제14주: 테이블과 범주형 자료 분석

1. cut(): 데이터를 구간으로 나누기

- 양적 변수를 구간으로 나누어 범주형 변수로 변환하고자 한다.

예제: 아래는 30명의 신생아의 체중 자료이다 (단위: lb). 이 자료를 “6미만”, “6이상~7미만”, “7이상~8미만”, “8이상~9미만”, “9이상”의 5가지 범주를 갖는 자료로 변환하시오.

```
7.2, 7.8, 6.8, 6.2, 8.2, 8.0, 8.2, 5.6, 8.6, 7.1, 8.2, 7.7, 7.5, 7.2, 7.7, 5.8,
6.8, 6.8, 8.5, 7.5, 6.1, 7.9, 9.4, 9.0, 7.8, 8.5, 9.0,7.7, 6.7, 7.7
```

cut()함수 사용법

cut(x, breaks, labels = NULL, right = TRUE, ...)

- x: 수치형 벡터. cut을 통해 팩터 형으로 변환됨.
- breaks: (1) 구간의 개수를 나타내는 한 개의 숫자, 또는 (2) cut지점을 나타내는 2개 이상의 고유한 숫자값들 벡터
- labels: cut 수행 결과 얻게되는 범주 level들에 대한 label들
- right: 각 구간이 오른쪽으로 닫힌 구간이면 TRUE, 아니면 FALSE

예시 1. 1부터 10까지의 수 (10, 1, 2, ..., 9)를 (0, 5], (5, 10] 두 개의 구간으로 나누는 다음 예를 보자.

```
cut(c(10, 1:9), breaks = c(0, 5, 10))
 [1] (5,10] (0,5]  (0,5]  (0,5]  (0,5]  (0,5]  (5,10] (5,10] (5,10] (5,10]
Levels: (0,5] (5,10]
## 여기서, (0, 5]는 0<x≤5 의 형태이므로 breaks의 값을 c(1,5,10)으로 지정
## 해서는 안된다. 이 경우 10이 어떤 구간에도 속하지 않기 때문이다.
```

예시 2. 3개의 구간으로 나누는 예

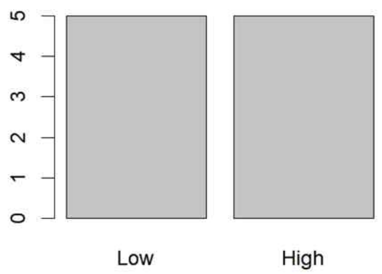
```
cut(c(10, 1:9), breaks = 3)
 [1] (7,10]  (0.991,4]  (0.991,4]  (0.991,4]  (0.991,4]  (4,7]  (4,7] (4,7]
 [9] (7,10]  (7,10]
Levels: (0.991,4] (4,7] (7,10)
## 이처럼 구간의 수를 지정하면 동일한 너비의 구간이 자동으로 구해진다.
```

예시 3. 각 구간의 하한값을 폐구간으로 하여 나누는 예

```
cut(c(10, 1:9), breaks = c(0, 5, 10), right = FALSE)
[1] <NA>    [0,5)   [0,5)   [0,5)   [0,5)   [5,10) [5,10) [5,10) [5,10) [5,10)
## 10 이 어떤 구간에도 속하지 않으므로 NA로 설정됨
cut(c(10, 1:9), breaks = c(1, 6, 11), right = FALSE)
[1] [6,11) [1,6)   [1,6)   [1,6)   [1,6)   [1,6)   [6,11) [6,11) [6,11) [6,11)
```

예시 4. 두 개의 구간으로 나누고 각 구간(범주 수준)에 label을 “Low” 및 “High”로 설정하는 예

```
cut(c(10, 1:9), breaks = c(0, 5, 10), labels = c("Low", "High"))
[1] High Low  Low  Low  Low  Low  High High High High
Levels: Low High
x <- cut(c(10, 1:9),
   breaks = c(0, 5, 10),
   labels = c("Low", "High"))
barplot(table(x))
```



분할표(Contingency Table)

분할표는 명목형 또는 순서형 자료의 도수를 표 형태로 기록한 것이다.

