Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

08

χ^2-분포와 χ^2-검정

경북대학교 배준현 교수
(joonion@knu.ac.kr)

08. χ^2-분포와 χ^2-검정

- **독립성검정과 적합성검정**
  - 교차표를 통해 범주형으로 수집된 두 변수의 조합별 빈도 파악
  - 독립성 검정: *independence* test
    - 두 범주형 변수 간의 관련성이 모집단에서 존재하는지 검정
  - 적합성 검정: *goodness-of-fit* test
    - 두 개 이상의 범주를 갖는 범주형 변수의 범주별 비율의 분포를 관측
    - 관측된 범주별 빈도로 모집단에서 기대되는 비율 분포가 존재하는지 검정

08. χ^2-분포와 χ^2-검정

- **카이스퀘어 검정:** *chi-square* test
  - **교차표상의 응답 빈도를 바탕으로 범주형 변수 간의 관련성 검정**
  - **예) 안전벨트 착용과 승객 안전 간의 관계 분석**
    - **교통사고 환자의 안전벨트 착용 유무와 환자 상태를 조사한 교차표**

```
> survivors <- matrix(c(1443, 151, 47, 1781, 312, 135), ncol=2)
> dimnames(survivors) <- list("Status"=c("minor injury", "serious injury", "dead"),
                              "Seatbelt"=c("With Seatbelt", "Without Seatbelt"))
> survivors
                Seatbelt
Status           With Seatbelt Without Seatbelt
  minor injury            1443             1781
  serious injury           151              312
  dead                      47              135
```

08. χ^2-분포와 χ^2-검정

- **두 변수 간의 관계 파악을 위해 교차표에 합계와 비율을 추가**

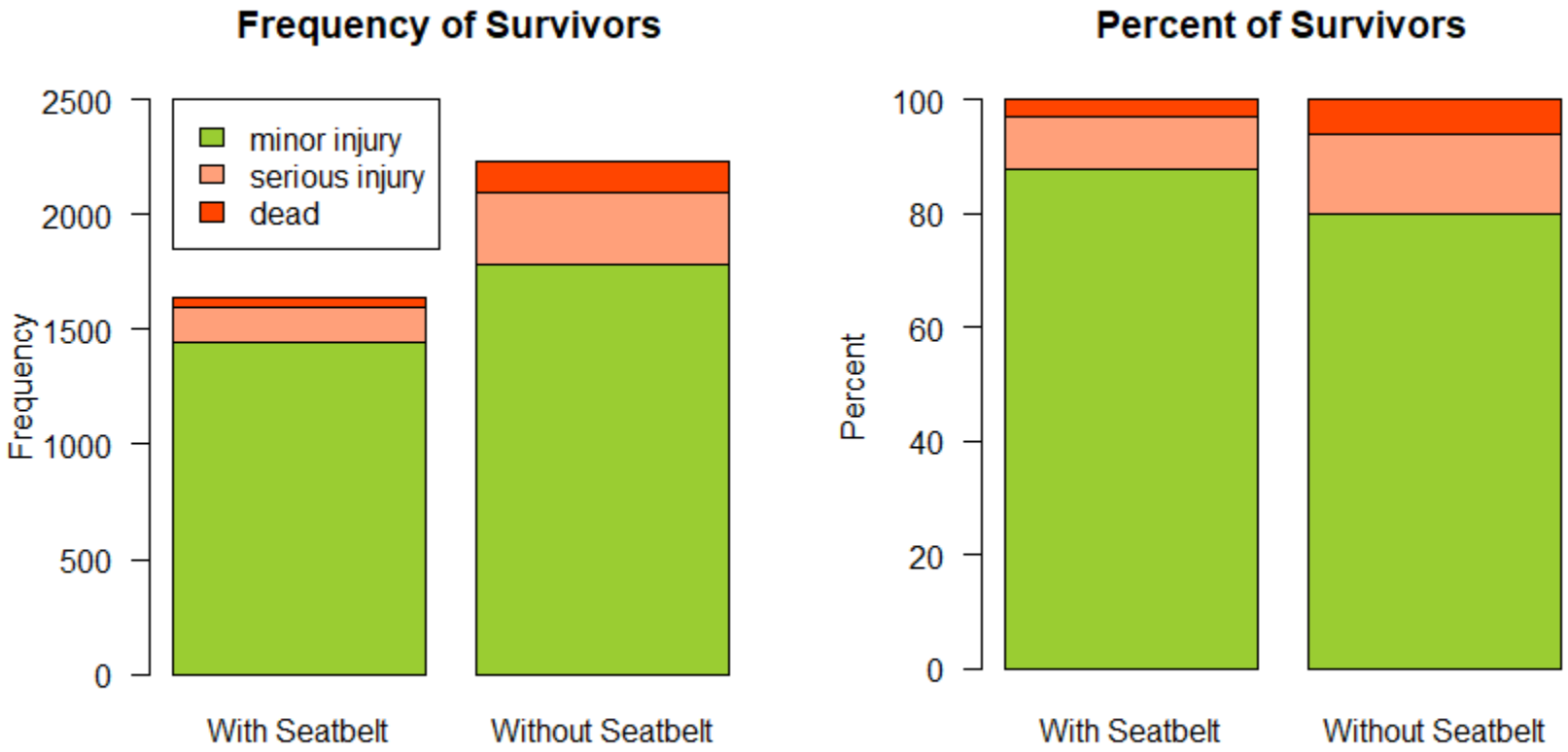
```
> addmargins(survivors)
                Seatbelt
Status           With Seatbelt Without Seatbelt  Sum
  minor injury            1443              1781 3224
  serious injury           151               312  463
  dead                      47               135  182
  Sum                     1641              2228 3869

> addmargins(prop.table(addmargins(survivors, 2), 2), 1)
                Seatbelt
Status           With Seatbelt Without Seatbelt        Sum
  minor injury      0.87934186       0.79937163 0.83329026
  serious injury    0.09201706       0.14003591 0.11966917
  dead              0.02864107       0.06059246 0.04704058
  Sum               1.00000000       1.00000000 1.00000000
```

