기초통계학II

2. 단일모집단의 기본

(1) 모평균에 대한 추론

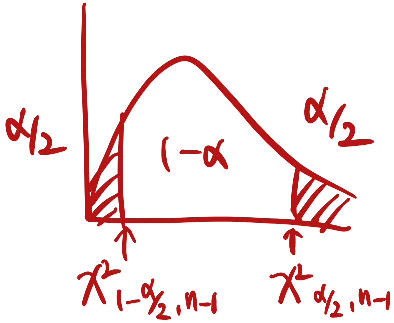
(2) 모분산에 대한 추론

(3) 모비율에 대한 추론

$S^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$

■ σ^2에 대한 추정

$X_1, X_2, \cdots, X_n \overset{i.i.d}{\longrightarrow} N(\mu, \sigma^2)$: 정규모집단 가정



점추정 : $\widehat{\sigma^2} = S^2$

어떤 분포?

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

○ σ^2의 $100(1-\alpha)\%$ 신뢰구간

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}} \right)$$

○ σ의 $100(1-\alpha)\%$ 신뢰구간

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}} \right)$$

■ σ^2에 대한 가설검정

① 가설

귀무가설	대립가설	
	$H_1 : \sigma^2 \neq \sigma_0^2$	양측검정(two-sided test)
$H_0 : \sigma^2 = \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	단측검정(one-sided test)
	$H_1 : \sigma^2 < \sigma_0^2$	

가설검정에서 범할수있는 오류의 max.

② 유의수준 α 정하기 : 0.05, 0.01

③ 검정통계량 귀무가설이 $H_0 : \sigma^2 = \sigma_0^2$ 인 경우 검정통계량

가설이 맞는지 틀린지 알기위해 표본으로부터 얻는것

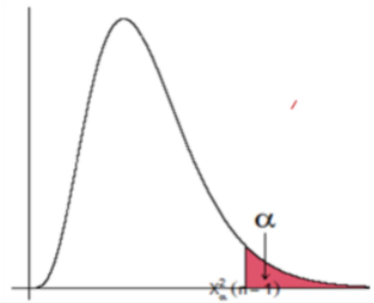
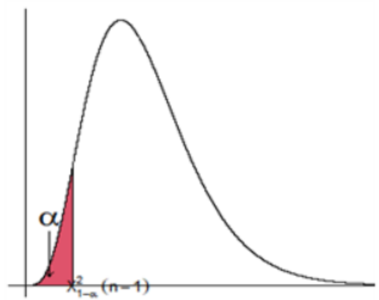
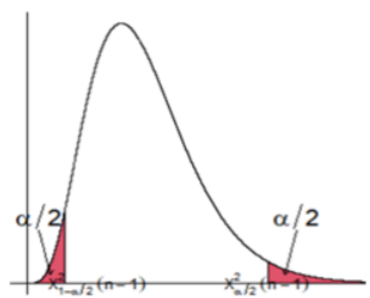
$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$$

점추정량의 분포를 알고있는 확률변수사용

중심축량

④ 기각역 : 채택/기각의 기준값

$$\boxed{\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}}$$

	대립가설		
	$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$
			
기각역	$R : \{\chi^2 \geq \chi^2_{\alpha}(n-1)\}$	$R : \{\chi^2 \leq \chi^2_{1-\alpha}(n-1)\}$	$R : \left\{\begin{matrix} \chi^2 \geq \chi^2_{\alpha/2}(n-1), \\ \chi^2 \leq \chi^2_{1-\alpha/2}(n-1) \end{matrix}\right\}$

표준편차가 작다.

(예제) 특정 지질형성에서 생긴 바위의 견본을 발굴해서 카드뮴 함유량을 알아보기 위해 화학분석을 하고자 한다. 광석의 질에 대한 지표로 광물질의 함유량의 균일함이 있다고 하자. 만약에 카드뮴 함유량의 표준편차가 4미만이면 광석의 질은 만족스러운 것으로 볼 수 있다고 한다. 25개의 견본을 대상으로 조사한 결과 카드뮴 함유량의 평균과 표준편차가 각각 10.2와 3.1로 나타났다고 한다.

$\hat{\sigma}^2 = 3.1^2$

(1) σ^2에 대한 98% 신뢰구간을 구하라.

