**SAS 데이터 분석 입문 10장**

**2019020650 김형욱**

**\* 10장 예제문제**

**<예 10-1>**

**data** sasadv.diff;

input diff @@;

cards;

-21.7 8.8 9.6 3.1 4.1 0.8

1.8 3.3 0.7 10.3 1.4 2.5

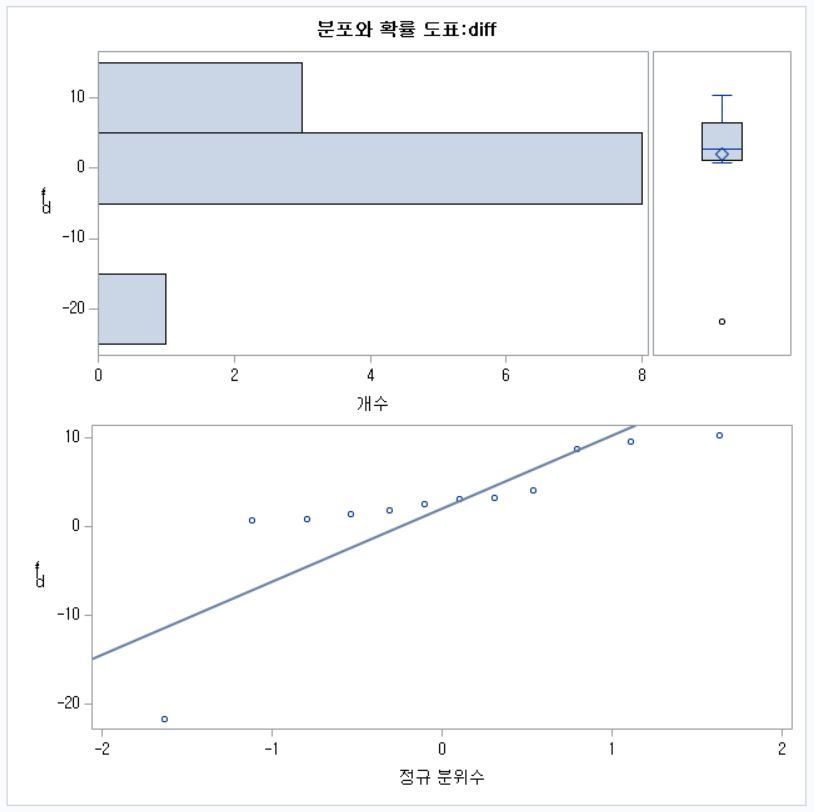
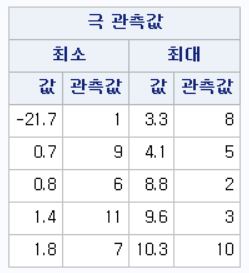
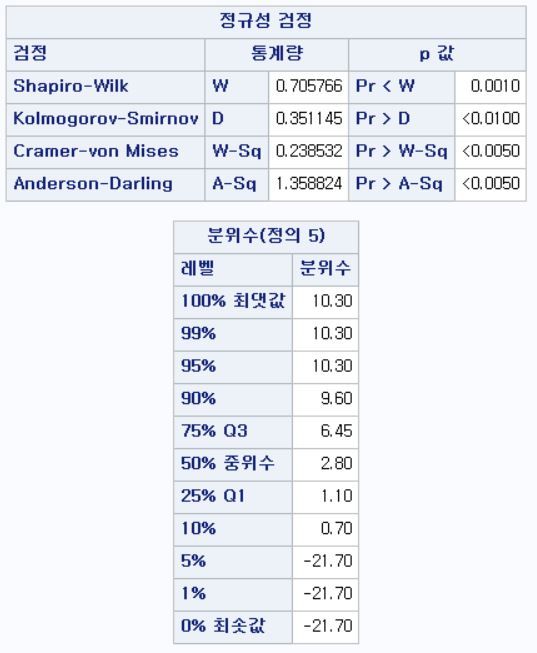
;

**run**;

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

var diff;

**run**;



해석 : 위 예제는 데이터의 정규성을 확인하기 위한 방법에 대한 것이다. 여기서 귀무가설은 “모집단의 분포가 정규분포이다”이고, 대립가설은 “모집단의 분포가 정규분포가 아니다”이다. Shapiro-Wilk test와 Kolmogorv-Smirnov test에서의 유의확률이 각각 0.001과 0.01보다 작은 값이므로 유의수즌 5% 하에서 귀무가설을 기각한다. 따라서 모집단의 분포가 정규분포라고 할 수 없다. 실제로 자료의 분포를 시각적으로 보여주는 ‘상자그림’과 ‘줄기-잎 그림’을 살펴보면, 데이터의 분포가 한쪽으로 치우쳐져 있고 매우 작은 극단값(-21.7)이 존재한다는 것을 알 수 있다.

**<예 10-2>**

**proc** **univariate** data=sasadv.cholest;

by gender;

probplot age / normal(mu=est sigma=est);

inset max min mean median std;

**run**;

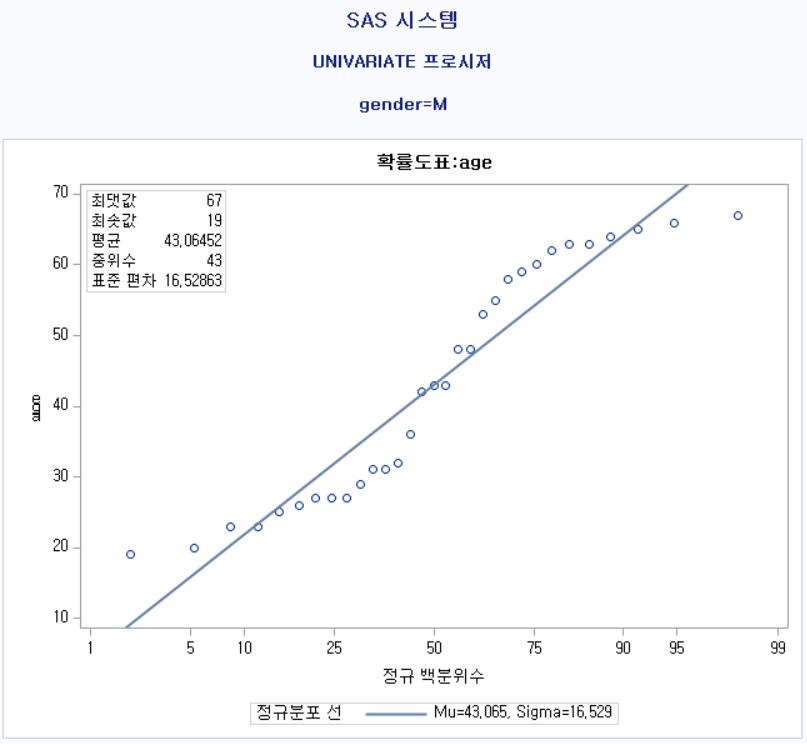
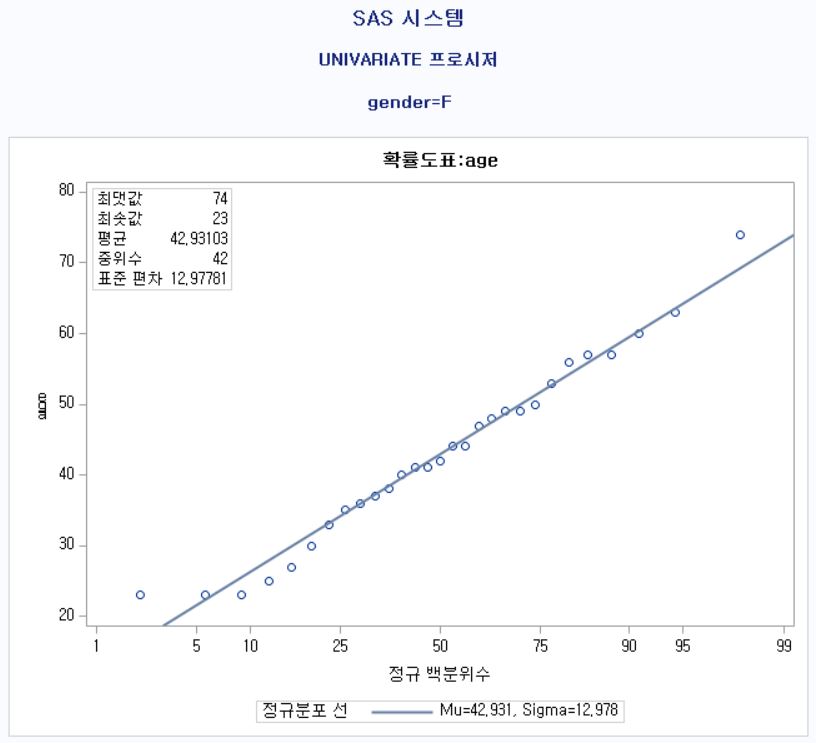
**proc** **univariate** data=sasadv.cholest;

