

검정 Test

CE730 통계와 금융

조남운

목차

- 검정의 이해
- 통계적 검정
- 좋은 검정법
- 좋은 검정법의 선택
- 모평균과 모비율에 대한 검정
- 카이제곱분포를 이용한 검정법

$$\text{Sharpe Ratio} := \frac{\bar{R}_p - \bar{R}_F}{\sigma_p}$$

검정 Test

- 임의표본을 통해 모집단의 특성을 합리적으로 파악하고자 하는 일련의 절차
- Question Example
 - 투자자문회사 A는 올해 운용중인 포트폴리오 샤프비 평균이 작년의 3.0보다 높은지 알기 위해 자신의 포트폴리오 중 일부의 SR 평균을 산출한뒤 3.4가 나왔음을 확인함.
 - ⇒ 자, 그럼 올해는 3.0보다 높다고 할 수 있을까?

다른 사례: Fischer's Lady Tasting Tea

- 밀크티 := 홍차 + 우유
- 1920년대에 통계학자들도 동석하고 있던 한 티파티에서 한 여성이 다음과 같은 주장을 했음
- “나는 우유를 봇고 홍차를 부은 밀크티와 홍차를 부은 뒤 우유를 부은 밀크티를 구분할 수 있다”
- 이에 즉석에서 실험이 진행됨



실험 디자인

- 가설: 이 부인은 두 타입의 밀크티를 식별할 수 있는가?
 - 또는, 밀크티는 어떤 것을 먼저 부었는지에 따라 맛이 달라지나?
- 8잔의 밀크티를 준비
 - 4잔은 우유먼저, 나머지 4잔은 홍차를 먼저 부음
 - 순서를 뒤섞고 각 밀크티의 일련번호와 타입을 기록해둠 (부인은 알 수 없는 방식으로)
- 부인은 한잔씩 맛을 보고 어떤 타입인지 식별
- 통계적 가설검정을 통해 가설을 **검정**

Question

- “만일 이 부인이 7잔은 맞추고 1잔을 틀렸다면, 우리는 이 부인이 두 버전의 홍차를 구별할 수 있다고 볼 수 있을까?

사례3: Unfair Dice?

- 상대방의 패가 유난히 잘 나온다
 - 주사위 6이 5연속 나왔다
 - Q: 상대방은 속임수를 쓴 것일까?



Answer

- 그럴 수도 있고, 아닐 수도 있다!
 - 모집단의 분산에 달려있다
 - 모분산을 모른다면?
 - 모분산도 추정할 수 있다
- 어떻게?
 - CLT를 이용: 표본평균의 분산은 표본 크기가 커질때 정규분포에 수렴한다!

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

모분포는 모르지만,
표본평균의 분포는
알 수 있다!

모분산을 모를 경우

- 만일 모분산을 모른다면?
(대부분의 경우가 여기에 해당됨)
- 표본분산 s^2 를 사용했을 때의 분포를 알고 있다!
(t분포)

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \xrightarrow{d} t(n-1)$$

모분포는 모르지만,
표본평균의 분포는
알 수 있다!

다시, 첫번째 질문으로

- 첫번째 문제: “표본평균이 3.4였다”
- Case1
 - 표본의 수는 31개
 - 모분산이 0.8^2 이었다
$$\frac{3.4 - 3.0}{0.7/\sqrt{31}} \approx 2.74$$
$$\Rightarrow P(Z > 2.74) = 0.00339$$
- Case2
 - 표본의 수는 31개
 - 모분산은 모르지만
표본분산은 1^2 이었다
$$\frac{3.4 - 3.0}{1/\sqrt{31}} \approx 2.227$$
$$\Rightarrow P(t(df = 30) > 2.27) \approx 0.015278$$

계산한 확률값의 의미

$$\Rightarrow P(Z > 2.74) = 0.00339$$

$$\Rightarrow P(t(df = 30) > 2.27) \approx 0.015278$$

- p-value
- 같은 모집단에서 추출된 두 샘플들의 차이가 0.4만큼 나타날 확률
 - 다시 말하면, 변동이 없다는 전제하에 이런 결과를 얻을 확률이 0.0039 (Case1), 0.015278 (Case2)라는 것
 - 즉, 일어날 확률이 0에 가깝다 \Rightarrow 차이나는 것은 모집단이 다를 가능성이 높다! \Rightarrow 유의미한 차이다!
- 검정의 핵심 아이디어

통계적 검정의 절차

- 검정을 위한 가설을 세운다 (귀무가설, 대립가설)
 - 연구가설: 대립가설, 검증가설: 귀무가설
 - ex) $\mu > 3$, Lady는 홍차맛 구별 가능
- 표본으로부터 모평균의 좋은 추정량을 도출
 - ex) 모평균 μ 의 좋은 추정량으로 표본평균 선정
- p-value 산출
 - ex) 검정을 위한 귀무가설 전제에서 표본추정량이 관측될 확률을 구함
 - 0에 충분히 가까우면 귀무가설 기각

귀무가설, 대립가설

- 대립가설 (Alternative Hypothesis)
 - 연구자가 알고자 하는 모수영역
 - 예1: 올해의 SR은 3 이상인가?
 - 예2: 저 사람은 홍차맛을 구별할 줄 아는가?
- 귀무가설 (Null Hypothesis)
 - 대립가설에 반대되는 모수영역
 - 예1: 올해의 SR은 3과 차이가 없다
 - 예2: 저 사람은 홍차맛을 구별하지 못한다

p-value의 의미

$$P(\bar{X} \geq 3.4 | \mu = 3.0)$$

- 앞의 예에서 구한 p-value는 귀무가설($\mu = 3.0$) 하에서 표본추정량(3.4)이 관측될 확률임
- 귀무가설은 알고 싶은 것을 뒤집어 만든 것
 - 낮은 p-value: 아닌 것이 확실함
 - 높은 p-value: 맞는 것이 확실한 것은 아님
- 확률론적 검정은 오직 기각만을 할 수 있음
 - 귀무가설을 만드는 이유

아닐 때에만 진실을 이야기하는 거울을 다루는 법

- 만일 질문에 대한 답이 “아니다” 일 때에만 진실을 이야기하는 마술 거울이 있다고 생각해보자
- B는 A가 자신을 좋아하는지 알고자 한다.
 - 질문을 어떻게 만들어야 할까?
- “A는 B를 좋아하니?”
 - 좋아할 경우 \Rightarrow 무응답 (진실을 알 수 없음)
 - 좋아하지 않을 경우 \Rightarrow 응답 (아니오) \Rightarrow 진실을 알았으나 알고자 하는 것과는 다름



1종, 2종 오류

Type I, II Error

- 1종 오류
 - “아버지를 아버지라 부르지 못하고..”
 - 진실은 TRUE인데 FALSE로 추정하는 오류
- 2종 오류
 - “아버지가 아닌데 아버지라 부르고..”
 - 진실은 FALSE인데 TRUE로 추정하는 오류

