전산통계학 실습자료
(SAS, R언어 프로그래밍)

---

1. SAS 설치 및 실행, 파일 다루기

2. SAS 기본 명령어, 연산자 및 함수

3. SAS 프로시저 및 기술통계량

4. SAS 그래프 출력

5. SAS 통계분석

1. SAS 설치 및 실행, 파일 다루기

1-1. 통계 프로그래밍 언어

- 수집한 데이터를 손쉽게 다루고 분석 결과물을 산출해내는 것을 도와주어 통계에 특화된 프로그래밍 언어를 일컫는다. 대표적으로 SAS, R 등의 통계 프로그래밍 언어가 있다.
- 과거 통계를 위해 SPSS라는 프로그램이 있었으나, 정해진 분석 결과물을 나타내었다. 따라서 사용자들이 원하는 결과를 직접 얻는 것을 목적으로 통계 프로그래밍 언어가 개발되었다.

1-2. SAS (Statistical Analysis System)

- SAS 연구소에 의하여 개발된 통계분석 패키지
- C와 JAVA 등의 컴퓨터 언어를 배우지 않아도 사용 가능
- 대용량 데이터의 처리에 높은 성능을 가짐

1-3. SAS University Edition 설치

- 아래 링크를 따라 가입/로그인

https://support.sas.com/edownload/software/DPUNVE001_VirtualBox

- unvbasicvapp*.ova 최신버전 파일이 자동으로 다운로드
- 자동으로 다운로드 되지 않을 경우 아래 링크로 직접 다운로드

http://www.sas.com/ko_kr/software/university-edition.html

1-4. VirtualBox 설치

- 아래 링크를 따라 최신 버전의 VirtualBox를 다운로드

https://www.virtualbox.org/wiki/Downloads

- 운영체제에 맞도록 다운로드

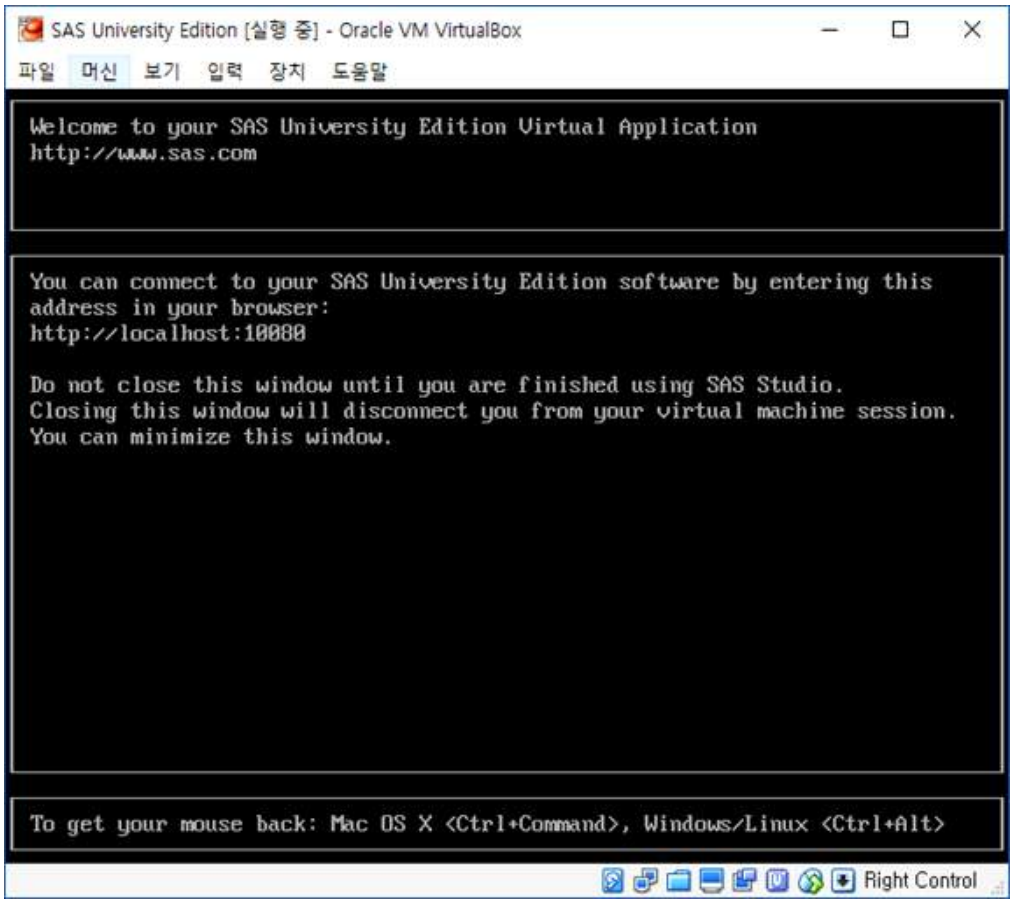
1-5. VirtualBox에 SAS 설치

- VirtualBox 실행
- File > Import Application (가상 시스템 가져오기) 클릭
- 다운로드 받은 unvbasicvapp*.ova 파일 선택하여 설치 진행
- VirtualBox 목록에 SAS University Edition이 생성되면 완료

1-6. VirtualBox 공유 폴더 설정

- 원하는 위치에 'SASUniversityEdition' 폴더를 생성하고, 해당 폴더 안에 'myfolders' 하위 폴더를 생성
- VirtualBox 목록에서 SAS University Edition 우측 클릭 > 설정
- 공유 폴더 탭 > 폴더 추가 아이콘 > 미리 생성한 'myfolders' 폴더 선택
- '읽기 전용'은 체크하지 않고, '자동 마운트'를 체크하고 확인하면 완료

1-7. SAS 실행

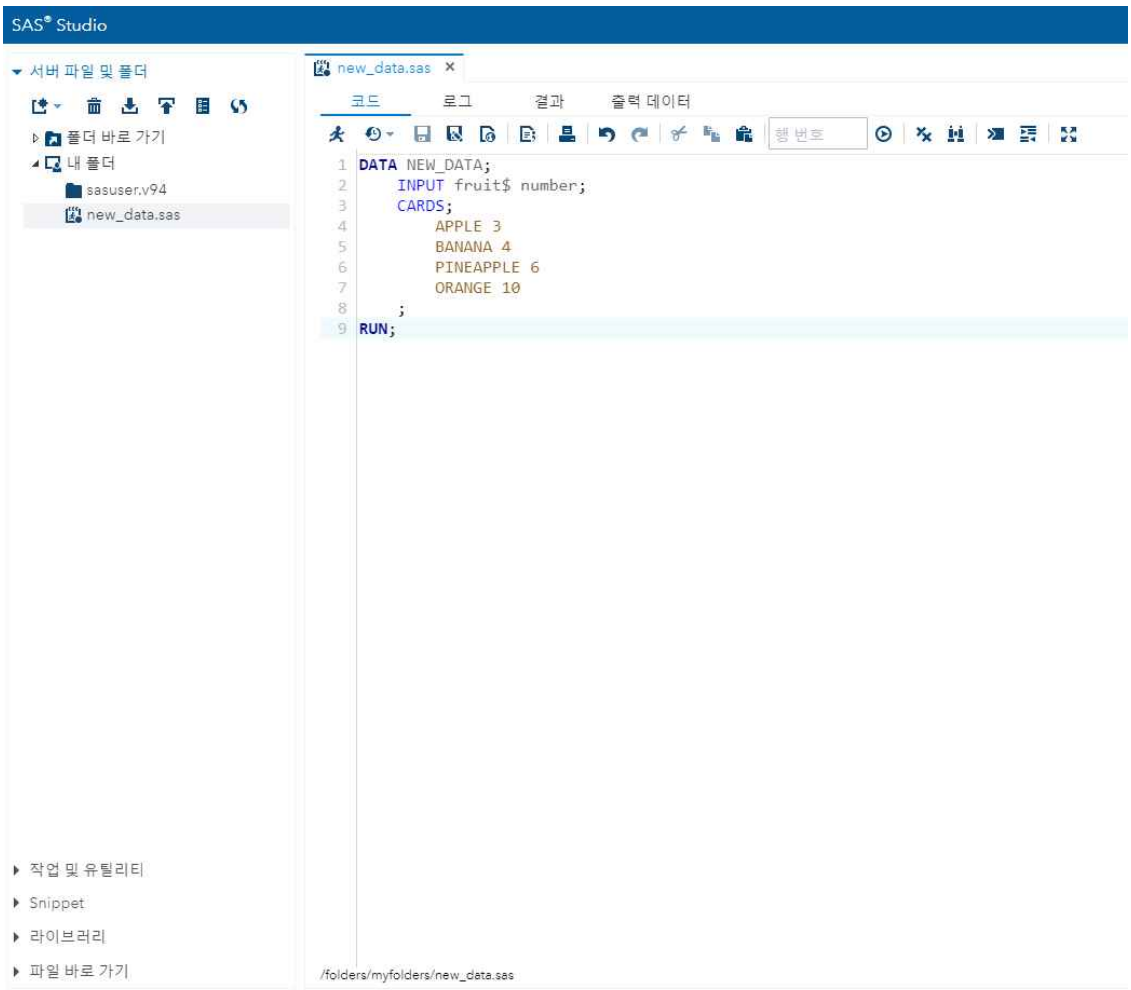


- SAS University Edition 실행
- 로딩이 모두 완료되어 위와 같은 화면이 뜨면 실행 성공

1-8. SAS 시작

- 웹 브라우저를 이용하여 http://localhost:10080 으로 접속
- SAS Studio 시작 버튼을 눌러 초기화면으로 접속

1-9. SAS 초기화면



- 좌측: 파일 목록 출력 / 파일 생성 화면 / 라이브러리
- 우측: 실제 코드 작성 화면 / 코드 실행 / 출력 결과 및 데이터
- '내 폴더'에서 작업 시 설정된 공유 폴더로 자동 파일 연결 및 저장
- 코드 작성 이후 실행(F3)을 이용하여 출력 데이터 혹은 결과를 확인

1-10. 자료(데이터)

- 자료는 수, 단어 등 의미를 가진 형태 단위
- 통계학은 많은 자료를 수집, 관찰하여 정리, 분석하는 수학의 한 분야
- 자료의 행(row): 행에는 수집 조건을 만족하는 '관측치'가 나열된다.
- 자료의 열(column): 열에는 관측치가 가지는 '속성'이 나열된다.
- 관측치는 자료의 속성에 대해 수치 혹은 단어로 특성이 드러난다.

1-11. SAS 파일 생성하기

- SAS 프로그램의 확장자는 (.sas) 이다.
- SAS의 모든 명령문은 세미콜론(;)으로 마무리해야 한다.
- SAS에서 사용되는 자료는 크게 '문자자료'와 '수치자료'로 구분된다. SAS 파일 작성 시 SAS 프로그램이 입력된 자료가 문자자료인지 인식하기 위해 자료입력 명령어인 INPUT 명령어 뒤에 나열되는 문자 변수 이름 뒤에는 반드시 $ 기호를 표기해야 한다.
- SAS에서 자료를 읽어오는 것은 줄 별로 수행되므로 한 줄에는 하나의 관측치 자료가 존재해야 한다. 만약 한 줄에 하나 이상의 자료가 같이 입력되어 있는 경우 INPUT 명령어의 마지막에 @@ 기호를 이용하여 INPUT 명령어 뒤에 기술된 변수의 수만큼 읽는 과정을 반복하도록 한다.
- SAS 프로그램에 자료를 입력하는 방법은 직접 입력하는 방법과 이전에 작성된 다른 확장자의 파일로부터 불러오는 방법이 있다.

1-12. SAS 파일 자료 직접 입력

```
DATA [새로운 SAS 자료이름];
        INPUT [변수이름1] [변수이름2] [변수이름3] ~;
        CARDS;
[자료1-1] [자료1-2] [자료1-3] ~
[자료2-1] [자료2-2] [자료2-2] ~
        ;
RUN;
```

- DATA: 자료를 입력하는 경우 반드시 필요한 SAS 파일 생성 명령어로, 다음에 기재된 자료 이름으로 자료(파일)가 생성된다.
- INPUT: 자료의 속성을 나타내는 변수를 지정하는 명령어로, 다음에 기재된 변수이름으로 변수들을 구분하고 속성을 나타낼 수 있다. 이러한 속성은 완성된 자료에서 열(column)로써 표현된다.
- CARDS: 직접 자료를 입력하는 것을 나타내는 명령어로, 자료 입력이 끝난 후 반드시 다음 줄에 세미콜론(;)으로 입력 종료를 나타내야 한다.
- RUN: SAS 프로그램에서 작성된 명령문의 모든 종료를 의미하며, 해당 명령어가 프로그램에 포함되어 있어야 프로그램이 수행될 수 있다.

1-13. SAS 파일 자료 외부로부터 불러와 저장하기

```
DATA [새로운 SAS 자료이름];
        INFILE "자료의 경로";
        INPUT [변수이름1] [변수이름2] [변수이름3] ~;
RUN;
```

- INFILE: 외부로부터 파일을 불러와 자료를 입력하는 명령어로, 다음에 기재되는 경로로부터 파일을 접근하여 자료를 새롭게 저장한다. 또한, 명령어 뒤와 세미콜론(;) 사이에 아래의 추가 옵션 등을 이용하여 저장하는 자료를 나누는 구분자 및 불러오는 위치를 결정할 수 있다.

