Chapter. 09

변수가 어떻게 생겼나: 기초 통계 분석

기초 통계 분석을 해야 하는 이유

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 확률 변수의 정의

- 변수 (variable): 특정 조건에 따라 변하는 값

- 확률 변수 (random variable): 특정 값(범위)을 확률에 따라 취하는 변수
  - ➢ 예시: 주사위를 던졌을 때 나오는 결과를 나타내는 변수 X

값	1	2	3	4	5	6
확률	1/6	1/6	1/6	1/6	1/6	1/6

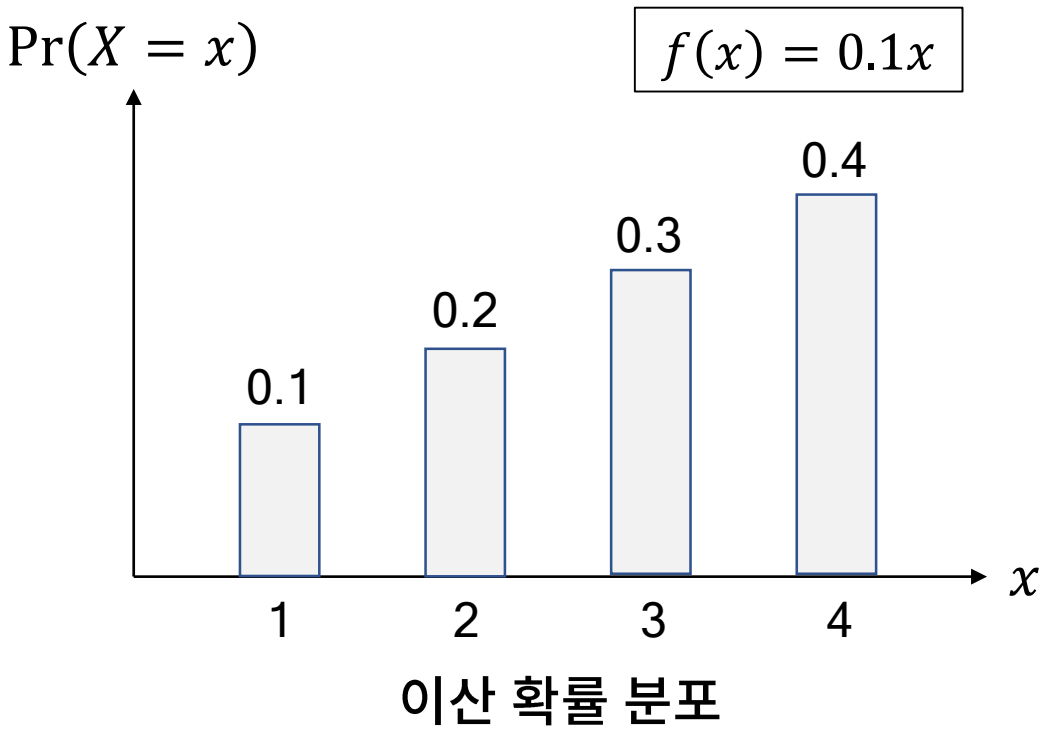
➡ 상태 공간

- 상태 공간의 크기가 무한한 변수를 연속 확률 변수, 유한한 변수를 이산 확률 변수라고 함

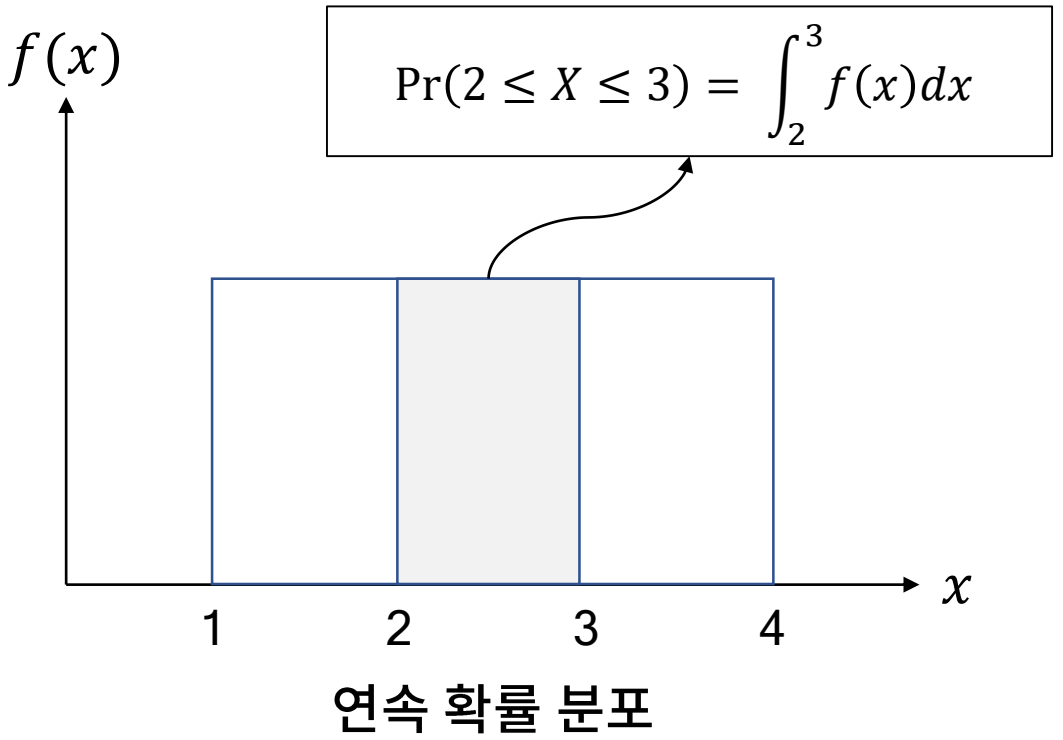
- 우리가 관측하는 데이터에 있는 변수는 특별한 경우를 제외하곤 모두 확률 변수임

I 확률 분포의 정의

- 확률 분포 (probability distribution)는 확률 변수가 특정한 값을 취할 확률을 나타내는 **함수**를 의미



이산 확률 분포



연속 확률 분포

I 확률 분포의 확인 방법

- 한 변수가 따르는 확률 분포를 확인했을 때의 효과
