전산통계 과제#1

컴퓨터소프트웨어 학부
2018008559
신상윤

2-2

코드

```
data protein;
input amount @@;
cards;
10.1 8.9 13.5 7.8 9.7 10.6 8.4 9.5 18.0 10.2
5.3 13.9 9.0 9.5 9.4 6.9 6.2 6.2 7.1 9.9
13.1 17.4 9.3 11.4 4.4
;
run;

proc univariate data=protein plot;
run;
```

결과

UNIVARIATE 프로시저
변수: amount

적률			
N	25	가중합	25
평균	9.828	관측값 합	245.7
표준 편차	3.34707833	분산	11.2029333
왜도	0.88142979	첨도	0.89709682
제곱합	2683.61	수정 제곱합	268.8704
변동계수	34.056556	평균의 표준 오차	0.66941567

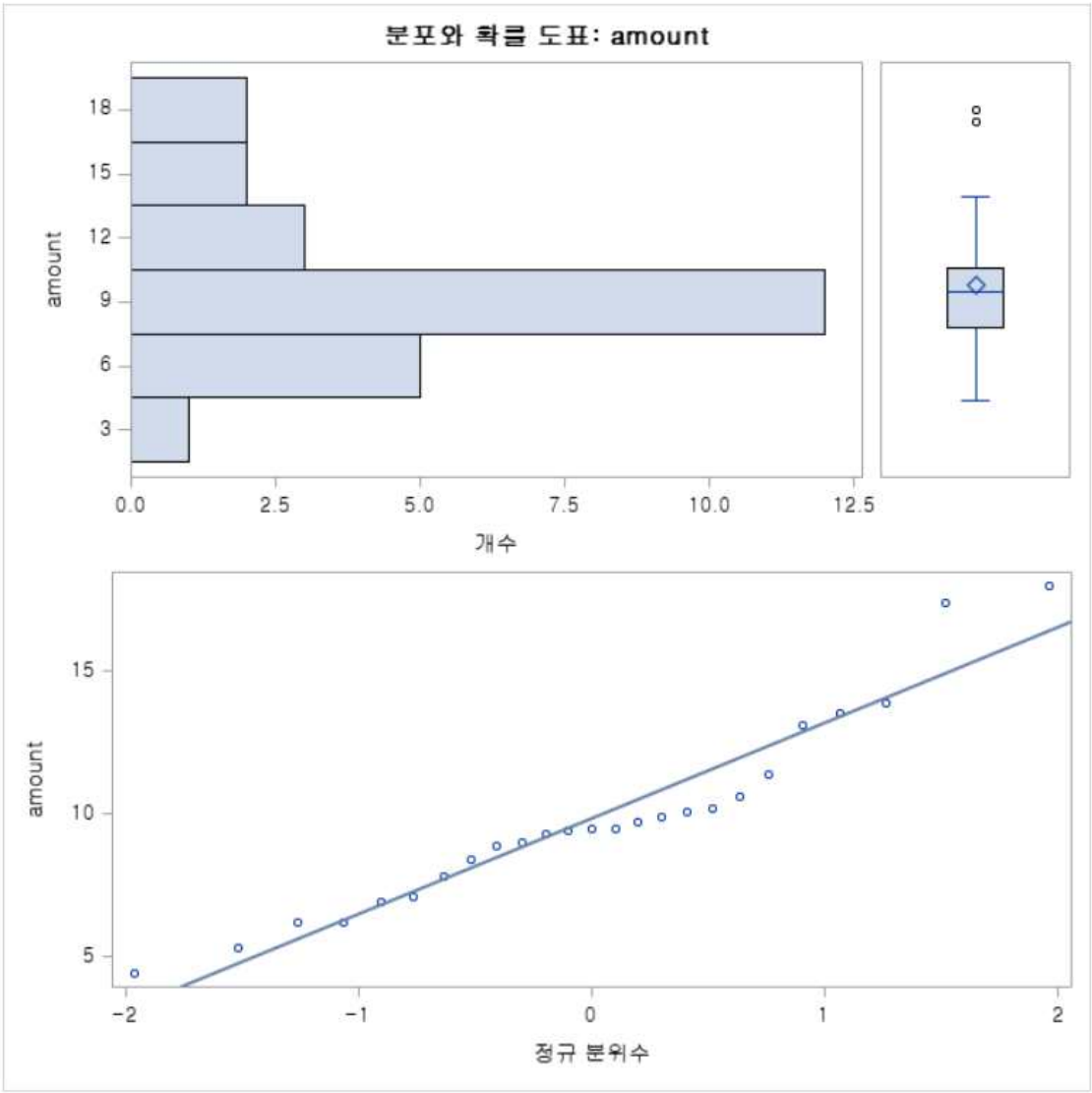
기본 통계 측도			
위치측도		변이측도	
평균	9.828000	표준 편차	3.34708
중위수	9.500000	분산	11.20293
최빈값	6.200000	범위	13.60000
		사분위수 범위	2.80000

Note: 표시된 최빈값은 2개의 최빈값(개수: 2) 중에 가장 작습니다.

