사회통계연습
통계적 추정

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

October 8, 2021

진행 순서

신뢰구간과 오차범위

신뢰구간과 오차범위

아까 배운 확률이론과 현실 통계분석의 관심 문제에는 차이가 있다.

- 앞서 배운 확률이론의 기초에서는 일단 모집단의 평균(μ)과 표준편차(σ)를 안다고 전제하고 있었다.
- 예컨대 A학교 전교생의 IQ 점수분포가 $\mu = 105$, $\sigma = 15$인 정규분포를 이룬다면 25명을 랜덤 샘플링하고 그 평균을 구해보니 108 이상으로 나올 확률은 $\mathrm{P}(\bar{\mathrm{X}}) = \mathrm{P}\left(\mathrm{Z} \geq \frac{108-105}{15/\sqrt{25}}\right) = 1 - \mathrm{P}(\mathrm{Z} < 1) = 0.159$, 즉 15.9%다.
- 그런데 가만 생각해보면 이건 완전 웃긴 이야기다!
- "이미 모집단의 평균(μ)과 표준편차(σ) 같은 걸 다 아는데 뭐하러 샘플링 따위를 하나?"
- "아니 게다가 표본을 무한히 뽑는다니 미친거 아닌가? 차라리 모집단을 전수 조사하고 말지..."
- 현실의 통계분석에서는 오히려 "반대로" 표본에 근거해 바로 그 모집단의 μ와 σ를 추리(inference)하고자 한다.

신뢰구간과 오차범위

표본으로부터 모집단의 성격을 추리(inference)하는 방식에는 두 가지가 있다.

- 하나는 오늘 배울 추정(estimation)이고, 다른 하나는 다음 주에 배울 가설검정(hypothesis test)이다.
- 다시 추정(estimation)에는 점추정(point estimation)과 구간추정(interval estimation)으로 나뉜다.

추정값(estimate)이란 추정된 구체적인 값이다.

- 맥주병 회사에서 일하는 문빈은 제조된 병 가운데 30개의 랜덤 샘플을 확인해 보았다. "음, 병 하나당 용량 평균은 330ml로군."
- 이를 들은 동혁이 물었다: "확실해?"
- 쫄린 문빈이 얼버무렸다: "아니, 꼭 그렇진 않고... 내 샘플에서 표준오차가 10ml 이니, 90% 신뢰구간은 313.6ml에서 346.4ml 사이야."

신뢰구간과 오차범위

점추정값과 구간추정값의 차이는 간단하다.

- 표본의 평균인 330ml를 가지고 모집단 평균을 예측했다면, (그래프 상에 점을 찍듯) 하나의 숫자로 구한 값이므로 점추정값(point estimate)이다.
- 나중에 문빈이 “313.6ml에서 346.4ml 사이”라고 말한 것은 구간추정값(interval estimate)이 된다. 문빈이 말한 90% 신뢰구간(confidence interval; CI)은 90% 구간추정값과 같은 것이다.

그런데 추정값(estimate)과 추정량(estimator)이 좀 헷갈린다.

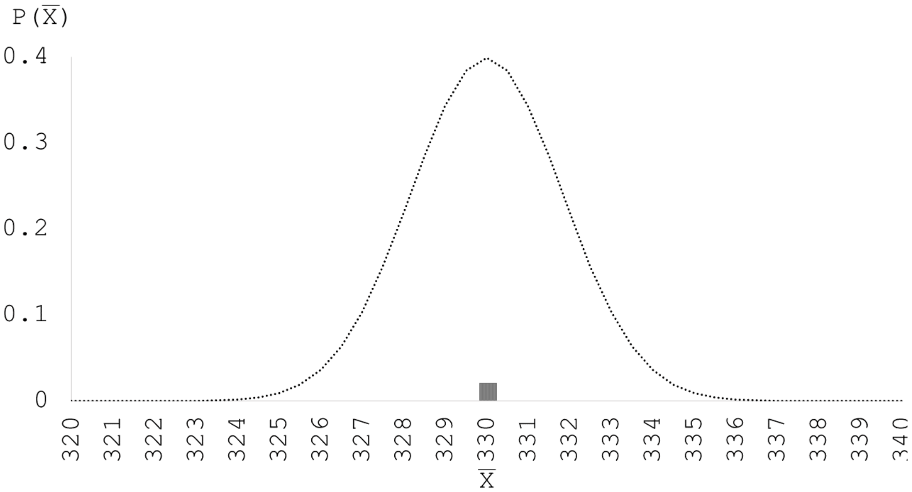
- 추정값(estimate)을 구하는 통계량(statistic)을 추정량(estimator)이라고 부른다.
- 앞서 문빈과 동혁의 예제에서 구체적인 수치인 330ml는 추정값(estimate)였고, 병에 담긴 용량의 표본평균(sample mean)이 추정량(estimator)였다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

