2023 年第八届“数维杯”大学生
数学建模挑战赛论文

题　　目　受试者感受与节育器质量评价模型

摘　要

随着医疗手段的进步和发展，宫内节育器作为一种相对安全、有效、经济、可逆、简便的节育器具越来越被广大妇女所选用。本文基于两个医院的节育器的临床数据等，通过运用**熵权法、t 检验、斯皮尔曼相关系数分析法、岭回归、主成分分析法、灰色关联分析法和二分类逻辑回归**建立模型，对问题一到四进行解决。

针对问题一，我们首先建立了 **t 检验**模型，分别对两个医院的节育器理化指标、被试的身体全面情况、两个医院随访时被试的不适程度分别进行**独立样本 t 检验**，初步得出两个医院的临床数据存在显著性差异的结论。其次，通过对节育器理化指标、受试者的身体全面情况、不适程度三个指标进行**描述性统计，**得出导致**两个医院临床数据具有显著性差异的原因**。

针对问题二，我们运用斯皮尔曼相关性分析对身体指标和随访主诉情况指标进行分析，并对样本间的指标进行假设检验，得到身体指标和**随访主诉情况中的不适状况程度具有很大联系**。同时，对不适状况程度和身体指标等指标进行岭回归分析，**得出身体指标是受试者出现不适状况的主要因素**。

针对问题三，我们首先通过**主成分分析法**计算组成身体状况因素的六个指标的贡献率作为对应权重，再用 CRITIC **权重法**确定包含身体状况等七个指标的权重，接着使用**灰色关联分析法**计算加权灰色关联度作为两类节育器的质量综合得分。最终得到**第二组的被试的灰色加权关联度均值小于第三组的被试的灰色加权关联度均值。**根据均值，**我们认为 VCu380 比 VCu260 记忆型宫内节育器的质量更优**，更适合生产。

针对问题四，我们采用**二分类逻辑回归算法**，基于问题三，对影响节育器质量的决定性因素进行分析，将使用 VCu260 还是 VCu380 这个分类变量作为二分类逻辑回归的被解释变量，将其余七个指标作为解释变量进行回归。最后根据**混淆矩阵热力图和 ROC 曲线**，证明模型效果较好并保证了**影响宫内节育器质量的决定因素是是否怀孕**这一结果的可靠性。

关键词　节育器；t 检验；斯皮尔曼相关系数；灰色关联分析；二分类逻辑回归

目 录

一. 问题重述

1.1 背景介绍

据统计，约 70%选用 IUD 作为避孕方法，在世界 IUD 避孕总人数中占 80%。某公司研发了 VCu260 和 VCu380 两种型号的记忆型宫内节育器。此类节育器不仅有独特形状记忆功能，还具有抗腐蚀、耐磨损、超弹性和对身体副作用较小等特点。然而，节育器在给女性带来方便的同时，也可能会引起疼痛、不适、脱落或者出血等症状[1]。

基于此，将这两种型号的节育器与已经被临床应用的 MCu 功能性宫内节育器一起做临床试验，用于探究这两种型号的节育器是否适合投入生产。实验共分为三组，MCu 功能性宫内节育器作为对照组（标记为 1 组），VCu260 和 VCu380 记忆型宫内节育器均为试验组（分别标记为 2、3 组），每组入组均为 525 例，并且医院会在第一、三、六、十二个月就主诉情况进行随访记录。本次实验选择在两家医院同时进行临床试验，其中，临床指标中的节育器的理化指标、受试者的身体指标、随访时的主诉情况均能在一定程度上反应节育器的质量。

1.2 需要解决的问题

问题 1：根据附件中给出的数据，分析两所医院的临床数据是否有显著性差异，并对导致这种差异的因素进行分析。

问题 2：根据附件中给出的关于身体指标和受试者出现不适状况的数据，分析身体指标与随访主诉情况的联系，并说明受试者的身体指标是否为导致受试者身体出现不适状况的因素。

问题 3：根据题目中给出的有关受试者的身体指标、节育器的理化指标与随访时的主诉情况，建立节育器质量模型，分析 VCu260 和 VCu280 这两种节育器哪种质量更优、更适合生产。

问题 4：根据问题 3 中建立的节育器质量模型，探究影响宫内节育器质量的决定因素。

二、问题分析

IUD 作为一种节育工具，因其安全、有效、经济、可逆、简便的特性，成为了广大妇女较容易接受的节育器具，目前已经成为了我国育龄妇女的主要避孕措施[2]。随着医疗水平的发展和进步，节育器的形式和功能也逐渐多样化。然而节育器对女性带来的便利的同时，也会带来不适症状[3]。VCu260 和 VCu380 作为两种新型的节育器，是否适合投入生产，还需要进行临床检验，由此探究节育器的质量以及女性对其的适应程度[4]。

2.1 问题 1 的分析

问题一要求结合数据分析两个医院的临床数据的差异。我们选取节育器的理化指标、受试者的身体指标以及随访时的主诉情况进行分析。由于数据中存在被使者数据缺失的现象，导致结果不准确。因此，将数据中被使者失访以及未进行随访的数据剔除。然后，采用 t 检验的方法，分析两个医院的数据差异，并对导致差异的因素进行描述性统计分析[5]。

2.2 问题 2 的分析

问题二要求结合数据分析受试者的身体指标与随访主诉情况的联系，根据已有资料，我们将受试者的年龄、月经周期/经期、既往使用节育器的情况、宫腔深度纳入身体状况的衡量指标，而受试者的不适状况由非经期出血、疼痛、经量多、分泌物增多、经期/周期异常五个指标来体现。我们采用斯皮尔曼相关系数分析法，由此说明受试者的身体指标是否是受试者出现不适状况的主要因素[6]。

2.3 问题 3 的分析

问题三要求建立节育器质量模型比较 VCu260 和 VCu380 的这两类记忆型宫内节育器的质量。根据已有的数据，首先我们运用主成分分析法计算衡量身体健康的各个指标的贡献率，接着，我们运用灰色关联分析法计算出 VCu260 记忆型宫内节育器的质量综合得分低于 VCu380 记忆型宫内节育器的质量综合得分，由此认为，VCu380 比 VCu260 更适合投入生产。

2.4 问题 4 的分析

问题四要求在问题三所建立的模型的基础上分析节育器质量的决定性因素。我们采用二分类逻辑回归的方法，将 VCu260 和 VCu380 这两种型号的节育器作为二分因变量，将身体状况、使用节育器型号、不适程度、最终是否不适、是否怀孕、是否因症取出、是否脱落这七项指标作为自变量，输入到 SPSSPRO 中得出结论：是否怀孕是影响节育器质量的决定性因素，而对节育器的适应程度和节育器是否脱落也是重要的考量因素。

三、模型假设

假设 1：假设题目中所给的数据真实可靠；
假设 2：假设对异常值处理时所产生的误差可以忽略不计；
假设 3：假设年龄对节育器的适应性程度呈反向变动

四、定义与符号说明

符号定义	符号说明
x_{ij}	第 i 个被试的第 j 项指标
r	斯皮尔曼相关系数
d_i	指标 X_{1i} 与 X_{2i} 的等级差
λ	正则化参数，决定回归系数
b_j	第 j 个指标的信息贡献率
ρ	分辨系数，题中 ρ 取 0.5

五、模型的建立与求解

5.1 数据的预处理：

题目所给出的附件数据非常庞大和复杂，为了更好地利用所给数据，使建立的模型能够更好分析新型节育器的质量以及女性对其的适应性问题，在对问题进行正式求解之前，对数据进行一定的分析，并利用相应的数学方法对数据进行预处理。具体步骤如下：

（1）附件二:剔除随访时失访的数据，得到医院一剩余被试 408 个和医院二剩余被试 743 个。

序号	组别	随访时的主诉情况																																	
		失访				脱落				因症取出				怀孕				非月经期出血				疼痛				经量多				分泌物增多				经期/周期异常				有不适人数			
		1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12	1	3	6	12
10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	1	1	1	0	1	1	1	0
36	1	0	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	0	0	0	1	0	0	0	0	0	0	0	1	0	0
69	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	1	1	0	0
75	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
89	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
98	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
103	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1

图 1 处理后的附件二

（2）附件一：仅保留处理后的附件二中存在的被试，得到医院一剩余 408 个和医院二 743 个，与附件二相一致。

序号	组别	年龄（岁）	初潮年龄（岁）	月经周期（天）	月经经期（天）	既往应用节育器情况			宫腔深度（cm）	使用节育器型号情况			放置节育器时宫颈扩张情况
						IUD	无	其它		小	中	大	
10	1	33	17	28	4	1	0	0	8	0	1	0	1
17	1	29	13	28	5	1	0	0	8	0	1	0	1
36	1	38	16	31	4	1	0	0	8	0	1	0	0
69	1	34	13	29	4	1	0	0	7	0	1	0	1
75	1	39	15	29	4	1	0	0	8	0	1	0	1
89	1	30	13	32	5	0	0	1	8	0	1	0	1
98	1	23	14	30	3	0	1	0	7	0	1	0	0
103	1	34	14	30	4	1	0	0	8	0	1	0	1

