청년을 응원합니다!
K-Digital Training

기초통계 및 경영통계

강사 김경하

고용노동부 | 삼정 KPMG | 현대경제연구원 | i-Growth

기초 통계 with 데이터분석, 머신러닝

학습 내용

- EDA와 통계량 이해
- 추론 통계학 이해
- 통계적 가설 검정 이해
- 다양한 가설검정 방법 살펴보기

학습 내용

- EDA 주요 활동에 대해 안다.
- 가설 검정 방법과 절차에 대해 알아본다.
- 영 가설과 대립 가설의 의미를 안다.
- 유의확률(p-value)의 의미를 안다.
- 가설검정에서 발생가능한 오류를 이해한다.
- 다양한 가설검정 방법을 이해하고, 실습한다.

EDA와 통계량 이해

탐색적 데이터 분석(EDA)

- EDA는 그래프를 통한 시각화와 통계 분석등을 바탕으로 수집한 데이터를 다양한 각도에서 관찰하고 이해하는 방법

EDA의 필요성과 효과

- 데이터를 여러 각도에서 살펴보면서 데이터의 전체적인 양상과 보이지 않던 현상을 더 잘 이해할 수 있도록 도움
- 문제 정의 단계에서 발견하지 못한 패턴을 발견하고, 이를 바탕으로 데이터 전처리 및 모델에 관한 가설을 추가하거나 수정할 수 있음
- 즉 본격적인 데이터 분석에 앞서 구체적인 분석 계획을 수립하는데 도움이 됨

EDA 주요 활동

구분	활동 내용
분석 목적 및 변수 확인	- 대략적인 문제 정의 - 변수별 의미를 코드북 및 도메인 지식을 통해 확인
데이터 전체적으로 살펴보기	- 데이터 자체에 문제가 없는지를 확인하기 - 데이터의 일부를 샘플링하여 하나하나 살펴보기 - 이상치 및 결측치 탐색 - 군집화 및 빈발 패턴 추출등을 통한 주요 패턴 파악
개별 속성값 확인	- 개별 변수에 대한 빈도분석 및 기술 통계 - 그래프시각화(히스토그램, 파이차트, 박스플롯) - 분포 적합성 검정
속성 간 관계 파악	- 카이제곱 검정 - 상관관계 분석 - T-검정 및 일원분산분석 - 산점도 및 히트맵 시각화

확률 변수의 정의

- 변수(variable) : 특정 조건에 따라 변하는 값
- 확률 변수(random variable) : 특정 값(범위)을 확률에 따라 취하는 변수 → 예시 : 주사위를 던졌을 때 나오는 결과를 나타내는 변수 X

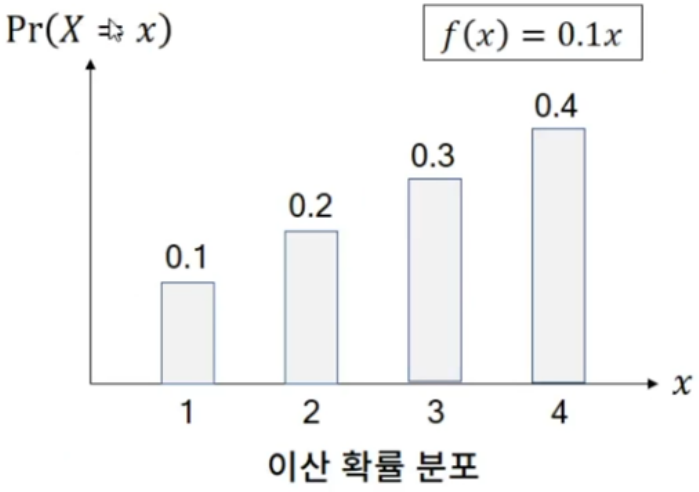
값	1	2	3	4	5	6
확률	1/6	1/6	1/6	1/6	1/6	1/6

→ **상태 공간**

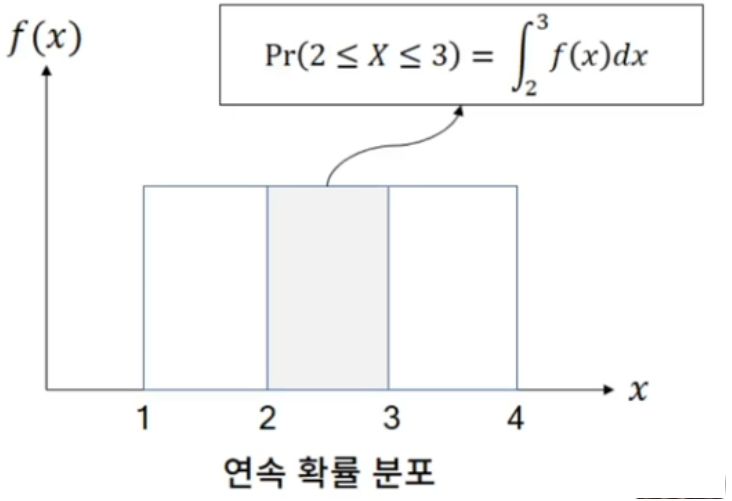
- **상태 공간**의 크기에 따라
  - **무한**한 변수: **연속 확률** 변수
  - **유한**한 변수: **이산 확률** 변수

확률 분포의 정의

- 확률 분포(probability distribution)
  - 확률 변수가 특정한 값을 취할 확률을 나타내는 **함수**를 의미함



확률 질량 함수



확률 밀도 함수

확률 분포의 확인

- 한 변수가 따르는 확률 분포를 확인했을 때의 효과
  - 현재 수집한 데이터가 어떻게 생겼는지를 이해할 수 있음
  - 새로 데이터가 들어오면 어떻게 들어올 것인지 예상할 수 있음
- 그러나 가지고 있는 데이터는 **샘플 데이터**이므로 **절대로 정확히 한 변수가 따르는 확률 분포를 알 수 없음**.
- **그래프**를 이용하여 확인하거나 **적합성 검정을 사용**하여 **확률 분포를 확인**해야 하는데, **이 작업은 굉장히 많은 노력과 시간이 필요함**.

통계량의 필요성 : 간단하게 확률 분포 확인

- 통계량은 확률 분포의 특성을 나타내는 지표를 의미함
- 통계량을 계산하는 기초 통계분석(기술 통계 분석)을 바탕으로 확률 분포를 간단하게 확인 가능함
  (단, 반드시 각 통계량이 나타내는 의미를 이해해야 함.)
- 특히, 변수가 많은 경우에 훨씬 효율적으로 사용 가능함

통계량의 종류

- 통계량은 크게 대표 통계량, 산포 통계량, 분포 통계량으로 구분

구분	내용	예시
대표 통계량	데이터의 중심 및 집중경향을 나타내는 통계량	평균, 중앙값, 최빈값 등
산포 통계량	데이터의 퍼진 정도를 나타내는 통계량	분산, 범위, 표준편차
분포 통계량	데이터의 위치 정보 및 모양을 나타내는 통계량	외도, 첨도, 사분위수, 최대값, 최소값