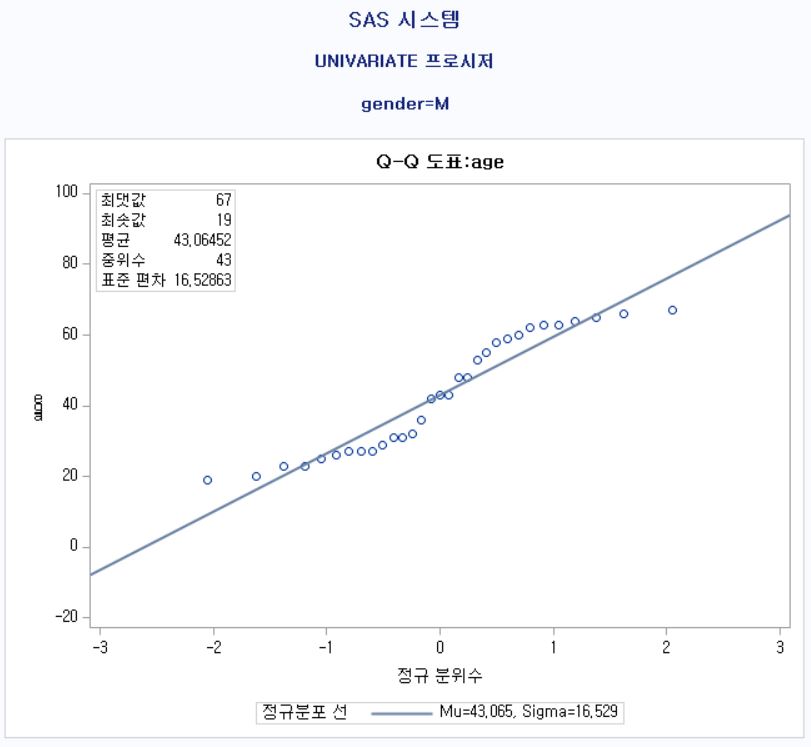
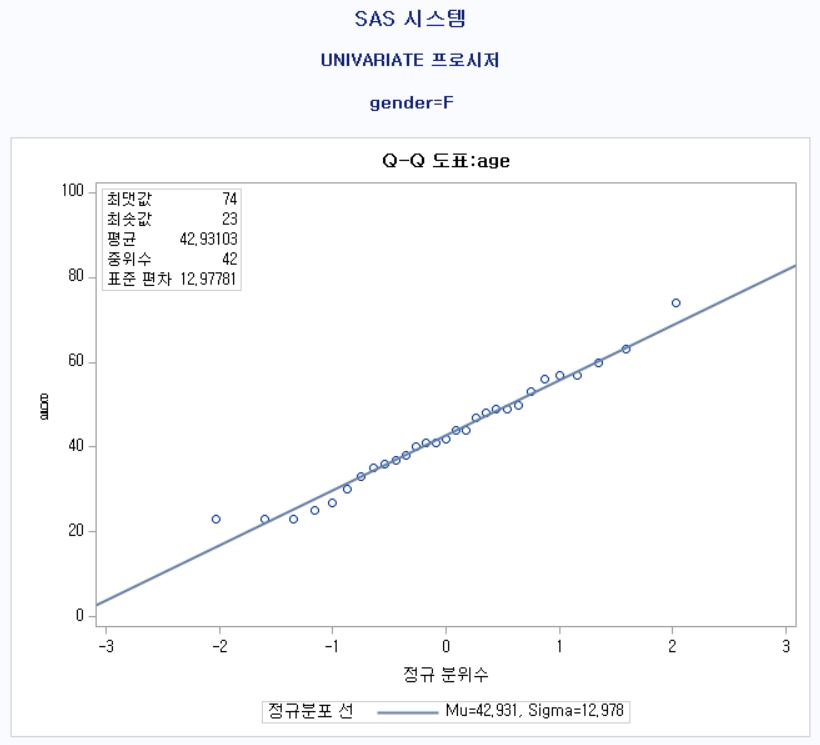
by gender;

qqplot age / normal(mu=est sigma=est);

inset max min mean median std;

**run**;





해석 : 위 예제는 ‘콜레스테롤 과포화율’ 데이터의 변수 age(나이)에 대해서 P-P plot과 Q-Q plot을 작성한 것이다. 그 결과, 여자집단의 경우 나이(age)의 분포가 정규분포와 잘 부합된다고 할 수 있지만, 남자집단의 경우에는 데이터 값을 나타내는 점들이 직선으로부터의 이탈 정도가 심하여 정규분포에 따른다고 가정하기가 어렵다는 것을 알 수 있다.

**<예 10-3>**

**proc** **univariate** data=sasadv.cholest;

class gender;

histogram age / normal(mu=est sigma=est)

outhistogram=outhisto

cfill=orange

vaxis=**0** to **40** by **5**

midpoints=**10** to **80** by **10**;

inset normal(ksd ksdpval);

**run**;

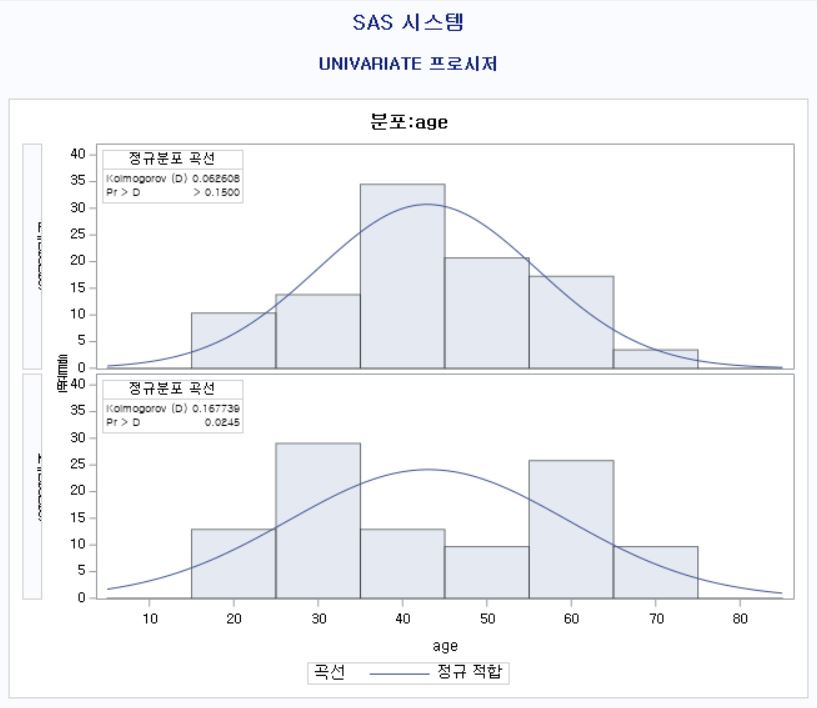
**proc** **univariate** data=sasadv.cholest;

