모집단에 관한 추론

모비율에 관한 추론

성공횟수와 성공비율 표본분포의 정규분포 근사

성공확률이 p인 베르누이 실험을 반복하는 베르누이 과정 $\{X_n; n=1, 2, \ldots\}$을 생각하자. 베르누이 확률변수 X_n의 평균은 p이고 표준편차는 $\sqrt{\pi(1-\pi)}$ 이다. 이는 시행을 n번 반복하는 베르누이 과정에 표본분포의 모집단분포이다.

- 경험법칙에 따르면 $n\pi \geq 10$이고 $n\pi(1-\pi) \geq 10$이면, 성공횟수와 성공비율 표본분포의 근사분포로서 정규분포를 사용할 수 있다.
- 성공횟수 $N_n = X_1 + X_2 + \ldots + X_n$ 의 표본분포는 이항분포를 따른다. 즉 $N_n \sim B(n, \pi)$. 성공횟수 표본분포의 평균은 $n\pi$ 이고 표준편차는 $\sqrt{n\pi(1-\pi)}$ 이다. 경험법칙을 만족하면 정규분포로 근사한다. 즉 $N_n \sim N\left(n\pi, \sqrt{n\pi(1-\pi)}^2\right)$ 이다..
- 성공비율 $p=N_n/n$ 표본분포의 평균은 π 이고 표준편차는 $\sqrt{\pi(1-\pi)}/\sqrt{n}$이다. 경험법칙을 만족하면 정규분포로 근사한다. 즉 $p \sim N\left(\pi, \sqrt{\pi(1-\pi)/n}^2\right)$ 이다.

모비율에 관한 추론

◆ 모비율의 점 추정치

- $p = N_n / n$ (N_n 은 표본중 관심있는 사건의 발생횟수, n은 표본의 수)

◆ 모비율의 구간 추정과 신뢰구간

연구 문제 8-1-1
대통령 선거 직전에 한 방송사가 실시한 여론조사결과는 다음과 같았다고 하자. 조사대상자 2,500명 중 1,300명이 야당 후보를 지지한다고 대답했고, 나머지 1,200명이 여당 후보를 지지한다고 답변하였다. 유권자 전체의 야당 후보에 대한 지지율을 추정하고, 95% 신뢰구간을 구하여라.

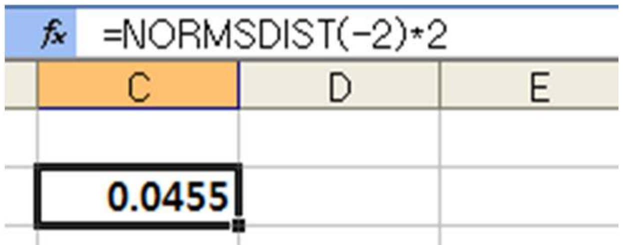
$$p = \frac{N_n}{n} = \frac{1{,}300}{2{,}500} = 0.52 \text{ 혹은 } 52\%$$

모비율에 관한 추론

◆ **표본비율의 표본오차**

- $\sqrt{\pi(1-\pi)/n} \quad \rightarrow \sqrt{p(1-p)/n} = \sqrt{\frac{0.52(1-0.52)}{2500}} = 0.01$

