2023 年第八届“数维杯”大学生
数学建模挑战赛论文

题　　目　**统计学观点下宫内节育器的生产评价分析**

摘　要

在生物医疗技术迅猛发展之下，宫内节育器（IUD）因其具有相对安全、有效和简便等优点逐渐受到广大妇女青睐，目前已成为我国育龄妇女的主要避孕措施。本文利用多种统计学方法对受试者的身体指标、节育器的理化指标和随访主诉情况进行建模分析，通过综合评价的手段在 Vcu260 型和 VCu380 型节育器中选择一个更优的产品进行生产，并且给出了影响节育器质量的相关因素。

对于**数据处理**方面，对数据的缺失值进行了替换填充，并且对于数据特征中重复、异常的数据点进行了处理，利用题目所给的条件对附件 2 的数据进行了预处理。

针对问题 1，观察数据发现身体指标为定量数据、理化指标和主诉情况为定性数据；首先通过数据形态验证了身体指标服从**正态分布**，并对身体指标进行了**独立样本 t 检验**，对理化指标和主诉情况进行了**卡方检验**，发现只有身体指标中的初潮年龄、月经周期和主诉情况中的脱落、因症取出…等 8 个指标没有显著性差异，其余的 9 个指标存在显著性差异，最后还分析了出现差异性可能的原因。

针对问题 2，首先对数据进行了皮尔逊相关性分析，初步分析了身体指标与主诉情况的联系，然后利用**典型相关分析**模型再次衡量了身体指标与主诉情况的联系:发现**宫颈深度**使主诉情况**倾向好**的方面，**月经经期**使主诉情况**倾向不好**的方面，并且通过分析主诉情况发现**节育环对怀孕的抑制是十分有效的**，负面的主要影响是**分泌物增多和经期/周期异常**。最后通过比较身体指标与主诉情况、理化指标与主诉情况的典型相关系数大小判断出**身体指标不是其主要原因**，理化指标影响力比身体指标**多 12.64%**。

针对问题 3，首先建立了**二级指标评价体系**，并将所有指标都转变为了正向型指标，接着利用**熵权法**确定各个指标的权重，然后对样本按照 VCu260 型和 VCu380 型节育器的标准进行了分类，利用 **Topsis 评价模型**对节育器质量进行多级评价，最后按加权平均的方法计算了每个样本对节育器质量的总评分，以两个类别的样本总评分的均值来判断哪个类型的节育器更优。最后计算得到 VCu260 型和 VCu380 型节育器均分为 0.6921 和 0.7001，**VCu380 型节育器优于 VCu260 型节育器，更适合生产**。

针对问题 4，首先利用问题 3 的权重模型初步分析出理化指标是影响节育器质量的决定因素；接着利用身体指标和理化性质对于是否出现不适症状进行了 **Logistic 回归分析**，同样可以发现**节育器的理化指标是影响节育器质量的决定因素**，其中**使用大型节育器是负向因素，放置节育器时宫颈扩张情况是正向因素**。

在模型的验证中，利用斯皮尔曼相关系数对皮尔逊相关系数进行了**误差分析**，发现这两种方法计算出来的相关系数相差不大，最大误差不超过 3%，进而验证了模型的**稳定性**。

关键词　独立样本 t 检验；卡方检验；相关性分析；熵权法；Topsis；Logistic 回归

目 录

一、 问题重述

在生物医疗技术迅猛发展之下，宫内节育器（IUD）因其具有相对安全、有效和简便等优点逐渐受到广大妇女青睐，成为我国育龄妇女的主要避孕措施。据悉，我国约有 70 %的妇女选用 IUD 作为避孕方法，该人数占到世界 IUD 避孕总人数的 80 %，可见投入生产 IUD 并广泛运用于各大医院，不仅有利于进一步提高我国的医疗水平，同时给广大妇女及其家庭带来许多的便利。

当前某公司利用镍钛记忆合金丝支架研发了两种型号的 IUD，但是目前发明的宫内节育器在给女性带来便利的同时，也可能会引起一些不良的生理症状，因此探究 VCu260 与 VCu380 这两种型号哪个更适合生产对于人们的生命健康具有十分重要的意义。

问题 1: 基于附件 1 和附件 2，试着分析两个医院的临床数据有无显著性差异，若结果显示存在显著性差异，进一步分析导致这种差异的因素。

问题 2: 利用附件 1 和附件 2，分析受试者的身体指标与随访主述情况二者之间的联系，同时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。

问题 3: 根据受试者的身体指标、节育器的理想化指标与随访时的主述情况等相关数据，搭建节育器质量模型，进一步分析对比 VCu260 与 VCu380 这两种型号的 VCu 记忆型宫内节育器的质量哪个更优，更适合投入生产。

问题 4: 基于问题 3 构建的节育器质量模型，探究影响宫内节育器的决定因素。

二、 问题分析

2.1 问题 1 的分析

对于问题 1，分析两个医院的临床数据有无显著性差异，并针对有显著差异的指标原因探究。观察附件一的数据可以发现，临床数据指标主要分为两个方面，一是受试者的身体指标包括年龄、初潮年龄、月经周期、月经经期和宫颈深度；二是节育器的理化指标包括既往节育器使用情况（IUD、无和其他）、节育器型号（小、中和大）和放置节育器时扩张情况，其中身体指标为定量变量，理化指标为定性变量。通常差异性分析可以通过 t 检验、卡方检验和非参数检验进行，这三种方法的适用范围不同，t 检验要求变量为定量变量且服从正态分布，卡方检验主要用于定性变量，非参数检验主要用于不满足正态分布的数据。首先对数据进行预处理后，将数据分为定量和定性两类，定量的数据再接着进行正态分布检验，若服从正态分布，则采用 t 检验，反之则采用非参数检验；对于定性的数据采用卡方检验。经检验后，对于存在显著性差异的指标进行探究性因素分析，给出导致存在差异性可能的原因。

2.2 问题 2 的分析

对于问题 2，分析受试者的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是受试者出现不适症状的主要原因。首先，由于某些受试者的数据缺失，为此要对这些缺失值进行处理，并且如果第一、三个月有不适症状，但第六、十二没有不适症状，则认为该症状对该样本无影响。考虑到受试者的身体指标与随访主诉情况这两类指标里均包含多个子指标，而典型相关分析正好对应这种“多对多”的指标相关性分析，因此采用典型相关分析来分析这两类指标之间的联系。同样地，也可以将受试者的身体指标与随访主诉情况进行典型相关性分析，比较受试者的身体指标与随访主诉情况和节育器的理化指标与随访主诉情况典型相关性的值大小即可以判断受试者的身体指标是否是受试者出现不适症状的主要原因。

2.3 问题 3 的分析

对于问题 3，需要根据受试者的身体指标、节育器的理化指标与随访主诉情况，建立质量评价模型，来判断 VCu260 型节育器和 VCu380 节育器哪个更适合生产。首先，建立多级评价指标体系，并需要考虑指标的性质，将全部指标都转变为正向型指标；由于每一个指标的影响程度都不相同，因此需要确定各个指标的权重，本文采用熵权法确实权重；接着建立 Topsis 节育器质量评价体系，分别计算 3 个一级指标的评分，再对这 3 个一级指标评分按加权求和计算每一样本的总评分。最后，按 VCu260 型节育器和 VCu380 节育器的样本进行分类，计算这两类节育器的评分均分，均分高的则更适合生产。

2.4 问题 4 的分析

对于问题 4，需要结合问题 3 的节育器质量模型探究影响节育器质量的决定因素。由于问题 3 已经通过熵权法计算了每一个指标的权重，故可以依据这个指标权重来判断哪个指标对节育器质量的影响更大。为进一步验证指标的重要性，建立受试者的身体指标和节育器的理化指标与是否出现不适症状的 Logistic 回归模型，通过其指标的影响系数以及显著性进一步说明哪个指标对节育器质量影响更大。

三、 模型假设

1) 假设数据来源真实有效；
2) 假设除题目所给指标外，其他方面的指标对节育器的质量无影响；
3) 假设两个医院的数据为独立数据，不相互影响。

四、 定义与符号说明

符号定义	符号说明
S	样本标准差
σ^2	方差
N_O	观察频数
N_T	理论频数
R	相关系数
$\mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})$	数组$\boldsymbol{X}$和$\boldsymbol{Y}$的协方差
$\widetilde{\boldsymbol{X}}(\tilde{x})$	对$\widetilde{\boldsymbol{X}}(\tilde{x})$标准化后的数组(值)
$\boldsymbol{P}$	概率矩阵
$w(W)$	权重
$Score_i$	第i个样本对节育器质量的评分

五、 模型的建立与求解

5.1 数据处理

在对数据进行建模之前，首先要检查数据是否缺失以及需要将数据处理成模型所需要的数据，下面首先对数据的缺失值进行处理。

5.1.1 缺失值处理

由于部分受试者在随访时难以联系，造成失访，会有部分数据缺失，一院随访主诉情况部分缺失值如下:

表 1 随访主诉情况缺失值

序号	组别	失访(1 月)	失访(3 月)	失访(6 月)	失访(12 月)
49	1	0	0	0	1
163	1	0	0	1	0
185	1	0	0	0	1
197	1	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮

由表 1 可以看到序号 49、163、185 和 197 的样本有在 6、12、12 和 12 月份的缺失值。一般来说，定性数据的缺失值采用众数进行填充，但由于存在缺失值的样本比例较小(见表 2)，并且若本来数据没有缺失时，该月份对应的症状发病率也非常小，因此存在缺失值的样本在该月份认为没有发病。

表 2 各月份缺失值比例

缺失月份	1 月	3 月	6 月	12 月
缺失比例(%)	0.2342	2.4980	4.1374	3.4348

经检查，其他数据无缺失情况。

5.1.2 重复值、异常值处理

经检查，发现附件 2 的二院随访的节育器使用后主诉情况的数据中出现重复的序号和组别，以及在“有不适人数”指标的统计方法不统一，有些位置只统计该月份是否出现不适症状，有些位置却统计该月份出现不适症状的次数，或者统计数据不对，这些原因造成的数据的异常。

表 3 重复值、异常值部分数据

Excel 索引	序号	组别	有不适人数(1 月)	有不适人数(3 月)	有不适人数(6 月)	有不适人数(12 月)
235	9	2	1	**3**	0	0
429	9	2	1	0	0	0
236	15	2	0	1	**3**	0
430	15	2	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

经统计，在附件 2 的二院随访的节育器使用后主诉情况中重复值的比例为 15.0899%(占总体比例为 3.5884%)，异常值比例为 13.9506%(占总体比例为 6.1289%)。

对于同一个数据面板，数据类型应为同一个类型，考虑到重复值和异常值的比例较小，因此对于重复值采取删除处理，对于异常值重新修正成为统计该月份是否出现不适症状。

5.1.3 数据预处理

附件 1 和附件 2 提供的定性数据对于未填写的位置采取保留空白处理，填写了的位置则为“1”,空白的位置会导致软件读取数据或者计算时带来不必要的麻烦，因此对于空白的位置采用“0”填充，使该定性变量转化为 01 变量。

另外，在问题附录中提到“若受试者在使用宫内节育器前期出现不适症状，但后期症状消失，可认为受试者适应该节育器。”，因此对于第六、十二月没有不适症状的样本认为该症状对节育器质量无影响。因此，根据此说明可以对数据进行处理，将第 1、3、6 和 12 月统一成为一个变量，处理过程见表 4。

表 4 随访主诉情况数据预处理

序号	组别	疼痛				疼痛改	经量多				经量多改
		1	3	6	12		1	3	6	12	
270	1	1	1	0	0	0	1	0	0	0	0

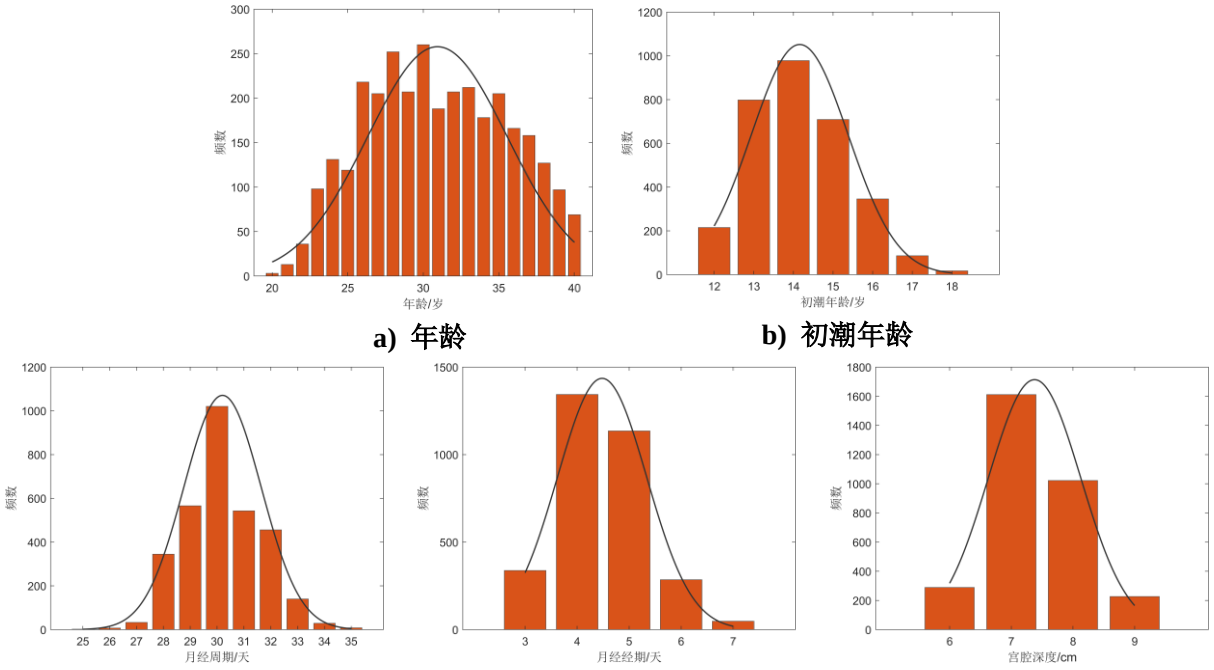
281	1	1	0	0	0	0	1	1	1	1	1
352	1	0	1	0	0	0	0	1	0	0	0
389	1	1	1	1	0	1	1	0	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