1. 분할표 만들기: table()

분할표를 작성하는 기본 함수는 table()이다.

```
## 주어진 한 개의 범주형 벡터에서 a, b, c의 출현 횟수를 세는 간단한 예
table(c("a","b","b","b","c","c","d"))
a b c d
1 3 2 1
```

범주형 변수의 개수가 2개 일 경우에는 행렬형태의 분할표를 생성

```
## warpbreaks 자료는 wool의 종류와 장력(L, M, H)에 따른 breaks의 수를
## 저장한 자료이다.
head(warpbreaks)
```

```
  breaks wool tension
1     26    A      L
2     30    A      L
3     54    A      L
4     25    A      L
5     70    A      L
6     52    A      L

## warpbreaks 자료에 대하여, wool과 tension 별로 몇 개의 자료가
## 관측되었는지 표로 출력한다.
with(warpbreaks, table(wool, tension))
    tension
wool L M H
   A 9 9 9
   B 9 9 9
```

2. 분할표로부터 각 칸(cell)의 비율 계산: prop.table()

prop.table()은 분할표로부터 각 칸의 비율을 계산한다.

```
## 한 개의 범주형 변수 벡터인 경우 상대도수 분포표 계산
prop.table(table(c("a","b","b","b","c","c","d")))
        a         b         c         d
0.1428571 0.4285714 0.2857143 0.1428571

## 두 개의 범주형 변수 벡터인 경우 상대도수 분포표 구하기
prop.table(with(warpbreaks, table(wool, tension)))
    tension
wool         L         M         H
   A 0.1666667 0.1666667 0.1666667
   B 0.1666667 0.1666667 0.1666667

## 두 개의 범주형 변수 벡터인 경우, 행 또는 열 별 조건부 상대도수 분포표
prop.table(with(warpbreaks, table(wool, tension)), 1)
    tension
wool         L         M         H
   A 0.3333333 0.3333333 0.3333333
```

```
   B 0.3333333 0.3333333 0.3333333
prop.table(with(warpbreaks, table(wool, tension)), 2)
    tension
wool    L    M    H
   A 0.5 0.5 0.5
   B 0.5 0.5 0.5
```

3. 각 칸의 도수를 알고 있는 경우 분할표 만들기: xtabs()

분할표를 만드는 데에 table()도 좋지만 xtabs() 함수가 더 유연하다. xtabs()는 각 칸의 도수를 알고 있는 경우에도 사용할 수 있다. 사용법은 "xtabs(도수변수 ~ 행변수 + 열변수, 데이터)"의 형태로 사용한다.

```
## 기본 예: x, y라는 두 가지 범주형 변수가 있고, (x,y)에 대한 도수가 num에
## 저장되어 있을 때
d <- data.frame( x = c("1","2","2","1"),
                 y = c("A","B","A","B"),
                 num = c(3,5,8,7))
d
xt <- xtabs(num ~ x + y, data = d)
xt
     y
x    A B
  1  3 7
  2  8 5
## num의 값이 분할표의 해당하는 각 칸 (x, y)에 입력이 되었음을 확인 가능

## 만약 도수를 나타내는 칼럼이 따로 없는 경우, 그리고 각 관찰 결과가 서로
## 다른 행에 표현되어 있다면 "~ 변수 + 변수 ..." 형태로 formula를 작성
d2 <- data.frame( x = c("A","A","A","B","B"),
                 result = c(5,1,4,7,6))
xtabs(~ x, d2)
x
A B
3 2
## x의 범주 수준 별로 도수가 분할표의 각 칸에 입력 되었음을 확인 가능
```

범주형 자료분석
범주형 변수를 통해 작성한 테이블 구조를 바탕으로 범주형 자료분석을 수행한다.