		진실	
		“타짜”	속임수 안씀
거짓	거짓	귀무가설 기각X	OK
	眞	귀무가설 기각	Type I Error

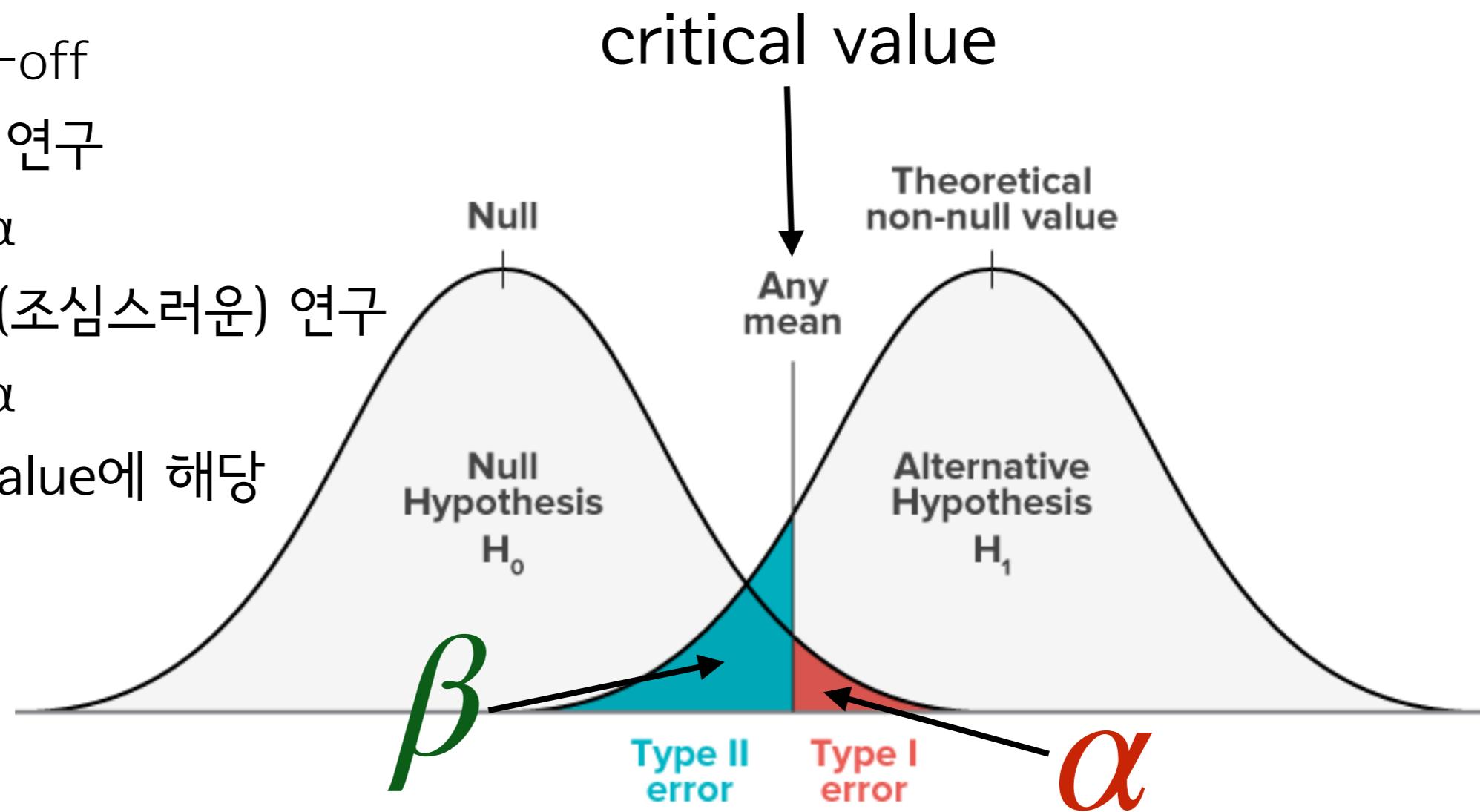
		Reality	
		H_0 Is True	H_1 Is True
Conclusion	Do Not Reject H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

1,2종 오류의 의미

$$\alpha := \Pr(\bar{X} > c | \mu)$$

$$\beta := \Pr(\bar{X} < c | \text{not } \mu)$$

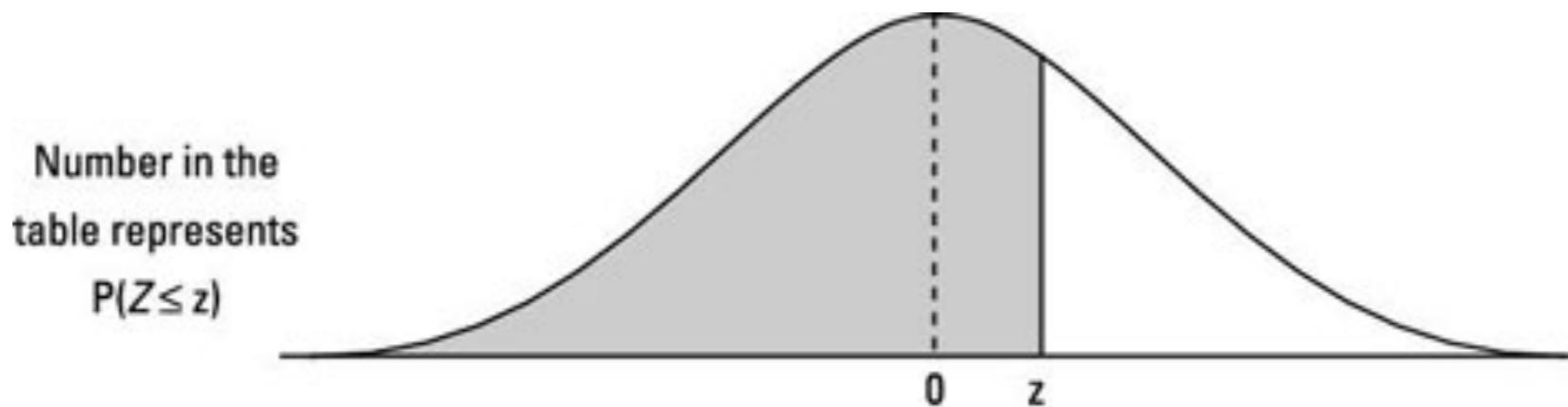
- 1,2종 오류를 모두 줄일 수는 없다
 - trade-off
- 급진적인 연구
 - 높은 α
- 보수적인(조심스러운) 연구
 - 낮은 α
- α 가 p-value에 해당



추정 관례

- 제1종 오류 수준 (α)를 고정한 상태에서
- 제2종 오류 수준 (β)를 최소화하는 검정법
- 분야마다 α 의 수준은 다름
 - 자연과학계열: 대체로 낮은 편 $10e-6$
 - 사회과학계열: 대체로 높은 편 0.05

예제7.1



- SR 문제에서 SR이 $N(3, 0.644^2)$ 를 따르고 25개의 평균을 구했을 때 α 가 0.01에서 임계치를 구하기

$$\begin{aligned}\alpha &= \Pr(\bar{X} > c | \mu = 3) = 0.01 \\ &= \Pr\left(Z > \frac{c - 3}{0.644/\sqrt{25}}\right) = 0.01\end{aligned}$$

$$Z = 2.33 \Rightarrow c = 3.21$$

표준정규분포표로부터 산출

검정력

$$\alpha := \Pr(\bar{X} > c | \mu)$$

$$\beta := \Pr(\bar{X} < c | \text{not } \mu)$$

$$\Rightarrow 1 - \beta = \Pr(\bar{X} \geq c | \mu > 3)$$

- 검정력의 정의: $1 - \beta$
- 예제 7.2
 - $\mu = 3.1, 3.2, \dots$
 - 3.0 (귀무가설)에서 멀어질수록 검정력이 강해진다 \Rightarrow 바람직한 성질

검정 절차

- (1단계) 귀무가설과 대립가설을 세운다
- (2단계) 표본으로부터 검정통계량을 구한다
 - 가능한한 성질이 좋은 추정량을 사용
- (3단계) 귀무가설이 참이라는 가정 하에 임계치를 α (p-value, 혹은 유의수준 significance level)로 구하고 크기 비교
 - 크면 귀무가설 기각
 - 작으면 귀무가설 기각하지 않음

1단계: 귀무가설과 대립가설

- 귀무가설과 대립가설은 관심있는 모수가 가질 수 있는 모든 값의 범위를 2개 영역으로 분리한 것
- 샤프비(SR)의 예
 - 궁금한 것: 포트폴리오 수익률평균은 3보다 큰가?
 - 대립가설 (H_a): $\mu > 3.0$
 - 귀무가설 (H_0): $\mu \leq 3.0$
 - 단측검정
 - 양측검정: $H_0: \theta = \theta_0$ vs. $H_a: \theta \neq \theta_0$

SR case

- 그런데, 우리는 앞에서 귀무가설로 $\mu=3.0$ 을 사용하지 않았는가? 그러면 양측검정인가?
 - Answer:
 - 현재 관찰된 값은 $3.4 > 3.0$
 - 이 관찰된 값을 기반으로 모평균 $\mu = 3.0$ 일 확률은 $\mu = 2.9, 2.8, \dots$ 일 확률보다 높음
 - 따라서, $H_0: \mu = 3.0 | \bar{X} = 3.4$ 을 기각한다는 것은 그 가능성의 부분집합인 $\mu < 3.0$ 도 기각한다는 것은 자연스러운 추론

일반화

- 일반화: 검정통계량의 분포가 모수에 대하여 단조 증가[감소] 함수일 경우에는 아래와 같은 가설은 동일함

$$H_0 : \theta \leq \theta_0 \quad vs. H_a : \theta > \theta_0$$

$$H_0 : \theta = \theta_0 \quad vs. H_a : \theta > \theta_0$$

$$H_0 : \theta \geq \theta_0 \quad vs. H_a : \theta < \theta_0$$

$$H_0 : \theta = \theta_0 \quad vs. H_a : \theta < \theta_0$$

2단계: 표본을 통한 추정량 산출

- 표본으로부터 검정통계량을 구하되, 이 검정통계량의 표본분포가 추정하고자 하는 모수 (μ)에 "만"의 존해야 함
- SR의 예: 표본 포트폴리오 수익률의 표본평균은 모평균에 대한 가장 좋은 통계량임이 증명되어 있음 (Sample mean is the MVUE: Minimum-Variance Unbiased Estimator)
- 예제7.1: 표본분산이 0.8²인 경우
$$(\bar{X} - \mu) / (0.8 / \sqrt{25}) \sim t(df = 25 - 1)$$

