JOURNAL OF WINE RESEARCH, 2017
VOL. 28, NO. 4, 313–325
https://doi.org/10.1080/09571264.2017.1392291

Routledge
Taylor & Francis Group

The 2016 Wines of Portugal Challenge: general implications of more than 8400 wine-score observations

Jeffrey Bodington[a,b] and Manuel Malfeito-Ferreira[b]

[a]Bodington & Company, 50 California Street #630, San Francisco, CA 94111, USA; [b]Linking Landscape, Environment, Agriculture and Food Research Centre (LEAF), Instituto Superior de Agronomia, University of Lisbon, Lisboa, Portugal

ABSTRACT

The Wines of Portugal Challenge is an annual competition among wines produced by over 1000 vintners in over 30 of the country's wine growing regions. In 2016, judges assigned scores to over 1300 wines resulting in over 8400 wine-score observations. Analysis of that large sample yields implications about wine judges' ratings that are difficult to detect with statistical significance in the small samples that are typical of most wine tastings. The Challenge's frequency distribution of scores showed left skewness and local peaks just below the score thresholds for bronze, silver and gold awards. Student's *t*-tests showed that there were no significant differences in scores assigned by gender-of-judge, nationality-of-judge and to wines from different regions. However, judges did assign higher scores to sweet wines than to other types of wine. While the dispersion in scores was material, *p*-values showed that the aggregate order of rating was very unlikely to be random and the distributions of mean scores showed that the strengths of judges' preferences against the least-preferred wines were stronger than those in favor of the most-preferred wines. Ties between wines' mean scores were common and could be broken by several methods including the preference probabilities implied by a Plackett-Luce model.

ARTICLE HISTORY
Received 10 November 2016
Accepted 3 October 2017

KEYWORDS
Portugal; wine tasting; OIV score sheet; statistics; preference; ranking models

1. Introduction

Dozens of wine competitions are held each year that, in part, promote wine producers, brands, regions and countries (Peattie, 1995). Judges at those competitions grant awards that they intend to reflect the quality of each wine. Consumers then employ those awards as cues that influence their wine purchasing and consumption decisions (Herbst & Von Arnim, 2009). However, those awards are often a subject of controversy due to the low reliability of judges (Cliff & King, 1997, 1999; Hodgson, 2008 Honoré-Chedozeau, Ballester, Chatelet, & Valérie Lempereur, 2015; Scaman, Dou, Cliff, Yuksel, & King, 2001), a lack of reliable statistical analysis (Ashenfelter & Quandt, 2012) and preferences in some cases for international commercial wine styles (Loureiro, Brasil, & Malfeito-Ferreira, 2016).

CONTACT Jeffrey Bodington jcb@bodingtonandcompany.com

In an effort to standardize the methodologies employed in wine competitions, the International Organization of Vine and Wine (OIV) published rules that competitions must follow to obtain recognition by the OIV (2009). Among many aspects of a competition, the rules apply to selecting judges, recording results on a form and the method employed to award medals. OIV prescribes that the aggregate score for each wine is the arithmetic mean of the scores given by the judges, and awards are then made according to that arithmetic mean. That process can lead to many ties. Although ties can be broken by considering more decimal places, some ties remain and sums-of-scores methods can lead to odd results because judges' may use different implicit scoring scales. González, Sánchez-Sáenz, and Mejias-Barrera (2014) proposed a method to score and break ties but did not test the method on actual tasting data. On that foundation, this article presents an examination of the distribution of judges' scores under OIV rules, compares the scores that different types of judges assign to different types of wine, and then proposes several methods of breaking ties in wines' mean scores.

The tasting data employed here are the results of the 2016 Wines of Portugal (WoP) Challenge. Wine judges, according to the OIV protocol, assigned a score to each of over 1300 wines. The resulting sample of over 8400 wine-score observations, in addition to specific information about Portuguese wines, enabled large-sample analysis of judges' scoring behavior that is also applicable to other wine tastings. The large sample also enabled the observation of results that are not evident, or cannot be tested to widely accepted levels of statistical significance, in the small samples that are typical of wine tastings. The WoP tasting protocol and results are described and tested for randomness in Section 2. The distribution of judges' scores, skewness, local peaks, judge-gender bias and judge-nationality bias are evaluated in Section 3. In Section 4, judges' scores for wines of certain types and from certain regions are analyzed. Section 5 presents an analysis of the potential for randomness in the assignments of bronze, silver and gold awards, and Section 6 then presents methods of breaking ties between wines when granting such awards. Conclusions follow in Section 7.

2. Description of the WoP Challenge

The Wines of Portugal (again, WoP) Challenge is an annual wine tasting event organized by the Instituto da Vinha e do Vinho, I.P. (IVV). IVV is the Portuguese government institution that coordinates and controls the Portuguese wine industry, with central offices in Lisbon, Portugal.

