Journal of Wine Economics, Volume 10, Number 1, 2015, Pages 4–30
doi:10.1017/jwe.2014.26

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Analyses of Wine-Tasting Data: A Tutorial*

Ingram Olkin [a], Ying Lou [b], Lynne Stokes [c] and Jing Cao [d]

Abstract

The purpose of this paper is to provide a tutorial of data analysis methods for answering questions that arise in analyzing data from wine-tasting events: (i) measuring agreement of two judges and its extension to *m* judges; (ii) making comparisons of judges across years; (iii) comparing two wines; (iv) designing tasting procedures to reduce burden of multiple tastings; (v) ranking of judges; and (vi) assessing causes of disagreement. In each case we describe one or more analyses and make recommendations on the conditions of use for each. (JEL Classifications: C10, C12, C13, C59, C90)

Keywords: correlation, design of experiments, judge agreement, measures of concordance, meta-analysis, multiple judges.

I. Introduction

Many studies produce data consisting of rankings or ratings of *n* objects by *m* judges. Almost every award and admission to a competitive program is based on such data. In this paper, we are concerned with measuring the quality of rankings and ratings specifically for the judging of wines. Because both rankings and ratings are typically produced using the assessments of two or more judges, their consistency is one measure of the quality of the resulting ranking or rating. We discuss advantages and disadvantages of a variety of measures that have been or could be used for evaluating wine judge consistency.

* The authors wish to thank the editor Karl Storchmann and the reviewer for their comments and suggestions. Jing Cao was partially supported by NSF grant IIS-1302564.

[a] Ingram Olkin is Professor Emeritus, Department of Statistics, Stanford University; e-mail: olkin@stanford.edu.
[b] Ying Lou is Manager Statistician, Lilly Suzhou Pharmaceutical Co, Shanghai Branch; e-mail: lou_ying@lilly.com.
[c] Lynne Stokes is Professor, Department of Statistical Science, Southern Methodist University; e-mail: slstokes@smu.edu.
[d] Jing Cao is Associate Professor, Department of Statistical Science, Southern Methodist University; e-mail: jcao@smu.edu.

© American Association of Wine Economists, 2014

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 1
Example of Wine-Rating System

Score	*Rating system*
95–100	Classic; a great wine
90–94	Outstanding; superior character and style
80–89	Good to very good; wine with special qualities
70–79	Average; drinkable wine that may have minor flaws
60–69	Below average; drinkable but not recommended
50–59	Poor; undrinkable, not recommended

Ranking means that the objects are arranged in order of some attribute, from largest to smallest or vice versa. Rating means that objects are sorted into categories with similarly defined attributes. Rankings and ratings are usually based on an underlying score assigned by a judge or judges. The scores can be interval, meaning that the differences in scores are comparable, or ordinal, in which case they are not. Whether the score is interval or ordinal, and whether ranking or rating is the goal, determines the data analysis needed to assess consistency of judging.

The scoring systems for ranking or rating wines may assign numbers between 80 and 100 or employ a 100-point scale; scores are regarded as ordinal rather than numerical. The scoring of wine quality depends on many factors: aesthetics, pleasure, complexity, color, appearance, odor, aroma, bouquet, tartness, and the interactions with the senses of these characteristics. Table 1 illustrates the wine-rating system employed by a well-known wine journal, *Wine Spectator*. Other wine journals have different rating systems: a summary of such systems can be found at www.wine.com. It is notable that so many variables have been condensed to such a gross scoring system. Many studies examining the judging of wine have found consistency to be lower than desired (Cao and Stokes, 2010; Gawel and Godden, 2008; Hodgson, 2008, 2009). The multifactor nature of scoring may explain this lack of consistency. However, inconsistency may compromise the credibility of wine competitions and thereby lead to a loss of consumer confidence in the competition results.

Neither rankings nor ratings are meaningful unless there is a minimum level of consistency (or concordance) among judges. A maximum ranking consistency ignores the closeness of the scores and occurs when all judges rank the wines in the same order. It can be achieved even if the judges do not have perfectly consistent rankings but requires the additional condition that judges' assigned scores are close enough that they fall in the same category. If a set of judges produces rankings that are independent of one another, then the ranks provide no information of the merits of the wines. An exception to this occurs if information on the relative competency of the judges is available or can be discerned from the rankings themselves.

Measuring the consistency of ratings has been studied in different disciplines for decades, and there is a long bibliography on methods. Some of the key descriptive

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

terms that apply are: methods of concordance, measures of agreement, multirater data, ranking methods, interrater reliability, paired comparisons, and rater agreement (or judge agreement). Judge agreement, which we use as a generic term, can refer to exact matching for nominal ratings or consistency of the ranking. There is little consensus about which statistical methods are best for measuring judge agreement.

The goal of this paper is to describe a variety of those methods and what they can be used for, to describe explicitly what they measure, and to provide guidance on their proper usage in analyzing wine-tasting data. Wine tasting is a popular event for both the specialist and the general public. For example, the website (www.liquidasset.com) created and maintained by Orley Ashenfelter and Richard Quandt provides rich wine-tasting data, wine-tasting analyses, and articles. We single out several papers that provide statistical analyses that relate to the current tutorial. Some of the discussions overlap with our paper, so we suggest that these papers be viewed as complementary to the current paper (Ashenfelter and Quandt, 1999; Ginsburgh and Zang, 2012; Quandt, 2007). Specifically, Quandt (2006) published a paper in the *Journal of Wine Economics* titled "Measurement and Inference in Wine Tasting." The author discussed three topics: (1) testing ranking of wines in blind tastings, (2) evaluating judge agreement, and (3) testing how well judges identify the wines. In doing so, he provided a valuable introduction of statistical methods on these topics. Quandt used Kendall's coefficient of concordance to measure judge agreement. Our contribution in the paper is describing and comparing a number of measures of judge agreement in different scenarios (i.e., two judges, more than two judges, wine tastings across years). By doing this, we hope to provide more depth on measuring and testing judge agreement in wine tasting.

The methods we review range from simple to complex and from familiar to new. In general, there is a tradeoff between simplicity of the measure and richness of the information about wine judging that it provides. In some cases, a simple method that provides information about a single feature of the scoring is best; in other cases, a need to distinguish more detail about the scoring process calls for a more complex method.

In this paper, assume that the goal for judge consistency is that the order of rankings or ratings agrees. We first review the simplest case of two judges and n wines (Section II) to fix ideas and then generalize to $m > 2$ judges and n wines (Section III), which is the usual situation. These analyses provide measures of judge consistency within one study. However, tasting of the same wines may take place at different localities or across different years. For the analyses of these data, we provide a survey of methods to compare judge agreement over multiple datasets (Section IV). In Section V, we discuss a number of analyses other than judge agreement, which includes comparing two specific wines evaluated by m judges, the method of paired comparisons that generates a network, and the design of a wine-tasting event using a balanced incomplete block design.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

II. Analysis for Two Judges and *n* Wines

The simplest wine-tasting event is the case in which two judges each taste n wines and provide a score. The scoring can be a numerical value (discussed in Section II.A) or a rank (discussed in Sections II.B through II.D).

A. A Parametric Analysis: Pearson's Product-Moment Correlation

Pearson's product-moment correlation r is a measure of the linear correlation between two continuous variables. Suppose (x_i, y_i), $i = 1, \ldots, n$, are the scores on the ith wine by judge 1 (the x's) and by judge 2 (the y's). Then the sample product-moment correlation is defined as:

$$r = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ denote the means of the x's and y's, respectively. This measure is scale and location invariant, so that r does not change if the x's and y's are changed to $ax_i + b$ and $cy_i + d$, respectively, provided that $ac > 0$. The range of r is between -1 and $+1$; the midpoint of 0 indicates the lack of a linear relationship in the scores. The value $r = 1$ indicates not only that the order of ranking is identical for the judges but that gaps between the scores of two wines for one judge is a constant proportion of the gaps between their scores for the other judge. When $r = -1$ the gap relationship is still true, but the order of ranking of the two judges is exactly reversed. When the x's and y's have the same ranking but a less-than-perfect linear relationship, this measure will produce less than perfect agreement between the two judges even if their rankings are the same. The focus of r on measuring the linear relationship between the x's and y's may be considered too stringent a rule for comparison of wine judges.

