Homework 9

抹茶奶绿

April 22, 2025

Exercise 1.

尝试以简要框架形式给出概率部分知识的总结，并指出自己掌握起来相对困难的知识点.

Solution.

- 概率公理及基本性质：条件概率、全概率公式、贝叶斯公式.
- 随机变量及分布函数：常见离散分布和连续分布，联合分布.
- 数学期望与方差：定义、性质、矩母函数及应用.
- 大数定律与中心极限定理：概率不等式，大数定律、中心极限定理.

Exercise 2.

给出一个抽样调查实例，试指出你认为的其可能的不当之处.

Solution.

实例：在商场中对顾客进行消费满意度调查，通过在周末下午在人流量最大的地方随机发放问卷.

可能不当之处：

- 抽样框不全：仅在周末和特定位置发放，可能忽略工作日或商场其他区域的顾客，样本代表性不足.
- 自愿响应偏倚：部分顾客主动参与意愿更强，可能造成回答倾向性.
- 时间窗口偏倚：特定时段（下午）无法覆盖早晚时间段顾客特征.
- 无放回约束：未记录已调查人员，可能重复调查同一人，破坏独立性.

Exercise 3.

设总体的大小为 N，总体均值和方差分别为 μ, σ^2，X_i $(i=1,\dots,n)$ 为无放回抽取的简单随机样本.

I 证明：$E[X_i]=\mu$, $\mathrm{Var}(X_i)=\sigma^2$.

II 证明：$E[\bar{X}]=\mu$, $\mathrm{Var}(\bar{X})=\dfrac{\sigma^2}{n}\dfrac{N-n}{N-1}$.

Solution.

I 首先，X_i 的分布为：

$$P(X_i=x_r)=\frac{1}{N},\quad r=1,2,\dots,N.$$

因此

$$E[X_i]=\sum_{r=1}^{N}x_r\,P(X_i=x_r)=\frac{1}{N}\sum_{r=1}^{N}x_r=\mu$$

$$E[X_i^2]=\frac{1}{N}\sum_{r=1}^{N}x_r^2=\mu^2+\sigma^2$$

故

$$\mathrm{Var}(X_i)=E[X_i^2]-\big(E[X_i]\big)^2=\sigma^2$$

II 令 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, 则

$$E[\bar{X}]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]=\mu$$

对于 $i \neq j$ 的协方差，有

$$P(X_i = x_r, X_j = x_s) = \begin{cases} \frac{1}{N(N-1)}, & r \neq s \\ 0, & r = s \end{cases}$$

所以

$$E[X_iX_j] = \sum_{r\neq s} x_r x_s \frac{1}{N(N-1)} = \frac{1}{N(N-1)}\Big(\sum_{r,s} x_r x_s - \sum_r x_r^2\Big) = \frac{N\mu^2 - \sum_r x_r^2}{N(N-1)}$$

于是

$$\mathrm{Cov}(X_i, X_j) = E[X_iX_j] - \mu^2 = \frac{N\mu^2 - (\mu^2+\sigma^2)N}{N(N-1)} = -\frac{\sigma^2}{N-1}$$

那么

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\Big(\sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{1\le i<j\le n}\mathrm{Cov}(X_i,X_j)\Big) \\ &= \frac{1}{n^2}\Big(n\sigma^2 + n(n-1)(-\sigma^2 N - 1)\Big) = \frac{\sigma^2}{n}\frac{N-n}{N-1}\end{aligned}$$

Exercise 4.

设随机样本 X_i $(i = 1, \dots, n)$ 来自二项总体 $\mathrm{B}(k, p)$.

I 给出参数 k 和 p 的矩估计.

II 讨论上述估计的不足之处.

Solution.

I 样本一阶矩 $\mu_1 = \bar{X}$, 二阶中心矩 $m_2 = \dfrac{1}{n}\sum(X_i - \bar{X})^2$.

而对于二项分布 $E[X] = kp$, $\mathrm{Var}(X) = kp(1-p)$.

矩估计通过

$$\mu_1 = kp, \quad m_2 = kp(1-p)$$

即

$$\hat{p} = 1 - \frac{m_2}{\mu_1}, \qquad \hat{k} = \frac{\mu_1^2}{\mu_1 - m_2}$$

II 不足之处：

- 易受极端值影响.

Exercise 5.

设随机样本 X_i $(i = 1, \dots, n)$ 来自均匀分布 $U(\theta, 2\theta)$，求 θ 的矩估计和极大似然估计.

