기초통계학II

3. 두 모집단의 비교

(1) 평균 비교

- 독립표본, 짝표본

(2) 분산 비교

(3) 비율 비교

$\hat{\theta} = P = \frac{\Sigma X_i}{n}$ $\quad (\Sigma X_i \sim B(n, \theta))$

모비율은 대표본인 경우만 고려한다 $\sim N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$

◈ 모비율 추론에 대한 복습

(1) θ의 100(1-α)% 신뢰구간

$\sim N(0,1)$

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

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

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

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

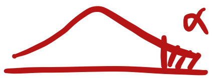
◈ 모비율 추론에 대한 복습

(2) 가설검정

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

○ 유의수준을 α라고 하면, 기각역



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

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

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

[예제] A 전자회사는 판매중인 노트북의 고객 만족도를 알기위해, 노트북을 사용 중인 고객 200명에게 만족 (1)과 불만족 (0)으로 설문 조사하였다. 이중 150명의 고객이 만족한다고 답변하였다.

(1) 전체 고객의 만족도 θ의 점 추정치를 구하라.

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{150}{200} = 0.75$$

(2) 모집단 전체의 만족도 θ의 99% 신뢰구간을 구하라.

($n = 200$, $\hat{\theta} = 0.75$, $z_{0.005} = 2.575$)

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

$$0.75 - 2.575\sqrt{0.75 \times 0.25/200} < \theta < 0.75 + 2.575\sqrt{0.75 \times 0.25/200}$$

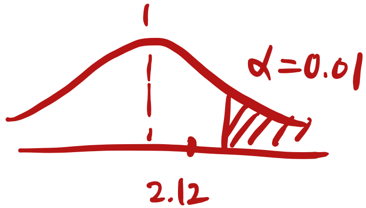
$$\Rightarrow \quad 0.67 < \theta < 0.83$$

(3) A 전자회사의 과거 조사에 의하면 이전 모델의 고객 만족도는 68%였다. 현재 모델의 고객 만족도가 이전 모델에 비해 향상되었는지, 귀무가설과 대립가설을 세우고, 200명 고객의 설문조사 결과를 바탕으로 유의수준 $\alpha = 0.01$로 검정하라.

- 가설 : $H_0 : \theta = 0.68 \;\; v.s. \;\; H_1 : \theta > 0.68$

- 검정통계치 : $z = \dfrac{p - \theta_0}{\sqrt{\dfrac{\theta_0(1-\theta_0)}{n}}} = \dfrac{0.75 - 0.68}{\sqrt{\dfrac{(0.68)(0.32)}{200}}} = 2.12$

검정통계량의 값



- 결론 : 이 값은 $z_{0.01} = 2.33$보다 작으므로 유의수준 1%에서 귀무가설을 기각할 수 없다. 즉, 현재 모델의 고객 만족도가 이전 모델에 비해 향상되었다고 할 수 없다.

- 유의확률 : P-값 = $P(Z \geq 2.12) = 0.017$ > 0.01 → H_0 기각불가능

+) $\alpha = 0.05$일때는 H_0 기각가능!
유의수준에따라 결론이달라짐 → 결론작성시 유의수준제시하는것이 좀더 정확하다.

(4) 노트북 사용자 전체의 만족도 θ를 추정하고자할 때, 99%의 확신으로 추정의 오차가 0.025을 넘지 않게 하려면, 몇 명의 고객을 설문 조사하면 되는가?

