

# 강의교안 이용 안내

- 본 강의교안의 저작권은 이윤환과 한빛아카데미(주)에 있습니다.
- 이 자료를 무단으로 전제하거나 배포할 경우 저작권법 136조에 의거하여 벌금에 처할 수 있고 이를 병과(併科)할 수도 있습니다.





제대로 알고 쓰는

**R 통계분석**

## CHAPTER 08

# 범주형 자료분석

# Contents

## 8.1 적합도 검정

- 관찰도수와 기대도수에 대해 학습한다.
- 적합도 검정에 대해 학습한다.

## 8.2 동질성 검정과 독립성 검정

- 분할표에 대해 학습한다.
- 동질성 검정과 독립성 검정에 대해 학습한다.

## 9장을 위한 준비



# 01. 적합도 검정

- 관찰도수와 기대도수에 대해 학습한다.
- 적합도 검정에 대해 학습한다.

# 범주형 자료분석

- 표본으로부터 수집된 자료의 유형이 값으로부터 사칙연산 등을 수행할 수 있는 양적자료에 대해 평균, 표준편차, 비율 등 특성에 대한 가설검정을 알아보았습니다.
- 값으로부터 직접 계산을 할 수 없는 질적자료의 경우 실시할 수 있는 분석에 대해 알아보시다.
  - 여기서 중요한 계산은 질적자료의 각 범주의 개수를 세는 것으로 R에서 이를 위해 앞 장에서 table을 만들어 보았습니다.

# 적합도 검정

- 멘델이 제안한 여러 유전 법칙 중 다음의 법칙을 잘 알고 계실 것입니다.
  - 순종의 둥글고 황색인 완두(RRYY)콩과 주름지고 녹색인 완두(rryy)콩을 교배하면,
  - 제1대에서는 잡종인 둥글고 황색인 완두(RrYy)콩만 나타나고,
  - 이 잡종 1대를 자화수분시키면 제2대에서는 나타날 수 있는 경우가
  - 둥글고 황색, 둥글고 녹색, 주름지고 황색, 주름지고 녹색의 네 가지이며,
  - 이들의 출현 비율은 9:3:3:1이 된다.
- 멘델의 실험
  - 556개의 완두콩 관찰
  - 둥글고 황색인 콩 315개, 둥글고 녹색인 콩 108개, 주름지고 황색인 콩 101개, 주름지고 녹색인 콩 32개

# 적합도 검정

## 예제 1 멘델의 유전법칙

- 순종의 둥글고 황색인 완두(RRYY)콩과 주름지고 녹색인 완두(rryy)콩을 교배하면 2대째 발현되는 완두콩의 형질은 둥글고 황색, 둥글고 녹색, 주름지고 황색, 주름지고 녹색이 9:3:3:1의 비율로 나타난다.
- 다음은 관찰실험의 결과입니다.

구분	둥글고 황색	주름지고 황색	둥글고 녹색	주름지고 녹색	합
개체 수	315	101	108	32	556
비율(%)	56.7%	18.2%	19.4%	5.8%	100%

- 2대째 발현되는 완두콩의 형질은 4개의 범주를 갖는 범주형 자료로 모집단을 네 개의 범주로 나누고, 표본에서 관찰되는 각 범주에 해당하는 개수를 통해 모집단의 비율을 추정할 수 있습니다

# 적합도 검정

- 멘델의 실험 자료를 통해 우리가 잘 알고 있는 9:3:3:1의 비율에 맞게 나타난 것인지 통계적 가설검정을 통해 확인해 봅시다.
- 가설수립
  - 영가설 : 완두콩의 모양과 색깔의 2대 유전은 9:3:3:1로 나타난다.

$$H_0 : p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16}$$

- 대안가설 : 완두콩의 모양과 색깔의 2대 유전은 9:3:3:1이 아니다.

$$H_1: not H_0$$



# 범주형 자료분석

- 검정통계량

- 기대도수( $E_i$ )

- 검정통계량의 계산은 영가설 하에서 실시함을 잘 알고 있을 것입니다.
    - 영가설이 참, 즉 9:3:3:1의 비율로 2대의 형질이 발현됐다고 하면,
    - 556개의 실험 대상에 대해 둥글고 황색인 콩은  $556 \times \frac{9}{16}$ 개, 둥글고 녹색인 콩과 주름지고 황색인 콩은  $556 \times \frac{3}{16}$ 개 개, 주름지고 녹색인 콩은  $556 \times \frac{1}{16}$ 개 개가 관찰될 것입니다.
    - 각 범주의 영가설 하에서의 비율을 곱해 계산된 수를 기대도수라고 합니다

- 관찰도수( $O_i$ )

- 앞서 실험을 통해 관찰한 각 형질의 개수와 같이 각 범주별로 관찰한 개수를 관찰도수라고 합니다.

# 범주형 자료분석

구분	등글고 황색	주름지고 황색	등글고 녹색	주름지고 녹색	합
영가설하 의 비율	9/16	3/16	3/16	1/16	1
기대도수	312.75	104.25	104.25	34.75	556
관찰도수	315	101	108	32	556

- 기대도수와 관찰도수의 차이를 생각해봅시다.
  - 만일 영가설이 참이라면 관찰도수와 기대도수의 차이는 크지 않을 것입니다.

## ① 관찰도수와 기대도수의 차이

- 등글고 황색 :  $O_1 - E_1 = 315 - 312.75 = 2.25$
- 주름지고 황색 :  $O_2 - E_2 = 101 - 104.25 = -3.25$
- 등글고 녹색 :  $O_3 - E_3 = 108 - 104.25 = 3.75$
- 주름지고 황색 :  $O_4 - E_4 = 32 - 34.75 = -2.75$

# 범주형 자료분석

- ▣ 차이의 총합을 구해 그 크기가 크지 않다면, 영가설을 채택하고자 하는데 분산을 구할 때와 마찬가지로 부호(+, -)에 의해 차이가 상쇄될 수 있습니다.
- ▣ 이에 이 차이의 제곱을 구합니다.

