1장. 범주형 데이터 정리

범주형 자료의 탐색

내용

1. 일변량 범주형 자료의 도수분포표 작성
2. 이변량 범주형 자료의 2차원 분할표 작성
3. 일변량 범주형 자료를 위한 그래프
4. 이변량 범주형 자료를 위한 그래프

통계 자료의 유형

- 양적 자료(연속형 자료)
  - 예: 키, 몸무게, 소득, 강수량, 자녀의 수

- 질적 자료(범주형 자료)
  1. 명목형 자료
     예: 성별 구분(남성=1, 여성=2). 지역 구분(서울=1, 경기=2, ...)
  2. 순서형 자료
     예: 강의평가(매우 그렇다=1, 그렇다=2, ... ,전혀 안 그렇다=5)

  - 이항 자료(binary data): 두 개의 범주를 갖고 있는 범주형 자료

예제: 패키지 vcd의 데이터 프레임 Arthritis

관절염 환자들을 대상으로 새로운 치료법의 효과를 알아보기 위해 이루어진 임상실험 자료

```
> 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 ...
 $ Sex        : Factor w/ 2 levels "Female","Male": 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 ...
```

```
> head(Arthritis, n=5)
  ID   Treatment  Sex  Age  Improved
1 57     Treated Male   27      Some
2 46     Treated Male   29      None
3 77     Treated Male   30      None
4 17     Treated Male   32    Marked
5 36     Treated Male   46    Marked
```

- 범주형 변수: Treatment(Placebo, Treated), Sex(Male, Female), Improved(None, Some, Marked)
- 연속형 변수: Age

- 설명변수: Treatment, Sex, Age
- 반응변수: Improved

- 범주형 변수 사이의 연관성 탐색
- 분석의 첫 단계: 분할표 형태로의 정리

1. 분할표 작성

- 범주형 자료를 도수분포표 혹은 분할표로 정리하기 위한 R 함수

함수	대략의 기능
table(var1, var2, . . . , varN)	N개의 범주형 변수로 N차원 분할표 작성
prop.table(table)	상대도수 분할표(두 변수의 결합분포) 작성
prop.table(table, margins)	margins로 정의된 방향으로 조건분포 작성

1) 1차원 도수분포표 및 상대도수분포표 작성

- 예: 변수 Improved의 도수분포표 작성

```
> with(Arthritis, table(Improved))
Improved
     None      Some    Marked
       42        14        28
```

- 예: 변수 Improved의 상대도수 분포표 작성

```
> my_table1 <- with(Arthritis, table(Improved) )
> prop.table(my_table1)
Improved
     None      Some    Marked
0.5000000 0.1666667 0.3333333
```

소수점 자릿수 조절이 필요한 상황

```
> options("digits")              # 디폴트 소수점 자릿수
$digits
[1] 7
> options("digits"=2)            # 소수점 자릿수 조절

> prop.table(my_table1)
Improved
  None    Some  Marked
  0.50    0.17    0.33
```

2) 2차원 분할표 및 상대도수 분할표 작성

- 예: 변수 Treatment와 Improved의 2차원 분할표 작성

```
> my_table2 <- with(Arthritis, table(Treatment, Improved))

> my_table2
                Improved
Treatment        None      Some    Marked
   Placebo         29         7         7
   Treated         13         7        21
```

- 예: 변수 Treatment와 Improved의 2차원 상대도수 분할표 작성

```
> prop.table(my_table2)
                Improved
Treatment         None    Some    Marked
    Placebo     0.345   0.083     0.083
    Treated     0.155   0.083     0.250
```

두 변수의 결합분포

3) 2차원 조건분포 분할표 작성

- 두 범주형 변수의 관계 규명에 결합분포보다 더 유용함
- 작성 방법:

prop.table(table, margin)

- table : 함수 table( )로 작성된 분할표
- margin: 조건변수 지정
  - margin=1 : 행 변수가 조건변수
  - margin=2 : 열 변수가 조건변수

- 예: Treatment를 조건변수로 하여 Improved와 상대분할표 작성

```
> prop.table(my_table2, margin=1)
                 Improved
Treatment      None   Some   Marked
    Placebo    0.67   0.16     0.16
    Treated    0.32   0.17     0.51
```

  - 변수 Treatment는 설명변수
  - Treatment가 Placebo인 경우 67%가 None
    Treated인 경우 51%가 Marked

- 예: Improved를 조건변수로 하여 Treatment와 상대분할표 작성

```
> prop.table(my_table2, margin=2)
                Improved
Treatment       None   Some   Marked
    Placebo     0.69   0.50     0.25
    Treated     0.31   0.50     0.75
```

  - 변수 Improved는 반응변수
  - Improved가 None인 경우 69%가 Placebo, 31% Treated
    Marked인 경우 25%가 Placebo, 75% Treated

2. 범주형 데이터를 위한 그래프

- 분할표: 자료의 특성을 정확하게 판단하기 어려움
- 자료의 특성 파악을 위해 적절한 그래프 이용이 필수
- 범주형 데이터에 적합한 그래프
  - 막대 그래프
  - 파이 그래프
  - Mosaic plot(이변량 이상의 경우 적합)

1) 일변량 범주형 자료를 위한 그래프

- 막대 그래프 작성을 위한 함수
  - 패키지 graphics
    - 함수 plot( ): 요인을 자료로 입력
    - 함수 barplot( ): 도수분포표를 자료로 입력
  - 패키지 ggplot2
    - 함수 geom_bar( )

- 파이 그래프 작성을 위한 함수
  - 패키지 graphics
    - 함수 pie( ): 도수분포표를 자료로 입력
  - 패키지 ggplot2
    - 함수 geom_bar( )와 coord_polar( ) 생략

