중앙값 69..? 최대 100

lower bound θ upper bound

→ 100번 했을때 몇번을 맞추는지? ex) 95번 ; 신뢰수준 95%.

$\hat{\theta}$ 좋이 좋아야 구간폭이 짧다 (분산↓)

7장 구간추정 (Interval Estimation)

상한.하한에 해당하는 통계량 : 점추정량 ± E

신뢰수준을 고정시켰을때 구간의 폭이 좁은게 좋음!

⇒ MLE 또는 최소분산비편향추정량

o 모수(parameter) θ ☞ 추정량(estimator): $\hat{\theta} = g(X_1, X_2, ..., X_n)$ 대신

$[L(X_1, X_2, ..., X_n),\ U(X_1, X_2, ..., X_n)]$ 형식의 추정 방법 사용

- $L(X_1, X_2, ..., X_n)$: lower bound
- $U(X_1, X_2, ..., X_n)$: upper bound

o 구간 $[L(X_1, X_2, ..., X_n),\ U(X_1, X_2, ..., X_n)]$ 으로 θ 추정

⇨ $P[L(X_1, X_2, ..., X_n) \le \theta \le U(X_1, X_2, ..., X_n)] = 1-\alpha$

↳ 신뢰수준을 만족시키며 구간의 폭이 가장 작은 지점

ex) 0.95일때가 0.99일때보다 구간이 더 좁음

: θ에 대한 $100(1-\alpha)$% **신뢰구간**

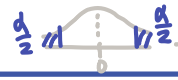
☞ $1-\alpha$: **신뢰계수**(confidence coefficient)

신뢰수준(confidence level) : $P[L \le \theta \le R] \ge 1-\alpha$

↳ 유의수준을 만족시키는 신뢰구간

ex) $X_1 \cdots X_n$ rs from $N(\mu, \sigma^2)$, 이때 $\theta = \mu$

$\hat\theta = \bar X$ 이용 → $\frac{\bar X - \mu}{\sigma/\sqrt n} \sim N(0,1)$: $\hat\theta$과 θ의 func.이면서 분포를알고있는 quantity



→ 무한히 많은 구간 중 구간폭이 최소이려면 좌우대칭

7장. 구간추정 (Interval Estimation)

⇒ $P[-z_{\alpha/2} \le \frac{\bar X - \mu}{\sigma/\sqrt n} \le z_{\alpha/2}] = 1-\alpha$ 또는 $P[\bar X - z_{\alpha/2}\frac{\sigma_0}{\sqrt n} \le \theta \le \bar X + z_{\alpha/2}\frac{\sigma_0}{\sqrt n}] = 1-\alpha$

NOTE: 바람직한 구간추정?

- 주어진 신뢰계수 $(1-\alpha)$일 때, **구간의 길이를 최소**로 하는 **구간추정** → 점추정이 좋아야함 (MVUE or MLE)

  ① 효율적인 추정량 $\hat\theta = g(X_1, X_2, \dots, X_n)$ 사용

  ② $\hat\theta = g(X_1, X_2, \dots, X_n)$의 분포를 이용해 신뢰계수를 만족하는 구간 설정

  ⇨ "pivotal quantity (중심 축량)" $\boxed{Q = h(\hat\theta, \theta)}$ 활용 → 추정량 & parameter로 이루어진 함수

  ≠ statistic

  ↳ 분포를 알아야함!

$X_1 \dots X_n$ rs from $N(\mu, \sigma^2)$, 이때 $\theta = \sigma^2$

(예) $Q = \frac{\bar X - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$ ⇨ $P[\bar X - E \le \mu \le \bar X + E] = 1 - \alpha$

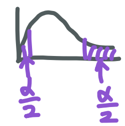
↳ 즉 $[\hat\theta - E, \hat\theta + E]$ 형태

$P[-t_{\frac{\alpha}{2}}(n-1) \le \frac{\bar X - \mu}{S/\sqrt n} \le t_{\frac{\alpha}{2}}(n-1)] = 1-\alpha$

+) 분산추정에서 χ^2 사용

▶ 일반적인 형태 ($\hat\theta$ 의 분포가 θ를 중심으로 좌우 대칭인 경우)

$$L(X_1, X_2, \dots, X_n) = \hat\theta - E \quad \& \quad U(X_1, X_2, \dots, X_n) = \hat\theta + E$$



→ 좌우대칭 X

구간의 길이가 최소는 아님

높이가 같으면서 면적의 합이 α인 지점을 찾는다.