  - (1) 현재 수집한 데이터가 어떻게 생겼는지를 이해할 수 있음
  - (2) 새로 데이터가 들어오면 어떻게 들어올 것인지 예상할 수 있음

- 그러나 가지고 있는 데이터는 샘플 데이터이므로 절대로 **정확히 한 변수가 따르는 확률 분포를 알 수 없음**

- 그래프를 이용하여 확인하거나 적합성 검정을 사용하여 확률 분포를 확인해야 하는데, 이 작업은 굉장히 많은 노력이 필요함

I 통계량의 필요성: 간단하게 확률 분포 확인

- 통계량은 **확률 분포의 특성**을 나타내는 지표를 의미함
- 통계량을 계산하는 **기초 통계 분석 (기술 통계 분석)**을 바탕으로 확률 분포를 **간단하게 확인** 가능함 (단, 반드시 각 통계량이 나타내는 의미를 이해해야 함)
- 특히, 변수가 많은 경우에 훨씬 효율적으로 사용 가능함

I 통계량의 종류

- 통계량은 크게 대표 통계량, 산포 통계량, 분포 통계량으로 구분할 수 있음

구분	내용	예시
대표 통계량	데이터의 중심 및 집중경향을 나타내는 통계량	평균, 최빈값 등
산포 통계량	데이터의 퍼진 정도를 나타내는 통계량	분산, 범위, 표준편차
분포 통계량	데이터의 위치 정보 및 모양을 나타내는 통계량	왜도, 첨도, 사분위수, 최대값, 최소값