图 2 处理后的附件一

5.2 问题 1 的模型建立与求解

5.2.1 思路分析

对于问题一，我们决定从节育器理化指标，被试身体全面状况，不适程度三个角度分别用 t 检验分析两医院的这些指标有无显著性差异。

对于节育器理化指标。直接作为多选题的形式录入 SPSS，将医院一和医院二作为分组变量进行独立样本 t 检验。

对于身体全面状况。首先，我们选择了是否使用过节育器、放置节育器时宫颈扩张情况、宫腔深度、月经经期、年龄、初潮年龄、月经周期这七个指标，其次对各个指标标准化，再次，用熵权法计算各个指标权重，用权重与对应指标相乘再将各指标值相加，得到每个被试的身体全面情况量化值，最后，对被试身体全面情况进行独立样本 t 检验。

对于不适程度。首先，我们选取了非月经期出血、疼痛、经量多、分泌物增多、经期/周期异常这五个指标，其次，对这五个指标用熵权法计算权重，之后用权重与对应项相乘再将各个指标相加得到被试不适程度评价值，最后，对两医院被试不适程度进行独立样本 t 检验。

最后进行描述统计，针对结果分析两医院临床数据的显著性差异。

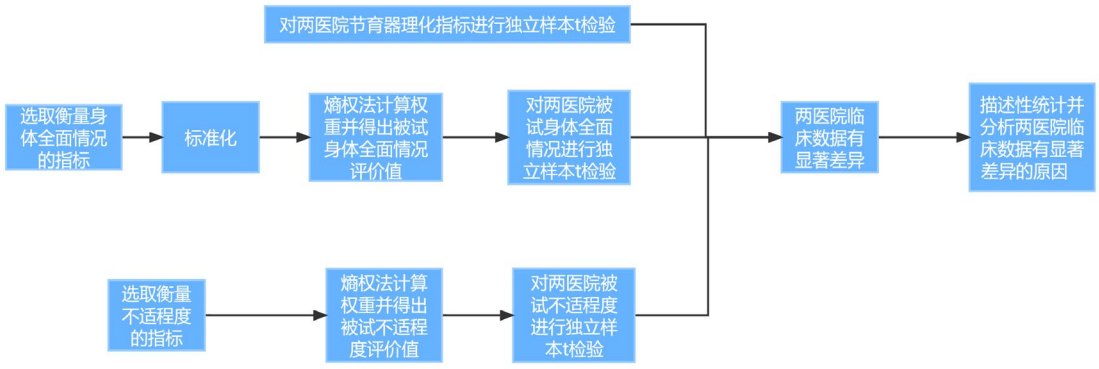

图 3 问题一求解流程图

5.2.2 t 检验模型的建立

t 检验是用 t 分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著的一种假设检验方法。因为此题分组变量为医院一和医院二，我们选择运用独立样本 t 检验，t 检验的统计量为：

$$t=\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{(n_1-1)\ S_1^2+(n_2-1)\ S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (1)$$

其中，S_1^2 和 S_2^2 为医院一和医院二数据的方差，n_1 和 n_2 分别为医院一和医院二样本容量。

我们需要解决的问题是分析两个医院间的临床数据是否有显著差异，并说明导致这种差异的因素。为了确保数据的有效性以及分析结果的准确性，我们采用预处理之后的数据，建立 t 检验的模型，分别对两个医院的节育器理化指标、被试的身体全面情况、两个医院随访时被试的不适程度分别进行独立样本 t 检验。

5.2.3 t 检验模型的求解

（1）对两个节育器的理化指标进行独立样本 t 检验，我们设：

H_0=两个医院的节育器理化指标没有显著差异

H_1=两个医院的节育器理化指标有显著差异

利用 SPSS 进行分析，将节育器型号以多选题形式录入，作为检验变量；医院一和医院二作为分组变量，进行独立样本 t 检验，得到的结果如下所示：

		莱文方差等同性检验		平等值等同性t检验					差值95%置信区间	
		F	显著性	t	自由度	Sig	平均值差值	标准误差差值	下限	上限
节育器型号（小）	假定等方差	1638.891	0.000	-16.51	1149	0.000	-0.442	0.027	-0.494	-0.389
	不假定等方差			-19.294	1148.915	0.000	-0.442	0.023	-0.486	-0.397
节育器型号（中）	假定等方差	115.111	0.000	35.056	1149	0.000	0.713	0.020	0.673	0.753
	不假定等方差			31.869	638.539	0.000	0.713	0.022	0.669	0.757
节育器型号（大）	假定等方差	4.651	0.031	-36.124	1149	0.000	-0.748	0.021	-0.788	-0.707
	不假定等方差			-36.915	892.318	0.000	-0.748	0.020	-0.787	-0.708

图 4 节育器的理化指标独立样本 t 检验结果

结果显示，无论节育器型号是什么，P 值均为 0.000，小于显著性水平，所以拒绝原假设，可以认为两个医院的节育器理化指标存在显著差异。

（2）对两个医院的被试的身体全面情况进行独立样本 t 检验，我们设：

H_0=两个医院的被试的身体全面情况没有显著差异

H_1=两个医院的被试的身体全面情况有显著差异

在本次检测中，我们选取附件一中的是否使用过节育器、放置节育器时宫颈扩张情况、宫腔深度、月经经期、年龄、初潮年龄、月经周期七个指标作为身体情况的衡量标准。对其中部分处理过的指标的定义详细说明：关于是否使用过节育器，我们设 1 为用过，0 代表没用过；放置节育器时宫颈扩张情况，1 为扩张，0 为没扩张。

由于指标的单位各不相同，我们决定对其进行正态分布标准化。

$$X_i^{'} = \frac{x_i - \mu}{\sigma} \qquad (2)$$

在对指标进行定义后，我们运用熵权法计算各指标的权重：

算得各指标的权重如下：

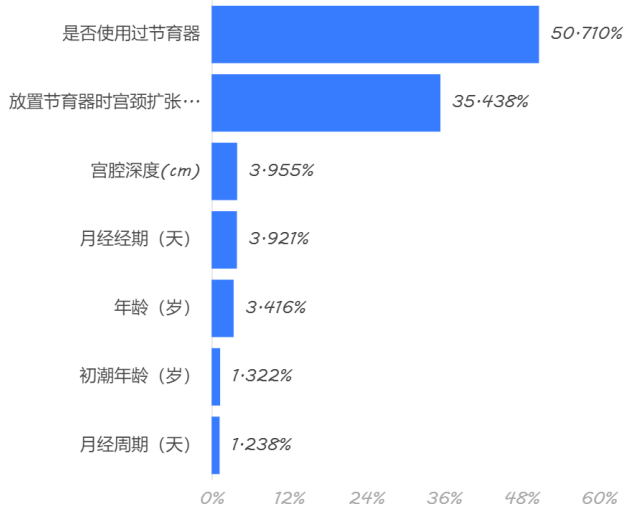

图 5 身体全面情况各个指标权重图

将其作为权重和对应项相乘再相加，得到每个医院各个被试身体状况的量化结果。将身体状况和所在医院数据输入到 SPSS 中，分组变量为医院一和医院二，检验变量为身体状况。得到 t 检验的结果：

		莱文方差等同性检验		平等值等同性t检验					差值95%置信区间	
		F	显著性	t	自由度	Sig	平均值差值	标准误差差值	下限	上限
被试身体全面情况	假定等方差	221.956	0.000	4.964	1149	0.000	0.19147821	0.038570906	0.115800906	0.267155514
	不假定等方差			4.278	556.456	0.000	0.19147821	0.044761012	0.103557005	0.279399414

图 6 身体全面情况独立样本 t 检验结果

如上图显示，P 值为 0.000，小于显著性水平，所以拒绝原假设，可以认为两个医院的被试身体全面情况存在显著差异。

（3）在对两个医院的被试随访的不适状况程度进行独立样本 t 检验时，我们设：

H_0=两个医院的被试的不适状况程度没有显著差异

H_1=两个医院的被试的不适状况程度有显著差差异

对于被试的不适状况程度，我们将非月经期出血、疼痛、经量多、分泌物增多、经期/周期异常这五个变量纳入衡量范围。将各个指标各自第 1、3、6、12 月的值相加，用熵权法求得各个指标的权重如下：

