전산통계 과제#3

컴퓨터소프트웨어 학부
2018008559
신상윤

4-2

코드

```
data deer;
        input hindleg foreleg @@;
        diff = hindleg - foreleg;
cards;
142 138    140 136    144 147    144 139    142 143
146 141    149 143    150 145    142 136    148 146
;
run;
proc univariate data = deer normal plot;
        var diff;
run;
```

결과

(가) diff = 뒷다리 – 앞다리로 설정한 것으로 보아 평균적으로 사슴의 뒷다리가 일반적으로 앞다리보다 길다는 생각으로 표본을 생산한 것 같다.

(나)

귀무가설(H0) : 평균적으로 뒷다리 <= 앞다리

대립가설(H1) : 평균적으로 뒷다리 > 앞다리

m : diff 들의 모평균이라 할 때

H0 : m <= 0, H1 : m > 0

(다)

1종오류 : H0이 참이지만 H0을 기각

m <= 0이지만 m > 0으로 결론 냄

2종오류 : H0이 거짓이지만 H0을 기각하지 않음

m > 0이지만 m <= 0으로 결론 냄

(라) 유의수준을 5%로 한다는 의미는 1종오류 즉, m <= 0이지만 m > 0으로 결론 낼 확률은 최대 5%라는 의미이다. 1종오류로 정하는 이유는 1종오류가 더 심각한 오류이기 때문이다. 2종오류는 자신의 연구가 맞는데 틀렸다고 생각하여 다시 연구할 가능성이 있지만, 1종오류는 자신의 연구가 틀렸는데 맞았다고 생각하여 피해를 줄 수 있다.

(마)

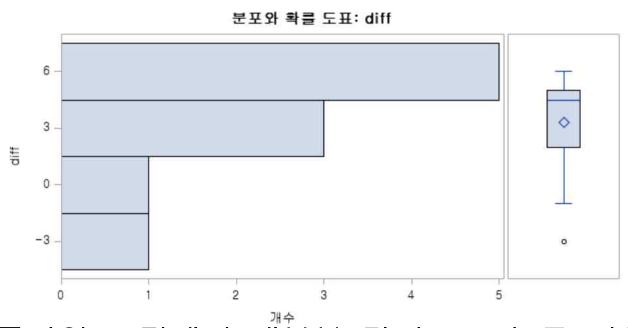
적률			
N	10	가중합	10
평균	3.3	관측값 합	33
표준 편차	3.0568684	분산	9.34444444
왜도	-1.3466487	첨도	0.78935489
제곱합	193	수정 제곱합	84.1
변동계수	92.6323759	평균의 표준 오차	0.96666667

기본 통계 측도			
위치측도		변이측도	
평균	3.300000	표준 편차	3.05687
중위수	4.500000	분산	9.34444
최빈값	5.000000	범위	9.00000
		사분위수 범위	3.00000

분위수(정의 5)	
레벨	분위수
100% 최댓값	6.0
99%	6.0
95%	6.0
90%	6.0
75% Q3	5.0
50% 중위수	4.5
25% Q1	2.0
10%	-2.0
5%	-3.0
1%	-3.0
0% 최솟값	-3.0

평균 3.3, 중앙값 4.5, 최빈값 5, 표준편차 3.057이다. Q1의 값이 2로 대부분 값들이 0보다 크다. 최소값은 –3이고, 최대값은 6으로 평균에는 6이더 가깝다. 4, 5, 6이 2번 이상 나타났고, 그중 5는 3번으로 최빈값이다.

(바)



줄기잎 그림에서 대부분 값이 0보다 큰 것을 알 수 있고, 상자 그림에서 –3이 극단값임을 알 수 있다. 요약하면 3이상인 값이 많이 분포하고, -3인 극단값을 가진다.

(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.

