多元统计小测复盘

Yooki

November 2021

第一章　预备知识

1.1 Categorical Distribution

categorical 分布也称广义伯努利分布、multinoulli 分布;

1. Parameter:
   - k: number of categories(integer)
   - $p_1, \cdots, p_k$: event probabilities
   - $p_i \geq 0, \sum p_i = 1$
2. Support:
   - $x \in \{1, \cdots, k\}$
3. PMF
   - $p(x = i) = p_i$
   - $p(x) = p_1^{[x=1]} \cdots p_k^{[x=k]}$
   - $p(x) = [x = 1] \cdot p_1 + \cdots + [x = k] \cdot p_k, [x = i]\ \textit{is the Iverson bracket}$
   - $[x = i] = \mathbf{I}_{\{x=i\}}(x)$

MLE

设总体分布 X 为 categorical 分布 (K 类)，样本为 x_n，样本容量为 N

$$P(x_n|p_1, \cdots, p_k) = \prod_{k=1}^{K} p_k^{[x_n=k]} \tag{1.1}$$

Likelihood function：

$$L(\mathbf{x}|\mathbf{p}) = \prod_{n=1}^{N} P(x_n|\mathbf{p}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p_k^{[x_n=k]} = \prod_{k=1}^{K} p_k^{\sum_{n=1}^{N}[x_n=k]} \tag{1.2}$$

$$lnL(\mathbf{x}|\mathbf{p}) = \sum_{k=1}^{K}(\sum_{n=1}^{N}[x_n = k]) \cdot ln(p_k) \tag{1.3}$$

Larange Multiplier:

$$\Lambda(\mathbf{p}, \lambda) = lnL(\mathbf{x}|\mathbf{p}) + \lambda(1 - \sum_{k=1}^{K} p_k) \tag{1.4}$$

$$\frac{\partial lnL}{\partial p_k} = \frac{\sum_{n=1}^{N}[x_n = k]}{p_k} - \lambda = 0 \tag{1.5}$$

$$\Longrightarrow p_k = \frac{\sum_{n=1}^{N}[x_n = k]}{\lambda}$$

$$\because \sum_{k=1}^{K} p_k = 1 \therefore \lambda = \sum_{k=1}^{K}\sum_{n=1}^{N}[x_n = k] = \sum_{n=1}^{N}\sum_{k=1}^{K}[x_n = k] = N$$

$$\hat{p}_{k,MLE} = \frac{\sum_{n=1}^{N}[x_n = k]}{N} \tag{1.6}$$

1.2 Multinomial Distribution

MLE

设总体分布 X 为 Multinomial 分布 (K 类)，样本为 $\mathbf{x}=(x_1,\cdots,x_K)$，其中 x_i 为第 i 类出现的次数，样本容量为 1(对 $\tilde{N}$ 次独立试验的一次观测 | 一次多项分布试验)

$$L(\mathbf{x}|\mathbf{p})=P(\mathbf{x}|\mathbf{p})=P(x_1=N_1,\cdots,x_K=N_K|\mathbf{p})=\frac{\tilde{N}!}{N_1!\cdots N_K!}\prod_{k=1}^{K}p_k^{N_k} \tag{1.7}$$

$$\sum_{k=1}^{K}p_k=1 \tag{1.8}$$

$$\sum_{k=1}^{K}N_k=\tilde{N} \tag{1.9}$$

注意到 $L(\mathbf{x}|\mathbf{p})$ 与 categorical 分布关于 $\mathbf{p}$ 的部分 (kernel) 是类似的，故可做与 (1.1) 节类似的处理；

通过简单的计算，我们得到

$$\hat{p}_{k,MLE}=\frac{N_k}{\tilde{N}} \tag{1.10}$$

1.3 中心极限定理

《概率论基础第二版》(李贤平)

设 $\xi_1,\cdots,\xi_n,\cdots$ 是一个相互独立的随机变量序列，他们具有有限的数学期望和方差：

$$E\xi_k,\ D\xi_k \quad (k=1,\cdots,n,\cdots) \tag{1.11}$$

$$\zeta_n=\sum_{k=1}^{n}\frac{\xi_k-E\xi_k}{D\xi_k}\ \dot{\sim}\ N(0,1) \tag{1.12}$$

第二章　第二题回顾

2.1　问题

为了解决某地大学毕业生的去向问题，对来自该地多所大学的 1000 名学生组成的随机样本进行了调查，搜集的数据如 Table1 所示。令 $p1, p2, p3$ 分别表示该地区大学毕业生“继续深造”、“就业”、“其他去向”的概率;

1. 根据调查背景，确定样本的类型，容量和总体分布
2. 寻找参数 $\mathbf{p} = (p_1, p_2, p_3)^T$ 的估计量 $\hat{\mathbf{p}}$
3. 根据中心极限定理写出 $\hat{\mathbf{p}}$ 的近似分布

2.2　解答

2.2.1　问题 1

1. 角度一:
   - 总体分布：categorical Distribution
   - 样本：x_n 学生 n 的毕业选择 x_n
   - 类别：$K = 3$，三种毕业选择，编码为 {1:” 继续深造”, 2:” 就业”, 3:” 其它”}
   - 样本容量：$N = 1000$
2. 角度二：

- 总体分布：Multinomial Distribution
- 样本：$\mathbf{x} = (x_1, \cdots, x_K)$ 对 N 个学生的毕业选择的一次调查
- x_i：类别 i 在一次观测中出现的次数
- 类别：$K = 3$
- 样本容量：1

2.2.2 问题 2

从极大似然估计的角度出发：

1. 角度一

   由第一章的预备知识1.6，我们得到

$$\hat{p}_{k,MLE} = \frac{\sum_{n=1}^{N}[x_n = k]}{N}, \quad k = 1, 2, 3; N = 1000 \tag{2.1}$$

2. 角度二

   由第一章的预备知识1.10，我们得到

$$\hat{p}_{k,MLE} = \frac{N_k}{\tilde{N}}, \quad k = 1, 2, 3; N = 1000 \tag{2.2}$$

2.2.3 问题 3

易知，固定 k 时有

$$[x_n = k] = \mathbf{I}_{\{x_n=k\}}(x) \overset{i.i.d}{\sim} b(1, p_k), \quad n = 1, \cdots, N \tag{2.3}$$

由2.1知，$\hat{p}_k$ 由 N 个独立同分布的随机变量序列构成；根据中心极限定理1.12知

$$E(\hat{p}_k) = p_k, \quad D(\hat{p}_k) = \frac{p_k(1-p_k)}{N} \tag{2.4}$$

$$\hat{p}_k \dot{\sim} N(p_k, \frac{p_k(1-p_k)}{N}) \tag{2.5}$$

疑惑：$\because [x_n = k]$ 与 $[x_n = j]$ 之间不独立，$\therefore \hat{p}_k$ 与 $\hat{p}_j$ 之间不独立，如何使用中心极限定理做出 $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3)^T$ 的近似分布？