통계학 실험 과제 #1

2017-17497 김현규

예제 1.

> cdc<-read.table("C:\\cdc.txt", header=T)

> table(cdc$genhlth)

excellent fair good poor very good

4657 2019 5675 677 6972

> summary(cdc$genhlth)

excellent fair good poor very good

4657 2019 5675 677 6972

범주형 자료이므로 분할표를 작성해야 하며 R에서 table함수를 통해 얻을 수 있다. 작성 결과 위와 같은 결과를 얻을 수 있었으며 summary를 이용하였을 경우에도 같은 결과를 얻었음을 알 수 있다.

예제2.

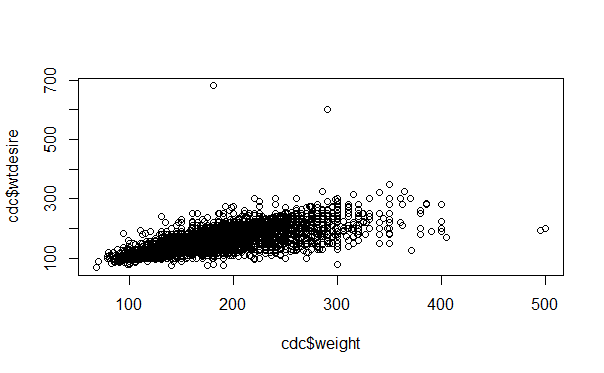
> summary(cdc$weight)

Min. 1st Qu. Median Mean 3rd Qu. Max.

68.0 140.0 165.0 169.7 190.0 500.0

Weight 변수는 수치로 나타나는 변수로 평균이나 중앙값, 사분위수와 같은 값들을 알아낼 수 있고 이 값은 summary 함수를 통해 얻을 수 있다.

예제3.

> plot(cdc$weight, cdc$wtdesire)

> cor(cdc$weight, cdc$wtdesire)

[1] 0.8000521

Weight 변수와 wtdesire 변수의 산점도는 위와 같다. 산점도에서 볼 수 있듯이 weight와 weightdesire에는 양의 상관관계가 있다고 보여지고, 두 변수의 상관계수를 cor함수를 통해 구하면 0.8정도로 두 변수가 강한 양의 상관관계가 있음을 알 수 있다.

예제4.

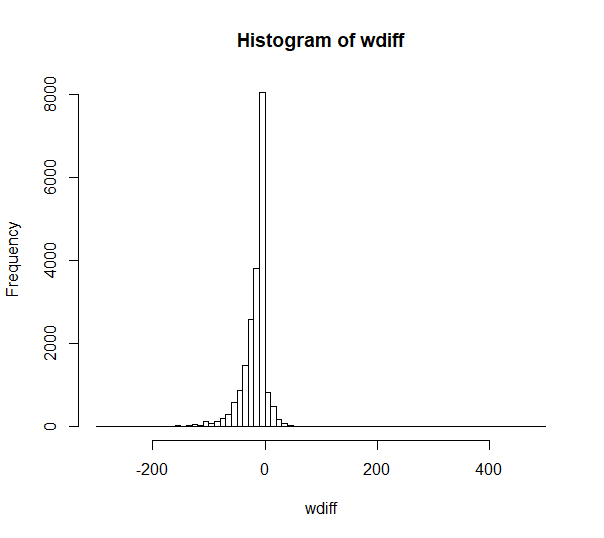
> wdiff = cdc$wtdesire - cdc$weight

> summary(wdiff)