Chapter. 09

변수가 어떻게 생겼나: 기초 통계 분석

대표 통계량

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 평균

- 산술 평균: 가장 널리 사용되는 평균으로 연속형 변수에 대해 사용

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

x_i : i번째 관측치

n : 관측치의 개수

➢ 이진 변수에 대한 산술 평균은 **1의 비율**과 같음

➢ 다른 관측치에 비해 매우 크거나 작은 값에 크게 영향을 받음

- 조화 평균: 비율 및 변화율 등에 대한 평균을 계산할 때 사용 (데이터의 역수의 산술 평균의 역수)

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

- 절사 평균: 데이터에서 α ~ $1-\alpha$의 범위에 속하는 데이터에 대해서만 평균을 낸 것

$$\frac{\sum_{i=\lfloor n\times\alpha \rfloor}^{\lfloor n\times(1-\alpha) \rfloor} x_i}{\lfloor n\times(1-\alpha) \rfloor - \lfloor n\times\alpha \rfloor}$$

x_i : i번째 관측치

n : 관측치의 개수

➢ 매우 크거나 작은 값에 의한 영향을 줄이기 위해 고안됨

파이썬을 이용한 평균 계산

구분	구현 코드
산술 평균	• numpy.mean(x) • numpy.array(x).mean() • Series(x).mean()
조화 평균	• len(x) / numpy.sum(1/x) • scipy.stats.hmean(x)
절사 평균	• scipy.stats.trim_mean(x, proportiontocut) ➢ proportiontocut: 절단할 비율

I 최빈값

- 한 변수가 **가장 많이 취한 값**을 의미하며, **범주형 변수**에 대해서만 적용
- 파이썬을 이용한 최빈값 계산
  - scipy.stats.mode(x)
  - Series.value_counts().index[0] (주의: 최빈값이 둘 이상이면 사용 불가)

