Evaluation of Wine Judge Performance Based on a Simple T-Test

Jing Cao and Lynne Stokes

In a previous study on wine judge performance, Cao and Stokes (2010) introduced a Bayesian ordinal model to identify which one or ones of three judge characteristics, bias, discrimination ability, and random variation, are responsible when judges disagree. Judge bias measures the systematic difference between a judge’s score and the average score from all the judges. Judge discrimination measures a judge’s ability to distinguish wines based on their quality. Judge variation measures the size of the random component of variability in a judge’s assessment of wine quality. The model provides insight into what makes judges differ. However, the estimates of the parameters do not have a closed form and they are produced via sampling-based Bayesian inference. This may present an obstacle for wine practitioners on the evaluation of judge performance. In this paper, we propose to use simple t-tests to evaluate wine judge performance through the same three characteristics.

The t-test requires replicate samples on a set of wines. We use the data from the California State Fair Wine Competition in 2009. The judges were instructed to provide letter scores (e.g., No award, Bronze, Bronze+). Letter scores were later transformed to numerical scores ranging from 80 points to 100 points. In the dataset, 8 distinct scores were recorded (80, 84, 86, 88, 90, 92, 94, 96). In each of the 17 panels, there are three samples of 4 different wines, where each triplicate was poured from the same bottle. The data from one of the 17 panels is presented in Table 1. Following the notation used by Hodgson (2008), we use $J_1$ to $J_4$ to denote the four judges in a panel. The three $R_1$ values represent the scores given to the replicates of the first wine tested, with the scores of the remaining wines denoted by $R_2$, $R_3$, and $R_4$. 
Table 1: Raw Data from Panel A

<table>
<thead>
<tr>
<th></th>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>80</td>
<td>86</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>R1</td>
<td>84</td>
<td>86</td>
<td>80</td>
<td>86</td>
</tr>
<tr>
<td>R1</td>
<td>84</td>
<td>80</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>R2</td>
<td>86</td>
<td>86</td>
<td>86</td>
<td>84</td>
</tr>
<tr>
<td>R2</td>
<td>84</td>
<td>88</td>
<td>92</td>
<td>84</td>
</tr>
<tr>
<td>R2</td>
<td>80</td>
<td>96</td>
<td>86</td>
<td>86</td>
</tr>
<tr>
<td>R3</td>
<td>84</td>
<td>90</td>
<td>86</td>
<td>92</td>
</tr>
<tr>
<td>R3</td>
<td>80</td>
<td>84</td>
<td>80</td>
<td>86</td>
</tr>
<tr>
<td>R3</td>
<td>86</td>
<td>88</td>
<td>96</td>
<td>94</td>
</tr>
<tr>
<td>R4</td>
<td>90</td>
<td>86</td>
<td>90</td>
<td>94</td>
</tr>
<tr>
<td>R4</td>
<td>80</td>
<td>84</td>
<td>96</td>
<td>90</td>
</tr>
<tr>
<td>R4</td>
<td>84</td>
<td>84</td>
<td>86</td>
<td>88</td>
</tr>
</tbody>
</table>

\[ \bar{y}_{j.} = 83.5 \]
\[ \bar{y}_{-j.} = 87 \]
\[ t^b_j = -7 \]
\[ p\text{-value} = 0.02 \]

**T-test on Judge Bias.** Suppose there are \( M \) judges who are instructed to assign scores to \( N \) wines, where each wine has \( K \) replicate samples. In Table 1, we have \( M = 4 \), \( N = 4 \), and \( K = 3 \). Let \( y_{ijk} \) be the observed score assigned by judge \( j \) on the \( k \)th replicate of wine \( i \), and \( \bar{y}_{j.} \) be the mean score judge \( j \) assigned to the set of wines including the replicates. The t-test statistic on judge bias for judge \( j \) is

\[
t^b_j = \frac{\bar{y}_{j.} - \bar{y}_{-j.}}{s^b_{-j.}} \quad j = 1, \ldots, M,
\]

where \( \bar{y}_{-j.} = (\sum_{r=1}^{M} \bar{y}_{r.} - \bar{y}_{j.})/(M - 1) \) is the mean score assigned by all the judges excluding judge \( j \) to the whole set of wines and \( s^b_{-j.} \) is the standard deviation of \( \{\bar{y}_{r.}, r \neq j\} \), which are the mean scores assigned by all the judges excluding judge \( j \) in the panel. The t-test has
\((M - 2)\) degrees of freedom.

