통계학 실습

4. 그래프 출력

3주차 – 기술통계 UNIVARIATE, MEANS

- UNIVARIATE와 MEANS 명령어에서 기술통계량을 볼 때, 기준이 되는 변수의 내부 값이 하나로만 되어 있을 때는 'BY'를, 여러 값을 가지고 있을 때는 'CLASS'를 사용합니다.

- Example)
  - SASHELP.CLASS에서 'AGE' 변수는 나이를 여러 개 가지고 있으므로, 각 나이에 대한 분석을 보고 싶다면 CLASS AGE; 를 통해 확인하여야 합니다.

그래프 출력

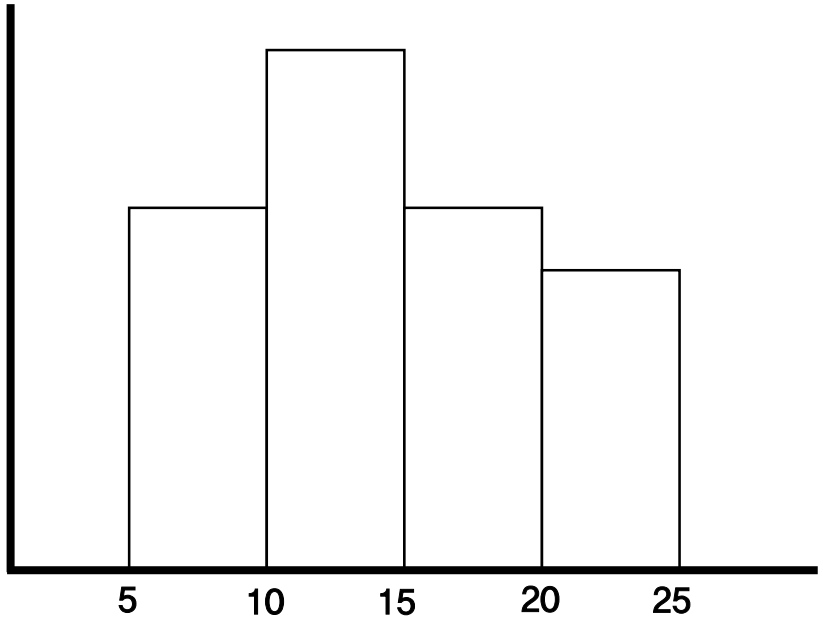
- 데이터를 분석하는 것도 중요하지만 분석된 내용을 전달하기 쉽게 표현하는 것도 중요하다. 그래프를 이용하여 한 눈에 보기 쉽게 내용을 전달하는 것이 좋다.

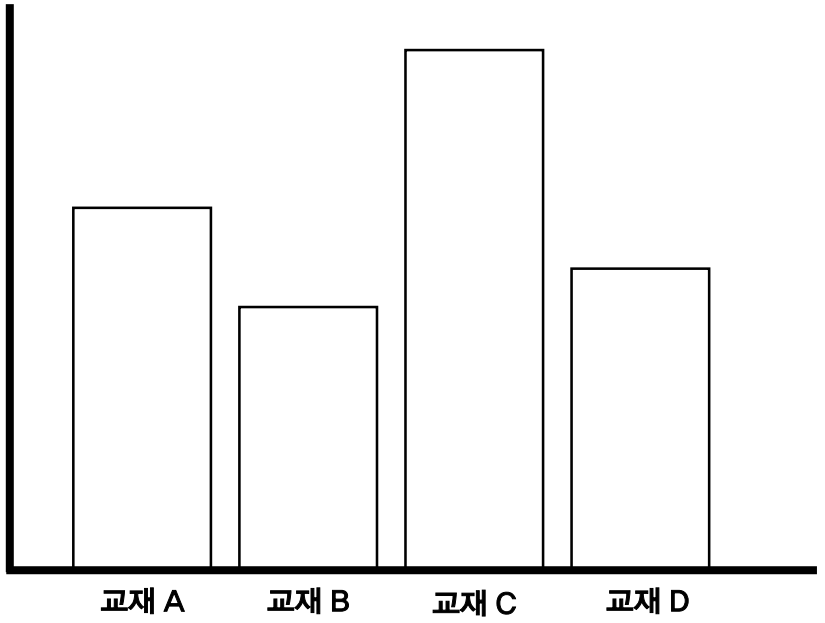
- 다양한 그래프 중에 데이터를 분석하여 히스토그램, 바 차트, 원형 차트를 그려본다.