In any case, inference procedures based on the correlation r are available if the judges' scores can be assumed to have a bivariate normal distribution. The variance r of depends on the true correlation ρ of the scores. It was noted by R.A. Fisher that the variance of

$$Z = \frac{1}{2} \log \frac{1 + r}{1 - r}. \tag{2}$$

is independent of ρ and depends only on the sample size. This transformation is known as Fisher's variance stabilizing transformation and has the benefit that Z is approximately normally distributed when n is large, with mean

$$\zeta = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho}, \tag{3}$$

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 2
Scores for 2 Judges on 12 Wines

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Judge 1	80	84	84	84	88	88	88	88	90	92	92	94
Judge 2	80	80	80	84	80	80	96	96	92	94	94	90

and variance $1/(n-3)$. A simple example of how to use Z to construct confidence intervals for ρ is illustrated with the data from two judges, shown in Table 2.

In this case $r = 0.660$ and $Z = 0.793$. An approximate 95% confidence interval for ζ is given by $0.793 \pm 1.96\sqrt{1/(12-3)}$, which yields the interval (0.140, 1.446). Reversing the transformation (Equation (3)) yields that $\rho = [\exp(2\zeta) - 1]/[exp(2\zeta) + 1]$, so that a 95% confidence interval for ρ can be obtained as $[\exp(2 \times 0.140) - 1]/\exp(2 \times 0.140) + 1]$ to $[\exp(2 \times 1.446) - 1]/\exp(2 \times 1.446) + 1]$, or (0.139, 0.895). This interval is very wide due to the small number of wines assessed, so there is little information from this sample about the consistency of the judges. However, it does support the hypothesis that the judges are not performing independently of one another, but this is not of particular interest because experts are expected to score wines similarly.

B. A Nonparametric Analysis: Spearman's Rank Correlation

The inference procedures for the product moment correlation r assume that the judges' scores follow a bivariate normal distribution. Because wine scores are ordinal, this assumption does not hold exactly, which is why nonparametric procedures can be used for comparing the consistency of wine judges. Spearman's rank correlation, denoted by r_s, is one of the earliest nonparametric methods measuring correlation (Spearman, 1904). This method is described most easily as the product moment correlation calculated from ranks of the scores rather than the scores themselves. It measures how well the relationship between the two variables can be described using a monotonic function of the scores. Because r_s can be computed as a product-moment correlation of ranks, it, too, has a range of -1 to $+1$, and $r_s = 0$ indicates the lack of a linear relationship in the ranks. However, r_s is not an estimator of any parameter of the underlying distribution of scores (x_i, y_i) in the way that r is an estimator of the parameter ρ of an underlying bivariate normal distribution. Rather, r_s is most usefully thought of as an index of the similarity of two rankings, where only the three values -1, 0, and $+1$ have interpretable meanings for the underlying distribution of the scores. A rank correlation of 0 for two judges would be that, for any ranking given by judge 1, every possible ranking by judge 2 has the same probability, namely, $1/n!$ Thus, inference to a population based on the statistic r_s is meaningful only for testing the hypothesis that the rank correlation in the population is 0.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 3
Rank Data for 2 Judges and 12 Wines

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Judge 1	1	3	3	3	6.5	6.5	6.5	6.5	9	10.5	10.5	12
Judge 2	3	3	3	6	3	3	11.5	11.5	8	9.5	9.5	7

Let u_i and v_i denote the ranks given to wine i by judges 1 and 2, respectively, and let $d_i = u_i - v_i$, $i = 1, \ldots, n$. Then, rather than using Equation (1) applied to the ranks, Spearman's rank correlation can be computed more easily as

$$r_s = 1 - \frac{6\sum_1^n d_i^2}{n(n^2-1)}. \tag{4}$$

The statistic r_s is approximately normally distributed when n is large. Under the hypothesis that the rank correlation in the population is 0, r_s has a mean of 0 and a variance of $1/(n-1)$ (see, e.g., Kendall, 1962).

Table 3 shows the ranks of the data in Table 2, where ties are shown as the mean of the tied ranks. Using Equation (4), we obtain $r_s = 0.596$. The rank correlation is smaller than the product moment correlation, which in this case is due to ties. In general, ties decrease the correlation. A test that the population rank correlation is 0 can be carried out by calculating the test statistic $z = r_s/\sqrt{1/(n-1)} = 0.596/\sqrt{1/11} = 1.98$, which has a p-value = 0.04. Therefore we can again reject (at level $\alpha = .05$) the hypothesis of a rank correlation of 0 in the population.

(1) Discussion of Ties

A discussion of ties is necessary because wine scores invariably contain ties for each of the judges. The customary recommendation for ties is to use the mean rank of the tied values when computing r_s, as in the example above. This is reasonable when there are few ties and n is large but can be misleading especially if the number of ties is large relative to n. The following simple example clarifies the issue:

	Wine 1	Wine 2	Wine 3
U (Judge 1)	1	2	3
V (Judge 2)	2	2	2

Here $r_s = 1/2$, but by creating a scatter plot of the data, it is apparent that a conclusion of no relationship and a value of 0 for r_s would be more accurate. Suppose all possible rankings of the tied values are enumerated and r_s is computed for each. This results in the following 6 (= 3!) arrangements,

U	V	U	V	U	V	U	V	U	V	U	V
1	1	1	1	1	2	1	2	1	3	1	3
2	2	2	3	2	1	2	3	2	1	2	2
3	3	3	2	3	3	3	1	3	2	3	1

with the resulting values for r_s : 1, $\frac{1}{2}$, $\frac{1}{2}$, $-\frac{1}{2}$, $-\frac{1}{2}$, -1. Now use the average of these values, $\bar{r}_s = 0$, as a measure of correlation. This calculation requires $m!$ repetitions for each set of m ties. For Table 3 this requires (3!)(4!)(2!) for the ties for judge 1 and (5!)(2!)(2!) for the ties for judge 2, for a total of 138,240 computations of r_s and yields $\bar{r}_s = 0.531$. For this example, the computations are feasible. However, a complete enumeration of the permutations for n wines can be computationally intensive for large samples with many ties.

When there are ties, the statistic r_s no longer can attain the values of -1 and $+1$. Kendall (1962) proposed a modification of r_s that adjusts the denominator when ties are present so that the statistic still can achieve these boundary values. Define

$$T_s = \sum t(t^2 - 1)/12, \quad U_s = \sum u(u^2 - 1)/12, \tag{5}$$

where t and u are the lengths of the tied sequences in the xs and ys, respectively. For the data in Table 3, we have

$$T_s = [4(15) + 3(8) + 2(3)]/12 = 7.5,$$
$$U_s = [5(24) + 2(3) + 2(3)]/12 = 11.$$

The adjusted statistic is

$$r_{as} = \frac{\frac{n^3-n}{6} - T_s - U_s - \sum_1^n d_i^2}{\sqrt{\frac{n^3-n}{6} - 2T_s}\sqrt{\frac{n^3-n}{6} - 2U_s}} = \frac{286 - 7.5 - 11 - 115.5}{(16.46)(16.25)} = 0.568.$$

Either of these methods provides a better adjustment for ties for r_s than using the mean ranks and reduces the estimate of 0.596 to either 0.531 or 0.568.