2단계: 표본을 통한 추정량 산출

- 표본으로부터 검정통계량을 구하되, 이 검정통계량의 표본분포가 추정하고자 하는 모수 (μ)에 "만"의 존해야 함
- SR의 예: 표본 포트폴리오 수익률의 표본평균은 모평균에 대한 가장 좋은 통계량임이 증명되어 있음 (Sample mean is the MVUE: Minimum-Variance Unbiased Estimator)
- 예제7.1: 표본분산이 0.8²인 경우

$$(\bar{X} - \mu) / (0.8 / \sqrt{25}) \sim t(df = 25 - 1)$$

안다

2단계: 표본을 통한 추정량 산출

- 표본으로부터 검정통계량을 구하되, 이 검정통계량의 표본분포가 추정하고자 하는 모수 (μ)에 "만"의 존해야 함
- SR의 예: 표본 포트폴리오 수익률의 표본평균은 모평균에 대한 가장 좋은 통계량임이 증명되어 있음 (Sample mean is the MVUE: Minimum-Variance Unbiased Estimator)
- 예제7.1: 표본분산이 0.8²인 경우

$$(\bar{X} - \mu) / (0.8 / \sqrt{25}) \sim t(df = 25 - 1)$$

안다

안다

2단계: 표본을 통한 추정량 산출

- 표본으로부터 검정통계량을 구하되, 이 검정통계량의 표본분포가 추정하고자 하는 모수 (μ)에 "만"의 존해야 함
- SR의 예: 표본 포트폴리오 수익률의 표본평균은 모평균에 대한 가장 좋은 통계량임이 증명되어 있음 (Sample mean is the MVUE: Minimum-Variance Unbiased Estimator)
- 예제7.1: 표본분산이 0.8²인 경우

$$(\bar{X} - \mu) / (0.8 / \sqrt{25}) \sim t(df = 25 - 1)$$

안다

안다

안다

2단계: 표본을 통한 추정량 산출

- 표본으로부터 검정통계량을 구하되, 이 검정통계량의 표본분포가 추정하고자 하는 모수 (μ)에 "만"의 존해야 함
- SR의 예: 표본 포트폴리오 수익률의 표본평균은 모평균에 대한 가장 좋은 통계량임이 증명되어 있음 (Sample mean is the MVUE: Minimum-Variance Unbiased Estimator)
- 예제7.1: 표본분산이 0.8²인 경우

$$(\bar{X} - \mu) / (0.8 / \sqrt{25}) \sim t(df = 25 - 1)$$

안다

안다

안다

안다

3단계: Test 유의수준 α 인 임계치 c 구하기

$$P(\bar{X} > c | \mu = 3.0) = 0.05$$

- 위의 예에서 $c=3.21$ 임
- 표본평균이 3.21보다 클 경우 $\Rightarrow H_0$ 전제 하에서 이런 표본평균을 구할 확률(p-value)이 낮다 \Rightarrow 귀무가설 기각한다 \Rightarrow 귀무가설은 틀렸다 \Rightarrow 대립가설이 말이 된다 \Rightarrow 현재 전체 포트폴리오 SR은 과거보다 높다 유의수준 α 의 수준에서 이야기할 수 있다
- 그렇지 않을 경우 $\Rightarrow H_0$ 을 기각하지 못한다 \Rightarrow 귀무가설이 맞는지 틀린지 알지 못한다 \Rightarrow 현재의 포트폴리오 SR이 과거보다 높은지 모른다

UMP test

Uniformly Most Powerful test

- 하나의 가설검정에 대해 유의수준 α 를 만족하는 검정법은 무수히 많음
- UMP검정: 이 중 검정력 (대립가설 하에서 귀무가설을 기각할 확률 $1-\beta$)이 가장 큰 검정법
- UMPU: UMP보다 약한 검정법
 - UMP Unbiased test
 - 검정력이 유의수준 α 보다 큰 검정법 중 검정력이 가장 큰 검정법
 - 일치성(consistency): 표본크기가 커질 수록 $1-\beta$ 가 1이 수렴하는 (검정력이 커지는) 성질
- 자세한 내용은 수리통계학에서 다룸

p-value

$$p\text{-}value := P(\bar{X} > \bar{x} \mid \mu = 3.0)$$

- 귀무가설의 전제 하에서 관측된 표본평균 (small bar \bar{x})보다 큰 값이 나올 확률
- 이 값이 유의수준보다 낮으면 \Rightarrow 임계치보다 높은 값임 \Rightarrow 귀무가설 기각
- 아니면 \Rightarrow 임계치보다 낮은 값임 \Rightarrow 귀무가설 기각 못함 (즉, 이 데이터로는 확증할 수 없음)

$$p\text{-value} \left\{ \begin{array}{l} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{array} \right..$$

모평균 및 모비율에 대한 검정

Topics

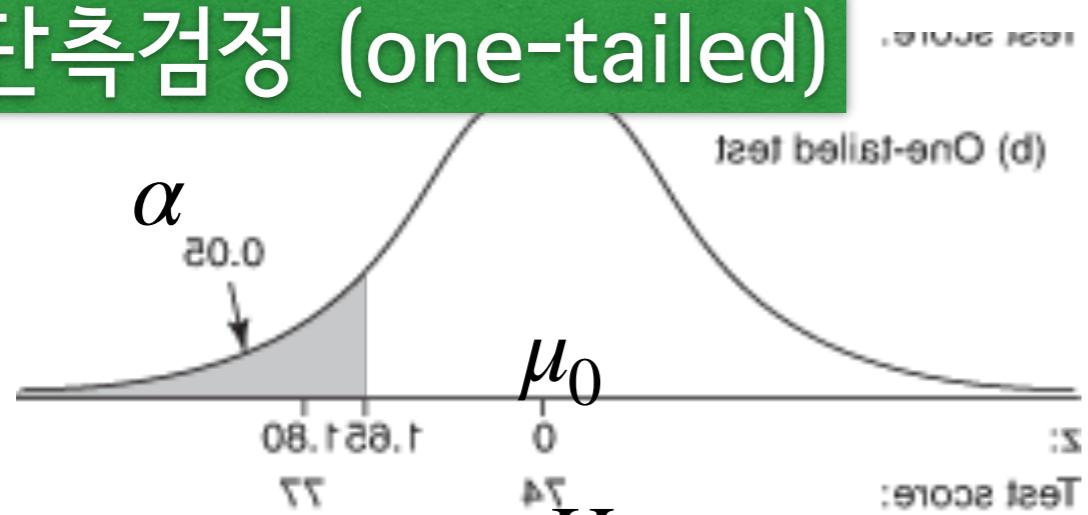
- 모평균 μ 에 대한 검정
 - 모분산이 알려진 경우
 - 표본평균이 정규분포를 따르지만 분산을 모르는 경우
- 모비율에 대한 검정
- 신뢰구간과 양측검정, 표본크기
- 두 개의 모집단 관련 검정

모평균 μ 에 대한 검정: 모분산이 알려진 경우 (Case 1)

$$p\text{-value} := P(\bar{X} \geq [\leq] \bar{x} | H_0)$$

- 기각역 (Rejection region)
 - H_0 전제하에서 현재의 샘플로 구성한 추정량이 나타날 확률이 유의수준 α 이하인 영역

(c) 단측검정 (one-tailed)