다양한 그래프

- 히스토그램
  - 도수 분포표의 하나로써 가로축의 빈도를 보기 위한 그래프이다. 도수 분포의 상태를 직사각형의 기둥 모양으로 나타낸 형태이다.
- 바 차트
  - 히스토그램과 마찬가지로 가로축의 빈도를 보기 위한 그래프이며, 똑같이 직사각형의 기둥 모양으로 나타낸 형태이다. 히스토그램과 바 차트의 차이는 가로축에 존재하는 데이터 간의 연속성이 존재하느냐 이다. (다음 장에 설명)
- 원형 차트
  - 전체에서 해당 데이터가 차지하는 비율을 부채꼴 모양으로 나타낸 그래프이다. 모든 데이터가 전체 원에 해당하며, 부채꼴의 중심각이 전체에서 차지하는 비율을 나타낸다.

히스토그램과 바 차트의 차이





- 히스토그램
  - 히스토그램은 막대가 서로 붙어 있다. 따라서 가로축에서 비교가 중요하며, 해당 데이터들이 서로 간에 연관성(연속성) 등이 존재하게 된다.
  - 예시로 학생들 수에 따른 수업에 대한 이해도가 될 수 있다.
- 바 차트
  - 바 차트는 굳이 가로축의 크기(너비)가 중요한 것이 아니다. 각 가로축 데이터 간의 서로 다른 항목들이 존재하게 된다.
  - 예시로 서로 다른 교재에 따른 선호하는 학생들의 수 등이 될 수 있다.

히스토그램

```
PROC UNIVARAITE DATA=데이터 NOPRINT;
  HISTOGRAM 변수명
  / MIDPOINTS = 숫자 TO 숫자 BY 숫자;
RUN;
```

- NOPRINT : UNIVARAITE 출력 제거
- HISTOGRAM : 해당 변수를 히스토그램에 출력해준다.
- MIDPOINTS : 가로축에 해당되는 자료의 단위 값을 BY ~ 에 의하여 설정할 수 있다. 숫자 TO 숫자는 히스토그램의 중간들의 값의 범위가 된다.
  - 200 TO 300 BY 30이면, 해당 자료가 30 단위로 히스토그램이 그려지며, 200 ~ 300에 해당하는 자료가 히스토그램의 중앙에 놓이게 된다.

히스토그램 예시와 커브 피팅

```
PROC UNIVARIATE DATA=SASHELP.CARS
NOPRINT;
  HISTOGRAM horsepower
  / NORMAL MIDPOINTS = 70 TO 500 BY 50;
RUN;
```

- 히스토그램의 옵션에서 NORMAL을 이용하면 해당 히스토그램에 분포 곡선이 추가된다.

바 차트

```
PROC SGPLOT DATA=데이터;
  VBAR 변수명;
  TITLE “제목” ;
RUN;
```

- SGPLOT : 바 차트를 출력하여 준다.
- VBAR : 그래프를 출력하는데 사용되는 변수가 따라온다.
- TITLE : 바 차트의 제목이 설정된다.

바 차트 예시

```
PROC SQL;
create table CARS_DATA as
SELECT horsepower
FROM SASHELP.CARS
WHERE make in ('Audi', 'BMW');
RUN;


PROC SGPLOT DATA=work.CARS_DATA;
  VBAR horsepower;
  TITLE “horsepower of cars” ;
RUN;
```

- SGPLOT을 통해 자동차의 마력을 바 차트로 출력한다.
  - 위의 PROC SQL은 단순히 추가하면 된다.

누적 바 차트

```
PLOC SGPLOT DATA=SAS데이터;
  VBAR 변수명1 / GROUP = 변수명2;
  TITLE “제목” ;
RUN;
```

- 누적 바 차트는 데이터에서 어떤 변수가 다른 변수에 대해 산출된 그래프로써 분류별 비교가 가능하게 된다.
- GROUP : VBAR에 해당하는 변수명을 GROUP 변수에 따라 묶어 표현한다.
- GROUP을 포함하여 한 막대 내에 여러 분류들이 서로 다른 색깔로 모두 출력된다.
  - ex) ‘자동차 종류별 길이의 빈도’는 한 막대에 여러 자동차 종류가 모두 포함되어 출력된다.

