최강력 검정

정의

- 개요

1. 가설검정을 실시할 때 무한개에 가까운 구간중에서 가장 효율적인 추정을 할 수 있도록 최선의 구간을 정하는 것
2. 즉, $H_0 : \theta \in W_0,\ H_1 : \theta \in W_1$ 이라고 할 때

1) 가설의 채택 기각을 이 분포의 범위 S에 기반할 때

(1) 범위 S 중 $\mathrm{X} \in C$ 이면 H_1 채택, $\mathrm{X} \in C^c$이면 H_0 채택이라는 규칙을 세울 때,

(2) 이 때 C를 기각역이라고 한다.

2) 이를 좀더 형식적으로 정의하면

(1) C가 기각역이고, $P_{\theta_0}[\mathrm{X} \in C] = \alpha$ 라고 할 때

(2) S 의 모든 임의의 부분집합 A에 대하여 $P_{\theta_1}[\mathrm{X} \in A]$ 를 결정하고

(3) 이 때, 모든 A에 대하여 $P_{\theta_1}[\mathrm{X} \in C] \geq P_{\theta_1}[\mathrm{X} \in A]$ 일때 이를 최강기각역이라고 한다.

정의

- 개요

3. 즉 이를 다시 정리하면

1) $f(x;\theta_0)$를 H_0에 해당하는 모수를 가지는 분포의 pdf라고 하자.

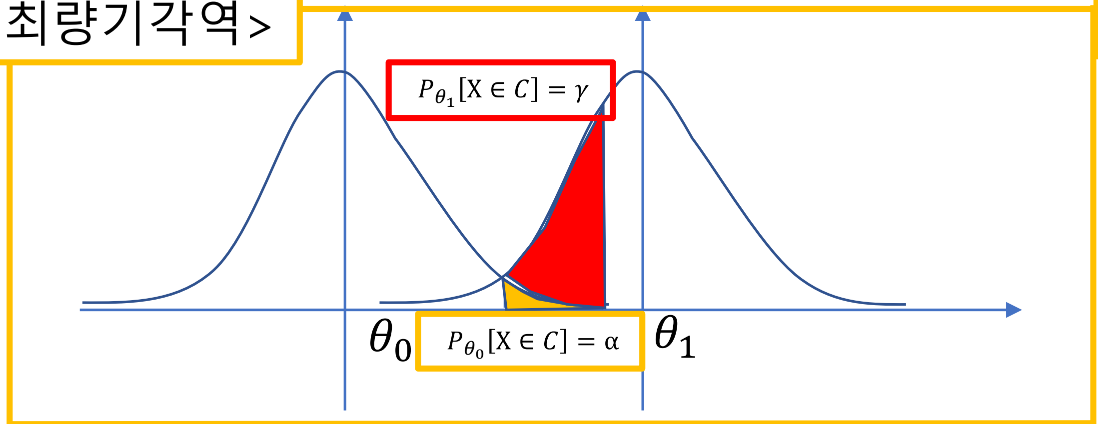
2) 이 H_0분포에서 확률 α를 가지는 부분집합 C는 매우, 무수히 많다.

3) 이때, 중요한것은 기회 비용이다.

