STAT201B Lecture 12 Multinomial Tests

1 Tests for Multinomial Data

1.1 Definition: Multinomial distribution

令:

- random variable $Z \in \{1, \dots, k\}$
- parameter $p = (p_1, \dots, p_k)$, 其中 $p_j = \mathbb{P}(Z = j)$

若:

- $Z_1, \dots, Z_n$ 为一组 iid sample
- $X_j = \#\{Z_i : Z_i = j\}, \quad j = 1, \dots, k$ ($\sum_{j=1}^k X_j = n$)

则 $X = (X_1, \dots, X_k) \sim Multinomial(n, p)$, PMF 为

$$f(x_1, \dots, x_k; p) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Remark

- Binomial distribution 为 Multinomial distribution 的一个特例
- $Z \in \{1, \dots, k\}$ 表示 Z 有 k 个 labels 可以选择, 具体的 labels 可以不属于 $\{1, \dots, k\}$
- parameter p 是 $k-1$ **dimensional** 的, 因为 $\sum_{j=1}^k p_j = 1$
- p 的 MLE 为 $(\hat{p}_1, \dots, \hat{p}_k) = (X_1/n, \dots, X_k/n)$

Example

假设袋子里有三种颜色的球: 红色 ($Z = 1$), 黄色 ($Z = 2$), 蓝色 ($Z = 3$) , 抽中的概率分别为 p_1, p_2, p_3 ($p_1 + p_2 + p_3 = 1$)

从袋中有放回地抽 n 次球, 令 $Z_1, \dots, Z_n$ 表示这 n 次抽取的颜色, 令:

$$\begin{aligned} X_1 &= \#\text{ of red balls sampled} = \#\{Z_i : Z_i = 1\}, \\ X_2 &= \#\text{ of yellow balls sampled} = \#\{Z_i : Z_i = 2\}, \\ X_3 &= \#\text{ of blue balls sampled} = \#\{Z_i : Z_i = 3\}, \end{aligned}$$

则

$$(X_1, X_2, X_3) \sim Multinomial(n, p_1, p_2, p_3)$$

Logic

接下来我们考虑以下假设:

$$H_0 : (p_1, \dots, p_k) = (p_{01}, \dots, p_{0k}) := p_0 \quad v.s. \quad H_1 : (p_1, \dots, p_k) \neq (p_{01}, \dots, p_{0k})$$

1.2 方法一: likelihood ratio test

由于 Multinomial distribution 的 PMF 为:

$$f(x_1,\dots,x_k;p)=\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$$

因此, 令 MLE 为 $\hat{p}=(\hat{p}_1,\dots,\hat{p}_k)=(X_1/n,\dots,X_k/n)$, 则likelihood ratio test 的 test statistic 为:

$$T(X)=\frac{\mathcal{L}_n(\hat{p})}{\mathcal{L}(p_0)}=\prod_{j=1}^{k}\left(\frac{\hat{p}_j}{p_{0j}}\right)^{X_j}$$

使用 Wilk's theorem, 有

$$\lambda(X)=2\log T(X)=2\sum_{j=1}^{k}X_j\log\left(\frac{\hat{p}_j}{p_{0j}}\right)\xrightarrow{d}\chi^2_{k-1}$$

因此 the approximate size α LRT reject H_0 when $\lambda(X)\geq\chi^2_{k-1,\alpha}$

> ⚠ **Remark**
>
> 自由度为 $k-1$ 是由于 $dim(\Theta)=k-1$ 且 $dim(\Theta_0)=0$

> **Example**
>
> ***问题:***
> STAT201B Lecture 12 Multinomial-1.png
>
> ***解答:***
> 问题等价于以下 hypothesis testing problem:
>
> $$H_0:p_1=p_2=p_3=\frac{1}{3}\quad v.s.\quad H_1:\text{not }(p_1=p_2=p_3=\frac{1}{3})$$
>
> 因此, under H_0, 我们有:
>
> $$\lambda(X)=2\log T(X)=2\sum_{j=1}^{3}X_j\left(\log\left(\frac{3X_i}{n}\right)\right)\xrightarrow{d}\chi^2_2$$