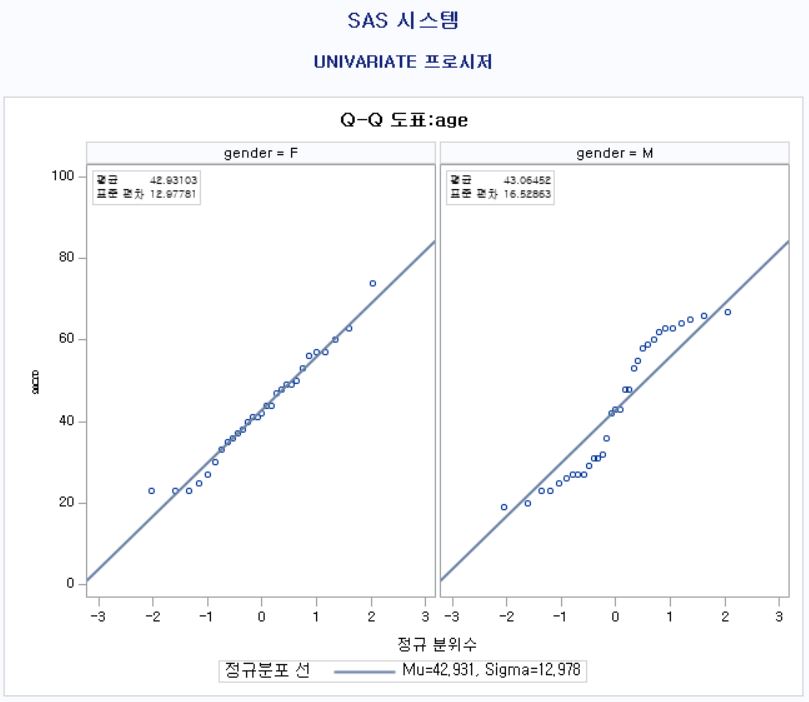
class gender;

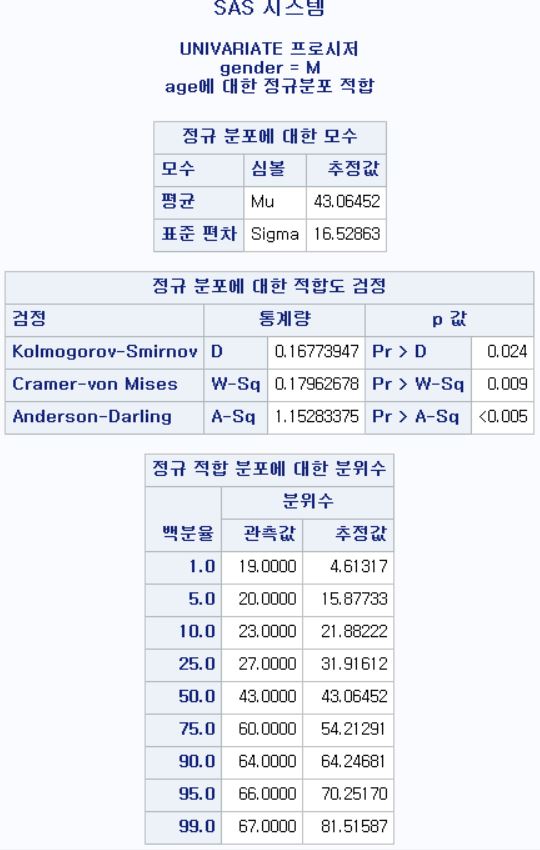
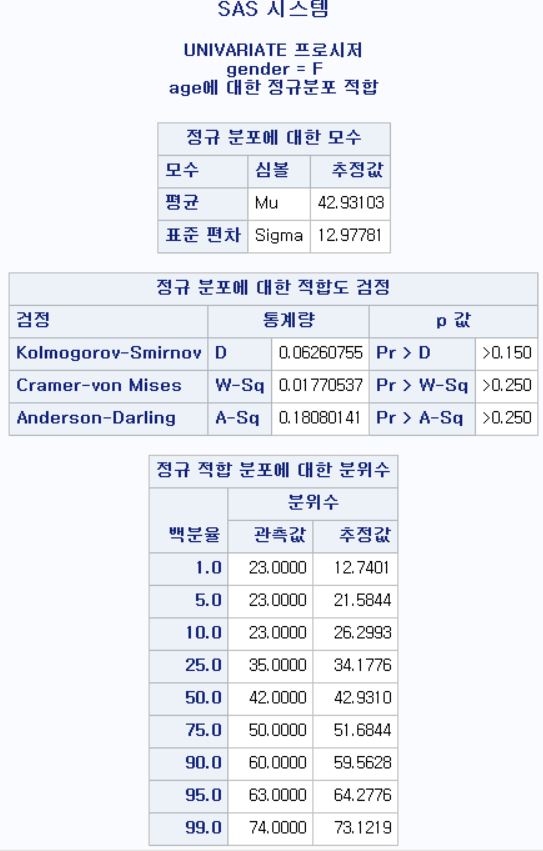
qqplot age / normal(mu=est sigma=est) nrow=**1** ncol=**2**;

inset mean std;

**run**;







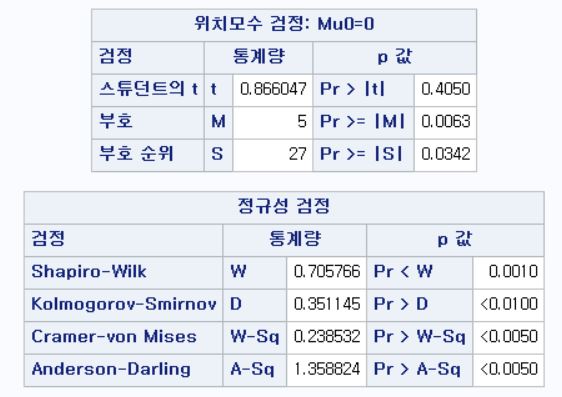
해석 : 위 예제는 ‘콜레스테롤 과포화율’ 데이터의 변수 age(나이)에 대해서 히스토그램을 작성한 것이다. 또한 HISTOGRAM 명령문을 사용하여 데이터가 가정된 특정 분포에 잘 부합되는지를 검정해보았다. 분석 결과에서 ‘정규 적합 분포에 대한 분위수’ 부분을 살펴보았다. 여자집단의 경우에는 실제 분위수들과 정규분포 하에서 추정된 분위수들이 비슷한 값들을 갖고, 검정통계량들의 유의확률이 비교적 커서 유의수준 5% 하에서 귀무가설(데이터가 정규분포를 따른다)을 기각하지 못하기 때문에 데이터가 정규분포에 잘 부합된다고 할 수 있다. 반면, 남자집단의 경우에는 실제 분위수들과 정규분포 하에서 추정된 분위수들이 비슷한 값들을 갖지 않고, 검정통계량들의 유의확률이 작아서 유의수준 5% 하에서 귀무가설(데이터가 정규분포를 따른다)을 기각하기 때문에 데이터가 정규분포에 잘 부합되지 않는다고 할 수 있다.

**<예 10-4>**

**proc** **univariate** data=sasadv.diff normal mu0=**0**;

var diff;

**run**;



해석 : 위 예제는 어떤 건강보조음료의 효능을 알아보기 위해 12명의 사람들로부터 실험 전(pre)과 후(post)의 신체특성의 차이(diff=post-pre)를 부호검정 한 것이다. 검정 결과, 부호검정의 검정통계량 값이 5이고 유의확률이 0.0063이므로 유의수준 5% 하에서 귀무가설(diff=0)을 기각한다. 따라서 실험 전(pre)과 후(post)의 신체특성의 차이(diff=post-pre)가 있다고 할 수 있다. 실제로 자료를 살펴보면, 12개의 표본 중 첫 번째 표본의 경우만 실험 후(post)의 수치가 실험 전(pre)보다 작으며(-21.7), 수치의 차이가 다른 표본들에 비해서 매우 크다는 것을 알 수 있다. 즉, 첫 번째 표본은 일종의 특이치(outlier)로 간주할 수 있는데, 이 경우 t-검정의 결과는 첫 번째 표본에 큰 영향을 받는다는 것을 알 수 있다. 반면에 부호검정을 사용하면, 부호만을 사용하기 때문에 위와 같이 특이치(outlier)에 큰 영향을 받지 않는다는 것을 알 수 있다.

