Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

01

데이터 탐색과 통계분석

경북대학교 배준현 교수
(joonion@knu.ac.kr)

데이터에 대한 두 가지 접근법: EDA *.vs.* CDA

- **탐색적** 데이터 분석: EDA, *exploratory* data analysis
  - 정해진 가설과 모형없이 데이터의 구조와 특성을 통해 통찰을 얻는 분석 기법
  - *John Tukey*: EDA는 우리가 존재한다고 믿는 것들은 물론이고, 존재하지 않는다고 믿는 것들을 발견하려는 태도, 유연성, 그리고 자발성이다.
- **확증적** 데이터 분석: CDA, *confirmatory* data analysis
  - 가설을 수립하고 데이터를 통해 통계적 유의성을 검정하는 전통적 분석 기법
  - *Ronald Fisher*: 가설검정, 신뢰구간, 유의확률, 유의수준(*p-value*)

- **통계분석**: *statistical analysis*
  - **기술적 통계**: *descriptive* statistics
    - 수집한 데이터의 특성을 수치로 요약하거나 시각적으로 표현하는 통계분석 방법
    - 평균, 표준편차, 교차표, 히스토그램, 막대그래프 등
  - **추론적 통계**: *inferential* statistics
    - 수집한 표본집단으로부터 모집단의 특성을 추정하기 위한 통계분석 방법
    - 가설검정, 평균검정, 분산분석, 카이제곱검정, 회귀분석 등

- **데이터**: *datum* or *data*
  - 데이터: 관찰, 측정, 실험, 또는 조사를 통해 얻는 실체적 사실이나 정보
  - 변수(변량): *variable* or *variate*
    - 변수: 관찰, 측정, 실험, 또는 조사의 대상이 되는 수량
    - 관측값(*observations*): 변수(변량)에 대한 관측을 통해 얻는 값
  - 변수의 종류:
    - 연속형(*continuous*): 수치로 표현할 수 있는 변량. 예) 키, 몸무게
    - 범주형(*categorical*): 범주로 표현할 수 있는 변량. 예) 성별, 혈액형
  - 변수의 구분:
    - 독립변수(*independent*): 종속변수에 영향을 주는 변수. 예) 부모의 키
    - 종속변수(*dependent*): 독립변수로부터 영향을 받는 변수. 예) 자녀의 키

▪ 혼돈의 카오스: 비슷하고 헷갈리는 용어들

수치형 자료 *numeric* data

범주형 자료 *categorical* data

양적 자료 *quantitative* data

질적 자료 *qualitative* data

연속형 자료 *continuous* data

이산형 자료 *discrete* data

명목형 자료 *nominal* data

순서형 자료 *ordinal* data

독립변수 *independent* variable

종속변수 *dependent* variable

특징변수 *feature* variable

목적변수 *target* variable

설명변수 *explanatory* variable

반응변수 *response* variable

예측변수 *predictor* variable

결과변수 *outcome* variable

■ 데이터의 수집: *data collection*

- 연구주제에 관련된 데이터를 관찰하고 정리: 설문응답 or 실험결과
- 데이터셋: 주로 2차원 테이블(데이터프레임) 형태로 정리된 데이터

설 문 지

1. 당신의 성별은? ( ) ① 남성 ② 여성

2. 글을 쓰실 때 어느 쪽 손을 주로 쓰십니까? ( ) ① 오른손 ② 왼손

3. 오른쪽 손을 폈을 때, 엄지 끝에서 새끼 손가락 끝까지의 길이를 써주세요. ( ) 단위: cm

4. 왼쪽 손을 폈을 때, 엄지 끝에서 새끼 손가락 끝까지의 길이를 써주세요. ( ) 단위: cm

5. 당신의 키는 얼마입니까? 단위(cm/m, feet/inch)와 함께 적어주세요. ( )

…… (이하 생략)