## ② 관찰도수와 기댓도수의 차이의 제곱

- 등글고 황색 :  $(O_1 - E_1)^2 = 2.25^2 = 5.6025$
- 주름지고 황색 :  $(O_2 - E_2)^2 = (-3.25)^2 = 10.5625$
- 등글고 녹색 :  $(O_3 - E_3)^2 = 3.75^2 = 14.0625$
- 주름지고 황색 :  $(O_4 - E_4)^2 = (-2.75)^2 = 7.5625$
- ▣ 관찰도수와 기댓도수 차이는 기댓도수 대비한 차이를 구합니다.
  - 예로써 다음과 같이 차이의 제곱이 1로 동일한 경우를 생각해봅시다.

구분	$(O_i - E_i)^2$	관찰도수	기댓도수
경우 1	1	10001	10000
경우 2	1	1	2

# 범주형 자료분석

- ▣ 경우 1과 경우 2 모두 관찰도수와 기댓도수의 차이의 제곱이 1입니다.
  - 관찰도수와 기댓도수의 차이가 미치는 영향은 다릅니다.
  - 각 범주별로 발생하는 차이의 크기는 각 기댓도수에 대한 관찰도수와 기댓도수의 차이 제곱의 상대적 비중으로 측정합니다.
    - 경우 1에서는 그 크기가  $1/10000$ 이고 경우 2에서는 그 크기가  $1/2$ 입니다.

## ③ 기댓도수에 대한 관찰도수와 기댓도수의 차이 제곱의 상대적 비중

- 둥글고 황색 :  $\frac{(O_1 - E_1)^2}{E_1} = \frac{5.6025}{312.75} = 0.016187$
- 주름지고 황색 :  $\frac{(O_2 - E_2)^2}{E_2} = \frac{10.5625}{104.25} = 0.101319$
- 둥글고 녹색 :  $\frac{(O_3 - E_3)^2}{E_3} = \frac{14.0625}{104.25} = 0.134892$
- 주름지고 황색 :  $\frac{(O_4 - E_4)^2}{E_4} = \frac{7.5625}{34.75} = 0.217626$

# 범주형 자료분석

- 마지막으로 이렇게 구한 값들을 모두 더한 값이 검정통계량입니다.
- 이를 정리하면 다음 표와 같으며, 검정통계량은 자유도가 3인  $\chi^2$ -분포를 따릅니다.
  - 범주의 개수가 k인 경우 자유도가 (k-1)인  $\chi^2$ -분포를 따르고 우리의 예에서는 각 형질이 4가지로 4-1=3으로, 자유도가 3입니다.

구분	관찰도수	기대도수	$O_i - E_i$	$(O_i - E_i)^2 / E_i$
둥글고 황색	315	312.75	2.25	0.016187
주름지고 황색	101	104.25	-3.25	0.101319
둥글고 녹색	108	104.25	3.75	0.134892
주름지고 녹색	32	34.75	-2.75	0.217626
합	556	556		0.470024

# 범주형 자료분석

## 검정통계량

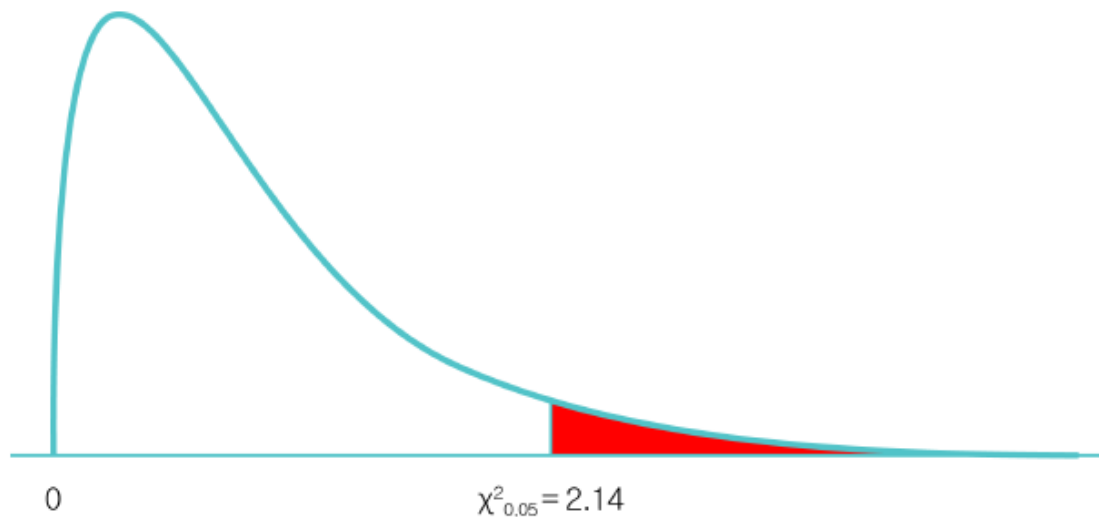
모집단을  $k$ 로 나누는 범주형 자료에 대해,  $i$ 번째 범주의 관찰도수를  $O_i$ , 기대도수를  $E_i$ 라고 하면, 다음의 검정통계량  $\chi_0^2$ 은 자유도가  $(k-1)$ 인  $\chi^2$ -분포를 따릅니다.

$$\chi_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1) \quad (8.1)$$

- 멘델의 자료로부터 구한 검정통계량은 자유도가 3인  $\chi^2$ -분포에서 0.470024로 나타났습니다.
- 이 값은 영가설 하에서 기대도수와 관찰도수의 차이로, 그 값이 크면 영가설 하의 기대도수가 참이 아닐 가능성이 커집니다.