추론 통계학 이해

추론 통계학

- 모집단의 성격을 모르는 상황에서 모집단의 성격을 규명해야할 때
  - 어느 회사 제품의 평균 용량을 알기를 원하다고 하면,
    - 제품 전체를 분석한다면 시간적 제약과 경비등의 문제로 불가능함
    - 표본의 특성을 나타내는 통계량을 계산하여 모집단의 모수를 찾아냄
- 가능한 모수를 정확하게 추정하는 것이 통계적 분석에서 가장 중요함
  - 이 분야를 다루는 통계학을 **추론 통계학**이라고 함
  - 추론 통계학이 다루는 영역 : 통계적 추정, 가설검정

* 모집단 : 우리가 관심을 가지고 연구하고 싶어 하는 전체 대상
* 모수 : 모집단의 특징을 나타내는 수치적인 값(평균, 표준편차, 비율 등의 지표)

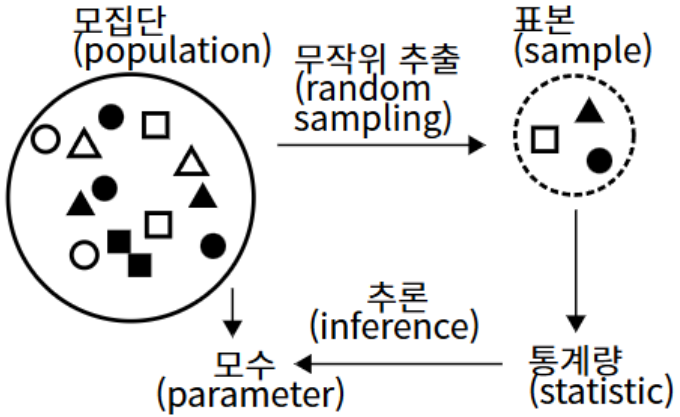
추론 통계학

- 통계적 추정(statistical **estimation**)
  - 통계량을 기초로하여 **모수를 추정**하는 **통계적 분석** 방법
  - 표본으로부터 계산한 통계량을 사용하여 모수의 특성을 규명하는 것

- 가설검정(**hypothesis** test)
  - 모수에 대하여 **특정한 가설**을 세워 놓고, **모집단으로 부터 추출된 표본을 분석함**으로써 그 가설의 **타당성 여부를 결정**하는 것

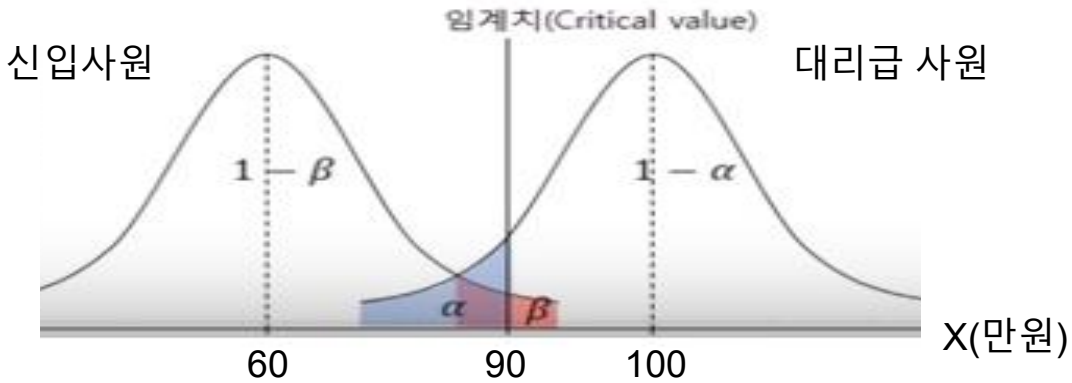
추론 통계학 – 통계적 추정 예시

- 서울 시내의 모든 대학의 신입생을 대상으로 수학능력시험 점수를 알아보려고 하는데, 시간적인 제약과 경비 때문에 서울 시내의 모든 대학의 신입생을 대상으로 수학능력시험을 조사할 수는 없다.
- 이런 경우 각 대학의 신입생 중에서 적정한 수의 학생을 표본으로 뽑아 그들의 수학능력 시험 점수를 조사하여 본 결과 평균 점수가 250점 이었다고 하고, 다음과 같이 답하면,
  - 250점일 것이다.
  - 225점~275점 사이일 것이다.
  - 170~330점 사이일 것이다.

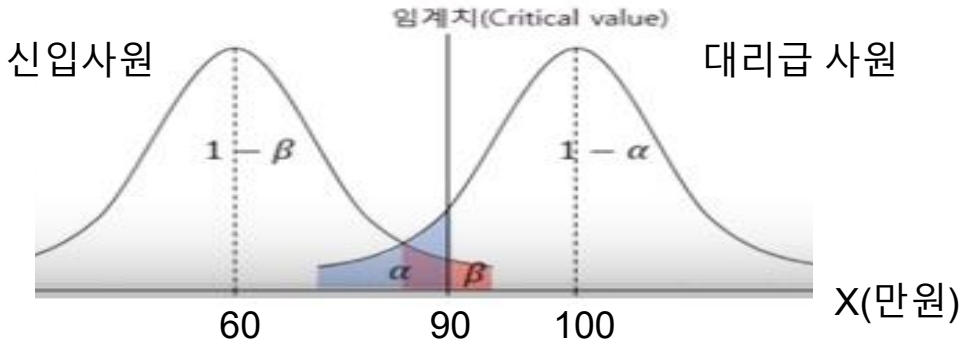


추론 통계학 – 가설검정 예시

- 어느 기업의 신입사원들과 입사 후 연차가 어느정도 쌓인 대리급 사원들의 같은 해 명절 상여금을 비교해 본 결과 명절 상여금의 분포가 아래 그림과 같다.
- 대리급 사원들의 평균 상여금은 100만원이었고, 신입 사원들의 평균 상여금은 60만원, 두 분포는 정규분포를 이룬다고 한다.
- 인사팀장이 한 사원을 선택해 직급을 확인하지 않고 그 사원의 상여금을 조사하니, 이번 명절 상여금으로 80만원을 받았다고 한다면,
  그 사원은 신입사원일까? 대리일까?



추론 통계학 – 가설검정 예시



- **귀무가설** : 그 사원은 대리 직급이다. **대립가설** : 그 사원은 신입사원이다.
- 몇 만원 이상을 대리직급으로 보야야 할지 **기준 필요** -> 90만원 이상 대리
- 80만원 받은 직원에 대한 의사결정, **귀무가설은 기각**됨
- 결과적으로 그 직원은 신입사원이라는 **대립가설이 채택**됨
- α와 β 구간의 의사결정 **오류 존재할 가능성**

가설 수립과 검정하기

- 탐색적 데이터 분석은 **가설(질문) 수립과 검정의 반복**으로 구성됨

가설 수립과 검정하기(예시)

- 문제 상황
  - 한 보험사에서 고객의 **이탈이 크게 늘고 있음**
  - 이탈한 고객들을 미리 식별하는 모델을 만들고 **이탈할 것이라 예상**되는 고객을 관리하여 **이탈율을 줄이고자** 함
- 가장 먼저 선행되어야 하는 작업은?

가설 수립과 검정하기(예시)

- 문제 상황
  - 한 보험사에서 고객의 **이탈이 크게 늘고 있음**
  - 이탈한 고객들을 미리 식별하는 모델을 만들고 **이탈할 것이라 예상**되는 고객을 관리하여 **이탈율을 줄이고자** 함
- 가장 먼저 선행되어야 하는 작업은?
  - 문제 정의 및 목표 설정
  - **고객의 이탈 원인을 파악**하는 것이고, 파악된 원인을 모델의 **특징**으로 사용해야 함.

