컴퓨터 응용통계
8장 이표본 t-test

최경미

이표본 t-test

사과냐?　　배냐?

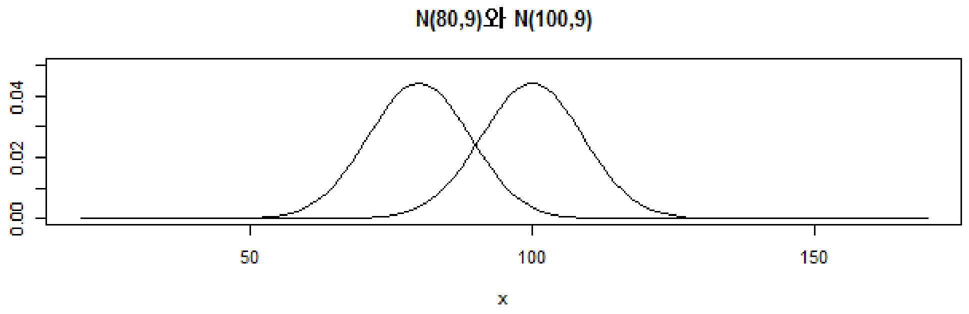
H_0:  $=$ 

H_1:  $\neq$ 

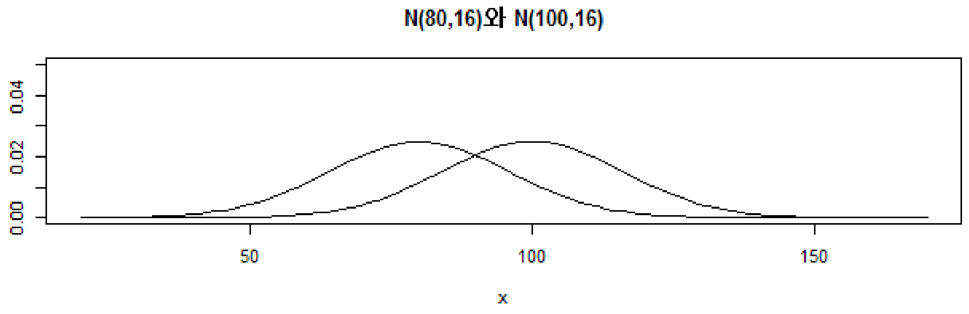
두 집단의 평균 비교

- **가설** H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$

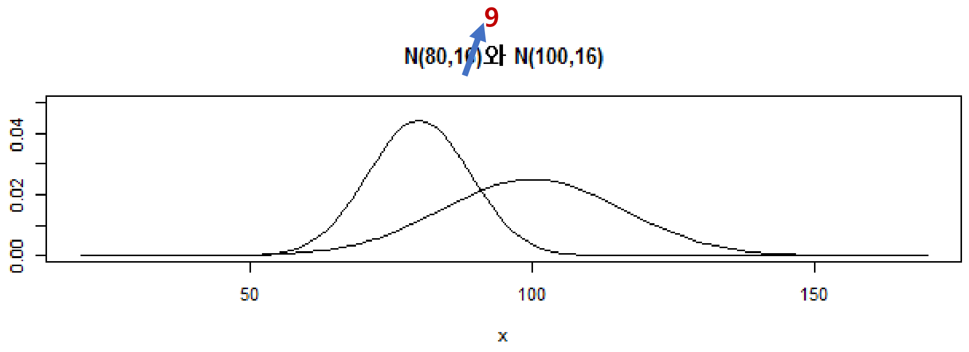
- 분산이 작으면, 평균이 달라 보이고,
  분산이 크면, 평균이 같아 보임.
- 분산이 다를 때는 ……



등분산



등분산



이분산

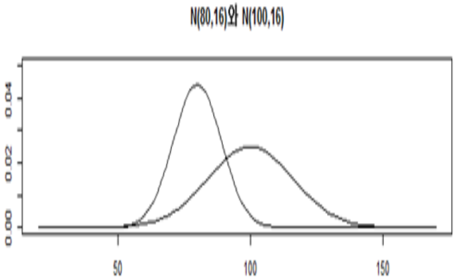
독립인 두 표본의 평균비교

- 등분산성(equal variances) 검정

① 등분산이면 등분산 이표본 $T-$검정

② 분산이 다르면 이분산(non-equal variances) 이표본 $T-$검정

8.1 이분산 T -검정



- $X_1, X_2, \cdots, X_{n_1} \sim$ iid $N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \cdots, Y_{n_2} \sim$ iid $N(\mu_2, \sigma_2^2)$
- 독립표본
- $\sigma_1^2 \neq \sigma_2^2$ 가정
- **가설** H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$

$$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \ , \bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2 \ , \quad S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$$