- 예제: state.region 미국 50개주를 4개 지역 범주로 구분한 요인

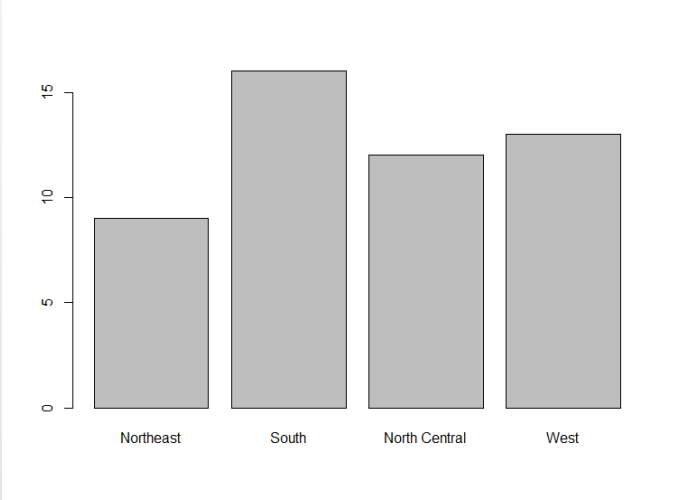
```
> state.region
 [1] South         West          West          South         West
 [6] West          Northeast     South         South         South
[11] West          West          North Central North Central North Central
[16] North Central South         South         Northeast     South
[21] Northeast     North Central North Central South         North Central
[26] West          North Central West          Northeast     Northeast
[31] West          Northeast     South         North Central North Central
[36] South         West          Northeast     Northeast     South
[41] North Central South         South         West          Northeast
[46] South         West          South         North Central West
Levels: Northeast South North Central West
```

```
> str(state.region)
 Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 ...
```

숫자의 의미는?

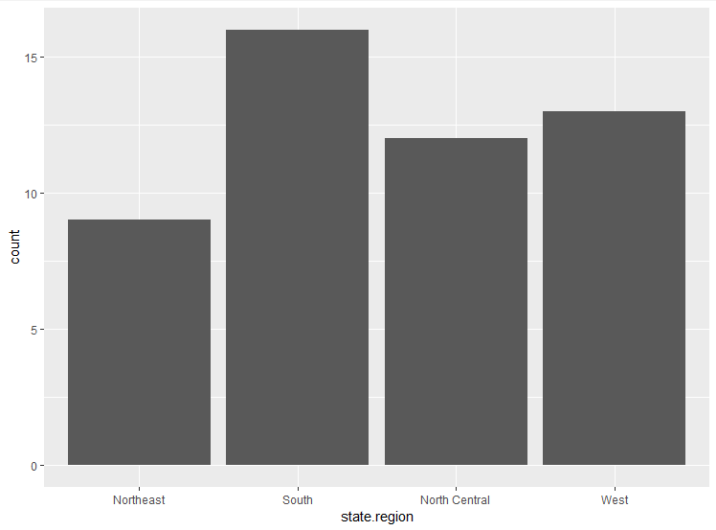
- 막대 그래프
  - 요인을 자료로 이용하는 경우
    - 함수 plot( )으로 작성

```
> plot(state.region)
```



- 패키지 ggplot2으로 작성

```
> library(ggplot2)
> ggplot(data.frame(state.region), aes(x=state.region)) +
    geom_bar()
```



함수 ggplot( ):
- 데이터 프레임 지정
- 시각적 요소와 변수 mapping

시각적 요소:
- x축 변수
- y축 변수
- 점의 색, 크기, 모양 등등

함수 aes( ):
- 시각적 요소와 변수 mapping(연결)

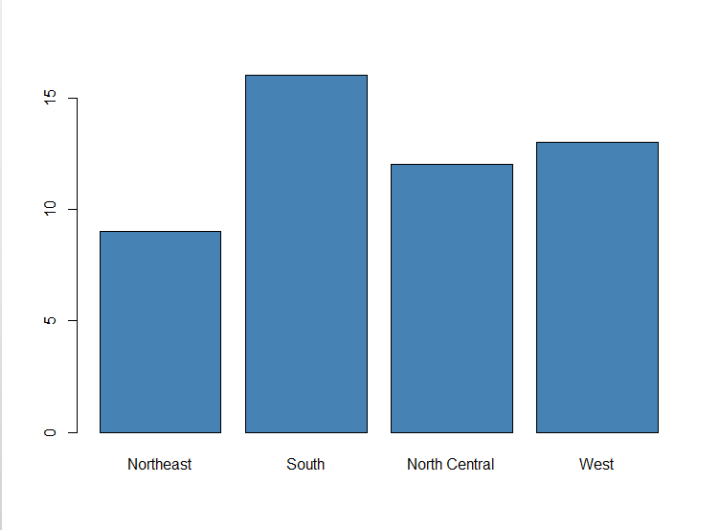
함수 geom_bar( ):
- 막대 그래프 작성

- 도수분포표를 자료로 이용하는 경우

```
> counts <- table(state.region)
> counts
state.region
    Northeast         South North Central          West
            9            16            12            13
```

- 함수 barplot( )으로 작성

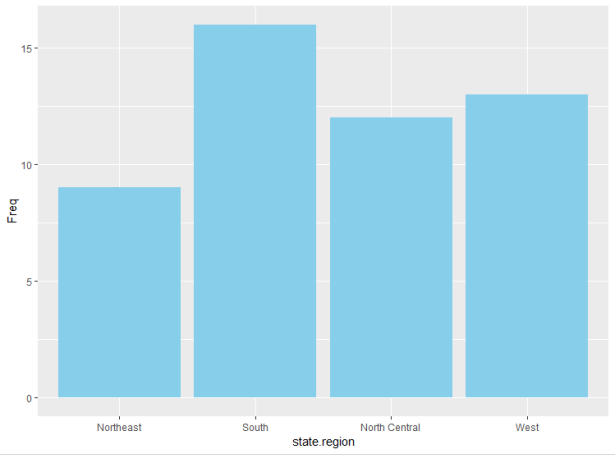
```
> barplot(counts, col="steelblue")
```



