

赛区评阅编号（由赛区组委会填写）：

2019 年高教社杯全国大学生数学建模竞赛

承 诺 书

我们仔细阅读了《全国大学生数学建模竞赛章程》和《全国大学生数学建模竞赛参赛规则》（以下简称“竞赛章程和参赛规则”，可从 <http://www.mcm.edu.cn> 下载）。

我们完全清楚，在竞赛开始后参赛队员不能以任何方式，包括电话、电子邮件、“贴吧”、QQ 群、微信群等，与队外的任何人（包括指导教师）交流、讨论与赛题有关的问题；无论主动参与讨论还是被动接收讨论信息都是严重违反竞赛纪律的行为。

我们完全清楚，在竞赛中必须合法合规地使用文献资料和软件工具，不能有任何侵犯知识产权的行为。否则我们将失去评奖资格，并可能受到严肃处理。

我们以中国大学生名誉和诚信郑重承诺，严格遵守竞赛章程和参赛规则，以保证竞赛的公正、公平性。如有违反竞赛章程和参赛规则的行为，我们将受到严肃处理。

我们授权全国大学生数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

我们参赛选择的题号（从 A/B/C/D/E 中选择一项填写）：_____

我们的报名参赛队号（12 位数字全国统一编号）：_____

参赛学校（完整的学校全称，不含院系名）：_____

参赛队员（打印并签名）：1. 杨国旭

2. _____

3. _____

指导教师或指导教师组负责人（打印并签名）：_____

（指导教师签名意味着对参赛队的行为和论文的真实性负责）

日期： 2020 年 07 月 15 日

赛区评阅编号：
(由赛区填写)

全国评阅编号：
(全国组委会填写)

2019 年高教社杯全国大学生数学建模竞赛

编 号 专 用 页

赛区评阅记录（可供赛区评阅时使用）：

评 阅 人						
备 注						

送全国评阅统一编号：
(赛区组委会填写)

基于典型相关分析和多元线性回归的葡萄酒评价论文

摘 要

随着综合国力的不断提升,人们不再满足于传统的葡萄酒,如何建立针对葡萄酒质量的评价标准,进而提升其质量成为了社会的焦点。本文考虑了多种因素对葡萄酒质量的影响,建立了合适完整的评价模型,可以很好的评定葡萄酒等级,给出合适的分数。

针对任务 1 要求比较两组评酒员的评价结果有无显著性差异,并建立合理的评价模型来判断两组结果在可信度方面的优劣,进而选出更可信的组。首先,进行数据预处理,将异常数据和少测漏测数据修正;然后,选取适用于小样本的 Shapiro-wilk 检验、直方图、正态 Q-Q 图进行正态分布验证,发现四组全部符合正态分布;最后,采用 T 检验、检查两组评分均值是否存在显著性差异,通过比较 p 值与标准差,来评定可信度高的组别。比较发现两组评酒员对于红葡萄酒的评价没有显著性差异,对白葡萄酒的评价存在显著性差异,可第二组的标准差及标准误差平均值均小于第一组,因此,认为第二组更加可信。因此,本文后续分析中,对于葡萄酒质量评价结果均选用第二组评酒员评分。

任务 2 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,确定评价酿酒质量的主要指标,建立综合评价模型。首先,根据资料筛选附表中的数据,确定对酿酒葡萄影响较大的理化指标,由于第二组的评价结果更可信,选取第二组的评价结果作为葡萄酒的质量;然后,由于主成分分析能将多个指标转换为少数几个主成分,且筛选出的主成分彼此之间互不相关,能反映出原始数据的大部分信息,采用主成分分析法,筛选出 7 个主成分,用来简化评价指标;最后,建立对酿酒葡萄分级的综合评价模型,实现对酿酒葡萄的优劣的分级。将葡萄酒分为优、良、中、差四个类别,酿酒红葡萄中品质最优的为样品 2, 5, 9, 26;酿酒白葡萄中品质最优的为样品 3, 15, 24, 27, 28。

任务 3 要求分析酿酒葡萄与葡萄酒的理化指标之间的联系。酿酒葡萄和葡萄酒的理化指标是两个包含多项数据指标变量的组别,如果直接对这些变量计算 spearman 相关系数或者 person 相关系数进行两两分析,耗时较长,且很难得到关于这两组变量之间关系的一个明确清晰的结果。而典型相关分析可以把多个变量与多个变量之间的相关化为两个具有代表性的变量之间的相关,能够揭示出两组变量之间的内在联系。首先,建立典型相关分析模型,分析酿酒葡萄与葡萄酒的理化指标之间的典型相关关系;然后,根据上述分析给出酿酒葡萄与葡萄酒的理化指标之间的联系。令酿酒葡萄为输入变量,葡萄酒的理化指标为输出变量。求解发现花色苷可以让葡萄酒变红,酿酒葡萄的 DPPH 自由基可以转化为葡萄酒中的 DPPH 半抑制体积。

针对任务 4 要求计算酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响,并论证用葡萄和葡萄酒的理化指标来评价葡萄酒的质量是否合适。同时,多元线性回归可以定量的描述某一现象和某些因素之间的函数关系,因此,选用多元线性回归方程来求解任务 4。首先,对样本进行随即筛选,选取其中一部分进行分析;其次,为了确保指标之间互不相关,进行初步筛选;然后,利用筛选后的指标和葡萄酒质量的评价结果,建立多元线性回归方程;最后,利用余下的样本的酿酒葡萄和葡萄酒的理化指标,对葡萄酒质量求解得到的多元线性回归方程进行验证。经过显著性检验和带入数据计算误差发现,计算的回归方程可以很哈皮的表达两者之间的关系。

本文采用典型相关分析、T 检验、多元线性回归、主成分分析等方法,运用 MATLAB、SPSS、EXCEL 等软件,对葡萄酒质量的评价问题进行了多角度的分析,并给出利用理化指标评价葡萄酒质量的模型,具有很大的推广价值。

关键词: Shapiro-wilk 检验; T 检验; 典型相关分析; 多元线性回归; 主成分分析;

1 问题重述

1.1 问题背景

葡萄酒的酿造可追溯至几千年前，它是一种世界通畅性酒种，有着广泛交流的基础，如今已成为世界最重要的酒种之一。葡萄酒的品质测定又叫品酒、评酒，是指评酒员通过眼、鼻、口等感觉器官对葡萄酒的外观、香气、滋味及典型性等感官特性进行分析评定的一种分析方法。一方面，评酒员必须要抛开个人喜好，排除时间、地点、环境和情绪等的影响，像一台精密的仪器一样进行感官分析；另一方面，因为葡萄酒的复杂多样及变化性，评酒员又必须充分发挥主观能动性，将获得的感觉与大脑中贮存的感官质量标准进行比较分析。只有兼顾以上两个方面，才能保证结果的精确性。同时各个评酒员之间还必须保证分析结果的一致性，一致性和精确性是正确性的必要保证。

怎样面对不同的影响因素对葡萄酒进行品评，成为一个困扰着无数人的问题。

1.2 问题提出

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。针对葡萄酒的品评，提出了以下四个问题：

1. 分析附件1中两组评酒员的评价结果有无显著性差异，哪一组结果更可信？
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？

1.3 待解决问题

根据上述情况，本文需要解决以下四项任务：

1. 分析两组评酒员的评价结果有无显著性差异，选出更可信的组数。
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. 计算酿酒葡萄与葡萄酒的理化指标之间的联系。
4. 计算酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，论证用葡萄和葡萄酒的理化指标来评价葡萄酒的质量是否合适。

2 问题分析

2.1 整体性分析

本题是关于葡萄酒质量评价的问题，目的是要在找出的可靠葡萄酒评分数据的基础上，设计一种可以很好评价葡萄酒质量的方法，建立适当的数学模型并用算法进行实现，从而很好的评价葡萄酒质量。对于不同任务的方案设计情况，需要考虑的因素有所差别。但不论如何最终目的都是建立一种合适准确的评价模型。四个问题整体呈现出层层递进性，相互关联，前一个任务很好的完成是后一个任务拥有良好结果的必要前提和保证。因此，分步处理好每一个任务，在求解单个任务时，尽可能多的进行验证，力求实验结果贴合实际，是十分必要的。

问题解决方案的整体执行流程，如图 1 所示。

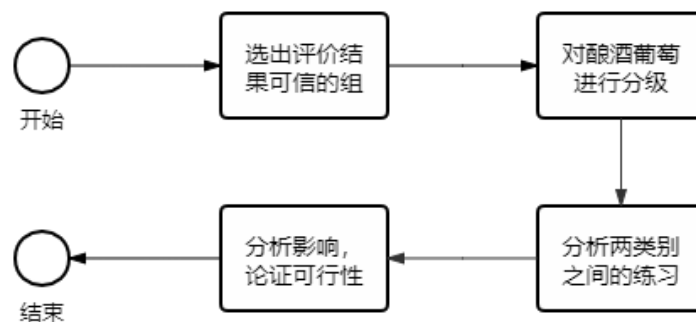


图 1 解决方案整体执行流程图

2.2 对评价体系的分析

2.2.1 对评酒员评价的分析

评酒员是具备较熟练的评酒技能，熟知各类酒固有风格，可以准确给出所品尝酒评人的专业人员，因此，采用评酒员的打分对葡萄酒的质量进行分析、了解，该途径拥有很高的技术支持，准确度、可信度高。对葡萄酒进行评价的评酒员共有两组，每组各有 10 位评酒员，分别对白葡萄酒和红葡萄酒的外观、香气、口感、平衡/整体等四大部分、澄清度、色调、纯正度、浓度、质量、持久度等 9 个小部分进行打分。多组多名评酒员对两种葡萄酒进行多维打分，可信度高，可以减少口味不同而造成的误差影响。

2.2.2 对酿酒葡萄的分析

酿酒葡萄是指以酿造葡萄酒为主要生产目的的葡萄品种，本题中酿酒葡萄共分为红葡萄和白葡萄两类。红葡萄和白葡萄拥有多种指标来进行评级，其中包含氨基酸总量、蛋白质、VC 含量、花色苷、酒石酸、苹果酸、柠檬酸、多酚氧化酶活力、褐变度、DPPH 自由基 1/IC50、总酚、单宁、葡萄总黄酮、白藜芦醇、黄酮醇、总糖、还原糖、可溶性固体物、PH 值、可滴酸、固酸比、干物质含量、果穗质量、百粒质量、果梗比、出汗率、果皮质量、果皮颜色等 28 项一级指标，天门冬氨酸、苏氨酸、丝氨酸、谷氨酸、脯氨酸等 29 项二级指标。

针对不同品种的红、白葡萄，为了减少个体差异性的影响，分别测试 27 组和 28 组葡萄样品，可通过计算求解各个样本之间的共性。而且，给出的数据中，如 VC 含量、花色苷鲜重等指标重复测量多次，在解决问题进行数据处理时，应求取多次测量的平均值，用来减少误差。

2.2.3 对葡萄酒理化指标的分析

葡萄酒分为红葡萄酒和白葡萄酒两种。在两者的理化指标中，红葡萄酒比白葡萄酒多了花色苷这一指。同时，两者又都包含单宁、总酚、酒总黄酮、白藜芦醇(mg/L)、DPPH 半抑制体积、色泽共 6 项一级理化指标，反式白藜芦醇苷、顺式白藜芦醇苷、反式白藜芦醇、顺式白藜芦醇苷、色泽 H、色泽 C 等共 7 项二级指标。在数据的处理中，针对拥有多组测量数据的 DPPH 半抑制体积等，进行多组数据求平均值的计算。

2.3 对三种任务的分析

2.3.1 任务 1 的分析

任务 1 要求比较两组评酒员的评价结果有无显著性差异，并建立合理的评价模型来判断两组结果在可信度方面的优劣，进而选出更可信的组。首先，由于给出的数据是人工测得的，可能会存在操作失误进而测错或者漏测、少测的情况发生，因此，在进行实际的处理之前，需要进行数据预处理，将异常数据和少测漏测数据修正；然后，选取适用于小样本的 Shapiro-wilk 检验，验证是否符合正态分布，分析评分是否合理准确；最后，可以采用 T 检验进行分析，通过比较 p 值，评定可信度高的组别，实现任务要求。

2.3.2 任务 2 的分析

任务 2 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，确定评价酿酒质量的主要指标，建立综合评价模型。改评价体系主要包含两个方面：葡萄酒质量和酿酒葡萄自身的理化指标。可是，由于两个方面所包含指标成分过多，如果选取全部作为评选依据的话，工程量较大，且不易评价。因此，需要选取可靠部分数据作为主要数据，来进行评级。

同时，主成分分析能将多个指标转换为少数几个主成分，且筛选出的主成分彼此之间互不相关，能反映出原始数据的大部分信息。因此，采用主成分分析可以很好的实现降维，筛选出主成分，构建评价模型的指标体系。然后，根据筛选出来的数据进行综合评价，即可完成对酿酒葡萄的分级。

2.3.3 任务 3 的分析

任务 3 要求分析酿酒葡萄与葡萄酒的理化指标之间的联系。酿酒葡萄和葡萄酒的理化指标是两个包含多项数据指标变量的组别，如果直接对这些变量计算 spearman 相关系数或者 person 相关系数进行两两分析，过程复杂，耗时较长，且很难得到关于这两组变量之间关系的一个明确清晰的结果。所以，需要改进模型，采用其他方法进行求解。

同时，典型相关分析可以把多个变量与多个变量之间的相关化为两个具有代表性的变量之间的相关，能够揭示出两组变量之间的内在联系。因此，建立典型相关分析模型，分析酿酒葡萄与葡萄酒的理化指标之间的典型相关关系是十分可靠的。然后，可以根据上述分析，得到酿酒葡萄与葡萄酒的理化指标之间的联系。

2.3.4 任务 4 的分析

任务 4 要求计算酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证用葡萄和葡萄酒的理化指标来评价葡萄酒的质量是否合适。酿酒葡萄和葡萄酒的理化指标是多个因素，葡萄酒质量只有一个结果，同时，多元线性回归是含有多个解释变量的线性回

归模型，可用于解释被解释的变量与其他多个变量解释变量之间的线性关系，所以采用多元线性回归模型十分合适。

在建立好线性回归方程之后，进行可行性检验，分析所得结果和理想值之间的误差，若误差值在可控范围内，则证明建立的线性回归方程准确可靠。因此，需要预留出来一部分数据用于线性回归结果的验证。为了实现这一目的，可在建立线性回归方程前，对样本进行随机选择，选中的样本用于建立线性回归方程，未选中的样本用于结果的检验。

3 问题假设

- (1) 假设每位评酒员的打分不受之前品酒样品的影响；
- (2) 假设每位评酒员的专业程度高，受过很好的品酒培训；
- (3) 假设本题葡萄酒质量不受葡萄产地、栽培条件、成熟度、树龄等各因素影响；
- (4) 假设本题仅仅考虑酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响；
- (5) 假设酿酒葡萄的理化指标的测量误差处于合理范围之内，可以忽略不计；
- (6) 假设本问题所使用的数据均真实有效，具有系统分析价值。

4 符号说明

本文中所涉及的主要符号及其含义如表 1 所示：

表 1 主要符号对照表

符号名称	符号含义	符号单位
m	评酒员人数	人
n	样本数	个
x_i	评酒员评定分数的异常值	分
x_m	评酒员评定分数的缺失值	分
j	样本序数	个
y	酿酒葡萄质量综合评价	—
U	7 个主成分的各个指标的载荷	—
w_i	第 i 个样品的综合得分	分
B	每一酿酒葡萄样本所在级别	—
x_k	酿酒葡萄的理化指标	—
y_k	葡萄酒的理化指标	—
β	线性回归系数	—
V	典型变量	—
W	解释变量	—

注：上表中只列举了文中所涉及的部分符号，其它符号在文中给出。其中，变量及函数名用斜体表示，向量用斜体加粗表示。

5 模型的建立与求解

5.1 任务 1 的模型建立与求解

任务 1 要求比较两组评酒员的评价结果有无显著性差异，并建立合理的评价模型来判断两组结果在可信度方面的优劣，进而选出更可信的组。求解任务 1 的策略：首先，进行数据预处理，将异常数据和少测漏测数据修正；然后，选取适用于小样本的 Shapiro-wilk 检验、直方图、正态 Q-Q 图，验证是否符合正态分布，分析评分是否合理准确；最后，采用 T 检验检查两组给出的评分均值是否有显著性差异，通过比较 p 值，来评定可信度高的组别。

5.1.1 数据预处理

由于给出的数据是自行测得的，可能会存在操作失误进而测错或者漏测、少测的情况发生。而且，经过对数据的浏览分析，我们发现有些数据存在缺失，并且部分数据存在异常。因此，需要进行异常数据的修正和缺失数据的增添，这样才能确保实验数据和实验结果的正确性。

