# 属性数据分析作业

sword

2019-09-21

## 目录

# 1 第一周作业

1. 教材 14 页习题一第 1 题。

**解**

(1) 这些是定性数据。

(2) 作出这些交通事故原因的频数频率分布表如下。

表 1: 50 起交通事故原因的频数频率分布表

| 驾驶因素 | 频数 | 频率（%） | 累计频率（%） |
|---|---|---|---|
| 驾驶错误 | 11 | 22 | 22 |
| 察觉得晚 | 21 | 42 | 64 |
| 判断失误 | 14 | 28 | 92 |
| 酒后或疲劳驾驶 | 3 | 6 | 98 |
| 偏离规定的行驶路线 | 1 | 2 | 100 |
| 合计 | 50 | 100 | |

(3) 借助 Python 软件绘制条形图与圆形图如下。

(i) 条形图

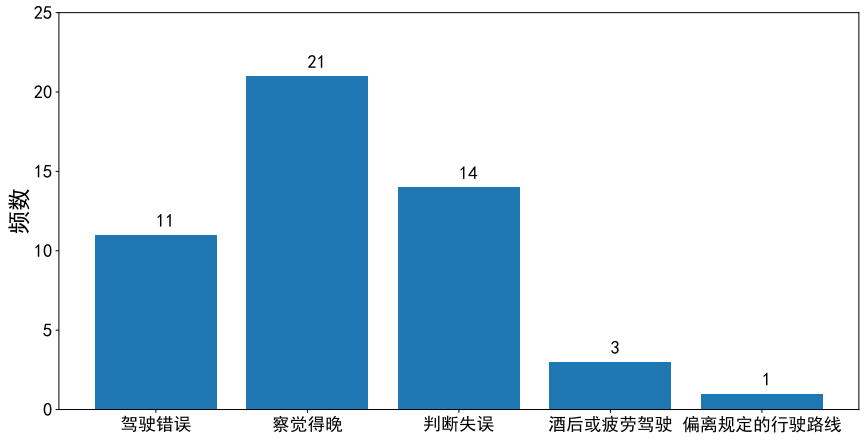


图 1: 50 起交通事故原因的条形图

(ii) 圆形图

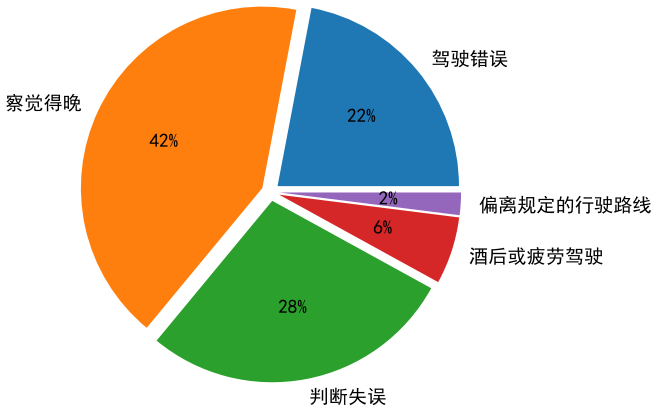


图 2: 50 起交通事故原因的圆形图

(4) 造成交通事故的驾驶因素中，察觉得晚最主要，判断失误其次，驾驶错误再次。

(5) 描述数据的中心位置可用众数、中位数和百分位数，这里给出众数和中位数。
众数：察觉得晚。
中位数：判断失误。
描述数据的离散程度可用离异比率、G-S 系数和熵。
离异比率：$V = 1 - f_{\text{察觉得晚}} = 58\%$ 。
G-S 系数：$\mathrm{G-S}(\xi) = 1 - \sum_{i=1}^{5} p_i^2 = 0.6928$ 。
熵：$H(\xi) = -\sum_{i=1}^{5} p_i \ln p_i = 1.30$ 。

2. 教材 29 页习题二第 6 题。

**解**

原假设为

$$\mathrm{H}_0 : p_1 = r^2 , p_2 = p^2 + 2pr , p_3 = q^2 + 2qr , p_4 = 2pq$$

同时满足 $r + p + q = 1$ 。先求出原假设成立下 $r , p , q$ 的极大似然估计。由已知，得似然函数满足

$$L(r , p , q) \propto r^{748}(p^2 + 2pr)^{436}(q^2 + 2qr)^{132}(pq)^{58}$$

约束条件为

$$r+p+q=1$$
$$0\leq r\,,p\,,q\leq 1$$

这里借助 MATLAB 软件进行非线性规划求解，得到结果为

$$r=0.61 \qquad p=0.29 \qquad q=0.10$$

由此可进行卡方拟合优度检验如下。

表 2: 血型遗传学模型的卡方检验

| 血型 | 观测值 | 理论值 | $\chi^2$ 统计量 | 自由度 | p 值 |
|---|---|---|---|---|---|
| O 型 | 374 | 372.1 | 0.0179 | 1 | 0.8936 |
| A 型 | 436 | 437.9 | | | |
| B 型 | 132 | 132 | | | |
| AB 型 | 58 | 58 | | | |

