

데이터통계분석

김영주
wrapup@dongyang.ac.kr

교수 소개

- 김영주
- 회사: (주) 푸드테크
- 부서: Data Intelligence



데이터 직군 비교

같은 회사 안에서도 경계가 흐릿하다. 기업마다 다르므로 채용공고의 직무기술서를 확인한다.

직군	핵심 역할	하는 일	초점
데이터 분석가	데이터 정리·분석·시각화	Raw 데이터 정리 및 구성 · 기술 통계를 활용한 분석 · 해석과 시각화·대시보드 구성 · 분석 결과를 비즈니스 팀에 발표	과거와 현재의 데이터 이해
데이터 과학자	머신러닝·AI 알고리즘 모델을 개발해 예측 제공	통계 모델 평가 · 머신러닝으로 예측 알고리즘 구축 · 모델의 지속적 학습 및 최적화 운영	미래 예측 생성
데이터 엔지니어	데이터 플랫폼과 파이프라인 아키텍처 개발·운영	데이터 소비를 위한 API 구축 · 데이터 플랫폼 아키텍처 구성 · 파이프라인 개발 · 운영 · 머신러닝 모델 feature store 제공	소프트웨어 개발·분산처리 기술 의존도 높음

출처: "데이터분석가 vs 데이터엔지니어 vs 데이터과학자 차이가 뭘까? (1) 역할과 정의" <https://socrates-dissatisfied.tistory.com/5>

무엇을 배우나요? : 데이터 통계 분석

■ 이론 → 수식 → 실습의 3단계로 데이터를 분석

기초 통계학에 대한 핵심적인 이론을 학습하고, 수식을 통해 개념이 의미하는 바를 파악합니다. 그런 다음, 파이썬으로 실습하면서 다양한 형태의 결과물(값, 표, 그림 등)을 살펴봅니다.

■ 복잡한 수식 없이 데이터 분석의 기초를 학습

파이썬을 활용해 복잡한 수식이 없이도 통계분석을 학습할 수 있습니다.

■ 파이썬 소스코드 제공

- 책에 수록된 파이썬 소스코드는 한빛 아카데미 홈페이지

(https://www.hanbit.co.kr/support/supplement_list.html) 를 통해 다운받을 수 있습니다.

AI 시대에 데이터 통계학, 왜 배우나

분석은 맡길 수 있다. 무엇을 시킬지, 결과가 맞는지, 왜 그런지는 여전히 사람 몫이다.

1 질문을 정확히 시킬 수 있다 — "데이터 분석해줘"는 내가 원하는 분석을 주지 않는다

2 용어가 곧 압축이다 — "이상치 빼고 중앙값으로" 한 줄이 설명 수십 줄을 대신한다

3 지표를 고를 줄 안다 — 같은 데이터도 무엇으로 보느냐에 따라 다른 이야기가 된다

4 결과를 검증한다 — 그럴듯한 답과 맞는 답은 다르다

5 설명 책임을 진다 — "AI가 그랬는데요"는 근거가 아니다

같은 데이터, 다른 얼굴 — 평균과 최빈값

소득처럼 한쪽에 몰리고 꼬리가 긴 분포에서는, 대푯값마다 서로 다른 값을 가리킨다.

지표	무엇을 말하나	예시의 값	무엇을 조심하나
평균	모두 더해 사람 수로 나눈 값	380만원	소수의 고소득에 끌려 올라간다
중앙값	줄 세웠을 때 한가운데 사람의 값	200만원	극단값에 잘 흔들리지 않는다
최빈값	가장 많이 나온 값	200만원	가장 흔한 사람의 몫에 가깝다

예시 — 10명 중 9명이 월 200만원, 1명이 2,000만원을 번다. 평균만 보면 "월 380만원 버는 집단"이 된다. 어떤 지표를 고르느냐가 곧 어떤 이야기를 하느냐다.

"AI가 그랬는데요"는 근거가 아니다

분석을 맡길 수는 있어도, 설명 책임은 넘길 수 없다.

- 1 "근거가 뭐야?" — 어떤 데이터를 어떤 방법으로 봤는지 말할 수 있어야 한다
- 2 "이게 뭘 뜻하는데?" — 숫자를 의사결정의 언어로 옮기는 일은 사람이 한다

통계는 AI를 대체하는 지식이 아니라, AI를 부리고 그 결과를 책임지는 언어다.

