**宫颈癌相关因素调研报告**

#### 袁雨

PB20151805

**摘要**

本次调研对2012年至2013年间在加拉加斯的加拉加斯大学医院接受妇科服务的患者的医疗记录，使用了SPSS与Python进行数据分析，通过观察数据概况、数据预处理与相关性分析等，得到了宫颈癌的病因分析、诊断技术分析，可以让更多人了解到宫颈癌的诱因及诊断，做好预防与治疗。并给出了预测宫颈癌的机器学习模型，可以帮助临床预防与筛查。

关键词：宫颈癌；数据分析；SPSS；Python；机器学习

**Abstract**

In this survey, SPSS and python were used to analyze the medical records of patients who received gynecological services in the University Hospital of Caracas from 2012 to 2013. Through observing the data overview, data preprocessing and correlation analysis, the cause analysis and diagnostic technology analysis of cervical cancer were obtained, which can let more people understand the inducement and diagnosis of cervical cancer, and do a good job in prevention and treatment. And a machine learning model of cervical cancer is given, which can help clinical prevention and screening.

*Keywords: Cervical cancer; Data analysis; SPSS； Python； Machine learning*

# 概述

## 研究目的

根据世卫组织的调查显示，中国育龄女性中有 41% 患有不同程度的妇科疾病，已婚妇女患病率更是高达 70%。此外，每年有 20～30 万中国女性因恶性妇科疾病导致死亡，其中卵巢癌、宫颈癌都是女性头号杀手。在如此令人心悸的数据面前，很少女性可以意识到生理健康知识的重要性。放眼全球，每年有超过 50 万的宫颈癌病例，其中有超过 25 万的患者死亡。

然而，宫颈癌可以通过人类乳头瘤病毒 (HPV) 疫苗和常规的低成本筛查项目来预防，筛查程序中最广泛使用的两种技术是传统宫颈细胞学涂片、液基细胞学和阴道镜。此外，大多数情况下，这种癌症可以通过在早期阶段切除受影响的组织来治愈。

宫颈癌的发展通常是缓慢的，并在子宫颈异常 (不典型增生) 之前。但是，如果没有早期症状的诊断，可能会导致预防的疏忽。此外，在缺乏资源的发展中国家，由于问题意识低，患者通常不太进行常规筛查。虽然改进首次就诊时的病变切除对参加筛查项目的患者有直接影响，但最脆弱的人群很少坚持治疗项目，主要原因可能是患者缺乏对这一问题的认识和对医疗过程的不适。此外，在低收入国家，这一问题可能是由于弱势群体难以获得信息和接近医疗中心。

因此，个体患者风险的计算预测在此背景下具有关键作用。识别出宫颈癌发病风险高的患者，可以提高宫颈癌筛查方案的靶向效果。本次调研通过分析宫颈癌指标与诊断，得出宫颈癌的病因分析与诊断技术分析，让更多人可以了解到宫颈癌的诱因及诊断，正确认识到妇科疾病的重要性，做好预防与治疗。并给出预测宫颈癌的模型，帮助临床预防与筛查。

## 研究方案

首先通过查找资料为原始变量添加标签，然后使用Python的pandas、numpy等库对数据进行预处理，包括对缺失值的处理以及异常值的处理等。

接下来对处理过的数据使用SPSS与Python进行特征与目标变量之间的相关性分析。使用SPSS进行交叉列联表分析、两独立样本t检验、多因素方差分析等，同时结合Python的matplotlib、seaborn等库进行可视化。

最后使用Python的机器学习库对目标变量进行预测。

## 研究依据

使用来自 UCI 机器学习存储库的数据集。该数据集侧重于宫颈癌指标或诊断的预测，是在委内瑞拉加拉加斯的加拉加斯大学医院收集的，包含 858 名患者的人口统计信息、习惯和历史医疗记录。由于隐私问题（缺失值），一些患者决定不回答一些问题。

通过分析数据集并与关于宫颈癌的科学研究结合，得到宫颈癌的主要影响病因分析与诊断技术分析，并得到预测宫颈癌的模型。

## 研究工具

本次调研使用了统计软件 IBM SPSS Statistics 26 , 并通过JupyterLab运行Python程序。

# 数据分析

## 数据概况

### **数据简介**

该数据集包含858名患者的病历，是2012年至2013年间在加拉加斯的加拉加斯大学医院接受妇科服务的患者的随机抽样。患者的年龄在13至84岁之间，平均年龄为27岁；所有的患者都拥有活跃的性行为，其中大多数(98%)至少怀孕过一次；筛查方法包括传统细胞学检查、阴道镜醋酸评估和和Schiller试验(Lugol碘溶液)；活检表示最终确诊与否；病历包括患者年龄、性生活（性伴侣数量和首次性交年龄）、怀孕次数、吸烟行为、避孕用具（激素和宫内节育器）的使用以及性传播疾病的历史记录。一些患者由于个人隐私原因没有回答所有的问题。

查找资料后为变量补充标签。

|  |  |  |
| --- | --- | --- |
| **表1 数据集简介** | | |
| 特征 | 类型 | 标签 |
| Age | int | 年龄 |
| Number of sexual partners | int | 性伴侣数量 |
| First sexual intercourse | int | 第一次性交的年龄 |
| Num of pregnancies | int | 怀孕的次数 |
| Smokes | bool | 是否吸烟 |
| Smokes (years) | float | 吸烟的年数 |
| Smokes (packs/year) | float | 每年吸烟的包数 |
| Hormonal Contraceptives | bool | 是否使用激素避孕药 |
| Hormonal Contraceptives (years) | float | 使用激素避孕药的年数 |
| IUD | bool | 是否使用宫内节育器 |
| IUD (years) | float | 使用宫内节育器的年数 |
| STDs | bool | 是否感染性传播疾病 |
| STDs (number) | int | 感染性传播疾病的数量 |
| STDs:condylomatosis | bool | 是否感染湿疣病 |
| STDs:cervical condylomatosis | bool | 是否感染宫颈湿疣病 |
| STDs:vaginal condylomatosis | bool | 是否感染阴道湿疣病 |
| STDs:vulvo-perineal condylomatosis | bool | 是否感染外阴会阴湿疣病 |
| STDs:syphilis | bool | 是否感染梅毒 |
| STDs:pelvic inflammatory disease | bool | 是否感染盆腔炎 |
| STDs:genital herpes | bool | 是否感染生殖器疱疹 |
| STDs:molluscum contagiosum | bool | 是否感染 |
| STDs:AIDS | bool | 是否感染艾滋病 |
| STDs:HIV | bool | 是否感染HIV病毒 |
| STDs:Hepatitis B | bool | 是否感染乙肝病毒 |
| STDs:HPV | bool | 是否感染HPV病毒 |
| STDs: Number of diagnosis | int | 性传播疾病的诊断次数 |
| STDs: Time since first diagnosis | int | 自首次诊断以来的年数 |
| STDs: Time since last diagnosis | int | 自上次诊断以来的年数 |
| Dx:Cancer | bool | 预测是否有癌症复发的可能 |
| Dx:CIN | bool | 预测是否有宫颈上皮内瘤变复发的可能 |
| Dx:HPV | bool | 预测是否有人乳头瘤病毒复发的可能 |
| Dx | bool | 预测是否有复发的可能 |
| Hinselmann | bool | 阴道镜筛查结果 |
| Schiller | bool | Schiller碘试验筛查结果 |
| Cytology | bool | 细胞学筛查结果 |
| Biopsy | bool | 宫颈活检 |

### **数据分布**

使用箱图观察数据的大致分布，如下图所示。

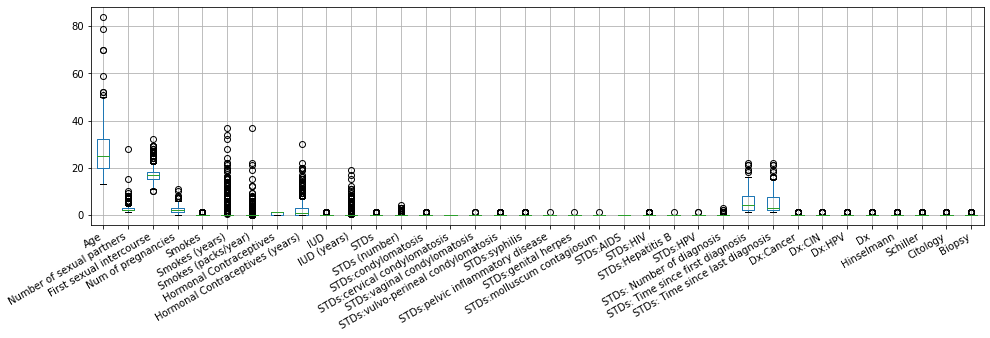


图1 数据集分布箱图

将数据导入JupterLab,将“？”替换为“NaN”。统计数据的具体分布，如下表所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表2 数据集分布** | | | | | |
| 特征 | 数量 | 均值 | 标准差 | 最小值 | 最大值 |
| Age | 858 | 26.821 | 8.498 | 13 | 84 |
| Number of sexual partners | 832 | 2.528 | 1.668 | 1 | 28 |
| First sexual intercourse | 851 | 16.995 | 2.803 | 10 | 32 |
| Num of pregnancies | 802 | 2.276 | 1.447 | 0 | 11 |
| Smokes | 845 | 0.146 | 0.353 | 0 | 1 |
| Smokes (years) | 845 | 1.220 | 4.089 | 0 | 37 |
| Smokes (packs/year) | 845 | 0.453 | 2.227 | 0 | 37 |
| Hormonal Contraceptives | 750 | 0.641 | 0.480 | 0 | 1 |
| Hormonal Contraceptives (years) | 750 | 2.256 | 3.764 | 0 | 30 |
| IUD | 741 | 0.112 | 0.316 | 0 | 1 |
| IUD (years) | 741 | 0.515 | 1.943 | 0 | 19 |
| STDs | 753 | 0.105 | 0.307 | 0 | 1 |
| STDs (number) | 753 | 0.177 | 0.562 | 0 | 4 |
| STDs:condylomatosis | 753 | 0.058 | 0.235 | 0 | 1 |
| STDs:cervical condylomatosis | 753 | 0.000 | 0.000 | 0 | 0 |
| STDs:vaginal condylomatosis | 753 | 0.005 | 0.073 | 0 | 1 |
| STDs:vulvo-perineal condylomatosis | 753 | 0.057 | 0.232 | 0 | 1 |
| STDs:syphilis | 753 | 0.024 | 0.153 | 0 | 1 |
| STDs:pelvic inflammatory disease | 753 | 0.001 | 0.036 | 0 | 1 |
| STDs:genital herpes | 753 | 0.001 | 0.153 | 0 | 1 |
| STDs:molluscum contagiosum | 753 | 0.001 | 0.036 | 0 | 1 |
| STDs:AIDS | 753 | 0.000 | 0.000 | 0 | 0 |
| STDs:HIV | 753 | 0.024 | 0.153 | 0 | 1 |
| STDs:Hepatitis B | 753 | 0.001 | 0.036 | 0 | 1 |
| STDs:HPV | 753 | 0.003 | 0.052 | 0 | 1 |
| STDs: Number of diagnosis | 858 | 0.087 | 0.303 | 0 | 1 |
| STDs: Time since first diagnosis | 71 | 6.141 | 5.895 | 1 | 22 |
| STDs: Time since last diagnosis | 71 | 5.817 | 5.755 | 1 | 22 |
| Dx:Cancer | 858 | 0.021 | 0.143 | 0 | 1 |
| Dx:CIN | 858 | 0.010 | 0.102 | 0 | 1 |
| Dx:HPV | 858 | 0.021 | 0.143 | 0 | 1 |
| Dx | 858 | 0.028 | 0.165 | 0 | 1 |
| Hinselmann | 858 | 0.041 | 0.198 | 0 | 1 |
| Schiller | 858 | 0.086 | 0.281 | 0 | 1 |
| Cytology | 858 | 0.051 | 0.221 | 0 | 1 |
| Biopsy | 858 | 0.064 | 0.245 | 0 | 1 |

## 数据预处理

### **缺失值处理**

使用python 统计每列的缺失值数量，如下表所示。

|  |  |
| --- | --- |
| **表3 数据集缺失值** | |
| 特征 | 类型 |
| Age | 0 |
| Number of sexual partners | 26 |
| First sexual intercourse | 7 |
| Num of pregnancies | 56 |
| Smokes | 13 |
| Smokes (years) | 13 |
| Smokes (packs/year) | 13 |
| Hormonal Contraceptives | 108 |
| Hormonal Contraceptives (years) | 108 |
| IUD | 117 |
| IUD (years) | 117 |
| STDs | 105 |
| STDs (number) | 105 |
| STDs:condylomatosis | 105 |
| STDs:cervical condylomatosis | 105 |
| STDs:vaginal condylomatosis | 105 |
| STDs:vulvo-perineal condylomatosis | 105 |
| STDs:syphilis | 105 |
| STDs:pelvic inflammatory disease | 105 |
| STDs:genital herpes | 105 |
| STDs:molluscum contagiosum | 105 |
| STDs:AIDS | 105 |
| STDs:HIV | 105 |
| STDs:Hepatitis B | 105 |
| STDs:HPV | 105 |
| STDs: Number of diagnosis | 0 |
| STDs: Time since first diagnosis | 787 |
| STDs: Time since last diagnosis | 787 |
| Dx:Cancer | 0 |
| Dx:CIN | 0 |
| Dx:HPV | 0 |
| Dx | 0 |
| Hinselmann | 0 |
| Schiller | 0 |
| Cytology | 0 |
| Biopsy | 0 |

