범주형 자료분석 개론 1.4 ~ 1.6

2021210088 허지혜

1.4 이산형 자료에 대한 통계적 추론

모수를 통계적으로 추론하는 세가지 표준적인 방법

(1) 추정된 표준오차인 $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ 를 이용하여 신뢰구간 $\hat{\pi} \pm z_{\alpha/2}$ 을 구한다.

(2) 귀무가설 하의 표준오차인 $SE_0 = \sqrt{\pi_0(1-\pi_0)/n}$ 를 이용한 유의성검정통계량 $z = (\hat{\pi} - \pi_0)/SE_0$ 를 이용해 신뢰구간을 구한다.

(3) 가능도 함수를 이용한 추론 방법

설명변수의 선형적인 효과를 나타내는 임의의 모수 β
다음 귀무가설($H_0 : \beta = \beta_0$)에 대해 유의성 검정을 해보자.

첫번째 방법)

간단한 검정통계량 = 최대가능도추정량 $\hat{\beta}$ 의 대표본 정규성을 이용하는 것

최대가능도추정값을 대입하여 구한 $\hat{\beta}$ 의 **표준오차**를 SE 라고 하자.

→ 왈드 통계량
(Wald statistic)

귀무가설이 참일때 검정통계량은 $z = (\hat{\beta} - \beta_0)/SE$ 이며
근사적으로 표준정규분포를 따른다.
동등하게 z^2 은 근사적으로 자유도가 1인 카이제곱분포를 따른다.

→ 왈드 검정
(Wald test)

1.4 이산형 자료에 대한 통계적 추론 1.4.1 왈드, 가능도비, 스코어 추론

단측검정이나 양측검정에서의 P-값은 표준정규분포 z 에서 찾아 구할 수 있다.

$H_0: \beta \neq \beta_0$ 에 대해서 동등하게 z^2 은 근사적으로 자유도가 1인 카이제곱분포를 따른다.
P-값은 카이제곱확률분포에서 관측값의 오른쪽 부분에 해당하는 확률이다.

표준정규분포에서 $\pm z$ 의 양쪽 꼬리부분 확률 = 자유고가 1인 카이제곱분포에서 z^2 의 오른쪽 꼬리부분 확률

예시)

표준정규분포에서 0.05에 해당하는 양측검정 확률은 -1.96의 왼쪽과 1.96의 오른쪽 꼬리확률 부분

자유도 1에 해당하는 카이제곱분포에서 3.84의 오른쪽 꼬리부분의 확률과 동일

표준분포의 누적확률

```
> 2*pnorm(-1.96)
[1] 0.04999579
> pchisq(1.96^2,1)
[1] 0.9500042
> 1-pchisq(1.96^2,1)
[1] 0.04999579
```

카이제곱분포 누적확률

설명변수의 선형적인 효과를 나타내는 임의의 모수
다음 귀무가설($H_0 : \beta = \beta_0$)에 대해 유의성 검정을 해보자.

두번째 방법)

스코어 검정(score test)
추정된 표준오차값이 아닌, 귀무가설 H_0이 참일 때 타당한 표준오차값들을 이용한다.

예를 들어서 이항모수에 대한 z 검정에서 귀무가설 하의 표준오차값, $SE_0 = \sqrt{\hat{\pi}_0(1-\hat{\pi}_0)/n}$ 을 사용하는 것.

설명변수의 선형적인 효과를 나타내는 임의의 모수
다음 귀무가설($H_0 : \beta = \beta_0$)에 대해 유의성 검정을 해보자.

세번째 방법)

가능도비(likelihood-ratio) 검정통계량 = $2\log(l_1/l_0)$
하나의 모수 β 가 있을 때
(1) 귀무가설이 참일 경우 l_0 값 (l_0 = β_0 에 의해 계산되는 가능도 함수)
(2) 모든 모수값에 대하여 구한 가능도함수의 최댓값 l_1
(l_1 = 최대가능도추정값 $\hat{\beta}$ 에 의해 계산되는 가능도 함수)

로그 변환을 해주는 이유 = 대략적인 카이제곱 표본분포를 따르게 되기 때문.
귀무가설 $H_0 : \beta = \beta_0$ 하에서 가능도비 검정통계량은 자유도가 1인 대표본 카이제곱분포를 따르게 된다.