- MASS 패키지: survey 데이터셋

```
> library(MASS)
> str(survey)
'data.frame':  237 obs. of  12 variables:
 $ Sex   : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
 $ Wr.Hnd: num  18.5 19.5 18 18.8 20 18 17.7 17 20 18.5 ...
 $ NW.Hnd: num  18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5 18.5 ...
 $ W.Hnd : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
 $ Fold  : Factor w/ 3 levels "L on R","Neither",..: 3 3 1 3 2 1 1 3 3 3 ...
 $ Pulse : int  92 104 87 NA 35 64 83 74 72 90 ...
 $ Clap  : Factor w/ 3 levels "Left","Neither",..: 1 1 2 2 3 3 3 3 3 3 ...
 $ Exer  : Factor w/ 3 levels "Freq","None",..: 3 2 2 2 3 3 1 1 3 3 ...
 $ Smoke : Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
 $ Height: num  173 178 NA 160 165 ...
 $ M.I   : Factor w/ 2 levels "Imperial","Metric": 2 1 NA 2 2 1 1 2 2 2 ...
 $ Age   : num  18.2 17.6 16.9 20.3 23.7 ...
```

```
> ?survey
```

Student Survey Data: This data frame contains the responses of 237 Statistics I students at the University of Adelaide to a number of questions.

Sex	The sex of the student. (Factor with levels "Male" and "Female".)
Wr.Hnd	span (distance from tip of thumb to tip of little finger of spread hand) of writing hand, in centimetres.
NW.Hnd	span of non-writing hand.
W.Hnd	writing hand of student. (Factor, with levels "Left" and "Right".)
Fold	“Fold your arms! Which is on top” (Factor, with levels "R on L", "L on R", "Neither".)
Pulse	pulse rate of student (beats per minute).
Clap	‘Clap your hands! Which hand is on top?’ (Factor, with levels "Right", "Left", "Neither".)
Exer	how often the student exercises. (Factor, with levels "Freq" (frequently), "Some", "None".)
Smoke	how much the student smokes. (Factor, levels "Heavy", "Regul" (regularly), "Occas" (occasionally), "Never".)
Height	height of the student in centimetres.
M.I	whether the student expressed height in imperial (feet/inches) or metric (centimetres/metres) units. (Factor, levels "Metric", "Imperial".)
Age	age of the student in years.

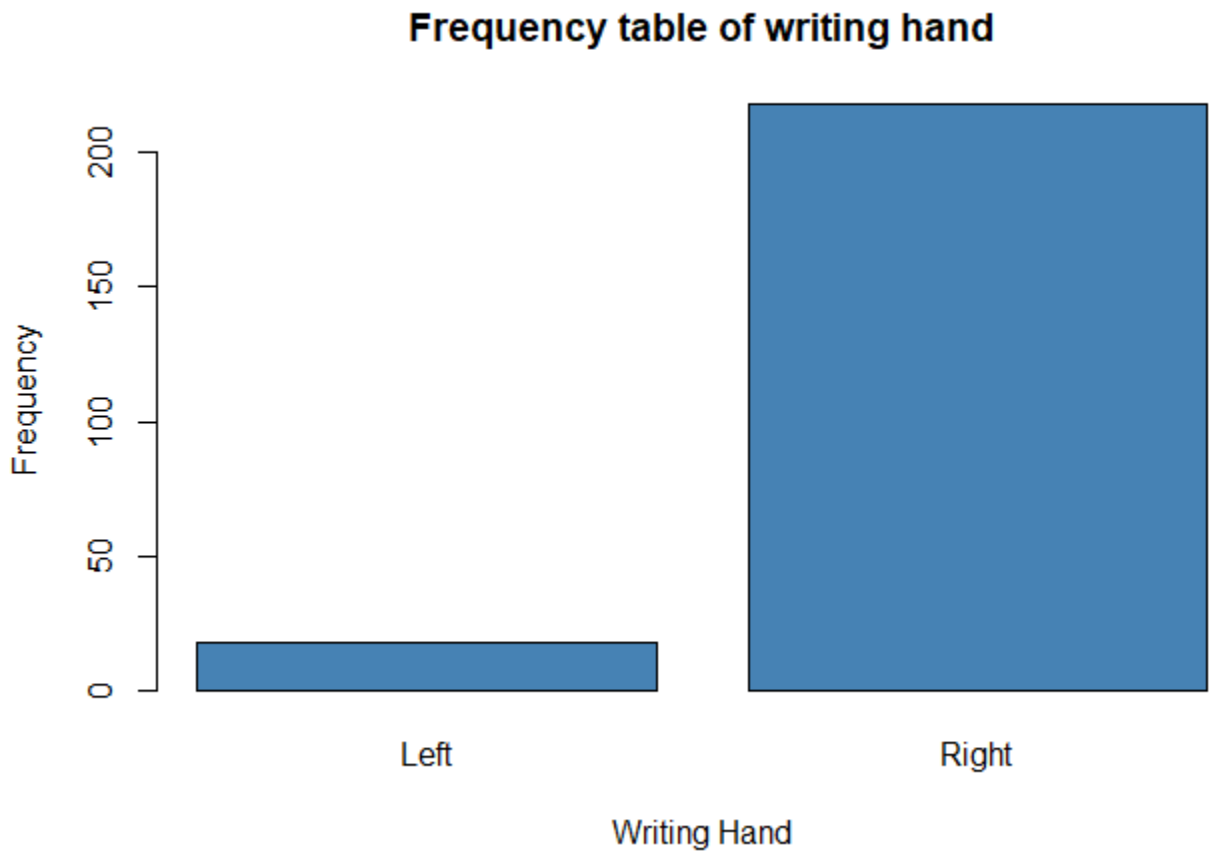
- **범주형 변수의 데이터 탐색**

```
> levels(survey$W.Hnd)
[1] "Left"  "Right"

> freq.tab <- table(survey$W.Hnd)
> freq.tab
 Left Right
   18   218

> freq.prop <- prop.table(freq.tab)
> freq.prop
      Left      Right
0.07627119 0.92372881
```

```
> barplot(freq.tab, col = "steelblue",
          xlab = "Writing Hand", ylab = "Frequency",
          main = "Frequency table of wrting hand")
```