统计可见，变量“STDs: Time since first diagnosis” 与“STDs: Time since last diagnosis”缺失值过多，故舍弃这两个变量。其他变量的缺失值并不多，为了确保分析的准确性，对于其他缺失值，舍弃其所在个案。

### **异常值处理**

异常值是指样本中的个别值，其数值明显偏离所属样本的绝大部分观测值。异常值会影响数据集的均值和标准差，在统计上给出错误的结果，故对异常值进行处理。

作出缺失值处理后的箱图。

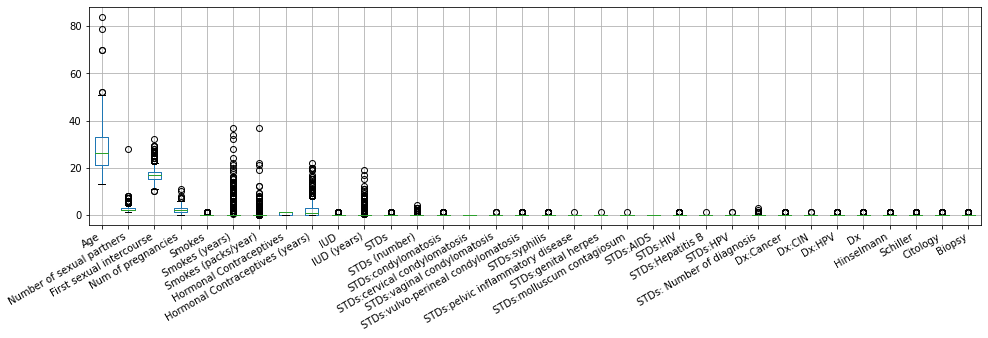


图2 缺失值处理后的数据集分布箱图

可见该数据集仍存在异常值。对数值型变量Age、Number of sexual partners、First sexual intercourse、

Num of pregnancies的异常值进行处理，其他的数值型变量因为含0值过多，异常值判断不准确，选择不做处理。使用的具体方法是IQR法，将数据三等分，Q1，Q3分别为等分点，任何小于Q1− k(Q3 − Q1) , 或大于Q3 + k(Q3− Q1)的点均定义为极值，这里取k=3。

处理后的数据分布如下。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **表4 预处理后的数据集分布** | | | | | | |
| 特征 | 数量 | 均值 | 标准差 | 最小值 | 最大值 |
| Age | 646 | 26.927 | 7.907 | 13 | 52 |
| Number of sexual partners | 646 | 2.426 | 1.181 | 1 | 6 |
| First sexual intercourse | 646 | 17.009 | 2.517 | 10 | 27 |
| Num of pregnancies | 646 | 2.310 | 1.399 | 0 | 8 |
| Smokes | 646 | 0.136 | 0.343 | 0 | 1 |
| Smokes (years) | 646 | 1.144 | 4.061 | 0 | 37 |
| Smokes (packs/year) | 646 | 0.435 | 2.342 | 0 | 37 |
| Hormonal Contraceptives | 646 | 0.649 | 0.478 | 0 | 1 |
| Hormonal Contraceptives (years) | 646 | 2.324 | 3.757 | 0 | 22 |
| IUD | 646 | 0.113 | 0.317 | 0 | 1 |
| IUD (years) | 646 | 0.543 | 2.031 | 0 | 19 |
| STDs | 646 | 0.096 | 0.295 | 0 | 1 |
| STDs (number) | 646 | 0.161 | 0.539 | 0 | 4 |
| STDs:condylomatosis | 646 | 0.053 | 0.223 | 0 | 1 |
| STDs:cervical condylomatosis | 646 | 0.000 | 0.000 | 0 | 0 |
| STDs:vaginal condylomatosis | 646 | 0.005 | 0.068 | 0 | 1 |
| STDs:vulvo-perineal condylomatosis | 646 | 0.051 | 0.220 | 0 | 1 |
| STDs:syphilis | 646 | 0.023 | 0.151 | 0 | 1 |
| STDs:pelvic inflammatory disease | 646 | 0.002 | 0.039 | 0 | 1 |
| STDs:genital herpes | 646 | 0.002 | 0.039 | 0 | 1 |
| STDs:molluscum contagiosum | 646 | 0.002 | 0.039 | 0 | 1 |
| STDs:AIDS | 646 | 0.000 | 0.000 | 0 | 0 |
| STDs:HIV | 646 | 0.020 | 0.141 | 0 | 1 |
| STDs:Hepatitis B | 646 | 0.002 | 0.039 | 0 | 1 |
| STDs:HPV | 646 | 0.003 | 0.056 | 0 | 1 |
| STDs: Number of diagnosis | 646 | 0.091 | 0.309 | 0 | 1 |
| Dx:Cancer | 646 | 0.026 | 0.160 | 0 | 1 |
| Dx:CIN | 646 | 0.005 | 0.068 | 0 | 1 |
| Dx:HPV | 646 | 0.025 | 0.156 | 0 | 1 |
| Dx | 646 | 0.025 | 0.156 | 0 | 1 |
| Hinselmann | 646 | 0.046 | 0.211 | 0 | 1 |
| Schiller | 646 | 0.094 | 0.293 | 0 | 1 |
| Cytology | 646 | 0.059 | 0.235 | 0 | 1 |
| Biopsy | 646 | 0.068 | 0.252 | 0 | 1 |

## 相关性分析

首先使用热图来直观感受各变量的相关性，数字表示相关系数，颜色深浅表示相关性大小。

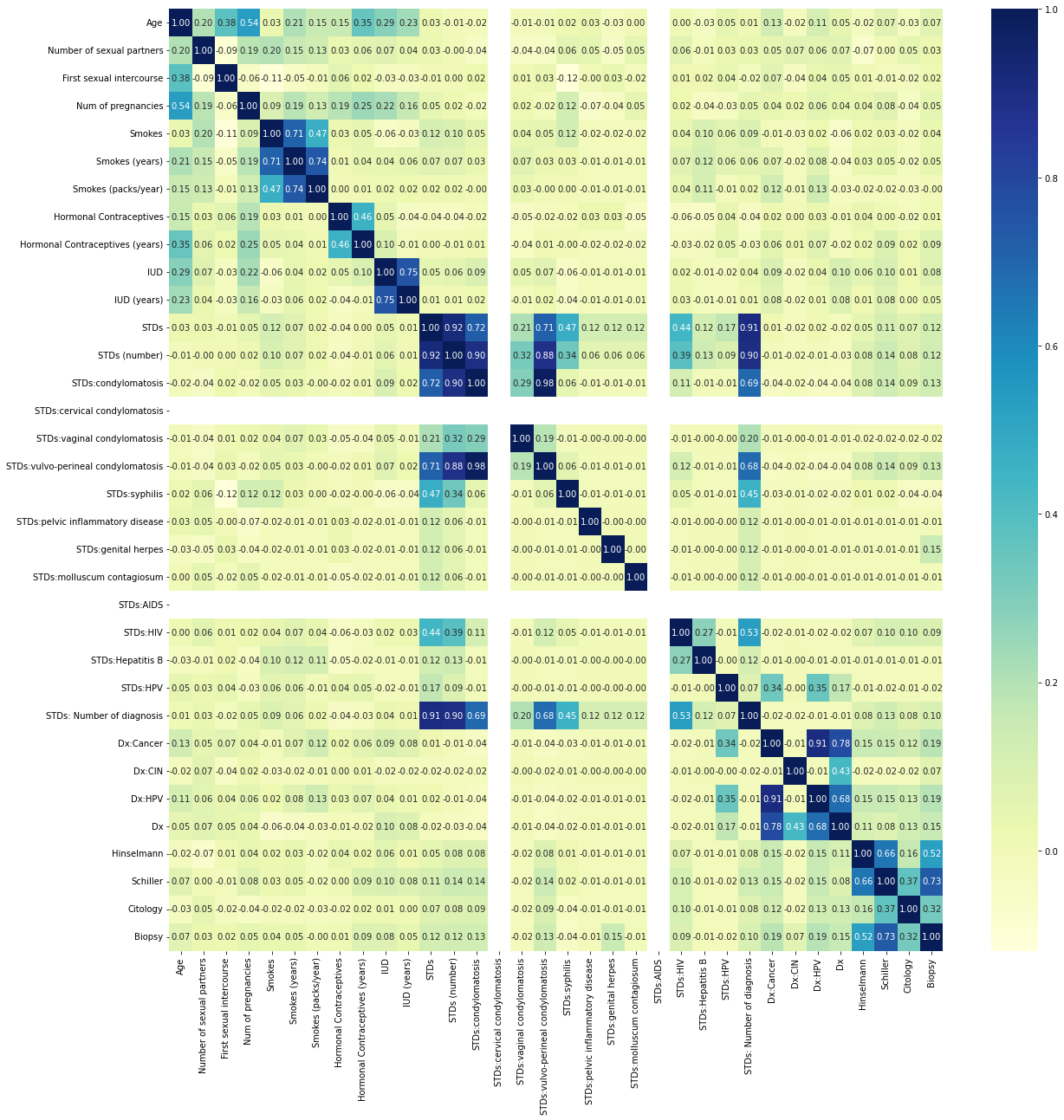


图3 数据集变量相关性热图

### **2.3.1 年龄**

绘制散点图，在目标变量值上叠加一个范围在(-0.1,0.1)之间的随机数，将点的颜色设置成透明，这样重叠的部分就会比较明显。可见活检为阳性的患者年龄分布较均匀，活检为阴性的患者整体年龄稍偏小。

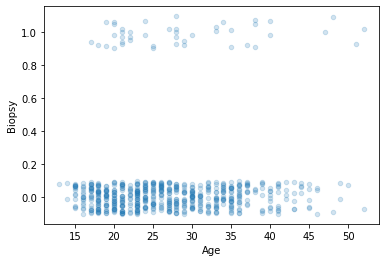


图4 Age 和 Biopsy散点图

将年龄看作近似服从正态分布的总体的随机独立样本，采用两独立样本t检验的方法来研究活检结果为阳性的患者与阴性患者的年龄的平均值是否存在显著差异。原假设是活检阳性与阴性患者的总体均值无显著差异，备择假设为有显著差异。结果如下表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表5 Age 和 Biopsy的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| Age | 0 | 602 | 26.78 | 7.782 | 0.317 |
| 1 | 44 | 29.00 | 9.306 | 1.403 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **表6 Age 和 Biopsy两独立样本t检验结果** | | | | | | | | | | |
|  | | 莱文方差等同性验 | | 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 | 差值 95% 置信区间 | |
| 下限 | 上限 |
| Age | 假定等方差 | 2.369 | 0.124 | -1.805 | 644 | 0.072 | -2.224 | 1.233 | -4.645 | 0.196 |
| 不假定等方差 |  |  | -1.546 | 47.499 | 0.129 | -2.224 | 1.438 | -5.117 | 0.669 |

活检结果为阳性的患者年龄均值为29.00，阴性患者的年龄均值为26.78，二者有一定差距。且均值标准误不大，说明这种差异的造成应该是系统性的。

首先看两总体方差是否相等的莱文F检验。该检验的F统计量的观测值为2.369，对应的p-值为0.124。如果显著性水平α为0.05，由于概率p-值大于0.05，可以认为两总体的方差没有显著差异。

然后看两总体均值差的检验。由于两总体方差没有显著差异，应看假定等方差t检验结果。其中，t统计量的观测值为-1.805，对应的双侧概率P-值为0.72，如果显著性水平α为0.05，由于概率p-值近似0.05，可以认为两总体的均值有一定差异，即活检结果为阳性的患者与阴性患者的年龄的平均值有一定差异。