In Table 1, we list \(\bar{y}_{j}\) and \(\bar{y}_{-j}\) for the 4 judges in the panel. Compared to the other three judges, judge \(J_1\) tends to assign lower scores to the group of wines. Based on test (1), we have

\[
\bar{t}^0 \equiv \frac{83.5 - \frac{86.5 + 87 + 87.5}{3}}{SD(86.5, 87, 87.5)} = \frac{83.5 - 87}{0.5} = -7,
\]

where \(SD()\) denotes the function to calculate standard deviation. The \(p\)-value for a two-sided test is 0.02, indicating that the bias from judge \(J_1\)'s scoring pattern is significant. We also report the test statistic and the \(p\)-value for the other 3 judges: no obvious bias is observed.

**T-test on Judge Discrimination.** A single score assigned by a judge on a wine contains both the judge’s true evaluation of the wine and the measurement error that occurred on this specific tasting. The judge’s true evaluation of the wine can be ideally obtained by the average of scores on the same wine if the judge has tasted it many times. Based on this rationale, we assume that a judge’s mean score over the replicate samples for each individual wine (i.e., \(\bar{y}_{ij} = \sum_{k=1}^{K} y_{ijk}\)) contains the judge’s true evaluation of the wine. Then the distribution of the mean scores, \(\bar{y}_{ij}\), can tell us how well the judge can discriminate among different wines. Let \(d_j = SD(\bar{y}_{ij}, i = 1, \ldots, N)\) be the standard deviation of the mean scores \((\bar{y}_{ij})\) over the replicate samples for judge \(j\). We use it to measure the spread of judge \(j\)'s evaluation of different wines. The t-test statistic on judge discrimination for judge \(j\) is

\[
t^d_j = \frac{d_j - \bar{d}_-}{s^d_{-j}} \quad j = 1, \ldots, M,
\]

where \(\bar{d}_- = \frac{\sum_{r=1}^{M} d_r - d_j}{(M - 1)}\) is the average discrimination level of all the judges excluding judge \(j\) and \(s^d_{-j}\) is the standard deviation of \(\{d_r, r \neq j\}\). The t-test has \((M - 2)\) degrees of freedom.

In Table 2, we list the data from another panel. For each judge in the panel, we also report the average score of the three samples for each wine. Note that judge \(J_2\) has a range of about 5 points \((91.33 - 86)\) to differentiate the four wines. By comparison, judges \(J_1, J_3,\) and \(J_4\) all have a range of more than 10 points \((90.67 - 80, 96 - 85.33, 96 - 85.33)\) in their evaluation. Based on test (2), \(d_2 = SD(86, 88, 91.33, 89.33) = 2.24\), we have

\[
t^d_2 = \frac{2.24 - \frac{5.11 + 4.67 + 4.65}{3}}{SD(5.11, 4.67, 4.65)} = \frac{2.24 - 4.81}{0.26} = -9.88.
\]
Since we are only concerned with small discrimination, we compute the $p$-value for the one-sided hypothesis, i.e., that judge $j$’s discrimination level is below the average level from the other judges. The $p$-value for judge $J2$ is 0.01, indicating that the judge has a significantly lower discrimination level than the others in the panel.

Table 2: Raw Data from Panel B

<table>
<thead>
<tr>
<th></th>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>80</td>
<td>88</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>R1</td>
<td>80</td>
<td>90</td>
<td>90</td>
<td>96</td>
</tr>
<tr>
<td>R1</td>
<td>80</td>
<td>80</td>
<td>88</td>
<td>92</td>
</tr>
<tr>
<td>R2</td>
<td>92</td>
<td>92</td>
<td>86</td>
<td>96</td>
</tr>
<tr>
<td>R2</td>
<td>84</td>
<td>86</td>
<td>84</td>
<td>96</td>
</tr>
<tr>
<td>R2</td>
<td>96</td>
<td>86</td>
<td>86</td>
<td>90</td>
</tr>
<tr>
<td>R3</td>
<td>90</td>
<td>90</td>
<td>92</td>
<td>96</td>
</tr>
<tr>
<td>R3</td>
<td>80</td>
<td>92</td>
<td>92</td>
<td>80</td>
</tr>
<tr>
<td>R4</td>
<td>86</td>
<td>92</td>
<td>96</td>
<td>80</td>
</tr>
<tr>
<td>R4</td>
<td>80</td>
<td>86</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>R4</td>
<td>80</td>
<td>90</td>
<td>96</td>
<td>96</td>
</tr>
</tbody>
</table>

$d_j$ 5.11  2.24  4.67  4.65
$\bar{d}_{-j}$ 3.85  4.81  4.00  4.01
$t^d_j$ 0.90  -9.88  0.43  0.42
$p$-value 0.77  0.01  0.65  0.64

Note: The average scores of triplicate samples for each judge on each wine are listed in ( ).