The Challenge is open only to wines that are made from grapes grown, vinified and bottled in Portugal. Those wines comply with the label requirements of European Union Regulation 1224/2007 through Portuguese Order 239/2012 and the OIV Standard for International Wine and Spirituous Beverages of Viticultural Origins under Resolution OIV/Concours 332A/2009. In sum, those regulations mean that wines labeled to be from different regions are actually made in the respective regions and are not the same wine under different labels. That standard is a foundation for the regional comparisons made below in Section 4.

The IVV Board of Directors selects judges from among wine professionals who include winemakers, sommeliers, oenophiles, gastronomes, journalists and others by invitation. IVV divides the judges into panels of five to seven members. A Portuguese winemaker

who has recognized expertise directs each panel, and each panel also includes two foreign judges. Once tasting begins, wines are tasted in flights of approximately eight samples. While the wines are tasted blind to vintner and price, the judges are told the vintage year, category and grape variety of each sample. The judges taste and score each wine, and each judge fills out a computerized tasting form. To minimize the effect of color differences on relative scores, IVV instructs judges to give maximum sub-scores to visual parameters. Differences in scores are thus due primarily to judges' assessments of non-visual wine qualities. Mineral water and unsalted crackers are provided to attenuate palate fatigue. The director of each panel has access to all scores and, after evaluating each wine, has the option to initiate a short discussion amongst the judges. Each judge may then, but is not required to, enter a revised score. Under OIV rules, the aggregate score for each wine is the arithmetic mean of the judges' final scores.

IVV awards four different medals. No medals are awarded to the wines with a mean scores lower than 80. Bronze medals are awarded to wines with mean scores of 80 or more but less than 85 (up to a maximum of 25% of all prized wines including Gold and Silver), Silver medals are awarded to wines with mean scores of 85 or more but less than 90 (up to a maximum of 12% of all wines entered in the WoP), and Gold medals are awarded to wines with mean scores of 90 points or more (up to a maximum of 6% of all wines entered in the WoP). 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). The Grand Jury is selected by IVV and is composed of three Portuguese and three foreign judges. The percentage limits on Bronze, Silver, Gold and Great Gold have the effect of eliminating 'score inflation' and a resulting imbalance and devaluation of the medals. These percentage limits mean that the score thresholds for each of the medals are in practice higher than 80, 85 and 90.

The 2016 Challenge was held from May 9th through 12th, 2016 at the National Agriculture Fair in Santarém, Portugal. The wines and tasters are summarized in Table 1. In sum, 151 judges sampled 1328 wines and turned in a total of 8445 scores.

Before turning to evaluate specific implications of the Challenge results, Marden (1995, Chapters 3 and 4) and Alvo and Yu (2014, Section 2.3) advise beginning with tests for randomness. If judges' expressions of preference are merely random then there is little point

Table 1. Summary of the 2016 WoP Challenge.

Wines	Number	Description
Total	1328	
Vintages	26	1952, 1966, 1967, 1971, 1972, 1974, 1978, 1984, 1987, 1994, 1996, 1999, 2000, 2001, 2003, 2004, 2005, 2007-2015 and non-vintage
Categories	8	Red, red sparkling, rosé, rosé sparkling, white, white sparkling, late harvest and sweet
Origin standard DOP	17	Açores, Alenquer, Alentejo, Bairrada, Beira Interior, Bucelas, Carcavelos, Dão, Douro, Madeirense, Óbidos, Palmela, Pico, Távora-Varosa, Tejo, Trás-os-Montes and Vinho Verde
DOP Fortified	4	Madeira, Moscatel Douro, Moscatel Setúbal, Porto
IGP	11	Alentejano, Algarve, Beira Atlântico, Duriense, Lisboa, Minho, Península de Setúbal, Tejo, Terras da Beira, Terras do Dão and Transmontano
Vintage year/ variety	10	Non-DOP and IGP wines that have quality control and may show the vintage year or the grape variety on the label
Judges		
Total	151	
Scores/wine	6.4	Average number of judges scoring each wine
Sample size	8445	Aggregate total of scores assigned to wines by the judges

in further analysis. On that basis, the frequency distribution of judges' scores appears in Figure 1 and several tests of randomness are presented below.

The distribution in Figure 1 does not have the characteristic flat shape of a uniform random distribution. Random scoring between 0 and 100 would yield a flat distribution with a mean score of 50 but the observed distribution had a mean score of 82.8, standard deviation (SD) of 7.2 and left-hand skewness of −1.1. The possibility remains that scores are random with a bounded and skewed but bell-shaped distribution for behavioral, protocol and other reasons. That possibility was tested here using the test in Equations (1) through (3) below. Equation (1) expresses the likelihood ratio statistic (LRS) for the difference between the maximum likelihood estimate ($\mathcal{L}_{\text{MLE}}$) and the random likelihood ($\mathcal{L}_0$). The LRS has, asymptotically, a chi-square distribution with $W-1$ degrees of freedom. See, an example of this test in Marden (1995, p. 58, 216). Equation (2) expresses $\mathcal{L}_{MLE}$ using a Plackett-Luce rank preference model of each taster's scores. Among many applications, Plackett-Luce was employed to evaluate taste test results for sushi by Chen (2014), animal feed by Marden (1995) and wine by Bodington (2015a, 2015b). Plackett-Luce employs a preference probability for each wine ($\hat{\rho}_i$ for wine i taster t and totals of W wines and T tasters) that expresses the chance that the wine is most-preferred among the alternatives. Visualize the machinery in Equation (2) as calculating the probability of one branch on a probability tree. See Luce (1977), Plackett (1975), Marden (1995, p. 118), Alvo and Yu (2014, p. 151) and a simple, replicable example here in Section 6.