另外，由于随访时只对出现不适情况的受试者进行记录,对于未出现不适情况的受试者不进行记录，导致了附件 1 和附件二的数据长度不一致，破坏了数据的完备性。因此，对于未记录的样本也需要进行补充，均表示为“0”表示未出现不适症状。

5.2 问题 1 的模型建立与求解

5.2.1 正态分布检验

对数据进行差异性检验分析之前，首先要做各指标的正态分布检验，针对于受试者的身体指标(年龄、初潮年龄、月均周期、月经经期和宫颈深度)这 5 个定量指标进行正态分布检验:



a) 年龄　b) 初潮年龄

c) 月均周期　d) 月经经期　e) 宫颈深度

图 1 正态分布检验

由图 1 可以看到，对于这 5 个定量指标，其数据形态均呈现“钟形”特征，并且与正态分布趋势线趋势相吻合，因此可认为受试者的身体指标服从正态分布。

5.2.2 独立样本 t 检验

t 检验是通过均值比较两组定量数据之间的差异。t 检验的前提是两组数据来自正态分布的群体，满足独立性，满足方差齐性，。独立样本 t 检验，该检验用于检验两组非相关样本被试所获得的数据的差异性。

t 检验的前提之一服从正态分布已在 5.2.1 节进行了验证，而两组样本独立的条件

在假设的第三点已说明，为此还缺少方差齐性的条件。对于方差齐性，并不是硬性要求，这是由于无论方差齐性还是非方差齐性同样可以构造 t 统计量，只是形式不同而已。

设$S_{\bar{x}_1}, S_{\bar{x}_2}$为两组独立样本的样本标准差，满足下式

$$S_{\bar{x}_1}=\sqrt{\frac{\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)}{n_1-1}}, S_{\bar{x}_2}=\sqrt{\frac{\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)}{n_2-1}} \tag{1}$$

式中，$\bar{x}_1, \bar{x}_2$为第一和第二组样本的均值，n_1, n_2为第一和第二组样本量，则对应的方差为$\sigma^2_{\bar{x}_1}, \sigma^2_{\bar{x}_2}$为

$$\sigma^2_{\bar{x}_1}=\frac{S^2_{\bar{x}_1}}{n_1}, \sigma^2_{\bar{x}_2}=\frac{S^2_{\bar{x}_2}}{n_2} \tag{2}$$

- 若为方差齐性时,有均值标准差的采用加权平均法处理，并且由独立性假设可得

$$\begin{aligned} S^2_{\bar{x}_1-\bar{x}_2} &= wS^2_{\bar{x}_1}+w_2S^2_{\bar{x}_2} \\ &= \frac{(n_1-1)S^2_{\bar{x}_1}+(n_2-1)S^2_{\bar{x}_2}}{(n_1-1)+(n_2-1)} \end{aligned} \tag{3}$$

则对应的标准误(混合标准误)为

$$\begin{aligned} \sigma_{\bar{x}_1-\bar{x}_2} &= \sqrt{\frac{S^2_{\bar{x}_1-\bar{x}_2}}{n_1}+\frac{S^2_{\bar{x}_1-\bar{x}_2}}{n_2}} \\ &= \sqrt{\left[\frac{(n_1-1)S^2_{\bar{x}_1}+(n_2-1)S^2_{\bar{x}_2}}{(n_1-1)+(n_2-1)}\right]\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \end{aligned} \tag{4}$$

- 若为非方差齐性时,直接按独立性相加即可得其标准误

$$\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{S^2_{\bar{x}_1}}{n_1}+\frac{S^2_{\bar{x}_2}}{n_2}} \tag{5}$$

采用 t 统计量代替标准误可以得到齐性方差和非齐性方差的 t 统计量

$$t=\begin{cases} \sqrt{\left[\frac{(n_1-1)S^2_{\bar{x}_1}+(n_2-1)S^2_{\bar{x}_2}}{(n_1-1)+(n_2-1)}\right]\times\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} & \text{方差齐性} \\ \sqrt{\frac{S^2_{\bar{x}_1}}{n_1}+\frac{S^2_{\bar{x}_2}}{n_2}} & \text{非方差齐性} \end{cases} \tag{6}$$

对应的原假设与备择假设为

H_0:两组变量无显著差异, H_1:两组变量有显著差异

利用 SPSS 可得

表 5 身体指标独立样本 t 检验结果

指标	方差齐性类型	莱文方差等同性检验		平均值等同性 t 检验				
		F	显著性	t	自由度	Sig.（双尾）	平均值差值	标准误差差值

年龄	假定等方差	200.369	0.000	5.841	3147	0.000	0.958	0.164
	假定等不方差	-	-	5.842	2964.170	**0.000**	0.958	0.164
初潮年龄	假定等方差	0.000	0.988	0.041	3147	**0.967**	0.002	0.044
	假定不等方差	-	-	0.041	3146.886	0.967	0.002	0.044
月经周期	假定等方差	51.394	0.000	-0.151	3147	0.880	-0.008	0.051
	假定不等方差	-	-	-0.151	3073.322	**0.880**	-0.008	0.051
月经经期	假定等方差	13.206	0.000	-12.805	3147	0.000	-0.382	0.030
	假定不等方差	-	-	-12.804	3131.315	**0.000**	-0.382	0.030
宫颈深度	假定等方差	32.656	0.000	2.535	3147	0.011	0.068	0.027
	假定不等方差	-	-	2.535	3111.514	**0.011**	0.068	0.027

由表 5 可以看到(加粗的位置)，年龄、月经经期和宫颈深度的 p 值为 0.000、0.000 和 0.011 小于 0.05，在 95%的置信水平下，拒绝原假设H_0，接受备择假设H_1，即认为两个医院样本的年龄、月经经期和宫颈深度有显著差异。而初潮年龄和月经周期的 p 值为 0.967 和 0.880 大于 0.05，在 95%的置信水平下接受原假设H_0，即认为两个医院样本的初潮年龄和月经周期无显著差异。

5.2.3 卡方检验

卡方检验是一种用途很广的计数资料的假设检验方法，属于非参数检验。主要是两个分类变量的关联性分析。根本思想在于比较理论频数和观测频数的吻合程度或者拟合优度问题，卡方检验统计量为

$$\chi^2 = \sum \frac{(N_O - N_T)^2}{N_T} \tag{7}$$

式中，N_O为观测频数，N_T为理论频数，对应的原假设与备择假设于独立样本 t 检验相同，同样利用 SPSS 可得

表 6 理化指标卡方检验结果

指标	皮尔逊卡方值	自由度	渐进显著性(双侧)
既往应用节育器情况	134.858	2	0.000
使用节育器型号情况	2043.662	3	0.000
放置节育器时宫颈扩张情况	750.021	1	0.000

由表 6 可知，既往应用节育器情况、使用节育器型号情况和放置节育器时宫颈扩

张情况的 p 值均为 0.000 小于 0.05，在 95%的置信水平下，拒绝原假设H_0，接受备择假设H_1，即认为两个医院样本的既往应用节育器情况、使用节育器型号情况和放置节育器时宫颈扩张情况有显著差异。

表 7 随访主诉情况卡方检验结果

指标	皮尔逊卡方值	自由度	渐进显著性(双侧)	指标	皮尔逊卡方值	自由度	渐进显著性(双侧)
脱落	0.030	1	0.864	经量多	0.424	1	0.515
因症取出	1.129	1	0.288	分泌物增多	5.305a	1	0.021
怀孕	7.230	1	0.001	经期/周期异常	0.026	1	0.873
非月经期出血	0.012	1	0.913	有不适人数	2.886	1	0.089
疼痛	12.007	1	0.001				

由表 7 可知，随访主诉情况中的怀孕、疼痛和分泌物增多的 p 值为 0.001、0.001 和 0.021，均小于 0.05，在 95%的置信水平下，拒绝原假设H_0，接受备择假设H_1，即认为两个医院样本随访主诉情况中的怀孕、疼痛和分泌物增多有显著差异。而随访主诉情况中的脱落、因症取出、非月经期出血、经量多、经期/周期异常和有不适人数的 p 值为 0.864、0.288、0.913、0.515、0.873 和 0.089，均大于 0.05，在 95%的置信水平下，接受原假设H_0，即认为两个医院样本随访主诉情况中的脱落、因症取出、非月经期出血、经量多、经期/周期异常和有不适人数无显著差异。

5.2.4 差异性检验结果分析

综合独立样本 t 检验和卡方检验的结果可得如下差异性检验结果:

表 8 临床数据差异性检验结果

一级指标	二级指标	是否有显著性差异
受试者的身体指标	年龄	是
	初潮年龄	否
	月经周期	否
	月经经期	是
	宫颈深度	是
节育器的理化指标	既往应用节育器情况	是
	使用节育器型号情况	是
	放置节育器时宫颈扩张情况	是
随访主诉情况	脱落	否
	因症取出	否

怀孕	是
非月经期出血	**否**
疼痛	是
经量多	**否**
分泌物增多	是
经期/周期异常	**否**
有不适人数	否

由表 8 可知，只有受试者的身体指标中的初潮年龄、月经周期和随访主诉情况中的脱落、因症取出、非月经期出血、经量多、经期/周期异常、有不适人数共 8 个指标没有显著性差异，其余的 9 个指标均有显著性差异，特别是节育器的理化指标这一类指标均为有显著性差异。

导致这些**差异性出现的原因**可能是这两家医院进行**样本抽取时不够科学**、**样本量不够多**以及**偶然群体误差**。

5.3 问题 2 的模型建立与求解

5.3.1 相关性分析

描述指标与指标之间的联系最合适的就是相关性分析了，而计算相关性又有多种方法，例如皮尔逊相关系数、斯皮尔曼相关系数等等。由于在 5.2.1 节已经对受试者的身体指标进行了正态分布的检验，验证了其数据是满足正态分布的，这就满足了皮尔逊相关系数的使用条件之一。皮尔逊相关系数还有一个条件是数据趋势呈现线性分布，其矩阵散点图如下:

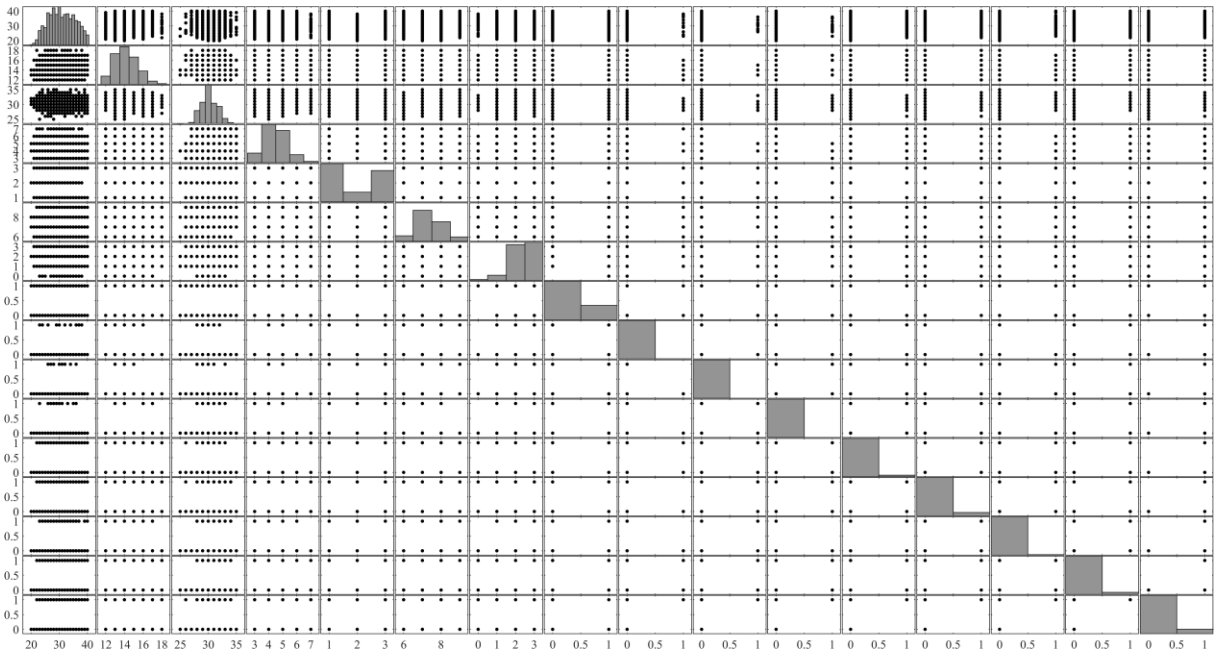


图 2 数据趋势矩阵散点图

由图 2 可以看到，数据趋势的“线性”关系几乎没有，为此对于数据呈现线性的这个条件并不满足。但是通过 6.1 节的分析后可以发现，皮尔逊相关系数和斯皮尔曼相关系数的误差非常小，因此这里先采用皮尔逊相关系数进行相关系数的计算。