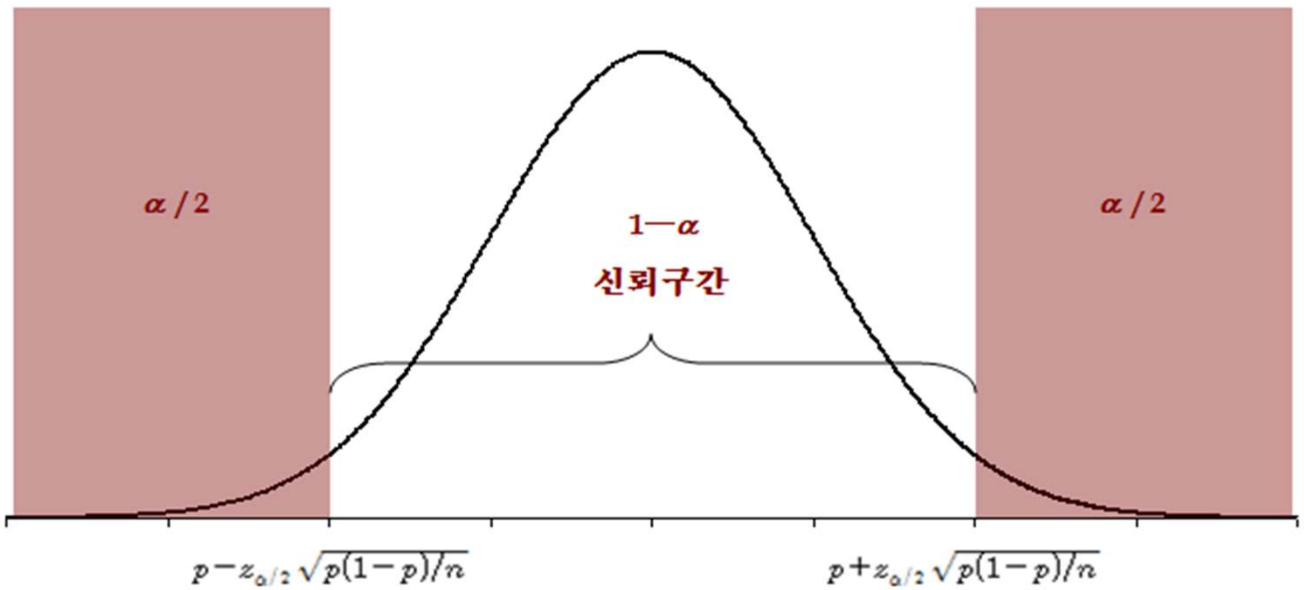
- 표본이 충분히 크면 $p \sim N\,(0.52,\ 0.01^2)$
- 실제 투표결과 만일 52%에서 2%이상 차이날 확률은?
  - ❖ $P(\pi < 0.5$ 또는 $\pi > 0.54) = P(Z < -2$ 또는 $Z > 2) = 2 * P(Z < -2)$



모비율에 관한 추론

◆ **π에 대한 신뢰구간 추정**

- $\mathrm{p} \sim N\left(p, \sqrt{\pi(1-\pi)/n}^{2}\right)$으로 해야 하지만
- $\mathrm{p} \sim N\left(p, \sqrt{p(1-p)/n}^{2}\right)$ 또는 $\mathrm{Z} = \frac{p-\pi}{\sqrt{p(1-p)/n}}$ 을 대신 사용
- 따라서 신뢰구간은 $\left(p - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < \pi < p + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right)$



모비율에 관한 추론

π 에 대한 신뢰구간 추정량

표본이 충분히 클 때, π에 대한 $(1-\alpha)\times100\%$ 신뢰구간 추정량은 다음과 같다.

$$p \pm z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$$

- ❑ 90% 신뢰구간: $Z_{0.050}=1.645$
- ❑ 98% 신뢰구간: $Z_{0.010}=2.326$
- ❑ 95% 신뢰구간: $Z_{0.025}=1.960$
- ❑ 99% 신뢰구간: $Z_{0.005}=2.576$

엑셀 활용

표본이 충분히 클 때, p에 대한 $(1-\alpha)\times100\%$ 신뢰구간은 다음과 같다.

=*p*-NORMSINV(1-α/2)*SQRT(*p**(1-*p*)/*n*)

여기서 =SQRT는 제곱근을 구하는 엑셀함수이다. 선거 문제로 예시하면 다음과 같다.

95% 신뢰구간의 하한: =0.52-NORMSINV(0.975)*SQRT(0.52*(1-0.52)/2500)

95% 신뢰구간의 상한: =0.52+NORMSINV(0.975)*SQRT(0.52*(1-0.52)/2500)

모비율의 검정과 신뢰구간

대통령 선거 직전 2,500명을 대상으로 한 여론조사에서, 야당 후보에 대한 지지율은 52%이고 여당 후보에 대한 지지율은 48%로 나타났다. 그러면 이러한 차이가 통계적으로 유의한 차이인가? 유권자 전체 모집단의 여당 후보에 대한 지지율을 π라고 하면, 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \pi = 0.5 \qquad\qquad H_1 : \pi \neq 0.5$$