가설 예시	검정 방법
보험 가입 기간이 긴 고객일수록 이탈율이 줄어들 것이다.	보험 가입 기간과 고객 이탈 여부 간의 관계 분석 독립표본 t검정(보험 가입 기간이 다른 두 그룹 이탈율 비교)
고객 여정과 이탈율은 관계가 있을 것이다.	주요 고객 여정 추출 및 주요 고객 여정 여부와 이탈율 간 카이 제곱 검정 및 히트맵 시각화

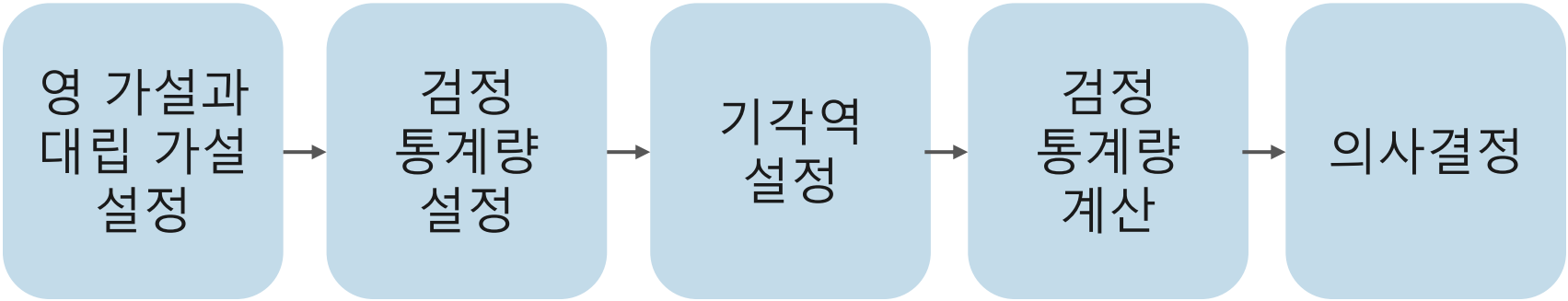
통계적 가설 검정 이해
(Statistical Hypothesis Test)

통계적 가설 검정 개요

- 갖고 있는 **샘플**을 가지고 **모집단의 특성**에 대한 가설의 **통계적 유의성**을 **검정**하는 일련의 과정
  - 수집한 데이터의 매우 특별한 경우를 제외하고, 대부분 샘플이며, 모집단은 정확히 알 수 없는 경우가 더 많음.
  - **통계적 유의성**
    - 어떤 실험결과(데이터)가 확률적으로 봐서 단순한 **우연이 아니라고 판단될 정도로 의미가 있는 차이라는 것을 검증하는 것**
    - 의사결정의 근거 제공 → 단순한 숫자 차이가 아니라 신뢰할 수 있는 차이인지 확인
    - 비즈니스 적용 가능성 → 데이터 기반 의사결정시 과학적 근거 마련

통계적 가설 검정 개요

- 통계적 가설 검정 5단계



가설 수립

가설 검정

가설 성립한다
성립하지 않는다.
기각역 설정

설정한 검정
통계량을
계산함.

가설의 유의성을
판단하여 의사결정

통계적 가설 검정 개요

- 영 가설(null hypothesis)과 대립 가설(alternative hypothesis)로 구분하여, 가설을 수립하여야 함

	영 가설 또는 기무가설 (null hypothesis)	대립 가설 (alternative hypothesis)
정의	- 특별한 증거가 없으면 참으로 추정되는 가설 • 우리의 **관심 대상이 아닌** 가설 • 검정을 통해, 기각하고 싶어함	- 특별한 증거가 없으면 거짓으로 추정되는 가설 • 우리의 **관심 대상**인 가설 • 검정을 통해, 채택하고 싶어함
표기	H_0	H_1 또는 H_a
설정 방법	모집단에 대한 특성을 등호로 표기 =	모집단에 대한 특성을 부등호로 표기 (단측 검정 ,양측 검정) >,<, <=, >=, !=

통계적 가설 검정 개요

- 영 가설(null hypothesis)과 대립 가설(alternative hypothesis)로 구분하여, 가설을 수립하여야 함

구분	영 가설 또는 귀무가설 (null hypothesis)	대립 가설 (alternative hypothesis)
예시	H_0 : 대한민국 성인 남성의 키의 평균은 175cm 이다. H_0 : 성인 남성의 키는 성인 여성의 키와 같다	H_1 : 대한민국 성인 남성 키의 평균은 175cm와 같지 않다.(양측 검정) H_1 : 성인 남성의 키는 성인 여성의 키보다 크다(단측 검정)

통계적 가설 검정 - 가설 검정의 오류

- 가설 검정에서 발생하는 오류
  - 제1종 오류(Type 1 Error) : 귀무가설 참을 거짓이라 결정 함
  - 제2종 오류(Type 2 Error) : 귀무가설 거짓을 참이라고 결정 함

		결정	
	구분	**귀무가설 기각함 (F)**	**귀무가설 기각 안함 (T)**
실제 사실	**귀무가설 거짓 (F)**	올바른 선택	제2종 오류
	귀무가설 참 (T)	제1종 오류	올바른 선택

통계적 가설 검정 – p-value

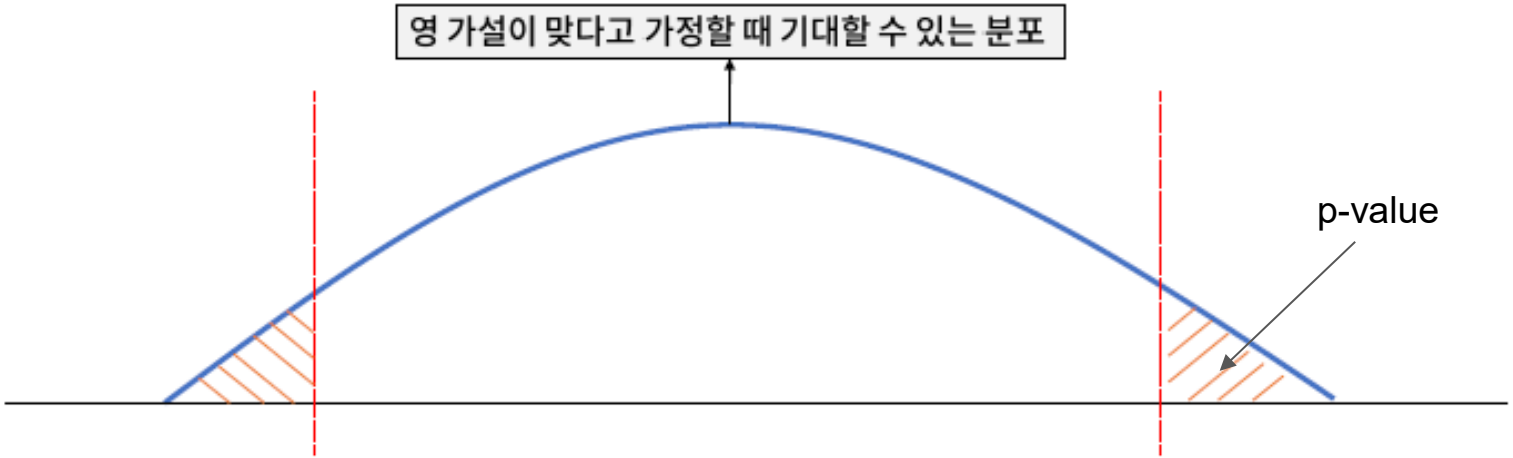
가설 검정에서 P-value 의미

- p-value(유의 확률)
  - 귀무가설이 참이라고 가정 했을 때, 우리가 현재 관측한 데이터보다 극단적인 결과가 나올 확률을 의미함.
  - **P-value가 작다는 것**은 귀무가설이 맞다고 보기 어려우며, 대립가설이 더 타당할 가능성이 높다는 뜻
- p-value 해석
  - $p < 0.05$: 귀무가설 기각 → 대립가설 채택(통계적으로 유의미함)
  - $p \geq 0.05$: 귀무가설 채택 → 대립가설을 채택할 증거 부족

