Chapter. 10

둘 사이에는 무슨 관계가 있을까: 가설 검정

영 가설과 대립 가설, p-value

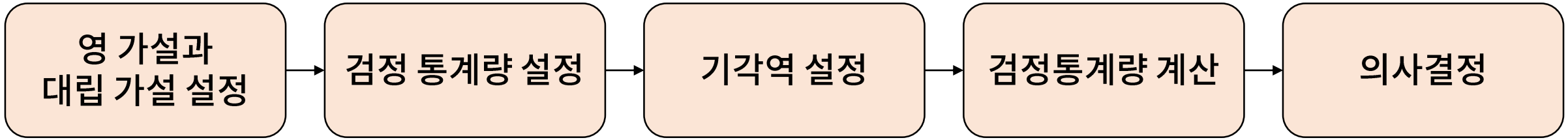
FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 통계적 가설 검정 개요

- 갖고 있는 샘플을 가지고 **모집단의 특성**에 대한 가설에 대한 통계적 유의성을 검정하는 일련의 과정
  - ➢ 수집한 데이터는 매우 특별한 경우를 제외하고는 전부 샘플이며, 모집단은 정확히 알 수 없는 경우가 더 많음
  - ➢ 통계적 유의성: 어떤 실험 결과 (데이터)가 확률적으로 봐서 단순한 우연이 아니라고 판단될 정도로 의미가 있음

- 통계적 가설 검정 과정은 다음과 같이 5단계로 구성됨



I 영 가설과 대립 가설

- 영 가설 (null hypothesis)와 대립 가설(alternative hypothesis)로 구분하여, 가설을 수립해야 함

	영 가설	대립 가설
정의	• 특별한 증거가 없으면 **참으로 추정**되는 가설 • 우리의 관심 대상이 **아닌** 가설 (검정을 통해, **영 가설을 기각**하고 싶어함)	• 특별한 증거가 없으면 **거짓으로 추정**되는 가설 • 우리의 **관심 대상**인 가설
표기	H_0	H_1 또는 H_a
설정 방법	모집단에 대한 특성을 **등호**로 표기	모집단에 대한 특성을 **부등호**로 표기 (단측 검정, 양측 검정)
예시	H_0: 대한민국 성인 남성의 키의 평균은 173cm이다. H_1: 대한민국 성인 남성의 키의 평균은 173cm과 같지 않다 (양측 검정) H_0: 성인 남성의 키는 성인 여성의 키와 같다. H_1: 성인 남성의 키는 성인 여성의 키보다 크다 (단측 검정)	

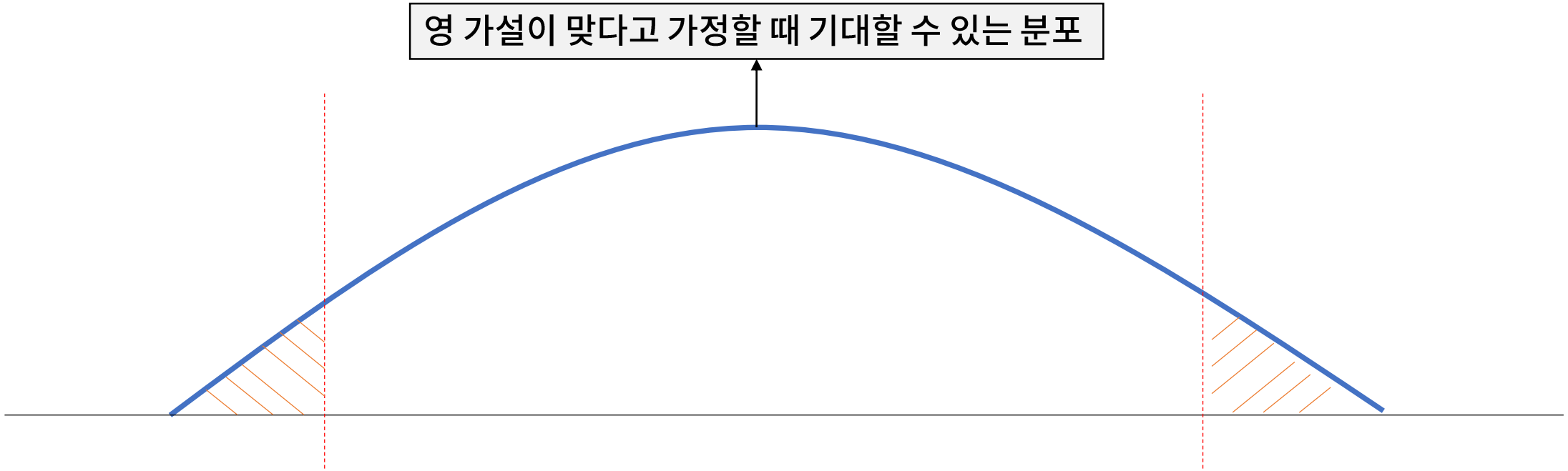
I 오류의 구분

- 가설 검정에서 발생하는 오류는 참을 거짓이라 하는 제 1종 오류 (type 1 Error)와 거짓을 참이라 하는 제 2종 오류로 구분됨

	영 가설 기각 X	영 가설 기각 O
영 가설 참	올바른 결정	제 1종 오류
영 가설 거짓	제 2종 오류	올바른 결정

I 유의 확률, p-value

- 유의 확률 p-value: 영 가설이 맞다고 가정할 때 얻은 결과와 다른 결과가 관측될 확률로, 그 값이 작을수록 **영 가설을 기각할 근거가 됨**

