専門教育教材

SAS 初級課程

1996.

統計廳
統計研修院

총 목 차

SAS 기 본 운 영 법

목 차

1. SAS 개요

가. SAS 란 ?

SAS(STRATEGIC APPLICATION SOFTWARE 또는 STATISTICAL ANALYSIS SYSTEM)의 약어로서 미국 North Carolina에 있는 SAS연구소에 의해 개발된 통계분석 Package이다.

나. SAS 특징

(1) 자료이용(DATA ACCESS)
- 직접적으로 DATABASE의 자료를 이용
- 원자료를 직접이용하고, 여러 가지 형태로 변환 또는 처리가 용이함.

(2) 자료관리(DATA MANAGE)
- 자료의 정렬, 결합, 분류,수정, 검색, 입력등이 편리함.

(3) 자료분석(DATA ANALYSIS)
- 통계분석,의사결정,모형설정 및 예측 등

(4) 자료표현(DATA PRESENT)
- 보고서 작성을 위한 여러가지 그래프(차트, 지도, 3차원그래프)
- 여러가지 통계표 작성

다. SAS의 운영환경

SAS는 MVA(Multi-Vendor- Architectur)개발 철학에 근거하여 개발되었다. 따라서 사용자의 H/W환경과 운영체계(OS)와 상관없이 동일한 개발환경을 지원하며 사용되는 문장이나 명령어도 어느환경이나 동일하다.
- 대형컴퓨터, 미니컴퓨터, 워크스테이션, 개인용컴퓨터 등

라. SAS 종류

SAS/BASE,	SAS/STAT,	SAS/ETS,	SAS/IML,	SAS/OR,	SAS/QC,
SAS/GRAPH,	SAS/AF,	SAS/FSP,	SAS/ASSIST,	SAS/EIS,	SAS/CALC,
SAS/CONNECT,	SAS/ACCESS,	SAS/CPE,	SAS/INSIGHT,	SAS/LAB,	SAS/PH-CLINICAL

2. PC/SAS 이용법

PC/SAS를 가동시키려면 CONFIG.SYS 화일에 FILES와 BUFFERS를 지정해야 하는데 최소값은 다음과 같다.

FILES=50

BUFFERS=20

가. PC/SAS 시작과 종료

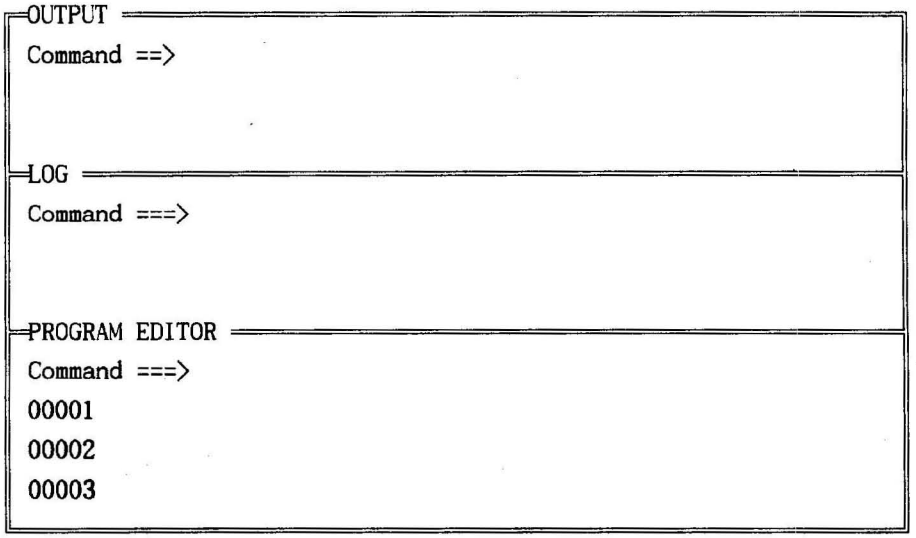
PC에서 SAS를 시작하려면 프로그램이 하드디스크의 C DRIVE에 있고 SAS라는 이름의 DIRECTORY에 있을 때

C:\>CD SAS ↵

C:\SAS>SAS ↵ 하면

다음과같이 세가지 창(WINDOW)으로 나뉘어져 화면관리시스템(DMS,Display Manager System)상태로 나타난다.

PC/SAS의 주화면



SAS의 모든 작업은 이 윈도우를 통하여 이루어진다. DATA를 입력하고 SAS프로그램을 만들고, 실행(SUBMIT)하고, 에러 발생시 프로그램을 수정하고 또한 분석결과를 보고, 프린트하는 등 모든 작업이 이 윈도우를 통하여 이루어진다.

PC/SAS를 종료하려면 PROGRAM EDITOR의 Command line에서 BYE ↵하면 된다. 또는 1행에 endsas;를 치고 Command line에서 SUBMIT ↵하면 된다.

나. SAS DMS 이용법

(1) DMS 기본 WINDOW : 사용자가 임의로 열고 닫을 수 없는 기본윈도우.

- OUTPUT WINDOW : 프로그램의 실행 결과 출력.
- LOG WINDOW : 프로그램실행메시지, 오류발생메시지 출력.
- PROGRAM EDITOR WINDOW : 실제 SAS프로그램의 작성과 실행하는 곳.

(2) 기타 DMS WINDOW : 필요에 따라 사용자가 열고 닫을수 있는 윈도우

- KEYS WINDOW : 기능키에 지정된 DMS명령어의 조회 및 변경.
- OPTIONS WINDOW : 시스템의 전반적인 초기옵션의 조회 및 변경.
- HELP WINDOW : SAS 프로그램 전반에 걸친 도움말을 보는 곳.

(3) 각각의 윈도우를 열고 닫는 방법.

- 윈도우 열기 : Command line에서 해당윈도우의 이름을 입력한다. 즉, OUTPUT, LOG, PGM, KEYS등을 치거나 KEYS WINDOW 에 지정된 기능키(FUNCTION KEY)를 사용하면 된다.
- 윈도우 닫기 : Command line에서 END 를 입력한다.

(4) SAS DMS명령어

표1) COMMAND LINE 명령어

명령문	설 명	예 제
SUBMIT	프로그램을 실행할때	FILE 'A:TEST1'
ZOOM	윈도우 화면 크기를 조정할때	FILE 'A:TEST '
FIND	문자열을 찾을 때	FIND 'SAS'
CHANGE	문자열을 변환할 때	C 'SSS' 'SAS'
FILE	프로그램을 외부화일로 보관(SAVE)할 때 또는	FILE 'A:TEST1'
	프로그램 또는 프로그램의 결과를 프린트할때	FILE 'PRN'
INCLUDE	외부화일로 지정된 프로그램을 불러올 때	INCLUDE 'A:TEST1'
CLEAR	화면에 나타난 결과를 지우고자 할때	
RECALL	바로 전 실행된 SAS프로그램을 불러오자 할 때	
X	DOS로 벗어남. DOS에서 EXIT를 치면 다시 SAS로 돌아옴	
BYE	SAS를 끝낸다. 또는 00001 LINE에 ENDSAS;를 치고 SUBMIT	

표2) LINE COMMAND 명령어(TEXT EDITOR 명령어)

명령문	기 능	예 제
I, I{n}	I{n} : LINE을 n행 삽입	I3001 3행삽입
M, M{n},	M{n} : LINE을 n행 이동(A,B로 장소지정)	M2002 2,3행을 5행
MM···MM	BLOCK 단위로 이동	A0005 뒤에 이동
C, C{n},	C{n} : LINE을 n행 복사.	C0004 4행을 10행뒤
CC···CC	BLOCK 단위로 복사	B0010 에 복사
D, D{n},	D{n}: LINE을 n행 삭제	D2006 6,7행 삭제
DD···DD	BLOCK 단위로 삭제	00007
R, R{n},	R{n}: 현재 LINE을 다음 LINE에 N행 복사.	R0008 8 행을 9행에
RR··RR	BLOCK 단위로 다음 LINE에 복사	00009 복사
A(After)	복사, 이동 장소를 LINE 앞에 지정	00010
B(Before)	복사, 이동 장소를 LINE 뒤에 지정	00011

3. SAS 프로그램의 형태

SAS는 SAS의 자체 문장으로 구성되는 명령어에 의해 운영되며, SAS 자신만에 언어(SAS LANGUAGE)를 가지고 있다. 각 문장은 다음과 같은 SAS 언어의 규칙에 따른다.

- SAS DATASET이름으로 8자내의 임의의 영문자과 숫자를 사용한다. 영문자는 대, 소문자에 구애받지 않는다. #, $, @ 등 특수문자를 사용할 수 없다.
- SAS DATASET을 구성하는 변수이름은 8자이내이고, 변수명 첫자는 숫자로 시작하면 안된다. 중간에 공백, -(하이픈)은 사용할 수 없다.
- SAS 문장은 세미콜론(;)으로 구분한다.
- SAS 문장은 어떤 열(Column)에서도 시작할 수 있고 여러 행에 걸쳐서 써도 된다.
- 한 행(Line)에 여러 문장을 쓸 수 있다. 단, 문장마다 세미콜론으로 구분해야 한다.
- 한문장을 주석으로 사용하고자 하는 경우에는 * 으로 시작하여 ;로 닫는다.
- 여러문장을 주석으로 사용하고자 하는 경우에는 /* 으로 시작하여 */로 닫는다.

  단 /*

  /* 전달문 */

  */ 인 경우는 안된다.

가. SAS 프로그램의 구조

SAS는 자료입력부분(DATA STEP)과 자료분석부분(PROC STEP)으로 되어있다.

(1) DATA STEP

- SAS DATASET의 생성 및 수정, 변경

예) DATA ONE ;

```
 INPUT X Y @@; 입력자료 변수명 ┐  DESCRIPTION PORTION
 SUM = X+Y ;   새로운변수생성   │  DATASET에 대한 일반적인 정보수록
CARDS ;        직접입력지시문  ┘
22 23 25 22 23 25 ──────────┐  DATA PORTION
26 27 28 26 27 28           │  실제 자료 입력
30 40 50 30 40 50 ──────────┘
RUN ;
```

(2) PROC STEP

- 생성된 SAS DATASET를 대상으로 자료처리 및 분석작업 수행

예) PROC PRINT ; 입력자료를 출력하라는 명령어
VAR SUM X Y Z ; 출력할 자료 변수명
RUN;

나. 자료읽기

SAS로 자료를 읽는 방법은 프로그램내에서 자료를 직접읽는 방법과, 외부화일에 있는 화일을 불러들여 읽는 방법이 있다.

(1) 프로그램 내에서의 자료 읽기

```
DATA 데이터세트 명 ;
 INPUT 변수명 [변수타입] 변수명 [변수타입] ... ;
CARDS ; 데이터입력시작
 DATA LINE
 . . . . .
RUN ; DATA STEP 종료 ;
예) DATA ONE ; ONE
     INPUT REGION $ SALES  ;
    CARDS ;
    SEOUL  9664
    BUSAN 22969
    DAEGU 18941
    RUN ;
```

(2) 외부화일에서의 자료읽기

```
DATA 데이터세트 명 ;
 INFILE 외부데이타화일의 위치와 명칭 [OPTION] ;
INPUT 변수명 [변수타입] 변수명 [변수타입] ... ;
RUN ; DATA STEP 종료 ;
예) DATA ONE ;                          결과) ----+----+----+----
     INFILE  'C:\CITY.DAT'  ;                  SEOUL     123456
    INPUT REGION $ 1-8 SALES 11-20 ;           BUSAN     67543
    RUN ;
```

다. 자료읽기에 사용되는 명령문

(1) DATA 문 : DATA [SAS dataset [(옵션들)]

SAS 데이타셋이름으로 8자내의 임의의 영문자와 숫자를 사용한다. 단 #, $, @ 등 특수문자를 사용할 수 없다. SAS Dataset 사용법의 응용형태로 _NULL_ 이 사용된 형태, Dataset 이름이 생략된 형태, Dataset 이름이 여러 개 있는 경우 등이 있다.

- DATA 문의 전형적인 형태 : DATA ONE ;

```
DATA ONE ;
  INPUT A B ;
CARDS;
11 21
31 41
```

- DATA _NULL_ 의 형태 : DATA _NULL_;

- DATA 셋 이름이 없는 경우 : DATA ;
  Dataset 명이 없을 때, SAS를 실행하는 동안 생겨나는 SAS Dataset 순서에 따라 DATA1,DATA2, ··· 이름의 SAS Dataset이 생긴다.

- DATA 셋 이름을 여러 개 지정하는 경우 : DATA A1 A2 A3 ;
  DATA 문에서 Dataset을 경우에 따라서는 여러 개를 사용할 수 있다.
  예) DATA A1 A2 A3;

```
IF       X=1 THEN OUTPUT A1;
ELSE IF X=2 THEN OUTPUT A2;
ELSE             OUTPUT A3;
```

- DATA 셋 이름이 두 단어로 되는 경우 : DATA ONE.A1 ;

- DATA문에 사용되는 선택적 명령
  DROP=변수명 : 불필요한 변수제거명령, DATA ONE(DROP=REG) ;
  FIRSTOBS=N : 첫번째 RECODE의 시작위치
  OBS=N : RECORD의 개수를 지정함,
  DATA ONE ; SET A1(FIRSTOBS=100 OBS=20);
  KEEP=변수명 : 필요한변수만 지정할 때, DATA ONE(KEEP=REG) ;
  RENAME=(OLD 변수명=NEW 변수명): 변수의이름을 바꾸고자 할 때

(2) INPUT 문

INPUT 문은 읽어 들이고자 하는 DATA의 변수의 이름과 입력형식를 나타내는 문장(statement)이다.

- 자유형식 (Free Format)
  데이타가 1칸 또는 그 이상의 빈칸으로 분리되어 있는 경우 사용하며 그 형식은 다음과 같다. 문자형변수인 경우에는 변수명 다음에 $를 붙인다.
  INPUT variable [$];
  예) DATA ONE ;
  INPUT REGION $ SALES ;
  CARDS;
  SEOUL 123456 28292
  BUSAN 67543 8499

- 변수의 위치를 지정하는 방법 (column input)
  자유형식의 데이타는 변수사이에 한칸 이상 빈칸을 넣어야 하므로 데이타가 많은 장소를 차지하게 된다. 연속되어 입력된 경우에는 변수가 입력된 위치를 지정하는 방법을 이용한다.
  INPUT variable [$] startcolumn-endcolumn [.decimals];

  예) INPUT SALES 7-12 .2 ;

입력	결과
2314	23.14
2	0.02
400	4.00
-140	-1.40
12.234	12.234
12.2	12.2

- 변수가 차지하는 열의 크기를 지정하는 방법(formatted input) :
  변수의 차지하는 열의 위치를 지정하지 않고 크기만을 지정하는 방법이다. 이 방법은 변수명을 X1, X2,..,X10와 같이 연속적인 변수명을 사용하는 경우에 상당히 편리하다.
  INPUT (변수명, 변수명,..) (변수특성 열길이);

예1) DATA ONE ;
INPUT (REGION SALES INVENT)($5. +1 6. +1 5.) ;
CARDS;
SEOUL 123456 28292
BUSAN 67543 8499
RUN ;

여기서 (REGION SALES INVENT)($5. +1 6. +1 5.) : 변수 REGION는 문자 5칸($5. 또는 $CHAR5.), 다음 1칸 띄고(+1), SALES 6칸(6.), 1칸 띄고(+1), INVENT 5칸(5.)을 차지한다.

예2) DATA A1 ; INPUT A1 1-3 A2 4-6 A3 7-10 ···. ;
=> DATA A1 ; INPUT (A1-A3)(2*3. 4.);

여기서 A1, A2 두 개의 변수는 3칸을 차지하고 (2*3.), 변수 A3은 4칸을 차지한다.

- 변수의 위치를 지정하는 방법(pointer input) :
Input {pointcontrol}valiable={$}{informat}...;

(가) COLUMN 위치지정(@)
예) DATA ONE ;
INPUT @1 REGION $5. @7 SALES 6. @12 INVENT 5. ;
CARDS;
SEOUL 123456 28292
BUSAN 67543 8499
RUN ;
여기서,
@1 REGION $5. : 문자변수로 1번째 열부터 시작하여 5열을 차지
@7 SALES 6. : 숫자변수로 7번째 열부터 시작하여 6열을 차지.
@12 INVENT 5. : 숫자변수로 12번째 열부터 시작하여 5열을 차지.

(나) LINE 위치지정(#,/)
입력하고자 하는 변수의 갯수가 많아서 1줄에 입력할 수없는 경우가 있다. 이 때에는 "#"이나 "/"를 이용하여 다음 행이라는 것을 지정해야 한다.
예1) DATA ONE ;
INPUT (x1-x3)(3*2.) #2 (y1-y4)(4*6.);
예2) DATA ONE ;
INPUT x1 1-2 x2 3-4 x3 5-6 /
y1 1-6 y2 7-12 y3 13-18 y4 19-24;

(다) LINE holding(@, @@)

· 변수 format 뒤에 @ 하나 사용 : 계속하여 같은라인을 read 하고자 할 경우

예) DATA ONE ; INPUT type $ 1 @ ;
if type='c' then input @1 course $1. @4 rof 3. ;
else if type='d' then input @2 name $2. @8 id $1.;

· 변수 format 뒤에 @@ 하나 사용 : 계속하여 라인에 구애없이 data를 read 하고자 할 경우

예) DATA ONE ; INPUT name $ age @@;
CARDS ;
jone 33 chun 13 kim 20 tom 10 lee 20

- INPUT 문의 DATA 읽기형식

표1) 숫자변수 읽는 형식

형 식	설 명	자리수	소숫점	default
w.	standard numeric	1-32		
w.d			0-31	
BZw.	blank are zeros	1-32	0-31	1
COMMAw.d	commas in numbers	2-32	0-31	1
Ew.	scientific notation	7-32	0-31	12
HEXw.	numeric hexadecimal	1-16		8
IBw.d	integer binary	1-8	0-10	4
MRBw.d	Microsoft real BASIC binary	2-8	0-10	4
PDw.d	packed decimal	1-16	0-10	1
PIBw.d	positive integer binary	1-8	0-10	1
PKw.d	unsigned packed decimal	1-16	0-10	1
RBw.d	real binary(floating point)	2-8	0-10	4
S370FIBw.d	IBM 370 integer binary	1-8	0-10	4
S370FPDw.d	IBM 370 packed decimal	1-16	0-10	1
S370FRBw.d	IBM 370 real binary	2-8	0-10	6
S370PIBw.d	IBM 370 positive integer binary	1-8	0-10	4
ZDw.d	zoned decimal	1-16	0-10	1
ZDBw.d	zoned decimal	1-32	0-10	1

주) w : data 자리수, d : 소숫점 자리수

표2) 문자변수 읽는 형식

형 식	설 명	자리수	default
$w.	standard character	1-200	1 or length of value
$CHARw.	characters with blanks	1-200	1 or length of value
$EBCDICw.	EBCDIC to ASCII	1-200	1
$HEXw.	character hexadecimal	1-200	2
$VARYINGw.	varying-length character VALUES	1-200	8 or length of value

표3) 날짜변수 읽는 형식

형 식	설 명	자리수	default
DATEw.	date of form ddMMMyy	7-32	7
DATETIMEw.	date-time values	13-40	18
DDMMYYw.	date values	6-32	8
JULIANw.	Julian dates	5-32	5
MMDDTYYw.	date values	6-32	8
MONYYw.	month and year	5-32	5
NENGOw.	Japanese dates	7-32	105
TIMEw.d	time values	5-32	8
YYMMDDw.	date values	6-32	8
YYQW.	year and quarter	4-32	4

(3) INFILE 문

SAS 프로그램내에서 DATA를 직접입력하려면 CARDS ; 문을 사용하고 디스크에 보관되어 있는 외부자료를 SAS 프로그램 안으로 불러올 때 INFILE문을 사용한다.

- OPTIONS
  외부화일명 : 'C:\SAMPLE.SDAT' ;
  FIRSTOBS=라인수 : INFILE 'C:\SAMPLE.DAT' FIRSTOBS=100 ;
  OBS=라인수 :
  LRECL=N : 읽고자하는 RECORD 길이(1~32767)를 지정.
  PAD : 가변길이 RECORD를 읽을 때, LRECL= 과 함께 사용.

MISSOVER : READ하는 RECORD에 MISSING VALUE가 있으면 변수에 '.'을 SET 하고 다음 LINE으로 READ POINTER를 옮김.
STOPOVER : READ하는 RECORD에 MISSING이 있으면 READ 명령이 중단됨.
TRUNCOVER : 비정상적인 화일을 정상적으로 읽을 때.

라. 자료변환에 사용되는 명령문

(1) KEEP 문

KEEP 변수명 : 변수명은 SAS Dataset에 보관하고자 하는 변수명들이다.

(2) DROP 문

DROP 변수명 : 변수명은 SAS Dataset에서 제외하고자 하는 변수명들이다.

(3) RENAME 문

RENAME 원래 변수명= 바꾸고자 하는 새로운 변수명 ···;

(4) IF-THEN, IF-THEN/ELSE

- THEN을 사용하는 경우 :IF expression THEN statement ;
  IF 뒤의 조건에 만족하는 경우의 관측치에만 적용된다. 여기에 ELSE 문을 첨가하여 IF에서 지정하지 않은 관측치에 대해서도 자료처리를 할 수 있다.

- THEN을 사용하지 않는 경우 :IF expression ;
  THEN 문을 사용하지 않는 경우에는 IF 조건에 맞는 관측치의 경우에만 자료처리를 적용하여 Dataset을 만든다.

(5) DELETE

DELETE는 어떤 조건식에 맞는 관측치를 삭제하는 경우에 사용한다. DELETE가 DROP과 다른 점은 DROP은 관측치를 삭제하는것이 아니라 관측치는 그대로 두고 변수를 삭제한다.

IF 조건식 THEN DELETE;

(6) STOP 문

STOP 문은 SAS DATA STEP을 중지하고자 할 때 사용한다. STOP 문을 실행하게 되는 관측치는 SAS Dataset에 포함되지 않는다. 수만 개의 자료를 분석하는 SAS 프로그램을 만드는 경우 우선 몇 개의 샘플만 이용한 SAS 프로그램을 만들어 실행하여 프로그램에 이상이 있는지를 먼저 체크한 후에 전 데이타를 실행하는 것이 훨씬 효율적이다. 이 때 효과적으로 사용할 수 있는것이 STOP 문이다.

```
IF _N_ = 100 THEN STOP;
```

(7) OUTPUT 문

INPUT 문에 의한 자료를 읽어 들이지 않고 DATASET를 만들거나 조건문에 의해 여러개의 DATASET를 만들 때 사용한다.

```
예1) DATA ONE ;
     DO I=1 TO 10 ; OUTPUT ; END ;
     PROC PRINT DATA=ONE ;
     OBS    I
      1     1
      2     2
      ..    ..
     10    10

예2) DATA ONE TWO ;
      INPUT OD $ X Y Z ;
      IF ID='A' THEN OUTPUT ONE ;
      ELSE IF ID='B' THEN OUTPUT TWO ;
      CARDS ;
      A  2 5 4
      B  7 8 9
      RUN ;
     PROC PRINT DATA=ONE;      PROC PRINT DATA=TWO ;

     결과) OBS ID  X  Y  Z     결과) OBS ID  X  Y  Z
           1   A  2  5  4           1   B  7  8  9
```

(8) DO 문

DO 문은 END 문이 나올 때까지 사이에 있는 여러 개의 명령문들을 실행하고자 할 때 사용한다. DO 문에는 단순 DO 문, 반복을 표시하는 DO 문, DO WHILE 그리고 DO UNTIL 문이 있다.

- 단순 DO 문

  단순 DO 문은 「DO; ··· END;」 사이의 명령문을 실행하고자 할 때 사용하는 것으로 보통 IF ··· THEN/ELSE 문과 같이 사용된다.

```
DO ;
  SAS 문장들;
END;
 DO I=2,3,5,7,9 ;
 DO MONTH='JAN','FEB','MAR';
```

- 반복 DO 문

  반복 DO 문은 DO 문 안에 색인을 나타내는 색인(index)변수를 지정하여 색인변수의 크기에 선택적으로 반복회수를 지정한다.

```
DO 색인변수=시작 [TO 끝 [BY 증가분] [WHILE 또는 UNTIL(표현식)]  ];
  SAS 문장들;
END;
DO I=1 TO 8 BY 2,12,14,17;
DO I=1 TO 20 BY 5;
```

- DO WHILE 문

  DO WHILE 문은 WHILE 다음의 조건이 맞는 한 계속해서 실행된다.

```
DO WHILE (표현식);
  SAS 문장들
 END;
 DO I=1 TO 10 WHILE(X<Y); DO I=10 TO 0 BY -1 WHILE(MONTH='JAN') ;
```

- DO UNTIL 문

  DO UNTIL문은 UNTIL 다음의 조건이 맞을 때 실행을 중지한다. 조건은 처음부터 체크하는 것이 아니고 SAS 문장을 실행한 후에 체크를 한다. 그러므로 어떠한 경우에도 적어도 1번은 DO-END 문사이의 SAS명령을 실행한다.

```
DO UNTIL (표현식);
  SAS 문장들
 END;
DO I=2 TO 20 BY 2 UNTIL ((X/3)>Y);
```

(9) ARRAY 문(배열문)

여러 변수에 동일한 명령을 적용시키고자 하는 경우에 이 변수군을 배열문으로 지정하여 각 변수를 배열문의 원소로 지정할 수 있다. 일종의 dimension 개념이다. 이 ARRAY 문의 사용은 시스템마다(극히) 조금씩 다를 수있다.

ARRAY array명[{n}] [$] [length] [변수들 이름];

(9) TITLE 문

TITLE 문은 SAS 결과를 프린트할 때 프린트 윗부분에 프린트하는 기능으로 10줄까지 지정할 수 있다.

TITLE[n] ['title'];

(10) FOOTNOTE 문

FOOTNOTE 문은 SAS 결과를 인쇄하고자 할 때 프린트용지 제일 아래 부분에 프린트하는 기능으로 10줄까지 지정할 수 있다. 프린트용지의 제일 아래에서 몇 줄 위에 text를 프린트할 것인지를 지정한다.

FOOTNOTE[n] ['text'];

(11) LABEL 문

변수에 특정이름을 주고자 할 때 사용한다.

LABEL variable= 'label' ···;

variable : Label을 지정하고자 하는 변수명

label : 빈칸을 포함하여 40자까지 가능하며 인용부호가 앞뒤에 있어야 한다.

(12) FORMAT 문

데이타스텝에서 변수가 가지는 값의 크기에 따라 특정이름을 주고자 할 때 사용한다.

FORMAT variables [format] ···;

variable : 특정이름을 지정하고자 하는 변수명

format : 변수의 값에 따라 프린트하고자 하는 양식(foramt)

(18) SAS 내장함수

SAS에는 다른 소프터웨어처럼 여러가지 기능을 가진 내장함수를 갖고 있어서 이를 알아두면 상당히 편리하게 사용된다. 또한 이를 모르면 Data 처리가 불가능한 것도 많다.

표1) 통계에 관한 함수

함 수	사 용 법	결 과 치
CSS(수정된 자승합)	X=CSS(4,2,3.5,6)	X=8.1875
CV (변이계수)	X=CV(5,8,9,3,6)	X=38.50754
KURTOSIS(첨도)	X=KURTOSIS(1,0,1,0)	X=-6
MAX(최대값)	X=MAX(5,6,7,8)	X=8
MIN(최소값)	X=MIN(5,6,7,8)	X=5
MEAN(평균값)	X=MEAN(5,6,7,8,9)	X=7
N(자료의 개수)	X=N(1,2,3,4,6)	X=5
NMISS(Missing 자료수)	X=NMISS(1,.,2,4,.)	X=2
RANGE(범위)	X=RANGE(2,3,4,5)	X=3
SKEWNESS(왜도)	X=SKEWNESS(0,1,1)	X=-1.73205
STD(표준편차)	X=STD(2,4,6,3,1)	X=1.923538
STDERR(표준오차)	X=STDERR(2,6,3,4)	X=0.8539126
SUM(합계)	X=SUM(2,4,6,3,1)	X=16
USS(자승합)	X=USS(4,2,3.5,6)	X=68.25
VAR(분산)	X=VAR(4,2,3.5,6)	X=2.729167

표2) 연산(Arithematic) 기능을 가진 함수

함 수	사 용 법	결 과 치
ABS(절대값)	X= -5 Y=ABS(X)	Y=5
MOD(나눗셈에서의 나머지)	X=MOD(10,3)	X=1
SIGN(변수의 부호 또는 0)		
SQRT(제곱근)	X=SQRT(25)	X=5

표3) Truncation 함수

함 수	기 능 설 명	사용법	결과치
CEIL	변수보다 큰 수이면서 제일 작은 정수	X=CEIL(2.1)	X=3
FLOOR	변수보다 적은 수이면서 제일 큰 정수	X=FLOOR(2.1)	X=2
INT	정수부분만을 취한다.	X=INT(2.1)	X=2
		X=INT(0.999)	X=1
ROUND	반올림의 값	X=ROUND(223.456,1)	X=223

표4) 수학적 연산함수

함 수	사 용 법	결 과 치
EXP(지수함수)	X=EXP(1)	X=2.71828
LOG(자연로그(base가 e))	X=LOG(1)	X=2.30259
LOGn(base가 n인 로그 n=10 이면 상용로그)		

표5) 변수변형에 관한 함수

함 수	기 능 설 명 및 사 용 법	결과치
LENGTH	변수의 길이 : X=LENGTH('ABCDEF')	X=6
SUBSTR	문자에서의 일부를 추출 : X='AB123CD'	
	Y=SUBSTR(X,3,3)	Y=123
UPCASE	대문자로 변환 : X='abc' Y=UPCASE(X)	X='ABC'
COMPRESS	특정문자 제거후 압축하기: X='ABC DEF'	
	Y=COMPRESS(Y)	Y=ABCDEF
INDEX	찾고자하는 문자들의 위치 구하기 : X='ABC.DEF(X=Y)'	
	Y=INDEX(X,'DEF')	Y=4
INDEXC	찾고자하는 문자들중에서 발견된 첫번째 위치구하기	
	X='ABC.DEF(X=Y)' Y=INDEX(X,'0123456789',';=.B')	Y=2
LEFT	문자열을 좌로 정렬하기 : X=' HI THERE'	
	Y=LEFT(X)	Y='HI THERE '
REPEAT	문자반복구하기 : X=REPEAT('ONE',2)	X=ONEONE
REVERSE	문자순서 바꾸기 : X='ABC' Y=REVERSE(X)	X='CBA'

표5) 변수변형에 관한 함수(계속)

함 수	기 능 설 명 및 사 용 법	결과치
RIGHT	문자열을 우로 정열하기 : X='HI THERE ' Y=RIGHT(X)	Y=' HI THERE'
TRANSLATE	문자열중 명시된 문자로 바꾸기 : X='XYZW' Y=TRANSLATE(X,'AB','VW')	Y='XYZB'
VERIFY	문자열 탐색하기 : X='XYZ' Y='12345' Z=VERIFY(Y,X) Y문자열 중에서 X문자중 XYZ을 찾아보고 첫 번째위치를 찾는다. Y문자열 중에서 X문자열 XYZ가 없으므로 Z=0가 된다.	Z=0
TRIM	문자열 뒷 공백 없애기 : X='HONG' Y=KIL-DONG' Z='TRIM(X)¦¦' '¦¦TRIM(B)'	Z='HONG KIL-DONG'

표6) 특수기능함수

함 수	기 능 설 명	결과치
DIFn	n-th LAG 차이를 계산한다.	OBS X Y=DIF(X) Z=DIF2(X) 1 1 . . 2 2 1 . 3 6 4 5 4 4 -2 2
LAGn	n-th Lagged 값을 계산한다.	OBS X Y=LAG(X) Z=LAG2(X) 1 1 . . 2 2 1 . 3 6 2 1 4 4 6 2
SYMGET	마이크로값 치환하기	

```
%LET SYM1=AAA ; %LET SYM2=BBB ; %LET SYM3=CCC ;
  DATA ONE ; INPUT CODE $ ;
  X=SYMGET(CODE) ;        결과)  OBS CODE  X
  CARDS ;                         1  SYM2 BBB
  SYM2                            2  SYM3 CCC
  SYM3                            3  SYM1 AAA
  SYM1
  PROC PRINT ;
```

(19) SAS 연산자(Operators)

SAS 연산자는 자료의 산술계산, 비교 및 논리식의 비교 등을 행하는데 필요한 기호이다.

- 산술연산자 : + , - , *, /,**
  예) Y = $X^3 + 3 \times X - 3$ ==> Y = X**3 + 3*X - 3;

- 비교연산자 : =(EQ), ^=(NE), >(GT), <(LT),
  <=(LE), >=(GE), ^>(NG), ^< (NL)
  예1) IF AGE < 10 THEN AGEG=1;
  IF 10 <= AGE < 30 THEN AGEG=2;
  IF X < Y THEN C=5; ELSE C=12;
  IF SEX='M' THEN DELETE;
  C=5*(X<Y) + 12*(X>Y);
  여기서 (X<Y)는 X가 Y보다 크면 0, X가 Y보다 작으면 1의 값을 가진다. 그러므로 X=6, Y=8인 경우에는 C=5*(1) + 12*(0)=5가 된다.

  예2) IF REG ='11' OR REG='21' OR REG='31' 또는
  IF REG IN ('11','21', '31')
  여기서 IN은 변수 state가 가지는 값중에 'ny','pa', '1a' 중 어느 하나의 문자를가지는 자료를 추출한다.

  예3) IF NAME > 'S'; 변수 NAME 을 「S」와 비교한다.
  IF NAME > 'S ': 변수 NAME 을 「S 」와 비교한다.
  IF NAME >:'S'; 변수 NAME 의 첫 글자를 S와 비교한다.

- 논리연산자 : &(AND), |(OR), ^(NOT)
  예1) IF REG='11' AND SEX='F' THEN ;
  NOT(A=B & C>D) 또는 A NE B OR C LE D

- 기타 연산자 : ><(MIN) , <> (MAX), || (Concatenation)
  · A>< B : A와 B 중 작은 값을 가진다.
  예) K = A >< B : A=5이고 B=4인 경우 K=4의 값을 가지게 된다.

  · SI||GUSIGUN : SI 와 GUSIGUN의(문자) 값을 옆으로 붙인다.
  예) REG=SI||GUSIGUN;
  여기서 SI='11' 이고 GUSIGUN='050'인 경우 REG는 '11050'

마. 자료저장 및 출력에 사용되는 명령문

작업의 최종결과 또는 중간결과의 데이타를 외부화일로 보관하여 놓으면 이 자료를 이용하는 다음 작업시 처음부터 새로 작업하지 않아도 된다. SAS 데이타 셋은 SAS 작업중에는 WORK.xxx 이라는 이름의 화일이 임시로 생겼다가 SAS 작업이 끝나면 WORK.xxx 화일이 없어진다.

FILE 문과 PUT 문을 사용하여 외부화일을 만들면 된다. 외부화일을 읽어 들이는 경우 INFILE 문과 INPUT 문과 비교하여 보면 「IN」만 생략된 형태이다. 작업이 끝나면 변수를 보관하는 외부화일(DOS)이 생긴다.

(1) FILE 문 : FILE file-spec [옵션들]

FILE 문은 현재 DATA STEP에서 PUT 문에 의해 생긴 OUTPUT 화일을 지정하는 명령이다. INFILE 문의 반대개념으로 이해하면 된다.

- OPTIONS

LINESIZE(LS)=line-size : 1행에 프린트할 수 있는 문자의 수
PAGESIZE(PS)=value : 한 페이지에 인쇄하고자 하는 행의 수
NOTITLES : 이미 사용된 타이틀을 인쇄하지 않을 때 사용

```
예) TITLE 'TEST PROGRAM TITLE' ;
      DATA _NULL_ ; SET ONE ;
         FILE PRINT NOTITLE ;
      PUT @1 REGION  $5.
          @10 SALES   6.  ;
      RUN ;
결과)  ----+----+----+
       SEOUL    12456
```

LINESLEFT=VARIABLE : 사용 가능한 LINE 수가 변수에 저장됨

```
예) DATA  _NULL_ ; SET ONE ;
     FILE PRINT LINESLEFT=L ;
    PUT @1 REGION  $5.
        @10 SALES   6.  ;
    IF  L < 7  THEN _PAGE_ @ ; RUN ;
```

LRECL=logical-record-length

HEADER=label : 각 page 마다 타이틀을 작성하는데 사용

```
예) DATA  _NULL_ ; SET ONE ;
     FILE PRINT HEADER=HTITLE ;
     PUT @1 REGION  $5.
         @10 SALES   6.  ;  RUN ;
```

LINE=variable : PRINT한 현재의 LINE위치가 변수에 저장

COLUMN=variable : PRINT한 현재의 COLUMN위치가 변수에 저장

N=PS(N=PAGESIZE, N=VALUE) : PRINT할 PAGE크기를 지정함.

```
예) OPTIONS PS=50 ;
    DATA  _NULL_ ; SET ONE ;
      FILE PRINT N=PS ;
      PUT @1 REGION  $5.
          @10 SALES   6.  ; RUN ;
```

(2) PUT 문 : PUT 변수명 과 출력양식

PUT 문은 SAS log, OUTPUT 화일, 또는 FILE 문에 의해 지정된 외부화일에 내용을 기록하고자 할 때 사용한다. PUT 문 앞에 FILE 문이 없으면 결과는 SAS log에 기록된다. INPUT 문의 반대개념으로 이해하면 된다. 형식은 INPUT 문과 꼭 같다.

```
예1)  DATA _NULL_;   X=11; Y=15;
      PUT X 10-19 .1  Y 20-29 .1; RUN;
결과) ----+----1----+----2----+----3----+--
               11.0       15.0
```

```
예2) DATA ONE;
      INPUT name & $15. score1 score2;
     PUT   name $15. +3 score1 5. score2 5.;
    CARDS;
    bill perkins   102  115
    roger          87   91
   RUN;
```

```
예3)  DATA ONE ; INPUT X Y;    결과) ----+----1----+----2-
      PUT x y;                       102 115
     CARDS;                          89 91
     102 115
     87  91
```

바. 만들어진 DATASET 이용

이미 만들어진 하나 또는 여러개의 데이터세트를 읽어 새로운 SAS데이타세트를 생성하고자 하는 경우에 SET/MERGE/UPDATE 문을 사용한다.

(1) SET 문

SET 문은 이미 만들어진 하나 또는 여러 개의 SAS Dataset로 새로운 SAS Dataset을 만들고자 하는 경우에 사용한다.

```
DATA 새로운 Dataset이름;
 SET 기존 Dataset이름(들);
자료변환등 SAS 프로그램문장들
예) DATA ALL ;
     SET A1 A2;  : Dataset A1, A2를 아래위로 합한다.
    DATA ONE ;
      SET ALL ;
     IF REG='11';  : Dataset ALL 중에서 지역이 11인 경우를 ONE으로
    DATA ALL;
      MERGE A1 A2 ;
      BY REG;  : Dataset A1, A2를 지역별(REG)로 옆으로 합친다.
```

- 데이타 셋의 합침(아래위로 합침)

A1

x	y
1	21
2	56

A2

x	y	z
1	33	11
2	56	21

ALL

x	y	z
1	21	.
2	56	.
1	33	11
2	56	21

- BY 문을 이용한 SET 문

A1

code	x	y
1	11	21
3	22	56

A2

code	x	y	z
2	33	11	5
3	45	6	7
5	56	21	6
7	76	43	7

ALL

code	x	y	z
1	11	21	.
2	33	11	5
3	22	56	.
3	45	6	7
5	56	21	6
7	76	43	7

(2) MERGE 문

앞에서 설명한 SET 문은 단순히 SAS Dataset을 아래위로 합하는 것에 지나지 않으나 MERGE 문은 옆으로 합치는 기능으로 변수를 추가하는 기능으로 생각하면 된다.

```
DATA 새로운이름;
  MERGE  기존이름들; ( BY 변수명; )
```

- 동일변수가 없는 경우

예)
```
DATA A1;
  INPUT X Y;
CARDS;
1 21
2 56
```

```
DATA A2 ;
  INPUT Z @@;
CARDS;
11 22 33
```

```
DATA ALL; MERGE A1 A2;
```

A1

x	y
1	21
2	56

A2

z
11
22
33

ALL

x	y	z
1	21	11
2	56	22
.	.	33

- 동일 변수(Y)가 있는 경우

예) DATA A1;
INPUT X Y @@ ;
CARDS;
1 21 2 56

DATA A2;
INPUT Y Z@@;
CARDS;
11 5 22 6 33 9

DATA ALL ; MERGE A1 A2;

A1

x	y
1	21
2	56

A2

y	z
11	5
22	6
33	9

ALL

x	y	z
1	11	5
2	22	6
.	33	9

- BY 문을 이용한 MERGE

MERGE를 이용하는 경우 가장 유용하면서 강력한 기능을 발휘하는 것이 바로 이 BY를 이용하는 MERGE 문의 경우이다. 예를들면 데이타의 분류코드체계가 서로 다른 경우 새로운 코드 또는 하나의 일관성있는 코드로 조정할 필요가 있는데 이 때 유용하게 사용될 수 있다.

예) DATA A1 ;
INPUT CODE1 $ X Y ;
CARDS;
11 1 21
51 2 56
41 3 77

DATA A2;
INPUT CODE1 CODE2 $;
CARDS;
11 A1
31 C1
61 D1

PROC SORT DATA=A1; BY CODE1;
PROC SORT DATA=A2; BY CODE1;
DATA ALL;
MERGE A1 A2; BY CODE1 ;

A1

code1	x	y
11	1	21
41	3	77
51	2	56

A2

code1	code2
11	a1
31	c1
61	d1

ALL

code1	x	y	code2
11	1	21	a1
31	.	.	c1
41	3	77	
51	2	56	
61	.	.	d1

(3) UPDATE 문

UPDATE 는 MERGE의 경우와 거의 유사하나 UPDATE는 반드시 BY 문을 사용해야 한다. Missing 데이타를 처리하는데 있어 앞 Dataset의 변수가 Missing이 아니고 뒤에 오는 Dataset의 변수가 Missing인 경우, MERGE를 쓰면 Missing이 되나 UPDATE는 원래의 데이타 값을 가진다.

```
DATA 새로운이름;
 UPDATE 기존이름들; BY 변수명;
```

예)

```
DATA A1;                          DATA A2;
   INPUT CODE $ X Y;                INPUT CODE $ Y Z ;
  CARDS;                           CARDS;
  11 1 2                           11   1  111
  51 2 5                           21   .  222
  21 4 3                           31   3  333
  41 3 7                           41   4  777
                                   71   7  777
 PROC SORT DATA=A1 ; BY CODE ;
 PROC SORT DATA=A2 ;BY CODE ;
 DATA ALL ;
  UPDATE A1 A2;  BY CODE ;
```

A1

code1	x	y
11	1	2
21	4	③
41	3	7
51	2	5

A2

code1	y	z
11	1	111
21	.	222
31	3	333
41	4	777
71	7	777

ALL

code1	x	y	z
11	1	1	111
21	4	③	222
31	.	3	333
41	3	4	777
51	2	5	.
71	.	7	777

4. PROC의 종류(SAS/BASE)

가. PROC PRINT

PROC PRINT;는 단순히 SAS Dataset에 있는 DATA를 프린트하는 PROC이다.

```
PROC PRINT options;
  VAR     variables;
  ID      variables;
  BY      variables;
  PAGEBY  byvariable;
  SUM     variables;
  SUMBY   byvariable;
```

- OPTIONS

DATA=SAS-dataset : 실행될 SAS-dataset 이름, 생략시는 가장 최근에 생성된 SAS-dataset를 사용.

DOUBLE , D : 출력결과가 한행씩 간격을 둔다.

NOOBS : 관측값들의 번호가 프린트되지 않는다.

LABEL : 변수명 대신에 해당 LABEL을 프린트한다.

N : Dataset 에 들어있는 관측치의 수를 프린트한다.

SPLIT=' *' : 프린트할 때 label로 준 이름을 두줄로 프린트할 경우

```
예) PROC PRINT split='*' ;          결과) this is
    label x='this is * a name' ;                a name
```

- VAR variables; 프린트될 변수명을 지정한다. 이 VAR문이 사용하지 않으면 Dataset에 들어있는 모든 변수들이 프린트된다.

- ID variable; 제시된 변수들의 값들을 출력결과에서 관찰값들을 구분하기 위해 관찰값들의 순서 대신에 사용한다. 하나의 표본 관찰값들이 너무 길어 한행에 인쇄할 수 없을 때 ID변수들의 값들을 관찰값에 대한 자료값을 포함하는 모든행의 처음에 프린트한다.

- BY variables; BY 변수들에 의해 분류된 집단(GROUP)마다 관찰값들이 분리되어 프린트된다. BY 문장을 사용할 경우에는 분석될 변수명으로 정렬(SORT)되어 있어야 한다. 정렬되어 있지 않으면 PROC SORT를 이용하여 먼저 정렬하여야 한다.

- PAGEBY byvariable; BY에서 지정된 변수의 값이 바뀔 때 새로운 page에 프린트된다.

- SUM variable; 지정한 변수의 합계를 프린트한다.

- SUMBY byvariable; BY에서 지정된 변수의 값이 바뀔 때마다 합계을 출력하고자 할 때 사용한다.

나. PROC SORT

PROC SORT는 SAS Dataset을 한개이상의 변수를 크기순으로 정렬하여 새로운 SAS Dataset을 만드는데 사용한다. SAS의 모든 PROC에서 BY 문을 사용하는 경우 BY 뒤의 변수명으로 정렬되어 있어야 하는데 이 때 PROC SORT를 이용하여 미리 정렬해야 한다.

정렬하고자 하는 변수가 숫자인 경우에는 missing, 음수, 영, 양수의 순으로 정렬된다. 문자인 경우에는 다음과 같은 순서로 정렬된다. 이 순서는 다음의 OPTIONS에 「REVERSE」를 사용하든지 「DESCENDING」을 사용하여 역으로 정렬할 수 있다.

빈칸!"#$%&'()*+-./0123456789:;<=>?@
ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]_
abcdefghijklmnopqrstuvwxyz{|}~

```
PROC SORT options;
  BY option variable option variable ···;
```

- OPTIONS
  DATA=SAS-dataset
  OUT=SAS Dataset : 실행된 결과를 보관할 Dataset 명을 지정, 생략시에는 정렬된 자료가 원래 Dataset 이름으로 저장된다.

NODUPKEY : Output Dataset을 만들 때 BY 뒤에 있는 변수의 값이 중복되는 관측치는 제외한다.

NODUPREC 또는 NODUP : Output Dataset 을 만들 때 중복된 관측치는 제외한다. 이는 관측치의 모든 변수의 값들이 정확하게 일치하는 경우에만 제외한다는 것을 의미한다.

- BY option variable option variable ···; option에는 DESCENDING(내림차순)을 사용할 수 있으며 생략시에는 ASCENDING 옵션이 사용된다.

예) PROC SORT DATA=one ; BY DESCENDING size age;
변수 size는 내림차순으로 정렬하고, 변수 age는 올림차순으로 정렬한다.

다. PROC TRANSPOSE

PROC TRANSPOSE는 SAS Dataset을 전치(Transpose)하는 기능으로 변수를 관측치로, 관측치를 변수로 전치한다. 즉 행을 열로, 열을 행으로 바꾼다. 일반적인 DATABASE 관리하는데 편리하게 사용된다.

```
PROC TRANSPOSE OPTIONS;
  VAR      variable-list;
  ID       variable;
  IDLABEL  variable;
  COPY     variable-list;
  BY       variable-list;
```

- OPTIONS

DATA=SAS-dataset :

OUT=SAS Dataset : 실행한 뒤의 Dataset 을 보관하는 문장

PREFIX=접두어 : PREFIX를 지정하지 않으면 전치된 Dataset의 변수명은 COL1, COL2, ··· 란 이름으로 자동적으로 지정된다.