其计算公式为

$$R_{XY}=\frac{\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})}{S_X S_Y} \tag{8}$$

式中， S_X, S_Y 为样本标准差计算见式(1)， $\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})$ 为协方差计算公式如下:

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{n-1} \tag{9}$$

利用式(8)可以计算得相关系数如下图 3:

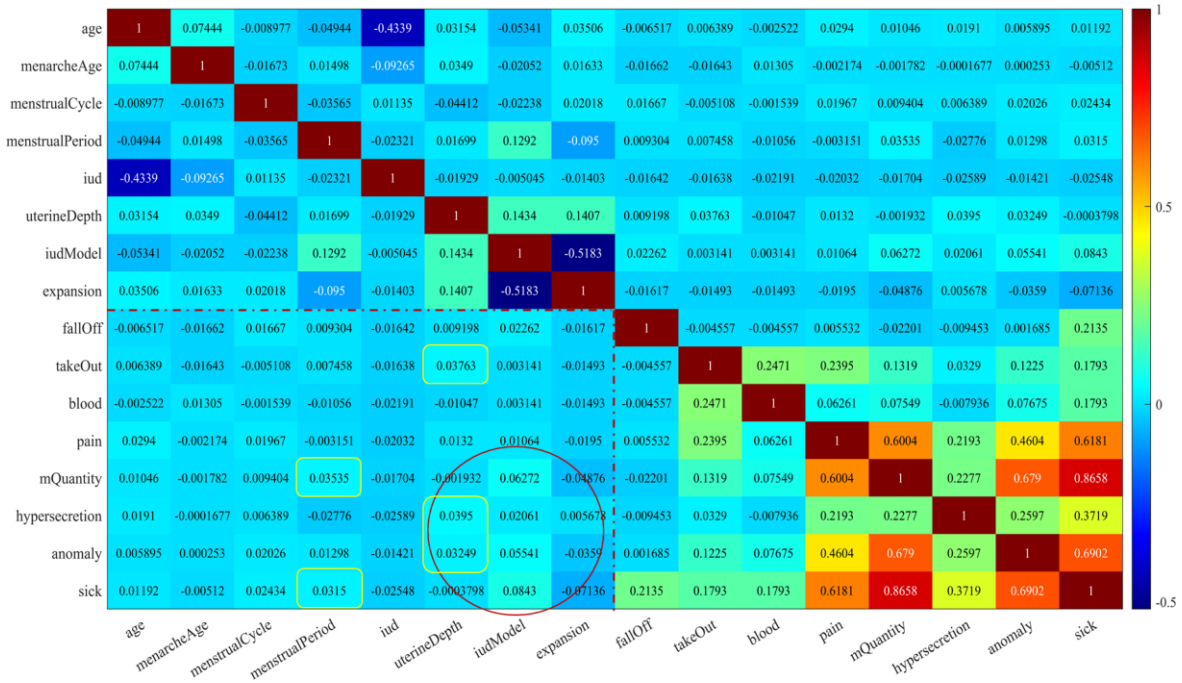


图 3 皮尔逊相关系数

由图 3 可以看到，在与随访主诉情况有关的相关系数区域(图 3 中用红色虚线框起来的区域)，受试者的身体指标与随访主诉情况相关系数值大于 0.03 的有 5 个位置(黄色矩形框框住的位置，有一个框同时框住两个位置)，分别是宫颈深度与因症取出(0.03763)、宫颈深度与分泌物增多(0.0395)、宫颈深度与经期/周期异常(0.03249)、月经经期与经量多(0.03535)、月经经期与有不适症状(0.03150)，这 5 个位置的系数且都为正，说明都是正相关的。

考虑受试者的身体指标是否是出现不适症状的主要原因，图 3 中相关系数值最大的为节育器的理化指标(图 3 中红色圆圈的位置，最大为 0.0843)，因此可以初步判断节育器的理化指标才是受试者身体出现不适状况的主要原因。

下面将对皮尔逊相关系数进行进一步拓展，采用典型相关分析分析其联系来验证分析的合理性。

5.3.2 典型相关分析模型的建立

典型相关分析是用于分析两组随机变量之间的相关程度(利用皮尔逊相关系数)的一种统计方法，它能够有效地揭示两组随机变量之间的相互关系。方法要求数据服从正态分布，而在 5.2.1 节中以及分析了受试者的 5 个身体指标服从正态分布，满足能够进行典型相关分析的条件；并且还要求数据能够计算皮尔逊相关系数，在 5.3.1 节中以及计算出了皮尔逊相关系数，并且在 6.1 节验证了皮尔逊相关系数与斯皮尔曼相关系数相差无几，因此可以进行典型相关分析。

假设$\boldsymbol{X}=(X_1,X_2,\cdots,X_p),\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_q)$是两个相互关联的随机向量，分别在两组变量中选取若干有代表性的综合变量$\boldsymbol{U}_i,\boldsymbol{V}_i$，使得每一个综合变量是原变量的线性组合,即

$$\begin{cases}\boldsymbol{U}_i=a_1^{(i)}X_1+a_2^{(i)}X_2+\cdots+a_p^{(i)}X_p=\boldsymbol{a}^{(i)}\boldsymbol{X}\\ \boldsymbol{V}_i=b_1^{(i)}Y_1+b_2^{(i)}Y_2+\cdots+b_q^{(i)}Y_q=\boldsymbol{b}^{(i)}\boldsymbol{Y}\end{cases}\tag{10}$$

其中

$$\boldsymbol{X}=\begin{bmatrix}X_{11}&X_{12}&\cdots&X_{1p}\\X_{21}&X_{22}&\cdots&X_{2p}\\\vdots&\vdots&\ddots&\vdots\\X_{n1}&X_{n2}&\cdots&X_{np}\end{bmatrix},\boldsymbol{Y}=\begin{bmatrix}Y_{11}&Y_{12}&\cdots&Y_{1q}\\Y_{21}&Y_{22}&\cdots&Y_{2q}\\\vdots&\vdots&\ddots&\vdots\\Y_{n1}&Y_{n2}&\cdots&Y_{nq}\end{bmatrix}\tag{11}$$

假设$p\leqslant q$，令$\boldsymbol{Z}_{(p+q)\times1}=\begin{bmatrix}\boldsymbol{X}_{p\times1}\\\hline\boldsymbol{Y}_{q\times1}\end{bmatrix}$服从正态分布，设样本量为$n$，则对$\boldsymbol{Z}_{(p+q)\times1}$标准化后，设对应的皮尔逊相关系数矩阵为$\boldsymbol{R}$，则有

$$\boldsymbol{R}=\left[\begin{array}{c|c}\underset{(p\times p)}{\boldsymbol{R}_{11}}&\underset{(p\times q)}{\boldsymbol{R}_{12}}\\\hline\underset{(q\times p)}{\boldsymbol{R}_{21}}&\underset{(q\times q)}{\boldsymbol{R}_{22}}\end{array}\right]\tag{12}$$

则可以得到矩阵$\boldsymbol{A}$和$\boldsymbol{B}$的样本估计：

$$\begin{cases}\boldsymbol{A}_{p\times p}=\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\\\boldsymbol{B}_{q\times q}=\boldsymbol{R}_{22}^{-1}\boldsymbol{R}_{21}\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\end{cases}\tag{13}$$

矩阵$\boldsymbol{A}$和$\boldsymbol{B}$具有相同的特征根λ^2，$\boldsymbol{a},\boldsymbol{b}$为对应的特征向量，即典型变量(原变量)线性组合的系数。

5.3.3 身体指标和主诉情况联系分析

利用 SPSS 可得

表 9 受试者的身体指标与随访主诉情况典型相关分析结果

典型相关个数	相关性	特征值	威尔克统计	F	分子自由度	分母自由度	显著性

1	0.081	0.007	0.988	1.101	35.000	13198.597	0.313
2	0.038	0.003	0.994	0.728	24.000	10948.384	0.828
3	0.033	0.001	0.998	0.457	15.000	8665.798	0.961
4	0.026	0.001	0.999	0.396	8.000	6280.000	0.923
5	0.018	0.000	1.000	0.327	3.000	3141.000	0.806

由于这 5 对典型变量均不显著，观察第一对典型变量，显著性水平为 0.313 还算可以接受并且其相关系数也是最大的。为此，姑且取第一对典型变量进行分析，其标准化典型相关系数如下图 4。

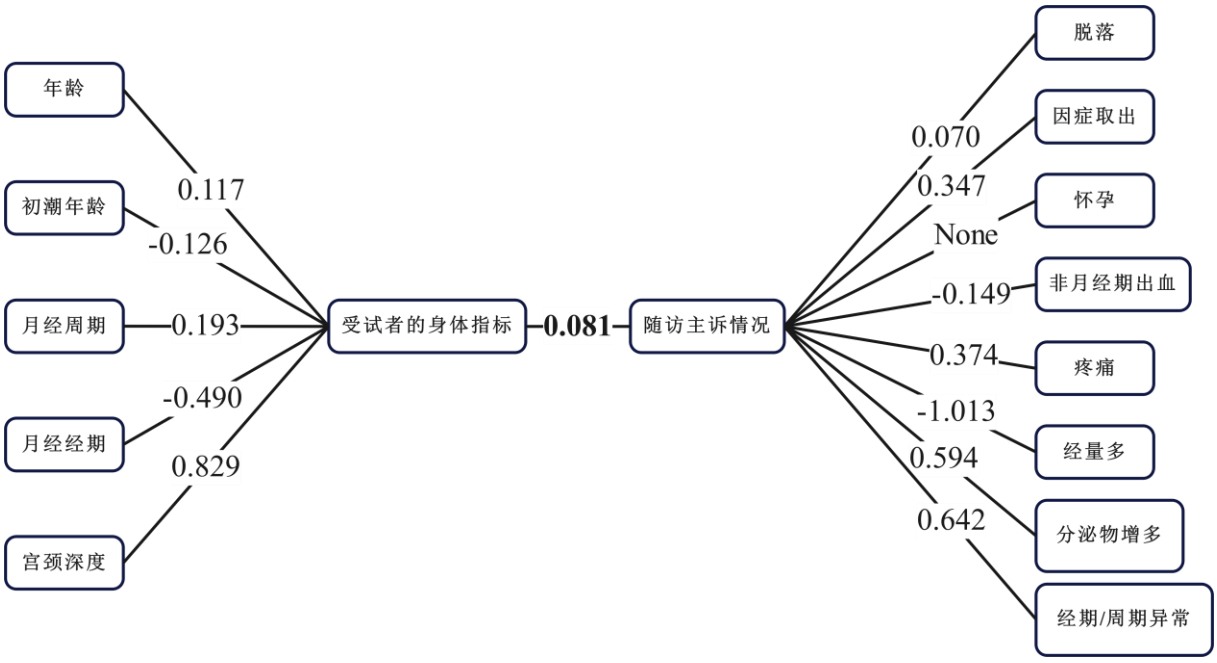


图 4 受试者的身体指标与随访主诉情况典型相关系数

由图 4 可知，受试者的身体指标与随访主诉情况直接的典型相关系数为 0.081>0，为正相关。对于受试者的身体指标，年龄、月经周期和宫颈深度相关系数为 0.117、0.193 和 0.829，这三个指标的相关系数大于 0，说明年龄、月经周期和宫颈深度正向影响受试者的身体指标，并且宫颈深度的相关系数绝对值最大，这也说明了**宫颈深度**有很大可能促进受访者倾向节育器**质量好**的一个重要原因；而初潮年龄和月经经期的相关系数为-0.126 和-0.490，小于 0 表现为负相关，说明初潮年龄和月经经期负向影响受试者的身体指标，其中月经经期的相关系数绝对值最大，这也说明了**月经经期**是导致受访者倾向节育器**质量差**的一个重要原因。

对于随访主诉情况，由于在随访主诉情况中的怀孕的人几乎没有，代入计算时导致了系数矩阵奇异，为此，需将怀孕这个指标排除在外进行计算。并且由此可知，节育环**对怀孕的抑制是十分有效的**。其中相关系数绝对值最大的为**经量多**，即表示经量多是人们对于节育器主诉的**最主要的方面**，接着是**分泌物增多**和**经期/周期异常**，相关系数为 0.594 和 0.642，说明了这两个方面是节育器对受试者**次要影响**。而**脱落**的相关系数为 0.070，是随访主诉情况中最小的指标，说明了节育器脱落的**现象并不是很多**。

