DSL 심화스터디 <기초반>

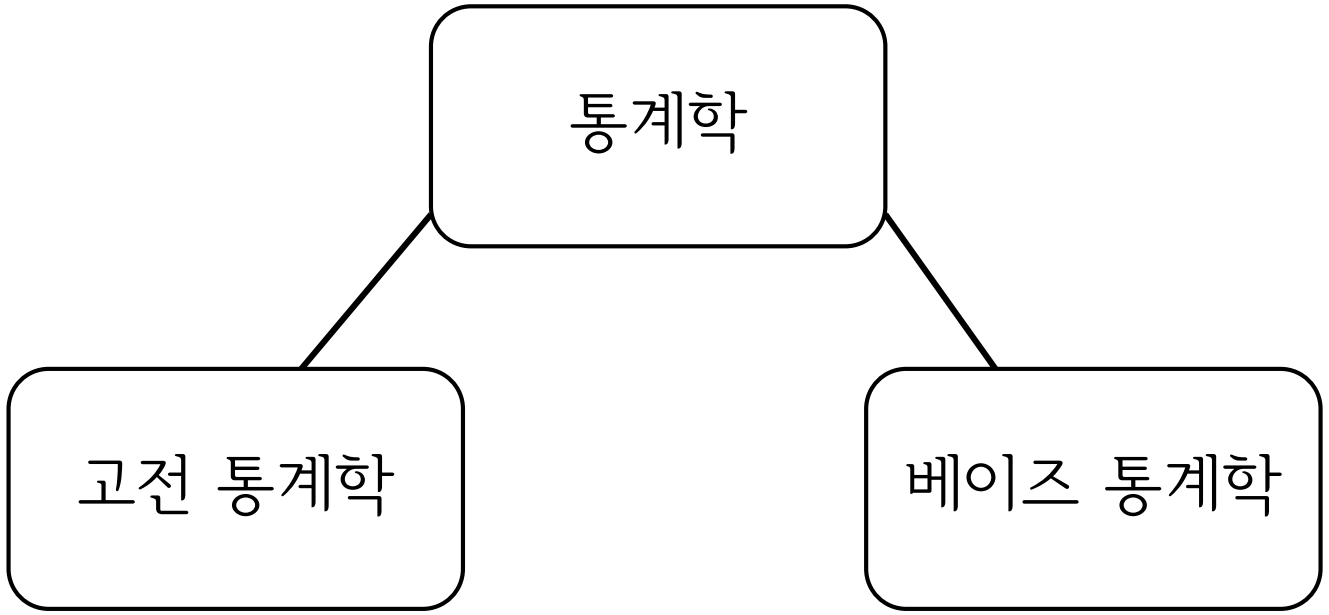
베이지안 통계

Day1

발표자 : 5기 박채은

베이지안 통계란?

- 베이즈가 주창한 통계적 방법 => 고전 통계학과 구분됨.
- 표집에서 얻은 정보뿐만 아니라 연구자가 가지고 있는 사전 지식이나 주관적 의견 또는 신념과 같은 정보도 포함시키는 추리 통계의 한 방법



베이지안 통계학이 이렇게 중요합니다!(강승호 교수님의 답변)

통계학의 미래에 대한 질문에서 교수님의 답변

(답변) 전체적으로 위 질문은 통계학의 미래를 묻는 질문이라고 생각합니다. 요즘처럼 컴퓨터 연산속도는 기하급수적으로 빨라지고 (퀀텀 컴퓨터가 사용화된다고 합니다), 인공지능이 개발되는 시대에 통계학의 미래가 어떻게 될지 저도 개인적으로 너무나 궁금합니다. 걱정도 좀 되고 기대도 좀 되기도 하고요.

최근 큰 인기를 끌고 있는 머신러닝, 딥러닝 같은 분야들은 "예측 (prediction)" 분야입니다. 딥러닝을 사용하면 알파고처럼 예측이 상당히 정확하다고 하는데, 왜 정확해지는지는 아직 이론적으로는 밝혀지지 않았다고 합니다. 결국 통계학이 그에 대한 답을 찾아내지 않을까 추측합니다.

미래를 다룬 영화를 보면, 주인공이 인공지능에게 말로 "xxx에 대해 분석해봐" 라고 지시하면, 슈퍼컴퓨터를 갖춘 인공지능이 수 초 내에 "xxx일 확률이 60%입니다"라고 답변하는 것을 볼 수 있습니다. 그 미래의 인공지능이 어떤 알고리즘으로 그런 예측을 하게 될지는 저로서는 지금 알 수 없지만, 결국 예측은 100% 확실한 것은 없고 불확실성을 내포할 수 밖에 없기에 "확률"이 포함될 수 밖에 없으리라 생각합니다. 그런데 "확률"을 다루는 학문이 무언인가요? 네, 바로 통계학입니다. 그래서 매우 복잡하고 어려운 방법으로 예측을 하겠지만, 결국 통계학이 깊숙이 관련되어 있으리라 예상합니다.

그러한 복잡한 예측 방법에는 여러 통계방법들이 핵심적인 역할을 하겠지만, 저는 개인적으로 베이지안 (Bayesian) 통계학이 핵심적인 역할을 하지 않을까 추측합니다.

