葡萄酒的质量分析及评价

摘 要

目前葡萄酒的质量主要通过评酒员进行打分判断，本文通过建立统计学模型对两组评酒员评价结果可信度进行了讨论，同时给出了酿酒葡萄的分级策略以及对酿酒葡萄与葡萄酒的理化性质之间的联系进行了研究。

针对问题一，首先对附件一的数据进行了**异常值处理**。为了判断两组品酒员的评分结果是否具有显著性差异，需要酒样数据进行**正态检验**，若满足正态分布则使用配样 *t* 检验进行显著性分析，反之则使用非参数检验。经检验酒样数据服从正态分布，因此采用**配样 *t* 检验**进行显著性分析，得到了两组评酒员的评酒结果**具有显著性差异**的结论。同时通过**方差分析**两组评酒员的打分，基于此得到了第二组评酒员的评价结果更可信。

针对问题二，利用**标准化处理**后的葡萄酒评分，我们建立葡萄的理化指标与葡萄酒评分之间的**逐步回归**关系，通过该两数关系得出葡萄酒的预期评分；接着利用葡萄酒得分对葡萄酒分级通过进行 ***K-means*** 聚类分析，划分出所给葡萄酒样品的等级，进而通过分析酿酒葡萄所酿葡萄酒预期得分与葡萄酒样品各等级之间的“距离”末对酿酒葡萄进行分级。

针对问题三，需要我们建立葡萄酒和理化指标之间的联系，由于葡萄的理化指标众多，先通过**相关系数**矩阵确定了葡萄酒与葡萄理化指标中具有较大相关性的指标，从而实现了对葡萄理化**指标筛选**。再利用**多元线性回归**的方法确定了葡萄酒理化指标与葡萄理化指标间一对多的函数关系，通过分析得到葡萄酒理化指标与葡萄理化指标之间具有较强相关性的结论。

针对问题四，沿用第三题的思想分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并根据结果求解的相关性系数进行分析，得到了不能用葡萄和葡萄酒的理化指标来评价葡萄酒的质量的结论。

关键词：配样 *t* 检验　逐步回归　*K-means* 聚类　相关性分析　多元线性回归

一、 问题重述

1.1 问题背景

当今社会，随着人们生活水平的提高，人们对作为时尚品的葡萄酒的质量要求也越来越高。在确定葡萄酒质量时，人们一般会聘请一批资深的评酒员进行评比，根据不同指标所得分数求和得到总分，以此确定葡萄酒的质量。其中，酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系，酿酒葡萄的理化指标和葡萄酒的理化指标会在一定程度上反映葡萄和葡萄酒的质量。

本题给出了 3 份材料，附件 1 是不同评酒员对不同样品的评价结果，附件 2 给出了白葡萄、红葡萄的理化指标和白葡萄酒和红葡萄酒的理化指标，附件 3 给出了葡萄和葡萄酒的芳香物质。

1.2 问题提出

（1）尝试建立数学模型，分析附件 1 中两组评酒员的评价结果有无显著性差异，哪一组结果更可信。

（2）根据附件 2，即根据酿酒葡萄的理化指标和葡萄酒的质量，建立模型对这些酿酒葡萄的品质进行分级。

（3）建立数学模型分析酿酒葡萄理化指标和葡萄酒理化指标的关系。

（4）探讨酿酒葡萄和葡萄酒的理化指标对葡萄酒的质量的影响，并且论证能否利用酿酒葡萄和葡萄酒的理化指标来评价葡萄酒的质量。

二、 模型假设

1. 评酒员的资质较高，不存在故意乱打分的情况。
2. 各个样品酒原产地相似，酿酒葡萄的产地对葡萄酒的质量影响相同。
3. 仪器对样本理化指标和所含芳香物质的测试不存在随机误差，且附件所给数据真实、准确、可靠。
4. 酒样品容量较大时，认为各组样本服从正太分布且相互独立。
5. 两种葡萄酒和酿造葡萄的分级标准相同，且葡萄酒分为优、良、合格、不合格四个级别。（仅供参考，根据后续需要可适当修改）
6. 假设附件 1 中，酒样品为一级指标；外观、口感、香气分析和整体评价为二级指标；澄清度、色调、纯正度、浓度、持久性和质量为三级指标。

三、 符号说明

符号	说明
σ	标准差
α	显著性水平
$N_{1..2}$	酒样总数/评酒员总数
s_d	配对样本差值的标准偏差
D_n	***K-S*** 检验统计量
Sig	双尾响应值
r	相关系数

四、 问题分析

4.1 问题一的分析

要想比较两组评酒员的评价结果是否存在差异，并建立合理的评价模型以判断两组结果在可信程度的优劣，我们可以先对附件 1 的数据进行观察分析，易知葡萄酒样品评分为百分制，外观、口感等指标占据一定比例。对于评价结果是否有显著性差异的判断，我们要先判断样本数据是否满足正态分布，若采取正态分布，我们可以继续利用配对 t 检验法对每种酒的最终得分进行分布检验；反之，则可以采取非参数检验法。而对于两组评酒员评价结果可靠性的判断，我们可以选择方差开根号，即标准偏差来反映，标准偏差越小，可靠性越大。

4.2 问题二的分析

问题二要求我们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。由常识可以知道，葡萄酒的质量很大程度上取决于酿酒葡萄的质量，优质的葡萄酒对应优质的酿酒葡萄，劣质的葡萄酒对应的酿酒葡萄质量也相应较差，因此我们考虑利用附件一中所给的葡萄酒质量评分作为参考标准建立聚类分析模型对葡萄进行分级。

4.3 问题三的分析

问题三要求我们分析酿酒葡萄与葡萄酒的理化指标之问的联系，由于葡萄酒与酿酒葡萄有多个理化指标，因此简单的两指标问相关分析不再适用。分析可知酿酒葡萄的理化指标影响了葡萄酒的理化指标，它们之问并不是互相影响而是一种因果关系，因此考虑建立模型，描述多个葡萄酒理化指标与酿酒葡萄的多个理

化指标之间的联系，通过这种联系分析酿酒葡萄理化指标对葡萄酒理化指标的影响。根据附件二可知酿酒葡萄理化指标数量较多，而样本量较小，取过多的酿酒葡萄指标进行分析难免产生较大的误差，因此必须先对酿酒葡萄的理化指标进行筛选，再建立多元线性回归方程求解。

4.4 问题四的分析

要想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，我们依然采用第三问的思考方法。经过第三间的分析与求解，我们得出的结论是:葡萄酒理化指标与酿酒葡萄的理化指标之间具有比较高度的相关性。并且，分析可知葡萄酒的理化指标对葡萄酒质量的影响更为直接，而酿酒葡萄的理化指标必须通过葡萄酒的理化指标米间接影响葡萄酒的质量，因此我们考虑分析葡萄酒的理化指标对葡萄酒质量的影响，进而利用葡萄理化指标与葡萄酒理化指标之问高度的相关性来分析葡萄的理化指标对葡萄酒质量的影响。通过查阅文献与网络资料，我们得知葡萄中的芳香物质对所酿出的葡萄酒的气味、口感等方面有比较大的影响，初步分析可以通过葡萄酒或葡萄中的芳香物质来评价葡萄酒的质量。同时，由于附件三中的芳香物质种类众多，必须对芳香物质进行筛选。葡萄酒的香气与口感占评分体系的比重较大，且通过文献资料可知芳香物质对香气与口感确实有比较大的影响，因此考虑利用芳香物质对香气分析、口感分析评分的相关程度作为筛选的标准。