（一）异常数据的修补

附件 1 的数据中，有的数据明显比两侧的数据过大或过小，显然是不合理数据。

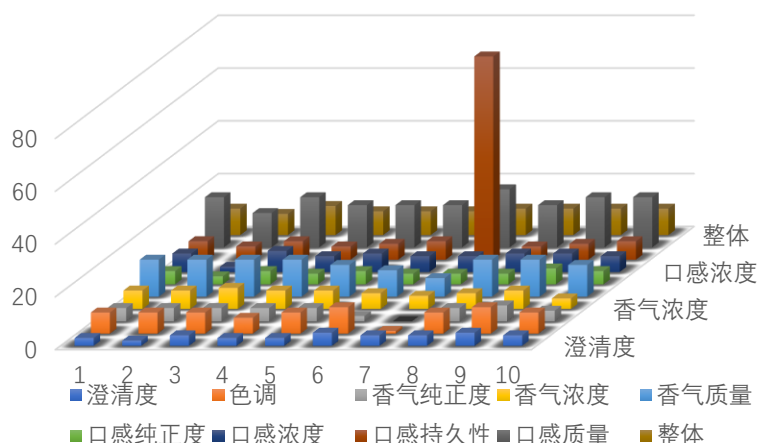


图 2 第一组白葡萄酒样品 3 得分数据分析图

上图 2 是第一组白葡萄酒样品 3 得分数据分析图，可以直观的看到评酒员 7 对样品 3 口感持久性评分的数据，相较于相邻各评酒员的评分发生了明显的突变现象。这种数据异常有可能对数据挖掘的结果产生不利影响。因此，需要对该异常数据进行修补。

而且，评酒员的评分情况是针对单个葡萄酒样品的某一特征的专一评分，葡萄酒各项指标的打分情况个体差异性较大，可不同评酒员对同一种样本的某一特征的打分情况则相似度较高，差异不大。因此，我们采用均值替换法，求取除去异常数据之外其余数据的平均值，作为替补值对异常数据进行修补，具体公式如下所示：

$$x_i = \frac{1}{9} \sum_{k=1, k \neq i}^9 x_k, \quad (i=1,2,\dots,10) \quad (1-1)$$

其中， x_i 为数据异常值， i 为数据异常的评酒员标号。

在对上述数据进行修补之后，修改后的数据如下表 2 所示。

表 2 第一组白葡萄酒样品 3 持久性数据修补表

品酒员	1	2	3	4	5	6	7	8	9	10
增添后	7	5	7	5	6	7	7	5	6	7

(二) 缺失数据的增添

附件 1 的数据中，有的数据明显缺失，显然是不合理的。

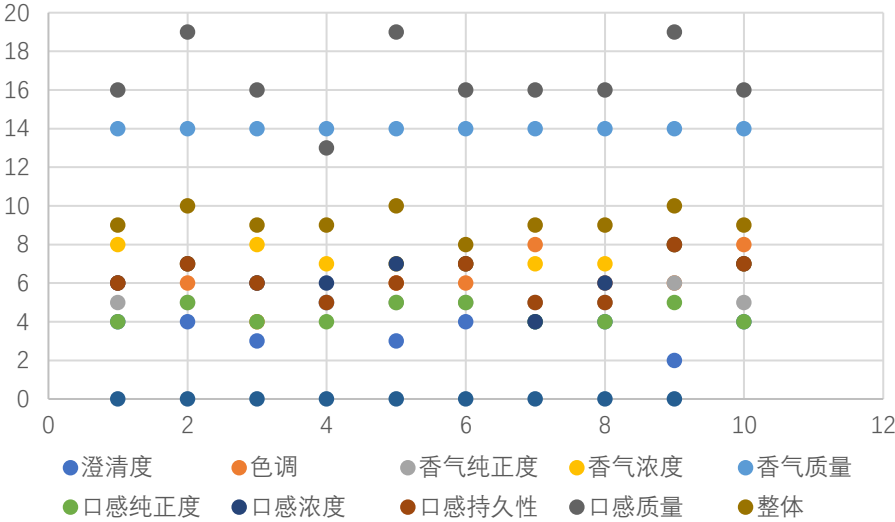


图 3 第二组红葡萄酒样品 20 得分数据散点图

上图 3 是第二组红葡萄酒样品 20 得分数据散点图，可以直观的看到评酒员 4 对样品 20 的评分数据中，没有色调特征的数据，所以存在数据缺失情况。这种数据缺失可能会对后续模型建立、问题分析产生不利影响。因此，需要对该缺失数据进行修补。

对于数据中存在的缺失现象，我们采用和解决数据异常相同的方法——均值替换法，对缺失数据进行处理，具体公式如下所示：

$$x_i = \frac{1}{9} \sum_{k=1, k \neq i}^9 x_k, \quad (i = 1, 2, \dots, 10) \tag{1-2}$$

其中， x_i 为缺失值， i 为数据异常的评酒员标号。

对缺失数据增添后，第一组红葡萄酒样品 20 中色调的数据增添情况如下表 3 所示。

表 2 第一组红葡萄酒样品 20 色调 10 的数据增添表

品酒员	1	2	3	4	5	6	7	8	9	10
增添前	6	6	4	—	6	6	8	6	6	8
增添后	6	6	4	6	6	6	8	6	6	8

其中，— 代表数据缺失。

5.1.2 评分的正态分布检验

对两组评酒员评分的差异性评价分析要求样本数据满足正态分布。因此，我们首先对样本数据的正态分布进行检验。此处我们对两组评分的均值数据进行检验，查看是否符合正态分布。

主要步骤如下所示：

Step1: 统计每个酒样品评酒员的评分均值

首先，计算每一个酒样品的 10 名评酒员的评分均值，均值的公式如下所示：

$$x_m = \frac{1}{10} \sum_{k=1}^{10} x_k, (i = 1, 2, \dots, 27) \quad (1-3)$$

其中， x_m 为缺失值， m 为酒样品的组号。

得到第一组红葡萄酒样品、第一组白葡萄酒样品、第二组红葡萄酒样品、第二组白葡萄酒样品的评酒员打分均值情况。绘制成表格，第一组红葡萄酒样品的部分数据如下表 3（四张表完整版见附录 8.1.1 表 19-22）所示：

表 3 第一组红葡萄酒样品的部分数据汇总表

酒样品	评酒员										均值
	1	2	3	4	5	6	7	8	9	10	
1	51	66	49	54	77	61	72	61	74	62	62.7
2	71	81	86	74	91	80	83	79	85	73	80.3
3	80	85	89	76	69	89	73	83	84	76	80.4
4	52	64	65	66	58	82	76	63	83	77	68.6
5	74	74	72	62	84	63	68	84	81	71	73.3
6	72	69	71	61	82	69	69	64	81	84	72.2
7	63	70	76	64	59	84	72	59	84	84	71.5
8	64	76	65	65	76	72	69	85	75	76	72.3
9	77	78	76	82	85	90	76	92	80	79	81.5
10	67	82	83	68	75	73	75	68	76	75	74.2
11	73	60	72	63	63	71	70	66	90	73	70.1
12	54	42	40	55	53	60	47	61	58	69	53.9
13	69	84	79	59	73	77	77	76	75	77	74.6
14	70	77	70	70	80	59	76	76	76	76	73

Step2: 进行 Shapiro-wilk 检验

由题意得，第一组红葡萄酒、第一组白葡萄酒、第二组红葡萄酒和第二组白葡萄酒样品的酒样品数目分别为 27、28、27、28。所以，采用适用于小范围的 Shapiro-wilk 检验对样本进行检验，并绘制 Q-Q 图。

假设 H_0 表示该组均值服从正态分布， H_1 表示该组均值不服从正态分布。

使用 SPSS 软件，选取置信水平为 95%，对四组数据的均值 Shapiro-wilk 检验。以第一组白葡萄酒为例，它的 Shapiro-wilk 检验、直方图和 Q-Q 图结果分别如下表 4 和图 4、图 5 所示。

表 4 第一组白葡萄酒评分均值的 Shapiro-wilk 检验表

	统计	自由度	p 值
第一组白葡萄酒均值	0.979	28	0.824

经测量, $p = 0.824 > 0.05$, 则不能拒绝 H_0 假设, 即可以认为第一组白葡萄酒均值服从正态分布。

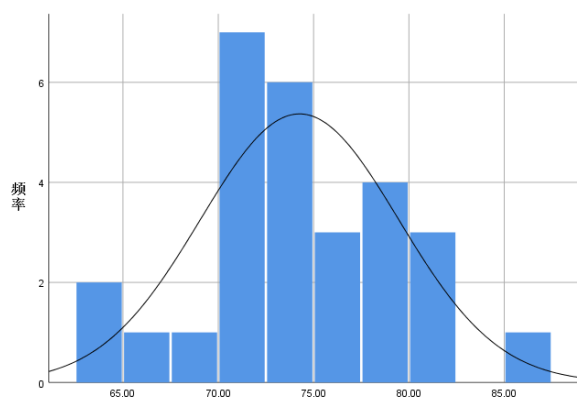


图4 第一组白葡萄酒均值的直方图

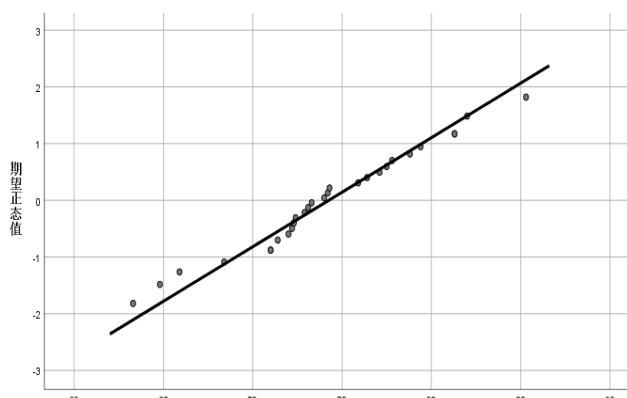


图5 第一组白葡萄酒均值的正态 Q-Q 图

观察上述图 4, 可发现第一组白葡萄酒的均值可以近似的看成正态分布。观察图 5, 发现均值点呈直线散列分布, 即可以近似看成一条直线, 即该组均值符合正态分布。

综上所述, 经过 Shapiro-wilk 检验、直方图和 Q-Q 图的重重验证, 我们可以近似的认为第一组白葡萄酒的评分均值符合正态分布。

经测量, 第一组红葡萄酒、第二组红葡萄酒样和第二组白葡萄酒样品的评分均值也符合正态分布。(相关表格见附件 8.1.2 表 23-25, 图 9-14)

5.1.3 显著差异检验与可信度评估

经 Shapiro-wilk 检验、直方图和 Q-Q 图验证, 得第一组、第二组的红、白葡萄酒评分均值情况服从正态分布。同时, 为了说明评酒员评分的科学性以及两组评分的可信度, 我们检查两组给出的评分均值是否有显著性差异, 即对数据进行显著性检验。

(一) T 检验模型的建立

T 检验是用于小样本, 总体标准差 σ 未知的正态分布总体的两个平均值差异程度的检验方法。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。因此, 可以采用 T 检验来判定两个组别的红、白葡萄酒的评分均值是否存在显著差异。

T 检验的基本理论如下所示:

Step1: 进行假设

设 X_1, X_2, \dots, X_{n1} 是来自总体 $N(\mu_1, \sigma_1^2)$ 的第一组酒样品评分均值, Y_1, Y_2, \dots, Y_{n2} 是来自总体 $N(\mu_2, \sigma_2^2)$ 的第二组酒样品评分均值, 且两组酒样品的评分均值相互独立。为检验两个总体的均值是否有显著差异, 给出假设:

$$H_0: \mu_1 - \mu_2 = 0 \quad (1-4)$$

$$H_1: \mu_1 - \mu_2 \neq 0 \quad (1-5)$$

其中, H_0 表示两组评酒员对酒样品的评价结果无显著性关系; H_1 表示两组评酒员对酒样品的评价结果有显著性关系。

Step2: 计算统计量

当方差未知但相等时, 所用统计量为:

$$T = \frac{\bar{X} - \bar{Y}}{S_{\omega} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} : t(n_1 + n_2 - 2) \quad (1-6)$$

其中,

$$S_{\omega}^2 = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2] \\ = \frac{1}{n_1 + n_2 - 2} \left[\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2 \right] \quad (1-7)$$

Step3: 计算拒绝域

$$t > t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \quad (1-8)$$

$$P\left\{ |T| > t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2) \right\} = 1 - \alpha \quad (1-9)$$

(二) 进行 T 检验

(1) 红葡萄可信度的评定

使用SPSS软件进行独立样本的T检验, 得到的T检验分析结果如表5、表6所示:

表5 第一组和第二组红葡萄酒的组统计表

	组数	个案数	平均值	标准差	标准误差平均值
均值	1.00	27	73.6148	6.76255	1.30145
	2.00	27	70.5148	3.97799	0.76556

表6 第一组和第二组红葡萄酒的独立样本检验表

莱文方差等同性检验					平均值等同性 t 检验			
	F	显著性	t	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
假定等 方差	2.245	0.140	2.053	0.245	3.10000	1.50992	0.0701	6.1298
不假定 等方差			2.053	0.246	3.10000	1.50992	0.0530	6.1470

分析实验结果, 发现莱文方差等同性检验结果为 $F = 2.245$, 显著性概率为 $0.245 > 5\%$, 因此, 接收假设 H_0 , 认为两组评酒员对酒样品的评价结果无显著性关系。

而且, 可以看出第一组数据的标准差 $6.76255 >$ 第二组的标准差 3.97799 , 标准误差平均值第一组 $1.30145 >$ 第二组 0.76556 , 因此, 第二组的评分更加稳定, 数据的波动小。所以, 针对红葡萄酒而言, 可认为第二组的评价结果更可信。

(2) 白葡萄可信度的评定

使用SPSS软件进行独立样本的T检验, 得到的T检验分析结果如表7、表8所示:

表7 第一组和第二组白葡萄酒的组统计表

	组数	平均值	标准差	标准误差平均值
均值	1.00	74.2607	5.20123	0.98294
	2.00	76.5321	3.17094	0.59925

表8 第一组和第二组白葡萄酒的独立样本检验结果表

莱文检验		平均值等同性 t 检验						
	F	显著性	t	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
假定等 均 方差	5.044	.029	-1.973	0.044	-2.27143	1.15121	-4.57946	0.03660
值 不假定 等 方差			-1.973	0.045	-2.27143	1.15121	-4.59060	0.04774

分析实验结果，发现莱文方差等同性检验结果为 $F = 5.044$ ，显著性概率为 $0.044 < 5\%$ ，因此，接收假设 H_1 ，认为两组评酒员对酒样品的评价结果有显著性关系。

而且，可以看出第一组数据的标准差 $5.20123 >$ 第二组的标准差 3.17094 ，标准误差平均值第一组 $0.98294 >$ 第二组 0.59925 ，因此，第二组的评分更加稳定，数据的波动小。所以，针对白葡萄酒而言，可认为第二组的评价结果更可信。

综上所述，两组评酒员对于红葡萄酒的评价没有显著差异，对白葡萄酒的评价有显著差异，第二组的结果更可信。因此，本文后续分析中，对于红葡萄酒和白葡萄酒质量评价结果均选用第二组评酒员评分。

5.2 任务 2 的模型建立与求解

任务 2 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，关键是确定评价酿酒质量的主要指标，建立综合评价模型。首先，根据资料筛选附表中的数据，确定对酿酒葡萄影响较大的理化指标，由于第二组的评价结果更可信，选取第二组的评价结果作为葡萄酒的质量；然后，将数据进行标准化处理，采用主成分分析法，找出主成分，以简化评价指标；最后，以所得的主成分作为新的评价指标，建立对酿酒葡萄分级的综合评价模型，对酿酒葡萄的优劣进行分级。

5.2.1 评价指标简化

由于附表 2 的数据量过大，对数据进行初步的筛选便十分必要。经过文献阅读和相关调查，我们发现酿酒葡萄中的氨基酸、蛋白质、花色苷、有机酸、酚类、醇类、还原糖、果穗、出汁率、多酚氧化酶活力、DPPH 自由基、可溶性固形物这 12 个指标是对酿酒葡萄影响较大的理化指标。其次，由于第二组的评价结果更为可信，选用第二组的评价结果作为酿酒葡萄的葡萄酒的质量指标。综上，初步确定了酿酒葡萄的 13 个指标。

由于选取的指标数 13 仍旧很多，且主观性较强，容易使它们提供的整体信息发生重叠，不易得出简明的规律。主成分分析能将多个指标转换为少数几个主成分，且筛选出的主成分彼此之间互不相关，能反映出原始数据的大部分信息。因此，我们采用主成分分析可以很好的实现降维，筛选出主成分，构建评价模型的指标体系，进而实现分级。

