Journal of Contemporary Educational Research

Research Article

# Research on Liquor Storage Based on Statistical Analysis Model

Xuhua Xu

Sichuan University Jinjiang College, Chengdu 620860, Sichuan Province, China

**Abstract:** According to the physical and chemical indexes of different periods, different storage containers and different storage sites, a two overall T test was used to show that there were significant differences in physical and chemical indexes of liquor body in different detection sites. Correlation analysis of wine storage in different storage methods by Spearman correlation coefficient. By using the principal component analysis method, the comprehensive evaluation index system of the quality of the wine body was constructed, and the classification model of the detection location based on the comprehensive evaluation was established by using the index system. The detection sites were classified, and the results showed that the detection sites were divided into four grades.

**Keywords:** Normal; Standardization; Correlation coefficient; Principal component analysis.

***Publication date:*** March, 2021

***Publication online:*** 31 March, 2021

***Corresponding author:*** Xuhua Xu, 1009610896@qq.com

## 1 Introduction

Climate, wine storage conditions, storage location, time and other factors have great influence on the quality of finished liquor[1]. Sichuan Guchuan Liquor Co., Ltd. divided the detection sites into ground, bamboo forest, underground (divided into Longchang, Yixing), unglazed 6 Jin, osmanthus forest, glaze 1 jin, glaze 5 jin, open-air stainless steel and thatched grass. The physical and chemical indexes of the wine body of different detection points were detected after a quarter of brewing. There are many factors affecting the physical and chemical indexes. In the case of wine, the physical and chemical indexes and the aroma components of the wine depend mainly on the grape variety, the ecological conditions of the grape producing area, and the factors of the wine brewing technology and the year[2,3,4]. Multivariate statistical analysis has many important applications in the analysis of wine body change and detection. Based on statistical analysis, Huo Dan Qun put forward a method for distinguishing liquor based on gas chromatography[5]. Qian Chong combined with cluster analysis (CA) and principal component analysis (PCA) to analyze the flavor groups of several flavor liquors[6].

## 2 On the collection and processing of data

### 2.1 Dealing with missing data

For missing data, the mean shift method is applied to deal with the missing data.The mean substitution method is to replace the missing data with the average value of the residual data after removing the abnormal data[7].

$$x^* = \frac{1}{N}\sum_{m=1}^{N} x_m$$

$x^*$ is the default value, and  is the valid sample number of the group.

### 2.2 Normality test of data

The statistical law shows that the normal distribution has an extremely broad practical background. The probability distribution of many random variables in the production and scientific experiments can be described approximately by the normal distribution[9,10]. Therefore, the normality test of the physical and chemical index data of the original wine and the eight generation liquor is helpful to verify the rationality of the analysis results. In statistics,

Shapiro-Wilk test, Kolmogorov-Smirnov test, P-P map and other methods are used to test the normal hypothesis of the data. The first two tests have high requirements for the sample. When the sample is too few, the test results are not sensitive, and the test results are too sensitive when the sample is too many. Therefore, we use data skewness and kurtosis coefficient to detect the normality of data.

If the distribution of a set of data is symmetrical, then the partial state coefficient $SK$=0, if 0.5<$SK$<1 or -1<$SK$<-0.5, can be called the medium partial distribution, the closer the $SK$ is to 0, the lower the skew degree.

$$SK = \frac{n\sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3} \quad (1)$$

Where $\bar{x}$ is the sample mean, $s$ is the standard deviation, and is the sample number.

The kurtosis is usually compared with the standard normal distribution. If a group of data obeys the standard normal distribution, the peak state coefficient $K$=0. it is the peak distribution when $K$>0 and is the flat distribution when $K$<0.

$$K = \frac{n(n+1)\sum (x_i - \bar{x})^4 - 3\left[\sum (x_i - \bar{x})^2\right]^2 (n-1)}{(n-1)(n-2)(n-3)s^4} \quad (2)$$

Substituting the processed data into (1), (2) or using spss20.0 software, the corresponding skewness and kurtosis values can be calculated, The above table shows that 0.5<$SK$<1 or -1<$SK$<-0.5, -1<$K$<-1, except isobutanol and ethyl hexanoate are not too conforming to normal distribution, and the physical and chemical indexes of other components conform to the normal distribution hypothesis.