특징) 음수값을 가질수 없고, P-값은 카이제곱의 오른쪽 꼬리부분 확률이 됨.
l_1/l_0 가 클수록 P-값은 작아져 귀무가설을 받아들일 수 없는 강한 증거가 됨.

1.4 이산형 자료에 대한 통계적 추론 1.4.1 왈드, 가능도비, 스코어 추론

Y에 대한 정규분포를 가정한 일반적인 회귀모형에서도 왈드, 가능도비, 스코어, 가능도비 검정법들은 동일한 검정통계량과 P-값을 보여준다.

귀무가설이 참일때 1) 표본 n이 클 경우, 유사한 결과를 가져옴.
2) 표본 n이 작거나 중간 크기일 경우, 왈드 검정의 신뢰도가 가장 떨어짐.
=> 가능도비랑 스코어 검정은 실제오류 확률을 사용하기 때문.

또한 각 방법들에 대하여 신뢰구간을 계산할 수 있다.

모수 β 에 대한 95% 신뢰구간은 귀무가설 $H_0 : \beta = \beta_0$에 대한 유의성검정에서 0.05보다 큰 P-값들에 대한 β 값들의 집합이다.

예시) 95% 왈드 신뢰구간은 $z = (\hat{\beta} - \beta_0)/SE$ 일때 $|z| < 1.96$ 를 만족하는 β_0 값들의 집합

$$= \hat{\beta} \pm 1.96(SE)$$

10번중 9번의 성공을 거두는 새로운 치료법을 평가하기 위한 임상 실험
귀무가설 : $H_0 : \pi = 0.50$
대립가설 : $H_a : \pi \neq 0.50$ 을 검정하기 위한 방법 (n = 10, $\hat{\pi}$ = 0.90)

1) 왈드 검정

왈드 검정에서 추정된 표준 오차 : $SE = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{0.90(0.10)/10} = 0.095$
왈드 검정에서 추정된 z 검정통계량 : $z = (\hat{\pi} - \pi_0)/SE = (0.90 - 0.50)/0.095 = 4.22$
왈드 검정에서 추정된 카이제곱통계량 : 자유도가 1이고 $(4.22)^2 = 17.78$ 으로 P-값 < 0.001이다.

2) 스코어 통계량

스코어 통계량에서 추정된 표준 오차 : $SE_0 = \sqrt{\pi_0(1-\pi_0)/n} = \sqrt{[0.50(0.50)/10]} = 0.158$
스코어 통계량에서 추정된 z 검정통계량 : $z = (\hat{\pi} - \pi_0)/SE_0 = (0.90 - 0.50)/0.185 = 2.53$
스코어 통계량에서 추정된 카이제곱통계량 : 자유도가 1이고 $(2.53)^2 = 6.4$ 으로 P-값은 0.011이다.

10번중 9번의 성공을 거두는 새로운 치료법을 평가하기 위한 임상 실험
귀무가설 : $H_0 : \pi = 0.50$
대립가설 : $H_a : \pi \neq 0.50$ 을 검정하기 위한 방법 (n = 10, $\hat{\pi}$ = 0.90)

3) 가능도비

10번중 9번만 성공을 관측한 이항확률 = 가능도 함수 => $l(\pi) = \frac{10!}{9!1!}(\pi^9(1-\pi)^1) = 10\pi^9(1-\pi)$

위 귀무가설이 참일때, $l_0 = 10(0.50)^9(0.50) = 0.00977$
$\hat{\pi} = 0.90$ 일 때의 최대가능도추정값에 대한 기능도함수의 값 $l_1 = 10(0.90)^9(0.10) = 0.3874$ 과 비교한다.

