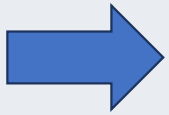
범주형 자료분석 개론

2021210088 허지혜

2.5 순서형 자료의 독립성 검정

자연스러운 순서를 가짐

분할표의 행,열이 순서형일때
검정통계량 X^2 과 G^2 카이제곱 독립성검정은
순서형 정보를 사용 x



순서적인 개념을 활용하는
검정통계량 사용

2.5.1 독립성가설과 선형추세의 대립가설

행변수 X
열변수 Y

=> 추세연관성

X : 증가 => Y : 증가 or 감소

하나의 모수를 사용하여 순서형 추세연관성을 설명

일반적인 분석 방법)
범주 수준에 점수를 할당하여 선형추세나 상관관계를 측정하는 것이다.

X 와 Y 사이의 관계에서 양 또는 음의 선형추세에 민감한 검정통계량을 제시하고자 하는데,
이 통계량은 자료의 상관관계에 대한 정보를 이용하고 있다.

2.5.1 독립성가설과 선형추세의 대립가설

행점수 : $u_1 \le u_1 \le \cdots \le u_r$
열점수 : $v_1 \le v_1 \le \cdots \le v_c$

=

범주의 수준과 같은 순서를 갖는 단조점수

행점수 평균 표본 : $\bar{u} = \sum u_i \widehat{\pi_{i+}}$
열점수 평균 표본 : $\bar{v} = \sum v_j \widehat{\pi_{+j}}$

점수를 정할 때 범주 사이의 거리를 반영하도록 해야함,
범주 간 거리가 클수록 서로 더 멀리 떨어진 것으로 간주

X와 Y 사이의 표본상관계수 -1≤R≤1

$$R = \frac{\sum(u_i - \bar{u})(v_i - \bar{v})\widehat{\pi_{ij}}}{\sqrt{[\sum(u_i - \bar{u})^2 \widehat{\pi_{i+}}][\sum(v_i - \bar{v})^2 \widehat{\pi_{+j}}]}}$$

두 변수 사이의 독립성
= 상관계수의 참값이 0

절대값이 클수록 독립성으로부터
선형적으로 멀어지는 것으로 나타낸다.

2.5.1 독립성가설과 선형추세의 대립가설

상관계수의 참값이 0이라는 독립성가설과 상관계수의 참값이 0이 아니라는 양측 대립가설을 검정하기 위한 검정통계량은 $M^2 = (n-1)R^2$ 이다.
층화된 순서형 분할표에 대한 통계량의 특별한 경우에 해당한다.

이 통계량은 표본상관계수 $|R|$ 의 크기가 증가하고 표본크기 n이 증가하면 함께 증가한다.
또한 표본크기가 크면 이 통계량은 근사적으로 자유도 df=1인 카이제곱분포를 따른다.

통계량이 큰 값을 가지면 X^2 와 G^2 와 같이 독립성을 반증하며 이때 P-값은 관측값의 오른쪽 꼬리 부분의 확률이 된다. 또한 $M = \sqrt{(n-1)R}$ 은 근사적으로 표준정규분포를 따른다.
이 통계량은 두 변수 사이에 상관관계의 방향성까지 함께 나타내는 단측대립가설에 적용할 수 있다.

X^2 와 G^2 와 같이 M^2 도 반응변수와 설명변수를 구별하지 않는다.
즉, 열과 행을 바꾸어도 동일한 M^2 의 값을 얻게 된다.

2.5.2 예제 : 음주와 영아의 기형성에 관한 연구

알코올 소비량	기형아		합계	기형아 발생 퍼센티지	표준화된 잔차
	없음	있음			
0	17066	48	17114	0.28	-0.18
<1	14464	38	14502	0.26	-0.71
1-2	788	5	793	0.63	1.84
3-5	126	1	127	0.79	1.06
≥6	37	1	38	2.63	2.71

임산부의 음주의 효과에 대한 전향적 연구이다.
표본 추출된 임산부들이 임신 3개월이 경과했을 때를 대상으로 알코올 소비량에 관한 설문조사를 실시하였다. 그 후 영아들을 대상으로 기형아인지 아닌지 여부를 관측한 결과를 기록하였다.

