Dealing with categorical variables

동질성 검정 (Test of Homogeneity)

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

데이터 불러오기 (K colors)

```
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)

```
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(angry,j) = P(happy,j) = P(relaxed,j) = P(sad,j) (j: black, darkslategray, gainsboro)
H1: Not H0

검정통계량과 유의확률 직접 구하기

감정별 개체 수: n(i,.)

```
(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)

```
(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

```
(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)

```
(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)

```
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

```
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

```
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

```
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.

독립성 검정 (Test of Independence)

독립성 검정은 두 개의 범주형 변수가 서로 연관이 있는지를 검정하는 것이다.

분할표(Contingency table)

```
c.tab  # 위 데이터와 동일
```

```
         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)과 세 가지의 공통된 색(color)은 관련이 없다(서로 독립이다).
대립가설: 감정(emotion)과 세 가지의 공통된 색(color)은 관련이 있다(서로 독립이 아니다).

H0: P(i,j) = P(i,.) * P(.,j) (i: angry, happy, relaxed, sad), (j: black, darkslategray, gainsboro)
H1: Not H0

검정통계량과 유의확률 직접 구하기

감정별 개체 수: n(i,.)

```
(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)

```
(g.n <- margin.table(c.tab, margin=2))  # magrin=2: 각 행의 합
```

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

감정별 개체의 비율: P(i,.)

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

```
emotion
    angry     happy   relaxed       sad 
0.2435744 0.2256179 0.2796634 0.2511443 
```

색깔별 개체의 비율: P(.,j)

```
(g.p <- g.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 * P(i,.) * P(.j)

```
(expected <- margin.table(c.tab) * (a.p %*% t(g.p)))  # margin.table(): 총 개체 수(n)
```

```
         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)

```
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

```
alpha = 0.05  # 유의수준 5%
df = (4-1)*(3-1)  # 독립성 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)과 세 가지의 공통된 색(color)은 관련이 있다(서로 독립이 아니다)는 통계적으로 유의한 결과를 얻는다.

유의확률

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

```
[1] 0
```

유의수준 5%에서 유의확률이 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 감정(emotion)과 세 가지의 공통된 색(color)은 관련이 있다(서로 독립이 아니다)는 통계적으로 유의한 결과를 얻는다.

R 함수를 이용한 검정

```
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 유의확률이 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 즉, 감정(emotion)과 세 가지의 공통된 색(color)은 관련이 있다(서로 독립이 아니다)는 통계적으로 유의한 결과를 얻는다.

참고

참고로, 분할표가 2X2 형태인 경우, chisq.test()는 자동으로 'Yates의 연속성 수정'을 통해 카이제곱 검정통계량을 구해서 검정한다. 이는 검정통계량의 분자부분을 (|O(i,j) - E(i,j)| - 0.5)^2 이렇게 수정해준 것이다. 이를 사용하지 않으려면 correct=FALSE를 전달하면 된다.

이는 2X2 자료의 경우 각 항에 나타나는 확률이 이항분포를 따르지만, 여러 번 반복할 경우 이항분포가 정규분포와 비슷한 형태를 가지게 되어 정규분포로 가정하고, 이산형 분포를 연속형 분포로 가정 시 발생하는 차이를 수정하기 위함이다.

또한, 동질성 검정과 독립성 검정이 검정통계량이 같은데 이는 검정통계량의 최종 과정만 똑같을 뿐 알고자 하는 바는 두 검정이 서로 다르니, 가설을 수립하고 검정통계량을 유도하는 과정에서의 두 검정의 차이를 잘 알아둘 필요가 있다.

이 둘의 차이 설명
이 둘의 차이 설명2

결론적으로, 동질성 검정은 한 모집단 내의 여러 그룹에 대한 분포의 동질성 검정이고, 독립성 검정은 한 모집단에 대해 두 변수간의 독립성에 대한 검정이다. 따라서, 위 데이터로는 사실 독립성 검정을 한 것이 잘못된 것 같다. 그냥 검정통계량을 구하고 검정하는 방법만 참고 하자.