山西大学学报（自然科学版）14(3)：249-253，1991
Journal of Shanxi University (Nat. Sci. Ed.)

定性资料的参数估计

张信东　刘维奇

（山西大学数学系）

摘要　张尧庭等人引入了定性资料分析的多元统计方法．本文讨论了定性资料的参数估计及其大样本性质．

关键词　定性资料，最大似然估计

0　引言

定性资料的统计分析在许多实际问题中愈来愈显示出其重要性，引起了不少专家学者的重视．对定性资料的处理通常有两种途径：一是看成列联表，用非参数方法处理；一是定性资料定量化，然后用定量资料的分析方法。张尧庭和胡飞芳（见[1]）首次把多元统计方法系统地用于定性资料的分布．

设 ξ 是一个离散的随机变量，它只取 a_1，…，a_k 这有限个值，相应的概率分别为 p_1，…，p_k，即 $P(\xi=a_i)=p_i$，$i=1$，…，K．令

$$\xi^{(i)}=\begin{cases}1 & \text{当 } \xi=a_i\\ 0 & \text{当 } \xi\neq a_i\end{cases}\qquad i=1,\ \cdots,\ K \tag{0.1}$$

于是 $\xi^{(i)}$ 表示 ξ 是否取到 a_i（个体是否具有第 i 类属性）的指示随机变量．这样一个离散的随机变量 ξ 就与一个 $K\times 1$ 的随机向量 $\underline{\xi}=(\xi^{(1)},\ \cdots,\ \xi^{(k)})^{\tau}$——对应．简记为 $\underline{\xi}\to\xi$．由[1]可知

$$E\underline{\xi}=\underline{p}=(p_1,\ \cdots,\ p_k)^{\tau} \tag{0.2}$$

$$V_{ar}(\underline{\xi})=D-\underline{p}\,\underline{p}^{\tau} \tag{0.3}$$

其中 $D=diag(p_1,\ \cdots p_k)$．

本文将讨论定性资料参数 p_1，…，p_k 的矩估计和最大似然估计，同时证明了它们的无编性和大样本性质，即强相合性和渐近正态性．更为重要的是在含有定性资料的统计模型中，利用定性资料之一一对应的随机向量代替定性资料而对统计模型进行研究，这一点有待进一步讨论．

收稿日期：1990．12．25

1 参数估计

设X_1, …, X_n是定性资料ξ的样本，则由上文中提到的对应关系得到ξ的观察向量$\underline{X}_1$, …, $\underline{X}_n$. 类似地$\underline{X}_j=(X_j^{(1)}, \cdots, X_j^{(k)})^\tau$, $j=1, \cdots, K$.

1.1 矩估计 根据(0.2)式不难得到参数向量$\underline{p}$的矩估计

$$\hat{\underline{p}}_M=\frac{1}{n}\sum_{j=1}^{n}\underline{X}_j \tag{1.1}$$

1.2 最大似然估计 根据ξ的定义可知其只取K个值，即$(1, 0\cdots, 0)^\tau$, …, $(0, 0\cdots, 1)^\tau$. 记$X=(X_1, \cdots, X_n)^\tau$，则似然函数

$$L(X; \underline{p})=p_1^{\sum X_j^{(1)}}\cdots p_k^{\sum X_j^{(k)}} \tag{1.2}$$

由于$\sum_{i=1}^{k}p_i=1$，对数似然函数

$$\begin{aligned} lnL(X; \underline{p}) &= \sum_{i=1}^{k}\left(\sum_{j=1}^{n}X_j^{(i)}\right)\ln p_i \\ &= \sum_{i=1}^{k-1}\left(\sum_{j=1}^{n}X_j^{(i)}\right)\ln p_i+\sum_{j=1}^{n}X_j^{(k)}\ln(1-p_1\cdots\cdots p_{k-1}) \end{aligned} \tag{1.3}$$

