

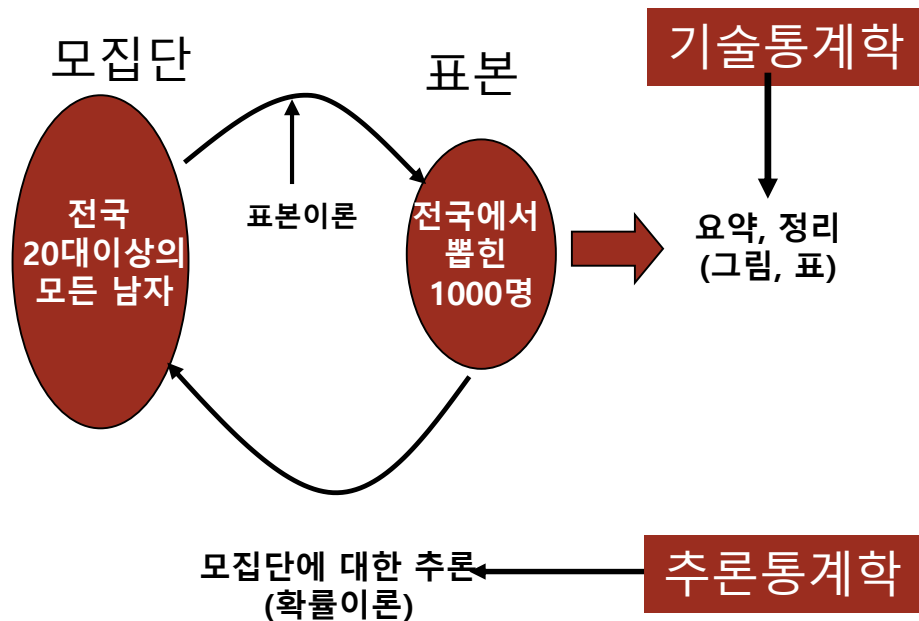
통계계산소프트웨어

# 데이터 요약 및 표현

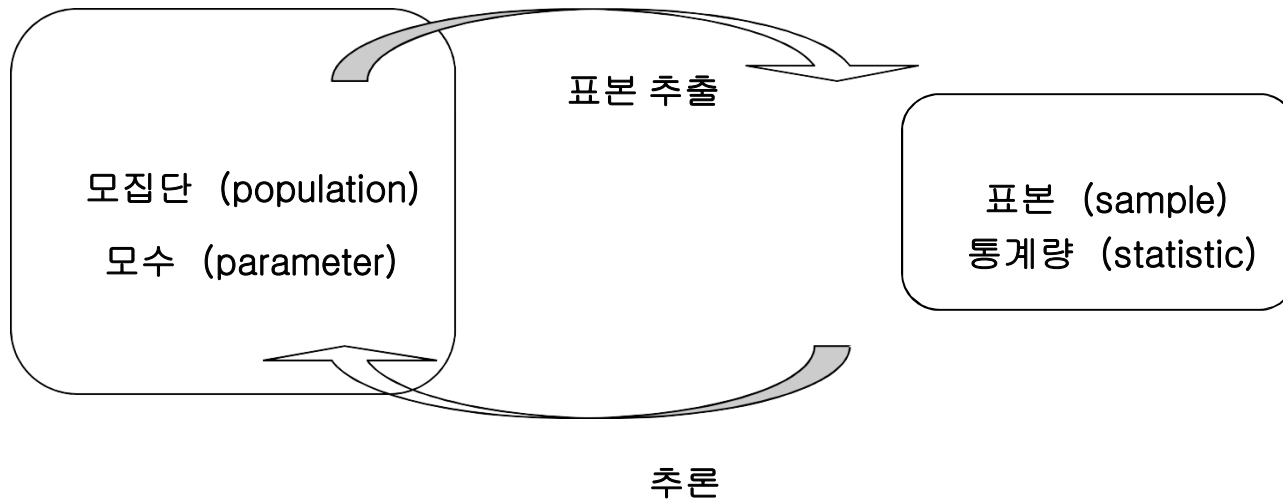
2018. 10.

# 통계학의 두 방향

예) 우리나라 성인남자의 키 조사(1000명의 표본)



# 모집단과 표본 사이의 관계



통계학은 표본의 자료를 수집, 정리, 요약하고 나아가 요약된 자료를 토대로 그 자료의 모태가 되는 모집단에 대해 짐작, 추측해 보는 작업을 포함한다. 일반적으로 통계자료란 표본으로 추출된 관측값 들의 집합을 의미하며, 이러한 통계자료를 요약 및 정리하는 방법은

첫째, 통계자료들을 도수분포표 또는 다양한 형태의 그래프로 표현  
둘째, 주어진 자료를 대표하는 대표값과 산포도로 자료를 요약

# 목 차



**PROC MEANS**



**PROC UNIVARIATE**



**PROC FREQ**



**PROC STANDARD**



## PROC MEANS

# PROC MEANS

## (기초통계)

```
PROC MEANS DATA=sas-dataset-name <dataset-options > ;
  VAR      variablelist ;
  OUTPUT   OUT=libref.dsn  stat1=var1 statn=varn / options ;
  CLASS    variablelist / options ;
  FREQ     variable ;      WEIGHT variable ;
  TYPES    requests ;      WAYS    lists ;
  BY       variablelist ;  ID      variablelist ;
RUN;
```

| STATEMENT          | 내용  |
|--------------------|---|
| <b>VAR</b>         | 통계량을 산출할 연속형 변수 지정  |
| <b>CLASS</b>       | 통계량을 산출할 그룹지정, 2 <sup>k</sup> 만큼 _TYPE_ 값이 산출됨            |
| <b>FREQ</b>        | 동일한 데이터가 변수로 지정된 수만큼 있다고 가정하고 계산                          |
| <b>WEIGHT</b>      | 중요도 또는 수정계수 등의 가중치  |
| <b>TYPES</b>       | CLASS 그룹 중에서 출력할 그룹 A*B B*C 형태, 반드시 CLASS문이 있어야함.         |
| <b>WAYS</b>        | CLASS 로 결합되는 변수 수 지정, If 2 , then A*B B*C C*A 등           |
| <b>OUTPUT OUT=</b> | 통계량을 저장할 데이터셋 이름 지정                                       |
|                    | N=n1-n3 mean=ave1-ave3 std=std1-std3 sum=sum1-sum3 통계량=변수 |

# PROC MEANS

- 모집단의 평균비교와 같은 가설검정 등의 통계적 추론을 수행하기 전에  
먼저 수집된 자료(표본)의 통계량을 정리하고 요약하는 것이 중요
- 모집단 분포의 특성을 추정할 수 있는 수치적 측도로  
대표값, 산포도, 형상측도가 있다.
  - 대표값 : 분포의 중심경향을 나타냄. 평균값, 중앙값, 최빈값 등
  - 산포도 : 자료가 대표값을 중심으로 얼마나 퍼져있는가를 나타냄.  
분산, 표준편차, 범위 등
  - 형상측도 : 분포의 모양을 설명. 왜도(비대칭 정도), 첨도(분포가 얼마나 첨예한지)  
예) 왜도가 0이면 좌우가 완전 대칭인 분포. 왼쪽으로 치우친 분포는 0보다 큼.  
정규분포의 첨도는 3이지만 SAS에서는 첨도를 0으로 정의. 첨도가 0보다  
크면 정규분포보다 뾰족한 분포라고 예측
- MEANS 절차는 숫자형 변수의 관측치의 개수, 평균, 표준편차, 최소값, 최대값 등의  
기술통계량을 구해준다. OUTPUT 부명령문을 이용하여 여러 기술통계량을  
SAS 데이터 셋의 변수로 저장할 수 있다.

# PROC MEANS

## ■ PROC MEANS 형식

```
PROC MEANS DATA=SAS-data-set <옵션(들)> <statistic-keyword(s)>;  
  BY 변수(들);  
  CLASS 변수(들); 변수의 자릿값에 따른 개별 통계량을 구하고자 하는 변수 지정  
  OUTPUT OUT=SAS-data-set 통계량=저장변수명...;  
  VAR 변수(들); 분석에 사용할 숫자변수들 지정  
                  (대부분의 분석 프로시저에서 사용됨)  
                  VAR 변수명 안쓰면 숫자형 변수에 대해서는 다 나옴  
  WEIGHT 변수;  
RUN;
```

- ✓ BY와 CLASS 모두 변수별로 구할 때 사용하는 같은 의미의 명령어이지만 출력 결과 형식이 다르며, BY문은 이전에 **SORT로 정렬**되어야 함  
CLASS 문은 지정한 변수(CLASS 변수)의 값으로 부그룹(subgroup)을 만들고 부그룹별로 통계량을 따로 구하고자 할 때 사용
- ✓ 관측치 개수(결측제외), 평균, 표준편차, 최소값 그리고 최대값이 기본으로 출력되는 통계량임
- ✓ MEANS 프로시저는 리포트 생성에, SUMMARY 프로시저는 data set 생성을 위해 주로 사용함. SUMMARY는 print 설정해야 출력됨



## PROC MEANS 절차의 옵션들

| 구분         | KEYWORD                                       | 내용  |
|------------|---|---|
| 프로시저<br>옵션 | <b>MAXDEC=n</b>                               | 출력되는 통계량의 값을 소수 n자리까지만 출력   |
|            | <b>통계량</b>                                    | Default : N, MEAN, STD, MIN, MAX<br>기타 : CV, RANGE, VAR, USS, CSS, SKEWNESS,... |
|            | <b>NOPRINT</b>                                | 인쇄하지 마세요. (PROC SUMMARY)  |
|            | <b>DESCENDING</b>                             | _TYPE_ 내림차순으로 출력(default=ASCENDING)   |
|            | <b>NWAY</b>                                   | _TYPE_가 가장 큰 값만 출력  |
| 출력<br>제어문  | WHERE<br>LABEL<br>FORMAT<br>TITLE<br>FOOTNOTE |   |

# PROC MEANS

- Means 절차의 statistic-keyword (지정하지 않으면 N, MEAN, STD, MIN MAX 출력)
  - ✓ N : 각 변수에 대해서 관찰값(observation)의 개수를 출력  
만일 데이터에 결측치가 있을 경우는 결측치가 없는 관찰값의 개수만을 출력
  - ✓ NMISS : 각 변수별로 결측값(missing value) 의 개수를 출력
  - ✓ MEAN : 평균값
  - ✓ SUM : 관찰값(observation) 의 합계
  - ✓ STD : 표본표준편차(Standard deviation)
  - ✓ MIN : 최소값(Minimum)
  - ✓ MAX : 최대값(Maximum)
  - ✓ RANGE : 범위 (MAX - MIN )
  - ✓ VAR : 분산(Variance)
  - ✓ USS(Uncorrected Sum of Squares) : 각 관찰값 들의 제곱의 합
  - ✓ CSS(Corrected Sum of Squares) : 평균에 의해 수정된 관찰값 들의 제곱의 합

# PROC MEANS

## ■ Means 절차의 statistic-keyword

- ✓ STDERR(Standard error) : 표준오차, 평균에 관한 표준편차
- ✓ CV : 변이(변동) 계수 (Coefficient of Variation)  
표준편차/표본평균\*100  
변이계수를 이용하면 평균치가 다른 집단이나 단위가 다른  
집단의 산포도(散布度)를 비교할 수 있음  
이 계수가 작을수록 평균치 가까이에 분포하고 있음
- ✓ T : 해당 변수의 모집단의 평균이 0 이란 귀무가설하 에서의 Student' s  
t 값을 출력
- ✓ PRT : 유의확률(p-value) : '실제는 아닌데도 오차나 우연에 의해  
데이터와 같은 차이가 생길 확률'. P값이 작을때 유의하다라고 판단
- ✓ SKEWNESS : 비대칭도(왜도). 확률분포 곡선의 비대칭 정도를  
나타내는 척도
- ✓ KURTOSIS : 뾰족함의 정도(첨도).