5.2.2 主成分分析及综合评价模型的建立

以红葡萄酒为例进行主成分分析。

Step1: 对13个指标的观测值进行标准化

令 $x = (x_1, x_2, \dots, x_{13})^T$ ，建立数据标准化模型：

$$X = \frac{x - \bar{x}}{\sigma} \quad (1-10)$$

利用MATLAB命令 $x = \text{zscore}(x)$ ，得到标准化矩阵。

Step2: 建立各个指标间的相关系数矩阵

利用MATLAB命令 $R = \text{corrcoef}(X)$ 加以实现。

Step3: 计算相关系数矩阵R的特征根、特征向量及特征根的贡献率

利用MATLAB命令 $[\text{vec}, \text{val}] = \text{eig}(R)$ ，求取特征根和贡献率。

Step4: 计算累计贡献率

取前7个主程序的贡献率相加，与85%进行比较，符合则选取这7个部分为主成分。

Step5: 建立综合评价模型

计算各样品的每个主成分值，将每个样品的所有主成分值加到一起即为该样品的综合得分。建立样品关于7个主成分的主成分值的评价模型，为：

$$W = XU = (w_{ij})_{27 \times 7} \quad (1-11)$$

$$w_i = \sum_{j=1}^7 w_{ij}, i = 1, 2, \dots, 27 \quad (1-12)$$

其中， X 为27个酿酒葡萄(红)的标准化结果矩阵， U 为7个主成分的各个指标的载荷， w_i 为第*i*个样品的综合得分。

5.2.3 评价系统求解

使用 Matlab 软件进行求解，具体步骤如下所示：

以红葡萄酒为例求解。

Step1: 对13个指标的观测值进行标准化

令 $x = (x_1, x_2, \dots, x_{13})^T$ ，建立数据标准化模型：

$$X = \frac{x - \bar{x}}{\sigma} \quad (1-13)$$

利用MATLAB命令 $x = \text{zscore}(x)$ ，得到标准化矩阵。

Step2: 建立各个指标间的相关系数矩阵

利用MATLAB命令 $R = \text{corrcoef}(X)$ 加以实现。得到相关系数矩阵如下表9所示。颜色代表的深浅，反应了相关系数的大小情况，越小蓝色越深，越大红色越深。

表9 红葡萄酒的相关系数矩阵表

1.00 0	0.02 5	0.08 1	0.36 4	0.26 6	0.46 5	0.43 6	-0.3 83	-0.0 09	-0.1 18	0.10 6	0.55 3	0.14 9
0.02 5	1.00 0	0.42 8	0.47 0	0.54 1	0.39 5	-0.2 95	0.06 8	0.42 1	0.11 0	0.74 5	0.02 6	0.52 3
0.08 1	0.42 8	1.00 0	0.09 3	0.71 3	0.35 3	-0.0 13	-0.1 17	0.52 3	0.40 0	0.65 5	0.05 3	0.27 3
0.36 4	0.47 0	0.09 3	1.00 0	0.27 5	0.14 3	0.11 9	-0.0 25	0.08 4	-0.1 28	0.27 6	0.22 1	0.28 8

0.26 6	0.54 1	0.71 3	0.27 5	1.00 0	0.33 7	-0.0 48	-0.2 47	0.51 4	0.07 4	0.85 3	0.30 3	0.65 2
0.46 5	0.39 5	0.35 3	0.14 3	0.33 7	1.00 0	0.15 1	-0.2 01	0.04 5	0.28 9	0.42 8	0.15 6	0.26 4
0.43 6	-0.2 95	-0.0 13	0.11 9	-0.0 48	0.15 1	1.00 0	-0.2 52	0.02 5	0.06 4	-0.2 87	0.44 8	-0.0 53
-0.3 83	0.06 8	-0.1 17	-0.0 25	-0.2 47	-0.2 01	-0.2 52	1.00 0	0.08 5	0.02 3	0.00 6	-0.3 17	0.20 0
-0.0 09	0.42 1	0.52 3	0.08 4	0.51 4	0.04 5	0.02 5	0.08 5	1.00 0	0.04 2	0.46 5	0.14 1	0.31 2
-0.1 18	0.11 0	0.40 0	-0.1 28	0.07 4	0.28 9	0.06 4	0.02 3	0.04 2	1.00 0	0.09 6	0.17 5	-0.2 30
0.10 6	0.74 5	0.65 5	0.27 6	0.85 3	0.42 8	-0.2 87	0.00 6	0.46 5	0.09 6	1.00 0	0.03 5	0.60 5
0.55 3	0.02 6	0.05 3	0.22 1	0.30 3	0.15 6	0.44 8	-0.3 17	0.14 1	0.17 5	0.03 5	1.00 0	0.02 5
0.14 9	0.52 3	0.27 3	0.28 8	0.65 2	0.26 4	-0.0 53	0.20 0	0.31 2	-0.2 30	0.60 5	0.02 5	1.00 0

Step3: 计算相关系数矩阵R的特征根、特征向量及特征根的贡献率

利用MATLAB命令 [vec, val]=eig(R), 求取特征根和贡献率。经过汇总, 红葡萄酒的特征向量、特征根、贡献率等数据如下表10所示。

表10 红葡萄酒相关数据汇总表

特征 向量	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13
x1	0.16	-0.4	0.20	0.12	-0.0	-0.1	0.02	0.22	0.54	-0.2	0.35	-0.3	-0.0
	0338	9988	0924	6056	4207	4079	4391	819	5537	7553	7473	0141	4169
x2	0.37	0.18	0.10	0.23	0.14	0.27	-0.0	0.27	-0.4	0.09	0.59	0.02	-0.2
	4739	6969	0607	5679	9317	2036	3564	5825	099	0478	4416	5775	3474
x3	0.35	0.05	-0.3	-0.1	-0.0	-0.0	-0.3	-0.2	0.38	0.08	0.29	0.51	0.17
	6685	8481	8374	3435	7483	0313	0318	6465	4787	8674	7519	2176	2385
x4	0.20	-0.1	0.42	0.18	0.36	0.52	-0.3	-0.2	0.12	-0.0	-0.3	0.07	0.07
	9015	3398	1216	407	8107	035	253	4631	1897	968	4595	2893	1271
x5	0.43	-0.0	-0.0	-0.1	-0.2	-0.0	0.18	-0.2	0.05	0.02	-0.2	-0.1	-0.7
	9385	1104	1518	9221	37	0134	9739	911	5901	1433	3394	056	2337
x6	0.26	-0.1	-0.1	0.55	-0.0	-0.4	-0.1	0.33	-0.1	0.09	-0.4	0.27	-0.0
	4129	9789	5687	8414	0734	0195	323	396	0813	4205	0234	8376	6366
x7	-0.0	-0.4	-0.0	-0.2	0.29	-0.3	-0.4	-0.2	-0.3	0.30	0.13	-0.2	-0.0
	1818	8069	2848	5971	7058	1937	0079	6259	479	8461	7085	04	6259
x8	-0.0	0.38	0.11	-0.0	0.67	-0.3	0.15	0.03	0.37	0.23	0.01	0.02	-0.2
	5768	6769	837	121	1696	3567	6571	9464	8345	6341	7314	5296	0752
x9	0.28	0.11	-0.0	-0.5	0.14	0.01	-0.2	0.56	-0.0	-0.2	-0.2	-0.0	0.06
	1305	5782	9749	9067	1641	4112	5957	6369	3615	6567	3855	8746	491
x10	0.07	-0.0	-0.6	0.21	0.38	0.12	0.16	-0.2	-0.0	-0.4	-0.0	-0.3	0.03
	5428	3425	6752	4709	9742	5046	7515	0061	7497	0144	2397	1145	5574

x11	0.43	0.17	-0.0	0.07	-0.1	0.00	0.10	-0.0	0.08	0.48	-0.1	-0.5	0.46
	4985	4565	0484	2306	3516	0916	8453	4041	2668	5893	0929	2282	3884
x12	0.12	-0.4	-0.0	-0.2	0.21	0.20	0.63	0.12	-0.0	0.22	-0.0	0.33	0.15
	9926	6067	2773	3868	8192	4675	9903	8734	6748	7042	4038	8393	93
x13	0.32	0.12	0.35	-0.0	0.03	-0.4	0.21	-0.3	-0.2	-0.4	0.06	0.13	0.29
	4893	8813	9557	5546	6803	4881	2279	0845	7229	6793	2702	2092	6681
特征	4.26	2.41	1.53	1.07	0.95	0.78	0.58	0.49	0.40	0.18	0.15	0.13	0.01
值	269	2766	584	7871	4473	9314	4123	1326	0458	5425	2745	5587	7382
贡献	0.32	0.18	0.11	0.08	0.07	0.06	0.04	0.03	0.03	0.01	0.01	0.01	0.00
率	7899	5597	8142	2913	3421	0716	4933	7794	0804	4263	175	043	1337
累计	0.32	0.51	0.63	0.71	0.78	0.84	0.89	0.93	0.96	0.97	0.98	0.99	
贡献	7899	3497	1638	4551	7972	8689	3621	1416	222	6484	8233	8663	1
率													

其中, x_1 、 x_2 、 x_3 、 x_4 、 x_5 、 x_6 、 x_7 、 x_8 、 x_9 、 x_{10} 、 x_{11} 、 x_{12} 、 x_{13} 分别表示氨基酸、蛋白质、花色苷、有机酸、酚类、醇类、还原糖、果穗、出汁率、多酚氧化酶活力、DPPH 自由基、可溶性固形物及第二组评价结果。

Step4: 计算累计贡献率, 得出主成分

根据上述得出的数据可得, 前7个主成分的累计贡献率达89.3621%, 因此, 我们考虑只取前面7个能够很好地概括原始变量的主成分。

$$F_1 = 0.160338x_1 + 0.374739x_2 + 0.356685x_3 + 0.209015x_4 + 0.439385x_5 \\ \textcircled{1} \quad + 0.264129x_6 - 0.01818x_7 - 0.05768x_8 + 0.281305x_9 + 0.075428x_{10} \\ + 0.434985x_{11} + 0.129926x_{12} + 0.324893x_{13}$$

$$F_2 = -0.49988x_1 + 0.186969x_2 + 0.058481x_3 - 0.13398x_4 - 0.01104x_5 \\ \textcircled{2} \quad - 0.19789x_6 - 0.48069x_7 + 0.386769x_8 + 0.115782x_9 - 0.03425x_{10} \\ + 0.174565x_{11} - 0.46067x_{12} + 0.128813x_{13}$$

$$F_3 = 0.200924x_1 + 0.100607x_2 - 0.38374x_3 + 0.421216x_4 - 0.01518x_5 \\ \textcircled{3} \quad - 0.15687x_6 - 0.02848x_7 + 0.11837x_8 - 0.09749x_9 - 0.66752x_{10} \\ - 0.00484x_{11} - 0.02773x_{12} + 0.359557x_{13}$$

$$F_4 = 0.126056x_1 + 0.235679x_2 - 0.13435x_3 + 0.18407x_4 - 0.19221x_5 \\ \textcircled{4} \quad + 0.558414x_6 - 0.25971x_7 - 0.0121x_8 - 0.59067x_9 + 0.214709x_{10} \\ + 0.072306x_{11} - 0.23868x_{12} - 0.05546x_{13}$$

$$F_5 = -0.04207x_1 + 0.149317x_2 - 0.07483x_3 + 0.368107x_4 - 0.237x_5 \\ \textcircled{5} \quad - 0.00734x_6 + 0.297058x_7 + 0.671696x_8 + 0.141641x_9 + 0.389742x_{10} \\ - 0.13516x_{11} + 0.218192x_{12} + 0.036803x_{13}$$

$$F_6 = -0.14079x_1 + 0.272036x_2 - 0.00313x_3 + 0.52035x_4 - 0.00134x_5 \\ \textcircled{6} \quad - 0.40195x_6 - 0.31937x_7 - 0.33567x_8 + 0.014112x_9 + 0.125046x_{10} \\ + 0.000916x_{11} + 0.204675x_{12} - 0.44881x_{13}$$

$$F_7 = 0.024391x_1 - 0.03564x_2 - 0.30318x_3 - 0.3253x_4 + 0.189739x_5 \\ \textcircled{7} - 0.1323x_6 - 0.40079x_7 + 0.156571x_8 - 0.25957x_9 + 0.167515x_{10} \\ + 0.108453x_{11} + 0.639903x_{12} + 0.212279x_{13}$$

经分析，各主成分的代表部分如下所示：

① 在第一个主成分的表达式中，第2、3、5、11、13项的系数比较大，也即这5项指标对酿酒红葡萄酒质量的影响较大。

② 在第二个主成分的表达式中，第8、2、11项的系数比较大，也即果穗、蛋白质、DPPH自由基对酿酒红葡萄酒质量的影响较大。

③ 在第三个主成分的表达式中，第4、13项的系数比较大，也即有机酸和第二组评价结果对酿酒红葡萄酒质量的影响较大。

④ 在第四个主成分的表达式中，第6、2、10项的系数比较大，也即醇类、蛋白质和多酚氧化酶活力对酿酒红葡萄酒质量的影响较大。

⑤ 在第五个主成分的表达式中，第8、4、10项的系数比较大，也即果穗、有机酸和多酚氧化酶活力对酿酒红葡萄酒质量的影响较大。

⑥ 在第六个主成分的表达式中，第2、12项的系数比较大，也即蛋白质和DPPH自由基对酿酒红葡萄酒质量的影响较大。

⑦ 在第七个主成分的表达式中，第12、13、5、10项的系数比较大，也即DPPH自由基、第二组评价结果、酚类和多酚氧化酶活力对酿酒红葡萄酒质量的影响较大。

绘制成主成分图如下图6所示：



图6 酿酒葡萄质量评价主成分图

由于酿制红葡萄酒和白葡萄酒的酿酒葡萄的理化指标相同，因此用同样的方法建立类似的模型，详情请见附录 8.2.3 表 26-27。

Step5: 计算综合得分

计算酿酒葡萄（红、白）的综合得分见下表 11。

表 11 酿酒葡萄（红、白）的综合得分表

样品序号	酿酒葡萄（红）	酿酒葡萄（白）
1	-2.225	-4.1133
2	6.7132	-1.9846
3	1.3257	6.6028
4	-4.5751	-1.1605
5	6.8865	1.7796
6	-4.5472	3.3718
7	-11.0391	0.0148
8	0.9562	-9.7736
9	16.4566	1.5913
10	-0.9663	1.2441
11	-3.8886	-4.1178
12	-6.4915	3.1442
13	1.8892	-1.1444
14	1.3634	-3.3251
15	-3.9208	5.563
16	1.1547	-11.2427
17	0.7015	-1.4854
18	-8.1752	-3.0263
19	0.2676	-5.4498
20	-0.3751	0.7967
21	3.1488	-3.1586
22	-3.7008	0.2626
23	4.4183	1.2363
24	-0.069	6.4794
25	-0.1012	0.0616
26	6.0829	-0.5375
27	-1.2898	13.258
28		5.1135

利用 EXCEL 作出酿酒红、白葡萄的综合得分条形图如下图图 7、图 8 所示：

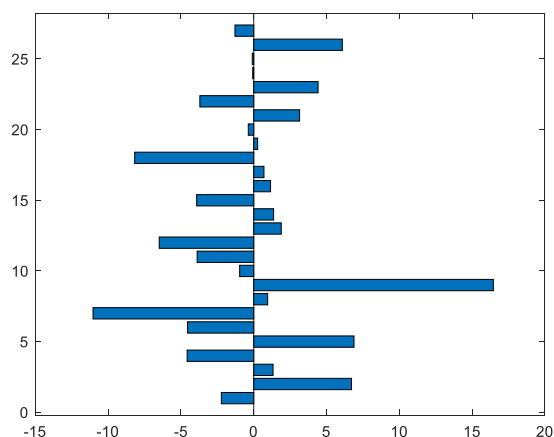


图7 酿酒红葡萄综合得分图

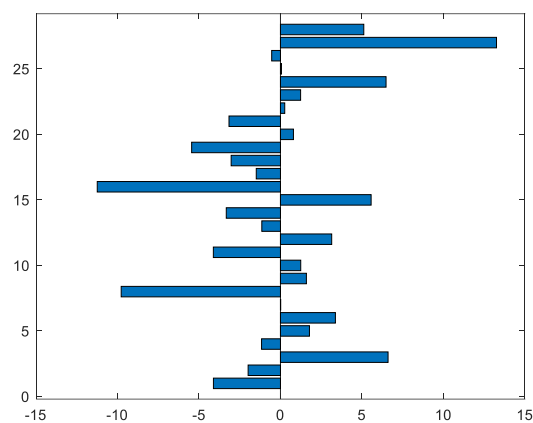


图8 酿酒白葡萄综合得分图

分析图7和图8，发现大部分酿酒葡萄的质量落在(-5, 5)之间，只有少部分葡萄在(-10, -5)和(5, 15)之间。这也验证了现实中酿酒葡萄的种植情况，极好的葡萄由于难以种植，因此数量较少，极差的葡萄由于某些葡萄酒的理化指标需要又必须种植，一般的葡萄由于成本和品质同时可以满足需求，故会大量种植。

