全国大学生数学建模竞赛
湖南赛区优秀论文集
(2012)

全国大学生数学建模竞赛湖南赛区组委会 编
吴孟达 主编

清华大学出版社

全国大学生数学建模竞赛
湖南赛区优秀论文集
(2012)

全国大学生数学建模竞赛湖南赛区组委会 编
吴孟达 主编

清华大学出版社
北京

内容简介

本书收录了2012年全国大学生数学建模竞赛湖南赛区获得全国一、二等奖的部分优秀论文。这些论文分别围绕“葡萄酒的评价”、“太阳能小屋的设计”、“脑卒中发病环境因素分析及干预”和“机器人避障问题”等四个实际问题展开研究，从不同角度出发，综合运用多种数学方法，建立了各具特色的数学模型。为了保持论文原貌，本书只做了符号和文字上的订正，没有进行大的改动。同时，每篇论文都附有指导教师点评。

本书可作为高等院校数学建模课程的参考用书，也可作为数学建模竞赛的培训资料。

图书在版编目(CIP)数据

全国大学生数学建模竞赛湖南赛区优秀论文集. 2012/全国大学生数学建模竞赛湖南赛区组委会编. --北京：清华大学出版社，2013

ISBN 978-7-302-31867-5

Ⅰ. ①全… Ⅱ. ①全… Ⅲ. ①数学模型－文集 Ⅳ. ①O22－53

中国版本图书馆CIP数据核字(2013)第072017号

责任编辑：石　磊　洪　英
封面设计：傅瑞学
责任校对：王淑云
责任印制：何　芊

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦A座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：三河市金元印装有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　张**：14.75　　**字　数**：356千字
版　　次：2013年6月第1版　　**印　次**：2013年6月第1次印刷
印　　数：1～3000
定　　价：28.50元

产品编号：050961-01

编　委　会

FOREWORD 前言

数学建模是连接数学与实际问题的桥梁。数学建模在应用数学方法解决各类实际问题过程中的作用越来越重要，已经成为现代应用数学的一个重要领域。培养大学生数学建模这一必备的技能和素质，对于培养高质量、高层次人才具有重要意义。

由教育部高等教育司和中国工业与应用数学学会主办的全国大学生数学建模竞赛提供了一个实战平台，参赛同学有机会亲身体验用数学方法解决一个具体实际问题的创造和发现的全过程。在这个过程中，像实际的科研活动一样，不但要发挥个人的主观能动性和创造力，还要全队密切配合，分工明确，协同工作，这样才有可能高质量地完成解答。该项竞赛已有二十多年的历史，一直深受广大师生的欢迎，不断向前发展，目前是全国高校中规模最大的课外科技活动，有力地推动了大学数学教育改革。

根据《关于组织举办2012年全省普通高校大学生学科竞赛活动的通知》(湘教通[2012]264号)和全国大学生数学建模竞赛组委会的有关要求，湖南省教育厅于2012年9月7日至10日组织举办了湖南省大学生数学建模竞赛暨全国大学生数学建模竞赛湖南赛区的比赛。全省共有45所普通高校的633个代表队参赛。参赛同学围绕“葡萄酒的评价”、“太阳能小屋的设计”、“脑卒中发病环境因素分析与干预”和“机器人避障问题”等四个实际问题展开研究。在竞赛过程中，同学们的聪明才智和创新精神得到了充分的发挥，提交了不少出色的答卷，涌现出一批优秀的参赛队。经竞赛组委会组织专家评审并报送全国组委会评审，共获得全国一等奖8个，二等奖52个。

本书收录了湖南赛区本科组全国一等奖论文6篇，全国二等奖论文4篇，专科组全国一等奖论文2篇，全国二等奖论文1篇(排名不分先后)。这些论文分别从不同角度出发，综合运用多种数学方法，建立了各具特色的数学模型，很好地完成了题目中提出的各项要求。本书既可以作为高等院校数学建模课程的参考用书，也可以作为数学建模竞赛培训资料。希望本书的出版，能够让更多的大学生、教师以及关心数学建模竞赛的人们了解这项赛事，并给准备参加竞赛的同学们以适当的启发。同时，对于有意继续研究这些问题的师生，本书也是一本很好的参考书。为保持论文原貌，本书只做了符号和文字上的订正，没有进行大的改动。同时，每篇论文都附有指导教师点评。