# 범주형 자료분석

- **유의수준 0.05에서의 기각역과 유의확률**
  - ▣ 검정통계량이 작으면 작을수록 영가설이 참으로 받아들여지고, 크면 클수록 영가설을 참으로 받아들이기 어려워집니다.
    - 기각역은 오른쪽에 위치하는 한쪽검정입니다.
    - 자유도가 3인  $\chi^2$ -분포에서 유의수준을 0.05로 했을 때의 임계값은 약 2.14입니다.



[그림 8-1] 자유도가 3인  $\chi^2$ -분포에서 기각역( $\alpha = 0.05$ )

# 범주형 자료분석

- 자유도가 3인  $\chi^2$ -분포에서 검정통계량 0.40024의 유의확률을 R 코드를 이용하여 구하면 약 0.9254입니다.
  - $\chi^2$ -분포의 R에서의 함수이름은 “chisq”입니다.
  - 유의확률은 0.470024보다 클 확률로 1에서  $P(X < x)$ 를 구하는 함수를 나타내는 “p”를 분포함수 이름앞에 붙힌 “pchisq()”의 값을 빼서 구합니다.

```
> 1-pchisq(0.470024, df=3)  
[1] 0.9254259
```



# 범주형 자료분석

- R함수 `chisq.test()`를 이용한 검정

```
1. x <- c(315, 101, 108, 32)
2. chisq.test(x, p=c(9, 3, 3, 1)/16)
```

- 1줄 : 각 범주별 개수를 담고 있는 벡터를 만듭니다.
- 2줄 : `chisq.test()` 함수를 이용하여  $\chi^2$ -분포를 따르는 검정을 실시합니다.
  - 첫 전달인자로 각 범주의 개수를 담고 있는 벡터 혹은 테이블 자료를 전달합니다.
    - 예에서 각 범주의 개수를 담은 벡터를 사용했습니다.
    - 데이터셋으로부터 `table()`, `xtabs()` 와 같은 함수를 이용하여 table로 만든 자료를 전달하는 것이 조금 더 일반적일 것입니다.
  - p로 전달되는 값은 각 범주별 영가설 하의 확률이 저장된 벡터입니다.

# 범주형 자료분석

Chi-squared test for given probabilities

data: x

**X-squared = 0.47002, df = 3, p-value = 0.9254**

- 앞서 직접 구한 값과 동일함을 알 수 있습니다.


## • 판정

- 기각역을 이용한 판정
  - 검정통계량은 0.47002로 채택역에 속하므로 영가설을 채택합니다.
- 유의확률을 이용한 판정
  - 검정통계량으로부터 유의확률  $P(X^2 > \chi^2)$ 은 약 0.9254로 유의수준 0.05보다 크므로 영가설을 채택합니다.

# 범주형 자료분석

- **적합도 검정**

- 앞서 우리는 멘델의 자료를 이용하여 잘 알려진 비율인 9:3:3:1을 따르는지 살펴보았습니다.
- 이와 같이 자연현상이나 각종 실험을 통해 관찰되는 도수들이 영가설 하의 분포(범주형 자료의 각 수준별 비율)에 얼마나 일치하는지의 적합성에 대한 검정을 “적합도 검정”이라 합니다.



## 02. 동질성 검정과 독립성 검정

- 두 범주형 자료를 요약하는 분할표에 대해 학습한다.
- 동질성 검정과 독립성 검정으로 가설검정하는 방법에 대해 학습한다.

## $r \times c$ 분할표 : 두 범주형 자료의 요약

### • $r \times c$ 분할표

- $r$ 개의 행과  $c$ 개의 열로 구성된 표 구조로 같은 두 가지로 구성할 수 있습니다.
- $r$ 개의 하위 모집단에 대한  $c$ 개의 범주별 분류표
  - 모집단  $\Omega$ 는  $r$ 개의 하위 모집단  $\Omega_1, \Omega_2, \dots, \Omega_r$ 로 구성되고, 임의의 속성(변수)  $k$ 는  $c$ 개의 범주로 구성되어 있다고 합시다.
  - 이때 하위 모집단별로 속성  $k$ 에 의해  $c$ 개의 범주로 분류될 때, 각 하위 모집단별로 속성  $k$ 에 의한  $c$ 개의 범주별 응답수를 표로 구성합니다.
  - 행을 구성하는 하위 모집단별로 추출한 확률표본들로 구성되어 있는 경우로 행별 합은 각각의 표본의 크기로 고정되어 있습니다.
  - 예) 연령대별 SNS 사용 현황
    - 연령대별로 가장 많이 이용하는 SNS 서비스를 조사한 내용을 분할표로 정리해 보았습니다.
    - 연령대에 따라 하위 모집단을 세 그룹으로 나누고, 각 연령대별로 SNS 서비스 사 5개의 이용자 수를 기록하였습니다.

$r \times c$  분할표 : 두 범주형 자료의 요약

구분	F 사	T 사	K 사	C 사	기타	합
20대	207	117	111	81	16	532
30대	107	104	236	109	15	571
40대	78	76	133	32	17	336
합	392	297	480	222	48	1,439

## $r \times c$ 분할표 : 두 범주형 자료의 요약

- ▣ 두 개의 범주형 변수에 대한 교차표
  - 단일 모집단에서 표본을 추출하여 두 개의 서로 다른 범주형 변수의 교차표를 구합니다.
  - 임의의 변수  $k_1$ 은  $r$ 대의 범주를, 임의의 변수  $k_2$ 는  $c$ 개의 범주를 가질 때 각 범주별 도수를 표로 구성하는 경우입니다.
  - 행과 열별로 합의 크기는 고정되어 있으며 각 값은 고정된 전체 합 내에서 추출된 확률표본입니다.
  - 예) 모 대학원의 성별 대학원 합격 여부
    - 모 대학원의 지원자 자료에서 성별로 합격 여부를 조사했습니다. 이 경우 전체 지원자 자료에 성별 변수와 합격 여부 변수가 있습니다.

