Journal of Intelligent Manufacturing, Vol.1 Iss.1 2025, pp.20-27
ISSN :3078-4115

Wine tasting

Yifei Zhao *

*Corresponding author

Abstract: Wine evaluation remains a critical challenge in the winemaking industry. This study developed three analytical models—variance analysis, principal component analysis (PCA), and multiple linear regression—to address discrepancies among tasters, grape grading systems, correlations between grape characteristics and wine quality parameters, as well as the impact of physicochemical indicators on wine quality. The research established a scientific evaluation framework through three key steps: First, incomplete or erroneous data were eliminated through analysis, followed by the construction of a single-factor variance analysis model using Matlab. Results revealed no significant differences in red wine scoring among tasters, while notable variations were observed in white wine scoring. Next, principal component analysis was employed to identify key components, with min-max normalization applied to derive comprehensive scores for both red and white wines, enabling proper classification. Multiple linear regression models demonstrated linear correlations between compounds in grape varieties and those in wine products. Finally, correlation analysis and regression analysis confirmed that physicochemical indicators influence wine quality but cannot serve as standalone evaluation criteria. The proposed models enhanced credibility by eliminating errors and optimizing principal component analysis processes.

Key words: One-way ANOVA, ANOVA, normalization, multiple linear regression

1. Introduction

The quality of wine is determined through evaluation by certified wine tasters[1]. Each taster first samples the wine, then scores various indicators, calculates the total score, and finally determines the wine's quality. The quality of wine is directly related to the quality of the grapes used for winemaking[2]. The physicochemical properties tested on both the wine and grapes can, to some extent, reflect their respective quality levels.

Appendix 1 presents evaluation results of wines from a specific year, while Appendices 2 and 3 provide component data for both the wines and their grape varieties. The mathematical model is designed to address the following questions: 1. Analyze whether there are significant differences in the wine evaluators 'results between the two groups listed in Appendix 1, and determine which group's findings are more reliable; 2. Classify grapes based on their physicochemical properties and wine quality; 3. Explore correlations between grape and wine physicochemical parameters; 4. Evaluate how these parameters influence wine quality, and demonstrate the feasibility of using them as quality assessment criteria.

2. basic assumption

It is assumed that there is no error in the wine samples presented to the tasters and no emergency occurs during the tasting process;

It is assumed that the brewing process and storage conditions have no effect on the quality and physical and chemical indexes of the wine;

It is assumed that the physical and chemical indexes and aromatic substances of the wine grapes and wine do not change within a certain period of time;

3. Model building and solving of problem 1

3.1 Problem analysis

Given the errors in the provided data, we need to exclude obviously erroneous data from the given dataset to accurately reflect the differences and credibility issues between the two groups of wine tasters. Due to the large volume of data, we

need to integrate and simplify the overall evaluations of wine samples by each group of tasters.

As the influencing factor is only the number of groups, we plan to establish a single-factor ANOVA model and use the ANOVA1 command in Matlab software to process the balanced data to obtain the return value p to evaluate whether there is a significant difference between the two groups of wine tasters.

To determine which group's results are more reliable, we need to assess the fluctuation around the mean through member ratings[3]. Therefore, we plan to establish a variance analysis model and implement it using Matlab programming. By calculating the variances of red and white wine samples' ratings from Group 1 and Group 2, we can compare them. The group with the smaller variance will be deemed more credible.

3.2 Data processing

In the first group of red wine tasting evaluation forms in Appendix 1, Taster No.4 failed to evaluate the hue characteristics of Sample No.20, which led us to exclude this sample's hue data. For the white wine evaluation forms in Appendix 1, Taster No.6 scored Sample No.3's persistence beyond the maximum value, while Taster No.9 rated Sample No.8's persistence above the full score – both instances necessitated data exclusion.

Next, calculate the total scores for each red wine sample by summing up the individual ratings from all tasters in Group 1. The average of these totals yields the overall average score for each red wine sample. Similarly, calculate the average scores for tasters in Group 2. After arranging the red wine samples in numerical order, compile the average score tables for both groups (see Table 1 in Appendix 1). Applying the same method to the white wine scoring data, we obtain the corresponding average score tables for both groups (see Table 2 in Appendix 1).

3.3 Model establishment and solution

3.3.1 One-way ANOVA model

To determine whether there are significant differences in the evaluation results between the two groups of wine tasters, we can establish two single-factor ANOVA models by considering only factor A (group differences) influencing red and white wine quality assessments. For red wine analysis: Factor A has two levels (A1 and A2). Under level Ai, the total Xi follows a normal distribution N(ui, a2), where i=1,2 with unknown parameters a2 and u_i. While u_i values may differ, it is assumed that all xi samples share identical variance. Additionally, 27 independent trials (comprising 27 red wine samples) were conducted at each level Ai, with all other influencing factors except A remaining constant throughout the experiment.