원형 차트

```
PROC TEMPLATE;
  DEFINE STATGRAPH pie;
    BEGINGRAPH;
      LAYOUT REGION;
        PIECHART CATEGORY = 변수명 /
          DATALABELLOCATION = OUTSIDE
          CATEGORYDIRECTION = CLOCKWISE
          START = 180 NAME = 'pie';
        DISCRETELEGEND 'pie' /
          TITLE = ‘제목’ ;
      ENDLAYOUT;
    ENDGRAPH;
  END;
RUN;
```

- 좌측 코드를 통하여 그려질 원형 차트를 미리 설정해두고, 다음 장의 코드를 이용하여 데이터를 선택하여 원형 차트를 출력한다.
- 좌측 코드에서 CATEGORY에 차트에서 보고자 하는 변수명을 넣어 그 비율을 확인한다.

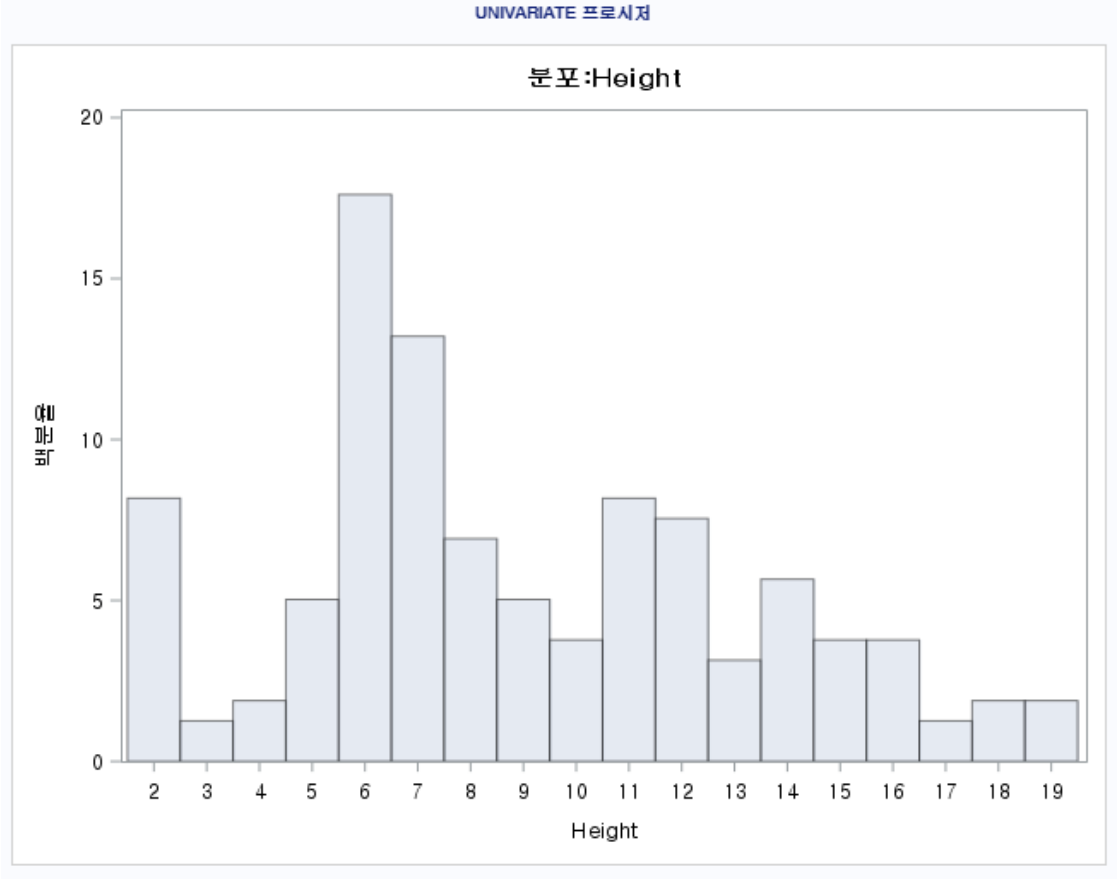
원형 차트 (계속)

```
PROC SGRENDER DATA=SAS데이터
  TEMPLATE = pie;
RUN;
```

- 여기서 불러온 데이터에서 원형 차트로 보고 싶은 항목이 이전 코드의 CATEGORY에 해당된다.
- 원형 차트로 만들 데이터를 선택하여, 전에 만들어진 코드에 이어 쓰면 원형 차트가 출력된다.
- 여기서 TEPLATE는 SGRENDER에 속한 옵션이다.

