多項分類上 偏倚에 關한 硏究

(A Note on the Bias in the Multi-nomial Classification)

尹 龍 雲*

Abstract

If two inspectors classify items in a lot into m classes, it is possible that each of them makes wrong classification in some cases, thus causing bias. Expressions have been obtained for tne limits of this bias in estimating the proportion of the different classes. From the results of tne classification they obtained limit for the estimates of proportions have been worked out, based on assumption regarding the magnitudes of propabilities of misclassification.

Now we suppose that P_{ti} $(t=1,2)$ is the probability that t the inspector classifies correctly an item in class A_i and $q_{t/i}$ is the probability that he missclassifies in A_j an item actually belonging to A_i, therefor.

$$P_{ti}+\sum_{j\neq i} q_{t/i}=1$$

An estimate for the proportion P_k of the class A_k in the lot would be

$$\hat{P}_k=r_{1k}+\left(\frac{1}{2}\right)\sum_{j\neq k} r_{kj}+r_{jk}$$

The % Bias in proportion $\hat{P}_k$ is

$$\frac{E(\hat{P}_k)-P_k}{P_k}\times 100$$

1 序

두 檢査量이 m階級에서 하나의 「모트」의 單位로 分類되어진다고 假定하면 어떠한 경우 各 階級이 잘못 分類되어짐으로써 偏倚가 發生하게 되는 可能性이 있게 되는 것인데 이러한 것은 서로 相異한 階級의 比率을 推定함에 있어서의 偏倚限界를 求하고자 함을 意味하게 된다. 特히 $m=3$인 特殊한 경우에 있어서 그 比率의 推定値에 對한 上限, 下限의 限界가 導出되어지는 바, 이는 缺測分類上의 確率의 크기에 關한 假定如何에 따라 달라지는 것이다.

이제 m階級 $A_1, A_2, \cdots\cdots, A_m$中의 하나가 屬해 있는 各 單位인 大母集團의 n單位의 標本이 두 檢査量에 의하여 變한다고 假定하자. 이것은 各 標本의 單位로 分類되어지는 反面 약간의 缺測分類가 되어진다. 이러한 分類結果에서 이것은 서로 相異한 階級의 比率을 推定하는데 있어서의 偏倚에 對한 限界를 求하고자 하는데 있다.

2 階級 $m=2$일 때의 2個의 檢査量에 依한 分類上의 偏倚

母集團이 相當히 큰 P_i에서 各 單位의 比率이 階級 A_i에 屬한다고 假定하자. n單位에서 x_{ij}單位가 檢査量 2에 依하여 A_i에 屬하는 것으로, 또 檢査量 1에 依하여 A_j에 屬하는 것으로 分類한다.

지금 第t次 檢査量이 階級 A_i와 $q_{t/i}$에서 하나의 單位가 正確하게 分類되어지는 確率을 $P_{ti}(t=1.2)$이라하고 階級 A_j에서 實際的으로 A_i에 屬하여 있는 하나의 單位가 缺測分類되어지는 確率을 생각하자. 그러면

* 木浦工業專門學校 工業經營學科 專任講師(1978. 6. 7 接受)

$$P_{ii}+\sum_{j\neq i}q_{ij|i}=1 \quad \cdots\cdots(1)$$

이다. 即

$$\frac{x_{ij}}{n}=r_{ij}$$

라 定義할 수 있다.

또한 階級의 度數에 對한 期待値를 計算하면

$$E(r_{ij})=P_iq_{1j|i}P_{2i}+P_jP_{1j}q_{2i|j}+\sum_{s\neq i,j}P_sq_{1j|s}q_{2i|s} \quad \cdots\cdots(2)$$

$$E(r_{ii})=P_iP_{1i}P_{2i}+\sum_{j\neq i}P_jq_{1i|j}q_{2i|j} \quad \cdots\cdots(3)$$

이 된다.

이 「lot」에서 階級 A_k의 比率 P_k에 對한 推定値는

$$\hat{P}_k=r_{kk}+\left(\frac{1}{2}\right)\sum_{j\neq k}(r_{kj}+r_{jk}) \quad \cdots\cdots(4)$$

이다.

여기서

$$2E(\hat{P}_k)=P_k\left[2-\sum_{j\neq k}(q_{1j|k}+q_{2j|k})\right]+\sum_{j\neq k}P_j(q_{1k|j}+q_{2k|j}) \quad \cdots\cdots(5)$$

이다.