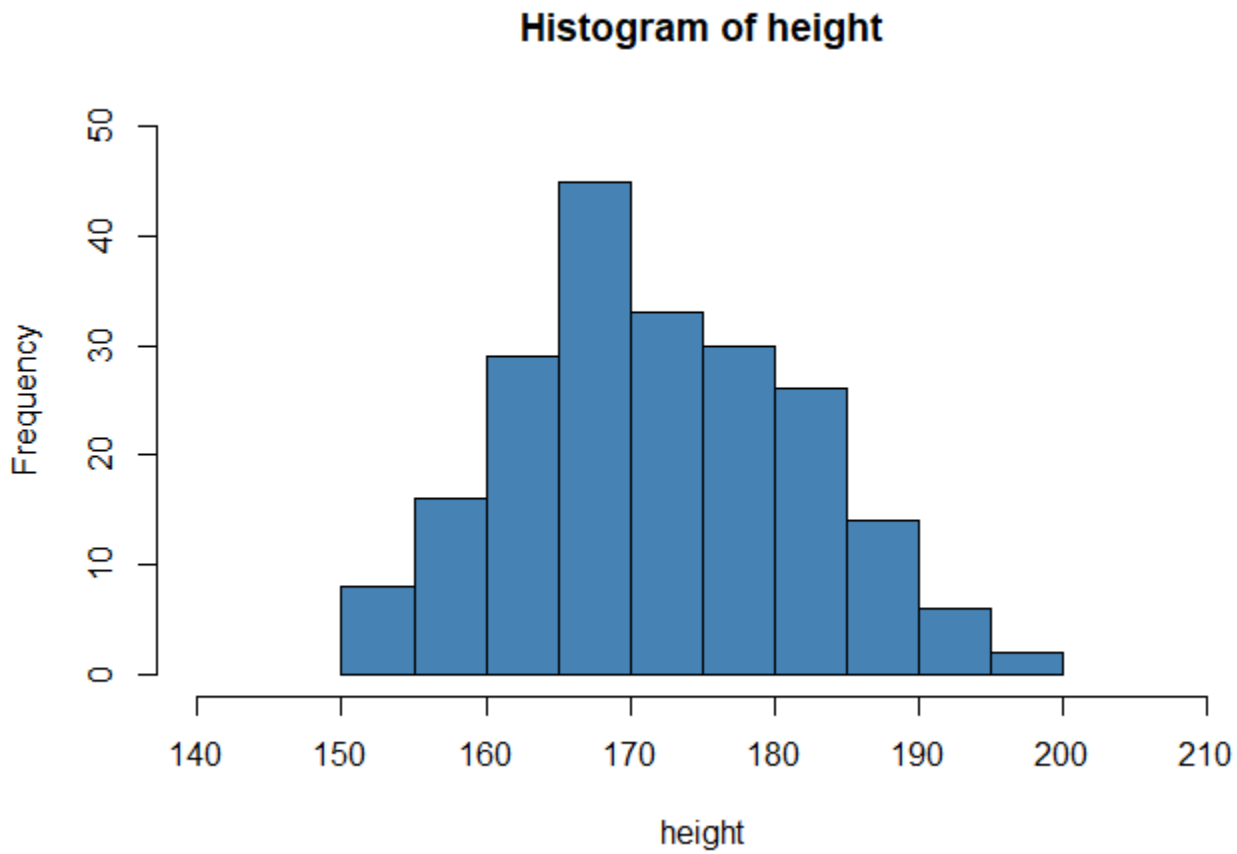


- **연속형 변수의 데이터 탐색**

```
> survey$Height
  [1] 173.00 177.80     NA 160.00 165.00 172.72 182.88 157.00 175.00
 [10] 167.00 156.20     NA 155.00 155.00     NA 156.00 157.00 182.88
 [19] 190.50 177.00 190.50 180.34 180.34 184.00     NA     NA 172.72
 [28] 175.26     NA 167.00     NA 180.00 166.40 180.00     NA 190.00
 [37] 168.00 182.50 185.00 171.00 169.00 154.94 172.00 176.50 180.34
 [46] 180.34 180.00 170.00 168.00 165.00 200.00 190.00 170.18 179.00
 [55] 182.00 171.00 157.48     NA 177.80 175.26 187.00 167.64 178.00
 [64] 170.00 164.00 183.00 172.00     NA 180.00     NA 170.00 176.00
 [73] 171.00 167.64 165.00 170.00 165.00 165.10 165.10 185.42     NA
 [82] 176.50     NA     NA 167.64 167.00 162.56 170.00 179.00     NA
 [91] 183.00     NA 165.00 168.00 179.00     NA 190.00 166.50 165.00
[100] 175.26 187.00 170.00 159.00 175.00 163.00 170.00 172.00     NA
[109] 180.00 180.34 175.00 190.50 170.18 185.00 162.56 158.00 159.00
[118] 193.04 171.00 184.00     NA 177.00 172.00 180.00 175.26 180.34
[127] 172.72 178.50 157.00 152.00 187.96 178.00     NA 160.02 175.26
[136] 189.00 172.00 182.88 170.00 167.00 175.00 165.00 172.72 180.00
......(이하 생략)
```

```
> height <- survey$Height
> length(height)
[1] 237
> mean(height)
[1] NA
> mean(height, na.rm = T)
[1] 172.3809
> median(height, na.rm = T)
[1] 171
> max(height, na.rm = T)
[1] 200
> min(height, na.rm = T)
[1] 150
> quantile(height, probs = 0.9, na.rm = T)
 90%
185.42
