SAS의 통계처리방법(I)

1984. 11

경제기획원조사통계국

일 러 두 기

이 책은 SAS BASIC에 관한 내용을 직원 교육용으로 활용하기 위하여 간략하게 간추린 것입니다. 통계국에서 1983년부터 설치운영하고 있는 SAS는 기본적인 통계분석 뿐만아니라 ETS (Econometion and Time Series), IMS/DL-1 (Information Management System/Data Language 1.)과도 연결이 가능하며 Graphic OR (Operations Research) 등 다양한 분석기법을 취급할 수 있는 program package인바 SAS를 이해하고자 하는 사람에게 도움이 되고져 SAS에 관한 최신정보를 입수하고 실무중심의 예제를 첨부하여 실용적인 참고서를 만들려고 노력했읍니다.

그러나 SAS가 가히 Software의 혁명이라고 할만큼 기능이 다양하고 end user에게 쉽게 사용할 수 있도록 만들어졌는데 반하여 이 책이 과연 입문서로서의 역할을 다할 수 있을런지는 자신할 수 없으나 부족한 점은 계속하여 보완하고 개정하여 나가도록 노력할 계획이오니 양지하시기 바랍니다.

끝으로 이 책을 편찬하는데 심혈을 기울인 이경의처리관의 노고를 치하하는 바입니다.

자료처리과장

목 차

통 계 측 정 방 법

통 계 측 정 방 법

(1) 명목측정(Nominal Measurement)

명목측정이라 함은 수로서의 특성을 전혀 갖지 않고 관찰대상의 고유한 속성을 분류하기 위해 숫자 등의 기호를 부여한 것이다. 예를 들면, 지역별(1. 서울, 2. 부산, 3. 대구, 4. 인천), 성별(1. 남, 2. 녀), 종교(1. 기독교, 2. 불교, 3. 유교, 4. 이스람교)등을 측정할 경우다. 즉, 명목측정은 같다. 다르다. 이외에는 상대적 크기를 알 수 없음은 물론 덧셈, 곱셈과 같은 수학적 계산을 할 수 없는 것으로서 단순히 동일값을 표시하기 위한 수단의 역할만 하는 것이다. 따라서 숫자간의 크기나, 서열을 가정하는 통계적 분석기법에서는 명목측정의 성격을 가진 자료는 이용될 수 없다.

(2) 서열측정(Ordinal Measurement)

서열측정은 명목측정과는 달리 같다, 다르다, 차이 뿐만아니라 크기 서열까지를 제공하는 것으로서 예를들면, 직업분류에 있어서 1. 노무직, 2. 감독직, 3. 관리직을 각각 상, 중, 하로 분류하는 경우이다. 그러나 이 측정은 크다, 높다라는 사실이외에는 어느정도의 차이가 나는가에 대한 정보는 제공하지 못한다. 따라서 서열측정 역시 수의 완전한 특성을 보유하지 못함으로 통계적 기법의 적용에는 한계가 있다고 하였다.

(3) 등간측정(Interval Measurement)

등간측정이란 크다, 작다하는 서열적 성격 뿐만 아니라 얼마만큼 큰

가하는 차이 (distance)에 관한 정보를 갖는 측정이다. 예를 들면 20세, 40세, 60세라는 세사람이 존재한다면 20세와 40세의 차이는 40세와 60세의 차이와 같다고 말할 수 있다. 그러나 60세는 20세보다 세배 늙었다고 말할 수 없다. 그 이유는 등간측정은 절대영점 (Absolute Zero Point)이 아닌 임의영점 (Arbitrary Zero Point)만 갖기 때문이다. 즉, 등간측정은 두 대상의 차이는 나타낼 수 있지만 비율에 대해서는 설명하지 못한다.

(4) 비율측정 (Ratio Measurement)

비율측정은 서열, 등간이외에 한 측정치는 다른 측정치의 두배 또는 세배 등의 비율에 관한 정보까지 제공한다. 즉, 비율측정은 절대영점에 정해져 있는 측정이다. 예를 들면, 무게 (길이)등에서 20 *kg*은 10 *kg*의 두배, 10 *cm*는 5 *cm*의 두배라 말할 수 있다.

1) 집중 경향의 지수

최빈도 (Mode), 중앙치 (Median), 평균 (Mean)

최빈도 (Mode)

최빈도란 한분포에서 빈도가 가장 높은 수치를 말한다. 이때 모든 점수가 똑같은 빈도를 갖는 분포가 있을 수 있다. 이러한 경우에 최빈도는 없다고 본다. 또한 가장 높은 빈도를 2개이상 갖는 분포도 있다. 이러한 분포를 상봉분포 (bimodal distribution) 삼봉분포 (trimodal distribution), 다봉분포 (multimodal distribution)을 이룬다.

중앙치 (Median)

중앙치는 역시 집중경향의 지수의 하나로서 한 분포안에 포함된 전체사례 (N)을 양등분하는 점에 해당하는 수치를 말한다. 즉, 점수들이 순서대로 나열되었을 때 제일 가운데 있는 점수가 중앙치가 된다.

평균 (Mean)

평균은 모든 점수를 다합한 값을 전체 사례수로 나눈 값이다. 평균치는 가장 민감한 집중경향의 지수로서 대부분의 통계절차에 밀접한 관계가 있다. 여기서 평균이 가지고 있는 몇가지 속성은 다음과 같다.

① 각 사례의 점수로부터 평균값을 뺀 즉 편차의 합은 0이다.
  즉 $\Sigma(X-M)=0$

② 평균은 특히 분포가 Skewness 되어 있을때 극단수치(extremes core)에 매우 민감하다.
  즉 평균은 각 수치들의 무게의 중심으로서 지렛대의 균형점 구실을 하는데 한두개의 수치라도 평균으로부터 상대적으로 먼지점에 걸려 있을때 평균은 그 분포의 대표적 수치로서 기능을 다할 수 없는 것이다.

③ 평균치로부터의 편차의 자승의 합 즉 $\Sigma(X-M)^2$ 은 다른 어떤 점수들을 기준으로 해서 얻어진 편차의 제곱들의 총합보다 항상 적다. 평균치는 편차의 자승의 값이 최소가 되는 집중 경향치로서 최소자승의 합을 구함으로써 평균치를 알아낼 수 있다. 이러한 방법이 최소자승법 (Least Squares Method)이다.

2) 분산도의 지수

범위 (Range), 분산 (Variance), 표본편차 (Standard Diviation)

표준점수 (Standard Score)

분산도 (dispersion)란 한 분포안의 사례들이 집중경향치를 중심으로 얼마나 밀집, 혹은 분산되어 있는가 하는 정도를 말한다.

분산도는 집중경향치와 함께 분포를 해석 기술하는데 중요한 역할을 하며 지수를 통해 측정되는데 가장 빈번히 사용되는 지수들은 범위, 분산, 표준편차 등이다.

범위 (Range)

범위는 한 분포내에서 최고치와 최저치 사이의 간격으로 최고치에서 최저치를 뺀값을 말하며 보통 R로 표시한다. 즉 R = Max − Min 이다.

분산 (Variance)

분산 또는 변량이란 편차를 자승하여 모두 합한 후에 이 값을 (전체사례수 − 1)로 나눈 값이다. 한 분포안의 값들이 평균으로 부터 떨어져 있는 방향이 편차 (variance)이다. 이들 편차를 모두 합하면 $\sum d = 0$이 된다. 이들 편차의 종합이 0이 되지 않도록 하기 위하여 편차를 자승한 값이 바로 분산이다.

$$S^2 = \frac{\sum (X - M)^2}{N - 1}$$

표본의 변량 또는 표준편차를 계산할 때 분모는 보통 N - 1이 되는데 이것은 모집단의 변량과 표준편차에 대한 불편 추정치 (un-biased estimated)가 되기 때문이다. 사례수 (N)이 작을 때에는

불편추정치 N-1을 사용하지만 N이 클때에는 N-1 대신 N을 사용해도 별차이는 없다. 변량을 표시하는 방법으로 S^2과 σ^2이 사용되는데 통상 S^2은 표본의 변량을 말하고 σ^2은 모집단의 변량을 말한다.

표준편차 (Standard Diviation)

분산도의 지수로서 변량보다 더 일반적으로 사용되는 것이 표준편차이다. 이는 변량에 제곱근을 씌운 이유는 변량을 구하기 위하여 편차를 자승하여 합쳤으므로 그 반대로 다시 제곱근을 씌운 것이다. 표준편차를 나타내는 기초로 모집단의 경우에 σ, 표본의 경우 S를 사용하는 것이 보통이다.

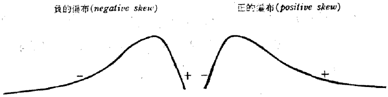
표준점수 (Standard Score)

편차, 분산, 표준편차들은 어떤 점수들이 한 분포내에서 차지하는 상대적 위치를 파악하고자 하는 경우였으나 분포가 다른 주 점수들은 비교해 보고자하는 경우가 있다. 코끼리무게와 사람의 몸무게와 같이 평균치와 표준편차가 다른 두 분포내의 어떤 두 점수들을 비교하고자 하는 경우에 있다. 이와 같이 한 분포내에서 차지하는 상대적 위치나 분포가 다른 두 점수들을 직접 비교하는 경우에 사용되는데 이를 표준점수 또는 Z score라고도 한다.

일반공식은 $Z = \frac{X - M}{S}$

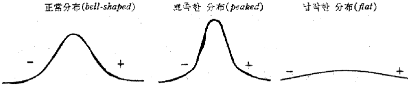
X =원점수, M=평균, S =표준편차

3) SKEWNESS , KURTOSIS



Skewness

어떤 분포가 정상분포에서의 대칭성(symmetry)을 어느 정도 만족시키는 가를 나타내는데 이 역시 정상분포일때 0의 값을 갖는다. skewness가 0보다 크다(S〉0)는 것은 다수의 값이 평균보다 낮은곳에 집중하고 극소수의 값들이 평균보다 높은곳에 위치하고 있어 positive skewness를 이룬다. skewness가 0보다 작을 때(S〈0)에는 이와 정반대로 해석하면 된다.



Kurtosis

분포가 어느 정도 뾰족하게 솟아 있는지를 표시하는 것으로서 정상 분포일 때 0의 값을 갖는다. kurtosis가 0보다 클때(K〉0) 분포는 정상 분포보다 뾰족한(peaked) 모양을 나타내고 0보다 작을때 (K〈0)는 평평한 (flat) 모양을 나타내게 된다.

CHAPTER 1. INTRODUCTION TO SAS

CHAPTER 1. INTRODUCTION TO SAS

1. SAS PRODUCTS

(1) Statistical Analysis System

What is SAS?

SAS is a computer software system that consists of several products that provide tools for data entry, data management, and data analysis:

RETRIEVAL	flexible input techniques.
TRANSFORMATIONS	programming language with statistical and mathematical functions.
MAINTENANCE	storing, documenting, updating, and editing.
MANIPULATION	sorting, subsetting, concatenation, and merging.
REPORT WRITING	printing information using program statements.
PRINTER GRAPHICS	charts and two-dimensional plots.
DATA REDUCTION AND SUMMARIZATION	descriptive statistics.
STATISTICAL ANALYSIS	from simple crosstabulations to complex multivariate techniques.

(2) Additional SAS Products: SAS/GRAPH

A device-intelligent color graphics system that features:

o charts

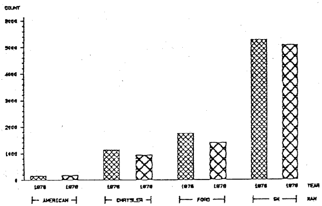
o plots

o maps

o slides

o reports

U.S. Passenger Car Production

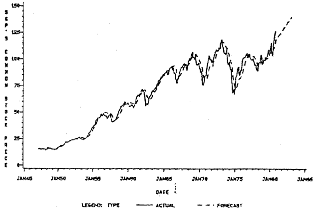


(3) Additional SAS Products: SAS/ETS

An econometric and time series analysis system that features:

- o time series forecasting
- o time series regression techniques
- o simultaneous equation modeling techniques
- o financial reports

PROC FORECAST Output
Displayed with PROC GPLOT

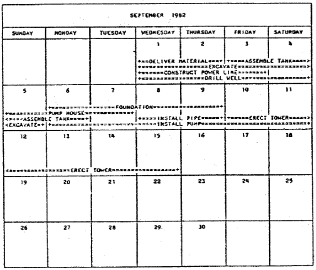


(4) Additional SAS Products: SAS/OR

A systems analysis and operations research system that features:

- o linear programming
- o critical path methods
- o network modeling techniques

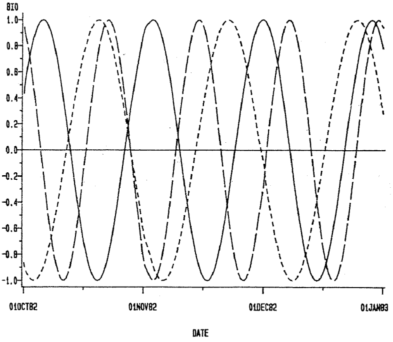
PROC CPM Output
Displayed with PROC CALENDAR



(5) Biorhythm Chart

Birthday Sept. 10, 1952.

BIORHYTHM CHART

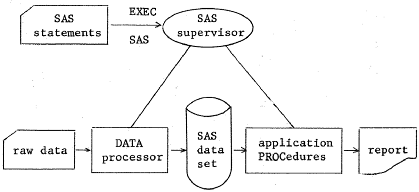


LEGEND: PHYTHM —— EMOTION --------- INTELLECT ——— PHYSICAL

2. THE DATA STEP

(1) SAS Processing

SAS consists of a data-handling language and a library of procedures that work together as a system.



- o A supervisor program, written in re-entrant Assembler, directs the execution.
- o All storage is allocated as needed.
- o The supervisor links to the programs on the library.
- o Each procedure is one or two separate load modules on the library.

(2) SAS Jobs

All SAS jobs are a sequence of SAS steps.

There are only two kinds of SAS steps:

- o DATA steps prepare SAS data sets
- o PROCedure steps analyze or process SAS data sets.

Example:

Given : A company has recored revenue and expense data yearly from 1978 through 1981.

Objective : Compute the difference between expenses and revenue (income) for each year and compute the average income across all the years.

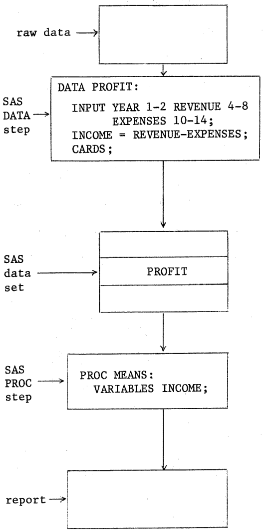
data fields ⟶ | year | | revenue | | expenses |

raw data ⟶

1	2	3	4	5	6	7	8	9	10	11	12	13	14
7	8		4	5	0	0	0		3	2	0	0	0
7	9		5	3	0	0	0		3	6	0	0	0
8	0		5	4	0	0	0		3	8	0	0	0
8	1		6	5	0	0	0		4	3	0	0	0

	YEAR	REVENUE	EXPENSES	INCOME= REVENUE−EXPENSES
variables ⟶				
SAS data set ⟶	78	45000	32000	13000
	79	53000	36000	17000
	80	54000	38000	16000
	81	65000	43000	22000

(3) SAS Jobs



When processing data, the flexibility of programming statements is needed.

Once the data are set up, SAS is told what to do—not how to do it.

