**SAS 데이터 분석 입문 6장**

**2019020650 김형욱**

**\* 6장 예제문제**

**<예 6-1>**

**data** sasadv.harvest;

input fertil $ yield @@;

cards;

F1 148 F1 76 F1 134 F1 98

F2 166 F2 153 F2 255

F3 264 F3 214 F3 327 F3 304

F4 335 F4 436 F4 423 F4 380 F4 465

;

**run**;

**proc** **anova** data=sasadv.harvest;

class fertil;

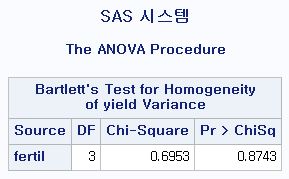
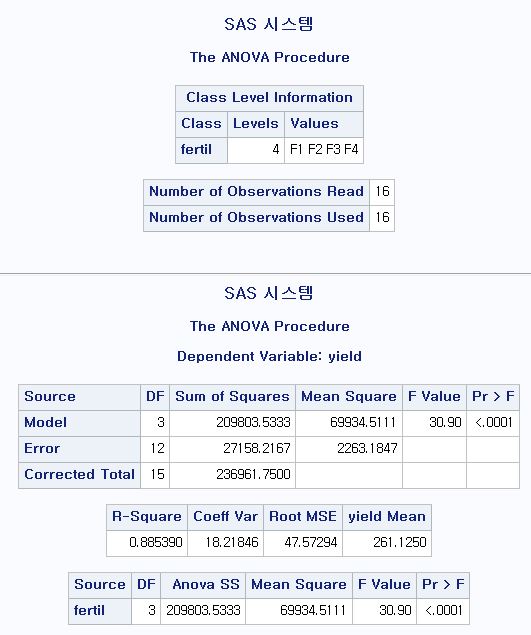
model yield=fertil;

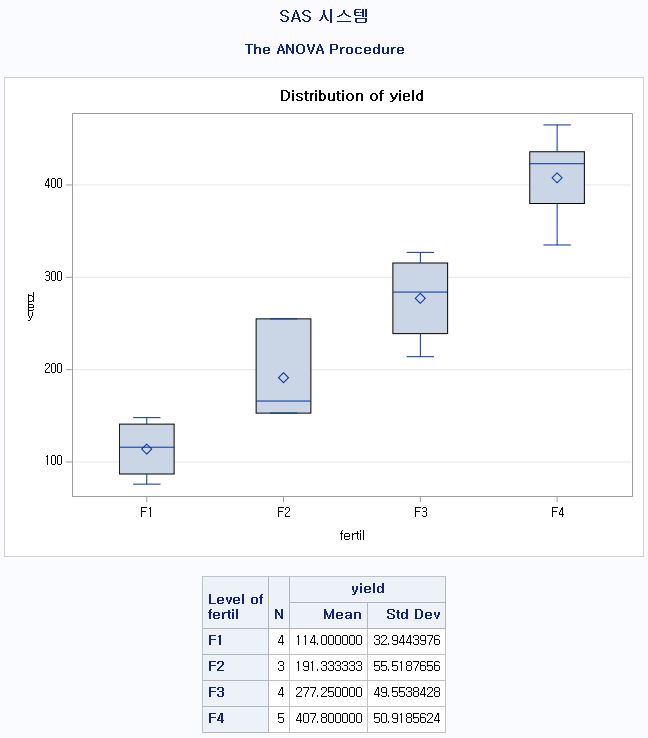
means fertil / hovtest=bartlett;

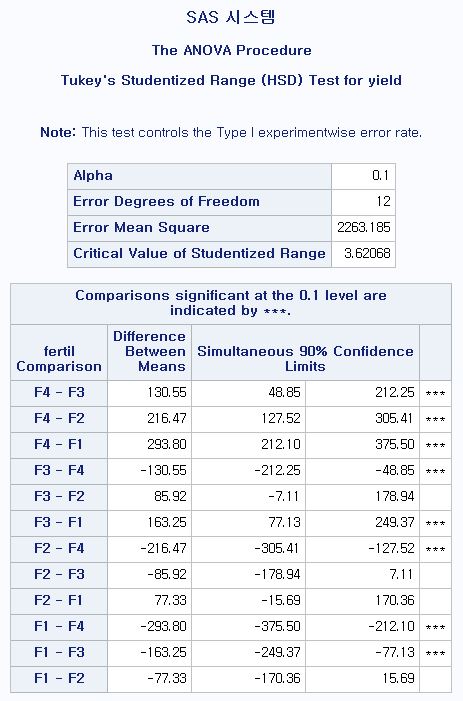
means fertil / tukey cldiff alpha=**0.10**;

means fertil / tukey lines alpha=**0.10**;

**run**;









해석 : 위 문제는 ‘수확량’ 데이터에 대해 일원분류 분산분석을 돌린 결과이다. 귀무가설은 “비료에 따라 농작물의 평균 수확량에 차이가 없다”이고, 대립가설은 “비료에 따라 농작물의 평균 수확량에 차이가 있다”이다. 우선 모분산의 동일성 검정 결과를 보면, Bartlett 검정에서 유의확률이 0.8743으로 유의수준 5% 하에서 모분산이 동일하는 귀무가설을 기각하지 못한다. 따라서 각 처리수준에서의 모분산들이 동일하다는 가정을 만족한다고 할 수 있다. 또한 분산분석 결과, F-값이 30.90이고 유의확률이 0.0001이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 비료에 따라 농작물의 평균 수확량에 차이가 있다고 할 수 있다. 그래서 사후분석으로 다중비교를 해본 결과, 비료의 종류 (F4, F3), (F4, F2), (F4, F1), (F3, F1)간에는 수확량의 모평균에 차이가 있다고 나왔다.

**<예 6-2>**

**data** sasadv.prefer;

do product = 'A1', 'A2', 'A3', 'A4';

do customer = **1** to **5** by **1**;

input prefer @@;

output;

end;

end;

cards;

5 7 9 10 8

2 3 4 5 2

4 7 6 5 7

6 4 2 2 1

;

**run**;

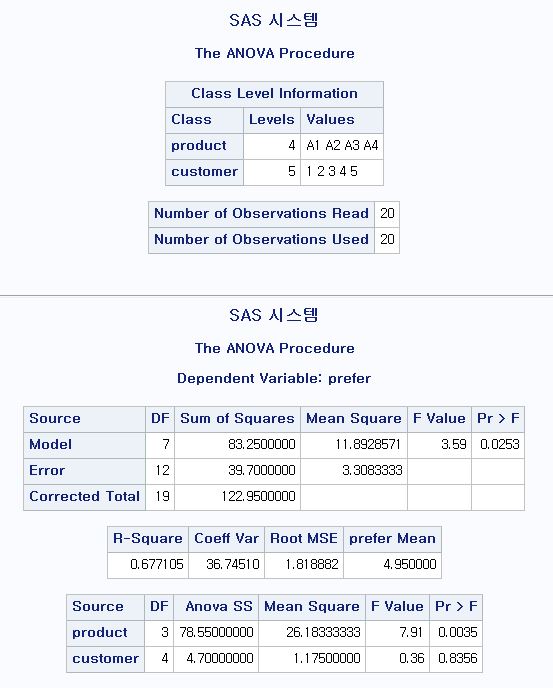
**proc** **anova** data=sasadv.prefer;

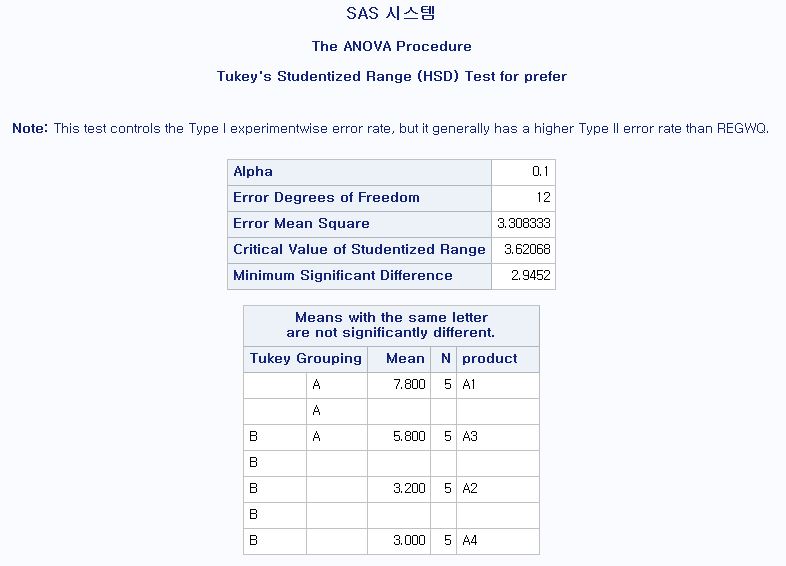
class product customer;

model prefer=product customer;

means product / duncan tukey alpha=**0.10**;