08. χ^2-분포와 χ^2-검정

- 안전벨트 착용과 안전 간의 관계 파악을 위한 막대그래프 그리기

```
par(mfrow=c(1, 2))
barplot(survivors, ylim=c(0, 2500), las=1,
        col=c("yellowgreen", "lightsalmon", "orangered"),
        ylab="Frequency", main="Frequency of Survivors")
legend(0.2, 2500, rownames(survivors),
       fill=c("yellowgreen", "lightsalmon", "orangered"))
survivors.prop <- prop.table(survivors, 2)
barplot(survivors.prop*100, las=1, col=c("yellowgreen", "lightsalmon", "orangered"),
        ylab="Percent", main="Percent of Survivors")
par(mfrow=c(1, 1))
```

08. χ^2-분포와 χ^2-검정



08. χ^2-분포와 χ^2-검정

- **통계적 가설검정:**
  - 귀무가설: 안전벨트 착용과 승객 안전 간에는 관련이 없다
  - 관측빈도: *observed* count
    - 실제로 관측된 값. 교차표상의 셀값
  - 기대빈도: *expected* count
    - 귀무가설이 참이라는 전제하에 우리가 기대할 수 있는 빈도
  - 기대빈도의 계산
    - 귀무가설이 참이라면 안전벨트 착용 여부와 관계없이
    - 두 집단 모두에게서 환자의 상태별 비율이 동일하게 나타날 것

- 환자 상태별 기대빈도 계산:
  - 경상 환자의 비율: 83.3%
    - 안전벨트 착용자의 경상 기대빈도: 0.83*1641=1367
    - 안전벨트 미착용자의 경상 기대빈도: 0.83*2228=1855.9
  - 중상 환자의 비율: 12.0%
    - 안전벨트 착용자의 중상 기대빈도: 0.12*1641=196.9
    - 안전벨트 미착용자의 중상 기대빈도: 0.12*2228=267.4
  - 사망 환자의 비율: 4.7%
    - 안전벨트 착용자의 사망 기대빈도: 0.047*1641=77.1
    - 안전벨트 미착용자의 사망 기대빈도: 0.047*2228=104.7

08. χ^2-분포와 χ^2-검정

- **카이스퀘어 검정 절차:**
  - $\chi^2 - value$: 기대빈도와 관측빈도의 비교를 통해 계산되는 값
    - $\chi^2 = \sum_{ij} \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$, o_{ij}: 관측빈도, e_{ij}:기대빈도, i, j: 행과 열
  - χ^2-검정:
    - 표본으로부터 산출된 χ^2값이 귀무가설이 참이라는 가정하에
    - χ^2분포상에서 얼마나 나타나기 어려운 희박한 경우인지
      - 혹은 흔하게 관찰될 수 있는 경우인지 평가

