사회통계
통계적 추정

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

추정의 논리

추정의 논리

앞서 배운 표집분포의 이론은 현실 통계 분석의 목적과는 동떨어져 있다.

- 표집분포의 이론은 일단 모집단의 평균 μ와 표준편차 σ를 안다고 전제하고 있었다.
- 가령 "A학교 전교생의 IQ 점수 분포가 $\mu = 105$, $\sigma = 15$인 정규분포일 때, 25명 임의표본의 평균이 108 이상으로 나올 확률은 얼마인가?"

$$P(\bar{X}) = P\left(Z > \frac{108 - 105}{15/\sqrt{25}}\right) = 1 - P(Z < 1) = 0.159$$

- "그럴 확률은 15.9%이다."

추정의 논리

- 그런데 가만 생각해보면 이건 완전 웃긴 이야기다!
- "이미 모집단의 평균(μ)과 표준편차(σ) 같은 걸 다 아는데 뭐하러 표본을 새로 수집하나?"
- "아니, 그건 좋다구 치자. 표본을 무한히 뽑는다니 미친거 아냐? 차라리 모집단을 전수 조사하고 말지!"
- 현실의 통계분석에서는 오히려 "반대로" 표본에 근거해 바로 그 모집단의 μ와 σ를 추론(inference)해야 할 필요가 있다.

추정의 논리

표본으로부터 모집단의 성격을 추론하는 방식에는 두 가지가 있다.

- 하나는 오늘 배울 추정(estimation)이고, 다른 하나는 나중에 배울 가설검정(hypothesis test)이다.
- 다시 추정(estimation)은 점 추정(point estimation)과 구간 추정(interval estimation)으로 나뉜다.

추정의 논리

먼저 추정에 관한 기초적인 용어를 살펴보자.

맥주병 회사에서 일하는 고래는 제조된 병 가운데 임의표본($n = 30$)에서 평균값을 확인해 보았다.

"음, 병 하나당 용량 평균은 330 ml로군."

이를 들은 새우가 캐물었다.

"확실해?"

쫄린 고래는 얼버무렸다.

"아니, 꼭 그렇진 않고... 90% 신뢰구간은 313.6 ml에서 346.4 ml 사이야."

추정의 논리

- (처음에 고래가 그러했듯) 표본의 평균인 330 ml를 가지고 모집단 평균을 예측했다면, (그래프 상에 점을 찍듯) 하나의 숫자로 구한 값이므로 점 추정값(point estimate)이다.
- 나중에 고래가 "313.6 ml에서 346.4 ml 사이"라고 말한 것은 구간 추정값(interval estimate)이 된다.
- 90% 신뢰구간(confidence interval; CI)은 구간 추정값의 신뢰할 수 있는 폭에 관한 것이다.
- 추정값(estimate)이란 추정된 구체적인 값이고, 추정값(estimate)을 구하는데 사용되는 통계량(statistic)을 추정량(estimator)이라고 부른다.
- (앞서 고래와 새우의 대화에서) 330 ml라는 구체적인 수치는 추정값(estimate)이고, 병에 담긴 용량의 '표본평균(sample mean)'이 곧 추정량(estimator)이다.

점 추정

점 추정

대체 어떻게 고래는 점 추정값을 제시할 수 있었나?

