통계 분석

2019년 2학기

강봉주

데이터 기술

데이터 기술

[변수 유형]

변수는 모집단 특성에 대한 하나의 측정값의 집합

구분	내용	예	비고
명목척도 (nominal scale)	이름으로 객체를 구분할 때 즉, 주로 분류를 위하여 사용	성별, 국적, 주민번호, 인종, 언어, 장르, 품사 등	질적(qualitative) 척도. 수학연산(=,≠) 적용 가능. 최빈값(mode) 계산가능
순서척도 (ordinal scale)	순위(rank)를 갖고 있으며 정렬(sorting) 가능하나 객체간의 차이의 정도를 표현하지 않음	성적(A, ..., F), 설문조사 응답(완전동의, 대부분동의, ..., 완전비동의) 등	최빈값, 중간값(median) 계산가능. 수학연산(>, <) 적용가능.
구간척도 (interval scale)	객체 간의 차이의 정도를 표현하지만, 비율은 의미가 없음	기온(섭씨), 성적(0-100), 날짜, 위치(좌표) 등	최빈값, 중간값, 평균, 범위, 표준편차 적용가능. 수학연산(+,-) 적용가능
비율척도 (ratio scale)	객체 간의 차이의 정도도 표현하며 비율 또한 의미가 있음. 의미 있는 유일한 0 값을 가지고 있음.	질량, 길이, 경과기간, 각도, 충전량 등	기하평균 등 적용 가능. 수학연산(*, /) 적용 가능

데이터 기술

[변수 유형]

실제는 연속형(continuous, interval)과 범주형(categorical, discrete)으로 주로 구분

데이터 기술

[BANK 데이터]

추가 판매를 위한 대상 고객을 추출하기 위한 데이터

#	변수	유형	설명	비고	한글 설명
1	age	numeric			나이
2	job	categorical	type of job	"admin.","unknown","unemployed","management","housemaid","entrepreneur","student","blue-collar","self-employed","retired","technician","services")	직업
3	marital	categorical	marital status	"married", "divorced", "single"	결혼상태
4	education	categorical		"unknown", "secondary", "primary", "tertiary")	교육수준
5	default	binary	has credit in default ?		채무불이행유무

데이터 기술

[BANK 데이터]

추가 판매를 위한 대상 고객을 추출하기 위한 데이터

6	balance	numeric	average yearly balance, in euros		년잔액
7	housing	binary	has housing loan?		주택대출유무
8	loan	binary	has personal loan?		개인대출유무
9	contact	categorical	contact communication type	"unknown", "telephone", "cellular"	접촉채널
10	day	numeric	last contact day of the month		최근접촉일
11	month	categorical	last contact month of year	"jan", "feb", "mar", ..., "nov", "dec"	최근접촉월
12	duration	numeric	last contact duration, in seconds		최근접촉시간(초)

데이터 기술

[BANK 데이터]

추가 판매를 위한 대상 고객을 추출하기 위한 데이터

13	campaign	numeric	number of contacts performed during this campaign and for this client		캠페인기간내 접촉건수
14	pdays	numeric	number of days that passed by after the client was last contacted from a previous campaign		최근캠페인후 경과일수
15	previous	numeric	number of contacts performed before this campaign and for this client		과거캠페인접촉건수
16	poutcome	categorical	outcome of the previous marketing campaign	"unknown", "other", "failure", "success"	최근캠페인결과
17	y	binary	has the client subscribed a term deposit?		정기예금가입여부

데이터 기술

[기술 통계량]

모집단 또는 표본의 특성을 하나의 숫자로 요약

```
# install.packages('Hmisc'): Harrell Miscellaneous
library(Hmisc)
describe(df)
```

데이터 기술

[기술 통계량]

모집단 또는 표본의 특성을 하나의 숫자로 요약

Gini's Mean Difference

$$\frac{\sum\sum|x_i - x_j|}{n(n-1)}$$

```
---------------------------------------------------------------------------
balance
       n  missing distinct     Info     Mean      Gmd      .05
    4521        0     2353        1     1423     2150     -162
     .10      .25      .50      .75      .90      .95
       0       69      444     1480     3913     6102

lowest : -3313 -2082 -1746 -1680 -1400, highest: 27069 27359 27733 42045 71188
---------------------------------------------------------------------------
housing
       n  missing distinct
    4521        0        2

Value         no   yes
Frequency   1962  2559
Proportion 0.434 0.566
---------------------------------------------------------------------------
```

