**SAS 데이터 분석 입문 4장**

**2019020650 김형욱**

**\* 4장 예제문제**

**<예 4-1>**

**data** sasadv.csi;

input csi @@;

label csi='소비자 만족도 지수';

cards;

75 63 49 86 53 80 70 72 81 80 69 76 85 95 66 77 77 63 58 74

68 90 82 59 60

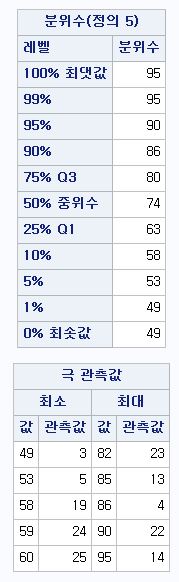
;

**run**;

**proc** **univariate** data=sasadv.csi cibasic alpha=**0.05**;

var csi;

**run**;

****

**<예 4-2>**

**proc** **means** data=sasadv.csi mean std clm alpha=**0.05**;

var csi;

**run**;

****

**<예 4-3>**

**data** sasadv.poll;

input yesno $ count;

cards;

YES 250

NO 150

;

**run**;

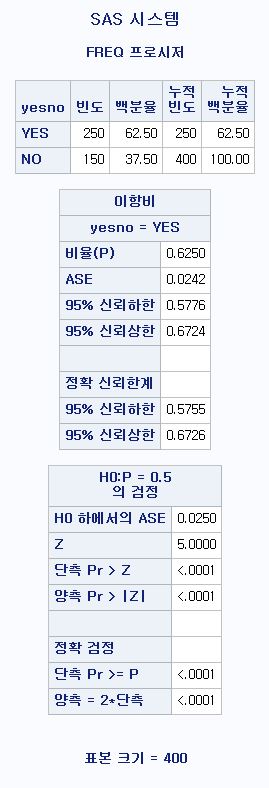
**proc** **freq** data=sasadv.poll order=data;

weight count;

exact binomial;

tables yesno / alpha=**0.05**;

**run**;

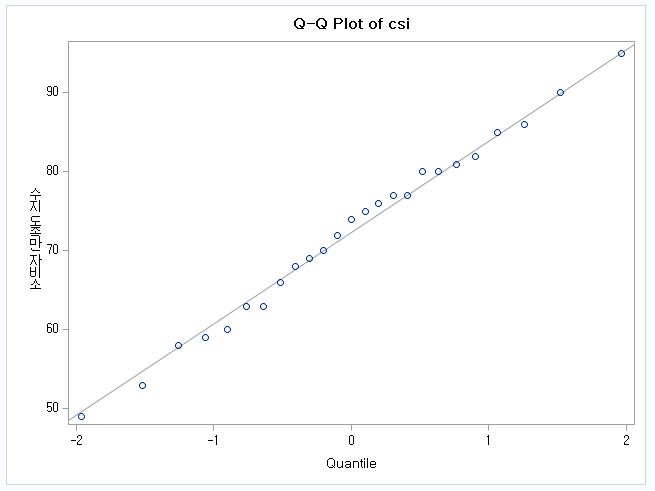
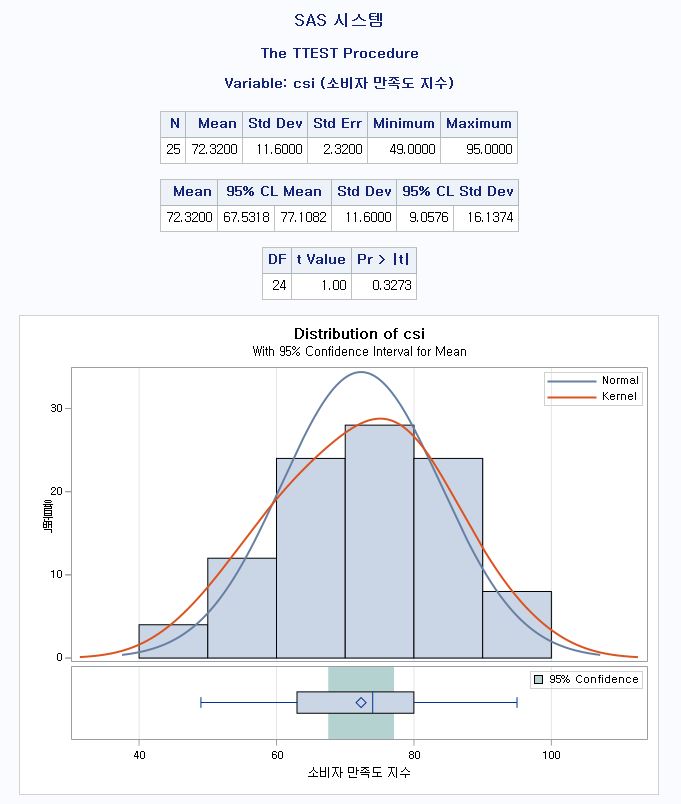
****