综上所述，可以将酿酒葡萄分为四个等级：优(5, 15)、良(0, 5)、中(-5, 0)、差(-10, -5)、对酿酒葡萄的得分进行分级，即得酿酒葡萄的分级如表12所示。

表12 酿酒葡萄的分级

级别	酿酒葡萄（红）	酿酒葡萄（白）
优	2,5,9,26	3,15,24,27,28
良	3,8,13,14,16,17,19,21,23,	5,6,7,9,10,12,20,22,23,25
中	1,4,6,10,11,15,20,22,24,25,27	1,2,4,11,13,14,17,18,21,26
差	7,12,18	8,16,19

结合以上图表可以得到：

(1) 27个酿酒红葡萄样品中品质最优的为样本2, 5, 9, 26，品质最劣的为样本7, 12, 18；28个酿酒白葡萄样品中品质最优的为样本3, 15, 24, 27, 28，品质最劣的为样8, 16, 19。

(2) 最优的要求相对较高，而越低级别的酿酒葡萄对各项指标远离最优的要求也相对较高，因此，要求越高，达到标准的样本数越少。

5.3 任务3的模型建立与求解

任务3要求分析酿酒葡萄与葡萄酒的理化指标之间的联系。其中，酿酒葡萄和葡萄酒的理化指标是两个包含多项数据指标变量的组别。求解任务3的策略：首先，建立典型相关分析模型，分析酿酒葡萄与葡萄酒的理化指标之间的典型相关关系；然后，根据上述分析给出酿酒葡萄与葡萄酒的理化指标之间的联系。

5.3.1 典型相关分析模型的建立

酿酒葡萄和葡萄酒的理化指标是两个包含多项数据指标变量的组别，如果直接对这些变量计算 spearman 相关系数或者 person 相关系数进行两两分析，过程复杂，耗时较长，且很难得到关于这两组变量之间关系的一个明确清晰的结果。所以，需要改进模型，采用其他方法进行求解。

同时，典型相关分析可以把多个变量与多个变量之间的相关化为两个具有代表性的变量之间的相关，是研究两组变量之间相关关系的一种多元统计方法，能够揭示出两组变量之间的内在联系。因此，采用典型相关分析的方法可以很好的研究酿酒葡萄和葡萄酒的理化指标之间的相关性。令酿酒葡萄为输入变量，葡萄酒的理化指标为输出变量。

主要步骤如下所示：

① 建立原始矩阵

根据表格中原有数据，我们设酿酒葡萄的理化指标记为 $\mathbf{X} = (X_1, X_2, \dots, X_{55})'$ ，葡萄酒的理化指标记为 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_9)'$ ， \mathbf{Z} 为 12+9 总体观测数矩阵：

$$\mathbf{Z} = \begin{bmatrix} X_{1,1} & \cdots & X_{1,12} & Y_{1,1} & \cdots & Y_{1,9} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{55,1} & \cdots & X_{55,12} & Y_{55,1} & \cdots & Y_{55,9} \end{bmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \end{pmatrix}'$$

② 对原始数据进行标准化变换并计算相关系数矩阵

我们利用问题二中公式对酿酒葡萄和葡萄酒的理化指标数据进行标准化处理，然后计算两样本间的相关系数矩阵 \mathbf{R} ，并将 \mathbf{R} 分为

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}$$

其中， \mathbf{R}_{11} 、 \mathbf{R}_{22} 分别为酿酒葡萄和葡萄酒指标内的相关系数阵， \mathbf{R}_{12} 、 \mathbf{R}_{21} 为酿酒葡萄指标与葡萄酒指标间的相关系数阵。

③ 求典型相关系数及典型变量

首先求 $\mathbf{A} = \mathbf{R}_{11}^{-1} \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}$ 的特征根 λ_i^2 ，特征向量 $\mathbf{S}_1 \alpha_i$ ； $\mathbf{B} = \mathbf{R}_{22}^{-1} \mathbf{R}_{21} \mathbf{R}_{11}^{-1} \mathbf{R}_{12}$ 的特征根 λ_i^2 ，特征向量 $\mathbf{S}_2 \beta_i$ ，则有

$$\alpha_i = \mathbf{S}_1^{-1}(\mathbf{S}_1 \alpha_i), \beta_i = \mathbf{S}_2^{-1}(\mathbf{S}_2 \beta_i)$$

则随机变量酿酒葡萄的理化指标 \mathbf{X} 和葡萄酒的理化指标 \mathbf{Y} 的典型相关系数为 λ ，典型变量为

$$\begin{aligned} V_1 &= \alpha_1' \mathbf{X}; & V_2 &= \alpha_2' \mathbf{X}; & \dots & V_i &= \alpha_i' \mathbf{X} \\ W_1 &= \beta_1' \mathbf{Y}; & W_2 &= \beta_2' \mathbf{Y}; & \dots & W_i &= \beta_i' \mathbf{Y} \end{aligned} \quad (i \leq 55)$$

④ 检验各典型相关系数的显著性

对典型相关系数 λ_i 进行显著性检验。在作两组变量酿酒葡萄的理化指标 \mathbf{X} 和葡萄酒的理化指标 \mathbf{Y} 的典型相关分析之前，首先应检验两组变量是否相关；如果不相关，即 $\text{cov}(\mathbf{X}, \mathbf{Y})=0$ ，则讨论的两组变量的典型相关就毫无意义。

5.3.2 典型相关分析模型的求解

附件 2 中包含 27 个红葡萄样本和 28 个白葡萄样本 2 组指标共 21 个指标的原始数据。其中，13 个是酿酒葡萄的理化指标，如下表所示。

表 12 酿酒葡萄的理想化指标表

x_1	x_2	x_3	x_4
氨基酸总量	蛋白质	花色苷	有机酸
x_5	x_6	x_7	x_8
酚类	醇类	还原糖	果穗

x_9	x_{10}	x_{11}	x_{12}
出汁率	多酚氧化酶活力	DPPH 自由基	可溶性固形物

除此之外，9 个是葡萄酒的理化指标：

$x_{13} \rightarrow$ 花色苷(mg/L)， $x_{14} \rightarrow$ 单宁(mmol/L)， $x_{15} \rightarrow$ 总酚(mmol/L)， $x_{16} \rightarrow$ 酒总黄酮(mmol/L)， $x_{17} \rightarrow$ 白藜芦醇(mg/L)， $x_{18} \rightarrow$ DPPH 半抑制体积 (IV50) 1/IV50(uL)， $x_{19} \rightarrow$ L*(D65)， $x_{20} \rightarrow$ a*(D65)， $x_{21} \rightarrow$ b*(D65)。

利用 SPSS 软件对酿酒葡萄和葡萄酒的理化指标进行了典型相关分析，结果如下表 13 所示：

表 13 典型相关系数表

	相关系数	特征值	威尔克统计	F	分子自由度	分母自由度	p值
1	0.988	41.664	0.000	2.428	108.000	56.202	0.000
2	0.960	11.744	0.000	1.805	88.000	55.362	0.010
3	0.951	9.514	0.002	1.497	70.000	53.464	0.062
4	0.931	6.522	0.019	1.104	54.000	50.485	0.362
5	0.756	1.338	0.141	0.657	40.000	46.383	0.911
6	0.647	0.720	0.330	0.528	28.000	41.083	0.961
7	0.484	0.306	0.568	0.424	18.000	34.426	0.972
8	0.450	0.254	0.741	0.419	10.000	26.000	0.924
9	0.265	0.076	0.930	0.265	4.000	14.000	0.896

利用第1对、第2对、第3对、第4对典型变量之间的典型相关系数均大于90%。即表明这4对典型变量的解释能力较强，相应典型变量之间密切相关。可是观察 p 值，发现第1对、第2对和第3对的 p 值分别为0.000、0.010、0.062，均小于0.1，所以前三对典型变量之间相关关系显著。第四对的 p 值为0.362，大于0.1，因此不符合要求，需要舍去。

标准化后的典型变量的系数来建立典型相关模型

$$\begin{aligned} \textcircled{1} \quad V_1 &= 0.17x_{10} - 0.16x_{19} - 0.18x_{20} - 1.02x_{27} \\ W_1 &= 4.13x_{18} \end{aligned}$$

$$\begin{aligned} \textcircled{2} \quad V_2 &= 0.14x_3 - 0.29x_4 + 0.25x_5 - 0.58x_6 + 0.37x_{11} - 0.47x_{12} + 0.54x_{13} + 0.36x_{20} - 0.35x_{28} \\ W_2 &= -1.90x_{31} + 0.72x_{32} + 0.31x_{33} + 0.53x_{36} - 0.61x_{37} \end{aligned}$$

$$\begin{aligned} \textcircled{3} \quad V_3 &= 0.26x_2 + 0.36x_4 + 0.25x_9 + 0.56x_{10} + 0.55x_{11} + 0.56x_{13} \\ &\quad + 0.30x_{19} - 0.28x_{20} + 0.28x_{21} + 0.44x_{26} \\ W_3 &= 0.27x_{31} + 0.24x_{32} + 0.30x_{33} + 0.58x_{34} + 0.2x_{35} + 0.43x_{36} \\ &\quad - 0.35x_{37} - 0.19x_{38} \end{aligned}$$

根据典型变量重要程度及系数大小，从建立的典型相关模型可以看出，葡萄酒各指标受酿酒葡萄各指标变动的作用程度可用三对典型相关变量予以综合描述。

第一对，典型变量主要将DPPH半抑制体积从各种酿酒葡萄指标中分离出来（典型载荷为4.13），与果皮质量呈现最大相关（相应典型载荷为-1.02）。由此可见，葡萄酒中的DPPH半抑制体积主要来自于葡萄的果皮，同时，酿酒葡萄中的DPPH自由基含量、PH值、可滴定酸含量也对其有一定的影响。因此，增大酿酒葡萄果皮的含量对葡萄酒中DPPH半抑制体积含量的增加有重要影响。

第二对，典型变量将花色苷及单宁从9个葡萄酒指标中分离出来（典型载荷为-1.90和0.72），酿酒葡萄指标中与之相对应的解释变量是苹果酸、葡萄总黄酮和单宁（典型载荷为-0.58、0.54和-0.47）。显而易见的，葡萄酒中和酿酒葡萄中的单宁具有较强的相关性，葡萄酒中的花色苷（类黄酮化合物）主要来源于酿酒葡萄中的葡萄总黄酮。值得注意的是，酿酒葡萄中的苹果酸不仅使得葡萄的发酵顺利进行，还保护着对红葡萄酒起主要呈色作用的花色苷和对花色苷起中等辅色作用的单宁物质，使得红葡萄酒呈现漂亮的宝石红色。

第三对，典型变量将酒总黄酮和DPPH半抑制体积从9个葡萄酒指标中分离出来（典型载荷为0.58和0.43），酿酒葡萄指标中与之相对应的解释变量是DPPH自由基、葡萄总黄酮和总酚（典型载荷为0.56、0.56和0.55）。在葡萄总黄酮消除自由基的抗氧化作用和总酚对清除自由基保护的共同作用下，酿酒葡萄中的DPPH自由基转化为葡萄酒中的DPPH半抑制体积。

5.4 任务4的模型建立与求解

任务4要求计算酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证用葡萄和葡萄酒的理化指标来评价葡萄酒的质量是否合适。同时，多元线性回归可以定量的描述某一现象和某些因素之间的函数关系，因此，选用多元线性回归方程来求解任务4。求解任务4的策略：首先，对样本进行随即筛选，选取其中一部分进行分析；其次，为了确保指标之间互不相关，进行初步筛选；然后，利用筛选后的指标和葡萄酒质量的评价结果，建立多元线性回归方程；最后，利用余下的样本的酿酒葡萄和葡萄酒的理化指标，对葡萄酒质量求解得到的多元线性回归方程进行验证。

5.4.1 多元线性回归模型的建立

多元线性回归模型是指含有多个解释变量的线性回归模型，用于解释被解释的变量与其他多个变量解释变量之间的线性关系。同时，自变量酿酒葡萄和葡萄酒的理化指标是多个因素，因变量葡萄酒质量只有一个，所以采用多元线性回归模型十分合适。

主要步骤如下所示：

① 建立多元线性模型

因变量葡萄酒质量受 k 个自变量 x_1, x_2, \dots, x_n 的影响，其 n 组观测值为 $(y_a, x_{1a}, x_{2a}, \dots, x_{ka})$ ， $a = 1, 2, \dots, n$ 。即，建立的多元线性回归模型的结构形式如下所示：

$$y_a = \beta_0 + \beta_1 x_{1a} + \beta_2 x_{2a} + \dots + \beta_k x_{ka} + \varepsilon_a \quad (1-10)$$

其中， $\beta_0, \beta_1, \dots, \beta_k$ 为待定参数， ε_a 为随机变量。

为了方便，引入矩阵

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

其中，X 成为模型设计矩阵，是常数矩阵，Y 与 ε 是随机向量，且

$$Y: N_n(X\beta, \sigma^2 I), \varepsilon: N_n(0, \sigma^2 I) \quad (1-14)$$

(I 为 n 阶单位阵)

ε 是不可观测的随机误差向量， β 是回归系数构成的向量，是未知待定的常数向量。

② 回归系数 β 的最小二乘估计

选取 β 的一个估计值，记为 $\hat{\beta}$ ，使随机误差 ε 的平方和达到最小，即

$$\begin{aligned} \min_{\beta} \varepsilon^T \varepsilon &= \min_{\beta} (Y - X\beta)^T (Y - X\beta) \\ &= \min_{\beta} Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta \end{aligned} \quad (1-15)$$

由最小二乘法的要求，由多元函数取得极值的必要条件可求解回归参数的标准方程如下：

$$\begin{cases} \left. \frac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0} = 0 \\ \left. \frac{\partial Q}{\partial \beta_j} \right|_{\beta_j = \hat{\beta}_j} = 0 \end{cases} \quad (j=1, 2, \dots, p)$$

可以证明任意给定的 X, Y，正规方程组总有解，虽然当 X 不满秩时，其解不唯一，但对任意一组解 $\hat{\beta}$ 都能使残差平方和最小， $Q(\hat{\beta}) = \min_{\beta} Q(\beta)$

特别地，当 X 秩时，即 $r(X) = r(X^T X) = p$ ，则正规方程组的解为 $\hat{\beta} = (X^T X)^{-1} X^T Y$ ，即为回归系数的估计值。

③ 逐步回归分析

由于建立回归模型时，不是每一个因子对 y 的影响程度都很大。从而我们通过逐步回归的方法（剔除优选法）来对因子进行筛选。

5.4.2 多元线性回归模型的求解

主要求解步骤如下所示：

① 利用 MATLAB 软件中的 randperm 函数在 27 个酿酒红葡萄和红葡萄酒的样品中随机抽取 20 个样品。抽取的 20 个样品如下表 14 所示：

表 14 样品抽取情况表

抽取样品详情																			
7	1	2	22	21	12	27	18	5	25	9	8	14	13	23	26	15	20	19	4

② 利用 SPSS 软件对所抽取的样品进行求解，得到酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的线性回归方程系数如下表所示。

表15 线性回归方程系数情况表

模型	未标准化系数		标准化系数		t	显著性
	B	标准错误	Beta			
(常量)	58.155	14.271			4.075	0.005
x1	0.000	0.001	-0.148		-0.474	0.650
x2	0.041	0.028	0.619		1.479	0.183
x3	-0.009	0.012	-0.268		-0.762	0.471
x4	0.053	0.280	0.052		0.189	0.855
x5	0.759	0.267	1.400		2.841	0.025
x6	0.023	0.026	0.232		0.885	0.405
x7	0.029	0.035	0.284		0.816	0.442
x8	0.011	0.004	0.596		2.612	0.035
x9	-0.155	0.155	-0.346		-1.004	0.349
x10	-0.153	0.078	-0.438		-1.966	0.090
x11	-22.215	21.995	-0.717		-1.010	0.346
x12	-0.039	0.049	-0.176		-0.789	0.456

易得酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的十二元线性回归方程:

$$y = 58.155 + 0.041x_2 - 0.009x_3 + 0.053x_4 + 0.759x_5 + 0.023x_6 + 0.029x_7 + 0.011x_8 - 0.153x_9 - 0.153x_{10} - 22.215x_{11} - 0.039x_{12}$$

其中,最后筛选和剔除后剩下评价葡萄酒质量的指标为:氨基酸总量 x_1 、蛋白质 x_2 、花色苷 x_3 、有机酸 x_4 、酚类 x_5 、醇类(mg/kg) x_6 、还原糖 x_7 、果穗 x_8 、出汁率 x_9 、多酚氧化酶活力 x_{10} 、DPPH自由基 x_{11} 、可溶性固形物 x_{12} 。

③ 回归模型的显著性检验

利用 SPSS 软件对求解得到的线性回归模型进行了显著性检验,充分表明酿酒葡萄和葡萄酒的理化指标与葡萄酒质量之间的密切关系,结果如下:

表 16 模型摘要表

模型	R	R 方	调整后 R 方	标准估算的错误
1	0.930 ^a	0.865	0.634	2.07302

通过表 X 可以得到,复相关系数 $R=0.930$,多重判定系数 $R^2=0.865$,调整后 $R^2=0.634$ 。调整后 R^2 的值越大,模型的拟合效果越好。

表 17 ANOVA 表

模型		平方和	自由度	均方	F	显著性
1	回归	192.948	12	16.079	3.742	0.045 ^b
	残差	30.082	7	4.297		
	总计	223.029	19			

由上表可知 $F=3.742$, $F>F_{0.05}(7,20)=2.51$,回归方程显著,即自变量和因变量存在明显的函数关系。

5.4.3 多元线性回归模型的验证

在线性回归模型的基础上,将未选中的7个样本作为检验样本组,对模型的可行性进行验证,样本质量值的评分如下:

表 18 检验样本组 7 个样本的葡萄酒质量值表

红葡萄	样品3	样品6	样品10	样品11	样品15	样品16	样品24
计算值	71.64	65.01	71.90	64.22	73.25	78.10	70.35
实际值	71.2	66	72.6	65.7	71.6	77.1	71.5

计算值与实际值相差不大,所以,求得的回归方程符合实际,酿酒葡萄和葡萄酒的理化指标与葡萄酒的质量之间的关系足够密切。

6 模型评价

6.2 模型实用性分析

该模型采用了 T 检验、F 检验、主成分分析、典型相关分析、多元线性回归等方法,对两组评酒员的评价结果进行判别,检查其是否有显著性差异。进行正态分布检验的时候,采用了 Shapiro-wilk 检验、直方图和 Q-Q 图三种检验方法,有效地避免了人为观测产生的误差。本文运用了 MATLAB、SPSS、EXCEL 等软件,对葡萄酒质量的评价问题进行了多角度的分析,并给出了利用理化指标评价葡萄酒质量的模型,具有很大的推广价值。

6.3 模型推广

针对从网络上收集到的数据,该模型求出的结果极大程度上与数据相吻合。本文分析了多种影响葡萄酒质量的因素,并给出了各个因素对葡萄酒质量影响的比重,比重越大,影响程度越深。这一模型可以很好的用于葡萄酒质量的判断,在稍微修改之后,运用同样的方法建模,可以将模型更好的运用于数量预测、质量评定等其他多个方面。

7 参考文献

- [1] 李叶凤,王圣仪,谭丹,刘亚新,齐鹏宇,刘敦华,张昂. 基于挥发性成分分析的葡萄酒产地及品种鉴别研究进展[J/OL]. 食品科学:1-13[2021-06-20]
- [2] 李猛. 红葡萄酒品质特点及鉴别方法分析[J]. 中国果菜, 2020, 40 (09) :65-68.
- [3] 田伟业,杨和财,张军翔,李甲贵,房玉林,陶永胜. 中外葡萄酒产品等级分级评价及启示[J]. 中国酿造, 2020, 39 (10) :220-224.
- [4] 陈虹瑶,杨易,季俊霖,徐绍荣,杨晓杰,潘立臣. 基于 AHP 的葡萄酒产业发展影响因素评价[J]. 酿酒科技, 2020 (07) :124-128.
- [5] 朱哲,董星池,许力戈,许明辉,高翔,魏楠. 葡萄酒的检验与评价[J]. 中国新通信, 2019, 21 (06) :237.
- [6] 黄筱鹞. 中国葡萄酒感官评价体系正式发布[J]. 酿酒科技, 2018 (12) :44.
- [7] 许珊珊,谭兵,李屹,何婷. 葡萄酒质量的评价[J]. 现代商贸工业, 2018, 39 (09) :54-56.
- [8] 杨学山,刘琦,段卫朋,宋茹茹,韩舜愈,祝霞. 不同浸渍工艺对贵人香干白葡萄酒香

- 气品质的影响[J]. 核农学报, 2020, 34(10):2271-2279.
- [9] 李俊, 侯晓楠, 杨明岸, 欧嘉仪, 张琳. 红葡萄酒的品质鉴定研究进展[J]. 中外葡萄与葡萄酒, 2020(03):55-59.

附 录

8.1 任务 1 附录

8.1.1 各组葡萄酒样品评分均值表格

表 19 第一组红葡萄酒样品的数据汇总表

酒样品	品酒员										均值
	1	2	3	4	5	6	7	8	9	10	
1	51	66	49	54	77	61	72	61	74	62	62.7
2	71	81	86	74	91	80	83	79	85	73	80.3
3	80	85	89	76	69	89	73	83	84	76	80.4
4	52	64	65	66	58	82	76	63	83	77	68.6
5	74	74	72	62	84	63	68	84	81	71	73.3
6	72	69	71	61	82	69	69	64	81	84	72.2
7	63	70	76	64	59	84	72	59	84	84	71.5
8	64	76	65	65	76	72	69	85	75	76	72.3
9	77	78	76	82	85	90	76	92	80	79	81.5
10	67	82	83	68	75	73	75	68	76	75	74.2
11	73	60	72	63	63	71	70	66	90	73	70.1
12	54	42	40	55	53	60	47	61	58	69	53.9
13	69	84	79	59	73	77	77	76	75	77	74.6
14	70	77	70	70	80	59	76	76	76	76	73
15	70	77	70	70	80	59	76	76	76	76	73
16	72	80	80	71	69	71	80	74	78	74	74.9
17	70	79	91	68	97	82	69	80	81	76	79.3
18	63	65	51	55	52	57	62	58	70	68	60.1
19	76	84	84	66	68	87	80	78	82	81	78.6
20	78	84	76	74	82	79	76	76	86	81	79.2
21	73	90	96	71	69	60	79	73	86	74	77.1
22	73	83	72	68	93	72	75	77	79	80	77.2
23	83	85	86	80	95	93	81	91	84	78	85.6
24	70	85	90	68	90	84	70	75	78	70	78
25	60	78	81	62	70	67	64	62	81	67	69.2
26	73	80	71	61	78	71	72	76	79	77	73.8
27	70	77	63	64	80	76	73	67	85	75	73

表 20 第一组白葡萄酒样品的数据汇总表

酒样品	品酒员										均值
	1	2	3	4	5	6	7	8	9	10	
1	85	80	88	61	76	93	83	80	95	79	82
2	78	47	86	54	79	91	85	68	73	81	74.2
3	85	67	89	75	78	75	136	79	90	79	85.3
4	75	77	80	65	77	83	88	78	85	86	79.4
5	84	47	77	60	79	62	74	74	79	74	71
6	61	45	83	65	78	56	80	67	65	84	68.4
7	84	81	83	66	74	80	80	68	77	82	77.5
8	75	46	81	54	81	59	73	77	85	83	71.4
9	79	69	81	60	70	55	73	81	76	85	72.9
10	75	42	86	60	87	75	83	73	91	71	74.3
11	79	46	85	60	74	71	86	62	88	72	72.3
12	64	42	75	52	67	62	77	56	68	70	63.3
13	82	42	83	49	66	65	76	62	65	69	65.9
14	78	48	84	67	79	64	78	68	81	73	72
15	74	48	87	71	81	61	79	67	74	82	72.4
16	69	49	86	65	70	91	87	62	84	77	74
17	81	54	90	70	78	71	87	74	92	91	78.8
18	86	44	83	71	72	71	85	64	74	81	73.1
19	75	66	83	68	73	64	80	63	73	77	72.2
20	80	68	82	71	83	81	84	62	87	80	77.8
21	84	49	85	59	76	86	83	70	88	84	76.4
22	65	48	90	58	72	77	76	70	80	74	71
23	71	66	80	69	80	82	78	71	87	75	75.9
24	82	56	79	73	67	59	68	78	86	85	73.3
25	86	80	82	69	74	67	77	78	77	81	77.1
26	75	66	82	75	93	91	81	76	90	84	81.3
27	58	40	79	67	59	55	66	74	73	77	64.8
28	66	75	89	69	88	87	85	76	88	90	81.3

表 21 第二组红葡萄酒样品的数据汇总表

酒样品	品酒员										均值
	1	2	3	4	5	6	7	8	9	10	
1	68	71	80	52	53	76	71	73	70	67	68.1
2	75	76	76	71	68	74	83	73	73	71	74
3	82	69	80	78	63	75	72	77	74	76	74.6
4	75	79	73	72	60	77	73	73	60	70	71.2
5	66	68	77	75	76	73	72	72	74	68	72.1
6	65	67	75	61	58	66	70	67	67	67	66.3
7	68	65	68	65	47	70	57	74	72	67	65.3
8	71	70	78	51	62	69	73	59	68	59	66
9	81	83	85	76	69	80	83	77	75	73	78.2
10	67	73	82	62	63	66	66	72	65	72	68.8
11	64	61	67	62	50	66	64	51	67	64	61.6
12	67	68	75	58	63	73	67	72	69	71	68.3
13	74	64	68	65	70	67	70	76	69	65	68.8
14	71	71	78	64	67	76	74	80	73	72	72.6
15	62	60	73	54	59	71	71	70	68	69	65.7
16	71	65	78	70	64	73	66	75	68	69	69.9
17	72	73	75	74	75	77	79	76	76	68	74.5
18	67	65	80	55	62	64	62	74	60	65	65.4
19	72	65	82	61	64	81	76	80	74	71	72.6
20	80	75	80	66	70	84	79	83	71	70	75.8
21	80	72	75	72	62	77	63	70	73	78	72.2
22	77	79	75	62	68	69	73	71	69	73	71.6
23	79	77	80	83	67	79	80	71	81	74	77.1
24	66	69	72	73	73	68	72	76	76	70	71.5
25	68	68	84	62	60	66	69	73	66	66	68.2
26	68	67	83	64	73	74	77	78	63	73	72
27	71	64	72	71	69	71	82	73	73	69	71.5

表 22 第二组白葡萄酒样品的数据汇总表

酒样品	品酒员										均值
	1	2	3	4	5	6	7	8	9	10	
1	84	78	82	75	79	84	81	69	75	72	77.9
2	79	76	77	85	77	79	80	59	76	70	75.8
3	85	74	71	87	79	79	80	45	83	73	75.6
4	84	78	74	83	69	82	84	66	77	72	76.9
5	83	79	79	80	77	87	82	73	84	91	81.5
6	83	75	74	69	75	77	80	67	77	78	75.5
7	78	79	74	69	69	82	80	61	72	78	74.2
8	74	78	74	67	73	77	79	66	73	62	72.3
9	77	78	89	88	84	89	85	54	79	81	80.4
10	86	77	77	82	81	87	84	61	73	90	79.8
11	79	83	78	63	60	73	81	61	60	76	71.4
12	73	81	73	79	67	79	80	44	64	84	72.4
13	68	78	79	81	78	72	75	62	65	81	73.9
14	75	77	76	76	78	82	79	68	78	82	77.1
15	83	77	88	80	84	83	80	63	76	70	78.4
16	68	63	75	60	67	86	67	71	52	64	67.3
17	77	69	79	83	79	87	88	75	78	88	80.3
18	75	83	82	79	74	84	78	71	74	67	76.7
19	76	75	78	70	81	80	83	66	78	77	76.4
20	86	74	75	78	85	81	78	61	73	75	76.6
21	81	80	79	85	83	76	80	58	85	85	79.2
22	80	76	82	88	75	89	80	66	72	86	79.4
23	74	80	80	80	74	79	75	73	83	76	77.4
24	67	80	77	77	79	78	83	65	72	83	76.1
25	79	76	79	86	83	88	83	52	85	84	79.5
26	80	72	75	83	71	83	83	53	62	81	74.3
27	72	79	84	79	76	83	77	63	79	78	77
28	75	82	81	81	78	84	79	71	76	89	79.6

8.2.1 各组葡萄酒样品正态分布检验结果

(一) 第一组红葡萄酒

表 23 第一组红葡萄酒均值的正态性检验表

	柯尔莫戈洛夫-斯米诺夫(V) ^a			夏皮洛-威尔克		
	统计	自由度	p值	统计	自由度	p值
第一组红葡萄酒均值	.158	27	.200*	.922	27	.054

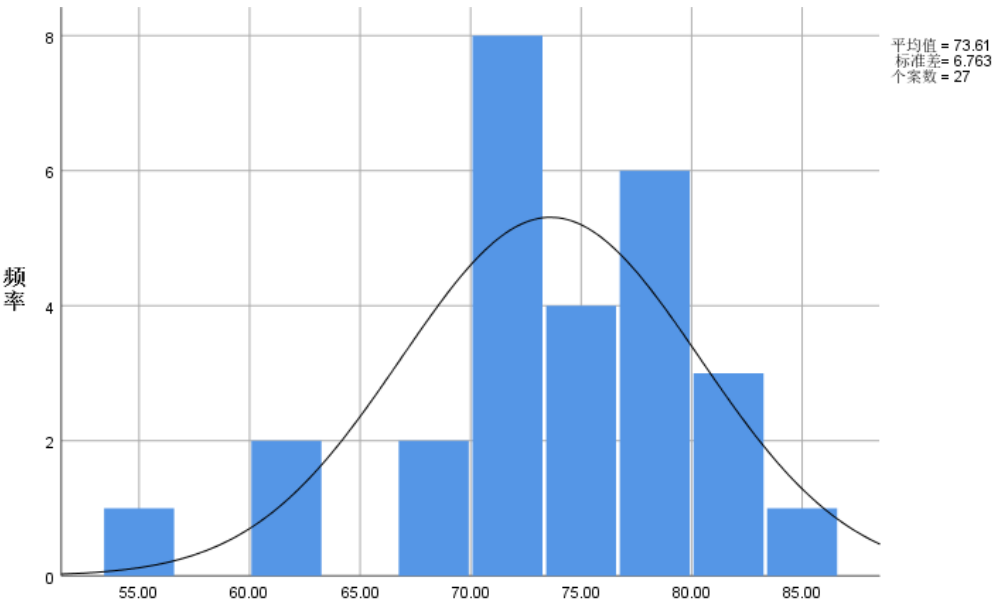


图 9 第一组红葡萄酒均值的直方图

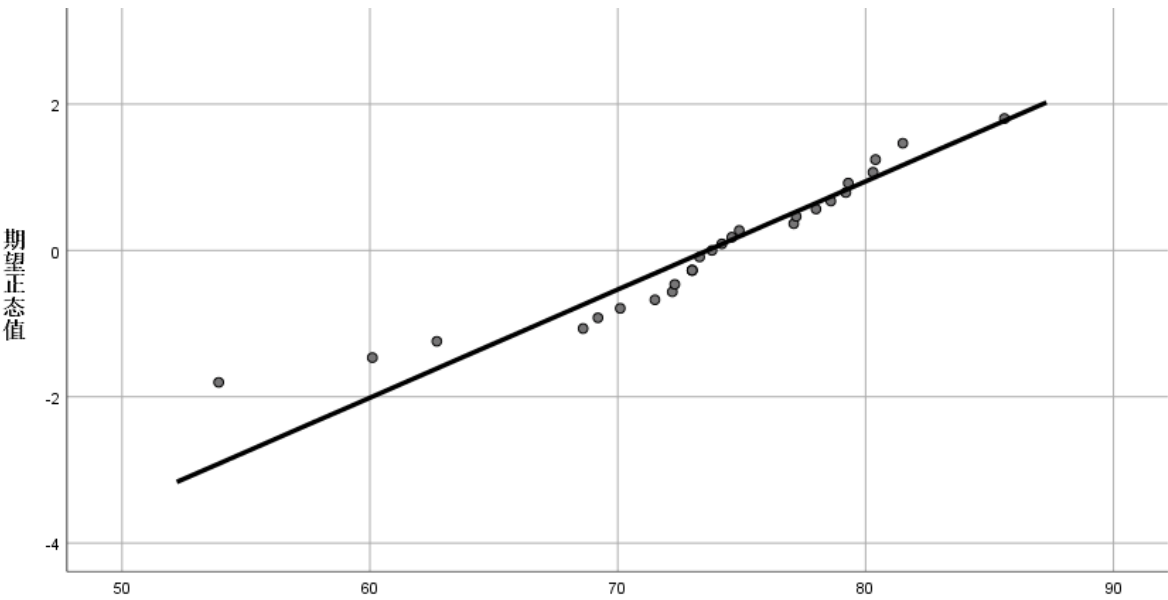


图 10 第一组红葡萄酒均值的正态 Q-Q 图

（二）第二组红葡萄酒

表 24 第二组红葡萄酒均值的正态性检验表

	柯尔莫戈洛夫-斯米诺夫(V) ^a			夏皮洛-威尔克		
	统计	自由度	p值	统计	自由度	p值
第二组红葡萄酒均值	.124	27	.200 [*]	.980	27	.868

