평균에 대한 비교

- 하나의 집단의 평균을 특정 수치와 비교:
  - One-sample t-test
  - 변수: 양적자료
- 두 개의 독립적인 집단의 평균 비교:
  - Two sample t-test
  - 변수: 양적자료(관심변수), 질적자료 (집단표시)
- 쌍을 이룬 두 집단의 평균차이 비교:
  - Paired t-test
  - 변수: 양적자료, 양적자료 (쌍을 이룬 자료들)
  - 변수: 양적자료, 질적자료 (개체표시)

만일 우리의 관심사가 무언가의 평균이 아니라 무언가의 비율이라면? 즉, 양적자료가 아니라 질적자료가 관심 대상이라면?

비율 (Proportions)

R 명령어

함수명	내용
binom.test(x,n)	하나의 비율: Binomial exact test
prop.test(x,n)	하나의 비율: Normal approximation
chisq.test(x,p)	적합성 검정 (여러 개의 비율 동시검정)
prop.test(c(x1,x2),c(n1,n2))	두 비율의 차이

모집단 비율에 대한 양측검정

- 예: 국립 안전심의회(NSC)

  국립 안전심의회(NSC)는 크리스마스와 연초 기간에 교통사고로 500명이 사망하고 25,000명이 부상을 입는다고 추정 하였다.
  NSC는 사고의 50%가 음주 운전으로 발생한다고 주장 하였다.

모집단 비율에 대한 양측검정

- 예: 국립 안전심의회(NSC)

120건의 교통사고를 표본으로 조사한 결과 67건이 음주운전으로 일어난 사고였다. 이 자료를 바탕으로 유의수준 $a = .05$에서 NSC의 주장을 검정하시오.

하나의 비율에 관한 검정

- Binomial exact test
  - 자료의 수가 적을 때 사용
  - 최근 컴퓨팅 기술의 향상으로 큰 자료 수에도 계산시간이 오래 걸리지 않는다.
  - `binom.test()` 사용
- 정규근사 test
  - 자료의 수가 지나치게 많아서 각 binomial density 값을 계산하는데 시간이 많이 걸릴 경우 사용
  - `prop.test()` 사용