NAME=변수명 : TRANSPOSE하기전에 변수명들이 있는 GROUP의 변수명

- VAR variables; 열과 행을 바꾸고자 하는 변수을 지정한다.

- ID variable: 변수값으로 변수명을 사용하고자 할 때, 지정하지 않으면 변수명이 COL1 COL2 ... COLn이 된다.

- BY variables: BY 변수들에 의해 분류된 집단(GROUP)별로 열과 행을 바꾸고자 할 때 사용한다.

- PROC TRANSPOSE의 예

```
예) DATA a1;
      INPUT a b c;
    CARDS;
      1  2  3
      4  5  6
      7  8  9
      10 11 12
 PROC TRANSPOSE DATA=A1 OUT=A2 ;
 PROC TRANSPOSE DATA=A2 OUT=A3 ;
 PROC PRINT DATA=A1;
 PROC PRINT DATA=A2;
 PROC PRINT DATA=A3;
```

결과 A1)

OBS	A	B	C
1	1	2	3
2	4	5	6
3	7	8	9
4	10	11	12

결과 A2)

OBS	_NAME_	COL1	COL2	COL3	COL4
1	A	1	4	7	11
2	B	2	5	8	22
3	C	3	6	9	33

결과 A3)

OBS	_NAME_	A	B	C
1	COL1	1	2	3
2	COL2	4	5	6
3	COL3	7	8	9
4	COL4	10	11	12

라. PROC FREQ

PROC FREQ 는 빈도 및 누적빈도를 알아보는 데 사용되는데, 일반적인 설문지 분석에서 가장 많이 사용되고 있다.

```
PROC FREQ OPTIONS1;
   TABLES  requests/OPRIONS2;
   WEIGHT  variable;
   BY      variables;
```

- OPTIONS1

  DATA=SAS-dataset ;

  ORDER=FREQ, DATA, INTERNAL, FORMATTED

  ORDER=FREQ : 계급에 해당하는 값의 관측치가 많은 순서대로

  ORDER=DATA : 데이타셋에 나타나는 계급값의 순서대로

  ORDER=INTERNAL : 계급변수가 갖는 값 순서대로

  ORDER=FORMATTED : 계급변수가 가지는 여러 값에 지정한 이름순서대로

- TABLE requests/OPTIONS2;

  하나의 PROC FREQ 문에 여러 개의 TABLES 문을 사용할 수 있다. TABLES 문이 없으면 Dataset에 들어있는 모든 변수에 대해서 1차원(Oneway) 빈도분포를 프린트한다. 1차원 도수분포표에서는 기본적으로 빈도수, 누적빈도수, 구성비, 누적구성비를 프린트하고 2차원 도수분포표에서는 Cell빈도, Cell 퍼센트, 행 퍼센트, 열 퍼센트를 프린트한다.

  Missing 데이타는 제외되고 각 도수분포표 아래에 Missing 데이타의 갯수를 별도로 프린트한다. TABLE 형식을 지정하는 「requests 문」은 프린트하고자 하는 변수의 분할표 형태를 지정한다.

- OPTIONS2

  MISSING : Missing 값을 의미있는 수준(level)으로 간주하여 별도의 그룹을 형성하게 한다. 생략하면 고려대상에서 제외된다.

  FORMAT=formatname : 각 테이블셀(Table cell)의 형식을 지정하는데 사용한다. Default 는 12.2 이다.

  EXPECTED : 각 CELL의 기대빈도를 프린트한다.

  DEVIATION : 각 CELL의 실제 빈도-기대빈도를 프린트한다.

CELLCHI2 : 각 CELL 마다 (실제빈도-기대빈도)**2/기대빈도를 프린트하며, 이는 자유도 1인 $\chi 2$분포를 따르며, 모든 CELL에 대한 합이 $\chi 2$값이다.

CUMCOL : 각 CELL의 누적행 퍼센트를 프린트한다.

CHISQ : 독립성검정을 하는데 필요한 $\chi 2$의 결과를 프린트 한다.

ALL : 모든 통계치를 프린트한다.

NOFREQ : 분할표에서 실제 빈도를 프린트하지 않는다.

NOPERCENT : 분할표에서 CELL빈도 / 전체 빈도 를 프린트하지 않는다

NOROW : 분할표에서 CELL빈도 / 행빈도 를 프린트하지 않는다

NOCOL : 분할표에서 CELL빈도 / 열빈도 를 프린트하지 않는다

MISSING : 빈도계산에서 missing 값을 포함하여 계산한다.

OUT=SAS-dataset : 변수이름과 해당빈도를 SAS Dataset으로 보관하고자 할 때 사용한다. 각 수준에 해당하는 관측치 수를 TABLES 뒤의 형식에 따라 "COUNT"란 변수명으로 저장된다. 이 옵션은 일반사용자들이 잘 모르는 것으로 자료 처리에 상당히 유용하게 사용될 수 있다.

NOPRINT : 이는 PROC FREQ의 결과를 프린트하지 않을 때 사용하는 것으로 OUT=SAS-dataset을 사용할 때와 같이 SAS Dataset만을 만들고자 하는 경우 사용한다.

표1) 분할표의 여러가지 형태

표 현 방 법	설 명
TABLE A	A의 단순빈도표를 프린트한다.
TABLE A*B	A, B의 2 차원 빈도표를 프린트한다.
TABLE A*B*C	A, B, C의 3 차원 빈도표를 프린트하는데 A의 각 값에 따라 B*C의 2 차원 분할표를 프린트한다.
TABLE A*(B C)	TABLES A*B A*C와 동일
TABLES (A B)*(C D)	TABLES A*C B*C A*D B*D와 동일
TABLES A--C	TABLES A B C와 동일
TABLES (A--C)*D	TABLES A*D B*D C*D와 동일
TABLES A--C*D	에러

- WEIGHT variable; 각 관측값에 가중치를 적용하여 분석하고자할 때 사용.

- BY variables; BY 변수들로 구분된 각 수준(level)별로 분석하고자 할 때 사용.

마. PROC TABULATE

PROC TABULATE는 통계보고서를 작성할 때 사용되는 과정이다.

```
PROC TABULATE options1 ;
  CLASS     variables;
  VAR       variables;
  FREQ      variable ;
  WEIGHT    variable ;
  FORMAT    variable format. ···;
  LABEL     variable=label···;
  BY        variables;
  TABLE     dimension_expression/options2
  KEYLABEL keyword='text'···;
```

- OPTIONS1

  DATA=SAS-dataset

  MISSING : Missing값을 의미있는 수준으로 간주하여 별도의 그룹을 형성하게 끔 한다. 생략하면 고려대상에서 제외된다.

  FORMAT=format-name : format=w.d,commaw. default(f=12.2)

  ORDER=FREQ, DATA, INTERNAL, 또는 FORMATTED :

  ORDER=FREQ : 계급에 해당하는 값의 관측치가 많은 순서대로

  ORDER=DATA : 데이타셋에 나타나는 계급값의 순서대로

  ORDER=INTERNAL : 계급변수가 갖는 값 순서대로

  ORDER=FORMATTED : 계급변수가 가지는 여러값에 지정한 이름순서대로

- CLASS variables; 변수명들은 문자 , 숫자에 관계없이 GROUP화 할 수 있는 변수.

- VAR variables; 분석될 변수(NUMERIC)를 지정.

- FORMAT variable format. ···; CLASS에서 적용된 GROUP변수를 PROC FORMAT ; 에서 정의된 변수명으로 대치시켜 줄 경우.

  예) PROC FORMAT ;
  VALUE SEXFMT 1='남자' 2='여자' ;
  PROC TABULATE ; CLASS SEX ; TABLE SEX ;
  FORMAT SEX $SEXFMT. ;

- LABEL variable=label···; 변수명을 다른 text로 바꿀 때.
  예) LABEL REG='지역' ;

- BY variables; BY 변수들로 구분된 각 수준(level)별로 분석하고자 할 때.

- TABLE dimension-expression/options; :

TABLE 문은 3차원까지 사용가능하고 컴마로 구분하며 각각 차원에 대해 옵션을 지정하게 된다. 3차원인 경우 가장 왼쪽의 차원은 페이지를 지정하고 두 번째 차원은 행, 가장 오른 쪽의 차원은 열을 가리키게 된다. 1차원만 지정된 경우에는 열을 의미한다.
TABLE의 표현방식은 X*(Y Z)와 같이 하나 이상의 Operand와 연산자(Operator)로 구성된다.

- 통계량을 지정하는 Operand
Operand는 분석하고자 하는 변수에 SUM, STD과 같이 통계량을 지정하는 것과 퍼센트를 나타내는 PCTN(빈도율), PCTSUM(합의비율), 그외에 N, NMISS, MEAN, MIN, MAX, RANGE, VAR, USS, CSS, STDERR CV, T, PRT, SUMWGT 등이 있다. 단, PCTN는 빈도(N)의 퍼센트를, PCTSUM은 합계(SUM)의 퍼센트를 지정한다.

예) TABLE X, Y * Z *SUM;

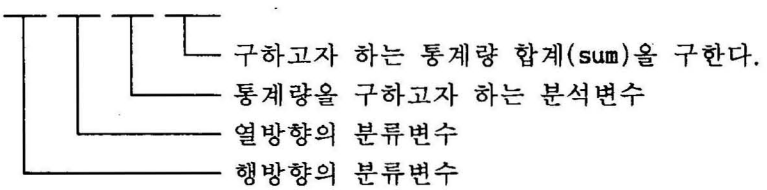


표1) 연산자의 표현방법

표현방법	설 명
*	부집단(Subgroup)
빈칸(space)	테이블의 연속
괄호()	그룹 또는 순서를 지정

예) 분류변수 A,B,C,D이고 통계량을 구하고자 하는 변수가 X인 유효한 표현들

A	:	분류변수를 지정하는 경우
X	:	통계량을 구하는 변수를 지정하는 경우
A*B	:	분류변수 A, B의 2차원으로 구하는 경우
A*B*C	:	분류변수 A, B, C의 3차원으로 구하는 경우
A*MEAN*X	:	분류변수 A에 대하여 X의 평균값을 구하는 경우
A B	:	A와 B에 대해서 1차원으로 연속해서 구하는 경우
A B*C	:	A에 대해 1차원, B와 C에 대해 2차원으로 연속해서 구하는 경우
(A B)*C	:	(A와 B) 그리고 C에 대해서 2차원적으로 구하는 경우 A*C B*C와 같다.
(A B)*(C D)*E	:	(A와 B),(C와 D) 그리고 E에 대해서 3차원으로 구한 경우

- OPTIONS2

RTS=n : row stub의 size를 표시함, default로는 linsize의 1/4값임.

FUZZ=n : n보다 적은 분석변수를 print하는 과정에서 '0'으로 대치됨.

BOX='text' : table이 2차원인 경우에 row stub 상단과 column stub 좌우부분 box에 설명 text를 삽입할 수 있다.

CONDENSE : 가능한 한페이지에 표가 나타나도록 한다.

(9) KEYLABEL

TABLE문에 주어진 분석변수를 rename시킬 경우

예) keylabel all='전국' sum='합계' mean='평균' ;

바. PROC MEANS, SUMMARY, UNIVARIATE

각 변수들의 기술통계량을 구하는 과정이다. 즉 평균, 분산, 표준편차, 최대값, 최소값, 범위, 자승합, 표준오차 등을 구할 수 있다.

(1) PROC MEANS의 형태

```
PROC MEANS OPTIONS ;
  VAR     variables;
  BY      variables;
  CLASS   variables;
  FREQ    variable;
  WEIGHT  variable;
  ID      variables;
  OUTPUT OUT=SASdataset Keyword=변수명;
```

- OPTIONS
  DATA=SAS DATASET
  NOPRINT
  MAXDEC=N
  VARDEF=DF,WGT,N,WDF(DEFULT=DF)

- OUTPUT OUT=SASDATASET
  keyword=변수명 : 구하고자하는 통계치 지정
  N(관측치수), MEAN (평균), STD(표준편차), MIN(최소값),
  MAX(최대값), RANGE(범위), SUM(합계), VAR(분산),
  USS(수정안된 자승합), CSS(수정된 자승합), STDERR(표준오차),
  CV(변이계수), T(t통계량), PRT(t통계량 유의확률),
  SUMWGT(가중치 합), NMISS(missing value의 개수),

- CLASS variables; 부그룹별로 기술통계량을 구하고자 할 때 부그룹을 지정한다. CLASS 문은 BY 문과 그 기능이 거의 같으나 출력양식이 조금 다르고, BY 문은 정렬(Sort)이 되어 있어야 사용가능하다.

(2) PROC UNIVARIATE의 형태

```
PROC UNIVARIATE options;
  VAR      variables;
  BY       variables;
  FREQ     variables;
  WEIGHT   variables;
  ID       variables;
  OUTPUT OUT=SASdataset keyword=변수명;
```

- OPTIONS

  DATA=SAS-dataset :

  NOPRINT : 프린트하지 않을 경우

  PLOT : stem-and-leaf 그림, box-plot, 정규분포 plot를 그린다.

  FREQ : 도수분포, 빈도, 퍼센트, 누적퍼센트를 구한다.

  NORMAL : 입력자료가 정규분포를 따르는지에 대한 검정통계량을 구한다.

- OUTPUT OUT=SASdataset Keyword=변수명;

  OUT=SAS-dataset

  Keyword=변수명 : 구하고자 하는 통계치 지정

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

(3) PROC SUMMARY의 형태

```
PROC SUMMARY options;
  CLASS    variables;
  VAR      variables;
  BY       variables;
  FREQ     variable;
  WEIGHT   variable;
  ID       variables;
  OUTPUT   OUT=SASdataset Keyword=변수명;
```

- OPTIONS

  DATA=SAS-dataset

  PRINT : 기술통계량을 프린트할 경우에 사용한다.

  MAXDEC=n : 통계치를 프린트할 때 소수 몇자리까지 할 것인가?

  FW=n : 통계치를 프린트할 때 각 통계치의 필드의 폭을 지정한다.

  MISSING : CLASS variables에서 지정한 변수가 Missing인 경우에도 별도의 그룹을 형성하게 끔 한다.

  NWAY : 가장 높은 _TYPE_ 값을 가진 통계량만을 OUTPUT으로 한다.

- OUTPUT OUT=SASdataset Keyword=변수명;

  OUT=SAS-dataset

  keyword=변수명 : 구하고자하는 통계치 지정

  N(관측치수), NMISS(missing value의 개수), MEAN(평균),
  STD(표준편차), MIN(최소값), MAX(최대값), RANGE(범위),
  SUM(합계), VAR(분산), USS(수정안된 자승합),
  CSS(수정된 자승합), STDERR(표준오차), CV(변이계수),
  T(t통계량), PRT(t통계량의유의확률), SUMWGT(가중치의 합)

- CLASS variables: 부그룹별로 기술통계량을 구하고자 할 때 부그룹을 지정한다. CLASS 문은 BY 문과 그 기능이 거의 같으나 출력양식이 조금 다르고, BY 문은 정렬(Sort)이 되어 있어야 사용가능하다.

- CLASS에 기술된 GROUP변수에 따른 _type_ 형태

예1) CLASS x ; 인 경우

type

0 : 총괄적인 통계치들

1 : x의 각 group별 통계치들

예2) CLASS x y; 인 경우

type

0 : 총괄적인 통계치들

1 : y의 각 group별 통계치들

2 : x의 각 group별 통계치들

3 : x의 각 group에 따른 y의 group별 통계치들(x*y)

예3) CLASS x y z; 인 경우

type

0 : 총괄적인 통계치들

1 : z의 각 group별 통계치들

2 : y의 각 group별 통계치들

3 : y의 각 group에 따른 z의 group별 통계치들(y*z)

4 : x의 각 group별 통계치들

5 : x의 각 group에 따른 z의 group별 통계치들(x*z)

6 : x의 각 group에 따른 y의 group별 통계치들(x*y)

7 : x의 각 group에 따르고 y의 group에따른 z의 group별 통계치들 (x*y*z)

- VAR variables; 구하고자 하는 변수를 지정.

- FREQ variable; VAR변수들의 각 관측값들의 빈도수는 대응하는 FREQ변수의 값으로 나타난다. FREQ 뒤의 변수의 값이 Missing이거나 1 보다 적으면 계산에서 제외되고 정수가 아닌 경우에는 정수 부분만을 고려한다.

- WEIGHT variable; 각 관측값에 가중치를 적용하여 분석하고자할 때 사용.

- BY variables; BY 변수들로 구분된 각 수준(level)별로 분석할 때 사용.

- ID variable; 프린트에는 관측치의 번호가 프린트되는데 이 관측치의 일련 번호대신 ID 뒤에 지정한 변수값이 프린트된다.

사. PROC PLOT

PLOT는 2차원의 그래프를 그리는 PROC으로 자료분포, 자료에러체크, 상관분석, 회귀분석 등에 많이 사용된다.

```
PROC PLOT OPTIONS1;
  BY    variables;
  PLOT  request-list/OPTIONS2;
```

- OPTIONS1

  DATA=SAS-dataset

  VPERCENT 또는 VPCT=value : 한 페이지 수직크기를 지정하는 것으로 지정된 값은 % 를 나타내며 해당하는 크기만큼 한 페이지에 축소하여 그림을 그린다.

  예) VPERCENT=33 : 수직으로 한 페이지에 3개의 그림을 그릴 수 있는 크기만큼 축소한다.

  VPERCENT=50 25 25 : 한 페이지에 3개의 그림을 그리는데 첫 번째 그림은 다른 것의 두배이다.

  HPERCENT 또는 HPCT =value : 한 페이지 수평크기를 지정하는 것으로 지정된 값은 % 를 나타내며 해당하는 크기만큼 한 페이지에 축소하여 그림을 그린다.

  UNIFORM : PLOT 문에서 BY 문을 사용하면 BY 뒤의 변수수준마다 그래프를 그리는데 각 그래프마다 Scale이 다르게 프린트된다. 이를 균등한 Scale로 하는 경우에 사용한다.

  NOMISS : 가로, 세로축을 고려할 때 Missing 데이타를 고려대상에서 제외시킨다.

- PLOT request-list/OPTIONS2;

  · VERTICAL * HORIZONTAL : 수직축과 수평축에 나타내고자 하는 변수명을 지정한다.

  예) PROC PLOT ; PLOT Y*X;

  Y는 수직축, X는 수평축에 그린다.

· VERTICAL * HORIZONTAL='CHARACTER' : 그래프상에 한개 이상의 관측치를 표시할 때 지정된 문자로 위치를 나타낸다.

예) PROC PLOT;PLOT Y*X='+';
Y는 수직축, X는 수평축에 그리는데 단 그 위치는 + 로 표시된다.

· VERTICAL * HORIZONTAL=VARIABLE : 그래프상에 한개 이상의 관측치를 표시할 때 지정된 변수의 첫 글자로 위치를 나타낸다.

예) PROC PLOT;PLOT HEIGHT*WEIGHT=SEX;

- OPTIONS2

VAXIS=숫자 : Y축에 등간격으로 표시하고자 하는 수

예) PROC PLOT;PLOT y*x/VAXIS=0 TO 12 BY 2;
PROC PLOT;PLOT y*x='+'/VAXIS= 0 TO 12 BY 2;

HAXIS=숫자 : X축에 등간격으로 표시하고자 하는 수
VREF =숫자 : Y축에 그리고자 하는 값의 범위
HREF =숫자 : X축에 그리고자 하는 값의 범위
BOX : 그래프주위에 경계선을 그리고자 할 때

예) PROC PLOT;PLOT y*x=z/HAXIS= 0 TO 12 BY 2 BOX;
PROC PLOT;PLOT y*x=z/HAXIS= 0 TO 12 BY 2 HREF=4 8;

OVERLAY : 한 페이지에 두가지 이상의 그래프를 그리고자 할 때

예) PROC PLOT;PLOT y*x=z a*b/OVERLAY;

아. PROC CHART

수평, 수직의 히스토그래프, 블럭차트(Block Chart), 파이차트(Pie chart)등을 그리는 데 사용한다.

```
PROC CHART  OPTIONS1 ;
  BY      variables;
  VBAR    variables/OPTIONS2;
  HBAR    variables/OPTIONS2;
  BLOCK   variables/OPTIONS2;
  PIE     variables/OPTIONS2;
  STAR    variables/OPTIONS2;
```

- OPTIONS옵션 1
  DATA=SAS-dataset
  LPI =p : 결과를 프린트할 때 그래프의 간격을 조정한다.
  p는 (1인치당 라인수/1인치당 열수)*10 이다.(Default=6)
  예) 1인치당 8행, 12열을 인쇄하는 프린터의 경우에는
  LPI=6.6667 (8 / 12 * 10) 이다.

- BY variables; 분류된 GROUP별로 그림표를 그릴 때.

- VBAR variables/OPTIONS2; 수직 막대그림표.

- HBAR variables/OPTIONS2; 수평 막대그림표

- BLOCK variables/OPTIONS2; 3차원으로 구성분포

- PIE variables/OPTIONS2; 원형의 구성분포.

- STAR variables/OPTIONS2; 별모양의 구성분포.

- VBAR, HBAR, BLOCK, STAR에 사용되는 OPTIONS2
  DISCRETE : 숫자차트변수가 연속적이기 보다는 이산적인 경우에 사용한다.
  생략되면 PROC CHART는 모든 변수가 연속적인 것으로 간주한다.
  TYPE=FREQ : 도수, DEFAULT.

TYPE=PERCENT 또는 PCT : 백분율.
TYPE=CFREQ : 누적도수.
TYPE=CPERCENT 또는 CPCT : 누적백분율.
TYPE=SUM : SUMVAR= 에서 지정된 변수의 합계.
TYPE=MEAN : SUMVAR= 에서 지정된 변수의 평균.
SUMVAR=변수명
예) VBAR dept/ TYPE=MEAN SUMVAR=sales;
각 dept의 값에 대하여 변수 sales의 평균값.
MIDPOINT=값 : 차트에 사용되는 변수의 중앙값 지정.
예) VBAR x / MIDPOINTS= 10 20 30 40 50;
5개의 막대를 그릴 수 있는데 첫 번째 막대는 중앙값 10을 가진다.
FREQ=variable : 관찰값에 대한 도수를 나타낼 때.
AXIS=value : 축을 만드는데 사용하는 최대값을 지정
LEVELS=n : 막대그림표의 수준수를 명시

- VBAR, HBAR, BLOCK에 사용되는 추가 OPTIONS2
GROUP=변수명 : 병렬도표를 만들 때, 변수는 문자나 숫자 일 수 있고 이산형이다.
예) VBAR sex/ GROUP=dept;
각 dept 별로 남녀의 빈도차트를 그린다.
SUBGROUP=변수명 : 한 막대를 변수명의 각 값에 따라 분리·표시하여 그린다.
SYMBOL='char' : 그림표의 구조에 사용되는 기호를 명시.
NOSYMBOL : 도표밑에 인쇄된 기호에 관한 설명을 삭제.
NOPZEROS : ZERO값을 가진 그림표를 삭제.

- VBAR과 HBAR에 사용되는 추가 OPTIONS2
ASCENDING : 그래프를 그리는데 그룹내에서 오름차순으로 그래프와 관련통계량을 프린트한다.
DESCENDING : 그래프를 그리는데 그룹내에서 내림차순으로 그래프와 관련통계량을 프린트한다.

- HBAR 에 사용되는 추가옵션 2
NOSTAT : 통계량들을 프린트하지 않을 때.
FREQ : 수평막대그림표에서 도수가 도표옆에 인쇄할 때.
CFREQ : 누적도수가 인쇄되도록 할 때.
PERCENT : 관측값들의 백분율을 프린트할 때.
CPERCENT : 누적백분율을 인쇄할 때.
SUM : 관찰값의 총합을 명시.
MEAN : 관찰값의 평균을 명시.

가 설 검 정

목 차

Ⅰ. 표본분포

1. 표본의 추출

통계조사의 대상이 되는 집단 즉 모집단은 일반적으로 아주 크다. 그러므로, 전체 모집단을 모두 조사하는 것은 엄청난 비용과 시간을 필요로 한다. 이 때 모집단에서 일부를 추출한 표본을 이용하여 전체 모집단의 속성을 예측하는 것을 **추측통계**(inferential statistics)라 한다. 그러나, 모집단의 조사결과와 표본의 조사결과는 차이가 있기 마련이다. 이러한 차이를 줄이기 위해 여러 가지 표본의 추출방법이 연구되어 왔는데, 이 중 통계학에서 많이 사용되는 추출법은 단순확률추출법(simple random sampling)이다.

> **단순확률추출법**
>
> 모집단의 모든 원소가 표본으로 뽑힐 확률이 같도록 표본을 추출하는 방법

단순확률 표본추출시 한번 추출한 원소를 다시 모집단에 포함시키는 복원추출(with replacement)이나, 추출된 원소를 다시 모집단에 넣지 않는 비복원추출(without replacement) 모두 가능하나 실제 거의 모든 표본추출은 비복원추출로 이루어진다.

실제 표본추출시 모집단의 각 원소가 표본으로 뽑힐 확률이 같도록 하려면 어떠한 수단이 필요한데, 대개 난수표(random number table)를 많이 사용한다. 난수표란 0 에서 9까지의 숫자를 특별한 규칙성이나 편중성이 없이 흩어 놓은 표이다. <표 5.1>은 부록에 실은 난수표의 일부분이다.

<표 5.1> 난수표

85967	73152	14511	85285	36009	95892	36962	67835	...
07483	51453	11649	86348	76431	81594	95848	36738	...
96283	01898	61414	83525	04231	13604	75339	11730	...
.....	.....	.....	.....	.....	.....	.....	.....	...
.....	.....	.....	.....	.....	.....	.....	.....	...

이러한 난수표를 이용한 다음의 단순확률 표본추출 예제를 살펴보자.

[예 5.1] 어느 공장에 50명의 생산직 근로자의 주당 제품 생산량이 아래와 같다. 이 모집단에서 10명의 표본을 비복원으로 단순확률 추출하라.

번호	생산량	번호	생산량	번호	생산량	번호	생산량	번호	생산량
01.	30	11.	49	21.	58	31.	48	41.	59
02.	38	12.	64	22.	38	32.	35	42.	56
03.	33	13.	62	23.	71	33.	26	43.	65
04.	49	14.	37	24.	47	34.	62	44.	50
05.	33	15.	25	25.	65	35.	51	45.	54
06.	43	16.	32	26.	54	36.	67	46.	61
07.	60	17.	65	27.	74	37.	30	47.	57
08.	31	18.	56	28.	36	38.	57	48.	55
09.	34	19.	55	29.	62	39.	50	49.	26
10.	61	20.	43	30.	31	40.	62	50.	41

(실제로는 50명의 작은 모집단에서 표본을 추출할 필요가 없지만 설명을 위한 예이다.)

〈풀이〉

50명중 10명의 표본을 추출하기 위해서는 난수표의 아무데서나 시작해서 두 자리 숫자(모집단의 크기 50이 두 자리 수이기 때문)를 밑으로 읽어 내려간다. 만일에 왼쪽 상단부터 밑으로 두 자리 수를 읽어 적어보면 다음과 같다.

85 07 96 49 97 90 28 25 28 84 41 67 72 92 29 74 03 75 09 75
21 65 84 46 59 31 82 01 32 59 ...

여기서 차례로 근로자 번호를 추출하면 되는데, 만일에 난수가 01에서 50 사이를 벗어나면 버린다. 또 같은 숫자가 나오면 비복원추출이므로 뒤에 나온 숫자는 버린다. 그러면 10명의 단순확률 추출된 표본은

근로자 번호 : 07 49 28 25 41 29 03 09 21 46

이고, 각 근로자의 생산량은

생 산 량 : 60 26 36 65 59 62 33 34 58 61

이러한 난수표 이용법은 너무 많은 난수를 버리게 되므로 난수를 모집단의 크기로 나눈 나머지 수(modulo)를 사용하기도 한다. 즉, 처음 10개의 두 자리 난수를 모집단의 크기 50으로 나눈 나머지 수

35 07 46 49 47 40 28 25 28 34

를 추출할 근로자 번호로 정하는 것이다. 나머지가 0 이 되면 번호 50의 표본을 추출한다. 같은 숫자가 나올 경우는 버리고, 다음 난수를 이용한다. □

난수를 만드는 방법은 여러 가지가 있는데 최근에는 컴퓨터를 이용하여 난수를 많이 만든다. CATS를 이용하여 필요한 난수를 추출할 수도 있는데 5절에서 살펴보기로 하자.

2. 표본평균의 분포

통계적 실험이나 조사의 목적은 모집단에 대한 정보를 알아보려고 하는 것이다. 모집단의 정보란 대개 모평균, 모분산, 모비율 등과 같은 모집단의 특성값을 말한다. 이러한 모집단의 특성값을 모수(parameter)라고 한다. 모집단 전체를 조사하는 것은 불가능하거나 시간, 경비가 많이 들기 때문에, 대개 모수는 표본을 추출하여 표본평균, 표본분산, 표본비율과 같은 표본의 특성값을 이용하여 추정하게 된다. 이러한 표본의 특성값을 통계량(statistic)이라 부르고, 표본통계량의 분포를 표본분포(sampling distribution)라 한다. 이 표본분포는 표본통계량과 모수 사이의 관계를 규명해 주기 때문에 모수의 추정과 검정을 가능케 한다. 이 절에서는 먼저 표본평균의 분포를 다음의 예를 이용하여 알아보자.

[예 5.2] 한 회사의 영업사원 10명을 모집단이라 하자. 관심 있는 확률변수는 이 회사에서의 근무년수인데 다음과 같다.

3, 6, 2, 4, 8, 7, 9, 5, 1, 10

이 모집단의 평균과 분산을 구하고, 여기서 표본의 크기(n)가 2인 모든 가능한 표본들을 단순확률 복원추출하여 그 표본평균들의 분포를 구하라. (이렇게 작은 모집단은 실제로는 굳이 표본을 추출할 필요가 없지만, 여기서는 표본평균의 분포를 설명하기 위한 예이다.)

<풀이>

모집단의 모평균은 $\mu = 5.5$, 모분산은 $\sigma^2 = 8.25$ 이다. 모든 가능한 표본의 개수는 $10 \times 10 = 100$개인데, 이들 각각의 표본들과 그 표본평균 모두를 적어보면 <표 5.2>와 같고 <표 5.3>은 표본평균들의 도수분포표다. 이러한 표본평균의 도수분포표(<표 5.3>)를 n = 2 일 때의 표본평균의 분포(sampling distribution of sample means)라고 한다. <그림 5.1>은 모집단의 분포와 표본평균의 분포를 막대그래프로 나타낸 것이다. 표에서 보듯이 각각의 표본평균과 모평균은 서로 다르다. 하지만 이 표본평균들은 모평균 5.5 주위에 많이 몰려 있음을 <표 5.3>을 살펴보면 알 수 있다. 또 100개의 모든 표본평균들의 평균은 5.5 이고 분산은 4.125 이다.

이 표본평균의 분포와 모집단의 분포를 자세히 살펴보면 다음의 세 가지 중요한 사실을 관찰할 수 있다.

〈표 5.2〉 모집단에서 추출 가능한 n=2인 모든 표본들과 표본평균

표본	$\bar{x}$	표본	$\bar{x}$	표본	$\bar{x}$	표본	$\bar{x}$	표본	$\bar{x}$
1,1	1.0	3,1	2.0	5,1	3.0	7,1	4.0	9,1	5.0
1,2	1.5	3,2	2.5	5,2	3.5	7,2	4.5	9,2	5.5
1,3	2.0	3,3	3.0	5,3	4.0	7,3	5.0	9,3	6.0
1,4	2.5	3,4	3.5	5,4	4.5	7,4	5.5	9,4	6.5
1,5	3.0	3,5	4.0	5,5	5.0	7,5	6.0	9,5	7.0
1,6	3.5	3,6	4.5	5,6	5.5	7,6	6.5	9,6	7.5
1,7	4.0	3,7	5.0	5,7	6.0	7,7	7.0	9,7	8.0
1,8	4.5	3,8	5.5	5,8	6.5	7,8	7.5	9,8	8.5
1,9	5.0	3,9	6.0	5,9	7.0	7,9	8.0	9,9	9.0
1,10	5.5	3,10	6.5	5,10	7.5	7,10	8.5	9,10	9.5
2,1	1.5	4,1	2.5	6,1	3.5	8,1	4.5	10,1	5.5
2,2	2.0	4,2	3.0	6,2	4.0	8,2	5.0	10,2	6.0
2,3	2.5	4,3	3.5	6,3	4.5	8,3	5.5	10,3	6.5
2,4	3.0	4,4	4.0	6,4	5.0	8,4	6.0	10,4	7.0
2,5	3.5	4,5	4.5	6,5	5.5	8,5	6.5	10,5	7.5
2,6	4.0	4,6	5.0	6,6	6.0	8,6	7.0	10,6	8.0
2,7	4.5	4,7	5.5	6,7	6.5	8,7	7.5	10,7	8.5
2,8	5.0	4,8	6.0	6,8	7.0	8,8	8.0	10,8	9.0
2,9	5.5	4,9	6.5	6,9	7.5	8,9	8.5	10,9	9.5
2,10	6.0	4,10	7.0	6,10	8.0	8,10	9.0	10,10	10.0

(1) 모든 가능한 표본평균들의 평균($\mu_{\bar{x}}$ 로 표시)은 모평균과 같다. 즉 위의 〈표 5.1〉에 계산된 100개 표본평균들의 평균은 모평균과 같은 5.5가 된다.

$$\mu_{\bar{x}} = \frac{1 + 1.5 + \ldots + 10}{100} = 5.5$$

(2) 모든 가능한 표본평균들의 분산($\sigma_{\bar{x}}^2$ 로 표시)은 모분산을 표본의 크기로 나눈 값이다. 즉

$$\sigma_{\bar{x}}^2 = \frac{(1-5.5)^2 + (1.5-5.5)^2 + \ldots + (10-5.5)^2}{100} = 4.125$$

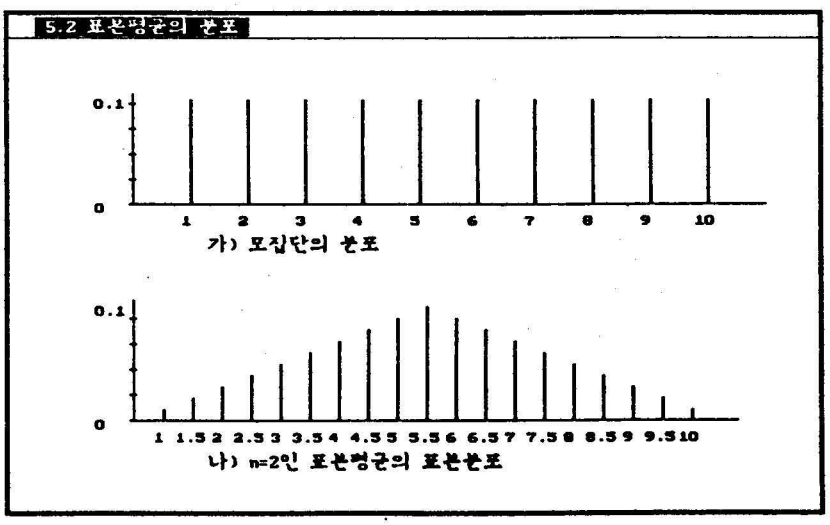
는 모분산 $\sigma^2 = 8.25$ 을 표본의 크기 2로 나눈 값이다. 따라서 표본의 크기가 커질수록 $\sigma_{\bar{x}}^2$ 은 점점 작아져 표본평균의 분포가 모평균에 점차로 밀집되어진다.

(3) 모든 가능한 표본평균들의 분포는 $\mu_{\bar{x}} = 5.5$를 중심으로 대칭형이며 정규분포와 유사한 형태를 띄운다.

□

〈표 5.3〉 표본평균의 도수분포(표본분포)

$\bar{x}$	도수	상대도수
1	1	0.01
1.5	2	0.02
2	3	0.03
2.5	4	0.04
3	5	0.05
3.5	6	0.06
4	7	0.07
4.5	8	0.08
5	9	0.09
5.5	10	0.10
6	9	0.09
6.5	8	0.08
7	7	0.07
7.5	6	0.06
8	5	0.05
8.5	4	0.04
9	3	0.03
9.5	2	0.02
10	1	0.01
	100	100/100=1



〈그림 5.1〉 모집단의 분포와 표본평균의 분포

만일 모집단이 매우 크면 위의 예와 같이 모든 가능한 표본들을 모두 찾아서 그 표본평균의 분포를 찾는 것은 불가능하다. 하지만 위에서 관찰한 사실들은 이 예제에만 국한되지 않고 모집단이 크거나 다른 분포형태를 가져도 관찰되고, 또 수학적으로 증명할 수 있는 사실들이다. 특히 모집단이 정규분포라면 표본평균의 분포는 역시 정규분포(모평균 주위에 분산이 적은)이다.

표본평균의 분포 (복원추출) --- 모집단이 정규분포인 경우

모집단이 모평균 μ 모분산 σ^2 인 정규분포를 따를 때
크기가 n 인 표본을 단순확률 복원추출하면
(1) 모든 가능한 표본평균들의 평균($\mu_{\bar{x}}$)은 모평균과 같다.($\mu_{\bar{x}} = \mu$)
(2) 모든 가능한 표본평균들의 분산($\sigma_{\bar{x}}^2$)은 모분산을 n으로 나눈 값이다.
($\sigma_{\bar{x}}^2 = \sigma^2/n$)
(3) 모든 가능한 표본평균들의 분포는 정규분포이다.
위의 사실을 간단히

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

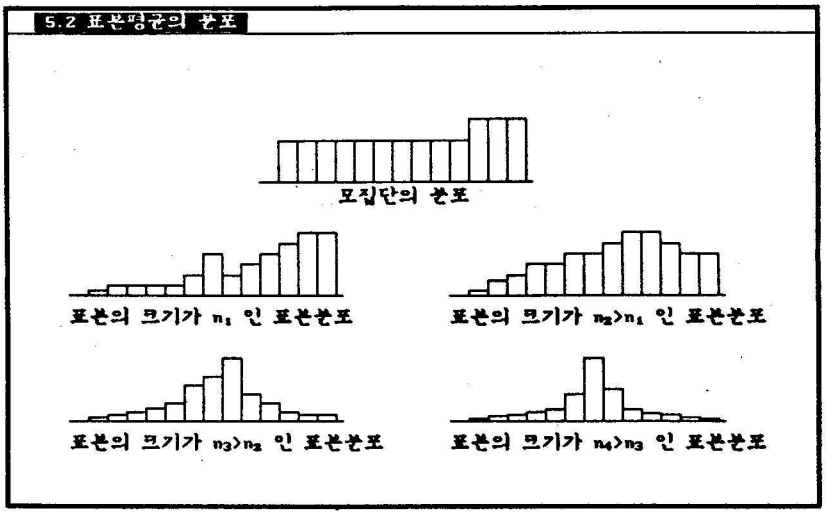
로 적기도 한다.

만일에 모집단이 무한모집단이고 표본의 크기(n)가 충분히 크면 모집단이 정규분포가 아니라도 표본평균의 분포는 근사적으로 정규분포임을 보일 수 있다. 이를 중심극한정리(Central Limit Theorem)라고 하는데 구체적으로 요약하면 다음과 같다.

[중심극한정리]
표본평균의 분포 (복원추출) --- 모집단이 정규분포가 아닌 경우

모집단이 평균 μ 분산 σ^2 이고 정규분포가 아닐 경우 무한모집단이면
표본평균의 분포는 n 이 충분히 클 때 근사적으로 정규분포 $N(\mu, \sigma^2/n)$를 따른다.

<그림 5.2>는 모집단이 평균에서 왼쪽으로 편중된 분포일 때 서로 다른 표본의 크기에 대한 표본평균의 분포들이 표본의 크기가 커짐에 따라 정규분포에 가까워짐을 보여 주고 있다.



<그림 5.2> 표본평균의 분포 ---모집단이 정규분포가 아닌 경우

비복원으로 표본추출을 하는 경우에 무한모집단이라면 위의 중심극한정리가 그대로 적용된다. 하지만 비복원추출시 유한모집단이라면 주의하여야 한다. 이 경우 모든 표본평균의 분산($\sigma_{\bar{x}}^2$)은

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\,\frac{N-n}{N-1}$$

가 된다. 위 식의 우측에 나타나는 항 (N-n)/(N-1)을 유한모집단 수정계수 (finite population correction factor)라고 하는데, N 이 n 보다 충분히 크면 이 계수는 1 에 가까워져 $\sigma_{\bar{x}}^2 = \sigma^2 / n$ 으로 근사된다. 대개 n/N 이 0.02 보다 작으면 수정계수는 무시한다.

표본평균의 분포 (비복원추출) --- 유한모집단이며 정규분포인 경우

모집단이 모평균 μ 모분산 σ^2 인 정규분포이며 유한모집단이면
표본평균의 분포는 다음과 같은 정규분포를 따른다.

$$\bar{x} \sim N\left(\mu,\ \frac{\sigma^2}{n}\,\frac{N-n}{N-1}\right)$$

비복원추출시 유한모집단이면서 정규분포가 아닌 경우에도 N 과 n 이 충분히 크다면, 근사적으로 위의 정규분포를 사용하여 만족할만한 결과를 얻을 수 있다.

앞에서 연구한 사실을 종합하면, 모집단의 크기가 표본의 크기보다 아주 크

고 또 표본의 크기도 충분히 크다면 (대개 30이상), 표본평균의 분포는 모집단의 분포나 추출방법에 관계없이 개략적으로 다음과 같이 적을 수 있다.

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

이러한 표본평균의 분포를 이용하는 예제를 살펴보자.

[예 5.3] 어느 회사에서 나오는 전구의 수명은 평균이 2000시간, 표준편차가 50시간이라 알려져 있다. 만일 100개의 전구를 단순확률 추출하였을 때 표본평균이 1990시간 이었다면 이 표본평균값이 모든 가능한 표본평균중에서 어디에 위치하는지 알아보고 싶다. 표본평균이 1990시간 미만이 될 확률은?

<풀이>

전구가 계속 생산된다고 할 때 무한모집단으로 가정할 수 있으므로 중심극한정리에 의해 표본평균은 평균이 $\mu_{\bar{x}} = 2000$, 분산이 $\sigma_{\bar{x}}^2 = 50^2/100 = 25$ 인 정규분포이다. 즉,

$$\bar{x} \sim N(2000, 25).$$

따라서 표본평균이 199시간 미만이 될 확률은

$$P(\bar{x} < 1990) = P(Z < (1990-2000)/5) = P(Z < -2) = 0.0228$$

이다. 즉 모든 가능한 표본평균 중 추출된 표본평균은 아래서 2.28% 지점에 있으므로 드문 경우에 속한다고 볼 수 있다. □

[예 5.4] 위의 예제에서 어느날 생산된 500개의 제품중 100개를 비복원추출하여 얻은 표본평균이 1990시간이었다고 하자. 표본평균이 1990시간 미만일 확률을 구하라.

<풀이>

N = 500 인 유한모집단에서 비복원추출하였으므로 유한모집단 수정계수를 고려하여야 한다. 즉 표본평균은 개략적으로 평균이 $\mu_{\bar{x}} = 2000$, 분산이

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{2500}{100}\frac{500-100}{500-1} = 20.04$$

인 정규분포이다. 즉, $\bar{x} \sim N(2000, 20.04)$. 따라서 표본평균이 199시간 미만이 될 확률은 $\bar{x}$의 표준편차가 $\sqrt{20.04} = 4.477$ 이므로

$$P(\bar{x} < 1990) = P(Z < (1990-2000)/4.477) = P(Z < -2.23) = 0.0129$$

이다. 즉, 모든 가능한 표본평균 중 추출된 표본평균은 아래서 1.29% 지점에 있으므로 더욱 드문 경우에 속한다고 볼 수 있다. □

3. 표본분산의 분포

모집단의 모분산과 표본에서 얻어지는 표본분산 사이의 관계를 알 수 있다면 역시 미지의 모분산을 추정하는데 많은 도움이 된다. 아래의 예를 가지고 표본분산의 분포(sampling distribution of sample variances)를 알아보자.

[예 5.5] 앞 절의 [예 5.2]에서 사용한 영업사원 10명의 근무년수 모집단을 다시 생각하자.

3, 6, 2, 4, 8, 7, 9, 5, 1, 10

여기서 표본의 크기(n)가 2인 모든 가능한 표본들을 단순확률 복원추출하여 그 표본분산들의 분포를 구하라. (역시 이렇게 작은 모집단은 굳이 표본을 추출할 필요가 없지만, 여기서는 표본분산의 분포를 설명하기 위한 예이다.)

〈풀이〉

표본의 크기 2로 복원추출한 모든 가능한 표본분산들을 적어보면 〈표 5.4〉와 같다.

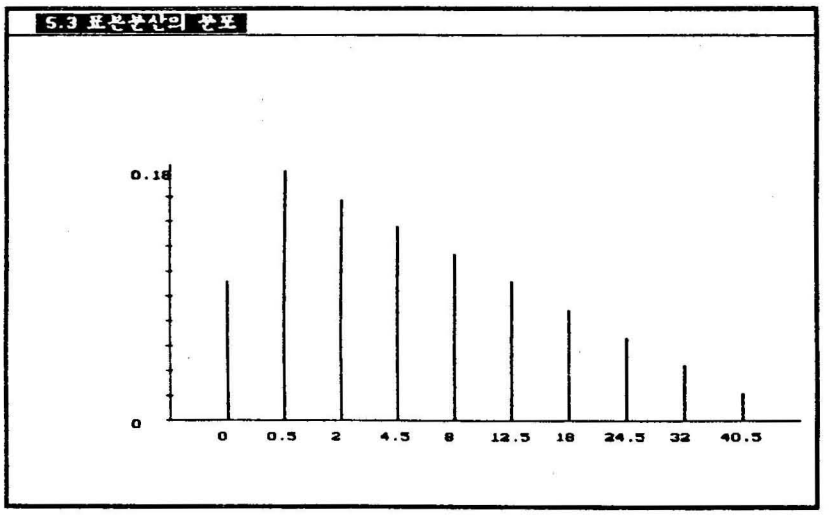
〈표 5.4〉 N=10인 모집단에서 추출가능한 n=2인 모든 표본들의 표본분산

표본	s^2	표본	s^2	표본	s^2	표본	s^2	표본	s^2
1,1	0.0	3,1	2.0	5,1	8.0	7,1	18.0	9,1	32.0
1,2	0.5	3,2	0.5	5,2	4.5	7,2	12.5	9,2	24.5
1,3	2.0	3,3	0.0	5,3	2.0	7,3	8.0	9,3	18.0
1,4	4.5	3,4	0.5	5,4	0.5	7,4	4.5	9,4	12.5
1,5	8.0	3,5	2.0	5,5	0.0	7,5	2.0	9,5	8.0
1,6	12.5	3,6	4.5	5,6	0.5	7,6	0.5	9,6	4.5
1,7	18.0	3,7	8.0	5,7	2.0	7,7	0.0	9,7	2.0
1,8	24.5	3,8	12.5	5,8	4.5	7,8	0.5	9,8	0.5
1,9	32.0	3,9	18.0	5,9	8.0	7,9	2.0	9,9	0.0
1,10	40.5	3,10	24.5	5,10	12.5	7,10	4.5	9,10	0.5
2,1	0.5	4,1	4.5	6,1	12.5	8,1	24.5	10,1	40.5
2,2	0.0	4,2	2.0	6,2	8.0	8,2	18.0	10,2	32.0
2,3	0.5	4,3	0.5	6,3	4.5	8,3	12.5	10,3	24.5
2,4	2.0	4,4	0.0	6,4	2.0	8,4	8.0	10,4	18.0
2,5	4.5	4,5	0.5	6,5	0.5	8,5	4.5	10,5	12.5
2,6	8.0	4,6	2.0	6,6	0.0	8,6	2.0	10,6	8.0
2,7	12.5	4,7	4.5	6,7	0.5	8,7	0.5	10,7	4.5
2,8	18.0	4,8	8.0	6,8	2.0	8,8	0.0	10,8	2.0
2,9	24.5	4,9	12.5	6,9	4.5	8,9	0.5	10,9	0.5
2,10	32.0	4,10	18.0	6,10	8.0	8,10	2.0	10,10	0.0

〈표 5.5〉는 표본분산들의 도수분포표이고 〈그림 5.3〉은 그 막대그래프이다.