- 3 "다른 해석은 없어?" — 지표를 바꿔 보고 반례를 찾으려면 개념을 알아야 한다

- 4 "AI한테 물어보고 오겠습니다" — 회의는 그 자리에서 멈춘다

평가

평가 항목	평가 비율(%)
중간 고사	30
기말 고사	40
과제	10
출결	20

강의 계획 (실습 언어 : Python)

1 교과목 소개 및 데이터의 개념과 변수

2 [기본통계학] 1차원 데이터 - 지표 계산, 시각화 실습

3 [기본통계학] 2차원 데이터 - 지표 계산, 시각화 실습

4 추측통계의 기본

5 이산형 확률변수와 대표적인 이산형 확률분포

6 연속형 확률변수 와 대표적인 연속형 확률 분포

7 중간고사 전 내용 정리 및 회고

8 중간 고사

9 독립동일분포

10 통계적 추정

11 통계적 가설검정

12 회귀분석 - 단순회귀모형, 중회귀모형

13 회귀분석 - 모형의 선택, 모형의 타당성

14 기말고사 전 내용 정리 및 회고

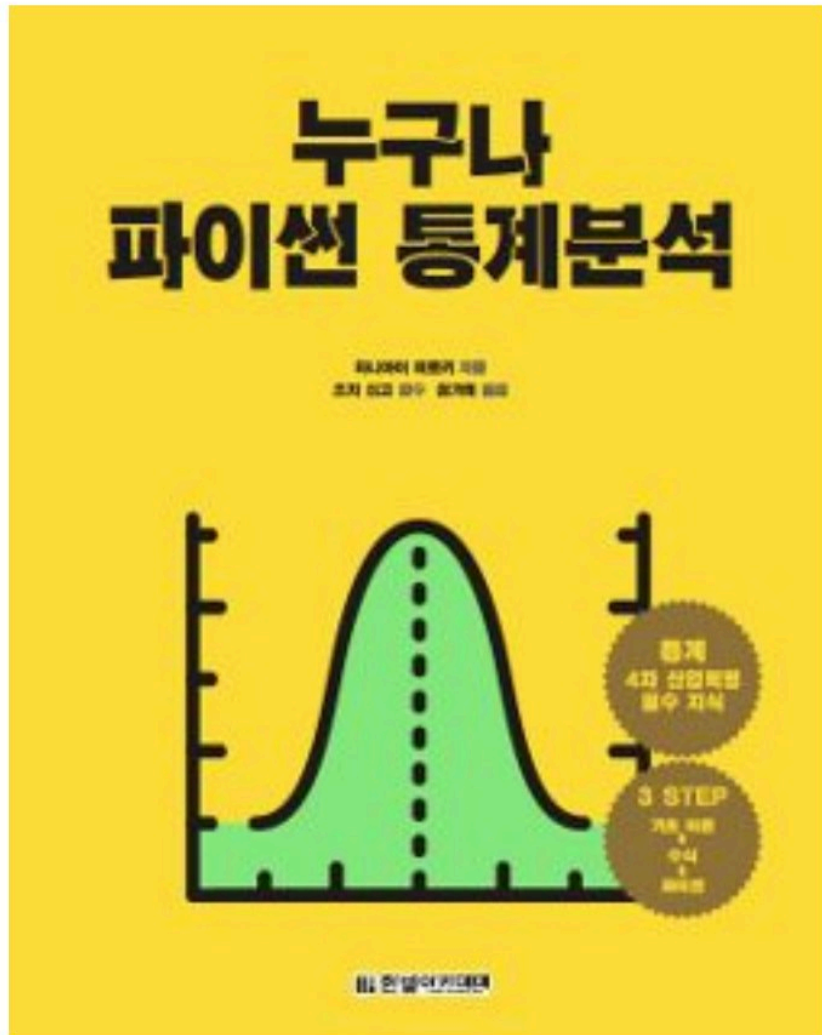
15 기말 고사

중간,기말 고사

- 중간 고사, 기말 고사

문제에 대한 답안 작성, 실기 시험(Jupyter notebook)

- 과제 (미정)



- 도서명: 누구나 파이썬 통계분석
- ISBN: 979-11-5664-488-0 93310
- 저자: 타니아이 히로키
- 감수자: 츠지 신고
- 역자: 권기태
- 출판사: 한빛아카데미(주)
- 페이지: 384p
- 정가: 26,000원

CHAPTER 01

데이터

CONTENTS

- 1.1 데이터의 크기
- 1.2 변수의 종류
- 1.3 정리

데이터: 실습 데이터 살펴보기

표 데이터 처리에 특화된 통계분석 라이브러리 Pandas의 `read_csv` 함수를 사용하여 체력 테스트 결과를 담은 `csv` 파일을 읽어들이м

In [1]:

```
# Pandas를 pd라는 이름으로 임포트
import pandas as pd
```

In [2]:

```
# 학생번호를 인덱스로 하여 csv 파일을 읽어 들여, 변수 df에 저장
df = pd.read_csv( '../data/ch1_sport_test.csv',
                  index_col= '학생번호' )
```

```
# 변수 df를 표시
df
```

