

Contents

1 Unit 1. 통계적 모델링과 우도 (Modeling, Identifiability, Likelihood)	5
1.0.1 1. 통계적 모델 (Statistical Model)	5
1.0.2 2. 식별 가능성 (Identifiability)	6
1.0.3 3. 우도 (Likelihood)	7
1.1 로그 우도 (Log-Likelihood)와 실전 적용	8
1.2 자주 묻는 질문 (FAQ)	9
2 Unit 2. 점 추정 (Point Estimation)	13
2.0.1 1. 추정량(Estimator)의 정의와 본질	13
2.0.2 2. 최대우도추정 (MLE, Maximum Likelihood Estimation)	14
2.0.3 3. 적률법 (MoM, Method of Moments)	15
2.0.4 4. 추정량의 성질 (Properties): 성적표 매기기	15
2.1 실전 시나리오: 모바일 게임 가챠 확률 검증	16
2.2 자주 묻는 질문 (FAQ)	16
3 Unit 3. 점근적 성질 (Asymptotic Properties)	21
3.0.1 1. 일치성 (Consistency)	21
3.0.2 2. 점근적 정규성 (Asymptotic Normality)	22
3.0.3 3. 피셔 정보와 크래머-라오 하한 (CRLB)	22
3.0.4 4. 델타 방법 (Delta Method)	23
3.1 실전 시나리오: 모바일 게임 리텐션(재접속률) 예측	23
3.2 자주 묻는 질문 (FAQ)	24
4 Unit 4. 신뢰 구간 (Confidence Intervals)	27
4.0.1 1. 신뢰 구간의 정의와 철학	27
4.0.2 2. 점근적 신뢰 구간 구성법 (The Recipe)	28
4.0.3 3. 보수적 신뢰 구간 (Conservative CI)	29
4.1 실전 시나리오: A/B 테스트 신뢰 구간	29
4.2 자주 묻는 질문 (FAQ)	30
5 Unit 2 (Part B). 가설 검정 (Hypothesis Testing)	33
5.0.1 1. 가설 검정의 구조: 비대칭적 싸움	33
5.0.2 2. 두 가지 종류의 오류 (Type I & Type II Error)	34
5.0.3 3. 유의 수준(α)과 검정력(Power): 최적화 문제	34
5.0.4 4. p-value의 정확한 정의와 오해	35
5.1 실전 시나리오: 넥슨 신규 아이템 매출 분석	36
5.2 자주 묻는 질문 (FAQ)	36

6 Unit 6. 검정 방법론 (Testing Methodology)	41
6.0.1 1. 네이만-피어슨 보조정리 (Neyman-Pearson Lemma)	41
6.0.2 2. 우도비 검정 (LRT)과 월크스 정리	42
6.0.3 3. 통계적 검정의 삼위일체 (The Holy Trinity)	43
6.0.4 4. t-test: 데이터가 적을 때의 현실	43
6.1 실전 계산: 불공정 동전 판별 (Wald Test)	44
6.2 자주 묻는 질문 (FAQ)	44
7 Unit 7. 적합도 검정 (Goodness of Fit)	49
7.0.1 1. 이산 데이터와 다항 분포 (The Setup)	49
7.0.2 2. 피어슨의 카이제곱 검정 (Pearson's χ^2 Test)	50
7.0.3 3. 복합 가설 검정 (Composite Goodness of Fit)	50
7.1 실전 시나리오: 확률형 아이템(가차) 조작 의혹 검증	51
7.2 자주 묻는 질문 (FAQ)	52
8 Unit 3. 선형 회귀 (Linear Regression)	55
8.0.1 1. 선형 회귀 모델의 구조 (The Setup)	55
8.0.2 2. 최소자승법 (Least Squares Estimation, LSE)	56
8.0.3 3. 기하학적 해석: 직교 투영 (Projection)	56
8.0.4 4. 가우스-마르코프 정리 (Gauss-Markov Theorem)	57
8.0.5 5. 추론 (Inference): 가설 검정	57
8.1 실전 시나리오: 넷플릭스 시청 시간 예측	58
8.2 자주 묻는 질문 (FAQ)	58
9 Unit 9. 일반화 선형 모형 (GLM)	63
9.0.1 1. 지수족 분포 (Exponential Family)	63
9.0.2 2. 연결 함수 (Link Function)	64
9.0.3 3. 추정 알고리즘: IRLS	65
9.1 실전 시나리오: 대학원 합격 예측 (로지스틱)	65
9.2 자주 묻는 질문 (FAQ)	66
10 Unit 4. 밀도 추정 (Density Estimation)	69
10.0.11. 히스토그램 (Histogram)	69
10.0.22. 커널 밀도 추정 (KDE, Kernel Density Estimation)	70
10.0.33. 편향-분산 트레이드오프 (Bias-Variance Trade-off)	70
10.0.44. 최적의 대역폭 선택 (Optimal Bandwidth)	71
10.1 실전 시나리오: 게임 유저 플레이 타임 분석	71
10.2 자주 묻는 질문 (FAQ)	72
11 Unit 5. 주성분 분석 (PCA)	75
11.0.11. 차원 축소의 철학: 분산은 정보다	75
11.0.22. 기하학적 해석: 최적의 좌표축 찾기	76
11.0.33. 수학적 엔진: 공분산 행렬과 고유값 분해	76
11.0.44. 차원 축소의 실행 (Dimension Reduction)	77
11.1 실전 시나리오: 넥슨 유저 세분화 (User Segmentation)	77
11.2 자주 묻는 질문 (FAQ)	78

12 Unit 12. 고차원 회귀와 희소성 (High-dimensional Regression & Sparsity)	81
12.0.11. 문제의 본질: $p > n$ 일 때 발생하는 일	81
12.0.22. 희소성 가정 (Sparsity Assumption)	82
12.0.33. 규제화 (Regularization): Ridge vs LASSO	82
12.0.44. LASSO의 기하학적 해석 (The Geometry of Sparsity)	83
12.0.55. 이론적 보장 (Theoretical Guarantees)	83
12.1실전 시나리오: 넥슨 게임 로그 분석	84
12.2자주 묻는 질문 (FAQ)	84

Course Structure & Current Focus

- Unit 0: Probability Review (Prerequisites)
- Unit 1: Modeling, Identifiability, Likelihood (현재 단원)
 - 1.1 Statistical Model
 - 1.2 Identifiability
 - 1.3 Likelihood Function
- Unit 2: Maximum Likelihood Estimation (MLE)
- Unit 3: Method of Moments
- Unit 4: Hypothesis Testing

Chapter 1

Unit 1. 통계적 모델링과 우도 (Modeling, Identifiability, Likelihood)

지난 시간(Unit 0)까지 우리는 확률론의 기초와 큰 수의 법칙(LLN), 중심극한정리(CLT)라는 무기를 손에 넣었습니다. 이제 이 무기를 가지고 현실 세계의 데이터를 분석하기 위해, 가장 먼저 해야 할 일은 **'데이터를 담을 수학적 그릇(모델)'**을 만드는 것입니다.

□ 개요 (Overview)

통계학을 제대로 하기 위해서는 현실의 불확실한 데이터를 수학적으로 엄밀하게 정의해야 합니다. 이 단원에서는 데이터를 수학적 집합으로 변환하는 **통계적 모델(Statistical Model)**의 정의, 모델의 파라미터가 유일한지 확인하는 **식별 가능성(Identifiability)**, 그리고 데이터를 통해 파라미터를 추정하기 위한 도구인 **우도(Likelihood)**의 개념을 배웁니다. 이는 추후 배울 '최대우도추정(MLE)'을 위한 필수적인 기초 작업입니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Sample Space (E)	데이터가 나올 수 있는 모든 후보들의 집합 (무대)
Parameter Space (Θ)	우리가 찾고 싶은 정답(θ)이 숨어있는 범위
Statistical Model	데이터가 발생할 수 있는 확률 규칙들의 모음집
Identifiability	범인의 지문이 유일한가? (파라미터 구별 가능 여부)
Likelihood (L_n)	이 데이터가 관측되었을 때, 범인이 θ 일 가능성

1.0.1 1. 통계적 모델 (Statistical Model)

[개념] 개념 1: 통계적 모델이란 무엇인가?

한 줄 요약: 현실의 막연한 데이터를 엄밀한 수학적 집합(Set)과 확률(Probability)로 번역하는 과정입니다.

6CHAPTER 1. UNIT 1. 통계적 모델링과 우도 (MODELING, IDENTIFIABILITY, LIKELIHOOD)

1) 직관적 비유: 맞춤 정장 만들기

현실의 데이터는 '사람의 몸'과 같습니다. 통계적 모델은 이 몸에 딱 맞는 '정장(수트)'을 만드는 과정입니다.

- **표본 공간 (E):** 옷감이 될 수 있는 재료들입니다. (면인가? 실크인가?)
- **파라미터 공간 (Θ):** 옷의 치수 조절 범위입니다. (허리 20~40인치 사이)
- **확률 분포 족 ($\{P_\theta\}$):** 치수(θ)에 따라 만들어지는 다양한 옷 디자인들입니다.

우리는 이 중에서 '내 몸(데이터)'에 가장 잘 맞는 옷(θ)을 찾고 싶은 것입니다.

2) 기술적 정의 및 수학적 구조

통계적 모델은 쌍(Pair) $(E, \{P_\theta\}_{\theta \in \Theta})$ 으로 정의됩니다.

1. **Sample Space (E):** 관측 데이터 X 가 취할 수 있는 값들의 집합.
2. **Parameter Space (Θ):** 미지수 θ 가 속한 집합. $\Theta \subseteq \mathbb{R}^k$.
3. **Probability Family (P_θ):** θ 값 하나가 정해지면, 그에 대응하는 확률 분포 하나가 결정됩니다.

3) 구체적 예시: 동전 던지기 (베르누이 모델)

우리가 앞면(1), 뒷면(0)이 나오는 동전을 던진다고 합시다.

- $E = \{0, 1\}$ (동전은 0 아니면 1만 나옵니다.)
- $\Theta = [0, 1]$ (앞면이 나올 확률 p 는 0과 1 사이입니다.)
- $\{P_\theta\}$ = 베르누이 분포 $\text{Ber}(p)$ 들의 집합.

[주의] 오해 방지: 잘 정의된 모델(Well-specified)이란?

"모델을 세운다"는 것은 가정을 한다는 뜻입니다. 만약 실제 데이터는 '주사위(1~6)'인데, 모델을 '동전(0,1)'으로 세우면 어떻게 될까요?

- 실제 데이터 생성 원리 P 가 우리 모델 집합 $\{P_\theta\}$ 안에 없을 때, 이를 **Misspecified Model**이라고 합니다.
- 반대로, 실제 자연의 법칙 P 가 우리 모델 안에 포함되어 있다면 **Well-specified Model**이라고 합니다.

—

1.0.2 2. 식별 가능성 (Identifiability)

[개념] 개념 2: 이 모델로 정답을 찾을 수 있는가?

한 줄 요약: 서로 다른 원인(파라미터)이 서로 다른 결과(분포)를 만들어야만, 결과를 보고 원인을 역추적할 수 있습니다.

1) 직관적 비유: 범인의 지문

범죄 현장에서 지문을 채취했습니다(데이터).

- **식별 가능:** 모든 사람의 지문이 다르다면, 지문을 보고 범인을 특정할 수 있습니다.
- **식별 불가능:** 만약 철수와 영희의 지문이 똑같이 생겼다면? 지문을 확보해도 둘 중 누가 범인인지 알 수 없습니다.

2) 기술적 정의

함수 $\theta \mapsto P_\theta$ 가 **단사함수(Injective)**여야 합니다. 즉,

$$\theta \neq \theta' \implies P_\theta \neq P_{\theta'}$$

만약 서로 다른 θ 가 같은 P_θ 를 만든다면, 데이터가 무한히 많아도 θ 를 추정할 수 없습니다.

3) 숫자 예시: 식별 불가능한 경우 (Non-identifiable)

어떤 데이터 X 가 평균이 $\mu_1 + \mu_2$ 인 정규분포를 따른다고 가정해 봅시다.

$$X \sim \mathcal{N}(\mu_1 + \mu_2, 1)$$

우리가 데이터에서 평균이 5라는 사실을 알아냈습니다.

- 경우 A: $\mu_1 = 2, \mu_2 = 3 \rightarrow \text{합 } 5$
- 경우 B: $\mu_1 = 1, \mu_2 = 4 \rightarrow \text{합 } 5$

μ_1 과 μ_2 의 조합은 무수히 많습니다. 데이터만으로는 절대 진짜 (μ_1, μ_2) 를 찾을 수 없습니다. 이것이 **식별 불가능**입니다.

—

1.0.3 3. 우도 (Likelihood)

[개념] 개념 3: 관점의 대전환

한 줄 요약: ”데이터가 주어졌을 때, 이 파라미터가 정답일 점수는 몇 점인가?”를 계산하는 함수입니다.

1) 직관적 비유: 명탐정 코난

- **확률(PDF):** 범인(θ)이 정해져 있을 때, 어떤 증거(x)를 남길지 예측하는 것. (미래 예측)
- **우도(Likelihood):** 증거(x)가 이미 확보되었을 때, 누가 범인(θ)일지 추리하는 것. (과거 추론)

8CHAPTER 1. UNIT 1. 통계적 모델링과 우도 (MODELING, IDENTIFIABILITY, LIKELIHOOD)

2) 기술적 정의: i.i.d와 결합 확률

데이터 X_1, \dots, X_n 이 서로 독립(Independent)이고 같은 분포(Identically Distributed)를 따른다고 가정합니다. 우도 함수 $L_n(\theta)$ 는 결합 확률 밀도 함수(Joint PDF)와 식은 같지만, **주인공(변수)**이 다릅니다.

$$L_n(\theta) = \prod_{i=1}^n f(X_i; \theta)$$

여기서 X_i 는 이미 관측된 고정값(상수)이고, θ 가 변수입니다.

3) 계산 예시: 불량품 찾기

공장에서 부품을 3개 뽑았는데 (양품, 양품, 불량)이 나왔습니다. 양품확률을 p 라고 합시다. (1=양품, 0=불량) 데이터: $X_1 = 1, X_2 = 1, X_3 = 0$.

- **가설 A ($p = 0.5$):** 공장이 반반 확률로 만듦.

$$L_3(0.5) = 0.5 \times 0.5 \times (1 - 0.5) = 0.125$$

- **가설 B ($p = 0.9$):** 공장이 90% 확률로 잘 만듦.

$$L_3(0.9) = 0.9 \times 0.9 \times (1 - 0.9) = 0.081$$

결과: 이 데이터(양,양,불) 기준으로는 가설 A($p = 0.5$)의 우도(0.125)가 더 높습니다. 즉, $p = 0.5$ 일 가능성이 더 높다고 추론할 수 있습니다.

1.1 로그 우도 (Log-Likelihood)와 실전 적용

왜 로그를 취하는가?

우도 L_n 은 확률의 곱셈(\prod)입니다. n 이 커지면 값이 0에 너무 가깝게 작아져서 컴퓨터가 계산하지 못합니다(Underflow). 따라서 로그를 취해 **덧셈(\sum)**으로 바꿉니다. 로그는 단조증가 함수이므로 최대값의 위치는 변하지 않습니다.

$$\ell_n(\theta) = \sum_{i=1}^n \log f(X_i; \theta)$$

실전 시나리오: 사용자 체류 시간 분석

당신이 앱 서비스 기획자라고 가정합시다. 사용자의 체류 시간(T)이 지수 분포 $f(t; \lambda) = \lambda e^{-\lambda t}$ 을 따른다고 모델링했습니다.

1. **데이터 수집:** 3명의 유저가 각각 2분, 3분, 5분을 머물렀습니다. ($x_1 = 2, x_2 = 3, x_3 = 5$)

2. **우도 함수 구성:**

$$L(\lambda) = (\lambda e^{-2\lambda}) \times (\lambda e^{-3\lambda}) \times (\lambda e^{-5\lambda}) = \lambda^3 e^{-10\lambda}$$

3. **로그 우도:**

$$\ell(\lambda) = \log(\lambda^3 e^{-10\lambda}) = 3 \log \lambda - 10\lambda$$

4. **최적화:** 이것을 미분해서 0이 되는 λ 를 찾으면, 그것이 바로 가장 그럴듯한(Likely) 파라미터입니다.

1.2 자주 묻는 질문 (FAQ)

Q1. 확률(Probability)과 우도(Likelihood)는 같은 거 아닌가요? A. 아니요, 정반대입니다!