<풀이> $\hat{\theta}=0.75$를 사전 정보로 이용

$$n \geq \left(\frac{z_{\alpha/2}}{d}\right)^2 \hat{\theta}(1-\hat{\theta}) = \left(\frac{2.575}{0.025}\right)^2 0.75 \times 0.25 = 1989.2$$

$n=1990$ 이상 조사하면 99%의 확신으로 추정의 오차가 0.025을 넘지 않게 되는 것을 보장

■ 두 모집단의 비교

○ 모집단 비교의 예

① A반과 B반의 기초 통계학 시험결과 비교

② 도시와 시골의 소득비교

③ 운동전과 운동 후의 폐활량 비교

④ 쌍둥이를 대상으로 다른 diet방법을 실시한 후 효과의 비교

◦ ① ② : 별개의 두 집단 비교 ⇨ 독립표본

◦ ③ ④ : 쌍을 이룬 집단의 비교 ⇨ 짝비교(대응표본)

전, 후

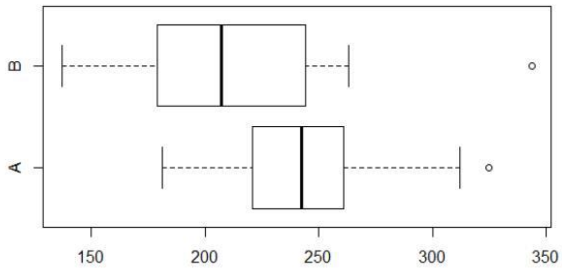
□ 평균비교 - 독립표본

▣ Western Collaborative Group Study

○ 1960-61년 미국 California에서 3,154명의 중년남성 중 행동타입에 따라 분류된 40명의 콜레스테롤을 측정

반대

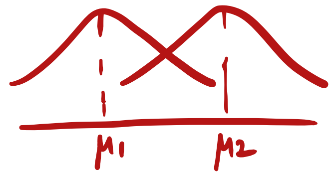
Type A -urgency -aggression -ambition	233 291 312 250 246 197 268 224 239 239 254 276 234 181 248 252 202 218 212 325
Type B -relaxed -non-competitive -less hurried	344 185 263 246 224 212 188 250 148 169 226 175 242 252 153 183 137 202 194 213



Q? Type A가 Type B보다 콜레스테놀이 높다고 할 수 있는가?

○ 가정 I

- $X_1,\ldots,X_m \sim$ iid $N(\mu_1,\sigma^2)$, $Y_1,\ldots,Y_n \sim$ iid $N(\mu_2,\sigma^2)$
- $X_1,\ldots,X_m$와 $Y_1,\ldots,Y_n$는 서로 독립 (공통분산 σ^2 unknown)



○ 점추정

- 모수: $\mu_1 - \mu_2$ ⇦ $\bar{X} - \bar{Y}$: 표본평균의 차

분산이다를경우?

○ 표본평균의 차 $\bar{X} - \bar{Y}$의 성질

- $E(\bar{X}) = \mu_1$, $E(\bar{Y}) = \mu_2$
- $Var(\bar{X}) = \sigma^2/m$, $Var(\bar{Y}) = \sigma^2/n$

모평균의 차이만 고려할수 없음

- $\overline{X} \sim N(\mu_1, \sigma^2/m)$, $\overline{Y} \sim N(\mu_2, \sigma^2/n)$, $\overline{X}$와 $\overline{Y}$는 독립

$$\Rightarrow \overline{X} - \overline{Y} \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right)$$

$E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y})$
$= \mu_1 - \mu_2$
↳ 불편추정량 $\overline{X} - \overline{Y}$

$Var(\overline{X} - \overline{Y})$
$= Var(\overline{X}) + Var(\overline{Y})$

- 표준화: $Z = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{1/m + 1/n}} \sim N(0, 1)$

- σ^2의 추정

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

- $S_X^2 = \dfrac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2$, $S_Y^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

- $S_p^2 = \dfrac{\sum(X_i - \overline{X})^2 + \sum(Y_i - \overline{Y})^2}{m+n-2} = \dfrac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$

↳ 합동표본분산(pooled sample variance)

자유도: $(m-1) + (n-1) = m+n-2$

○ 중심축량

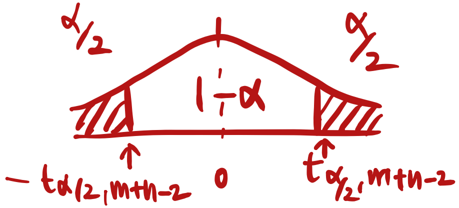
$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

추정이므로 t분포

- 각 모집단에 대한 정규성 확인
  - Jarque-Bera test, Shapiro-Wilk test, ...

