葡萄酒的质量分析及评价

摘 要

本文主要对两组评酒员的评价结果及可信度、酿酒葡萄的分级、酿酒葡萄与葡萄酒的理化性质之间的联系和是否影响葡萄酒的质量进行分析及研究。

针对问题一，利用附件一中评酒员群体对红、白葡萄酒进行两次评分的数据，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。经过检验，发现样本数据符合正态分布，则采用配对t检验法对每种酒的最终得分进行分布检验，最后发现两组评酒员评价结果具有显著性差异。而对于两组评酒员评价结果可靠性的判断，我们可以选择标准差来反映，标准差越小，可靠性越大。最后发现第二组评酒员的的评价结果更可信。

针对问题二，采用第二组评酒员对葡萄的评价结果作为依据，通过聚类分析法以组距为标准对葡萄酒划分为四个等级，然后以酿酒葡萄的理化指标为辅助，综合两者的结果来对酿酒葡萄进行分级，再通过对理化指标的分析将红葡萄和白葡萄各分为优、良、中等、差四级。

针对问题三，由于葡萄的理化指标众多，我们先使用了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的第一步筛选。接着利用多元线性回归的方法拟合了葡萄酒理化指标与葡萄理化指标间对多的函数关系，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，首先要求分析葡萄和葡萄酒理化指标对葡萄酒质量的影响，我们通过分析葡萄酒理化指标与葡萄酒质量的函数关系，并利用第三问的结论，说明了葡萄与葡萄酒的理化指标只在一定程度上对葡萄酒质量有影响。问题第二部分要求我们论证能否用理化指标来评价葡萄酒，我们得出的结论为:不能仅仅通过葡萄与葡萄酒的理化指标对葡萄酒质量进行评价。

关键词：正态分布　　配对t检验　　标准差　　聚类分析　　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、中等、不合格四个级别。

三、 符号说明

符号	符号说明
α	显著性水平
d_i	配对样本的偏差
$\bar{d}$	配对样本的偏差的平均值
s_d	配对样本差值的标准偏差
n	配对样本数
X_{ij}	第j号评酒员对第i号酒样的评价分数
x_{ij}	理化指标的标准化数据

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、 模型的建立及求解

5.1 问题一模型的建立和求解

5.1.1 第一部分：检验显著性差异

一般来说，在两个样本显著性差异检验时，常用的方法是对试验的样本均值进行参数检验，如方差分析等。然而，这些检验方法需要明确样本总体所服从的分布，如正态分布、二项分布等，并且要求方差齐性。并且，由于统计规律表明，正态分布有极其广泛的实际背景，生产与科学实验中很多随机变量的概率分布都可以近似地用正态分布来描述，对葡萄酒质量的评分进行正态性检验有助于我们分析得出该评分是否科学、合理。因此我们先需要判断两组样本是否满足正态分布。