매일 평균 음주량으로 측정된 알코올 소비량 = 순서형 범주를 가지는 반응변수
기형 여부 = 이항 변수

2.5.2 예제 : 음주와 영아의 기형성에 관한 연구

알코올 소비량	기형아		합계	기형아 발생 퍼센티지	표준화된 잔차
	없음	있음			
0	17066	48	17114	0.28	-0.18
<1	14464	38	14502	0.26	-0.71
1-2	788	5	793	0.63	1.84
3-5	126	1	127	0.79	1.06
≥6	37	1	38	2.63	2.71

어떤 변수가 두 범주만 갖고 있는 이항변수인 경우, 순서형 변수를 다루는 M^2 와 같은 통계량을 사용할 수 있다.

예로, 기형 여부에 대하여 '없음'의 범주를 낮음, '있음'의 범주를 높음으로 간주할 수 있다.
두 범주에 대하여 어떤 점수를 선택하더라도 같은 M^2 값을 갖기 때문에
간단하게 '없음'이 0을, '있음'을 1이라는 값으로 사용하도록 하자.

2.5.2 예제 : 음주와 영아의 기형성에 관한 연구

알코올 소비량	기형아		합계	기형아 발생 퍼센티지	표준화된 잔차
	없음	있음			
0	17066	48	17114	0.28	-0.18
<1	14464	38	14502	0.26	-0.71
1-2	788	5	793	0.63	1.84
3-5	126	1	127	0.79	1.06
≥6	37	1	38	2.63	2.71

표의 관측도수는 아주 작은 수부터 큰 수까지 골고루 혼합되어 있다.
이 경우 표본의 크기는 왕 크지만 X^2 나 G^2 의 실제 표본추출분포는 카이제곱분포와 가깝지 않다.

통계량을 각각 구해보면 X^2 = 6.2(P=0.19)이며 G^2 = 12.1(P=0.02) => 상반된 결과
두 통계량은 모두 알코올 소비량의 순서적인 개념을 고려하지 않은 것이다.

2.5.2 예제 : 음주와 영아의 기형성에 관한 연구

알코올 소비량	기형아		합계	기형아 발생 퍼센티지	표준화된 잔차
	없음	있음			
0	17066	48	17114	0.28	-0.18
<1	14464	38	14502	0.26	-0.71
1-2	788	5	793	0.63	1.84
3-5	126	1	127	0.79	1.06
≥6	37	1	38	2.63	2.71

알코올 소비의 각 수준에서 기형아가 발생한 백분율과 기형 여부의 '있음'의 범주에 대하여 구한 표준화된 잔차값을 보여 주고 있다. 이들은 모두 알코올 소비량이 증가할수록 기형아가 더 많이 태어나는 추세를 보여준다.
순서통계량 M^2 을 계산하기 위해 알코올 소비량의 수준에 대하여 점수를 정할 필요가 있다. 이런 경우, 범주의 중간값을 점수로 사용하는 것이 합리적!

2.5.2 예제 : 음주와 영아의 기형성에 관한 연구

```
> CMHtest(Malform, rscores=c(0,0.5,1.5,4.0,7.0))
Cochran-Mantel-Haenszel Statistics

                AltHypothesis   Chisq Df     Prob
cor         Nonzero correlation  6.5699  1 0.010372
rmeans   Row mean scores differ 12.0817  4 0.016754
cmeans   Col mean scores differ  6.5699  1 0.010372
general     General association 12.0817  4 0.016754

> sqrt(6.5699)
[1] 2.563182
> 1-pnorm(2.56318)
[1] 0.005185913
```

v값들을 정해보자.

v값들을 정한다.
알코올 소비량과 기형아 발생 간의 표본상관계순 R = 0.0142이다.
여기서 선택한 점수를 이용해 구한 상관계수값은 매우 작아보인다.
그러나 이 예제와 같이 이항변수가 매우 다른 주변합을 갖고 있는 경우 큰 R값을 구하는 것은 불가능하다.
여기서 R은 검정을 위한 추세 정보를 요약하는 역할을 한다.

