일반통계학
제 5장
통계적 추론

통계학 2016.1학기 정혜영

1.1 점추정

-모수를 하나의 값으로 택하는 것

(1) $X_1, \ldots, X_n$이 랜덤표본 일 때, $x_1, \ldots, x_n$ 은 관측값, θ는 모수(parameter)

(2) 추정량 : 미지의 모수 θ의 추정에 사용되는 통계량 $\hat{\theta} = \mathrm{t}(X_1, \ldots, X_n)$

추정치 : 추정량에 표본 관측값을 대입한 값 $\mathrm{t}(x_1, \ldots, x_n)$

(예) $\hat{\mu} = \bar{X}$(모평균 μ에 대한 추정량) , $\bar{x}$ (모평균 μ에 대한 추정치)

(3) 표준오차 : 추정량 $\hat{\theta}$의 표준편차 S.E($\hat{\theta}$) = $\sqrt{Var(\hat{\theta})}$

(4) 평균제곱오차 (mean square error, MSE)

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

1.2 좋은 점추정량의 성질

(1) 불편성 : $E(\hat{\theta}) = \theta$ 일 때, $\hat{\theta}$은 θ의 불편추정량 편중되지 않은 정도

(2) 효율성 : θ 에 대한 두 추정량 $\widehat{\theta_1}$과 $\widehat{\theta_2}$에 대해서 Var($\widehat{\theta_1}$) < Var($\widehat{\theta_2}$) 일 때,

$\widehat{\theta_1}$이 $\widehat{\theta_2}$보다 더 효율적인 추정량이 된다. 비교의 개념.

(3) 일치성 : $\lim_{n\to\infty} P\{|\widehat{\theta_n} - \theta| > \varepsilon\} = 0$ 일 때, $\widehat{\theta_n}$은 θ의 일치추정량

즉, 표본크기가 점차 커짐에 따라 점추정량의 값이 모수에 근접

-> 모수가 아닐 확률이 0으로 간다.

- 모평균 μ 의 점추정량 $\widehat{\mu}$ = 표본평균 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n} X_i$, S.E($\widehat{\mu}$)=$\frac{\sigma}{\sqrt{n}}$
- 모비율 p 의 점추정량 $\widehat{p}$=표본비율 $\bar{P}=\frac{X}{n}$, (X : 성공의 수~B(n, p))

최소불편분산추정량들

S.E ($\widehat{p}$)=$\sqrt{\frac{p(1-p)}{n}}$

- 모분산 σ^2의 점추정량 $\widehat{\sigma^2}$ = 표본분산 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

1 추정

- 불편성 증명

$E(\bar{X})=$ mu

$E(\bar{P})=$ p

$E(S^2)=$ sigma^2

- 불편추정량에 대한 MSE

$$\mathrm{MSE}(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

 점추정은 모수와 얼마나 가까이 있는지 알 수가 없다.

1.3 구간추정

- 모수 θ에 대하여 통계량 $L_\theta(X_1, \ldots, X_n)$ 과 $U_\theta(X_1, \ldots, X_n)$이 있어서

$$P(\theta \in (L_\theta, U_\theta)) = \mathrm{P}(L_\theta < \theta < U_\theta) = 1 - \alpha$$

일 때, 구간 (L_θ, U_θ)의 관측값 $(L_\theta(x_1, \ldots, x_n), U_\theta(x_1, \ldots, x_n))$를 θ의 $100(1-\alpha)\%$ 신뢰구간 또는 구간 추정값이라고 한다. 구간이 통계량이다.

- 신뢰수준 1-α는 신뢰구간을 여러 번 반복해서 얻을 때, 100(1-α)%의 신뢰구간들이 모수 θ를 포함함을 의미한다.
- 동일한 신뢰수준하에서 신뢰구간의 길이가 짧을 수록 더욱 정교한 구간추정이 된다.

  100번 실험하면 구간(통계량)이 100개 나오고, 이중 95개의 구간이 모수를 포함한다.