- 패키지 ggplot2에 의한 작성

```
> df_1 <- as.data.frame(counts)
> df_1
   state.region Freq
1     Northeast    9
2         South   16
3 North Central   12
4          West   13
```

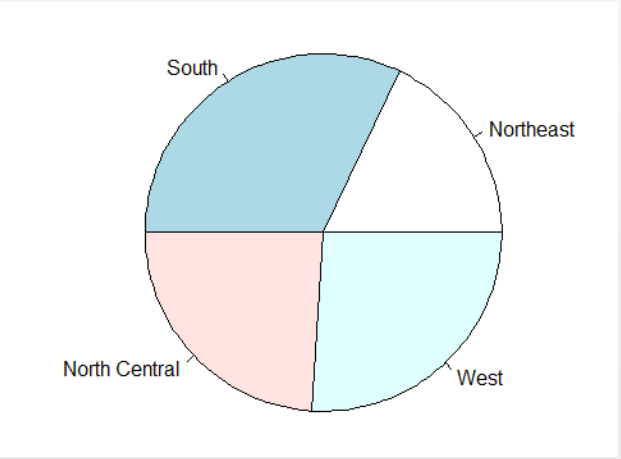
```
> ggplot(df_1, aes(x=state.region, y=Freq)) +
    geom_bar(stat="identity", fill="skyblue")
```



- 모든 geom 함수: 디폴트 stat이 있음
- stat: 입력 자료를 이용한 변환 방식 지정
  - geom_bar( ): 디폴트 stat="count"
  - 도수분포 계산하여 그래프 작성
- 도수분포가 자료로 입력
  - stat="identity": '있는 그대로' 그래프 작성

- 파이 그래프

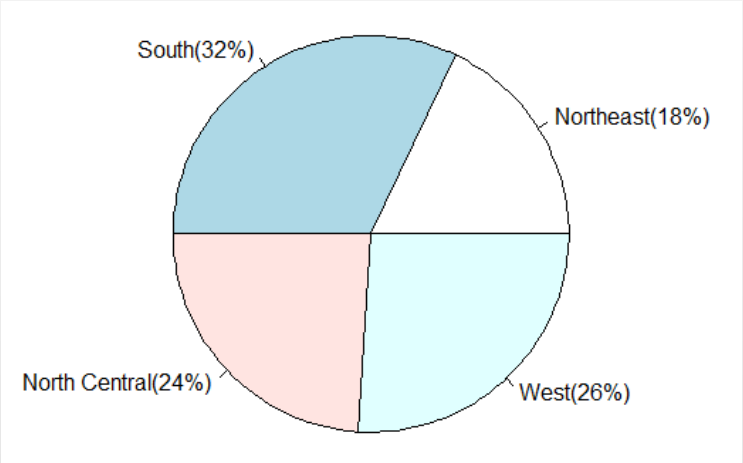
```
> counts <- table(state.region)
> pie(counts)
```



- 면적으로 빈도수 구분
- 차이 구분의 정확성: 길이 vs 면적
- "North Central" vs "West"
- 좋은 그래프는 아님

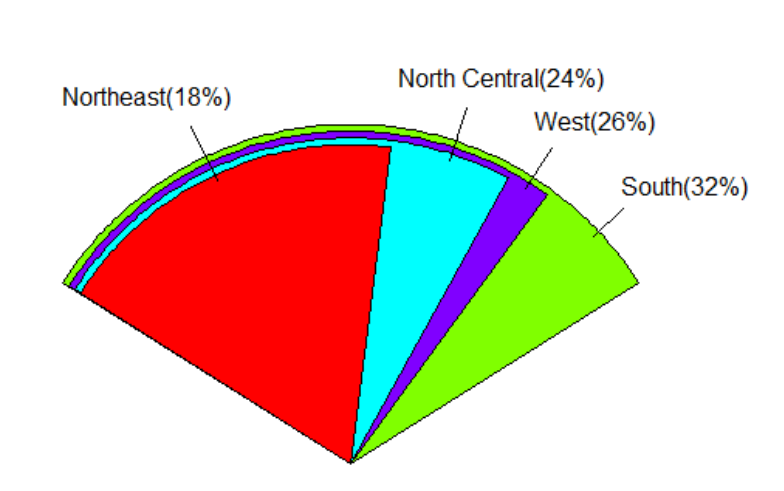
- 각 파이 조각에 라벨 추가

```
> pct <- prop.table(counts)*100
> region <- paste0(names(pct), "(", pct, "%)")
> region
[1] "Northeast(18%)"  "South(32%)"  "North Central(24%)"
[4] "West(26%)"
> pie(counts, labels=region)
```



- Fan plot:

```
> library(plotrix)
> fan.plot(counts, labels=region)
```



2) 이변량 범주형 자료를 위한 그래프

- 막대 그래프
  - 쌓아 올린 막대 그래프
  - 옆으로 붙여 놓은 막대 그래프

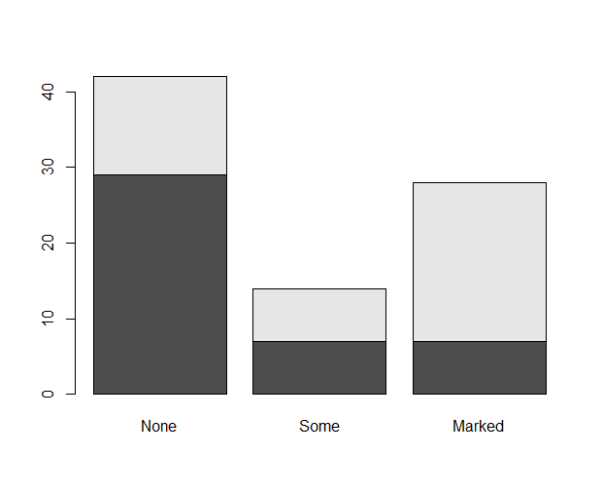
- Mosaic plot
  - 두 개 이상의 범주형 변수 관계 탐색에 유용한 그래프