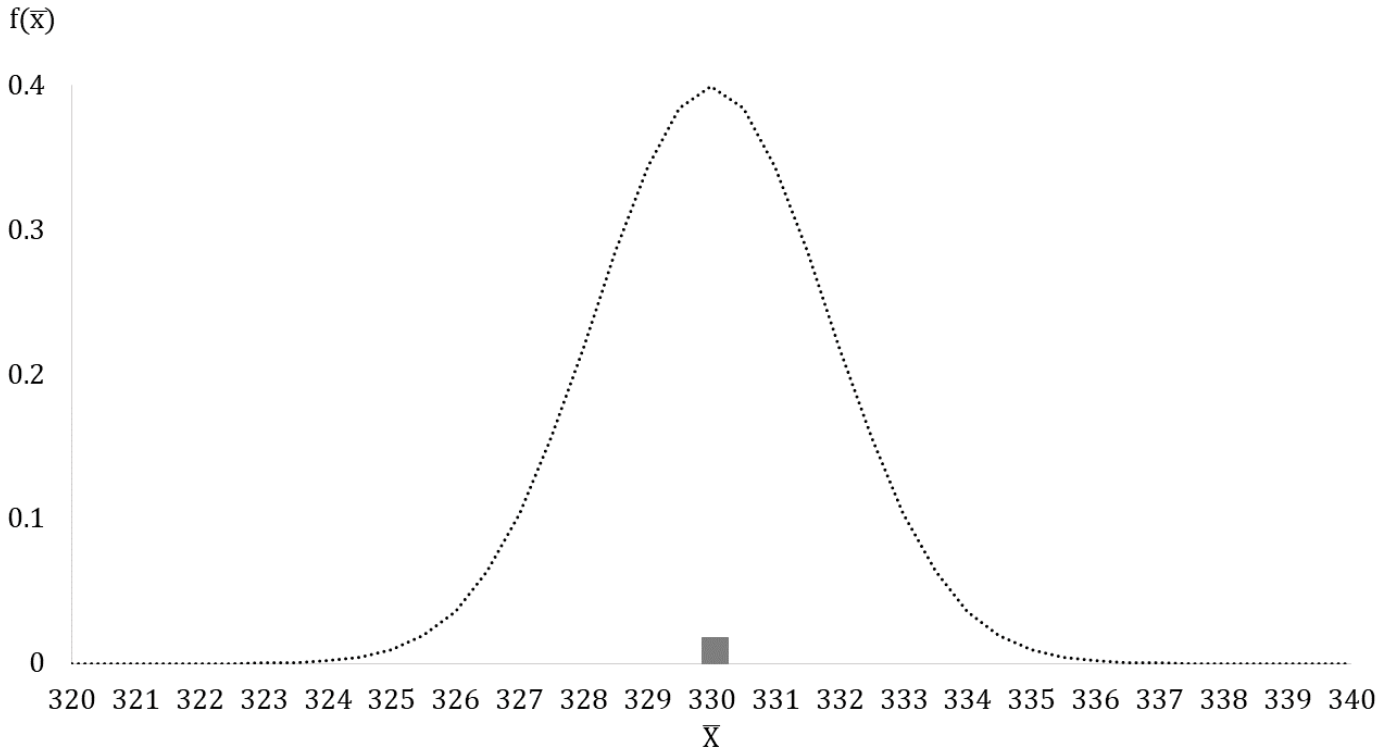
- 중심극한정리를 통해 (모집단의 분포와 무관하게) 표본 크기만 충분히 크면 다음이 성립함을 알 수 있다.

$$\mu_{\bar{X}} = \mu \qquad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- 그런데 표집분포의 이론을 거꾸로 생각해보면, (1) 설령 모평균 μ에 대한 정보가 사전에 주어지지 않았고, (2) 표본을 무한히 뽑지 않았더라도,
- 내가 지금 가지고 있는 (하나의 표본로부터 얻은) 평균값(a sample mean)은 (굳이 어느 쪽인가 하면) μ를 반영하고 있을 가능성이 제일 높다!
- 다시 말해, 임의표본을 한 번만 뽑고 거기서 평균을 구해 $\bar{x}$를 얻었다면 일단 이것이 모평균에 대한 최선의 추정량(best estimator)인 것이다!

점 추정

- 만약 고래가 무한히 표본을 추출하여(단 $n \geq 30$) 그 표본평균들의 확률분포를 구해보면 다음 점선과 같을 것이다.
- 그렇기에 만일 하나의 표본평균을 구한 뒤, 그것이 표집분포 위의 어디에 위치할 것인지 딱 한군데만 판돈을 건다면 μ라고 추측하는 것이 최선이다(Why?).