**run**;





해석 : 위 문제는 ‘제품선호도’ 데이터에 대해서 반복이 없는 이원배치 분산분석을 한 결과이다. 변수 product(제품)에 대한 귀무가설은 “제품의 종류에 따른 모평균에 차이가 없다”이고, 대립가설은 “제품의 종류에 따른 모평균에 차이가 있다”이다. 변수 customer(소비자)에 대한 귀무가설은 “소비자에 따라 제품의 선호도에 차이가 없다”이고, 대립가설은 “소비자에 따라 제품의 선호도에 차이가 있다”이다. 분산분석 결과, 변수 product(제품)에 대한 F-값은 7.91이고 유의확률은 0.0035이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 제품의 종류에 따른 모평균의 차이는 있다고 할 수 있다. 또한 변수 customer(소비자)에 대한 F-값은 0.36이고 유의확률은 0.8356이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 소비자에 따라 제품의 선호도에 차이가 없다고 할 수 있다. 마지막으로 통계적으로 유의한 차이를 보인 변수 product(제품)에 대하여 다중비교를 실시한 결과를 보면, 제품 ‘A1’과 ‘A3’의 경우가 ‘A2’와 ‘A4’에 비해서 선호도의 평균이 높다는 것을 알 수 있다.

**<예 6-3>**

**data** sasadv.sales;

do city = 'Lagre ', 'Middle', 'Small ';

do design = 'A', 'B', 'C';

do rep = **1**, **2**, **3**;

input sales @@;

output;

end;

end;

end;

cards;

23 20 21 22 19 20 19 18 21

22 20 19 24 25 22 20 19 22

18 18 16 21 23 20 20 22 24

;

**run**;

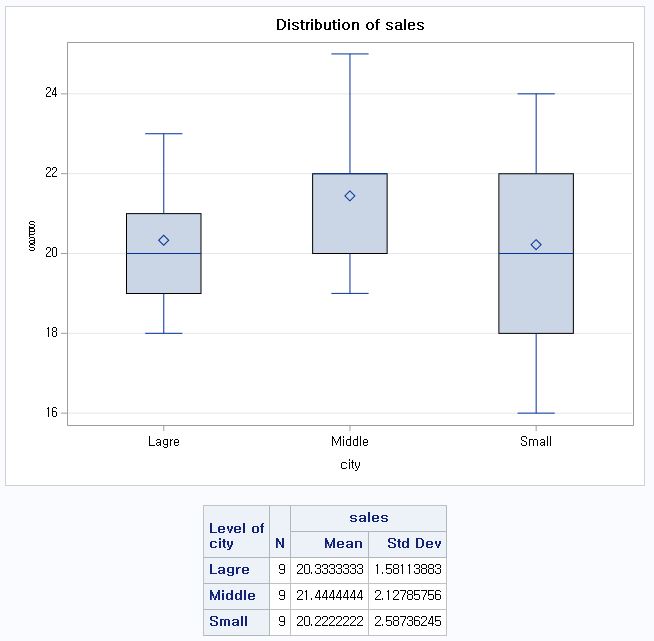
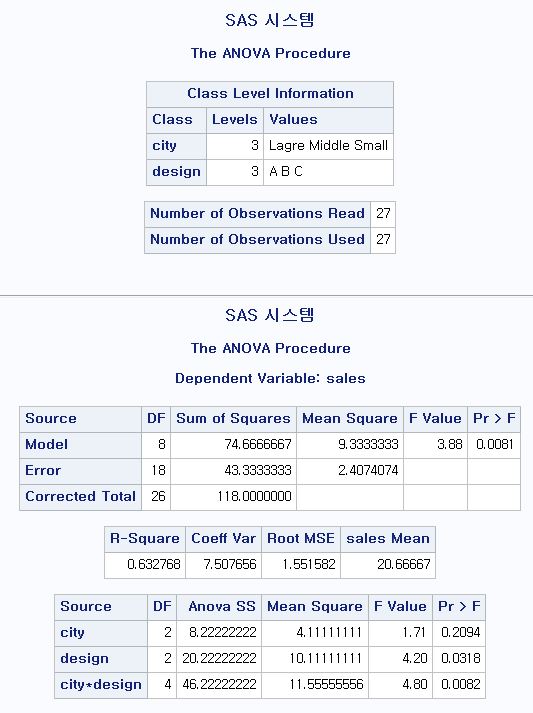
**proc** **anova** data=sasadv.sales;

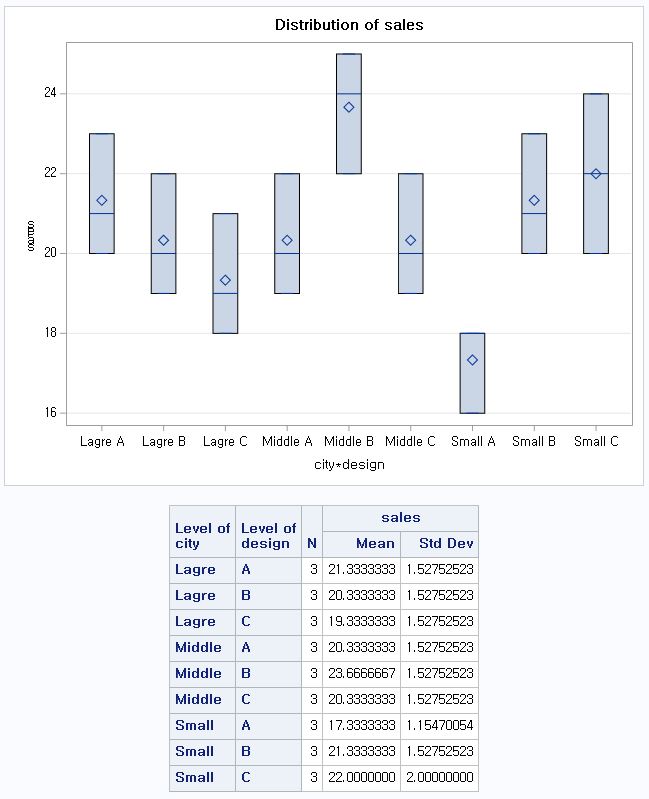
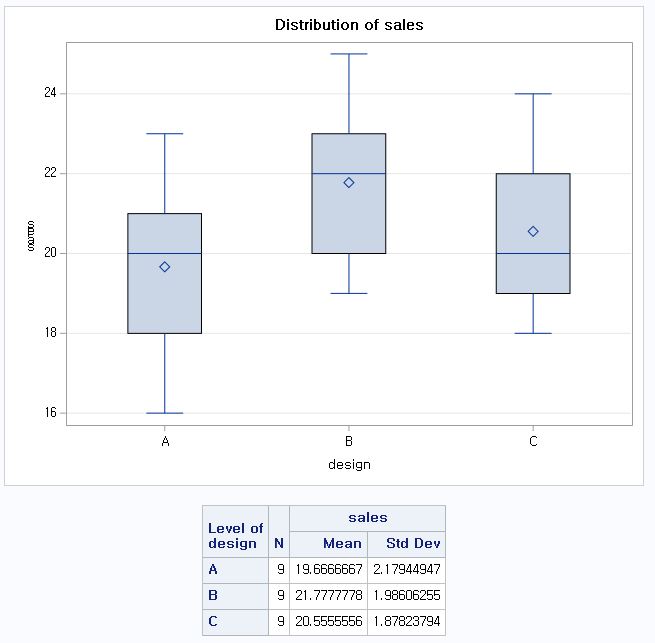
class city design;

model sales=city design city\*design;

means city design city\*design;

**run**;





해석 : 상호작용 ‘city\*design’에 대응되는 유의확률이 0.0082이므로 유의수전 5% 하에서 귀무가설(H0=상호작용 효과가 없다)을 기각한다. 따라서 이 경우 상호작용의 효과가 존재한다고 할 수 있다. 상호작용의 의미를 보다 자세히 살펴보면, 대도시(Large)에서는 디자인 A의 만족도가 가장 높은 반면, 중도시(Middle)에서는 디자인 B가 그리고 소도시(Small)에서는 디자인 C의 판매량이 가장 많다는 것을 알 수 있다. 따라서 도시 구분에 따라서 디자인의 효과가 변화하고 있으며, 두 처리에 의한 상호작용의 효과가 존재한다고 할 수 있다.