*. 这是真显著性的下限。

a. 里利氏显著性修正

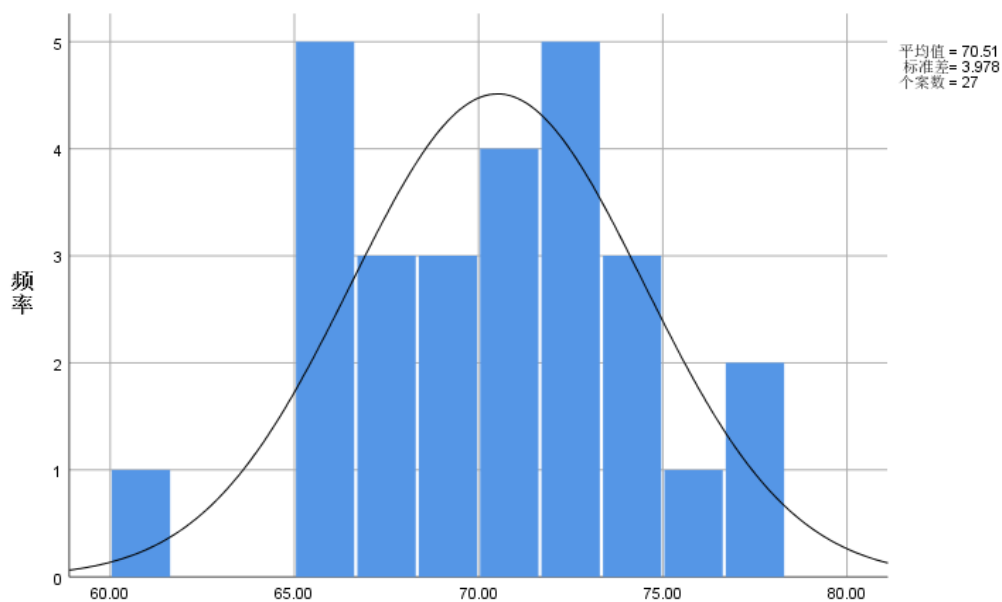


图 11 第二组红葡萄酒均值的直方图

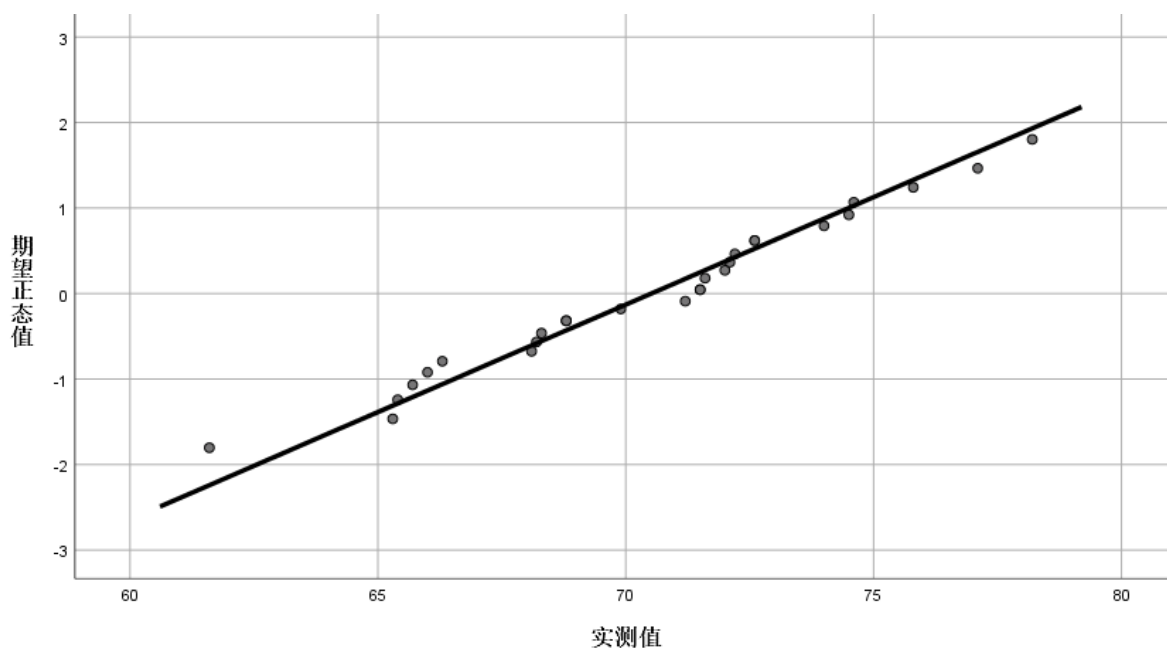


图 12 第二组红葡萄酒均值的正态 Q-Q 图

（二）第二组白葡萄酒

表 25 第二组白葡萄酒均值的正态性检验表

	柯尔莫戈洛夫-斯米诺夫(V) ^a			夏皮洛-威尔克		
	统计	自由度	p值	统计	自由度	p值
第二组白葡萄酒均值	.122	28	.200 [*]	.946	28	.153

*. 这是真显著性的下限。

a. 里利氏显著性修正

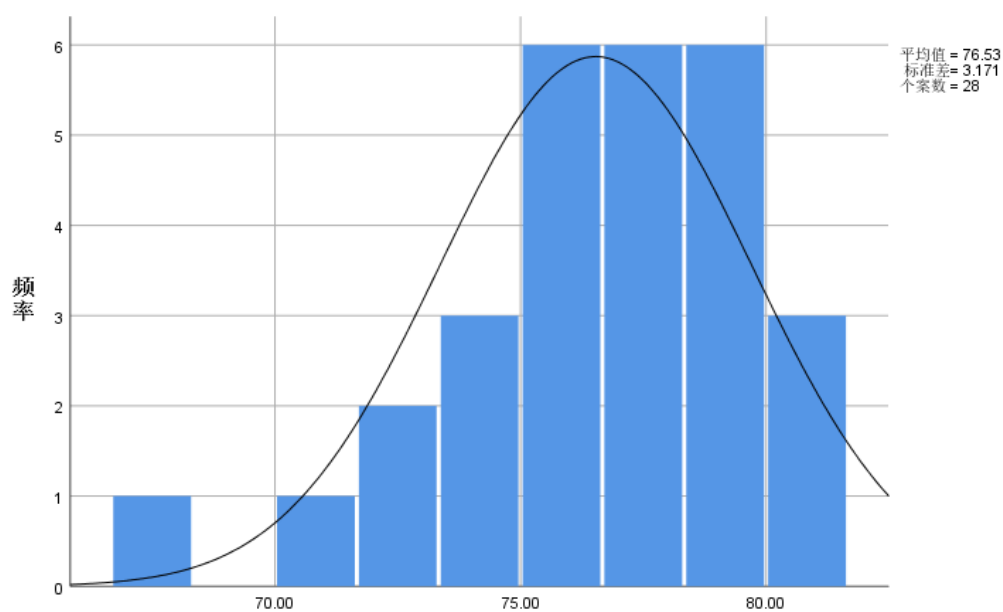


图 13 第二组白葡萄酒均值的直方图

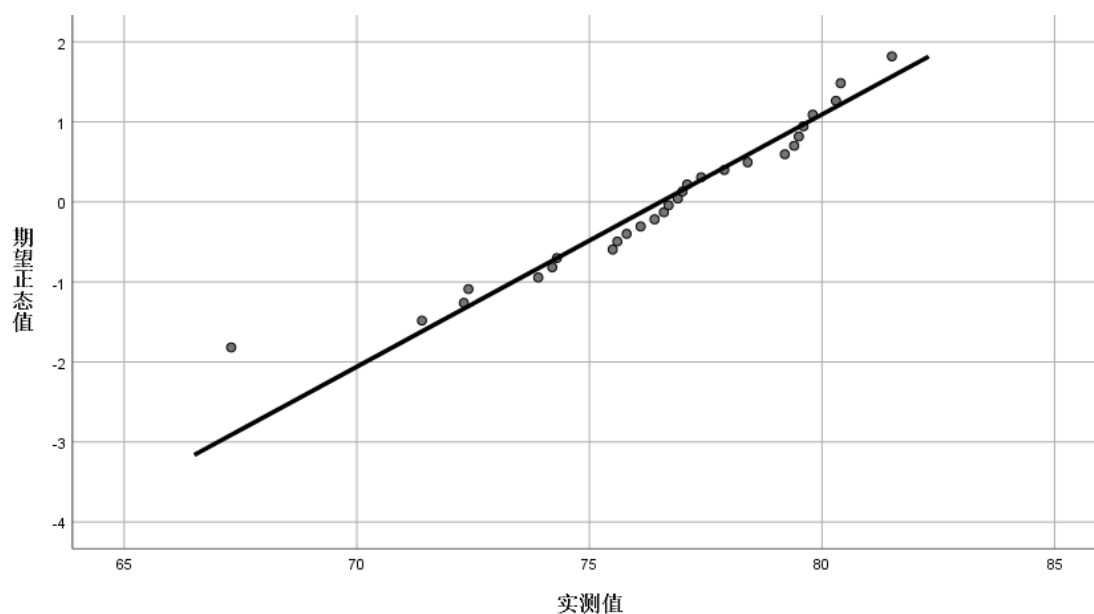


图 14 第二组白葡萄酒均值的正态 Q-Q 图

8.2 任务 2 附录

8.2.1 酿酒白葡萄程序及结果

(一) 酿酒白葡萄程序

%酿酒白葡萄的 13 个主因素数据

```
x=[1279.30 497.953 0.410 6.040 5.703 0.3977 186.849 267.60 72.3
51.245 0.3263 199.3 77.9
1870.93 539.254 0.404 5.420 5.016 1.6627 222.278 97.60 69.5
10.176 0.3058 209.1 75.8
5022.14 468.656 2.511 11.790 6.981 1.6333 232.429 203.96 74.7
12.474 0.2835 206.8 75.6
2085.76 496.948 1.625 6.920 5.342 1.3188 212.442 106.60 72.6
32.459 0.3681 213.6 76.9
2658.04 468.603 0.361 9.200 6.415 3.1705 256.02 106.70 69.0
14.196 0.2460 210.3 81.5
1847.12 492.088 0.417 8.320 10.733 8.0864 220.124 296.36 78.7
14.730 0.3366 198.8 75.5
1721.58 559.560 2.064 4.240 10.550 4.7952 242.2 234.10 68.6
17.211 0.3705 171.5 74.2
1273.22 493.739 2.455 7.590 5.322 5.6080 196.156 281.80 79.0
58.633 0.1045 174.8 72.3
1927.42 484.390 3.275 10.290 6.078 0.7518 220.415 138.40 64.9
32.220 0.3783 219.5 80.4
2095.61 516.492 0.422 7.340 7.854 1.1776 243.081 118.40 70.2
31.785 0.3418 225.2 79.8
1566.97 559.059 0.407 7.200 7.910 1.8589 174.03 323.30 72.0
38.482 0.2318 185.2 71.4
1724.16 459.045 0.409 4.640 8.643 8.4308 250.51 286.42 71.9
18.370 0.4484 221.7 72.4
664.96 460.830 2.090 4.900 11.697 0.2103 177.755 286.08 77.0 26.424
0.4089 186.7 73.9
1542.17 522.480 0.404 3.310 5.083 1.3006 253.4 144.20 77.4
31.728 0.4172 191.9 77.1
2669.22 549.854 1.222 5.550 8.684 11.8734 212.712 498.13 70.0
8.288 0.3536 189.7 78.4
991.92 464.819 3.693 5.750 4.998 5.3140 216.733 259.20 71.4 52.591
0.0976 185.1 67.3
1167.29 416.876 1.671 10.210 6.604 6.0135 182.553 104.78 75.0
20.481 0.2597 198.8 80.3
1289.93 583.403 0.812 3.980 12.078 2.7771 198.668 285.90 83.0
25.628 0.2112 182.8 76.7
817.81 449.107 2.072 9.380 4.877 0.8535 210.105 164.26 69.9 30.204
0.2003 204.5 76.4
2045.24 480.667 0.829 10.650 5.241 6.1633 235.097 202.38 75.9
29.330 0.3347 209.1 76.6
1554.02 586.240 0.819 8.500 4.386 0.7158 228.364 92.40 69.6
23.920 0.2524 207.9 79.2
1457.67 407.990 1.651 8.970 6.220 8.9296 225.523 73.52 68.3
36.357 0.3656 216.0 79.4
1522.52 504.624 1.244 8.280 5.184 12.4106 253.514 195.70 68.2
20.055 0.3596 210.3 77.4
3068.34 630.561 0.372 6.960 10.563 3.2138 227.25 195.60 66.7
14.021 0.4061 224.1 76.1
2350.79 519.210 0.402 6.950 7.611 8.6195 255.29 141.20 74.2
34.358 0.2123 228.7 79.5
2073.33 496.835 1.657 6.810 5.768 4.3089 259.877 102.20 56.3
```

```

18.194  0.2995  212.5   74.3
      2475.21    636.101    0.420   8.300   17.254  50.4966    224.944  165.70  63.8
19.723  0.3846  191.9   77
      3785.57    451.833    1.645   8.890   6.592   5.7323  239.325  108.00  65.9
46.060  0.3650  247.7   79.6

```

```

];
X=zscore(x); %数据标准化
r=corrcoef(X); %计算相关系数矩阵
[vec,val]=eig(r) %求特征值(val)及特征向量(vec)
newval=diag(val) ;
[y,i]=sort(newval) ; %对特征根进行排序, y 为排序结果, i 为索引
fprintf('特征根排序: \n')
for z=1:length(y)
    newy(z)=y(length(y)+1-z);
end
fprintf('%g\n',newy)
rate=y/sum(y);
fprintf('\n 贡献率: \n')
newrate=newy/sum(newy)
sumrate=0;
newi=[];
for k=length(y):-1:1
    sumrate=sumrate+rate(k);
    newi(length(y)+1-k)=i(k);
    if sumrate>0.85 break;
end
end %记下累积贡献率大 85%的特征值的序号放入 newi 中
fprintf('主成分数: %g\n\n',length(newi));
fprintf('主成分载荷: \n')
for p=1:length(newi)
    for q=1:length(y)
        result(q,p)=sqrt(newval(newi(p)))*vec(q,newi(p));
    end
end %计算载荷
disp(result)
A=(result) '*X';
B=[]
for i=1:28
    B(i,1)=sum(A(:,i)); %计算酿酒葡萄得分
end
end
B

```

(二) 酿酒白葡萄运行结果

```

vec =
    0.0419   -0.4796    0.1038   -0.0001   -0.2178    0.1252    0.3187   -0.3640    0.5064    0.1611    0.2510
0.0630    0.3331
    0.1266    0.4166    0.2526    0.0671   -0.2058    0.4603   -0.1035   -0.4838   -0.1787   -0.1024    0.0030
0.4479    0.0023
    0.2582    0.2683   -0.0776    0.2264   -0.4940    0.0887   -0.1255    0.2663    0.2665   -0.2159    0.4700
-0.3175   -0.1622
   -0.3332    0.3839    0.3115   -0.1569    0.2623    0.0009   -0.1336   -0.0407    0.0069    0.4445    0.4530
-0.2702    0.2363
   -0.4406    0.0994   -0.4757   -0.2436   -0.3376   -0.0537    0.1408    0.1990   -0.0636    0.1633    0.2400
0.4950   -0.0280
    0.4601   -0.1131    0.1401    0.1352    0.1785   -0.3941    0.1465    0.0975   -0.3760   -0.0012    0.4973
0.3480    0.0985
   -0.3029    0.2796    0.0209    0.2725   -0.1014   -0.5801    0.0231   -0.2460    0.1340   -0.4249   -0.1037
0.0174    0.3715
   -0.0477    0.2227   -0.1590    0.4831    0.4950    0.0760    0.2280    0.0200    0.4443    0.1076    0.0468

```

0.2506	-0.3351										
	0.2524	0.2381	0.1806	-0.0971	-0.3406	-0.4437	0.1666	-0.1272	0.1223	0.5273	-0.2963
-0.0066	-0.3162										
	-0.2885	-0.0164	0.1614	0.2019	-0.1240	0.1257	0.6959	-0.0688	-0.4085	-0.0941	0.0679
-0.3079	-0.2333										
	-0.0425	0.0918	0.5253	-0.0024	-0.0869	0.1373	0.2066	0.6482	0.1751	-0.0597	-0.2336
0.2541	0.2668										
	0.3949	0.3939	-0.4206	-0.2413	0.1448	0.1379	0.4211	0.0103	-0.0042	-0.0002	-0.1277
-0.1746	0.4354										
	-0.0080	-0.0783	-0.1984	0.6527	-0.1906	0.1163	-0.1682	0.0981	-0.2593	0.4585	-0.1724
-0.0504	0.3619										
val =											
	0.1096	0	0	0	0	0	0	0	0	0	0
0	0	0									
	0	0.1622	0	0	0	0	0	0	0	0	0
0	0	0									
	0	0	0.2420	0	0	0	0	0	0	0	0
0	0	0									
	0	0	0	0.3393	0	0	0	0	0	0	0
0	0	0									
	0	0	0	0	0.3983	0	0	0	0	0	0
0	0	0									
	0	0	0	0	0	0	0.5591	0	0	0	0
0	0	0									
	0	0	0	0	0	0	0	0.6566	0	0	0
0	0	0									
	0	0	0	0	0	0	0	0	0.8474	0	0
0	0	0									
	0	0	0	0	0	0	0	0	0	0.9681	0
0	0	0									
	0	0	0	0	0	0	0	0	0	0	1.1151
0	0	0									
	0	0	0	0	0	0	0	0	0	0	0
1.3258	0	0									
	0	0	0	0	0	0	0	0	0	0	0
0	2.8402	0									
	0	0	0	0	0	0	0	0	0	0	0
0	0	3.4362									