C. A Nonparametric Analysis: Kendall's Tau

Spearman's statistic is useful because it takes advantage of the familiarity of Pearson's correlation coefficient and eliminates the assumption of normality required for inference. However, as noted, the statistic itself does not have an interpretation as an estimator of any parameter of the joint distribution of scores. An alternative statistic that does not have this drawback is Kendall's tau (Kendall, 1938). It is an unbiased estimator of the probability of concordance minus the probability of discordance for a randomly chosen pair of bivariate observations from any distribution. This parameter (difference in probabilities of concordance and

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

discordance) is denoted by τ. Concordance means that the order of x_i and x_j are the same as that of y_i and y_j, and discordance means the opposite. If $\tau = 0$, it means that concordance or discordance is equally likely for the chosen pair; a value of +1 or −1 means that every pair chosen is concordant or discordant, respectively.

Tau can be estimated from a sample of size n by scoring each of the $n(n-1)/2$ pairs (x_i, y_i) and (x_j, y_j) as a +1 or −1 according to whether they are concordant or discordant, that is whether $(x_i - x_j)(y_i - y_j) > 0$ or < 0. Call the total score S, then the estimator of τ is

$$\hat{\tau} = 2S/n(n-1). \tag{6}$$

This statistic is approximately normally distributed when n is large, and its variance is approximately

$$V(\hat{\tau}) = \frac{2(2n+5)}{9n(n-1)}. \tag{7}$$

Because $V(\hat{\tau})$ does not depend on τ, an approximate 95% confidence interval for τ is given by $\hat{\tau} \pm 1.96\sqrt{V(\hat{\tau})}$. A test of the hypothesis H_0: $\tau = \tau_0$ vs. the alternative hypothesis H_a:$\tau \neq \tau_0$ is given by comparing the statistic $z = (\hat{\tau} - \tau_0)/\sqrt{V(\hat{\tau})}$ to a standard normal variable.

As with r_s, ties can cause a problem with $\hat{\tau}$. An adjustment proposed by Kendall (1945) has the positive feature that it maintains the range of −1 to 1, which otherwise does not hold for $\hat{\tau}$ when there are ties and is implemented as follows. When a tie occurs (i.e., $(x_i = x_j)$ or $(y_i = y_j)$), the pair is scored 0 in the computation of S, and the denominator is adjusted accordingly, resulting in the statistic

$$\hat{\tau}_b = \frac{S}{\sqrt{(\frac{1}{2}n(n-1) - T_\tau)(\frac{1}{2}n(n-1) - U_\tau)}}, \tag{8}$$

where $T_\tau = \sum t(t-1)/2$ and $U_\tau = \sum u(u-1)/2$, with t and u the lengths of the tied sequences in the xs and ys. When the number of ties is not too large compared to n, the expression for $V(\hat{\tau})$ in Equation (7) is a reasonable approximation.

For the data in Table 2, $\hat{\tau} = 0.400$ and $V(\hat{\tau}_b) = 0.049$, providing a 95% confidence interval of (0, 0.833). Note that the definition of τ does not use the size of the discordance. For example, the pair (90, 92) as compared with the pair (92, 95) or (93, 97) yields the same value of $\hat{\tau}$. By comparison, slightly disconcordant pairs have a smaller impact on r and r_s than seriously disconcordant pairs. Thus, if distinguishing the degree of disconcordance is important to the user, Kendall's tau may not be a suitable measure.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

D. A Nonparametric Analysis: Cohen's (Weighted) Kappa

Cohen's kappa κ is a nonparametric measure of exact matching consistency for judges, classifying items into c categories that are unordered (Cohen, 1960). Let p_{ij} denote the proportion of wines that are placed in the ith level by the first judge and jth level by the second judge, and let $p_{i+} = \sum_{j=1}^{c} p_{ij}$ and $p_{+j} = \sum_{i=1}^{c} p_{ij}$ denote the row and column proportions. The proportion of exact matches is $p = \sum_{i=1}^{c} p_{ii}$. The innovation of κ was that it corrected the judges' match rate p for chance matching, meaning that which would occur if each judge assigned categories independently according to his or her own pattern. With independent matching, the probability of a match would be $p_e = \sum_{i=1}^{c} p_{i+}p_{+i}$. Thus κ is defined as the increase in the match rate occurring beyond that of chance as a fraction of the total possible increase; that is,

$$\kappa = \frac{p - p_e}{1 - p_e} = 1 - \frac{1 - p}{1 - p_e}. \tag{9}$$

The coefficient κ is similar in spirit to an R^2 value in regression, because it provides a quantification of a ratio of the improvement in a measure of unpredictability. Like R^2, the maximum value of κ is 1, which occurs when the judges always agree. The value $\kappa = 0$ occurs when the judges assign scores in a way that is perfectly consistent with chance matching, in which case $p = p_e$. Unlike R^2, κ can be negative if the match rate p is less than the match rate p_e that one would expect to obtain by chance.

Like Spearman's correlation r_s, κ is an index and not an estimator of any population parameter. In fact, the value of κ depends entirely on how categories are defined. For example, the data in Table 2 can be categorized by an individual score, in which case $c = 7$ (80, 84, 88, 90, 92, 94, 96); $p = 2/12$ and $p_e = \left(\frac{5}{12}\right)\left(\frac{1}{12}\right) + \ldots + \left(\frac{2}{12}\right)\left(\frac{0}{12}\right) = \frac{13}{144}$, yielding from Equation (9) that $\kappa = 0.084$. However, if the categories are defined by the four intervals (80 – 84, 85 – 89, 90 – 95, 96 – 100), then $p = 8/12$ and $p_e = \left(\frac{4}{12}\right)\left(\frac{6}{12}\right) + \ldots + \left(\frac{0}{12}\right)\left(\frac{2}{12}\right) = \frac{40}{144}$, yielding $\kappa = 0.88$. Even when the same categories are used, the value of κ is related to both the sample size and the distribution of scores that the judges use. As a result, the notion of inference from a sample value of κ to a population is generally not meaningful. The one exception is that κ provides a test for a null hypothesis of chance matching. If the categorization by the two judges is independent, then an estimate, $\hat{v}(\kappa)$, of variance of κ is

$$\hat{v}(\kappa) = \frac{p_e(1 - p_e)}{n(1 - p_e)^2} = \frac{p_e}{n(1 - p_e)}. \tag{10}$$

A test statistic is given by $z = \kappa/\sqrt{\hat{v}(\kappa)}$, which has an approximate standard normal distribution under the null hypothesis of chance agreement (Cohen, 1968, eq. 3). To test whether the two judges in Table 2 are behaving as if there is chance agreement for the seven categories, we calculate $z = 0.084/\sqrt{0.0083} = 0.92$, which has a

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

(one-sided) p-value of 0.18, indicating that we cannot reject the hypothesis of chance agreement. To test whether the two judges are behaving as if there is chance agreement for the four categories defined in Table 1, this test statistic becomes $z = 0.88/\sqrt{0.032} = 4.92$ ($p < 0.001$), indicating that the judges are agreeing more than chance predicts. This illustrates the sensitivity of κ to the categories chosen by the analyst.

A refinement of κ, known as weighted kappa and denoted by κ_w, was introduced by Cohen to make the measure more meaningful for ordinal data (Cohen, 1968). For κ_w, disagreements between judges are weighted by the size of their discrepancy. A weight w_{ij} is assigned to the (i,j)th cell to reflect the degree of disagreement between the ith and jth level; that is, $w_{ii} = 0$ and w_{ij} increases as the distance between i and j increases. Cohen's weighted kappa is defined as

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} p_{ij}}{\sum_{i,j} w_{ij} p_{i+} p_{+j}} \equiv 1 - \frac{\bar{p}_w}{\bar{p}_{ew}}. \tag{11}$$

A test of the null hypothesis of chance matching can be carried out using the fact that under this hypothesis, the variance of κ_w can be estimated by

$$\hat{v}(\kappa_w) = \frac{\sum_{i,j} w_{ij}^2 p_{ij} - \bar{p}_w^2}{n \bar{p}_{ew}^2} \tag{12}$$

yielding a test statistic $z = \kappa_w / \sqrt{\hat{v}(\kappa_w)}$, which has approximately a standard normal distribution under the null hypothesis of chance agreement (Cohen, 1968, eq. 13).