五、 数据预处理

对于大数据题目我们首先需要进行原始数据进行预处理，对于附件可能出现的如空值，超出取值范围的异常值，不合理的重复值等进行处理，以达到降低异常数据对模型分析影响的目的。

5.1.1 原始数据分析

（*a*）异常值分析

对于本题我们通过绘制附件一中各种指标对应的箱型图，通过观察各组数据的离散情况，发现存在异常值如下图所示：

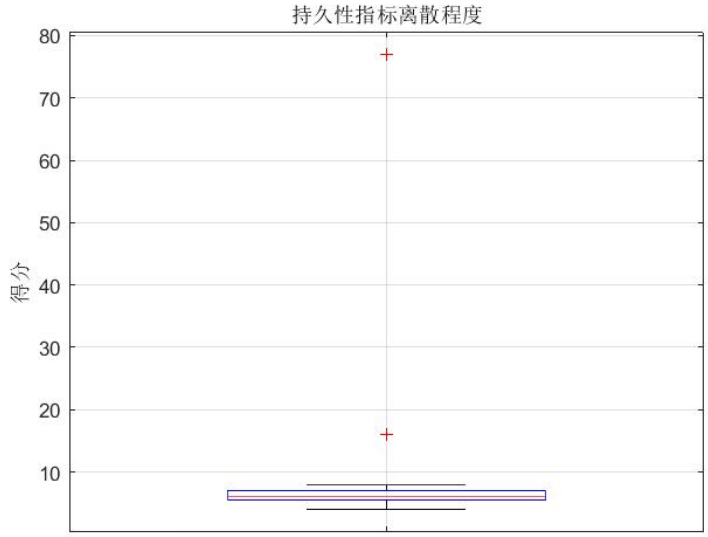

图 1：持续性指标箱型图

可以发现对于满分为 8 分的持久性指标，存在两个超过指标的异常值分别为第一组白葡萄酒 3 号样品 9 号评酒员的 77 分，第一组白葡萄酒 8 号样品 9 号评酒员员的 16 分。

（*b*）空值分析

本文通过对附件一数据采用各组酒样不同的评酒员打分的平均值作为特征进行讨论。通过 *Matlab* 使用 *xlsread* 进行数据读取，截取目标区域的数据，再利用 IF（ISNAN）进行空值判断。

基于此，可以得到对于附件一中存在一组空值数据，为第一组红葡萄酒 20 号酒样 4 号评酒员对酒的色调评分。

5.1.2 原始数据预处理

通过上文分析，我们找到了附件数据的异常值，为了降低异常数据对模型分析的影响，采用平均值对异常数据进行修正具体方案如下：

(1)第一组白葡萄酒 9 号评酒员对 3 号酒持久性评价“77”分更换为“7”分。

(2)第一组白葡萄酒 9 号评酒员对 8 号酒持久性评价“16”分更换为“6”分。

(3)第一组红葡萄酒 4 号评酒员对 20 号酒的色调评分更换为“6”分。

六、 模型的建立及求解

6.1 问题一模型的建立和求解

问题一要求分析附件 1 中两组评酒员的评价结果有无显著性差异，并判断哪一组更为可信。对于显著性差异可以通过对附件数据是否服从正态分布，采用配样 *t* 检验和非参数检验的方法进行判断；对于评酒员可信度，可以采用方差分析的方法，从方差大小判断两组评酒员的打分客观程度。

6.1.1 *PartI* 检验显著性差异

基于上文，为了检验两组评酒员的评分结果是否显著性差异，需要根据对附件一的数据进行正态分布的检验。基于附件一数据的服从情况，选用不同的检验模型，具体给出流程图如下：