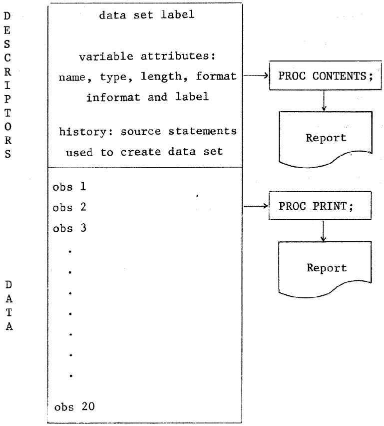
(4) Documenting SAS Data Sets

A SAS data set contains:

o descriptor records

o data records.

SAS data set



(5) THE SAS DATA SET

- DATA steps create SAS data sets.
- SAS data sets are rectangular or flat.
- You refer to a SAS data set by its name. You refer to SAS variables by their names.
- Every SAS data set has a name and is physically stored on disk or tape. In simple jobs, the SAS data sets are stored on temporary space, but they can be stored permanently.
- Data must be in the form of a SAS data set before they can be analyzed by SAS procedures.
- There is no limit to the number of DATA steps that can be included in a single job.
- If a raw file is to be produced, the DATA step can create a null SAS data set, saving time and space.

(6) SAS Statements are Free-Format

- All SAS statements end with a semicolon (;).
- Statements begin and end anywhere.
- Several statements can be on the same line.
- One statement can continue over several lines.
- SAS words are separated by one or more blanks.

(7) A SAS Job with a DATA Step and Two PROC Steps.

DATA step →

```
DATA BIGTOWNS;

  INPUT RANK80 RANK70
    CITY $14.  POP80;

LABEL
  POP80 = 'IN THOUSANDS';

CARDS;

 1   1   New York          7135
 2   2   Chicago           3049
 3   3   Los Angeles       2787
 4   4   Philadelphia      1755
 5   6   Houston           1573
 6   5   Detroit           1258
 7   8   Dallas             847
 8  15   San Diego          817
 9  14   San Antonio        798
10   7   Baltimore          792
11  18   Honolulu           723
12  21   Indianapolis       704
13  20   Phoenix            681
14   9   Washington DC      671
15  16   Memphis            664
16  13   San Francisco      649
17  12   Milwaukee          633
18  17   Boston             597
19  10   Cleveland          595
20  29   San Jose           593
;
```

PROC step →

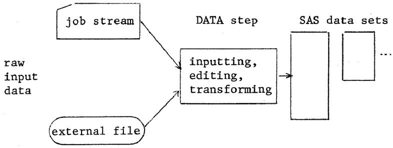
```
TITLE;

PROC PRINT;
  ID RANK80;
  FORMAT POP80 COMMA5.;
```

PROC step →

```
PROC CONTENTS;
```

(8) Overview of the DATA Step

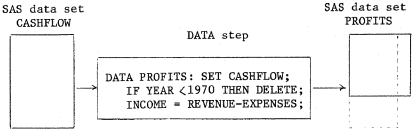
Retrieval



Transferring with reshaping

Given : SAS data set CASFLOW contains the variables YEAR, REVENUE, and EXPENSES.

Objective: Delete observations for years before 1970 and compute the income for each remaining year.

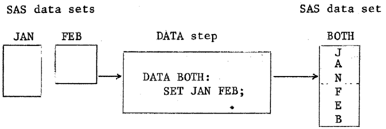


(9) Overview of the DATA Step

Concatenation

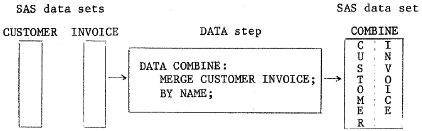
Given : SAS data set JAN contains January sales data and SAS data set FEB contains February sales data.

Objective : Concatenate the two SAS data sets into one SAS data set that contains the sales data for both months.



Merging

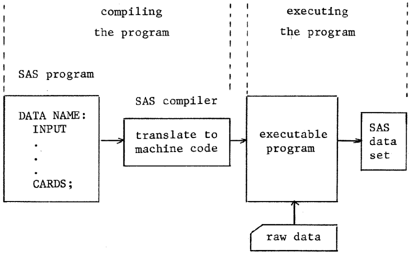
Given : SAS data set CUSTOMER contains names and addresses for customers and SAS data set INVOICE contains names and invoices for recent sales to the customers.



(10) Overview of the DATA Step

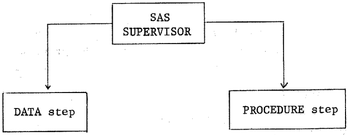
The DATA step involves writing a compact SAS program to process data. SAS processes the program in two steps.

- o The program is compiled.
- o The program is executed.



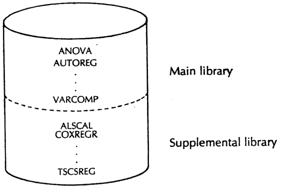
(11) SAS Job

- o DATA steps prepare SAS data sets.
- o PROCEDURE (or PROC) steps process SAS data sets.

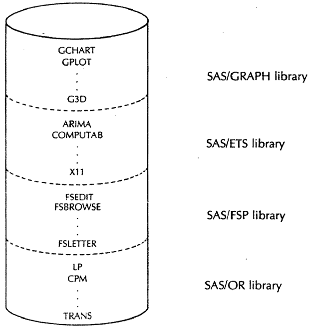


3. SAS PROCEDURE LIBRARY JCL

- Procedures are programs designed to process SAS data sets.
- A procedure is called by name from the SAS procedure library.



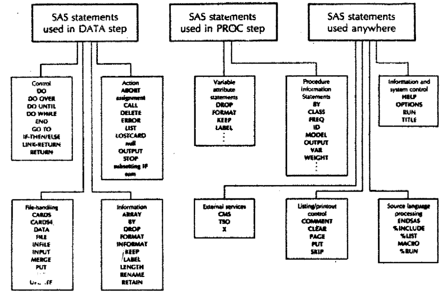
SAS offers optional 'add-on' procedure libraries

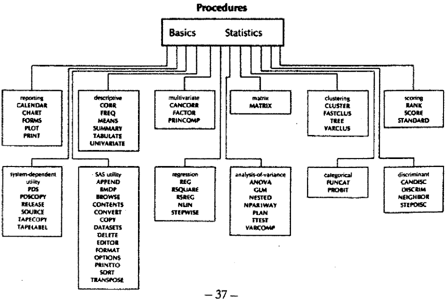


4. SAS JCL

```
//SASSAS30   JOB
//SAS        EXEC  SAS
//SYSIN         DD  *


      SAS PROGRAM


/*
//
```

(1) SAS VIEW


SAS statements used in DATA step
SAS statements used in PROC step
SAS statements used anywhere
Control DO DO OVER DO UNTIL DO WHILE END GO TO IF-THEN/ELSE LINK-RETURN RETURN
Action ABORT assignment CALL DELETE ERROR LIST LOSTCARD null OUTPUT STOP subsetting IF sum
Variable attribute statements DROP FORMAT KEEP LABEL
Procedure information Statements BY CLASS FREQ ID MODEL OUTPUT VAR WEIGHT
Information and system control HELP OPTIONS RUN TITLE
File-handling CARDS CARDS4 DATA FILE INFILE INPUT MERGE PUT
Information ARRAY BY DROP FORMAT INFORMAT KEEP LABEL LENGTH RENAME RETAIN
External services CMS TSO X
Listing/printout control COMMENT CLEAR PAGE PUT SKIP
Source language processing ENDSAS %INCLUDE %LIST MACRO %RUN


Procedures
Basics Statistics
reporting CALENDAR CHART FORMS PLOT PRINT
descriptive CORR FREQ MEANS SUMMARY TABULATE UNIVARIATE
multivariate CANCORR FACTOR PRINCOMP
matrix MATRIX
clustering CLUSTER FASTCLUS TREE VARCLUS
scoring RANK SCORE STANDARD
system-dependent utility PDS PDSCOPY RELEASE SOURCE TAPECOPY TAPELABEL
SAS utility APPEND BMDP BROWSE CONTENTS CONVERT COPY DATASETS DELETE EDITOR FORMAT OPTIONS PRINTTO SORT TRANSPOSE
regression REG RSQUARE RSREG NLIN STEPWISE
analysis-of-variance ANOVA GLM NESTED NPAR1WAY PLAN TTEST VARCOMP
categorical FUNCAT PROBIT
discriminant CANDISC DISCRIM NEIGHBOR STEPDISC

(2) A simple SAS job

Example :

Given : The name, sex, age, height, and weight of each student in a summer camp were recorded and stored with the following record format.

DATA FIELD DESCRIPTION	FIELD POSITION
NAME	1-10
SEX	11
AGE	13-14
HEIGHT	16-19
WEIGHT	21-25

Objectives : 1. Obtain a listing of the data.

2. Compute some descriptive statistics for the heights and weights such as the mean, minimum, and maximum.

3. Produce a plot of height versus weight.

(3) A Listing of the Raw Data

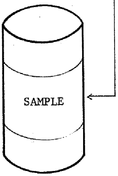
NAME	SEX	AGE	HEIGHT	WEIGHT
KYUNGEUI	M	15	65.7	110.9
JOHN	M	12	59.0	99.5
JAMES	M	12	57.3	83.0
ALFRED	M	14	69.0	112.5
WILLIAM	M	15	66.5	112.0
JEFFREY	M	13	62.5	84.0
RONALD	M	15	67.0	133.0
THOMAS	M	11	57.5	85.0
PHILIP	M	16	72.0	150.0
ROBERT	M	12	64.8	128.0
HENRY	M	14	63.5	102.5
JANET	F	15	62.5	112.5
JOYCE	F	11	51.3	50.5
JUDY	F	14	64.3	90.0
CAROL	F	14	62.8	102.5
JANE	F	12	59.8	84.5
LOUISE	F	12	56.3	77.0
BARBARA	F	13	65.3	98.0
HARY	F	15	66.5	112.0
ALICE	F	13	56.5	84.0

Note: These data can be on cards, disk, tape, or lines entered on a terminal.

5 Temporary Data Sets

default DDname = WORK

```
DATA SAMPLE;
 INPUT NAME $ 1-10 SEX $ 11 AGE 13-14
       HEIGHT 16-19 WEIGHT 21-25;
CARDS;
KYUNGEUI  M 15 65.7 110.9
JOHN      M 12 59.0  99.5
JAMES     M 12 57.3  83.0
ALFERD    M 14 69.0 112.5
WILLIAM   M 15 66.5 112.0
JEFFREY   M 13 62.5  84.0
RONALD    M 15 67.0 133.0
THOMAS    M 11 57.5  85.5
PHILIP    M 16 72.0 150.0
ROBELT    M 12 64.8 128.0
HENRY     M 14 65.3 102.5
JANET     F 15 62.5 112.5
JOYCE     F 11 51.3  50.5
JUDY      F 14 64.3  90.0
CAROL     F 14 62.8 102.5
JANE      F 12 59.8  84.5
LOUISE    F 12 56.3  77.0
BARBARA   F 13 65.3  98.0
MARY      F 15 66.5 112.0
ALICE     F 13 56.5  84.0
```



The note on the SAS log for this DATA step would read:

NOTE: DATA SET WORK. SAMPLE HAS 10 OBSERVATIONS AND 5 VARIABLES. 504 OBS/TRK.

The data set WORK. SAMPLE has been stored in work space and will be deleted at the end of the job or session.

The data set may be referred to as SAMPLE or WORK. SAMPLE

PROC PRINT DATA=WORK. SAMPLE;

is equivalent to

PROC PRINT DATA=SAMPLE;

6. A Simple SAS Job: OS Batch

```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SAMPLE     DD  DSN=SAS.SAMPLE,DISP=SHR
//SYSIN        DD  *
DATA SAMPLE;
 INPUT NAME $ 1-10 SEX $ 11 AGE 13-14
       HEIGHT 16-19 WEIGHT 21-25;
CARDS;
KYUNGEUI   M 15 65.7 110.9
JOHN       M 12 59.0  99.5
JAMES      M 12 57.3  83.0
ALFERD     M 14 69.0 112.5
WILLIAM    M 15 66.5 112.0
JEFFREY    M 13 62.5  84.0
RONALD     M 15 67.0 133.0
THOMAS     M 11 57.5  85.5
PHILIP     M 16 72.0 150.0
ROBELT     M 12 64.8 128.0
HENRY      M 14 65.3 102.5
JANET      F 15 62.5 112.5
JOYCE      F 11 51.3  50.5
JUDY       F 14 64.3  90.0
CAROL      F 14 62.8 102.5
JANE       F 12 59.8  84.5
LOUISE     F 12 56.3  77.0
BARBARA    F 13 65.3  98.0
MARY       F 15 66.5 112.0
ALICE      F 13 56.5  84.0
;
PROC PRINT DATA=SAMPLE;
PROC MEANS DATA=SAMPLE;
     VAR   WEIGHT HEIGHT;
PROC PLOT DATA=SAMPLE;
     PLOT WEIGHT*HEIGHT;
/*
//
```

(1) PROC PRINT output: OS batch

```
                         14:47 WEDNESDAY, OCTOBER 31, 1984     1
                          SAS
```

OBS	NAME	SEX	AGE	HEIGHT	WEIGHT
1	KYUNGEUI	M	15	65.7	110.9
2	JOHN	M	12	59.0	99.5
3	JAMES	M	12	57.3	83.0
4	ALFERD	M	14	69.0	112.5
5	WILLIAM	M	15	66.5	112.0
6	JEFFREY	M	13	62.5	84.0
7	RONALD	M	15	67.0	133.0
8	THOMAS	M	11	57.5	85.5
9	PHILIP	M	16	72.0	150.0
10	ROBELT	M	12	64.8	128.0
11	HENRY	M	14	65.3	102.5
12	JANET	F	15	62.5	112.5
13	JOYCE	F	11	51.3	50.5
14	JUDY	F	14	64.3	90.0
15	CAROL	F	14	62.8	102.5
16	JANE	F	12	59.8	84.5
17	LOUISE	F	12	56.3	77.0
18	BARBARA	F	13	65.3	98.0
19	MARY	F	15	66.5	112.0
20	ALICE	F	13	56.5	84.0

(2) PROC MEANS output: OS batch

```
                    SAS                    14:47 WEDNESDAY, OCTOBER 31, 1984     2
```

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	STD ERROR OF MEAN	SUM	VARIANCE	C.V.
WEIGHT	20	100.59500000	22.28130776	50.50000000	150.00000000	4.98226579	2011.9000000	496.45944737	22.150
HEIGHT	20	62.59500000	5.03184403	51.30000000	72.00000000	1.13620965	1251.9000000	25.31944737	8.118

7. A Simple Interactive SAS Job: TSO

```
READY
SAS
NOTE: SAS RELEASE 82.2
      AT SAS INSTITUTE INC. (0) (00000).
  1?
DATA CLASS;
  2?
   INPUT NAME $ 1-8 SEX $ 11 AGE 13-14
  3?
         HEIGHT 16-19 WEIHGT 21-25;
  4?
CARDS;
  5>
KYUNGEUI M 15   65.7   110.9
  6>
JONH     M 12   59.0    99.5
  7>
JAMES    M 12   57.3    83.0
  8>
ALERED   M 14   69.0   112.5
  9>
WILLEAM  M 15   66.5   112.0
 10>
JEEEREY  M 13   62.5    84.0
 11>
RONALD   M 15   67.0   133.0
 12>
THOMAS   M 11   57.5    85.0
 13>
PHILIP   M 16   72.0   150.0
 14>
ROBERT   M 12   64.8   128.0
 15>
HENRY    M 14   63.5   128.0
 16>
JAMET    E 15   62.5   112.5
 17>
JOYCE    E 11   51.3    50.5
 18>
JUDY     E 14   64.3    90.0
 19>
CAROL    E 14   62.8   102.5
 20>
JANE     E 12   59.8    84.5
 21>
LOUISE   E 12   56.3    77.0
 22>
BARBARA  E 13   65.3    98.0
 23>
MARY     E 15   66.5   112.0
 24>
ALICE    E 13   56.5    84.0
 25>
RUN;
NOTE: DATA SET WORK CLASS HAS 19 OBSERVATIONS
```

8. SAS Syntax and SAS Data Sets

(1) Rules for writing SAS statements

SAS statements begin with an identifying keyword and end with a semicolon.

```
DATA CLASS:
   INPUT NAME $1-10 SEX $11 AGE 13-14
        HEIGHT 16-19 WEIGHT 21-25;
   CARDS;
data lines
PROC PRINT DATA=CLASS;
PROC MEANS DATA=CLASS;
   VARIABLES HEIGHT WEIGHT;
PROC PLOT DATA=CLASS;
   PLOT WEIGHT*HEIGHT;
```

SAS statements are free-format.