In practice, the choice of weights for κ_w is important. Cohen (1968) recommended that a committee of experts decide them, but in practice this only increases subjectivity. Another of his suggestions was to use linear weights proportional to the number of categories apart ($|i-j|$) or quadratic weights proportional to the square of the number of categories apart. The quadratic weights $w_{ij} = (i-j)^2$, produces a value of κ_w close to Spearman's rank correlation (Cohen, 1968).

In our example, using quadratic weights and categories defined as the seven scores as in Table 2 results in $\kappa_w = 0.50$ and $\hat{v}(\kappa_w) = 0.038$, yielding a test statistic $z = 0.50/\sqrt{0.038} = 2.56$ with a (one-sided) p-value = 0.005. For categories defined by the four intervals (80–84, 85–89, 90–95, 96–100), we obtain $\kappa_w = 0.615$ and $\hat{v}(\kappa_w) = 0.038$, yielding a test statistic $z = 0.615/\sqrt{0.038} = 3.15$ with a (one-sided) p-value = 0.001. So the use of weights provides a more similar conclusion about consistency of the judges for the two categorizations.

III. Analysis for *m* Judges and *n* Wines

In this section, we introduce four measures of consistency of a group of $m > 2$ judges of n wines. Three of them are parametric, and one is nonparametric; two of them

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 4
Example of Ratings on Bordeaux Wines

	RP	*NM*	*TA*	*JL*	*JA*	$\bar{y}_{i.}$
W1	90	92	92	92	91	91.4
W2	96	93	93	94	95	94.2
W3	91	88	92	91	89	90.2
W4	95	92	93	93	92	93.0
W5	93	94	93	94	90	92.8
W6	91	93	88	89	90	90.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
W108	93	91	91	91	87	90.6
$\bar{y}_{j}$	91.4	91.4	91.0	91.9	90.7	

provide only a group measure of consistency, while the other two provide estimates of individual judge characteristics. To illustrate their use, we examine data from the website bordoverview.com, which contains ratings assigned by world-renowned critics from the United States and Europe for hundreds of Bordeaux wines from 2004 to 2012 vintages. In this study, we use the dataset from 2012, which contains 108 wines, each tasted by five wine critics: Robert Parker (RP, American), Neal Martin (NM, English), Tim Atkin (TA, English), Jeff Leve (JL, American), and Jane Anson (JA, English). The ratings consist of exact scores (e.g., 85) and range data, which we converted to their maximum bound (e.g., 95 for 93–95). The wines were rated by each critic and placed into one of 17 ordered categories ranging from 82 to 98, where a higher score means a better wine quality. Table 4 shows a subset of the data for the 5 judges of the 108 wines.

A. A Parametric Analysis: Intraclass Correlation (ICC)

The intraclass correlation can be used to assess the consistency of measurements made by multiple judges measuring the same quantity. ICC can be defined in different ways, depending on whether judges are considered random or fixed effects in an analysis of variance (ANOVA) model. The conceptual difference in these two models is defined by whether the judges are regarded as a random sample from a population of judges or whether the judges providing scores are themselves of interest. For the data in Table 4, we are interested in measuring consistency of these specific judges, so we consider the judges as fixed effects.

Let y_{ij} be the score of wine i assigned by judge j. McGraw and Wong (1996) proposed a two-way mixed effects model

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad i = 1, \cdots, n; \; j = 1, \cdots, m, \tag{13}$$

where μ denotes the population mean, α_i is the wine random effect, assumed identically distributed with mean 0 and variance σ_α^2, β_j is the judge fixed effect, and ϵ_{ij} is the

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

residual effect, also assumed identically distributed with mean 0 and variance σ_e^2. The ICC is then defined as

$$\rho_{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}. \tag{14}$$

ICC ranges from 0 and 1 and can be interpreted as the proportion of total variance "between wines." A larger value of ICC, or equivalently, a smaller value of σ_e^2, indicates a higher level of judge agreement among judges.

An estimator of ρ_{ICC} is

$$\hat{\rho}_{ICC} = \frac{MS_\alpha - MS_e}{MS_\alpha + (m-1)MS_e}, \tag{15}$$

where $MS_\alpha = \sum_{i=1}^{n} (\bar{y}_{i+} - \bar{y}_{++})^2/(n-1)$ is the mean squared error for the wine factor. $\bar{y}_{i+}$ denotes the mean score assigned by all the judges to the ith wine, $\bar{y}_{+j}$ denotes the mean score assigned by the jth judge on all the wines, and $\bar{y}_{++}$ denotes the grand mean over all y_{ij}s. The term $MS_e = \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i+} - \bar{y}_{+j} + \bar{y}_{++})^2/[(n-1)(m-1)]$ is the mean squared error for the residual effect. The estimate of ρ_{ICC} for the Bordeaux data is

$$\hat{\rho}_{ICC} = \frac{25.249 - 4.211}{25.249 + (5-1)4.211} = 0.500,$$

which shows a medium level of concordance among the five wine critics.

To obtain a confidence interval (γ_L, γ_U) for ρ_{ICC}, define $\nu_1 = (n-1)(m-1)$ and $\nu_2 = n-1$, and

$$F_O = \frac{MS_\alpha}{MS_e}, \quad F_L = \frac{F_O}{F^*}, \quad F_U = F_O F^*,$$

where $F^* = F_{1-\alpha/2,\nu_1,\nu_2}$ is the upper $100(\alpha/2)\%$ tail of the F-distribution with ν_1 and ν_2 degrees of freedom (d.f.) (Shrout and Fleiss, 1979). Then

$$\gamma_L = \frac{F_L - 1}{F_L + m - 1}, \quad \gamma_U = \frac{F_U - 1}{F_U + m - 1}. \tag{16}$$

Based on these data, $F_O = 5.996$, $F_L = 4.499$, and $F_U = 8.213$, and the 95% confidence interval for ρ_{ICC} is (0.412, 0.591).

It should be noted the ICC in Equation (14) is defined as a ratio of variances, which must be nonnegative. Yet, when MS_α is small relative to MS_e in the ANOVA model, $\hat{\rho}_{ICC}$ will be negative, with a lower bound of $-1/(m-1)$. This

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

generally will not occur, but when it does, it can be interpreted as indicating that ρ_{ICC} is low or that Equation (13) does not fit the data well.

B. A Parametric Analysis: Product-Moment Correlations

Pearson's product-moment correlation (Equation (1)) can be used to measure judge agreement among m judges. Calculate the product-moment correlation r_{ij} for each pair of judges J_i and J_j ($i \neq j; i, j = 1, \ldots, m$). The composite yields a correlation matrix

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{21} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & \cdots & \cdots & 1 \end{pmatrix}, \qquad \text{where } r_{ij} = r_{ji}.$$

One overall measure of concordance is the average correlation among all pairs:

$$g = \sum_{i \neq j} r_{ij} / m(m-1). \tag{17}$$

The value $g = 1$ is that of complete concordance, and a value of $g = 0$ is that of no correlation between the judges. However, the lower bound of g is constrained (because R is positive semidefinite) to $g = -1/(m-1)$, which occurs when all pairs have $r_{ij} = -1/(m-1) r_{ij} = -1/(m-1)$. The lower bound of $g = -1$ can occur only for $m = 2$ judges. Intuitively, the reason it is not possible for more discordance to occur among $m > 2$ judges is that if J_1 is perfectly negatively correlated with both J_2 and J_3, then J_2 and J_3 must be perfectly positively correlated.

