1강 통계적 추론의 개요와 확률의 개념

학습내용
1. 통계적 추론의 기초를 이해한다.
2. 통계학의 역사를 이해한다.
3. 확률의 개념을 이해한다.
4. 조건부 확률을 이해한다.

01 통계적 추론의 기초 개념

1. 통계적 추론의 개요

- 통계학
    - (기술 통계학) 관심대상으로부터 데이터를 수집, 요약
    - (추론 통계학) 데이터로부터 일반성을 찾아내고, 이를 근거로 불확실한 사실에 대한 결론 및 규칙성 도출 → 통계적 추론
- 통계적 추론의 출발점
    - 관심 대상은 불확실 → 불확실성은 확률로 표현
    - 관심 대상을 모두 측정할 수 없음 → 일부를 측정해서 관심 대상 전체를 추론

2. 통계적 추론의 용어

- 모집단과 표본
    - 모집단(population): 관심 대상 전체
    - 표본(sample): 모집단의 일부 → 확률표본
- 확률변수와 데이터
    - 확률변수(random variable): 사건을 실수로 바꾸어 주는 함수
        - 확률변수는 확률분포에 따라 결정된다
    - 데이터: 관측값
- 확률분포
    - 확률변수는 확률분포를 따름
    - 확률분포: 몇 개의 모수(parameter)를 가진 수학함수로 가정
- 통계량과 표본분포
    - 통계량: 표본의 함수
        - (예) 표본평균, 표본분산

- 표본분포: 통계량의 확률분포

3. 통계적 추론의 구조

- 통계적 추론: 추정과 검정
  - 추정
    - 수 많은 꽃씨 중 흰색 꽃씨의 비율
    - 호수에 사는 A 어종의 수
    - 지지율 조사
  - 검정
    - 동전은 평평한가?
    - 신약은 효과 있는가?
- 통계적 추론의 원리
  - 가장 가능성 높은 결론을 도출한다 → 최대 가능도 추정
  - 가능성 낮은 일을 믿지 않는다
- 통계적 추론의 과정
  - 모집단: 모수, 확률분포
  - 표본: 통계량, 표본분포

    [메모] ①모집단에서 표본을 추출하고, ②표본에서 통계량과 표본분포를 구한다. ③통계추론방법을 활용하여(베이즈주의자는 사전정보를 추가로 활용) 모수와 확률분포를 추론한다.
- 통계적 추론의 과정
  - 어떤 모집단의 확률변수 X 는 $f(x|\theta)$ 라는 확률분포를 따른다.
    - $X \sim f(x|\theta)$
  - 그 확률변수에서 표본을 n개 추출하면, 그 표본도 $f(x|\theta)$ 라는 동일한 확률분포를 따르겠지
    - 확률표본: $X_1, X_2, \cdots, X_n \sim f(x|\theta)$
  - 가장 가능성 높은 결론을 도출하기 위해서 가능도 함수를 구한다.
    - 가능도 함수는 결합확률밀도함수랑 똑같다.
    - $L(\theta) = f(x_1, \cdots, x_n|\theta)$
    - 서로 독립적이라고 하면,
      $$L(\theta) = f(x_1, \cdots, x_n|\theta) = \prod_{x=1}^{n} f(x_i|\theta)$$
    - 라고 표현할 수 있다.
  - 가능도 함수의 값을 가장 크게하는 통계량

θ 을 구한다.
    - (수학적으로는) 가능도 함수를 미분해서 미분값이 0이 되는 통계량을 구한다.
  - 예를 들어, 모집단의 확률분포가 정규분포 $N(\mu, \sigma^2)$ 였다면,
    - μ 를 위한 추정량은 $\bar{X}$ 가 도출된다.
  - $\bar{X}$ 를 가지고 확률분포를 이용해서 추정과 검정을 한다.

