2021 年秋统计学习题 07 参考解答

1. 独立投掷硬币 10 次，记硬币正面朝上的概率为 p，假设检验问题为 $H_0: p = \frac{1}{2} \leftrightarrow p \neq \frac{1}{2}$. 如果正面出现 0 次或 10 次时拒绝检验原假设.
   (a). 检验的显著性水平是多少?
   (b). 如果 $p = 0.1$，检验的势是多少?

解 记 X 为 10 次投掷中硬币正面朝上的次数.

拒绝域 $C = \{0, 10\}$：当试验结果，即 10 次投掷中正面次数 $x \in C$ 时，拒绝原假设.

(a) 显著性水平为拒真的概率：

$$\alpha = \mathbb{P}(X \in C \mid p = \tfrac{1}{2})$$

$$= \mathbb{P}(X = 0 \text{ 或 } X = 10 \mid p = \tfrac{1}{2})$$

$$= \mathbb{P}(X = 0 \mid p = \tfrac{1}{2}) + \mathbb{P}(X = 10 \mid p = \tfrac{1}{2})$$

$$= \left(\tfrac{1}{2}\right)^{10} \times 2 = \frac{1}{2^9}.$$

(b) 检验的势为拒假的概率：

$$p(0.1) = \mathbb{P}(X \in C \mid p = 0.1)$$

$$= \mathbb{P}(X = 0 \mid p = 0.1) + \mathbb{P}(X = 10 \mid p = 0.1)$$

$$= \binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{10}(0.1)^{10}(0.9)^0$$

$$= 0.9^{10} + 0.1^{10}.$$

2. 设 $X_1, \cdots, X_n$ 为来自参数为 λ 的 Poisson 分布的样本.
   (a). 考虑检验问题 $H_0\colon \lambda = \lambda_0 \leftrightarrow H_1\colon \lambda = \lambda_1$（$\lambda_0 < \lambda_1$）. 求检验的似然比.
   (b). 利用 Poisson 分布的可加性解释如何确定上述假设检验问题的显著性水平为 α 的拒绝域.
   (c). 证明对于假设 $H_0\colon \lambda = \lambda_0 \leftrightarrow H_1\colon \lambda > \lambda_0$，上述检验是一致最优势的.

解 (a) 似然函数为

$$L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

似然比为

$$\Lambda(X) = \frac{L(\lambda_0)}{L(\lambda_1)} = \frac{\prod_{i=1}^{n} \lambda_0^{x_i} e^{-\lambda_0}}{\prod_{i=1}^{n} \lambda_1^{x_i} e^{-\lambda_1}} = \frac{\lambda_0^{\sum_{i=1}^{n} x_i}}{\lambda_1^{\sum_{i=1}^{n} x_i}} e^{\lambda_1-\lambda_0}$$

小写

$$= \left(\frac{\lambda_0}{\lambda_1}\right)^{\sum_{i=1}^{n} x_i} e^{\lambda_1-\lambda_0}$$

(b) 因 $\lambda_0<\lambda_1$, 似然比检验的拒绝域

$$\{x: \Lambda(x) \le k\}$$

等价于

$$\{x: \sum_{i=1}^{n} x_i \ge c\}$$

其中 c 满足

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \ge c \,\middle|\, \lambda=\lambda_0\right) = \alpha$$

其中 $\sum_{i=1}^{n} X_i$ 服从参数为 $n\lambda$ 的 Poisson 分布.

c) 由Neyman-Pearson引理，上述检验对任一检验问题

$$H_0: \lambda=\lambda_0 \longleftrightarrow H_1: \lambda=\lambda_1,$$

其中$\lambda_1>\lambda_0$，都是最优势检验，且拒绝域与λ_1无关，因而是一致最优势检验。

3. 设样本 $X_1, X_2, \cdots, X_n$ 来自总体 $X \sim B(1,\theta)$. 现在有假设检验

$$H_0: \theta = 0.48 \leftrightarrow H_1: \theta = 0.52.$$

检验方法如下：若 $\sum_{i=1}^n x_i$ 较大则拒绝 H_0. 利用中心极限定理进行近似计算，求样本量应该至少为多大才能使两类错误发生的概率都约为 0.01.

