기초 통계 과제

백준호

July 29, 2024

1 Frequentist와 Bayesian의 차이점

Frequentist 방법과 Bayesian 방법론은 통계학에서 확률과 추론을 다루는 두 가지 주요 접근법입니다.

- **확률 해석**: Frequentist는 확률을 반복 시행에서의 장기적인 빈도로 정의하고, Bayesian은 주관적 믿음의 정도로 해석합니다.
- **모수 추정**: Frequentist는 모수를 고정된 상수로 가정하고 점추정과 신뢰구간을 사용합니다. 반면 Bayesian은 모수를 하나의 확률 분포로 보고 사전 분포와 데이터로부터 사후 분포를 계산합니다.
- **사전 정보 사용**: Frequentist는 사전 정보를 사용하지 않고 데이터만으로 추론하지만, Bayesian은 사전 분포를 사용하여 기존 지식을 반영합니다.
- **가설 검정**: Frequentist는 p-value를 사용하여 귀무가설을 검정하지만, Bayesian은 사후 확률을 통해 가설의 신뢰도를 평가합니다.
- **계산 복잡성**: Bayesian 방법은 사후 분포 계산이 필요한데, 손으로 계산하기 어려운 경우가 많아 Frequentist 방법보다 계산적으로 더 복잡할 수 있습니다.

2 사전분포 $g(\theta)$를 완벽하게 안다고 가정하는 베이지안의 문제점에 대한 해결법

- **Nonparametric Bayes**: 사전분포를 특정하지 않고 데이터로부터 직접 추정하는 비모수적 베이지안 접근법을 사용할 수 있습니다.
- **Non-informative prior**: 불확실성을 반영한 non-informative prior를 사용하여 편향을 줄일 수 있습니다.
- **Bayesian Update**: 초기 사전분포의 불확실성을 데이터에 기반한 사후분포로 계속 업데이트하여 점진적으로 정확도를 높일 수 있습니다.
- **Sensitivity Analysis**: 여러 사전분포를 사용하여 결과의 민감도를 분석하고, 사전분포에 따른 결과 변화를 평가할 수 있습니다.
- **Hierarchical Modeling**: 계층적 사전분포를 도입하여 더 일반적인 모형을 구축할 수 있습니다.
- **Empirical Bayes**: 전문적 의견과 기존 데이터로부터 사전분포를 구성하여 보다 현실적인 사전 정보를 반영할 수 있습니다.

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

다음과 같은 확률밀도함수 $f_n(x)$가 있다고 가정해 봅시다:

$$f_n(x) = \begin{cases} n^2 x & \text{if } 0 \le x \le \frac{1}{n}, \\ n & \text{if } \frac{1}{n} < x \le \frac{2}{n}, \\ 0 & \text{otherwise.} \end{cases}$$

이 함수들의 누적분포함수 $F_n(x)$는 다음과 같습니다:

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{n^2 x^2}{2} & \text{if } 0 \le x \le \frac{1}{n}, \\ \frac{1}{2} + n(x - \frac{1}{n}) & \text{if } \frac{1}{n} < x \le \frac{2}{n}, \\ 1 & \text{if } x > \frac{2}{n}. \end{cases}$$

$n \to \infty$일 때, $F_n(x)$는 다음과 같이 수렴합니다:

$$F_n(x) \to \begin{cases} 0 & \text{if } x < 0, \\ 0 & \text{if } 0 \le x < \frac{1}{2}, \\ 1 & \text{if } x \ge \frac{1}{2}. \end{cases}$$

이 결과는 $x = \frac{1}{2}$에서 불연속점이 생기는 누적분포함수로 수렴합니다. 그러나 확률밀도함수 $f_n(x)$의 경우, $n \to \infty$일 때 다음과 같이 됩니다:

$$f_n(x) = \begin{cases} n^2 x & \text{if } 0 \le x \le \frac{1}{n}, \\ n & \text{if } \frac{1}{n} < x \le \frac{2}{n}, \\ 0 & \text{otherwise.} \end{cases}$$

이는 $n \to \infty$일 때 무한대로 발산하는 부분이 존재하며, 하나의 의미 있는 함수로 수렴하지 않습니다. 따라서, 확률밀도함수를 사용하여 분포수렴을 정의할 수 없는 이유는 확률밀도함수가 특정한 형태로 수렴하지 않기 때문입니다. 반면, 누적분포함수는 항상 유한한 값을 가지며 분포수렴을 의미 있게 정의할 수 있습니다.

4 MGF를 이용한 CLT의 증명

정리 (중심극한정리). $X_1, \ldots, X_n$이 평균 μ와 분산 σ^2를 가지는 분포로부터 추출된 독립동일 분포 표본이라고 가정합니다. 그러면 다음 확률변수

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

는 표준 정규 분포로 수렴합니다.

증명. $Z_i = \frac{X_i - \mu}{\sigma}$라고 하고, Z_i들의 공통 적률생성함수(MGF)를 $M_Z(t)$라고 합시다. 그러면,

$$M_{Y_n}(t) = M_{\frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i}(t) = M_{\sum_{i=1}^n Z_i}\left(\frac{t}{\sqrt{n}}\right) = \left(M_Z\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

테일러 급수를 사용하면, $M_Z\left(\frac{t}{\sqrt{n}}\right)$는 다음과 같이 근사할 수 있습니다:

$$M_Z\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} \frac{M_Z^{(k)}(0)}{k!}\left(\frac{t}{\sqrt{n}}\right)^k = 1 + \frac{t}{\sqrt{n}}M_Z'(0) + \frac{t^2}{2n}M_Z''(0) + o\left(\left(\frac{t}{\sqrt{n}}\right)^2\right)$$

Z_i들이 표준화된 변수이므로 $M_Z'(0) = 0$이고 $M_Z''(0) = 1$입니다. 따라서,

$$M_Z\left(\frac{t}{\sqrt{n}}\right) \approx 1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

이제 이 결과를 n제곱하여,

$$\left(1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{t^2/2}, \text{ as } n \to \infty$$

따라서 Y_n의 MGF가 표준 정규 분포의 MGF로 수렴함을 보였습니다. 이는 Y_n이 분포상으로 표준 정규 분포로 수렴함을 의미합니다.