    + 히스토그램, Q-Q plot

- 등분산성 확인
  - $\mathrm{N}(\mu_1, \sigma_1^2)$ 이고 $\mathrm{N}(\mu_2, \sigma_2^2)$ 일 때 $\sigma_1^2 = \sigma_2^2$ 인지를 검정
  - 간이검정: $0.5 < s_{\mathrm{X}}/s_{\mathrm{Y}} < 2$ 이면 $\sigma_1^2 = \sigma_2^2$ 를 만족



○ 구간추정

- $100(1-\alpha)\%$ 신뢰구간

$$
\begin{aligned}
1-\alpha &= P(-t_{\alpha/2,m+n-2} < T < t_{\alpha/2,m+n-2}) \\
&= P\left(-t_{\alpha/2,m+n-2} < \frac{\overline{X}-\overline{Y}-\boxed{(\mu_1-\mu_2)}}{S_p\sqrt{1/m+1/n}} < t_{\alpha/2,m+n-2}\right) \\
&= P\Bigg(\overline{X}-\overline{Y}-t_{\alpha/2,m+n-2}S_p\sqrt{\frac{1}{m}+\frac{1}{n}} < \\
&\qquad \mu_1-\mu_2 < \overline{X}-\overline{Y}+t_{\alpha/2,m+n-2}S_p\sqrt{\frac{1}{m}+\frac{1}{n}}\Bigg)
\end{aligned}
$$

$$
\Rightarrow \left(\overline{X}-\overline{Y}-t_{\alpha/2,m+n-2}S_p\sqrt{\frac{1}{m}+\frac{1}{n}},\ \overline{X}-\overline{Y}+t_{\alpha/2,m+n-2}S_p\sqrt{\frac{1}{m}+\frac{1}{n}}\right)
$$

$s_X^2 = \frac{1}{m-1}\Sigma(x_i-\bar{x})^2 = \frac{1}{m-1}\left(\Sigma x_i^2 - n\bar{x}^2\right)$

Type A를 X, Type B를 Y 라고 하면

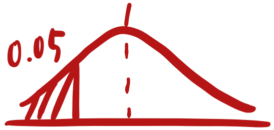
- $\bar{x} = 245.05$, $\bar{y} = 210.30$, $\sum x_i^2 = 1226495$, $\sum y_i^2 = 928920$
- $s_X = 36.64$, $s_Y = 48.34$ ⇨ $s_X / s_Y = 0.758$ ⇒ 등분산성 만족

$\hat{\sigma}^2 \doteq s_p^2 = \dfrac{69903.15}{38} = 1839.557$ ⇨ $s_p = 42.89$

- $t_{0.025,38} = 2.02$, $t_{0.05,38} = 1.686$ 38일때 정보X → 컴퓨터로 아님 그냥 40일때로
- 95% 신뢰구간(양측)

$$\bar{x} - \bar{y} \pm t_{0.025,38} s_p \sqrt{\frac{1}{20} + \frac{1}{20}} \Rightarrow (7.29, 62.21)$$

- 95% 신뢰구간(단측) $P(L < \mu_1 - \mu_2) = 1 - \alpha$



$$\left(\bar{x} - \bar{y} - t_{0.05,38} s_p \sqrt{\frac{1}{20} + \frac{1}{20}}, \infty\right) \Rightarrow (11.88, \infty)$$

○ 가설검정

○ $H_0 : \begin{cases} ① \mu_1 \le \mu_2 \\ ② \mu_1 \ge \mu_2 \\ ③ \mu_1 = \mu_2 \end{cases}$ 또는 $H_0 : \mu_1 = \mu_2$ vs $H_1 : \begin{cases} ① \mu_1 > \mu_2 \\ ② \mu_1 < \mu_2 \\ ③ \mu_1 \ne \mu_2 \end{cases}$