- 환자 상태별
  - 환자 상태별 기대빈도:
    - $\chi^2 = \frac{(1443-1367)^2}{1367} + \frac{(1781-1855.9)^2}{1855.9} + \frac{(151-196.9)^2}{196.9} + \frac{(312-267.4)^2}{267.4}$

$$+ \frac{(47-77.1)^2}{77.1} + \frac{(135-104.7)^2}{104.7}$$

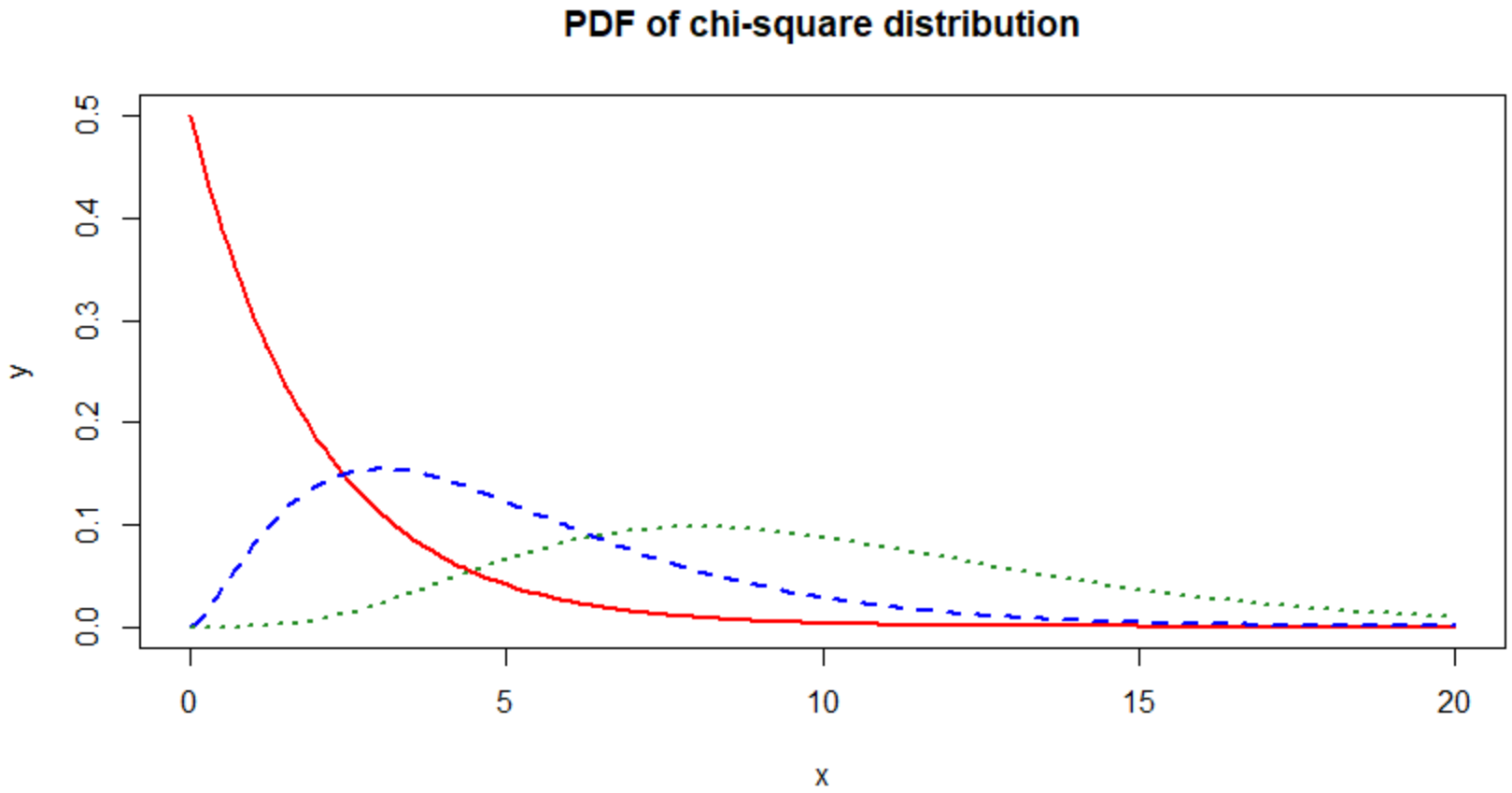
$$= 45.91$$

08. χ^2-분포와 χ^2-검정

```
> expected <- matrix(c(1367, 1855.9, 196.9, 267.4, 77.1, 104.7), ncol=2, byrow = T)
> dimnames(expected) <- list("Status"=c("minor injury", "serious injury", "dead"),
                             "Seatbelt"=c("With Seatbelt", "Without Seatbelt"))
> expected
                Seatbelt
Status           With Seatbelt Without Seatbelt