Under the hypothesis of multivariate normality of the underlying scores and no concordance, g is approximately normally distributed with mean zero and variance $\mathrm{Var}(g) \approx m(m-1)/n$. This allows inference of the hypothesis that the average ρ_{ij} among the m judges is 0, based on the sample of wines rated. However, as noted previously, this is not usually an hypothesis of interest for expert wine judges. Rather, one is more likely to be interested in determining which judge is most consistent with the other judges in terms of the values given to each wine. To do this, define the column means from the correlation matrix,

$$r_i = \sum_{i \neq j} r_{ij} / (m-1), \tag{18}$$

for $i = 1, \ldots, m$. Now order the column means

$$r_{[1]} \geq r_{[2]} \geq \ldots \geq r_{[m]}.$$

This ranking provides a measure of consistency in the sense that $J_{[1]}$ is in more agreement with the other $m-1$ judges than $J_{[2]}$; $J_{[2]}$ is in more agreement with the others

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 5
Pairwise Pearson's Correlation

	TA	*JA*	*RP*	*NM*	*JL*
TA	1.000	0.338	0.413	0.493	0.468
JA	0.338	1.000	0.435	0.479	0.531
RP	0.413	0.435	1.000	0.603	0.748
NM	0.493	0.479	0.603	1.000	0.707
JL	0.468	0.531	0.748	0.707	1.000
r_i	0.428	0.446	0.550	0.571	0.614

than $J_{[3]}$, and so on. The gap $r_{[i]} - r_{[i-1]}$ provides a measure of the degree to which $J_{[i]}$ dominates $J_{[i-1]}$.

For the Bordeaux data, the matrix R and the column means r_i are presented in Table 5, where the columns of R representing judges are ordered by their consistency. The average between judge correlation is $g = 0.522$. Jeff Leve ($r_4 = 0.614$) shows the most agreement with the other judges, and Tim Atkin ($r_3 = 0.428$) shows the least. This information is based on pairwise comparisons, which provides more information than a single measure of concordance, such as the ICC. In the next section, we describe a method that provides still more information about the individual differences in the judge's score characteristics.

C. A Parametric Analysis: Bayesian Ordinal Model

Cao and Stokes (2010) proposed a Bayesian ordinal model to evaluate wine judge performance by identifying which of the three judge characteristics—bias, discrimination ability, and random variation—are responsible for disagreement among judges. Judge bias measures the systematic difference between a judge's score and the average score of all judges. Based on the bias, one can identify whether a judge is relatively generous, neutral, or stringent in his or her assignment of scores. Judge discrimination measures a judge's ability to distinguish wines based on their quality. Judges with a higher degree of discriminating power can distinguish wines more easily. Judge variation measures the amount of judgment error in a judge's assessment of wine quality.

Suppose judges assign scores by first estimating the wine quality on a continuously valued scale and then comparing it with category cutoffs, which are the same for all judges. Let $\theta_i (i = 1, \ldots, n)$ denote the underlying unobserved quality of wine i, x_{ij} the estimate of θ_i by judge j, and y_{ij} the observed ordinal score assigned by judge j on wine i. Then the continuous estimate of the ith wine's quality θ_i by judge j is modeled as

$$x_{ij} = \alpha_j + \beta_j \theta_i + e_{ij}. \tag{19}$$

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 6
Analysis Based on the Bayesian Ordinal Model

	TA	*JA*	*NM*	*RP*	*JL*
$\hat{\alpha}$	−0.055	−0.133	0.037	0.061	0.186
$\hat{\beta}$	0.694	0.692	0.745	0.895	0.884
σ^2	0.716	0.704	0.246	0.329	0.103
r_B	0.570	0.591	0.825	0.838	0.954

The quantity e_{ij} is the judgment error made by judge j on his assessment of wine i, and is assumed to have a normal distribution with mean zero and a judge-specific variance σ_j^2 for all wines. The parameter α_j quantifies the bias for judge j, and β_j measures the discrimination ability of judge j. Then y_{ij} takes the ordinal score of s if x_{ij} falls between category cutoffs c_{s-1} and c_s, that is

$$y_{ij} = s \quad \text{if and only if} \quad c_{s-1} < x_{ij} \leq c_s.$$

The quantities c_s denote the upper cutoff of ordinal level s for all judges. For more details on the prior assignment of the parameters and Bayesian computation, see Cao and Stokes (2010).

A judge's performance can be evaluated based on the correlation between his/her assigned ratings and the estimates of wine latent quality θ_i's, and the correlation is denoted by r_B. Table 6 shows the results of the Bayesian ordinal model. Jeff Leve's ratings have the highest agreement (i.e., value of r_B) with the estimates of latent quality of the wines (i.e., $\hat{\theta}_i$'s), which is attributed to his larger discrimination level and smaller measurement error. The ratings assigned by Tim Atkin have the least agreement with the $\hat{\theta}_i$'s due to his smaller discrimination level and larger measurement error.

Tables 5 and 6 agree on the most and least consistent of the judges but reverse the order of the middle pair (i.e., NM and RP). Compared to the previous measures (i.e., $\hat{\rho}_{ICC}$ and g), which provide an aggregate assessment of judge agreement, the Bayesian model provides insight into what makes judges differ.

D. A Nonparametric Analysis: Kendall's Coefficient of Concordance

We now return to global measures of consistency for multiple judges but without the assumption of normality that the parametric measures require. Kendall's coefficient of concordance W is such a measure (Kendall and Smith, 1939). Let R_{ij} denote the rank of wine i ($i = 1, L, n$) based on the ratings assigned by judge j ($j = 1, L, m$), and $R_{i+} = \sum_{j=1}^{m} R_{ij}$ ($j \neq i$) denote the total rank given to wine i by the m judges. The mean of the total of the ranks is $\bar{R} = m(n+1)/2$ and denote the sum of squared deviations of each total rank from the overall mean by $S = \sum_{i=1}^{n} (R_{i+} - \bar{R})^2$.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Kendall's W is defined as

$$W = \frac{12S}{m^2 n(n^2-1)}, \tag{20}$$

where W is the ratio of the variance of the row sums of ranks R_{i+} divided by the maximum value of the variance under complete judge agreement. Kendall's W ranges from 0 to 1, where $W = 1$ indicates that all judges have ranked the wines in the same order. When the ratings from the m judges are independent (no concordance), W is not exactly 0. In this case, W has an expectation of $E(W) = 1/m$ with variance $Var(W) = 2(m-1)/m^3(n-1)$ (Kendall, 1962). For example, independent rankings from three judges will yield $W = 0.33$ on average. The statistic W approaches zero as m increases (by the order of $1/m^2$) and also as n increases. Doubling the number of judges decreases the variance by 1/4.

A simple function of W makes it interpretable as a nonparametric equivalent of g in Equation (17). Specifically, the average of Spearman's rank correlations, $\bar{r}_s$, over the $m(m-1)/2$ pairs of judges can be computed as,

$$\bar{r}_s = \frac{mW-1}{m-1}. \tag{21}$$

In the case of perfect judge agreement, $\bar{r}_s = 1$ and $W = 1$. When $W = 0$, $\bar{r}_s = -1/(m-1)$, which means there is actually disagreement among judges.

Siegel (1956) proposed an adjusted Kendall's W when ties are present. Define

$$T_j = \sum_{i=1}^{g_j} (t_i^3 - t_i),$$

where t_i is the number of tied ranks in the group of ranks for wine i, and g_j is the number of groups of ties of the ranks given by judge j. Then Kendall's W corrected for ties is

$$W = \frac{12S}{m^2 n(n^2-1) - m\sum_1^m T_j}. \tag{22}$$

Tables of critical values exist for small values of m and n (Siegel, 1956). For larger samples, an approximate test of significance is obtained from the statistic $m(n-1)W$, which has a chi-square distribution with $n-1$ degrees of freedom under the assumption of no association (Friedman, 1937). For the Bordeaux data, we have

$$W = \frac{12 \times 1585745}{5^2 \times 108 \times (108^2-1) - 5 \times 81432} = 0.612$$

with a p-value of 3.34E–24, indicating there is a significant amount of agreement among the judges.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 7
Example of Ratings on Bordeaux Wines over Three Years

	2010			*2011*			*2012*		
	RP	*NM*	*JA*	*RP*	*NM*	*JA*	*RP*	*NM*	*JA*
W1	91	95	93	91	92	90	91	92	92
W2	95	94	94	94	94	88	97	89	89
W3	92	96	94	89	91	92	91	92	89
W4	86	92	88	83	92	93	89	89	89
W5	95	93	94	93	90	92	90	90	88
W6	96	96	92	93	90	92	92	90	88
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
W70	93	92	92	89	88	86	92	91	87

IV. Meta-Analysis for *m* Judges and *n* Wines

The term *meta-analysis* has been used in methods for combining the results of independent studies (Cooper et al., 2009). In the medical context, it relates to how to combine the results of independent studies, each one designed to evaluate a treatment effect.