- 확률: θ 가 고정 \rightarrow 데이터 x 가 변수. (적분하면 1)
- 우도: 데이터 x 가 고정 \rightarrow 파라미터 θ 가 변수. (적분해도 1이 아님)

Q2. 왜 식별 가능성(Identifiability)을 먼저 확인하나요? A. 식별 불가능한 모델을 가지고 우도 계산을 하는 것은, 답이 없는 문제를 열심히 푸는 것과 같습니다. 열심히 계산해서 우도가 최대인 지점을 찾아도, 그 θ 가 유일한 정답이라고 확신할 수 없기 때문입니다.

Next Step: 우리는 이제 '우도 함수'라는 강력한 점수판을 만들었습니다. 다음 단원인 Unit 2에서는 미분(Calculus)을 사용하여 이 우도 함수를 **최대화(Maximum)** 시키는 θ , 즉 **MLE(Maximum Likelihood Estimator)**를 구하는 법을 본격적으로 배웁니다.

[요약] Unit 1 핵심 요약

- **통계적 모델:** 데이터를 설명하기 위한 수학적 가정 ($E, \{P_\theta\}$).
- **식별 가능성:** 파라미터가 다르면 분포도 달라야 한다 ($\theta \neq \theta' \implies P_\theta \neq P_{\theta'}$). 이것이 보장되어야 추정이 가능하다.
- **우도(Likelihood):** 관측된 데이터 X 를 고정하고, 파라미터 θ 를 변화시키며 '가능성'을 측정하는 함수.
- **로그 우도:** 계산의 편의성과 미분을 위해 우도에 로그를 취한 형태 ($\sum \log f$).

(a4paper, 11pt)book fontspec amsmath, amssymb, amsthm geometry graphicx xcolor
hyperref booktabs array
left=25mm, right=25mm, top=30mm, bottom=30mm

10CHAPTER 1. UNIT 1. 통계적 모델링과 우도 (MODELING, IDENTIFIABILITY, LIKELIHOOD)

Contents

Course Structure & Current Focus

- Unit 1: Modeling, Identifiability, Likelihood (완료)
- Unit 2: Point Estimation (현재 단원)
 - 2.1 Estimator: Definition & Intuition
 - 2.2 Maximum Likelihood Estimation (MLE)
 - 2.3 Method of Moments (MoM)
 - 2.4 Properties: Bias, Variance, MSE
- Unit 3: Confidence Intervals (구간 추정)
- Unit 4: Hypothesis Testing (가설 검정)

Chapter 2

Unit 2. 점 추정 (Point Estimation)

Unit 1에서 우리는 '우도(Likelihood)'라는 강력한 도구를 만들었습니다. 이제 질문을 던질 차례입니다. "그래서 그 우도를 가장 높게 만드는 범인(θ)은 정확히 누구인가?" 이번 단원에서는 미지의 파라미터를 하나의 숫자(Point)로 찍어 맞추는 방법과, 그 방법의 성적을 매기는 기준을 배웁니다.

□ 개요 (Overview)

점 추정은 데이터를 함수에 넣어 미지의 파라미터 θ 에 대한 최적의 추측값 $\hat{\theta}$ 를 계산하는 과정입니다. 우리는 두 가지 주요 방법론인 **MLE(최적화 접근)**와 **MoM(대수의 법칙 접근)**을 배우고, 추정량이 좋은지 나쁜지를 판별하는 **MSE(Bias-Variance Trade-off)** 프레임워크를 익힙니다. 이것은 머신러닝의 손실 함수(Loss Function) 개념의 기초가 됩니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Estimator ($\hat{\theta}_n$)	데이터를 입력받아 추정값을 뱉어내는 '함수' (공식 그 자체)
Estimate	실제 데이터를 넣어 계산된 구체적인 '숫자'
Bias (편향)	영점 조절 상태. 평균적으로 정답을 맞히는가?
Variance (분산)	정밀도. 데이터가 조금 바뀔 때 추정값이 얼마나 흔들리는가?
MSE	종합 점수. 편향의 제곱과 분산을 합친 총 에러.

—

2.0.1 1. 추정량(Estimator)의 정의와 본질

[개념] 개념 1: 추정량은 숫자가 아니라 '함수'이자 '확률변수'다

한 줄 요약: 추정량은 데이터가 들어오면 정답을 내놓는 '기계'이며, 들어오는 데이터가 랜덤하므로 기계가 내놓는 답도 랜덤하게 변합니다.

1) 직관적 비유: 여론조사 기관

대통령 지지율(θ)을 알고 싶습니다.

- **추정량($\hat{\theta}$):** "지나가는 사람 100명에게 물어보고 찬성 비율을 계산한다"라는 **규칙**입니다.

- **확률변수로서의 성격:** 오늘 100명을 조사했을 때와, 내일 100명을 조사했을 때 결과는 다를 것입니다. 즉, $\hat{\theta}$ 는 고정된 값이 아니라 **분포(Distribution)**을 가집니다.

2) 기술적 정의

추정량 $\hat{\theta}_n$ 은 표본 X_1, \dots, X_n 의 함수입니다.

$$\hat{\theta}_n = g(X_1, \dots, X_n)$$

중요 조건: 이 함수 g 안에는 미지의 파라미터 θ 가 들어 있으면 안 됩니다. (우리가 모르는 값을 이용해 계산할 수는 없으니까요.)

2.0.2 2. 최대우도추정 (MLE, Maximum Likelihood Estimation)

(개념) 개념 2: 가장 그럴듯한 범인 찾기

한 줄 요약: “현재의 데이터가 관찰될 확률을 수학적으로 가장 높여주는 θ 값을 정답으로 채택하자.”

1) 직관적 비유: 등산하기

안개 낀 산(로그 우도 함수)에서 가장 높은 봉우리(최대값)를 찾아가는 과정입니다.

- 산의 높이: $\ell_n(\theta)$ (로그 우도)
- 전략: 기울기(미분값)가 0이 되는 지점을 찾는다.

2) 수학적 절차 (Algorithm)

- **Log-Likelihood:** $\ell_n(\theta) = \sum_{i=1}^n \log f(X_i; \theta)$ 를 구합니다.
- **Differentiate:** θ 에 대해 미분합니다. $\frac{\partial}{\partial \theta} \ell_n(\theta)$.
- **Solve:** 미분값이 0이 되는 방정식(Estimating Equation)을 풁니다.
- **Check:** 두 번 미분하여 음수인지(위로 볼록, Concave) 확인하여 최대값임을 보장합니다.

3) 계산 예시: 포아송 분포 (웹사이트 방문자 수)

하루 방문자 수 X 가 포아송 분포 $\text{Pois}(\lambda)$ 를 따른다고 가정합시다. ($\theta = \lambda$)

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

데이터: X_1, \dots, X_n

- 로그 우도: $\ell_n(\lambda) = \sum(-\lambda + X_i \ln \lambda - \ln X_i!)$ = $-n\lambda + (\sum X_i) \ln \lambda - C$
- 미분: $\frac{\partial}{\partial \lambda} \ell_n(\lambda) = -n + \frac{\sum X_i}{\lambda}$
- 0으로 놓기: $-n + \frac{\sum X_i}{\lambda} = 0 \implies \hat{\lambda}_{MLE} = \frac{1}{n} \sum X_i = \bar{X}$

결론: 포아송 분포의 MLE는 직관적이게도 **표본 평균**입니다.

2.0.3 3. 적률법 (MoM, Method of Moments)

(개념) 개념 3: 평균을 평균에 맞춘다

한 줄 요약: ”이론적인 평균(Population Mean)이 실제 데이터의 평균(Sample Mean)과 같아야 한다”는 단순한 믿음에서 출발합니다.

1) 직관적 비유: 요리 간 맞추기

국물 맛(데이터)을 봅니다.

- 이론: 소금(θ)을 1스푼 넣으면 짠맛 농도($E[X]$)가 10이 되어야 한다.
- 실제: 국물을 떠서 맛보니 짠맛 농도(\bar{X})가 20이다.
- 추론: ”아, 소금이 2스푼 들어갔겠구나($\hat{\theta} = 2$).”

2) 수학적 정의

k 번째 이론적 적률(Moment) $m_k(\theta) = \mathbb{E}[X^k]$ 을 구하고, 이를 표본 적률 $\frac{1}{n} \sum X_i^k$ 과 같다고 합니다.

$$\mathbb{E}[X] = \bar{X} \quad (1\text{차 모멘트 매칭})$$

미지수가 2개면 2차 모멘트($E[X^2]$)까지 사용합니다.

(주의) MLE vs MoM: 누가 더 좋은가?

일반적으로 **MLE가 더 정밀(Efficient)**합니다. 하지만 MLE는 계산이 복잡하거나 미분이 불가능할 수 있습니다. 이때 계산이 쉬운 **MoM**을 먼저 구해서 MLE를 찾기 위한 **초기값(Initial Guess)**으로 사용하는 경우가 많습니다.

2.0.4 4. 추정량의 성질 (Properties): 성적표 매기기

(개념) 개념 4: Bias-Variance Trade-off

한 줄 요약: 좋은 추정량은 영점이 잘 잡혀있어야 하고(Low Bias), 쓸 때마다 탄착군이 좁게 모여야 합니다(Low Variance).

1) 편향 (Bias): 영점 조절

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

추정량을 무한히 반복했을 때, 그 **평균**이 과연 정답(θ)과 일치하는가?

- $Bias = 0$: 비편향 추정량 (Unbiased). 영점이 정확함.
- $Bias \neq 0$: 편향 추정량 (Biased). 영점이 틀어짐.

2) 분산 (Variance): 탄착군 크기

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

데이터가 바뀔 때마다 추정값이 얼마나 들쭉날쭉한가? (안정성)

3) MSE (Mean Squared Error): 최종 점수

우리의 목표는 에러의 총합을 줄이는 것입니다.

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \underbrace{\text{Bias}(\hat{\theta})^2}_{\text{영점 오차}} + \underbrace{\text{Var}(\hat{\theta})}_{\text{흔들림}}$$

인사이트: 때로는 영점이 조금 틀어지더라도(Bias 존재), 탄착군을 획기적으로 좁힐 수 있다면(Variance 감소), 그게 더 좋은 추정량일 수 있습니다. (이것이 현대 머신러닝의 핵심 철학입니다.)

2.1 실전 시나리오: 모바일 게임 가챠 확률 검증

[시나리오] Scenario: 유저들의 불만

당신은 게임 회사 넥슨의 PM입니다. ”전설 아이템 획득 확률(p)이 1%라고 했는데, 너무 안 나온다!”라는 유저 불만이 폭주합니다. 데이터를 확인해보니, 100명의 유저가 가챠를 돌렸고 그중 딱 1명만 성공했습니다. ($n = 100, \sum x_i = 1$)

1. **모델링:** 성공/실패이므로 베르누이 분포 $X \sim \text{Ber}(p)$ 입니다.
2. **MLE 적용:** 베르누이의 MLE는 표본 평균입니다.

$$\hat{p}_{MLE} = \frac{1}{100} = 0.01 (1\%)$$

회사 측 입장: ”데이터로 추정해 보니 1%가 맞습니다. 시스템은 정상입니다.”

3. **Bayesian 관점 (심화 예고):** 하지만 만약 유저가 3명만 돌려서 0명이 나왔다면? $\hat{p} = 0\%$. 확률이 0이라고 단정할 수 있을까요? 이때는 ”과거의 경험(Prior)”을 섞는 베이지안 방식이 필요할 수 있습니다.
-

2.2 자주 묻는 질문 (FAQ)

Q1. 비편향(Unbiased) 추정량이 무조건 좋은 건가요? A. 아닙니다! 비편향이지만 분산이 태평양만큼 넓다면 쓸모가 없습니다. 약간의 편향을 감수하더라도 분산이 매우 작은 추정량(예: Ridge Regression)을 선택하는 경우가 많습니다. 결국 **MSE(총 에러)**가 작은 것이 장땡입니다.

Q2. MLE는 항상 정답을 주나요? A. 데이터가 무한히 많다면($n \rightarrow \infty$) MLE는 참값으로 수렴합니다(Consistent). 하지만 데이터가 적을 때(n 이 작을 때)는 MLE도 틀릴 수 있고, 심지어 편향(Biased)되어 있을 수도 있습니다. (예: 정규분포 분산의 MLE는 편향되어 있음).

Next Step: 우리는 점 추정을 통해 $\hat{\theta} = 0.01$ 이라는 숫자를 얻었습니다. 하지만 이 숫자가 **얼마나 확실한지**는 아직 모릅니다. 0.01이라곤 했지만, 실제로는 0.005일 수도 있고 0.015일 수도 있지 않을까요? 다음 Unit 3에서는 이 불확실성을 감안하여 정답이 있을 만한 **구간(Interval)**을 구하는 법을 배웁니다.

(요약) Unit 2 핵심 요약

- **추정량(Estimator):** 데이터의 함수이자 확률변수. 분포를 가진다.
- **MLE:** 우도(Likelihood)를 최대화하는 값을 찾는다. (미분 $\nabla\ell = 0$)
- **MoM:** 표본 평균을 이론적 평균과 같다고 둔다. (계산이 쉬움)
- **Bias-Variance Trade-off:** $MSE = Bias^2 + Var.$ 편향과 분산의 균형을 맞춰 전체 에러를 줄이는 것이 목표다.

(a4paper, 11pt)book fonts spec amsmath, amssymb, amsthm geometry graphicx xcolor
hyperref booktabs array
left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 1: Modeling (무대 세팅)
- Unit 2: Point Estimation (범인 지목)
- Unit 3: Asymptotic Properties (현재 단원: 무한대로 확장)
 - 3.1 Consistency (일치성)
 - 3.2 Asymptotic Normality (점근적 정규성)
 - 3.3 Fisher Information & CRLB (추정의 한계)
 - 3.4 Delta Method (함수의 추정)
- Unit 4: Hypothesis Testing (가설 검정)

Chapter 3

Unit 3. 점근적 성질 (Asymptotic Properties)

Unit 2에서 우리는 $\hat{\theta}$ 라는 '추정량'을 만들었습니다. 하지만 데이터 개수(n)가 적을 때는 이 추정량이 얼마나 정확한지 계산하기 어렵습니다(분산 공식이 복잡함). 그래서 우리는 질문을 바꿉니다. "만약 데이터가 무한히 많아진다면($n \rightarrow \infty$), 이 추정량은 결국 정답을 맞히게 될까?" 이것이 통계적 추론의 이론적 보증수표가 됩니다.

□ 개요 (Overview)

이 단원에서는 데이터의 크기 n 이 커질 때 추정량이 가지는 극한의 성질을 다룹니다. 추정량이 참값으로 수렴하는지(**Consistency**), 오차의 분포가 정규분포를 따르는지(**Asymptotic Normality**) 확인하고, 추정 정밀도의 절대적 한계선(**CRLB**)을 수학적으로 규명합니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Consistency (일치성)	데이터를 많이 모으면, 결국 정답을 찾아내는가?
Asymptotic Normality	오차의 모양이 종 모양(정규분포)으로 예쁘게 모이는가?
Fisher Information (I)	데이터 하나가 파라미터에 대해 주는 힌트의 양 (곡률).
CRLB	신(God)도 이보다 더 정확하게 추정할 순 없다 (분산의 하한선).
Delta Method	$\hat{\theta}$ 의 분포를 알 때, $f(\hat{\theta})$ 의 분포를 근사하는 기술.

—

3.0.1 1. 일치성 (Consistency)

(개념) 개념 1: 좋은 추정량의 최소 자격 요건

한 줄 요약: 데이터가 쌓일수록 추정값의 오차가 0으로 줄어들어, 결국 참값과 같아져야 합니다.

1) 직관적 비유: 디지털 사진의 해상도

- $n \rightarrow \infty$ 작을 때: 저화질 픽셀 사진. 무엇인지 흐릿하게 보입니다.

- $n \rightarrow \infty$: 고화질 4K 사진. 실제 피사체(참값 θ^*)와 완전히 똑같이 보입니다.
- 만약 데이터를 무한히 모았는데도 사진이 흐릿하다면? 그 카메라는 고장난 것입니다(Inconsistent).

2) 기술적 정의: 확률적 수렴

추정량 $\hat{\theta}_n$ 이 참값 θ^* 로 **확률** 수렴(Converge in Probability) **해야 합니다.

$$\forall \epsilon > 0, \lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta^*| > \epsilon) = 0$$

의미: 오차가 ϵ 보다 클 가능성이, 데이터가 늘어날수록 0이 된다. (주로 **대수의 법칙(LLN)**에 의해 보장됩니다.)

3.0.2 2. 점근적 정규성 (Asymptotic Normality)

(개념) 개념 2: 모든 길은 정규분포로 통한다

한 줄 요약: 원래 데이터 분포가 무엇이든 상관없이, "추정 오차"의 분포는 종 모양(Bell Curve)을 따릅니다.