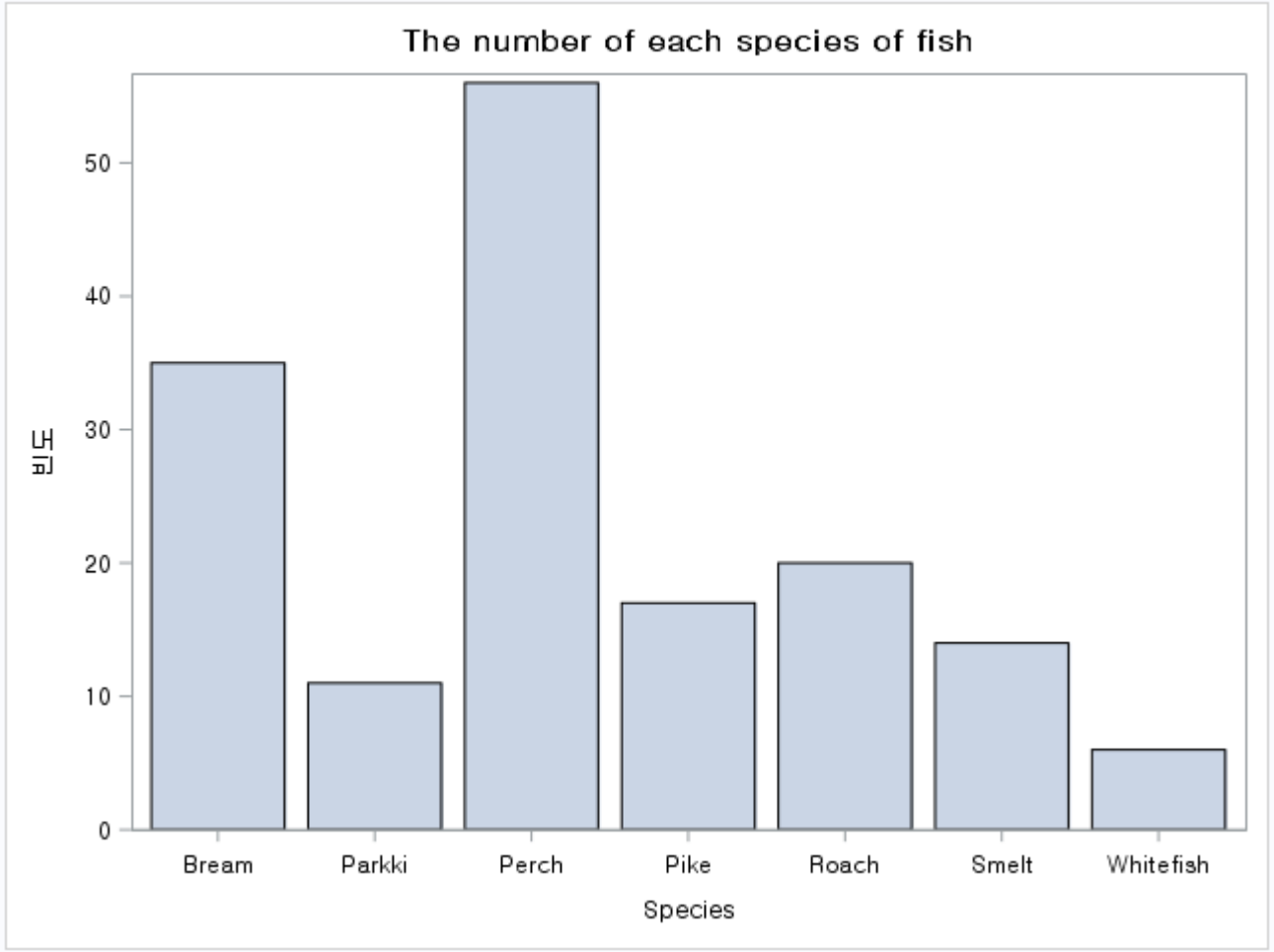
실습

- 1. SASHELP에 있는 FISH 라이브러리를 확인하여 column 명을 확인한다.
- 2. 물고기의 길이(Height)에 따른 히스토그램을 그려본다.



