Probability & Statistics

통계적 추론

강사 양석환

통계적 추론 개요

통계적 추론(statistical inference)

- 통계적 추론(statistical inference)
  - 모집단으로부터 생성된 관측데이터로부터 모집단의 분포의 속성(중심위치, 산포 등)에 대한 가설을 검증하거나 추정하는 등의 과정
  - 대표적인 통계적 추론 과정은 추정과 검정임

추정 (Estimation)

- 추정(Estimation) 이란?
  - 표본 데이터를 이용해 모집단의 특성 값인 모수(Population Parameter)를 확률적으로 추론하는 과정
  - 대표적인 추정 방법
    - MME (Method of Moment Estimation, 적률 추정량)
    - MLE (Maximum Likelihood Estimation)
    - MAP (Maximum a Posteriori Estimation, 최대 사후 추정)

- **MME (Method of Moment Estimation, 적률 추정량)**
  - **적률(Moment) 개념을 사용하여 추정하는 방법**
    - 적률: 확률 분포의 위치나 모양 따위의 특성을 나타내는 기댓값
  - **확률변수가 확률밀도(질량)함수 f(x)의 표본이라고 했을때**
    - k차 표본 적률(Simple Moment)과 모적률(Population Moment)을 일치 시킴으로써 모수를 추정할 수 있음
    - k차의 표본 적률

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

    - k차의 모적률

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k = m_k = M_k = E\left(X^k\right)$$

적률 추정법

표본 적률과 모적률을 일치 시킴으로써 모수를 추정하는 방법

Artificial Intelligence & Data Analysis Laboratory

MME (적률 추정량)

- 적률 추정량 구해보기 1
  - 확률 변수 $X_1, X_2, \dots, X_n$이 이항 분포 $Binomial(1,\ p\ )$의 표본일 때, p 와 $p(1-p)$ 의 적률추정량
    - p의 MME → p 는 이항 분포를 따르는 확률 변수 X의 기댓값에 해당함
    - 따라서 p 는 1차 모적률에 해당함 → $M_1 = E\left(X^1\right) = E(X) = p$
    - 1차 표본 적률 계산 → $m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^1 = \frac{1}{n}\sum_{i=1}^{n} X_i$
    - 1차 모적률과 표본적률 일치시키기 → $\hat{p} = M_1 = m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ → $\therefore \hat{p} = \bar{X}$
    - $p(1-p)$ 의 MME → 모적률 → $p(1-p) = Var(X) = E\left(X^2\right) - [E(X)]^2 = M_2 - M_1^2$
    - 따라서 표준 적률을 이용해 MME를 구하면

$$p\widehat{(1-p)} = M_2 - M_1^2 = m_2 - m_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

실습 코드 참고

AiDA Lab.

MME (적률 추정량)

- 적률 추정량 구해보기 2
  - 확률 변수 $X_1, X_2, \dots, X_n$이 감마 분포 $\Gamma(\alpha, \beta)$ 의 표본일 때, α 와 β 의 적률추정량 구하기
    - $\Gamma(\alpha, \beta)$의 기댓값은 $\alpha\beta$ 이고 분산은 $\alpha\beta^2$ 임
    - 따라서 모적률 → $E(X) = M_1 = \alpha\beta \rightarrow Var(X) = E(X^2) - [E(X)]^2 = M_2 - M_1^2 = \alpha\beta^2$
    - 정리하면 → $M_1 = \alpha\beta$ 이고 $M_2 - M_1^2 = \alpha\beta^2$
    - 위의 식을 α 와 β 에 대하여 표현하면 → $\alpha = \frac{M_1^2}{M_2 - M_1^2}$, $\beta = \frac{M_2 - M_1^2}{M_1}$
    - 따라서 표본 적률과 모적률을 일치시켜 MME를 구하면

$$\hat{\alpha} = \frac{M_1^2}{M_2 - M_1^2} = \frac{m_1^2}{m_2 - m_1^2} = \frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2}, \qquad \hat{\beta} = \frac{M_2 - M_1^2}{M_1} = \frac{m_2 - m_1^2}{m_1} = \frac{\frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2}{\bar{X}}$$

실습 코드 참고

AiDA Lab.