$$\text{LRS} = 2 \cdot (\mathcal{L}_{\text{MLE}} - \mathcal{L}_0) \tag{1}$$

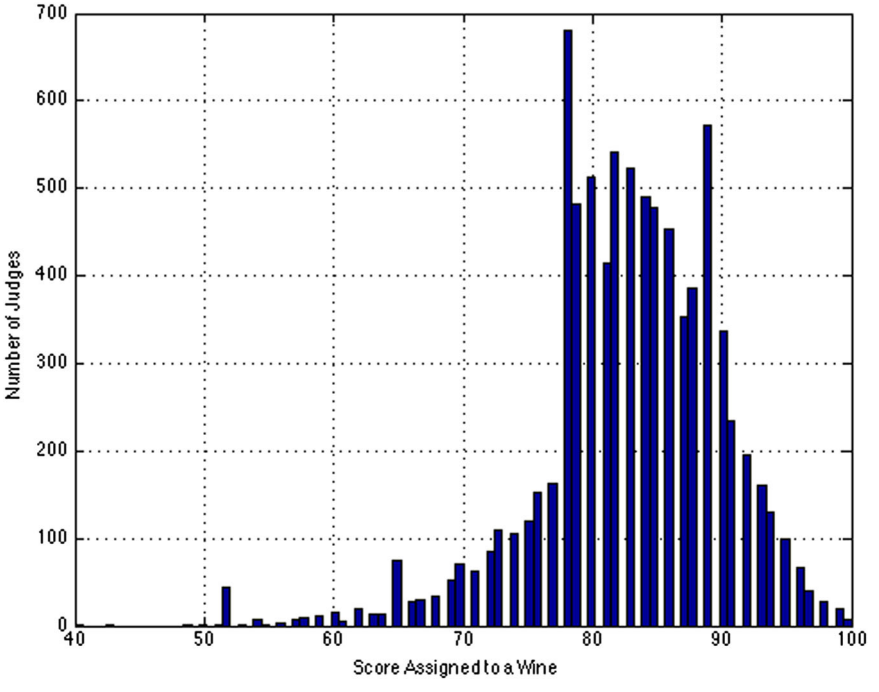

Figure 1. Frequency distribution of judges' scores.

$$\mathcal{L}_{\text{MLE}} = \sum_{t=1}^{T} \ln\left[\prod_{i=1}^{W_t}\left(\frac{\hat{\rho}_i}{\sum_{j=i}^{W_t}(\hat{\rho}_j)} | \boldsymbol{s}_t, \hat{\boldsymbol{\rho}}\right)\right] \tag{2}$$

$$\mathcal{L}_0 = \sum_{t=1}^{T} \ln(1/W_t!) \tag{3}$$

MATLAB code written by the authors to implement Equations (1) through (3) is available on request. Results were checked by replicating Marden (1995, p. 216) using rank data, equivalent score data and equivalent scores with ties. No taster scored every wine thus the number of wines in Equations (2) and (3) is taster-specific (W_t with a maximum W). See modeling of partial rankings in Marden (1995, p. 284). Ties were modeled by employing the expectation of $\hat{\rho}_i$ for the wines with tied scores and thus tied ranks. That approach is equivalent to but avoids the impossibility of evaluating trillions of rank vector permutations. For example, let $y_t = (A\overline{BC})$ be the rank order vector for a taster who assigned the same scores to wines B and C. For preference probabilities $\rho = (0.6,\ 0.3,\ 0.1)$ the Plackett-Luce model probability of (ABC) is 0.45 and of (ACB) is 0.15 for an expectation of 0.30. Setting $\rho = (0.6,\ 0.2,\ 0.2)$, calculating the Plackett-Luce probability also yields 0.30.

To conclude this overview of the 2016 Challenge, the p-value of the LRS in Equation (1) for the Challenge results is less than 10^{-6}. The preference orders implied by the judges' scores thus appear very unlikely to be the result of random assignments.

3. Local peaks, judge gender and judge nationality

In addition to showing that the pattern of judges' scores was not random, Figure 1 shows local peaks in frequency just below scores of 80, 85 and 90. Those peaks in scores correspond to just below the score thresholds for Bronze, Silver and Gold medals. According to IVV, that bunching of scores is due to the WoP practice of allowing judges to re-score wines to obtain what they consider a fair overall distribution of medals. It may also be due, in part, to a tendency to assign scores just below a medal threshold.