데이터 기술

[기술 통계량]

[범주형 데이터의 요약]

```
> table(df$marital)

divorced  married   single
     528     2797     1196
```

일원 빈도표
(one-way frequency table)

데이터 기술

[기술 통계량]

[범주형 데이터의 요약]

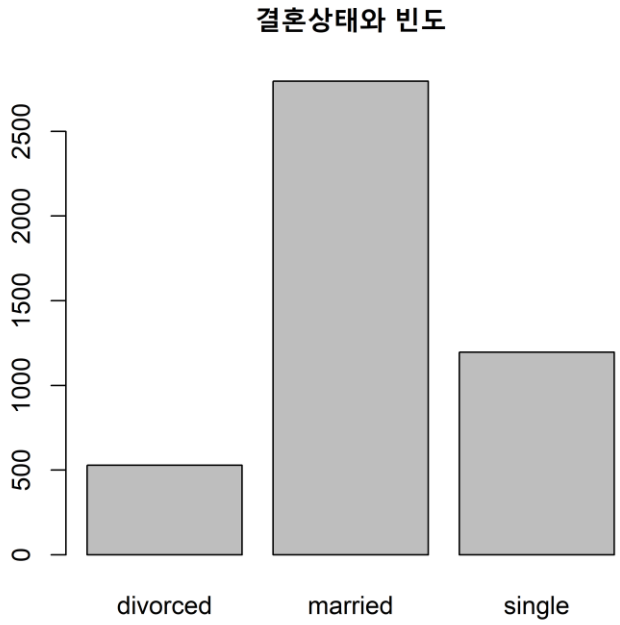
```
> print(table(df$marital)/sum(table(df$marital)), digits=4)

divorced  married   single
  0.1168   0.6187   0.2645
```

데이터 기술

[기술 통계량]

[범주형 데이터의 요약]



데이터 기술

[기술 통계량]

[범주형 데이터의 요약]

빈도표에 대한 막대 도표와 파이 도표를 생성하세요.

데이터 기술

[기술 통계량]

[범주형 데이터의 요약]

1) “y”, “marital” 변수에 대한 2원 빈도표를 생성하세요.
2) 행 합에 대한 각 셀의 비율을 계산하세요.

데이터 기술

[기술 통계량]

[범주형 데이터의 요약]

‘y’ 변수의 각각의 값에 대한 ‘pdays’의 평균 값을 구하세요.

```
     y    pdays
1   no 36.00600
2  yes 68.63916
```

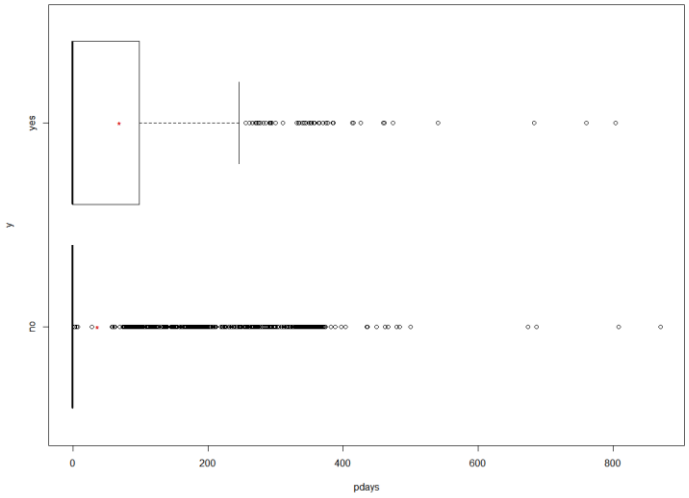
데이터 기술

[기술 통계량]

[범주형 데이터의 요약]