**<예 6-4>**

**proc** **summary** data=sasadv.sales nway;

class city design;

var sales;

output out=meanout mean(sales)=mean;

**run**;

symbol1 i=join w=**1** v=dot cv=black h=**2**;

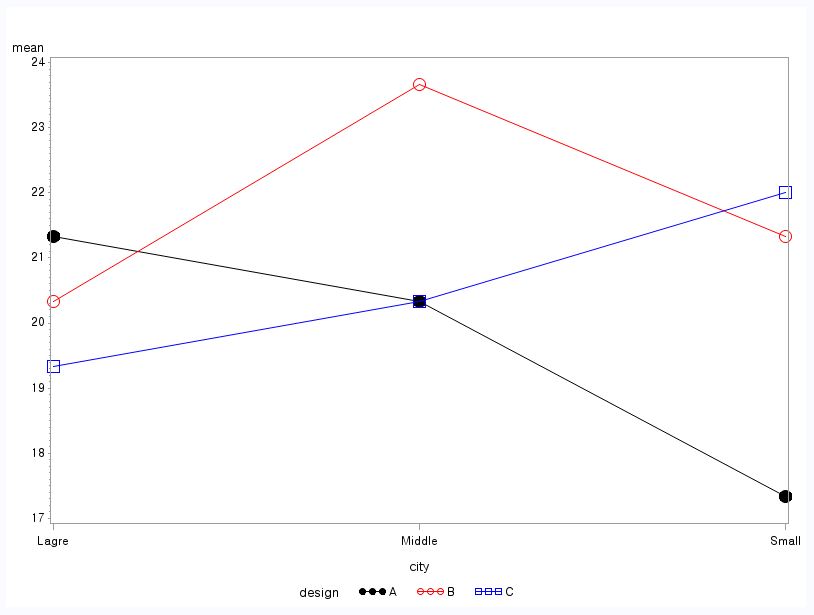
symbol2 i=join w=**1** v=circle cv=red h=**2**;

symbol3 i=join w=**1** v=square cv=blue h=**2**;

**proc** **gplot** data=meanout;

plot mean\*city=design;

**run**;



해석 : 상호작용 효과는 평균 프로파일 도표를 이용하면 보다 쉽게 파악할 수 있다. 위의 평균 프로파일 도표는 <예 6-3>에서 사용된 데이터에 대한 것으로, 도시 구분에 따라 디자인의 효과가 달라진다는 것을 쉽게 살펴볼 수 있다. 만약 상호작용 효과가 존재하지 않는다면, 평균 프로파일 도표가 일정한 간격을 두고 평행선을 이루는 형태를 가지게 될 것이다.

**<예 6-5>**

**data** sasadv.ancova;

input group $ age score @@;

cards;

A 31 30 A 28 0 A 25 10 A 34 40 A 39 55

A 26 20 A 30 65 A 26 5 A 31 40 A 23 0

B 36 65 B 33 50 B 31 90 B 29 25 B 41 99

B 36 60 B 32 25 B 32 80 B 27 5 B 32 99

;

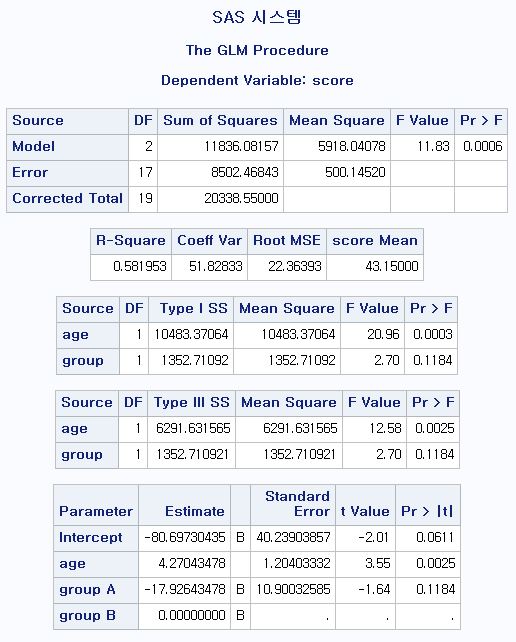
**run**;

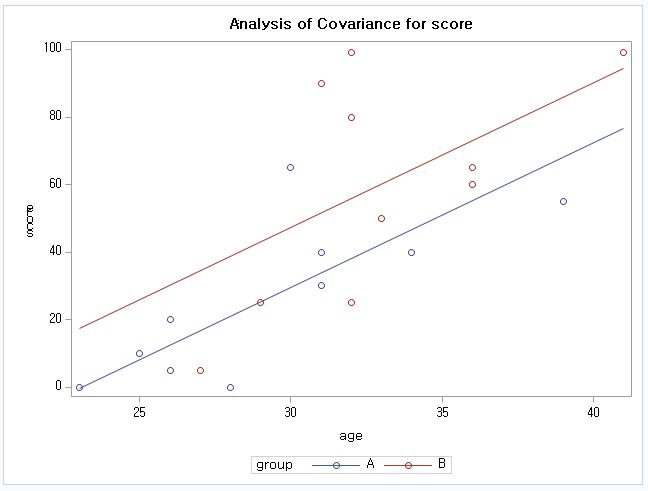
**proc** **glm** data=sasadv.ancova;

class group;

model score=age group / solution;

**run**;





해석 : ‘Type Ⅲ SS’를 보면, 변수 age(나이)의 유의확률이 0.0025이므로 유의수준 5% 하에서 유의하지만 변수 group(집단)의 유의확률은 0.1184로 유의수준 5% 하에서 유의하지 않다. 즉, 나이가 같도록 고정(통제)된다면 두 집단에 대한 심리검사 점수의 평균 차이는 유의하지 않다고 할 수 있다. 또한 ‘각 모수에 대한 최소제곱추정치(Parameter estimate)’를 살펴보면, 공변량 age에 대한 기울기가 4.270이고 group A에 대한 추정치가 -17.926으로 추정되었다. 즉, 나이가 같은 경우 집단에 따른 심리검사 점수의 모평균 차이에 대한 추정값이 -17.926이라는 의미이다.

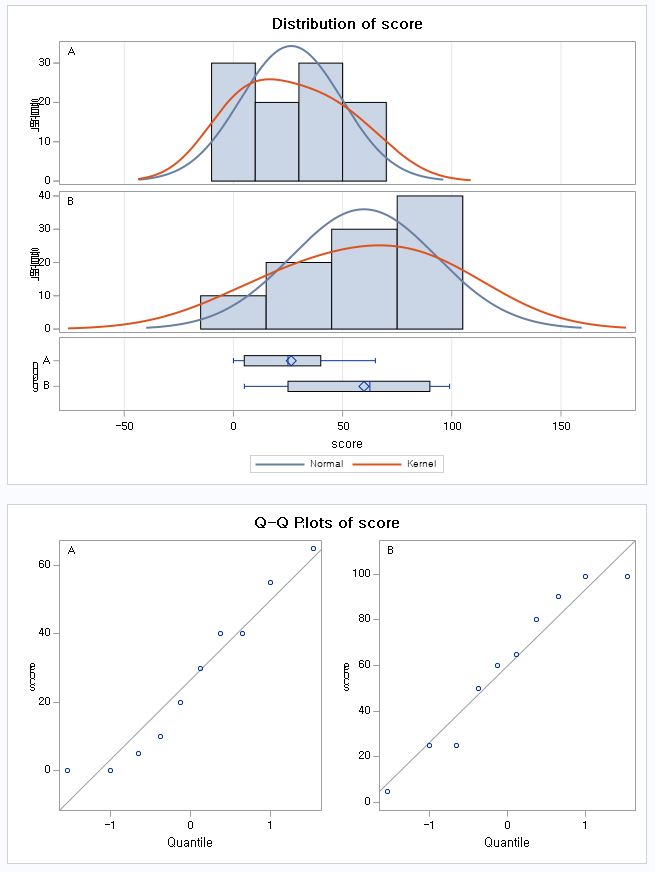
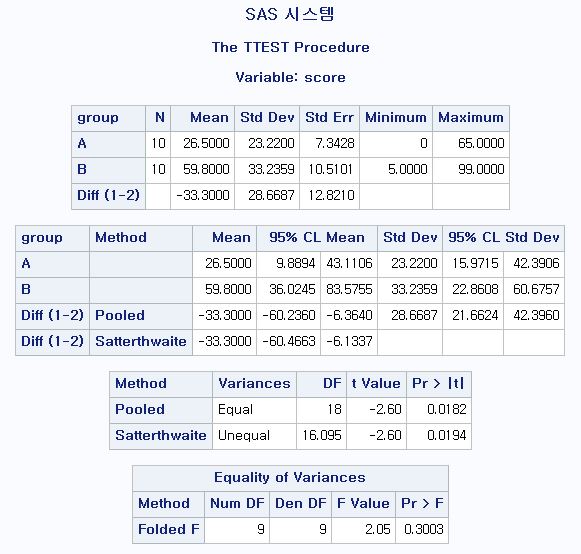
**<예 6-6>**

