Chapter 3. 데이터 요약 및 표현

❖ 프로시저 단계

- 이미 생성되어 있는 데이터 셋에 대하여 필요한 연산과 분석을 수행
- 분석된 결과를 인쇄
- 특별한 형태의 새로운 데이터 셋을 생성

❖ PROC 명령문

PROC 명령문
PROC procedure-name DATA=SAS-data-set options;

- DATA=옵션이 생략되면 가장 마지막으로 생성되거나 사용된 데이터 셋이 처리대상이 됨

기초통계량의 계산

❖ MEANS 프로시저

- 숫자변수에 대한 기본적인 일변량 기술통계량을 제공
- 계산된 여러 통계량을 포함하는 새로운 데이터 셋을 생성, 이를 다음 분석의 입력데이터로 사용

MEANS 프로시저에 의한 기술통계량

```
PROC MEANS DATA=mysas.htwt MAXDEC=2 MEAN STD CV RANGE;
CLASS gender;
VAR age height weight;
TITLE '>>>>>>성별 기초 통계량<<<<<';
RUN;
```

- MAXDEC=: 분석결과를 인쇄할 때 사용될 최대 소수점 자릿수 지정, 디폴트=8
- 옵션에 어떤 통계량도 지정하지 않으면 N, MEAN, STD, MIN, MAX가 출력
- CLASS 명령문: 변수의 자료값에 따른 개별 통계량을 구하고자 하는 변수를 지정
- VAR 명령문: 분석에 사용할 숫자변수들을 지정하는 것, 대부분의 분석 프로시저에서 사용
- TITLE명령문: 출력결과에 제목을 붙이기 위해 사용

- SUMMARY 프로시저
  - MEANS 프로시저와 거의 유사한 기능을 가짐
  - 주로 계산된 기술통계량들에 의해 새로운 데이터 셋을 생성하기 위해 사용

일변량 통계량의 출력

- UNIVARIATE 프로시저
  - MEANS나 SUMMARY 프로시저와 마찬가지로 일변량 기술통계량을 제공
  - 해당 변수의 분포에 대해 보다 다양한 분석결과를 제공
    - 적률, 극단값, 중위수, 4분위수
    - 위치모수, 척도모수
    - 추정값에 대한 신뢰구간
    - 정규성에 대한 검정
    - 정규확률그림, 상자그림, 줄기-잎 그림등과 같은 분포에 대한 그림 출력

UNIVARIATE 프로시저에 의한 기술통계량

```
PROC SORT DATA=mysas.htwt;
BY gender;
RUN;
PROC UNIVARIATE DATA=mysas.htwt  NORMAL PLOTS;
BY gender;
VAR age height;
HISTOGRAM age/NORMAL;
PROBPLOT height / NORMAL;
QQPLOT age / EXPONENTIAL;
RUN;
```

일변량 통계량의 출력

- NORMAL 과 PLOTS 옵션
  - 정규성 검정의 결과와 정규확률그림을 출력하도록 지정
- HISTOGRAM
  - 막대그래프와 모수적 또는 비모수적 밀도곡선을 출력
- PROBPLOT
  - 특정 분포에 근거한 확률그림을 출력
- QQPLOT
  - 특정분포에 근거한 분위수-분위수 그림을 출력
- BY 명령문
  - BY 변수의 자료값에 따라 독립적으로 해당 프로시저를 수행하고자 할 때 사용
  - CLASS 명령문과 유사한 기능
  - 단, BY 명령문을 사용할 때는 BY 변수에 의해서 정렬되어 있어야 함
  - BY 명령문은 거의 모든 프로시저에 대해 제한 없이 사용할 수 있지만 CLASS 명령문은 사용할 수 있는 프로시저와 기능이 다소 제한적
- BOXPLOT 프로시저
  - 보다 근사한 형태의 상자그림을 얻을 수 있음

BOXPLOT 프로시저에 의한 상자그림

```
PROC BOXPLOT DATA=mysas.htwt;
PLOT height*gender / BOXSTYLE=SKELETAL;
RUN;
```