**<예 10-5>**

**data** sasadv.mice;

input drug $ days @@;

cards;

A 1 A 1 A 3 A 3 A 4

B 3 B 4 B 4 B 10 B 10 B 26

;

**run**;

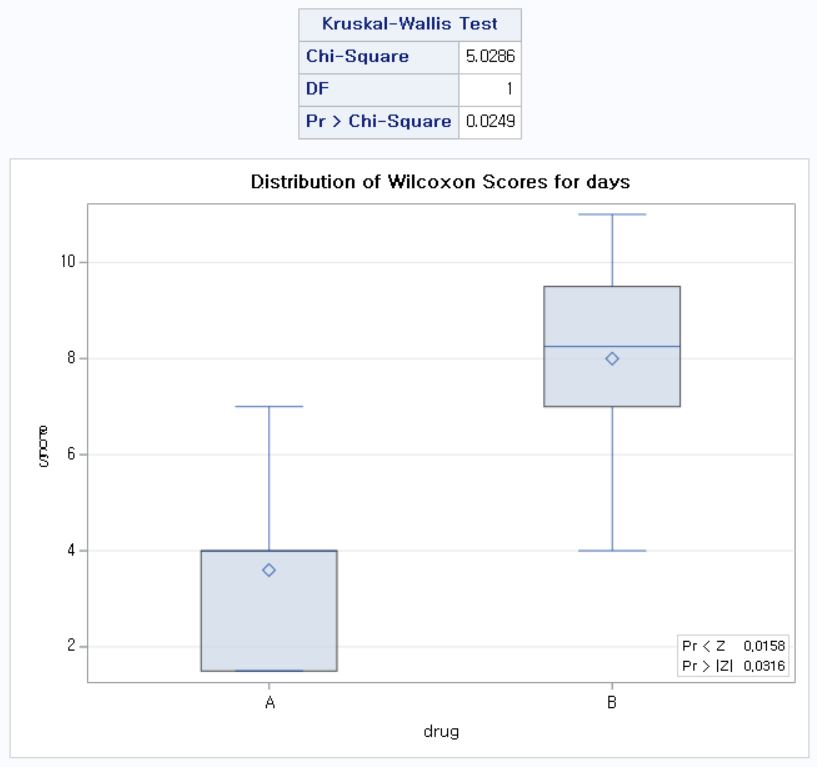
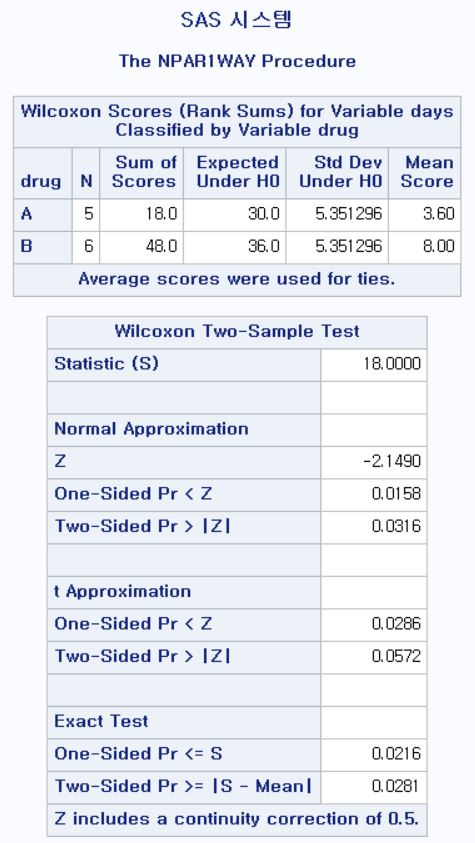
**proc** **npar1way** data=sasadv.mice wilcoxon;

exact wilcoxon;

class drug;

var days;

**run**;



해석 : 위 예제는 두 개의 약품(drug : A, B)을 각각 5마리와 6마리의 쥐에게 투여한 후 생존시간(days)을 측정한 데이터를 사용하였고, 두 약품에 대한 평균 생존시간이 동일한지를 확인하기 위해 Wilcoxon’s Rank Sum Test를 한 것이다. 여기서 귀무가설은 “두 약품에 대한 평균 생존시간이 동일하다”이고, 대립가설은 “두 약품에 대한 평균 생존시간이 동일하지 않다”이다. 검정 결과, 검정통계량 값은 18이고 근사 유의확률이 0.0316, 정확한 유의확률이 0.0281이므로 모두 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 두 약품에 대한 평균 생존시간이 동일하지 않다고 할 수 있다.

**<예 10-6>**

**data** sasadv.perfume;

input company $ weight @@;

cards;

A 117.1 A 121.3 A 127.8 A 121.9

A 117.4 A 124.5 A 119.5 A 115.1

B 123.5 B 125.3 B 126.5 B 127.9

B 122.1 B 125.6 B 129.8 B 117.2

;

**run**;

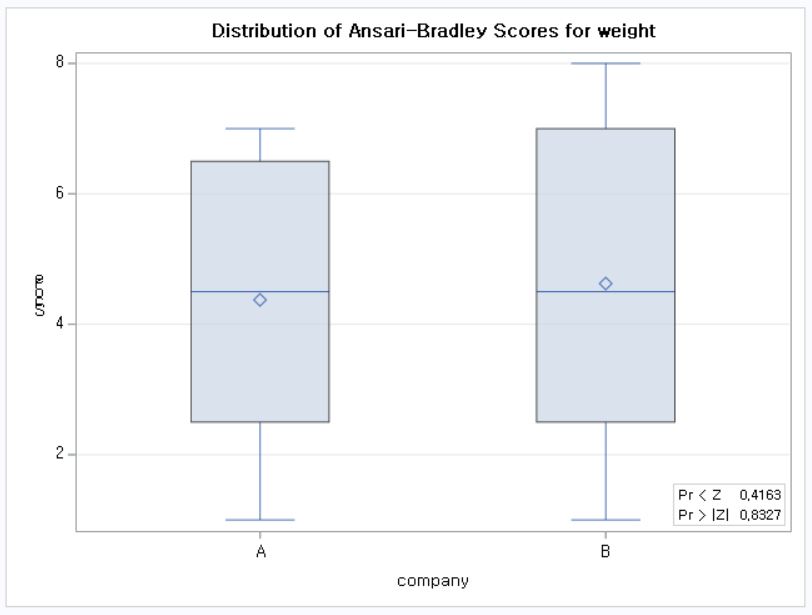
**proc** **npar1way** data=sasadv.perfume AB;

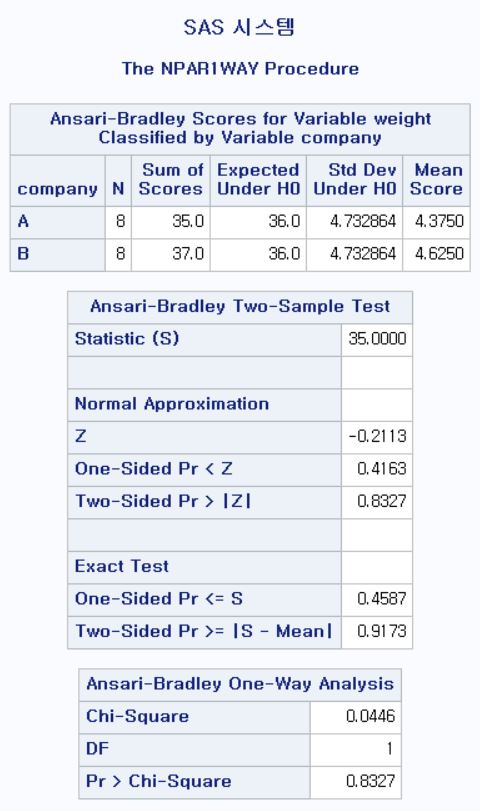
exact AB;

class company;

var weight;

**run**;





해석 : 위 예제는 두 회사(company : A, B)에서 제조한 계피향료의 무게를 조사한 데이터에 대해 무게(weight)의 분산이 같은지를 알아보고자 하는 것이다. 여기서 귀무가설은 “두 회사에서 제조한 계피향료 무게의 분산은 동일하다”이고, 대립가설은 “두 회사에서 제조한 계피향료 무게의 분산은 동일하지 않다”이다. Ansari-Bradley Test 결과, 검정통계량 값이 35이고 근사 유의확률은 0.8327, 정확한 유의확률은 0.9173이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 두 회사에서 제조한 계피향료 무게의 분산은 동일하다고 할 수 있다.

**<예 10-7>**

**data** sasadv.health;

input pre post @@;

diff=post-pre;

cards;

63.1 41.4 46.5 55.3 40.1 49.7 48.5 51.6

33.2 37.3 38.8 39.6 48.3 50.1 40.5 43.8

36.9 37.6 40.5 50.8 42.2 43.6 34.9 37.4

;

**run**;

**proc** **univariate** data=sasadv.health;

var diff;

**run**;



해석 : 위 예제는 어떤 건강보조음료의 효능을 알아보기 위해 12명의 사람들로부터 실험 전(pre)과 후(post)의 신체특성의 차이(diff=post-pre)를 부호검정 및 부호순위검정을 수행한 것이다. 여기서 귀무가설은 “실험 전과 후의 신체특성의 차이가 없다”이고, 대립가설은 “실험 전과 후의 신체특성의 차이가 있다”이다. 검정 결과, 부호검정과 부호순위검정의 유의확률이 각각 0.0063과 0.0342이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 실험 전과 후의 신체특성의 차이가 있다고 할 수 있다.