**proc** **ttest** data=sasadv.ancova;

class group;

var score;

**run**;



해석 : <예 6-6>은 <예 6-5>의 데이터를 이용해서, 변수 group(집단)만을 고려하여 독립표본 t-검정을 수행한 결과이다. 우선 모분산의 동일성 검정 결과, 유의확률이 0.3003이므로 유의수준 5% 하에서 두 집단의 모분산이 동일하다는 귀무가설을 기각하지 못하므로 모분산의 동일성 가정이 만족한다는 것을 알 수 있다. 모분산이 동일하다는 가정 하에서의 t-값에 대한 유의확률이 0.0182로서 유의수준 5% 하에서 유의하다는 것을 알 수 있다. 따라서 공변량인 나이(age)를 고려하지 않고 단순히 t-검정을 수행한다면 잘못된 결론을 유도할 수 있음을 알 수 있다. 또한 모평균의 차이에 대한 추정값이 -33.3으로, 공분산분석에서의 추정값인 -17.926보다 훨씬 크게 추정된다는 것을 알 수 있다.

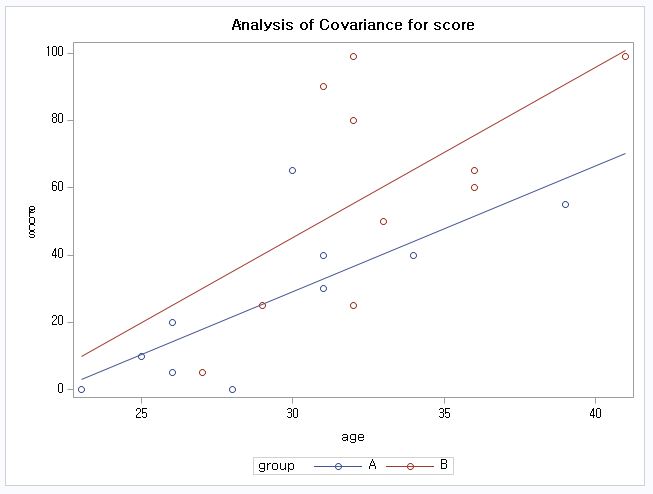
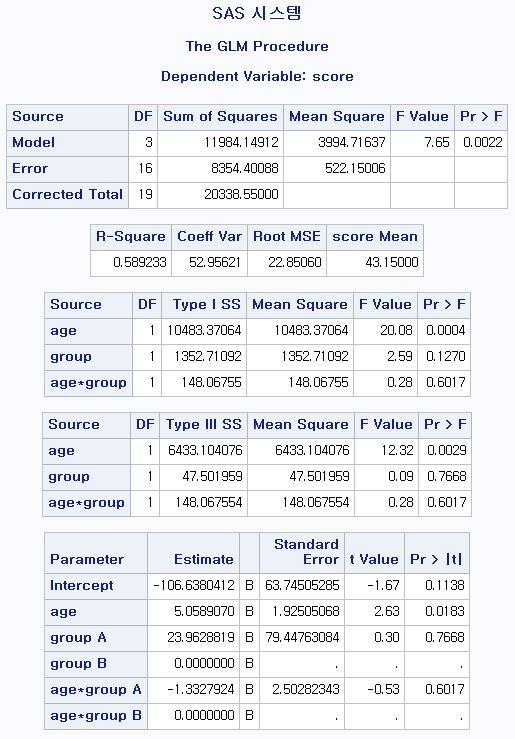
**<예 6-7>**

**proc** **glm** data=sasadv.ancova;

class group;

model score=age group age\*group / solution;

**run**;



해석 : 위 문제에서는 <예 6-5>의 데이터를 사용하였고, 기울기의 동일성에 대한 검정을 수행하기 위해서 MODEL 명령문에 공변량(age)과 요인(group)의 상호작용 효과(age\*group)를 추가해주었다. ‘Type Ⅲ SS’에서 상호작용 효과에 해당하는 ‘age\*group’에 대한 F-검정 결과를 살펴보면, 유의확률이 0.6017이므로 유의수준 5% 하에서 귀무가설(H0=기울기가 동일하다)을 기각하지 못한다. 따라서 이제 기울기의 동일성을 가정할 수 있으므로 상호작용 효과를 모형에서 제거하여 <예 6-5>와 같이 공분산분석을 수행하면 된다.

**\* 6장 연습문제**

**<연습문제 6-1>**

**data** sasadv.ex6\_1;

input cotton $ tensile;

cards;

15% 7

15% 7

15% 15

15% 11

15% 9

20% 12

20% 17

20% 12

20% 18

20% 18

25% 14

25% 18

25% 18

25% 19

25% 19

30% 19

30% 25

30% 22

30% 19

30% 23

35% 7

35% 10

35% 11

35% 15

35% 11

;

**run**;

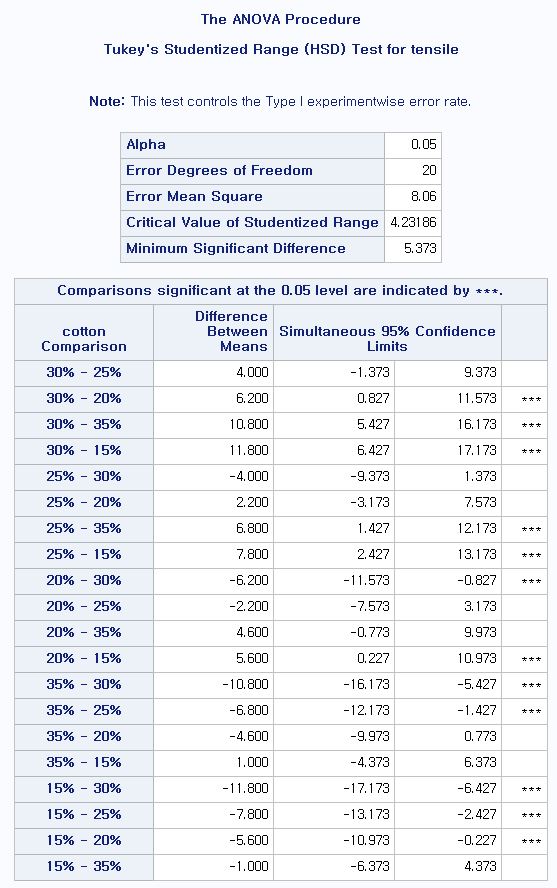
**proc** **anova** data=sasadv.ex6\_1;

class cotton;

model tensile=cotton;

means cotton / tukey cldiff;

**run**;



해석 : 위 문제는 합성섬유의 장력강도가 면(cotton)의 함유량에 따라 어떻게 다른지를 알아보기위한 실험이다. 귀무가설은 “면의 함유량에 따라 평균 장력강도에 차이가 없다”이고, 대립가설은 “면의 함유량에 따라 장력강도에 차이가 있다”이다. 분산분석 결과, F-값이 14.76이고 유의확률이 0.001이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 면의 함유량에 따라 평균 장력강도에 차이가 있다고 할 수 있다. 사후검정으로 다중비교를 해 본 결과, 면(cotton)의 함유량 (30%, 20%), (30%, 35%), (30%, 15%), (25%, 35%), (25%, 15%), (20%, 15%)는 각각 유의한 차이가 존재한다고 나왔다.