구분	남학생	여학생	합
합격	1,198명	557명	1,755명
불합격	1,493명	1,278명	2,771명
합	2,691명	1,835명	4,526명

# 동질성 검정

## • 동질성 검정

- 모집단이 범주형 변수  $R$ 에 의해  $\Omega_1, \Omega_2, \dots, \Omega_r$ 의  $r$ 개의 하위 모집단으로 나누어진다고 할 때, 각 하위 모집단별로  $c$ 개의 범주를 갖는 범주형 변수  $k(k = \{k_1, k_2, \dots, k_c\})$ 에 대해 하위 모집단  $\Omega_1, \Omega_2, \dots, \Omega_r$ 이 속성(변수)  $k$ 의 각 범주별로 그 비율이 동일한지 검정하는 것을 “동질성 검정”이라고 합니다.

### 예제 2 연령대별 SNS 이용률의 동질성 검정

- 20대에서 40대까지 연령대별로 서로 조금씩 그 특성이 다른 SNS 서비스들에 대해 이용 현황을 조사한 자료를 바탕으로 연령대별로 홍보 전략을 세우고자 합니다.
- 연령대별로 이용현황이 서로 동일한지 검정해 봅시다.



# 동질성 검정

구분	F 사	T 사	K 사	C 사	기타	합
20대	207	117	111	81	16	532
30대	107	104	236	109	15	571
40대	78	76	133	32	17	336
합	392	297	480	222	48	1,439

## 가설수립

- ▣ 영가설 : 연령대별로 SNS 서비스별 이용 현황은 동일합니다.
- ▣ 대안가설 : 연령대별로 SNS 서비스별 이용 현황은 동일하지 않습니다.
- ▣ 영가설이 나타내는 수식 표현을 살펴 봅시다.
  - 연령대 변수를 R, SNS 서비스 제공사 변수를 C라 할 때,
  - 분할표 상에 관찰수를  $n_{ij}$ 라 하고, 연령대별 서비스 이용률은  $p_{ij}$ 라고 하면 다음과 같이 나타낼 수 있을 것입니다.

# 동질성 검정

$R \backslash C$	F 사	T 사	K 사	C 사	기타	합
20대	$n_{11}(p_{11})$	$n_{12}(p_{12})$	$n_{13}(p_{13})$	$n_{14}(p_{14})$	$n_{15}(p_{15})$	$n_{1.}(p_{1.} = 1)$
30대	$n_{21}(p_{21})$	$n_{22}(p_{22})$	$n_{23}(p_{23})$	$n_{24}(p_{24})$	$n_{25}(p_{25})$	$n_{2.}(p_{2.} = 1)$
40대	$n_{31}(p_{31})$	$n_{32}(p_{32})$	$n_{33}(p_{33})$	$n_{34}(p_{34})$	$n_{35}(p_{35})$	$n_{3.}(p_{3.} = 1)$
합	$n_{.1}(p_{.1})$	$n_{.2}(p_{.2})$	$n_{.3}(p_{.3})$	$n_{.4}(p_{.4})$	$n_{.5}(p_{.5})$	$n_{..} = n$

- 영가설이 나타내는 것이 연령대별로 j번째 서비스 이용률은 서로 같다는 것이고, 이는 다음과 같이 나타낼 수 있습니다.

$$H_0 : p_{.j} = p_{1j} = p_{2j} = p_{3j}, \quad j = 1, 2, 3, \dots, c, p_{.j} = \frac{n_{.j}}{n}$$

- 대안가설은 영가설이 아닌 경우로 이는 적어도 한 행은 서비스 이용률의 분포가 다름을 나타냅니다

# 동질성 검정

## • (동질성 검정에서의)검정통계량

- 앞서 적합도 검정에서 알아본 기대도수와 관찰도수의 차이를 마찬가지로 구합니다.
- 먼저 영가설이 참이라는 가정 하에 기대도수를 구해보시다.
  - 예) 20대의 F 사 서비스 이용자 수의 기대도수( $E_{11}$ )는 다음과 같이 구합니다.
    - 20대 전체 인원 532명 중에 20대의 F 사 서비스 이용률을 곱하여 구합니다.
      - 영가설이 참이라는 가정 하에 20대의 F 사 서비스 이용률은 전체의 F 사 서비스 이용률과 같습니다.
      - 조사 대상 전체의 F 사 서비스 이용률은 전체 인원수 중 F 사를 이용한다고 응답한 인원수로  $\frac{n_{.1}}{n}$ 으로 계산할 수 있습니다.
      - 이로부터 20대 전체 인원 중에 20대의 F 사 서비스 이용자 수의 기대도수는

$$E_{11} = 20\text{대 전체 인원수} \times F\text{사 이용률} = n_{1.} \times \frac{n_{.1}}{n} = 532 \times \frac{392}{1439} \cong 144.92$$

# 동질성 검정

- ▣ 모든 셀의 기대도수를 R을 이용해 구해봅시다
  - [8장을 위한 준비]에서 사용한 sns.c 이용

```

12. c.tab <- table(sns.c$age.c, sns.c$service.c)
13. (a.n <- margin.table(c.tab, margin=1))
14. (s.n <- margin.table(c.tab, margin=2))
15. (s.p <- s.n / margin.table(c.tab))
16. (expected <- a.n %*% t(s.p))

```

- 12줄 : 연령대를 행으로, 서비스 제공사를 열로 하는 분할표를 생성합니다.
- 13줄 : 조사대상이 된 연령대별 사용자 수  $n_{i.}$  ( $i = 1, 2, 3$ )를 구해 a.n에 저장하고 출력합니다
  - margin.table() 함수에 margin=1로 하여 행별 합을 구합니다.
  - rowSums() 함수를 이용하여 동일한 결과(행별 합)를 얻을 수 있으며 margin.table() 함수의 결과와 다른 점은 결과값의 형태로 margin.table() 함수는 table로 rowSums() 함수는 이름있는 벡터로 그 결과를 돌려줍니다.

