清华大学本科生考试试题专用纸

考试课程：贝叶斯统计导论　　　样题

姓名:__________　　学号:__________　　院系名称:______________　　成绩:_______

请注意：本试卷包括五个大题，满分 100 分.

一. (25 分) 假设有一枚硬币，每次抛掷它得到正面的概率为 θ，得到反面的概率为 $1-\theta$. 现连续抛掷 n 次 ($n>1$)，一共得到了 y 个正面.

(1)(5 分) 采用区间 [0,1] 上的均匀分布作为 θ 的先验分布，请写出对应的 θ 的后验分布.

(2)(5 分) 采用区间 [0,1] 上的均匀分布作为 θ 的先验分布，利用上问结论，请证明对应的 θ 的后验方差总是小于 θ 的先验方差.

(3)(5 分) 采用区间 [0,1] 上的均匀分布作为 θ 的先验分布，请推导出 y 的先验预测分布 (prior predictive distribution): $Pr(y=k)=\int_0^1 Pr(y=k|\theta)d\theta$, 其中 $k=0,1...,n$.

(4)(5 分) 采用 $Beta(\alpha,\beta)$ 作为 θ 的先验分布，请写出对应的 θ 的后验分布.

(5)(5 分) 采用 $Beta(\alpha,\beta)$ 作为 θ 的先验分布，利用上问结论，请证明对应的 θ 的后验均值的大小总是在 θ 的先验均值 ($\frac{\alpha}{\alpha+\beta}$) 和观测均值 ($\frac{y}{n}$) 之间.

二. (5 分) (Jeffrey's 无信息先验) 假设 $y|\theta \sim Poisson(\theta)$. 求 θ 的 Jeffrey's 先验密度.

三. (15 分) 设数据 $(y_1,...,y_J)$ 来自参数为 $(\theta_1,...,\theta_J)$ 的多项分布，其中 $J>2$ 是已知给定的整数. 取定 $\theta=(\theta_1,...,\theta_J)$ 的先验分布为参数为 $(a_1,...,a_J)$ 的 Dirichlet 分布. 记 $\alpha=\frac{\theta_1}{\theta_1+\theta_2}$.

(1)(8 分) 求出 (θ_1,θ_2) 的边缘后验分布.

(2)(7 分) 求出 (α) 的边缘后验分布.

四. (20 分) 判断题 (判断为错的题目，请简要说明理由；若缺少理由或理由错误，不得分):

(1)(2 分) 共轭先验与模型 (抽样分布) 来自同一个分布族.

(2)(2 分) 参数 θ 的无信息先验就是指在 θ 的取值范围上的均匀分布.

(3)(2 分) 参数 $(\theta_1,...,\theta_m)$ 满足可交换性 (Exchangeability) 等价于 $(\theta_1,...,\theta_m)$ 之间相互独立.

(4)(2 分) 假定观测值 y 来自于参数为 θ 的某个分布，如果 $\hat{\theta}$ 是后验均值，则 $log(\hat{\theta}|y)$ 不是随机变量.

(5)(2 分) 假定观测值 $y=\{y_1,...,y_n\}$ 来自于参数为 θ 的某个分布. 只要样本量 n 足够大，就有 $\theta|y$ 是渐近正态的，即 θ 的后验分布渐近收敛到某个正态分布.

(6)(2 分) 贝叶斯学派和频率学派都有关于参数的渐近正态性理论，前者针对给定样本下参数的后验分布，后者针对给定参数真值下参数估计的分布.

(7)(2 分) 模型检验中，常用的一种内部验证方法是通过后验预测分布生成重复数据 y^{rep}，基于某个针对特定假设而选择的统计量 $T(y,\theta)$，计算 $p=Pr(T(y^{rep},\theta)\geq T(y,\theta)|\theta)$ 来考察该假定是否合理；如果 p 值太小 (<0.05)，我们就要拒绝该模型，在除了该模型之外的剩余模型中选择一个模型重新计算.

(8)(2 分) 上问中提到生成重复数据 y^{rep}，抽样时只需控制样本量和原数据一样即可.

(9)(2 分) 模型比较中，由于要修正重复使用样本导致的偏差，引入了 AIC, DIC, WAIC, Cross Validation 等基于预测效果进行模型比较的指标 (也称，对预测精度的度量, measures of predictive accuracy).

(10)(2 分) 假设抛硬币得到 9 个正面 3 个反面，用 θ 表示硬币出现正面的概率. 考虑假设检验 $H_0 : \theta = \frac{1}{2}$ 和 $H_1 : \theta > \frac{1}{2}$. 由于存在似然原则 (Likelihood Principle)，无论抽样模型是二项分布还是负二项分布，频率学派或贝叶斯学派下，我们都将得到一致的结论. 例如，在贝叶斯学派下，我们可以使用 Bayes Factor 来进行比较. 所以当我们确定获得 9 个正面和 3 个反面后，就无法区分抽样模型是二项分布还是负二项分布了.

五. (35 分) 现研究我校 J 个不同院系的同学们的英语六级成绩. J 个院系中，第 j 个院系有 n_j 个同学考了托福，其成绩分别为 $y_{ij}, i = 1, ..., n_j$. 考虑用一个层次模型来研究该问题. 假设 y_{ij} 来自均值为 θ_j，方差已知的正态总体，具体地，有 $y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma^2), i = 1, ..., n_j, j = 1, ..., J$; 其中 σ^2 已知. 同时假定 $\{\theta_j, j = 1, ..., J\}$ 的先验分布都是 $\mathcal{N}(\mu, \tau^2)$，且给定 μ, τ 下是独立的. 取超参数 (μ, τ) 的先验分布为 $p(\mu, \tau) \propto 1$。

(1)(15 分) 求 θ_j 的条件后验分布，即 $p(\theta_j|\mu, \tau, y)$. (Hint: 先基于 $\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$ 的分布求联合后验分布 $p(\theta, \mu, \tau|y)$).

(2)(10 分) 求 $E(\theta_j|\tau, y)$, $var(\theta_j|\tau, y)$.

(3)(10 分) 现发现除了这 J 个院系，还有一个系随机挑选了一位同学 B 去参加英语六级考试. 请问，如何根据上述模型和已观测数据 $(y_{ij}, i = 1, ..., n_j, j = 1, ..., J)$，通过模拟得到这位同学 B 的英语成绩的后验预测分布和后验均值? (需逐步写明从哪个分布抽样，例如，从 $p(\theta_j|\mu, \tau, y)$ 中抽取 θ_j；但无需写出对应分布的具体概率密度表达式)