R컴퓨팅

11강

R 데이터 탐색 II

정보통계학과 장영재 교수

1 기술통계 함수

2 도수분포표와 교차표

1 기술통계 함수

1 기술통계 함수

1 사분위수와 백분위수 및 분위수(quantile)

- 0과 1사이의 값인 p에 대해서 100p% 백분위수(percentile)는 자료를 크기 순으로(오름차순) 배열할 때 100p% 에 해당되는 자료이며 p분위 수는 100p% 백분위수(percentile)

- 사분위수는 자료를 순서대로 배열할 때 25%, 50% 및 75%에 해당되는 세 값을 말하며 각각 제1사분위수, 제2사분위수(또는 중앙값), 제3사분위수라고 하며 각각 Q_1,Q_2, 및 Q_3로 표시

```
quantile(x, probs = seq(0, 1, 0.25), ...)
```

1 기술통계 함수

1 사분위수와 백분위수 및 분위수(quantile)

- x: 분위수를 계산할 자료를 저장한 벡터를 설정하며 na.rm이 TRUE가 아니면 x에 NA나 NaN이 사용될 수 없음

- probs: 100p% 백분위수를 얻을 p의 벡터를 설정

1 기술통계 함수

1 사분위수와 백분위수 및 분위수(quantile)

› 보기 11-1: 1부터 100까지의 값을 갖는 100개의 자료로 분위수 계산

```
> x <- 1:100
> quantile(x)
  0%  25%  50%  75% 100%
1.00 25.75 50.50 75.25 100.00
> quantile(x, prob=c(0.05, 0.1, 0.9, 0.95))
   5%   10%   90%   95%
 5.95 10.90 90.10 95.05
```

1 기술통계 함수

1 사분위수와 백분위수 및 분위수(quantile)

› 보기 11-2: BMI 자료에서 키에 대한 5%, 10%, 25%, 50%, 75%, 90%, 95% 백분위수는

```
〉 quantile(BMI$height, prob=c(0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95))
  5%  10%  25%  50%  75%  90%  95%
153.0 155.0 158.0 162.0 165.0 168.4 172.2
```

1 기술통계 함수

1 사분위수와 백분위수 및 분위수(quantile)

[1] 산술평균과 절삭평균

- n개의 자료의 중심에 대한 척도로 사용하는 값들로 평균이나 절삭평균 (또는 절사평균; trimmed mean)을 사용

- 평균은 자료 중에서 지나치게 크거나 작은 한두 개의 값에 민감하게 반응하므로 이에 대한 보정으로 절삭평균을 사용하며 R-언어에서 절삭평균 및 평균을 계산하는 함수는

```
mean(x, trim = 0, na.rm=, ...)
```

1 기술통계 함수

2 자료의 중심에 대한 측도

- x: 평균을 계산할 자료를 저장하고 있는 벡터를 설정
- trim: % 절삭평균을 계산할 때 절삭할 비율 를 설정
- na.rm: 논리값을 설정하며 자료 x에 NA가 포함되었을 때 NA를 제외

1 기술통계 함수

2 자료의 중심에 대한 측도

› 보기 11-3: 1부터 100까지의 수와 1000이 포함된 101개의 자료에 대해서

```
> x<- c(1:100, 1000)    # 자료 설정
> mean(x)
[1] 59.90099              # 평균값
> mean(x, trim=0.05)     # 5% 절삭평균
[1] 51
```

으로 전체자료의 평균값은 59.9이지만 5% 절삭평균은 51

1 기술통계 함수

2 자료의 중심에 대한 측도

› 보기 11-4: BMI자료에서 여자의 몸무게의 평균과 5% 절삭평균 구하기

› 몸무게인 자료는 2번 열이므로

```
〉BMI[BMI[,5] == "F", 2]
```

여자 몸무게의 평균은

```
〉 mean(BMI[BMI[,5]== "F", 2])
[1] 51.87342
```

같은 방법으로 5% 절삭평균은 trim 속성에 0.05를 설정하여

```
〉 mean(BMI[BMI[,5]== "F", 2], trim=0.05)
[1] 51.81944
```

