제 2장. 모집단과 표본

기술 통계학 (descriptive statistics)
: 자료의 특성을 표, 그림, 통계량 등을 사용하여 쉽게 파악할 수 있도록 정리, 요약하는 방법을 다루는 분야

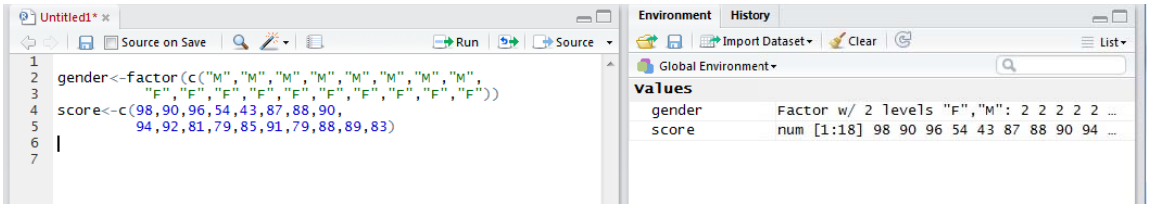
2.1 일변량 자료의 요약 - 그래프를 이용한 요약

2.1.1 자료 입력하기

다음은 18명의 학생들의 성별과 점수에 대한 자료이다.

성별	M	M	M	M	M	M	M	M	F	F	F	F	F	F	F	F	F	F
점수	98	90	96	54	43	87	88	90	94	92	81	79	85	91	79	88	89	83

다음과 같이 자료를 입력한다.

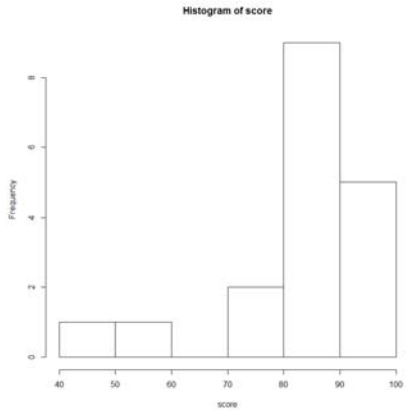


성별 자료의 경우, 문자로 표시된 범주형 자료이기 때문에 factor() 함수를 사용하여 요인(factor) 변수임을 나타내었다. 즉, gender 변수는 "M"과 "F"의 두 가지 값을 갖는 범주형 자료가 된다.

2.1.2 히스토그램

일변량 자료의 분포를 알아보는데 유용한 그래프는 히스토그램이다. 히스토그램은 hist(x)의 명령으로 그린다. 결과는 plots 창에서 확인할 수 있다.

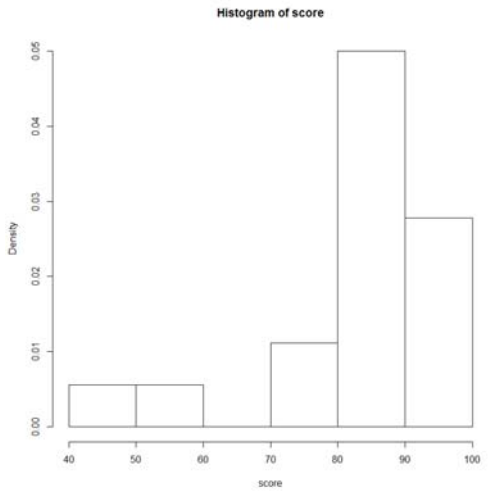
```
> hist(score)
```



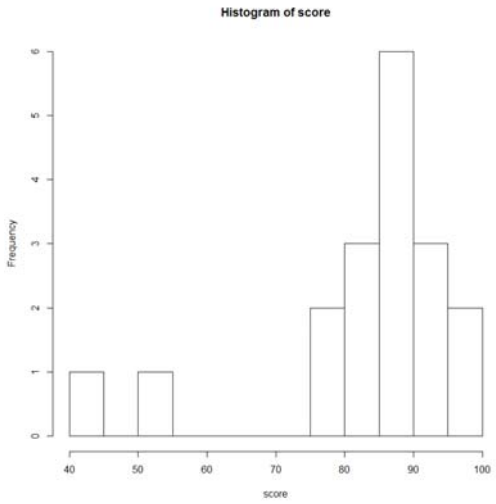
R에서 사용하는 함수들은 다양한 파라미터(parameters)들을 갖고 있고 이 파라미터들을 적절하게 지정할 수 있다. 히스토그램의 모양을 결정짓는 요소들 중 한 가지는 막대의 너비, 즉 각 막대가 나타내는 구간이다. 또한 히스토그램의 Y축을 각 구간별 자료의 개수로 표시하거나 각 구간의 확률밀도로 나타낼 수도 있다. 이러한 다양한 요소들이 hist()의 파라미터로 지정되어 있으며 이 파라미터들에 적절한 값을 설정할 수 있는 것이다. 파라미터를 지정하지 않는 경우에는 기본적으로 입력된 값으로 실행된 결과를 출력한다.