$M^2 = 6.57$

R의 P-값은 0.010으로 상관관계가 0이 아니라는 강력한 증거를 제시한다.
또한 표준정규통계량 M=2.56은 양의 상관관계를 나타내는 대립가설에 대해 0.005의 P-값을 갖는다.

2.5.2 예제 : 음주와 영아의 기형성에 관한 연구

```
> Malform <- matrix(c(17066, 14464, 788, 126, 37, 48, 38, 5, 1, 1),ncol=2)
> Malform
      [,1] [,2]
[1,] 17066   48
[2,] 14464   38
[3,]   788    5
[4,]   126    1
[5,]    37    1
>
> library(vcdExtra)
필요한 패키지를 로딩중입니다: vcd
필요한 패키지를 로딩중입니다: grid
필요한 패키지를 로딩중입니다: gnm
```

```
> CMHtest(Malform, rscores=c(0,0.5,1.5,4.0,7.0))
Cochran-Mantel-Haenszel Statistics

                 AltHypothesis   Chisq Df     Prob
cor         Nonzero correlation  6.5699  1 0.010372
rmeans   Row mean scores differ 12.0817  4 0.016754
cmeans   Col mean scores differ  6.5699  1 0.010372
general    General association 12.0817  4 0.016754

> sqrt(6.5699)
[1] 2.563182
> 1-pnorm(2.56318)
[1] 0.005185913
```

v값들을 정해보자.

2.5.2 예제 : 음주와 영아의 기형성에 관한 연구

다른 장에서는 모형에 근거한 분석 방법 중 하나로 M^2 검정을 다룬다.
모형에 근거한 분석방법은 칸확률의 평활 추정값 뿐만 아니라 그 효과의 크기까지 추정해준다.
이러한 추정값들은 단순한 유의성검정보다 더 많은 정보를 제공한다.

2.5.3 순서형 검정통계량의 추가적인 검정력

독립성검정을 할 때 X^2 와 G^2 은 가능한 한 가장 일반적인 형태의 대립가설을 고려하지만 칸확률은 어떤 특정한 형태의 종속성을 나타낼 수 있다.
(r-1)(c-1)개의 자유도는 대립가설이 귀무가설보다 (r-1)(c-1)개나 많은 모수를 갖고 있다는 것을 반영 하고 있다.

이러한 통계량은 일반적인 경향에 대한 검정을 할 수 있는 반면에 어떤 특별한 유형의 연관성을 알아내는 것에는 민감하지 않다.

검정통계량 M^2 는 선형추세를 나타내는 상관계수에 기반한 하나의 추가 모수만을 사용하여 연관성을 설명할 수 있다.

카이제곱 검정통계량이 이 하나의 모수에 대한 검정이면 df = 1이 된다.

2.5.3 순서형 검정통계량의 추가적인 검정력

실제로 양이나 음의 연관성이 있을 때
M^2 를 사용하는 순서형 검정법 > X^2 , G^2 검정력이다.
자유도 df 는 카이제곱분포의 평균과 같으므로 df = 1 을 갖는 M^2 는 df = (r-1)(c-1)인 X^2 나 G^2 와 비슷한 값을 갖는 경우에 상대적으로 오른쪽 꼬리부분으로 훨씬더 먼 곳에 위치하게 된다.

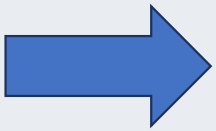
따라서 더 작은 P-값을 갖게 된다.

실제로 선형추세가 있을 때 M^2 는 X^2 , G^2 와 비슷한 크기의 값을 갖는 경향이 있으므로 결과적으로 더 작은 P-값을 제공하여 더 높은 검정력을 갖게 된다.

2.5.4 점수의 선택

대부분 자료는 점수의 선택에 영향을 안끼친다.

여러 유형의 단조점수는
유사한 결과를 보여줌.