빈도표의 작성

❖ FREQ 프로시저

- 일차원 또는 다차원 빈도표와 변수들간의 상호연관성을 재는 측도들을 제공
- 각 변수값들의 분포와 연관도에 관한 정보를 요약하여 제공

FREQ 프로시저에 의한 빈도표

```
PROC FREQ DATA=mysas.htwt;
TABLES dept gender*dept;
RUN;
```

- TABLES 명령문: TABLES 명령문 내에 여러 개의 빈도표 형식을 지정할 수 있음
  - TABLES age: 변수 age에 관한 1차원 빈도표
  - TABLES age*gender: 변수 age와 gender의 각 수준에 대한 2차원 빈도표
  - TABLES a*b*c*d: 4차원 빈도표
  - TALBES a*(b c): a*b와 a*c의 빈도표
  - TABLES (a b)*(c d): a*c, a*d, b*c, b*d의 빈도표
  - TABLES (a b c)*d: a*d, b*d, c*d의 빈도표
  - TABLES a-c: a, b, c 각각의 빈도표
  - TABLES (a-c)*d: (a b c)*d와 동일

빈도표의 작성

drink 데이터 셋

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

FREQ 프로시저에 의한 빈도표

```
PROC FREQ DATA=mysas.drink;
WEIGHT count;
TABLES age drink age*drink / NOCOL NOPERCENT;
RUN;
```

빈도표의 작성

- 출력제어 옵션
  - NOROW: 행 퍼센트를 출력하지 않음
  - NOCOL: 열 퍼센트를 출력하지 않음
  - NOPERCENT: 각 칸의 퍼센트를 출력하지 않음
- 연관성 측도에 관련된 통계량을 출력하는 옵션
  - CHISQ: 독립성 또는 동질성 검정하기 위한 카이제곱 통계량
  - AGREE: 일치도 통계량
- WEIGHT 명령문
  - 변수값들의 각 조합에 대한 빈도를 나타내는 변수가 데이터 셋 내에 포함되어 있는 경우
  - 빈도변수를 지정하는 명령문은 프로시저에 따라 다름
    - Ex. MEAN, SUMMARY, UNIVARIATE 프로시저에서는 FREQ 명령문을 사용하여 빈도변수를 지정

그림 그리기

❖ PLOT 프로시저

- 두 변수 사이의 관계를 플롯 형식으로 표현하는 기능

PLOT 프로시저의 사용

```
PROC PLOT DATA=mysas.htwt;
PLOT height*age='H' weight*age=gender/HPOS=50 VPOS=15 OVERLAY;
RUN;
```

- 그리고자 하는 수직축(height, weight)과 수평축(age)을 지정
- 하나의 PLOT 명령문 내에 여러 개의 플롯 형식을 지정할 수 있음
- 플롯 형식 지정은 TABLES명령문과 유사
- 플롯에 문자또는 변수를 지정할 수도 있음
  - Ex. Height*age 플롯에서는 문자 H, weight*age 플롯에는 변수 gender의 자료값 사용
- OVERLAY 옵션
  - 동일한 PLOT 명령문 안에 있는 모든 플롯을 하나의 그림에 겹쳐서 그림
- HPOS=, VPOS=
  - 수평축과 수직축의 길이를 지정

그림 그리기

❖ CHART 프로시저

- HBAR(수평막대도표), VBAR(수직막대도표), PIE(파이도표), STAR(별도표), BLOCK(블록도표) EMD의 명령문을 이용하여 여러 가지 형태의 도표를 그리는 기능

PLOT 프로시저의 사용

```
PROC CHART DATA=mysas.htwt;
HBAR dept;
PIE age / DISCRETE;
RUN;
```

- DISCRETE 옵션
  - 숫자변수에 대해서 모든 자료값을 각각의 막대 또는 파이로 표현하도록 지정
  - 생략될 경우 SAS가 적절한 구간으로 나누어 도표를 표현

SUMBAR=, TYPE= 옵션의 사용

```
PROC CHART DATA=mysas.htwt;
HBAR dept gender / SUMVAR= age  TYPE=MEAN;
RUN;
```

