2021.04.07.
김은영

제1장 기초통계이론

1.4 통계적 추정과 검정

1.4.1 추정량의 종류와 성질

1. 추정량(estimator)의 정의

(1) 정의

확률변수 $Y_1, Y_2, \cdots, Y_n$들의 함수가 미지의 모수 θ를 추정하는 공식으로 사용되는 경우

(2) 표기: $\hat{\theta}\left(Y_1, Y_2, \cdots, Y_n\right)$ 또는 $\hat{\theta}$

2. 추정량의 성질

(1) 불편성(unbiasedness)

$$\hat{\theta}\text{의 편의: } Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$$Bias(\hat{\theta}) = 0 \rightarrow \hat{\theta}\text{: } \theta\text{의 불편 추정량(unbiased estimator)}$$

(2) 일치성(consistency)

$\lim_{n \to \infty} P(|\hat{\theta} - \theta| < \epsilon) = 0 \rightarrow \hat{\theta}$: θ의 일치 추정량(consistent estimator)

때때로 $\hat{\theta}$은 확률적으로 θ에 수렴합니다.

(3) 최소분산성(minimum variance)

$Var(\hat{\theta}) \leq Var(\hat{\theta}^*)$, 여기서 $\hat{\theta}^*$: θ의 다른 추정량 → $\hat{\theta}$은 최소분산을 가진다.

3. 모수 추정법

(1) 최소제곱법(method of least squares)

$$Y_i = \theta + \epsilon_i, \quad i = 1, 2, \cdots, n, \quad \epsilon\text{: 오차항}$$

$\hat{\theta}$이 다음 조건을 만족하면 ‘최소제곱 추정량(Least Squares Estimator: LSE)’이다.

$$\min \sum_{i=1}^{n} \left(Y_i - \theta\right)^2 = \min \sum_{i=1}^{n} \epsilon_i^2 \rightarrow \hat{\theta} = \underset{\theta}{\arg} \min \sum_{i=1}^{n} \epsilon_i^2$$

(2) 최대우도 추정량(Maximum Likelihood Estimator: MLE)

$Y_1, Y_2, \cdots, Y_n$: $pdf \;\; f(y|\theta)$로부터의 확률표본

우도함수(likelihood function): $\mathscr{L}(\theta | Y_1, Y_2, \cdots, Y_n) = \prod_{i=1}^{n} f(y_i|\theta)$

로그우도함수: $\ell(\theta) = \log \mathscr{L}(\theta|y)$

θ의 최대우도 추정량: $\hat{\theta}_{MLE} = \underset{\theta}{\arg} \mathscr{L}(\theta|y)$

로그함수: 단조증가함수 → $\hat{\theta}_{MLE} = \underset{\theta}{\arg} \max \ell(\theta)$

[예제]

$Y_1, Y_2, \cdots, Y_n$: $P(\lambda)$로부터의 확률표본

$$\mathscr{L}(\lambda|y) = \prod_{i=1}^{n} f(y_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-n\lambda}\lambda^{\sum y_i}}{y_1! \cdots y_n!}$$

$$\ell(\lambda) = \log \mathscr{L}(\lambda|y) = -n\lambda + \sum_{i=1}^{n} \log \lambda - \sum_{i=1}^{n} \log y_i!$$

$$\frac{\partial l(\lambda)}{\partial \lambda} = -n + \sum_{i=1}^{n} y_i \frac{1}{\lambda} = 0 \;\rightarrow\; \lambda = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{Y} \Rightarrow \hat{\lambda}_{MLE} = \overline{Y}$$

[예제]

$Y_1, Y_2, \cdots, Y_n$: $U(0, \theta)$로부터의 확률표본, θ의 MLE는?

$$f(y_i|\theta) = \frac{1}{\theta}, \qquad\qquad (0 \le y_i \le \theta)$$

$$= \frac{1}{\theta} I(0 \le y_i \le 1)$$

$$\mathscr{L}(\theta|y) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 \le y_i \le \theta)$$

$$= \theta^{-n} I(0 \le y_1 \le \theta) I(0 \le y_2 \le \theta) \cdots I(0 \le y_n \le \theta)$$

$$= \theta^{-n} I(0 \le y_{(n)} \le \theta), \qquad\qquad y_{(n)} = \max(y_1, \cdots, y_n)$$

$$y_{(n)} = \underset{\theta}{\arg} max \mathscr{L}(\theta|y)$$

< 최대우도 추정량의 성질 >

① 점근적 정규성(asymptotic normality)

$$\sqrt{n}\left(\hat{\theta}_{MLE}-\theta\right) \rightarrow N\left(0, I^{-1}(\theta)\right)$$