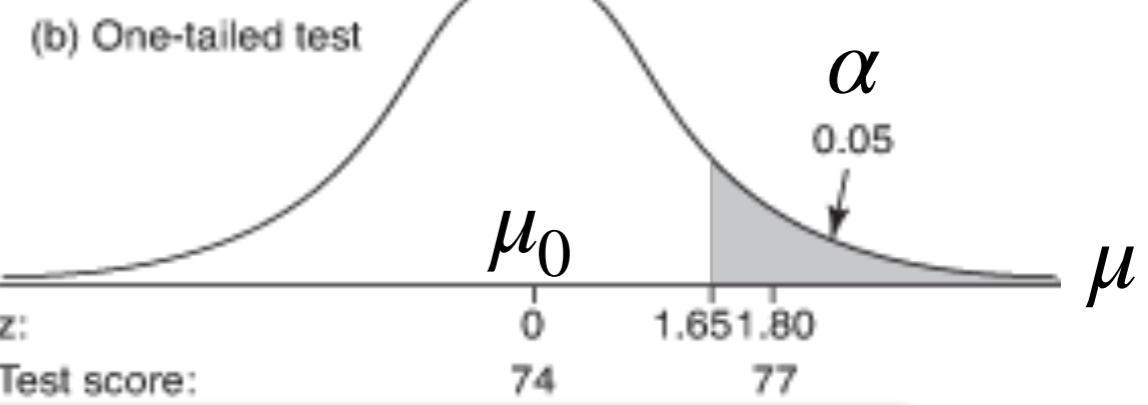
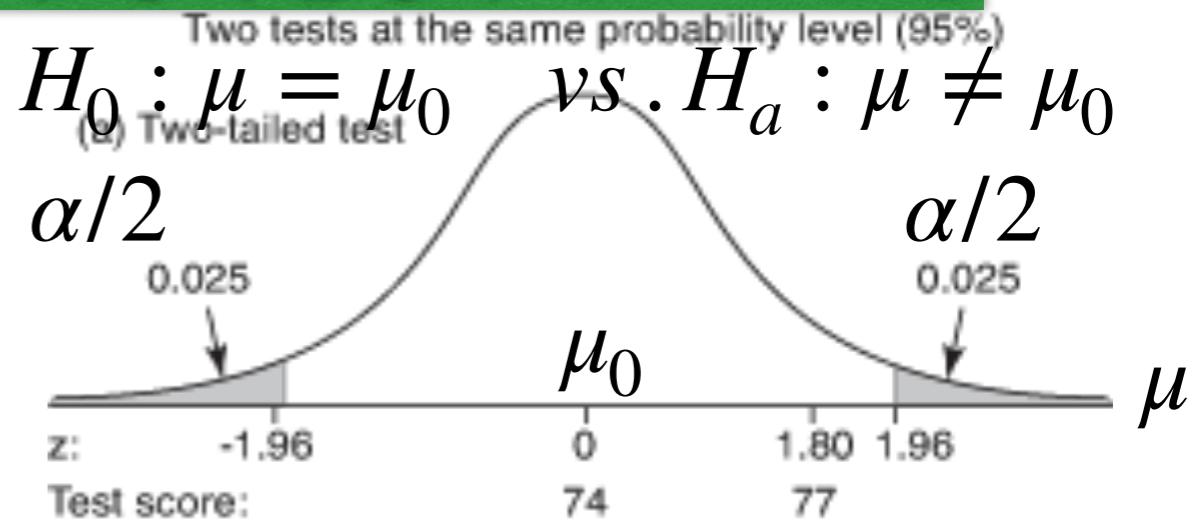


$$H_0 : \mu = \mu_0 \quad vs. \quad H_a : \mu < \mu_0$$

(a) 단측검정 (one-tailed)

$$H_0 : \mu = \mu_0 \quad vs. \quad H_a : \mu > \mu_0$$

(b) 양측검정 (two-tailed)



p-value Approach

- 샘플로부터 구한 통계량 x
 - 관측한 값 \bar{X}
- 샘플통계량 분포 \bar{X}
 - 모르는 값 \Rightarrow 확률변수
- p-value 구하기
 - 귀무가설이 참이라는 전제 하에 도출한 샘플통계량 분포상에서 관측한 통계량 x 가 나타날 확률

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

CLT

$$p\text{-value} \left\{ \begin{array}{l} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{array} \right..$$

p-value Approach

- 샘플로부터 구한 통계량 x
 - 관측한 값 \underline{x}
- 샘플통계량 분포 \bar{X}
 - 모르는 값 \Rightarrow 확률변수
- p-value 구하기
 - 귀무가설이 참이라는 전제 하에 도출한 샘플통계량 분포상에서 관측한 통계량 x 가 나타날 확률

안다

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

CLT

$$p\text{-value} \left\{ \begin{array}{l} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{array} \right.$$

p-value Approach

- 샘플로부터 구한 통계량 x
 - 관측한 값 \underline{x}
- 샘플통계량 분포 \bar{X}
 - 모르는 값 \Rightarrow 확률변수
- p-value 구하기
 - 귀무가설이 참이라는 전제 하에 도출한 샘플통계량 분포상에서 관측한 통계량 x 가 나타날 확률

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

안다

안다

CLT

$$p\text{-value} \left\{ \begin{array}{l} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{array} \right.$$

p-value Approach

- 샘플로부터 구한 통계량 x
 - 관측한 값 \bar{X}
- 샘플통계량 분포 \bar{X}
 - 모르는 값 \Rightarrow 확률변수
- p-value 구하기
 - 귀무가설이 참이라는 전제 하에 도출한 샘플통계량 분포상에서 관측한 통계량 x 가 나타날 확률

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

CLT

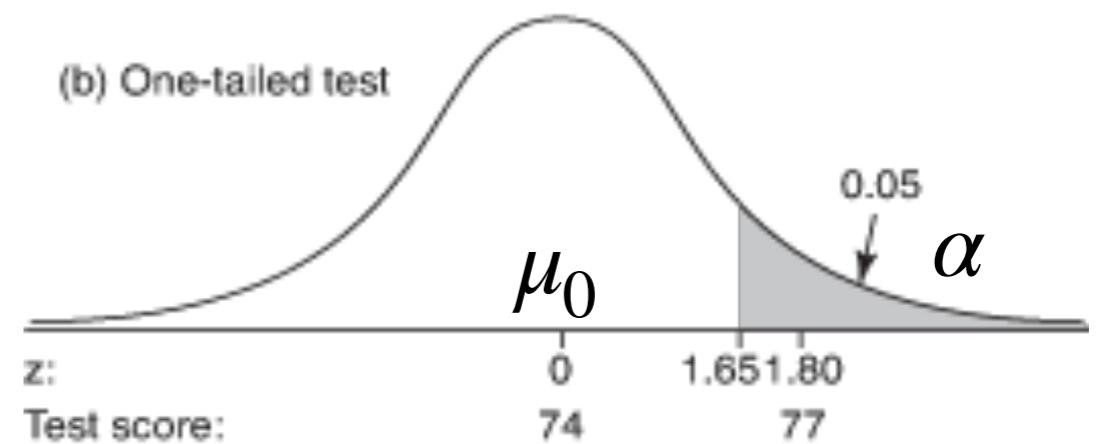
안다
안다
모른다

$$p\text{-value} \left\{ \begin{array}{l} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{array} \right.$$

Case (1-a)

p-value =

$$\begin{aligned} P(\bar{X} > \bar{x} | \mu = \mu_0) &= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$



(a) 단측검정 (one-tailed)

$$H_0 : \mu = \mu_0 \quad vs. \quad H_a : \mu > \mu_0$$

$$p\text{-value} \left\{ \begin{array}{l} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{array} \right..$$

Case (1-b)

When $\bar{x} > \mu_0$:

p-value =

$$\begin{aligned} P(\bar{X} > \bar{x} | \mu = \mu_0) &= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

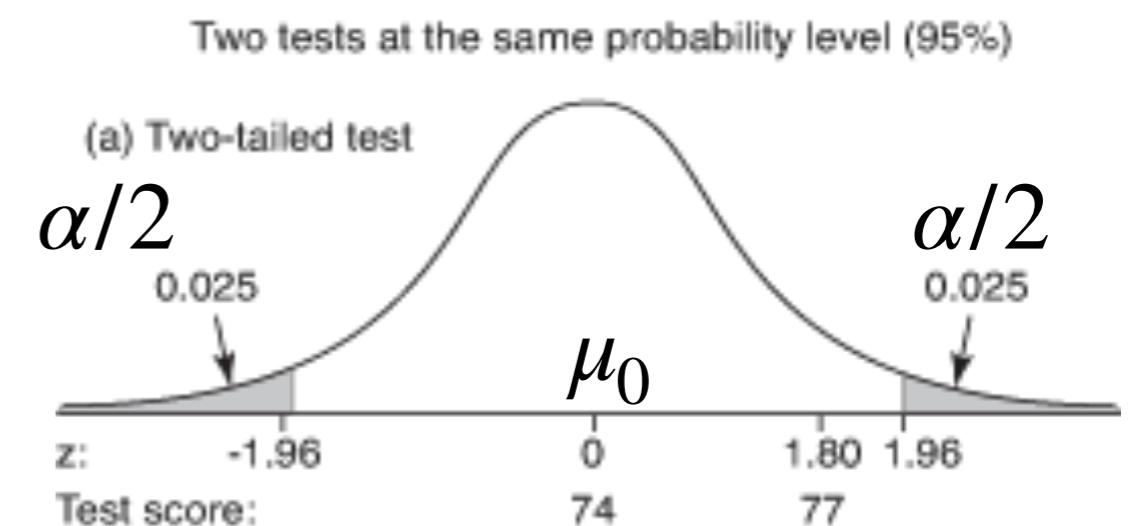
When $\bar{x} < \mu_0$:

p-value =

$$\begin{aligned} P(\bar{X} < \bar{x} | \mu = \mu_0) &= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

$$p\text{-value} \begin{cases} < \alpha/2 \Rightarrow \text{REJECT } H_0, \\ \geq \alpha/2 \Rightarrow \text{DO NOT REJECT } H_0 \end{cases} \cdot$$