# PROC MEANS\_예제

```
DATA htw;  
  INPUT name $ sex $ dept $ age height weight;  
CARDS;  
김철수 M Stat 25 170 67  
강민호 M Stat 20 169 70  
이영희 F Math 19 160 58  
박지수 F Econ 21 160 59  
최병호 M Math 28 177 62  
장순미 F Stat 22 173 60  
이상호 M Econ 19 170 71  
김미숙 F Math 16 150 48  
박흥식 M Econ 20 165 53  
유은영 F Stat 16 169 57  
;  
RUN;
```

# PROC MEANS\_예제

```
/* PROC MEANS ~ by 사용 */
```

```
PROC SORT DATA=htwt;  
BY sex ; RUN;
```

```
PROC MEANS DATA=htwt MAXDEC=2 ;  
  BY sex ;  
  VAR age height weight;  
RUN;
```

*SAS 시스템*

*MEANS 프로시저*

*sex=F*

| 변수     | N | 평균     | 표준편차 | 최솟값    | 최댓값    |
|--------|---|--------|------|--------|--------|
| age    | 5 | 18,80  | 2,77 | 16,00  | 22,00  |
| height | 5 | 162,40 | 8,96 | 150,00 | 173,00 |
| weight | 5 | 56,40  | 4,83 | 48,00  | 60,00  |

*sex=M*

| 변수     | N | 평균     | 표준편차 | 최솟값    | 최댓값    |
|--------|---|--------|------|--------|--------|
| age    | 5 | 22,40  | 3,91 | 19,00  | 28,00  |
| height | 5 | 170,20 | 4,32 | 165,00 | 177,00 |
| weight | 5 | 64,60  | 7,37 | 53,00  | 71,00  |

# PROC MEANS\_예제

```
/* PROC MEANS ~ class 사용 */
```

```
PROC MEANS DATA=htwt MAXDEC=2;  
  CLASS sex ;  
  VAR age height weight;  
RUN;
```

*SAS 시스템*

*MEANS 프로시저*

| sex | 관측값 수 | 변수     | N | 평균     | 표준편차 | 최솟값    | 최댓값    |
|-----|-------|--------|---|--------|------|--------|--------|
| F   | 5     | age    | 5 | 18,80  | 2,77 | 16,00  | 22,00  |
|     |       | height | 5 | 162,40 | 8,96 | 150,00 | 173,00 |
|     |       | weight | 5 | 56,40  | 4,83 | 48,00  | 60,00  |
| M   | 5     | age    | 5 | 22,40  | 3,91 | 19,00  | 28,00  |
|     |       | height | 5 | 170,20 | 4,32 | 165,00 | 177,00 |
|     |       | weight | 5 | 64,60  | 7,37 | 53,00  | 71,00  |

# PROC MEANS\_예제 8.1

/\* PROC MEANS ~ OUTPUT 명령문의 사용 \*/

```
PROC MEANS DATA=htwt NOPRINT;  
  CLASS sex dept;  
  VAR age height weight;  
  OUTPUT OUT=htwt_m MEAN=;  
RUN;
```

Noprint : output문을 통해 단순히 새로운  
sas 데이터셋을 얻고자 할 때 사용

\_TYPE\_ : CLASS문과 관련된 자동변수  
\_FREQ\_ : 통계량이 계산될 때 사용되는  
관측수를 의미하는 자동변수

| VIEWTABLE: Work.Htwt_m |     |      |        |        |       |              |        |
|------------------------|-----|------|--------|--------|-------|--------------|--------|
|                        | sex | dept | _TYPE_ | _FREQ_ | age   | height       | weight |
| 1                      |     |      | 0      | 10     | 20,6  | 166,3        | 60,5   |
| 2                      |     | Econ | 1      | 3      | 20    | 165          | 61     |
| 3                      |     | Math | 1      | 3      | 21    | 162,33333333 | 56     |
| 4                      |     | Stat | 1      | 4      | 20,75 | 170,25       | 63,5   |
| 5                      | F   |      | 2      | 5      | 18,8  | 162,4        | 56,4   |
| 6                      | M   |      | 2      | 5      | 22,4  | 170,2        | 64,6   |
| 7                      | F   | Econ | 3      | 1      | 21    | 160          | 59     |
| 8                      | F   | Math | 3      | 2      | 17,5  | 155          | 53     |
| 9                      | F   | Stat | 3      | 2      | 19    | 171          | 58,5   |
| 10                     | M   | Econ | 3      | 2      | 19,5  | 167,5        | 62     |
| 11                     | M   | Math | 3      | 1      | 28    | 177          | 62     |
| 12                     | M   | Stat | 3      | 2      | 22,5  | 169,5        | 68,5   |

# PROC MEANS\_예제

/\* PROC MEANS ~ OUTPUT 명령문의 사용 \*/

```
PROC MEANS DATA=htwt NOPRINT NWAY;  
  CLASS sex dept;  
  VAR age height weight;  
  OUTPUT OUT=htwt_m MEAN=;  
RUN;
```

/\* NWAY : 마지막 TYPE에 대해서만 출력  
최대 조합구성(제일 큰 \_TYPE\_)만  
출력 데이터 셋에 생성 \*/

| VIEWTABLE: Work.Htwt_m |     |      |        |        |      |        |        |
|------------------------|-----|------|--------|--------|------|--------|--------|
|                        | sex | dept | _TYPE_ | _FREQ_ | age  | height | weight |
| 1                      | F   | Econ | 3      | 1      | 21   | 160    | 59     |
| 2                      | F   | Math | 3      | 2      | 17,5 | 155    | 53     |
| 3                      | F   | Stat | 3      | 2      | 19   | 171    | 58,5   |
| 4                      | M   | Econ | 3      | 2      | 19,5 | 167,5  | 62     |
| 5                      | M   | Math | 3      | 1      | 28   | 177    | 62     |
| 6                      | M   | Stat | 3      | 2      | 22,5 | 169,5  | 68,5   |



## PROC MEANS\_예제 8.2

/\* OUTPUT 명령문\_새로운 변수 이름의 지정 \*/

```
PROC MEANS DATA=htwt NOPRINT;
  CLASS sex dept;
  VAR age height weight;
  OUTPUT OUT=htwt_m1
    MEAN(age height weight) = mean_a mean_h mean_w
    SUM(age height weight) = sum_a sum_h sum_w;
RUN;
```

|    | sex | dept | _TYPE_ | _FREQ_ | mean_a | mean_h       | mean_w | sum_a | sum_h | sum_w |
|----|-----|------|--------|--------|--------|--------------|--------|-------|-------|-------|
| 1  |     |      | 0      | 10     | 20,6   | 166,3        | 60,5   | 206   | 1663  | 605   |
| 2  |     | Econ | 1      | 3      | 20     | 165          | 61     | 60    | 495   | 183   |
| 3  |     | Math | 1      | 3      | 21     | 162,33333333 | 56     | 63    | 487   | 168   |
| 4  |     | Stat | 1      | 4      | 20,75  | 170,25       | 63,5   | 83    | 681   | 254   |
| 5  | F   |      | 2      | 5      | 18,8   | 162,4        | 56,4   | 94    | 812   | 282   |
| 6  | M   |      | 2      | 5      | 22,4   | 170,2        | 64,6   | 112   | 851   | 323   |
| 7  | F   | Econ | 3      | 1      | 21     | 160          | 59     | 21    | 160   | 59    |
| 8  | F   | Math | 3      | 2      | 17,5   | 155          | 53     | 35    | 310   | 106   |
| 9  | F   | Stat | 3      | 2      | 19     | 171          | 58,5   | 38    | 342   | 117   |
| 10 | M   | Econ | 3      | 2      | 19,5   | 167,5        | 62     | 39    | 335   | 124   |
| 11 | M   | Math | 3      | 1      | 28     | 177          | 62     | 28    | 177   | 62    |
| 12 | M   | Stat | 3      | 2      | 22,5   | 169,5        | 68,5   | 45    | 339   | 137   |



## PROC UNIVARIATE

# PROC UNIVARIATE

## (일변량통계)

```
PROC UNIVARIATE DATA=sas-dataset-name <dataset-options> ;  
  VAR      variablelist ;  
  OUTPUT   OUT=libref.dsn  stat1=var1 statn=varn / options ;  
  CLASS    variablelist ;  
  FREQ     variable ;      WEIGHT variable ;  
  BY       variablelist ;  ID      variablelist ;  
  HISTOGRAM variablelist /options ;  
  PROBLOT  variablelist /options ;  CDFPLOT variablelist /options ;  
  QQPLOT   variablelist /options ;  PPLOT   variablelist /options ;  
RUN;
```

| STATEMENT | 내용                       |
|-----------|--------------------------|
| HISTOGRAM | 줄기잎 그림, 상자그림             |
| PROBPLOT  | 특정 분포에 근거한 확률 그림         |
| CDFPLOT   | 특정 분포에 근거한 누적확률 그림       |
| QQPLOT    | 특정분포에 근거한 분위수-분위수 관계 그림. |
| PPLOT     | 특정분포의 범위와 관측치의 잔차와의 관계   |