下面考察受试者的身体指标是否是受试者出现不良状况的主要原因。

5.3.4 身体指标与理化指标的影响力分析

同样利用典型相关分析计算节育器的理化指标与随访主诉情况相关性，利用 SPSS 可得

表 10 节育器的理化指标与随访主诉情况典型相关分析结果

典型相关个数	相关性	特征值	威尔克统计	F	分子自由度	分母自由度	显著性
1	0.082	0.007	0.991	1.376	21.000	9014.063	0.117
2	0.058	0.001	0.997	0.669	12.000	6280.000	0.782
3	0.034	0.001	0.999	0.704	5.000	3141.000	0.620

由表 10 可知，这 3 对典型相关系数仍然不显著，同样考察第一对典型变量，其显著性水平为 0.117 接近 90%的置信水平还算可以接受并且其相关系数也是最大的。比较第一对典型相关系数的值大小，发现 $R_1^{body}=0.081<R_1^{iud}=0.082$，这说明节育器的理化指标对受试者出现不良状况的影响更大。

不过，这两个相关系数只相差了 0.001，相对误差为 1.22%，这说服力确实不太够，接着考虑节育器的理化指标与随访主诉情况所有(3 个)典型相关系数值之和

$$R_1^{iud}+R_2^{iud}+R_3^{iud}=0.174$$

也同样考虑受访者的身体指标与随访主诉情况的前 3 个典型相关系数值之和

$$R_1^{body}+R_2^{body}+R_3^{body}=0.152$$

显然 0.174>0.152，相对误差为 12.64%，这说明理化指标是受试者出现不良状况的主要原因，并且节育器的理化指标比受访者的身体指标的影响程度多 12.64%。

5.3.5 模型检验

下面对受试者的身体指标与随访主诉情况典型相关分析进行检验，这两个典型相关分析相应的方差解释比例如下表 11 和表 12。

表 11 身体指标与主诉情况典型方差比例

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	**0.498**	0.001	**0.305**	0.001
2	0.199	0.001	0.130	0.000
3	0.195	0.000	0.194	0.000
4	0.194	0.000	0.096	0.000
5	0.214	0.000	0.190	0.000

注:集合 1 为身体指标，集合 2 为主诉情况.

表 12 理化指标与主诉情况典型方差比例

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	**0.458**	0.003	**0.308**	0.001
2	0.281	0.000	0.163	0.000
3	0.261	0.000	0.127	0.000

注:集合 1 为理化指标，集合 2 为主诉情况.

由于前文只考虑了第一对典型相关变量，因此检验也只考虑第一对典型相关变量，由表 11 和表 12 可知，身体指标(理化指标)方差解释比例达到 0.498(0.458)，主诉情况的解释比例达到 0.305(0.308)，相对于其他的典型相关变量的比例较高，说明了采用第

一对典型相关变量来解释该类指标是具有代表性意义的。

5.4 问题 3 的模型建立与求解

5.4.1 评价指标体系的建立

问题 3 要求我们建立节育器的质量评价模型，分析 VCu260 和 VCu380 记忆性宫内节育器的质量优劣，而节育器的质量是与评价受试者的身体指标、节育器的理化指标和随访主诉情况有关的，这就需要建立这 3 类指标与节育器的质量的关系。

考虑到指标的类型并不统一，有正向型、中间型和负向型的指标，为此首先要将中间型的指标和负向型的指标转化为正向型的指标。

对于中间型的指标，可做如下变换

$$\tilde{x}_i = 1 - \frac{|x_i - x_{best}|}{M} \tag{14}$$

其中$M = \max\{|x_i - x_{best}|\}$表示为中间型的指标中最大的偏差值。

对于负向型的指标，可做如下变换

$$\tilde{x}_i = \max\{x_i\} - x_i \tag{15}$$

利用式(15)和式(14)即可将中间型的指标和负向型的指标转化为正向型的指标。

表 13 评价指标体系

一级指标	二级指标	指标类型	均值(众数)
受试者的身体指标	年龄	中间型	30.9387
	初潮年龄	中间型	14.1585
	月经周期	中间型	30.2026
	月经经期	中间型	4.4798
	宫颈深度	中间型	7.3769
节育器的理化指标	既往应用节育器情况	中间型	1
	使用节育器型号情况	中间型	3
	放置节育器时宫颈扩张情况	中间型	0
随访主诉情况	脱落	负向型	-
	因症取出	负向型	-
	非月经期出血	负向型	-
	疼痛	负向型	-
	经量多	负向型	-
	分泌物增多	负向型	-
	经期/周期异常	负向型	-

由于受试者的身体指标和节育器的理化指标无法明确判断指标类型，为此采取中间型处理，其最好的中间值为平均数或者众数，而随访主诉情况很明显是负向型指标。由于每个指标的重要性(权重)并不一致，为此，需要对经过指标正向化后的指标进行权重确定。

下面将采用熵权法确定指标的权重。

5.4.2 熵权法确定权重

熵权法的基本思路是根据指标变异性的大小来确定客观权重。对于某项指标，可以用熵值来判断某个指标的离散程度，其信息熵值越小，指标的离散程度越大，该指标对综合评价的影响(即权重)就越大。

其过程如下:

Step1: 对于已经正向化的n个指标和m个样本的矩阵$\boldsymbol{X}$

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{16}$$

对其进行标准化，设其标准化后的矩阵为$\widetilde{\boldsymbol{X}}$，则$\widetilde{x}_{ij}$可以表示为

$$\widetilde{x}_{ij}=\frac{x_{ij}-\min\{x_{1j},x_{2j},\cdots,x_{nj}\}}{\max\{x_{1j},x_{2j},\cdots,x_{nj}\}-\min\{x_{1j},x_{2j},\cdots,x_{nj}\}} \tag{17}$$

Step2: 计算概率矩阵$\boldsymbol{P}$，其中p_{ij}可以表示为

$$p_{ij}=\frac{\widetilde{x}_{ij}}{\sum_{i=1}^{n}\widetilde{x}_{ij}} \tag{18}$$

Step3: 计算第j个指标的信息熵e_j

$$e_j=-\frac{1}{\ln n}\sum_{i=1}^{n}p_{ij}\ln(p_{ij}) \tag{19}$$

其中若$p_{ij}=0$,则令$\ln(p_{ij})=0$。

Step4: 得到归一化权重，设第j个指标的信息效用值为d_j，则$d_j=1-e_j$，进而可得到归一化权重

$$w_j=\frac{d_j}{\sum_{k=1}^{m}d_k} \tag{20}$$

按照 Step1~Step4 的步骤进行计算可得各指标权重如下图 5。

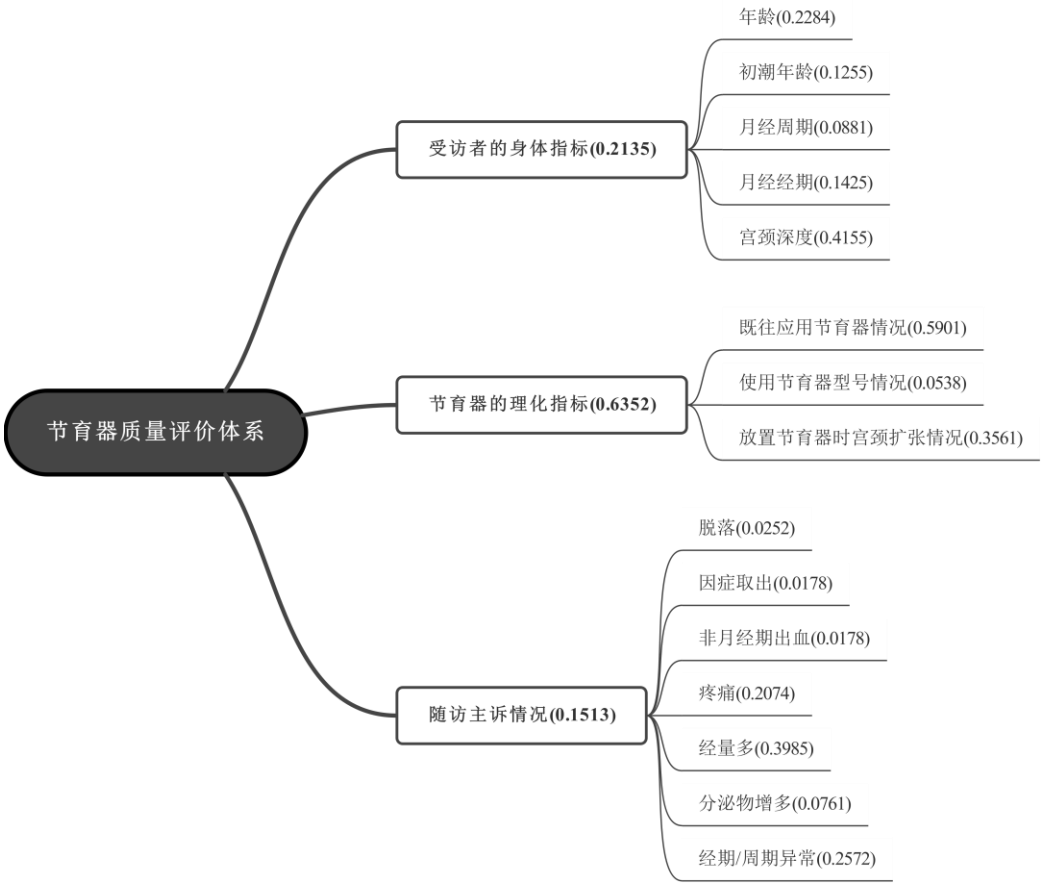


图 5 指标权重

5.4.3 Topsis 评价模型的建立

TOPSIS 法又称为优劣解距离法，是一种常用的组内综合评价方法，能充分利用原始数据的信息，其结果能精确地反映各样本对节育器质量评价之间的差距。基本过程为先对原始数据矩阵进行指标正向化后，再对正向化的矩阵进行标准化处理消除量纲的影响，然后分别计算各评价样本与最优样本和最劣样本间的距离，获得各评价样本与最优样本的相对接近程度，以此作为评价节育器优劣的依据。

首先由式(17)可得标准化后的数据矩阵$\widetilde{\boldsymbol{X}}_{n\times m}$，设$\widetilde{X}_j^-,\widetilde{X}_j^+$表示第$j$个指标的最小值和最大值

$$\begin{cases}\widetilde{X}_j^- = \min\left\{[\widetilde{x}_{1j},\widetilde{x}_{2j},\cdots,\widetilde{x}_{nj}]^T\right\}\\ \widetilde{X}_j^+ = \max\left\{[\widetilde{x}_{1j},\widetilde{x}_{2j},\cdots,\widetilde{x}_{nj}]^T\right\}\end{cases} \tag{21}$$

则第i个样本距离最小值和最大值的距离为

$$\begin{cases}D_j^- = \sqrt{\sum_{j=1}^{m} w_j\left(\widetilde{X}_j^- - \widetilde{x}_{ij}\right)^2}\\ D_j^+ = \sqrt{\sum_{j=1}^{m} w_j\left(\widetilde{X}_j^+ - \widetilde{x}_{ij}\right)^2}\end{cases} \tag{22}$$

式中，w_j为第j个指标的权重，具体值见图 5。则第i个样本距离对其节育器的评价分数为

$$Score_i = \frac{D_i^-}{D_i^- + D_i^+} \tag{23}$$

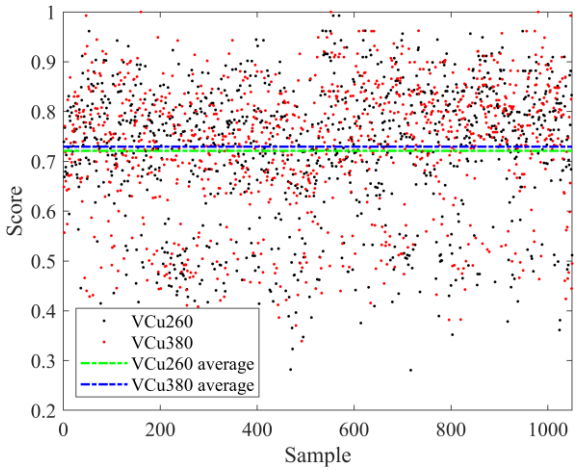
由于表 13 的评价体系为多级指标评价体系，因此需要对 Topsis 进行相应的改进。令$Score_i^{body}, Score_i^{iud}, Score_i^{state}$分别表示第$i$个样本的受试者的身体指标、节育器的理化指标和随访主诉情况的评价分数，$W^{body}, W^{iud}, W^{state}$表示这三个一级指标的权重，每个样本最后的总评分为这 3 个一级指标的加权评分

$$Score_j = W^{body} Score_j^{body} + W^{iud} Score_j^{iud} + W^{state} Score_j^{state} \tag{24}$$