解 由中心极限定理,

$$Z := \frac{\sum_{i=1}^n X_i - n\theta}{\sqrt{n\theta(1-\theta)}}$$

近似服从标准正态分布. 设拒绝域为

$$\mathbb{C} = \{x : \sum_{i=1}^n x_i > c\}, \quad c \text{ 为常数},$$

则第一类错误发生的概率为

$$\mathbb{P}(X \in C \mid \theta = 0.48) \approx 0.01,$$

第二类错误发生的概率为

$$1 - \mathbb{P}(X \in C \mid \theta = 0.52) \approx 0.01,$$

即

$$\mathbb{P}\left(Z > \frac{c - n\times 0.48}{\sqrt{n\times 0.48\times 0.52}}\right) \approx 0.01,$$

$$\mathbb{P}\left(Z > \frac{c - n\times 0.52}{\sqrt{n\times 0.52\times 0.48}}\right) \approx 0.99,$$

由此可知

$$\frac{c-n\times 0.48}{\sqrt{n\times 0.48\times 0.52}}\approx 2.33 \qquad (1)$$

$$\frac{c-n\times 0.52}{\sqrt{n\times 0.52\times 0.48}}=-2.33 \qquad (2)$$

(1), (2) 两式相加可得

$$2c=n. \qquad (3)$$

(1), (2) 两式相减可得

$$\frac{n\times 0.04}{\sqrt{n\times 0.52\times 0.48}}=4.66,$$

即

$$\boxed{n=\frac{4.66^2\times 0.52\times 0.48}{0.04^2}\approx 3388.}$$

再由(3)可得

$$c\approx 1694.$$

注：本题仅要求计算n，只需(1)，(2)两式相减即可.

4. 设 $X_1, X_2, \cdots, X_n$ 为来自 $\mathrm{Beta}(\mu, 1)$ 的样本，$Y_1, Y_2, \cdots, Y_m$ 为来自 $\mathrm{Beta}(\theta, 1)$ 的样本，且假设这两个样本之间也相互独立.

(a). 求 $H_0: \theta = \mu \leftrightarrow H_1: \theta \neq \mu$ 的广义似然比检验.

(b). 证明上述检验可以基于如下统计量

$$T = \frac{\sum_{i=1}^n \log X_i}{\sum_{i=1}^n \log X_i + \sum_{i=1}^m \log Y_i}.$$

解 设 $X \sim \mathrm{Beta}(\mu, 1)$，则它的概率密度函数为

$$f(x|\mu) = \frac{\Gamma(\mu+1)}{\Gamma(\mu)\cdot\Gamma(1)} x^{\mu-1}(1-x)^{1-1} \mathbb{1}_{(0,1)}(x)$$

$$= \mu x^{\mu-1} \mathbb{1}_{(0,1)}(x),$$

类似地，设 $Y \sim \mathrm{Beta}(\theta, 1)$，则 Y 的概率密度函数为

$$f(y|\theta) = \theta y^{\theta-1} \mathbb{1}_{(0,1)}(y).$$

大写正体

用 X，Y 分别表示样本 $(X_1, \cdots, X_n)$，$(Y_1, \cdots, Y_m)$，

小写，正体

x，y 分别表示 X，Y 对应的样本点.

样本的似然函数为（设 $x_i, y_i \in (0,1)$）：

$$L(\mu, \theta | \mathrm{x}, \mathrm{y}) = \prod_{i=1}^n \mu x_i^{\mu-1} \prod_{i=1}^m \theta y_i^{\theta-1}$$

$$= \mu^n \theta^m \left(\prod_{i=1}^n x_i\right)^{\mu-1} \left(\prod_{i=1}^m y_i\right)^{\theta-1}.$$

对数似然函数为 $(\mu, \theta) \in \Theta := \{(\mu, \theta): \mu, \theta \in \mathbb{R}\}$.

$$l(\mu, \theta | \mathrm{x}, \mathrm{y}) = \log L(\mu, \theta | \mathrm{X}, \mathrm{Y})$$

$$= n \log \mu + m \log \theta + (\mu-1)\sum_{i=1}^n \log x_i + (\theta-1)\sum_{i=1}^m \log y_i$$

分别对μ, θ求导并令导数为零可得

$$\frac{\partial l}{\partial \mu} = \frac{n}{\mu} + \sum_{i=1}^{n} \log x_i = 0,$$

$$\frac{\partial l}{\partial \theta} = \frac{m}{\theta} + \sum_{i=1}^{m} \log y_i = 0,$$

由此可得μ, θ的MLE如下：

$$\hat{\mu} = -\frac{n}{\sum_{i=1}^{n} \log x_i}, \quad \hat{\theta} = -\frac{m}{\sum_{i=1}^{n} \log y_i}.$$

原假设H_0可表示为

$$(\mu, \theta) \in \Theta_0 := \{(\mu, \mu) : \mu \in \mathbb{R}\}.$$

在H_0下，似然函数为

$$L(\mu, \mu \mid X, Y) = \mu^{n+m} \left(\prod_{i=1}^{n} x_i \prod_{j=1}^{m} y_i \right)^{\mu - 1}$$

