전산통계 과제#4

컴퓨터소프트웨어 학부
2018008559
신상윤

5-1

코드

```
data exe5_1;
input rank $ fp @@;
cards;
good 0.58 good 2.8 good 2.77 good 3.5 good 2.67 good 2.97
good 2.18 good 3.24 good 1.49 good 2.19 good 2.7 good 2.57
bad 2.28 bad 1.06 bad 1.08 bad 0.07 bad 0.16 bad 0.7
bad 0.75 bad 1.61 bad 0.34 bad 1.15 bad 0.44 bad 0.86
;
run;
proc ttest data = exe5_1 cochran alpha = 0.05 order=data;
        class rank;
        var fp;
run;
```

결과

(가) 먼저 모분산을 모르므로 모분산 동일성에 대하여 검정해야 한다.

검정통계량 F = $\dfrac{0.792038^2}{0.6286566^2}$ =1.587322156

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	11	11	1.59	0.4558

유의수준 5%에서 기각역에 포함되지 않으므로 H0를 채택한다.
따라서 모분산이 동일하다 할 수 있다(pooled)

이때 검정통계량은 $s_p^2 = \dfrac{11*0.792038^2 + 11*0.6286566^2}{22}$

T = $\dfrac{2.4717 - 0.875 - 0(H0: \mu_1 = \mu_2)}{\sqrt{s_p^2 * (\frac{1}{12} + \frac{1}{12})}}$ = 5.469847461

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	22	5.47	<.0001
Satterthwaite	Unequal	20.922	5.47	<.0001
Cochran	Unequal	11	5.47	0.0002

기각역에 포함되므로 유의수준 5%에서 H0를 기각한다.
따라서 모평균이 같다 할 수 없다. 즉, 모평균에 차이가 있다.

(나)

신뢰구간은 위에서 구한 합동 추정량을 이용해 구할 수 있다.

$\overline{x_1} - \overline{x_2} = 2.4717 - 0.875$ = 1.5967

$t_{0.025} * 0.71502913\sqrt{\frac{1}{12} + \frac{1}{12}}$ = 0.605420137

95% 신뢰구간 : (0.991279863, 2.202120137)

rank	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
good		2.4717	1.9684	2.9749	0.7920	0.5611	1.3448
bad		0.8750	0.4756	1.2744	0.6287	0.4453	1.0674
Diff (1-2)	Pooled	1.5967	0.9913	2.2020	0.7150	0.5530	1.0120
Diff (1-2)	Satterthwaite	1.5967	0.9895	2.2039			

5-2

코드

```
data exe5_2;
input M $ D @@;
cards;
A 0.95 A 0.82 A 0.78 A 0.96 A 0.71 A 0.86 A 0.99
B 0.89 B 0.91 B 0.94 B 0.91 B 0.90 B 0.89
;
run;
proc ttest data = exe5_2 alpha = 0.05 order=data;
        class M;
        var D;
run;
```

결과

H0 : $\sigma_1^2 = \sigma_2^2$ H1 : $\sigma_1^2 \neq \sigma_2^2$

검정통계량 F = $\dfrac{0.1042^2}{0.0186^2}$ = 31.384

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6	5	31.32	0.0016

기각역에 포함되므로 H0를 기각한다.
유의수준 5% 하에서 두 모집단의 분산이 동일하다고 할 수 없다.

5-3

코드

```
data exe5_3;
input id nut nut_fried @@;
cards;
1 61 56 2 60 54 3 56 47 4 63 59 5 56 51
6 63 51 7 69 57 8 56 54 9 44 63 10 61 58
;
run;
proc ttest data = exe5_3 alpha = 0.05 order=data;
        paired nut * nut_fried;
run;
```

결과

H0 : $\mu_1 - \mu_2 = 0$ H1 : $\mu_1 - \mu_2 \neq 0$

검정통계량 T = $\frac{3.9 - 0(\because H0)}{8.7743/\sqrt{10}}$ = 1.405568863

DF	t Value	Pr > |t|
9	1.41	0.1934

기각역에 포함되지 않으므로 H0를 채택한다.
유의수준 5% 하에서 단백질 함량의 평균에 차이가 없다.