> quantile(height, probs = c(0.25, 0.75), na.rm = T)
25% 75%
165 180
```

```
> hist(height, col = "steelblue", breaks = 15,
       xlim = c(140, 210), ylim = c(0, 50))
```



- **연속형 변수의 통계량 요약 정보**

```
> df <- subset(survey, select = c(2, 3, 6, 10, 12))
> summary(df)
     Wr.Hnd           NW.Hnd          Pulse            Height           Age       
 Min.   :13.00   Min.   :12.50   Min.   : 35.00   Min.   :150.0   Min.   :16.75  
 1st Qu.:17.50   1st Qu.:17.50   1st Qu.: 66.00   1st Qu.:165.0   1st Qu.:17.67  
 Median :18.50   Median :18.50   Median : 72.50   Median :171.0   Median :18.58  
 Mean   :18.67   Mean   :18.58   Mean   : 74.15   Mean   :172.4   Mean   :20.37  
 3rd Qu.:19.80   3rd Qu.:19.73   3rd Qu.: 80.00   3rd Qu.:180.0   3rd Qu.:20.17  
 Max.   :23.20   Max.   :23.50   Max.   :104.00   Max.   :200.0   Max.   :73.00  
 NA's   :1       NA's   :1       NA's   :45       NA's   :28                     
```

```
> library(stargazer)
> stargazer(survey, type="text", title = "Summary of survey dataset")

Summary of survey dataset
================================================
Statistic  N   Mean   St. Dev.   Min     Max
------------------------------------------------
Wr.Hnd    236 18.669   1.879   13.000  23.200
NW.Hnd    236 18.583   1.967   12.500  23.500
Pulse     192 74.151   11.687    35      104
Height    209 172.381  9.848   150.000 200.000
Age       237 20.375   6.474   16.750  73.000
------------------------------------------------
```