$n=25,\ \chi^2_{0.01}(24)=42.98,\ \chi^2_{0.99}(24)=10.86$

σ^2의 98% 신뢰구간 : $\left(\dfrac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)},\ \dfrac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$

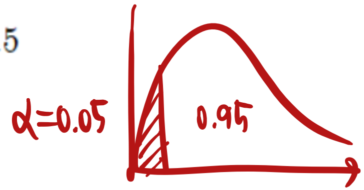
$$= \left(\frac{24\times(3.1)^2}{42.98}, \frac{24\times(3.1)^2}{10.86}\right) = (5.366, 21.238)$$

σ의 98% 신뢰구간 : $(\sqrt{5.366}, \sqrt{21.238}) = (2.316, 4.608)$

(2) 모표준편차 σ가 4보다 작다는 증거가 있는지를 결정하기 위한 가설을 검정하라.($\alpha = 0.05$)

① **가설:** $H_0 : \sigma = 4$ vs $H_1 : \sigma < 4$ ($H_0 : \sigma^2 = 16$, $H_1 : \sigma^2 < 16$)

② **검정통계치 :** $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2} = \dfrac{24 \times 3.1^2}{4^2} = 14.415$



③ **기각역** $R : \chi^2 \le \chi^2_{0.95}(24) = 13.85$

④ **결론 :** 검정통계량의 값이 기각역에 포함되지 않으므로 주어진 자료로부터 σ가 4보다 작다고 할 수 없다. 광석의 질이 만족스럽다고 할 수 없다.

(H_0 채택, H_1 기각)

□ 모비율

θ는 관심을 갖는 모비율, $\theta = \dfrac{\sum_{i=1}^{N} X_i}{N}$

- 표본크기가 큰 경우(**대표본**) $\boxed{n\theta \geq 5 \text{ and } n(1-\theta) \geq 5}$
- 베르누이 확률표본 $X_1, X_2, \ldots, X_n \sim$ iid $B(\theta)$, 성공이면 1, 실패면 0 → 베르누이 시행

$$f(x) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1, \quad 0 < \theta < 1$$

- 성공횟수 $X = X_1 + \cdots + X_n \sim Bin(n, \theta)$

$$\hat{\theta} = p = \frac{\sum_{i=1}^{n} X_i}{n}$$

○ 점추정량

- 모수 θ ⇦ $P = X/n$: 표본비율
  - 대부분의 교재에서는 θ는 p, P를 $\hat{p}$로 표시하고 있음

모비율 표본비율

기호주의

$p=\theta$, $\hat{p}=P$

○ 표본비율의 정규근사

- ○ 표본비율 $P = X/n = (X_1 + \cdots + X_n)/n$ 은 **표본평균**
- ○ $E(X_i) = \theta, Var(X_i) = \theta(1-\theta) \Rightarrow E(P) = \theta, Var(P) = \theta(1-\theta)/n$

$\frac{1}{n^2}(Var(X_1)+\cdots+Var(X_n))$

$=\frac{1}{n^2}(n \cdot \theta(1-\theta))$

- ○ n 이 큰 경우, 중심극한정리에 의해

$$P \simeq N\left(\theta, \frac{\theta(1-\theta)}{n}\right) \Rightarrow \frac{P-\theta}{\sqrt{\theta(1-\theta)/n}} \simeq N(0,1)$$

⊕ 이항분포이용

$X \sim Bin(n, \theta)$

$E(X) = n\theta$, $Var(X) = n\theta(1-\theta)$

$\rightarrow E(P) = \frac{1}{n} \cdot E(X) = \frac{1}{n} \cdot n\theta$

$Var(P) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} \cdot n\theta(1-\theta)$

- 정규근사는 n과 θ에 영향을 받음
  - ○ $n\theta \geq 5$ and $n(1-\theta) \geq 5$ 이면 적절

P(L<θ<U) = 1−α 가 되는 (L, U) 구하기

$\hat{\theta}$ = P 표준화 (정규분포 이용)

○ 구간추정

$$\circ \ 1-\alpha \simeq P\left(-z_{\alpha/2} < \frac{P-\theta}{\sqrt{\theta(1-\theta)/n}} < z_{\alpha/2}\right)$$

$$= P\left(P-z_{\alpha/2}\boxed{\sqrt{\frac{\theta(1-\theta)}{n}}} < \theta < P+z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}\right)$$

