[엑셀과 파이썬을 이용한 통계분석]

[1] 통계

1. 통계

: 정보 또는 데이터를 규정하고 구성하며 해석하는 데 사용되는 여러 도구와 기법

1) 기술통계 : 수집된 자료의 특성을 정리 및 요약하여 설명(describe)

Ex. 평균, 최빈값

2) 추론통계(우리가 주로 하는) : 표본으로 모집단을 추정하는 것 – 가설검정

2. 기본 용어의 정의

- 자료와 단위

1) 자료 : 분석 대상 관찰 속성을 기록한 수열

2) 관찰단위 : 자료 관찰 및 수집 단위 ex. 가구

3) 분석단위 : 분석 수행 및 발견이 일반화 될 수 있는 단위 ex. 가구원

3. 변수(variable) : 변동하는 수량, 변할 수 있는 숫자 <-> 상수

Ex. 변수 : 연령, 주소(데이터 분석의 대상) / 상수 : 국적(데이터 분석 대상이 아님)

1) 모델

a. 독립변수 : x변수, 설명변수

b. 종속변수 : y변수, 반응변수

2) 특성

a. 양적변수 = 수치형 변수 : 등간척도, 비율척도

-등간 : 절대0 없음(온도가 0도라고 해서 온도가 없는 것이 아님)

-비율 : 절대0 있음(소득이나 무게가 0이면, 그것이 없음을 뜻함)

-이산변수, 연속변수

b. 질적변수 = 범주형 변수 : 명목척도, 서열척도(서열 ex. 설문조사 – 리커트 척도)

4. 자료의 이해

1) 자료의 구조

: 행렬 구조(matrix)

-Pandas에서는 DataFrame / Oracle에서는 Table

-열 = 변수 / 행 = 관측치

ID	famale	socialclass	IQ	income
1	1	1	135	250
2	0	2	110	310

(변수 : 4개 / 관측치 : 2개)

-ID는 고유키(ex. 주민등록번호)이기 때문에, 변수로 취급하지 않음!

-성별에서 1-여자, 0-남자임 > '여자'를 기준으로 더미변수를 만든 것 > 범주형 변수

-socialclass : 순위

2) 자료 유형의 중요성

-자료 유형에 따라 적절한 통계분석의 접근방법이 달라짐

ID	famale	socialclass	IQ	income
1	1	1	135	250
2	0	2	110	310
3	1	3	128	1500
4	0	2	98	122
5	1	2	106	450
6	0	1	102	190

(2종의 자료 유형 : 범주, 수치 / 4종의 척도 : 'y/n', group, IQ테스트, 소득 단위)

-범주형 : famale - 명목 / socialclass – 순위

-수치형 : IQ - 등간 / income – 비율

5. 수체계의 속성에 따른 유형 분류 **범주/수치 구분!

1) 정체성 : 숫자마다 구분할 수 있는 특별한 의미 – 범주형 中 명목

2) 크기 : 내재된 순서에 따라 크고 작음을 판단 – 범주형 中 순위

3) 간격 : 숫자 사이의 차이 크기가 일정 – 수치형 中 등간

4) 절대영점 : 절대 0이라는 기준점 – 수치형 中 비율

6. 척도의 활용

: 동일 개념을 서로 다른 유형의 척도로 측정 가능

-양적 변수 -> 질적변수 변환 가능 (반대는 불가능)

Ex. 나이(양적) -> 연령대(질적) 변환 가능!

-양적변수 : 평균으로 분석 가능 / 질적변수 : 빈도만 산출 가능

-정보량의 크기(정확성의 정도) : 비율 > 등간 > 서열 > 명목

7. 자료의 수치 요약

1) 중심경향의 측도 : 평균, 중앙값, 최빈값

2) 산포의 측도 : 범위, 분산, 표준편차

3) 비대칭의 측도 : 왜도, 첨도

[2] 기술통계

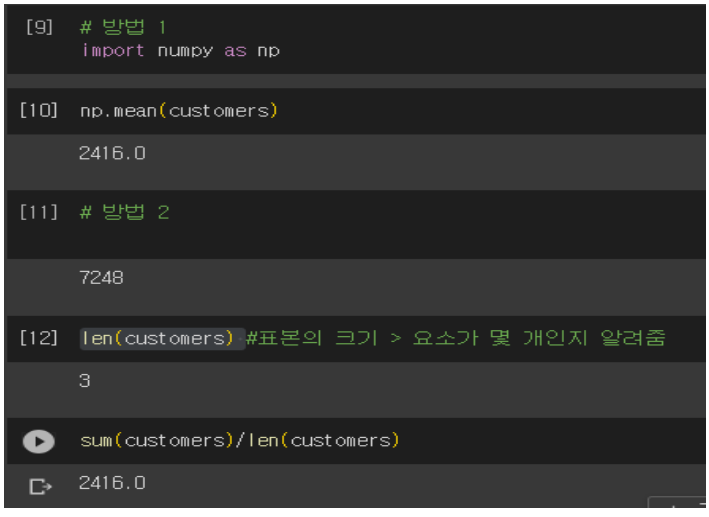
1. 중심경향의 측도

1) 평균 (보통 산술평균을 이야기 함 > 가장 많이 사용됨)

