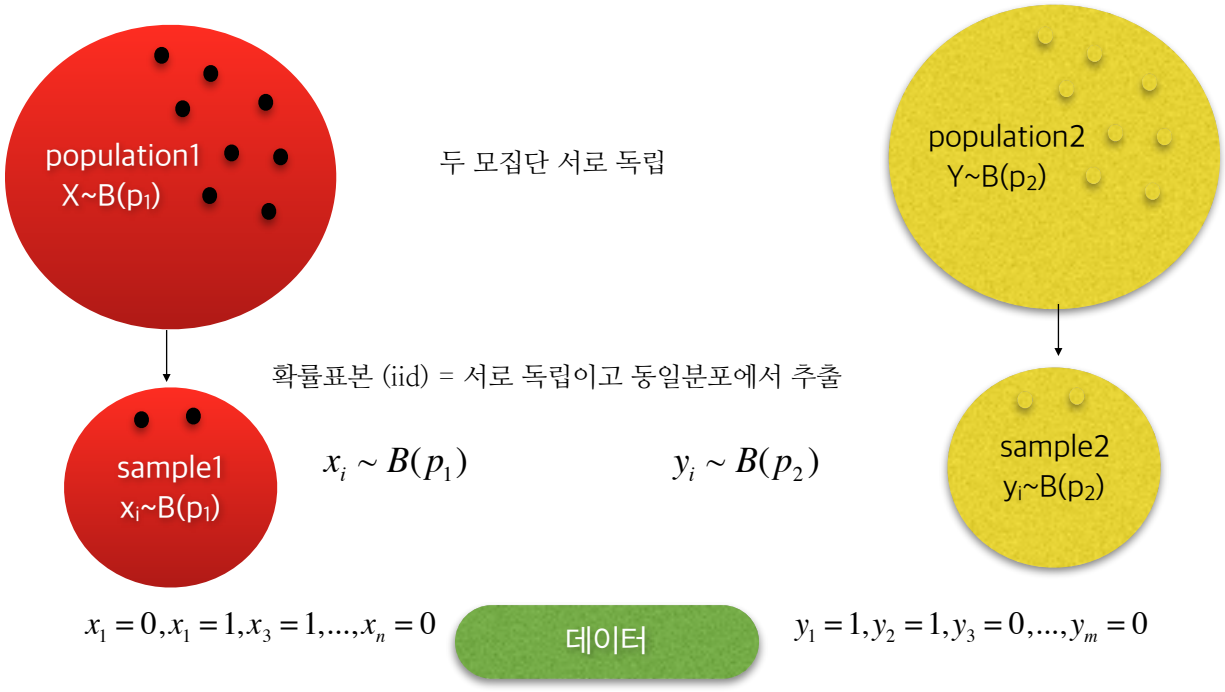
두 모집단 모비율 추론 개념



- 관심모수 : $\theta = (p_1 - p_2)$ – 두 모집단 비율 차이

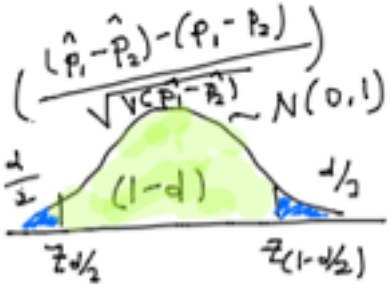
MVUE for $\theta = p_1 - p_2$

- 모수의 최소분산불편추정량은 표본비율의 차이
- $\hat{\theta} = \hat{p_1 - p_2} = \frac{\sum x_i}{n} - \frac{\sum y_i}{m} = \frac{\#of\ 1inx}{n} - \frac{\#of\ 1iny}{m}$
- MVUE=(표본비율 차이) 평균과 분산 : $E(\hat{\theta}) = p_1 - p_2, V(\hat{\theta}) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$

MVUE 샘플링분포

- $\sum x_i \sim B(n, p_1)$, $\sum y_i \sim B(m, p_2)$ 이고 독립이다.
- 대표본 이론 (Large Sample Theory) : 표본크기 n이 충분히 크면($\min(np, nq) \geq 5$) 표본 합(비율)의 분포는 정규분포에 근사한다.
- 그러므로 표본비율의 차이는 정규분포를 따른다. ~ $(\hat{p}_1 - \hat{p}_2) \sim N(p_1 - p_2, \frac{p_1 q_1}{n} + \frac{p_2 q_2}{m})$

- 피봇 통계량 : $\dfrac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\dfrac{p_1q_1}{n}+\dfrac{p_2q_2}{m}}} \sim N(0,1)$



100(1−α)% 신뢰구간

$$(\hat{p}_1-\hat{p}_2)-z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n}+\frac{\hat{p}_2\hat{q}_2}{m}} \le (p_1-p_2) \le (\hat{p}_1-\hat{p}_2)+z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n}+\frac{\hat{p}_2\hat{q}_2}{m}}$$

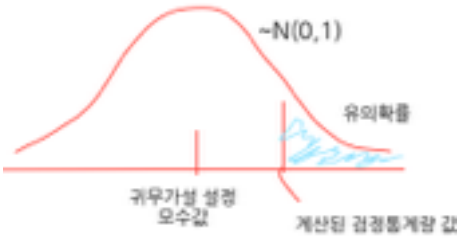
가설검정

1) 귀무가설 : $H_0 : p_1 - p_2 = 0$ (두 모집단 비율은 동일하다, 차이가 없다)

2) 대립가설 : $H_0 : p_1 - p_2 \neq 0$ (차이가 있다) – 양측 대립가설

대립가설 : $H_0 : p_1 - p_2 > 0$ (모집단1 비율이 모집단2보다 크다)