위치모수 검정: Mu0=0				
검정	통계량		p 값	
스튜던트의 t	t	14.68146	Pr > |t|	<.0001
부호	M	12.5	Pr >= |M|	<.0001
부호 순위	S	162.5	Pr >= |S|	<.0001

분위수(정의 5)	
레벨	분위수
100% 최댓값	18.0
99%	18.0
95%	17.4
90%	13.9
75% Q3	10.6
50% 중위수	9.5
25% Q1	7.8
10%	6.2
5%	5.3
1%	4.4
0% 최솟값	4.4

극 관측값			
최소		최대	
값	관측값	값	관측값
4.4	25	13.1	21
5.3	11	13.5	3
6.2	18	13.9	12
6.2	17	17.4	22
6.9	16	18.0	9



univariate로 기술통계량을 출력하고, 옵션으로 plot을 추가하여 줄기-잎 그림, 상자그림, 정규확률도를 그림으로 나타나도록 하였다.
상자그림을 보면 극단값인 18.0 17.4가 있는 것을 알 수 있다.

2-3

코드

```
data king;
input name:$9. 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;
proc univariate data=king plot;
        class war;
        var age;
run;
proc means data=king maxdec=2 max min median mean stddev;
        class war;
        var age;
        output out=kingout
        max(age)=max_age min(age)=min_age median(age)=mid_age
mean(age)=m_age std(age)=s_age;
run;
proc print data=kingout;
run;
proc freq data=king;
        table war agegroup war*agegroup;
run;
```

결과

(가)

UNIVARIATE 프로시저
변수: age
war = After

적률			
N	13	가중합	13
평균	48.8461538	관측값 합	635
표준 편차	16.7970541	분산	282.141026
왜도	0.36684329	첨도	-0.2695737
제곱합	34403	수정 제곱합	3385.69231
변동계수	34.3876698	평균의 표준 오차	4.6586646

기본 통계 측도			
위치측도		변이측도	
평균	48.84615	표준 편차	16.79705
중위수	48.00000	분산	282.14103
최빈값	.	범위	60.00000
		사분위수 범위	23.00000

위치모수 검정: Mu0=0				
검정	통계량		p 값	
스튜던트의 t	t	10.48501	Pr > |t|	<.0001
부호	M	6.5	Pr >= |M|	0.0002
부호 순위	S	45.5	Pr >= |S|	0.0002

분위수(정의 5)	
레벨	분위수
100% 최댓값	82
99%	82
95%	82
90%	67
75% Q3	59
50% 중위수	48
25% Q1	36
10%	32
5%	22
1%	22
0% 최솟값	22

극 관측값			
최소		최대	
값	관측값	값	관측값
22	24	54	16
32	25	59	19
33	18	66	15
36	20	67	26
40	17	82	21

UNIVARIATE 프로시저
변수: age
war = Before

적률			
N	14	가중합	14
평균	43.4285714	관측값 합	608
표준 편차	15.7319581	분산	247.494505
왜도	0.15770705	첨도	-0.5496137
제곱합	29622	수정 제곱합	3217.42857
변동계수	36.2249035	평균의 표준 오차	4.20454266

기본 통계 측도			
위치측도		변이측도	
평균	43.42857	표준 편차	15.73196
중위수	41.50000	분산	247.49451
최빈값	30.00000	범위	57.00000
		사분위수 범위	26.00000

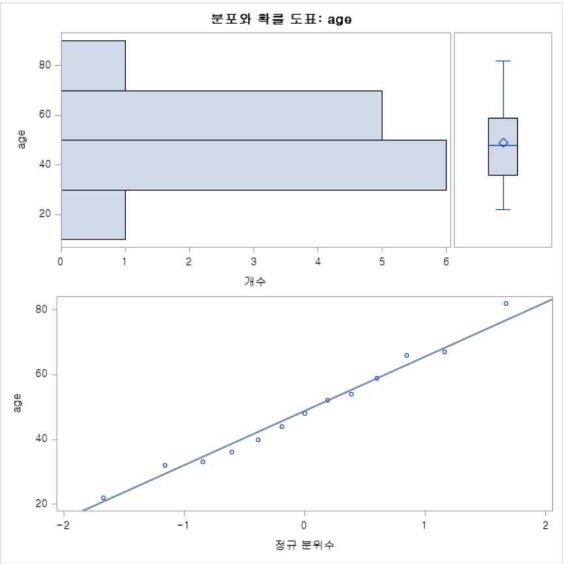
Note: 표시된 최빈값은 2개의 최빈값(개수: 2) 중에 가장 작습니다.

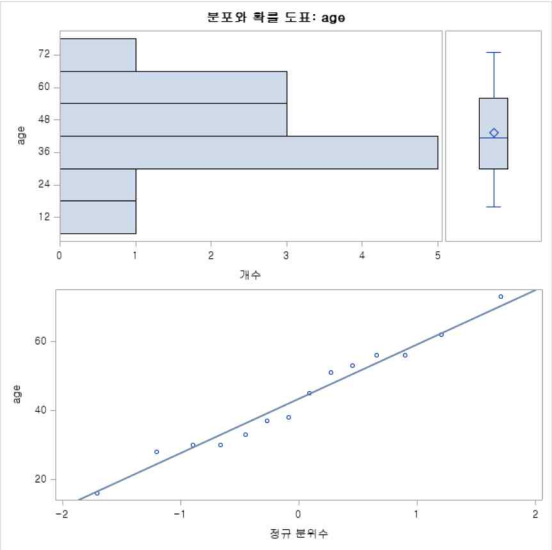
위치모수 검정: Mu0=0				
검정	통계량		p 값	
스튜던트의 t	t	10.32896	Pr > |t|	<.0001
부호	M	7	Pr >= |M|	0.0001
부호 순위	S	52.5	Pr >= |S|	0.0001

