일반화된 네이만－피어슨판별규칙의 추정에 대한 모의비교

정현성, 림창호

경애하는 **김정은**동지께서는 다음과 같이 말씀하시였다.

《첨단돌파전을 힘있게 벌려야 나라의 과학기술전반을 빨리 발전시키고 지식경제의 토대를 구축해나갈수 있습니다.》

론문에서는 일반화된 네이만－피어슨판별규칙의 추정에 대한 모의비교문제를 론의하였다.

선행연구[2－4]에서는 2개의 모집단에 대하여 제1종의 오판별확률이 α를 넘지 않는 조건밑에서 제2종의 오판별확률이 최소로 되는 네이만－피어슨판별규칙과 그에 대한 추론문제를 취급하였으며 선행연구[1]에서는 일반화된 네이만－피어슨판별규칙의 한가지 추정문제에 대하여 론의하였다.

표본 $\boldsymbol{x}=(x_1, \cdots, x_p)^{\mathrm{T}}$가 모집단 $G=\{G_1, \cdots, G_m\}$에 속한다고 하고 G_r의 밀도함수를 $f_r(\boldsymbol{x})$ $(r=1, \cdots, m)$라고 하자.

표본공간 (R^p, B^p)의 어떤 분할규칙 $R^p=R_1^p+\cdots+R_m^p$ $(R_r^p\in B^p$ $(r=1, \cdots, m))$를 생각하고 $\boldsymbol{x}\in R_r^p\Rightarrow \boldsymbol{x}\in G_r$로 판별한다. 이때 $\boldsymbol{x}\in G_r$일 때 $\boldsymbol{x}\notin G_r$라고 잘못 판별하는 오판별확률은 $\alpha_r=\int\limits_{\overline{R}_r^p} f_r(\boldsymbol{x})d\boldsymbol{x}$, $\overline{R}_r^p+R_r^p=R^p$이다.

일반화된 네이만－피어슨판별규칙[1]

$$((R_1^P)^*, \cdots, (R_m^P)^*)_{(\alpha_1, \cdots, \alpha_{m-1})}\in \arg\min_{\Phi(\alpha_1, \cdots, \alpha_{m-1})} \int\limits_{\overline{R}_m^P} f_m(\boldsymbol{x})d\boldsymbol{x}$$

에서 $\alpha_m=\int\limits_{(\overline{R}_m^P)^*} f_m(\boldsymbol{x})d\boldsymbol{x}$는 $(\alpha_1, \cdots, \alpha_{m-1})\in\Delta(\alpha_1, \cdots, \alpha_{m-1})$들로 결정되는 함수이며

$$\Delta(\alpha_1, \cdots, \alpha_{m-1})=\left\{(\alpha_1, \cdots, \alpha_{m-1})\middle| 0\le\alpha_1\le 1, \int\limits_{(R_1^p)^*} f_2(\boldsymbol{x})d\boldsymbol{x}\le\alpha_2\le 1, \cdots, \int\limits_{(R_1^p)^*+\cdots+(R_{m-2}^p)^*} f_{m-1}(\boldsymbol{x})d\boldsymbol{x}\le\alpha_{m-1}\le 1\right\}$$

이다.

$$\Phi(\alpha_1, \cdots, \alpha_{m-1})=\left\{(R_1^p, \cdots, R_{m-1}^p)\middle| \int\limits_{\overline{R}_1^p} f_1(\boldsymbol{x})d\boldsymbol{x}\le\alpha_1, \cdots, \int\limits_{\overline{R}_{m-1}^p} f_{m-1}(\boldsymbol{x})d\boldsymbol{x}\le\alpha_{m-1}, R_1^p, \cdots, R_{m-1}^p\in B^p\right\}$$

$$\overline{R}_r^p=R^p-R_r^p \quad (r=1, \cdots, m)$$

이다.

표본자료 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$이 주어졌을 때 주어진 $\alpha_1, \cdots, \alpha_{m-1}$에 대하여 추정된 밀도함수를 가지는 일반화된 네이만－피어슨판별규칙[1]은 다음과 같다.