를 얻음

1 기술통계 함수

2 자료의 중심에 대한 측도

[2] 중앙값

- 중앙값(median)은 중위수라고도 부르며 자료를 크기 순서대로 오름차순으로 배열할 때 중앙에 위치한 값
- 자료의 수 n이 홀수이면 순서대로 배열할 때 중앙에 위치하는 하나의 값이 중앙값이며 n이 짝수이면 중앙에 위치하는 값이 두 개 이므로 이 두 값의 평균
- 중앙값은 median 함수에 의해서 얻을 수 있으며 사용법은

```
median(x, na.rm = FALSE)
```

1 기술통계 함수

2 자료의 중심에 대한 측도

- 보기 11-5: 1부터 100까지의 수와 1000이 포함된 101개의 자료에 대해서

```
> x<- c(1:100, 1000)    # 자료 설정
> median(x)
[1] 51
> y <- c(x, NA)         # 위에 x에서 마지막 값이 NA
> median(y)
[1] NA
> median(y, na.rm=T)  # 결측값 제외한 자료만으로 중앙값 계산
[1] 51
```

1 기술통계 함수

2 자료의 중심에 대한 측도

› 보기 11-6: 백분위수를 구하는 방법으로 중앙값 구하기

BMI 자료에서 키의 중앙값은

```
> median(BMI$h)
[1] 162
```

이며 중앙값은 제 50% 백분위수이기도 하므로

```
> quantile(BMI$h, probs=0.5)
50%
162
```

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

- 자료의 흩어짐에 대한 측도로서, 분산, 표준편차, 범위, 사분위수 범위 등이 있음

[1] 분산과 표준편차

- 표준편차는 분산의 양의 제곱근으로 보통 s로 표현한다. 분산과 표준편차는 각각 var과 sd 함수로 얻을 수 있음

```
sd(x, na.rm = FALSE)
var(x, y = NULL, na.rm = FALSE)
```

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

- x: 표준편차 또는 분산을 계산할 수치자료가 저장된 행렬 또는 벡터를 설정
- y: y는 기본값으로 설정하지 않는 것이나 행렬, 벡터 또는 data frame
- na.rm: 결측값을 제외할지 설정하는 논리값

- 보기 11-7: BMI 자료에서 키의 분산, 표준편차를 계산

```
> var(BMI$height)
[1] 29.95095
> sd(BMI$height)
[1] 5.472746
```

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

[2] 범위 및 최대 최소

› 범위 및 최대, 최소는 다음의 함수로 얻을 수 있음

```
range(..., na.rm = FALSE)
max(..., na.rm = FALSE)
min(..., na.rm = FALSE)
```

› range(x)[1]은 x의 가장 작은 값, range(x)[2]는 x의 가장 큰 값으로 통계학에서 말하는 범위는 range(x)[2] - range(x)[1]

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

› 보기 11-8: BMI자료에서 키의 최대, 최소 및 범위

```
〉 range(BMI[,1])     # 키의 최대/최소
[1] 150 180
```

› 남자인 경우 키의 최대, 최소는 성별이 다섯 번째 열이고, 키가 첫 번째 열

```
〉 maxmin 〈– range(BMI[BMI[,5] == "M",1])  # 키의 최대/최소
〉 maxmin
[1] 160 180
〉 rng 〈– maxmin[2] – maxmin[1]
〉 rng
[1] 20
```

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

[3] 사분위수 범위

- 사분위수범위(InterQuatile Range; IQR)는 제3사분위수 Q_3와 제1사분위수 Q_1의 차이로 R-언어에서는

```
IQR(x, na.rm = FALSE)
```

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

› 보기 11-9: 키, 몸무게 자료 BMI에서 몸무게의 사분위수 범위는

```
> IQR(BMI$weight)
[1] 8
```

이고 quantile 함수로 사분위수를 계산해 보면

```
> quantile(BMI$weight)
  0%  25%  50%  75% 100%
  40   49   52   57   80
```

이므로 임

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

[4] 상관계수와 공분산

- 자료가 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$로 n개의 짝으로 얻어진 경우 x가 증가(감소)할 때 y가 증가(감소)하면 공분산은 양의 값을,