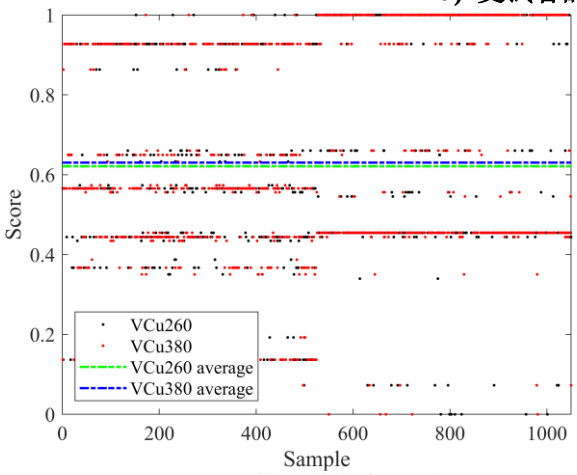
得到了每一个样本对节育器的评分后，再按每 VCu260 和 VCu380 节育器进行分类，即按组别为 2 和 3 进行分类，分类后计算每一类的均分，均分高的那一类节育器则质量更优。

5.4.4 评价结果

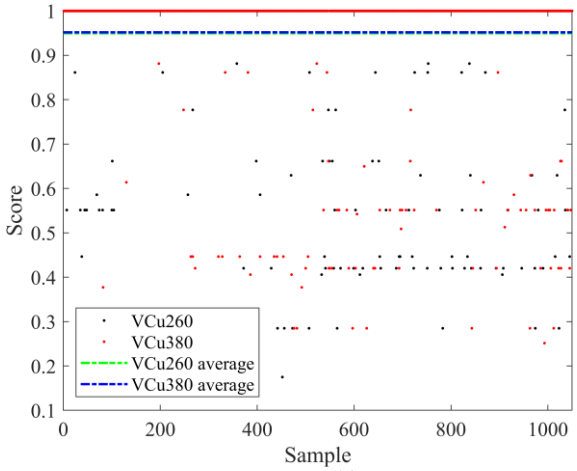
利用 Topsis 评价体系计算每一个样本对节育器质量的评分，并按照 VCu260 和 VCu380 节育器进行分类后可得如下图 6 的结果。



a) 受试者的身体指标评分



b) 节育器的理化指标评分



c) 随访主诉情况

图 6 一级指标评分

由图 6 可以看到，在一级指标中， VCu380 型节育器评分均大于 VCu260 型节育器，下面按照式(24)计算总评分，得到图 7。

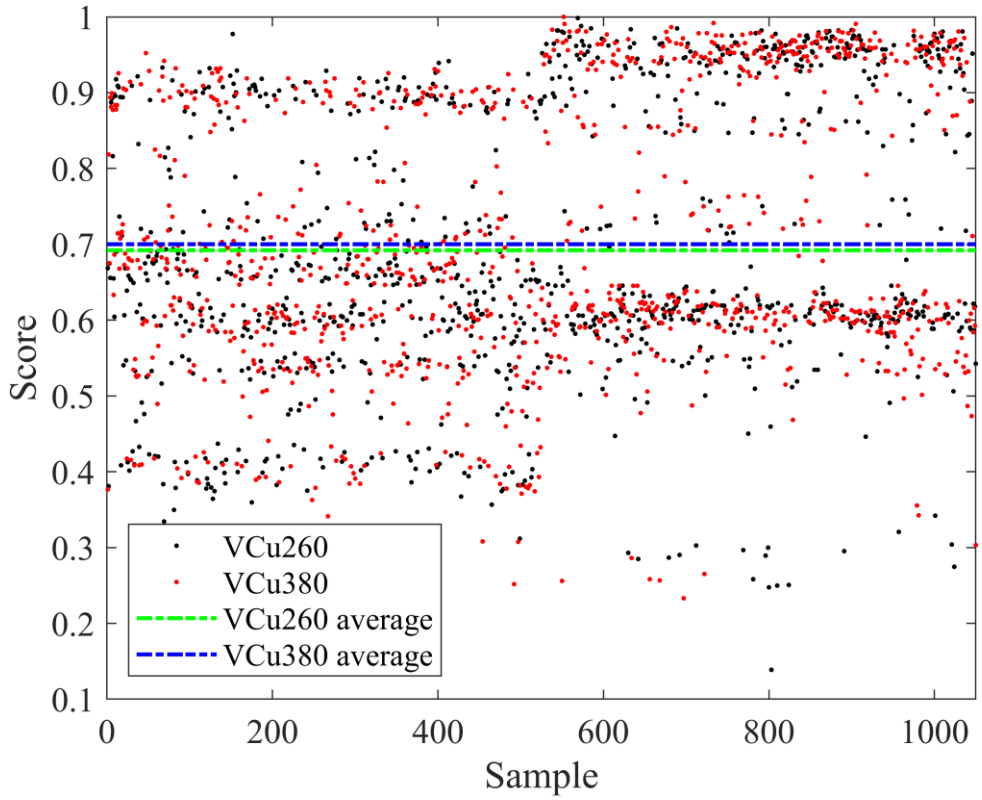


图 7 节育器质量总评分

由图 7 可知，VCu380 型节育器评分均分为 0.7001，Vcu260 型节育器评分均分为 0.6921，显然 0.7001>0.6921，说明了 **VCu380 型节育器优于 VCu260 型节育器**，更适合生产。

5.5 问题 4 的模型建立与求解

5.5.1 节育器质量影响探究分析

每个指标对节育器质量的影响程度都不一样，由 5.4.2 节以及图 5 可知，节育器的理化指标权重为 0.6352，受试者的身体指标权重为 0.2135，随访主诉情况权重为 0.1513。显然，权重越大对节育器的影响程度越大，则节育器的理化指标是影响节育器质量的主要因素，而节育器的理化指标又分为既往应用节育器情况、使用节育器型号情况和放置节育器时宫颈扩张情况，这 3 个二级指标的权重分别为 0.5901、0.0538 和 0.3561。故**影响最大的因素为既往应用节育器情况，其次是放置节育器时宫颈扩张情况**。

上述只是单单从权重方面进行分析，说服力有些欠缺，考虑到这 3 个一级指标的逻辑关系，受试者的身体指标和节育器的理化指标相当于自变量，而随访主诉情况相当于因变量，即身体指标和理化指标与主诉情况存在函数关系。而在随访主诉情况中，我们只关注是否出现不适症状，若出现不适症状则说明为节育器质量不好(记为 0)，若

没有出现不适症状则说明节育器质量好(记为 1)，这样就可以把因变量表示为 01 变量。下面将建立身体指标和理化指标与是否出现不适症状逻辑回归关系来进一步探究影响节育器质量的决定性因素。

5.5.2 Logistic 回归模型的建立

设各指标之间满足线性函数

$$\begin{aligned} y_i &= a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_m x_{im} \\ &= x_i' a \end{aligned} \tag{25}$$

对于节育器适应情况$\eta \in [0,1]$，若$\eta \geqslant 0.5$则表示为没有不适症状适应节育器，反之则为有不适症状不适应节育器。给定x的情况下，考虑y的两点分布

$$\begin{cases} P(y=1\,|x) = f(x,a) \\ P(y=0\,|x) = 1 - f(x,a) \end{cases} \tag{26}$$

写成一般形式

$$P(ture) = (f(x,a))^{y_i}(1-f(x,a))^{1-y_i} = 1-\eta \tag{27}$$

其中$f(x,a)$被称为连接函数，且$f(x_i'a) = f(x,a)$，$f(x,a)$函数的取法一般有两种：

(1) 标准正态分布的累积密度函数

$$f(x,a) = \int_{-\infty}^{x_i'a} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{28}$$

(2).*Sigmoid* 函数

$$f(x,a) = \frac{e^{x_i'a}}{1+e^{x_i'a}} = \frac{1}{e^{-x_i'a}+1} \tag{29}$$

由于式(29)有具体的解析表达式，而式(28)没有，因此采用式(29)作为连接函数。

5.5.3 Logistic 回归模型的求解

采用(29)式作为连接函数后，需对该模型进行求解，采用最大似然估计法,记

$$h_\theta(x) = \frac{e^{x_i'a}}{1+e^{x_i'a}} \tag{30}$$

则似然函数为

$$L(\theta) = \prod_{i=1}^{n} (h_\theta(x_i))^{y_i}(1-h_\theta(x_i))^{1-y_i} \tag{31}$$

对式(31)两边同时取对数

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \left(y_i \ln h_\theta(x_i) + (1-y_i)\ln(1-h_\theta(x_i))\right) \tag{32}$$

当$l(\theta)$最大时的a值就是所求的a。一般地，在函数最优化的时候习惯上求最小值，

因此，在$l(\theta)$前面添加一个负号得:

$$J(\theta)=-l(\theta)=-\sum_{i=1}^{n}\left(y_i\ln h_\theta(x_i)+(1-y_i)\ln(1-h_\theta(x_i))\right) \tag{33}$$

$J(\theta)$称为逻辑回归的损失函数，$J(\theta)$的值越小，拟合函数的效果就越好，下面对$J(\theta)$的最小值进行求解：

$J(\theta)$最小值的求解采用梯度下降法：

Step1: 随机给定一组a

Step2: 将a代入$J(\theta)$，使得得到的点沿着负梯度发方向移动，梯度的求法如下

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}J(\theta)&=-\frac{1}{n}\sum_{i=1}^{n}\left(y_i\frac{1}{h_\theta(x_i)}\frac{\partial}{\partial\theta_j}h_\theta(x_i)-(1-y_i)\frac{1}{1-h_\theta(x_i)}\frac{\partial}{\partial\theta_j}h_\theta(x_i)\right)\\
&=-\frac{1}{n}\sum_{i=1}^{n}\left(y_i\frac{1}{f(\theta' x_i)}-(1-y_i)\frac{1}{1-f(\theta' x_i)}\right)\frac{\partial}{\partial\theta_j}f(\theta' x_i)\\
&=-\frac{1}{n}\sum_{i=1}^{n}\left(y_i\frac{1}{f(\theta' x_i)}-(1-y_i)\frac{1}{1-f(\theta' x_i)}\right)f(\theta' x_i)\left(1-f(\theta' x_i)\right)\frac{\partial}{\partial\theta_j}\theta' x_i\\
&=-\frac{1}{n}\sum_{i=1}^{n}\left(y_i(1-f(\theta' x_i))-(1-y_i)f(\theta' x_i)\right)x_i^j\\
&=-\frac{1}{n}\sum_{i=1}^{n}(y_i-f(\theta' x_i))x_i^j\\
&=\frac{1}{n}\sum_{i=1}^{n}(h_\theta(x_i)-y_i)x_i^j
\end{aligned} \tag{34}$$

Step3: 循环 Step2,直到$J(\theta)$的值不再改变，此时的a就是所求的a。

5.5.4 结果分析

利用 Stata 进行 Logistic 回归可得如下结果

表 14 Logistic 模型检验结果

Number of obs	LR chi2(11)	Prob > chi2	Pseudo R2	Log likelihood
3149	37.70	0.0001	0.0168	-1049.2034

由表 14 可以看到，Prob > chi2 值为 0.0001<0.05，即在 95%的置信水平下拒绝原假设，说明该 Logistic 模型是显著的，回归结果如下表 15。

表 15 Logistic 模型结果

自变量	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
年龄	-0.0014	0.0165	-0.0900	0.9320	-0.0338	0.0310
初潮年龄	0.0149	0.0481	0.3100	0.7570	-0.0793	0.1091
月经周期	-0.0607	0.0387	-1.5700	0.1170	-0.1366	0.0153
月经经期	-0.0757	0.0676	-1.1200	0.2630	-0.2082	0.0568
使用其他节育器情况:IUD	-0.1740	0.1378	-1.2600	0.2070	-0.4442	0.0962
使用其他节育器情况:无	0.1085	0.2236	0.4900	0.6270	-0.3297	0.5468

使用其他节育器情况:其他	0.0000	(omitted)				
宫腔深度	0.0430	0.0892	0.4800	0.6300	-0.1318	0.2178
使用节育器型号情况:小	0.0906	0.1476	0.6100	0.5390	-0.1986	0.3798
使用节育器型号情况:中	**-0.4356**	0.3404	-1.2800	0.2010	-1.1027	0.2316
使用节育器型号情况:大	**-0.7986**	0.3402	-2.3500	**0.0190**	-1.4654	-0.1317
放置节育器时宫颈扩张情况	**0.2982**	0.1746	1.7100	0.0880	-0.0440	0.6404
常数	4.4016	1.7169	2.5600	0.0100	1.0365	7.7666

从影响系数的角度分析，由表 15 可以看到，使用大型节育器的影响系数最大，其值为-0.7986，其次是使用中型节育器，其值为-0.4356，再者是放置节育器时宫颈扩张情况，其值为 0.2982。并且中、大型节育器的影响系数为负数，与节育器质量呈现负相关，放置节育器时宫颈扩张情况的影响系数为正数，与节育器质量呈现正相关。