불균형한 경우,
결과가 달라진다.

동일한 간격의 행점수들에 대해 검정통계량은 매우 약한 연관성을 나타낸다.(M^2 = 1.83, P-값 : 0.18)
R 과 M^2 의 값은 동일한 간격을 유지하는 점수의 변환에 대해서는 변하지 않는다.

예로, (1,2,3,4,5) 대신에 (0,1,2,3,4),(2,4,6,8,10)과 같은 점수들을 사용하더라도
동일한 통계량값을 얻게 된다.

2.5.4 점수의 선택

다른 방법은 점수를 선택하지 않고 각 관측값에 순위를 매긴 후에 그 순위에 범주 점수로 사용하는 방법.
일반적으로 조사자의 판단에 따라 범주 간의 거리를 반영하는 점수를 선택하는 것이 더 바람직함.
점수 선택에 자신이 없으면 민감도 분석을 할 것!

순서형 r x c 분할표에 대한 다른 검정법들은 순서형변수 연관성측도들을 활용하는 방법.

예) 감마와 켄달의 타우비는 켄달의 타우라 불리는 순서형 변수의 연관성 측도를 일반적인 분할표에 응용한 것으로 분할표에서 연관성을 나타내는 측도로 사용된다.

이러한 측도의 표본값을 표준오차로 나눈 통계량은 독립성가정 하에서 대표본 표준정규분포를 따른다.
이러한 검정법들은 M^2 에 근거한 검정법처럼 연관성을 설명하기 위해 한 개의 모수를 사용하기 때문에 뛰어난 검정력을 갖는 장점이 있다.

2.5.5 r x 2 와 2 x c 명목형-순서형 분할표의 추제검정법

순서형 변수 X, 반응 변수 Y 라면 r x 2 가 된다.

X의 수준에 따라 Y의 성공할 비율이 어떻게 변화하는지 알아보는데 관심이 있다.

행 점수가 선택되면 M^2 는 이 비율의 선형추세를 감지하게 되는데 이 내용은 뒤 모형과 관련이 있다. M^2 를 사용하는 독립성검정에서 작은 P-값은 이 선형추세의 기울기가 0이 아니라는 것을 암시한다. 이러한 순서형 변수검정법을 Cochran-Armitage 추세검정법이라고 한다.

2.5.5 r x 2 와 2 x c 명목형-순서형 분할표의 추제검정법

알코올 소비량	기형아		합계	기형아 발생 퍼센티지	표준화된 잔차
	없음	있음			
0	17066	48	17114	0.28	-0.18
<1	14464	38	14502	0.26	-0.71
1-2	788	5	793	0.63	1.84
3-5	126	1	127	0.79	1.06
≥6	37	1	38	2.63	2.71

X가 이항변수인 경우, 분할표의 크기는 2 x c이다.

표에서는 두 행이 두가지 처리를 나타낼 때 두 그룹 간 비교를 할 수 있다.

표의 M^2 통계량은 두 행 사이의 열변수 Y의 점수 {V}의 평균 차이를 알아볼 수 있다.
M^2 를 사용한 독립성검정법에서 작은 P-값은 행 간의 평균 차이가 0이 아니라고 암시한다.

2.5.5 r x 2 와 2 x c 명목형-순서형 분할표의 추제검정법

X가 r>2개의 범주를 가지는 명목형 변수이고 Y가 순서형 변수일때

M^2 검정은 두 변수를 모두 순서형으로 간주하기 때문에 적합하지 않다.

이 경우에는 나중에 다루게 될, 명목형 설명변수를 가지는 순서형 반응변수에 대한 모형을 사용해야한다. 이때 독립성검정을 위한 통계량은 df = (r-1)을 갖는 대표본 카이제곱분포를 따른다.

2.6 소표본의 정확추론과 베이지안 추론

표본크기가 증가하면 pearson의 카이제곱통계량은 근사적으로 카이제곱에 가까운 분포를 가지게 된다.
그러나 표본크기가 작거나 불규칙성이 의심될 불균형한 자료일때는 대표본 근사 방법 보다는 **정확분포**를 사용해 추론한다.