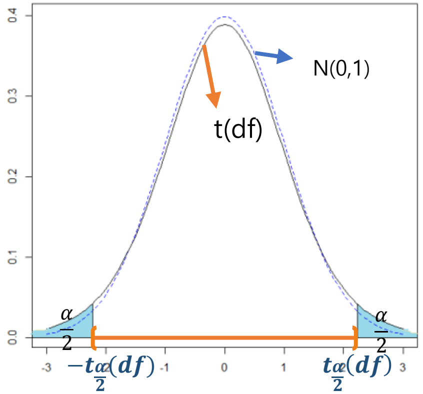
$$E[\bar{X} - \bar{Y}] = \mu_1 - \mu_2$$

$$Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + (-1)^2 Var(\bar{Y}) - 2Cov(\bar{X}, \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\widehat{Var(\bar{X} - \bar{Y})} = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

$$se(\bar{X} - \bar{Y}) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$t = \frac{Z}{\sqrt{\frac{\chi^2(r)}{r}}} \sim t(r)$$



검정통계량 $T = \dfrac{(\bar{X}-\bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \sim t(df)$

Satterthwaite의 자유도 $df = \dfrac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$

$\mu_1 - \mu_2$의 100(1- α)% **신뢰구간**

$$\left(\bar{X} - \bar{Y} - t_{\alpha/2}(df)\sqrt{S_1^2/n_1 + S_2^2/n_2}, \quad \bar{X} - \bar{Y} + t_{\alpha/2}(df)\sqrt{S_1^2/n_1 + S_2^2/n_2}\right)$$

기각역 R: $|T| \geq t_{\frac{\alpha}{2}}(df)$

유의확률 $p -$값 $= P(|T| \geq |t_0|)$

예제 8.2 mtcars자료 표8.1에서 4기통 차와 6기통 차의 평균연비 비교

> x<-c(22.8, 24.4, 22.8, 32.4, 30.4, 33.9, 21.5, 27.3, 26.0, 30.4, 21.4), $n_1 = 11$

> y<-c(21.0, 21.0, 21.4, 18.1, 19.2, 17.8, 19.7), $n_2 = 7$

> mean(x)

> mean(y)

> sd(x)

> sd(y)

> mydata<-c(x,y)

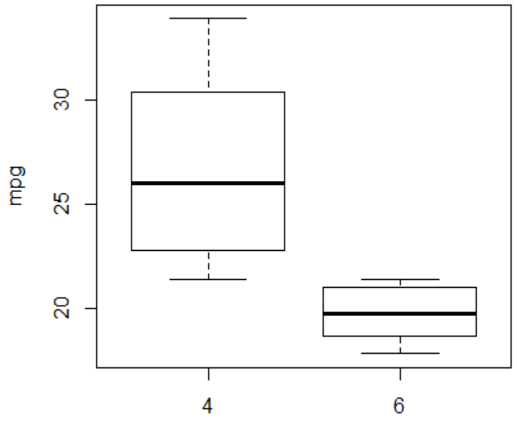
> m<-length(x)

> n<-length(y)

> mygroup<-c(rep("4",m),rep("6",n))

> boxplot(mydata~mygroup, ylab="mpg")

실린더 수	평균	표준편차
4	26.66	4.51
6	19.74	1.45



```
> # 평균 검정
> t.test(x,y)

        Welch Two Sample t-test

data:  x and y
t = 4.7191, df = 12.956, p-value = 0.0004048
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.751376 10.090182
sample estimates:
mean of x mean of y
 26.66364  19.74286
```

$H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

검정통계량, 자유도, 유의확률

< 0.05

∴ 평균 다름

$H_1: \mu_1 \neq \mu_2$

$\mu_1 - \mu_2$

$\mu_1 - \mu_2$

$0 \notin$

신뢰구간

표본평균

$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ 에 **대한 추론**

- $(\mu_1 - \mu_2)$ 의 95% 신뢰구간$= (3.75, 10.09)$

$$26.66 - 19.74 - t_{0.025}\,(12.956)\sqrt{\frac{(4.51)^2}{11} + \frac{(1.45)^2}{7}} = 3.75$$

$$26.66 - 19.74 + t_{0.025}\,(12.956)\sqrt{\frac{(4.51)^2}{11} + \frac{(1.45)^2}{7}} = 10.09$$

