Lecture 01 - Introduction to Statistics

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

SEOULTECH 국립 서울과학기술대학교 IT정책전문대학원

I. Motivation

Setting

- 동전을 던져서 앞면/뒷면을 맞추는 게임을 한다. 게임을 시작하기 전에 동전을 던져보는 테스트를 할 수 있다. 게임에 사용하기로 한 동전을 살펴보니 왠지 약간 찌그러져 있어서 fair coin (앞/뒤의 확률이 각각 50%인 동전)이 아닐 수도 있다는 생각이 든다.

- 여러번 동전을 던지는 실험의 수행한 결과로서 아래의 4가지 시나리오를 생각해보자.
  1. 10번을 던져서 6번 앞면이 나왔다.
  2. 100번을 던져서 60번 앞면이 나왔다.
  3. 1000번을 던져서 600번 앞면이 나왔다.
  4. 100번을 던져서 65번 앞면이 나왔다.

- 위의 4가지 경우중에서 어떤 시나리오가 "동전이 fair coin이 아니다."라는 가장 큰 **확신**을 주는가?

통계 분석을 통해서 답할 수 있는 질문들

- 실험결과에 대한 위의 네 가지의 시나리오 중에서 어떤 경우가 동전이 찌그러져 있다는 가장 높은 수준의 **확신**을 주는가?
- 만약에 네 개의 실험이 각기 다른 동전을 사용했다면, 네 개의 동전에 대해서 앞면이 나올 확률이 가장 높은 순서대로 동전의 **순위를 매길 수 있는가**?
- 통계적으로 분석하였을 때에 3번 시나리오가 동전이 찌그러져 있다는 가장 확실한 근거라고 한다면, **얼마나 강력한 확신**을 가질 수 있는가? 이를테면, 3번 시나리오 결과로서 동전이 찌그러져 있다는 100%의 확신을 가질 수 있는가? 아니면 90%?

Exercise 1

또 다른 어떤 질문에 답하는 것이 가능할까?

기초 통계학에서 다루는 것들

- 위와 같은 질문들에 답할 수 있는 수학적인 이론
- 높은 확신을 얻을 수 있도록 실험을 설계하는 법

응용 분야

- 의약품의 효과가 있는가?
- 광고의 효과가 확실한가?
- 신제품의 효과를 확실하기 위해서는 몇번의 실험을 수행해야 하고 어떤 결과를 얻어야 하는가?

II. 통계에 관련된 용어들

참값과 추정 (True value & Estimation)

- 통계학에서는 신이 이미 정해버린 것과 같은 **참 값**의 존재를 가정한다.
  - 참값을 true value라고 한다.
  - 인간은 true value에 대해서 알기를 원한다.

- 신이 관장하는 **큰 공간**에 비해 인간은 매우 **작은 공간**에서 살아가면서 관찰과 경험을 한다. 이를 통해서 신의 의도를 파악하기 위해서 노력한다.
  - 큰 공간을 **모집단 (population set)** 이라고 한다.
  - 작은 공간을 **표본집단 (sample space)**이라고 한다.
  - 의도를 파악하기 위해서 추론하는 것을 **추정 (estimation)**이라고 한다.

- **추정(estimation)**의 목적은 아래와 같이 크게 두 가지로 나눌 수 있다.
  1. 신이 정한 **참 값**을 알아내려고 한다. 예를 들어, 앞에 등장한 동전의 앞면이 나올 확률을 알아내려고 한다.
  2. 신이 정한 **참 값으로 인한 결과**를 알아내려고 한다. 예를 들어 특정한 동전을 10번 던졌을 때에 앞면이 7번 이상 나올 확률을 알아내려고 한다.

통계적 가설 검정 (Statistical Hypothesis Testing)

- 통계적 가설 검정은 **표본공간**의 관찰을 바탕으로, 이를테면 아래와 같은 결론을 내리는 것을 목표로 한다.
  - (애초에 신이 설계하기를) A 동전이 B 동전보다 앞면이 많이 나온다.
  - (애초에 신이 설계하기를) A 의약품이 B 의약품보다 효과가 좋다.
  - (애초에 신이 설계하기를) 일반적으로 사람들의 왼손의 크기보다 오른손의 크기가 크다.

- **신뢰 수준(confidence level)**이라는 개념을 추가하면 통계적 가설 검정의 결론을 얼마나 확신할 수 있는지를 포함한 결론을 내릴 수 있다. 예를 들어,
  - A 동전이 B 동전보다 앞면이 많이 나온다는 것을 90% 수준에서 확신할 수 있다.
  - A 의약품이 B 의약품보다 효과가 좋다는 것을 99% 수준에서 확신할 수 있다.
  - 일반적으로 사람들의 왼손의 크기보다 오른손의 크기가 크다는 것을 95% 수준에서 확신할 수 있다.

실험 계획 (Design of Experiment)

- 신이 관장하는 전체 집단에 해당하는 모집단 (Population)에 비해서 인간이 관찰할 수 있는 부분집합인 표본공간(Sample space)은 규모가 작다.
  - 만약에 모집단에 버금갈 정도로 넓은 표본공간이 주어진다면, 추정과 가설 검정에 있어서 더 높은 근거와 확신을 가질 수 있다.
  - 하지만 실험과 관찰에 비용이 수반되듯이, 넓은 표본공간을 확보하는 것은 때로는 돈, 시간, 인력을 포함한 높은 비용이 수반된다.

- 그렇기 때문에 적절한 규모의 표본공간을 구성하는 데에는 다음과 같은 고려가 필요하다. 이를 실험 계획 (Design of experiment)라고 한다.
  1. 결론을 내릴 수 있을 정도로 충분히 커야 한다.
  2. 비용을 최소화할 수 있도록 충분히 작은 표본공간을 구성하는 계획이 필요하다.

III. Datacamp

개요

- Python, R, SQL등의 프로그래밍 언어를 interactive한 tutorial 형식으로 배울수 있는 사이트입니다.
- E-class에 제공된 invitation link로 들어가서 @seoultech.ac.kr 계정을 사용하면 무료로 가입이 됩니다.
- 해당 링크는 이번학기 링크에 해당합니다. 매년 3월과 9월에 계정이 만료됩니다. 이번학기 수업이 끝난 이후에도 심민규 교수의 github.com 페이지에서 6개월 마다 업데이트 되는 새로운 링크를 제공합니다. (https://github.com/aceMKSim/teaching/)

과제

- 1개의 course는 약 4시간이 소요됩니다. 이번학기에 아래의 assignment가 부여됩니다.

- Introduction to R
- Intermediate R Course
- Introduction to Statistics in R
- Reporting with R Markdown
- Introduction to Regression in R
- Introduction to the Tidyverse

Survey

1. 이름/학번/이메일
2. 대학 전공
3. 간단한 직무 소개
4. 통계학 관련 수업 수강 경험
5. 프로그래밍 경험

6. R 사용 경험
   - 실무에서 사용
   - 간단한 프로젝트
   - 수업을 들은적 있다
   - 해본적 없다

7. 엑셀의 `averageif()` 혹은 `sumif()` 함수
   - 편하게 사용할 수 있다.
   - 충분한 시간이 주어진다면 할 수 있다.
   - 무엇을 하는 함수인지는 알 것 같다.
   - 전혀 모른다.

8. 엑셀의 `index()` - `match()` 함수
   - 편하게 사용할 수 있다.
   - 충분한 시간이 주어진다면 할 수 있다.
   - 무엇을 하는 함수인지는 알 것 같다.
   - 전혀 모른다.