  신뢰구간100%는 (-inf, inf)이다;;

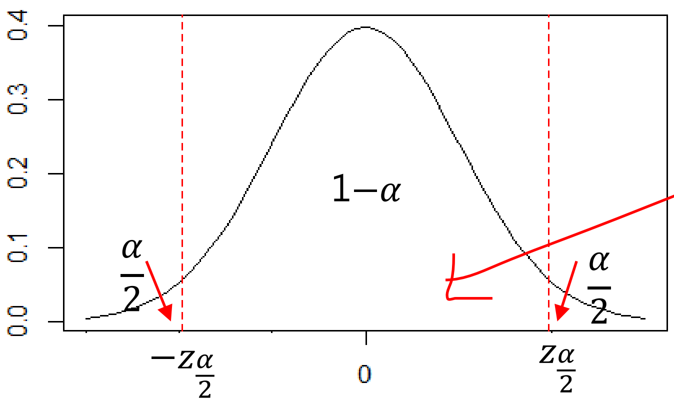
표본의 분포위에서 가운데는 모수(불편성),
동일한 신뢰수준에 신뢰구간이 짧기 위함. 따라서 대칭으로 만든다.

1.3 구간추정

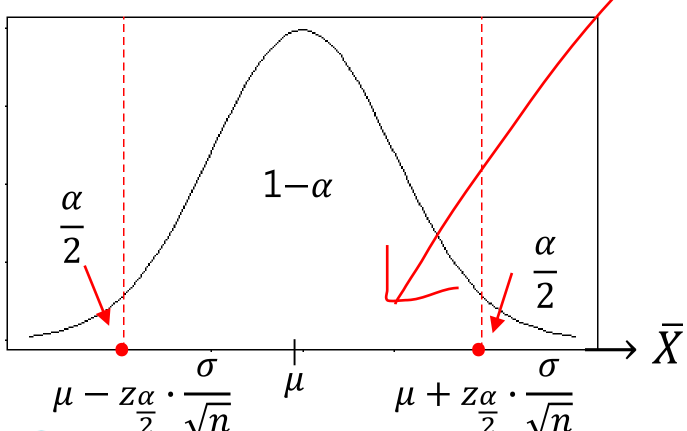
점추점에 대한 오차를 반영하는 것 이 구간추정이다.

- 모분산을 알 때 모평균 μ에 대한 $100(1-\alpha)\%$ 신뢰구간 유도

standard normal distribuion



Normal distribuion



$$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathrm{P}\left(\mu - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

(해석: 표본평균이 구간 사이에 존재할 확률)

$$\Leftrightarrow \mathrm{P}\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

(해석: 구간이 모평균 μ를 포함할 확률)

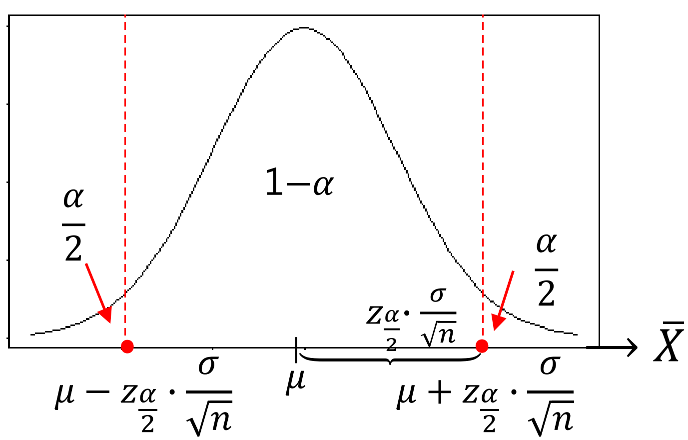
즉, 표본평균이 $1-\alpha$ 구간안에 존재하면 그 표본평균으로 구한 확률구간은 모평균을 포함하게 된다. => 우리가 원하는 구간추정의 의미를 갖게 됨.

1 추정 신뢰수준을 고정시키고 구간의 길이를 작게 하려면 n을 키우면 된다.