점 추정

대체 어떻게 고래는 90% 신뢰구간을 제시할 수 있었나?

- 다시 그림을 다시 잘 보면, 나의 표본평균에서 아주 살짝 옆으로 이동한 값이 진짜 모평균일 확률도 엄청 높긴 하다. 아무래도 전재산을 걸기엔 틀릴까봐 겁난다.
- 그러니 확률이 높은 순서대로 90%의 면적을 채워나가 보면 어떨까?
- 확률이 높은 쪽은 가운데에 몰려있으니 양 끄트머리를 빼고 90%를 채우는 것이 상식적이다(Why?).
- 정규분포를 직접 그려보고 색칠 공부를 해보자.
- 그 안에 포함되는 $\bar{x}$값들이 바로 90% 신뢰구간의 폭을 결정한다.

점 추정

- 양 끄트머리 5%씩을 빼고 가운데 90%의 면적을 계산하는 것이라면 엑셀로 할 수 있다.
- (모집단에서 계산된 평균 μ와 표준편차 σ가 아니라) 표본평균 $\mu_{\bar{X}}$와 표준오차 $\sigma_{\bar{X}}$를 가지고 Z 값으로 표준화하자.

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma/\sqrt{n}}$$

- 확률을 주고 Z 값을 받는 엑셀 함수는 NORM.INV$(\cdot)$ 함수임을 배웠다.
- NORM.INV(0.95, 0, 1)로 오른쪽 끄트머리 경계 Z 값을,
  NORM.INV(0.05, 0, 1)로 왼쪽 끄트머리 경계 Z 값을 구할 수 있다(Why?).

점 추정

- 엑셀에서 실제로 NORM.INV(0.95, 0, 1)와 NORM.INV(0.05, 0, 1)를 구해보면 각각 1.64와 -1.64가 나온다.
- 물론 이 값들은 표준화된 Z 값이므로 직관적으로 해석이 안된다. 다시 원점수로 돌리는 방법도 전에 배웠다.

$$\bar{X} = Z \cdot \sigma_{\bar{X}} + \mu_{\bar{X}}$$

- 그러므로 90% 신뢰구간은 다음과 같다.

$$\begin{aligned} & [Z_{0.05} \cdot \sigma_{\bar{X}} + \mu_{\bar{X}}, && Z_{0.95} \cdot \sigma_{\bar{X}} + \mu_{\bar{X}}] \\ =& [-1.64 \cdot 10 + 330, && 1.64 \cdot 10 + 330] \\ =& [313.6, && 346.4] \end{aligned}$$

점 추정

90% 신뢰구간은 어떻게 해석해야 할까?

- "[313.6, 346.4] 사이에 모집단의 평균이 놓일 확률이 90%이다."
  → 이것은 신뢰구간의 잘못된 해석이다(Why?).
- "표본을 무한히 추출했다면 각각의 90% 신뢰구간들(confidence intervals)도 무한히 구할 수 있다. 이 모든 신뢰구간들의 90%는 모평균인 330을 포함하고 있다."
  → 이것이 신뢰구간의 올바른 해석이다.

점 추정

첫번째 해석은 왜 틀릴까?

- 그것은 모평균 μ는 알려지지 않았을 뿐, 상수(constant)이기 때문이다. 모평균은 확률변수가 아니다(Why?).
- 오히려 확률변수인 쪽은 표본평균이므로, 표집마다 매번 달라지는 것은 (모평균이 아니라) 신뢰구간이다.
- 따라서 해석할 때 "기준이 되는 쪽"은 당연히 모평균 μ이고, 90%로 이를 맞출 수 있는 것은 신뢰구간들이다.

점 추정

개념을 이해했다면 이제 수식으로도 나타낼 수 있다.

