6차시
통계적 검증과 예측

20 . .

지은이: 이은재

Ⅰ. 통계적 추론

- 통계적 추론이란?

Statistical Inference: 표본 데이터를 바탕으로 모집단의 특성을 추정하거나 결론을 내리는 과정.

- 통계적 추론 사례

- 1,000명이 참여한 여론조사를 통해 국내 전체 유권자의 지지율을 추정
- 임상시험 표본 데이터를 통해 신약의 효과 판단
- 스마트팜 센서 일부 위치의 온도를 측정해 전체 온실 상태 추정
- 유튜브 조회수 일부 데이터를 기반으로 전체 추천 알고리즘의 방향성 파악

- 모수 검정 (Parametric Test)

: 데이터가 특정 분포(보통 **정규분포**)를 따른다고 가정하고 수행하는 통계적 검정.

구분	탐색적 분석 (EDA)
t-test	표본 집단 두 그룹간의 평균이나 샘플값을 비교
z-test	모집단을 아는 경우 평균이나 샘플값 비교
ANOVA	셋 이상의 그룹 간 평균과 샘플 비교
F-test	두 그룹간의 분산 비교

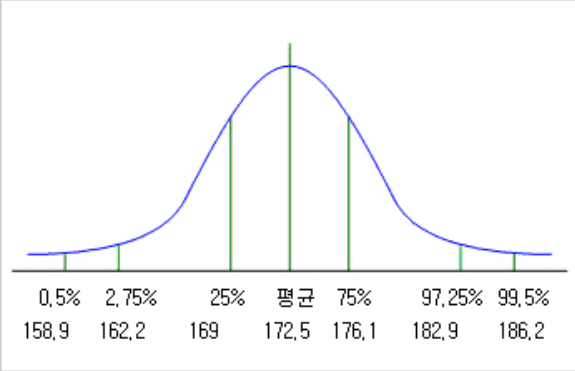
- 표준점수 (Z-score) 구하기

데이터가 평균에서 얼마나 떨어져 있는지 표준 편차를 사용해 변환한 점수.

z = x - 모집단의 평균 / 모집단의 표준 편차

- 누적분포 이해하기

표준정규분포 -> 평균이 0 / 표준편차가 1 이다.

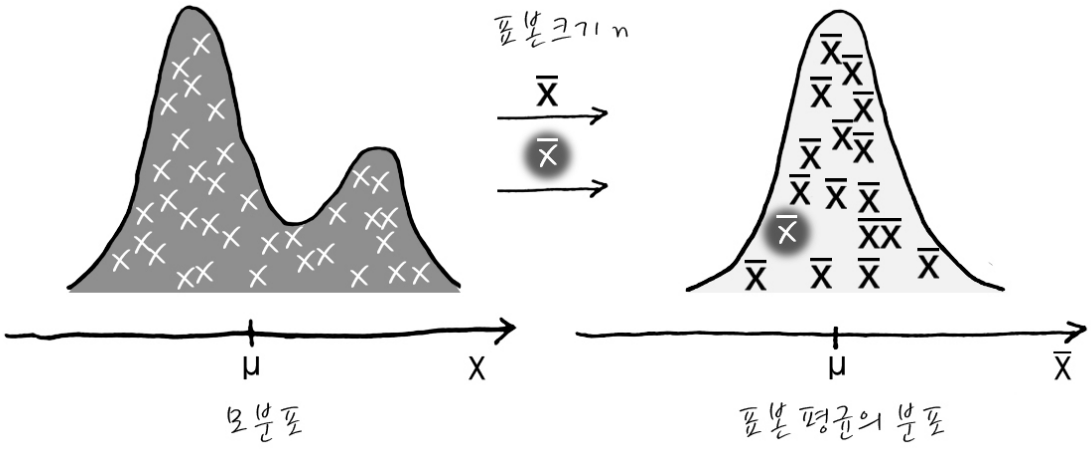


따라서, 표준정규분포를 따르는 데이터는 각 샘플의 z점수를 알면 전체 데이터가 어떻게 분포되어있는지 알 수 있다.

Q. 데이터가 표준정규분포를 따르지 않는다면?

- 중심극한정리

무작위로 샘플을 뽑아 만든 표본의 평균은 정규분포에 가깝다!



Ⅱ. 가설검정

표본 데이터를 이용해 어떤 주장이 통계적으로 타당한지 판단하는 절차.
"우리가 관찰한 차이가 우연인지, 아니면 의미 있는 변화인지"를 검증하는 과정이다.

