**SAS 데이터 분석 입문 2장**

**2019020650 김형욱**

**\* 2장 연습문제**

**<연습문제 2-3>**

**data** sasadv.king;

input name $ age @@;

if \_N\_<=**14** then war='Before';

else war='After';

agegroup=INT(age/**10**)\***10**;

cards;

태조 73 정종 62 태종 45 세종 53 문종 38

단종 16 세조 51 예종 28 성종 37 연산군 30

중종 56 인종 30 명종 33 선조 56 광해군 66

인조 54 효종 40 현종 33 숙종 59 경종 36

영조 82 정조 48 순조 44 헌종 22 철종 32

고종 67 순종 52

;

**run**;

/\* (가) \*/

**proc** **sort** data=sasadv.king;

by war;

**run**;

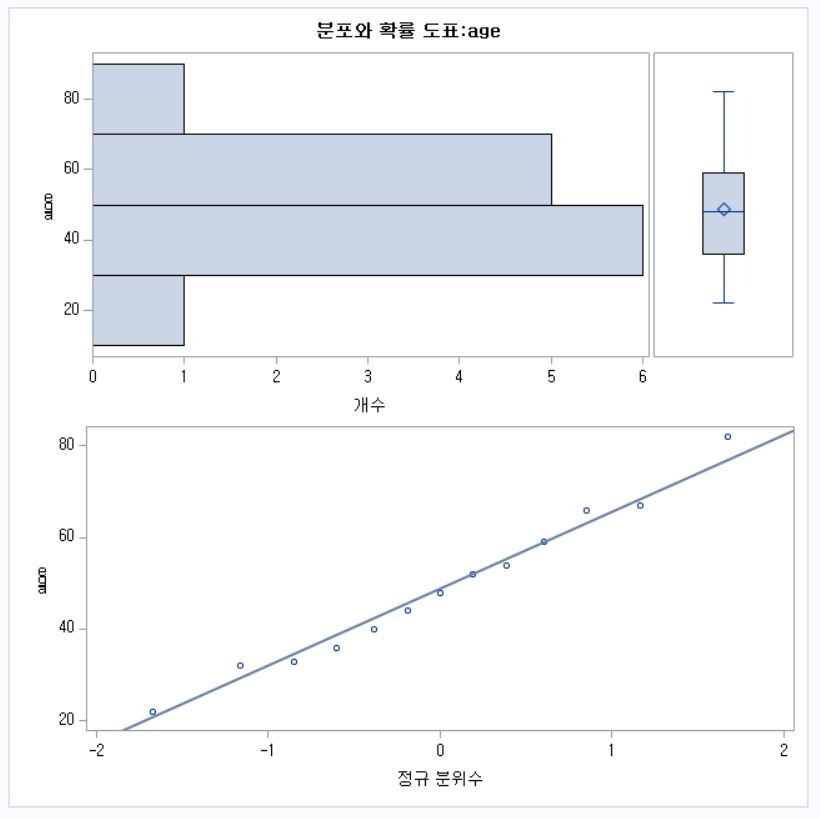
**proc** **univariate** data=sasadv.king plot;

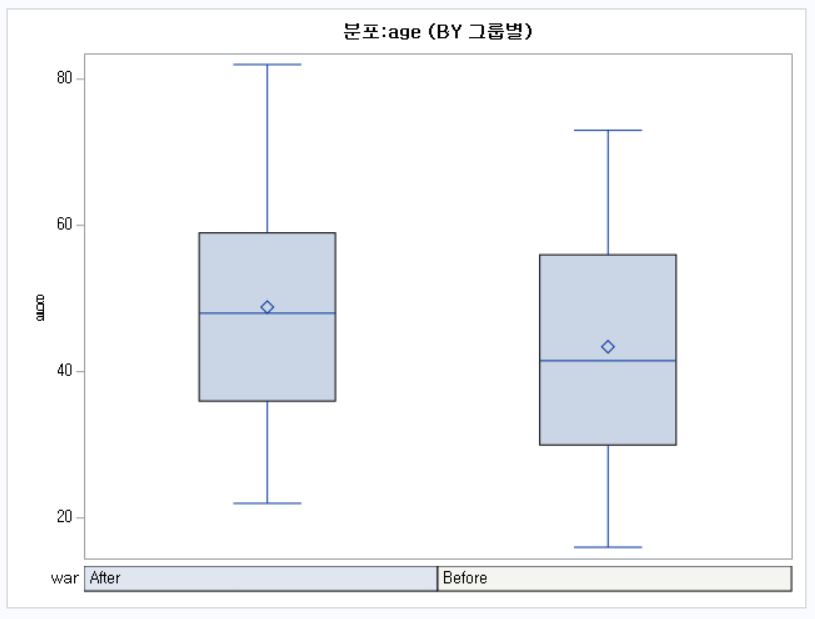
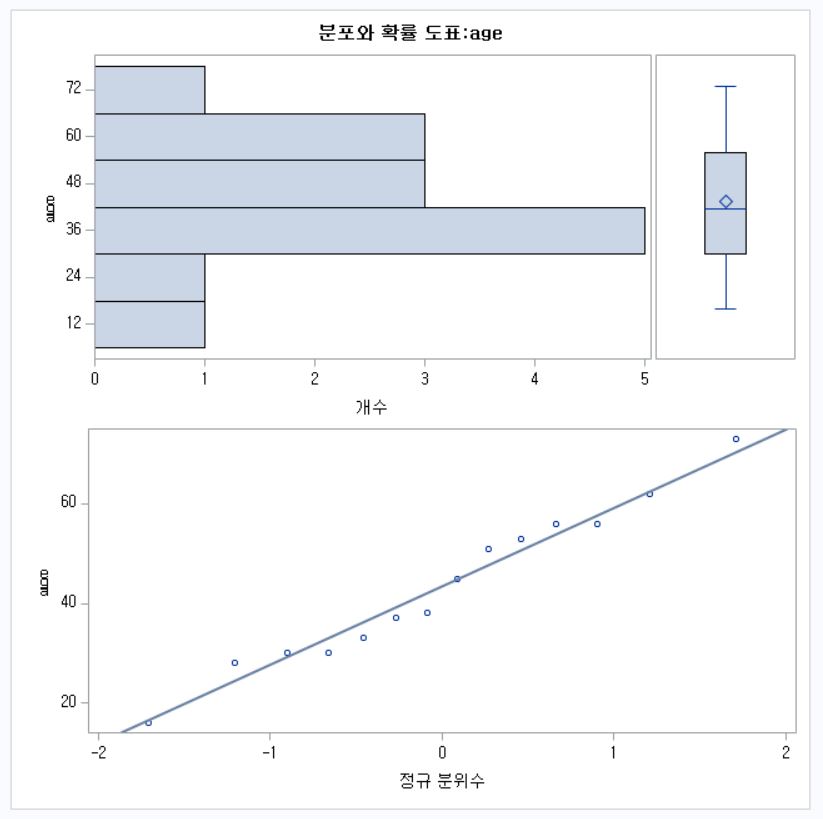
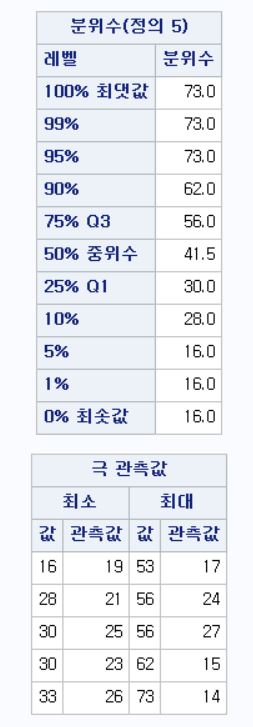
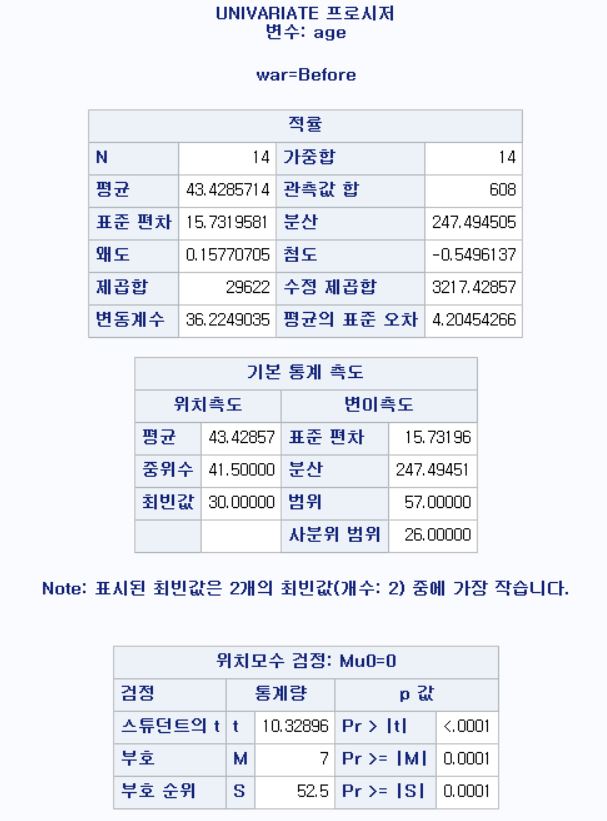
by war;