In the previous sections, we have studied different measures of judge agreement with two or more judges. In this section, we focus on testing whether the judge agreement patterns are homogeneous across different datasets. This falls in the category of meta-analysis. The data model in this section relates to the case in which there are wine tastings across k years (or at k sites). That is, there are the same m judges and n wines at each year. The Bordeaux data contain three judges' tasting scores (RP, NM, and JA) across three years (2010, 2011, and 2012) on the same set of 70 different wines (i.e., $k = 3$, $m = 3$, and $n = 70$). Table 7 shows a subset of the data.

A. A Parametric Analysis: Concordance Across Years Based on Correlations

From the Bordeaux wine data in Table 7 we generate three matrices of Pearson correlations

$$R_{2010} = \begin{pmatrix} 1 & 0.606 & 0.692 \\ 0.606 & 1 & 0.587 \\ 0.692 & 0.587 & 1 \end{pmatrix}, \quad R_{2011} = \begin{pmatrix} 1 & 0.580 & 0.492 \\ 0.580 & 1 & 0.447 \\ 0.492 & 0.447 & 1 \end{pmatrix},$$

$$R_{2012} = \begin{pmatrix} 1 & 0.585 & 0.566 \\ 0.585 & 1 & 0.509 \\ 0.566 & 0.509 & 1 \end{pmatrix},$$

from 2010, 2011, and 2012. Each matrix shows the correlation of the scores among the three judges, where the rows and columns correspond to RP, NM, and JA, in that

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 8
Test for the Homogeneity of Correlation Matrices

	Three-year test	*2010–2011 test*	*2010–2012 test*	*2011–2012 test*
Test statistic	3.412	1.812	1.216	0.596
p-value	0.452	0.196	0.532	0.910
d.f.	2	1	1	1
Conclusion	do not reject H_0	do not reject H_0	do not reject H_0	do not reject H_0

order. To investigate whether the years differ with respect to concordance, we test the homogeneity of correlation matrices.

There are several alternative tests for the homogeneity of correlation matrices, all based on the assumption that the correlation matrices arise from a multivariate normal model. The likelihood ratio test is rather complicated, and, instead, we adopt a simpler test proposed by Larntz and Perlman (1988) that still provides a correct significance level. Apply the Fisher transformation (Equation (2)) to each correlation coefficient $r_{ij}^{(h)}$ $(i, j = 1, \ldots, m)$ in the hth $(h = 1, \ldots, k)$ sample correlation matrix R_h, obtaining $z_{ij}^{(h)}$. Define the element-wise means $\bar{z}_{ij} = \sum_{h=1}^{k} z_{ij}^{(h)}/k$ and variances $v_{ij} = (n-3)\sum_{h=1}^{k}(z_{ij}^{(h)} - \bar{z}_{ij})^2$, where n is the number of wines. Then the test statistic is

$$T = max\{v_{ij},\ 1 \le i < j \le m\}. \tag{23}$$

The null hypothesis that the sample correlation matrices $R_1, \ldots, R_k$ come from a population with a common correlation matrix is rejected if $T > \chi^2_{k-1}(\tilde{\alpha})$, where $\chi^2_{k-1}(\tilde{\alpha})$ is the $\tilde{\alpha}$ th percentile of the χ^2_{k-1} distribution. To achieve a given significance level α, we compute $\tilde{\alpha} = (1-\alpha)^{2/m(m-1)}$ from α and use $\tilde{\alpha}$ in the chi-square test. Because the test statistic (Equation (23)) is based on the maximum of v_{ij} $(1 \le i < j \le m)$, the test is generally conservative. For large m, the data has to show a strong difference to reject the null hypothesis.

Table 8 provides a summary of the three-year and two-year comparisons. Based on the chi-square test, there is a lack of evidence to conclude that the agreement among the three judges is heterogeneous across the three years ($p = 0.452$). This conclusion confirms the noticeable similarity among the three correlation matrices. The subsequent pairwise comparisons yield similar test results (i.e., do not reject H_0), which is consistent with the conclusion on the across-the-three-year comparison. The direct comparison of correlation matrices R_{2010}, R_{2011}, and R_{2012} provides more details on the judge agreement patterns. Both the order and the magnitude of the pairwise correlations remain similar between 2011 and 2012, i.e., $r_{12} > r_{13} > r_{23}$, $g_{2011} = 0.51$, $g_{2012} = 0.55$, where g is the average correlation among all pairs in Equation (17). In contrast, the order of the pairwise correlations in 2010 has changed to $r_{13} > r_{12} > r_{23}$, and the average level of judge agreement is $g_{2010} = 0.63$.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

The fact that g_{2010} is the largest and g_{2011} is the smallest among the three gs and the order of the pairwise correlations differs between the two years suggests that the judge agreement between 2010 and 2011 is the weakest. This conclusion is confirmed by a smaller p-value between the 2010 and 2011 comparison ($p = 0.196$) compared to the other two-year comparisons (p-values are 0.532 and 0.910, respectively).

B. A Nonparametric Analysis: Concordance Across Years

In order to avoid a test based on the assumption of an underlying normal distribution, we can use a nonparametric procedure to assess judge consistency across years or sites. Specifically, we use Kendall's coefficient of concordance as a basis of comparison, rather than the product- moment correlation coefficient. Again we illustrate the method with the Bordeaux data from Table 7. As in the previous section, we test the hypothesis that judge agreement is consistent over time against the hypothesis that it is not. If the relationship among the judges' ratings were consistent, an exchange of a year's set of scores for another year's for each of the wines would not change the distribution of Kendall's W nor would it change the relationship among the coefficients from year to year. Based on this rationale, we propose a permutation test using Kendall's W (Equation (22)). For the Bordeaux data in Table 7, we permute the year indices for each wine i to produce a null dataset under the hypothesis that judge agreements are comparable over time. Note that each year produces a value of W. A test statistic that measures the discrepancy in consistency can be calculated from each permuted dataset and an empirical null distribution produced. The test statistic we propose is

$$RW = \frac{W_{min}}{W_{max}}, \tag{24}$$

where W_{min} and W_{max} are the minimum and the maximum of the Ws respectively. The p-value for this comparison is the tail probability of RW based on the observed data for the empirical null distribution. This permutation test can be extended to multiple years, where the year index are permutated for each wine.

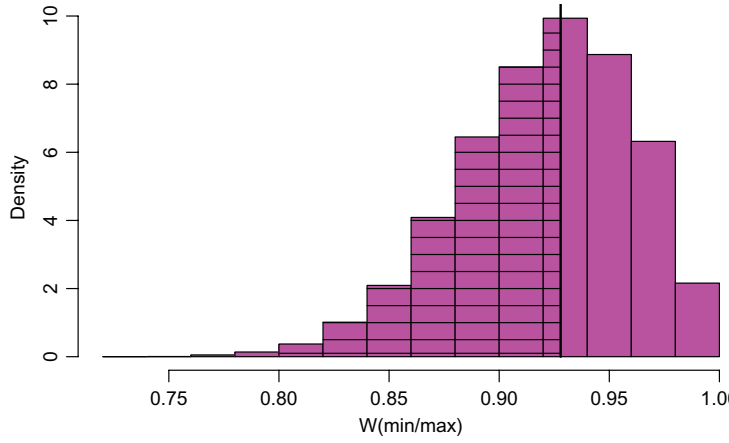
Applying this method to test whether the levels of judge agreement are similar across the three years for the Bordeaux wine data yielded the following results. The p-values, based on 20,000 permutations, are 0.525 for the three-year comparison, 0.260 for the 2010–2011 comparison, 0.535 for the 2010–2012 comparison, and 0.716 for the 2011–2012 comparison, respectively. Taken together, we can reach the same conclusion as the one in the parametric analysis—that is, levels of judge agreement are similar across the three years, where the similarity of judge agreement between 2010 and 2011 is relatively weaker.