Chapter. 09

변수가 어떻게 생겼나: 기초 통계 분석

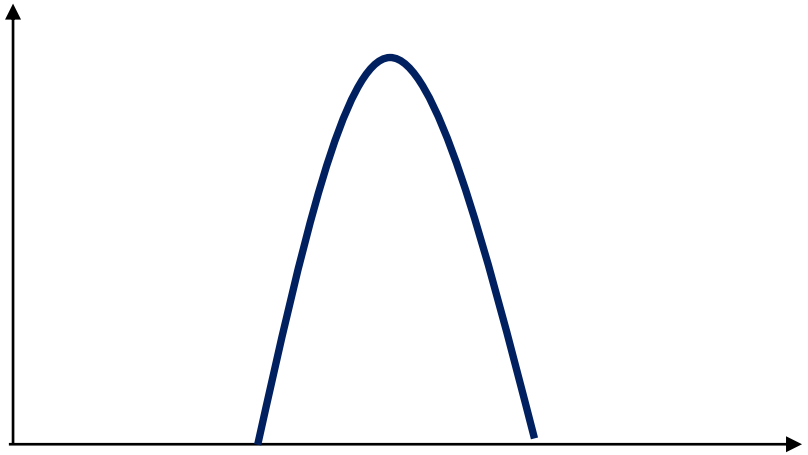
산포 통계량

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

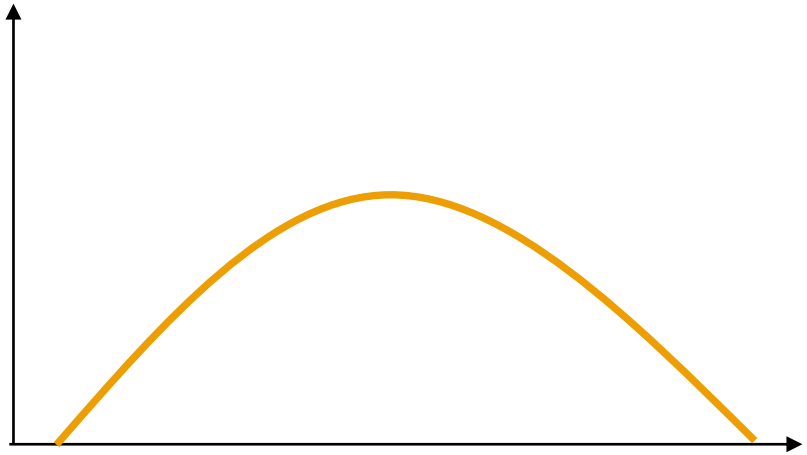
강사. 안길승

I 개요

- 산포란 **데이터가 얼마나 퍼져있는지**를 의미함



산포가 작은 변수



산포가 큰 변수

- 즉, 산포 통계량이란 데이터의 산포를 나타내는 통계량이라고 할 수 있음

I 분산, 표준편차

- 편차: 한 샘플이 평균으로부터 떨어진 거리, $x_i - \mu$ (x_i : i번째 관측치, μ : 평균)

- 분산: 편차의 제곱의 평균

$$\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

x_i : i번째 관측치
μ : 평균
n : 관측치의 개수

➢ 편차의 합은 항상 0이 되기 때문에, 제곱을 사용
➢ 자유도가 0 (모분산)이면 $n-1$으로 나누지 않고, n으로 나눔

- 표준편차: 분산에 루트를 씌운 것

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

x_i : i번째 관측치
μ : 평균
n : 관측치의 개수

➢ 분산에서 제곱의 영향을 없앤 지표

I 변동계수

- 분산과 표준편차 모두 값의 스케일에 크게 영향을 받아, 상대적인 산포를 보여주는데 부적합함

데이터	(1, 2, 3, 4, 5)	(10, 20, 30, 40, 50)
표준편차	1.414	14.14

- 따라서 변수를 스케일링한 뒤, 분산 혹은 표준편차를 구해야 함
- 만약 모든 데이터가 양수인 경우에는 변동계수(상대 표준편차)를 사용할 수 있음
  - ➢ 변동계수 = 표준편차 / 평균

파이썬을 이용한 분산, 표준편차, 변동계수 계산

구분	구현 코드
분산	• numpy.var(x, ddof) • numpy.array(x).var(ddof) • Series(x).var(ddof) # ddof: 자유도
표준편차	• numpy.std(x, ddof) • numpy.array(x, ddof).std() • Series.std(x, ddof) # ddof: 자유도
변동계수 계산	• numpy.std(x, ddof) / numpy.mean(x) • scipy.stats.variation(x) # ddof: 자유도

I (Tip) 둘 이상의 변수의 값을 상대적으로 비교할 때: 스케일링

- 국어 점수가 90점인 학생과 수학 점수가 80점인 학생 중 누가 더 잘했나?
  - ➢ 국어 점수 변수와 수학 점수 변수의 분포가 다르기 때문에 정확히 비교가 힘듦
  - ➢ (예) 국어 점수 평균이 95점이고 수학 점수 평균이 30점이라면, 당연히 수학 점수가 80점인 학생이 더 잘한 것임