var age;

**run**;







해석 : 임진왜란 전과 후 각각에 대하여 변수 age(나이)의 기술통계량을 확인해보니, 임진왜란 후에 왕들의 평균 수명이 늘어난 것을 확인할 수 있다. 또한 줄기-잎 그림과 상자그림을 통해서도 임진왜란 후에 왕들의 평균 수명이 늘어났다는 것을 눈으로 확인할 수 있다.

/\* (나) \*/

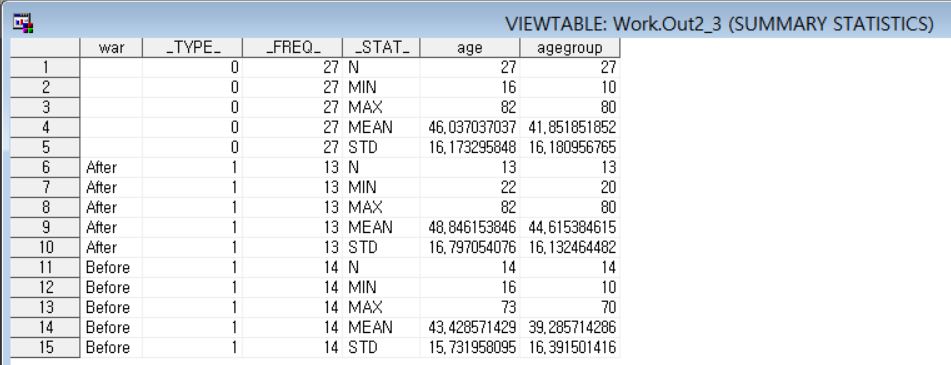
**proc** **means** data=sasadv.king

max min median mean std;

class war;

output out=out2\_3;

**run**;

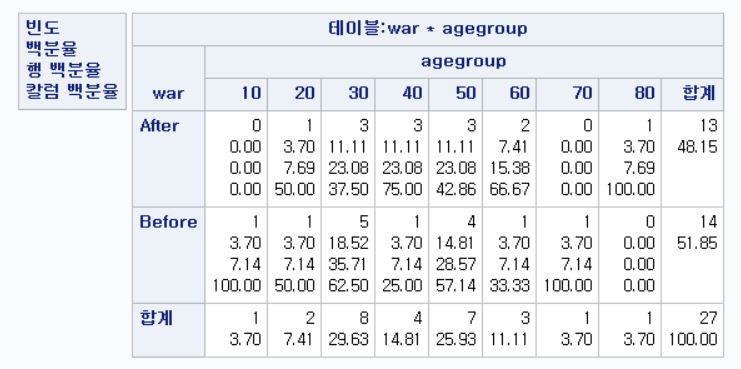
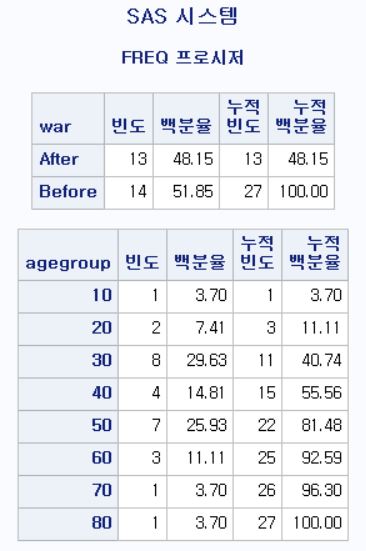


/\* (다) \*/

**proc** **freq** data=sasadv.king;

tables war agegroup war\*agegroup;

**run**;



**<연습문제 2-4>**

**data** sasadv.ex2\_4;

input size $ **1**-**13** manufact $ **14**-**26** model $ **27**-**42** mileage **43**-**54** reliable **55**-**64** index;

cards;

Small Chevrolet Geo Prizm 33 5 4

Small Honda Civic 29 5 4

Small Toyota Corolla 30 5 4

Small Ford Escort 27 3 3

Small Dodge Colt 34 . .

