KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

5. 확률과 통계적 추론 II: 정규분포와 신뢰구간

정치와 데이터분석

박상훈 (sh.park.poli@gmail.com)
강원대학교

오늘의 목표

10:10-11:00

지난 주차의 분포, 확률에 대한 개념 복습하기

11:10-12:05

모집단, 표본, 추론, 그리고 신뢰구간에 대해 이해하기 I.

12:15-12:45

모집단, 표본, 추론, 그리고 신뢰구간에 대해 이해하기 II.

실습과제 해설 및 질의응답

Part I. 잃어버린 확률과 분포를 찾아서

Recap the Last Class!

확률분포의 유형

확률분포: 확률변수의 모든 가능한 값과 그 확률들의 분포

- 주사위 눈금: 1~6 각 1/6 확률 → 균등한 분포 (이산형)

이산형 분포: 특정 값들이 떨어져 있음. 확률질량함수 (PMF)로 정의

- 동전 앞면 개수 $(0, 1, 2, \ldots, n)$, 사건 발생 횟수 $(0, 1, 2, \ldots)$ 등

연속형 분포: 값이 연속적인 범위. 확률밀도함수 (PDF)로 정의

- 키, 몸무게, 시험 점수 등의 분포 (연속값)

Recap the Last Class!

기대값과 분산

기대값 $\mathbb{E}[X]$: 확률변수 X의 평균적인 값 (확률 가중 평균)

- 이산형: $\mathbb{E}[X] = \sum x \cdot \Pr(X = x)$
- 연속형: $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$
- 분포의 중심 경향을 나타냄(장기적 평균).

분산 $\mathrm{Var}(X)$: 분포의 산포(변동성) 측도 = $\mathbb{E}[(X - \mathbb{E}[X])^2]$

- 표준편차 $\sigma = \sqrt{\mathrm{Var}(X)}$는 분산의 양의 제곱근 (단위 일치)
- 분산/표준편차가 클수록 분포가 퍼져 있음(데이터 변동성 높음).

Recap the Last Class!

이항(Binomial) 분포

이항분포: n번의 독립 시도에서 특정 사건의 성공 횟수 분포

- 각 시도 성공확률 p (변하지 않음), 실패확률 $1-p$
- 확률질량함수: $\Pr(X=k) = \binom{n}{k} p^k (1-p)^{,n-k}, \quad k = 0, 1, \ldots, n$
- 기댓값 $\mathbb{E}[X] = np$, 분산 $\mathrm{Var}(X) = np(1-p)$
  - 10개 선거구 중 후보 A가 선거구당 당선 확률 0.6일 때, 당선된 선거구 수 $X \sim \mathrm{Binomial}(n=10, p=0.6)$
  - $\Pr(X=6) = \binom{10}{6}(0.6)^6(0.4)^4 \approx 0.2508$ (가장 확률 높은 경우는 $k=6$)

Recap the Last Class!

포아송(Poisson) 분포

포아송분포: 일정 시간/공간 내에 발생하는 드문 사건의 횟수 모델

- 모수 λ = 단위 시간당 평균 발생횟수 (기댓값)
- 확률질량함수: $\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$
- 기댓값 $\mathbb{E}[X] = \lambda$, 분산 $\mathrm{Var}(X) = \lambda$ (평균=분산)
  - 한 시간에 콜센터로 걸려오는 전화 건수 $X \sim \mathrm{Poisson}(\lambda = 20)$ (평균 20건)
  - $\Pr(X = 25) = \frac{20^{25} e^{-20}}{25!} \approx 0.044$ (25건 발생할 확률)
  - 포아송분포는 n 크고 p 매우 작을 때 $\mathrm{Binomial}(n, p)$의 극한으로 수렴(평균 $\lambda = np$)

Recap the Last Class!

정규(Normal) 분포

정규분포(Gaussian 분포): 연속분포의 한 종류, 종(bell) 모양 대칭 곡선

- 모수: 평균 μ (중심)와 분산 σ^2 (퍼짐 정도)
- 밀도함수: $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp!\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- 평균을 중심으로 좌우 대칭, σ가 클수록 곡선이 낮고 넓게 퍼짐.
- **중심극한정리: 표본크기 n이 충분히 크면, 표본평균 분포가 정규분포에 근사**
- 여러 확률모형에서 정규분포는 실질적인 한계분포로 중요
    - 표본평균, 오차항 분포 등 다양한 통계량이 근사적으로 정규분포 따름.

Recap the Last Class!

정규분포의 특징

표준정규화: 정규분포 $X \sim N(\mu, \sigma^2)$에서 $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ (평균0, 분산1)

- 정규분포 확률 계산은 표준정규분포 $N(0,1)$ 표로부터 얻음
- 분포의 68-95-99% 법칙: 정규분포에서 평균 기준
  - 약 **68%**의 값이 $\mu \pm 1\sigma$ 범위 내
  - 약 **95%**의 값이 $\mu \pm 2\sigma$ (정확히 $\mu \pm 1.96\sigma$) 범위 내
  - 약 **99.7%**의 값이 $\mu \pm 3\sigma$ 범위 내

정규분포는 극단치(outlier) 확률이 낮음(꼬리 쪽 급격히 감소).

Recap the Last Class!