5-5

코드

```
data exe5_5;
input M $ T @@;
cards;
A 2.0 A 2.1 A 2.5 A 3.0 A 3.3 A 4.2 A 4.2 A 4.3 A 6.8 A 7.6
B 3.6 B 3.7 B 3.9 B 3.9 B 3.9 B 4.0 B 4.0 B 4.0 B 4.1 B 4.2 B 4.3
;
run;
proc ttest data = exe5_5 alpha = 0.05 order=data;
        class M;
        var T;
run;
```

결과

(가) H0 : $\sigma_A^2 \leq \sigma_B^2$ H1 : $\sigma_A^2 > \sigma_B^2$

검정통계량 F = $\dfrac{1.895^2}{0.2014^2}$ = 88.5318

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	10	88.57	<.0001

기각역에 포함되므로 H0를 기각한다.
유의수준 5% 하에서 A방법의 분산이 B방법의 분산보다 크다 할 수 있다.

(나) 위에 결과를 이용하여 두 분산의 비에 대한 95% 신뢰구간을 구하면

$\left(\dfrac{1}{F_{0.025}(9,10)}\dfrac{s_1^2}{s_2^2}, F_{0.025}(10,9)\dfrac{s_1^2}{s_2^2}\right)$ = (23.43, 350.93)이다.

5-6

코드

```
data deer;
input id hindleg foreleg @@;
cards;
1 142 138    2 140 136    3 144 147    4 144 139    5 142 143
6 146 141    7 149 143    8 150 145    9 142 136    10 148 146
;
run;
proc ttest data = deer alpha = 0.05 order=data;
          paired hindleg*foreleg;
run;
```

결과

H0 : $\mu_1 - \mu_2 = 0$ H1 : $\mu_1 - \mu_2 \neq 0$

검정통계량 T = $\dfrac{3.3}{3.0569/\sqrt{10}}$ = 3.41375

DF	t Value	Pr > |t|
9	3.41	0.0077

기각역에 포함되므로 H0를 기각한다.
따라서 유의수준 5% 하에서 평균에 차이가 있다고 할 수 있다.
이는 4-2 결과와 같다.

(사)

위치모수 검정: Mu0=0				
검정	통계량		p 값	
스튜던트의 t	t	3.413793	Pr > |t|	0.0077
부호	M	3	Pr >= |M|	0.1094
부호 순위	S	23.5	Pr >= |S|	0.0117

검정통계량 : 3.413793, 유의수준 5%에서 기각역 : T >= $T_{\alpha}(9)$ = 1.83311

$\alpha = 0.05$하에서 H0를 기각한다. 따라서 평균적으로 사슴의 뒷다리가 앞다리보다 길다고 할 수 있다.

5-7

결과

철사 종류 A : σ_1= 3, 여기서 표본 20개 추출 $\overline{x_1}$ = 32.5

전기 처리한 철사 종류 A : σ_2= 4, 여기서 표본 25개 추출 $\overline{x_2}$ = 36.4

(가)

$$32.5 - 1.96 * \frac{3}{\sqrt{20}} \le \mu_1 \le 32.5 + 1.96 * \frac{3}{\sqrt{20}}$$

31.19 $\le \mu_1 \le$ 33.81

마찬가지로 34.83 $\le \mu_2 \le$ 37.97

(나)

$$32.5 - 36.4 - 1.96\sqrt{\frac{3^2}{20} + \frac{4^2}{25}} \le \mu_1 - \mu_2 \le 32.5 - 36.4 + 1.96\sqrt{\frac{3^2}{20} + \frac{4^2}{25}}$$

-5.95 $\le \mu_1 - \mu_2 \le$ -1.85

(다)

검정통계량 Z = $\dfrac{32.5-36.4}{\sqrt{\dfrac{9}{20}+\dfrac{16}{25}}}$ = -3.74

기각역에 포함되므로 H0를 기각한다.

따라서 유의수준 5% 하에서 $\mu_1 < \mu_2$라 할 수 있다.

이때 p_value는 0.00009이다.

5-9

코드

```
data exe5_9;
input type $ amount @@;
cards;
A 31 A 34 A 29 A 26 A 32 A 35 A 38 A 34 A 30 A 29 A 32 A 31
B 26 B 24 B 28 B 29 B 30 B 29 B 32 B 26 B 31 B 29 B 32 B 28
;
run;
proc ttest data = exe5_9 alpha=0.05 order=data;
          class type;
          var amount;
run;
```

결과

(가) 먼저 모분산을 모르므로 모분산 동일성에 대하여 검정해야 한다.