## 3 Saliency test

The above test shows that the physical and chemical indexes of each component belong to the normal population. In order to further say the quality of the original wine and the eight generation wine and the reliability of various storage methods, it is necessary to examine whether there is a significant difference between the physical and chemical indexes of the original wine and the eight generation of wine. That is the significant test of the data, the commonly used $t$ test, the $F$ test, and the $\chi^2$ test, The use and use conditions of the three methods are different, but the basic principles are the same. According to the situation, the use of $t$ test is more appropriate. As the experimental group is divided into two groups of original wine and eight generation wine, a double general $t$ test can be used, that is, whether the difference between the average number of the sample of the physical and chemical indexes and the overall difference between the two sets of physical and chemical indexes is obvious.

Step1.suggest a hypothesis: $H_0$:$\mu_1$=$\mu_2$, there was no significant difference in the physical and chemical indexes between the original liquor and the eight generation liquor;

H1:$\mu_1 \neq \mu_2$, there are significant differences in physical and chemical indexes between the original wine and the eight generation of wine.

Step2. Selected level of significance $a$=0.05, $n_1$=$n_2$=$n$=17

Step3. Calculating the $t$ value,

where

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}} \qquad S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2}{n(n-1)}}$$

Among them, $\bar{x}_1$-$\bar{x}_2$ was the mean standard error. $\bar{x}_1$ and $\bar{x}_2$ were the average of physicochemical indexes of the two groups respectively, and $n$ was the sample content and the degree of freedom $df$=2($n$-1). Check the critical value: $t_{0.05}$, $t_{0.01}$, compare the absolute value of the calculated t value with the inference.In fact, after the normal analysis of the data, a paired $t$ test, a Matlab programming solution or a spss20.0 software for calculating the $p$ value can solve the above problems.its $p$ value is distributed between the interval [0.029, 0.043] and obviously $P$<0.05. Under the premise of the confidence level of , there is a significant difference between the physical and chemical indexes of the original wine and the eight generation of wine.

## 4 Correlation analysis (spearman correlation coefficient)

Pearson Correlation Coefficient and Spearman Correlation Coefficient can be used to indicate the direction and degree of the change trend between two variables. The value range is -1 to +1, and the 0 indicates that two variables are unrelated, positive value means positive correlation, negative value means negative correlation, the greater the absolute value is, the stronger the correlation is. The Pearson

correlation coefficient has a high requirement for the data. The experimental data is usually assumed to be a pair from the population of normal distribution, which is influenced by the abnormal value, because the detection methods of all kinds of data are not linear, with the interference of the experimental error, the detection error and so on, the correlation of Pearson's correlation may fall. And Spearman level correlation can make up the above defects, it is based on the order of the original data to solve the problem. This paper uses the Spearman correlation coefficient to analyze.

Spearman hierarchy correlation formula

$$r_s = 1 - \frac{6\sum d_i^2}{N(N^2-1)}$$

Among them, $d_i$ is the difference between the grades of physical and chemical indicators. $N$ is the effective sample number of the group. Taking the Osmanthus fragrans as an example, the specific indexes of the original liquor and the eight generation liquor are shown in Table 1(Table 1).

**Table 1.** Spearman grades

| Physicochemic-al index | Score | | ranking | | Grade differenced | $d^2$ |
|---|---|---|---|---|---|---|
| | Osmanthus 1 | Osmanthus 2 | Osmanthus 1 | Osmanthus 2 | | |
| Ethyl hexanoate | 440.3 | 138.4 | 1 | 1 | 0 | 0 |
| ethyl acetate | 293.0 | 40.9 | 2 | 4 | -2 | 4 |
| Ethyl lactate | 218.2 | 87.8 | 3 | 2 | 1 | 1 |
| Ethyl butyrate | 77.3 | 12.1 | 4 | 8 | -4 | 16 |
| Acetal | 49.8 | 15.4 | 5 | 7 | -2 | 4 |
| N-butanol | 47.4 | 7.4 | 6 | 10 | -4 | 16 |
| Propanol | 44.8 | 43.6 | 7 | 3 | 4 | 16 |
| acetaldehyde | 26.5 | 18.8 | 8 | 6 | 2 | 4 |
| Isoamyl alcohol | 24.6 | 19.2 | 9 | 5 | 4 | 16 |
| Ethyl valerate | 21.5 | 5.1 | 10 | 13 | -3 | 9 |
| methanol | 17.8 | 7.8 | 11 | 9 | 2 | 4 |
| SEC butanol | 17.3 | 5.4 | 12 | 12 | 0 | 0 |
| Isobutanol | 12.3 | 7.3 | 13 | 11 | 2 | 4 |
| Hexanol | 10.3 | 3.2 | 14 | 14 | 0 | 0 |

Substituting the specific data in Table 1 into the Spearman rank correlation formula. The calculation shows that 75.8% $\leqslant r_s \leqslant$ 1( $\leqslant$ s $\leqslant$ 196), It can be seen there has a strong correlation between the physical and chemical indexes. Under the condition of confidence (bilateral) , the correlation between them can be considered significant by testing the hypothesis, thus it can be seen that various storage methods are brewing. The main components of liquor are highly correlated, and the data detected by experimental liquor are relatively accurate.