a.正态分布的检验

（1）绘制两组频数分布图进行分布初步分析

通过附件 1 得到葡萄酒质量评价数值后，对葡萄酒质量及其对应的酒样品数目分布进行分析，通过软件绘制出红、白葡萄酒在两组评酒员评价结果下的评分分布直方图：

得到第一组红酒和第二组红酒的评数分布图

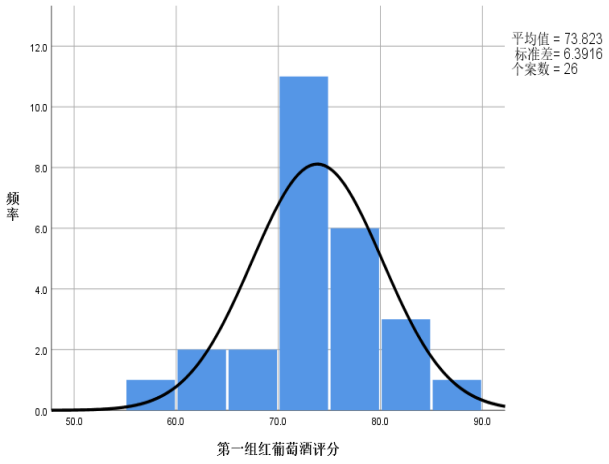


图 1：第一组红葡萄酒品尝得分

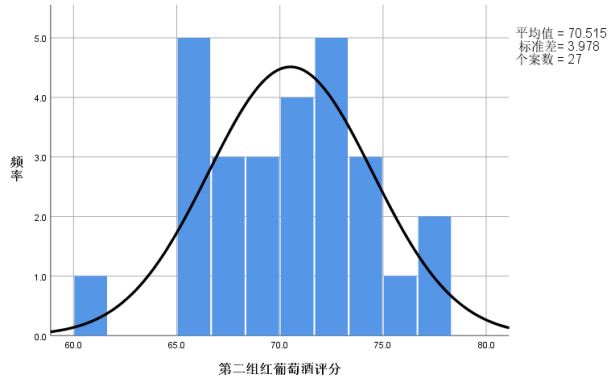


图 2：第二组红葡萄酒品尝得分

得到第一组白酒和第二组白酒的评数分布

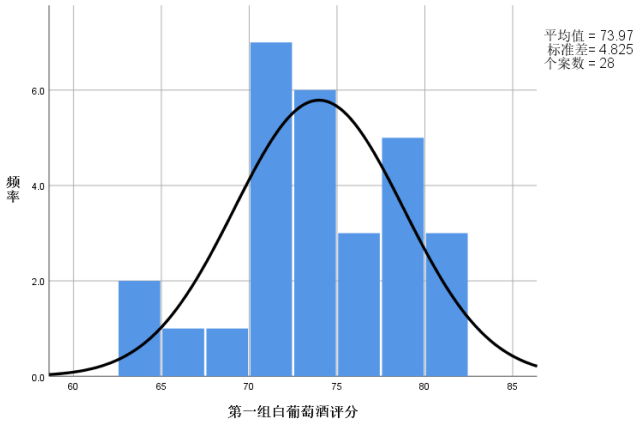


图 3：第一组白葡萄酒品尝得分

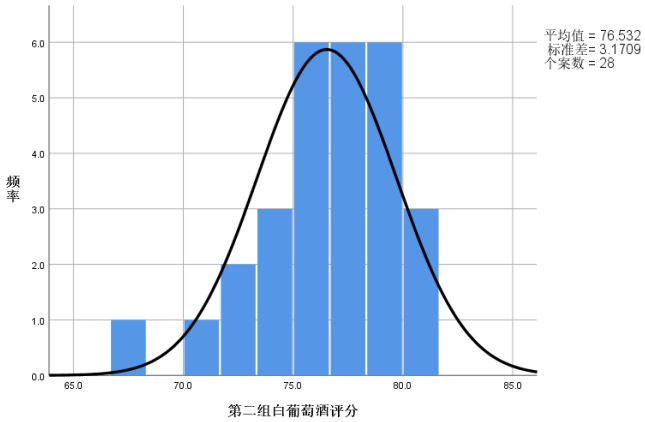
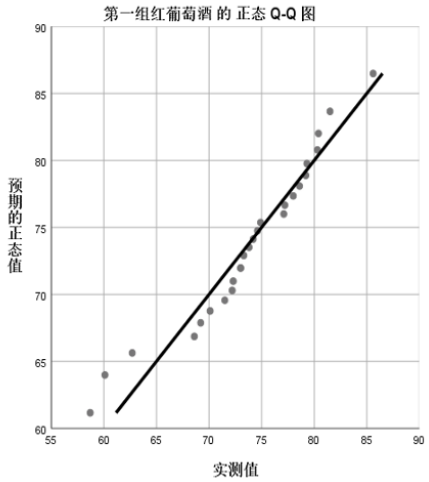


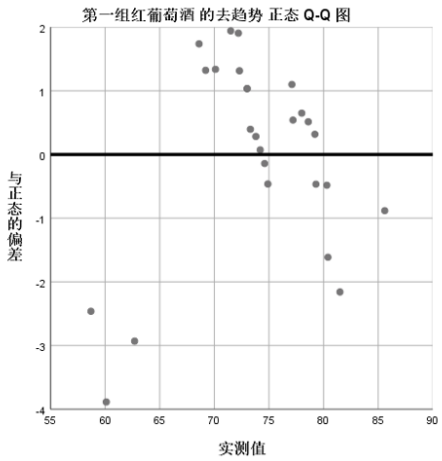
图 4：第二组白葡萄酒品尝得分

通过对图像观察分析，可以大致预测葡萄酒样品质量及其对应的数量分布呈现正态分布。接下来进行数据分布的正态性检验。

（2）基于 Q-Q 图及 K-S 检验检验正态分布

利用 SPSS 统计软件中的 Q-Q 图及单样本 K-S 检验，对数据两组品酒员分别对红、白葡萄酒品尝得到的四组评价结果进行了正态分布检验，若样点在正态分布 Q-Q 图上呈直线散布，则被检验数据基本上成一条直线。





