# 任务报告

## 不同基线数据下患者出院状态分布图及其统计学描述

### 1.1 性别与出院状态分布



图1 性别与患者出院状态柱状分布图

Fig.1 Histogram of gender and patient discharge status

### 1.2 年龄与出院状态分布



图2 年龄与患者出院状态柱状分布图

Fig.2 Histogram of age and discharge status of patients

### 1.3 体重与出院状态分布



图3 体重与患者出院状态柱状分布图

Fig.3 Histogram of weight and discharge status of patients

### 身高与出院状态分布



图4 身高与患者出院状态柱状分布图

Fig.4 Histogram of weight and discharge status of patients

## 差异性分析

在完成了对不同基线数据下患者出院状态的分布图后，需要进一步利用差异性分析方法来对不同基线数据与患者出院状态的关联性进行评估。当前常见的差异性分析方法包括ANOVA(方差分析)、T-test(T检验)以及Chi-Square Analysis(卡方分析)这三种。由于不同基线数据（性别、年龄、体重、身高）与患者出院状态两类数据均为分类数据，综合考虑三种差异性分析方法应用条件，选择Chi-Square Analysis对不同基线数据与患者出院状态进行关联性分析。

### 2.1 提出假设

以基线数据中的性别为例，建立原假设H0与备择假设H1，假设内容如表1。原假设认为性别对患者出院状态无显著影响，备择假设认为性别对患者出院状态有显著影响。当原假设被推翻后，则认为备择假设成立。

表1 卡方分析的假设

Tab.1 The hypothesis of the Chi-Square Analysis

|  |  |
| --- | --- |
| Hypothesis | Hypothetical content |
| H0 | Gender has no significant impact on mortality at hospital discharge |
| H1 | Gender has a significant impact on mortality at hospital discharge |

### 2.2 计算卡方值

对不同基线数据下患者出院状态的数据进行分类预处理后，利用R中的chiq.test()函数进行计算出各组的χ2(卡方值)以及对应的*P*值(显著性水平)，结果如表2所示。

表2 不同基线数据患者出院状态的卡方检验结果

Tab.2 Chi-square analysis of patient discharge status with different baseline data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Baseline data | Number | Flaga=0 | Flag=1 | χ2 | *P* |
| Gender |  |  |  |  |  |
| Male | 3604 | 2452 | 1152 | 0.015798 | 0.9 |
| Female | 2669 | 1811 | 858 |
| Age |  |  |  |  |  |
| <=20 | 14 | 12 | 2 | 84.907 | 0.0004998 |
| 20~30 | 167 | 137 | 30 |
| 30~40 | 268 | 209 | 59 |
| 40~50 | 535 | 404 | 131 |
| 50~60 | 1192 | 849 | 343 |
| 60~70 | 1486 | 1006 | 480 |
| 70~80 | 1369 | 892 | 477 |
| 80~90 | 979 | 599 | 380 |
| >90 | 263 | 155 | 108 |
| Weight/kg |  |  |  | 10.004 | 0.2648 |
| <=20 | 63 | 47 | 16 |
| 20~30 | 978 | 639 | 339 |
| 30~40 | 2294 | 1576 | 718 |
| 40~50 | 1781 | 1199 | 582 |
| 50~60 | 715 | 482 | 233 |
| 60~70 | 252 | 182 | 70 |
| 70~80 | 93 | 70 | 23 |
| 80~90 | 52 | 36 | 16 |
| >90 | 45 | 32 | 13 |
| Height/cm |  |  |  |  |  |
| <=150 | 137 | 169 | 68 | 7.4 | 0.1925 |
| 150~160 | 1382 | 921 | 461 |
| 160~170 | 1942 | 1294 | 648 |
| 170~180 | 1929 | 1333 | 596 |
| 180~190 | 674 | 465 | 209 |
| >190 | 109 | 81 | 28 |

aFlag denotes patients discharge status. “Flag=1” denotes discharge status is death.