### **2.3.2 性行为**

#### **Number of sexual partners**

将Number of sexual partners进行分组，离散化到[0，2，4，6]，对各个区间的Biopsy取均值。

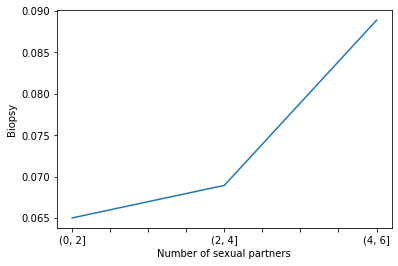


图5 Number of sexual partners 和 Biopsy 折线图

可见性伴侣数量与活检阳性的概率近似呈正相关。

**First sexual intercourse**

将First sexual intercourse进行分组，离散化到[10，15，20，25，30]，对各个区间的Biopsy取均值，得到下图。

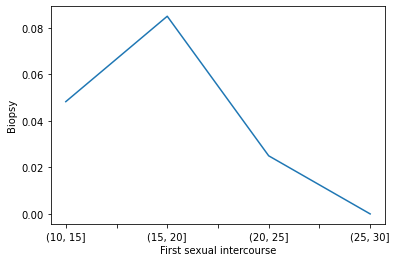


图6 First sexual intercourse 和 Biopsy 折线图

可见第一次性交的年龄近似与活检阳性的概率近似呈负相关。

#### **Num of pregnancies**

对怀孕次数根据活检阳性与否进行分组统计平均值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表7 Num of pregnancies 和 Biopsy 的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| Num of pregnancies | 0 | 602 | 2.29 | 1.401 | 0.057 |
| 1 | 44 | 2.59 | 1.352 | 0.204 |

可见活检为阴性的患者怀孕次数的平均值为2.29，活检为阳性的患者怀孕次数的平均值为2.59，后者略大于前者。

### **2.3.3 吸烟**

#### **Smokes**

利用交叉分组下的频数分析来判断吸烟与否在患宫颈癌的风险方面是否有差异。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表8 Smokes 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| Smokes | 0 | 计数 | 522 | 36 | 558 |
| 期望计数 | 520.0 | 38.0 | 558.0 |
| 占 Smokes 的百分比 | 93.5% | 6.5% | 100.0% |
| 占 Biopsy 的百分比 | 86.7% | 81.8% | 86.4% |
| 占总计的百分比 | 80.8% | 5.6% | 86.4% |
| 残差 | 2.0 | -2.0 |  |
| 1 | 计数 | 80 | 8 | 88 |
| 期望计数 | 82.0 | 6.0 | 88.0 |
| 占 Smokes 的百分比 | 90.9% | 9.1% | 100.0% |
| 占 Biopsy 的百分比 | 13.3% | 18.2% | 13.6% |
| 占总计的百分比 | 12.4% | 1.2% | 13.6% |
| 残差 | -2.0 | 2.0 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 Smokes 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

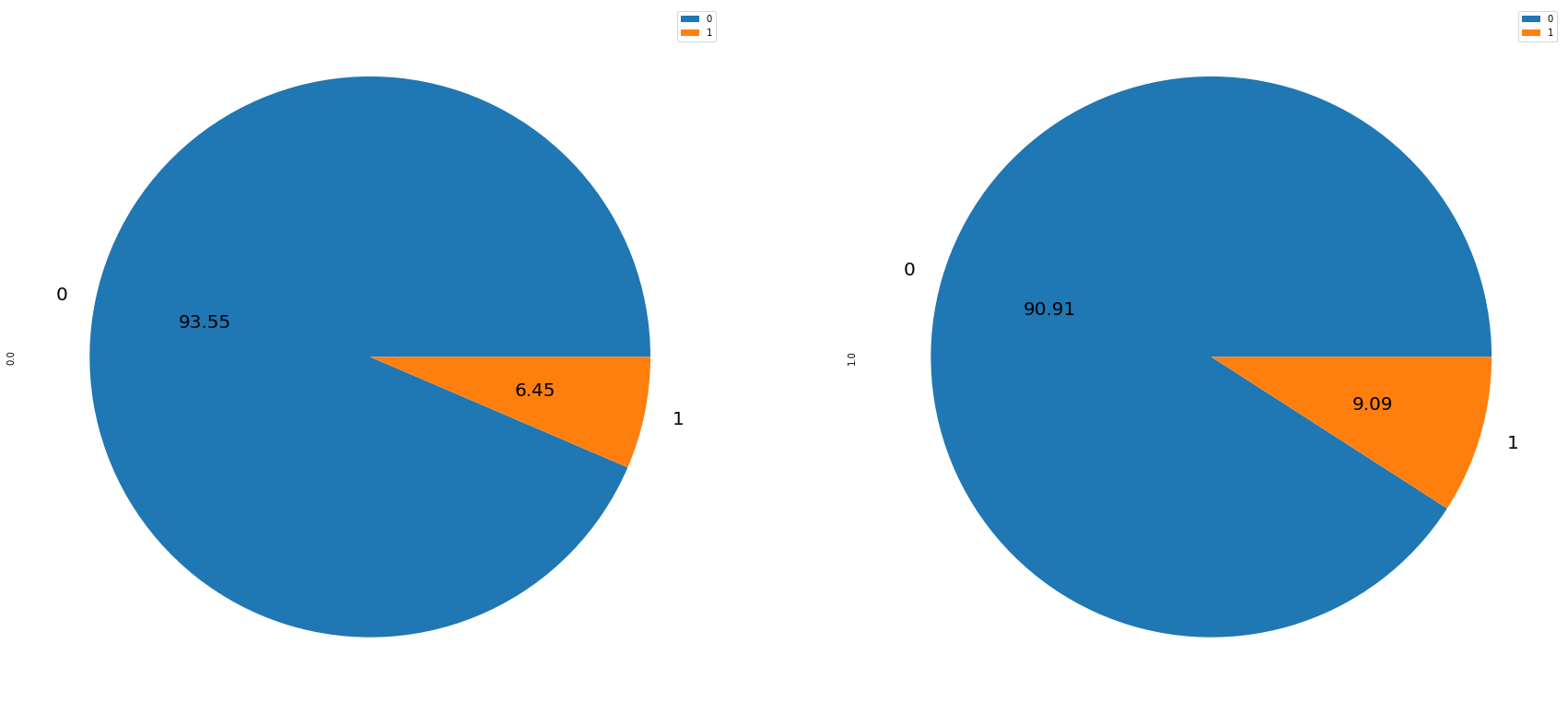


图7 Smokes 和 Biopsy 交叉饼图

由表可见，在646名患者中，不吸烟的有558人，吸烟的有88人，分别占总人数的86.4%和13.6%。宫颈癌活检为阴性的有602人，阳性的有44人，分别占总人数的93.2%和6.8%。对吸烟与否进行分析，在558个不吸烟的患者中，宫颈癌活检为阳性占6.5%；在88个吸烟的患者中，宫颈癌活检为阳性占9.1%，近似为不吸烟患者的1.5倍。可见，吸烟能一定程度下提高患宫颈癌的风险。

#### **Smokes(years)**

对吸烟年数根据活检阳性与否进行分组统计平均值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表9 Smokes (years) 和 Biopsy 的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| Smokes (years) | 0 | 602 | 1.093 | 3.897 | 0.159 |
| 1 | 44 | 1.847 | 5.878 | 0.886 |

可见活检为阴性的患者平均吸烟年数为1.093，活检为阳性的患者平均吸烟年数为1.847，后者近似为前者的两倍，可认为吸烟年数对患宫颈癌的风险有一定影响。

#### **Smokes（packs/years）**

将每年吸烟的包数看作近似服从正态分布的总体的随机独立样本，采用两独立样本t检验的方法来研究活检结果为阳性的患者与阴性患者的每年吸烟包数的平均值是否存在显著差异。原假设是活检阳性与阴性患者的总体均值无显著差异，备择假设为有显著差异。结果如下表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表10 Smokes (packs/years) 和 Biopsy 的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| Smokes (packs/year) | 0 | 602 | 0.438 | 2.406 | 0.098 |
| 1 | 44 | 0.400 | 1.154 | 0.174 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **表11 Smokes (packs/year) 和 Biopsy 的两独立样本t检验结果** | | | | | | | | | | |
|  | | 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 | 差值 95% 置信区间 | |
| 下限 | 上限 |
| Smokes (packs/year) | 假定等方差 | 0.095 | 0.758 | 0.103 | 644 | 0.918 | 0.038 | 0.366 | -0.681 | 0.756 |
| 不假定等方差 |  |  | 0.189 | 74.108 | 0.851 | 0.038 | 0.200 | -0.360 | 0.436 |

可见活检结果为阳性的患者与阴性患者的包数的平均值相近。

首先看两总体方差是否相等的莱文F检验。该检验的F统计量的观测值为0.095，对应的P-值为0.758。如果显著性水平a为0.05，由于概率P-值大于0.05，可以认为两总体的方差没有显著差异。

然后看两总体均值差的检验。由于两总体方差没有显著差异，应看假定等方差t检验结果。其中，t统计量的观测值为0.103，对应的双侧概率P-值为0.918，如果显著性水平a为0.05，由于概率P-值大于0.05，可以认为两总体的均值没有显著差异，即活检结果为阳性的患者与阴性患者的每年吸烟包数的平均值没有显著差异。

### **2.3.4 避孕**

#### **Hormonal Contraceptives**

利用交叉分组下的频数分析来判断使用激素避孕药或不使用的患者在患宫颈癌的风险方面是否有差异。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表12 Hormonal Contraceptives 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| Hormonal Contraceptives | 0 | 计数 | 212 | 15 | 227 |
| 期望计数 | 211.5 | 15.5 | 227.0 |
| 占 Hormonal Contraceptives 的百分比 | 93.4% | 6.6% | 100.0% |
| 占 Biopsy 的百分比 | 35.2% | 34.1% | 35.1% |
| 占总计的百分比 | 32.8% | 2.3% | 35.1% |
| 残差 | 0.5 | -0.5 |  |
| 1 | 计数 | 390 | 29 | 419 |
| 期望计数 | 390.5 | 28.5 | 419.0 |
| 占 Hormonal Contraceptives 的百分比 | 93.1% | 6.9% | 100.0% |
| 占 Biopsy 的百分比 | 64.8% | 65.9% | 64.9% |
| 占总计的百分比 | 60.4% | 4.5% | 64.9% |
| 残差 | -0.5 | 0.5 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 Hormonal Contraceptives 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

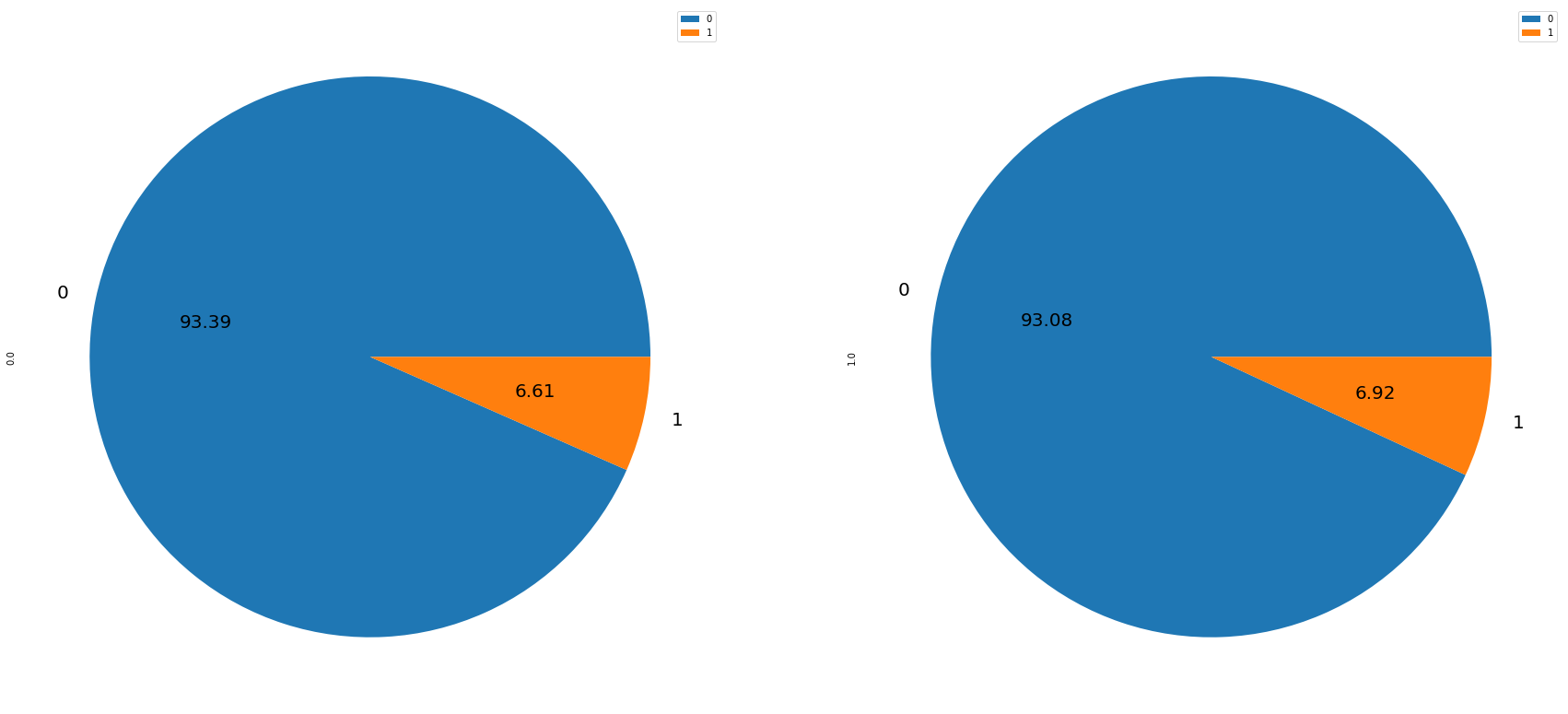


图8 Hormonal Contraceptives 和 Biopsy 交叉饼图

由表可见，在646名患者中，不使用激素避孕药的的有227人，使用的有419人，分别占总人数的35.1%和64.9%。宫颈癌活检为阴性的有602人，阳性的有44人，分别占总人数的93.2%和6.8%。对吸烟与否进行分析，在227个不使用激素避孕药的患者中，宫颈癌活检为阳性占6.6%；在88个吸烟的患者中，宫颈癌活检为阳性占6.9%，可见，使用激素避孕药对患宫颈癌的风险影响不大。