- 예제: 패키지 vcd의 Arthritis

```
> library(vcd)
> my_table <- with(Arthritis, table(Treatment, Improved))
> my_table
         Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21
```

- 쌓아 올린 막대 그래프
  - 함수 barplot( )으로 작성: 분할표를 자료로 입력
  - 설명변수 위주로 쌓아 올리는 것이 더 효과적

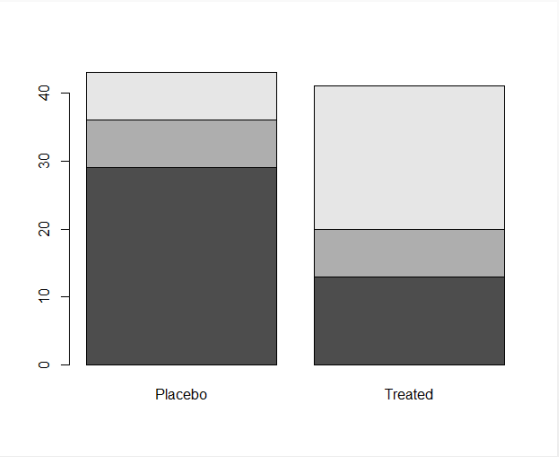
```
> barplot(my_table)
```



- 반응변수 위주로 쌓아 올린 막대: 입력된 분할표의 형태 유지
- 두 변수의 분할표의 열과 행을 바꿔 입력
- 함수 t( ): 전치 행렬

```
> my_table
           Improved
Treatment None Some Marked
  Placebo   29    7      7
  Treated   13    7     21
```

```
> barplot(t(my_table))
```



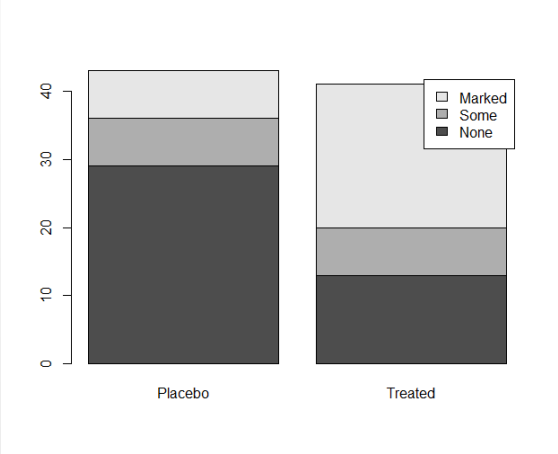
- 실제 치료가 이루어진 그룹과 위약 그룹 안에서 병세 호전 정도의 분포 확인
- 병세 호전 정도의 차이 비교 가능
- 쌓아 올려진 조각의 범례 필요

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

각 조각의 면적은 두 변수의 결합확률에 근거

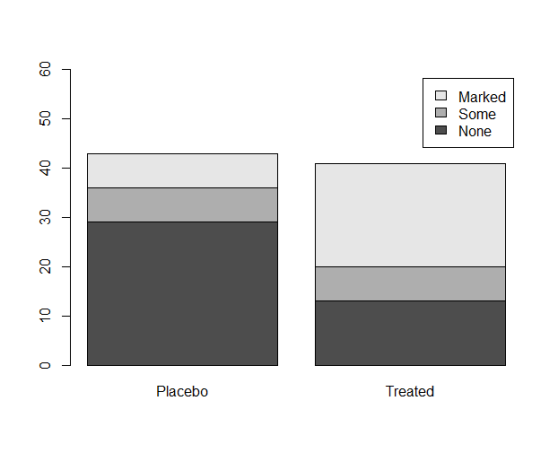
- 범례 추가

```
> barplot(t(my_table), legend.text=TRUE)
```



- 범례가 그래프와 겹쳐진 경우
- Y축 범위 확대

```
> barplot(t(my_table), legend.text=TRUE, ylim=c(0,60))
```



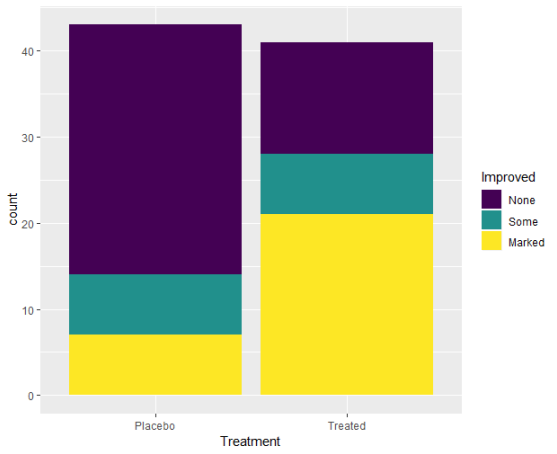
- 패키지 ggplot2로 작성

원자료 사용

```
> library(ggplot2)
> ggplot(Arthritis, aes(x=Treatment, fill=Improved)) +
    geom_bar()
```

분할표 사용

```
> ggplot(as.data.frame(my_table),
                aes(x=Treatment, y=Freq, fill=Improved)) +
    geom_bar(stat="identity")
```

