DSL Seminar: 통입+통방 (1)

Kyung-han Kim

Data Science Lab

January, 2023

목차

- 세미나 오리엔테이션
- 통계학은 무엇일까?
- 각종 기초 개념 설명
- 조건부확률과 베이즈 정리 (시간이 되면)

오리엔테이션 - 기본 스케줄

- 1주차: 오리엔테이션
- 2주차: 확률변수와 확률분포 (이산형)
- 3주차: 확률분포 (연속형)
- 4주차: 추정량, 표본평균과 표본분산, 통계적 추정, 모평균의 구간 추정
- 5주차: 가설 검정
- 6주차: 카이제곱 검정, 분산분석(1)
- 7주차: 분산분석(2), 회귀분석(1)
- 8주차: 회귀분석(2), QnA

통계학은 무엇일까?

- 여러 전공 수업에서 교수님들께서 남기신 말씀들을 종합해 보자면 통계학을 아래와 같이 정의할 수 있을 것입니다.
- 통계학은,
  모집단에서 **표본**을 **추출**하고,
  그 표본에서 얻은 **추정량**으로 **모수**를 **추정**하는 학문이다.
- 우리 학교 통계학입문 교재에서는,
  "통계학은 추론 과정에서 필연적으로 수반되는 **오차**의 크기를 계산하고, 그것을 줄이는 방법을 찾는 학문이다." 라고 정리한다.
- 지금부터는 통계학이 무엇인지 이해하기 위해,
  통계학을 정의하는 데에 사용된 용어들을 설명합니다.

통계학 기초 개념(1): 모집단과 표본

- 모집단 (Population): 우리가 조사하고자 하는, 관심 있는 집단 전체
  - 모집단 전체를 조사하는 것을 전수 조사라고 한다.
  - 전수조사는 시간적, 비용적 문제로 사실상 시행이 불가능한 경우가 많다.
- 표본 (Sample): 모집단의 특성을 알아내기 위해 뽑은 모집단의 일부분
- 우리는 전체 집단의 성격을 알아내고 싶지만,
  그러기 위해 집단 자체를 통째로 조사하는 것은 거의 불가능하다.
- 따라서 그 중 일부에만 접근하면서도 최대한 정확하게 모집단의
  특성을 알아내는 것이 통계학의 목적이다.
  - 모집단의 일부만을 조사하는 방식을 표본조사라고 한다.
- 추출 (Sampling): 표본을 뽑는 과정, 절차, 규칙
  - 데이터사이언스표본추출이론 수업에서 자세히 다룹니다.

통계학 기초 개념(2): 모수, 추정, 추정량, 오차

- 모수 (Parameter): 모집단의 성질을 나타내는 특정한 값
  - 확률변수와 확률분포에 대해 배우면 더 잘 이해할 수 있습니다.
  - 모집단을 특정 확률분포를 따르는 확률변수로 표현할 수 있다면, 해당 확률변수가 따르는 확률분포의 모수를 알아내는 것이 곧 모집단의 성질을 설명하는 것과 동등한(equivalent) 문제가 됩니다.
- 추정 (Estimation): 하나의 값으로 다른 값을 유추하는 것
- 추정량 (Estimator): 모수를 추정하기 위해 계산하는 값
- 통계학에서는 추정량을 이용해 모수를 추정하는 통계적 추정 (Statistical estimation)을 사용합니다.
- 오차 (Error): 모수와 추정량의 차이. 작으면 작을수록 좋다.

통계학 기초 개념(3): 대푯값

- 대푯값은 모집단을 단 하나의 값으로 요약하는 값입니다.
- 가장 대중적인 대푯값으로는 평균(Mean), 중위수/중간값(Median), 최빈값(Mode)이 있습니다.
- 평균 (Mean): 모든 값을 더해서 표본 수로 나눠 구함

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- 중위수/중간값 (Median): 모든 값을 크기 순서대로 나열했을 때 정확히 가운데에 위치하는 값
  - 데이터의 개수가 짝수일 경우, 가장 가운데에 위치하는 두 값의 평균으로 구한다.
- 최빈값 (Mode): 전체 값 중에서 가장 자주 출현한 값
- 이 중에서 평균을 가장 중점적으로 보게 됩니다.
  평균은 기댓값(Expectation)이라고도 합니다.

통계학 기초 개념(4): 산포도

- 산포도는 데이터가 얼마나 서로 흩어져 있는지를 나타내는 값입니다.
- 통계학입문/통계방법론에서 중점적으로 다루는 산포도는 사실상 분산/표준편차가 유일합니다.
- 분산 (Variance): 편차 (Deviance)의 제곱의 평균
  - 편차는 변량(데이터) - 평균 으로 정의된다.
  - 편차를 제곱하지 않고 더하면 항상 0이 되기 때문에 무의미하다.

$$V[X] = E[(X - \mu)^2] = E[X^2] - E[X]^2$$

- 표준편차 (Standard Deviation): 분산의 양의 제곱근.
  편차를 제곱하면서 단위가 달라지기 때문에 단위를 원상복구하기 위해 표준편차를 사용한다.

$$\sigma[X] = \sqrt{V[X]}$$

확률이란 무엇인가?