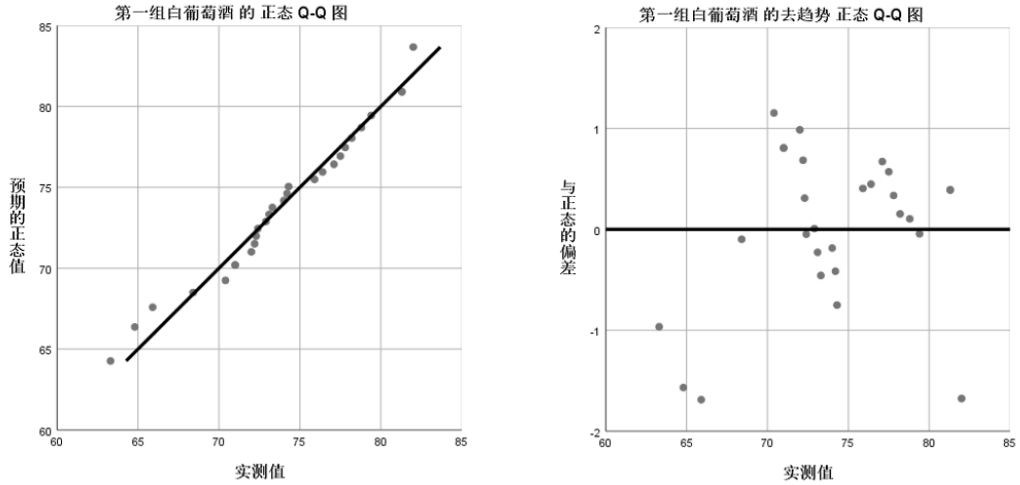


图 5：第一组红、白葡萄酒的正态 Q-Q 图

单样本柯尔莫戈洛夫-斯米诺夫检验

		红葡萄酒1评分	红葡萄酒2评分	白葡萄酒1评分	白葡萄酒2评分
个案数		27	27	28	28
正态参数[a,b]	平均值	73.085	70.515	73.97	76.532
	标准 偏差	7.3472	3.9780	4.825	3.1709
最极端差值	绝对	.156	.124	.091	.122
	正	.089	.078	.080	.076
	负	-.156	-.124	-.091	-.122
检验统计		.156	.124	.091	.122
渐近显著性（双尾）		.091[c]	.200[c,d]	.200[c,d]	.200[c,d]

图 6：单样本 K-S 检验结果图

从图 5 可以看出数据的散点分别近似为一条直线，且与对角线大致重叠；双边检验结果$p_1 = 0.091 > 0.05$，$p_2 = p_3 = p_4 = 0.200 > 0.05$。因此可以认为品酒员对葡萄酒的评分服从正态分布。

b.配对t检验模型的建立

由于品酒员对葡萄酒的评分服从正态分布，我们可以采用配对t检验模型来进行两对样本的显著性差异的检验。以第一组和第二组红葡萄酒为例（白葡萄酒与其处理方法相同）。

1) 提出假设。

原假设$H_0: \mu_1 = \mu_2$，即两组评酒员对葡萄酒的评分的平均值相等。

备择假设$H_1: \mu_1 \neq \mu_2$，即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。

2) 确定显著性水平

依据小概率原理，规定显著性水平$\alpha = 0.05$。

3) 计算检验统计量，确定概率值做出判断。

设$x_i, y_i(i = 1,2,\cdots,27)$分别表示第一、二组各个红葡萄酒样品的总分。

$$d_i = x_i - y_i$$

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

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

$$t = \frac{\bar{d} - \mu_1}{s_d/\sqrt{n}} \quad (2)$$

其中，d_i为偏差，$\bar{d}$为配对样本差值的平均值，s_d为配对样本差值的标准偏差，n为配对样本数，t为样本统计量，即在零假说：$\mu = \mu_1$为真的条件下服从自由度为$n-1$的t分布。

4) 作出推断结论：是否接受假设。

根据最后得出的的结果p，判断其是否小于假设的$\alpha = 0.05$。若小于，则拒绝原假设，即存在显著性差异；若大于，则接受原假设，不存在显著性差异。

c.配对t检验模型的求解

将附件 1 表格中的具体指代入到配对t检验模型中，使用 SPSS 统计软件进行统计计算，结果如下表：

表 1：配对样本检验结果

		平均值	标准偏差	配对差值	差值 95%置信区间		t	自由度	$Sig.$ 双尾
				标准误差平均值	下限	上限			
配对 1	红 1-红 2	2.5704	5.3628	1.0321	.4489	4.6918	2.490	26	.019
配对 2	白 1-白 2	-2.5607	5.0707	.9583	-4.5269	-.5945	-2.672	27	.013