- 90% 신뢰구간은 다음과 같이 수학적으로 표현할 수 있다.

$$P(Z_{0.05} \leq \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \leq Z_{0.95}) = .90$$

또는,

$$P(\bar{X} - Z_{0.95} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} - Z_{0.05} \cdot \sigma_{\bar{X}}) = .90$$

또는,

$$P(\bar{X} + Z_{0.05} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + Z_{0.95} \cdot \sigma_{\bar{X}}) = .90$$

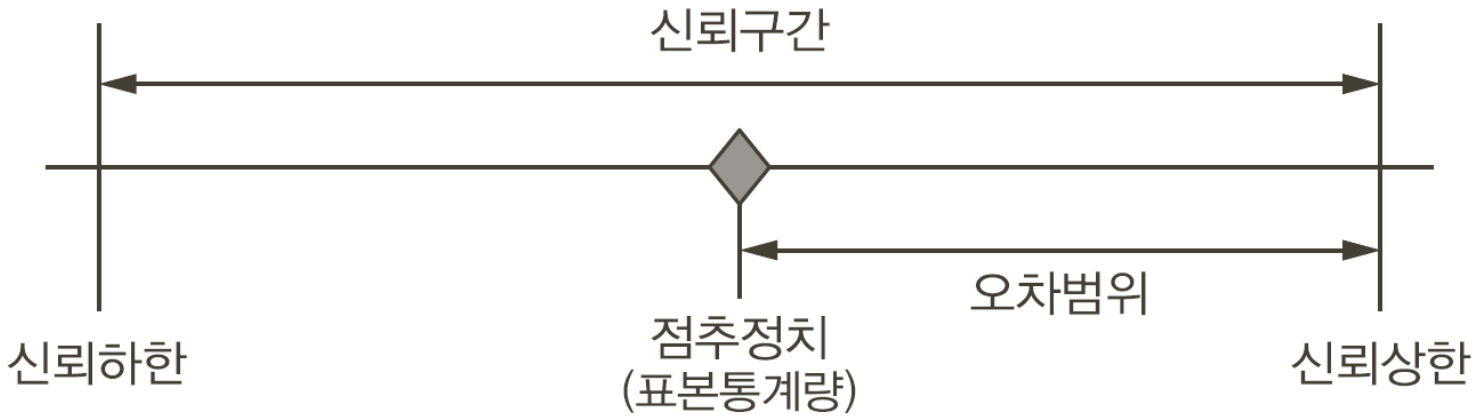
- (혼동스럽게 보이겠지만) 세 수식은 모두 같은 것이다(Why?).

점 추정

- 신뢰구간을 나타낼 때 해석과 일치시키기 위해서 반드시 기준점을 $\bar{X}$가 아닌 μ로 삼아야 한다!
- 이때, 우변의 90%라는 기준을 신뢰수준(confidence level)이라고 부른다. 또 90% 말고 95%나 99% 등 다양한 신뢰수준을 사용할 수도 있다.
- 1에서 신뢰수준을 뺀 값을 유의수준(significance level) α라고 부른다. 또 10% 말고 5%나 1% 등 다양한 유의수준을 사용할 수도 있다.
- 위에서 $Z \cdot \sigma_{\bar{X}}$ 부분을 특별히 오차범위(margin of error)라고 부른다.

점 추정

- 점 추정값, 신뢰구간, 오차범위 등 용어가 혼동스러울 수 있으므로 세심히 공부하자.



점 추정

놓치기 쉬운 사실은 신뢰구간이 추정의 오류에 관해 말해준다는 점이다.

- 우리는 현실적으로 모집단 μ를 모른다. 표본을 무한히 뽑을 수도 없다. 그래서 하는 수 없이 표본을 한 번만 추출해서 $\bar{x}$를 계산했다.
- "과연 μ를 추정하기에 그 $\bar{x}$는 얼마나 믿을 만한가?"
- 이 질문이 무엇을 묻고 있는지 곰곰이 생각해보자!
- 애초에 신뢰구간 추정의 목적은 "표본에서의 계산된 우리의 통계량(statistic) $\bar{x}$가 얼마나 믿을만한가(reliable)?"를 파악하는데 있는 셈이다.