귀무가설(영가설 H_0) : 기존 상태, 검정을 시작할 때 기본적으로 설정하는 가설
대립가설(연구가설 H_1) : 연구자가 입증하고 싶은 가설

통계적 가설 검정 – p-value

- **p-value(유의 확률)**
  귀무가설이 맞다고 가정할 때 얻은 결과와 **다른 결과가 관측될 확률**로, **그 값(p-value)이 작을 수록 귀무가설을 기각할 근거가 됨**

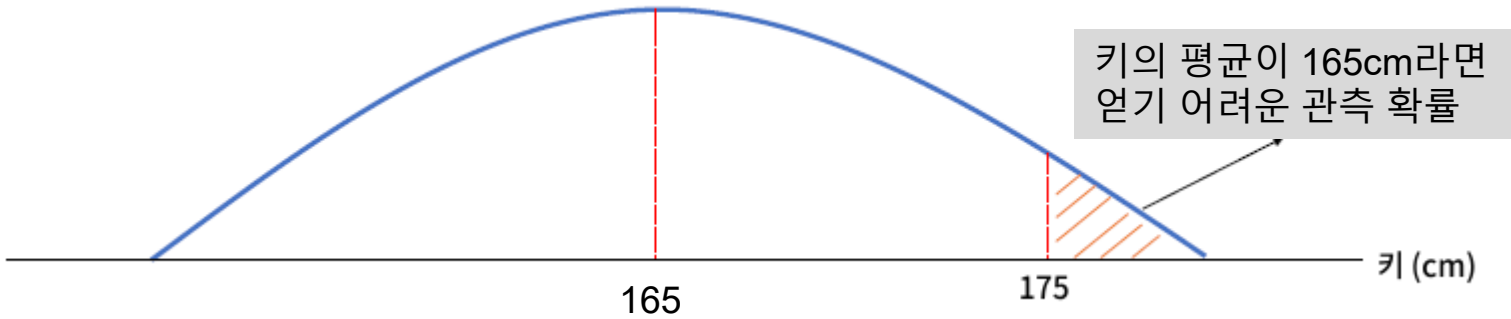


p-value가 0.05(5%) 미만이면 귀무가설(영 가설)을 기각함

통계적 가설 검정 – p-value

유의 확률, p-value(예시1) – 한 모집단

- 영 가설(H_0) : 대한민국 성인 남성의 평균 키는 165cm일 것이다.
- 대립 가설(H_1) : 대한민국 성인 남성의 평균 키는 165cm 이상일 것이다.
- 관측한 대한민국 성인 남성의 키는 평균 175cm, 표준편차 1cm이다.
- 관측값 개수 n = 20명



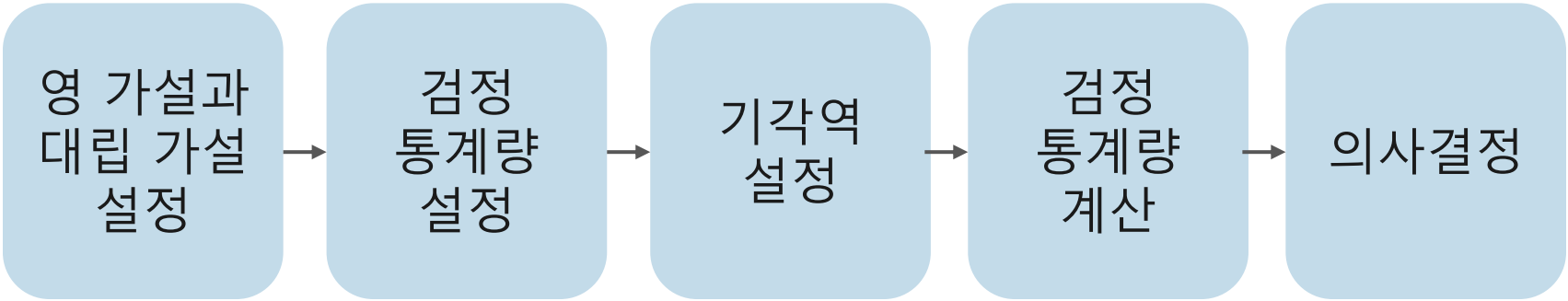
가설 검정의 종류

- 가설 검정과 사용 목적 설명

가설검정 종류	사용 목적
단일 표본 t-검정	한 그룹의 평균이 기준 값과 차이가 있는지를 확인
독립 표본 t-검정	서로 다른 두 그룹의 데이터 평균 비교
쌍체 표본 t-검정	특정 실험 및 조치 등의 효과가 유의한지 확인
일원분산(ANOVA) 분석	셋 이상의 그룹 간 차이가 존재하는지 확인
상관분석	두 연속형 변수 간에 어떠한 선형 관계를 가지는지 파악
카이제곱 검정	두 범주형 변수가 독립인지 파악

통계적 가설 검정

- 통계적 가설 검정 5단계



가설 수립

가설 검정

가설 성립한다
성립하지 않는다.
기각역 설정

설정한 검정
통계량을
계산함.

가설의 유의성을
판단하여 의사결정

단일 표본 t-검정
(One-Sample t-test)

단일표본 t검정

- 목적 : 그룹의 평균이 기준 값과 차이가 있는지를 확인하는 검정방법
- 영가설과 대립 가설

모집단 평균

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

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

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

$H_0: \bar{x} = 10$

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

단일표본 t검정

단일표본 t검정의 선행 조건

- 단일 표본 t-검정은 **해당 변수가 정규분포를 따라야** 수행할 수 있음
- Kolmogorov-Smornov 또 Shapiro-Wilk를 사용한 **정규성 검정이 선행**되어야 함
- 보통 샘플 수가 많을 수록(30개 이상) 정규성을 띌 가능성이 높아지므로, 샘플 수가 부족한 경우에만 정규성 검정을 수행한 뒤, 정규성을 띄지 않는다 라고 판단된다면 비모수적 방법인 부호검정(sign test)나 윌콕슨 부호-순위 검정을 수행해야 함.