# 동질성 검정

```
> (a.n <- margin.table(c.tab, margin=1))
```

```
20대 30대 40대
532  571  336
```

```
> rowSums(c.tab)
```

```
20대 30대 40대
532  571  336
```

- 전체 인원 중 서비스별 이용자 수의 비율을 구합니다
  - 14줄 : 조사대상으로부터 서비스별 이용자 수  $n_j (j = 1, 2, 3, 4, 5)$  를 구해 s.n에 저장하고 출력합니다.
    - margin.table() 함수에 margin=2로 하여 열별 합을 구합니다
    - colSums() 함수를 이용하여 구할수도 있습니다.
  - 15줄 : 서비스별 이용자 수를 전체 응답자 수로 나눠 서비스별 이용자 수의 비율을 s.p에 저장하고 출력합니다.
    - margin.table() 에 margin 값을 지정하지 않으면 전체 합을 구합니다.

# 동질성 검정

```
> (s.n <- margin.table(c.tab, margin=2))
  F    T    K    C    E
392 297 480 222  48

> (s.p <- s.n / margin.table(c.tab))
           F           T           K           C           E
0.2724114 0.2063933 0.3335650 0.1542738 0.0333565
```

- 기대도수를 구합니다.
  - R의 연산자 '%\*%'는 행렬의 곱을 구하는 연산자입니다.
  - R에서 a.n이나 s.p 같이 단일 factor형 변수의 각 수준별 수를 나타내는 table은 원소의 개수로 factor의 수준 수만큼을 갖는 (열)벡터처럼 사용할 수 있습니다.
    - rowSums()와 colSums()는 벡터를 반환하므로 조금 더 명확할 것입니다.
  - R의 t() 함수는 행과 열의 위치를 바꾸는 전치행렬을 구해줍니다.
  - 16줄 : 열로 구성된 s.p를 전치하여 행 벡터를 구하고, 이렇게 구한 행 벡터를 열로 구성된 a.n과 행렬 곱셈을 하여  $r \times c$  행렬을 만듭니다. 이렇게 생성된  $r \times c$  행렬을 변수 expected에 저장하고 출력합니다

# 동질성 검정

```
> (expected <- a.n %*% t(s.p))
```

	F	T	K	C	E
20대	144.92286	109.80125	177.4566	82.07366	17.74566
30대	155.54691	117.85059	190.4656	88.09034	19.04656
40대	91.53023	69.34816	112.0778	51.83600	11.20778

```
> round(expected, 2)
```

	F	T	K	C	E
20대	144.92	109.80	177.46	82.07	17.75
30대	155.55	117.85	190.47	88.09	19.05
40대	91.53	69.35	112.08	51.84	11.21

- round() 함수를 이용하여 소수점 두째자리까지 표시하도록 하였습니다.
- 이를 정리하면 다음 표와 같습니다.

# 동질성 검정

서비스 연령대	F 사	T 사	K 사	C 사	기타	합
20대	144.92	109.80	177.46	82.07	17.75	532
30대	155.55	117.85	190.47	88.09	19.05	571
40대	91.53	69.35	112.08	51.84	11.21	337
합	392	297	480	222	48	1439

- 검정통계량은 적합도 검정에서와 동일하게 기대도수를 이용하여 20대에 대해서 관찰도수와 차이 제곱을 기대도수로 나눈 값들을 모두 더하고, 동일한 방법으로 다른 하위 모집단인 30대와 40대에 대해서도 값을 구하며, 이는 다음과 같은 분포를 따릅니다.

동질성 검정에서의 검정통계량은 자유도가 ‘(행의 개수-1)×(열의 개수-1)’인 자유도를 갖는  $\chi^2$ -분포를 따릅니다.

$$\chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1) \times (c-1)) \quad (8.2)$$



# 동질성 검정

구분		F 사	T 사	K 사	C 사	기타	합
20대	$O_{1j}$	207	117	111	81	16	532
	$E_{1j}$	144.92	109.80	177.46	82.07	17.75	532
	$\frac{(O_{1j} - E_{1j})^2}{E_{1j}}$	26.59	0.47	24.89	0.01	0.17	52.14
30대	$O_{2j}$	107	104	236	109	15	571
	$E_{2j}$	155.55	117.85	190.47	88.09	19.05	571
	$\frac{(O_{2j} - E_{2j})^2}{E_{2j}}$	15.15	1.63	10.89	4.96	0.86	33.49
40대	$O_{3j}$	78	76	133	32	17	336
	$E_{3j}$	91.53	69.35	112.08	51.84	11.21	336
	$\frac{(O_{3j} - E_{3j})^2}{E_{3j}}$	2.00	0.64	3.91	7.59	2.99	17.13

# 동질성 검정

## 검정통계량 계산하기

```
18.( o.e <- c.tab-expected )
19.( t.t <- sum( (o.e)^2 / expected ) )
```

- 18줄 : '관찰도수 table - 기대도수 table'의 연산에서 사용한 뺄셈 연산자('-')는 서로 다른 table의 행과 열별로 일치하는 셀 간의 차이를 구합니다. 이를 o.e에 저장하고 출력합니다.
- 19줄 : 검정통계량은 '관찰도수-기대도수' 값의 제곱을 기대도수로 나눈 모든 값을 합하는 것으로 이를 구해 변수 t.t에 저장하고 출력합니다

```
> (o.e <- c.tab-expected)
      F      T      K      C      E
20대 62.077137  7.198749 -66.456567 -1.073662 -1.745657
30대 -48.546908 -13.850591 45.534399 20.909659 -4.046560
40대 -13.530229  6.651842 20.922168 -19.835997  5.792217
> (t.t <- sum( (o.e)^2 / expected ) )
[1] 102.752
```