가능도비 검정통계량 : $2\log(l_1/l_0) = 2\{\log(0.3874/0.00977)\} = 7.36$
자유도 1인 카이제곱분포로부터 구한 이 통계량은 P-값은 0.007이다.

n의 크기가 작거나 최대가능도추정량이 모수 공간에 위치할 때와 같이, 세 통계량이 눈에 띄게 다른 값을 가질 경우 최대가능도추정량의 분포가 정규성을 만족하지 못하고 표준오차 추정값도 좋지 않다고 볼 수 있다.
이런 경우, 소표본 방법이 대표본 방법을 사용하는 것보다 신뢰도가 높다.

이항모수에 대한 통계적 추정에서, 대표본 가능도비와 양측 z 스코어 검정과
이러한 검정을 기초한 신뢰구간은 nπ>=5, n(1-π)>=5 를 만족할 때 정확한 편이다.

만족하지 않으면, 이항분포를 직접 사용하는 것이 더 정확하다.

최근 소프트웨어를 사용하면 어떤 표본의 크기에 대해서도 이와 같은 직접적인 접근법을 사용할 수 있다.

예시)

10번중 9번의 성공을 거두는 새로운 치료법을 평가하기 위한 임상 실험
귀무가설 : $H_0 : \pi = 0.50$
대립가설 : $H_a : \pi > 0.50$ 을 검정하기 위한 방법 ($\pi = 0.50$ 에서 이항분포를 가정)

단측검정에서 구한 P-값 = $P(Y \geq 9) = P(9) + P(10) = \frac{10!}{9!1!}(0.50)^9(0.50) + \frac{10!}{10!0!}(0.50)^{10}(0.50)^0 = 0.011$

양측검정에서 구한 P-값 = $P(Y \geq 9 \text{ or } Y \leq 1) = 2[P(Y \geq 10)] = 0.021$

이산확률분포에서 일반적인 P-값을 이용한 소표본추정은 매우 보수적!
=
귀무가설이 참일때 P-값이 정확이 0.05가 아니라 ≤0.05라는 것

10번중 9번의 성공을 거두는 새로운 치료법을 평가하기 위한 임상 실험
귀무가설 : $H_0 : \pi = 0.50$
대립가설 : $H_a : \pi > 0.50$ 을 검정하기 위한 방법 ($\pi = 0.50$ 에서 이항분포를 가짐)

y=9 또는 y=10 일때만 꼬리에서의 P-값이 ≤0.05인 값을 가진다.
0.010 + 0.001 = 0.011로 H_0 를 기각할 확률은 0.011에 불구하다.

=> 0.05의 유의수준을 목표로 했지만 실제 제1종 확률은 0.011이라는 것이다.
의도보다 훨씬 더 작은 값을 가지게 된다.

검정통계량이 이산형 분포를 가질 때 유의성검정 문제점을 보여준다.

1.4 이산형 자료에 대한 통계적 추론 1.4.3 소표본 이항추정과 중앙 P-값

P-값을 어떻게 달라지는지 생각해보자

검정통계량 = 연속형 분포

P-값의 분포는 [0,1]에서 균동분포를 가진다.
즉, 어떤값을 가질 확률은 동일하다.
이 경우 P-값이 0.05보다 작은 값을 가질 확률은 정확히 0.05이고 기대값고 0.50이다.

검정통계량 = 이산형 분포

귀무가설 하에서 P-값의 분포는
이산형이고 기대값은 0.50보다 커진다.
평균화하는 과정에서 이산형 분포에 대한
P-값은 커지게된다.