1.3 方法二: Pearson's χ^2 test

> **Logic**
>
> 此处的 Pearson's χ^2 test 也被称为卡方拟合度检验/单因素卡方检验, 用于**检验一个分类变量的预期频率与观测到的频率之间是否存在显著差异**

另一个常见的 test 是 Pearson's χ^2 test

Test statistic:

$$T=\sum_{j=1}^{k}\frac{(X_j-np_{0j})^2}{np_{0j}}=\sum_{j=1}^{k}\frac{(X_j-E_j)^2}{E_j}$$

其中 $E_j=\mathbb{E}[X_j]=np_{0j}$ 为 X_j 在 H_0 下的期望值

Limiting distribution:

在 H_0 下, $T\to\chi^2_{k-1}$

⚠ **Remark**

- LRT 和 Pearson's χ^2 为 asymptotically equivalent
- 相较于 LRT, Pearson's χ^2 statistic 更快地收敛到 χ^2_{k-1}, 因此对于 small n 的情况, 更应该使用 Pearson's χ^2 test

Proof: LRT 和 Pearson's χ^2 asymptotically equivalent 及其 asymptotic distribution (补充)

LRT statistic 为:

$$\lambda(X) = 2\sum_{j=1}^{k} X_j \log \frac{\hat{p}_j}{p_{j0}}.$$

我们在 p_{j0} 附近对 $\log(\hat{p}_j/p_{j0})$ 作二阶泰勒展开:

$$\log \frac{\hat{p}_j}{p_{j0}} = -\frac{p_{j0}-\hat{p}_j}{\hat{p}_j} + \frac{1}{2}\frac{(p_{j0}-\hat{p}_j)^2}{\hat{p}_j^2} + o\big((p_{j0}-\hat{p}_j)^2\big).$$

代回定义式得

$$\lambda(X) = -2\sum_{j=1}^{k} X_j \frac{p_{j0}-\hat{p}_j}{\hat{p}_j} + \sum_{j=1}^{k} X_j \frac{(p_{j0}-\hat{p}_j)^2}{\hat{p}_j^2} + o_p(1).$$

由于 $\hat{p}_j = X_j/n$ 且 $\sum_j p_{j0} = \sum_j \hat{p}_j = 1$, 一次项和为零, 即:

$$-2\sum_{j=1}^{k} X_j \frac{p_{j0}-\hat{p}_j}{\hat{p}_j} = -2n\sum_{j=1}^{k}(p_{j0}-\hat{p}_j) = 0$$

因此,

$$\lambda(X) = \sum_{j=1}^{k} \frac{n^2}{X_j}\Big(p_{j0} - \frac{X_j}{n}\Big)^2 + o_p(1).$$

在 H_0 下, 有 $X_j/n \xrightarrow{p} p_{j0}$, 故根据 Slutsky theorem,

$$\frac{n^2}{X_j} \xrightarrow{p} \frac{n}{p_{j0}} \quad \Rightarrow \quad \frac{n^2}{X_j}\Big(p_{j0} - \frac{X_j}{n}\Big)^2 \xrightarrow{p} \frac{n(\hat{p}_j - p_{j0})^2}{p_{j0}}.$$

于是,

$$\lambda(X) \xrightarrow{p} \sum_{j=1}^{k} \frac{(X_j - np_{j0})^2}{np_{j0}} =: T_{\text{Pearson}}.$$

由多项分布的中心极限定理,

$$\sqrt{n}(\hat{p} - p_0) \xrightarrow{d} \mathcal{N}(0, \Sigma), \qquad \Sigma = \operatorname{diag}(p_0) - p_0 p_0^\top, \qquad \mathbf{1}^\top(\hat{p} - p_0) = 0,$$

因此,

$$T_{\text{Pearson}} = n(\hat{p} - p_0)^\top \operatorname{diag}(p_0)^{-1}(\hat{p} - p_0) \xrightarrow{d} \chi^2_{k-1}.$$

综上,

$$\lambda(X) = 2\sum_{j=1}^{k} X_j \log \frac{\hat{p}_j}{p_{j0}} \xrightarrow{d} \chi^2_{k-1}.$$