- They can begin anywhere, end anywhere
- One statement can continue over several lines.
- Several statements can be on a line.
- Blanks-as many as you want-separate fields. Special characters also separate fields.

Note: We recommend that DATA and PROC statements start in column 1 and that other statements be indented.

(2) Structure of SAS Data Sets

A SAS data set is a collection of data values arranged in a rectangular table.

VARIABLES

	NAME	SEX	AGE	HEIGHR	WEIGHT
observation 1	KYUNGEUI	M	15	65.7	110.9
observation 2	JOHN	M	12	59.0	99.5
observation 3	JAMES	M	12	57.3	83.0
.	.	.	.	.	.
observation 20	ALICE	F	13	56.5	84.0

The columans in the table are called variables.

o Variables correspond to fields of data.

o Each variable is given a name.

o There are two kinds of variables.

character variables: NAME, SEX (1-200 long)

numeric variables: AGE, HEIGHT, WEIGHT (floating)

The rows in the table are called observations (or records).

There is no limit on the number of observations.

Note: The rectangular structure of a SAS data set implies that every variable must exist for each observation.

(3) Naming SAS Data Sets and SAS Variables

Rules for names:

- o 1-8 characters
- o start with A-Z or ___
- o continue with numbers, letters, or underscores.

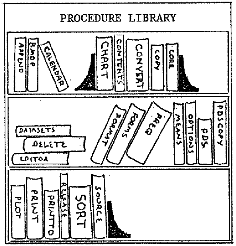
Suggestion: Choose meaningful data set and variable names.

Note: Under CMS SAS, the underscore, '__', is not allowed in SAS data set names.

Example:

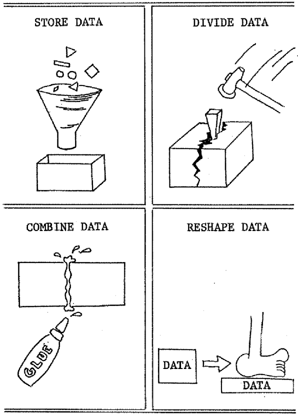
```
DATA CLASS;
   INPUR NAME $ 1-10 SEX $11 AGE 13-14
        HEIGHT 16-19 WEIGHT 21-25;
   CARDS;
```

(4) Procedure and Data Step.



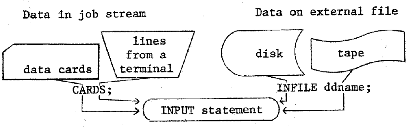
- The SAS Procedure Library contains PROCs or procedures to process SAS data sets.
- Each is accessed by name from the SAS procedure library.

DATA STEP: DATA HANDLING



(5) Inputting Raw Data

SAS can handle data in virtually any form, from almost any input file.



Note: Regardless of where the data are stored, the same INPUT statement is used.

Functions of the CARDS, INFILE and INPUT statements

- o Reading the data.

INPUT statement reads raw data lines

assigns names to the SAS variables that correspond to the fields

has three modes: COLUME, LIST, and FORMATTED.

- o Pointing to the data file.

CARDS statement indicates to SAS that data records follow immediately

INFILE statement points SAS to an external file where the raw data are stored.

(6) Formatted Input

Specify the starting location and field widths (similar to FORTRAN and PL/I). Move an input "pointer" to the starting position of the field, then specify the variable and an informat.

INPUT pointer control variable informat;

informats:	W.	numeric width
	W.d	numeric with decimal
	$W.	character
pointer controls:	@n	go to column n
	+n	move the pointer n positions
	W.	informats advance the pointer

```
INPUT NAME $8.  @11  SEX $1.  +1  AGE 2.
     +1 HEIGHT 4.  +1 WEIGHT 5.;
```

NAME	SEX	AGE	HEIGHT	WEIGHT
KYUNGEUI	M	15	65.7	110.9
JOHN	M	12	59.0	99.5
JAMES	M	12	57.3	83.0
ALFRED	M	14	69.0	112.5

With formatted input you can read data in nonstandard numeric or character formats.

(7) Selected Informats:

w.	standard numeric
w.d	standard numeric with decimal
$w.	standard character
$CHARw.	characters with blanks
HEXw.	numeric hexadecimal
$HEXw.	character hexadecimal
IBw.d	integer binary
PIBw.d	positive integer binary
PDw.d	packed decimal
PKw.	unsigned packed decimal
RBw.d	real binary (floating point)
ZDw.d	zoned decimal
ZDBw.d	zoned decimal with blanks
CBw.d	
$CBw.	column binary
PUNCHd.	
ROWw.d	
COMMAw.d	commas in numbers
Ew.	scientific notation
BZw.d	blanks are zeros
$VARYINGw.	varying-length character values

See pages 388-398 in the SAS User's Guide: Basics, 1982 Edition for additional information on SAS informats.

CHAPTER 2. SAS DATA FILE에 관하여

CHAPTER 2. SAS DATA FILE에 관하여

1. CREATING VARIABLES AND EDITING VALUES

(1) Assignment Statements

Assignment statements are used to create new variables and to modify values of existing variables. SAS evaluates an expression and assigns the result to a variable.

```
variable=expression;
```

(2) Example

- o Read three variable (YEAR, REVENUE, and EXPENSES) into a SAS data set.
- o Add a variable named INCOME, which is the difference between REVENUE and EXPENSES.
- o Change the values of YEAR from 2 digits to 4 digits.

```
//SPSSAS30  JOB CLASS=V
//SAS       EXEC  SAS
//SAMPLE  DD DSN=SAS.SAMPLE,DISP=SHR
//SYSIN         DD  *
DATA PROFITS;
     INPUT YEAR REVENUE EXPENSES;
     INCOME=REVENUE-EXPENSES;
     YEAR=YEAR+1900;
 CARDS;
80 5650 1050
81 6280 1140
;
PROC PRINT;
/*
//
```

SAS

OBS	YEAR	REVENUE	EXPENSES	INCOME
1	1980	5650	1050	4600
2	1981	6280	1140	5140

Program data vector

YEAR	REVENUE	EXPENSES	INCOME

Note: Any variable defined by an assignment statement is included in the program data vector.

(3) Types of Expressions

o simple arithmetic operation: + - * / **

X2=X;	move the value
SUM=X+Y	addition
DIF=X-Y	subtraction
TWICE=X*2;	multiplication
HALF=X/2;	division
CUBIC=X**3;	exponentiation
Y=-X;	change the sign

o constants

N=0;	numeric constant
SEX='FEMALE';	character constant

o complex expressions

priority of evaluation () ** * / + -

A=X+Y+Z;	left to right
A=X+Y*Z;	operator precedence
A=X/Y/Z;	left to right
A=X/(Y/Z);	parenthetical

o functions

```
variable=FUNCTIONNAME (argument1, argument2,...);
S=SQRT(X);
A=ABS(X);
Z=ABS (SQRT(X)-2);
```

(4) SAS Functions

o Selected functions that compute simple statistics.

SUM	sum
MEAN	arithmetic mean
VAR	variance
MIN	minimum value
MAX	maximum value
STD	standard deviation

Example:

Given : Temperature data a specific location are recorded every hour on the hour for several days. Each record in a file represents one day and contains the date and the 24 recorded temperatures for that date.

Objective: Create a SAS data set that contains the date, the 24 hourly temperatures, the average temperature, the minimum temperature and the maximum temperature for each day.

```
DATA TEMP;
   INPUT DATE $ 1-7 @11 (T1-24) (2.);
   AVGTEMP=MEAN (OF T1-T24);
   MINTEMP=MIN  (OF T1-T24);
   MAXTEMP=MAX  (OF T1-T24);
   CARDS:
data lines
```

program data vector

DATE	T1	...	AVGTEMP	MINTEMP	MAXTEMP

(5) DO and END Statements

Example : Execute several statements when a condition is met.

Given : Salary information for a group of employees.

Objective : If the department number is 201, then define the department name to be SALES and set gross pay equal to salary plus commission. Otherwise, the department name is ADMIN and the gross pay is simply equal to salary.

(6) Sample

```
//SPSSAS30  JOB CLASS=V
//SAS        EXEC  SAS
//SAMPLE  DD DSN=SAS.SAMPLE,DISP=SHR
//SYSIN        DD  *
DATA EMPLOY;
     INPUT NAME$ 1-8 DEPTNO 10-12
           COM 14-17 SALARY 19-23;
     IF DEPTNO=201 THEN
        DO;
           DEPT='SALES';
           GROSSPAY=COM+SALARY;
        END;
     ELSE
        DO;
          DEPT='ADMIN';
           GROSSPAY=SALARY;
        END;
 CARDS;
JOHNSON  201 1500 18000
MOSSER   101      21000
LARKIN   101      24000
GARRETT  201 4800 18000
;
PROC PRINT;
```

SAS

OBS	NAME	DEPTNO	COM	SALARY	DEPT	GROSSPAY
1	JOHNSON	201	1500	18000	SALES	19500
2	MOSSER	101	.	21000	ADMIN	21000
3	LARKIN	101	.	24000	ADMIN	24000
4	GARRETT	201	4800	18000	SALES	22800

(7) Comments on Missing Values

Missing values propagate through arithmetic expressions.

```
DATA;
   INPUT CHECKING SAVINGS;
   TOTAL1=CHECKING+SAVINGS;
   TOTAL2=SUM (CHECKING, SAVINGS);
   CARDS;
```

```
100  2000
300
```

CHECKING	SAVINGS	TOTAL1	TOTAL2
100	2000	2100	2100
300	.	.	300

Missing values compare as minus infinity.

```
DATA;
   INPUT PAYMENT DUE;
   IF PAYMENT <DUE THEN STATUS='PAST DUE';
   ELSE STATUS='PAID';
   CARDS;
10  20
25  25
    10
```

PAYMENT	DUE	STATUS
10	20	PAST DUE
25	25	PAID
.	10	PAST DUE

(8) Selecting Observations

The subsetting IF statement

General form of the subsetting IF statement:

IF expression;

The subsetting IF statement is equivalent to:

IF ¬ (expression) THEN DELETE;

The subsetting If statement tells SAS which observations to include in the output SAS data set. The statement works like a gate; it allows an observation to pass when the expression is true.

Example: Take a subset of the data.

```
DATA HISTORY;
   INPUT YEAR REVENUE;
   IF YEAR 1970 THEN DELETE;
   CARDS;

DATA HISTORY;
   INPUT YEAR REVENUE;
   IF YEAR  =1970;
   CARDS;
```

(9) The OUTPUT Statement

```
DATA HISCHOOL COLLEGE;
   INPUT NAME $ 1-8 SEX $ 10 YRS-EDUC 12-13;
   IF YRS_EDUC <= 12 THEN OUTPUT HISCHOOL;
   IF YRS_EDUC > 12 THEN OUTPUT COLLEGE;
   CARDS;
KATHRYN   F  16
GEORGE    M  12
WILLIAM   M  18
JENNIFER  F  12
CYNTHIA   F  16
```

input buffer

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...	80

program data vector

NAME	SEX	YRS_EDUC

SAS data set HISCHOOL

NAME	SEX	YRS_EDUCE
GEORGE	M	12
JENNIFER	F	12

SAS data set COLLEGE

NAME	SEX	YRS_EDUCE
KATHRYN	F	16
WILLIAM	M	18
CYNTHIA	F	16

(10) Executing the DATA Step

```
DATA NEWSMP;
   SET SAMPLE;
   BIRTH-YR=1984-AGE;
```

SAS data set
SAMPLE

NAME	SEX	AGE	HEIGHT	WEIGHT
KYUNGEUI	M	15	65.7	110.9
JOHN	M	12	59.0	99.5
JAMES	M	12	57.3	83.0
ALFRED	M	14	69.0	112.5

program data vector

NAME	SEX	AGE	HEIGHT	WEIGHT

SAS data set
NEWSAMP

NAME	SEX	AGE	HEIGHT	WEIGHT	BIRTH_YR
KYUNGEUI	M	15	65.7	110.9	1969
JOHN	M	12	59.0	99.5	1972
JAMES	M	12	57.3	83.0	1972
ALFRED	M	14	69.0	112.5	1970

(11) Sample

```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SYSIN         DD  *
DATA SAMPLE;
 INPUT NAME $ 1-10 SEX $ 11 AGE 13-14
       HEIGHT 16-19 WEIGHT 21-25;
CARDS;
KYUNGEUI  M 15 65.7 110.9
JOHN      M 12 59.0  99.5
JAMES     M 12 57.3  83.0
ALFERD    M 14 69.0 112.5
WILLIAM   M 15 66.5 112.0
JEFFREY   M 13 62.5  84.0
RONALD    M 15 67.0 133.0
THOMAS    M 11 57.5  85.5
PHILIP    M 16 72.0 150.0
ROBELT    M 12 64.8 128.0
HENRY     M 14 65.3 102.5
JANET     F 15 62.5 112.5
JOYCE     F 11 51.3  50.5
JUDY      F 14 64.3  90.0
CAROL     F 14 62.8 102.5
JANE      F 12 59.8  84.5
LOUISE    F 12 56.3  77.0
BARBARA   F 13 65.3  98.0
MARY      F 15 66.5 112.0
ALICE     F 13 56.5  84.0
;
DATA NEWSAMP;
     SET SAMPLE;
     BIRTH_YR=1984-AGE;
PROC PRINT DATA=NEWSAMP;
/*
//
```

(12) Print the Data Set

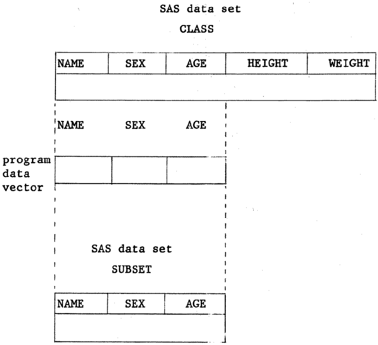
SAS

OBS	NAME	SEX	AGE	HEIGHT	WEIGHT	BIRTH_YR
1	KYUNGEUI	M	15	65.7	110.9	1969
2	JOHN	M	12	59.0	99.5	1972
3	JAMES	M	12	57.3	83.0	1972
4	ALFERD	M	14	69.0	112.5	1970
5	WILLIAM	M	15	66.5	112.0	1969
6	JEFFREY	M	13	62.5	84.0	1971
7	RONALD	M	15	67.0	133.0	1969
8	THOMAS	M	11	57.5	85.5	1973
9	PHILIP	M	16	72.0	150.0	1968
10	ROBELT	M	12	64.8	128.0	1972
11	HENRY	M	14	65.3	102.5	1970
12	JANET	F	15	62.5	112.5	1969
13	JOYCE	F	11	51.3	50.5	1973
14	JUDY	F	14	64.3	90.0	1970
15	CAROL	F	14	62.8	102.5	1970
16	JANE	F	12	59.8	84.5	1972
17	LOUISE	F	12	56.3	77.0	1972
18	BARBARA	F	13	65.3	98.0	1971
19	MARY	F	15	66.5	112.0	1969
20	ALICE	F	13	56.5	84.0	1971