추가 옵션	설명	기본 값
DELIMITER="기호"	지정한 기호를 기준으로 자료 구분	공백
FIRSTOBS=위치	지정한 위치로부터 자료 입력 시작	1

- 공유 폴더 설정 시 "/folders/myfolders/" 경로가 파일의 기본 경로가 되며, 좌측 파일 목록에서 우측 클릭 > 속성을 통해서 확인할 수도 있다.
- INPUT: 외부로부터 INFILE 명령어를 통해 자료를 불러오는 경우 프로그램은 해당 자료의 속성이 무엇인지 모르기 때문에 반드시 입력해야 한다.

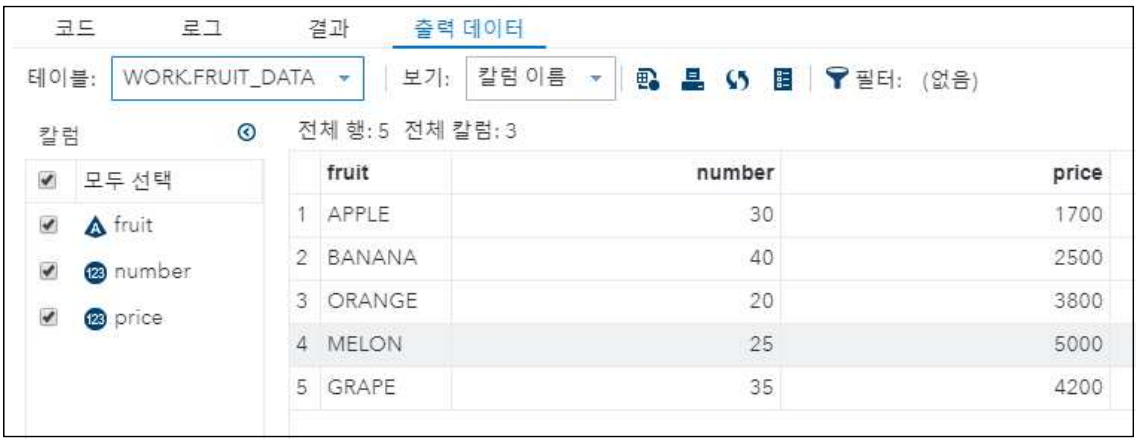
<실습1> SAS 파일 자료 직접 입력

- 아래의 과일 가게 자료를 데이터로 직접 저장해본다.
- 문자자료와 수치자료의 구분을 통해 변수이름 입력 시 $를 적절히 사용
- 프로그램 실행 후 '출력 데이터' 탭에서 저장된 자료를 확인 가능

[과일 가게 자료 (fruit_data)]

과일이름 (fruit)	개수 (number)	가격 (price)
APPLE	30	1,700
BANANA	40	2,500
ORANGE	20	3,800
MELON	25	5,000
GRAPE	35	4,200

[출력 결과]



<실습2> SAS 파일 자료 외부로부터 불러와 저장하기

- 아래의 예시 데이터를 저장하여 텍스트 파일(.txt)을 생성한다.
- 저장된 텍스트 파일로부터 자료를 불러와 입력한다.
- 텍스트 파일의 첫 줄을 보고 자료 이름을 연동한다.
- 자료의 구분 및 추가 옵션의 사용으로 자료를 적절히 입력한다.

[예시 데이터]

```
CLASS|NAME|KOR|ENG|MAT|SCI
1|AMY|67|87|90|98
1|BECKY|45|45|56|98
1|CAESAR|95|59|96|88
1|DAVID|65|94|89|98
1|ERIC|45|65|78|98
1|FLORIA|78|76|98|89
2|GEORGE|87|67|65|56
2|HARRY|89|98|78|78
2|IAN|100|78|56|65
2|JACK|99|89|87|87
2|KELLY|98|45|99|97
2|LINDA|89|99|72|86
```

[출력 결과]

전체 행: 12 전체 칼럼: 6

	CLASS	NAME	KOR	ENG	MAT	SCI
1	1	AMY	67	87	90	98
2	1	BECKY	45	45	56	98
3	1	CAESAR	95	59	96	88
4	1	DAVID	65	94	89	98
5	1	ERIC	45	65	78	98
6	1	FLORIA	78	76	98	89
7	2	GEORGE	87	67	65	56
8	2	HARRY	89	98	78	78
9	2	IAN	100	78	56	65
10	2	JACK	99	89	87	87
11	2	KELLY	98	45	99	97
12	2	LINDA	89	99	72	86

2. SAS 기본 명령어, 연산자 및 함수

2-1. INPUT 명령어 (자료 입력)

- 프로그램으로 자료를 입력할 때, 프로그램은 주어지는 자료의 속성(분류) 등이 무엇인지 모르기 때문에 이를 인식시키기 위해 어떤 자료가 입력되는지 정의하기 위한 명령어
- 일반적으로 공백을 기준으로 데이터를 구분하여 입력할 수 있지만, 입력되는 자료가 일정한 형식의 자료라면 아래의 다양한 방법들을 이용하여 유연하게 자료를 입력할 수 있다.

예시	설명
INPUT a b c d;	변수이름을 지정하는 일반적인 형태
INPUT a1-a5;	연속된 변수이름에 저장
INPUT str $ num1 num2;	자료가 문자인 경우 꼭 $를 이용하여 저장
INPUT a 1-2 b 3-5 c 10;	주어진 문자열에서 원하는 위치를 저장
INPUT a 4. b 5. c 2.;	주어진 문자열에서 원하는 수만큼 저장

- 자료를 직접 입력하는 경우, 예시 중 위의 3가지는 공백을 기준으로 구분하여 저장하는 자료를 자유롭게 작성해도 되나, 아래 3가지와 같이 원하는 위치 및 수를 기준으로 저장하는 경우 공백도 범위에 포함되기 때문에 이를 고려하여 자료를 직접 입력해야 한다.
- 또한, 자료가 일정한 형식을 띄지 않는 경우 입력할 수 있는 문자자료 길이의 기본 값이 ‘8’이므로 더 긴 문자열을 입력하려면 length 명령어를 이용하여 해당 문자자료의 변수이름에 대한 길이를 미리 할당해야 한다.

2-2. SET 명령어 (자료 재정의)

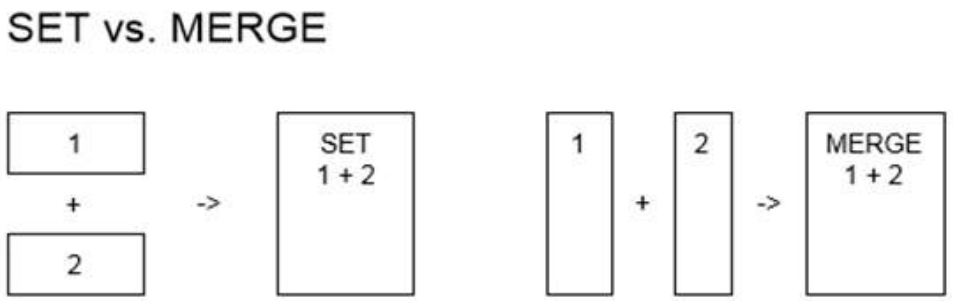
```
DATA [새로운 SAS 자료이름];
      SET [기존 SAS 자료이름];

DATA [새로운 SAS 자료이름];
      SET [기존 SAS 자료이름1] [기존 SAS 자료이름2] ~;

RUN;
```

- SET: 명령어 뒤에 기재되는 자료이름이 하나인 경우 기존에 존재하던 자료를 새로운 이름으로 복사하여 저장하는 재정의 역할을 수행한다. SET 명령어를 이용하면 기존 자료로부터 필요한 자료만을 일부 추출하여 새로운 자료로 만드는 기능을 수행할 수 있다.
- SET 명령어 뒤에 기재되는 자료이름이 여러 개인 경우 기존에 존재하던 여러 자료들을 하나의 자료로 만들어 저장한다. 단, SET에서 하나의 자료가 되는 것은 자료 행(row)의 증가로 수행되므로 해당 자료들에 저장되는 속성(변수이름) 열(column)이 서로 동일해야 한다. 일치하지 않으면 기존 자료에서 분류되지 않았던 관측치가 추가되므로 missing value가 발생할 수 있다.

2-3. MERGE 명령어 (자료 합치기)

```
DATA [새로운 SAS 자료이름];
      MERGE [기존 SAS 자료이름1] [기존 SAS 자료이름2] ~;
RUN;
```

- MERGE: 명령어 뒤에 기재된 두 가지 이상의 SAS 자료들을 합쳐서 새로운 자료로 생성한다. SET에서 여러 자료가 하나의 자료로 추가되어 합쳐지는 것과 달리, MERGE에서는 합쳐지는 대상에 존재하는 모든 변수를 포함하여 서로 연관된 관측치 간의 자료를 기반으로 합친다.



2-4. KEEP / DROP 명령어 (자료 추출)

```
DATA [새로운 SAS 자료이름];
        SET [기존 SAS 자료이름];
        KEEP [변수이름1] [변수이름2] ~;
RUN;

DATA [새로운 SAS 자료이름];
        (KEEP = [변수이름1] [변수이름2] ~;)
        SET [기존 SAS 자료이름];
RUN;
```

- KEEP: 기존 SAS 자료 중 특정 변수만 포함하는 새로운 자료로 재정의

```
DATA [새로운 SAS 자료이름];
        SET [기존 SAS 자료이름];
        DROP [변수이름1] [변수이름2] ~;
RUN;

DATA [새로운 SAS 자료이름];
        (DROP = [변수이름1] [변수이름2] ~;)
        SET [기존 SAS 자료이름];
RUN;
```

- DROP: 기존 SAS 자료 중 특정 변수만 제거 후 새로운 자료로 재정의

2-5. 조건문

```
DATA [새로운 SAS 자료이름];
        ~
        IF 조건1 THEN 문장1;
        ELSE IF 조건2 THEN 문장2;
        ~
        ELSE 문장3;
RUN;
```

- IF 조건 THEN 문장: 조건에 부합하는 경우 문장이 수행된다.
- 한 번에 여러 조건을 나열하여 사용하는 경우 ELSE IF (혹은 ELIF)를 이용하여 여러 조건을 동시에 수행하도록 할 수 있다.

2-6. 연산자

```
/* SCORE1: 3-5. (실습) 조건문 예시 */

DATA TOT_AVG_SCORE;
      SET SCORE1;
      TOT = KOR + ENG + MAT + SCI;
      AVG = TOT / 4;
RUN;
```

- 수치자료들과 연산자를 이용한 결과를 새로운 수치자료로 만들 수 있다.
- 조건문을 사용하는 것과 같이 연산 결과도 입력 정의 없이 사용가능하다.

[산술 연산자]

연산자	설명	연산자	설명
+	더하기	-	빼기
*	곱하기	/	나누기
**	거듭제곱		

[비교 연산자]

연산자	설명	연산자	설명
=	같다	^= 또는 ~=	같지 않다
>	크다	<	작다
>=	크거나 같다	<=	작거나 같다

[논리 연산자]

연산자	설명	연산자	설명
&	AND	|	OR
^( ) 또는 ~( )	NOT		

2-7. SAS 함수

- 자주 사용되는 수식 및 필요한 계산들을 함수로 미리 구현하여 제공
- 일반적인 함수 사용 방법: 변수이름 = 함수이름(매개변수1, 매개변수2, ~);

[SAS 내장 함수]

함수	설명	함수	설명
COS(X)	코사인	CEIL(X)	천장 함수
SIN(X)	사인	FLOOR(X)	바닥 함수
TAN(X)	탄젠트	INT(X)	정수 부분
MEAN(...)	평균	ROUND(X, 단위)	단위부터 반올림
SUM(...)	합	EXP(X)	지수(e)의 X승
STD(...)	표준편차	GAMMA(X)	감마함수
VAR(...)	분산	LOG(X)	자연로그
CV(...)	변이계수	LOG10(X)	상용로그
RANGE(...)	범위	PROBBNML(P, N, M)	이항분포
STDERR(...)	표준오차	POISSON(λ, M)	포아송분포
ABS(X)	절댓값	PROBBETA(X, A, B)	베타분포
MAX(...)	최댓값	PROBGAM(X, A)	감마분포
MIN(...)	최솟값	PROBNORM(X)	정규분포
MOD(X, Y)	Modulo 연산	PROBCHI(X, DF)	카이제곱분포
SIGN(X)	양수 1, 음수 -1	PROBT(X, DF)	t-분포
SQRT(X)	제곱근	PROBF(X, DF1, DF2)	f-분포

- 연산자를 사용하는 것과 같이, 적절한 내장 함수로 자료를 추가할 수 있다.

<실습1> INPUT 명령어 (자료 입력)

[INPUT 명령어 예시 1]

```
DATA class;
        length classname $ 12;
        INPUT classname $ number;
        CARDS;
KOREA 45
SCIENCE 37
MATH 40
ENGLISH 60
PROGRAMMING 70
      ;
RUN;
```

- 공백을 기준으로 자료를 구분하고, 입력하는 자료의 길이를 고려하여 지정

[INPUT 명령어 예시 2]

```
DATA class;
        INPUT code $ 2. classname $ 3-5 number;
        CARDS;
01KOR 45
02SCI 37
03MAT 40
04ENG 60
05PRO 70
      ;
RUN;
```

- 동일한 형식을 가진 자료로부터 필요한 자료의 속성을 유연하게 지정

[예시 1과 2의 실행 결과]

OBS	classname	number
1	KOREA	45
2	SCIENCE	37
3	MATH	40
4	ENGLISH	60
5	PROGRAMMING	70

OBS	code	classname	number
1	01	KOR	45
2	02	SCI	37
3	03	MAT	40
4	04	ENG	60
5	05	PRO	70

<실습2> SET 명령어 (자료 재정의)

- 학생들의 점수(SCORES)를 나타낸 데이터(텍스트파일)를 다운로드한다.

https://drive.google.com/open?id=1bwmzuOHcpkciUIxBfwIKADUzg3wftYJd

[SET 명령어 예시]

```
DATA SCORE1;
        INFILE "/folders/myfolders/score1.txt";
        INPUT GEN $ KOR MAT ENG SCI;

DATA SCORE2;
        INFILE "/folders/myfolders/score1.txt";
        INPUT GEN $ KOR MAT ENG SCI;

DATA SCORES;
        SET SCORE1 SCORE2;
RUN;
```

- 주어진 score1.txt, score2.txt 자료로부터 하나의 자료를 만든다.
- 생성된 새로운 자료 전체 행의 개수가 각 자료 행의 개수의 합이 된다.