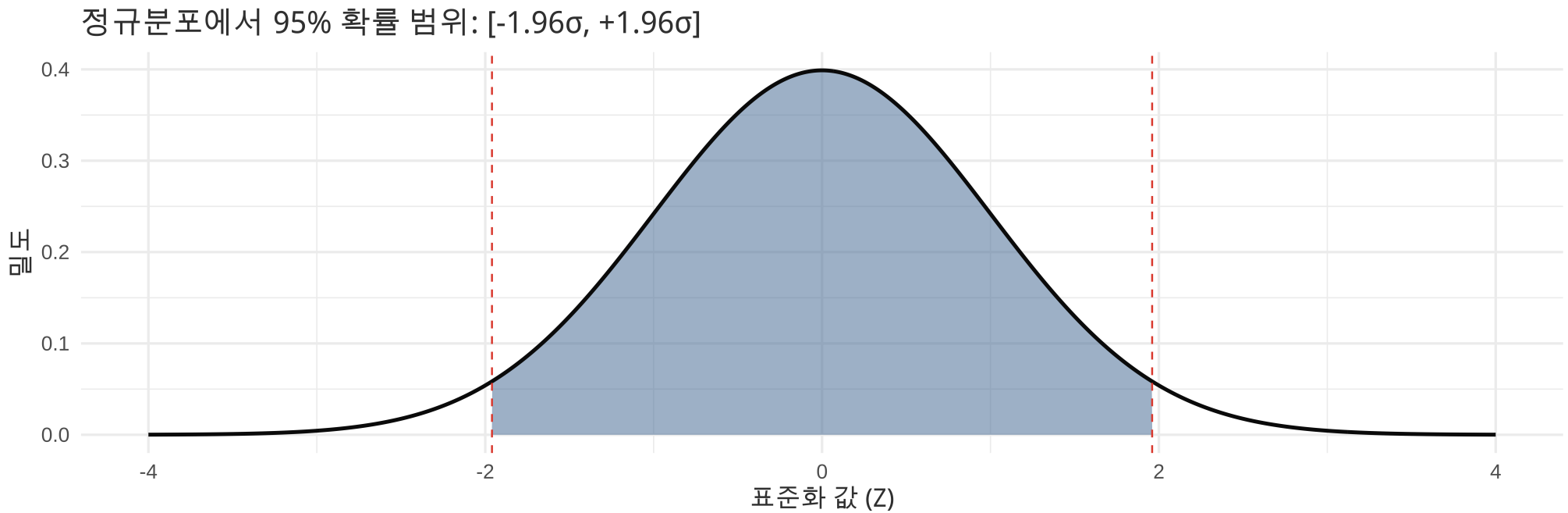
정규분포의 특징

95% 범위와 $\pm 1.96\sigma$

- 정규분포에서 평균 주변 약 95% 확률질량이 $\mu \pm 1.96\sigma$ 구간에 포함됨.
  - 즉, $P(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) \approx 0.95$
- 1.96은 표준정규분포에서 누적확률 0.975에 해당하는 값 (상위 2.5% 지점)
- 이를 활용해 표본평균의 신뢰구간을 구축할 수 있음.

Recap the Last Class!

정규분포의 특징



표준정규분포 $N(0,1)$에서 파란 부분은 전체 확률의 약 95%를 차지

빨간 점선은 $Z = \pm 1.96$ 위치 (경계), 양쪽 꼬리 각 2.5% 확률 영역

Recap the Last Class!

모집단과 표본

과학적 연구의 목표는 **모집단(population)** 전체에 대한 결론을 내리는 것

하지만 현실적으로 **모집단 전체를 조사하는 것은 불가능**

따라서 우리는 **표본(sample)** 을 사용

- 표본은 모집단을 축소한, '관측한 모집단의 하위집단'이지만, 모집단의 특성을 제대로 반영해야 함.
- 이를 **대표성 있는** 표본(representative sample)이라고 함.

Recap the Last Class!

모집단과 표본

💡 **대표성(representativeness)** 이란?

> 표본의 특성이 모집단의 특성과 유사한 정도

- 대표성이 떨어지면 아무리 복잡한 통계기법을 써도 **잘못된 결론**을 얻을 수 있음.

대표성을 담보하는 핵심: 무작위화(Randomization)

대표성을 확보하는 가장 기본적인 방법: **무작위추출(random sampling)**

- 모집단의 모든 구성원이 **동등한 확률**로 표본에 포함될 기회를 갖도록 하는 것
- 특정 집단이 과대 혹은 과소 대표되지 않도록 함.
- 체계적 편향(systematic bias)을 방지하고, 표본 오차(sampling error)를 계산 가능하게 만들어 줌.

Recap the Last Class!

모집단과 표본

💡 대표성(representativeness) 이란?

> 표본의 특성이 모집단의 특성과 유사한 정도

- 대표성이 떨어지면 아무리 복잡한 통계기법을 써도 **잘못된 결론**을 얻을 수 있음.

대표성을 담보하는 핵심: 무작위화(Randomization)

대표성을 확보하는 가장 기본적인 방법: **무작위추출(random sampling)**

🔍 예시

- "대학생 여론조사"를 할 때, 오직 오전 수업에만 참여한 학생들을 조사한다면 편향된 표본
- 반면, 학과·학년·시간대와 무관하게 무작위로 선정하면 **대표성 확보**

Recap the Last Class!

모집단과 표본

💡 **대표성(representativeness)** 이란?

> 표본의 특성이 모집단의 특성과 유사한 정도

- 대표성이 떨어지면 아무리 복잡한 통계기법을 써도 **잘못된 결론**을 얻을 수 있음.

대표성을 담보하는 핵심: 무작위화(Randomization)

대표성을 확보하는 가장 기본적인 방법: **무작위추출(random sampling)**

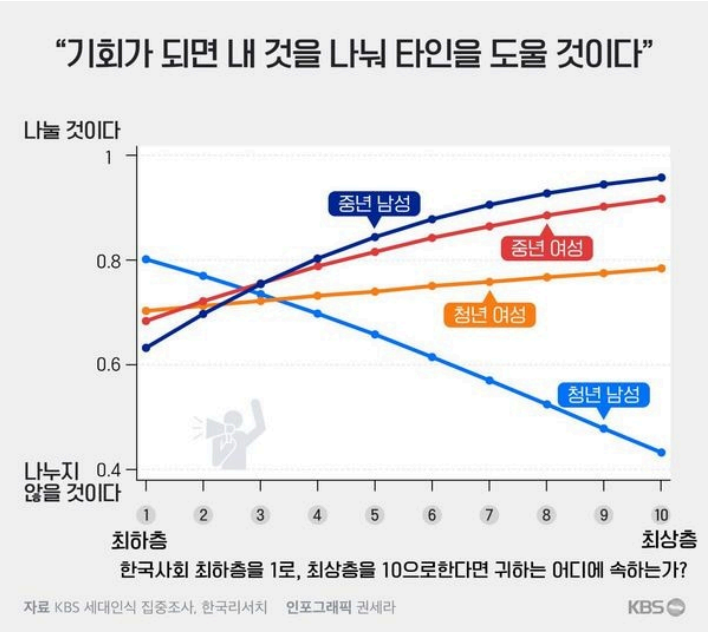
🎯 무작위화는 '운에 맡긴다'가 아니라, **편향을 줄이고 과학적 추론의 기반을 만드는 것**

Part II. 정규분포와 신뢰구간

확률과 통계적 추론 II: 정규분포와 신뢰구간

KBS 세대 인식조사

KBS 세대 인식조사 자료를 통해 분석한 "주관적 계층 의식과 세대 및 성별 간 관계"



KBS-한국리서치 온라인 설문자료

청년세대(20-34세)와 586세대 간 상호인식이 어떠한가?

- 종속변수: 도움을 줄 의사가 있다(1) or 없다(0)

연구진의 주장:

"50대 남녀 및 20-34세 여성과 달리, 20-34세 남성은 자신이 소속한 계층이 높다고 생각할수록 우리 사회의 어려운 사람들을 위해 내가 가진 것을 나누어주고 싶다는 생각을 덜 한다."

확률과 통계적 추론 II: 정규분포와 신뢰구간

무엇이 문제였을까?

확률과 통계적 추론 II: 정규분포와 신뢰구간

무엇이 문제였을까?