**<예 10-8>**

**data** sasadv.fly;

input type $ effect @@;

cards;

A 72 A 65 A 67 A 75 A 62 A 73

B 55 B 59 B 68 B 70 B 53 B 50

C 64 C 74 C 61 C 58 C 51 C 69

;

**run**;

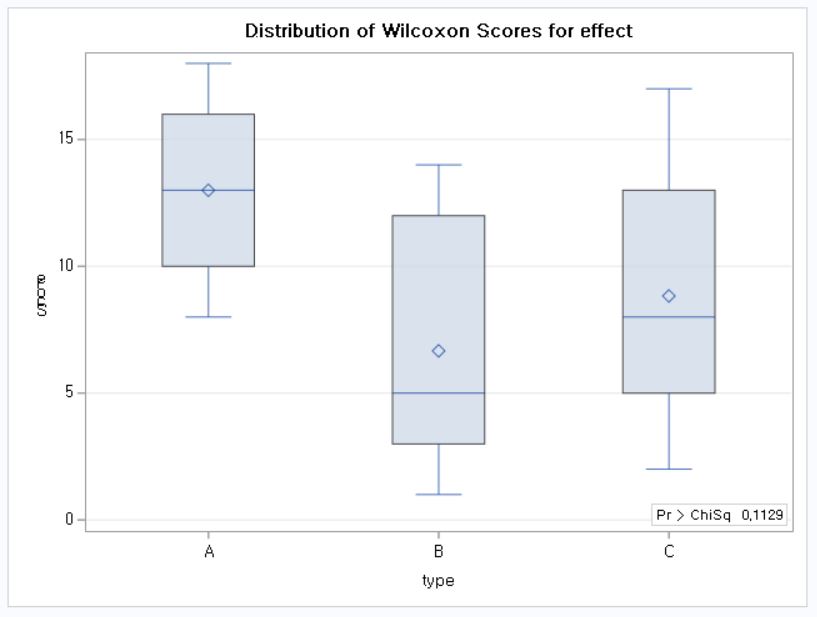
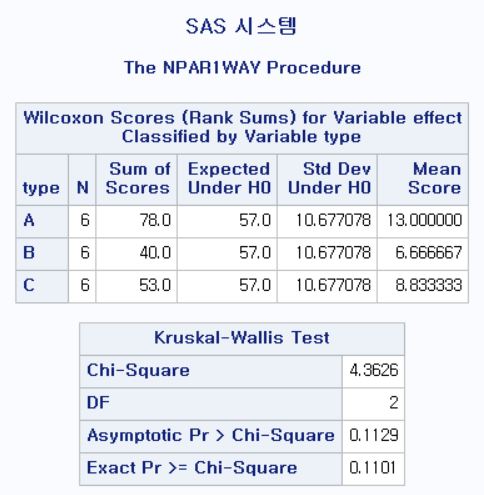
**proc** **npar1way** data=sasadv.fly wilcoxon;

exact wilcoxon;

class type;

var effect;

**run**;



해석 : 위 예제는 세 종류의 파리약(type : A, B, C)의 살충효과(effect, %)를 조사한 데이터에 대해, 파리약의 종류에 따른 살충효과의 차이를 알아보고자 Kruskal-Wallis Test를 수행한 것이다. 여기서 귀무가설은 “파리약의 종류에 따른 살충효과가 차이가 없다”이고, 대립가설은 “파리약의 종류에 따른 살충효과가 차이가 있다”이다. 검정 결과, 검정통계량 값이 4.3626이고 근사 유의확률은 0.1129, 정확한 유의확률은 0.1101이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 파리약의 종류에 따른 살충효과가 차이가 없다고 할 수 있다.

**\* 10장 연습문제**

**<연습문제 10-1>**

**data** sasadv.ex10\_1;

input Zn @@;

cards;

437 358 72 43 107 223 60 72 54 35 70 20 34 24 24 51 23

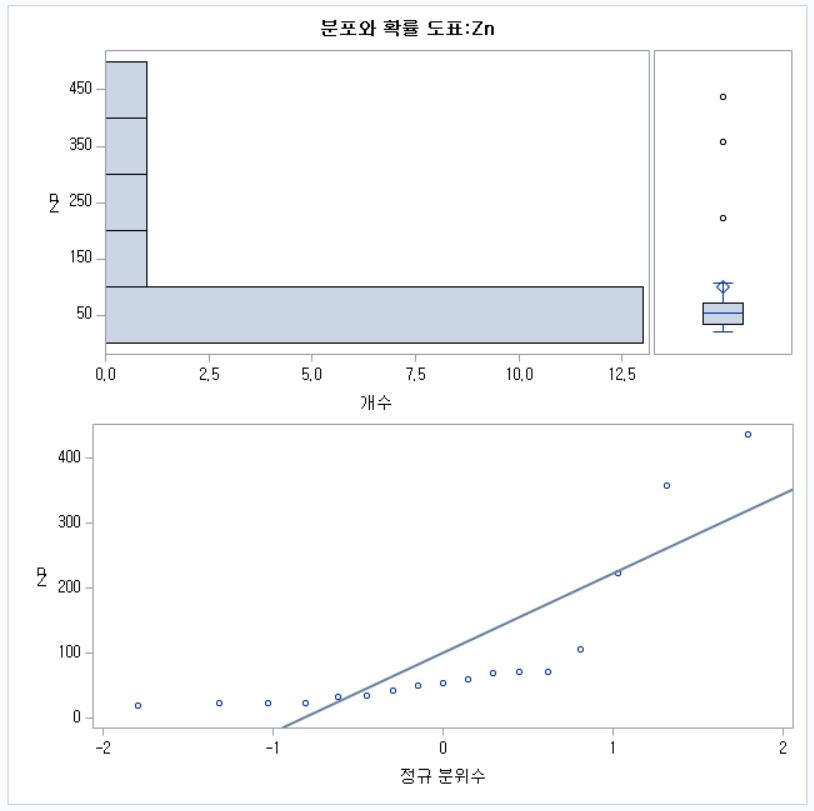
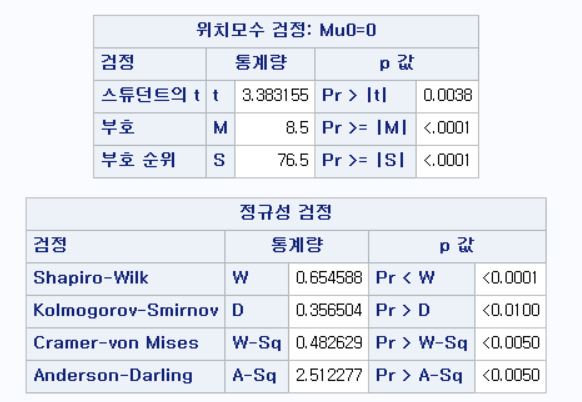
;

**run**;

**proc** **univariate** data=sasadv.ex10\_1 normal plot;

var Zn;

**run**;



해석 : 위 문제에서는 17명의 정상인에 대해 피부생체조직의 아연집중도를 조사한 데이터를 사용하였고, 아연집중도는 정규분포를 따른다고 할 수 있는지를 알아보기 위해 적합도 검정을 수행한 것이다. 여기서 귀무가설은 “아연집중도는 정규분포를 따른다”이고, 대립가설은 “아연집중도는 정규분포를 따르지 않는다”이다. Shapiro-Wilk Test 및 Kolmogorov-Smirnov Test 결과, 모두 유의확률이 유의수준 0.05보다 작으므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 아연집중도는 정규분포를 따른다고 할 수 없다. 또한 ‘상자그림’ 및 ‘줄기-잎 그림’을 통해 데이터의 분포를 시각적으로 확인해봐도 정규분포를 따르지 않는다는 것을 알 수 있다.