从显著性的角度分析，由表 15 可以看到，在 95%置信水平下显著的只有使用大型节育器的影响系数，勉强显著的是放置节育器时宫颈扩张情况。

综合影响系数和显著性的分析，可以得出如下结论:使用大型节育器是影响节育器质量的最主要的负向因素，放置节育器时宫颈扩张情况是是影响节育器质量的最主要的正向因素。

再综合 5.5.1 节的分析，最终可以得到如下结论: **影响宫内节育器质量的决定因素是节育器的理化指标，而在节育器的理化指标中，使用大型节育器是影响节育器质量的最主要的负向因素，放置节育器时宫颈扩张情况是是影响节育器质量的最主要的正向因素。**

六、 模型的验证与评价

6.1 模型验证分析

在 5.3.1 中采用了皮尔逊相关系数来计算相关性，但是发现数据的“线性”趋势并不强，下面将采用斯皮尔曼相关系数的方法来重新计算相关性，分析其结果与皮尔逊相关系数的误差。

斯皮尔曼相关系数采用秩相关，也称等级相关，来描述两个变量之间的关联程度与方向，该方法对原始变量分布不作要求，属于非参数统计方法。

其计算公式为

$$R_{XY}=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)} \tag{35}$$

式中，d_i为X_i与Y_i的等级差，而一个数的等级就是将它所在的一列数按照从小到大排序后，这个数所在的位置。

利用式(8)和式(35)可得图 8 相关系数矩阵(上三角为斯皮尔曼相关系数，下三角为皮尔逊相关系数)

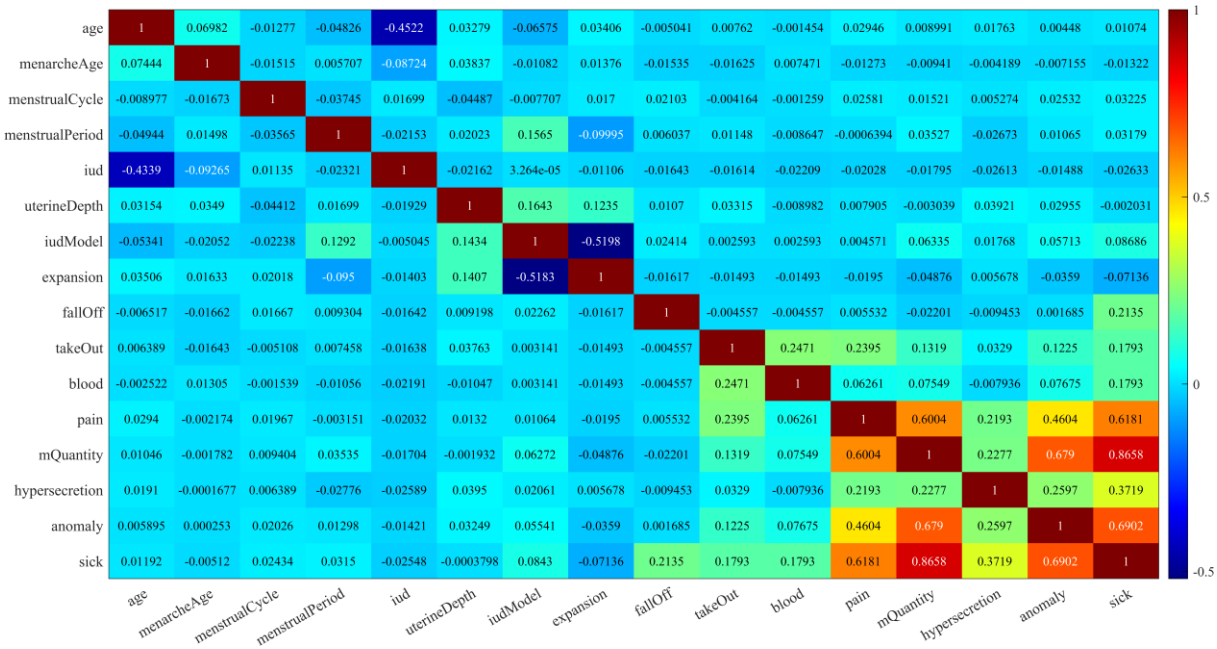


图 8 皮尔逊相关系数与斯皮尔曼相关系数

观察图 8 可以看到，皮尔逊相关系数与斯皮尔曼相关系数其结果几乎相差无几。按行展开其系数矩阵后可得如下图 9。

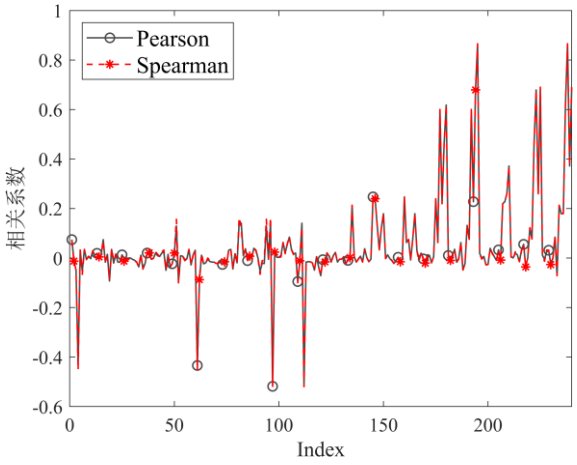


图 9 相关系数对比图

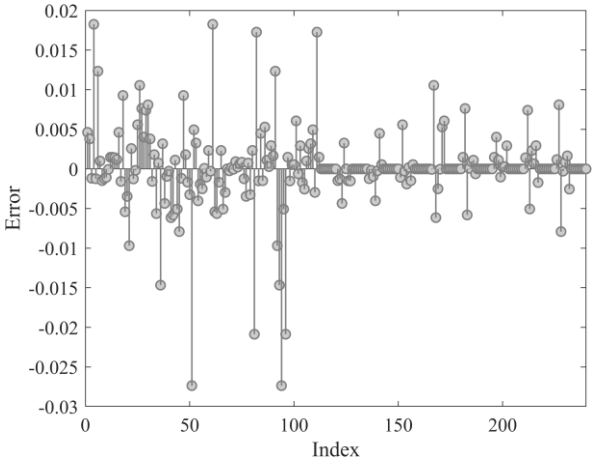


图 10 相关系数误差

由图 9 可以看到，皮尔逊相关系数与斯皮尔曼相关系数的趋势保持一致，只在某些位置出现了一点误差。图 10 给出了皮尔逊相关系数与斯皮尔曼相关系数的绝对误差线，可以看到，这两种方法计算出来的相关系数误差最大不超过 3%。

从以上分析可以说明在 5.3.1 节直接采用皮尔逊相关系数是可行的。

6.2 模型的优点

1) 本文对数据进行了预处理，减小了数据的分析难度。
2) 本文所采用的模型都是先进行数据验证，当数据满足模型要求后才对其建模，提高了模型运用合理性。

3) 本文针对每一个问题，都采用了两种方法结合进行分析，提高了模型结合创新性。
4) 本文对模型进行了验证，提高了模型的可靠性。

6.3 模型的缺点

1) 本文在对数据进行处理时，对于缺失的数据直接采用了删除处理，减低了数据样本量，可能会对模型计算结果造成一定的误差。
2) 本文所采用的模型偏于传统，未能采用目前一些主流的模型，例如机器学习、深度学习等等。

6.4 模型的推广

对于本文的模型，同样可以用在其他方法的数据分析中。例如本文的问题 1 独立样本 t 检验和卡方检验几乎可以检查任何数据的差异性；问题 2 的典型相关分析可以用于检验两组指标之间的相关性；问题 3 的 Topsis 与熵权法的结合适用于几乎所有面板数据的评价；问题 4 的 Logistic 回归模型适应于几乎所有分类数据的建模。

此外，对于问题 1 还可以采用非参数检验(如 Mann-Whitney U 检验)来说明数据之间的差异性；问题 2 的相关性也可以通过建立多变量回归方程进行描述；问题 3 的权重确定方法还可以采用层次分析法、变异系数法、随机森林等，评价模型可以采用模型综合评价；问题 4 的主要因素探究还可以采用机器学习的方法，例如 SVM 支持向量机、神经网络、决策树、KNN 最近邻分类等等。

参考文献

[1]高艺祥,杨民红,李兰会.独立样本 t 检验的 Excel 和 SPSS 分析[J].畜牧与饲料科学,2018,39(10):79-82.

[2]陈银梦,詹倩.运用双样本 t 检验的若干误区与正确条件[J].统计与管理,2019(2):40-42.

[3]房祥忠.卡方分布与卡方检验[J].中国统计,2022(5):29-31.

[4]赵晨飞. 基于谱特征分析和卡方检验的特征选择方法研究[D].天津师范大学,2019.

[5]程娟娟.高校科研与教学关系实证研究——基于皮尔逊相关系数的分析[J].中国高校科技,2022(10):46-52.DOI:10.16209/j.cnki.cust.2022.10.016.

[6]王晓燕,李美洲.浅谈等级相关系数与斯皮尔曼等级相关系数[J].广东轻工职业技术学院学报,2006(04):26-27.

[7]张伟锋. 斯皮尔曼简捷相关系数与基尼伽玛相关系数的统计特性分析[D].广东工业大学,2020.DOI:10.27029/d.cnki.ggdgu.2020.001333.

[8]魏赟.兰州市农村居民人均可支配收入的典型相关分析[J].湖北农业科学,2021,60(13):156-161+169.DOI:10.14088/j.cnki.issn0439-8114.2021.13.032.

[9]高允锁,王小丹,郭敏,蒋湘玲,高芳,黄涛,符林梅.3295 名旅游岛居民心理健康与总体幸福感典型相关分析[J].中国卫生统计,2016,33(05):865-867.

[10]刘晓君,鲁晶涵.基于 topsis-熵权法的绿色施工节地技术综合效益评价研究[J].数学的实践与认识,2023,53(01):1-10.

[11]黄莲琴,刘明玥,梁晨.基于熵权 TOPSIS 法的公司绿色治理观测指标与评价研究[J].电子科技大学学报(社科版),2023,25(02):95-106.DOI:10.14071/j.1008-8105(2023)-1003.

[12]谭英,王闯.基于逻辑回归分类算法的大学生就业去向模型研究[J].创新创业理论研究与实践,2023,6(03):6-12.

[13]田崇文.基于逻辑回归算法的债券市场实质违约因素分析与实质违约风险预警[J].营销界,2021(26):41-45.

[14]周志华.机器学习 = Machine learning[M].清华大学出版社,2016.