예를 들어, 설문조사나 여론조사를 수행하면 피조사자는 문항별로 여러 개의 선택사항 가운데 가장 적절한 한 개를 택하는 경우가 일반적이다. 이 경우 문항은 질적(qualitative)변수로서, 보통 선택사항 개수가 제한된 범주형(categorical) 변수이다. 범주형 변수는 명목형(nominal) 이거나 또는 순위형(ordinal) 일 수 있다.
(-> 12주차 강의 참고)

범주형 자료에 대해서는 기초적인 통계량(빈도수, 최빈값)들을 살펴볼 수도 있고 변수들 사이의 관계를 분석할 수도 있는데, 후자의 경우가 범주형 자료분석에 해당한다고 볼 수 있다. 예를 들어, 1000명에게 모병제에 대한 찬성 여부(찬성, 반대, 중립)을 조사하여 성별(남성, 여성)과 어떤 관계가 있는가, 또는 소득수준과 어떤 관계가 있는가 등에 대해 통계적 가설검정을 실시할 수 있다.

1. 이항분포와 다항분포

1-1. 이항분포 (Binomial Distribution).

표본 공간에서 표본 공간이 {성공, 실패}와 같이 단 두 개의 가능한 경우로만 구성되어있는 경우를 가정하자. 한 번의 시행의 결과가 성공이 나올 확률이 p이고 실패가 나올 확률이 $1-p$이라고 하고, 총 n번의 시행을 통해 얻게 되는 성공의 개수를 X라고 하자. 예를 들어, 전국 만19세 이상 성인의 성별은 {남성, 여성} 둘 중 한 가지만 가능하며, $n=100$명의 성인을 무작위로 선택하는 경우 모집단 수가 표본 수보다 상대적으로 훨씬 커서 남성이 선택될 확률 p가 매 시행마다 거의 일정하다고 볼 수 있다.

확률표

남성	여성	합계
p	$1-p$	1

도수표의 예

남성	여성	합계
58	42	100

이항분포에서는 성공의 개수 X가 가질 수 있는 값은 0, 1, ..., n이다. 반대로, 실패의 개수는 $Y=n-X$라고 할 수 있다. 성공의 개수가 $X=x$일 확률은 다음과 같이 구할 수 있다:

$$P(X=x)=\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},\ x=0,1,...,n.$$

* 이항분포의 확률질량함수와 누적확률분포함수를 계산하는 R함수를 찾아보자.

1-2. 다항분포 (Multinomial Distribution)

다항분포는 표본공간의 원소가 세 개 이상인 경우로 이항분포를 확장시킨 분포이다. 예를 들어, 직장생활 만족도 문항에서 가능한 응답이 {불만족, 보통, 만족} 인 경우를 들 수 있다.

확률표

불만족	보통	만족	합계
p_1	p_2	p_3	1

도수표의 예

불만족	보통	만족	합계
23	51	26	100

일반적으로, 한 번의 시행에서 나타날 수 있는 결과가 k개이고 서로 배반이며, 각각의 결과를 얻을 확률을 $p_1, p_2, \ldots, p_k$ 라고 하자. 여기서, 확률의 합은 1이어야 하므로, $p_k = 1-(p_1+p_2+\cdots+p_{k-1})$로 둘 수 있다. 이때, n번의 독립적인 시행을 통해 i $(i=1,2,\ldots,k)$번째 결과가 나온 개수를 X_i라고 하자. i번째 결과가 $X_i = x_i$번 나올 확률은

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

$$x_1 + x_2 + \cdots + x_k = n$$

$$p_1 + p_2 + \cdots + p_k = 1$$

여기서, 다항분포는 확률변수 한 개가 아닌 여러 개의 확률변수 $X_1, \ldots, X_{k-1}$가 가질 수 있는 값들에 대한 분포임을 알 수 있다:

$$(X_1, X_2, \ldots, X_{k-1}) \sim Mltinomial(n, p_1, p_2, \ldots, p_{k-1})$$

* R함수 rmultinom(n, size, prob)을 이용해 다항분포로부터 시행횟수 $n=100$이고 확률 $p_1 = p_2 = p_3 = 1/3$인 표본을 추출해보자. 주의: 시행 횟수값은 size=에 전달함. 또한, prob=에는 길이가 k인 벡터 $(p_1, p_2, \ldots, p_k)$를 전달함.