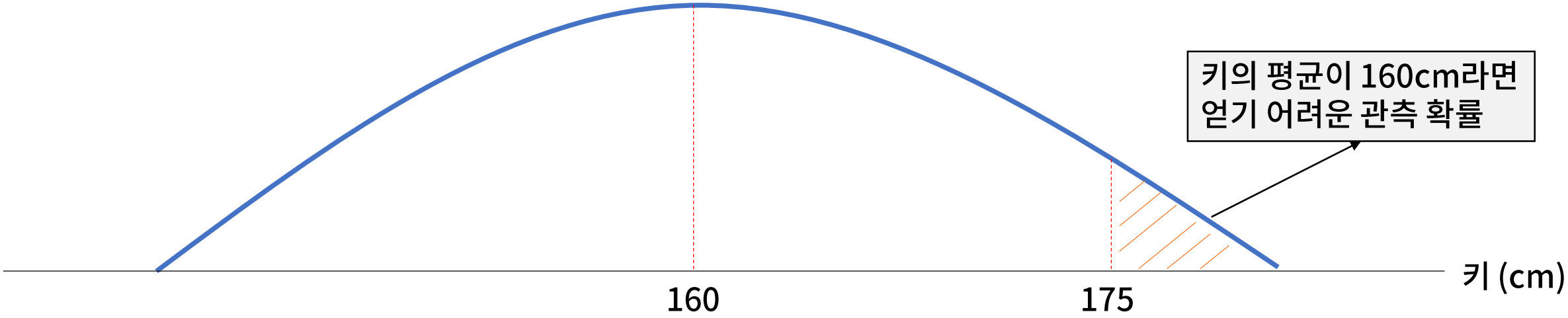


- 보통, **p-value가 0.05 혹은 0.01 미만이면 영 가설을 기각함**

l 유의 확률, p-value (예시 1)

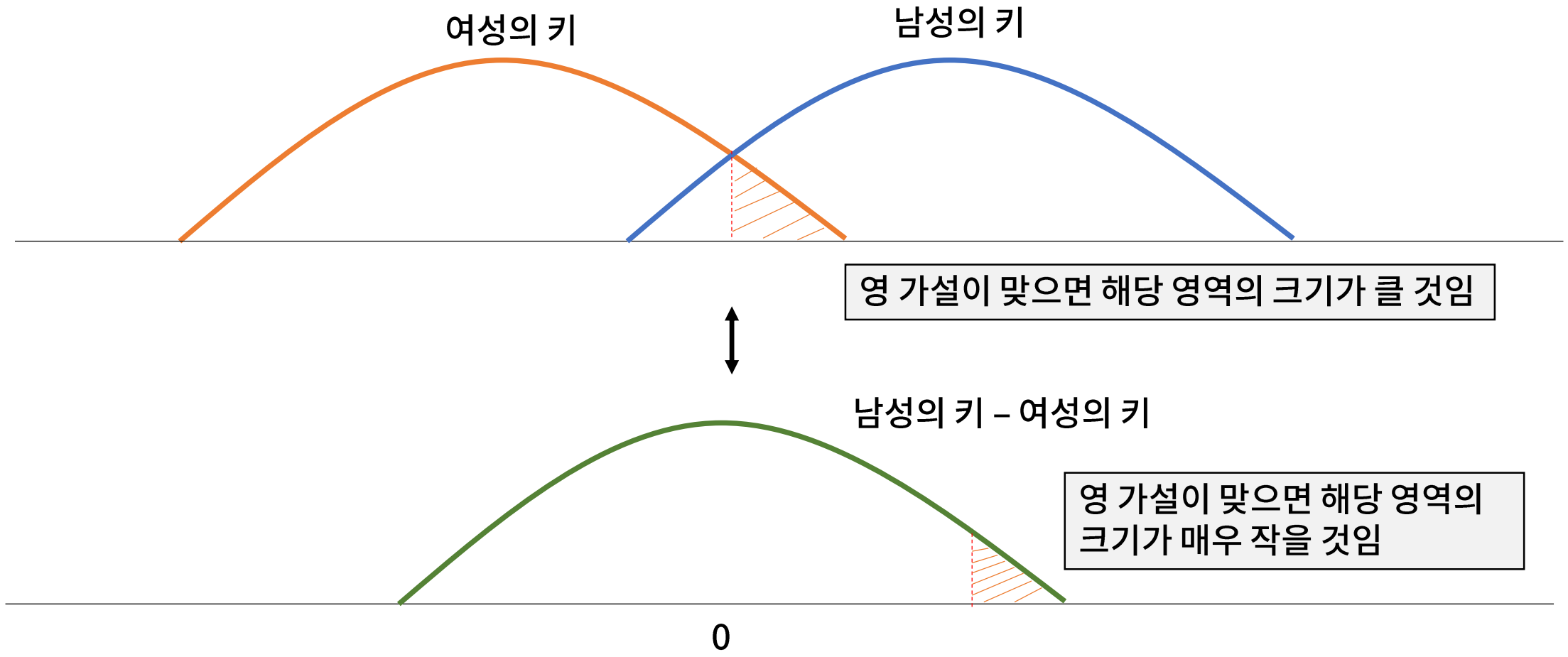
- 영 가설: 대한민국 성인 남성의 키는 160cm일 것이다.
- 대립 가설: 대한민국 성인 남성의 키는 160cm 이상일 것이다.

- 관측한 대한민국 성인 남성의 키의 평균은 175cm, 표준편차는 1cm이다.



