认知建模基础作业4

> **Note**
>
> 1. 本次作业有三道题，总分按100分计，请在**JupyterLab**中使用 *python* 或者 *R* 在本文档中完成本次作业。
> 2. 请在做答时描述你的解题思路，并附上相应代码，我们将会根据结果的正确及清晰程度进行评分。
>   - 请确保文字结果及代码清晰、易读。
>   - 代码应有简要注释以及运行结果，运行结果应有相应的说明。请注意：注释允许使用中文/英文，可以使用co-pilot自动生成的注释，但要求对关键语句进行详细的注释，包括但不限于选择该语句的原因、参数设置的原因等。**代码的注释会纳入评分依据**
>   - 图片结果应当有相应的标题。
> 3. 最终提交时请重命名此ipynb文件。将运行好的结果导出成html，并将相应的文件打包上传，文件命名格式请参照 “作业N_编程语言_本科生/研究生_姓名_学号 ”，使用英文和拼音，例：“homework4_python_undergraduate_zhangsan_2101111111”(本科生)或者 “homework4_python_postgraduate_zhangsan_2101111111”(研究生)。

---

1 Bayesian Factor

在全部题目均为判断题的考试中，共有100道题目，答对0-59道题目为`不及格(F)`，答对60-84道题目为`及格(B)`，答对85-100道题目为`优秀(A)`，同学 X 前三次考试均为`及格(B)`，第四、五次考试成绩为`不及格(F)`，假设同学 X 的能力 θ （即做对每一道题的概率，不同题目是否做对是独立的，不同考试之间不发生改变）的先验是 $U(0,1)$。

1. 请你将成绩视作删失数据（censored data 可参考课上Cha Sa soon的例子），用MCMC采样的方法，报告同学 X 能力的后验分布，并基于后验分布，给出如果该同学额外参加一次考试，成绩为 `优秀(A)``及格(B)``不及格(F)` 概率的后验预测分布（10分）。
2. 有两个模型可以解释同学 X 的行为，$\mathcal{H}_0:\theta=0.5$ 和 $\mathcal{H}_1:\theta\sim U(0,1)$，对于这样的嵌套模型，我们可以用Savage-Dickey方法估计贝叶斯因子，具体做法是，在 $\mathcal{H}_1$的假设下进行MCMC采样后，使用参数 θ 的先验分布和后验分布计算：

$$BF_{01}=\frac{p(\mathcal{D}|\mathcal{H}_0)}{p(\mathcal{D}|\mathcal{H}_1)}=\frac{p(\theta=0.5|\mathcal{D},\mathcal{H}_1)}{p(\theta=0.5|\mathcal{H}_1)}$$

   其中先验分布（分母）可以直接获得，而后验分布（分子）可以对采样样本进行平滑（如直方图或者高斯核估计等）得到，请你自行查阅资料，解释算得贝叶斯因子对两个假设的支持程度（10分）。

3. 对于先验的选择，$\mathcal{H}_2:\theta\sim U(0.5,1)$ 可能更加合理，请你预测使用在这一先验下贝叶斯因子 BF_{02} 相比 BF_{01} 如何改变（你可以选择修改模型后重新采样的结果，或者使用 $\mathcal{H}_1$的结果通过数值计算估计新的贝叶斯因子的值，10分）

---

2 Bayesian Model Selection

在本题目中，我们希望通过一个现实情形的类比，帮助大家理解贝叶斯模型选择的概念。某中心的老师希望探究 A 同学的课堂参与情况，有两个模型进行解释：

- $\mathcal{M}_1$：该同学是**出席者**；
- $\mathcal{M}_0$：该同学是**缺席者**。
  检验模型的数据为 A 同学在16周课程中的签到情况 $\mathcal{D}=\{\mathcal{D}_i\in\{0,1\}\}_{i=1}^{16}$：`0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0`，其中 `1` 表示签到成功，`0` 表示签到失败。我们对模型证据作出如下假定：出席者签到成功的概率为99%，而缺席者签到成功的概率为40%，不同观测之间独立。即模型证据函数为：

$$p(\mathcal{D}_i|\mathcal{M}_1)=\begin{cases}0.99,\text{ if }\mathcal{D}_i=1\\0.01,\text{ if }\mathcal{D}_i=0\end{cases}$$
$$p(\mathcal{D}_i|\mathcal{M}_0)=\begin{cases}0.4,\text{ if }\mathcal{D}_i=1\\0.6,\text{ if }\mathcal{D}_i=0\end{cases}$$