(b) 양측검정 (two-tailed)



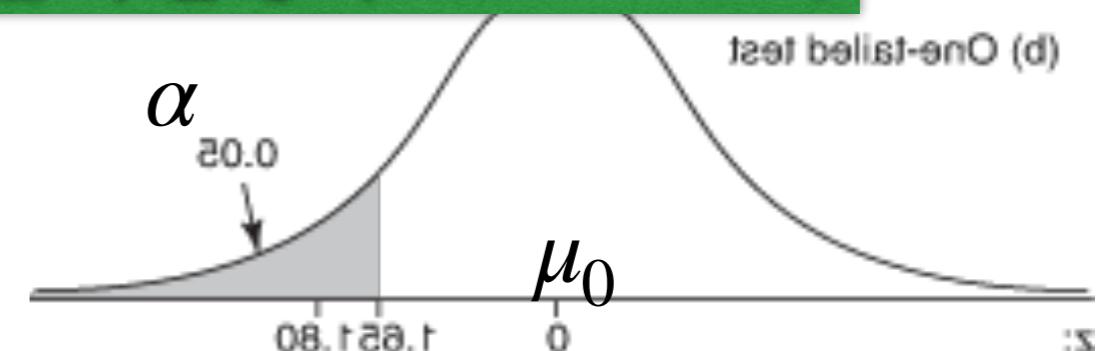
$$H_0 : \mu = \mu_0 \quad vs. \quad H_a : \mu \neq \mu_0$$

Case (1-c)

p-value =

$$\begin{aligned} P(\bar{X} < \bar{x} | \mu = \mu_0) &= P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z < \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

(c) 단측검정 (one-tailed)



$$H_0 : \mu = \mu_0 \quad vs. \quad H_a : \mu < \mu_0$$

$$p\text{-value} \left\{ \begin{array}{l} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{array} \right.$$

모평균 μ 에 대한 검정: 표본평균이 정규분포를 따르나 분산을 모르는 경우 (Case 2)

- 만일 n 이 크다면 (통상 30 이상) CLT에 의해 표본 평균은 정규분포를 따름 (본 케이스에 해당)
- 만일 CLT가 적용되지 않을 정도로 작은 표본이라면
 - 정규분포일 경우 $\rightarrow t$ 분포로 푼다 (본 케이스)
 - 어떤 분포인지 모르는 경우 \rightarrow 풀 수 없다.
 - 검정이 불가능함을 의미함

Case 1 vs. Case 2

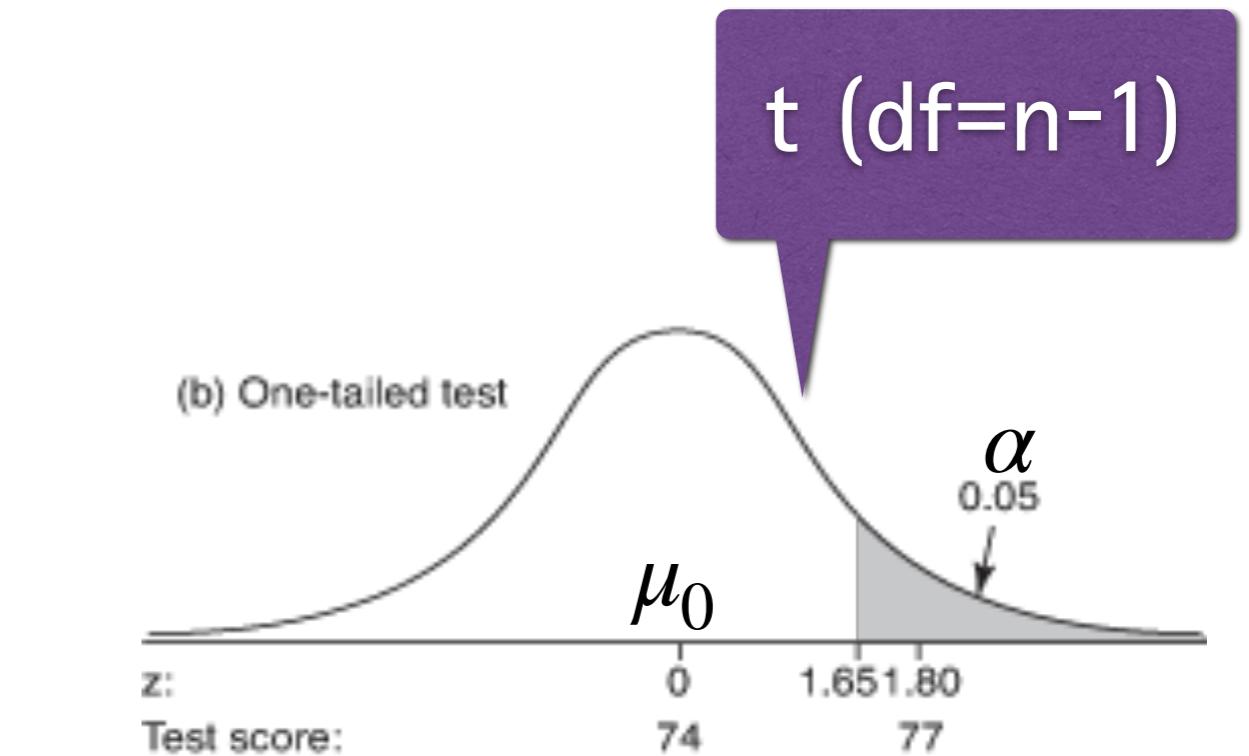
- 아래의 차이점을 제외하고는 Case 1과 동일한 과정
- 표본분포는 정규분포가 아닌, t 분포를 따른다
 - ⇒ 정규분포가 아닌 t분포로 p-value 구한다
- 모분산에 대한 정보가 없다
 - ⇒ 모분산에 대한 추정치로 표본분산을 사용한다

Case 2-a

p-value =

$$P(\bar{X} > \bar{x} | \mu = \mu_0) = P\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} > \frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right)$$

$$= P\left(T > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$



(a) 단측검정 (one-tailed)

$$H_0 : \mu = \mu_0 \quad vs. \quad H_a : \mu > \mu_0$$

$$p\text{-value} \begin{cases} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{cases}.$$

모비율에 대한 검정: 예

- 여론조사: 국민의 정책선호도를 알고자 1000명을 뽑아 정책선호도를 물어보았다
 - 알고자 하는 것: 정책선호도:= [정책선호하는 사람의 수]/[국민수]
 - 표본비율: 임의 배정된 1000명의 정책선호도
- 적중률이 60%를 넘는다는 투자자문회사의 주장이 타당한지 검토하기 위해 투자자문 자료 100개를 뽑아 분석하였다.
 - 알고자 하는 것: 투자자문회사의 적중률:= [예측에 성공한 자문수]/[총 자문수]
 - 표본비율: 임의 추출된 투자자문 100개 자료의 적중율

모비율 검정법

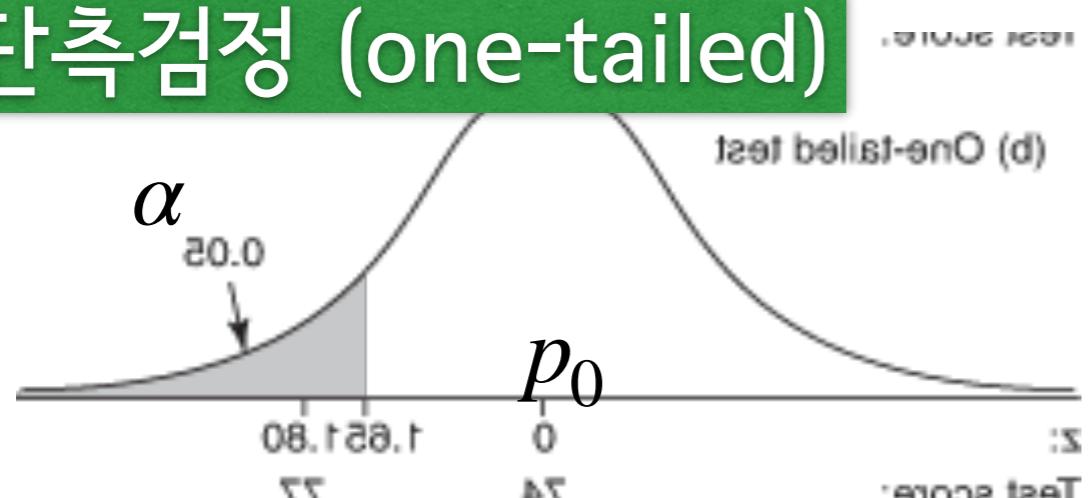
- 모비율을 확률로 하는 베르누이 시행으로 해석 가능
 - 베르누이 분포는 이산분포이지만 n 이 커질수록 정규분포에 수렴함
$$Y := n\hat{P} \sim B(n, p) \xrightarrow{d} N(np, np(1 - p))$$
$$\hat{P} \xrightarrow{d} N(p, p(1 - p)/n)$$
- p 에 대한 좋은 추정량인 표본비율 (\hat{p})을 사용
 - μ 대신 p 를 추정하는 것으로 근원적으로 Case 1과 동일한 문제