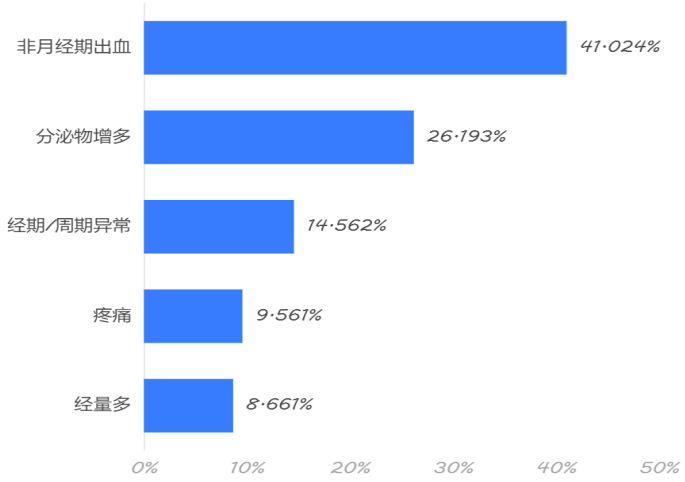

图 7 不适状况程度各个指标权重图

将对应项相乘再相加得到两个医院各个被试的不适状况程度。将数据结果输入到 SPSS 中，分组变量为医院 1 和医院 2，检验变量为不适状况程度。得到 t 检验的结果：

		莱文方差等同性检验		平等值等同性t检验					差值95%置信区间	
		F	显著性	t	自由度	Sig	平均值差值	标准误差差值	下限	上限
不适程度	假定等方差	27.088	0.000	3.630	1149	0.000	0.01676332	0.02795254	0.02795254	0.09373284
	不假定等方差			3.388	686.828	0.001	0.0179579	0.02558371	0.02558371	0.09610167

图 8 不适状况程度独立样本 t 检验结果

结果显示，假定等方差时 P 值为 0.000，不假定等方差时 P 值为 0.001，均小于显著性水平，所以拒绝原假设，可以认为两个医院的被试身体全面情况存在显著差异。

综上，据上述分析可以得知，两个医院的临床数据之间存在显著性差异。

5.2.4 描述性统计的求解

为了进一步探究导致这种差异的因素，我们对两医院的节育器理化指标，被试身体全面情况，不适程度这三个指标依次进行描述性统计。

（1）节育器的理化指标描述性统计：

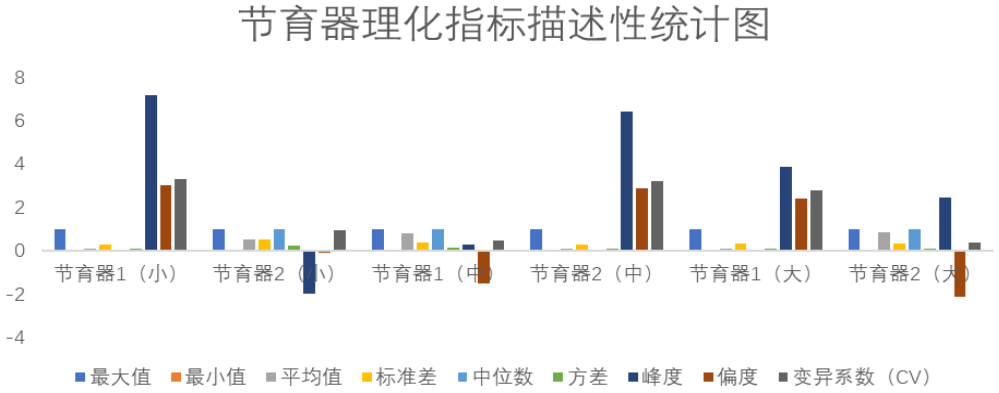

图 9 节育器理化指标描述性统计图

节育器（小）1 和节育器（小）2 的偏度都比较大，分别为 3.026 和-0.1，表明数据分布偏左和偏右；而节育器（中）1 和节育器（中）2 的偏度分别为-1.517 和 2.896，表明它们的分布形态相对比较对称。节育器（小）1、节育器（中）1 和节育器（大）1 的峰度都比较大，分别为 7.193、0.303 和 3.873，表明它们的分布比正态分布更尖峭；而节育器（小）2、节育器（中）2 和节育器（大）2 的峰度较小，表明它们的分布比正态分布更平坦。

他们的使用节育器（中）和（大）的均值和极值相差较大，但是标准差相差不大，说明数据相对为集中。

总的来说，医院 1 和医院 2 的节育器在不同型号均值和极值，以及峰度和偏度有较大差异，说明数据分布存在较大区别，这就造成医院 1 和 2 在节育器理化特性存在显著性差异的因素。

（2）被试身体全面情况描述性统计：

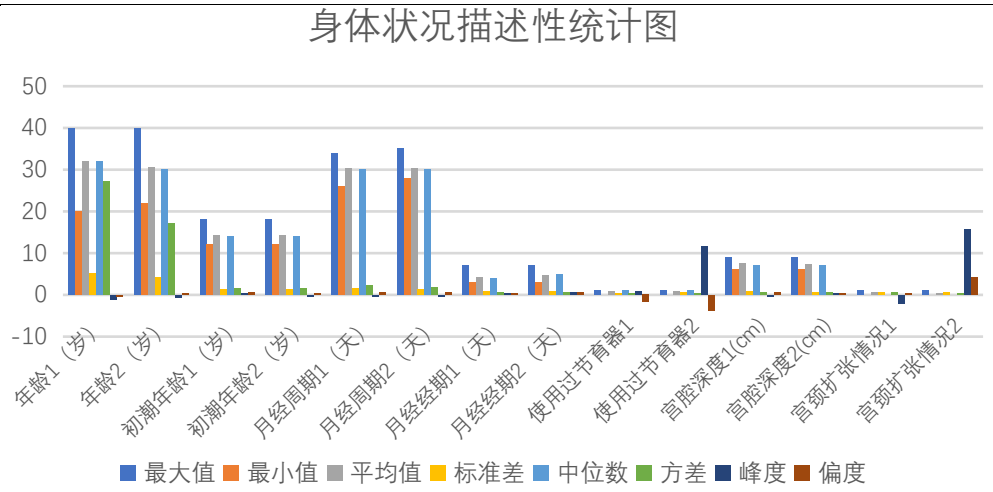

图 10 身体状况描述性统计图

年龄 1 平均值略高于年龄 2，但差距不大，标准差也比年龄 2 大一些；初潮年龄 1 和 2 两组数据相差不大，但是初潮年龄 1 的峰度和偏度较高；月经周期 1 和 2 两组数据相差不大，标准差也比较小；月经经期 1 和 2 两组数据相差不大，但月经经期 1 的峰度和偏度较高；使用过节育器 1 的平均值略低于使用过节育器 2，但标准差比使用过节育器 2 大一些；宫腔深度 1 和 2 两组数据相差不大，但宫腔深度 1 的峰度和偏度较高；宫颈扩张情况 1 和 2 两组数据相差较大，宫颈扩张情况 2 的峰度和偏度较高。

综上分析，医院 1 的初潮年龄和月经经期，宫腔深度的峰度和偏度和医院 2 相比较高，以及医院 2 的防止节育器时宫扩情况的峰度和偏度较大，这就说明了医院 1 和 2 在初潮年龄和月经经期，宫腔深度，防止节育器时宫扩情况的数据存在较大差异，所以我们认为这是造成医院 1 和医院 2 临床数据在身体状况指标上存在显著性差异的原因。

（3）不适程度描述性统计

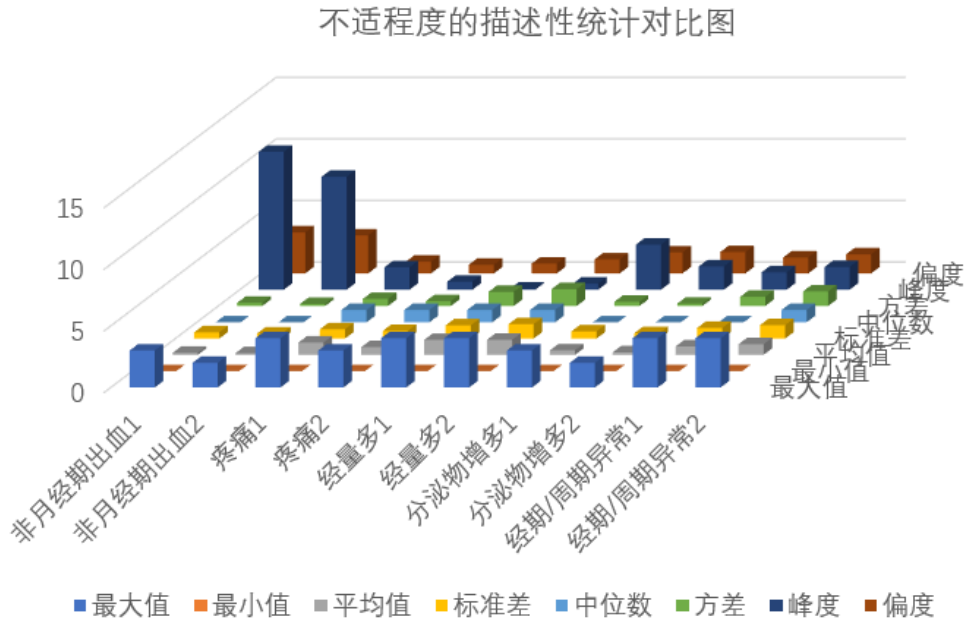

图 11 不适程度的描述性统计对比图

根据结果，我们认为：

非月经期出血：平均每个周期内出血次数很少，其中非月经期出血 1 发生的次数更少，最大值为 3，说明个别情况下可能会有较多的出血。偏度和峰度较高，说明数据分布较为不均匀。