---

2 Tests for Multinomial Data 的变式: Independence Test 和 Goodness of Fit Test

> **Logic**
>
> 除了上述情形之外, LRT 和 Pearson's χ^2 test 还可以用作 tests of independence, 因为 contingency table 中的所有 elements 共同构成了 multinomial distribution

2.1 Tests of independence

2.1.1 Two binary variables 的情况

问题设置:

令 random variables Y 和 Z 为 binary, 并考虑以下数据和概率表:

	$Y=0$	$Y=1$	
$Z=0$	X_{00}	X_{01}	$X_{0\cdot}$
$Z=1$	X_{10}	X_{11}	$X_{1\cdot}$
	$X_{\cdot 0}$	$X_{\cdot 1}$	n

	$Y=0$	$Y=1$	
$Z=0$	p_{00}	p_{01}	$p_{0\cdot}$
$Z=1$	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	1

可以视 $X=(X_{00},X_{01},X_{10},X_{11}) \sim Multinomial(n,p_{00},p_{01},p_{10},p_{11})$
考虑检验:

$$H_0: Y \text{ and } Z \text{ are independent} \quad v.s. \quad H_1: Y \text{ and } Z \text{ are not independent}$$

即等价于检验:

$$H_0: p_{ij}=p_{i\cdot}p_{\cdot j} \quad \forall\, i,j \in \{0,1\} \quad v.s. \quad H_1: \text{otherwise}$$

解决方法 (LRT):

在 H_1 下, MLE 为:

$$\hat{p}_{ij}=\frac{X_{ij}}{n}.$$

在 H_0 下,

$$p_{ij}=p_{i\cdot}p_{\cdot j}, \quad \hat{p}_{i\cdot}=\frac{X_{i\cdot}}{n}, \quad \hat{p}_{\cdot j}=\frac{X_{\cdot j}}{n},$$

因此 MLE 为:

$$\hat{p}_{ij}^{(0)}=\hat{p}_{i\cdot}\hat{p}_{\cdot j}=\frac{X_{i\cdot}X_{\cdot j}}{n^2}.$$

因此 LRT statistic 为:

$$\lambda(X)=2\sum_{i=0}^{1}\sum_{j=0}^{1}X_{ij}\log\frac{\hat{p}_{ij}}{\hat{p}_{ij}^{(0)}}=2\sum_{i=0}^{1}\sum_{j=0}^{1}X_{ij}\log\frac{nX_{ij}}{X_{i\cdot}X_{\cdot j}}.$$

且 H_0 下有,

$$\lambda(X)\xrightarrow{d}\chi_1^2.$$

> **Remark**
>
> 在考虑自由度时, 有 $dim(\Theta)=4-1=3, dim(\Theta_0)=(2-1)+(2-1)=2$, 因此自由度为 $dim(\Theta)-dim(\Theta_0)=1$

解决方法 (Pearson's χ^2 test):

H_0 下的 Expectation 为:

$$E_{ij} = n\hat{p}_{ij}^{(0)} = \frac{X_{i\cdot}X_{\cdot j}}{n}$$

因此 Pearson χ^2 statistic 为:

$$T = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{(X_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=0}^{1}\sum_{j=0}^{1}\frac{(X_{ij}-X_{i\cdot}X_{\cdot j}/n)^2}{X_{i\cdot}X_{\cdot j}/n}$$

且 H_0 下有,

$$T \xrightarrow{d} \chi_1^2$$

2.1.2 Two categorical variables 的情况

问题设置:

令随机变量 Y 取 I 个类别、Z 取 J 个类别。记

$$X_{ij} = \text{cell count for } (Y=i, Z=j), \quad X_{i\cdot} = \sum_{j=1}^{J} X_{ij}, \quad X_{\cdot j} = \sum_{i=1}^{I} X_{ij}, \quad n = \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}$$

则

$$X = (X_{11}, \ldots, X_{IJ}) \sim \text{Multinomial}\big(n;, p_{11}, \ldots, p_{IJ}\big),$$

