cross section data, in conjunction with Equation (4), to estimate the parameters in a PMF. The best method is likely to depend on the form of the data and the analysis intended.

[13] For example, consider modeling a judge as a signal processor. Using the *Variance Sum Law*, $\sigma_{ij}^2 = \sigma_j^2 + \eta_j^2\sigma_i^2$ where $\eta_j \geq 0$ is a parameter representing the skill of judge (j) in assessing wine (i).

VI. Conclusion

Wine judges assess wines and assign ratings that are discrete and within bounded sets of scores, ordered categories, or ranks. But much evidence shows that, although they are not merely random, those assignments are both stochastic and heteroscedastic. Those assignments may also be affected by anchoring, expectations, and serial position biases. The distribution of a rating observed is then wine- and judge-specific. Estimating the distribution of a rating to support ratings-related research, or to calculate consensus among judges, is thus acutely difficult because the sample size drawn from a latent wine- and judge-specific distribution is usually one.

The parameters in a PMF expressing the latent distribution of a rating observed can be estimated using the simple maximum entropy estimator in Equation (4). That estimator incorporates the information from none, one or more observations and it relies on a minimum of additional assumptions. An example yields results in Figure 1 that are consistent with the results of experiments in which blind replicates are embedded within flights of wines that are evaluated by trained and tested judges.

Equation (4) is intended as a tool to support ratings-related research and assessments of consensus among judges. Research may lead to improvements in, or application-specific variations of, Equation (4). Further tests of the estimator could include estimating parameters in PMFs from cross section ratings data. On that foundation, the assumption that ratings are deterministic and/or IID that is implicit in much current research, and most calculations of multi-judge consensus, could be relaxed. Aggregates that depend on sums, and research uses ratings in transformations or regressions, can be re-framed, using Equation (4), as maximum likelihood functions that are explicit about the uncertainty surrounding a rating observed.

References:

Agresti, A. (2007). An Introduction to Categorical Data Analysis, Second Addition. John Wiley & Sons, Hoboken NJ, 373 pages.

Ashton, R.H. (2014). Nothing good vver came from New Jersey: Expectations and the sensory perception of wine. *Journal of Wine Economics*, 9(3), 304-319.

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.

Barnett, W. (2003). The modern theory of consumer behavior: Ordinal or cardinal? *Quarterly Journal of Austrian Economics*, 6(1), pp. 41-65.

Bodington, J. C. (2017). The distribution of ratings assigned to blind replicates. *Journal of Wine Economics*, 12(4), 363-369.

Bodington, J. C. (2020). The latent distribution of a rating observed. AAWE Working Paper No 259, October 2020, 15 pages.

Bodington, J., and Malfeito-Ferreira, M. (2017). The 2016 wines of Portugal challenge: General implications of more than 8400 wine-score observations. *Journal of Wine Research*, 2(4), pp. 313-325.

Capehart, K.W. (2019). Does blind tasting work? Another look. *Journal of Wine Economics*, 14(3), 298-308.

Casella, G. and Strawderman, W.E. (1981). Estimating a normal bounded mean. *Annals of Statistics*, 9, 870-878.

Chen, H. & Wang, N. (2014). The assignment of scores for ordinal categorical data. *Scientific World Journal*, 2014(304217), 7 pages.

Cicchetti, D. & Cicchetti, A. (2014). Categorizing A Wine Rating Scale: 2, 3, 4, or More: Is There One We Should Go For? *Journal of Business and Economics*, 5(8), pp. 1199-1204.

Cook, L. and Harslett, P. (2015). An introduction to entropy estimation of parameters in economic models. Australian Productivity Commission, 18th Annual Conference on Global Economic Analysis, Melbourne, June 17-19, 2015, 46 pages.

Corsi, A. and Ashenfelter, O. (2019). Predicting Italian wine quality from weather data and expert ratings. *Journal of Wine Economics*,14(3), 231-233.

de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, 118, 245-260.

Filipello, F. (1955). Small panel taste testing of wine. *American Journal of Enology*, 6(4), 26-32.

Filipello, F. (1956). Factors in the analysis of mass panel wine-preference data. *Food Technology*, 10, 321-326.

Filipello, F. (1957). Organoleptic wine-quality evaluation II. Performance of judges. *Food Technology*, 11, 51-53.

