葡萄酒评价

摘 要

葡萄酒在生活中非常常见，对于葡萄酒的好坏，有评酒师进行打分，从而得到对葡萄酒的划分。本题从两组评酒师之间是否存在差异性、根据葡萄酒的理化指标和葡萄酒的质量对葡萄酒的等级划分、酿酒葡萄与葡萄酒的理化指标之间的关系，以及是否能用葡萄酒的理化指标评判葡萄酒，解决以下几个问题：

问题一分析评酒员对葡萄酒评价结果的差异性，根据题目所给数据首先进行处理，首先判断各评酒员对于各组葡萄酒的评分属于正态分布，然后进行 t 检验判断没有显著性差异。最后根据信度分析得到第一组评酒员更可靠。

问题二要求对酿酒葡萄进行等级区分，酿酒葡萄影响着葡萄酒的质量，一瓶葡萄酒的质量，不仅仅只有评酒员单方面评价，也要从许多理化指标考虑，由于所给指标量大，所以首先进行主成分分析，将所给数据进行降维处理减少指标个数。再讨论两者之间的关系，通过建立主成分回归模型，得到各主成分综合得分，按分数高低，并结合质量大小，采用个案排秩对其进行等级划分，最终分成五个等级。

问题三使用逐步线性回归模型，忽略与葡萄酒质量相关性很小的理化指标，建立酿酒葡萄、葡萄酒的理化指标与葡萄酒质量的关系式，求出拟合程度来判断酿酒葡萄、葡萄酒的理化指标对葡萄酒质量的影响程度。得出红葡萄酒理化性质和酿酒红葡萄的各理化指标相关性较强，白葡萄酒理化性质和酿酒红葡萄的各理化指标相关性较弱的结论。

问题四继续沿用问题三中的逐步线性回归模型，由于评酒员是通过外观分析、香气分析和口感分析三个角度对葡萄酒的质量进行评价，因此针对理化指标对葡萄酒质量的影响分析可从该三个角度着手进行探究。建立酿酒葡萄、葡萄酒的理化指标与外观分析、香气分析和口感分析的关系式。得出无论是红葡萄酒和白葡萄酒，它们的质量都与酿酒葡萄和葡萄酒的理化指标存在良好的相关性。因此，我们有理由认为，能够用葡萄和葡萄酒的理化指标来评价葡萄酒的质量

关键词：正态分布 t 检验 差异性分析 信度分析 主成分分析 逐步线性回归

一、问题重述

确定葡萄酒质量时一般是通过聘请一批有资质的评酒员进行品评。每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到其总分，从而确定葡萄酒的质量。酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系，葡萄酒和酿酒葡萄检测的理化指标会在一定程度上反映葡萄酒和葡萄的质量。附件中给出了某一年份一些葡萄酒的评价结果，并分别给出了该年份这些葡萄酒的和酿酒葡萄的成分数据。我们需要建立数学模型并且讨论下列问题:

1、分析附件 1 中两组评酒员的评价结果有无显著性差异，并确定哪一组的评价结果更可信。

2、根据酿酒葡萄的理化指标和葡萄酒的质量对这些酿酒葡萄进行分级。

3、分析酿酒葡萄与葡萄酒的理化指标之间的联系。

4、分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。

二、问题分析

问题一根据附件一所提供的两组评酒师所打出的分数,对于两组评酒员的评分绘制出正态分布直方图，根据 Q-Q 图以及正态性检验发现符合正态分布；然后再根据 t 检验计算出显著性 Sig 的值均大于 0.05，从而得出两组评酒员的评价结果无显著性差异，再利用信度分析通过比较 α 系数，判断两组中哪一组更具有可靠性。

问题二要求对葡萄酒进行等级区分，酿酒葡萄影响着葡萄酒的质量，一瓶葡萄酒的质量，不仅仅只有评酒员单方面评价，也要从许多理化指标考虑。由第一问可以得出存在一组的评价结果更可信，考虑采用可信度高的那一组对葡萄酒的评价结果，以此作为第二问的数据依据，简化数据量。附件二中提供的酿酒葡萄的理化指标，考虑到其数目过多，且部分指标对各自的品质影响小，考虑采用主成分分析法对各项指标进行降维处理，减少指标个数。得到酿酒葡萄的综合指标，最后通过个案排秩加法排名对其进行等级划分。

问题三由于酿酒葡萄理化指标众多,考虑到葡萄酒的理化指标与有些酿酒葡萄的理化指标相关性很小，可以忽略这些理化指标使模型变得简洁。考虑运用逐步回归分析，挑选出对葡萄酒质量有显著影响的自变量，构造最优的回归方程，找出酿酒葡萄与蕾萄酒的理化指标的关系式,进而求得酿酒葡萄与葡萄酒的理化指标之间的拟合程度。拟合程度代表酿酒葡萄与葡萄酒的理化指标间的关系。

问题四继续沿用逐步线性回归模型，忽略与葡萄酒质量相关性很小的理化指标，建立酿酒葡萄、葡萄酒的理化指标与葡萄酒质量的关系式。附件一中评酒员通过外观品质、香气品质和口感品质三种类型的指标对葡萄酒的质量进行评价，因此针对理化指标对葡萄酒质量的影响分析，可从该三种类型着手进行探究。建立酿酒葡萄、葡萄酒的理化指标与外观品质、香气品质和口感品质的关系式，求出拟合程度来判断酿酒葡萄、葡萄酒的理化指标对葡萄酒质量的影响程度。要论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量，考虑利用拟合程度来判别，若拟合程度较大，葡萄酒与酿酒葡萄和葡萄酒的联系越紧密，则可以用来论证葡萄酒的质量。若不能，则要进一步考虑芳香物质对葡萄酒质量的影响。

三、模型假设

（1）假设数据来源真实有效
（2）假设酿酒工艺条件相同，无其他人为因素影响
（3）假设酿酒工艺和贮存条件等对葡萄酒质量及理化指标无影响
（4）假设对理化指标的检测误差在可接受范围之内

四、符号说明

x_{ij}	第i个评价对象的第j个指标的取值	u_m	特征向量
r_{ij}	第i各指标于第j个指标的相关系数	x_i	葡萄样品的理化指标
$\bar{x}_j$	样本均值	y_i	葡萄酒的理化指标
s_j^2	样本方差	z_i	葡萄酒中的芳香物质
b_j	第j个主成分的信息贡献率	R^2	判定系数
y_m	第m主要成分	u_m	特征向量

五、模型建立与求解

5.1 两组评酒员数据的差异性分析对于数据的处理

数据的预处理：使用 excel 将每组葡萄酒评酒员的评分取平均值，得到红葡萄酒 27 组数据，白葡萄酒 28 组数据。然后利用 SPSS 对于两组样本进行 t 检验和信度分析。