어떻게 문빈은 점추정값(point estimate)을 제시할 수 있나?

- 아까 중심극한정리를 통해 분포가 어떻게 생긴 모집단이건 표본 크기만 충분히 크면, $\mathrm{E}(\bar{\mathrm{X}}) = \mu$ 이고 $\mathrm{se}(\bar{\mathrm{X}}) = \sigma/\sqrt{\mathrm{n}}$ 이다는 점을 배웠다.
- 그런데 (1) "모집단의 평균(μ)에 대한 정보가 사전에 주어지지 않았더라도" 그리고 (2) "설령 샘플을 무한히 뽑지 않아도", 가만히 생각해보면 내가 지금 하나의 샘플로부터 얻은 평균값은 (어느 쪽이냐하면) μ를 반영할 확률이 제일 높다!
- 다시 말해, 랜덤 샘플을 한 번만 뽑고 거기서 평균을 구해 $\bar{\mathrm{X}}$를 얻었다면 일단 이것이 모집단 평균(population mean)에 대한 최선의 추정량(best estimator)인 것이다.

신뢰구간과 오차범위


P($\bar{X}$)
0.4
0.3
0.2
0.1
0
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
$\bar{X}$

신뢰구간과 오차범위

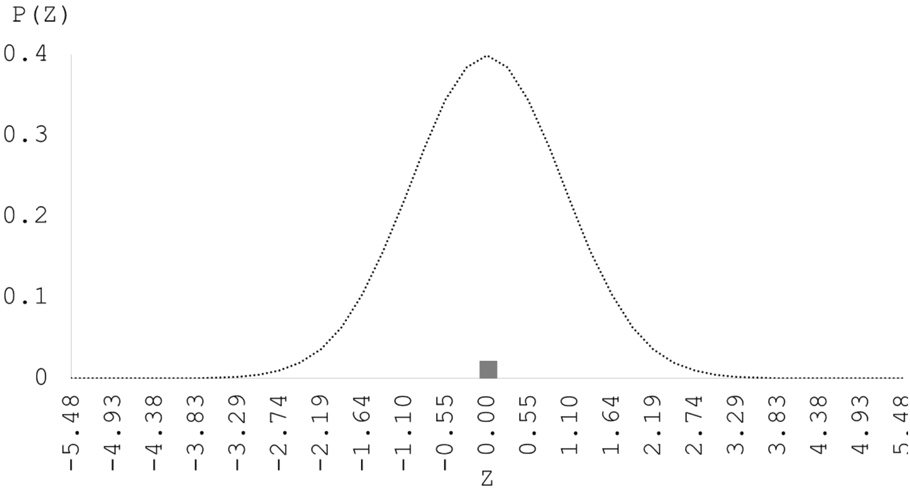
그런데 90% 신뢰구간 같은 건 어떻게 계산할 수 있나?

- 근데 그림을 다시 잘보면, 표본평균(sample mean)에서 아주 조금만 옆으로 이동해도 그게 모평균(population mean)일 확률이 제법 높긴 하다(쫄~).
- 그러면 확률이 높은 순서대로 90%의 면적을 채워나가보자. 거기에 대응하는 X값이 곧 90% 신뢰구간이다!
- 확률이 높은 쪽은 가운데에 몰려있으니 양 끄트머리를 빼고 90%를 채우는 것이 상식적이다.
- 이렇게 양 끄트머리 5%씩을 빼고 가운데 90%의 면적을 계산하는 것이라면 이미 지난 주에 엑셀을 통해 배워 할 수 있다.
- 이제 모집단에서 계산된 평균(μ)과 표준편차(σ)가 아니라, 표본에서 계산된 평균($\mu_{\bar{X}}$)과 표준오차($SE_{\bar{X}}$)를 가지고 Z-점수 표준화를 먼저 하자.