本书在编辑出版过程中，得到了各参赛院校和清华大学出版社的大力支持，在此表示衷心感谢。

全国大学生数学建模竞赛
湖南赛区组委会
2013年5月

CONTENTS 目录

2012年全国大学生数学建模竞赛A题：葡萄酒的评价

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评分，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1给出了某一年份一些葡萄酒的评价结果，附件2和附件3分别给出了该年份这些葡萄酒和酿酒葡萄的成分数据。请尝试建立数学模型讨论下列问题。

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？

问题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。

问题3：分析酿酒葡萄与葡萄酒的理化指标之间的联系。

问题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。

附件1：葡萄酒品尝评分表

附件2：葡萄和葡萄酒的理化指标

附件3：葡萄和葡萄酒的芳香物质

（附件略，请到 http://www.shumo.com 下载）

基于理化指标的葡萄酒质量评价

刘阳洋，兰天鹏，余奇

指导教师：指导教师组

（国防科技大学 湖南长沙 410073）

摘要

本文通过对葡萄酒评酒分数以及葡萄酒和葡萄的理化指标的统计数据进行多元统计分析，采用经验建模加模型验证，即首先通过对统计数据的分析得到已知的经验，然后把这些经验再代入数据中进行模型的验证，综合运用多种统计分析方法研究酿酒葡萄与葡萄酒理化指标的统计意义，分析理化指标的相关性，筛选出主要影响指标，并建立对葡萄酒的质量评定的客观评价体系，得到了令人满意的结果。

针对问题1，我们建立了统一的评酒员可信因子计算方法，然后计算出每组评酒员可信因子的期望和方差，由于第二评酒小组成员可信因子相对第一评酒小组成员可信因子的期望高而方差低，因此我们判定第二评酒小组的结果更可信，并且两评酒小组评酒结果有显著性差异。

针对问题2，利用改进的主成分分析法，从葡萄的30个一级理化指标中提取出主成分，主成分很好地保留了理化指标所携带的信息。然后我们建立了一个综合评价函数，通过该评价函数给出葡萄样品基于理化指标的客观评分，再结合评酒员给出的主观评价得分，对葡萄样品进行分级。我们得到的结果是：

1、2、3、8、9、23号红葡萄样品为一级；14、21号红葡萄样品为二级；5、6、7、15、16、17、19、22、24号红葡萄样品为三级；4、10、11、12、13、18、20、25、26、27号红葡萄样品为四级。

5、21、22、23、27、28号白葡萄样品为一级；2、3、4、6、9、10、12、14、17、20、24、25、26号白葡萄样品为二级；7、8、11、15、18、19号白葡萄样品为三级；1、13、16号白葡萄样品为四级。

针对问题3，我们首先考虑能否用葡萄的理化指标来线性拟合出葡萄酒的理化指标，对于不能很好地用线性组合近似的葡萄酒的理化指标，我们采用相关性分析得出与之相关性较大的非线性指标，通过非线性指标的引入来更好地解释葡萄酒的理化指标和酿酒葡萄之间的关系。结果我们发现红葡萄酒9个一级理化指标中有7个能够很好地用葡萄的理化指标线性表出，其余两个由于受到非线性因素的影响只能部分确定它们之间的联系。

针对问题4，我们选用Fisher型逐步判别法筛选出酿酒葡萄和葡萄酒的关键理化指标，然后建立了量化的判别方程式，通过显著性检验及回代准确率论证了此方案的可行性。对于红葡萄酒我们提取了总酚、还原糖、果梗质量、出汁率4个理化指标，分别得到了三级葡萄酒与四级葡萄酒的判别方程：

$$Y_3 = 0.516X_{11} + 0.256X_{17} + 0.036X_{23} + 1.002X_{26} - 73.768$$

$$Y_4 = -0.1X_{11} + 0.204X_{17} + 0.018X_{23} + 1.239X_{26} - 65.315$$

用这个判别方程对红葡萄酒质量评级的准确率达到 88.9%，基于这些工作我们说明了利用理化指标建立质量评价体系的可行性，并对葡萄酒厂家提出了建议。

关键词：可信因子　改进的主成分分析法　线性建模　Fisher 型逐步判别法

1. 问题重述

为了确定葡萄酒质量，一般通过聘请一批有资质的评酒员进行品评。根据评酒员给出的总分，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。根据附件所给数据，建立数学模型讨论下列问题。