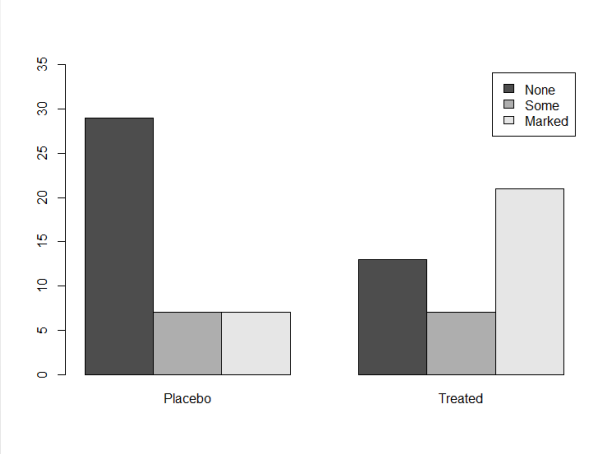


- 함수 aes( ) 안에서 시각적 요소 fill에 반응변수 mapping
- 변수 Improved의 범주별 조각이 쌓아 올려짐

- 옆으로 붙여 놓은 막대 그래프

  - 함수 barplot( )으로 작성: 분할표 입력

```
> barplot(t(my_table), beside=TRUE, legend.text=TRUE,
          ylim=c(0,35))
```



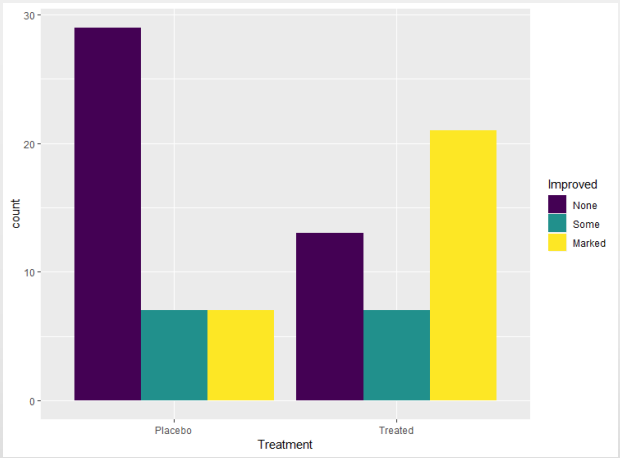
  - 옵션 beside=TRUE 사용

- ggplot2에서 작성

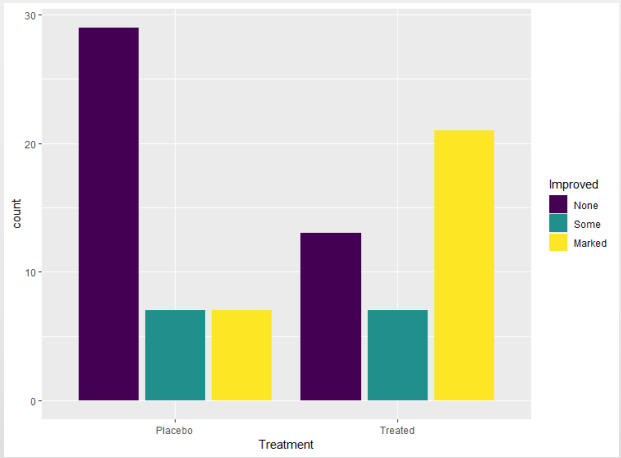
  함수 geom_bar( )에 position="dodge" 지정

  쌓아 올린 막대 그래프 경우: position="stacked" (디폴트)

```
> pp <- ggplot(Arthritis, aes(x=Treatment, fill=Improved))
> pp + geom_bar(position="dodge")
> pp + geom_bar(position="dodge2")
```



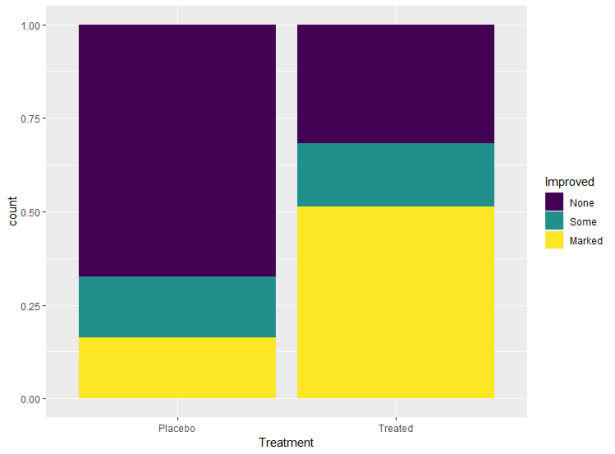
position="dodge"



position="dodge2"

- 함수 geom_bar( )에 position="fill" 지정

```
> pp + geom_bar(position="fill")
```



- 각 조각 면적: 행 변수 Treatment를 조건으로 하는 열 변수 Improved의 조건부 확률
- 두 막대의 높이는 1
- 두 변수의 관계 파악에 더 효율적

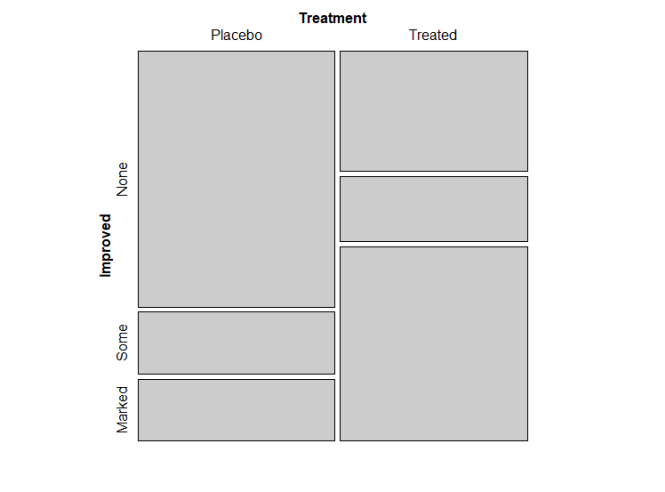
```
> prop.table(my_table, 1)
          Improved
Treatment       None      Some    Marked
  Placebo 0.6744186 0.1627907 0.1627907
  Treated 0.3170732 0.1707317 0.5121951
```