- 일반식: $H_0 : \mu_1 - \mu_2 = \boxed{\overset{=0}{\delta}}$ vs $H_1 : \begin{cases} ① \mu_1 - \mu_2 > \delta \\ ② \mu_1 - \mu_2 < \delta \\ ③ \mu_1 - \mu_2 \ne \delta \end{cases}$

○ 검정통계량 $T = \dfrac{\overline{X} - \overline{Y} - \delta}{S_p \sqrt{1/n + 1/m}} \sim t_{n+m-2}$

중심축량

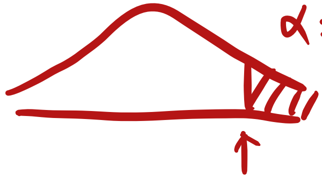
○ 유의수준을 α라고 하면, 기각역

① $[t_{\alpha,m+n-2}, \infty) \Leftrightarrow t \ge t_{\alpha,m+n-2}$

② $(-\infty, -t_{\alpha,m+n-2}] \Leftrightarrow t \le -t_{\alpha,m+n-2}$

③ $(-\infty, -t_{\alpha/2,m+n-2}]$, $[t_{\alpha/2,m+n-2}, \infty) \Leftrightarrow |t| \ge t_{\alpha/2,m+n-2}$

◉ Type A의 평균 콜레스테롤 수치가 Type B의 평균 콜레스테롤 수치보다 높다고 할 수 있는가?



- $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X > \mu_Y$ $(\delta = 0)$

기각역

- $$t = \frac{245.05 - 210.30}{42.89\sqrt{\frac{1}{20} + \frac{1}{20}}} = 2.562 \geq t_{0.05,38} = 1.686$$

- 5% 유의수준에서 Type A의 평균이 Type B의 평균보다 크다고 할 수 있음 (H0 기각)

유의확률

- p-값: $P(T > 2.562) = 0.007$ ⇨ 1%유의수준에서도 같은 결과

□ 평균비교 (독립표본)

- $\sigma_1^2 = \sigma_2^2$를 만족하지 않는 경우(이분산)

○ 가정 II

- $X_1, \ldots, X_m \sim$ iid $N(\mu_1, \sigma_1^2)$
- $Y_1, \ldots, Y_n \sim$ iid $N(\mu_2, \sigma_2^2)$
- $X_1, \ldots, X_m$와 $Y_1, \ldots, Y_n$는 서로 독립

○ 점추정

- 모수: $\mu_1 - \mu_2 \quad \Leftarrow \quad \overline{X} - \overline{Y}$: 표본평균의 차

○ 표본평균의 차 $\bar{X} - \bar{Y}$의 성질

- $E(\bar{X}) = \mu_1$, $E(\bar{Y}) = \mu_2$
- $Var(\bar{X}) = \sigma_1^2/m$, $Var(\bar{Y}) = \sigma_2^2/n$
- $\bar{X} \sim N(\mu_1, \sigma_1^2/m)$, $\bar{Y} \sim N(\mu_2, \sigma_2^2/n)$, $\bar{X}$와 $\bar{Y}$는 독립

$$\Rightarrow \bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

- 표준화: $\boxed{Z = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \sim N(0,1)}$

모분산에 대한 정보가 주어졌다면 표준화된 식이 중심축량
↓
구간추정 → 가설검정

- $T = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_X^2/m + S_Y^2/n}} \sim ?$ ⇐ Behrens-Fisher problem

추정량

어떤 분포를 따르는지 정확히 알 수 없음.

○ $T \simeq t_{\nu}$ 모분산을 알지못하므로 t분포, 하지만 자유도를 명확히 정의할 수 없음

① $\nu = \min(m-1, n-1)$ ⇦ 주로 이용

② Welch-Satterthwaite equation

$$\nu = \frac{(S_X^2/m + S_Y^2/n)^2}{(S_X^2/m)^2/(m-1) + (S_Y^2/n)^2/(n-1)}$$

표본분산값으로 만듦.