〈표 5.5〉 표본분산의 도수분포표

s^2	도수	상대도수
0	10	0.10
0.5	18	0.18
2	16	0.16
4.5	14	0.14
8	12	0.12
12.5	10	0.10
18	8	0.08
24.5	6	0.06
32	4	0.04
40.5	2	0.02
계	100	1.0



〈그림 5.3〉 표본분산의 분포

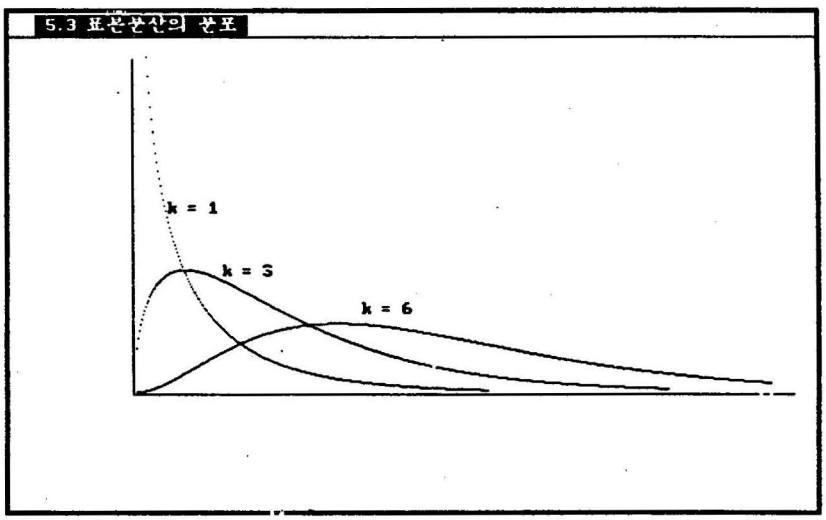
이 그림에서 관찰할 수 있는 사실은

(1) 작은 표본분산들이 많고 큰 표본분산이 적은 비대칭분포이다.

(2) 모든 표본분산들의 평균(μ_S) 은 모분산(σ^2=8.25)과 같다. 즉,

$$\mu_S = \frac{0.0 + 0.5 + \ldots + 0.0}{100} = 8.25$$ 이다. □

일반적으로 표본분산의 분포는 모집단이 정규분포이고 모분산이 σ^2일 때, 표본분산의 상수곱이 카이제곱분포(chi-sqare; χ^2 distribution)를 따른다. 이 카이제곱분포는 자유도(degree of freedom)라는 정수에 따라, 자유도 1인 카이제곱분포 (χ^2_1로 표시), 자유도 2인 카이제곱분포 (χ^2_2로 표시), ... , 자유도가 27인 카이제곱분포(χ^2_{27}로 표시), ... 등으로 나누어 지는 분포군이다. 이러한 카이제곱분포군은 비대칭분포인데 <그림 5.4>는 여러 가지 자유도에 대한 분포함수 그림이다. 각 자유도에 따른 카이제곱분포의 몇 가지 백분위수에 대한 값이 부록에 수록되어 있다. CATS의 부록에서는 카이제곱분포의 모든 누적확률과 백분위수의 값을 즉시 컴퓨터가 알려 준다.



<그림 5.4> 여러 자유도에 대한 카이제곱분포

구체적으로 표본분산의 분포를 요약하면 다음과 같다.

표본분산의 분포

모집단이 모분산 σ^2 인 정규분포를 따를 때 크기가 n인 표본을 단순확률복원추출하면, 표본분산 s^2 의 특정한 상수곱 $(n-1)s^2/\sigma^2$ 은 자유도가 (n-1)인 카이제곱분포를 따른다. 즉,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

이 표본분산의 분포를 이용하는 예제를 살펴보자.

[예 5.6] 한 과자의 포장에 표시되어 있는 무게는 200g 이다. 이 과자의 무게가 모분산 σ^2 = 100인 정규분포를 따를 때 크기가 26인 표본을 단순확률복원추출하였더니 표본분산 s^2 이 90 이었다. 이 표본분산이 모든 가능한 표본분산중에서 어디에 위치하는지 알아보고 싶다.

1) 표본분산이 90미만이 될 확률은?

2) 표본분산이 90에서 110사이일 확률은?

<풀이>

모든 가능한 표본분산의 상수곱 $(n-1)s^2/\sigma^2$ 은 자유도가 (n-1)인 카이제곱분포를 따르므로, 이 문제에서는 $(26-1)s^2/100$, 즉 $s^2/4$ 은 자유도가 (26-1)인 카이제곱분포를 따른다. 따라서, 1) 표본평균이 90시간 미만이 될 확률은

$$P(s^2 < 90) = P((s^2/4) < (90/4)) = P(\chi^2_{25} < 22.5) = 0.3933$$

이고 2) 표본평균이 90시간에서 110시간 사이일 확률은

$$P(90 < s^2 < 110) = P((90/4) < (s^2/4) < (110/4)) = P(22.5 < \chi^2_{25} < 27.5) = P(\chi^2_{25} < 27.5) - P(\chi^2_{25} < 22.5) = 0.6686 - 0.3933 = 0.2753$$

이다. 여기서 확률 $P(\chi^2_{25} < 27.5)$ 와 $P(\chi^2_{25} < 22.5)$는 CATS의 부록을 이용하여 구할 수 있다. □

4. 표본비율의 분포

모집단의 모비율과 표본에서 얻어지는 표본비율 사이의 관계를 알 수 있다면 역시 미지의 모비율을 추정하는데 많은 도움이 된다. 아래의 예를 가지고 모든 가능한 표본비율의 분포(sampling distribution of sample proportions)를 알아보자.

[예 5.7] 어느 회사의 사원 10명을 모집단이라 하자. 사원들의 회사에 대한 만족도를 조사하여 만족을 1로 불만을 0으로 표시하였을 때 다음과 같다.

1 0 1 1 0 1 1 0 0 1

즉 모집단에서 만족하는 비율(p)은 p=0.6이다. 여기서 크기가 5인 모든 표본을 복원추출하여 그 표본비율의 분포를 구하라. (역시 이렇게 작은 모집단은 굳이 표본을 추출할 필요가 없지만, 여기서는 표본비율의 분포를 설명하기 위한 예이다.)

〈풀이〉

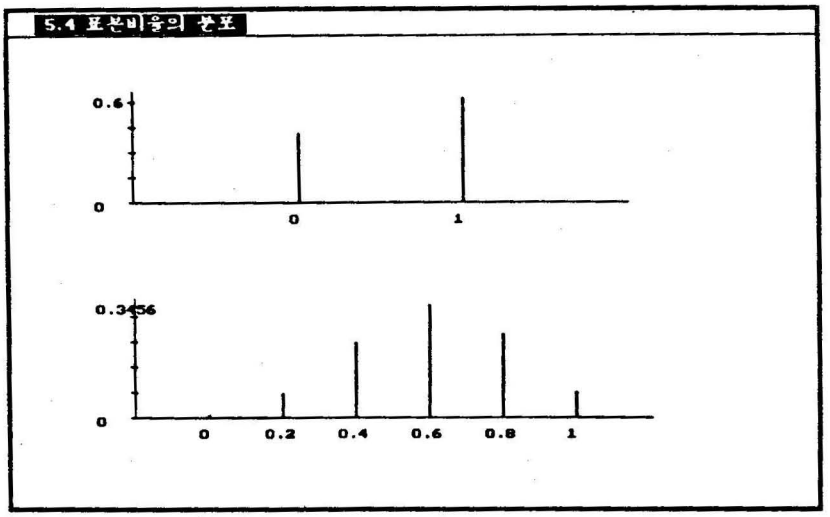
크기가 5인 모든 가능한 복원추출 표본의 개수는 10×10×10×10×10 = 100000개나 된다. 이 중 서로 다른 표본의 종류와 각 경우의 수는 아래와 같다.

표본의 종류	경우의 수	
모두 불만족(0,0,0,0,0)	${}_5C_0 \times 4 \times 4 \times 4 \times 4 \times 4$ =	1024
한 사람 만족(0,0,0,0,1)	${}_5C_1 \times 4 \times 4 \times 4 \times 4 \times 6$ =	7680
두 사람 만족(0,0,0,1,1)	${}_5C_2 \times 4 \times 4 \times 4 \times 6 \times 6$ =	23040
세 사람 만족(0,0,1,1,1)	${}_5C_3 \times 4 \times 4 \times 6 \times 6 \times 6$ =	34560
네 사람 만족(0,1,1,1,1)	${}_5C_4 \times 4 \times 6 \times 6 \times 6 \times 6$ =	25920
다섯사람 만족(1,1,1,1,1)	${}_5C_5 \times 6 \times 6 \times 6 \times 6 \times 6$ =	7776
	계	100000

따라서 각각의 표본에서 표본비율($\hat{p}$)을 구해 도수분포표를 만들면 〈표 5.6〉과 같다. 세 사람이 만족하는 경우(표본비율 0.6)일 경우가 제일 많음을 알 수 있다. 〈그림 5.5〉는 모집단의 분포와 표본비율의 분포를 비교한 것이다. 이러한 표본비율의 분포는 표본의 크기가 커지면 표본평균의 분포와 유사하게 모비율 p=0.6을 중심으로 대칭이며, 정규분포와 유사한 형태를 보이는 것을 알 수 있다. □

〈표 5.6〉 표본비율의 도수분포표

표 본	$\hat{p}$	도수	상대도수
모두 불만족	0.0	1024	0.01024
한 사람 만족	0.2	7680	0.07680
두 사람 만족	0.4	23040	0.23040
세 사람 만족	0.6	34560	0.34560
네 사람 만족	0.8	25920	0.25920
다섯사람 만족	1.0	7776	0.07776
	계	100000	1.0



<그림 5.5> 모집단의 분포와 표본비율의 분포

일반적으로 표본의 크기가 큰 경우 표본비율의 분포는 다음과 같다.

표본비율의 분포 --- 무한모집단의 경우

모집단의 모비율을 p라 하자. 일반적으로 표본의 크기가 충분히 클 때 표본비율의 분포는 근사적으로 평균이 p, 분산이 p(1-p)/n인 정규분포다.

$$\hat{p} \sim N(p, \frac{p(1-p)}{n})$$

유한개의 모집단에서 비복원추출할 경우에는 분산에 역시 유한모집단 수정계수 (N-n)/(N-1)을 곱하여야 한다.

[예 5.8] 한 반도체 공장에서 만들어내는 반도체의 3%가 불량품이라고 하자. 300개의 표본을 비복원추출하였을 때 불량품의 표본비율이 2%이었다. 이 표본비율이 모든 가능한 표본비율 중에서 어디에 위치하는지 알아보고 싶다. 표본비율이 2% 이상일 확률은?

<풀이>

표본비율의 분포는 근사적으로 $\hat{p} \sim N(0.03, 0.03(1-0.03)/300)$ 이므로

$P(\hat{p} > 0.02) = P(Z > (0.02-0.03)/0.00985)$

$$= P(Z > -1.02)$$
$$= 1 - P(Z \leq -1.02)$$
$$= 1 - 0.1539 = 0.8461$$

이다. □

연 습 문 제

5.1.1 어느 공장에 70명의 근로자가 있다. 주당 제품 생산량을 조사하기 위하여 10명의 표본을 단순확률 비복원추출하려고 한다. 부록의 난수표를 사용하여 표본추출을 하라.

5.1.2 한 대학의 신입생 1900명중 10%인 190명을 단순확률 비복원 추출하려고 한다. CATS를 사용하여 표본추출을 하라.

5.1.3 다섯 사람 A,B,C,D,E 중에서 세 사람을 복원추출하는 모든 경우를 나열하라.

5.2.1 어느 선의 장력이 평균이 99.8이고 표준편차가 5.48인 정규분포를 따른다.

(1) 표본의 크기가 100일 때 표본평균의 분포를 구하여라.

(2) 이 모집단으로부터 16개의 값을 단순확률 복원추출했을 때 이 표본의 평균이 98.8과 100.9사이에 있을 확률을 구하여라.

5.2.2 직업소개소에서는 지원자에 대해 적성검사를 행하는데 요구되는 평균시간이 24.5분이고 표준편차는 4.5분이라고 한다.

(1) 이 모집단으로부터 81개의 표본을 뽑았을 때 이 표본평균의 평균과 표준편차를 구하여라.

(2) 81명의 지원자 화일을 택해서 이 지원자들의 적성검사 평균시간이 25분보다 클 확률을 구하여라.

5.2.3 어느 회사에 1500명의 사원이 있다. 각 사원당 자선냄비에 기부하는 평균 기부금의 액수는 2,575원이고 표준편차는 525원이다. 100명의 사원을 단순확률 비복원추출하였을 때 표본평균이 2,500원과 2,700원사이에 있을 확률을 구하여라.

5.2.4 1200명의 간부가 있는 어느 집단에서 하루에 점심식사에 쓰는 평균 돈의

액수가 6,500원이고 표준편차는 6,000원이다. 이 집단으로부터 36명의 간부를 단순확률 추출했을 때 평균 돈의 액수가 5,000원과 10,000원 사이에 있을 확률을 구하여라.

5.2.5 모집단이 2,4,6,8,10 의 다섯 개 숫자로 구성되어 있다고 하자. 비복원추출로 표본의 크기가 2인 표본을 뽑아서 x의 표본분포를 작성하여라. 모집단과 표본평균의 평균과 분산을 구하라.

5.2.6 어느 회사 기능공들의 평균 고용기간이 2.5년이고 표준편차는 3년이라고 한다. 40명을 단순 확률추출로 뽑았을 때 평균 고용기간이 3.5년 이상일 확률을 구하여라.

5.3.1 한 알의 감기약에 들어있는 성분 A의 함량은 20mg 이어야 하는데 분산이 1 이내이면 정확히 그 성분을 포함하고 있는 것으로 간주한다. 과연 성분 A의 함량이 정확한가 조사하기 위하여 생산라인에서 41개의 감기약을 단순확률 복원추출해 성분 A를 조사해보니 표본분산이 1.09^2 이었다. 모집단이 분산이 1인 정규분포를 따른다고 가정할 때 표본분산이 1.09^2 미만이 될 확률은?

5.3.2 1년동안 기른 숭어의 길이가 표준편차가 4.35cm인 정규분포를 한다고 한다. 숭어 25마리를 잡아 표준편차를 계산하였을 때

1) 표본분산이 20미만이 될 확률은?
2) 표본분산이 15에서 22사이일 확률은?

5.3.3 한 열 전도체는 열을 가해 온도를 재는 실험을 하였을 때 온도의 분산이 25라고 한다. 이 전도체에 실험을 여섯 번 하였을 때

1) 표본분산이 30이상이 될 확률은?
2) 표본분산이 15에서 25사이일 확률은?

5.3.4 모집단이 2,4,6,8,10 의 다섯 개 숫자로 구성되어 있다고 하자. 비복원추출로 표본의 크기가 2인 표본을 뽑아서 표본분산 s^2의 표본분포를 작성하여라.

5.4.1 어느 금융회사는 고객의 60%가 그들의 계정잔액의 확인을 위한 조사서에 답해준다는 것을 알았다. 24명의 고객을 단순확률 추출해서 계정잔액에 대한 확인을 위해 조사서를 보냈을 때 50%이상이 응답해 줄 확률은?

5.4.2 어느 타자수에 의해 작성된 문서의 5%가 적어도 하나의 오자를 포함하고 있다는 것을 알았다. 475개의 문서를 조사했을 때 적어도 하나의 오자를 포함하고 있을 문서가 3%에서 7.5%사이에 있을 확률을 구하여라.

5.4.3 시청자의 25%가 TV 프로그램에서 너무 많은 폭력적 요소를 포함하고 있다고 생각한다는 것을 알았다. 이 집단으로부터 200명을 뽑았을 때 이 의견에 동의하는 비율이 0.24와 0.28사이에 있을 확률을 구하여라.

5.4.4 어느 광고 회사에서는 어느 성인 집단의 20%만이 이 광고 회사에서 만든 선전문구를 귀기울여 듣지 않는다고 주장한다. 이 집단으로부터 100명의 성인을 추출했을 때 그 중에서 24명이 귀기울여 듣지 않는다고 답했다. 만약 이 회사의 주장이 사실이라면 이 표본에서 얻어진 사람의 수보다 같거나 많은 사람의 수를 포함하고 있을 확률은 얼마이겠는가?

5.4.5 여론 조사기관에서 4000명의 주부로부터 250명을 선택하여 조사해본 결과 143명이 1대 이상의 TV수상기를 가지고 있다고 답했다. 모비율이 0.55이라면 이 조사에서 얻은 비율이상의 표본비율을 얻을 확률은 얼마이겠는가?

5.4.6 어느 회사의 사원 10명의 회사에 대한 만족도를 조사하여 만족을 1 로 불만을 0 으로 표시하였을 때 다음과 같다.

1 0 1 1 0 0 1 0 0 0

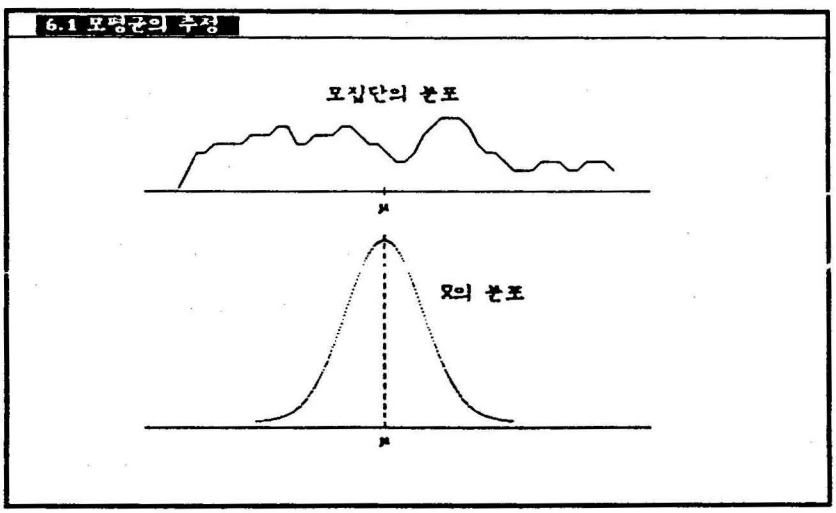
즉 모집단에서 만족하는 비율(p)은 p=0.4이다. 여기서 크기가 5인 모든 표본을 복원추출하여 그 표본비율의 분포를 구하라.

Ⅱ. 모수의 추정

1. 좋은 추정량이란?

어느 경제연구소에서 대졸 신입사원의 평균 초임이 얼마인지 알아보려고 한다. 이를 위해 금년도 대졸자 전체를 조사하기에는 너무 비용과 시간이 많이 들어 표본을 추출하였다. 상식적으로 전체 대졸자 초임의 평균에 대한 추정값으로서는 표본의 평균이 먼저 머리에 떠오른다. 과연 이 표본평균이 모평균을 잘 예측하여 줄 수 있을까?

위 질문에 대한 답은 '맞다'이다. 왜냐하면 5장에서 살펴보았듯이 모든 가능한 표본평균의 분포가 표본의 크기가 충분히 크다면 모평균주위에 분산이 아주 적은 정규분포를 이루기 때문이다(<그림 6.1>). 따라서 표본을 하나 추출하여 구한 표본평균은 대개 모평균에 아주 가까운 값이 된다.



<그림 6.1> 모집단의 분포와 표본평균들의 분포

표본평균과 같이 모수를 추정하는데 사용되는 표본의 통계량을 추정량(estimator)이라 하고, 그 추정량의 관측된 값을 추정값(estimate)이라 한다. 한 모수의 추정량은 여러 개가 있을 수 있다. 예를 들어 표본평균이나 표본의 중앙값은 모두 모평균의 추정량이다. 한 모수의 추정량이 여러 개인 경우 '어느 추정량이 더 좋은가?' 또 '좋다는 기준은 무엇인가?'라는 의문을 갖게 된다. 좋은 추정량이 되는 조건에는 여러 가지가 있는데 이 책에서는 '불편', '상대적 효율', '일치'의 세 가지 조건을 소개한다.

모든 가능한 표본평균($\bar{x}$)들의 기대값이 모평균(μ)이 되기 때문에 $\bar{x}$ 를 μ 의 불편추정량(unbiased estimator)이라고 한다. 불편추정량이란 표본통계량의 관측된 값들이 추정하려는 모수의 어느 한 쪽에 편중되지 않는다는 뜻으로 일반적으로 정의하면 다음과 같다.

불편추정량

미지 모수 θ 의 추정량 $\hat{\theta}$ 가 $E(\hat{\theta}) = \theta$ 를 만족할 때 $\hat{\theta}$ 를 θ 의 불편추정량이라 한다.

[예 6.1] 지금까지 배운 표본통계량 중에서 어느 것이 불편추정량이고, 어느 것이 아닌지 살펴보아라.

<풀이>

아래의 사실은 수학적으로 모두 증명이 가능한데 증명은 이 책의 수준을 넘으므로 설명만 한다.

1) 위에서 보았듯이 표본평균($\bar{x}$)은 모평균(μ)의 불편추정량이다. 즉 $E(\bar{x})=\mu$.
2) 일반적으로 표본의 중앙값(m)은 모평균(μ)의 불편추정량이 아니다. 그러나 모집단의 분포가 대칭인 경우는 표본의 중앙값이 모평균의 불편추정량이다.
3) 무한모집단의 경우 표본분산(s^2)은 모분산(σ^2)의 불편추정량이다. 즉 $E(s^2) = \sigma^2$. 표본분산 공식의 분모가 n 대신 n-1 인 이유중의 하나는 바로 표본분산을 모분산의 불편추정량으로 만들기 위함이다.
4) 표본표준편차(s)는 모표준편차(σ)의 불편추정량이 아니다. 그러나 표본의 크기가 충분히 크면 표본표준편차는 모표준편차에 매우 가까워진다.
5) 표본비율($\hat{p}$)은 모비율(p)의 불편추정량이다. 즉, $E(\hat{p}) = p$. 이 사실은 5장의 표본비율의 분포를 살펴보면 알 수 있다. □

만일 모수 θ 의 불편추정량이 두 개 이상 있을 때는, 두 추정량의 표본분포를 비교하여 모수 θ 에 더 밀집되어 있는 추정량을 선택하면 좋을 것이다. 즉, 추정량의 분산이 더 작은 것이 상대적으로 모수에 근접할 가능성이 높은 추정량이다. 이와 같은 추정량을 상대적인 **효율추정량**(relative efficient estimator)이라 한다.

상대적 효율추정량

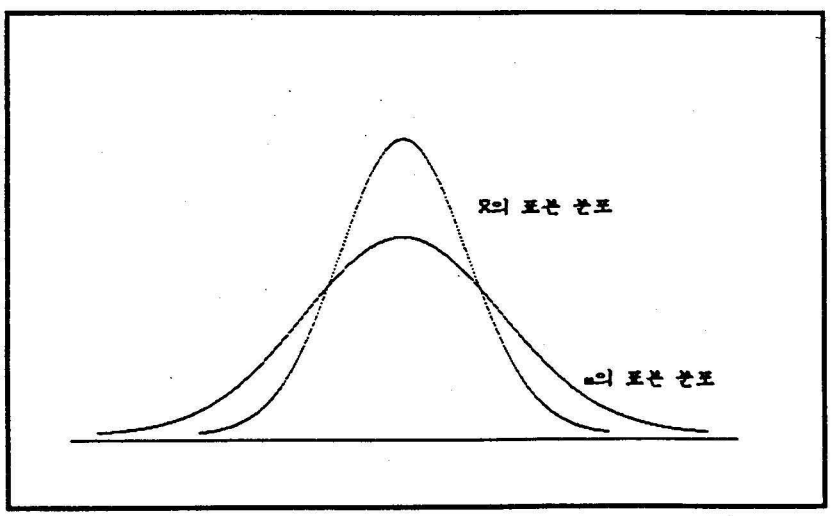
표본통계량 $\hat{\theta}_1$ 과 $\hat{\theta}_2$ 가 모두 모수 θ 의 불편추정량일 때 $\hat{\theta}_1$의 분산이 $\hat{\theta}_2$의 분산보다 작다면 $\hat{\theta}_1$ 을 $\hat{\theta}_2$ 보다 상대적으로 효율이 높은 추정량이라 한다. $\hat{\theta}_1$ 의 분산이 모수 θ 의 다른 어느 불편추정량의 분산보다 작을 때 $\hat{\theta}_1$ 을 효율추정량이라 한다.

위에서 보듯이 어떤 추정량의 분산 또는 표준편차는 그 추정량의 정밀도를 나타내는 측도가 될 수 있다. 그래서 이 추정량의 표준편차를 표준오차(standard error)라고 한다. 즉 $\bar{x}$ 의 표준오차는 $\sigma/\sqrt{n}$ 이고, $\hat{p}$의 표준오차는 $\sqrt{p(1-p)/n}$ 이다.

[예 6.2] 타이어의 수명이 평균 μ, 분산 σ^2 인 정규분포 모집단이라 가정하자. 이 모집단에서 크기가 n 인 표본을 추출할 때 모평균 μ 의 추정량으로 표본평균($\bar{x}$)과 표본중앙값(m)중 어느 것이 더 좋은가 살펴보아라.

<풀이>

모집단이 정규분포일 때는 $\bar{x}$ 나 m 모두 모평균 μ 의 불편추정량임을 수학적으로 보일 수 있다. 5장의 표본평균의 분포에서 $\bar{x}$ 의 분산은 σ^2/n 임을 알았다. 통계이론에서 표본중앙값 m 의 분산은 표본의 크기가 충분히 큰 경우 근사적으로 $1.57(\sigma^2/n)$ 임을 보일 수 있다. 따라서 모평균 μ 의 추정량으로는 표본평균($\bar{x}$)이 표본중앙값(m) 보다 상대적으로 효율이 높은 추정량이다. <그림 6.2>가 이를 잘 보여주고 있다. □



<그림 6.2> 상대적으로 효율이 높은 추정량

만일에 한 불편추정량의 분산이 다른 모든 불편추정량의 분산보다 작다면 이를 단순히 효율추정량이라 한다. 수학적으로 표본평균 $\bar{x}$ 는 모평균 μ 의 다른 모든 불편추정량보다 분산이 작거나 같음을 보일 수 있다. 즉 $\bar{x}$ 는 μ 의 효율추정량이다.

또 다른 좋은 추정량의 조건은 표본의 크기가 커질수록 추정량의 값과 모수

가 더 가까워져야 한다는 것이다. 이러한 성질을 갖는 추정량을 일치추정량(consistent estimator)이라고 한다. 아래의 정의에서 '확률적으로 같다'는 말을 좀 더 자세히 설명하자면 이 책의 수준을 넘는 확률적 수렴을 정의하여야 하므로 독자들은 단순한 수렴으로 일단 이해하기 바란다.

일치추정량

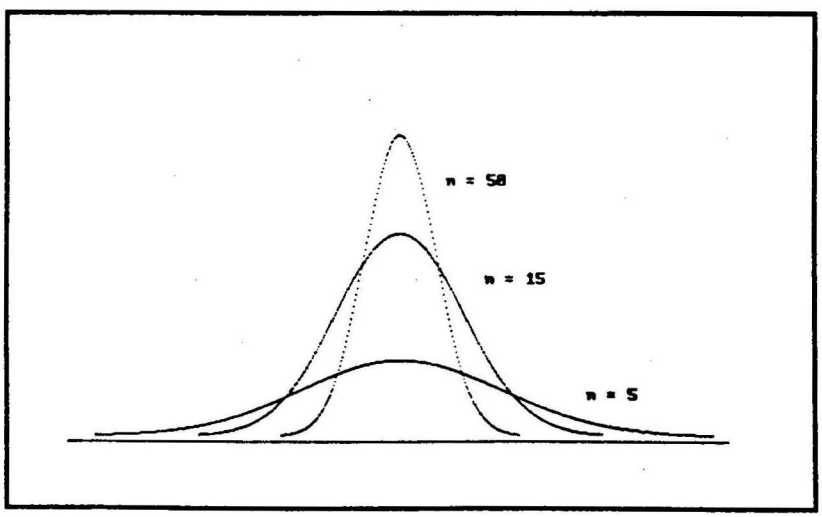
표본추정량 $\hat{\theta}$ 가 표본의 크기가 커짐에 따라 미지 모수 θ 와 확률적으로 같아질 때 $\hat{\theta}$ 을 θ 의 일치추정량이라 한다.

[예 6.3] 표본평균 $\bar{x}$ 가 모평균 μ의 일치추정량인가 살펴보아라.

<풀이>

5장에서 살펴본 표본평균의 분포에서 $\bar{x}$의 평균이 μ, 분산이 σ^2/n 임을 알았다. 따라서 표본의 크기 n 이 증가하면 분산 σ^2/n 이 점차로 적어져 0 에 가까워진다. 즉, 표본평균의 값이 모평균 μ 에 점점 가까워지게 된다(<그림 6.3>). 그러므로 표본평균 $\bar{x}$ 는 모평균 μ 의 일치추정량이다. □

이밖에 표본분산(s^2)은 모분산(σ^2)의 일치추정량이고, 표본비율($\hat{p}$)도 모비율(p)의 일치추정량이다



<그림 6.3> 일치추정량 $\bar{x}$

2. 모평균의 추정

미지인 모집단의 평균을 추정하기 위한 실용적인 예는 우리 주변에 수없이 많다. 몇 가지 예를 들면

1) 최근 창립한 어느 회사가 대졸 신입사원의 임금을 결정하기 위해 같은 업종 회사들의 대졸 신입사원 임금의 평균을 알아보려고 한다.
2) 한국전력회사에서 한 도시에 사는 전체 가정의 일일 평균 전력소비량을 조사하여 그 도시 전체 전력수요를 알아보려고 한다.
3) 어느 타이어 회사에서 최근 개발한 새 타이어의 평균수명을 알아보아 과거의 모형과 비교하려고 한다.

앞 절에서 알아보았듯이 표본평균은 모집단의 평균을 추정하기 위한 좋은 추정량의 성질을 모두 만족한다. 따라서 모평균 μ 의 추정에는 표본평균 $\bar{x}$ 를 사용하는 것이 좋다. 이 때 단지 관측된 표본평균의 하나의 값이 모평균의 추정값이라고 하는 것을 모평균의 점추정(point estimation, 하나의 점(수치)으로 추정한다는 뜻)이라 한다.

> 모평균(μ) 의 점추정 - 표본평균($\bar{x}$)
>
> $\bar{x}$ 는 μ 의 불편, 일치추정량, $\bar{x}$ 의 표준오차 = $\sigma / \sqrt{n}$

점추정 이외의 방법에 구간으로서 모평균을 추정하는 구간추정(interval estimation)이 있다. 만일 모집단이 평균이 μ 분산이 σ^2 인 정규분포라면, 표본평균 $\bar{x}$의 분포는 평균이 μ 분산이 σ^2/n인 정규분포이므로, 한 $\bar{x}$의 추정값이 구간 $\mu \pm 1.96\sigma/\sqrt{n}$ 에 포함될 확률은 95% 이다. 즉,

$$P(\mu - 1.96\sigma/\sqrt{n} \le \bar{x} \le \mu + 1.96\sigma/\sqrt{n}) = 0.95$$

이 식을 다시 정리하면

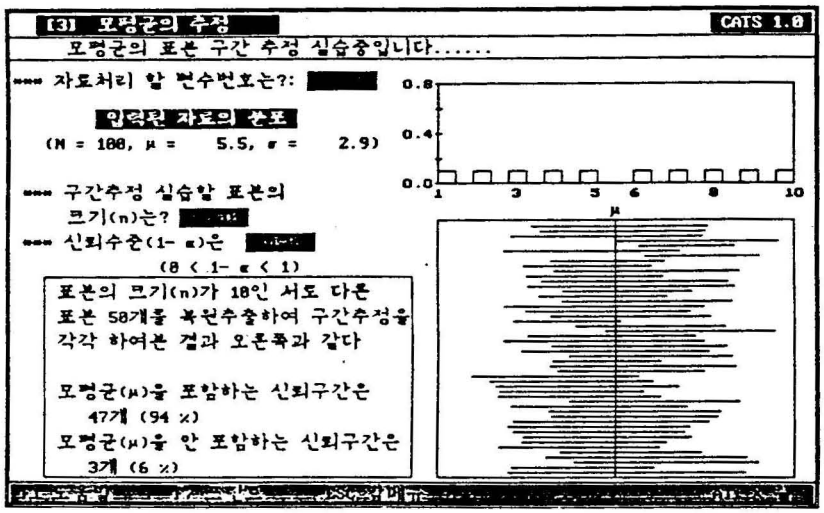
$$P(\bar{x} - 1.96\sigma/\sqrt{n} \le \mu \le \bar{x} + 1.96\sigma/\sqrt{n}) = 0.95$$

로 쓸 수 있다. 이 식의 의미는 모든 가능한 표본평균에 대해 구간공식 (σ 는 알려져 있다고 가정)

$$[\bar{x} - 1.96\sigma/\sqrt{n},\ \bar{x} + 1.96\sigma/\sqrt{n}]$$

을 적용하였을 때 얻어지는 모든 가능한 구간들 중 95%의 구간들이 모평균 μ 를

포함한다는 것이다. 이 구간공식을 모평균의 95% 신뢰구간이라 한다. <그림 6.4>는 모평균의 95% 구간추정을 설명한 그림이다. 단, 모든 가능한 표본을 모두 추출하는 것은 너무 경우의 수가 많기 때문에 50개의 표본을 복원추출하여 구간추정을 하였다.



<그림 6.4> 모평균의 95% 구간추정의 의미

일반적으로 모평균 μ 의 신뢰구간은, α 를 모평균이 포함 안될 확률이라고 할 때, 모집단이 정규분포라면 다음과 같다.

모평균 μ 의 $100(1-\alpha)\%$ 구간추정 --- 모집단이 정규분포 $N(\mu,\sigma^2)$인 경우

$$[\bar{x} - Z_{1-\alpha/2}\ (\sigma/\sqrt{n}),\ \bar{x} + Z_{1-\alpha/2}\ (\sigma/\sqrt{n})]$$

여기서 $1-\alpha$ 는 신뢰도(confidence level)라고도 하는데, 이 구간공식에 의해 산출된 모든 구간들 중에서 모평균이 포함되어 있을 구간들의 확률을 뜻한다. 대개 α는 0.01 또는 0.05를 사용한다. Z_α는 표준정규분포의 왼쪽꼬리에서부터의 누적확률이 α 가 되는 점, 즉 $100\alpha\%$ 백분율을 말한다. 예를 들면 $Z_{0.025} = -1.96$ $Z_{0.95} = 1.645$ 이다.

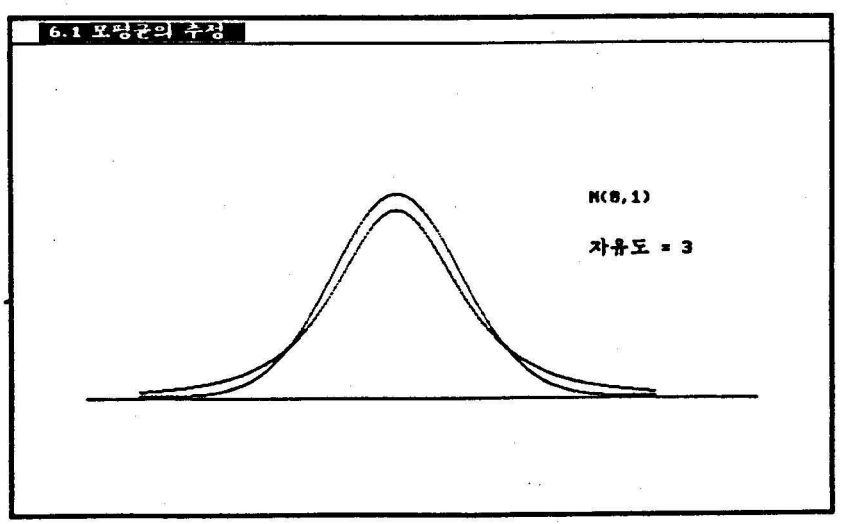
위의 구간공식으로 미지 모평균을 추정할 때의 문제점은 대개 모표준편차 σ 를 모른다는 것이다. 하지만 표본의 크기가 충분히 클 경우 $\bar{x}$ 는 중심극한정리

에 의해 $N(\mu, \sigma^2/n)$에 근사하므로 $(\bar{x} - \mu) / (\sigma/\sqrt{n})$ 는 $N(0,1)$에 근사한다. 이 경우 (표본의 크기가 약 30보다 큰 경우), σ 를 s로 대치하면 $(\bar{x} - \mu) / (s/\sqrt{n})$ 도 역시 $N(0,1)$에 근사함을 수학적으로 보일 수 있다. 따라서 표본의 크기가 충분히 클 때 모평균 μ 의 신뢰구간은 근사적으로 다음과 같이 구할 수 있다.

모평균 μ 의 $100(1-\alpha)\%$ 구간추정 --- σ 를 모르나 대표본인 경우

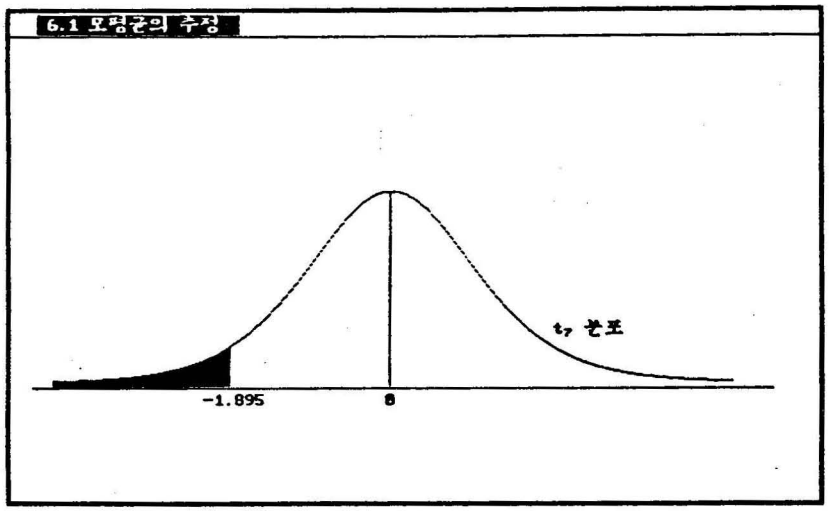
$$[\bar{x} - Z_{1-\alpha/2}\ (s/\sqrt{n}),\ \bar{x} + Z_{1-\alpha/2}\ (s/\sqrt{n})]$$

표본의 크기가 작다면 위의 근사 신뢰구간 공식을 사용할 수 없다. 그러나 표본의 크기가 작더라도 이를 이용하여 신뢰구간 공식을 유도할 수 있다. 모집단이 정규분포인 경우는 $(\bar{x} - \mu) / (s/\sqrt{n})$ 의 정확한 분포는 자유도가 n-1인 t분포이다. t 분포는 아일랜드의 한 양조회사에서 근무하던 통계학자 W.S.Gosset에 의해 연구되었는데, 회사의 규정상 이름을 밝힐 수 없어 스튜던트(Student)라는 가명으로 1907년에 연구결과를 발표하였다. 그래서 흔히 이 분포를 스튜던트 t분포(Student's t distribution)라 부른다. t분포는 단 하나의 분포가 아니라, 자유도라는 모수에 따라 t_1, t_2, ... , t_{30}, ... 등 무수히 많은 분포군을 뜻한다. t분포의 모양은 표준정규분포와 흡사하다. t분포는 0을 중심으로 좌우대칭이며 표준정규분포보다 두터운 꼬리를 갖는 특징을 갖고 있다. <그림 6.5>가 표준정규분포와 자유도가 3인 t분포를 동시에 보여 주고 있다.



<그림 6.5> t분포와 정규분포의 비교

또 t분포는 자유도가 증가할 수록 표준정규분포에 가까워지는데, 대개 자유도가 30이 넘으면 비슷하다. 이것이 대표본일 때 정규분포를 사용하여 신뢰구간 공식을 유도한 근거이다. 자유도가 n인 t분포의 왼쪽 꼬리에서부터의 누적확률이 α 가 되는 점을 $t_{n,\alpha}$ 로 표시하자. 예를 들어 $t_{7,0.05}$ 는 t_7 분포 <그림 6.6> 왼쪽 서부터의 누적확률이 0.05가 되는 점을 뜻하는데, 이 값을 컴퓨터를 이용하여 계산하면 -1.895이다. 표준정규분포에서는 이 값이 -1.645이었다. t분포는 대칭이므로 $t_{n,1-\alpha} = - t_{n,\alpha}$ 이다. 즉 $t_{7,0.95} = 1.895$ 가 된다.



<그림 6.6> t_7 분포의 5% 백분율 $t_{7,0.05}$ 의 의미

σ 를 s 로 대치하였을 때, 표본의 크기가 작고 모집단이 정규분포인 경우 t분포를 이용한 다음의 신뢰구간 공식을 사용하여야 한다.

모평균 μ 의 $100(1-\alpha)\%$ 구간추정
--- σ 를 모르며 소표본이고 모집단이 정규분포인 경우

$$[\bar{x} - t_{n-1,1-\alpha/2}\,(s/\sqrt{n}),\ \bar{x} + t_{n-1,1-\alpha/2}\,(s/\sqrt{n})]$$

만일 소표본이고 모집단이 정규분포가 아니라면 위의 공식을 사용해서는 안되고 다른 통계적방법을 고려하여야 한다.

[예 6.4] 금년도 대졸자의 초임을 알아보기 위하여 100명을 단순확률추출하여 조사하니 평균이 45만원, 표준편차가 5만원이었다.

1) 전체 대졸자 초임의 평균을 점추정하라.
2) 전체 대졸자 초임의 평균을 95%의 신뢰도로 구간추정하라.
3) 전체 대졸자 초임의 평균을 99%의 신뢰도로 구간추정하라. 이 구간의 너비가 95% 신뢰구간과 비교해 어떠한가?
4) 표본의 크기가 400명 이었을 때 전체 대졸자 초임의 95% 신뢰구간을 구하라. 문제 2)와 비교해 구간의 너비가 어떠한가?

<풀이>

1) 전체 대졸자 초임 평균의 점추정은 표본평균이므로 45만원이다.
2) 95% 신뢰도의 구간추정은 표본의 크기가 30보다 크므로 대표본 공식을 이용하여도 좋다. 여기서 95%의 신뢰구간은 $\alpha = 0.05$ 를 의미하므로 $Z_{1-\alpha/2} = Z_{1-0.05/2} = Z_{0.975} = 1.96$ 이다. 따라서 95% 신뢰구간은

$$[\bar{x} - Z_{1-\alpha/2}(s/\sqrt{n}),\ \bar{x} + Z_{1-\alpha/2}(s/\sqrt{n})]$$
$$[\ 45 - 1.96(5/10)\ ,\ 45 + 1.96(5/10)\ ]$$
$$[\ 44.02\ ,\ 45.98\ ]$$

3) 99%의 신뢰구간은 $\alpha = 0.01$ 를 의미하므로 $Z_{1-\alpha/2} = Z_{1-0.01/2} = Z_{0.995} = 2.575$ 이다. 따라서 99% 신뢰구간은

$$[\bar{x} - Z_{1-\alpha/2}(s/\sqrt{n}),\ \bar{x} + Z_{1-\alpha/2}(s/\sqrt{n})]$$
$$[\ 45 - 2.575(5/10)\ ,\ 45 + 2.575(5/10)\ ]$$
$$[\ 43.71\ ,\ 46.29\ ]$$

그러므로 신뢰도가 증가하면 구간의 너비가 넓어진다.

4) 표본의 크기가 400명이라면 95% 신뢰구간은

$$[\bar{x} - Z_{1-\alpha/2}(s/\sqrt{n}),\ \bar{x} + Z_{1-\alpha/2}(s/\sqrt{n})]$$
$$[\ 45 - 1.96(5/20)\ ,\ 45 + 1.96(5/20)\ ]$$

[44.51 , 45.49] 이다. 그러므로 표본의 크기가 증가하면 구간의 너비가 좁아진다. 즉 보다 정확한 추정이 된다. □

[예 6.5] 위의 예제에서 만일 표본의 크기가 25명이고 모집단이 정규분포일 경우 전체 대졸자 초임의 평균을 점추정, 또 95% 신뢰도로 구간추정하여라.

<풀이>

대졸자 초임 평균의 점추정은 역시 표본평균 45만원이다. 구간추정은 표본의 크기가 작으므로 소표본 공식을 사용하여야 한다. 따라서 95% 신뢰구간은 $t_{n-1.1-\alpha/2} = t_{25-1.1-0.05/2} = t_{24.0.975} = 2.0639$ 이므로

$$[\bar{x} - t_{n-1.1-\alpha/2}(s/\sqrt{n}),\ \bar{x} + t_{n-1.1-\alpha/2}(s/\sqrt{n})]$$
$$[\ 45 - 2.0639(5/5)\ ,\ 45 + 2.0639(5/5)\ ]$$
$$[\ 42.9361\ ,\ 47.0639\ ]$$

표본의 크기가 작아지면 구간의 너비가 넓어짐에 유의하라. □

3. 모분산의 추정

미지의 모집단의 분산을 추정하기 위한 몇 가지 실용적인 예를 들어보자.

1) 두 볼트회사에서 한 자동차회사에 볼트를 납품한다. 볼트는 직경이 너무 크거나 작아도 불량품이다. 자동차회사에서 각 볼트 회사에서 납품하는 볼트 직경의 분산을 알아보아 의사결정 자료로 하고자 한다.
2) 금년도 대입학력고사의 난이도를 평가하기위해 학력고사 성적의 분산을 알아보고 싶다.

6.1절에서 살펴 보았듯이 무한모집단의 경우 표본분산(s^2)은 모분산(σ^2)의 불편추정량이다. 따라서 모분산(σ^2)의 점추정에는 표본분산(s^2)이 이용된다. 그리고 모표준편차(σ)의 추정에는 표본표준편차(s)가 이용되는데 주의할 것은 표본표준편차는 모표준편차의 불편추정량이 아니라는 것이다. 그러나, 표본의 크기가 커지면 s를 σ 의 추정량으로 사용해도 큰 오차가 없다.

모분산(σ^2)의 점추정 - 표본분산(s^2)
s^2 은 σ^2 의 불편추정량
모표준편차(σ)의 점추정 - 표본표준편차(s)

5장에서 우리는 모집단이 정규분포일 때 $(n-1)s^2/\sigma^2$ 이 자유도가 $(n-1)$인 χ^2 분포를 따르는 것을 알았다. 이를 이용하여 모분산(σ^2)과 모표준편차(s)의 구간추정은 다음과 같이 할 수 있다.

