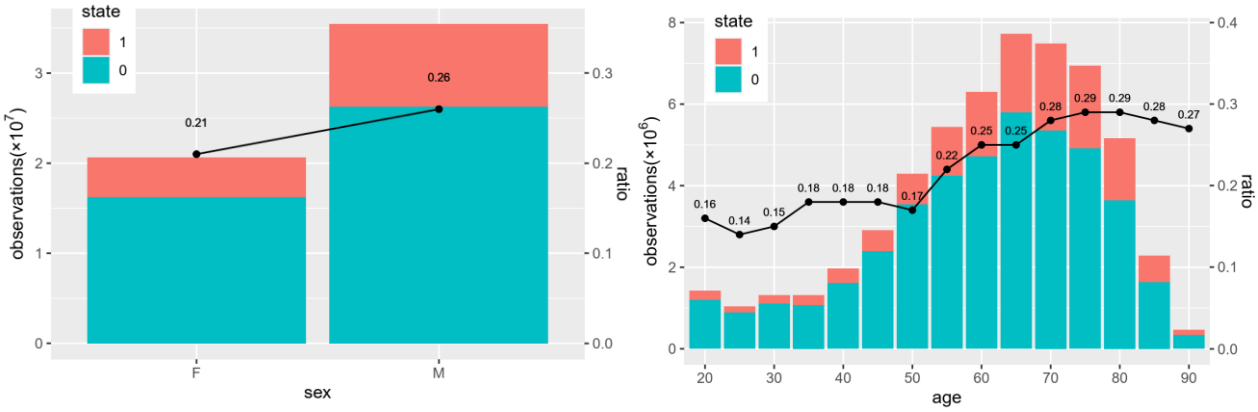
数据分析报告任务

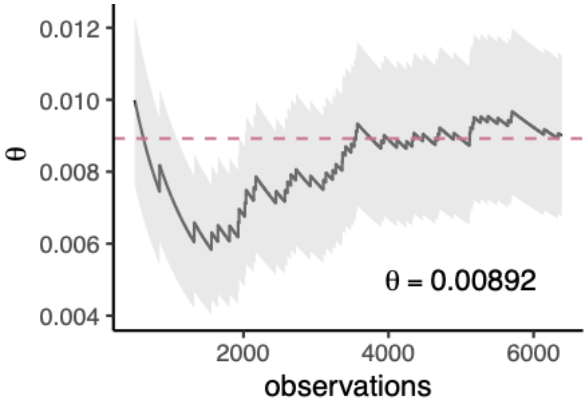
在 mydata 这个截面数据中，包含了患者的出院状态、住院时长等因变量，患者的基本统计数据（基线数据）、患者入院首日的数据等。使用该数据集，完成以下任务（先后顺序可调换）：

1、使用 ggplot 画出不同基线数据（身高、体重、年龄、性别等）下的患者出院状态的分布图，并用文字做统计描述，示例图如下：



2、选用合适的差异性分析方法如 ANOVA、卡方检验等，对不同基线数据下的患者出院状态做差异性分析，并对检验结果做统计描述。

3、在贝叶斯的框架下，请给出患者出院状态时的死亡概率θ的（1）后验分布，（2）画出θ的点估计值随时间变动的曲线图，（3）并添加相应的贝叶斯可信区间，示例如图所示。



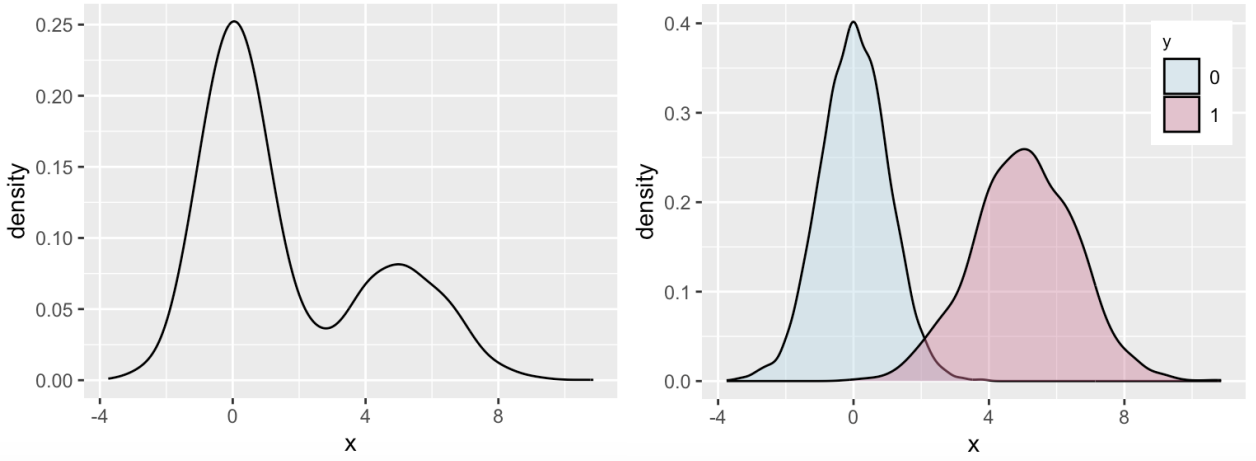
提示：可设置θ的先验分布为 Beta 分布，经验即患者出院状态 y 的观测，可认为服从二项分布，则θ的后验分布是什么分布，参数值为多少。选择前 100 个 y 的观测，生成θ先验分布的参数初始值。当第 101 个 y 的观测出现时，则可根据θ的后验分布更新其参数，并根据后验分布计算θ的点估计值和标准差（可选择后验期望），当第 102 个 y 观测出现时，重复该过程，则可得到一列随时间变化的点估计值和标准差，据此给出θ的点估计值曲线，并添加 95%的可信区间。

