Homework 11

抹茶奶绿

May 12, 2025

Exercise 1.

从一批次产品随机地取 100 个样品进行检测，发现 40 个不合格，求这批产品合格率 p 的 95% 置信区间估计.

Solution.

记样本合格率 $\hat{p} = 0.60$，正态近似

$$\hat{p} \sim N\Big(p,\ \frac{p(1-p)}{n}\Big)$$

由

$$P\Big(\big|\hat{p} - p\big| \le z_{0.975}\sqrt{\frac{p(1-p)}{n}}\Big) \approx 0.95$$

可得

$$n(\hat{p}-p)^2 \le z^2\, p(1-p) \quad \Longrightarrow \quad (n+z^2)p^2 - (2n\hat{p}+z^2)p + n\hat{p}^2 \le 0$$

于是

$$p = \frac{2n\hat{p} + z^2 \ \pm\ z\sqrt{z^2 + 4n\hat{p}(1-\hat{p})}}{2\,(n+z^2)}$$

代入 $n = 100,\ \hat{p} = 0.60,\ z^2 = 1.96^2 \approx 3.8416$

$$p_{\mathrm{L}} \approx \frac{123.8416 - 1.96\sqrt{99.8416}}{207.6832} \approx 0.502 \quad p_{\mathrm{U}} \approx \frac{123.8416 + 1.96\sqrt{99.8416}}{207.6832} \approx 0.690$$

因此在该近似下，95% 置信区间约为 $(0.502,\ 0.690)$.

Exercise 2.

设随机样本 $X_1, \ldots, X_n$ 来自概率密度

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1, \\ 0, & \text{其它}, \end{cases}$$

其中未知参数 $\theta > 0$，$\hat{\theta}$ 为 θ 的极大似然估计量.

（1）当 $n = 4$ 时，一组样本观测值为 $X = (1/2, 1/3, 1/5, 1/5)$，利用 Fisher 信息量和该组观测值给出 $\hat{\theta}$ 的标准误差估计.

（2）基于（1）的估计给出满足 95% 要求的 θ 的置信区间.

Solution.

(1) 似然函数

$$L(\theta) = \prod_{i=1}^{4} \theta\, x_i^{\theta-1} = \theta^4 \Big(\prod_{i=1}^{4} x_i\Big)^{\theta-1}$$

对数似然

$$\ell(\theta) = 4\ln\theta + (\theta-1)\sum_{i=1}^{4}\ln x_i \implies \hat{\theta} = -\frac{4}{\sum_{i=1}^{4}\ln x_i}$$

计算 $\sum \ln x_i = \ln\dfrac{1}{2} + \ln\dfrac{1}{3} + 2\ln\dfrac{1}{5} = -\ln 2 - \ln 3 - 2\ln 5 \approx -4.6052$，故

$$\hat{\theta} \approx \frac{4}{4.6052} \approx 0.8685$$

Fisher 信息量

$$I(\theta) = -E\Big[\frac{\partial^2 \ell}{\partial \theta^2}\Big] = \frac{4}{\theta^2}$$

因此

$$\mathrm{se}(\hat\theta) \approx \frac{1}{\sqrt{I(\hat\theta)}} = \frac{\hat\theta}{\sqrt{4}} = \frac{0.8685}{2} = 0.4343$$

(2) 近似 $\hat\theta \sim N\big(\theta,\ \mathrm{se}^2(\hat\theta)\big)$，取 $z_{0.975} = 1.96$，则

$$\hat\theta \pm z_{0.975}\,\mathrm{se}(\hat\theta) = 0.8685 \pm 1.96 \times 0.4343 = 0.8685 \pm 0.8504$$

故 95% 置信区间约为

$$(0.8685 - 0.8504,\ 0.8685 + 0.8504) \approx (0.0181,\ 1.7189)$$

$n = 4$ 太小导致误差较大 (?

Exercise 3.
假设总体服从 $N(\mu, \sigma^2)$，参数 μ 已知，σ^2 未知，$X_1, \ldots, X_n$ 为其随机样本，常数 $0 < \alpha < 1$.
(1) 求 σ 的极大似然估计 $\hat\sigma$.
(2) 利用 Fisher 信息量给出 $\hat\sigma$ 的标准误差估计.
(3) 利用 $\hat\sigma$ 给出 $\log\sigma$ 的 $1-\alpha$ 置信区间估计.

Solution.