# 동질성 검정

- 유의수준 0.05에서의 기각역과 검정통계량으로부터 구한 유의확률
  - 행 수준의 수는 3이고, 열 수준의 수는 3이므로 검정통계량은 자유도가  $(3-1) \times (3-1)$ 인 4입니다.
  - 자유도가 4인  $\chi^2$ -분포에서 (오른쪽) 한쪽 검정의 임계값은 약 9.488 입니다.
  - `qchisq(0.95, df=4)`
  - 유의확률은 다음과 같으며 0으로 나왔습니다. (값이 너무 작아 0으로 표시)

```
> 1-pchisq(t.t, df=4)
[1] 0
```

# 동질성 검정

- R 함수 `chisq.test()`를 이용한 검정

```
> chisq.test(c.tab)
```

Pearson's Chi-squared test

data: c.tab

**X-squared = 102.75, df = 8, p-value < 2.2e-16**

- `chisq.test()` 함수에 전달하는 값은 관찰도수를 담고 있는 table 자료이며, 검정의 결과로 검정통계량 102.75, 유의확률은  $2.2e-16$ 보다 작은 것으로 표시됩니다.
- $2.2e-16$ 은  $2.2 \times 10^{-16}$ 에 대한 과학식 표현으로 0에 상당히 가까운 수로 R이 구분해 낼 수 없을 정도로 작은 값을 나타냅니다.

# 동질성 검정

## • 판정

- 앞서와 마찬가지로 검정통계량의 기각역 위치 여부 혹은 유의확률과 유의수준 비교를 통해 판정을 내릴 수 있습니다.
- 유의확률과 유의수준을 비교해보면, 유의확률은 거의 0으로 유의수준보다 작아 영가설을 기각합니다.

## • 결론

- 연령대별로 SNS 서비스 이용현황이 동일한지 1,439명으로부터 설문조사한 결과
- 검정통계량 102.75, 유의확률은  $< 0.000$ 으로 유의수준 0.05에서 연령대별로 SNS 서비스별 이용현황이 동일하다고 볼 수 없는 것으로 나타났습니다.
- 즉, 연령대에 따라 주로 사용하는 SNS 서비스는 서로 다른 것으로 볼 수 있습니다.

# 동질성 검정

## • x.test()에 숨겨진 값들

- 앞서 사용한 t.test() 혹은 chisq.test()를 실행하면 검정통계량, 자유도, 유의확률을 등을 볼 수 있습니다.
- 검정을 위해 사용하는 함수들은 보다 많은 정보를 담고 있습니다.
  - 다음은 chisq.test() 함수 실행 후 전달해 주는 정보입니다.
  - 검정을 위해 중간 계산에 사용하는 각종 결과물과 최종 결과물 출력을 위해 계산된 정보들이 들어가 있습니다.

```
> ct.info <- chisq.test(c.tab)
> names(ct.info)
[1] "statistic" "parameter" "p.value"    "method"    "data.name"
[6] "observed"  "expected"  "residuals" "stdres"
```

# 동질성 검정

- 기대도수를 구해 봅시다.
  - 기대도수는 `expected`에 저장되어 있습니다.

```
> addmargins(ct.info$expected)
```

	F	T	K	C	E	Sum
20대	144.92286	109.80125	177.4566	82.07366	17.74566	532
30대	155.54691	117.85059	190.4656	88.09034	19.04656	571
40대	91.53023	69.34816	112.0778	51.83600	11.20778	336
Sum	392.00000	297.00000	480.0000	222.00000	48.00000	1439

- `chisq.test()` 결과를 변수로 저장한 경우 다음과 같이 변수 이름을 입력하면 앞서 살펴본 `chisq.test()` 함수의 결과 화면을 볼 수 있습니다.

```
> ct.info
```

```

      Pearson's Chi-squared test
data:  c.tab
X-squared = 102.75, df = 8, p-value < 2.2e-16

```

# 독립성 검정

- 독립성 검정
  - 어떤 모집단에서 관찰한 두 개의 속성 R과 C가 범주형 변수일 때 두 변수가 서로 연관이 있는지를 검정하는 것을 독립성 검정이라고 합니다.
  - 독립성 검정은 독립인 두 사건의 곱사건을 이용하는 검정으로, 앞서 3장에서 학습한 두 변수가 서로 독립이라면 두 사건의 곱사건의 확률이  $P(R \cap C) = P(R) \cdot P(C)$ 가 됨을 이용하여 검정합니다.

## 예제 3 성별에 따른 대학원 입학 여부의 독립성 검정

어느 대학원 입시에서 입학 과정 중 여성에 대한 성차별이 있었다는 내용의 소송이 발생했습니다. 고소인은 근거 자료로 성별에 따라 합격에 영향을 미쳤음을 주장했는데, 그 근거를 다같이 살펴 봅시다.



# 독립성 검정

## • R의 내장자료 UCBAdmissions

- 1973년 미국 버클리 대학원에서 실제 있었던 사건으로 모집 인원이 가장 많은 상위 6개 학과의 성별 합격 여부를 저장한 자료입니다.

구분	남성	여성	합
합격	1,198명 (44.5%)	557명 (30.4%)	1,755명 (59.5%)
불합격	1,493명 (55.5%)	1,278명 (69.4%)	2,771명 (40.5%)
합	2,691명 (61.2%)	1,835명 (38.8%)	4,526명

- 이 자료를 바탕으로 성별에 따른 합격 여부가 연관이 있는지 검정을 실시해봅시다.

# 독립성 검정

## 가설 수립

- ▣ 영가설 : 성별과 합격 여부는 관련이 없습니다(서로 독립입니다).
- ▣ 대안가설 : 성별과 합격 여부는 관련이 있습니다(서로 연관이 있습니다).
- ▣ 합격 여부 변수를 R, 성별 변수를 C라 할 때, 분할표 상의 각 셀의 관찰수를  $n_{ij}$ , 확률을  $p_{ij}$ 로 나타내면,
- ▣ 합격 여부와 성별 차이가 관련이 없다는 영가설은 두 변수가 독립임을 나타내고, 두 사건이 독립일 때  $P(R \cap C) = P(R) \cdot P(C)$ 로 계산됨을 이용하여

$$H_0 : p_{ij} = p_{i.} \times p_{.j}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

- ▣ 대안가설은 영가설이 아닌 즉  $H_1: not H_0$ 로 표현하고, 이는 두 변수가 서로 연관이 있어(독립이 아닌것으로) 곱사건의 확률을 각 사건의 확률끼리의 곱으로 나타낼 수 없음을 뜻합니다.