4-5

코드

```
data exe5;
	L1 = 171 - 1.96 * 6 / 10;
	R1 = 171 + 1.96 * 6 / 10;
	L2 = 171 - 1.96 * 6 / 20;
	R2 = 171 + 1.96 * 6 / 20;
	ok1 = 0;
	ok2 = 0;
	do i = 1 to 100000;
		sum = 0;
		do j = 1 to 100;
			x = rand('normal',171,6);
			sum = sum + x;
		end;
		x = sum / 100; /* 점추정 */
		if(L1 <= x and x <= R1) then do;
			ok1 = ok1 + 1;
		end;
		sum = 0;
		do j = 1 to 400;
			x = rand('normal',171,6);
			sum = sum + x;
		end;
		x = sum / 400; /* 점추정 */
		if(L2 <= x and x <= R2) then do;
			ok2 = ok2 + 1;
		end;
	end;
	p1 = ok1 / 100000;
	p2 = ok2 / 100000;
run;
proc print data = exe5(drop = i j x sum ok1 ok2);id;run;
```

결과

L1	R1	L2	R2	p1	p2
169.824	172.176	170.412	171.588	0.94919	0.95046

(가) 구간은 (169.824, 172.176)이다. 실제로 점 추정 해봤을 때 약 95%가 포함되었다.

(나) 구간은 (170.412, 171.588)이다. 마찬가지로 실제로 약 95%가 포함되었다. (나)의 경우 더 많은 표본을 뽑아 평균이 모평균에 가까울 확률이 높으므로 구간이 (가)보다 작아진다. 이는 식에서도 확인할 수 있다.

4-7

코드

```
data exe7;
	t_inv = tinv(0.975,49);
	L = 114 - t_inv * 8.4 / sqrt(50);
	R = 114 + t_inv * 8.4 / sqrt(50);
run;
proc print data = exe7;id;run;
```

결과

t_inv	L	R
2.00958	111.613	116.387

모분산이 주어지지 않았으므로 T분포로 근사한다. T = $\frac{\overline{X}-\mu}{s/\sqrt{n}}$이고,

95%의 신뢰구간은 $(114 - t_{\alpha/2}(49)*\frac{s}{\sqrt{n}}, 114 + t_{\alpha/2}(49)*\frac{s}{\sqrt{n}})$이다.

$\alpha = 0.025$이므로 신뢰구간은 (111.613, 116.387)이다.

4-8

코드

```
data exe8;
	input name $ cnt;
cards;
A 55
B 45
;
run;
data exe8_1;
	L = 0.55 - 1.96 * sqrt(0.55 * 0.45 / 100);
	R = 0.55 + 1.96 * sqrt(0.55 * 0.45 / 100);
	L1 = 1;
	do while(1 - cdf('binomial',54,L1,100) > 0.025);
		L1 = L1 - 0.0001;
	end;
	R1= 0;
	do while(cdf('binomial',55,R1,100) > 0.025);
		R1 = R1 + 0.0001;
	end;
run;
proc print data = exe8_1;id;run;
proc freq data = exe8 order = data;
	weight cnt;
	exact binomial;
	table name / alpha = 0.05;
run;
```

결과

이항비	
name = A	
비율(P)	0.5500
ASE	0.0497
95% 신뢰하한	0.4525
95% 신뢰상한	0.6475
정확 신뢰한계	
95% 신뢰하한	0.4473
95% 신뢰상한	0.6497

L	R	L1	R1
0.45249	0.64751	0.4472	0.6497

정규근사 신뢰구간은 (0.4525, 0.6475)
정확한 신뢰구간은 (0.4473, 0.6497)로 직접 구한 값과 차이가 없었다.
주의할 점은
$P(X \geq 55) = 1 - P(X \leq 54)$라는 것이다.

4-12

코드

```
data exe12;
        reject = quantile('chisquared',0.05,70);
        test = 70 * 16 / 30;
run;
proc print data = exe12;id;run;
```

결과

reject	test
51.7393	37.3333

기각역은 $\chi^2 \leq 51.7393$이고, 검정통계량은 37.333이다. 기각역에 포함되므로, α = 5% 하에서 H0는 기각된다. 따라서 분산이 30보다 작다고 할 수 있다.

4-13

코드

```
data exe13;
        n = 2 * 1.96 / 0.1;
        n = n*n;
run;
proc print data = exe13;id;run;
```

결과

n
1536.64

표본은 1537개 이상 추출해야 한다.

4-16

H0 : 맥박의 변화가 없다. μ = 60

H1 : 맥박의 변화가 있다(증가했다). μ > 60

$\overline{X} \sim N(60, 5^2/16)$, $Z = \dfrac{70-60}{5/4} = 8$. 기각역 : $Z \geq Z_{0.05} = 1.645$

8$\geq$1.645이므로 기각역에 포함 H0기각.