(1) 总体密度

$$f(x;\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\}$$

对数似然函数

$$\ell(\sigma) = \sum_{i=1}^n\Big[-\ln\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}\Big] = -n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2$$

求导

$$\frac{\partial\ell}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (x_i-\mu)^2 \overset{!}{=} 0 \quad\Longrightarrow\quad \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2,\quad \hat\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2}$$

(2) Fisher 信息量

$$I(\sigma) = -E\Big[\frac{\partial^2\ell}{\partial\sigma^2}\Big] = -E\Big[\frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^n (x_i-\mu)^2\Big] = -\Big(\frac{n}{\sigma^2} - \frac{3n\sigma^2}{\sigma^4}\Big) = \frac{2n}{\sigma^2}$$

因此

$$\mathrm{se}(\hat\sigma) \approx \frac{1}{\sqrt{I(\hat\sigma)}} = \frac{1}{\sqrt{2n}/\,\hat\sigma^{-1}} = \frac{\hat\sigma}{\sqrt{2n}}$$

(3) 我们关心 $\tau = \log\sigma$. 由 Delta 方法

$$\mathrm{Var}(\log\hat\sigma) \approx \Big(\frac{d}{d\sigma}\log\sigma\Big|_{\sigma=\hat\sigma}\Big)^2 \mathrm{Var}(\hat\sigma) = \frac{1}{\hat\sigma^2}\Big(\frac{\hat\sigma^2}{2n}\Big) = \frac{1}{2n}$$

于是近似 $\log\hat\sigma \sim N\big(\log\sigma,\ 1/(2n)\big)$. 取 $z_{1-\alpha/2}$ 为标准正态上 $1-\alpha/2$ 分位点，可得

$$\Big[\log\hat\sigma\ -\ \frac{z_{1-\alpha/2}}{\sqrt{2n}}\ ,\ \log\hat\sigma\ +\ \frac{z_{1-\alpha/2}}{\sqrt{2n}}\Big]$$

Exercise 4.

为提高某化学生产过程的得率，原催化剂 20 次试验的得率均值为 91.73，样本方差为 3.89；新催化剂 30 次试验的得率均值为 93.75，样本方差为 4.02. 假设两总体正态且独立，不假设方差相等，求两总体均值差的 95% 置信区间.

Solution.

构造标准正态变量

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

用样本方差替代总体方差后，置信区间变为

$$\mu_1 - \mu_2 \in (\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

那么置信区间为

$$(-3.1434,\ -0.8966)$$

Exercise 5.

李雷和韩梅梅开始约会，韩梅梅迟到时间 $X \sim \mathrm{U}[0,\theta]$，先验 $\theta \sim \mathrm{U}[0,1]$（小时）. 若观测到 $X = x$，求更新后的 θ 的后验分布.

Solution.

观测模型为

$$f(x \mid \theta) = \begin{cases} \dfrac{1}{\theta}, & 0 < x < \theta, \\ 0, & \text{otherwise.} \end{cases}$$

先验为

$$\pi(\theta) = \begin{cases} 1, & 0 < \theta < 1, \\ 0, & \text{otherwise} \end{cases}$$

后验密度（未标准化）

$$f(\theta \mid x) \propto \frac{1}{\theta} \cdot \mathbf{1}_{x<\theta<1}$$

归一化常数

$$C = \int_x^1 \frac{1}{\theta}\, d\theta = -\ln x$$

因此

$$f(\theta \mid x) = \begin{cases} -\dfrac{1}{\theta \cdot \ln x}, & x < \theta < 1, \\ 0, & \text{otherwise} \end{cases}$$

Exercise 6.

计算硬币正面朝上概率 θ 的后验众数估计（MAP）. 并给出 $n = 20, x = 13$ 时的具体值，讨论其与 MLE 的关系.

Solution.

先验密度 $f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$，似然 $\theta^x(1-\theta)^{n-x}$. 后验 $\mathrm{Beta}(\alpha+x, \beta+n-x)$.
于是

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\alpha+x-1}{\alpha+\beta+n-2}$$

若采用非信息先验 $\alpha=\beta=1$，则 $\hat{\theta}_{\mathrm{MAP}} = \frac{x}{n} = \hat{\theta}_{\mathrm{MLE}}$. 当 $n=20, x=13$，MAP $= 13/20 = 0.65$.

Exercise 7.

假设总体 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，$\sigma^2$ 已知，先验 $\mu \sim N(\mu_0, \sigma_0^2)$.
(1) 求 μ 的 MAP 估计；(2) 求 μ 的后验均值估计.

Solution.

似然函数

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

先验分布

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

由共轭性，后验分布仍为正态分布

$$\mu \mid \boldsymbol{X} \sim N\left(\mu_n, \sigma_n^2\right)$$

其中

$$\mu_n = \frac{n\sigma_0^2\bar{X} + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2}, \qquad \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}$$

(1) **MAP** 估计：即后验分布的众数，由于后验是正态分布，其众数就是均值，因此

$$\hat{\mu}_{\mathrm{MAP}} = \mu_n = \frac{n\sigma_0^2\bar{X} + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2}$$

(2) 后验均值估计：正态分布的期望就是其均值，因此：

$$\hat{\mu}_{\mathrm{Bayes}} = E[\mu \mid \boldsymbol{X}] = \mu_n = \frac{n\sigma_0^2\bar{X} + \sigma^2\mu_0}{n\sigma_0^2+\sigma^2}$$

Exercise 8.