# PROC UNIVARIATE

- UNIVARIATE 절차는 숫자 변수들에 대한 여러가지 기술 통계량을 출력해 준다. MEANS 절차는 5개의 통계량을 디폴트로 출력하지만, UNIVARIATE 절차는 기본적으로 적률 등의 통계량과 기본 위치 측도, 위치 모수의 가설검정, 분위수, 변수의 극단값에 대한 정보를 자세히 제공한다.
- NORMAL 등 기타 옵션을 지정하면 정규성 검정이나 상자도표 등 간단한 도표를 얻을 수 있다.
  - 적률에 기초한 다양한 기술 통계량을 계산
  - 극단값, 중위수, 4분위수 등에 대한 상세한 정보를 계산
  - 위치모수나 척도모수에 대해서 절사평균(trimmed mean)과 같은 로버스트(robust) 추정값을 계산
  - 추정값에 대한 신뢰구간 계산
  - 정규성 검정 수행
  - 줄기와 잎 그림, 상자도표 그림 출력

# PROC UNIVARIATE

## ■ PROC UNIVARIATE

```
PROC UNIVARIATE 옵션들;  
  VAR    variables;  
  BY     variables;  
  FREQ   variables;  
  WEIGHT variables;  
  ID     variables;  
  OUTPUT OUT=SAS-dataset keyword=변수명;  
  HISTOGRAM variables/ options;  
  PROBPLOT variables/ options; 확률(백분위) 분포  
  QQPLOT variables/ options; 4분위 분포  
RUN;
```

- ✓ 사용형식은 옵션을 제외하고는 MEANS 절차의 형식과 거의 동일
- ✓ PROC UNIVARIATE는 기술통계량을 구하는 프로시저인 MEANS, SUMMARY, TABULATE에 비해 보다 세부적인 통계량을 구할 수 있음

# PROC UNIVARIATE

## ■ PROC UNIVARIATE 명령어

- ✓ **FREQ variable** : 각 관측치가 FREQ 뒤에 지정되어 있는 변수의 개수를 계산  
FREQ 뒤의 변수의 값이 Missing이거나 1보다 작으면 계산에서 제외되고 정수가 아닌 경우에는 정수부분만을 고려
- ✓ **WEIGHT variable** : 각 관측치에 가중치를 지정하고자 할 때 사용  
이 때 가중치의 값이 0과 같거나 작으면 관측치는 계산에서 제외
- ✓ **OUTPUT OUT=SASdataset Keyword=변수명;**
  - **OUT=SAS-dataset**  
Output을 받아내는 SAS Dataset 이름을 지정
  - **Keyword=변수명**  
Output을 받아내는 SAS Dataset에 보관하고자 하는 Keyword

# PROC UNIVARIATE

## ■ PROC UNIVARIATE 명령어

👉 Keyword= 변수명

N ,      NMISS,      NOBS,      MEAN,      STDMEAN,  
SUM,      STD,      VAR,      CV,      USS,  
CSS,      SKEWNESS,      KURTOSIS,      SUMWGT,      MAX,  
MIN,      RANGE,      Q3,      MEDIAN,      Q1,  
QRANGE,      P1,      P5,      P10,      P90,  
P95,      P99,      MODE,      T,      PROBT,  
MSIGN,      PROBM,      SIGNRANK,      PROBS,      NORMAL,      PROBN,  
등등

# PROC UNIVARIATE

## ■ PROC UNIVARIATE 옵션들

- ✓ DATA=SAS-dataset

PROC UNIVARIATE를 실행하고자 하는 SAS-dataset를 지정

생략시는 가장 최근의 SAS-dataset이 이용

- ✓ NOPRINT

PROC UNIVARIATE의 결과인 기술통계량을 프린트하지 않을 경우에 사용하는 것으로 PROC UNIVARIATE의 목적이 단지 기술통계량을 보관하는 SAS Dataset을 구하고자 하는 경우에만 사용

- ✓ PLOT : stem-and-leaf 그림, box-plot, 정규분포 plot를 그림
- ✓ FREQ : 도수분포, 빈도, 퍼센트, 누적퍼센트를 구함
- ✓ NORMAL : 입력자료가 정규분포를 따르는지에 대한 검정통계량을 구함



## PROC UNIVARIATE\_예제 3.2

```
PROC SORT DATA=htwt;  
BY sex; RUN; /* univariate는 by 사용. Sort 필요. Class는 불필요 */
```

```
PROC UNIVARIATE DATA=htwt NORMAL PLOTS;  
  BY sex;  
  VAR age height weight;  
  HISTOGRAM age/NORMAL;  
  PROBPLOT height weight/NORMAL;  
  QQPLOT age/EXPONENTIAL;  
RUN;
```

```
/* probplot 확률(백분위)분포  
   qqplot 4분위분포/ 지수분포그래프와 비교  
*/
```

SAS 시스템

UNIVARIATE 프로시저  
변수: age

sex=F

| 적률    |            |           |            |
|-------|------------|-----------|------------|
| N     | 5          | 가중합       | 5          |
| 평균    | 18,8       | 관측값 합     | 94         |
| 표준 편차 | 2,77488739 | 분산        | 7,7        |
| 왜도    | -0,0093604 | 첨도        | -2,70366   |
| 제곱합   | 1798       | 수정 제곱합    | 30,8       |
| 변동계수  | 14,7600393 | 평균의 표준 오차 | 1,24096736 |

| 기본 통계 측도 |          |        |         |
|----------|----------|--------|---------|
| 위치측도     |          | 변이측도   |         |
| 평균       | 18,80000 | 표준 편차  | 2,77489 |
| 중위수      | 19,00000 | 분산     | 7,70000 |
| 최빈값      | 16,00000 | 범위     | 6,00000 |
|          |          | 사분위 범위 | 5,00000 |

정규성검정

표본 2000 이하 : Shapiro-Wilk 통계량

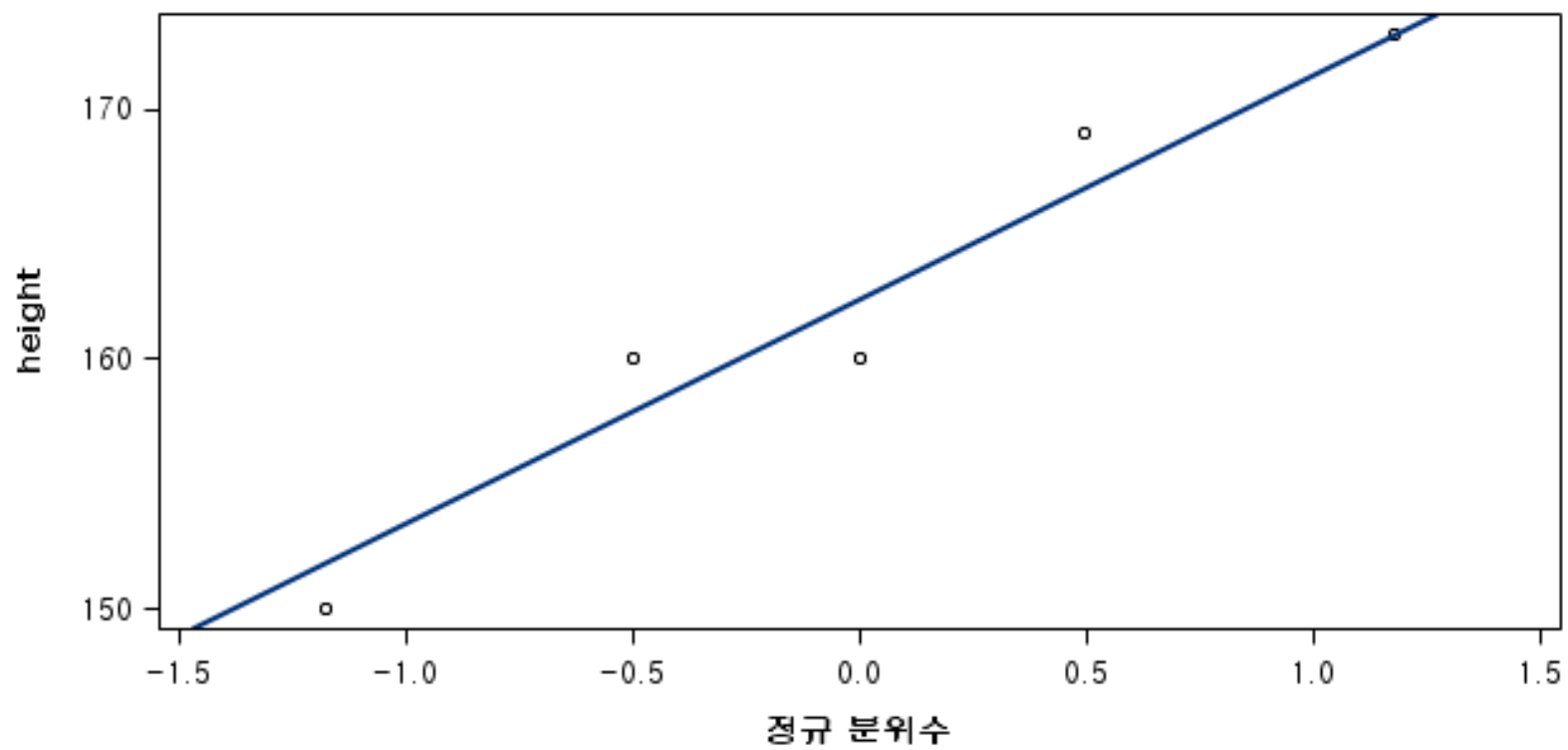
표본 2000 이상 : K-S 통계량

귀무가설 : 분포가 정규분포이다

| 위치모수 검정: $\mu_0=0$ |     |          |             |        |
|--------------------|-----|----------|-------------|--------|
| 검정                 | 통계량 |          | p 값         |        |
| 스튜던트의 t            | t   | 15,14947 | $Pr >  t $  | 0,0001 |
| 부호                 | M   | 2,5      | $Pr >=  M $ | 0,0625 |
| 부호 순위              | S   | 7,5      | $Pr >=  S $ | 0,0625 |

| 정규성 검정             |      |          |             |            |
|--------------------|------|----------|-------------|------------|
| 검정                 | 통계량  |          | p 값         |            |
| Shapiro-Wilk       | W    | 0,876112 | $Pr < W$    | 0,2921     |
| Kolmogorov-Smirnov | D    | 0,243525 | $Pr > D$    | $> 0,1500$ |
| Cramer-von Mises   | W-Sq | 0,049283 | $Pr > W-Sq$ | $> 0,2500$ |
| Anderson-Darling   | A-Sq | 0,336245 | $Pr > A-Sq$ | $> 0,2500$ |

| 분위수(정의 5) |     |
|-----------|-----|
| 레벨        | 분위수 |
| 100% 최댓값  | 22  |
| 99%       | 22  |
| 95%       | 22  |
| 90%       | 22  |
| 75% Q3    | 21  |