- 귀무가설 (영가설, null hypothesis)

우리가 세계에 대해 이미 알고 있는 바를 나타내는 가설.
모든 통계적 검정은 먼저 이 귀무가설을 *기본 진실(default truth)*로 놓고 출발한다.
귀무가설이 참으로 확인되면, "차이가 없다", "효과가 없다", "변화가 없다"는 의미이다.

예시)
- 두 집단 평균이 동일하다.
- 새로운 비료는 기존 비료와 작물 생산량 차이가 없다.
- 신약은 기존 약과 효과 차이가 없다.

귀무가설은 검정 과정에서 기각되거나, 기각되지 않거나 둘 중 하나이다.
"증거가 부족하면 유지된다."

- 대립가설 (alternative hypothesis)

우리가 세계에 대해 이미 알고 있는 바를 대체할 수 있는 내용을 나타내는 가설.
귀무가설과 논리적으로 반대되며 상호 배타적인 주장을 담고 있다.

대립가설이 참으로 확인되면, "차이가 있다", "효과가 있다", "변화가 있다"는 의미이므로
새로운 가능성을 발견할 수 있다!

귀무가설이 기각될 때만 채택될 수 있다.

대립가설에는 두 형태가 있다:

양측검정: 차이의 유무를 확인하기 (두 평균이 *다르다)*
단측검정: 특정 방향으로의 차이를 확인하기 (한쪽 방향으로 *크다 / 작다)*

예:
한 약의 효과가 더 크다(H_1: $\mu_1 > \mu_2$)
두 집단 평균이 다르다(H_1: $\mu_1 \neq \mu_2$)

- 유의 수준

“우연일 가능성을 몇 %까지 허용할 것인가?”를 정하는 값.
유의수준은 귀무가설을 기각할 기준선 역할을 한다.

보통 $\alpha = 0.05$ 또는 0.01을 사용한다.
귀무가설이 사실인데도 기각할 확률을 5%까지만 허용하겠다.
즉, 5% 확률의 ‘우연한 극단적 상황’은 그냥 잡음으로 보고 넘어가겠다는 의미로 사용한다.

- p-value

귀무가설이 참이라고 가정했을 때, 지금 데이터처럼 극단적인 결과가 나올 **확률**.

p-value가 작을수록, 지금 관측된 결과는 “우연이라 보기 어렵다”는 의미가 된다.

판단 기준:
p-value $< \alpha \rightarrow$ 귀무가설 **기각**, 대립가설 **채택**
p-value $\geq \alpha \rightarrow$ 귀무가설 **기각 실패**(단, 채택이 아님!)

중요:
p-value는 ‘귀무가설이 참일 확률’이 아니다.

- z-점수 검정과 t-점수 검정

- Z-검정 (Z-test)

모집단의 분산 또는 표준편차를 알고 있는 경우,
또는 표본 크기가 충분히 큰 경우($n \geq 30$)사용하는 검정.

예:
검사 장비가 충분히 정확해 모집단 표준편차를 알고 있을 때
QC(품질관리)에서 공정 평균을 계속 모니터링할 때

-> 모집단을 아는 경우에만 사용할 수 있으므로 실제 사용은 제한적

- t-검정 (t-test)

모집단 분산을 모를 때, 또는 표본 크기가 작을 때($n < 30$)사용.

특징:
t-분포 사용
표본의 불확실성을 고려하므로 z-분포보다 두꺼운 양쪽 꼬리를 가진다.
실제 연구·분석에서는 거의 대부분 t-검정을 사용한다.

표본 크기가 커지면 t-분포는 정규분포에 수렴하므로,
결국 t-test와 z-test 결과는 비슷해진다.

Ⅲ. 모델링과 머신러닝

- 모델링이란?

데이터 속에 숨어 있는 **관계, 패턴, 규칙**을 수학적·통계적 구조로 표현하는 과정이다.
즉, 현실 세계의 복잡한 현상을 **수식이나 알고리즘 형태로 단순화해 설명하고 예측하는 작업**을 말한다.

1. 왜 모델링을 하는가?

모델링의 핵심 목적은 크게 세 가지다.

① 설명(Explanation)

"무엇이 결과에 영향을 주는가?"
→ 변수들 사이의 관계를 이해하기 위해 모델을 만든다.
예: 비료 투여량이 증가할수록 작물 생산량이 얼마나 변하는가?