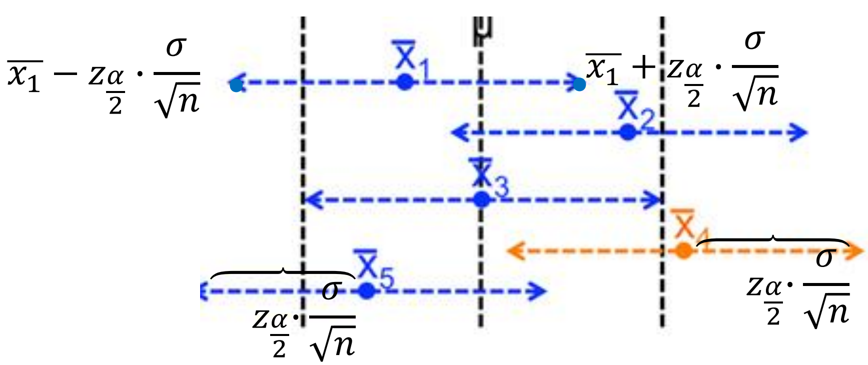
1.3 구간추정

- 모분산을 알 때 모평균 μ에 대한 100(1−α)% 신뢰구간=$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$

Normal distribuion



길이가 변하지 않는다.



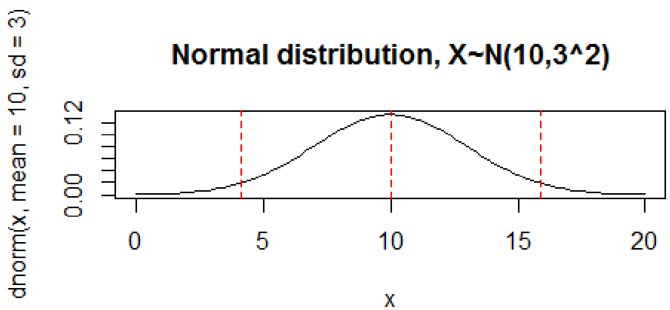
< 95% 신뢰구간의 의미 >

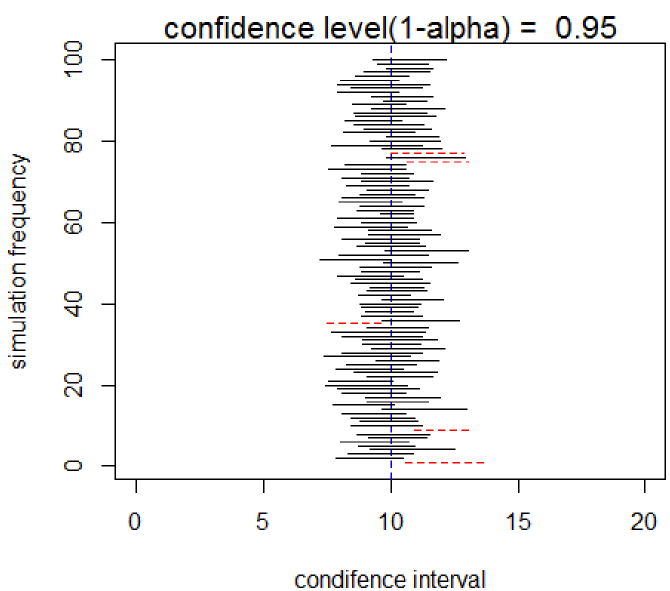
: 100번의 실험결과 표본평균 100개가 얻어지는데 이것으로 만든 신뢰구간 중 95개는 모평균을 포함하고 5개는 모평균을 포함하지 않는데 이 100개의 구간 중 1개를 의미.

(즉, 100개의 표본평균 중 95개는

$\left(\mu - z_{\frac{0.05}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \mu + z_{\frac{0.05}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right)$ 구간 안에 있고 5개는 구간 밖에 있는데 그 100개 중 1개의 표본평균을 얻었다.)