  minor injury          1367.0           1855.9
  serious injury         196.9            267.4
  dead                    77.1            104.7

> chisqr <- sum((survivors - expected)^2 / expected)
> chisqr
[1] 45.90677
```

- **카이스퀘어 분포**
  - 자유도에 따라 분포의 모양이 달라지며, 대체로 오른쪽으로 긴 꼬리를 가짐
  - χ^2분포의 자유도: 교차표를 구성하는 두 변수의 범주의 개수에 의해 결정
    - 자유도 $=$ (행 변수의 범주의 개수 -1) $\times$ (열 변수의 범주의 개수 -1)
    - 자유도 $=$ (교차표의 행 개수 -1) $\times$ (교차표의 열 개수 -1)

08. χ^2-분포와 χ^2-검정



08. χ^2-분포와 χ^2-검정

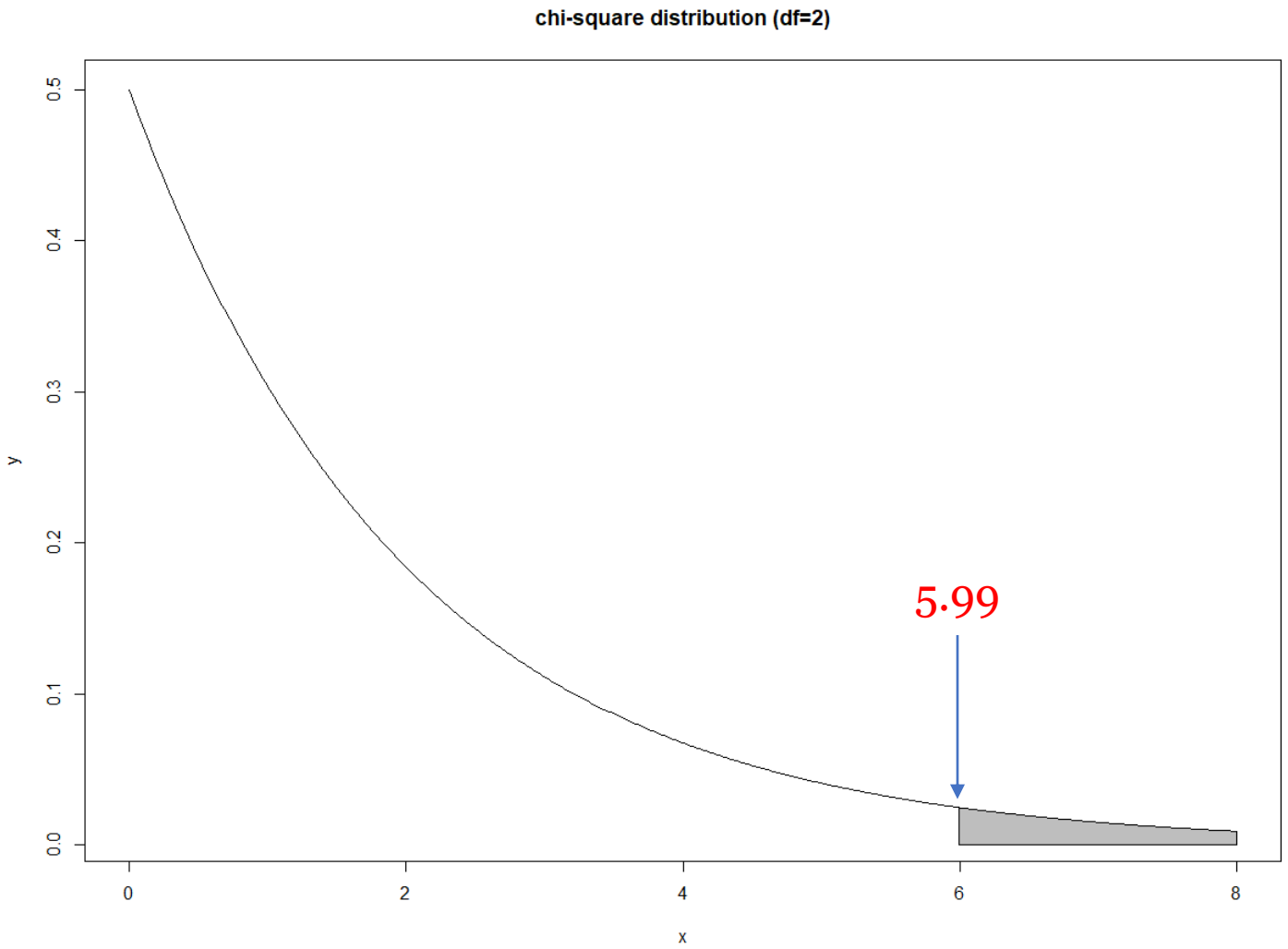
- pchisq() 함수를 이용하여 특정 χ^2값에 대응되는 유의확률을 구할 수 있음

```
> pchisq(45.91, df=2, lower.tail=FALSE)
[1] 1.073421e-10
```