令

$$\frac{\delta \ln L(X; \underline{p})}{\delta p_\alpha}=0, \quad \alpha=1, 2\cdots, K-1 \tag{1.4}$$

得到正规方程

$$\frac{\sum_{j=1}^{n}X_j^{(\alpha)}}{\hat{p}_\alpha}=\frac{\sum_{j=1}^{n}X_j^{(k)}}{\hat{p}_k}, \quad \alpha=1, 2\cdots, K-1 \tag{1.5}$$

改写为

$$\hat{p}_\alpha=\frac{\sum_{j=1}^{n}X_j^{(\alpha)}}{\sum_{j=1}^{n}X_j^{(k)}}\hat{p}_k, \quad \alpha=1, 2, \cdots, K-1 \tag{1.6}$$

两边对α求和得

$$\hat{p}_k=\frac{\sum_{j=1}^{n}X_j^{(k)}}{\sum_{\alpha=1}^{k}\sum_{j=1}^{n}X_j^{(\alpha)}} \tag{1.7}$$

从(0.1)注意到

$$\sum_{i=1}^{k}\xi^{(i)}=1$$

故　$$\sum_{a=1}^{k} X_j^{(a)} = 1,\ j = 1,\ 2,\ \cdots,\ n \tag{1.8}$$

利用(1.8)将(1.7)代入(1.6)得

$$\hat{p}_a = \frac{1}{n}\sum_{j=1}^{n} X_j^{(a)},\ a = 1,\ 2,\ \cdots K$$

故参数向量 $\underline{p}$ 的最大似然估计

$$\hat{\underline{p}}_L = \frac{1}{n}\sum_{j=1}^{n} \underline{X}_j \tag{1.9}$$

显然参数 $\underline{p}$ 的最大似然估计就是其矩估计简记为 $\hat{\underline{p}}$.

2　估计 $\hat{\underline{p}}$ 的一些性质

在这一节中将证明定性资料参数估计的无偏性、强相合性和渐近正态性.

定理2.1　由(1.9)式确定的估计量 $\hat{\underline{p}}$ 是参数 $\underline{p}$ 的无偏估计. 即　$E\hat{\underline{p}} = \underline{p}$.

证明　由(0.2)式和(1.9)式即得.

定理2.2　由(1.9)式确定的估计量 $\hat{\underline{p}}$ 是参数 $\underline{p}$ 的强相合估计，即当样本容量 n 趋于无穷时

$$\hat{\underline{p}} \longrightarrow \underline{p}\quad a.s.$$

证明　因为样本 $X_1,\ \cdots,\ X_n$ 是独立随机变量且同 ξ 具有相同分布，故 $\{\underline{X}_j\}$ 是 $i.i.d$ 随机向量序列，从而 $\{X_j^{(i)}\}$，$i = 1,\ \cdots,\ K$ 是 K 个 $i.i.d$ 随机变量序列，且

$$E|X_j^{(i)}| = p_i \leqslant 1\quad i = 1,\ \cdots,\ K \tag{2.1}$$

故由柯尔莫哥洛夫独立同分布的强大数定律有

$$\frac{1}{n}\sum_{j=1}^{n} \underline{X}_j \xrightarrow{n\to\infty} \underline{p}\quad a.s. \tag{2.2}$$

即 $\hat{\underline{p}}$ 是 $\underline{p}$ 的强相合估计.

在讨论估计的渐近正态性之前引入首先如下引理

引理　设 $Y_n = (Y_{n1},\ \cdots,\ Y_{nk})^\tau$，$n = 1,\ 2,\ \cdots$，是一个随机向量序列，已知对任何实数 $C_1,\ \cdots,\ C_k$，

$$C^\tau Y_n \xrightarrow[n\to\infty]{d} N(0,\ C^\tau \Sigma C) \tag{2.3}$$

其中　$C = (C_1,\ \cdots,\ C_k)^\tau$，则　$Y_n \xrightarrow[n\to\infty]{d} N(0,\ \Sigma)$.