- Mosaic plot

  - 두 개 이상의 범주형 변수 관계 탐색에 유용한 그래프
  - 패키지 vcd의 함수 mosaic( )으로 작성

분할표 입력

```
> mosaic(my_table, direction="v")
```



작성 방법

- 행 변수(Treatment)의 상대도수의 비율로 정사각형을 수직으로 분리 (direction="v")
- 수직으로 분리된 두 조각을 행 변수를 조건으로 하는 열 변수의 조건부 확률에 비례하여 수평 방향으로 분리

옵션 direction

- 첫 번째 분할 방향
- 디폴트: direction="h"

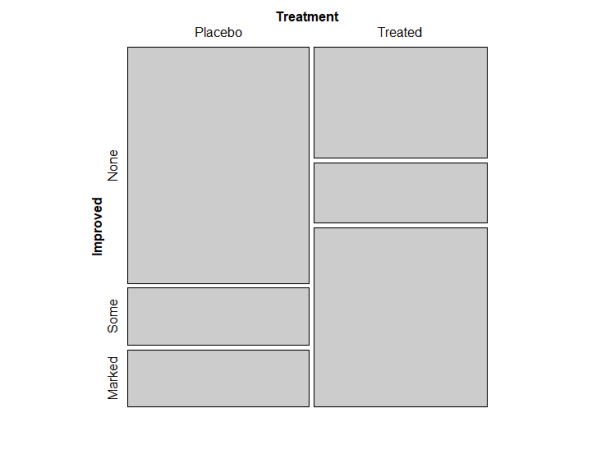
- 원자료 입력: R 공식으로 변수 선언

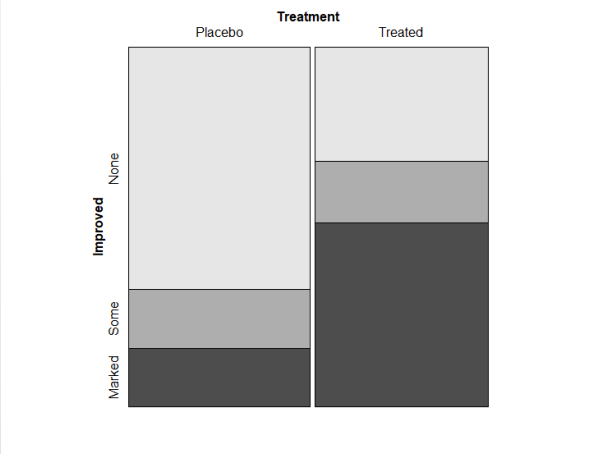
'~ 변수 + 변수' 형태

```
> mosaic(~ Treatment + Improved, data=Arthritis, direction="v")
```

'반응변수 ~ 설명변수' 형태

```
> mosaic(Improved ~ Treatment, data=Arthritis, direction="v")
```





반응변수의 수준에 따라 조각이 다른 색으로 채워짐

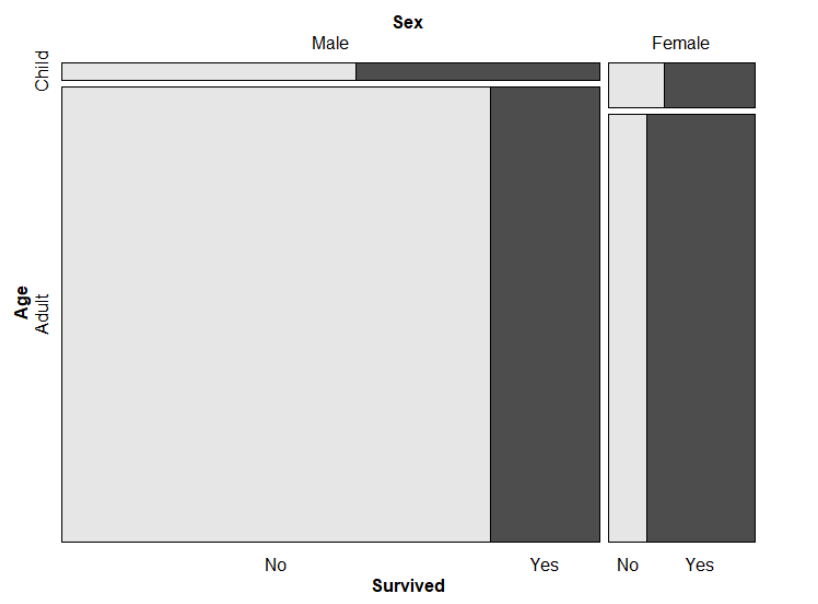
- 예제: Titanic

```
> str(Titanic)
 'table' num [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
 - attr(*, "dimnames")=List of 4
  ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
  ..$ Sex     : chr [1:2] "Male" "Female"
  ..$ Age     : chr [1:2] "Child" "Adult"
  ..$ Survived: chr [1:2] "No" "Yes"
```

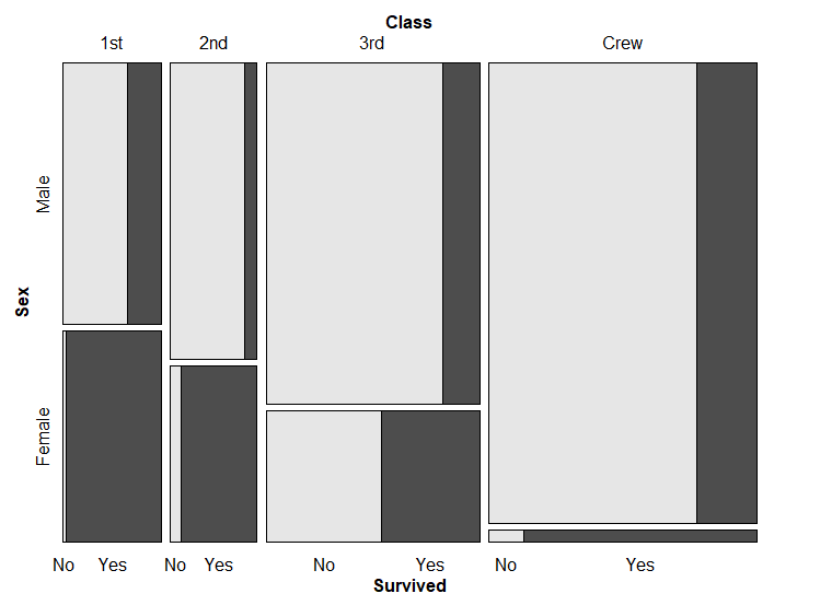
  - 반응변수: Survived
  - 설명변수: Class, Sex, Age

  생존에 큰 영향을 미친 변수는?