5.1.1 正态分布检验

根据附件一的两组红白葡萄酒数据，通过 SPSS[1]绘制出两组红葡萄酒与白葡萄酒评分的正态分布直方图以及评分的正态 Q-Q 图：

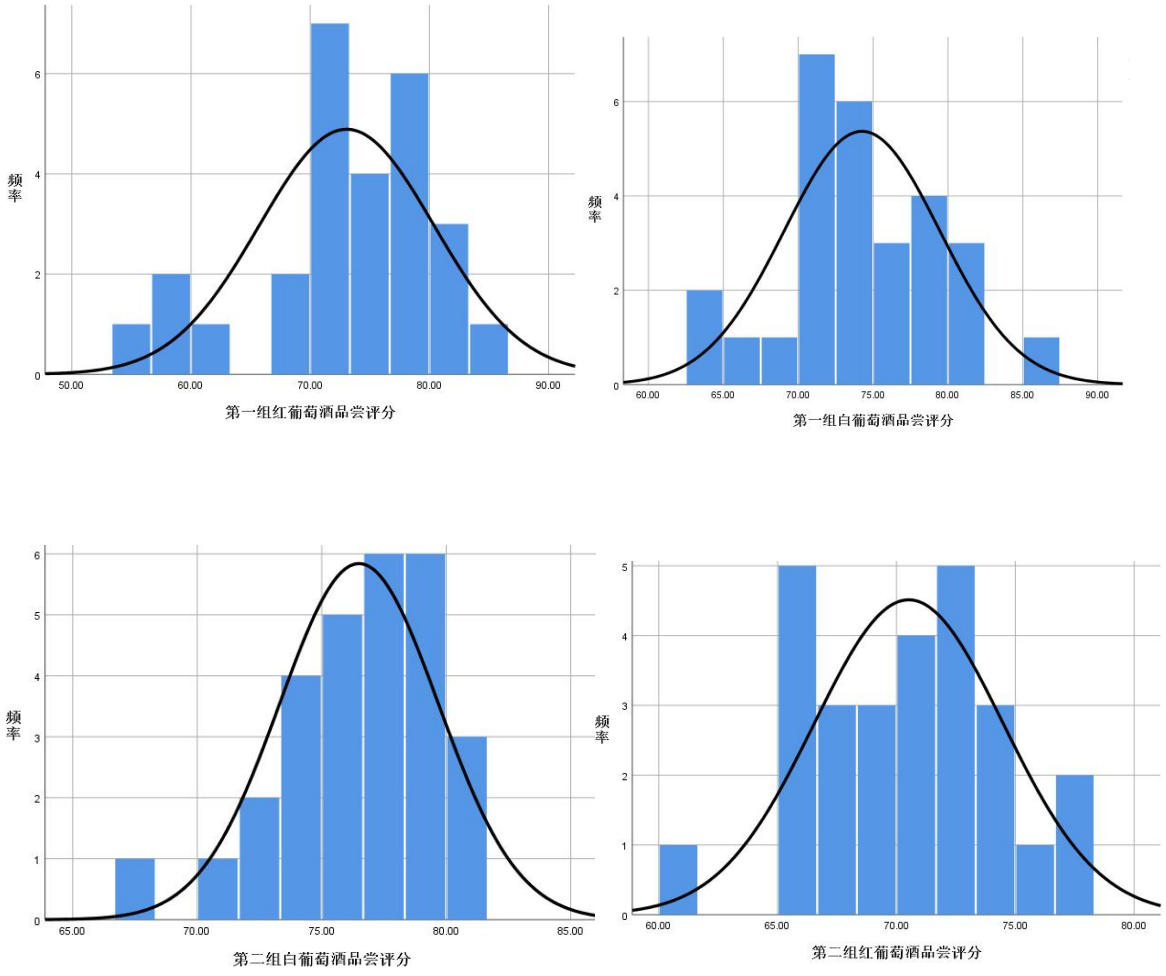


图 1 各组葡萄酒品尝评分分布直方图

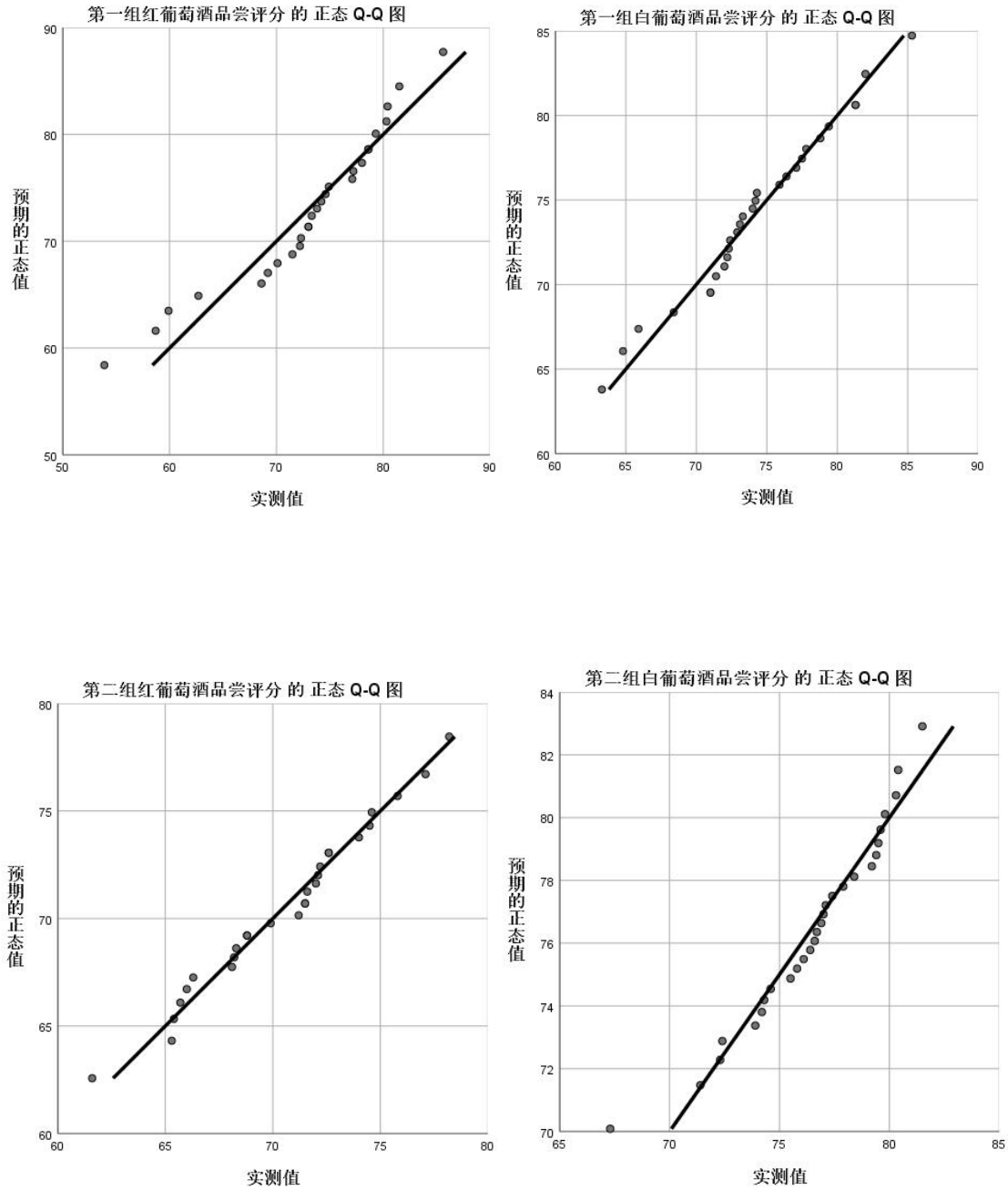


图 2 各组葡萄酒品尝评分的正态 Q-Q 图

根据所绘制出的图像，得出两组评酒员对葡萄酒评分符合正态分布。

图 1 通过两组评酒员的品尝评分分布直方图，以及图 2 的品尝评分的正态 Q-Q 图得到每一组评酒员的得分近似在一条直线附近说明葡萄酒的评分可能符合正态分布。

为了准确性，对两组评酒员的评分使用 SPSS 进行正态性检验，如下表：

表 1 正态性检验

	柯尔莫戈洛夫-斯米诺夫(V)[a]			夏皮洛-威尔克		
	统计	自由度	显著性	统计	自由度	显著性
第一组红葡萄酒品尝评分	0.157	27	0.085	0.925	27	0.052
第二组红葡萄酒品尝评分	0.124	27	0.200*	0.980	27	0.868
第一组白葡萄酒品尝评分	0.122	27	0.200*	0.980	27	0.853
第二组白葡萄酒品尝评分	0.097	27	0.200*	0.950	27	0.218