(13) Reading Selected Variables

Example: DROP and KEEP input data set options.

```
DATA SUBSET;
   SET CLASS (DROP=HEIGHT WEIGHT);

DATA SUBSET;
   SET CLASS (KEEP=NAME SEX AGE);
```



(14) The PUT Function in PROG Step

Example:

```
//SPCSAS30  JOB CLASS=V
//SAS         EXEC  SAS
//SYSIN          DD  *
DATA EXAMPLE;
   INPUT DEC @@;
   CNUM = PUT(DEC,4.);
   HEX  = PUT(DEC,HEX.);
   ROMAN= PUT(DEC,ROMAN.);
   WORDS= PUT(DEC,WORDS15.);
CARDS;
1 2 3 4 5 6 7 8 9 10
 50 100 1000
;
PROC PRINT ;
   VAR CNUM HEX ROMAN WORDS DEC;
   ID DEC;
 TITLE PUT FUNCTION;
PROC CONTENTS;
 RUN;
/*
//
```

PUT FUNCTION

DEC	CNUM	HEX	ROMAN	WORDS	DEC
1	1	00000001	I	ONE	1
2	2	00000002	II	TWO	2
3	3	00000003	III	THREE	3
4	4	00000004	IV	FOUR	4
5	5	00000005	V	FIVE	5
6	6	00000006	VI	SIX	6
7	7	00000007	VII	SEVEN	7
8	8	00000008	VIII	EIGHT	8
9	9	00000009	IX	NINE	9
10	10	0000000A	X	TEN	10
50	50	00000032	L	FIFTY	50
100	100	00000064	C	ONE HUNDRED	100
1000	1000	000003E8	M	ONE THOUSAND	1000

```
                                        PUT FUNCTION                    17:39 THURSDAY, NOVEMBER 15, 1984    2
                               CONTENTS OF SAS DATA SET WORK.EXAMPLE
TRACKS USED=2  SUBEXTENTS=2  OBSERVATIONS=13  CREATED BY OS JOB SPCSAS30  ON CPUID 00-3031-000000
AT 17:39 THURSDAY, NOVEMBER 15, 1984    BY SAS RELEASE 82.3  DSNAME=SYS84320.T173842.RF103.SPCSAS30.R0000001  BLKSIZE=19039
LRECL=45  OBSERVATIONS PER TRACK=423  GENERATED BY DATA
                                    ALPHABETIC LIST OF VARIABLES

          #  VARIABLE  TYPE LENGTH POSITION  FORMAT        INFORMAT  LABEL

          2  CNUM      CHAR     4       12
          1  DEC       NUM      8        4
          3  HEX       CHAR     8       16
          4  ROMAN     CHAR     6       24
          5  WORDS     CHAR    15       30

              +-------------------------- SOURCE STATEMENTS ---------------------------+
              |DATA EXAMPLE;                                                           |
              |    INPUT DEC @@;                                                       |
              |    CNUM = PUT(DEC,4.);                                                 |
              |    HEX  = PUT(DEC,HEX.);                                               |
              |    ROMAN= PUT(DEC,ROMAN.);                                             |
              |    WORDS= PUT(DEC,WORDS15.);                                           |
              |CARDS;                                                                  |
              +------------------------------------------------------------------------+
```

2. CREATING FORMATS

(1) The FORMAT Procedure

The FORMAT procedure is used to create user-defined formats.

These user-defined formats can be used:

- o in a PUT statement
- o in a FORMAT statement with a procedure.

(2) PROC FORMAT Options;

Selected options:

PRINT

DDNAME=ddname

Statements used with PROC FORMAT:

```
VALUE name (options)
     range 1 = 'label1'
     range 2 = 'label2'
     ... ;

PICTURE name (options)
     range 1 = 'picture1' (options)
     range 2 = 'picture2' (options)
```

(3) The VALUE Statement

Assigning values:

o single numbers

```
VALUE Q 1 = 'AGREE' 2 = 'DISAGREE'
```

o ranges of numbers

```
VALUE AGEFMT  0-12 = CHILD
             13-19 = 'TEEN'
             20-HIGH = 'ADULT';
```

o several values

```
VALUE SEXFMT 1 = 'FEMALE'
             2 = 'MALE'
             0, 3-9 = 'MISCODED'
```

o character values and ranges of characters

```
VALUE $GRADE A = 'GOOD'
             B-D = 'FAIR'
             E = 'POOR'
             I,U = 'SEE INSTRUCTOR';
```

o character values with special characters.

```
VALUE $CODEX 'A*1' = 'FIRST'
             'A*2' = 'SECOND'
              A___ = 'MORE THAN TWO';
```

Assigning labels:

o maximum length of 40 characters

o labels must be enclosed in single quotes.

(4) Date, Time, and Datetime Informats and Formats

Name of format or informat	Example	Informat, format, or both
DATEw.	04JUL1976	both
YYMMDDw.	76-07-04	both
MMDDDYYw.	7/4/76	both
DDMMYYw.	4/7/76	both
MONYYw.	Jul76	both
YYQw.	76Q3	both
WEEKDATEw.	Monday, July 4, 1976	format
WORDDATEw.	July 4, 1976	format
HHMMw.d	23:45	format
HOURw.d	23	format
MMSSw.d	45:23.4	format
MSECw.	TIME MIC values	informat
PDTIMEw.	packed-decimal time from RMF records	informat
RMFDURw.	RMF time interval measurements	informat
TIMEw.d	23:45:23.5	both
TODw.	23:45:23.4	format
TUw.	timer units	informat
DATETIMEw.	0JUL1976:23:45:23.5	both
RMFSTAMPw.	RMF time-date field	informat
SMFSTAMPw.	SMF time-date field	informat
TODSTAMPw.	8-byte time-of-day stamp	informat

Note: See pages 409-421 in the SAS User's Guide: Basics, 1982 Edition.

(5) Date, Time, and Datetime Values

Example : A company wants to determine the length of Employment in years for each of its employees who resigned in 1981.

```
//SPSSAS30   JOB CLASS=V
//SAS         EXEC  SAS
//SAMPLE   DD DSN=SAS.SAMPLE,DISP=SHR
//SYSIN          DD  *
DATA RESIGNED;
     INPUT NAME $ 10. +1 FIRSTDAY MMDDYY8.
      +1 LASTDAY MMDDYY8.;
     DAYS=LASTDAY-FIRSTDAY;
 CARDS;
ARONTH, J. 12/01/73  4/30/81
DESTER, L.  7/14/61 12/31/81
HARLIN, M.  8/03/77  6/15/81
;
PROC PRINT;
/*
//
```

o Remember that SAS date and time values have implicit units.

SAS

CBS	NAME	FIRSTDAY	LASTDAY	DAYS
1	ARONTH, J.	5083	7790	2707
2	DESTER, L.	560	8035	7475
3	HARLIN, M.	6424	7836	1412

(6) Concatenating SAS Data Sets

Example : Combine the MALES and FEMALES data sets into one data set.

MALES DATA SET

OBS	DEPT	NAME	SEX
1	918	EPERT	M
2	917	HIGGINS	M
3	916	RICHARDS	M
4	914	RYAN	M

FEMALES DATA SET

OBS	DEPT	NAME	SEX
1	917	DOOB	F
2	914	LARSON	F
3	918	POWELL	F

```
DATA BOTH;
   SET MALES FEMALES;
PROC PRINT;
   TITLE 'RESULT OF THE CONCATENATION';
```

	DEPT	NAME	SEX
program data vector			

BOTH DATA SET

RESULT OF THE CONCATENATION

OBS	DEPT	NAME	SEX
1	918	EPERT	M
2	917	HIGGINS	M
3	916	RICHARDS	M
4	914	RYAN	M
5	917	DOOB	F
6	914	LARSON	F
7	918	POWELL	F

(7) Interleaving SAS Data Sets

Example : Combine the MALES and FEMALES data sets such that the resulting data set has its observations arranged in alphabetical order.

```
PROC SORT DATA=MALES;   BY NAME;
PROC SORT DATA=FEMALES; BY NAME;
```

MALES DATA SET

OBS	DEPT	NAME	SEX
1	918	EPERT	M
2	917	HIGGINS	M
3	916	RICHARDS	M
4	914	RYAN	M

FEMALES DATA SET

OBS	DEPT	NAME	SEX
1	917	DOOB	F
2	914	LARSON	F
3	918	POWELL	F

```
DATA BOTHSORT;
   SET MALES FEMALES;
   BY NAME;
PROC PRINT;
   TITLE 'RESULT OF INTERLEAVING';
```

program data vector

DEPT	NAME	SEX

BOTHSORT DATA SET

RESULT OF INTERLEAVING

OBS	DEPT	NAME	SEX
1	917	DOOB	F
2	918	EPERT	M
3	917	HIGGINS	M
4	914	LARSON	F
5	918	POWELL	F
6	916	RICHARDS	M
7	914	RYAN	M

(8) Match Merging

Data sets with unequal numbers of observations

Example : Suppose Larson is missing from the SALARY data set. (Note: Both data sets are already sorted by name.)

GENERAL DATA SET

OBS	DEPT	NAME	SEX
1	917	DOOB	F
2	918	EPERT	M
3	917	HIGGINS	M
4	914	LARSON	F
5	918	POWELL	F
6	916	RICHARDS	M
7	914	RYAN	M

SALARY DATA SET

OBS	NAME	NETPAY	GROSSPAY
1	DOOB	169.06	272.29
2	EPERT	224.36	310.40
3	HIGGINS	777.50	1235.46
4	POWELL	189.39	271.54
5	RICHARDS	219.27	352.84
6	RYAN	291.56	399.20

```
DATA MERGED;
   MERGE GENERAL SALARY;
   BY NAME;
PROC PRINT;
   TITLE 'MATCH MERGING';
   TITLE2 'UNEQUAL NUMBERS OF OBSERVATIONS';
```

	DEPT	NAME	SEX	NETPAY	GROSSPAY
program data vector					

MERGED DATA SET

MATCH MERGING
UNEQUAL NUMBERS OF OBSERVATIONS

OBS	DEPT	NAME	SEX	NETPAY	GROSSPAY
1	917	DOOB	F	169.06	272.29
2	918	EPERT	M	224.36	310.40
3	917	HIGGINS	M	777.50	1235.46
4	914	LARSON	F	.	.
5	918	POWELL	F	189.39	271.54
6	916	RICHARDS	M	219.27	352.84
7	914	RYAN	M	291.56	399.20

(9) Match Merging

Identical variable names

CLOTHES DATA SET

OBS	DATE	SALES
1	18OCT82	223.93
2	19OCT82	387.82
3	20OCT82	229.28
4	21OCT82	318.32
5	22OCT82	519.07

EQUIP DATA SET

OBS	DATE	SALES
1	18OCT82	492.28
2	19OCT82	228.20
3	20OCT82	542.98
4	21OCT82	325.02
5	22OCT82	733.60

```
DATA ALLSALES;
    MERGE CLOTHES EQUIP;
    BY DATE;

PROC PRINT;
    TITLE 'MERGING DATA SETS';
    TITLE2 'WITH IDENTICAL VARIABLE NAMES';
```

program data vector

DATE	SALES

ALLSALES DATA SET

MERGING DATA SETS
WITH IDENTICAL VARIABLE NAMES

OBS	DATE	SALES
1	18OCT82	492.28
2	19OCT82	228.20
3	20OCT82	542.98
4	21OCT82	325.02
5	22OCT82	733.60

(10) Match Merging

Identical variable names

Example : Repeat the preceding example using the RENAME data set option.

CLOTHES DATA SET

OBS	DATE	SALES
1	18OCT82	223.93
2	19OCT82	387.82
3	20OCT82	229.28
4	21OCT82	318.32
5	22OCT82	519.07

EQUIP DATA SET

OBS	DATE	SALES
1	18OCT82	492.28
2	19OCT82	228.20
3	20OCT82	542.98
4	21OCT82	325.02
5	22OCT82	733.60

```
DATA ALLSALES;
   MERGE CLOTHES (RENAME (SALES=CL_SALES))
         EQUIP   (RENAME=(SALES=EQ_SALES));
   BY DATE;
PROC PRINT;
   TITLE 'RESULT OF MERGING DATA SETS';
   TITLE2 'WITH IDENTICAL VARIABLE NAMES';
   TITLE3 'USING THE RENAME DATA SET OPTION';
```

program data vector

DATE	CL_SALES	EQ_SALES

ALLSALES DATA SET

RESULT OF MERGING DATA SETS WITH IDENTICAL VARIABLE NAMES USING THE RENAME DATA SET OPTION

OBS	DATE	CL_SALES	EQ_SALES
1	18OCT82	223.93	492.28
2	19OCT82	387.82	228.20
3	20OCT82	229.28	542.98
4	21OCT82	318.32	325.02
5	22OCT82	519.07	733.60

(11) UPDATE Application

Example : Add new employees and make changes to existing values for old employees in the PAYROLL data set.
(Note: Both data sets have already been sorted by Name.)

PAYROLL DATA SET

OBS	DEPT	NAME	SEX	NETPAY	GROSSPAY
1	917	DOOB	F	169.06	272.29
2	918	EPERT	M	224.36	310.40
3	917	HIGGINS	M	777.50	1235.46
4	914	LARSON	F	215.47	283.92
5	918	POWELL	F	189.39	271.54
6	916	RICHARDS	M	219.27	352.84
7	914	RYAN	M	291.56	399.20

```
DATA NEWINFO;
   INPUT DEPT 1-3 NAME $ 5-12 SEX $14
        NETPAY 16-22 GROSSPAY 24-30;
   CARDS;
    POWELL    221.75  310.62
916 SERPANT  M  207.22  398.65
    DOOB        191.65  252.57
918 ARCHER   F  315.17  420.00
PROC SORT DATA=NEWINFO;
   BY NAME;
PROC PRINT DATA=NEWINFO;
```

NEWINFO DATA SET

OBS	DEPT	NAME	SEX	NETPAY	GROSSPAY
1	918	ARCHER	F	315.17	420.00
2	.	DOOB		191.65	252.57
3	.	POWELL		221.75	310.62
4	916	SERPANT	M	207.22	398.65

(12) UPDATE Application

Example : Update the PAYROLL data set with the NEWINFO data set.

```
DATA PAYROLL2;
   UPDATE PAYROLL NEWINFO;
   BY NAME;
PROC PRINT;
   TITLE 'PAYROLL2 DATA SET';
```

PAYROLL2 DATA SET

OBS	DEPT	NAME	SEX	NETPAY	GROSSPAY
1	918	ARCHER	F	315.17	420.00
2	917	DOOB	F	191.65	252.57
3	918	EPERT	M	224.36	310.40
4	917	HIGGINS	M	777.50	1235.46
5	914	LARSON	F	215.47	283.92
6	918	POWELL	F	221.75	310.62
7	916	RICHARDS	M	219.27	352.84
8	914	RYAN	M	291.56	399.20
9	916	SERPANT	M	207.22	398.65

(13) SAS Array

Example: Create the 4 Variables with a SAS Array.

```
//SPSSAS30  JOB CLASS=V
//SAS        EXEC  SAS
//SAMPLE  DD DSN=SAS.SAMPLE,DISP=SHR
//SYSIN         DD  *
DATA CONVERT;
 INPUT N1-N5 ;
 ARRAY PERCENT N1-N5;
   TOTAL=SUM(OF N1-N5);

 DO OVER PERCENT;
   PERCENT=ROUND((PERCENT/TOTAL*100),1);
 END;
   TOTAL=SUM(OF N1-N5);
 CARDS;
11  5 26  9  49
19  5  4 36  37
43  3  9  8  36
 6 25 28 34   7
44 13  6  5  32
;
PROC PRINT;
 TITLE ARRAY SAMPLE;
/*
//
```

10:59 FRIDAY, NOVEMBER 16, 1

ARRAY SAMPLE

OBS	N1	N2	N3	N4	N5	TOTAL
1	11	5	26	9	49	100
2	19	5	4	36	37	101
3	43	3	9	8	36	99
4	6	25	28	34	7	100
5	44	13	6	5	32	100

CHAPTER 3. 기 초 통 계 처 리

CHAPTER 3. 기초통계처리

1. CHART

2. PLOT

3. Discriptive statistic

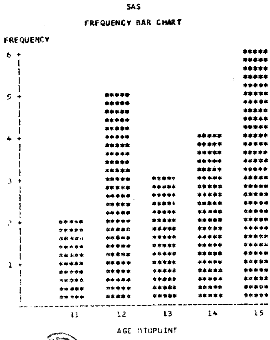
1) CORR	computes bivariate correlations and other measures of association for continuous variables.
2) FREQ	print tables of frequency counts, cross tabulations, and bivariate measures of association for categorial variables.
3) MEANS	computes and prints means and other descriptive statistics.
4) UNIVARIATE	computes univariate statistics including quantiles.