**<예 4-4>**

**proc** **ttest** data=sasadv.csi h0=**70**;

var csi;

**run**;

****

해석 : 가설검정을 위해 귀무가설을 ‘서비스 만족도의 모평균이 70이다’, 대립가설을 ‘서비스 만족도의 모평균이 70보다 높다’라고 두었다. 그리고 검정 결과, 검정통계량의 값이 1로 임계치인 1.711보다 크지 않으므로 귀무가설을 기각할 수 없다. 따라서 ‘서비스 만족도의 모평균이 70 이상’이라고 할 수 없다.

**<예 4-5>**

**proc** **univariate** data=sasadv.csi mu0=**70** alpha=**0.05** cibasic;

var csi;

**run**;



해석 : 예 4-4와 마찬가지로 귀무가설을 ‘서비스 만족도의 모평균이 70이다’, 대립가설을 ‘서비스 만족도의 모평균이 70보다 높다’라고 두고 가설 검정을 한 결과, 검정통계량의 값이 1로 임계치인 1.711보다 크지 않으므로 귀무가설을 기각할 수 없다. 따라서 ‘서비스 만족도의 모평균이 70 이상’이라고 할 수 없다.

**<예 4-6>**

**data** sasadv.goods;

input state $ count @@;

cards;

Poor 54 Good 346

;

**run**;

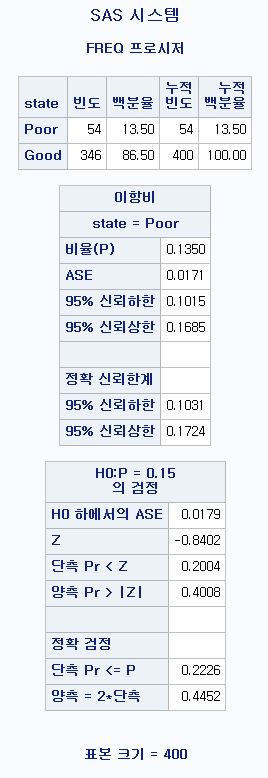
**proc** **freq** data=sasadv.goods order=data;

weight count;

exact binomial;

tables state / binomial(p=**0.15**) alpha=**0.05**;

**run**;



해석 : 어느 회사의 제품에 대한 불량률은 15% 정도이며, 이 블량률을 줄이기 위해 새로운 재료를 사용하려고 한다. 새로운 재료를 사용한 제품 중 400개를 뽑아 조사한 결과, 54개의 불량품이 발견되었다. 이때 불량률이 종전(15%)에 비해 낮아졌다고 할 수 있는지를 유의수준 5%하에서 검정하기 위해, 귀무가설을 ‘p(=불량률)=0.15’, 대립가설을 ‘p(=불량률)<0.15’라고 두었다. 검정 결과, 검정통계량의 값이 -0.84로 임계치인 -1.645보다 크므로 귀무가설을 기각할 수 없다. 따라서 ‘불량률이 종전(15%)보다 낮아졌다고 할 수 없다.