I 유의 확률, p-value (예시 2)

- 영 가설: 대한민국 성인 남성의 키는 여성의 키와 같을 것이다.
- 대립 가설: 대한민국 성인 남성의 키는 여성의 키보다 작다.



Chapter. 10

둘 사이에는 무슨 관계가 있을까: 가설 검정

단일 표본 t 검정과 독립 표본 t 검정

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 단일 표본 t-검정 개요

- 목적: 그룹의 평균이 기준 값과 차이가 있는지를 확인

- 영 가설과 대립 가설

$H_0: \bar{x} = \mu$ ($\bar{x}$: 표본 평균, μ: 기준 값)

$H_1: \bar{x} > \mu$ or $\bar{x} < \mu$ or $\bar{x} \neq \mu$

- 가설 수립 예시: 당신이 한 웹 사이트를 운영하고 있는데, 고객이 웹사이트에서 체류하는 평균 시간이 10분인지 아닌지를 알고 싶어 다음과 같이 가설을 수립하였다.

$H_0: \bar{x} = 10$

$H_1: \bar{x} \neq 10$

I 단일 표본 t-검정의 선행 조건

- 단일 표본 t – 검정은 해당 변수가 정규 분포를 따라야 수행할 수 있으므로, Kolmogorov-Smornov나 Shapiro-Wilk를 사용한 정규성 검정이 선행되어야 함

- 그렇지만 보통 샘플 수가 많을수록 정규성을 띌 가능성이 높아지므로, 샘플 수가 부족한 경우에만 정규성 검정을 수행한 뒤, 정규성을 띄지 않는다 라고 판단된다면 비모수적 방법인 부호 검정 (sign test)나 윌콕슨 부호 – 순위 검정을 수행해야 함

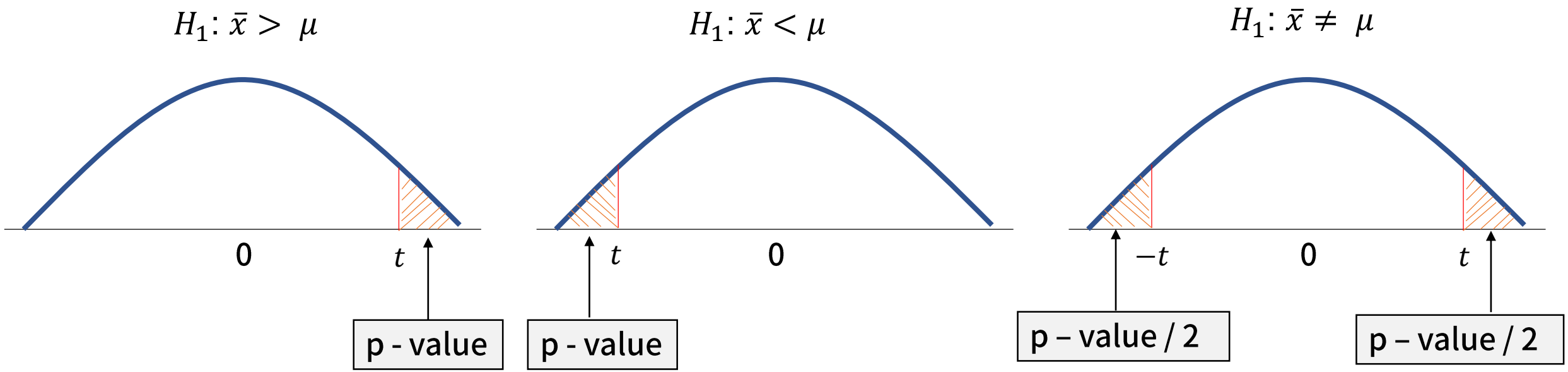
I 단일 표본 t-검정 통계량

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}},$$

✓ $\bar{x}$: 표본 평균 ✓ n: 표본 수

✓ μ: 기준 값 ✓ s: 표본 표준편차

- 위에 제시된 통계량을 t분포 상에 위치시키는 방식으로 p-value를 계산



I 정규성 검정: Kolmogorov-Smornov

- Kolmogorov-Smornov 검정 (이하 KS test)은 관측한 샘플들이 특정 분포를 따르는지 확인하기 위한 검정 방법임

- KS test는 특정 분포를 따른다면 나올 것이라 예상되는 값과 실제 값의 차이가 유의한지를 확인하는 방법으로, 해당 특정 분포를 정규 분포로 설정하여 정규성 검정에도 사용함

I 파이썬을 이용한 단일 표본 t – 검정

구분	코드	결과 해석
정규성 검정 (KS test)	scipy.stats.kstest(x, 'norm')	• result = (statistics, pvalue)의 튜플 형태 • pvalue가 특정 수치 미만이면 정규성을 따른다고 판단
단일 표본 t - 검정	scipy.stats.ttest_1samp (x, popmean)	• result = (statistics, pvalue)의 튜플 형태 • statistics가 양수면 x의 평균이 popmean보다 큰 것이며, 음수면 x의 평균이 popmean보다 작음을 의미 • pvalue가 특정 수치 미만이면 x는 popmean과 같지 않다고 판단
윌콕슨 부호 – 순위 검정	scipy.stats.wilcoxon(x)	• result = (statistics, pvalue)의 튜플 형태 • 단일 표본 t-검정과 결과 해석이 같음 (단, popmean은 x의 중위수로 설정됨)

I 독립 표본 t-검정 개요

- 목적: 서로 다른 두 그룹의 데이터 평균 비교

- 영 가설과 대립 가설

H_0: $\mu_a = \mu_b$ (μ_a: 그룹 a의 표본 평균, μ_b: 그룹 b의 표본 평균)

H_1: $\mu_a > \mu_b$ or $\mu_a < \mu_b$ or $\mu_a \neq \mu_b$

- 가설 수립 예시: 2020년 7월 한 달 간 지점 A의 일별 판매량과 지점 B의 일별 판매량이 아래와 같다면, 지점 A와 지점 B의 7월 **판매량 간 유의미한 차이**가 있는가?