分析结果可知，由于$p_1 = 0.019 < 0.05$，$p_2 = 0.013 < 0.05$，均拒绝原假设，所以两组评酒员的评价结果都有显著性差异。

5.1.2 第二部分：评价结果的可靠性

a.模型的建立

同样先以两组红葡萄酒的评价为例（两组白葡萄酒的处理方法相同），对两

组数据的可靠性进行分析，在葡萄酒的感官评价中，由于品酒员的评价尺度等方面的差异，会导致不同品酒员对同一酒样的评价差异很大，从而不能真实地反应酒样的差异。

因此，我们假设X_{ij}为第j号评酒员对第i号酒样的评价分数，每一酒样的评价误差程度为:

$$S_i = \sqrt{\frac{1}{10}\sum_{j=1}^{10}\left(X_{ij} - \bar{X_i}\right)^2} \quad (3)$$

$$\bar{X_i} = \frac{1}{10}\sum_{j=1}^{10} X_i$$

其中，$\bar{X_i}$为对于第i号酒样 10 个评酒员的评价总分的平均值$(i = 1,2,\cdots,27)$。此处的D_i为 10 名品酒员对第i号酒样评价的离散指标。

我们通常认为，当不同专家对同一样品酒的评分差距越小时，说明该评分是越能被大多数人接受的。而评价误差程度的平均值（标准差）为：

$$S = \frac{1}{27}\sum_{i=1}^{27} S_i \quad (4)$$

因此综上所述，我们建立我们的可靠性评价指标：对于同一样品酒的评价误差程度的平均值S越小越好。

b.模型的求解

使用$Excel$软件对附件 1 的具体值进行上述的数据求解处理，得到S的值如下表：

表 2：评价分数标准差求解结果

S	红葡萄酒	白葡萄酒
第一组	7.413148	10.55195
第二组	5.620081	7.069039

分析表中数据可知，第二组的样品酒的离散程度更小，即更稳定。因此第二组品酒员对样品酒的一致性更高，评价结果更可靠。

5. 2 问题二模型的建立和求解

5. 2. 1 模型建立

对于样品和指标（变量）进行分类主要采用聚类分析法，而求取样品以及类之间的距离有多种方法，本文主要采用欧几里得距离和最短距离法。

a. 数据标准化

因为所使用的数据的量纲不同，各组数据之间差异较大，为了减少因量纲不同而导致的分类结果误差交大问题，在进行聚类分析之前，先对所使用到的数据进行标准化处理。

标准方差：

$$s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{ij})^2}$$

标准化后：

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{s_j}(s_j\neq 0) \qquad (5)$$

b. 空间距离

若每种葡萄酒有 p 个指标，那么可以将每个样品看作为 p 维空间内的一个点，这个点对应的 p 个指标及为这个点的坐标，那么 n 个样品及可组成 p 维空间的 n 个点。此时就可以用各坐标间的距离来衡量样品之间的接近程度。

令表示第 i 个样品的第 j 个指标，表示第 i 个样品与第 j 个样品之间的距离，最常见的计算距离的方法为：

闵氏（Minkowski）距离，即：

$$d_q(x,y)=[\sum_{k=1}^{p}|x_{ik}-y_{jk}|]^{\frac{1}{q}} \qquad (6)$$

在$q=1,2$，或者$q\to+\infty$时，可分辨得到：

[1] 绝对值距离：

$$d_{ij}(1)=\sum_{k=1}^{p}|x_k-y_k|$$

[2] 欧几里得(Euclid)距离：

$$d_{ij}(2)=[\sum_{k=1}^{p}|x_k-y_k|^2]^{\frac{1}{2}}$$

[3] 切比雪夫(Chebyshew)距离：

$$d_{ij}(\infty)=\max|x_{ik}-x_{jk}|$$

c. 系统聚类

将 n 个样品各自看成一类，然后规定样品之间的距离和类与类之间的距离。刚开始，每个样品各自成类，类与类之间的距离与样品与样品之间的距离相等，选择距离最小的一对并成一个新类，计算新类与其他类的距离，再将距离最近的两类合并，这样每次减少一类，直至所有的样品都成一类为止，最终完成样品的分类。计算类与类之间的距离主要有：

（1） 最短距离法：

设G_p、G_q、G_r分别为一类样本，则最短距离的计算公式为：

$$D_k(p,q)=\min\left\{d_{il}\mid j\in G_p,l\in G_q\right\}$$

此时将G_p与G_q合并为G_r，则任意的类G_k和G_r的距离公式为：

$$D_{kr}^2=\min d_{ij}=\min\left\{\min d_{ij},\min d_{ij}\right\}=\min\left\{D_{kp},D_{kq}\right\} \qquad (7)$$