보수적인 경향을 해결하기 위해 다른 종류의 P-값을 사용
= 중앙 P-값
= 더 극단적인 결과의 확률에 관측된 결과의 확률의 절반을 더한 값

10번중 9번의 성공을 거두는 새로운 치료법을 평가하기 위한 임상 실험
귀무가설 : $H_0 : \pi = 0.50$
대립가설 : $H_a : \pi > 0.50$ 을 검정하기 위한 방법 ($\pi = 0.50$ 에서 이항분포를 가짐)

일반적인 P-값 : $P(9) + P(10) = 0.010 + 0.001 = 0.011$
중앙 P-값 : $(P(9)/2) + P(10) = 0.010/2 + 0.001 = 0.006$

$\pi \neq 0.50$
양측 중앙 P-값 : 0.012

중앙 P-값은 귀무가설 하의 기댓값 0.50을 가지며 이는 연속분포 하에서의 검정통계량의 일반적인 P-값과 동일한 값이다.

두 개의 단측 중앙 P-값의 합 = 1
그치만 관측된 값 각각의 꼬리에서 계산된 일반적인 단측 P-값의 합은 1이 넘는다.

중앙 P-값에 기초한 추론은 소표본 방법의 보수적 경향성과 대표본 방법들의 잠재적 비적합성 사이에서 타협점을 제시한다.
중앙 P-값을 이용한 이항검정에서 기각되지 않는 π_0 값들의 집합으로부터 π 에 대한 신뢰구간을 만드는 것도 가능하다.

1.5 비율에 대한 베이지안 통계

빈도론자로 불리는 전통적인 통계적 추론 방법을 사용
= 모수값들을 고정된 값으로, 자료값들을 확률분포로 가지는 확률 변수의 실현값으로 간주
= 모수값이 주어질 때 자료가 가질 수 있는 가능한 값들을 확률의 형태로 나타낸다.

최근에는 모수를 확률변수로 간주하고
자료뿐만 아니라 모수 역시 확률분포 하에 있다고 가정하는 베이지안 방법에 대한 인기가 높아지고 있다.
= 관측된 자료가 주어질 때 모수가 가질 수 있는 가능한 값들을 확률의 형태로 나타낸다.

1.5 비율에 대한 베이지안 통계 1.5.1 통계적 추론을 위한 베이지안 방법

베이지안 방법에서는 사전분포를 가정함.
이 확률분포는 주관적인 사전적인 믿음, 다른 연구자로부터 얻은 모수값들에 대한 정보 반영
또는 아무런 정보 없이 자료에만 거의 전적으로 기초한 더욱 객관적인 추론 결과를 얻을 수 있음

사전 분포	+	가능도함수	=	모수들에 대한 사후분포를 생성

사후분포 = 사전분포와 연구에서 관측된 자료에 기반해 모수에 대한 정보를 반영함

모수 β 와 y로 표시된 자료에 대하여

$f(\beta)$ = β 의 사전분포를 나타내는 확률함수
$p(y|\beta)$ = 자료에 대한 확률함수
$g(\beta|y)$ = 자료를 관측한 이후 β 에 대한 사후분포

베이즈 정리로부터

$g(\beta|y)$는 $p(y|\beta)f(\beta)$에 비례한다.

이제, 자료를 관측한 이후, $p(y|\beta)$ 를 모수의 함수로 볼때, 이는 기능도함수 l(β)이다.
=> 모수의 사후분포 = 가능도함수 * 사전분포에 대한 확률함수 로 결정

사전분포가 비교적 수평인 경우, 자료분석가들이 실제 자료의 분석에서 선택하는 사전분포의 경우에는, 모수에 대한 사후분포 역시 가능도함수와 비슷한 모양을 가지게 된다.

위 사례들을 제외하면 사후분포는 쉽게 계산될 수 없으며 이를 근사시키기 위해 소프트웨어를 통해 시뮬레이션 방법들을 찾는다.

1) MCMC(Markow Chain Monte Carlo)

소프트웨어를 통해 사후분포에 대한 근사분포로부터 순차적으로 아주 오랫동안 확률변수값들을 생성시킨다.