MLE (최대 가능도 추정)

- MLE (Maximum Likelihood Estimation)
  - 주어진 샘플 데이터의 Likelihood(가능도, 우도)를 이용하여 모집단의 분포를 추정하는 방법
    - 가능도를 최대화하는 방향으로 추정함
    - 이때 가장 그럴듯한 추정을 선택함
      - 분포의 파라미터가 될 수 있는 후보 값은 여러 가지가 있는데 해당 샘플을 얻을 가능성이 높은 가장 그럴듯한 파라미터를 추정한다는 의미

정규분포에서
평균을 추정한다는 것은 → 분포의 중심이 어디인지를 추정한다는 의미
분산을 추정한다는 것은 → 분포가 중심으로부터 얼마나 넓게 퍼져 있는지를 추정한다는 의미

- 확률과 가능도의 차이
  - 확률: 고정된 확률 분포에서 우리가 관심있는 영역을 의미함
  - 가능도: 우리가 얻은 샘플은 고정되어 있는 상황에서 확률 분포를 정의하는 파라미터가 바뀔 때마다 측정하는 값

실습 코드 참고

Artificial Intelligence & Data Analysis Laboratory

MLE (최대 가능도 추정)

- **최대 가능도의 추정이란?**
  - **파라미터 θ에 대한 가능도 함수 $L(\theta)$을 최대화할 수 있는 θ값을 구하는 것을 의미함**
  - **최대 값을 가지는 지점**
    - 주어진 수식(데이터)를 1계 미분했을 때 $\frac{dL(\theta)}{d\theta} = 0$ 이 되는 지점 → 이때의 θ를 구하면 됨
  - **문제점**
    - 일반적인 이산 확률 분포의 식은 많은 경우, 확률의 곱으로 표현됨 → 미분이 어려운 경우가 많음
    - 이때는 가능도 함수에 자연로그를 적용하여 로그 가능도 함수 $\log_e L(\theta)$를 만들어서 적용할 수 있음
    - 로그 가능도 함수로 바꿔더라도 이 식을 최대로 만들어주는 θ가 가능도 함수 $L(\theta)$도 함께 최대로 만들어주기 때문에 문제 없음

$$\frac{d}{d\theta}\log_e L(\theta) = 0$$

log를 적용하는 이유
log를 적용하면 곱셈을 덧셈으로 바꿀 수 있기 때문에
고차 방정식을 단숨에 1차 방정식으로 만들어 줌으로써
수식을 다루는 난이도를 낮출 수 있음

Artificial Intelligence & Data Analysis Laboratory

MLE (최대 가능도 추정)

- AI에서 MLE는 어떻게 활용되나?
  - MLE는 이미 확보한 데이터를 사용해서 미처 발견하지 못한 확률 모델의 파라미터를 추정할 때 사용하는 통계기법이므로
  - 과거의 데이터로부터 미래를 예측할 때 많이 사용함

MAP(최대 사후 추정)

- MAP(Maximum a Posteriori Estimation, 최대 사후 추정)
  - 베이즈 추정
    - 데이터의 빈도에 따라 계산하는 일반적인 빈도주의에서는 확률분포의 파라미터를 상수로 취급
    - 베이즈 추정에서는 파라미터 θ가 확률 변수로 취급 → θ 가 확률 변수이므로 θ의 확률 $P(\theta)$의 계산이 가능함
    - 사건이 발생하기 전 파라미터 θ의 확률인 $P(\theta)$ → 사전 확률 밀도 함수(=사전분포)
    - 확률 변수가 연속 확률 변수라고 가정하고, 다음과 같이 지정
      - 파라미터 θ의 사전 확률 밀도 함수를 $f(\theta)$
      - 파라미터 θ가 주어졌을 때 표본 x가 관측될 확률 밀도 함수를 $f(x|\theta)$
    - 파라미터 θ가 주어졌을 때 표본 x가 얻어질 확률 → $P(x|\theta)$
    - 파라미터 θ와 표본 x가 동시에 얻어질 확률 → $f(\theta, x) = f(x|\theta)f(\theta)$