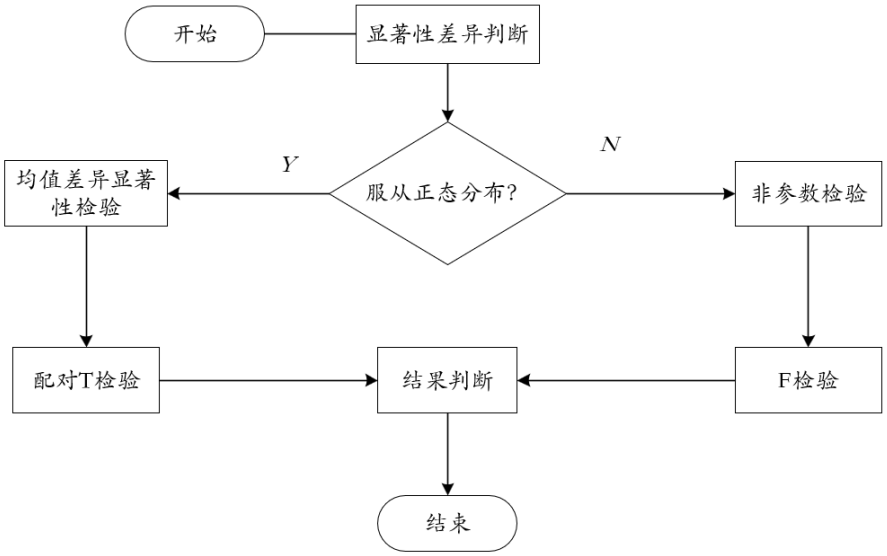

图 2：显著性差异判断流程

（*a*）正态分布检验

基于上文，为了检验两组评酒员打分结果之间是否具有显著性差异，首先需要对附件一数据进行正态分布检验，通过数据预处理后我们可以得到两组红白葡萄酒评酒员的打分均值。

基于此，分布绘制出两组红白葡萄酒的评酒员的打分频数分布直方图进行分析，可以得到直方图结果如下图所示：

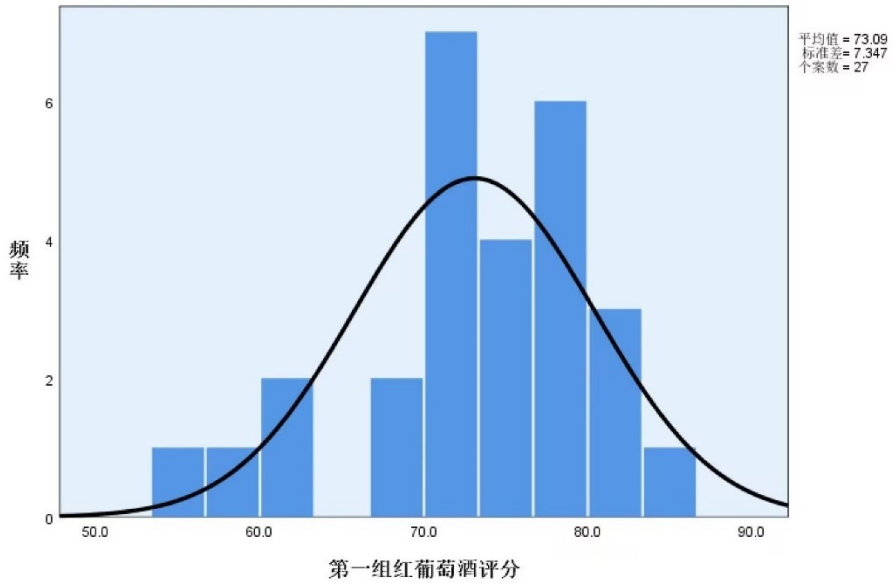

图 3：第一组红葡萄频率分布直方图

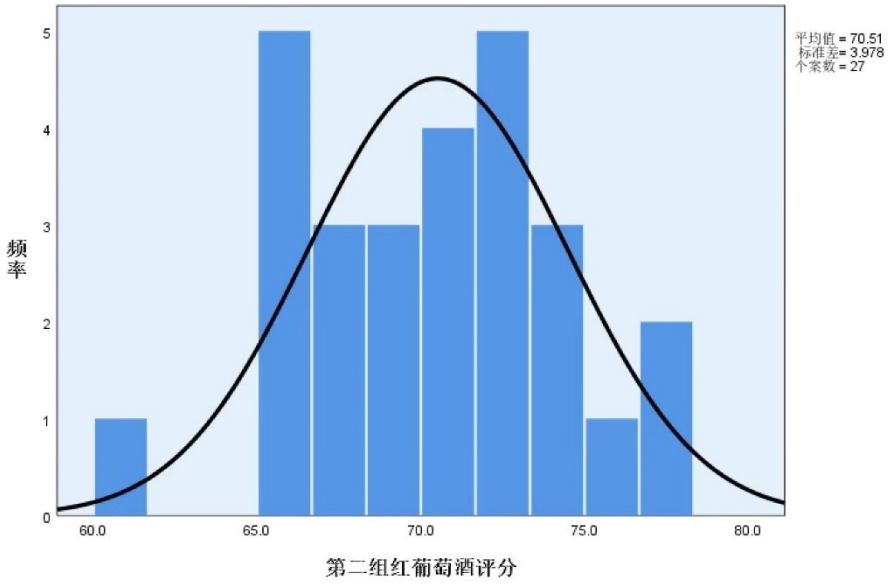

图 4：第二组红葡萄频率分布直方图

同理绘制出两组白葡萄酒的频率分布直方图如下：

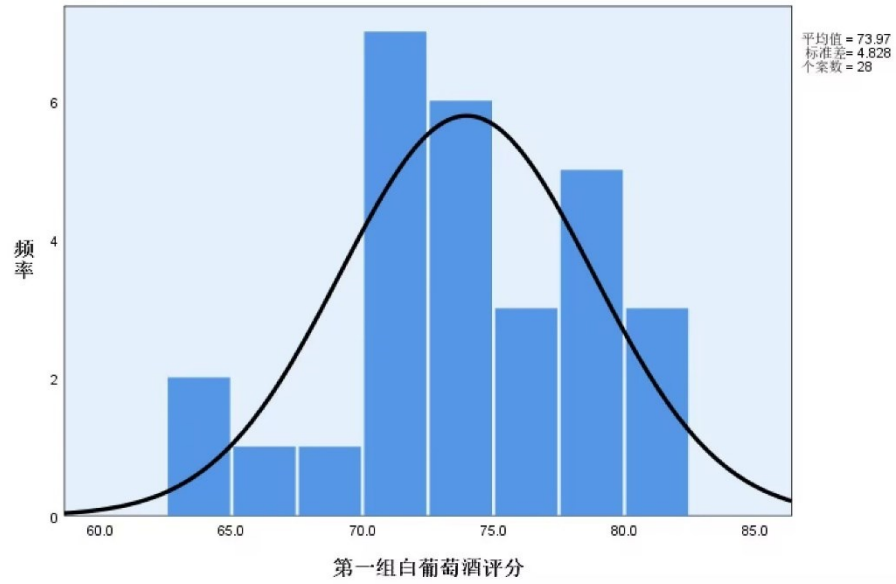

图 5：第一组白葡萄频率分布直方图

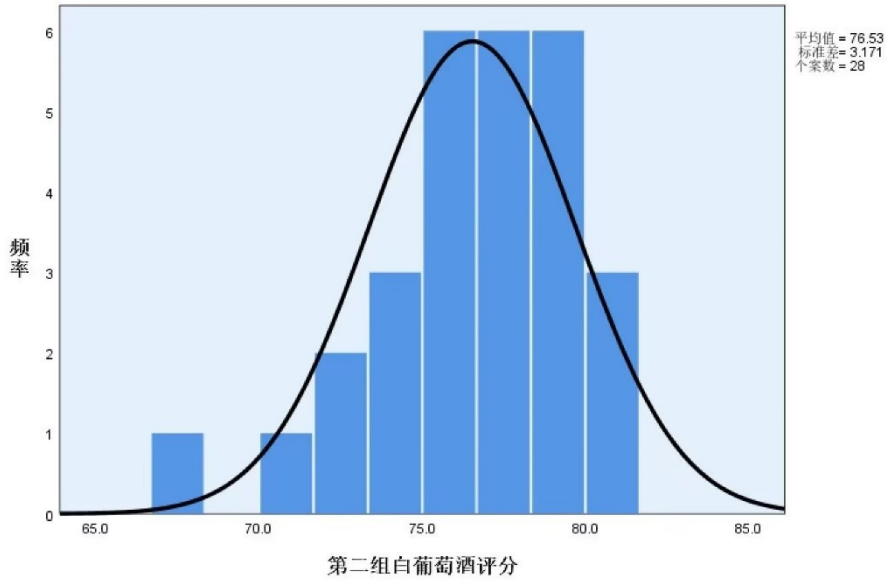

图 6：第二组白葡萄频率分布直方图

根据两组红白葡萄酒的得分频率分布直方图可以得知，葡萄酒样品的质量大致服从正态分布，同时为了进一步建立数据服从正态分布的依据，可以采用数值统计分析 ***K-S*** 检验法对结论进行验证。

(*b*) *Kolmogorov-Smirnov* 统计验证

为了进一步确立数据服从正态分布，可以采用数值统计分析 ***K-S*** 检验法对结论进行验证，本文以第一组红葡萄酒为例建立 *K-S* 检验模型，并在设立显著性水平为 $\alpha = 0.05$ 的基础下进行讨论。

Step1 提出假设：

原假设 H_0：数据集服从正态分布。

备择假设 H_1：数据集不服从正态分布。

Step2 确定显著性水平：

基于上文，依据小概率原则规定显著性水平 $\alpha = 0.05$。

Step3 计算检验统计量

设经验分布函数为 F_{exp}，对每个评酒员的打分情况进行排序得到随机变量函数 $F_{obs}(y_i) = \dfrac{i}{n}$ $(i = 1,2,3....,27)$，基于此给出 *K-S* 检查法的检验统计量记为 D_n 数学表达式如下：

$$D_n = max\,|F_{exp}(x) - F_{obs}| \tag{1}$$

Step4 进行假设检验给出结论：

基于上文，在给定显著性水平$\alpha = 0.05$的前提下，根据 *K-S* 检验统计量计算出双尾响应 **Sig**，若响应值大于 0.05 则认为其满足原假设，反之为备择假设。

根据上文，将附件 1 中的处理通过 *SPSS* 进行处理求解 *K-S* 模型可以得到模型相关参数如下表 1 所示：

表 1：K-S 检验求解结果

		一红	二红	一白	二白
正态参数	平均值	73.085	70.515	74.315	76.500
	标准偏差	7.3472	3.9780	4.5698	3.2267
	绝对	.156	.124	.094	.119
	正	.089	.078	.094	.076
	负	-.156	-.124	-.086	-.119
检验统计		.156	.124	.094	.119
渐近显著性（双尾）		.091	.200	.200	.200