**<연습문제 10-2>**

**data** sasadv.ex10\_2;

input paint $ time;

cards;

Old 6.6

Old 5.8

Old 7.8

Old 5.7

Old 6.0

Old 8.4

Old 8.8

Old 8.4

Old 7.3

Old 5.8

Old 5.8

Old 6.5

New 6.4

New 5.8

New 7.4

New 5.5

New 6.3

New 7.8

New 8.6

New 8.2

New 7.0

New 4.9

New 5.9

New 6.5

;

**run**;

/\* 데이터의 정규성 검정 \*/

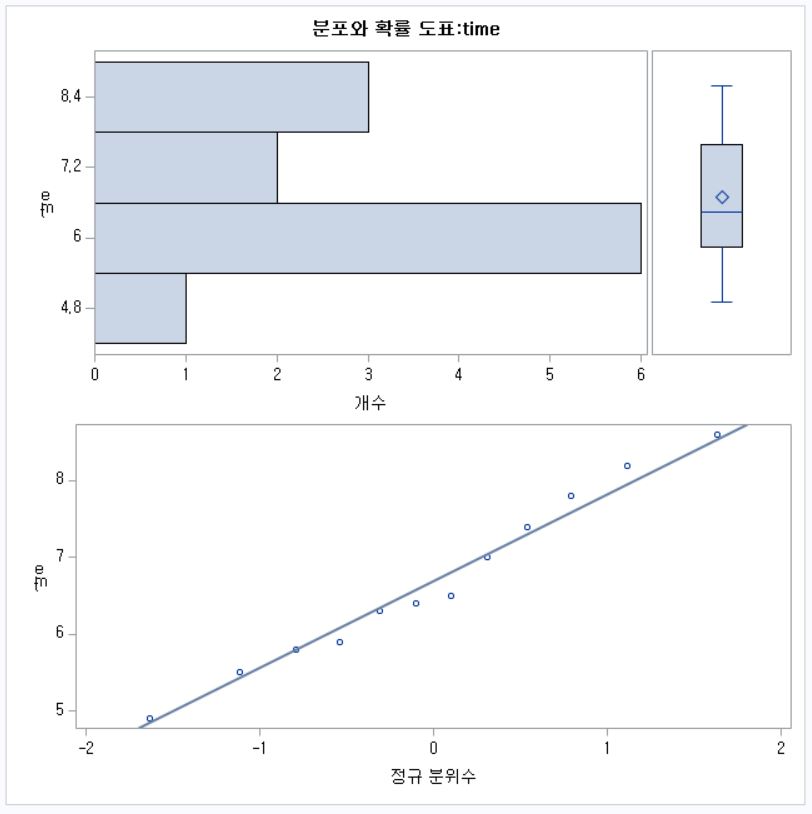
**proc** **univariate** data=sasadv.ex10\_2 normal plot;

class paint;

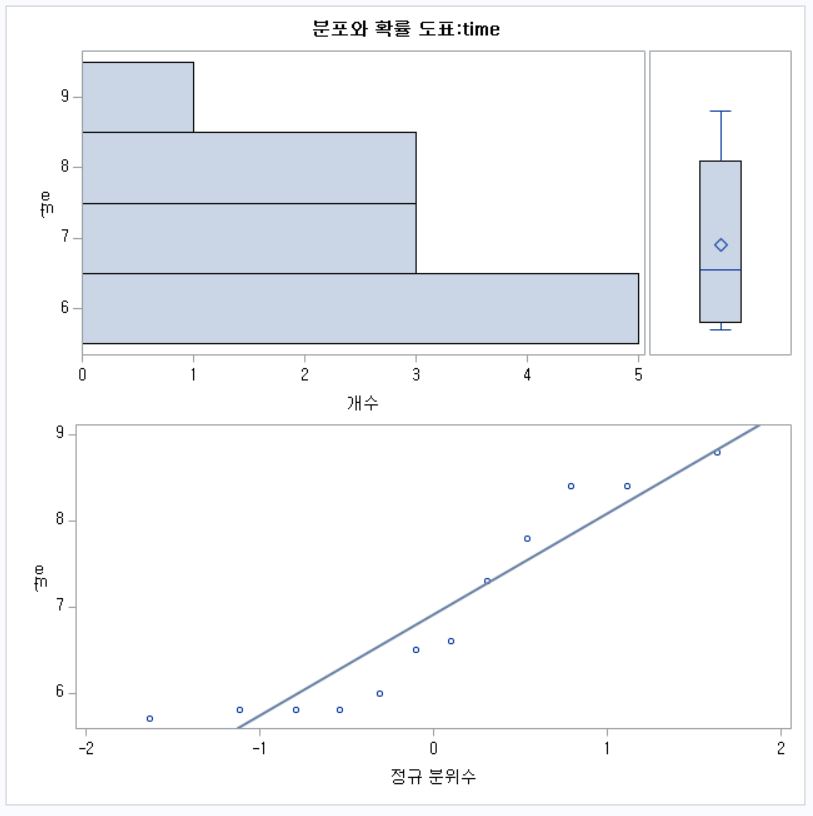
var time;

**run**;

**1. 새로운 페인트(New)의 경우**



**2. 기존의 페인트(Old)인 경우**



해석 : 먼저 새로운 페인트를 사용한 경우 데이터의 정규성 검정 결과, Shapiro-Wilk Test의 유의확률이 0.9333이므로 유의수준 5% 하에서 귀무가설(데이터가 정규분포를 따른다)을 기각하지 못한다. 따라서 데이터가 정규분포를 따른다고 할 수 있다. 또한 ‘상자그림’ 및 ‘줄기-잎 그림’을 살펴보아도, 데이터가 정규분포를 따름을 알 수 있다. 반면에 기존의 페인트를 사용한 경우 데이터의 정규성 검정 결과, Shapiro-Wilk Test의 유의확률이 0.0462이므로 유의수준 5% 하에서 귀무가설(데이터가 정규분포를 따른다)을 기각한다. 따라서 데이터가 정규분포를 따른다고 할 수 없다. 또한 ‘상자그림’ 및 ‘줄기-잎 그림’을 살펴보아도, 데이터가 정규분포를 따르지 않음을 알 수 있다. 그러므로 비모수적인 방법을 사용하는 것이 적절해보인다.