任意로 選擇된 하나의 單位가 缺測分類되어지는 第 t 次 檢査量의 確率을 $q_t(t=1.2)$라 하자.

$$\bar{q}_t=\sum P_i\sum_{j\neq i}\bar{q}_{ji|i} \quad \cdots\cdots(6)$$

그러면 確率抽出單位가 缺測分類되어지는 平均確率은

$$\bar{q}=\left(\frac{1}{2}\right)\sum P_i\sum_{j\neq i}(\bar{q}_{1j|i}+q_{2i|i}) \quad \cdots\cdots(7)$$

이다.

推定된 P값 即 P_k를 갖는 項을 分離하거나 또한 그러한 階級(q_{1k1}, q_{1k2}, ……, q_{2k1}, q_{2k2}, ……와 같은 項)에서 缺測分類와 一致하게 되는 q項을 分離하게 되면 (7)式에서

$$2\bar{q}=\hat{P}_k\sum_{j\neq k}(q_{1j|k}+q_{2j|k})+\sum_{j\neq k}P_j(q_{1k|j}+q_{2k|j})+\sum_{s\neq k}P_s\sum_{j\neq s,k}(q_{1j|s}+q_{2j|s}) \quad \cdots\cdots(8)$$

을 얻게 된다.

(8式)에서의 $\sum_{j=k}P_j(q_{1k|j}+q_{2k|j})$의 값을 (5式)에 代入하여 풀면

$$E(\hat{P}_k)=\bar{q}+P_k\left[1-\sum_{j\neq k}(q_{1j|k}+q_{2j|k})\right]-\left(\frac{1}{2}\right)\sum_{i\neq k}P_s\sum_{j\neq s,k}(q_{1j|s}+q_{2j|s}) \quad \cdots\cdots(9)$$

을 얻는다. 따라서

$$P_k=\frac{E(\hat{P}_k)-\bar{q}+\left(\frac{1}{2}\right)\sum_{i\neq k}P_s\sum_{j\neq s,k}(q_{1j|s}+q_{2j|s})}{1-\sum_{j\neq k}(q_{1j|k}+q_{2j|k})} \quad \cdots\cdots(10)$$

이다. 또한 比率 $\hat{P}_k$의 百分率偏倚(% Bias)는

$$\frac{E(\hat{P}_k)-P_k}{P_k}\times 100(\%) \quad \cdots\cdots(11)$$

이다. 그러므로

$$\%\ \text{Bias}=\frac{\bar{q}-\sum_{j\neq k}(q_{1j|k}+q_{2j|k})E(P_k)-\left(\frac{1}{2}\right)\sum_{s\neq k}P_s\sum_{j\neq s,k}(q_{1j|s}+q_{2j|s})}{E(P_k)-\bar{q}+\left(\frac{1}{2}\right)\sum_{i\neq k}P_s\sum_{j\neq s,k}(q_{1j|s}+q_{2j|s})}\times 100 \quad \cdots\cdots(12)$$

이다.

$\hat{P}_k$에 있어서의 最大値 陽의 偏倚 및 最小値 陰의 偏倚는 만일 階級 $As(s=1.2\cdots\cdots m$ 및 $s\neq k)$가 屬하는 單位가 階級 $A_j(j=1.2\cdots\cdots.m$ 및 $j\neq s.k)$에서 缺測·分類되어진다고 한다면 아무런 影響을 받지 않는다. 따라서 $\hat{P}_k$의 偏倚를 計算함에 있어서 이러한 類型의 缺測分類를 하지 않는 一般的인 損失을 갖지 않고도 假定할 수 있게 되는 것이다. 即

$$q_{1j\cdot s}=q_{2j\cdot s}=0 \quad (j\neq s.k \text{ 및 } s\neq k) \quad \cdots\cdots(13)$$

이다. 따라서 比率 $\hat{P}_k$의 %偏倚는

$$\%\ \text{Bias}=\frac{\bar{q}-\sum_{j\neq k}(q_{1j\cdot k}+q_{2j\cdot k})E(\hat{P}_k)}{E(\hat{P}_k)-\bar{q}}\times 100 \quad \cdots\cdots(14)$$

이다.

$\hat{P}_k$의 가장 큰 陽의 偏倚는 두 檢査量 各各의 A_k에서 A_j에 屬하는 하나의 單位가 缺測分類되는 陽의 確率을 나타낼 때 發生되며 또한 階級 A_k에 屬하는 하나의 單位가 缺測分類되는 "0 確率"(zero probability)을 가질 때 생긴다. 即

$$q_{1j\cdot k}=q_{2j\cdot k}=0 \quad (j\neq k) \quad \cdots\cdots(15)$$

이다.