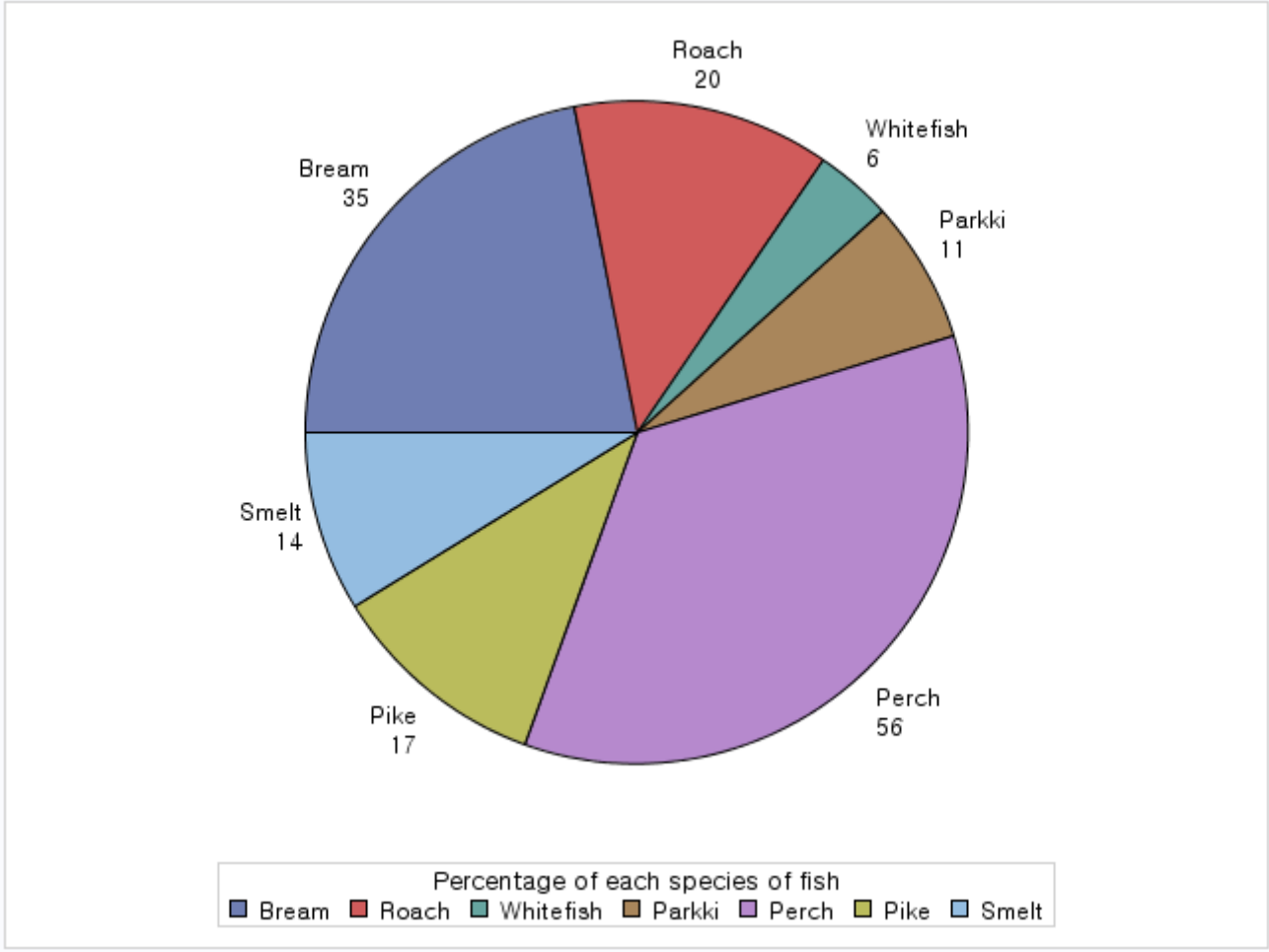
실습 (2)

- 3. 물고기의 각 종에 대한 개체 수를 바 차트로 그려본다.



실습 (3)

- 4. 물고기의 개체 수의 비율을 확인할 원형 차트를 그려본다.



문제

- SASHELP 라이브러리의 Holiday 데이터에는 각 휴일과 날짜들이 저장되어 있다. 해당 데이터에서 ‘월별 휴일의 수’를 통하여 가장 많은 휴일이 있는 달을 한 눈에 확인하고 싶다. 여러 종류의 그래프 중 하나를 선택하여 보고자 하는 데이터를 한 눈에 보이도록 해보자.