2021년 6월 29일 조선일보 기사에 정리된 다섯 가지 질문들

- Q1. X 축의 10개로 나뉜 소득 수준마다 충분한 응답 수가 모였는지?
- Q2. 2030 남성 응답자 300명이 충분한 표본크기인지?
- Q3. 설문 결과가 이렇게 예쁜 선으로 표현되는 게 가능한지?
- Q4. 회귀분석 대신, 각 구간별 '네'라고 답한 비율을 표시해주는 간단한 방식을 사용하는 것은 어떨지?
- Q5. 이 그래프가 제대로 된 결과물일 가능성?

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란?

모집단의 정확한 값을 아는 것은 거의 불가능

- 우리는 표본을 통해 **'그럴 듯한 범위'**를 추정(estimate)
- 이때 사용하는 것이 **신뢰구간(confidence interval)**

일반적으로 사용하는 95% 신뢰구간이란, 같은 방식으로 100번 표본을 뽑으면 약 **95번은 실제 모집단의 값을 포함하게 되는 구간**을 의미. 💬 쉽게 말하면,

- 정확한 한 점(점추정)을 맞추는 대신, 그 근처의 **믿을 만한 구간**을 제시하는 방법

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란? 점추정과 구간추정

점추정(point estimate)은 **창으로 물고기 한 마리를 찌르려는 시도**

신뢰구간(interval estimate)은 **그물을 던져서 그 안에 물고기가 들어오기를 기대하는 방법**

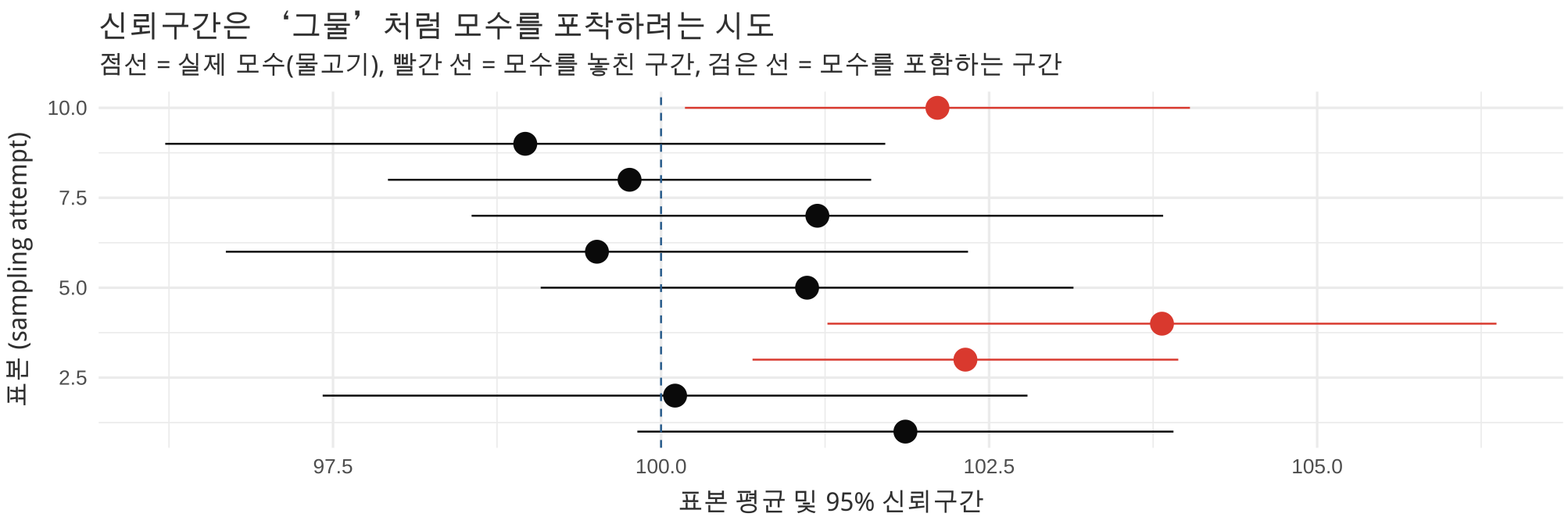
- 그물은 넓지만 실제 물고기를 잡을 가능성이 높음.
- 표본이 많을수록 그물은 조밀해지고, 신뢰구간은 좁아짐.
- 신뢰수준이 높을수록(예: 99%) 그물은 더 넓어짐.

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란? 점추정과 구간추정

점추정(point estimate)은 **창으로 물고기 한 마리를 찌르려는 시도**

신뢰구간(interval estimate)은 **그물을 던져서 그 안에 물고기가 들어오기를 기대하는 방법**



확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란?

예시로 이해하기: 대학생의 하루 평균 수면시간

대학생 100명을 조사했더니 하루 평균 수면시간은 **6.8시간**, 표준편차는 **0.9시간**이었다고 하자.

이때, 95% 신뢰구간 계산은 다음과 같음.

$$6.8 \pm 1.96 \times \frac{0.9}{\sqrt{100}} = [6.62, 6.98]$$

- 이 구간은 진짜 평균이 이 안에 있을 확률이 95%라는 뜻이 **아님**.
- **이 과정을 100번 반복하면 95개의 구간이 실제 평균을 포함한다**는 뜻
  - 즉, 신뢰구간은 하나의 구간이 아니라, **추정 절차의 신뢰도**

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란?

예시로 이해하기: 공장 생산제품의 내구도

어떤 공장에서 생산되는 물건의 평균 수명을 알고 싶어, 25개를 랜덤 추출해 **95% 신뢰도로 구간추정**을 했더니 결과는 **[98.3, 102.2]일**이었다고 하자. 💡 이 신뢰구간에 대해 올바른 해석은 무엇일까?

A. 해당 물건의 평균 수명은 95% 확률로 98.3~102.2일 사이에 있다.

B. [98.3, 102.2]일이라는 신뢰구간은 95% 확률로 정확하다.

C. 이 신뢰구간 하나에 대한 확률적 진술은 불가능하다.

D. [98.3, 102.2]일이라는 신뢰구간은 95% 믿을 수 있다.

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란?

예시로 이해하기: 공장 생산제품의 내구도

어떤 공장에서 생산되는 물건의 평균 수명을 알고 싶어, 25개를 랜덤 추출해 **95% 신뢰도로 구간추정**을 했더니 결과는 **[98.3, 102.2]일**이었다고 하자. 💡 이 신뢰구간에 대해 올바른 해석은 무엇일까?

A. 해당 물건의 평균 수명은 95% 확률로 98.3~102.2일 사이에 있다.

B. [98.3, 102.2]일이라는 신뢰구간은 95% 확률로 정확하다.

C. 이 신뢰구간 하나에 대한 확률적 진술은 불가능하다.

D. [98.3, 102.2]일이라는 신뢰구간은 95% 믿을 수 있다.

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란?