분위수(정의 5)	
레벨	분위수
100% 최댓값	73.0
99%	73.0
95%	73.0
90%	62.0
75% Q3	56.0
50% 중위수	41.5
25% Q1	30.0
10%	28.0
5%	16.0
1%	16.0
0% 최솟값	16.0

극 관측값			
최소		최대	
값	관측값	값	관측값
16	6	53	4
28	8	56	11
30	12	56	14
30	10	62	2
33	13	73	1

전쟁 이후에는 이전보다 왕들의 평균 수명이 5살 높았다.
표준편차는 비슷한 걸로 보아 전쟁 전후의 수명분포는 비슷하다 볼 수 있다. 중간값도 전쟁 이후 수명이 더 높았다.
수명의 최솟값, 최댓값은 각각
전쟁전 16, 73세
전쟁후 22, 82세로
두 경우 모두 전쟁 후가 높았다.
전쟁 전 수명의 최댓값인 73세는 전체 데이터 중 2번째로 큰 값이다.





왼쪽이 전쟁 후, 오른쪽이 전쟁 전이다. 분포가 비슷하다는 것을 볼 수 있다.

(나)

MEANS 프로시저

분석 변수: age						
war	관측값 수	최댓값	최솟값	중위수	평균	표준편차
After	13	82.00	22.00	48.00	48.85	16.80
Before	14	73.00	16.00	41.50	43.43	15.73

OBS	war	_TYPE_	_FREQ_	max_age	min_age	mid_age	m_age	s_age
1		0	27	82	16	45.0	46.0370	16.1733
2	After	1	13	82	22	48.0	48.8462	16.7971
3	Before	1	14	73	16	41.5	43.4286	15.7320

(저장된 데이터)

(다)

FREQ 프로시저

war	빈도	백분율	누적 빈도	누적 백분율
After	13	48.15	13	48.15
Before	14	51.85	27	100.00

agegroup	빈도	백분율	누적 빈도	누적 백분율
10	1	3.70	1	3.70
20	2	7.41	3	11.11
30	8	29.63	11	40.74
40	4	14.81	15	55.56
50	7	25.93	22	81.48
60	3	11.11	25	92.59
70	1	3.70	26	96.30
80	1	3.70	27	100.00

빈도
백분율
행 백분율
칼럼 백분율

테이블 war * agegroup									
	agegroup								
war	10	20	30	40	50	60	70	80	합계
After	0 0.00 0.00 0.00	1 3.70 7.69 50.00	3 11.11 23.08 37.50	3 11.11 23.08 75.00	3 11.11 23.08 42.86	2 7.41 15.38 66.67	0 0.00 0.00 0.00	1 3.70 7.69 100.00	13 48.15
Before	1 3.70 7.14 100.00	1 3.70 7.14 50.00	5 18.52 35.71 62.50	1 3.70 7.14 25.00	4 14.81 28.57 57.14	1 3.70 7.14 33.33	1 3.70 7.14 100.00	0 0.00 0.00 0.00	14 51.85
합계	1 3.70	2 7.41	8 29.63	4 14.81	7 25.93	3 11.11	1 3.70	1 3.70	27 100.00

2-4

코드

```
data car;
input size $ manufact $ model $ mileage reliable index;
cards;
Small Chevrolet GeoPrizm 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 LeBaron 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;
proc means data=car maxdec=3 mean stddev;
        var mileage reliable;
        output out = carout
        mean(mileage reliable)=m_mileage m_reliable
        std(mileage reliable)=s_mileage s_reliable;
run;
proc print data=carout;
run;
proc freq data=car;
        table size index size*index;
run;
```

결과

(가)

MEANS 프로시저

변수	평균	표준편차
mileage	26,000	4,169
reliable	3,385	1,502

결측치는 빼고 계산을 한 것을 알 수 있다.

또한 저장한 데이터를 따로 출력해주었다.

OBS	_TYPE_	_FREQ_	m_mileage	m_reliable	s_mileage	s_reliable
1	0	14	26	3.38462	4.16949	1.50214

(나)

FREQ 프로시저

size	빈도	백분율	누적 빈도	누적 백분율
Compact	6	42.86	6	42.86
Mid-Size	3	21.43	9	64.29
Small	5	35.71	14	100.00

index	빈도	백분율	누적 빈도	누적 백분율
3	8	61.54	8	61.54
4	5	38.46	13	100.00
결측값 빈도 = 1				

마찬가지로 결측치는 따로 나타내어 준다.