Xij is the jth independent test in group i.

To determine whether the two levels of A have a significant effect on the score is equivalent to the following hypothesis test:

h0:u1=u2:h1 (Where u1 and u2 are not equal)

Since the value of x ij is affected by Ai and random factor &ij, it needs to be decomposed:

Xij=ui+&ij，i=1，2;j=1,2,...,27 (1)

Let i,j ∈ N(0,a ²) and be independent. Let denote the total mean of red wine sample scores, and ai represents the effect of level Ai on the score. Then:

$$u = \frac{1}{n}\sum_{i=1}^{2} n_i u_i , n = \sum_{i=1}^{2} n_i, a_i = u_i - u, i = 1,2$$

(2)

The model can be expressed as (1) and (2)

$$\begin{cases} x_{ij} = u + a_i + \&_{ij} \\ \sum_{i=1}^{2} a_i = 0 \\ \&_{ij} \sim N(0, a^2), i = 1,2; j = 1,2, ..., n_i \end{cases}$$

The original hypothesis is h0: a1=a2=0

When a=0.01, the null hypothesis H_0 is rejected, indicating that the effect of factor A is highly significant; when a=0.01, the null hypothesis H_0 is not rejected, but when a=0.05, the null hypothesis H_0 is rejected, indicating that the effect of factor A is significant; when a=0.05, the null hypothesis H_0 is not rejected, indicating that the effect of factor A is not significant.

In this model, we use the ANOVA1 command of the single-factor analysis of variance in the MATLAB statistical toolbox to solve. The data of this problem is balanced and the processing method is

P=anoval(x)

The return value p represents a probability that accepts null hypothesis (H_0) when p> α. The matrix X contains the first column showing the average scores given by the first group of tasters for each red wine sample, and the second column corresponding to the average scores from the second group. The Matlab program is detailed in Appendix 1, with operational results presented there. Analysis revealed p=0.1159> α = 0.05, indicating no significant difference in scoring between the two groups. Subsequently, we conducted within-group comparisons using one-way ANOVA with the Matlab's anova1 command, yielding p-values of 0.0006 for the first group and 0 for the second group, demonstrating significant differences across all 10 tasters in both groups. White grape data processed similarly yielded p = 0.0226, confirming significant scoring differences between the two groups. Our analysis through one-way ANOVA models concludes: No significant differences exist in red wine scoring between groups, while white wine scoring shows statistically significant differences.

4. Model establishment and solution of problem 2

4.1 Problem analysis

The classification of wine grapes is related to their physicochemical indicators and wine quality [3]. For the physicochemical indicators of wine grapes, given the numerous parameters, we plan to use principal component analysis (PCA) for dimensionality reduction. By calculating the contribution rates and cumulative contribution rates of PCA components, we will eliminate indicators with minimal impact on classification. Subsequently, we rank the scores of each PCA component. Regarding wine quality, since the second group's scoring obtained from the first question proves more reliable, we select the data from this group. The average score of 10 tasters' ratings in the second group is calculated and standardized into a normalized value Y2 red. We then apply min-max normalization to the raw data to obtain standardized total scores. Through calculations, we derive the comprehensive score Y1 red for red grapes and the quality index value Y2 red for red wine. These two values are weighted in a 7:3 ratio to calculate the total score Yred, which is finally classified according to score intervals.

4.2 Model establishment

The process of establishing the principal component analysis model is as follows: Taking red grapes as an example, we have 27 samples, each sample has 30 variables, and the original data is written into a 27×30 data matrix.

$$X = \begin{bmatrix} X_{11} & \cdots & X_{130} \\ \vdots & \ddots & \vdots \\ X_{271} & \cdots & X_{2730} \end{bmatrix}$$

Xij represents the data of the j-th variable for the i-th sample. Step 1: Standardize the matrix using the min-max method. Step 2: Calculate the correlation coefficient matrix. The formula is as follows:

$$r_{ij} = \frac{\sum_{k=1}^{27}(x_{ki} - x_i)(x_{kj} - x_j)}{\sqrt{\sum_{k=1}^{27}(x_{ki} - x_i)x^2 \sum_{k=1}^{27}(x_{kj} - x_j)x^2}}$$

Xkj represents the data of the j-th variable for the i-th sample xij. This correlation coefficient matrix is obtained as follows:

$$R = \begin{bmatrix} r_{11} & \cdots & r_{130} \\ \vdots & \ddots & \vdots \\ r_{301} & \cdots & r_{3030} \end{bmatrix}$$

In this formula, rij (i, j=1,2,...,30) represents the correlation coefficient between the original variables xi and xj. Step 3 involves calculating the eigenvalues and eigenvectors of R. After solving the characteristic equation, the eigenvalues are sorted from largest to smallest. The corresponding eigenvectors are then calculated for each eigenvalue. Step 4 calculates the principal component contribution rates (bi, i=1,2,3,...,10) and their cumulative contribution rates.