아래의 결과를 확인해보자.

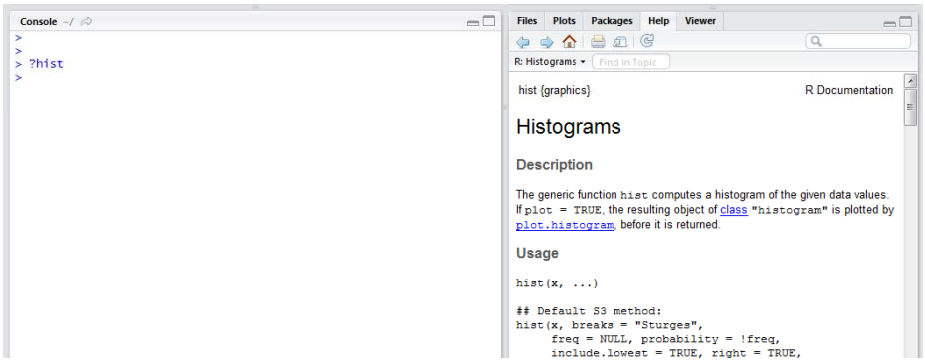
```
> hist(score, freq=F)
```




```
> hist(score, breaks=10)
```



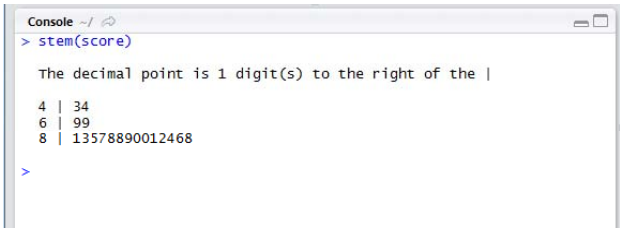
함수와 파라미터에 대한 설명은 도움말을 통해 확인할 수 있다. R에서 도움말을 보기 위해서는 help(함수명) 또는 ?함수명을 입력하면 된다.



2.1.3 줄기-잎 그림

줄기-잎 그림은 stem(x) 명령으로 그린다.

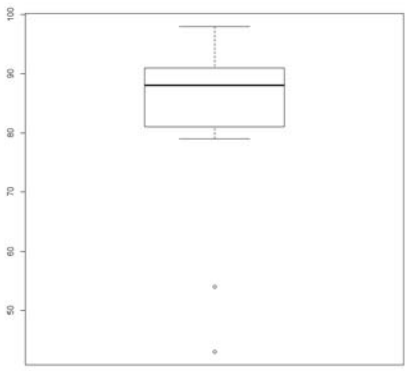
```
> stem(score)
```



2.1.4 상자그림

상자그림은 데이터의 분포를 보여주는 그림으로 가운데 상자는 제 1사분위수, 중앙값, 제 3사분위수를 보여준다. 상자그림은 boxplot(x)로 그릴 수 있다.

```
> boxplot(score)
```



2.2 일변량 자료의 요약 - 수치를 이용한 요약

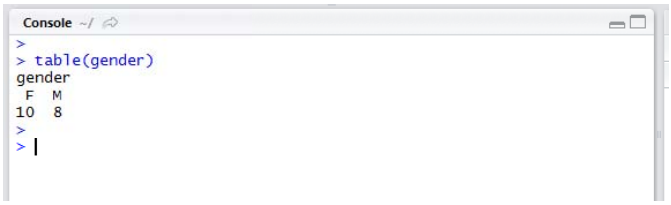
통계량 : 표본으로부터 계산되는 표본의 특성값

- 중심위치의 측도 : 평균, 중앙값
- 산포의 측도 : 분산, 표준편차, 사분위수범위

2.2.1 범주형 자료의 요약

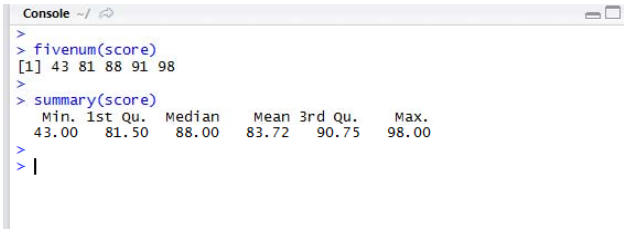
범주형 자료의 요약은 분할표를 이용할 수 있다. 분할표를 작성하는 기본 함수는 table()을 사용한다.

```
> table(gender)
```



2.2.2 숫자형 자료의 요약

다섯 수치 요약은 데이터를 '최소값, 제 1사분위수, 중앙값, 제 3사분위수, 최대값'으로 요약한다. 다섯 수치 요약은 fivenum(x)를 이용할 수 있고 summary(x)는 다섯 수치 요약에 더해 평균까지 계산해준다.