빈도
백분율
행 백분율
칼럼 백분율

테이블 size * index			
size	index 3	index 4	합계
Compact	6 46.15 100.00 75.00	0 0.00 0.00 0.00	6 46.15
Mid-Size	1 7.69 33.33 12.50	2 15.38 66.67 40.00	3 23.08
Small	1 7.69 25.00 12.50	3 23.08 75.00 60.00	4 30.77
합계	8 61.54	5 38.46	13 100.00
결측값 빈도 = 1			

2-7

코드

```
data b_smoking;
input ppm @@;
cards;
72 70 68 67 73 71 72 70 69 70 68
72 69 66 73 71 70 72 70 69 72 73
;
run;
data a_smoking;
input ppm @@;
cards;
74 72 69 68 72 72 72 71 67 73 69
71 68 74 73 70 74 68 71 74 74 69
;
run;
proc freq data=b_smoking;
        table ppm;
run;
proc freq data=a_smoking;
        table ppm;
run;
proc univariate data=b_smoking;
        histogram ppm / vaxis=0 to 30 by 5
                                midpoints=65 to 75 by 1;
run;
proc univariate data=a_smoking;
        histogram ppm / vaxis=0 to 30 by 5
                                midpoints=65 to 75 by 1;
run;
proc means data=b_smoking maxdec=2 mean median std;
run;
proc means data=a_smoking maxdec=2 mean median std;
run;
```

결과

전

FREQ 프로시저

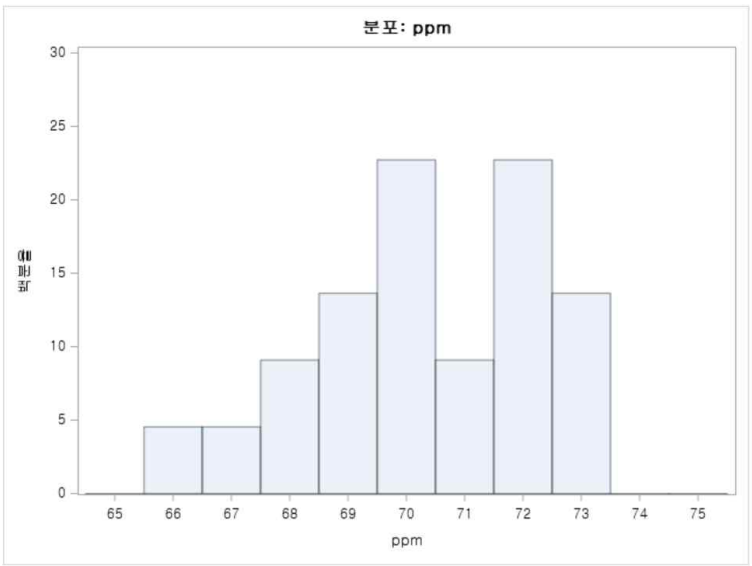
ppm	빈도	백분율	누적 빈도	누적 백분율
66	1	4.55	1	4.55
67	1	4.55	2	9.09
68	2	9.09	4	18.18
69	3	13.64	7	31.82
70	5	22.73	12	54.55
71	2	9.09	14	63.64
72	5	22.73	19	86.36
73	3	13.64	22	100.00

후

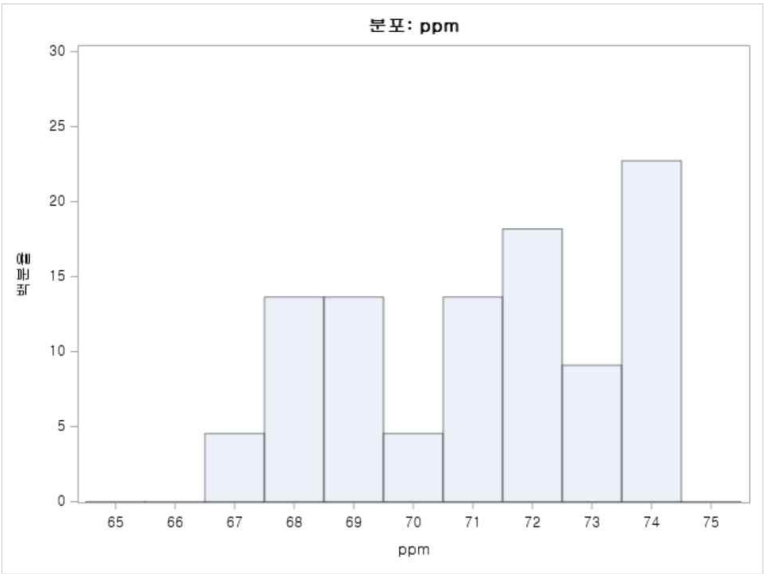
FREQ 프로시저

ppm	빈도	백분율	누적 빈도	누적 백분율
67	1	4.55	1	4.55
68	3	13.64	4	18.18
69	3	13.64	7	31.82
70	1	4.55	8	36.36
71	3	13.64	11	50.00
72	4	18.18	15	68.18
73	2	9.09	17	77.27
74	5	22.73	22	100.00

전



후



비교를 편하게 하기 위해 따로 중요 기술 통계량을 나타냈다.

분석 변수: ppm		
평균	중위수	표준편차
70.32	70.00	1.99

분석 변수: ppm		
평균	중위수	표준편차
71.14	71.50	2.32

담배를 피운 후의 맥박수는 평균도 높았고, 중위수도 높았다. 이는 담배를 피운 후 전체적으로 맥박수가 높아졌다는 것을 의미한다.
그러나 평균은 0.82 정도 올라갔는데 담배를 피우든, 피우지 않든 표준편차가 2 이상인 것으로 보아 유의미한 결과는 아닌 것 같다.

다시 정리하면 평균, 중위수 등을 비교해보면 담배를 피운 후 전체적으로 맥박수는 높아졌다고 할 수 있으나 그 차이는 표준편차보다 작아 담배 때문에 맥박수가 높아졌다고 보기는 어려울 것 같다.