Of 151 judges, 40 were women and 111 were men. The women assigned a mean score of 82.2 with an SD of 7.2 and skewness of −1.1. The men assigned a mean score of 83.1 with an SD of 7.2 and skewness of −0.9. A two-sample, two-tailed Student's t-test of a difference in means showed that the distributions of the scores assigned by the genders were not significantly different. The average difference between the scores assigned by women and men to the same wines was −0.4 with an SD of 4.6 in a total of 1217 wines that were tasted by both genders. A one-sided t-test showed that difference was not significantly different from zero. Those results imply that women and men appear to assign about the same scores, and award the same medals, to the same wines. Those findings are also consistent with the food-related results in Corbin (2006) and the wine-related findings in Bodington (2017).

In addition to judges of both genders, the WoP's judges included both Portuguese and foreign nationals. A total of 127 judges were Portuguese and 24 were from other countries. The Portuguese judges assigned a mean score of 83.2 with an SD of 6.7 and skewness of −0.9. Foreign judges assigned a mean score of 81.7 with an SD of 8.3 and skewness of −0.9.

The *p*-value for a *t*-test of a difference in those means is approximately 0.80. Next, the average difference between the scores assigned to the same wine by the Portuguese and the foreign judges was 1.6 with an SD of 5.7 in a sample that includes all 1328 wines. As with the test concerning gender differences in scores above, a one-sided *t*-test showed that difference in scores is not significantly different from zero. Those results imply that Portuguese and foreign judges appear to assign about the same scores, and thus assign the same medals, to the same wines.

4. Preferences for types of wine and regions

Results for several types of WoP wine appear in Table 2 below. The types in Table 2 are defined by color, sparkling, sweetness and Origin Standard. The Origin Standard types are Protected Denomination of Origin (DOP, wine made from grapes grown and vinified in one of several small regions in Portugal) and Protected Geographic Denomination (IGP, wines vinified using at least 85% of grapes from one of several larger regions in Portugal). See www.winesofportugal.info for a map of the small DOP and larger IGP regions.

Inspection of Table 2 indicates that the means of scores for each type of wine were similar and within one SD of each other. The skewness of scores assigned to each type and the *p*-value of the LRS for the judges' scores were also calculated but they are not shown in Table 2 because the results are nearly uniform. All of the skewnesses are left except for red sparkling and Late Harvest with samples sizes too small to be statistically significant. All of the LRS *p*-values are less than 0.05, most are less than 0.001, and they indicate that the judges' scores are very unlikely to be random. As a check, the sample sizes shown in Table 2 are often less than the product of the number of wines multiplied by the number of tasters because not all tasters actually tasted all of the wines of each type.

Do judges assign higher scores to some types of wines? As a result, do judges tend to award better medals to some types of wines? Those questions were answered by calculating *p*-values for two-sample *t*-tests for differences in means. The result is the diagonal matrix of *p*-values in Table 3. Most of the *p*-values imply that the mean scores that

Table 2. Comparison of scores for different wine categories and origin standards.

Segment	Wines, #	Judges, #	Sample size, #	Mean score (SD)
All wines and tasters	1328	151	8445	82.8 (7.2)
Type				
Red	692	131	4384	82.1 (7.0)
Red sparkling	2	7	14	87.6 (4.5)
Rosé	51	47	321	80.8 (6.0)
Rosé sparkling	10	7	70	84.4 (4.8)
White	427	147	2710	82.5 (7.2)
White sparkling	45	12	294	83.7 (5.5)
Late Harvest	4	7	23	83.0 (4.4)
Sweet	97	24	629	89.5 (6.5)
Total	1328	–	8445	–
Origin standard				
DOP	836	151	5315	83.1 (7.4)
IGP	482	145	3065	82.4 (6.8)
Vintage year/variety	10	55	65	84.2 (6.0)
Total	1328	–	8445	–

Table 3. Two-sample *T*-test for a difference in mean score, *p*-value [*note to Editor, need to fit top row titles in same as row titles*].

	Red	…							DOP	…	
Type											
Red	1.00										
Red sparkling	0.06	1.00									
Rosé	0.42	0.04	1.00								
Rosé sparkling	0.31	0.25	0.16	1.00							
White	0.74	0.07	0.34	0.39	1.00						
White sparkling	0.42	0.16	0.21	0.78	0.54	1.00					
Late Harvest	0.65	0.11	0.32	0.59	0.80	0.77	1.00				
Sweet	0.03	0.43	0.02	0.08	0.03	0.05	0.04	1.00			
Origin standard											
DOP	0.50	0.09	0.26	0.54	0.64	0.75	0.96	0.05	1.00		
IGP	0.80	0.07	0.36	0.37	0.93	0.50	0.76	0.04	0.60	1.00	
Vintage year/variety	0.28	0.17	0.15	0.92	0.34	0.80	0.56	0.06	0.47	0.32	1.00

judges assign to different types of wine are about the same. However, sweet wines have the highest mean score in Table 2 and the lowest *p*-values in Table 3. The judges appear to prefer sweet wines. In addition, note that the difference in the sweet mean and all-wines mean straddles the 85-score line between Bronze and Silver. Sweet wines are getting more Silver medals than other wines. The mean score for red sparkling wines is also higher than all but sweet and some of its *p*-values are low, however, the sample size is small with just two wines.