예시로 이해하기: 공장 생산제품의 내구도

95% 신뢰수준/신뢰구간이란 [98.3, 102.2]이라는 **구간이 참값을 포함할 확률이 95%라는 뜻이 아님.**

- 신뢰수준은 **신뢰구간을 만드는 절차 자체의 정확성**을 의미
    - 즉, 같은 방식으로 표본을 반복 추출해 신뢰구간을 여러 번 만들면, 그 중 **약 95%의 구간이 모평균의 참값을 포함**하게 됨.
- [98.3, 102.2]라는 구간은 **이미 고정된 값**
    - 참 모평균은 그 안에 **들어 있거나(1)**, **안 들어 있거나(0)** 둘 중 하나일 뿐

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간, 모집단, 표본

모집단(population): 우리가 알고 싶은 전체 집단의 특성으로, 예를 들면 "이 공장에서 생산된 모든 제품의 평균 수명"

표본(sample): 모집단의 일부를 추출한 것으로, 예를 들면 "그중 25개의 제품을 무작위로 뽑아 관찰한 것"

구분	모집단	표본
대상	전체 (모든 단위)	일부 (추출된 단위)
값	모수(parameter)	통계치(statistic)
특성	고정(fixed)	확률(random)

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간, 모집단, 표본: 고정된 모수, 확률적 통계치

모수(parameter): 모집단의 **진짜** 평균, 비율, 회귀계수 등

- 실제로 존재하지만 우리가 직접 알 수 없음. 따라서 **고정값(fixed value)**

통계치(statistic): 표본으로부터 계산한 추정치. 표본의 평균, 비율, $\hat{\beta}$ 등

- 표본을 다르게 뽑으면 값이 달라질 수 있으므로 **확률변수(random variable)**

> 📊 **즉,** 모수는 움직이지 않지만, 통계치는 표본이 바뀔 때마다 '흔들리는' 값

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간, 모집단, 표본: 고정된 모수, 확률적 통계치

확률적이다?

확률적이라는 말은 **우리가 뽑은 표본이 달라질 수 있다**는 뜻

- 표본평균 $\bar{X}$는 매번 달라질 수 있지만 그 분포는 예측 가능: **표본평균의 분포(표집분포)**
- 결국, 우리는 모집단의 모수를 직접 알 수 없기 때문에
  - 표본의 변동성을 통계적으로 모델링하여
  - "이 안에 참값이 있을 것이다"라는 구간(신뢰구간)을 설정

📈 **따라서 신뢰구간의 확률은** 구간이 아니라, **그 구간을 만드는 절차가 신뢰할 만한 확률적 성질**을 가진다는 의미

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간이란?

신뢰구간(confidence interval)

- 모수(parameter)가 특정 확률적 신뢰수준으로 포함되는 값의 범위
  - "모평균은 95% 신뢰수준에서 [a, b] 사이에 있다"고 표현

확률과 통계적 추론 II: 정규분포와 신뢰구간

95% 신뢰구간

신뢰수준 95%란 의미

- 동일한 추정 과정을 100번 반복하면 약 95번은 구간이 참된 모수를 포함
- 주의할 점은 특정 연구에서 구한 하나의 신뢰구간이 모수를 포함할 확률이 95%라는 의미는 아님 (모수는 고정되어 있고, 구간이 확률적).

올바른 해석: 추출된 표본에 기반한 추정 방법의 장기적 신뢰도가 95%라는 것이며 일반적으로 신뢰구간 = 점추정 $\pm$ (임계값) $\cdot$ (표준오차) 형태로 산출

확률과 통계적 추론 II: 정규분포와 신뢰구간

95% 신뢰구간

모집단 평균 μ에 대한 95% 신뢰구간을 유도: 표본평균 $\bar{X}$ 사용

표본평균의 분포 (표본크기 n 충분히 크다고 가정): $\bar{X} \sim N!\left(\mu, \frac{\sigma^2}{n}\right)$

- 표준오차 (모평균 추정의 표준편차) = $\frac{\sigma}{\sqrt{n}}$

확률과 통계적 추론 II: 정규분포와 신뢰구간

95% 신뢰구간

정규분포 성질 이용: $\Pr\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$

이 부등식을 μ에 대해서 풀면: $\Pr\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$

따라서 **95% 신뢰구간**: $\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

- 모표준편차 σ를 알 경우 (또는 n이 충분히 커서 추정 $\hat{\sigma}$로 대체) 적용 가능

확률과 통계적 추론 II: 정규분포와 신뢰구간

95% 신뢰구간: 여론조사에서 모비율 신뢰구간

한 여론조사에서 표본 비율 $\bar{p} = 0.43$ (43%)를 얻었다고 하자(응답자 수 $n = 900$).

- 관심 모수: 모집단 비율 p (전체 유권자 중 지지율)
- 표본비율의 표준오차: $\mathrm{SE} = \sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}} \approx \sqrt{\dfrac{0.43 \times 0.57}{900}} \approx 0.0165$ (1.65%p)
- 95% 신뢰구간: $\bar{p} \pm 1.96 \times \mathrm{SE} = 0.43 \pm 1.96(0.0165)$
- 계산: $1.96 \times 0.0165 \approx 0.0323$ (약 3.2%p)
- CI: [0.398, 0.462], 즉 39.8% ~ 46.2%

확률과 통계적 추론 II: 정규분포와 신뢰구간

95% 신뢰구간: 여론조사에서 모비율 신뢰구간

해석: "모집단 지지율 p가 95% 신뢰수준에서 약 40%에서 46% 사이"

- **신뢰수준 95%**로 모비율 p가 39.8%~46.2% 사이에 있다고 표현
- 우리가 이런 조사/추정을 반복하면, 95%의 경우 참 p가 해당 범위에 들 것이라는 의미

잘못된 해석: "참 값이 95% 확률로 이 구간 안에 있다." (X)

- 모수 p는 고정값이므로 틀린 진술 (확률개념은 구간 추출 과정에 대한 것)

신뢰구간은 표본오차의 범위를 직관적으로 보여줌: 표본평균 $\pm$ 오차범위(margin of error)

- 위 예에서 $\pm$ 3.2%p가 오차범위; 언론에서는 "표본오차 $\pm$ 3.2%p (95% 신뢰수준)"으로 표기

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준(confidence level)과 유의수준(significance level)

신뢰수준 ($1-\alpha$): 구간추정이나 검정에서 우리가 **틀리지 않기를 바라는 정도**

유의수준 (α): 우리가 허용하는 최대 오류 확률 (보통 0.05, 즉 5%)

