금융인을 위한

통계와 데이터 분석 입문

데이터의 형태 이해하기

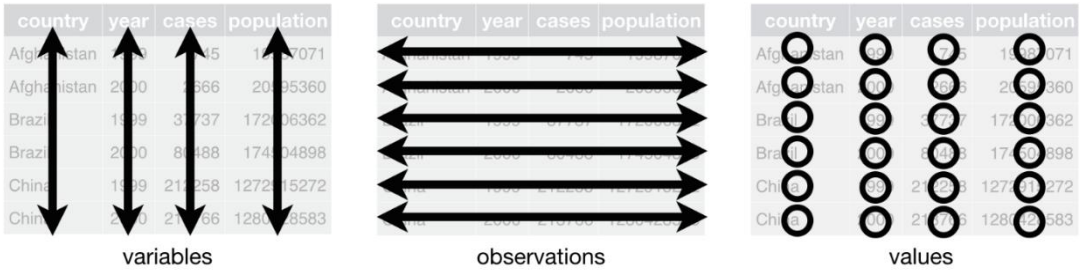
한국금융연수원
KOREA BANKING INSTITUTE

학습 내용

1. 데이터의 구성
2. 데이터를 구성하고 있는 변수의 종류
   - 수치형 변수, 범주형 변수
3. 변수의 종류에 따른 대표적인 통계적 분석 기법들
4. 탐색적 자료분석이란?
5. (실습) 파이썬을 활용한 데이터 구조 파악

데이터의 구성

- 데이터를 구성하고 있는 것은 변수(variable)들로
  변수는 속성(attribute), 특징(feature)이라고도 불림
- 변수들에 대해 관측치(observation) 별로 값이 입력됨
- 일반적으로 가로 방향은 변수, 세로 방향은 관측치를 나타냄
  → 데이터의 크기는 변수의 개수와 관측치의 개수로 결정됨



출처 https://r4ds.had.co.nz/tidy-data.html

범주형 변수

- 변수는 크게 범주형 변수와 연속형 변수로 구분

범주형 변수

변수의 값이 특정 범주를 나타냄

- 서열이 있는 경우

  예 성적(A, B, C, D, F),
  경제적 수준(상, 중, 하), 만족도(나쁨, 보통, 좋음) 등

- 서열이 없는 경우

  예 성별(남, 여), 인종(황인종, 흑인종, 백인종),
  혈액형(A, B, O, AB),
  산업군(제조업, 서비스업, 금융업 등) 등

  - 숫자로 표기되기도 하지만 구분하기 위한 목적일 뿐 크기를 의미하지 않음

연속형 변수

연속형 변수

변수의 값이 **크기**나 **양**을 나타내고 숫자로 표현됨

예

1인당 국민 총생산, 가구당 월평균 소득,
일일 주가지수 수익률 등

- 일반적으로 범주형 변수, 연속형 변수가 혼재되어 데이터에 포함되어 있음

변수의 종류에 따른 대표적인 통계적 분석 기법들

연속형 변수

T-test

가정한 값과 통계적으로 유의한 **평균 차이**가 있는지를 검정하는 방법

예

특정 주가 지수의 평균 수익률이 3%보다 통계적으로 유의하게 큰지 여부

변수의 종류에 따른 대표적인 통계적 분석 기법들

연속형 변수

회귀분석(Regression)

예측하고자 하는 연속형 변수(종속변수)와
다른 변수(독립변수)들 간의 **관계**를 나타내는 분석 방법

예

주택 가격에 영향을 줄 수 있는 여러가지 변수들
(학군, 범죄율, 역세권, 공원근접성 등)과
주택 가격 간의 관계를 찾고 주택가격 예측

- 독립변수에 범주형 변수가 포함될 수도 있음

변수의 종류에 따른 대표적인 통계적 분석 기법들

범주형 변수

독립성 검정

두 범주형 변수가 **독립**인지 아닌지를 검정하는 방법

예

「학력」(초등졸, 중등졸, 고등졸, 대졸, 대학원졸)이라는
범주형 변수와 「연소득」(상, 중, 하)이라는
범주형 변수가 서로 관련 있는지 아니면 독립인지 여부

변수의 종류에 따른 대표적인 통계적 분석 기법들

연속형, 범주형 변수

분산분석(ANOVA)

3개 이상의 범주로 나뉘어진 범주형 변수의 값에 따라 관심있는 연속형 변수의 **평균에 차이가 있는지**를 검정하는 방법

예

3개로 나뉜 고객의 등급에 따라 서비스 만족도 평균에 통계적으로 유의한 차이가 있는지 여부

변수의 종류에 따른 대표적인 통계적 분석 기법들

연속형, 범주형 변수

로지스틱 회귀분석

대표적인 분류(classification) 기법으로
예측하고자 하는 범주형 변수(종속변수)와
다른 변수(독립변수)들과의 관계를 나타내는 분석 방법

예

대출 상환 여부 예측, 온라인 광고에서 클릭 여부 예측

- 로지스틱 회귀 분석 이외에도 여러가지 분류 방법 존재
  - 서포트 벡터머신(support vector machine),
    의사결정나무(decision tree) 등

탐색적 자료분석(exploratory data analysis, EDA)

탐색적 자료분석

통계적 분석 기법을 적용하기 전에 데이터에 포함된
변수들을 **요약**하고 **시각적으로 표현**해보는 과정

- 데이터에 대한 감각을 익힐 수 있음
- 통계 분석 기법들에는 가정(assumption)이 포함되어 있음
  - 데이터의 분포 및 요약값들을 검토하면서
    가정에 대한 확인 가능
- 다양한 각도에서 데이터를 살펴보는 과정을 통해
  문제 정의 단계에서 미처 고려하지 못한 다양한 패턴을
  발견하고 문제를 수정할 수 있음

학습정리

- **데이터**는 **변수**와 **관측치**로 구성되어 있음
- 변수의 종류에는 **수치형 변수**, **범주형 변수**가 있고
  어떤 변수들이 분석 대상이 되는지에 따라서 활용할 수 있는
  통계적인 분석 기법들이 달라질 수 있음
- 실제 통계 분석 기법을 적용하기 전에 **데이터를 요약**하고
  시각적으로 탐색하는 과정이 필요함
- 파이썬의 pandas의 여러 가지 함수들
  (shape, columns, info, head, tail, isnull 등)을 활용하여
  데이터의 구조를 파악할 수 있음