모비율 p 에 대한 검정: 모분산이 알려진 경우 (Case 1*)

$$p\text{-value} := P(\hat{p} \geq [\leq] p_0 | H_0)$$

- 기각역 (Rejection region)
 - H_0 전제하에서 현재의 샘플로 구성한 추정량이 나타날 확률이 유의수준 α 이하인 영역

(c) 단측검정 (one-tailed)

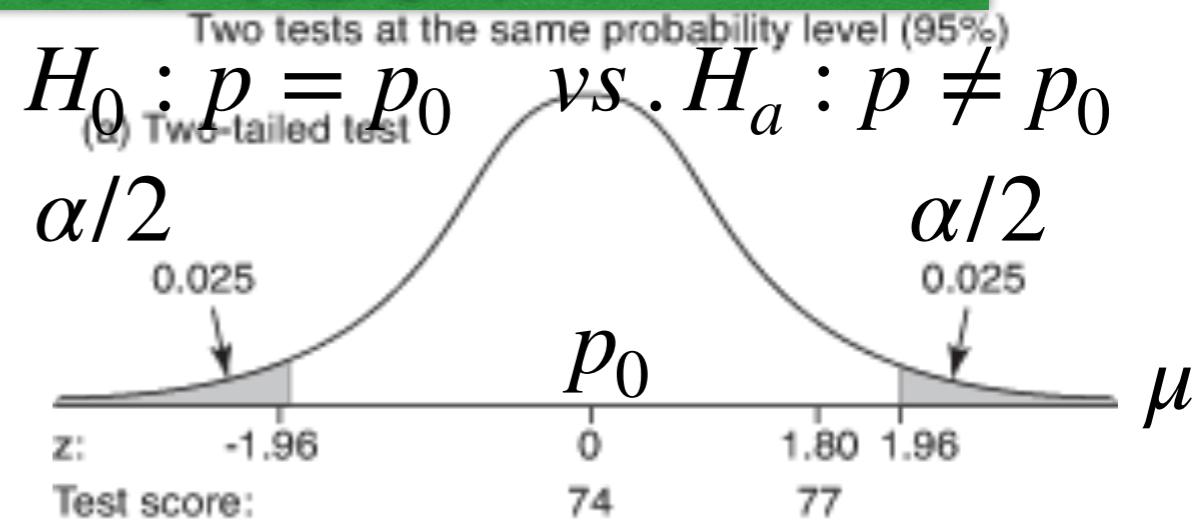


$$H_0 : p = p_0 \quad vs. \quad H_a : p < p_0$$

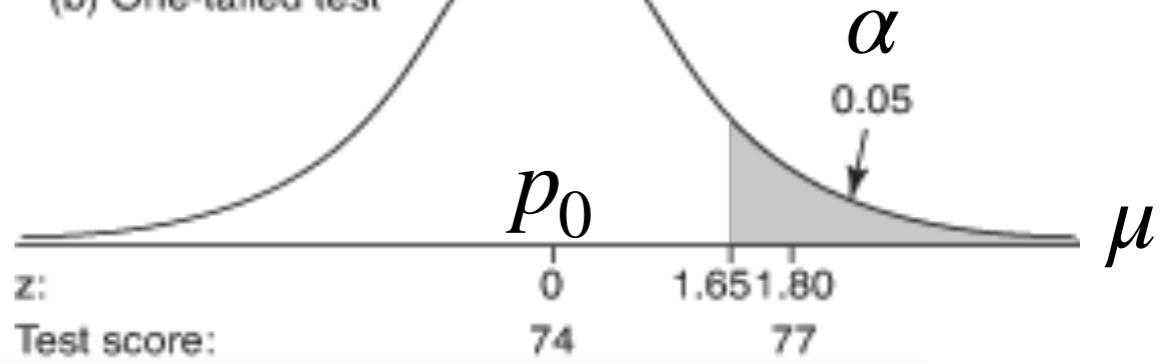
(a) 단측검정 (one-tailed)

$$H_0 : p = p_0 \quad vs. \quad H_a : p > p_0$$

(b) 양측검정 (two-tailed)



(b) One-tailed test



Ex 7.5

- 어떤 연구기관이 작년 주식 투자자 중 기껏해야 30% 만이 이익을 냈다고 주장

$$H_0 : p = 0.3 \quad vs. \quad H_a : p < 0.3$$

- 이를 95% 신뢰수준으로 검증하기 위해 임의추출된 100명의 투자자들 중 수익을 낸 투자자의 비율을 조사해봄: 25명이 이익을 냈음

$$\hat{p} = 0.25, \quad n = 100$$

Ex 7.5: 허오!

$$Y := n\hat{P} \sim B(n, p) \xrightarrow{d} N(np, np(1-p))$$
$$\hat{P} \xrightarrow{d} N(p, p(1-p)/n)$$

$$\text{p-value} = P(\bar{p} < \hat{p} | p = p_0) = P\left(\frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$
$$= P\left(Z < \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

$$p\text{-value} \begin{cases} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{cases}.$$

Ex 7.5: 풀이

$$\begin{aligned}\text{p-value} &= P\left(Z < \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}\right) \\ &= P\left(Z < \frac{0.25 - 0.3}{\sqrt{0.3(1 - 0.3)/100}}\right) = P(Z < -1.0911) \\ &= 1 - 0.8621 = 0.1379 > \alpha = 0.05\end{aligned}$$

$$p\text{-value} \begin{cases} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{cases}.$$

Ex 7.5: 풀이

$$\begin{aligned} \text{p-value} &= P\left(Z < \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}\right) \\ &= P\left(Z < \frac{0.25 - 0.3}{\sqrt{0.3(1 - 0.3)/100}}\right) = P(Z < -1.0911) \\ &= 1 - 0.8621 = 0.1379 > \alpha = 0.05 \end{aligned}$$

p-value $\begin{cases} < \alpha \Rightarrow \text{REJECT } H_0, \\ \geq \alpha \Rightarrow \text{DO NOT REJECT } H_0 \end{cases}$

신뢰구간과 양측검정, 표본의 크기

- 추정을 위해 사용한 방법은 크게 세가지
 - 임계값: 귀무가설하에서 유의수준에 맞는 통계량 찾기 (c_1, c_2)
 - p-value: 귀무가설 하에서 관측값이 관측될 확률 구하기
 - 신뢰구간: 귀무가설 하에서 유의수준을 충족하는 범위 구하기 CI: $\{x \mid c_1 \leq x \leq c_2\}$
- 이 방법들은 사실상 동일한 풀이의 다른 표현일 뿐임

두 모집단 관련 검정: $\mu_1 - \mu_2$ 에 대한 검정

서로 독립

$$X_1, \dots, X_n \sim iidN(\mu_1, \sigma_1^2) \quad Y_1, \dots, Y_m \sim iidN(\mu_2, \sigma_2^2)$$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

$$s_x^2 := \sum_i^n (X_i - \bar{X})^2 / (n - 1) \quad s_y^2 := \sum_i^m (Y_i - \bar{Y})^2 / (m - 1)$$
$$\frac{(n - 1)}{\sigma^2} s_x^2 \sim \chi^2(n - 1) \quad \frac{(m - 1)}{\sigma^2} s_y^2 \sim \chi^2(m - 1)$$

$\mu_1 - \mu_2$ 에 대한 검정:
분산은 모르지만 동분산인 경우
($n, m \leq 30$)

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

서로 독립

$$X_1, \dots, X_n \sim iid(\mu_1, \sigma^2) \quad Y_1, \dots, Y_m \sim iid(\mu_2, \sigma^2)$$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0, 1)$$

$$s_x^2 := \sum_i^n (X_i - \bar{X})^2 / (n - 1) \quad s_y^2 := \sum_i^m (Y_i - \bar{Y})^2 / (m - 1)$$

$$\frac{(n - 1)}{\sigma^2} s_x^2 \sim \chi^2(n - 1) \quad \frac{(m - 1)}{\sigma^2} s_y^2 \sim \chi^2(m - 1)$$

$$\frac{(n-1)}{\sigma^2} s_x^2 \sim \chi^2(n-1) \quad \frac{(m-1)}{\sigma^2} s_y^2 \sim \chi^2(m-1)$$

$$\frac{(n-1)}{\sigma^2} s_x^2 + \frac{(m-1)}{\sigma^2} s_y^2 \sim \chi^2(n-1+m-1)$$