2-8

코드

```
data temp;
input Temperature @@;
day = _N_;
cards;
24 27 28 32 30 35 26 29 31 30
35 37 34 33 34 29 35 32 34 34
29 30 29 36 35 33 37 29 40 32
;
run;
proc freq data=temp;
        table Temperature;
run;
proc univariate data=temp;
        var Temperature;
        histogram Temperature / vaxis=0 to 20 by 5
                                midpoints=24 to 40 by 1;
```

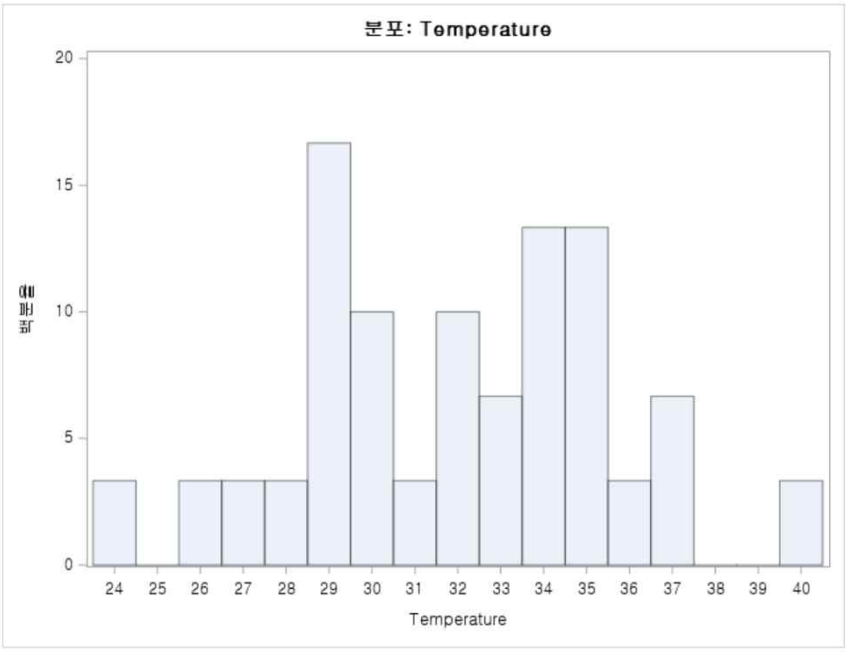
```
run;
proc sgplot data=temp;
        series x = day y = Temperature / MARKERS LINEATTRS =
(THICKNESS = 2);
        XAXIS TYPE = DISCRETE;
run;
proc means data=temp maxdec=2 mean median var std;
        var Temperature;
run;
```

결과

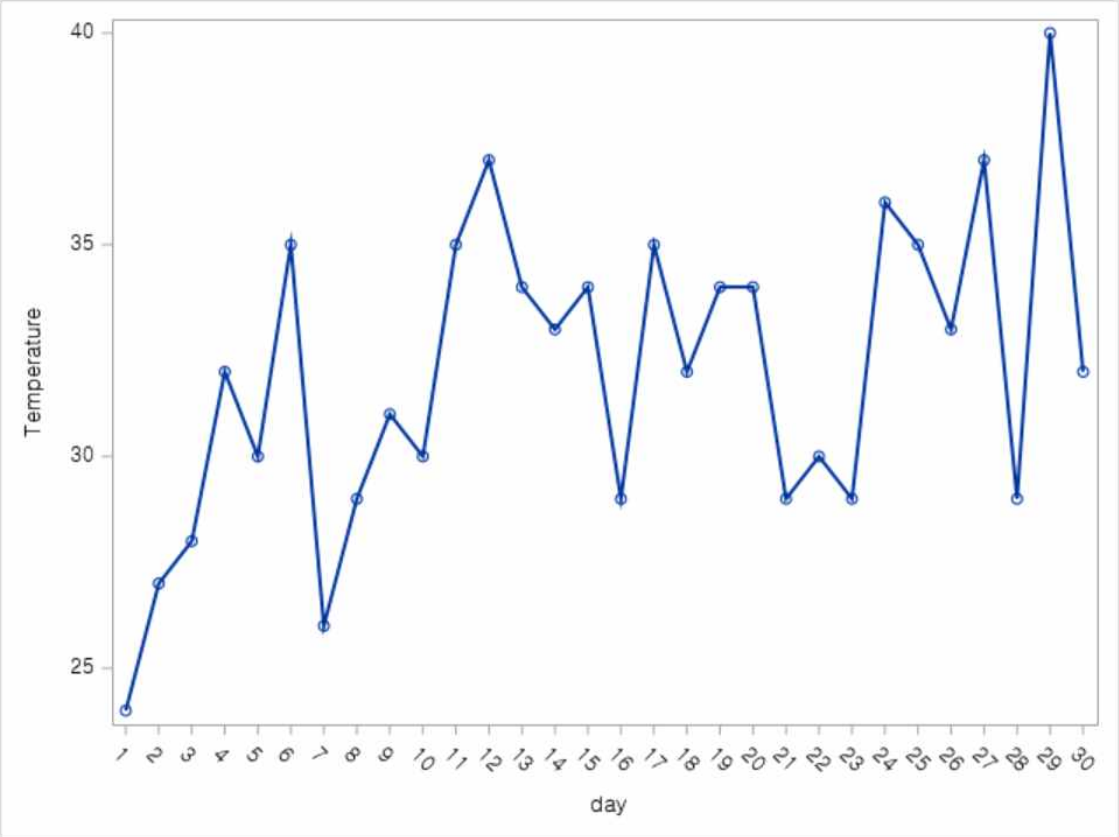
(가)