1) ‘y’ 변수의 각각의 값에 대한 ‘pdays’의 평균 값을 구하세요.
2) ‘y’에 값에 따른 상자 그림을 그려보세요.

```
     y    pdays
1   no 36.00600
2  yes 68.63916
```



데이터 기술

[기술 통계량]

[연속형 데이터의 요약]

```
# 모든 값에 따른 건수
table(df$age)

# 요약 함수
summary(df$age)

# 줄기-잎 그림
stem(df$age)
```

데이터 기술

[기술 통계량]

[연속형 데이터의 요약]

```
> summary(df$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   33.00   39.00   41.17   49.00   87.00
```

k 사분위수: $\Pr(X < x_k) \leq \frac{k}{4}$

$\Pr(X < 33) \leq \frac{1}{4}$

```
> nrow(subset(df, age <33 ))/nrow(df)
[1] 0.2333555
> nrow(subset(df, age <34 ))/nrow(df)
[1] 0.2744968
```

데이터 기술

[기술 통계량]

[연속형 데이터의 요약]

1 백분위수와 99 백분위수를 구하세요.

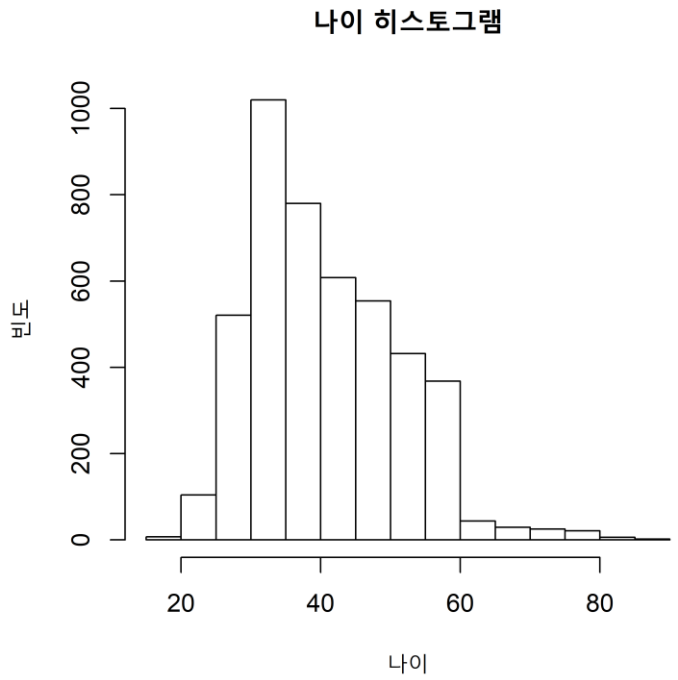
데이터 기술

[기술 통계량]

[연속형 데이터의 요약]

[히스토그램]

```
hist(df$age, main='나이
히스토그램', xlab='나이',
ylab='빈도')
```



데이터 기술

[기술 통계량]

[연속형 데이터의 요약]

[히스토그램]

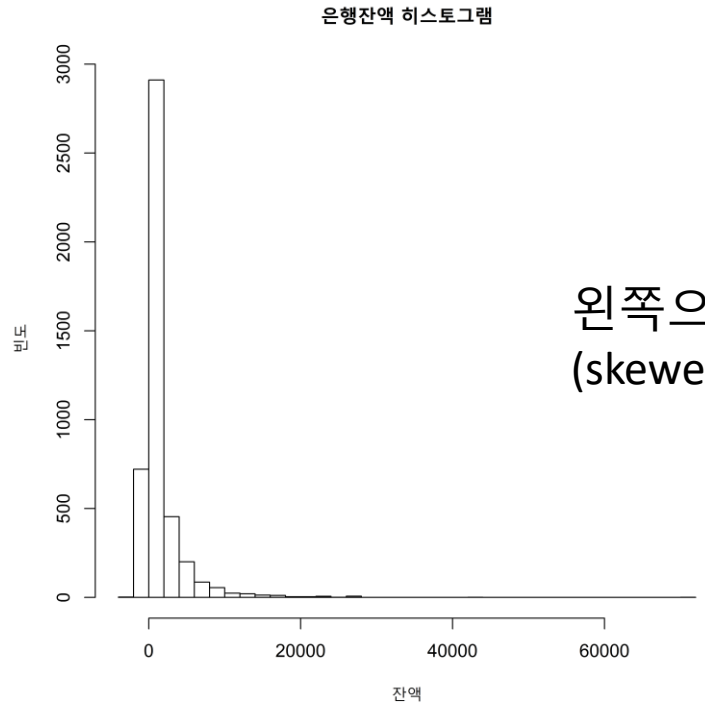
1) 셀의 수가 50개인 경우의 히스토그램을 생성하세요.
2) 셀의 수가 20개인 경우(균등 폭)의 각 셀의 경계값을 구하세요.
3) 2)의 각 경계 값으로 구간을 구성하고 각 구간별 빈도를 구하세요.