**T-test on Judge Variance.** The variation over the replicate samples of individual wines stems from measurement error in a judge’s assessment of the wines. Let $s_{ij}$ be the standard deviation over the replicate samples of the $i$th wine tasted by the $j$th judge, and $\bar{s}_j = \sum_{i=1}^{N} s_{ij}$. Then $\bar{s}_j$ directly measures the amount of measure error for judge $j$. The
t-test statistic on judge variation for judge $j$ is

$$ t^v_j = \frac{\bar{s}_{-j} - \bar{s}_{-j}}{s^v_{-j}} \quad j = 1, \ldots, M, $$

(3)

where $\bar{s}_{-j} = (\sum_{r=1}^{M} \bar{s}_{r} - \bar{s}_{-j})/(M - 1)$ is the average level of measurement error for all the judges excluding judge $j$, and $s^v_{-j}$ is the standard deviation of $\{\bar{s}_{r}, r \neq j\}$. The t-test has $(M - 2)$ degrees of freedom.

In Table 3, we present the data from a third panel. Along with the raw data, we also report the standard deviation of the scores over the triplicate samples for each wine. Judge $J2$ is consistent only in evaluating the second wine ($R2$), which was voted unanimously to be the worse. For the other three wines, judge $J2$ shows substantially more variation than the other three judges. Note that for two wines ($R3$ and $R4$), judge $J2$ assigned the lowest score (80) and the highest score (96) to the same wine. For the other wine ($R1$), the judge assigned the lowest score (80) and the sub-highest score (94) to the replicates.

Based on test (3), $\bar{s}_{.2} = Mean(7.57, 0, 8.72, 8.33) = 6.15$, we have

$$ t^d_2 = \frac{6.15 - \frac{2.23 + 1.44 + 1.37}{3}}{SD(2.23, 1.44, 1.37)} = \frac{6.15 - 1.68}{0.48} = 9.33. $$

Since we are only concerned with large variation, we compute the $p$-value for the one-sided hypothesis, i.e., that judge $j$’s variation level is above the average level from the other judges. The $p$-value for judge $J2$ is 0.01, indicating that the judge has a significantly higher variation level than the others in the panel.

**Discussion.** In this paper, we have used the same examples included in Cao and Stokes (2010). The conclusions based on the t-tests are consistent with the ones based on the Bayesian model. In the following we summarize the pros and cons of the proposed t-tests compared to the Bayesian analysis. The pros are: 1) the tests are very simple to understand and conduct; 2) the tests can provide useful information on the three characteristics of judge’s scoring patterns. The cons include: 1) the tests on judge discrimination and variation require replicate samples of wines; 2) the test on judge discrimination can not simultaneously identify judges with negative discrimination, whose evaluation criterion tends to be contrary to the majority opinion. Taking advantage of the modeling approach, the Bayesian model does not have these constraints. In addition, the t-tests can only be used to evaluate judge
Table 3: Raw Data from Panel C

<table>
<thead>
<tr>
<th></th>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>92</td>
<td>94</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>R1</td>
<td>94 (2)</td>
<td>80 (7.57)</td>
<td>88 (1.15)</td>
<td>94 (2)</td>
</tr>
<tr>
<td>R1</td>
<td>90</td>
<td>92</td>
<td>88</td>
<td>92</td>
</tr>
<tr>
<td>R2</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>R2</td>
<td>84 (2.31)</td>
<td>80 (0)</td>
<td>80 (0)</td>
<td>84 (2.31)</td>
</tr>
<tr>
<td>R2</td>
<td>84</td>
<td>80</td>
<td>80</td>
<td>84</td>
</tr>
<tr>
<td>R3</td>
<td>88</td>
<td>96</td>
<td>88</td>
<td>88</td>
</tr>
<tr>
<td>R3</td>
<td>92 (2.31)</td>
<td>94 (8.72)</td>
<td>84 (2.31)</td>
<td>88 (1.15)</td>
</tr>
<tr>
<td>R3</td>
<td>88</td>
<td>80</td>
<td>84</td>
<td>90</td>
</tr>
<tr>
<td>R4</td>
<td>84 (2)</td>
<td>84 (8.33)</td>
<td>80 (2.31)</td>
<td>90 (0)</td>
</tr>
<tr>
<td>R4</td>
<td>88</td>
<td>96</td>
<td>84</td>
<td>90</td>
</tr>
</tbody>
</table>

\[ \bar{s}_{j} = 2.23, \bar{s}_{-j} = 2.99, t^d_j = -0.28, p-value = 0.60 \]

Note: The standard deviations of the scores on triplicate samples listed in ( ).

performance, while the Bayesian model can produce more accurate ranking of wines after adjusting judges’ scoring patterns.

REFERENCES