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。

问题 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。

问题 3：分析酿酒葡萄与葡萄酒的理化指标之间的联系。

问题 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。

2. 模型假设

(1) 假设评酒员评价不同的酒是无差别的。

(2) 不考虑葡萄酒加工工艺的差异，即题目中葡萄酒样品的质量完全由酿酒葡萄决定。

(3) 假设表中所给的统计数据具有一定的代表性。

(4) 假设问题中设定的理化指标都可以反映实际情况。

(5) 假设酒的一个方面的指标不会影响评酒员对另一方面指标的评判。

3. 符号说明

μ_i：评分员对 i 类指标评分样本总体的均值。

σ_i：评分员对 i 类指标评分样本总体的方差。

b：评酒员的打分可信因子。

b_i：第 i 个评酒员的可信因子。

I_i：葡萄中第 i 个原始理化指标的影响力因子。

$\boldsymbol{X}_i$：第 i 个红葡萄样品的理化指标向量，$i=1,2,\cdots,27$。

$\boldsymbol{Y}_i$：第 i 个白葡萄样品的理化指标向量，$i=1,2,\cdots,28$。

x_{ij}：第 i 个红葡萄样品的第 j 项理化指标，$i=1,2,\cdots,27$；$j=1,2,\cdots,30$。

y_{ij}：第 i 个白葡萄样品第 j 项理化指标，$i=1,2,\cdots,28$；$j=1,2,\cdots,30$。

4. 问题分析

4.1 背景分析

传统的葡萄酒的评价的方式是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。这样的评价方式不免有些过于主观化，由于检测生产出来的葡萄酒的一些理化指标在实际生产过程中是很容易做到的，所以自然会想到能不能用葡萄酒的一些理化指标来对生产出来的葡萄酒进行初步的分级，然后再用评酒员评定的方法进行更精细的分级。而同时又知道酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系，所以在评价葡萄酒的好坏的时候不仅要考虑葡萄酒的理化指标，还要考虑酿酒葡萄的理化指标，几个指标间和与葡萄酒质量之间的待明确的关系如图 1 所示。

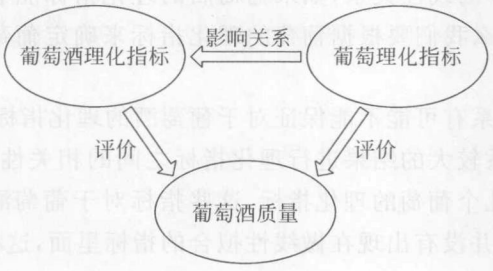

图 1　待明确的关系

这样自然产生了题中所给的问题，用葡萄的理化指标和葡萄酒的理化指标来估计葡萄酒的质量，从而使得葡萄酒的初步评价更加便于操作，对于葡萄酒的实际生产过程和酿酒葡萄的采购过程有极大的指导意义。

4.2 各问的分析

4.2.1 问题 1

由于不同的葡萄酒样品或者相同的葡萄酒样品的不同评分指标没有可比性，因此只能比较同一葡萄酒样品的同一评分指标。

为确定两组评酒员之间是否有显著性差异以及哪一组评酒员的评价结果更加可信，我们通过建立根据评价结果判断每个评酒员的可信程度的标准，得到每组每一位评酒员的可信因子后，比较两组评酒员可信度的均值和方差判断两组评酒员是否有显著性差异并确定两组评酒员哪组更加可信。

4.2.2 问题 2

涉及对酿酒葡萄的分级，我们需要考虑的主要有两方面因素的影响，一是葡萄的各项理化性质，如果皮颜色、含糖量等；二是用此类葡萄所酿造的葡萄酒的质量。我们所要完成的工作是综合葡萄酒的质量和葡萄的理化指标对于葡萄品质的依赖，试图给出基于葡萄理化指标进行评级的参考标准。

具体做法是利用问题1中我们确定的可信度较高的一组评分数据，由问题1的分析比较过程，我们有理由相信它能够在一定程度上反映葡萄酒质量在客观上的差异，这组评分数据可以作为我们建立所谓的基于理化指标的评级标准体系的一个重要参考。从题目的附件2中所给的各项理化指标我们抽取了全部一级指标来描述酿酒葡萄的品质差异，这些理化指标已经足以反映葡萄在各个方面的特点，尽量降低了分级模型的复杂度。利用多元统计中主成分分析的经典方法对这些指标进行分析，找出影响葡萄分级结果的最主要的理化指标，并研究将葡萄理化指标作为葡萄分级的直接标准的方法，通过数据检验论证该标准的科学性。

