Do expert tasters evaluate wines consistently? A statistical analysis and a proposal for improvement

Jonas De Maere

Abstract[1]

Choosing the right bottle out of hundreds of wines is quite a challenge for the average consumer. In supermarkets, where most retail wines are sold, it is all too often not an easy task for the occasional wine drinker. Therefore, consumers tend to turn to wine experts' judgments, brands, and awards to inform their purchasing decisions.

A bottle decorated with a medal distinguishes itself from the other bottles on the shelves. For the consumers, a medal stands for quality. This makes wine competitions commercially important for many producers. For example*, Concours Mondial de Bruxelles*, one of the world's largest wine competitions represents, with more than 8000 participating samples, about half a billion marketed bottles. Nonetheless, relatively little attention has been paid to the scoring and evaluation methods that are used to obtain the results of the competition.

We believe that the consumer can only rely on a medal if he/she can be sure that expert wine tasters judge objectively enough. Therefore, the reliability of evaluation by jury panels in wine competitions deserves special attention. Are wine competitions' results biased by individual taster deviations and inconsistent assessment? Earlier research has shown that assessment of quality is the result of an individual's conceptions and previous experiences, and even experienced wine tasters tend to incorporate personal taste and preferences in their judgments. These findings indicate that evaluations by professional wine tasters are not entirely consistent.

The objective of this thesis is twofold. First, we wish to get an insight into the degree of consistency with which expert tasters evaluate wines. This allows us to develop a correction that can be used to reduce the inconsistency in the taster evaluations. Second, we enquire whether this inconsistency affects the competition results. That is, would the competitions results be different if tasters judged wines more consistently?

We use state-of-the-art statistical methods to examine the pattern of judges' evaluations in the Best Belgian Wine Competition of 2012 to answer these two questions. The Best Belgian Wine Competition is organized by the Flemish Sommelier Association (VVS) and takes place annually in Bruges, Belgium. Based on the judges' evaluation results, the jury selects the best Belgian wine of the year, as well as the second and the third wine of the year. The judges are members of the VVS and can thus be considered as professional wine tasters.

We use a set of statistical techniques to examine the taster's judging consistency. As mentioned earlier, our first research question is to isolate taster deviations. We use box plots to graphically represent these deviations.

[1] Abstract of the full thesis with the same title submitted to the Weinakademie Österreich in partial fulfillment of the requirements for the degree of Weinakademiker.

For each wine, we compute the mean score and standard deviation. For each taster, we compute his/her deviation by subtracting a given wine's mean score from the taster score. We then compute each taster's relative deviation by dividing the taster's deviation by the wine's standard deviation. Finally, we use the absolute values of the relative taster deviation and calculate their mean per taster over the wines he/she evaluated. We call the resulting measure taster's mean relative deviation. It indicates by how much a taster deviates from the mean wine score *on average, relative to other tasters' deviations*, which answers our first question.

We then use the obtained measure to answer our second research question, whether tasting deviation affects the ranking of the wines in a competition. For each taster's score, we compute a correction. We multiply the taster's mean relative deviation by each wine's standard deviation and give it the same sign as in the taster's original deviation. This correction indicates by how much a taster would deviate in evaluating a particular wine, *if he/she always deviated to the same degree from the jury*.

We use it to compute the new, corrected score for each wine, defined as the average corrected score of jury members for that wine (i.e., in the same manner as the original competition scores, except that we use *corrected* individual taster's scores for the calculation). Finally, we rank the wines according to the corrected scores and compare it with the original ranking to see if increased consistency matters for the ranking.

Our analyses produce very interesting results. Our box plots show that there is considerable variance in the tasters' scores (the boxes in the box plots are "wide"), and the scores are scattered across quite a large range, meaning that individual taster's scores can be located relatively far from the "true" score (the mean relative deviation equals, on average, 0.77 standard deviations).

To answer our second research question, we applied our correction to compute the taster's corrected scores for each wine and generated a box plot for the corrected scores. Compared to original wine scores, the corrected wine scores have nearly the same average. However, the individual scores are much less scattered and are much closer to the "true" scores. In other words, the corrected scores take out some of the "noise" and make the overall evaluation more consistent, as there is more agreement among the corrected scores than among original scores.

The practical question now is: Does this affect the ranking of the wines? The answer is yes. Although the very best wines remain the same, the original rankings and the new rankings based on corrected scores generally differ to some extent. This is true for both red and white wines. Based on the corrected scores, we can say that individual taster biases affect the competition results.

Our results lay the basis for the development of a more objective, consistent and reliable method for the evaluation of wines by expert panels. First, we support the practices observed by the Flemish Sommelier Association. Organizers need to make sure that they select professional and experienced jury members, instruct them appropriately, and calibrate the marks before the tasting. Second, since even professional tasters exhibit some inconsistency in evaluating wines, it may be beneficial for wine competition organizers to correct their results for tasting inconsistencies. Besides the correction proposed here, we make suggestions for future research, such as decomposing taster deviations into two conceptually distinct components: random and systematic.

Our findings also have important implications for the wine market. The consumer can probably find medals of a hundred of wine competitions on the wine bottles on the supermarket shelves. However, it is very difficult for the consumer to tell how reliable the wine competitions are. Therefore, we advise competition organizers to communicate their practices and publicize their results for the benefit of the consumer.