- 95% 신뢰수준 $\rightarrow \alpha$ = 0.05
- 99% 신뢰수준 $\rightarrow \alpha$ = 0.01

신뢰수준과 유의수준은 서로 보완적: 신뢰수준 = $1-\alpha$

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준(confidence level)과 유의수준(significance level)

유의수준 (α): "우리가 잘못 기각할 위험을 감수하는 확률"

- α = 0.05 란 95%의 신뢰수준을 의미하며, 동시에 실제로 차이가 없는데도 5%의 확률로 "차이가 있다"고 결론낼 위험을 의미

구분	의미	예시	확률
1종 오류 (Type I)	H_0이 참인데 기각	코로나19에 걸리지 않았는데 양성이라고 판단 (false positive)	α
2종 오류 (Type II)	H_0이 거짓인데 기각하지 않음	코로나19에 걸렸는데 음성이라고 판단 (false negative)	β
검정력 (Power)	H_0이 거짓일 때 기각할 확률	코로나19에 걸렸을 때 양성으로 정확히 판단할 확률	1-β

α는 너무 엄격하게 설정해도 검정력이 줄어드는 상충관계(trade-off)가 존재

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준(confidence level)과 유의수준(significance level)

임계값(critical value)

임계값은 귀무가설을 기각할지 말지를 결정하는 **경계선**

- 양측검정에서 α = 0.05 $\rightarrow$ 각 꼬리에 0.025씩
- 표준정규분포 기준 임계값: $\pm$ 1.96

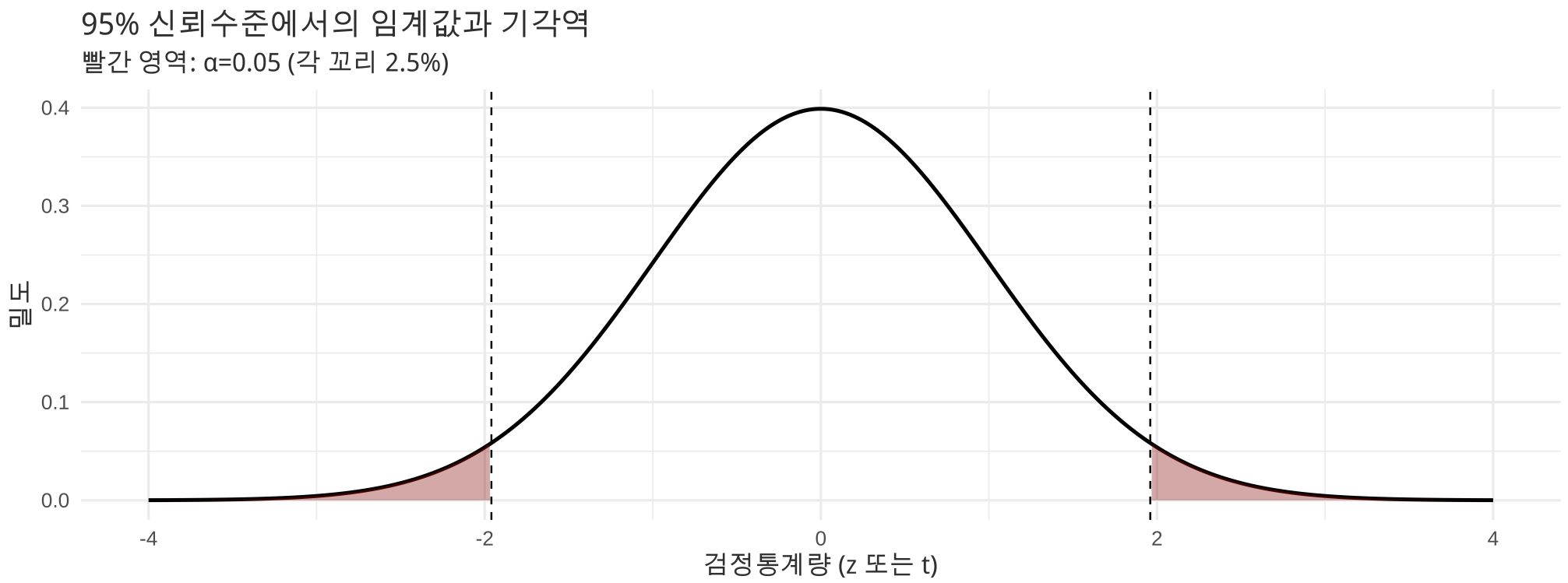
표본이 평균에서 $\pm$ 1.96 표준오차 이상 떨어져 있으면, 귀무가설로부터 너무 멀다 $\rightarrow$ 기각

- 다음 주의 가설검정 파트에서 더 자세히 배울 것임. 특히, "기각"(reject)이라는 것에 대해.
- 기본적으로 '마땅히 그러한 것'으로부터 아주 큰 차이가 있는 '보기 드문 결과'인지 여부를 확인하는 것이라고 이해하면 됨.

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준(confidence level)과 유의수준(significance level)

임계값(critical value)



확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준과 구간 폭

신뢰수준 선택에 따라 구간의 폭이 달라짐.

- 90% 신뢰구간: 약 $\bar{X} \pm 1.645, \mathrm{SE}$ (조금 더 좁은 구간)
- 95% 신뢰구간: $\bar{X} \pm 1.96, \mathrm{SE}$
- 99% 신뢰구간: 약 $\bar{X} \pm 2.576, \mathrm{SE}$ (더 넓은 구간)

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준과 구간 폭

신뢰수준 ↑ ⇒ 구간 폭 ↑ (더 높은 신뢰 확보 위해 범위를 넓게 잡음)

- 앞서 여론조사 예에서 99% 신뢰구간은 $43\% \pm 4.2\%p \to [38.8\%, 47.2\%]$ (약 8.4%p 폭)
- 90% 신뢰구간은 $43\% \pm 2.7\%p \to [40.3\%, 45.7\%]$ (약 5.4%p 폭)

연구 목적에 따라 90%, 95%, 99% 등 신뢰수준 선택 (95%가 관행적으로 많이 사용)