Moving from categories of wine to regions, judges' scores for wines from over 30 regions are summarized below in Table 4. Although again not shown, the skewnesses

Table 4. Comparison of scores for wines from different regions.

DOP, not fortified				IGP			
Region	Wines	Sample Size	Mean score (SD)	Region	Wines	Sample Size	Mean score (SD)
Açores	1	6	81.2 (4.3)	Alentejano	204	1305	82.6 (6.5)
Alenquer	2	13	78.1 (5.2)	Algarve	21	132	83.6 (6.7)
Alentejo	73	465	82.2 (6.1)	Beira Atlântico	13	85	83.7 (5.8)
Bairrada	76	495	82.3 (6.7)	Duriense	2	12	85.1 (4.0)
Beira Interior	17	111	81.0 (8.5)	Lisboa	97	607	81.6 (7.0)
Bucelas	3	20	84.7 (3.9)	Minho	29	186	82.7 (6.5)
Carcavelos	1	6	91.7 (3.4)	P. de Setúbal	56	366	82.0 (6.9)
Dão	114	722	81.9 (6.3)	Tejo	48	292	81.5 (7.4)
Douro	236	1471	82.6 (7.4)	Terras da Beira	1	6	77.0 (9.6)
Madeirense	2	12	76.3 (7.2)	Terras do Dão	4	27	76.1 (9.1)
Óbidos	2	13	75.5 (13.7)	Transmontano	7	47	82.4 (7.0)
Palmela	14	92	80.4 (10.0)	Total IGP	482	3065	–
Pico	1	6	82.0 (2.8)				
Tejo	50	308	83.7 (6.2)	V. Year/Grape	10	65	84.2 (6.0)
Távora-Varosa	3	20	85.6 (4.9)				
Trás-os-Montes	27	178	78.3 (8.6)				
Vinho Verde	122	778	82.6 (6.7)				
Subtotal	744	4716	–				
DOP, Fortified							
Madeira	4	24	87.6 (7.6)				
Moscatel Douro	4	24	88.2 (5.9)				
Moscatel Setúbal	11	66	89.4 (6.4)				
Porto	73	485	89.9 (6.4)				
Subtotal	92	599	–				

Wines check total: 744 + 92 + 482 + 10 = 1328 matches totals in Tables 1 and 2
Sample Size check total: 4716 + 599 + 3065 + 65 = 8445 matches totals in Tables 1 and 2

were all left-handed, the LRS *p*-values were all less than 0.05 and most of the *p*-values were less than 0.01.

The results in Table 4 show that judges do not appear to have a strong preference for wines from any particular region over another. While wines from a few regions, such as DOP Carcavelos and DOP Óbidos, do have relatively high or low scores, the sample sizes were small for those regions and the means were still within one SD of the aggregate mean. With one exception, none of the differences in means were statistically significant. The exception was marginal and it was the DOP Fortified wines Madeira, Moscatel Douro, Moscatel Setúbal and Porto. The *p*-values for one-sided *t*-tests that those wines' scores were higher than the WoP mean were approximately 0.15 and higher. That is a marginal finding of significance. It is tempting to conclude that these fortified wines are of higher quality than the table wines. However, Loureiro et al. (2016) reported that when using the OIV sheet with red table wines, tasters preferred those wines with higher residual sugar and smoother mouthfeel. Therefore, it may be that the higher scores for fortified wines are due to how those factors are recorded on the OIV tasting form.

5. Dispersion of scores and potential randomness in awards

None of the 1328 wines in the 2016 Challenge were assigned the same score by all of the six or seven of the judges who evaluated each wine. Cao (2014) and Bodington (2012, 2015a, 2015b) posited that observed scores are a mixture of consensus, idiosyncratic and random expressions of preference. Numerous evaluations showed that there is often a consensus among some judges that some wines are better than others. See, for example, Ashenfelter and Quandt (2012) concerning the 1976 Judgement of Paris and Bodington's (2017) analysis of 23 tastings that involved over 900 wines. Next, economics literature is rich in examples of consumers' idiosyncratic preferences. Fillipello and Berg (1959) showed that wine tasters have individual preferences for and against oaky, buttery, sweet, citrus, strong fruit, tannin and other flavors. Finally, regarding randomness, Hodgson (2008) and Cao (2014) showed that some of the awards granted by wine judges do appear to be random.

Figure 2 below is a graph of the mean score for each of the Challenge wines arranged in order from highest mean to lowest. The error bars in Figure 2 indicate the range of one sample SD of scores about their respective mean, and those error bars show that dispersion in scores is material. Except for a lower SD on the very most-preferred wines, there was no statistically significant trend in SD with decreasing mean score.

The lower SD on the most-preferred wines, and the obvious non-linear trend in the means of scores, may have a combination of explanations. First, a potential explanation is that the widely employed 0-to-100 score scale actually induces the skewness in the distributions that appear in Figures 1 and 2. If judges award an average wine a score in the 80s then judges have more room for lower than higher scores. Randomness in score assignments within the unequal intervals above and below that average would lead to some left-hand skewness. The authors do not suggest here that such bias is the only or even a material explanation for the shape of the distributions in Figures 1 and 2. The authors do suggest that without further research on how judges actually score on a 0-to-100 scale, the possibility of some downward bias cannot be dismissed. Note also that this is one of several arguments in favor of comparing ranks in addition to scores.