# Univariate 예제 : cars dataset

- 다음은 미국산, 유럽산, 일본산 자동차에 관한 변수와 그 내용이다.

| Variable | Description                               | Type | Codes   |
|----------|---|------|---|
| MPG      | Miles per gallon                          | Num  |   |
| ENGINE   | Engine displacement (cu in)               | Num  |   |
| HORSE    | Horsepower                                | Num  |   |
| WEIGHT   | Vehicle weight (lbs.)                     | Num  |   |
| ACCEL    | Time to accelerate from 0 to 60 mph (sec) | Num  |   |
| YEAR     | Model year (modulo 100)                   | Num  | 0 (Missing)<br>70 = 1970<br>71 = 1971<br>...<br>82 = 1982                               |
| ORIGIN   | Country of origin                         | Num  | 1 = American, 2 = European<br>3 = Japanese  |
| CYLINDER | Number of cylinders                       | Num  | 3 = 3 cylinders, 4 = 4 cylinders<br>5 = 5 cylinders, 6 = 6 cylinders<br>8 = 8 cylinders |

- 다음은 cars dataset의 처음 15 subject의 자료이다.

| Obs | mpg  | engine | horse | weight | accel | year | origin | cylinder |
|-----|------|--------|-------|--------|-------|------|--------|----------|
| 1   | 18.0 | 307.0  | 130   | 3504   | 12.0  | 70   | 1      | 8        |
| 2   | 15.0 | 350.0  | 165   | 3693   | 11.5  | 70   | 1      | 8        |
| 3   | 18.0 | 318.0  | 150   | 3436   | 11.0  | 70   | 1      | 8        |
| 4   | 16.0 | 304.0  | 150   | 3433   | 12.0  | 70   | 1      | 8        |
| 5   | 17.0 | 302.0  | 140   | 3449   | 10.5  | 70   | 1      | 8        |
| 6   | 15.0 | 429.0  | 198   | 4341   | 10.0  | 70   | 1      | 8        |
| 7   | 14.0 | 454.0  | 220   | 4354   | 9.0   | 70   | 1      | 8        |
| 8   | 14.0 | 440.0  | 215   | 4312   | 8.5   | 70   | 1      | 8        |
| 9   | 14.0 | 455.0  | 225   | 4425   | 10.0  | 70   | 1      | 8        |
| 10  | 15.0 | 390.0  | 190   | 3850   | 8.5   | 70   | 1      | 8        |
| 11  | .    | 133.0  | 115   | 3090   | 17.5  | 70   | 2      | 4        |
| 12  | .    | 350.0  | 165   | 4142   | 11.5  | 70   | 1      | 8        |
| 13  | .    | 351.0  | 153   | 4034   | 11.0  | 70   | 1      | 8        |
| 14  | .    | 383.0  | 175   | 4166   | 10.5  | 70   | 1      | 8        |
| 15  | .    | 360.0  | 175   | 3850   | 11.0  | 70   | 1      | 8        |

# UNIVARIATE 프로시저의 사용

# CARS dataset에서 차의 중량에 관해 자료를 요약해 보자.

```
* SAS PROGRAM;  
TITLE 'Use of PROC UNIVARIATE';  
PROC UNIVARIATE DATA=tmp1.cars NORMAL PLOT;  
  VAR weight;  
  HISTOGRAM weight;  
RUN;
```

# UNIVARIATE 프로시저의 사용

| Moments                |            |                         |            |
|------------------------|------------|-------------------------|------------|
| <b>N</b>               | 406        | <b>Sum Weights</b>      | 406        |
| <b>Mean</b>            | 2969.56158 | <b>Sum Observations</b> | 1205642    |
| <b>Std Deviation</b>   | 849.827166 | <b>Variance</b>         | 722206.212 |
| <b>Skewness</b>        | 0.46795831 | <b>Kurtosis</b>         | -0.7516387 |
| <b>Uncorrected SS</b>  | 3872721674 | <b>Corrected SS</b>     | 292493516  |
| <b>Coeff Variation</b> | 28.6179338 | <b>Std Error Mean</b>   | 42.1762141 |

| Tests for Location: Mu0=0 |           |          |                     |        |
|---------------------------|-----------|----------|---------------------|--------|
| Test                      | Statistic |          | p Value             |        |
| <b>Student's t</b>        | <b>t</b>  | 70.40844 | <b>Pr &gt;  t </b>  | <.0001 |
| <b>Sign</b>               | <b>M</b>  | 203      | <b>Pr &gt;=  M </b> | <.0001 |
| <b>Signed Rank</b>        | <b>S</b>  | 41310.5  | <b>Pr &gt;=  S </b> | <.0001 |

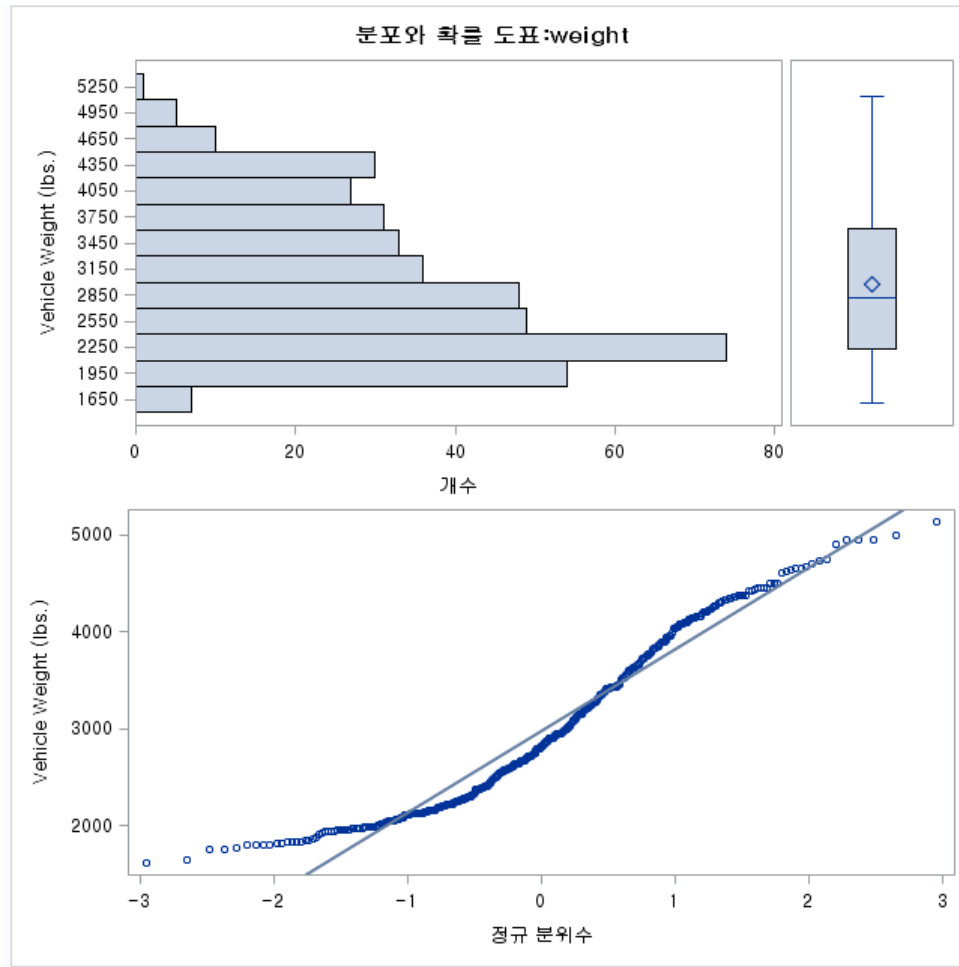
| Basic Statistical Measures |          |                            |           |
|----------------------------|----------|----------------------------|-----------|
| Location                   |          | Variability                |           |
| <b>Mean</b>                | 2969.562 | <b>Std Deviation</b>       | 849.82717 |
| <b>Median</b>              | 2811.000 | <b>Variance</b>            | 722206    |
| <b>Mode</b>                | 1985.000 | <b>Range</b>               | 4408      |
|                            |          | <b>Interquartile Range</b> | 1390      |

| Tests for Normality       |             |          |                     |         |
|---------------------------|-------------|----------|---------------------|---------|
| Test                      | Statistic   |          | p Value             |         |
| <b>Shapiro-Wilk</b>       | <b>W</b>    | 0.949775 | <b>Pr &lt; W</b>    | <0.0001 |
| <b>Kolmogorov-Smirnov</b> | <b>D</b>    | 0.090456 | <b>Pr &gt; D</b>    | <0.0100 |
| <b>Cramer-von Mises</b>   | <b>W-Sq</b> | 1.064994 | <b>Pr &gt; W-Sq</b> | <0.0050 |
| <b>Anderson-Darling</b>   | <b>A-Sq</b> | 6.835844 | <b>Pr &gt; A-Sq</b> | <0.0050 |