**<연습문제 6-3>**

**data** sasadv.ex6\_3;

input concentration $ tensile;

cards;

A1 46.8

A1 58.0

A1 51.4

A1 61.0

A1 45.8

A2 51.2

A2 62.4

A2 30.6

A2 46.0

A2 48.8

A3 50.2

A3 38.2

A3 46.8

A3 26.7

A3 22.7

A4 21.4

A4 22.1

A4 28.2

A4 42.7

A4 25.2

;

**run**;

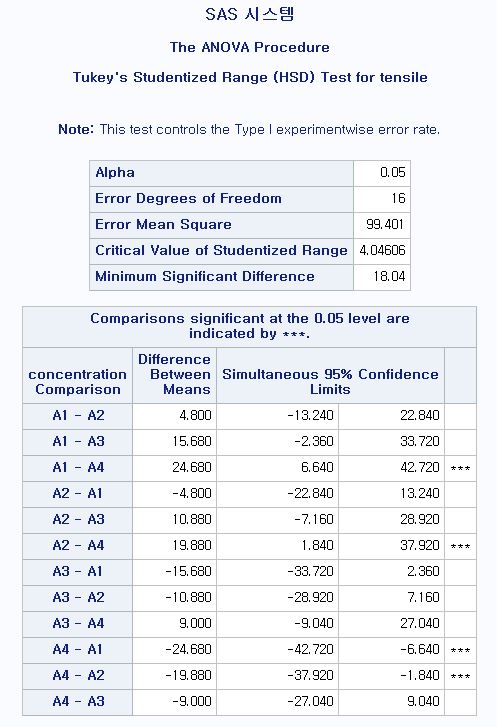
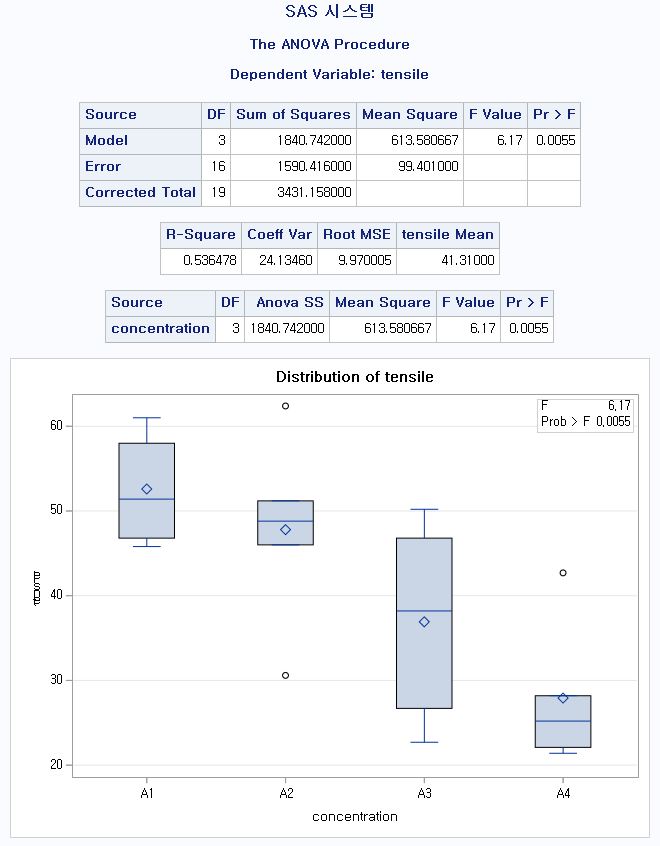
**proc** **anova** data=sasadv.ex6\_3;

class concentration;

model tensile=concentration;

means concentration / tukey cldiff;

**run**;



해석 : 위 문제는 가공시의 처리액의 농도에 따라 직물의 인장강도가 차이가 있는지를 알아보기 위한 실험이다. 귀무가설은 “가공시의 처리액의 농도에 따라 직물의 인장강도가 차이가 없다”이고, 대립가설은 “가공시의 처리액의 농도에 따라 직물의 인장강도가 차이가 있다”이다. 분산분석 결과, F-값이 6.17이고 유의확률이 0.0055이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 가공시의 처리액의 농도에 따라 직물의 인장강도가 차이가 있다고 할 수 있다. 사후검정으로 다중비교를 해 본 결과, 인장강도 (A1, A4), (A2, A4)는 각각 유의한 차이가 존재한다고 나왔다.

**<연습문제 6-5>**

**data** sasadv.ex6\_5;

input board $ insect\_num @@;

cards;

레몬색 45 레몬색 59 레몬색 48 레몬색 46 레몬색 38 레몬색 47

흰색 21 흰색 12 흰색 14 흰색 17 흰색 13 흰색 17

녹색 37 녹색 32 녹색 15 녹색 25 녹색 39 녹색 41

파란색 16 파란색 11 파란색 20 파란색 21 파란색 14 파란색 7

;

**run**;

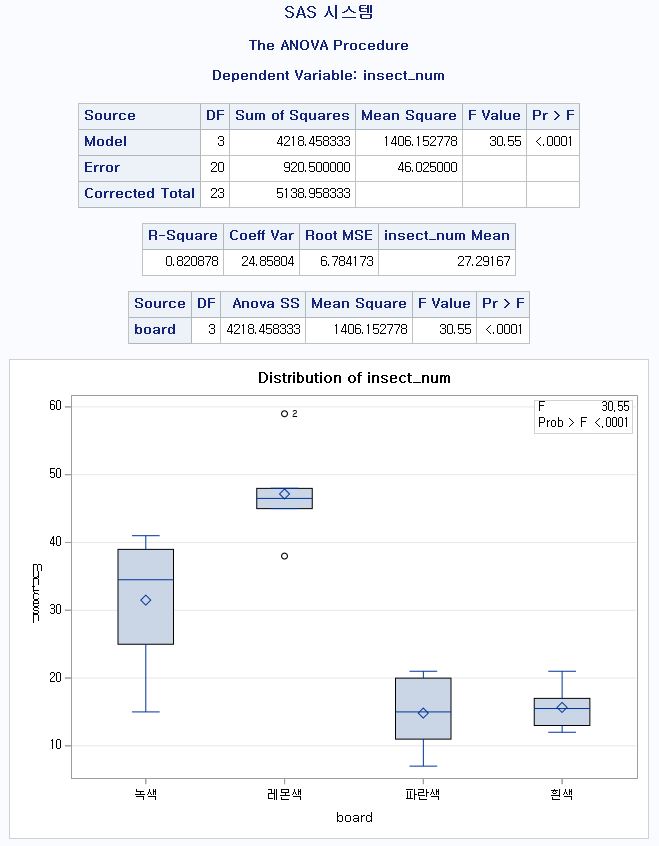
**proc** **anova** data=sasadv.ex6\_5;

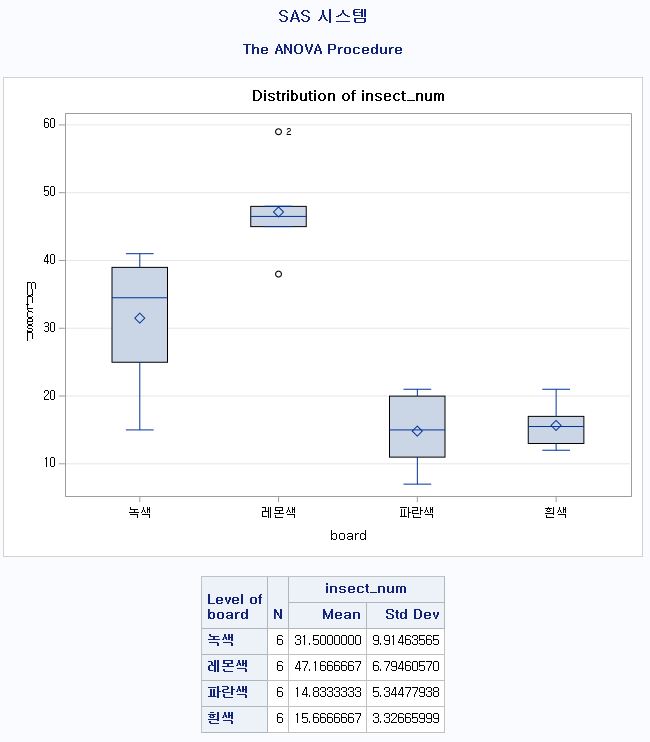
class board;

model insect\_num=board;

means board;

**run**;





해석 : 위 문제는 딱정투구벌레가 색상을 선호하는 정도가 다른지에 대한 연구실험이다. 귀무가설은 “판자의 색상에 따라 잡힌 딱정투구벌레의 수가 차이가 없다”이고, 대립가설을 “판자의 색상에 따라 잡힌 딱정투구벌레의 수가 차이가 있다”이다. 분산분석 결과, F-값이 30.55이고 유의확률이 0.001이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 판자의 색상에 따라 잡힌 딱정투구벌레의 수가 차이가 있다고 할 수 있다. 사후검정으로 다중비교를 해 본 결과, 딱정투구벌레의 색상 선호도는 ‘레몬색>녹색>흰색>파란색’ 순으로 나왔다.