α=0.05하에서 H0기각. 따라서 맥박의 변화가 있다고 할 수 있다.

4-18

코드

```
data exe18;
        input Mn @@;
cards;
1269 1271 1263 1265
;
```

```
run;
proc means data = exe18 n mean;run;
```

결과

분석 변수: Mn	
N	평균
4	1267.00

H0 : μ = 1267, H1 : $\mu \neq 1267$

$Z = \frac{1267-1267}{\sqrt{5}/\sqrt{4}} = 0$, 기각역 $|Z| \geq Z_{\alpha/2} = 1.645$ 기각역에 포함하지 않는다.

α=0.1하에서 H0기각하지 않는다. 따라서 망간이 녹는 평균 온도가 1267이 아니라고 할 수 없다.

4-20

코드

```
data exe20;
        input work $ cnt;
cards;
YES 160
NO 40
;
run;
proc freq data = exe20 order = data;
        weight cnt;
        exact binomial;
        table work / binomial(p=0.9) alpha=0.05;
run;
```

결과

H0: P = 0.9의 검정	
H0 하에서의 ASE	0.0212
Z	-4.7140
단측 Pr < Z	<.0001
양측 Pr > |Z|	<.0001
정확 검정	
단측검정 Pr <= P	<.0001
양측 = 2*단측	<.0001

H0 : p = 0.9, H1 : p $\neq$ 0.9

Z = -4.714, 기각역 : $|Z| \geq Z_{\alpha/2} = 1.96$ 기각역에 포함된다. H0기각

α=0.05하에서 H0를 기각한다. 따라서 평균 출근율과 다르다고 할 수 있다.

4-22

코드

```
data exe22;
	input six $ cnt;
cards;
YES 52
NO 188
;
run;
proc freq data = exe22 order=data;
	weight cnt;
	exact binomial;
	table six / binomial(p=0.1666666) alpha=0.1;
run;
```

결과

H0: P = 0.1666666의 검정	
H0 하에서의 ASE	0.0241
Z	2.0785
단측 Pr > Z	0.0188
양측 Pr > |Z|	0.0377
정확 검정	
단측검정 Pr >= P	0.0261
양측 = 2*단측	0.0521

H0 : 바른 주사위다. p = $\frac{1}{6}$

H1 : 바른 주사위가 아니다. p $\neq \frac{1}{6}$

Z = 2.0785, 기각역 : $|Z| \geq Z_{\alpha/2} = 1.645$ 기각역에 포함됨. H0 기각

α=0.1하에서 H0를 기각한다. 따라서 바른 주사위가 아니다.

4-24

코드

```
data exe24;
        input PF $ cnt;
cards;
P 66
F 54
;
run;
proc freq data = exe24 order=internal;
        weight cnt;
        exact binomial;
        table PF / binomial(p=0.4) alpha=0.05;
run;
```

결과

H0: P = 0.4의 검정	
H0 하에서의 ASE	0.0447
Z	1.1180
단측 Pr > Z	0.1318
양측 Pr > |Z|	0.2636
정확 검정	
단측검정 Pr >= P	0.1528
양측 = 2*단측	0.3055

H0 : 열등하지 않다. p $\leq$ 0.4
H1 : 열등하다. 불합격률이 높다. p > 0.4
Z = 1.118, 기각역 : $Z \geq Z_{\alpha} = 1.645$ 기각역에 포함되지 않는다.

α=0.05하에서 H0를 기각하지 않는다. 강남고등학교 학생들이 열등하다고 할 수 없다.

4-26

코드

```
data exe26;
        test = 8 * 0.038 / 0.04;
        reject = quantile('chisquared',0.05,8);
run;
proc print data = exe26;id;run;
```

결과

test	reject
7.6	2.73264

H0 : $\sigma^2 \geq$0.04 ,H1 : 좋은 제품 σ^2< 0.04

χ^2=7.6, 기각역 $\chi^2 \leq \chi^2_{0.95}(8)$=2.73264, 기각역에 포함되지 않음.

α=0.05하에서 H0 기각하지 않음. 따라서 좋은 제품이라 할 수 없다.