* Categorial data

1. CHART Procedure

(1) 기 본 형

```
PROC CHART;
VBAR AGE;
```



(2) 일반형

```
PROC  CHART options;
   BY        variables;
   VBAR      variables/options;
   HBAR      variables/options;
   BLOCK     variables/options;
   PIE       variables/options;
   STAR      variables/options;
```

(3) Options

1) Determining the values represented

TYPE = code (FREQ CFREQ PCT CPCT SUM or MEAN)

SUMVAR = variable

FREQ = variable

2) Grouping among and within bars

GROUP = variable

SUBGROUP = variable name

3) Classifying the observations into bars

DISCRETE MIDPOINTS LEVELS

4) Formatting the chart

NOSPACE SYMBOL = 'character' MISSING AXIS=values

NOSTAT FREQ CFREQ PERCENT CPERCENT SUM MEAN

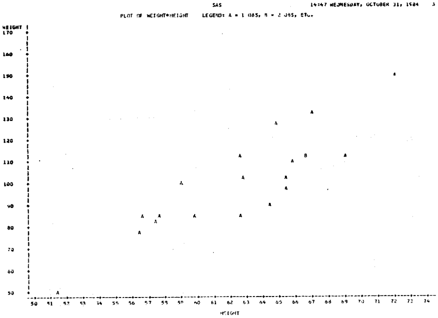
2. PLOT PROCEDURE

(1) 기본형

PROC PLOT;

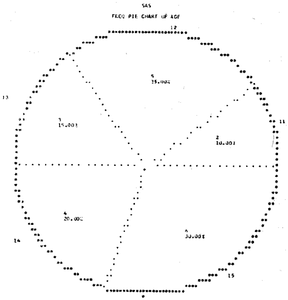
PLOT HEIGHT * WEIGHT;

1) PLOTS



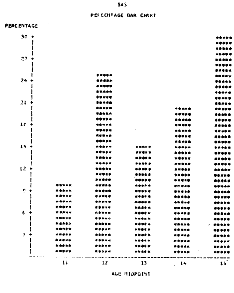
2) PROC CHART;

PIE AGE;



3) PROC CHART;

VBAR AGE;



(2) 일반형

```
PROC PLOT DATA=dataset UNIFORM NOLEGEND;
   BY    variables;
   PLOT  requests/options;
```

o request : vertical *horizontal=('character') or
  variable

o scale of axis options

  VAXIS HAXIS VZERO VREVERSE

o plot size options

  VPOS=n HPOS=n VSPACE=n HSPACE=n

o overlaying plots

  OVERLAY

o contour plots

  CONTOUR = n (n ranges 1 to n).

3. CORR procedure

(1) Product-moment Correlation (Pearson)

1) True product moment correlation

$$\rho_{xy} = \mathrm{cov}(x, y) / \sqrt{\mathrm{var}(x)\ \mathrm{var}(y)}$$
$$= E(x-E(x))\ (y-E(y)) / \sqrt{E(x-E(x))^2\ E(y-E(y)))^2}$$

2) Sample correlation estimates

$$= (x-\bar{x})(y-\bar{y}) / \sqrt{(x-\bar{x})^2\ \ (y-\bar{y})^2}$$

3) Spearman's rank order conrelation coefficient;

norparametric measure that is calculatears the correlation of the ranks of the data

$$\rho = \Sigma(r_i - \bar{r})\ (s_i - \bar{s})\ /\ \sqrt{\Sigma(r_i - \bar{r})\ \Sigma(s_i - \bar{s})}$$

4) Kendall's tau- τ

; measure calculated from concordances and discordances

$$\tau = \Sigma_{i\ j} Sgn(x_i - x_j)\ Sgn(y_i - y_j)\ /$$

$$\sqrt{((n(n-1/2 - t_i(t_i-1))\ (n(n-1)2- u_i(u_i-1)))}$$

(2) 기본형

```
PROC CORR;
     VAR HEIGHT WEIGHT;
     WITH AGE;
```

```
                                        SAS                  16:17 WEDNESDAY, OCTOBER 31, 1984   1
VARIABLE    N          MEAN        STD DEV            SUM         MINIMUM         MAXIMUM

HEIGHT     20   57.59500000     5.38128403   1251.90000300    51.30000000     72.00000000
WEIGHT     20  130.59500000    22.28136996   2311.90000000    50.50000000    156.00000000
AGE        20   13.40000000     1.50037694    268.00000000    11.00000000     16.00000000

          CORRELATION COEFFICIENTS / PROB > |R| UNDER HO:RHO=0 / N = 20
                              HEIGHT   WEIGHT

                    AGE      0.81600  0.73898
                              0.0001   0.0002
```

(3) 일반형

```
PROC CORR options;
  VAR      variables;
  WITH     variables;
  WEIGHT   variables;
  FREQ     variables;
  BY       variables;
```

(4) Options

SPEARMAN	BEST	NOMISS
KENDALL	NOSIMPLE	SSCP
PEARSON	NOPRINT	COV
RANK	NOPROB	NCORR

OUTP= SAS dataset: requests that CORR create a new SAS data set containing Pearson correlation

OUTS= SAS dataset: requests that CORR create a new SAS data set containing Spearman correlations

OUTK= SAS dataset: requests that CORR create a new SAS data set containing Kendall correlations.

4. FREQ procedure. (discrete variable에 사용)

(1) 기본형

PROC FREQ

TABLE AGE *SEX;

```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SAMPLE     DD  DSN=SAS.SAMPLE,DISP=SHR
//SYSIN         DD  *
DATA SAMPLE;
 INPUT NAME $ 1-10 SEX $ 11 AGE 13-14
      HEIGHT 16-19 WEIGHT 21-25;
CARDS;
KYUNGEUI  M 15 65.7 110.9
JOHN      M 12 59.0  99.5
JAMES     M 12 57.3  83.0
ALFERD    M 14 69.0 112.5
WILLIAM   M 15 66.5 112.0
JEFFREY   M 13 62.5  84.0
RONALD    M 15 67.0 133.0
THOMAS    M 11 57.5  85.5
PHILIP    M 16 72.0 150.0
ROBELT    M 12 64.8 128.0
HENRY     M 14 65.3 102.5
JANET     F 15 62.5 112.5
JOYCE     F 11 51.3  50.5
JUDY      F 14 64.3  90.0
CAROL     F 14 62.8 102.5
JANE      F 12 59.8  84.5
LOUISE    F 12 56.3  77.0
BARBARA   F 13 65.3  98.0
MARY      F 15 66.5 112.0
ALICE     F 13 56.5  84.0
;
PROC FREQ;
    TABLES AGE*SEX;
/*
//
```

```
                         SAS

                 TABLE OF AGE BY SEX

AGE          SEX

FREQUENCY|
 PERCENT |
 ROW PCT |
 COL PCT |F        |M        |  TOTAL
---------+---------+---------+
11       |       1 |       1 |      2
         |    5.00 |    5.00 |  10.00
         |   50.00 |   50.00 |
         |   11.11 |    9.09 |
---------+---------+---------+
12       |       2 |       3 |      5
         |   10.00 |   15.00 |  25.00
         |   40.00 |   60.00 |
         |   22.22 |   27.27 |
---------+---------+---------+
13       |       2 |       1 |      3
         |   10.00 |    5.00 |  15.00
         |   66.67 |   33.33 |
         |   22.22 |    9.09 |
---------+---------+---------+
14       |       2 |       2 |      4
         |   10.00 |   10.00 |  20.00
         |   50.00 |   50.00 |
         |   22.22 |   18.18 |
---------+---------+---------+
15       |       2 |       3 |      5
         |   10.00 |   15.00 |  25.00
         |   40.00 |   60.00 |
         |   22.22 |   27.27 |
---------+---------+---------+
16       |       0 |       1 |      1
         |    0.00 |    5.00 |   5.00
         |    0.00 |  100.00 |
         |    0.00 |    9.09 |
---------+---------+---------+
TOTAL            9        11       20
             45.00     55.00   100.00
```

(2) 일반형

```
PROC FREQ options;
  TABLES requests/options;
  WEIGHT variables;
  BY     variables;
```

(3) TABLES Options

EXPECTED — requests that the expected cell frequency under the hypothesis of independence (or homogeneity) be printed.

DEVIATION — requests that FREQ print deviation of the cell frequency from the expected value.

CELLCHI2 — requests that FREQ print the cell's contribution to the total x^2 statistic. This is compulted as (frequency-expected)**2/expected and is approximately distributed x^2 with 1 df.

CHISQ — requests a chi-square(x^2) test of homogeneity or independence for each two-way table requested in a TABLES statement. For 2 by 2 tables, Fisher's Exact Test is performed. The formula for x^2 is given in Chapter17, "Introduction to Descriptive Statistics."

ALL requests the basic set of measures of association popularized by Goodman and Kruskal for two-way tables, including some of the standard errors, the contingency coefficient, Cramer's V, gamma, Kendall's tau-ζ, Stuart's tau-c, Somer's D, and lambda asymmertric and Spearman correlations. Of course, not all statistics are appropriate for the data in a given table, (See ALL Option: Measures of Association below.)

NOFREQ suppresses printing the cell frequencies for a crosstabulation.

NOPERCENT suppresses printing the cell percentages for a crosstabulation.

NOROW suppresses printing the row percentages in cells of a crosstabulation.

NOCOL suppresses printing the column percentages in cells of a crosstabulation.

CUMCOL requests cumulative column percentages be printed in the cells.

LIST prints two-way to n-way tables in a list format rather than as crosstabulation tables. Expected cell frequencies are not printed when LIST is specified, even if EXPECTED is specified.

NOCUM suppresses the cumulative frequencies, percentages, and cumulative percentage columns for one-way frequencies and frequencies in list format when the LIST option is included.

MISSING asks FREQ to consider missing values like other values in calculations of percents and other statistics.

SPARSE causes the procedure to write out or print information about all possible combinations of levels of the variables in the table request, even when some combinations of levels do not occur in the data. This option affects printouts under the LIST option and output data sets.

NOPRINT suppresses all printed output except that controlled by CHISQ and ALL.

5. MEANS procedure.

(1) 기본형

```
PROC MEANS;
   VAR HEIGHT WEIGHT;
   BY SEX;
```

SAS 17:51 THURSDAY, NOVEMBER 15, 1984 1

VARIABLE	N	MEAN	STANDARD DEVIATION	MINIMUM VALUE	MAXIMUM VALUE	STD ERROR OF MEAN	SUM	VARIANCE	C.V.
SEX=F									
HEIGHT	9	60.58838889	5.01832752	51.30000000	66.50000000	1.67277584	565.30000000	25.18361111	8.283
WEIGHT	9	90.11111111	19.38391372	50.50000000	112.50000000	6.46130457	811.00000000	375.73611111	21.511
SEX=M									
HEIGHT	11	64.23636364	4.72488576	57.30000000	72.00000000	1.42460666	706.6000000	22.32454545	7.355
WEIGHT	11	109.17272727	21.51344189	83.30000000	150.00000000	6.48654679	1200.9000000	462.82818182	19.706

(2) 일반형

```
PROC MEANS options;
   BY variables;
   VAR variables;
   FREQ variables;
   ID variables;
   OUTPUT OUT = SAS dataset keyword = names...;
```

(3) PROC MEANS options;

N	the number of observations on which calculations are based
NMISS	the number of missing values
MEAN	the mean
STD	the standard deviation
MIN	the smallest value
MAX	the largest value
RANGE	the range
SUM	the sum
VAR	the variance
USS	the uncorrected sum of squares
CSS	the corrected sum of squares
STDERR	the standard error of the mean
CV	the coefficient of variation (percent)
SKEWNESS	the measure of skewness
KURTOSIS	the measure of kurtosis
T	the student's t value for testing the hypothesis that the population mean is 0
PRT	the probability of a greater absolute value of Student's t.

(4) STATISTICS

$$VAR = \Sigma(x-\overline{x})/n-1$$

$$SKEWNESS = m_3/m_2^{2/2}$$

$$KURTOSIS = m_4/m_2^2$$

$$T = \overline{x}/\sqrt{VAR/N}$$

$$PRT = Pro(|t| > T)$$

6 . UNIVARIATE procedure

(1) 기본형

PROC UNIVARIATE ;

VAR HEIGHT AGE ;

BY SEX ;

UNIVARIATE procedure에서는 다른 기초통계 procedure 에서 계산되지 않는 백분위수가 계산된다.

```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SAMPLE     DD  DSN=SAS.SAMPLE,DISP=SHR
//SYSIN        DD  *
DATA SAMPLE;
 INPUT NAME $ 1-10 SEX $ 11 AGE 13-14
       HEIGHT 16-19 WEIGHT 21-25;
CARDS;
KYUNGEUI   M 15 65.7 110.9
JOHN       M 12 59.0  99.5
JAMES      M 12 57.3  83.0
ALFERD     M 14 69.0 112.5
WILLIAM    M 15 66.5 112.0
JEFFREY    M 13 62.5  84.0
RONALD     M 15 67.0 133.0
THOMAS     M 11 57.5  85.5
PHILIP     M 16 72.0 150.0
ROBELT     M 12 64.8 128.0
HENRY      M 14 65.3 102.5
JANET      F 15 62.5 112.5
JOYCE      F 11 51.3  50.5
JUDY       F 14 64.3  90.0
CAROL      F 14 62.8 102.5
JANE       F 12 59.8  84.5
LOUISE     F 12 56.3  77.0
BARBARA    F 13 65.3  98.0
MARY       F 15 66.5 112.0
ALICE      F 13 56.5  84.0
;
PROC UNIVARIATE FREQ PLOT NORMAL;
     VAR HEIGHT AGE;
/*
//
```

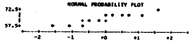
UNIVARIATE

VARIABLE=HEIGHT

```
                MOMENTS                                        QUANTILES(DEF=4)                          EXTREMES

N               11   SUM WGTS          11      100% MAX      72     99%     72            LOWEST    HIGHEST
MEAN       64.2364   SUM            706.6       75% Q3       67     95%     72              57.3       65.7
STD DEV    4.72489   VARIANCE     22.3245       50% MED    65.3     90%   71.4              57.5       66.5
SKEWNESS -0.207734   KURTOSIS   -0.718587       25% Q1       59     10%  57.34                59         67
USS        45617.7   CSS          223.245        0% MIN    57.3      5%   57.3              62.5         69
CV         7.35547   STD MEAN     1.42461                            1%   57.3              64.8         72
T:MEAN=0   45.0906   PROB>|T|      0.0001      RANGE       14.7
SGN RANK        33   PROB>|S| 0.00385729       Q3-Q1          8
NUM ^= 0        11                             MODE        57.3
W:NORMAL  0.940714   PROB<W         0.502

       STEM LEAF                  #
          7 2                     1
          6 556679                6
          6 2                     1
          5 779                   3
            ----+----+----+----+
            MULTIPLY STEM.LEAF BY 10**+01
```