**<연습문제 6-7>**

**data** sasadv.ex6\_7;

input treatment $ obs;

cards;

T1 31

T1 20

T1 18

T1 17

T1 9

T1 8

T1 10

T1 7

T1 120

T1 15

T2 18

T2 17

T2 14

T2 11

T2 10

T2 7

T2 5

T2 6

T2 88

T2 11

;

**run**;

/\* 독립표본 t-test \*/

**proc** **ttest** data=sasadv.ex6\_7;

class treatment;

var obs;

**run**;

/\* 일원배치 분산분석(One-way ANOVA) \*/

**proc** **anova** data=sasadv.ex6\_7;

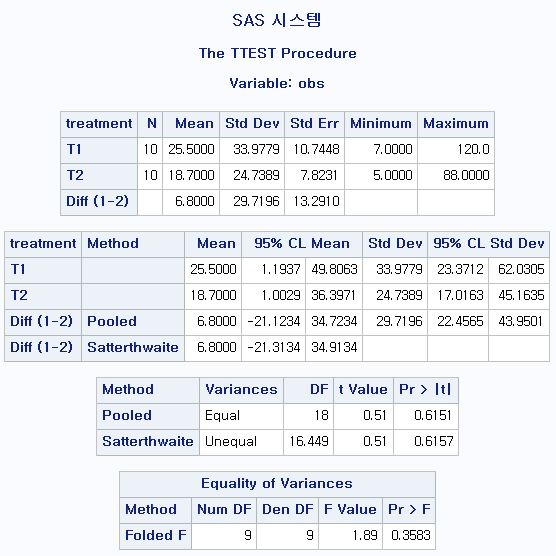
class treatment;

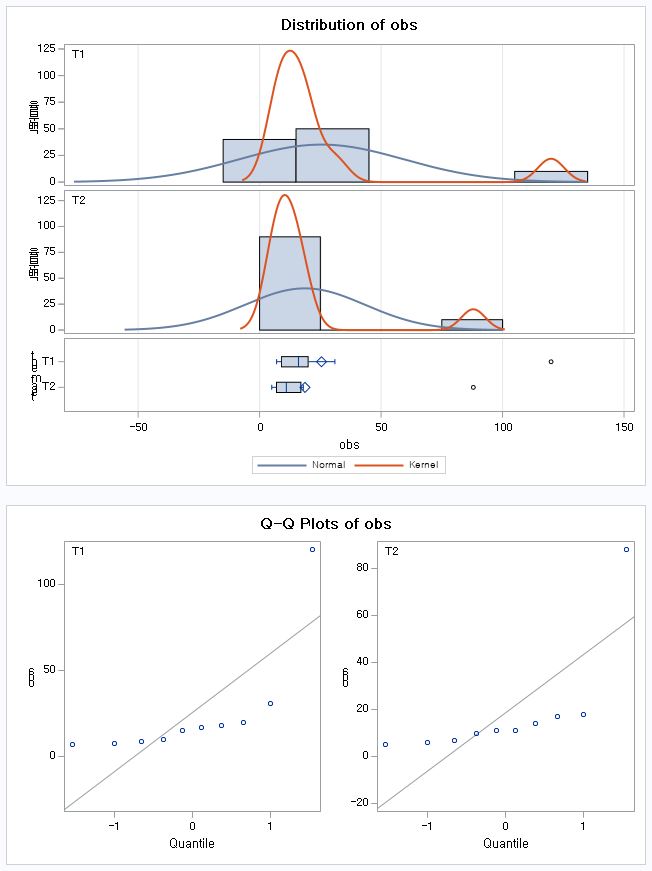
model obs=treatment;

means treatment / cldiff;

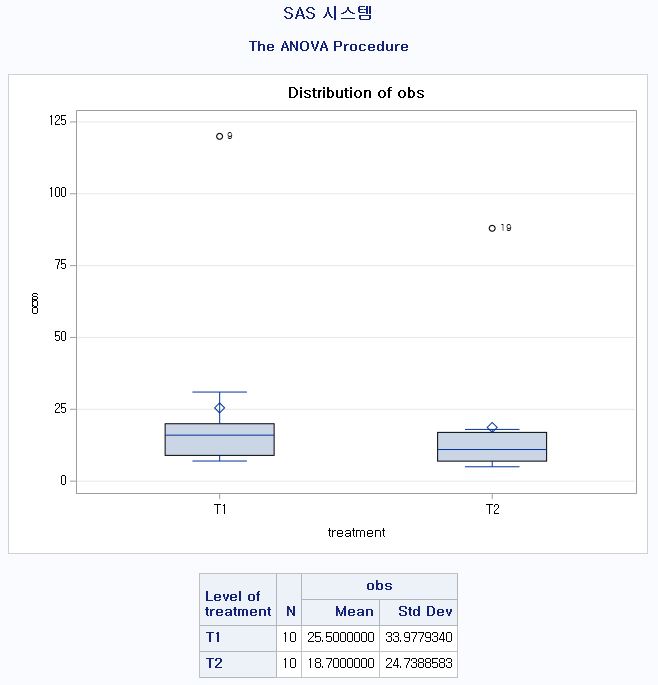
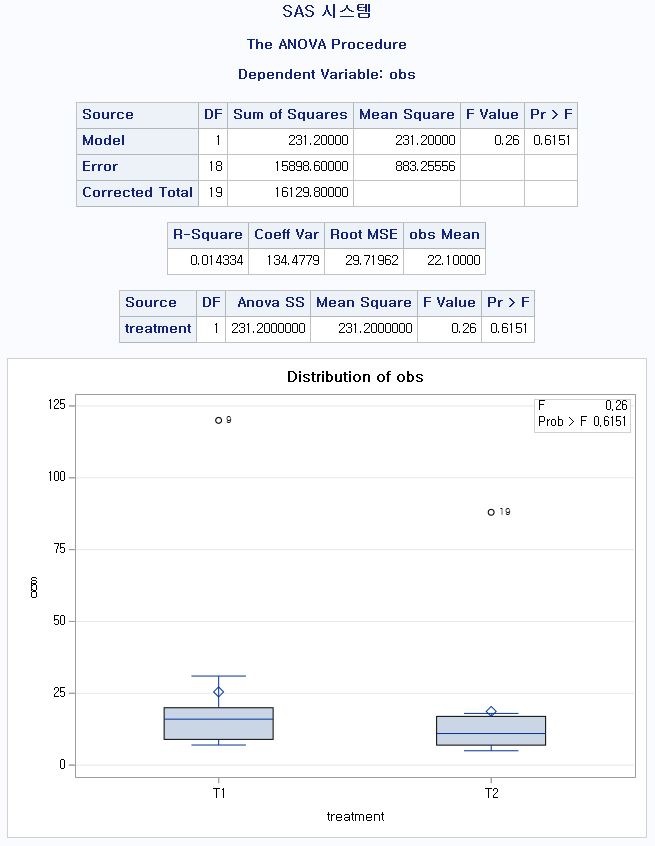
**run**;

**1. t-검정 결과**





**2. 분산분석(ANOVA) 결과**



해석 : 위 문제는 서로 다른 처리(1, 2)를 해준 두 집단의 관측치 모평균이 서로 같은지를 확인하는 문제이다. 귀무가설은 “두 집단의 관측치 모평균이 서로 같다”이고, 대립가설은 “두 집단의 관측치 모평균이 서로 다르다”이다. 먼저 t-검정 결과, ‘모분산의 동일성 검정’에서 유의확률이 0.3583이므로 유의수준 5% 하에서 귀무가설을 기각하지 못하기 때문에 두 집단의 모분산이 동일하다고 할 수 있다. 따라서 모분산의 동일성을 가정한 경우의 t-값에 대한 유의확률을 보면 0.6151로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 그러므로 두 집단의 관측치 모평균이 서로 다르다고 할 수 있다. 다음으로 분산분석(ANOVA)을 한 결과를 살펴보면, F-값에 해당하는 유의확률이 0.6151이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 두 집단의 관측치 모평균이 서로 다르다고 할 수 있다. 결과적으로 두 집단 간 모평균에 차이가 있는지를 검정하는 문제에서는, t-검정과 분산분석(ANOVA)의 결과가 동일하게 나온 것을 확인할 수 있다.