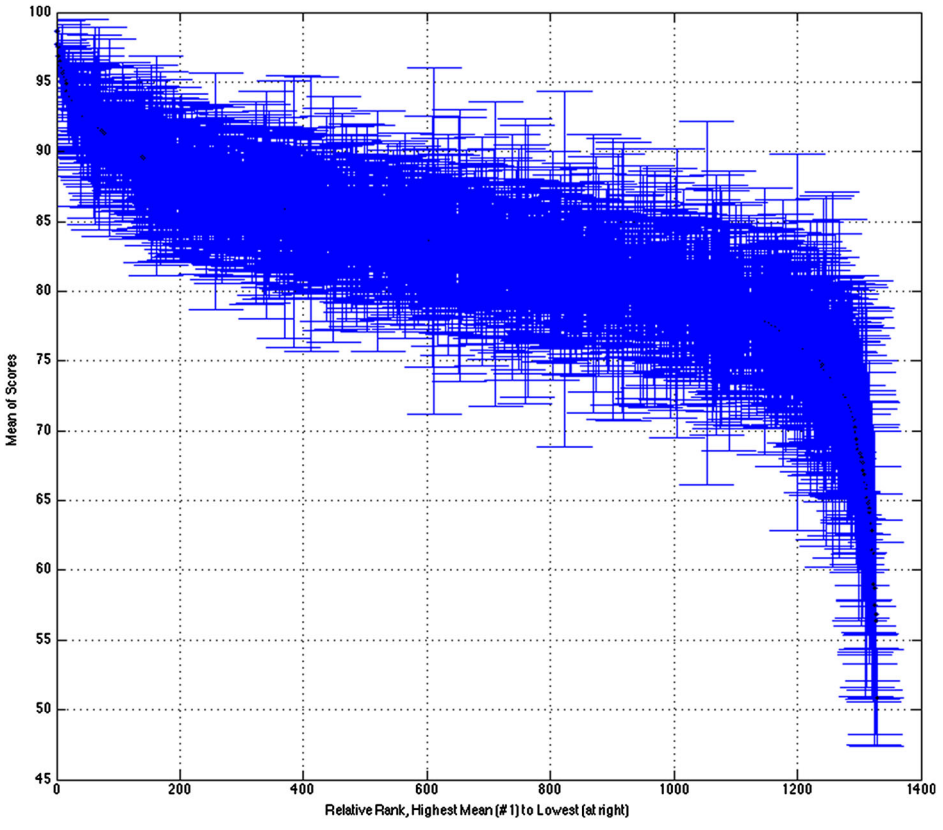

Figure 2. Means of scores for the 2016 challenge (with error bars to ± one SD of each wine's scores).

According to the WoP's tasting protocol, and subject to the percentage limitations described in Section 2, medals were granted to wines with mean scores over 80, 85 and 90. For every wine with a mean score ≥80 and arranged in descending order, the mean score for each wine appears in Figure 3 below. The mean scores ≥80 were the same as those in Figure 2 except that the DOP Fortified wines have been removed from the sample. As found in Section 4, the scores for the DOP Fortified wines were marginally higher-than-average and including them here may bias the results depicted in Figure 3. The resulting sample contains 902 wines and 5742 observations from 151 judges.

Marks (2015, p. 326) described a judge's observed score for a wine as a weighted average of latent sub-scores that the judge assigns to different aspects of a wine. There may be uncertainty in those sub-scores. Re-sampling, asking each judge to re-score each wine, could yield a different mean if there was randomness in the judges' assignments. Re-sampling with different judges could yield a different mean if there was a difference in the judges' non-random idiosyncratic preferences. Re-sampling with either the same or different judges could yield a different mean if any basis for assignments of scores was non-stationary. On that foundation, error bars in Figure 3 show the range of one SD about each sample mean. The errors bars thus show uncertainty about what award a wine may have earned. See calculation of the SD of a sample mean in, for example, Crawshaw and Chambers (2001, p. 438).

