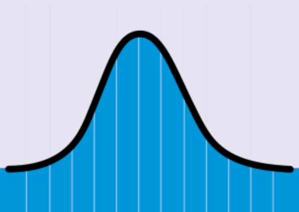
베이즈데이터분석 / 이재용 교수

04강

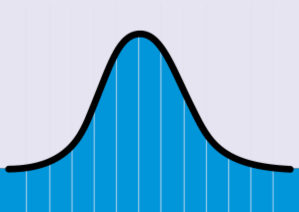
베이즈 가설 검정

한국방송통신대학교

목차



목차



가설검정 : 문제의 구조

가설

$\Theta = \Theta_0 \cup \Theta_1, \Theta_0 \cap \Theta_1 = \emptyset$

$H_0 : \theta \in \Theta_0$ 대 $H_1 : \theta \in \Theta_1$.

모형

$$x|\theta \sim f(x|\theta)$$

사후확률의 계산

$\alpha_0 = \pi(\Theta_0 | x), \alpha_1 = \pi(\Theta_1 | x)$

를 계산해보자.

사전분포

Θ 상에서의 사전분포 π는 다음과 같이 나타낼 수 있다.

$$\pi = \pi_0 \cdot g_0(\theta) + \pi_1 \cdot g_1(\theta).$$

여기서,

1. π_0는 H_0의 사전확률이고, π_1은 H_1의 사전확률이다.

$$0 \le \pi_0, \pi_1 \le 1, \pi_0 + \pi_1 = 1.$$

2. g_0는 H_0가 사실이라고 가정했을 때 사전분포이고,
   g_1은 H_1이 사실일 때 사전분포이다.
   다음과 같이 π를 이용해 나타낼 수 있다.

$$g_0(\theta) = \frac{\pi(\theta)}{\pi_0} I(\theta \in \Theta_0), \; g_1(\theta) = \frac{\pi(\theta)}{\pi_1} I(\theta \in \Theta_1)$$

몇 가지 용어들

사전 오즈비(prior odds ratio)

$$\frac{\pi_0}{\pi_1},$$

위에서 $\pi_0 = \pi(\Theta_0), \pi_1 = \pi(\Theta_1).$

사후 오즈비(posterior odds ratio)

$$\frac{\alpha_0}{\alpha_1} = \frac{\pi_0 \cdot m_0(x)}{\pi_1 \cdot m_1(x)} = \frac{\pi_0}{\pi_1} \times \frac{m_0(x)}{m_1(x)}$$

따라서

사후오즈비 = 사전오즈비 × 베이즈인자

와 같이 나타낼 수 있다.

베이즈인자(bayes factor)

$$B_{01} = \frac{m_0(x)}{m_1(x)}$$

$$B_{10} = \frac{m_1(x)}{m_0(x)} = \frac{1}{B_{01}}$$

제프리스의 기준(Jeffreys' criteria)

$\log_{10}\frac{\pi(H_1\|x)}{\pi(H_0\|x)}$	$\frac{\pi(H_1\|x)}{\pi(H_0\|x)}$	H_1에 대한 증거의 크기 (strength of the evidence for H_1)
$0 \sim \frac{1}{2}$	$1 \sim 3.2$	간단히 언급할 이상의 가치가 없다 (not worth than a bare mention)
$\frac{1}{2} \sim 1$	$3.2 \sim 10$	상당하다(substantial)
$1 \sim 2$	$10 \sim 100$	강하다(strong)
> 2	> 100	결정적이다(decisive)

목차

단순가설 대 단순가설(simple hypothesis)

가설

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta = \theta_1, \theta_0 \neq \theta_1$$

사전분포

$$\pi_0 = \pi\{\theta_0\}$$

$$\pi_1 = \pi\{\theta_1\}$$

단순가설 대 단순가설(simple hypothesis)