[실행 결과]

<실습3> MERGE 명령어 (자료 합치기)

[MERGE 명령어 예시]

```
/* SCORES: 3-2. (실습) SET 명령어 예시 */

DATA PASS;
        INFILE "/folders/myfolders/pass.txt";
        length PASS $ 9;
        INPUT GEN $ PASS $;

DATA SCORES_PASS;
        MERGE SCORES PASS;

RUN;
```

- 주어진 pass.txt 자료를 불러와 이전에 생성된 SCORES 자료에 합친다.
- 한글자료는 한 글자당 길이가 3으로 계산되므로 PASS 자료에 존재하는 최대 3글자 한글자료를 저장하기 위해 length를 지정한다.

[실행 결과]

<실습4> 조건문

[조건문 예시]

```
DATA SCORE1;
        INFILE "/folders/myfolders/score1.txt";
        INPUT GEN $ KOR MAT ENG SCI;

DATA CSE_SCORE1;
        SET SCORE1;
        IF MAT >= 90 THEN GRADE="A";
        ELSE IF MAT >= 80 THEN GRADE="B";
        ELSE IF MAT >=70 THEN GRADE="C";
        ELSE GRADE="D";
RUN;
```

- 수학 점수 자료로부터 등급(GRADE)이라는 새로운 자료가 추가되는 예시

[실행 결과]

- 위에서부터 조건이 만족되면 명령문이 실행된다.
- 추가되는 새로운 자료에 대해서는 미리 입력 정의를 해줄 필요가 없다.

<마무리> SAS 기본 명령어, 연산자 및 함수

- 주어진 score1.txt와 score2.txt를 하나의 데이터로 합친다.
- 주어진 pass.txt를 새로운 데이터로 저장하면서 아래의 내용을 변경한다.

- 주어진 pass.txt에서 합격/불합격을 나타내는 변수의 내용을, 합격일 경우 'T'로, 불합격일 경우 'F'로 저장되도록 한다. (변수를 추가하지 않고 변경)

- 위 두 가지 데이터를 병합하고, 모든 과목 점수를 더한 TOTAL 변수와 평균을 나타내는 AVG 변수를 추가하여 저장한다.
- AVG 변수를 통해 평균 80 이상인 학생에 대해 수상여부를 나타내는 PRIZE 변수를 추가한다. PRIZE 변수에는 평균이 80 이상인 경우 'YES'를, 아니면 'NO'를 저장한다.

[실행 결과]

전체 행: 600 전체 칼럼: 9 행 1-100

	GEN	KOR	MAT	ENG	SCI	PASS	TOTAL	AVG	PRIZE
1	F	79	90	77	6	F	252	63	NO
2	M	60	58	70	26	F	214	53.5	NO
3	F	76	9	80	36	T	201	50.25	NO
4	M	28	99	32	92	F	251	62.75	NO
5	F	78	52	95	47	F	272	68	NO
6	M	82	7	77	3	F	169	42.25	NO
7	F	18	98	38	77	F	231	57.75	NO
8	M	48	1	8	9	F	66	16.5	NO
9	F	92	8	11	40	F	151	37.75	NO
10	M	61	96	43	1	T	201	50.25	NO
11	F	20	63	6	57	F	146	36.5	NO
12	M	49	63	14	33	F	159	39.75	NO
13	F	47	12	60	1	F	120	30	NO
14	M	53	34	71	3	F	161	40.25	NO
15	F	59	99	48	30	F	236	59	NO
16	M	13	87	6	11	T	117	29.25	NO
17	F	27	3	47	32	T	109	27.25	NO
18	M	11	26	60	70	F	167	41.75	NO
19	F	45	32	78	80	F	235	58.75	NO
20	M	49	67	48	13	F	177	44.25	NO
21	F	68	43	91	88	T	290	72.5	NO
22	M	41	80	54	46	T	221	55.25	NO
23	F	97	44	37	58	T	236	59	NO

<과제1> SAS 기본 명령어, 연산자 및 함수

(1) 실습에서 사용한 score1.txt와 score2.txt를 하나의 데이터로 합친다.

- 데이터의 속성은 GEN / KOR / MAT / ENG / SCI로 통일한다.

(2) 점수의 비율이 국어:수학:영어:과학=2:4:2:2가 되어 총점이 1,000점이 되도록 계산된 TOTAL 속성을 추가한다.

(3) 새로운 NOTE 속성을 추가하여, TOTAL이 800점 이상인 경우 'GREAT', 600점 이상인 경우 'GOOD', 400점 이상인 경우 'OK', 이외에는 'FAIL'이 출력되도록 한다.

(4) MAT과 추가된 TOTAL, NOTE 속성만 존재하는 데이터를 만든다.

- 과제 결과: MAT, TOTAL, NOTE 속성을 가진 데이터

[실행 결과]

전체 행: 600 전체 칼럼: 3

	MAT	TOTAL	NOTE
1	90	684	GOOD
2	58	544	OK
3	9	420	OK
4	99	700	GOOD
5	52	648	GOOD
6	7	352	FAIL
7	98	658	GOOD
8	1	134	FAIL
9	8	318	FAIL
10	96	594	OK
11	63	418	OK
12	63	444	OK
13	12	264	FAIL
14	34	390	FAIL
15	99	670	GOOD
16	87	408	OK
17	3	224	FAIL
18	26	386	FAIL
19	32	534	OK
20	67	488	OK

(중요) 과제 제출

- 작성한 코드 파일(.sas)과 출력된 결과(캡처 혹은 저장)의 PDF(지원되지 않는 경우 Word 혹은 HWP 이용), 2개의 파일을 압축(.zip)하여 아래 서식으로 이름을 정하고 이메일 제목 또한 동일하게 하여 제출한다.
- 파일 이름: 통계학실습_과제X_학번_이름.zip
- 이메일 제목: 통계학실습_과제X_학번_이름
- 제출 이메일: gtsk623@gmail.com

3. SAS 프로시저 및 기술통계량

3-1. SAS 프로시저

- 저장된 SAS 자료나 데이터 파일 등을 다루어 내부 관측된 값들을 출력, 정렬, 관계 분석 등 다양한 결과로 보여주기 위해 사용되는 문법이다.

[출력 프로시저]

코드 로그 결과 출력 데이터

```
DATA CLASS;
    SET SASHELP.CLASS;

PROC PRINT DATA=CLASS;
RUN;
```

OBS	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83.0
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84.0
10	John	M	12	59.0	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90.0
13	Louise	F	12	56.3	77.0
14	Mary	F	15	66.5	112.0
15	Philip	M	16	72.0	150.0
16	Robert	M	12	64.8	128.0
17	Ronald	M	15	67.0	133.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

- 저장된 SAS 자료를 실제 테이블 형태로 출력한다.

3-2. SASHELP 라이브러리

- SAS 프로그램에서 다양한 예시 데이터를 제공하는 내장 라이브러리
- 내장 라이브러리 데이터의 덮어쓰기(overwrite)를 막기 위해 SET 명령어를 통해 새 데이터로 선언하여 사용한다.

[SASHELP 라이브러리]

3-3. 출력 프로시저

```
PROC PRINT DATA=[SAS 자료이름];
      VAR [변수이름1] [변수이름2] ~;
RUN;
```

- PROC: 프로시저를 실행하는 명령어로, 뒤에 붙는 기능을 수행한다.
- PRINT: 데이터에 포함되어 있는 모든 속성들을 포함하여 값을 출력한다.
- VAR: 명령어 뒤의 변수들만 포함하여 출력한다. (없다면, 전체 출력)

[출력 프로시저 예시]

```
PROC PRINT DATA=SASHELP.CLASS;
PROC PRINT DATA=SASHELP.CLASS;
      VAR NAME;
RUN;
```

[출력 결과]

OBS	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69.0	112.5
2	Alice	F	13	56.5	84.0
3	Barbara	F	13	65.3	98.0
4	Carol	F	14	62.8	102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83.0
7	Jane	F	12	59.8	84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	84.0
10	John	M	12	59.0	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90.0
13	Louise	F	12	56.3	77.0
14	Mary	F	15	66.5	112.0
15	Philip	M	16	72.0	150.0
16	Robert	M	12	64.8	128.0
17	Ronald	M	15	67.0	133.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

OBS	Name
1	Alfred
2	Alice
3	Barbara
4	Carol
5	Henry
6	James
7	Jane
8	Janet
9	Jeffrey
10	John
11	Joyce
12	Judy
13	Louise
14	Mary
15	Philip
16	Robert
17	Ronald
18	Thomas
19	William

- 좌측은 SASHELP.CLASS 라이브러리의 모든 데이터를 출력한 결과
- 우측은 SASHELP.CLASS 라이브러리에서 'NAME' 속성만을 출력한 결과

3-4. 정렬 프로시저

```
PROC SORT DATA=[SAS 자료이름];
        BY [DESCENDING] [변수이름1] [변수이름2] ~;
RUN;
```

- SORT: 데이터의 특정 속성을 기준으로 오름차순 혹은 내림차순 정렬 저장
- BY: 명령어 뒤에 나열되는 속성을 기준으로 순서대로 정렬 (필수 명령어)
- ASCENDING(오름차순)이 기본 값이며, DESCENDING(내림차순)은 BY 명령어 뒤에 추가하여 사용한다.

3-5. 산점도 프로시저 (그래프)

```
PROC PLOT DATA=[SAS 자료이름];
        PLOT [변수이름1]*[변수이름2]="기호" /
        VAXIS=num1 TO num2 BY num3
        HAXIS=num4 TO num5 BY num6;
RUN;
```

- 산점도란 어떤 두 변수(속성)의 관계를 시각적으로 파악하기 위해 실제 데이터 값(수치자료)을 직교 좌표계에 점으로 표시하여 나타내는 방법이다.
- PLOT: 자료의 속성 2개에 대한 관계를 산점도로 표현하는 명령어로, 전자가 세로축, 후자가 가로축으로 표현된다. 표시되는 각 점을 원하는 특정 기호로 나타내도록 할 수 있다.
- HAXIS, VAXIS: 가로, 세로축에 나타나는 자료의 눈금을 지정한다. 직접 범위를 지정하지 않을 경우 자동으로 최댓값과 최솟값 사이로 출력하고, 범위를 지정하는 경우 시작 값과 끝 값 사이에 포함되는 데이터만을 출력한다. BY 명령어는 눈금 단위 값으로써, 기본 값으로 1을 가진다.

```
PROC PLOT DATA=[SAS 자료이름];
        PLOT [변수이름1]*[변수이름2]="기호1"
        [변수이름3]*[변수이름4]="기호2" / OVERLAY;
RUN;
```

- OVERLAY: 동시에 여러 변수 쌍들의 관계를 한 화면에 겹쳐 표현한다. 각 변수 쌍들은 지정한 기호로 출력되어 구분된다.
- 일반적으로 한 속성에 대해 두 비슷한 속성들이 가지는 의미를 파악하기 위해 한 축을 통일한 뒤 다른 두 속성들을 OVERLAY한다.

3-6. 기술통계학

- 통계학의 자료는 많은 수의 관측치들로 이루어져 있기 때문에 분석에 앞서 자료의 특성을 파악하기 위해 정리 및 요약 등을 필요로 한다.
- 기술통계학은 분석 대상이 되는 자료들을 표, 그래프, 수치 등으로 요약하여 자료의 대체적인 내용을 파악하기 위한 분야이다.