BOXPLOT

NORMAL PROBABILITY PLOT



FREQUENCY TABLE

```
              PERCENTS                     PERCENTS                     PERCENTS                     PERCENTS
VALUE COUNT  CELL   CUM   VALUE COUNT  CELL   CUM   VALUE COUNT  CELL   CUM   VALUE COUNT  CELL   CUM
 57.3     1   9.1   9.1    62.5     1   9.1  36.4    65.7     1   9.1  63.6      69     1   9.1  90.9
 57.5     1   9.1  18.2    64.8     1   9.1  45.5    66.5     1   9.1  72.7      72     1   9.1 100.0
   59     1   9.1  27.3    65.3     1   9.1  54.5      67     1   9.1  81.8
```

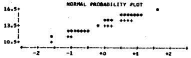

UNIVARIATE

VARIABLE=AGE

```
                MOMENTS                                        QUANTILES(DEF=4)                          EXTREMES

N               11   SUM WGTS          11      100% MAX      16     99%     16            LOWEST    HIGHEST
MEAN       13.5455   SUM              149       75% Q3       15     95%     16                11         14
STD DEV    1.63485   VARIANCE     2.67273       50% MED      14     90%   15.8                12         15
SKEWNESS-0.0084787   KURTOSIS    -1.36934       25% Q1       12     10%   11.2                12         15
USS           2045   CSS          26.7273        0% MIN      11      5%     11                12         15
CV         12.0693   STD MEAN    0.492925                            1%     11                13         16
T:MEAN=0   27.4797   PROB>|T|      0.0001      RANGE          5
SGN RANK        33   PROB>|S| 0.00370152       Q3-Q1          3
NUM ^= 0        11                             MODE          12
W:NORMAL   0.92463   PROB<W         0.403

       STEM LEAF                  #
         16 0                     1
         15 000                   3
         14 00                    2
         13 0                     1
         12 000                   3
         11 0                     1
         10
            ----+----+----+----+
```

BOXPLOT

NORMAL PROBABILITY PLOT



FREQUENCY TABLE

```
              PERCENTS                     PERCENTS                     PERCENTS
VALUE COUNT  CELL   CUM   VALUE COUNT  CELL   CUM   VALUE COUNT  CELL   CUM
   11     1   9.1   9.1      13     1   9.1  45.5      15     3  27.3  90.9
   12     3  27.3  36.4      14     2  18.2  63.6      16     1   9.1 100.0
```

(2) STATISTICS

N	the number of observations on which the calculations were based
NMISS	the number of missing balues
NOBS	the number of observations
MEAN	the mean
SUM	the sum
STD	the standard deviation
VAR	the variance
SKEWNESS	skewness
KURTOSIS	kurtosis
SUMWGT	the sum of the weights
MAX	the largest value
MIN	the smallest value
RANGE	the range
Q3	the upper quantile or the seventy-fifth percentile
MEDIAN	the median or the fiftieth percentile
Q1	the lower quartile or the twenty-fifth percentile
QRANGE	the difference between the upper and lower quartiles, that is, Q3-Q1
P1	the first percentile
P5	the fifth percentile
P10	the tenth percentile
P90	the ninetieth percentile
P95	the ninety-fifth percentile

P99	the ninety-ninth percentile
MODE	the most frequent value. If the mode is not unique, the smallest mode is used

(3) 일반형

```
PROC UNIVARIATE options;
   VAR variables;
   BY variables;
   FREQ variables;
   WEIGHT variables;
   ID variables;
   OUTPUT OUT = SAS data set key word = names,...;
```

(4) PROC UNIVARIATE options;

NOPRINT	suppresses all printed output. NOPRINT can be used when the only purpose for executing the procedure is to create new data sets.
PLOT	causes UNIVARIATE to produce a stem-and-leaf plot (or a horizontal bar char), a box plot, and a normal probability plot.
FREQ	requests a frequency table consisting of the variable values, frequencies, percentages, and cumulative percentages.

NORMAL causes UNIVARIATE to compute a test statistic for the hypothesis that the input data come from a normal distribution. The probability of a more extreme value of the test statistic is also printed.

DEF=value specifies which of the four definitions given below in the section Computational Method is to be used to calculate percentiles. The DEF value may be 1,2,3,4, or 5, If DEF= is omitted, definition 4 is used.

CHAPTER 4. 회귀분석과 분산분석에 관련된 기법

CHAPTER 4. 회귀분석과 분산분석에 관련된 기법

A. 회귀분석 (Regression Analysis)

1. 회귀분석

(1) 회귀분석은 종속변수 (dependent variable)와 독립변수 (independent variables)들간의 관계를 알아보기 위해 사용된다. 독립변수는 설명변수 (explaratory variable)라고도 한다. SAS의 REG procedure를 이용하면 선형 회귀분석을 할 수 있다.

(2) 선형회귀모형 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon$

2. REG procedure

(1) PROC REG options;

```
    MODEL dependents = independents / options;
    WEIGHT variable;
    ID variable;
    OUTPUT OUT = SAS dataset keyword = names...;
    TEST linear equations,...;
```

(2) PROC REG Statement

대표적인 예로는

PROC REG DATA = A;	SAS dataset A를 이용한다.
PROC REG;	가장 최근에 만들어진 SAS dataset를 이용 표준화된 printout을 만든다.
PROC REG ALL;	가장 자세한 printout을 만든다.

(3) MODEL statement의 대표적인 예로는

MODEL Y=X1 X2/XPX I ;

선형모형 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$과 관련된 각종 통계계산을 하고 option으로 X'X와 (X'X)를 printout 한다.

MODEL Y=X1 X2 /COLLIN ;

multicollinearity 분석을 option으로 수행한다.

MODEL Y=X1 X2 / PARTIAL

종속변수와 독립변수의 partial scatter plot을 만든다.

(4) WEIGHT variable; 가중 최소자승법에 의한 회귀분석을 한다.

(5) ID variable; 각 관측치를 표식하는 변수를 이용할 때 사용한다.

(6) OUTPUT statement의 대표적인 예로는

OUTPUT OUT = B

PREDICTED = YHAT

RESIDUAL = YRESID ;

종속변수의 predicted values을 YHAT, 잔여분 (residual)을 YRESID로 이름짓고 SAS dataset B에 넣는다.

(7) TEST statement의 대표적인 예로는

MODEL Y = X1 X2 X3 X4 ;

선형모형 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$에 관련된 회귀분석에서 통계적가설 $\beta_3 = \beta_4 = 0$을 검정한다.

※ 그 밖의 자세하고 전반적인 소개와 설명을 SAS User's Guide ; Statistics pp. 39 - 83을 참조할 것.

(8) Sample

다음은 83년도 도시가계조사 대상중 15가구에 대한 X(소속), Y1(소비지출), Y2(교육교양오락비)의 자료이다.

```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SYSIN         DD  *
DATA HOUSE;
 INPUT X Y1 Y2;
CARDS;
66618 118319 8534
122469 136736 9432
 171358 157592 10050
219049 184560 13376
270043 208717 15605
319007 243169 24881
370755 271723  26245
420115  305076 34969
472620 339056 38686
519782  355504 34544
572612 383370 42598
 640551 423327 53675
741887 461573 54641
844780 520846 53965
1221396 666157 91260
;
PROC REG ;
     MODEL Y1=X;
PROC REG ;
     MODEL Y2=X;
/*
//
```

DEP VARIABLE: Y1

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	333741990533	333741990533	1692.873	0.0001
ERROR	13	2562888776	197145290		
C TOTAL	14	336304879309			

ROOT MSE	14040.844	R-SQUARE	0.9924
DEP MEAN	318382	ADJ R-SQ	0.9918
C.V.	4.410067		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	85641.266	6718.686	12.747	0.0001
X	1	0.500658	0.012168	41.145	0.0001

DEP VARIABLE: Y2

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	1	7114011490	7114011490	449.167	0.0001
ERROR	13	205897037	15838234		
C TOTAL	14	7319908527			

ROOT MSE	3979.728	R-SQUARE	0.9719
DEP MEAN	34164.067	ADJ R-SQ	0.9697
C.V.	11.64887		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	184.047	1904.340	0.097	0.9245
X	1	0.073096	0.003448964	21.194	0.0001

DEP VARIABLE: OXY

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	6	202.659	33.776418	4.860	0.0175
ERROR	9	62.553917	6.950435		
C TOTAL	15	265.212			

ROOT MSE	2.636368	R-SQUARE	0.7641
DEP MEAN	46.284937	ADJ R-SQ	0.6069
C.V.	5.695952		

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	49.360100	26.377029	1.871	0.0941
RUNTIME	1	-0.748050	0.443428	-1.687	0.1259
AGE	1	-0.218662	0.142014	-1.540	0.1580
WEIGHT	1	0.103334	0.099252	1.041	0.3250
RUNPULSE	1	-0.314035	0.279587	-1.123	0.2904
MAXPULSE	1	0.454630	0.343396	1.324	0.2182
RSTPULSE	1	-0.331326	0.179302	-1.848	0.0977

COLLINEARITY DIAGNOSTICS VARIANCE PROPORTIONS

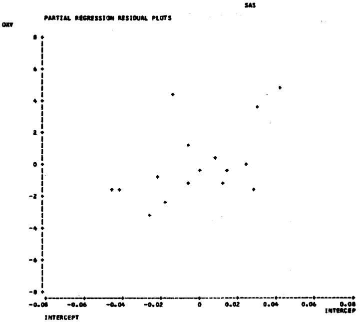
NUMBER	EIGENVALUE	CONDITION INDEX	PORTION INTERCEP	PORTION RUNTIME	PORTION AGE	PORTION WEIGHT	PORTION RUNPULSE	PORTION MAXPULSE	PORTION RSTPULSE
1	6.939	1.000	0.0000	0.0004	0.0002	0.0001	0.0000	0.0000	0.0001
2	0.031822	14.767	0.0008	0.3719	0.0299	0.0183	0.0001	0.0001	0.0100
3	0.015039	21.480	0.0000	0.0626	0.3548	0.1035	0.0000	0.0001	0.0430
4	0.009915	26.455	0.0001	0.2055	0.0001	0.3260	0.0013	0.0006	0.1389
5	0.003443	44.895	0.0227	0.0894	0.4816	0.1542	0.0097	0.0040	0.4079
6	0.0005113	116.497	0.8201	0.1042	0.1273	0.3458	0.0818	0.0105	0.2251
7	.000074964	304.248	0.1564	0.1661	0.0062	0.0520	0.9070	0.9847	0.1750

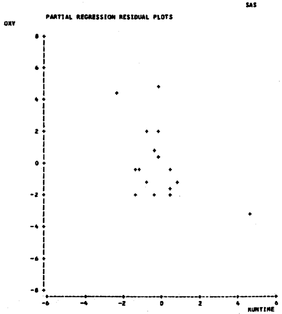
(9) Sample

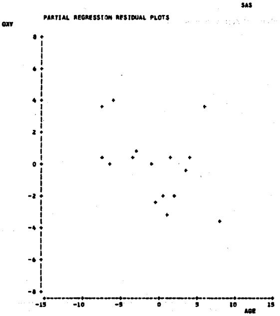
Aerobic Fitness Prediction :

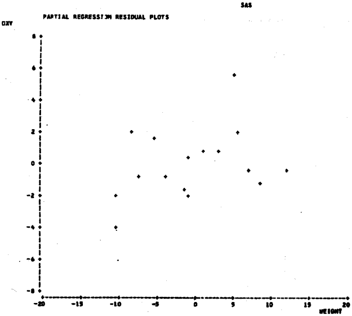
Aerobic fitness는 산소 소비능력으로 측정되는데 이 측정방법은 비용이 많이 들고 불편하므로 그 대신 간단한 운동측정으로 산소 소비능력을 예측하고자 한다. 회귀분석에 사용된 종속변수는 OXY (oxgen uptake rate 산소소비율), 설명변수로는 RUNTIME (1.5 마일 주행시간), AGE (나이), WEIGHT (체중), RUNPULSE (주행중의 맥박수), MAXPULSE (주행중의 최대맥박수), RSTPULSE (휴식중의 맥박수) 포함되었다.

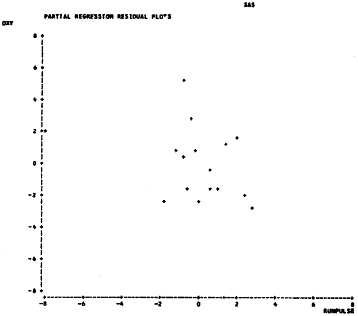
```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SYSIN         DD  *
DATA FITNESS;
 INPUT AGE WEIGHT OXY RUNTIME RSTPULSE
       RUNPULSE MAXPULSE;
CARDS;
44 89.47 44.609 11.37 62 178 182 40 75.07 45.313 10.07 62 185 185
44 85.84 54.297  8.65 45 156 168 42 68.15 59.571  8.17 40 166 172
38 89.02 49.874  9.22 55 178 180 47 77.45 44.811 11.63 58 176 176
40 75.98 45.681 11.95 70 176 180 43 81.19 49.091 10.85 64 162 170
44 81.42 39.442 13.08 63 174 176 38 81.87 60.055  8.63 48 170 186
44 73.03 50.541 10.13 45 168 168 45 87.66 37.388 14.03 56 186 192
45 66.45 44.754 11.12 51 176 176 47 79.15 47.273 10.60 47 162 164
54 83.12 51.855 10.33 50 166 170 49 81.42 49.156  8.95 44 180 185
51 69.63 40.836 10.95 57 168 172 51 77.91 46.672 10.00 48 162 168
48 91.63 46.774 10.25 48 162 164 49 73.37 50.388 10.08 67 168 168
57 73.37 39.407 12.63 58 174 176 54 79.38 46.080 11.17 62 156 165
52 76.32 45.441  9.63 48 164 166 50 70.87 54.625  8.92 48 146 155
51 67.25 45.118 11.08 48 172 172 54 91.63 39.203 12.88 44 168 172
51 73.71 45.790 10.47 59 186 188 57 59.08 50.545  9.93 49 148 155
49 76.32 48.673  9.40 56 186 188 48 61.24 47.920 11.50 52 170 176
52 82.78 47.467 10.50 53 170 172
;
PROC REG OUTEST=EST;
     MODEL OXY=RUNTIME AGE WEIGHT RUNPULSE
            MAXPULSE RSTPULSE/PARTIAL COLLIN;
/*
//
```

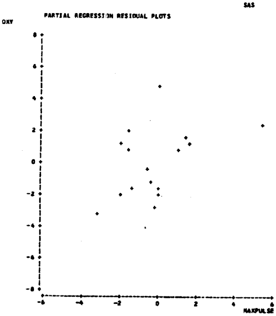

SAS
PARTIAL REGRESSION RESIDUAL PLOTS
OXY
8
6
4
2
0
-2
-4
-6
-8
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
INTERCEP
INTERCEPT


SAS
PARTIAL REGRESSION RESIDUAL PLOTS
OXY
8
6
4
2
0
-2
-4
-6
-8
-6
-4
-2
0
2
4
6
RUNTIME

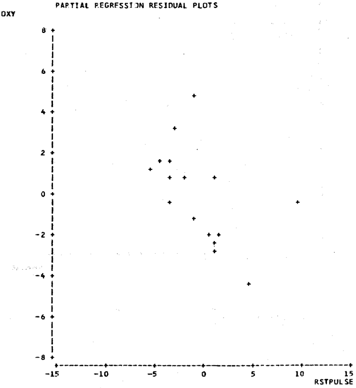

SAS
PARTIAL REGRESSION RESIDUAL PLOTS
OXY
8
6
4
2
0
-2
-4
-6
-8
-15
-10
-5
0
5
10
15
AGE


SAS
PARTIAL REGRESSION RESIDUAL PLOTS
OXY
8
6
4
2
0
-2
-4
-6
-8
-20
-15
-10
-5
0
5
10
15
20
WEIGHT


SAS
PARTIAL REGRESSION RESIDUAL PLOTS
OXY
RUNPULSE


SAS
PARTIAL REGRESSION RESIDUAL PLOTS
OXY
MAXPULSE

PARTIAL REGRESSION RESIDUAL PLOTS



3. STEPWISE procedure

(1) PROC STEPWISE option;

```
MODEL dependents = independent/options;
WEIGHT variable;
BY variable;
```

(2) PROC STEPWISE option

NOINT	prevents the procedure from automatically including an intercept term in the model.
FORWARD F	requests the forward-selection technique.
BACKWARD B	requests the backward-elimination technique.
STEPWISE	requests the stepwise technique, the default.
MAXR	requests the maximum R^2 improvement technique.
MINR	requests the minimum R^2 improvement technique.
SLENTRY = value SLE = value	specifies the significance level for entry into the model used in the forward-selection and stepwise techniques. If SLENTRY = is omitted, STEPWISE uses the SLENTRY = value .50 for forward selection, .15 for stepwise.