- 상대적으로 비교하기 위해 각 데이터에 있는 값을 상대적인 값을 갖도록 변환함

$$\frac{x-\mu}{\sigma}$$

Standard Scaling

$$\frac{x-\min(x)}{\max(x)-\min(x)}$$

Min-max Scaling

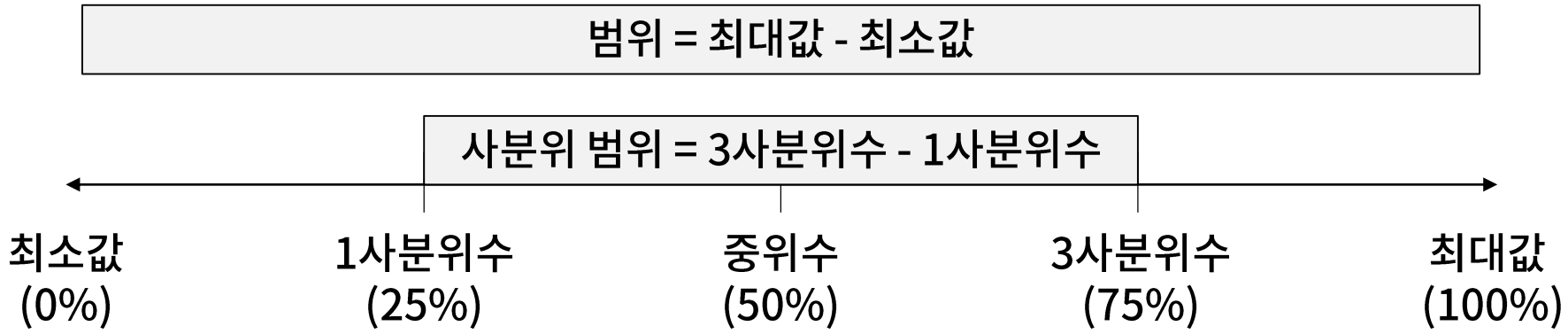
- 스케일링은 변수 간 비교 뿐만 아니라, 머신러닝에서도 널리 사용됨

I 파이썬을 이용한 스케일링

구분	구현 코드
Standard Scaling	• (x - x.mean()) / x.std() # x: ndarray • sklearn.preprocessing.StandardScaler
Min-max Scaling	• x - x.min() / (x.max() - x.min()) # x: ndarray • sklearn.preprocessing.MinMaxScaler

I 범위와 사분위 범위

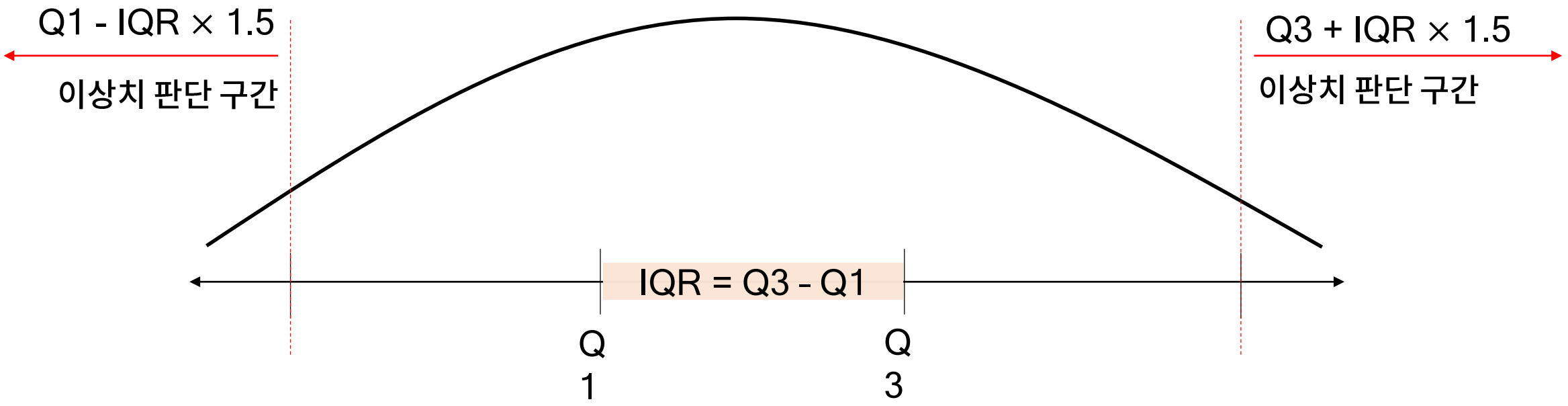
- 범위와 사분위 범위는 산포를 나타내는 가장 직관적인 지표 중 하나임



- 사분위 범위를 IQR (interquartile range)라고도 하며, 이상치를 탐색할 때도 사용됨 (참고: IQR Rule)