귀무가설이 맞는다는 가정 하에서 검정통계량의 값은

$$\mathrm{Z} = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}} = \frac{0.52 - 0.5}{\sqrt{0.5(1-0.5)/2500}} = 2$$

2라는 값이 나올 가능성, 즉 p-value는 0.046(=Pr(Z<-2 또는 Z>2))

모비율의 검정과 신뢰구간

연구 문제 8-1-2

대통령 선거 직전에 한 방송사가 실시한 여론조사결과는 다음과 같았다고 하자. 조사대상자 2,500명 중 1,300명이 야당 후보를 지지한다고 대답했고, 나머지 1,200명이 여당 후보를 지지한다고 답변하였다. 이 여론조사결과를 근거로 여당 후보가 승리할 것이라고 결론지을 수 있는가?

표본크기

- **오차한계 E=** $Z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$

$$\left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p)$$

- **표본의 크기 n=**$\boldsymbol{\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2 \times p(1-p)}$

신문이나 텔레비전 같은 대중매체의 여론조사에서는 p = 0.5라고 가정하고 모비율을 추정한다. 일반적으로 95%의 신뢰수준에서 오차한계를 ±3% 이내로 설정한다. 따라서 필요한 최소한의 표본크기는

$$\mathrm{n} = \boldsymbol{\left(\frac{1.96}{0.03}\right)^2 \times 0.5(1-0.5) = 1067.1}$$

이므로 1,068 개가 된다.

만일 E가 0.03에서 0.01로 더 정확해 지려면 표본의 수는?