Compact Ford Tempo 24 1 3

Compact Chrysler Le Baron 23 3 3

Compact Buick Skylark 21 3 3

Compact Plymouth Acclaim 24 3 3

Compact Chevrolet Corsica 25 2 3

Compact Pontiac Sunbird 24 1 3

Mid-Sized Toyota Camry 24 5 4

Mid-Sized Honda Accord 26 5 4

Mid-Sized Ford Taurus 20 3 3

;

**run**;

/\* (가) \*/

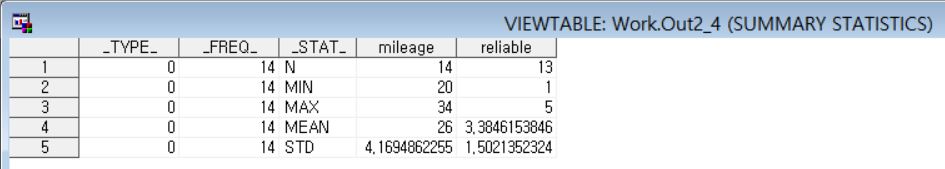
**proc** **means** data=sasadv.ex2\_4

mean std maxdec=**3**;

var mileage reliable;

output out=out2\_4;

**run**;



/\* (나) \*/

**proc** **freq** data=sasadv.ex2\_4;

tables size index size\*index;

**run**;



**<연습문제 2-6>**

**data** sasadv.ex2\_6;

input cholest @@;

cards;

239 161 210 179 212 195 301 357 233 256

234 195 199 284 245 174 310 286 176 212

297 249 282 233 205 286 269 305 247 292

;

**run**;

/\* (가) \*/

**proc** **freq** data=sasadv.ex2\_6;

tables cholest;

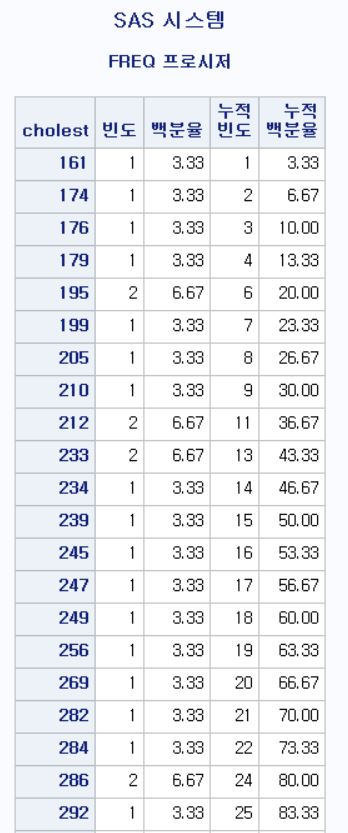
**run**;

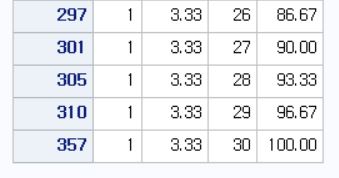
**proc** **univariate** data=sasadv.ex2\_6;

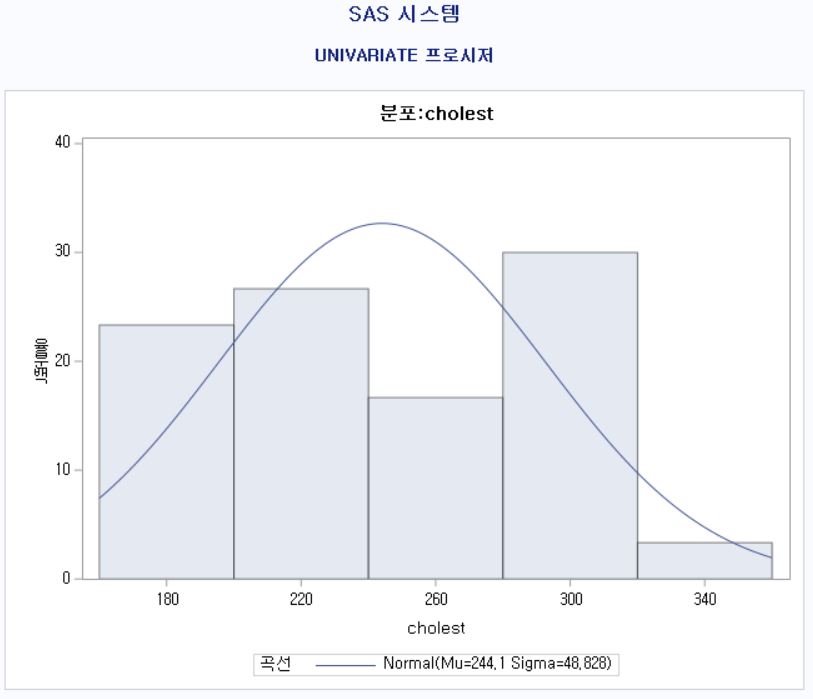
var cholest;

histogram cholest / normal;

**run**;







/\* (나) \*/

**proc** **means** data=sasadv.ex2\_6

mean median var std;

var cholest;

**run**;



**<연습문제 2-7>**

**data** sasadv.ex2\_7;

input before after;

cards;

72 74

70 72

68 69

67 68

73 72

71 72

72 72

70 71

69 67

70 73

68 69

72 71

69 68

66 74

73 73

71 70

70 74

72 68

70 71

69 74

72 74

73 69

;

**run**;

/\* (가) \*/

**proc** **freq** data=sasadv.ex2\_7;

tables before after;

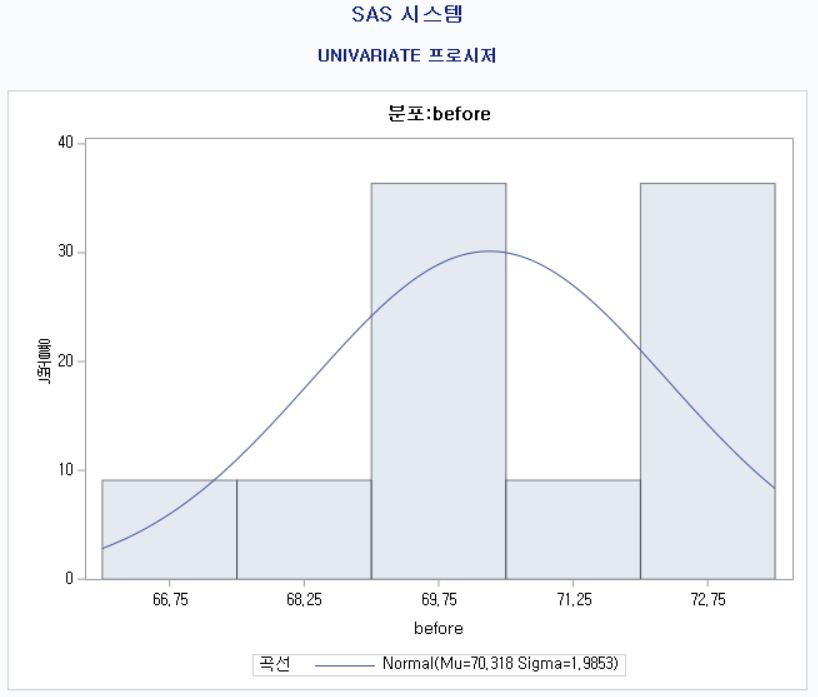
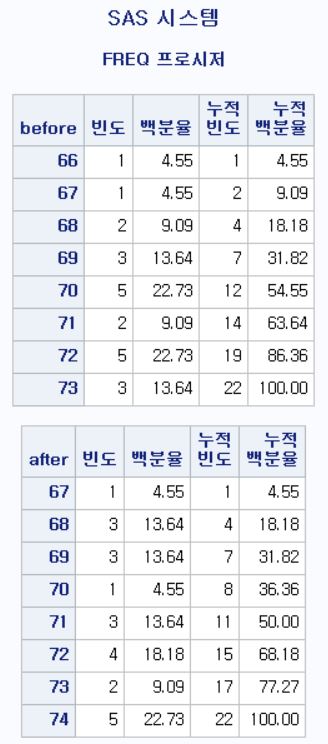
**run**;