※ **신뢰수준**(confidence level) : $(1-\alpha)$

$$P[L(X_1, \dots, X_n) \le \theta \le U(X_1, \dots, X_n)] \ge 1 - \alpha$$ ☞ 구간의 길이 최소화...?

7.1 평균에 대한 신뢰구간

$X_1, X_2, \cdots, X_n$ random sample from $N(\mu, \sigma_0^2)$: μ에 대한 구간추정

※ σ_0^2 : known (알고 있을 때)

▶ pivotal quantity : $Q = \dfrac{\overline{X} - \mu}{\sigma_0 / \sqrt{n}} \sim Z$ 이용

$$\Rightarrow P[\overline{X} - z_{\alpha/2}(\frac{\sigma_0}{\sqrt{n}}) \le \mu \le \overline{X} + z_{\alpha/2}(\frac{\sigma_0}{\sqrt{n}})] = 1 - \alpha$$

95%의 의미! → 100번 中 95번

95%의 확률로 맞는 추정이라고 신뢰

그 안에 있을 확률이 95%가 아님!

상·하한이 통계량이 아닌 관측값이므로 확률계산 불가

(참값은 저 구간 안에 있거나 아 없거나)

⇨ μ에 대한 100$(1-\alpha)$% 신뢰구간 : $[\overline{X} - z_{\alpha/2}(\frac{\sigma_0}{\sqrt{n}}), \overline{X} + z_{\alpha/2}(\frac{\sigma_0}{\sqrt{n}})]$

▶ 100$(1-\alpha)$% 의 의미?

☞ [예제 7.1-2] & 그림 7.1-1 (교재 p.323)

또는 n이 충.클

필수!

분자 → 표준정규분포 $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ 이용

① 중심극한정리 $\sim N(0,1)$

분모 → 카이제곱분포 $\frac{(n-1)S^2}{\sigma^2}$ 이용

→ t분포

7장. 구간추정 (Interval Estimation)

$X_1, X_2, \cdots, X_n$ random sample from $N(\mu, \sigma^2)$: μ 에 대한 구간추정

$\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}$

※ σ^2: unkwon (모르는 경우)

⇒ $\chi^2(n-1)$을 따르다 (정규가정필수!)

▶ pivotal quantity : $Q = \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$ 이용 ☜ **정규분포 가정**

⇒ 정규가정이 없는경우? t분포사용불가

⇨ μ에 대한 $100(1-\alpha)\%$ 신뢰구간 :

$$\left[\overline{X} - t_{\alpha/2,\,(n-1)}\left(\frac{S}{\sqrt{n}}\right),\ \overline{X} + t_{\alpha/2,\,(n-1)}\left(\frac{S}{\sqrt{n}}\right)\right]$$

② n이 충.클이면 $S^2 \to \sigma^2$으로 수렴

→ σ^2에 대한 일치추정량

→ 근사적으로 따르므로 S^2자리에 σ^2을 넣으면

근사적으로 $Q = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \mathrel{\dot\sim} N(0,1)$

o [예제 7.1-5] & [예제 7.1-6] (각자)

NOTE : 정규분포 가정이 없으면,

▶ pivotal quantity : $Q = \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \doteq Z$ 이용 ☜ **표본크기 n이 큰 경우**

"중심극한정리(CLT)"

→ 정규분포가정이 없는데 n이 작으면? 수리통계 수준이 아님.

(수) b10분야

변수변환으로 정규분포에 근사하게 만듦.