4.2.3 问题3

在假设存储和加工过程相同的情况下，我们可以大致认为所给的葡萄酒样品的各个理化指标只与酿造该葡萄酒的酿酒葡萄的理化指标密切相关，这样讨论葡萄酒的理化指标和葡萄的理化指标之间的定量关系才变得有意义。

首先我们考虑最简单的线性关系，如果葡萄酒的理化指标能够表示成为某几个葡萄的理化指标的线性组合，那么我们要根据葡萄的理化指标来确定葡萄酒的理化指标就变得简单、方便。

但是，简单的线性关系有可能不能保证对于葡萄酒的理化指标的精确刻画，所以我们对以上得到的偏离线性关系较大的结果进行理化指标之间的相关性分析，找到与这个葡萄酒的理化指标显著相关的几个葡萄的理化指标，这些指标对于葡萄酒的理化性质会产生显著影响，同时它们有些指标并没有出现在做线性拟合的指标里面，这样只考虑线性拟合显然会产生比较大的偏差。

我们采用的线性建模只是对现有的数据进行分析，从而找到葡萄酒的理化指标与葡萄的理化指标之间的经验关系，属于一种经验建模，但是它使得用葡萄的理化指标能够很容易地判断葡萄酒的理化指标的范围，使得对葡萄酒的进一步评价能够对应到对葡萄的理化指标的检测上来，对实际酒商购进原料酿酒葡萄提供一定的参考。

4.2.4 问题4

在背景分析中我们看到，问题4的解决将有很高的应用价值，而问题2、问题3的求解对问题4也会有很大帮助，基于对问题实际应用背景的理解，问题4的目标是试图建立一个根据酿酒葡萄与葡萄酒的理化指标判别葡萄酒质量的客观评价体系，在问题1中我们将确定出更可信的一组评酒员，有理由相信这个主观评分可以作为葡萄酒质量的评价，因此我们以国际通行的帕克评价体系作为葡萄酒质量的判别标准，对葡萄酒按质量进行分级。

在问题2、问题3中我们将分别得到影响葡萄的几个关键理化指标和这些指标与葡萄酒理化指标之间的相关性，利用这些相关性可以初步确定出一个关键指标集，然后采用Fisher型逐步判别法从关键指标集中筛选出因子并建立判别方程式，这个判别方程可以作为量化的评价指标。通过检验判别方程式的显著性和回代正确率来论证该评价体系的可行性。

5. 模型的建立和求解

5.1 问题 1

5.1.1 数据的分析

对于第一组第 4 号评酒员对第 20 号酒样的色调数据的缺失，以及第一组第 7 号评酒员对第 3 号酒样持久性数据异常(超出总分)，第一组第 9 号评酒员对第 8 号酒样持久性数据异常(超出总分)，我们考虑使用相应酒样相应指标的其余 19 个评酒员打分的平均值作为该处的分数，以消除该位置的数据对于后续计算的影响。

由于不同的酒样以及相同酒样的不同指标的评分相互之间没有可比性，并不能将两小组对不同酒样或相同酒样的不同指标的评分进行合并比较，而对同一酒样的同一评分每组只有 10 个分值的数据，相对而言，数据量太少，且对每组评分员的 10 项评分经过正态性检验，得到大部分的 10 项评分不符合正态分布(见附件表)，因此不能用两总体正态分布均值差的检验，或者单因子二水平方差分析。

5.1.2 基于可信度的组间评价策略模型建立

1) 可信度标准的建立

我们注意到，葡萄酒评价的每一个单独的指标都有各自的评价总分，各项指标之间的分数没有相互比较的意义，只能把不同评酒员的评分进行横向比较。而现在要建立一个全面统一的可信度标准来评价评酒员的可信度，就要使得全部的评分都具有相互比较的意义，所以我们考虑使用相对得分(标准化得分)——用每一项得分除以每项的总分数，得到的相对的值，也即

$$r_{p_i} = \frac{p_i}{P_i}$$

其中，r_{p_i} 是每一个值的相对分数；P_i 是所对应指标的总分数。

这样我们就把整个分数表统一化了，也就可以进行进一步的比较工作。同时我们注意到一个很重要的方面：红葡萄酒和白葡萄酒的评价具有相同的评价指标和相同的总分，所以在考虑相对得分的情况下红葡萄酒和白葡萄酒是统一起来的。