Filipello, F. and H.W. Berg (1958). The Present Status of Consumer Tests on Wine. Presentation to the Ninth Annual Meeting of the American Society of Enologists, Asilomar, Pacific Grove, California, June 27-28, 1958.

Gergaud, O., Ginsburgh, V. and Moreno-Ternero, J.D. (2021). Wine Ratings: Seeking a Consensus among Tasters via Normalization, Approval, and Aggregation. *Journal of Wine Economics*, 1-22.

Golan, A., Judge, G. and Miller, D. (1996). Maximum entropy econometrics. New York, John Wiley & Sons, 307 pages.

Hartley, R.V.L. (1928). Transmission of information. *Bell System Technical Journal*, July 1928, 535–563.

Hölle, D., Aufschnaiter, S., Bogon, J., Pfeuffer, C., Kiesel, A. and Thomaschke, R. (2020). Quality ratings of wine bottles in e-commerce: the influence of time delays and spatial arrangement. *Journal of Wine Research*, 31(2), 152-170.

Jaynes, E.T. (1957a). Information theory and statistical mechanics. *Physics Review*, 106, 620-630.

Jaynes, E.T. (1957b). Information theory and statistical mechanics II. *Physics Review*, 108, 171-190.

Kahneman, D., Sibony, O., and Sunstein, C.R. (2021). Noise, a flaw in human judgement. New York, Little, Brown Spark, Hachette Book Group, 454 pages.

Kliparchuck, K. (2013). What's wrong with wine ratings? MyWinePal, April 8, 2013.

Lam, J., Lambrechts, M., Pitt, C. and Afsharipour, A. (2019). When writing about wine: how ratings impact reviews. *Journal of Wine Research*, 30(4), 335-345.

Leaf, D.E., Hui, J., & Lui, C. (2009). Statistical inference with a single observation of N(θ,1). *Pakistani Journal of Statistics*, October, 2009, 17 pages.

Lombardi, O., Holik, F. and Vanni, L. (2015). What is Shannon information? *Synthese,* 193, 33 pages.

Lynch, S.M. (2007). Introduction to applied Bayesian statistics and estimation for social scientists. New York, Springer, 357 pages.

Malfeito-Ferreira, M., Diako, C. and Ross, C. F. (2019). Sensory and chemical characteristics of 'dry' wines awarded gold medals in an international wine competition. *Journal of Wine Research*, 30(3), 204-219.

Mantonakis, A., P. Rodero, I. Lesschaeve and R. Hastie (2009). Order in Choice: Effects of Serial Position on Preferences. *Psychological Science*, November 2009, 20(11), 1309-1312.

Mitchell, T. (1997, p. 57). Machine Learning. New York, McGraw Hill, 414 pages.

Primavera, J. (2011). Estimating a normal bounded mean. Reading seminar on classics, University of Paris Dauphine, November 21, 2011, 103 pgs. Accessed 1 March 2021 at https://www.slideshare.net/xianblog/bounded-normal-mean-minimax-estimation?ref=.

Rioul, O. (2008). This is IT: A primer on Shannon's Entropy and Information. *L'Information, Seminaire Poincare*, XXIII, 43-77.

Rodriguez, C.C. (1996). Confidence intervals from one observation. In: Skilling J., Sibisi S. (eds) Maximum Entropy and Bayesian Methods. Fundamental Theories of Physics (An International Book Series on The Fundamental Theories of Physics: Their Clarification, Development and Application), vol 70. Springer, Dordrecht, 8 pages.

Schapire, R. (2014). Theoretical machine learning, introduction to probability estimation. Accessed 7 June 2021 at https://arxiv.org/abs/0904.3664, 5 pages.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3), 379-423, 623-656. Reprinted with corrections, 33 pages, downloaded on 12 August 2020 from http://web.mit.edu/6.976/www/handout/shannon.pdf.

Shashua, A. (2008). Introduction to machine learning, Lecture 3: Maximum Likelihood / Maximum Entropy Duality. Accessed 7 June 2021 at https://pdf4pro.com/view/introduction-to-machine-learning-arxiv-2431ae.html, 5 pages.

Taber, G.M., (2005). Judgement of Paris. New York, Scribner, 326 pgs.

Veseth, M. (2008). Wine by the numbers. The Wine Economist. February 9, 2008.