**proc** **univariate** data=sasadv.ex2\_7;

var before after;

histogram before after / normal;

**run**;





/\* (나) \*/

**proc** **means** data=sasadv.ex2\_7

mean median std;

var before after;

**run**;

해석 : 담배를 피운 후의 분당 평균 맥박수가 피우기 전보다 더 높게나온 것을 확인할 수 있다.

**<연습문제 2-9>**

**data** sasadv.ex2\_9;

input region $ carbon;

cards;

A 95

A 96

A 92

A 102

A 103

A 93

A 101

A 92

A 95

A 90

B 184

B 202

B 215

B 204

B 195

B 201

B 169

B 182

B 192

C 215

C 214

C 197

C 216

C 215

C 208

C 228

C 208

C 216

C 214

C 227

D 155

D 142

D 146

D 149

D 146

D 152

D 159

;

**run**;

/\* (가) \*/

**proc** **freq** data=sasadv.ex2\_9;

tables carbon;

**run**;

**proc** **sort** data=sasadv.ex2\_9;

by region;

**run**;

**proc** **univariate** data=sasadv.ex2\_9;

by region;

var carbon;

histogram carbon / normal;

**run**;

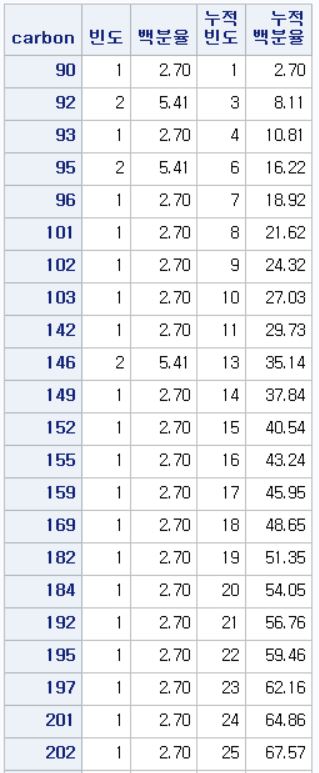
**proc** **means** data=sasadv.ex2\_9

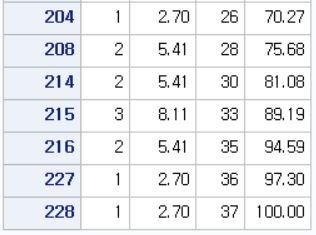
mean median std;

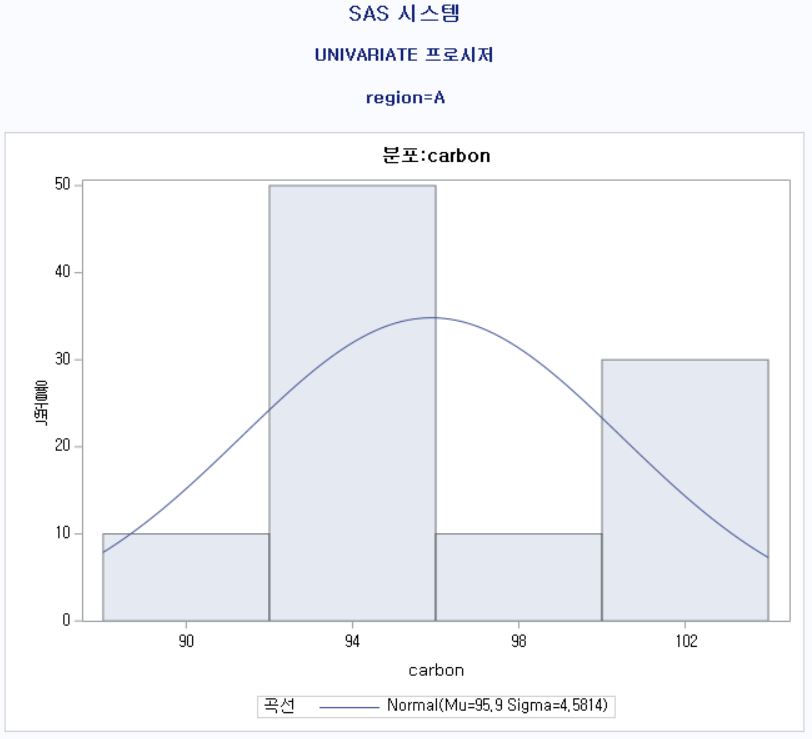
class region;

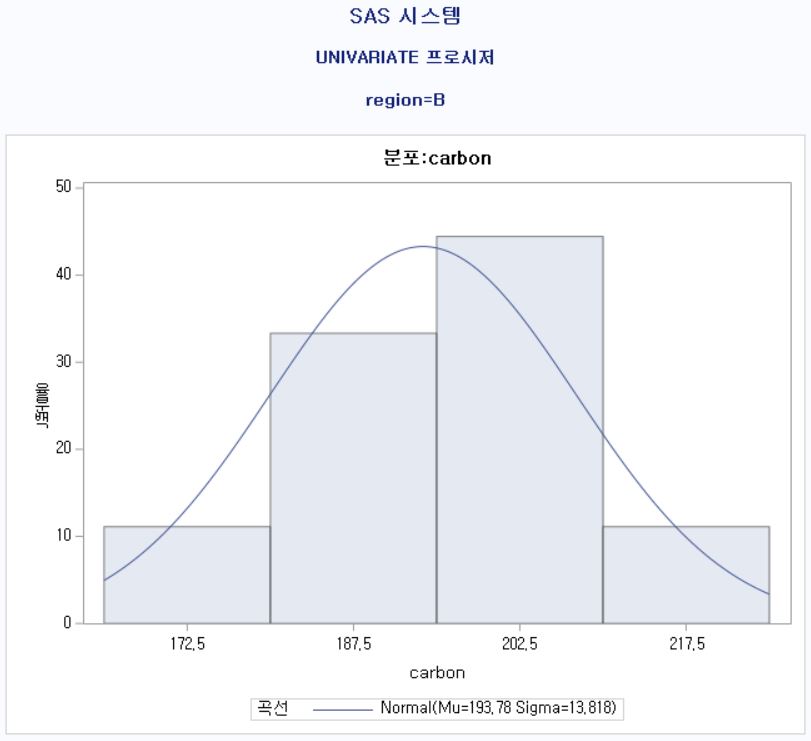
var carbon;