其中 $\sum_{i,j} p_{ij} = 1$

考虑检验:

$$H_0:\ p_{ij} = p_{i\cdot}p_{\cdot j}\ \forall\, i,j \quad v.s. \quad H_1: Y \text{ and } Z \text{ are not independent}$$

解决方法 (LRT):

在 H_1 下, MLE 为:

$$\hat{p}_{ij} = \frac{X_{ij}}{n}$$

在 H_0 下,

$$p_{ij} = p_{i\cdot}p_{\cdot j}, \quad \hat{p}_{i\cdot} = \frac{X_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{X_{\cdot j}}{n}$$

因此 MLE 为:

$$\hat{p}_{ij}^{(0)} = \hat{p}_{i\cdot}\hat{p}_{\cdot j} = \frac{X_{i\cdot}X_{\cdot j}}{n^2}.$$

因此 LRT statistic 为:

$$\lambda(X) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}\log\frac{\hat{p}_{ij}}{\hat{p}_{ij}^{(0)}} = 2\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}\log\frac{nX_{ij}}{X_{i\cdot}X_{\cdot j}}.$$

且 H_0 下有,

$$\lambda(X) \xrightarrow{d} \chi_\nu^2, \qquad \nu = (I-1)(J-1).$$

> ⚠ **Remark**
>
> 在考虑自由度时, 有 $dim(\Theta) = IJ-1, dim(\Theta_0) = (I-1)+(J-1) = I+J-2$, 因此自由度为 $dim(\Theta) - dim(\Theta_0) = IJ - I - J + 1 = (I-1)(J-1)$

解决方法 (Pearson's χ^2 test):

H_0 下的 Expectation 为:

$$E_{ij} = n\hat{p}_{ij}^{(0)} = \frac{X_{i\cdot}X_{\cdot j}}{n}.$$

因此 Pearson χ^2 statistic 为:

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(X_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(X_{ij}-X_{i\cdot}X_{\cdot j}/n)^2}{X_{i\cdot}X_{\cdot j}/n}.$$

且 H_0 下有,

$$T \xrightarrow{d} \chi^2_\nu, \qquad \nu = (I-1)(J-1).$$

Logic

除了 independence test 之外, 我们还可以将前面的思想推广, 用于检验样本是否来自一个给定 parametric model

2.2 Test of goodness of fit

Null hypothesis:

Observed data 来自某个 assumed parametric family, 即

$$H_0:\ F \in \mathcal{F} = f(x;\theta) : \theta \in \Theta.$$

思路:

将 data 与 model 均进行 discretize, 由此构造 multinomial distribution

流程:

1. 划分出 k 个互不相交的区间:

$$I_1,\ I_2,\ \ldots,\ I_k$$

2. 计算 observations 中落入第 j 个区间的个数, 即令

$$N_j = \#\{X_i \in I_j\}, \qquad j = 1,\ldots,k,$$

由此 $N = (N_1,\ldots,N_k)$ 满足:

$$N = (N_1,\ldots,N_k) \sim \text{Multinomial}\big(n;, p_1(\theta),\ldots,p_k(\theta)\big)$$

3. 对于每个区间 I_j，定义

$$p_j(\theta) = P_\theta(X \in I_j) = \int_{I_j} f(x;\theta),dx, \qquad j = 1,\ldots,k$$

并求出未知参数 θ 的 MLE:

$$\ell(\theta) = \sum_{j=1}^{k} N_j \log p_j(\theta), \qquad \tilde{\theta} = \arg\max_\theta \ell(\theta)$$

4. 在 H_0 下, observation 的频数应与 model prediction 的期望频数接近, 由此可以定义 Pearson 型统计量:

$$Q = \sum_{j=1}^{k}\frac{(N_j - np_j(\tilde{\theta}))^2}{np_j(\tilde{\theta})}$$