- **가설** H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} = \frac{(26.66 - 19.74) - 0}{\sqrt{(4.51)^2/11 + (1.45)^2/7}} = 4.7191,$$

$\mathrm{df} = 12.956$

$|T| \geq t_{0.025}(12.956) = qt(0.975, 12.956) = 2.16$

$p -$ 값 $= P(\,|T| \geq |4.7191|)$

$= 2 * \mathrm{pt}(-4.7191, 12.956) = 0.0004048 < 0.05$

H_0기각

$\therefore \mu_1 \neq \mu_2$

8.2 등분산 T -검정

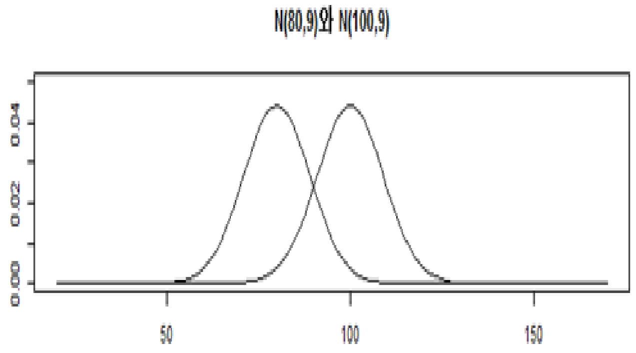
- $X_1, X_2, \cdots, X_{n_1} \sim \text{iid } N(\mu_1, \sigma_1^2)$
- $Y_1, Y_2, \cdots, Y_{n_2} \sim \text{iid } N(\mu_2, \sigma_2^2)$
- 독립표본
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 가정
- **가설** H_0: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$
- **공통분산 (pooled var)**

$$s_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}, \quad se(\bar{X}-\bar{Y}) = s_p\sqrt{1/n_1 + 1/n_2}$$

- **검정통계량** $T = \dfrac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{s_p\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$
- $\mu_1 - \mu_2$의 100(1- α)% **신뢰구간**

$$\left(\bar{X}-\bar{Y}-t_{\alpha/2}\,(n_1+n_2-2)s_p\sqrt{1/n_1+1/n_2},\right.$$
$$\left.\bar{X}-\bar{Y}+t_{\alpha/2}\,(n_1+n_2-2)s_p\sqrt{1/n_1+1/n_2}\right)$$

- **기각역** R: $|T| \geq t_{\alpha/2}\,(n_1+n_2-2)$
- **유의확률** $p-$값 $= P(|T| \geq |t_0|)$



NOTE

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2, \quad S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2$$

$$s_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$$

$$E[\bar{X}-\bar{Y}] = \mu_1 - \mu_2$$

$$Var(\bar{X}-\bar{Y}) = \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2} = \sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

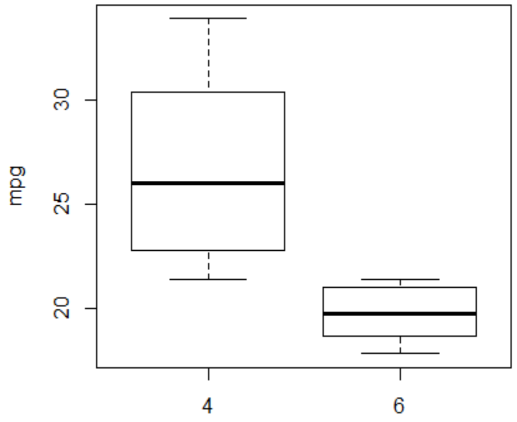
$$\widehat{Var(\bar{X}-\bar{Y})} = s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

$$se(\bar{X}-\bar{Y}) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

예제 8.2 mtcars자료 표8.1에서 4기통 차와 6기통 차의 평균연비 비교

```
> x<-c(22.8, 24.4, 22.8, 32.4, 30.4, 33.9, 21.5, 27.3, 26.0, 30.4, 21.4)

>y<-c(21.0, 21.0, 21.4, 18.1, 19.2, 17.8, 19.7)

> mean(x)

> mean(y)

> sd(x)

> sd(y)

> mydata<-c(x,y)
> m<-length(x)

> n<-length(y)

> mygroup<-c(rep("4",m),rep("6",n))

> boxplot(mydata~mygroup, ylab="mpg")
```