证明　据(2.3)式，当 $n\to\infty$ 时，对 $\forall t \in R$

$$f_{C^\tau Y_n}(t) = Ee^{itC^\tau Y_n} \longrightarrow e^{-\frac{1}{2}C^\tau \Sigma C t^2} \tag{2.4}$$

设　$t_1,\ \cdots,\ t_n \in R$，记 $T = (t_1,\ \cdots,\ t_n)^\tau$ 则 Y_n 的联合特征函数

$$f_{Y_n}(t_1, \cdots, t_n) = Ee^{iT^\tau Y_n}$$

取 $t=1$，由(2.4)式

$$Ee^{iT^\tau Y_n} \longrightarrow e^{-\frac{1}{2}T^\tau \Sigma T}$$

再由([2], *Theorem* 3)

$$Y_n \xrightarrow[n\to\infty]{d} N(0, \Sigma) \qquad \text{引理证毕.}$$

定理2.3 由(1.9)式确定的参数 $\underline{p}$ 的估计 $\hat{\underline{p}}$ 具有渐近正态性，即当样本容量 n 趋于无穷时

$$\sqrt{n}(\hat{\underline{p}} - \underline{p}) \xrightarrow{d} N(0, \Sigma)$$

其中 $$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_k \\ \cdots & & \cdots & \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{pmatrix}_{K\times K}$$

证明 令

$$M = \sqrt{n}(\hat{\underline{p}} - \underline{p}) \triangleq (M_1, \cdots, M_k)^\tau$$

则

$$EM = 0 \tag{2.4}$$

$$Var(M) = \frac{1}{n}\sum_{j=1}^{n} Var(\underline{X}_j) = D - \underline{p}\,\underline{p}^\tau = \Sigma \tag{2.5}$$

设 $C = (C_1, \cdots, C_k)^\tau$ 是任一实向量，则

$$EC^\tau M = 0 \tag{2.6}$$

$$VarC^\tau M = C^\tau \Sigma C \tag{2.7}$$

且与样本容量 n 无关. 又因为

$$C^\tau M = C^\tau \sqrt{n}\left(\frac{1}{n}\sum_{j=1}^{n}\underline{X}_j - \underline{p}\right)$$

$$= \frac{\sum_{j=1}^{n} C^\tau(\underline{X}_j - \underline{p}_j)}{\sqrt{nC^\tau\Sigma C}} \cdot \sqrt{C^\tau\Sigma C}$$

其中 $C^\tau(\underline{X}_j - \underline{p})$ 是 *i.i.d* 随机变量，并且

$$E[C^\tau(\underline{X}_j - \underline{p})]^2 \leqslant \sum_{i=1}^{k} C_i^2 \cdot \sum_{i=1}^{k} p_i(1-p_i)$$

$$\leqslant \frac{K}{4}\sum_{i=1}^{k} C_i^2 < \infty$$

故由([2], *Corollary 2 of Theorem* 1)

$$C^{\tau}M \xrightarrow[n\to\infty]{d} N(0,\ C^{\tau}\Sigma C)$$

再由引理可知

$$M=\sqrt{n}\,(\hat{\underline{p}}-\underline{p}) \xrightarrow[n\to\infty]{d} N(0,\ \Sigma)$$

即 $\underline{p}$ 的估计 $\hat{\underline{p}}$ 具有渐近正态性.

参 考 文 献

1 张尧庭，胡飞芳. 定性资料的多元分析方法. 应用概率统计，1990；6(4)：433～436

2 Chow Y, Teicher H. Probability Theory. New York, Springer-Verlag, 1978. 266, 294

PARAMETRIC ESTIMATION OF QUALITATIVE DATA

Zhang Xindong　Liu Weiqi

(*Department of Mathematics, Shanxi University*)

Abstract

Zhang Yaoting et al. discussed how to analyze the qualitative data by the multivariate statistical methods. In this paper, the parametric estimation of the qualitative data and its large sample property is discussed.

Key words qualitative data, maximum likelihood estimation