자료분석가는 순차적으로 생성된 확률변수값들로부터 구한 사후분포와 평균과 같은 추정값을 근사시키는 과정에서 Monte Carlo 오류가 아주 충분히 작아지도록 충분히 오랫동안 확률변수값들을 생성시킨다.

1.5 비율에 대한 베이지안 통계 1.5.1 통계적 추론을 위한 베이지안 방법

특정 모수에 대해, 사후분포를 이용하는 베이지안 추론 방법들은 빈도론자 추론 방법과 유사하다.
빈도론자들의 95% 신뢰구간과 유사하게, 사후분포에 대한 95% 신뢰구간을 잡을 수 있다.
이런 구간을 사후구간 or 신용구간 이라고 한다.
간단한 사후구간은 양쪽 꼬리에서의 각각의 동일한 확률과 함께 사후분포의 백분위수를 사용한다.

예시)

모수에 대한 양쪽 꼬리부분이 동일한 95% 사후구간은 사후분포의 2.5와 97.5 백분위수 사이의 구간이다.

사후분포의 평균은 모수의 베이지안 점추정값이다.

베이지안 추론에서 꼬리부분의 사후확률은 P-값을 대신하여,
사후확률분포에서 모수가 양의 값을 가지는 사후확률처럼 유용하게 사용된다.

이항모수 π에 대한 베이지안 추론은 **베타 분포(beta distribution)**를 사전분포로 사용
π에 대한 베타 확률밀도함수는 다음에 비례한다.

$$f(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1}, 0 \leq \pi \leq 1$$

이 분포는 초모수라 불리는 α>0와 β>0에 의존.
이 초모수는 추론의 목표가 되는 모수인 π와 구분되는 모수이다.

베타 분포의 평균은 $E(\pi) = \alpha/(\alpha+\beta)$ 이다.

베타 확률밀도함수는 다양한 형태를 가지고 있다.
1) α = β일 때, 0.50 근처에서 대칭
[0,1]에 위치하는 균등분포는 f(π) = 1 는 0과 1 사이 구간 내에서 균등하게 질량을 가지고 있다.
이는 베타 분포에서 α = β = 1인 특별한 경우라고 볼 수 있다.

2) α = β < 1 일 때, U자형 다봉분포를 형성

3) α = β > 1 일 때, 종형분포를 형성

α = β가 커질수록 분산은 작아진다.

π에 대한 사전지식의 부족은 균등 사전분포를 제시한다.
사후분포는 이항 가능도함수와 동일한 형태로 나타낸다.
이를 대신해서 베이지안에서 주로 쓰이는 사전분포는 **제프리 사전분포**이다.

제프리 사전분포는 모수에 대한 측정 단위들이 서로 다른 경우에도 사전분포들이 서로 동등하다.
이항모수인 경우, 제프리 사전분포는 $\alpha = \beta = 0.5$인 베타 분포이며 U자형의 대칭 형태를 가진다.
이는 수평은 아니지만, 사전분포가 비교적 정보를 반영하지 않고 있기 때문에
균등 분포에 비해서 분산이 더 크고 빈도론자 방법들 중 가장 좋은 결과를 나타내는 것과
비슷한 추론 결과를 나타낸다.
예를 들어, 사후구간들이 명목적인 수준에 가까운 실제 포함확률을 갖게 된다.

다른 방법을 사용할 이유가 있지 않는 한, 제프리 사전분포를 사용하거나
균등 사전분포를 사용할 것을 추천한다.