设随机变量 X 的分布列 $P(X=k|\theta) = \theta(1-\theta)^k$, $k=0,1,2,\dots$，先验 $\theta \sim \mathrm{U}(0,1)$. 三次独立观测得 $x_1=2$, $x_2=3$, $x_3=5$.
(1) 求 θ 的后验密度；(2) 求后验均值估计.

Solution.

三次观测值的联合似然函数

$$L(\theta) = \prod_{i=1}^{3} \theta(1-\theta)^{x_i} = \theta^3(1-\theta)^{\sum x_i} = \theta^3(1-\theta)^{10}$$

先验为 $\pi(\theta)=1$ 在 $(0,1)$ 上均匀，故后验密度

$$f(\theta \mid x) \propto \theta^3(1-\theta)^{10}$$

归一化

$$f(\theta \mid x) = \frac{\theta^3(1-\theta)^{10}}{\text{Beta}(4,\ 11)}$$

Beta(α, β) 分布的均值为 $\dfrac{\alpha}{\alpha+\beta}$，代入后验参数

$$\mathbb{E}[\theta \mid x_1, x_2, x_3] = \frac{4}{15}$$

Exercise 9.
（Bayes 区间估计）总体 $X_i \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$，$\sigma^2$ 已知，$0 < \alpha < 1$.
(1) 先验 $\mu \sim N(\mu_0, \sigma_0^2)$，求使后验包含概率 $1-\alpha$ 且长度最短的区间；
(2) 当 $\sigma_0^2 \to \infty$ 时，估计 (1) 中区间的极限情况，并比较其与经典置信区间，并直观解释；
(3) 若先验 $f(\mu) \propto 1$，分别给出 μ 的后验分布，MAP 区间和等尾可信区间，并与经典方法比较.

Solution.

(1) 似然函数 $\bar{X} \sim N(\mu, \sigma^2/n)$；先验 $\mu \sim N(\mu_0, \sigma_0^2)$
由共轭性，后验分布仍为正态

$$\mu \mid \boldsymbol{X} \sim N\left(\mu_n, \sigma_n^2\right)$$

其中

$$\mu_n = \frac{n\sigma_0^2\bar{X} + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2}, \qquad \sigma_n^2 = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

因为后验为正态分布，长度最短的 $1-\alpha$ 区间即为对称区间

$$\left[\mu_n - z_{1-\alpha/2}\cdot\sigma_n,\ \mu_n + z_{1-\alpha/2}\cdot\sigma_n\right]$$

(2) 当 $\sigma_0^2 \to \infty$，$\mu_n \to \bar{X}, \sigma_n^2 \to \dfrac{\sigma^2}{n}$, 所以后验分布变为

$$\mu \mid \boldsymbol{X} \sim N\left(\bar{X},\ \frac{\sigma^2}{n}\right)$$

于是 (1) 中区间退化为：

$$\left[\bar{X} - z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right]$$

这恰好是经典的 $1-\alpha$ 置信区间.
先验变得非常“平”，它对后验几乎没有影响.

(3) 若先验 $f(\mu) \propto 1$，这相当于 $\sigma_0^2 \to \infty$，即一个无信息先验，等价于第 (2) 问的极限情形.
因此此时后验分布为：

$$\mu \mid \boldsymbol{X} \sim N\left(\bar{X},\ \frac{\sigma^2}{n}\right)$$

MAP 与等尾可信区间均为经典区间

$$\left[\bar{X} - z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right]$$

Exercise 10.
（计算机实验续）利用作业 10-11 的结果给出 $\theta = e^{\mu}$ 的 95% 置信区间估计，并讨论其他构造置信区间的方法.

Solution.
代码如下：

```
from random import normalvariate as norm, choice
from math import exp, sqrt, log

def data_set(n: int):
    data = []
    observation = []
    for i in range(0, n):
        data += [norm(5, 1)]
    for i in range(0, n):
        observation += [choice(data)]
    return exp(sum(observation) / n)

theta = []
for i in range(0, 1000):
    theta += [data_set(100)]
mean = sum(theta)/1000
alpha = exp(1/200)
print("1:", [mean/(alpha+1.96*sqrt(alpha*(alpha-1)/1000)), mean/(alpha-1.96*sqrt(alpha*(
x = sum([log(data_set(100)) for i in range(0, 1000)]) / 1000
print("2:", [exp(x - 1.96 * (10**(-2.5))), exp(x + 1.96 * (10**(-2.5)))])
```

输出结果：

$$1 : [148.80640899688925, 150.11485329551672]$$

$$2 : [147.25682892324858, 149.09360441846417]$$

若先得到：

$$\mu \in \left[\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \ \bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

对两端取指数变换即可得到 θ 的置信区间：

$$\theta \in \left[e^{\bar{X} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}, \ e^{\bar{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}\right]$$