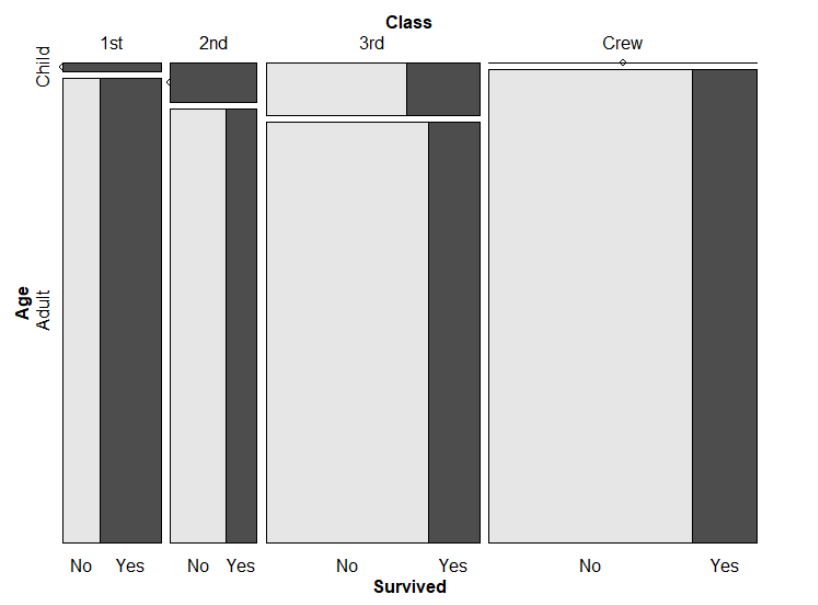
```
> mosaic(Survived ~ Sex + Age, data=Titanic, direction="v")
```



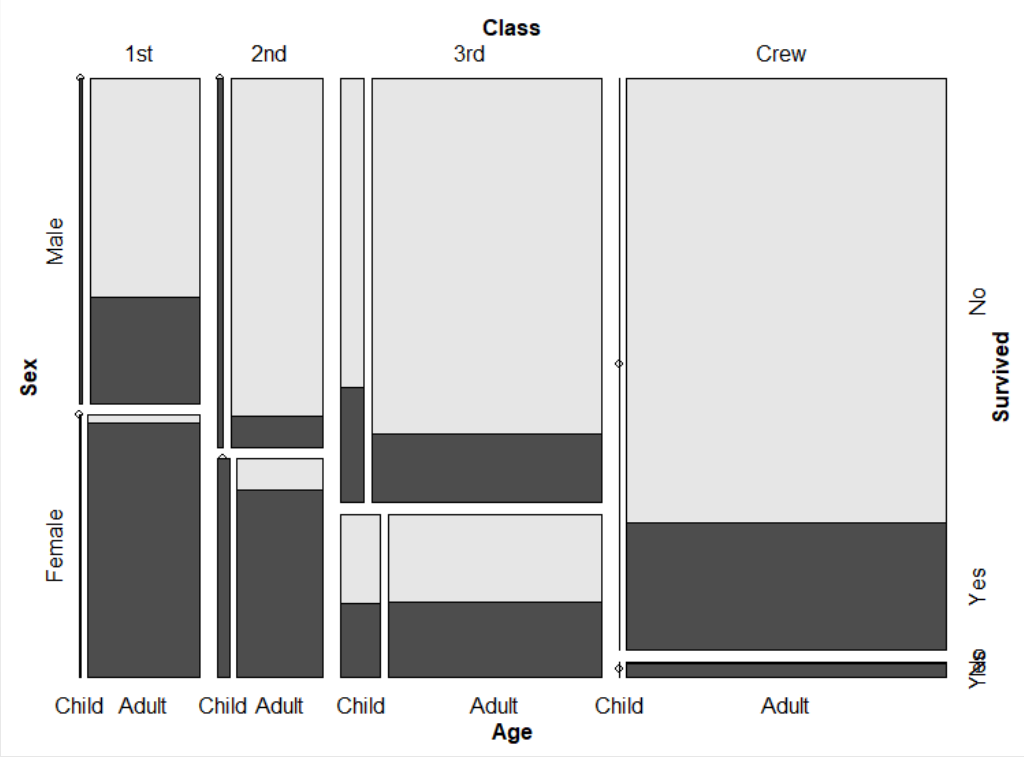
```
> mosaic(Survived ~ Class + Sex, data=Titanic, direction="v")
```



```
> mosaic(Survived ~ Class + Age, data=Titanic, direction="v")
```



```
> mosaic(Survived~., data=Titanic, direction="v")
```