且 H_0 下有:

$$Q \xrightarrow{d} \chi^2_{k-1-s}$$

其中 $s = \dim(\theta)$ 为模型分布参数的维度

Remark

1. 在进行分箱操作的时候, 我们通常会确保每个 category 至少有 5 个 observations, 且我们通常会确保各个 category 的 observations 数相互接近 (例如, 对于 continuous distribution, 我们可以考虑使用分位数进行分箱)
2. 若参数 θ 已知, 则可以不用求解 MLE, test statistic 及其 limiting distribution (under H_0) 分别为:

$$Q = \sum_{j=1}^{k} \frac{(N_j - np_j(\theta))^2}{np_j(\theta)} \xrightarrow{d} \chi^2_{k-1}$$

3. 和 Kolmogorov–Smirnov (KS) Test 的对比:
   - KS test 无需分箱
   - KS test 只适用于一维连续分布
   - KS test statistic 在 H_0 下的渐进分布为 Kolmogorov 分布
   - KS test 对于局部异常尤其敏感, 而 Pearson's χ^2 test 反映的是全局平均效应

   关于 KS test 的详细论述, 见 STA4100 Lecture 4

Example

问题:

给定 iid sample $X_1, .., X_n$, 检验 data 是否服从 Poisson distribution (with unknown parameter λ), 即:

$$H_0 : X_1, \ldots, X_n \sim Poisson(\lambda) \quad v.s. \quad H_1 : Otherwise$$

解答:

首先 construct K categories (确保每个 category 至少有 5 个 observations), 例如:

$$\{0\}, \{1\}, \ldots, \{K-2\}, \{\geq K-1\}$$

随后统计每个 category 的 observations 的数量:

$$\begin{aligned} Y_j &= \#\{X_i : X_i = j-1\}, \quad \text{for } j = 1, \ldots, K-1, \\ Y_K &= \#\{X_i : X_i \geq K-1\} \end{aligned}$$

计算 H_0 下 observation 落入每个区间的概率:

$$p_j(\lambda) = \begin{cases} e^{-\lambda} \cdot \frac{\lambda^{j-1}}{(j-1)!}, & j \leq K-1 \\ 1 - \sum_{j=1}^{K-1} e^{-\lambda} \cdot \frac{\lambda^{j-1}}{(j-1)!}, & j = K \end{cases}$$

并且有:

$$\mathbb{E}[Y_j] = n \cdot p_j(\lambda)$$

若 K 较大 (尾部概率可以忽略), 由于 discrete distribution 本身就是分箱好的形式, 我们可以直接使用原始 MLE:

$$\hat{\lambda}_{MLE} = \bar{X}_n$$

由此 test statistic 为:

$$T = \sum_{j=1}^{K} \frac{(Y_j - n \cdot p_j(\hat{\lambda}))^2}{n \cdot p_j(\hat{\lambda})} \xrightarrow{d} \chi^2_{K-1-1}$$

Remark

若 λ 已知, 则 test statistic 为

$$T = \sum_{j=1}^{K} \frac{(Y_j - n \cdot p_j(\lambda))^2}{n \cdot p_j(\lambda)} \xrightarrow{d} \chi^2_{K-1}$$

Example

问题:

给定 iid sample $X_1, .., X_n$, 检验 data 是否服从 Exponential distribution (with unknown parameter λ), 即:

$$H_0 : X_1, \ldots, X_n \sim Exp(\lambda) \quad v.s. \quad H_1 : Otherwise$$

解答:

H_0 下的 density 为:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0,$$

H_0 下的 CDF 为:

$$F(x;\lambda) = 1 - e^{-\lambda x}.$$

取若干区间 $I_1, \ldots, I_k$ (例如按理论分布的分位数划分，使各区间概率大致相等), 并令:

$$p_j(\lambda) = P_\lambda(X \in I_j) = F(b_j;\lambda) - F(a_j;\lambda),$$

其中 $I_j = (a_j, b_j]$
参数的 MLE 为:

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

观测频数与期望频数分别为:

$$N_j = \#\{X_i \in I_j\}, \qquad E_j = n\, p_j(\hat{\lambda}).$$

Pearson 统计量为:

$$T = \sum_{j=1}^{k} \frac{(N_j - E_j)^2}{E_j},$$

当 H_0 成立且样本量大时, 有:

$$T \xrightarrow{d} \chi^2_\nu, \qquad \nu = k - 1 - s,\ s = 1.$$