사후분포

$$\alpha_0 = \pi(\{\theta_0\}|x) = \frac{\pi_0 \cdot f(x|\theta_0)}{\pi_0 \cdot f(x|\theta_0) + \pi_1 \cdot f(x|\theta_1)}$$

$$\alpha_1 = 1 - \alpha_0$$

베이즈인자

$$B_{01} = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

예: 압정의 예

- 모형 $x|\theta \sim Bin(n = 10, \theta)$.
- $x = 7$을 관측했다고 하자.
- 다음의 가설 검정을 고려해보자.

$$H_0 : \theta = 1/2 \text{ vs } H_1 : \theta = 2/3.$$

- 사전확률

$$\pi_0 = \pi_1 = 1/2.$$

- 베이즈 인자와 가설들의 사후 확률을 구하자.

예: 정규 모형

- (모형)

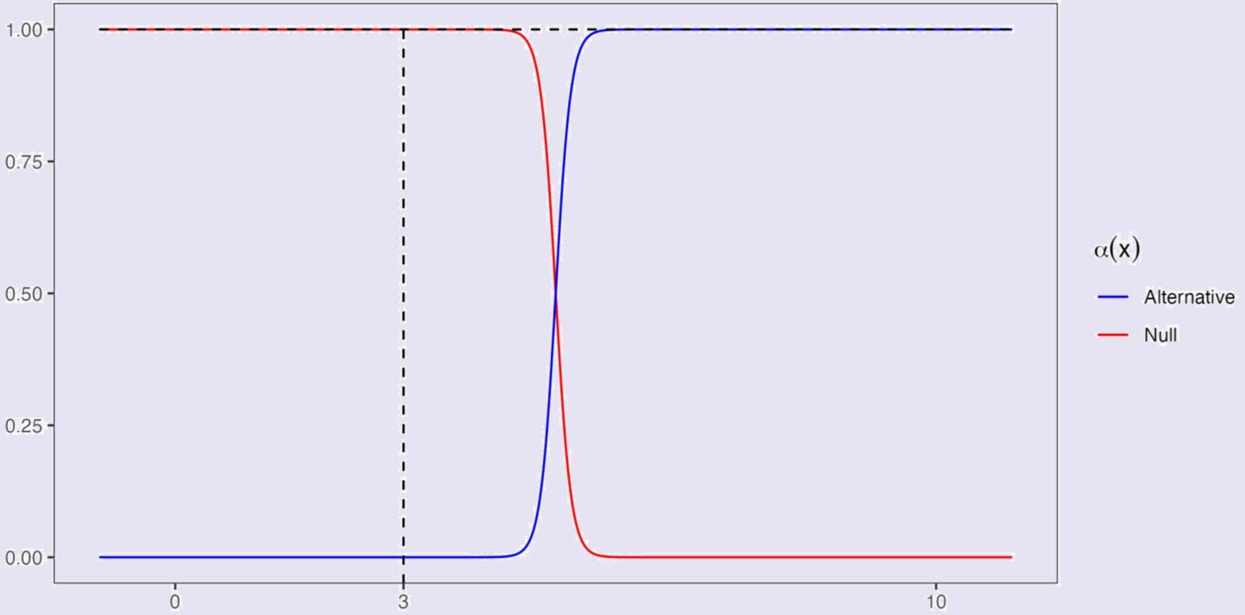
$$x|\theta \sim N(\theta, 1)$$

- (가설)

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta = 10$$

- 베이즈 인자와 가설들의 사후 확률을 구하자.

예: 정규 모형. 사후 확률의 그림



- **정규 모형의 가설 검정**

  $H_0 : \theta = 0$ vs $H_1 : \theta = 10$에서 x의 값에 따른 사후 확률들 $\alpha_0(x)$(빨간색)와 $\alpha_1(x)$(파란색)의 그림.

예: 정규 모형. 사후 확률의 그림

- $x = 3$을 관측하면, H_1의 사후 확률은 $\alpha_1(3) = 2.06 \times 10^{-9}$으로, H_1일 가능성이 거의 없고, H_0가 확실시 된다.