검정통계량 F = $\frac{3.1945^2}{2.4618^2}$ =1.6838

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	11	11	1.68	0.4009

기각역에 포함되지 않으므로 H0를 채택한다.

따라서 유의수준 5% 하에서 $\sigma_1^2 = \sigma_2^2$라 할 수 있다(pooled)

이때 검정통계량은

$$s_p^2 = \frac{11*3.1945^2 + 11*2.4618^2}{22}$$ = 8.132644

T = $\dfrac{31.75-28.6667}{\sqrt{s_p^2 * (\frac{1}{12}+\frac{1}{12})}}$ = 2.64835

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	22	2.65	0.0147
Satterthwaite	Unequal	20.659	2.65	0.0152

H0 : $\mu_1 \leq \mu_2$ H1 : $\mu_1 > \mu_2$

기각역에 포함되므로 H0를 기각한다.
따라서 유의수준 5% 하에서 A품종이 B품종보다 수확량이 많다고 할 수 있다.

(나) 수확량의 차에 대한 95% 신뢰구간은 위에서 구한 값들을 이용해 구할 수 있다.

$\overline{x_1} - \overline{x_2} = 31.75 - 28.6667$ = 3.0833

$2.074 * 2.8518\sqrt{\frac{1}{12}+\frac{1}{12}}$ = 2.415

95% 신뢰구간 : (0.6687, 5.4979)

type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
A		31.7500	29.7203	33.7797	3.1945	2.2629	5.4238
B		28.6667	27.1025	30.2308	2.4618	1.7439	4.1799
Diff (1-2)	Pooled	3.0833	0.6689	5.4978	2.8518	2.2055	4.0363
Diff (1-2)	Satterthwaite	3.0833	0.6598	5.5069			

5-10

코드

```
data exe5_10;
input id post comp @@;
cards;
1 4 3 2 6 4 3 8 6 4 3 2 5 3 5 6 7 8
7 4 3 8 5 5 9 9 10 10 10 9 11 9 8 12 7 5
;
run;
proc ttest data = exe5_10 alpha=0.05 order=data;
          paired post*comp;
run;
```

결과

H0 : $\mu_1 - \mu_2 \leq 0$ H1 : $\mu_1 - \mu_2 > 0$

검정통계량 T = $\dfrac{0.5833}{1.3114/\sqrt{12}}$ = 1.5408

DF	t Value	Pr > |t|
11	1.54	0.1516

기각역에 포함되지 않으므로 H0를 채택한다.
따라서 유의수준 5% 하에서 전문회사에 맡기는 것이 빠르다고 할 수 없다.

5-13

코드

```
data exe5_13;
input type $ zzz @@;
cards;
up 35.3 up 35.9 up 37.2 up 33.0 up 31.9 up 33.7 up 36.0 up 35.0
up 33.3 up 36.6 up 37.9 up 35.6 up 29.0 up 33.7 up 35.7
down 32.4 down 34.0 down 34.4 down 31.8 down 35.0 down 34.6 down
34.6 down 33.5
down 33.6 down 31.5 down 33.8
;
run;
proc ttest data = exe5_13 alpha=0.05 order=data;
        class type;
        var zzz;
run;
```

결과

(가) H0 : $\sigma_1^2 \le \sigma_2^2$ H1 : $\sigma_1^2 > \sigma_2^2$

검정통계량 F = $\dfrac{2.2778^2}{1.1784^2}$ = 3.7363

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	10	3.74	0.0420

기각역에 포함되므로 H0를 기각한다.
따라서 유의수준 5% 하에서 7시간 이상 수면을 취하는 집단이 7시간 미만 수면을 취하는 집단보다 에너지 소비량의 분산이 크다고 할 수 있다.

(나) 먼저 두 집단의 분산의 비에 대한 95% 신뢰구간을 구해보면

$$(\frac{1}{F_{0.025}(14,10)}*3.7363,\ F_{0.025}(10,14)*3.7363)$$

=(1.0523, 11.7577) 이고, 루트를 씌워주면 표준편차 비의 신뢰구간이 된다.

따라서 표준편차의 비에 대한 95% 신뢰구간은

$1.02585 \le \frac{\sigma_1}{\sigma_2} \le 3.42895$ 이다.