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준과 p-값

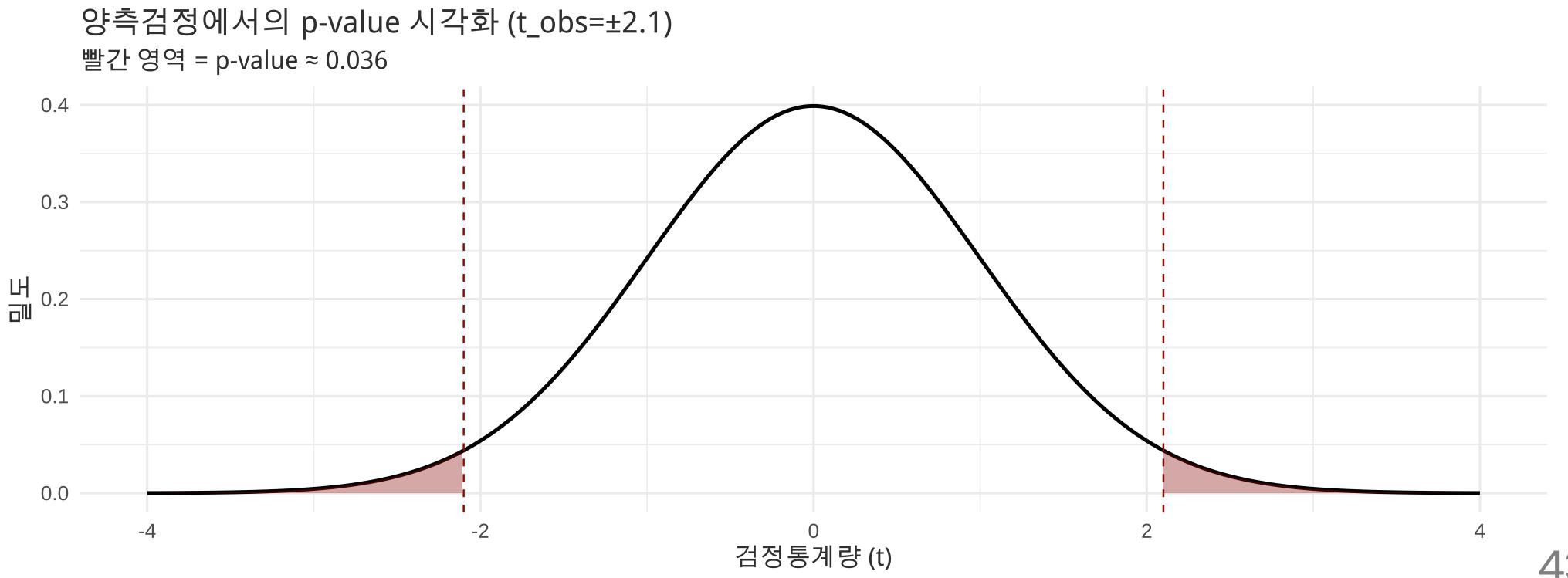
귀무가설이 참일 때, 현재의 데이터보다 더 극단적인 결과가 나올 확률

- 가설검정 파트에서 자세히 배우겠지만, 어떤 원인이 결과에 미치는 효과를 기대할 때 우리는 "그 효과가 없을 것"(H_0 = 0, 귀무가설 = 효과가 없음)이라고 가설을 설정
- 그리고 데이터를 통해 그러한 귀무가설(효과가 없다는 가설)의 주장과는 달리 0보다 유의미하게 더 큰 효과를 발견했을 때, "효과가 없다"는 경험적 근거가 없다(귀무가설의 기각)고 주장하며 기대한 효과를 간접적으로 검정
- p = 0.03일 때, 이는 '귀무가설이 참'(효과가 실제로 없다)일 때 이런 데이터가 나올 확률은 3%라는 것을 의미
    - 즉, 매우 드문 일이므로 귀무가설이 참(효과가 없다)이라는 가설을 기각
    - $p \leq \alpha$ 이면 귀무가설을 기각(유의미한 차이), $p > \alpha$이면 기각 불가

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준과 p-값

p-값은 표본에서 계산된 t-통계량의 위치를 기준으로 그보다 **더 극단적인 값이 나올 확률**을 면적으로 계산한 것



확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰수준, 임계값, 그리고 p-값

임계값 접근 vs p-값 접근

계산한 $|t|$ 값을 임계값과 비교

- $|t| \geq t$-통계량 $\leftrightarrow p \leq \alpha$ 이면 H_0 기각
- $|t| < t$-통계량 $\leftrightarrow p > \alpha$ 이면 H_0 기각 불가

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간의 의미 확인: 시뮬레이션

모집단 분포를 알고 있을 때, 구한 신뢰구간들이 얼마나 모수를 포함하는지 확인

- 모집단: $N(\mu = 100, \sigma = 15)$에서 20개의 표본 추출 (각 표본크기 $n = 30$)
- 각 표본마다 95% 신뢰구간 계산

20개 신뢰구간 중 대략 95%인 19개 정도가 모평균 100을 포함하리라 기대

확률과 통계적 추론 II: 정규분포와 신뢰구간

신뢰구간의 의미 확인: 시뮬레이션

R Code | Visualization | Plot | Results

```r
set.seed(123)  # 재현 가능하게 시드 설정
mu <- 100; sigma <- 15; n <- 30
B <- 20  # 표본 개수
ci_data <- data.frame(sample = 1:B, mean = numeric(B),
                      lower = numeric(B), upper = numeric(B))
for(i in 1:B) {
  samp <- rnorm(n, mean = mu, sd = sigma)
  xbar <- mean(samp); s <- sd(samp)
  ci_data$mean[i] <- xbar
  # 95% CI: xbar ± 1.96 * s/sqrt(n)
  ci_data$lower[i] <- xbar - 1.96 * (s / sqrt(n))
  ci_data$upper[i] <- xbar + 1.96 * (s / sqrt(n))
}
```

확률과 통계적 추론 II: 정규분포와 신뢰구간

표본크기의 영향

표본 크기 n에 따라 **표준오차(SE)**가 달라짐: $\mathrm{SE} \propto \dfrac{1}{\sqrt{n}}$

n이 커질수록 $\sqrt{n}$ 증가 $\rightarrow$ SE 감소 $\rightarrow$ 신뢰구간 폭 감소 (정밀도 증가)

같은 95% 신뢰수준에서 표본크기의 효과 변화

- $n \approx 1,000$일 때 오차범위 $\pm$ 3%p (여론조사에서 흔히 보는 크기)
- $n \approx 2,500$일 때 오차범위 $\pm$ 2%p (표본 2.5배로 약 1%p 감소)
- $n \approx 10,000$일 때 오차범위 ±1%p (표본 10배로 오차 ~1/3)

확률과 통계적 추론 II: 정규분포와 신뢰구간

표본크기의 영향

표본 크기 n에 따라 **표준오차(SE)**가 달라짐: $\mathrm{SE} \propto \frac{1}{\sqrt{n}}$

n이 커질수록 $\sqrt{n}$ 증가 $\rightarrow$ SE 감소 $\rightarrow$ 신뢰구간 폭 감소 (정밀도 증가)

같은 95% 신뢰수준에서 표본크기의 효과 변화

표본 증가의 한계: 정밀도 높이려면 n을 **크게** 늘려야 함 (비용 ↑)

통계적으로 유의미한 개선에는 큰 표본 증가 필요 (수확체감 발생)

확률과 통계적 추론 II: 정규분포와 신뢰구간

데이터 변동성과 구간 폭

모집단의 변동성(σ)이 클수록 표본평균의 표준오차도 커짐 $\rightarrow$ 신뢰구간이 넓어짐.