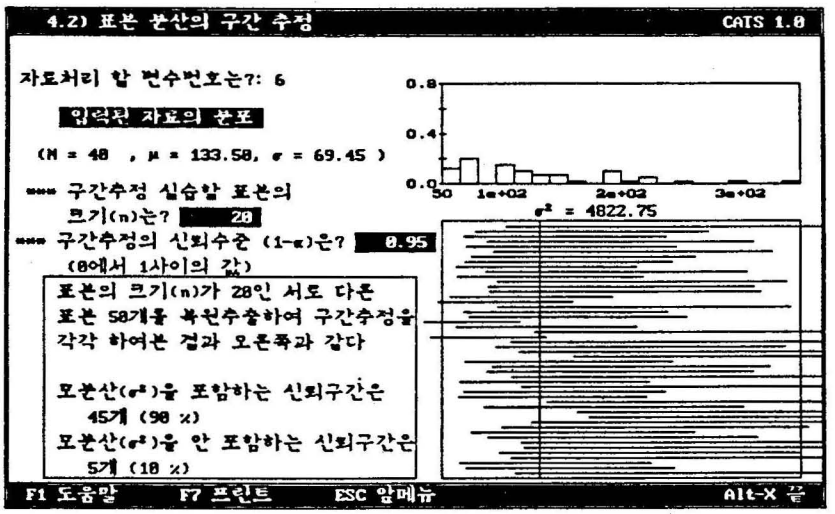
모분산(σ^2)의 $100(1-\alpha)\%$ 신뢰구간 - 모집단이 정규분포를 따르는 경우

$$\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \right]$$

모표준편차(σ)의 $100(1-\alpha)\%$ 신뢰구간 - 모집단이 정규분포를 따르는 경우

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}} \right]$$

<그림 6.7>은 모분산의 95% 신뢰구간을 실습한 것이다. 단, 모든 가능한 표본을 모두 추출하는 것은 너무 경우의 수가 많기 때문에 50개의 표본을 복원추출하여 구간추정을 하였다.



<그림 6.7> 모분산의 95% 구간추정의 의미

[예 6.6] 금년도 대졸자 초임을 조사하기 위하여 25명을 단순확률 복원추출하여 조사하였더니 표본표준편차가 5만원이다. 모분산, 모표준편차의 점추정과 95% 신뢰도로 구간추정을 하라. 단, 모집단이 정규분포를 한다고 가정하자.

<풀이>

대졸자 초임의 모분산의 점추정은 표본분산이므로 $s^2 = 5^2 = 25$이다. 그리고 모표준편차의 점추정은 표본표준편차이므로 $s = 5$ 이다. 모분산의 95% 신뢰구간은

$$\left[\frac{(n-1)s^2}{\chi^2_{25-1.1-0.05/2}} , \frac{(n-1)s^2}{\chi^2_{25-1.0.05/2}} \right]$$

$$\left[\frac{(25-1)5^2}{39.364} , \frac{(25-1)5^2}{12.401} \right]$$

$$[\ 15.242\ ,\ 48.383\ ]$$

이다. 그리고 모표준편차의 95% 신뢰구간은 $[\ \sqrt{15.242}\ ,\ \sqrt{48.383}\ ]$ 즉, [3.904 , 6.956] 이 된다. □

4. 모비율의 추정

미지 모집단의 비율을 추정하기 위한 몇 가지 실용적인 예를 들어보자.

1) 금년도 선거에서 어느 정당의 지지율은 몇 %나 될까?
2) 현재 우리나라의 실업률은 몇 %나 될까?
3) 자동차 부속품 1만개를 수입하는데 과연 여기에 불량품이 몇 %나 될까?

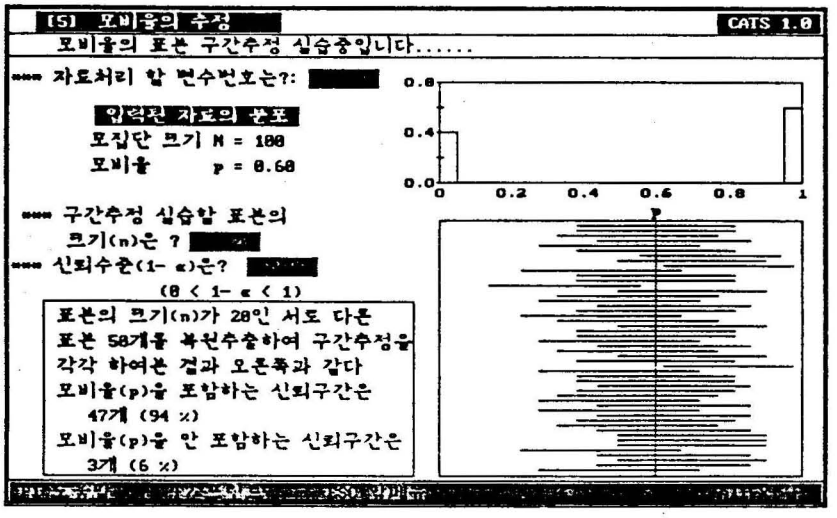
이렇게 모집단의 한 특성에 대한 비율을 추정하려는 것이 모비율(p)의 추정이다. 6.1절에서 살펴 보았듯이 표본비율($\hat{p}$)은 모비율(p) 추정시 좋은 추정량의 조건을 모두 만족하므로 모비율의 점추정에는 표본비율이 사용된다. 이때 표본비율의 표준오차는 5장의 표본비율의 분포에 의하면 $\sqrt{p(1-p)/n}$ 가 된다. 모비율 p는 미지수이므로 $\sqrt{\hat{p}(1-\hat{p})/n}$ 를 표준오차의 추정량으로 사용한다.

> 모비율(p)의 점추정 - 표본비율($\hat{p}$)
>
> $\hat{p}$ 는 p의 불편, 효율, 일치 추정량, $\hat{p}$ 의 표준오차 = $\sqrt{p(1-p)/n}$

표본의 크기가 충분히 크면 $\hat{p}$ 의 분포는 정규분포에 근사하게 된다는 (5장의 표본비율분포 참조) 사실로부터 모비율 p의 구간추정은 다음과 같이 할 수 있다.

> 모비율(p)의 100(1-α)% 신뢰구간 --- 대표본인 경우
>
> $$\left[\hat{p} - Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$
>
> 모비율의 구간추정시 표본의 크기 n이 충분히 큰 기준은 $n\hat{p} > 5$, $n(1-\hat{p}) > 5$ 일 때이다.

<그림 6.8>은 모비율의 95% 신뢰구간을 실습한 것이다. 단, 모든 가능한 표본을 모두 추출하는 것은 너무 경우의 수가 많기 때문에 50개의 표본을 복원추출하여 구간추정을 하였다.



<그림 6.8> 모비율의 95% 구간추정의 의미

[예 6.7] 어느 대학의 총 학생회장 선거에 입후보한 학생이 본인의 지지율을 알아 보기 위하여 200명의 학생을 단순확률 추출하여 질문하였더니 120명이 지지를 하였다. 모집단의 지지율을 점추정하고, 95%의 신뢰도로 구간추정을 하라.

<풀이>

전체 학생의 지지율의 점추정은 표본비율이므로

$$\hat{p} = 120/200 = 0.6$$

이다. 전체 지지율의 95% 신뢰구간은

$$\left[\hat{p} - Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} , \hat{p} + Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

이므로

$$\left[0.6 - 1.96\sqrt{\frac{0.6(1-0.6)}{200}} , 0.6 + 1.96\sqrt{\frac{0.6(1-0.6)}{200}} \right]$$

$$[\ 0.5321\ ,\ 0.6679\ ]$$

이 된다. □

5. 표본크기의 결정

지금까지는 주어진 표본을 이용하여 모수를 추정하는 것을 다루었다. 그러나 이러한 표본을 얻기 전에 표본의 크기를 얼마로 할 것인가를 먼저 결정해야 할 때가 많이 있다. 이 문제는 추정의 정밀도와 밀접한 연관이 있다. 우리는 앞절에서, 일반적으로 표본의 크기가 클수록 모수를 구간추정할 때 구간의 너비가 좁아짐(즉, 정밀도가 높아짐)을 알 수 있었다. 하지만 표본을 많이 추출하기 위해서는 비용이 많이 들기 때문에 대개는 연구자가 만족할만한 정밀도를 설정한 후 이 정밀도를 달성하기 위한 표본의 크기를 결정한다.

6.2절에서 보았듯이 평균이 μ 분산이 σ^2 인 모집단에서 모평균 $100(1-\alpha)\%$ 신뢰구간은

$$[\bar{x} - Z_{1-\alpha/2}\,(\sigma/\sqrt{n}),\ \bar{x} + Z_{1-\alpha/2}\,(\sigma/\sqrt{n})]$$

이었다. 이 때 $Z_{1-\alpha/2}(\sigma/\sqrt{n})$ 를 모평균 μ 추정에서의 **오차의 한계**(maximum allowable error)라고 한다. 따라서 오차의 한계를 d 로 하기위한 표본의 크기는 방정식

$$Z_{1-\alpha/2}\,(\sigma/\sqrt{n}) = d$$

을 n에 관하여 풀면 된다.

모평균 추정시 표본의 크기 결정

$$n = \left(\frac{Z_{1-\alpha/2}\,\sigma}{d} \right)^2$$

이 식에서 모표준편차 σ 는 미지수이므로 과거의 경험자료를 이용하던가, 또는 예비조사를 해 추정하기도 한다.

[예 6.8] 어느 공장에서 생산되는 전구 수명의 표준편차가 대개 100시간 이라고 한다. 전구의 평균수명을 95% 신뢰도로 추정하려고 하는데 오차의 한계가 20시간 이내가 되기 위한 표본의 크기를 구하라.

<풀이>

$$n = \left(\frac{Z_{1-\alpha/2}\,\sigma}{d} \right)^2 = \left(\frac{1.96 \times 100}{20} \right)^2 = 9.8^2 = 96.04$$

그러므로, 필요한 표본의 크기는 97개이다. □

비슷한 방법으로 모비율 p의 100(1-α)% 구간추정은

$$\left[\hat{p} - Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

이므로, 오차의 한계가 d가 되기 위해서는 방정식

$$Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = d$$

을 n에 관하여 풀면 된다.

모비율 추정시 표본의 크기 결정

$$n = \hat{p}(1-\hat{p}) \left[\frac{Z_{1-\alpha/2}}{d} \right]^2$$

이 때에 $\hat{p}$ 는 과거의 경험에 의해 추정된 값을 이용하던가, 예비조사를 하여 추정하기도 한다. 하지만 모비율에 대해 전혀 지식이 없을 때는 n의 값이 최대가 되는 $\hat{p} = 0.5$를 사용하기도 한다.

[예 6.9] 금년도 대통령 선거에 어느 후보의 지지율을 95% 신뢰도로 조사하려고 한다. 오차의 한계가 2.5% 이내가 되기 위한 표본의 크기를 구하라.

<풀이>

모비율에 대한 지식이 없으므로 0.5 로 가정하면

$$n = \hat{p}(1-\hat{p}) \left[\frac{Z_{1-\alpha/2}}{d} \right]^2$$

이므로

$$n = 0.5(1-0.5)\ 1.96^2 / 0.025^2 = 1536.6$$

이다. 우리나라에서 실시되는 각종 여론조사의 표본의 크기가 종종 1500명 정도인 것을 보는데 그 이유는 바로 단순확률 표본추출시 오차의 한계를 대략 2.5%로 하기 위한 조사이기 때문이다. □

연 습 문 제

6.2.1 어떤 냉동식품회사가 대량으로 사들인 옥수수의 평균 길이를 알고자 한다. 옥수수 200개를 단순확률 추출하여 측정을 하여보니 길이의 평균이 20.8cm가 된다는 것을 알 수 있었다. 모집단의 표준편차는 2cm이다. 모평균의 95% 신뢰구간을 구하여라.

6.2.2 한 전화 응답 서비스는 각 전화가 끝날 때마다 통화시간이 기록된 보고서를 작성한다. 단순확률 추출로 보고서 9개를 뽑아 평균 통화시간이 1.2분임을 알았고 모집단은 표준편차가 0.6분인 정규분포를 따른다고 알려졌다. 모평균의 99% 신뢰구간을 구하여라.

6.2.3 어떤 큰 제조회사의 품질 관리자는 5500개의 가공하지 않은 재료들의 평균무게를 알고자 한다. 250개를 단순확률 추출하여 측정한 결과 평균은 65kg이다. 모집단의 표준편차는 15kg이다. 미지의 모평균에 대한 95% 신뢰구간을 구하라.

6.2.4 한 신체 건강연구팀은 17세에서 21세 사이의 남자를 대상으로 일정한 표준 운동이후의 평균 산소 소모량을 측정하고자 한다. 연구에 의해 모분산은 0.0512라고 할 수 있다. 25명을 단순확률 추출한 결과는 다음과 같다.

2.87 2.05 2.90 2.41 2.93 2.94 2.26 2.21 2.20 2.88
2.51 2.51 2.56 2.59 2.52 2.51 2.50 2.58 2.52 2.58
2.44 2.48 2.43 2.46 2.46 (리터 /분)

산소 소모량이 정규분포를 따를 때 모평균의 95% 신뢰구간을 구하라.

6.2.5 한 산업 심리학자가 특정 모집단의 여성근로자의 평균 나이를 알고자 한다. 모집단으로부터 60명의 표본을 단순확률 추출한 나이의 평균이 23.67이었다. 모집단의 나이가 어떠한 분포인지 모르나 표준편차가 15일 때 모평균의 99% 신뢰구간을 구하라.

6.2.6 어느 기계에 유연성있는 플라스틱 호스를 사용할 수 있는지 결정하기 위한 연구에서 기사는 호스가 받는 평균압력을 추정하고자 한다. 기사는 24시간 간격으로 9번의 압력을 측정하였다. 표본의 평균과 표준편차가 각각 362, 45이고 압력은 근사적으로 정규분포를 따른다. 평균압력에 대한 99% 신뢰구간을 구하라.

6.2.7 30초동안의 삽입광고에 드는 라디오 방송비용을 추정하기위해 16개의 방송국을 단순확률 추출하였더니 표본평균이 15.5만원, 표본분산이 8이었다. 모든

라디오 방송국의 광고비용이 정규분포를 따른다고 할 때 모평균에 대한 95% 신뢰구간을 구하라.

6.3.1 한 옷감을 만드는데 쓰이는 실의 장력을 검사하려고 한다. 이 실 한 뭉치의 10군데를 단순확률 표본추출하여 장력 검사를 했더니 분산은 4이다. 분산의 구간추정을 위해 필요한 가정은 무엇인가? σ^2의 95% 신뢰구간을 구하여라.

6.3.2 어느 생산 관리자는 제품공정에서의 특정작업을 끝내는데 요구되어지는 시간을 알려고 한다. 25개의 관측치의 확률표본이 분석을 위해 사용되었다. 표본자료들로부터 계산된 분산은 0.3^2 이다.

(1) σ^2의 95% 신뢰구간을 구하여라.
(2) σ^2의 99% 신뢰구간을 구하여라.
(3) σ^2의 90% 신뢰구간을 구하여라
(4) 타당한 신뢰구간을 구하기 위해 필요한 가정은 무엇인가?

6.3.3 어느 생태학자가 공장지역이 위치한 강으로부터 15개 표본의 물을 구해 그 물이 함유하고 있는 특정 오염물질의 양을 측정하려 한다. 만일 오염물질의 양이 정규분포를 따르고, $\Sigma(x_i-\bar{x})^2 = 508.06$ 일 때 모분산의 95% 신뢰구간을 구하라.

6.4.1 노동청은 작업의 단조로움으로 인해 직업을 바꾸는 사무직 근로자들의 비율을 알고 싶어한다. 최근 직업을 바꾼 사무직 근로자 400명을 단순확률 추출하여 질문하였다. 그 중 200명이 직업의 단조로움 때문에 직업을 바꾸었다고 답했다. 직업을 바꾼 이들의 모비율에 대한 95% 신뢰구간을 구하라.

6.4.2 어느 회사에서 지난 주 모든 고용인들에게 배포된 안전관리 인쇄물을 읽어 기억하고 있는 고용인들의 비를 추정하고 싶어 한다. 300명의 고용인을 단순확률 추출하여 인쇄물의 내용을 기억하고 있는 정도를 측정하는 시험을 실시했다. 이 시험에 응한 사람들중 75명이 시험에 합격했다. 시험에 합격한 사람들의 모비율에 대한 95% 신뢰구간을 구하라.

6.4.3 근로자 이직의 원인 연구에서 조사자는 어느 회사에서 종사하였던 고용인 200명을 뽑아 조사하였다. 200명중 140명이 그들의 감독관과 화합할 수 없어서 이직했다고 답했다. 이런 이유 때문에 이직한 사람들에 대한 모비율의 95% 신뢰구간을 구하여라.

6.4.4 175명의 성인을 대상으로 '현재 우리나라에서 가장 심각하게 생각되어지는 사회문제가 무엇이냐'는 질문에 79명이 약물과 알콜의 남용이라고 답했다. 이런 의견을 낸 사람들의 전체 집단에 대한 모비율의 95% 신뢰구간을 구하여라.

6.5.1 어느 회사에서 생산된 플라스틱 제품의 충격에 대한 강도의 평균을 알고 싶어한다. 99% 신뢰도로 오차의 한계가 20psi이내에 있으려면 얼마나 많은 충격 강도의 실험을 해야겠는가? 이전의 경험에 의하면 σ^2의 추정치가 4900이라고 한다.

6.5.2 2500명의 근로자를 두고 있는 어느 회사에서 근로자들의 통근에 걸리는 평균시간을 알고 싶어한다. 조사자는 99% 신뢰도를 가지고 오차의 한계가 1분이내인 추청치를 원한다. 예비표본에서 얻은 분산이 25^2 이었다. 이 때 조사자가 얻어야 할 표본의 크기는 얼마인가?

6.5.3 어느 심리학자는 한 회사 직원들의 평균 IQ의 구간추정을 하려고 한다. 95% 신뢰구간일 때 오차의 한계가 5점 이내로 하려고 한다. 전에 경험에 의하면 이러한 집단의 IQ는 근사적으로 분산이 100인 정규분포를 따른다고 한다. 심리학자는 이 집단으로부터 얼마나 많은 표본을 택해야 하는가?

6.5.4 한 대학에서 학생들의 요구가 충분히 높다면 토요일 강의를 하려고 한다. 학교에서 토요일 강의를 제공했을 때 이 강의에 등록하는 학생의 비율을 95% 신뢰구간으로 추정하려고 한다. 오차의 한계가 0.05이내에 있기 위해 조사해야 할 학생들의 표본의 수는?

6.5.5 산업의학의 한 전문가는 전체 신발공장 중에서 3일 이상 병으로 결근할 때 의사의 진단서를 근로자들로부터 요구하는 공장의 비율을 알고 싶어한다. 95% 신뢰도를 가지고 0.05이내의 오차의 한계로 추정하려면 얼마만큼의 표본의 수가 필요한가? 단, 조사자는 모비율이 0.3을 넘지 못한다고 생각한다.

6.5.6 한 시장조사 분석가는 어느 동네에서 적어도 가족의 한 사람이 신문 광고를 보는 가구의 비율이 얼마인지를 알고 싶다. 분석가는 90% 신뢰도를 가지고 모비율의 오차의 한계가 0.04이내에 있기를 원한다. 20가구의 예비조사에서 응답가구의 35%가 적어도 가족중 한 사람이 신문광고를 본다는 것을 알았다. 이 때 표본가구의 수는 얼마로 해야 되는가?

Ⅲ. 한 모집의 가설검정

1. 모평균의 가설검정

모집단의 모수가 궁금하여 이를 6장처럼 추정하는 문제도 있지만 모수의 값에 대한 가설이 타당한지에 관심이 있을 수도 있다. 예를 들면,

1) 어느 과자제품의 겉봉지에 용량이 200g이라 표시되어 있다. 과연 표시된 용량만큼 과자가 들어있을까?
2) 어느 전구공장에서 새로 개발한 전구가 과거의 것보다 훨씬 전구 수명이 길다고 선전한다. 과연 이 선전이 믿을만 할까?

이와같은 의문(가설)에 대한 답을 주는것이 가설검정(hypothesis testing)이다. 즉 가설검정은 표본을 이용하여 미지의 모집단 모수에 대한 두 가지 가설을 놓고 어느 가설을 선택할 것인지 통계적으로 의사결정을 하는 것이다. 가설검정의 이론을 다음의 예를 들어 설명하자.

[예 7.1] 어느 전구 공장에서 기존의 생산방식으로 만들어진 전구의 평균수명은 1500시간, 표준편차는 200시간으로 알려져 있다. 최근 이 회사에서 새로운 생산방식을 도입하려는데, 이 방식으로 생산된 전구는 평균수명이 1600 시간이라고 주장한다. 이와같은 주장을 확인하기 위해 30개의 표본을 단순확률 추출하여 표본평균을 구하여 보니 $\bar{x} = 1555$ 이었다. 과연 새 방식의 전구수명이 1600시간이라고 할 수 있는가?

<풀이>

이 문제의 질문에 대한 답을 통계적으로 접근하는 방법은 먼저 모평균(μ)에 대한 서로 다른 주장에 대해 두개의 가설을 세운다. 즉,

$H_0: \mu = 1500$
$H_1: \mu = 1600$

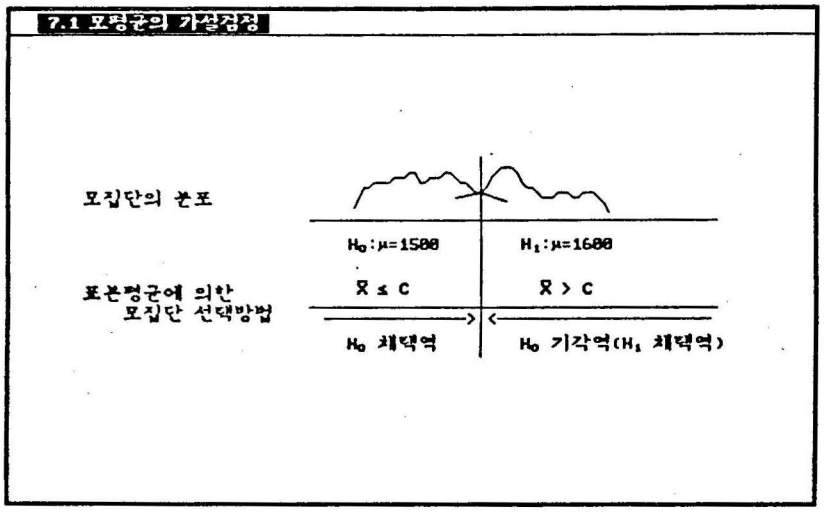
여기서 H_0를 귀무가설(null hypothesis), H_1을 대립가설(alternative hypothesis)이라 부른다. 대개의 경우에 귀무가설 H_0는 기존의 알려져 있는 사실로 정하고, 대립가설은 새로운 사실 또는 현재의 믿음에 변화가 있는 사실을 정한다. 그래서 두 가설중 하나를 선택할 때, '확실한 근거가 있기전에는 대립가설(변화된 사실)을 선택하지 않고 귀무가설(현재의 사실)을 받아들인다'는 것이 가설검정의 기본적인 생각이다. 이러한 가설검정을 보수적 의사결정방법 (conservative decision making)이라고 한다.

두 개의 가설중 하나를 선택하는 상식적인 기준은 '표본평균이 어느 가설에 더 가까운가'일 것이다. 이러한 거리개념를 이용하는 상식적 기준에 의하면 표본평균 1555이 $H_1: \mu = 1600$에 더 가까우므로 대립가설을 선택할 것이다. 통계적 가설검정도 상식적 기준에 근거한 것인데 다만 $\bar{x}$ 의 표본분포 이

론도 함께 고려한 것이다. 즉, 통계적 가설검정은 $\bar{x}$ 의 표본분포 이론에 근거한 기준값(critical value) C 를 선정한 후 (다음 쪽에 설명이 나옴)

'$\bar{x}$ 가 C보다 작으면 가설 H_0 를 채택하고, 아니면 H_0 를 기각(H_1 채택)'

이라는 선택기준(decision rule)으로 한 가설을 선택하게 된다. 이때 $\bar{x}$ 가 C보다 작은 영역 ($\{\bar{x} < C\}$)을 H_0 채택역 (acceptance region), $\bar{x}$가 C 보다 큰 영역 ($\{\bar{x} \geq C\}$)을 H_0 기각역(rejection region)이라 부른다.



<그림 7.1> 채택역과 기각역

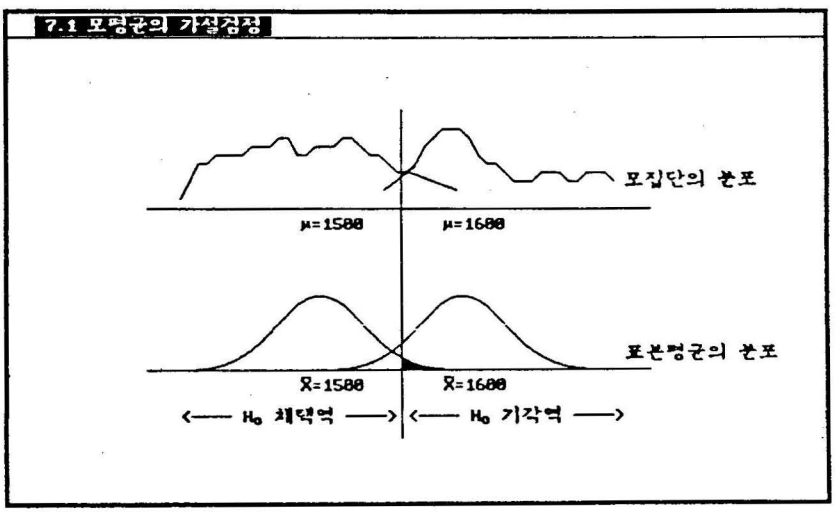
이러한 선택방법에 의해 한 가설을 선택하게 되면 반드시 그 결정에는 두 가지 오류의 가능성이 있다. 즉, H_0 가 참일 때 H_1 을 채택하는 제1종오류 (Type I Error)와 H_1이 참일 때 H_0를 채택하는 제2종오류(Type II Error)가 있다. 이들을 표로 요약하면 다음과 같다.

<표 7.1> 가설검정의 오류

		실 제 상 황	
		H_0 참	H_1 참
검정결과:	H_0 채택	옳은 결정	2 종 오류
	H_1 채택	1 종 오류	옳은 결정

이 두가지 오류는 표본의 크기가 일정할 때 어느 한 오류를 줄이려고 하면 다른 오류가 커지게 된다. 그래서 귀무가설 H_0 를 '과거나 현재의 사실'로 하고 '확실한 근거가 없는 한 귀무가설을 채택'하는 보수적 결정방식을 생각해 낸 것이다. 이러한 보수적 방식에서는 '귀무가설이 참인데 대립가설을 선택'하는 제1종오류가 우리에게 더 큰 손실을 가져오므로 이를 가능하면 줄이려고 노력한다. 즉, 통계적 가설검정에서는 제1종오류가 발생할 확률의 허용한계(대개 5%이나, 엄격한 검정에서는 1%를 사용)를 결정하여, 이 한계를 만족시키는 선택기준을 이용한다. 이 제1종오류가 발생할 확률의 허용한계를 유의수준(significance level)이라 하며 흔히 α 로 나타낸다. 제2종오류의 확률은 대개 β로 표시한다.

유의수준만 정하면 5장의 표본평균의 분포를 이용하여 두 가설에 대한 선택기준을 마련할 수 있다. <그림 7.2>는 두 개의 가설에 대한 가상적인 모집단의 분포와, 각 모집단에 대한 모든 가능한 표본평균의 분포를 그린 것이다.



<그림 7.2> 통계적 가설검정

표본평균의 분포는 중심극한정리에 의해 근사적으로 H_0: μ = 1500 의 모집단인 경우 $N(1500, 200^2/30)$이고, H_1: μ = 1600 의 모집단인 경우 $N(1600, 200^2/30)$이 된다. 여기서 각 모집단의 표준편차는 과거의 자료인 200으로 가정하였다. 이때

'$\bar{x}$ 가 C보다 적으면 H_0 를 채택하고, $\bar{x}$ 가 C보다 크면 H_1 을 채택'

하는 선택기준을 세우면 그림의 빗금친 부분이 제1종오류가 발생할 확률을

나타낸다. 만일에 유의수준, 즉 제1종오류 발생확률의 허용한계를 5%로 하면 (따라서 $P(\bar{x} < C) = 0.95$) C는 정규분포이론에 의해

$$1500 + (1.645)(200/\sqrt{30}) = 1560.06$$

가 된다. 따라서 선택기준은

'$\bar{x} < 1500 + (1.645)(200/\sqrt{30}) = 1560.06$ 이면 H_0 를 채택하고, 아니면 H_0 를 기각(H_1 을 채택)한다.'

이다. 위 문제에서는 $\bar{x} = 1555$ 이므로 H_0 를 채택한다. 즉, H_0: $\mu = 1500$ 인 가설이 맞다고 판정하는 것인데 이는 상식적 기준에 의한 결과와 상반되는 것이다. 그 이유는 보수적 의사결정 방식이기 때문에 표본의 결과가 귀무가설을 기각할 충분한 근거가 되지 못한다는 것이다. 위의 선택기준은 보수적 결정방식에 의한 결과라는 것을 강조하는 의미로 다음과 같이 쓰기도 한다.

'$\bar{x} < 1560.06$ 이면 H_0 를 기각하지 못하고, 아니면 H_0 를 기각한다.'

또 위의 선택기준은 계산편의상 아래와 같이 쓰기도 한다.

'$(\bar{x} - 1500)/(200/\sqrt{30}) < 1.645$ 이면 H_0 를 채택, 아니면 H_0 기각.'

이 경우 $\bar{x} = 1555$ 일때 $(1555-1500)/(200/\sqrt{30}) = 1.506$ 은 1.645보다 작으므로 H_0 를 채택한다.. □

위의 예에서 보았듯이 보수적 결정방식에 의한 가설검정에서는 제1종오류가 발생할 확률에 근거한 선택기준을 만들었기 때문에, 대립가설 H_1: $\mu = 1600$ 은 선택기준에 단지 '가설 H_0 의 모평균(μ= 1500)보다 크다'라는 점만 고려되었다. 즉 위의 예제에서 대립가설을 H_1: $\mu > 1500$ 이라 하여도 똑같은 계산에 의해 H_0 를 기각한다는 결론을 내릴 수 있다.

일반적으로 모평균에 대한 가설검정에서 대립가설의 형태는 크게

1) H_1: $\mu > \mu_0$ 2) H_1: $\mu < \mu_0$ 3) H_1: $\mu \neq \mu_0$

의 세가지이다. 1)은 가설 H_0 오른쪽에 기각역을 가지므로 **우측검정**(right-sided test), 2)는 가설 H_0 왼쪽에 기각역을 가지므로 **좌측검정**(left-sided test), 3)은 가설 H_0 양편에 기각역을 가지므로 **양측검정**(two-sided test)이라 부른다. 모표준편차를 알 경우(현실적이지는 못하지만 이론을 전개시키기 위한 것임) 각각의 형태에 대한 가설의 선택기준은 <표 7.2>와 같다. 여기서 α 는 유의수준이다.

<표 7.2> 모평균의 가설검정 - 모표준편차 σ 를 알 경우

가설의 종류	선 택 기 준
1) H_0: $\mu = \mu_0$ H_1: $\mu > \mu_0$	$\bar{x} > \mu_0 + Z_{1-\alpha}(\sigma/\sqrt{n})$ 이면 H_0 기각, 아니면 H_0 채택
2) H_0: $\mu = \mu_0$ H_1: $\mu < \mu_0$	$\bar{x} < \mu_0 - Z_{1-\alpha}(\sigma/\sqrt{n})$ 이면 H_0 기각, 아니면 H_0 채택
3) H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	$\bar{x} > \mu_0 + Z_{1-\alpha/2}(\sigma/\sqrt{n})$ 또는 $\bar{x} < \mu_0 - Z_{1-\alpha/2}(\sigma/\sqrt{n})$ 이면 H_0 기각, 아니면 H_0 채택

참고: 1)의 경우 H_0: $\mu \leq \mu_0$ 로, 2)의 경우 H_0: $\mu \geq \mu_0$ 로 쓸 수 있다.

위의 선택기준 공식은 간단히

1) $(\bar{x}-\mu_0)/(\sigma/\sqrt{n}) > Z_{1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택

2) $(\bar{x}-\mu_0)/(\sigma/\sqrt{n}) < -Z_{1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택

3) $|(\bar{x}-\mu_0)/(\sigma/\sqrt{n})| > Z_{1-\alpha/2}$ 이면 H_0 기각, 아니면 H_0 채택

으로 쓸 수 있다. 여기서 $(\bar{x}-\mu_0)/(\sigma/\sqrt{n})$ 를 검정통계량(test statistic)이라고 부르기도 한다. 일반적으로 모표준편차 σ 는 미지수이다. 그러나 표본의 크기가 충분히 클 경우(대략 30이상), $(\bar{x}-\mu_0)/(s/\sqrt{n})$ 는 정규분포에 근사하게 되므로 앞의 가설검정의 공식에서 σ대신 표본표준편차 s를 이용하여 모평균의 가설검정을 할 수 있다.

<표 7.3> 모평균의 가설검정 - σ 를 모르나 대표본인 경우

가설의 종류	선 택 기 준
1) H_0: $\mu = \mu_0$ H_1: $\mu > \mu_0$	$(\bar{x}-\mu_0)/(s/\sqrt{n}) > Z_{1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
2) H_0: $\mu = \mu_0$ H_1: $\mu < \mu_0$	$(\bar{x}-\mu_0)/(s/\sqrt{n}) < -Z_{1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
3) H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	$\lvert(\bar{x}-\mu_0)/(s/\sqrt{n})\rvert > Z_{1-\alpha/2}$ 이면 H_0 기각, 아니면 H_0 채택

[예 7.1]에서 $\bar{x}$가 1555 일 경우나 1540 일 경우 모두 가설 H_0 는 기각되지 못하지만 기각되지 못한 근거의 강력한 정도가 다르다. 가설이 기각되지 못한 근거의 정도는 관찰된 표본평균의 값을 기준값으로 하였을 때의 제1종오류 확률을 계산하면 알 수 있는데 이를 p-값(p-value)이라 한다. 즉, p-값은 측정된 표본평균이 모든 가능한 표본평균중에서 어디에 위치하고 있는지를 알려 준다. $\bar{x}$가 1540일 경우의 p-값은, $\bar{x}$가 1555일 경우의 p-값보다 크다. p-값은 더 클수록 기각되지 못한 강력한 근거가 된다(귀무가설 H_0가 기각되는 경우는 더 작을 수록 기각된 근거가 더 강력하다). 따라서 p-값이 분석자가 고려하는 유의수준보다 작으면 표본평균이 기각역에 있다는 것을 뜻하기 때문에 H_0 를 기각한다. 대부분의 통계패키지에서는 p-값을 계산하여 준다.

p-값을 이용한 가설 선택기준
p-값이 유의수준보다 작으면 H_0 기각, 아니면 H_0 채택

구체적으로 각각의 검정형태에 따른 p-값의 계산은 아래와 같다.

〈표 7.4〉 p-값의 계산

가설의 종류	p-값, 여기서 $\bar{x}_{obs}$ 는 관찰된 표본평균의 값
1) $H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$	$P(\bar{x} > \bar{x}_{obs})$
2) $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$P(\bar{x} < \bar{x}_{obs})$
3) $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	만약 $\bar{x}_{obs} > \mu_0$ 이면 $2P(\bar{x} > \bar{x}_{obs})$, 아니면 $2P(\bar{x} < \bar{x}_{obs})$

표본의 크기가 작을 경우 모집단이 정규분포를 따른다면 $(\bar{x}-\mu_0)/(s/\sqrt{n})$ 는 자유도가 (n-1)인 t분포를 따르므로 이 경우의 모평균의 가설검정은 위의 선택기준에서 표준정규분포대신 t 분포를 사용하여 다음과 같이 한다.

〈표 7.5〉 모평균의 가설검정 - 소표본의 경우 (모집단이 정규분포)

가설의 종류	선 택 기 준
1) H_0: $\mu = \mu_0$ H_1: $\mu > \mu_0$	$(\bar{x}-\mu_0)/(s/\sqrt{n}) > t_{n-1,1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
2) H_0: $\mu = \mu_0$ H_1: $\mu < \mu_0$	$(\bar{x}-\mu_0)/(s/\sqrt{n}) < -t_{n-1,1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
3) H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$	$\lvert(\bar{x}-\mu_0)/(s/\sqrt{n})\rvert > t_{n-1,1-\alpha/2}$ 이면 H_0 기각, 아니면 H_0 채택

참고: 1)의 경우 H_0: $\mu \leq \mu_0$ 로, 2)의 경우 H_0: $\mu \geq \mu_0$ 로 쓸 수 있다.

[예 7.2] 작년도의 대졸초임 모평균은 50만원이었다. 금년도 대졸 신입사원 100명을 조사하였더니 평균이 53만원, 표준편차가 10만원이었다.

1) 금년도 대졸 초임이 올랐는가 검정하고 p-값을 구하라. $\alpha = 1\%$

2) 금년도 대졸 초임이 작년과 같은지 검정하고 p-값을 구하라. $\alpha = 1\%$

〈풀이〉

1) 이 예의 가설은 H_0: $\mu = 50$, H_1: $\mu > 50$ 인 우측검정이다. 표본의 크기가 크므로 (n=100) 가설 선택기준은

'$(\bar{x}-\mu_0)/(s/\sqrt{n}) > Z_{1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택'

'$(\bar{x}-50)/(10/\sqrt{100}) > Z_{1-0.01}$ 이면 H_0 기각, 아니면 H_0 채택'

따라서

$(53-50) / (10/10) = 3$, $Z_{0.99} = 2.326$

이므로 H_0 는 기각이 된다. 가설의 선택기준을 아래와 같이 쓸 수도 있다.

'$\bar{x} > 50 + (2.3265)(10/\sqrt{100})$ 이면 H_0 기각, 아니면 H_0 채택'

p-값은 표본평균을 기준값으로 했을 때의 제1종 오류의 확률이므로 표본평균이 53보다 클 확률을 구하면 된다. H_0: $\mu = 50$ 가 참이라는 가정하에서 $\bar{x}$의 분포는 근사적으로 N(50, 100/100)인 분포이므로

$$p\text{-값} = P(\bar{x} > 53) = P(Z > (53-50)/(10/10)) = P(Z > 3) = 0.0013$$

이다. 여기서 $(53-50)/(10/10) = 3$ 는 위에서 이미 계산했던 검정통계량의 값임을 주목하면, p-값은 바로 검정통계량의 분포(여기서는 표준정규분포)를 이용하여 구할 수 있다.

2) 이 예의 가설은 H_0: $\mu = 50$, H_1: $\mu \neq 50$ 인 양측검정이다. 표본의 크기가 크므로 (n=100) 가설 선택기준은

' $|(\bar{x}-\mu_0)/(s/\sqrt{n})| > Z_{1-\alpha/2}$ 이면 H_0 기각, 아니면 H_0 채택'

' $|(\bar{x}-50)/(10/\sqrt{100})| > Z_{1-0.005}$ 이면 H_0 기각, 아니면 H_0 채택'

따라서

$(53-50) / (10/10) = 3, \quad Z_{0.995} = 2.575$

이므로 H_0 는 기각이 된다. p-값은 표본평균을 기준값으로 했을 때의 제1종오류의 확률이므로 표본평균이 53보다 클 확률을 구하여 2를 곱하면 된다. H_0: $\mu = 50$ 가 참이라는 가정하에서 $\bar{x}$의 분포는 근사적으로 N(50,100/100)인 분포이므로

$$p\text{-값} = 2P(\bar{x} > 53) = 2P(Z > (53-50)/(10/10)) = 2P(Z > 3) = 0.0026$$

이다. □

[예 7.3] 위 예제에서 표본의 크기가 16명 이고 모집단이 정규분포라면 금년도 대졸초임이 올랐는가 검정하라. 또 p-값은?

<풀이>

소표본이므로 가설 선택기준은

' $(\bar{x}-\mu_0)/(s/\sqrt{n}) > t_{n-1,1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택

' $(\bar{x}-50)/(10/\sqrt{16}) > t_{16-1,1-0.01}$ 이면 H_0 기각, 아니면 H_0 채택

따라서

$(53-50) / (10/\sqrt{16}) = 1.2, \quad t_{15,0.99} = 2.602$

이므로 H_0 가 채택이 된다. 가설 선택기준을

' $\bar{x} > 50 + (2.602)(10/\sqrt{16})$ 이면 H_0 기각, 아니면 H_0 채택

으로 쓸 수도 있음을 유의하라. p-값은 t_{15} 분포에서 검정통계량의 값 1.2 보다 클 확률이므로 프로그램을 이용하면 0.1244임을 알 수 있다. □

이 절에서는 표본의 크기가 이미 주어진 경우에 어떻게 가설검정하는가 알아보았다. 표본의 크기가 일정할 경우에 제1종오류의 가능성을 줄이려고 하면 제2종오류의 가능성이 커지므로 두 종류의 오류를 동시에 줄일 수는 없다. 따라서, 표본의 크기가 미리 정해졌거나, 자료가 주어진 경우에는 보수적 결정방법으로 제1종오류만을 고려하는 가설검정을 하였다. 그러나, 만일 표본의 크기를 조절할 수 있다면 표본의 크기를 적절히 크게함으로서 두 종류의 오류를 함께 고려하는 가설검정 방법도 있는데 이것에 대해서는 4절에서 알아보기로 한다.

2. 모분산의 가설검정

다음은 미지의 모집단의 분산을 가설검정하기 위한 몇 가지 예이다.

1) 한 자동차회사에 현재 볼트를 납품하는 부품회사의 볼트는 직경이 평균 7㎜, 분산이 0.25라고 한다. 최근 경쟁회사는 자기회사의 볼트는 직경의 평균이 7㎜, 분산이 0.16이라고 주장하면서 납품을 신청하고 있다. 과연 이 주장이 맞는지 어떻게 알아볼 수 있는가?
2) 작년도 대입 학력고사 수학점수의 분산이 100 이라 한다. 금년도 수학문제가 작년보다 너무 쉽다고 한다. 학력고사 성적의 분산이 작년보다 작아졌는지 어떻게 알아 볼 수 있나?

5.3절에서 우리는 표본분산(s^2)의 표본분포를 알아 보았다. 즉, 모집단이 분산이 σ^2 인 정규분포를 따를 때 표본의 크기가 n이라면 $(n-1)s^2/\sigma^2$ 은 자유도가 (n-1)인 카이제곱분포를 한다. 이 이론을 이용하면 모분산에 대한 가설검정을 다음과 같이 할 수 있다.

〈표 7.6〉 모분산의 가설검정 - 모집단이 정규분포인 경우

가설의 종류	선 택 기 준
1) $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	$(n-1)s^2/\sigma_0^2 > \chi^2_{n-1,1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
2) $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$	$(n-1)s^2/\sigma_0^2 < \chi^2_{n-1,\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
3) $H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$(n-1)s^2/\sigma_0^2 > \chi^2_{n-1,1-\alpha/2}$ 또는 $(n-1)s^2/\sigma_0^2 < \chi^2_{n-1,\alpha/2}$ 이면 H_0 기각, 아니면 H_0 채택

참고: 1)에서 $H_0: \sigma^2 \leq \sigma_0^2$ 으로, 2)에서 $H_0: \sigma^2 \geq \sigma_0^2$ 로 쓸 수 있다.

[예 7.4] 자동차 부속품중 볼트를 생산하는 회사가 있다. 이 볼트의 직경의 규격은 15㎜인데 분산이 0.0005^2이내라면 납품할 수 있다. 최근에 생산된 제품중 25개를 단순확률 표본추출하여 분산을 조사하였더니 0.0006^2이었다. 볼트의 직경이 정규분포를 따른다고 가정하였을 때, 최근 생산된 제품을 납품할 수 있는지 5% 유의수준으로 가설 검정을 하여라.

<풀이>

이 문제의 가설은 $H_0: \sigma^2 = 0.0005^2$ $H_1: \sigma^2 > 0.0005^2$이다. 따라서 선택기준은

' $(n-1)s^2/\sigma^2_0 > \chi^2_{n-1.1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택 '

이다. $s^2 = 0.0006^2$ 이므로

$(25-1)\times 0.0006^2/0.0005^2 = 34.56$, $\chi^2_{25-1.1-0.05} = \chi^2_{24.0.95} = 36.415$

따라서 가설 H_0 가 채택된다. □

3. 모비율의 가설검정

모집단의 미지비율을 가설검정하기 위한 몇 가지 예를 들어보자.

1) 금년도 대통령 선거에서 어느 후보의 지지율이 과연 50%를 넘을까?
2) 작년도 실업율이 7%였다고 한다. 올해의 실업률은 높아졌는가?
3) 자동차 부속품 1만개를 배로 수입하는데 과거의 경험으로 보아 이중 2%가 불량품이었다. 이번에도 불량품이 2% 일까?

표본의 크기가 충분히 클 때 표본비율($\hat{p}$)의 표본분포는 평균이 모비율(p)이고 분산이 $p(1-p)/n$ 인 정규분포에 근사하게 된다. 따라서 대표본일 때의 모평균의 가설검정과 유사하게 모비율의 가설검정을 다음과 같이 할 수 있다.

〈표 7.7〉 모비율의 가설검정 - 대표본일 경우

가설의 종류	선 택 기 준
1) $H_0: p = p_0$ $H_1: p > p_0$	$\dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > Z_{1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
2) $H_0: p = p_0$ $H_1: p < p_0$	$\dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < -Z_{1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
3) $H_0: p = p_0$ $H_1: p \neq p_0$	$\left\lvert \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right\rvert > Z_{1-\alpha/2}$ 이면 H_0 기각, 아니면 H_0채택

참고: 구간추정에서와 마찬가지로 대표본의 판정기준은 $np_0 > 5$, $n(1-p_0) > 5$ 임.
1)에서 $H_0: p \leq p_0$ 로, 2)의 경우 $H_0: p \geq p_0$ 로 쓸 수 있다.

[예 7.5] 한 지역의 의원선거에서 특정후보에 대한 지난달의 여론조사 결과는 지지율이 60%이었다. 최근에 지지율에 변동이 있는지 알아보기 위해 100명을 단순확률 추출하였더니 55명이 지지를 하였다. 유의수준 5%로 특정후보에 대한 현재 지지율이 60%인가 검정하라.

<풀이>

이 문제의 가설은 $H_0: p = 0.6$, $H_1: p \neq 0.6$ 이다. $np_0 = 60$, $n(1-p_0) = 40$ 이므로 대표본이라 할 수 있고, 따라서 선택기준은

$$\left| \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right| > Z_{1-\alpha/2} \text{ 이면 } H_0 \text{ 기각, 아니면 } H_0 \text{ 채택'}$$

이다. $\hat{p} = 55/100 = 0.55$ 이므로

$$\left| \frac{0.55 - 0.6}{\sqrt{.6(1-.6)/100}} \right| = 1.02, \quad Z_{1-0.05/2} = Z_{0.975} = 1.96$$

따라서 가설 H_0 는 채택된다. □

소표본인 경우는 이항분포를 사용하여 가설검정을 하는데 여기서는 생략하기로 한다.

4. SAS 실습

SAS에서 한 집단의 모평균에 대한 가설검정을 하기 위한 procedure는 2장에서 살펴본 MEANS에서 T PRT라는 options을 사용하면 된다. 예로서 2장의 'ex0201.dat'에서 모집단의 봉급평균이 100인지 알고 싶다고 하자. SAS로서 가설검정을 하기 위해서는 다음과 같이 salary-100이라는 새로운 변수를 만든 다음 이 변수에 대해 MEANS T PRT; 를 사용한다. SAS 출력은 <표 7.8>과 같다. 여기서 T 는 관측된 t값이고, Prob>|T| 는 양측검정에 대한 p-값이다.

```
DATA ex2;
   INFILE 'ex0201.dat';
   INPUT sex marital age occup educ salary;
RUN;
DATA new; SET ex2;
   newsal = salary-100; RUN;
PROC MEANS T PRT; VAR salary; RUN;
```

<표 7.8> 봉급에 대한 MEANS T PRT: procedure 출력

Analysis Variable : SALARY

N Obs	T	Prob>|T|
40	12.0051114	0.0001

연 습 문 제

7.1.1 모집단이 표준편차가 50인 정규분포를 따른다고 가정하자. 모집단에서 25개를 단순확률 표본추출하여 계산된 평균이 70이다. 유의수준 0.01 에서 H_0: μ = 100 이라는 가설을 검정하라.