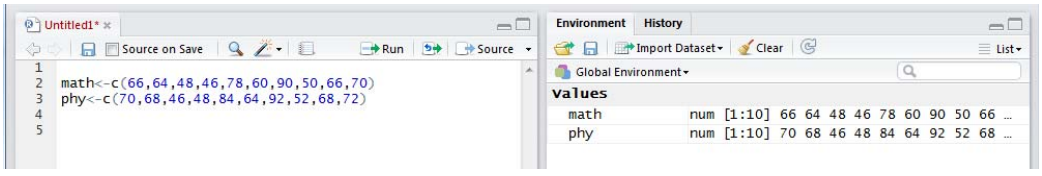
또한 다음과 같은 다양한 함수를 이용하여 통계량을 계산할 수 있다.

함수	기능
mean(x)	평균
sd(x)	표준편차
var(x)	분산
median(x)	중위수
quantile(x)	사분위값
sum(x)	합
min(x) / max(x)	최소 / 최대값

2.3 이변량 자료의 요약

다음은 어느 고등학교에서 랜덤하게 추출된 10명의 수학, 물리 성적이다.

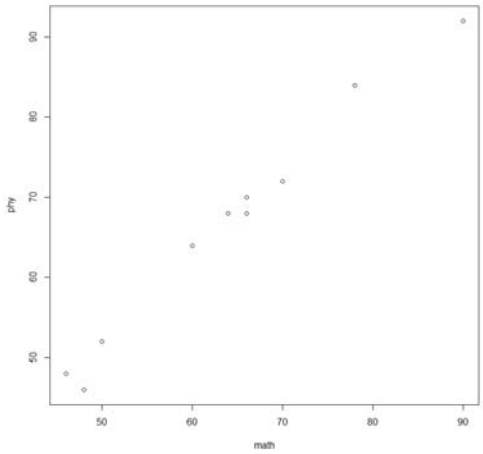
수학	66	64	48	46	78	60	90	50	66	70
물리	70	68	46	48	84	64	92	52	68	72



2.3.1 그래프를 이용한 요약

그래프를 이용한 이변량 자료의 요약은 산점도(scatter plot)를 이용할 수 있다. 산점도를 그리기 위해서는 plot() 함수를 사용할 수 있다.

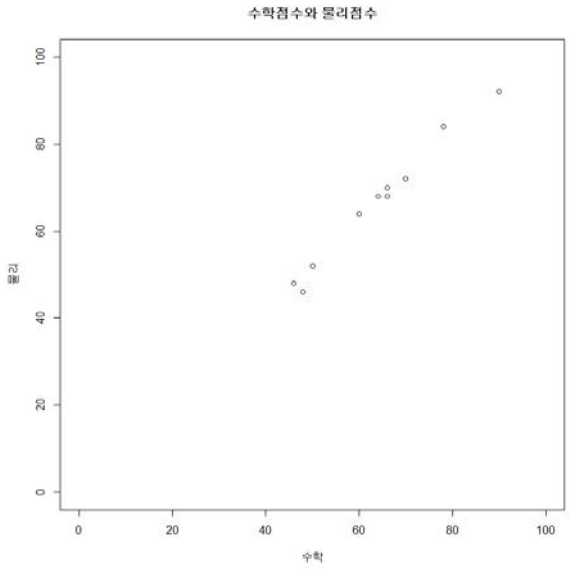
```
> plot(math, phy)
```



plot() 함수에는 다양한 하위레벨 선택사항들이 있는데 이를 사용하여 다양한 형태의 그래프를 그릴 수 있다. 다음은 몇 가지 주요 선택사항들에 대한 설명이다.

plot(x, y, main=‘ ’, sub=‘ ’, xlim=c(a,b), ylab=‘ ’, type=‘ ’)	
x	x 축 변수
y	y 축 변수
main	그래프 제목
sub	그래프 아래쪽의 제목
xlim / ylim	x축(y축) 좌표의 범위 지정
xlab / ylab	x축(y축) 제목 지정
type	p : 관측치를 점으로 표현 (기본값) l : 관측치를 선으로 이어서 표현 c : 관측치를 점선으로 그려서 표현 n : 관측치를 나타내지 않음

```
> plot(math,phy,main="수학점수와 물리점수", xlim=c(0,100), ylim=c(0,100),
        xlab="수학", ylab="물리")
```



2.3.2 상관계수를 이용한 요약

두 변수의 상관계수를 구하기 위해서는 cor() 함수를 사용할 수 있다.

```
Console ~/
>
> cor(math,phy)
[1] 0.9918056
>
>
```