예제

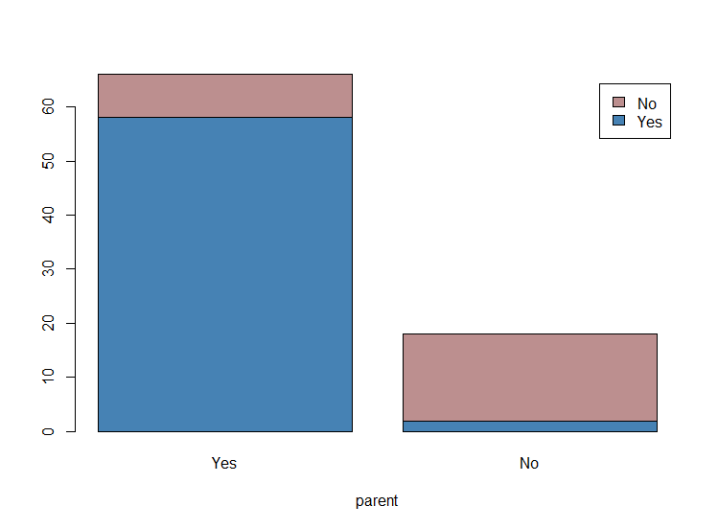
Parent	Child	
	Yes	No
Yes	58	8
No	2	16

- 부모와 어린 자녀의 안전벨트 착용여부에 대한 조사 데이터
- 부모의 안전벨트 착용여부(설명변수)가 자녀의 벨트 착용여부(반응변수)에 미치는 영향력 탐색이 주 목적

```
> belt <- matrix(c(58,2,8,16), ncol=2)
> dimnames(belt) <- list(parent=c("Yes", "No"),
                         child=c("Yes", "No"))
> belt
      child
parent Yes No
   Yes  58  8
   No    2 16
```

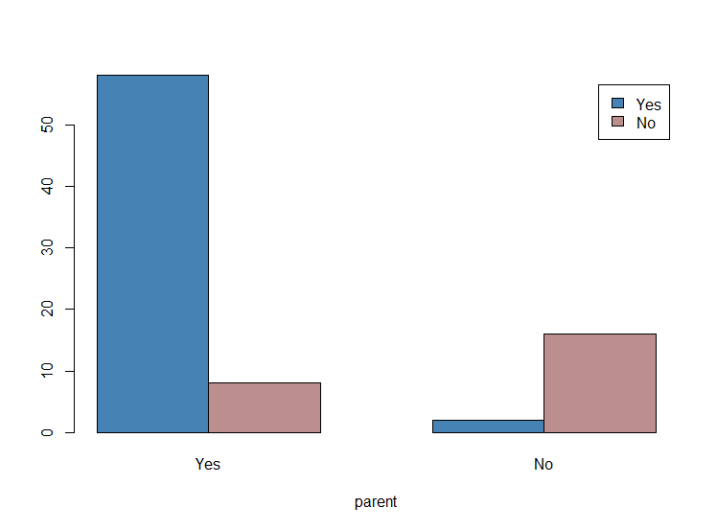
- 쌓아 올린 막대그림: 함수 barplot( )

```
> barplot(t(belt), xlab="parent", legend.text=TRUE,
          col=c("steelblue", "rosybrown"))
```



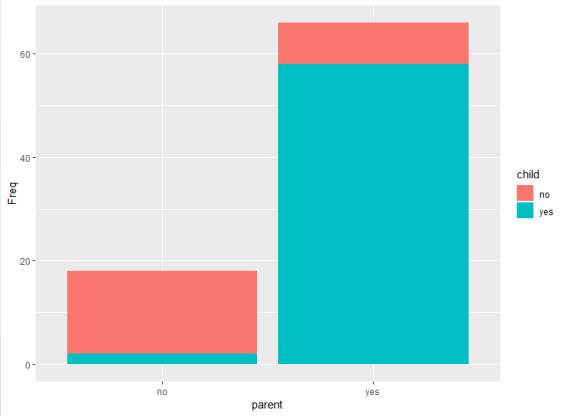
- 옆으로 붙여 놓은 막대그림: 함수 barplot( )

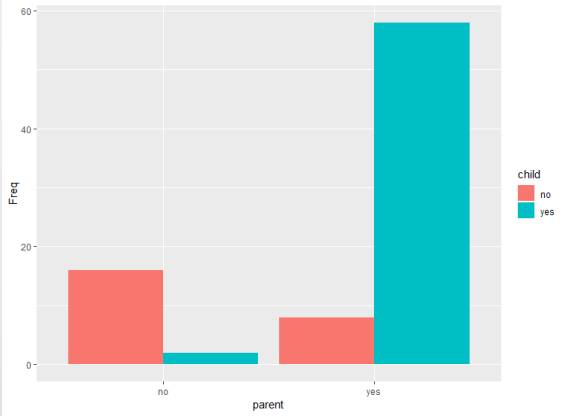
```
> barplot(t(belt), xlab="parent", beside=TRUE, legend.text=TRUE,
          col=c("steelblue","rosybrown"))
```

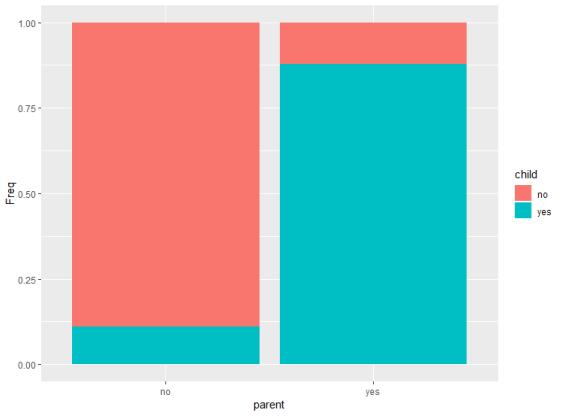


- 패키지 ggplot2로 작성

```
> df_1 <- data.frame(parent=c("yes", "yes", "no", "no"),
                     child=c("yes", "no", "yes", "no"),
                     Freq=c(58,8,2,16))
> p <- ggplot(df_1, aes(x=parent, y=Freq, fill=child))
> p + geom_bar(stat="identity")
> p + geom_bar(stat="identity", position="dodge")
> p + geom_bar(stat="identity", position="fill")
```







- Mosaic plot

```
> mosaic(belt, gp=gpar(fill=c("steelblue","rosybrown")),
         direction="v")
```