# 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

在提供的数据集mydata中，对性别(gender)、入院年龄(admission\_age)、入院体重(weight\_admit)以及身高(height)这四组基线数据下患者出院状态进行统计分析，绘制了这四组对应基线数据下患者出院状态的分布图，并进行了相应的统计描述。

### 1.1 性别与出院状态分布

在提供的数据集中共有6273份患者数据，其中男性（M）有3604份，女性（F）有2669份数据。在男性患者样本中，出院状态为死亡(flag=1)有1152份，死亡率为0.3196。在女性患者样本中，出院状态为死亡(flag=1)有1152份，死亡率为0.3215。上述分析数据分成男女两大类，绘制出堆叠式柱状图以及死亡率折线如图1所示。图中左侧纵轴表示人数，右侧纵轴表示患者出院状态的死亡率，横轴为性别。从图中直观上分析，男性与女性患者出院状态的死亡率差异很小，均在在0.32附近。



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

将数据集中6273为患者按照年龄进行分组，除了20岁以下以及90岁以上各自归为一组，年龄梯度按照5岁一个梯度进行划分，因此一共划分出了16个组别。将各个组别的出院状态进行统计并计算对应的死亡率，绘制成分布图如图2所示。对图2进行统计学描述如下：

1）样本数据数量在各年龄层存在一定差异，患者数量集中在40岁以上人群，从侧面反映出入院患者年龄上以中年、老年为主；

2）从出院状态死亡率折线图可以直观发现，患者出院状态死亡率随年龄增长呈现出增长趋势，在20岁以下仅为0.14，90岁以上达到0.41。



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

将数据集中6273名患者按照体重进行分组，除了30kg以下以及90kg以上各自归为一组，年龄梯度按照10kg一个梯度进行划分，因此一共划分出了8个组别。将各个组别的出院状态进行统计并计算对应的死亡率，绘制成分布图如图3所示。对图3进行统计学描述如下：

1）样本数据中患者的体重集中在40~70kg之间，约占到了97%

2）从出院状态死亡率折线图可以看出不同体重下患者出院状态死亡死亡率最大为0.34(30kg)以下，最低为0.25(70~80kg),差值为0.09。由于差异量并不十分显著，并且30kg以及70~80kg患者的数据量不是很充足，因此无法判断体重过轻对于患者出院状态死亡率是否有直接显著影响，需要后续进一步做差异性分析进行检验。



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

将数据集中6273名患者按照身高进行分组，除了150cm以下以及195cm以上各自归为一组，年龄梯度按照5cm一个梯度进行划分，因此一共划分出了11个组别。将各个组别的出院状态进行统计并计算对应的死亡率，绘制成分布图如图3所示。对图3进行统计学描述如下：

1）样本数据中患者的身高集中在150~185cm之间，约占到了90%

2）从出院状态死亡率折线图可以看出不同体重下患者出院状态死亡率最大为0.37(150~155cm)，最低为0.21(195cm)，差值为0.16。由于195cm以上身高患者数据样本数据量很少，导致计算出的最低死亡0.21存在一定偶然性，因此也需要进一步进行差异性分析去判断身高与患者出院状态死亡概率的关联性。



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。

结合伯努利分布似然函数以及Beta分布特点，当给定一个服从伯努利分布*bernoulli*(*θ*)的数据样本*X*=(*x*1,…,*x*n)以及某个共轭先验分布*θ~Beta*(*α,β*)时，可以得到后验分布：

 (2)

以前100名患者的出院状态作为经验，假设患者出院状态死亡概率先验分布服从参数为(alpha=35, beta=65)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

(a) Probability density function. (b) Cumulative density function

利用患者出院状态的观测数据更新先验分布，得到最新的后验分布的概率密度函数与累积分布函数如图6所示。后验分布仍然服从Beta分布，参数为*θ~Beta*(2012*,*4268)，其统计学指标如表4所示。



1. (b)

图6 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.6 Posterior distribution of mortality probability at discharge state

1. Probability density function. b) Cumulative density function

表4 后验分布参数与统计指标

Tab.4 Posterior distribution parameters and statistical indicators

|  |  |
| --- | --- |
| Statistical indicators | *θ* value |
| Mean | 0.3203822 |
| Median | 0.3203631 |
| Lower Confidence Interval(95%) | 0.308897 |
| Upper confidence interval(95%) | 0.3319757 |