점 추정

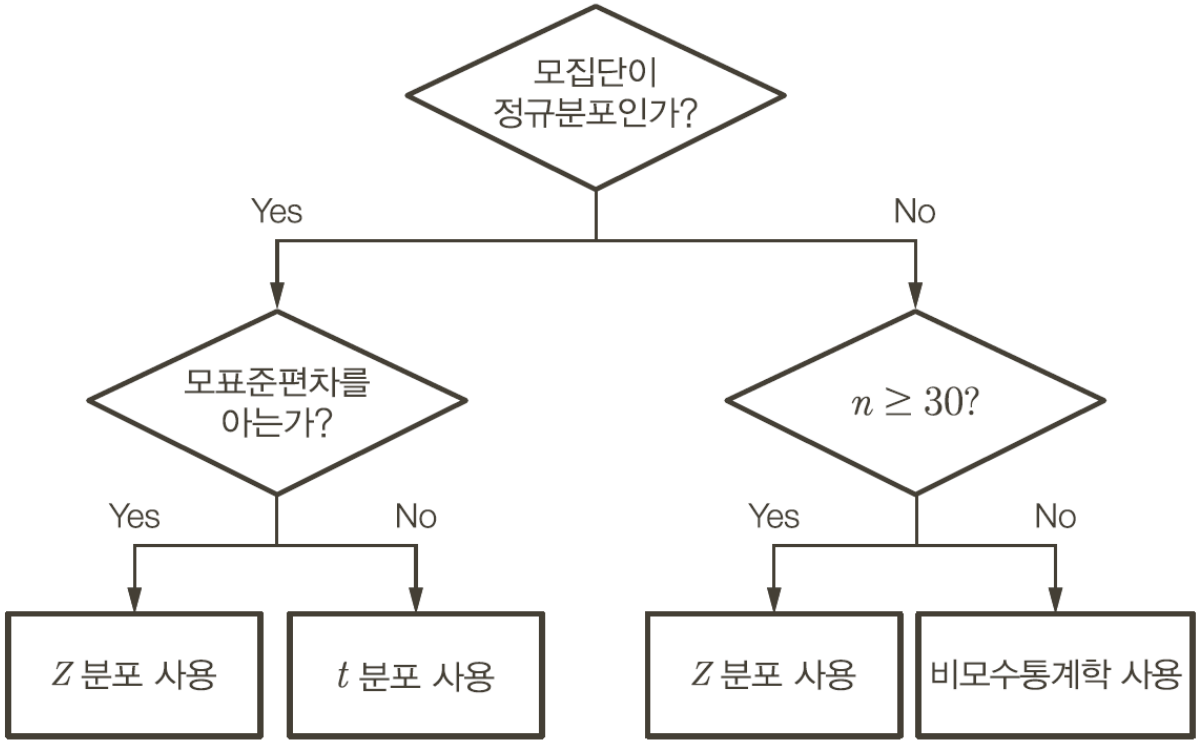
- $\sigma_{\bar{X}}$가 작다면 우리의 표본은 $\mu_{\bar{X}}$을 중심으로 아주 밀집되어 있다는 의미이므로, $\mu_{\bar{X}}$는 꽤 믿을만 할 것이며, 90% 신뢰구간도 좁을 것이다.
- $\sigma_{\bar{X}}$가 크다면 우리의 표본이 $\mu_{\bar{X}}$가 아닌 여러 값들로 퍼져있다는 의미이므로, $\mu_{\bar{X}}$는 믿기 어려울 것이며, 90% 신뢰구간도 넓을 것이다.
- 각각의 상황에 따른 가상의 표집분포를 그리고 색칠 공부도 해보자!
- 다시 말하지만, 표준오차 $\sigma_{\bar{X}}$는 표본을 이용하여 추정할 때 예상되는 "오류의 크기"를 나타낸다.

모평균의 신뢰구간

모평균의 신뢰구간

지금까지 공부한 내용이 사실 모평균의 신뢰구간의 핵심이다.

- 모평균의 신뢰구간(confidence intervals for the population mean)은 구체적인 조건에 따라 접근방법이 살짝 달라진다.