학생번호	학년	악력	윗몸일으키기	점수	순위
1	1	40.2	34	15.4	
2	1	34.2	14	7.10	
3	1	28.8	27	11.7	
4	2	39.0	27	14.5	
5	2	50.9	32	17.2	
6	2	36.5	20	9.9	
7	3	36.6	31	13.6	
8	3	49.2	37	18.1	
9	3	26.0	28	10.8	
10	3	47.4	32	16.3	

참고: 엑셀 포맷 - <https://xlsxwriter.readthedocs.io/worksheet.html>

데이터: 실습 데이터 살펴보기

체력 테스트 결과를 DataFrame으로 읽어들이는 결과

Out [2] :

학생번호	학년	악력	윗몸일으키기	점수	순위
1	1	40.2	34	15	4
2	1	34.2	14	7	10
3	1	28.8	27	11	7
4	2	39.0	27	14	5
5	2	50.9	32	17	2
6	2	36.5	20	9	9
7	3	36.6	31	13	6
8	3	49.2	37	18	1
9	3	26.0	28	10	8
10	3	47.4	32	16	3

학생번호,학년,악력,윗몸일으키기,점수,순위

1,1,40.2,34,15,4
2,1,34.2,14,7,10
3,1,28.8,27,11,7
4,2,39.0,27,14,5
5,2,50.9,32,17,2
6,2,36.5,20,9,9
7,3,36.6,31,13,6
8,3,49.2,37,18,1
9,3,26.0,28,10,8
10,3,47.4,32,16,3

데이터: 실습 데이터 살펴보기

DataFrame에서 악력에 대한 열을 추출하면 1차원 데이터 구조인 Series 반환

In [3]:

```
df[ ' 악력 ' ]
```

Out [3]:

학생번호

1	40.2
2	34.2
3	28.8
4	39.0
5	50.9
6	36.5
7	36.6
8	49.2
9	26.0
10	47.4

Name: 악력, dtype: float64

학생번호,학년,악력,윗몸일으키기,점수,순위

1	1	40.2	34	15	4
2	1	34.2	14	7	10
3	1	28.8	27	11	7
4	2	39.0	27	14	5
5	2	50.9	32	17	2
6	2	36.5	20	9	9
7	3	36.6	31	13	6
8	3	49.2	37	18	1
9	3	26.0	28	10	8
10	3	47.4	32	16	3

데이터: 데이터의 크기

DataFrame의 크기는 shape라는 인스턴스 변수 참조

```
In [4]:  
df.shape  
  
Out [4]:  
(10, 5)
```

데이터 개수 10 (레코드 수), 변수 개수 5 (컬럼 수)

```
학생번호,학년,약력,윗몸일으키기,점수,순위  
1,1,40.2,34,15,4  
2,1,34.2,14,7,10  
3,1,28.8,27,11,7  
4,2,39.0,27,14,5  
5,2,50.9,32,17,2  
6,2,36.5,20,9,9  
7,3,36.6,31,13,6  
8,3,49.2,37,18,1  
9,3,26.0,28,10,8  
10,3,47.4,32,16,3
```

확인 문제 — 데이터를 표로 읽고 크기를 세는 법

시험에 이런 형태로 나올 수 있는 문제입니다. 문제를 띄우면 답 장으로 넘기기 전에 먼저 손으로 풀어 보세요.

1 shape 가 (10, 5) 인 데이터에서 10 과 5 는 무엇인가?

2 DataFrame 에서 열 하나를 꺼내면 무엇이 되나?

shape 가 (10, 5) 인 데이터에서 10 과 5 는 무엇인가?

체력 테스트 결과를 DataFrame 으로 읽었더니 shape 가 (10, 5) 였습니다.

shape 가 (10, 5) 인 데이터에서 10 과 5 는 무엇인가?

10 = 데이터 개수(레코드, 행), 5 = 변수 개수(컬럼, 열)

shape 는 (행, 열) 순서입니다. 한 행이 학생 한 명, 한 열이 악력·윗몸일으키기 같은 측정 항목 하나입니다. 순서를 거꾸로 적는 실수가 가장 많습니다.

DataFrame 에서 열 하나를 꺼내면 무엇이 되나?

약력 열 하나만 꺼냈습니다.

- ① DataFrame — 2차원 그대로
- ② Series — 1차원
- ③ 숫자 하나
- ④ 리스트 5개

DataFrame 에서 열 하나를 꺼내면 무엇이 되나?

2번 — Series (1차원)

열 하나는 한 변수의 값을 사람 수만큼 늘어놓은 것이라 1차원입니다. 02장의 1차원 데이터 정리가 바로 이 Series 하나를 요약하는 일입니다.

Q&A

다음 강의

- 교재 : 02장 1차원 데이터 정리