按照以上定义的相对分数的概念，每一名评酒员相当于给出了 550 个评价分数(红 270 个，白 280 个)，记为 $p_i(i=1,2,3,\cdots,550)$，现在要根据这 550 次的评判来确定这个评酒员的评酒的可信程度，那么我们就来分析这 550 次评判结果。定义每一个评酒员的评判分数向量：

$$\boldsymbol{p}_i = \{r_{p_1}, r_{p_2}, r_{p_3}, \cdots, r_{p_{550}}\}, \quad i = 1,2,3,\cdots,10$$

以及评分的标准向量：

$$\boldsymbol{s} = \{\overline{p_1}, \overline{p_2}, \overline{p_3}, \cdots, \overline{p_{550}}\}$$

其中，$\overline{p_i}$ 为所对应项得到的 20 位评酒员所给分数的平均分。

2) 可信因子的定义

分析统计数据，确定可信阈值 w_0，只要相对分数的残差平方值 $S_i=(r_{p_i}-\overline{p_i})^2$ 大于此阈值，那么此对应位置的可信因子值就为 0，而当 $S_i=0$ 时，可信因子的值就为 1；而可信阈

值是在设定置信比率 $\beta=90\%$，也即能够使占总数 90%以上残差平方满足条件 $S_i \leqslant w_0$ 时的 w_0 值。

这样，每一个评酒员的可信因子：

$$b_j = 1 - \frac{\sum_{i=1}^{550} (S_i - w_0)^2}{550 w_0}, \quad j = 1,2,\cdots,20$$

由以上过程我们得到了每个评酒员的可信因子 b_j。

3）显著性差异分析

由于上述可信因子计算中采用的是 20 个人的总平均分，两组的可信度计算根据相同的标准进行，所以可以根据两组成员的可信度的均值判断两组评酒员的评分是否有显著性差异。

4）团体可信度评价

根据以上建立的可信度的评价标准，得到每组中每位评酒员的可信因子，分析所得数据，计算每组评酒员可信因子的期望与方差，期望值越高说明该组评酒员的可信度越高，方差越小说明该组评酒员的水平接近，不会出现较大的误判。我们根据可信因子的期望与方差对两个组的评酒能力进行整体上的比较。

5.1.3 模型求解

针对以上建立的模型，我们采用 C++编程一次性处理大量的数据信息，按照既定的模型对两组结果进行显著性分析，以及将所有分数标准化，然后计算每个评酒员的可信因子值。

在置信比率 $\beta=90\%$ 时，求得的可信阈值为 $w_0=0.17$，求得的各个评酒员的可信因子如表 1 所示（其中 R_{ij} 表示第 i 组第 j 个评酒员）。

表 1 评酒员可信因子

编号	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}
可信因子	0.941	0.908	0.928	0.924	0.930	0.919	0.933	0.945	0.917	0.949
编号	R_{20}	R_{21}	R_{22}	R_{23}	R_{24}	R_{25}	R_{26}	R_{27}	R_{28}	R_{29}
可信因子	0.955	0.944	0.942	0.942	0.940	0.939	0.940	0.928	0.951	0.939

求得两组评酒员可信因子的期望分别为：$\bar{b}_1=0.929$，$\bar{b}_2=0.942$，方差分别为 $S_1=1.5\times10^{-4}$，$S_2=4.7\times10^{-5}$。

组二的期望高于组一，同时组二的方差也相应地低于组一，说明组二的可信度较组一要高。由于一、二组的期望有显著不同，所以我们判定两组评分有显著性差异。

5.1.4 结果分析及检验

对于问题 1 的模型，其中涉及一个可变的参数——可信阈值，可信阈值的大小随着置信比率 β 设定值的不同而发生变化，我们分别取得不同的置信比率，求得最终每个人的可信因子如下。

$\beta=60\%$，$w_0=0.05$ 时，见表 2。

表 2 可信阈值为 0.05 时评酒员可信因子

编号	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}
可信因子	0.961	0.956	0.957	0.960	0.961	0.957	0.960	0.964	0.957	0.963
编号	R_{20}	R_{21}	R_{22}	R_{23}	R_{24}	R_{25}	R_{26}	R_{27}	R_{28}	R_{29}
可信因子	0.964	0.964	0.962	0.962	0.961	0.961	0.961	0.961	0.965	0.962