하나의 비율: Binomial Exact Test

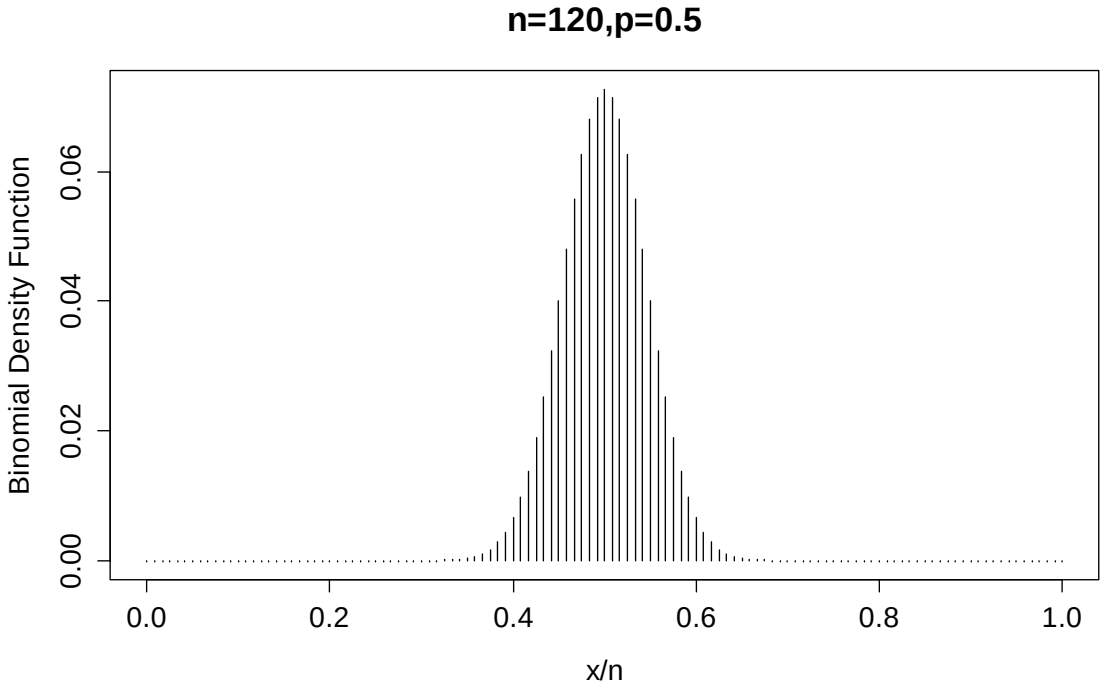
- 가설

$$H_0: p = p_0 \; vs. \; H_a: p \neq p_0$$

- 귀무가설 하에 ($p = p_0$) Binomial density 값을 구하여 표본비율이 귀무가설 하에서 얼마나 나올법한 값인가를 p-value를 통해 판단한다.