对该交叉列联表进行卡方检验。原假设是：使用激素避孕药对患宫颈癌没有影响，显著性水平设为0.05。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表13 Hormonal Contraceptives 和 Biopsy 的卡方检验结果** | | | | | |
|  | 值 | 自由度 | 渐进显著性（双侧） | 精确显著性（双侧） | 精确显著性（单侧） |
| 皮尔逊卡方 | 0.023a | 1 | 0.880 |  |  |
| 连续性修正b | 0.000 | 1 | 1.000 |  |  |
| 似然比 | 0.023 | 1 | 0.880 |  |  |
| 费希尔精确检验 |  |  |  | 1.000 | 0.511 |
| 线性关联 | 0.023 | 1 | 0.880 |  |  |
| 有效个案数 | 646 |  |  |  |  |
| a. 0 个单元格 (0.0%) 的期望计数小于 5。最小期望计数为 15.46。 | | | | | |
| b. 仅针对 2x2 表进行计算 | | | | | |

由表可得，χ2/*df =* 0.023/1 = 0.023。卡方的概率*P-*值0.880大于α，故接受原假设，认为使用激素避孕药对患宫颈癌没有影响。

#### **Hormonal Contraceptives (years)**

对使用激素避孕药的年数根据活检阳性与否进行分组统计平均值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表14 Hormonal Contraceptives (years) 和 Biopsy 的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| Hormonal Contraceptives (years) | 0 | 602 | 2.236 | 3.575 | 0.146 |
| 1 | 44 | 3.517 | 5.613 | 0.846 |

可见活检结果为阴性的患者使用激素避孕药的年数的平均值为2.236，阳性患者为3.517，阳性患者的均值是阴性患者的1.5倍。均值标准误不大，说明这种差异的造成应该是系统性的。

#### **IUD**

利用交叉分组下的频数分析方法来分析活检阳性与否的患者在使用宫内节育器的方面是否有所差异，结果如下表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表15 IUD 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| IUD | 0 | 计数 | 538 | 35 | 573 |
| 期望计数 | 534.0 | 39.0 | 573.0 |
| 占 IUD 的百分比 | 93.9% | 6.1% | 100.0% |
| 占 Biopsy 的百分比 | 89.4% | 79.5% | 88.7% |
| 占总计的百分比 | 83.3% | 5.4% | 88.7% |
| 残差 | 4.0 | -4.0 |  |
| 1 | 计数 | 64 | 9 | 73 |
| 期望计数 | 68.0 | 5.0 | 73.0 |
| 占 IUD 的百分比 | 87.7% | 12.3% | 100.0% |
| 占 Biopsy 的百分比 | 10.6% | 20.5% | 11.3% |
| 占总计的百分比 | 9.9% | 1.4% | 11.3% |
| 残差 | -4.0 | 4.0 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 IUD 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

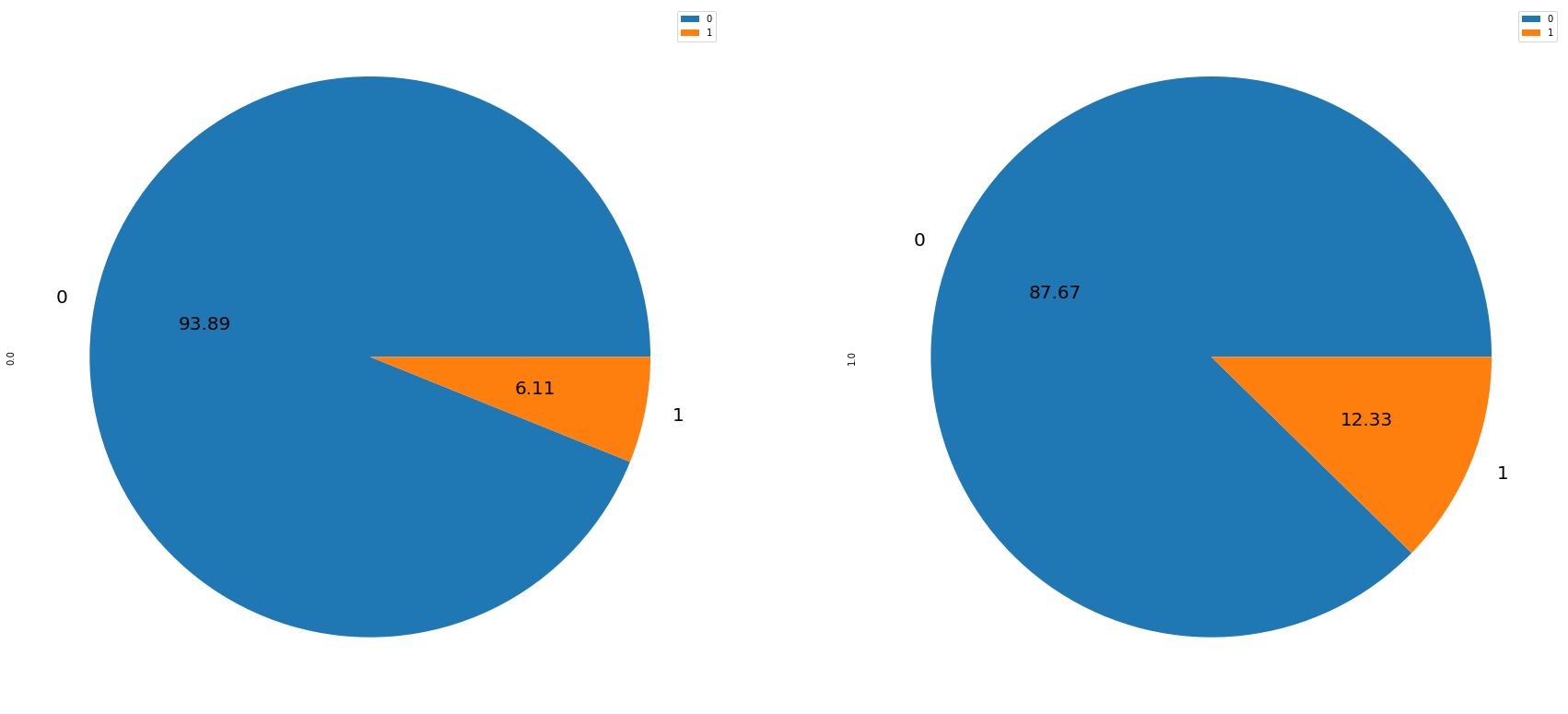


图9 IUD 和 Biopsy 交叉饼图

由表可见，在646名患者中，不使用宫内节育器的有573人，使用的有73人，分别占总人数的88.7%和11.3%。宫颈癌活检为阴性的有602人，阳性的有44人，分别占总人数的93.2%和6.8%。对使用宫内节育器与否进行分析，在573个不使用的患者中，宫颈癌活检为阳性占6.1%；在73个使用的患者中，宫颈癌活检为阳性占12.3%，是前者的两倍。

对该交叉列联表进行卡方检验。原假设是：使用宫内节育器与否对患宫颈癌没有影响，显著性水平设为0.05。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表16 IUD 和 Biopsy 的卡方检验结果** | | | | | |
|  | 值 | 自由度 | 渐进显著性（双侧） | 精确显著性（双侧） | 精确显著性（单侧） |
| 皮尔逊卡方 | 3.947a | 1 | 0.047 |  |  |
| 连续性修正b | 3.028 | 1 | 0.082 |  |  |
| 似然比 | 3.329 | 1 | 0.068 |  |  |
| 费希尔精确检验 |  |  |  | 0.078 | 0.048 |
| 线性关联 | 3.941 | 1 | 0.047 |  |  |
| 有效个案数 | 646 |  |  |  |  |
| a. 1 个单元格 (25.0%) 的期望计数小于 5。最小期望计数为 4.97。 | | | | | |
| b. 仅针对 2x2 表进行计算 | | | | | |

由表可得，χ2/*df =* 3.947/1 = 3.947。卡方的概率*P-*值0.047小于α，故可拒绝原假设，认为使用宫内节育器对患宫颈癌有显著影响。

对使用怀孕次数根据使用宫内节育器与否进行分组统计平均值。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表17 Num of pregnancies 和 IUD的基本描述统计量** | | | | | |
|  | IUD | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| Num of pregnancies | 0 | 573 | 2.20 | 1.384 | 0.058 |
| 1 | 73 | 3.16 | 1.214 | 0.142 |

可见不使用宫内节育器的患者怀孕次数的平均值为2.20，使用的为3.16，后者是前者的1.5倍。均值标准误不大，说明这种差异的造成应该是系统性的。

#### **IUD (years)**

将使用宫内节育器的年数看作近似服从正态分布的总体的随机独立样本，采用两独立样本t检验的方法来研究活检结果为阳性的患者与阴性患者的使用宫内节育器的年数的平均值是否存在显著差异。原假设是活检阳性与阴性患者的总体均值无显著差异，备择假设为有显著差异。结果如下表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表18 IUD (years) 和 Biopsy 的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| IUD (years) | 0 | 602 | 0.518 | 2.01555 | 0.082 |
| 1 | 44 | 0.886 | 2.22267 | 0.335 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **表19 IUD (years) 和 Biopsy 的两独立样本t检验结果** | | | | | | | | | | | |
|  | | 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 | 差值 95% 置信区间 | |
| 下限 | 上限 |
| IUD (years) | 假定等方差 | 2.957 | 0.086 | -1.162 | | 644 | 0.246 | -0.368 | 0.317 | -0.991 | 0.254 |
| 不假定等方差 |  |  | -1.067 | | 48.312 | 0.291 | -0.368 | 0.345 | -1.062 | 0.325 |

可见活检为阴性的患者平均使用宫内节育器年数为0.518，活检为阳性的患者平均使用宫内节育器年数为0.886，但均值标准误较大，说明这种差异应该是由抽样误差造成的。

首先看两总体方差是否相等的莱文F检验。该检验的F统计量的观测值为2.957，对应的P-值为0.086。如果显著性水平a为0.05，由于概率P-值大于0.05，可以认为两总体的方差没有显著差异。

然后看两总体均值差的检验。由于两总体方差没有显著差异，应看假定等方差t检验结果。其中，t统计量的观测值为-1.162，对应的双侧概率P-值为0.246，如果显著性水平a为0.05，由于概率P-值大于0.05，可以认为两总体的均值没有显著差异，即活检结果为阳性的患者与阴性患者的使用宫内节育器年数的平均值没有显著差异。

### **2.3.5 性传播疾病**

#### **STDs**

利用交叉分组下的频数分析方法来分析感染性传播疾病与否的患者在患宫颈癌的风险方面是否有所差异，结果如下表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表20 STDs 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| STDs | 0 | 计数 | 550 | 34 | 584 |
| 期望计数 | 544.2 | 39.8 | 584.0 |
| 占 STDs 的百分比 | 94.2% | 5.8% | 100.0% |
| 占 Biopsy 的百分比 | 91.4% | 77.3% | 90.4% |
| 占总计的百分比 | 85.1% | 5.3% | 90.4% |
| 残差 | 5.8 | -5.8 |  |
| 1 | 计数 | 52 | 10 | 62 |
| 期望计数 | 57.8 | 4.2 | 62.0 |
| 占 STDs 的百分比 | 83.9% | 16.1% | 100.0% |
| 占 Biopsy 的百分比 | 8.6% | 22.7% | 9.6% |
| 占总计的百分比 | 8.0% | 1.5% | 9.6% |
| 残差 | -5.8 | 5.8 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 STDs 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