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{SE_{\bar{X}}}$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

신뢰구간과 오차범위



충북대학교 CHUNGBUK NATIONAL UNIVERSITY

신뢰구간과 오차범위

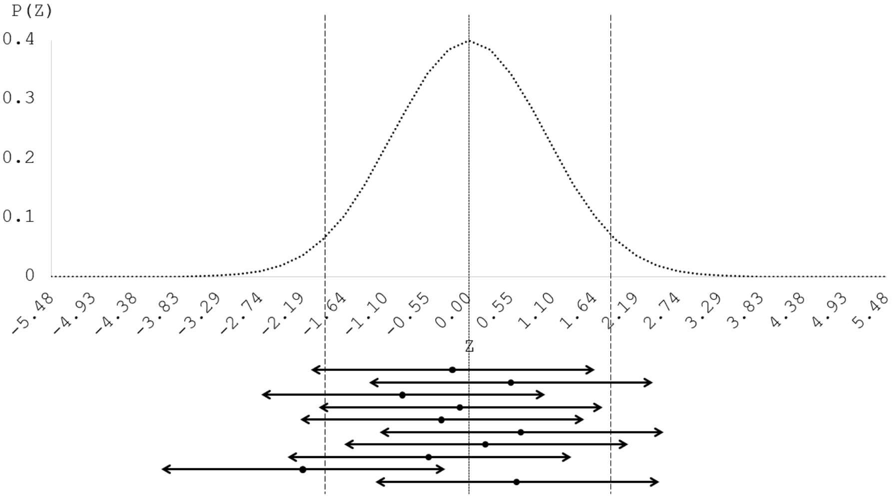
- 엑셀에서 "확률을 주었을때 Z-점수를 토해내는 함수"는 NORM.INV() 함수임을 배웠다.
- NORM.INV(0.95, 0, 1)에서 오른쪽 끄트머리 경계 Z값을, 다른 한편으로 NORM.INV(0.05, 0, 1)에서 왼쪽 끄트머리 경계 Z값을 구할 수 있다(Why?)
- 엑셀에서 실제로 NORM.INV(0.95, 0, 1)와 NORM.INV(0.05, 0, 1)를 구해보면 각각 1.64와 -1.64가 나온다.
- 물론 이 값들은 표준화된 Z-점수이므로 직관적으로 해석이 안된다. 다시 원점수로 돌리는 방법도 지난 주에 배웠다($\mathrm{x_i} = \mathrm{Z_i} \cdot \mathrm{SE}_{\bar{\mathrm{X}}} + \mu_{\bar{\mathrm{X}}}$)
- 즉, 문빈이 말한 90% 신뢰구간은 $[-1.64 \cdot 10 + 330, 1.64 \cdot 10 + 330] = [313.6, 346.4]$였던 것이다.

신뢰구간과 오차범위

90% 신뢰구간(confidence interval)은 어떻게 해석될까?

- "표본을 무한히 많이 추출하여 표본평균들(sample means)을 무한히 계산하였다면, 전체 표본평균들(sample means)의 90%가 [313.6, 346.4] 신뢰구간 사이에 놓인다."
  → 이것이 신뢰구간의 바른 해석이다. 색칠한 내용과도 잘 어울린다!
- "표본을 무한히 많이 뽑았다면 각각의 90% 신뢰구간들(confidence intervals)도 무한히 많이 구할 수도 있다. 이 모든 신뢰구간들의 90%는 모집단의 평균(population mean)인 330을 포함하고 있다."
  → 이것도 신뢰구간의 바른 해석이다.
- "[313.6, 346.4] 사이에 모집단의 평균이 놓일 확률이 90%이다."
  → 이것은 신뢰구간의 **잘못된** 해석이다.