I (참고) IQR Rule

- 변수별로 IQR 규칙을 만족하지 않는 샘플들을 판단하여 삭제하는 방법



파이썬을 이용한 범위 및 사분위 범위 계산

구분	구현 코드
범위	• numpy.ptp(x) • numpy.max(x) – numpy.min(x)
사분위 범위	• numpy.quantile(x, 0.75) - numpy.quantile(x, 0.25) • scipy.stats.iqr(x)

Chapter. 09

변수가 어떻게 생겼나: 기초 통계 분석

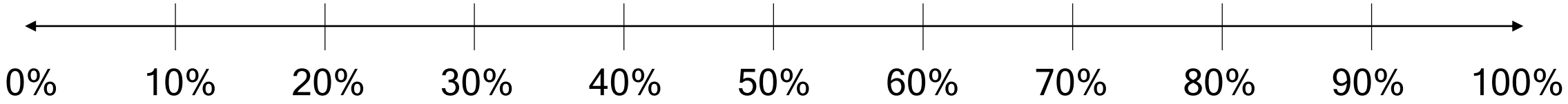
분포 통계량

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

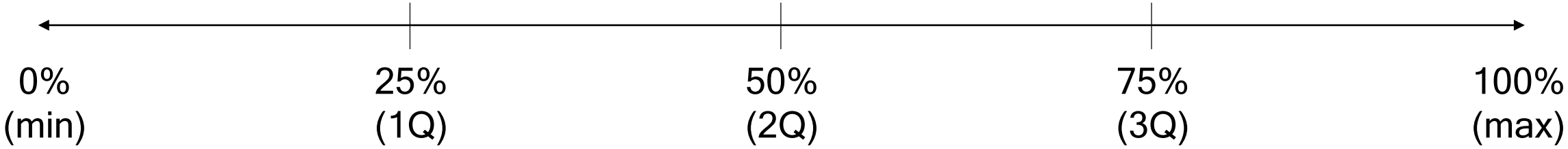
강사. 안길승

I 백분위수와 사분위수

- 백분위수 (percentile)는 데이터를 크기 순서대로 오름차순 정렬했을 때, 백분율로 나타낸 특정 위치의 값을 의미



- 사분위수 (quantile)는 데이터를 크기 순서대로 오름차순 정렬했을 때, 4등분한 위치의 값을 의미

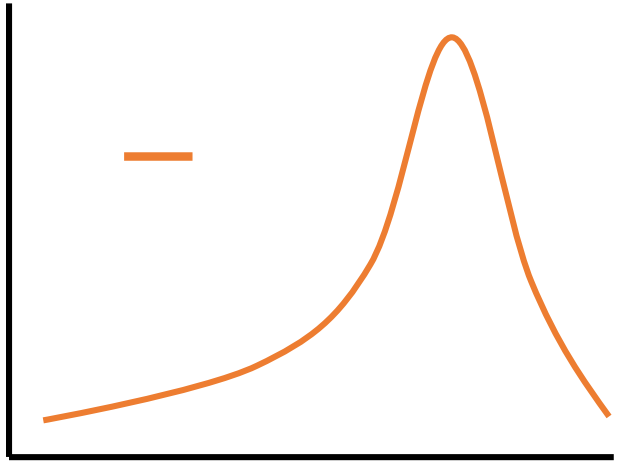


I 파이썬을 이용한 백분위수와 사분위수 계산

구분	구현 코드
백분위수	• numpy.percentile(x, q) # q: 위치 (0 ~ 100)
사분위수	• numpy.quantile(x, q) # q: 위치 (0 ~ 1)

I 왜도

- 왜도 (skewness): 분포의 비대칭도를 나타내는 통계량
- 왜도가 음수면 오른쪽으로 치우친 것을 의미하며, 양수면 왼쪽으로 치우침을 의미함