1.3 구간추정





- 모수는 우리가 모르긴 하지만 불변의 값인 반면, 신뢰구간은 표본에 따라 달라질 수 있으므로, '모수가 신뢰구간에 포함될 비율'이 아니라, 여러 번 표본을 뽑아 신뢰구간을 구하였을 때 '<u>신뢰구간이 모수를 포함할 비율</u>'을 고려하는 것 ⇒ 그 비율이 신뢰수준
- 100(1−α)% 신뢰수준을 갖는 신뢰구간

  $(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}})$: 이 구간이 모평균 μ를 포함한다고 신뢰할 만한 수준이 100(1−α)% 다.

- 오차한계 : $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ $(\underbrace{|\bar{x} - \mu|}_{\text{오차}} < z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$가 될 것이라 100(1−$\alpha$)% 신뢰할 만 한다.)

1.3 구간추정

- 표본크기 : 100(1−α)% 오차한계를 d 이하로 혹은 신뢰구간의 길이를 2d 이하로 하기 위한 표본의 크기는 (즉, $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le d$)

$$n \ge (z_{\frac{\alpha}{2}} \cdot \sigma/d)^2$$

- 표본의 크기를 늘이면 같은 신뢰수준하에서 오차한계를 줄일 수 있다. 즉, 신뢰구간의 길이를 줄여서 더 정교한 구간추정이 가능하다.

예제) σ=30일 때, 95% 오차한계를 5이하로 만들기 위해 필요한 표본크기는?

$$n \ge (z_{\frac{0.05}{2}} \cdot 30/5)^2 = 138.2976$$

설문지 응답자가 139명 이상은 되어야 설문지 결과로 얻은 표본평균의 값이 실제 모평균과 ±5이내에 있다고 95%정도 신뢰할 수 있다.

2.1 가설검정 용어 정리

-가설검정 : 표본으로부터 주어진 정보를 이용하여, 모수에 대한 예상, 주장 또는 단순한 추측 등의 옳고 그름을 판정하는 과정

(1) 가설 (hypothesis): 어떤 법칙을 설명하는 기술

• 귀무가설 (null hypothesis: H_0)
-반증을 찾기 위해 상정된 가설 또는 기존의 가설

• 대립가설 (alternative hypothesis: H_1)
- 자료로부터의 강력한 증거에 의하여 입증하고자 하는 가설
-양측가설과 단측가설이 있다.

예) 기존 공정에서 전구의 수명에 대한 평균이 120시간이고 표준편차가 10이다. 새로운 공법에 대하여 표본 25개를 뽑아 표본평균을 구해보니 124였다.

a. 새로운 공법이 평균을 증가시켰다고 말할 수 있는가?

$$H_0: \mu = 120 \;\; vs \; H_1: \mu > 120 \text{ (단측 가설)}$$

b. 새로운 공법과 기존의 공법의 평균 수명이 다르다고 할 수 있는가?

$$H_0: \mu = 120 \;\; vs \; H_1: \mu \neq 120 \text{ (양측 가설)}$$

(2) 가설검정 : 귀무가설의 반증에 대한 강도를 제공하여 귀무가설의 기각 여부를 판정하는 것.

(3) 검정통계량 (test statistic) : 가설검정에 사용되는 통계량, 검정하려는 모수의 점추정량이 되기도 하고, 이 점추정량을 표준화 한 것을 사용하기도 함