SLSTAY = value
SLS = value
: specifies the significance level for staying in the model for the backward elimination and stepwise techniques. If it is omitted, STEPWISE uses the SLSTAY = value .10 for backward elimination, .15 for stepwise.

INCLUDE = n
: forces the first n in dependent variables always to be included in the model. The selection techniques are performed on the other variables in the MODEL statement.

START = s
: is used to begin the comparing-and-switching process for a model containing the first s independent variables in the MODEL statement, where s is the START value. Consequently, no model is evaluated that contains fewer than s variables. This applies only to the MAXR or MINR methods.

STOP = s
: causes STEPWISE to stop when it has found the "best" s-variable model, where s is the STOP value. This applies only to the MAXR or MINR methods.

(3) Sample

```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SYSIN        DD  *
DATA FITNESS;
 INPUT AGE WEIGHT OXY RUNTIME RSTPULSE
      RUNPULSE MAXPULSE;
CARDS;
44 89.47 44.609 11.37 62 178 182 40 75.07 45.313 10.07 62 185 185
44 85.84 54.297  8.65 45 156 168 42 68.15 59.571  8.17 40 166 172
38 89.02 49.874  9.22 55 178 180 47 77.45 44.811 11.63 58 176 176
40 75.98 45.681 11.95 70 176 180 43 81.19 49.091 10.85 64 162 170
44 81.42 39.442 18.08 63 174 176 38 81.87 60.055  8.63 48 170 186
44 73.03 50.541 10.13 45 168 168 45 87.66 37.388 14.03 56 186 192
45 66.45 44.754 11.12 51 176 176 47 79.15 47.273 10.60 47 162 164
54 83.12 51.855 10.33 50 166 170 49 81.42 49.156  8.95 44 180 185
51 69.63 40.836 10.95 57 168 172 51 77.91 46.672 10.00 48 162 168
48 91.63 46.774 10.25 48 162 164 49 73.37 50.388 10.08 67 168 168
57 73.37 39.407 12.63 58 174 176 54 79.38 46.080 11.17 62 156 165
52 76.32 45.441  9.63 48 164 166 50 70.87 54.625  8.92 48 146 155
51 67.25 45.118 11.08 48 172 172 54 91.63 39.203 12.88 44 168 172
51 73.71 45.790 10.47 59 186 188 57 59.08 50.545  9.93 49 148 155
49 76.32 48.673  9.40 56 186 188 48 61.24 47.920 11.50 52 170 176
52 82.78 47.467 10.50 53 170 172
;
PROC STEPWISE;
     MODEL OXY=RUNTIME AGE WEIGHT RUNPULSE
            MAXPULSE RSTPULSE/FORWARD BACKWARD MAXR;
/*
//
```

SAS

FORWARD SELECTION PROCEDURE FOR DEPENDENT VARIABLE OXY

STEP 1 VARIABLE RUNTIME ENTERED R SQUARE = 0.51078375 C(P) = 6.66735310

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	1	135.46619793	135.46619793	14.62	0.0019
ERROR	14	129.74622900	9.26758779		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	81.64683345				
RUNTIME	-1.39844296	0.36577384	135.46619793	14.62	0.0019

STEP 2 VARIABLE WEIGHT ENTERED R SQUARE = 0.60399606 C(P) = 5.11058824

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	2	160.18726178	80.09363089	9.91	0.0024
ERROR	13	105.02516516	8.07885886		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	48.28695194				
RUNTIME	-1.33949135	0.34317170	123.07815246	15.23	0.0018
WEIGHT	0.16201964	0.09262092	24.72106384	3.06	0.1038

STEP 3 VARIABLE AGE ENTERED R SQUARE = 0.64683145 C(P) = 5.47608956

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	3	171.54773904	57.18257968	7.33	0.0047
ERROR	12	93.66468790	7.80539066		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	59.95801942				
RUNTIME	-1.37103355	0.33832785	128.17854038	16.42	0.0016
AGE	-0.17543689	0.14541860	11.36047726	1.46	0.2509
WEIGHT	0.12445818	0.09621645	13.05999045	1.67	0.2202

STEP 4 VARIABLE RSTPULSE ENTERED R SQUARE = 0.71725553 C(P) = 4.78887076

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	4	190.22507929	47.55626982	6.98	0.0048
ERROR	11	74.98734764	6.81703160		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	69.56755177				
RUNTIME	-1.02647845	0.37855301	50.12352050	7.35	0.0202
AGE	-0.23244101	0.14019594	18.73915164	2.75	0.1255
WEIGHT	0.12316517	0.08992202	12.78907088	1.88	0.1981
RSTPULSE	-0.19405853	0.11772252	18.67734026	2.74	0.1261

STEP 5 VARIABLE MAXPULSE ENTERED R SQUARE = 0.73107392 C(P) = 6.26159313

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	5	193.88988831	38.77797766	5.44	0.0113
ERROR	10	71.32253862	7.13225386		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	53.27960122				
RUNTIME	-0.89271791	0.42982434	30.76615923	4.31	0.0645
AGE	-0.22550054	0.14372717	17.55675515	2.46	0.1477
WEIGHT	0.14002908	0.09493859	15.51592637	2.18	0.1710
MAXPULSE	0.10425673	0.14544267	3.66480902	0.51	0.4899
RSTPULSE	-0.28827482	0.17743354	18.82646338	2.64	0.1353

NO OTHER VARIABLES MET THE 0.5000 SIGNIFICANCE LEVEL FOR ENTRY INTO THE MODEL.

SAS

FORWARD SELECTION PROCEDURE FOR DEPENDENT VARIABLE OXY

STEP 6 VARIABLE RUNPULSE ENTERED R SQUARE = 0.76413655 C(P) = 7.00000000

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	6	202.65850968	33.77641828	4.86	0.0175
ERROR	9	62.55391725	6.95043525		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	49.36009962				
RUNTIME	-0.74804963	0.44342815	19.78000062	2.85	0.1259
AGE	-0.21866196	0.14201393	16.47769907	2.37	0.1580
WEIGHT	0.10333411	0.09925160	7.53397871	1.08	0.3250
RUNPULSE	-0.31403453	0.27958750	8.76862137	1.26	0.2904
MAXPULSE	0.45463012	0.34339635	12.18251832	1.75	0.2182
RSTPULSE	-0.33132614	0.17930198	23.73303169	3.41	0.0977

STEP 0 ALL VARIABLES ENTERED R SQUARE = 0.76413655 C(P) = 7.00000000

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	6	202.65850968	33.77641828	4.86	0.0175
ERROR	9	62.55391725	6.95043525		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	49.36009962				
RUNTIME	-0.74804963	0.44342815	19.78000062	2.85	0.1259
AGE	-0.21866196	0.14201393	16.47769907	2.37	0.1580
WEIGHT	0.10333411	0.09925160	7.53397871	1.08	0.3250
RUNPULSE	-0.31403453	0.27958750	8.76862137	1.26	0.2904
MAXPULSE	0.45463012	0.34339635	12.18251832	1.75	0.2182
RSTPULSE	-0.33132614	0.17930198	23.73303169	3.41	0.0977

STEP 1 VARIABLE WEIGHT REMOVED R SQUARE = 0.73572922 C(P) = 6.08395783

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	5	195.12453097	39.02490619	5.57	0.0104
ERROR	10	70.08789596	7.00878960		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	62.87457464				
RUNTIME	-0.78284870	0.44401894	21.78691330	3.11	0.1084
AGE	-0.26148263	0.13649723	25.72059967	3.67	0.0844
RUNPULSE	-0.40984946	0.26511305	16.75056903	2.39	0.1532
MAXPULSE	0.52655465	0.33778480	17.03137533	2.43	0.1501
RSTPULSE	-0.31410422	0.17928524	21.51302479	3.07	0.1103

STEP 2 VARIABLE RUNPULSE REMOVED R SQUARE = 0.67257013 C(P) = 6.49396072

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	4	178.37396194	44.59349049	5.65	0.0101
ERROR	11	86.83846499	7.89440591		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	75.63600249				
RUNTIME	-1.01260690	0.44403645	41.06128972	5.20	0.0435
AGE	-0.29374928	0.14316093	33.23717104	4.21	0.0648
MAXPULSE	0.05109850	0.14824402	0.93795353	0.12	0.7368
RSTPULSE	-0.24213863	0.18374957	13.70865793	1.74	0.2144

STEP 3 VARIABLE MAXPULSE REMOVED R SQUARE = 0.66903354 C(P) = 4.62890961

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB>F
REGRESSION	3	177.43600842	59.14533614	8.09	0.0033
ERROR	12	87.77641852	7.31470154		
TOTAL	15	265.21242694			

	B VALUE	STD ERROR	TYPE II SS	F	PROB>F
INTERCEPT	82.73541290				
RUNTIME	-1.07499032	0.39040747	55.45867543	7.58	0.0175
AGE	-0.29308149	0.13779178	33.09228480	4.52	0.0548
RSTPULSE	-0.19625927	0.12193934	18.94825983	2.59	0.1335

ALL VARIABLES IN THE MODEL ARE SIGNIFICANT AT THE 0.1000 LEVEL.

B. 분산분석 (Analysis of Variance)

1. 분산분석

(1) 분산분석은 관측시 사이의 분산을 각 요인에 따른 분산으로 나누는 방법이다. 분산분석의 방법은 실험 (Experiment)이 어떻게 계획되었는가에 따라 다르며 여기에서는 Completely Randomized Design, Randomized Blocks Design에 대하여 설명하기로 한다.

Completely Randomized Design (CRD)

(2) 모형 $y_{ij} = \mu + \tau_i + \epsilon_{ij}$

- y_{ij}: 반응 (response)
- μ: 평균
- τ_i: 처리 i의효과 ($\sum_{i=1}^{t} \tau_i = 0$)
- ϵ_{ij}: 오차

(3) CRD는 t종류의 처리 (treatment)의 효과를 알아보기 위한 실험계획이며 각 처리에 n개의 subject를 임의로 배치한다.

2. ANOVA procedure.

(1) PROC ANOVA;

CLASS treatment variable;

MODEL response = treatment variable;

(2) SAS EXAMPLE

PROC ANOVA;

CLASS BRAND;

MODEL WEAR = BRAND;

MEANS BRAND/DUNCAN;

```
//SPCSAS30  JOB CLASS=V
//SAS         EXEC  SAS
//SAMPLE   DD DSN=SAS.SAMPLE,DISP=SHR
//SYSIN          DD  *
DATA;
 INPUT BRAND$ WEAR;
CARDS;
ACME  2.3
ACME  2.4
ACME  2.1
ACME  2.5
CHAMP 2.2
CHAMP 2.3
CHAMP 2.4
CHAMP 2.6
AJAXP 2.2
AJAXP 2.0
AJAXP 1.9
AJAXP 2.1
TUFFY 2.4
TUFFY 2.7
TUFFY 2.6
TUFFY 2.8
XTRA  2.4
XTRA  2.5
XTRA  2.3
XTRA  2.4
;
PROC PRINT;
PROC ANOVA;
   CLASS BRAND;
MODEL WEAR=BRAND;
   MEANS BRAND/DUNCAN;
;
/*
//
```

```
17:41 THURSDAY, NOVEMBER 15, 1984  1

              SAS

   OBS       BRAND     WEAR

     1       ACME      2.3
     2       ACME      2.4
     3       ACME      2.1
     4       ACME      2.5
     5       CHAMP     2.2
     6       CHAMP     2.3
     7       CHAMP     2.4
     8       CHAMP     2.6
     9       AJAXP     2.2
    10       AJAXP     2.0
    11       AJAXP     1.9
    12       AJAXP     2.1
    13       TUFFY     2.4
    14       TUFFY     2.7
    15       TUFFY     2.6
    16       TUFFY     2.8
    17       XTRA      2.4
    18       XTRA      2.5
    19       XTRA      2.3
    20       XTRA      2.4
```

```
17:41 THURSDAY, NOVEMBER 15, 1984    2

              SAS

    ANALYSIS OF VARIANCE PROCEDURE

        CLASS LEVEL INFORMATION

CLASS     LEVELS     VALUES

BRAND        5       ACME AJAXP CHAMP TUFFY XTRA

   NUMBER OF OBSERVATIONS IN DATA SET = 20
```

ANALYSIS OF VARIANCE PROCEDURE

DEPENDENT VARIABLE: WEAR

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	4	0.67700000	0.16925000	7.64	0.0015	0.670629	6.3221
ERROR	15	0.33250000	0.02216667		ROOT MSE		WEAR MEAN
CORRECTED TOTAL	19	1.00950000			0.14888474		2.35500000

SOURCE	DF	ANOVA SS	F VALUE	PR > F
BRAND	4	0.67700000	7.64	0.0015

SAS

ANALYSIS OF VARIANCE PROCEDURE

DUNCAN'S MULTIPLE RANGE TEST FOR VARIABLE: WEAR
NOTE: THIS TEST CONTROLS THE TYPE I COMPARISONWISE ERROR RATE, NOT THE EXPERIMENTWISE ERROR RATE.

ALPHA=0.05 DF=15 MSE=.0221667

MEANS WITH THE SAME LETTER ARE NOT SIGNIFICANTLY DIFFERENT.

DUNCAN GROUPING	MEAN	N	BRAND
A	2.6250	4	TUFFY
B	2.4000	4	XTRA
B			
B	2.3750	4	CHAMP
B			
B	2.3250	4	ACME
C	2.0500	4	AJAXP

3 . Randomized Block Design (RBD)

(1) 모 형 $y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$

- y_{ij} : 반응 (response)
- μ : 평균
- β_i : Biock $_i$ 의 효과 ($\sum_{i=1}^{b} \beta_i = 0$)
- τ_j : 처리 $_j$ 의효과 ($\sum_{j=1}^{t} \tau_j = 0$)
- ϵ_{ij} : 오차

(2) BRD는 t종류의 처리 (treatment)의 효과를 알아보기 위한 실험계획이며 각 처리에 같은 Block 안의 subject를 임의로 배치된다. 같은 Block 안의 subject는 서로 유사하는 공통점이 있다.