# UNIVARIATE 프로시저의 사용

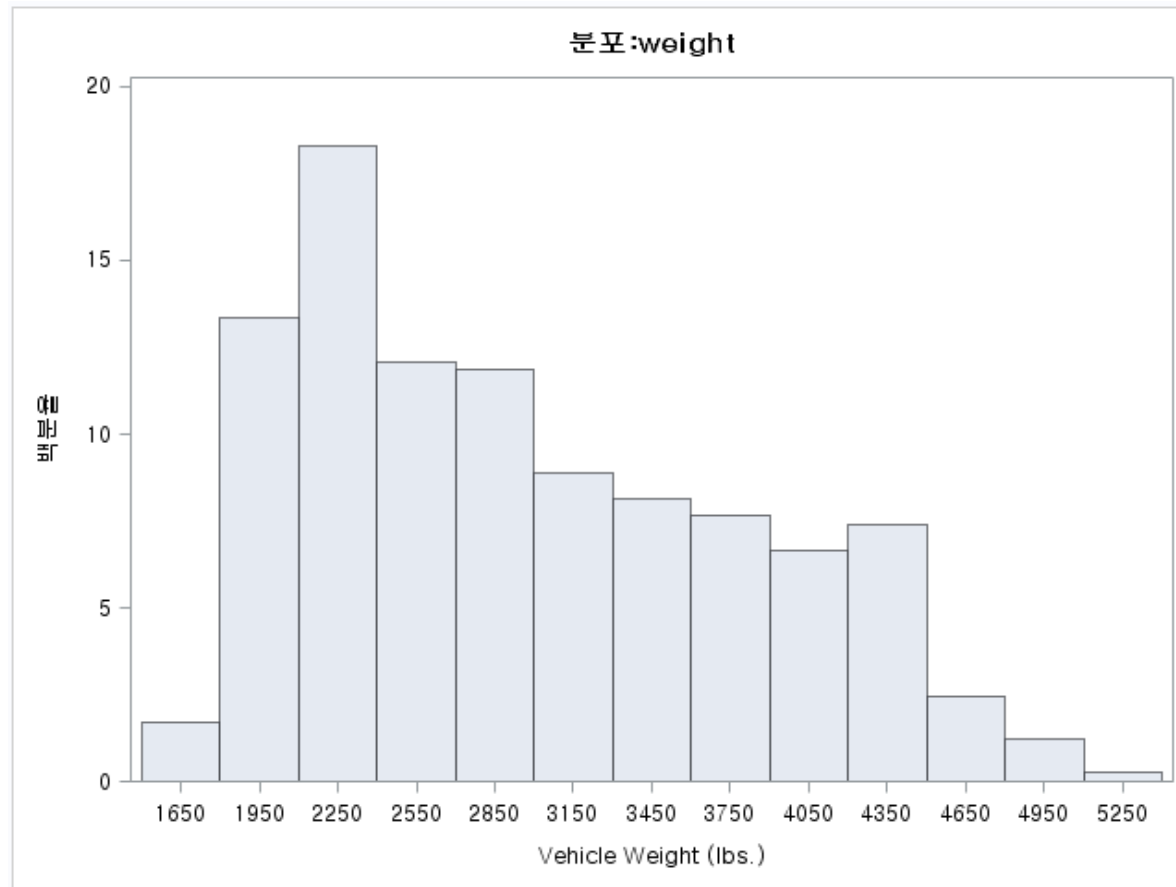
| Quantiles (Definition 5) |          |
|--------------------------|----------|
| Quantile                 | Estimate |
| 100% Max                 | 5140     |
| 99%                      | 4951     |
| 95%                      | 4457     |
| 90%                      | 4257     |
| 75% Q3                   | 3613     |
| 50% Median               | 2815     |
| 25% Q1                   | 2226     |
| 10%                      | 1985     |
| 5%                       | 1925     |
| 1%                       | 1773     |
| 0% Min                   | 1613     |

| Extreme Observations |     |         |     |
|----------------------|-----|---------|-----|
| Lowest               |     | Highest |     |
| Value                | Obs | Value   | Obs |
| 1613                 | 332 | 4951    | 75  |
| 1649                 | 344 | 4952    | 70  |
| 1755                 | 385 | 4955    | 38  |
| 1760                 | 386 | 4997    | 82  |
| 1773                 | 331 | 5140    | 40  |

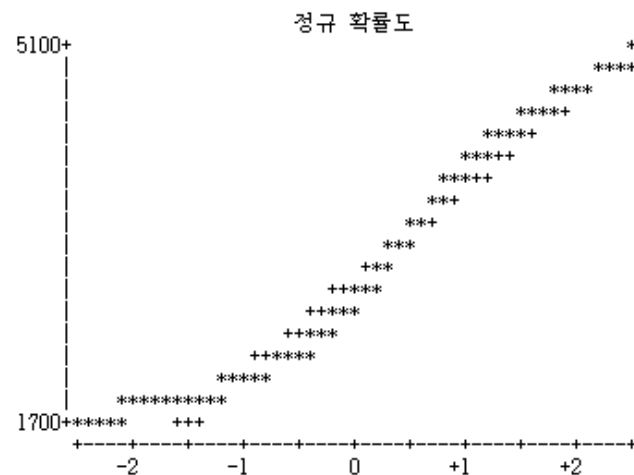
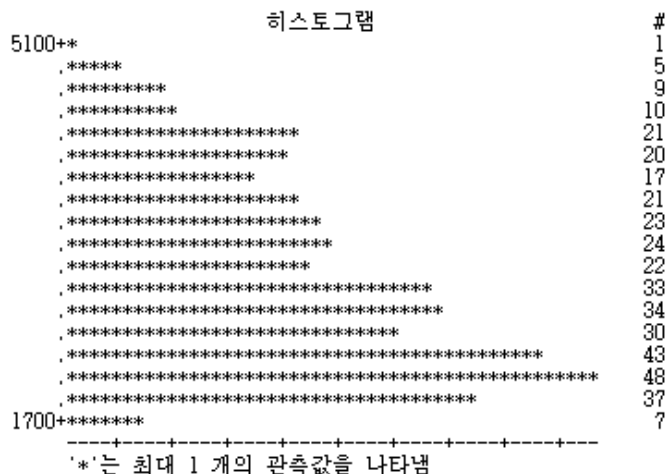




# UNIVARIATE 프로시저의 사용

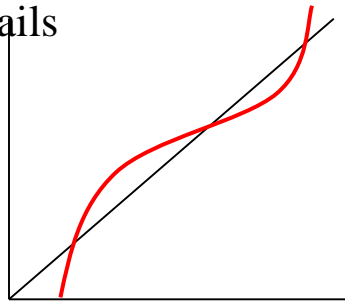


# UNIVARIATE 프로시저의 사용

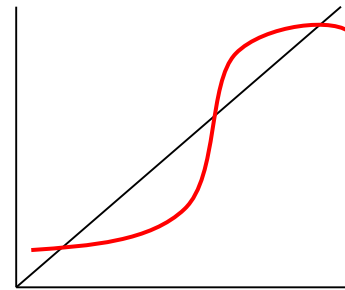


- ❖ + in the plot: Normal
- ❖ \* in the plot: Sample

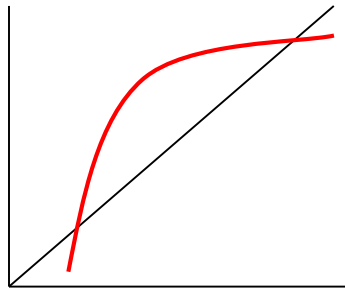
Heavy  
tails



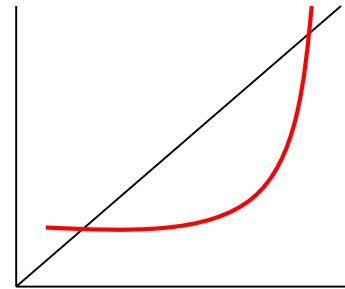
Light tails



Skewed to the left



Skewed to the right



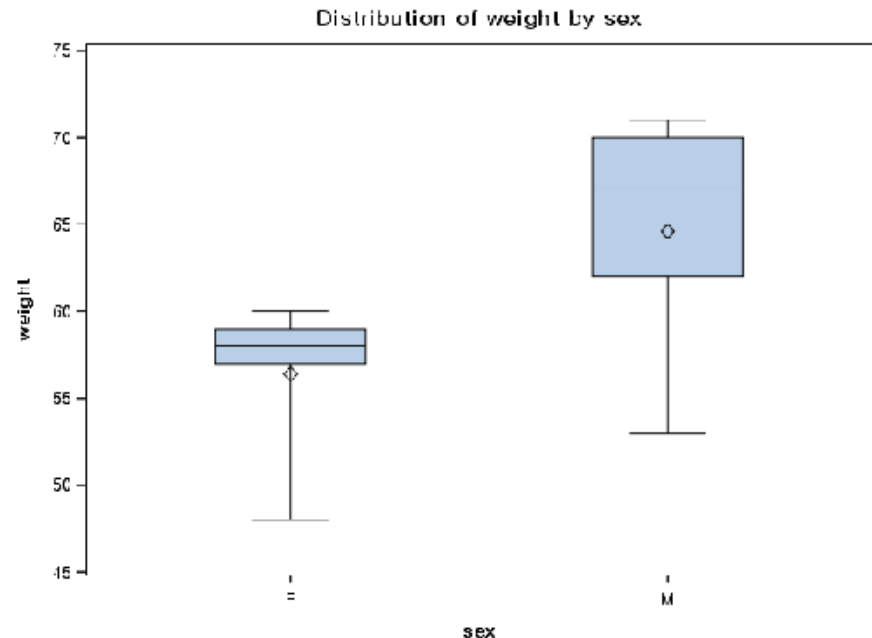
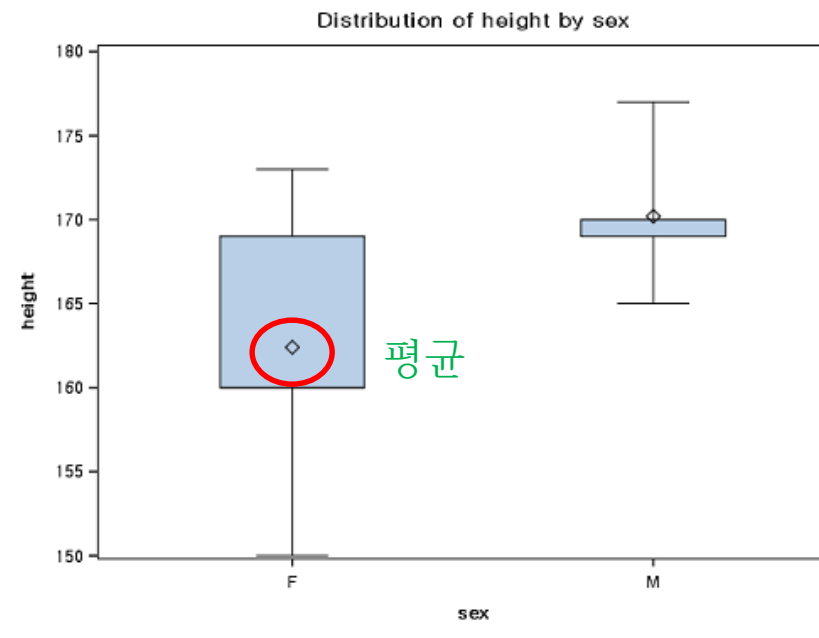
# BOXPLOT 프로시저에 의한 상자그림\_예제 3.3

```
PROC SORT DATA=htwt;
```

```
BY sex; RUN;
```

```
PROC BOXPLOT DATA=htwt;
```

```
  PLOT ( height weight)*sex / BOXSTYLE=SKELETAL; /* skeletal : 최대값과 최소값을 이용 */  
  RUN;                                           /* schematic : 1.5배 사분위범위에 포함되는 값만 이용*/
```



1. 다음은 어느 대학에서 개설된 교양통계학 강좌를 수강한 학생들의 학년별 성적자료이다. 이 자료에 대해서 성적의 변수명은 score, 학년의 변수명은 year(L=1학년, H=2,3,4학년) 로 지정하여 SAS파일 ex1\_1을 생성하고, 다음 문제를 해결하시오.

| 학년 | 성적   |
|----|--|
| L  | 88 67 77 75 84 90 95 73 69 80 79 82 95 62    |
| H  | 82 86 88 92 77 72 96 75 68 96 85 86 76 89 83 |

- 1) 이 성적자료에 대한 기술통계량을 구하시오
- 2) 줄기-잎 그림과 상자그림을 작성하시오
- 3) 위 자료가 정규분포를 따르는지 설명하시오
- 4) MEANS 프로시저를 이용하여 학년별 평균, 표준편차, 변이계수, 최대값, 최소값을 구하시오

# 실습-coding

```
DATA ex1_1;  
  INPUT year $ score @@;  
  CARDS;  
  L 88 L 67 L 77 L 75 L 84 L 90 L 95 L 73 L 69 L 80 L 79 L 82  
  L 95 L 62 H 82 H 86 H 88 H 92 H 77 H 72 H 96 H 75 H 68 H 96  
  H 85 H 86 H 76 H 89 H 83  
;RUN;  
  
/* ods graphics off; */  
PROC UNIVARIATE DATA=ex1_1 NORMAL PLOT;  
  VAR score;  
  HISTOGRAM score/NORMAL;  
  RUN;  
PROC MEANS DATA=ex1_1 mean std cv max min;  
  CLASS year;  
  VAR score;  
  RUN;
```



## PROC FREQ

범주형 자료로 구성된 분할표를  
작성하거나 교차분석을 수행하기  
위해서 FREQ 절차를 수행한다.