依次循环进行，最终完成对样品的分类。

（2） 最长距离法：

$$D_k(p,q)=\max\left\{d_{jl}\mid j\in G_p,l\in G_q\right\}$$

将类G_p与类G_q合并为类G_r，则任意的类G_k和G_r的距离公式为：

$$D_{kr}^2=\max d_{ij}=\min\left\{\max d_{ij},\max d_{ij}\right\}=\max\left\{D_{kp},D_{kq}\right\} \qquad (8)$$

（3） 重心法：

$$D_c(p,q)=d(x_i,x_j)$$

将类G_p与类G_q合并为类G_r，则任意的类G_k和G_r的距离公式为：

$$D_{kr}^2=\frac{n_p}{n_r}D_{kp}^2+\frac{n_q}{n_r}D_{kq}^2-\frac{n_qn_p}{n_r^2}D_{pq}^2 \qquad (9)$$

5.2.2 模型求解

1 .对红葡萄酒进行分类

由第一小题可知，第二组品酒员对样品酒的一致性更高，评价结果更可靠。故取第二组品酒员的评价标准，算出各项所占权重并加和，最终求出十位品酒员对每个葡萄酒样品的平均打分情况。作 27 种酒品的综合评价指标，品酒员的平均打分以及酿酒葡萄的理化性质指标形成一个 31 列 28 行的原始资料库，将其数据标准化。分别采用最短距离法、最长距离法、以及重心法对类与类之间的相似度进行判断。得出如下表格：

表 3：三种不同方法所推导出的分类表

方法	级数	酒品号
最大距离法	第一级	1、8、14
	第二级	2、3、9、23
	第三级	10、20、25、26
	第四级	4、5、6、7、11、12、13、15、16、17、18、19、21、22、24、27
最短距离法	第一级	1
	第二级	2、9、23
	第三级	4、5、6、7、8、10、11、12、13、14、15、16、17、18、19、20、21、22、24、25、26、27
	第四级	3
重心距离法	第一级	2、9、23
	第二级	1、8、14
	第三级	4、5、6、7、10、11、12、13、15、16、17、18、19、20、21、22、24、25、26、27
	第四级	3

由图可以看出，相较于最短距离法和及重心法，使用最大距离法能够更好的将所有酒品进行分类，使用最大距离法得出下方聚类分析树状图：

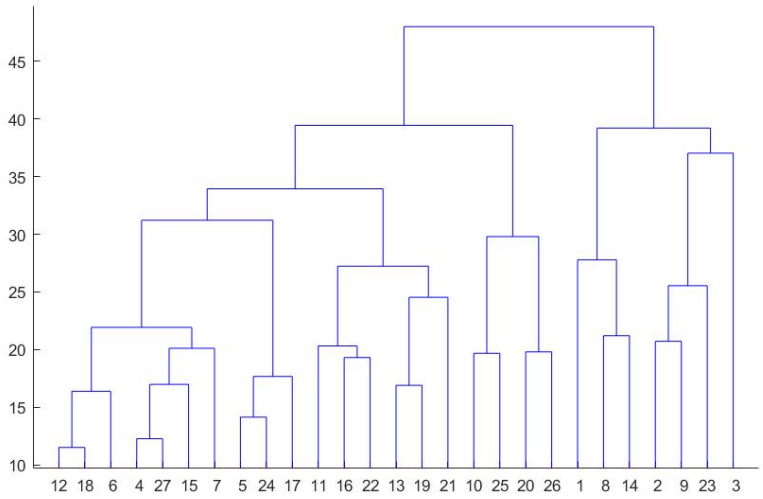


图 7：平均链接同组树状图

聚类分析只能将样品进行大致分类，并不能由此确定各级别的质量，而附件 2 里所给的理化指标以及品酒员的评分都可认为正相关于葡萄酒的质量。所以我们计算了每一级样品的每一指标的平均值，以及全部样品每个指标的品均值，此过程中共涉及 28 个指标。将各级样品指标的品均值减去全部样品指标的品均值，记录其中为正的数量，得出如下表格：

表 4：各等级高于总平均数的指标数

	第一级	第二级	第三级	第四级
高于平均数的个数	13	20	6	8

上表数据显示，第二级的酒品高于总平均数的指标数明显高于其他三级，第一级的酒品高于总平均数的指标数明显高于第三级和第四级，所以酒品质量排在第一、第二位的应该分别位第二级和第一级。第三级和第四级指标数相差并不明显，再次比较第三级和第四级的各指标大小，发现第四组的指标平均值普遍高于第三级，所以第三位为第四组，第四位为第三组。将上述一到四的排名划分为优，良，中等，差四个等级，可得到如下等级划分表：