**\* 4장 연습문제**

**<연습문제 4-1>**

**data** sasadv.ex4\_1;

input mileage @@;

cards;

21.0 22.7 25.8 20.6 18.5 21.4 19.3

17.6 22.7 20.6 17.9 18.3 24.7 23.3

24.3 21.5 20.0 19.8 22.9 19.9

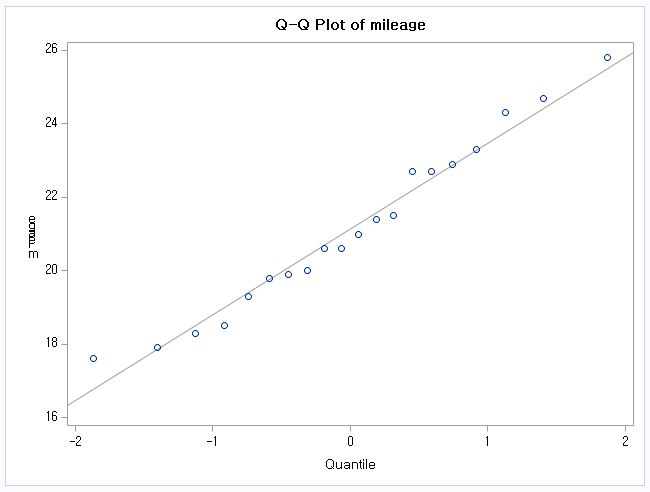
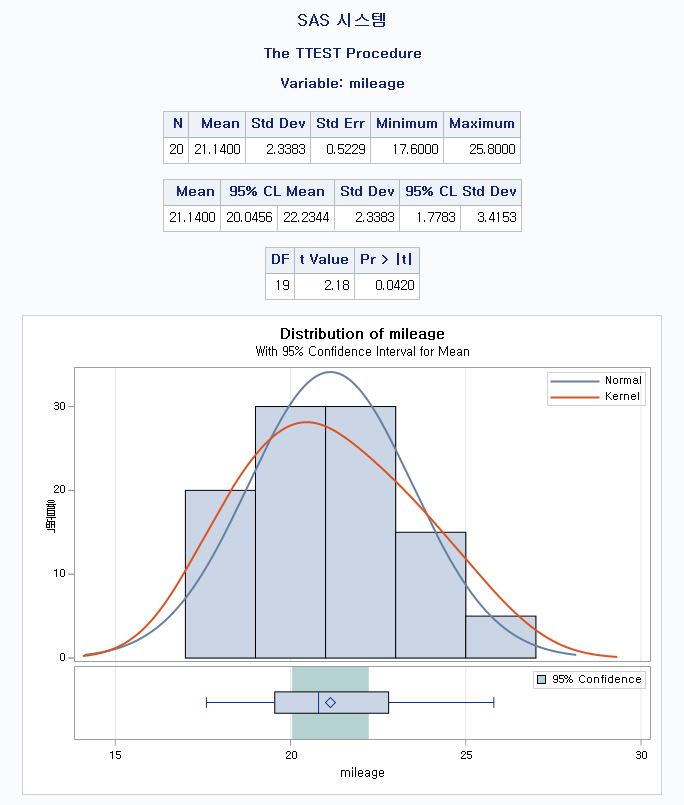
;

**run**;

**proc** **ttest** data=sasadv.ex4\_1 h0=**20** alpha=**0.05**;

var mileage;

**run**;



해석 : 가설 검정을 위해, 귀무가설을 ‘새 자동차의 평균 연비는 20km/l이다’, 대립가설을 ‘새 자동차의 평균 연비는 20km/l가 아니다’라고 두었다. 검정 결과, p-value가 0.0420이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 새 자동차의 연비가 평균적으로 리터당 20km라고 할 수 없다.

**<연습문제 4-2>**

**data** sasadv.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

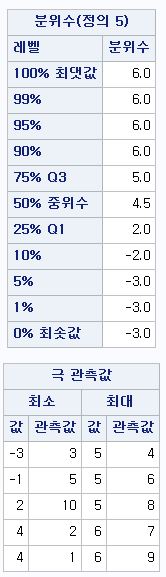
;