特征根排序:

3.43618

2.84017

1.32583

1.11508

0.968133

0.847368

0.656638

0.559135

0.398314

0.339316

0.242028

0.162164

0.109634

贡献率:

newrate =

	0.2643	0.2185	0.1020	0.0858	0.0745	0.0652	0.0505	0.0430	0.0306	0.0261	0.0186
0.0125	0.0084										

主成分数: 7

主成分载荷:

0.6175	0.1061	0.2890	0.1701	0.4983	-0.3351	0.2583
0.0043	0.7548	0.0034	-0.1081	-0.1759	-0.4454	-0.0839
-0.3006	-0.5350	0.5412	-0.2279	0.2622	0.2451	-0.1017
0.4380	-0.4553	0.5216	0.4694	0.0068	-0.0375	-0.1083
-0.0519	0.8342	0.2764	0.1724	-0.0625	0.1832	0.1141
0.1826	0.5865	0.5726	-0.0013	-0.3699	0.0898	0.1187
0.6887	0.0293	-0.1194	-0.4486	0.1319	-0.2264	0.0187
-0.6211	0.4223	0.0539	0.1136	0.4372	0.0184	0.1847
-0.5861	-0.0110	-0.3412	0.5568	0.1204	-0.1171	0.1350
-0.4324	-0.5188	0.0782	-0.0994	-0.4020	-0.0633	0.5639
0.4946	0.4282	-0.2690	-0.0631	0.1723	0.5967	0.1674
0.8071	-0.2943	-0.1471	-0.0003	-0.0042	0.0095	0.3412
0.6709	-0.0849	-0.1985	0.4842	-0.2552	0.0903	-0.1363

B =

[]

B =

-4.1133
-1.9846
6.6028
-1.1605
1.7796
3.3718
0.0148
-9.7736
1.5913
1.2441
-4.1178
3.1442
-1.1444
-3.3251
5.5630
-11.2427
-1.4854
-3.0263
-5.4498
0.7967
-3.1586
0.2626
1.2363
6.4794
0.0616
-0.5375
13.2580
5.1135

8.2.2 酿酒红葡萄程序及结果

（一）酿酒红葡萄程序

%酿酒红葡萄的 13 个主因素数据

```
x=[2027.96 552.323 407.608 2.060 23.344 17.6780 237.668 182.68 78.2
33.658 0.4296 214.8 68.1
2128.82 627.273 223.956 9.930 26.988 27.4550 229.136 80.60 77.5
30.811 0.4637 235.8 74
8397.28 586.637 157.512 8.080 21.714 164.9927 273.758 82.68 71.3
19.209 0.4085 250.7 74.6
2144.68 530.618 79.264 3.770 10.715 26.9679 237.766 139.70 53.3
15.435 0.2653 191.9 71.2
1844.00 583.934 120.184 9.490 17.734 6.6502 195.46 505.04 65.7
```

```

31.440 0.3956 212.5 72.1
    3434.17    538.182    45.766 2.830 10.604 7.7272 223.817 200.34 71.6
36.680 0.2748 240.3 66.3
    2391.16    486.456    60.356 5.820 9.115 9.8648 303.95 61.00 71.4
25.498 0.1755 214.8 65.3
    1950.76    559.364    240.972    5.710 14.660 115.5546 196.99 212.20 59.1
50.337 0.4143 214.8 66
    2262.72    700.004    240.430    13.230 30.513 58.5407 194.925 175.52 78.3
16.773 0.6644 214.8 78.2
    1364.14    544.508    43.786 2.450 9.258 28.7475 161.421 255.20 71.3
10.329 0.3252 181.5 68.8
    2355.69    541.872    7.377 9.290 6.046 25.5751 237.891 177.90 58.1
14.163 0.2788 214.8 61.6
    2556.79    494.192    31.923 6.080 12.261 2.4802 262.155 193.04 63.8
20.986 0.1971 248.7 68.3
    1416.11    607.043    64.902 4.300 14.097 40.7586 212.237 155.50 68.3
27.982 0.4400 221.7 68.8
    1237.81    600.680    139.841    5.730 14.591 134.6375 255.335 203.90 66.3
41.478 0.3593 197.7 72.6
    2177.91    522.340    52.376 6.230 11.734 9.7179 208.933 159.70 67.4
25.646 0.2188 233.4 65.7
    1553.50    582.571    60.248 9.030 11.184 8.1900 189.275 110.90 71.4
13.552 0.2366 203.3 69.9
    1713.65    549.606    58.999 5.880 15.745 43.8121 271.504 446.70 71.2
17.079 0.3581 238.2 74.5
    2398.38    514.567    38.850 3.60 7.239 6.5161 265.773 195.04 62.5
26.982 0.2254 203.3 65.4
    2463.60    545.215    115.301    5.560 18.398 31.2649 220.333 172.00 67.6
30.315 0.3791 212.5 72.6
    2273.63    560.145    23.110 3.510 11.981 9.6262 227.338 303.10 59.4
12.343 0.2817 214.8 75.8
    6346.83    562.996    88.879 15.510 16.746 47.2196 259.11 140.20 60.0
18.026 0.3789 224.1 72.2
    2566.61    488.030    73.606 6.490 17.340 13.8003 226.399 102.60 56.3
21.727 0.2834 219.5 71.6
    2380.81    542.161    172.217    4.080 29.704 44.7476 212.564 272.60 77.5
16.311 0.5715 205.7 77.1
    1638.83    532.951    144.453    8.360 8.721 14.3803 244.512 518.63 76.5
14.973 0.2828 191.9 71.5
    1409.70    536.278    49.236 2.870 12.205 30.2112 156.038 289.64 58.9
14.186 0.3505 193.4 68.2
    851.17 586.422 58.045 7.150 7.187 13.9166 197.377 793.60 68.7
31.934 0.3169 198.3 72
    1116.61    529.829    33.778 6.230 8.881 15.9809 213.216 277.16 59.4
22.936 0.2647 204.7 71.5
];
X=zscore(X); %数据标准化
r=corrcoef(X); %计算相关系数矩阵
[vec1,lamda,rate]=pcacov(r); %lamda为r的特征值,rate为各个主成分的贡献率
[vec,val]=eig(r) %求特征值(val)及特征向量(vec)
newval=diag(val);
[y,i]=sort(newval); %对特征根进行排序,y为排序结果,i为索引
fprintf('特征根排序: \n')
for z=1:length(y)
    newy(z)=y(length(y)+1-z);
end
fprintf('%g\n',newy)
rate=y/sum(y);

```

```

fprintf('\n 贡献率: \n')
newrate=newy/sum(newy)
sumrate=0;
newi=[];
for k=length(y):-1:1
    sumrate=sumrate+rate(k);
    newi(length(y)+1-k)=i(k);
    if sumrate>0.85 break;
end
end %记下累积贡献率大 85%的特征值的序号放入 newi 中
fprintf('主成分数: %g\n\n',length(newi));
fprintf('主成分载荷: \n')
for p=1:length(newi)
    for q=1:length(y)
        result(q,p)=sqrt(newval(newi(p)))*vec(q,newi(p));
    end
end %计算载荷
disp(result)
A=((result))'*X';
B=[]
for i=1:27
    B(i,1)=sum(A(:,i)); %计算酿酒葡萄得分
end
end

```

B (二) 酿酒红葡萄运行结果

```

vec =
    -0.0417    -0.3014     0.3575    -0.2755     0.5455     0.2282     0.0244    -0.1408    -0.0421     0.1261
0.2009    -0.4999     0.1603
    -0.2347     0.0258     0.5944     0.0905    -0.4099     0.2758    -0.0356     0.2720     0.1493     0.2357
0.1006     0.1870     0.3747
     0.1724     0.5122     0.2975     0.0887     0.3848    -0.2646    -0.3032    -0.0031    -0.0748    -0.1343
-0.3837     0.0585     0.3567
     0.0713     0.0729    -0.3459    -0.0968     0.1219    -0.2463    -0.3253     0.5204     0.3681     0.1841
0.4212    -0.1340     0.2090
    -0.7234    -0.1056    -0.2339     0.0214     0.0559    -0.2911     0.1897    -0.0013    -0.2370    -0.1922
-0.0152    -0.0110     0.4394
    -0.0637     0.2784    -0.4023     0.0942    -0.1081     0.3340    -0.1323    -0.4020    -0.0073     0.5584
-0.1569    -0.1979     0.2641
    -0.0626    -0.2040     0.1371     0.3085    -0.3479    -0.2626    -0.4008    -0.3194     0.2971    -0.2597
-0.0285    -0.4807    -0.0182
    -0.2075     0.0253     0.0173     0.2363     0.3783     0.0395     0.1566    -0.3357     0.6717    -0.0121
0.1184     0.3868    -0.0577
     0.0649    -0.0875    -0.2385    -0.2657    -0.0362     0.5664    -0.2596     0.0141     0.1416    -0.5907
-0.0975     0.1158     0.2813
     0.0356    -0.3115    -0.0240    -0.4014    -0.0750    -0.2006     0.1675     0.1250     0.3897     0.2147
-0.6675    -0.0342     0.0754
     0.4639    -0.5228    -0.1093     0.4859     0.0827    -0.0404     0.1085     0.0009    -0.1352     0.0723
-0.0048     0.1746     0.4350
     0.1593     0.3384    -0.0404     0.2270    -0.0675     0.1287     0.6399     0.2047     0.2182    -0.2387
-0.0277    -0.4607     0.1299
     0.2967     0.1321     0.0627    -0.4679    -0.2723    -0.3085     0.2123    -0.4488     0.0368    -0.0555
0.3596     0.1288     0.3249

val =
    0.0174         0         0         0         0         0         0         0         0         0
0         0         0         0
         0     0.1356         0         0         0         0         0         0         0
0         0         0         0
         0         0     0.1527         0         0         0         0         0         0

```

0	0	0	0	0	0	0.1854	0	0	0	0	0
0	0	0	0	0	0	0	0.4005	0	0	0	0
0	0	0	0	0	0	0	0	0.4913	0	0	0
0	0	0	0	0	0	0	0	0	0.5841	0	0
0	0	0	0	0	0	0	0	0	0	0.7893	0
0	0	0	0	0	0	0	0	0	0	0	0.9545
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
1.0779	0	0	0	0	0	0	0	0	0	0	0
0	1.5358	0	0	0	0	0	0	0	0	0	0
0	0	2.4128	0	0	0	0	0	0	0	0	0
0	0	0	4.2627	0	0	0	0	0	0	0	0

特征根排序:

4.26269

2.41277

1.53584

1.07787

0.954473

0.789314

0.584123

0.491326

0.400458

0.185425

0.152745

0.135587

0.0173824

贡献率:

newrate =

0.3279	0.1856	0.1181	0.0829	0.0734	0.0607	0.0449	0.0378	0.0308	0.0143
0.0117	0.0104	0.0013							

主成分数: 7

主成分载荷:

0.3310	-0.7765	0.2490	0.1309	-0.0411	-0.1251	0.0186
0.7737	0.2904	0.1247	0.2447	0.1459	0.2417	-0.0272
0.7364	0.0908	-0.4756	-0.1395	-0.0731	-0.0028	-0.2317
0.4315	-0.2081	0.5220	0.1911	0.3596	0.4623	-0.2486
0.9072	-0.0172	-0.0188	-0.1996	-0.2315	-0.0012	0.1450
0.5453	-0.3074	-0.1944	0.5797	-0.0072	-0.3571	-0.1011
-0.0375	-0.7467	-0.0353	-0.2696	0.2902	-0.2837	-0.3063
-0.1191	0.6008	0.1467	-0.0126	0.6562	-0.2982	0.1197
0.5808	0.1798	-0.1208	-0.6132	0.1384	0.0125	-0.1984
0.1557	-0.0532	-0.8273	0.2229	0.3808	0.1111	0.1280
0.8981	0.2712	-0.0060	0.0751	-0.1321	0.0008	0.0829
0.2682	-0.7156	-0.0344	-0.2478	0.2132	0.1818	0.4891
0.6708	0.2001	0.4456	-0.0576	0.0360	-0.3987	0.1622

B =

[]

B =

-2.2250
6.7132
1.3257
-4.5751
6.8865
-4.5472
-11.0391
0.9562
16.4566
-0.9663
-3.8886
-6.4915
1.8892
1.3634
-3.9208
1.1547
0.7015
-8.1752
0.2676
-0.3751
3.1488
-3.7008
4.4183
-0.0690
-0.1012
6.0829
-1.2898

8.2.3 酿酒白葡萄程序及结果

表 26 白葡萄酒相关系数矩阵表

1.00 0	0.13 0	-0.0 75	0.39 4	0.10 6	0.12 8	0.41 0	-0.1 04	-0.2 47	-0.3 26	0.21 5	0.46 5	0.26 3
0.13 0	1.00 0	-0.4 40	-0.3 51	0.51 1	0.35 5	0.03 0	0.19 8	-0.1 10	-0.2 79	0.10 1	-0.2 27	-0.0 55
-0.0 75	-0.4 40	1.00 0	0.22 9	-0.2 58	-0.1 84	-0.1 93	0.01 9	-0.0 89	0.32 9	-0.3 18	-0.2 02	-0.3 11
0.39 4	-0.3 51	0.22 9	1.00 0	-0.2 33	0.09 0	0.02 7	-0.3 71	-0.1 89	-0.0 22	-0.1 50	0.38 0	0.39 2
0.10 6	0.51 1	-0.2 58	-0.2 33	1.00 0	0.62 7	-0.1 44	0.33 2	0.04 7	-0.3 37	0.31 5	-0.2 55	-0.0 70
0.12 8	0.35 5	-0.1 84	0.09 0	0.62 7	1.00 0	0.12 2	0.05 6	-0.2 65	-0.1 64	0.17 5	-0.0 99	0.03 0
0.41 0	0.03 0	-0.1 93	0.02 7	-0.1 44	0.12 2	1.00 0	-0.4 05	-0.4 16	-0.3 13	0.27 1	0.49 8	0.23 5
-0.1 04	0.19 8	0.01 9	-0.3 71	0.33 2	0.05 6	-0.4 05	1.00 0	0.37 8	-0.0 22	-0.0 59	-0.5 49	-0.4 64
-0.2 47	-0.1 10	-0.0 89	-0.1 89	0.04 7	-0.2 65	-0.4 16	0.37 8	1.00 0	0.19 0	-0.2 61	-0.4 13	-0.1 51
-0.3 26	-0.2 79	0.32 9	-0.0 22	-0.3 37	-0.1 64	-0.3 13	-0.0 22	0.19 0	1.00 0	-0.4 28	-0.0 58	-0.2 35

0.21 5	0.10 1	-0.3 18	-0.1 50	0.31 5	0.17 5	0.27 1	-0.0 59	-0.2 61	-0.4 28	1.00 0	0.33 1	0.29 4
0.46 5	-0.2 27	-0.2 02	0.38 0	-0.2 55	-0.0 99	0.49 8	-0.5 49	-0.4 13	-0.0 58	0.33 1	1.00 0	0.51 0
0.26 3	-0.0 55	-0.3 11	0.39 2	-0.0 70	0.03 0	0.23 5	-0.4 64	-0.1 51	-0.2 35	0.29 4	0.51 0	1.00 0