1. 固定效应假说：该模型假设 A 同学要么是**出席者**（$\mathcal{M}_1$）要么是**缺席者**（$\mathcal{M}_0$），在整个学期不会发生改变。请分别计算似然函数 $p(\mathcal{D}|\mathcal{M}_1)$ 和 $p(\mathcal{D}|\mathcal{M}_0)$ ，并报告按照最大似然法选择的可以解释 A 同学行为的模型（5分）；
2. 在固定效应假说基础上，我们假定 A 同学是出席者的概率为 $p(\mathcal{M}_1)=\theta$, 而 θ 的先验概率为 $[0,1]$ 均匀分布，请画出图模型，并通过**MCMC**采样报告 θ 的后验分布和后验期望（5分）；
3. 随机效应假说：该模型假设 A 同学每节课前都会随机“选择”成为**出席者**（$\mathcal{M}_1$）或者**缺席者**（$\mathcal{M}_0$）。具体来讲， A 同学成为出席者的先验概率是 $r\sim U(0,1)$，隐变量 $z_i\sim Bernoulli(r)$ 表示每节课的状态（1对应出席者，0对应缺席者）。画出图模型，并通过MCMC采样报告 r 的后验分布和后验期望，同时报告哪那些周 A 同学更可能缺席了课程（5分）；
4. 在随机效应模型下，除了用 r 的后验期望表示 A 同学每节课前成为**出席者**的平均概率，我们还可以近似计算出席者模型($\mathcal{M}_1$)的胜出概率（probability of exceedance），其定义是：

$$\phi_1=p(r>1-r|\mathcal{D})=\int_{0.5}^1 p(r|\mathcal{D})dr$$

   我们可以用**MCMC**采样结果中 $r>0.5$ 的比例进行近似，请估计 $\mathcal{M}_1$ 的胜出概率（5分）
5. 下面我们考虑现实的模型比较问题，假设现在有 $\mathcal{M}_1$ 和 $\mathcal{M}_0$ 两个模型，分别对16名被试的数据进行逐被试的拟合得到两组AIC指标 `AIC1` 和 `AIC0`，我们可以用AIC指标对模型证

据进行近似：

$$\log p(\mathcal{D}_i|M_0) = -\frac{1}{2}AIC_0[i])$$
$$\log p(\mathcal{D}_i|M_1) = -\frac{1}{2}AIC_1[i])$$

除此之外均与2-4问中图模型结构一致，请你仿照2-4问的做法，（1）在固定效应假说下采样模型得到 θ 的后验分布和后验期望；（2）在随机效应假说下采样模型得到 r 的后验分布和后验期望，并报告模型 M_1 的胜出概率。（你可能需要用到pm.Potential函数，它接收的值应该是对数概率，10分）

6. 结合上面两个具体的例子，解释在固定效应和随机效应两种假设下，进行模型比较时会得到不同结果的可能原因，你认为在模型比较中更应该采用哪种思路？（5分）

Case study

3 Hierarchical Latent Mixture Model

在 `homework3` 中，我们引入了 `Coffee or Beer` 这样一道题目，题目内容如下：

> **Note**
>
> 在一项食品研究中，研究人员想要区分两类专业人士的口味偏好：
>
> - **咖啡师**：专长于咖啡的风味、烘焙程度、酸度等；
> - **酿酒师**：专长于啤酒的苦度、麦芽风味、酒精发酵工艺等。
>
> 咖啡师给咖啡评分(α_0)比啤酒分数(β_0)更高，而酿酒师给啤酒评分(α_1)比咖啡分数更高(β_1)。同时咖啡师给分极差更大($\beta_0 < \beta_1 < \alpha_1 < \alpha_0$)。在实际拟合数据时，**我们发现可能还存在第三类人：跨界专家**。他们是研究食品风味的科学家，对不同风味有自己研究，对啤酒和咖啡并无给分高低之分，但由于见多识广,评分(θ)不高不低($\beta_0 < \beta_1 < \theta < \alpha_1 < \alpha_0$)
> **另外，值得注意的是，第一名rater是酿酒师，第一杯饮料是咖啡**。在本题目中，**分数范围是0-100的连续值**。