1-3. 다항분포의 근사

다항분포에서 n가 충분히 크면 근사적으로 다음이 성립한다:

$$\sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \sim \chi^2_{k-1}$$

만약, 다항분포에서 $k=2$이면 이항분포에 해당하며, 위 식의 좌변을 다음과 같이 정리할 수 있다:

$$\sum_{i=1}^{2} \frac{(X_i - np_i)^2}{np_i} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{np_2}$$

$$= \frac{(X_1 - np_1)^2}{np_1 p_2} = \left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \right)^2$$

이항분포의 정규근사에 의해, $(X_1 - np_1)/\sqrt{np_1(1-p_1)} \sim N(0,1)$ 이므로, 카이제곱분포

의 정의에 따라 다음이 성립함을 알 수 있다:

$$\sum_{i=1}^{2}\frac{(X_i - np_i)^2}{np_i} \sim \chi_1^2$$

2. 적합도 검정(Goodness-of-Fit Test)

통계분석에서는 종종 데이터가 특정 분포를 따름을 가정한다. 이 장에서는 범주형 자료가 특정 분포를 따르는지를 검정하기 위한 적합도 검정법인 카이제곱 검정을 수행한다. 적합도검정에서의 귀무가설과 대립가설은 다음과 같다:

$$H_0 : p_1 = p_{10},\ \cdots, p_{k-1} = p_{k-1,0}$$

$$H_1 : H_0\text{는 사실이 아님}$$

귀무가설이 사실이라면 기대빈도는 다음과 같다:

$$E_{i0} = np_{i0},\quad i = 1,2,\ldots,k-1$$

$$E_{i0} = n\left(1 - \sum_{i=1}^{k-1} p_{i0}\right),\quad i = k$$

이를 위해 사용되는 검정통계량은 다음과 같다.

$$X_0^2 = \sum_{i=1}^{k}\frac{(O_i - E_{i0})^2}{E_{i0}} \sim \chi_{k-1}^2$$

k: 칸의 수

O_i: i번째 칸의 관측도수

E_{i0}: 귀무가설 하에서 i번째 칸의 기대도수

* 유의수준 α일 때 기각역을 정의해보시오.

* p-값을 정의해보시오.

```
## survey 자료를 사용해 글씨를 왼손으로 쓰는 사람과 오른손으로 쓰는 사람의 비율이 3:7인지의 여부를 분석해보자.
data(survey, package = "MASS")
table(survey$W.Hnd)

 Left Right 
   18   218 
chisq.test(table(survey$W.Hnd), p=c(0.3, 0.7))

        Chi-squared test for given probabilities

data:  table(survey$W.Hnd)
```

```
X-squared = 56.252, df = 1, p-value = 6.376e-14

## 귀무가설 하에서 검정통계량의 값을 직접 계산해보자.

## 기각역과 p-값을 각각 구하고, 위 결과와 동일한 지 비교해보자.
```

3. 독립성 검정: 두 변수 간의 연관 여부

분할표의 행에 나열된 속성과 열에 나열된 속성이 독립이라면 (i,j)셀의 확률 p_{ij}에 대해 아래 식이 성립한다.

$$p_{ij}=p_{i\cdot}p_{\cdot j}$$

두 범주형 변수 X_1, X_2의 독립성을 검정하기 위한 독립성 검정의 가설은 다음과 같다:

H_0 : 두 범주형 변수 X_1, X_2는 서로 독립이다.

H_1 : 두 범주형 변수 X_1, X_2는 서로 독립이 아니다.

따라서, 독립성 검정을 위한 가설은 다음과 같이 설정할 수 있다:

H_0 : $p_{ij}=p_{i\cdot}p_{\cdot j}$, for all i,j

H_1 : H_0가 사실이 아니다. (즉, 적어도 하나의 $p_{ij}\neq p_{i\cdot}p_{\cdot j}$)

귀무가설 하에서 검정통계량은 다음과 같다:

$$\begin{aligned} X_0^2 &= \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \\ &= \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(O_{ij}-np_{i\cdot}p_{\cdot j})^2}{np_{i\cdot}p_{\cdot j}} \sim \chi^2_{(k-1)(l-1)} \end{aligned}$$

k: 행의 수

l: 열의 수

O_{ij}: (i,j)번째 칸의 관측값(관측도수)

E_{ij}: 귀무가설이 참일 때 (i,j)번째 칸의 기댓값(기대도수).