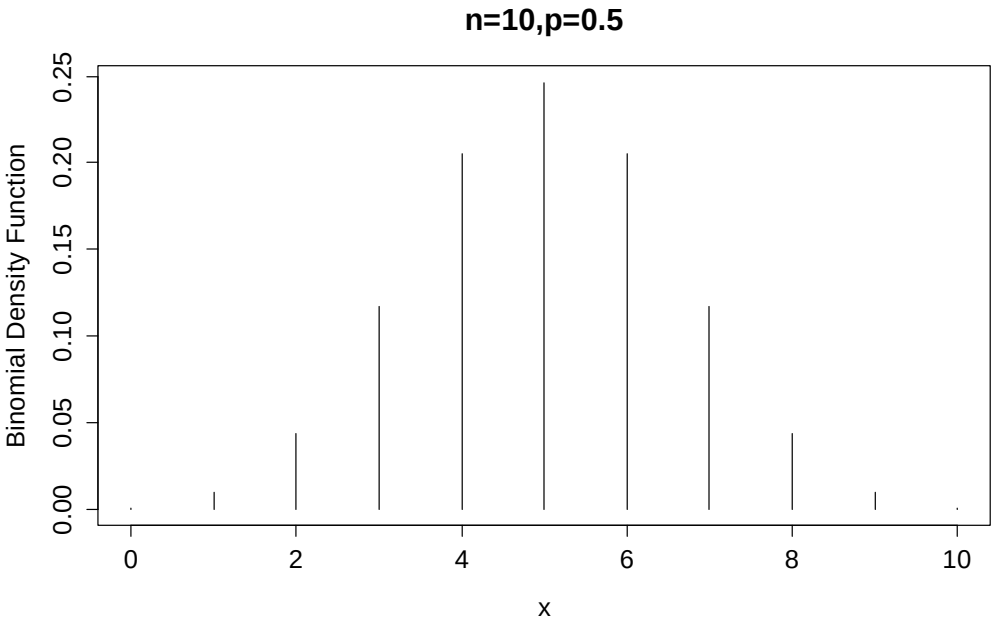
하나의 비율: Binomial Exact Test

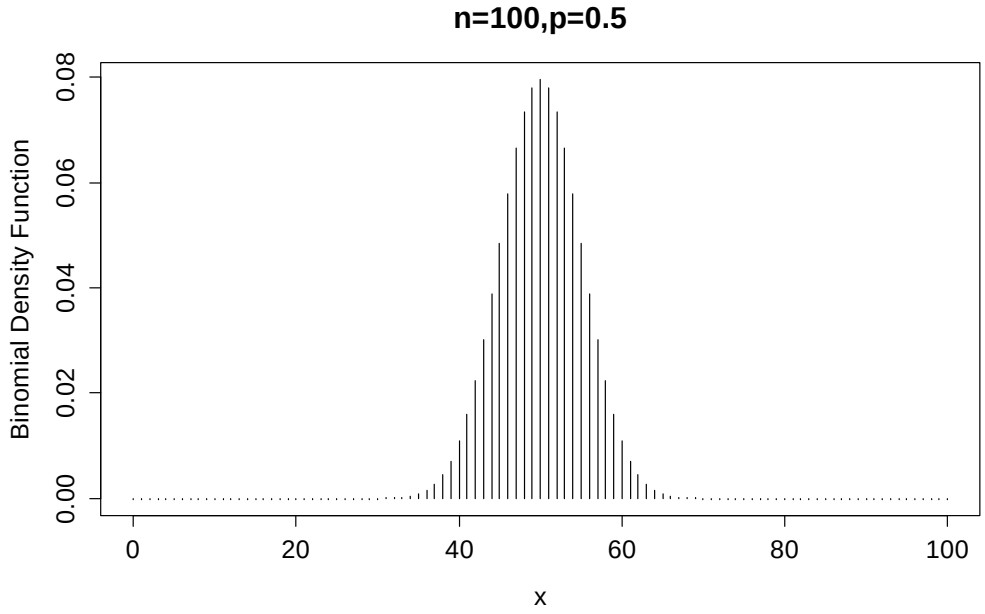
- $n =$ $, p_0 =$ $, \hat{p} =$
- 가설



하나의 비율: Binomial Exact Test

```
> binom.test(67,120)

        Exact binomial test

data:  67 and 120
number of successes = 67, number of trials = 120, p-value = 0.2352
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4648270 0.6488919
sample estimates:
probability of success
             0.5583333
```


n=10,p=0.5
Binomial Density Function
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
x

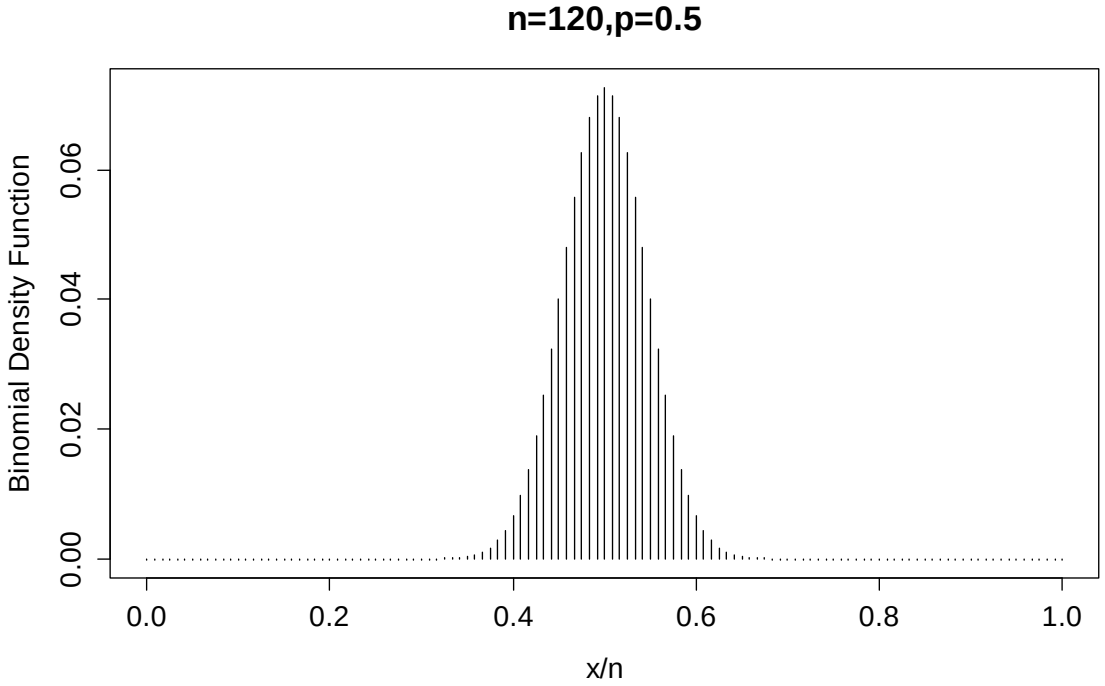

n=100,p=0.5
Binomial Density Function
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
x