- qchisq() 함수를 이용하여 특정 확률에 대응하는 χ^2값을 구할 수 있음

```
> qchisq(0.05, df=2, lower.tail=FALSE)
[1] 5.991465
```

08. χ^2-분포와 χ^2-검정



08. χ^2-분포와 χ^2-검정

- **독립성 검정:** *independence* test
  - 두 범주형 변수가 서로 독립인지 검정
    - 독립: 두 변수가 서로 관련이 없다.
  - 성별과 선호하는 도서 장르가 독립이다.
    - 성별에 따라서 좋아하는 도서 장르가 다르지 않다. (서로 관련이 없다)
  - 두 변수의 범주 조합별 빈도를 기록한 교차표를 토대로 χ^2-검정 절차를 수행

08. χ^2-분포와 χ^2-검정

- Titanic 데이터셋을 이용하여 독립성검정 수행

```
> 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“

> Titanic
, , Age = Child, Survived = No
 Sex
Class  Male Female
  1st     0      0
  2nd     0      0
  3rd    35     17
  Crew    0      0

, , Age = Adult, Survived = No
……
```

08. χ^2-분포와 χ^2-검정

- 타이타닉호 탑승객의 승객 구분(1등실, 2등실, 3등실, 승무원)에 따른 생존율 차이 분석

```
> Titanic.margin <- margin.table(Titanic, margin=c(4, 1))
> Titanic.margin
        Class
Survived 1st 2nd 3rd Crew
     No  122 167 528  673
     Yes 203 118 178  212
```

08. χ^2-분포와 χ^2-검정

- **카이스퀘어 분석을 위한 교차표 만들기**

```
> addmargins(Titanic.margin)
        Class
Survived  1st  2nd  3rd Crew  Sum
     No   122  167  528  673 1490
     Yes  203  118  178  212  711
     Sum  325  285  706  885 2201

> addmargins(prop.table(addmargins(Titanic.margin, 2), 2), 1)
        Class
Survived       1st       2nd       3rd      Crew       Sum
     No  0.3753846 0.5859649 0.7478754 0.7604520 0.6769650
     Yes 0.6246154 0.4140351 0.2521246 0.2395480 0.3230350
     Sum 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

08. χ^2-분포와 χ^2-검정

- 생존율 차이가 있는지 확인하기 위한 독립성 검정

```
> chisq.test(Titanic.margin)