根据上表可知双尾检验响应值 Sig 均大于 0.05，因此进一步验证了数据服从正态分布的结论，下文都将基于正态分布进行讨论。

(*c*) 配对 *t* 检验模型的建立

与上文，通过了直方图进行初步分析以及 *K-S* 检验法都验证了附件 1 数据服从正态分布的结论。根据显著性差异检验流程图，选用配对 *t* 检验进行对两组评酒员的显著性差异判断。

Step1 提出假设：

原假设H_0：两组评酒员对葡萄酒的评分的平均值相等。

备择假设H_1：两组评酒员对葡萄酒的评分的平均值不相等，亦即两组评酒员的评价结果存在显著性差异。

Step2 确定显著性水平：

基于上文，依据小概率原则规定显著性水平$\alpha = 0.05$。

Step3 计算检验统计量

分别设x_i, y_i $(i = 1,2,3...27)$为第一组和第二组的红葡萄第i酒样的总分，

基于此给出如下定义：

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} \tag{2}$$

$$t = \frac{\sqrt{n}(\bar{d} - \mu)}{s_d} \tag{3}$$

其中，$d_i, \bar{d}$ 分别为偏差和配对样本差值的平均值，s_d 为配对样本差值的标准偏差，t 为样本统计量。

Step4 进行假设检验给出结论：

基于上文，在给定显著性水平 $\alpha = 0.05$ 的前提下，根据配样 T 检验统计量计算出双尾响应 *Sig*，若响应值大于 0.05 则认为其满足原假设，反之为备择假设。

(*d*) 配样 *t* 检验模型求解

根据上文，将附件 1 中的处理通过 *SPSS* 进行处理求解配样 *t* 模型可以得到模型相关参数如下表 2 所示：

表 2：配对 *t* 样本检验结果

		平均值	标准偏差	配对差值	差值 95%区间		t	自由度	Sig
				标准误差平均值	下限	上限			
配对 1	红 1-红 2	2.5704	8.1126	1.5613	-.6389	5.7798	1.6446	26	.012
配对 2	白 1-白 2	-2.5571	6.1341	1.1592	-4.9357	-.1786	-2.2026	27	.036

分析结果可知，由于 $sig1 = 0.012 < 0.05$，$sig2 = 0.036 < 0.05$，均拒绝原假设。综上，两组评酒员的评价结果存在有显著性差异。

6.1.2 *PartII* 评酒员可信度判断

题目一中同时还要求我们对两组评酒员的评分结果进行可信度判断，考虑到评分结果带来的不可信因素主要来源于主观判断，因此可以采用方差分析分别在红、白葡萄酒的评分中对两组评酒员的主观程度给出量化。

（*a*）方差分析模型的建立

方差在概率论及统计学中描述的是一个随机变量的离散程度，即一组数字与其平均值之间的距离的度量，是随机变量与其总体均值或样本均值的离差的平方的期望值，给出方差数学表达式如下：

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 \tag{4}$$

其中 N 为各组评酒员人数，x_i 为不同评酒员对同一酒样的打分，$\bar{x}$ 为一组人对同一种酒样打分的均值。

本文认为，当不同专家对同一种酒样的评分差距越小时，说明该组评分是越客观的，定义一组专家的评价误差程度均值 S 如下：

$$\bar{\sigma} = \frac{1}{N_2}\sum_{i=1}^{N_2}\sigma_{ij} \quad (j=1,2) \tag{5}$$

其中 N_2 为酒样数，σ_{ij} 为对于酒样 i 号第 j 组品酒员的方差。

（*b*）模型的求解

基于上文，首先通过 *Matlab* 对 *Excel* 附件数据进行预处理得到每个酒样每组品酒员对应的标准差，再通过 *plot* 绘出两组评酒员分别在红、白葡萄酒的方差大小对比图如下：

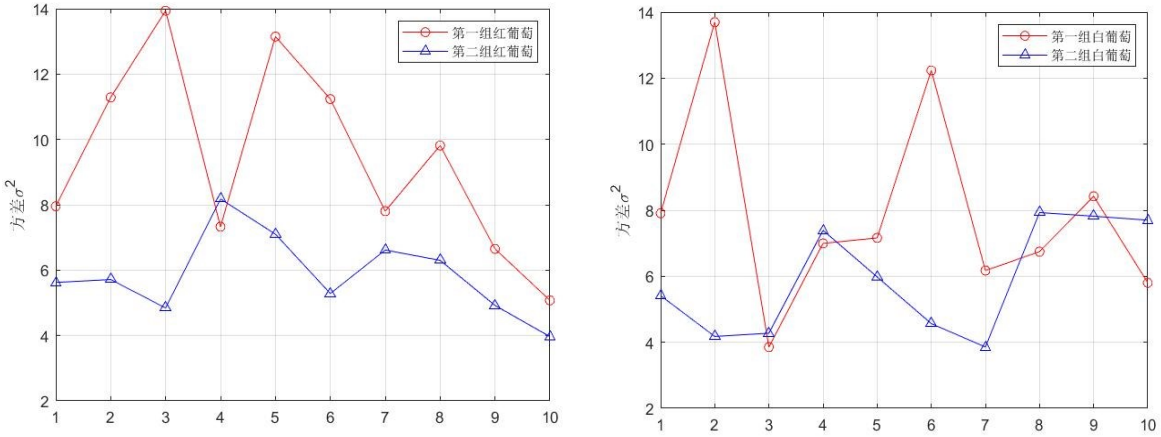

图 7：两组评酒员方差对比

根据图七，可以初步判断对于红葡萄酒第一组评酒员的方差明显高于第二组，对于白葡萄酒第一组评酒员有同样的趋势，因此为了进一步确定两组评酒员的方差特征给出均值方差如下表所示：

表 3：评价分数方差求解结果

$\bar{\sigma}$	红葡萄酒	白葡萄酒
第一组评酒员	7.413148	10.55195
第二组评酒员	5.620081	7.069039

根据表 3，可以明显得出第二组评酒员的方差无论在红葡萄酒还是白葡萄酒均小于第一组评酒员。也即，本文认为第二组评酒员的评价结果更为客观，相较于第一组评酒员更具有可信度。

6.2 问题二模型的建立和求解

问题二要求我们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。

根据常识可以知道酿酒葡萄的质量在很大程度上决定了葡萄酒的质量，优质的葡萄酒使用了优质的酿酒葡萄，相应的劣质的酿酒葡萄使用了劣质的酿酒葡萄，因此我们参考第一问中对葡萄酒质量的评分作为参考标准，对葡萄酒进行分级。

但是葡萄酒的质量也不完全取决于酿酒葡萄，葡萄酒和酿酒葡萄的理化指标在一定程度上也反映了葡萄酒和酿酒葡萄的质量，因此我们也会引入相应的理化指标作为参考标准。

6.2.1 数据的标准化处理

处理一 对附件一中数据再分析，很明显看到对葡萄酒的质量评定为品酒员的感官评定，这种评价为典型的人为评定，所以将会不可避免的产生误差。

因此我们需将数据进行标准化处理。

设为第 P_i 号酒样被第 j 品酒员评价的分数，$P_{ij}^{'}$ 为标准化后的评价分数，$\overline{P}_j$ 为第 j 号品酒员的平均打分，σ_j 为第 j 号品酒员打分的方差。

标准化公式：

$$P_{ij}^{'}=\frac{P_{ij}-\overline{P_j}}{\sigma_j} \tag{6}$$

处理二 由于理化指标的数值大小不同、数据波动范围不同，为了消除这些影响，我们先对理化指标进行标准化处理。其中 x 为某种指标的原始数据，σ_x 为改理化指标的方差。