表 5：红葡萄酒等级划分表

等级	酒品序号
优	2、3、9、23
良	1、8、14
中等	4、5、6、7、11、12、13、15、16、17、18、19、21、22、24、27
差	10、20、25、26

2 对白葡萄酒进行聚类分析

根据上一问对红葡萄酒进行分类的方法，利用聚类分析的欧式距离将 28 种

酒品分为四个组类如下表：

表 6：白葡萄酒级数划分表

级数	酒品序号
第一级	6、7、10、15、18、24
第二级	2、3、4、5、9、12、14、17、19、20、21、22、23、25、26、28
第三级	1、8、11、13、16
第四级	27

同样利用每一级的平均值与总平均值做差，对于无法确定的排名前后的级再进行单独比较， 其中酒品 27 在 29 项指标中有超过二十项高于平均水平，且其氨基酸、褐变度、总酚、单宁含量远高于平均水平，所以酒品第 27 号 d 单独为优，排名第二位的为第一级，第三名为第二级，最后为第三类，得出如下所示的酿酒葡萄(白)的等级划分：

表 7：白葡萄酒等级划分表

等级	酒品序号
优	27
良	6、7、10、15、18、24
中等	2、3、4、5、9、12、14、17、19、20、21、22、23、25、26、28
差	1、8、11、13、16

5.3 问题三模型的建立和求解

5.3.1 基于相关系数的酿酒葡萄理化指标的筛选

分析附件二的数据，我们发现有一些理化指标存在某种的关系。如红葡萄酒的理化指标中，带有红色字样的反式白藜芦醇苷，顺式白藜芦醇苷，反式白藜芦醇，顺式白藜芦醇的含量相加就是等于带有蓝色字样的白藜芦醇的含量。许多带有红色字样的理化指标经过一定的函数变换就是前面带有蓝色字样的理化指标。因此，我们只需要考虑带有蓝色字样的理化指标就足够了。

以红葡萄酒及红葡萄样本为例进行说明。（白葡萄酒及白葡萄样本的处理方法相同）

设红葡萄酒的各个理化指标花色苷，单宁…花色b分别为：y_1，$y_2 \cdots y_9$；红葡萄样本各个理化指标氨基酸，蛋白质…花色b为x_1，$x_2 \cdots x_{30}$。

为了建立单个葡萄酒理化指标与多个酿酒葡萄理化指标之间的函数关系，利用这多个酿酒葡萄理化指标与某个葡萄酒理化指标之间有比较密切的联系，反之则代表联系不够紧密，可以考虑将其排除在考虑范围。

根据附件 2 的具体数据，利用 Matlab 软件计算每个葡萄酒的理化指标与每个酿酒葡萄理化指标之间的相关系数，生成相关矩阵，如下表：

其中，*表示两理化指标间的相关度不大。

表 8：葡萄与葡萄酒相关系数矩阵表

	花色苷	单宁	总酚	酒总黄酮	白藜芦醇	DPPH	L	a	b
氨基酸	*	0.50	*	*	*	0.40	0.48	*	*
蛋白质	*	0.47	0.43	0.44	*	0.38	*	*	*
VC	*	*	*	*	*	*	*	*	*
花色苷	*	*	*	*	*	*	*	*	*
酒石酸	*	*	*	*	*	*	*	*	0.46
苹果酸	*	*	*	*	*	*	*	0.56	
柠檬酸	0.38	*	*	*	*	*	*	*	*
多酚氧化酶	0.48	*	*	*	*	*	0.41	*	*
褐变度	*	0.45	0.46	0.44	*	0.38	0.56	*	*
DPPH	0.57	*	*	*	0.42	*	*	*	*
总酚	*	*	*		0.46	*	*	*	*
单宁	*	*	*	*	*	*	*	*	*
总黄酮	0.44	*	*	*	0.57	*	*	*	*
白藜芦醇	*	*	*	*	*	*	*	0.45	*
黄酮醇	0.41	0.58	0.41	*	*	0.42	0.52	*	*
总糖	*	*	*	*	*	*	*	*	*
还原糖	*	*	*	*	*	*	*	*	0.57
可溶固形物	*	0.41	*	*	*	*	*	*	*
PH	*	*	*	*	*	*	*	*	*
可滴定酸	*	*	*	*	*	*	*	*	*
固酸比	*	*	*	*	*	*	*	0.44	0.39
干物质	*	0.42	*	*	*	*	*	*	*
果穗	*	*	*	*	*	*	*	*	*
百粒	*	*	*	*	*	*	*	*	*
果梗比	*	0.47	0.40	*	*	*	0.47	*	*
出汁率	*	*	0.39	0.48	*	0.42	0.44	*	*
果皮	*	*	*	*	*	*	*	*	*
L	*	0.46	0.44	*	*	0.39	0.49	*	*
a	*	*	*	*	*	*	0.59	0.54	0.06
b	*	*	*	*	*	*	*	*	*