        Pearson's Chi-squared test

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

08. χ^2-분포와 χ^2-검정

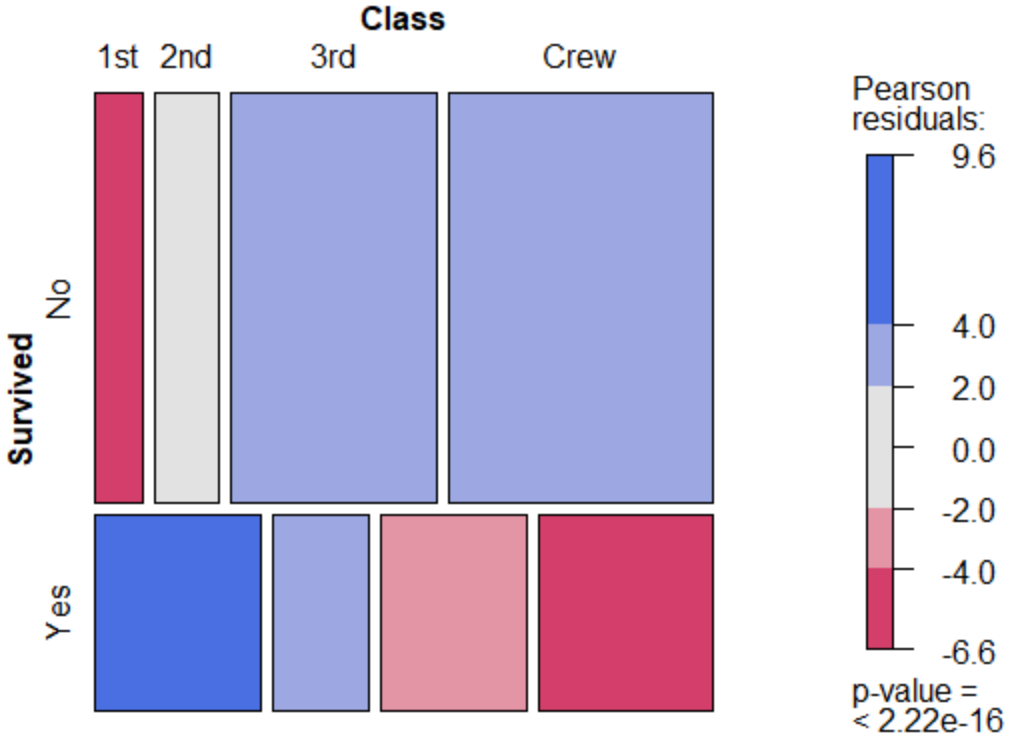
- 두 범주형 변수 간의 관련성의 강도를 평가하기

```
library(vcd)
assocstats(Titanic.margin)
                    X^2 df P(> X^2)
Likelihood Ratio 180.9  3         0
Pearson          190.4  3         0

Phi-Coefficient   : NA
Contingency Coeff.: 0.282
Cramer's V        : 0.294
```

08. χ^2-분포와 χ^2-검정

- 두 범주형 변수 간의 관계를 모자이크 플롯으로 시각화하기

```
library(vcd)
windows(width=7.0, height=5.5)
mosaic(Titanic.margin, shade=TRUE, legend=TRUE)
mosaic(~ Survived + Class, data=Titanic.margin, shade=TRUE, legend=TRUE)
```



08. χ^2-분포와 χ^2-검정

- 데이터프레임 형태의 데이터셋에 대해서는 교차표 생성 없이 직접 독립성검정 수행 가능함
- MASS 패키지의 survey 데이터셋에서 독립성 검정

```
> library(MASS)
> str(survey)
'data.frame':  237 obs. of  12 variables:
 $ Sex   : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
 $ Wr.Hnd: num  18.5 19.5 18 18.8 20 18 17.7 17 20 18.5 ...
 $ NW.Hnd: num  18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5 18.5 ...
 $ W.Hnd : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
 $ Fold  : Factor w/ 3 levels "L on R","Neither",..: 3 3 1 3 2 1 1 3 3 3 ...
 $ Pulse : int  92 104 87 NA 35 64 83 74 72 90 ...
 $ Clap  : Factor w/ 3 levels "Left","Neither",..: 1 1 2 2 3 3 3 3 3 3 ...
 $ Exer  : Factor w/ 3 levels "Freq","None",..: 3 2 2 2 3 3 1 1 3 3 ...
 $ Smoke : Factor w/ 4 levels "Heavy","Never",..: 2 4 3 2 2 2 2 2 2 2 ...
 $ Height: num  173 178 NA 160 165 ...
 $ M.I   : Factor w/ 2 levels "Imperial","Metric": 2 1 NA 2 2 1 1 2 2 2 ...
 $ Age   : num  18.2 17.6 16.9 20.3 23.7 ...
```