标准化公式：

$$x_i^{'}=\frac{x_i-\overline{x}}{\sigma_x}$$

6.2.2 模型的建立

利用标准化处理后的葡萄酒评分，我们建立葡萄的理化指标与葡萄酒评分之间的函数关系，通过该两数关系得出某一酿酒葡萄酿出的葡萄酒的预期评分;接着利用葡萄酒得分对葡萄酒进行聚类分析，划分出所给葡萄酒样品的等级，进而通过分析酿洒葡萄所酿葡萄洒预期得分与葡萄酒样品各等级之间的“距离”末对酿酒葡萄进行分级。从而既考虑到了所酿葡萄酒评分对葡萄分级的影响，又考虑到了酿酒葡萄的理化指标对葡萄分级的影响。

Step1 建立理化指标与评分的函数：

由于酿酒葡萄经处理后的理化指标多达 27 种，全部给子考虑将使计算十分繁琐，也不能较清晰地判断出主要影响酿酒葡萄质量的理化指标，因此我们对酿酒葡萄的理化指标进行筛选。

利用多元统计分析中逐步回归的思想，把葡萄酒评分作为因变量，对应的酿酒葡萄理化指标作为自变量，将酿酒葡萄的理化指标逐个加入到两数中进行拟合，若相应的统计量是检验显著的则保留该变量，检验不显著则剔除该变量。

该方法能够筛选出对葡萄酒评分有显著影响的酿酒葡萄理化指标，同时能够有效地减少酿酒葡萄理化指标之间的多重共线性 *VIF*。

令 y 为标准化处理后的葡萄酒评分，$x_1, x_2, ..., x_{27}$ 分别对应酿酒葡萄的 27 种理化指标，建立如下形式的函数：

$$y = B_0 + B_1 x_1 + B_2 x_2 + ... + B_i x_i$$

该函数求解得出的结果可以获得各个葡萄酒的预期评分。

Step2 酒样分类：

我们对各葡萄酒得分进行聚类，我们利用 *K*-均值聚类的方法对其进行分类，再利用各类平均分对其高低进行划分，具体步骤如下：

（1） 随机选择 K 个簇中心。

（2） 定义代价函数：

$$J(c,\mu) = \min_{\mu} \min_{c} \sum_{i=1}^{M} \| x_i - \mu_{c_i} \|^2$$

（3） 令 $t = 0,1,2,...$ 为迭代步数，重复下面两步直到 J 收敛：

对于每一个样本 x_i，将其分配到距离最近的簇

$$\arg\min_{\mu} \| x_i - \mu_k^{(t)} \|^2 \to c_i^{(t)}$$

对于每一个类簇 k，重新计算该类簇的中心

$$\arg\min_{\mu} \sum_{i:c_i^{(t)}=k} \| x_i - \mu \|^2 \to \mu_k^{(t+1)}$$

K 均值算法在迭代时，假设当前 J 没有达到最小值，那么首先固定簇中心 $\{\mu_k\}$，调整每个样例 x_i 所属类别 c_i，来让 J 函数减少，然后固定 $\{c_i\}$，调整簇中心 $\{\mu_k\}$ 使 J 减小。这两个过程交替循环，J 单调递减，当 J 递减到最小值时，$\{\mu_k\}$ 和 $\{c_i\}$ 也同时收敛。

Step3 酿酒葡萄等级划分

通过步骤一我们可以知道酿酒葡萄理化指标与葡萄酒评分的函数关系，得知了各个葡萄酒样本所酿葡萄酒的预期评分，并且我们也知道了葡萄酒评级的基准

分。我们可以测算出各个葡萄所酿酒的预期评分到葡萄酒各级基准分之间的空间距离，我们使用如下公式来计算其距离：

$$d_{i,j}=\sqrt{(y(i)-a(j))^2}$$

其中 $d_{i,j}$ 表示第 i 个葡萄所酿葡萄酒的预期评分到第 j 级葡萄酒基准分之间的距离，$y(i)$ 表示第 i 个葡萄所酿葡萄酒的预期评分，$a(j)$ 表示第 j 级葡萄酒基准分。其中同一个葡萄所酿葡萄酒样本到四个葡萄酒基准分之间的距离最小的那个，我们就将该样本分到那个级别。

6.2.3 模型的求解

***Step1* 的求解**

利用 *matlab* 对建立的函数进行逐步回归，得到的结果如下：

红葡萄酒： $R=0.908$

$$y=0.01x_2-0.016x_3+0.01x_5-0.021x_6-0.035x_7-0.053x_8+0.054x_{12}-0.04x_{14}$$
$$+0.019x_{15}+0.006x_{16}+0.042x_{18}+0.036x_{20}-0.002x_{21}+0.26x_{22}-0.004x_{23}$$

白葡萄酒： R=0.808

$$y=0.131x_3-0.34x_4+0.125x_7-0.061x_8+0.017x_9+0.085x_{10}-0.065x_{11}+0.044x_{14}$$
$$+0.082x_{16}-0.19x_{18}+0.1x_{20}+0.03x_{22}-0.039x_{24}+0.094x_{26}$$

由该回归方程可以看出酿酒葡萄理化指标决定酒样评分的主要指标，并且上述函数的 R 值均较接近 1，可以知道拟合程度较好，能够代表所有指标的变化。

利用上述函数关系，可以获得各个葡萄酒的预期评分。

***Step2* 的求解**

在各个级别内我们求其平均分可以得出下表：

表 4：葡萄酒分类标准

	一级	二级	三级	四级
红葡萄酒	1.12	0.25	-0.77	-0.86
白葡萄酒	0.56	-0.73	-0.2	-1.61

在该表中将葡萄酒划分为了四个等级，其实一级为最优等级，四级为最差的等级，在每个等级中都有划分详细的基准分。

根据该划分标准，可以将聚类后的各个酒的样品进行分级

表 5：红葡萄酒样本分级

葡萄酒样本	1	2	3	4	5	6	7	8	9	10	11	12	13	14
等级	三	一	一	二	二	三	四	三	一	二	四	二	四	二
葡萄酒样本	15	16	17	18	19	20	21	22	23	24	25	26	27	
等级	四	二	一	三	二	一	一	二	一	二	二	二	二	

表 6：白葡萄酒的样本分级

葡萄酒样本	1	2	3	4	5	6	7	8	9	10	11	12	13	14
等级	一	二	二	二	一	二	四	四	一	一	四	四	四	一
葡萄酒样本	15	16	17	18	19	20	21	22	23	24	25	26	27	28
等级	一	三	一	一	一	二	二	一	一	四	一	二	一	一

***Step3* 的求解**

利用距离法可以测算出以下分类：

表 7：红葡萄样本分级

葡萄样本	1	2	3	4	5	6	7	8	9	10	11	12	13	14
等级	二	一	一	二	二	三	四	三	一	二	四	二	四	二
葡萄样本	15	16	17	18	19	20	21	22	23	24	25	26	27	
等级	三	二	二	三	二	一	一	二	一	二	二	二	二	

表 8：白葡萄样本分级

葡萄酒样本	1	2	3	4	5	6	7	8	9	10	11	12	13	14
等级	一	三	二	二	一	二	三	四	一	一	四	四	四	一
葡萄酒样本	15	16	17	18	19	20	21	22	23	24	25	26	27	28
等级	二	三	一	二	一	二	二	一	一	四	一	二	一	二