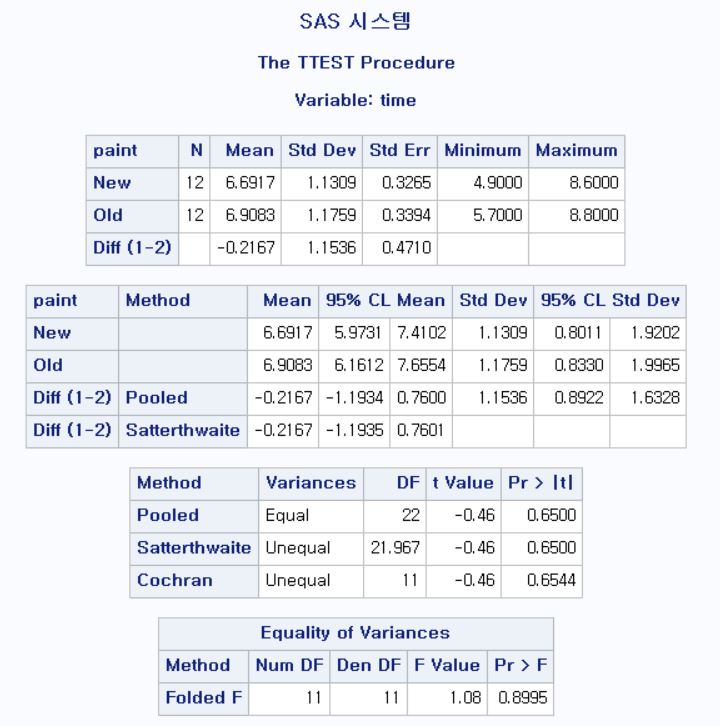
/\* 1. t-test \*/

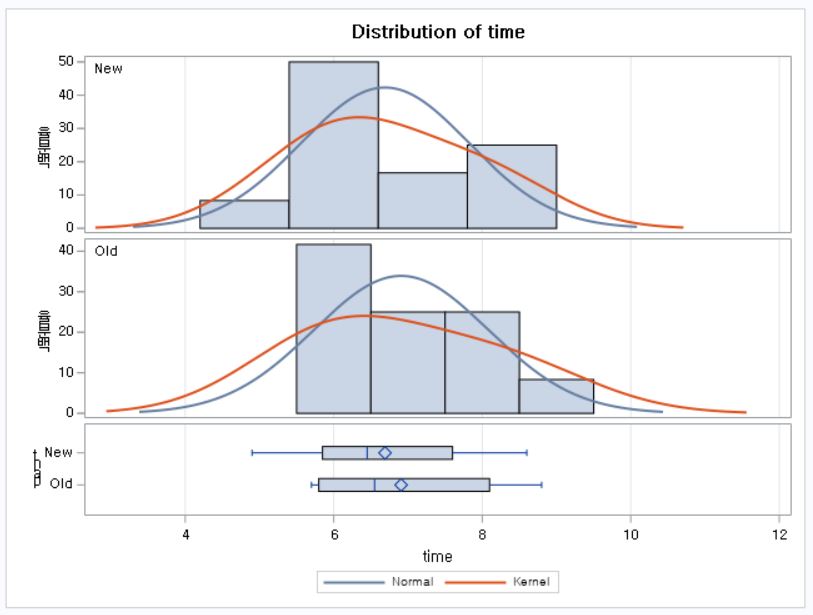
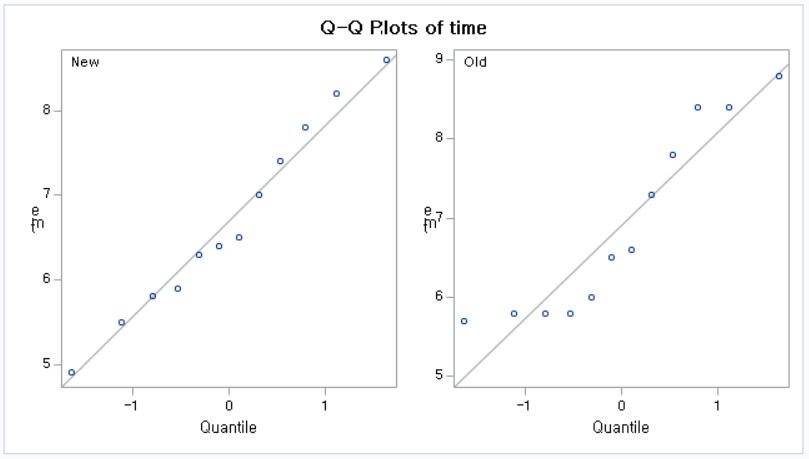
**proc** **ttest** data=sasadv.ex10\_2 cochran;

class paint;

var time;

**run**;



해석 : 먼저 모분산의 동일성에 대한 검정 결과를 살펴보면, 유의확률이 0.8995이므로 유의수준 5% 하에서 귀무가설(모분산이 동일하다)을 기각하지 못한다. 따라서 등분산의 가정을 만족한다고 할 수 있다. ‘Pooled’에 해당되는 유의확률 값을 확인해보면 0.65로 유의수준 5% 하에서 귀무가설(기존의 페인트와 새로운 페인트의 마르는 시간에는 차이가 없다)을 기각하지 못한다. 따라서 기존의 페인트와 새로운 페인트의 마르는 시간에는 차이가 없다고 할 수 있다.

/\* 2. rank-sum test \*/

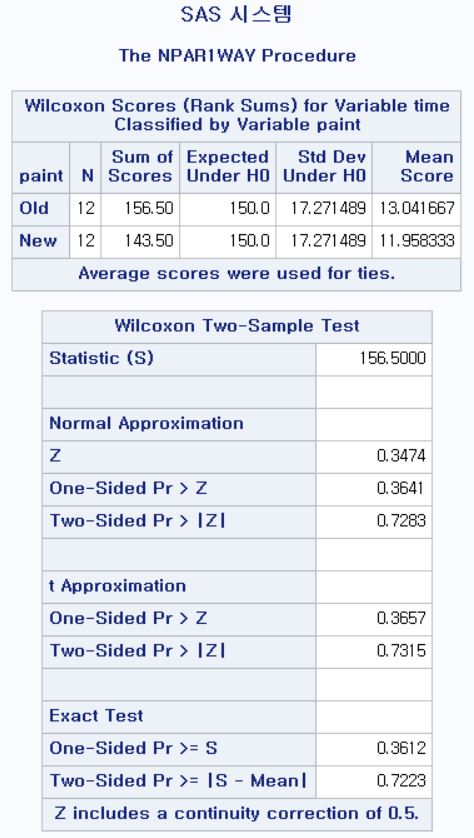
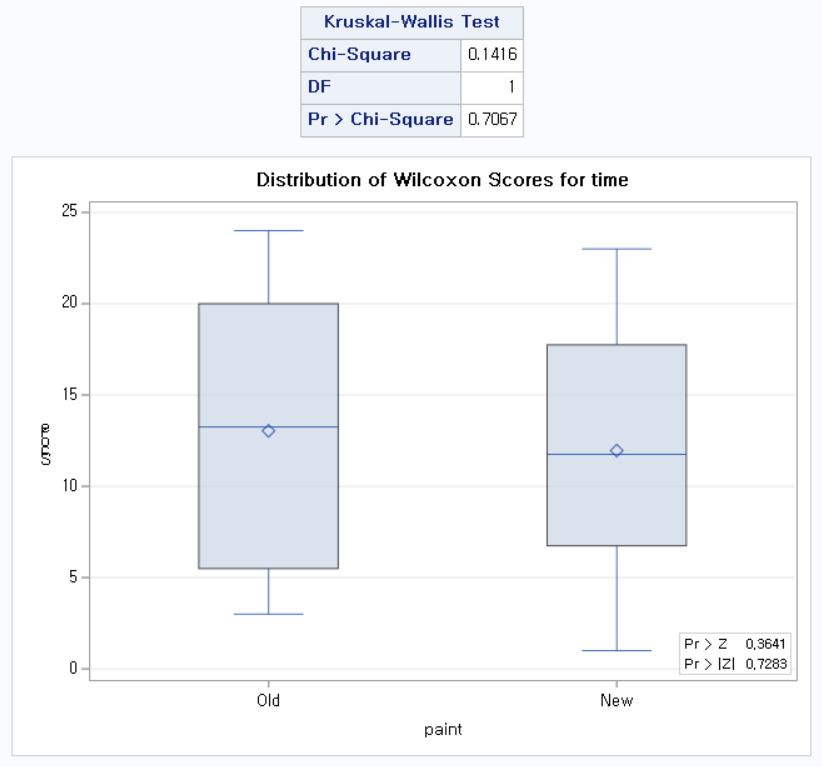
**proc** **npar1way** data=sasadv.ex10\_2 wilcoxon;

exact wilcoxon;

class paint;

var time;

**run**;

해석 : 이번에는 동일한 데이터에 대하여 Wilcoxon’s Rank Sum Test를 수행하였다. 여기서 귀무가설은 “기존의 페인트와 새로운 페인트의 마르는 시간에는 차이가 없다”이고, 대립가설은 “기존의 페인트와 새로운 페인트의 마르는 시간에는 차이가 있다”이다. 검정 결과, 정확한 유의확률은 0.7223이고 근사 유의확률은 0.7283이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 기존의 페인트와 새로운 페인트의 마르는 시간에는 차이가 없다고 할 수 있다.

**<연습문제 10-3>**

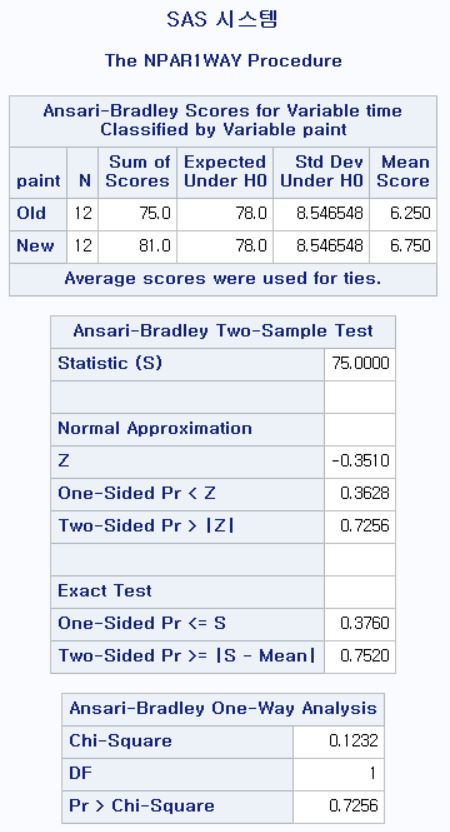
**proc** **npar1way** data=sasadv.ex10\_2 AB;

exact AB;

class paint;

var time;

**run**;



해석 : 위 문제는 <연습문제 10-2>의 데이터에 대해 Ansari-Bradley Test를 수행한 것이다. 여기서 귀무가설은 “두 페인트의 건조시간의 분산에 차이가 없다”이고, 대립가설은 “두 페인트의 건조시간의 분산에 차이가 있다”이다. 검정 결과, 정확한 유의확률은 0.7520이고 근사 유의확률은 0.7256이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 두 페인트의 건조시간의 분산에 차이가 없다고 할 수 있다.