因此，在 5% 的显著性水平下，不应拒绝原假设，即认为数据与遗传学理论相符合。

附：具体实现代码如下。

(i) 条形图、圆形图的绘制（环境：Python 3.7.2 & vscode 1.37.1）

```
import pylab
import matplotlib.pyplot as plt

pylab.mpl.rcParams['font.sans-serif'] = ['SimHei']
pylab.mpl.rcParams['axes.unicode_minus'] = False

name_list = ["驾驶错误", "察觉得晚", "判断失误",
             "酒后或疲劳驾驶", "偏离规定的行驶路线"]
num_list = [11, 21, 14, 3, 1]

rects = plt.bar(name_list, num_list)
plt.ylim(0, 25)
plt.ylabel(u"频数", fontsize=25)
for rect in rects:
```

```
    height = rect.get_height()
    plt.text(rect.get_x()+rect.get_width()/2,
             height+0.5, str(height), fontsize=20)
plt.xticks(fontsize=19)
plt.yticks(fontsize=20)
plt.show()

explode = [0.05, 0.05, 0.05, 0.05, 0.05]
plt.pie(num_list, labels=name_list, explode=explode,
        autopct="%1.f%%", textprops={"fontsize": 18})
plt.show()
```

(ii) 非线性规划求解 $r, p, q$ 的极大似然估计（环境：MATLAB 2019a）

```
%建立m函数
function f  = cdhm_1_2(x)
f = -(748*log(x(1))+436*log(x(2)^2+2*x(2)*x(1))+
    132*log(x(3)^2+2*x(3)*x(1))+58*log(x(2)*x(3)));

%命令行
[x, fval] = fmincon(@ cdhm_1_2,
                    [0.1 0.3 0.6],
                    [], [],
                    [1 1 1], [1],
                    [0 0 0], [1 1 1])
```

## 2 第三周作业

1. 教材 68 页习题三第 1 题。

**解**

原假设为疫苗无效，备择假设为疫苗有效。当原假设成立时，对照组中的发病率应与处理组中的发病率近似相等；而当备择假设成立时，对照组中的发病率应显著高于处理组中的发病率。由此，我们采用四格表的优比检验法，结果如下。

表 3: 疫苗有效性的优比检验

| U 统计量 | p 值 |
|---|---|
| 5.8048 | $3.2221 \times 10^{-9}$ |

由于 p 值小于 0.05，故在 0.05 的显著性水平下拒绝原假设，即认为疫苗显著有效。

2. 教材 68 页习题三第 2 题。

**解**

由题设，构建四格表如下。

表 4: A,B 肥料对植物生长状况的影响

| | 长势良好 | 长势一般 | |
|---|---|---|---|
| A 种肥料 | 53 | 47 | 100 |
| B 种肥料 | 783 | 117 | 900 |

对此检验，原假设为两种肥料效果无明显差异，备择假设为 B 种肥料效果比 A 种肥料好。由于施加每种肥料的植物数量是确定的，从而此检验为单侧给定下的独立性检验。检验结果如下表所示。

表 5: 两种肥料效果差异的显著性检验

| U 统计量 | p 值 |
|---|---|
| -8.7111 | $1.5047 \times 10^{-18}$ |

由于 p 值小于 0.05，故在 0.05 的显著性水平下拒绝原假设，即认为 B 种肥料显著有效。

3. 教材 68 页习题三第 4 题。

**解**

由题设，构建四格表如下。

表 6: 精神病患者与神经病患者的自杀倾向对比

| | 有自杀情绪 | 无自杀情绪 | |
|---|---|---|---|
| 精神病患者 | 3 | 22 | 25 |
| 神经病患者 | 9 | 16 | 25 |

对此检验，原假设为精神病患者的自杀倾向比例与神经病患者的自杀倾向比例相等，备择假设为精神病患者的自杀倾向比例与神经病患者的自杀倾向比例不等。由于两种病患的调查数量是确定的，从而此检验为单侧给定下的独立性检验。检验结果如下表所示。

表 7: 精神病患者与神经病患者自杀倾向比例的显著性检验

| $\chi^2$ 统计量 | 自由度 | p 值 |
|---|---|---|
| 3.9474 | 1 | 0.0469 |

由于 p 值小于 0.05，故在 0.05 的显著性水平下拒绝原假设，即认为精神病患者的自杀倾向比例与神经病患者的自杀倾向比例不等。

4. 教材 69 页习题三第 5 题。

**解**

记 $n_{11}$ 为左半球良性肿瘤数，$n_{+1}$ 为良性肿瘤总数。由费歇尔精确检验，在给定 $n_{+1}=10$ 的条件下，$n_{11}$ 服从超几何分布 $h(16\,,12\,,10)$ 。

原假设为左半球的良性肿瘤和右半球的良性肿瘤一样多，备择假设为左半球的良性肿瘤多而右半球的良性肿瘤少。进行费歇尔精确检验，结果如下表。

表 8: 左右半球良性肿瘤数的显著性检验

| $n_{11}$ | p 值 |
|---|---|
| 9 | 0.0082 |

由于 p 值小于 0.05，故在 0.05 的显著性水平下拒绝原假设，即认为左半球的良性肿瘤多而右半球的良性肿瘤少。

5. 教材 69 页习题三第 7 题。

**解**

测试者大费周章地对一个人的品酒能力进行测试，说明测试者认为被试者有一定的品酒能力。从而原假设为测试者极力希望拒绝的命题，即被试者没有品酒能力，备择假设为被试者有品酒能力。

记实际为黄酒且品尝结果正确的数量为 $n_{11}$ ，品尝结果为黄酒的总数量为 $n_{+1}$ 。则在 $n_{+1}=15$ 给定的条件下，$n_{11}$ 服从超几何分布 $h(30\,,15\,,15)$ 。对此双侧给定的四格表，进行费歇尔精确检验，结果如下表所示。

表 9: 被试者品酒能力的显著性检验

| $n_{11}$ | p 值 |
|---|---|
| 11 | 0.0014 |

由于 p 值小于 0.05，故在 0.05 的显著性水平下拒绝原假设，即认为被试者有品酒能力。