# 모비율에 대한 검정

한 모집단의 비율

# 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.

120건의 교통사고를 표본으로 조사한 결과 67건이 음주운전으로 일어난 사고였다.

- 이 자료를 바탕으로 음주운전으로 일어난 사고의 비율에 대한 95% 신뢰구간을 구하시오.
- 유의수준 $\alpha$ = .05에서 NSC의 주장을 검정하시오.

# 모비율에 대한 구간추정

$$\hat{p} \pm z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

```
> prop.test(67,120,p=0.5)

        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
```

# 모비율에 대한 가설검정

## 1. 귀무가설과 대립가설 설정

$H_0$:

$H_a$:

## 2. 가정체크

$$np \geq 5, \qquad n(1-p) \geq 5$$

## 3. 검정 통계량과 P-값

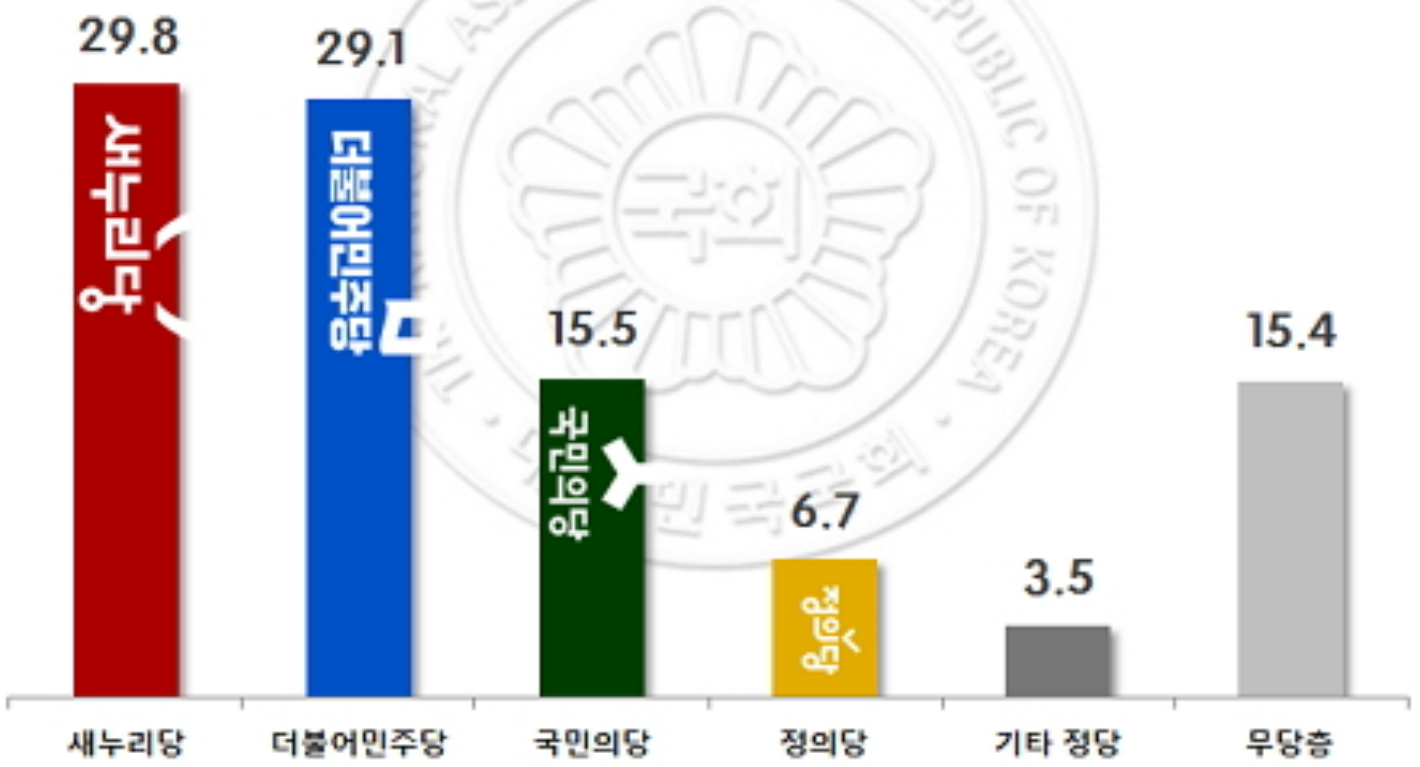
```
> prop.test(67,120,p=0.5)

        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
```