NOTE. $[L(X_1, X_2, \ldots, X_n),\ U(X_1, X_2, \ldots, X_n)]$ ☜ 양측 신뢰구간

ex) 지지율이 □%만 넘으면 된다

▶ **단측 신뢰구간** : $[L(X_1, X_2, \ldots, X_n) \le \theta]$ 또는 $[\theta \le U(X_1, X_2, \ldots, X_n)]$

ex) 발병률이 얼마이상인지 알고싶은경우

(예) $P[\overline{X} - z_{\alpha}(\frac{\sigma}{\sqrt{n}}) \le \mu\,] = 1-\alpha$

또는 $P[\,\mu \le \overline{X} + z_{\alpha}(\frac{\sigma}{\sqrt{n}})\,] = 1-\alpha$

※ **NOTE**

① $\overline{X}$ 를 사용함으로써 구간의 길이 최소화

② n 이 증가하면 구간의 길이 감소

Sample size 중요!

③ n 이 고정된 경우, $1-\alpha$ 감소하면 ⇒ 구간의 길이?

③ 구간 폭이 sample data(표본자료)에 따라 변할 수 있음

7.2 평균의 차이($\mu_x - \mu_y$)에 대한 신뢰구간

$X_1, X_2, \cdots, X_n$ random sample from $N(\mu_x, \sigma_x^2)$ A지역의 가계소득

$Y_1, Y_2, \cdots, Y_m$ random sample from $N(\mu_y, \sigma_y^2)$ B지역의 가계소득

→ 두 지역의 평균소득 차이는?

실제로는 평균을 모르는데 분산을 알지는 못함!

※ σ_x^2, σ_y^2 : known (알고 있을 때)

$\hat{\theta}$ (MLE) θ

▶ pivotal quantity : $Q = \dfrac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \sim Z$ → cov=0

n이 크면 S_x^2, S_y^2로 바꿀수있음

⇨ $(\mu_x - \mu_y)$에 대한 $100(1-\alpha)\%$ 신뢰구간 :

$$\left[(\overline{X} - \overline{Y}) - z_{\alpha/2}\sigma_w ,\ (\overline{X} - \overline{Y}) + z_{\alpha/2}\sigma_w \right], \quad \sigma_w = \sqrt{\sigma_x^2/n + \sigma_y^2/m}$$

$P(-z_{\alpha/2} \le Q \le z_{\alpha/2}) = 1-\alpha$

o [예제 7.2-1]

$\bar{X} \dot{\sim} N(\mu, \sigma_x^2)$ 이므로 정규가정이 없어도됨! ⇒ 근사적인 $100(1-\alpha)\%$ 신뢰구간
$\bar{Y} \dot{\sim} N(\mu, \sigma_y^2)$

[NOTE 1] σ_x^2, σ_y^2 unknown (모르는 경우) & n, m **이 큰 경우** **"중심극한정리(CLT)"**

▶ pivotal quantity : $Q = \dfrac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{S_x^2/n+S_y^2/m}} \dot{\sim} Z$ 이용

일치추정량 ; $S_x^2 \to \sigma_x^2$, $S_y^2 \to \sigma_y^2$

$$\frac{\dfrac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\dfrac{\sigma_x^2}{n}+\dfrac{\sigma_y^2}{m}}}}{\sqrt{\dfrac{\dfrac{(n-1)S_x^2}{\sigma_x^2}+\dfrac{(m-1)S_y^2}{\sigma_y^2}}{n+m-2}}} \sim t(n+m-2)$$

$\to N(0,1)$, $\to \chi^2(n+m-2)$

[NOTE 2] σ_x^2, σ_y^2 unknown (모르는 경우) & n, m **이 작은 경우**

☞ **가정** : $\sigma_x^2 = \sigma_y^2 = \sigma^2$ ※ **분산이 같지 않은 경우** (교재 p.331~332 참조)

▶ pivotal quantity : $Q = \dfrac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{S_p\sqrt{1/n+1/m}} \sim t_{(n+m-2)}$ 이용

정규분포가정필요!

여기서 합동추정량(pooled estimator) : $S_p^2 = \dfrac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2} = \dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{m}(Y_i-\bar{Y})^2}{n+m-2}$

↳ 등분산인 경우

o [예제 7.2-2]

11/21

$\theta = \mu_x - \mu_y = \mu_D \rightarrow E(D_i) = \mu_D$

$D_i = X_i - Y_i$

$\hat{\theta} = \frac{1}{n}\Sigma(X_i - Y_i) = \bar{D}$

약속!