-종류 : 산술평균, 가중평균, 기하평균, 조화평균

-모든 값의 합계를 값의 개수로 나눈 값 : $\overline{X}$ = ∑X / n

+) $\overline{X}$ (X bar) : 통계량 : 표본평균 / ∑X에서 X : 관측치 / n : 표본크기



-평균에 대해 몇 가지 기억해야 할 사항

a. 표본평균 : $\overline{X}$ / 모평균 : μ(mu)

b. 표본의 크기 : n / 모집단의 크기 : N

c. 평균은 이상치에 미감

-극단적으로 높은 값이 존재하면 > 평균이 높아짐

-극단적으로 낮은 값이 존재하면 > 평균이 낮아짐

-데이터에 이상치가 있으면 평균이 왜곡됨 > 대표성에 문제가 생김 > 이상치 꼭 살펴보기

-해결 방법 : 중앙값과 함께 고려하기

d. 편차 = 데이터 값 – 평균

-편차의 합 = 0

```
[14]  x = [3,4,5]

[17]  x_bar = np.mean(x);x_bar
      4.0

[22]  x_dev = x - x_bar;x_dev
      array([-1.,  0.,  1.])

[23]  sum(x_dev)
      0.0
```

2) 가중평균 = ∑ (데이터 값(관측치)*빈도수) / ∑표본의 개수

-빈도수 : 가중치

```
[24]  score = pd.Series([97,94,92,91,90,89,78,60])
      freq = pd.Series([4,11,12,21,30,12,9,1])

[26]  # 점수*빈도
      score_freq = score*freq

[27]  # 분자
      sum(score_freq)
      8967
```

```
[28]  # 분모
      sum(freq)
      100

[29]  # 가중평균 = 분자 / 분모
      sum(score_freq)/sum(freq)
      89.67

[31]  # 산술평균
      score.mean()
      86.375
```

+) 기하평균 (=비율평균)

Ex. 수익률(산술평균이 아닌, 기하평균을 적용함)

-주식 등 자산 = 포트폴리오

-100만원 투자

(1) 이익 50% : 100+50=150

(2) 손해 50% : 150-75=75

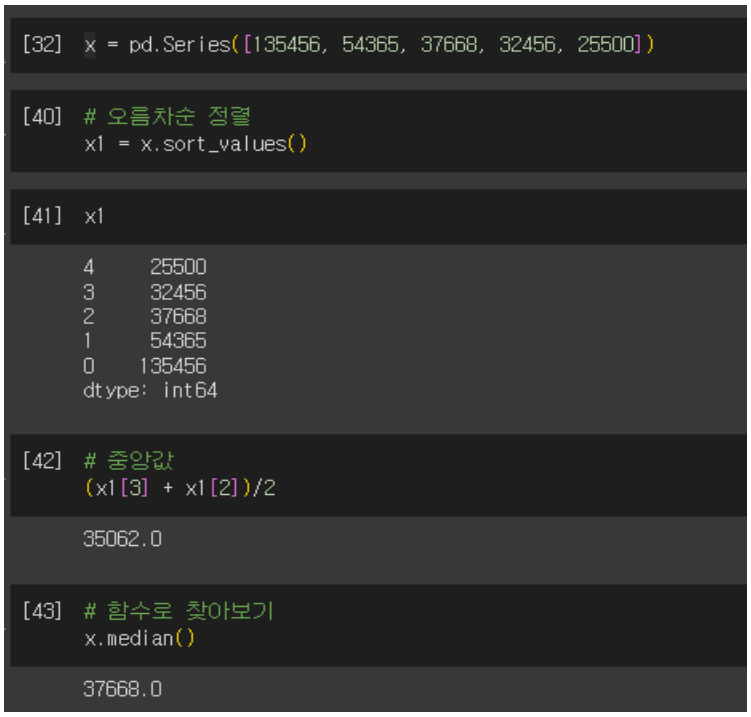
+) 조화평균 (F1 score의 공식)

3) 중앙값

-오름차순(작은 값 > 큰 값)으로 나열

a. 데이터 개수가 홀수 : 중간에 있는 값

b. 데이터 개수가 짝수 : 중간에 있는 두 값의 평균



2. 산포의 측도

1) 위치의 측도

-사분위수(quantile) : 자료의 위치를 25%씩 4개로 구분

-백분위수(percentile) : 자료의 위치를 1% 씩 100개로 구분

a. 중앙값 = 2번째 사분위수 = 50번째 백분위수

b. 1번째 사분위수 = 25번째 백분위수

c. 3번째 사분위수 = 75번째 백분위수