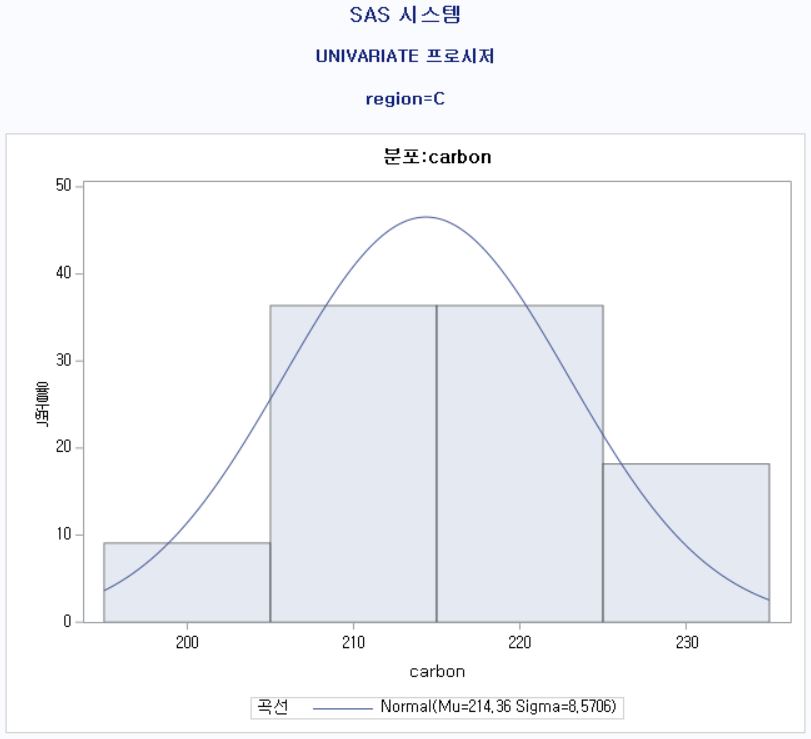
**run**;

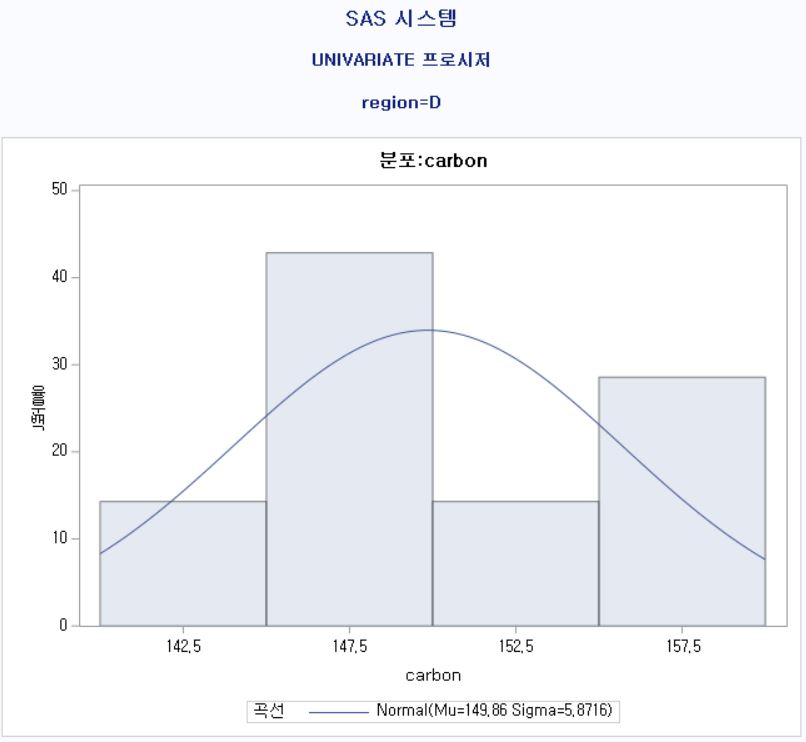














/\* (나) \*/

해석 : (가)의 결과를 보면, 지역 A가 평균 일산화탄소량이 가장 적고 지역 C의 평균 일산화탄소량이 다른 지역들보다 월등히 높게 나온 것을 확인할 수 있다. 결과적으로 지역별 평균 일산화탄소량을 크기순으로 나열해보면, “A<D<B<C”이다.

**\* 2장 예제문제**

**<예 2-1>**

**data** sasadv.cholest;

input gender $ age super @@;

cards;

M 23 40 M 64 88 M 66 110 M 31 86 M 55 137

M 48 78 M 58 111 M 31 88 M 27 80 M 25 86

M 20 80 M 32 47 M 63 106 M 23 65 M 62 74

M 43 66 M 43 79 M 36 58 M 67 123 M 27 87

M 29 88 M 48 90 M 63 56 M 27 73 M 19 112

M 59 110 M 65 118 M 26 52 M 53 106 M 42 67

M 60 57 F 30 66 F 25 69 F 40 65 F 38 52

F 57 84 F 33 86 F 23 35 F 42 116 F 49 76

F 35 55 F 49 73 F 44 89 F 50 127 F 60 87

F 63 142 F 47 77 F 23 76 F 27 58 F 36 91

F 48 107 F 23 98 F 74 128 F 44 84 F 56 146

F 53 75 F 37 120 F 41 80 F 41 82 F 57 123

;

**run**;

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

by gender;

**run**;

options ps=**55** ls=**65** nodate pageno=**1**;

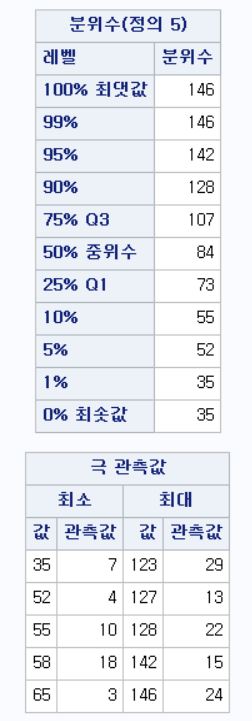
**proc** **univariate** data=sasadv.cholest;

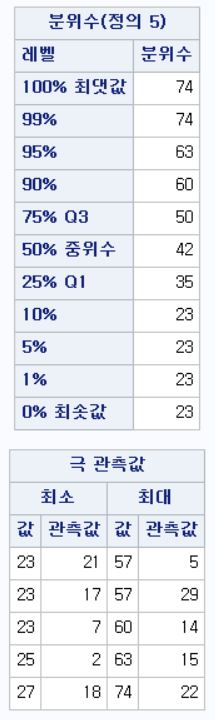
by gender;