② 예측(Prediction)

"새로운 데이터가 들어왔을 때 결과를 미리 예측할 수 있는가?"
예: 내일의 온실 온도, 유튜브 영상 조회수, 판매량 예측.

③ 의사결정(Decision Making)

모델을 기반으로 더 나은 전략 세우기, 리스크 관리, 자원 배분 등을 수행한다.

2. 모델링의 과정

문제 정의
무엇을 예측/설명하려는가?

데이터 수집 및 전처리
결측치 처리, 이상치 제거, 스케일링 등

모델 선택
선형 회귀, 로지스틱 회귀, 결정트리, 랜덤포레스트, 딥러닝 등

모델 학습(훈련)
데이터를 이용해 패턴을 학습시킴

평가(Evaluation)

훈련/테스트 데이터로 모델 성능을 검증
R^2, RMSE, 정확도, F1 등 지표 사용

튜닝(Tuning)
파라미터 조정, 특징(feature) 재선택, 과적합 방지

배포 및 활용
실제 환경에서 예측 모델로 사용
앱, 서비스, 비즈니스 분석 등에 활용

3. 모델링의 종류

① 통계적 모델링 (Statistical Modeling)
주로 "관계 설명"에 사용한다.
선형 회귀, 로지스틱 회귀처럼 해석 가능한 모델 중심
모델의 가정이 명확함 (정규성, 선형성 등)

② 머신러닝 모델링 (Machine Learning Modeling)
"예측 정확도"에 초점.
복잡한 패턴도 잘 잡아냄
랜덤포레스트, SVM, XGBoost, 신경망 등
해석보다 성능이 중요할 때 사용

4. 모델링의 핵심 개념
변수(Feature): 모델이 바라보는 정보 요소
선형성/비선형성: 관계가 직선인가, 곡선인가
과적합(Overfitting): 훈련 데이터에만 맞춰지고 일반화가 안 되는 현상
일반화(Generalization): 처음 보는 데이터에서도 성능 유지
가설 공간(Hypothesis Space): 모델이 표현할 수 있는 패턴의 범위

- 머신러닝을 사용한 모델링

- 지도학습과 비지도 학습

- 지도학습(Supervised Learning)
정답(label)이 있는 데이터로 학습
회귀(연속값 예측)
분류(범주 예측)
예: 집값 예측, 스팸 메일 분류

- 비지도학습(Unsupervised Learning)
정답이 없는 데이터를 구조화
군집(k-means)
차원축소(PCA)
예: 고객 세그먼트, 이상치 탐지

- 훈련 세트와 테스트 세트

훈련 세트 (Train set)
모델 학습용. 패턴을 배우는 데이터.

검증 세트 (Validation set)
모델 튜닝, 파라미터 조정, 구조 변경에 사용.
(딥러닝에서는 필수)

테스트 세트 (Test set)
최종 성능 평가용.
모델이 처음 보는 데이터.

- 결정계수(R^2)

회귀 모델의 설명력을 나타내는 지표.
$0 \leq R^2 \leq 1$
1에 가까울수록 데이터의 변동을 잘 설명
음수가 나올 수도 있음 → 모델이 형편없다는 뜻

- 선형 회귀 (Linear Regression)

연속적인 값을 예측하는 가장 기본 모델.

핵심 아이디어:
y = ax + b 형태의 관계를 찾기
오차를 최소화하는 a와 b를 계산 (최소제곱법)

장점:
해석이 쉽고 학습이 빠르다.

- 로지스틱 회귀 (Logistic Regression)

0/1 분류 문제를 위한 모델.
핵심 아이디어:
출력이 확률값으로 나오는 시그모이드 함수를 활성화 함수로 사용하자.
“특성이 증가하면 확률이 어떻게 변하는지” 직관적으로 이해 가능

예:
스팸 여부
구매할지 말지
진단 결과(양성/음성)

- 예측하기

훈련된 모델에 새로운 데이터를 입력 → 예측값 출력.

절차:
데이터 전처리(정규화, 결측값 처리)
모델에 입력
예측 결과 해석
필요하면 재학습 → 성능 향상

예측의 핵심은 **“과거 패턴을 이용해 미래를 추정하는 것”이다.**
- 과거 패턴을 그대로 따르는 데이터는 예측이 쉽지만,
- 패턴을 크게 따르지 않는 데이터일수록 예측이 어렵다.

Q. 어떤 예측이 쉽고 어려울까?