신뢰구간과 오차범위



P(Z)
0.4
0.3
0.2
0.1
0
-5.48
-4.93
-4.38
-3.83
-3.29
-2.74
-2.19
-1.64
-1.10
-0.55
0.00
0.55
1.10
1.64
2.19
2.74
3.29
3.83
4.38
4.93
5.48
Z

신뢰구간과 오차범위

왜 이렇게 신뢰구간(confidence interval) 해석이 헷갈릴까?

- 모평균(μ)는 알려지지 않았을 뿐, 상수(constant)임을 기억할 것! 확률변수가 아니다. 오히려 확률변수 쪽은 표본평균(sample means)과 그에 따른 신뢰구간(confidence interval)이다.
- 따라서 해석할 때 "기준이 되는 쪽"은 반드시 모평균(the population mean; μ)이고, 90%로 이를 맞출 확률을 갖는 것은 무한히 많은 표본평균들(sample means) 내지 신뢰구간들(confidence intervals)이다.

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

신뢰구간과 오차범위

표본평균과 신뢰구간을 이해했다면 이제 수식으로도 나타낼 수 있다.

- 여기서 유심히 봐야하는 부분은 가운데에 $\bar{X}$가 아닌 μ 가 기준으로 자리잡고 있다는 점이다.
- 엑셀에서 **NORM.INV(0.05, 0, 1)**을 통해 -1.64라는 값을 찾았고, 엑셀에서 **NORM.INV(0.95, 0, 1)**를 통해 1.64라는 값을 찾았다(숫자 주의!)
- 그러므로 90% 신뢰구간은 다음과 같다:

$$P(\bar{X} - 1.64 \cdot SE_{\bar{X}} \leq \mu \leq \bar{X} + 1.64 \cdot SE_{\bar{X}}) = .90$$

- 이때, 90% 를 신뢰수준(confidence level)이라고 표현하며, 90% 말고 95%나 99% 등 신뢰수준도 결국 같은 원리로 계산할 수 있다.
- $\pm 1.64 \cdot SE_{\bar{X}}$ 부분을 특별히 오차범위(margin of error)라고 부른다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모평균의 신뢰구간

근본적으로 굉장히 중요한 부분은 지금 우리가 의사결정의 오류를 다루고 있다는 사실이다.

- 제일 처음에 우리는 “표준오차(standard error)는 표본평균(sample means)을 이용해서 의사결정을 할 때 예상되는 오류의 크기를 나타낸다”고 말했다.
- 애초에 신뢰구간(confidence interval) 추정의 목적은 “표본에서의 계산된 우리의 통계량(statistic)이 얼마나 믿을만한가(reliable)?”를 파악하는데 있다.
- 아까 우리는 현실의 통계분석상 모집단의 μ를 모른다고 했다. 표본을 무한히 뽑을 수도 없다고 했다. 그래서 하는 수 없이 표본을 한 번 추출해서 $\mu_{\bar{X}}$를 계산했다. 과연 그 $\mu_{\bar{X}}$는 얼마나 믿을만 한가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모평균의 신뢰구간

- 우리의 표본이 $\mu_{\bar{X}}$ 을 중심으로 아주 밀집되어 있다면, $SE_{\bar{X}}$ 는 작고, $\mu_{\bar{X}}$ 는 꽤 믿을만 할 것이며, 90% 신뢰구간 $P(\bar{X} - 1.64 \cdot SE_{\bar{X}} \leq \mu \leq \bar{X} + 1.64 \cdot SE_{\bar{X}})$ 도 좁을 것이다.
- 우리의 표본이 $\mu_{\bar{X}}$ 가 아닌 여러 값들로 퍼져있다면, $SE_{\bar{X}}$ 는 크고, $\mu_{\bar{X}}$ 는 믿기 어려울 것이며, 90% 신뢰구간 $P(\bar{X} - 1.64 \cdot SE_{\bar{X}} \leq \mu \leq \bar{X} + 1.64 \cdot SE_{\bar{X}})$ 도 넓을 것이다.
- 각각의 상황에 따른 가상의 표집분포(sampling distribution)를 상상해보면 큰 도움이 된다!

