

동질성 검정(Test of Homogeneity)

Code ▼

- [분할표\(Contingency table\)](#)
- [가설\(유의수준 5%\)](#)
- [검정통계량과 유의확률 직접 구하기](#)
- [R 함수를 이용한 검정](#)

동질성 검정은 서로 다른 집단에 대해 특정 범주형 자료의 분포가 유사한지 검정하는 것이다.

데이터 불러오기 (K colors)

Hide

```
df <- read.csv(file = '/Users/jaeyonglee/Documents/College/CUAI/winter_conference/R/data/k_colors.csv', header = TRUE)
df
```

emotion <fctr>	color <fctr>
happy	darkslategray
happy	black
happy	darkslategray
happy	darkslategray
happy	black
happy	darkslategray
happy	darkslategray
happy	black
happy	darkslategray
happy	black
1-10 of 28,402 rows	
Previous 1 2 3 4 5 6 ... 100 Next	

분할표(Contingency table)

Hide

```
c.tab <- table(df)
c.tab
```

```
      color
emotion black darkslategray gainsboro
angry    4601           1858        459
happy    3767           1978        663
relaxed  4049           3056        838
sad      4276           2209        648
```

가설(유의수준 5%)

귀무가설: 감정(emotion)별로 세 가지의 공통된 색(color)의 비율이 같다.

대립가설: 감정(emotion)별로 세 가지의 공통된 색(color)의 비율이 같지 않다.

H0: $P(.,j) = P(\text{angry},j) = P(\text{happy},j) = P(\text{relaxed},j) = P(\text{sad},j)$ (j: black, darkslategray, gainsboro)

H1: Not H0

검정통계량과 유의확률 직접 구하기

감정별 개체 수: $n(i,.)$

Hide

```
(a.n <- margin.table(c.tab, margin=1)) # magrin=1: 각 행의 합
```

```
emotion
angry    happy relaxed    sad
 6918    6408    7943   7133
```

색깔별 개체 수: $n(.,j)$

Hide

```
(s.n <- margin.table(c.tab, margin=2)) # margin=2: 각 열의 합
```

```
color
      black darkslategray    gainsboro
16693           9101         2608
```

색깔별 개체의 비율: $P(.,j) = n(.,j) / n$

Hide

```
(s.p <- s.n / margin.table(c.tab)) # margin.table(): 총 개체 수(n)
```

```
color
      black darkslategray    gainsboro
0.58774030    0.32043518    0.09182452
```

기대도수표: $E(i,j) = n(i,.) * P(.,j)$

Hide

```
(expected <- a.n %*% t(s.p)) # t는 전치
```

```
      color
emotion black darkslategray gainsboro
angry    4065.987    2216.771    635.2420
happy    3766.240    2053.349    588.4115
relaxed  4668.421    2545.217    729.3622
sad      4192.352    2285.664    654.9843
```

카이제곱 검정통계량: $\sum_i \sum_j (O(i,j) - E(i,j))^2 / E(i,j)$

Hide

```
o.e <- c.tab - expected # 관찰도수와 기대도수의 차이
t.t <- sum(((o.e)^2) / expected)) # 그것의 제곱을 기대도수로 나눈 값들의 합
```

```
[1] 394.769
```

기각역

Hide

```
alpha = 0.05 # 유의수준 5%
df = (4-1)*(3-1) # 동질성 검정의 카이제곱의 자유도는 (행의 수 - 1)*(열의 수 - 1)

qchisq(1-alpha, df=df) # 누적확률이 1-alpha인 지점이 기각값이다
```

```
[1] 12.59159
```

유의수준 5%에서 검정통계량이 기각역에 있으므로 귀무가설을 기각한다. 즉, 감정(emotion)별로 세 가지의 공통된 색(color)의 비율이 같지 않다는 통계적으로 유의한 결과를 얻는다.

유의확률

Hide

```
1-pchisq(t.t, df=df)
```

```
[1] 0
```

유의수준 5%에서 유의확률(p-value)이 0.05보다 작으므로 귀무가설을 기각한다. 즉, 감정(emotion)별로 세 가지의 공통된 색(color)의 비율이 같지 않다는 통계적으로 유의한 결과를 얻는다.

R 함수를 이용한 검정

Hide

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

```
      Pearson's Chi-squared test

data:  c.tab
X-squared = 394.77, df = 6, p-value < 2.2e-16
```

유의수준 5%에서 유의확률(p-value)이 0.05보다 작으므로 귀무가설을 기각한다. 즉, 감정(emotion)별로 세 가지의 공통된 색(color)의 비율이 같지 않다는 통계적으로 유의한 결과를 얻는다.