- x가 증가(감소)할 때 y가 감소(증가)하면 공분산은 음의 값을 가짐

- 피어슨(Pearson)의 선형상관계수를 말하며 표본에서 얻은 상관계수는 보통 r로 표시

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

› 상관계수 및 공분산은 각각 cor과 cov함수를 사용하여 얻을 수 있으며, 이들 함수는 다음과 같이 사용

```
cov(x, y = NULL)
cor(x, y = NULL)
```

› x: 상관계수 또는 공분산을 계산할 벡터, 행렬, 데이터 프레임 등을 설정

› y: x가 벡터일 경우 y가 반드시 있어야 함

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

› 보기 11-10: BMI 자료에서 키와 몸무게의 상관계수 및 공분산을 계산해 보고, 상관계수를 공식을 적용하여 얻어 보자.

› 키와 몸무게의 공분산은

```
〉 cov(BMI$height, BMI$weight)
[1] 25.42062
```

이며 상관계수는

```
〉 cor(BMI$height, BMI$weight)
[1] 0.6473976
```

1 기술통계 함수

3 자료의 흩어짐에 대한 측도

› 보기 11-11: 표준정규분포로부터 100개씩의 난수를 세 번 얻어 이를 행렬에 저장하고 이의 상관계수를 계산하기

```
〉 x 〈– cbind(rnorm(100), rnorm(100), rnorm(100))
〉 # 100X3 행렬로 세 개의 열을 가진 자료를 x로 얻고
〉 cor(x)
             [,1]        [,2]        [,3]
[1,]  1.00000000  0.04526954 –0.01003914
[2,]  0.04526954  1.00000000 –0.08316823
[3,] –0.01003914 –0.08316823  1.00000000
```

2 도수분포표와 교차표

2 도수분포표와 교차표

1 table 함수

- R-언어의 table 함수는 매개변수가 한 개이면 도수분포표를, 둘 이상이면 교차표를 출력하는 함수로 사용법은

```
table(..., exclude = dnn = list.names(...), ...)
```

- ... : 한 개 이상의 객체를 설정할 수 있으며 한 개의 매개변수가 주어진 경우 도수분포표, 두 개의 이상의 객체가 주어지면 교차표가 얻어짐
- exclude : 교차표에서 제외할 수준의 값을 설정
- dnn : 교차표의 행변수 열변수 등에 사용할 이름

2 도수분포표와 교차표

1 table 함수

보기 11-12: BMI 자료에서 성별에 따른 도수분포표는

```
> table(BMI$gender) # 성별에 따른 도수분포표
F M
158 19
```

보기 11-13: 두 개의 변수가 설정된 경우 교차표가 출력되며 첫 번째 변수는 행, 두 번째 변수는 열이 됨

```
> table(BMI$gender, BMI$marr) # 행과 열에 주어진 변수
N Y
F 126 32
M 7 12
```

2 도수분포표와 교차표

1 table 함수

- 만일 세 개의 변수가 설정되면 첫 번째 변수는 행, 두 번째 변수는 열인 교차표를 세 번째 변수의 각각의 값에 대해서 출력

```
> table(BMI$gender, BMI$religion, BMI$marr)
, , = N
Bu C1 C2 No
F 17 34 23 52
M 1 2 0 4
, , = Y
Bu C1 C2 No
F 1 11 10 10
M 2 0 8 2
```

2 도수분포표와 교차표

1 table 함수

› 보기 11-14: 특정한 값을 제외한 교차표를 얻을 수 있음

```
〉 # 값이 “No"인 경우 제외한 교차표
〉 table(BMI$gender, BMI$religion, exclude="No", dnn=c("성별", "종교"))
종교
성별 Bu C1 C2
F 18 45 33
M 3 2 8
```

2 도수분포표와 교차표

2 ftable 함수

- 3차원 이상인 경우 ftable 함수가 일반적으로 더 유용.