Solution.

MOM：$E[X] = \dfrac{3}{2}\theta \approx \mu_1$，则 $\hat{\theta}_{\text{MOM}} = \dfrac{2}{3}\mu_1$.

MLE：$L(\theta) = \frac{1}{\theta^n}, \theta \le x_i \le 2\theta$，似然在 θ 最小时取最大，故 $\hat{\theta}_{\text{MLE}} = \max_i X_i/2$.

Exercise 6.
设总体概率密度函数

$$f(x;a,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\cdot\frac{(x-a)^2}{\sigma^2}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right),\quad x\in\mathbb{R}$$

其中 $a\in\mathbb{R}$, $\sigma>0$ 为参数.

I 验证 $f(x;a,\sigma)$ 的归一性;

II 设样本 X_i $(i=1,\ldots,n)$ 来自此总体，求 a 和 σ^2 的矩估计;

III 列出 a,σ^2 的极大似然估计所满足的方程，并指出一种迭代求解方法.

I 变换变量 $x=a+\sigma y$ 得

$$\int_{\mathbb{R}} f(x;a,\sigma)\,dx=\int_{\mathbb{R}}\frac{1}{\sigma\sqrt{2\pi}}\cdot\frac{(\sigma y)^2}{\sigma^2}e^{-y^2/2}\cdot\sigma\,dy=\int_{\mathbb{R}}\frac{y^2}{\sqrt{2\pi}}e^{-y^2/2}dy \xlongequal{Gaussian} 1$$

II 类似地，我们继续利用 $Gaussian$ 可得

$$E[X]=a,\quad Var(X)=3\sigma^2$$

那么

$$\hat{a}_{\text{MOM}}=\mu_1,\quad \hat{\sigma}^2_{\text{MOM}}=\frac{m_2}{3}$$

III 写出对数似然函数

$$\ell(a,\sigma^2)=-3n\ln\sigma+2\sum_{i=1}^n\ln|X_i-a|-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-a)^2+const$$

一阶偏导为

$$\frac{\partial\ell}{\partial\sigma^2}=-\frac{3n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum(X_i-a)^2=0\Rightarrow\hat{\sigma}^2=\frac{1}{3n}\sum(X_i-a)^2$$

$$\frac{\partial\ell}{\partial a}=-2\sum\frac{1}{X_i-a}+\frac{1}{\sigma^2}\sum(X_i-a)=0$$

初始令 $a^{(0)}=\bar{X}$, $\sigma^{2(0)}=\hat{\sigma}^2_{\text{MOM}}$，更新为

$$a^{(t+1)}=a^{(t)}-\left.\frac{\partial\ell/\partial a}{\partial^2\ell/\partial a^2}\right|_{a=a^{(t)}},\quad \sigma^{2(t+1)}=\frac{1}{3n}\sum_{i=1}^n(X_i-a^{(t+1)})^2$$

迭代即得极大似然估计.

Exercise 7.
设随机样本 X_i 来自 $Bernoulli$ 总体 $\mathrm{B}(p)$，请给出参数 p 的矩估计和极大似然估计.

Solution.

I MOM:

$$E[X]=p\approx\mu_1=\frac{1}{n}\sum_{i=1}^n X_i$$

即

$$\hat{p}_{\text{MOM}}=\mu_1$$

II MLE:

$$L(p)=\prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}=p^{\sum X_i}(1-p)^{n-\sum X_i}$$

取对

$$\ell(p)=\sum X_i \ln p+(n-\sum X_i)\ln(1-p)$$

求导

$$\frac{d\ell}{dp}=\frac{\sum X_i}{p}-\frac{n-\sum X_i}{1-p}=0$$

那么

$$\hat{p}_{\text{MLE}}=\frac{1}{n}\sum_{i=1}^{n} X_i=\mu_1$$

因此，矩估计与极大似然估计在此问题中完全一致：

$$\hat{p}_{\text{MOM}}=\hat{p}_{\text{MLE}}=\mu_1$$

Exercise 8.
设总体是 m 项多项分布，总数为 n，单元概率 p_i，样本频数 $X_i(i=1,\dots,m)$，求参数 p_i 的极大似然估计.

Solution.
多项分布似然

$$L\propto\prod_{i=1}^{m} p_i^{X_i},\quad \sum p_i=1.$$

$Lagrange$ 乘子易见

$$\frac{X_i}{p_i}=\frac{X_j}{p_j}$$

于是

$$\hat{p}_i=\frac{X_i}{n},\quad i=1,\dots,m.$$

Exercise 9.
设总体 X 的分布如下：

X	1	2	3
P	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

其中 $0<\theta<1$ 是未知参数. 已取样本 $x_1=1$, $x_2=2$, $x_3=3$，求 θ 的矩估计和极大似然估计.