계산법들이 발전해 표본크기에 상관없이 정확분포를 정확히 근사시킬 수 있다.
독립성검정을 위해 카이제곱검정 대신 항상 정확검정을 실시할 수 있다.
즉, 어떠한 표본크기도 베이지안 방법을 사용해 정확추론을 실시 가능하다.

2.6.1 2 x 2 분할표의 Fisher의 정확추론

독립성 가설은 오즈비(=1)이 되는 가설에 대응한다.

(가정) 칸도수 $\{n_{ij}\}$ 가 독립적인 이항표본이나 네 칸에 대한 다항분표표본에서 추출

미지의 모수에 종속되지 않은 귀무가설 하에서 칸도수의 소표본 확률분포는 관측된 자료와 동일한 행합과 열합을 가지는 분할표들의 집합들을 고려해 정의된다.
칸도수가 독립적인 이항표본으로부터 추출되었을 때 행합, 즉 이항분포의 모수 n들은 이미 고정되어 있다.
이때, 관측된 자료와 동일한 열합을 가지는 분할표를 고려해보자!

고정된 행합과 열합에 대해 칸도수의 분포 = 초기화 분포
행과 열의 주변합이 주어졌을 때 n_{11} 값은 나머지 세 개의 칸도수를 결정함.
따라서 확률은 다음과 같다.

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}} \qquad \binom{a}{b} = \frac{a!}{b!\,(a-b)!}$$

이 확률분포는 미지의 모수를 포함하고 있지 않기 때문에 근사적이 아니라 정확하게 P-값을 계산할 수 있다.

2.6.1 2 x 2 분할표의 Fisher의 정확추론

독립성을 검정하기 위한 P-값은 현재 관측된 결과 또는 이 결과보다 대립가설을 더 지지하는 결과들에 대한 초기화분포의 확률합이다.

$H_\alpha : \theta > 1$ 의 대립가설에 대해 P-값을 계산해보자.

주변합이 주어졌을 때 (1,1)칸에서 현재의 n_{11} 값보다 더 큰 값을 갖는 분할표는 현재의 표본 오즈비 $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$ 보다 더 큰 표본 오즈비를 갖게 되므로 이 대립가설을 뒷받침하는 강한 증거를 보여준다.

따라서 P-값은 현재의 n_{11} 의 관측값보다 더 큰 값을 갖게 될 초기화분포의 오른쪽 꼬리분포 확률과 같다.

2 x 2 표의 검정법은 1934년에 영국의 저명한 통계학자인 fisher에 의해 제안되어서 fisher의 정확검정법이라고 부른다.

2.6.2 Fisher의 차 맛보기 실험

우유와 차 중 어느것을 먼저 컵에 부었는지 구별할 수 있다는 주장에 실험이 시작되었다.
여덟 컵의 차를 맛보는 실험인데 처음 네 컵은 우유를 붓고 다음 네 컵은 차를 먼저 부었다.
유형마다 네 컵씩 있음을 알리고 그 중 우유를 먼저 부은 네 컵을 고르도록 하였다.
각 컵들은 랜덤한 순서로 맛을 내도록 하였다.

$H_0 : \theta = 1$
$H_\alpha : \theta > 1$ 대하여 정확검정

열의 주변합은 행의 주변합 (4,4)과 동일한데, 이것은 그녀가 네 컵에 우유를 먼저 넣었다는 사실을 이미 알고 있기 때문이다. 따라서 주변분포는 자연적으로 고정되어 버렸다.
n_{11} 의 귀무가설 하에서의 분포는 행과 열의 주변합을 갖는 모든 2 X 2표에서 정의되는 초기화분포이다.
n_{11} 의 가능한 값은 (0,1,2,3,4)이다.