1) 직관적 비유: 돋보기로 확대하기

n 이 커지면 $\hat{\theta}_n$ 은 θ^* 라는 한 점으로 뭉칩니다(분산 $\rightarrow 0$). 분포를 보려면 이 오차를 **확대**해야 합니다.

- 그냥 오차: $\hat{\theta}_n - \theta^* \rightarrow 0$ (너무 작아서 안 보임)
- \sqrt{n} 배 확대: $\sqrt{n}(\hat{\theta}_n - \theta^*)$ (이제 모양이 보임 \rightarrow 정규분포!)

2) 기술적 정의

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma_{asym}^2)$$

여기서 σ_{asym}^2 를 **점근 분산(Asymptotic Variance)**이라고 합니다. 이것을 알아야 신뢰구간을 만들 수 있습니다.

3.0.3 3. 피셔 정보와 크래머-라오 하한 (CRLB)

(개념) 개념 3: 추정 정밀도의 물리적 한계선

한 줄 요약: 정보가 많을수록(곡선이 뾰족할수록) 분산은 작아지며, 그 한계는 피셔 정보의 역수입니다.

1) 직관적 비유: 산봉우리의 뾰족함

로그 우도 함수(Likelihood)를 산이라고 생각합시다.

- **뾰족한 산 (High Curvature):** 정상(최대값)의 위치가 아주 명확합니다. \rightarrow **정보 많음, 분산 작음.**
- **평평한 산 (Low Curvature):** 어디가 정상인지 헷갈립니다. \rightarrow **정보 적음, 분산 큼.**

2) 공식 및 계산 예시 (동전 던지기)

파라미터 θ 에 대해 우도 함수의 '볼록한 정도(2계 도함수)'가 정보량입니다.

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right]$$

예시: 베르누이 분포 ($f(x; p) = p^x(1-p)^{1-x}$)

1. 로그 우도: $\log f = x \log p + (1-x) \log(1-p)$

2. 1차 미분: $\frac{x}{p} - \frac{1-x}{1-p}$

3. 2차 미분: $-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$

4. 기댓값 ($E[X] = p$): $I(p) = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$

결과 (CRLB): 어떤 비편향 추정량도 분산이 $\frac{1}{nI(p)} = \frac{p(1-p)}{n}$ 보다 작을 순 없습니다. (참고: 표본평균 \bar{X} 의 분산이 정확히 이 값입니다. 즉, \bar{X} 는 가장 효율적인 추정량입니다!)

3.0.4 4. 델타 방법 (Delta Method)

[개념] 개념 4: 함수를 통과한 추정량의 분포

한 줄 요약: $\hat{\theta}$ 가 정규분포라면, $g(\hat{\theta})$ 도 (근사적으로) 정규분포이며, 분산은 '기울기 제곱'만큼 변합니다.

1) 직관적 비유: 지구는 둥글지만 운동장은 평평하다

지구($g(\theta)$ 곡선)는 둥글지만, 우리가 서 있는 좁은 공간에서는 평평한 직선처럼 보입니다. 이 성질(선형 근사)을 이용해 분산을 예측합니다.

2) 공식

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 \cdot \sigma^2)$$

원래 분산 σ^2 에 **변환 함수의 미분값의 제곱($[g'(\theta)]^2$)**을 곱해주면 됩니다.

3.1 실전 시나리오: 모바일 게임 리텐션(재접속률) 예측

[시나리오] Scenario: 대규모 업데이트의 성과

넥슨에서 대규모 업데이트를 진행했습니다. 전체 유저의 재접속률 p 를 추정하려 합니다.

1. **데이터 수집:** $n = 10,000$ 명의 유저 로그를 분석했더니 $\hat{p} = 0.6$ (60%)이 나왔습니다.

2. **목표:** 우리는 단순히 p 가 아니라, **오즈(Odds) $\frac{p}{1-p}$ ** (접속할 확률이 접속 안 할 확률의 몇 배인가?)에 관심이 있습니다.

3. **문제:** \hat{p} 의 분산은 알겠는데, 오즈 $\frac{\hat{p}}{1-\hat{p}}$ 의 신뢰구간은 어떻게 구하죠?
4. **해결 (델타 방법):***
 - 변환 함수: $g(p) = \frac{p}{1-p}$
 - 미분: $g'(p) = \frac{1}{(1-p)^2}$
 - 새로운 분산: 원래 분산 $\frac{p(1-p)}{n}$ 에 $[g'(p)]^2$ 를 곱함.
 - 결론: 복잡한 시뮬레이션 없이도, 미분 한 번으로 오즈 값의 오차 범위를 즉시 계산할 수 있습니다.

3.2 자주 묻는 질문 (FAQ)

- Q1. 데이터가 적으면(n 이 작으면) 이 이론들은 쓸모 없나요? A. 완전히 쓸모없는 건 아닙니다. 통계학에서는 보통 $n \geq 30$ 정도면 중심극한정리가 작동한다고 봅니다. 하지만 n 이 작을 때는 정규분포 근사가 부정확할 수 있으므로, t-분포를 쓰거나 부트스트랩(Bootstrap) 같은 시뮬레이션 방법을 쓰는 것이 안전합니다.
- Q2. 왜 하필 \sqrt{n} 을 곱하나요? 그냥 n 을 곱하면 안 되나요? A. 좋은 질문입니다!

- 오차($\hat{\theta} - \theta$)는 대략 $\frac{1}{\sqrt{n}}$ 의 속도로 줄어듭니다.
- 그냥 두면 0이 되고, n 을 곱하면 발산(∞)해 버립니다.
- 딱 균형을 맞춰서 안정적인 분포(정규분포)를 만들기 위해 역수인 \sqrt{n} 을 곱하는 것입니다.

Next Step: 이제 우리는 추정량 $\hat{\theta}$ 가 정규분포를 따른다는 강력한 무기를 얻었습니다. 그렇다면 ”이 추정값이 0.5라고 주장하는 것이 타당한가?”를 판단할 수 있겠죠? 다음 **Unit 4**에서는 이 분포를 이용해 가설을 검증하는 **가설 검정(Hypothesis Testing)**으로 넘어갑니다.

(요약) Unit 3 핵심 요약

- **Consistency:** $n \rightarrow \infty$ 이면 추정량은 참값이 된다.
- **Asymptotic Normality:** 추정 오차는 \sqrt{n} 스케일에서 정규분포 $\mathcal{N}(0, I(\theta)^{-1})$ 를 따른다.
- **Fisher Info & CRLB:** 분산의 하한선은 정보량의 역수다. MLE는 이 한계에 도달하는 최적의 추정량(Efficient Estimator)이다.
- **Delta Method:** 변환된 추정량 $g(\hat{\theta})$ 의 분산은 $g'(\theta)^2 \times \text{Var}(\hat{\theta})$ 이다.

[a4paper, 11pt] book fontspec amsmath, amssymb, amsthm geometry graphicx xcolor hyperref booktabs array
left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 2: Point Estimation (범인 지목)
- Unit 3: Asymptotic Properties (이론적 무기 확보)
- Unit 4: Confidence Intervals (현재 단원: 수사망 펼치기)
 - 4.1 Concept & Philosophy (빈도주의적 해석)
 - 4.2 Construction with MLE (구간 만드는 법)
 - 4.3 Conservative CI (보수적 접근)
- Unit 5: Hypothesis Testing (최종 판결)

Chapter 4

Unit 4. 신뢰 구간 (Confidence Intervals)

Unit 2에서 우리는 $\hat{\theta}$ 라는 하나의 값(Point)을 구했고, Unit 3에서는 데이터가 많을수록 이 값이 정규분포를 그리며 참값에 다가간다는 사실을 증명했습니다. 이제 이 정규분포라는 지도를 펼쳐놓고, ”그래서 참값이 어디부터 어디 사이에 있는데?”라는 질문에 답할 차례입니다.

□ 개요 (Overview)

신뢰 구간은 점 추정값의 **'오차 범위(Margin of Error)'**를 수학적으로 계산하는 도구입니다. 이 단원에서는 신뢰 구간의 올바른 해석(철학), 중심극한정리를 이용한 구간 구성법(레시피), 그리고 미지의 파라미터를 처리하는 수학적 기법(Slutsky's Theorem)을 배웁니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Confidence Level ($1 - \alpha$)	신뢰 수준. 보통 95%(0.95)를 사용합니다.
Significance Level (α)	유의 수준. 틀릴 확률의 허용치 (보통 5%).
Quantile ($q_{\alpha/2}$)	정규분포에서 꼬리를 자르는 기준점 (예: 1.96).
Pivot	분포가 θ 에 의존하지 않는 통계량 (예: Z-score).
Slutsky's Theorem	n 이 클 땐, 모르는 참값 대신 추정값을 써도 된다는 허가증.

4.0.1 1. 신뢰 구간의 정의와 철학

(개념) 개념 1: 움직이는 것은 '고리(구간)'이지 '기둥(참값)'이 아니다

한 줄 요약: 참값 θ 는 신만이 아는 고정된 상수입니다. 우리가 데이터를 뽑을 때마다 변하는 것은 신뢰 구간입니다.

1) 직관적 비유: 고리 던지기 (Ring Toss)

- **기둥 (θ):** 바닥에 박혀 있습니다. 절대 움직이지 않습니다.
- **고리 (Interval):** 사람이 던집니다(데이터 수집). 던질 때마다 고리의 위치가 바뀝니다.
- **신뢰 수준 95%:** ”내가 고리를 100번 던지면, 그중 95번은 기둥에 걸린다”는 뜻입니다.

- **주의:** 이미 바닥에 떨어진 고리(계산된 구간)를 보고 ”기둥이 이 안에 들어올 확률 95%”라고 말하면 안 됩니다. 기둥은 들어와 있거나(1), 안 들어와 있거나(0) 둘 중 하나입니다.

[주의] 오개념 주의: Frequentist View

”이 구간 [0.4, 0.6] 안에 참값이 있을 확률은 95%다” → 틀렸습니다!

”수많은 평행우주에서 실험을 반복했을 때, 만들어진 구간들의 95%가 참값을 포함한다” → 맞습니다.

—

4.0.2 2. 점근적 신뢰 구간 구성법 (The Recipe)

(개념) 개념 2: 정규분포를 역이용하여 구간 만들기

한 줄 요약: $\hat{\theta}$ 가 정규분포를 따른다는 사실(Unit 3)을 이용해, 거꾸로 θ 의 범위를 추적합니다.

Step 1: 정규성 확보 (From Unit 3)

MLE $\hat{\theta}_n$ 은 점근적으로 정규분포를 따릅니다.

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right)$$

Step 2: 피벗(Pivot) 구성

위 식을 표준정규분포 $Z \sim \mathcal{N}(0, 1)$ 형태로 만듭니다.

$$\sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \approx Z$$

Step 3: 부등식 세우기

표준정규분포의 95%가 들어있는 구간은 $[-1.96, 1.96]$ 입니다. ($q_{0.025} \approx 1.96$)

$$P\left(-q_{\alpha/2} \leq \sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \leq q_{\alpha/2}\right) \approx 1 - \alpha$$

Step 4: 슬루츠키 정리 (Slutsky's Theorem) - 결정적 단계

위 식의 $I(\theta)$ 에는 여전히 모르는 값 θ 가 들어있습니다. 계산을 위해 ** $I(\theta)$ 를 $I(\hat{\theta}_n)$ 으로 바꿔치기(Plug-in)** 합니다.

- **논리:** 일치성(Consistency)에 의해 $\hat{\theta}_n \rightarrow \theta$ 이므로, $I(\hat{\theta}_n) \rightarrow I(\theta)$ 입니다.
- **결과:** n 이 충분히 크면 이 바꿔치기는 수학적으로 정당합니다.

Step 5: 최종 공식 (θ 에 대해 정리)

$$\mathcal{I} = \left[\hat{\theta}_n - \frac{q_{\alpha/2}}{\sqrt{nI(\hat{\theta}_n)}}, \quad \hat{\theta}_n + \frac{q_{\alpha/2}}{\sqrt{nI(\hat{\theta}_n)}} \right]$$

- **중심:** 점 추정값 $\hat{\theta}_n$
 - **폭(Width):** $\frac{1}{\sqrt{n}}$ (데이터 많으면 좁아짐) $\times \frac{1}{\sqrt{I}}$ (정보 많으면 좁아짐)
-

4.0.3 3. 보수적 신뢰 구간 (Conservative CI)

(개념) 개념 3: 잘 모를 땐 최악의 상황을 가정하라

한 줄 요약: $I(\theta)$ 를 정확히 계산하기 어렵거나 데이터가 적을 때, 분산이 가장 커지는 경우(Worst Case)를 대입하여 넓은 구간을 잡습니다.

예시: 선거 여론조사 (베르누이 분포)

지지율 p 에 대한 신뢰 구간을 구할 때, 분산 항은 $p(1-p)$ 입니다.

- **문제:** 아직 p 를 모릅니다. \hat{p} 를 대입하자니 오차가 걱정됩니다.
- **해결:** $p(1-p)$ 는 $p = 0.5$ 일 때 최댓값 0.25를 가집니다.
- **보수적 구간:** 그냥 무조건 0.25를 대입합니다.

$$\hat{p} \pm 1.96 \frac{\sqrt{0.25}}{\sqrt{n}} = \hat{p} \pm \frac{0.98}{\sqrt{n}} \approx \hat{p} \pm \frac{1}{\sqrt{n}}$$

- **실전 팁:** 언론에서 흔히 말하는 "표본오차 $\pm 3.1\%$ " 같은 표현이 이 보수적 구간(주로 $n = 1000$ 일 때 $\approx 3.1\%$)을 사용한 것입니다.
-

4.1 실전 시나리오: A/B 테스트 신뢰 구간

(시나리오) Scenario: 웹사이트 배너 클릭률 비교

당신은 e커머스 앱의 PM입니다. 기존 배너(A)와 신규 배너(B)의 클릭률(CTR) 차이를 분석합니다.

1. **데이터:**

- A안: 1000명 중 50명 클릭 ($\hat{p}_A = 0.05$)
- B안: 1000명 중 70명 클릭 ($\hat{p}_B = 0.07$)
- 차이: $\Delta = 0.02$ (2% 상승)

2. **질문:** 이 2% 차이가 진짜 실력 차이일까요, 아니면 우연일까요?

3. **신뢰 구간 계산 (95%):** 두 비율 차이의 표준오차(SE) $\approx \sqrt{\frac{p(1-p)}{n} + \frac{p(1-p)}{n}}$

$$SE \approx \sqrt{\frac{0.05 \times 0.95}{1000} + \frac{0.07 \times 0.93}{1000}} \approx 0.011$$

$$95\% CI = [0.02 - 1.96(0.011), \quad 0.02 + 1.96(0.011)] = [-0.001, 0.041]$$

4. **해석:** 구간 $[-0.1\%, 4.1\%]$ 사이에 **0이 포함**되어 있습니다. 즉, "차이가 0일 수도 있다(효과가 없다)"는 가능성을 배제할 수 없습니다. 통계적으로 유의미한 차이가 아닙니다.

4.2 자주 묻는 질문 (FAQ)

Q1. 신뢰 구간을 좁히려면 어떻게 해야 하나요? A. 두 가지 방법이 있습니다.

1. 데이터를 더 많이 모읍니다 (n 증가 \rightarrow 분모 커짐 \rightarrow 폭 감소). 가장 확실한 방법입니다.
2. 신뢰 수준을 낮춥니다 (99% \rightarrow 90%). 하지만 틀릴 위험(Risk)이 커집니다.

Q2. Slutsky 정리는 왜 중요한가요? A. 이론과 현실을 이어주는 다리이기 때문입니다. 이론적으로 분산 공식에는 참값 θ 가 들어가야 하지만, 현실에선 θ 를 모릅니다. Slutsky 정리가 "데이터가 많으면 추정값 $\hat{\theta}$ 를 대신 써도 괜찮아!"라고 수학적으로 허락해주기 때문에 우리가 실제로 숫자를 대입해서 계산할 수 있는 것입니다.

Next Step: 신뢰 구간을 구했더니 $[0.4, 0.6]$ 이 나왔습니다. 누군가 "평균이 0.5입니까?"라고 묻는다면 "그럴싸하다"고 답할 수 있겠죠. 하지만 "평균이 0.7입니까?"라고 묻는다면 "아니오"라고 할 것입니다. 이처럼 신뢰 구간은 주장의 옳고 그름을 판별하는 도구가 됩니다. 다음 **Unit 5**에서는 이를 공식화한 **가설 검정(Hypothesis Testing)**을 배웁니다.