seq, cut 함수 이용

데이터 기술

[기술 통계량]

[연속형 데이터의 요약]

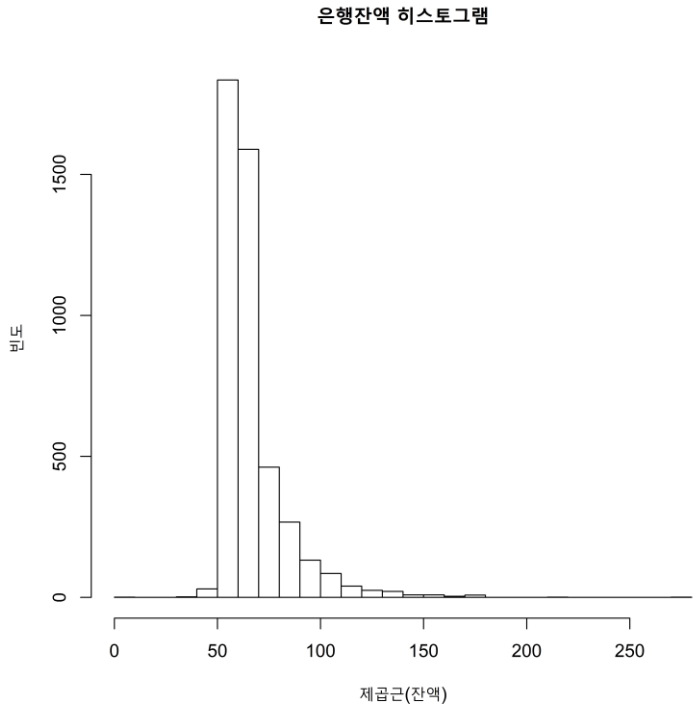


왼쪽으로 급하고 오른쪽으로 완만한 분포
(skewed to the right, right-skewed)

데이터 기술

[기술 통계량]

[연속형 데이터의 요약]