실린더 수	평균	표준편차
4	26.66	4.51
6	19.74	1.45



$H_0\!: \mu_1 = \mu_2 \;\; vs.\; H_1\!: \mu_1 \neq \mu_2$

```
> # 예제 8.3
> x <- mtcars[mtcars$cyl=="4", "mpg"]
> y <- mtcars[mtcars$cyl=="6", "mpg"]
> t.test(x,y, var.equal=TRUE)
            Two Sample t-test
data:   x and y
t = 3.8952, df = 16, p-value = 0.001287
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.154286 10.687272
sample estimates:
mean of x mean of y
 26.66364   19.74286
```

자료 가져오기

등분산 검정 표시하기

검정통계량, 자유도, 유의확률

< 0.05

∴ **평균 다름**

$H_1\!: \mu_1 \neq \mu_2$

$\mu_1 - \mu_2$

신뢰구간

$df = n_1 + n_2 - 2 = 16$

표본평균

$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ 에 대한 추론

df= (11-1)+(7-1)=16

$$s_p^2 = \frac{(11-1)(4.51)^2 + (7-1)(1.45)^2}{(11-1)+(7-1)} = 13.501$$

$(\mu_1 - \mu_2)$ 의 95% 신뢰구간

$$26.66 - 19.74 - t_{0.025}\,(16)\,\sqrt{13.501}\sqrt{\frac{1}{11} + \frac{1}{7}} = = 3.15$$

$$26.66 - 19.74 + t_{0.025}\,(16)\sqrt{13.501}\sqrt{\frac{1}{11} + \frac{1}{7}} = 10.69$$

$$CI = (3.15, 10.69)$$

$$H_0:\ \mu_1 = \mu_2\ \ vs.\ H_1:\ \mu_1 \neq \mu_2$$

$$T = T = \frac{(26.66-19.74)-0}{\sqrt{13.501}\sqrt{1/11+1/7}} = 3.8952,\ \ \mathrm{df}\ =\ 16$$

$$|T| \geq t_{0.025}(16) = qt(0.975, 16) = 2.12$$

$$p - \text{값} = P(\ |T| \geq |3.8952|) =\ 0.001287$$

8.3 등분산성 검정을 위한 F −통계량

- 독립표본

$X_1, X_2, \cdots, X_{n_1}$ ~iid $N(\mu_1, \sigma_1^2)$

$Y_1, Y_2, \cdots, Y_{n_2}$ ~iid $N(\mu_2, \sigma_2^2)$

- H_0: $\sigma_1^2 = \sigma_2^2$ H_1: $\sigma_1^2 \neq \sigma_2^2$ 또는 H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$

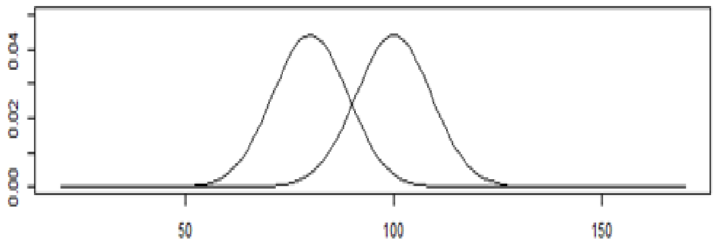
  $V_1 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1)$, $V_2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$

  $$F = \frac{V_1/r_1}{V_2/r_2} = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

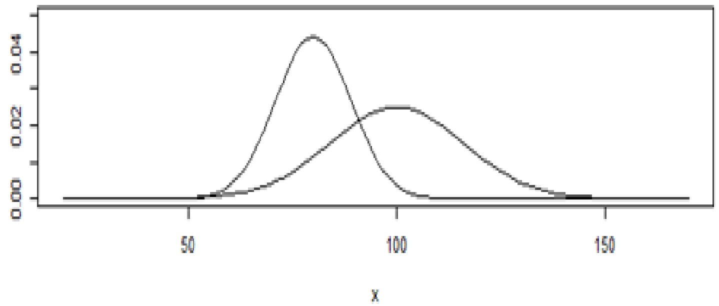
  검정통계량 $F = \frac{S_1^2}{S_2^2}$

- $p-$값 $= 2P(F \geq f_0)$ 또는 $2P(F \leq f_0)$

N(80,9)와 N(100,9)



N(80,16)와 N(100,16)