**<연습문제 10-4>**

**data** sasadv.ex10\_4;

input group $ score @@;

cards;

A 772 A 764 A 600 A 564

B 792 B 612 B 592

C 752 C 680 C 624 C 580 C 572

;

**run**;

/\* 1. 일원분류 분산분석 \*/

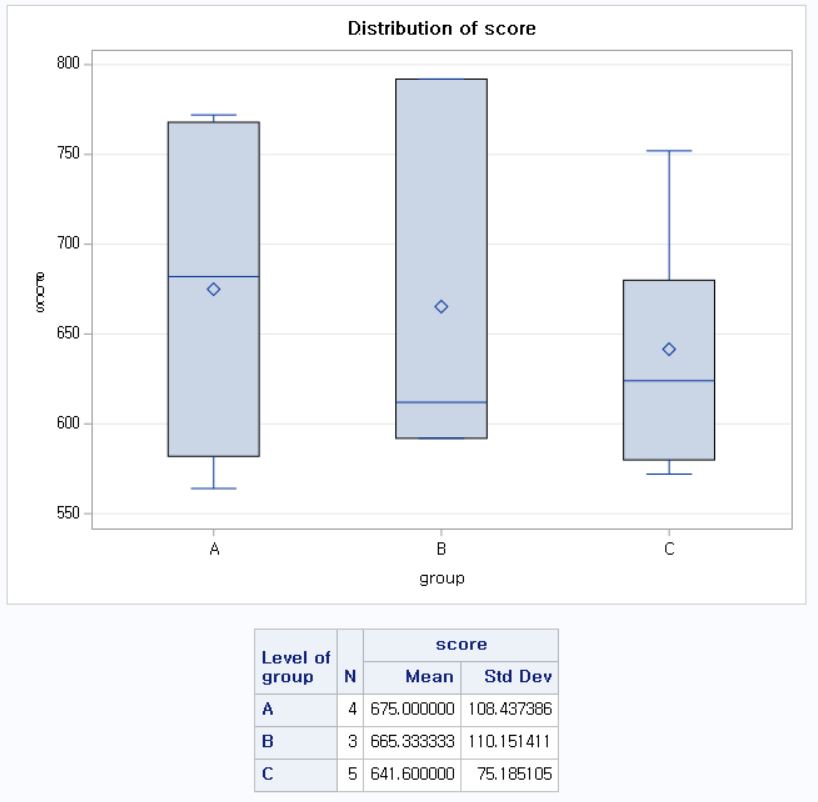
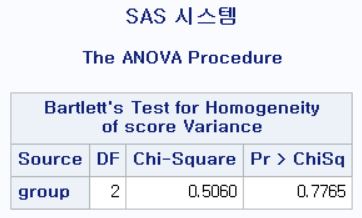
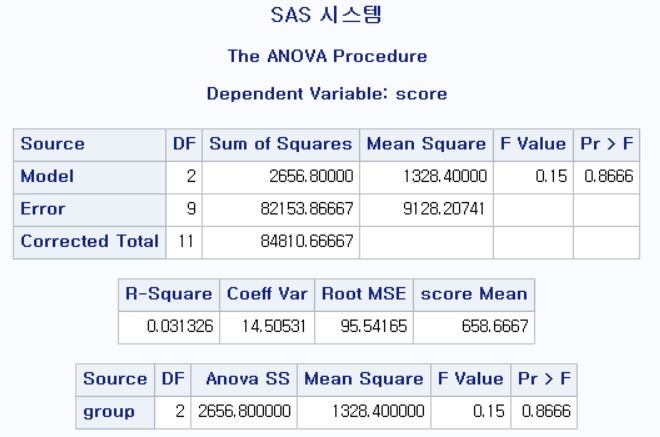
**proc** **anova** data=sasadv.ex10\_4;

class group;

model score=group;

means group / hovtest=bartlett;

**run**;



해석 : 먼저 세 그룹의 학생들에게 시험을 치른 다음 얻은 성적에 관한 데이터로 일월분류 분산분석을 수행한 것이다. 여기서 귀무가설은 “세 그룹 간 학생들의 평균 성적이 차이가 없다”이고, 대립가설은 “세 그룹 간 학생들의 평균 성적이 차이가 있다”이다. 우선 세 그룹의 모분산이 동일한지를 검정해본 결과, Bartlett 검정의 유의확률이 0.7765이므로 유의수준 5% 하에서 귀무가설(세 그룹의 모분산이 동일하다)을 기각하지 못한다. 따라서 세 그룹의 모분산이 동일하다고 할 수 있다. 그리고 분산분석 결과, 유의확률이 0.8666이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 세 그룹 간 학생들의 평균 성적이 차이가 없다고 할 수 있다.

/\* 2. Kruskal-Wallis test \*/

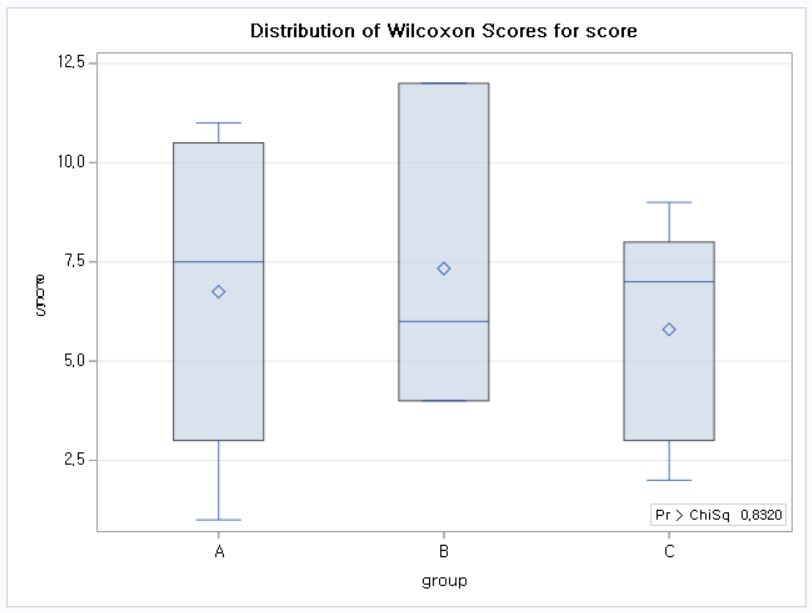
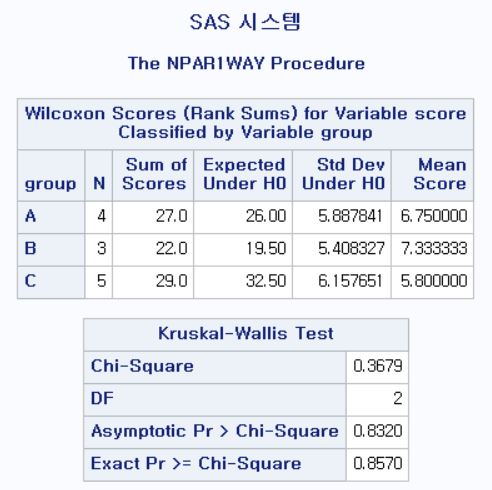
**proc** **npar1way** data=sasadv.ex10\_4 wilcoxon;

exact wilcoxon;

class group;

var score;

**run**;



해석 : 다음은 동일한 데이터에 대해 Kruskal-Wallis Test를 수행한 것이다. 여기서 귀무가설은 “세 그룹 간 학생들의 평균 성적이 차이가 없다”이고, 대립가설은 “세 그룹 간 학생들의 평균 성적이 차이가 있다”이다. 검정 결과, 검정통계량 값은 0.3679이고 정확한 유의확률은 0.8570이고 근사 유의확률은 0.8320으로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 세 그룹 간 학생들의 평균 성적이 차이가 없다고 할 수 있다.