疼痛：平均每个周期内有一次疼痛发生，其中疼痛 2 发生的次数更少，最大值为 3，说明个别情况下可能会有较多的疼痛。偏度和峰度较低，说明数据比较均匀。

经量多：平均每个周期内有一次经量较多，其中经量多 2 发生的次数稍多一些，最大值为 4。偏度和峰度较低，说明数据比较均匀。分泌物增多：平均每个周期内有 0.29 次分泌物增多，其中分泌物增多 1 发生的次数较多。偏度和峰度较高，说明数据分布不均匀。经期/周期异常：平均每个周期内有 0.69 次出现经期/周期异常，偏度和峰度较高，说明数据分布不均匀。

综合分析，对于非月经期出血、疼痛、分泌物增多和经期/周期异常这四个指标，两个医院的数据差异比较明显，而经量多这个指标，在两次观测中显示出的数据差异较小。所以我们认为，导致医院 1 和 2 在不适程度的指标上存在显著性差异的因素是两个医院在非月经期出血、疼痛、分泌物增多和经期/周期异常这四个指标的数据偏差较大。

5.2.5 结果

通过对模型的建立和求解，利用 SPSS 对筛选后的数据进行初步运算。首先通过 SPSS 对两个节育器理化指数、被试的身体全面情况、两个医院的随访的不适程度分别**进行独立样本 t 检验，得出两个医院的临床数据具有显著性差异。**

通过对指标的描述性统计，可以看出**医院一和医院二的不同型号节育器的均值和极值，以及峰度和偏度有较大差异；**医院一和医院二的被试在初潮年龄，月经经期，宫腔深度，放置节育器时宫腔扩张情况的数据存在较大差异；且两医院被试的非月经期出血、疼痛、分泌物增多和经期/周期异常这四个指标的数据偏差较大。所以最终导致两医院临床数据有显著差异。

5.3 问题 2 的模型建立与求解

5.3.1 数据预处理

将第一次处理过后的附件一和附件二中，医院一和医院二的数据按照医院一在上，医院 2 在下，整理得到附件一一和附件二一中。

5.3.2 思路分析

我们需要解决的问题是分析并说明受试者的身体指标是否是受试者出现不适状况的主要因素。

查阅相关资料后，我们将附件一一中的年龄、月经经期、既往节育器使用情况作为衡量身体健康状况的因素。

而对不适状况程度（同不适程度）的衡量则用熵权法进行加权，其中继续沿用附件二一中的非经期出血、疼痛、经量多、分泌物增多、经期/周期异常这五项作为衡量不适状况的依据。

其余指标为附件二一中脱落，因症取出，怀孕，放置节育器时宫颈扩张情况。

即我们共有六组数据：身体指标，不适状况程度，脱落，因症取出，怀孕，放置节育器时宫颈扩张情况。对这六组数据两两之间计算斯皮尔曼相关系数。并对数据使用岭回归，得到回归系数进一步验证各种指标的重要程度。

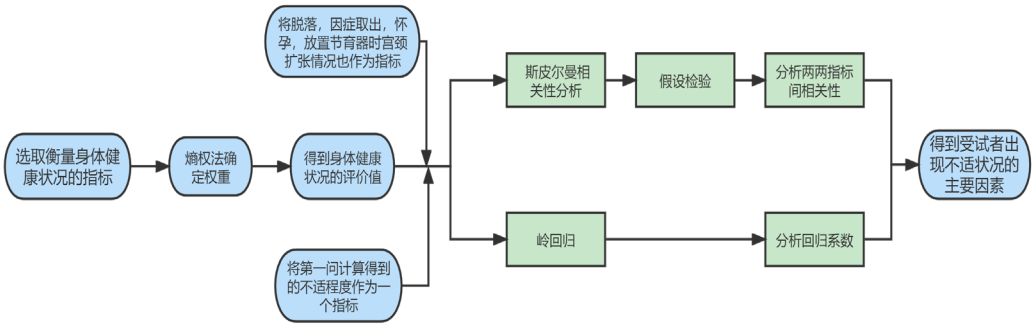

图 12 问题二求解流程图

5.3.3 斯皮尔曼相关系数模型的建立

斯皮尔曼等级相关系数用来估计两个变量 X、Y 之间的相关性，其中变量间的相关性可以使用单调函数来描述。如果两个变量取值的两个集合中均不存在相同的两个元素，那么，当其中一个变量可以表示为另一个变量的很好的单调函数时（即两个变量的变化趋势相同），两个变量之间的 ρ 可以达到+1 或-1。

假设两个随机变量集合分别为 X、Y，它们的元素个数均为 N，两个随即变量取的第 i（1<=i<=N）个值分别用 X_i 、Y_i 表示。对 X、Y 进行排序（同时为升序或降序），得到两个元素排行集合 x、y，其中元素 x_i 、y_i 分别为 Xi 在 X 中的排行以及 Yi 在 Y 中的排行。将集合 x、y 中的元素对应相减得到一个排行差分集合 d，其中 $d_i = x_i - y_i$，1<=i<=N。随机变量 X、Y 之间的斯皮尔曼等级相关系数可以由 x、y 或者 d 计算得到，其计算方式如下所示：

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N\left(N^2-1\right)} \text{ 或 } \rho = \frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2 \sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}} \tag{3}$$

在本题中，集合 X 和 Y 如下：

$$X = Y = \{\text{身体指标，不适状况程度，脱落，因症取出，怀孕，放置节育器时宫颈扩张情况}\}$$

5.3.4 岭回归模型的建立

岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的

最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

线性回归模型的目标函数为：

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij}\right)^2 \qquad (4)$$

RSS 随着特征数量 p 的增加逐渐减小，为了保证回归系数 β 可求，岭回归模型在目标函数上加了一个 L2 范数的惩罚项：

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_i x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \qquad (5)$$

其中，$\lambda >= 0$，用以控制 RSS 和 L2 范数的相对权重，其需要通过岭迹法或交叉验证法进行选择。

此外，由于岭回归是为了解决多重共线性的问题，必然存在多个变量，而多个变量一般都存在量纲不同的问题，所以在使用岭回归之前，需要对 X 用如下公式进行标准化：

$$\widetilde{x_{ij}} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \overline{x_j}\right)^2}} \qquad (6)$$