同理，求得两组评酒员可信因子的期望分别为：$\bar{b}_1=0.960$，$\bar{b}_2=0.962$，方差分别为 $S_1=6.2\times10^{-6}$，$S_2=1.6\times10^{-6}$。

$\beta=70\%$，$w_0=0.09$ 时，见表 3。

表 3 可信阈值为 0.09 时评酒员可信因子

编号	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}
可信因子	0.944	0.931	0.937	0.940	0.942	0.935	0.942	0.950	0.934	0.950
编号	R_{20}	R_{21}	R_{22}	R_{23}	R_{24}	R_{25}	R_{26}	R_{27}	R_{28}	R_{29}
可信因子	0.952	0.949	0.947	0.947	0.945	0.945	0.944	0.943	0.953	0.945

求得两组评酒员可信因子的期望分别为：$\bar{b}_1=0.940$，$\bar{b}_2=0.947$，方差分别为 $S_1=3.6\times10^{-5}$，$S_2=9.9\times10^{-6}$。

$\beta=80\%$，$w_0=0.11$ 时，见表 4。

表 4 可信阈值为 0.11 时评酒员可信因子

编号	R_{10}	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}	R_{17}	R_{18}	R_{19}
可信因子	0.940	0.923	0.931	0.933	0.936	0.928	0.937	0.946	0.926	0.947
编号	R_{20}	R_{21}	R_{22}	R_{23}	R_{24}	R_{25}	R_{26}	R_{27}	R_{28}	R_{29}
可信因子	0.950	0.945	0.943	0.944	0.941	0.941	0.940	0.937	0.950	0.941

求得两组评酒员可信因子的期望分别为：$\bar{b}_1=0.935$，$\bar{b}_2=0.943$，方差分别为 $S_1=5.9\times10^{-5}$，$S_2=1.7\times10^{-5}$。

分析上面改变参数可信阈值得到的三组数据，我们发现：

(1) 对于每一组来说，第二组的可信因子期望都高于第一组，并且第二组的方差低于第一组，也就是说第二组的评价更加可信；

(2) 随着置信比率 β 的减小，可信因子的期望值增大，二者不断接近，同时方差逐渐减小。

第一个结论就证明了，不管我们的可变参数取何种值，得到的结论是统一的，也就是第二组评酒员比第一组评酒员得到的结论更可信，同时说明我们采用的问题 1 的模型是完全正确的；第二个结论表示的是一种趋势，当置信比例减少，所取的数据的量减少，分布就更集中于平均值，每个评酒员的可信度就更接近，极限情况下所有评酒员的可信度都相等，反映在两组之间的差距在不断缩小，这也是符合客观规律的。

5.2 问题 2

5.2.1 改进型主成分分析法模型

1）酿酒葡萄理化指标的提取

基于题目中所给的理化指标特性数量太多，我们舍弃了葡萄的二级指标，选取葡萄的一级理化指标作为样品的属性值。分别为氨基酸总量 W_{amino}、蛋白质总量 $W_{protein}$、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、总酚、单宁、葡萄总黄酮、白藜芦醇、总糖、还原糖、可溶性固形物、pH、可滴定酸、固酸比、干物质含量、果穗质量、百粒质量、果梗比、出汁率、果皮质量、果皮颜色。对多次测定的理化特性，我们取平均值。

传统的主成分分析是一种线性降维技术，但本文中葡萄的各个理化指标呈现非线性，主成分分析的降维效果不理想，甚至出现评价偏差很大的结果。为此，我们通过对传统主成分进行改进，使其适用于非线性数据。

2）数据的线性化改进

在对数据进行标准化处理之前，为防止矩阵中有的数据为非正数，可以将所有数据加上一个略小于最小负数的相反数，这样平移不会改变结果，按平移后的矩阵进行如下对数变换：

$$x_{ij}^{*} = \ln(x_{ij} - x_{\min} + \Delta), \quad y_{ij}^{*} = \ln(y_{ij} - y_{\min} + \Delta)$$

这里 $x_{\min} = -1.567$，$y_{\min} = -6.068$，$\Delta = 0.001$。

通过对红葡萄的计算，我们将原始数据与线性化处理后的原始数据进行对比，可以看到在累积贡献率方面，传统方法要选择前 8 个才能达到 83%，而改进的主成分分析法只需要选择前 7 个就能达到 84%以上的累积贡献率。同时改进前第一主成分的贡献率为 23.2%，改进后的第一主成分的贡献率达到 38.1%，几乎是传统方法前两个主成分之和。这说明对初始数据进行线性化处理具有一定的优越性。改进前后的主成分累积贡献率对比见表 5。