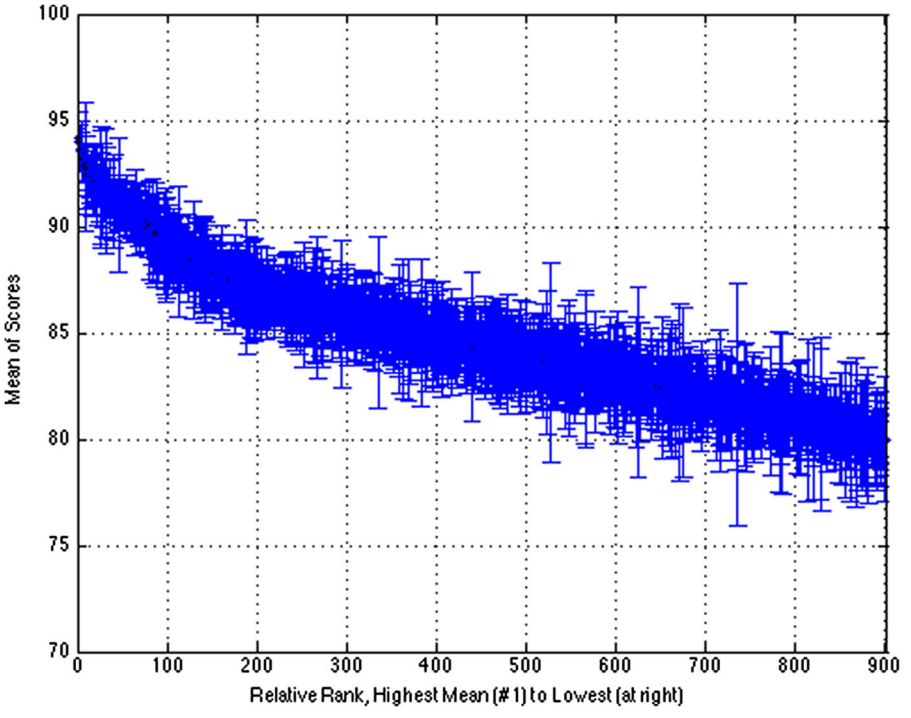

Figure 3. Non-fortified wines with means of scores ≥80 for the 2016 challenge (with error bars to ±one SD of each wine's sample mean of scores).

The results in Figure 3 showed that uncertainty in whether or not a wine's mean score is over the medals' score thresholds can be material. Many of the wines could trade places and some may have qualified for a particular medal by chance alone. However, none of those findings imply that medal winners are just random. As shown in Bodington (2012, p. 187) the greater the difference in mean score or rank, the lower the *p*-value for a hypothesis test that two wines are equally preferred. Close calls in Figure 3, such as low-Gold (near but >90) compared to high-Silver (near but <90), are more likely to be separated by randomness than high-Gold (near 100) from low-Gold (near but >90), high-Silver from low-Silver and so on.

6. Methods of breaking ties

Of 902 wines displayed in Figure 3 above, there were ties in mean scores for 709. Nearly 80% of the wines tied with at least one other wine. The WoP needs a reliable method of breaking those ties.

There are many methods of calculating judges' aggregate relative assessment of quality and preferences. Arrow's (1963) famous general possibility theorem is that there is no method of combining ranked individual expressions of preference into an aggregate that does not have logical flaws. Restated without the double negative, every method of combining expressions of individual rankings into an aggregate has logical flaws. See also Marden (1995, p. 134) for discussion of this issue. The advantage of the mean of scores employed in WoP is that a mean is easy to understand and calculate. Disadvantages are that it can lead to numerous ties and, in some cases, results that violate the choice

axiom of transitivity. To break ties and preserve transitivity, WoP could employ other methods of aggregating judges' scores into an index of relative quality or preference. Those methods include the Borda counts and Shapely values described in Ginsburgh and Zang (2014), the sign test described in Olkin, Lou, Stokes, and Cao (2015, p. 23) and ranking models, including Plackett-Luce, such as those described in Olkin et al. (2015, p. 24), Marden (1995) and Alvo and Yu (2014).

Using Plackett-Luce in Equation (2) as an example, the MLE solution yields a vector of preference probabilities ($\hat{\rho}_i$). There is one $\hat{\rho}_i$ for each wine, and each $\hat{\rho}_i$ expresses the judges' aggregate relative preference for each wine. Favoring wines that have the highest $\hat{\rho}_i$ can break ties. Although there were 709 ties among wines with a mean score over 80, the Plackett-Luce vector of $\hat{\rho}_i$ for those same wines contained no ties. As a check and simple example, let the scores (s_t) assigned by three judges to three wines (A, B, C) be $s_1 = (85,\ 90,\ 95)$, $s_2 = (90,\ 82,\ 98)$ and $s_3 = (87,\ 90,\ 85)$. The sums of scores are $s = (262,\ 262,\ 278)$ thus wines A and B are tied. The MLE solution for the Plackett-Luce preference probabilities is approximately $\hat{\rho} = (0.25,\ 0.30,\ 0.45)$ thus the tie is broken and the preference order is (C, B, A).

7. Conclusions

The results of the 2016 WoP Challenge provide a large sample that enables analysis of scoring behavior and wine judge preferences that are not evident in the small samples that are typical of most tastings or, due to small sample sizes, cannot be tested to widely accepted levels of statistical significance. This analysis showed that the distribution of scores and the *p*-value for a Plackett-Luce model test imply that the WoP results are very unlikely to be random. The distribution of scores showed local peaks just below the score thresholds for Gold, Silver and Bronze awards. Student's *t*-tests showed that there were no significant differences in scores assigned by gender-of-judge, nationality-of-judge and to wines from different regions. However, judges do appear to assign higher scores to sweet wines than to other types of wines. While dispersion in scores was material, results also showed that the strengths of judges' preferences against the least-preferred wines were stronger than those in favor of the most-preferred wines. The dispersion in scores also showed that some wines may have received their awards by chance.