7.1.2 한 볼트 제조업자는 볼트의 표준편차가 0.020 인치이고 평균길이가 4.5 인치라고 주장한다. 16개의 표본을 추출하여 4.512 라는 평균을 얻었을 때 볼트의 실제 평균 길이가 제조업자의 주장과 다르다고 말할 수 있나? 볼트의 길이는 정규분포를 따른다고 가정하고 유의수준 = 0.01이라 하자.

7.1.3 한 화학약품 제조업자는 고정된 양의 어떤 성분들에 증류수를 첨가한 화합물을 생산한다. 물의 필요량은 성분들의 순도에 달려 있다. 제조업자의 경험으로는 정상생산을 위한 물의 필요량이 6 리터이고 표준편차가 1 리터라고 한다. 제품 9개의 표본추출로 7리터란 표본평균을 얻었다. 유의수준=0.05 일 때 제품들이 정상생산이라고 볼 수 있는가?

7.1.4 한 심리학자가 신체장애 근로자들에 관한 연구를 하고 있다. 이 심리학자는 과거의 경험에 의거하여 이런 장애 근로자들 모집단의 평균 사교(교제)점수가 80 보다 크다고 믿었다. 점수 모집단으로부터 고용인 20명을 표본추출하여 다음의 결과를 얻었다.

99, 69, 91, 97, 70, 99, 72, 74, 74, 76,
96, 97, 68, 71, 99, 78, 76, 78, 83, 66.

심리학자는 모집단의 평균 사교점수가 옳을 지 알고 싶어한다. 모집단이 표준편차 10인 정규분포를 따를 때, 유의수준을 0.05로 하여 검정하라.

7.1.5 열 감지 장치의 평균작동 온도가 제조업자에 따르면 화씨 190도이다. 16개 장치의 표본에서 얻어진 작동 온도의 평균과 표준편차는 각각 195도와 8도이다. 이 자료로 평균 작동 온도가 제조업자의 주장보다 크다고 할 수 있는가? 유의수준은 0.05로 하고 작동 온도는 근사적으로 정규분포를 따른다고 가정하자.

7.1.6 햄버거 가게에서 팔리는 햄버거중 25개를 표본추출하여 평균 무게가 3.8온스이고 표준편차가 0.5임을 알았다. 이 자료로부터 모집단의 평균이 4온스 보다 작다고 할 수 있는가? 유의수준을 0.05로 놓고 햄버거의 무게는 근사적으로 정규분포를 따른다.

7.1.7 다음은 한 도매 식료품 회사의 선적부서에 일하는 고용인 중 10명을 단순확률 추출하여 얻은 몸무게이다.

154, 154, 186, 243, 159, 174, 183, 163, 192, 181.

이 자료에 근거하여 선적부서에서 일하는 고용인들의 평균 무게가 160보다 크다고 할 수 있는가? 유의수준은 0.05로 하라.

7.1.8 한 주택 중개인이 어느 지역의 평균 집값이 4억원보다 크다고 주장한다. 36집을 표본추출하여 평균이 5억원이고 표준편차가 1억원임을 알았다. 유의수준 0.05하에서 중개인의 주장이 옳은가 검정하라.

7.1.9 한 빌딩의 경영자는 주차장의 많은 차들이 주말에 평균 90분이상 주차한다고 생각한다. 주말에 오는 100대의 자동차를 표본추출하여 96분이란 평균 주차시간과 30분이란 표준편차를 얻었다. 유의수준=0.05에서 경영자의 주장이 옳은지 검정하라.

7.1.10 여러 전기 회사의 상담자로 일하는 한 산업 심리학자는, 어떤 일에 적응을 못한 비숙련 고용인 40명을 표본추출하여 얻은 적응점수가 다음과 같다.

73	57	96	78	74	42	55	44	91	91
50	65	46	63	82	60	97	79	85	79
92	50	42	46	86	81	81	83	64	76
40	57	78	66	84	96	94	70	70	81

모분산은 280이라고 알려졌다. 생산 감독자가 모든 고용인들의 평균 적응점수가 60보다 크다고 주장한다. 이 주장이 옳은지를 유의수준 0.05에서 검정하라. 또 p 값은 ?

7.1.11 지붕에 바르는 타르를 만드는 한 회사는 불순물의 비율이 평균 3% 가 넘지 않길 원한다. 타르 30통을 표본추출하여 불순물의 비율이 다음과 같을 때 이

자료에 의해 모평균이 3%보다 작다고 할 수 있겠는가?

3	3	1	1	0.5	2	2	4	5	4	5	3	1	3	1
4	1	1	4	2	5	3	1	1	1	0.75	1.5	3	3	2

7.2.1 정규분포를 따르는 모집단으로부터 크기 21인 표본을 추출하여 분산 10을 얻었다. 유의수준 0.05에서 귀무가설 $\sigma^2 = 15$, 대립가설 $\sigma^2 \neq 15$ 를 검정하라.

7.2.2 금속 세척기의 내부 지름은 만들어지는 공정이 관리하에 있을 때는 분산 0.00005^2 이하를 가진다고 한다. 조립라인으로부터 표본 31개를 추출하여 분산 0.000061^2 를 얻었다. 이 자료에 의하면 조립공정이 관리 밖에 있다고 할 수 있겠는가? 유의수준을 0.05라 하고 답을 얻기 위하여 어떤 가정이 필요한가?

7.2.3 어떤 제조업자가 물품을 들이려면 합성섬유의 장력강도는 분산이 5이하여야만 한다. 새 선적물로 부터 25개의 표본을 추출하여 분산이 7일 때 이 자료는 제조업자가 선적을 거절하기 위한 충분한 한계를 제공하는가? 유의수준이 0.05이고 섬유의 장력강도가 근사적으로 정규분포를 따른다고 가정하자.

7.3.1 학생수가 10,000명인 어떤 대학에서 학생용 주차장을 만들려고 한다. 학교 당국은 학생의 20% 이상이 자동차로 등교한다고 생각한다. 만약 250명을 표본추출하여 65명이 자동차 등교를 한다고 나왔다면 학교 당국의 생각이 옳았는지를 유의수준 0.05로 검정하라.

7.3.2 어떤 회사의 상임 회계사는 경비명세서에 대한 사무원의 실수를 주시하여 명세서의 20% 이상이 적어도 하나의 실수를 포함한다고 믿는다. 만약 400개의 명세서를 표본추출하여 100개에서 적어도 하나의 실수를 발견했다면 회계사의 믿음이 옳았는 지를 검정하라. 유의수준은 0.05로 놓아라.

7.3.3 전업에 대한 연구에서 한 연구자가, 작년동안 전업한 200명의 고급 고용인을 만나보았다. 그 중 30 명이 그들이 전업한 이유를 그들의 옛직업에서 승진에 대한 큰 전망을 기대할 수 없었기 때문이라 진술했다. 이 자료로부터 앞과 같은 이유로 직업을 바꾼 고용인이 전체의 20%보다 적다고 할 수 있는가? 유의수준은 0.05이다.

Ⅳ. 두 모집단의 가설검정

1. 두 모평균의 가설검정

두 모집단의 평균을 비교하는 문제들은 우리 주변에 아주 많이 있다. 예를 들어보자.

1) 금년도 대졸 사원의 초임이 남녀별로 차이가 있을까?
2) 두 생산라인에서 생산되는 제품들의 무게에 차이가 있을까?
3) 타자속도를 증가시키기 위하여 타자수에게 실시한 특별교육이 과연 타자속도의 증가를 가져 왔을까?

이와 같이 두 모집단의 평균(μ_1 과 μ_2)의 비교는 모평균의 차 $\mu_1 - \mu_2$ 가 0 보다 큰가, 작은가, 같은가 하는 가설을 검정함으로써 가능하다. 이러한 두 모평균의 비교는 각 모집단에서 추출된 표본들이 서로 독립적으로 추출되었을 경우와 아닌 경우(대응비교라 함)에 따라 검정방법이 다르다.

가. 표본들이 독립적으로 추출된 경우

모집단에서 서로 독립적으로 표본을 추출하였을 때 모평균의 차 $\mu_1 - \mu_2$ 의 추정량은 표본평균의 차 $\bar{x}_1 - \bar{x}_2$ 이며, 모든 가능한 표본평균의 차는 표본이 충분히 클 경우 근사적으로 평균이 $\mu_1 - \mu_2$ 이고 분산이 $\sigma_1^2/n_1 + \sigma_2^2/n_2$ 인 정규분포를 따르게 된다. 여기서 두 모집단의 분산 σ_1^2 과 σ_2^2 은 대개 알려져 있지 않으므로 대표본인 경우 $s_1^2/n_1 + s_2^2/n_2$ 이 $\bar{x}_1 - \bar{x}_2$ 에 대한 분산의 추정치로 사용된다. 이것을 이용하여 대표본인 경우 두 모평균의 차이에 대한 검정을 다음과 같이 할 수 있다.

〈표 8.1〉 두 모평균의 가설검정
- 대표본이고 ($n_1 \geq 30$, $n_2 \geq 30$) 표본이 서로 독립적으로 추출되었을 경우

가설의 종류	선 택 기 준
1) $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} > Z_{1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
2) $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} < -Z_{1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
3) $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$\left\lvert \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \right\rvert > Z_{1-\alpha/2}$ 이면 H_0기각, 아니면 H_0채택

[예 8.1] 금년에 기업체에 취업한 대졸 사원들의 초임을 남여별로 조사하였다. 단순확률 추출된 30명의 남자 대졸사원의 월별 초임의 평균은 452,000원, 표준편차는 22,000원 이었고, 35명의 여자 대졸사원의 초임평균은 395,000원, 표준편차는 31,000원 이었다. 남자와 여자의 초임이 차이가 있는가? 1%의 유의수준으로 검정하라. p-값은?

<풀이>

이 문제의 가설은 H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$ 이다. 따라서 선택기준은

$$\left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \right| > Z_{1-\alpha/2} \text{ 이면 } H_0 \text{ 기각, 아니면 } H_0 \text{ 채택}$$

이다.

$n_1 = 30$, $\bar{x}_1 = 452000$, $s_1 = 22000$,

$n_2 = 35$, $\bar{x}_2 = 395000$, $s_2 = 31000$

이므로

$$\left| \frac{452000 - 395000}{\sqrt{22000^2/30 + 31000^2/35}} \right| = 8.633, \quad Z_{1-0.01/2} = Z_{0.995} = 2.575$$

따라서 H_0 는 기각 된다. 즉, 남자와 여자의 초임이 차이가 있다고 할 수 있다. 이 문제의 p-값은 양측검정이므로 표준정규분포에서 표본통계량의 값 8.633보다 큰 확률을 구해 두 배를 해주면 된다. 즉,

p-값 = $2 \times P(Z > 8.633) = 2 \times 0.0000 = 0$

이다. □

표본의 크기가 작은 경우에는, 두 모집단이 정규분포를 따르고 분산이 같다는 가정하에 두 모평균의 가설검정에 다음과 같은 통계량을 사용한다.

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} \qquad \text{여기서} \quad s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

s_p^2 은 모분산의 추정량으로 s_1^2 과 s_2^2 의 표본의 크기에 가중치를 주어 모분산을 추정한 것으로 공통분산(pooled variance)이라 한다. 즉 공통분산은 두 모집단의 분산이 같다고 가정했으므로 두 분산의 표본크기에 비례한 가중평균이다. 위의 통계량은 자유도가 n_1+n_2-2 인 t분포를 하는데 이를 이용하여 소표본인 경우 두 모평균의 차이에 대한 검정을 다음과 같이 할 수 있다.

〈표 8.2〉 두 모평균의 가설검정

- 소표본($n_1<30$ 또는 $n_2<30$)이고, 표본이 서로 독립적으로 추출되었으며, 두 모집단이 정규분포를 따르고, 두 모분산이 같은 경우

가설의 종류	선 택 기 준
1) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$	$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} > t_{n_1+n_2-2,\,1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
2) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 < \mu_2$	$\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2/n_1 + s_p^2/n_2}} < -t_{n_1+n_2-2,\,1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
3) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$	$\left\lvert\dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}\right\rvert > t_{n_1+n_2-2,\,1-\alpha/2}$ 이면 H_0기각, 아니면 H_0채택

[예 8.2] [예 8.1]에서 남자 대졸사원이 표본수가 15명, 여자 대졸사원의 표본수가 14명이고 나머지 표본자료는 똑 같다고 하자. 남자와 여자의 초임이 차이가 있는지 1% 유의수준으로 검정하라. 초임은 정규분포를 따르고 두 모분산은 같다고 가정하라.

〈풀이〉

이 문제의 가설은 H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$ 이므로 선택기준은

$$\left\lvert\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}\right\rvert > t_{n_1+n_2-2,\,1-\alpha/2}$$ 이면 H_0 기각, 아니면 H_0 채택

이다. $n_1=15$, $\bar{x}_1=452000$, $s_1=22000$, $n_2=14$, $\bar{x}_2=395000$, $s_2=31000$ 이므로

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = \frac{(15-1)22000^2 + (14-1)31000^2}{15 + 14 - 2} = 26715^2$$

$$\left\lvert\frac{452000 - 395000}{\sqrt{26715^2/15 + 26715^2/14}}\right\rvert = 5.74$$

$$t_{15+14-2,\,1-0.01/2} = t_{27,\,0.995} = 2.7707$$

따라서 $5.74 > 2.7707$ 이므로, 가설 H_0 는 기각된다. □

만일에 표본의 크기가 작고 두 모집단의 분산이 다를 경우 모집단이 정규분포를 따르더라도 통계량

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

은 t 분포를 따르지 않는다. 두 모집단의 분산이 다른 경우 두 모평균의 가설검정을 Behrens-Fisher 문제라고 한다. 이 문제를 해결하기 위한 여러 가지 방법이 연구되었는데, 이 책에서는 t분포를 이용하여 근사적으로 다음과 같은 t'값을 계산하여 가설검정을 하는 방법을 소개한다.

$$t'_{n_1+n_2-2,\,p} = \frac{(s_1^2/n_1)t_{n_1-1,\,p} + (s_2^2/n_2)t_{n_2-1,\,p}}{s_1^2/n_1 + s_2^2/n_2}$$

이 t'은 각 표본평균의 분산(s_1^2/n_1 과 s_2^2/n_2)을 가중치로 하여 얻어진 t_{n_1-1}과 t_{n_2-1}의 가중평균이다. 이것을 이용하면 소표본이고 두 모분산이 서로 다른 경우 두 모평균의 가설검정을 아래와 같이 할 수 있다.

〈표 8.3〉 두 모평균의 가설검정

- 소표본($n_1<30$ 또는 $n_2<30$)이고, 표본이 서로 독립적으로 추출되었으며, 두 모집단이 정규분포를 따르고, 두 모분산이 다른 경우

가설의 종류	선 택 기 준
i) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} > t'_{n_1+n_2-2,\,1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
i) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 < \mu_2$	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} < -t'_{n_1+n_2-2,\,1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
i) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$	$\left\lvert\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}\right\rvert > t'_{n_1+n_2-2,\,1-\alpha/2}$ 이면 H_0기각, 아니면 H_0채택

[예 8.3] [예 8.2]에서 두 모분산이 서로 다를 경우에 남자와 여자의 초임이 차이가 있는지 1% 유의수준으로 검정하라.

〈풀이〉

이 문제의 가설은 H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$ 이므로 선택기준은

$$\left\lvert\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}\right\rvert > t'_{n_1+n_2-2,\,1-\alpha/2}$$ 이면 H_0 기각, 아니면 H_0 채택

이다. $n_1 = 15$, $\bar{x}_1 = 452000$, $s_1 = 22000$, $n_2 = 14$, $\bar{x}_2 = 395000$, $s_2 = 31000$ 이므로

$$\left| \frac{452000 - 395000}{\sqrt{22000^2/15 + 31000^2/14}} \right| = 5.74$$

$$t'_{n_1+n_2-2,1-\alpha/2} = t'_{15+14-2,1-0.01/2}$$

$$= \frac{(22000^2/15)t_{15-1,0.995} + (31000^2/14)t_{14-1,0.995}}{22000^2/15 + 31000^2/14}$$

$$= 3.001$$

따라서 가설 H_0는 기각된다. □

나. 대응비교

두 모평균을 비교하는 지금까지의 가설검정에서는 두 표본이 서로 독립적으로 추출된 경우를 다루었다. 하지만 어느 경우에는 두 표본을 독립적으로 추출하기가 힘들거나, 독립적으로 추출하였을 때 각 표본개체의 특성이 너무 차이가 나서 결과분석이 무의미할 때가 있다. 예를 들면 타자수에게 타자속도를 증가시키기 위한 특수교육을 시킨 후 과연 이 교육이 타자속도 증가에 효과가 있었는가 알아 보고 싶다고 하자. 이 때 교육전과 교육후에 서로 다른 표본을 추출하면 개인의 차가 심하기 때문에 교육의 효과를 측정하기가 어렵다. 이러한 경우 교육전에 표본추출되어 속도를 측정한 타자수에 대하여, 교육후에 속도를 측정하여 비교하면 특수교육의 효과를 잘 알아 낼 수가 있다. 이렇게 서로 독립적이지 않은, 비슷한 성질의 표본을 사용하여 두 모집단의 평균을 비교하는 가설검정을 대응비교(paired comparison)라고 한다. 대응비교일 때는 먼저 다음과 같이 관찰된 n쌍(pair)의 차(d_i)를 계산해서 평균($\bar{d}$)과 표준편차(s_d)를 구한다.

모집단 1의 표본(X_{i1})	모집단 2의 표본(X_{i2})	$d_i = X_{i1} - X_{i2}$
X_{11}	X_{12}	$d_1 = X_{11} - X_{12}$
X_{21}	X_{22}	$d_2 = X_{21} - X_{22}$
.	.	.
.	.	.
.	.	.
X_{n1}	X_{n2}	$d_n = X_{n1} - X_{n2}$

d_i 의 평균 $\bar{d} = \sum d_i / n$

d_i 의 분산 $s_d^2 = \sum (d_i - \bar{d})^2 / (n-1)$

두 모집단이, 모평균이 같은 정규분포일 때 $\bar{d}/(s_d/\sqrt{n})$ 는 자유도가 (n-1)인 t분포를 따르는데 이를 이용하여 대응비교인 경우 두 모평균의 차이에 대한 검정을 다음과 같이 할 수 있다.

〈표 8.4〉 두 모평균의 가설검정 (대응비교)
- 모집단이 정규분포이고 두 표본이 쌍(종속적)으로 추출되었을 경우

가설의 종류	선 택 기 준
1) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$	$\bar{d}/(s_d/\sqrt{n}) > t_{n-1,1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
2) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 < \mu_2$	$\bar{d}/(s_d/\sqrt{n}) < -t_{n-1,1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
3) H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 \neq \mu_2$	$\lvert \bar{d}/(s_d/\sqrt{n}) \rvert > t_{n-1,1-\alpha/2}$ 이면 H_0 기각, 아니면 H_0 채택

[예 8.4] 한글 타자속도를 빠르게 하기위한 교육을 8명의 타자수에게 실시하여 교육전과 후의 타자속도를 조사하였더니 아래와 같다. 타자교육이 속도를 증가시켰는지 5% 유의수준으로 검정하라. 단, 타자속도는 정규분포라고 가정하자.

타자수 번 호	타자속도(글자/분) 교육전	교육후	차($d_i=X_{i1}-X_{i2}$)
1	52	58	-6
2	60	62	-2
3	63	62	1
4	43	48	-5
5	46	50	-4
6	56	55	1
7	62	68	-6
8	50	57	-7

〈풀이〉

이 문제의 가설은 H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 < \mu_2$ 이므로 가설 선택기준은

$\bar{d}/(s_d/\sqrt{n}) < -t_{n-1,1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택

이다. 위의 자료에서 차 d_i의 평균($\bar{d}$)은 -3.5이고 표준편차(s_d)는 3.16 이므로

$$\bar{d}/(s_d/\sqrt{n}) = -3.5 / (3.16/\sqrt{8}) = -3.91$$
$$-t_{n-1,1-\alpha} = -t_{8-1,1-0.05} = -1.8946$$

따라서 가설 H_0 는 기각되므로 타자교육이 속도를 증가시켰다고 할 수 있다. □

2. 두 모분산의 가설검정

두 모분산의 비교를 하는 아래의 예를 살펴보자.

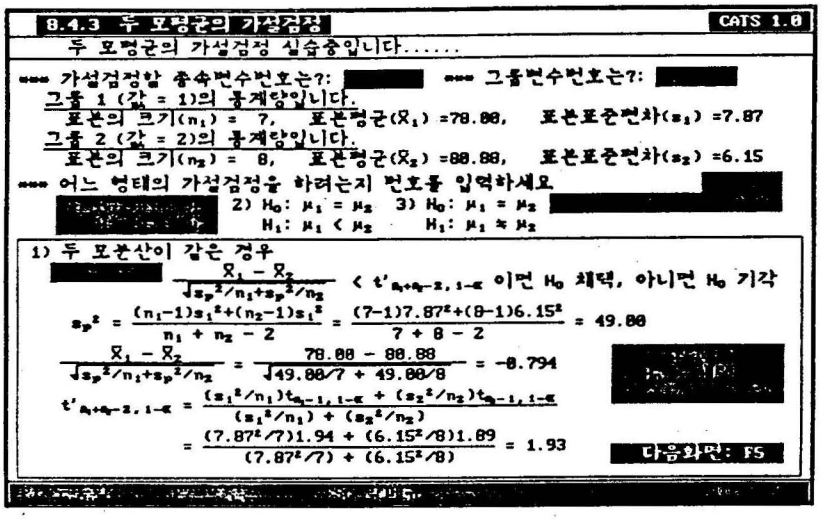
1) 앞 절에서 두 모평균을 비교할 경우 표본의 크기가 작다면 두 모분산이 같은지 다른지에 따라 가설검정의 선택기준이 다른 것을 알았다. 그러면 현실적으로 미지의 두 모분산이 같은지 어떻게 검정할 수 있나?

2) 자동차 조립에 쓰이는 볼트의 품질은 그 직경에 대한 규격을 엄격하게 지키느냐에 달려 있다. 두 회사에서 이 볼트를 납품하는데 직경의 평균은 같다고 한다. 따라서, 분산이 더 작은 제품이 우수하다고 볼 수 있는데 분산에 대한 비교를 어떻게 할 수 있나?

이러한 두 모집단의 분산(σ_1^2 과 σ_2^2)을 비교하는 경우에는 모평균의 비교처럼 분산의 차이($\sigma_1^2 - \sigma_2^2$)를 조사하지 않고 분산의 비(σ_1^2/σ_2^2)를 비교한다. 이 분산비가 1보다 큰가, 작은가, 같은가를 알아보면 σ_1^2 이 σ_2^2 보다 큰가, 작은가, 같은가를 알 수 있다. 분산의 차대신 분산비를 이용하는 이유는 표본분산비에 대한 분포를 수학적으로 찾아낼 수 있기 때문이다. 즉, 통계량

$$(s_1^2/\sigma_1^2) / (s_2^2/\sigma_2^2)$$

은 두 모집단이 각각 정규분포를 따를 경우 분자자유도 n_1-1, 분모자유도 n_2-1 인 F 분포를 따르는데 이 사실을 이용하여 모분산비에 대한 가설검정을 한다. F 분포는 비대칭인 분포군으로서 분모자유도, 분자자유도에 따라 서로 다른 분포를 갖는다. <그림 8.1>은 여러가지 자유도에 따른 F분포의 그림이다.



<그림 8.1> 여러가지 자유도에 따른 F분포의 그림

두 모분산의 가설검정은 F 분포를 이용하여 다음과 같이 할 수 있다.

<표 8.5> 두 모분산의 가설검정
- 두 모집단이 정규분포인 경우

가설의 종류	선 택 기 준
1) $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$(s_1^2 / s_2^2) > F_{n-1, n-1, 1-\alpha}$ 이면 H_0 기각, 아니면 H_0 채택
2) $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$	$(s_1^2 / s_2^2) < F_{n-1, n-1, \alpha}$ 이면 H_0 기각, 아니면 H_0 채택
3) $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$\frac{\text{큰 표본분산}}{\text{작은 표본분산}} > F_{x-1, y-1, 1-\alpha/2}$ 이면 H_0 기각, 아니면 H_0 채택 * 여기서 x 는 큰 표본분산에 해당하는 표본의 크기이고 y 는 작은 표본분산에 해당하는 표본의 크기이다.

[예 8.5] [예 8.2]에서 대졸 남녀사원의 초임의 분산이 같은가 유의수준 5%로서 검정하라.

〈풀이〉

이 문제의 가설은 $H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$ 이다. 따라서 선택기준은

$$\frac{\text{큰 표본분산}}{\text{작은 표본분산}} > F_{x-1.y-1.1-\alpha/2} \text{ 이면 } H_0 \text{ 기각, 아니면 } H_0 \text{ 채택}$$

큰 표본분산이 31000^2 이므로 x=14, 작은 표본분산이 22000^2 이므로 y=15 이다. 따라서

$$(\text{큰표본분산}/\text{작은 표본분산}) = 31000^2/22000^2 = 1.986$$

$$F_{x-1.y-1.1-\alpha/2} = F_{14-1.15-1.1-0.05/2} = F_{13.14.0.975} = 3.03$$

그러므로, 가정 H_0 는 채택이 된다. 즉, 두 모분산은 같다고 볼 수 있다. □

3. 두 모비율의 가설검정

두 모비율을 비교하는 아래의 예를 살펴보자.

1) 금년도 대통령 선거에서 특정후보에 대한 지지율이 남자와 여자 유권자 사이에 차이가 있는가?
2) 어느 공장에서 제품을 만들어 내는 두 대의 기계가 있는데 각 기계의 불량률이 서로 다른가?

이러한 두 모집단의 모비율(p_1 과 p_2)을 비교하는 것은, 모평균과 유사하게 두 모비율의 차 ($p_1 - p_2$)를 가설검정함으로써 가능하다. 두 모집단에서 서로 독립적으로 추출한 표본비율의 차 $\hat{p}_1 - \hat{p}_2$ 는 표본의 크기가 충분히 큰 경우 평균이 $p_1 - p_2$, 분산이 $p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ 인 정규분포를 따른다. 여기서 분산의 추정을 위해서는 p_1과 p_2를 모르므로 두 표본비율($\hat{p}_1$ 과 $\hat{p}_2$)의 표본의 크기를 가중치로 하는 가중평균 $\bar{p}$ 를 사용한다.

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

두 모비율의 차에 대한 검정은 통계량

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})/n_1 + \bar{p}(1-\bar{p})/n_2}}$$

을 이용하여 다음과 같이 한다.

〈표 8.6〉 두 모비율의 가설검정
- 대표본이고, 표본이 서로 독립적으로 추출되었을 경우 -

가설의 종류	선택기준
i) H_0: $p_1 = p_2$ H_1: $p_1 > p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})/n_1 + \bar{p}(1-\bar{p})/n_2}} > Z_{1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
i) H_0: $p_1 = p_2$ H_1: $p_1 < p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})/n_1 + \bar{p}(1-\bar{p})/n_2}} < -Z_{1-\alpha}$ 이면 H_0기각, 아니면 H_0채택
i) H_0: $p_1 = p_2$ H_1: $p_1 \neq p_2$	$\left\lvert\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})/n_1+\bar{p}(1-\bar{p})/n_2}}\right\rvert > Z_{1-\alpha/2}$ 이면 H_0기각, 아니면 H_0채택

[예 8.6] 금년도 대통령 선거에서 특정후보의 지지율을 남녀별로 표본조사하였더니 추출된 표본중 남자 225명중 54명이 지지를 하였고, 여자 175명중 52명이 지지를 하였다. 남녀의 지지율이 같은지 5% 유의수준으로 검정하라.

〈풀이〉

이 문제의 가설은 H_0: $p_1 = p_2$, H_1: $p_1 \neq p_2$ 이므로 선택기준은

$$\left|\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})/n_1+\bar{p}(1-\bar{p})/n_2}}\right| > Z_{1-\alpha/2} \text{ 이면 } H_0\text{기각, 아니면 } H_0\text{채택}$$

$\hat{p}_1 = 54/225 = 0.240$ $\hat{p}_2 = 52/175 = 0.297$ 이므로

$\bar{p} = (54+52) / (225+175) = 106/400 = 0.265$ 이다.

따라서

$$\left|\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})/n_1+\bar{p}(1-\bar{p})/n_2}}\right| = \left|\frac{0.240 - 0.297}{\sqrt{0.265(1-0.265)/225+0.265(1-0.265)/175}}\right|$$

$$= 1.28$$

$Z_{1-\alpha/2} = Z_{1-0.05/2} = Z_{0.975} = 1.96$

그러므로, 가설 H_0는 채택이 된다. 즉, 남녀별 특정후보의 지지율은 같다고 볼 수 있다. □

4. SAS 실습

SAS에서 두 집단의 모평균에 대한 가설검정을 하기 위한 procedure에는 TTEST가 있는데 그 일반형은 다음과 같다.

```
PROC TTEST options;
     CLASS variable;
     VAR variables;
     BY variables;
```

해설: 1) PLOT options;

options ---

① DATA=dsn 처리에 사용될 data set 이름을 정해준다.

② PLOT 줄기-잎 그림, 상자그림, normal probability plot을 그려준다.

③ FREQ 도수분포를 작성

2) CLASS variable;

두 수준의 값을 갖는 그룹변수

3) VAR variables;

측도의 계산을 원하는 변수를 나열한다.

4) BY variables;

BY 변수의 각 값에 대한 MEANS 표가 작성된다.

화일 'ex0201.dat'를 이용하여 월급에 대해 남녀별 모평균에 차이가 있는지 TTEST procedure로 실습한 결과가 <표 8.7>과 같다. 모분산이 같은 경우와 다른 경우에 대해 관측된 t 값(T)과 양측검정에 대한 p-값(Prob>|T|)을 보여준다.

```
PROC TTEST;
   CLASS sex;
   VAR salary;
```

<표 8.7> 남녀별 봉급에 대한 TTEST procedure 출력

Variable: SALARY

SEX	N	Mean	Std Dev	Std Error	Minimum	Maximum
1	27	134.8148148	68.01415975	13.08933115	60.00000000	350.0000000
2	13	130.7692308	77.72403516	21.55676878	50.00000000	300.0000000

Variances	T	DF	Prob>|T|
Unequal	0.1604	21.2	0.8741
Equal	0.1683	38.0	0.8673

For H0: Variances are equal, F' = 1.31 DF = (12,26) Prob>F' = 0.5469

연 습 문 제

8.1.1 종이제조 공장에서는 두 군데의 산림지 중 하나를 사려고 하고 있다. 각 산림지에서 표본으로 추출된 50그루씩의 나무의 직경 측정결과는 아래와 같다. 이 자료들은 유의수준 0.05에서 평균적으로 B지역의 나무들이 A지역의 나무들보다 작다고 하는 충분한 증거가 되겠는가? 이 검정의 p-값은?

A 지역	$\bar{x} = 28.25$	$s^2 = 25$
B 지역	$\bar{x} = 22.50$	$s^2 = 16$

8.1.2 한 분석가는 두가지 형태의 소매상들의 광고 방법에 대해서 연구하고 있다. 변수는 지난 일년 동안 광고에 소요된 액수의 합계이다. 각 형태의 소매상들로부터 독립적으로 추출된 확률표본이 아래와 같다.

(단위 천원)

A형태	$n = 60$	$\bar{x} = 14,800$	$s^2 = 180,000$
B형태	$n = 70$	$\bar{x} = 14,500$	$s^2 = 133,000$

이 자료들로부터 A형태의 소매상들이 B형태의 소매상들보다 광고에 더 많은 투자를 했다고 결론지을 수 있겠는가? (유의수준 = 0.05)

8.1.3 현재의 집에서 거주한 기간에 대하여, A지역에서 100가구 B지역에서 150가구를 표본추출하여 계산한 통계량이 아래와 같다. 이 자료는 A지역의 가구들이 B지역의 가구들보다 평균 거주기간이 짧다고 할 충분한 증거가 되겠는가? (유의수준 = 0.05)

지역 A	$\bar{x}$ = 33개월	s^2 = 900
지역 B	$\bar{x}$ = 49개월	s^2 = 1,050

8.1.4 한 광고 분석가가 라디오, 텔레비젼, 신문, 잡지등의 광고매체를 직장남성과 주부들이 얼마나 접하는지 표본조사를 하였다. 조사한 결과는 각 집단이 특정한 한 주에 접하게 된 광고의 회수인데 그 통계량이 아래의 표와 같다. 이 자료들은 주부들이 직장남성보다 평균적으로 광고를 더 많이 접하게 된다고 할 충분한 증거가 되겠는가?

집 단	n	접하는 광고의 평균 수	표준편차
작장 남성	100	200	50
무직 여성	144	225	60

8.1.5 전선 공장에서는 단위 길이당 저항의 측면에서 두가지 형태의 전선을 비교하려고 한다. 전선 1의 30개 표본과 전선 2의 35개 표본은 아래와 같은 결과를 내었다. (단위: ohms x 10^2)

전선 1									
55.2	53.5	52.3	54.1	52.4	50.5	53.5	46.9	52.9	57.1
55.7	51.2	55.2	57.4	53.9	58.1	50.6	59.4	51.8	50.8
56.9	56.3	59.1	52.7	56.1	58.2	53 1	50.6	53.1	59.7
전선 2									
46.9	50.6	47.3	48.0	49.2	48.4	48.5	48.6	48.2	50.2
47.2	50.3	49.1	48.2	47.4	48.1	49.4	47.4	49.7	49.1
49.3	50.3	50.8	48.3	47.7	48.5	51.1	50.9	49.5	49.7
51.4	48.1	49.7	50.9	48.6					

이 결과들을 근거로 두 표본집단의 평균저항이 다르다고 말할 수 있겠는가? (유의수준 = 0.05로 놓아라.)

8.1.6 한 공장에서는 두 개 회사의 모터 오일의 점착성을 비교하려고 한다. 각 회사의 제품 중 확률적으로 추출된 32개의 제품들이 검사받은 결과가 아래와 같다. 이 자료들을 근거로 하여 두 회사제품의 점착성 평균이 다르다고 결론지을 수 있겠는가? (유의수준 = 0.05로 놓아라.)

회사 A	13	21	60	35	38	10	36	24	35	35
	45	19	42	11	35	39	25	17	51	25
	52	25	11	11	55	44	25	41	16	47
	50	18								
회사 B	46	52	66	65	71	67	47	48	58	42
	66	69	60	80	45	47	69	75	43	46
	74	73	43	70	51	72	65	45	76	48
	56	64								

8.1.7 한 직물 제조업자는 두 판매회사 중 한 곳에서 어떤 종류의 실을 사야한다. 두 판매회사의 생산은 가격과 강도를 제외한 모든 점에서 비슷한 것이 명백하다. 제조업자는 판매회사 1의 생산품이 판매회사 2보다 평균적으로 낮은 강도를 갖는다고 할 만한 충분한 이유가 없다면 가격이 더 싼 판매회사 1로부터 실을 구입할 것이다. 두 회사의 생산품으로부터 표본을 추출한 결과가 아래와 같다. 강도가 근사적으로 정규분포를 따른다고 할 때

1) 유의수준 0.05에서 적당한 가설검정에 기초할 때 제조업자에게 더 싼 실을 구입하라고 충고하겠는가? 모분산이 같다고 가정하라.
2) 모분산이 같지 않다고 가정할 때 (1)를 반복하라.

판매사 1	$n = 10$	$\bar{x} = 94$	$s^2 = 14$
판매사 2	$n = 12$	$\bar{x} = 98$	$s^2 = 9$

8.1.8 다음 자료는 어떤 공장의 2교대 조로부터 얻은 표본의 결과이다. 변수는 어떤 일에 필요한 시간의 길이이다. 교대 2조의 평균 시간이 교대 1조보다 작다고 할 수 있겠는가 ? 유의수준은 0.05이다. 이 검정이 타당하기 위한 모든 가정을 명시하라.

교대 1	$n = 10$	$\bar{x} = 26.1$	$s^2 = 144$
교대 2	$n = 8$	$\bar{x} = 17.6$	$s^2 = 110$

8.1.9 한 수면제 제조업자는 새로운 처방 B와 지금 시중에서 유통되는 처방 A의 효과를 비교하고 있다. 사흘밤동안 25명에게 처방 B를 실험하고 독립적으로 25명에게 처방 A를 실험하였다. 변수는 약을 먹지 않았을 때와 비교하여 늘어난 평균 수면의 시간이다. 결과가 다음과 같을 때 처방 B가 처방 A보다 효과적이라고 할 수 있겠는가 ? 유의수준=0.05이다.

약	A	B
평균	1.4	1.9
표본분산	0.09	0.16

8.1.10 한 산업심리학자는 근로자들이 직업을 바꾸는 큰 요소가 근로자 개인의 일에 대한 자긍심이라고 생각한다. 그 학자는 직업을 자주 바꾼 근로자(그룹 A)들이 그렇지 않은 근로자(그룹 B)들보다 낮은 자긍심을 갖고 있다고 생각한다. 각 그룹에 대하여 표본을 독립적으로 추출해 자긍심의 점수를 측정한 자료가 다음과 같다.

그룹 A	60	45	42	62	68	54	52	55	44	41							
그룹 B	70	72	74	74	76	91	71	78	76	78	83	50	52	66	65	53	52

이 자료는 심리학자의 생각을 뒷받침해 줄 수 있겠는가? 모집단의 점수는 정규분포를 따르고, 모분산은 알려지지 않았지만 같다고 가정하자. 유의수준이 0.01

8.1.11 한 대학의 경영학과에서, 남자들이 여자보다 주식시장에 대한 지식이 많다는 몇사람의 주장에 대해 논쟁이 일어났다. 논쟁을 가라앉히기 위해 지도강사는 각 15명의 남녀를 독립적으로 표본 추출하여 주식시장에 대한 지식 측정검사를 하였다. 결과는 다음과 같다.

여자	73	96	74	55	91	50	46	82	43	79	79	50	46	81	83
남자	57	78	42	44	91	65	63	60	97	85	92	42	86	81	64

이 자료에 의하면 평균적으로 남자가 여자보다 주식시장에 대한 더 많은 지식을 갖고 있다고 말 할 수 있는가? 유의수준은 0.05. 어떤 가정이 필요한가?

8.1.12 한 석유 회사가 총 가솔린의 연비를 향상시키리라 믿는 가솔린 첨가제를 개발했다. 계획된 마케팅 프로그램을 지지할 정보를 얻기 위해서 16쌍의 자동차

를 이용하여 대응비교를 다룰 검사 기관을 고용하였다. 각 쌍은 구조, 모델, 엔진의 크기, 그리고 다른 관계특성까지의 세목들이 동일하다. 각 쌍의 한 자동차는 임의로 추출되어지고 첨가제를 넣은 가솔린을 사용하여 시험코스를 운전한다. 다른 한 차는 첨가제를 넣지 않은 가솔린을 사용하여 같은 코스를 운전한다. 모든 시험 운전이 끝나고 시험 코스에서 일 리터당 Km수가 다음과 같게 나타났다. 이 자료는 첨가제가 연비를 증가시킨다고 말할 수 있는 근거가 되나? 연비는 대략 정규분포를 한다고 가정하자.

쌍	첨가제 넣음(X1)	첨가제 안넣음(X2)	쌍	첨가제 넣음(X1)	첨가제 안넣음(X2)
1	17.1	16.3	9	10.8	10.1
2	12.7	11.6	10	14.9	13.7
3	11.6	11.2	11	19.7	18.3
4	15.8	14.9	12	11.4	11.0
5	14.0	12.8	13	11.4	10.5
6	17.8	17.1	14	9.3	8.7
7	14.7	13.4	15	19.0	17.9
8	16.3	15.4	16	10.1	9.4

8.1.13 한 연구는 다양한 장소에서 어떻게 효과적으로 거리조명등을 위치하게 함으로써 한 마을안의 자동차 사고를 줄일 수 있겠는가 하는 조사를 다루고 있다. 다음 표는 12 곳에 조명을 달기 일년 전과 일년 후의 매주 밤시간의 평균사고 수이다. 이 자료는 조명이 밤시간의 자동차 사고를 줄였다고 할 수 있는 근거를 제공하는가?

위 치	A	B	C	D	E	F	G	H	I	J	K	L
전의 사고수	8	12	5	4	6	3	4	3	2	6	6	9
후의 사고수	5	3	2	1	4	2	2	4	3	5	4	3

8.2.1 한 사람이 두 평균의 차이를 검정하고자 t 검정의 사용을 고려중이다. 두 표본의 크기가 각각 16이고 분산이 28.5, 9.5이다. 이 자료는 모분산이 같다는 가정하에서 t 검정의 사용이 적당치 않다고 할 수 있는가? 유의수준은 0.05이다.

8.2.2 스트레스성 직업에 종사하는 고용인들 사이에 긴장을 풀어주는 두 약품

을 비교하기 위한 어떤 연구가 계획되었다. 한 의료팀이 처리 첫 두달의 마지막 날에 두 처리 집단에서 실험 대상자의 긴장 정도에 대한 자료를 수집하여 표본으로부터 분산이 $s_1^2 = 2916$ $s_2^2 = 4624$를 얻었다. 각 집단의 크기는 8이다. 유의수준이 0.05라 할 때 이 자료는 두 모집단의 긴장 정도의 변화에 차이가 있다고 할 수 있는가 ? 필요한 가정을 말하라.

8.3.1 한 용단 제조업체는 화씨 250도 이상의 온도에서 견딜 수 있는 재료를 찾고 있는 중이다. 두가지의 물질이 있는데 하나는 천연물질이고 다른 하나는 값이 싼 인공물질인데 이 둘은 열에 견디는 정도를 제외하고는 같은 만족도를 갖고 있다. 두 물질에서 각각 250개의 표본을 독립적으로 추출하여 이 특성에 대한 실험을 했다. 천연물질에서 36개, 인공물질에서 45개의 표본이 화씨 250도이하의 온도에서 실패하였는데 이 자료로 부터 두 물질이 열에 견디는 관점에서 차이가 있다고 할 수 있겠는가? 유의수준은 0.05로 하라.

8.3.2 어느 회사의 노동조합은 대학교육을 받지 못한 150명의 판매사원중 63%가 대학교육을 지금이라도 다시 받고 싶어하는 것을 알아냈다. 그 회사는 10년 전에도 유사한 연구를 했었는데 그 때는 160명중 58%만이 원했었다. 유의수준을 0.05로 할 때 대학교육을 원하는 정도가 10년전과 다르지 않다는 귀무가설을 검정하라. 표본은 서로 독립적으로 추출되었다.

8.3.3 A유형의 회사 200개를 표본추출하여 조사하니, 전 판매의 1% 이상을 광고에 쓰는 회사가 그 중 12%에 해당됨을 알았다. B유형의 회사들에서 같은 크기의 표본을 독립적으로 추출하니 전 판매의 1% 이상을 광고에 쓰는 회사들이 15%이었다. 유의수준을 0.05로 놓고 다음을 검정하라.

$$H_0 : p_B \leq p_A \,,\; H_1 : p_B > p_A$$

8.3.4 한 회사의 판매사원과 비판매사원에 대한 여가활동에 대한 연구를 하였다. 판매사원과 비판매사원중 각각 독립적으로 400명을 추출하여 조사하니, 288명의 판매사원과 260명의 비판매사원들이 그들의 여가시간을 주로 스포츠활동에 보낸다고 한다. 이 자료에서 두 집단이 여가시간을 스포츠활동에 보내는 비율이 같다고 할 수 있는가? 유의수준은 0.05이다.