4. 통계적 추론의 구분

- 통계적 추론 이론과 데이터 분석
  - 통계적 추론 이론
    - 내용을 알고 있는 상자에서 무작위로(randomly) n개의 공을 뽑았을 때, 빨간 공이 x개 나올 확률은?
  - 데이터 분석
    - 내용을 모르는 상자에서 n개를 뽑았을 때 이 중 x개가 빨간 공이라면 이 상자에는 빨간 공이 몇 %일까?
- 통계적 추론
  - 이론적 부분: 연역적 추론
  - 데이터 분석: 귀납적 추론
- 통계적 추론의 구성
  - 확률 이론: 객관적 확률, 주관적 확률
  - 추론 이론: 빈도론적 추론, 베이즈 추론

02 통계학의 역사

1. 통계학의 역사

- 통계학의 역사(1)
  - 확률의 시대: 17~18세기
    - B. Pascal, Jacob Bernoulli, A. de Moivre, T. Bayes
  - 오차이론의 시대: 18세기 중~19세기 중
    - P. S. Laplace, C. F. Gauss
    - 정규분포, 중심극한정리, 최소제곱법
  - 통계의 시대: 19세기 중, 후

        - A. Quetelet, F. Galton
    - 통계적 추론의 시대: 20세기 초 중
        - Karl Pearson, R. A. Fisher, W. S. Gosset, J. Neyman
    - 데이터 과학의 시대: 21세기
- 통계학의 역사(2)
    - 1900년 Karl Pearson - χ^2 분포
    - 1908년 W. S. Gosset - t 검정 분포
    - 1925년 R. A. Fisher - 최대 가능도 추정법, 정보량, 충분성, 효율성, 유의성 검정 등
        - 저서: 연구자를 위한 통계학
    - 1933년 Neyman Pearson - 가설 검정 이론
    - 1937년 Neyman Pearson - 신뢰구간
    - 1950년 Wald - 의사 결정 이론

03 확률의 개념

1. 확률의 정의

- 확률적 실험
    - 어떤 실험이 반복될 때 개개의 실험 결과는 미리 알 수 없으나 반복과정에서 "규칙성"을 지니는 실험
        - 동전 던지기, 주사위 던지기 등
- 표본공간과 사건
    - 표본공간(sample space): 확률적 실험을 통해 일어날 수 있는 모든 가능한 결과의 집합
    - 사건(event): 표본공간의 부분집합
- 확률의 정의
    - 어떤 사건이 일어날 가능성을 0과 1사이의 실수로 표현
- 빈도론적 확률
    - n번 시행했을 때 사건 A가 일어날 확률
        - $P(A) = \lim_{n \to \infty} \frac{n(A)}{n}$
- 고전적 확률
    - $P(A) = \frac{n(A)}{n(S)}$
- 공리적 확률: 다음의 공리를 만족하는 P라는 측도
    - $0 \le P(A) \le 1$
    - $P(S) = 1$

    $A_1, A_2, \cdots$

- 가 서로 배반사건일 때,
  $$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$
  를 만족
  $A_1, A_2, \cdots$
  - 가 서로 배반사건일 때, 합집합의 확률이 각각의 확률의 합과 같다
- 확률의 계산
  - $P(A^c) = 1 - P(A)$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2. 조건부 확률

- 조건부 확률: 사건 B 조건 하에 사건 A 발생 확률
  $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
  - - B라는 사건이 발생했다는 조건 하에서 사건 A가 발생할 확률 ***P(A|B)*** 는
    - B라는 사건이 발생할 확률 ***P(B)*** 가 분모가 되고, B도 발생하고 A도 발생할 확률인 ***P(A∩B)*** 가 분자인 분수로 계산한다.
  - 식의 형태를 바꾸면 이렇게도 표현할 수 있다.
  - $P(A \cap B) = P(A|B)P(B)$
- 예) 주사위 눈이 짝수라는 조건 하에 주사위 눈이 3 이하의 숫자가 나올 확률은?
  - S = {1,2,3,4,5,6}, A = {1,2,3}, B = {2,4,6}, A∩B = {2} 이라고 하면
    $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$
  - 이다.
- 조건부 확률의 성질
  - P(B|A) ≥ 0
  - P(S|A) = 1
  - $$S = \bigcup_{i=1}^{\infty} B_i, B_i \Rightarrow P(\bigcup_{i=1}^{\infty} B_i|A) = \sum_{i=1}^{\infty} P(B_i|A)$$
- 역확률
  - $$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$
- 베이즈 정리
  $S = B \cup B^c$
  - 이고, A가 S의 부분집합이라고 하면 A를 다음과 같이 표현할 수 있다.
    - $A = (A \cap B) \cup (A \cap B^c)$
      $(A \cap B)$
    - 와
      $(A \cup B^c)$
      는 배반사건이므로 두 합집합의 확률은 더하기로 표현할 수 있다.