表 5 改进前后的主成分累积贡献率对比

成分	改进前			改进后		
	合计	方差的百分比/%	累积百分比/%	合计	方差的百分比/%	累积百分比/%
1	6.966	23.221	23.221	11.434	38.112	38.112
2	4.941	16.469	39.690	3.893	12.978	51.091
3	3.736	12.452	52.142	3.265	10.882	61.972
4	2.840	9.467	61.609	2.206	7.355	69.327
5	2.000	6.665	68.274	1.774	5.912	75.240
6	1.742	5.807	74.082	1.397	4.657	79.896
7	1.418	4.728	78.809	1.258	4.195	84.091
8	1.270	4.233	83.042	0.825	2.751	86.842
9	0.961	3.204	86.246	0.728	2.426	89.268
10	0.738	2.462	88.708	0.647	2.157	91.425
11	0.691	2.302	91.010	0.540	1.799	93.225
12	0.514	1.713	92.724	0.504	1.679	94.904
⋮	⋮	⋮	⋮	⋮	⋮	⋮

3）酿酒理化指标主成分的提取与综合评价函数

主成分分析法是一种降维的统计方法，它的工作目标是在力求数据信息丢失最少的原则下，对高维变量空间进行降维处理，在降低计算复杂度的同时又不失计算的准确性。参照文献[4]，它的主要步骤如下：

（1）为消除量纲的影响，首先需要将原始数据进行标准化。以红葡萄为例，第 i 类红葡萄样品的理化指标向量为 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{ip})$，$i=1,2,\cdots,n$，对数据进行如下的标准化变换：

$$z_{ij}=\frac{x_{ij}-\bar{x}_{ij}}{s_j},\quad i=1,2,\cdots,n;\ j=1,2,\cdots,p$$

其中，$\bar{x}_j$ 为样本平均值，$\bar{x}_j=\dfrac{\sum_{i=1}^{n}x_{ij}}{n}$；$s_j$ 为标准差，$s_j=\sqrt{\dfrac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1}}$；于是得到标准化矩阵 $\boldsymbol{Z}=[z_{ij}]_{n\times p}$。

（2）求出标准化矩阵 $\boldsymbol{Z}$ 的相关系数矩阵

$$\boldsymbol{R}=\frac{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}}{n-1}$$

（3）解相关矩阵 $\boldsymbol{R}$ 的特征方程，得到 p 个特征根 $\lambda_1,\lambda_2,\cdots,\lambda_p$，计算各主成分的方差贡献率和累积贡献率，用 g_i 表示方差贡献率，k_i 表示累积贡献率。

$$g_i=\frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i},\quad k_i=\sum_{n=1}^{i}g_n$$

根据累积贡献率的大小在保证数据累积贡献率超过 80% 的前提下，选取最少的 m 个主成分。

（4）构造综合评价函数，即综合评价指标。评价函数可表示为

$$F=\sum_{i=1}^{m}g_iq_i$$

其中，$q_1,q_2,\cdots,q_m$ 为提取出来的主成分值。

5.2.2 模型求解

正文中我们以红葡萄为例研究理化指标与葡萄之间的关系，主成分分析中每个原始指标对主成分的信息量提供反映在主成分载荷矩阵（见附表 1 和附表 2）中。从表中可知：

（1）氨基酸、蛋白质、花色苷、酒石酸、苹果酸、柠檬酸、多酚氧化酶、褐变度、DPPH 自由基、总酚、单宁、葡萄总黄酮、白藜芦醇、黄酮醇、总糖、果皮颜色在第一主成分中有较高的载荷，说明第一主成分基本反映了这 16 个指标。

（2）维生素 C 含量、还原糖、可溶性固形物、可滴定酸、估算比、干物质质量在第二主成分中有较高的载荷，说明第二主成分基本反映了这 6 个指标。

（3）果穗质量、百粒质量、果梗比、果皮质量在第三主成分中有较高的载荷，说明第三主成分基本反映了这 4 个指标。

（4）果皮颜色在第四主成分中有较高的载荷。

（5）pH 和固酸比在第五主成分中有较高的载荷。