하나의 비율: Binomial Exact Test

대립가설	R 명령어
$p \neq 0.5$	binom.test(x,n)
$p \neq p_0$	binom.test(x,n,p_0)
$p > p_0$	binom.test(x,n,p_0,alter="greater")
$p < p_0$	binom.test(x,n,p_0,alter="less")

하나의 비율: Normal Approximation Test

$$\frac{X}{n} = \hat{p} \sim N\left(p_0, \sqrt{\frac{p(1-p)}{n}}^2\right)$$

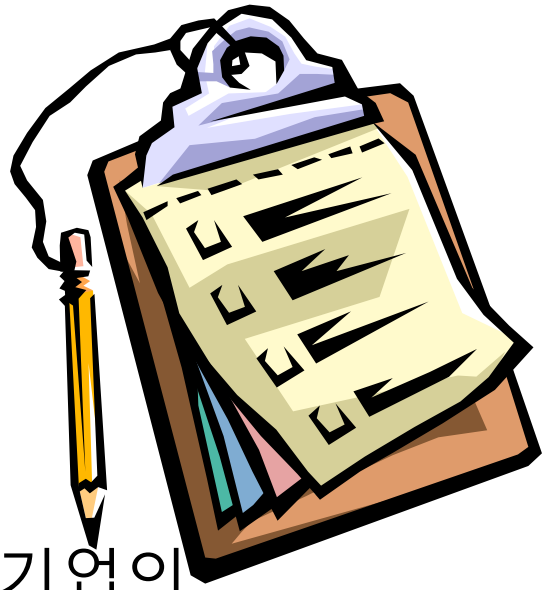


하나의 비율: Normal Approximation Test

```
> prop.test(67,120)

        1-sample proportions test with continuity correction

data:  67 out of 120, null probability 0.5
X-squared = 1.4083, df = 1, p-value = 0.2353
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4649273 0.6479534
sample estimates:
        p
0.5583333
```

두 비율의 차이

- 예: 시장조사 협회

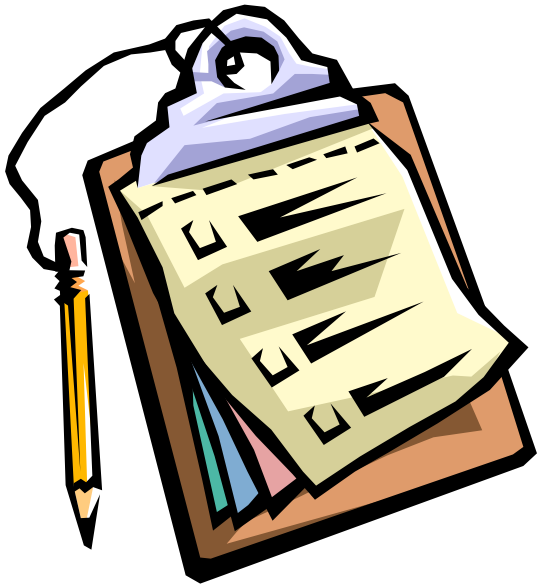


시장조사 협회는 의뢰기업의 새로운 광고 캠페인 효과를 측정하려고 한다. 새로운 캠페인이 시작 되기 전에 측정하고자 하는 시장지역의 150가구에 대하여 전화조사를 실시하였다.
조사결과, 150가구 중 60가구가 의뢰기업의 생산품에 대하여 알고 있었다.

새로운 캠페인은 TV와 신문을 통해 3주 동안 실시해 왔다.

두 비율의 차이

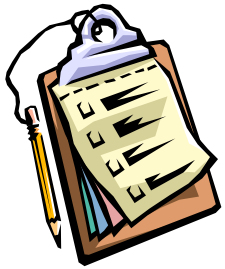
- 예 : 시장조사 협회



새로운 캠페인이 시작된 직후
실시된 조사에서는 250가구 중
120가구가 의뢰회사의 제품에 대하여
알고 있다고 한다.

이러한 자료는 '새로운 광고 캠페인이
의뢰회사의 제품에 대하여 인지도를 증가
시켰다'는 주장을 지지하는가?