(4) 검정오류

	H_0 **참**	H_1 **참**
H_0 채택	옳은 결정	제2종 오류(α)
H_0 기각	제1종 오류(β)	옳은 결정

(4) 검정오류

- 제 1종 오류 (Type I error, α $=P(H_0$ 기각 | H_0가 참) : 검정의 **유의수준)**

: 귀무가설이 맞는데도 잘못하여 이를 기각하고 대립가설을 채택할 확률

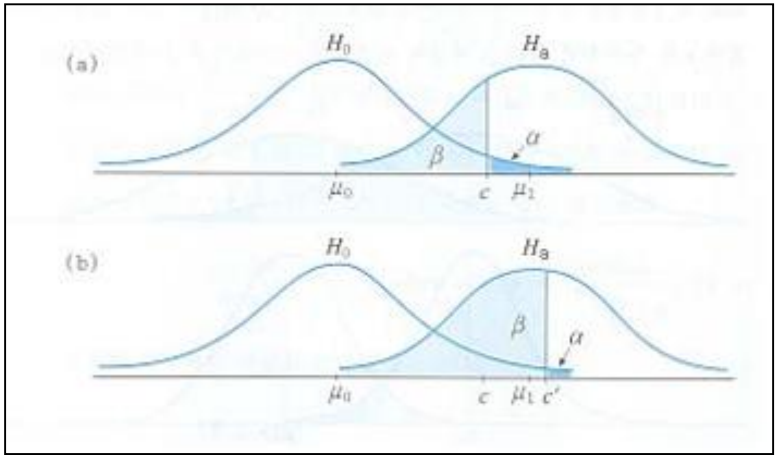
- 제 2종 오류 (Type II error, β= $P(H_0$ 채택 | H_1이 참))

: 대립가설이 사실임에도 불구하고 귀무가설을 채택하게 되는 확률

- 검정력 (1-β) $=P$ (H_1 채택 | H_1가 참)= $P(H_0$ 기각|H_0가 거짓)

: -대립가설이 참일 때 대립가설을 선택 (귀무가설이 거짓일 때 귀무가설 기각)

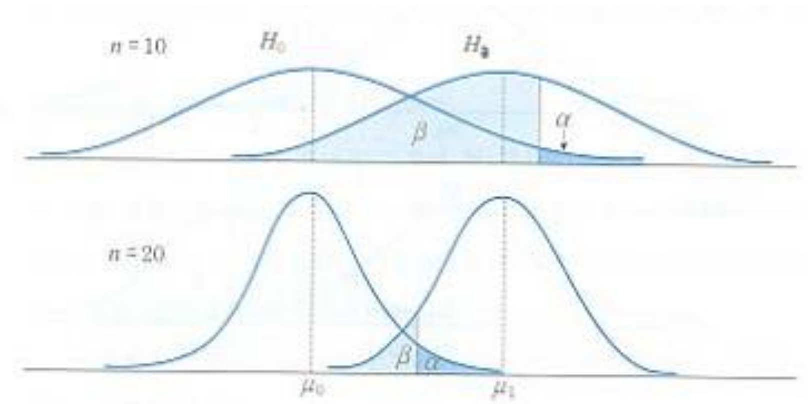
-검정력이 클수록 좋은 검정 방법임.



- 두 가지 검정오류인 α 와 β를 최소로 하는 기각역을 구하는 것이 최선 ⇒ 그러나 α 를 너무 작게 하려다 보면 β가 너무 커지는 모순관계가 있음.

- 이상적 목표는 α 와 β를 동시에 작게 하는 것이나, β는 수학적으로 다루기 매우 어렵다 ⇒ 실제문제에서는 주어진 α를 만족시키는 기각역 중에서 β를 최소로 하는 기각역을 선택 (=주어진 α 에 대하여 1-β를 크게 하는 검정을 선호)
- 검정통계량을 사용하여 검정하는 경우, 대립가설이 정의된 방향으로 기각역을 구하면 검정력이 가장 큰 최선의 기각역이 됨 (단측검정은 정의된 방향, 양측검증은 양쪽으로 기각역 구함)

<고정된 α에 대하여 표본수 증가에 따른 β의 변화>



- 표본의 수에 의해 두 검정오류를 동시에 줄이는 것이 가능하므로 표본수도 중요함.
- 적정한 검정오류를 만족시키는 표본의 수를 구하는 공식을 이용하여 적정한 표본수 산정 가능.

2.2 유의확률

- 귀무가설 H_0하에서 검정통계량이 실제 관측된 값보다 대립가설을 지지하는 방향으로 더욱 치우칠 확률