# 범주형 자료의 분석 – 카이제곱 검정

- ❖ 범주형 자료의 분석
  - 하나인 경우: 도수 분포표, 막대 그래프를 이용
  - 두개인 경우: 분할표를 이용 ( $\chi^2$  검정) : 교차분석

- ❖ 2차원 분할표 (contingency table)의 표현 방법

| Row Levels \ Column Levels | 1                        | ... | J                        | Total                    |
|----------------------------|--------------------------|-----|--------------------------|--------------------------|
| 1                          | $n_{11}$                 | ... | $n_{1J}$                 | $\sum_j n_{1j} = n_{1.}$ |
| ...                        | ...                      | ... | ...                      | ...                      |
| I                          | $n_{I1}$                 | ... | $n_{IJ}$                 | $\sum_j n_{Ij} = n_{I.}$ |
| Total                      | $\sum_i n_{i1} = n_{.1}$ | ... | $\sum_i n_{iJ} = n_{.J}$ | $n$                      |



# 테이블

| Food Type    | Pesticide Status |              | Total         |
|--------------|------------------|--------------|---------------|
|              | Present          | Not Present  |               |
| Organic      | 29               | 98           | 127           |
| Conventional | 19,485           | 7,086        | 26,571        |
| <b>Total</b> | <b>19,514</b>    | <b>7,184</b> | <b>26,698</b> |

| Wine         | Music     |           |           | Total      |
|--------------|-----------|-----------|-----------|------------|
|              | None      | French    | Italian   |            |
| French       | 30        | 39        | 30        | 99         |
| Italian      | 11        | 1         | 19        | 31         |
| Other        | 43        | 35        | 35        | 113        |
| <b>Total</b> | <b>84</b> | <b>75</b> | <b>84</b> | <b>243</b> |

# 카이제곱 검정

- ❖  $H_0$ : 두 변수 사이에 연관성이 없다 vs.  $H_1$ : 두 변수 사이에 연관성이 있다.
- ❖ 표본비율에 나타나는 차이가 실제 차이에 의한 것인지 랜덤추출에서 나오는 우연한 결과인지를 보고자 함.
- ❖ 표본에서 관측된 실제빈도와 귀무가설이 참일 때의(즉, 연관성이 없을 때) 기대빈도를 비교해서 그 차이가 크면 귀무가설을 기각.

- ❖ 귀무가설이 참일 때, 각 cell의 기대빈도는

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

|     | 1        | ... | J        |          |
|-----|----------|-----|----------|----------|
| 1   | $n_{11}$ | ... | $n_{1J}$ | $n_{1.}$ |
| ... | ...      | ... | ...      | ...      |
| I   | $n_{I1}$ | ... | $n_{IJ}$ | $n_{I.}$ |
|     | $n_{.1}$ | ... | $n_{.J}$ | $n$      |

# 동일성 검정과 독립성 검정

카이제곱 검정은 두 범주형 자료의 연관성을 검정하는 것.

- ❖ 동일성 검정: 몇 가지 다른 분포를 비교. 몇 개의 랜덤표본들이 뽑힌 모집단의 분포를 비교하게 됨. (실험계획에 따른 연구, experimental study)
- ❖ 독립성 검정: 하나의 랜덤표본을 뽑아서 두 개의 범주형 변수에 따라 분류한 후 두 변수가 독립인지를 검정. (관측연구, observational study)

|     | 1        | ... | J        |          |
|-----|----------|-----|----------|----------|
| 1   | $n_{11}$ | ... | $n_{1J}$ | $n_{1.}$ |
| ... | ...      | ... | ...      | ...      |
| I   | $n_{I1}$ | ... | $n_{IJ}$ | $n_{I.}$ |
|     | $n_{.1}$ | ... | $n_{.J}$ | $n$      |

# 카이제곱 검정

❖ 독립성 검정 -- Pearson  $\chi^2$  검정 (동일성검정의 경우에도 검정절차는 같음)

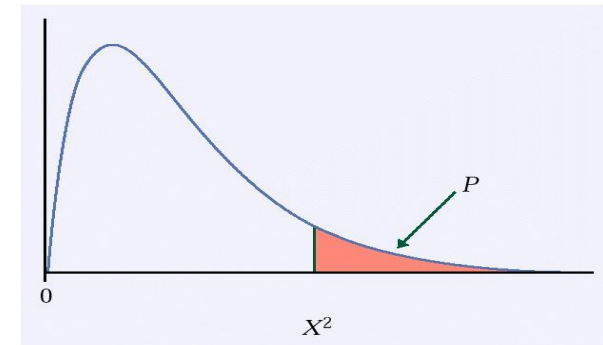
❖  $H_0$ : 행과 열은 독립이다 (즉, 행과 열 사이에 관련이 없다.)

❖  $H_1$ : 행과 열은 독립이 아니다. (즉, 행과 열 사이에 관련이 있다.)

❖  $n_{ij}$  = 관찰빈도,  $m_{ij}$  = 기대빈도 
$$m_{ij} = n \left( \frac{n_{i \cdot} n_{\cdot j}}{n^2} \right) = \frac{n_{i \cdot} n_{\cdot j}}{n}$$

- $\chi^2$  검정통계량: 다음 통계량은 귀무가설 하에서 근사적으로 자유도  $(I-1)(J-1)$ 을 가진  $\chi^2$  분포를 따른다. (각 셀의 기대빈도 > 5)

$$\begin{aligned} Q &= \sum_{i=1}^I \sum_{j=1}^J \frac{(\text{관찰도수}_{ij} - \text{기대도수}_{ij})^2}{\text{기대도수}_{ij}} \\ &= \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - m_{ij})^2}{m_{ij}} \stackrel{\text{Approx.}}{\sim} \chi^2_{(I-1)(J-1)} \end{aligned}$$



# 카이제곱 검정

❖ 예: 호흡기 질환에 관한 두 개의 처치법 (Treatment 과 Placebo)을 비교하는 무작위 임상 실험으로부터 얻은 자료가 옆  $2 \times 2$  표에 정리되어 있다.

| Outcome<br>Treatment | Favorable<br>(=1) | Unfavorable<br>(=0) | Total |
|----------------------|-------------------|---------------------|-------|
| Placebo              | 4                 | 12                  | 16    |
| Treatment            | 10                | 5                   | 15    |
| Total                | 14                | 17                  | 31    |

- $H_0$ : 처치법과 그에 따른 결과가 관련이 없다.
- $H_1$ : 처치법과 그에 따른 결과가 관련이 있다.
- $\chi^2$  검정결과 유의수준 0.05에서 귀무가설 기각. 즉, 처치법과 그에 따른 결과가 관련이 있으며 Treatment가 Placebo보다 favorable result의 비율이 높다고 할 수 있다.

$$Q = 5.43 > \chi_{1,0.05}^2 = 3.84 \quad p\text{-value} = 0.0198$$

# PROC FREQ

## (덧수분포표, 분할표)

```
PROC FREQ DATA=sas-dataset-name <dataset-options > ;
  TABLES variable-req ;
  EXACT statistic-keyword /options ;
  TEST statistic-keyword /options ;
  OUTPUT OUT=libref.dsn stat1=var1 statn=varn / options ;
  WEIGHT variable ;
  BY variablelist ;
RUN;
```

| 구분             |          | 내용  |
|----------------|----------|---|
| STATEMENT      | TABLES   | 분할표를 작성할 범주형 변수 지정 a , a b , a*b , a*(b c)          |
|                | EXACT    | 정확성 검정 방법 (AGREE, CHISQ, KENTB, KAPPA, FISHER, ...) |
|                | TEST     | 동일성 검정 방법 (AGREE, GAMMA, KAPPA, KENTB, PCORR,... )  |
| TABLES OPTIONS |          | 검정(AGREE, CHISQ) 출력제어 (NOCOL, NOROW)                |
| OPTIONS        | ORDER=   | 출력순위 (DATA, FORMATTED, FREQ, INTERNAL)              |
|                | NLEVELS  | 분할 갯수   |
|                | PAGE     | 한페이지 한 개의 테이블 출력                                    |
|                | COMPRESS | 여백이 충분하면 다음 테이블도 같은 페이지에 출력                         |

# PROC FREQ

## ■ PROC FREQ

```
PROC FREQ DATA=SAS-data-set <옵션(들)>;  
  BY 변수(들);  
  OUTPUT OUT=SAS-data-set 옵션(들);  
  TABLES 변수(들)< /옵션(들)>; 범주형 변수 입력  
  WEIGHT 변수/옵션; FREQ 절차에 포함되는 자료의 가중치변수 지정  
RUN; WEIGHT 문이 생략되면 관측치의 도수는 1
```

- ✓ 빈도(및 백분율) 생성함
- ✓ 도수분포표나 분할표를 만들어 주는 프로시저이며, 더불어 변수값들의 분포와 연관도에 관한 정보를 요약해 줌
- ✓ 특히 카이제곱 검정과 같이 독립성 및 적합도 검정에 유용하게 쓰임
- ✓ 결과를 리포트로 생성(자동)하며 SAS Data set으로도 저장 가능함
- ✓ 해당 변수에 결측값이 있을 경우 리포트 하단에 결측값이 몇 개인지를 표시함
- ✓ 1차 또는 2차 빈도 분석이 가능
- ✓ 주로 사용되는 변수는 명목형, 서열형 자료와 같은 이산형(discrete) 자료에 대응되는 변수

## 참고 : 독립성 검정

- 독립성검정은 자료에 포함된 두 가지 특성 사이에 어떠한 연관관계가 있는지를 검정하는 방법으로 독립성 검정에서의 귀무가설은 두 가지 특성들이 서로 독립이라는 것이다. 즉 두 특성 사이에 연관관계가 없다는 것이다.
- 독립성 검정을 위해서 카이제곱 검정이 사용되며, 이를 위한 검정통계량값과 유의 확률 값은 FREQ 프로시저를 이용하여 얻을 수 있다.
- 분석결과 두 특성이 독립이 아닌 경우 어느 정도 연관성을 가지고 있는지를 분석하기 위해 필요한 기본적인 연관성 측도(measure of association)로 파이계수, 분할계수, 크라머의 V를 제공한다.