베타 분포는 이항모수에 대한 추론을 위한 **켤레 사전분포(conjugate prior distribution)**이다.
이는 가능도함수와 결합되었을 때 사후분포가 사전분포와 같은 족이 되는 확률분포족을 의미한다.
만일 베타(α,β) 사전분포를 이항 가능도함수와 결합시킬 경우,
사후분포는 $\alpha^* = y + \alpha$, $\beta^* = n - y + \beta$ 인 베타 분포가 된다.
파이에 대한 베이지안 점추정은 사후분포의 평균이 된다.

$$\frac{\alpha^*}{\alpha^*+\beta^*} = \frac{y+\alpha}{n+\alpha+\beta} = \left(\frac{n}{n+\alpha+\beta}\right)\frac{y}{n} + \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right)\frac{\alpha}{\alpha+\beta}$$

이 추정값은 표본비율 $\hat{\pi} = y/n$ 와 사전분포의 평균 $\alpha/(\alpha+\beta)$ 의 가중평균값이다.
n이 커질수록 표본비율에 부여된 가중치 $n/(n+\alpha+\beta)$ 는 1에 가깝게 커진다.
α = β 일 때, 표본비율은 0.5에 가깝게 작아진다.

합법적 낙태에 대한 의견 n = 1810명의 사람들 중 y = 837명은 정책을 지지했고
n - y= 973명은 정책을 지지하지 않았다.
$\hat{\pi} = 0.462$ 이고 95% 스코어 신뢰구간은 (0.440,0.485)이다.
이를 베이지안 점추정값과 신뢰추정값과 비교하면 어떻게 될까요 ?

1) 제프리 베타(0.5,0.5) 사전분포

사후분포 = (α^* , β^*) 이며 $\alpha^* = y + \alpha = 837.5$, $\beta^* = n - y + \beta = 973.5$ 다.
π의 사후평균 추정값은 $\alpha^*/(\alpha^* + \beta^*) = 837.5/(837.5 + 973.5) = 0.462$ 이다.
뒤에 나오듯 소프트웨어(R)를 통해 (0.440,0.485) 신뢰구간을 구할 수있고,
양 끝점은 베타 사후밀도의 2.5와 97.5 백분위수가 된다.

베이지안 점추정과 사후구간의 최대가능도추정값은 빈도론자 95% 신뢰구간의 소수점 셋째자리 까지 동일하다.
n이 크거나 사전분포가 상당히 분산되어 있을 경우, 빈도론자와 베이지안 추론은 매우 비슷하다.
그러나 해석이 다르다.

실제 모수값 π는 (0.440,0.485) 안에 있거나 없다.

빈도론자 :
95% 신뢰구간 = 서로 다른 독립인 표본들로부터 반복적으로 신뢰구간들을 구했을때 이중 95%가 π에 포함.

베이지안 :
자료를 관측한 이후에 π가 0.440과 0.485 사이에 있을 확률이 0.95.

$H_a : \pi < 0.50$ 에 대한 $H_0 : \pi = 0.50$ 스코어 검정에서

빈도론자 :
P-값은 0.000695이다. 이러한 단측검정에서 내포된 귀모가설은 $H_0 : \pi \geq 0.50$ 이며 검정통계량값을 구하기 위해 경계값을 사용한다.
H_0이 참일 때 관측된 값이나 H_a 의 방향에서 더 극단적인 값과 같은 검정통계량을 얻을 확률이 0.000695라고 해석한다.

베이지안 :
P(π >= 0.50)이고 이 값은 0.000692이다.
자료를 관측한 이후에 π≥0.50일 확률이 0.000692라고 해석한다.

이항모수들에 대한 베이지안 방법은 베타 분포 이외에도 다른 분포들을 사전분포로 사용할 수 있다.

c > 2 인 다항모수들의 사전분포의 경우, 베타 분포는 디리슈레 분포(Dirichlet distribution)를 생성한다.
이는 합이 1인 음이 아닌 값들 ($\pi_1, \cdots, \pi_c$)에서 정의된다.
사후분포 역시 디리슈레 분포가 된다.

제시된 모형들의 효과모수들은 대부분 실수값을 갖는다.
이러한 모수들에 대하 베이지안 정규 사전분포를 사용한다.