[요약] Unit 4 핵심 요약

- **철학:** 신뢰 구간은 랜덤한 구간이며, 참값 θ 는 고정되어 있다. "95% 확률로 포함한다"는 것은 반복 실험 시의 성공률을 의미한다.
- **구성법:** $\hat{\theta} \pm 1.96 \times \text{Standard Error}$.
- **Slutsky's Theorem:** 표준오차(Standard Error) 계산 시 미지의 θ 대신 $\hat{\theta}$ 를 대입할 수 있게 해주는 정리.
- **보수적 구간:** 분산의 최댓값을 사용하여 계산. 안정적이지만 구간이 넓다.

(a4paper, 11pt)book fontspec amsmath, amssymb, amsthm geometry graphicx xcolor
hyperref booktabs array
left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 1: Modeling (무대 세팅)
- Unit 2 (Part A): Point Estimation (범인 추정)
- Unit 2 (Part B): Hypothesis Testing (현재 단원: 판결 내리기)
 - 2.5 Structure: H_0 vs H_1
 - 2.6 Two Types of Errors
 - 2.7 Level(α) and Power($1 - \beta$)
 - 2.8 p-value Interpretation
- Unit 3: Asymptotic Properties

Chapter 5

Unit 2 (Part B). 가설 검정 (Hypothesis Testing)

지난 파트(Estimation)에서 우리는 데이터를 통해 "성공 확률이 60%($\hat{p} = 0.6$)일 것이다"라고 추측했습니다. 하지만 누군가 땀지를 겁니다. "에이, 원래 50%인데 우연히 높게 나온 거 아냐?" 이 질문에 답하기 위해 우리는 단순히 값을 구하는 것을 넘어, **'Yes or No'로 결정을 내리는 체계**가 필요합니다.

□ 개요 (Overview)

가설 검정은 불확실한 데이터 속에서 두 가지 주장(H_0, H_1) 중 하나를 선택하는 과정입니다. 이 단원에서는 가설 검정을 **"제1종 오류(α)를 통제하면서 제2종 오류를 최소화(Power 최대화)하는 수학적 최적화 문제"**로 정의하고, 그 결과물인 p-value의 진정한 의미를 배웁니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Null Hypothesis (H_0)	귀무가설. "차이가 없다", "효과가 없다". (피고인은 무죄)
Alternative Hypothesis (H_1)	대립가설. "차이가 있다", "효과가 있다". (피고인은 유죄)
Type I Error (α)	멀쩡한 사람을 범인으로 잡음. (거짓 양성)
Type II Error (β)	진짜 범인을 놓아줌. (거짓 음성)
Power ($1 - \beta$)	진짜 범인을 잡아낼 확률. (검정력)
p-value	피고인이 무죄(H_0)라고 쳤을 때, 이런 증거가 나올 희박함의 정도.

—

5.0.1 1. 가설 검정의 구조: 비대칭적 싸움

(개념) 개념 1: 법정 공방 (Courtroom Trial)

한 줄 요약: 가설 검정은 두 가설이 대등하게 싸우는 것이 아니라, "무죄 추정의 원칙" 하에 유죄의 증거를 찾는 과정입니다.

1) 직관적 비유

- **검사 (H_1):** "이 약은 효과가 있습니다!" (입증하고 싶은 것)
- **변호사 (H_0):** "아닙니다. 그냥 물(Placebo)과 똑같습니다." (기본 상태)
- **판사 (ψ):** 증거(데이터)가 압도적으로 확실하지 않으면, 무죄(H_0)를 선고합니다. 즉, H_0 를 기각하기(Reject) 전까지는 H_0 가 참이라고 가정합니다.

2) 수학적 정의

- 파라미터 공간 Θ 를 두 개로 쪼갭니다. Θ_0 (귀무가설 영역) vs Θ_1 (대립가설 영역).
- **검정 함수 (Test Function) $\psi(X)$:

$$\psi(X) = \begin{cases} 1 & (H_0 \text{ 기각}, H_1 \text{ 채택}) \\ 0 & (H_0 \text{ 기각 실패}, H_0 \text{ 유지}) \end{cases}$$

—

5.0.2 2. 두 가지 종류의 오류 (Type I & Type II Error)

(개념) 개념 2: 억울한 누명 vs 범인 놓침

한 줄 요약: 우리는 신이 아니기에 오류를 피할 수 없습니다. 어떤 오류가 더 치명적인지 파악해야 합니다.

실제 진실 \ 우리의 결정		H_0 유지 (무죄 선고)	H_0 기각 (유죄 선고)
H_0 참 (무죄)		옳은 결정 (√)	Type I Error (α) (억울한 옥살이)
H_1 참 (유죄)		Type II Error (β) (범인 도주)	옳은 결정 ($1 - \beta$) (정의 구현 = Power)

과학계의 관점

과학계는 **Type I Error(거짓 발견)**를 훨씬 심각하게 봅니다.

- Type I: 효과 없는 약을 "효과 있다"고 팔아서 환자가 죽음. (치명적)
- Type II: 효과 있는 약을 발견 못하고 지나침. (아쉽지만 안전함)

따라서 통계학은 **"Type I Error를 철통같이 막는 것"**을 최우선으로 설계됩니다.

—

5.0.3 3. 유의 수준(α)과 검정력(Power): 최적화 문제

(개념) 개념 3: 제약 조건 하의 최적화

한 줄 요약: "무고한 사람을 가둘 확률은 5% 미만으로 하되($\alpha \leq 0.05$), 그 한도 내에서 최대한 많은 범인을 잡아라(Max Power)."

수학적 레시피 (Neyman-Pearson Lemma)

우리의 목표는 최고의 검정 함수 ψ 를 찾는 것입니다.

1. **Constraint (제약):** $P_{H_0}(\psi(X) = 1) \leq \alpha$ (보통 $\alpha = 0.05$)
2. **Objective (목표):** Maximize $P_{H_1}(\psi(X) = 1)$ (이를 **검정력(Power)**이라 함)

계산 예시: 동전 던지기

동전이 앞면($H_1, p > 0.5$)에 편향되어 있는지 검사합니다. $H_0 : p = 0.5$. 데이터: 10번 던져서 앞면 개수 X 를 셹니다.

- **전략:** ”앞면이 많이 나오면 기각하자.” (기각역: $X \geq k$)
 - ** α 설정 (0.05):** $H_0(p = 0.5)$ 하에서 확률 계산
 - $P(X \geq 9) = P(9) + P(10) \approx 0.0098 + 0.0010 = 0.0108 (< 0.05 \text{ OK})$
 - $P(X \geq 8) = P(8) + \dots \approx 0.0439 + 0.0108 = 0.0547 (> 0.05 \text{ Fail})$
 - **결정:** 기준(k)은 9입니다. 9번 이상 나와야만 ”사기 동전”이라고 부를 수 있습니다.
 - **검정력 계산:** 만약 실제 $p = 0.8$ 이라면? $\text{Power} = P_{p=0.8}(X \geq 9) \approx 0.37$. (범인이 $p = 0.8$ 정도면 37% 확률로만 검거 가능. 데이터(n)를 더 늘려야 함.)
-

5.0.4 4. p-value의 정확한 정의와 오해

(개념) 개념 4: 증거의 강도(Strength of Evidence)

한 줄 요약: ”피고인이 무죄(H_0)라고 치자. 근데 이런 데이터가 나올 확률이 로또 1등 당첨 확률만큼 낮네? 그럼 무죄가 아닌가보다.”

1) 정의

$$\text{p-value} = P(T(X) \geq t_{obs} \mid H_0 \text{ is true})$$

관측된 값(t_{obs})보다 **더 극단적인(More Extreme)** 값이 H_0 세상에서 나올 확률입니다.

2) 해석 가이드

- **Small p-value (< 0.05):** H_0 하에서는 거의 기적 같은 일이다. $\rightarrow H_0$ 가 틀렸다고 보자. (기각)
 - **Large p-value (> 0.05):** H_0 하에서도 흔히 일어날 법한 일이다. $\rightarrow H_0$ 를 유지하자.
 - **주의:** p-value는 ” H_0 가 참일 확률”이 아닙니다! (조건부 확률의 방향 혼동 금지)
-

5.1 실전 시나리오: 넥슨 신규 아이템 매출 분석

[시나리오] Scenario: 업데이트 효과 검증

당신은 '카트라이더'의 신규 카트바디를 출시했습니다. 기존 일평균 매출은 1억 원($\mu_0 = 1$)입니다. 업데이트 후 30일간 데이터를 보니 평균 1.05억 원($\bar{X} = 1.05$)이 되었습니다. 표준편차는 $\sigma = 0.2$ 입니다.

1. **가설 설정:**

- H_0 : 매출 변화 없다 ($\mu = 1$)
- H_1 : 매출 올랐다 ($\mu > 1$)

2. **검정 통계량 (Z-score):**

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.05 - 1}{0.2/\sqrt{30}} \approx \frac{0.05}{0.0365} \approx 1.37$$

3. **p-value 계산:** 표준정규분포에서 $Z \geq 1.37$ 일 확률은 약 0.085 (8.5%)입니다.

4. **결론:** $p\text{-value}(0.085) > \alpha(0.05)$.

5. **Report:** "매출이 500만 원 오르긴 했지만, 통계적으로 유의미하지 않습니다. 단순한 운(우연)이었을 가능성은 배제할 수 없습니다." (H_0 기각 실패)

—

5.2 자주 묻는 질문 (FAQ)

Q1. 왜 하필 0.05(5%)인가요? A. 역사적인 관습(Convention)입니다. 통계학의 아버지 로널드 피셔가 "20번에 1번 정도 틀리는 건 봐주자"라고 제안한 데서 유래했습니다. 의학이나 항공 우주처럼 안전이 중요한 분야는 0.01(1%)이나 그 이하를 쓰기도 합니다.

Q2. p-value가 0.0001이면 효과가 엄청나게 크다는 뜻인가요? A. 아닙니다! p-value는 "효과가 0이 아니라는 확신"의 정도이지, "효과의 크기(Effect Size)"가 아닙니다.

- 데이터가 100만 개면, 매출이 1원만 올랐어도 p-value는 0.00001이 될 수 있습니다.
- 따라서 실무에서는 p-value(유의성)와 함께 **Effect Size(실질적 변화량)**를 꼭 같이 봐야 합니다.

Next Step: 우리는 p-value를 통해 "효과가 있다/없다"를 판별하는 법을 배웠습니다. 하지만 n 이 무한히 커지면 모든 가설 검정이 어떻게 수렴할까요? 다음 Unit 3에서는 데이터가 많을 때($n \rightarrow \infty$) 가설 검정 통계량이 어떤 분포로 수렴하는지(점근적 성질)를 다룹니다.

[요약] Unit 2 (Part B) 핵심 요약

- **구조:** H_0 (보수적) vs H_1 (입증 목표). 무죄 추정의 원칙.

- **오류:** Type I(거짓 발견, α)이 Type II(놓침, β)보다 더 심각하게 다뤄진다.
- **최적화:** α 를 고정(0.05)하고, Power($1 - \beta$)을 최대화하는 기각역을 찾는다.
- **p-value:** H_0 가 참일 때 관측 데이터가 나올 확률. α 보다 작으면 ”너무 희박하므로 H_0 가 거짓말 같다”고 판단한다.

{a4paper, 11pt}book fonts spec amsmath, amssymb, amsthm geometry graphicx xcolor
hyperref booktabs array tikz
left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 4: Confidence Intervals (범위 추정)
- Unit 5: Hypothesis Testing Basics (검정의 철학: α, β)
- Unit 6: Testing Methodology (현재 단원: 구체적인 계산법)
 - 6.1 Neyman-Pearson Lemma (검정의 황금률)
 - 6.2 Likelihood Ratio Test (LRT) & Wilks' Theorem
 - 6.3 The Holy Trinity (LRT, Wald, Score)
 - 6.4 t-test (소표본 검정)
- Unit 7: Goodness of Fit (적합도 검정)

Chapter 6

Unit 6. 검정 방법론 (Testing Methodology)

Unit 5에서 우리는 가설 검정이 ”1종 오류(α)를 둑어두고 검정력(Power)을 최대화하는 최적화 문제”라는 것을 배웠습니다. 하지만 ”그래서 구체적으로 어떤 식을 계산해야 검정력이 최대가 되는데?”라는 질문에는 아직 답하지 않았습니다. 이번 단원에서는 우도(Likelihood)를 이용해 그 **최강의 공식**을 만들어냅니다.

□ 개요 (Overview)

이 단원에서는 검정력을 수학적으로 최대화하는 **네이만-피어슨 보조정리**를 시작으로, 이를 일반화한 **LRT(우도비 검정)**를 배웁니다. 특히 우도 함수(Likelihood Function)의 기하학적 형태를 분석하는 3가지 방법(**LRT, Wald, Score**)의 관계를 이해하고, 데이터가 적을 때 사용하는 **t-test**까지 다룹니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Neyman-Pearson Lemma	”단순 vs 단순” 가설에서 우도비(Ratio)가 짱이다.
Likelihood Ratio (λ)	(H_0 설명력) / (H_1 설명력). 작을수록 H_0 기각.
Wilks’ Theorem	LRT 통계량은 데이터가 많으면 카이제곱 분포(χ^2)가 된다.
Wald Test	MLE와 H_0 사이의 **수평 거리**를 잰다.
Score Test	H_0 지점에서의 **기울기(Slope)**를 잰다.

—

6.0.1 1. 네이만-피어슨 보조정리 (Neyman-Pearson Lemma)

[개념] 개념 1: 가장 강력한 검정(UMP)을 만드는 레시피

한 줄 요약: 두 개의 점(θ_0 vs θ_1)을 비교할 때, 직관이나 감이 아니라 **”우도의 비율”**을 기준으로 삼는 것이 수학적으로 가장 정확하다.

1) 직관적 비유: 비밀번호 매칭

두 사람이 각자 자신이 진짜 계정 주인이라고 주장합니다.

- 철수(H_0): "비밀번호는 1234야."
- 영희(H_1): "비밀번호는 5678이야."

우리가 가진 데이터(입력된 키로그)가 1234와 얼마나 비슷한지($L(\theta_0)$), 5678과 얼마나 비슷한지($L(\theta_1)$) 확률을 계산해서 **비율**을 봅니다. 이 비율만큼 확실한 증거는 없습니다.

2) 수학적 결론

기각역(Rejection Region)을 다음과 같이 설정할 때 검정력(Power)이 최대가 됩니다.

$$\frac{L(\theta_1)}{L(\theta_0)} > k \quad (\text{즉, } H_1 \text{의 우도가 압도적으로 높을 때})$$

이것을 **최강력 검정(UMP: Uniformly Most Powerful test)**이라고 합니다.

6.0.2 2. 우도비 검정 (LRT)과 윌크스 정리

(개념) 개념 2: 만능열쇠 (Universal Key)

한 줄 요약: 복잡한 모델이라도 "제약 있는 우도"와 "제약 없는 우도"의 비율만 계산하면, 카이제곱 분포를 이용해 바로 검정할 수 있습니다.

1) 검정 통계량의 구성

$$\lambda = \frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} = \frac{L(\hat{\theta}_{H_0})}{L(\hat{\theta}_{MLE})}$$

- 분자: H_0 라는 족쇄를 차고 낼 수 있는 최대 점수.
- 분모: 족쇄를 풀고(전체 공간) 낼 수 있는 최대 점수.
- 해석: 비율 λ 가 1에 가까우면 족쇄(H_0)가 별로 방해가 안 된 것이니 H_0 채택. 0에 가까우면 기각.

2) Wilks' Theorem (핵심 정리)

$n \rightarrow \infty$ 일 때, 다음 통계량은 **카이제곱 분포**를 따릅니다.

$$-2 \log \lambda \xrightarrow{d} \chi_d^2$$

(d : 파라미터 개수의 차이. 예: H_0 는 고정값, H_1 은 자유면 $d = 1$) 이 정리 덕분에 우리는 복잡한 시뮬레이션 없이도 p-value를 구할 수 있습니다.

6.0.3 3. 통계적 검정의 삼위일체 (The Holy Trinity)

(개념) 개념 3: 우도 산(Mountain) 등반하기

한 줄 요약: H_0 가 틀렸다는 것을 입증하기 위해 산의 높이(LRT), 너비(Wald), 기울기(Score) 중 하나를 잡니다.

우도 함수(Likelihood Function)를 **하나의 산봉우리**라고 상상해 봅시다.