### ★ 연관성 측도(measures 옵션)

1) 순서형 자료 : 감마, 켄달의 타우, 스튜어트의 타우 등

\*\* -1에서 1사이의 값을 가지며 절대값이 클수록 연관성이 크다.

2) 명목형 자료 : 파이계수, 크래머의 V, 분할계수, 람다 등

\*\* 일반적으로 0과 1사이에 존재하며 이 값이 클수록 연관성이 크다.



# PROC FREQ

```
proc freq data=orion.sales;  
run;
```

data set의 모든  
변수 각각에  
대해 일원빈도  
생성

```
proc freq data=orion.sales;  
  tables Gender Country;  
run;
```

Gender와  
Country 각각의  
일원빈도 생성

```
proc freq data=orion.sales;  
  tables Gender* Country;  
run;
```

rows

columns

Gender와  
Country에 대한  
이원빈도 생성

# PROC FREQ

## ■ PROC FREQ와 함께 쓰이는 명령어

- ✓ BY 변수명

- ✓ TABLES 변수명 : 인쇄할 도수분포표 및 분할표의 양식을 지정

원하는 대로 많은 TABLES문을 쓸 수 있음

예 1) tables a; ..... a 라는 변수의 1차원 도수표

예 2) tables a\*b; ..... a와 b의 2차원 도수분할표

예 3) tables (a b) \* C; ..... a\*c , b\*c 의 각각 분할표




예 4) tables a\*b\*c; ..... 3차원 분할표 출력

여기서 맨 처음 명시한 a라는 변수를 조절변수(control variable) 라고 하는데 이 변수가 가지는 값에 따라 나머지 두 변수의 분할표를 만들어 줌

예 5) tables x1-x30; .... x1 에서 x30 까지의 1차원 도수표

# PROC FREQ

## ■ PROC FREQ와 함께 쓰이는 명령어

- ✓ WEIGHT 변수명 :  일반적으로 하나의 관찰값은 도수계산에서 한 개로 처리되는 반면에 weight문을 쓰면 가중값을 나타내는 변수를 사용하여 도수를 계산하도록 지정하는 명령문
-  만일 분할표를 그대로 입력하여 카이제곱 검정을 해야 한다면 그 cell의 숫자는 그 수준에서 관찰값의 개수를 의미하므로 반드시 weight문을 써야 함
-  만일 한 변수가 가지고 있는 수준이 a라고 할 때 이 관측값이 20개가 있다면 가중치를 나타내기 위하여 weight문에 이를 SAS에 알려주는 변수를 정의하여야 함

# PROC FREQ

## ■ 표시할 통계량 제어 옵션들

TABLES 변수(들)< /옵션(들)>;

| Option         | 내용                      |
|----------------|-------------------------|
| NOCUM          | 누적 빈도와 누적 백분율 표시 안함     |
| NOPERCENT      | 백분율, 누적백분율, 총 백분율 표시 안함 |
| NOFREQ         | 셀 빈도와 총 빈도 표시 안함        |
| NOROW          | 행 백분율 표시 안함             |
| NOCOL          | 칼럼 백분율 표시 안함            |
| Out=data set 명 | 변수 값과 빈도를 data set으로 생성 |

# PROC FREQ

## ■ 표시할 통계량 제어 옵션들(기타 내용)

TABLES 변수(들) < /옵션(들) >;

| Option    | 내용                                     |
|-----------|--|
| EXPECTED  | 각 cell의 기대 도수(expected frequency)      |
| DEVIATION | (관찰도수-기대도수)의 절대값                       |
| CHISQ     | 각 2차원 도수분포표에서 얻어진 전체 카이제곱 통계량과 유의확률 표시 |
| ALL       | 연관 관계의 모든 척도(또는 통계량)를 인쇄               |
| LIST      | 도수분포표를 분할표 형식이 아니라 List 형태로 인쇄         |
| MISSING   | 도수에 관한 통계량의 계산에서 결측값의 개수를 포함           |
| CELLCHI2  | 각 cell에 대한 카이제곱 통계량                    |

# PROC FREQ

## ■ PROC FREQ의 옵션들

| Option  | 내용   |
|---------|--|
| NLEVELS | TABLES 문장에 기술된 각 변수의 유일값의 레벨수(개수)를 제공하는 테이블을 표시  |
| ORDER=  | 변수 값을 표시하는 순서(order)를 지정<br>예) order = freq : 빈도의 내림차순<br>예) order = data : 데이터 셋에 나타나는 범주의 순서대로 분할표를 출력 |

# PROC FREQ 예제 3.4

```
PROC FREQ data=htwt;
  TABLES dept sex*dept;
RUN;
```

SAS 시스템

FREQ 프로시저

| dept | 빈도 | 백분율   | 누적 빈도 | 누적 백분율 |
|------|----|-------|-------|--------|
| Econ | 3  | 30,00 | 3     | 30,00  |
| Math | 3  | 30,00 | 6     | 60,00  |
| Stat | 4  | 40,00 | 10    | 100,00 |

빈도  
백분율  
행 백분율  
칼럼 백분율

| 테이블: sex * dept |       |       |       |        |
|-----------------|-------|-------|-------|--------|
| sex             | dept  |       |       |        |
|                 | Econ  | Math  | Stat  | 합계     |
| F               | 1     | 2     | 2     | 5      |
|                 | 10,00 | 20,00 | 20,00 | 50,00  |
|                 | 20,00 | 40,00 | 40,00 |        |
|                 | 33,33 | 66,67 | 50,00 |        |
| M               | 2     | 1     | 2     | 5      |
|                 | 20,00 | 10,00 | 20,00 | 50,00  |
|                 | 40,00 | 20,00 | 40,00 |        |
|                 | 66,67 | 33,33 | 50,00 |        |
| 합계              | 3     | 3     | 4     | 10     |
|                 | 30,00 | 30,00 | 40,00 | 100,00 |

# PROC FREQ : 연령별 음료수 선호도 예제 3.5

[표] Drink 데이터셋 (총 108명의 음료수 선호도 조사)

| age | drink | Count |
|-----|-------|-------|
| 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    |



# PROC FREQ : 연령별 음료수 선호도 예제 3.5

/\* WEIGHT 명령문의 사용 \*/

```
DATA 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=drink;
  WEIGHT count;
  TABLES age*drink/ NOCL NOPERCENT CHISQ MEASURES;
RUN;
```

/\* NOCOL : 열 퍼센트를 출력하지 않음 \*/  
/\* NOPERCENT : 각 칸의 퍼센트를 출력하지 않음 \*/  
/\* MEASURES : 연관성의 기본적인 측도가 출력된다 \*/

**\*\* 독립성 검정**

귀무가설 : 연령과 음료수 선호도에는 차이가 없다

# PROC FREQ 예제

FREQ 프로시저

| 빈도<br>행 백분율 | 테이블: age * drink |       |       |       |     |
|-------------|------------------|-------|-------|-------|-----|
|             | drink            |       |       |       | 합계  |
|             | age              | A     | B     | C     |     |
| 18          | 10               | 14    | 2     | 12    | 38  |
|             | 26.32            | 36.84 | 5.26  | 31.58 |     |
| 19          | 13               | 7     | 10    | 8     | 38  |
|             | 34.21            | 18.42 | 26.32 | 21.05 |     |
| 20          | 12               | 4     | 6     | 10    | 32  |
|             | 37.50            | 12.50 | 18.75 | 31.25 |     |
| 합계          | 35               | 25    | 18    | 30    | 108 |

age \* drink 테이블에 대한 통계량

| 통계량                  | 자유도 | 값       | Prob   |
|----------------------|-----|---------|--------|
| 카이제곱                 | 6   | 11.8683 | 0.0650 |
| 우도비 카이제곱             | 6   | 12.5675 | 0.0504 |
| Mantel-Haenszel 카이제곱 | 1   | 0.0015  | 0.9692 |
| 파이 계수                |     | 0.3315  |        |
| 우발성 계수               |     | 0.3147  |        |
| 크래머의 V               |     | 0.2344  |        |

| 통계량           | 값       | ASE    |
|---------------|---------|--------|
| 감마            | -0.0110 | 0.1217 |
| Kendall의 타우-b | -0.0078 | 0.0865 |
| Stuart의 타우-c  | -0.0082 | 0.0908 |

P-value가 0.065 이므로  
유의수준 5%에서  
귀무가설이 기각되지  
않는다. 즉 연령별  
음료수선호도에 차이가  
없다.

# 실습

2. 어떤 회사가 두 종류의 신상품 A, B를 개발하여 소비자들이 어떤 종류의 신상품을 더 좋아하는가를 조사하기 위하여 40명을 임의로 추출하여 다음과 같은 자료를 얻었다. 여기서, M은 남자, F는 여자를 나타내고, 1은 신상품 A를 2는 신상품 B를 나타낸다.

|   |
|---|
| M 2 F 1 M 1 F 2 F 2 M 2 M 1 F 1 M 1 F 2 |
| M 2 M 1 M 1 F 2 F 1 M 1 F 1 M 1 F 2 F 2 |
| M 2 F 2 M 1 M 1 M 1 F 2 F 2 F 2 M 1 F 2 |
| M 2 M 2 F 2 F 1 M 1 M 1 F 1 F 2 M 2 F 2 |

- 1) SAS 파일명은 ex\_2로 하고, 성별은 gender, 선호상품의 변수명은 favor로 하여 자료를 입력하시오.
- 2) 위 자료에 대한 도수분포표를 작성하시오.
- 3) 성별에 따른 선호하는 신상품의 분포를 알아보기 위한 분할표를 작성하시오.