**<연습문제 6-9>**

**data** sasadv.ex6\_9;

do lab='A', 'B', 'C', 'D';

do food='음식 1', '음식 2', '음식 3';

input cholesterol @@;

output;

end;

end;

cards;

3.4 2.6 2.8

3.0 2.7 3.1

3.3 3.0 3.4

3.5 3.1 3.7

;

**run**;

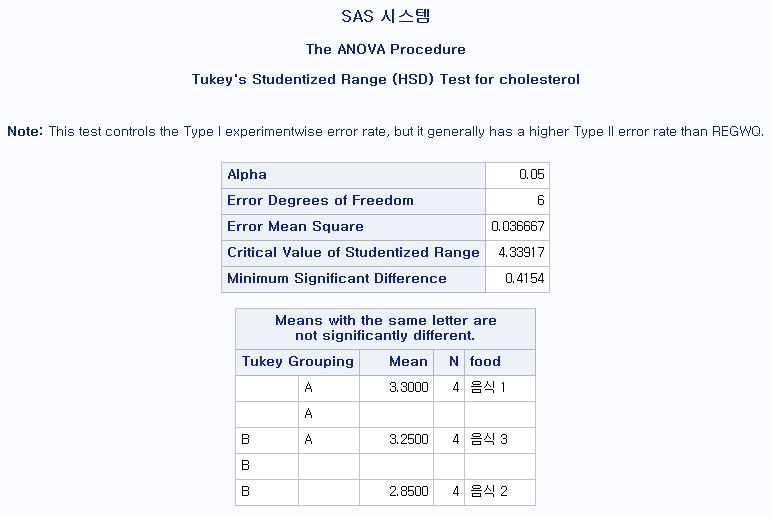
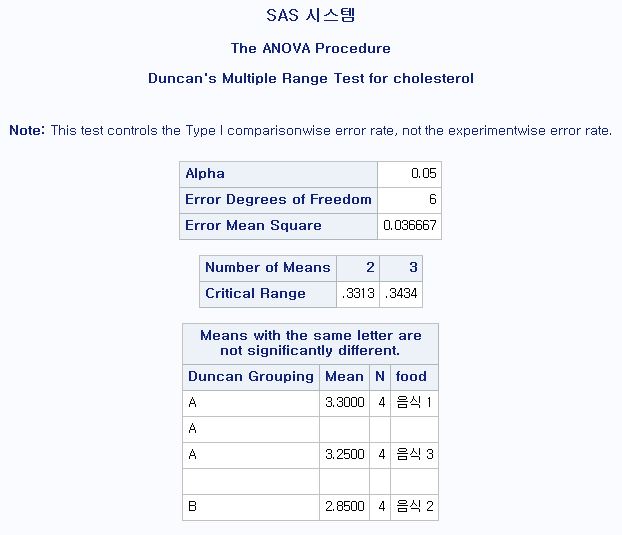
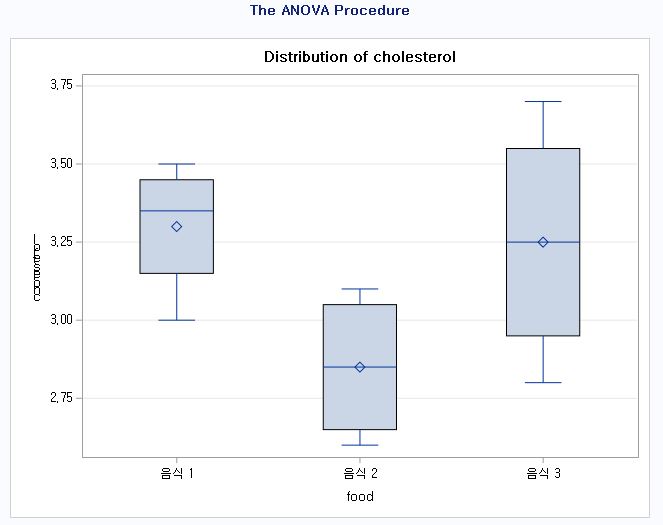
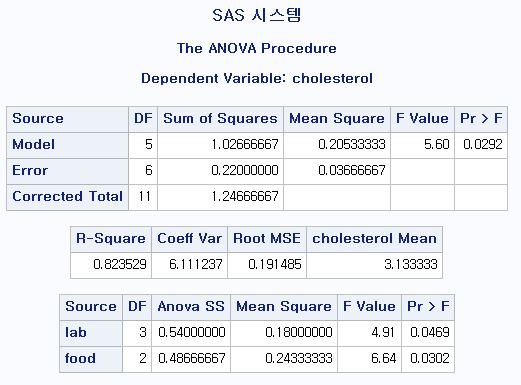
**proc** **anova** data=sasadv.ex6\_9;

class lab food;

model cholesterol=lab food;

means food / duncan tukey alpha=**0.05**;

**run**;



해석 : 위 문제는 다이어트 음식과 실험실에 따라 콜레스테롤 함량에 차이가 있는지를 검정하는 문제이다. 귀무가설은 “다이어트 음식과 실험실에 따라 콜레스테롤 함량에 차이가 없다”이고, 대립가설은 “다이어트 음식과 실험실에 따라 콜레스테롤 함량에 차이가 있다”이다. 분산분석 결과, F-값이 5.60이고 유의확률이 0.0292이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 다이어트 음식과 실험실에 따라 콜레스테롤 함량에 차이가 있다고 할 수 있다. 사후검정으로 다중비교(Duncan, Tukey 검정)를 해 본 결과, ‘음식1’과 ‘음식3’이 ‘음식2’에 비해 평균 콜레스테롤 함유량이 높았다.

**<연습문제 6-11>**

**data** sasadv.ex6\_11;

input element level weight\_plus;

cards;

1 1 73

1 1 102

1 1 118

1 1 104

1 1 81

1 1 107

1 1 100

1 1 87

1 1 117

1 1 111

2 1 98

2 1 74

2 1 56

2 1 111

2 1 95

2 1 88

2 1 82

2 1 77

2 1 86

2 1 92

3 1 94

3 1 79

3 1 96

3 1 98

3 1 102

3 1 102

3 1 108

3 1 91

3 1 120

3 1 105

1 2 90

1 2 76

1 2 90

1 2 64

1 2 86

1 2 51

1 2 72

1 2 90

1 2 95

1 2 78

2 2 107

2 2 95

2 2 97

2 2 80

2 2 98

2 2 74

2 2 74

2 2 67

2 2 89

2 2 58

3 2 49

3 2 82

3 2 73

3 2 86

3 2 81

3 2 97

3 2 106

3 2 70

3 2 61

3 2 82

;

**run**;

**proc** **anova** data=sasadv.ex6\_11;

class element level;

model weight\_plus=element level element\*level;

means element level element\*level;

**run**;

/\* 평균 프로파일 도표(상호작용 효과 확인을 위해) \*/

**proc** **summary** data=sasadv.ex6\_11 nway;

class element level;

var weight\_plus;

output out=meanout6\_11 mean(weight\_plus)=mean;

**run**;

symbol1 i=join w=**1** v=dot cv=black h=**2**;

