Evaluation of Wine Judge Performance through Three Characteristics: Bias, Discrimination, and Variation

Jing Cao (jcao@mail.smu.edu) and
Lynne Stokes (slstokes@mail.smu.edu)

Department of Statistical Science
Southern Methodist University

Judge performance is a critical component of a wine competition's success. A number of studies have shown that wine judges may differ considerably in their opinions. We proposed to use an ordinal modeling approach to evaluate wine judge performance more closely through three characteristics of judge's performance: bias, discrimination ability, and variation. We applied the method to analyze the data from a major U.S. Wine Competition.

Our model uses three judge characteristics to describe judging behavior, which then can be used to explain judge disagreements. Judge bias measures the systematic difference between a judge's score and the average score from all the judges. Based on the bias, we 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 degree of randomness in a judge's assessment of wine quality. Each judge receives a score on each characteristic of his judging performance. This provides insight into what makes judges differ.

Our main findings from the data analysis are that 1) there are a number of judges with positive or negative bias and a low level of discrimination, and 2) there are a few judges showing a negative level of discrimination or having a large variance. Among these problems regarding judging styles, bias and a low level of discrimination may be easier to overcome. Judges with these two issues still rank the wines in a similar order compared to the other judges, and they only need to adjust the center and scale of their scores to be more consistent with the rest of the panel. Under the assumption that the majority opinion is in conformity with wine quality, the more serious problems for judges are having a negative level of discrimination and a large variance, where the former indicates the judges are using criteria contrary to the majority opinion and the latter indicates that there is much randomness in the judges' score.

The purpose of this analysis was to provide more insight on wine judge performance. With the information, wine judges may be trained more efficiently, competition coordinators in wine competition will have clearer guidelines in future judge selection, and eventually, the accuracy of wine competition may be improved.

In the paper, we also described three main advantages of the ordinal model which are not demonstrated by the current data analysis and which improve on current methods of

judge ranking. The first is that the model enables us to extract information about judge performance without replicate data. The second is that the model can provide estimates of wine quality, which can be used to rank the wines. The third is that because of the modeling approach, we can provide a way to answer a variety of hypothetical questions in a structured way. For example, the model can be used to determine 1) how many judges are needed to achieve a certain quality of evaluation (i.e., giving a "good" wine a sufficiently high chance of being evaluated correctly); 2) whether there are hidden biases for or against certain grape varieties (i.e., Chardonnay, Merlot, and White Zinfandel) by individual judges; 3) which of the different competition arrangements will give the most reliable results. Thus the competition manager could, through the model, virtually alter the competition design, say by having one additional judge per panel, or by reducing the bias of all judges by training them to use the scale in a particular way, or some other method of his choosing, and predict the effect of changes to competition design.