Artificial Intelligence & Data Analysis Laboratory

MAP(최대 사후 추정)

- 파라미터 θ가 주어졌을 때 표본 x가 관측될 확률 밀도 함수를 $f(x|\theta)$라고 했는데 표본 x가 주어진다면 파라미터 θ가 얻어질 확률은?

$$f(\theta|x) = \frac{f(\theta, x)}{f(x)} = \frac{f(x|\theta)f(\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

- 위 식에서 $f(\theta|x)$를 사후 확률 밀도 함수(Posterior Probability Density Function, =사후 분포)
- 위 식에서 $f(x)$는 파라미터 θ의 영향을 받지 않음 → $f(x)$는 파라미터 θ와 무관 → $f(x)$는 상수취급

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$

- 사후 확률 밀도 함수 $f(\theta|x)$는 상수인 $f(x)$와 무관하며 오직 $f(x|\theta)f(\theta)$의 값에만 영향을 받음
  → 베이즈 추정의 핵심 개념의 하나

Artificial Intelligence & Data Analysis Laboratory

MAP(최대 사후 추정)

- MAP 적용의 목적은 주어진 표본 x를 이용하여 파라미터 θ를 추정하는 것

→ MAP는 사후 확률 밀도 함수 $f(\theta|x)$를 최대화 시키는 파라미터 θ를 추정값으로 구함

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} f(\theta|x) = \underset{\theta}{\operatorname{argmax}} \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \propto \underset{\theta}{\operatorname{argmax}} f(x|\theta)f(\theta)$$

실습 코드 참고

비편향(Unbiased) 추정

- 하나의 통계량이 있어서 이로부터 모수를 추정한다고 할 때
  - 표본 평균 통계량
    - 모집단 분포로부터 추출된 표본으로부터 생성된 하나의 확률 변수
    - 표본 평균 통계량의 분포: $N(\theta, \frac{1}{n}1^2)$
  - 모집단의 모수 추정
    - 모집단으로부터 무한히 많은 샘플을 뽑아서 그 분포를 구성한 후 추정하면 되지만 현실적으로 불가능하므로
    - 단 한 번의 표본으로부터 모집단의 모수를 추정해야 함
    - 이때 통계량은 모수에 대한 하나의 추정량이 되며, 표본으로부터 계산된 하나의 값은 추정값이 됨
    - 모수를 추정하는 것이므로 추정된 기댓값은 모수와 일치하여야 하고, 이를 만족하는 값을 비편향 추정량이라고 함
    - 그러나 비편향 추정은 현실적으로 어려우므로 가능도(Likelihood)를 이용한 방식(MLE 등)이 주로 사용됨

기타 추정 방법

- 점 추정 (Point Estimation)
  - **하나의 값으로 모집단 특성을 추정하는 방법으로 모수의 단일 추정치를 제공함**
    - 예: 표본 평균을 사용하여 모집단 평균을 추정]
  - **점 추정 방법의 종류**
    - **표본 평균 추정:** 모집단의 평균을 추정하는데 표본의 평균을 사용함. 표본 평균은 모집단 평균의 추정치로 사용됨.
    - **표본 분산 추정:** 모집단의 분산을 추정하는데 표본의 분산을 사용함. 표본 분산은 모집단 분산의 추정치로 사용됨.
    - **표본 비율 추정:** 모집단 비율을 추정하는데 표본의 비율을 사용함. 표본 비율은 모집단 비율의 추정치로 사용됨.