R 예제 $H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$ 또는 $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

```
> # 예제 8.4
> x<-c(22.8, 24.4, 22.8, 32.4, 30.4, 33.9, 21.5, 27.3, 26.0, 30.4, 21.4)
> y<-c(21.0, 21.0, 21.4, 18.1, 19.2, 17.8, 19.7)
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 9.6261, num df = 10, denom df = 6, p-value = 0.01182
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.762592 39.198688
sample estimates:
ratio of variances
          9.626086
```

$n_1 = 11$

$n_2 = 7$

검정통계량, 자유도 n_1, n_2, 유의확률

$< \alpha$=0.05 이므로, H_0 기각함

$H_1: \sigma_1^2 \neq \sigma_2^2$

$\sigma_1^2 \,/\, \sigma_2^2$

$1 \notin$

신뢰구간

$\sigma_1^2 \,/\, \sigma_2^2$

s_1^2/s_2^2

혼자 풀기

1. R의 mtcars에서 트랜스미션 자동(automatic) 19대와 수동(manual) 13대 차량의 mpg (연비)를 하여 SAS PROC TTEST을 실행한 결과, <표1>과 <표2>를 얻었다. 틀린 설명은 어느 것인가?

<표1> 등분산 검정 $H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \neq \sigma_2^2$

Method	Num DF	Den DF	F Value	Pr > F
F	12	18	2.59	p= 0.0669

$> \alpha = 0.05$ ∴ 등분산

<표2> ttest $H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	30	-4.11	p= 0.0003
Satterthwaite	~~Unequal~~	18.332	-3.77	p= 0.0014

$< \alpha = 0.05$ ∴ 평균 다름

① 독립인 이표본에서 평균을 비교하는 문제이고 검정통계량은 t이다.
② 유의수준 0.05에서 자동과 수동의 분산이 같다고 보기 어렵다.
③ 검정통계량은 t=-4.11이다.
④ 유의수준 0.05에서 두 집단의 평균이 다르다고 말할 수 있다.
⑤ 답 없음

혼자 풀기 (시험)

14. R의 InsectSprays에서 spray B와 spray F를 뿌렸을 때 죽은 벌레 수가 다음과 같이 주어졌다. 정규분포를 가정할 때, spray B와 F를 뿌릴 때, 죽은 벌레 수가 같다고 말할 수 있을지 검정하자. 옳은 것을 모두 고르라. var.test(x,y) ; t.test(x,y); t.test(x,y,var.equal=T)를 사용하자.

a. 등분산 검정의 유의확률은 p= 0.2294이다.

b. 두 평균의 동일성 검정에 대한 검정통계량 $t = -0.61259$의 자유도는 19.498이고, 유의확률은 p=0.5472이다.

c. 두 평균의 동일성 검정에 대한 검정통계량 $t=-0.61259$의 자유도는 22이고, 유의확률은 p= 0.5464이다.

d. 유의수준 0.05에서 spray B와 F를 뿌릴 때, 죽은 벌레 수의 차이가 유의하게 다르다.

① a b ② a c ③ a b d ④ a c d ⑤ 위 보기 중 답 없음

B	11, 17, 21, 11, 16, 14, 17, 17, 19, 21, 7, 13
F	11, 9, 15, 22, 15, 16, 13, 10, 26, 26, 24, 13

```
x <- subset(InsectSprays, spray=="B", c(count) )
y <- subset(InsectSprays, spray==“F", c(count) )
```

```
x<-InsectSprays[InsectSprays$spray=="B", "count"]
y<-InsectSprays[InsectSprays$spray=="F", "count"]
```

```
x <- c(11, 17, 21, 11, 16, 14, 17, 17, 19, 21, 7, 13)
y <- c(11, 9, 15, 22, 15, 16, 13, 10, 26, 26, 24, 13)
```

실전 보고서 (과제)

R의 InsectSprays에서 B,F를 뿌릴 때, 죽는 벌레 수가 동일한지 검정하기 위하여, 유의수준 0.05에서 이표본 T-검정을 실시해보자. 그림1은 자료의 상자도표이다.

커저를 상자도표 그림에 놓고, 복사 붙이기

자르기로 크기 조절, 가운데 정렬

그림1. 살충제 B,F의 상자도표

두 스프레이를 뿌릴 때 죽은 평균 벌레수가 동일한지 알아보기 위하여, 다음과 같이 가설을 세우자.

$$H_0: \ \mu_B = \mu_F \quad H_1: \mu_B \neq \mu_F$$

삽입 => 수식 => 선택해서 입력

표본크기는 각각 $n_1, =?\ n_2 =??$이고, 표본평균은 $\bar{x} = ?, \bar{y} = ?$이고, 표본표준편차는 $s_X =?$, $s_Y =?$이다. 등분산 검정에 대한 유의확률 $p =?$가 유의수준 $\alpha = 0.05$보다 (크므로, 작으므로), (등분산이다, 이분산이다.)