Ties between means of scores can be frequent in a large sample. Nearly 80% the WoP wines tied with at least one other wine. Without abandoning using means of scores to order wines and grant awards, because that metric is easy to calculate and communicate, ties can be broken by employing one or more of many other methods of transforming scores into relative measures of quality or preference.

Acknowledgements

The authors thank two reviewers for their insightful and constructive comments. Remaining errors are the responsibility of the authors alone. Authors are also grateful to Mr. Frederico Falcão, IVV President, for supplying WoP Challenge results.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Fundação para a Ciência e a Tecnologia: [Grant Number UID/AGR/04129/2013 (LEAF)].

ORCID

Manuel Malfeito-Ferreira http://orcid.org/0000-0002-7985-963X

References

Alvo, M., & Yu, L. H. (2014). *Statistical methods for ranking data*. New York: Springer. 273 pp.
Arrow, K. J. (1963). *Social choice and individual values* (2nd ed.). New York, NY: Wiley.
Ashenfelter, O., & Quandt, R. E. (2012). Analyzing a wine tasting statistically. *Chance*, *12*(3), 16–20.
Bodington, J. (2012). 804 tastes: Evidence on randomness, preferences and value from blind tastings. *Journal of Wine Economics*, *7*(2), 181–191.
Bodington, J. (2015a). Evaluating wine-tasting results and randomness with a mixture of rank preference models. *Journal of Wine Economics*, *10*(01), 31–46.
Bodington, J. (2015b). Testing a mixture of rank preference models on judges' scores in Paris and Princeton. *Journal of Wine Economics*, *10*(2), 173–189.
Bodington, J. (2017). Wine, women, men and type II error. *Journal of Wine Economics*. doi:10.1017/jwe.2017.8
Cao, J. (2014). Quantifying randomness versus consensus in wine quality ratings. *Journal of Wine Economics*, *9*(2), 202–213.
Chen, W. (2014). *How to order sushi* (PhD dissertation). Harvard University, Cambridge, MA.
Cliff, M., & King, M. (1997). The evaluation of judges at wine competitions: The application of eggshell plots. *Journal of Wine Research*, *8*, 75–80.
Cliff, M., & King, M. (1999). Use of principal component analysis for the evaluation of judge performance at wine competitions. *Journal of Wine Research*, *10*, 25–32.
Corbin, C. (2006). *Sex differences in taste preferences in humans. Literature review for psychology 451 under D. Pittman*. Spartanburg, SC: Wofford College.
Crawshaw & Chambers. (2001). *Advanced level statistics*. Cheltenham: Nelson Thornes. 688 pp.
Fillipello, F., & Berg, H. W. (1959). The present Status of consumer tests on wine. *American Journal of Enology and Viticulture*, *10*(1), 8–12.
Ginsburgh, V., & Zang, I. (2014). Shapley ranking of wines. *Journal of Wine Economics*, *7*(2), 169–180.
González, J., Sánchez-Sáenz, C., & Mejias-Barrera, P. (2014). Stochastic model for the process of wine award: Visualization and quantification. *Ciência e Técnica Vitivinícola*, *29*, 53–59.
Herbst, F., & Von Arnim, C. (2009). The role and influence of wine awards as perceived by the South African wine consumers. *Acta Commercii*, *9*, 90–101.
Hodgson, R. T. (2008). An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, *3*(2), 105–113.
Honoré-Chedozeau, C., Ballester, J., Chatelet, B., & Valérie Lempereur, V. (2015). *Wine competition: From between-juries consistency to sensory perception of consumers*. Bio Web of conferences. 5, 03009.
Loureiro, V., Brasil, R., & Malfeito-Ferreira, M. (2016). A New wine tasting approach based on emotional responses to rapidly recognize classic European wine styles. *Beverages*, *2*(1), 6.
Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*(3), 215–233.
Marden, J. I. (1995). *Analyzing and modeling rank data*. London: Chapman & Hall, 329 pp.
Marks, D. (2015). Seeking the veritas about the vino: Fine wine ratings as wine knowledge. *Journal of Wine Research*, *26*(4), 319–335.
OIV. (2009). *OIV Standard for international wine competitions and spirituous beverages of vitivinicultural origin* OIV-CONCOURS 332A-2009 Retrieved from http://www.oiv.int/public/medias/4661/oiv-concours-332a-2009-en.pdf

Olkin, I., Lou, Y., Stokes, L., & Cao, J. (2015). Analyses of wine-tasting data: A tutorial. *Journal of Wine Economics*, *10*(1), 4–30.

Peattie, S. (1995). Promotional competitions – A winning technique for wine marketing. *International Journal of Wine Marketing*, *7*(3), 31–48.

Plackett, R. L. (1975). The analysis of permutations. *Applied Statistics*, *24*, 193–202.

Scaman, C., Dou, J., Cliff, M., Yuksel, D., & King, M. (2001). Evaluation of wine competition judge performance using principal component similarity analysis. *Journal of Sensory Studies*, *16*, 287–300.