表27 红葡萄酒相关数据汇总表

特征 向量	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13
x1	0.33	0.06	0.25	0.16	0.50	-0.3	0.31	0.12	-0.2	-0.0	0.10	-0.4	0.04
	31	30	10	11	64	640	87	52	178	001	38	796	19
x2	0.00	0.44	0.00	-0.1	-0.1	-0.4	-0.1	0.46	-0.2	0.06	0.25	0.41	0.12
	23	79	30	024	787	838	035	03	058	71	26	66	66
x3	-0.1	-0.3	0.47	-0.2	0.26	0.26	-0.1	0.08	-0.4	0.22	-0.0	0.26	0.25
	622	175	00	159	65	63	255	87	940	64	776	83	82
x4	0.23	-0.2	0.45	0.44	0.00	-0.0	-0.1	0.00	0.26	-0.1	0.31	0.38	-0.3
	63	702	30	45	69	407	336	09	23	569	15	39	332
x5	-0.0	0.49	0.24	0.16	-0.0	0.19	0.14	-0.0	-0.3	-0.2	-0.4	0.09	-0.4
	280	50	00	33	636	90	08	537	376	436	757	94	406
x6	0.09	0.34	0.49	-0.0	-0.3	0.09	0.14	-0.3	0.17	0.13	0.14	-0.1	0.46
	85	80	73	012	760	75	65	941	85	52	01	131	01
x7	0.37	0.01	-0.1	-0.4	0.13	-0.2	0.02	-0.5	-0.1	0.27	0.02	0.27	-0.3
	15	74	037	249	40	460	31	801	014	25	09	96	029
x8	-0.3	0.25	0.04	0.10	0.44	0.02	0.22	0.07	0.49	0.48	-0.1	0.22	-0.0
	351	06	68	76	43	00	80	60	50	31	590	27	477
x9	-0.3	-0.0	-0.2	0.52	0.12	-0.1	0.16	-0.4	-0.3	-0.0	0.18	0.23	0.25
	162	066	963	73	23	272	66	437	406	971	06	81	24
x10	-0.2	-0.3	0.06	-0.0	-0.4	-0.0	0.69	0.12	-0.1	0.20	0.16	-0.0	-0.2
	333	079	79	941	085	688	59	57	240	19	14	164	885
x11	0.26	0.25	-0.2	-0.0	0.17	0.64	0.20	0.13	-0.0	-0.0	0.52	0.09	-0.0
	68	41	336	597	51	82	66	73	869	024	53	18	425
x12	0.43	-0.1	-0.1	-0.0	-0.0	0.01	0.42	0.13	0.14	-0.2	-0.4	0.39	0.39
	54	746	277	002	042	03	11	79	48	413	206	39	49
x13	0.36	-0.0	-0.1	0.45	-0.2	0.09	-0.1	0.11	-0.1	0.65	-0.1	-0.0	0.29
	19	504	724	85	593	81	682	63	906	27	984	783	67
特征 值	3.43 62	2.84 02	1.32 58	1.11 51	0.96 81	0.84 74	0.65 66	0.55 91	0.39 83	0.33 93	0.24 20	0.16 22	0.10 96
贡献 率	0.26 43	0.21 85	0.10 20	0.08 58	0.07 45	0.06 52	0.05 05	0.04 30	0.03 06	0.02 61	0.01 86	0.01 25	0.00 84
累计 贡献 率	0.32 79	0.48 28	0.58 48	0.67 06	0.74 50	0.81 02	0.86 07	0.90 37	0.93 44	0.96 05	0.97 91	0.99 16	1.00 00

8.3 任务 3 附录

(1) SPSS 求得结果

```
STATS CANCORR SET1=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 SET2=x13 x14
x15 x16 x17 x18 x19 x20 x21
/OPTIONS COMPUTECVARS=NO
/PRINT PAIRWISECORR=NO LOADINGS=YES VARPROP=YES
COEFFICIENTS=YES.
```

Canonical Correlations

备注

已创建输出		20-JUN-2021 02:49:05
注释		
输入	活动数据集	数据集 0
	过滤器	<无>
	权重	<无>
	拆分文件	<无>
语法		BEGIN PROGRAM '# !.
资源	处理程序时间	00:00:00.34
	耗用时间	00:00:00.95

[数据集 0]

典型相关性设置

值

集合 1 变量	x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12
集合 2 变量	x13 x14 x15 x16 x17 x18 x19 x20 x21
集中的数据集	无
评分语法	无
用于评分的相关性	9

典型相关系数

	相关 系数	特征 值	威尔克 统计	F	分子自 由度	分母自 由度	p 值
1	.988	41.664	.000	2.428	108.000	56.202	.000
2	.960	11.744	.000	1.805	88.000	55.362	.010
3	.951	9.514	.002	1.497	70.000	53.464	.062
4	.931	6.522	.019	1.104	54.000	50.485	.362
5	.756	1.338	.141	.657	40.000	46.383	.911
6	.647	.720	.330	.528	28.000	41.083	.961
7	.484	.306	.568	.424	18.000	34.426	.972
8	.450	.254	.741	.419	10.000	26.000	.924
9	.265	.076	.930	.265	4.000	14.000	.896

H0 for Wilks 检验是指当前行和后续行中的相关性均为零

集合 1 标准化典型相关变量对应的线性相关系数

变 量	1	2	3	4	5	6	7	8	
1 x	-.062	.402	-.226	.275	-.490	-.681	.261	.112	
2 x	.152	-.031	.175	-.800	1.172	-.445	.676	.424	
3 x	-1.17	-.671	.391	-.745	-.589	.127	.250	.678	
4 x	.016	.006	.262	.509	-.253	.136	.443	-.466	
5 x	.383	.584	-.190	1.279	1.662	-.565	-.781	-.199	
6 x	-.057	.111	.414	-.243	.049	-.569	-.585	-.439	
7 x	.296	.296	.453	.231	.281	.389	.262	.913	
8 x	-.008	.134	.054	.251	-.008	-.627	-.075	.284	
9 x	.082	.093	-.342	.232	.157	-.481	-.410	-.053	
10 x	.096	-.162	.181	.941	.282	-.095	.040	-.569	
11 x1	-.254	.480	-.435	-.161	-1.59	1.447	.569	-.114	
12 x	-.162	-.048	.383	-.998	-.288	.529	-.416	-.226	

集合 2 标准化典型相关变量对应的线性相关系数

变 量	1	2	3	4	5	6	7	8	
x 13	-.190	-1.45 8	1.549	1.448	-.967	1.281	-1.70 8	-1.17 3	
x 14	-.236	.868	.997	-1.51 4	-.375	-.957	-.148	-.722	
x 15	-.218	-.484	-1.23 0	.516	-.715	2.204	3.130	-2.05 3	
x 16	-.172	-.330	-.070	.146	1.931	.453	-.490	.150	
x 17	-.166	.204	-.034	.583	-.421	-.569	-.124	.460	
x 18	.743	.821	-.249	.165	-.898	.008	-2.50 3	1.130	
x 19	.718	-.813	.776	.708	-1.37 2	2.473	-1.30 9	-1.96 9	
x 20	.247	-.346	-.121	.289	-.138	.501	-.552	-1.66 6	
x 21	.362	-.063	1.135	.742	.067	.600	-.113	-.306	

集合 1 非标准化典型相关变量对应的线性相关系数

变 量	1	2	3	4	5	6	7	8	
x 1	.000	.000	.000	.000	.000	.000	.000	.000	
x 2	.003	-.001	.004	-.018	.026	-.010	.015	.009	
x 3	-.013	-.007	.004	-.008	-.007	.001	.003	.008	
x 4	.005	.002	.081	.158	-.079	.042	.138	-.145	
x 5	.057	.087	-.028	.190	.247	-.084	-.116	-.030	
x 6	-.001	.003	.010	-.006	.001	-.014	-.014	-.011	
x 7	.009	.009	.013	.007	.008	.011	.008	.026	

8	x	.000	.001	.000	.002	.000	-.004	.000	.002	
9	x	.011	.013	-.047	.032	.021	-.066	-.056	-.007	
10	x	.010	-.017	.018	.096	.029	-.010	.004	-.058	
11	x1	-2.27	4.296	-3.89	-1.44	-14.2	12.94	5.087	-1.01	
12	x	3	2	0	26	3	8			
	x	-.009	-.003	.022	-.056	-.016	.030	-.023	-.013	

集合 2 非标准化典型相关变量对应的线性相关系数

变 量		1	2	3	4	5	6	7	8	
13	x	-.001	-.006	.007	.006	-.004	.006	-.007	-.005	
14	x	-.081	.299	.343	-.521	-.129	-.329	-.051	-.249	
15	x	-.086	-.192	-.487	.204	-.283	.873	1.239	-.813	
16	x	-.057	-.111	-.024	.049	.647	.152	-.164	.050	
17	x	-.057	.071	-.012	.202	-.146	-.197	-.043	.159	
18	x	5.827	6.439	-1.95	1.291	-7.04	.062	-19.6	8.865	
19	x	.034	-.038	.036	.033	-.064	.116	-.061	-.092	
20	x	.019	-.026	-.009	.022	-.010	.038	-.042	-.126	
21	x	.048	-.008	.149	.098	.009	.079	-.015	-.040	

集合 1 典型载荷

变 量		1	2	3	4	5	6	7	8	
1	x	-.068	.676	.371	-.070	-.372	-.177	-.049	-.030	
2	x	-.395	.354	-.048	-.273	.527	-.153	.507	-.210	

3	x	-.953	.081	.099	.116	.188	.045	-.004	.142	
4	x	-.023	.434	.270	.136	.037	.004	.638	-.193	
5	x	-.630	.618	-.118	.123	.338	.226	-.086	-.082	
6	x	-.364	.420	.457	-.148	.007	-.350	-.082	-.284	
7	x	.225	.277	.686	.161	-.022	.141	-.165	.538	
8	x	.064	-.162	-.199	.196	-.057	-.460	.177	.163	
9	x	-.400	.260	-.244	.043	.420	-.201	-.180	.255	
10	x	-.368	-.343	.496	.322	.225	-.015	-.121	-.333	
11	x1	-.651	.550	-.243	.010	.223	.108	.193	-.176	
12	x	.011	.425	.502	-.227	.040	.256	-.362	-.142	

集合 2 典型载荷

变 量		1	2	3	4	5	6	7	8	
13	x	-.954	-.008	.195	.085	-.011	.094	-.145	-.042	
14	x	-.715	.606	.179	-.111	-.002	.187	-.112	-.170	
15	x	-.754	.541	-.013	.158	.029	.316	.015	-.096	
16	x	-.648	.498	-.098	.103	.327	.392	-.218	-.021	
17	x	-.166	.569	-.329	.636	-.251	-.095	-.077	-.034	
18	x	-.629	.664	-.017	.145	.061	.278	-.219	-.033	
19	x	.796	-.329	-.158	-.268	-.248	.169	.007	.186	
20	x	.362	.029	-.216	.315	.219	-.531	-.109	-.614	

x	.425	.486	.510	.426	.267	-.040	.232	-.099	
21									

集合 1 交叉载荷

变 量	1	2	3	4	5	6	7	8	
x	-.067	.649	.353	-.065	-.282	-.114	-.024	-.013	
1									
x	-.390	.340	-.046	-.254	.399	-.099	.246	-.094	
2									
x	-.942	.077	.094	.108	.142	.029	-.002	.064	
3									
x	-.023	.417	.257	.127	.028	.003	.309	-.087	
4									
x	-.623	.593	-.113	.114	.256	.146	-.041	-.037	
5									
x	-.359	.403	.435	-.138	.005	-.226	-.040	-.128	
6									
x	.222	.266	.653	.150	-.017	.091	-.080	.242	
7									
x	.063	-.156	-.189	.183	-.043	-.297	.086	.073	
8									
x	-.395	.249	-.232	.040	.318	-.130	-.087	.115	
9									
x	-.364	-.329	.472	.300	.171	-.010	-.058	-.150	
10									
x1	-.643	.528	-.232	.009	.169	.070	.093	-.079	
1									
x	.011	.408	.478	-.212	.030	.166	-.175	-.064	
12									

集合 2 交叉载荷

变 量	1	2	3	4	5	6	7	8	
x	-.943	-.007	.185	.079	-.008	.061	-.070	-.019	
13									
x	-.707	.582	.170	-.103	-.002	.121	-.054	-.077	
14									
x	-.745	.519	-.012	.147	.022	.204	.007	-.043	
15									

16	x	-.641	.478	-.093	.096	.247	.254	-.106	-.009	
17	x	-.164	.546	-.313	.592	-.190	-.062	-.037	-.015	
18	x	-.621	.637	-.016	.135	.046	.180	-.106	-.015	
19	x	.786	-.316	-.151	-.249	-.188	.110	.003	.084	
20	x	.357	.027	-.206	.294	.166	-.344	-.053	-.276	
21	x	.420	.466	.485	.397	.202	-.026	.112	-.045	

已解释的方差比例

典型变量	集合 1 * 自身	集合 1 * 集合 2	集合 2 * 自身	集合 2 * 集合 1
1	.198	.193	.420	.410
2	.176	.162	.224	.207
3	.131	.119	.058	.052
4	.031	.027	.093	.080
5	.071	.041	.039	.023
6	.048	.020	.077	.032
7	.080	.019	.022	.005
8	.061	.012	.052	.010
9	.068	.005	.015	.001

8.4 任务 4 附录

SPSS 软件运行结果

```
FREQUENCIES VARIABLES=VAR00014
/STATISTICS=RANGE MINIMUM MAXIMUM MODE
/ORDER=ANALYSIS.
```

频率

备注

已创建输出	20-JUN-2021 03:37:54
注释	
输入	活动数据集
	数据集 0
	过滤器
	<无>
	权重
	<无>
	拆分文件
	<无>
工作数据文件中的行数	20

理	缺失值处	对缺失的定义	将用户定义的缺失值视为缺失。
		使用的个案数	统计基于所有具有有效数据的个案。
语法			FREQUENCIES VARIABLES=VAR00014 /STATISTICS=RANGE MINIMUM MAXIMUM MODE /ORDER=ANALYSIS.
资源		处理程序时间	00:00:00.00
		耗用时间	00:00:00.03

[数据集 0]

统计

VAR00014

个 案 有	20
数 效	
缺 失	0
众数	71.50 ^a
范围	12.90
最小值	65.30
最大值	78.20

a. 存在多个众数。显示了最小的值。

VAR00014

		频率	百分比	有效百分比	累积百分比
有效	有 65.30	1	5.0	5.0	5.0
	65.40	1	5.0	5.0	10.0
	66.00	1	5.0	5.0	15.0
	68.10	1	5.0	5.0	20.0
	68.20	1	5.0	5.0	25.0
	68.30	1	5.0	5.0	30.0

68.	1	5.0	5.0	35.0
80				
71.	1	5.0	5.0	40.0
20				
71.	2	10.0	10.0	50.0
50				
71.	1	5.0	5.0	55.0
60				
72.	1	5.0	5.0	60.0
00				
72.	1	5.0	5.0	65.0
10				
72.	1	5.0	5.0	70.0
20				
72.	2	10.0	10.0	80.0
60				
74.	1	5.0	5.0	85.0
00				
74.	1	5.0	5.0	90.0
50				
75.	1	5.0	5.0	95.0
80				
78.	1	5.0	5.0	100.0
20				
总	20	100.0	100.0	
计				

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT y

/METHOD=ENTER x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12.

回归

备注

已创建输出	20-JUN-2021 03:39:15
注释	
输入	活动数据集
	数据集 0
	过滤器
	<无>
	权重
	<无>
	拆分文件
	<无>
	工作数据文件中的行
数	20

理	缺失值处	对缺失的定义	将用户定义的缺失值视为缺失。
		使用的个案数	统计基于那些对于任何所用变量都不具有缺失值的个案。
语法			REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT y /METHOD=ENTER x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12.
资源	处理程序时间		00:00:00.03
	耗用时间		00:00:00.05
	所需内存量		11376 字节
	残差图需要更多内存		0 字节

输入/除去的变量^a

模 型	输入的变量	除去的变 量	方法
1	x12, x10, x2, x8, x1, x9, x7, x6, x4, x3, x5, x11 ^b	.	输入

a. 因变量: y

b. 已输入所请求的所有变量。

模型摘要

模 型	R	R 方	调整后 R 方	标准估算的 错误
1	.930 ^a	.865	.634	2.07302

a. 预测变量: (常量), x12, x10, x2, x8, x1, x9, x7, x6, x4, x3, x5, x11

ANOVA^a

模型		平方和	自由 度	均方	F	显著性
1	回	192.94	12	16.079	3.742	.045 ^b
	归	8				
	残	30.082	7	4.297		
	差					

总	223.02	19			
计	9				

a. 因变量: y

b. 预测变量: (常量), x12, x10, x2, x8, x1, x9, x7, x6, x4, x3, x5, x11

系数^a

模型		未标准化系数		标准化	t	显著性
		B	标准 错误	Beta		
1	(常量)	58.155	14.271		4.075	.005
	x1	.000	.001	-.148	-.474	.650
	x2	.041	.028	.619	1.479	.183
	x3	-.009	.012	-.268	-.762	.471
	x4	.053	.280	.052	.189	.855
	x5	.759	.267	1.400	2.841	.025
	x6	.023	.026	.232	.885	.405
	x7	.029	.035	.284	.816	.442
	x8	.011	.004	.596	2.612	.035
	x9	-.155	.155	-.346	-1.004	.349
	x10	-.153	.078	-.438	-1.966	.090
	x11	-22.21	21.995	-.717	-1.010	.346
	5					
	x12	-.039	.049	-.176	-.789	.456

a. 因变量: y