두 모집단 비율의 차이에 대한 점추정량



p_1 = 새로운 캠페인 실시 전 제품에 대해 인지를 하고 있는 가구의 모집단의 비율

p_2 = 새로운 캠페인 실시 후 제품에 대해 인지를 하고 있는 가구의 모집단의 비율

$\bar{p}_1$ = 새로운 캠페인 실시 전 제품에 대하여 인지를 하고 있는 가구의 표본 비율

$\bar{p}_2$ = 새로운 캠페인 실시 후 제품에 대하여 인지를 하고 있는 가구의 표본 비율

두 비율의 차이

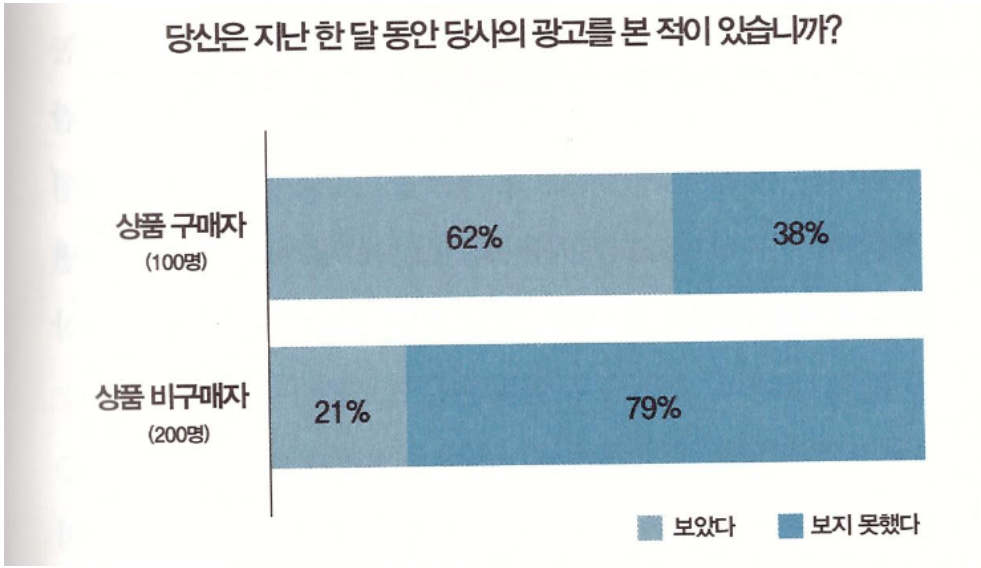
- 가설

$$H_0: p_1 = p_2 \quad vs. \quad H_a: p_1 < p_2$$

	캠페인 전	캠페인 후
인지 O	60	120
인지 X	90	130
계	150	250

```
> prop.test(c(60,120),c(150,250),alter="less")

        2-sample test for equality of proportions with continuity correction

data:  c(60, 120) out of c(150, 250)
X-squared = 2.1118, df = 1, p-value = 0.07308
alternative hypothesis: less
95 percent confidence interval:
 -1.00000000  0.00917893
sample estimates:
prop 1 prop 2
  0.40   0.48
```

여기서 잠깐: 집단 비교와 인과관계



- 보았기 때문에 샀다?
- 샀기 때문에 보았다?

여기서 잠깐: 집단 비교와 인과관계

- 폭력게임과 소년범죄
  - 범죄를 저지른 아이들은 폭력적인 게임을 즐김.
  - 폭력게임이 범죄의 원인?
  - 다른 영향요소
    - 부모의 성향
    - 타고난 본성
    - 가정환경

‘공정한 비교’를 위한 방법

- 관찰연구(Observational study)
  - 부모의 성격, 가정환경, 심리적 경향 등 ‘관련있는 조건’을 추적조사, 측정된 조건에 한해서 ‘공정한 비교’

- 실험연구(Experimental study)
  - 데이터 수집을 최대한 ‘공정’하게
  - Ex) 연수 대상자를 임의로 반반씩 나누어 한쪽은 특별연수, 한쪽은 일반연수, 업무성과를 수치화해 비교
  - 윤리적 논란 가능성 (ex. 흡연자와 비흡연자의 폐암발생 여부 비교)