相比于最小二乘法，岭回归以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法。

5.3.4 问题二模型的求解

(1)第一问中熵权法求出不适状况程度的各部分权重。

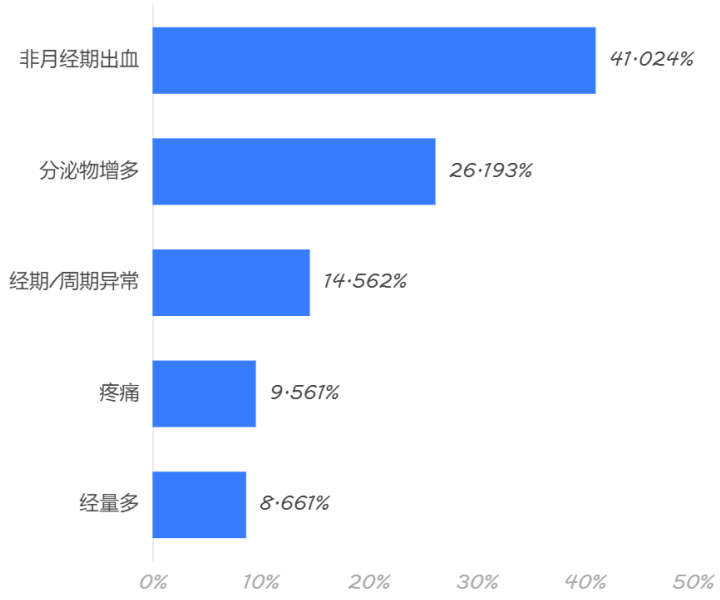

图 13 不适状况程度的各部分权重

（2）通过 matlab 代码(spearman.m)来求解斯皮尔曼相关系数，并绘制相关系数热力图如下所示：

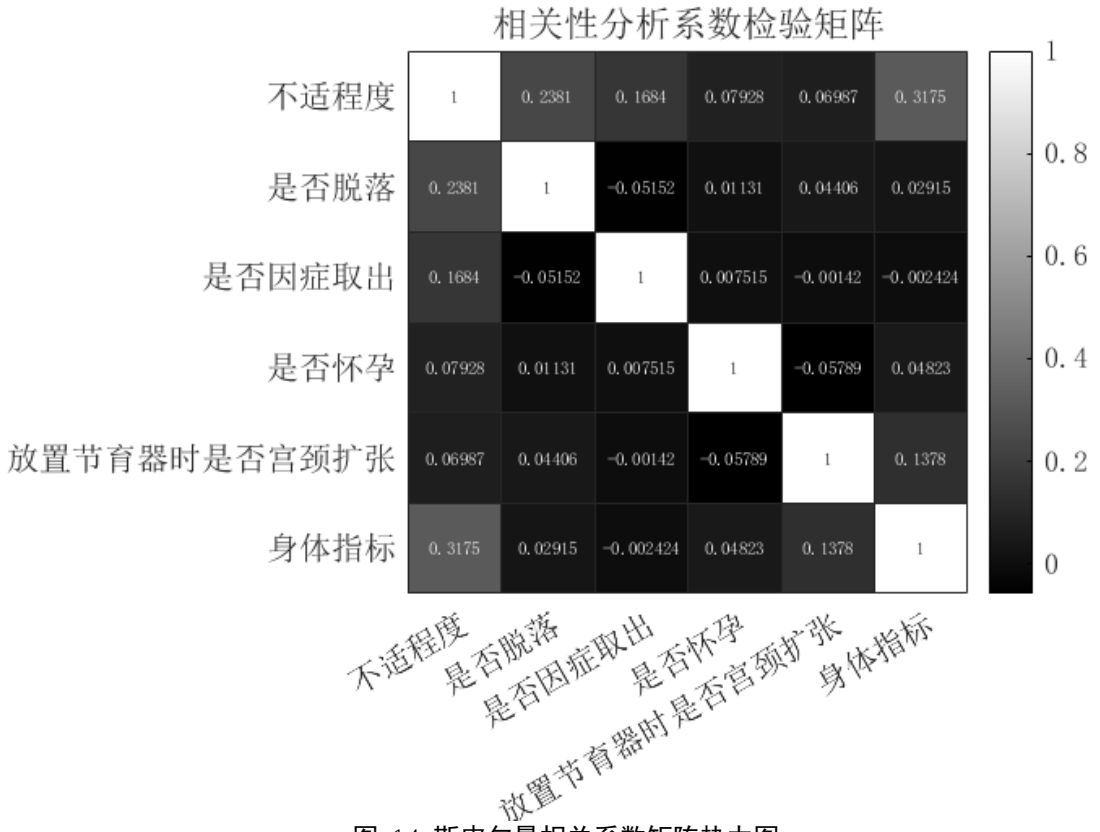

图 14 斯皮尔曼相关系数矩阵热力图

基于样本较大，我们对两两指标之间进行假设检验，计算检验值：

$$r_s\sqrt{n-1} \sim N(0,1)$$

并求出对应的 P 值，原假设为没有显著差异，备择假设为有显著差异，若 P 值大于 0.05 则接受原假设，否则接受备择假设。

根据热力图我们可以看出：身体指标与其他五项指标（不适状况程度，脱落，因症取出，怀孕，放置节育器时宫颈扩张情况）的相关系数分别为 0.3175，0.02915，-0.002424，0.04823，0.1378。而不适状况程度由非经期出血、疼痛、经量多、分泌物增多、经期/周期异常构成，所以依据相关系数矩阵身体指标和随访主诉情况有以下关系：

① 身体指标与随访主诉情况中不适状况程度有最大关联，其中突出的指标为非经期出血，和分泌物增多，因为身体指标的数值越小代表越健康即不适程度低，并且身体状况越不健康，非经期出血和分泌物会增多，符合常识。

② 身体指标与其他的指标，比如疼痛，是否脱落等等，相关程度不是很大，这些指标可能受到节育器的质量等其他因素的影响。

（3）我们使用 SPSSPRO 软件，求解了岭回归对相同的六项指标进行分析，来判断受试者的身体状况是否是受试者出现不适状况程度的主要因素。其中因变量为不适状况程度，自变量为身体指标，脱落，因症取出，怀孕，放置节育器时宫颈扩张情况。得到的结果如下：

表 1 岭回归结果

K=0.022	非标准化系数		标准化系数	t	P	R^2	调整 R^2	F
	B	标准误	Beta					
常数	0.434	0.018	-	24.288	0.000***			
是否脱落	-0.171	0.036	-0.103	-7.383	0.000***			
是否因症取出	-0.150	0.034	-0.101	7.383	0.000***			
是否怀孕	0.091	0.066	0.052	1.89	0.059*	0.951	0.943	25.21(0.000***)
放置节育器时是否宫颈扩张	0.056	0.019	0.082	2.97	0.003***			
身体指标	0.201	0.004	0.191	-1.277	0.002***			
	因变量：不适程度							

注：***、**、*分别代表 1%、5%、10%的显著性水平

岭回归的结果显示：基于 F 检验显著性 P 值为 0.000***，水平上呈现显著性，拒绝原假设，表明自变量与因变量之间存在着回归关系。同时，模型的拟合优度 R^2 为 0.951，模型表现为较为较好。模型的公式：不适程度（同不适状况程度）=0.434-0.171×是否脱落-0.15×是否因症取出＋0.091×是否怀孕＋0.056×放置节育器时是否宫颈扩张+0.201×身体指标。

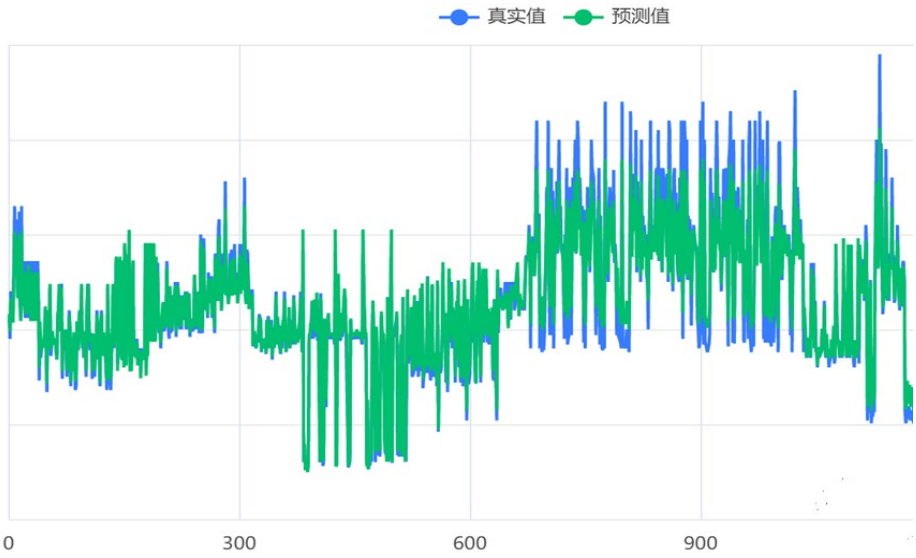

图 15 岭回归真实值和预测值对比图

并且我们通过观察到以不适程度为因变量，身体指标，脱落，因症取出，怀孕，放置节育器时宫颈扩张情况为自变量模型的预测结果与真实值大致趋势一致，即此模型有较强的适配性。

我们可以得出结论：由于岭回归中受试者的身体状况的比重最大为 0.201，所以我们认为受试者的身体指标是否是受试者出现不适状况的主要因素。

5.3.5 结果

针对问题，我们采用斯皮尔曼相关系数模型**对身体指标和随访主诉情况即五项指标（不适状况程度，脱落，因症取出，怀孕，放置节育器时宫颈扩张情况）的相关系数分别为 0.3175，0.02915，-0.002424，0.04823，0.1378。**得到结果：身体指标与随访主诉情况中不适状况程度有最大关联，其中突出的指标为非经期出血，和分泌物增多。

还使用岭回归模型对相同的六项指标进行分析，来判断受试者的身体状况是否是受试者出现不适状况程度的主要因素。**身体指标的回归系数（0.201）是 5 个因变量中最大的，所以我们认为受试者的身体指标是否是受试者出现不适状况的主要因素。**

5.4 问题 3 的模型建立与求解

5.4.1 数据预处理

根据题目的要求，我们先剔除掉附件一一，附件二一中作为对照组的使用 MCu 的功能性宫内节育器的被试得到附件三一，对第二组 408 个被试和第三组的 323 个被试，共计 803 个样本数据进行分析。