2.4 자료를 이용한 예제 (cdc.txt)

행동위험요인 감시시스템(The Behavioral Risk Factor Surveillance System)은 매년 미국에서 시행되는 대규모 전화 설문 조사이다. 이 조사에서는 응답자들의 현재 건강 상태 및 그들의 건강과 관련된 생활 습관 등을 조사한다. 이 조사에 관한 자세한 내용은 BRFSS 의 웹사이트에서 확인할 수 있다. (http://www.cdc.gov/brfss)

주어진 자료는 2000년도에 시행된 20,000명의 BRFSS 조사 데이터의 일부이며 전체 200개 이상의 항목 중에서 간추린 9개의 항목을 포함하고 있다. 각 변수에 대한 설명은 다음과 같다.

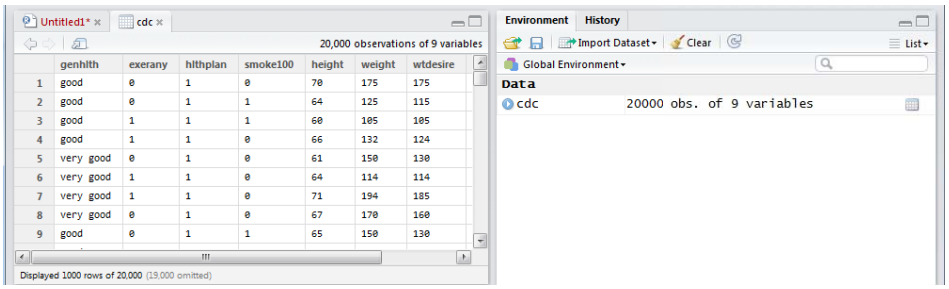
- genhlth : 범주형 자료, 전반적인 건강상태 (excellent/very good/good/fair/poor)
- exerany : 범주형 자료, 지난달의 운동 여부 (1=yes, 0=no)
- hlthplan : 범주형 자료, 건강보험 가입 여부 (1=yes, 0=no)
- smoke100 : 범주형 자료, 현재까지 최소 100개피 이상의 담배 흡연 여부 (1=yes, 0=no)
- height : 숫자형 자료, 신장 (inch)
- weight : 숫자형 자료, 체중 (pound)

- wtdesire : 숫자형 자료, 응답자가 생각하는 본인의 이상적인 체중 (pound)
- age : 숫자형 자료, 나이 (year)
- gender : 범주형 자료, 성별 (m=남성, f=여성)

주어진 자료는 텍스트 형태의 자료이다. 텍스트 자료를 불러오기 위해서는 read.table()의 명령어를 사용한다. 주로 자료의 첫 번째 행은 변수명을 나타내는 경우가 많은데, 이처럼 첫 번째 행을 관측치가 아닌 변수명으로 인식하도록 하기 위해서는 header=T 라는 파라미터를 사용한다.

```
> read.table(" 파일 저장 경로 ", header=T)
```

다음은 cdc.txt 파일을 불러와서 cdc 라는 변수명으로 저장한 결과이다. workspace를 확인하면 cdc 라는 이름의 자료가 생성된 것을 확인할 수 있고 dataview 창을 통해 내용을 확인할 수도 있다.



이렇게 불러온 자료는 데이터 프레임의 형태로 저장되고 자료의 각 열(column)은 각각의 변수를 나타낸다. 데이터 프레임의 각 열을 사용하기 위해서는 $변수명을 사용한다. 예를 들어 cdc 데이터의 첫 번째 열인 genhlth 변수를 사용하기 위해서는 cdc$genhlth를 입력한다.

예제 1. genhlth 변수에 대해 적절한 방법을 이용하여 요약해보자. 범주형 자료의 경우에는 어떠한 요약 방법을 사용할 수 있는가?
예제 2. weight 변수에 대한 수치적 요약 값을 구해보자. 전체 응답자의 평균 몸무게는 얼마인가?
예제 3. weight 변수와 wtdesire 변수의 산점도를 그려보자. 두 변수 사이에는 어떠한 관계가 존재한다고 보여지는가? 두 변수의 상관계수는 무엇은 나타내고 있는가?
예제 4. wtdesire 변수와 weight 변수의 차를 계산하여 새로운 변수 wdiff 를 만들어보자. wdiff 의 분포는 어떠한가? 수치적 요약과 그래프 요약을 통해 살펴보자. 이것이 의미하는 바는 무엇인가?
예제 5. age 변수를 이용하여 히스토그램을 그려보자. 그리고 구간의 수를 50, 100으로 바꿔가며 동일한 히스토그램을 그린 후 비교해보자.
(참고) 히스토그램은 자료의 형태를 파악하기 위한 쉬운 방법이지만 구간의 수가 달라짐에 따라 그 모양이 조금씩 달라질 수 있다.