[통계학에서의 용어 정의]

용어	설명
모집단	모든 관측치들의 집합
표본	분석을 위해 추출한 일부 관측치들의 집합
모수	모집단의 특성을 나타내는 여러 값들
통계량	표본의 특성을 나타내는 표본의 함수

3-7. 기술통계량: UNIVARIATE

```
PROC UNIVARIATE DATA=[SAS 자료이름] NORMAL PLOT;
        BY [변수이름1] [변수이름2] ~;
        VAR [변수이름3] [변수이름4] ~;
RUN;
```

- UNIVARIATE: 일변량 자료에 대한 다양한 기술통계량들을 제공하는 프로시저로써 여러 가지 대표 값, 산포도, 모수에 대한 신뢰구간 및 정규성 검정 자료, 상자그림, 줄기-잎 그림 등의 분포에 관련된 그래프 등을 출력한다. 일변량 자료란 분석 대상이 되는 변수가 1개인 자료를 의미한다.
- NORMAL: 정규성 검정을 위한 검정통계량 값 계산 및 유의확률 값 등을 제공하는 추가 명령어
- PLOT: 정규성 검정을 위한 정규확률도 및 자료의 전체적인 형태 및 대칭성 등의 파악을 위한 상자그림, 줄기-잎 그림 등을 제공하는 추가 명령어
- BY: 주어진 변수들을 집단으로 구분하여 집단별로 분석 결과를 출력해주는 명령어로, 입력되지 않는 경우 전체 표본 집단에 대해 기술통계량이 산출되며 입력되는 경우 반드시 해당 집단들의 구분을 위해 정렬해주어야 한다.
- VAR: 일변량 분석 결과를 얻고자 하는 대상 변수

[UNIVARIATE 예시]

```
DATA CLASS;
        SET SASHELP.CLASS;

PROC SORT DATA=CLASS;
        BY AGE;

PROC UNIVARIATE DATA=CLASS;
        BY AGE;
        VAR HEIGHT;
RUN;
```

[출력 결과]

UNIVARIATE 프로시저
변수: Height

Age=11

적률			
N	2	가중합	2
평균	54.4	관측값 합	108.8
표준 편차	4.38406204	분산	19.22
왜도	.	첨도	.
제곱합	5937.94	수정 제곱합	19.22
변동계수	8.05893758	평균의 표준 오차	3.1

기본 통계 측도			
위치측도		변이측도	
평균	54.40000	표준 편차	4.38406
중위수	54.40000	분산	19.22000
최빈값	.	범위	6.20000
		사분위 범위	6.20000

UNIVARIATE 프로시저
변수: Height

Age=12

적률			
N	5	가중합	5
평균	59.44	관측값 합	297.2
표준 편차	3.29742324	분산	10.873
왜도	1.31547581	첨도	1.97728439
제곱합	17709.06	수정 제곱합	43.492
변동계수	5.54748189	평균의 표준 오차	1.4746525

기본 통계 측도			
위치측도		변이측도	
평균	59.44000	표준 편차	3.29742
중위수	59.00000	분산	10.87300
최빈값	.	범위	8.50000
		사분위 범위	2.50000

- 학생들을 나이별로 집단을 구분하여, 각 집단에서 신장에 대해 일변량 분석된 기술통계량을 출력한다.
- 분석에서 나이별 결과를 얻기 위해 BY 명령어를 사용하였으므로 반드시 분석 이전 나이에 대한 변수에 대해 정렬을 수행해야 한다.

3-8. 기술통계량: MEANS

```
PROC MEANS DATA=[SAS 자료이름];
        CLASS [변수이름1] [변수이름2] ~;
        VAR [변수이름3] [변수이름4] ~;
RUN;
```

- MEANS: UNIVARIATE처럼 일변량 자료에 대한 기술통계량을 제공하는 명령어이다. 기본적으로 관측치의 수, 평균, 표준편차, 최솟값, 최댓값 등을 출력한다. 아래 표의 다양한 키워드를 DATA 명령어 뒤에 입력하여 추가적인 여러 가지 기술통계량을 출력할 수 있다.
- CLASS: 주어진 변수들에 대해 그룹으로 자동구분지어 분류하여 기술통계량 결과를 출력한다. 앞선 UNIVARIATE에서의 BY와 역할이 같으나, 미리 정렬할 필요가 없다는 장점이 있다. (UNIVARIATE에서도 사용가능)
- VAR: 일변량 분석 결과를 얻고자 하는 대상 변수

[MEANS 키워드]

키워드	설명	키워드	설명
CLM	평균 신뢰구간	RANGE	범위
N	관측치수	NMISS	결측치수
CV	변이계수	VAR	분산
STDDEV	표준편차	STDERR	표준오차
MEDIAN(Pn)	중앙값 - n백분위수	Qm(Pn)	제m사분위수 - n백분위수
LCLM	평균 95% 신뢰구간의 하한	UCLM	평균 95% 신뢰구간의 상한
SUM / SUMWGT	합 / 가중합	Pn	백분위수 (n=1,5,90,95,99)
MEAN	평균	PROBT	유의확률
T	검정통계량		

<실습1> 정렬 프로시저

- SASHELP 라이브러리의 CLASS 데이터를 이용한다.
- CLASS 데이터를 'CLASS_DATA'라는 이름의 새로운 데이터로 지정한다.
- 위 CLASS_DATA를 나이에 대해 오름차순으로 정렬한다.
- 정렬된 데이터를 PRINT(출력 프로시저)로 출력한다.

[출력 결과]

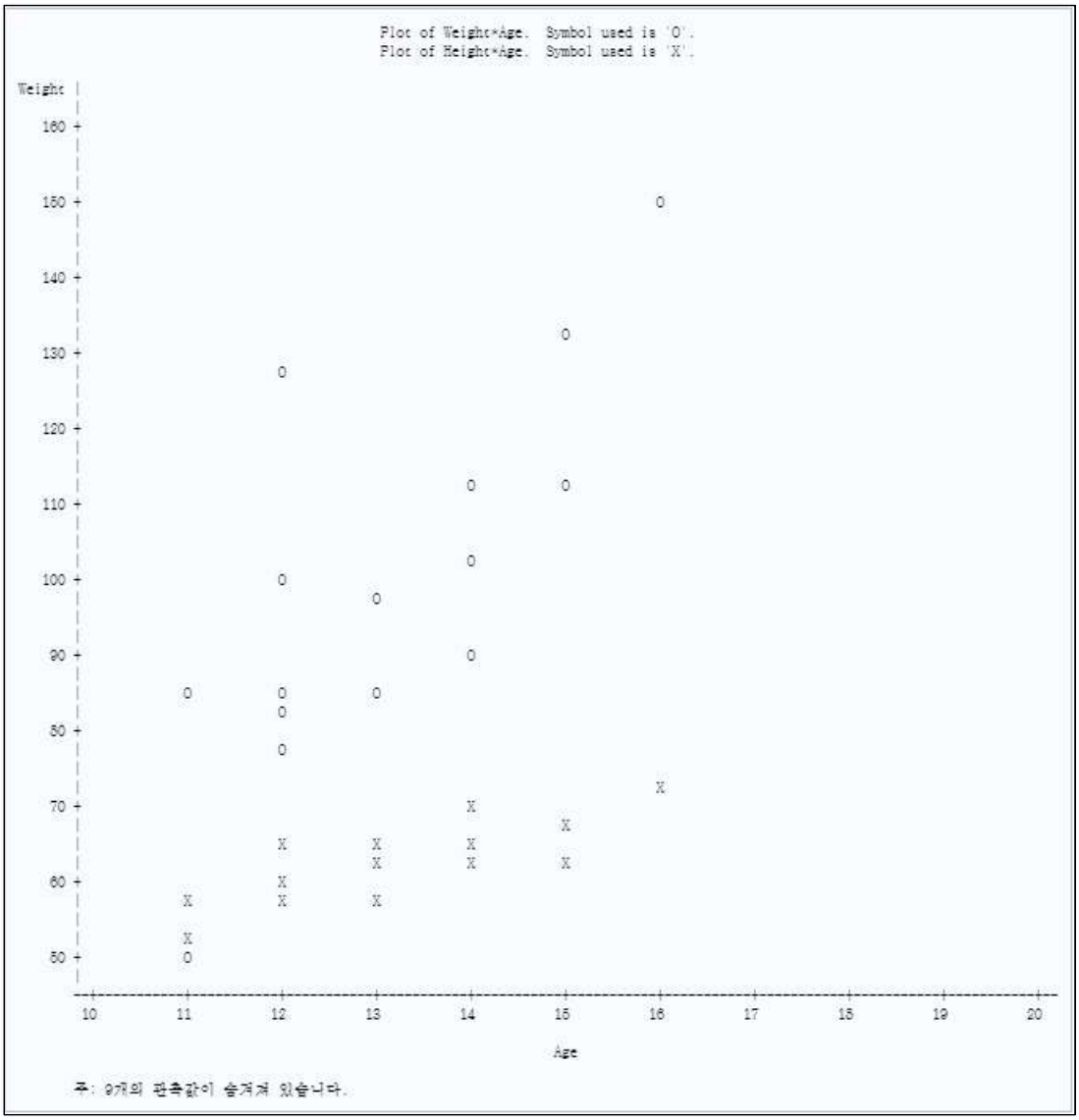
OBS	Name	Sex	Age	Height	Weight
1	Joyce	F	11	51.3	50.5
2	Thomas	M	11	57.5	85.0
3	James	M	12	57.3	83.0
4	Jane	F	12	59.8	84.5
5	John	M	12	59.0	99.5
6	Louise	F	12	56.3	77.0
7	Robert	M	12	64.8	128.0
8	Alice	F	13	56.5	84.0
9	Barbara	F	13	65.3	98.0
10	Jeffrey	M	13	62.5	84.0
11	Alfred	M	14	69.0	112.5
12	Carol	F	14	62.8	102.5
13	Henry	M	14	63.5	102.5
14	Judy	F	14	64.3	90.0
15	Janet	F	15	62.5	112.5
16	Mary	F	15	66.5	112.0
17	Ronald	M	15	67.0	133.0
18	William	M	15	66.5	112.0
19	Philip	M	16	72.0	150.0

<실습2> 산점도 프로시저 (그래프)

- SASHELP 라이브러리의 CLASS 데이터를 이용하여 산점도를 출력한다.
- CLASS 데이터의 WEIGHT, HEIGHT, AGE 변수를 이용한다.
- 학생들의 나이에 따른 몸무게, 신장 산점도를 OVERLAY하여 출력한다.

- 가로축: 나이 (시작 값 10, 끝 값 20, 단위 1)
- 세로축: 몸무게(기호 'O'), 신장(기호 'X') (시작 값 50, 끝 값 160, 단위 10)

[출력 결과]



- 학생들의 나이에 대하여 신장과 몸무게 각각의 관계가 나타난 산점도
- 학생들의 신장과 몸무게는 나이에 대해 비교적 비례 관계를 형성하고 있다.

<마무리> SAS 프로시저 및 기술통계량

- SASHELP 라이브러리의 CLASS 데이터를 이용한다.

- 학생들을 나이가 많은 순으로 정렬하여 PRINT 출력한다.
- 학생들의 신장을 모두 포함하는 숫자 범위(시작 값, 끝 값)를 확인한다.

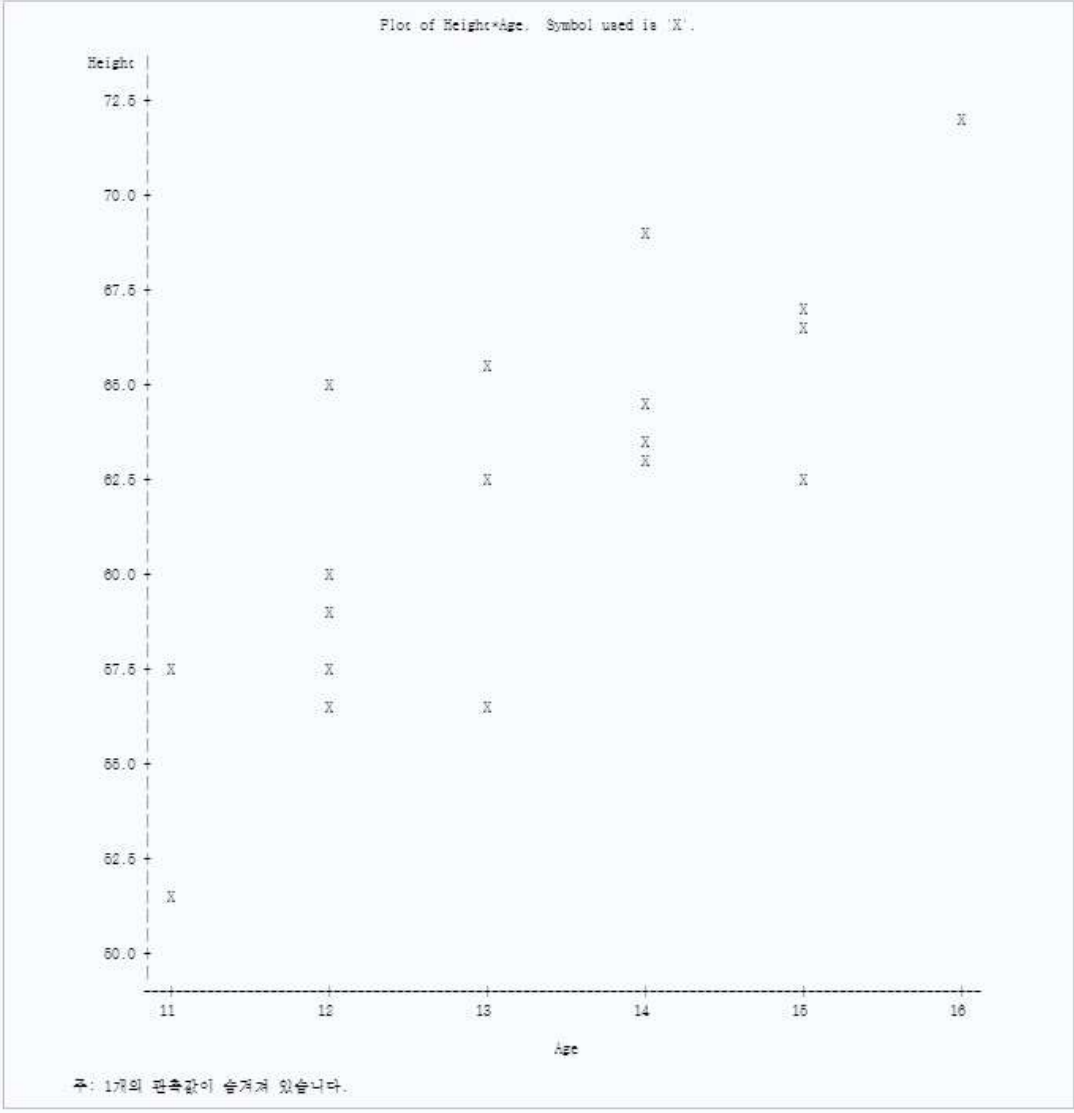
- 학생들의 나이에 따른 신장 관계를 표현한 산점도를 그린다.