- **정상(Peak):** $\hat{\theta}_{MLE}$ (데이터가 가장 잘 설명되는 지점)
- **현재 위치:** θ_0 (귀무가설이 주장하는 지점)

A. LRT (Likelihood Ratio Test) - "높이 차이"

- **질문:** "정상의 고도($L(\hat{\theta})$)와 현재 위치의 고도($L(\theta_0)$) 차이가 많이 나는가?"
- **특징:** 가장 정확하지만, 두 지점의 우도를 모두 계산해야 합니다.
- **수식:** $2[\ell(\hat{\theta}) - \ell(\theta_0)]$ (로그 우도의 차이)

B. Wald Test - "수평 거리"

- **질문:** "정상($\hat{\theta}$)까지 수평으로 얼마나 걸어가야 하는가?" ($\hat{\theta} - \theta_0$)
- **특징:** 직관적입니다. 하지만 θ 를 제곱하거나 로그를 취하면 거리가 변하는 단점이 있습니다.
- **수식:** $\frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})}$ (거리를 표준오차로 나눈 것의 제곱)

C. Score Test (LM Test) - "경사도(Slope)"

- **질문:** "지금 서 있는 곳(θ_0)의 경사가 가파른가?"
- **특징:** 정상을 밟아볼 필요가 없습니다! (MLE 계산 불필요). 경사가 평평하면(≈ 0) 여기가 정상이니 H_0 유지, 가파르면 정상이 저 위에 있다는 뜻이니 H_0 기각.
- **수식:** H_0 지점에서의 미분값(Score)의 제곱을 정보량(Information)으로 나눈 것.

—

6.0.4 4. t-test: 데이터가 적을 때의 현실

(개념) 개념 4: 모르는 게 있으면 꼬리가 두꺼워진다

한 줄 요약: 분산 σ^2 을 몰라서 표본분산 S^2 을 대충 써야 할 때, 그 불안감만큼 기각 기준을 엄격하게(꼬리를 두껍게) 만듭니다.

상황

데이터 $X_i \sim \mathcal{N}(\mu, \sigma^2)$ 인데, μ 를 검정하고 싶지만 σ 도 모릅니다.

- Z -test (이상적): $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ (분산을 앎 \rightarrow 정규분포)
- t -test (현실적): $\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ (분산을 추정함 \rightarrow t -분포)

$n \rightarrow \infty$ 가 되면 $S \rightarrow \sigma$ 가 되므로 t -분포는 정규분포와 같아집니다. 즉, t -test는 **LRT의 유한 표본(Finite Sample) 버전**입니다.

6.1 실전 계산: 불공정 동전 판별 (Wald Test)

[시나리오] Scenario: 카지노의 의심

카지노에서 어떤 동전이 앞면(p)이 너무 많이 나온다는 제보가 들어왔습니다. $H_0 : p = 0.5$ vs $H_1 : p \neq 0.5$. 데이터: $n = 100$ 번 던져서 앞면이 60번 나옴 ($\hat{p} = 0.6$).

1. **도구 선택:** 계산이 편한 **Wald Test**를 사용합시다.

2. **거리 측정:** $|\hat{p} - p_0| = |0.6 - 0.5| = 0.1$.

3. **불확실성(표준오차) 계산:** H_0 하에서의 표준오차는 $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5 \times 0.5}{100}} = 0.05$.

4. **Wald 통계량 (Z -score):**

$$W = \frac{\text{거리}}{\text{표준오차}} = \frac{0.1}{0.05} = 2.0$$

5. **판정:** $|W| = 2.0 > 1.96$ (유의수준 5% 기준값).

6. **결론:** ”2표준편차만큼 벗어났으므로 통계적으로 유의미하다. H_0 기각. 이 동전은 불공정하다.”

6.2 자주 묻는 질문 (FAQ)

Q1. LRT, Wald, Score 중에 뭘 써야 하나요? A. 데이터가 무한히 많다면 ($n \rightarrow \infty$) 셋의 결과는 같습니다.

- **정확도:** LRT $>$ Wald \approx Score. (LRT가 분포 근사가 제일 좋습니다.)
- **편의성:** Wald가 제일 쉽습니다 (신뢰구간 구할 때 좋음).
- **계산 비용:** 모델이 너무 복잡해서 MLE를 구하기 힘들 땐 Score가 짱입니다.

Q2. 월크스 정리는 왜 -2 를 곱하나요? A. 수학적 유도 과정에서 정규분포의 지수승 ($e^{-x^2/2}$)에 로그를 취하면 $-x^2/2$ 가 나옵니다. 여기서 x^2 (카이제곱) 모양을 맞추기 위해 -2 를 곱해서 계수를 없애주는 것입니다.

Next Step: 지금까지는 "파라미터 θ 가 0이냐 아니냐"를 검정했습니다(Parametric Test). 그런데 만약 "데이터가 정규분포를 따르긴 하는가?"처럼 분포의 모양 자체를 의심해야 한다면 어떻게 할까요? 다음 **Unit 7**에서는 모델의 가정 자체를 검증하는 **적합도 검정(Goodness of Fit)**을 배웁니다.

(요약) Unit 6 핵심 요약

- **Neyman-Pearson:** 단순 가설 대립 시 우도비(Likelihood Ratio)가 검정력 최강(UMP)이다.
- **Wilks' Theorem:** $-2\log(\text{우도비}) \sim \chi^2$. 복잡한 모델 검정의 만능열쇠.
- **The Holy Trinity:**
 - LRT: 높이 차이 (정확함)
 - Wald: 수평 거리 (직관적, $\hat{\theta}$ 필요)
 - Score: 기울기 (계산 빠름, $\hat{\theta}$ 불필요)
- **t-test:** 분산을 모를 때 사용하는 LRT의 엄밀한 버전.

{a4paper, 11pt}book fonts spec amsmath, amssymb, amsthm geometry graphicx xcolor
hyperref booktabs array
left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 5, 6: Parametric Hypothesis Testing (모델을 믿고 θ 검정)
- Unit 7: Goodness of Fit (현재 단원: 모델 \mathcal{P} 자체를 검증)
 - 7.1 Discrete Setting & Multinomial Distribution
 - 7.2 Pearson's χ^2 Test (핵심 공식)
 - 7.3 Composite Hypothesis (파라미터를 모를 때)
 - 7.4 CDF-based Test (KS Test)
- Unit 8: General Linear Models (회귀 분석의 확장)

Chapter 7

Unit 7. 적합도 검정 (Goodness of Fit)

Unit 6까지 우리는 ”이 데이터는 정규분포를 따른다”는 가정하에 평균이 0인지 아닌지를 싸웠습니다(LRT, Wald, Score). 하지만 누군가 근본적인 질문을 던집니다. ”애초에 데이터가 정규분포가 아니면 어떡할 건데?” 이번 단원에서는 특정 파라미터가 아니라, 데이터의 분포 모양 자체가 이론과 맞는지 확인하는 법을 배웁니다.

□ 개요 (Overview)

적합도 검정은 **”관측된 데이터(Reality)”**와 **”모델이 예측한 데이터(Theory)”** 사이의 거리를 측정하는 과정입니다. 데이터를 범주(Bin)로 나누어 **다항 분포** 문제로 치환하고, **피어슨 카이제곱 통계량**을 사용해 오차를 분석합니다. 이때 파라미터를 추정해서 검정할 경우 자유도가 감소하는 원리까지 다룹니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Multinomial Dist.	주사위 던지기. 데이터가 여러 칸(Bin) 중 하나에 들어감.
Observed (O_j)	실제 데이터가 각 칸에 들어간 횟수 (N_j).
Expected (E_j)	이론적으로 들어갔어야 할 횟수 (np_j).
χ^2 Statistic	$\sum \frac{(O-E)^2}{E}$. 현실과 이론의 괴리감 점수.
Degrees of Freedom (df)	자유롭게 움직일 수 있는 데이터의 개수 ($K - 1 - d$).

—

7.0.1 1. 이산 데이터와 다항 분포 (The Setup)

(개념) 개념 1: 연속을 이산으로 (Discretization)

한 줄 요약: 모양을 비교하기 가장 쉬운 방법은, 구간(Bin)을 나누어 각 구간에 몇 개가 들어갔는지 세는 것입니다. (히스토그램 만들기)

1) 설정 (Setup)

데이터가 K 개의 범주(Category) 중 하나에 떨어집니다.

- 예: 혈액형 (A, B, O, AB → $K = 4$)

- 예: 키 (160이하, 160-170, 170-180, 180이상 $\rightarrow K = 4$)

각 범주에 관측된 횟수를 N_1, \dots, N_K 라고 하면, 이 벡터는 **다항 분포(Multinomial Distribution)**를 따릅니다.

7.0.2 2. 피어슨의 카이제곱 검정 (Pearson's χ^2 Test)

(개념) 개념 2: 현실(Observed)과 이론(Expected)의 거리

한 줄 요약: 단순히 차이를 제곱해서 더하는 게 아니라, "기대값이 큰 곳은 오차도 크다"는 점을 감안하여 **표준화**한 거리입니다.

1) 직관적 비유: 시험 점수 보정

- 수학(평균 50점)에서 10점 차이 나는 것과,
- 체육(평균 90점)에서 10점 차이 나는 것은 무게감이 다릅니다.
- 오차의 절대적인 크기($(O - E)^2$)가 아니라, 그 동네의 규모(E) 대비 얼마나 큰지를 봐야 합니다.

2) 핵심 공식 (T_n)

$$T_n = \sum_{j=1}^K \frac{(N_j - np_j^0)^2}{np_j^0} = \sum_{j=1}^K \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

- **분자 $(O - E)^2$:** 오차의 크기.
- **분모 E :** 분산(Variance) 역할. (포아송/다항분포에서 분산은 평균과 비례함).
- **의미:** $\frac{\text{오차}}{\text{표준편차}} \approx Z$ (표준정규분포)가 되므로, 이것의 제곱 합은 χ^2 가 됩니다.

3) 점근적 성질 (Asymptotic Property)

$n \rightarrow \infty$ 일 때,

$$T_n \xrightarrow{d} \chi_{K-1}^2$$

왜 K 가 아니라 $K - 1$ 인가? 데이터의 총합은 반드시 n 이어야 합니다 ($\sum N_j = n$). 마지막 칸(N_K)은 앞의 $K - 1$ 개 값이 정해지면 자동으로 결정되므로 자유가 없습니다.

7.0.3 3. 복합 가설 검정 (Composite Goodness of Fit)

(개념) 개념 3: 옷을 몸에 맞췄다면, 심사 기준을 높여라

한 줄 요약: 파라미터를 몰라서 데이터로 추정($\hat{\theta}$)한 뒤 검정할 때는, 이미 데이터를 한 번 훔쳐본 셈이므로 자유도(df)를 깎아서 패널티를 줍니다.

1) 문제 상황

”이 데이터가 포아송 분포를 따르는가?”라고 묻지만, λ 가 3인지 5인지 모릅니다.

- 해결: 데이터에서 먼저 MLE $\hat{\lambda}$ 를 구합니다.
- 적용: 그 $\hat{\lambda}$ 를 이용해 기대 확률 $p_j(\hat{\lambda})$ 를 계산합니다.

2) 자유도의 변화 (핵심 이론)

$$T_n = \sum \frac{(N_j - np_j(\hat{\theta}))^2}{np_j(\hat{\theta})} \xrightarrow{d} \chi_{K-1-d}^2$$

- d : 추정한 파라미터의 개수.
 - **직관:** 파라미터를 추정하는 과정에서 데이터의 정보(자유도)를 d 만큼 소모했습니다. 모델을 데이터에 ”끼워 맞췄기(Fitting)” 때문에, 오차($O-E$)가 자연 상태보다 인위적으로 줄어듭니다. 이를 보정하기 위해 분포의 기준을 더 빽빽하게(df 감소) 잡는 것입니다.
-

7.1 실전 시나리오: 확률형 아이템(가챠) 조작 의혹 검증

[시나리오] Scenario: 유저들의 봉기

당신은 넥슨의 게임 운영자입니다. 유저들이 ”전설/영웅/일반 아이템 확률이 공지된 것과 다르다!”고 소송을 제기했습니다. 공지 확률: 전설(10%), 영웅(30%), 일반(60%). ($K = 3$)

1. **데이터 수집:** 100회 뽑기 결과 ($n = 100$)

- 관측(O): 전설 5개, 영웅 25개, 일반 70개.

2. **기대값 계산 (E):**

- 전설: $100 \times 0.1 = 10$
- 영웅: $100 \times 0.3 = 30$
- 일반: $100 \times 0.6 = 60$

3. **카이제곱 통계량 계산:**

$$\begin{aligned} \chi^2 &= \frac{(5 - 10)^2}{10} + \frac{(25 - 30)^2}{30} + \frac{(70 - 60)^2}{60} \\ &= \frac{25}{10} + \frac{25}{30} + \frac{100}{60} = 2.5 + 0.83 + 1.67 = 5.0 \end{aligned}$$

4. **판정 (자유도 $K - 1 = 2$):** 유의수준 5%에서 χ_2^2 의 임계값은 **5.99**입니다.

$$5.0 < 5.99 \quad (\text{기각 실패})$$

5. **결론:** ”관측된 차이가 다소 있긴 하지만(전설이 적게 나옴), 통계적으로 ’조작’이라 단정할 만큼 극단적이진 않습니다. 우연의 범위 내입니다.” (유저들의 분노는 가라앉지 않겠지만, 수학적으로 무죄입니다.)

7.2 자주 묻는 질문 (FAQ)

- Q1. 데이터가 적어도 이 방법을 쓸 수 있나요? A. 위험합니다. 보통 각 칸(Bin)의 기대 빈도(E_j)가 최소 5 이상은 되어야 카이제곱 분포 근사가 잘 맞습니다. 만약 기대 빈도가 너무 작으면, 인접한 칸들을 합쳐서(Merge) $E \geq 5$ 를 만들어야 합니다.
- Q2. 그냥 $(O - E)$ 만 보면 안 되나요? 왜 제곱하고 나누나요? A. $(O - E)$ 를 그냥 더하면 양수와 음수가 상쇄되어 항상 0이 됩니다. 그래서 제곱을 합니다. 그리고 E 로 나누는 것은 "가중치" 때문입니다. 1000개 중 10개 차이나는 것과, 20개 중 10개 차이나는 것은 다르기 때문입니다.
- Q3. 연속형 데이터는 어떻게 하나요? A. 카이제곱 검정을 쓰려면 강제로 구간(Bin)을 나눠야 합니다(정보 손실 발생). 그게 싫다면 구간을 나누지 않고 누적 분포 함수(CDF) 자체를 비교하는 **콜모고로프-스미르노프(KS) 검정**을 사용하면 됩니다.

Next Step: 우리는 이제 θ 를 찾는 것(Unit 2 6)을 넘어 모델 전체를 검증(Unit 7)했습니다. 통계학의 기초는 끝났습니다. 이제 통계학의 꽃이자 머신러닝의 시초인 **Unit 8: 선형 회귀 분석 (Linear Regression)**으로 넘어가, 변수와 변수 사이의 관계($X \rightarrow Y$)를 모델링합니다.

(요약) Unit 7 핵심 요약

- **관점 전환:** 파라미터 값이 아니라 분포의 모양(Shape)을 검정한다.
- **카이제곱 통계량:** $\sum \frac{(O-E)^2}{E}$. 관측값과 기대값의 표준화된 거리.
- **자유도(df):** 기본적으로 $K-1$. 단, 파라미터를 데이터로 추정했다면 $K-1-(\text{추정 파라미터 수})$ 로 줄어든다.
- **주의점:** 각 칸의 기대 빈도가 너무 작으면(5 미만) 신뢰할 수 없다.

[a4paper, 11pt]book fontspec amsmath, amssymb, amsthm geometry graphicx xcolor
hyperref booktabs array bm
left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 1, 2: Estimation (현재 상태 파악)
- Unit 3: Linear Regression (현재 단원: 미래 예측과 관계 규명)
 - 3.1 The Setup: Matrix Formulation
 - 3.2 Least Squares Estimation (LSE)
 - 3.3 Geometric Interpretation (Projection)
 - 3.4 Gauss-Markov Theorem (BLUE)
 - 3.5 Inference (t-test, F-test)
- Unit 4: Hypothesis Testing (심화)

Chapter 8

Unit 3. 선형 회귀 (Linear Regression)

지금까지 우리는 하나의 변수(예: 동전 앞면 확률 p , 평균 키 μ)를 추정하는 데 집중했습니다. 하지만 현실 세계는 여러 변수가 얹혀 있습니다. 키는 유전, 영양, 운동량에 영향을 받죠. 이제 우리는 ”변수 X 가 변할 때 결과 Y 는 어떻게 변하는가?”를 수학적으로 모델링하고, 이를 통해 보이지 않는 미래를 예측(Prediction)하는 단계로 나아갑니다.