- 표준오차에 θ가 포함되어 있음($\Rightarrow$ θ의 추정량 P로 대체)

대표본이기 때문에 오차 영향 무시

$$\Rightarrow \left(P-z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}},\ P+z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}\right)$$

◉ 독일 Saxony 지역의 1889년 병원기록

○ 이 지역에서 출생한 73380명 중 아들은 38100명

○ 이 지역의 아들의 출생비율 θ에 대한 95% 신뢰구간

- $p = 38100/73380 = 0.519$

- S.~~D~~E. 추정값 $= \sqrt{\dfrac{p(1-p)}{n}} = \sqrt{\dfrac{0.519(1-0.519)}{73380}} = 0.00184$

표준오차

⇨ $(0.519 - 1.96 \times 0.0018,\ 0.519 + 1.96 \times 0.0018) = (0.5156,\ 0.5228)$

$p \pm z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$

○ 표본크기 결정

이번 조사는 M**가 R**에 의뢰해 지난 20일 전국 19세 이상 성인남녀 1천명을 전화설문한 결과로 신뢰수준 95%에 표본오차는 ±3.1%입니다.

표본크기를 조정해서 오차 수정가능.

- ○ 표본의 크기는 모수추정의 정확도 및 신뢰도에 영향을 줌
  - ① **신뢰수준** ⇦ 신뢰도
  - ② **오차범위** (오차: $\mathrm{P}-\theta$) ⇦ 정확도
- ○ $100(1-\alpha)\%$ 신뢰수준에서 허용오차범위가 $\pm\delta$일 때

표본비율 P
모비율 θ

$$1-\alpha = P\left(\underbrace{|P-\theta|}_{\text{오차}} < z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}} \right)$$

$$\Rightarrow \delta \geq z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}} \quad ⇨ \quad n \geq \left(\frac{z_{\alpha/2}}{\delta}\right)^2 \theta(1-\theta)$$

$$P\left(-z_{\alpha/2} < \frac{P-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} < z_{\alpha/2}\right) = 1-\alpha$$

○ θ는 과거 조사 기록, pilot survey 등으로 추정. θ에 대한 정보가 없는 경우 모든 θ에 대해 성립하도록 n을 결정

- $\theta = 0.5$ 일 때 $\theta(1-\theta)$가 가장 큼 ⇨ $n = \left(\dfrac{z_{\alpha/2}}{\delta}\right)^2 \dfrac{1}{4}$

n이 가장 크게

※ $\theta(1-\theta) = -(\theta - \frac{1}{2})^2 + \frac{1}{4}$

◉ 95%신뢰수준에서 오차범위가 $\pm 5\%$ $(\delta = 0.05)$

$$1.96\sqrt{\frac{\theta(1-\theta)}{n}} \le 0.05 \quad \Rightarrow \quad n \ge \left(\frac{1.96}{0.05}\right)^2 \theta(1-\theta)$$

⇨ $\theta = 0.5$일 때 $n = 384.16$이므로 최소한 385명을 추출해야 함

※ $n = 1{,}000$일 때, 표본오차는 $1.96\sqrt{\dfrac{0.5(1-0.5)}{1{,}000}} = 0.031$

○ 가설검정

간단히

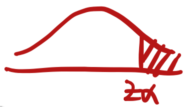
○ H_0 : $\theta = \theta_0$ vs H_1 : $\begin{cases} ① \ \theta > \theta_0 \\ ② \ \theta < \theta_0 \\ ③ \ \theta \neq \theta_0 \end{cases}$

○ **검정통계량**: $\boxed{Z = \dfrac{P - \theta_0}{\sqrt{\dfrac{\theta_0(1-\theta_0)}{n}}} \sim N(0,1)}$

점추정량(표본비율)의 표준화된 식.