- 나이를 가로축으로, 신장을 세로축으로 한다.
- VAXIS/HAXIS를 사용하여 범위를 설정하고, 단위 값은 10으로 한다.

- 일변량 분석

- UNIVARIATE를 이용하여 학생들의 나이에 대해 일변량 분석한다.
- MEANS를 이용하여 학생들의 신장의 합/평균/분산을 출력한다.

[산점도 출력 결과]



[일변량 분석 출력 결과]

UNIVARIATE 프로시저
변수: Age

적률			
N	19	가중합	19
평균	13.3157895	관측값 합	253
표준편차	1.49267216	분산	2.22807018
왜도	0.05361167	첨도	-1.1109255
제곱합	3409	수정 제곱합	40.1052632
변동계수	11.2097909	평균의 표준 오차	0.34244248

MEANS 프로시저

분석 변수: Height		
합계	평균	분산
1184.40	62.3368421	26.2869006

<과제2> SAS 프로시저 및 기술통계량

(1) SASHELP 라이브러리의 FISH 데이터를 이용한다.

- 종(SPECIES)에 따른 무게(WEIGHT)를 산점도로 출력하는 코드를 작성하고, 가장 넓은 무게 범위 폭을 가지는 종이 무엇인지 확인해본다.
- ‘MEANS’ 명령어를 이용하여 종 별로 무게의 표준편차를 확인하는 코드를 작성한다. 산점도에서 확인한 것과 같이 가장 넓은 무게 범위 폭을 가지는 종이 가장 큰 편차를 가지는지 확인해본다.

(2) SASHELP 라이브러리의 CLASS 데이터를 이용한다.

- ‘UNIVARIATE’ 명령어를 이용하여 나이별 집단의 몸무게에 대해 일변량 분석을 출력하는 코드를 작성한다. (단, CLASS를 사용하지 말고 BY 사용)
- 나이와 몸무게의 관계를 나타내는 산점도를 출력한다. 이 때, 나이의 범위는 10~20 (단위=1), 몸무게의 범위는 50~160 (단위=10)으로 지정한다.

(중요) 과제 제출

- 작성한 코드 파일(.sas)과 출력된 결과(캡처 혹은 저장)의 PDF(지원되지 않는 경우 Word 혹은 HWP 이용), 2개의 파일을 압축(.zip)하여 아래 서식으로 이름을 정하고 이메일 제목 또한 동일하게 하여 제출한다.
- 파일 이름: 통계학실습_과제X_학번_이름.zip
- 이메일 제목: 통계학실습_과제X_학번_이름
- 제출 이메일: gtsk623@gmail.com

4. SAS 그래프 출력

4-1. 그래프

- 다양한 데이터에 속하는 변수(속성)들로부터 산출되는 값 혹은 관계 등을 시각적으로 표현하기 위해 도형으로 나타내는 방법
- 다양한 결과에 대해 전달하고자 하는 내용을 효율적으로 전달하기 위해 다양한 그래프 중 적절한 그래프를 선택할 수 있어야 한다.
- 그래프의 종류에는 산점도, 상자그림, 줄기-잎 그림, 히스토그램, 바 차트, 파이 차트 등 다양한 그래프가 있다.

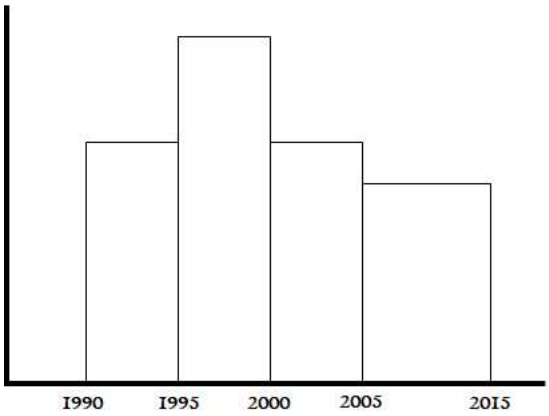
4-2. 자료 분포 파악을 위한 그래프

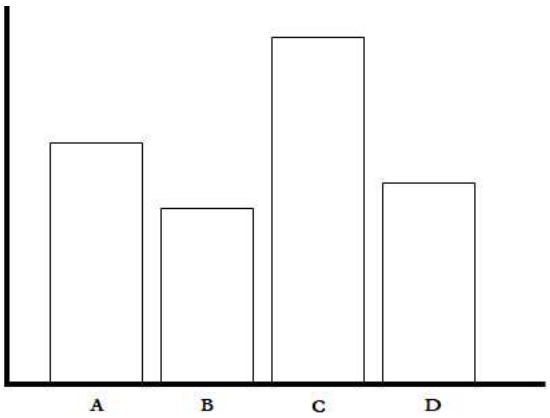
- 히스토그램(막대): 도수분포표로써 주어진 데이터의 분포를 표현하기 위해 사용한다. 도수란 데이터에서 일정한 구간으로 나눈 계급의 각각에 속하는 변량의 개수를 뜻한다. 즉, 도수분포표와 히스토그램을 사용하여 주어진 데이터가 속하는 계급 및 분포 등을 한 눈에 확인할 수 있다.
- 바 차트(막대): 데이터를 분류하여 해당되는 분류 항목의 빈도를 정확하게 표현하기 위해 사용한다. 히스토그램과 같이 직사각형 기둥으로 표현된다.
- 파이 차트(원형): 전체 데이터를 원으로 표현하고, 전체에서 원하는 해당 데이터가 차지하는 비율을 부채꼴 모양으로 나타내어 분류된 항목의 빈도를 표현하기 위해 사용한다. 부채꼴의 중심각은 전체에서 차지하는 해당 데이터의 비율을 나타낸다.

4-3. 히스토그램과 바 차트의 차이

항목	히스토그램	바 차트
세로축	해당하는 자료의 빈도 수	해당하는 자료의 빈도 수
가로축	계급의 크기 (범위)	분류 항목
변수 특성	연속적 변수	불연속적 변수
막대 간 비교	면적 비교	높이 비교
형태 특징	막대가 서로 붙어 있음, 막대의 폭이 계급 크기에 따라 서로 다를 수 있음	막대가 서로 떨어져 있음, 막대의 폭이 항상 같고 높이가 데이터를 표현함

[히스토그램과 바 차트 예시]





- 히스토그램과 바 차트 중 그래프를 선택하는 방법 (예시)

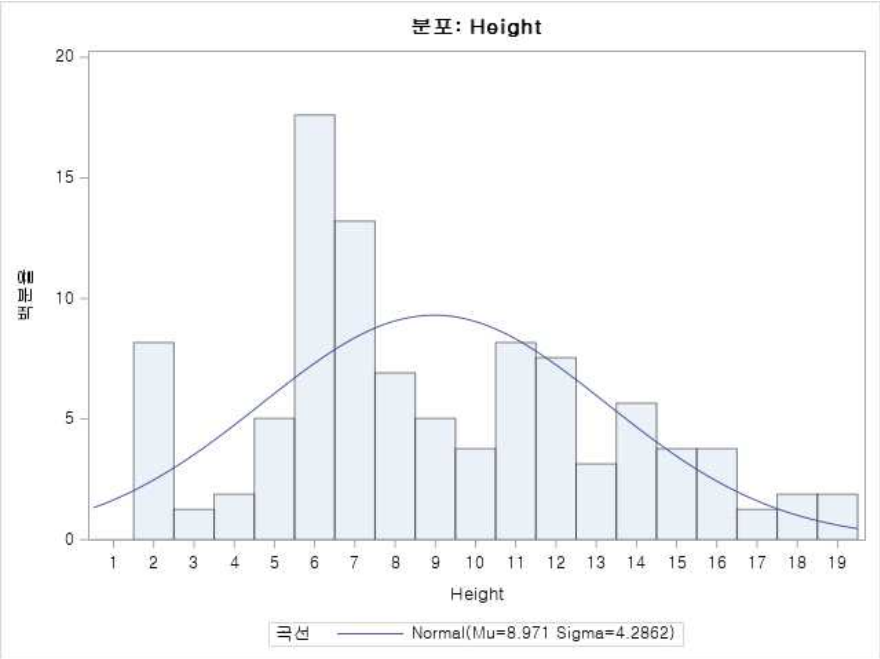
- 영화를 100편 시청하고 관련 그래프를 그려 데이터를 표현하고자 한다.
- 표현하는 방법 중 '장르 별 영화 편수' 혹은 '제작연도 별 영화 편수' 등으로 데이터를 표현할 수 있다.
- '장르 별 영화 편수'의 경우 '장르'가 분류된 불연속적 변수이므로, 바 차트를 통해 높이로 시청 편수의 빈도를 표현하는 것이 유리하다.
- '제작연도 별 영화 편수'의 경우 '연도'가 연속적 변수이므로, 히스토그램을 통해 너비로 시청 편수의 빈도를 표현하는 것이 유리하다.

4-4. 히스토그램

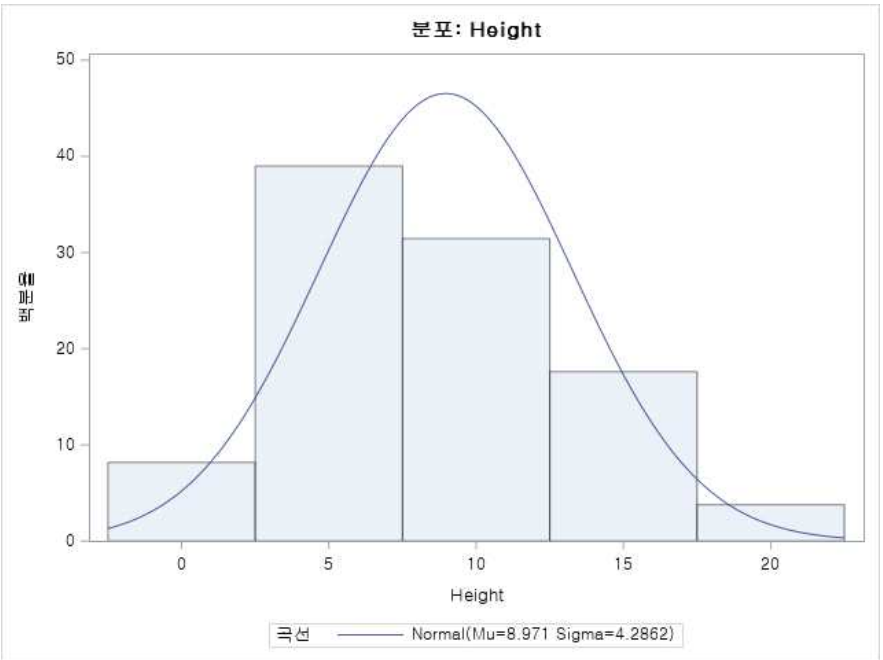
```
PROC UNIVARIATE DATA=[SAS 자료이름] NOPRINT;
        HISTOGRAM [변수이름]
        / NORMAL MIDPOINTS=num1 TO num2 BY num3;
RUN;
```

- NOPRINT: UNIVARIATE 기술통계량이 제공하는 여러 출력을 제거한다.
- HISTOGRAM: 해당 변수에 대한 히스토그램을 출력한다.
- NORMAL: 정규분포 곡선이 같이 출력된다.
- MIDPOINTS: 중간점이 되는 범위와 계급의 크기를 결정한다. 계급의 중간점을 지정하고, 각 계급 간 중간점이 BY 명령어 단위만큼 증가하며 분포표가 표현된다.

[히스토그램 MIDPOINTS 예시]



- MIDPOINTS=1 TO 19 BY 1;
- 1을 중간점으로 하는 계급이 선택되고, 19까지 중간점의 값이 1씩 증가
- 범위의 남은 값들을 표현하기 위해 자동으로 계급이 추가된다.



- MIDPOINTS=5 TO 10 BY 5;
- 표현해야 하는 데이터가 더 남은 경우, 해당 단위로 계속 더하여 표현된다.

4-5. 바 차트

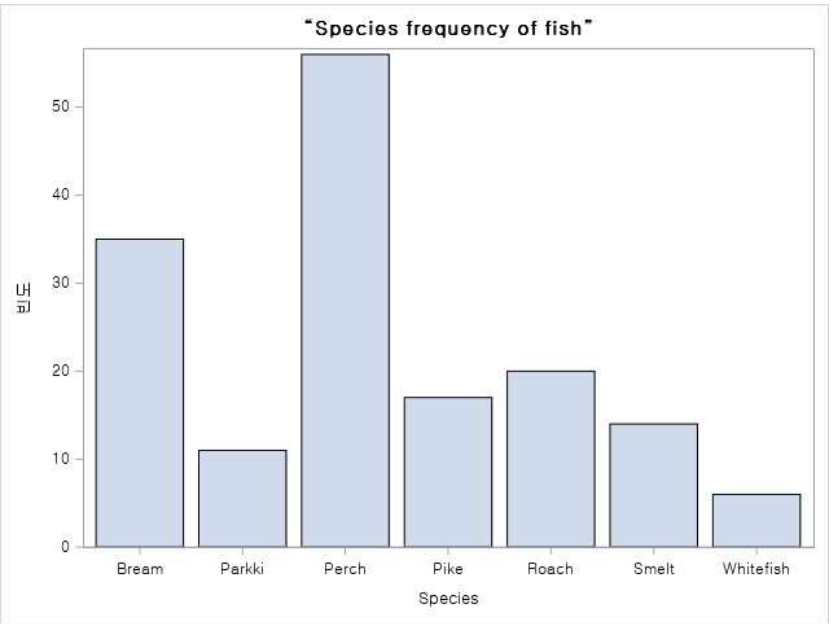
```
PROC SGPLOT DATA=[SAS 자료이름];
        VBAR [변수이름1] / GROUP=[변수이름2];
        TITLE "제목";
RUN;
```

- SGPLOT: 변수를 지정하는 명령어를 통해 다양한 그래프를 출력한다.
- VBAR: 바 차트 출력
- VBOX: 상자그림 출력
- HISTOGRAM: 히스토그램 출력
- SCATTER: 산점도 출력
- GROUP: 출력하고자 하는 변수의 종속변인이 되는 변수를 지정한다. 지정한 종속변인을 통해 각 세부 그룹으로 한 단계 더 구분되어 출력된다.
- TITLE: 그래프의 제목을 지정한다.