08. χ^2-분포와 χ^2-검정

- 성별에 따라 팔짱을 끼었을 때의 손 위치에 차이가 있는지 독립성 검정

```
> with(survey, chisq.test(Fold, Sex))
        Pearson's Chi-squared test

data:  Fold and Sex
X-squared = 2.5741, df = 2, p-value = 0.2761

> crosstab <- with(survey, table(Fold, Sex))
> crosstab
          Sex
Fold       Female Male
  L on R       48   50
  Neither       6   12
  R on L       64   56
> chisq.test(crosstab)
        Pearson's Chi-squared test

data:  crosstab
X-squared = 2.5741, df = 2, p-value = 0.2761
```

08. χ^2-분포와 χ^2-검정

- **적합성 검정:** *goodness-of-fit* test
  - 범주형 변수가 하나일 경우에는 범주별 비율 분포에 대한 가설검정을 할 수 있음
  - 적합성 검정: 관측한 빈도를 토대로 모집단에서의 비율 분포를 검정
  - 예) 세 이동통신회사의 시장점유율이 동일한지 검증
    - 150명의 휴대전화 사용자를 대상으로 이용하고 있는 이동통신회사를 조사
    - 조사 결과, A=60명, B=55명, C=35명으로 집계되었다면?

08. χ^2-분포와 χ^2-검정

- **적합성 검정:** *goodness-of-fit* test
  - 범주형 변수가 하나일 경우에는 범주별 비율 분포에 대한 가설검정을 할 수 있음
  - 적합성 검정: 관측한 빈도를 토대로 모집단에서의 비율 분포를 검정
  - 예) 세 이동통신회사의 시장점유율이 동일한지 검증
    - 150명의 휴대전화 사용자를 대상으로 이용하고 있는 이동통신회사를 조사
    - 조사 결과, A=60명, B=55명, C=35명으로 집계되었다면?

08. χ^2-분포와 χ^2-검정

- **세 이동통신회사의 시장점유율이 동일하지 적합성검정 수행**

```
> chisq.test(c(60, 55, 35))

        Chi-squared test for given probabilities

data:  c(60, 55, 35)
X-squared = 7, df = 2, p-value = 0.0302
```

08. χ^2-분포와 χ^2-검정

- **시장점유율이 각각 45%, 30%, 25% 라는 주장에 대한 적합성검정**

```
> oc <- c(60, 55, 35)
> null.p <- c(0.45, 0.30, 0.25)
> chisq.test(oc, p=null.p)

        Chi-squared test for given probabilities

data:  oc
X-squared = 3.2222, df = 2, p-value = 0.1997
```

08. χ^2-분포와 χ^2-검정

- 작년에는 조사한 결과가 있을때, 올해의 조사 결과가 작년과 동일한지 적합성 검정

```
> oc <- c(60, 55, 35)
> chisq.test(oc, p=c(45, 25, 15)/85)
        Chi-squared test for given probabilities

data:  oc
X-squared = 10.178, df = 2, p-value = 0.006165
```

08. χ^2-분포와 χ^2-검정

- **다차원 데이터를 1차원 데이터로 축약: HairEyeColor 데이터셋에서 적합성 검정**

```
> str(HairEyeColor)
 'table' num [1:4, 1:4, 1:2] 32 53 10 3 11 50 10 30 10 25 ...
 - attr(*, "dimnames")=List of 3
  ..$ Hair: chr [1:4] "Black" "Brown" "Red" "Blond"
  ..$ Eye : chr [1:4] "Brown" "Blue" "Hazel" "Green"
  ..$ Sex : chr [1:2] "Male" "Female"
```

08. χ^2-분포와 χ^2-검정

- **머리색이 갈색, 검은색, 금발이 각각 50%, 25%, 15%라는 주장에 대한 적합성 검정**

```
> hairs <- margin.table(, margin=1)
> hairsHairEyeColor
Hair
Black Brown   Red Blond
  108   286    71   127

> chisq.test(hairs, p=c(0.25, 0.50, 0.10, 0.15))
        Chi-squared test for given probabilities

data:  hairs
X-squared = 29.934, df = 3, p-value = 1.425e-06
```

08. χ^2-분포와 χ^2-검정

- survey 데이터셋: 비흡연자 70%, 나머지 유형 흡연자가 각각 10%라는 주장에 대한 적합성 검정

```
> library(MASS)
> smokers <- table(survey$Smoke)
> smokers
Heavy Never Occas Regul
   11   189    19    17

> chisq.test(smokers, p=c(0.1, 0.7, 0.10, 0.10))
        Chi-squared test for given probabilities

data:  smokers
X-squared = 12.898, df = 3, p-value = 0.004862
```

Any Questions?