FREQ 프로시저

Temperature	빈도	백분율	누적 빈도	누적 백분율
24	1	3.33	1	3.33
26	1	3.33	2	6.67
27	1	3.33	3	10.00
28	1	3.33	4	13.33
29	5	16.67	9	30.00
30	3	10.00	12	40.00
31	1	3.33	13	43.33
32	3	10.00	16	53.33
33	2	6.67	18	60.00
34	4	13.33	22	73.33
35	4	13.33	26	86.67
36	1	3.33	27	90.00
37	2	6.67	29	96.67
40	1	3.33	30	100.00



(나)



(다)

MEANS 프로시저

분석 변수: Temperature			
평균	중위수	분산	표준편차
31.97	32.00	13.21	3.63

2-9

코드

```
data  Aco;
input  val  @@;
cards;
95  96  92  102  103  93  101  92  95  90
;
```

```
run;
data Bco;
input val @@;
cards;
184 202 215 204 195 201 169 182 192
;
run;
data Cco;
input val @@;
cards;
215 214 197 216 215 208 228 208 216 214 227
;
run;
data Dco;
input val @@;
cards;
155 142 146 149 146 152 159
;
run;
proc freq data=Aco;
        table val;
run;
proc freq data=Bco;
        table val;
run;
proc freq data=Cco;
        table val;
run;
proc freq data=Dco;
        table val;
run;
proc univariate data=Aco;
        histogram val / vaxis=0 to 40 by 10
                        midpoints=90 to 105 by 5;
```

```
run;
proc univariate data=Bco;
        histogram val / vaxis=0 to 40 by 5
                        midpoints=165 to 215 by 10;
run;
proc univariate data=Cco;
        histogram val / vaxis=0 to 40 by 10
                        midpoints=190 to 230 by 10;
run;
proc univariate data=Dco;
        histogram val / vaxis=0 to 40 by 10
                        midpoints=140 to 160 by 5;
run;
proc means data=Aco maxdec=2 mean median std;
run;
proc means data=Bco maxdec=2 mean median std;
run;
proc means data=Cco maxdec=2 mean median std;
run;
proc means data=Dco maxdec=2 mean median std;
run;
```

결과

(가)

A

val	빈도	백분율	누적 빈도	누적 백분율
90	1	10.00	1	10.00
92	2	20.00	3	30.00
93	1	10.00	4	40.00
95	2	20.00	6	60.00
96	1	10.00	7	70.00
101	1	10.00	8	80.00
102	1	10.00	9	90.00
103	1	10.00	10	100.00

B

val	빈도	백분율	누적 빈도	누적 백분율
169	1	11.11	1	11.11
182	1	11.11	2	22.22
184	1	11.11	3	33.33
192	1	11.11	4	44.44
195	1	11.11	5	55.56
201	1	11.11	6	66.67
202	1	11.11	7	77.78
204	1	11.11	8	88.89
215	1	11.11	9	100.00