## 5 Principal component analysis

In the storage process of wine body, different storage methods will produce different physical and chemical indexes of liquor body. There are 17 indexes such as ethyl acetate in this paper. Through the above analysis, we can see that there is a certain correlation between the 17 indexes. If the whole index is considered as a whole, it will not only increase the complexity of the analysis problem, but also appear to overlap the information and cause information inaccuracy. In this paper, the principal component analysis (PCA) is used to recombine the numerous physical and chemical indexes with a certain correlation into a new group of independent variables, and can represent the vast majority of the information of the original variable.

### 5.1 Data standardization processing

Suppose the sample observation data matrix is as follows

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

The raw data can be standardized according to the following methods

$$x_{ij}^* = \frac{x_{ij} - \overline{x_j}}{\sqrt{Var(x_j)}} (i = 1,2...,n; j = 1,2,...p) \quad (3)$$

Where $\overline{x_j}=\frac{1}{n}\sum_{i=1}^{n}x_{ij}; Var(x_j)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x_j})^2.$

the standardized data can be obtained by substituting the sample data into (3) formula

### 5.2 Calculation of sample correlation coefficient matrix

The correlation coefficient matrix of data after standardization is shown below.

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{np} \end{bmatrix}$$

where

$$r_{ij}=\mathrm{cov}(x_i,x_j)=\frac{\sum_{k=1}^{k=n}(x_i-\overline{x_i})(x_j-\overline{x_j})}{n-1}, n>1 \qquad (4)$$

Replace the sample data into (4) form to get normalized correlation coefficient

### 5.3 Calculate the eigenvalues and corresponding eigenvectors of the correlation coefficient matrix R.

Through simple mathematical knowledge, we can find the eigenvalues and corresponding eigenvectors of , or use the spss20.0 software to carry out the principal component analysis on the standardized physical and chemical indexes after the standardization, and also find out the corresponding solutions.

We know the eigenvalues $\lambda_1$=15.519, $\lambda_2$=1.136, $P(P>0)$ principal components can be obtained by principal component analysis. As the variance of each principal component is diminishing, the amount of information contained is also diminishing, in actual analysis, the $P$ principal components are generally not taken, but the former K principal components are taken according to the size of the cumulative contribution rate $Q$ of each principal component. where

$$Q=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i}$$

The larger the $Q$, the more information of the original variables included in the principal component, the selection of the number of principal components is mainly determined by the cumulative contribution rate of the principal component. In order to ensure that the comprehensive variable can maintain the vast majority of the information of the original variable, the cumulative contribution rate of the principal component is generally above . From table 7, we can see that the contribution rate of the cumulative variance of the first two principal components is , that is, covering most of the information. This indicates that the first two principal components can represent the initial 17 indicators to analyze the liquor index level of different storage points of wine body, so the first two indicators can be extracted. The principal components are recorded as $F_1$, $F_2$ and $L_1$, $L_2$ are their coefficients, respectively, of which $L_1=\frac{\lambda_1}{\lambda_1+\lambda_2}, L_2=\frac{\lambda_2}{\lambda_1+\lambda_2}$ or the percentage of the respective variances divided by the percentage of the cumulative variance,where

$$L_1=\frac{0.91288}{0.97971}, L_2=\frac{0.06683}{0.97971}$$

So we have

$$F=L_1F_1+L_2F_2=0.9318F_1+0.6821F_2 \qquad (5)$$

### 5.4 Calculation of principal component score

According to the standardized original data, according to each index, the principal component expression can be replaced respectively, and the new data of each index under each principal component can be obtained, that is, the main component score, the specific form is as follows:

$$F=\begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1k} \\ F_{21} & F_{22} & \cdots & F_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ F_{n1} & F_{n2} & \cdots & F_{nk} \end{bmatrix}$$

where

$$F_{ij}=a_{j1}X_{i1}+a_{j2}X_{i2}+\ldots+a_{jp}X_{ip}, i=1,2\ldots.n; j=1,2\ldots k \qquad (6)$$

$a_{jp}$ is the component score coefficient, which can be defined as the square root of the eigenvalue of the principal component loa d matrix by the principal component, and the component score coefficient of each variable can be obtained by calculation.then the function expression of $F_1$, $F_2$ can be obtained.