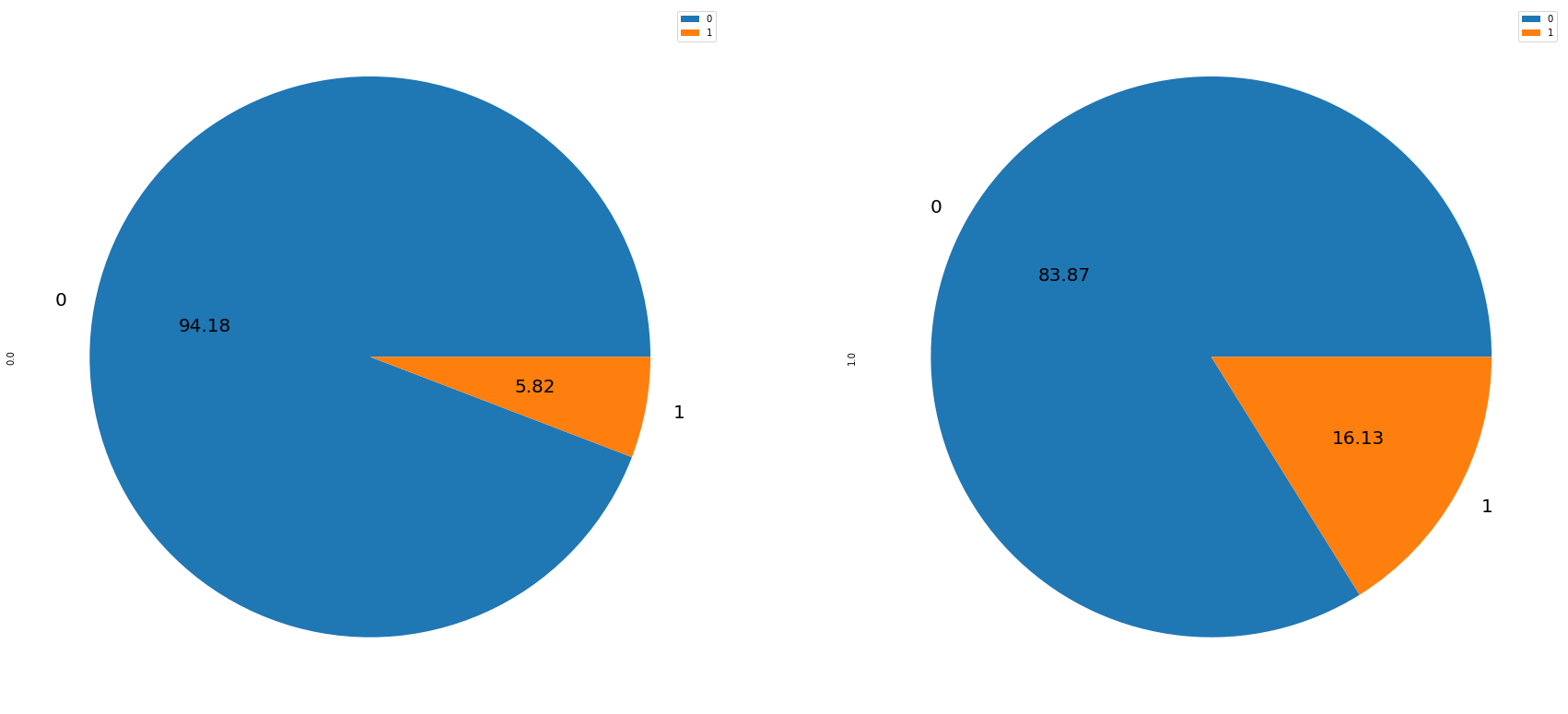


图10 STDs 和 Biopsy 交叉饼图

可见，在646名患者中，未感染性传播疾病的有584人，未感染的的有62人，分别占总人数的90.4%和9.6%。宫颈癌活检为阴性的有602人，阳性的有44人，分别占总人数的93.2%和6.8%。对感染性传播疾病与否进行分析，在584个未感染的患者中，宫颈癌活检为阳性占5.8%；在62个使用的患者中，宫颈癌活检为阳性占16.1%，是前者的3倍。可见，感染性传播疾病能提高患宫颈癌的风险。

对该交叉列联表进行卡方检验。原假设是：感染性传播疾病与否对患宫颈癌没有影响，显著性水平设为0.05。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表21 STDs 和 Biopsy 的卡方检验结果** | | | | | |
|  | 值 | 自由度 | 渐进显著性（双侧） | 精确显著性（双侧） | 精确显著性（单侧） |
| 皮尔逊卡方 | 9.381a | 1 | 0.002 |  |  |
| 连续性修正b | 7.828 | 1 | 0.005 |  |  |
| 似然比 | 7.229 | 1 | 0.007 |  |  |
| 费希尔精确检验 |  |  |  | 0.006 | 0.006 |
| 线性关联 | 9.367 | 1 | 0.002 |  |  |
| 有效个案数 | 646 |  |  |  |  |
| a. 1 个单元格 (25.0%) 的期望计数小于 5。最小期望计数为 4.22。 | | | | | |
| b. 仅针对 2x2 表进行计算 | | | | | |

由表可得，χ2/*df =* 9.381/1 = 9.381。卡方的概率*P-*值0.002小于α，故可拒绝原假设，认为感染性传播疾病对患宫颈癌有显著影响。

#### **STDs (number)**

将感染性传播疾病的数量看作近似服从正态分布的总体的随机独立样本，采用两独立样本t检验的方法来研究活检结果为阳性的患者与阴性患者的感染性传播疾病的数量的平均值是否存在显著差异。原假设是活检阳性与阴性患者的总体均值无显著差异，备择假设为有显著差异。结果如下表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表22 STDs (number) 和 Biopsy 的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| STDs (number) | 0 | 602 | 0.14 | 0.509 | 0.021 |
| 1 | 44 | 0.41 | 0.816 | 0.123 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **表23** **STDs (number) 和 Biopsy 的两独立样本t检验结果** | | | | | | | | | | |
|  | | 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 | 差值 95% 置信区间 | |
| 下限 | 上限 |
| STDs (number) | 假定等方差 | 28.935 | 0.000 | -3.186 | 644 | 0.002 | -0.266 | 0.084 | -0.430 | -0.102 |
| 不假定等方差 |  |  | -2.134 | 45.478 | 0.038 | -0.266 | 0.125 | -0.517 | -0.015 |

可见活检结果为阴性的患者感染性传播疾病的数量的平均值为0.14，阳性患者为0.41，是阴性患者的3倍。均值标准误不大，说明这种差异的造成应该是系统性的。

首先看两总体方差是否相等的莱文F检验。该检验的F统计量的观测值为28.935，对应的P-值<0.001。如果显著性水平α为0.05，由于概率P-值小于0.05，可以认为两总体的方差有显著差异。

然后看两总体均值差的检验。由于两总体方差有显著差异，应看不假定等方差t检验结果。其中，t统计量的观测值为2.134，对应的双侧概率P-值为0.038，如果显著性水平α为0.05，由于概率P-值小于0.05，可以认为两总体的均值有显著差异，即活检结果为阳性的患者与阴性患者的感染性传播疾病的数量的平均值有显著差异。

#### **STDs:condylomatosis~ STDs:HPV**

以STDs:condylomatosis~ STDs:HPV为控制变量，Biopsy为观测变量，利用多因素方差分析进行研究。其中原假设为：感染湿疣病与否没有对患宫颈癌产生显著影响（其他同理）。

|  |  |  |
| --- | --- | --- |
| **表24 主体间因子** | | |
|  | | 个案数 |
| STDs:condylomatosis | 0 | 612 |
| 1 | 34 |
| STDs:cervical condylomatosis | 0 | 646 |
| STDs:vaginal condylomatosis | 0 | 643 |
| 1 | 3 |
| STDs:vulvo-perineal condylomatosis | 0 | 613 |
| 1 | 33 |
| STDs:syphilis | 0 | 631 |
| 1 | 15 |
| STDs:pelvic inflammatory disease | 0 | 645 |
| 1 | 1 |
| STDs:genital herpes | 0 | 645 |
| 1 | 1 |
| STDs:molluscum contagiosum | 0 | 645 |
| 1 | 1 |
| STDs:AIDS | 0 | 646 |
| STDs:HIV | 0 | 633 |
| 1 | 13 |
| STDs:Hepatitis B | 0 | 645 |
| 1 | 1 |
| STDs:HPV | 0 | 644 |
| 1 | 2 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表25 Biopsy多因素方差分析结果** | | | | | |
| 因变量: Biopsy | | | | | |
| 源 | III 类平方和 | 自由度 | 均方 | F | 显著性 |
| 修正模型 | 2.118a | 10 | 0.212 | 3.459 | 0.000 |
| 截距 | .089 | 1 | 0.089 | 1.450 | 0.229 |
| STDscondylomatosis | .016 | 1 | 0.016 | 0.263 | 0.608 |
| STDscervicalcondylomatosis | .000 | 0 | . | . | . |
| STDsvaginalcondylomatosis | .088 | 1 | 0.088 | 1.432 | 0.232 |
| STDsvulvoperinealcondylomatosis | .000 | 1 | 0.000 | 0.000 | 1.000 |
| STDssyphilis | .119 | 1 | 0.119 | 1.950 | 0.163 |
| STDspelvicinflammatorydisease | .003 | 1 | 0.003 | 0.057 | 0.812 |
| STDsgenitalherpes | .884 | 1 | 0.884 | 14.436 | 0.000 |
| STDsmolluscumcontagiosum | .003 | 1 | 0.003 | 0.057 | 0.812 |
| STDsAIDS | .000 | 0 | . | . | . |
| STDsHIV | .293 | 1 | 0.293 | 4.783 | 0.029 |
| STDsHepatitisB | .044 | 1 | 0.044 | 0.716 | 0.398 |
| STDsHPV | .007 | 1 | 0.007 | 0.113 | 0.737 |
| 误差 | 38.885 | 635 | 0.061 |  |  |
| 总计 | 44.000 | 646 |  |  |  |
| 修正后总计 | 41.003 | 645 |  |  |  |
| a. R 方 = .052（调整后 R 方 = .037） | | | | | |

由表可得，观测变量Biopsy的总变差SST为41.003，它被分解为13个部分。这些变差除以各自的自由度后，得到各自的方差，并可计算出各F检验统计量的观测值和各自由度下的概率P-值。其中STDscervicalcondylomatosis、STDsAIDS由于组内样本数过少导致结果不具有统计学意义，不显示。如果显著性水平α为0.05，由于STDsgenitalherpes、STDsHIV的概率P-值<0.05，应拒绝原假设，可以认为STDsgenitalherpes、STDsHIV对Biopsy的结果存在显著差异，即感染生殖器疱疹、HIV病毒对患宫颈癌有显著影响。

#### **STDs: Number of diagnosis**

首先过滤掉STDs=0的个案。

将性传播疾病的诊断次数看作近似服从正态分布的总体的随机独立样本，采用两独立样本t检验的方法来研究活检结果为阳性的患者与阴性患者的性传播疾病的诊断次数的平均值是否存在显著差异。原假设是活检阳性与阴性患者的总体均值无显著差异，备择假设为有显著差异。结果如下表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表26 STDs: Number of diagnosis 和 Biopsy 的基本描述统计量** | | | | | |
|  | Biopsy | 个案数 | 平均值 | 标准 偏差 | 标准 误差平均值 |
| STDs: Number of diagnosis | 0 | 52 | 0.96 | 0.441 | 0.061 |
| 1 | 10 | 0.90 | 0.316 | 0.100 |

可见除去未感染的患者，活检结果为阴性的患者性传播疾病的诊断次数的平均值为0.96，阳性患者为0.90，阳性患者的均值略小于阴性患者。

### **2.3.6 疾病复发预测**

#### **Dx:Cancer**

利用交叉分组下的频数分析来判断癌症预测复发与否的患者在患宫颈癌的风险方面是否有差异。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表27 Dx:Cancer 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| Dx:Cancer | 0 | 计数 | 591 | 38 | 629 |
| 期望计数 | 586.2 | 42.8 | 629.0 |
| 占 Dx:Cancer 的百分比 | 94.0% | 6.0% | 100.0% |
| 占 Biopsy 的百分比 | 98.2% | 86.4% | 97.4% |
| 占总计的百分比 | 91.5% | 5.9% | 97.4% |
| 残差 | 4.8 | -4.8 |  |
| 1 | 计数 | 11 | 6 | 17 |
| 期望计数 | 15.8 | 1.2 | 17.0 |
| 占 Dx:Cancer 的百分比 | 64.7% | 35.3% | 100.0% |
| 占 Biopsy 的百分比 | 1.8% | 13.6% | 2.6% |
| 占总计的百分比 | 1.7% | 0.9% | 2.6% |
| 残差 | -4.8 | 4.8 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 Dx:Cancer 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