### 3.2出院死亡概率估计随观察量变化

3.1中直接给出了最终由先验分布最终更新得到的后验分布，本节将给出后验分布更新过程中死亡概率*θ*的估计值，具体过程如下：

1）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）；

2）在先验分布基础上根据数据集中患者出院状态不断更新观测，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7 出院状态死亡概率随观察量变化估计

Fig. 7 Estimation of the probability of death at discharge with observation

## 4、二元混合高斯分布参数估计与模型拟合

### 4.1 随机数x概率密度

根据数据集中患者出院状态y生成一列符合混合高斯分布的随机数x，其中y=0所对应的x服从参数为N(0,2)的高斯分布，y=1所对应的x服从参数为N(4,2)的高斯分布，最终得到了具有混合高斯分布的随机数x的概率密度如8所示，其中图8(a)为全体x的概率密度图，8(b)为x在不同y类别下的概率密度图。

 

1. (b)

图8 混合高斯分布随机数x的概率密度分布

(a) 不按y分类 (b) 按y分类

Fig. 8 Probability density distribution plot of mixed Gaussian distribution random number x. (a) Not categorized by y. (b) Categorized by y

### 4.2 不同来源随机数x的概率密度分布与相对概率

针对4.1节中生成的随机数序列x分别计算其来源于高斯分布N(0, 2)与N(4, 2)的概率密度f0与f1以及其来源于出院状态y=1的相对概率f，结果如图9(a-c)所示。从图9(a, b)中不难发现x在两个高斯分布中的概率密度峰值出现在两个不同值，这也与4.1节生成的混合高斯分布相吻合。图9(c)中相对概率f从随着x值增大逐渐由0增至1。



图9 (a)观测值x来源于N(0, 2)的概率密度分布 (b) 观测值x来源于N(4, 2)的概率密度分布 (c) 观测值x来源于N(μ1, σ1)的相对概率

Fig.9 (a) Probability density distribution of observed value x from N(μ0, σ0) (b) Probability density distribution of observed value x from N(μ1, σ1) (c) Relative probability of observation x derived from N(μ1, σ1).

### 4.3 分类错误概率与ROC曲线

根据4.2节计算出的相对概率*f*，设定判断阈值为0.5，即观测值对应的y=1，*f*>0.5则认为分类正确，反之分类错误。由此可以得到混淆矩阵如表5所示。由表5中TP、FP、FN与TN的值进一步计算出当前阈值条件下的第一类错误、第二类错误概率以及F1值分别为0.084、0.079、0.879。

表5 混淆矩阵

Tab.5 Confusion matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Actual | |
| Death | Survival |
| Predicted | Judging death | True Positive=1841 | False Positive=337 |
| Judging survival | False Negative=169 | True Negative=3926 |

当模型判定阈值由0以步长为0.01不断变化时，根据模型敏感性（FPR）与精确性（TPR）计算公式得到ROC曲线如图10所示，图中显示了当前预测模型AUC值(十次平均结果)为0.9771，蓝色虚线为AUC=0.5的ROC线。由于当前模型的AUC=0.9771>0.5，这表明该分类模型在合理设定阈值的条件下具有预测价值。



图10 ROC曲线

Fig.10 ROC curve

### 4.4 EM算法估计二元混合高斯分布参数

对4.1节生成的符合二元混合高斯分的随机数x利用EM(Expectation-Maximum)算法进行参数估计。首先设置估计的初始化参数如表6：

表6 估计初始化参数

Tab.6 Estimate initialization parameters

|  |  |  |  |
| --- | --- | --- | --- |
| Distribution | Mean | Variance | Weight |
| 1 | -0.5 | 1.5 | 0.5 |
| 2 | 4.5 | 1.5 | 0.5 |

进一步在上述初始条件下进行EM算法迭代估计，由于每次迭代结果略有差异，因此取三次估计结果的均值作为最终得到随机数x混合高斯分布的参数估计与实际参数对比如表7所示：

表7 EM估计参数与真实参数对比

Tab.7 EM estimation parameters compared with real parameters

|  |  |  |  |
| --- | --- | --- | --- |
| Distribution | Mean | Variance | Weight |
| 1(Real) | 0 | 2 | 0.680 |
| 2(Real) | 4 | 2 | 0.320 |
| 1(Estimated) | 0.052 | 1.438 | 0.689 |
| 2(Estimated) | 4.12 | 1.401 | 0.311 |