○ 유의수준을 α라고 하면, 기각역 : 유의수준과 대립가설에 의해 위치결정

① $[z_{\alpha}, \infty)$ $\Leftrightarrow$ $z \geq z_{\alpha}$



$\alpha = 0.05$

② $(-\infty, -z_{\alpha}]$ $\Leftrightarrow$ $z \leq -z_{\alpha}$

③ $(-\infty, -z_{\alpha/2}]$, $[z_{\alpha/2}, \infty)$ $\Leftrightarrow$ $z \leq -z_{\alpha/2}$, $z \geq z_{\alpha/2}$

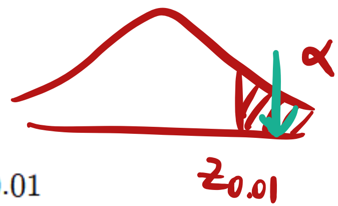
◉ 독일 Saxony Geissler 지역의 출생자료를 이용하여 그 당시 아들의 출생비율이 딸의 출생비율보다 높은지를 검정

- 가설: $\mathrm{H}_0 : \theta = 0.5$ vs $\mathrm{H}_1 : \theta > 0.5$
- 검정통계량

$$\frac{p-\theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$$

$$Z = \frac{P - 0.5}{\sqrt{0.5 \times 0.5 / 73380}} \simeq N(0,1)$$

- 1% 유의수준:

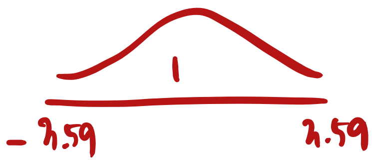


$$z = \frac{0.519 - 0.5}{\sqrt{0.5 \times 0.5 / 73380}} = 10.41 \geq 2.326 = Z_{0.01}$$

⇨ H_0 기각: 아들의 출생비율이 딸의 출생비율보다 높다.

유의확률 $z = 10.41$의 p-값: $\mathrm{P}(\mathrm{Z} \geq 10.41) \simeq 0$

부록표에 (−3.59, 3.59) 만있음.
여기를 벗어나면 0으로간주



예제1) 어떤 특정 암의 경우에 수술을 시행한 후 완치되는 비율(5년 이상 생존비율)이 30%라고 한다. 이 암에 걸린 60명의 환자를 대상으로 수술 뿐 아니라 수술 전후에 일정기간 방사선치료를 병행하였더니 60명 중 27명이 완치되었다고 한다.
이 자료로부터 수술만 하는 것보다 방사선 치료를 병행하는 것이 암의 완치율(p)을 높이는데 효과가 있다고 할 수 있는지 검정하라.(유의수준 5%)

$$\frac{p-\theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}}$$

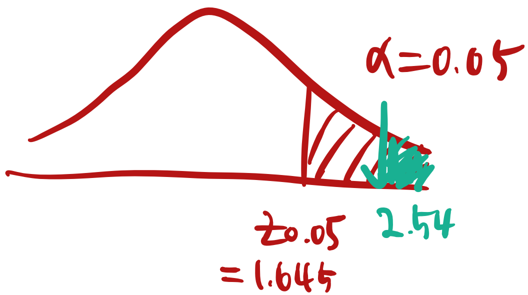
① **가설 :** $H_0 : \theta = 0.3\ , H_1 : \theta > 0.3$

표본비율

② **검정통계량의 값 :** $z = \dfrac{0.45 - 0.3}{\sqrt{0.3(1-0.3)/60}} = 2.535,\qquad P = \dfrac{27}{60} = 0.45$

③ **기각역 :** $R : \{Z \geq z_{0.05} = 1.645\}$

유의확률(p-값) : $P(Z \geq 2.54) = 0.0055$



④ 결론

(i) **기각역기준 :** 검정통계량의 값 z=2.535는 기각역 1.645보다 크므로 기각역에 속한다. 따라서 유의수준 5%에서 귀무가설을 기각할 수 있다. 즉, 이 자료에 의하면 유의수준 5%에서 수술만 하는 것보다 방사선 치료를 병행하는 것이 암의 완치율(θ) 을 높이는데 효과가 있다고 할 수 있다.

(ii) **유의확률기준:** p값이 0.0055로 유의수준 0.05보다 작으므로 귀무가설을 기각할 수 있다. 즉, 이 자료에 의하면 유의수준 5%에서 수술만 하는 것보다 방사선 치료를 병행하는 것이 암의 완치율(θ)을 높이는데 효과가 있다고 할 수 있다.

=> 유의수준 0.0055까지 귀무가설을 기각할 수 있다.