- SUMVAR=, TYPE= 옵션
  - 막대 또는 파이에 대해서 다른 변수의 통계량을 표현

프로시저와 함께 사용되는 명령문들

❖ OPTIONS 명령문

- 출력결과의 형식을 규정하거나 SAS의 여러 가지 사용환경을 설정하기 위한 옵션들을 지정하는 명령문

OPTIONS 명령문의 사용

```
OPTIONS LINESIZE=80 PAGESIZE=50 NODATA PAGENO=1;
PROC PRINT DATA=mysas.htwt;
RUN;
```

- CENTER, NOCENTER
  - 출력결과를 가운데 정렬할 것인지 아니면 왼쪽 정렬할 것인지를 지정
- DATA, NODATA
  - 현재 날자의 출력 여부를 지정
- NUMBER, NONUMBER
  - 페이지 번호의 출력 여부를 지정
- LINESIZE=n
  - 출력결과에서 한 줄의 길이를 지정
- PAGESIZE=n
  - 출력결과에서 한 페이지에 들어가는 줄 수를 지정
- PAGENO=n
  - 새로운 페이지 번호를 지정

프로시저와 함께 사용되는 명령문들

❖ TITLE, FOOTNOTE 명령문

- 출력되는 결과에 제목과 주석을 붙이기 위한 명령문
- 두 개의 명령문은 프로시저 단계의 내부에 사용할 수 있고, 특정 프로시저와 관계없이 독자적으로 사용할 수도 있음

TITLE와 FOOTNOTE 명령문의 일반적 사용형식
TITLE*n* ‘TEXT’; 또는 FOOTNOTE*n* ‘TEXT’;

- 나타나는 제목이나 주석은 여러 줄에 걸쳐 겹쳐 출력되도록 처리할 수 있음
  - 사용형식의 *n*: 출력되는 제목이나 주석의 줄 번호를 나타냄
- OPTIONS 명령문의 경우와 마찬가지로 특정 프로시저에만 영향을 주는 것이 아니라 이후의 모든 프로시저의 출력에 영향을 줌
- 새로운 TITLE*n* 명령문이 사용되면 기존의 제목 중 *n*보다 큰 줄 번호를 사용한 TITLE 명령문에 의해서 붙여진 제목은 모두 무시됨

프로시저와 함께 사용되는 명령문들

- 제목에 인용부호(‘ ’)가 포함된 경우에는 이중 인용부호(“ ”)를 사용해서 인용부호가 제목의 일부임을 구별해 주어야 함

TITLE와 FOOTNOTE 명령문의 사용

```
OPTIONS PAGESIZE=30 NODATA PAGENO=1;
TITLE “---Data-Set hwtw ---”;
PROC FREQ DATA=mysas.htwt;
TITLE2 “>>> 성별 빈도 <<<“;
TABLES gender;
RUN;
TITLE2 “>>> 기초통계량<<<“;
TITLE3 “*** ‘성별’ ***”;
PROC MEANS DATA=mysas.htwt MEAN;
CLASS gender;
VAR height weight;
FOOTNOTE “키와 몸무게”;
RUN;
```

프로시저와 함께 사용되는 명령문들

❖ WHERE 명령문

- 특정한 조건을 만족하는 데이터에 대해서만 해당 프로시저를 수행시키기 위해서 사용
- 비교연산자나 IN 또는 CONTAINS 명령어를 사용하여 조건식을 지정할 수 있음

비교 연산자

기 호	약 어	기 능	기 호	약 어	기 능
=	EQ	같다	>=	GE	크거나 같다
^= 또는 ~=	NE	같지 않다	<=	LE	작거나 같다
>	GT	크다	&	AND	그리고
<	LT	작다	| 또는 !	OR	또는

WHERE 명령문의 사용

```
OPTIONS NODATE PAGENO=1;
TITLE '---Data-Set htwt---';
TITLE2 '**** 연령 > 20 ****';
PROC MEANS DATA=mysas.htwt MEAN STD;
WHERE age>20;
CLASS gender;
VAR height weight;
RUN;
```