따라서 (14式)에서 %偏倚上限境界

$$\overline{\%\ \text{Bias}}=\frac{\bar{q}}{E(\hat{P}_k)-\bar{q}}\times 100 \quad \cdots\cdots(16)$$

을 얻게 된다.

한편 가장 작은 陰의 偏倚는 두 檢査量의 各各이 階級 A_k에서 階級 $A_j(j=1.2.\cdots\cdots.m$ 및 $J\neq k)$의 하나의 單位가 缺測分類되는 "0"確率을 갖일 때 생기며 그리고 A_j에서 A_k의 하나의 單位가 缺測分類되는 陽의 確率을 가질 때 생기게 되는 것이다. 即

$$q_{1k\cdot j}=q_{2kj}=0 \quad (j\neq k) \quad \cdots\cdots(17)$$

이다. 따라서 (14式)으로 부터의 %偏倚 下限境界인 $\underline{\%\ \text{bias}}$는 (17式)에 의하여 評價되어 진다.

이제 (8式)과 (9式)의 方法을 利用하여 (14式)을 單純化시키면

$$\underline{\%\ \text{偏倚}} = -\frac{q}{E(\hat{P}_k)+\bar{q}}\times 100 \quad \cdots\cdots(18)$$

을 얻게 된다. (4式)에서

$$E\left[\left(\frac{1}{2}\right)\sum_{j\neq k}(r_{kj}+r_{jk})\right]=E(\hat{P}_k)-E(r_{kk}) \quad \cdots(19)$$

이다. (9式)과 (3式)을 (19式)에 代入하여 (13式)과 (1式)의 方法을 利用하여 單純化시키면

$$E\left[\left(\frac{1}{2}\right)\sum_{j\neq k}(r_{kj}+r_{jk})\right]=\bar{q}-P_k\sum_{j\neq k}q_{1jk}\sum_{j\neq k}q_{2jk}-\sum_{j\neq k}P_jq_{1kj}q_{2kj} \quad \cdots\cdots(20)$$

에 對立하게 된다.

따라서 이는

$$\bar{q}=E\left[\left(\frac{1}{2}\right)\sum_{j\neq k}(r_{kj}+r_{jk})\right]+P_k\sum_{j\neq k}q_{1jk}\sum_{j\neq k}q_{2jk}+\sum_{j\neq k}P_jq_{1kj}q_{2kj} \quad \cdots\cdots(21)$$

로 된다.

또한 (15式)의 方法을 (21式)에 利用하고 이를 (16式)에다 代入하면 % 偏倚上限境界는

$$\underline{\%\ \text{Bias}}=\frac{E\left[\left(\frac{1}{2}\right)\sum_{j\neq k}(r_{kj}+r_{jk})\right]+\sum_{j\neq k}P_jq_{1kj}q_{2kj}}{E(r_{kk})\sum_{j\neq k}P_jq_{1kj}q_{2kj}}\times 100 \quad \cdots\cdots(22)$$

을 얻는다.

다음 (17式)을 (21式)에 利用하고 이를 (18式)에 代入하면 % Bias 下限境界는

$$\underline{\%\ \text{Bias}}=\frac{E\left[\left(\frac{1}{2}\right)\sum_{j\neq k}(r_{kj}+r_{jk})\right]+P_k\sum_{j\neq k}q_{1jk}\ \sum q_{2jk}}{E\left[r_{kk}+\sum_{j\neq k}(r_{kj}+r_{jk})\right]+P_k\sum_{j\neq k}\ q_{1jk}\sum_{j\neq k}q_{2jk}}\times 100 \quad \cdots\cdots(23)$$

을 얻게 된다.

3 $m=3$일때의 2檢査量에 依한 分類上의 偏倚

(22式) 및 (23式)에서 $m=3$으로 놓으면 P_1의 推定値에 對하여

$$\underline{\%\ \text{Bias}}=\frac{\left(\frac{1}{2}\right)E(r_{12}+r_{13}+r_{21}+r_{31})+P_2q_{112}}{E(r_{11})-P_2q_{112}q_{212}-\frac{q_{212}+P_3q_{113}q_{213}}{P_3q_{113}q_{213}}}\times 100 \quad \cdots\cdots(24)$$

$$\underline{\%\ \text{Bias}}=-\frac{\left(\frac{1}{2}\right)E(r_{12}+r_{13}+r_{21}+r_{31})+P_1}{E(r_{11}+r_{12}+r_{13}+r_{21}+r_{31})+P_1}\ \frac{(q_{121}+q_{131})(q_{221}+q_{231})}{(q_{121}+q_{131})(q_{221}+q_{231})}\times 100 \quad \cdots\cdots(25)$$

을 얻게 된다.