### 2.3 确定显著性水平

在进行假设检验前，需要事先设定显著性水平阈值α用于评估各组计算出的*P*值是否足够小以拒绝原假设。本文设置显著性水平阈值值α为0.05。当*P*值小于α时，表示认为有足够的证据拒绝原假设，反之认为观察到的差异是不显著的，无足够证据拒绝原假设。

### 卡方分析结果

将表2中各组数据计算出的*P*值与*α*值进行比较，结果如表3所示。从表中可以看出年龄组的P值<*α*，这表明年龄组的原假设被推翻，它的备择假设成立，即年龄对患者出院状态有显著影响。性别、体重以及身高组的P值>*α*，则原假设成立，即性别、体重以及身高对患者出院状态没有显著性影响。

表3 卡方分析结果

Tab.3 Chi-Square Analysis Results

|  |  |  |
| --- | --- | --- |
| Baseline data | *P* | Test Results |
| Gender | >*α* | Gender has no significant impact on mortality at hospital discharge |
| Age | <*α* | Age has a significant impact on mortality at hospital discharge |
| Weight | >*α* | Weight has no significant impact on mortality at hospital discharge |
| Height | >*α* | Height has no significant impact on mortality at hospital discharge |

## 3、出院死亡概率后验分布以及估计值随时间变化曲线

### 3.1 出院死亡概率*θ*后验分布

首先构建一个参数为*θ*的伯努利分布概率模型*X*~*bernoulli*(*θ*)，其中*X*为样本数据中患者出院状态。由贝叶斯定理有：

 (1)

式中为所求的后验分布，与分别给定*θ*下的数据样本的似然估计函数与*θ*的先验分布，分母的积分项为与*θ*无关的归一化常数c，常称为证据。因此理论上要求解后验分布的需要分别计算出、以及c即可求出后验分布。

#### 1）先验分布选择

假设患者出院状态死亡概率先验分布服从参数为(alpha=2, beta=5)的Beta分布，概率密度函数与累积分布函数如图5所示。



1. (b)

图5 出院状态死亡概率先验分布

a)概率密度函数 b) 累积分布函数

Fig.5 Prior Beta distribution of mortality probability at discharge state

1. Probability density function. b) Cumulative density function

#### 2）计算样本似然度



图6 出院状态死亡概率似然估计

#### 3）计算证据

#### 4）计算后验分布

利用患者出院状态的观测数据更新先验分布，得到最新的



1. (b)

图7 出院状态死亡概率后验分布

a)概率密度函数 b) 累积分布函数

Fig.5 Posterior distribution of mortality probability at discharge state

1. Probability density function. b) Cumulative density function

#### 5）出院死亡概率估计随观察量变化

首先选取前100个观测量计算出院死亡概率*θ*的先验Beta分布，得到初始的先验Beta分布参数为Beta（*θ,*35,65），然后逐个不断更新观测，不断调整后验Beta分布的参数，直至更新完所有样本，更新过程中的*θ*变化如图8所示。图中蓝色曲线为更新过程中*θ*的期望估计，灰色区域为*θ*估计的95%置信区间，红色虚线为更新完所有样本后*θ*的后验分布期望，其值为0.3204。



图8 出院状态死亡概率随观察量变化估计

Fig. 8 Estimation of the probability of death at discharge with observation

首先设定了先验分布的参数（在这里我们使用了均匀分布作为先验），然后计算了似然度（二项分布的概率质量函数）和证据（通过积分计算）。接下来，通过加权更新先验分布的参数得到后验分布的参数。最后，使用dbeta()函数计算后验分布的概率密度函数，并使用plot()函数绘制后验分布的概率密度图。同时，还绘制了先验分布的概率密度曲线和真实值Theta=0.5的竖线作为参考。最后，使用posterior\_alpha和posterior\_beta计算了后验分布的均值、中位数和95%置信区间，并将结果输出。