利用EM算法估计得到的分布参数对y进行概率密度估计，得到真实y的概率密度与估计分布参数的概率密度图对比分别如图11(a, b)所示。



图11 真实概率密度与估计概率密度的对比

Fig.11 True probability density versus estimated probability density

利用估计出的分布参数，根据4.3节中的步骤求解出当前分类模型的ROC曲线如图12所示，图中曲线的Auc取10次平均后为0.9768。图12与4.3中利用真实y分布得到的ROC曲线相比基本一致，Auc值前者略低一些，但都在0.97左右大于0.5，表明EM算法估计出的高斯混合分布参数与原分布具有较高的一致性，得到的预测分布模型可以在合理设定阈值的条件下具有预测价值。



图12 ROC曲线（EM算法参数估计）

Fig.12 ROC curve (EM algorithm parameter estimation)

### 4.5训练集模型拟合与测试集预测

1)线性判别(Linear Discriminate Analysis, LDA)

线性判别分析进行建模前需要假设需要分类的*k*个总体分别服从高斯分布且各总体方差相同，然后分别求解自变量在*k*个类中的概率作为判断依据，将概率进行对数化并舍去不含*k*的共同项，得到线性判别函数如式(3)：

 (3)

式中的参数、与为待估计建模的参数，估计公式如式(4)：

 (4)

随机选取70%的样本作为训练集代入上式进行参数估计，结果如表8所示。

表8 LDA函数参数估计

Tab.8 LDA function parameter estimation

|  |  |
| --- | --- |
| Parameter | Estimated value |
|  | 0.679 |
|  | 0.321 |
|  | -0.027 |
|  | 4.013 |
|  | 2.031 |

根据估计的参数，分别得到随机数分属于两个类别的判别函数，将训练集与测试集代入判别函数，可以计算出两个数据集中随机数属于状态1的相对概率，按照0.01的步进不断改变判断阈值，得到两数据集的ROC曲线如图13(a, b)所示，其中图13(a)为训练样本集的ROC曲线，对应Auc值为0.9773，图13(b)为测试样本集的ROC曲线，对应Auc值为0.977。



图13 ROC曲线(LDA) (a) 训练集 (b) 测试集

Fig.13 ROC curve(LDA) (a) training set (b) test set

2）Fisher判别

Fisher判别也是一种线性判别分析方法，基本思想是将样本进行投影，使投影后类内方差最小，类间方差最大。在R中调用Mass库中的lda()函数可以实现模型的预测，随机选取样本中70%的样本作为训练集代入进行训练，得到模型的线性判别系数为0.719。在此基础上将训练样本集以及测试样本集代入训练完成的模型进行预测，最终得到了图14(a, b)所示的ROC曲线，其中图14(a)为训练样本集的ROC曲线，对应Auc值为0.9779，图14(b)为测试样本集的ROC曲线，对应Auc值为0.9697。



图14 ROC曲线(Fisher判别) (a) 训练集 (b) 测试集

Fig.14 ROC curve(Fisher Discrimination) (a) training set (b) test set

3）二次判别（Quadratic Discriminant Analysis Algorithm, QDA）

在进行LDA分析时假设的是各分类总体具有相同的协方差矩阵，而利用QDA则更加灵活，允许每个类别协方差矩阵独立估计，但参数估计相对复杂。在R中调用qda()函数进行模型预测，随机选取70%的样本作为训练集代入进行训练，最终得到训练集与测试集样本的ROC曲线分别如图15(a, b)所示，对应的Auc值均为0.9776。



图15 ROC曲线(QDA) (a) 训练集 (b) 测试集

Fig.15 ROC curve(QDA) (a) training set (b) test set

4）逻辑回归（Logistic Regression）

逻辑回归常用于解决二分类问题，利用逻辑函数(Sigmoid函数)的变换来建立模型，将输入的线性组合通过一个非线性映射，将结果限制在0和1之间，预测样本属于某个类别的概率。在R中利用glm()函数实现逻辑回归预测，同样随机选取70%的样本作为训练集代入进行训练，最终得到训练集与测试集样本的ROC曲线分别如图16(a, b)所示，对应的Auc值分别为0.9787与0.9740。



图16 ROC曲线(Logistic Regression) (a) 训练集 (b) 测试集

Fig.16 ROC curve(Logistic Regression) (a) training set (b) test set

由于本次进行训练的随机数服从的二元混合高斯分布，并且两高斯分布的方差相同，因此利用四种模型拟合方法得到的ROC曲线Auc值没有显著差异，而实际应用需要根据不同的分布选择对应合适的模型拟合方法进行分类处理方能实现更好的效果。