Day1 강의 목표

- 확률의 세 가지 정의를 이해하기
- 조건부 확률에 대해 이해하기
- 조건부 확률에 대한 이해를 바탕으로 베이즈 정리 계산하기
- 이산형, 연속형 확률 분포에 대해 이해하기

확률의 세 가지 정의

1. Classical: equally likely
-같은 가능성 가지는 outcome-> 같은 확률 가짐.
-예: fair six sided die 던질 때 4가 나올 확률은 1/6.

2. Frequentist: relative frequency
-어떤 시행을 반복하였을 때 일어나는 상대 빈도수
-예: fair six sided die 계속 던질 때, 여섯 번 중 한 번 꼴로 4가 나오면 4가 나올 확률은 1/6.

3. Bayesian: personal perspective
-확률 계산할 때 해당 문제에 관하여 알고 있는 정보를 반영
-예: 4지 선다형 문제 답 모를 때 익숙한 단어 있는 선지를 고르면 맞을 확률이 ¼보다 커질 수 있음.

+ odds

A라는 사건 있을 때, odds는 $\frac{P(A)}{P(A^c)}$

두 가지 결과만 나오는 사건에서
두 결과가 나올 확률의 비

예) fair six sided die 던질 때 4가 나오는 사건 : A
P(A) = 1/6
Odds for A = $\frac{1/6}{5/6} = \frac{1}{5}$ 또는 1:5로 표현됨.

확률이 part / whole 이라면 odds는 part : part임.

Odds => probability
Odds: a : b => probability: $\frac{a}{(a+b)}$

Conditional probability(조건부 확률)

P(A|B) = 사건 B가 일어났다고 가정할 때, 사건 A가 일어날 확률

즉, B를 새로운 표본공간으로 생각하고 B에서 (A ∩ B)가 일어날 확률

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

예: F: Female, CS: CS major

P(F | CS) = 1/3

P(CS | F) = 4/9

	Female	Not Female	Total
CS major	4	8	12
Not CS major	5	13	18
	9	21	30

사건의 독립과 종속

독립(independence)
사건 B가 일어나거나 일어나지 않는 것이 사건A가 일어날 확률에 영향을 주지 않음.

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

종속
-독립이 아닌 경우

베이즈 정리

베이지안 통계에서 하는 대부분의 일에 대한 이론적 뒷받침

two discrete events:

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

가능한 outcome이 A와 A^c 뿐만이 아니라 세 가지 경우(A_1, A_2, A_3)라면? (+A_1, A_2, A_3 중 하나는 반드시 일어남.)

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)}$$

Bayes thereom for continuous distributions

continuous random variable θ

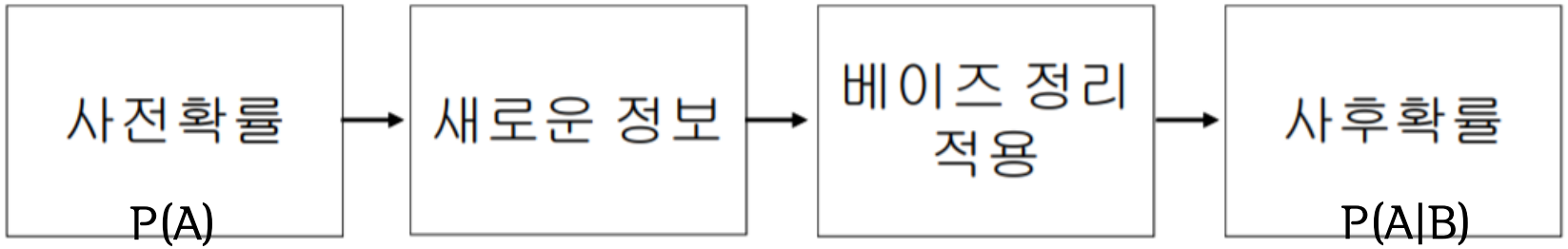
conditional density for θ given y

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta}.$$

베이즈 정리

"결과로부터 원인을 추론하기"
예) 원인(가설): 질병에 걸림 혹은 질병에 걸리지 않음.
결과(데이터): 증상 유무
=> 증상 유무(결과)을 통해 질병 걸렸는지(원인)를 추론할 수 있다.
-경험적 데이터로부터 원인(가설)의 진위를 알아볼 때 이용되는 정리

예) A: 병에 걸린 사건
B: 증상이 있을 사건
P(A) -> p(A|B)



비판 O
튜링의 암호 해독

Pdf(probability density function)의 key rule

1. 모든 실수값에 대해 $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Distributions

- distributions
  - discrete — Bernoulli, binomial, geometric, multinomial, poisson 등
  - continuous — Uniform, normal, exponential, gamma, beta, t 등

3. Review of distributions

Discrete 예시) Poisson
- **고정된 지역, 시간 또는 부피** 등에서 관심 있는 **사건의 관찰 수 또는 발생 횟수**
- 예: **오늘 하루** 우리 학교 컴퓨터에 접속한 **사용자의 수**
  : **어느 주말 일요일에** 발생한 **교통사고 사망자의 수**
  : **어느 하루** 동안 지정된 생산 라인에서 발생한 **불량품의 개수**

Continuous 예시) exponential
- **랜덤 시간**의 분포를 설명하기 위한 분포 중 하나
- 예: 전구의 **수명**
  : 어느 방사능 원소가 **분해될 때까지 걸리는 시간**

중심극한정리

★★통계학에서 매우매우 중요한 정리★★

$X_{1,\ldots,}\ X_n$: a random sample from a distribution which has mean μ and variance σ^2 $(0 < \sigma^2 < \infty)$

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \Rightarrow N(0,1)\,.$$

감사합니다

1) In what ways could the frequentist paradigm be considered objective? In what ways could the Bayesian paradigm be considered objective?

2) Identify ways in which each paradigm might be considered subjective.