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \end{pmatrix}$ Common factor for red grapes;

$A = \begin{bmatrix} a_{11} & \cdots & a_{130} \\ \vdots & \ddots & \vdots \\ a_{i1} & \cdots & a_{i30} \end{bmatrix}$ The red grape factor score coefficient matrix;;

$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{30} \end{pmatrix}$ Indicates 30 indicators for each sample;

Y=AX

In the fifth step, the scores of each principal component were obtained. The comprehensive score Y1 red of each red grape sample was calculated by weighted cumulative variance contribution rate.

$$Y_{1\,红} = b_1y_1 + b_2y_2 + b_3y_3 + ... + b_iy_i$$

Yi is the main role.

$$Y_{红} = 0.7Y_{1\,红} + 0.3Y_{2\,红}$$

4.3 Model solving

The Matlab program for data dimensionless and principal component analysis is shown in Appendix 2. According to the running results, the cumulative contribution rate of the first 10 standardized samples reached 89%, so we take the first 10 as the principal components.

主成分	1	2	3	4	5
贡献率	0. 4507098	0. 291162	0. 216053	0. 165769	0. 133675
主成分	6	7	8	9	10
贡献率	0. 098457	0. 093197	0. 06944	0. 051791	0. 044838

Based on the contribution rates of these ten principal components, we calculated the comprehensive score Y1red for each red grape sample using cumulative variance contribution weighting. The composite scores Yred for red grapes and red wine samples were then determined with a 7:3 weight ratio. As shown in Appendix 2, we ranked the composite scores from highest to lowest and divided them into equal intervals to establish four quality grades for the grapes:

等级	等级分数段	红葡萄酒样品号
1	(1.11，1.30)	1，2，3，8
2	(0.81，1.11)	9，12，14，16，17，22，23
3	(0.51，0.80)	4，5，6，7，11，15，18，19，20，21，24，27
4	(0.20，0.50)	10，13，25，26

等级	等级分数段	白葡萄样本号
1	(0.501, 0.75)	5, 20, 21, 23，24，27, 28
2	(0.251, 0.500)	1, 2, 4, 6, 7, 10, 12, 14, 17, 18, 22, 26
3	(0.001, 0.250)	3, 9, 11, 13, 15, 25
4	(-0.250, 0.00)	8, 16, 19

As can be seen from the table, we divided red and white grapes into four grades. For red grapes, the grape category showed a distribution of more in the middle and less at both ends, with medium quality grapes in the majority. For white grapes, the first three grades of grapes were mostly strategic, and fewer poor-quality grapes.

5. Model establishment and solution of problem 3

5.1 Problem analysis

To analyze the correlation between physicochemical indicators in grape varieties and wine quality, we aim to establish a linear regression equation to examine their relationship[4]. Given the complexity of these indicators, we will utilize the principal components derived from red grapes and red wine data in Question 2 to simplify the model. The proposed equation positions the principal component of red grape physicochemical indicators as the dependent variable and the principal component of red wine physicochemical indicators as the independent variable. Through analysis of regression coefficients, we can determine whether specific compounds in grape varieties are closely associated with corresponding substances in wine products.

5.2 Model establishment

Taking red grapes and red wine as examples, the method for establishing a multiple regression model is as follows: Let the principal factors of physicochemical indicators in red grapes (anthocyanins, total phenols, tannins, and soluble solids) be the dependent variables Yi (i=1,2,3,4), while the principal factors of physicochemical indicators in red wine (anthocyanins, tannins, total phenols, and total flavonoids) are the independent variables Xi (i=1,2,3,4). The equations established are as follows:

$$\begin{cases} Y_1 = L_{11}X_1 + L_{12}X_2 + \ldots + L_{19}X_9 \\ Y_2 = L_{21}X_1 + L_{22}X_2 + \ldots + L_{29}X_9 \\ Y_3 = L_{31}X_1 + L_{32}X_2 + \ldots + L_{39}X_9 \\ Y_4 = L_{41}X_1 + L_{42}X_2 + \ldots + L_{49}X_9 \end{cases}$$

Lij is the regression coefficient for row i and column j.