5.3.2 多元线性回归方程模型的建立

针对酿酒葡萄的这些理化指标，使用多元线性回归方程建立其与葡萄酒各理化指标之间的关系。

多元线性回归分析的模型为：

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \quad (10)$$

其中，$i = 1,2,\cdots,9$，式中$\beta_0,\beta_1,\cdots,\beta_m$都是与$x_1,x_2,\cdots,x_m$无关的未知参数，$\beta_0,\beta_1,\cdots,\beta_m$叫做回归系数，$\varepsilon$是随机误差项。

现有n个独立观测数据（y_i，$x_{i1}\cdots x_{im}$），得：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i$$

设：

$$\varepsilon = [\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_{30}]^T,\ \beta = [\beta_0\ \beta_1\ \cdots\ \beta_m]^T$$

$$X = \begin{bmatrix} 1 \cdots x_{1m} \\ \vdots \quad \vdots \\ 1 \cdots x_{1m} \end{bmatrix}\ Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}$$

模型可表示为：

$$Y = X\beta + \varepsilon \quad (11)$$

5.3.3 模型的求解

将已知数据$i = 1,2,\cdots,9$等代入模型中，使用 SPSS 软件进行求解。y_1求解结果如下：

表 9：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

为了去除量纲的影响，我们可以取标准回归系数。由于常数的均值是其本身，经过标准化就为 0。有：

$$y_1 = 0{,}308x_7 + 0.363x_8 - 0.072x_{15} + 0.173x_{13} + 0.195x_{25} + 0.321x_{10}$$

同理我们可以得到$y_2,\cdots,y_9$。

综上所述，基于红葡萄酒和红葡萄理化指标之间的多元线性回归方程为：

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

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

$$y_3 = 0.194x_2 - 0.271x_{28} + 0.129x_{26} + 0.240x_9 + 0.105x_{25} + 0.110x_{15}$$

$$y_4 = 0.248x_2 + 0.327x_9 + 0.313x_{26}$$
$$y_5 = 0.821x_{13} - 0.112x_{10} - 0.180x_{11}$$
$$y_6 = 0.510x_1 - 0.072x_{28} + 0.340x_2 - 0.115x_{25} - 0.084x_{15} + 0.470x_9$$
$$y_7 = 0.153x_{25} - 0.166x_2 - 0.086x_8 - 0.378x_{15} + 0.475x_{29} + 0.202x_{27} - 0.162x_{26} - 0.245x_9$$
$$y_8 = -0.428x_2 - 0.024x_{14} - 0.455x_{29} - 0.272x_{21}$$
$$y_9 = 0.791x_{17} + 0.466x_5 - 0.179x_{29} - 0.337x_{22}$$

基于白葡萄酒和白葡萄理化指标之间的多元线性回归方程为：

$$y_1 = 0.572x_{12} + 0.193x_{27} - 0.717x_{15} + 0.42x_{16} + 0.169x_{21} + 1.118x_{11} + 1.111x_{13} - 0.415x_{22} - 0.274x_{25} + 0.32x_{26} + 0.135x_{10}$$
$$y_2 = 0.540x_{13} + 0.305x_{18} + 0.201x_1 + 0.171x_{27} - 0.156x_4 - 0.119x_{26}$$
$$y_3 = 0.249x_{11} + 0.168x_2 + 0.262x_6 + 0.332x_{15} - 0.247x_{25}$$
$$y_4 = 0.348x_{16} - 0.254x_6 - 0.161x_5$$
$$y_5 = 1.291x_{13} + 0.435x_{16} + 0.507x_3 - 0.25x_8 - 0.96x_{11} + 0.309x_{15}$$
$$y_6 = -0.69x_{22} + 0.254x_{26} - 0.345x_{14} + 0.402x_{17} + 1.128x_{20} + 1.078x_{21} - 0.162x_9$$
$$y_7 = 0.486x_{24} - 0.387x_{19} - 0.697x_{25} + 0.513x_{26} - 0.219x_7 + 0.126x_2 + 0.529x_5 - 0.51x_{27} - 0.255x_{22} + 0.324x_{18} + 0.161x_{13}$$
$$y_8 = 0.756x_{26} - 0.796x_{24} + 0.251x_{22} - 0.115x_{20} - 0.313x_{27} + 0.288x_{16} + 0.449x_{23} + 0.866x_{21} + 0.123x_5 - 0.075x_{19}$$