6.2.4 模型检验

经过分析，该分级模型的误差可能来自以下几个方面：

利用逐步回归建立两数关系时产生的误差：回归分析的目的是提高拟合程度，因此为了提高拟合程度有可能将过多的理化指标变量选入函数中进行拟合。虽然显示的拟合程度很高，但由于样本量个数不大，引入过多的变量将对模型效果产生影响。

对葡萄酒聚类和分级时产生的误差：附件中提供的葡萄酒样品质量并没有覆盖所有可能的范围，在此基础上直接对葡萄酒进行分级难免欠缺妥当；并且红葡萄酒样本与白葡萄酒样本质量未必处于同一个档次，两种葡萄酒样本都分成四级可能会产生误差。

评价酿酒葡萄好坏的最主要因素是所酿葡萄酒的质量好坏，因此在大样本的基础上，酿酒葡萄的分级与葡萄酒的分级不应该有显著性的差异，据此可以检验该分级模型的有效性。可知红葡萄酒与红葡样本评级的相似度很高，白葡萄酒与白葡样评级的相似度也很高，可知该评价模型效果较好。

6.3 问题三模型的建立和求解

6.3.1 问题转化

问题三要求我们分析酿酒葡萄与葡萄酒的理化指标之间的联系，首先通过对附件的数据进行初步分析可以发现部分数据可以由其他数据直接进行表示，因此需要先进指标选择。

选择好模型的决策变量后，可以发现部分指标对结果的影响并不大，同时酿酒葡萄的理化指标多达六十余项，因此可以先通过相关性分析对理化指标进行筛选。同时根据一个自变量多个因变量的特性选用多元线性回归模型进行求解，进一步确定酿酒葡萄和葡萄酒之间的关系。

6.3.2 基于相关性分析的指标筛选

根据上文，可以得知对于酿酒葡萄的理化指标存在一定的重复性，即部分理化指标可以由其他已知的理化指标来计算得出，因此首先需要对这类指标进行剔除处理以降低对模型分析的影响。

初步筛选指标后，仍然存在许多对葡萄酒指标影响不打的理化指标，对于这类指标，需要逐一进行相关性分析，对于相关性过低的理化指标也即对结果影响不大的指标进行剔除处理。

因此，定义皮尔逊相关系数如下：

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} \tag{7}$$

也即样本相关系数 r_{xy}：

$$r_{xy}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{8}$$

如果双变量的相关系数越趋近于 1 或-1，则说明两变量之间存在较强的相关性，否则相关性较弱。其中 $\mathrm{cov}(X,Y)$ 为两变量对应的协方差，$\sigma_X\sigma_Y$ 为两变量标准差的乘积。

基于此，根据附件二的数据给出相关性矩阵如下表所示：

根据上表的相关性矩阵，可以得知对于葡萄酒影响较大的酿酒葡萄指标，对于影响较小的指标进行剔除处理以降低无关变量对模型分析的影响，并基于此建立葡萄酒和酿酒葡萄之间的关系函数。

6.3.3 多元线性回归方程模型的建立

题目三要求讨论葡萄酒和酿酒葡萄之间的联系，上文通过相关性分析对于不同的葡萄酒指标的酿酒葡萄理化指标进行了筛选。问题可以转化为对于一个自变量多个决策变量的关系模型，因此可以采用多元线性回归进行分析，建立多元线性回归模型如下：

$$y_i=\sum_{j=0}^{N_i}\beta_j x_{ij}+\varepsilon_i \quad (i=1,2\ldots9) \tag{9}$$

其中 y_i 表示葡萄酒的理化指标，x_{ij} 表示第 i 组第 j 个相关性指标，ε_i 表示第 i 组回归结果的随机误差，N_i 表示第 i 组的总指标数。

若设：

$$\boldsymbol{\beta}=(\boldsymbol{\beta}_0,\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_{N_i})^T,\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_i)^T,\mathbf{Y}=(y_1,y_2,\ldots,y_n)^T$$

$$\mathbf{X}=(x_{ij})=\begin{bmatrix}1 & x_{11} & x_{12} & \cdots & x_{1N_1}\\ 1 & x_{21} & x_{22} & \cdots & x_{2N_2}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & x_{91} & x_{92} & \cdots & x_{9N_9}\end{bmatrix}$$

则有最优解如下：

$$\begin{cases}\boldsymbol{\beta}_i=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\\ \mathbf{Y}=\mathbf{X}\boldsymbol{\beta}_i+\boldsymbol{\varepsilon}_i\end{cases}\tag{10}$$

6.3.4 模型的求解

基于上文，对附件 2 的数据进行 *SPSS* 导入选择好每个葡萄酒指标对于的酿酒葡萄理化指标。通过分析—回归—线性—选择对应的因变量、多个自变量—共线性诊断—*R* 方描述进行求解。

以红葡萄酒花色苷 y_1 为例有：

表 10: 关于y_1的多元线性回归方程求解结果

模型		未标准化系数		标准化系数		
		系数	标准误差	Beta	t	显著性
1	(常量)	-483.054	161.555		-2.990	.007
	x7	95.886	44.908	.308	2.135	.045
	x8	8.517	3.499	.363	2.434	.024
	x15	-.408	1.058	-.072	-.386	.704
	x13	8.165	12.671	.173	.644	.527
	x25	40.580	40.832	.195	.994	.332
	x10	659.296	612.732	.321	1.076	.295

为了去除酿酒葡萄各理化指标之间存在的数量级差异，选用标准化回归系数进一步优化模型，同时由于常数的标准化系数为 0，因此可以得到一个过原点的多元回归函数如下：

$$y_1=0.308x_7+0.363x_8-0.072x_{15}+0.173x_{13}+0.195x_{25}+0.321x_{10}$$

同理分别对红白葡萄酒的其余指标进行处理有：

红葡萄酒：

$$y_1=0.308x_7+0.363x_8-0.072x_{15}+0.173x_{13}+0.195x_{25}+0.321x_{10}$$

$$y_2=0.305x_1+0.392x_2-0.134x_{28}+0.103x_{25}+0.149x_{22}+0.431x_{18}+0.028x_{15}$$

$$y_3=0.194x_2-0.271x_{28}+0.129x_{25}+0.240x_9+0.105x_{25}+0.110x_{15}$$

$$y_4=0.248x_2+0.327x_9+0.313x_{26}$$

$$y_5=0.821x_{13}-0.112x_{10}-0.180x_{11}$$

$$y_6=0.510x_1-0.072x_{28}+0.340x_2-0.115x_{25}-0.084x_{15}+0.470x_9$$

$$y_7 = 0.153x_{25} - 0.166x_2 - 0.086x_8 - 0.378x_{15} + 0.475x_{29} + 0.202x_{27} - 0.162x_{26} - 0.245x_9$$

$$y_8 = -0.428x_2 - 0.024x_{14} - 0.455x_{29} - 0.272x_{21}$$

$$y_9 = 0.791x_{17} + 0.466x_5 - 0.179x_{29} - 0.337x_{22}$$

白葡萄酒：

$$y_1 = 0.572x_{12} + 0.193x_{27} - 0.717x_{15} + 0.42x_{16} + 0.169x_{21} + 1.118x_{11} + 1.111x_{13} - 0.415x_{22} - 0.274x_{25} + 0.32x_{26} + 0.135x_{10}$$

$$y_2 = 0.540x_{13} + 0.305x_{18} + 0.201x_1 + 0.171x_{27} - 0.156x_4 - 0.119x_{26}$$