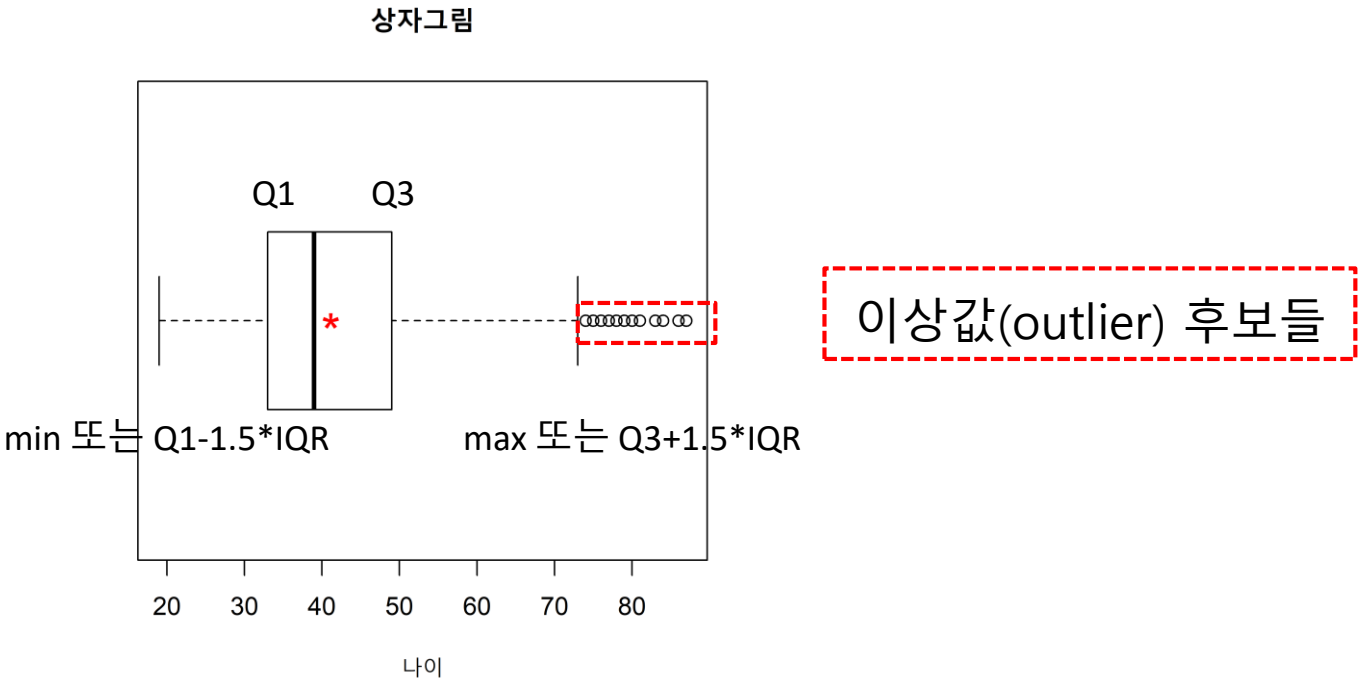


적절한 변환을 통하여 정상적인 분포로 유도

데이터 기술

[기술 통계량]

[연속형 데이터의 요약]



데이터 기술

[기술 통계량]

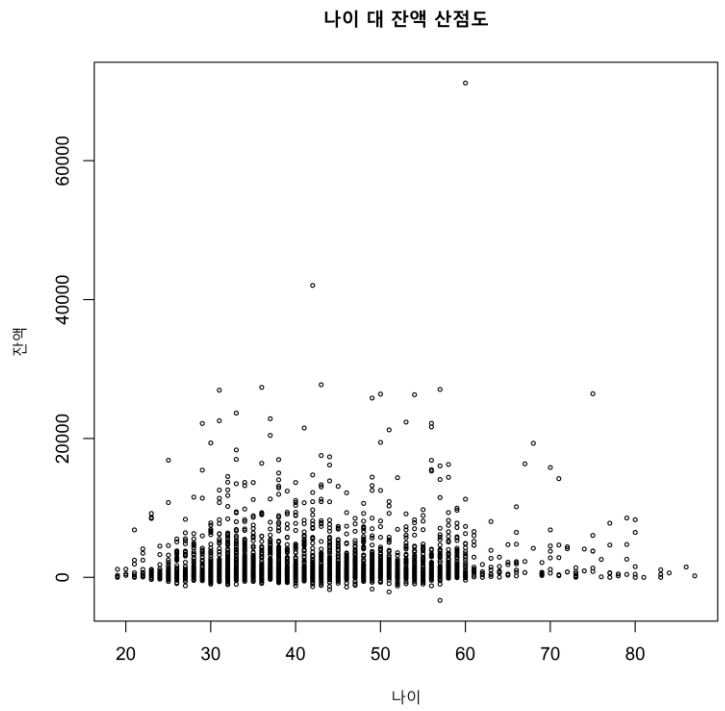
[연속형 데이터의 요약]

1) [BANK] 데이터의 모든 연속형 변수 목록을 생성하세요.
2) 모든 연속형 변수에 대한 히스토그램을 생성하세요.

데이터 기술

[기술 통계량]

[산점도: 변수 간의 관련성(주로 연속형 변수)]




```
plot(df$age, df$balance,
    cex=0.5,
    main='나이 대 잔액 산점도',
    xlab='나이',
    ylab='잔액')
```

데이터 기술

[기술 통계량]

[산점도 행렬]