# 독립성 검정

## • 검정통계량

- ▣ 동질성 검정과 마찬가지로 영가설이 참이란 가정 하에 기댓도수를 구합니다.
- ▣ 여성이면서 합격할 경우의 기댓도수( $E_{12}$ )를 구하는 과정입니다.

① 전체 인원 중에서 여성이면서 합격하는 사람들의 비율을 구합니다.

② 여성이면서 합격한 사람들의 비율( $p_{12}$ )은 영가설 하에서  $p_{1.} \times p_{.2}$ 입니다.

•  $p_{1.}$ 은 합격할 확률이며 전체 조사 대상 중 합격자의 비율입니다.  $\rightarrow p_{1.} = \frac{n_{1.}}{n_{..}}$

•  $p_{.2}$ 는 여성일 확률로 전체 조사 대상 중 여성의 비율입니다.  $\rightarrow p_{.2} = \frac{n_{.2}}{n_{..}}$

• 이로부터 영가설하에서 여성이면서 합격한 비율은  $p_{12} = \frac{n_{1.}}{n_{..}} \times \frac{n_{.2}}{n_{..}}$  입니다.

③ 전체 조사 대상수( $n_{..}$ )를 ②에서 구한  $p_{12}$ 를 곱해, 여성이면서 합격한 경우의 기댓도수  $E_{12}$ 를 구합니다.

$$E_{12} = n_{..} \times p_{12} = n_{..} \times \frac{n_{1.}}{n_{..}} \times \frac{n_{.2}}{n_{..}} = \frac{n_{1.} \times n_{.2}}{n_{..}} = \frac{1755 \times 1835}{4526} \cong 711.54$$

# 독립성 검정

성별 \ 합격 여부	남성	여성	합
합격	1,043.46	711.54	1,755
불합격	1,647.54	1,123.46	2,771
합	2,691	1,835	4,526

## ▣ R을 이용하여 기대도수 구하기

- 첨부 파일의 04.independece.R에서 R이 기본으로 제공하는 UCBA admissions를 이용하여 합격 여부와 성별 테이블을 만들어 ucba.tab에 저장하였습니다.

```

7. (a.n <- margin.table(ucba.tab, margin=1))
8. (g.n <- margin.table(ucba.tab, margin=2))
9.
10.(a.p <- a.n / margin.table(ucba.tab))
11.(g.p <- g.n / margin.table(ucba.tab))
12.
13.(expected <- margin.table(ucba.tab) * (a.p %*% t(g.p)))
14.addmargins( expected )

```

# 독립성 검정

- ▣ Step #1) 성별, 합격 여부의 각 수준(여자, 남자, 합격, 불합격)별 합 구하기
  - 7줄 : 합격 여부별 수준인 합격자와 불합격자의 합( $n_{i.}$ )을 구하기 위해 전달인자 margin을 1(행별 합)로 하는 margin.table()을 사용하고, 그 값을 a.n에 저장하고 출력합니다.
  - 8줄 : 성별 수준인 남성과 여성의 수의 합( $n_{.j}$ )을 구하기 위해 전달인자 margin을 2(열별 합)로 하는 margin.table()을 사용하고, 그 값을 g.n에 저장하고 출력합니다.

```
> (a.n <- margin.table(ucba.tab, margin=1))
Admit
Admitted Rejected
      1755      2771
> (g.n <- margin.table(ucba.tab, margin=2))
Gender
Male Female
  2691   1835
```

# 독립성 검정

- Step #2) 성별, 합격 여부의 각 수준(여자, 남자, 합격, 불합격)별 비율을 구합니다
  - 10줄 : 전체 합격자와 불합격자의 비율( $p_{i.}$ )을 구하기 위해 `margin.table()`로 얻는 전체지원자 수( $n_{..}$ )로 합격 여부별 지원자 수가 저장된 변수 `a.n( $n_{i.}$ )`을 나누고( $n_{i.} / n_{..}$ ), 이를 `a.p`에 저장하고 출력합니다.
  - 11줄 : 여성과 남성의 비율( $p_{.j}$ )을 구하기 위해 `margin.table()`로 얻는 전체지원자 수( $n_{..}$ )로 성별 지원자 수가 저장된 변수 `g.n( $n_{.j}$ )`을 나누고( $n_{.j} / n_{..}$ ), 이를 `g.p`에 저장하고 출력합니다.

```
> (a.p <- a.n / margin.table(ucba.tab))
Admit
  Admitted  Rejected
0.3877596 0.6122404
> (g.p <- g.n / margin.table(ucba.tab))
Gender
      Male      Female
0.5945647 0.4054353
```

# 독립성 검정

- Step #3) 기대빈도는 '전체 지원자 수 × 합격 여부 비율 × 성별 비율'로 구합니다
- 13줄 : 동질성 검정에서와 같이 '합격 여부 비율 × 성별 비율'을 '2행 × 2열'짜리 행렬로 만들기 위해 행렬 곱셈(%\*%)을 이용하여 계산하고, 전체 지원자 수를 곱해 변수 expected에 저장하고 출력합니다.
- 14줄 : 행과 열의 합까지 함께 만들어 출력해 봅시다.

```
> (expected <- margin.table(ucba.tab) * (a.p %*% t(g.p)))
```

Gender

Admit	Male	Female
Admitted	1043.461	711.5389
Rejected	1647.539	1123.4611

```
> addmargins( expected )
```

Gender

Admit	Male	Female	Sum
Admitted	1043.461	711.5389	1755
Rejected	1647.539	1123.4611	2771
Sum	2691.000	1835.0000	4526

# 독립성 검정

- 동질성 검정과 동일하게 '관찰도수와 기대도수의 차이 제곱'을 기대도수로 나눈 값들을 모두 더하여 검정통계량을 구합니다.