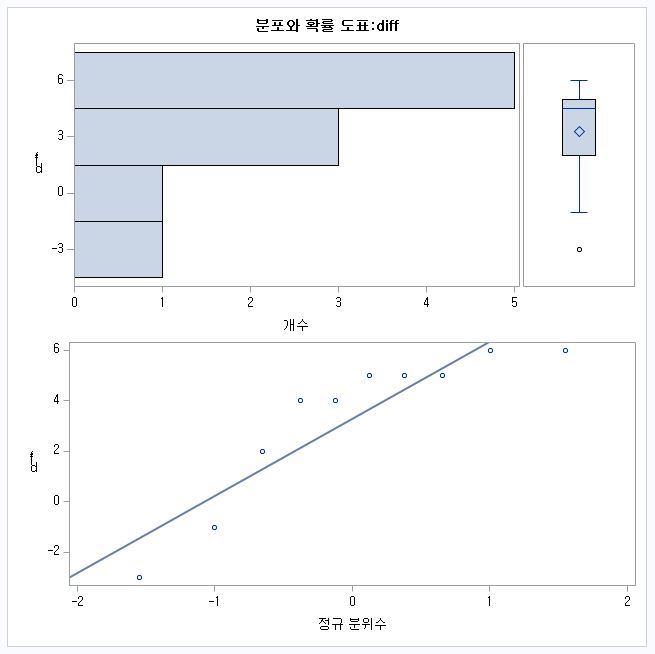
**run**;

**proc** **univariate** data=sasadv.deer normal plot;

var diff;

**run**;





**(가)**

연구자가 모든 사슴들(전체 모집단)의 앞다리 및 뒷다리를 조사하기에는 시간과 비용이 많이 들기 때문에, 적당한 표본(여기서는 사슴 10마리, 즉 n=10)을 추출하여 조사에 임하였다.

**(나)**

귀무가설 : 사슴의 앞다리와 뒷다리의 평균 다리길이 차이는 없다.

대립가설 : 사슴의 앞다리와 뒷다리의 평균 다리길이 차이가 있다.

**(다)**

통계적 가설에 대해 생각할 수 있는 두 가지 유형의 오류

1. 제 1종 오류 : 귀무가설이 사실인데 귀무가설을 기각하는 오류. 즉, 실제로는 사슴의 앞다리와 뒷다리의 평균 다리길이 차이가 없는데, 차이가 있다고 하는 경우이다.

2. 제 2종 오류 : 귀무가설이 거짓인데 귀무가설을 기각하지 않는 오류. 즉, 실제로는 사슴의 앞다리와 뒷다리의 평균 다리길이 차이가 있는데, 차이가 없다고 하는 경우이다.

**(라)**

“유의수준을 5%로 한다”라는 말의 의미는 제 1종오류를 범할 확률의 최대 허용한계를 5%로 하겠다는 의미이다. 즉, 실제로는 사슴의 앞다리와 뒷다리의 평균 다리길이 차이가 없는데, 차이가 있다고 할 확률의 최대 허용한계를 5%로 정하겠다는 말이다. 또한 제 1종 오류를 범할 확률을 미리 지정된 확률 이하로 하는 검정방법을 사용하는 이유는, 실제적으로 제 1종 오류가 보다 중요한 의미를 가지기 때문이다. 즉, 실제로는 사슴의 앞다리와 뒷다리의 평균 다리길이 차이가 없는데도 불구하고, 차이가 있다는 연구자의 주장을 받아들이게 되는 오류를 최소화하는 것이 중요하다.

**(마)**

기초통계량들을 살펴보면, 사슴의 앞다리와 뒷다리의 평균 다리길이 차이가 3.3이다. 또한 대부분의 값들이 양의 값을 가지는 것으로 보아, 일반적으로 사슴의 뒷다리의 길이가 앞다리의 길이보다 길다는 것을 알 수 있다. 그리고 제 1 사분위수와 제 3 사분위수를 보면 각각 2.0과 5.0으로, 전체 표본의 50%가 다리길이 차이 2.0~5.0 정도를 보이고 있음을 알 수 있다.

**(바)**

즐기-잎 그림과 상자그림을 보면 변수 diff의 분포형태가 왼쪽으로 치우진 분포라는 것을 알 수 있다.

**(사)**

가설 검정 결과를 보면, 검정통계량(t값) 값이 3.413793으로 유의수준 5% 하에서의 임계치인 1.383보다 크고 p-value 또한 0.0077이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 사슴의 앞다리와 뒷다리의 평균 다리길이 차이가 있다고 할 수 있다.