적합성 검정 (Goodness of Fit Test)

- 관찰값이 기대되는 값과 일치하는지 조사하는 검정
- 질적자료(범주형자료)에서 각 카테고리에 속하는 자료의 비율이 기대하는 바와 일치하는지를 검정
- 가설

$$H_0: p_0 = p_{10}, p_2 = p_{20}, \cdots, p_k = p_{k0}$$

$$H_a: not\ H_0$$

적합도 검정

- 예 : Finger Lakes Homes (A)

Finger Lakes Homes 는
2층 콜론얼, 통나무집, 스프릿 레벨,
A프레임이라는 4가지 모델의 조립식
주택을 만들어 내고 있다.
생산계획을 수립하기 위해, 경영진은
이전 고객들의 구매형태에 어떤 선호하는
모델이 있는지를 확인하고 싶어 한다.

적합도 검정

- 예: Finger Lakes Homes (A)

  지난 2년 동안 판매된 100채의 주택이 모델별로 아래에 나타나 있다.

모델	콜로니얼	통나무	스프릿-레벨	A-프레임
판매수량	30	20	35	15

적합도 검정

- 가설

H_0: $p_C = p_L = p_S = p_A = .25$

H_a: 모집단의 비율은 $p_C = .25$, $p_L = .25$, $p_S = .25$, 그리고 $p_A = .25$ 이 아니다.

여기서:

p_C = 콜로니얼을 구매한 모집단의 비율
p_L = 통나무집을 구매한 모집단의 비율
p_S = 스플릿 레벨을 구매한 모집단의 비율
p_A = A프레임을 구매한 모집단의 비율

적합도 검정

- 검정통계량

$$\chi^2 = \sum_i \frac{(\text{관찰빈도}_i - \text{기대빈도}_i)^2}{\text{기대빈도}_i}$$

자유도=k-1

```
> x=c(30,20,35,15)
> chisq.test(x,p=c(0.25,0.25,0.25,0.25))

        Chi-squared test for given probabilities

data:  x
X-squared = 10, df = 3, p-value = 0.01857
```

독립성검정

- 두 범주형 자료가 독립인지 검정

분할표 (독립성) 검정

- 예: Finger Lakes Homes (B)

  Finger Lakes Homes 가 판매한 각 주택은 가격과 스타일에 따라 분류될 수 있다.
  Finger Lakes의 경영자는 주택의 가격과 스타일이 서로 독립적인 변수인지를 확인하고 싶어한다.

분할표 (독립성) 검정

- 예: Finger Lakes Homes (B)

  지난 2년 동안 판매된 주택들을 가격과 모델에 따라 아래에 분류하였다.
  편의상, 가격은 $99,000 보다 큰 경우와 $99,000 보다 작거나 같은 경우로 나타낸다.

가격	콜로니얼	통나무	스프릿	A-프레임
≤ $99,000	18	6	19	12
> $99,000	12	14	16	3

분할표 (독립성) 검정

- 가설

H_0: 주택의 가격은 구매된 주택의 스타일과 독립적이다.

H_a: 주택의 가격은 구매된 주택의 스타일과 독립적이지 않다.

```
> chisq.test(matrix(c(18,12,6,14,19,16,12,3),ncol=4))

        Pearson's Chi-squared test

data:  matrix(c(18, 12, 6, 14, 19, 16, 12, 3), ncol = 4)
X-squared = 9.1486, df = 3, p-value = 0.02738
```

분할표 (독립성) 검정

- 검정통계량

$$\chi^2 = \sum_i \frac{(\text{관찰빈도}_i - \text{기대빈도}_i)^2}{\text{기대빈도}_i}$$

자유도= (k-1)*(m-1)

```
> chisq.test(matrix(c(18,12,6,14,19,16,12,3),ncol=4))

        Pearson's Chi-squared test

data:  matrix(c(18, 12, 6, 14, 19, 16, 12, 3), ncol = 4)
X-squared = 9.1486, df = 3, p-value = 0.02738
```

분할표 (Contingency Table)만들기

- 두 개의 범주형 자료
- 각 변수 당 2개 이상의 카테고리 존재 (예, m개, k개)
- 총 m*k개의 cell

분할표 (Contingency Table)만들기

- 예) 124명 (test group: 60, placebo:64)을 대상으로 병세가 호전되는지 (outcome=1)그렇지 않은지 (outcome=0) 조사하여 아래와 같은 결과를 얻었다.