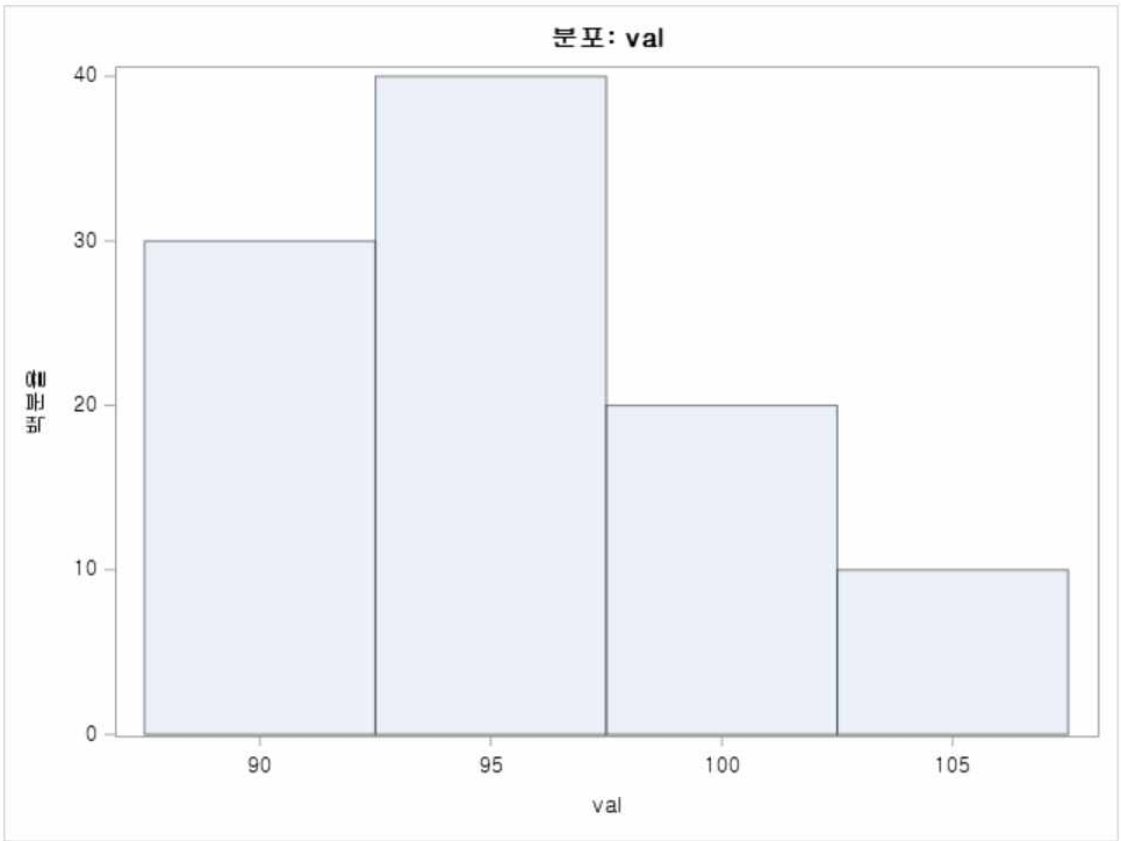
C

val	빈도	백분율	누적 빈도	누적 백분율
197	1	9.09	1	9.09
208	2	18.18	3	27.27
214	2	18.18	5	45.45
215	2	18.18	7	63.64
216	2	18.18	9	81.82
227	1	9.09	10	90.91
228	1	9.09	11	100.00

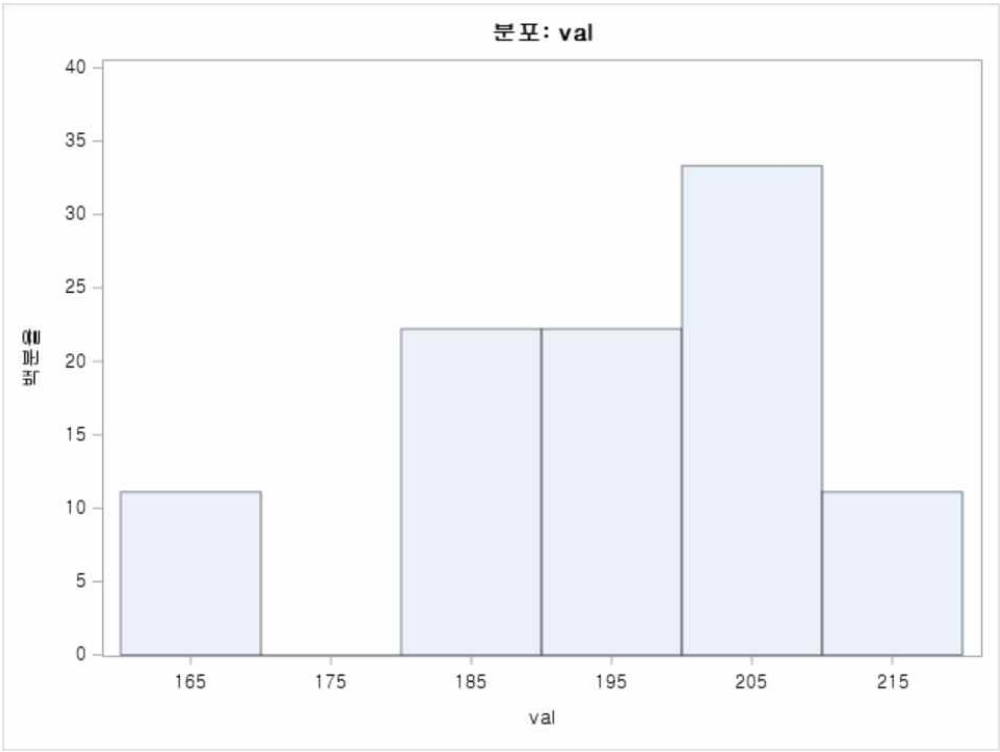
D

val	빈도	백분율	누적 빈도	누적 백분율
142	1	14.29	1	14.29
146	2	28.57	3	42.86
149	1	14.29	4	57.14
152	1	14.29	5	71.43
155	1	14.29	6	85.71
159	1	14.29	7	100.00