$F_1=0.062x_1+0.064x_2+0.064x_3+0.002x_4+0.064x_5+0.064x_6+0.064x_7+0.064x_8+0.064x_9+0.063x_{10}+0.058x_{11}+0.064x_{12}+0.063x_{13}+0.064x_{14}+0.064x_{15}+0.064x_{16}+0.063x_{17}$

$F_2=0.142x_1+0.028x_2-0.056x_3+0.877x_4-0.045x_5-0.036x_6+0.019x_7-0.016x_8-0.073x_9+0.139x_{10}+0.156x_{11}+0.059x_{12}-0.009x_{13}-0.020x_{14}-0.098x_{15}-0.095x_{16}+0.107x_{17}$

From a statistical standpoint, date shows the ranking of liquor quality in different storage locations. Compared with the eight generation liquor,

the original wine can be seen from various indicators. Besides the underground Yixing 1 and thatched grass 1, the original liquor body is obviously superior to the eight generation liquor.

## 6 Conclusion

In order to compare the influence of different wine spots on the quality of liquor quality, the same hidden wine spots has original wine and eight generation of wine. The Ranking of different hidden wine spots can be obtained by calculating the weighted average of the score of the original wine and the eight generation of wine. In general, the collection points are divided into four grades, the first grade is underground Longchang and Glaze 1 jin, second grade for full buried, bamboo forest, unglazed 6 Jin, third grade of Osmanthus forest, underground Yixing, fourth grades of glaze 5 jin, thatch.

Taking underground wine as an example, the eight generation of wine has the highest score in the underground Yixing, while the original wine has a third score in the underground Longchang. This shows that the quality of the underground storage is higher, which can be seen from the key indicators, such as the underground Longchang 1. After storage for one quarter, the key index is that the content of ethyl hexanoate is up, only slightly below ground and buried in bamboo forest, the content of ethyl butyrate is only below ground, the content of ethyl lactate is only lower than that of bamboo forest, the content of ethyl acetate is the highest, the content of ethyl valerate is the highest, the content of total ester is the highest, the content of hexanol (fruit aroma) is the highest and isobutanol(special aroma) is the highest content, n-butanol (smell bad) moderate content. From the wine body sense impression, the original wine: the colorless transparent, the grain fragrance, the aldehyde taste, the cellar fragrance is more prominent, the wine body is full and full, the sweet and clean, Long aftertaste, typical style. The eight generation of wine: colorless, transparent, elegant, alcohol and alcohol, soft and refreshing, long aftertaste, outstanding style. Considering the taste and flavor characteristics, the storage quality of underground Longchang 1 is better than that of other storage wine.

## References


[1] Zhang J H. Using environmental factors to optimize liquor quality[J]. Wine Making, 2010, 37 (6):52-54.

[2] Ribereau-Gayon P, Glories Y, Maujean A, et al. Handbook of Enology,The Chemistry of Wine: Stabilization and Treatments[M]. Hoboken: John Wiley & Sons, 2000.

[3] Rodrıíguez-Bencomo J J, Conde J E, Rodrıguez-Delgado M A,et al. Determination of esters in dry and sweet white wines by headspace solid-phase micro-extraction and gas chromatography[J]. Journal of Chromatography A, 2002, 963(1):213-223.

[4] Salinas M R, Garijo J, Pardo F, et al. Color, polyphenol, and aroma compounds in rosé wines after prefermentative maceration and enzymatic treatments[J]. American Journal of Enology and Viticulture, 2003, 54(3):195-202.

[5] Huo D Q,Zhang M M, Hou C J. Based on principal component analysis and discriminant analysis of liquor brand identification method[J]. Agricultural Engineering Journal, 2011, 27 (2):297-301.

[6] Qian C, Liao Y H, Liu M Y. Cluster analysis and principal component analysis of different flavor liquor[J]. Chinese Journal of Food Science, 2017:243-255.

[7] Li Y, Li J M, Jiang Z J. Application of Statistical Analysis in Wine Quality Evaluation[J]. Brewing Technology, 2009, 178 (4):79-82.

[8] Zhao H, Shao S H, Xie D P. Method of Handling Outliers in Analytical Data[J]. Journal of Zhoukou Normal University, 2004, 21(5):70-72.

[9] Jia J P, He X Q, Jin Y J. Statistics (Fifth Edition)[M]. Beijing: Renmin University of China Press, 2012:57-58.

[10] Sheng J,Xie S Q, Pan C Y. Probability Theory and Mathematical Statistics (Third Edition)[M]. Beijing: Higher Education Press, 2001:144-153.