附录

附录1

介绍： 数据处理+正态分布检验+相关性分析 MATLAB 代码

```
clc;clear;close all
set(0,'defaultfigurecolor','w');

%% loaddata
Type11 = '附件 1';
Path11 = '.\Data\附件 1：两个院临床受试者和节育器的基本数据.xlsx';
Sheet11 = "一院临床受试者和节育器的基本数据";
Data11 = loadData(Path11, Sheet11, Type11);

Type12 = '附件 1';
Path12 = '.\Data\附件 1：两个院临床受试者和节育器的基本数据.xlsx';
Sheet12 = "二院临床受试者和节育器的基本数据";
Data12 = loadData(Path12,Sheet12,Type12);

Type21 = '附件 2';
Path21 = '.\Data\附件 2：两个医院随访的节育器使用后主诉情况.xlsx';
Sheet21 = "一院随访的节育器使用后主诉情况";
Data21 = loadData(Path21,Sheet21,Type21);

Type22 = '附件 2';
Path22 = '.\Data\附件 2：两个医院随访的节育器使用后主诉情况.xlsx';
Sheet22 = "二院随访的节育器使用后主诉情况";
Data22 = loadData(Path22,Sheet22,Type22);

%% 预处理
Data11f = fillmissing(Data11,"constant",0);
Data12f = fillmissing(Data12,"constant",0);
Data21f = fillmissing(Data21,"constant",0);
Data22f = fillmissing(Data22,"constant",0);
Data22f(233:426,:) = [];
loss1 = sum(Data21f.loss1) + sum(Data22f.loss1);
loss3 = sum(Data21f.loss3) + sum(Data22f.loss3);
loss6 = sum(Data21f.loss6) + sum(Data22f.loss6);
loss12 = sum(Data21f.loss1) + sum(Data22f.loss12);
n_data = size(Data21f,1) + size(Data22f,1);
alpha1 = loss1 / n_data;
alpha3 = loss3 / n_data;
alpha6 = loss6 / n_data;
alpha12 = loss12 / n_data;
fprintf('alpha1=%.4f%%,\nalpha3=%.4f%%,\nalpha6=%.4f%%,\nalpha12=%.4f%%\
n',alpha1*100,alpha3*100,alpha6*100,alpha12*100)


%% 合并分组
Data11f.category = ones(size(Data11f,1),1);
Data12f.category = ones(size(Data12f,1),1)*2;
Data1All = [Data11f;Data12f];
Data21f.category = ones(size(Data21f,1),1);
Data22f.category = ones(size(Data22f,1),1)*2;
Data2All = [Data21f;Data22f(2:end,:)];
% DataAll = [Data1All,Data2All];
Data2AllTemp = Data2All;
```

```
Data2AllTemp = Data2AllTemp(2:end,7:42);
%% 再处理
outputNames2 =
{'num','group','fallOff','takeOut','pregnant','blood','pain',...
    'mQuantity','hypersecretion' ,'anomaly', 'sick','category'};
outputData2All = table('Size',[size(Data2All,1)-
1,length(outputNames2)],'VariableTypes',{'double','double','double', ...

'double','double','double','double','double','double','double','double',
'double'},'VariableNames', outputNames2);
outputData2All.num = Data2All.num(2:end);
outputData2All.group = Data2All.group(2:end);
for i = 1:size(Data2AllTemp,2)/4
    DataTemp = Data2AllTemp(:,i*4-3:i*4);
    sum12 = sum(table2array(DataTemp(:,1:2)),2);
    sum34 = sum(table2array(DataTemp(:,3:4)),2);
    idx12 = find(sum12 > 0);
    idx34 = find(sum34 > 0);
    outputData2All.(outputNames2{i+2})(idx34) = 1;
end
outputData2All.category = Data2All.category(2:end);
outputNames1 =
{'num','group','age','menarcheAge','menstrualCycle','menstrualPeriod','i
ud',...
    'uterineDepth','iudModel' ,'expansion', 'category'};
outputData1All =
table('Size',[size(Data1All,1),11],'VariableTypes',{'double','double','d
ouble', ...

'double','double','double','double','double','double','double','double'}
,'VariableNames', outputNames1);
outputData1All.num = Data1All.num;
outputData1All.group = Data1All.group;
outputData1All.age = Data1All.age;
outputData1All.menarcheAge = Data1All.menarcheAge;
outputData1All.menstrualCycle = Data1All.menstrualCycle;
outputData1All.menstrualPeriod = Data1All.menstrualPeriod;
outputData1All.uterineDepth = Data1All.uterineDepth;
outputData1All.expansion = Data1All.expansion;
outputData1All.category = Data1All.category;

% Data1AllTemp = Data1All;
idx1 = find(Data1All.iud == 1);
idx2 = find(Data1All.iudNone == 1);
idx3 = find(Data1All.iudOther == 1);
outputData1All.iud(idx1) = 1;
outputData1All.iud(idx2) = 2;
outputData1All.iud(idx3) = 3;

idx1 = find(Data1All.iudModelS == 1);
idx2 = find(Data1All.iudModelM == 1);
idx3 = find(Data1All.iudModelB == 1);
outputData1All.iudModel(idx1) = 1;
outputData1All.iudModel(idx2) = 2;
outputData1All.iudModel(idx3) = 3;

% 补充主诉情况的数据
outputData2AllNew =
table('Size',[size(Data1All,1),length(outputNames2)],'VariableTypes',{'d
ouble','double','double', ...
```

```
'double','double','double','double','double','double','double','double',
'double'},'VariableNames', outputNames2);
outputData2AllNew.num = outputData1All.num;
outputData2AllNew.group = outputData1All.group;
outputData2AllNew.category = outputData1All.category;
% [UU,idxUU1] = intersect(outputData2AllNew.num,outputData2All.num);
% idxNew = find(outputData2AllNew.num == outputData2All.num);
k = 1;
% aa = [];
for i = 1:size(outputData2AllNew,1)
    for j = 1:size(outputData2All,1)
        if outputData2AllNew.num(i) == outputData2All.num(j) &&
outputData2AllNew.group(i) == outputData2All.group(j)...
                && outputData2AllNew.category(i) ==
outputData2All.category(j)
            outputData2AllNew(i,:) = outputData2All(j,:);
            k = k + 1;
%             aa = [aa,j];
            break
        end
    end
end
outputDataAll = [outputData1All,outputData2AllNew(:,3:end-1)];
outputDataAll1 = [Data1All,outputData2AllNew(:,3:end-1)];
outputDataAll1.sick = ~outputDataAll1.sick;
% aaa = sum(outputDataAll1{:,16:23},2);
% length(find(aaa>0))
%% 正态分布检验
Name =
{'age','menarcheAge','menstrualCycle','menstrualPeriod','uterineDepth'};
Label = {'年龄/岁','初潮年龄/岁','月经周期/天','月经经期/天','宫腔深度/cm'};
Beta = [11.5,3.3,3.8,2.3,2.0];
nName = length(Name);
for i = 1:nName
    U = unique(Data1All.(Name{i}));
    C = zeros(length(U),1);
    Xticks = cell(1,length(U));
    for j = 1:length(U)
        C(j) = length(find(Data1All.(Name{i}) == U(j)));
        Xticks{j} = num2str(U(j));
    end
    figure(i)
    b = bar(U,C);
    b.FaceColor = [0.8500 0.3250 0.0980];
    b.EdgeColor = [0.3,0.3,0.3];
    hold on
    [miu,sigma] = normfit(Data1All.(Name{i}));
    xx = min(U):(max(U)-min(U))*0.005:max(U);
    yy = normpdf(xx,miu,sigma)*max(C)*Beta(i);
    plot(xx,yy,'-','Color',[0.2 0.2 0.2],'LineWidth',1.3)

    xlabel(Label{i})
    ylabel('频数')
    set(gca,'FontName','TimesSimSun','FontSize',11) %设置坐标轴刻度字体名
称，大小
    exportgraphics(gca,['.\Fig\bar',Name{i},'.png'], 'Resolution',600)

end
%% 绘制相关系数矩阵
```

```
idxP = [3:10,12,13,15:20];
[R,P] = corrcoef(table2array(outputDataAll(:,idxP)));
Label = outputDataAll.Properties.VariableNames(idxP);
scrsz = get(0,'ScreenSize'); %%%% 获取屏幕的尺寸
figure('Name','hotR','Position',[0 30 scrsz(3) scrsz(4)-95]);
hot_figure = heatmap(Label,Label,R);
hot_figure.GridVisible = 'off';
colormap(gca,'jet')
set(gca,'FontName','Times New Roman','FontSize',12.5) %设置坐标轴刻度字体名
称，大小
exportgraphics(gca,'./Fig/Pearson.png', 'Resolution',600)

R1 = corr(table2array(outputDataAll(:,idxP)),'type' , 'Spearman');
figure('Name','hotR_P','Position',[0 30 scrsz(3) scrsz(4)-95]);
hot_figure = heatmap(Label,Label,R1);
hot_figure.GridVisible = 'off';
colormap(gca,'jet')
set(gca,'FontName','Times New Roman','FontSize',12.5) %设置坐标轴刻度字体名
称，大小
exportgraphics(gca,'./Fig/Pilsman.png', 'Resolution',600)

Rl = tril(R);
Ru = triu(R1);
RR = Rl + Ru;
for i = 1:length(RR)
    RR(i,i) = 1;
end

f = figure('Name','hotRR','Position',[0 30 scrsz(3) scrsz(4)-95]);
hot_figure = heatmap(Label,Label,RR);
hot_figure.GridVisible = 'off';
colormap(gca,'jet')
set(gca,'FontName','Times New Roman','FontSize',12.5) %设置坐标轴刻度字体名
称，大小
exportgraphics(gca,'./Fig/PP.png', 'Resolution',600)

idxR = find(R ~= 1);
Rexpension = R(idxR);

idxR1 = find(R1 ~= 1);
Rexpension1 = R1(idxR1);

E = Rexpension - Rexpension1;
figure('Name','Rline')
p1 = plot(Rexpension,'-
ko','MarkerIndices',1:round(length(Rexpension)*0.05):length(Rexpension),
'LineWidth',1.0);
p1.Color = [0.3,0.3,0.3];
hold on
p2 = plot(Rexpension1,'--
r*','MarkerIndices',2:round(length(Rexpension)*0.05):length(Rexpension1)
,'LineWidth',1.0);
set(gca,'FontName','Times New Roman','FontSize',12.5) %设置坐标轴刻度字体名
称，大小
legend('Pearson','Spearman','Location','northwest','fontSize',15)
xlabel('Index')
ylabel('相关系数','FontName','宋体')
xlim([0,length(Rexpension)])
```

```
exportgraphics(gca,'./Fig/相关系数对比.png', 'Resolution',600)

figure('Name','error')
% plot(E)
s = stem(E);
s.Color = [0.5,0.5,0.5];
s.LineWidth = 1.0;
s.MarkerFaceColor = [0.8,0.8,0.8];
set(gca,'FontName','Times New Roman','FontSize',12.5) %设置坐标轴刻度字体名称，大小
xlim([0,length(E)])
xlabel('Index')
ylabel('Error')
exportgraphics(gca,'./Fig/相关系数误差.png', 'Resolution',600)

%% 矩阵散点图
f = figure('Name','figMatrix','Position',[0 30 scrsz(3) scrsz(4)-95]);
[Sf,AX,BigAx,H,HAx] = plotmatrix(table2array(outputDataAll(:,idxP)));
for i = 1:length(idxP)
    H(i).FaceColor = [0.3,0.3,0.3];
    H(i).EdgeColor = [0.1,0.1,0.1];
end
for i = 1:length(idxP)
    for  j = 1:length(idxP)
        Sf(i,j).MarkerSize = 2;
        Sf(i,j).Color = 'k';
        Sf(i,j).Marker = 'o';
        Sf(i,j).MarkerFaceColor = 'k';
        set(AX(i,j),'FontName','Times New Roman','FontSize',9) %设置坐标轴刻度字体名称，大小
    end
end
exportgraphics(f,'./Fig/散点矩阵.jpg', 'Resolution',600)


%% % 保存数据
writetable(Data11f,'.\Data\附件 1：两个院临床受试者和节育器的基本数据改.xlsx','Sheet','一院临床受试者和节育器的基本数据','WriteMode','overwritesheet');
writetable(Data12f,'.\Data\附件 1：两个院临床受试者和节育器的基本数据改.xlsx','Sheet','二院临床受试者和节育器的基本数据','WriteMode','overwritesheet');
writetable(Data21f,'.\Data\附件 2：两个医院随访的节育器使用后主诉情况改.xlsx','Sheet','一院随访的节育器使用后主诉情况','WriteMode','overwritesheet');
writetable(Data22f,'.\Data\附件 2：两个医院随访的节育器使用后主诉情况改.xlsx','Sheet','二院随访的节育器使用后主诉情况','WriteMode','overwritesheet');
writetable(Data1All,'.\Data\附件 1 汇总.xlsx','Sheet','Sheet1','WriteMode','overwritesheet');
writetable(outputData1All,'.\Data\附件 1 汇总.xlsx','Sheet','Sheet1','WriteMode','overwritesheet');
writetable(outputData2All,'.\Data\附件 2 汇总.xlsx','Sheet','Sheet1','WriteMode','overwritesheet');
writetable(outputDataAll,'.\Data\附件 12 汇总.xlsx','Sheet','Sheet1','WriteMode','overwritesheet');
writetable(outputDataAll1,'.\Data\附件 12 汇总无合
```

```
并.xlsx','Sheet','Sheet1','WriteMode','overwritesheet');
fprintf('数据已保存至.\\Data\\中!\n')
```

附录2

介绍： 导入数据子函数 MATLAB 代码

```
function Data = loadData(Path,Sheet,type)
switch type
    case '附件 1'
        opts = spreadsheetImportOptions("NumVariables", 14);
        % 指定工作表和范围
    %     opts.Sheet = "一院临床受试者和节育器的基本数据";
        opts.Sheet = Sheet;
%         opts.DataRange = "A3:N1577";
        % 指定列名称和类型
        opts.VariableNames = ["num", "group", "age", "menarcheAge",
"menstrualCycle",...
            "menstrualPeriod", "iud", "iudNone", "iudOther",
"uterineDepth", "iudModelS",...
            "iudModelM", "iudModelB", "expansion"];
        opts.VariableTypes = ["double", "double", "double", "double",
"double", "double",...
            "double", "double", "double", "double", "double", "double",
"double", "double"];

        % 导入数据
        Data = readtable(Path, opts, "UseExcel", false);
        Data(1:2,:) = [];


    case '附件 2'
        opts = spreadsheetImportOptions("NumVariables", 42);
%         opts.Sheet = "一院随访的节育器使用后主诉情况";
        opts.Sheet = Sheet;
%         opts.DataRange = "A4:AP472";

        opts.VariableNames = ["num", "group", "loss1", "loss3", "loss6",
"loss12",...
            "fallOff1", "fallOff3", "fallOff6", "fallOff12", "takeOut1",
"takeOut3",...
            "takeOut6", "takeOut12", "pregnant1", "pregnant3",
"pregnant6", "pregnant12",...
            "blood1", "blood3", "blood6", "blood12", "pain1", "pain3",
"pain6", "pain12",...
            "mQuantity1", "mQuantity3", "mQuantity6", "mQuantity12",
"hypersecretion1",...
            "hypersecretion3", "hypersecretion6", "hypersecretion12",
"anomaly1", "anomaly3",...
            "anomaly6", "anomaly12", "sick1", "sick3", "sick6",
"sick12"];
        opts.VariableTypes = ["double", "double", "double", "double",
"double", "double",...
            "double", "double", "double", "double", "double", "double",
"double", "double", ...
            "double", "double", "double", "double", "double", "double",
"double", "double", ...
```

```
            "double", "double", "double", "double", "double", "double",
"double", "double", ...
            "double", "double", "double", "double", "double", "double",
"double", "double", ...
            "double", "double", "double", "double"];

        % 导入数据
        Data = readtable(Path, opts, "UseExcel", false);
        Data(1:2,:) = [];

end



%% 清除临时变量
clear opts
    end
```