- 확률(Probability)은 어떤 사건이 발생할 가능성을 0과 1 사이의 숫자로 수치화한 것을 의미한다.
  - 사건(Event): 결과가 우연에 의해 결정되는 실험 혹은 관찰.
- 수치화하는 방식에 따라 크게는 전통적 접근, 상대적 비율 접근, 주관적 접근으로 나눌 수 있다.
- 1) 전통적 접근: 발생할 가능성이 똑같은 사건에는 똑같은 확률을 부여하는 방식
- 2) 상대적 비율 접근: 똑같은 시행을 무수히 많이 반복했을 때 해당 사건이 발생하는 횟수의 비율이 수렴해가는 값을 확률로 취급하는 방식 (Frequentist approach)
- 3) 주관적 접근: 개인적인 믿음의 정도 (Degree of belief)에 따라 확률을 부여하는 방식 (Bayesian approach)

cf] Frequentist and Bayesian

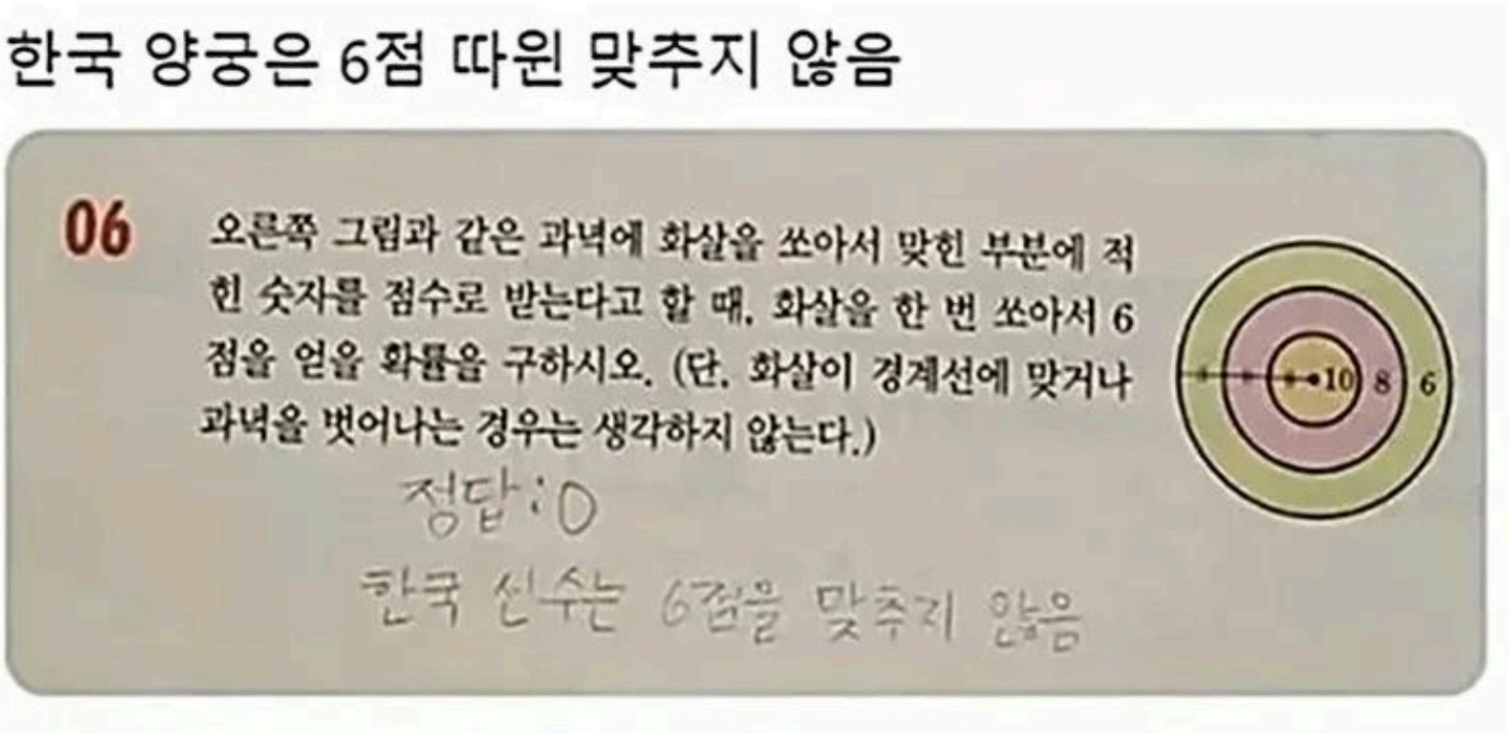

Figure 1: Degree of belief

조건부 확률

- 조건부 확률 (Conditional Probability)은 한 사건(A)이 발생했다는 것이 조건으로 주어졌을 때, 다른 사건(B)이 발생할 확률로 정의된다. $P(B|A)$로 표기한다.

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

- 만약 사건 A가 사건 B에 영향을 주지 못한다면, 사건 A가 발생했다는 정보가 B의 발생 확률에 영향을 주지 않을 것이다.

$$\text{A and B are independent if and only if } P(B|A) = P(B)$$

$$\text{This condition is equivalent with } P(A)P(B) = P(A \cap B)$$

베이즈 정리

- 조건부 확률의 정의식을 변형해 베이즈 정리 (Bayes' Theorem)를 얻을 수 있다.
- 베이즈 정리의 목표는 조건과 관심 있는 사건의 위치를 뒤바꾸는 것이다. (A와 B의 위치를 바꾸는 것)

$$\text{Bayes' Theorem: } P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)P(A|B)}{P(A)}$$

- 베이즈 정리를 이용하면 $P(A|B)$와 P(A), P(B)를 알 때 $P(B|A)$를 구할 수 있다!
- 이 정리로부터 베이즈 통계 (Bayesian Statistics)의 뼈대가 완성된다.
  - Prior와 Likelihood를 이용해 Posterior를 Update한다.