[바 차트 예시]

```
PROC SGPLOT DATA=SASHELP.FISH;
        VBAR SPECIES;
        TITLE "Species frequency of fish";
RUN;
```

[출력 결과]



4-6. 파이 차트

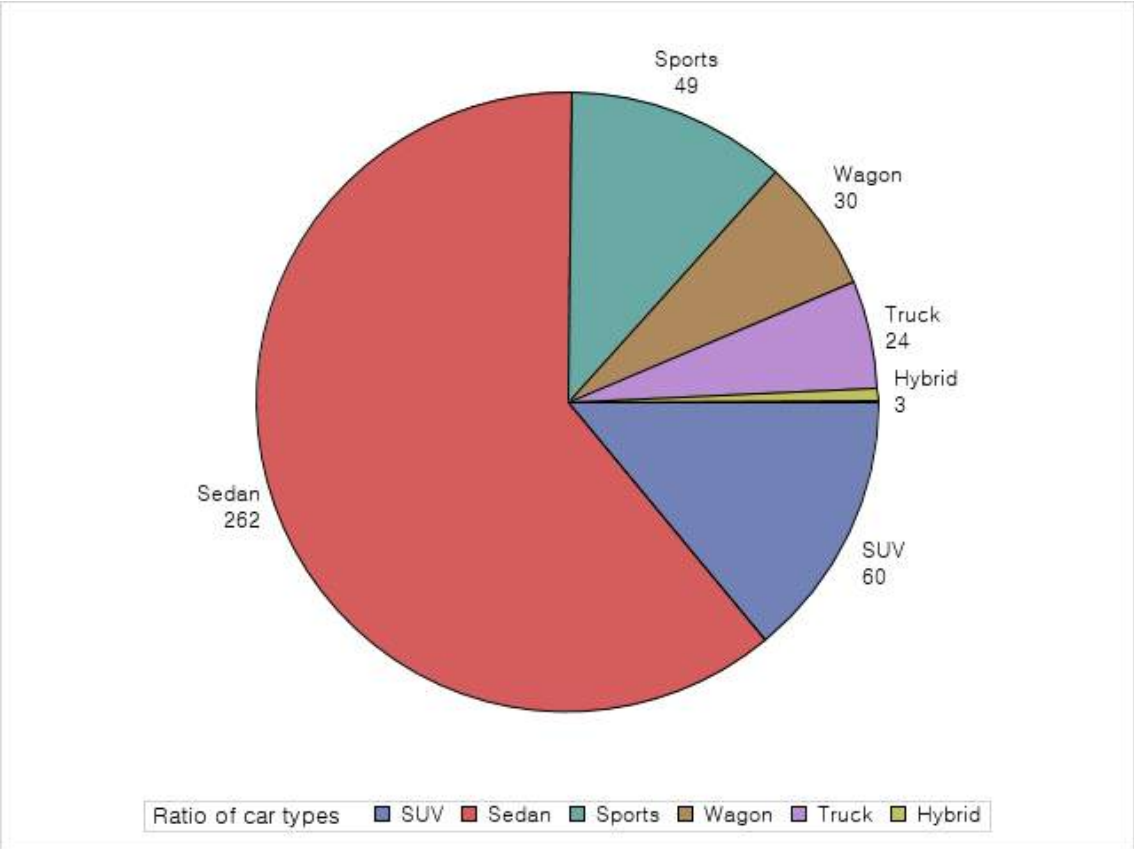
```
PROC TEMPLATE;
        DEFINE STATGRAPH [템플릿이름];
        BEGINGRAPH;
        LAYOUT REGION;
        PIECHART CATEGORY=[변수이름] /
                DATALABELLOCATION=[위치]
                CATEGORYDIRECTION=[방향]
                START=[시작각도] NAME=[차트이름];
                DISCRETELEGEND [차트이름] /
                TITLE= "범례 제목";
        TITLE "그래프 제목";
        ENDLAYOUT;
        ENDGRAPH;
END;

PROC SGRENDER DATA=[SAS데이터] TEMPLATE=[템플릿이름];
RUN;
```

- TEMPLATE: 파이 차트를 출력할 템플릿을 지정한다. 출력할 데이터의 변수를 미리 템플릿에 지정하여야 하며, 이후 템플릿에 출력하고자 하는 데이터를 연결하고 템플릿을 불러와 출력한다.
- DEFINE STATGRAPH: 데이터를 출력할 템플릿 이름을 지정한다.
- LAYOUT REGION: 출력하는 그래프의 레이아웃을 지정한다.
- PIECHART CATEGORY: 데이터의 출력할 변수를 미리 지정한다.
- DATALABELLOCATION: 차트에서 각 분류에 대한 범례와 실수치를 출력할 위치를 지정한다. 내부는 INSIDE, 외부는 OUTSIDE 옵션을 지정한다.
- CATEGORYDIRECTION: 차트를 채우는 데이터의 출력 방향을, 시계방향(CLOCKWISE) 혹은 반시계방향(COUNTERCLOCKWISE)으로 지정한다.
- START: 차트를 채우는 데이터의 출력 시작각도(0~360)를 지정한다.
- NAME: 범례를 따로 지정할 때 불러오기 위해 차트 이름을 지정한다.
- DISCRETELEGEND: 차트 이름을 통해 외부에서 범례를 따로 출력한다.
- DISCRETELEGEND / TITLE: 범례의 제목을 지정한다.
- TITLE: 전체 템플릿(차트)의 제목을 지정하여 출력한다.
- SGRENDER: 출력하고자 하는 데이터를 지정하고 미리 설정된 템플릿을 통해 차트를 출력한다.

[파이 차트 예시]

```
PROC TEMPLATE;
        DEFINE STATGRAPH pie;
        BEGINGRAPH;
        LAYOUT REGION;
        PIECHART CATEGORY=TYPE /
                DATALABELLOCATION=OUTSIDE
                CATEGORYDIRECTION=CLOCKWISE
                START=0 NAME='pie';
                DISCRETELEGEND 'pie' /
                TITLE= "Ratio of car types";
        TITLE "Ratio of car types";
        ENDLAYOUT;
        ENDGRAPH;
END;

PROC SGRENDER DATA=SASHELP.CARS TEMPLATE=pie;
RUN;
```

[출력 결과]



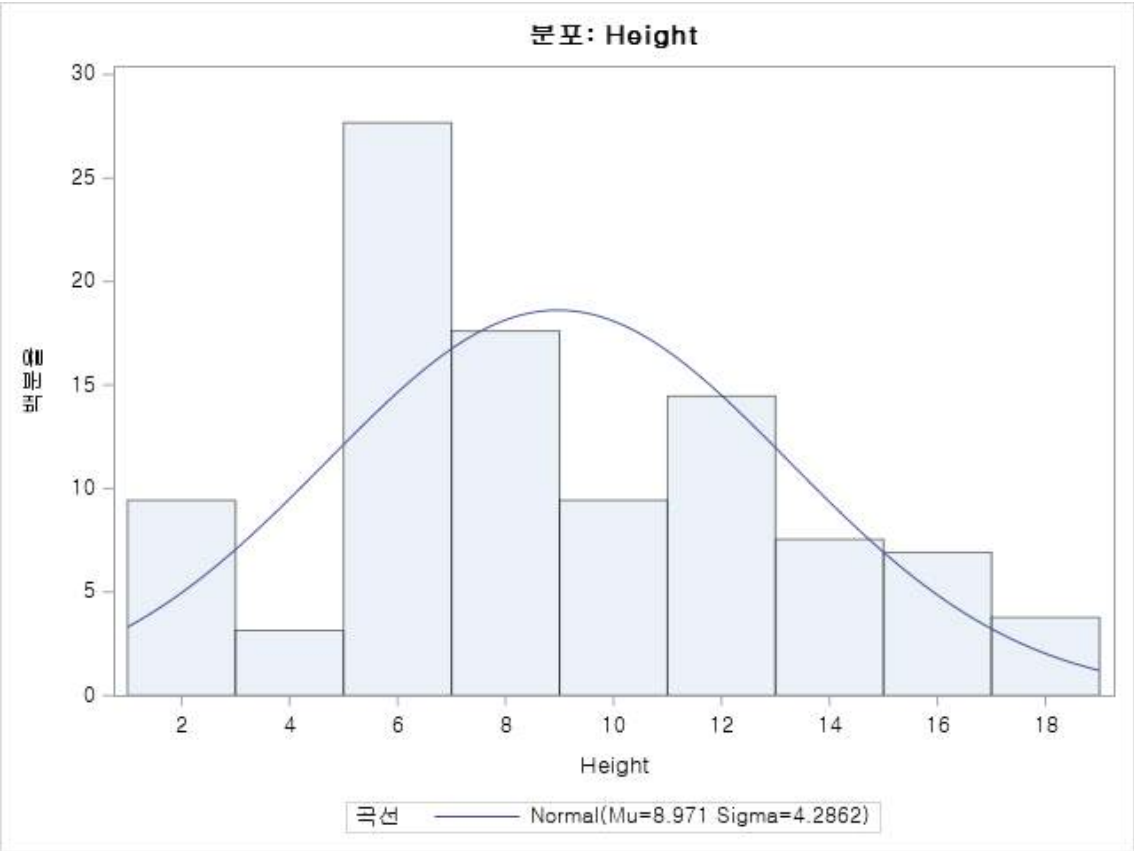
<실습1> 히스토그램

[히스토그램 예시]

```
DATA HFISH;
        SET SASHELP.FISH;
        KEEP HEIGHT;

PROC MEANS DATA=HFISH MAX MIN;

PROC UNIVARIATE DATA=HFISH NOPRINT;
        HISTOGRAM HEIGHT
        / NORMAL;
RUN;
```

- SASHELP의 FISH 데이터에서 HEIGHT 변수의 히스토그램을 출력한다.
- MEANS의 MAX, MIN 옵션을 통해 전체 데이터의 범위를 파악한다.
- NORMAL 옵션을 통해 분포 곡선을 추가한다.

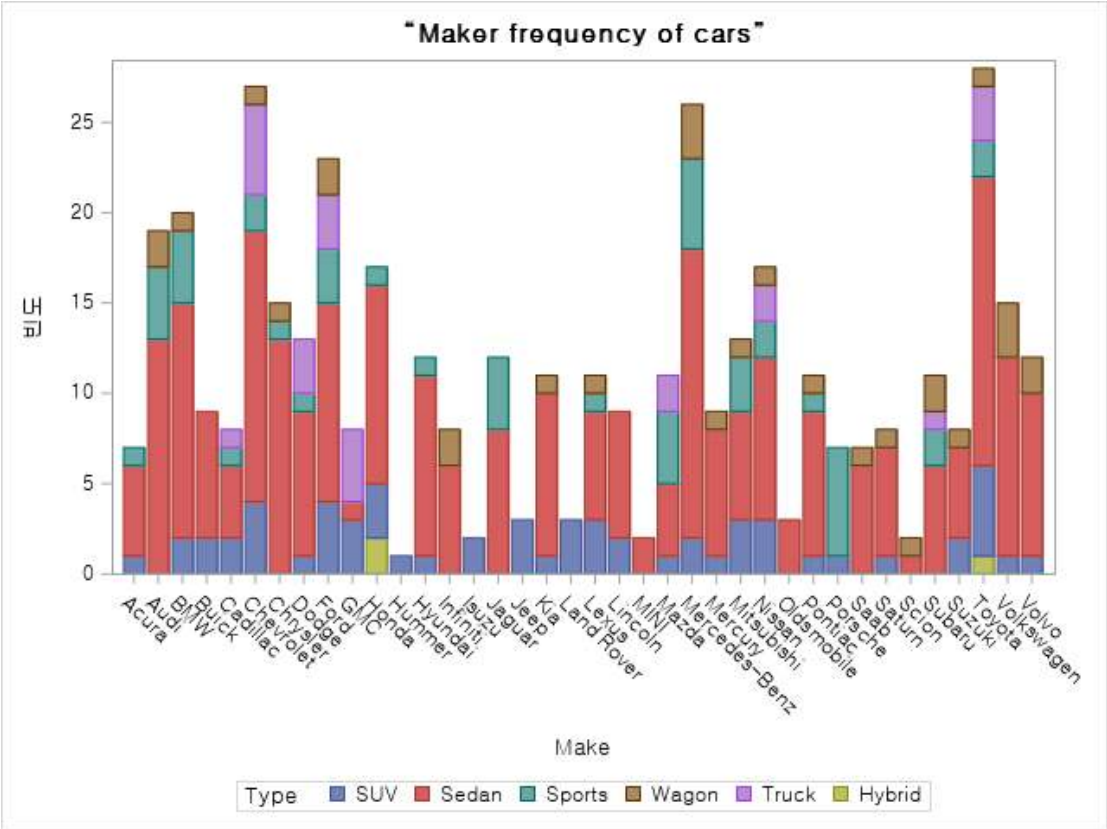
[출력 결과]



<실습2> 바 차트

- 자동차 브랜드와 그 차종에 따른 빈도를 바 차트로 표현한다.
- 여러 자동차 데이터(SASHELP.CARS)에서 각 자동차의 브랜드(MAKE) 별로 빈도를 출력한다. 또한 동시에 각 브랜드의 내부적으로 자동차 종류(TYPE)를 구분하여 '중첩된' 바 차트로 출력되도록 한다.
- 차트의 제목은 "Maker frequency of cars"로 지정한다.