여기서, $I(\theta)=E\left[\left(\dfrac{\partial \log f(y,\theta)}{\partial\theta}\right)^2\right]=-E\left[\left(\dfrac{\partial^2 \log f(y,\theta)}{\partial\theta^2}\right)\right]$

: Fisher Information Matrix (피셔정보행렬)

② 함수적 불변성(functional invariance)

$\hat{\theta}$: θ의 MLE → $g(\hat{\theta})$: $g(\theta)$의 MLE, 여기서 g: 잘 정의된 함수(well-defined function)

[예제]

$Y_1, Y_2, \cdots, Y_n$: $P(\lambda)$로부터의 확률표본

$$\mathscr{L}(\lambda|y)=\prod_{i=1}^{n} f(y_i|\lambda)=\prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{y_i}}{y_i!}=\frac{e^{-n\lambda}\lambda^{\sum y_i}}{y_1!\cdots y_n!}$$

$$\ell(\lambda)=\log\mathscr{L}(\lambda|y)=-n\lambda+\sum_{i=1}^{n}\log\lambda-\sum_{i=1}^{n}\log y_i!$$

$$\frac{\partial l(\lambda)}{\partial\lambda}=-n+\sum_{i=1}^{n} y_i\frac{1}{\lambda}=0 \rightarrow \lambda=\frac{\sum_{i=1}^{n} y_i}{n}=\overline{Y} \Rightarrow \hat{\lambda}_{MLE}=\overline{Y}$$

- $\overline{Y}$: λ의 MLE
- $\log(\overline{Y})$: $\log(\lambda)$의 MLE
- $\overline{Y}^2$: λ^2의 MLE

1.4.4 정규모집단에 대한 추론

$Y_1, Y_2, \cdots, Y_n$: $i.i.d.\ N(\mu, \sigma^2)$

- $\overline{Y} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$
- $\overline{Y}$, s^2: indep.
- $(n-1)\dfrac{s^2}{\sigma^2} \sim \chi^2(n-1)$

1. σ를 알 수 있을 때

$$Z = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

μ에 대한 $100(1-\alpha)\%$ 신뢰구간

$$P\left[-z_{\alpha/2} \le \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \le z_{\alpha/2}\right] = 1 - \alpha$$

$$P\left[\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

2. σ를 알 수 없을 때

$$T = \frac{\dfrac{\overline{Y} - \mu}{\sigma / \sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2} / (n-1)}} \sim t(n-1)$$

$$= \frac{\overline{Y} - \mu}{s / \sqrt{n}} \sim t(n-1)$$

여기서, $Z \sim N(0,1)$, $V \sim \chi^2(n-1)$, Z, V: 독립

μ에 대한 $100(1-\alpha)\%$ 신뢰구간

$$P\left[-t_{\alpha/2}(n-1) \le \frac{\overline{Y} - \mu}{s / \sqrt{n}} \le t_{\alpha/2}(n-1)\right] = 1 - \alpha$$

$$P\left[\overline{Y} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \le \mu \le \overline{Y} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

1.6 다중 검정(Multiple Testing)

1.6.0 통계적 가설검정

1. 가설검정

(1) 가설의 종류

H_0 : 귀무가설(null hypothesis) [예] $H_0 : \theta = \theta_0$

H_1 : 대립가설(alternative hypothesis) [예] $H_1 : \theta \neq \theta_0$

(2) 검정 통계량(test statistics): $T(y)$

(3) 기각역(rejection[critical] region): R

(4) 통계적 가설검정의 결론

① $T(y) \in R \rightarrow H_0$을 기각

② $T(y) \notin R \rightarrow H_0$을 기각할 수 없음

실제상태 / 가설검정의 결론	H_0이 참	H_1이 참
H_0을 기각할 수 없음	옳은 결론	잘못된 결론 TYPE Ⅱ ERROR
H_0을 기각	잘못된 결론 TYPE Ⅰ ERROR	옳은 결론

(5) 가설검정과 관련된 두 가지 오류

$\alpha = P_{H_0}(T \in R)$: 제1종 오류가 발생할 확률

$\beta = P_{H_0}(T \notin R)$: 제2종 오류가 발생할 확률

(6) 가설검정이란?