일자	지점 A	지점 B
2020.07.01	160	170
2020.07.02	220	180
⋮	⋮	⋮
2020.07.31	190	150
평균	200	180

I 독립 표본 t-검정 선행 조건

- 독립성: 두 그룹은 서로 독립적이어야 함
- 정규성: 데이터는 정규분포를 따라야 함
  - ➢ 정규성을 따르지 않으면 비모수 검정인 Mann-Whitney 검정을 수행해야 함
- 등분산성: 두 그룹의 데이터에 대한 분산이 같아야 함
  - ➢ Levene의 등분산 검정: p-value가 0.05 미만이면 분산이 다르다고 판단
  - ➢ 분산이 같은지 다른지에 따라 사용하는 통계량이 달라지므로, 설정만 달리해주면 됨

I 독립 표본 t-검정 통계량

- 두 그룹의 분산이 같은 경우

$$t = \frac{\bar{x}_a - \bar{x}_b}{s\Big/\sqrt{\frac{1}{n_a}+\frac{1}{n_b}}},$$

✓ $\bar{x}_a$: 그룹 a의 표본 평균 ✓ n_a: 그룹 a의 샘플 수 ✓ s: 통합 분산

✓ $\bar{x}_b$: 그룹 b의 표본 평균 ✓ n_b: 그룹 b의 샘플 수

$$s = \sqrt{\frac{(n_a-1)\times s_a^2+(n_b-1)s_b^2}{n_a+n_b-2}},$$

✓ s_a: 그룹 a의 표준편차

✓ s_b: 그룹 b의 표준편차

- 두 그룹의 분산이 다른 경우

$$t = \frac{\bar{x}_a - \bar{x}_b}{s},$$

✓ $\bar{x}_a$: 그룹 a의 표본 평균

✓ $\bar{x}_b$: 그룹 b의 표본 평균

$$s = \sqrt{\frac{s_a^2}{n_a}+\frac{s_b^2}{n_b}},$$

✓ n_a: 그룹 a의 샘플 수 ✓ s_a: 그룹 a의 표준편차

✓ n_b: 그룹 b의 샘플 수 ✓ s_b: 그룹 b의 표준편차

I 파이썬을 이용한 독립 표본 t-검정

구분	코드	결과 해석
정규성 검정 (KS test)	scipy.stats.kstest(x, 'norm')	• pvalue가 특정 수치 미만이면 정규성을 따른다고 판단
등분산성 검정 (Levene test)	scipy.stats.levene(s1, s2, s3, …) # s1, s2, …: 샘플 (배열)	• pvalue가 특정 수치 미만이면 샘플 간 분산이 같지 않다고 판단
독립 표본 t - 검정	scipy.stats.ttest_ind (a, b, equal_var) # a, b: 두 그룹의 데이터 (배열) # equal_var: 등분산성을 만족하는지 여부	• statistics가 양수면 a의 평균이 더 크다고 판단 • pvalue가 특정 수치 미만이면 a와 b의 평균이 같지 않다고 판단
Mann – Whitneyu 검정	scipy.stats.mannwhitneyu(a, b) # a, b: 두 그룹의 데이터 (배열)	• result = (statistics, pvalue)의 튜플 형태 • pvalue가 특정 수치 미만이면 a와 b의 평균이 같지 않다고 판단