C2 $X + Y \sim \chi^2(n_1 + n_2)$

$$s_p^2 := \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

Pooled Variance (합동 표본분산)

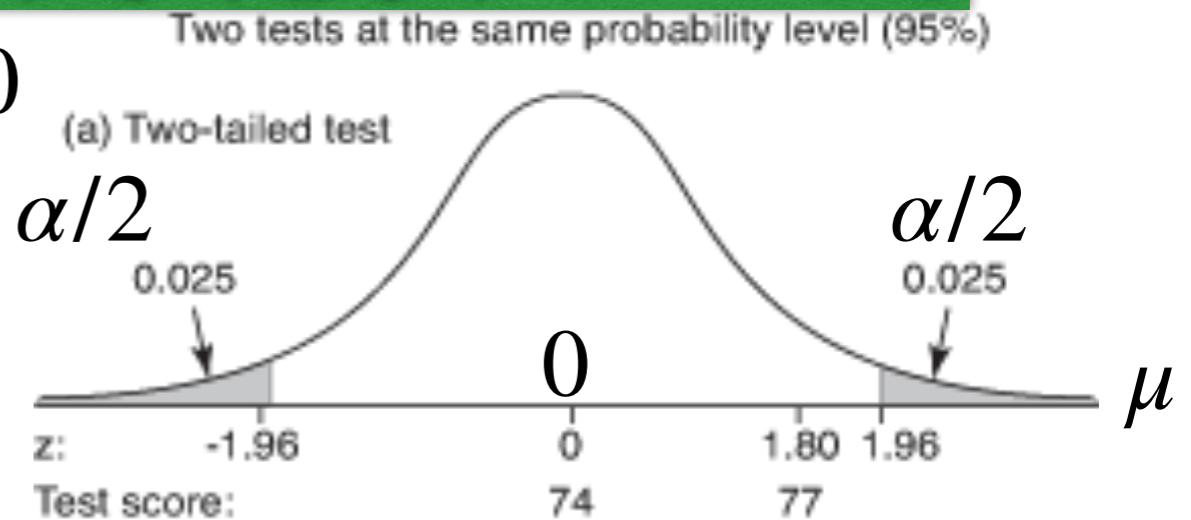
C3 $\frac{Z}{\sqrt{X/n_1}} \sim t(n_1)$

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2)$$

가설검정

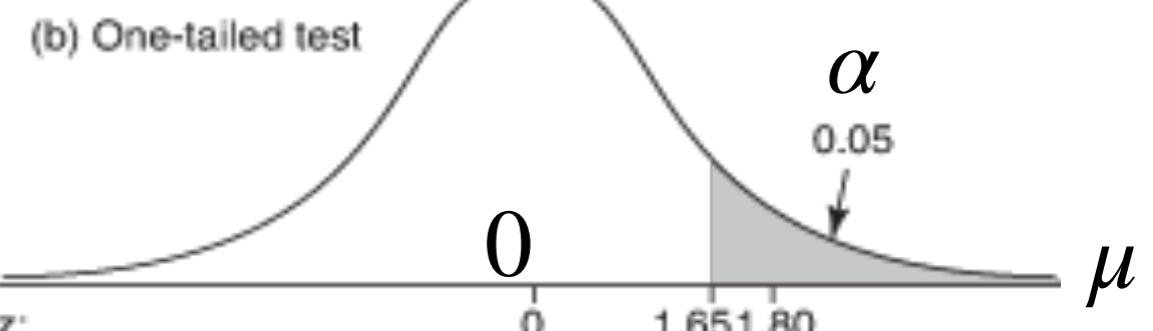
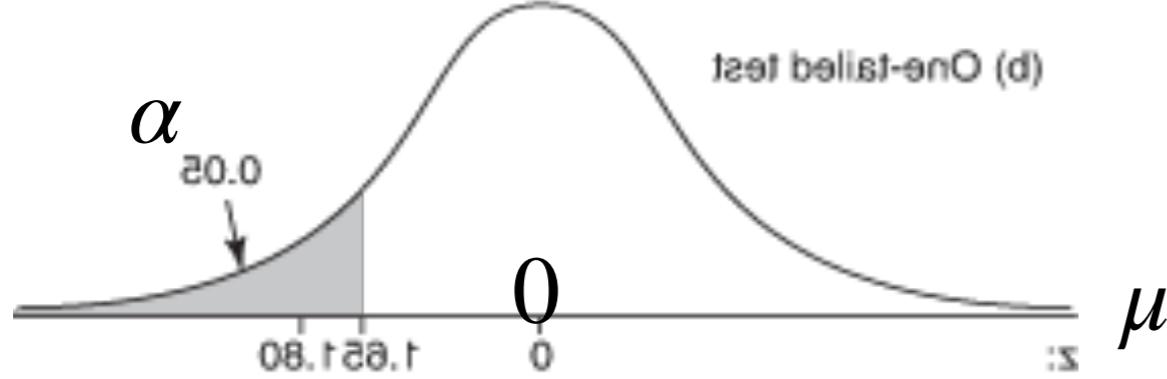
$$p\text{-value} := P(\hat{p} \geq [\leq] p_0 | H_0)$$

(b) 양측검정 (two-tailed)



분포: $t(df=n+m-2)$

(c) 단측검정 (one-tailed)



(a) 단측검정 (one-tailed)

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_a : \mu_1 - \mu_2 < 0$$

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_a : \mu_1 - \mu_2 > 0$$

(참고) 분산 모르고 이분산인 경우

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \sim t(df = [\nu])$$

$$\nu := \frac{(s_1^2/n + s_2^2/m)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}}$$

짝진 표본 검정

test for paired sample

- 하나의 모집단으로부터 2개의 변수가 관찰되는 경우
 - 예: 신발 내구도 테스트를 위해 실험자들에게 양 발에 각각 실험신발과 대조군신발(baseline, or reference)을 신김 \Rightarrow 마모율 측정
 - 이러한 경우 한 샘플(실험참가자)로부터 두 개의 측정값을 얻게 됨
- 이때의 관측값을 $d_i := X_i - Y_i$ 라고 정의
 - X_i, Y_i : i번째 참가자로부터 측정한 실험신발과 대조군신발의 마모율

짝진표본검정

$$H_0 : \mu_d \geq \mu_0 \quad vs. \quad \mu_d < \mu_0$$

$$\mu_d := \mu_1 - \mu_2$$

$$T = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}} \sim t(df = n - 1) \quad (\text{di 가 정규분포일 경우})$$

$$T = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}} \sim N(\mu_0, 1) \quad (\text{di 분포 모르지만 } n \text{이 클 때})$$

$$\bar{d} := \frac{1}{n} \sum_i^n d_i$$

$$s_d^2 := \frac{\sum_i^n (d_i - \bar{d})^2}{n - 1}$$

χ^2 분포를 이용한 검정법

Topics

- 모분산에 대한 검정
- 분할표를 이용한 검정
- 분산비에 대한 검정
- F-test

모분산에 대한 검정

$$H_0 : \sigma^2 \geq \sigma_0^2 \quad vs. \quad \sigma^2 < \sigma_0^2$$

- 모분산에 대한 좋은 추정량은 표본분산임
- χ^2 분포는 비대칭분포이므로 양측검정시 왼쪽, 오른쪽 통계량 모두 따로 구해야 함

$$s_x^2 := \sum_i^n (X_i - \bar{X})^2 / (n - 1)$$
$$\frac{(n - 1)}{\sigma^2} s_x^2 \sim \chi^2(n - 1)$$