## 4. 결론

MK 매일경제 MBN

# 정당 지지도: '16년 6월 4주차 주간집계

단위: %

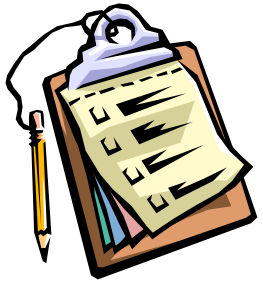



•조사기관: 리얼미터 •총응답자: 전국 성인 2,539명 •응답률: 8.4%
•조사방법: 무선(70%)-유선(30%) RDD 전화면접·스마트폰앱·자동응답 혼용 •표집오차: 95% 신뢰수준 ±1.9%p
•조사기간: 2016년 6월 20일(월)~24일(금)

# 모비율에 대한 검정

두 모집단의 비율 비교

# 두 비율의 차이



- 예: 시장조사 협회

시장조사 협회는 의뢰기업의 새로운 광고 캠페인 효과를 측정 하려고 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. 새로운 캠페인이 시작된 직후 실시된 조사에서는 250가구 중 120가구가 의뢰회사의 제품에 대하여 알고 있다고 한다.

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

# 두 모집단 비율의 차이에 대한 추정량

- 모수: $p_1 - p_2$
  - $p_1$ = 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

- 추정량: $\widehat{p_1} - \widehat{p_2}$
  - $\widehat{p_1}$=새로운 캠페인 실시 전 제품에 대하여 인지하고 있는 가구의 표본 비율
  - $\widehat{p_2}$=새로운 캠페인 실시 후 제품에 대하여 인지하고 있는 가구의 표본 비율

## 1. 귀무가설과 대립가설 설정

$$H_0: p_1 - p_2 = 0 \quad vs. \quad H_a: p_1 - p_2 < 0$$

## 2. 가정체크

$$n_1 p_1 \geq 5, \qquad n_1(1 - p_1) \geq 5$$
$$n_2 p_2 \geq 5, \qquad n_2(1 - p_2) \geq 5$$

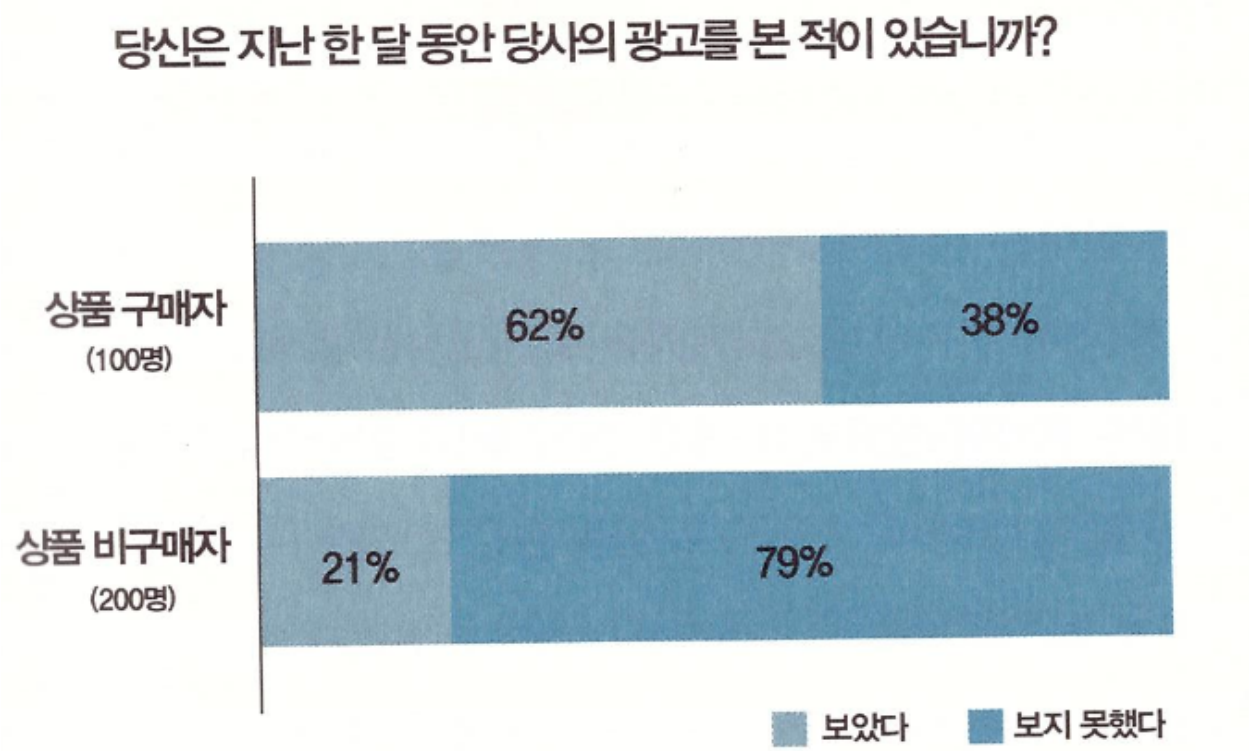
| | 캠페인 전 | 캠페인 후 |
|---|---|---|
| 인지 O | 60 | 120 |
| 인지 X | 90 | 130 |
| 계 | 150 | 250 |