프로시저와 함께 사용되는 명령문들

- CONTAINS 명령어
  - 변수가 지정된 특정한 문자열을 포함하고 있는지를 검토하여 이를 만족하는 데이터에만 프로시저를 적용

CONTATINS 명령어의 사용

```
TITLE2 '**** Dept : Stat or Math ****';
PROC PRINT DATA=mysas.htwt;
WHERE dept CONTAINS "at";
RUN;
```

- IN 명령어
  - WHERE 명령문에 여러 개의 자료값을 나열할 때 매우 유용하게 사용
  - 콤마 또는 하나 이상의 공백으로 구분하여 자료값을 나열
  - 문자변수에 대해서는 각 문자열을 인용부호(" ")로 닫아주어야 함

IN 명령어의 사용

```
WHERE age IN (16, 19, 21, 28)
WHERE age IN ("강민호", "최병호", "장순미", "김미숙")
```

프로시저와 함께 사용되는 명령문들

- FORMAT 명령문
  - 변수에 대한 출력포맷을 지정하는 것
  - 자료값 자체를 변경하는 것이 아니라 프로시저를 이용하여 데이터를 출력하거나 VIEWTABLE 윈도우에 데이터 셋을 디스플레이 할 때만 자료값을 바꾸어 보여주는 것이기 때문에 데이터의 저장이나 관리에 매우 효율적으로 사용될 수 있음
- LABEL 명령문
  - 각 변수에 변수에 대한 설명(레이블)을 부여하기 위해 사용
  - 데이터 단계에서 출력 포맷이나 레이블이 지정되어 있으면 이 데이터셋을 사용하는 모든 프로시저의 수행에 영향을 줌
  - 프로시저 내에 FORMAT 명령문이나 LABEL 명령문을 사용하면 해당 프로시저가 수행되는 동안에만 영향을 줌

FORMAT, LABEL 명령문의 사용

```
TITLE >>>>>> 출력포맷의 이용 <<<<<<;
PROC PRINT DATA=mysas.htwt LABEL;
FORMAT dept $1. height 5.1 weight 5.1;
LABEL height=신장(cm) weight=체중(kg);
RUN;
```

  - dept는 문자변수로서 1바이트가 할당, 변수 height, weight는 전체 자리수를 5자리로 하고 소수점 이하 자리수를 1로 출력하도록 지정

연습문제

- 7개의 변수(name, gender, status, year, section, score, finalscore)로 이루어진 자료값이 다음과 같을 때, 이 자료값을 SAS 데이터 셋으로 변환시키고, status와 year를 구분하여 score 변수에 대한 기초통계량을 출력하여라.

- 그리고 score 변수와 finalscore 변수의 관계를 PLOT 프로시저를 이용하여 나타내어라. 이 때 gender를 고려하여 플롯으로 표현하여라.

Data

name	gender	status	year	section	score	finalscore
Abbott	F	2	97	A	90	87
Branford	M	1	98	A	92	97
Crandell	M	2	98	B	81	71
Dennison	M	1	97	A	85	72
Edgar	F	1	98	B	89	80
Faust	M	1	97	B	78	73
Greeley	F	2	97	A	82	91
Hart	F	1	98	B	84	80
Isley	M	2	97	A	88	86
Jasper	M	1	97	B	91	93

연습문제

❖ 14종류의 자동차에 대해서 6개의 변수를 조사한 것이다.

- MEANS 프로시저를 이용하여 변수 mileage와 reliable에 대한 평균, 표준편차, 합계를 소수점 이하 세 자리까지 출력하여라.
- UNIVARATE 프로시저를 이용하여 변수 size의 각 수준별로 변수 mileage와 reliable에 대한 일변량 기술통계량들을 출력하여라.
- size*manufact, size*index의 2차원 분할표를 출력하여라.
- 플롯에 표현할 문자변수로 model을 지정하여 두 변수 mileage와 reliable의 플롯을 작성하여라.

size	manufact	model	mileage	reliable	index
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	Tauru	20	3	3