(3) PROC ANOVA ;

CLASS block variable treatment ;

MODEL response = block treatment ;

```
//SPSSAS30   JOB CLASS=V
//SAS         EXEC  SAS
//SAMPLE   DD DSN=SAS.SAMPLE,DISP=SHR
//SYSIN          DD  *
DATA;
 INPUT BLOCK BLEND$ PCTLOSS;
CARDS;
1  B  18.2
1  A  19.6
1  C  17.0
1  E  18.3
1  D  15.1
2  A  16.5
2  E  18.3
2  B  19.2
2  C  18.1
2  D  16.0
3  B  17.1
3  D  17.8
3  C  17.3
3  E  19.8
3  A  17.5
;
PROC PRINT;
PROC ANOVA;
   CLASS BLOCK BLEND;
MODEL PCTLOSS = BLOCK BLEND;
   MEANS BLEND/WALLER;
;
/*
//
```

SAS 14:41 WEDNESDAY, OCTOBER 31, 1984 1

OBS	BLOCK	BLEND	PCTLOSS
1	1	B	18.2
2	1	A	19.6
3	1	C	17.0
4	1	E	18.3
5	1	D	15.1
6	2	A	16.5
7	2	E	18.3
8	2	B	19.2
9	2	C	18.1
10	2	D	16.0
11	3	B	17.1
12	3	D	17.8
13	3	C	17.3
14	3	E	19.8
15	3	A	17.5

SAS 14:41 WEDNESDAY, OCTOBER 31, 1984 2

ANALYSIS OF VARIANCE PROCEDURE

CLASS LEVEL INFORMATION

CLASS	LEVELS	VALUES
BLOCK	3	1 2 3
BLEND	5	A B C D E

NUMBER OF OBSERVATIONS IN DATA SET = 15

SAS 14:41 WEDNESDAY, OCTOBER 31, 1984 3

ANALYSIS OF VARIANCE PROCEDURE

DEPENDENT VARIABLE: PCTLOSS

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PR > F	R-SQUARE	C.V.
MODEL	6	10.64800000	1.77466667	1.10	0.4370	0.452260	7.1650
ERROR	8	12.89600000	1.61200000		ROOT MSE		PCTLOSS MEAN
CORRECTED TOTAL	14	23.54400000			1.26964562		17.72000000

SOURCE	DF	ANOVA SS	F VALUE	PR > F
BLOCK	2	0.24400000	0.08	0.9278
BLEND	4	10.40400000	1.61	0.2615

(4) Example

```
PROC ANOVA;
   CLASS BLOCK BLEND;
   MODEL PCTLOSS = BLOCK BLEND;
   MEANS BLEND/WALLER;
```

Note: PCTLOSS = percent loss of insects

BLOCK = locations (b=3)

BLEND = blends of household insecticide (t-s)

C . t -검정 (t - Test)

1 . Two relased samples

(1) 두 측정치가 쌍 (pair)을 이루고 있는 경우의 t = test를 말한다. 예를 들면 실험의 subject로 15마리의 동물이 채택되고 각 동물이 흥분안정제를 주입받는다고 할 때 투약전의 맥박수를 x, 투약후의 맥박수를 y라 하면 (x , y)는 쌍을 이룬다. x와 y의 값에 차이가 있는지의 여부를 알기 위한 t statistic 은

$$t = \bar{d}\sqrt{s_d^2/n}$$

이고 여기에서 $\bar{d} = \bar{y} - \bar{x}$, $s_d^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2/(n-1)$,

$d_i = y_i - x_i$ ($i=1,\ldots,n$)이다.

(2) SAS : Example

```
//SPCSAS30  JOB CLASS=V
//SAS       EXEC  SAS
//SYSIN        DD  *
DATA SAMPLE;
 INPUT ID PRETEST POSTTEST;
       DIFF=POSTTEST-PRETEST;
CARDS;
 1  80  82
 2  73  71
 3  70  95
 4  60  69
 5  88 100
 6  84  71
 7  65  75
 8  37  60
 9  91  95
10  98  99
11  52  65
12  78  83
13  40  60
14  79  86
15  59  62
;
PROC PRINT;
PROC MEANS MEAN STDERR T PRT;
      VAR DIFF;
TITLE PAIRED-COMPARISONS T TEST;
/*
//
```

SAS

OBS	ID	PRETEST	POSTTEST	DIFF
1	1	80	82	2
2	2	73	71	-2
3	3	70	95	25
4	4	60	69	9
5	5	88	100	12
6	6	84	71	-13
7	7	65	75	10
8	8	37	60	23
9	9	91	95	4
10	10	98	99	1
11	11	52	65	13
12	12	78	83	5
13	13	40	60	20
14	14	79	86	7
15	15	59	62	3

PAIRED-COMPARISONS T TEST

VARIABLE	MEAN	STD ERROR OF MEAN	T	PR>|T|
DIFF	7.93333333	2.56434651	3.09	0.0079

2. Two Independent Samples

(1) 두 표본이 독립적으로 얻어진 경우 평균치 간의 차이가 있는지의 여부를 알아보는 t - test 이다. 첫번째 표본을 ($x_1, \ldots, x_{n_1}$) 이라 하고 두번째의 표본을 ($y_1, \ldots, y_{n_2}$)이라고 하면

$$t = (\bar{y} - \bar{x}) / \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

이고 여기에서 $s^2 = [\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2] / (n_1 + n_2 - 2)$ 이다.

(2) 예를 들면 두 종류의 탄약으로 만들어진 총알의 속도를 비교해 보는 경우 첫번째 표본은 한 종류의 탄약으로 만들어진 총알의 속도 측정치로 이루어져 있고 두번째 표본은 다른 종류의 탄환으로 만들어진 총알의 속도 측정치로 이루어져 있다고 하자.

(3) SAS : Example

```
//SPCSAS30   JOB CLASS=V
//SAS        EXEC  SAS
//SAMPLE   DD DSN=SAS.SAMPLE,DISP=SHR
//SYSIN         DD  *
DATA;
 INPUT POWDER VELOC;
CARDS;
 1  27.3
 1  28.1
 1  27.4
 1  27.7
 1  28.0
 1  27.4
 1  27.1
 1  28.1
 2  28.3
 2  27.9
 2  28.2
 2  28.4
 2  27.9
 2  27.7
 2  28.5
 2  27.9
 2  28.4
 2  27.8
;
PROC PRINT;
PROC TTEST;
   CLASS POWDER;
   TITLE T-TEST SAMPLE;
```

SAS 14:40 WEDNESDAY, OCTOBER 31, 1984 1

OBS	POWDER	VELOC
1	1	27.3
2	1	28.1
3	1	27.4
4	1	27.7
5	1	28.0
6	1	27.4
7	1	27.1
8	1	28.1
9	2	28.3
10	2	27.9
11	2	28.2
12	2	28.4
13	2	27.9
14	2	27.7
15	2	28.5
16	2	27.9
17	2	28.4
18	2	27.8

T-TEST SAMPLE 14:40 WEDNESDAY, OCTOBER 31, 1984 2

TTEST PROCEDURE

VARIABLE: VELOC

POWDER	N	MEAN	STD DEV	STD ERROR	MINIMUM	MAXIMUM	VARIANCES	T	DF	PROB > |T|
1	8	27.63750000	0.39256483	0.13879263	27.10000000	28.10000000	UNEQUAL	-2.7785	12.6	0.0161
2	10	28.10000000	0.29059326	0.09189366	27.70000000	28.50000000	EQUAL	-2.8762	16.0	0.0110

FOR H0: VARIANCES ARE EQUAL, F'= 1.82 WITH 7 AND 9 DF PROB > F'= 0.3940

CHAPTER 5. 다 변 량 분 석

CHAPTER 5. 다변량분석

A. DISCRIMINANT Analysis

1. Introduction

(1) Data: Multivariate (correlated) numerical data with pre-defined group identifier (class variable).

(2) Object: 1. Develope a decision rule (discriminant model) which can separate maximally based on the information of the sample.

2. Assign some new observations with unknown origin into one of the known groups.

2. Related Procedures

(1) DISCRIM: For the approximate multivariate normal population within group (with equal or not-equal variance-covariance matrix).

(2) NEIGHBOR: For radically non-normal population using non-parametric nearest neighbor method.

(3) CANDISC: Dimension-reduction technique related principal component and canonical correlation by taking linear composites of response variables with cleass variable, which give maximum between group deviation.

(4) STEPDISC: As a variable selection technique using forward selection, backward elimination or stepwise selection method to find a subset of variables that best discriminates group differences.

3. Alternative Procedure

(1) FUNCAT: Fitting categorical linear model with classification variable as the dependent variable.

(2) ANOVA: Series of Univariate Analysis of variance technique.

4. Background of DISCRIM Procedure

(1) Distribution in each group approx multivariate normal

$N(\underline{\mu t}, \Sigma t)$ or $N(\underline{\mu t}, \Sigma)$

for t=1,2,...k (number of groups).

μt Zt Σ : Population parameter

$\updownarrow$ $\updownarrow$ $\updownarrow$

Mt   St   S : Sample estimator

(2) Homogeneity test on Σ_t (t=1,2,...k) : Use the Bartlett's likelihood ratio test Statistic (Approximate Chi-square test) and the results of test determine whether the criterion is based on S_t or S.

(3) Decision Rule: Generalized Squared distance from $\underline{x}$ to group t such that

$$D_t^2(\underline{x}) = g_1(\underline{x},\ t) + g_2(t)$$,where

$$g_1(\underline{x}, t) = (\underline{x} - \underline{m}t)' S_t^{-1} (\underline{x} - m_t) + \log_e |S_t|$$

or

$$(\underline{x} - \underline{m}_t)'\ S\ (\underline{x} - \underline{m}_t)$$

"Sample Mahalanobis distance"

and g_2 (t) = g -2 $\log_e$qt, where

qt = Prior Probability of group t

If qt = constant for all t then qt = o.

Assign a new $\underline{x}$ to group t if

$$D_t^2(\underline{x}) = \min(D_1^2, D_2^2, \cdots\cdots D_k^2)$$

Equivalently SAS uses posterior probability of $\underline{x}$ belonging to group u, where

Posterior Probability

$$P_u(\underline{x}) = e^{-\frac{1}{2}P_u(\underline{x})} / \sum_{t=1}^{k} e^{-\frac{1}{2}D_t(x)}$$

and an observation X is assigned to group u if

$$P_u(\underline{x}) = \max\{P_1(x), \cdots\cdots P_k(x)\}$$

5. Outline of Use

PROC DISCRIM DATA = XX SIMPLE POOL = TEST WCORR OUT = OUT_{XX}; CLASS GROUP;

PROC DISCRIM DATA = OUTXX TESTDATA = YY;

TESTCLASS GROUP;

B. FACTOR Analysis

1. Introduction

(1) Data: Multivariate Numerical data, Correlation matrix, Covariance matrix, Factor Pattern matrix, or Scoring coefficient matrix.

(2) Object: Investigate structural relationship among response variables.

Several, Difficult to interprete, Correlated		variables

Using common factor	Few, Conceptually meaningful, Relatively independent	Hidden factors generating the dependence in responses

analysis with rotation.

2. Related Procedures (in the sense of structural analysis)

(1) CANCORR: Relationship between two sets of variables by finding a small number of linear composites for each set (canonical variables).

(2) PRINCOMP: Obtain a small number of linear composites of original variables (Principal Components) by orthogonal transformation that has as much the information in the original variables as possible (variance-oriented).

(3) FACTOR:

3. Background of FACTOR Procedure

Model: $x = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots\cdots \lambda_{ik} f_k + e^i, i=1,2,\ldots,$ p (# of responses) and $P \rangle \rangle k$

, where

$f_1, f_2 \cdots, f_k : k$ - common-factor variates

(factor loadings) : parameters representing the importance of the j'th factor on i'th response (need not to be orthogonal)

$e^{(i)}$: i'th specific factor variate.

$x^{(i)}$: linear function of a small number of unobservable (hidden) latent, common factor)) variables and a single (latent) specific factor. (i.e., response variable).

$\underline{x} = \Lambda f + \underline{e}$, where

$f = N(\underline{O}, I)$, $\underline{e} = N(O, D\varphi_i)$ φ_i is i'th specific variance (specificity) and $\underline{y}$ and $\underline{e}$ are stochastically independent.

Then the fundamental representation of Z under factor model is

$$\Sigma = \Lambda\Lambda' + D\varphi_i$$

Thus

$\mathrm{Var}(x^{(i)}) = \sigma_i^2 - \varphi_i = \sum^{k} \lambda_{ij}^2$ (communality of the i'th response)

$$\mathrm{Cov}(x^{(i)}, x^{(j)}) = \sigma_{ij} = \sum_{l=1}^{k} \lambda_{il} \lambda_{jl}$$

and $\mathrm{cov}(\underline{x}, f') = \Lambda$ (or correlation of $\underline{x}$ and $\underline{f}$ if standardized)

4. Factor Rotation

Let $\Lambda T = \Lambda^*$ such that $T\,T' = I$ (orthogonal) then $\Sigma = \Lambda^* \Lambda^* = D_P = M + D_\varphi$ Thus there exists infinitely many such T(ije, infinitely many factor loading matrix generating the same Σ.

Thusrstone (1945) proposed the concept of "simple structure" as a means of selecting Λ most meaningful and interpretable (simplicity and parsimony).

Orthogonal rotation Oblique rotation	Varimax, quartimax, equamax, procrustean

e.g., Varimax (Kaiser) : Minimize the sum of variances of the squared loadings within each column of factor matrix.

5. Factor Extraction and Goodness of Fit Test on the Factor Model

(1) Maximum likelihood factor analysis:

Factor extraction by maximum likelihood technique, and goodness of fit test for the number of factors extracted by the generalized likelihood ratio principle. (Lawley and HOME).

(2) Principal component analysis.

(3) Principal factor analysis.

(4) Iterative Principal Factor analysis.

(5) Alpha factor analysis.

(6) Image analysis.

6. Outline of Use

(1) Principal Component Analysis:

```
PROC FACTOR SCORE OUTSTAT = XX
   SCREE REORDER;
PROC SCORE DATA = A SCORE =
   OUT = SCDATA:
PROC PLOT; PLOT FACTOR 2 * FACTOR 1
```

(2) Principal factor analysis (Simplest and Computationally efficient) :

```
PROC FACTOR ... ; PRIORS  { SMC } ; (for singular
                          { NAX }    correlation matrip)

PROC FACTOR N = #
   ROTATE = VARIMAX ROUND SCORE ... ;
PROC SCORE ... ;
PROC PLOT  ... ;
PROC REG   ... ;
```

(3) Maximum likelihood factor analysis: (best in statistical point of view)

1) Rigid test for the number of factors.

2) Desirable asymptotic properties on the estimators.

3) Not require normality but expensive.

```
PROC FACTOR METHOD = ML;
```

C. CLUSTER Analysis

1. Introduction

(1) Data: 1. Multivariate numerical data without pre-defined group information.

2. Similarity matrix (squared distance matrix, correlation type matrix).

(2) Object: Assign (cluster) observations (or variable) into groups in such a way that (high) similarity within group and (high) dissimilarity between groups in some sense.

(3) Types of clusters

1. Disjoint cluster
2. Hierachical cluster
3. Overlapping cluster
4. Fuzzy cluster (combined)

2. Related Procedures

(1) CLUSTER: Find hierachical clusters of observations using centroid, Ward-Hook, or average linkage method on squared Euclidean distance.

(2) FASTCLUS: Disjoint cluster based on K-means method for large data.

(3) VARCLUS: Hierachical and disjoint clustering of variables.

(4) TREE: Drew a tree diagram (dendrogram) using output from CLUSTER or VARCLUS.

3. CLUSTER Procedure

Use the three standard agglomerative hierachical clustering algorithms based on different distance measures as a measure of similarity.

(1) Centroid : Euclidean distance between cluster centroids (Robust to outliers).

(2) Ward-Hook Method : Error sum of squares within clusters, within clusters,

$x_{ij} \quad k \quad$ (i = 1,2,..., p (# of responses)

j = 1,2,..., n_k (size of kth cluster)

k = 1,2,..., c (# of clusters)).

$$(SSE)_k = \sum_{i=1}^{P} \sum_{j=1}^{xk} (x_{ijk} - \bar{x}_{ik})^2$$

$$SSE = \sum_{k=1}^{c} (SSE)_k$$