- **집단별 기술통계량**: 집단의 특성을 파악하거나 집단 간의 차이를 비교하고자 할 때.

```
> mean(survey$Pulse, na.rm = T)
[1] 74.15104

> table(survey$Sex)

Female   Male
   118    118

> table(survey$Exer)

Freq None Some
 115   24   98
```

```
> tapply(survey$Pulse, INDEX = survey$Sex, FUN = mean, na.rm = T)

> with(survey, tapply(Pulse, Sex, mean, na.rm = T))
  Female     Male
75.12632 73.19792

> with(survey, tapply(Pulse, Exer, mean, na.rm = T))
    Freq     None     Some
71.96842 76.76471 76.18750

> with(survey, tapply(Pulse, list(Sex, Exer), mean, na.rm = T))
           Freq     None    Some
Female 73.60976 71.42857 77.0000
Male   70.67925 80.50000 75.0303
```

```
> aggregate(survey$Pulse, by = survey$Exer, FUN = mean, na.rm = T)
Error in aggregate.data.frame(as.data.frame(x), ...) :
  'by' must be a list

> aggregate(survey$Pulse, by = list(survey$Exer), FUN = mean, na.rm = T)

> with(survey, aggregate(Pulse, list(Exer), mean, na.rm = T))
 Group.1        x
1    Freq 71.96842
2    None 76.76471
3    Some 76.18750

> with(survey, aggregate(Pulse, list(Exercise=Exer), mean, na.rm = T))
  Exercise        x
1     Freq 71.96842
2     None 76.76471
3     Some 76.18750
```

- vcd 패키지: Arthritis 데이터셋

```
> library(vcd)
> str(Arthritis)
'data.frame':  84 obs. of  5 variables:
 $ ID       : int  57 46 77 17 36 23 75 39 33 55 ...
 $ Treatment: Factor w/ 2 levels "Placebo","Treated": 2 2 2 2 2 2 2 2 2 2 ...
 $ Sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
 $ Age      : int  27 29 30 32 46 58 59 59 63 63 ...
 $ Improved : Ord.factor w/ 3 levels "None"<"Some"<..: 2 1 1 3 3 3 1 3 1 1 ...
```

```
> ?Arthritis
```

Arthritis Treatment Data: Data from Koch & Edwards (1988) from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis.

ID	patient ID.
Treatment	factor indicating treatment (Placebo, Treated).
Sex	factor indicating sex (Female, Male).
Age	age of patient.
Improved	ordered factor indicating treatment outcome (None, Some, Marked).

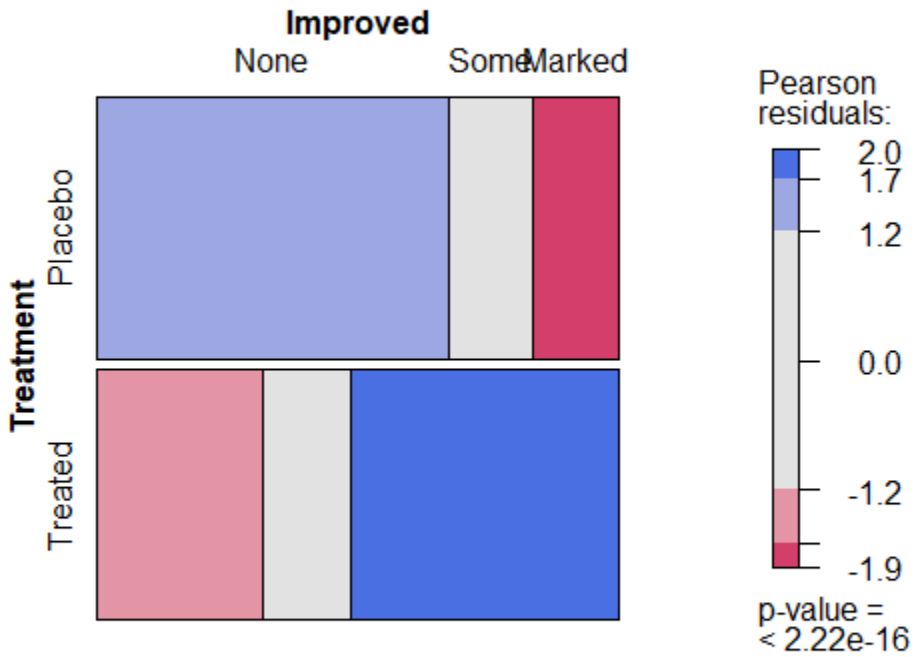
- **교차표**(cross table): **두 변수의 범주별 빈도수를 통해 범주형 변수 간의 관계를 파악**

```
> table(Arthritis$Improved, Arthritis$Treatment)

> with(Arthritis, table(Improved, Treatment))

> xtabs(~ Improved + Treatment, data = Arthritis)
Treatment
Improved Placebo Treated
  None        29      13
  Some         7       7
  Marked       7      21
```

```
> mosaic(Improved ~ Treatment, data = Arthritis, gp = shading_max)
```