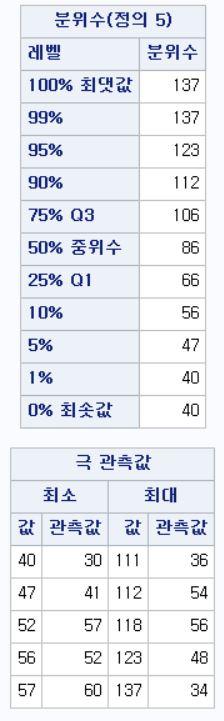
var super age;

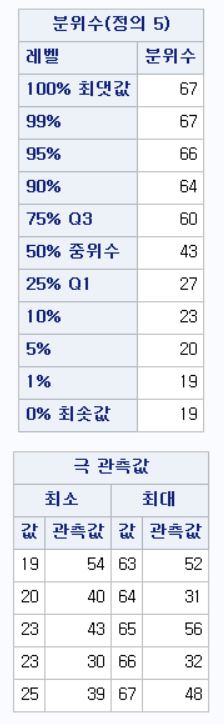
label super='콜레스테롤 과포화율' age='나이';

**run**;









해석 : 기술통계량을 살펴보면, 여자가 남자보다 평균 콜레스테롤 과포화율(super)이 조금 더 높은 것을 알 수 있다.

**<예 2-2>**

**proc** **rank** data=sasadv.cholest out=rankout ties=low;

where gender='F';

var super;

ranks r\_super;

**run**;

**proc** **sort** data=rankout;

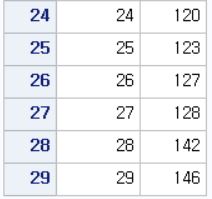
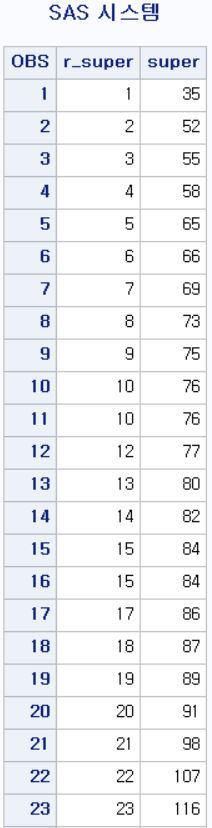
by r\_super;

**run**;

**proc** **print** data=rankout;

var r\_super super;

**run**;



* RANK 프로시저를 사용하면 하나 또는 그 이상의 숫자변수에 대한 순위(rank)를 얻거나, 원래변수의 값들 대신 이들 순위를 포함하는 새로운 SAS 데이터셋을 생성할 수 있다.

**<예 2-3>**

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

by gender;

var super age;

output out=univ\_out

mean=s\_mean std=s\_std a\_std

pctlpts=**33.3** **66.6** pctlpre=s\_p a\_p;

label super='콜레스테롤 과포화율' age='나이';

**run**;

**proc** **print** data=univ\_out;

**run**;



**<예 2-4>**

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

class gender;

histogram age / outhistogram=outhisto

cfill=orange

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

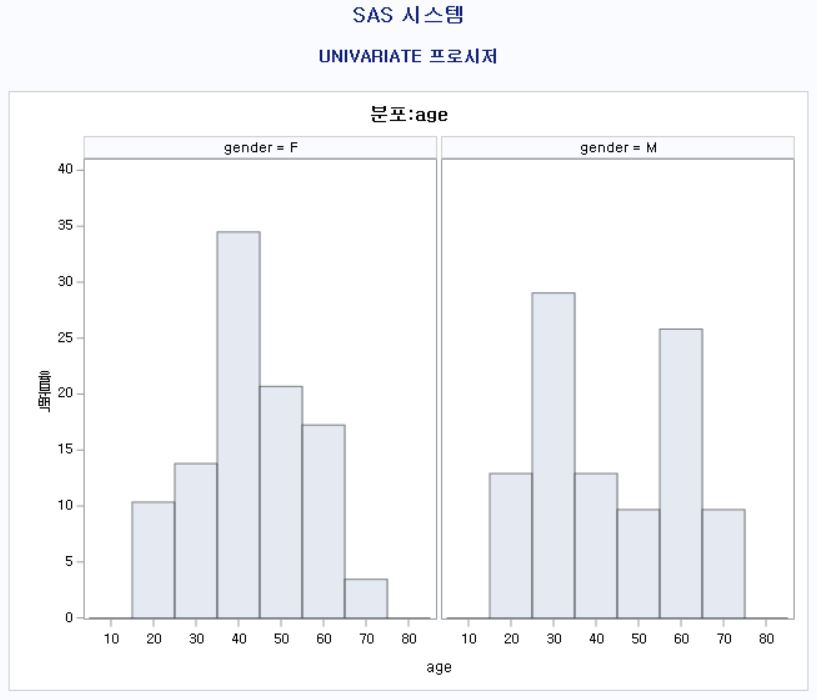
midpoints=**10** to **80** by **10**

nrow=**1** ncol=**2**;

**run**;

**proc** **print** data=outhisto label;

**run**;





해석 : 위 결과로부터 여성의 연령대는 비교적 중간(40~50대)에 많이 분포하고 남성의 연령대는 양쪽 끝(20대, 60대)에 많이 분포한다는 것을 알 수 있다.