$$
\begin{aligned}
(\hat{R}_1^P)^* &= \{\boldsymbol{x} \mid q_2\hat{f}_2(\boldsymbol{x}) \le q_1\hat{f}_1(\boldsymbol{x}),\ \cdots,\ q_m\hat{f}_m(\boldsymbol{x}) \le q_1\hat{f}_1(\boldsymbol{x})\} \\
(\hat{R}_2^P)^* &= \{\boldsymbol{x} \mid q_2\hat{f}_2(\boldsymbol{x}) > q_1\hat{f}_1(\boldsymbol{x}),\ \cdots,\ q_m\hat{f}_m(\boldsymbol{x}) \le q_2\hat{f}_2(\boldsymbol{x})\} \\
&\cdots \qquad \cdots \qquad \cdots \\
(\hat{R}_m^P)^* &= \{\boldsymbol{x} \mid q_m\hat{f}_m(\boldsymbol{x}) > q_1\hat{f}_1(\boldsymbol{x}),\ \cdots,\ q_m\hat{f}_m(\boldsymbol{x}) > q_{m-1}\hat{f}_{m-1}(\boldsymbol{x})\}
\end{aligned} \tag{1}
$$

일반성을 잃지 않고 $\{\boldsymbol{x} \mid \hat{f}_1(\boldsymbol{x}) = \cdots = \hat{f}_m(\boldsymbol{x}) = 0\} = \varnothing$ 이라고 가정하며 $\hat{f}_1(\boldsymbol{x}),\ \cdots,\ \hat{f}_m(\boldsymbol{x})$ 은 R^p 에서 련속이라고 가정한다.

보조정리 $Q_{m-1} := \{(q_1,\ \cdots,\ q_{m-1}) \mid 0 \le q_1,\ \cdots,\ q_{m-1},\ q_m \le 1,\ q_1 + \cdots + q_{m-1} + q_m = 1\}$

$$
S_{m-1} := \left\{ (\alpha_1,\ \cdots,\ \alpha_{m-1}) \left| \int_{(\hat{R}_1^P)^*} \hat{f}_1(\boldsymbol{x})d\boldsymbol{x} = \alpha_1,\ \cdots,\ \int_{(\hat{R}_{m-1}^P)^*} \hat{f}_{m-1}(\boldsymbol{x})d\boldsymbol{x} = \alpha_{m-1} \right. \right\}
$$

로 표시할 때 넘기기 $T : Q_{m-1} \to S_{m-1}$ 은 우로의 단일넘기기이다. 여기서

$$
T(q_1,\ \cdots,\ q_{m-1}) = (T_1(q_1,\ \cdots,\ q_{m-1}),\ \cdots,\ T_{m-1}(q_1,\ \cdots,\ q_{m-1})) = (\alpha_1,\ \cdots,\ \alpha_{m-1})
$$

일반화된 네이만－피어슨판별규칙의 추정알고리듬은 다음과 같다.

① 구역 $Q_{m-1}^1 := \{(q_1,\ \cdots,\ q_{m-1}) \mid 0 \le q_1,\ \cdots,\ q_{m-1},\ q_m \le 1\}$ 을 한변의 길이가 $1/(m-1)$ 이고 등간격인 $(m-1)^{m-1}$ 개의 정립방체구역들로 분할한다. 이때 서로 다른 정점들의 개수는 m^m 이다.

② 서로 다른 정점 $\boldsymbol{q}' = (q_1',\ \cdots,\ q_{m-1}')$ 들가운데서 $q_1' + \cdots + q_{m-1}' \le 1$ 을 만족시키는 정점 $\boldsymbol{q}'$ 에 대해서는 $\boldsymbol{\alpha}' = (\alpha_1',\ \cdots,\ \alpha_{m-1}')$ 를 모두 구하고 $\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\| \Rightarrow \min$ 으로 되는 $\boldsymbol{\alpha}^1 = (\alpha_1^1,\ \cdots,\ \alpha_{m-1}^1)$ 과 그에 대응하는 $\boldsymbol{q}^1 = (q_1^1,\ \cdots,\ q_{m-1}^1)$ 을 구한다. 그러나 서로 다른 모든 정점 $\boldsymbol{q}'$ 에 대하여 $q_1' + \cdots + q_{m-1}' > 1$ 이면 주어진 $\boldsymbol{\alpha} = (\alpha_1,\ \cdots,\ \alpha_{m-1})$ 에 대한 일반화된 네이만－피어슨판별규칙은 존재하지 않는다.

③ α_i^1 과 α_i $(i = 1,\ \cdots,\ m-1)$ 의 크기관계를 비교하여 $\alpha_i^1 = \alpha_i$ 이면 $q_i'' = q_i^1 + 1/(m-1)$ (또는 $q_i'' = q_i^1 - 1/(m-1)$)로, $\alpha_i^0 < \alpha_i$ 이면 $q_i'' = q_i^1 - 1/(m-1)$ 로, $\alpha_i^0 > \alpha_i$ 이면 $q_i'' = q_i^1 + 1/(m-1)$ 로 잡고 $\overrightarrow{q_1^1 q_1''},\ \cdots,\ \overrightarrow{q_{m-1}^1 q_{m-1}''}$ 를 모서리로 하는 정립방체를 ①의 $(m-1)^{m-1}$ 개의 정립방체들가운데서 찾는다.