参考：https://web.stanford.edu/class/stats200/Lecture20.pdf

关于可信区间：理论计算难，建议采用数值方法，用 R 中的 qbeta 计算。

4、将患者出院状态 y 单独放入一个新的 excel 表格中，生成一列符合混合高斯分布的随机数，命名为 x。其中，y=0 所对应的 x 具有高斯分布$N(\mu 0, \sigma 0)$，y=1 所对应的 x 具有高斯分布$N(\mu 1, \sigma 1)$，具体参数数值自己决定。则该混合高斯具有形式：$\alpha N(\mu 1, \sigma 1) + (1-\alpha)N(\mu 2, \sigma 2)$，针对[x,y]这个数据，做如下任务：

（**4.1**）使用 ggplot 画出 **x** 的概率密度图以及 x 在给定 y 的情况下各类别的概率密度图，示例如下：



（**4.2**）对于 x 中的所有观测，分别计算其来源于$N(\mu 0, \sigma 0)$和$N(\mu 1, \sigma 1)$的概率密度，并进一步计算其来源于 1 的相对概率 f=f1/(f0+f1)。若该观测对应的 y=1，而 x 的 f 大于 0.5，则

认为分类正确，否则分类错误。使用概率密度 f 以及 y 计算第一类错误和第二类错误的概率、**auc** 值、**f1** 值等，并画出 **ROC** 曲线。

（**4.3**）在实际数据中，我们往往只有 x 的值，并不知晓 y 的值。在 x 服从二元混合高斯分布的假设下，使用 **EM** 算法（可拓展至其他算法）（1）计算$\alpha N(\mu 1, \sigma 1) + (1-\alpha)N(\mu 2, \sigma 2)$中的参数数值，并与真实的参数值做对比，（2）给出在 EM 算法下的 y 的估计值（概率密度）与真实 y 对比的 **ROC** 曲线，并与（4.2）的 ROC 曲线做对比。

（**4.4**）使用 x 与 y 的值，在将数据分为训练样本集和测试样本集的前提下，通过线性判别、Fisher 判别、Logit 回归分别对训练样本集进行模型拟合，后用于测试样本进行预测，并给出训练样本集、测试样本集的 **ROC** 曲线。

5、给出 mydata 数据中，首日检验数据以及出院状态共 109 个变量的相关性检验分析（此步无需结果），并给出与出院状态相关性最高的前 10 个变量，以及出院状态共 11 个变量的相关性图，并标注相关系数，且相关性越强，颜色越深。

6、从 5 中可以看出，虽然很多变量与出院状态 y 具有很强的相关性，但这些变量自身也具有很强的相关性，这导致很多模型不满足前提假设，强行使用会导致结果不可信等一系列问题。使用 Logit 回归模型对 5 中的 y 以及前 10 个变量做拟合，并与单独使用每个变量做拟合进行对比，描述各变量的系数、显著性的区别。

7、给出处理多重共线性的解决方案，并对 5 中的 109 个变量进行处理（缺失值等预先自行处理），并使用 Logit 回归模型进行拟合，解决方案应包括但不限于使用：Ridge，Lasso，PCA 等。对比不同解决方案下的模型结果。

8、选择一种 7 中的解决方案，并使用除去经典的 Logit 回归外，其他的分类模型对 6 中的数据进行拟合，应包括但不限于：SVM、随机森林、AdaBoost、XGBoost 等。对结果进行分析并对比这些模型的效果，建议以规范的表格和图的形式呈现。

9、在 8 中的众多模型中，往往涉及到模型中含有多个参量的情况，请使用（不限于）交叉验证等方法寻找最佳的参数组合，并与使用默认参数（default value）的模型结果进行对比，指出同一个模型在不同参数下的效果差异，建议以规范的表格和图的形式呈现。

10、将 mydata 中的基线数据纳入现有解释变量中，进行最佳参数组合下的模型构建。

提示：基线数据中含有多个类别型变量，它们在不同的模型中需要采取不同的预处理方法，如决策树类的模型可以直接处理类别型变量，无需自己预处理，而 Logit 回归则可处理数值型变量，需将类别型变量预处理后进行模型构建。若为二分类变量，则可使用数值 0-1 进行替换。对于多分类变量，如 n 类别变量，则需要将其转化为 n-1 个 0-1 数值变量。

综上：在完成 1-10 的任务后，大家可以在此基础上自由发挥，并最终以论文或数据报告等形式呈现。小组组员可自行决定如何分配报告得分，如报告 90 分，组员 A96 分，组员 B90 分，组员 C84 分，平均 90 即可。默认组员分数相同，如需重新分配请提前告诉我。报告暂定最后一次课前提交到 QQ 邮箱，并现场汇报最终报告（可辅助以 PPT）。

文章格式不限，但要满足学术规范，可参考：https://www.nature.com/articles/s41591-020-0789-4

数据来源：https://physionet.org/content/mimiciv/2.0/

其中，mydata 为 mimic III 中 患病名称（long_title）中包含有 sepsis 字样的截面数据。