5.4.2 思路分析

我们需要解决的问题是在 VCu260 和 VCu380 中选择质量较好的一个。查阅相关文献后，我们认为衡量节育器质量要考虑身体健康程度，使用节育器的型号,是否成功避孕，是否脱落，是否因症取出，最终是否适应，不适程度，这七项指标[7]。

首先，对于身体指标。我们考虑到受试者的身体健康程度也会影响使用节育器的感受及效果。选取会影响被试身体状况的：年龄，月经周期，月经经期，既往应用节育器情况，宫腔深度和放置节育器时宫颈扩张情况六项指标。并采用主成分分析法计算六个指标的贡献率，作为身体指标中相应指标的权重，加权求和得到身体指标。

我们用 CRITIC 权重法对身体指标，使用节育器型号情况、最终是否不适、不适程度、是否怀孕、是否因症取出、是否脱落等七个指标确定对应权重，最后运用灰色关联分析法，计算第二组和第三组的被试的灰色加权关联度，作为 VCu260 和 VCu380 两类节育器的质量综合得分。最后综合得分高的，质量更优，更适合生产[9]。

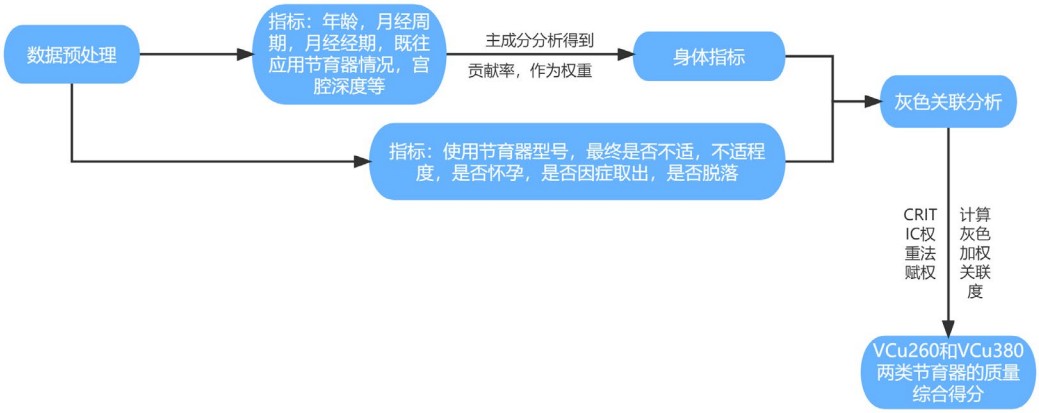

图 16 问题三流程图

5.4.3 主成分分析模型的建立

用主成分分析计算这六项指标各自的贡献率并作为权重：

（1）对原始六项数据标准化处理。

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j},\quad i=1,2,3,4,5,6,\quad j=1,2,3...803 \tag{7}$$

其中，$\mu_i=\frac{1}{803}\sum_{j=1}^{803}a_{ij}$；$s_i=\sqrt{\frac{1}{803-1}\sum_{j=1}^{803}(a_{ij}-\mu_i)^2}$，$i=1,2,3,4,5,6$

（2）计算相关系数矩阵 R。相关系数矩阵 $R=(r_{ij})_{6\times6}$，有

$$r_{ij}=\frac{\sum_{k=1}^{803}\tilde{a}_{ki}\cdot\tilde{a}_{kj}}{803-1},i,j=1,2,3...6 \tag{8}$$

其中，$r_{ii}=1$；$r_{ij}=r_{ji}$，r_{ij}为第 i 个指标与第 j 个指标的相关系数。

（3）计算特征值和特征向量。计算相关系数矩阵 R 的特征值 $\lambda_1\geq\lambda_2\geq\lambda_3\geq\lambda_4\geq\lambda_5\geq\lambda_6\geq0$，及对应的标准化特征向量 $\mu_1,\mu_2,\mu_3,\mu_4,\mu_5,\mu_6$，其中 $\mu_j=\left[\mu_{1j},\mu_{2j},\mu_{3j},\mu_{4j},\mu_{5j},\mu_{6j}\right]^T$，由特征向量组成 5 个新的指标变量：

$$y_1=\mu_{11}\tilde{x}_1+\mu_{21}\tilde{x}_2+...+\mu_{61}\tilde{x}_6,$$

$$y_2=\mu_{12}\tilde{x}_1+\mu_{22}\tilde{x}_2+...+\mu_{62}\tilde{x}_6,$$

$$\cdots$$

$$y_6=\mu_{16}\tilde{x}_1+\mu_{26}\tilde{x}_2+...+\mu_{66}\tilde{x}_6,$$

式中：y_1为第 1 主成分；y_2为第 2 主成分…y_6为第 6 主成分。

(4)计算特征值 λ_j，(j=1, 2…6)的信息贡献率。

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{6} \lambda_k}, j = 1,2,...,6 \qquad (9)$$

为主成分 y_j 的信息贡献率。

5.4.4 灰色关联分析模型的建立

为了确定两类节育器间的关联度，我们采用灰色关联度分析的方法进行计算。

（1）确定比较对象(评价对象)和参考数列(评价标准)。设评价对象有 803 个，评价指标有 7 个，

得出参考数列为:

$$x_0 = \{x_0(k) \mid k = 1,2,..., 7\}$$

比较数列为

$$x_i = \{x_i(k) \mid k = 1,2,..., 7\}, i = 1,2,..., 803$$

（2）我们采用 CRITIC 权重法确定各指标值对应的权重。设权重数组为 $w = [w_1,...,w_7]$ 其中 $w_k(k = 1,2,..., 7)$ 为第 k 个评价指标对应的权重。

随后，计算灰色关联系数：

$$\xi_i(k) = \frac{\min\limits_s \min\limits_t |x_0(t) - x_s(t)| + \rho \max\limits_s \max\limits_t |x_0(t) - x_s(t)|}{|x_0(t) - x_i(t)| + \rho \max\limits_s \max\limits_t |x_0(t) - x_s(t)|} \qquad (10)$$

为比较数列 x_i 对参考数列在第 x_0 个指标上的关联系数其中 $\rho \in [0,1]$ 为分辨系数其中，称 $\min\limits_s \min\limits_t |x_0(t) - x_s(t)|$，$\max\limits_s \max\limits_t |x_0(t) - x_s(t)|$ 分别为两级最小差及两级最大差。题中 ρ 取 0.5。

（3）最后，通过灰色加权关联度计算公式进行加权并进行评价分析。

5.4.5 问题三模型的求解

（1）通过 matlab 代码(PCA.m)来进行主成分分析，计算这六项指标各自的贡献率并作为权重：

权重如下图：

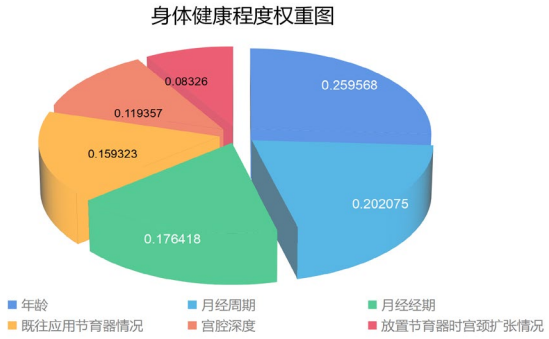

图 17 身体健康程度权重图

将这六项指标的值标准化后，用权重与对应的值相乘再将各项相加得到关于身体健康程度的值。

其次，对于节育器的理化指标。我们考虑被试所使用的节育器型号分为小，中，大三种。对这三种型号依次赋权 2,3,4，这是因为这样的权重可以清晰的表示出受试者使用节育器的各种情况。与对应指标相乘再相加作为此被试使用的节育器型号。

最后，对于随访时的主诉情况。由于题目所给的附录说明中提到若前期不适，后面月份适应，则可认为前期为适应期，不考虑节育器质量问题。所以我们只考虑十二月被试的适应情况，若十二月适应，则可认为此节育器适应性较好，若十二月不适应，则认为此节育器适应性较差。同时，要衡量节育器的质量，是否成功避孕，是否脱落，是否因症取出，还有第一问计算的不适程度都应该作为判断节育器质量的指标[10]。

（2）灰色关联分析的参考数列的确定和各种指标（身体状况、使用节育器型号情况、最终是否不适、不适程度、是否怀孕、是否因症取出、是否脱落）权重的确定。

① 灰色关联分析的参考数列的确定

表 2 参考数列一

身体状况	使用节育器型号情况	最终是否不适	不适程度	是否怀孕	是否因症取出	是否脱落
13.68349	0	0	0	0	0	0

针对参考数列我们给出如下解释：

身体状况：我们认为年龄和月经经期占身体状况的很大权重，所以年龄较年轻身体越好，月经经期来的较早说明身体状况较好，故最佳的值为 13.68349。

使用节育器型号情况：我们认为没有使用过节育器为最佳，更能反应节育器的作用。故最佳的值为 0。

最终是否不适和不适程度：最终没有不适，不适程度越低越好是使用节育器最好的情况，所以最佳的值为 0。

是否怀孕和是否因症取出，是否脱落：不怀孕和不因症取出，不脱落，说明节育器表现良好，所以最佳的值为 0。

考虑到灰色关联系数的计算，取 0 会导致异常值，会存在分母为 0 的异常情况，所以我们取接近于 0 的数字 0.01.