上面的建模却忽略了个体差异，**比如不同咖啡师给咖啡打分均值和标准差可能也是不同的**。在这道题目中，我们利用**分层贝叶斯建模**，即考虑群体差异同时考虑个体差异，在上述假设基础上，本题增加额外假设如下：

> **Note**
>
> - **不同群体对不同饮料给分均值** μ ： 服从以下关系($\beta_0 < \beta_1 < \theta < \alpha_1 < \alpha_0$)，处于同一群体的不同个体对同一类别饮料给分均值相同。

- **不同群体对饮料给分幅度均值** $\bar{\sigma}_i$：由于酒精的作用，酿酒师给分相对来说比较随意，其对于饮品给分幅度均值($\bar{\sigma}_1$)相对较大，大于等于咖啡师给分幅度均值($\bar{\sigma}_0$)和跨界专家给分均值($\bar{\sigma}_2$)，也即($\bar{\sigma}_0 \leq \bar{\sigma}_1$， $\bar{\sigma}_2 \leq \bar{\sigma}_1$)。同时 $0 < \bar{\sigma}_1 < 5$。
- **每名个体给分幅度** σ_j： 服从高斯分布，均值为个体所在群体给分幅度均值$\bar{\sigma}_i$，标准差为定值 0.1，例如对某位酿酒师：$\sigma_j \sim \mathcal{N}(\bar{\sigma}_1, 0.1^2)$，并且同一个体对不同饮料给分幅度相同（即同一个个体给咖啡或者啤酒的标准差没有差异)；
- 每名个体 j 给不同饮料的评分均服从高斯分布 $\mathcal{N}(\mu, \sigma_j^2)$。

可参考课件中`6.2 Exam scores with individual differences`代码进行作答：

1. 请你根据本题目的假设写出模型并绘制出概率图模型(5分)
2. `homework3`题目`coffee or beer`中`h2`的概率图模型见附件`cb_h2.pdf`，请你比较和(1)中生成的概率图模型。请说明本题目中**群体差异**和**个体差异**是如何在概率图模型中体现(5分)
3. 请采用MCMC进行采样报告参数α_0，α_1，β_0，β_1，θ，σ_0，σ_1, σ_2的期望并绘制出其对应的后验分布（10分)【采样时请设置`tune=3000`， `random_seed=422`，`target_accept=0.9`】
4. 利用(3)中的后验分布，请绘制出综合所有链与所有采样的σ_j的箱线图(5分)
5. 斯丢皮德认为，本题目的模型在理想采样情况下likelihood会不小于homework3中h2对应的理想采样结果对应的likelihood。请你根据如下理由进行思考，**斯丢皮德说的是否正确?第三次作业中h2是否是本次作业模型的嵌套?** （5分)
   他的理由如下：

   > **Tips**
   > 从模型关系上来看，作业3中h2的假设对应的模型是本次作业假设对应模型的嵌套，homework3的h2对应的模型并未考虑到不同类别人群的群体差异以及群体内部的个体差异，而本次作业对应模型则考虑到这一点，因此表征能力更强。即使尚未存在群体差异与个体差异，本文对应模型在理想状态下对应参数也会收敛到作业3中h2到模型，因此本作业对应模型likelihood应该不小于上次作业模型的likelihood。

6. 如果我们**没有第一杯饮料是咖啡**这样的假设，请你重新进行采样，在`tune=2000`, `random_seed=422`, `target_accept=0.9`的情况下重新进行实验，请报告`xi`， `sigmaj`的r hat值，请问为什么会出现这样的情况(5分)