5.3 Model solving

After solving the model with Matlab, the following four regression equations are obtained respectively

$$\begin{cases} Y_1 = 0.0206 + 0.7715X_1\_0.1368X_2 + 0.1167X_3 + 0.1500X_4 \\ Y_2 = 0.1093 - 0.0888X_1 + 0.1335X_2 + 0.3002X_3 + 0.5672X_4 \\ Y_3 = 0.2004 + 0.2670X_1 + 0.1968X_2 + 0.1666X_3 + 0.3409X_4 \\ Y_4 = 0.2088 + 0.0004X_1 + 1.1575X_2 - 1.4123X_3 + 0.6867X_4 \end{cases}$$

And the correlation coefficient R2 of the above four equations is calculated, which are respectively

R_1^2=0.8729,R_2^2=0.8256,R_3^2=0.8401,R_4^2=0.8565;

As R^2 approaches 1, the regression equation demonstrates better fit. This confirms the validity of the four equations and the established regression model. The correlation between key indicators of red grapes and primary physicochemical properties of red wine can be

analyzed through coefficient comparisons: 1. Overall, key grape indicators show positive correlations with red wine indicators. 2. Anthocyanin levels in red grapes exhibit strong positive correlation with tannin content in red wine, while showing weak association with other indicators. 3. Total phenol content in red grapes shows significant positive correlation with red wine, though minimal interaction with other indicators. 4. Tannin levels in red grapes demonstrate strong positive correlation with total flavonoids in red wine, with limited interaction with other indicators. 5. Soluble solids content in red grapes exhibits strong negative correlation with total phenol content in red wine, showing minimal interaction with other indicators. Applying the same method reveals relationships between white grape and white wine physicochemical properties: The regression equation is：

$$\begin{cases} Y_1 = -0.18732 + 0.121646X_1 + 0.317797X_2 + 0.171833X_3 + 0.29537X_4 \\ Y_2 = 0.353575 + 0.442474X_1 + 0.100466X_2 - 0.13278X_3 + 0.071923X_4 \\ Y_3 = 0.283085 + 0.080963X_1 + 0.289992X_2 + 0.220219X_3 - 0.0507X_4 \\ Y_4 = 0.586443 + 0.409801X_1 - 0.06597X_2 - 0.32662X_3 + 0.041714X_4 \end{cases}$$

The R^2 values were 0.8379,0.8009,0.7447, and 0.8145 respectively, indicating good regression performance with the model being applicable. Comparative analysis of coefficients in the regression equations reveals the following correlations between key indicators of red grapes and primary physicochemical properties of wine: 1. Overall positive correlation exists between key red grape indicators and red wine indicators. 2. The flavonoid content in white grapes shows a strong positive correlation with total phenolic compounds in white wine. 3. Total sugar content in white grapes demonstrates significant positive correlation with tannin levels in white wine. 4. Titratable acid content in white grapes exhibits close positive correlation with both total phenolic compounds and total flavonoids in white wine. 5. Dry matter content in white grapes shows marked positive correlation with tannin levels and color intensity in white wine, while demonstrating a clear negative correlation with total flavonoids.

6. Model establishment and solution of problem 4

6.1 Problem analysis

Since the third question established a linear relationship between physicochemical parameters of wine grapes and wine quality, and considering that wine quality is determined by sommelier scoring with distinctions between physicochemical and aromatic indicators[5], we can reframe the problem as analyzing how these physicochemical parameters influence scoring. To examine their correlation, we conduct a multivariate linear regression analysis using physicochemical parameters as independent variables and physicochemical scores as dependent variables. We retain independent variables with high correlation coefficients while eliminating those with low coefficients to simplify the equation. Subsequently, we perform multiple linear regression on the remaining independent variables to derive new equations. After obtaining regression models, we calculate updated physicochemical scores based on original parameters, then proportionally scale this score M to obtain the overall quality rating. We plan to compare the new rating with the original one through fitting methods. If the fit is poor, it indicates that physicochemical parameters alone cannot adequately evaluate wine quality; conversely, they can be used effectively.