먼저 부은 4개의 컵 중 3개를 맞출 확률 = $P(3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = 16/70 = 0.229$

2.6.2 Fisher의 차 맛보기 실험

n_{11} =3 보다도 대립가설을 지지하는 분할표는 n_{11} =4인 경우!
이때 네 칸의 도수는 $n_{11} = n_{22}$ = 4 이고 $n_{12} = n_{21}$ =0이다.

n_{11}	확률	P-값	X^2
0	0.014	1.000	8.0
1	0.229	0.986	2.0
2	0.514	0.757	0.0
3	0.229	0.243	2.0
4	0.014	0.014	8.0

$$P(4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 1/70 = 0.014$$

모든 가능한 값과 그에 대응하는 확률들이다.

독립성가설을 기각할 만큼 확실한 증거라고 할 수 없다.

위 표는 Pearson X^2 통계량이 카이제곱분포에 근사하는 표본분포를 따르고 있지 않음을 보여준다.
주변 도수가 주어져 있을 때, X^2 통계량이 가질 수 있는 값은 0,2,8 뿐이다.

2.6.2 Fisher의 차 맛보기 실험

n_{11}	확률	P-값	X^2
0	0.014	1.000	8.0
1	0.229	0.986	2.0
2	0.514	0.757	0.0
3	0.229	0.243	2.0
4	0.014	0.014	8.0

만약 모든 컵의 넣는 순서를 올바르게 예측했다면 관측 결과는 초기화분포의 오른쪽 꼬리의 가장 극단적인 값이 되었으므로 P-값은 0.014가 되어 대립가설의 주장을 믿을 수 있는 강한 근거를 제시한다.

2.6.2 Fisher의 차 맛보기 실험

양측대립가설은 카이제곱검정에서 다루는 일반적인 형태의 통계적 종속성과 관련되어 있다.
이 검정법의 정확한 P-값은 일반적으로 관측된 도수보다 실현 가능성이 더 적은 도수들에 대응되는 확률의 양측꼬리의 합으로 정의된다.

실제 관측값의 확률 P(3)=0.229이다.
크지 않는 모든 확률을 합치면 0.485이다.
행과 열의 주변합이 같은 때 초기화분포는 단봉이며 대칭이므로 양측 P-값은 단측의 두배가 된다.

2.6.2 Fisher의 차 맛보기 실험

```
> tea <- matrix(c(3,1,1,3),ncol=2)
> fisher.test(tea)

        Fisher's Exact Test for Count Data

data:  tea
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

양측검정

```
> fisher.test(tea,alternative="greater")

        Fisher's Exact Test for Count Data

data:  tea
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309
```

단측검정

2.6.3 실제 P-값(제 1종 오류)의 보수성 : 중앙 P-값

앞에서 했던 정확검정 결과는 보수적이라고 여겨지는데 이는 0.05와 같이 원래 의도했던 명목형 제1종 오류의 확률보다 작은 실제 제 1종 오류율을 갖기 때문이다.

이러한 보수성을 줄이기 위해 중간 P-값을 사용하는 것을 추천한다.
차 맛보기 예제에서 n_{11} =3인 검정자료에서 단측중간 P-값은 $P(3)/2 + P(4) = 0.229/2 + 0.014 = 0.129$
로 기존의 P-값인 0.243보다 작은 값을 갖는다.

```
> library(epitools)
> ormidp.test(3,1,1,3,or=1)
  one.sided two.sided
1 0.1285714 0.2571429
```

독립성 검정 중앙 P-값

2.6.4 오즈비에 대한 소표본신뢰구간

오즈비에 대한 신뢰구간은 정확분포로부터 구할 수 있다.
이 신뢰구간은 θ_0 에 대하여 $H_0 : \theta = \theta_0$ 을 검정하기 위한 Fisher의 정확검정법을 일반화시킨 검정과 관련되어있다.
θ 의 95% 신뢰구간은 $H_0 : \theta = \theta_0$ 의 정확검정에서 P-값>0.05를 만족하는 모든 θ_0 의 값
즉, 유의수준 0.05에서 귀무가설을 기각하지 않는 모든 θ_0 의 값을 포함하는 구간이다.

정확검정에서 이산성 때문에 이런 신뢰구간 역시 보수적이다.
실제로 신뢰수준은 0.95와 같은 명목상의 신뢰수준보다 상당히 커질 수 있다.
이러한 보수성을 줄이기 위해 중간 P-값을 사용하는 검정법을 사용해 신뢰구간을 구할 수 있다.
이 신뢰구간은 중간 P-값이 0.05를 초과하는 모든 θ_0 의 값들로 이루어져 있으며
원래의 신뢰구간보다 길이가 짧고 실제 포함확률이 명목적인 수준에 더 가깝다.