Chapter. 10

둘 사이에는 무슨 관계가 있을까: 가설 검정

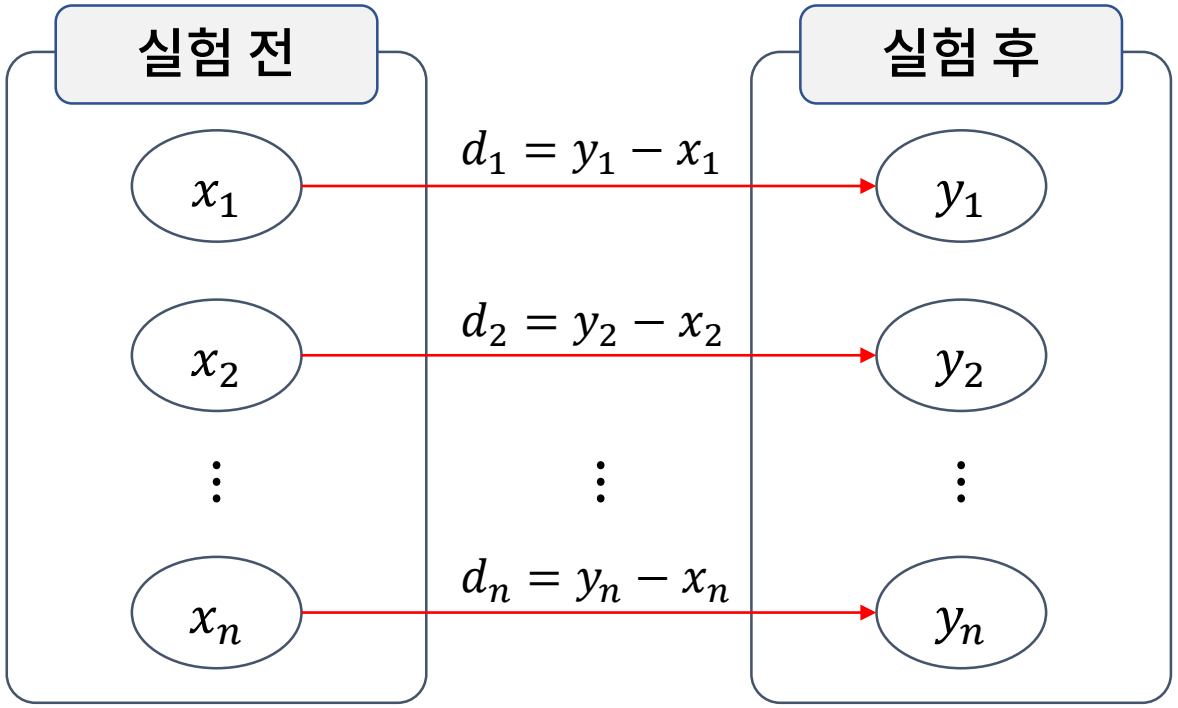
쌍체 표본 t검정

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 쌍체 표본 t-검정 개요

- 목적: 특정 실험 및 조치 등의 효과가 유의한지를 확인



$$H_0: d = 0$$

$$H_1: d > 0 \text{ or } d < 0 \text{ or } d \neq 0$$

I 쌍체 표본 t-검정의 선행 조건

- 실험 전과 후의 측정 값 (즉, X와 Y)은 정규 분포를 따르지 않아도 무방함
- 그러나 측정 값의 차이인 d는 정규성을 갖고 있어야 함

l 쌍체 표본 t-검정의 통계량

$$t = \frac{\bar{d}}{s_d/\sqrt{n}},$$

✓ $\bar{d}$: d의 평균

✓ s_d: d의 표준편차

I 파이썬을 이용한 쌍체 표본 t-검정

구분	코드	결과 해석
정규성 검정 (KS test)	scipy.stats.kstest(x, 'norm')	• pvalue가 특정 수치 미만이면 정규성을 따른다고 판단
쌍체 표본 t 검정	scipy.stats.ttest_rel(a, b) # a, b: 실험 전 후 결과 (주의: 반드시 길이가 같아야 함)	• pvalue가 특정 수치 미만이면 그룹 a와 그룹 b간 차이가 존재한다고 판단 (즉, 특정 실험의 효과가 존재) • statistics가 양수면 양의 효과 (d > 0)가 있다고 판단하며, 음수면 음의 효과 (d < 0)가 있다고 판단

Chapter. 10

둘 사이에는 무슨 관계가 있을까: 가설 검정

일원분산분석

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 일원분산분석 개요

- 목적: **셋 이상의 그룹 간 차이가 존재**하는지를 확인하기 위한 가설 검정 방법임

- 영 가설과 대립 가설

> H_0: $\mu_a = \mu_b = \mu_c$ (μ_a: 그룹 a의 표본 평균, μ_b: 그룹 b의 표본 평균, μ_c: 그룹 c의 표본 평균)
>
> H_1: 최소한 한 개 그룹에는 차이를 보인다

- 가설 수립 예시: 2020년 7월 한 달 간 지점 A, B, C의 일별 판매량이 아래와 같다면, 지점별 7월 **판매량 간 유의미한 차이**가 있는가? 또한, 어느 지점 간에는 유의미한 판매량 차이가 존재하지 않는가?

일자	지점 A	지점 B	지점 C
2020.07.01	160	170	180
2020.07.02	220	180	185
⋮	⋮	⋮	⋮
2020.07.31	190	150	140
평균	200	190	180

I 독립 표본 t검정을 사용하면 안 되는 이유

- 일원분산분석은 **독립 표본 t검정을 여러 번** 사용한 것과 같은 결과를 낼 것 처럼 보임

일원분산분석

$H_0: \mu_a = \mu_b = \mu_c$

H_1: 최소한 한 개 그룹에는 차이를 보인다

⟷

독립 표본 t 검정

$H_0: \mu_a = \mu_b$ $H_1: \mu_a \neq \mu_b$

$H_0: \mu_b = \mu_c$ $H_1: \mu_b \neq \mu_c$

$H_0: \mu_c = \mu_a$ $H_1: \mu_c \neq \mu_a$

- 독립 표본 t 검정에서 하나 이상의 영 가설이 기각되면, 자연스레 일원분산분석의 영가설 역시 기각되므로, 기각된 원인까지 알 수 있으므로 일원분산분석이 필요하지 않아 보일 수 있음

- 그러나 독립 표본 t 검정을 여러 번 했을 때, 아무리 높은 p-value가 나오더라도 그 신뢰성에 문제가 생길 수 있어, 일원분산분석이 필요함
  - 각 가설의 p-value가 0.95이고, 그룹의 개수가 k일 때 모든 영가설이 참일 확률: $(0.95)^k$
  - 그룹의 개수가 3개만 되어도 그 확률이 0.857로 크게 감소하며, 그룹의 개수가 14개가 되면 그 확률이 0.5 미만으로 떨어짐

일원분산분석의 선행 조건

- 독립성: 모든 그룹은 서로 독립적이어야 함
- 정규성: 모든 그룹의 데이터는 정규분포를 따라야 함
  - ➢ 그렇지 않으면 비모수적인 방법인 Kruskal-Wallis H Test를 수행해야 함
- 등분산성: 모든 그룹에 데이터에 대한 분산이 같아야 함
  - ➢ 그렇지 않으면 비모수적인 방법인 Kruskal-Wallis H Test를 수행해야 함