6.2 Model establishment

We took the physical and chemical indexes of wine as the independent variable and the physical and chemical score of wine as the dependent variable to establish the multiple linear regression equation.

$$\begin{cases} Y_1 = C_{11}X_1 + C_{12}X_2 + \ldots + C_jX_j \\ Y_2 = C_{21}X_1 + C_{22}X_2 + \ldots + C_{2j}X_j \\ \ldots \\ Y_i = C_{i1}X_1 + C_{i2}X_2 + \ldots + C_{ij}X_j \end{cases}$$

Cij represents the regression coefficient in the i-th row and j-th column. Xi denotes the physicochemical indicators of the independent variable, while Yj is the dependent variable. We need to retain the independent variables with high correlation coefficients and ignore those with low coefficients. Subsequently, we perform multiple linear regression on the remaining independent variables to obtain a new equation.

$$\begin{cases} Y_1 = C_{11}X_1 + C_{12}X_2 + \ldots + C_{1m}X_m \\ Y_2 = C_{21}X_1 + C_{22}X_2 + \ldots + C_{2m}X_m \\ \ldots \\ Y_i = C_{i1}X_1 + C_{i2}X_2 + \ldots + C_{im}X_m \end{cases}$$

CIM represents the correlation coefficient in the m-th column of the i-th row, where m denotes the number of retained independent variables. After obtaining the regression equation, we calculate the new rationalization score using a principled indicator, which we intend to call the rationalization score M. Subsequently, we proportionally amplify this score M:

$$M' = \frac{M}{0.6}$$

CIM represents the correlation coefficient in the m-th column of the i-th row, where m denotes the number of retained independent variables. After obtaining the regression equation, we calculate the new rationalization score using a principled indicator, which we intend to call the rationalization score M. Subsequently, we proportionally amplify this score M:

$$K = \frac{M' - B}{B}$$

B is the total score of the wine samples obtained in the first question.

6.3 Solution of the model

The multivariate linear regression equation was processed using Matlab, with the program details provided in Appendix 4. After removing independent variables with low correlation coefficients, we re-established the equation and performed linear fitting between the derived equation and data. The fitting program and results are shown in Appendix 4. The linear fitting graph reveals a clear linear trend between the two variables, indicating that physicochemical indicators of grape varieties and wine quality influence wine quality. However, the poor fit suggests that evaluating wine quality solely based on these indicators is inappropriate. We then conducted logarithmic fitting between the derived equation and data, with the program and fitting results presented in Appendix 4. Although both variables showed consistent increases/decreases overall, the fit remained unsatisfactory. Through this process, we conclude that while physicochemical indicators of grape varieties and wine quality affect wine quality, they do not directly determine wine quality itself.

7. Evaluation of the model

7.1Advantages of the model

1. The variance analysis model developed in Problem One transforms credibility comparisons into variance magnitude comparisons, which remains applicable when the number of tasters increases; 2. The principal component analysis model established in Problem Two provides a market-oriented grape grading method suitable for large-scale product classification and widely applied in quality assessment; 3. The multiple linear regression models constructed in Problems Three and Four respectively demonstrate the relationships between physicochemical indicators of wine grapes and wine quality, as well as the correlation between wine quality and physicochemical parameters. These models concretize abstract concepts, allowing clear visualization of grape selection criteria required

to produce wines meeting specific physicochemical standards. This effectively addresses the challenge of selecting optimal grape varieties for winemakers.

7.2 Disadvantages and improvements of the model

1. For Question 1, we can conduct a variance analysis on each taster's scores within Group 2 to identify which tasters demonstrate greater reliability through analyzing variance levels. 2. Regarding Question 4, we can incorporate wine aroma parameters and perform another multiple linear regression analysis. This approach evaluates wine quality by integrating physicochemical properties, aroma metrics, and grape characteristics. The methodology remains consistent with Question 4, with final evaluation based on model fit. A high R^2 value indicates that the combined assessment of grape characteristics, wine physicochemical properties, and aroma parameters effectively evaluates wine quality.

Reference

[1] Xu Yihan. Wine evaluation based on statistical analysis model [J]. Modern Food, 2021,(18):214-216.DOI:10.16736/j.cnki.cn41-1434/ts.2021.18.058.

[2] Sun Pingshuang. Cluster analysis —— Taking wine classification as an example [J]. Computer Products and Circulation, 2020, (10):282.

[3] Jiang Junlin. Wine evaluation [J]. Modern Food, 2019,(19):146-148.DOI:10.16736/j.cnki.cn41-1434/ts.2019.19.048.

[4] Mao Yuanyuan. Wine evaluation based on multiple statistical analysis models [J]. China Brewing, 2018,37(04):159-163.

[5] Cheng Ran, Jue Dengke, Wang Yiwei. Grading and evaluation model for wine quality [J]. China High-tech Zone, 2018, (03):58.