구분		남성	여성	합
합격	$O_{1j}$	1,198	557	1,755
	$E_{1j}$	1,043.46	711.54	1,755
	$\frac{(O_{1j} - E_{1j})^2}{E_{1j}}$	22.89	33.56	56.45
불합격	$O_{2j}$	1,493	1,278	2,771
	$E_{2j}$	1,647.54	1,123.46	2,771
	$\frac{(O_{2j} - E_{2j})^2}{E_{2j}}$	14.50	21.26	35.76
합		2,691	1,835	4,526

- 각 범주별로 구한 기대빈도를 모두 더하면 우리가 얻고자 하는 검정통계량이 됩니다. 예에서는  $22.89 + 33.56 + 14.50 + 21.26 \approx 92.21$ 입니다



# 독립성 검정

독립성 검정에서의 검정통계량은 자유도가 ‘(행의 개수 - 1) × (열의 개수 - 1)’인 자유도를 갖는  $\chi^2$ -분포를 따릅니다.

$$\chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1) \times (c-1)) \quad (8.3)$$

- ▣ 독립성 검정에서의 검정통계량은 위와 같으며 이는 동질성 검정의 검정통계량과 동일합니다.
- ▣ 동질성 검정은 하위 모집단 사이 특정 변수에 대한 분포의 동질성을 검정하는 것이고, 독립성 검정은 두 변수 사이의 연관성을 검정하는 것입니다.
  - 이는 분명히 다른 검정이지만 분할표의 형태를 하고 있으며, 기대도수를 구하는 과정이 가설의 차이에 따라 계산의 중간과정만 다를 뿐 최종 값은 동일하게 계산되는 검정통계량을 사용합니다.
  - 그러나 검정통계량을 구하는 과정만 같을 뿐 검정을 시작하는 배경이 다르기에 본인이 원하는 검정이 동질성 검정인지, 독립성 검정인지 잘 판단해야 합니다.

# 독립성 검정

- 유의수준 0.05에서의 기각역과 검정통계량으로부터 구한 유의확률
  - 행 수준의 수는 2, 열 수준의 수는 2이므로 검정통계량은 자유도가  $(2-1) \times (2-1)$  인 1입니다.
  - 자유도가 1인  $\chi^2$ - 분포에서 (오른쪽) 한쪽검정의 임계값은 약 3.841 로 기각역은  $\chi^2 > 3.841$  인 영역입니다.
  - 검정통계량 92.21에 대한 자유도가 1인  $\chi^2$ -분포에서의 유의확률은 거의 0에 가까운 값으로 나타납니다

```
> 1-pchisq(92.21, df=1)  
[1] 0
```

# 독립성 검정

- R의 `chisq.test()`를 이용한 검정
  - 동질성 검정과 마찬가지로 `chisq.test()`를 사용합니다.

```
> chisq.test(ucba.tab)
```

Pearson's Chi-squared test with **Yates' continuity correction**

```
data: ucba.tab
```

```
X-squared = 91.61, df = 1, p-value < 2.2e-16
```

- `chisq.test()` 함수에 전달하는 전달인자는 관찰도수를 담고 있는 table 자료이며, 검정의 결과로 검정통계량 91.61 유의확률은 2.2e-16보다 작은, 즉 0에 아주 가까운 값으로 나타났습니다.
- 우리가 구한 검정통계량 92.21과 미세하게 차이납니다.
- `chisq.test()`는 2×2 분할표에 대해서는 '연속성 수정'을 통해 카이제곱 통계량을 구하는 것이 기본으로 되어 있습니다.

# 독립성 검정

- 연속성 수정을 통한 검정통계량은 다음과 같습니다.

$$corrected \chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}} \sim \chi^2((r-1) \times (c-1))$$

- 이를 R을 통해 직접 구해보시다.

```
> o.e2 <- (abs(ucba.tab - expected) - 0.5)^2 / expected
> margin.table(o.e2)
[1] 91.6096
```

- abs() 함수는 절대값을 구하는 함수이며 관찰빈도와 기대빈도의 차이에 절대값을 구해 0.5를 뺀 값을 제공해서 기대빈도로 나누고 이를 합하여 연속성 수정을 통한 검정통계량을 구합니다.

# 독립성 검정

- 연속성 수정을 하지 않고 값을 구하려면 다음처럼 `chisq.test()` 함수에 `correct=FALSE` 를 넣어 구합니다.

```
> chisq.test(ucba.tab, correct=FALSE)
```

Pearson's Chi-squared test

data: ucba.tab

X-squared = **92.205**, df = 1, p-value < 2.2e-16

- 우리가 앞서 구한 값과 동일한 결과를 얻었습니다.

## 판정

- 검정통계량의 기각역 위치 여부는 검정통계량이 92.21로 임계값(3.841)을 벗어난 기각역에 있으므로 영가설을 기각합니다.
- 또한 유의확률과 유의수준을 비교한다면 유의확률은 거의 0으로 유의수준(0.05)보다 작아 영가설을 기각합니다.

# 독립성 검정


## ▣ 결론

- 성별과 입학 여부 간에 연관이 있는지를(즉, 입학에 있어 성차별이 있었는지를) 알아보기 위해 대학원 입학 정원의 상위 6개 학과에 지원한 4,526명에 대해 성별, 입학 여부별로 분류하여 독립성 검정을 실시한 결과, 검정통계량은 92.21, 유의확률은 거의 0에 가까운 것으로 나타나 성별과 입학 여부는 연관이 있다는 통계적으로 유의한 결론을 얻었습니다. 즉 대학원 입학에서 입학여부와 성별은 연관이 있었던 것으로 판단됩니다.



# 9장을 위한 준비

: 아버지와 아들의 키



# Q & A



수고하셨습니다.