$Var(D_i) = \sigma_x^2 + \sigma_y^2 - 2Cov(X_i, Y_i)$

$Var(X_i - Y_i) = \sigma_x^2 + \sigma_y^2$

$\rightarrow$ $Var(X_i - Y_i) > Var(D_i)$이므로 쌍을만들어 D_i를 쓰는게 좋음

■ 쌍 비교 (paired comparison) → 효과의 compounding을 막기위해

$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ random sample from **bivariate Normal dist.**

$\Rightarrow D_i = X_i - Y_i,\ D_1, D_2, \ldots, D_n$ random sample from $N(\mu_D, \sigma_D^2)$ ☜

$\hat{\sigma}_D^2 = S_D^2 = \frac{1}{n-1}\Sigma(D_i - \bar{D})^2$

▶ pivotal quantity : $Q = \dfrac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \sim t_{(n-1)}$ 이용

위의것과 완전 다름!

쌍비교는 독립X $\rightarrow$ $\sigma_D^2 = \sigma_x^2 + \sigma_y^2 - 2Cov(X_i, Y_i)$

평균차이는 독립 $\rightarrow$ $\sigma^2 = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$

⇨ $\mu_D = (\mu_x - \mu_y)$에 대한 100$(1-\alpha)$% 신뢰구간 :

$$\left[\bar{D} - t_{\alpha/2,\,(n-1)}\left(\frac{S_D}{\sqrt{n}}\right),\ \bar{D} + t_{\alpha/2,\,(n-1)}\left(\frac{S_D}{\sqrt{n}}\right)\right]$$

Q. 쌍 비교(paired comparison)를 하는 이유는? ⇨ $Var(\bar{D}) = \ldots$ 그게 분산이 더 작아서 효율적이니까!

[예제 7.2-4]

7.3 비율(P)에 대한 신뢰구간

$\hat{p} = \frac{1}{n}\sum X_i$

$X_1, X_2, \cdots, X_n$ random sample from Ber(p) & 표본크기 n이 큰 경우 →

▶ pivotal quantity : $Q = \dfrac{Y - np}{\sqrt{np(1-p)}} = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}} \doteqdot Z$ 이용

⇨ P에 대한 100($1-\alpha$)% 신뢰구간:

$P(-1.96 \le Q \le 1.96) \doteqdot 0.95$

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ \right]$$

o [예제 7.3-1] ~ [예제 7.3-3]

7.3-3 식

$P\left(\dfrac{|\hat{p}-p|}{\sqrt{\frac{p(1-p)}{n}}} \le 1.96\right)$을 근의공식으로 풂.

※ **NOTE** : 다른 형태의 신뢰구간 (p. 339 참조)

부등식 $\dfrac{|\hat{p} - p|}{\sqrt{p(1-p)/n}} \le z_{\alpha/2}$ ☞ 이차함수 부등식의 해(구간)

※ NOTE

$\hat{p}$가 매우 작아서 하한값이 음수가 될 수 있음!

1) $p \approx 0$인 경우 신뢰구간 ☞ 교재 p. 341

2) 모비율의 차이($p_1 - p_2$)에 대한 신뢰구간 ☞ p.342

E ; $p_1 - p_2$

V ; $\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$

→ n, m이 크면 정규근사

$$Q = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}} \doteq Z$$

▶ 표본크기 결정 문제

- (최대) 허용 오차를 만족시키는 표본크기의 결정 ☞ 7.4절

$\bar{X} \pm E$ 형태 → 같은 신뢰수준에서 더 짧은 신뢰구간을!

ex) $E = z_{\alpha/2} \frac{\sigma^2}{\sqrt{n}} = 100 \rightarrow n = ?$

$E = z_{\alpha/2} \frac{\sigma^2}{\sqrt{n}} = 50 \rightarrow n = ?$

▶ 분산(σ^2)에 대한 구간추정 등

↳ $\hat{\sigma}^2 = S^2$ (MVUE)

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

$\Rightarrow P\left(a \le \frac{(n-1)S^2}{\sigma^2} \le b\right) = 1-\alpha$인 a와 b를 찾아 σ^2에 대한 식으로 바꾸기

↳ χ^2분포는 좌우대칭 X

구간을 최소로 하는 값이 $\frac{\alpha}{2}$ 지점이 아님!

or