모평균의 신뢰구간

모평균의 신뢰구간

사실 지금까지 공부한 내용이 모평균의 신뢰구간(confidence intervals for the population mean)이다.

- 계산에 매몰되면 큰 그림을 놓치기 쉬우니 다시 돌이켜보자.
- 현실의 통계분석에서 우리는 모집단의 평균(μ)과 표준편차(σ)를 알지 못한다. 표본을 무한히 뽑지도 않는다.
- 대신 크기가 n인 표본을 딱 한 번만 추출하고 거기에서 (가상적인) 표집분포의 평균($\mu_{\bar{X}}$)과 표준오차($SE_{\bar{X}}$)를 추리한다.
- 중심극한명제에 의지하여 우리는 하나의 표본으로 구한 $\mu_{\bar{X}}$가 모평균(population mean)의 best estimator임을 안다(딱 하나의 숫자만 고른다면 어찌되었든 이것이 최선이기 때문이다!).

충북대학교 CHUNGBUK NATIONAL UNIVERSITY

모평균의 신뢰구간

- 그런데 만일 모평균에 대한 점추정량(point estimator)을 넘어 구간추정량(interval estimator)을 생각한다면 어떨까?
- 또다시 중심극한명제에 의지하여 우리는 정규분포에서 (확률이 높은 부분만을 최대한 골라) 90%의 면적을 계산할 수 있다.
- 면적의 x축에 대응하는 Z-점수에 원점수를 표준화(또는 역표준화)하여 정확히 구간이 어디에서 어디까지인지 밝힌다.
- (1) μ가 기준이 되도록 주의하면서, 또는 (2) $\mu_{\bar{X}}$와 $S_{\bar{X}}$가 확률적으로 변화한다는 사실에 유념하면서 신뢰구간의 의미를 해석한다.

예제 3. 네오청주 장학사 재범은 모교인 C대학에서 K교수를 질타하기 위해 〈사회통계연습〉 교육 효과가 영 부진하다고 의심한다. 그는 이 과목 이수자로부터 30명의 표본을 랜덤하게 뽑아, 평균이 75점이고 표준편차가 10점임을 확인한 뒤 두 눈을 의심하였다. 그의 표본에 따르면 모평균의 90% 신뢰구간(confidence interval)은 어떻게 되는가?

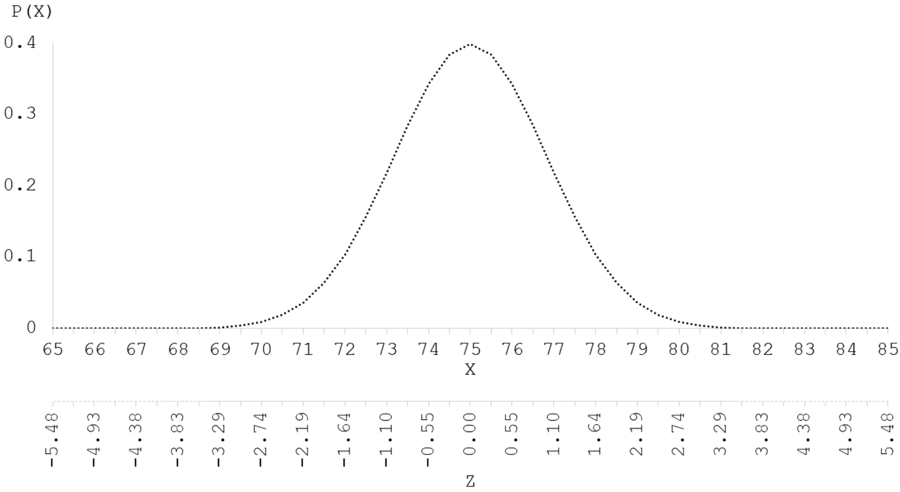
- 이 문제에서 주어진 정보는 다음과 같다:

