기초 통계 세션 과제

김준호

August 14, 2024

1 Frequentist와 Bayesian의 차이점

Frequentist 방법과 Bayesian 방법론은 통계학에서 확률과 추론을 다루는 두 가지 주요 접근법이다. 먼저 확률 해석적 측면에서 Frequentist는 확률을 반복 시행에서의 장기적인 빈도로 정의하고, Bayesian은 주관적 믿음의 정도로 해석한다. 이는 곧 모수 추정의 방법론 차이로 이어지는데, Frequentist는 모수를 고정된 상수로 가정하고, 샘플 데이터로부터 점추정과 신뢰구간을 사용한다. 반면 Bayesian은 모수를 하나의 확률 분포로 보고 사전 분포와 데이터로부터 사후 분포를 계산한다. Frequentist는 사전 정보를 사용하지 않고 데이터만으로 추론하지만 Bayesian은 사전 분포를 사용하여 기존 지식을 반영한다.

가설 검정 측면에서 Frequentist는 p-value를 사용하여 귀무가설을 검정하지만, Bayesian은 사후 확률을 통해 가설의 신뢰도를 평가한다.

특징적으로 Bayesian 방법은 사후 분포 계산이 필요한데, 손으로 계산하기 어려운 경우가 많아 Frequentist 방법보다 계산적으로 더 복잡할 수 있다

이러한 차이점들로 인해 각 방법론은 상황에 따라 서로 다른 강점을 가지며, 적절히 선택하여 사용해야 한다.

2 사전분포 g(θ)를 완벽하게 안다고 가정하는 베이지안의 문제점에 대한 해결법

사전분포를 특정하지 않고 데이터로부터 직접 추정하는 Nonparametric Bayes, 즉 비모수적 베이지안 접근법을 사용할 수 있다.

불확실성을 반영한 non-informative prior를 사용하여 편향을 줄일 수 있다.

Sensitivity Analysis를 통해 여러 사전분포를 사용하여 결과의 민감도를 분석하고, 사전분포에 따른 결과 변화를 평가할 수 있다.

Hierarchical Modeling으로 계층적 사전분포를 도입하여 더 general한 모형을 구축할 수 있다.

Empirical Bayes(데이터 기반 사전분포): 전문적 의견과 기존 데이터로부터 사전분포를 구성하여 보다 현실적인 사전 정보를 반영할 수 있다.

Bayesian Update: 초기 사전분포의 불확실성을 데이터에 기반한 사후분포로 계속 업데이트하여 점진적으로 정확도를 높일 수 있다. 이러한 방법들을 통해 베이지안 방법론에서 사전분포에 대한 가정의 문제점을 완화할 수 있다.

3 분포수렴은 CDF를 통해 정의된다. PDF를 이용하여 분포수렴을 정의할 수 없는 이유를 반례를 통해 서술하라.

확률밀도함수(Probability Density Function, PDF)를 이용하여 분포수렴을 정의할 수 없는 이유를 반례를 통해 설명할 수 있다.

먼저, 분포수렴의 정의를 보자. 확률변수 X_n이 어떤 확률변수 X로 분포수렴한다고 할 때, 이는 모든 실수 x에 대해 X_n의 누적분포함수 $F_{X_n}(x)$가 X의 누적분포함수 $F_X(x)$로 수렴함을 의미한다. 즉,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \text{for all } x \in R$$

여기서 중요한 점은 누적분포함수 $F_X(x)$는 항상 연속적이고 비감소하는 함수라는 것이다. 반면, 확률밀도함수 $f_X(x)$는 이러한 연속성과 항상 일치하지 않을 수 있다.

반례: 디랙 델타 함수

디랙 델타 함수 $\delta(x)$는 특정한 점 $x=0$에서 무한대의 값을 가지며, 전체 면적이 1인 일반화된 함수로, 이를 통해 특정한 값에 질량이 집중된 확률분포를 표현할 수 있다. 예를 들어, X_n이 평균이 0이고 분산이 $\frac{1}{n}$인 정규분포를 따를 때, X_n의 PDF는 다음과 같이 표현된다:

$$f_{X_n}(x) = \frac{1}{\sqrt{2\pi\frac{1}{n}}} \exp\left(-\frac{x^2}{2\frac{1}{n}}\right)$$

이때, n이 커짐에 따라 이 정규분포는 점점 좁아지고 높아지며, 그 결과 디랙 델타 함수로 수렴한다:

$$\lim_{n\to\infty} f_{X_n}(x) = \delta(x)$$

그러나, 디랙 델타 함수 $\delta(x)$는 전통적인 의미의 PDF가 아니다. 따라서 PDF로는 분포수렴을 정의하기 어렵다. 반면, 이 확률변수의 누적분포함수는 다음과 같이 수렴한다:

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

여기서 $F_X(x)$는 x가 0보다 크면 1, 작으면 0인 함수로, X_n이 0에서 모든 질량을 가지는 경우를 나타낸다.

결론적으로, PDF를 이용하여 분포수렴을 정의할 수 없는 이유는 PDF가 불연속적이거나 디랙 델타 함수처럼 전통적인 의미의 함수가 아닌 경우가 있기 때문이다. 반면, CDF는 항상 존재하며, 분포의 누적적인 특성을 통해 분포수렴을 올바르게 정의할 수 있다.

4 MGF를 이용한 CLT의 증명

- **Theorem (CLT).** $X_1, \ldots, X_n$ 이 μ와 σ^2을 평균과 분산으로 갖는 분포로부터의 iid random sample 이라고 하자. 이 때 확률변수

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

은 표준정규분포로 분포수렴한다.

Proof. $Z_i = (X_i - \mu)/\sigma$이고, $M_Z(t)$ 가 Z_i의 common mgf라고 하자. 이 때, $n \to \infty$이면 다음이 성립한다:

$$\begin{aligned}
M_{Y_n}(t) &= M_{\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i}(t) \\
&= M_{\sum_{i=1}^n Z_i}\left(\frac{t}{\sqrt{n}}\right) \\
&= \left(M_Z\left(\frac{t}{\sqrt{n}}\right)\right)^n \\
&= \left\{\sum_{k=0}^{\infty} \frac{M_Z^{(k)}(0)}{k!}\left(\frac{t}{\sqrt{n}}\right)^k\right\}^n \\
&= \left\{1 + \frac{t}{\sqrt{n}}\cdot M_Z'(0) + \frac{t^2}{2n}\cdot M_Z''(0) + o\left(\left(\frac{t}{\sqrt{n}}\right)^2\right)\right\}^n \\
&= \left\{1 + 0 + \frac{t^2}{2n}\cdot 1 + o\left(\frac{1}{n}\right)\right\}^n \\
&= \left\{1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right\}^n \\
&\to e^{t^2/2},
\end{aligned}$$

즉, Y_n의 mgf는 standard normal random variable의 mgf로 수렴한다.