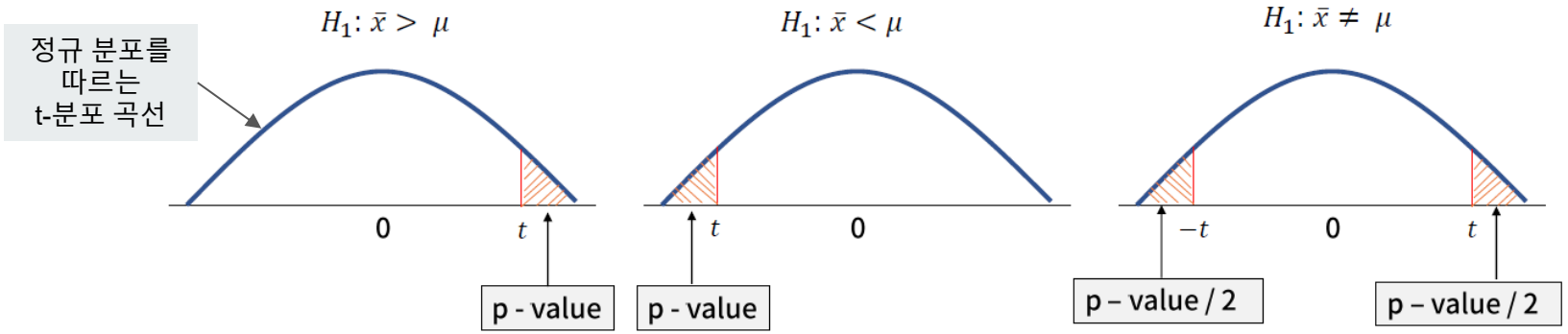
단일표본 t검정

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 t-분포 상에 위치시키는 방식으로 p-value를 계산함



- 단측 검정: **t 통계량**이 임계값보다 크거나 작은 경우 영가설을 기각함
- 양측 검정: $|t|$ 가 임계값보다 큰 경우 영가설을 기각함

정규성 검정 : Kolmogorov-Smornov

- Kolmogorov-Smornov검정(KS test)
  - **관측한 샘플들이 특정분포를 따르는지 확인**하기 위한 검정 방법
  - 해당 특정 분포를 정규 분포로 설정하여 **정규성 검정에도 사용**함
- KS test는 특정 분포를 따른다면 나올 것이라 예상되는 값과 실제 값의 차이가 유의한지를 확인하는 방법
  - 우리가 관측한 값과 특정 분포를 따랐을 때 나올 값과의 차이가 크면 클 수록 특정 분포를 따르지 않는다라고 볼 수 있음

단일 표본 t-검정

- 파이썬을 이용한 단일 표본 t-검정

구분	코드	결과해석
정규성 검정 (KS test)	scipy.stats.kstest(x, 'norm', args=(평균, 표준편차))	• result = (statistics, pvalue)의 튜플 형태 • p-value > **특정 수치(0.05) : 정규성을 따른**다고 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 의미함 • p-value <= **특정 수치(0.05) 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 **중위수**로 설정됨)

단일 표본 t-검정

- 2-1. **단일 표본 t검정**과 독립 표본 t검정 실습

독립 표본 t-검정
(independent sample t-test)

독립 표본 t-검정

- 목적 : **서로 다른 두 그룹의 데이터 평균 비교**
- 영가설과 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 :
  2024년 7월 한달 간 지점 A의 일별 판매량과 지점 B의 일별 판매량이 아래와 같다면, 지점A와 지점B의 7월 **판매량 간 유의미한 차이**가 있는가?

일자	지점A	지점B
2024.07.1	160	170
2024.07.2	220	180
…	…	…
2024.07.31	190	150
평균	200	180

독립 표본 t-검정

독립 표본 t-검정 선행 조건

- **독립성** : **두 그룹은 서로 독립적**이어야 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 0.05미만이면 분산이 다르다고 판단(등분산성이 없음)
  - 분산이 같은지 다른지에 따라 사용하는 통계량이 달라지므로, 설정만 달리해주면 됨

- 비모수 : 특정한 분포(예: 정규분포)에 대한 가정이 없거나, 그런 가정을 하지 않는 검정 방법을 의미
- 등분산성 검정은 두 개 이상의 그룹이 같은 분산을 갖는지 평가하는 검정

독립 표본 t-검정 통계량

- 두 그룹의 분산이 같은 경우

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

✓ $\bar{x}_a$: 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

독립 표본 t-검정 통계량

- 파이썬을 이용한 독립 표본 t-검정

구분	코드	결과해석
정규성 검정 (KS test)	scipy.stats.kstest(x, 'norm', args=(평균, 표준편차))	• p-value > **특정 수치(0.05) : 정규성**을 따른다고 판단할 수 있음
등분산성 검정 (Levene test)	scipy.stats.levene(s1,s2,s3,...) # s1, s2, ... : 샘플(배열)	• p-value > **특정 수치(0.05) : 샘플 간 분산이 유사하다고 판단할 수 있음(등분산성 만족)** p-value <= **특정 수치(0.05) : 샘플 간 분산이 같지 않다고** 판단할 수 있음
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 • pvalue <= **특정 수치(0.05) : a와 b의 평균이 같지 않다**고 판단
Mann-Whitneyu검정	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 않다**고 판단

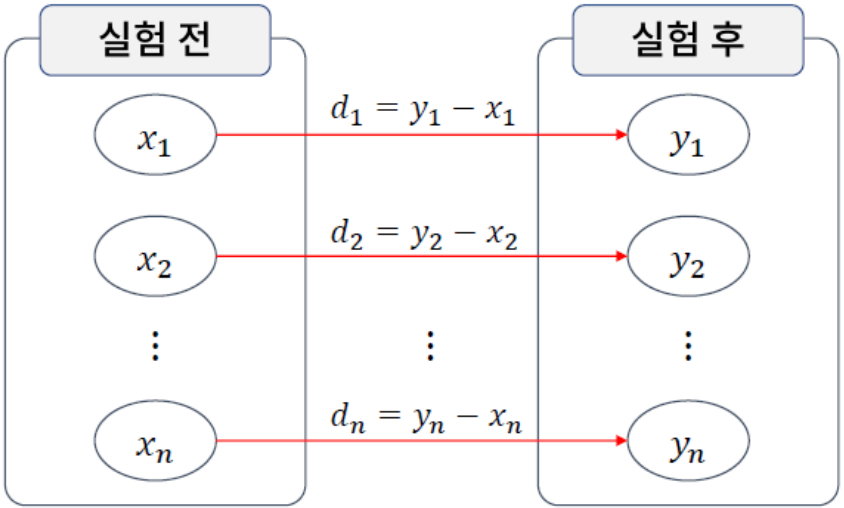
독립 표본 t-검정

- 2-1. 단일 표본 t검정과 **독립 표본 t검정** 실습

쌍체 표본 t-검정
(paired sample t-test)

쌍체 표본 t-검정

- 목적 : 특정 실험 및 조치 등의 **효과가 유의미한지**를 확인



$H_0: d = 0$

$H_1: d > 0$ or $d < 0$ or $d \neq 0$

쌍체 표본 t-검정

가설 수립 예시

- 문제 상황
  - 어떤 제약회사가 신약을 투여하기 전과 후의 혈압 변화를 관찰하고자 합니다.
  - 같은 환자 10명을 대상으로 신약 투여 전후의 혈압을 측정한 데이터가 있다고 가정합니다.
- 목표
  - 신약 복용 전과 후의 평균 혈압 간 유의미한 차이가 있는지를 검정한다

[실험 데이터]

대상자	복용 전	복용 후	차이
1	145	138	7
2	150	142	8
3	138	135	3
4	142	140	2
5	148	145	3
...	...	...	...
9	151	146	5
10	144	139	5

쌍체 표본 t-검정

- 쌍체 표본 t-검정의 선행 조건
  - 데이터는 쌍(pair)d으로 구성되어 있어야 함.
  - 실험 전과 후의 측정 값(즉, X와 Y)은 정규 분포를 따르지 않아도 무방함. 그러나 **측정 값의 차이인 d는 정규성**을 갖고 있어야 함.
  - 이상치가 없어야 함(검정 결과의 왜곡을 피하기 위함.