모평균의 신뢰구간

예제 4. 로즈는 자신의 모교에서 사회통계의 중간시험 평균점수를 추정하고자 한다. 그녀는 사회통계 과목 전체 이수자 가운데 30명의 표본을 임의로 추출하였고, 이 표본에서 평균은 75점이고 표준편차는 10점임을 확인하였다. 모평균의 95% 신뢰구간을 구하시오.

모평균의 신뢰구간

- 가장 먼저 문제가 무엇을 묻고 있는지 파악해야 한다.

$$P(? \leq \mu \leq ?) = .95$$

- 다음에 문제에서 주어진 정보를 정리해보자.

$$\mu = ?, \quad \sigma = ?, \quad \bar{x} = 75, \quad s = 10, \quad n = 30$$

- $n \geq 30$이므로 중심극한명제에 따라 로즈의 표집분포는 정규분포함을 알 수 있다.
- 이 표집분포의 평균 $\mu_{\bar{X}}$는 무엇인가? 주어진 (한 번의) 표본평균 $\bar{x} = 75$가 현실적으로 모평균 μ에 대한 최선의 추정량(best estimator)이다. 그러므로,

$$\mu_{\bar{X}} = \mu = 75$$

- 이 표집분포의 표준오차 $\sigma_{\bar{X}}$는 무엇인가? 주어진 (한 번의) 표본의 표준편차 s는 10점이고 표본 크기는 $n = 30$이므로,

$$\sigma_{\bar{X}} = \frac{10}{\sqrt{30}}$$

모평균의 신뢰구간

- 일단 (가상적인) 표본평균의 확률분포, 즉 표집분포를 그려본 뒤 구하려는 신뢰구간 95%에 대해 대충 색칠 공부를 해보자.
- 정확한 95% 면적을 결정하는 Z 값은 엑셀의 NORM.INV(0.975, 0, 1)과 NORM.INV(0.025, 0, 1)로 특정한다(Why?).
- 이제 그 값들을 x축의 Z 값에 표시하자.
- 아까 NORM.INV$(\cdot)$ 함수를 통해 구한 값들은 Z 값이므로 다시 원점수로 환원해야 한다($X = \mu + Z \cdot \sigma_{\bar{X}}$).
- 환원된 그 원점수들을 다시 원점수 축에 표시하고 신뢰구간을 보고한다.

$$P(\bar{X} + Z_{0.025} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + Z_{0.975} \cdot \sigma_{\bar{X}}) =$$

$$P(75 - 1.96 \cdot \frac{10}{\sqrt{30}} \leq \mu \leq 75 + 1.96 \cdot \frac{10}{\sqrt{30}}) =$$

$$P(71.42 \leq \mu \leq 78.58) = .95$$

모비율의 신뢰구간

모비율의 신뢰구간

이제 모비율의 신뢰구간에 대해서 이야기할 차례이다.

- 앞서 "표본평균"과 "표본비율"의 표집분포들을 대조하여 이야기한 것과 마찬가지로, "모평균의 신뢰구간"과 모비율의 신뢰구간(confidence intervals for the population proportion)을 대조할 수 있다.
- 현실에서 모비율을 추정해야 하는 사례는 얼마든지 있다:

  (1) 학자금 채무불이행 비율
  (2) 비영리 단체에서 기부요청 이메일을 보내자 이에 화답한 회원의 비율
  (3) 제조과정에서 불량품 발견률
  (4) 길거리에서 조사한 무단횡단의 비율

모비율의 신뢰구간

다행히 아까와 마찬가지로 논리 구조는 거의 똑같다.

- 현실적으로 (1) 우리는 모비율 π를 알지 못한다. (2) 표본도 무한히 뽑을 수 없다.
- 대신 크기가 n인 임의표본을 한 번만 수집하여, 그로부터 표집분포의 비율 p과 표준오차 σ_p를 계산한다.
- 중심극한명제에 의지하여 우리는 하나의 표본에서 계산한 p가 (어쨌든) 모비율 π에 대한 최선의 추정량임을 안다.