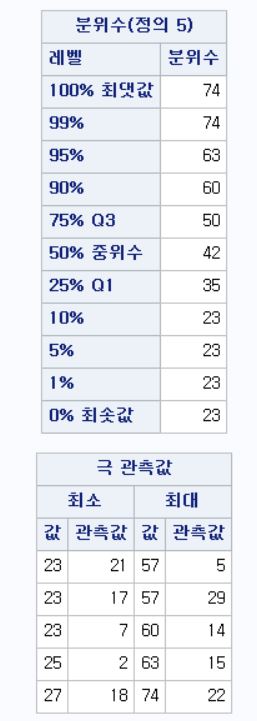
**<예 2-5>**

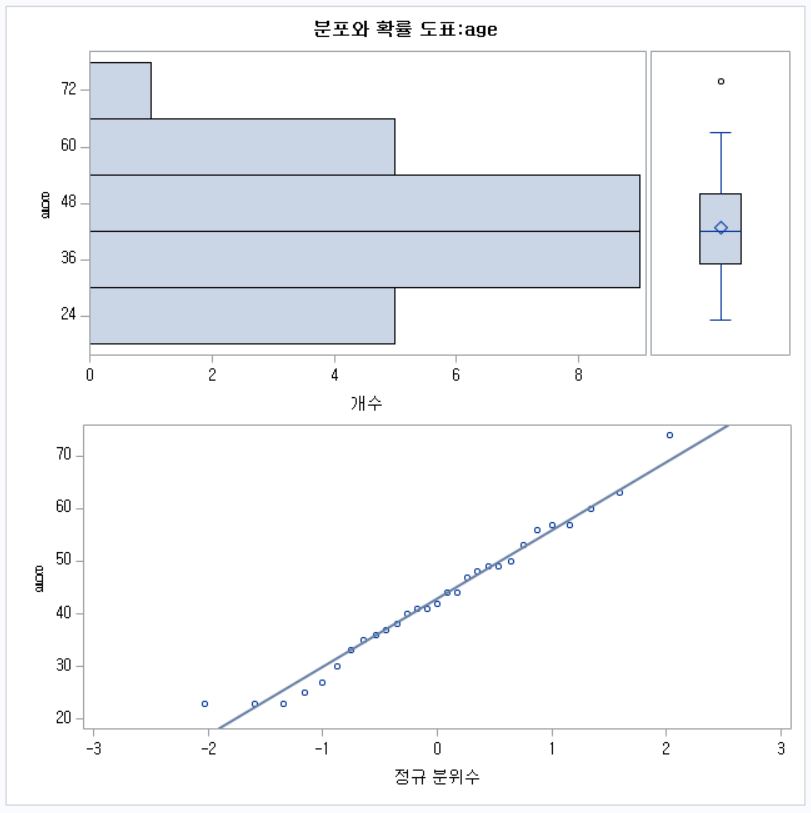
**proc** **univariate** data=sasadv.cholest plot;

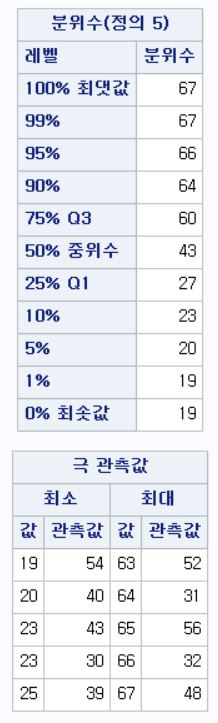
by gender;

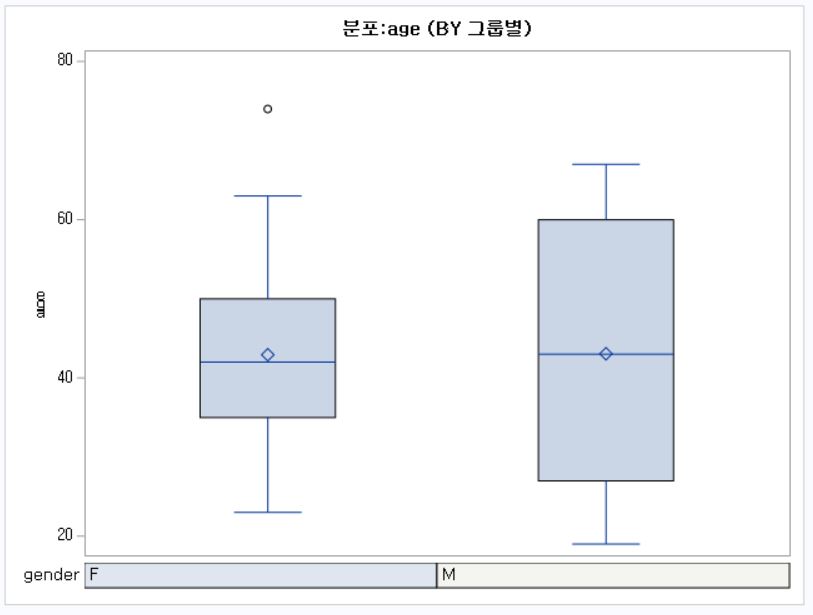
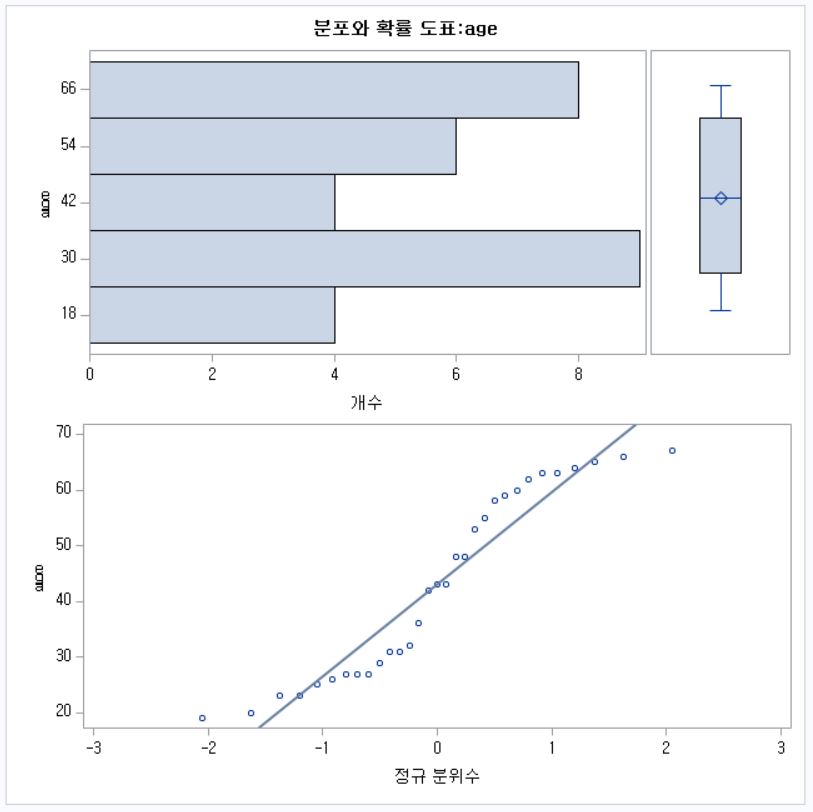
var age;