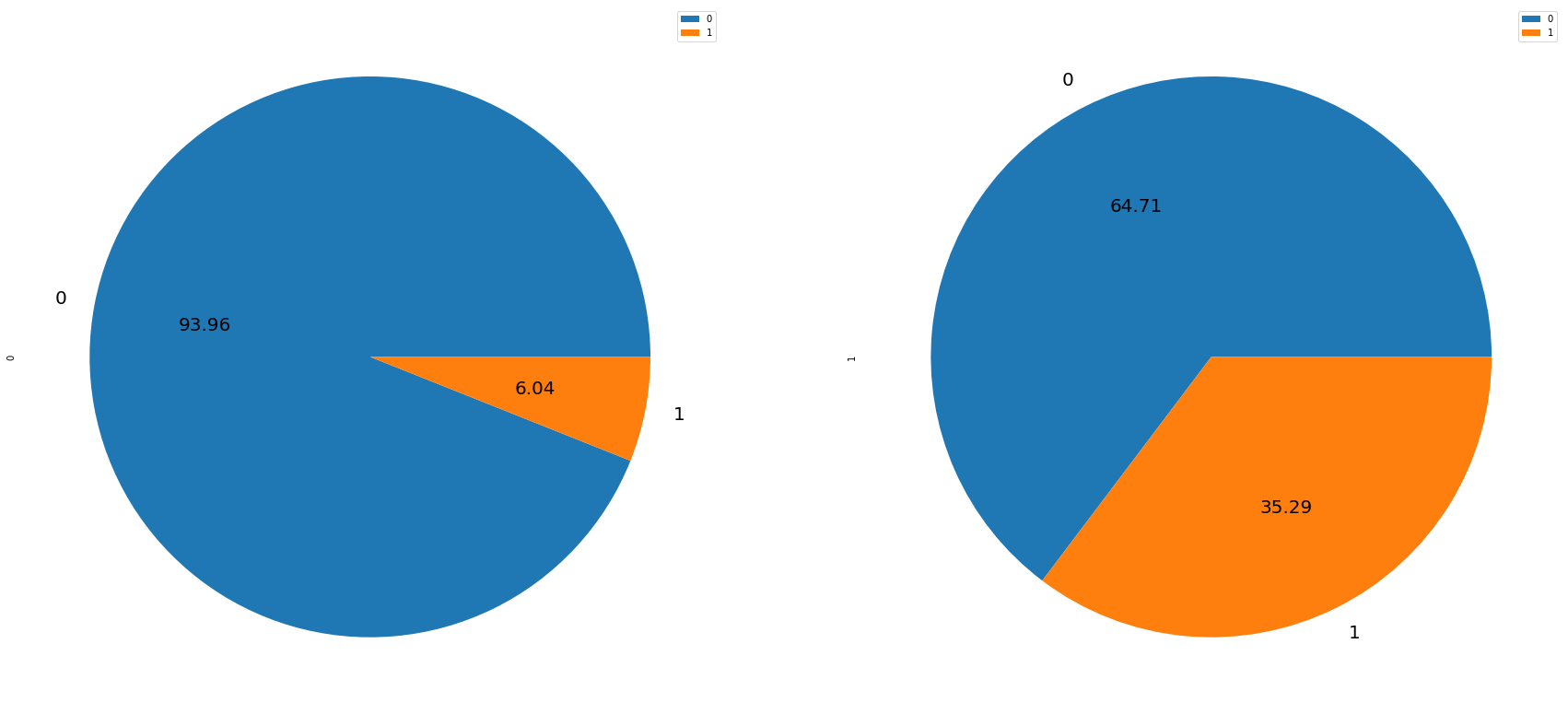


图11 Dx:Cancer 和 Biopsy 交叉饼图

由表可见，在646名患者中，癌症不会复发的有629人，会复发的有17人，分别占总人数的97.4%和2.6%。宫颈癌活检为阴性的有602人，阳性的有44人，分别占总人数的93.2%和6.8%。对癌症预测复发与否进行分析，在629个癌症不会复发的患者中，宫颈癌活检为阳性占6.0%；在17个癌症会复发的患者中，宫颈癌活检为阳性占35.3%，是前者的6倍。可见，癌症预测复发对患宫颈癌的风险影响很大。

对该交叉列联表进行卡方检验。原假设是：癌症预测复发对患宫颈癌没有影响，显著性水平设为0.05。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表28 Dx:Cancer 和 Biopsy 的卡方检验结果** | | | | | |
|  | 值 | 自由度 | 渐进显著性（双侧） | 精确显著性（双侧） | 精确显著性（单侧） |
| 皮尔逊卡方 | 22.316a | 1 | 0.000 |  |  |
| 连续性修正b | 17.945 | 1 | 0.000 |  |  |
| 似然比 | 12.326 | 1 | 0.000 |  |  |
| 费希尔精确检验 |  |  |  | 0.000 | 0.000 |
| 线性关联 | 22.282 | 1 | 0.000 |  |  |
| 有效个案数 | 646 |  |  |  |  |
| a. 1 个单元格 (25.0%) 的期望计数小于 5。最小期望计数为 1.16。 | | | | | |
| b. 仅针对 2x2 表进行计算 | | | | | |

由表可得，χ2/*df =* 22.316/1 = 22.316。卡方的概率*P-*值<0.001，故拒绝原假设，认为癌症预测复发对患宫颈癌有显著影响。

#### **Dx:CIN**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表29 Dx:CIN 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| Dx:CIN | 0 | 计数 | 600 | 43 | 643 |
| 期望计数 | 599.2 | 43.8 | 643.0 |
| 占 Dx:CIN 的百分比 | 93.3% | 6.7% | 100.0% |
| 占 Biopsy 的百分比 | 99.7% | 97.7% | 99.5% |
| 占总计的百分比 | 92.9% | 6.7% | 99.5% |
| 残差 | 0.8 | -0.8 |  |
| 1 | 计数 | 2 | 1 | 3 |
| 期望计数 | 2.8 | .2 | 3.0 |
| 占 Dx:CIN 的百分比 | 66.7% | 33.3% | 100.0% |
| 占 Biopsy 的百分比 | 0.3% | 2.3% | 0.5% |
| 占总计的百分比 | 0.3% | 0.2% | 0.5% |
| 残差 | -0.8 | 0.8 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 Dx:CIN 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

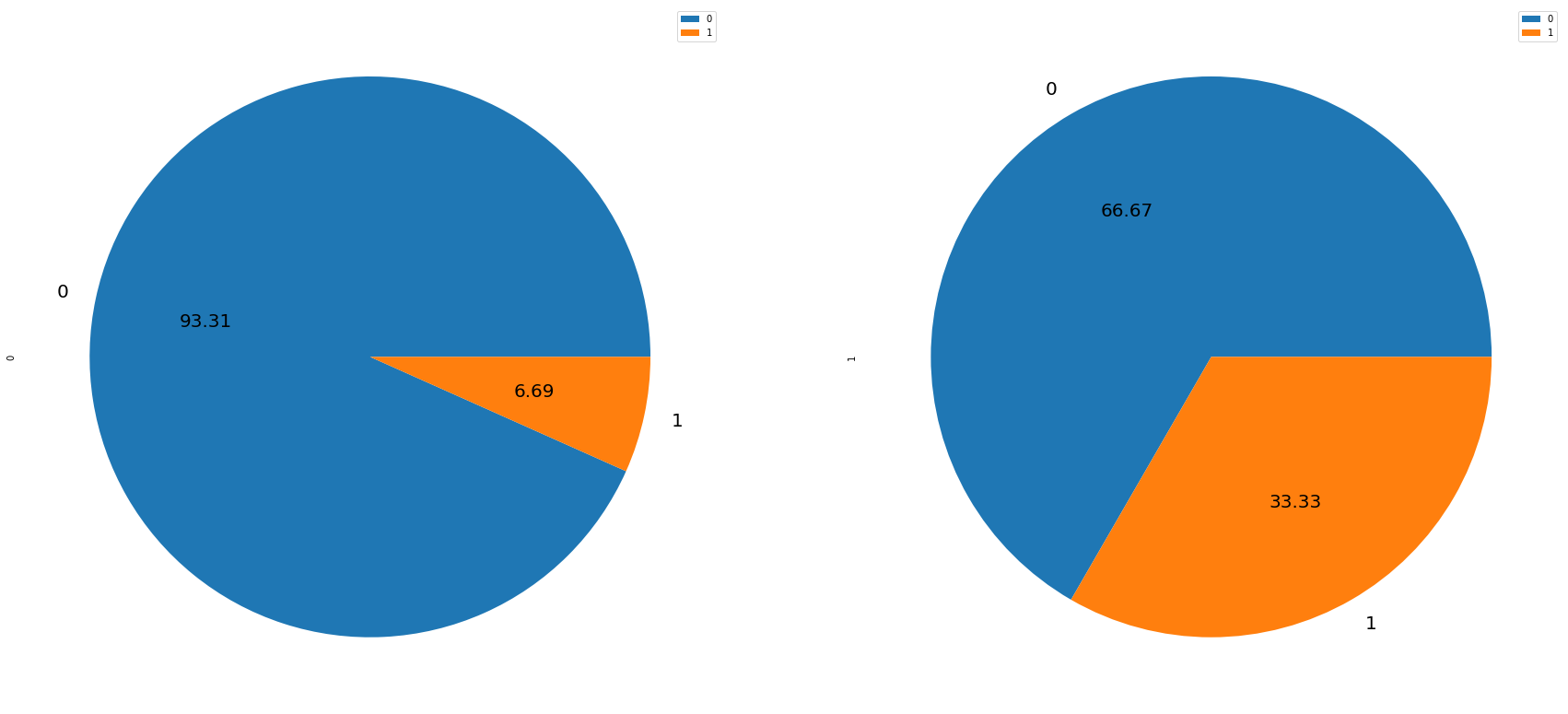


图12 Dx:CIN 和 Biopsy 交叉饼图

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表30 Dx:CIN 和 Biopsy 的卡方检验结果** | | | | | |
|  | 值 | 自由度 | 渐进显著性（双侧） | 精确显著性（双侧） | 精确显著性（单侧） |
| 皮尔逊卡方 | 3.340a | 1 | 0.068 |  |  |
| 连续性修正b | 0.461 | 1 | 0.497 |  |  |
| 似然比 | 1.852 | 1 | 0.174 |  |  |
| 费希尔精确检验 |  |  |  | 0.191 | 0.191 |
| 线性关联 | 3.335 | 1 | 0.068 |  |  |
| 有效个案数 | 646 |  |  |  |  |
| a. 2 个单元格 (50.0%) 的期望计数小于 5。最小期望计数为 .20。 | | | | | |
| b. 仅针对 2x2 表进行计算 | | | | | |

与Dx:Cancaer同理，可见宫颈上皮内瘤变会复发的患者活检为阳性的比例不会复发的患者的5倍，通过卡方检验可近似认为宫颈上皮内瘤变复发对患宫颈癌有显著影响。

#### **Dx:HPV**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表31 Dx:HPV 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| Dx:HPV | 0 | 计数 | 592 | 38 | 630 |
| 期望计数 | 587.1 | 42.9 | 630.0 |
| 占 Dx:HPV 的百分比 | 94.0% | 6.0% | 100.0% |
| 占 Biopsy 的百分比 | 98.3% | 86.4% | 97.5% |
| 占总计的百分比 | 91.6% | 5.9% | 97.5% |
| 残差 | 4.9 | -4.9 |  |
| 1 | 计数 | 10 | 6 | 16 |
| 期望计数 | 14.9 | 1.1 | 16.0 |
| 占 Dx:HPV 的百分比 | 62.5% | 37.5% | 100.0% |
| 占 Biopsy 的百分比 | 1.7% | 13.6% | 2.5% |
| 占总计的百分比 | 1.5% | 0.9% | 2.5% |
| 残差 | -4.9 | 4.9 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 Dx:HPV 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