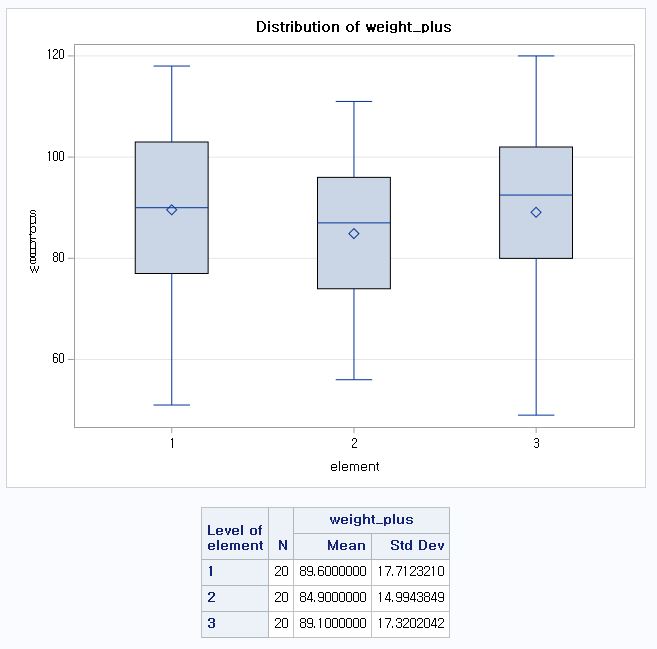
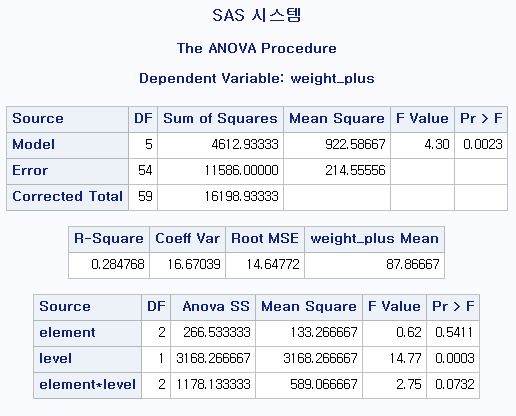
symbol2 i=join w=**1** v=circle cv=red h=**2**;

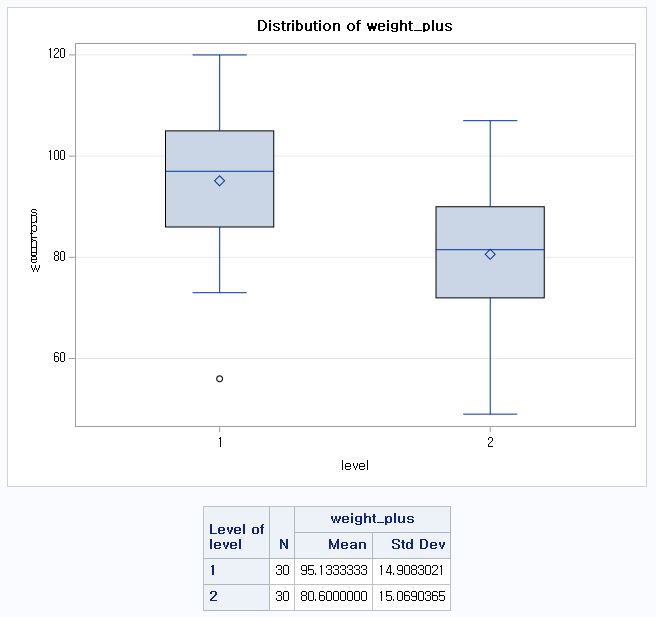
symbol3 i=join w=**1** v=square cv=blue h=**2**;

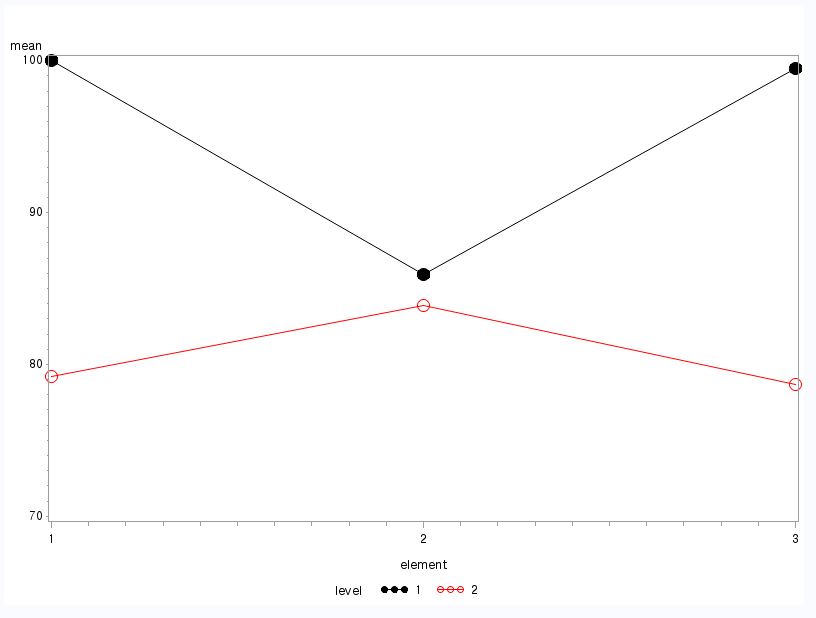
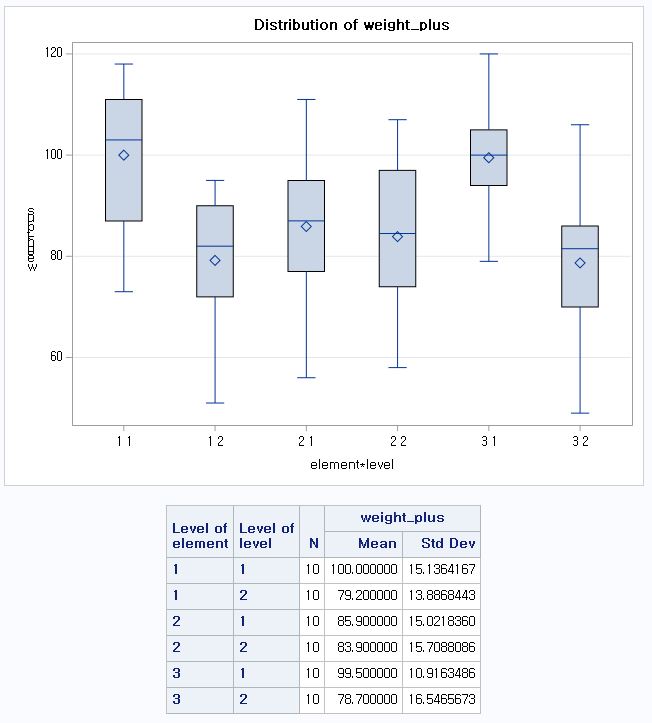
**proc** **gplot** data=meanout6\_11;

plot mean\*element=level;

**run**;







해석 : 위 문제는 여섯 종류의 사료를 쥐에게 먹였을 때 쥐의 체중증가에 어떤 영향을 주는지 알아보기 위해 단백질의 요소에 대한 세 수준(A:쇠고기, 곡류, 돼지고기)과 단백질의 수준에 대한 두 수준(B:높음, 낮음)을 고려한 후 실험을 한 것이다. 요인 A, B 및 상호작용의 효과가 있는지를 분석하기 위해, 다음과 같이 총 3개의 가설을 설정하였다.

**1. 귀무가설 : 단백질 요소(요인 A)에 따른 쥐의 체중증가량이 차이가 없다.**

**대립가설 : 단백질 요소(요인 A)에 따른 쥐의 체중증가량이 차이가 있다.**

**2. 귀무가설 : 단백질 수준(요인 B)에 따른 쥐의 체중증가량이 차이가 없다.**

**대립가설 : 단백질 수준(요인 B)에 따른 쥐의 체중증가량이 차이가 있다.**

**3. 귀무가설 : 요인 A와 B간에 상호작용 효과가 존재하지 않는다.**

**대립가설 : 요인 A와 B간에 상호작용 효과가 존재한다.**

먼저 첫 번째 가설 검정 결과를 살펴보면, F-값이 0.62이고 유의확률이 0.5411로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 단백질 요소(요인 A)에 따른 쥐의 체중증가량이 차이가 없다고 할 수 있다. 두 번째 가설 검정 결과를 살펴보면, F-값이 14.77이고 유의확률이 0.0003이므로 유의수준 5% 하에서 귀무가설을 기각한다. 따라서 단백질 수준(요인 B)에 따른 쥐의 체중증가량이 차이가 있다고 할 수 있다. 마지막으로 상호작용 효과 존재 여부에 대한 검정 결과를 살펴보면, F-값이 2.75이고 유의확률이 0.0732이므로 유의수준 5% 하에서 귀무가설을 기각하지 못한다. 따라서 유의수준 5% 하에서는 요인 A와 B간에 상호작용 효과가 존재하지 않는다고 할 수 있다. 하지만 평균 프로파일 도표를 보면 선들이 서로 평행선을 이루지 않으므로 상호작용 효과가 존재하지 않는다고 보기는 어려워 보인다. 따라서 유의수준을 10%로 늘려서 귀무가설을 기각하게끔 하여 결과를 해석하는 것이 적절해 보인다.