表 3 参考数列二

身体状况	使用节育器型号情况	最终是否不适	不适程度	是否怀孕	是否因症取出	是否脱落
13.68349	0.01	0.01	0.01	0.01	0.01	0.01

所以最终的参考数列为

$$x_0 = \{13.6835, 0.0100, 3.0000, 0.0100, 0.0100, 0.0100, 0.0100\}$$

②通过采用 CRITIC 权重法，各种指标的权重 $w=[w_1,...,w_7]$ 更新为

$w=[0.31542,0.36191,0.11427,0.06484,0.03427,0.05236,0.05692]$。

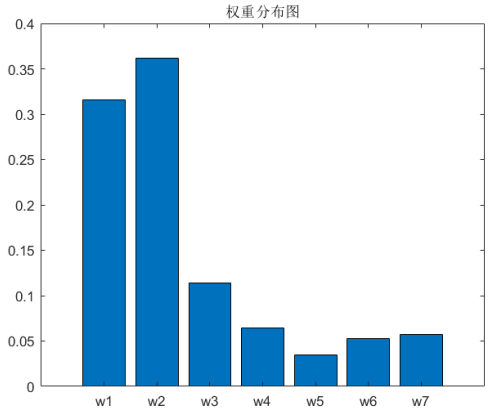

图 18 权重分布图

（3）通过 matlab 代码(PCA.m)来进行灰色关联分析。

根据灰色加权关联度的大小,对第二组的被试和第三组的被试进行排序,建立了关联序，而采用第二组的被试和第三组的被试加权关联度均值作为对 VCu260 记忆型宫内节育器和 VCu380 记忆型宫内节育器质量的综合得分，所以第二组的被试和第三组的被试关联度的均值越大，评价结果越好，对应的节育器的质量就更佳。

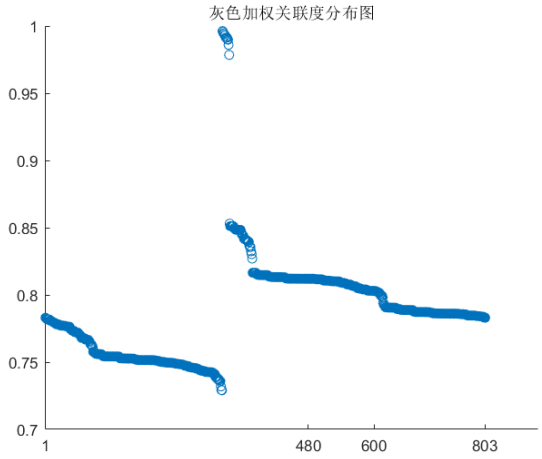

图 19 灰色加权关联度分布图

需要注意的是：前 480 个为第 2 组，后 323 为第 3 组。

经计算，第 2 组的均值即 VCu260 记忆型宫内节育器的质量综合得分为 0.756045849，第 3 组的均值即 VCu380 记忆型宫内节育器的质量综合得分为 0.80898292。

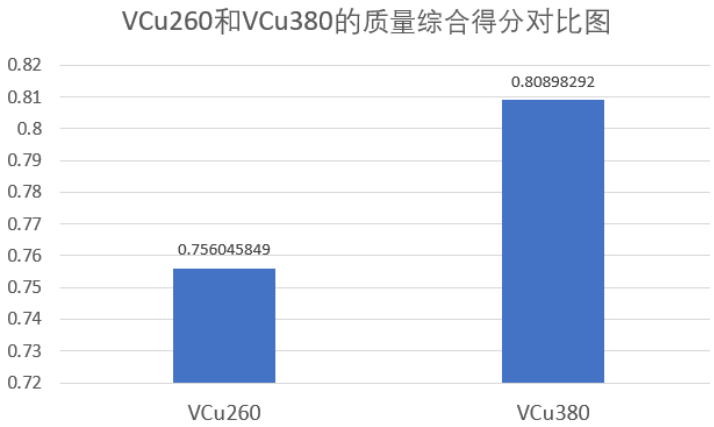

图 20 VCu260 和 VCu380 的质量综合得分对比图

5.4.5 结果

我们认为采用第二组的被试和第三组的被试加权关联度均值作为对 VCu260 记忆型宫内节育器和 VCu380 记忆型宫内节育器质量的综合得分，根据上图均值情况，**可以看出第二组的被试的灰色加权关联度均值小于第三组的被试的灰色加权关联度均值，所以 VCu260 记忆型宫内节育器的质量综合得分低于 VCu380 记忆型宫内节育器的质量综合得分**，所以我们认为 VCu380 比 VCu260 记忆型宫内节育器的质量更优，更适合生产。

5.5 问题 4 的模型建立与求解

5.5.1 数据预处理

数据同问题二。

5.5.2 思路分析

根据第三问建立的节育器质量模型，我们得到的结论是 **VCu380 的质量好于 VCu260。**将身体状况、使用节育器的型号、不适程度、最终是否不适、是否怀孕、是否因症取出、是否脱落作为七项指标，衡量对质量好坏的影响程度。为了探究影响宫内节育器质量的决定性因素，我们选用二分类逻辑回归的方法。将使用 VCu260 还是 VCu380 这个分类变量作为二分类逻辑回归的因变量。将上述提到的衡量节育器质量状况的七个指标作为自变量[8]。

图 21 问题四求解流程图

5.5.3 二分类逻辑回归模型的建立

二分类逻辑回归是一种基于概率模型的分类方法，是一种研究二分类因变量与一些影响因素之间关系的一种多变量分析方法。输出标记为 2 或 3（即使用 VCu260 或 VCu380），找到一个单调可微函数将线性回归模型的预测值映射到分类任务的真实标记 y 上，单调可微函数：

$$y = h_\theta(x) = g(\theta^T X) \qquad (11)$$

我们希望，$0 \le h_\theta(x) \le 1$，Sigmoid 函数即可以用作上式中的 $g(\bullet)$，Sigmoid 函数将 z 值转化为一个介于 0 和 1 之间的 y 值，将 Sigmoid 函数作为 $g(\bullet)$ 代入

$h_\theta(x)=g(\theta^T X)$中，得到：

$$h_\theta(x)=\frac{1}{1+e^{-\theta^T X}} \qquad (12)$$

以上对应的模型称为逻辑回归。利用 Logistic 回归进行分类的主要思想是：根据现有数据对决策边界线建立回归公式，以此进行分类。虽然其实质是一种分类方法，但是由于本题经过第三问，我们已知 VCu380 质量较好，我们决定根据此方法建立回归公式，将输出标记为使用这两种节育器，将所选指标作为自变量，根据观察回归系数的大小判断决定质量的主要因素。

5.5.4 二分类逻辑回归的求解

将自变量和因变量输入到 SPSSPRO 中得到模型的评价结果：

表 4 模型评价

似然比卡方值	P	AIC	BIC
1029.214	0.000***	1045.214	1082.721

注：***、**、*分别代表 1%、5%、10%的显著性水平

模型的似然比卡方检验的结果显示，显著性 P 值为 0.000***，水平上呈现显著性，拒绝原假设，因而模型是有效的。

基于此，我们得出的回归结果如下：

表 5 二分类逻辑回归结果

项	回归系数	标准误差	Wald	P	OR	OR 值 95% 置信区间 上限	OR 值 95% 置信区间 下限
常数	0.731	1.101	0.441	0.507	2.078	0.24	17.98
身体状况	0.007	0.063	0.013	0.908	1.007	0.89	1.141
使用节育器型号情况	-0.209	0.055	14.608	0.000***	0.811	0.729	0.903
不适程度	-0.587	0.308	3.631	0.057*	0.556	0.304	1.017
最终是否不适	-0.681	0.196	12.074	0.001***	0.506	0.345	0.743
是否怀孕	-16.723	1436.467	0	0.991	0	0	
是否因症取出	0.277	0.366	0.575	0.448	1.319	0.644	2.701
是否脱落	-0.51	0.367	1.935	0.164	0.601	0.293	1.232
因变量：组别							

注：***、**、*分别代表 1%、5%、10%的显著性水平

根据结果显示影响宫内节育器质量的决定因素是是否怀孕，也正符合题目附件中提到的生产节育器的前提是可以很大程度上避免怀孕。若使用节育器时怀孕，证明节育器质量有较大问题。