	OUTCOME		
Treat	**0**	**1**	**합**
Placebo	48	16	64
Test	20	40	60

분할표 (Contingency Table)만들기

- 각 셀에 해당하는 자료를 알고 있을 때
  - matrix(벡터, ncol=열의 개수)

```
> matrix(c(48,20,16,40),ncol=2)
     [,1] [,2]
[1,]   48   16
[2,]   20   40
```

분할표 (Contingency Table)만들기

- 각 셀에 해당하는 자료의 수가 두 변수의 카테고리를 나타내는 변수와 함께 열로 나타나 있을 때
  - xtabs(도수 ~ 가로+세로)

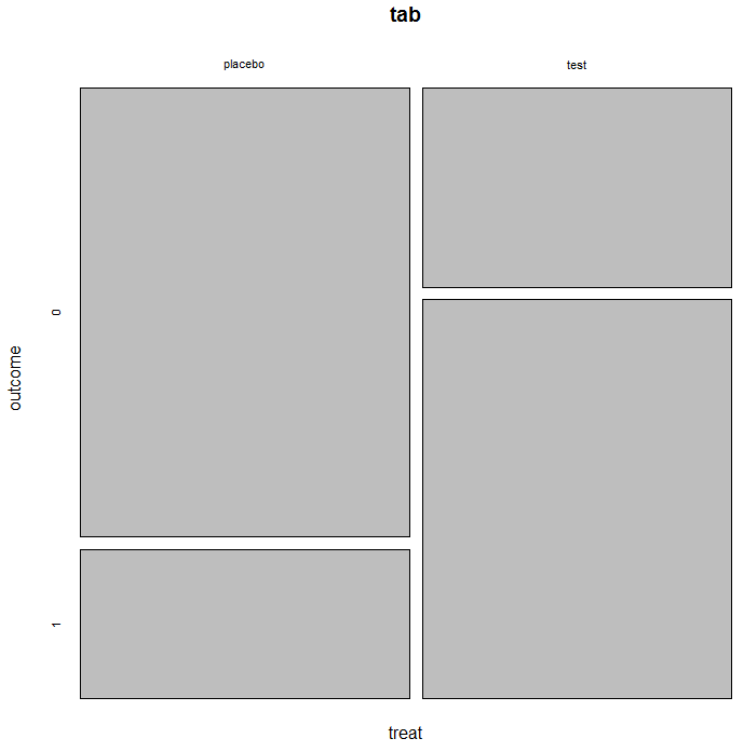
```
> respire
    treat outcome count
1 placebo       1    16
2 placebo       0    48
3    test       1    40
4   test        0    20

> xtabs(count~treat+outcome,data=respire)
         outcome
treat      0  1
  placebo 48 16
  test    20 40
```

분할표 (Contingency Table)만들기

- 각 셀에 해당하는 자료의 수가 계산되어 있지 않고 모든 관찰치에 대해 두 변수의 값이 나열되어 있을 때
  - xtabs(~ 가로+세로)

```
> respire2
      treat outcome
1   placebo       1
2   placebo       1
3   placebo       1
4   placebo       1
5   placebo       1
6   placebo       1
7   placebo       1
…
> xtabs(~treat+outcome,data=respire2)
         outcome
treat      0  1
  placebo 48 16
  test    20 40
```

Mosaic plot



- 막대폭=treatment의 빈도에 비례
- 막대길이=outcome의 빈도에 비례

```
> tab=xtabs(~treat+outcome,data=respire2)
>
> mosaicplot(tab)
```