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용

통계 소프트웨어 패키지 이용 => 범주형 자료분석 실시

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용 1.6.1 자료 파일 읽기와 패키지 실편하기

```
> Clinical <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Clinical.dat",
+ header=TRUE)
> Clinical
   subject response
1        1        1
2        2        1
3        3        1
4        4        1
5        5        1
6        6        1
7        7        1
8        8        1
9        9        1
10      10        0
> |
```

header = TRUE : R이 보내준 첫 번째 행에 변수 이름 포함

response : 이행 결과 실패(0) 성공(1)

이항모수들에 대한 통계적 추론을 실시할 수 있는 패키지 = binom

```
> install.packages("binom")
install.packages("binom")에서 경고가 발생했습니다 :
  'lib = "C:/Program Files/R/R-4.0.4/library"'는 기록이 가능하지 않습니다
--- 현재 세션에서 사용할 CRAN 미러를 선택해 주세요 ---
URL 'https://cran.seoul.go.kr/bin/windows/contrib/4.0/binom_1.1-1.zip'을 시$
Content type 'application/zip' length 411521 bytes (401 KB)
downloaded 401 KB

패키지 'binom'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다

다운로드된 바이너리 패키지들은 다음의 위치에 있습니다
        C:\Users\HeoJiHae\AppData\Local\Temp\RtmpwT8YCq\downloaded_packages
> library("binom")
```

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용 1.6.1 비율에 대한 통계적 추론을 위한 R 이용

n = 1810명 중 837명이 찬성하는 이항 도수를 가지는 합법적인 낙태에 대한 의견을 나타내는 예제

양측,단측 유의성검정을 실시하는 방법과 모비율에 대한 신뢰구간을 구하는 방법

```
> prop.test(837,1810,p=0.50,alternative="two.sided",correct=FALSE)

        1-sample proportions test without continuity correction

data:  837 out of 1810, null probability 0.5
X-squared = 10.219, df = 1, p-value = 0.00139
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4395653 0.4854557
sample estimates:
        p
0.4624309

> prop.test(837,1810,p=0.50,alternative="less",correct=FALSE)

        1-sample proportions test without continuity correction

data:  837 out of 1810, null probability 0.5
X-squared = 10.219, df = 1, p-value = 0.0006951
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.0000000 0.4817492
sample estimates:
        p
0.4624309
```

<양측>

<단측>

스코어 신뢰구간

correct = FALSE : 연속성 수정을 사용 못함
=> 추정이 보수적으로 이루어지는 경향 때문

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용


```
> Clinical
   subject response
1        1        1
2        2        1
3        3        1
4        4        1
5        5        1
6        6        1
7        7        1
8        8        1
9        9        1
10      10        0
> attach(Clinical)
> y <- sum(response)
> prop.test(y,n=10,conf.level=0.95,correct=FALSE)

        1-sample proportions test without continuity correction

data:  y out of 10, null probability 0.5
X-squared = 6.4, df = 1, p-value = 0.01141
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5958500 0.9821238
sample estimates:
  p
0.9
```

성공의 수를 얻기 위해 0과 1 값을 합산

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용 1.6.1 비율에 대한 통계적 추론을 위한 R 이용

```
> prop.test(sum(Clinical$response),10,correct=FALSE)$conf.int
[1] 0.5958500 0.9821238
attr(,"conf.level")
[1] 0.95
> with(Clinical, prop.test(sum(response),10,correct=FALSE)$conf.int)
[1] 0.5958500 0.9821238
attr(,"conf.level")
[1] 0.95
```

자료 파일을 attach한 경우, 자료 파일 안에 있는 변수와 동일한 이름을 가지는 변수가 R 세션에 이미 정의되어 있을때 혼란을 유발할 수 있으므로 명령문 자체에 자료파일의 이름을 명시하거나 자료 파일의 이름과 함께하고자 하는 명령을 with 함수와 함께 적을 수 있다.