모평균의 추정과 STUDENT'S T-DISTRIBUTION

◆ **표본평균의 분포(Case 1: σ를 알고 n이 충분히 큰 경우)**

● $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

◆ **표본평균의 분포(Case 2: σ를 모르고 n이 충분히 큰 경우(n>=30))**

● $\bar{X} \sim N\left(\mu, \frac{S^2}{n}\right)$

◆ **표본평균의 분포(Case 3: σ를 모르고 n이 작은 경우(n<30))**

● $\bar{X} \sim t\left(\mu, \frac{S^2}{n}\right)$

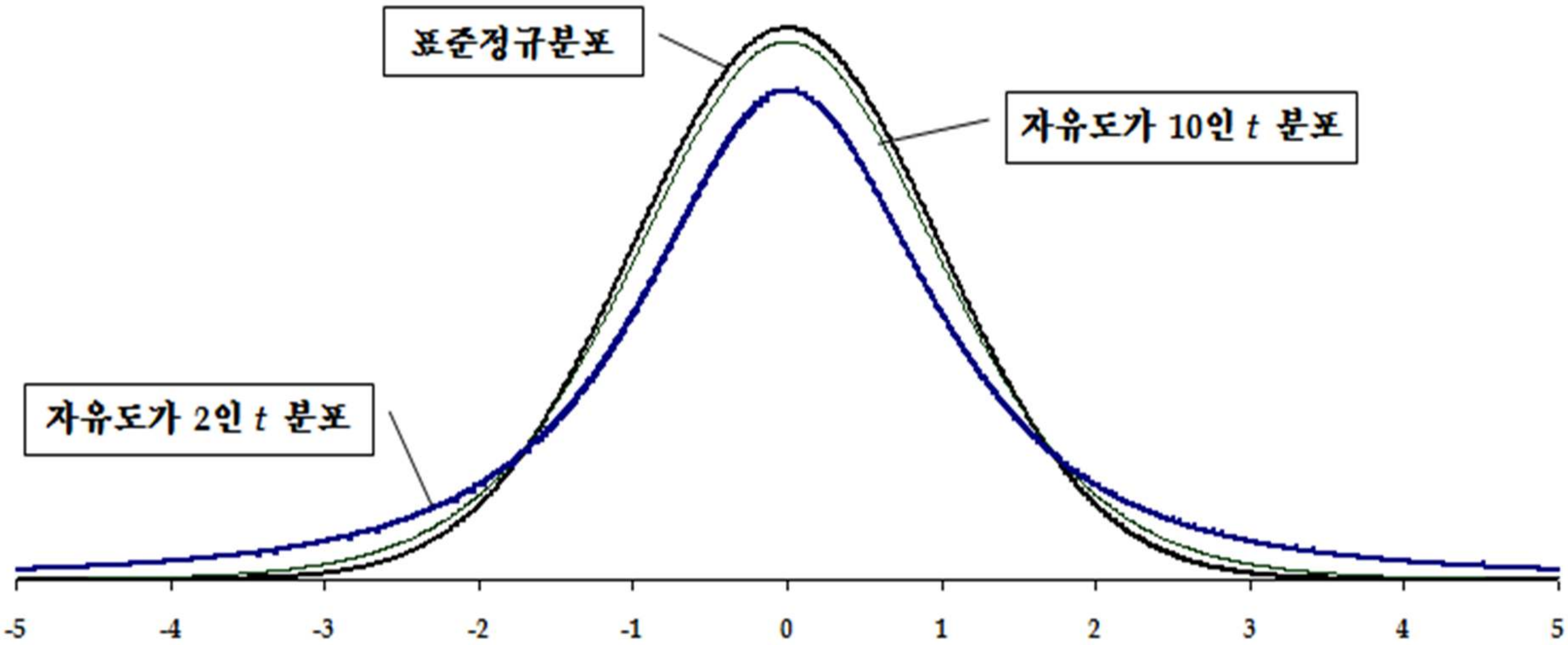
◆ **유의 수준 α에서의 신뢰구간**

● Case 1 : $\mu \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$, Case 2: $\mu \pm z_{\alpha/2} \times \frac{S}{\sqrt{n}}$, Case 3: $\mu \pm t_{\frac{\alpha}{2};n-1} \times \frac{S}{\sqrt{n}}$

◆ **유의 수준 α에서의 검정통계량**

● Case 1 : $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, Case 2: $Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$, Case 3: $t = \frac{\bar{X}-\mu}{S/\sqrt{n}}$

T-분포와 표준정규분포



T-검정 예

엑셀 활용

확률변수 T는 자유도가 df인 스튜던트 t 분포를 따른다고 하자. 그러면

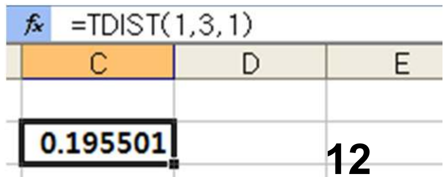
- ❑ $Pr(T > t)$를 구할 때, =TDIST $(t, df, 1)$
- ❑ $Pr(T < -t \text{ or } T > t)$를 구할 때, =TDIST $(t, df, 2)$

이제 정규분포를 따르는 모집단으로부터 작은 표본을 추출한 경우, 표준정규분포와 스튜던트 t 분포의 확률값이 어떻게 다른지 비교해보자. 술이라면 사족을 못 쓰는 50명의 애주가들의 모임이 있다. 이들이 한 번 모일 때, 마시는 소주의 양은 정규분포를 따른다고 하자. 최근 4번의 모임에서, 마신 소주병의 수는 다음과 같았다.

109 94 89 108

4개 표본의 평균은 100병이고 표준편차는 10병이다. 그러면 표본평균이 105병보다 클 확률은 얼마인가? 모평균은 표본평균 100병으로 추정할 수 있다. 모표준편차를 표본표준편차로 추정하였으므로

$\mathrm{t} = \frac{\bar{X}-100}{10/\sqrt{4}} = \frac{105-100}{10/\sqrt{4}} = 1$, $Pr\,(T > 1)$=



연구 문제 8-2-3

이제까지 여성의 T-값을 산출하는 기준으로 평균 1.12g/㎠과 표준편차 0.12g/㎠이 사용되어 왔다. 골밀도가 정상인 20~30대의 여성들 중 9명을 표본으로 추출한 결과, 표본평균 = 1g/㎠이고 표준편차 s = 0.12g/㎠이었다. 이를 근거로 평균이 변했다고 말할 수 있는가?

가설검정을 위한 귀무가설과 대립가설은 다음과 같다

$$H_0 : m = 1.12 \qquad H_1 : m \neq 1.12$$

귀무가설이 맞는다는 가정 하에서 검정통계량의 값은?

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{1 - 1.12}{0.12/\sqrt{9}} = -3$$

양측 검정의 p-값은 $Pr\ (T < -3 \text{ or } T > 3)$