$$P(A) = P((A \cap B) \cup (A \cap B^c))$$

$$= P(A \cap B) + P(A \cap B^c)$$

$$= P(A|B)P(B) + P(A|B^c)P(B^c)$$

따라서,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

이다.

- 베이즈 정리는 다음과 같이 나타낼 수 있다.

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

- 베이즈 정리를 더 일반화하면 다음과 같이 표현할 수 있다.

  $$S = \bigcup_{i=1}^{k} B_i, B_j$$

  - 는 배반사건
  - $$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$
  - $$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$
- 예) 전체 인구의 10%가 어떤 질병을 앓고 있다. 이 질병의 진단시약을 조사한 결과, 질병에 걸린 사람 중 90%는 양성 반응, 질병에 걸리지 않은 사람 중 80%는 음성 반응. 어떤 사람의 진단 시약 검사 결과가 양성 반응일 때 이 사람이 질별에 걸렸을 확률은?
  - 질병에 걸렸을 확률: P(D)=0.1, 질병에 걸리지 않았을 확률
    $P(D^c) = 0.9$
  - 질병에 걸렸는데 진단결과 양성일 확률: $P(T^+|D) = 0.9$, 질병에 걸렸는데 진단결과 음성일 확률: $P(T^-|D) = 0.1$
  - 질병에 안 걸렸는데 진단결과 음성일 확률:
    $P(T^-|D^c) = 0.8$
    , 질병에 안 걸렸는데 진단결과 양성일 확률:
    $P(T^+|D^c) = 0.2$
  - 진단 시약 검사가 양성 반응일 때 이 사람이 질병에 걸렸을 확률: P(D|T+)

- $P(D|T^+) = \dfrac{P(D \cap T^+)}{P(T^+)}$

$$P(T^+) = P((T^+ \cap D) \cup (T^+ \cap D^c))$$

$$= P(T^+ \cap D) + P(T^+ \cap D^c)$$

$$= P(T^+|D)P(D) + P(T^+|D^c)P(D^c)$$

= 0.9 × 0.1 + 0.2 × 0.9

$$P(D|T^+) = \frac{0.9 \times 0.1}{0.9 \times 0.1 + 0.2 \times 0.9} = \frac{1}{3}$$

- 독립
  - 두 사건 A,B 간 독립

$$P(A|B) = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$$

- 예) A,B가 독립사건이고, P(A)=P(B)=0.5 일 때 P(A∪B)?

P(A∪B) = P(A) + P(B) - P(A∩B)
= 0.5 + 0.5 - (0.5 × 0.5) = 0.75

04 정리하기

- 통계추론은 데이터를 기반으로 불확실한 사실에 대한 결론이나 예측을 하는데 필요한 이론과 방법에 관한 학문 분야로 통계학의 중심이다.
- 확률변수는 확률적 실험에서 실험결과를 관심의 대상이 되는 수 값으로 나타낸 것이다.
- 확률은 어떤 사건이 일어날 가능성을 0과 1사이의 실수로 표현한 것이다. 확률은 빈도론적, 고전적, 공리적으로 정의된다.
- 조건부 확률:
  $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
- 베이즈 정리:
  $P(B_i|A) = \dfrac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}, S = \bigcup_{i=1}^{k} B_i, B_j$
  배반 사건
- 독립성:
  $P(A|B) = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$