부록: 통계분포표

1. 표준정규분포표
2. t 분포의 p 백분위수표
3. 카이자승(χ^2)분포의 p 백분위수표
4. F 분포의 p 백분위수표
5. 난수표

1. 표준정규분포표

$P(-\infty < Z < z)$

z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	0.00
-3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005
-3.2	0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007
-3.1	0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010
-3.0	0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013
-2.9	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019
-2.8	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026
-2.7	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035
-2.6	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047
-2.5	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062
-2.4	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082
-2.3	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107
-2.2	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139
-2.1	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179
-2.0	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228
-1.9	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287
-1.8	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359
-1.7	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446
-1.6	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548
-1.5	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668
-1.4	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808
-1.3	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968
-1.2	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151
-1.1	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357
-1.0	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587
-0.9	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841
-0.8	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090	0.2119
-0.7	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389	0.2420
-0.6	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709	0.2743
-0.5	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050	0.3085
-0.4	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409	0.3446
-0.3	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783	0.3821
-0.2	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168	0.4207
-0.1	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562	0.4602
-0.0	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960	0.5000

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

2. t 분포의 p 백분위수표

	p					
	0.75	0.90	0.95	0.975	0.99	0.995
자유도						
df 1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.637	2.347	3.159	4.454	5.646
4	0.741	1.533	2.130	2.769	3.721	4.547
5	0.727	1.476	2.014	2.568	3.355	4.010
6	0.718	1.440	1.943	2.446	3.138	3.697
7	0.711	1.415	1.894	2.364	2.995	3.494
8	0.706	1.397	1.859	2.306	2.895	3.352
9	0.703	1.383	1.833	2.262	2.821	3.248
10	0.700	1.372	1.812	2.228	2.763	3.168
11	0.697	1.363	1.796	2.201	2.718	3.105
12	0.695	1.356	1.782	2.179	2.681	3.054
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.946
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
limit	0.674	1.282	1.645	1.960	2.326	2.576

3. 카이자승(χ^2)분포의 p 백분위수표

	p							
	0.005	0.025	0.05	0.9	0.95	0.975	0.99	0.995
df 1	0.00003	0.00098	0.00393	2.706	3.841	5.024	6.635	7.879
2	0.0100	0.0506	0.103	4.605	5.991	7.378	9.210	10.597
3	0.0717	0.216	0.352	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	0.711	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.145	9.236	11.070	12.833	15.086	16.750
6	0.676	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	2.733	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	3.325	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	3.940	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	4.575	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	5.226	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	5.892	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	6.571	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	7.261	22.307	24.996	27.488	30.578	32.801
16	5.142	6.908	7.962	23.542	26.296	28.845	32.000	34.267
17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401
22	8.643	10.982	12.338	30.813	33.924	36.781	40.289	42.796
23	9.260	11.689	13.091	32.007	35.172	38.076	41.638	44.181
24	9.886	12.401	13.848	33.196	36.415	39.364	42.980	45.559
25	10.520	13.120	14.611	34.382	37.652	40.646	44.314	46.928
26	11.160	13.844	15.379	35.563	38.885	41.923	45.642	48.290
27	11.808	14.573	16.151	36.741	40.113	43.195	46.963	49.645
28	12.461	15.308	16.928	37.916	41.337	44.461	48.278	50.993
29	13.121	16.047	17.708	39.087	42.557	45.722	49.588	52.336
30	13.787	16.791	18.493	40.256	43.773	46.979	50.892	53.672
35	17.192	20.569	22.465	46.059	49.802	53.203	57.342	60.275
40	20.707	24.433	26.509	51.805	55.758	59.342	63.691	66.766
45	24.311	28.366	30.612	57.505	61.656	65.410	69.957	73.166
50	27.991	32.357	34.764	63.167	67.505	71.420	76.154	79.490
60	35.534	40.482	43.188	74.397	79.082	83.298	88.379	91.952
70	43.275	48.758	51.739	85.527	90.531	95.023	100.425	104.215
80	51.172	57.153	60.391	96.578	101.879	106.629	112.329	116.321
90	59.196	65.647	69.126	107.565	113.145	118.136	124.116	128.299
100	67.328	74.222	77.929	118.498	124.342	129.561	135.807	140.169

4. F분포의 p 백분위수표

p = 0.95

	분자 df								
	1	2	3	4	5	6	7	8	9
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93

p = 0.95

	분자 df									
	10	12	15	20	24	30	40	60	120	200
1	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	253.68
2	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.49
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.54
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.65
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.39
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.69
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.25
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.95
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.73
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.56
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.43
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.32
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.23
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.16
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.10
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.04
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.99
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.95
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.91
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.88
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.84
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.82
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.79
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.77
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.75
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.73
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.71
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.69
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.67
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.66
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.55
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.44
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.32
200	1.88	1.80	1.72	1.62	1.57	1.52	1.46	1.39	1.30	1.26

5. 난수표

		열									
		12345	67890	1 12345	67890	2 12345	67890	3 12345	67890	4 12345	67890
행	01	41061	06462	50931	92621	81915	20895	33925	94062	12002	67659
	02	60838	83273	53038	18042	04516	17562	91464	31690	98696	96376
	03	81155	64589	19627	28271	21929	81536	27757	65190	15141	63172
	04	67477	62625	94544	97255	05702	33550	71454	51346	17405	00778
	05	71496	89506	21862	48619	60194	99177	20767	27125	05676	69068
	06	28042	67711	28387	79997	09614	75357	29592	24664	86804	67350
	07	27205	72456	45979	62214	72725	47409	65222	42179	19555	06509
	08	83476	96770	92824	51706	46173	79393	16136	28240	40273	44724
	09	44883	99390	31386	25418	77886	69148	41159	01688	19665	77392
	10	70433	28739	88040	20363	81520	33471	56061	33187	11550	38645
	11	93491	68777	65704	76302	98538	09850	33922	78059	66464	00488
	12	18697	77720	76963	88576	11672	51522	19346	43860	79790	94486
	13	41660	90504	95917	44779	34205	27343	81723	07717	68855	46933
	14	42587	93698	39612	69925	85275	83327	61038	01146	89341	96466
	15	73626	95256	82855	81391	89718	21931	60765	08502	62313	91868
	16	33652	23129	93169	58865	30751	88744	21341	66923	39061	69423
	17	22563	98608	02855	95476	27424	25113	46059	13876	05877	17883
	18	74582	65845	49222	41336	50412	64635	84492	32837	47144	08382
	19	98250	55988	97272	07072	90937	98489	58976	87289	86702	88319
	20	26388	13141	46025	81323	55792	17962	04046	37117	01243	43133
	21	03222	28804	97645	37276	90469	82296	20414	82964	12547	07368
	22	62137	15874	68933	27956	18795	82459	28323	05010	83014	17455
	23	16234	65292	67482	57111	46930	39806	49921	46525	06983	20562
	24	52928	68899	78625	27429	85147	13224	76966	15161	75295	44272
	25	09301	81907	44992	41539	98827	13821	88747	00747	23536	95676
	26	85211	02896	15424	49824	68430	33886	88349	33169	81143	86452
	27	49585	06990	53532	64820	02487	48734	14087	76536	36393	19034
	28	47103	04073	95249	39759	46037	83551	62706	39086	88738	98311
	29	40702	96885	37758	51886	71448	45677	23705	55169	11738	38372
	30	98748	25574	34126	48795	75926	27466	31126	49776	08617	94444
	31	49886	54171	52268	05494	39271	48074	10753	88472	88036	44567
	32	44831	95143	26031	24887	86463	58848	57563	40299	03465	89048
	33	11401	15528	96406	04414	76189	08753	71062	33484	74176	37530
	34	11737	84028	11126	01676	55840	00815	92400	27926	54235	52014
	35	46361	57232	70634	27083	81650	44313	68519	97054	24712	38752
	36	00648	50227	33331	06596	24640	74105	75815	72254	75508	11211
	37	85724	01516	52064	12752	35480	55806	07250	66933	22514	79698
	38	21195	56967	66375	48666	26444	46124	50767	35179	36501	28161
	39	44383	91950	55413	02795	36062	73456	09119	75633	83354	40844
	40	09217	40294	70403	11478	70597	64233	48396	60896	04546	92358
	41	49111	21695	96266	27045	87624	43068	22403	89640	67511	29117
	42	20065	93440	05984	72956	91464	61687	64053	23069	67133	27133
	43	65544	58146	76292	45532	21763	25705	26488	39166	66707	12738
	44	31392	85988	42838	87067	52391	13271	01035	85790	56273	75090
	45	78857	39376	74091	50315	44076	00450	16194	60223	09141	52546

회 귀 분 석
(선형회귀)

목 차

<선형회귀>

선형회귀

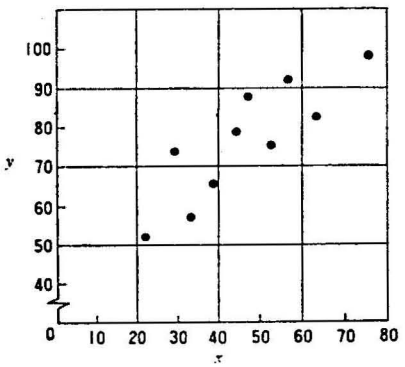
대학에 입학하지 않은 고등학교 3학년 학생의 대학 1학년 학기말 성적을 예측하려고 한다면 상당히 흥미있는 일일 것이며, 광고비용을 얼마나 투입하면 판매액이 얼마나 될 것인가 하는 것을 예측하는 문제도 기업주에게는 매우 중요한 일일 것이다. 이와 같이 두 개 이상의 변수들 간의 관계를 규명하여 분석하고자 하는 방법을 회귀분석(regression analysis)이라 하며, 변수들 간에는 다른 변수에 영향을 주는 변수도 있을 것이고 영향을 받는 변수도 있을 것이다. 회귀분석에서는 다른 변수들에 의하여 영향을 받는 변수를 종속변수(dependent variable) 또는 반응변수(response variable), 영향을 주는 변수를 독립변수(independent variable),또는 설명변수(explanatory variable)라 한다. 위에 예에서 대학의 성적이나 판매액은 종속변수가 될 것이고, 학력고사성적이나 광고비용은 독립변수가 될 것이다. 일반적으로 회귀분석은 한 개의 독립변수와 한 개의 종속변수 간의 관계만을 분석하는 것이 아니고 여러개의 변수들 간의 함수관계를 규명하는데 많이 쓰이고 있다. 이 장에서는 독립변수를 이용하여 종속변수를 예측하는 문제들을 주로 다루게 될 것이며, 종속변수를 y 독립변수를 x로 표시하겠다.

1. 단순회귀모형

먼저 대학입학 학력고사 수학성적과 대학에 입학하여 1학년 말의 평균성적의 관계를 살펴보기로 하자. <표 1>은 대학에 입학한 신입생의 모집단으로부터 표본으로 임의로 10명을 선출하여 그들의 학력고사 성적과 1학년말 평균성적을 나타낸 표이다. 두 변수들 간의 관계를 규명하기 위하여 우리가 제일 먼저 할것은 학생의 학년말 평균성적을 y, 학력고사 성적을 x로 놓고 그래프로 도시해 보는 것이다. [그림 1]은 이 관계를 도시한 그림이다. 이 그림으로 부터 우리는 x가 증가함에 따라 y도 증가함을 알 수 있으며 그 관계는 직선적이라는 사실도 알 수 있다.

학생	수학성적(x)	1학년말 평균성적(y)
1	39	65
2	43	78
3	21	52
4	64	82
5	57	92
6	47	89
7	28	73
8	75	98
9	34	56
10	52	75

<표 1> 수학성적과 1학년 말의 평균성적



[그림 1] <표 1>의 점산도

그러나 모든 점들이 직선상에 존재하는 것이 아니어서 y와 x의 관계식을 정확히 찾아내기 어려우므로 몇 가지 가정을 전제조건으로 하여 x로 부터 y를 예측할수 있는 함수식을 얻을 수가 있으며, 종속변수에 영향을 주는 독립변수가 1개 있을때의 회귀분석을 단순회귀분석(simple regression analysis)이라 하며, 이 때 변수들 간의 관계를 다음과 같은 함수식으로 표현하여 이 모형을 단순회귀모형(simple linear regression model)이라 한다.

단순회귀모형 :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

여기서 $y_i = i$ 번째 측정된 y의 값

β_0 , β_1 = 모집단의 회귀계수로서 β_0는 절편 β_1 은 기울기이다.

x_i = i번째 주어진 고정된 독립변수의 값

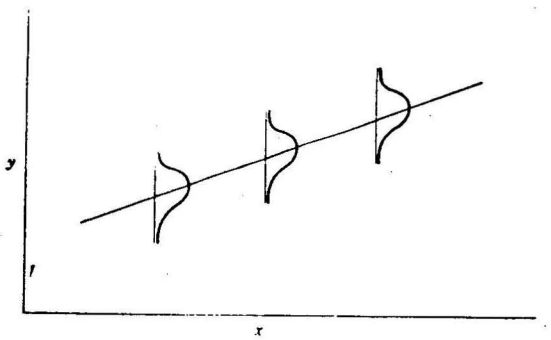
ε_i = i번째 측정된 y의 오차항으로 확률분포는 $N(0, \sigma^2)$이며 다른 오차항과는 서로 독립이다.

즉 $cov\ (\varepsilon_i, \varepsilon_j) = 0$ (단, $i \neq j$)

이 모형은 x와 y의 직선관계를 나타내지만 주어진 x_i 에서 측정치 y_i는 오차 ε_i 때문에 확률적으로 변함을 나타낸다. 즉 y_i는

$$y_i \sim N(\beta_0 + \beta_1 x_i\ , \sigma^2)$$

인 분포를 하며 모든 x_i에서 y_i의 분산은 ε_i의 분산과 동일하다. 이 관계를 그림으로 표현하면 [그림 2]와 같이 된다.



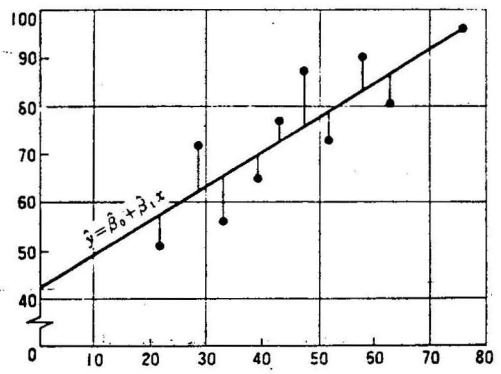
[그림 2] 선형회귀모형

2. 최소제곱법

단순회귀모형에서 회귀계수 β_0 와 β_1은 미지수이며 ε_i는 확률변수이므로 관측된 y_i들을 가장 잘 나타낼 수 있는 선형식(best fitting straight line)을 구하는 것이 필요할 것이다. 이 식을 구하기 위하여 y의 예측된 값을 $\hat{y}$이라 놓고 β_0와 β_1의 추정값을 각각 $\hat{\beta}_0$와 $\hat{\beta}_1$라 하였을경우 예측방정식은 다음과 같이 주어진다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

[그림 3]은 <표 1>의 자료들에 대한 예측선을 표현한 그림이다.



[그림 3] 선형예측모형

예측선으로부터 세로로 표시된 선은 y의 예측치와 관찰치의 差 즉, 잔차(residual)를 말한다. 따라서 i번째 점의 잔차는

$$y_i - \hat{y_i} \quad (\text{단}, \hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i)$$

이다.

그런데 최적회귀선(best fitting regression line)을 구하는 방법으로서 잔차제곱합

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

을 최소로 하는 $\hat{\beta}_0$와 $\hat{\beta}_1$를 구하는 방법으로 최소제곱법(least square method)이 있다.

다시 말해 잔차제곱합 $SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$ 을 최소로 하는 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 을 적합된 회귀직선의 절편과 기울기로 택하는 방법이다. 최소제곱법을 좀더 구체적으로 설명하면 다음과 같다.

잔차제곱합(SSE)에 $\hat{y_i}$의 값을 대입하면

$$SSE = \sum_{i=1}^{n} [\ y_i - (\ \hat{\beta}_0 + \hat{\beta}_1 x_i)]^{\ 2}$$

이므로 SSE를 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 에 대하여 편미분하여 0으로 놓으면 SSE의 값을 최소화 시키는 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 을 구할 수 있다.

여기에서 얻어진 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 을 최소제곱추정량(least square estimator)이라 한다. 최소제곱추정량 $\hat{\beta}_0$ 와 $\hat{\beta}_1$ 을 이용하여 최소제곱 예측방정식(least squares prediction equation)또는 추정회귀직선

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

을 얻을 수 있다.

최소제곱추정량과 추정회귀직선 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x_i})(y - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2}$$

$$\hat{\beta}_o = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

---

3. 오차분산 σ^2의 추정

앞에서 가정한 회귀모형

$$y = \beta_0 + \beta_1 x + \varepsilon$$

에서 오차항 ε은 평균이 0이고 분산이 σ^2인 확률변수이며 또한 관측값 y는 평

균이 $\beta_0+\beta_1 x$ 이고 분산이 σ^2 인 정규확률변수이다. 추정회귀직선 $\hat{y_i}$는 y_i의 추정량이기 보다 y의 평균의 추정량으로 보아야 한다. 오차 $\epsilon = y - (\beta_0+\beta_1 x)$로서 오차의 추정량은

$$\hat{\epsilon} = y - \hat{y}$$
$$= y - (\hat{\beta}_0 + \hat{\beta}_1 x)$$

로 될것이며 이를 잔차(residual)라 한다.
잔차들은 회귀모형을 분석하는데 긴요하게 이용되는데 회귀모형의 가정에서 관측값들의 분산을 σ^2로 하였으나 일반적으로 σ^2는 모르는 경우가 대부분이므로 σ^2의 추정치를 구하여야 하는데 이때 잔차제곱합(residual sum of squares)을 $SSE = \sum_{i=1}^{n}(y_i-\hat{y_i})^2$ 로 표현하면

분산 σ^2의 추정량 :

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{1}{n-2}\sum(y_i - \hat{y_i})^2$$

* MSE는 평균제곱오차(mean squared error)라 한다.

---

4. 회귀직선의 기울기 β_1에 대한 추론

y와 x의 관계를 연구하는 데 있어서 먼저 고려하여야 할 점은 두 변수 사이에 함수관계가 존재하는지에 대한 문제로서 x가 y의 예측에 대한 정보를 어느 정도 제공할 것인가를 구하는 문제이다.
즉 x가 주어진 자료로부터 증가 또는 감소함에 따라서 선형적으로 y도 증가 또는 감소한다는 정보를 얻는 것은 중요한 의미를 갖는다. 따라서 만약 선형관계가 확실히 성립된다면 x가 한 단위 변화하는데 따라 y가 어느정도 변하는가를 관찰하는 문제, 즉 기울기 β_1에 관하여 분석해 볼 필요가 있는 것이므로 이 절에서는 β_1에 관하여 좀더 살펴보기로 하겠다.

x가 증가(감소)함에 따라 y도 증가(감소)한다는 말은 $\beta_1 \neq 0$ ($\beta_1 > 0$ 또는 $\beta_1 < 0$) 라는 의미와 동일하므로 귀무가설 $\beta_1 = 0$ 에 대한 검정을 할 필요가 있을 것이다.

그런데 β_1에 대하여 검정을 하기 위해서는 β_1의 추정량 $\hat{\beta}_1$에 대하여 먼저 살펴보아야 하므로 모집단에서 얻어진 n개의 표본으로부터 얻어지는 추정량 $\hat{\beta}_1$에 대한 분포를 살펴보자.

만약 오차 ε가 정규분포를 한다면 추정량 $\hat{\beta}_0$와 $\hat{\beta}_1$가 y의 함수이고 정규분포를 할 것 이므로 $\hat{\beta}_1$의 기대값과 분산을 구하면 다음과 같이 얻어진다.

$$E(\hat{\beta}_1) = E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-y)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \beta_1$$

또한

$$Var(\hat{\beta}_1) = Var\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \frac{\sigma^2}{SS_x}, \qquad SS_x = \sum_{i=1}^{n}(x_i-\bar{x})^2.$$

따라서 β_1의 불편추정량은 $\hat{\beta}_1$임을 알수 있고 일반적으로 분산 σ^2을 모르는 경우가 대부분이므로 σ^2의 추정량 s^2를 이용하면 통계량

$$t = \frac{\hat{\beta}_1-\beta_1}{s/\sqrt{SS_x}}$$

는 자유도 $n-2$인 t-분포를 따른다.

회귀직선의 기울기에 대한 가설검정 :

귀무가설 $H_0 : \beta_1 = \beta_{10}$

대립가설 $H_1 : \beta_1 \neq \beta_{10}$

검정통계량 : $t = \frac{\hat{\beta}_1-\beta_{10}}{s}\sqrt{SS_x}$

예제

<표 1>의 자료를 이용하여 고등학생의 학력고사 성적 x와 1학년말 평균성적 y 사이에 선형관계가 존재하는지를 조사해 보아라.

<풀이> 기울기 β_1에 대하여 다음과 같은 가설을 검정하면 된다.

귀무가설 $H_0 : \beta_1 = 0$

대립가설 $H_1 : \beta_1 \neq 0$

에 대하여 검정통계량 t는

$$t = \frac{\hat{\beta}_1 - 0}{s}\sqrt{SS_x}$$

$$= \frac{0.765562}{8.70}\sqrt{2474}$$

$$= 4.377$$

그런데 자유도가 8이고 α=0.05인 경우 t분포표로 부터 $t_{0.05} = 2.306$ 이므로 검정통계량 t의 값이 $t > 2.306$ 혹은 $t < -2.306$인경우에 H_0를 기각한다.

따라서 $t = 4.377 > 2.306$ 이므로 귀무가설 $\beta_1 = 0$를 기각하고, 학력고사 수학성적과 학년말성적은 선형관계가 있다고 결론을 내릴 수가 있다.

기울기 β_1에 대한 신뢰구간 :

β_1의 신뢰구간은 다음과 같다.

$\hat{\beta}_1$의 $100(1-\alpha)\%$의 신뢰구간 :

$$\hat{\beta}_1 \pm \frac{t_{\alpha/2}s}{\sqrt{SS_x}}$$

예제

<표 1>의 자료를 이용하여 β_1의 95% 신뢰구간을 설정하여라.

<풀이> β_1의 95% 신뢰구간은 앞의 식에 의하여 자유도가 8이므로

$$\beta_1 \pm \frac{t_{0.05}s}{\sqrt{SS_x}} = 0.77 \pm \frac{(2.306)(8.70)}{\sqrt{2474}}$$

즉, 0.77 ± 0.40 이다.

5. 주어진 x 값에 대한 y 의 기대값 추정

만약 기업의 이익(y)이 광고지출비(x)에 선형적인 관계가 있다면 그 기업은 지출된 광고비에 대하여 평균이익을 추정하는 문제에 관심이 있을 것이며, 어떤 특정한 약의 투약(x)이 인체의 반응(y)에 어떻게 나타날 것인가 하는 문제는 의사에게 상당한 관심을 줄 것이다. 또한 학력고사 수학성적이 50점(x)인 학생의 대학 1학년말 성적(y)은 평균 얼마나 될까 하는 것도 관심의 대상이 될 것이다.

이와 같이 주어진 x값에 대한 y의 평균값 $E(y|x)$을 추정하는 문제를 생각해 보는 것은 중요한 일이다.

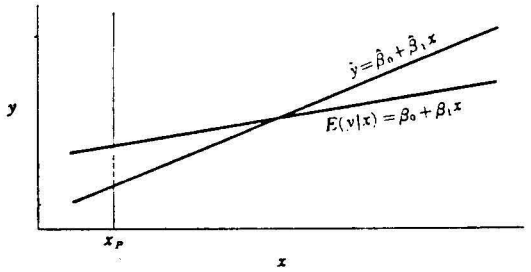
1절에서 정의한 확률모형에 따라 x와 y가 선형적인 관계를 갖는 다면 주어진 x에 대한 y의 평균치는 다음과 같이 나타낼 수 있다.

$$E(y|x) = \beta_0 + \beta_1 x$$

그런데 적합된 회귀직선이

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

이므로 $\hat{y}$를 x의 특정한 값에서 y의 기대값 $E(y|x)$를 추정하는 데 사용할 수 있을 것이다.



[그림 4] y의 값에 대한 기대값과 예측값

[그림 4]에 도시된 두 직선을 살펴보자. 한 직선은 x와 y의 실제 관계를 나타내는

$$E(y|x) = \beta_0 + \beta_1 x$$

의 식이고 다른 하나의 직선은 표본으로부터 적합된 예측직선(fitted prediction line)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

이다. 이 그림에서 $x = x_p$ 일때 y의 기대값에 대한 추정오차는 x_p 위의 점선 부분으로서 두 직선의 차에 해당되며, 이 두 직선 간의 차는 x가 측정된 구간의 극점으로 갈수록 커짐을 알 수 있다.

그런데 예측치 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 는 $\hat{\beta}_0$와 $\hat{\beta}_1$의 함수 이므로 정규분포를 하는 확률변수이며 그 평균과 분산은 다음과 같다.

$$E(\hat{y}) = \beta_0 + \beta_1 x$$

이고

$$\sigma^2_{\hat{y}} = \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}\right]$$

오차분산 σ^2를 모르는 경우 $\hat{y}$의 추정분산을 앞에서와 같이 σ^2대신에 s^2을 사용할 수 있으므로 앞에서와 같이 주어진 x값, 즉 x_p 에서 y의 평균치에 대한 가설검정을 할 수 있을 것이다. 즉,

귀무가설 $H_0: E(y|x = x_p) = E_0$

(단, E_0는 $x = x_p$ 일때 $E(y)$의 가정된 값)

에 대하여 검정통계량은

$$t = \frac{\hat{y} - E_0}{\sigma_{\hat{y}}\text{의추정값}} = \frac{\hat{y} - E_0}{s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}}}$$

이며 t는 자유도 $n-2$로서 t-분포를 따르므로 다음과 같이 검정할 수 있다.

$E(y|x)$의 검정 :

귀무가설 $H_0 : E(y|x = x_p) = E_0$

대립가설 $H_a : E(y|x = x_p) > E_0$

혹은 $H_a : E(y|x = x_p) < E_0$ (단측검정)

$H_a : E(y|x = x_p) \neq E_0$ (양측검정)

검정통계량 : $t = \dfrac{\hat{y} - E_0}{s\sqrt{\dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{SS_x}}}$

기각역 : 자유도 $n-2$인 경우에 t에서 t의 임계값을 이용한다.

즉, $t > t_a$ 혹은 $t < -t_a$ (단측검정)

$t > t_{a/2}$ or $t < -t_{a/2}$ (양측검정)

$E(y|x)$에 대한 신뢰구간 : 신뢰계수 $(1-\alpha)$일때 $x=x_p$에서 $E(y|x_p)$에 대한 신뢰구간은 다음과 같다.

$E(y|x)$에 대한 $100(1-\alpha)$ % 신뢰구간

$$\hat{y} \pm t_{a/2}\, s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}}$$

예제

앞의 예에서, 수학성적이 $x = 50$ 일때, 1학년말의 성적 y의 기대값에 대한 90% 신뢰구간을 구하라.

<풀이> $x_p = 50$ 이므로 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p$를 이용하면 $\hat{y} = 40.78+(0.77)(50)$ $= 79.28$ 이다. 따라서 95%신뢰구간은

$$\hat{y} \pm t_{0.025} S\sqrt{\frac{1}{n} + \frac{(x_p-\bar{x})^2}{SS_x}}$$

이므로 다음과 같이 구할 수 있다.

$$79.28 \pm (2.306)(8.70)\sqrt{\frac{1}{10} + \frac{(50-46)^2}{2,474}}$$

혹은

$$79.28 \pm 6.55$$

6. 주어진 x값에 대한 y의 예측

모집단으로부터 표본으로 추출되지않은 어떤 새로운 학생의 1학년말 성적을 예측하고자 할 때 앞의 <표 1>에서 10개의 측정치로부터 얻어지는 예측방정식 (prediction equation)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

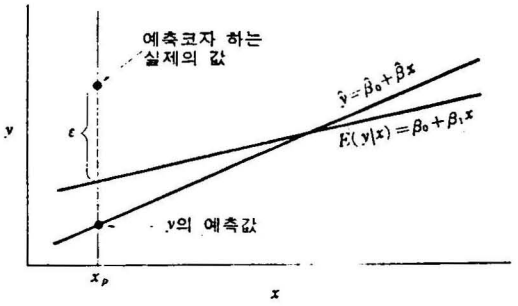
를 이용할 수 있다.

그런데 만약 수학성적이 x_p였다면 예측방정식에 의해 추정된 $\hat{y}$와 실제성적 y간의 차, 즉 예측오차(error of prediction)는 두 개의 원소로 구성됨을 알 수 있다. 즉 새로운 학생의 학년말 성적은

$$y = \beta_0 + \beta_1 x_p + \varepsilon$$

이기 때문에 $(y - \hat{y})$은 $\hat{y}$와 $E(y|x_p)$의 편차에 기대값으로부터의 편차를 나타내는 양 ε을 더한 것과 같은 것이다.

$$y - \hat{y} = [\ y - E(y|x_p)] + [\ \hat{y} - E(y|x_p)]$$



[그림 10.5] y값의 예측오차

따라서 하나의 값 y를 예측하는 데 대한 오차의 분산은 y의 기대치를 추정하는 데 대한 분산보다 더 클 것이며 $x=x_p$일 때 y의 어떤 특별한 값을 예측하는 데 대한 오차의 분산은 다음과 같이 얻어진다.

$$\sigma^2_{y-\hat{y}} = \sigma^2\left[1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_x}\right]$$

n 이 상당히 커지면 괄호 속의 둘째, 세째항은 적어질 것이고 예측오차의 분산은 σ^2에 근사할 것이다. 이 결과를 이용하여 주어진 $x=x_p$일때 y에 대한 예측구간 (prediction interval)을 구할 수 있다.

---

y에 대한 예측구간 :

신뢰계수 $1-\alpha$ 일때 y에 대한 예측구간은 다음과 같이 주어진다.

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_x}}$$

예제

학력고사 수학성적이 y = 50인 학생에 대하여 1학년말의 성적을 예측하여라.

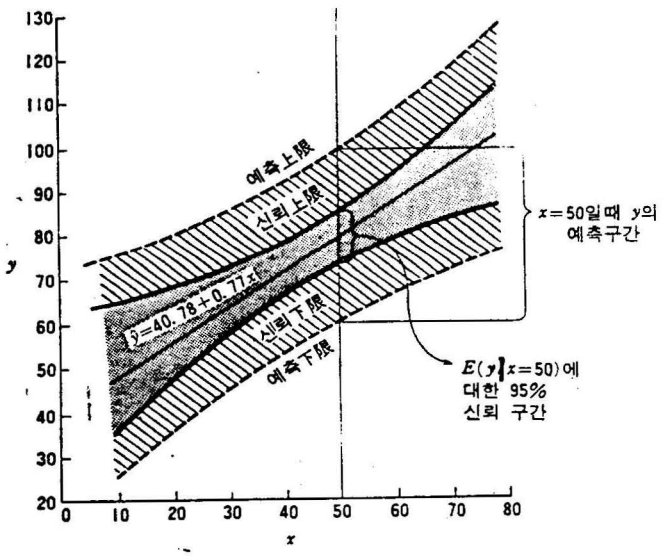
<풀이> x_p = 50 일때 y에 대한 신뢰구간은

$$\hat{y} \pm t_{\alpha/2}\, s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_x}}$$

$$= 79.28 \pm (2.306)(8.70)\sqrt{1+\frac{1}{10}+\frac{(50-46)^2}{2474}}$$

혹은 79.28 ± 21.10 이다.

[그림 6] 은 <표 1>의 자료들을 이용하여 $E(y|x)$에 대한 신뢰구간과 y의 측정한 값에서의 예측구간을 도시한 그림이고 실선은 신뢰구간을 나타내며 점선은 예측구간을 나타낸다.



[그림 6] <표 1>로 부터 $E(y|x)$에 대한
신뢰구간과 y의 예측구간

7. 상관계수

두 변수 x와 y의 선형(직선)관계의 정도를 크기로 나타내고자 할 때가 있다. 이것을 x와 y의 선형상관의 척도(measure of the linear correlation)라 하며, 이 선형관계의 척도를 상관계수(coefficient of correlation)라 하고 r로 표시하며 다음과 같이 정의한다.

표본상관계수 :

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$= \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

예제

<표 1>의 자료에 대하여 1학년말 성적과 학력고사 수학성적에 대한 상관계수를 구하여라.

<풀이>

$$SS_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = 1894$$

$$SS_x = \sum_{i=1}^{n}(x_i-\bar{x})^2 = 2474$$

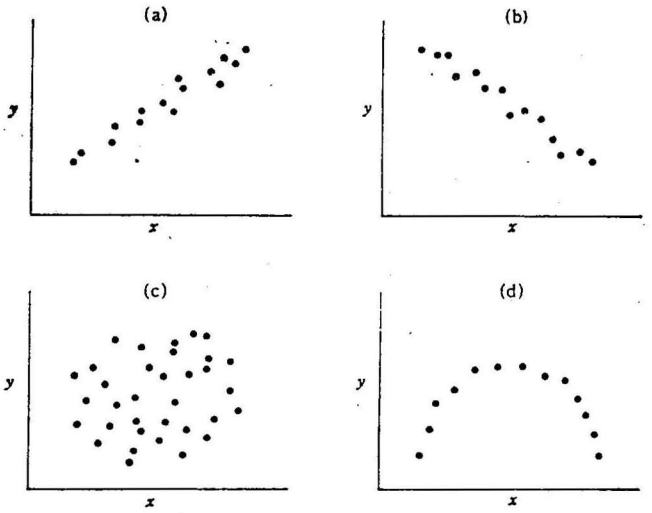
$$SS_y = \sum_{i=1}^{n}(y_i-\bar{y})^2 = 2056$$

이므로

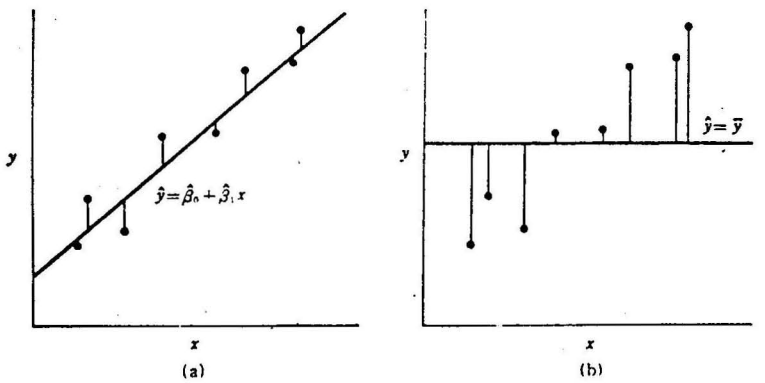
$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{1894}{\sqrt{2474\times 2056}} = 0.84$$

상관계수 r와 기울기의 추정량 $\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$ 의 식을 비교해 보면 분모는 항상 양의 수를 가지고 분자는 동일하게 SS_{xy}이므로 r과 $\hat{\beta}_1$의 부호는 SS_{xy}의 부호와 동일하며 $\hat{\beta}_1 = 0$ 이면 $SS_{xy} = 0$이므로 상관계수는 $r = 0$ 이됨을 알 수 있다.

따라서 $r = 0$ ($SS_{xy} = 0$)일 때는 x와 y사이에 상관관계가 없음을 나타낸다. 또한 r의 값이 양의 수로 얻어지면 기울기는 양이 되며 x가 증가하면 r도 증가하는 관계를 가지며, 이 때를 양의 상관(positive correlation)이라 하고[그림 7(a)] r의 값이 음의 수를 가지면 기울기는 음이 되어 x가 증가할 때 y는 감소하는 관계를 갖는데 [그림 7(b)] 이를 음의 상관(negative correlation)이라한다. $r = 0$에 가까우면 x와 y사이에는 상관이 없다고 하고 $r = 0$일 때는 무상관이라 한다. [그림 7(c)].



[그림 7] 점산도



[그림 8] 동일자료에대한 두 개의 적합선

표본상관계수의 성질 :

① 표본상관계수 r은 항상 -1과 1 사이에 있다.

② r의 절대값의 크기는 선형관계의 강도를 나타낸다.

③ $r > 0$: 직선의 기울기가 양(+)수로서 x가 증가할때 y도 증가하는 경향을 나타내며 $r = 1$일때 양(+)의 완전상관이라 한다.(그림 7a)

④ $r < 0$: 직선의 기울기가 음(-)수로서 x가 증가할때 y도 감소하는 경향을 나타내며 $r = -1$일때 음(-)의 완전상관이라 한다.(그림 7b)

⑤ $r = 0$: r의 값이 0에 가까울 경우 직선관계가 매우 약함을 나타내며 $r < 0$일때를 무상관이라 한다.(그림 7c)

* 두변수사이에 곡선관계가 있을때 r의 값이 0에 가까울수 있다. (그림 7d)

선형회귀모형을 적합시킴으로서 줄어드는 변동인

$$SS_y - SSE = \hat{\beta}_1 SS_{xy}$$

와 편차제곱합 SS_y의 비를 R^2이라 놓으면

$$R^2 = \frac{SS_y - SSE}{SS_y}$$

이때 R^2을 결정계수(coefficient of determination)라 하고, R^2은 다음구간에 있으며

$$0 \leq R^2 \leq 1$$

이 결정계수는 x와 y의 관계를 설명하는데 상관계수보다도 더 유용하게 쓰인다.

결정계수 :

$$R^2 = \frac{SS_y - SSE}{SS_y}$$

확률변수 X와 Y 사이의 모상관계수 ρ의 검정에 상관계수 r을 이용한다.

상관계수의 검정 :

귀무가설 : $H_0 : \rho = 0$

검정통계량 : $\delta = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$

기각역 : ① $H_1 : \rho < 0$ 일때 $\delta \geq t_{(n-2,\alpha)}$

② $H_1 : \rho > 0$ 일때 $\delta \leq -t_{(n-2,\alpha)}$

③ $H_1 : \rho \neq 0$ 일때 $|\delta| \geq t_{(n-2,\frac{\alpha}{2})}$

8. 중회귀모형

종속변수 y가 2개이상의 독립변수 $x_1, \cdots, x_k$에 의하여 영향을 받을때 종속변수와 독립변수의 관계를 단순회귀모형에서와 같은 방법으로 다음과 같이 가정하면

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

여기서

$$\varepsilon \sim N(0, \sigma^2).$$

이모형을 중회귀모형(multiple regression model)이라 하며, 예측방정식은 단순선형모형에서 사용했던 방법과 동일하게 최소자승법에 의해 구할 수 있다.

예를 들어 학생들의 1학년 말 평균점수를 y, 고등학교에서의 등수를 x_1, 수학능력시험에서의 수리탐구 영역을 x_2, 언어탐구 영역을 x_3라 하고(물론 다른 변수들도 있으나 여기서는 세개의 독립변수만을 생각한다), 다음과 같은 중회귀모형

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

을 적합시키고자 한다.

이 모형으로 부터 적합된 회귀선을 구하기위해 최소자승법으로 $\beta_0, \beta_1, \beta_2, \beta_3$를 추정해보자.

$\beta_0, \beta_1, \beta_2, \beta_3$의 추정량을 각각 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$라 하면 적합된 회귀선은

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \varepsilon$$

이 되며 적합된 회귀선에 의한 예측값 $\hat{y}$와 관측치 y와의 차, 즉 잔차를 최소로 하는 회귀선은 잔차자승합

$$Q = SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i-1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i})^2$$

를 최소로 하는 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$로 부터 구해진다. 따라서 Q를 각각 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$에 대해 편미분을 하여 0이라 놓으면 다음과 같은 정규방정식을 구할 수 있다.

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i-1}^{n} x_{1i} + \hat{\beta}_2 \sum_{i-1}^{n} x_{2i} + \hat{\beta}_3 \sum_{i-1}^{n} x_{3i}$$

$$\sum_{i=1}^{n} x_{1i} y_i = \hat{\beta}_0 \sum_{i-1}^{n} x_{1i} + \hat{\beta}_1 \sum_{i-1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i-1}^{n} x_{1i} x_{2i} + \hat{\beta}_3 \sum_{i-1}^{n} x_{1i} x_{3i}$$

$$\sum_{i=1}^{n} x_{2i} y_i = \hat{\beta}_0 \sum_{i-1}^{n} x_{2i} + \hat{\beta}_1 \sum_{i-1}^{n} x_{1i} x_{2i} + \hat{\beta}_2 \sum_{i-1}^{n} x_{2i}^2 + \hat{\beta}_3 \sum_{i-1}^{n} x_{2i} x_{3i}$$

$$\sum_{i=1}^{n} x_{3i} y_i = \hat{\beta}_0 \sum_{i-1}^{n} x_{3i} + \hat{\beta}_1 \sum_{i-1}^{n} x_{1i} x_{3i} + \hat{\beta}_2 \sum_{i-1}^{n} x_{2i} x_{3i} + \hat{\beta}_3 \sum_{i-1}^{n} x_{3i}^3$$

이 정규방정식을 풀어 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$를 구하면 되나 그 과정이 복잡하고 또한 예측치들을 분석하는 문제는 복잡하므로 이 책에서는 생략하기로 하고 독립변수의 수가 k이면 정규방정식의 수는 $(k+1)$개가 되어 회귀계수를 구하는 일이 상당히 어려우나 콤퓨터의 발달로 인하여 쉽게 계산할 수 있는 프로그램이 많이 개발되어 있다.

9. PROC REG의 예제 프로그램

가. 단순회귀분석

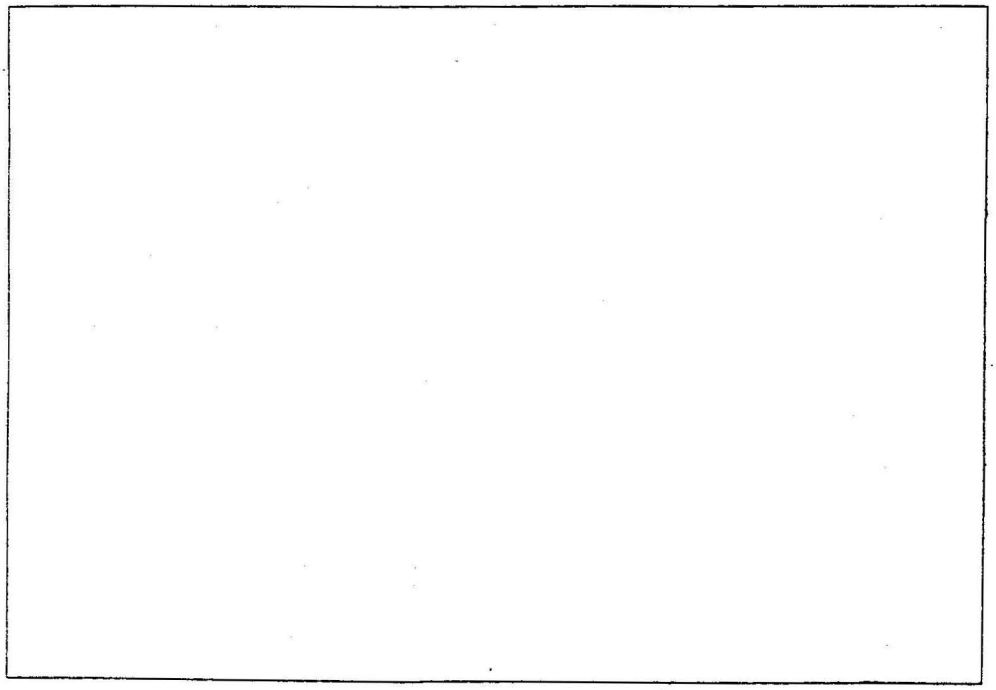
(1) SAS데이타셋의 구성

```
OPTIONS PS=60 NODATE PAGENO=1;
DATA Liking1;
  INPUT Y X @@ ;
  LABEL Y='선호정도' X='과외활동정도';
  CARDS;
3 4 7 9 2 3 1 1 6 3 2 4 8 7 3 3 9 8 2 1
RUN;
```

(2) 데이타의 그래프화

각 변수가 선형형태를 보이는가를 보기 위해 PROC GPLOT를 통해 산포도 및 회귀선을 도시시켜 보자. PROC GPLOT에서는 한글을 처리할 수 없기 때문에 레이블을 영문으로 수정했다. GOPTIONS문을 통해 그래프를 휴렛팩커드 레이저 프린터로 출력했다. SYMBOL문의 I=RL은 회귀선을 그래프에 삽입하기 위한 옵션이다. 산포도나 삽입된 회귀선을 볼 경우 선형성이 있는 것으로 생각할 수 있을 것이다.

```
DATA TEMP; SET LIKING1;
  LABEL Y='Preference'  X='Extraactivity'; RUN;
GOPTIONS DEVICE=HPLJS2 GUNIT=CELLS FTEXT=SWISS HTEXT=2
        HORIGIN=1.2IN VORIGIN=3IN HSIZE=5IN VSIZE=4IN;
SYMBOL1 I=RL V=DOT;
PROC GPLOT DATA=TEMP;
     PLOT Y*X=1; RUN;
```



(3) 단순회귀분석

```
PROC CORR DATA=Liking1 NOSIMPLE;
  TITLE '자료에 대한 상관관계분석';
  VAR Y;  WITH X;  RUN;
PROC REG DATA=Liking1;
  TITLE '자료에 대한 단순회귀분석';
  MODEL Y=X;  RUN;
```

(4) 단순회귀 분석결과

출력결과 1페이지를 보면 Y와 X변수간의 상관관계를 보여주고 있다. 상관관계계수가 0.85로서 의미가 있다($p < 0.002$). 따라서 우리가 수행하는 회귀분석도 의미가 있을 것이라는 것을 예측할 수 있다. 이러한 결과는 앞의 그래프에서도 볼 수 있으며 데이타간의 선형관계가 있다는 것을 나타낸다.

자료에 대한 상관관계분석 1

CORRELATION ANALYSIS

1 'WITH' Variables: X
1 'VAR' Variables: Y

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 10

	Y
X	0.85024
과외활동정도	0.0018

자료에 대한 단순회귀분석 2

Model: MODEL1
Dependent Variable: Y 선호정도

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	55.01298	55.01298	20.871	0.0018
Error	8	21.08702	2.63588		
C Total	9	76.10000			

Root MSE	1.62354	R-square	0.7229
Dep Mean	4.30000	Adj R-sq	0.6883
C.V.	37.75671		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.490728	0.97920512	0.501	0.6298
X	1	0.885877	0.19391151	4.568	0.0018

Variable	DF	Variable Label
INTERCEP	1	Intercept
X	1	과외활동정도

Y와 X변수간의 회귀분석결과가 출력결과 2페이지에 제시되어 있다. 결과를 보면 분산분석표에서 F값이 20.87($p < 0.002$)로서 회귀식이 의미가 있으며 (이는 상관관계분석과도 일치한다), R^2 값도 0.73으로서 현재의 회귀식이 전체 변동의 73%정도를 설명하고 있다. 회귀식에 대한 평균표준편차 (MSE(Y))는 1.63이고 Y의 평균은 4.3이다. 추정된 회귀식을 정리하면,

$$Y = \underset{(0.98)}{0.49} + \underset{(0.19)}{0.89}\, X \quad (R^2=0.73)$$

로서 절편이 의미가 없음을 알 수 있다. 따라서 회귀식의 절편은 0이라고 보더라도 지장이 없다고 할 수 있다.

나. 다중회귀분석

다음과 같이 라디오광고와 TV광고의 효과를 알아보기 위해서 라디오광고의 횟수와 TV광고의 횟수에 대한 광고내용의 기억정도를 측정했다.

라디오광고	TV광고	광고기억정도	라디오광고	TV광고	광고기억정도
1	2	54	3	2	61
1	4	60	3	4	66
1	6	62	3	6	70
1	8	65	3	8	67
2	2	59	4	2	58
2	4	64	4	4	67
2	6	65	4	6	66
2	8	67	4	8	65

이 데이타를 가지고 독립변수가 2개인 다중회귀모형을 분석해 보자.

(1) SAS데이타셋의 구성

[화일명:9-12REGR.SAS]

```
DATA ADTYPE;
    INPUT RADIO TV ADRECALL @@;
    LABEL RADIO='라디오광고횟수' TV='TV광고횟수'
          ADRECALL='광고기억정도';
    CARDS;
 1 2 54 3 2 61 1 4 60 3 4 66 1 6 62 3 6 70 1 8 65 3 8 67
 2 2 59 4 2 58 2 4 64 4 4 67 2 6 65 4 6 66 2 8 67 4 8 65
 ;
```

(2) 다중회귀분석 프로그램

[화일명:9-12REGR.SAS]

```
PROC REG DATA=ADTYPE;
   MODEL ADRECALL=RADIO TV; RUN;
```

(3) 다중회귀분석 수행결과

//모형의 적절성//

① 모형의 분산분석결과

모형의 모수들에 대한 귀무가설은,

H_0: $\beta_i = 0$, $i = 1, \ldots, p$

이에 대한 검정결과를 보기 위해 출력결과 1페이지의 분산분석표를 보면 F값이 11.575($p < 0.0013$)이다. 따라서 현재의 귀무가설을 기각할 수 있다. 즉 이에 대한 대립가설 "H_1: 적어도 하나이상의 β_i는 0이 아니다"라고 볼 수 있다.

SAS 1

Model: MODEL1
Dependent Variable: ADRECALL 광고기억정도

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	166.50000	83.25000	11.575	0.0013
Error	13	93.50000	7.19231		
C Total	15	260.00000			

Root MSE	2.68185	R-square	0.6404
Dep Mean	63.50000	Adj R-sq	0.5851
C.V.	4.22338		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	53.750000	2.22367074	24.172	0.0001
RADIO	1	1.350000	0.59967940	2.251	0.0423
TV	1	1.275000	0.29983970	4.252	0.0009

Variable	DF	Variable Label
INTERCEP	1	Intercept
RADIO	1	라디오광고횟수
TV	1	TV광고횟수

② R^2값 또는 조정 R^2값

R^2값이 0.6404이고 또는 조정 R^2값이 0.5851이어서 매우 높다고 할 수 있다. R^2의 계산식은

$$R^2 = \frac{SSR}{SST} = \frac{(SST-SSE)}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{93.50}{260.00} = 0.6404$$

로 계산이 되며, 조정 R^2의 계산식은

$$\text{조정 } R^2 = \frac{(n-1)R^2 - p}{n-k-1} = \frac{15X0.6404-2}{13} = 0.5851$$

로 계산이 된다.

③ 각 모수추정치의 유의도

각 추정치에 대한 유의도를 알아보고자 각 추정치에 대한 t값의 유의도를 보았다. 각 모수에 대한 t값이 모두 $p < 0.05$ 수준에서 의미가 있다.

④ 전반적인 모형의 적절성

이러한 결과들로 보아 우리가 세운 모형은 어느 정도 적절하다고 할 수 있다. 그러나 모형이 정말 적합한지를 보고자 하면, 잔차에 대한 검정을 실시해야 한다.

//모형 추정식//

회귀분석을 통한 추정모형은 다음과 같이 정리할 수 있다. 식아래의 괄호 안에 있는 숫자는 추정치에 대한 표준오차이다.

$$\underset{}{\text{ADRECALL}} = \underset{}{53.750} + \underset{(2.224)}{1.350}\ \text{RADIO} + \underset{(0.600)}{1.275}\ \text{TV}_{(0.300)}$$

다. 회귀모형의 잔차분석

설정한 모형의 가정이 적절한지를 알아보기 위해서는 잔차항들이 서로 독립이며 정규분포를 한다는 것을 본다. 잔차를 검정하는 방법은 더빈-왓슨(Durbin-Watson) 통계량을 이용해서 알아보는 방법과 기술통계학에서 살펴본 PROC UNIVARIATE를 이용해서 잔차에 대한 연검정(run test)과 샤피로-윌크(Shapiro-Wilk) 검정을 실시해 보는 것이다.