$$\mu =?, \sigma =?, \bar{\mathrm{X}} = 75, \mathrm{SD} = 10, \mathrm{n} = 30$$

- 중심극한명제에 따라 재범의 (가상적인) 표집분포(sampling distribution)는 정규분포함을 알 수 있다.
- 이 표집분포의 평균은 어떤가? 주어진 (한 번의) 표본의 평균 $\bar{\mathrm{X}} = 75$가 현실적으로 모평균(population mean)의 best estimator임을 안다. 따라서 재범의 (가상적인) 표집분포에서 평균, 즉 $\mathrm{E}(\bar{\mathrm{X}})$은 75일 것이다.
- 이 표집분포의 표준편차는 어떤가? 주어진 (한 번의) 표본의 표준편차(SD)는 10점이고 $\mathrm{n} = 30$이므로, 재범의 (가상적인) 표집분포에서 표준오차($\mathrm{SE}_{\bar{\mathrm{X}}}$)는 $75/\sqrt{30}$일 것이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모평균의 신뢰구간


P(X)
0.4
0.3
0.2
0.1
0
65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85
X
-5.48 -4.93 -4.38 -3.83 -3.29 -2.74 -2.19 -1.64 -1.10 -0.55 0.00 0.55 1.10 1.64 2.19 2.74 3.29 3.83 4.38 4.93 5.48
Z

모평균의 신뢰구간

- 일단 (가상적인) 표본평균의 확률분포, 즉 표집분포를 그려본 뒤 거기에서 문제되는 신뢰구간 90%를 대충 한 번 가늠해본다.
- 정확한 90% 면적을 결정하는 Z-점수는 엑셀의 **NORM.INV(0.95, 0, 1)**과 **NORM.INV(0.05, 0, 1)**로 특정한다.
- 이제 그 값들을 Z-점수 축에 표시하고 90% 면적을 실제로 색칠한다.
- 아까 **NORM.INV()** 함수를 통해 구한 값들은 Z-점수이므로 직관적으로 알 수 없다. 다시 원점수로 환원한다($x_i = Z_i \cdot SE_{\bar{X}} + \mu_{\bar{X}}$).
- 환원된 그 값들을 X-점수 축에 표시하고 신뢰구간을 보고한다.

$$
\begin{aligned}
P(\bar{X} - 1.64 \cdot SE_{\bar{X}} \le \mu \le \bar{X} + 1.64 \cdot SE_{\bar{X}}) &= \\
P(75 - 1.64 \cdot \frac{10}{\sqrt{30}} \le \mu \le 75 + 1.64 \cdot \frac{10}{\sqrt{30}}) &= \\
P(72 \le \mu \le 78) &= .90
\end{aligned}
$$

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모비율의 신뢰구간

모비율의 신뢰구간

이제 모비율의 신뢰구간(confidence intervals for the population proportion)에 대해서 이야기할 차례이다.

- 오늘 앞선 시간에 표본평균과 표본비율의 표집분포(sampling distribution)를 대조하여 이야기한 것과 마찬가지 맥락이다.
- 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 살펴본 무단횡단의 비율 등.

모비율의 신뢰구간

- 다행히 아까와 마찬가지로 논리 구조는 완전히 똑같다. 심지어 중심극한정리가 성립하는 것까지도!
- (1) 현실의 통계분석에서 우리는 모집단의 비율(π)과 표준편차(σ_π)를 알지 못한다. (2) 표본도 무한히 뽑지 않는다. 대신 크기가 n인 표본을 하나만 골라 거기에서 (가상적인) 표집분포의 비율(p)과 표준오차(SE_p)를 추리한다.
- 중심극한명제에 의지하여 우리는 하나의 표본으로 구한 p가 모비율(population proportion)의 best estimator임을 안다(딱 하나만 고른다면 어찌되었든 이것이 최선이기 때문이다!).