To add more detail to the permutation test, we include Figure 1, which is the histogram of RW's empirical null distribution for the three-year comparison. The vertical line shows the observed RW, which is $RW = 0.675/0.727 = 0.928$. The p-value is the proportion of RW under the empirical null distribution that is less

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Figure 1

The Empirical Null Distribution of R_W for the Three-Year Comparison



than the observed *RW*, which is the marked area below the vertical line in the histogram. As mentioned above, the *p*-value is 0.525 for the three-year comparison.

V. Other Novel Analyses on Wine Tasting

Tasting events provide a rich source of alternative scenarios and analyses. They raise many issues in addition to judge agreement or consistency. We here introduce several different models for future consideration.

A. Which Wine is Better? Analysis for Two Wines and m *Judges*

In this tasting model, two wines (A and B) are to be compared by m judges. The central goal is to determine which wine is superior. Suppose the data are obtained from a blinded tasting of two wines from m judges. Each judge gives a score of 1 if wine A is preferable to wine B, and a 2 if wine B is preferable to A. Ties are denoted by 0 and are omitted from the analysis. Table 9 provides an illustrative example of such data for two wines and 10 judges.

Define n_1 as the number of ones (i.e., number of judges who prefer wine A) and n_2 as the number of twos (i.e., number of judges who prefer wine B). Then a test statistic for deciding whether the differences in the preference between the two wines are due to chance is

$$C = \frac{(|n_1 - n_2| - 1)^2}{n_1 + n_2}. \tag{25}$$

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 9
Scores for 2 Wines and 10 Judges

Judge	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Score	1	1	2	1	2	0	0	1	1	1

Under the null hypothesis of no preference, the statistic C has an approximate chi-square distribution with 1 d.f. If the test result is significant, then it indicates that one wine is significantly preferable to the other. Otherwise, the preference is not statistically significant.

This test, called the sign test, was introduced by Dixon and Mood (1946). For small samples, exact significance levels are provided by Dixon and Massey (1969). Define $p = n_1/(n_1 + n_2)$ as the proportion of ones. then $|n_1 - n_2|/(n_1 + n_2) = 2|p - 0.5|$. So the statistic C measures the departure from equality of zeros and ones. The -1 in C is a continuity correction to make the distribution closer to the chi-square distribution.

For the data in Table 9, $n_1 = 6$, $n_2 = 2$, and $C=(|6 - 2| - 1)^2/8 = 1.125$, which has a p-value of 0.29 obtained from the chi-square distribution table. This indicates that no difference in preference between the two wines was detected. In general, a larger sample size of judges is required to detect significance. For a 5% level of significance, the chi-square critical value is 3.81. Thus, for $n_1 = 12$, $n_2 = 4$ the value of C is 3.06, which is not significant. However, for $n_1 = 18$, $n_2 = 6$ (a threefold increase) the value of C is 5.04, which is highly significant.

This sign test is to determine which of two wines is better. Quandt (2007) considered the case in which there are two types of wines being tasted and each type may contain a number of wines. Quandt proposed a permutation test based on the rank sums to determine whether the wines of one type on the whole are more favored than the wines of the other type.

B. Paired Comparison: Network

Paired comparisons is an efficient method for ranking n items. It involves comparing items two at a time. This is a recognized method that optometrists use to find the single best prescription for eyeglasses. However, there is a price in that the method requires $n(n-1)/2$ comparisons instead of n. For example, a comparison of 10 wines requires 45 comparisons instead of 10. This design has the advantage in terms of achieving a more accurate evaluation in that most people are capable of ranking only a small number of items (i.e., five to nine) at a time (Miller, 1956).

Pairwise comparisons also provide a method for the analysis of consistency. Denote the preference "A is preferred to B" by $A \rightarrow B$. For three wines, A, B, and C, if A is preferred to B, B is preferred to C, and C is preferred to A, then there is a loop and an

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

inconsistency. For a moderate number of wines with similar scores, such inconsistencies are to be expected. To further clarify this point, there are eight possible preference configurations among three wines:

$$(1)A \to B \to C \to A, (2)A \to B \to C, (3)B \to C \to A, (4)C \to A \to B,$$
$$(5)A \to C \to B \to A, (6)A \to C \to B, (7)C \to B \to A, (8)B \to A \to C.$$

Of these, only (1) and (5) are inconsistent, so with a random configuration, we expect a 25% chance of inconsistency among the preference configurations.

We present a network approach to study judge agreement based on pairwise comparisons of wines. In the network graph, the wines to be ranked are the vertices, and the judges' preferences are the edge flows connecting the vertices. Figure 2 provides an example of the network involving six wines. When judges all agree with one another, there will be no inconsistent triples in the network. The more inconsistent triples there are, the more serious is the lack of agreement among judges. So we can use the proportion of inconsistent triples among all triples as the test statistic (denoted as p_{inc}) to measure judge agreement.

We use the example in Figure 2 to demonstrate the idea. The network also can be presented by an $n \times n$ matrix

$$V = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

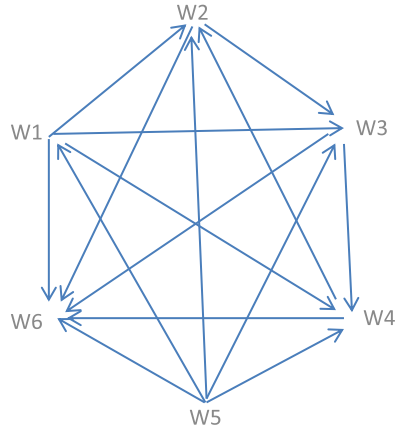
in which the element $v_{ij} = 1 (i \neq j; i, j = 1, \ldots, n; n = 6))$ indicates that vertex i is preferred to vertex j. The diagonal element $v_{ii} = 0$ and $v_{ji} = 1 - v_{ij} (i \neq j)$ is the symmetric comparison. There are several ways to define v_{ij}: by direct pairwise comparison, by comparing the score average between each pair of wines, or by the majority of votes between each pair, and so on. From Figure 2, the triples that consist of wines $W2$, $W3$, and $W4$ are inconsistent. Specifically, we have $W2 \to W3 \to W4 \to W2$.

The pairwise ranking matrix when there is no agreement between judges (denoted as V_{random}) can be constructed by randomly assigning a 0 and 1 to each $v_{ij} (i < j)$, the upper diagonal elements in V_{random}, and setting $v_{ji} = 1 - v_{ij}$. The empirical null distribution of the proportion p_{inc} of inconsistencies is then obtained as the percentage of inconsistencies over all random choices of preferences. The p-value is the proportion of p_{inc} under the empirical null distribution that is less than the observed p_{inc}. In the context of this example, the p-value is 0.047, which indicates that the judge agreement based on the triple-wise comparison is significant.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Figure 2

The Network of Pairwise Comparisons for Six Wines



C. The Design of a Wine-Tasting Event

As the number of wines to be tasted increases, it becomes more difficult to score or rank the wines. For example, about 2,500 different wines competed in the California State Fair Wine Competition in 2009. Altogether 68 judges were assigned to 17 panels, with 4 judges in each panel, and each panel evaluated about 150 wines over two days. The wines were organized into flights, which are collections of wines of a similar type. Each panel of judges evaluated a number of flights per day, ranging from three flights of large size (with more than 30 wines) to 10 flights of small size (with less than 10 wines). With the intensive tasting task, judges become exhausted, and the wine-tasting results can become less reliable. This raises the question of whether an analysis can be generated with fewer tastings, which in turn implies that there will be less difficulty in judging the wines.