- 교차표의 행과 열을 기준으로 빈도수의 합과 비율을 계산

```
> cross.tab <- with(Arthritis, table(Improved, Treatment))
> margin.table(cross.tab, margin = 1)
Improved
  None   Some Marked
    42     14     28

> margin.table(cross.tab, margin = 2)
Treatment
Placebo Treated
     43      41
```

```
> prop.table(cross.tab)
          Treatment
Improved     Placebo    Treated
  None   0.34523810 0.15476190
  Some   0.08333333 0.08333333
  Marked 0.08333333 0.25000000

> prop.table(cross.tab, margin = 1)
          Treatment
Improved     Placebo   Treated
  None   0.6904762 0.3095238
  Some   0.5000000 0.5000000
  Marked 0.2500000 0.7500000

> prop.table(cross.tab, margin = 2)
          Treatment
Improved     Placebo   Treated
  None   0.6744186 0.3170732
  Some   0.1627907 0.1707317
  Marked 0.1627907 0.5121951
```

```
> addmargins(cross.tab)
         Treatment
Improved Placebo Treated Sum
  None        29      13  42
  Some         7       7  14
  Marked       7      21  28
  Sum         43      41  84

> addmargins(cross.tab, margin = 1)
         Treatment
Improved Placebo Treated
  None        29      13
  Some         7       7
  Marked       7      21
  Sum         43      41

> addmargins(cross.tab, margin = 2)
         Treatment
Improved Placebo Treated Sum
  None        29      13  42
  Some         7       7  14
  Marked       7      21  28
```

- gmodels 패키지의 CrossTable() 함수: 교차분석을 위한 다양한 정보를 담은 교차표 생성

```
> library(gmodels)
> with(Arthritis, CrossTable(Improved, Treatment,
                             prop.r = F, prop.c = F, prop.t = T, prop.chisq = F))
```

```
Total Observations in Table:  84

             | Treatment
    Improved |   Placebo |   Treated | Row Total |
-------------|-----------|-----------|-----------|
        None |        29 |        13 |        42 |
             |     0.345 |     0.155 |           |
-------------|-----------|-----------|-----------|
        Some |         7 |         7 |        14 |
             |     0.083 |     0.083 |           |
-------------|-----------|-----------|-----------|
      Marked |         7 |        21 |        28 |
             |     0.083 |     0.250 |           |
-------------|-----------|-----------|-----------|
Column Total |        43 |        41 |        84 |
-------------|-----------|-----------|-----------|
```

- 다차원 테이블: 세 개의 범주형 변수 간의 관계를 파악

```
> cross.tab <- with(Arthritis, table(Improved, Sex, Treatment))
> ftable(cross.tab)
                Treatment Placebo Treated
Improved Sex
None     Female                19       6
         Male                  10       7
Some     Female                 7       5
         Male                   0       2
Marked   Female                 6      16
         Male                   1       5

> ftable(cross.tab, row.vars = c(2, 3))
                 Improved None Some Marked
Sex    Treatment
Female Placebo              19    7      6
       Treated               6    5     16
Male   Placebo              10    0      1
       Treated               7    2      5
```

```
> ftable(prop.table(cross.tab, margin = c(2, 3)))
                 Treatment    Placebo    Treated
Improved Sex
None     Female            0.59375000 0.22222222
         Male              0.90909091 0.50000000
Some     Female            0.21875000 0.18518519
         Male              0.00000000 0.14285714
Marked   Female            0.18750000 0.59259259
         Male              0.09090909 0.35714286
```

신약 처방을 받은 여성의 59.3%는 현저한 정상의 개선(Marked)이 있으며,
반면에 남성의 경우에는 그 비율이 35.7%에 그쳤다.

Any Questions?