Artificial Intelligence & Data Analysis Laboratory

기타 추정 방법

- 구간 추정 (Interval Estimation)
  - 추정 값이 어느 범위 내에 있을 가능성을 나타내는 방법
  - 구간 추정은 추정의 불확실성을 고려하는 데 도움이 됨
  - 신뢰구간 (Confidence Interval)
    - 대표적인 구간 추정 방법
    - 신뢰구간은 모수가 어떤 범위 내에 있을 것으로 추정되는 것을 나타냄
    - 일반적으로 95% 또는 99% 신뢰수준을 사용하며, 이 수준에서 모수가 포함될 가능성을 나타냄
      - 예: "평균 IQ는 95% 신뢰구간에서 90에서 110 사이에 있을 것으로 추정됩니다."

통계적 가설

통계적 가설

- 성적 등과 같은 어떤 확률 변수가 주어졌을 때
  - 과거의 다양한 예를 통해서 볼 때 $\theta = 75$라고 평가되었으나
  - 어떤 이유 또는 무엇인가의 변화에 의해 모평균의 값이 75보다 큰 값이라고 의심이 되는 경우,
  - 그러나 이에 대한 데이터 기반의 물증은 존재하지 않는 상황
  - 이때, $\theta > 75$일 것이라는 추측(Conjecture) 또는 주장을 통계 가설(Statistical Hypothesis)이라고 부름
  - 그러나 $\theta > 75$라는 가설은 틀릴지도 모르며 그 반대인 $\theta \leq 75$가 맞을 수도 있음
  - 따라서 통계 가설은 2가지가 존재함: 대립가설, 귀무가설

가설 검정 (Hypothesis Testing)

- 검정(Testing)이란?
  - 데이터 뒤에 숨어있는 확률 변수의 분포에 대한 가설이 맞는지 틀리는지 정량적으로 증명하는 작업
- 가설과 검정
  - 특정한 확률 분포를 가진 확률 변수로 데이터를 모형화 하면 모수를 추정할 수 있음
  - 모수의 추정 후에는 데이터 뒤에 숨어있는 확률 변수가 정말로 그 모숫값을 가졌는지 검증해 보아야 함 (해당 확률 변수가 그 모숫값을 가졌다는 주장을 논리적으로 증명해야 함)
  - 가설(Hypothesis) : 확률 분포에 대한 어떤 주장
  - 통계적 가설 검정(Statistical Hypothesis Testing, =검정): 가설을 증명하는 행위
  - 모수 검정(Parameter Testing): 확률 분포의 모숫값이 특정한 값을 가진다는 가설을 검정하는 것

가설 검정 (Hypothesis Testing)

- **가설 검정**
  - 특정 가설에 대한 증거를 수집하고, 그 가설이 참인지 거짓인지를 결정하는 과정
  - 일반적으로 "귀무 가설(null hypothesis)"과 "대립 가설(alternative hypothesis)"을 설정하고, 수집된 데이터를 통해 귀무 가설을 기각하거나 채택하는 결정을 내림
  - 예시: 어떤 새로운 약의 효과를 조사할 때
    - 귀무 가설: "이 약은 아무런 효과가 없다"
    - 대립 가설: "이 약은 효과가 있다"

Artificial Intelligence & Data Analysis Laboratory

검정 통계량(Test Statistics)

- 귀무가설이 맞거나 틀렸다는 것을 증명하려면 증거가 있어야 함
  - 예시
    - '어떤 병에 걸렸다'라는 가설을 증명하려면 환자의 혈액을 채취하여 혈액 내의 특정 성분의 수치를 측정해야 한다고 가정하면, 이때 해당 수치가 바로 검정 통계량이 됨
    - '어떤 학생이 우등상을 받을 수 있는 우등생이다'라는 가설을 증명하려면 시험 성적을 측정하면 되며, 이 시험 성적을 검정 통계량이라고 할 수 있음
- 검정 통계량
  - 표본 데이터 집합을 입력으로 계산되는 함수의 값
  - 확률 변수 X의 표본에서 계산된 함수의 값이므로 어떤 값이 나올지 정확하게 예측할 수 없음
  - 따라서 검정 통계량 t도 검정 통계량 확률 변수 T 라는 새로운 확률 변수의 표본으로 볼 수 있음