任意의 缺測分類의 確率이 작고 만일 그 分類가 類似한 品質에 依存된다고 期待하는 것이 妥當하면 A_3의 缺測分類確率에 있어서는 하나의 單位가 A_1部類에 屬하며 또한 이것은 逆으로 매우 작게 나타나게 된다. 이러한 根據에 의해 缺測分類確率이 q_{113}, q_{213}, q_{131} 및 q_{231}이 0.05보다 적으며 또한 q_{112}, q_{121}, q_{123}, q_{132}, q_{212}, q_{221}, q_{223} 및 q_{232}가 0.1보다 작다고 假定하자.

(2式) (13式) 및 (15式)에서 %偏倚上限境界에 對하여

$$E(r_{12})=P_2p_{12}q_{212}>0.90P_2q_{212} \quad \cdots\cdots(26)$$

$$E(r_{13})=P_3p_{13}q_{213}>0.95\,P_3q_{213} \quad \cdots\cdots(27)$$

$$E(r_{21})=P_2q_{112}p_{22}>0.90\,P_2q_{112} \quad \cdots\cdots(28)$$

$$E(r_{31})=P\,q_{113}p_{23}>0.95\,P_3q_{113} \quad \cdots\cdots(29)$$

를 얻을 수가 있다.

그래서

$$E(r_{12})E(r_{21})>0.8100P_2[P_2q_{112}q_{212}] \cdots(30)$$

및 $$E(r_{13})E(r_{31})>0.9025P_3[P_3q_{113}q_{213}] \cdots(31)$$

이 된다.

$\hat{P}_1$에서 最大陽의 偏倚를 推定하면 比率 P_2, P_3는 (13式), (15式)의 定理에 의하여 過少하게 推定되어진다.

그러므로

$$P_2>E(r_{12}+r_{21}+r_{22}+r_{23}+r_{32})$$

$$P_3>E(r_{13}+r_{23}+r_{31}+r_{32}+r_{33})$$

이다.

따라서 (30式)과 (31式)에서

$$P_2q_{112}q_{212}<\frac{E(r_{12})E(r_{21})}{0.8100P_2}<\frac{E(r_{12})E(r_{21})}{0.8100E(r_{12}+r_{21}+r_{22}+r_{23}+r_{32})} \quad \cdots(32)$$

및 $$P_3q_{113}q_{213}<\frac{E(r_{13})E(r_{31})}{0.9025P_3}<\frac{E(r_{13})E(r_{31})}{0.9025\,E(r_{13}+r_{23}+r_{31}+r_{32}+r_{33})} \quad \cdots(33)$$

을 導出하게 된다.

또한 (32式)과 (33式)을 (24式)에 代入하면

$$\underline{\%\ \text{Bias}}<\frac{\left(\frac{1}{2}\right)E(r_{12}+r_{13}+r_{21}+r_{31})}{E(r_{11})}+\frac{\frac{E(r_{12})E(r_{21})}{0.8100\,E(r_{12}+r_{21}+r_{22}+r_{23}+r_{32})}-\frac{E(r_{12})E(r_{21})}{0.8100E(r_{12}+r_{21}+r_{22}+r_{23}+r_{32})}+\frac{E(r_{13})E(r_{31})}{0.9025\,E(r_{13}+r_{23}+r_{31}+r_{32}+r_{33})}}{-\frac{E(r_{13}E(r_{31})}{0.9025E(r_{13}+r_{23}+r_{31}+r_{32}+r_{33})}}$$

$\times 100$ ……(34)

을 얻게 된다.

(2式), (13式), (17式)에서 % 偏倚 下限境界는

$$E(r_{12})=P_2q_{121}p_{21}>0.85p_1q_{121} \quad \cdots\cdots(35)$$

$$E(r_{13})=P_1q_{131}p_{21}>0.85\,P_1q_{131} \quad \cdots\cdots(36)$$

$$E(r_{21})=P_1p_{11}q_{221}>0.85\,P_1q_{221} \quad \cdots\cdots(37)$$

$$E(r_{31})=P_1p_{11}q_{231}>0.85\,P_1q_{231} \quad \cdots\cdots(38)$$

을 갖게 된다.

그래서

$$0.7225\,P_1[P_1(q_{121}+q_{131})(q_{221}+q_{231})]$$
$$<E(r_{12}+r_{13})E(r_{21}+r_{31}) \quad \cdots\cdots(39)$$

이다.

한편 $\hat{P}$에서 最小陰의 偏倚를 推定하면 P_1은 過少하게 推定되어진다.