- 표본 100명으로 평균 키 추정 vs 평균 수입 추정: 수입이 변동성이 더 크면 CI도 더 넓음.

신뢰구간 폭은 세 요인의 함수: 구간폭 $\propto$ z값 $\times \sigma/\sqrt{n}$

- z: 신뢰수준에 따른 임계값 (높은 신뢰수준일수록 큼)
- σ: 모표준편차 (데이터 산포)
- n: 표본크기

확률과 통계적 추론 II: 정규분포와 신뢰구간

데이터 변동성과 구간 폭

구간폭 줄이는 법

- 신뢰수준을 낮추거나 ($z \downarrow$), 하지만 신뢰 하락
- 표본크기 크게 ($n \uparrow$), 비용/시간 제약 고려
- 변동성 낮은 모집단 연구, 연구 주제에 따라 한계 있음

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 검정

우리는 '차이'를 알고 싶다

신뢰구간은 모평균이 어디쯤 있을지를 알려주는 일종의 '추정'

하지만 연구자는 종종 이러한 질문을 가지게 됨:

> 우리가 가진 이 표본'들'이 하나의 모집단에서 비롯된 것일까?
>
> 혹은 하나의 모집단에서 서로 다른 표본들이 '우연히' 차이가 나게된 것일까?

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 검정

우리는 '차이'를 알고 싶다

사실, **신뢰구간과 (가설)검정은 같은 논리의 두 표현**

접근	질문	해석
신뢰구간	참값이 이 범위 안에 있을까?	기준값과 통계적으로 다른가?
가설검정	범위 안/밖으로 판단	기각 / 비기각으로 판단

두 집단의 차이가 존재하는지를 확인하고자 할 때, 95% 신뢰구간이 [6.2, 7.1]일 때

- 관측한 두 집단의 차이가 7.00이면 → **유의미한 차이가 없음** (구간 안)
- 관측한 두 집단의 차이가 8.00이면 → **유의미한 차이가 있음** (구간 밖)

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 검정

일표본 검정 (One-sample t-test)

하나의 표본 평균이 **특정 기준(모평균)**과 다른지를 검정하는 방법

- 귀무가설 H_0: $\mu = \mu_0$
- 대립가설 H_A: $\mu \neq \mu_0$

> 어떤 에너지드링크의 카페인 함량 표준은 80mg이라 할 때, 표본 30개의 평균이 83mg, 표준편차 10mg이었다면?

95% 신뢰구간:

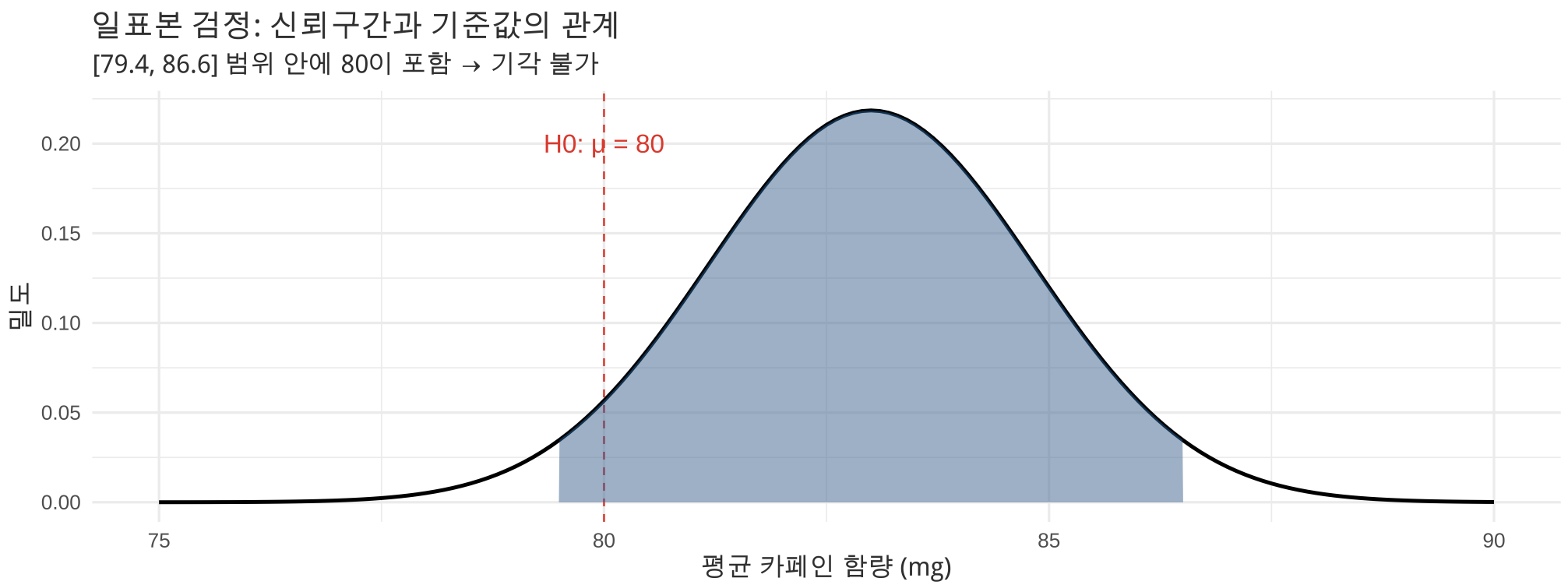
$$83 \pm 1.96 \times \frac{10}{\sqrt{30}} = [79.4, 86.6]$$

80mg이 구간 안에 있음 $\rightarrow$ H_0 **기각 불가**

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 검정

일표본 검정 (One-sample t-test)



확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 검정

일표본 검정 (One-sample t-test)

신뢰구간 관점: 기준값(μ_0)이 95% 신뢰구간 안에 있다면, 유의미한 차이 없음.

검정통계량 관점: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$

- 예시에서 $t = \frac{83-80}{10/\sqrt{30}} = 1.64$
- 임계값 1.96보다 작음 $\rightarrow$ H_0 기각 불가

*두 접근은 동일한 결론을 준다.**: 표본평균이 기준값과 다르다고 할 근거는 없다.