5.3.4 结果分析

由以上结果可知红葡萄酒理化指标中的白藜芦醇与红酒葡萄理化指标的相关程度较弱，色泽 b 理化指标与酿酒葡萄理化指标的相关程度一股；白葡萄酒理化指标中的白藜芦醇与酿酒葡萄理化指标相关程度弱，仅与苹果酸和总糖有些许相关关系。而葡萄酒的其他理化指标大多能够与酿酒葡萄的某几种理化指标之间建立起函数关系，且拟合程度不错。

5.4 问题四模型的建立和求解

5.4.1 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数

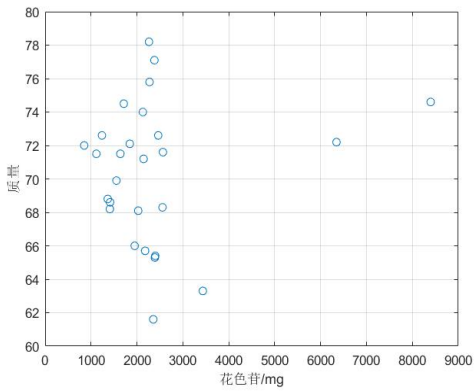


图 8：理化指标花色苷的散点图

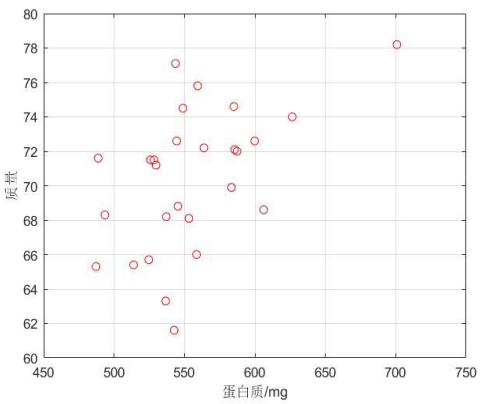


图 9：理化指标蛋白质的散点图

据的分布规律，以花色苷和蛋白质为例如下：

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。

5.4.2 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，否则反之。

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：

$$\mu = a_0 + a_1 y_1 + a_2 y_2 + a_3 y_3 + \cdots + a_8 y_8 + a_9 y_9 \qquad (12)$$

白葡萄酒：

$$v = b_0 + b_1 y_1 + b_2 y_2 + \cdots + b_7 y_7 + b_8 y_8 \qquad (13)$$

5.4.3 模型的求解

使用 Eviews 软件，代入附件 2 和附件 3 的数据进行多元线性回归，结果如下：

红葡萄酒：

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

表 10：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。

白葡萄酒：

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

表 11：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。

5.4.4 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。

六、 模型的评价及推广

6.1 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。

6.2 模型的缺点

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。

6.3 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，接着通过多元线性回归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。可以推广到其他领域，如生物科学、数理科学等，分析两个变量间的关系。

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

附录 1

基于问题二聚类分析的 matlab 程序

```
    load a.txt;
a1=zscore(a);
y=pdist(a1,'cityblock');
yc=squareform(y)
z=linkage(y)
dendrogram(z)
T=cluster(z,'maxclust',3)
for i=1:3
    tm=find(T==i);
    tm=reshape(tm,1,lengh(tm));
    fprintf('µÚ%dÀàµÄÓÐ%s\n',i,int2str(tm));
end

load  b.txt;
b1=zscore(b);
y=pdist(b1,'cityblock');
yc=squareform(y)
z=linkage(y,'complete')
dendrogram(z)
T=cluster(z,'maxclust',4)
for i=1:4
    tm=find(T==i);
    tm=reshape(tm,1,length(tm));
    fprintf('µÚ%dÀàµÄÓÐ%s\n',i,int2str(tm));
    end
```

附录 2

基于问题四绘制散点图的 matlab 程序

```
%导入 Test.xlsx 中 y1、Y1、x1、x2 同变量命名
plot(x1,y1,'o')
grid on
hold on
xlabel('氨基酸总量/mg')
ylabel('花色苷/mg')
grid off
hold off
%%
close all
plot(x2,y1,'ro')
grid on
hold on
xlabel('蛋白质/mg')
ylabel('花色苷/mg')
%%
close all
plot(x1,Y1,'o')
grid on
hold on
xlabel('花色苷/mg')
ylabel('质量')
%%
close all
plot(x2,Y1,'ro').
grid on
hold on
xlabel('蛋白质/mg)
ylabel('质量')
```