○ 구간추정

- $100(1-\alpha)\%$ 신뢰구간

$$1-\alpha \approx P(-t_{\alpha/2,\nu} < T < t_{\alpha/2,\nu})$$

$$= P\left(-t_{\alpha/2,\nu} < \frac{\overline{X}-\overline{Y}-\boxed{(\mu_1-\mu_2)}}{\sqrt{S_X^2/m+S_Y^2/n}} < t_{\alpha/2,\nu}\right)$$

$$= P\left(\overline{X}-\overline{Y}-t_{\alpha/2,\nu}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}} < \mu_1-\mu_2 < \overline{X}-\overline{Y}+t_{\alpha/2,\nu}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}\right)$$

$$\Rightarrow \left(\underline{\overline{X}-\overline{Y}}-t_{\alpha/2,\nu}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}},\ \overline{X}-\overline{Y}+t_{\alpha/2,\nu}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}\right)$$

$\mu \pm t_{\alpha/2,\nu} \times \frac{\sigma}{\sqrt{n}}$

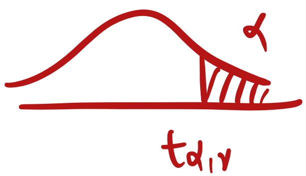
○ 가설검정

○ $H_0 : \begin{cases} ① \mu_1 \le \mu_2 \\ ② \mu_1 \ge \mu_2 \\ ③ \mu_1 = \mu_2 \end{cases}$ 또는 $H_0 : \mu_1 = \mu_2$ vs $H_1 : \begin{cases} ① \mu_1 > \mu_2 \\ ② \mu_1 < \mu_2 \\ ③ \mu_1 \ne \mu_2 \end{cases}$

○ 검정통계량 $T = \dfrac{\overline{X} - \overline{Y} - \delta}{\sqrt{S_X^2/m + S_Y^2/n}} \simeq t_\nu$, $\nu = \min(m-1, n-1)$

○ 유의수준을 α라고 하면, 기각역은

① $[\, t_{\alpha,\nu}, \infty)$ ② $(-\infty, -t_{\alpha,\nu}]$ ③ $(-\infty, -t_{\alpha/2,\nu}]$, $[\, t_{\alpha/2,\nu}, \infty)$



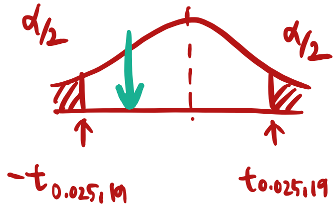
◉ 수능 수리영역 비교($\nu = \min(m-1, n-1)$ 사용)

- 수리영역 성적은 정규분포를 따른다고 가정
- A지역과 B지역에서 각각 20명을 무작위로 선택

통계량	A	B
평균	65	71
표준편차	23	11

표본으로부터 구한 값
→ 통계량이다.

- $s_A / s_B = 2.091 \Rightarrow$ 등분산성 만족하지 않음



- $H_0 : \mu_A - \mu_B = 0$ vs $H_1 : \mu_A - \mu_B \neq 0$

ν=19

$$t = \frac{65-71}{\sqrt{23^2/20 + 11^2/20}} = -0.781 \Rightarrow |-0.781| < t_{0.025,19} = 2.093$$

H0 기각불가능

⇒ 두 지역 간 수능 수리영역 평균에 유의한 차이 없음

◉ 수능 수리영역 비교($\nu = \min(m-1, n-1)$ 사용)

$\bar{X}-\bar{Y} = 65-71 = -6$

95% 신뢰구간

$(\bar{X}-\bar{Y}) \pm t_{\alpha/2,\nu}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}$