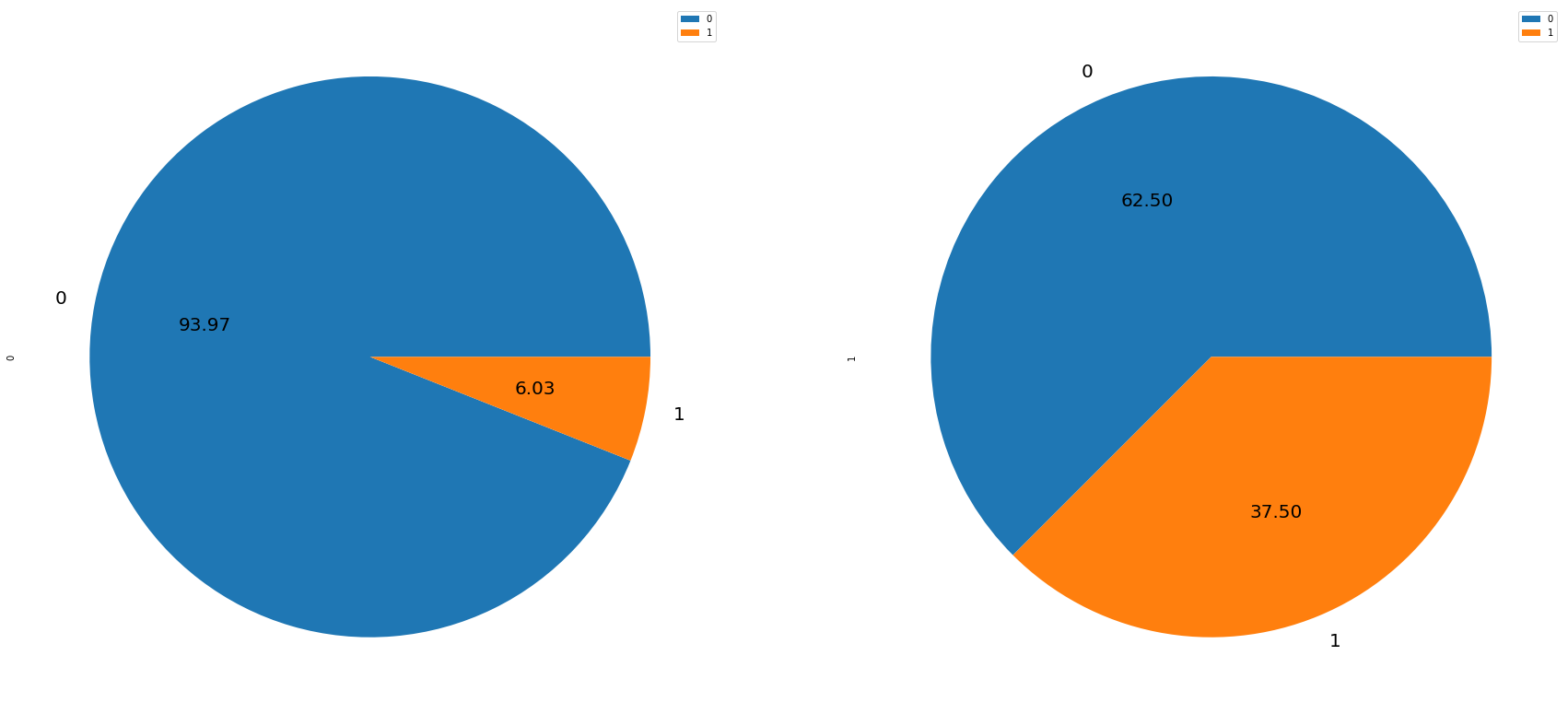


图13 Dx:HPV 和 Biopsy 交叉饼图

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表32 Dx:HPV 和 Biopsy 的卡方检验结果** | | | | | |
|  | 值 | 自由度 | 渐进显著性（双侧） | 精确显著性（双侧） | 精确显著性（单侧） |
| 皮尔逊卡方 | 24.344a | 1 | 0.000 |  |  |
| 连续性修正b | 19.638 | 1 | 0.000 |  |  |
| 似然比 | 13.106 | 1 | 0.000 |  |  |
| 费希尔精确检验 |  |  |  | 0.000 | 0.000 |
| 线性关联 | 24.306 | 1 | 0.000 |  |  |
| 有效个案数 | 646 |  |  |  |  |
| a. 1 个单元格 (25.0%) 的期望计数小于 5。最小期望计数为 1.09。 | | | | | |
| b. 仅针对 2x2 表进行计算 | | | | | |

与Dx:Cancaer同理，可见人乳头瘤病毒会复发的患者活检为阳性的比例不会复发的患者的6倍，通过卡方检验可近似认为人乳头瘤病毒复发对患宫颈癌有显著影响。

#### **Dx**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表33 Dx 和 Biopsy 的交叉列联表** | | | | | |
|  | | | Biopsy | | 总计 |
| 0 | 1 |
| Dx | 0 | 计数 | 591 | 39 | 630 |
| 期望计数 | 587.1 | 42.9 | 630.0 |
| 占 Dx 的百分比 | 93.8% | 6.2% | 100.0% |
| 占 Biopsy 的百分比 | 98.2% | 88.6% | 97.5% |
| 占总计的百分比 | 91.5% | 6.0% | 97.5% |
| 残差 | 3.9 | -3.9 |  |
| 1 | 计数 | 11 | 5 | 16 |
| 期望计数 | 14.9 | 1.1 | 16.0 |
| 占 Dx 的百分比 | 68.8% | 31.3% | 100.0% |
| 占 Biopsy 的百分比 | 1.8% | 11.4% | 2.5% |
| 占总计的百分比 | 1.7% | 0.8% | 2.5% |
| 残差 | -3.9 | 3.9 |  |
| 总计 | | 计数 | 602 | 44 | 646 |
| 期望计数 | 602.0 | 44.0 | 646.0 |
| 占 Dx 的百分比 | 93.2% | 6.8% | 100.0% |
| 占 Biopsy 的百分比 | 100.0% | 100.0% | 100.0% |
| 占总计的百分比 | 93.2% | 6.8% | 100.0% |

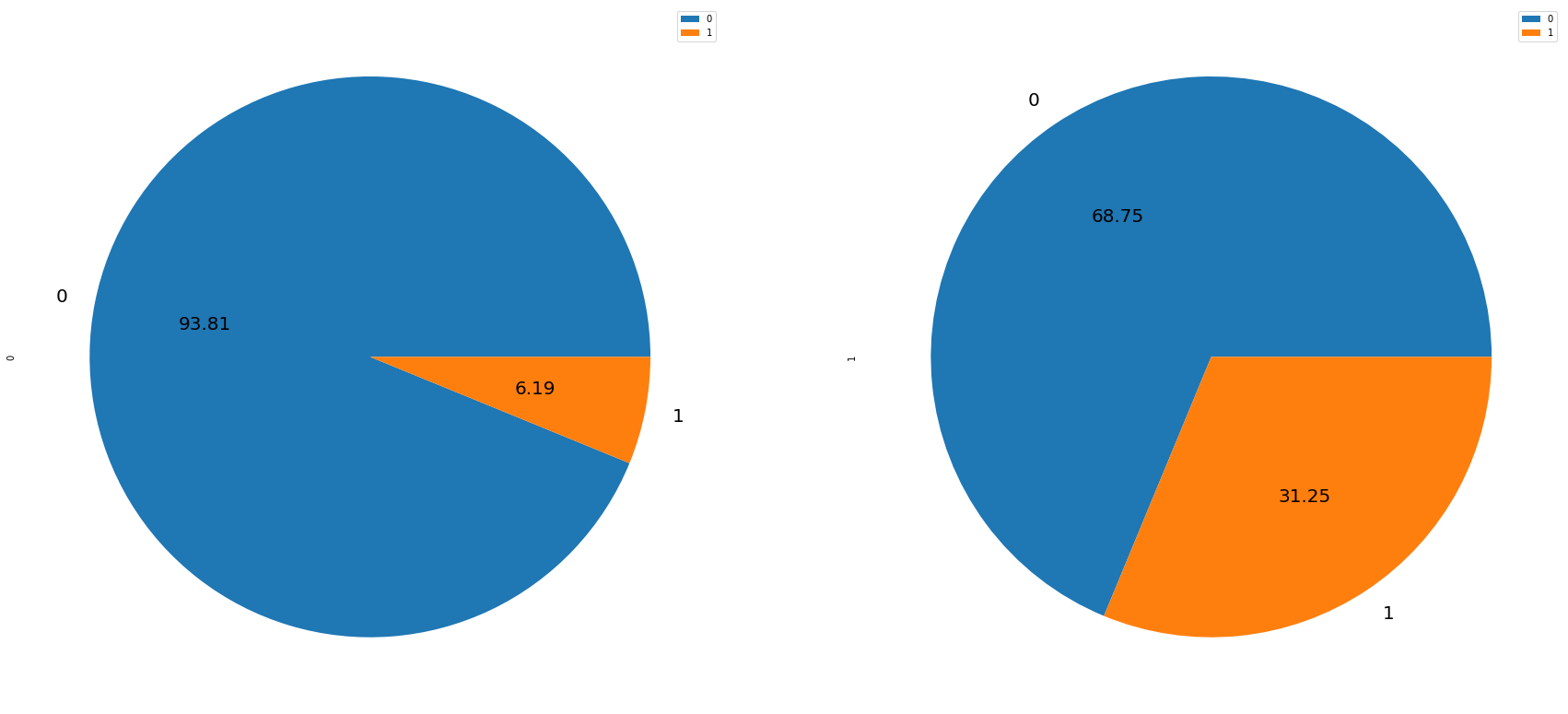


图14 Dx 和 Biopsy 交叉饼图

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表34 Dx 和 Biopsy 的卡方检验结果** | | | | | |
|  | 值 | 自由度 | 渐进显著性（双侧） | 精确显著性（双侧） | 精确显著性（单侧） |
| 皮尔逊卡方 | 15.438a | 1 | 0.000 |  |  |
| 连续性修正b | 11.742 | 1 | 0.001 |  |  |
| 似然比 | 8.937 | 1 | 0.003 |  |  |
| 费希尔精确检验 |  |  |  | 0.003 | 0.003 |
| 线性关联 | 15.414 | 1 | 0.000 |  |  |
| 有效个案数 | 646 |  |  |  |  |
| a. 1 个单元格 (25.0%) 的期望计数小于 5。最小期望计数为 1.09。 | | | | | |
| b. 仅针对 2x2 表进行计算 | | | | | |

与Dx:Cancaer同理，可见人乳头瘤病毒会复发的患者活检为阳性的比例不会复发的患者的6倍，通过卡方检验可近似认为人乳头瘤病毒复发对患宫颈癌有显著影响。

### **2.3.7 筛查方法**

#### **Hinselmann、Schiller与Citology**

将Biopsy作为观测变量，将Hinselmann、Schiller与Citology作为控制变量，利用多因素方差分析方法，研究筛查方法Hinselmann、Schiller与Citology是如何影响控制变量的，并进一步研究哪几种筛查方法的组合是预测Biopsy的最佳组合。

|  |  |  |
| --- | --- | --- |
| **表35 主体间因子** | | |
|  | | 个案数 |
| Hinselmann | 0 | 616 |
| 1 | 30 |
| Schiller | 0 | 585 |
| 1 | 61 |
| Citology | 0 | 608 |
| 1 | 38 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **表36 Biopsy多因素方差分析结果** | | | | | |
| 因变量: Biopsy | | | | | |
| 源 | III 类平方和 | 自由度 | 均方 | F | 显著性 |
| 修正模型 | 22.964a | 6 | 3.827 | 135.582 | 0.000 |
| 截距 | 2.388 | 1 | 2.388 | 84.599 | 0.000 |
| Hinselmann | 0.005 | 1 | 0.005 | 0.173 | 0.677 |
| Schiller | 2.353 | 1 | 2.353 | 83.336 | 0.000 |
| Citology | 0.006 | 1 | 0.006 | 0.208 | 0.648 |
| Hinselmann \* Schiller | 0.062 | 1 | 0.062 | 2.188 | 0.140 |
| Hinselmann \* Citology | 0.641 | 1 | 0.641 | 22.707 | 0.000 |
| Schiller \* Citology | 0.327 | 1 | 0.327 | 11.570 | 0.001 |
| Hinselmann \* Schiller \* Citology | 0.000 | 0 | . | . | . |
| 误差 | 18.039 | 639 | 0.028 |  |  |
| 总计 | 44.000 | 646 |  |  |  |
| 修正后总计 | 41.003 | 645 |  |  |  |
| a. R 方 = .560（调整后 R 方 = .556） | | | | | |

由表可得，观测变量Biopsy的总变差SST为41.003，它被分解为8个部分，分别是：由Hinselmann引起的变差0.005，由Schiller引起的变差2.353，由Citology引起的变差0.006，由Hinselmann和Schiller的交互作用引起的变差0.062，由Hinselmann 和 Citology的交互作用引起的变差0.641，由Schiller 和 Citology引起的变差0.327，由Hinselmann 、Schiller 和 Citology的交互作用引起的变差0.000，以及由随机因素引起的变差18.039.这些变差除以各自的自由度后，得到各自的方差，并可计算出各F检验统计量 的观测值和各自由度下的概率P-值。其中Hinselmann \* Schiller \* Citology由于组内样本数过少导致结果不具有统计学意义，不显示。由于Schiller、Hinselemann\* Citology、Schiller \* Citology的概率P-值<0.001，小于显著性水平α，应拒绝原假设，可以认为Schiller的筛查结果阴性或阳性对Biopsy的结果存在显著差异，且Hinselemann和Citology、Schiller 和 Citology对Biopsy产生显著的交互作用。

## 预测

### **2.4.1 基础模型**

使用python的sklearn库中的机器学习模型进行预测。先把数据集划分为训练集和测试集，再将特征标准化，然后分别使用LogisticRegression、Decision Tree、Random Forest、GaussianNB、KNN进行训练与预测，分别算出Train\_Score、Test\_accuracy、f1score、recall、precision、roc\_auc。其中结果如下。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **表37 基础模型训练测试结果** | | | | | | | |
| 模型 | Train\_Score | | Test\_accuracy | f1score | recall | precision | roc\_auc |
| LogisticRegression | | 0.954 | 0.954 | 0.571 | 0.429 | 0.857 | 0.712 |
| Decision Tree | | 1.000 | 0.964 | 0.720 | 0.643 | 0.818 | 0.816 |
| Random Forest | | 1.000 | 0.964 | 0.696 | 0.571 | 0.889 | 0.783 |
| GaussianNB | | 0.330 | 0.294 | 0.170 | 1.000 | 0.093 | 0.619 |
| KNN | | 0.947 | 0.933 | 0.316 | 0.214 | 0.600 | 0.602 |