[출력 결과]

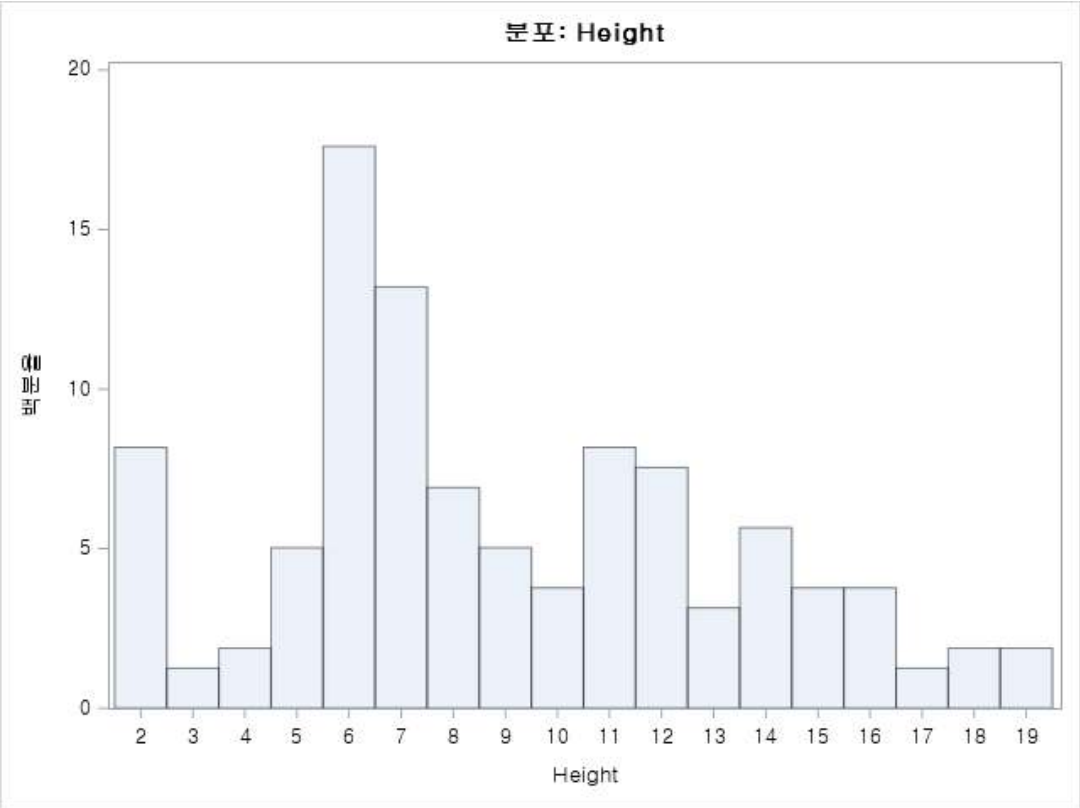


<마무리> SAS 그래프 출력

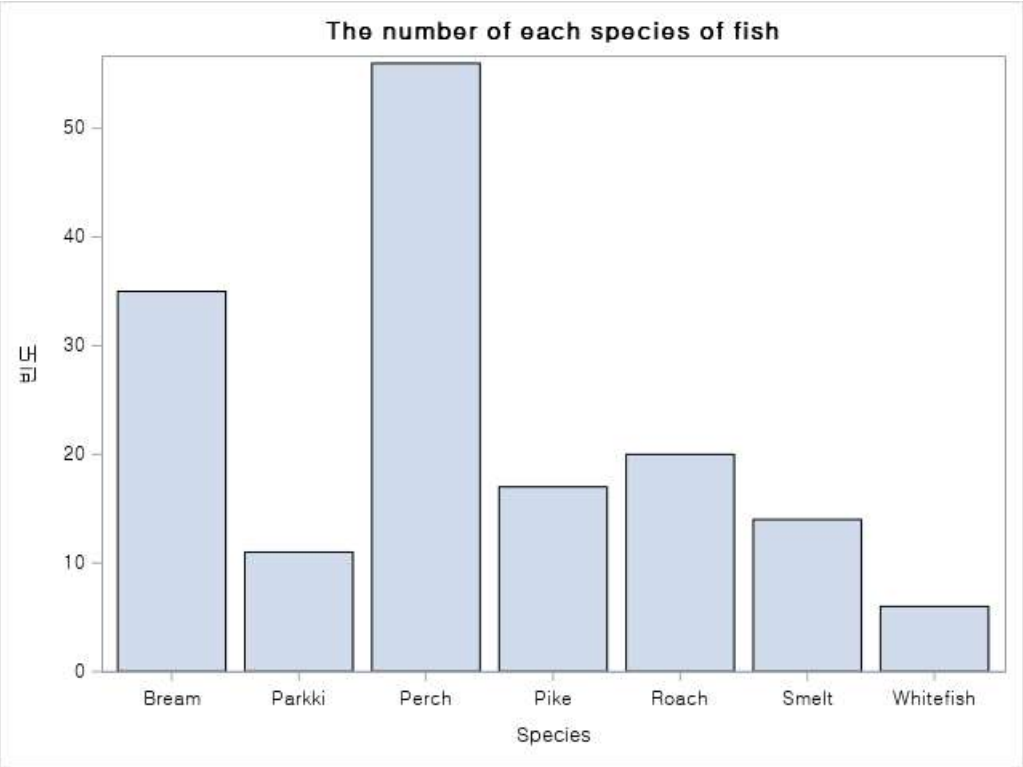
- SASHELP 라이브러리에 있는 FISH 데이터를 이용한다.
- 아래 출력 결과와 같은 결과가 나오도록 각 그래프를 출력한다.

- 물고기의 길이(HEIGHT)에 따른 히스토그램을 출력한다.
- 물고기의 각 종(SPECIES)에 따른 개체 수를 바 차트로 출력한다.
- 물고기의 개체 수의 비율을 나타내는 파이 차트를 출력한다.

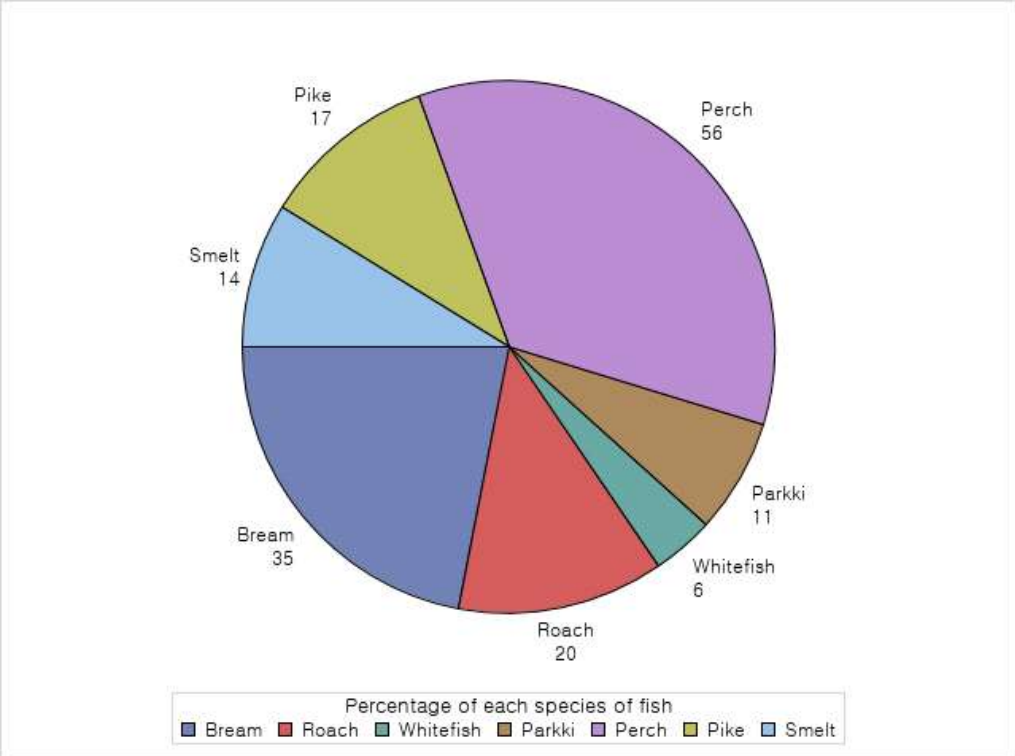
[히스토그램 출력 결과]



[바 차트 출력 결과]



[파이 차트 출력 결과]



<과제3> SAS 그래프 출력

(1) SASHELP 라이브러리의 HOLIDAY 데이터를 이용한다.

- HOLIDAY 데이터에는 각 휴일과 날짜들이 저장되어 있다. 해당 데이터 중 가장 많은 휴일이 있는 달을 한 눈에 확인하고자 한다. 여러 종류의 그래프 중 적절한 그래프를 선택하여 가장 많은 휴일이 있는 달을 확인하자.

(2) SASHELP 라이브러리의 SHOES 데이터를 이용한다.

- SHOES 데이터에는 지역, 제품 종류 등 신발 제품에 대한 데이터가 있다.
- 중첩 바 차트를 통해, 각 제품 종류별 빈도를 출력하고 내부적으로는 지역별 빈도를 출력하여, 동시에 제품 종류 및 지역별 비율을 확인하자.

(중요) 과제 제출

- 작성한 코드 파일(.sas)과 출력된 결과(캡처 혹은 저장)의 PDF(지원되지 않는 경우 Word 혹은 HWP 이용), 2개의 파일을 압축(.zip)하여 아래 서식으로 이름을 정하고 이메일 제목 또한 동일하게 하여 제출한다.
- 파일 이름: 통계학실습_과제X_학번_이름.zip
- 이메일 제목: 통계학실습_과제X_학번_이름
- 제출 이메일: gtsk623@gmail.com

5. SAS 통계분석

5-1. 통계분석

- 가설이란 어떤 자료들을 근거로 유추하여 예측되는 잠정적인 결론을 의미한다. 가설은 예측이기 때문에 논리적으로 혹은 실험적으로 증명이 되어야 한다. 이러한 가설을 검증하기 위해 사용하는 방법이 통계분석방법이다.
- 통계분석이란, 특정집단으로부터 자료를 수집하고 대상에 대한 적절한 통계분석방법을 활용하여 통계적 추론으로써 가설을 검증하는 것을 의미한다.
- 통계분석방법에는 빈도분석, 평균분석, 분산분석, 상관분석, 회귀분석 등이 존재한다. 각 분석 방법의 차이는 아래와 같다.

[통계분석방법의 종류]

통계분석방법	분석방법의 목적
빈도분석	집단 간의 빈도 비교를 통한 분석
평균분석	두 집단 간의 평균 비교를 통한 분석
분산분석	둘 이상 집단 간의 분산 비교를 통한 분석
상관분석	두 변수 간 상관관계 분석
회귀분석	독립변수와 종속변수의 인과관계 분석

5-2. 상관관계와 인과관계

- 상관관계란, 변수들 간의 상관성의 유무를 나타낸다. 즉, 두 변수가 얼마나 강하게 혹은 약하게 관계되어 있는지를 파악하는 용도이다.
- 인과관계란, 독립변수가 종속변수에 대한 영향의 유무를 나타낸다. 즉, 원인인 여러 독립변수로 인한 종속변수의 결과를 파악하는 용도이다.
- 예를 들어, 심장병과 암 발병은 ‘상관’이 있다고 할 때, 심장병과 암 발병 중 어떤 것이 원인이고 결과인지, 혹은 다른 어떤 변수가 해당 두 변수에 모두 영향을 미치는지 등은 좀 더 분석해봐야 알 수 있다. 따라서 상관관계가 인과관계보다 더 큰 개념을 나타내고 있음을 알 수 있다.

5-3. 상관분석

- 상관분석은 상관계수를 이용하여 상관관계의 유무와 정도를 파악할 수 있는 분석방법이다. 상관계수는 -1에서 +1 사이의 값을 가지며, 절댓값이 1에 가까울수록 두 변수는 더 강한 상관관계를 가지고 있다. 또한, -1에 가까울수록 음의 상관관계를 띄며, +1에 가까울수록 양의 상관관계를 띈다.
- 상관계수란, 두 변수 간의 상관관계의 정도를 나타내는 계수이다. 상관계수를 통해 두 변수가 함께 변화하는 경향이 있는 범위를 측정하고 상관관계의 정도와 방향 등을 설명할 수 있다. 상관계수는 여러 방법으로 계산되는데 대표적으로 Pearson 상관계수와 Spearman 상관계수가 이용된다.
- Pearson 상관계수는 두 변수가 모두 정규성을 따른다는 가정이 필요하며, 두 변수 관계의 선형성을 반영한 계수이다. 선형성은 두 변수가 함께 변화하며 변화하는 비율을 포함한다. 따라서 이 상관계수는 두 변수의 값이 함께 변하는 정도 혹은 따로 변하는 정도 등을 나타낸다.
- Spearman 상관계수는 두 변수가 정규성을 따르지 않는 경우 사용되며, 두 변수 관계의 단조성을 반영한 계수이다. 단조성은 두 변수가 함께 변화하는 경향이 있되 반드시 일정한 비율로 변화하지 않는 것을 포함한다. 따라서 이 상관계수는 두 변수 값에 순위를 매겨 순위에 대해 변하는 정도 등을 나타낸다.