(1) 잔차분석 프로그램

잔차에 대한 더빈-왓슨 통계량을 알아보기위해서는 MODEL ADRECALL = RADIO TV / DW; 와 같이 /와 DW 옵션을 추가해야 하며, 예측치와 잔차에 대한 산포도를 알아보기 위해서는 PLOT R. * P.; 문을 추가하면 된다. 연검정과 샤피로-윌크 검정을 알아보기 위해 PROC UNIVARIATE에서 잔차에 대한 정규성 검정(NORMAL 옵션)과 정규분포에 대한 그래프(PLOT 옵션)를 살펴보았다.

[화일명:9-131REG.SAS]

```
PROC REG DATA=ADTYPE;
    MODEL ADRECALL=RADIO TV /DW;
    PLOT R. * P.;
    OUTPUT OUT=ERORR R=RESID;
PROC UNIVARIATE DATA=ERROR NORMAL PLOT;
    VAR RESID; RUN;
```

(2) 잔차분석 프로그램 수행결과

① 잔차에 대한 더빈-왓슨값

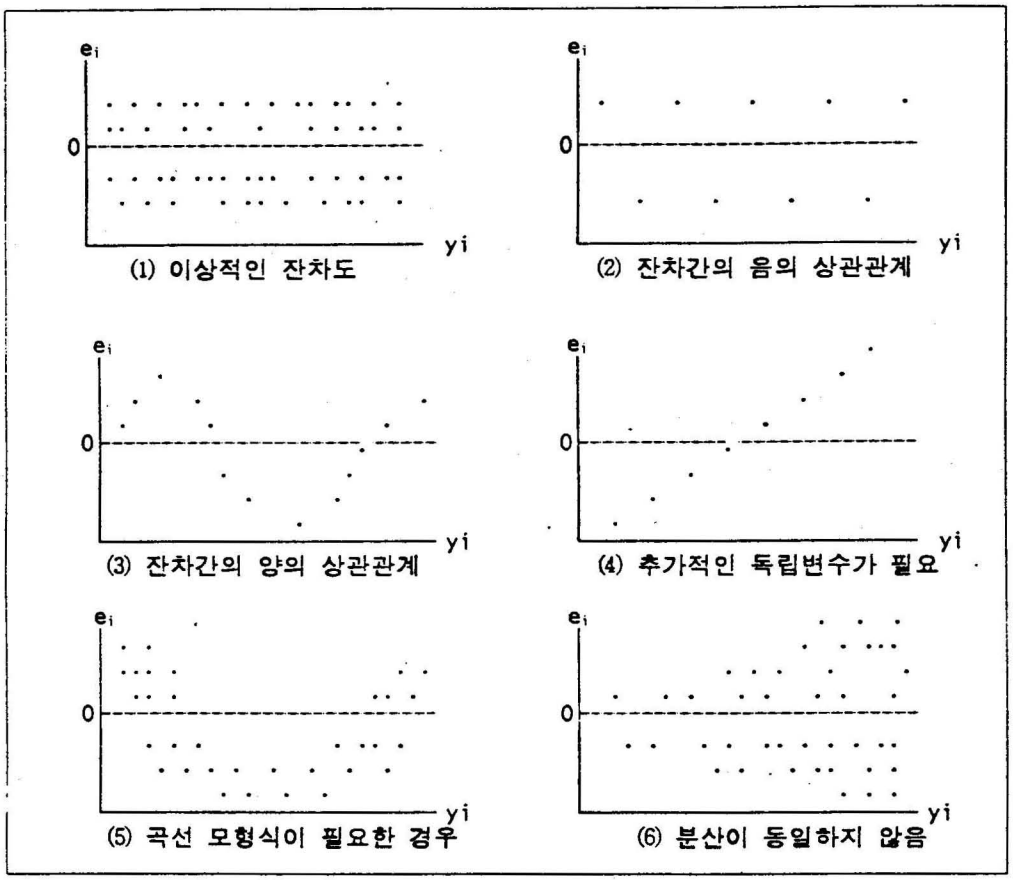
출력결과 2페이지에서 보면 잔차에 대한 더빈-왓슨값은 1.918로서 기준값이 2이므로 매우 적절하다고 할 수 있다. 즉 잔차는 정규분포를 하고 있다고 볼 수 있다. 또한 잔차에 대한 1차 자기상관계수값도 -0.132로서 낮은 값이다. 더빈-와슨값은 DW = 2(1- ρ)로서 2인 경우에 잔차에 대한 상관관계가 없음을 알 수 있으며, 0에 가까울수록 양의 상관관계를 나타내며 4에 가까울수록 음의 상관관관계를 나타낸다. DW값이 0에 가깝거나 4에 가까우면 잔차들간에 상관관계가 있어 모형이 적합하다고 할 수 없다. 따라서 본 모형은 이러한 기준에서 볼 때 적절하다고 볼 수 있다.

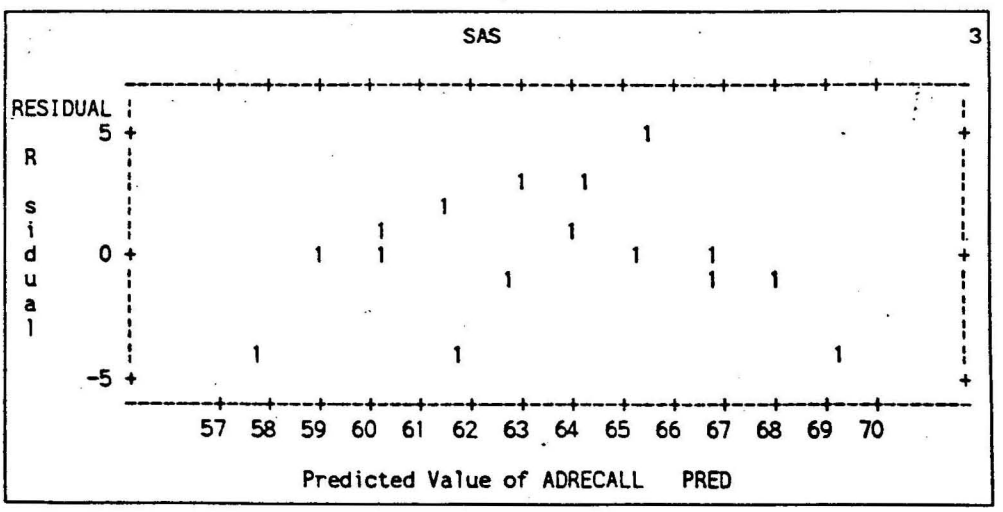
```
                                SAS                                2

Durbin-Watson D                 1.918
(For Number of Obs.)               16
1st Order Autocorrelation       -0.132
```

② 잔차에 대한 산포도

종속변수 또는 예측치와 잔차에 대한 산포도 형태가 (1)번과 같은 형태를 보이면 이상적이라고 볼 수 있으며, 기타 잔차의 모양이 (2), (3)과 같은 경

우는 잔차가 음 또는 양의 상관관계를 가지고 있으며, (4)와 같은 경우는 새로운 변수를 추가해야 한다는 것을 의미한다.





(5)와 같은 경우는 2차항 형태의 모형이 필요하다는 것을 의미하며, (6)과 같은 경우는 분산이 일정하다는 가정이 어긋나고 있다는 것을 의미한다. 따라서 이러한 경우는 가중회귀모형으로 분석하는 것이 바람직할 것이다. 잔차에 대한 산포도를 나타내는 출력결과 3페이지를 볼 경우 잔차가 특정한 패턴을 가지고 있다고 보기에는 미흡하다. .따라서 잔차에 대한 산포도에서도 적절한 형태를 띄고 있다고 볼 수 있다.

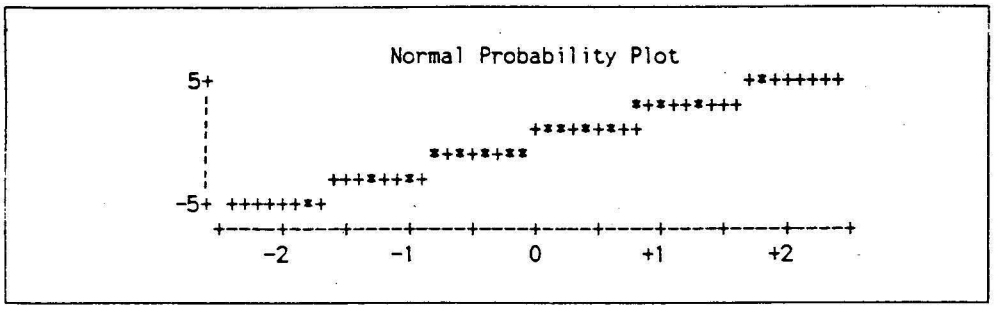
③ 잔차에 대한 정규성 검정

PROC UNIVARIATE으로 연검정과 샤피로-윌크검정을 한 결과 출력결과 4페이지에서 연검정 값이 -1($p < 0.9780$), 샤피로-윌크의 W값이 0.955($p < 0.5562$) 로서 잔차가 정규분포를 하고 있다는 것을 보여준다. 또한 이러한 결과는 출력결과 5페이지의 정규분포에 대한 그림에서도 거의 정규분포로 볼 수 있다. 결론적으로 "잔차가 정규분포를 한다"는 가정은 적절하다고 볼 수 있다.

```
                                SAS                                    4

                        UNIVARIATE PROCEDURE

Variable=RESID       Residual

                             Moments

          N                  16  Sum Wgts           16
          Mean                0  Sum                 0
          Std Dev      2.496664  Variance      6.233333
          Skewness    -0.11272  Kurtosis     -0.27338
          USS              93.5  CSS                93.5
          CV                  .  Std Mean      0.624166
          T:Mean=0            0  Prob>|T|        1.0000
          Sgn Rank           -1  Prob>|S|        0.9780
          Num ^= 0           15
          W:Normal     0.955069  Prob<W          0.5562

  * 이하 결과는 생략함
```

```
                                SAS                                    5
                        UNIVARIATE PROCEDURE

Variable=RESID       Residual

          Stem Leaf                    #          Boxplot
             4 6                       1             |
             2 481                     3             |
             0 0469                    4          +--+--+
            -0 08832                   5          *-----*
            -2 76                      2             |
            -4 4                       1             |
               ----+----+----+----+
```

(출력결과 5페이지 계속)



(3) 각 관찰치의 잔차검정

각 관찰치에 대하여 잔차를 분석하기 위해서는 MODEL문에서 R 옵션을 사용하면 된다. 이에 대한 프로그램은 다음과 같다.

[화일명:9-132REG.SAS]

```
PROC REG DATA=ADTYPE;
   MODEL ADRECALL=RADIO TV /R; RUN;
```

각 관찰치에 대한 분석의 결과 중에 관심을 가지고 보아야 할 것은 Rstudent라는 통계량과 Cook's D 통계량이다. Rstudent는 Residual(=Dep Var-Predicted Value) 값을 표준화한 T 값이다. 따라서 이 값은 자유도 n-p-1인 t-분포를 이루며, 일반적으로 2이상이면 이상적인 관찰치로서 모수 추정에 영향을 준다고 볼 수 있다.

Cook's D 통계량은 특정관찰치가 회귀계수 전반에 영향을 주는 정도에 관한 통계량으로서 이 값은

$$D_i = \left(\frac{e_i}{\sqrt{MSE(1-\rho_{ij})}} \right)^2 = \frac{\rho_{ij}}{1-\rho_{ij}} \frac{1}{p}$$

로 계산되며, 현재의 결과를 볼 경우 1보다 큰 값이 없기 때문에 영향력있는 관찰치는 없다고 볼 수 있다.

Press값은 각 관찰치마다 이 관찰치를 뺀 후, 나머지 관찰치를 가지고 회귀식을 적합시킨 후에 예측값을 계산하고 실제관측치와의 차이를 합한 결과이다. 이 값이 작을수록 모형이 잘 추정됐다고 볼 수 있다. 이 통계량은 여러 개의 모형 중에 하나 또는 그 이상의 모형을 선택할 때의 기준으로 R^2나 조정 R^2 기준과 같이 사용한다.

SAS 6

Obs	Dep Var ADRECALL	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual
1	54.0000	57.6500	1.438	-3.6500	2.264	-1.612
2	61.0000	60.3500	1.161	0.6500	2.417	0.269
3	60.0000	60.2000	1.161	-0.2000	2.417	-0.083
4	66.0000	62.9000	0.793	3.1000	2.562	1.210
5	62.0000	62.7500	1.161	-0.7500	2.417	-0.310
6	70.0000	65.4500	0.793	4.5500	2.562	1.776
7	65.0000	65.3000	1.438	-0.3000	2.264	-0.133
8	67.0000	68.0000	1.161	-1.0000	2.417	-0.414
9	59.0000	59.0000	1.161	0	2.417	0.000
10	58.0000	61.7000	1.438	-3.7000	2.264	-1.634
11	64.0000	61.5500	0.793	2.4500	2.562	0.956
12	67.0000	64.2500	1.161	2.7500	2.417	1.138
13	65.0000	64.1000	0.793	0.9000	2.562	0.351
14	66.0000	66.8000	1.161	-0.8000	2.417	-0.331
15	67.0000	66.6500	1.161	0.3500	2.417	0.145
16	65.0000	69.3500	1.438	-4.3500	2.264	-1.922

Obs	-2-1-0 1 2	Cook's D
1	***	0.350
2		0.006
3		0.001
4	**	0.047
5		0.007
6	***	0.101
7		0.002
8		0.013
9		0.000
10	***	0.359
11	*	0.029
12	**	0.100
13		0.004
14		0.008
15		0.002
16	***	0.497

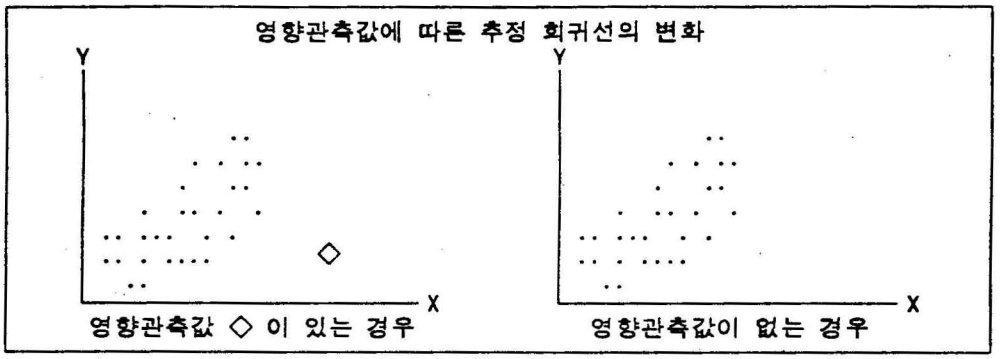
Sum of Residuals	0
Sum of Squared Residuals	93.5000
Predicted Resid SS (Press)	150.9261

(4) 특정 관찰치의 영향 관측값

특정 관찰치가 모수추정치에 영향을 주는지 주지 않는지를 보고자 하는 것이 영향 관측값이다. 특정 관측치가 모수추정치에 영향을 줄수록 이 변수를 포함시킨 모형과 그렇지 않은 모형간에 모수추정값에 심각한 차이가 발생할 것이다. 영향관측값은 그림과 같이 ◇가 있느냐 없느냐에 따라 모수추정치가 달라지기 때문에 그 값을 표시한 것이다. 이를 알아보기 위해서는 MODEL문의 옵션 중에 INFLUENCE옵션을 사용해야 한다.

[화일명:9-133REG.SAS]

```
PROC REG DATA=ADTYPE;
   MODEL ADRECALL=RADIO TV /INFLUENCE; RUN;
```



SAS 7

Obs	Residual	Rstudent	Hat Diag H	Cov Ratio	Dffits	INTERCEP Dfbetas	RADIO Dfbetas	TV Dfbetas
1	-3.6500	-1.7319	0.2875	0.9137	-1.1002	-1.0826	0.6882	0.6882
2	0.6500	0.2591	0.1875	1.5389	0.1244	0.0650	0.0321	-0.0964
3	-0.2000	-0.0795	0.1875	1.5623	-0.0382	-0.0332	0.0296	0.0099
4	3.1000	1.2342	0.0875	0.9735	0.3822	0.0974	0.1445	-0.1445
5	-0.7500	-0.2992	0.1875	1.5303	-0.1437	-0.0751	0.1113	-0.0371
6	4.5500	1.9608	0.0875	0.6053	0.6072	-0.1547	0.2295	0.2295
7	-0.3000	-0.1274	0.2875	1.7772	-0.0809	-0.0114	0.0506	-0.0506
8	-1.0000	-0.4001	0.1875	1.5038	-0.1922	0.1004	-0.0496	-0.1489
9	0	0.0000	0.1875	1.5648	0.0000	0.0000	0.0000	0.0000
10	-3.7000	-1.7618	0.2875	0.8949	-1.1191	-0.1573	-0.7001	0.7001
11	2.4500	0.9530	0.0875	1.1195	0.2951	0.2256	-0.1115	-0.1115
12	2.7500	1.1518	0.1875	1.1425	0.5533	-0.0963	0.4286	-0.1429
13	0.9000	0.3391	0.0875	1.3540	0.1050	0.0268	-0.0397	0.0397
14	-0.8000	-0.3193	0.1875	1.5256	-0.1534	0.0801	-0.1188	-0.0396
15	0.3500	0.1392	0.1875	1.5573	0.0669	-0.0116	-0.0173	0.0518
16	-4.3500	-2.1819	0.2875	0.6549	-1.3860	0.9742	-0.8670	-0.8670

출력결과 7페이지에서 보면 전체적으로 보아 Rstudent값이 2보다 큰값이 16번 관찰치이나 모자대각(Hat Diag H)값이 모두 1이하며 Dffits값도 모두 2이하이다. 또한 각 변수에 대한 Dfbetas값들도 모두 2이하이다. 따라서 영향력을 갖는 관찰치는 없다고 보아도 무방할 것이다.

라. 회귀모형을 이용한 예측, 신뢰구간

현재 추정된 모형으로서 라디오광고와 TV광고의 효과를 예측하기 위해서는 우선 각 변수의 구간이 예측한 모형과 비슷해야 한다. 라디오광고 횟수는 4번 이내이며 TV광고는 2번에서 8번 이내의 값을 가지고 있는 것이 바

람직할 것이다. 예를들어 라디오 광고의 횟수가 3번이고 TV광고의 횟수가 5번이라면 광고기억정도에 대한 예측치는,

$$ADRECALL = 53.750 + 1.350\ (3) + 1.270\ (5) = 64.15$$

로 예측이 된다.

각 관찰치의 광고기억정도에 대한 예측값을 얻기 위해서는 MODEL문에서 P옵션을 사용한다. CLI, CLM, R과 같은 옵션이 사용되면 없어도 된다. 광고기억정도의 기대값에 대한 95% 신뢰구간을 보고자하면 CLM옵션을 사용하며, 각 관찰치의 예측치에 대한 95% 신뢰구간을 보고자 하면 CLI옵션을 사용한다.

[화일명:9-14REGR.SAS]

```
PROC REG DATA=ADTYPE;
   MODEL ADRECALL=RADIO TV /CLI CLM; RUN;
```

출력결과 8페이지에서 보면 각 관찰치에 대한 예측치(Predicted Value)와 예측치에 대한 표준오차(Std Err Predict)와 기대치 (Mean) 에 대한 95% 상한/하한 신뢰구간, 예측치(Predict)에 대한 95% 상한/하한 신뢰구간 등이 출력되어 있다. 이들 결과를 볼 경우 기대치에 대한 신뢰구간이 예측치에 대한 신뢰구간보다 상한값이 더 작으며 하한값이 더 큼을 알 수 있다. 기대치에 대한 신뢰구간의 폭(Width)이 예측치에 대한 신뢰구간의 폭보다 좁다는 것을 의미한다. 이러한 결과가 나타나는 것은 예측치의 불확실성을 반영하기 때문이다. 기대값에 대한 95%신뢰구간의 추정식은

SAS 8

Obs	Dep Var ADRECALL	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict
1	54.0000	57.6500	1.438	54.5434	60.7566	51.0759	64.2241
2	61.0000	60.3500	1.161	57.8412	62.8588	54.0364	66.6636
3	60.0000	60.2000	1.161	57.6912	62.7088	53.8864	66.5136
4	66.0000	62.9000	0.793	61.1862	64.6138	56.8581	68.9419
5	62.0000	62.7500	1.161	60.2412	65.2588	56.4364	69.0636
6	70.0000	65.4500	0.793	63.7362	67.1638	59.4081	71.4919
7	65.0000	65.3000	1.438	62.1934	68.4066	58.7259	71.8741
8	67.0000	68.0000	1.161	65.4912	70.5088	61.6864	74.3136
9	59.0000	59.0000	1.161	56.4912	61.5088	52.6864	65.3136
10	58.0000	61.7000	1.438	58.5934	64.8066	55.1259	68.2741
11	64.0000	61.5500	0.793	59.8362	63.2638	55.5081	67.5919
12	67.0000	64.2500	1.161	61.7412	66.7588	57.9364	70.5636
13	65.0000	64.1000	0.793	62.3862	65.8138	58.0581	70.1419
14	66.0000	66.8000	1.161	64.2912	69.3088	60.4864	73.1136
15	67.0000	66.6500	1.161	64.1412	69.1588	60.3364	72.9636
16	65.0000	69.3500	1.438	66.2434	72.4566	62.7759	75.9241

$$x_0'b \pm t_{(n-p,\ \alpha/2)}\sqrt{MSE\ x_o'(X'X)^{-1}x_0}$$

예측치에 대한 95%신뢰구간의 추정식은

$$x_0'b \pm t_{(n-p,\ \alpha/2)}\sqrt{MSE\ (1+x_o'(X'X)^{-1}x_0)}$$

이기 때문에 신뢰구간의 폭이 서로 다르게 나타난다.

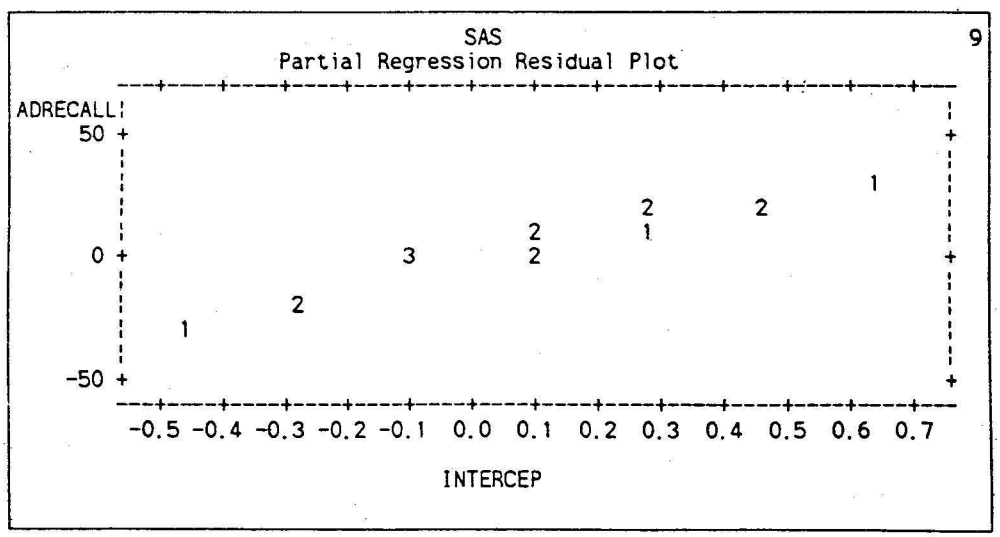
마. 회귀모형에서 각 변수의 선형성(편회귀 지뢰도)검정

단순회귀모형에서는 하나의 종속변수와 독립변수간의 산포도를 그려봄으로써 각 독립변수가 종속변수에 대해서 선형성에 대한 가정을 만족하는지를 알아볼 수 있다. 다중회귀모형에서 이를 알아보기 위해서는 여러 개의 산포도를 그려야 하나, SAS에서는 이를 알아보기 위해서 MODEL문에 PARTIAL이라는 옵션을 통해 간단히 알아 볼 수 있다.

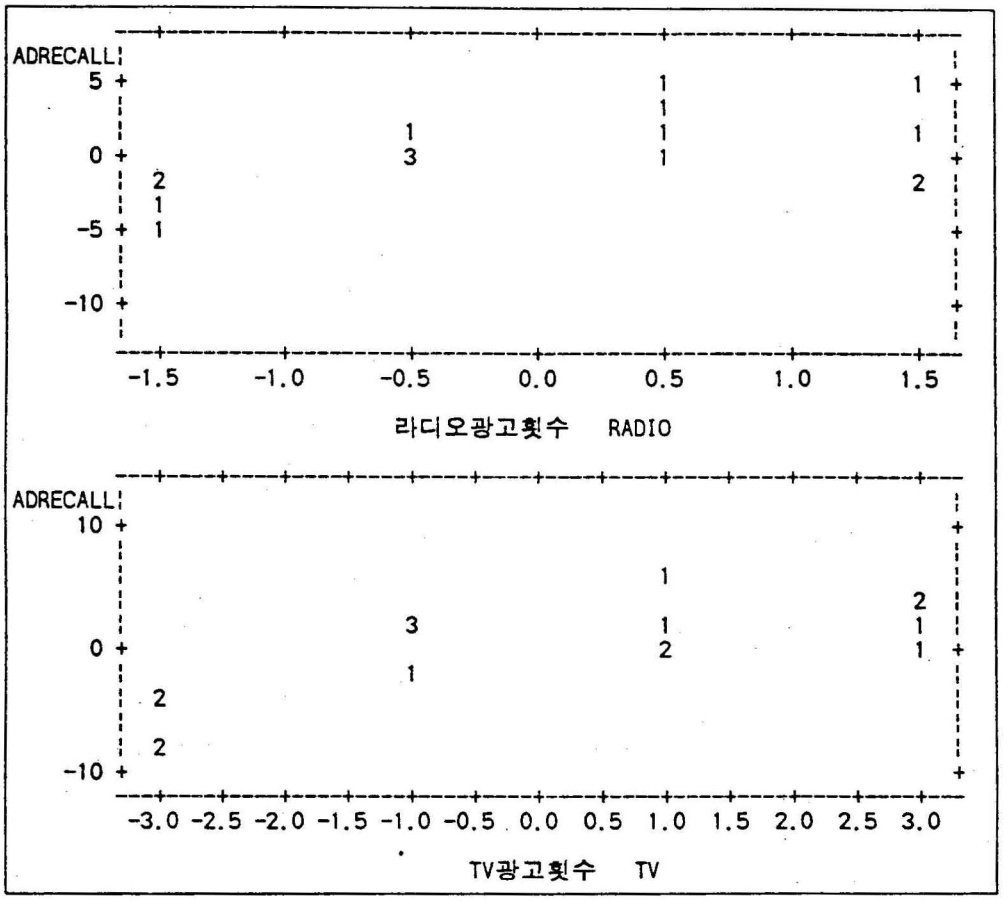
[화일명:9-15REGR.SAS]

```
PROC REG DATA=ADTYPE;
   MODEL ADRECALL=RADIO TV /PARTIAL; RUN;
```

출력결과 9페이지를 볼 경우 모든 변수가 종속변수에 대해 선형적이라고 할 수 있다.



(출력결과 9페이지 계속)



바. 회귀모형에서 각 모수추정치의 중요도

회귀분석을 하는 과정에서 각 독립변수의 단위나 표준편차가 다른 경우 각 독립변수가 차지하는 중요도는 단순히 추정된 모수추정치가 아니다. 따라서 독립변수간의 중요성을 파악할 수 있는 표준화된 모수추정치가 필요한데 이를 수행하는 옵션이 MODEL문에서 STB 옵션이다.

[화일명:9-16REGR.SAS]

```
PROC REG DATA=ADTYPE;
    MODEL ADRECALL=RADIO TV/STB; RUN;
```

출력결과 10페이지로부터 라디오광고 변수의 중요도가 0.37정도이며

TV광고 변수의 중요도가 0.71로서 TV광고가 광고기억정도에 더 영향을 미치는 것으로 볼 수 있으며, 그 영향을 미치는 정도가 약 2배가 정도라고 할 수 있다. 표준화된 모수에 의한 추정식은 다음과 같이 정리된다. 회귀식은 절편이 없는 식이며 각 변수들은 각 변수의 평균을 빼주고 각 변수의 표준편차로 나누어 준 값(이런 의미에서 각 변수에 "'"를 표시)이다.

ADRECALL' = 0.374 RADIO' + 0.707 TV'

```
                                SAS                                    10

                    Standardized  Variable
Variable  DF            Estimate     Label

INTERCEP   1          0.00000000  Intercept
RADIO      1          0.37442263  라디오광고횟수
TV         1          0.70724275  TV광고횟수
```

사. 이분산성에 대한 검정

이분산성(Heteroscedasticity)에 대한 검정은 모수추정치의 통계량들이 동일한 분산을 가지고 있는지에 대한 검정이다. 이분산성에 대한 통계량을 산출하는 옵션은 MODEL문에 ACOV(근사 공분산행렬을 출력), SPEC(동분산성에 대한 χ^2검정)를 지정함으로써 얻을 수 있다.

[화일명:9-17REGR.SAS]

```
PROC REG DATA=ADTYPE;
   MODEL ADRECALL=RADIO TV/ACOV SPEC; RUN;
```

```
                                SAS                                    11

                    Consistent Covariance of Estimates

ACOV                    INTERCEP              RADIO                 TV

INTERCEP          4.910546875          -0.9194375         -0.51403125
RADIO              -0.9194375           0.3325875          0.05046875
TV                 -0.51403125          0.05046875         0.074071875

          Test of First and Second Moment Specification
 DF:     5   Chisq Value: 52.033644104    Prob>Chisq:    0.0000
```

출력결과 11페이지를 보면 각 모수추정치들에 대한 분산과 동분산성에 대한 검정값이 나와 있는데, 공분산 행렬을 보더라도 분산이 동일하지 않을

것이라는 것을 예상할 수 있다. 이러한 결과는 분산이 동일하다는 귀무가설에 대한 검정결과를 보아도 비슷한 결과가 나타난다. χ^2값이 52.03($p < 0.0000$)으로서 모수추정치간의 분산이 동일하다는 가설이 기각된다.

아. 다중회귀모형에서 다중공선성 검정

독립변수들간에 상관계수가 높은 경우 변수들간의 상관성으로 말미암아 모형이 적절하게 예측된 것 같이 R^2값이 높게 나타난다. 또한 각 변수에 대한 모수추정치가 0으로(각 모수추정치에 대한 t값으로 판단) 나타난다. 이러한 경우 변수들 사이에 다중공선성(multicollinearity)이 있다고 할 수 있다. SAS에서는 이 들 다중공선성을 검정할 수 있는 옵션을 제공하고 있다. 이 옵션은 MODEL문에서 COLLIN(절편(intercept)이 없을 경우에는 COLLINOINT), VIF, TOL이라는 세 가지 옵션을 제공하고 있다. 이들 옵션을 이용한 예제를 보면 다음과 같다.

[화일명:9-18REGR.SAS]

```
DATA COLLIN;
   INPUT X1-X4 Y;
   CARDS;
130  30 60  35  785
145  75 52   5  743
155  40 47  55  876
155 110 44   5  725
200 115 34  10  838
235  20 26 105 1159
260  30 33  35  959
270  90 22  10  931
275  45 22  55 1092
280  40 20  55 1043
305  85  6  15 1027
330  45 12  55 1139
340  40 12  50 1094
;
PROC CORR DATA=COLLIN;
   VAR X1-X4 Y;
PROC REG DATA=COLLIN;
   MODEL Y = X1 - X4 / COLLIN VIF TOL; RUN;
```

출력결과 12페이지에서 보면 상관계수를 보면 X1과 X3의 상관계수가 -0.96으로 나타나 독립변수들을 회귀식에 모두 포함시키면 회귀식이 다중공선성이 있다는 것을 의심할 수 있다. 또한 X2와 X4변수간에도 다중공선성이 있을 가능성이 있어 보인다(Green 1978).

SAS 12

CORRELATION ANALYSIS

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 13

	X1	X2	X3	X4	Y
X1	1.00000 0.0	-0.19566 0.5218	-0.95831 0.0001	0.28482 0.3456	0.84602 0.0003
X2	-0.19566 0.5218	1.00000 0.0	0.02954 0.9237	-0.81110 0.0008	-0.53397 0.0602
X3	-0.95831 0.0001	0.02954 0.9237	1.00000 0.0	-0.24564 0.4186	-0.82159 0.0006
X4	0.28482 0.3456	-0.81110 0.0008	-0.24564 0.4186	1.00000 0.0	0.72947 0.0047

SAS 13

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	269425.07048	67356.26762	124.454	0.0001
Error	8	4329.69875	541.21234		
C Total	12	273754.76923			

Root MSE	23.26397	R-square	0.9842
Dep Mean	954.69231	Adj R-sq	0.9763
C.V.	2.43680		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	878.198559	248.96416928	3.527	0.0078
X1	1	0.540806	0.54371805	0.995	0.3490
X2	1	-0.389572	0.62229087	-0.626	0.5487
X3	1	-4.081240	2.44461367	-1.669	0.1336
X4	1	2.486616	0.58391606	4.259	0.0028

출력결과 13페이지에서 보면 분산분석표에서 모형의 적합성을 나타내는 F값이 124.454($p < 0.001$)이고 R^2값이 0.9842로서 매우 높다는 것을 알 수 있다. 각 독립변수에 대한 모수추정치의 결과를 보면 유의한 변수가 별로 없음을 알 수 있다. 특히 독립변수 X1과 X3는 종속변수 Y와 상관계수가 높게 나타났음에도 불구하고 유의하지 않게 나타났다. 이러한 경우에는 다중공선성을 의심하지 않을 수 없다.

Variable	DF	Tolerance	Variance Inflation
INTERCEP	1	.	0.00000000
X1	1	0.02881790	34.70065151
X2	1	0.11355446	8.80634699
X3	1	0.02693702	37.12363453
X4	1	0.15746106	6.35077631

Collinearity Diagnostics

Number	Eigenvalue	Condition Number	Var Prop INTERCEP	Var Prop X1	Var Prop X2	Var Prop X3	Var Prop X4
1	4.14179	1.00000	0.0000	0.0001	0.0011	0.0003	0.0022
2	0.54225	2.76373	0.0000	0.0001	0.0146	0.0008	0.0587
3	0.27655	3.86999	0.0000	0.0020	0.0103	0.0136	0.0035
4	0.03894	10.31307	0.0009	0.0187	0.2223	0.0052	0.3457
5	0.0004726	93.61826	0.9990	0.9791	0.7516	0.9801	0.5899

출력결과 14페이지에서 보면 모수추정치에 대한 허용도(Tolerance) 값이 0.1이하이면 다중공선성이 있다고 할 수 있는데 변수 X1과 X3가 다중공선성이 있는 것으로 나타났다. 분산영향요인의 값도 변수 X1과 X3가 10보다 높게 나타났으며, 고유값(Eigenvalue)이 0.01보다 작거나 조건지표(Condition Number)가 100이상인 경우에는 다중공선성이 있다고 볼 수 있는데 여기에서도 비슷한 결과가 나타났다. 특히 분산의 분할(Var Prop 변수)에 관한 정보에서도 변수 X1과 X3가 관련성이 높은 것으로 나타났다.

다중공선성이 나타난 경우에는 X1이나 X3 중에 하나의 변수를 제거시키고 변수를 추정하던지 (여러 가지 값이 X3가 나쁜 것으로 나타나 X3를 먼저 제거시키는 것이 바람직할 것같다), 변수선정방법(2.9란 참조)에 의해 변수를 제거시키는 방법을 사용할 수 있다.

어쩔수 없이 변수간에 상관관계가 높을 수 밖에 없는 상황(예 소득과 소비가 모두 독립변수로 취급될 때)에는 RIDGE 회귀분석이나 요인분석(factor analysis)나 PROC PRINCOMP의 주성분분석(Principal Component Analysis)을 통해 다중공선성을 줄일 수 있는 방법으로 회귀분석을 수행해야 할 것이다.

자. 다중회귀모형에서 변수선택

다중회귀모형에서 회귀모형에 대한 추정결과 유의도가 없는 독립변수들이 존재하거나 다중공선성이 존재하는 경우 변수를 선택하는 방법을 사용할 수 있다. SAS에서 모형선택방법은 9가지가 있으며 이를 간단히 살펴보면

다음과 같다. MODEL문에서 SELECTION= 옵션을 지정한다.

NONE: 디폴트로 모형선택 방법을 사용하지 않고 Full 모형을 추정.

BACKWARD(후방소거법): 모든 변수가 포함된 Full모형에서 출발하여 독립변수들 중 모형이 삭제되었을 경우 회귀모형 적합에 가장 작게 기여를 하는(R^2 등을 최소로 감소시키는) 변수를 단계적으로 삭제시켜나가는 방법.

FORWARD(전방선택법): 남아있는 독립변수들 모형이 추가되었을 경우에 회귀모형 적합에 가장 큰 기여를 할 수 있는(R^2 등을 최대로 증가시키는) 변수를 단계적으로 추가시켜나가는 방법.

STEPWISE(단계별회귀법): 앞의 두 가지 방법의 개념을 합한 것으로 회귀모형의 적합을 증가시킬 수 있는 변수를 추가시키기도 하고, 일단 모형에 추가되었어도 모형의 적합에 도움이 안되는 변수는 삭제하는 방법.

ADJRSQ(조정 R^2 선택법): 모든 가능한 회귀모형에 대해 조정 R^2값을 계산.

RSQUARE(R^2 선택법): 모든 가능한 회귀모형에 대해 R^2값을 계산.

CP(맬로우스의 CP 선택법): 모든 가능한 회귀모형에 대해 CP값을 계산.

MAXR(최대 R^2 증가법): 전방선택법의 일종으로 각 단계마다 R^2를 최대로 증가시키는 변수를 추가함. 모수추정치의 갯수별로 가장 좋은 모형 1개만을 출력.

MINR(최소 R^2 증가법): 전방선택법의 일종으로 각 단계마다 R^2를 최소로 증가시키는 변수를 추가함. 모수추정치의 갯수별로 가장 좋은 모형 1개만을 출력한다.

SELECTION 옵션의 사용예를 보면 다음과 같다.

```
PROC REG DATA=COLLIN;
  MODEL Y = X1 - X4 / SELECTION=BACKWARD;
  MODEL Y = X1 - X4 / SELECTION=FORWARD;
  MODEL Y = X1 - X4 / SELECTION=STEPWISE;
  MODEL Y = X1 - X4 / SELECTION=ADJRSQ;
  MODEL Y = X1 - X4 / SELECTION=RSQUARE;
  MODEL Y = X1 - X4 / SELECTION=CP;
  MODEL Y = X1 - X4 / SELECTION=MAXR;
  MODEL Y = X1 - X4 / SELECTION=MINR;
  MODEL Y = X1 - X4 / SELECTION=NONE;
```

그러면 여기서는 BACKWARD, FORWARD, STEPWISE방법에 대해서만 그 결과를 보기로 하자.

[화일명:9-19REGR.SAS]

```
PROC REG DATA=COLLIN;
   MODEL Y = X1 - X4 / SELECTION=BACKWARD;
   MODEL Y = X1 - X4 / SELECTION=FORWARD;
   MODEL Y = X1 - X4 / SELECTION=STEPWISE;
```

출력결과 15페이지에서는 BACKWARD방법에 의한 모형선정결과이고, 출력결과 16페이지는 FORWARD방법에 의한 모형선정결과이며, 출력결과 17페이지는 STEPWISE방법에 의한 모형선정결과이다. 3가지 모형선택방법으로부터 선정된 모형은 모두 X1, X3, X4를 포함하는 식이다. 이는 단순히 COLLIN이나 VIF, TOL의 옵션을 통해 다중공선성을 본 결과와는 다르다. 아마도 이러한 현상이 나타나는 것은 관찰치의 수가 적기 때문인 것으로 생각된다. 추정된 모형은

$$Y = \underset{(113.98)}{740.97} + \underset{(0.32)}{0.81}\,X1 - \underset{(1.36)}{2.83}\,X3 + \underset{(0.23)}{2.82}\,X4$$

이며, 이때 CP 는 3.39로서 모수추정치의 갯수에 가까우며, R^2도 0.98로서 매우 높다. 모형적합성에 대한 분산분석결과도 F값이 177.82($p < 0.0001$)로서 매우 의미가 있다고 볼 수 있다.

SAS 15

Backward Elimination Procedure for Dependent Variable Y

Step 0 All Variables Entered R-square = 0.98418402 C(p) = 5.00000000

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	4	269425.07048426	67356.26762107	124.45	0.0001
Error	8	4329.69874650	541.21234331		
Total	12	273754.76923077			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	878.19855906	248.96416928	6734.09806826	12.44	0.0078
X1	0.54080616	0.54371805	535.43092283	0.99	0.3490
X2	-0.38957160	0.62229087	212.10720344	0.39	0.5487
X3	-4.08124044	2.44461367	1508.45230431	2.79	0.1336
X4	2.48661594	0.58391606	9814.86392305	18.13	0.0028

Bounds on condition number: 37.12363, 347.9256

Step 1 Variable X2 Removed R-square = 0.98340922 C(p) = 3.39191125

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	269212.96328083	89737.65442694	177.82	0.0001
Error	9	4541.80594994	504.64510555		
Total	12	273754.76923077			

(출력결과 15페이지 계속)

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	740.97038280	113.97891932	21327.43828991	42.26	0.0001
X1	0.81213719	0.31701287	3312.00571517	6.56	0.0306
X3	-2.83159141	1.36269798	2178.95376047	4.32	0.0675
X4	2.81902679	0.23457787	72880.34289117	144.42	0.0001

Bounds on condition number: 12.65102, 78.36422

All variables in the model are significant at the 0.1000 level.

Summary of Backward Elimination Procedure for Dependent Variable Y

Step	Variable Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X2	3	0.0008	0.9834	3.3919	0.3919	0.5487

Forward Selection Procedure for Dependent Variable Y

Step 1 Variable X1 Entered R-square = 0.71574346 C(p) =134.78198467

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	195938.18437880	195938.18437880	27.70	0.0003
Error	11	77816.58485197	7074.23498654		
Total	12	273754.76923077			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	538.60162257	82.43168828	302013.53534973	42.69	0.0001
X1	1.75622692	0.33370355	195938.18437880	27.70	0.0003

Bounds on condition number: 1, 1

---

* Step 2 는 생략

Step 3 Variable X3 Entered R-square = 0.98340922 C(p) = 3.39191125

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	269212.96328083	89737.65442694	177.82	0.0001
Error	9	4541.80594994	504.64510555		
Total	12	273754.76923077			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	740.97038280	113.97891932	21327.43828991	42.26	0.0001
X1	0.81213719	0.31701287	3312.00571517	6.56	0.0306
X3	-2.83159141	1.36269798	2178.95376047	4.32	0.0675
X4	2.81902679	0.23457787	72880.34289117	144.42	0.0001

Bounds on condition number: 12.65102, 78.36422

(출력결과 16페이지 계속)

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection Procedure for Dependent Variable Y

Step	Variable Entered	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X1	1	0.7157	0.7157	134.7820	27.6974	0.0003
2	X4	2	0.2597	0.9754	5.4180	105.7854	0.0001
3	X3	3	0.0080	0.9834	3.3919	4.3178	0.0675

Stepwise Procedure for Dependent Variable Y

Step 1 Variable X1 Entered R-square = 0.71574346 C(p) =134.78198467

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	1	195938.18437880	195938.18437880	27.70	0.0003
Error	11	77816.58485197	7074.23498654		
Total	12	273754.76923077			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	538.60162257	82.43168828	302013.53534973	42.69	0.0001
X1	1.75622692	0.33370355	195938.18437880	27.70	0.0003

Bounds on condition number: 1, 1

---

* Step 2 는 생략

Step 3 Variable X3 Entered R-square = 0.98340922 C(p) = 3.39191125

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	269212.96328083	89737.65442694	177.82	0.0001
Error	9	4541.80594994	504.64510555		
Total	12	273754.76923077			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	740.97038280	113.97891932	21327.43828991	42.26	0.0001
X1	0.81213719	0.31701287	3312.00571517	6.56	0.0306
X3	-2.83159141	1.36269798	2178.95376047	4.32	0.0675
X4	2.81902679	0.23457787	72880.342[illegible]17	144.42	0.0001

Bounds on condition number: 12.65102, 78.36422

All variables in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable Y

Step	Variable Entered	Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	X1		1	0.7157	0.7157	134.7820	27.6974	0.0003
2	X4		2	0.2597	0.9754	5.4180	105.7854	0.0001
3	X3		3	0.0080	0.9834	3.3919	4.3178	0.0675

SAS 기 술 통 계

목 차

1. 통계학이란?

우리는 일상생활에서 통계라고 불리는 여러가지 형태의 자료들을 접하며 살아가고 있다. 여기서 통계란 어떤 자연현상이나 사회현상의 특성을 나타내는 요약된 형태의 숫자를 의미한다.

그렇다면 통계학이란 무엇인가? 통계학이란 좁은 의미로 데이타를 다루는 학문으로 볼 수 있다. 그러나 이보다 광의의 통계학은 데이타를 얻어내는 과정을 포함한 실험을 다룬다고 볼 수 있으며, 이 과정에서 얻어진 정보(데이타)를 효율적으로 수집 정리 분석하여 불확실한 사실에 대하여 최적의 의사결정을 하기 위한 학문이라 할 수 있다.

통계실험을 통하여 얻어진 데이타는 도표나 대표적 수치(측도)로 이를 정리 요약 분석하여야 그 자료의 특성을 파악 할 수 있다. 이러한 분야를 통계학에서는 기술통계(학)이라 부른다.

통계실험은 아무리 많이 반복하여도 그 결과가 똑같을 수 없으며 일어날 수 있는 모든 가능한 결과를 관찰 할 수는 없다. 따라서 우리가 얻는 데이타는 실험을 통해 얻을 수 있는 모든 가능한 결과중 일부분만을 관찰하는 셈이다. 여기서 실험에 의해 생기는 모든 가능한 결과들의 집단을 모집단 또는 표본공간이라 한다. 또한 실험결과로 얻은 데이타는 표본이라 하며 이는 표본공간의 일부가 된다. 현대의 통계학에서는 표본(부분)을 가지고 그 표본이 속한 모집단(전체)의 특성을 파악하는 추측통계학에 주로 관심을 갖는다. 이러한 추측통계학에서는 부분을 가지고 전체를 설명하기 때문에 여기에는 반드시 잘못 판단할 오류가 수반된다. 따라서 통계적 추론은 확률이라는 객관적 척도를 이용하여 이러한 불확실성의 정도를 나타내 준다.

2. 자료의 정리

가. 자료의 종류

서울의 어느 중학교 3학년 남녀학생들의 체력에 관한 자료를 얻기 위하여 다음과 같은 양식의 설문지를 이용한다고 하자.

체력검사 조사표

1. 성별 구분은? (M:남성, F:여성)
2. 체중은 몇 Kg 인가? ( )Kg
3. 신장은 몇 cm 인가? ( )cm
4. 즐겨 먹는 음식은 무엇인가?
   (1) 육류 (2) 생선류 (3) 채소류
5. 100m 달리기 속도는?
   (1) 13초이내 (2) 13초-18초 (3) 19초이상
6. 투포환 거리는 몇 m인가? ( )m
7. 악력(握力)은 몇 Kg인가? ( )Kg

여기서 조사대상인 학생들을 관찰단위(observation) 또는 케이스(case)라고 하며 조사표의 각 항목을 변수(variable)라고 한다. 변수는 측정수준에 따라 다음의 4 종류로 구분된다.

<표 2.1> 중학 3학년학생 20명의 체력검사데이타

sex(성별)	wt(체중,Kg)	ht(신장,cm)	food(즐기는음식)
m	55.7	168	3
m	60.0	172	1
f	45.3	162	3
m	49.8	167	2
f	47.3	159	2
f	51.7	170	1
m	48.5	165	3
m	56.3	174	1
f	59.6	167	1
m	53.9	165	2
f	43.4	156	3
f	45.6	158	1
m	47.3	162	2
m	50.8	159	3
m	52.6	166	3
f	50.7	160	2
m	64.6	178	1
f	54.1	163	1
f	43.7	156	2
m	65.8	173	2

(1) 명목척도(nominal scale)변수

위의 조사표에서 항목-1 과 항목-4 에 해당되는 변수로서 변수값이 범주형으로 주어지며 범주간에 순서를 부여 할 수 없다. 또한 명목척도로 측정된 변수는 4칙연산이 불가능하다.