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 검정

이표본 검정 (Independent two-sample test)

두 개의 독립된 집단 평균이 통계적으로 같은지 다른지를 검정

- 귀무가설 H_0: $\mu_1 = \mu_2$
- 대립가설 H_A: $\mu_1 \neq \mu_2$

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 가설검정

이표본 검정 (Independent two-sample test)

남학생 40명: 평균 172.5cm, SD 6.8

여학생 35명: 평균 168.3cm, SD 7.1

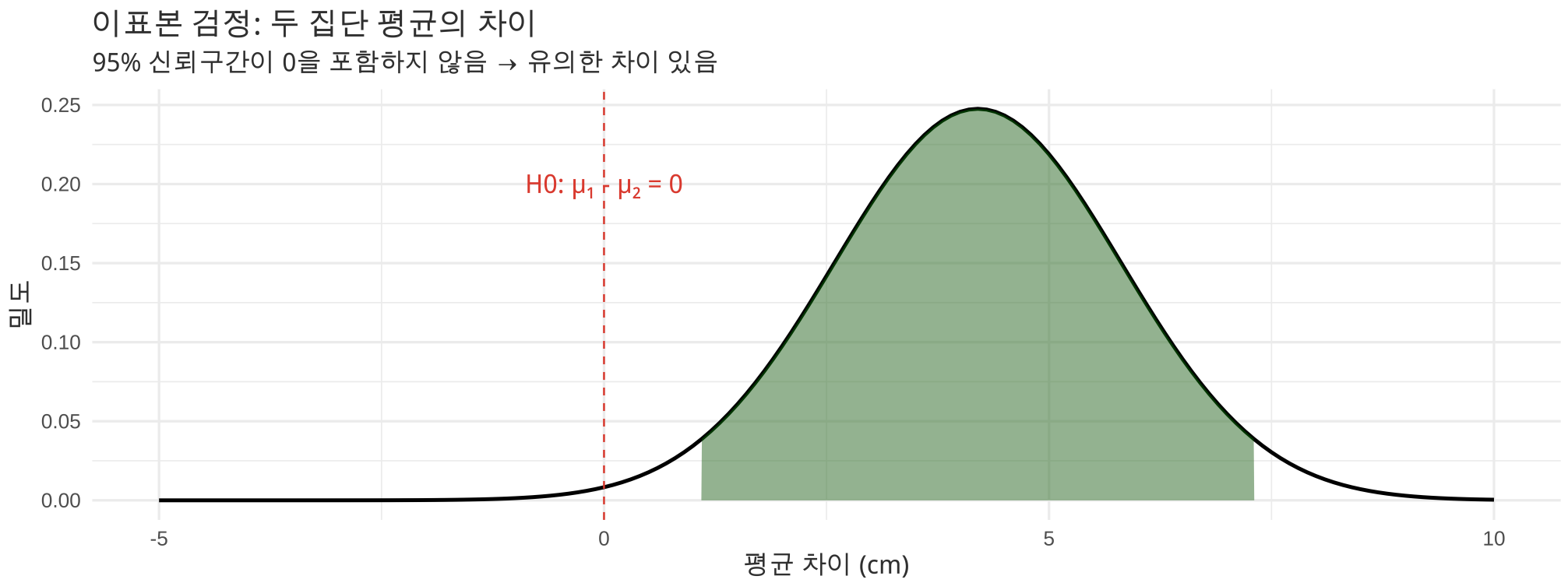
95% 신뢰구간 $(\mu_1 - \mu_2)$: $(172.5 - 168.3) \pm 1.96 \times SE = [2.1, 6.3]$

- 0이 포함되지 않으므로 통계적으로 **유의미한 차이 있음.**

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 가설검정

이표본 검정 (Independent two-sample test)



확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 가설검정

이표본 검정 (Independent two-sample test)

신뢰구간 접근: $(\mu_1 - \mu_2)$의 신뢰구간이 0을 포함하지 않으면, 두 집단 평균의 차이가 통계적으로 유의함.

검정통계량 접근: $t = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$

- 예시에서 $t = \frac{4.2}{1.05} = 4.0$
- 임계값 1.96보다 큼 → H_0 기각 (p < 0.001)

신뢰구간과 t검정은 같은 이야기를 서로 다른 언어로 표현한 것

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 가설검정

일표본 vs 이표본 검정

구분	일표본 검정	이표본 검정
비교 대상	기준값(μ_0)	두 집단의 평균(μ_1 vs μ_2)
귀무가설	$\mu = \mu_0$	$\mu_1 = \mu_2$
검정 기준	μ_0가 신뢰구간 안에 있는가?	($\mu_1-\mu_2$)의 신뢰구간에 0이 포함되는가?
주요 해석	한 집단의 평균이 특정 기준과 다른가	두 집단의 평균이 서로 다른가

🎯 **신뢰구간과 가설검정은 서로 보완적**: 구간 안에 있다/없다 = 기각한다/하지 않는다

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 가설검정

퀴즈: 일표본 검정

표본: 30개 제품의 평균 수명 83일, 표준편차 10일

가정: H_0: μ = 80

이때의 t-통계량은? **HINT**: $t = \frac{\bar{X}-\mu}{\s/\sqrt{n}}$

- $t = \frac{83-80}{10/\sqrt{30}} = 1.64$
- p = 0.11 (단측검정), p = 0.22 (양측검정)
    - α = 0.05 기준으로는 p > 0.05이므로 H_0 기각 불가. 즉, 차이가 있다고 보기 어렵다.

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 가설검정

퀴즈: 이표본 검정

남학생 평균 172.5, 여학생 평균 168.3, $n_{\text{남학생}} = 40$, $n_{\text{여학생}} = 35$, $s_{\text{남학생}} = 6.8$, $n_{\text{여학생}} = 7.1$

이때의 t-통계량은? **HINT**: $t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

- $t = \frac{172.5 - 168.3}{\sqrt{6.8^2/40 + 7.1^2/35}} \approx 2.61$
- p $\approx$ 0.001 (양측검정): 두 집단 평균은 통계적으로 다르다(H_0 기각)

확률과 통계적 추론 II: 정규분포와 신뢰구간

약간 맛보기: 신뢰구간과 가설검정

신뢰수준, 유의수준, p-값의 연결구조

신뢰수준이 높을수록(예: 99%) 유의수준이 작아지며 기각하기 어려워짐.

유의수준 α가 작을수록 임계값이 커지므로 더 엄격한 검정을 수행하는 것이 됨.

p-값은 데이터가 얼마나 극단적인지 나타내며, p-값이 유의수준보다 작거나 같으면 귀무가설을 기각

신뢰구간 밖에 있다는 이야기는 p-값이 유의수준보다 작거나 같다는 얘기와 같으며, 통계적으로 유의미한 차이가 존재한다는 것

🎯 **결론**: 신뢰구간, 임계값, p-value는 모두 같은 질문을 다른 방식으로 표현한 것

이 표본이 정말 우연일까?”

Part III. R을 이용한 시각화 실습 및 질의응답

다음 강의에는 가설검정의 논리에 대해 살펴볼 것

KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

감사합니다!

궁금한 것이 있으면 언제든 연락하세요.

강사 연락처

연락처	박상훈
	sh.park.poli@gmail.com
	sanghoon-park.com/
	영상바이오관 405