*. 这是真显著性的下限。

a. 里利氏显著性修正

当样本量 3 ≤n ≤5000 时，结果以 Shapiro - Wilk (W 检验) 为准.在常态检验结果中，由于样本数量为 27 和 28，结果以 Shapiro - Wilk (W 检验) 为准，在置信区间 95%（α=0.05）的情况下，P 均>0.05，不拒绝原假设，可认为变量服从正态分布。

5.1.2 t 检验（显著性检验）

由于两组评酒员对葡萄酒评分的样本都比较少，因此采用 t 检验来判断两组评酒员的评价结果有无显著性差异。

t 检验是为了比较数据样本之间是否具有显著性的差异，它是用 T 分布理论来推断差异发生的概率,从而判定两个平均数的差异是否显著,主要是用于样本含量小(小于 30 个)，总体标准差 δ 未知的正态分布。

用 SPSS 得出 t 检验的结果如下：

表 2 一二组白葡萄酒评分独立样本检验

		平均值等同性 t 检验						
		t	自由度	Sig.（双尾）	平均值差值	标准误差差值	差值 95% 置信区间	
							下限	上限
白葡萄酒	假定等方差	-1.939	54	0.058	-2.23571	1.15283	-4.54700	0.07557
	不假定等方差	-1.939	44.773	0.059	-2.23571	1.15283	-4.55796	0.08653

表 3 一二组红葡萄酒评分独立样本检验

		平均值等同性 t 检验						
		t	自由度	Sig.（双尾）	平均值差值	标准误差差值	差值 95% 置信区间	
							下限	上限
红葡萄酒	假定等方差	1.581	52	0.120	2.54074	1.60714	-0.68	5.765
	不假定等方差	1.581	40.05	0.122	2.54074	1.60714	-0.70	5.788

通过观察表 2，表 3 的 t 检验中两组红葡萄酒显著性 Sig 的值均大于 0.05，两组白葡萄酒显著性 Sig 的值均大于 0.05,得到两组数据没有显著性差异。

5.1.3 信度分析（可靠性分析）

Cronbach α系数是目前最常用的信度系数，它表明量表中每一组得分间的一致性，Cronbach α系数公式为：

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum_{i=1}^{k} S_i^2}{S_x^2})$$

式中k为测量的组数；S_i为第i组得分数的方差；S_x为测验总分的方差。

α系数，不是对测量的单独划分进行计算，而是对测量所有可能的花费的均值和折半系数（self-half reliability，将受测题目分成两半，然后估计参与者对这两部分反应的一致性）用于评估一个测量单独执行的可靠性。α 系数是衡量量表或检验信度的一种方法。一般来说，α>0.7 可以接受，0.7<α<0.98 属于高信度，α<0.35 属于低信度，予以拒绝。取值范围是 0<α<1。α系数一般用于测量量表的内部一致性，至少需要两个项目（变量）的得分，所有的项目应当测量同样的特点或特征。

以下是用 SPSS 对两组十名评酒员的α一致性检验：

表 4 可靠性统计

第一组红葡萄酒		**第二组红葡萄酒**		**第一组白葡萄酒**		**第二组白葡萄酒**	
克隆巴赫 Alpha	项数	克隆巴赫 Alpha	项数	克隆巴赫 Alpha	项数	克隆巴赫 Alpha	项数
0.916	10	0.861	10	0.773	10	0.697	10

通过表 4 可以看出在两种葡萄酒的评分中，第一组评酒员的α系数均大于第二组评酒员的α系数，且都大于 0.7，属于高信度，因此我们有充分的理由认为第一组评酒员打分的结果更可信。

5.2 根据酿酒葡萄的理化指标和葡萄酒的质量对酿酒葡萄进行分级

对葡萄酒进行等级区分，由第一问可以得出存在一组的评价结果更可信，考虑采用可信度高的那一组对葡萄酒的评价结果，以此作为第二问的数据依据，简化数据量。附件二中提供的酿酒葡萄的理化指标，考虑到其数目过多，且部分指标对各自的品质影响小，考虑采用主成分分析法对各项指标进行降维处理，减少指标个数。得到酿酒葡萄的综合指标，最后通过个案排秩加法排名对其进行等级划分。

5.2.1 主成分分析法[2]进行数据降维

采用主成分分析法首先对酿酒葡萄进行数据降维，并建立酿酒葡萄的理化指标和葡萄酒质量与酿酒葡萄质量的线性回归方程，求出评价得分，从而对酿酒葡萄进行分级。

主成分分析主要原理是采用降维的方式把原来多个变量转变为几个变量。

（1）对原始数据标准化处理：为消除不同变量的量纲影响，首先需要对变量进行标准化处理，进行主成分分析的指标变量有 m 个：$x_1, x_2, \cdots x_m$，共有 n 个评价对象，第 i 个评价对象的第 j 个指标的取值为 x_{ij}。x_{ij} 标准化后的指标为 x_{ij}，$x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, (i = 1,2,\cdots,m)$，其中样本均值 $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$，样本方差为 $s_j^2 = \frac{1}{n-1}\sum_{n-1}^{n}(x_{ij} - \bar{x}_j)^2, \left(j = 1,2,\cdots m\right)$

（2）计算相关系数矩阵 R

相关系数代表了不同指标之间的相关程度，绝对值越大代表相关性越高。相关性较高的变量之间存在信息上的重叠，信息重叠在很大程度上会影响评价结果的客观性，因此相关性矩阵可以证明进行主成分分析的必要性。

相关系数矩阵 $R = (r_{ij})_{m \times m}$，即

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1j} \\ r_{21} & r_{22} & \cdots & r_{2j} \\ \vdots & \vdots & & \vdots \\ r_{i1} & r_{i2} & \cdots & r_{ij} \end{bmatrix}$$

，$r_{ij} = \frac{\sum_{k=1}^{n} x_{ki} \cdot x_{kj}}{n-1}, (i, j = 1,2,\cdots m)$ 式中 $r_{ii} = 1, r_{ij} = r_{ji}, r_{ij}$ 是第 i 各指标于第 j 个指标的相关系数。