因为这是一份医疗数据集，我们的目的是要精准预测患者是否患有宫颈癌，所以应该更加重视recall，综合考虑recall与roc\_auc，我们选择Decision Tree 与 Random Forest作为基础模型。

### **2.4.2 特征选择**

选择RFE技术进行特征选择，递归删除特征，使用剩余特征构建模型，并计算模型的准确率。分别对Decision Tree和Random Forest使用RFE，并分别找到这两种模型的最佳特征。特征的选择基于recall。

对于Decision Tree，选出Age, Number of sexual partners, First sexual intercourse, Num of pregnancies, Smokes (years), Smokes (packs/year), Hormonal Contraceptives (years), IUD (years), STDs, STDs:genital herpes, STDs:HIV, Dx:CIN, Hinselmann, Schiller, Citology。

对于Random Forest，选出了Age, Number of sexual partners, First sexual intercourse, Num of pregnancies, Smokes (packs/year),Hormonal Contraceptives (years), IUD (years), Hinselmann, Schiller, Citology。

分别计算两种模型的分数，结果如下。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **表37 最终模型训练测试结果** | | | | | | | |
| 模型 | Train\_Score | | Test\_accuracy | f1score | recall | precision | roc\_auc |
| Decision Tree after Feature Selection | | 1.000 | 0.964 | 0.759 | 0.786 | 0.733 | 0.882 |
| Random Forest after Feature Selection | | 1.000 | 0.964 | 0.741 | 0.714 | 0.769 | 0.849 |

综合考量，最终选择Decision Tree after Feature Selection模型。

# 小结

## 研究结果

### **宫颈癌的病因**

#### **3.1.1.1 生物学因素**

##### 性传播疾病

通过对数据集的分析以及机器学习模型的特征选择，可知性传播疾病能促进宫颈癌发病，患有性传播疾病的患者患上宫颈癌的概率是未患性传播疾病患者的3倍。而在各种性传播疾病中，生殖器疱疹和HIV病毒的影响又格外大。

某些最普通的性传播疾病衣原体，可以提高宫颈癌的可能性高达6倍。HPV与其他微生物作用或可改变阴道内环境，增强复制能力和持久作用，从而引发感染。

生殖器疱疹是由于单纯疱疹病毒所引起，如果女性混合感染或者合并有单纯疱疹病毒，以及人乳头瘤病毒感染时，出现癌前病变或者宫颈癌的可能性有所增加。

艾滋病是由人免疫缺陷病毒（HIV）感染引起的慢性传染性疾病。在艾滋病晚期，人体免疫功能低下，会增加HPV感染的风险，免疫功能正常的女性感染 HPV 后发展为宫颈癌需要 15 到 20 年的时间，对于免疫功能异常的女性，仅需 5 到 10 年的时间。

##### 疾病复发概率

通过对数据集的分析以及机器学习模型的特征选择，发现容易复发癌症、宫颈上皮内瘤变、人乳头瘤病毒（HPV）的病人患上宫颈癌的比例是不易复发患者的5~6倍。

患有癌症可能说明此时人体的免疫功能低下，会增加宫颈癌的风险。此外，其他病毒也可能与HPV病毒具有协同作用。

宫颈上皮内瘤变是介于宫颈炎症与宫颈癌之间的一类病变，与HPV的感染密切相关。并且宫颈上皮内瘤病变是宫颈癌发展过程中一个重要阶段。

大约超过80%的女性一生中都感染过HPV，大多数HPV感染都是一过性的，我们的身体可以自行清除病毒，但若有其他因素导致HPV持续感染，则易发展为宫颈癌。99%以上宫颈癌患者的宫颈脱落细胞标本中，都可以检测到HPV感染。

#### **3.1.1.2 行为因素**

##### 性行为

通过对数据集的分析以及机器学习模型的特征选择，发现不洁的性行为会促进宫颈癌的发病，如同时交往多个性伴侣、初次性交年龄过小、怀孕次数过多等。

HPV是一种DNA病毒，主要通过皮肤或黏膜的直接接触传染。所以开始性生活后，HPV的感染率就会大大升高，而且显然初次性交年龄越小、性伴侣越多，HPV感染风险就越高。

研究发现，使用避孕套可以降低女性感染HPV的机会，它能够遮住私处的部分皮肤，但是仅对HPV的阻挡有意义而无法做到高效保护。若怀孕次数过多则可能使用避孕套次数较少，接触皮肤较多，增加感染HPV的概率；或者人工流产次数过多，可能发生宫颈损伤或炎症，最后导致宫颈糜烂，增加感染HPV的概率。

##### 避孕

通过对数据集的分析以及机器学习模型的特征选择，发现使用激素避孕药对患宫颈癌没有显著影响，但是使用年数长，会增加患宫颈癌的风险；使用宫内节育环会促进宫颈癌的发病，但是使用年数没有显著影响。

避孕药的主要成分是雌激素和孕激素复合制剂，作用主要是抑制排卵，改善宫颈粘连的一些症状。有研究发现，高水平的雌激素和HPV感染可能起到协同宫颈癌的发生的作用，故如果长期服用避孕药，会增加患宫颈癌的风险。

西班牙加泰罗尼亚Bellvitge生物医学研究所的Xavier Castellsagué医生及其同事报告，通过对26项流行病学研究进行汇总分析发现，宫内节育器(IUD) 虽然不能预防人类乳头瘤病毒(HPV)感染，却似乎可以降低HPV感染进展为宫颈癌的几率。使用IUD的女性在使用的第1年内罹患宫颈癌的风险减半，并且这种程度的保护作用在使用10年期间持续存在，与数据集的分析不符。但进一步数据分析发现，使用宫内节育器的患者怀孕次数是未使用患者的1.5倍，且使用宫内节育器年数均值为0.5，猜测当地的患者是因为怀孕次数过多后才选择了使用宫内节育器，且使用时间不长，在这之前可能已经感染了HPV或患上了宫颈癌。因为信息不足，没有控制变量等，所以造成了结果的不符。

##### 吸烟

通过对数据集的分析以及机器学习模型的特征选择，吸烟能促进宫颈癌发病，吸烟的患者患有宫颈癌的比例是不吸烟患者的1.5倍。而且吸烟的年数多会增加患宫颈癌的风险。

Pate 等研究指出排除HPV感染因素，主动吸烟可使宫颈癌的发生率增加2-5倍不等。马超等对被动吸烟与宫颈癌关系的Meta分析研究结果表明被动吸烟者患宫颈癌的可能性是非吸烟者的1.62倍。烟草中含有许多致癌物质，会阻碍抑癌基因起作用，从而增加人体的患癌几率。比如尼古丁和可替宁(尼古丁代谢产物)等可以通过呼吸道黏膜吸收入血，再作用到身体的各个部位，若集中分布在宫颈黏膜，就会诱导宫颈上皮发生细胞学异常反应，削弱局部免疫。

#### **社会学因素**

##### 诊断

通过对数据集的分析以及机器学习模型的特征选择，发现宫颈癌活检结果为阳性的患者患有性传播疾病的数量均值大于阴性患者。但是除去未感染性传播疾病的患者，阳性患者性传播疾病的诊断次数的平均值反而小于阴性患者。说明参与诊断治疗越少，患宫颈癌的风险越高。

绝大多数宫颈癌的病例发生在发展中国家，人们对常规筛查的问题意识较低，缺乏医学专业知识、拥有有限的医疗设备。而大多数发达国家已经执行了很多年宫颈癌筛查规划，可以在保健工作的不同环节帮助干预宫颈癌的预防及筛查。无论在发达国家或发展中国家，宫颈癌主要见于社会经济地位较低、受教育程度较低的女性中。研究表明，这些弱势群体难以获得信息，缺乏防癌意识，卫生状况较差，难以接近医疗中心，参与诊断治疗少，就诊时中晚期病例达80%以上。

### **宫颈癌的诊断技术**

通过对数据集的分析以及机器学习模型的特征选择，发现Schiller、Hinselemann 与 Citology结合、Schiller 与 Citology 对宫颈癌的诊断都效果显著。

宫颈细胞学检查是将宫颈的脱落细胞收集起来后制成细胞图片，通过细胞形态学的观察。

Schiller发明碘试验，将碘溶液涂在宫颈上观察其染色的部位，于碘未着色区域指导活检。

Hinselmann发明阴道镜，是妇科内窥镜之一，可将病变放大数倍到数十倍，有利于在镜下做一定的初步判断、取活检等。

目前宫颈病变的规范化诊断程序是首先通过“细胞学检查±HPV检测”进行初筛，然后通过阴道镜检查进行“四象限多点活检±ECC”，最后通过组织病理学检查进行确诊。

### **预测模型**

使用Python的 sklearn 库，将数据集划分为训练集和测试集，得到了一个预测模型：先将特征标准化，再进行特征选择，最后使用Decision Tree分类器，测试集的accuracy可达到0.964，recall可达到0.786，roc 可达到0.882。

## 研究启示

本次调研使用了委内瑞拉加拉加斯的加拉加斯大学医院收集的包含 858 名患者的人口统计信息、习惯和历史医疗记录，但因隐私问题，一些患者决定不回答一些问题，且部分异常值会影响后续分析，经删减后仅纳入646例医疗记录，在进行复杂分析时因样本量不够充分，且没有控制变量、缺少对照等等，部分分析结果与实际仍有一定误差。在后续研究中应增大数据集，以减少误差程度；宫颈癌的影响因素很多，本次调研的因素仍然不够全面，在后续研究中可以增加对雌激素、宫颈的炎症和创伤、流产次数、被动吸烟、治疗方法等的调研与分析，针对具体因素的研究可以控制变量、设置对照组等等。

通过研究结果可知，宫颈癌的病因包括生物学、行为、社会学等因素。

对于生物学因素，建议临床诊断时可更多关注性传播疾病，尤其是生殖器疱疹、HIV病毒、HPV病毒等。并对于癌症、宫颈上皮内瘤变、HPV病毒等的复发予以关注。呼吁大家接种疫苗，增加疫苗供应量等。

对于行为因素，建议在没有生育要求时的性接触使用避孕套、性接触前进行清洁、初次性交年龄不宜过小，性伴侣数量不宜过多，以切断性传播途径；建议使用避孕套，非必要不服用避孕药等，以减少协同作用；建议主动戒烟、劝说身边的人戒烟、远离吸烟者、进行规律的锻炼等，以减少主动吸烟及被动吸烟量。

对于社会学因素，建议提高广大公民的受教育程度，普及生理健康知识，提高防癌意识，加强性卫生教育及婚前健康检查与指导，鼓励或组织参加定期体检、宫颈筛查、接种HPV疫苗等等，积极完善我国的宫颈癌筛查与预防体系。若确诊癌症，应尽早积极治疗。

本次调研还提供了一个机器学习的预测模型，可以帮助临床诊断。但该模型仍可以继续优化，如调整参数、集成等。

希望本次调研可以让更多人了解到宫颈癌的诱因及诊断方法，正确认识到其重要性，做好预防或积极治疗。希望以后有机会做更深入的研究。愿更多的人不受癌症的困扰，多喜乐、长安宁。

# 致谢

诚挚地感谢我的论文指导老师刘燊老师。在刘老师的《社会科学统计软件与应用：SPSS》课上，我学会了统计分析方法、SPSS的基本使用方法、Prism的基本使用方法、论文的写作规范以及Word有关论文的使用方法。在论文选题与撰写过程中遇到了问题，刘老师也耐心解答。

诚挚地感谢助教黄瑞轩，为我论文写作过程中遇到的问题作出了耐心解答。

# 参考文献

1. 薛薇. 统计分析与 SPSS 的应用. Vol. 2. 中国人民大学出版社, 2014.
2. 姜建东. “基于数据挖掘的江苏省中医院宫颈癌辨治经验总结”. MA thesis. 南京中医药大学, 2019.
3. 贾海燕. “年轻女性宫颈癌临床及病理特征的Meta 分析”. MA thesis. 吉林大学, 2013.
4. 高淑平. “宫颈癌诊断与治疗的新进展”. In: 医学理论与实践 25.12 (2012), pp. 1446–1448.
5. Kelwin Fernandes et al. “Supervised deep learning embeddings for the prediction of cervical cancer diagnosis”. In: *PeerJ Computer Science* 4 (2018), e154.
6. Wen Wu and Hao Zhou. “Data-driven diagnosis of cervical cancer with support vector machine-based approaches”. In:

*IEEE Access* 5 (2017), pp. 25189–25195.