□ 개요 (Overview)

선형 회귀는 입력(X)과 출력(Y)의 관계를 선형 방정식($Y = X\beta$)으로 설명하는 기법입니다. 이 단원에서는 **최소자승법(LSE)**이 기하학적으로는 **직교 투영(Orthogonal Projection)**임을 이해하고, 데이터가 정규분포를 따르지 않아도 LSE가 최적의 추정량(**BLUE**)이 됨을 증명하는 **가우스-마르코프 정리**를 배웁니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Design Matrix (X)	요리 재료들의 목록표 ($n \times p$ 행렬).
Coefficient Vector (β)	각 재료가 맛에 미치는 영향력 (우리가 구해야 할 레시피).
Residual (e or \hat{e})	실제 맛(Y)과 레시피대로 만든 맛(\hat{Y})의 차이.
Projection Matrix (P)	Y 를 X 의 공간 위로 수직으로 내리꽂는 그림자 생성기.
BLUE	Best Linear Unbiased Estimator. (가장 믿을만한 선형 추정량).

—

8.0.1 1. 선형 회귀 모델의 구조 (The Setup)

(개념) 개념 1: 행렬로 세상을 표현하다

한 줄 요약: 복잡한 연립방정식 문제를 $Y = X\beta + \epsilon$ 이라는 우아한 행렬식 하나로 압축합니다.

1) 직관적 비유: 아파트 가격 맞추기

- Y (결과): 아파트 가격들 (10억, 15억, ...)

- X (재료): (평수, 역까지 거리, 학군 점수)
- β (영향력): (평당 가격, 1km당 감가액, 학군 프리미엄)
- ϵ (잡음): 옆집이 시세보다 싸게 내놓은 급매물 같은 예측 불가능한 요인.

2) 수학적 정의

$$Y = X\beta + \epsilon$$

여기서 X 는 $n \times p$ 행렬입니다 (n : 데이터 개수, p : 변수 개수). 주요 가정:

1. $\mathbb{E}[\epsilon] = 0$ (잡음의 평균은 0이다.)
 2. $\text{Var}(\epsilon) = \sigma^2 I_n$ (모든 데이터의 잡음 수준은 일정하고 독립적이다.)
-

8.0.2 2. 최소자승법 (Least Squares Estimation, LSE)

(개념) 개념 2: 오차의 제곱을 최소화하라

한 줄 요약: 모든 데이터 점들과 직선 사이의 거리(잔차) 제곱합을 최소로 만드는 β 를 찾습니다.

1) 최적화 문제

우리의 목표는 잔차 벡터의 길이 ($\|Y - X\beta\|^2$)를 최소화하는 $\hat{\beta}$ 를 찾는 것입니다.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2$$

2) 해 구하기 (Normal Equations)

위 식을 β 로 미분하여 0이 되는 지점을 찾으면 다음과 같은 **정규 방정식(Normal Equations)**이 나옵니다.

$$X^T X \beta = X^T Y$$

만약 $(X^T X)$ 의 역행렬이 존재한다면, 우리는 **단 한 번의 행렬 연산**으로 정답을 얻습니다.

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

8.0.3 3. 기하학적 해석: 직교 투영 (Projection)

(개념) 개념 3: 그림자 놀이

한 줄 요약: 우리의 예측값 \hat{Y} 는 실제 데이터 Y 를 X 가 만드는 평면(공간) 위로 **수직으로 내리꽂은 그림자**입니다.

1) 공간의 이해

- 데이터 Y 는 n 차원 공간에 떠 있는 하나의 점입니다.
- 우리가 가진 재료 X 들로 만들 수 있는 모든 예측값들의 집합을 **열 공간(Column Space, $\mathcal{C}(X)$)**이라고 합니다. 이는 n 차원 공간 속의 작은 평면입니다.

2) 직교성 (Orthogonality)

점 Y 에서 평면 $\mathcal{C}(X)$ 까지 거리가 가장 짧으려면? 당연히 **수직(Orthogonal)**으로 내려야 합니다. 즉, 잔차 벡터 $e = Y - \hat{Y}$ 는 평면 $\mathcal{C}(X)$ 와 직교합니다.

$$X^T e = 0 \implies X^T(Y - X\hat{\beta}) = 0$$

이 식을 풀면 앞서 본 정규 방정식 $X^T X \beta = X^T Y$ 가 바로 유도됩니다!

3) 투영 행렬 (Hat Matrix)

예측값 \hat{Y} 는 Y 에 **투영 행렬 P **를 곱한 것입니다.

$$\hat{Y} = PY, \quad \text{where } P = X(X^T X)^{-1} X^T$$

(P 를 Hat Matrix라고 부르는 이유는 Y 머리 위에 모자 \hat{Y} 를 써워주기 때문입니다.)

8.0.4 4. 가우스-마르코프 정리 (Gauss-Markov Theorem)

(개념) 개념 4: 왜 하필 LSE인가?

한 줄 요약: 데이터가 정규분포가 아니더라도, 오차의 평균이 0이고 분산이 일정하다면 LSE가 **"가장 분산이 작은(정밀한) 선형 추정량"**임이 수학적으로 보장됩니다.

BLUE (Best Linear Unbiased Estimator)

LSE 추정량 $\hat{\beta}$ 는 다음 조건을 만족하는 챔피언입니다.

- **Linear:** 데이터 Y 의 선형 결합으로 계산됨.
- **Unbiased:** 평균적으로 참값 β 를 맞춤 ($\mathbb{E}[\hat{\beta}] = \beta$).
- **Best:** 위의 두 조건을 만족하는 애들 중에서 **분산이 가장 작음(Minimum Variance)**.

즉, 다른 방법을 쓰면 영점이 흔들리거나(Biased), 탄착군이 넓어집니다(High Variance).

8.0.5 5. 추론 (Inference): 가설 검정

이제 $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ 라는 **정규분포 가정**을 추가합니다. 그래야 신뢰 구간과 p-value를 구할 수 있습니다.

1) 추정량의 분포

$\hat{\beta}$ 는 정규분포 따르는 Y 의 선형 변환인므로, 역시 정규분포를 따릅니다.

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

2) t-test (변수의 유의성 검정)

”학군(β_3)이 집값에 진짜 영향이 있나?” ($H_0 : \beta_3 = 0$)

$$T = \frac{\hat{\beta}_j - 0}{\text{SE}(\hat{\beta}_j)} \sim t_{n-p}$$

여기서 $\text{SE}(\hat{\beta}_j)$ 는 $\hat{\sigma} \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$ 입니다.

[주의] 코크란의 정리 (Cochran's Theorem)

우리가 t-분포를 쓸 수 있는 이유는 **”추정된 계수($\hat{\beta}$)와 잔차($\hat{\epsilon}$)가 서로 독립”**이라는 수학적 성질 덕분입니다. 이는 기하학적으로 투영된 그림자(\hat{Y})와 수직선(e)이 직교하기 때문에 성립합니다.

—

8.1 실전 시나리오: 넷플릭스 시청 시간 예측

(시나리오) Scenario: 새로운 드라마의 성공 예측

당신은 넷플릭스의 데이터 과학자입니다. 신규 드라마의 ’첫 달 시청 시간(Y)’을 예측하려 합니다.

1. **변수 설정 (X):**

- X_1 : 제작비 (억 원)
- X_2 : 주연 배우의 인스타 팔로워 수 (만 명)
- X_3 : 에피소드 수

2. **모델링:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

3. **LSE 결과:** $\hat{\beta} = [100, 2.5, 0.1, 50]^T$

- 해석: 제작비 1억 늘면 시청 시간 2.5만 시간 증가.
- 해석: 팔로워 1만 명당 시청 시간 0.1만 시간 증가. (생각보다 적음)

4. **t-test:** X_2 (팔로워 수)의 p-value가 0.35가 나옴.

- 결론: ”주연 배우의 팔로워 수는 시청 시간에 통계적으로 유의미한 영향이 없습니다.”
→ 마케팅 팀에 ”인플루언서 섭외보다 제작비 증액이 낫습니다”라고 제안.

—

8.2 자주 묻는 질문 (FAQ)

Q1. ($\mathbf{X}^T \mathbf{X}$)의 역행렬이 없으면 어떡하죠? A. 이를 **다중공선성(Multicollinearity)** 문제라고 합니다. 예를 들어 ’키(cm)’와 ’키(m)’를 동시에 변수로 넣으면, 두 변수가 완벽하게 겹쳐서 수학적으로 해를 구할 수 없습니다. 이때는 변수를 하나 제거하거나, **Ridge Regression** 같은 기법을 써야 합니다.

Q2. 왜 잔차 제곱의 합을 쓰나요? 절댓값의 합을 쓰면 안 되나요? A. 써도 됩니다(LAD 회귀). 하지만 절댓값은 미분이 불가능한 점(뾰족한 점)이 있어서 수학적으로 다루기 어렵습니다. 반면 제곱(L^2)은 미분이 깔끔하고, 기하학적으로 '직교 투영'이라는 완벽한 해석이 가능하기 때문에 통계학의 표준이 되었습니다.

Next Step: 우리는 연속형 숫자(Y)를 예측하는 선형 회귀를 배웠습니다. 그런데 만약 결과가 숫자가 아니라 **"합격/불합격", "암/정상"** 같은 범주라면 어떡할까요? 선형 회귀로는 0과 1 사이의 확률을 표현하기 어렵습니다. 다음 시간에는 이를 해결하는 **로지스틱 회귀(Logistic Regression)**와 일반화 선형 모형(GLM)을 배웁니다.

(요약) Unit 3 핵심 요약

- **모델:** $Y = X\beta + \epsilon$. 현실을 선형 결합으로 근사한다.
- **LSE:** 잔차 제곱합을 최소화하는 방법. $\hat{\beta} = (X^T X)^{-1} X^T Y$.
- **기하학:** 예측값 \hat{Y} 는 데이터 Y 를 열 공간 $C(X)$ 에 **수직 투영(Orthogonal Projection)**한 것이다.
- **Gauss-Markov:** 정규성 가정이 없어도 LSE는 분산이 가장 작은 최적의 추정량(BLUE)이다.

(a4paper, 11pt) book fontspec amsmath, amssymb, amsthm geometry graphicx xcolor
hyperref booktabs array
left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 3: Linear Regression (연속형 데이터 예측)
- Unit 7: Goodness of Fit (모델 검증)
- Unit 9: Generalized Linear Models (현재 단원: 확장팩 설치)
 - 9.1 Exponential Family (공통의 조상)
 - 9.2 Link Function (데이터 범위 해결)
 - 9.3 Logistic Regression (이진 분류)
 - 9.4 Poisson Regression (카운트 데이터)
 - 9.5 IRLS Algorithm (해를 구하는 법)
- Unit 10: Classification (머신러닝으로의 연결)

Chapter 9

Unit 9. 일반화 선형 모형 (GLM)

Unit 3에서 배운 선형 회귀($Y = X\beta$)는 강력하지만 치명적인 약점이 있습니다. Y 가 정규분포를 따라야 한다는 점이죠. 만약 "내일 비가 올까(0/1)?"를 선형 회귀로 예측했더니 "확률이 120%"라거나 "-30%"라는 황당한 답이 나온다면 어떡할까요? GLM은 이 문제를 해결하기 위해 선형 회귀에 '유연한 연결고리(Link)'를 달아주는 과정입니다.

□ 개요 (Overview)

GLM은 정규분포, 베르누이 분포, 포아송 분포 등을 **지수족(Exponential Family)**이라는 하나의 수학적 틀로 묶고, **연결 함수(Link Function)**을 통해 선형 예측($X\beta$)을 데이터의 특성에 맞는 범위(예: 0~1)로 매핑하는 기법입니다. 이를 통해 분류(Classification)와 회귀(Regression)를 통합적으로 다룹니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Exponential Family	"다르게 생겼지만 사실은 한 가족". 미분과 계산이 편한 분포들의 모임.
Link Function (g)	$X\beta$ (무한대 범위)를 μ (제한된 범위)로 통역해주는 번역기.
Logit Link	확률(0~1)을 $\pm\infty$ 로 펴주는 함수. (로지스틱 회귀용)
Log Link	양수(0~ ∞)를 전체 실수로 펴주는 함수. (포아송 회귀용)
IRLS	GLM은 공식 한 방에 안 풀리므로, 조금씩 정답을 찾아가는 반복 알고리즘.

—

9.0.1 1. 지수족 분포 (Exponential Family)

[개념] 개념 1: 통계학의 만능 플랫폼

한 줄 요약: 정규분포, 베르누이, 포아송 등은 겉보기엔 다르지만, 사실 $f(y) = h(y)e^{\eta T(y) - A(\eta)}$ 라는 **공통 DNA**를 가진 가족입니다.

1) 왜 배우는가?

서로 다른 분포들을 매번 따로 연구할 필요가 없습니다. 이 "가족"에 속하기만 하면 다음과 같은 강력한 성질(VIP 혜택)을 공유하기 때문입니다.

- **Convexity:** 로그 우도 함수가 항상 위로 볼록(Concave)합니다. 즉, 최적화할 때 **'가짜 봉우리(Local Optima)'가 없어서** 안심하고 답을 찾을 수 있습니다.
- **Automatic Moments:** 복잡한 적분 없이 미분만으로 평균과 분산을 구할 수 있습니다. ($A'(\eta) = \text{평균}, A''(\eta) = \text{분산}$)

2) 수학적 구조

$$f(y; \theta) = h(y) \exp(\eta(\theta) \cdot T(y) - A(\theta))$$

- η : 자연 파라미터 (Natural Parameter). 분포의 모양을 결정하는 핵심.
 - $T(y)$: 충분 통계량.
 - $A(\theta)$: 로그 분배 함수 (정규화 상수). 확률의 합이 1이 되게 맞추는 역할.
-

9.0.2 2. 연결 함수 (Link Function)

(개념) 개념 2: 무한의 세계와 유한의 세계를 잇다

한 줄 요약: 선형 회귀 결과값 $X\beta$ 는 $-\infty \sim \infty$ 범위를 가지지만, 우리가 원하는 평균 μ 는 범위가 제한적입니다(확률은 0~1). 이 둘을 이어주는 다리입니다.

1) 로지스틱 회귀 (Logistic Regression)

- **상황:** 합격(1) vs 불합격(0). $Y \sim \text{Bernoulli}(p)$.
- **문제:** $X\beta = 100$ 이 나오면 확률이 1을 넘어버림.
- **해결 (Logit Link):** 확률(p) 대신 **로그 오즈(Log-Odds)**를 예측합니다.

$$g(p) = \log\left(\frac{p}{1-p}\right) = X\beta$$

- **복원:** 역함수를 취하면 그 유명한 **시그모이드(Sigmoid)** 함수가 됩니다.

$$p = \frac{e^{X\beta}}{1 + e^{X\beta}} = \frac{1}{1 + e^{-X\beta}}$$

2) 포아송 회귀 (Poisson Regression)

- **상황:** 하루 교통사고 건수. $Y \in \{0, 1, 2, \dots\}$. (음수 불가능)
- **문제:** 선형 회귀는 음수 값을 예측할 수도 있음.
- **해결 (Log Link):** 평균(λ)에 로그를 씌워서 예측합니다.

$$g(\lambda) = \log(\lambda) = X\beta \implies \lambda = e^{X\beta}$$

- **의미:** X 가 증가하면 Y 는 **기하급수적(Multiplicative)**으로 증가합니다.
-

9.0.3 3. 추정 알고리즘: IRLS

(개념) 개념 3: 산을 오르는 반복적인 걸음

한 줄 요약: 선형 회귀처럼 한 번에 답(β)이 나오지 않습니다. 직선을 긋고(Linear), 가중치를 조절하고(Reweight), 다시 긋는 과정을 반복(Iterative)합니다.

1) 문제점

GLM의 로그 우도 함수를 미분하면, β 가 지수함수($e^{X\beta}$) 안에 갇혀 있어서 깔끔하게 정리가 안 됩니다 (No closed-form solution).

2) 해결책: 뉴턴-랩슨 IRLS

근사적으로 해를 구하는 수치해석 기법을 씁니다.

- **Iteratively Reweighted Least Squares (IRLS):** 매 단계마다 분산이 다른 것을 고려하여 **가중치(Weight)**를 둔 선형 회귀(WLS)**를 푸는 문제로 바꿔서 풁니다. 컴퓨터가 아주 빠르게 반복해서 최적의 β 를 찾아냅니다.
-

9.1 실전 시나리오: 대학원 합격 예측 (로지스틱)

(시나리오) Scenario: GPA와 합격률의 관계

학생들의 GPA(X)에 따른 합격 여부($Y \in \{0, 1\}$)를 분석합니다.