I 일원분산분석의 통계량

$$F = \frac{\text{집단 간 분산}}{\text{집단 내 분산}}$$

$$\text{집단 간 분산} = \frac{\sum_{g=1}^{G}\left(\left(\bar{x}_g - \bar{x}\right)^2 \times n_g\right)}{G-1}$$

- ✓ G: 그룹 개수
- ✓ n_g: 그룹 g에 속한 샘플 수
- ✓ $\bar{x}$: 모든 샘플의 평균
- ✓ $\bar{x}_g$: 그룹 g에 속한 샘플의 평균

$$\text{집단 내 분산} = \frac{\sum_{g=1}^{G}\left(s_g \times \left(n_g - 1\right)\right)}{n-G}$$

- ✓ n: 샘플 개수
- ✓ s_g: 그룹 g에 속한 샘플의 표준편차

I 사후분석: Tukey HSD test

- Tukey HSD (honestly significant difference) test는 일원분산분석에서 두 그룹 a와 b간 차이가 유의한 지 파악하는 사후 분석 방법임

$$\mathrm{HSD}_{a,b} = \frac{\max(\mu_a, \mu_b) - \min(\mu_a, \mu_b)}{\mathrm{SE}}$$

- ✓ μ_a: 그룹 a의 평균
- ✓ μ_b: 그룹 b의 평균
- ✓ SE: 그룹 a와 b의 표준 오차

- 만약, $\mathrm{HSD}_{a,b}$가 유의 수준보다 크면 두 차이가 유의하다고 간주

파이썬을 이용한 일원분산분석

구분	코드	결과 해석
정규성 검정 (KS test)	scipy.stats.kstest(x, 'norm')	• pvalue가 특정 수치 미만이면 정규성을 따른다고 판단
일원분산분석	scipy.stats.f_oneway(sample1, sample2, sample3, …)	• pvalue가 특정 수치 미만이면 최소 하나의 그룹은 다른 그룹의 평균과 다르다고 판단 (즉, 특정 실험의 효과가 존재)
사후분석	statsmodels.stats.multicomp. pairwise_tukeyhsd(Data, Group)	• 각 그룹 간 reject 결과 확인. Reject 컬럼이 True면 두 그룹 간 차이가 유의하다고 할 수 있음

Chapter. 10

둘 사이에는 무슨 관계가 있을까: 가설 검정

상관분석과 카이제곱검정

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 상관분석 개요

- 목적: 두 **연속형** 변수 간에 어떠한 **선형 관계**를 가지는지 파악
- 영 가설과 대립 가설

> H_0: 두 변수 간에는 유의미한 상관성이 존재하지 않는다
>
> H_1: 두 변수 간에는 유의미한 상관성이 존재한다

- 시각화 방법 : 산점도 (scatter plot)

I 피어슨 상관 계수

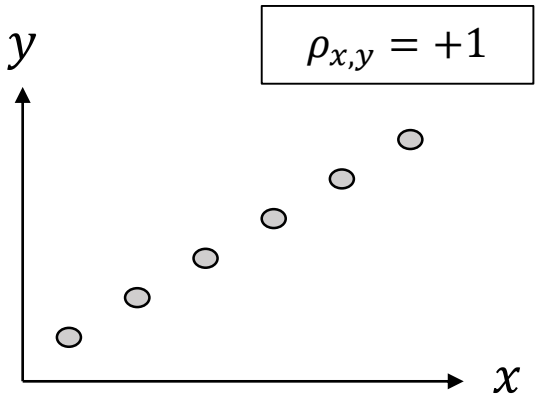
- 두 변수 모두 **연속형 변수**일 때 사용하는 상관 계수로 x와 y에 대한 상관 계수 $\rho_{x,y}$는 다음과 같이 정의됨

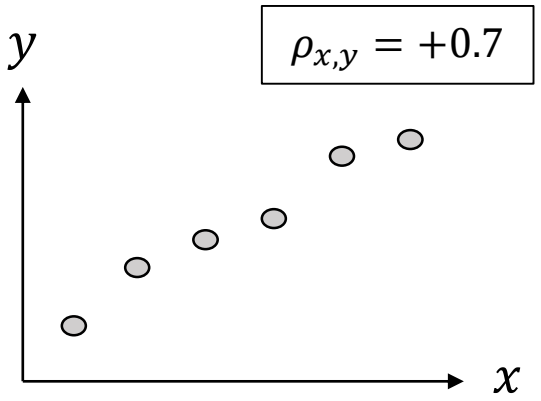
$$\rho_{x,y} = \frac{cov(x,y)}{\sqrt{var(x) \times var(y)}}$$

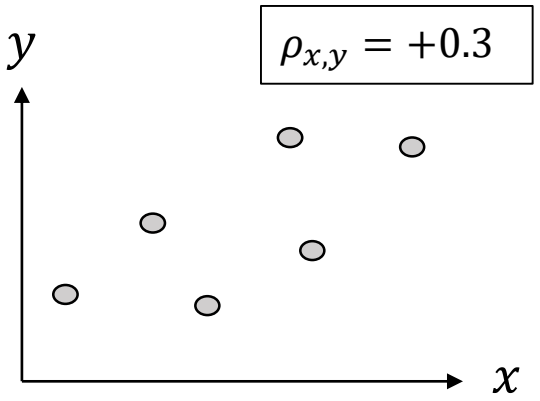
- ✓ $cov(x,y)$: x와 y의 공분산, $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-y)}{n-1}$
- ✓ $var(x)$: x의 분산