其次最终如果没有适应节育器，证明节育器质量也较差；不适程度越大，对节育器质量也会产生负向影响；如果节育器脱落，证明节育器固定效果不好，对节育器也会产生负向影响。即这三项指标在衡量质量时较重要，重要性仅次于是否怀孕。

根据 SPSSPRO 的图表显示：

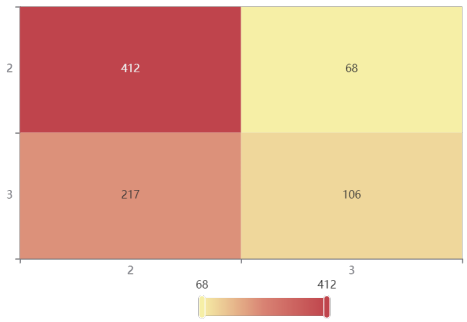

图 22 **混淆矩阵热力图**

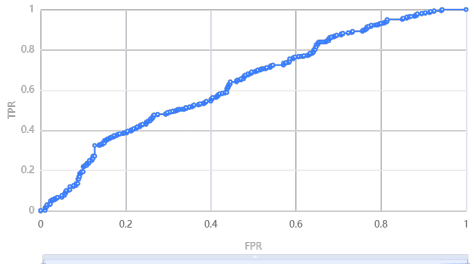

图 23 ROC **曲线**

从左图可以看出，如果根据这些指标推断节育器是 VCu260 还是 VCu380，有 518 个样本可以准确推断出来，只有 285 个样本会推断错误，再一次说明了这两个节育器的质量有较大差异。右图是图展示了 ROC 曲线图，用于衡量逻辑回归的分类效果。

右图的 ROC 曲线可以用于衡量逻辑回归的分类效果，ROC 曲线图把灵敏度（TPR）和特异度（FPR）结合，可以同时衡量两者关系。理想情况下，TPR 应该接近 1，FPR 应该接近 0。对右图分析，可以看出本模型分类效果较好，这两种节育器质量存在较大差异。

5.5.5 结果

综合上述分析，我们可以得出结论：建立的二分类逻辑回归模型具有较好的效果，根据此模型，可以得到影响宫内节育器质量的决定因素是是否怀孕。

而且节育器佩戴后被试舒适程度也很重要，所以如果到最后都没有适应节育器，对质量的影响也较大；且佩戴的不适程度越大，对节育器质量也会产生较大影响；如果节育器脱落，证明节育器固定效果不好，对节育器也会产生比较大的负面影响。这三项指标在衡量质量时较为重要，其重要性仅次于是否怀孕。

六、模型的评价及优化

6.1 误差分析

6.1.1 针对于问题 1 的误差分析

问题一中采用 T 检验的模型对数据进行分析，存在与重复测量相关的问题，会导致遗留效应。此外，T 检验虽然能确定样本间的区别，但不能帮助控制环境的影响，环境的变化可能会影响 T 检验的输出。

6.1.2 针对于问题 2 的误差分析

问题二中采用的斯皮尔曼相关系数的计算方法只能衡量两个变量间的单一关系，而不能运用于多个变量之间，在对数据进行排序和等级计算时的计算量较大，准确性会降低。

6.1.3 针对于问题 3 的误差分析

问题三中需要对数据中的指标赋予权重，由于没有过多参考只能通过数据的占比赋予权重，后面对数据进行降维，使数据的权重比重容易看出，二次计算出来的信息准确得到了有效提升。

6.1.4 针对问题 4 的误差分析

逻辑回归模型只能处理线性问题，而问题四中我们所使用的数据的决策边界不一定是线性的，逻辑回归模型的预测效果会不如其他非线性机器学习模型。

6.2 模型的优点（建模方法创新、求解特色等）

（1）基于附件所给的数据，对数据进行了全面的预处理，包括数据的完整性分析、数据的有效性及变量的分析与处理，为数据的正确使用打下了良好的基础；

（2）主成分分析法减少了变量的个数，又避免了确定权重时的主观随意性；

（3）通过熵权法，充分利用数据信息，模型自动给出了权重指标，使结果更加合理有效。

6.3 模型的缺点

（1）熵权法确定权重取决于数据本身，但不是最好的确定指标权重的方法；

（2）指标种类较多，可能存在多重共线性，对结果造成不利的影响。

6.4 模型的改进

（1）针对缺点 1，在时间和资源允许的前提下，进行基于专家打分的层次分析法确定各个指标的权重，从而保证权重的有效性；

（2）针对缺点 2，可先对数据进行多重共线性变量的判断和剔除，使得结果更合理。

参考文献

[1] 王一琳,石琴,张燕华.节育环危机[J].叙事医学,2021,4(05):364-366.

[2] 马小凤.妇科计划生育中放环手术时应用风险管理的效果研究[J].名医,2022(10):50-52.

[3] 李景海.放置宫内节育器对女性生殖健康的影响因素分析[J].世界最新医学信息文摘,2017,17(07):83.

[4] 鲍海霞.六种宫内节育器临床效果观察[J].大家健康(学术版),2014,8(12):190-191.

[5] 陈都,李圆媛,陈彧.基于 t 检验和逐步网络搜索的有向基因调控网络推断算法[J/OL].计算机应用:1-8[2023-05-14].http://kns.cnki.net/kcms/detail/51.1307.TP.20230424.1522.004.html

[6] 李元,刘雨田,冯立伟.基于斯皮尔曼相关分析的非线性动态过程特征提取与故障检测[J].山东科技大学学报(自然科学版),2023,42(02):98-107.DOI:10.16452/j.cnki.sdkjzk.2023.02.011.

[7] 王琦.基于层次分析法的职业女性二孩生育意愿影响因素研究[J].经济研究导刊,2022(01):144-148.

[8] 邹媛. 二元逻辑回归模型中几类一阶近似刀切估计的研究[D].贵州民族大学,2021.DOI:10.27807/d.cnki.cgzmz.2021.000106.

[9] 杨钊,郑治波,周浩明,徐顶巧,乐世俊,唐于平.基于层次分析-熵权法和网络药理学的蓝布正质量标志物研究[J/OL].中国中药杂志:1-11[2023-05-14].DOI:10.19540/j.cnki.cjcmm.20230428.201.

[10]杨月华,周健,施雯慧,张敏,许豪勤,杭桂芳.宫内节育器严重伤害事件报告质量评估及影响因素分析[J].中国医药导报,2018,15(08):129-133.

附 录

附录一：

spearman.m 问题二：进行斯皮尔曼相关性分析

```
T = table2array(readtable("斯皮尔曼相关性分析数据.xlsx","Range",'A2:F1152'));
data = zscore(T);
[xiangguan,p_value] = corr(data,'type','Spearman');

%绘画热力图
index_name = {'不适程度','是否脱落','是否因症取出','是否怀孕','放置节育器时是否宫颈扩张','身体指标'};
y_index = index_name;
x_index = index_name;

figure(1)

% 字号 12 字体宋体，随意改变，显示默认配色
H = heatmap(x_index,y_index,xiangguan,'FontSize',12,'FontName','宋体');
H.Title = '相关性分析系数检验矩阵';
colormap("colorcube")
```

附录二：

PCA.m 问题三：进行主成分分析

```
%% 读取数据
A=xlsread('附件三.xlsx', 'C3:H805');

% Transfer orginal data to standard data
a=size(A,1);    % Get the row number of A
b=size(A,2);    % Get the column number of A
for i=1:b
      SA(:,i)=(A(:,i)-mean(A(:,i)))/std(A(:,i));   % Matrix normalization
end

% Calculate correlation matrix of A.
CM=corrcoef(SA);

% Calculate eigenvectors and eigenvalues of correlation matrix.
[V, D]=eig(CM);

% Get the eigenvalue sequence according to descending and the corrosponding
% attribution rates and accumulation rates.
for j=1:b
      DS(j,1)=D(b+1-j, b+1-j);
end
for i=1:b
      DS(i,2)=DS(i,1)/sum(DS(:,1));
      DS(i,3)=sum(DS(1:i,1))/sum(DS(:,1));
end

% Calculate the numvber of principal components.
```

```
T=0.9;   % set the threshold value for evaluating information preservation level.
for K=1:b
      if DS(K,3)>=T
            Com_num=K;
            break;
      end
end

% Get the eigenvectors of the Com_num principal components
for j=1:Com_num
      PV(:,j)=V(:,b+1-j);
end

% Calculate the new socres of the orginal items
new_score=SA*PV;

for i=1:a
      total_score(i,2)=sum(new_score(i,:));
      total_score(i,1)=i;
end
new_score_s=sortrows(total_score,-2);
%%

disp('17：58')
% 显示结果
disp('特征值及贡献率：')
DS
disp('阀值 T 对应的主成分数与特征向量：')
Com_num
PV
disp('主成分分数：')
new_score
disp('主成分分数排序：')
new_score_s
```