```
> library(epitools)
> or.midp(c(3,1,1,3),conf.level=0.95)$conf.int
[1]    0.3100508 306.6338538
```

중앙 P-값을 사용한 구간

오즈비의 정확한 95% 신뢰구간 = (0.21,626.17)

구간이 넓은 이유?
표본크기가 작기 때문이다.

2.6.5 연관성 측정값에 대한 베이지안 추정

베이지안 방법들은 분할표에 대한 연관성 측정값들을 쉽게 추정한다.
2 x 2 분할표에 있는 독립인 두 이항표본들의 모수를 비교해보자.
가정) 첫번째 행의 성공횟수 Y_1 = 시행횟수가 n_1 고 모수가 π_1 인 이항분포를 따른다.
두번째 행의 성공횟수 Y_2 = 시행횟수가 n_2 고 모수가 π_2 인 이항분포를 따른다.

컬러 베이지안 방법에서는 초모수 값들이 모두 1(균등분포)이거나 0.50(제프리 사전)인 서로 독립인 베타 사전분포들을 가장 일반적으로 이용한다.
π_1 이 beta(α_1, β_1), π_2 가 beta(α_2, β_2)인 사전분포를 따를 때, 이 사전분포들의 베타를 어떤 값으로 정하느냐에 따라 서로 독립인 π_1, π_2 에 대한 베타 사후분포 beta($y_i + \alpha_i , n_i - y_i + \beta_i$)가 결정된다.

이로부터 비율의 차이, 상대위험도, 오즈비 각각에 대한 대응되는 사후분포를 구할 수 있다.
통계 프로그램을 이용해 이와 같은 연관성 척도들에 대한 신뢰구간을 구할 수 있다.

베이지안 추론에서 디폴트로 제프리 사전분포를 설정하면 모수 공간 전체에 걸쳐서 실제 포함확률이 명목적인 수준에 가깝게 유지되도록 하는 빈도주의 추론 결과를 얻을 수 있다.

2.6.5 연관성 측정값에 대한 베이지안 추정

경고)
몇몇 통계 프로그램에서는 대안적인 사후구간으로 최고사후밀도구간을 제시한다.
최고사후밀도구간은 신뢰구간 안의 모든 값에서 신뢰구간 바깥의 값보다 더 높은 사후밀도를 가지고 있다.
이 방법은 주어진 신뢰수준에서 가장 길이가 짧은 신뢰구간을 제시한다.
하지만 오즈비와 상대위험도와 같은 확률의 비선형 함수에 대한 최고사후밀도구간은 1/X와 같이 이를 비선형 번형시킨 최고사후밀도구간의 역매핑이 아니기 때문이다.

예를 들어, 오즈비θ 에 대한 95% 최고사후밀도구간이 (2.0,3.0)이라고 가정해 보자.
이 경우 비교의 대상이 되는 두 그룹을 서로 바꾸어서 표시하거나 성공, 실패를 반대로 표시했을 때와 관련된, 1/θ 의 사후분포에 대한 95% 최고사후밀도구간은 (1/2,1/3)이 아니다.
그러므로 양쪽 꼬리부분의 길이가 같은 사후신뢰구간을 사용하는 것이 더 바람직하다.

2.6.3 예제 : 소표본 임상 실험에서의 베이지안 추정

질병 치료를 위한 생물의학 연구에서는 하나의 집단이 새로운 치료를 받게 하고,
다른 집단은 표준 치료를 받거나 위약을 복용하도록 하여 새로운 치료의 효과를 검증한다.
귀무가설 $\pi_1 \leq \pi_2$ 에 대하여 대립가설을 검증하고자 할 때, 사후확률 $P(\pi_1 \leq \pi_2)$는 일종의 베이지안 P-값이다.

소표본에서는 사전분포에 대한 선택에 따라 결과값이 크게 달라질 수 있다.
환자들에게 치료법을 할당하기 위한 항아리 표본추출 방법을 사용한 임상실험 예제를 통해 이를 확인해보자.
실험처리집단에 할당된 11명의 환자들은 모두 성공이었고 대조처리집단에 할당된 1명의 환자만 실패였다.
즉, 2 x 2분할표에서 실험처리집단을 나타내는 첫 번째 행의 도수는 (11,0)이고 비교집단을 나타내는 두 번째 행의 도수는 (0,1)이다.

π_1, π_2 에 대한 beta(0.5,0.5) 사전분포에 대해서 π_1 에 대한 사후분포는 beta($y_1 + 0.5, n_1 - y_1 + 0.5$) = beta(11.5,0.5)
π_2 에 대한 사후분포는 beta($y_2 + 0.5, n_2 - y_2 + 0.5$) = beta(11.5,0.5)

2.6.3 예제 : 소표본 임상 실험에서의 베이지안 추정

이러한 사전분포를 사용하여 통계 프로그램에서 오즈비에 대해 양측 꼬리부분이 동일한 95% 사후신뢰구간을 구하면 (3.3, 1,361,274)이다.
반면에 균등사전분포를 사용해서 95% 사후신뢰구간을 구하면 (1.7, 4677)이다.
즉, 어떤 사전분포를 선택하느냐에 따라 결과값인 사후신뢰구간의 길이가 크게 달라지는 것을 확인할 수 있다.
그러나 표본의 크기가 작은 경우에는 빈도주의 방법들의 경우에도 서로 다른 방법들마다 매우 다른 결과값을 낼 수 있다.

사전분포 beta(0.5, 0.5)를 이용한 예제에서 꼬리부분이 동일한 95% 신뢰구간의 매우 정확한 시뮬레이션 결과는 다음과 같다.