Solution.
MOM：

$$\mu_1=\bar{X}=(1+2+3)/3=2,\quad E[X]=1\cdot\theta^2+2\cdot 2\theta(1-\theta)+3\cdot(1-\theta)^2=3-2\theta$$

那么

$$3-2\theta=2\ \Rightarrow\ \hat{\theta}_{\text{MOM}}=\frac{1}{2}$$

MLE：

$$L=\theta^{2\cdot 1}[2\theta(1-\theta)]^1(1-\theta)^{2\cdot 1}=2\,\theta^3(1-\theta)^3$$

求导得 $3/\theta-3/(1-\theta)=0$，即 $\hat{\theta}_{\text{MLE}}=1/2$.

Exercise 10.

设随机样本 $X_1, \ldots, X_n$ 来自

$$f(x) = \theta\, x^{\theta-1}, \quad 0 < x < 1,\ \theta > 0.$$

1. 求 θ 的矩估计 $\hat{\theta}_{\text{MOM}}$.
2. 求极大似然估计 $\hat{\theta}_{\text{MLE}}$.

Solution.

I $E[X] = \dfrac{\theta}{\theta+1} \approx \mu_1$，故 $\hat{\theta}_{\text{MOM}} = \dfrac{\mu_1}{1-\mu_1}$.

II 似然函数

$$L = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$

求导得 $n/\theta + \sum \ln x_i = 0$，即

$$\hat{\theta}_{\text{MLE}} = -\frac{n}{\sum_{i=1}^{n} \ln x_i}$$

Exercise 11.

（计算机实验）考虑第 4 题，分别取 $k = 10, p = 0.01$ 与 $k = 10, p = 0.5$，样本容量 $n = 10, 1000$，生成 $\text{B}(k, p)$ 样本，给出 k, p 的矩估计值. 多次尝试，观察是否有不合理结果？

Solution.

代码如下：

```python
import random
import numpy as np

def experiment(k, p, n):
    data = np.array([sum(1 for _ in range(k) if random.random() < p) for _ in range(n)])
    a_1 = data.mean()
    m_2 = data.var()
    if a_1 == 0:
        return 0, 0
    return (1 - m_2 / a_1), a_1 / (1 - m_2 / a_1)

print("\nBinomial: k = 10, p = 0.01, n = 1000")
print("Moment Estimation Result\n    p     k    ")
for i in range(0, 10):
    print("%d %.4f %.4f" % ((i,) + experiment(10, 0.01, 1000)))
```

模拟结果如下：

```
Binomial: k = 10, p = 0.5, n = 1000
Moment Estimation Result
    p     k
0 0.530972 9.399743
1 0.502381 9.914781
2 0.460832 10.821732
3 0.503798 10.045696
4 0.480193 10.408318
5 0.483703 10.369992
6 0.524598 9.506331
7 0.513090 9.696146
8 0.505161 9.818646
9 0.502389 9.715578
```

```
Binomial: k = 10, p = 0.5, n = 10
Moment Estimation Result
    p     k
0 0.629412 8.102804
1 0.591667 8.112676
2 0.400000 12.500000
3 0.657447 7.148867
4 0.188889 23.823529
5 0.814286 6.877193
6 0.716667 6.697674
7 0.455556 9.878049
8 0.771698 6.867971
9 0.340678 17.318408
```

```
Binomial: k = 10, p = 0.01, n = 1000
Moment Estimation Result
    p     k
0 0.030528 3.472188
1 0.084000 1.000000
2 0.049989 1.860400
3 0.000909 121.000000
4 -0.005150 -20.778584
5 0.011660 9.090615
6 0.045556 1.975610
7 0.033500 2.865672
8 -0.024449 -4.008347
9 0.080000 1.000000
```

```
Binomial: k = 10, p = 0.01, n = 10
Moment Estimation Result
    p     k
0 0.2000 1.0000
1 0.1000 1.0000
2 0.1000 1.0000
3 0.0000 0.0000
4 0.2000 1.0000
5 0.2000 1.0000
6 0.1000 1.0000
7 0.1000 1.0000
8 0.1000 1.0000
9 0.1000 1.0000
```

可以看出样本容量较小时容易出现不合理估计.