분산비에 대한 검정

- 단측검정
- $H_0: \sigma_x^2 \leq \sigma_y^2$ vs.
 $H_a: \sigma_x^2 > \sigma_y^2$
- $H_0: \sigma_x^2 \geq \sigma_y^2$ vs.
 $H_a: \sigma_x^2 < \sigma_y^2$
- 양측검정
- $H_0: \sigma_x^2 = \sigma_y^2$ vs.
 $H_a: \sigma_x^2 \neq \sigma_y^2$

$$s_x^2 := \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$$

$$s_y^2 := \sum_{i=1}^m (Y_i - \bar{Y})^2 / (m - 1)$$

$$\frac{(n-1)}{\sigma_x^2} s_x^2 \sim \chi^2(n-1)$$

$$\frac{(m-1)}{\sigma_y^2} s_y^2 \sim \chi^2(m-1)$$

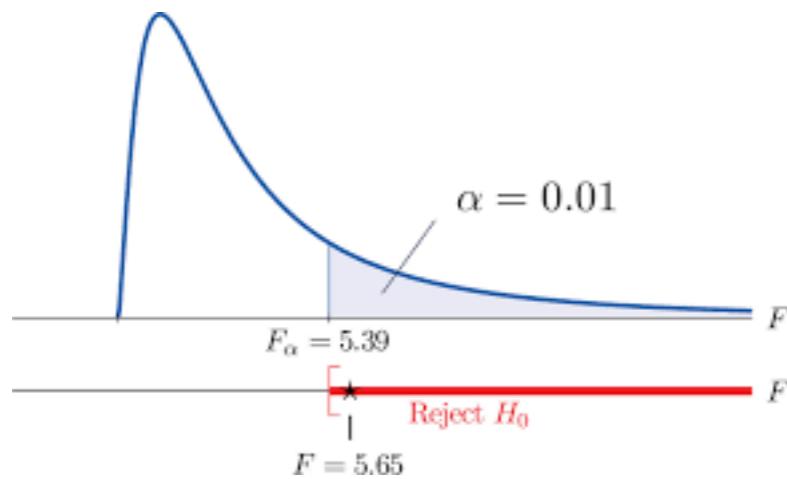
$$F = \frac{s_x^2 / \sigma_x^2}{s_y^2 / \sigma_y^2} \sim F(n-1, m-1)$$

$H_0: \sigma_x^2 \leq \sigma_y^2$ vs.

$H_a: \sigma_x^2 > \sigma_y^2$

$$P\left(\frac{s_x^2}{s_y^2} > c_1 \mid \sigma_x^2 = \sigma_y^2\right)$$

- 여기에서도 역시 귀무가설 $H_0: \sigma_x^2 \leq \sigma_y^2$ 은 $H_0: \sigma_x^2 = \sigma_y^2$ 과 같음



$$= P\left(\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} > c_1 \frac{\sigma_y^2}{\sigma_x^2} \mid \sigma_x^2 = \sigma_y^2\right)$$

$$= P\left(F > c_1 \frac{\sigma_y^2}{\sigma_x^2} \mid \sigma_x^2 = \sigma_y^2\right)$$

$$= P(F > c_1) = \alpha$$

OLS

Ordinary Least Squares (OLS)

OLS

Find $y = \mathbf{x}\beta + c$ for given N data $X = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$ satisfying:

$$\arg \min_{\beta, c} \sum_i^N ((\mathbf{x}_i \beta + c) - y_i)^2 \quad (\text{Least Square})$$

$$y = \mathbf{x}\beta + c = x_1\beta_1 + \cdots + x_m\beta_m + c$$

Note: given points – $(y_1, \mathbf{x}_1) = (y_1, x_{11}, x_{12}, \dots, x_{1m})$,
 $(y_2, \mathbf{x}_2), \dots, (y_N, \mathbf{x}_N)$ – are not variables. Our object is to find β^*, c^*
(linear equation) from given data X, Y .

OLS: Solution

Let our object function $f(\beta_1, \dots, \beta_n, c) := \sum_i^N ((\mathbf{x}_i \beta + c) - y_i)^2$. Then FOC is:

$$Df_{\beta,c}(\beta^*, c^*) = \mathbf{0} \quad (\text{FOC})$$

This leads to $m + 1$ equations:

$$\frac{\partial f}{\partial \beta_1}(\beta^*, c^*) = 2(\mathbf{x}_1 \beta^* + c^* - y_1)x_{11} + 2(\mathbf{x}_2 \beta^* + c^* - y_2)x_{21} + \dots + 2(\mathbf{x}_N \beta^* + c^* - y_N)x_{N1}$$

...

$$\frac{\partial f}{\partial \beta_m}(\beta^*, c^*) = 2(\mathbf{x}_1 \beta^* + c^* - y_1)x_{1m} + 2(\mathbf{x}_2 \beta^* + c^* - y_2)x_{2m} + \dots + 2(\mathbf{x}_N \beta^* + c^* - y_N)x_{Nm}$$

$$\Rightarrow 2(\mathbf{x}_1 \beta^* + c^* - y_1)\mathbf{x}_1^T + 2(\mathbf{x}_2 \beta^* + c^* - y_2)\mathbf{x}_2^T + \dots + 2(\mathbf{x}_N \beta^* + c^* - y_N)\mathbf{x}_N^T = \mathbf{0} \quad (\text{B})$$

$$\frac{\partial f}{\partial c}(\beta^*, c^*) = 2(\mathbf{x}_1 \beta^* + c^* - y_1)1 + 2(\mathbf{x}_2 \beta^* + c^* - y_2)1 + \dots + 2(\mathbf{x}_N \beta^* + c^* - y_N)1 = 0 \quad (\text{C})$$

OLS (2)

Remember $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nm} \end{pmatrix} = (C_1 \ \cdots \ C_m) = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$

Rearrange FOCs:

$$(\mathbf{x}_1^T \ \mathbf{x}_2^T \ \cdots \ \mathbf{x}_N^T) \begin{pmatrix} \mathbf{x}_1\beta^* + c^* - y_1 \\ \vdots \\ \mathbf{x}_N\beta^* + c^* - y_N \end{pmatrix} = X^T(X\beta^* + \mathbf{1}_{N \times 1}c^* - Y) = 0 \quad (\text{B2})$$

$$c^* = \frac{1}{N} (1 \ 1 \ \cdots \ 1) \begin{pmatrix} y_1 - \mathbf{x}_1\beta^* \\ \vdots \\ y_N - \mathbf{x}_N\beta^* \end{pmatrix} = \frac{1}{N} \mathbf{1}_{1 \times N} (Y - X\beta^*) \quad (\text{C2})$$

OLS (3)

From (C2) and (B2),

$$X^T \left(X\beta^* + \mathbf{1}_{N \times 1} \frac{1}{N} \mathbf{1}_{1 \times N} (Y - X\beta^*) - Y \right) = \mathbf{0}_{m \times 1} \quad (\text{D})$$

Rearrange (D) with regard to β^* yields:

$$X^T (X - \frac{1}{N} \mathbf{1}_{N \times 1} \mathbf{1}_{1 \times N} X) \beta^* = X^T (I_N - \frac{1}{N} \mathbf{1}_{N \times 1} \mathbf{1}_{1 \times N}) Y$$

Let $\mathbf{1}_{N \times 1} \mathbf{1}_{1 \times N} = \mathbf{1}_N$. ($N \times N$ matrix with all elements are 1)

$$\beta^* = \left(X^T \left(X - \frac{1}{N} \mathbf{1}_N X \right) \right)^{-1} \left(X^T \left(Y - \frac{1}{N} \mathbf{1}_N Y \right) \right)$$

$$= \left(X^T \left(I_N - \frac{1}{N} \mathbf{1}_N \right) X \right)^{-1} \left(X^T \left(I_N - \frac{1}{N} \mathbf{1}_N \right) Y \right)$$

OLS (4)

Sample Mean \bar{X}, \bar{Y}

$$\frac{1}{N} \mathbf{1}_N X = \frac{1}{N} \begin{pmatrix} 1 & \cdots & 1 \\ 1 & \cdots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots \\ x_{N1} & \cdots & x_{Nm} \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{x}}_1 & \cdots & \bar{\mathbf{x}}_m \\ \bar{\mathbf{x}}_1 & \cdots & \bar{\mathbf{x}}_m \\ \vdots & \ddots & \vdots \\ \bar{\mathbf{x}}_1 & \cdots & \bar{\mathbf{x}}_m \end{pmatrix} = \bar{X}$$

$$\frac{1}{N} \mathbf{1}_N Y = \frac{1}{N} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = \bar{Y}$$

Here $\bar{\mathbf{x}}_j, \bar{y}$ means sample mean of x_{ij}, y_i

$$\bar{\mathbf{x}}_j := \frac{1}{N} \sum_i^N x_{ij}, \quad \bar{y} := \frac{1}{N} \sum_i^N y_i$$

OLS (5)

Therefore, β^* is:

$$\beta^* = (X^T(X - \bar{X}))^{-1}X^T(Y - \bar{Y})$$

Note1: If $N \rightarrow \infty$, then $I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T \rightarrow I_N$ and

$$\beta^* \rightarrow (X^T X)^{-1} X^T Y$$

Note2: We should check SOC: whether $H = D^2 f_{\beta,c}(\beta^*, c^*)$ is PD or not.
Our object function has quadratic form with positive sign with regard to
 β, c when x_j is independent with each other and this means f is PD (when
 x_j is independent with each other: covariance with other variables are 0).
Note3: Some researchers denote X^T by X'

Next Topics

- 상관분석과 회귀분석

수고하셨습니다!



수고하셨습니다!