5-4. SAS 상관분석

```
PROC CORR DATA=[SAS자료이름] PEARSON SPEARMAN;
      VAR [변수이름1] [변수이름2];
RUN;
```

- PROC CORR: 두 변수 간 단순통계량과 상관계수를 나타내는 프로시저
- PEARSON, SPEARMAN: 해당하는 상관계수를 나타내는 옵션
- VAR: 관계를 분석할 두 변수를 입력

[SAS 상관분석 예시]

```
PROC CORR DATA=SASHELP.CARS PEARSON SPEARMAN;
      VAR HORSEPOWER ENGINESIZE;
RUN;
```

[실행 결과]

CORR 프로시저

2 개의 변수:	Horsepower EngineSize

단순 통계량

변수	N	평균	표준편차	중위수	최솟값	최댓값	레이블
Horsepower	428	215.88551	71.83603	210.00000	73.00000	500.00000	
EngineSize	428	3.19673	1.10859	3.00000	1.30000	8.30000	Engine Size (L)

피어슨 상관 계수, N = 428
H0: Rho=0 가정하에서 Prob > |r|

	Horsepower	EngineSize
Horsepower	1.00000	0.78743 <.0001
EngineSize Engine Size (L)	0.78743 <.0001	1.00000

스피어만 상관 계수, N = 428
H0: Rho=0 가정하에서 Prob > |r|

	Horsepower	EngineSize
Horsepower	1.00000	0.80774 <.0001
EngineSize Engine Size (L)	0.80774 <.0001	1.00000

- Pearson 상관계수=0.78743, Spearman 상관계수=0.80774
- 상관계수가 모두 양수이며, 1에 가깝기 때문에 강한 양의 상관관계
- 마력(HORSEPOWER)과 엔진크기(ENGINESIZE)는 상관성이 있는 것으로 예측할 수 있는데, 상관계수를 통해 해당 예측을 확인할 수 있다.

5-5. SAS 상관관계 매트릭스

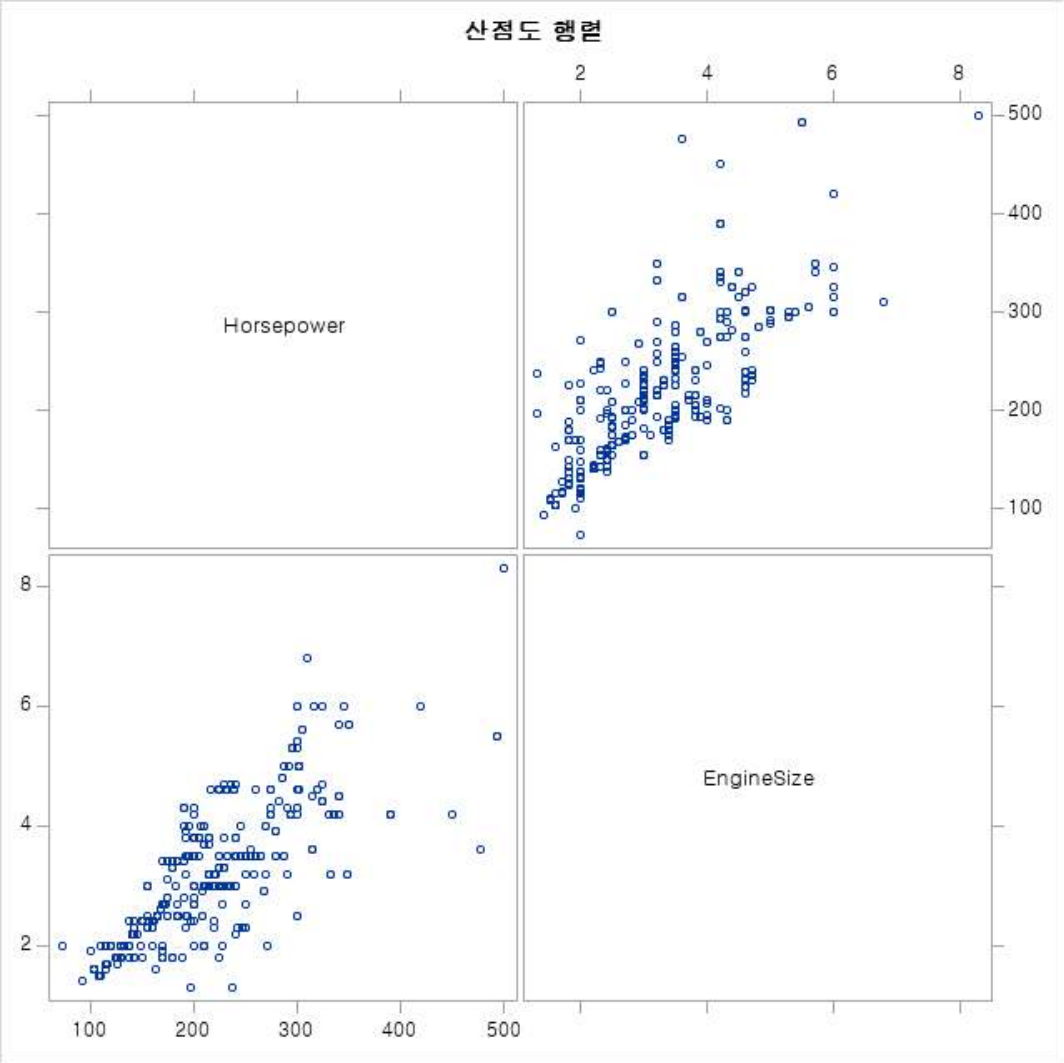
```
PROC CORR DATA=[SAS자료이름] PLOTS=MATRIX;
      VAR [변수이름1] [변수이름2];
RUN;
```

- PLOTS=MATRIX: 두 변수 간의 상관관계 매트릭스를 출력한다.

[SAS 상관관계 매트릭스 예시]

```
PROC CORR DATA=SASHELP.CARS PLOTS=MATRIX;
      VAR HORSEPOWER ENGINESIZE;
RUN;
```

[실행 결과]



5-6. 회귀분석

- 회귀분석은 독립변수가 종속변수에 미치는 영향에 대해 파악할 수 있는 분석방법이다. 관측된 데이터로부터 여러 독립변수들과 종속변수 간의 인과관계를 함수식으로 표현하여 설명한다. 회귀분석에서는 독립변수와 종속변수로 구한 상관계수의 제곱 값인 결정계수를 이용하여 독립변수로부터 의미있는 종속변수를 예측할 수 있는지 판별한다.
- 회귀분석에는 단순회귀분석과 다중회귀분석이 존재한다. 단순회귀분석은 독립변수가 한 개인 회귀분석이며, 다중회귀분석은 독립변수가 여러 개인 회귀분석을 의미한다.

5-7. SAS 단순회귀분석

```
PROC REG DATA=[SAS자료이름];
      MODEL [종속변수이름]=[독립변수이름];
RUN;
```

- PROC REG: 회귀분석 결과를 나타내는 프로시저
- MODEL: 종속변수와 독립변수를 지정

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1366287	1366287	695.21	<.0001
Error	426	837210	1965.28281		
Corrected Total	427	2203497			

Root MSE	44.33151	R-Square	0.6201
Dependent Mean	215.88551	Adj R-Sq	0.6192
Coeff Var	20.53473		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	52.77197	6.54692	8.06	<.0001
EngineSize	Engine Size (L)	1	51.02514	1.93520	26.37	<.0001

- 분산분석표(Analysis of Variance : ANOVA): 자료에서 추정된 회귀직선 모델(함수식)이 실제 데이터를 잘 설명하는지 파악하는 값을 나타낸다.
- 계산된 'R-Square'는 0과 1 사이의 값으로 현재 데이터가 전체를 얼마나 잘 설명하는지 나타내는 결정계수이며, 계산된 'Adj R-square'는 변수가 많아질수록 결정계수가 증가하는 단점을 보완하기 위한 수정결정계수이다.

[SAS 단순회귀분석 예시]

```
PROC REG DATA=SASHELP.CARS;
      MODEL HORSEPOWER=ENGINESIZE;
RUN;
```

[실행 결과]

The REG Procedure
Model: MODEL1
Dependent Variable: Horsepower

Number of Observations Read	428
Number of Observations Used	428

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1366287	1366287	695.21	<.0001
Error	426	837210	1965.28281		
Corrected Total	427	2203497			

Root MSE	44.33151	R-Square	0.6201
Dependent Mean	215.68551	Adj R-Sq	0.6192
Coeff Var	20.53473		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	52.77197	6.54692	8.06	<.0001
EngineSize	Engine Size (L)	1	51.02514	1.93520	26.37	<.0001

- 독립변수 엔진크기(ENGINESIZE)와 종속변수 마력(HORSEPOWER) 사이의 인과관계를 설명할 수 있는 값들을 출력한다.
- 가설검정 과정(귀무가설 기각 혹은 채택)을 통해 독립변수가 종속변수를 설명하는데 유의한 영향을 끼치는지 등을 파악할 수 있다.
- ‘Pr > F’와 ‘Pr > |t|’에서 설명되는 유의확률이 0.0001보다 작기 때문에 귀무가설을 기각할 수 있고, 독립변수로 종속변수를 유의미하게 설명할 수 있다는 뜻이 된다.
- 결정계수 R-Square 값이 0.6201이므로 현재 모델은 전체 데이터의 약 62%를 설명하고 있다는 뜻이 된다.

<실습1> 상관분석

- SASHELP.CARS 데이터에서 자동차의 마력(HORSEPOWER)과 자동차의 길이(LENGTH) 간의 상관계수를 구하여 상관관계의 정도를 파악한다.

[출력 결과]

CORR 프로시저

2 개의 변수:	Horsepower Length

단순 통계량

변수	N	평균	표준편차	합	최솟값	최댓값	레이블
Horsepower	428	215.88551	71.83603	92399	73.00000	500.00000	
Length	428	186.36215	14.35799	79763	143.00000	238.00000	Length (IN)

피어슨 상관 계수, N = 428
H0: Rho=0 가정하에서 Prob > |r|

	Horsepower	Length
Horsepower	1.00000	0.38155 <.0001
Length Length (IN)	0.38155 <.0001	1.00000

- 상관계수=0.38155
- 상관계수가 양수이며 1보다 0에 더 가깝다.
- 자동차의 마력과 길이는 약한 양의 상관관계를 가진다.

<실습2> 단순회귀분석

- SASHELP.HEART 데이터에서 콜레스테롤(Cholesterol)이 몸무게(Weight)에 영향을 미치는지 단순회귀분석을 통해 파악한다.
- 데이터의 양을 조절하기 위해 SET 데이터를 통해 (OBS=1000) 옵션을 추가하여 데이터 행(row)의 개수를 1,000개로 줄인다.

[출력 결과]

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	1000
Number of Observations Used	931
Number of Observations with Missing Values	69

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	798.95468	798.95468	1.06	0.3029
Error	929	698577	751.96630		
Corrected Total	930	699376			

Root MSE	27.42200	R-Square	0.0011
Dependent Mean	152.01933	Adj R-Sq	0.0001
Coeff Var	18.03850		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	147.30953	4.65676	31.63	<.0001
Cholesterol	1	0.02051	0.01990	1.03	0.3029

- 유의확률이 큰 수준이며 R-Square 값이 매우 작으므로, 귀무가설을 기각하기 어렵고 해당 데이터가 전체 데이터를 설명하기 어렵다는 뜻이 된다.
- 따라서 데이터에서 추출된 1,000개의 데이터로는 몸무게(Weight) 변수를 설명하는 것에 있어 콜레스테롤(Cholesterol)이 유의한 변수라고 보기 어렵다는 것을 나타낸다.

<과제4> SAS 통계분석

(1) SASHELP 라이브러리의 BASEBALL 데이터를 이용한다.

http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_sashelp_sect002.htm

(2) 관계를 확인하고 싶은 두 변수를 자유롭게 선택한다.

(3) 두 변수에 대하여 상관분석(매트릭스) 및 회귀분석을 수행하고 출력한다.

- 상관분석에서는 두 변수 간의 상관관계를 분석한다.
- 회귀분석에서는 두 변수 중 한 변수를 독립변수로, 나머지 하나를 종속변수로 설정하여 인과관계를 분석한다.

(4) 두 분석 절차와 분석 결과를 아래 내용을 포함한 보고서로 작성한다.

- 제목: 야구에서 '변수1'과 '변수2'의 관계 (지정한 변수들로 제목 기재)
- 학번, 이름
- 상관관계, 인과관계 예측 (예를 들어, 두 변수는 강한 양의 상관관계를 띨 것이다, 이 독립변수는 저 종속변수에 유의미한 영향을 줄 것이다.)
- 상관분석과 회귀분석 코드 (글 혹은 이미지)
- 상관분석과 회귀분석 출력 결과 (이미지)
- 상관분석 결과내용: 상관계수로부터 강한/약한 양/음의 상관관계 파악
- 회귀분석 결과내용: 유의확률과 결정계수로부터 인과관계의 유의미성 파악

(중요) 과제 제출

- 작성한 보고서를 PDF로 만들고 압축하여, 아래 서식으로 이름을 정하고 이메일 제목 또한 동일하게 하여 제출한다.
- 파일 이름: 통계학실습_과제X_학번_이름.zip
- 이메일 제목: 통계학실습_과제X_학번_이름
- 제출 이메일: gtsk623@gmail.com