1. **데이터:** GPA 3.0은 불합격, 3.5는 합격... 이런 데이터 100개.

2. **모델링:** $Y \sim \text{Bernoulli}(p)$, $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 X$.

3. **결과:** $\hat{\beta}_0 = -10$, $\hat{\beta}_1 = 3$.

4. **예측 계산 (GPA 3.8인 학생):**

- 선형 예측값: $\eta = -10 + 3(3.8) = -10 + 11.4 = 1.4$.
- 이것은 확률이 아니라 **로그 오즈**입니다.
- 확률 변환: $p = \frac{1}{1+e^{-1.4}} \approx \frac{1}{1+0.246} \approx 0.80$.

5. **해석:** GPA가 3.8인 학생은 합격 확률이 약 80%입니다.

6. **계수 해석:** $\beta_1 = 3$ 은 양수이므로, GPA가 높을수록 합격 오즈가 $e^3 \approx 20$ 배 증가합니다 (매우 강력한 영향).

9.2 자주 묻는 질문 (FAQ)

- Q1. 왜 확률을 바로 $X\beta$ 로 두지 않고 굳이 '오즈'에 로그를 씌우나요? A. 만약 $P(Y = 1) = \beta_0 + \beta_1 X$ 로 두면, X 가 아주 커졌을 때 확률이 1.5가 되거나 -0.2가 되는 모순이 발생합니다. 로그 오즈는 범위가 $(-\infty, \infty)$ 이므로, $X\beta$ 와 매칭시키기에 수학적으로 가장 안전하고 자연스럽습니다.
- Q2. 정준 연결 함수(Canonical Link)가 뭔가요? A. '순정 부품' 같은 겁니다. 수학적으로 가장 깔끔하게 떨어지는 연결 함수 조합입니다.

- Normal \rightarrow Identity ($Y = X\beta$)
- Bernoulli \rightarrow Logit
- Poisson \rightarrow Log

다른 걸 써도 되지만(예: Probit), 정준 링크를 쓰면 계산이 편하고 해석이 명확합니다.

Next Step: 우리는 GLM을 통해 Y 가 0/1인 경우(로지스틱)까지 다루었습니다. 하지만 만약 Y 가 0/1이 아니라 **개, 고양이, 사자"처럼 3개 이상의 범주**라면 어떡할까요? 다음 **Unit 10**에서는 이를 해결하는 다중 분류(Multiclass Classification)와 현대적 분류 기법들을 배웁니다.

(요약) Unit 9 핵심 요약

- **지수족(Exponential Family):** GLM의 수학적 기반. 로그 우도가 오목(Concave)하여 최적화가 쉽다.
- **연결 함수(Link Function):** $g(\mu) = X\beta$. 선형 예측값을 데이터 범위에 맞게 변환한다.
- **로지스틱 회귀:** $g = \text{logit}$. 이진 분류에 사용. 결과는 로그 오즈.
- **포아송 회귀:** $g = \text{log}$. 카운트 데이터에 사용.
- **IRLS:** GLM의 해를 구하는 반복적 알고리즘.

{a4paper, 11pt}book fontspec amsmath, amssymb, amsthm geometry graphicx xcolor hyperref booktabs array

left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 1-3: Parametric Statistics (정규분포 가정 하의 추정)
- Unit 4: Non-parametric Statistics (현재 단원: 가정 없는 추정)
 - 10.1 Histograms (가장 기초적인 방법)
 - 10.2 Kernel Density Estimation (KDE, 스무딩)
 - 10.3 Bias-Variance Trade-off (핵심 딜레마)
 - 10.4 Optimal Bandwidth Selection
- Unit 5: PCA (차원 축소)

Chapter 10

Unit 4. 밀도 추정 (Density Estimation)

Unit 3까지는 “키는 정규분포를 따른다”고 가정하고 평균(μ)만 찾았습니다. 하지만 만약 분포가 낙타 등처럼 봉우리가 두 개(Bimodal)라면요? 평균만으로는 데이터의 진짜 모습을 설명할 수 없습니다. 이제 우리는 파라미터 몇 개를 찾는 게 아니라, **분포의 모양(함수 f) 자체를 그려내는 방법**을 배웁니다.

□ 개요 (Overview)

밀도 추정은 관측된 데이터로부터 미지의 확률 밀도 함수 f 를 복원하는 기술입니다. 가장 단순한 **히스토그램**의 한계를 극복하기 위해 **커널 밀도 추정(KDE)**을 도입하며, 이때 발생하는 **편향(Bias)**과 **분산(Variance)**의 트레이드오프를 조절하여 최적의 스무딩 파라미터(대역폭 h)를 찾는 것이 핵심입니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Density Estimation	데이터 점들을 보고 원래 어떻게 분포해 있었는지 지도를 그리는 것.
Histogram	데이터를 구간(Bin)별로 나눠 벽돌 쌓기. (각지고 거친)
KDE (Kernel Density Estimation)	데이터를 벽돌 대신 부드러운 모래무덤으로 쌓기. (부드러움)
Bandwidth (h)	모래무덤의 평균 정도. 작으면 뾰족, 크면 평평.
Bias-Variance Trade-off	디테일을 살릴 것인가(Low Bias), 노이즈를 없앨 것인가(Low Variance)?

10.0.1 1. 히스토그램 (Histogram)

(개념) 개념 1: 디지털 모자이크

한 줄 요약: 데이터 공간을 격자(Bin)로 나누고, 각 칸에 떨어진 데이터 개수만큼 높이를 올리는 계단 함수입니다.

1) 한계점

- **불연속성:** 현실의 확률은 부드러운 곡선인데, 히스토그램은 각진 계단 모양입니다.
 - **시작점 의존성:** 구간을 0부터 시작하느냐, 0.5부터 시작하느냐에 따라 모양이 완전히 달라집니다.
-

10.0.2 2. 커널 밀도 추정 (KDE, Kernel Density Estimation)

[개념] 개념 2: 벽돌 대신 모래를 쌓자

한 줄 요약: 각 데이터 포인트 위치에 '커널(Kernel)'이라는 작은 확률의 언덕을 쌓고, 이 언덕들을 모두 더해서 전체 지형을 만듭니다.

1) 직관적 비유: 모래무덤 쌓기

데이터 점이 하나(X_i) 찍힐 때마다, 그 위에 모래 한 줌(K)을 놓습니다.

- 데이터가 몰려 있는 곳: 모래가 많이 쌓여 높은 산이 됩니다.
- 데이터가 없는 곳: 모래가 없어 평지가 됩니다.
- 결과적으로 아주 부드러운 곡선(Smooth Curve)이 완성됩니다.

2) 수학적 정의 (The Estimator)

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

- $K(\cdot)$: 커널 함수. 보통 정규분포(Gaussian) 모양을 씁니다. ($\int K = 1$)
 - h : **대역폭(Bandwidth)**. 커널의 폭을 결정하는 가장 중요한 변수입니다.
 - 수학적 의미: 경험적 분포와 커널 함수의 **합성곱(Convolution)**입니다.
-

10.0.3 3. 편향-분산 트레이드오프 (Bias-Variance Trade-off)

[개념] 개념 3: 너무 섬세해도, 너무 둔감해도 안 된다

한 줄 요약: 대역폭 h 가 작으면 노이즈까지 따라그리고(과적합), h 가 크면 중요 특징을 뭉개버립니다(과소적합)

우리의 목표는 추정함수 \hat{f} 와 실제함수 f 사이의 차이(MSE)를 줄이는 것입니다.

$$\text{MSE} = \text{Bias}^2 + \text{Variance}$$

1) 대역폭 h 가 작을 때 (Narrow Bandwidth)

- **현상:** 그래프가 자글자글하고 뾰족합니다 (Spiky).
- **Bias (낮음):** 데이터가 있는 위치를 아주 정확히 찌릅니다.
- **Variance (높음):** 데이터 하나만 바뀌어도 그래프 모양이 요동칩니다. (Overfitting)

2) 대역폭 h 가 클 때 (Wide Bandwidth)

- **현상:** 그래프가 평퍼짐하고 뒷걸음치는 듯합니다 (Smooth).
 - **Bias (높음):** 뾰족한 봉우리(Peak)를 깎아먹고 평탄화시킵니다. (Underfitting)
 - **Variance (낮음):** 데이터가 좀 바뀌어도 둔감하게 반응합니다.
-

10.0.4 4. 최적의 대역폭 선택 (Optimal Bandwidth)

[개념] 개념 4: 중庸(Golden Mean)을 찾아서

한 줄 요약: Bias와 Variance의 합이 최소가 되는 지점을 미분으로 찾습니다.

수학적 결론

MISE(Mean Integrated Squared Error)를 최소화하는 최적의 h 는 데이터 개수 n 과 다음과 같은 관계를 가집니다.

$$h_{opt} \propto n^{-1/5}$$

- 의미: 데이터 n 이 많아지면 h 를 천천히 줄여서 디테일을 살려야 합니다.
 - **주의:** 모수적 방법($1/\sqrt{n}$)보다 수렴 속도($n^{-2/5}$)가 느립니다. 즉, 함수 전체를 추정하려면 훨씬 더 많은 데이터가 필요합니다.
-

10.1 실전 시나리오: 게임 유저 플레이 타임 분석

(시나리오) Scenario: 평균의 함정 탈출

당신은 MMORPG 게임의 PM입니다. "유저들의 평균 플레이 타임은 2시간입니다"라는 보고를 받았습니다. 그래서 2시간짜리 콘텐츠를 만들었는데 망했습니다. 왜일까요?

- **데이터 시각화 (KDE 적용):** 단순 평균(μ) 대신 KDE로 분포를 그려보았습니다.
 - **발견 (Bimodal Distribution):** 분포가 낙타 등처럼 **봉우리가 두 개**였습니다.
 - 그룹 A (라이트 유저): 30분 플레이 (숙제만 하고 끔)
 - 그룹 B (헤비 유저): 5시간 플레이 (레이드 뛴)
 평균인 '2시간'에 해당하는 유저는 사실 아무도 없었습니다!
 - ** h 파라미터 튜닝:**
 - h 를 너무 크게 잡으면: 두 봉우리가 뭉개져서 하나로 보임 (평균 2시간의 함정 재현).
 - h 를 적절히 잡으면: 30분과 5시간의 두 봉우리가 명확히 분리됨.
 - **전략 수정:** 2시간짜리 콘텐츠 대신, **"30분짜리 일일 퀘스트"**와 **"4시간짜리 레이드"**로 콘텐츠를 이원화하여 대성공을 거둡니다.
-

10.2 자주 묻는 질문 (FAQ)

- Q1. 커널 함수 K 의 모양(정규분포, 삼각형 등)이 중요한가요? A. 별로 안 중요합니다. 커널의 모양보다는 **대역폭 h 의 크기**가 결과에 100배는 더 큰 영향을 미칩니다. 그냥 미분하기 편한 가우시안 커널을 쓰면 됩니다.
- Q2. 왜 비모수적 방법은 데이터가 많이 필요한가요? A. 모수적 방법은 ”종 모양이다”라는 강력한 힌트(가정)를 가지고 시작하므로, 중심(μ)과 폭(σ)만 맞추면 됩니다. 하지만 비모수적 방법은 아무런 힌트 없이 백지상태에서 지도를 그려야 하므로, 빈 공간을 채우기 위해 훨씬 많은 정보(데이터)가 필요합니다.

Next Step: 우리는 데이터의 분포(밀도)를 알아냈습니다. 이제 이 고차원 데이터를 다루기 쉽게 압축할 수는 없을까요? 다음 시간에는 비모수 통계의 또 다른 축인 **Unit 5: 주성분 분석 (PCA, Principal Component Analysis)**을 통해 데이터의 차원을 축소하고 숨겨진 구조를 시각화하는 법을 배웁니다.

[요약] Unit 4 핵심 요약

- **비모수 통계:** 특정 분포(\mathcal{N})를 가정하지 않고 함수 f 자체를 추정한다.
- **히스토그램:** 쉽지만 각지고, 구간 설정에 민감하다.
- **KDE:** 커널(K)과 대역폭(h)을 이용해 스무딩하는 기법.
- **Bias-Variance Trade-off:** h 가 작으면 과적합(Variance \uparrow), 크면 과소적합(Bias \uparrow).
- **최적의 h :** $n^{-1/5}$ 에 비례하여 설정한다.

[a4paper, 11pt]book fonts spec amsmath, amssymb, amsthm geometry graphicx xcolor
 hyperref booktabs array bm
 left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 3: Linear Regression (정답 Y 를 맞추는 지도 학습)
- Unit 4: Density Estimation (데이터의 분포 모양 추정)
- Unit 5: Principal Component Analysis (현재 단원: 데이터 구조 파악 및 압축)
 - 5.1 Philosophy: Variance is Information
 - 5.2 Geometric Interpretation
 - 5.3 Covariance Matrix & Eigen-decomposition
 - 5.4 Dimension Reduction Process
- Unit 6: Clustering (K-means, Hierarchical)

Chapter 11

Unit 5. 주성분 분석 (PCA)

Unit 4에서 우리는 데이터의 밀도(Density)를 추정했습니다. 하지만 데이터의 차원(p)이 커지면 밀도 추정은 급격히 어려워집니다(차원의 저주). ”변수가 100개나 되는데, 이걸 다 써야 하나? 중요한 것 몇 개만 추릴 순 없을까?” 이 질문에 답하기 위해, 우리는 데이터의 핵심 정보만 남기고 꺽데기를 버리는 **차원 축소(Dimensionality Reduction)**의 세계로 들어갑니다.

□ 개요 (Overview)

PCA는 고차원 데이터의 정보를 최대한 보존하면서 저차원으로 압축하는 기법입니다. **”분산(Variance)이 곧 정보(Information)”**라는 철학을 바탕으로, 데이터의 분산을 최대화하는 새로운 축(주성분)을 찾습니다. 이 과정은 수학적으로 **공분산 행렬의 고유값 분해(Eigenvalue Decomposition)** 문제로 귀결됩니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
Variance (분산)	데이터가 펴진 정도. PCA에서는 이를 ’정보량’으로 해석함.
Principal Component (PC)	데이터를 가장 잘 설명하는 새로운 좌표축 (방향).
Orthogonality (직교성)	PC1과 PC2는 수직이다. 즉, 서로 겹치는 정보가 없다.
Eigenvector (v)	새로운 축의 ’방향’.
Eigenvalue (λ)	그 축이 설명하는 정보의 ’크기’ (분산의 양).

—

11.0.1 1. 차원 축소의 철학: 분산은 정보다

[개념] 개념 1: 심장 박동기 모니터

한 줄 요약: 변화가 없는(분산 0) 데이터는 죽은 데이터입니다. 요동치는(분산 큰) 데이터가 살아있는 정보입니다.

1) 직관적 비유

병원 모니터의 심전도 그래프를 상상해 보세요.

- **일직선 ($Var = 0$):** 환자가 사망했습니다. 정보가 없습니다.

- **위아래로 뛸 ($Var > 0$):** 환자가 살아있습니다. 분석할 정보가 있습니다.
- **결론:** 데이터가 넓게 퍼져 있을수록, 서로를 구별할 수 있는 정보가 많습니다. 따라서 우리는 **분산이 가장 큰 방향**을 찾고 싶습니다.

2) 차원의 저주 탈출

변수(차원)가 많으면 계산도 힘들고 시각화도 불가능합니다. 정보량이 적은(분산이 작은) 변수는 과감히 버려서, 핵심만 남기는 것이 목표입니다.

11.0.2 2. 기하학적 해석: 최적의 좌표축 찾기

(개념) 개념 2: 럭비공 돌리기

한 줄 요약: 데이터 구름(Cloud)의 모양에 맞춰서 좌표축을 회전시킵니다. 가장 길쭉한 쪽이 x 축(PC1), 그 다음 넓은 쪽이 y 축(PC2)이 되도록요.

1) 첫 번째 주성분 (PC1)

데이터 점들을 가장 길게 통과하는 직선입니다.

- 기하학적으로는, 데이터 점들을 직선에 투영(Projection)했을 때 그 그림자들이 가장 넓게 퍼지는 선입니다.
- 동시에, 데이터 점들과 직선 사이의 수직 거리(Reconstruction Error)를 최소화하는 선입니다.

2) 두 번째 주성분 (PC2)

PC1과 반드시 **수직(Orthogonal)**이면서, 남은 분산을 가장 잘 설명하는 방향입니다.