进行降维处理，即用较少的几个综合指标代替原来较多的变量指标，而且使这些

较少的综合指标既能尽量多地反映原来较多变量指标所反映的信息,同时它们之间又是彼此独立的。

（3）计算相关系数矩阵 R 的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda m \geq 0$，与对应的特征向量 $u_1, u_2, \cdots u_m$，其中 $u_j = (u_{1j}, u_{2j}, \cdots, u_{nj})^T$，由特征向量组成 m 个新的指标变量

$$\begin{cases} y_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{n1}x_n \\ y_1 = u_{12}x_1 + u_{22}x_2 + \cdots + u_{n2}x_n \\ \cdots\cdots \\ y_1 = u_{1m}x_1 + u_{2m}x_2 + \cdots + u_{nm}x_n \end{cases}$$

上式中 y_1 是第一主要成分，y_2 是第二主要成分，y_m 是第 m 主要成分。

（4）选择 $p(p \leq m)$ 个主成分，计算综合评价值

A. 计算特征值 $\lambda_j (j = 1,2,\cdots,m)$ 的信息贡献率和累积贡献率。

主成分 y_j 的信息贡献率为：$b_j = \dfrac{\lambda_j}{\sum_{k=1}^{m} \lambda_k} (j = 1,2,\cdots,m)$

主成分 $y_1, y_2, \cdots, y_p$ 的累积贡献率为：$\alpha_p = \dfrac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}$

当 α_p 接近于 1（α_p=0.85，0.90，0.95）时，则选择前 P 个指标变量 $y_1, y_2, \cdots, y_p$ 作为 P 个主成分，代替原来 m 个指标变量，从而对 p 个主成分进行综合分析

B. 计算综合得分：$Z = \sum_{j=1}^{p} b_j y_j$

其中 b_j 为第 j 个主成分的信息贡献率，根据综合得分进行评价。

我们使用 matlab[3]对原始数据进行主成分分析得到如下结果：

在主成分的选取上，对应的特征值大小是一个重要衡量因素，普遍的做法是保存特征值要大于 1 的主成分，舍弃特征值小于 1 的主成分，因此最终的主成分个数会小于指标个数 n。也可以根据贡献度大小，累计贡献度达到某个程度，不同标准有 70%以上，85%以上或其他。这里选取所有特征值大于 1 的主成分，红葡萄一共有 11 个主成分,白葡萄一共有 12 个主成分。

表 5 红葡萄各主成分在原始数据上的特征向量

原始数据	x_1	x_2	x_3	……	x_{41}	x_{42}
第 1 主成分	0.117	0.222	-0.012		0.022	-0.098
第 2 主成分	0.230	-0.143	-0.174		0.038	0.014
第 3 主成分	0.050	0.030	0.046		-0.004	0.367
第 4 主成分	-0.132	-0.132	-0.009		0.011	-0.030
第 5 主成分	0.237	0.021	0.079		-0.084	-0.044
第 6 主成分	0.043	-0.254	0.216		0.102	0.109
第 7 主成分	-0.051	0.111	-0.097		0.362	0.140
第 8 主成分	-0.146	-0.141	-0.205		-0.379	0.072
第 9 主成分	-0.136	0.018	0.387		0.226	0.016
第 10 主成分	0.176	0.004	0.042		0.100	0.009
第 11 主成分	0.186	-0.145	0.443		-0.361	-0.100

表 6 白葡萄各主成分在原始数据上的特征向量

原始数据	x_1	x_2	x_3	……	x_{41}	x_{42}
第 1 主成分	0.165	0.042	-0.124		-0.019	0.182
第 2 主成分	0.077	0.235	-0.089		0.289	-0.148
第 3 主成分	0.021	-0.123	-0.257		-0.11	-0.238
第 4 主成分	0.035	-0.068	0.066		-0.065	0.218
第 5 主成分	0.071	0.166	-0.025		-0.125	0.141
第 6 主成分	-0.332	0.165	0.118		-0.102	0.003
第 7 主成分	0.048	-0.072	-0.147		-0.016	0.022
第 8 主成分	-0.155	-0.325	0.227		0.163	-0.008
第 9 主成分	0.214	-0.140	0.191		0.253	0.056
第 10 主成分	0.170	-0.039	-0.103		-0.289	-0.167
第 11 主成分	0.249	0.162	-0.030		-0.123	0.082

将特征贡献率作为系数，对应的指标作为自变量，可以得出每一个主成分的计算表达式。将标准化数据 Xi 代入表达式，就可以得到对应的主成分值。

红葡萄各个主成分的得分：

$$F_1 = 0.11767x_1 + 0.22247x_2 \ldots\ldots - 0.098502x_{42}$$
$$F_2 = 0.2301x_1 - 0.1437x_2 \ldots\ldots + 0.014381x_{42}$$
$$F_3 = 0.050632x_1 + 0.30433x_2 \ldots\ldots + 0.36772x_{42}$$
$$F_4 = -0.13297x_1 - 0.13207x_2 \ldots\ldots - 0.030006x_{42}$$
$$F_5 = 0.23731x_1 + 0.021458x_2 \ldots\ldots - 0.044583x_{42}$$

$$F_6 = 0.043919x_1 - 0.25415x_2 \ldots\ldots + 0.10973x_{42}$$
$$F_7 = -0.051667x_1 + 0.11129x_2 \ldots\ldots + 0.14087x_{42}$$
$$F_8 = -0.14627x_1 - 0.14136x_2 \ldots\ldots + 0.072333x_{42}$$
$$F_9 = -0.13689x_1 + 0.018001x_2 \ldots\ldots + 0.016694x_{42}$$
$$F_{10} = 0.17684x_1 + 0.004213x_2 \ldots\ldots + 0.0093286x_{42}$$
$$F_{11} = 0.18692x_1 - 0.14574x_2 \ldots\ldots - 0.10029x_{42}$$

将特征值作为系数，对应的主成分作为自变量，可以确定综合评价值的表达式，

红葡萄评价得分：

$$P = 8.9202F_1 + 6.5764F_2 + 5.7148F_3 + 3.7323F_4 + 2.8154F_5 + 2.2836F_6 + 1.8948F_7 + 1.7523F_8 + 1.4504F_9 + 1.1709F_{10} + 1.0514F_{11}$$

同理可得：

白葡萄各个主成分的得分：

$$F_1 = 0.16542x_1 + 0.041977x_2 \ldots\ldots + 0.18199x_{42}$$
$$F_2 = 0.077005x_1 + 0.23511x_2 \ldots\ldots - 0.14758x_{42}$$
$$F_3 = 0.021346x_1 - 0.1229x_2 \ldots\ldots - 0.23751x_{42}$$
$$F_4 = 0.035167x_1 - 0.068402x_2 \ldots\ldots + 0.21829x_{42}$$
$$F_5 = -0.071257x_1 - 0.16642x_2 \ldots\ldots - 0.14053x_{42}$$
$$F_6 = -0.33242x_1 + 0.1645x_2 \ldots\ldots + 0.0026946x_{42}$$
$$F_7 = 0.048405x_1 - 0.071726x_2 \ldots\ldots + 0.022399x_{42}$$
$$F_8 = -0.15463x_1 - 0.32468x_2 \ldots\ldots - 0.008373x_{42}$$
$$F_9 = 0.21411x_1 - 0.13999x_2 \ldots\ldots + 0.056344x_{42}$$
$$F_{10} = 0.16983x_1 - 0.039429x_2 \ldots\ldots - 0.16688x_{42}$$
$$F_{12} = 0.19685x_1 + 0.17012x_2 \ldots\ldots + 0.025571x_{42}$$