모비율의 신뢰구간

- 그런데 만일 모비율에 대한 점추정량(point estimator)을 넘어 구간추정량(interval estimator)을 생각한다면 어떨까?
- 또다시 중심극한명제에 의지하여 우리는 정규분포에서 (가장 확률이 높은 부분만을 골라) 90%의 면적을 계산할 수 있다.
- 면적의 x축에 대응하는 Z-점수에 원점수를 표준화(또는 역표준화)하여 정확히 구간이 어디에서 어디까지인지 밝힌다.
- (1) p가 기준이 되도록 주의하면서, 또는 (2) p와 SE_p가 확률적으로 변화한다는 점에 유념하면서, 신뢰구간의 의미를 해석한다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

예제 4. 미국 Opinion Today 2019년 7월 1일자에 따르면 트럼프 대통령의 경제정책을 지지하는 사람의 비율은 47%였다. 이 표본은 1,116 명의 성인을 대상으로 설계된 것이다. 트럼프의 경제정책에 지지하는 전체 미국인의 비율에 대한 99% 신뢰구간을 구하여라.

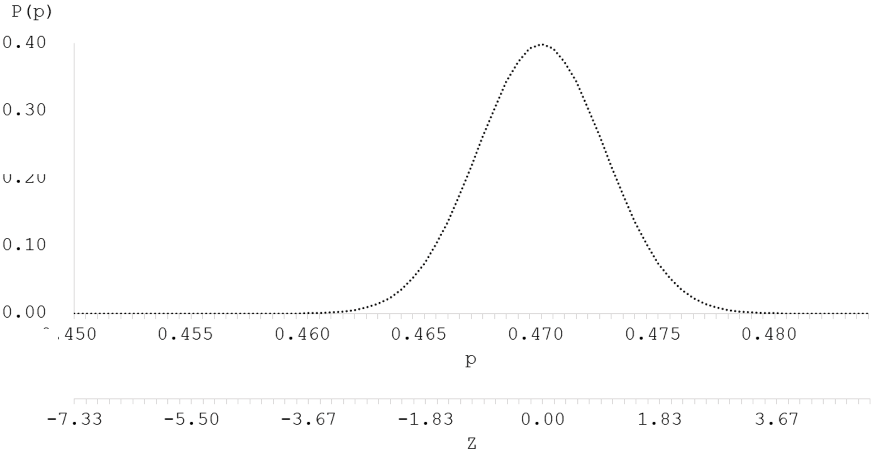
- 이 문제에서 얻은 정보는 다음과 같다:

$$\pi = ?, \sigma_\pi = ?, \mathrm{p} = .47, \mathrm{n} = 1116$$

- 중심극한명제에 따라 이 (가상적인) 표집분포(sampling distribution)는 정규분포함을 알 수 있다.
- 이 표집분포의 평균은 어떤가? 주어진 (한 번의) 표본의 평균 p=.47가 현실적으로 모비율(population proportion)의 best estimator임을 안다. 따라서 이 서베이의 (가상적인) 표집분포에서 비율, 즉 E(p)은 .47일 것이다.
- 이 표집분포의 표준편차는 어떤가? 이 서베이의 (가상적인) 표집분포에서 표준편차는 $\mathrm{se(p)} = \sqrt{\frac{\mathrm{p(1-p)}}{\mathrm{n}}} = \sqrt{\frac{0.47\cdot(1-0.47)}{1116}}$ 일 것이다.

충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모평균의 신뢰구간



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

모평균의 신뢰구간

- 일단 (가상적인) 표본비율의 확률분포, 즉 표집분포를 그려본 뒤 거기에서 문제되는 신뢰구간 99%를 대충 한 번 가늠해본다.
- 정확한 99% 면적을 결정하는 Z-점수는 엑셀의 **NORM.INV(0.995, 0, 1)**과 **NORM.INV(0.005, 0, 1)**로 특정한다.
- 이제 그 값들을 Z-점수 축에 표시하고 99% 면적을 실제로 색칠한다.
- 아까 **NORM.INV()** 함수를 통해 구한 값들은 Z-점수이므로 직관적으로 알 수 없다. 다시 원점수로 환원한다($p_i = Z_i \cdot SE_p + p$).
- 환원된 그 값들을 X-점수 축에 표시하고 신뢰구간을 보고한다.

$$P(p - 2.58 \cdot SE_p \leq \pi \leq p + 2.58 \cdot SE_p) =$$

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

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

끝!

와~ 이번 숙제도 엑셀 문제네~

충북대학교
CHUNGBUK NATIONAL UNIVERSITY