④ ③에서 찾은 정립방체에서 $\boldsymbol{q}^1$ 을 원점으로 하는 자리표계를 다시 설정하면 $Q_{m-1}^2 := \{(q_1,\ \cdots,\ q_{m-1}) \mid 0 \le q_1,\ \cdots,\ q_{m-1},\ q_m \le 1/(m-1)\}$ 이다.

⑤ Q_{m-1}^2 를 한변의 길이가 $(1/(m-1))^2$ 이고 등간격인 $(m-1)^{m-1}$ 개의 정립방체구역들로 분할하고 ②로부터 ④까지 반복하면 $\boldsymbol{q}^2$ 과 Q_{m-1}^3 이 얻어진다.

⑥ 마찬가지로 반복하면 점렬 $\{\boldsymbol{q}^n\}$ 이 얻어진다.

정리 일반화된 네이만－피어슨판별규칙의 추정알고리듬에 의한 점렬 $\{\boldsymbol{q}^n\}$ 은 $n \to \infty$ 일 때 주어진 $\boldsymbol{\alpha} = (\alpha_1,\ \cdots,\ \alpha_{m-1})$ 에 대한 일반화된 네이만－피어슨판별규칙의 $\boldsymbol{q}^* = (q_1^*,\ \cdots,\ q_{m-1}^*)$ 에로 수렴한다.

모집단에 대한 밀도함수를 알고있을 때의 일반화된 네이만－피어슨판별규칙과 모르고 있을 때의 추정된 일반화된 네이만－피어슨판별규칙을 추정알고리듬으로써 모의비교한다.

표본 x가 모집단 $G=\{G_1,\ G_2,\ G_3\}$에 속한다고 하고 G_r $(r=1,\ 2,\ 3)$의 밀도함수를 다음과 같이 준다.

$$f_1(x)=\frac{1}{\sqrt{10}}\left(\int_0^1 x^4(1-x)^{-1/2}dx\right)^{-1}\left(1+\frac{x^2}{10}\right)^{-11/2}=\frac{1}{\sqrt{10}}\frac{315}{256}\left(1+\frac{x^2}{10}\right)^{-11/2}$$: $S(10)$의 밀도함수

$$f_2(x)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-3)^2}{2}\right\}$$: N(3, 1)의 밀도함수

$$f_3(x)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-6)^2}{2}\right\}$$: N(6, 1)의 밀도함수

모집단 G_1, G_2는 분포형태를, 모집단 G_3은 평균과 분산을 모른다고 가정하자.

모집단의 사전확률은 모두 같다고 하자.

매 모집단에서 우연적으로 취한 10개의 표본자료

G_1: −0.409, −0.464, 1.716, 1.206, −0.097, −0.137, −0.486, 0.299, −1.376, 2.174

G_2: 2.219, 1.637, 3.592, 2.710, 2.370, 3.442, 3.500 6, 3.396 8, 3.673 8, 3.136 4

G_3: 7.371, 6.629, 6.835 2, 6.541 6, 5.765 4, 5.979 6, 3.977 9, 5.659 5, 4.343 6, 6.027 8

에 대하여 모집단밀도함수를 추정하면 다음과 같다.

$$\hat{f}_1(x)=\frac{1}{\sqrt{2\pi}\cdot 1.11}\exp\left\{-\frac{(x-0.24)^2}{2(1.11)^2}\right\},\ \hat{f}_2(x)=\frac{1}{\sqrt{2\pi}\cdot 0.7}\exp\left\{-\frac{(x-2.97)^2}{2(0.7)^2}\right\}$$

$$\hat{f}_3(x)=\frac{1}{\sqrt{2\pi}\cdot 1.06}\exp\left\{-\frac{(x-5.91)^2}{2(1.06)^2}\right\}$$

모집단밀도함수를 알고있을 때의 일반화된 네이만－피어슨판별규칙에 대한 추정값은 표 1과 같고 추정된 일반화된 네이만－피어슨판별규칙에 대한 추정값은 표 2와 같다.