白葡萄综合评价得分：

$$P = 9.4355F_1 + 6.3156F_2 + 4.3542F_3 + 3.0395F_4 + 2.9242F_5 + 2.4305F_6 + 2.071F_7 + 1.7047F_8 + 1.4428F_9 + 1.288F_{10} + 1.1869F_{11} + 1.0201F_{12}$$

5.2.2 每个样本的综合评价值

带入各主成分代表的理化性质值，得到每个样本的综合评价值，如下表：

表 7 酿酒葡萄的 综合评价值

红葡萄	评价得分	白葡萄	评价得分
葡萄样品 1	0.73	葡萄样品 1	-1.58
葡萄样品 2	0.45	葡萄样品 2	0.05
葡萄样品 3	2.32	葡萄样品 3	0.36
葡萄样品 4	-0.48	葡萄样品 4	-0.68
葡萄样品 5	-0.74	葡萄样品 5	0.90
葡萄样品 6	-0.18	葡萄样品 6	0.36
葡萄样品 7	-0.06	葡萄样品 7	0.94
葡萄样品 8	1.79	葡萄样品 8	-0.73
葡萄样品 9	0.52	葡萄样品 9	-0.43
葡萄样品 10	-0.67	葡萄样品 10	0.04
葡萄样品 11	1.09	葡萄样品 11	-1.36
葡萄样品 12	-0.16	葡萄样品 12	0.42
葡萄样品 13	0.08	葡萄样品 13	-1.56
葡萄样品 14	0.86	葡萄样品 14	0.31
葡萄样品 15	-0.54	葡萄样品 15	0.95
葡萄样品 16	-0.49	葡萄样品 16	-0.27
葡萄样品 17	-0.07	葡萄样品 17	-1.27
葡萄样品 18	-0.29	葡萄样品 18	-0.26
葡萄样品 19	-0.26	葡萄样品 19	-0.83
葡萄样品 20	-0.90	葡萄样品 20	0.51
葡萄样品 21	0.62	葡萄样品 21	-0.35
葡萄样品 22	-0.16	葡萄样品 22	-0.24
葡萄样品 23	0.29	葡萄样品 23	1.06
葡萄样品 24	-0.43	葡萄样品 24	0.10
葡萄样品 25	-1.37	葡萄样品 25	0.61
葡萄样品 26	-1.31	葡萄样品 26	0.56
葡萄样品 27	-0.64	葡萄样品 27	1.78
		葡萄样品 28	0.62

5.2.3 个案排秩加法排名得到分级结果

我们在对对象进行评价排名时，需要从多个不同的角度进行客观评价，而评价的角度越多，进行综合排名的难度就越高，不同的评价方法得出来的结论也会有所不同，因此我们采用个案排秩加法排名。

下面我们使用 SPSS，针对葡萄酒的质量和综合评价值，对它们进行一个综合排名。得到下表：

表 8 综合得分

红葡萄	综合得分(秩)	白葡萄	综合得分(秩)
葡萄样品 1	29.00	葡萄样品 1	28.00
葡萄样品 2	13.00	葡萄样品 2	30.00
葡萄样品 3	4.00	葡萄样品 3	45.50
葡萄样品 4	37.00	葡萄样品 4	31.00
葡萄样品 5	37.00	葡萄样品 5	29.50
葡萄样品 6	34.00	葡萄样品 6	21.50
葡萄样品 7	33.00	葡萄样品 7	46.00
葡萄样品 8	22.00	葡萄样品 8	13.00
葡萄样品 9	9.00	葡萄样品 9	20.00
葡萄样品 10	40.00	葡萄样品 10	30.00
葡萄样品 11	26.00	葡萄样品 11	13.00
葡萄样品 12	40.50	葡萄样品 12	20.00
葡萄样品 13	26.00	葡萄样品 13	5.00
葡萄样品 14	16.00	葡萄样品 14	24.00
葡萄样品 15	47.00	葡萄样品 15	37.00
葡萄样品 16	34.00	葡萄样品 16	25.00
葡萄样品 17	18.00	葡萄样品 17	27.00
葡萄样品 18	42.00	葡萄样品 18	24.00
葡萄样品 19	23.00	葡萄样品 19	14.00
葡萄样品 20	29.00	葡萄样品 20	42.00
葡萄样品 21	15.00	葡萄样品 21	28.00
葡萄样品 22	23.50	葡萄样品 22	17.50
葡萄样品 23	10.00	葡萄样品 23	45.00
葡萄样品 24	26.00	葡萄样品 24	29.00
葡萄样品 25	48.00	葡萄样品 25	42.00
葡萄样品 26	37.00	葡萄样品 26	46.50
葡萄样品 27	37.00	葡萄样品 27	30.00
		葡萄样品 28	48.50

在“个案排秩”的对话框里勾选将秩 1 指定给最大值。如 100 分，那么它的秩就是 1，那么 99 就是 2.000，因此秩值越小排名越靠前。对酿酒葡萄进行以下分级：

表 9 酿酒红葡萄的分级表

等级	编号
A(0-10)	3 9 23
B(11-20)	2 21 14 17
C(21-30)	8 19 22 11 13 24 1 20
D(31-40)	7 6 16 4 5 26 27 10
E(41-50)	12 18 15 25

根据以上模型得出，酿酒红葡萄等级 A 的编号为 3.9,23,等级 E 的编号为 12,18,15,25

酿酒白葡萄的分级如表所示：

表 10 酿酒白葡萄的分级表

等级	编号
A(0-10)	13
B(11-20)	8 11 19 22 9 12
C(21-30)	6 14 18 16 17 1 21 24 5 2 10 27
D(31-40)	4 15
E(41-50)	20 25 23 3 7 26 28

根据以上模型得出，酿酒白葡萄等级 A 的编号为 13,等级 E 的编号为 20,25,23,3,7,26,28

红葡萄与白葡萄相比，红葡萄等级 A 的葡萄样品较多，存在较多的等级 D 的葡萄样品；白葡萄等级 B 和等级 C 的葡萄样品居多。

5.3 讨论酿酒葡萄与葡萄酒的理化指标之间的联系

由于酿酒葡萄理化指标众多,考虑到葡萄酒的理化指标与有些酿酒葡萄的理化指标相关性很小，可以忽略这些理化指标使模型变得简洁。考虑运用逐步回归分析，挑选出对葡萄酒质量有显著影响的自变量，构造最优的回归方程，找出酿酒葡萄与蕾萄酒的理化指标的关系式，进而求得酿酒葡萄与葡萄酒的理化指标之间的拟合程度。拟合程度代表酿酒葡萄与葡萄酒的理化指标间的关系。

5.3.1 逐步线性回归模型的建立

本问采用逐步线性回归方法来分析酿酒葡萄与葡萄酒的理化指标之间的联系。将葡萄酒的理化指标作为 y，酿酒葡萄的理化指标作为 x 做线性回归，发现葡萄酒的理化指标与有些酿酒葡萄的理化指标相关性很小,可以忽略这些理化指标使模型变得简洁，故使用逐步线性回归的方法求出酿酒葡萄与葡萄酒之间主要理化指标的联系。