  하지만 유의성 검정의 경우

  $x = 3$의 p 값은 $\mathbb{P}(X \geq 3) = 0.00134$으로

  유의 수준 0.01에서 H_0를 기각한다.

  즉 유의성 검정에서는

  H_1이 유의한 증거가 있다고 결론을 내려서,

  베이즈 검정과 완전히 상반된 결론을 내린다.

목차

단순 가설 대 복합 가설(composite hypothesis)

가설

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

사전분포

$$\pi_0 = \pi\{\theta_0\}$$

$$\pi_1 = 1 - \pi_0 = \pi\{\theta \neq \theta_0\}$$

$g_1(\theta) : \Theta_1$ 상에서 사전밀도함수

$$\pi = \pi_0 \delta_{\theta_0} + \pi_1 g_1(\theta).$$

여기서, δ_{θ_0}은 θ_0에서 확률 1을 갖는

디락확률분포이다. 즉, $\delta_{\theta_0}(\{\theta_0\}) = 1.$

단순 가설 대 복합 가설(composite hypothesis)

사후확률

$$m_1(x) = \int_{\Theta_1} f(x|\theta) \cdot g_1(\theta) d\theta.$$

$$\alpha_0 = \frac{\pi_0 \cdot f(x|\theta_0)}{\pi_0 \cdot f(x|\theta_0) + \pi_1 \cdot m_1(x)}$$

$$\alpha_1 = \frac{\pi_1 \cdot m_1(x)}{\pi_0 \cdot f(x|\theta_0) + \pi_1 \cdot m_1(x)}$$

베이즈인자

$$B_{01} = \frac{f(x|\theta_0)}{m_1(x)}$$

예 : 정규모형 단순귀무가설(normal point null)

- (모형) $X_1, \cdot\cdot\cdot, X_n|\theta \sim \mathrm{N}(\theta, \sigma^2)$
- (가설) $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$
- (사전분포) $\pi_0 = \pi(H_0), \ \pi_1 = \pi(H_1).$

  $\theta|H_1 \sim N(\theta_0, \tau^2).\mathrm{i.e.}\, g_1(\theta) = N(\theta_0, \tau^2)$
- (베이즈인자)

$$B_{10} = (1 + \frac{n\tau^2}{\sigma^2})^{-\frac{1}{2}} e^{\frac{1}{2}z^2 \frac{1}{1+\frac{\sigma^2}{n\tau^2}}}$$

- (사후확률)

$$\alpha_0 = [1 + \frac{\pi_1}{\pi_0}(1 + \frac{n\tau^2}{\sigma^2})^{-\frac{1}{2}} e^{\frac{1}{2}z^2 \frac{1}{1+\frac{\sigma^2}{n\tau^2}}}]^{-1}$$

예 : 정규모형 단순귀무가설(normal point null)

사후확률의 극한 성질

- $|z| \uparrow \infty$로 가면, $B_{10} \longrightarrow \infty$ 이고 $\alpha_0 \to 0.$

- $|z| \downarrow 0$로 가면, $B_{10} \longrightarrow (1+\frac{n\tau^2}{\sigma^2})^{-\frac{1}{2}}$ 이고

$$\alpha_0 \to \frac{1}{1+\frac{\pi_1}{\pi_0}\cdot\frac{1}{\sqrt{1+\frac{n\cdot\tau^2}{\sigma^2}}}}$$

예 : 정규모형 단순귀무가설(normal point null)

참고

- 이 값은 $n, \tau^2, \sigma^2, \pi_0, \pi_1$이 고정되었을 때, z값의 변화로 얻을 수 있는 α_0의 최대값이다.

- $n \uparrow \infty$로 가면,

$$\left(1 + \frac{\pi_1}{\pi_0} \cdot \frac{1}{\sqrt{1 + \frac{n \cdot \tau^2}{\sigma^2}}}\right)^{-1} \uparrow 1$$

- $n \uparrow \infty$로 가면, $\alpha_0(z) \downarrow 0$, for $z \neq 0$.