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용 1.6.1 비율에 대한 통계적 추론을 위한 R 이용

10번중 9번의 성공을 거두는 새로운 치료법을 평가하기 위한 임상 실험
귀무가설 : $H_0 : \pi = 0.50$
대립가설 : $H_a : \pi > 0.50$ 을 검정하기 위한 방법

10번중 9번의 성공한 결과를 제시하고자 한다.

```
> library(binom)
> binom.confint(9,10,conf.level=0.95,method="asymptotic")
      method x  n mean     lower    upper
1 asymptotic 9 10  0.9 0.7140615 1.085939
> binom.confint(9,10,conf.level=0.95,method="wilson")
  method x  n mean    lower     upper
1 wilson 9 10  0.9 0.59585 0.9821238
> binom.confint(9,10,conf.level=0.95,method="agresti-coull")
         method x  n mean     lower    upper
1 agresti-coull 9 10  0.9 0.5740323 1.003941
```

asymptotic : 왈드 신뢰구간

wilson : 스코어 구간

Agresti-Coull 구간

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용 1.6.1 비율에 대한 통계적 추론을 위한 R 이용

y = 9 와 y = 10일 때, $H_a: \pi \neq 0.50$와 $H_a: \pi > 0.50$ 에 대하여 $H_0: \pi = 0.50$ 에 대한 P-값을 찾을 수 있다.

```
> binom.test(9,10,0.50,alternative="two.sided")

        Exact binomial test

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5549639 0.9974714
sample estimates:
probability of success
                   0.9

> binom.test(9,10,0.50,alternative="greater")

        Exact binomial test

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.01074
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6058367 1.0000000
sample estimates:
probability of success
                   0.9
```

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용 1.6.1 비율에 대한 통계적 추론을 위한 R 이용

n = 10 의 시행 중 y = 9의 성공을 다음과 같이 구할 수 있다.

```
> library(exactci)
필요한 패키지를 로딩중입니다: ssanv

다음의 패키지를 부착합니다: 'exactci'

The following object is masked from 'package:binom':

    binom.exact

> binom.exact(9,10,0.50,alternative="greater",midp=TRUE)

        Exact one-sided binomial test, mid-p version

data:  9 and 10
number of successes = 9, number of trials = 10, p-value = 0.005859
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6504873 1.0000000
sample estimates:
probability of success
                   0.9
```

```
> library(PropCIs)
> midPci(9,10,0.95)



data:

95 percent confidence interval:
 0.5966 0.9946
```

중앙 P-값을 사용하는 이항검정과 중앙 P-값을 사용하는 이항신뢰구간은 exatci 패키지를 이용한다.
PropCIs 패키지도 사용 가능.

1.6 비율에 대한 통계적 추론을 위한 R 소프트웨어 사용 1.6.1 비율에 대한 통계적 추론을 위한 R 이용

베타(α,β)의 사전 분포와 y번의 성공, n-y번의 실패를 나타내는 베이지안 사후구간은
$\alpha^* = y + \alpha$ 와 $\beta^* = n - y + \beta$ 를 모수로 가지는 베타 분포의 분위수를 qbeta 분위 함수를 사용해서 구한다.

꼬리부분의 확률을 구하기 위해 pbeta 누적확률함수를 사용할 수 있다.

```
> qbeta(c(0.025,0.975), 837.5, 973.5)
[1] 0.4395369 0.4854450
> pbeta(0.50, 837.5, 973.5)
[1] 0.9993082
```

요약)

이항모수에 대한 추론을 실시하기 위해 여러 방법들을 사용할 수 있다.
분석사가 어떤 방법을 사용할지 결정하는 것이 쉽지 않을 수 있다.
현대의 빠른 컴퓨팅 연산력으로 더 이상 대표본 정규 근사 혹은 카이제곱 근사에 기초한 방법들에 의존하는 간단한 경우만 고려하지 않아도 된다.
유의성검정 혹은 신뢰구간에 대한 빈도론자 방법은 중앙 P-값을 사용하는 정확 이항 추론을 권장한다.
베이지안 방법은 베타(0.5,0.5) 사전분포에 의해 유도되는 베타 사후분포를 사용하는 추론을 권장한다.

끝 ~~!