$$y_3 = 0.249x_{11} + 0.168x_2 + 0.262x_6 + 0.332x_{15} - 0.247x_{25}$$

$$y_4 = 0.348x_{16} - 0.254x_6 - 0.161x_5$$

$$y_5 = 1.291x_{13} + 0.435x_{16} + 0.507x_3 - 0.25x_8 - 0.96x_{11} + 0.309x_{15}$$

$$y_6 = -0.69x_{22} + 0.254x_{26} - 0.345x_{14} + 0.402x_{17} + 1.128x_{20} + 1.078x_{21} - 0.162x_9$$

$$y_7 = 0.486x_{24} - 0.387x_{19} - 0.697x_{25} + 0.513x_{26} - 0.219x_7 + 0.126x_2 + 0.529x_5 - 0.51x_{27} - 0.255x_{22} + 0.324x_{18} + 0.161x_{13}$$

$$y_8 = 0.756x_{26} - 0.796x_{24} + 0.251x_{22} - 0.115x_{20} - 0.313x_{27} + 0.288x_{16} + 0.449x_{23} + 0.866x_{21} + 0.123x_5 - 0.075x_{19}$$

6.4 问题四模型的建立和求解

6.4.1 原始数据分布规律

题目要求我们讨论的对象是酿酒葡萄、葡萄酒与葡萄酒质量，目标是探讨前两者对后者的影响，因此，我们在分析数据前，先绘制其对应量的散点图观察据的分布规律，以花色苷和蛋白质为例如下：

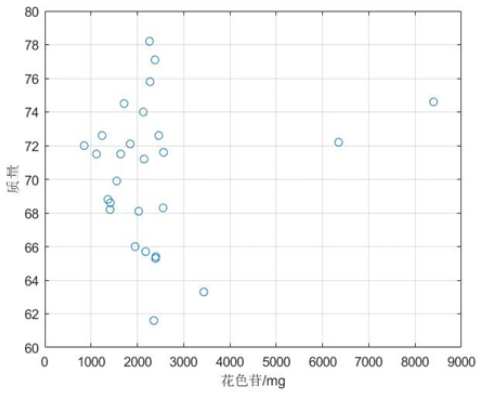

图 8：理化指标花色苷的散点图

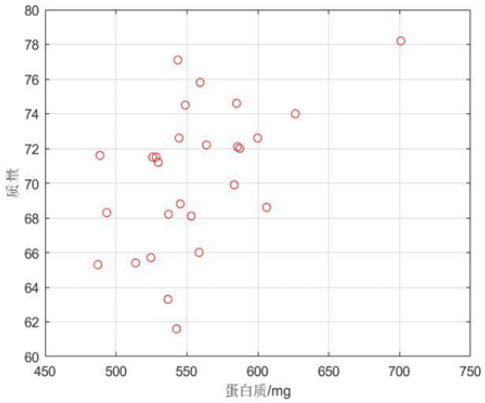

图 9：理化指标蛋白质的散点图

发现其均具有线性关系，但回归直线的斜率显然不相同，因此单一的线性秒速不足以反应题目要求。因此同样的，我们可以进行多元线性回归。

6.4.2 模型的建立

(*a*) 分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响

首先问题要求我们分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，由问题三结论知：葡萄的理化指标和葡萄酒的理化指标存在显著的多元线性关系，因此我们可用认为两者实际上是可以相互表示、等价的一组参数；即只需要分析葡萄酒的理化指标对葡萄酒质量的影响，就可说明酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响。

通过对数据的初步分析，发现葡萄酒质量大致符合正态分布且各变量散点图大致呈线性，因此沿用问题三的思想，通过构造葡萄酒质量和葡萄酒的理化指标之间的线性函数，并以此函数表示两者之间的关系。

(*b*) 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

其次问题要求我们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，由上述可知我们可以通过构建线性函数来表示两者之间的关系。得到函数后通过问题三的相关性分析判断各自变量对因变量的影响是否显著，若通过显著水平检验，则能够用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。

以红葡萄酒为例，由上述知：数据满足正态分布且具有较强的线性关系，因此我们建立葡萄酒质量与葡萄酒理化指标之间的函数关系，分析葡萄酒理化指标对葡萄酒质量的影响程度，建立线性关系如下：

$$\mu = a_0 + a_1 y_1 + a_2 y_2 + a_3 y_3 + \cdots + a_8 y_8 + a_9 y_9 \tag{11}$$

白葡萄酒：

$$v = b_0 + b_1 y_1 + b_2 y_2 + \cdots + b_7 y_7 + b_8 y_8 \tag{12}$$

6.4.3 模型的求解

同问题三使用 *SPSS* 红葡萄有如下结果：

$$\mu = 87.5 - 0.021y_1 + 0.692y_2 - 0.679y_3 + 0.592y_4 + 0.408y_5 - 5.416y_6 \\ -0.223y_7 - 0.066y_8 - 0.132y_9$$

表 11:红葡萄酒回归参数

自变量	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	C
标准差	0.011	0.775	1.303	0.709	0.407	2.447	0.129	0.1	0.143	6.4
T 统计值	-1.928	0.894	-0.521	0.835	1.002	-0.205	-1.773	-0.661	-0.992	5.354
显著性	0.071	0.384	0.609	0.415	0.330	0.740	0.101	0.517	0.369	0.03

由相关性分析判别可知：虽然有显著性大于 0.7，但于大部分自变量系数在统计上不具有显著影响水平。因此可以说明，红葡萄酒理化指标在一定程度上影响了红葡萄酒的质量，但影响有限，不能仅仅通过红葡萄酒的理化指标来评价葡萄酒的质量。

同理有白葡萄酒如下：

$$v = 86.5 - 0.017y_1 + 0.718y_2 + 2.010y_3 - 0.918y_4 - 0.379y_5 + 9.727y_6 \\ -0.133y_7 + 0.056y_8$$

表 12:白葡萄酒回归参数

自变量	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	C
标准差	0.016	1.096	1.843	1.002	0.575	37.451	0.182	0.141	7.114
T 统计值	-1.119	-0.665	1.091	-0.916	-0.658	0.260	-0.730	0.395	3.73
显著性	0.279	0.521	0.291	0.372	0.519	0.798	0.475	0.698	0.02

同样由相关性分析判别可知：虽然有显著性接近 0.8，但于大部分自变量系数在统计上不不具有显著影响水平，因此可以说明，白葡萄酒理化指标在一定程度上影响了白葡萄酒的质量，但影响有限，不能仅仅通过白葡萄酒的理化指标来评价葡萄酒的质量。

6.4.4 结果分析

综上所述，无论是在红葡萄酒还是白葡萄酒中由相关性分析知：葡萄及葡萄酒的理化指标在一定程度上影响了葡萄酒的质量，但影响水平均不通过显著性水平检验，因此不能使用葡萄和葡萄酒的理化指标来评价葡萄酒的质量。

七、 模型的评价及推广

7.1 模型的优点

1.本文使用 Excel 软件对表格数据进行处理并做出各种图表，简便、直观、快捷。
2.运用多种数学软件和统计软件对数据进行处理和显著性差异检验，如 Matlab，Excel，Spss 等，取长补短，使计算结果更加准确。
3.本文将定量和定性方法相结合，对酿酒葡萄和葡萄酒的理化指标进行了全面分析和总结。
4．本文建立的模型与实际紧密联系，充分考虑现实情况的不同阶段，从而使模型更贴近实际，通用性强。