유의 확률(p-value)

- 유의 확률
  - 확률 분포와 확률 분포의 표본 값 n개가 주어졌을 때, 그 확률 분포에서 해당 표본 값 혹은 더 희귀한 값이 나올 수 있는 확률
  - 어떤 표본 데이터가 해당 확률 분포에서 나오기 쉬운 값인지, 나오기 어려운 값인지를 숫자로 정량화 한 것
  - 유의 확률의 값은 확률 밀도 함수에서 표본 값을 기준으로 만들어진 양측 꼬리 부분에 해당하는 영역의 면적
  - 누적 확률 분포 함수를 사용할 경우, 유의 확률은 통계량 분포의 양 끝단의 면적을 구하기 때문에 양측 검정 유의 확률(Two-sided Test p-value)이라고 부름
  - 검정의 관점에서 유의 확률은 '귀무가설이 맞음에도 불구하고 현재 검정 통계량 값과 같은, 혹은 대립가설을 더 옹호하는 검정 통계량 값이 나올 확률'이라고 봄

AIDA Lab.

유의 확률(p-value)

- 유의 수준과 기각역
  - 유희 확률 값이 아주 작다는 것 = 귀무가설이 맞다는 가정 하에 현재의 검정 통계량 값이 나올 가능성이 매우 적다는 의미 → 유의 확률 값이 아주 작으면 귀무가설을 기각하고 대립가설을 채택할 수 있음
  - 유의 수준: 계산된 유의 확률 값에 대해 귀무가설을 기각할 것인지 채택할 것인지를 결정할 수 있는 기준값
    - 유의 확률이 유의 수준보다 작으면 귀무가설을 기각하고 대립가설을 채택함
  - 기각역: 유의 수준에 대해 계산된 검정 통계량
    - 기각역을 알고 있다면 유희 확률을 유의수준과 비교하는 것이 아니라 검정 통계량을 직접 기각역과 비교하여 기각 여부를 판단할 수도 있음

통계적 가설 검정의 진행

- 주로 다음과 같은 단계로 진행
  1. 가설 설정: 귀무 가설과 대립 가설 설정
  2. 표본 데이터 수집: 연구나 실험을 통해 필요한 데이터 수집
  3. 통계량 계산: 수집된 데이터를 사용하여 특정 통계량(예: t-통계량, z-통계량) 계산
  4. p-값(p-value) 계산: 통계량을 사용하여 p-값 계산
     - p-값은 귀무 가설이 참일 때, 관측된 결과나 더 극단적인 결과가 나타날 확률을 나타냄
  5. 결정: p-값을 사용하여 귀무 가설을 기각하거나 기각하지 않음
     - 일반적으로 p-값이 특정 유의수준(alpha)보다 작을 때 귀무 가설을 기각함

통계적 추론의 사용

- 통계적 추론을 사용하여 데이터로부터 의사결정을 내릴 때
  - 불확실성을 고려하고 통계적으로 유의한 결과를 얻는 것이 중요
  - 이러한 접근 방식은 연구 및 의사결정 과정에서 신뢰성 있는 결론을 도출하는 데 도움이 됨

- 통계적 추론은 다양한 분야에서 사용됨
  - 의학, 경제학, 사회과학, 공학, 생물학 등 다양한 연구 및 의사결정 과정에서 중요한 도구로 활용
  - 통계적 추론을 통해 데이터로부터 의미 있는 정보를 도출하고 불확실성을 고려하여 결정을 내릴 수 있음

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

통계적 추론에 사용되는 주요 도구와 개념

- 표본 추출 (Sampling)
  - 모집단에서 대표적인 표본을 추출하는 과정
  - 표본이 모집단을 잘 대표해야 함

- 통계량 (Statistics)
  - 표본 데이터로부터 추정된 모집단의 특성을 나타내는 수치
  - 평균, 분산, 표준편차 등이 있음

AiDA Lab.