제프리스의 역설 : 귀무가설의 사후확률과 p 값의 차이

- 정규모형의 가설검정 예에서 다음과 같이 정한다. Set $\tau^2 = \sigma^2$ and $\pi_0 = \pi_1 = \frac{1}{2}$

$$\alpha_0 = [1 + \frac{e^{\frac{1}{2}z^2\frac{n}{n+1}}}{\sqrt{1+n}}]^{-1}$$

z (p-value)	1	5	10	20	50	100	1000
1.645(0.1)	0.42	0.44	0.49	0.56	0.65	0.72	0.89
1.96(0.05)	0.35	0.33	0.37	0.42	0.52	0.60	0.80
2.576(0.01)	0.21	0.13	0.14	0.16	0.22	0.27	0.53
3.291(0.001)	0.086	0.026	0.024	0.026	0.034	0.045	0.124

제프리스의 역설 : 귀무가설의 사후확률과 p 값의 차이

제프리스의 파라독스(Jeffreys' paradox) 혹은 린들리의 파라독스(Lindley's paradox)

- n이 매우 크면 단순귀무가설일 때 p 값은 매우 작아서 유의성검정에서는 매우 강하게 귀무가설을 기각하나, 귀무가설의 사후확률이 1에 가까워서 베이지안 가설검정에서는 귀무가설이 거의 확실하다고 결론내리는 예들이 있다. 이와 같이 유의성 검정과 베이즈검정의 결과가 매우 다른 때, 제프리스의 역설혹은 린들리의 역설이라고 한다.

평균과 분산을 모르는 정규 모형의 가설 검정

- (모형)

$$x_1, \ldots, x_n | \mu, \sigma^2 \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$$

- (가설)

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0 \qquad (1)$$

- (제프리스-젤너-시오우 사전 분포)

$$\delta = \frac{\mu - \mu_0}{\sigma} \sim Ca(0,1) \qquad (2)$$

$$\pi(\sigma^2) d\sigma^2 \propto \frac{1}{\sigma^2} d\sigma^2$$

δ 의 분포는 다음과 동일하다.

$$\delta = \frac{\mu - \mu_0}{\sigma} \sim N\left(0, \sigma_\delta^2\right) \qquad (3)$$

$$\sigma^2 \sim Inv - \boldsymbol{X}_1^2$$

- 사후 확률의 계산은 R을 이용하는 것이 좋다.

예: 수면제 데이터

- (데이터) R의 데이터 sleep은 10명의 환자들에게 두 종류의 수면제를 투여하고 수면제를 투여하지 않았을 때에 비해 증가되는 수면 시간을 기록하였다. extra는 증가된 시간이고 ID는 환자의 번호, group은 수면제의 종류를 말한다.

$$x_1,\ldots, x_n|\mu, \sigma^2 \overset{i.i.d.}{\sim} N(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$$

- (모형)

X_i = 두번째 수면제의 수면 시간 − 첫번째 수면제의 수면 시간

$$x_1, \ldots, x_n|\mu, \sigma^2 \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$$

예: 수면제 데이터

- (가설)

$$H_0 : \mu = 0 \text{ vs } H_1 : \mu \neq 0$$

- 제프리스-젤너-시오우 사전 분포를 이용해서 베이즈 인자와 사후 확률을 구하라.

예측추론(predictive inference)

$$\theta \sim \pi(\theta)$$

$$x|\theta \sim f(x|\theta)$$

일 때, $z \sim g(z|\theta)$에 대해 예측하고자 한다.

이 때, x가 주어졌을 때, z의 예측분포(predictive density)는

$$p(z|x) = \int_{\Theta} g(z|\theta) \cdot \pi(\theta|x)\, d\theta.$$

이다.

다음시간

05 강

사전분포