- ftable 함수는

```
ftable(..., exclude = c(NA, NaN), row.vars = NULL, col.vars
= NULL)
```

또는

```
ftable(x, ...)
```

2 도수분포표와 교차표

2 ftable 함수

- x 또는 ...: 교차표를 생성할 수 있는 R의 개체를 설정하거나 형식을 설정
- exclude: 각 변수에서 제외할 값을 설정
- row.vars: 행에 사용할 변수의 번호 또는 변수의 이름을 설정
- col.vars: 열에 사용할 변수의 번호 또는 변수의 이름을 설정

2 도수분포표와 교차표

2 ftable 함수

보기 11-15: BMI 자료에서 종교와 성별의 교차표

```
> ftable (religion~gender, data=BMI)
        religion Bu C1 C2 No
gender
F                18 45 33 62
M                 3  2  8  6
```

보기 11-16: BMI 자료에서 5번째 열인 성별이 행에 사용되도록 설정

```
> ftable(BMI[,4:6], row.vars=2)
```

2 도수분포표와 교차표

2 ftable 함수

```
            religion      Bu     C1     C2     No
            marriage   N  Y  N  Y  N  Y  N  Y
gender
F                     17  1 34 11 23 10 52 10
M                      1  2  2  0  0  8  4  2
```

row.vars에 열의 이름을 설정하여도 같은 결과를 얻음

```
> ftable(BMI[,4:6], row.vars="marr")
            religion      Bu     C1     C2     No
            gender     F  M  F  M  F  M  F  M
marriage
N                     17  1 34  2 23  0 52  4
Y                      1  2 11  0 10  8 10  2
```

2 도수분포표와 교차표

3 prop.table 함수

- 도수분포표 또는 교차표의 상대도수값을 얻고자 할 때는 prop.table 함수 사용

```
prop.table(x, margin = NULL)
```

- x: table 또는 ftable등의 결과로 얻은 도수분포표 또는 교차표
- margin: 상대도수를 계산할 때 분모로 사용할 변수 지정

2 도수분포표와 교차표

3 prop.table 함수

› 보기 11-17: 전체 비율에 대한 상대도수 구하기

```
> prop.table(table(BMI$gender, BMI$religion))
                Bu         C1         C2         No
  F 0.10169492 0.25423729 0.18644068 0.35028249
  M 0.01694915 0.01129944 0.04519774 0.03389831
```

행에 대한 상대도수(행별로 합하면 각 행에 대하여 모두 1)는 다음과 같이 margin에 1을 설정

```
> prop.table(table(BMI$gender, BMI$religion), margin=1)
                Bu         C1         C2         No
  F 0.1139241 0.2848101 0.2088608 0.3924051
  M 0.1578947 0.1052632 0.4210526 0.3157895
```

2 도수분포표와 교차표

4 addmargins 함수와 table, ftable, prop.table 함수

- table, prop.table 및 ftable 함수는 행과 열의 합을 계산하지 않는다. 이 합을 계산하기 위한 함수로 addmargins 함수가 있으며 이 함수는

```
addmargins(A, margin = seq_along(dim(A)), ...)
```

로 사용

- A: ftable, table 또는 prop.table 함수의 결과인 객체 또는 array를 설정하고A에 dim이나 dimnames가 설정되어 있는 경우 addmargins 함수는 이를 사용

2 도수분포표와 교차표

4 addmargins 함수와 table, ftable, prop.table 함수

- 보기 11-18: BMI에서 gender와 religion의 교차표에서 합을 추가하기 위해 addmargins 함수를 다음과 같이 적용

```
> addmargins(table (BMI$gender, BMI$religion))
      Bu  C1  C2  No Sum
F     18  45  33  62 158
M      3   2   8   6  19
Sum   21  47  41  68 177
```

2 도수분포표와 교차표

4 addmargins 함수와 table, ftable, prop.table 함수

› addmargins 함수는 prop.table의 출력에도 적용할 수 있음

```
> addmargins(prop.table(table(BMI$gender, BMI$religion)))

            Bu         C1         C2         No        Sum
F   0.10169492 0.25423729 0.18644068 0.35028249 0.89265537
M   0.01694915 0.01129944 0.04519774 0.03389831 0.10734463
Sum 0.11864407 0.26553672 0.23163842 0.38418079 1.00000000
```

R컴퓨팅