통계적 추론에 사용되는 주요 도구와 개념

- 확률분포 (Probability Distribution)
  - 데이터가 어떤 확률분포를 따르는지를 알아내는 것이 중요함
  - 정규분포, 이항분포, t-분포 등 다양한 분포가 사용됨

- 신뢰구간 (Confidence Interval)
  - 표본 데이터를 기반으로 모집단 특성에 대한 추정치를 제시하는데 사용
  - 일반적으로 추정값과 신뢰수준을 함께 표현

통계적 추론에 사용되는 주요 도구와 개념

- 가설 검정 (Hypothesis Testing)
  - 귀무 가설과 대립 가설을 설정하고
  - 표본 데이터를 사용하여 두 가설을 비교함으로써 통계적으로 유의한 결과를 얻는 과정

- p-값 (p-value)
  - 가설 검정에서 사용되며, 귀무 가설이 참일 때 나올 확률을 나타냄
  - 작은 p-값은 귀무 가설을 기각하는 데 사용

모집단에 대한 분포 추론

모집단에 대한 분포 추론

- 모집단에 대한 분포 추론(population distribution inference)
  - 통계학에서 모집단의 확률분포에 대한 정보를 추정하고 이해하는 과정을 의미
  - 모집단(population): 특정 관심 대상 집단의 모든 개체나 단위
  - 모집단의 분포: 이러한 개체나 단위의 특성에 대한 확률적인 정보를 담고 있음
  - 모집단 분포를 이해하고 추론하는 것은 다양한 분야에서 중요한 역할을 수행하고 있음

  - 분포 추론은 다양한 분야에서 활용되며, 예측, 품질 관리, 의사결정 등에 중요한 역할을 수행함
  - 이를 통해 모집단에 대한 정보를 얻고, 이 정보를 활용하여 효과적인 의사결정을 내릴 수 있음

모집단 분포 추론의 진행

1. 표본 추출 (Sampling)
   - 실제로 모집단 전체를 조사하는 것은 매우 어렵거나 불가능할 수 있음
   - 따라서 모집단에서 무작위로 표본(sample)을 추출하여 수행함
   - 표본은 모집단의 부분 집합이며, 모집단을 대표할 수 있는 특성을 가지고 있어야 함
2. 표본 데이터 수집
   - 표본을 추출한 후, 표본 데이터를 수집
   - 표본 데이터는 모집단의 특성을 반영하고 있어야 하며
   - 통계량(예: 평균, 분산)을 계산하여 모집단의 특성을 추정하는 데 사용됨

모집단 분포 추론의 진행

3. 통계량 계산

- 표본 데이터를 사용하여 모집단의 특성에 대한 통계량(예: 표본 평균, 표본 분산)을 계산
- 이러한 통계량은 모집단 특성의 추정치(estimates)로 사용

4. 모집단 분포 가정

- 모집단의 분포에 대한 가정을 수립
- 대표적으로 사용되는 가정은 정규분포, 이항분포, 포아송분포 등
- 이러한 가정은 모집단의 실제 분포를 근사화하는 데 사용됨

모집단 분포 추론의 진행

5. 추론 및 신뢰구간

- 표본 데이터와 모집단 분포 가정을 사용하여 모집단의 특성에 대한 추론을 수행
- 이때 신뢰구간(confidence interval)을 계산하여 모집단 특성이 어떤 범위 내에 있을 것으로 추정
- 신뢰구간은 추정값과 추정의 불확실성을 동시에 나타내는 중요한 개념

6. 가설 검정

- 표본 데이터를 사용하여 모집단 분포에 대한 가설 검정을 수행할 수 있음
- 예: 모집단 분포의 평균이 어떤 특정 값과 다른지에 대한 검정을 할 수 있음

7. 결과 해석

- 추론 결과 해석, 모집단에 대한 확률분포 정보 요약을 통하여 의사결정이나 정책 결정에 활용

AiDA Lab.
Artificial Intelligence & Data Analysis Laboratory

THANK YOU