When each judge tastes each wine, the design is called a *completely balanced block design*. By comparison, a *balanced incomplete block design* (BIB) permits an analysis that yields similar comparisons but saves observations. The development of BIB originated in an agricultural context in the 1930s by R.A. Fisher and F. Yates (1974) and is well documented in books on the design of experiments. See, for example, Amerine and Roessler (1976), who proposed using a BIB for wine-tasting events. We now describe the method.

Because we desire a balance in the design, not all combinations of the number of judges and the number of wines permit a BIB. To fix ideas and to clarify the ingredients in the design, we focus on Table 10, in which $n = 6$ wines are scored by $m = 10$ judges, in which each judge tastes $j = 3$ wines, and each wine is scored by $w = 5$ judges. Furthermore, each wine is compared twice ($\lambda = 2$) with every other wine.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 10
BIB Design for the Configuration: $n = 6$, $m = 10$, $j = 3$, $w = 5$, $\lambda = 2$

	W1	*W2*	*W3*	*W4*	*W5*	*W6*	*Row Mean*
J1	5	4			6		5.0
J2	7	6				5	6.0
J3	6		8	5			6.3
J4	7		6			8	7.0
J5	4			6	5		5.0
J6			5		4	9	6.0
J7		7	8	6			7.0
J8		9	4		6		6.3
J9		6		5		10	7.0
J10				4	4	8	5.3
Col. Mean	5.8	6.4	6.2	5.2	5.0	8.0	

Not all choices of n, m, j, w, λ can be constructed to yield a BIB design. There are two restrictions:

$$Mj = nw \tag{26}$$

$$\lambda(n-1) = w(j-1). \tag{27}$$

The book *Statistical Tables for Biological, Agricultural and Medical Research* (Fisher and Yates, 1974, p. 56) and Cochran and Cox (1957) provide examples of BIB designs for different values of n, m, j, w, and λ. For example, the configuration $n = 6$, $m = 10$, $j = 3$, $w = 5$, $\lambda = 2$ satisfies the restrictions, so this BIB design has a total of 30 tastings instead of 60 under a completely balanced block design.

Table 10 presents an example of the above configuration, in which the scoring is on a 10-point scale with 1 being the lowest and 10 being the highest score. Here, we see that wines W1 and W2 are tasted by judges J1 and J2, that wines W1 and W5 are tasted by judges J1 and J5, and so on. Furthermore, wines W4 and W6 are compared twice, once by judge J9 and once by judge J10.

The data is analyzed using ANOVA under the BIB design. Let y_{il} be the score given by the lth judge to the ith wine, where $i = 1, L, n$, and $l = 1, L, m$. Further, suppose each score has a wine effect and a judge effect, so that

$$y_{il} = \mu + \alpha_i + \beta_l + \epsilon_{il},$$

where μ is the overall mean, α_i is the mean effect of the ith wine, β_l is the mean effect of the lth judge, and ϵ_{il} is the random effect. Note that the degrees of freedom for the judges is $m - 1$, and for the wines is $n - 1$. The degrees of freedom for the total is $N - 1$, where $N = nw = mj$. Table 11 shows the ANOVA table of results in which $MSE = SSE/d.f.$ The test of judge effect $(\beta_1 \ldots \beta_{10})$ is obtained from Table 11 as

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

Table 11
The ANOVA Analysis

Source	*SSE*	*d.f.*	*MSE*	*F-test*	*p-value*
Wine	29.1	5	5.82	2.37	0.09
Judge	16.7	9	1.86	0.75	0.66
Error	36.9	15	2.46		
Total	82.7	29			

$MSE(judges)/MSE(error) = 1.86/2.46 = 0.75$ and yields a p-value from the F-test as 0.66. With a significance level of 0.05, we reject the null hypothesis that the mean effects of judges are the same.

It should be noted that the test based on this ANOVA model is not for judge agreement in terms of concordance of wine ratings. It is to test whether judges have similar average ratings. Judges who have similar average ratings may have poor consensus. For example, if one judge assigns score 4, 5, 6 to three wines, and another judge assigns score 6, 5, 4 to the same three wines, then the average ratings between the two judges are the same, yet the orders of the ratings are completely opposite.

VI. Discussion

In this paper, we have provided a tutorial on data analysis of wine tasting events. Specifically, we focused on evaluating judge agreement in the cases of two judges, m judges, and comparison across years in Sections II to IV. In Section V, we provided other analyses on comparing two wines, pairwise comparison of wines based on a network and balanced incomplete block design to reduce the tasting load for judges. We plan to further develop some of the three topics in the future. In particular, one topic is how to use the network to produce an aggregate ranking of the wines and a valid measure on the significance of the ranking. A second area of research is to compare the effects of using a balanced incomplete block design with a completely balanced block design.

All the analyses in the paper are conducted in R, free statistical software. The R code is provided in the online supplementary file.

Supplementary Material

For supplementary material accompanying this paper visit http://dx.doi.org/10.1017/S1931436114000261.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press

References

Amerine, M.A., and Roessler, E.B. (1976). *Wines, Their Sensory Evaluation*. San Francisco: W.H. Freeman.

Ashenfelter, O., and Quandt, R. (1999). Analyzing a wine tasting statistically. *Chance*, 12(3), 16–20.

Cao, J., and Stokes, L. (2010). Evaluation of wine judge performance through three characteristics: Bias, discrimination, and variation. *Journal of Wine Economics*, 5(1), 132–142.

Cochran, W.G., and Cox, G.M. (1957). *Experimental Designs*. New York: Wiley.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220.

Cooper, H., Hedges, L. V., and Valentine, J. C. (2009). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd ed. New York: Russell Sage Foundation.

Dixon, W.J., and Massey, F.J., Jr. (1969). *Introduction to Statistical Analysis*, 3rd ed. New York: McGraw-Hill.

Dixon, W.J., and Mood, A.M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41(236), 557–566.

Fisher, R.A., and Yates, F. (1974). *Statistical Tables for Biological, Agricultural and Medical Research*, 2nd ed. Edinburgh: Oliver and Boyd.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200), 675–701.

Gawel, R., and Godden, P.W. (2008). Evaluation of the consistency of wine quality assessments from expert wine tasters. *Australian Journal of Grape and Wine Research*, 14(1), 1–9.

Ginsburgh, V., and Zang, I. (2012). Shapley ranking of wines. *Journal of Wine Economics*, 7(2), 169–180.

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. (2009). An analysis of the concordance among 13 U.S. wine competitions. *Journal of Wine Economics*, 4(1), 1–9.

Kendall, M.G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2), 81–93.

Kendall, M.G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3), 239–251.

Kendall, M.G. (1962). *Rank Correlation Methods*, 3rd ed. London: Charles Griffin.

Kendall, M.G., and Smith, B.B. (1939). The problem of m rankings. *Annals of Mathematical Statistics*, 10(3), 275–287.

Larntz, K., and Perlman, A. (1988). A simple test for equality of correlation matrices. In *Statistical Decision Theory and Related Topics IV*, J.O. Berger and S.S. Gupta (Eds.), 289–298.

Legendre, P. (2005). Species associations: The Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2), 226–245.

McGraw, K.O., and Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46.

Miller, G.A. (1956). The magical number seven, plus or minus two some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.

Quandt, R. (2006). Measurement and inference in wine tasting. *Journal of Wine Economics*, 1(1), 7–30.

Quandt, R. (2007). A Note on a test for the sum of ranksums. *Journal of Wine Economics*, 2(1), 98–102.

Shrout, P.E., and Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101.

https://doi.org/10.1017/jwe.2014.26 Published online by Cambridge University Press