- **Why Orthogonal?** PC1이 이미 설명한 정보와 겹치지 않는, 완전히 **새로운 정보(Independent Information)**만 담기 위해서입니다. 이를 통해 변수 간의 상관관계를 제거(Decorrelation)합니다.
-

11.0.3 3. 수학적 엔진: 공분산 행렬과 고유값 분해

(개념) 개념 3: 모든 것은 고유값 문제로 통한다

한 줄 요약: ”분산을 최대화하는 축을 찾아라”라는 복잡한 미적분 문제가, 놀랍게도 선형대수의 $Av = \lambda v$ 를 푸는 문제로 바뀝니다.

Step 1: 표본 공분산 행렬 (S)

데이터 X (중심화됨, Mean=0)에 대해:

$$S = \frac{1}{n-1} X^T X$$

이 행렬은 변수들끼리 얼마나 같이 움직이는지(Correlation)를 담고 있습니다. 우리의 목표는 이 상관성을 없애는(대각화하는) 것입니다.

Step 2: 스펙트럼 정리 (Spectral Theorem)

S 는 대칭 행렬 (Symmetric Matrix)입니다. 선형대수학의 스펙트럼 정리에 의해, 대칭 행렬은 반드시 **실수 고유값**을 가지며, 그 **고유벡터들은 서로 직교**합니다.

- 즉, 우리가 억지로 직교하는 축을 찾을 필요가 없습니다. 수학적으로 이미 직교하는 축(고유벡터)이 존재함이 보장됩니다.

Step 3: 고유값 분해 (The Solution)

$$Sv_j = \lambda_j v_j$$

- **고유벡터 (v_j):** 우리가 찾던 새로운 축의 **방향** (Principal Component).
 - **고유값 (λ_j):** 그 축 방향으로의 **분산의 크기** (정보량).
-

11.0.4 4. 차원 축소의 실행 (Dimension Reduction)

[개념] 개념 4: 정보의 압축

한 줄 요약: 중요도가 낮은(고유값이 작은) 축은 과감히 버립니다.

절차

- 고유값들을 크기순으로 나열합니다: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$.
- **Scree Plot**을 그려보거나, 누적 기여율 (Cumulative Variance Ratio)을 확인합니다.

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^p \lambda_j} \geq 0.95 \quad (95\% \text{ 정보 보존})$$

- 상위 k 개의 고유벡터만 남기고 나머지는 버립니다.
 - 데이터를 이 k 개의 축으로 투영 (Projection)하여 차원을 $p \rightarrow k$ 로 줄입니다.
-

11.1 실전 시나리오: 넥슨 유저 세분화 (User Segmentation)

(시나리오) Scenario: 너무 많은 게임 지표들

당신은 넥슨의 데이터 분석가입니다. 유저들의 플레이 패턴을 분석하려고 보니 지표가 너무 많습니다. (변수 20개: 총 퀸 수, 데스 수, 어시스트, 가한 피해량, 받은 피해량, 힐량, 타워 철거 수, CS, 골드 획득량...) 변수가 20개라 시각화도 안 되고, 군집화 (Clustering)를 돌려도 결과가 엉망입니다.

- **문제:** 변수 간 상관관계가 너무 높습니다. (퀸 많이 하면 골드도 많고 피해량도 높음).
- **PCA 적용:** 20차원 데이터를 입력하여 주성분 분석을 수행합니다.

3. **결과 해석:**

- **PC1 (설명력 60%):** 퀼, 딜량, 골드 등 '전투/성장' 관련 변수들이 모두 양의 가중치를 가짐. → 해석: **"실력(Skill) 지표"**
 - **PC2 (설명력 20%):** 받은 피해량, 힐량은 높고 퀼은 낮음. → 해석: **"탱킹/서포팅 성향"**
4. **활용:** 20개의 변수 대신, (Skill, Style)이라는 단 2개의 변수로 유저를 2차원 평면에 찍어봅니다. 유저 그룹이 명확하게 보이기 시작합니다.
-

11.2 자주 묻는 질문 (FAQ)

- Q1. 축을 돌리면 데이터의 의미가 변하지 않나요? A. 네, 변합니다. 원래 변수는 "키", "몸무게"처럼 물리적 의미가 명확했지만, PCA로 만들어진 PC1은 $0.7 \times \text{키} + 0.3 \times \text{몸무게}$ 같은 섞인 값이 됩니다. 이를 "크기(Size)"라고 해석할지는 분석가의 몫입니다. **해석력(Interpretability)을 잃는 대신, 요약력(Summarization)을 얻는 것**입니다.
- Q2. 데이터 스케일링(Scaling)을 꼭 해야 하나요? A. 필수입니다! 만약 키는 cm(170), 몸무게는 kg(60) 단위인데, 연봉을 원(50,000,000) 단위로 넣으면, 연봉의 분산이 압도적으로 커서 PC1이 그냥 "연봉" 축이 되어버립니다. 반드시 모든 변수를 표준화(Standardization, 평균 0 분산 1)한 뒤 PCA를 돌려야 공평한 비교가 됩니다.

Next Step: 우리는 PCA를 통해 데이터를 압축하고 시각화했습니다. 이제 이 잘 정리된 데이터들을 끼리끼리 묶어볼 수 없을까요? 다음 **Unit 6**에서는 비지도 학습의 또 다른 핵심인 **군집화(Clustering: K-means, Hierarchical)**를 통해 데이터 속에 숨어있는 그룹을 찾아냅니다.

(요약) Unit 5 핵심 요약

- **철학:** 분산(Variance)이 클수록 정보가 많다.
- **목표:** 정보를 최대한 보존하면서 변수의 개수(차원)를 줄이자.
- **방법:** 공분산 행렬 S 의 고유벡터(Eigenvector)를 찾는다.
- **PC의 성질:** 서로 직교(Orthogonal)하며, 상관관계가 제거된다.
- **주의:** 해석이 어려워질 수 있으며, 스케일링(정규화)이 필수적이다.

(a4paper, 11pt) book fontspec amsmath, amssymb, amsthm geometry graphicx xcolor
 hyperref booktabs array bm
 left=25mm, right=25mm, top=30mm, bottom=30mm

Contents

Course Structure & Current Focus

- Unit 8: Linear Regression (변수 p 가 작고 고정됨)
- Unit 11: PCA (변수를 합쳐서 줄임)
- Unit 12: High-dimensional Regression (현재 단원: 변수가 너무 많음 $p \gg n$)
 - 12.1 The Curse of Dimensionality ($p > n$)
 - 12.2 Sparsity Assumption (Occam's Razor)
 - 12.3 Regularization: Ridge vs LASSO
 - 12.4 Geometry of LASSO (왜 0이 되는가?)
- Unit 13: Robustness (이상치 대응)

Chapter 12

Unit 12. 고차원 회귀와 희소성 (High-dimensional Regression & Sparsity)

Unit 8에서 우리는 $Y = X\beta + \epsilon$ 을 풀기 위해 최소자승법(OLS)을 배웠습니다. 그때는 데이터(n)가 변수(p)보다 충분히 많았습니다. 하지만 현대의 유전체학이나 텍스트 분석 데이터는 변수는 수만 개인데 샘플은 수십 개뿐입니다. "미지수가 식보다 많은 연립방정식", 과연 풀 수 있을까요? 이 불가능해 보이는 문제를 풀기 위해 우리는 '희소성(Sparsity)'이라는 강력한 믿음을 도입합니다.

□ 개요 (Overview)

이 단원에서는 변수의 수(p)가 데이터의 수(n)보다 큰 고차원 상황($p \gg n$)에서 발생하는 문제점(Singularity, Overfitting)을 진단합니다. 이를 해결하기 위해 **"진짜 중요한 변수는 소수일 것이다"**라는 희소성 가정을 도입하고, **LASSO(L1 Regularization)**을 통해 변수 선택(Variable Selection)과 추정(Estimation)을 동시에 수행하는 기하학적 원리를 배웁니다.

□ 핵심 용어 사전

용어 (Term)	직관적 의미 (Meaning)
High-dimensional ($p \gg n$)	방정식보다 미지수가 더 많은 난감한 상황.
Sparsity (희소성)	"진짜 범인은 이 수만 명 중에 딱 한두 명이다."
Regularization (규제)	정답을 맞추는 것뿐만 아니라, 답안지(계수)가 깔끔해야 점수를 주는 채점 방식.
Ridge (L2)	계수들을 0에 가깝게 만듦 (Shrinkage).
LASSO (L1)	중요하지 않은 계수를 아예 0으로 만듦 (Selection).

—

12.0.1 1. 문제의 본질: $p > n$ 일 때 발생하는 일

[개념] 개념 1: 식보다 미지수가 많다

한 줄 요약: 해가 하나로 정해지지 않고 무수히 많으며, 컴퓨터는 그중에서 "노이즈까지 외워버린" 최악의 해를 선택하게 됩니다.

1) 수학적 불능 (Singularity)

OLS 해 $\hat{\beta} = (X^T X)^{-1} X^T Y$ 를 구하려면 역행렬이 존재해야 합니다. 하지만 $p > n$ 이면 $X^T X$ 는 **비가역 행렬(Singular Matrix)**이 되어 역행렬이 없습니다.

- 예: $x + y + z = 10$. 미지수는 3개, 식은 1개. $(1, 1, 8), (2, 3, 5) \dots$ 해가 무한함.

2) 완벽한 과적합 (Perfect Overfitting)

변수가 많으면 모델은 모든 데이터 포인트를 완벽하게 지나가는 구불구불한 곡선을 만들 수 있습니다. (Training Error = 0). 하지만 새로운 데이터가 들어오면 예측력이 0에 수렴합니다.

12.0.2 2. 희소성 가정 (Sparsity Assumption)

[개념] 개념 2: 오컴의 면도날 (Occam's Razor)

한 줄 요약: ”현상은 복잡해 보이지만, 실제로 결과(Y)를 조종하는 핵심 변수($\beta_j \neq 0$)는 극소수(s)일 것이다.”

수학적 정의

참값 벡터 β^* 는 p 차원이지만, 그중 0이 아닌 성분의 개수 s 는 n 보다 훨씬 작다고 가정합니다 ($s \ll n$).

$$\|\beta^*\|_0 := \sum_{j=1}^p \mathbb{I}(\beta_j^* \neq 0) = s \ll n$$

이 믿음이 있어야만 우리는 p 차원이라는 거대한 우주에서 s 개의 별을 찾을 수 있습니다.

12.0.3 3. 규제화 (Regularization): Ridge vs LASSO

[개념] 개념 3: 벌점(Penalty) 시스템

한 줄 요약: ”문제를 잘 푸는 것(RSS 최소화)”도 중요하지만, ”답안지가 복잡하면(계수가 크면)” 감점을 시키는 새로운 채점 기준을 도입합니다.

최적화 문제의 변화

$$\hat{\beta} = \operatorname{argmin}_{\beta} (\|Y - X\beta\|^2 + \lambda \cdot \text{Penalty}(\beta))$$

λ 는 규제의 강도입니다. λ 가 클수록 계수를 더 강하게 억누릅니다.

1) Ridge 회귀 (L2 Norm)

$$\text{Penalty} = \sum \beta_j^2$$

- 효과: 모든 계수를 균일하게 줄여줍니다 (Shrinkage).
- 한계: 0이 되지는 않습니다. (0.0001 같은 작은 값으로 살아남음). 변수 선택 기능이 없습니다.

2) LASSO 회귀 (L1 Norm)

$$\text{Penalty} = \sum |\beta_j|$$

- 효과: 덜 중요한 변수의 계수를 **정확히 0**으로 만듭니다 (Selection).
- 의의: 모델 해석이 용이해지고, 희소성 가정을 구현할 수 있습니다.

—

12.0.4 4. LASSO의 기하학적 해석 (The Geometry of Sparsity)

(개념) 개념 4: 둥근 원과 뾰족한 마름모

한 줄 요약: L1 페널티 영역(마름모)은 모서리가 뾰족해서, 최적해가 축(Axis) 위에서 형성될 확률이 매우 높습니다. 축 위에 있다는 것은 해당 좌표값이 0이라는 뜻입니다.

이 그림은 Unit 12의 하이라이트입니다.

- **등고선 (RSS):** 타원형으로 퍼져나가는 ”데이터가 원하는 답”.
- **제약 영역 (Penalty):** 원점 주변의 ”우리가 허용하는 답의 범위”. (λ 에 의해 결정됨).

상황 비교

1. **Ridge (원형):** $\beta_1^2 + \beta_2^2 \leq C$. 타원이 둥근 원과 만납니다. 점점은 주로 1사분면 어딘가($\beta_1 \neq 0, \beta_2 \neq 0$)에서 생깁니다. 0이 되지 않습니다.
2. **LASSO (마름모):** $|\beta_1| + |\beta_2| \leq C$. 타원이 퍼져나가다가 **뾰족한 모서리(Corner)**에 먼저 닿습니다. 이 모서리는 축 위에 존재합니다 (예: $\beta_1 = C, \beta_2 = 0$). 즉, 자연스럽게 β_2 가 0이 됩니다. 이것이 LASSO가 변수를 제거하는 수학적/기하학적 원리입니다.

—

12.0.5 5. 이론적 보장 (Theoretical Guarantees)

(개념) 개념 5: 신(Oracle)과의 대결

한 줄 요약: LASSO는 우리가 ”진짜 중요한 변수가 무엇인지 미리 알고 있을 때(Oracle)” 수행하는 추정만큼이나 훌륭한 성능을 냅니다.

오라클 부등식 (Oracle Inequalities)

LASSO의 오차는 다음을 만족합니다 (확률적으로).

$$\|\hat{\beta}_{LASSO} - \beta^*\|^2 \leq C \frac{s \log p}{n}$$

- 오차는 전체 변수 p 가 아니라, **중요 변수 개수 s **에 비례합니다.
- $\log p$ 는 p 가 아주 커져도 매우 천천히 증가하므로, $p \gg n$ 상황에서도 오차가 작게 유지됩니다.

—

12.1 실전 시나리오: 넥슨 게임 로그 분석

(시나리오) Scenario: 이탈 유저를 찾아라

당신은 넥슨의 데이터 분석가입니다. '메이플스토리' 유저 중 누가 다음 달에 게임을 접을지(Churn) 예측하고 싶습니다.

1. **데이터 상황:**

- $n = 500$ (최근 이탈한 유저 샘플 수)
 - $p = 10,000$ (수집된 행동 로그 종류: 점프 횟수, 채팅 수, 물약 사용, 특정 맵 방문 등 무수히 많음)
2. **문제:** $p \gg n$ 이므로 일반 회귀분석을 돌리면 모든 변수가 중요하다고 나오거나 에러가 납니다.
 3. **LASSO 적용:** L1 규제를 걸고 회귀분석을 수행합니다.
 4. **결과:** 10,000개의 변수 중 9,995개의 계수가 0이 되었습니다.
 - 살아남은 변수 5개: [길드 탈퇴 여부], [친구 목록 삭제 수], [고가 아이템 판매], [접속 시간 급감], [고객센터 불만 접수]
 5. **인사이트:** "수만 가지 행동 중, 이 5가지만 모니터링하면 이탈을 90% 예측할 수 있구나!" → 해당 유저들에게 쿠폰 발송(Action Item).
-

12.2 자주 묻는 질문 (FAQ)

Q1. 중요한 변수가 0이 되어버리면 어떡하나요? A. 그럴 위험이 있습니다. 만약 변수들끼리 상관관계가 높다면(예: 원발 움직임 vs 오른발 움직임), LASSO는 둘 중 하나만 남기고 하나는 0으로 죽여버립니다. 이를 방지하기 위해 Ridge와 LASSO를 섞은 **Elastic Net**을 사용하기도 합니다.

Q2. λ (규제 강도)는 어떻게 정하나요? A. λ 가 너무 크면 다 0이 되고(Underfitting), 너무 작으면 OLS랑 같아집니다(Overfitting). 보통 **교차 검증(Cross-Validation)**을 통해 예측 에러가 가장 작은 최적의 λ 를 찾습니다.

Next Step: 우리는 고차원 데이터에서 희소성을 이용해 '중요한 변수'를 찾아냈습니다. 그런데 만약 데이터에 단순히 변수가 많은 게 아니라, **악의적인 노이즈(Outlier)**가 섞여 있다면 어떡할까요? 다음 **Unit 13**에서는 데이터가 오염되어도 흔들리지 않는 **로버스트 통계(Robust Statistics)**를 배웁니다.

[요약] Unit 12 핵심 요약

- **문제:** $p \gg n$ 이면 OLS는 불가능하거나 과적합된다.
- **가정:** 희소성(Sparsity). 중요한 변수는 소수(s)다.

- **LASSO (L1):** 절댓값 페널티를 사용하여 변수 선택과 추정을 동시에 한다.
- **기하학:** L1의 뾰족한 모서리(Corner)가 최적해를 0으로 유도한다.
- **이론:** $s \log p/n$ 속도로 수렴하여 고차원에서도 작동한다.