7.2 模型的不足

1.对于一些数据进行了必要的处理，会带来一些误差。
2.模型中为使计算简便，使所得结果更加理想，忽略了一些次要影响的因素。
3.问题三建立的多元线性回归模型未在文后进行相应的显著性检验。

7.3 模型的推广

1.在问题一第一部分中，本文所建立的配对 t 检验模型不仅适用于判断两组评酒员的评价结果的显著性差异，还可适用于其他各种经济、社会、体育等活动两组评分结果的显著性差异判断。

2.在问题一第二部分中，建立的模型不仅适用于判断葡萄酒评价结果的可信度，还适用于其他需要判断多组结果可信度的情况。比如：各种比赛中，为防止由于各种原因导致的打分不公正，可采取多组裁判共同打分的方式。最后，就可利用本文的方差分析法判断每组打分结果的可信度，从而给出一个较为公平合理的和结果。

3.在问题二中，本文运用聚类分析，分别建立了红、白葡萄理化指标的分级标准。这个分级标准可以帮助酿造人员对采购来的葡萄进行快速分级，从而为后续不同等级葡萄酒的酿造奠定良好的基础。当今世界人们追逐高品质的生活，每年葡萄酒消耗量日益剧增，葡萄酒供应商的葡萄酒酿造数量更为惊人。因此，这个红、白葡萄理化指标分级标准应用前景广阔。

4.在问题三、四中，我们首先利用葡萄酒理化指标与酿酒葡萄理化指标之间相关多元线性回归矩阵，筛选出对葡萄酒某一理化指标相关程度较大的酿酒葡萄理化指标，接着通过的方法建立了葡萄酒理化指标与酿酒葡萄理化指标之间的函数关系。可以看到，理化指标之间的大部分函数关系拟合效果都不错，能够反映理化指标之间的变化关系。

参考文献

[1]赵建国,何嘉玉,李怡婷,祝利杰.数学建模经典案例分析——以葡萄酒质量评价为例[J].无线互联科技,2018,15(09):105-106.

[2]王强,汪丹丹.基于多元线性回归的葡萄酒质量评价[J].渭南师范学院学报,2013,28(09):126-130.

[3]刘令,熊奕达,赵云龙.影响葡萄酒质量的因子相关分析[J].吉林建筑工程学院学报,2013,30(05):72-74.

[4]钱圳冰,黄鸿基,冯帆,周行洲.酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响[J].中国市场,2017(18):205-206.

[5]马烁.葡萄酒评价的数学模型[J].太原师范学院学报(自然科学版),2019,18(01):15-20+52.

[6]陈晓青,蒋辉,邓伟民,贾文超.葡萄酒评价的差异分析与酿酒葡萄的分级方法——2012 年全国数学建模竞赛的数据分析[J].酿酒科技,2013(07):28-32.

附录

```
%% 提取EXCEL对应数据

sheet = "第一组白葡萄酒品尝评分";

filename = "C:\Users\19595\Desktop\A\附件1-葡萄酒品尝评分表.xls";

data = xlsread(filename,sheet);

%% 各组评分求平均

ans = [];
for i = 0 : 100
    res = 0;
    u = i *13 + 1;
    if(u >= 375)
        break;
    end
    for j = u : u + 9
        res = res + sum(data(j,1:10));
    end
    res = res / 10;
    if isnan(res)
    disp(u);
    end
    ans = [ans;res];
end

%%
% 创建一些示例数据


% 创建箱线图

boxplot(data, 'Labels', {''}, 'Whisker', 1.5, 'Widths', 0.5);


gridColor = [0.1 0.1 0.1]; % 更深的灰色，可以根据需要调整颜色值

gridLineWidth = 5; % 设置网格线宽度

grid on;
hGrid = findobj(gca, 'Tag', 'GridLines');
set(hGrid, 'Color', gridColor, 'LineWidth', gridLineWidth);
```

```
% 修改箱线图的线型和颜色

% 添加标题和轴标签
title('持久性指标离散程度');
ylabel('得分');

%% 第二组各酒样打分

filename = "C:\Users\19595\Desktop\A\附件1-葡萄酒品尝评分表.xls";
sheet = "第二组红葡萄酒品尝评分";
data = xlsread(filename,sheet);
%% 附件1.1

filename = "C:\Users\19595\Desktop\A\附件1-葡萄酒品尝评分表.xls";
sheet = "第一组红葡萄酒品尝评分";
data = xlsread(filename,sheet);
%% 附件1.3

filename = "C:\Users\19595\Desktop\A\附件1-葡萄酒品尝评分表.xls";
sheet = "第一组白葡萄酒品尝评分";
data = xlsread(filename,sheet);
data = data(2:end,[4:13]);
%% 附件1.4
```

```
filename = "C:\Users\19595\Desktop\A\附件1-葡萄酒品尝评分表.xls";

sheet = "第二组白葡萄酒品尝评分";


data = xlsread(filename,sheet);

data = data(:,[5:14]);

%% 逻辑处理红

ans = [];

for i = 0 : 100
    ans_2 = [];
    u = i *14 + 1;
    if(u >= 375) break; end
    for j = 1 : 10
        ans_2 = [ans_2 sum(data(u:u+9,j))];
    end
    ans = [ans; ans_2];
end


%% 逻辑处理白一

ans = [];

for i = 0 : 100
    ans_2 = [];
    u = i *13 + 1;
    if(u >= 375) break; end
    for j = 1 : 10
        ans_2 = [ans_2 sum(data(u:u+9,j))];
    end
    ans = [ans; ans_2];
end

%% 逻辑处理白二

ans = [];
for i = 0 : 100
    ans_2 = [];
    u = i *12 + 1;
    if(u >= 400) break; end
    for j = 1 : 10
        ans_2 = [ans_2 sum(data(u:u+9,j))];
```

```
    end
    ans = [ans; ans_2];
end


%% 标准差对比

filename = "C:\Users\19595\Documents\Tencent Files\1959558509\FileRecv\标准

差.xlsx";

sheet = "Sheet1";
data = xlsread(filename,sheet);

%%
x = 1 : 10;
y_2 = data(2:end,13);
y_1 = data(2:end,14);

y_3 = data(2:end,15);
y_4 = data(2:end,16);

figure(1)
plot(x,y_2,"ro-");
hold on
grid on
plot(x,y_1,"b^-");

ylabel("方差\sigma^2");

legend("第一组红葡萄","第二组红葡萄");



figure(2)
plot(x,y_3,"ro-");
hold on
grid on

ylabel("方差\sigma^2");

plot(x,y_4,"b^-");

legend("第一组白葡萄","第二组白葡萄");


%% 逐步回归
```

```
% 创建示例数据
x1=[5.5 2.5 8 3 3 2.9 8 9 4 6.5 5.5 5 6 5 3.5 8 6 4 7.5
    7]';%%(20维）
x2=[31 55 67 50 38 71 30 56 42 73 60 44 50 39 55 70 40 50 62
    59]';
x3=[10 8 12 7 8 12 12 5 8 5 11 12 6 10 10 6 11 11 9
    9]';
x4=[8 6 9 16 15 17 8 10 4 16 7 12 6 4 4 14 6 8 13
    11]';
y=[79.3 200.1 163.2 200.1 146 177.7 30.9 291.9 160 339.4 159.6
    86.3 237.5 107.2 155 201.4 100.2 135.8 223.3 195]';
X=[x1,x2,x3,x4];
stepwise(X(:,:), y)
```