(1) 对 p 个回归自变量[5] $X_1, X_2, \cdots, X_p$ 分别同因变量 Y 建立一元回归模型

$$Y = \beta_0 + \beta_i X_i + \varepsilon, i = 1,2,\cdots,p$$

计算变量 X，以及相应的回归系数的 F 检验统计量的值，记为 $F_1^{(1)}, \cdots, F_P^{(1)}$,取其中的最大值 F_{i1}，即 $F_{i1}^{(1)} = \max\left\{F_1^{(1)}, \cdots, F_P^{(1)}\right\}$。

对给定的显著性水平α，记相应的临界值为 $F^{(1)}$， $Fi_1^{(1)} \geq F^1$，则将 X_{i1} 引入回归模型，记 I_1 为选入变量指标集合。

(2)建立因变量 Y 与自变量子集$\{X_{i1},X_1\},\cdots,\{X_{i1},X_{i1-1}\},\{X_{i1},X_{i1+1}\},\cdots,\{X_{i1},X_p\}$的二元回归模型，共有$P-1$个。

计算变量的回归系数 F 检验的统计量值，记为$F_{K^{(2)}}(k=I_1)$，选其中最大者记为$F^{(2)}{}_{i2}$，对应自变量标记为i_2，即$F^{(2)}{}_{i2}=\{F_1{}^{(2)},\cdots,F^{(2)}{}_{i1-1},F^{(2)}{}_{i1+1},\cdots,F^{(2)}{}_p\}$。

对给定的显著性水平α，记相应的临界值为F^1，$F^{(2)}{}_{i1}\geq F^2$，则将变量X_{i2}，引入回归模型。否则，终止变量引入过程。

(2) 考虑因变量对因变量子集$\{X_{i1}，X_{i2}，X_p\}$的回归，重复(2)。

5.3.2 模型方法求解

使用 SPSS 进行逐步回归[4]分析，求解结果如下：

其中结果中的判定系数 R^2，也称为拟合优度或决定系数，即相关系数 R 的平方，用于表示拟合得到的模型能解释因变量变化的百分比，R^2 越接近 1，表示回归模型拟合效果越好。如果R^2= 0.666，说明模型拟合效果一般，但也可以接受。

红葡萄酒的理化指标与酿酒红葡萄各理化指标的关系：

花色苷：$y=2.058x_4+78.345x_{15}+7.575$

$R^2=0.942$

单　宁：$y=0.154x_{11}+4.059x_{17}+0.205x_{23}+0.051x_{28}-8.024$

$R^2=0.894$

总　酚：$y=0.320x_{11}+0.796x_{14}+1.156$

$R^2=0.859$

酒总黄酮：$y=0.398x_{11}-0.950$

$R^2=0.780$

白藜芦醇：$y=0.001x_1+6.326x_{18}+0.005x_{33}-0.658$

$R^2=0.663$

DPPH：$y=2.015E-5x_1+0.017x_{11}+0.043x_{17}-0.026x_{35}+0.014$

$R^2=0.874$

L*(D65)：$y=-0.135x_4-1.092x_5-0.797x_{23}+3.386x_{42}+59.390$

$R^2=0.883$

a*(D65)：$y=0.094x_4-4.818x_{39}+69.167$

$R^2=0.640$

b*(D65)：$y=-0.041x_4+1.169x_5-0.688x_6+0.448x_8+7.548x_{18}+0.256x_{27}-0.668x_{32}-2.123x_{38}-0.117x_{41}$

$R^2=0.939$

白葡萄酒的理化指标与酿酒白葡萄各理化指标的关系：

单　宁：$y = 0.170x_{12} + 0.792x_{23} + 4.062x_{37} + 0.318$

$R^2 = 0.758$

总　酚：$y = 0.000141x_1 + 0.158x_{13} + 0.009x_{24} - 1.201$

$R^2 = 0.647$

酒总黄酮：$y = -0.236x_5 + 0.333x_6 + 3.652x_{23} + 1.732$

$R^2 = 0.790$

白藜芦醇：无

DPPH：$y = 0.042x_{23} + 0.050$

$R^2 = 0.225$

L*(D65)：$y = -0.091x_{20} - 0.009x_{24} + 103.670$

$R^2 = 0.423$

a*(D65)：$y = 0.002x_2 - 0.007x_{26} - 0.738$

$R^2 = 0.342$

b*(D65)：$y = -0.359x_4 + 0.141x_5 - 0.195x_6 + 0.493x_{15} + 0.934x_{23} + 0.055x_{26} - 2.914$

$R^2 = 0.781$

我们通过比较葡萄酒的理化指标与酿酒葡萄各理化指标的拟合优度 R^2，可以得到：

红葡萄酒的花色苷、单宁、总酚、酒总黄酮、DPPH、L*(D65)、b*(D65)与酿酒红葡萄的各理化指标有很强的相关性，红葡萄酒的白藜芦醇、a*(D65)与酿酒红葡萄的各理化指标相关性一般，但可以接受。

白葡萄酒的单宁、酒总黄酮、b*(D65)与酿酒白葡萄的各理化指标有良好的相关性；白葡萄酒的总酚与酿酒白葡萄的各理化指标相关性一般，但可以接受；白葡萄酒的 L*(D65)、a*(D65)与酿酒白葡萄的各理化指标相关性极差，白藜芦醇与酿酒白葡萄的各理化指标无相关性。

总体来看，红葡萄酒理化性质和酿酒红葡萄的各理化指标相关性较强，白葡萄酒理化性质和酿酒红葡萄的各理化指标相关性较弱。

5.4 分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响

问题四继续沿用问题三中的逐步线性回归模型，由于评酒员是通过外观分析、香气分析和口感分析三个角度对葡萄酒的质量进行评价，因此针对理化指标对葡萄酒质量的影响分析可从该三个角度着手进行探究。建立酿酒葡萄、葡萄酒的理化指标与外观分析、香气分析和口感分析的关系式。

5.4.1 逐步线性回归模型的建立

因为要分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，因此，在问题三分析酿酒葡萄与葡萄酒的理化指标间联系的基础上，沿用第三问所建立的多元线性回归模型，并进行相应的改进。

考虑到酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的共同影响，记酿酒葡萄的理

化指标数值为 x，葡萄酒的理化指标数值为 y，葡萄酒的质量因素为

S_i（分别为外观、香气、口感），涉及个自变量的多元线性回归模型可表示为

$$\begin{cases} S_i = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p + \beta_0 + \beta_1 y_1 + \cdots + \beta_p y_p + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases}$$

式中 $\alpha_0, \alpha_1, \cdots, \alpha_p, \beta_0, \beta_1, \cdots, \beta_p, \sigma^2$ 都是与 $x_1, x_2, \cdots, x_p, y_0, y_1, \cdots, y_p$ 无关的未知参数，其中 $\alpha_0, \alpha_1, \cdots, \alpha_p, \beta_0, \beta_1, \cdots, \beta_p$ 为回归系数。

5.4.2 模型方法求解

按照上文建立的逐步线性回归模型，对酿酒葡萄的理化指标、葡萄酒的理化指标和葡萄酒质量进行逐步线性回归分析，利用 spss 求解出酿酒葡萄的理化指标和葡萄酒的理化指标葡萄酒质量之间的关系，即酿酒葡萄和葡萄酒的理化指标与葡萄酒的外观、香气、口感的关系。
结果如下:

红葡萄酒的质量与酿酒葡萄和葡萄酒的理化指标间的关系
外观：$S_1 = 0.054y_{12} + 7.776$
$R^2 = 0.183$<0.6

香气：$S_2 = 16.552x_{10} + 16.932$
$R^2 = 0.419$<0.6

口感：$S_3 = -0.504x_6 + 0.812y_4 + 29.907$
$R^2 = 0.543$<0.6

质量(总评分)：$S = -0.603x_6 + 0.703x_{13} + 0.248x_{31} + 61.002$
$R^2 = 0.670$>0.6

根据以上各式可以看出红葡萄酒的外观、香气、口感与酿酒葡萄和葡萄酒的理化指标拟合优度很低，因此我们认为红葡萄酒的外观、香气、口感与酿酒葡萄和葡萄酒的理化指标没有直接关联；虽然质量与酿酒葡萄和葡萄酒的理化指标拟合优度一般，但可以接受，总体质量与酿酒葡萄和葡萄酒的理化指标存在一定的相关性。

白葡萄酒的质量与酿酒葡萄和葡萄酒的理化指标间的关系：
外观：$S_1 = -0.069x_{26} - 4.225y_{11} + 447.504$
$R^2 = 0.483$<0.6

香气：$S_2 = 0.544x_{12} - 0.773x_{13} - 2.646x_{29} + 34.088$
$R^2 = 0.634$>0.6

口感：$S_3 = 0.002x_1 - 0.542x_7 - 0.486x_{11} - 6.429x_{19} - 0.072x_{27} + 63.280$
$R^2 = 0.706$>0.6

质量（总评分）：$S = 0.462x_5 + 21.717x_{17} - 3.069x_{23} - 0.013x_{33} + 0.013x_{34} + 0.583x_{42} + 21.307y_8 - 12.657y_8 - 12.657y_{11} - 2.527y_{15} + 1360.147$

$R^2 = 0.891 > 0.6$

根据以上各式可以看出白葡萄酒的香气、口感与酿酒葡萄和葡萄酒的理化指标拟合优度良好，且质量与酿酒葡萄和葡萄酒的理化指标拟合优度极好，但外观与酿酒葡萄和葡萄酒的理化指标拟合优度很差，因此我们认为，白葡萄酒外观与酿酒葡萄和葡萄酒的理化指标没有直接关系，白葡萄酒的香气、口感和总体质量与酿酒葡萄和葡萄酒的理化指标存在良好的相关性。

由于香气对葡萄酒质量的影响，可能芳香物质在一定程度上影响了葡萄酒的质量，因此我们就继续采用逐步回归的方法利用 spss 求解出酿酒葡萄的芳香物质和葡萄酒的香气之间的关系，结果如下:

红葡萄酒芳香物质对香气分析的影响：

$S_2^{'} = 2.116z_5 + 2.557z_{30} - 7.675z_{56} + 21.554$

$R^2 = 0.669 > 0.6$

可以得出，红葡萄酒的香气与红葡萄酒芳香物质中的乙酸正丙酯、乙酸 2-吡咯烷酮拟合优度良好，存在良好的相关性

白葡萄酒芳香物质对香气分析的影响：

$S_2^{'} = 0.329z_{19} - 0.520z_{30} - 0.728z_{60} + 23.794$

$R^2 = 0.441 < 0.6$

可以得出，红葡萄酒的香气与红葡萄酒芳香物质的拟合优度很差，因此可以认为白葡萄酒芳香物质和香气分析之间没有关系。

总体来看，无论是红葡萄酒和白葡萄酒，它们的质量都与酿酒葡萄和葡萄酒的理化指标存在良好的相关性。因此，我们有理由认为，能够用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。

六、模型的评价

6.1 模型的优点

对于问题一，首先进行绘制出两组红葡萄酒与白葡萄酒评分的正态分布直方图和评分的正态 Q-Q 图。并且运用 正态性检验来判断给定数据服从正态分布。使论文得出的结论更有说服力；其次，运用显著性检验，对两组评酒员的评分进行假设检验，得出两组评酒员对于葡萄酒的评价没有显著性差异的结论；最后通过信度分析求出两组评酒员 α 系数得出，第一组评酒员打分的结果更可信。

对于问题二，首先利用主成分分析法，有效的降低了数据的维数。再次利用主成分估计得主成分回归方程,最后带入酿酒葡萄理化指标得到酿酒葡萄的综合评价分数，对评价分数进行分级得到酿酒葡萄的评价等级。此模型将抽象的问题通过主成分，回归等一系列方法将抽象的问题转换为具体的分数，操作简单，模型可实现度高。

对于问题三,通过逐步分析法将葡萄酒的理化指标与酿酒葡萄的主要理化指标联系起来，将数十项指标简化为少量的指标，使模型更加简化，增加了模型的深度和广度。

对于问题四，延用问题三的模型，使问题简单可行，但值得注意的是，由于香气对葡萄酒质量的影响，可能芳香物质在一定程度上影响了葡萄酒的质量。

6.2 模型的缺点

问题三和四中的理化指标并没有采用附件中的第一指标，而是将所有的理化指标都代入计算。且问题四中最终的结论没有将外观、香气、口感和质量结合到一起，最终结论不够综合。

七、参考文献

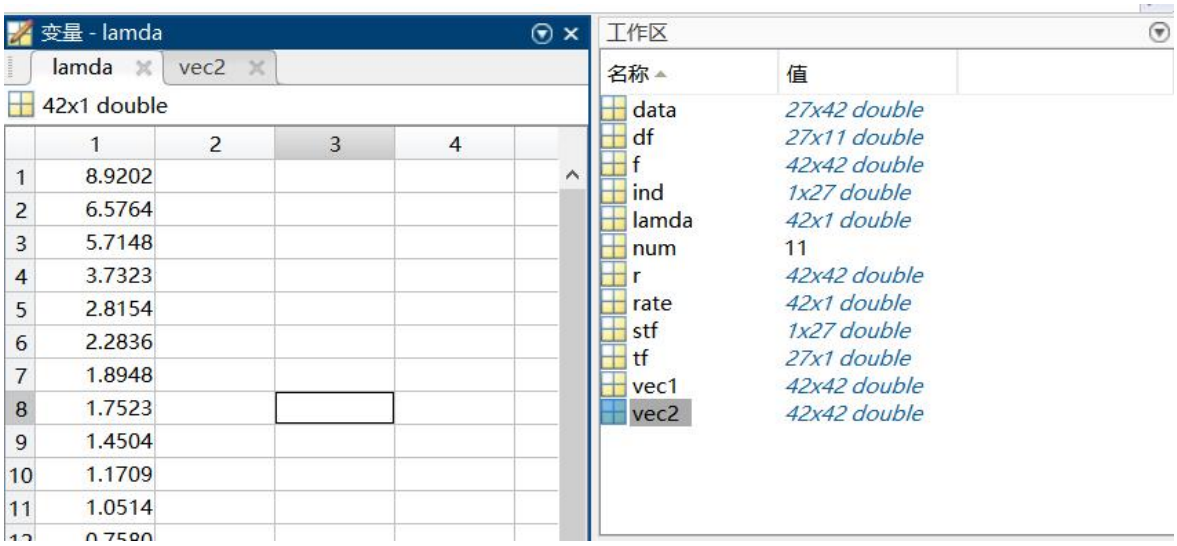
[1]SPSS 统计分析从基础到实践[M]. 北京：电子工业出版社 ， 罗应婷, 2007
[2]数学建模方法及其应用[M]. 北京：高等教育出版社 ， 韩中庚, 2005
[3]数学建模算法与应用[M].北京：国防工业出版社,司守奎、孙玺菁,2011
[4]SPSS 统计分析从入门到精通[M].北京：人民邮电出版社，杜强、贾丽艳，2009
[5]刘立祥.线性回归模型中自变量的选择与逐步回归方法[J]. 统计与决策,2015(21期)：82-86

