**[Data analysis] 최선의 데이터 분석법, RCT (데이터 분석의 힘 chapter.2) — 나무늘보의 개발 블로그**

**노트북:** 첫 번째 노트북
**만든 날짜:** 2021-03-14 오후 11:40
**URL:** https://continuous-development.tistory.com/231

Data scientist/Data analysis

# [Data analysis] 최선의 데이터 분석법, RCT (데이터 분석의 힘 chapter.2)

2021. 3. 11. 21:36 수정 삭제 공개

# 최선의 데이터 분석법, RCT

여기서 가정을 해본다 '전력 가격을 올리면 절전으로 이어지는가?' 라는 가정이 있다.

여기서 인과관계는 가격 인상이 소비량에 어떤 영향을 미치는가 이다.

가격 인상 이후 A의 전력 소비량을 y1 이라고 하자
가격 인상이 없었을 경우의 A의 전력 소비량을 y2라고 한다.

루빈의 정의에 따르면 가격 인상이 y1과 y2 의 차이인 **개입효과**에 의해 정의할 수 있다.

하지만 두 가지 데이터를 관측하는 것은 실제로는 불가능하다 => 인과적 추론의 근본 문제 이기 때문이다.(만약은 의미가 없다)

이렇게 관측이 불가능한 결과를 '실제로는 일어나지 않은 잠재적 결과'(반사실의 잠재적 결과)라고 한다.

여기서 해결책은 **개입 집단과 비교 집단이라는 사고방식**이다.

루빈은 한 사람에 대한 개입 효과는 측정할 수 없지만 여러 사람에 대한 개입 효과를 평균한 값인 '평균 개입효과'는 측정할 수 있다고 한다.
개입 집단 - 개입을 받는 집단
비교 집단 - 개입을 받지 않는 집단

이 두 집단을 사용해서 평균적인 개입 효과를 알 수 있다.

자 여기까지 우리는 두 집단(개입 집단과 비교 집단)을 이용하면 평균 개입 효과를 알 수 있다는 사실을 알게 되었다.
하지만 무턱대고 집단을 나눈다고 가능하지는 않다. 집단을 잘못 나누는 예를 생각해보자.

희망에 따라 개입 - 두 집단의 근본적인 차이가 있는 경우(미국에서 영어를 쓸 확률과 한국에서 영어를 쓸 확률)

애초에 근본적으로 다를 가능성이 높다. 이처럼 자신의 의지로 개입을 받아 들이느냐 마느냐를 판단하는 것을 자기 선택이라고 부른다.

자기 선택에 의해 형성된 집단은 다른 특성을 가질 확률이 높다.

자 이제 여기서 우리는 생각한다. 자기 선택을 가지지 않고 집단을 나누면 될 것 같다라는 생각이 든다.

## 이것이 바로 **무작위 비교 시행(RCT)**이다.

여기서 핵심은 집단을 나눌 때 반드시 무작위로 나눈다는 것이다. 이 집단을 A와 B 집단으로서 AB 테스트를 시행한다.

무작위로 집단을 나누게 된다면 중심 극한 정리 (동일한 확률분포를 가진 독립 확률 변수 n개의 평균의 분포는 n이 적당히 크다면 정규분포에 가까워진다는 **정리**) 에 따라 일정량의 표본만 충족한다면 두 집단의 통계는 정규 분포에 가까워진다. 이게 핵심이다.

두 개의 집단이 비슷한 분포를 가진다는 것은 통계적 동일 집단일 확률이 높아지며 집단 간의 동질성이 확보 된다.

더욱 더 디테일한 설명은 책을 참조하길 바란다.

RCT를 하기 위해서는 세가지 원칙이 반드시 지켜져야 한다.

## 1.적절하게 집단을 나눈다.

- > 해결하려는 문제의 답이 나오도록 집단을 적절하게 나눠야 한다.

## 2.집단은 반드시 무작위로 나눈다.

-> ex) 아까 나눈 것처럼 기준을 잡고 나누게 되면 의미가 없게 된다.

## 3.집단별로 충분한 표본수를 채운다.

-> 10명을 가지고 표본을 할 경우 이 값이 이상치에 큰 영향을 받을 확률이 높다 => 표본수가 클수록 평균값 계산에 표준오차가 작아지고 평균값의 신뢰성이 커진다.

이러한 RCT로 얻은 결과를 검증하고 비교하기 위해 가장 흔히 쓰이는 통계분석은 개입효과의 평균값을 분석하는 것이다.

1.실험 후 집단 별로 평균 값을 계산한다
2.평균 값의 차이를 비교한다.

## RCT의 강점과 약점

강점

- 인과관계를 과학적으로 보여준다.
- 분석 기법과 결과가 투명하다.

단점

- 비용, 시간, 노력이 많이 들고 각 기관의 협력도 필요하다.

---

이 내용들은 **데이터 분석의 힘** 이라는 책의 내용을 정리 및 요약한 내용입니다.

'Data scientist > Data analysis' 카테고리의 다른 글

**[Data analysis] 최선의 데이터 분석법, RCT (데이터 분석의 힘 chapter.2)**

[Data analysis] 인과 관계와 상관 관계 (데이터 분석의 힘 chapter.1)

rct RCT란 데이터 분석의 힘 무작위 비교 시행

나아무늘보

혼자 끄적끄적하는 블로그 입니다.