– 단측 대립가설



3) 검정통계량 : $TS = \dfrac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}\hat{q}(\frac{1}{n}+\frac{1}{m})}} \sim N(0,1)$, $\hat{p} = \dfrac{x+y}{n+m}$ –통합 비율

4) 유의확률 : $p(Z \le ts)$ (단측 왼쪽 대립가설, $H_0 : p_1 - p_2 < 0$) – alternative=c("less")

$p(Z \ge ts)$ (단측 오른쪽 대립가설, $H_0 : p_1 - p_2 > 0$) – alternative=c("greater")

양측 대립가설의 경우에는 위의 식 중 하나로 계산하여 2배함

예제 (홈런 경쟁) Keller "Managerial Statistics" 8th edition

세미소사와 맥과이어 홈런 경쟁으로 인하여 여성 팬이 증가하였다고 주장한다. 이를 알아보기 위하여 CNN/USA이 다음 조사를 하였다. 1995년 1008명 여성 중 413이 팬이라고 대답했고, 홈런 경쟁이 있은 1998년에는 1082 여성 중 681명이 팬이라고 답하였다. 이를 이용하여 주장에 대해 답하시오.

1. 연구문제 및 통계적 문제 정의

- 1998년 여성 팬 비율이 1995년보다 높다면 여성 팬이 증가하였다고 결론 내릴 수 있음
  - 독립인 두 모집단 비율 차이 검정 문제

2. 수집 데이터 정리 및 검증

- 1995년 n=1,008명 중 413명이 팬이라고 응답

- 1998년 n=1,082명 중 681명이 팬이라고 응답

3. 통계적 가설 설정

- 귀무가설 : 1995년 여성팬 비율과 1998년 여성팬 비율은 동일하다. $p_1 - p_2 = 0$
- 대립가설 : 1995년 여성팬 비율과 1998년 여성팬 비율은 차이가 있다. $p_1 - p_2 \neq 0$

4. 검정통계량 및 유의확률 계산

```
baseball=matrix(c(413,595,681,401),ncol=2,byrow=TRUE) #데이터 행렬로 입력함
colnames(baseball)=c("Yes","No")
rownames(baseball)=c("1995","1998")
as.table(baseball)
prop.test(as.table(baseball), correct=FALSE) #Yeates 연속보정 하지 않음
sqrt(100.95) #카이제곱 통계량이므로 정규분포 통계량은 제곱근 해야 함
```

```
> as.table(baseball)
     Yes  No
1995 413 595
1998 681 401
> prop.test(as.table(baseball), correct=FALSE)

        2-sample test for equality of proportions without
        continuity correction

data:  as.table(baseball)
X-squared = 100.9463, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2614987 -0.1778369
sample estimates:
   prop 1    prop 2
0.4097222 0.6293900
```

Yeates Continuity Correction : 이항분포는 이산형, 정규분포는 연속형이므로 이항분포를 연속형인 정규분포를 이용하여 확률을 계산하는 경우 연속 보정이 필요함

(이유) 임의의 한 값에서 이산형 확률은 0보다 크지만 연속형은 0이다.

(방법) $p(X \leq 10 \mid X \sim B) \approx p(X \leq 10.5 \mid X \sim N)$, $p(X \geq 10 \mid X \sim B) \approx p(X \geq 9.5 \mid X \sim N)$

- 검정 통계량 : $TS = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(\frac{1}{n} + \frac{1}{m})}} = \sqrt{100.94} = 10.04 \sim N(0,1)$, $\hat{p} = \dfrac{x+y}{m+n}$ (왜냐하면 귀무가설 하에서는 모집단 비율이 동일하므로 모비율이 같은지에 대한 검정에서는 이를 사용해야 한다.

- 유의확률 〈 0.001 – 매우 유의, 귀무가설 기각 – 1995년 여성 팬 비율과 1998년 여성 팬 비율은 차이가 있다. 1998년 여성팬 비율이 높으므로 결론적으로 소사와 맥과이어 홈런 경쟁으로 여성 팬이 늘었다고 할 수 있다.
- 95% 신뢰구간 : $-1.96 < \frac{(0.4097-0.6293)-(p_1-p_2)}{\sqrt{(\frac{(0.4097)(1-0.4097)}{n=1008}+\frac{(0.6293)(1-0.6293)}{m=1082})}} < 1.96$ – 신뢰구간 구할 때

  는 분모 아래가 (p_1, p_2)를 사용해야 하나, 계산이 복잡해져 그냥 $\hat{p}_1, \hat{p}_2$ 사용한다.

5. 결론 및 활용

1995년 여성팬 비율	1998년 여성팬 비율	검정통계량	유의확률	95% 신뢰구간
0.4097	0.6293	10.04	〈0.001	(0.36, 0.42)

- 귀무가설 기각 : 매우 유의, 홈런 경쟁으로 여성팬이 증가하였다고 할 수 있다.
- (활용) 이슈를 만들 수 있는 경기가 팬들을 만들 수 있다.

문제 #1 (Keller 9판 12.5 예제문제) package.csv

비누 회사에서 포장지 디자인을 2개 개발하였다. 고객들이 어느 디자인을 선호하는지 알아보기 위하여 자회사 비누 판매율이 비슷한 두 수퍼마켓을 임의 선택하여, 수퍼마켓에 각각 서로 다른 디자인의 비누를 진열하여 판매비율을 조사하였다. 이 회사 비누의 바 코드 “9077”이다. 데이터에는 각 슈퍼마켓에서 팔린 비누의 코드가 있다.

(1) 고객은 어느 디자인을 선호하는가? 유의수준 = 5%

(2) 슈퍼마켓 1의 디자인이 슈퍼마켓 2의 디자인보다 3% 이상 선호되고 있는가? 유의수준 5%

(3) 두 디자인의 판매비율 차이의 95% 신뢰구간을 구하시오.