- 쌍체 표본 t-검정의 통계량

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

✓ $\bar{d}$: d의 평균

✓ s_d: d의 표준편차

$s_d/\sqrt{n}$ 표준오차 SE

쌍체 표본 t-검정

- 파이썬을 이용한 쌍체 표본 t-검정

구분	코드	결과해석
정규성 검정	- Shapiro-Wilk 검정 scipy.stats.shapiro(d)	• 소표본에 강하고 정규성 검정 전용 • p-value > 특정 수치(0.05) : **정규성을 따른다고 판단할 수 있음**
	- D'Agostino-Pearson 검정 scipy.stats.normaltest(d)	• 샘플 수가 (n >= 20) 일 때 사용하는 권장 • p-value > 특정 수치(0.05) : 정규성을 따른다고 판단할 수 있음
쌍체 표본 t 검정	scipy.stats.ttest_rel(a, b) # a,b : 실험 전 후 결과 (주의 : 반드시 길이가 같아야 함)	• pvalue <= **특정 수치(0.05): 그룹 a와 그룹 b간 차이가 존재**한다고 판단 (즉, 특정 실험의 효과가 존재함) • statistics가 양수면 양의 효과(d>0)가 있다고 판단하며, 음수면 음의 효과(d<0)가 있다고 판단

쌍체 표본 t-검정

- 2-2. 쌍체 표본 t검정 실습
- 문제 : 다이어트 전과 후의 차이가 유의한지 검정하기

일원분산(ANOVA) 분석
(One-way ANalysis Of Variance)

일원분산(ANOVA) 분석

- 목적 : **셋 이상의 그룹 간 차이가 존재**하는지를 확인하기 위한 가설 검정 방법임(ANOVA 분산 분석)
- 영 가설과 대립가설

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

[Tip]

* ANOVA(**AN**alysis **O**f **Va**riance) : 분산분석
* 일원분산분석(One-way ANOVA) : 분석에 사용된 요인이 한 개(one factor)

용어	설명	예시
요인(factor)	그룹을 나누는 기준	지점, 브랜드, 수업 방식 등
수준(level)	요인의 하위 분류	A지점, B지점, C지점

일원분산(ANOVA) 분석

가설 수립 예시

- 문제 상황
  - 2024년 7월 한 달 동안 지점 A, B, C의 일별 판매량이 각각 기록되었다.
  - 세 지점간 유의미한 차이가 있는지 궁금하다.
- 목표
  - 전체 지점 간 유의미한 차이가 존재 여부 검정
  - 개별 지점 간 차이 여부 파악

일자	지점A	지점B	지점C
2024.07.1	160	170	180
2024.07.2	220	180	185
…	…	…	…
2024.07.31	190	150	140
평균	200	190	180

일원분산(ANOVA) 분석

일원분산분석의 선행 조건

- **독립성** : 모든 그룹은 서로 **독립적**이어야 함(실험 설계, 데이터 수집 시)
- **정규성** : 모든 **그룹의 데이터는 정규분포를 따라야 함**
  - Shapiro-Wilk 사용, Q-Q((Quantile-Quantile) Plot 확인
  - 그렇지 않으면 비모수적인 방법인 Kruskal-Wallis H Test를 수행해야 함.
- **등분산성** : 모든 그룹에 데이터에 대한 분산이 같아야 함.
  - Levene test(레빈 검정) 사용
  - 그렇지 않으면 등분산성 가정 없이 평균 차이를 비교하는 Welch's ANOVA 수행

일원분산(ANOVA) 분석

- 일원 분산분석의 통계량

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

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

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

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

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

일원분산(ANOVA) 분석

- 사후 분석 : Tukey HSD test
  - Tukey HSD(honestly significant difference) test는 일원분산분석에서 전체 그룹 간 평균 차이가 유의할 때, 두 그룹 a와 b간 차이가 유의한 지 파악하는 사후 분석 방법임

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

✓ μ_a: 그룹 a의 평균 ✓ SE: 그룹 a와 b의 표준 오차

✓ μ_b: 그룹 b의 평균

  - 만약, $HSD_{a,b}$가 유의 수준(임계값)보다 크면, 행당 그룹 간에는 유의미한 차이가 있다고 간주 함.

* 사후 분석 : 실험이나 통계 분석에서 일차적인 검정(primary test) 후에 추가로 수행하는 분석을 의미
* 유의하다 : “단순한 우연이나 무작위로 생긴 결과가 아니라,실제로 어떤 차이 또는 효과가 존재할 가능성이 크다”는 뜻

일원분산(ANOVA) 분석

- 파이썬을 이용한 일원분산분석

구분	코드	결과해석
정규성 검정	- Shapiro-Wilk 검정 scipy.stats.shapiro(d)	• p-value > **특정 수치(0.05) : 정규성**을 따른다고 판단할 수 있음
등분산성 검정 (Levene test)	scipy.stats.levene(s1,s2,s3,...) # s1, s2, ... : 샘플(배열)	• p-value > **특정 수치(0.05) : 샘플 간 분산이 유사하다고 판단할 수 있음(등분산성 만족).**
일원분산분석	scipy.stats.f_oneway(sample1, sample2, sample3, ...)	• p-value <= **특정 수치(0.05) : 최소 하나의 그룹은 다른 그룹의 평균**과 다르다고 판단 (집단 간 차이를 유발하는 요인/효과가 존재한다고 볼 수 있음)
사후분석	- 모수적 사후분석 statsmodels.stats.multicomp.pairwise_tukeyhsd(endog, groups, alpha=0.05)	• 정규성을 따를 때 사용 • endog : 분석 대상이 되는값(점수, 판매량), groups : 각 값이 속한 그룹 이름(A,B,C 등), alpha : 유의수준 • reject가 True이고, p-adj <= 0.05이면, 두 그룹 간 통계적으로 유의한 평균 차이가 있다고 판단함.
	- 비모수적 사후분석 scipy.stats.mannwhitneyu(a, b) # a,b : 두 그룹의 데이터(배열)	• 정규성을 따르지 않거나, 이상치가 많을 때, 중앙값 차이 또는 순위 차이를 비교하기 위한 방법 • p-value <= 특정 수치(0.05) : 두 집단 간에는 통계적으로 유의한 차이가 있다고 판단함.