모비율의 신뢰구간

모비율의 신뢰구간은 어떻게 구할 수 있을까?

- 중심극한명제에 의지하여 우리는 정규분포에서 (가장 확률이 높은 부분만을 골라) 신뢰구간(e.g., 90%, 95%, 99%, 99.9%)의 면적을 계산할 수 있다.
- 면적의 x축에 대응하는 Z 값에 원점수를 표준화하여 정확히 구간이 어디에서 어디까지인지 밝힌다.
- 신뢰구간의 기준은 어디까지나 모비율 π이 되어야 한다. (표본에 따라) 확률적으로 변화할 수 있는 것은 p임에 주의하면서 신뢰구간의 의미를 해석한다.

모비율의 신뢰구간

예제 5. 미국 Opinion Today 2019년 7월 1일자에 따르면 트럼프 대통령의 경제정책을 지지하는 사람의 비율은 47%였다. 이 표본은 1,116명의 성인을 대상으로 설계된 것이다. 트럼프의 경제정책에 지지하는 전체 미국인의 비율에 대한 99% 신뢰구간을 구하시오.

모비율의 신뢰구간

- 가장 먼저 문제가 무엇을 묻고 있는지 파악해야 한다.

$$P(? \leq \pi \leq ?) = .99$$

- 다음으로 이 문제에서 얻은 정보는 아래와 같다.

$$\pi = ?, \qquad p = .47, \quad n = 1116$$

- 중심극한명제에 따라 이 표집분포는 정규분포함을 알 수 있다.
- 이 (표본비율의) 표집분포의 평균 $E(p)$는 무엇인가? 주어진 (하나의) 임의표본의 평균 $p = 0.47$이 현실적으로 모비율에 대한 최선의 추정량이다. 그러므로,

$$E(p) = 0.47$$

- 이 표집분포의 표준오차 σ_p는 무엇인가? 주어진 (한 번의) 표본비율 p는 0.47이고 표본 크기는 $n = 1116$이므로,

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.47 \cdot (1 - 0.47)}{1116}}$$

모비율의 신뢰구간

- 일단 표본비율의 확률분포, 즉 표집분포를 그려본 뒤 구하려는 신뢰구간 99%에 대해 대충 색칠 공부를 해보자.
- 정확한 99% 면적을 결정하는 Z 값은 엑셀의 NORM.INV(0.995, 0, 1)과 NORM.INV(0.005, 0, 1)로 얻는다(Why?).
- 이제 그 값들을 x축 위의 Z 값에 표시하자.
- 아까 NORM.INV$(\cdot)$ 함수를 통해 구한 값들은 Z 값이므로 다시 원점수로 환원해야 한다($p = E(p) + Z \cdot \sigma_p$).
- 환원된 그 원점수들을 다시 그래프 위에 표시하고 신뢰구간을 보고한다.

$$P(p + Z_{0.005} \cdot \sigma_p \leq \pi \leq p + Z_{0.995} \cdot \sigma_p) =$$

$$P\left(0.47 - 2.58 \cdot \sqrt{\frac{0.47 \cdot (1 - 0.47)}{1116}} \leq \pi \leq 0.47 + 2.58 \cdot \sqrt{\frac{0.47 \cdot (1 - 0.47)}{1116}}\right) =$$

$$P(.43 \leq \pi \leq .51) = .99$$

모비율의 신뢰구간

문제를 풀 때는 $E(p)$와 σ_p의 성격을 다소 주의해서 공부해야 한다.

- (표본평균과는 달리) 표본비율에 관한 추정에서는 표본표준편차 σ_p를 별개로 제공받을 필요가 없다(Why?).

$$\sigma_p = Var(p) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- 모평균의 신뢰구간을 계산하는 방식으로 모비율의 신뢰구간도 추정할 수 있다(Why?). 이때 사례 수가 커질수록 두 신뢰구간은 비슷해진다.