표 1. 모집단밀도함수를 알고있을 때의 일반화된 네이만－피어슨판별규칙에 대한 추정값

걸음	$\tilde{q}_1$	$\tilde{q}_2$	$\tilde{q}_3$	$\tilde{\alpha}_1=0.1$	$\tilde{\alpha}_2=0.1$	$\tilde{\alpha}_3$
1	0.5	0.5	0	0.076 7	0.077 8	0.498 8
4	0.1875	0.437 5	0.375	0.125 5	0.101 8	0.073 5
7	0.2422	0.445 3	0.312 5	0.110 3	0.102 2	0.083 4
10	0.249	0.452 1	0.298 8	0.109 3	0.100 6	0.086 6
13	0.2499	0.453	0.297 1	0.109 3	0.100 6	0.086 6
19	0.25	0.453 1	0.296 9	0.109 3	0.100 3	0.087
25	0.25	0.453 1	0.296 9	0.109 3	0.100 3	0.087

표 2. 일반화된 네이만－피어슨판별규칙에 대한 추정값(표본의 개수가 10개인 경우)

걸음	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{\alpha}_1=0.1$	$\hat{\alpha}_2=0.1$	$\hat{\alpha}_3$
1	0.5	0.5	0	0.064 9	0.064 3	0.498 9
4	0.187 5	0.312 5	0.5	0.082 5	0.110 1	0.037 1
7	0.132 8	0.335 9	0.531 3	0.099 1	0.096 9	0.037 4
10	0.127 9	0.329 1	0.543 0	0.099 5	0.098 8	0.036 5
13	0.127 1	0.328 2	0.544 7	0.099 9	0.099	0.036 3
19	0.127	0.328 1	0.544 9	0.099 9	0.099	0.036 3
25	0.127	0.328 1	0.544 9	0.099 9	0.099	0.036 3

마찬가지방법으로 매 모집단에서 우연적으로 취한 100개의 표본자료에 대하여 모집단밀도함수를 추정하면 다음과 같다.

$$\hat{f}_1(x)=\frac{1}{\sqrt{2\pi}\cdot 1.1}\exp\left\{-\frac{(x-(-0.06))^2}{2(1.1)^2}\right\},\ \hat{f}_2(x)=\frac{1}{\sqrt{2\pi}\cdot 1.01}\exp\left\{-\frac{(x-2.95)^2}{2(1.01)^2}\right\}$$

$$\hat{f}_3(x)=\frac{1}{\sqrt{2\pi}\cdot 1.04}\exp\left\{-\frac{(x-5.93)^2}{2(1.02)^2}\right\}$$

이때 추정된 일반화된 네이만－피어슨판별규칙에 대한 추정값은 표 3과 같다.

표 3. 일반화된 네이만－피어슨판별규칙에 대한 추정값(표본의 개수가 100개인 경우)

결음	$\hat{q}_1$	$\hat{q}_2$	$\hat{q}_3$	$\hat{\alpha}_1=0.1$	$\hat{\alpha}_2=0.1$	$\hat{\alpha}_3$
1	0.5	0.5	0	0.077 2	0.078 1	0.498 9
4	0.187 5	0.437 5	0.375 0	0.126 4	0.105 9	0.078 1
7	0.242 2	0.460 9	0.296 9	0.113	0.101 3	0.093 3
10	0.249	0.465 8	0.285 2	0.112	0.1	0.096 1
13	0.249 9	0.466 2	0.283 9	0.111 5	0.1	0.096 6
19	0.25	0.466 3	0.283 7	0.111 5	0.1	0.096 6
25	0.25	0.466 3	0.283 7	0.111 5	0.1	0.096 6

모의자료에 대한 결과에서 보는바와 같이 밀도함수를 알고있을 때의 일반화된 네이만－피어슨판별규칙과 모르고있을 때의 추정된 일반화된 네이만－피어슨판별규칙에서 표본의 개수가 10일 때의 α_3에 대한 오차의 절대값은 0.050 7이며 표본의 개수가 100일 때의 오차의 절대값은 0.000 96이라는것을 알수 있다.

또한 표본의 개수가 증가할 때 판별경계를 나타내는 q_1, q_2, q_3 값들의 차도 작아진다는것을 알수 있다.

참 고 문 헌

[1] 림창호, 정현성; 조선민주주의인민공화국 과학원통보 **5**, 17, 주체108(2019).

[2] A. Zhao et al.; Journal of Machine Learning Research, **17**, 213, 1, 2016.

[3] Xin Tong et al.; arXiv:1802.02557v1 [stat.ME], 7 Feb 2018.

[4] Xin Tong et al.; arXiv:1802.02557v3 [stat.ME], 16 Jun 2018.

주체109(2020)년 9월 5일 원고접수

Simulation Comparisons for the Estimation of the Generalized Neyman-Pearson Discriminant Rule

Jong Hyon Song, Rim Chang Ho

In this paper, we demonstrated the estimation of the generalized Neyman-Pearson discriminant rule with sample data for the several populations by simulation experiments.

Keyword: Neyman-Pearson discriminant rule