일원분산(ANOVA) 분석

- 2-3. 일원분산분석 실습
- 문제 : 세 지점의 일별 판매량의 평균이 유의한 차이가 존재하는지 검정하기

상관분석
(Correlation Analysis)

상관분석

- 목적 : **두 연속형 변수** 간에 어떠한 **선형 관계**가 존재하는지, 그리고 그 관계의 방향과 강도를 파악하기 위함
- 영 가설과 대립 가설

H_0: 두 변수 간에는 유의미한 상관성이 존재하지 않는다	r = 0
H_1: 두 변수 간에는 유의미한 상관성이 존재한다	r != 0, r < 0, r > 0

- 방법
  - 피어슨 상관계수 : 정규분포를 따르는 연속형 변수 일 때, 이상치 민감
  - 스피어만 상관계수 : 서열형(순위형), 정규성을 몰라도 됨
- 시각화 방법 : 산점도(scatter plot), 히트맵(heatmap plot)

상관분석 - 피어슨 상관 계수

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

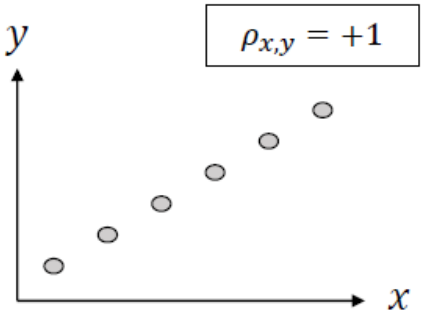
$$\rho_{x,y} = \frac{cov(x,y)}{\sqrt{var(x) \times var(y)}}$$

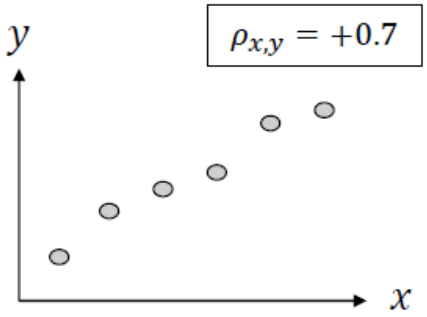
- ✓ $cov(x,y)$: x와 y의 공분산, $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-y)}{n-1}$
- ✓ $var(x)$: x의 분산

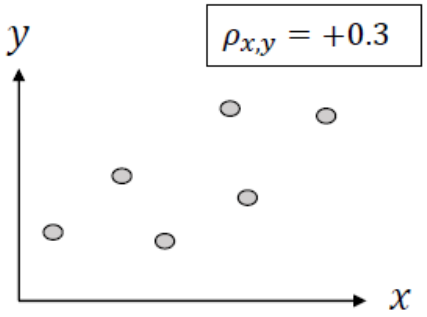
- 상관 계수가
  - 1에 가까울 수록 양의 상관관계가 강하다고 하며,
  - -1에 가까울 수록 음의 상관관계가 강하다고 함.
  - 0에 가까울 수록 상관관계가 약하다고 함.

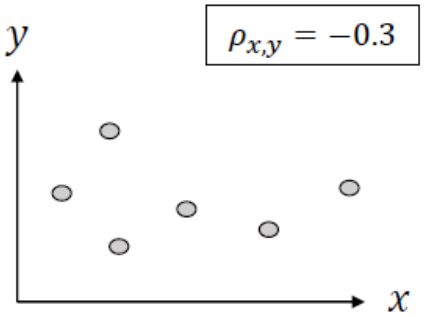
* 공분산(Covariance)은 두 개의 변수가 얼마나 함께 변하는지를 측정하는 통계량

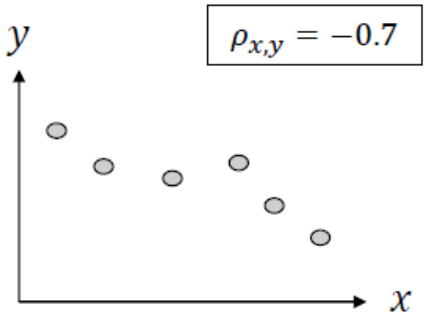
피어슨 상관 계수 산점도 그래프

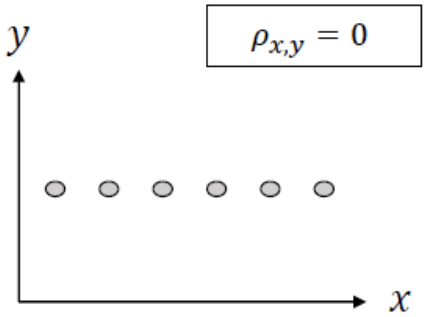










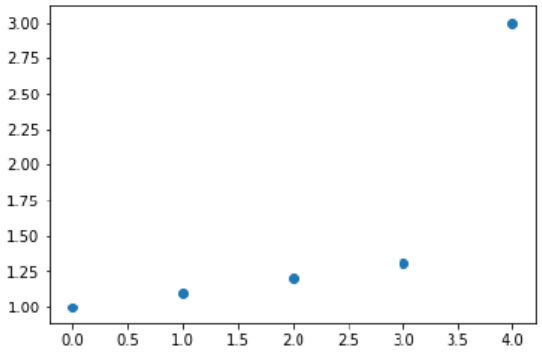


상관분석 - 스피어만 상관 계수

- 두 변수의 순위 사이의 **단조 관련성을 측정**하는 상관 계수로 x, y에 대한 스피어만 상관 계수 $S_{x,y}$는 다음과 같이 정의됨

$S_{x,y} = \rho_{r(x),r(y)}$ ✓ $r(x)$: x**의 요소의 개별 순위** (x의 순위와 y의 순위 사이의 상관계수)

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



$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

*** 단조 관련성 :** 한 변수가 증가할 때 다른 변수도 함께 증가하거나 감소하는 경향이 있는 관계를 말함

상관분석 - 피어슨 상관 계수

- 파이썬을 이용한 상관분석

구분	코드	결과해석
피어슨 상관계수	scipy.stats.pearsonr(x, y)	• 결과 (statistics, p-value) • statistics: 피어슨 상관계수(-1 ~ 1) (H_0 : 모집단 상관계수 $\rho = 0$) • p-value <= 0.05 : 유의한 선형 상관성이 있다고 판단
스피어만 상관계수	scipy.stats.spearmanr(x, y)	• 결과 (statistics, p-value) • statistics: 스피어만 상관계수(-1 ~ 1) (H_0 : 모집단 상관계수 $\rho_s = 0$) • p-value <= 0.05 : 유의한 순위 기반 상관성이 있다고 판단
사후분석	DataFrame.corr(method) #method: **pearson**, spearman	• 컬럼 간 상관계수 계산 방법(2D 행렬)

상관분석 실습

- 2-4. 상관분석 실습
- 문제 : 금, 은, 달러의 상관성 분석

카이제곱 검정
(Chi-Square Test)

카이제곱 검정

- 목적 : 두 **범주형** 변수가 서로 독립적인지 검정
- 영가설과 대립 가설

H_0: 두 변수가 서로 독립이다 (연관 없음)

H_1: 두 변수가 서로 종속된다 (연관 있음)

- 시각화 방법 : 교차 테이블
- 적용 상황 예)
  - 성별(남/여)과 구매 여부(구매/비구매)의 관계
  - 지역(서울/부산/대구)과 투표 여부(투표/비투표)의 관계
  - 학년(1~4학년)과 스마트폰 브랜드 선호(애플/삼성/기타)의 관계

카이제곱 검정

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;상태공간 = 2x3 = 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: $\frac{100 \times 90}{200} = 45$