(2) 순서척도(ordinal scale)변수

항목-5 가 이에 해당되는 변수로 변수값이 범주형으로 주어지며 범주간에 순위가 의미있는 변수이다.

(3) 구간척도(interval scale)변수

구간척도로 측정된 변수값은 변수값 상호간의 순서뿐 아니라 차이(distance)에도 의미를 부여할 수 있다. 그러나 절대영점을 갖지않으므로 비율의 의미는 갖지 못한다. 섭씨 또는 화씨로 측정된 온도는 이 척도에 속한다.

(4) 비율척도(ratio scale)변수

항목-2,3,6,7 은 비율척도로 측정된 변수이다. 절대영점을 가지며 변수값 상호간의 순서, 차이, 비율이 의미있다. 따라서 이 척도로 측정된 변수는 실수의 모든 성질을 만족한다.

위의 체력검사조사표로부터 부분적으로 정리된 데이타의 내용은 <표 2.1>과 같다.

나. 도수분포표

원시데이타를 수집한 후 이 데이타의 특성을 파악하기 위하여는 데이타를 분류 정리하여야 하며 데이타가 갖고 있는 정보를 요약하여야 한다. 이러한 단계를 기술 통계학이라 하며, 이를 기초로 하여 데이타가 추출된 모집단에 대한 추론을 할 수 있다.

앞의 체력검사조사표에서 항목-1, 4, 5 에 해당되는 변수는 변수값을 셀

수 있으며, 항목-2, 3, 6, 7 에 해당되는 변수는 변수값을 셀 수 없다. 이와같이 변수값을 셀 수 있거나 또는 몇개의 범주로 구별되는 변수를 이산형 변수(discrete variable) 또는 범주형 변수(categorical variable) 라고 하며, 변수값을 셀 수 없는 변수를 연속형 변수(continuous variable)라고한다.

이산형 변수에 대해서는 각 변수값에 해당되는 빈도수를 세어 도수분포표를 작성할 수 있으며, 연속형 변수에 대해서는 몇개의 계급(class)으로 분류하여 그룹화(grooping)한 후 도수분포표를 작성할 수 있다. 도수분포표에서는 각각의 변수값에 대한 빈도수, 백분율, 누적백분율을 제공해 준다.

(문제2.2.1) <표 2.1>의 체력검사데이타에서 이산형 변수인 SEX 를 이용하여 성별 도수분포표를 작성하여라.

다. 도수분포도의 작성

(1) 막대그림표(barchart)

이산형 변수에 대한 도수분포표를 그림표로 작성하면 시각적인 전달효과가 높아진다. 막대그림표(barchart)는 이산형 변수의 각 변수값에 대한 빈도수 또는 백분율을 그림으로 표현한 것이다.

(문제2.3.1) <표 2.1>의 체력검사데이타에서 성별 막대그림표를 작성하여라.

(2) 기둥그림표(histogram)

연속형 변수로 된 데이타는 몇개의 계급(class)으로 분류하여 각 계급에 속하는 빈도수, 백분율 등을 작성함으로써 데이타가 갖고있는 정보를 정리요약할 수 있다. 이와같이 연속형 변수에 대한 계급별 빈도수 또는 백분율을 그림으로 나타낸 것이 기둥그림표(histogram)이다.

기둥그림표를 작성하기 위하여는 계급의 수(number of classes), 구간의 폭(width of interval), 각 구간의 경계값(bondary of interval)을 정해야 한다. 계급의 수는 관찰단위의 수를 고려하여 정해야 하며 일반적으로 5-20 정도의 값을 택하면 된다. 구간의 폭은 계급의 수에 따라 등간격으로 정하면 되며, 데이타의 최소값을 포함하도록 계급의 하한치를 정한 후 각 구간

의 경계값을 정한다. 이때 각 구간의 대표값은 구간의 중앙값(midpoint)으로 계산한다.

(문제2.3.2) <표 2.1>의 체력검사데이타에서 전체 학생의 체중에 대한 기둥그림표를 작성하여라.

라. 줄기와 잎 그림표(stem-and-leaf diagram)

연속형 자료를 그림으로 요약표현하는 방법의 하나이다. 관측치의 상위 자리수를 줄기로 하고 하위자릿수를 잎으로 하여 줄기에 잎을 달아나가는 방식으로 자료를 요약 정리하여 자료의 전체적인 분포양상과 대칭성의 유무, 각 구간에 속하는 관측치의 분포까지도 알아볼 수 있는 탐색적 자료분석 방법이다.

<표2.1>의 자료에서 신장에 관한 줄기와 잎 그림표를 작성하면 다음과 같다. 여기서 줄기의 '·' 는 일자리 숫자가 0 에서 4까지를 의미하며, '*'는 5에서 9까지를 의미한다.

줄기	잎
15·	
15*	96896
16·	2203
16*	875756
17·	2043
17*	8

(문제2.4.1) 다음의 자료는 어느학급의 통계학점수에 대한 자료이다. 이를 줄기와 잎 그림표로 요약하여라

75 59 77 66 82 63 80 49 76 77 60 86 85 80 95 63 53 72

69 70 68 58 90 74 81

마. 측도에 의한 자료의 기술

막대그림표나 기둥그림표를 통해서 분포의 모양(shape of distribution)을 파악할 수도 있으나, 분포의 특성(characteristics)을 단일한 수치로 나타냄으로써 데이타가 갖고있는 분포에 대한 정보를 보다 압축된 형태로 표현할 수 있다. 이와같이 얻어진 표본의 특성치를 통계량(statistic)이라 한다.

(1) 중심경향의 측도(measure of central tendency)

표본의 중심위치에 대한 정보를 주는 측도로는 평균(mean), 중앙값(median), 최빈값(mode)이 있다.

가. 평균(mean)

관측치들의 합을 전체표본의 갯수로 나누어준 것을 평균이라 한다. 예를 들어 주어진 관측치들이 $x_1, x_2, \ldots, x_n$ 일때, 평균은

$$\text{평균}(x) = \Sigma x_i / n$$

으로 계산된다.

나. 중앙값(median)

관측치들을 작은 값부터 크기순으로 배열 했을때 이들의 중앙에 위치하는 값을 중앙값이라한다. 예를들어 <표2.1>의 자료에서 처음 7명 신장의 중앙값을 구해보자. 처음 7명의 신장 데이타를 크기순으로 배열하면

159 162 165 167 168 170 172

이므로 중앙값은 167이다. 자료의 갯수가 짝수일 때는 가운데 놓인 두개의 관측치를 평균하여 중앙값으로 정한다.

다. 최빈수(mode)

자료중 가장 빈도가 높은 값을 최빈수라 한다.

(2) 산포의 측도(measure of dispersion)

표본의 흩어진 정도를 재는 측도로 분산(variance), 범위(range) 등이 이용된다.

분산 $(S^2) = \Sigma (X_i - \bar{X}) / (n-1)$, n=표본의 크기

범위 (R) = 최대값 - 최소값

분산의 제곱근을 표준편차(standard deviation, S)라 하며 이는 표본의 단위와 같은 단위로 측정된 산포의 측도가 된다.

두개의 표본집단간의 상대적인 산포를 비교할 때는 변이계수(coefficient of variation, CV)를 이용한다.

$CV = S/\bar{X}$, (S= 표준편차, $\bar{X}$= 평균)

(3) 비대칭도(skewness)

분포의 모양이 대칭(symmetry)을 벗어나 어느 한쪽으로 기울어진 정도를 재는 측도로 3차적률을 이용하여 계산된다.

$$\text{비대칭도(SK)}=\frac{\Sigma\ (X_i-\bar{X})^3}{nS^3}$$

분포가 대칭이면 SK=0 이며, 분포가 오른쪽으로 기울어진(right-skewed) 경우에는 SK>0 이 된다.

(4) 첨도(kurtosis)

분포의 뾰족한 정도를 나타내는 측도로 4차적률을 이용하여 계산된다.

$$\text{첨도(KT)}=\frac{\Sigma(X_i-\bar{X})^4}{nS^4}-3$$

정규분포의 경우 KT=0 이며, KT>0 이면 일반적으로 정규분포보다 뾰족한 모양을 한다.

(5) 표준화 점수(standard score)

관찰단위의 변수값 X_i에 대해 그 분포내에서의 상대적인 위치가 어느정도인지를 기술하고자 할때, 다음과 같은 표준화 점수 Z_i를 계산하여 이용한다.

$$Z_i=(X_i-X)/S$$

(6) 백분위수(percentile)와 n분위수(n-tile)

데이타를 오름차순으로 정렬한 후 백등분 하였을때 등분된 값을 백분위수라 한다. 예를들면 25백분위수란 그 값보다 같거나 작은 데이타의 비율이 25%인 점을 의미하며, 50백분위수는 중앙값(median)을 의미한다.

n분위수(n-tile)란 데이타를 오름차순으로 정렬한 후 n 등분하였을 때 등분된 값을 나타낸다. 예를들어 4분위수는 오름차순 데이타의 4등분된 값을 의미하며 25, 50, 75백분위수와 같은 결과가 됨을 알 수 있다.

(문제2.5.1) <표2.1>의 체력검사데이타에서 신장에 대한 모든 가능한 통계량을 구하여라.

3. SAS의 활용

가. SAS 란?

SAS System의 단위 Software

SAS/CORE

SAS/BASE

SAS/STAT

SAS/IML

SAS/GRAPH

SAS/ETS

SAS/OR

SAS/QC, 등

나. SAS의 시작

C:\>cd sas

C:\sas>sas

다. SAS 화면관리 씨스템(Display Manager System)

OUTPUT Window

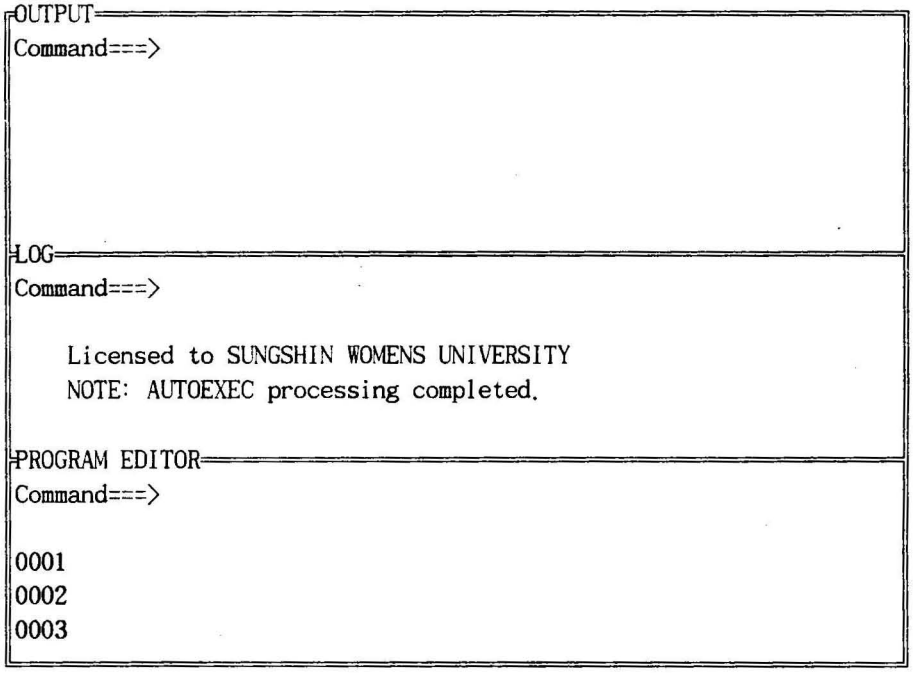
LOG Window

PROGRAM EDITOR Window

* SAS 화면관리 씨스템에서 쓸수 있는 명령어는 반드시 Command Line에서 입력시켜야 한다.
Command===>

* Command Line 으로 되돌아 가고 싶을때..... HOME Key

* Cursor 가 있는 Window를 확대하고 싶을때....F7 (zoom key)



라. SAS 프로그램의 편집 및 실행

* SAS 프로그램의 작성은 PGM Window에서 입력
  Command Line 에서 ENTER Key를 쳐서
  커서를 줄번호 00001로 이동
  필요시 F7(zoom key)를 친다

* SAS 프로그램 작성규칙

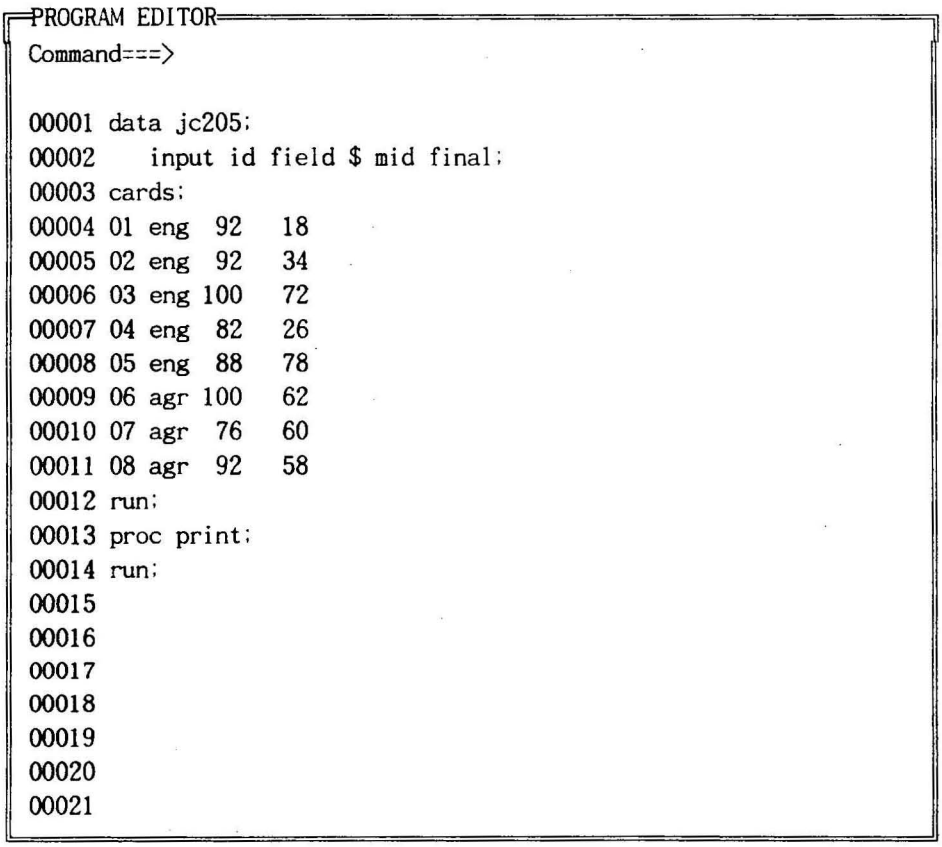
* SAS 프로그램의 실행

  function key F10 (submit key)을 누른다.
  실행결과는 OUTPUT Window에
  처리과정 및 오류상황은 LOGWindow에

* function key(기능키)의 활용

  F1...help
  F2...key window
  F3...log window
  F4...output window
  F5...next

F6...program window
F7...zoom
F8...subtop
F9...recall
F10...submit



* 각 window의 내용(프로그램 또는 출력결과)을 저장하고 싶을때

커서를 commmand line으로 보낸후 (home key 이용)
command line===> file 'filename'
예를 들면
command line===> file 'c:\sas\out\test.out'
command line===> file 'c:\sas\pgm\test.pgm'

* window의 내용을 프린터로 인쇄하고 싶을 때

command line===> file 'prn'

* window의 내용을 지우고 싶을 때

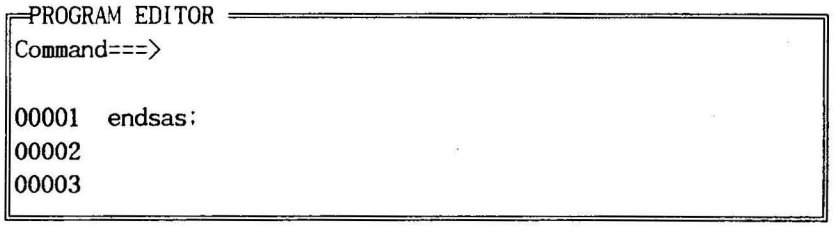
command line===> CLEAR

* 디스켓 또는 hard disk 에 있는 화일을 window로 호출

command line===> INCLUDE 'filename'
예를들면
command line===> INCLUDE 'c:\class_3\sample.pgm'

마. SAS 작업의 종료

PROGRAM EDITOR window 로 돌아간다(기능키 F6 이용)
command line===> BYE
또는
command line===> ENDSAS
또는
command line===> enter 키를 쳐서
커서를 00001 로 옮긴후
ENDSAS; 를 입력하고
기능키 F10을 눌러 실행한다



4. SAS작업의 예

* SAS작업의 구성

DATA --- PROC ---PROC --- DATA --- PROC ---

* PROC단계의 일반적 형식

PROC 모듈이름 [DATA=SASdsn] [options];
이 모듈과 함께하는 SAS문;
(예) PROC PRINT DATA=jc205;
PROC PRINT;

가. PROC UNIVARIATE

일변량 기술통계량 및 QUANTILE, 줄기와 잎 그림표, 상자그림표, 정규확률그림 등이 출력

(형식)

```
PROC UNIVARIATE options;
     사용되는 options들
       DATA=SASdsn
       PLOT....stem and leaf plot, box plot, normal prob plot
       FREQ....자료의 도수표
       NORMAL...정규성 검정 통계량
  VAR varlist;
  BY varlist;
  FREQ var;
  ID varlist;
  OUTPUT OUT=SASdsn keyword=names...;
```

(사용예)

```
OPTIONS PAGESIZE=60;
DATA JOB_1;
       INFILE 'C:\DATA\BODY.DAT':
       INPUT dist grip wt ht;
PROC PRINT;
PROC UNIVARIATE PLOT DATA=job_1;
       VAR wt ht;
```

***** FILENAME문과 INFILE문**********

```
FILENAME KIM 'C:\DATA\BODY.DAT';
INFILE KIM;
```

(출력결과)

SAS 15:19 Tuesday, November 1, 1994 1

UNIVARIATE PROCEDURE

Variable=WT

Moments

N	20	Sum Wgts	20
Mean	52.335	Sum	1046.7
Std Dev	6.498848	Variance	42.23503
Skewness	0.614203	Kurtosis	-0.29432
USS	55581.51	CSS	802.4655
CV	12.41779	Std Mean	1.453187
T:Mean=0	36.01396	Prob>|T|	0.0001
Sgn Rank	105	Prob>|S|	0.0001
Num ^= 0	20		

Quantiles(Def=5)

100% Max	65.8	99%	65.8
75% Q3	56	95%	65.2
50% Med	51.25	90%	62.3
25% Q1	47.3	10%	44.5
0% Min	43.4	5%	43.55
		1%	43.4
Range	22.4		
Q3-Q1	8.7		
Mode	47.3		

Extremes

Lowest	Obs	Highest	Obs
43.4(	11)	56.3(	8)
43.7(	19)	59.6(	9)
45.3(	3)	60(	2)
45.6(	12)	64.6(	17)
47.3(	13)	65.8(	20)

UNIVARIATE PROCEDURE

Variable=WT

Stem	Leaf	#	Boxplot
6	56	2	¦
6	00	2	¦
5	66	2	+-----+
5	0112344	7	*--+--*
4	56778	5	+-----+
4	34	2	¦

----+----+----+----+

Multiply Stem.Leaf by 10**+1

나. PROC PLOT

두변수간의 SCATTER PLOT을 그리는 절차로서, Overlay plot, contour plot등을 출력

(형식)

PROC PLOT options;

* 사용되는 option들
DATA=SASdsn

VPERCENT=values

HPERCENT=values

BY varlist;

PLOT 수직축변수*수평축변수[=변수 또는 '문자']/options;

* PLOT문에서 사용되는 option들

VAXIS=values HAXIS=values(예: VAXIS=0 to 50 by 10)

VREF=values HREF=values (예:VREF=10 20 30)

VPOS=n HPOS=n

BOX

OVERLAY

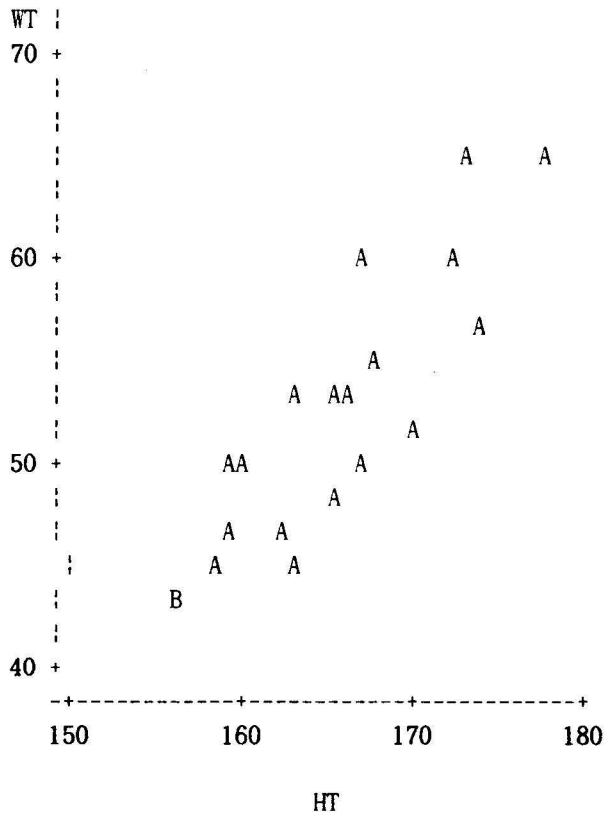
CONTOUR=value

(사용예-1)

```
PROC PLOT DATA=job_1;
        PLOT et*ht/vpos=20 hpos=40;
RUN;
```

SAS 15:04 Wednesday, November 2, 1994
3

Plot of WT*HT. Legend: A = 1 obs, B = 2 obs, etc.



***** 영구적 SAS dataset 만드는법******

```
LIBNAME myproc 'c:\class_3';
DATA myproc.test;
        INFILE 'c:\tongae\data\body.dat';
        INPUT dist grip wt ht;
RUN;
```

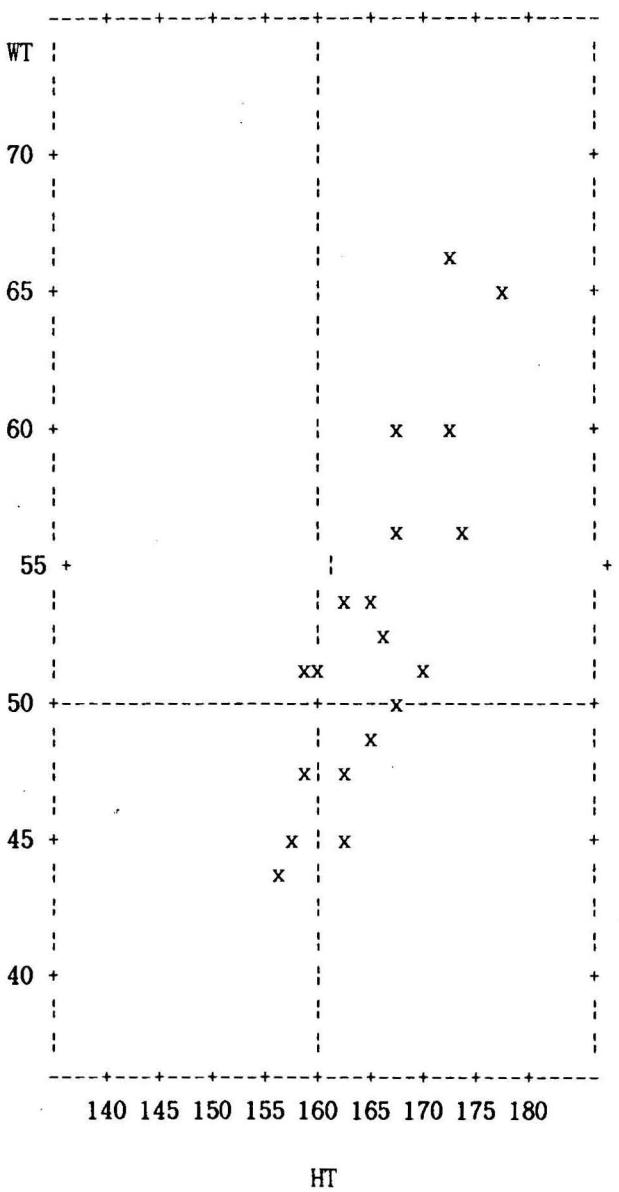
***** 영구 SAS dataset으로부터 새로운 dataset 형성 ******

```
LIBNAME my 'c:\class_3';
DATA SAM_1;
   GET my.test;
```

(사용예-2)

```
PROC PLOT DATA=SAM_1;
        PLOT wt*ht='x'/ HAXIS=140 TO 180 BY 5
                        VAXIS=40 TO 70 BY 5
                        BOX
                        HREF=160 VREF=50;
RUN;
```

Plot of WT*HT. Symbol used is 'x'.

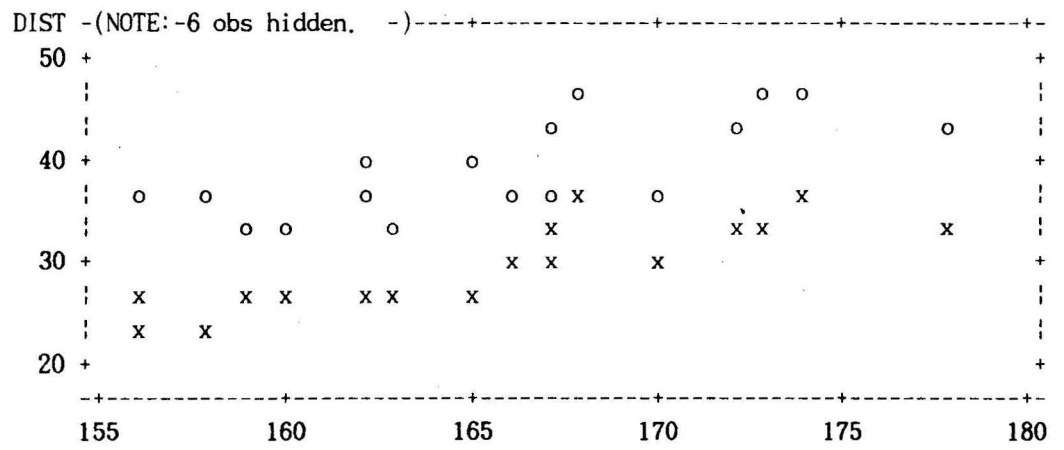


(사용예-3)

```
      LIBNAME MY 'C:\CLASS3';
      DATA SAMPLE-1;
      SET MY.TEST;
      PROC PLOT DATA=SAMPLE-1;
***** 또는 위의 4문장 대신에***************
      LIBNAME MY 'C:\CLASS3';
      PROC PLOT DATA=MY.TEST; <==== C:\CLASS3\TEST.SSD
*****************************************
            PLOT DIST*HT='X' GRIP*HT='O'/BOX
                                  OVERLAY;
      RUN;
```

SAS 15:04 Wednesday, November 2, 1994 10

Plot of DIST*HT. Symbol used is 'x'.
Plot of GRIP*HT. Symbol used is 'o'.



SAS 15:04 Wednesday, November 2, 1994

(사용예-4)

```
PROC PLOT HPERCENT=50 VPERCENT=50;
        PLOT WT*HT='X';
        PLOT WT*HT='+' /HAXIS=150 TO 175 BY 5;
        PLOT WT*HT='O' /BOX;
        PLOT DIST*WT='X' GRIP*WT='O' /OVERLAY;
RUN;
```

```
                                    SAS      15:04 Wednesday, November 2, 1994
                                                                          12

 Plot of WT*HT.  Symbol used is 'x'.
WT  (NOTE: 2 obs hidden.)
80 +
60 +           xx  x xxxx x xxx   x
40 +        x xx  x  x x
   --+---------+---------+---------+-
    150       160       170       180
                      HT

 Plot of WT*HT.  Symbol used is '+'.
WT  (NOTE: 3 gone.)
80 +
60 +              ++   + +++ + + + ++
40 +         +  ++  +   + +
   --+-----+-----+-----+-----+-----+-
    150   155   160   165   170   175
                      HT

 Plot of WT*HT.  Symbol used is 'o'.

WT -(NOTE:-5 obs hidden.  -)------+--
80 +                       oo    o   +
40 +------o-ooo-oo-oooo-o---o-------+
   -+---------+---------+---------+--
   150       160       170       180
                      HT

Plot of DIST*WT.  Symbol used is 'x'.
Plot of GRIP*WT.  Symbol used is 'o'.
DIST  (NOTE: 23 obs hidden.)
  50 +   oxxox xxxxxx x   x    xx
   0 +   x  x
     -+---------+---------+---------+
      40        50        60        70
                      WT
```

******* OUTPUT문의 용법 **********

한개의 입력line으로부터 여러개의 관찰단위를 만들때, 하나의 data step에서 여러개의 sas data set을 만들때, 입력자료의 정보를 결합할때, 등

예1.
```
DATA TEST; INPUT REP X1 X2 X3;

DROP X1-X3;

TRT=1; VALUE=X1; OUTPUT;

TRT=2; VALUE=X2; OUTPUT

TRT=3; VALUE=X3; OUTPUT;

CARDS;
1 20 25 26
2 15 17 13
;

PROC PRINT;
```

(결과)

OBS	REP	TRT	VALUE
1	1	1	20
2	1	2	25
3	1	3	26
4	2	1	15
5	2	2	17
6	2	3	13

예2.
```
DATA all male female;
INPUT sex $ 1-2 id 4-5 grade 7 mid 9-11 .1;
  IF sex=' ' THEN DO;
                    LIST;
                    DELETE;
                  END;
OUTPUT all;
  IF sex='m' THEN OUTPUT male;
  IF sex='f' THEN OUTPUT female;
CARDS;
   11 a 456
 f 22 b 658
 m 33 b 854
 m 44 a 379
; PROC PRINT DATA =all; PROC PRINT DATA=male;
PROC PRINT DATA=female;
```

(각 SASdatset의 내용)

dataset ALL	OBS	SEX	ID	GRADE	MID
	1	F	22	B	65.8
	2	M	33	B	85.4
	3	M	44	A	37.9

dataset male	OBS	SEX	ID	GRADE	MID
	1	M	33	B	85.4
	2	M	44	A	37.9

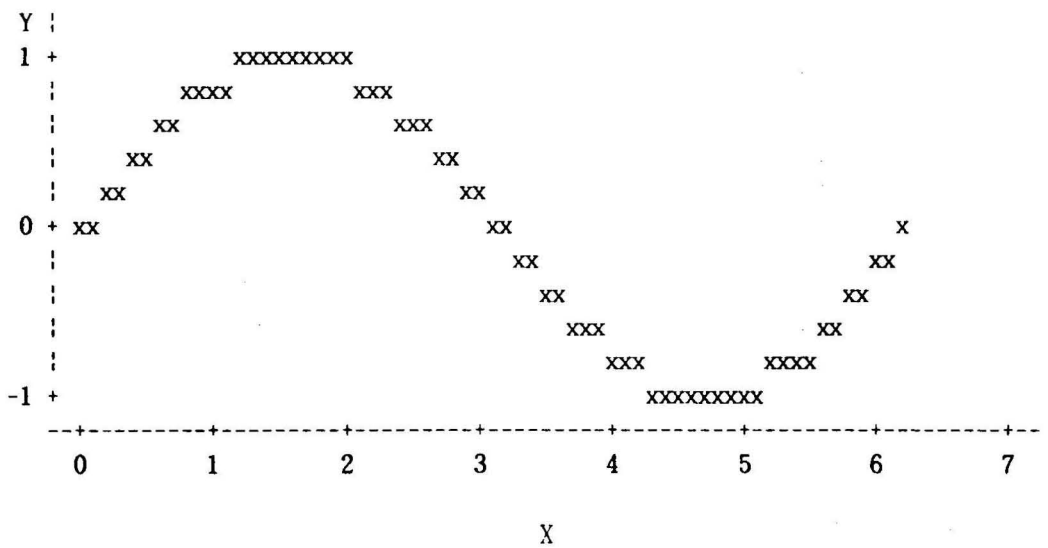
dataset female	OBS	SEX	ID	GRADE	MID
	1	F	22	B	65.8

(사용예-5) 함수의 그래프

```
DATA sineplot;
   pi=3.14159265359;

   DO x=0 TO 2*pi BY 0.1;
      y=SIN(X);

      OUTPUT;

   END;

RUN;

PROC PLOT DATA=sineplot;
   PLOT y*x='x'; RUN;
```

SAS 15:04 Wednesday, November 2, 1994
13

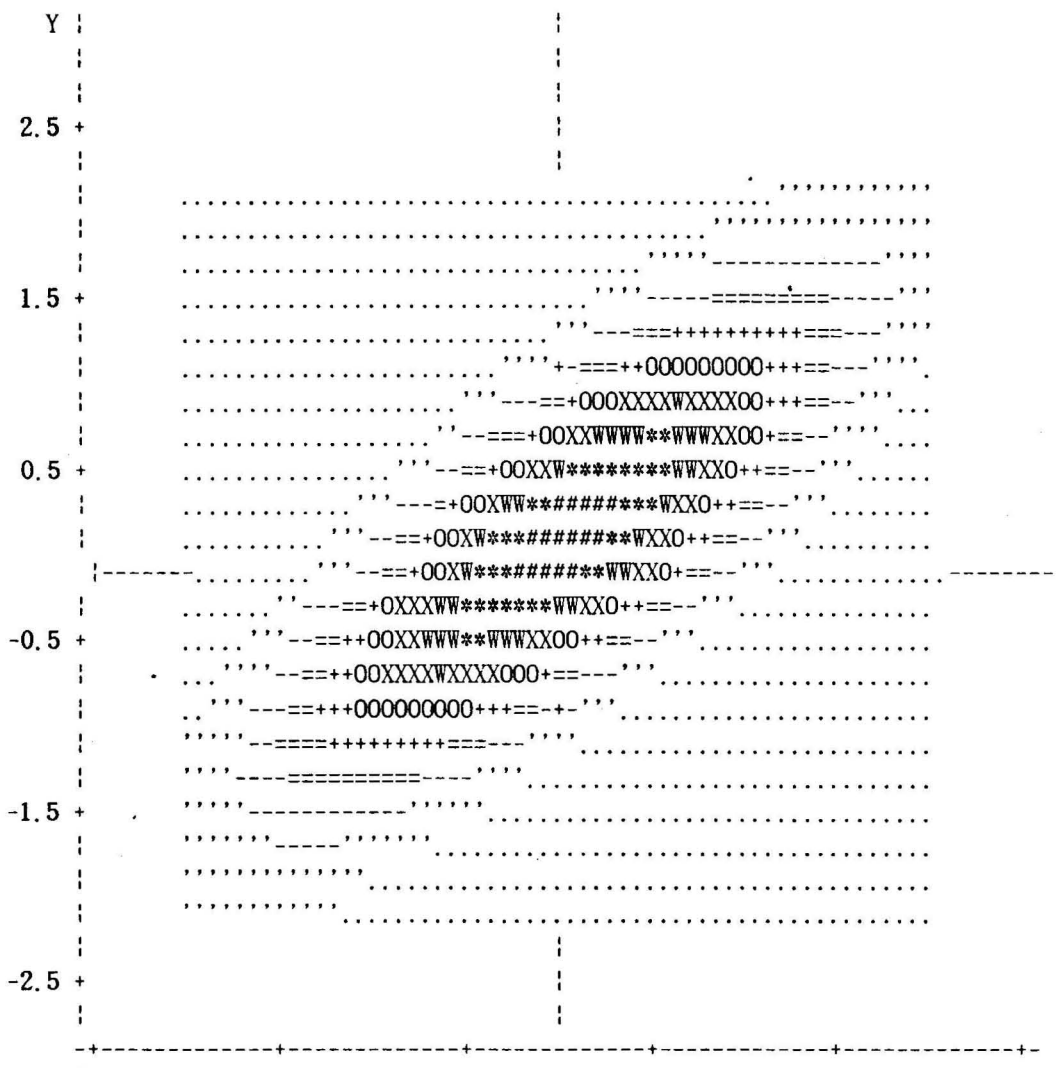
Plot of Y*X. Symbol used is 'x'.



(사용예-6) 등고선그림

```
DATA  normal;

   FORMAT z 5.2;
   pi=3.14159;
   r=0.8;

   DO x=-2 TO 2 BY 0.05;

      DO y= -2 TO 2 BY 0.05;

           z= 1/(2*pi*sqrt(1-r*r))
              *EXP(-(x*x-2*r*x*y+y*y)/2/(1-r*r));

           OUTPUT;

      END;

   END;  RUN;

PROC PLOT DATA=normal;

   PLOT y*x=z / HAXIS=-2.5 TO 2.5

                VAXIS=-2.5 TO 2.5

                HREF=0

                VREF=0

                CONTOUR=10; RUN;
```

Contour plot of Y*X.



Symbol	Z	Symbol	Z	Symbol	Z
.....	0.00 - 0.03	+++++	0.11 - 0.14	*****	0.22 - 0.25
'''''	0.03 - 0.05	00000	0.14 - 0.17	#####	0.25 - 0.28
-----	0.05 - 0.08	XXXXX	0.17 - 0.19		
=====	0.08 - 0.11	WWWWW	0.19 - 0.22		

NOTE: 5307 obs hidden.

**** INPUT 문에서 @의 용법 ****

(예1)

```
DATA one ; INPUT x;
CARDS;
1 2 3
4 5
6 4
;
PROC PRINT;
 RUN;
```

결과

OBS	X
1	1
2	4
3	6

(예2)

```
DATA two; INPUT x y @@;
CARDS;
12 43 13 36 15 35
21 65
; PROC PRINT; RUN;
```

결과

OBS	X	Y
1	12	43
2	13	36
3	15	35
4	21	65

(예3)

```
DATA three; INPUT x @;
INPUT Y; CARDS;
3 5
6 9
; PROC PRINT;
RUN;
```

결과

OBS	X	Y
1	3	5
2	6	9

다. PROC SORT

SAS dataset을 지정된 변수(들)의 관측값에 따라 오름차순 또는 내림차순으로 분류할 때 사용, 분류된 dataset은 PROC PRINT를 이용하여 출력

(형식)

```
PROC SORT [DATA=SASdsn] [OUT=SASdsn];
   BY [DESCENDING] varlist;
```

(예-1)

```
LIBNAME my 'c:\class3';
PROC SORT DATA=my.sam OUT=test;
  BY wt; RUN;
PROC PRINT DATA=test; RUN;
```

(예-2)

```
PROC PRINT DATA=body;
PROC SORT DATA=body;
   BY sex DESCENDING wt;
PROC PRINT DATA=body;
RUN;
```

**** SAS dataset의 관리(복사, 결합 등) ****

*** SET문의 활용 ***
기존의 dataset으로부터 복사, 일부 발췌, 또는 두개 이상의 dataset을 (끼워)연결할 때 사용

예1. DATA new; SET old;

예2.

```
DATA new; SET old; DROP x y;
DATA new; SET old; KEEP a b;
DATA new; SET old; IF sex='m';
```

예3.

```
DATA all; SET male fem; PROC PRINT DATA=all;
male: OBS  SEX  RANK
      1    m    3
      2    m    4
fem:  OBS  SEX  RANK
      1    f    1
all:  OBS  SEX  RANK
      1    m    3
      2    m    4
      3    f    1
```

예4.

```
PROC SORT DATA=male; BY rank;
PROC SORT DATA=fem; BY rank;
DATA all; SET male fem; BY rank;
PROC PRINT DATA=all;
```

*** MERGE문의 활용 ***

대응하는 관찰단위에 대해 새로운 변수들을 추가로 합성하고자 할때

예1. 1:1 Merging

```
DATA all; MERGE set1 set2;
PROC PRINT DATA=all;
```

set1:

OBS	X1	X2	X3
1	1	2	3
2	6	4	8
3	7	6	9

set2:

OBS	X4
1	4
2	5

all:

OBS	X1	X2	X3	X4
1	1	2	3	4
2	6	4	8	5
3	7	6	9	.

예2. BY변수를 이용한 Merging

```
PROC SORT DATA=aa;BY id;
PROC SORT DATA=bb;BY id;
DATA cc; MERGE aa bb; BY id;
PROC PRINT DATA=cc;RUN;
```

aa:

OBS	ID	X
1	1	3.2
2	2	5.1
3	3	4.0
4	5	3.2

bb:

OBS	ID	Y
1	1	6.1
2	2	9.5
3	4	8.3
4	5	11.6

cc:

OBS	ID	X	Y
1	1	3.2	6.1
2	2	5.1	9.5
3	3	4.0	.
4	4	.	8.3
5	5	3.2	11.6

*** UPDATE 문의 활용 ***

master file을 transaction file로 update할때

예.

```
DATA new;UPDATE aa bb; BY subj;
PROC PRINT DATA=new;
```

aa:

OBS	SUBJ	X1	X2
1	1	5	2
2	2	4	1
3	3	3	1
4	4	7	

bb:

OBS	SUBJ	X2
1	1	5
2	3	4
3	4	8
4	5	2

new:

OBS	SUBJ	X1	X2
1	1	5	5
2	2	4	1
3	3	3	4
4	4	7	8
5	5		2

라. PROC MEANS

숫자변수에 대한 기술통계량을 선택적으로 출력할때

(형식)

```
PROC MEANS [DATA=SASdsn] options;
   ** 사용되는 option들 **
      NOPRINT
      FW=n      <----- 디폴트는 12
      MAXDEC=m
       출력하고자 하는 통계량(OUTPUT문에도 사용됨)
       N*   NMISS  MEAN*  STD*  MIN*  MAX* RANGE SUM  VAR
       USS    CSS   STDERR  CV  SKEWNESS   KURTOSIS
      T (t 통계량 Ho:평균=0)
      PRT (통계량 T에 의한 양측 p-value)
   VAR varlist;
   BY varlist;   <--- 사전에 sort되야함
   CLASS varlist; <---사전에 sort될 필요없음
   FREQ var;
   WEIGHT var;
   OUTPUT OUT=SASdsn 통계량=namelist
                     통계량=namelist ...;
```

(예)

```
PROC MEANS DATA=my.sample N MEAN STD;
   CLASS sex;   VAR ht wt;
   OUTPUT OUT=stat1 MEAN=htave wtave
                    STD=sdht sdwt;
```

**** 산술할당문 ****

(형식) 변수= SAS식;

(예)
```
age=wt/plant*0.25;  total=total+x;
DATA one; INPUT x y; z=x+y; CARDS;
chr='abc';
```

**** 누적(sum) ****

(형식) 변수+SAS식; 또는
변수=SUM(변수리스트);

(예)
```
sum+x; sumsq+x*x;    sumvar=SUM(x1-x5);
```

**** IF문 ****

(형식-1) IF 논리식 THEN 명령문;

(예)
```
IF trt=5 THEN score=score-10;
```

(형식-2) IF 논리식 THEN 명령문;
ELSE 명령문;

(예) 1)
```
IF time GT 7 THEN class=2;
   ELSE class=1;
```

2)
```
IF(25< age <= 30) THEN group='teen';
   ELSE group='old';
```

(형식-3) IF 논리식; <---논리식이 참인 관찰단위만 추출하여 dataset을 구성

(예)
```
1) IF trt=5;
2) IF sex='f';
3) IF profit >= 250;
```

마. PROC FREQ

범주형 자료분석을 위한 procedure로서 독립성검정 또는 동질성검정을 수행 할 수 있다.

(형식) PROC FREQ options;
TABLES requests/options;
WEIGHT var;
BY var;

1. PROC FREQ문에서의 옵션
DATA=SASdsn

2-1. TABLES문의 requests
1) x*y 2) class*x*y 3) (x y)*(a b)

2-2. TABLES문에서의 옵션
LIST... 분할표가 아닌 도수표 형식 출력
OUT... PROC 결과를 dataset으로 출력
CHISQ... 카이제곱 통계량
MEASURES... 연관성측도
EXPECTED
DEVIATION
CELLCHI2
NOFREQ NOPERCENT NOROW NOCOL
NOCUM NOPRINT

*** PUT문의 용법 ***

PUT varlist; <--- SASlog에 출력
FILE PRINT;PUT varlist;<---OUTPUT window에 출력
FILE filespec; PUT varlist;<--- filespec에 출력
(비교) INFILE---INPUT문, INPUT---CARDS문

*** DO문의 활용 ***

1)DO--END문

```
data doexam;
  infile 'c:\tongae\data\body.dat';
  input grip dist wt ht;
if (wt gt 55) then do; class=2; group='heavy';
                    end;
else do; class=1;group='light';
      end;
put _all_;
run;
proc print;run;
```

(결과)

OBS	GRIP	DIST	WT	HT	CLASS	GROUP
1	38	48	55.7	168	2	heavy
2	33	44	60.0	172	2	heavy
3	26	41	45.3	162	1	light
4	29	36	49.8	167	1	light
5	27	34	47.3	159	1	light

2)DO index=start TO stop BY increment-- END

```
data random(drop=n);
  do n=1 to 10;
    x=UNIFORM(0);
    output;
  end;
proc print data=random;
run;
```

3)DO--WHILE--END

```
data;
n=0;
do while(n<3);
   nn=n*n;
   put n nn;
   n+1;
end;
```

4)DO--UNTIL--END

```
data;
  n=0;
  do until(n>=3); nn=n*n; put n nn;
     n+1;
  end;
run;
```

*** SELECT-WHEN/OTHERWISE-END문의 활용***

```
DATA stest; INPUT a x y @@;
  SELECT (a);
    WHEN (1) z=x+y;
    WHEN (2) z=x-y;
    OTHERWISE z=x*y;
  END;
CARDS;
1 2 3     2 3 4     3 5 7
;RUN;
```

*** GOTO문의 활용***

```
DATA class; INPUT x y @@;
  IF y=0 THEN GOTO noway;
  z=x/y;
  noway: q=x+y;
CARDS;
1 2    3  0
;RUN;
```

출력

X	Y	Z	Q
1	2	0.5	3
3	0	.	3

```
DATA class; INPUT x y;
  IF y=0 THEN GOTO noway;
  z=x/y;
  RETURN;
  noway: q=x+y;
CARDS;
1 2   3  0
; RUN;
```

출력

X	Y	Z	Q
1	2	0.5	.
3	0	.	3

**** ARRAY문의 용법 ****

(형식) ARRAY 배열이름{n|*} [$] 배열원소변수들;

(예)

```
DATA sample;
   ARRAY Y(4) X1-X4;
   INPUT X1-X4;
     cnt=0;
     DO i=1 TO 4;
       IF y(i) ne . THEN cnt+1;
     END;
   DROP i;
CARDS;
1 2 . 4
2 3 . .
4 5 6 8
. . . 5
;
PROC PRINT; RUN;
```

(결과)

OBS	X1	X2	X3	X4	CNT
1	1	2	.	4	3
2	2	3	.	.	2
3	4	5	6	8	4
4	.	.	.	5	1

**** %INCLUDE문의 용법 ****

외부 화일에 보관된 SAS프로그램을 호출하여 일괄처리 할때 사용

(형식) %INCLUDE fileref; <---fileref는 FILENAME문에 의해 지정됨

또는

%INCLUDE 'filename';

(비교) INCLUDE 'filename' command는 PGM EDITOR window의 command line(===>)에서 사용

**** COMMENT 문 ****

SAS 프로그램에 대한 주석을 달고자 할 때 이용

(형식)

* message;

또는

/* message */