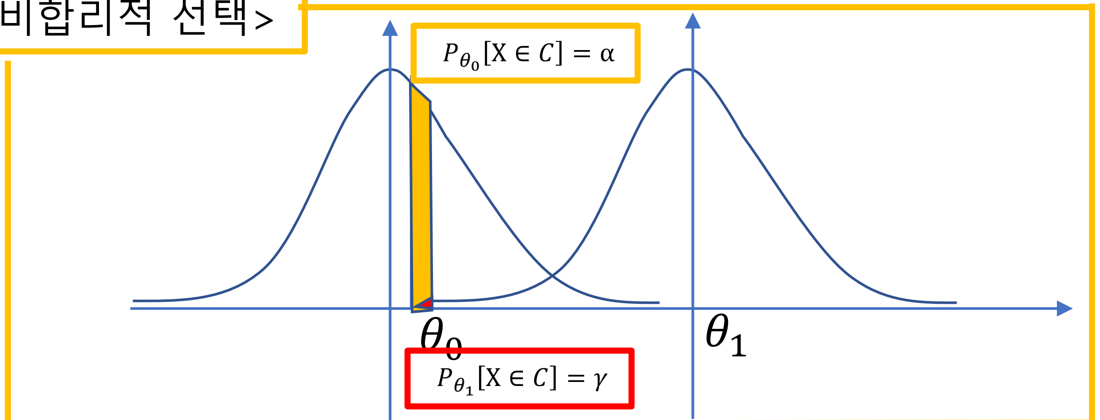
(1) H_0를 기각했는데 사실 H_0가 정답인 경우, 즉 2종 오류의 확률이 α로 동일하다면

(2) 그 중 검정력인 H_1이 정답일 확률이 같거나 클 경우를 선택하는것이 합리적

<최량기각역>



<비합리적 선택>



정의

- 네이만-피어슨 정리

1. 최량기각역을 정의하는 형식적이고 체계적인 방법론을 제공하는 정리
2. k 를 임의의 상수라고 하고
   $L(\theta_0; X) = \prod_{i=1}^{n} f(x_i;\ \theta)$, $L(\theta_1; X) = \prod_{i=1}^{n} f(x_i; \theta_1)$ 인 우도함수를 정의할 때
3. 다음의 경우, S의 부분집합 C를 가설 검정을 위한 최량기각역으로 정의한다.

1) $\frac{L(\theta_0;X)}{L(\theta_1;X)} \le k$, $X \in C$

2) $\frac{L(\theta_0;X)}{L(\theta_1;X)} > k$, $X \in C^c$

3) $P_{\theta_0}[X \in C] = \alpha$

정의

- 불편 검정(unbiased Test)

1. $H_0 : \theta \in W_0$, $H_1 : \theta \in W_1$ 의 검정이 존재하고, $\vec{X} = \begin{bmatrix} X_1 \\ \dots \\ X_n \end{bmatrix}$ 인 확률벡터라고 하자.

X가 pdf $f(\vec{X};\theta)$를 가진다고 할 때,

1) $\theta \in W_1$인 경우에 대하여 $P_{\theta \in W_1}\left[\vec{X} \in C\right] \geq \alpha$ 를 모든 경우에서 만족한다면

2) 이를 불편 검정이라고 한다.

2. 최강력검정과 불편검정의 관계

1) 최강력검정은 불편검정이다.

2) 즉, C를 최강력검정의 기각역이라고 가정하고, α를 이때의 유의수준이라고 한다면

(1) $\gamma(\theta_1) = P_{\theta_1}[\mathrm{X} \in C]$ 인 검정력 함수라고 정의할 때,

(2) 모든 경우에서 기각역을 C로 하는 검정은 불편검정이다.

정의

- 최강력 검정의 특징

1. 모수가 하나일 필요가 없다. 즉, 다중 모수의 경우에도 검정이 가능하다.
2. 가설 검정의 관심사가 굳이 모수의 차이 여부일 필요가 없다.
3. 즉, 단순히 $H_0 : \mathrm{g}(X_1, \dots, X_n)$ Vs $H_1 : \mathrm{h}(X_1, \dots, X_n)$ 를 검증하는 가설이라면

1) $\frac{\mathrm{g}(X_1,\dots,X_n)}{\mathrm{h}(X_1,\dots,X_n)} \leq k$ if $\mathrm{X} \in C$

2) $\frac{\mathrm{g}(X_1,\dots,X_n)}{\mathrm{h}(X_1,\dots,X_n)} > k$ if $\mathrm{X} \in C^c$

3) $P_{H_0}[\mathrm{X} \in C] = \alpha$

4. 위 조건을 만족하는 C는 최량기각역이 된다.