```
> library(PropCIs)
> orci.bayes(11,11,0,1,0.5,0.5,0.5,0.5,0.95, nsim=1000000)
[1] 3.276438e+00 1.361274e+06
> diffci.bayes(11,11,0,1,0.5,0.5,0.5,0.5,0.95, nsim=1000000)
[1] 0.09899729 0.99327276
```

2.6.3 예제 : 소표본 임상 실험에서의 베이지안 추정

정확도는 떠러지지만 오즈비와 비율의 차이와 같은 척도들의 신뢰구간을 근사시키는 간단한 방법은 직접 시뮬레이션 하는 것이다.

π_1, π_2 에 대한 사후 베타 분포에서 많은 수의 베타 확률변수들을 생성시키고,
생성된 확률변수에 대한 척도를 계산한 다음, 구하고자 하는 확률값에 대응하는 분위수를 찾는다.
$\pi_1 \leq \pi_2$ 를 만족하는 시뮬레이션 결과값들의 비율에 근사시켜 사후확률을 구할수도 있다.

이번엔 사전분포 beta(0.5, 0.5)와 이에 대응하는π_1, π_2에 대한 사후분포 beta(11.5, 0.5)와 beta(0.5, 1.5)를 이용해 구해보자.

2.6.3 예제 : 소표본 임상 실험에서의 베이지안 추정

```
> pi1 <- rbeta(100000000,11.5,0.5)
> pi2 <- rbeta(100000000,0.5,1.5)
> or <- pi1*(1-pi2)/((1-pi1)*pi2)
> quantile(or, c(0..25, 0.975))
에러: 예상하지 못한 수치형 상수(numeric constant)입니다. in "quantile(or, c($
> quantile(or, c(0.25, 0.975))
         25%         97.5%
5.799895e+01 1.363736e+06
> quantile(pi1 - pi2, c(0.025, 0.975))
      2.5%      97.5%
0.09896292 0.99327771
> mean(pi1 < pi2)
[1] 0.00587557
```