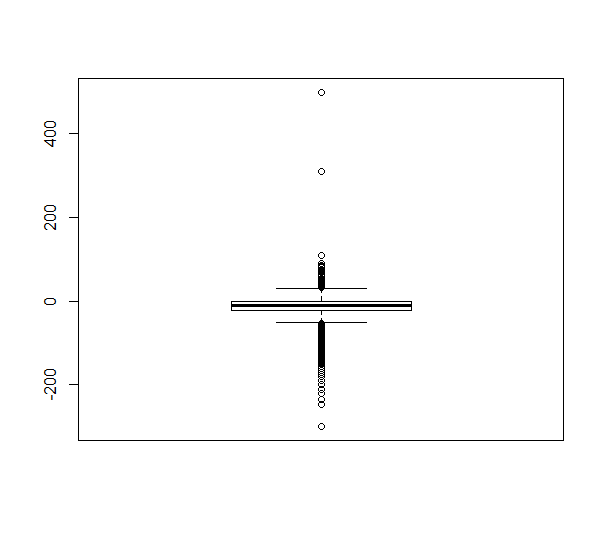
Min. 1st Qu. Median Mean 3rd Qu. Max.

-300.00 -21.00 -10.00 -14.59 0.00 500.00

> hist(wdiff, breaks=100)



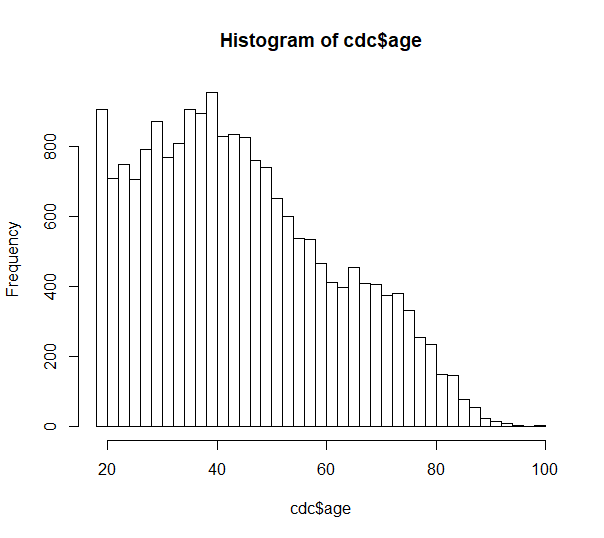
> boxplot(wdiff)



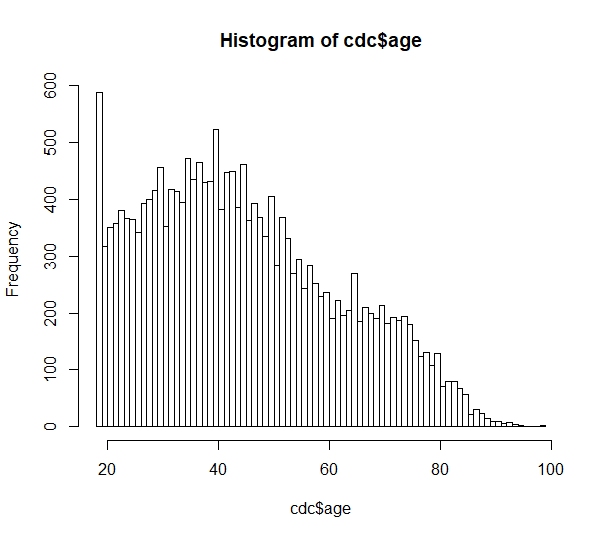
wdiff변수를 만들었다. 이 변수는 수치적인 변수이므로 summary 함수를 통해 중앙값, 평균값, 수분위수 등을 구할 수 있다. 최소값과 최대값은 각각 -300과 500이지만 산점도에서 볼 수 있듯이 큰 의미를 가지지 않는 아웃라이어이다. 중앙값과 평균값이 각각 -10.0, -14.59로 일반적으로 사람들은 자신의 몸무게보다 더 가벼운 몸무게를 원한다는 사실을 알 수 있다. 히스토그램과 상자그림에서도 이 사실을 알 수 있다.

예제5.

> hist(cdc$age, breaks=50)



> hist(cdc$age, breaks=100)



나이를 히스토그램으로 나타낸 결과는 위와 같다. breaks인자에 50과 100을 전달하여 구간의 수를 조절한 것을 알 수 있다. 두 그래프의 개형은 유사하게 나타남을 알 수 있다.