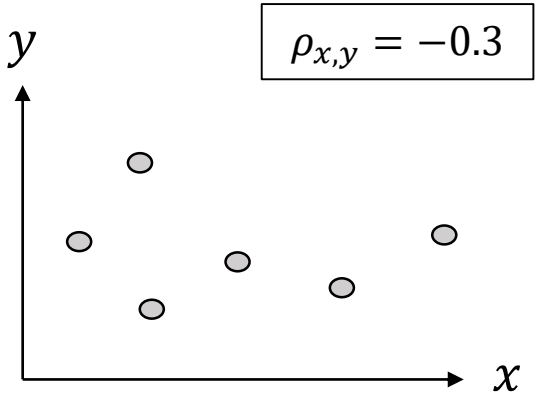
- 상관 계수가 1에 가까울수록 양의 상관관계가 강하다고 하며, -1에 가까울수록 음의 상관관계가 강하다고 함. 또한, 0에 가까울수록 상관관계가 약하다고 함

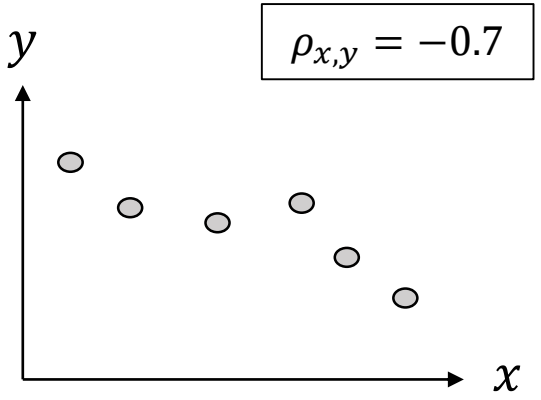
I 피어슨 상관 계수

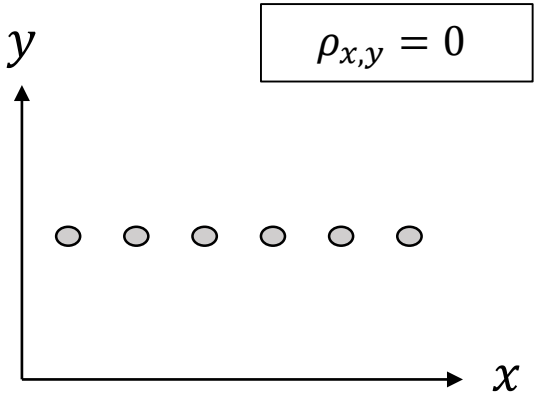












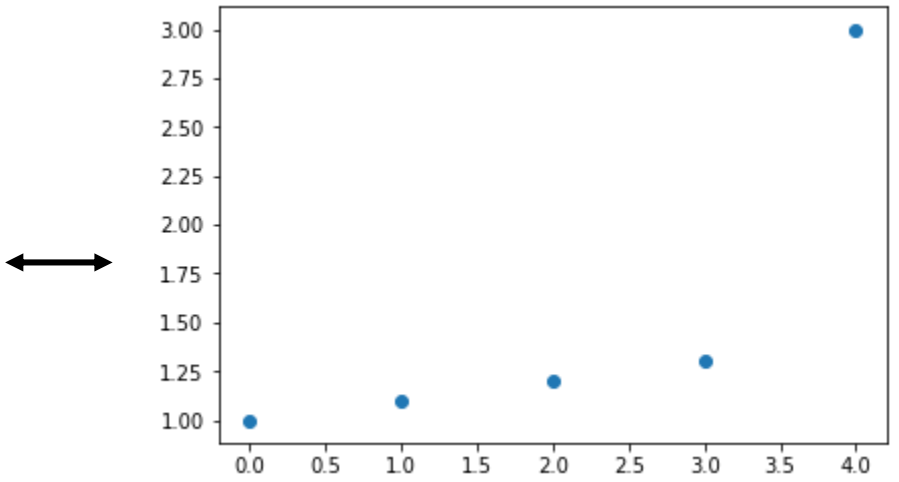
I 스피어만 상관 계수

- 두 변수의 순위 사이의 **단조 관련성**을 측정하는 상관 계수로 x와 y에 대한 스피어만 상관 계수 $S_{x,y}$는 다음과 같이 정의됨

$$S_{x,y} = \rho_{r(x),r(y)}$$

✓ $r(x)$: x의 요소의 개별 순위

x	y	$r(x)$	$r(y)$
0	1.0	1	1
1	1.1	2	2
2	1.2	3	3
3	1.3	4	4
4	3.0	5	5



- $\rho_{x,y} = 0.795$
- $S_{x,y} = 1.000$

파이썬을 이용한 상관분석

구분	코드	결과 해석
피어슨 상관계수 계산	scipy.stats.pearsonr(x, y)	• result = (statistics, pvalue) • statistics: 피어슨 상관계수 • pvalue: 0.05미만이면 유의한 상관성이 있다고 봄
스피어만 상관계수 계산	scipy.stats.spearmanr(x, y)	• result = (statistics, pvalue) • statistics: 스피어만 상관계수 • pvalue: 0.05미만이면 유의한 상관성이 있다고 봄
상관 행렬	DataFrame.corr(method) # method: pearson, spearman	• 컬럼 간 상관계수를 나타내는 행렬