它的对数似然函数为

$$\begin{aligned} l(\mu, \mu \mid X, Y) &= (n+m) \log \mu + (\mu - 1) \log \prod_{i=1}^{n} x_i \prod_{i=1}^{m} y_i \\ &= (n+m) \log \mu + (\mu - 1) \left[\sum_{i=1}^{n} \log x_i + \sum_{i=1}^{m} \log y_i \right]. \end{aligned}$$

由

$$0 = \frac{\partial l(\mu, \mu \mid X, Y)}{\partial \mu} = \frac{n+m}{\mu} + \sum_{i=1}^{n} \log x_i + \sum_{i=1}^{m} \log y_i$$

所以在 $(\mu,\theta)\in\Theta_0$ 条件下，$\mu=\theta$ 的 MLE 为

$$\hat{\mu}_0 = -\frac{n+m}{\sum_{i=1}^{n}\log x_i + \sum_{i=1}^{m}\log y_i}$$

备择假设可表示为

$$(\mu,\theta)\in\Theta_1=\{(\mu,\theta):\ \mu\in\mathbb{R},\ \theta\in\mathbb{R},\ \mu\neq\theta\}.$$

于是 $\Theta=\Theta_0\cup\Theta_1$. 广义似然比为

$$\Lambda(x,y)=\frac{\max_{(\mu,\theta)\in\Theta_0} L(\mu,\theta\mid x,y)}{\max_{(\mu,\theta)\in\Theta} L(\mu,\theta\mid x,y)}$$

$$=\frac{L(\hat{\mu}_0,\hat{\mu}_0\mid x,y)}{L(\hat{\mu},\hat{\theta}\mid x,y)}$$

$$=\frac{\hat{\mu}_0^{\,n+m}}{\hat{\mu}^{\,n}\,\hat{\theta}^{\,m}}\left(\prod_{i=1}^{n}x_i\right)^{\hat{\mu}_0-\hat{\mu}}\left(\prod_{i=1}^{m}y_i\right)^{\hat{\mu}_0-\hat{\theta}}.$$

为方便起见，记

$$a=\sum_{i=1}^{n}\log x_i,\qquad b=\sum_{i=1}^{m}\log b_i,$$

则

$$\hat{\mu}=-\frac{n}{a},\quad \hat{\theta}=-\frac{m}{b},\quad \hat{\mu}_0=-\frac{n+m}{a+b}.$$

$$\log\left[\left(\prod_{i=1}^{n} x_i\right)^{\hat{\mu}_0-\hat{\mu}}\left(\prod_{i=1}^{m} y_i\right)^{\hat{\mu}_0-\hat{\theta}}\right]$$

$$= (\hat{\mu}_0-\hat{\mu})\, a + (\hat{\mu}_0-\hat{\theta})\, b$$

$$= \hat{\mu}_0 (a+b) - a\hat{\mu} - b\hat{\theta}$$

$$= -(n+m) + n + m$$

$$= 0 .$$

这表明

$$\left(\prod_{i=1}^{n} x_i\right)^{\hat{\mu}_0-\hat{\mu}}\left(\prod_{i=1}^{m} y_i\right)^{\hat{\mu}_0-\hat{\theta}} = 1.$$

由此可知

$$\Lambda(X,Y) = \frac{\hat{\mu}_0^{\,n+m}}{\hat{\mu}^{n}\cdot\hat{\theta}^{m}} = \frac{\left(\frac{-(n+m)}{a+b}\right)^{n+m}}{\left(\frac{-n}{a}\right)^{n}\left(\frac{-m}{b}\right)^{m}}$$

$$= \left(\frac{n+m}{n}\right)^{n}\left(\frac{n+m}{m}\right)^{m}\left(\frac{a}{a+b}\right)^{n}\left(\frac{b}{a+b}\right)^{m}$$

$$= \left(\frac{n+m}{n}\right)^{n}\left(\frac{n+m}{m}\right)^{m}\left(\frac{a}{a+b}\right)^{n}\left(1-\frac{a}{a+b}\right)^{m}.$$

故 $\Lambda(X,Y) = \left(\frac{n+m}{n}\right)^{n}\left(\frac{n+m}{m}\right)^{m} T^{n}(1-T)^{m}$

即广义似然比基于 T.

10

5. 假设基因频率是均衡的，基因型 AA,Aa,aa 出现的频率分比为 $(1-\theta)^2$，$2\theta(1-\theta)$，θ^2. Plato 等发表了如下 190 个人类样本中触珠蛋白型的数据：

Hp1-1	Hp1-2	Hp2-2
10	68	112

表 1 触珠蛋白型

请验证遗传模型的拟合优度.