(등분산 T-검정, 이분산 T-검정)을 이용하여 계산한 평균차이 $(\mu_B - \mu_F)$에 대한 95% 신뢰구간은 $(?, ??)$이고, 검정통계량은 $T = ?$이며, 유의확률은 $p =?$이다. 따라서 유의수준 0.05 에서 귀무가설을 (기각한다, 기각하지 않는다.)즉, 유의수준 0.05에서 살충제 B와 F의 효과는 (같다, 다르다.)

부록

코드와 결과 붙이기

삽입 => 표 => 한 칸 선택

(시험) **문제** 두 표본이 $X_1, X_2, \cdots, X_{n_1} \sim \text{iid } \mathrm{N}(\mu_1, \sigma_1^2)$, $Y_1, Y_2, \cdots, Y_{n_2} \sim \text{iid } \mathrm{N}(\mu_2, \sigma_2^2)$일 때,

$\bar{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i$, $\bar{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$, $S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$, $S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})^2$, $s_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$ 에 대한 옳은 설명은 무엇인가?

① $\frac{\bar{X}-\mu_1}{\sigma_1/\sqrt{n_1}} \sim N(0,1)$

② $\frac{(n-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$

③ $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 일 때, $T = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{s_p\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$

④ $\sigma_1^2 \neq \sigma_2^2$ 일 때, $T = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{S_1^2/n_1+S_2^2/n_2}} \sim \mathrm{t}(df)$, $df = \frac{\left(S_1^2/n_1+S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1}+\frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$

⑤ $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \mathrm{F}(n_1-1, n_2-1)$

8.5 쌍체 비교법 (paired t-test) (생략)

- 두 약 A와 B의 생물학적 동등성을 검정하는 문제를 생각해보자.
- 체중이나 체질에 따라서 사람마다 두 약의 약효가 다르게 나타날 수 있기 때문에, 한 사람에게 두 약을 모두 투여한 후, 그 차이가 0인지를 검정한다.
- 이때, 앞서 투약되는 약의 잔여효과(carryover effect)가 나중에 투약되는 약의 효과에 영향을 미치지 않도록, 각 사람에게 투약되는 두 약의 순서를 랜덤하게 정해야 한다.

$\boxed{A_1 \quad B_1}$ $\boxed{A_2 \quad B_2}$ $\boxed{B_3 \quad A_3}$ … $\boxed{A_{n-1} \; B_{n-1}}$ $\boxed{B_n \; A_n}$

$d_i = A_i - B_i$, i=1,…,n

H_0: $\mu_d = 0$ vs. H_1: $\mu_d \neq 0$

$$T = \frac{\bar{d}}{s_d/\sqrt{n}}$$

$$\bar{d} \pm t_{\alpha/2}\,(n-1)\frac{s_d}{\sqrt{n}}$$

R: $|T| \geq \; t_{\alpha/2}\,(n-1)$

$p-$값 $= \mathrm{P}(|T| \geq |t_0|)$

예제 8.6 두 간호사의 심장 신호를 측정 비교 (생략)

Id:	1	2	3	4	5	6	7	8	9	10	11	12
A	4.8,	5.6,	6.0,	6.4,	6.5,	6.6,	6.8,	7.0,	7.0,	7.2,	7.4,	7.6
B	5.8,	6.1,	7.7,	7.8,	7.6,	8.1,	8.0,	8.1,	6.6,	8.1,	9.5,	9.6

```
> # 예제 8.6
> A<-c(4.8,5.6,6.0,6.4,6.5,6.6,6.8,7.0,7.0,7.2,7.4,7.6)
> B<-c(5.8,6.1,7.7,7.8,7.6,8.1,8.0,8.1,6.6,8.1,9.5,9.6)
> t.test(A,B, paired=T)
        Paired t-test
data:   A and B
t = -6.0237, df = 11, p-value = 8.628e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6043289 -0.7456711
sample estimates:
mean of the differences
               -1.175
```

쌍체비교

검정통계량, 자유도, 유의확률

신뢰구간

평균차이 $\bar{d}$