제1종 오류가 발생할 확률은 미리 정해진 α 수준보다 작거나 같도록 통제함면서[$P_{H_0}(T \in R) \le \alpha$] 검정력[$1-\beta$]을 최대화하는 것(→ β를 최소화하는 것)

1.6.1 다중 검정(Multiple Testing)

one-way classification [ANOVA]

$$Y_{ij} = \mu + a_i + \epsilon_{ij} \qquad \text{여기서, } i = 1, 2, 3$$
$$j = 1, 2, \cdots, n$$

replication / treatment	1	2	$\cdots$	n
1	Y_{11}	Y_{12}	$\cdots$	Y_{1n}
2	Y_{21}	Y_{22}	$\cdots$	Y_{2n}
3	Y_{31}	Y_{32}	$\cdots$	Y_{3n}

[Q] 세 개의 treatment 간에 진정한 차이가 있을까?

α_i : i번째 treatment effect $(i = 1, 2, 3)$

통계적 가설검정을 해 보자.

H_0 : $\alpha_1 = \alpha_2 = \alpha_3$ (→ 세 개의 treatment 간에는 진정한 차이가 없다.)

$$H_1: \text{not } H_0 \Rightarrow H_1: \begin{array}{l} \alpha_1 = \alpha_2 > \alpha_3 \\ \alpha_1 = \alpha_2 < \alpha_3 \\ \alpha_1 < \alpha_2 = \alpha_3 \\ \alpha_1 > \alpha_2 = \alpha_3 \\ \quad\vdots \\ \alpha_1 \neq \alpha_2 \neq \alpha_3 \end{array}$$

대안은?

H_{01} : $\alpha_1 = \alpha_2$ → 다중 검정(Multiple Testing)
H_{02} : $\alpha_1 = \alpha_3$
H_{03} : $\alpha_2 = \alpha_3$

1.6.2 제1종 오류

다음과 같은 다중 가설검정을 생각해 보자.

$H_{0i} : \mu_{1i} = \mu_{2i} \qquad (i = 1, \cdots, m)$

다음과 같은 사건 E_i를 생각해 보자.

$E_i = \{H_{0i}$를 기각$| H_{0i}$가 참$\} \qquad (i = 1, \cdots, m)$

→ 다중 검정에서 제1종 오류는

$$P\left(\bigcup_{i=1}^{m} E_i\right) = 1 - P\left(\bigcap_{i=1}^{m} E_i^C\right) = 1 - \prod_{i=1}^{m} P\left(E_i^C\right) \leq 1 - (1-\alpha)^m$$

↑ 상한

E_i들이 독립, $P(E_i) \leq \alpha$

[예제] $\alpha = 0.05$

$m = 2 \rightarrow 0.0975$
$m = 3 \rightarrow 0.1426$
$m = 100 \rightarrow 0.9941$

1.6.3 Family-Wise Error Rate (FWER)

가설검정의 결론 실제상태	H_0을 기각할 수 없음	H_0을 기각	
H_0이 참	U	V	m_0
H_1이 참	T	S	m_1
합계	$m-R$	R	m

FWER은 $P(V \geq 1) \leq \alpha$을 조정한다.
여기서, $V \geq 1$: 적어도 1개의 잘못된 결론이 도출되는 경우

1. Bonferroni Correction

$$P(V \geq 1) = P\left(\bigcup_{i=1}^{m} E_i\right) \leq \sum_{i=1}^{m} P(E_i)$$

↑

Bonferroni Inequality

Bonferroni는 각 귀무가설에 대한 유의수준을 $\alpha^* = \dfrac{\alpha}{m}$로 사용할 것을 제안했다.

< 한계점 >
m이 작으면, Bonferroni Correction: good
m이 크면, Bonferroni Correction: too bad

일반적으로 Bonferroni Correction은 매우 보수적이다.
→ 귀무가설을 기각하는 것이 매우 어렵다.

2. Sidak Procedure

$$P\left(\bigcup_{i=1}^{m} E_i\right) = 1 - P\left(\bigcap_{i=1}^{m} E_i^C\right) = 1 - \prod_{i=1}^{m} P\left(E_i^C\right)$$

$$= 1 - \{1 - P(E_i)\}^m = 1 - \left(1 - \alpha^*\right)^m \leq \alpha$$

$\Rightarrow \alpha^* = 1 - (1-\alpha)^{1/m}$
< 한계점 >
m이 작으면, Sidak Procedure: good
m이 크면, Sidak Procedure: not good

1.6.4 False Discovery Rate (FDR)

Benjamini & Hochberg (1995)

$$E\left(\frac{V}{R}\right) \le q$$

여기서, $\frac{V}{R}$: 잘못 기각한 귀무가설의 비율

q: 유의수준에 해당하는 미리 할당한 값

< 결론 >

m이 크면, FDR: good

m이 작으면, FWER, FDR: both are OK