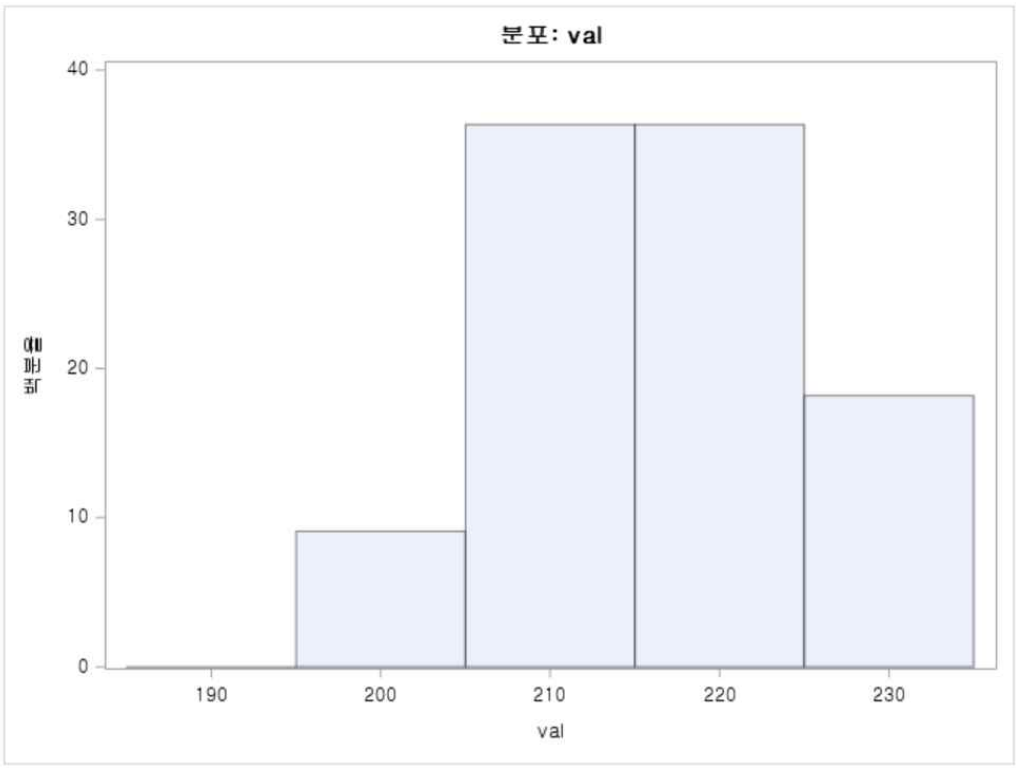
A



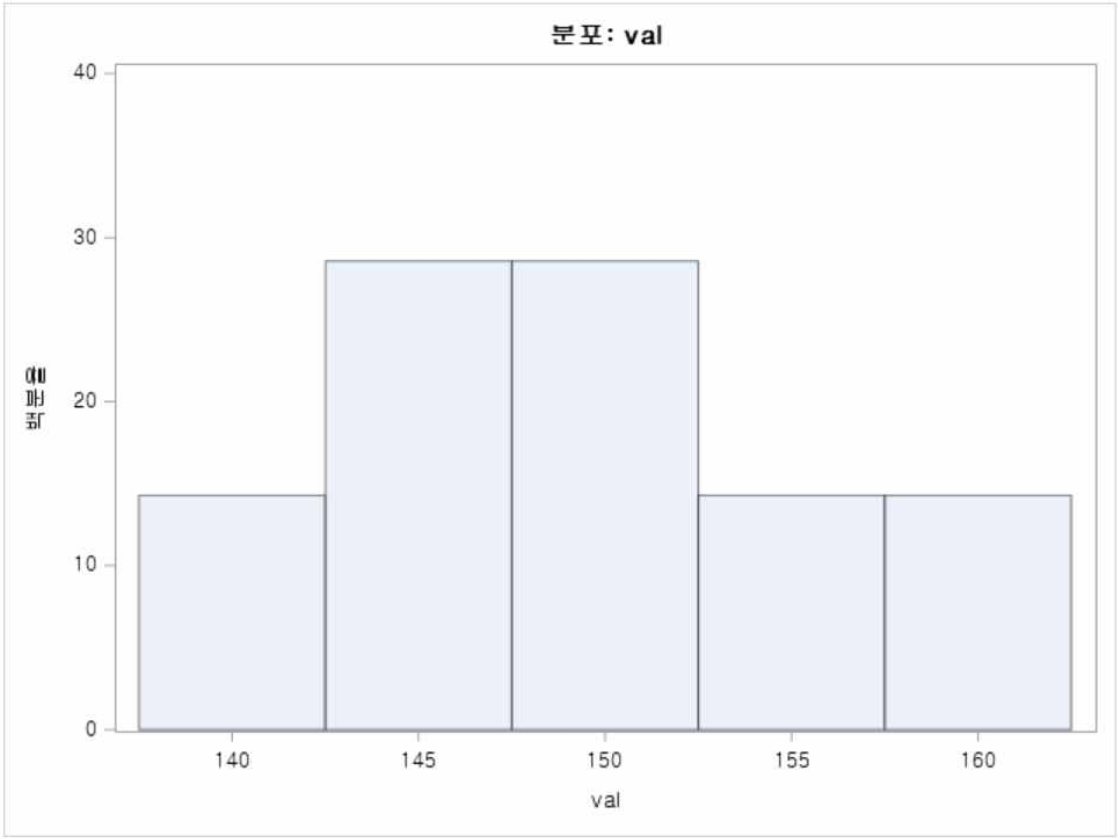
B


분포: val
40
35
30
25
20
15
10
5
0
퍼센트
165
175
185
195
205
215
val

C


분포: val
40
30
20
10
0
퍼센트
190
200
210
220
230
val

D



A

분석 변수: val		
평균	중위수	표준편차
95.90	95.00	4.58

B

분석 변수: val		
평균	중위수	표준편차
193.78	195.00	13.82

C

분석 변수: val		
평균	중위수	표준편차
214.36	215.00	8.57

D

분석 변수: val		
평균	중위수	표준편차
149.86	149.00	5.87

(나)

공기 중 일산화탄소량은 A < D < B < C 순서로 많다.
A지역과 D지역, D지역과 B지역은 각각 평균이 54, 44정도 차이가 나는데 이는 큰 차이로 확실하게 A지역이 D지역보다 일산화탄소량이 적다, D지역이 B지역보다 일산화탄소량이 적다고 할 수 있다. 그러나 B, C 지역은 평균이 20 정도 차이가 나지만 상대적으로 큰 B, C 지역의 표준편차까지 고려하면 두 지역은 공기 중 일산화탄소량이 비슷하게 많다고 볼 수 있다. 즉, 3단계로 구별을 하면 A는 1단계, D는 2단계, B, C는 3단계라 할 수 있다.