Problem Set 2

Due on November 09

说明：请给出计算、证明或构造步骤，以及相应的程序和运行结果。

习题 1

用随机模拟法计算二重积分 $\int_0^1 \int_0^1 e^{(x+y)^2}\, dy\, dx$, 用对立变量法改善精度。

习题 2

设 $p(x,\mu,\sigma)$ 表示 $\mathrm{N}(\mu,\sigma^2)$ 的分布密度函数，设总体 X 分布密度为

$$f(x)=\lambda p(x,\mu,\sigma)+(1-\lambda)p(x,\mu,k\sigma)$$

其中 $0<\lambda<1$，$k>1$。这称为混合正态分布。从总体中抽取简单随机样本 $X_1,X_2,\ldots,X_n$，对参数 μ 进行估计，有两种估计方法：

(1) 样本均值 $\bar{x}$；
(2) 取 $0<p<0.5$，计算删去最小的 p 比例和最大的 p 比例的样本值之后的样本平均值，记作 $\bar{x}_p$，称为 Winsorized mean。

这两个估计都是无偏的，哪一个的方差更小？理论上不容易解决。用随机模拟方法研究。需要设定一些不同的 μ, σ, λ, k, p, n 的值，估计需要的重复次数 N，比较不同情况下两种方法的估计方差。

习题 3

设 $\{X_t, t=0,1,\ldots\}$ 为状态取值于 $S=\{1,2,\ldots,m\}$ 的马氏链，有平稳分布 $\pi_j, j=1,2,\ldots,m$，若 X_0 服从平稳分布，证明所有 X_t 都服从平稳分布 $(t\geq 0)$。

习题 4

设 $X \sim \mathrm{b}(1,p)$, $X_1, X_2, \ldots, X_n$ 为样本。令 $S_0 = \sum_{i=1}^n X_i$, $\hat{p} = S_0/n = \frac{1}{n}\sum_{i=1}^n X_i$。用模拟方法比较如下五种置信区间：

(1) 利用正态近似。当 n 很大时

$$\frac{\hat{p}-p}{\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}}$$

近似服从 $\mathrm{N}(0,1)$，于是得置信区间

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}.$$

(2) 利用正态近似，令

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}),$$

n 很大时

$$\frac{\hat{p}-p}{\sqrt{\frac{1}{n}S^2}}$$

近似服从 $\mathrm{N}(0,1)$，于是得置信区间

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n}S^2}.$$

(3) Wilson 置信区间。利用正态近似，n 很大时

$$\frac{\hat{p}-p}{\sqrt{\frac{1}{n}p(1-p)}}$$

近似服从 $\mathrm{N}(0,1)$，解关于 p 的不等式

$$\left|\frac{\hat{p}-p}{\sqrt{\frac{1}{n}p(1-p)}}\right| \le z_{1-\frac{\alpha}{2}},$$

得置信区间 ($\lambda = z_{1-\frac{\alpha}{2}}$)

$$\frac{\hat{p}+\frac{\lambda^2}{2n}}{1+\frac{\lambda^2}{n}} \pm \frac{\lambda\sqrt{\frac{\lambda^2}{4n}+\hat{p}(1-\hat{p})}}{\sqrt{n}\left(1+\frac{\lambda^2}{n}\right)}.$$

(4) Agresti and Coull(1998) 方法。先计算 Wilson 区间中心 ($\lambda = z_{1-\frac{\alpha}{2}}$)

$$\tilde{p} = \frac{\hat{p} + \frac{\lambda^2}{2n}}{1 + \frac{\lambda^2}{n}},$$

取置信区间为

$$\tilde{p} \pm \lambda \frac{\sqrt{\tilde{p}(1-\tilde{p})}}{\sqrt{n}\sqrt{1 + \frac{\lambda^2}{n}}}.$$

(5) 用二项分布导出 p 的置信区间。令

$$p_1 = \left\{1 + \frac{n - S_0 + 1}{S_0} F_{1-\frac{\alpha}{2}}(2(n - S_0 + 1), 2S_0)\right\}^{-1}$$
$$p_2 = \left\{1 + \frac{n - S_0}{S_0 + 1} \frac{1}{F_{1-\frac{\alpha}{2}}(2(S_0 + 1), 2(n - S_0))}\right\}^{-1}$$

取置信区间为 $[p_1, p_2]$，其中 $F_q(n_1, n_2)$ 表示 $\mathrm{F}(n_1, n_2)$ 分布的 q 分位数。

对不同 $n, p, 1-\alpha$，比较这五种置信区间的优劣。

习题 5

考虑下的非线性回归模型 (logistic 曲线)

$$y = A(1 + e^{-bx}) + \varepsilon,\ \varepsilon \sim \mathrm{N}(0, \sigma^2),$$

其中 $A > 0, b > 0, \sigma^2 > 0$ 是未知参数。设有独立样本 $(x_i, y_i), i = 1, 2, \ldots, n$, 可以用最小二乘方法估计 A, b, σ^2, 估计值记为 $\hat{A}, \hat{b}, \hat{\sigma}^2$。

(1) 设真实的 $A = 10, b = 1, \sigma = 1$。编写程序模拟一组样本（取 $x_i = -10, -9, \ldots, 9, 10$），计算 $\hat{A}, \hat{b}, \hat{\sigma}^2$，然后用 bootstrap 方法估计 $\hat{A}, \hat{b}, \hat{\sigma}^2$ 的标准误差和偏差。

(2) 重复生成 N 组模型的样本，从这 N 组样本中分别得到估计值 $\hat{A}^{(j)}, \hat{b}^{(j)}, (\hat{\sigma}^{(j)})^2$，$j = 1, 2, \ldots, N$, 用得到的这些估计值作为 $\hat{A}, \hat{b}, \hat{\sigma}^2$ 的抽样分布的样本，估计其标准误差和偏差并与 bootstrap 方法得到的结果进行对比。

习题 6

设向量 $x=(x_1,\ldots,x_n)$, 每个 x_i 取值为 $+1$ 或 -1，记这样的 x 的集合为 $\mathcal{X}$。设随机向量 X 有如下概率质量函数

$$\pi(x)=P(X=x)=\frac{1}{Z}\exp\left\{\mu\sum_{i=1}^{n-1}x_ix_{i+1}\right\}$$

其中 μ 已知，Z 为归一化常数。这样的模型称为一维 Ising 模型。取 $\mu=1,n=50$, 设计 MCMC 算法从 $\pi(x)$ 抽样。

习题 7

设随机变量 X 和 Y 都取值于 $(0,B)$ 区间 (B 已知)。设 $Y=y$ 条件下 X 的条件分布密度为

$$f(x|y)\propto e^{-yx},\ x\in(0,B),$$

$X=x$ 条件下 Y 的条件分布密度为

$$f(y|x)\propto e^{-xy},\ y\in(0,B),$$

编写 R 程序用 Gibbs 抽样方法对 (X,Y) 抽样，估计 EX 和 $\rho(X,Y)$。