I 카이제곱 검정 개요

- 목적: 두 범주형 변수가 서로 독립적인지 검정
- 영 가설과 대립 가설

H_0: 두 변수가 서로 독립이다

H_1: 두 변수가 서로 종속된다

- 시각화 방법: 교차 테이블

I 교차 테이블과 기대값

- 교차 테이블(contingency table)은 두 변수가 취할 수 있는 값의 조합의 출현 빈도를 나타냄
- 예시: 성별에 따른 강의 만족도 (카테고리 수 = $2 \times 3 = 6$)

	만족	보통	불만족	합계
남성	50 (45)	40 (35)	10 (20)	100
여성	40 (45)	30 (35)	30 (20)	100
합계	90	70	40	200

⟷

- 여성이면서 강의에 보통이라고 응답한 사람이 30명
- 남성이면서 강의에 불만족을 느낀 사람 수에 대한 기대값이 20명

- 카테고리 $C_{i,j}$에 대한 기대값 = $\frac{N_i \times N_j}{N}$ (N: 전체 샘플 수, N_i: 값 i를 갖는 샘플 수, N_j: 값 j를 갖는 샘플 수)
- 예시) 성별 = 남성, 강의 = 만족에 대한 기대 값: $\frac{100 \times 90}{200} = 45$

I 카이제곱 통계량

- 카이제곱 검정에 사용하는 카이제곱 통계량은 **기대값과 실제값의 차이**를 바탕으로 정의됨

$$\chi^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}$$

- ✓ c: 카테고리 개수 (두 변수의 상태 공간의 곱)
- ✓ O_j: 카테고리 j의 실제 값 (관측 값)
- ✓ E_j: 카테고리 j의 기대 값

- 기대값과 실제값의 차이가 클수록 통계량이 커지며, 통계량이 커질수록 영 가설이 기각될 가능성이 높아짐 (즉, p-value가 감소함)

파이썬을 이용한 카이제곱 검정

구분	코드	결과 해석
교차 테이블 생성	pandas.crosstab (S1, S2)	• Series S1과 S2로 구성된 교차 테이블을 생성
카이제곱 검정	scipy.stats.chi2_contingency(obs) # obs: 실제값	• 교차 테이블의 실제값에 대한 기대값 계산 • 보통 pandas.crosstable의 결과에 대한 values를 입력으로 투입 • result = (chi2, pvalue, dof, expected)

Chapter. 10

둘 사이에는 무슨 관계가 있을까: 가설 검정

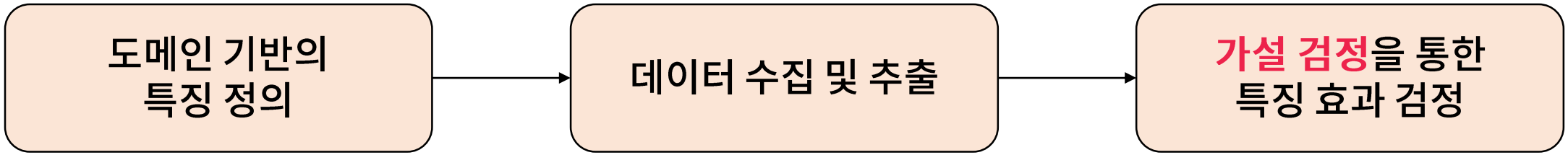
머신러닝에서의 가설검정

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

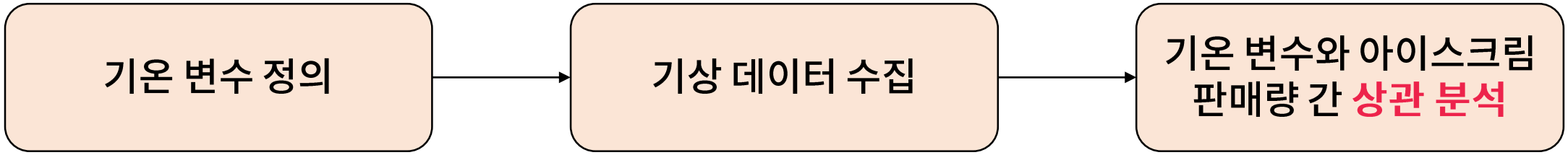
강사. 안길승

특징 정의 및 추출

- 예측 및 분류에 효과적인 특징을 정의하고 추출하는 과정은 다음과 같음



- (예시) 아이스크림 판매량 예측



특징 선택

- 특징 선택이란 **예측 및 분류에 효과적인 특징**을 선택하여 차원을 축소하는 기법임
- 특징의 효과성(클래스 관련성)을 측정하기 위해 가설 검정에서 사용하는 **통계량**을 사용함

특징
x_1
x_2
x_3
x_4
x_5
x_6
x_7

클래스 관련성 측정 함수 $R(\cdot)$

클래스 관련성
$R(x_1) = 1.2$
$R(x_2) = 1.4$
$R(x_3) = 0.6$
$R(x_4) = 0.9$
$R(x_5) = 2.3$
$R(x_6) = 4.1$
$R(x_7) = 3.3$

4개의 특징 선택

선택된 특징
x_6
x_7
x_5
x_2

| 특징 선택

- 클래스 관련성 척도는 특징과 라벨의 유형에 따라 선택함

통계량	특징 유형	라벨 유형
카이제곱 통계량	이진형	이진형 (분류)
상호 정보량	이진형	이진형 (분류)
	연속형	
	이진형	연속형 (예측)
	연속형	
F – 통계량	연속형	이진형 (분류)
	연속형	연속형 (예측)

Chapter.

둘 사이에는 무슨 관계가 있을까: 가설 검정

감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승