## PROC STANDARD

---

숫자변수들의 전부 또는 일부를 주어진 평균값과 표준편차를 이용하여 표준화시키고, 그 표준화된 값들을 포함하는 새로운 데이터셋을 생성할 때 사용

# STANDARD 프로시저 – 숫자 변수의 표준화 및 표준화 변수 생성

- **사용형식**

```
PROC STANDARD DATA=SAS-data-set option(s);
```

```
  BY variables;
```

```
  FREQ variables;
```

```
  VAR variable(s);
```

```
  WEIGHT variable;
```

```
  RUN;
```

- **PROC STANDARD 예제 8.6**

```
PROC STANDARD DATA=htwt OUT=stand_hw MEAN=0 STD=1;
```

```
  VAR height weight;
```

```
/* OUT옵션을 사용하지 않으면 새로운 자료값으로  
   기존의 데이터셋을 대체 */
```

```
  RUN;
```

## PROC STANDARD 예제

|    | name | sex | dept | age | height       | weight       |
|----|------|-----|------|-----|--------------|--------------|
| 1  | 김철수  | M   | Stat | 25  | 0,4741252424 | 0,8914431257 |
| 2  | 강민호  | M   | Stat | 20  | 0,345983285  | 1,3028784145 |
| 3  | 이영희  | F   | Math | 19  | -0,807294332 | -0,342862741 |
| 4  | 박지수  | F   | Econ | 21  | -0,807294332 | -0,205717644 |
| 5  | 최병호  | M   | Math | 28  | 1,3711189442 | 0,2057176444 |
| 6  | 장순미  | F   | Stat | 22  | 0,8585511146 | -0,068572548 |
| 7  | 이상호  | M   | Econ | 19  | 0,4741252424 | 1,4400235108 |
| 8  | 김미숙  | F   | Math | 16  | -2,088713906 | -1,714313703 |
| 9  | 박홍식  | M   | Econ | 20  | -0,166584545 | -1,028588222 |
| 10 | 유은영  | F   | Stat | 16  | 0,345983285  | -0,480007837 |

## PROC STANDARD 예제 8.7 (REPLACE 옵션의 사용)

```
PROC STANDARD DATA=htwt OUT=stand_hw REPLACE;  
  VAR height weight;  
RUN;
```

**REPLACE** : 결측값에 대해서 그 변수의 평균값으로 대체된다.

(\*\* REPLACE MEAN=옵션 : 결측값은 이 옵션에 주어진 값으로 대체)

# PROC PLOT (산점도)

```
PROC PLOT DATA=sas-dataset-name <dataset-options> ;
  PLOT variable-req ;
  BY variablelist ;
RUN;
```

| 구분           |           | 내용   |
|--------------|-----------|--|
| STATEMENT    | PLOT      | PLOT을 그릴 수평/수직축 변수 지정 $y*x='Symbol'$   |
|              | BY        | 정확성 검정 방법 (AGREE, CHISQ, KAPPA, FISHER, ...)   |
|              | TEST      | 동일성 검정 방법 (AGREE, GAMMA, KAPPA, PCORR,... )  |
| PLOT OPTIONS |           | HAXIS=n to m by k 수평축의 구간과 길이 지정<br>VAXIS=n to r by q 수직축의 구간과 길이 지정<br>OVERLAY 동일한 PLOT 명령문내의 플롯들을 합쳐서 표현 |
| OPTIONS      | VPERCENT= | PLOT의 배치비율 ( 50 , 50 25 25 , 300 )   |
|              | HPERCENT= | PLOT의 배치비율   |
|              | PAGE      | 한페이지 한 개의 테이블 출력   |
|              | COMPRESS  | 여백이 충분하면 다음 테이블도 같은 페이지에 출력  |



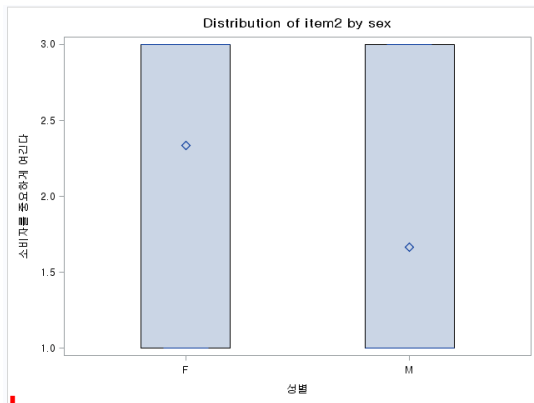
# PROC BOXPLOT

```
PROC BOXPLOT DATA=sas-dataset-name <dataset-options> ;  
  PLOT variable-req /options ;  
  INSET ;  
  INSETGROUP ;  
  BY variablelist ;  
  ID variable ;  
RUN;
```

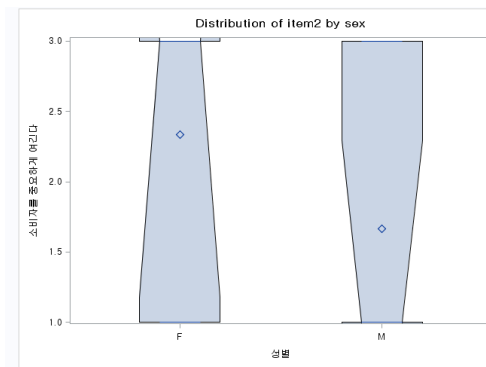
| 구분           |           | 내용                            |
|--------------|-----------|-------------------------------|
| STATEMENT    | PLOT      | PLOT을 그릴 수평/수직축 변수 지정 a*b='H' |
|              | INSET     | PLOT위에 표시하는 POI               |
| PLOT OPTIONS | BOXSTYLE= | SKELETAL SCHEMATIC            |
|              | NOTCHES   |                               |

# PROC BOXPLOT

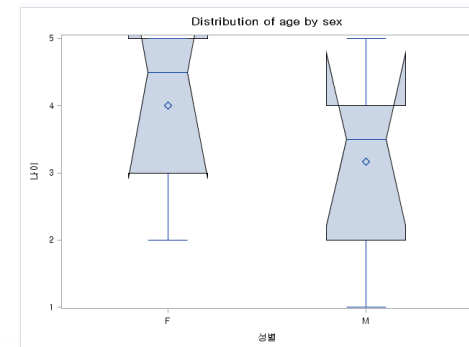
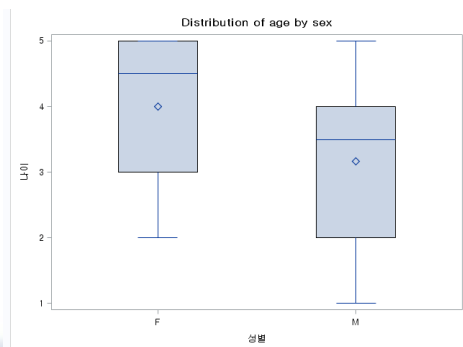
```
PROC BOXPLOT DATA=company1 ;  
  PLOT ( item2 age)*sex /BOXSTYLE=SCHEMATIC ;  
  PLOT ( item2 age)*sex / NOCHES ;  
RUN;
```



극한값 표시



Folding Effect in Small Size



# PROC CHART

## (차트/도표)

```
PROC CHART DATA=sas-dataset-name <dataset-options> ;
  BLOCK variablelist /options ;
  HBAR variablelist /options ;
  VBAR variablelist /options ;
  PIE variablelist /options ;
  STAR variablelist /options ;
  BY variablelist ;
```

RUN;

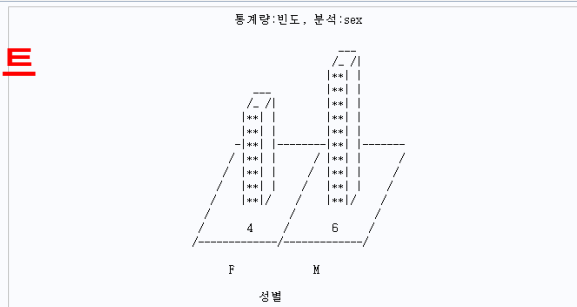
| 구분              | 내용  |
|-----------------|---|
| STATEMENT       | BLOCK 블록 차트                                   |
|                 | HBAR 수평바 차트                                   |
|                 | VBAR 수직바 차트                                   |
|                 | PIE 파이차트                                      |
|                 | STAR 스타 차트                                    |
| Chart<br>OPTION | AXIS= n TO m BY r n 에서 m 까지 r 마다 (HBAR, VBAR) |
|                 | GROUP= 분류(BY)                                 |
|                 | SUMVAR= 집계할 값                                 |
|                 | TYPE= 평균, 합계, 비율 등의 표시할 값 (default는 FREQ)     |
| OPTIONS         | LPI= (Line길이/coloumn길이) * 10 (default 는 6 )   |

# PROC CHART

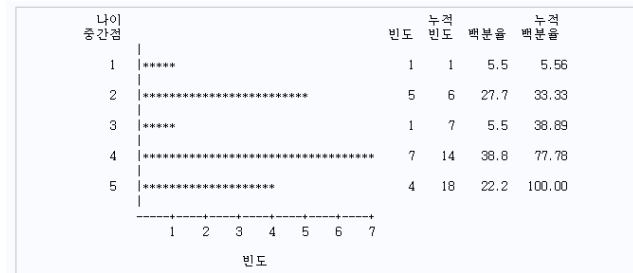
## (차트/도표)

```
PROC CHART DATA=company1 LPI=3 ;
  BLOCK SEX ;
  HBAR AGE / FREQ=item1 ;
  VBAR AGE / GROUP=sex ;
  PIE AGE / SUMVAR=item1 TYPE=MEAN ;
RUN;
```

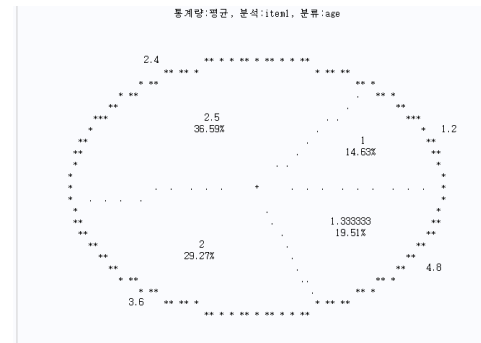
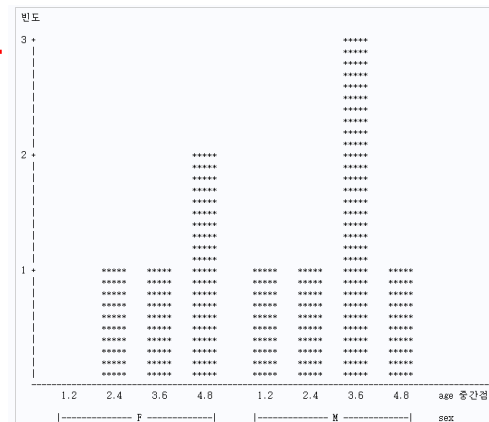
블록차트



수평바차트  
덧수가중



수직바차트  
그룹



파이차트  
LTI비율 축소