카이제곱 검정 – 카이제곱 통계량

- 카이제곱 검정에 사용하는 카이제곱 통계량은 **기대값과 실제값의 차이**를 바탕으로 정의됨 χ : 카이(그리스어)

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

- ✓ 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 변수간 연관성이 존재할 가능성이 높음
  - p-value <= 0.05
  - H_0: 두 변수는 독립이다 (변수 간에 연관성이나 상관성 없음)

카이제곱 검정

- 파이썬을 이용한 카이제곱 검정

구분	코드	결과해석
교차 테이블 생성	pandas.crosstable(S1, S2)	• Series S1과 S2로 구성된 교차테이블 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)

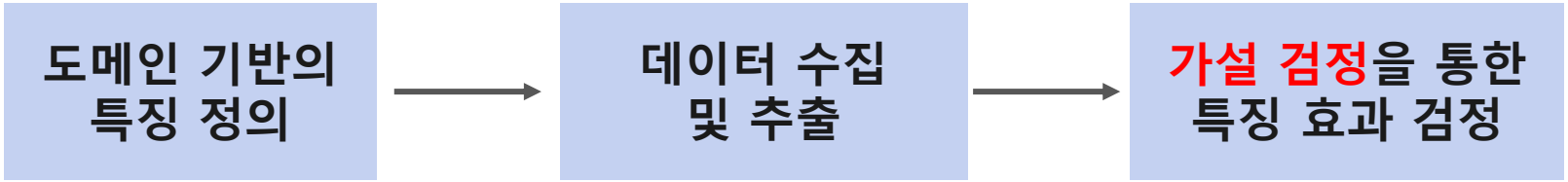
카이제곱 검정

- 2-5. 카이제곱검정 실습
- 문제 : 성별과 만족도는 서로 관련이 있는가?

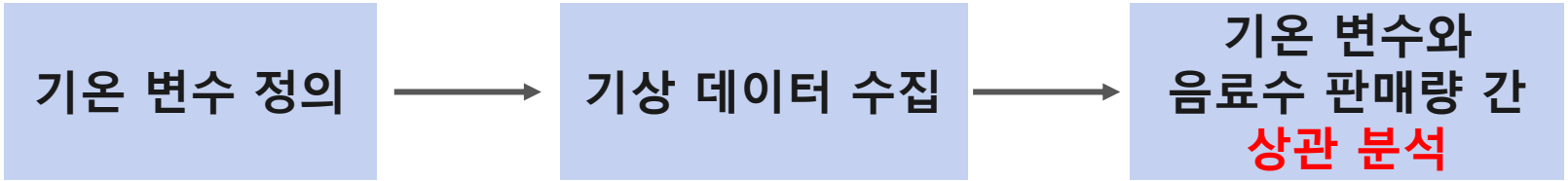
머신러닝에서의 가설검정

특징 정의 및 추출

- 예측 및 분류에 효과적인 특징을 정의하고 추출하는 과정

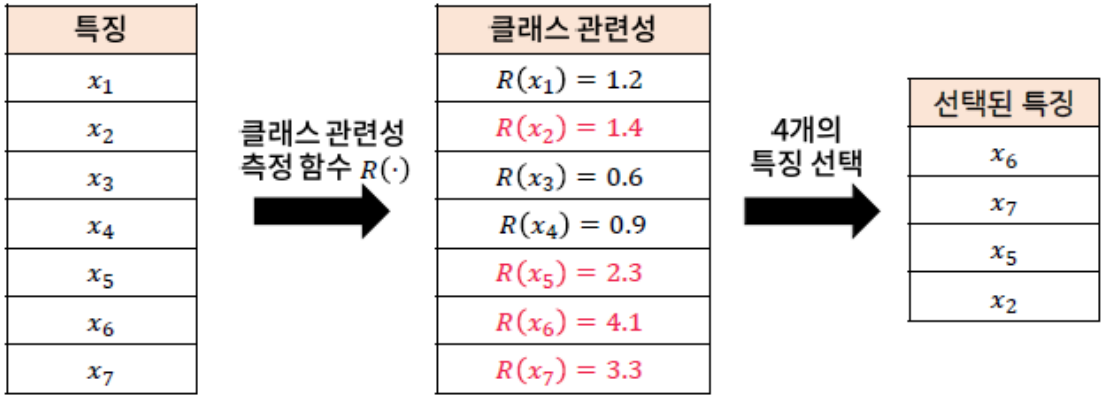


- 예시) 음료수 판매량 예측



특징 선택

- 특징 선택(Feature Selection)
  - **예측 및 분류에 효과적인 특징**을 선택하여 모델의 성능을 높이고, 차원의 저주를 완화하며 과적합을 줄이기 위한 기법
- 특징 효과성(클래스 관련성) 평가
  - 가설 검정에서 사용하는 통계량을 사용할 수 있음.
  - 연속형 목표 변수(t-통계량, ANOVA), 범주형 목표 변수(카이제곱 검정)



정리

가설 검정의 종류 정리

- 가설 검정과 사용 목적 설명

가설검정 종류	사용 목적
단일 표본 t-검정	한 그룹의 평균이 기준 값과 차이가 있는지를 확인
독립 표본 t-검정	서로 다른 두 그룹의 데이터 평균 비교
쌍체 표본 t-검정	특정 실험 및 조치 등의 효과가 유의한지 확인
일원분산(ANOVA) 분석	셋 이상의 그룹 간 차이가 존재하는지 확인
상관분석	두 연속형 변수 간에 어떠한 선형 관계를 가지는지 파악
카이제곱 검정	두 범주형 변수가 독립인지 파악

통계개념 정리

- 가설검정은 귀무가설(H_0)과 대립가설(H_1)을 설정하고, 표본 데이터를 통해 귀무가설이 통계적으로 기각될 수 있는지를 검정하는 절차
  - ① 귀무가설 설정, ② 표본 수집, ③ 검정 통계량 계산, ④ 유의확률(p-value) 비교 ⑤ 귀무가설을 기각하거나 채택함
- p-value(유의 확률) : 귀무가설이 옳다는 전제 하에 표본에서 실제로 관측된 통계값과 같거나 더 극단적인 통계값이 우연히 관측될 확률임
  - p-value <= 0.05 : 관찰된 결과가 귀무가설 하에서 매우 드물게(5% 이하의 확률로) 나타남 이런 극단적인 결과는 거의 일어나지 않음 → 귀무가설 기각
  - p-value > 0.05 : 관찰된 결과가 귀무가설 하에서도 충분히 일어날 수 있는 5% 초과의 확률이라는 의미. 귀무가설을 기각할 충분한 근거가 없다고 판단 → 귀무가설을 기각하지 못함

통계개념 정리

- 평균이 뮤(μ)이고 분산이 시그마^2(σ^2)인 모집단으로부터 가능한 모든 n개의 조합을 표본으로 추출하면 표본 평균들의 분포는 점점 정규분포에 가까워짐
  → 중심극한 정리,
  → **정규분포 기반의 통계기법**을 다양한 상황에 적용
  → 표본 크기 n >= 30 이면 대부분 정규분포에 충분히 근사함.

- 카이제곱 검정은 범주형 데이터의 통계 분석에 사용되는 검정으로, 2개의 범주형 변수 사이에 독립성을 판단함

- ANOVA 분산분석은 모집단이 셋 이상인 경우, 이들의 평균이 서로 동일한지 테스트함.


감사합니다