Appendix A: Literature review regarding what can be deduced about the shape of a latent distribution from one observation?

Rodriguez (1996) reports that, in 1964, Robert Machol derived a confidence interval (CI) for an estimate of the mean of an unbounded distribution that is symmetric about zero. Casella and Strawderman (1981) assumed that $x = \mu + \delta$ has a uniform prior distribution that is discrete,

bounded ($\pm a$) and symmetric about zero. Under those assumptions, they employed Bayes Rule to calculate that the posterior mean of the distribution is a hyperbolic tangent function (tanh) of the single observation and the distance to the bounds; $\hat{\mu} = \tanh(ax)$. Rodriguez (1996) re-stated Machol's result, presented an unpublished non-parametric CI for $\hat{\mu}$ that he attributed to Herbert Robbins, and he also derived a CI for the standard deviation ($\hat{\sigma}$) of an unbounded symmetric distribution.

Golan, Judge & Miller (1996, p. 115-123) examined least squares, maximum likelihood, Bayesian and maximum entropy methods of estimating, for one observation, the mean of a bounded distribution that is symmetric about zero. Based on what those authors describe as "standard sampling theory," the least squares and maximum likelihood methods yield $\hat{\mu} = x^o$. Their results using Bayes Rule cite and match Casella and Strawderman (1981). Their maximum entropy estimate of $\hat{\mu}$ is a Lagrange function of information entropy and an exogenous constraint on the difference between x^o and the unobserved μ. Both the Bayesian and the maximum entropy solutions show, as the range of x increases, that $\hat{\mu}$ tends away from x^o toward the center of the range. Golan, Judge & Miller compare the accuracy of the four methods assuming several types of error distribution, and they conclude that the maximum entropy method both relies on the least restrictive assumptions and is the most accurate.

Leaf, Hui & Lui (2009) examined Bayesian estimates of distribution parameters using a single observation. The authors found that reasonable Bayesian results depend very much on starting with a reasonable prior and they recommend further development of axiomatic or so-called fiducial inference. Primavera (2011) addressed estimating the mean of a bounded distribution from one observation using a sample mean, maximum likelihood estimator, Bayesian inference and game theory. The sample mean for one observation implies merely $\hat{\mu} = x^o$. A maximum likelihood estimate yields the same result but interposes a distribution function, and Bayes Rule yields results, like those found by Leaf *et al*., that depend on the prior and the form of the distribution function.

Cook & Harslett (2015, p. 11-22) used one observation and cross entropy to estimate the intercept and slope of a linear equation. Following Bayes, they assumed prior probability distributions for

the two parameters and then selected values for those parameters that minimized an exogenously-weighted cross entropy subject to the Lagrange constraints that the probabilities sum to unity and that the linear function of the parameters yields x^o.

Appendix B: Derivation of Equation (4)

See definitions for algebra in Section IV.

For an arbitrary distribution $\hat{p}$:

$$H(q|x^0) \leq H(u) \qquad \text{(B1)}$$
$$H(u) - H(q|x^0) \leq I(q|x^0, \hat{p}) + I(u, \hat{p}) \qquad \text{(B2)}$$

Two conditions must apply. The symbol for approximately equal ($\approx$) indicates that the solution will be a minimum difference but, due to the form of a PMF, may not be a precise equality ($=$).

$$\hat{p} \approx u|\, n = 0 \qquad \text{(B3)}$$
$$\hat{p} \approx q|x^0 \mid n \to \infty \qquad \text{(B4)}$$

Impose the conditions above on $\hat{p}$ and maximize entropy, subject to x^o, by minimizing the weighted sum of the cross entropies. See also Golan, Judge & Miller (1996, p. 41) regarding cross entropy as a distance and then minimizing cross entropy to calculate the PMF that is consistent with the data but has entropy closest to, in this case, the maximum $H(u)$. Weighting the cross entropy by sample size in Equation (B5) to obtain the conditions in Equations (B3) and (B4) is the author's addition. See other uses of weighting cross entropies by sample size in Mitchell (1997, p. 57) and Golan, Judge & Miller (1996, p. 110).

$$\arg[\hat{p}] = argmin[\, I(u, \hat{p}) + n \cdot I(q|x^o, \hat{p}) \,] \qquad \text{(B5) = (4)}$$