그러므로

$$P_1>E(r_{11}+r_{12}+r_{13}+r_{21}+r_{31})$$

이다. 即 (39式)에서

$$P_1(q_{121}+q_{131})(q_{221}+q_{231})<\frac{E(r_{12}+r_{13})E(r_{21}+r_{31})}{0.7225P_1}$$

$$<\frac{E(r_{12}+r_{13})E(r_{21}+r_{31})}{0.7225\,E(r_{11}+r_{12}+r_{13}+r_{21}+r_{31})} \quad \cdots\cdots(40)$$

을 가지며 또한 (40式)을 (25式)에 代入하면

$$\%\ \text{Bias}>-\frac{\left(\frac{1}{2}\right)E(r_{12}+r_{13}+r_{21}+r_{31})}{E(r_{11}+r_{12}+r_{13}+r_{21}+r_{31})}$$

$$+\frac{\dfrac{E(r_{12}+r_{13})E(r_{21}+r_{31})}{0.7225(r_{11}+r_{12}+r_{13}+r_{21}+r_{31})}}{+\dfrac{E(r_{12}+r_{13})E(r_{21}+r_{31})}{0.7225E(r_{11}+r_{12}+r_{13}+r_{21}+r_{31})}}\times 100 \cdots(41)$$

을 導出할 수가 있는 것이다.

만일 標本크기가 相當히 有效하다면 (34式)과 (41式)의 期待値는 觀察値로써 代替할 수 있다.

P_1에 對한 近似値 95% 信賴限界는

$$S\cdot E=(4n)^{-\frac{1}{2}}\{4r_{11}(1-r_{11})+r_{12}(1-r_{12})$$
$$+r_{21}(1-r_{21})+r_{13}(1-r_{13})+r_{31}(1-r_{31})$$
$$+4r_{11}r_{12}-4r_{11}r_{21}-4r_{11}r_{13}-4r_{11}r_{31}$$
$$-2r_{12}r_{21}-2r_{12}r_{13}-2r_{12}r_{31}-2r_{21}r_{13}$$
$$-2r_{21}r_{31}-2r_{13}r_{31}\}^{\frac{1}{2}} \quad \text{일 때}$$

$$\hat{P}_1-2S\cdot E-(0.01)\overline{\%\,\text{bias}}\hat{P}_1(\hat{P}_1+2S\cdot E)<P_1<P_1$$
$$+eS\cdot E-(0.01)\%\ \text{bias}\,\hat{P}_1(\hat{P}+2S\cdot E)$$

로써 주어진다.

P_2 및 P_3에 對한 信賴限界도 이와같은 類似한 方法으로 얻어질 수가 있다.

4 數理的 事例

100單位의 標本을 〈表 -1〉의 結果로서 두 檢査量에 依한 3個의 階級—最適, 適合, 不適으로 分類하여 생각해 보자.

〈表-1〉 2 檢査量의 分類結果表

1檢査量 / 2檢査量		檢査量 1			計
		最 適	適 合	不 適	
檢査量 2	最適	74	6	1	81
	適合	3	5	2	10
	不適	2	4	3	9
計		79	15	6	100

最適階級比率의 推定値에 對한 結果를 應用함으로써

$P_1=0.80$ $\quad S\cdot E=0.0361$

% Bias=10.03%(% 偏倚上限境界)

% Bias=−7.59%(% 偏倚下限境界)

이며 이때의 信賴限界는

$0.6403<p_1<0.9384$로 주어진다.

參考 文獻

Barlow, R.E. and Proschan, F. Mathematica Theory of Reliability, John Wiley & Sons New York, 1965.

Walker, H.M. and Lev, Joseph, Statistical Inference, Henry Holt and Co., New York, 1953.

Siegel, Sidney, Nonparametric Statistics, McGraw-Hill Book Co., New York, 1956.

Dixon, W.J and Massey, F.J., Jr., Introduction to Statistical Analysis, McGraw-Hill Book Co., New York, 1957.

Tias, G.C., and Tan, W.Y., "Bayesian Analysis of Random-Effect Models…," Technical Report No. 30., University of Wisconsin, 1964.

Savage, L. J., The Foundations of Statistics, John Wiley & Sons, New York, 1954.

Wilks, S.S. Mathematical Statistics, John Wiley & Sons., New York, 1962.

Kendall, M.G.: "Advanced Theory of Statistics," Vol. I, Charles Griffin & Co., Ltd., London, 1948.

Hogg, R., and A. Craig: "Introduction to Mathematical Statistics," The Macmillan Company, New York, 1959.

Snedecor, G.W., Statistical Methods, 5th ed., Iowa State College Press, Ames, Iowa, 1956.