附录3

介绍：熵权法计算权重子函数 MATLAB 代码

```
function [w,Z] = ShangQuan(X)
    % 标准化，消除量纲影响
    [m,n] = size(X);
    Z = zeros(m,n);
    X_max = max(X);
    X_min = min(X);
    for j = 1:n
        for i = 1:m
            Z(i,j) = (X(i,j) - X_min(j))/(X_max(j) - X_min(j));%Z 为标准化
后的矩阵
        end
    end
    % 熵权法计算权重
    d = zeros(1,n);
    ln_p = zeros(m,n);
    for j = 1:n
        z = Z(:,j);
        p = z / sum(z);
        for i = 1 : m
            if p(i)>0
                ln_p(i,j)=log(p(i));
            end
        end
        e = -sum(p.*ln_p(:,j)) / log(m);
        d(j) = 1 - e;
    end
    w = d./sum(d);
   end
```

附录4

介绍：绘制样本评分对比图子函数 MATLAB 代码

```
function FigTopsis(Name,s1_260,s1_380)
    figure('Name',Name)
    xrange = 1:length(s1_380);
    s1Mean260 = mean(s1_260);
    s1Mean380 = mean(s1_380);
    plot(xrange,s1_260,'.k')
    hold on
    plot(xrange,s1_380,'.r')
    ps1_260 = plot([0,max(xrange)],[s1Mean260,s1Mean260],'-
.k','LineWidth',1.5);
    ps1_260.Color = 'g';
    ps1_380 = plot([0,max(xrange)],[s1Mean380,s1Mean380],'-
.r','LineWidth',1.5);
    ps1_380.Color = 'b';
    xlim([0,max(xrange)])
    legend('VCu260','VCu380','VCu260 average','VCu380
average','Location','southwest')
    xlabel('Sample')
    ylabel('Score')
    set(gca,'FontName','Times New Roman','FontSize',13) %设置坐标轴刻度字体
名称，大小
    exportgraphics(gca,['./Fig/',Name,'Topsis.png'], 'Resolution',600)
   end
```

附录5

介绍：Topsis 计算评分 MATLAB 代码

```
clc;clear;close all
set(0,'defaultfigurecolor','w');

%% loaddata
Type11 = '附件 1';
Path11 = '.\Data\附件 1：两个院临床受试者和节育器的基本数据.xlsx';
Sheet11 = "一院临床受试者和节育器的基本数据";
Data11 = loadData(Path11, Sheet11, Type11);

Type12 = '附件 1';
Path12 = '.\Data\附件 1：两个院临床受试者和节育器的基本数据.xlsx';
Sheet12 = "二院临床受试者和节育器的基本数据";
Data12 = loadData(Path12,Sheet12,Type12);

Type21 = '附件 2';
Path21 = '.\Data\附件 2：两个医院随访的节育器使用后主诉情况.xlsx';
Sheet21 = "一院随访的节育器使用后主诉情况";
Data21 = loadData(Path21,Sheet21,Type21);

Type22 = '附件 2';
Path22 = '.\Data\附件 2：两个医院随访的节育器使用后主诉情况.xlsx';
Sheet22 = "二院随访的节育器使用后主诉情况";
Data22 = loadData(Path22,Sheet22,Type22);

%% 预处理
```

```
Data11f = fillmissing(Data11,"constant",0);
Data12f = fillmissing(Data12,"constant",0);
Data21f = fillmissing(Data21,"constant",0);
Data22f = fillmissing(Data22,"constant",0);
Data22f(233:426,:) = [];

%% 合并分组
Data11f.category = ones(size(Data11f,1),1);
Data12f.category = ones(size(Data12f,1),1)*2;
Data1All = [Data11f;Data12f];
Data21f.category = ones(size(Data21f,1),1);
Data22f.category = ones(size(Data22f,1),1)*2;
Data2All = [Data21f;Data22f(2:end,:)];
Data2AllTemp = Data2All;
Data2AllTemp = Data2AllTemp(2:end,7:42);

%% 再处理
outputNames2 =
{'num','group','fallOff','takeOut','pregnant','blood','pain',...
    'mQuantity','hypersecretion' ,'anomaly', 'sick','category'};
outputData2All = table('Size',[size(Data2All,1)-
1,length(outputNames2)],'VariableTypes',{'double','double','double', ...

'double','double','double','double','double','double','double','double',
'double'},'VariableNames', outputNames2);
outputData2All.num = Data2All.num(2:end);
outputData2All.group = Data2All.group(2:end);
for i = 1:size(Data2AllTemp,2)/4
    DataTemp = Data2AllTemp(:,i*4-3:i*4);
    sum12 = sum(table2array(DataTemp(:,1:2)),2);
    sum34 = sum(table2array(DataTemp(:,3:4)),2);
    idx12 = find(sum12 > 0);
    idx34 = find(sum34 > 0);
    outputData2All.(outputNames2{i+2})(idx34) = 1;
end
outputData2All.category = Data2All.category(2:end);
outputNames1 =
{'num','group','age','menarcheAge','menstrualCycle','menstrualPeriod','i
ud',...
    'uterineDepth','iudModel' ,'expansion', 'category'};
outputData1All =
table('Size',[size(Data1All,1),11],'VariableTypes',{'double','double','d
ouble', ...

'double','double','double','double','double','double','double','double'}
,'VariableNames', outputNames1);
outputData1All.num = Data1All.num;
outputData1All.group = Data1All.group;
outputData1All.age = Data1All.age;
outputData1All.menarcheAge = Data1All.menarcheAge;
outputData1All.menstrualCycle = Data1All.menstrualCycle;
outputData1All.menstrualPeriod = Data1All.menstrualPeriod;
outputData1All.uterineDepth = Data1All.uterineDepth;
outputData1All.expansion = Data1All.expansion;
outputData1All.category = Data1All.category;

% Data1AllTemp = Data1All;
idx1 = find(Data1All.iud == 1);
idx2 = find(Data1All.iudNone == 1);
idx3 = find(Data1All.iudOther == 1);
```

```
outputData1All.iud(idx1) = 1;
outputData1All.iud(idx2) = 2;
outputData1All.iud(idx3) = 3;

idx1 = find(Data1All.iudModelS == 1);
idx2 = find(Data1All.iudModelM == 1);
idx3 = find(Data1All.iudModelB == 1);
outputData1All.iudModel(idx1) = 1;
outputData1All.iudModel(idx2) = 2;
outputData1All.iudModel(idx3) = 3;

% 补充主诉情况的数据
outputData2AllNew =
table('Size',[size(Data1All,1),length(outputNames2)],'VariableTypes',{'d
ouble','double','double', ...

'double','double','double','double','double','double','double','double',
'double'},'VariableNames', outputNames2);
outputData2AllNew.num = outputData1All.num;
outputData2AllNew.group = outputData1All.group;
outputData2AllNew.category = outputData1All.category;
k = 1;
% aa = [];
for i = 1:size(outputData2AllNew,1)
    for j = 1:size(outputData2All,1)
        if outputData2AllNew.num(i) == outputData2All.num(j) &&
outputData2AllNew.group(i) == outputData2All.group(j)...
                && outputData2AllNew.category(i) ==
outputData2All.category(j)
            outputData2AllNew(i,:) = outputData2All(j,:);
            k = k + 1;
%             aa = [aa,j];
            break
        end
    end
end
outputDataAll = [outputData1All,outputData2AllNew(:,3:end-1)];

%% 计算均值 or 众数
outputDataAllForward = outputDataAll;
Nmeans = 5;
Names = outputDataAll.Properties.VariableNames;
NamesMeans = {Names{[3,4,5,6,8]}};
DataMeans =
table('Size',[1,Nmeans],'VariableTypes',{'double','double','double','dou
ble','double'},'VariableNames',NamesMeans);
for i = 1:Nmeans
    DataMeans.(NamesMeans{i})(1) = mean(outputDataAll.(NamesMeans{i}));
    M = max(abs(outputDataAll.(NamesMeans{i}) -
DataMeans.(NamesMeans{i})(1)));
    outputDataAllForward.(NamesMeans{i}) = 1 -
abs(outputDataAll.(NamesMeans{i}) - DataMeans.(NamesMeans{i})(1)) / M;
end

Nmodes = 3;
NamesModes = {Names{[7,9,10]}};
DataModes =
table('Size',[1,Nmodes],'VariableTypes',{'double','double','double'},'Va
riableNames',NamesModes);
for i = 1:Nmodes
```

```
    DataModes.(NamesModes{i})(1) = mode(outputDataAll.(NamesModes{i}));
    M = max(abs(outputDataAll.(NamesModes{i}) -
DataModes.(NamesModes{i})(1)));
    outputDataAllForward.(NamesModes{i}) = 1 -
abs(outputDataAll.(NamesModes{i}) - DataModes.(NamesModes{i})(1)) / M;
end

Nmin = 7;
NamesMin = {Names{[12,13,15:19]}};
for i = 1:Nmin
    M = max(outputDataAll.(NamesMin{i}));
    outputDataAllForward.(NamesMin{i}) = M -
outputDataAll.(NamesMin{i});
end

ShangQuanidx = [3:10,12,13,15:19];
outputDataAllForward = outputDataAllForward(:,ShangQuanidx);

w = ShangQuan(table2array(outputDataAllForward));

category1 = [1:4,6];
category2 = [5,7,8];
category3 = [9:15];
w1 = w(category1);
w2 = w(category2);
w3 = w(category3);
% 计算一级指标权重
W1 = sum(w1) / sum(w);
W2 = sum(w2) / sum(w);
W3 = sum(w3) / sum(w);
% 计算二级指标权重
ww1 = w1 / sum(w1);
ww2 = w2 / sum(w2);
ww3 = w3 / sum(w3);

%% 提取出 VCu260 和 VCu380 的数据

idx11 = find(outputDataAll.group == 1);
outputDataAllForwardDel = outputDataAllForward;
outputDataAllForwardDel(idx11,:) = [];
GDel = outputDataAll.group;
GDel(idx11,:) = [];

idx260 = find(GDel == 2);
idx380 = find(GDel == 3);

outputDataAllForwardDel = table2array(outputDataAllForwardDel);
data1 = outputDataAllForwardDel(:,category1);
data2 = outputDataAllForwardDel(:,category2);
data3 = outputDataAllForwardDel(:,category3);

[~,Z1] = ShangQuan(data1);
[~,Z2] = ShangQuan(data2);
[~,Z3] = ShangQuan(data3);

m = size(outputDataAllForwardDel,1);

D_max1 = sum(((Z1 - repmat(max(Z1),m,1)) .^ 2 ) .*
repmat(ww1,m,1) ,2) .^ 0.5;
D_min1 = sum(((Z1 - repmat(min(Z1),m,1)) .^ 2 ) .*
```

```
repmat(ww1,m,1) ,2) .^ 0.5;
s1 = D_min1 ./ (D_max1+D_min1);
s1_260 = s1(idx260);
s1_380 = s1(idx380);

D_max2 = sum(((Z2 - repmat(max(Z2),m,1)) .^ 2 ) .*
repmat(ww2,m,1) ,2) .^ 0.5;
D_min2 = sum(((Z2 - repmat(min(Z2),m,1)) .^ 2 ) .*
repmat(ww2,m,1) ,2) .^ 0.5;
s2 = D_min2 ./ (D_max2+D_min2);
s2_260 = s2(idx260);
s2_380 = s2(idx380);

D_max3 = sum(((Z3 - repmat(max(Z3),m,1)) .^ 2 ) .*
repmat(ww3,m,1) ,2) .^ 0.5;
D_min3 = sum(((Z3 - repmat(min(Z3),m,1)) .^ 2 ) .*
repmat(ww3,m,1) ,2) .^ 0.5;
s3 = D_min3 ./ (D_max3+D_min3);
s3_260 = s3(idx260);
s3_380 = s3(idx380);

s = W1*s1 + W2*s2 + W3*s3;
s_260 = s(idx260);
s_380 = s(idx380);


%% fig
Name1 = 's1';
FigTopsis(Name1,s1_260,s1_380);

Name2 = 's2';
FigTopsis(Name2,s2_260,s2_380);

Name3 = 's3';
FigTopsis(Name3,s3_260,s3_380);

Name = 's';
FigTopsis(Name,s_260,s_380);
```

附录6

介绍：计算 Logistic 回归 STATA 代码

```
clear
cls
import excel ".\Data\附件 12 汇总无合并.xlsx", sheet("Sheet1") firstrow
summarize age menarcheAge menstrualCycle menstrualPeriod iud iudNone
iudOther uterineDepth iudModelS iudModelM iudModelB expansion fallOff
takeOut  blood pain mQuantity hypersecretion anomaly sick
logit sick age menarcheAge menstrualCycle menstrualPeriod iud iudNone
iudOther uterineDepth iudModelS iudModelM iudModelB expansion,r
estat clas
predict yhat
estat gof
```