$= -6 \pm 2.093\sqrt{\frac{23^2}{20}+\frac{11^2}{20}}$

$= -6 \pm 11.932$

$\rightarrow (-17.932, 5.932)$

- 수리영역 성적은 정규분포를 따른다고 가정
- A지역과 B지역에서 각각 20명을 무작위로 선택

통계량	A	B
평균	65	71
표준편차	23	11

차이가 얼마인지 궁금할땐 추정, 차이가 0일때 이게 유의미한지 가설검정으로 알아보기

- $s_A / s_B = 2.091 \Rightarrow$ 등분산성 만족하지 않음
- $H_0 : \mu_A - \mu_B = 0$ vs $H_1 : \mu_A - \mu_B \neq 0$

$$t = \frac{65-71}{\sqrt{23^2/20 + 11^2/20}} = -0.781 \Rightarrow |-0.781| < t_{0.025,19} = 2.093$$

$\Rightarrow$ 두 지역 간 수능 수리영역 평균에 유의한 차이 없음

□ 정규성을 만족하지 않는 경우

각각

○ **대표본의 경우** ($m \geq 30$ **and** $n \geq 30$)

- 중심극한정리에 의해

$$\overline{X} \simeq N(\mu_1, \sigma_1^2/m), \ \overline{Y} \simeq N(\mu_2, \sigma_2^2/n) \ \Rightarrow \ \overline{X} - \overline{Y} \simeq N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right)$$

$$\Rightarrow \ \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/m + \sigma_2^2/n}} \simeq \underline{N(0, 1)}$$

- σ_1^2와 σ_2^2를 S_X^2와 S_Y^2로 대체가능

대표본이므로 오차영향X → 표준정규분포를 따름

○ $\mu_1 - \mu_2$의 $100(1-\alpha)\%$ 신뢰구간

$$1-\alpha \approx P\left(-z_{\alpha/2} < \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{S_X^2/m+S_Y^2/n}} < z_{\alpha/2}\right)$$

$$= P\left(\overline{X}-\overline{Y}-z_{\alpha/2}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}} < \mu_1-\mu_2 < \overline{X}-\overline{Y}+z_{\alpha/2}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}\right)$$

$$\Rightarrow \left(\overline{X}-\overline{Y}-z_{\alpha/2}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}, \overline{X}-\overline{Y}+z_{\alpha/2}\sqrt{\frac{S_X^2}{m}+\frac{S_Y^2}{n}}\right)$$

○ $H_0 : \mu_1 - \mu_2 = \delta$ 가설검정의 검정통계량

$$\boxed{Z = \frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{S_X^2/m+S_Y^2/n}} \simeq N(0,1)}$$

◉ A 회사와 B회사에는 생산하고 있는 초코파이의 열량을 비교하기 위해, A회사에서 50개를 B회사에서 100개를 무작위로 추출

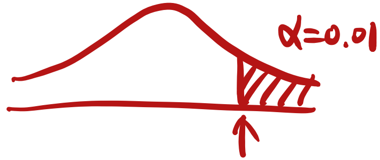
	A	B
평균	155Kcal	151Kcal
표준편차	7Kcal	6Kcal

○ 초코파이 열량의 평균 차 $\mu_A - \mu_B$에 대한 95% 신뢰구간

$$\left(\bar{x} - \bar{y} \pm z_{0.025}\sqrt{\frac{s_A^2}{m} + \frac{s_B^2}{n}}\right) = 155 - 151 \pm 1.96\sqrt{\frac{7^2}{50} + \frac{6^2}{100}}$$

⇨ (1.73, 6.27)

○ A의 초코파이 열량이 B회사 것보다 높은가?



- $H_0 : \mu_A = \mu_B$ vs $H_1 : \mu_A > \mu_B$ $(\alpha = 0.01)$

- $z = \dfrac{155 - 151}{\sqrt{7^2/50 + 6^2/100}} = 3.46 \geq z_{0.01} = 2.33$ H0기각 $z_{0.01}$

- 1% 유의수준에서 A회사의 것이 B회사의 것보다 열량이 높다고 할 수 있음

4kcal