**run**;









**<예 2-6>**

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

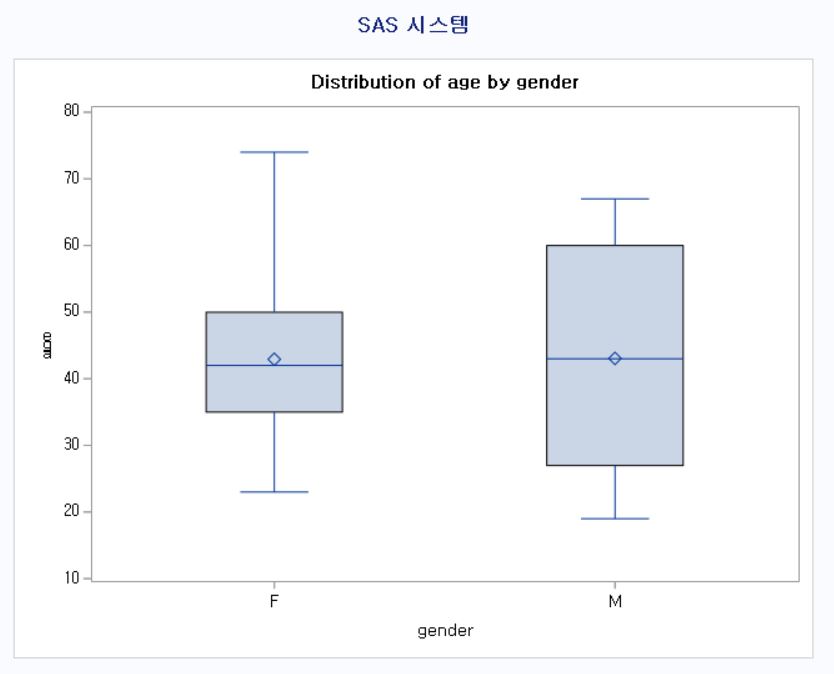
plot age\*gender / boxstyle=skeletal

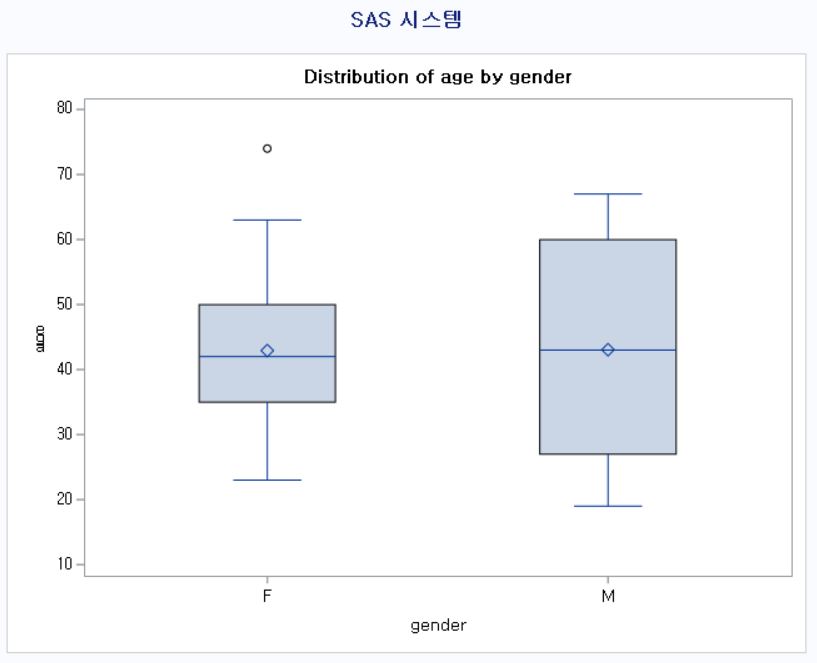
boxwidth=**20** hoffset=**15**;

plot age\*gender / boxstyle=schematic

boxwidth=**20** hoffset=**15**;

**run**;





**<예 2-7>**

**data** sasadv.score;

input dept $ gender $ age score @@;

cards;

Stat M 10 94 Stat F 10 96 Stat M 15 91 Stat M 15 86

Stat F 10 76 Stat M 20 88 Math M 20 71 Math F 20 66

Math M 15 81 Math F 10 77 Math F 15 55 Math F 20 78

;

**run**;

**proc** **means** data=sasadv.score;

class dept gender;

var age score;

**run**;

**proc** **means** data=sasadv.score maxdec=**2** max min mean;

class dept gender;

var age score;

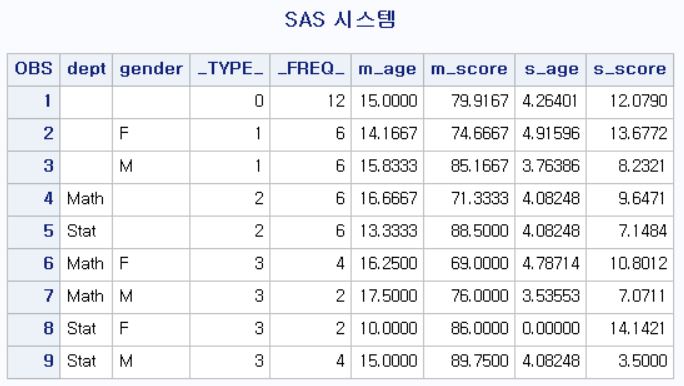
output out=scoreout

mean(age score)=m\_age m\_score std(age score)=s\_age s\_score;

**run**;

**proc** **print** data=scoreout;

**run**;



**<예 2-8>**

**proc** **summary** data=sasadv.score;

class dept gender;

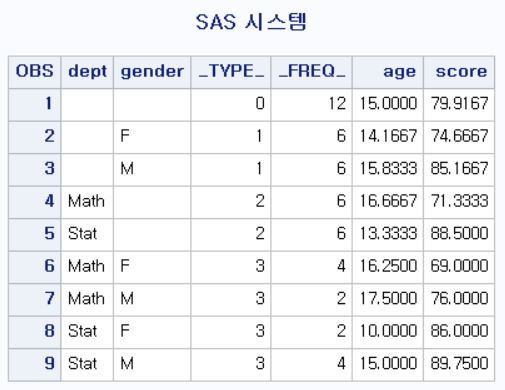
var age score;

output out=n\_score mean(age score)= ;

**run**;

**proc** **print** data=n\_score;

**run**;



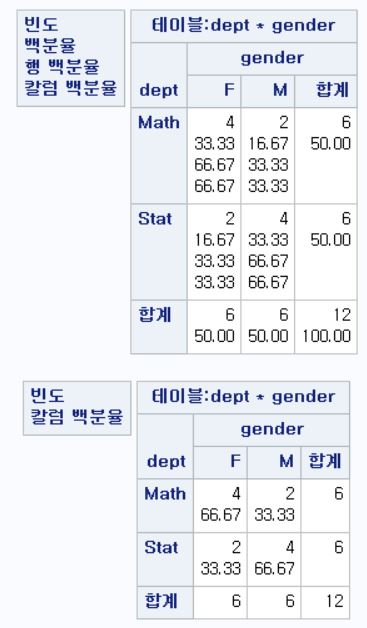
**<예 2-9>**

**proc** **freq** data=sasadv.score;

tables dept gender dept\*gender;

tables dept\*gender / norow nopercent;

**run**;



**<예 2-10>**

**data** sasadv.drink;

input age drink $ count @@;

cards;

18 A 10 19 A 13 20 A 12

18 B 14 19 B 7 20 B 4

18 C 2 19 C 10 20 C 6

18 D 12 19 D 8 20 D 10

;

**run**;

**proc** **freq** data=sasadv.drink;

weight count;

tables age age\*drink / nocol nopercent;

**run**;



- 위의 데이터와 같이, 빈도표의 대상이 되는 데이터세트에 각 칸별 빈도에 관한 정보가 미리 요약되어 있는 경우에는 FREQ 프로시저에서 WEIGHT 명령문에 해당 변수를 지정해줘야 한다.