예제

- 최량기각역의 예시

1. N = 5, $p = \theta$인 이항분포를 따르는 X에서
   $H_0 : \theta = \frac{1}{2}$ VS $H_0 : \theta = \frac{3}{4}$라는 가설을 검증한다.

1) 이 때, 그 PMF를 나타내면

(1)

x	0	1	2	3	4	5
$f(x;\frac{1}{2})$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
$f(x;\frac{3}{4})$	$\frac{1}{1024}$	$\frac{15}{1024}$	$\frac{90}{1024}$	$\frac{270}{1024}$	$\frac{405}{1024}$	$\frac{243}{1024}$
$\frac{f\left(x;\frac{1}{2}\right)}{f(x;\frac{3}{4})}$	$\frac{32}{1}$	$\frac{32}{3}$	$\frac{32}{9}$	$\frac{32}{27}$	$\frac{32}{81}$	$\frac{32}{243}$

(2) 이 때, $P_{H_0}[X \in C] = \frac{1}{32}$ 일 때 C 는 $C = \{x; 0,5\}$ 이고,

예제

- 최량기각역의 예시

1. N = 5, p = θ인 이항분포를 따르는 X에서
   $H_0 : \theta = \frac{1}{2}$ VS $H_0 : \theta = \frac{3}{4}$라는 가설을 검증한다.

(2) 이 때, $P_{H_0}[\mathrm{X} \in C] = \frac{1}{32}$ 일 때 C 는 $C = \{x; 0{,}5\}$ 이고

$A_1 = \{x; 0\}$, $A_2 = \{x; 5\}$ 라고 할 때

(3) $P_{H_1}[\mathrm{X} \in A_1] = \frac{1}{1024}$ 이고 $P_{H_1}[\mathrm{X} \in A_2] = \frac{243}{1024}$ 이다.

2) 이 경우,

(1) $P_{H_1}[\mathrm{X} \in A_2] = \frac{243}{1024} > P_{H_0}[\mathrm{X} \in C] = \frac{1}{32}$ 이지만

(2) $P_{H_1}[\mathrm{X} \in A_1] = \frac{1}{1024} < P_{H_0}[\mathrm{X} \in C] = \frac{1}{32}$ 이므로

(3) A_2를 선택하는것이 합리적이다.

3) 한편, $\frac{f\left(x; \frac{1}{2}\right)}{f\left(x; \frac{3}{4}\right)}$만 놓고 봤을때, A_2를 기각역으로 놓는 경우만 1보다 작은 것을 확인할 수 있다.

(1) 네이만-피어슨 정리로 표현하면, 이 때 k=1이다.

예제

- 알려진 분포를 이용한 최량검정

1. $X_1, \ldots, X_n$이 N(θ ,1) 을 따르는 분포에서 추출한 확률표본이라고 하자.

2. 이 때 $H_0 : \theta = 0$VS $H_0 : \theta = 1$라는 가설을 검증한다.

3. 우도비 검정 형식을 빌려오면

1) $\frac{L(\theta_0;X)}{L(\theta_1;X)} = \frac{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i)^2}{2\sigma^2}\right)}{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i-1)^2}{2}\right)} = \exp\left[-\sum x_i + \frac{n}{2}\right] \leq k$

2) $\exp\left[-\sum x_i + \frac{n}{2}\right] \leq k$를 로그변환하면

(1) $-\sum x_i + \frac{n}{2} \leq logk = \sum x_i \geq \frac{n}{2} - logk$

(2) 한편, $X_1, \ldots, X_n$이 N(θ ,1)를 따른다고 했으므로, 이 때 $\frac{\sum x_i}{n} = \overline{x} \sim$ N(θ, $\frac{1}{n}$)을 따른다.

3) $P\left[\overline{x} \geq \frac{\frac{n}{2} - logk}{n}\right] = \alpha$ 에서

(1) $\frac{\frac{n}{2} - logk}{n}$ = c는 최량기각역이 되고

(2) 이 때, $\frac{\frac{n}{2} - logk}{n}$ = c 는 Z_α 를 따른다.