## 3. 검정 통계량과 P-값

```
> 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
```

## 4. 결론

# 여기서 잠깐: 집단 비교와 인과관계



# 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)

# 모비율에 대한 검정

독립성검정

# 독립성(분할표) 검정

- 두 범주형 자료가 독립인지 검정

- 예
  - 맥주 선호도는 성별에 독립인가?
    맥주종류 (라이트맥주, 일반맥주, 흑맥주) vs. 성별
    - 독립적이라면 각 맥주를 선호하는 비율이 남녀가 동일

  - 생산현장에서 불량품 발생률이 작업교대조별(혹은 종류별)로 관련이 있는가?

  - 여론조사에서 특정 주제에 대한 의견이 성별(혹은 지역, 소득수준, 교육수준) 과 관련이 있는가?

# 예: 우울증

- 소득수준이 우울증에 영향을 미치는지 알기 위해 300명을 무작위 추출하여 다음의 결과를 얻었다. 소득수준이 우울증과 관련이 있다고 할 수 있는지를 유의수준 0.05에서 검정하라.

| 소득수준 | 우울증상 | |
|---|---|---|
| | 있다 | 없다 |
| 저소득 | 33 | 67 |
| 중간층 | 28 | 122 |
| 고소득 | 5 | 45 |

## 가설

$H_0$: 우울증과 소득수준은 독립이다.
$H_1$: 우울증과 소득수준은 독립적이지 않다.

## 검정통계량

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

자유도= (k-1)*(m-1)

```
> chisq.test(matrix(c(33,28,5,67,122,45),3,2))

 Pearson's Chi-squared test

data:  matrix(c(33, 28, 5, 67, 122, 45), 3, 2)
X-squared = 12.2183, df = 2, p-value = 0.002222
```

# 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)

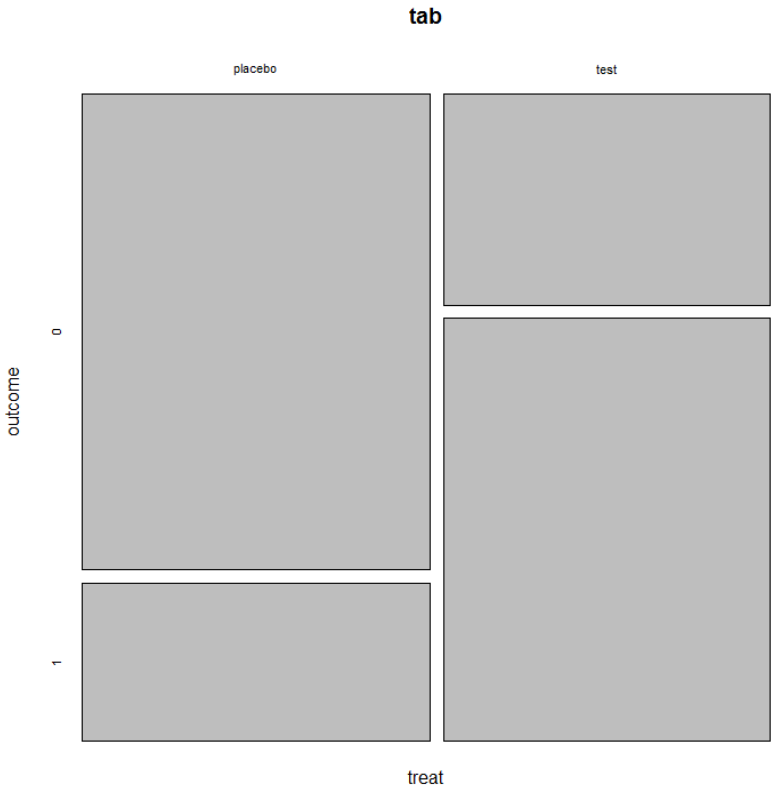
```
> respire
    treat outcome count
1 placebo       1    16
2 placebo       0    48
3    test       1    40
4   test        0    20

> 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




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

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

# 변수유형에 따른 분석기법

| | | 설명변수 (분석축) | | | |
|---|---|---|---|---|---|
| | | 한 그룹과 특정 숫자와의 비교 | 두 그룹의 비교 | 셋 이상 그룹의 비교 | 양적변수 (연속값)의 크기로 비교 |
| **반응변수 (분석하고 싶은 것)** | 양적변수 (연속값 등) | **한 집단의 평균 T-검정** | **독립표본 T-검정, 쌍체표본 T-검정** | **분산분석 (ANOVA)** | **회귀분석** |
| | 질적변수 (Yes/No) | **한 집단의 비율 Z-검정** | **두 집단의 비율 Z-검정** | **분할표 카이제곱 검정** | **로지스틱 회귀분석** |