解 按 John A. Rice "Math. Stat. & Data Analysis" §8.5.1 P272-274 中的计算. 取 $n=190$, X_1, X_2, X_3 分别表示基因型 AA, Aa, aa 出现的频数, 对应的观察值为

$$x_1 = 10, \quad x_2 = 68, \quad x_3 = 112.$$

θ 的 mle 为

$$\hat{\theta} = \frac{x_2 + 2x_3}{2(x_1 + x_2 + x_3)} = \frac{68 + 2\times 112}{2(10+68+112)} = 0.7684$$

	AA	Aa	aa	求和
观察频数 O_i	10	68	112	n 190
概率 $\hat{p}_i$	$(1-\hat{\theta})^2$ 0.053623	$2\hat{\theta}(1-\hat{\theta})$ 0.35590	$\hat{\theta}^2$ 0.59047	1
期望频数 E_i	$n\hat{p}_i$ 10.1894	67.6211	112.1894	190
$(O_i-E_i)^2/E_i$	0.00352	0.00212	0.00032	0.00596

设3个单元格的概率分别为 p_1, p_2, p_3，则检验遗传模型的拟合优度相当于如下假设检验问题：

H_0: $p_1=(1-\theta)^2$, $p_2=2\theta(1-\theta)$, $p_3=\theta^2$, $\theta\in[0,1]$.

H_1: $p_1+p_2+p_3=1$, $p_1, p_2, p_3\in[0,1]$.

用参数空间表示即

H_0: $(p_1, p_2, p_3)\in\Theta_0$.

H_1: $(p_1, p_2, p_3)\in\Theta_1$,

其中

$$\Theta_0=\{(p_1,p_2,p_3): p_1=(1-\theta)^2, p_2=2\theta(1-\theta), p_3=\theta^2, \theta\in[0,1]\},$$

$$\Theta_1=\{(p_1,p_2,p_3): p_1+p_2+p_3=1, p_1,p_2,p_3\in[0,1]\}.$$

由此可知自由度 $df=\dim\Theta_1-\dim\Theta_0=(3-1)-1=1$.

由于 $\chi^2=\sum_{i=1}^{3}\frac{(O_i-E_i)^2}{E_i}\sim\chi^2(1)$, 上分位点

显著性水平为 $\alpha=0.05$ 的拒绝域为 $\chi^2>\chi^2_{0.05}(1)=3.84$.

由于观测值 $0.00596<3.84$，故接受 H_0，即认为模型拟合好. [由p值 $P(\chi^2(1)>0.00596)=0.9384$ 也可判断]

12

6. Bhattacharjee 等为研究在盈月期间动物是否更易咬人而收集了某医疗结构处理被动物咬伤的病例. 月亮周期被划分为 10 个阶段, 每个阶段被咬伤的人数如下表所示. 29 日是盈月. 咬伤事件是否有时间趋势?

日期	2,3,4	5,6,7	8,9,10	11,12,13	14,15
人数	155	142	146	148	110

日期	16,17,18	19,20,21	22,23,24	25,26,27	28,29,1
人数	137	150	163	201	269

解: 记各阶段被咬伤的人数

解: 记各阶段人被咬伤的概率分别为 $p_1, p_2, \cdots, p_{10}$, 并且简记 $p=(p_1, p_2, \cdots, p_{10})$, 则咬伤事件是否有时间趋势相当于检验如下假设问题(无时间趋势当作零假设):

H_0: $p \in \Theta_0 = \{p: p_1 = p_2 = p_3 = p_4 = p_6 = p_7 = \cdots = p_{10} = \frac{3}{29}, p_5 = \frac{2}{29}\}$.

H_1: $p \in \Theta_1 = \{p: p_1 + p_2 + \cdots + p_{10} = 1, p_i \in (0,1)\}$.

χ^2 检验的自由度为 $\dim \Theta_1 - \dim \Theta_0 = (10-1) - 0 = 9$.

$\chi^2_{0.05}(9) = 16.919 < 85.48$, 故拒绝原假设, 认为有时间趋势. 也可用 P 值 $P(\chi^2(9) > 85.48) = 1.31 \times 10^{-14}$ 来判断: P 值 < 0.05, 接受备择假设. 检验值 85.48 计算

日期	2, 3, 4	5, 6, 7	8, 9, 10	11, 12, 1	14, 15	16, 17, 1	19, 20, 2	22, 23, 24	25, 26, 2	28, 29, 1	求和
观察到的人数Oi	137	150	163	201	269	155	142	146	148	110	1621
概率pi	3/29	3/29	3/29	3/29	2/29	3/29	3/29	3/29	3/29	3/29	1
期望的人数Ei	167.7	167.7	167.7	167.7	167.7	167.7	167.7	167.7	167.7	111.8	1621
(O-E)^2/E	5.62	1.87	0.13	6.62	61.21	0.96	3.94	2.81	2.31	0.03	85.48

如上表.