$E_{ij}=np_{i\cdot}p_{\cdot j}$. 하지만 알려져 있지 않은 값이 포함되어 있으므로, 추정값으로 대체하여 사용한다:

$$E_{ij}=n\left(\frac{O_{i\cdot}}{n}\right)\left(\frac{O_{\cdot j}}{n}\right)=\frac{O_{i\cdot}O_{\cdot j}}{n}$$

카이제곱 분포의 자유도는,

$(kl-1)-$[추정되는 모수의 수]$=(kl-1)-(k-1)-(l-1)=(k-1)(l-1)$

```
## survey 자료를 사용해 성별에 따른 운동량에 차이가 있는지 알아보자.
data(survey, package = "MASS")
head(survey[c("Sex","Exer")])
xtabs(~ Sex + Exer, data = survey)

## chisq.test()를 이용한다. 행렬자료인 경우 독립성검정을 수행한다.
chisq.test(xtabs(~ Sex + Exer, data = survey))

        Pearson's Chi-squared test

data:  xtabs(~Sex + Exer, data = survey)
X-squared = 5.7184, df = 2, p-value = 0.05731
```

4. 동일성 검정: 하위 모집단 사이의 분포가 동일한지 여부

동일성 검정 또한 독립성 검정처럼 “$(k \times l)$ 분할표”를 사용하고 카이제곱 통계량을 사용한다. 독립성 검정과 검정방법 또한 동일하지만 검정을 바라보는 관점이 다르다. 구체적으로, 독립성은 두 변수간의 연관이지만 동일성은 주로 행에 위치하는 하위 집단 별로 열 범주별 분포가 서로 같은지에 관심 있다. 설문조사 등에서 조사 집단 간 동일성 여부를 판단하는 데 많이 사용된다.

독립성 검정을 하는 것 보다 동일성검정을 하는 것이 더 타당한 경우가 있다.
(예1) 인문계 고등학생과 자연계 고등학생에 따라 국어/영어/수학에 대한 선호도에 차이가 있는가를 연구하고자 할 때, 이 경우는 고등학생 계열과 과목 선호도 간에 독립성 문제라기보다는, 인문계와 자연계에 따라 과목 선호도가 동일한가를 판단하는 동일성 문제로 보는 것이 타당하다.
(예2) 단과대학별로 남녀 분포가 같은지 알아보고자 할 때.

두 범주형 변수 X_1과 X_2의 동일성을 검정하는 가설은 다음과 같다:

H_0 : X_2의 각 수준 $\{R_1, R_2, ..., R_k\}$마다 X_1의 각 수준에 속할 확률은 동일하다.

H_1 : X_2의 각 수준 $\{R_1, R_2, ..., R_l\}$마다 X_1의 각 수준에 속할 확률은 동일하지 않다.

귀무가설 하에서의 검정통계량은

$$X_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(k-1)(l-1)}$$

$$E_{ij} = O_i . \times \left(\frac{O._j}{n} \right) = \frac{O_i . O._j}{n}$$

```
## Titanic 자료는 타이타닉 승객의 객실(Class), 성별(Sex), 연령대(Age),
## 생존여부(Survived)에 대한 자료이다.
## 생존여부에 따라 객실별 분포가 동일한지 유의수준 0.05에서 검정하라.
str(Titanic)
table [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"

## Class와 Survived별로 자료를 누적(합)한다.
apply(Titanic, c(1,4), sum)
    Survived
Class   No Yes
1st  122 203
2nd  167 118
3rd  528 178
Crew 673 212
## Survived변수를 행으로, 다른 변수를 열로 위치시키기 위해
## 전치(transpose)한다. Survived 값에 따른 하위 집단간 분포를 비교한다.
d2 <- t(apply(Titanic, c(1,4), sum))
        Class
Survived 1st 2nd 3rd Crew
     No  122 167 528  673
     Yes 203 118 178  212
chisq.test(d2)

        Pearson's Chi-squared test

data:  d2
X-squared = 190.4, df = 3, p-value < 2.2e-16
```