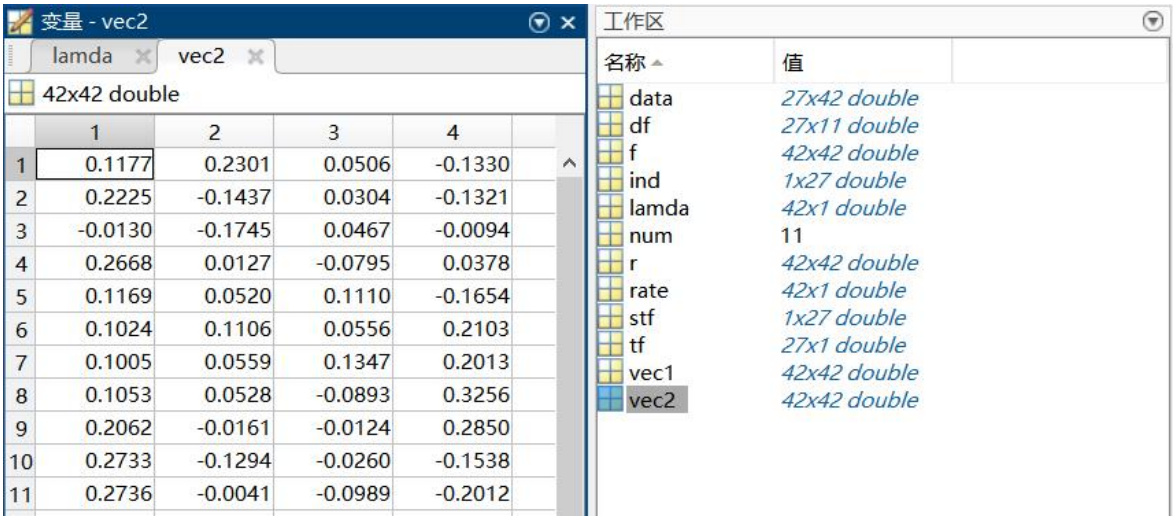
附录

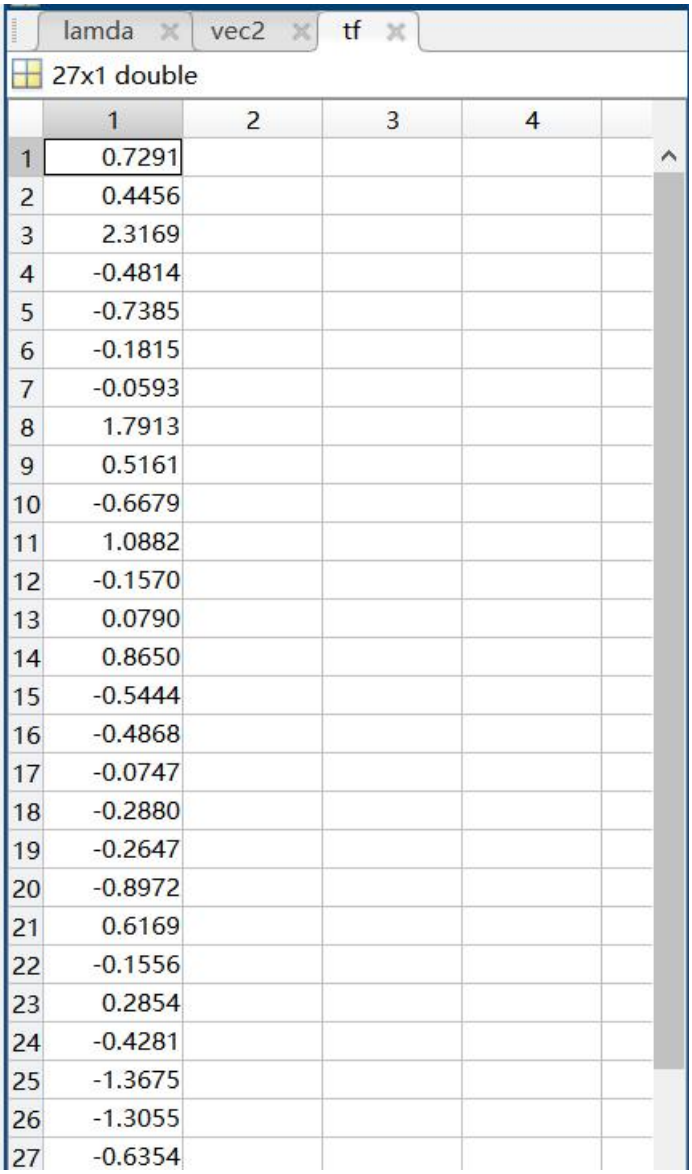
附录 1 问题二 解决主成分分析的 Matlab 程序

```
clc,clear
data = load('红理化.txt');%将原始数据保存在 txt 文件中
data=zscore(data);          %数据的标准化
r=corrcoef(data);           %计算相关系数矩阵 r
%下面利用相关系数矩阵进行主成分分析，vec1 的第一列为 r 的第一特征向量，即主成分的系数
[vec1,lamda,rate]=pcacov(r);                    %lamda 为 r 的特征值，rate 为各个主成分的贡献率
f=repmat(sign(sum(vec1)),size(vec1,1),1);       %构造与 vec1 同维数的元素为±1 的矩阵
vec2=vec1.*f;                    %修改特征向量的正负号，使得每个特征向量的分量和为正，即为最终的特征向量
num = max(find(lamda>1)); %num 为选取的主成分的个数,这里选取特征值大于 1 的
df=data*vec2(:,1:num);       %计算各个主成分的得分
tf=df*rate(1:num)/100;       %计算综合得分
[stf,ind]=sort(tf,'descend');   %把得分按照从高到低的次序排列
stf=stf'; ind=ind';                %stf 为得分从高到低排序，ind 为对应的样本编号
```

程序的运行结果：







附录 2 红白葡萄各个主成分的得分表达式

```
data = importdata('红主成分.txt');%保存好的文本数据
```

```
a = data.data;%所有的参数
b = data.textdata;%所有的字母变量
a = string(a);%转换为字符串格式
b = string(b);
[m,n] = size(a);%m 为每个公式中的变量个数，n 为公式个数
c = [];
for i=1:n
    d = '';
    for j=1:m
        if contains(a(j,i),'-')%判断是否含有减号
            d = strcat(d,a(j,i),b(j));
        else
            d = strcat(d,'+',a(j,i),b(j));%添加加号
        end
    end
    c = [c;d]
end
```