왜도 < 0

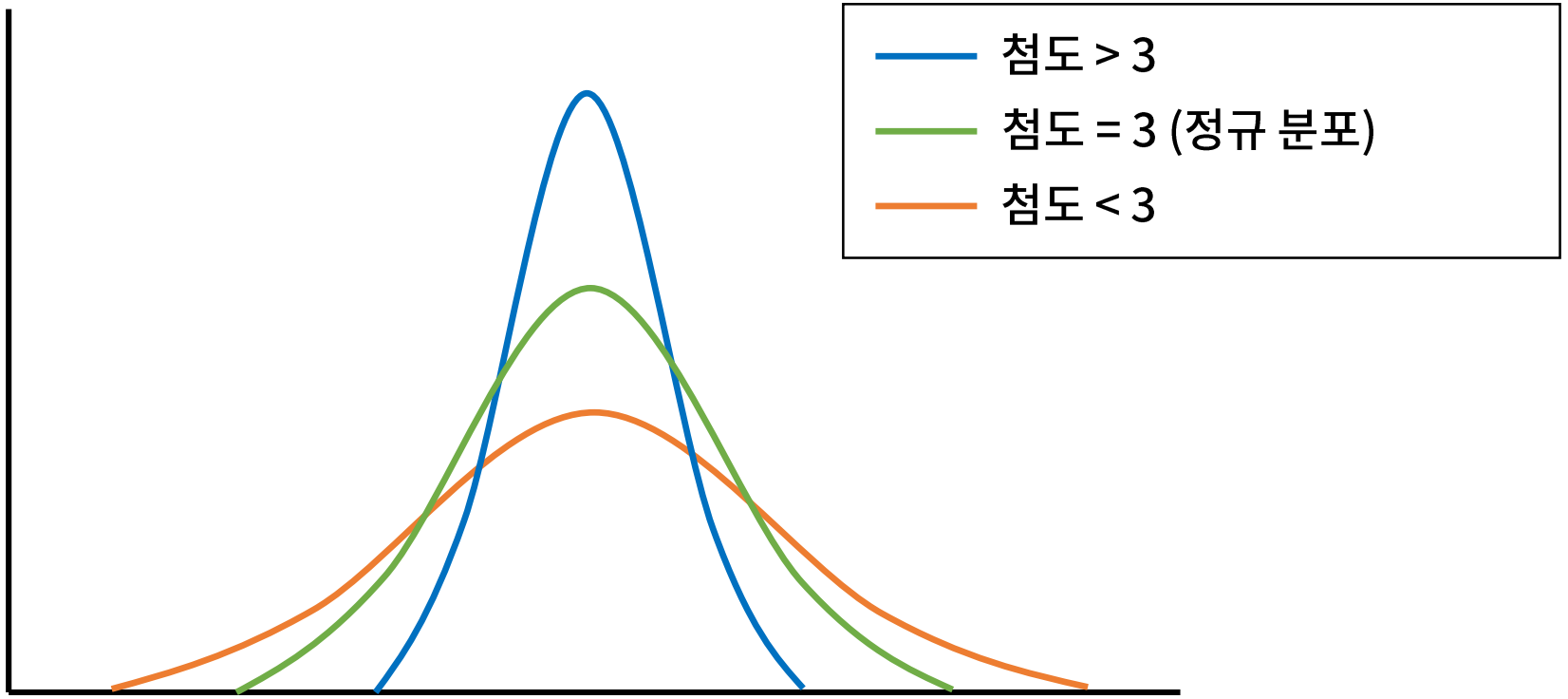
왜도 = 0

왜도 > 0

- 일반적으로 왜도의 절대값이 1.5 이상이면 치우쳤다고 봄

I 첨도

- 첨도 (kurtosis): 데이터의 분포가 얼마나 뾰족한지를 의미함. 즉, 첨도가 높을수록 이 변수가 좁은 범위에 많은 값들이 몰려있다고 할 수 있음



파이썬을 이용한 왜도와 첨도

구분	구현 코드
왜도	• scipy.stats.skew(x) • Series(x).skew()
첨도	• scipy.stats.kurtosis(x) • Series(x).kurtosis()

Chapter. 09

변수가 어떻게 생겼나: 기초 통계 분석

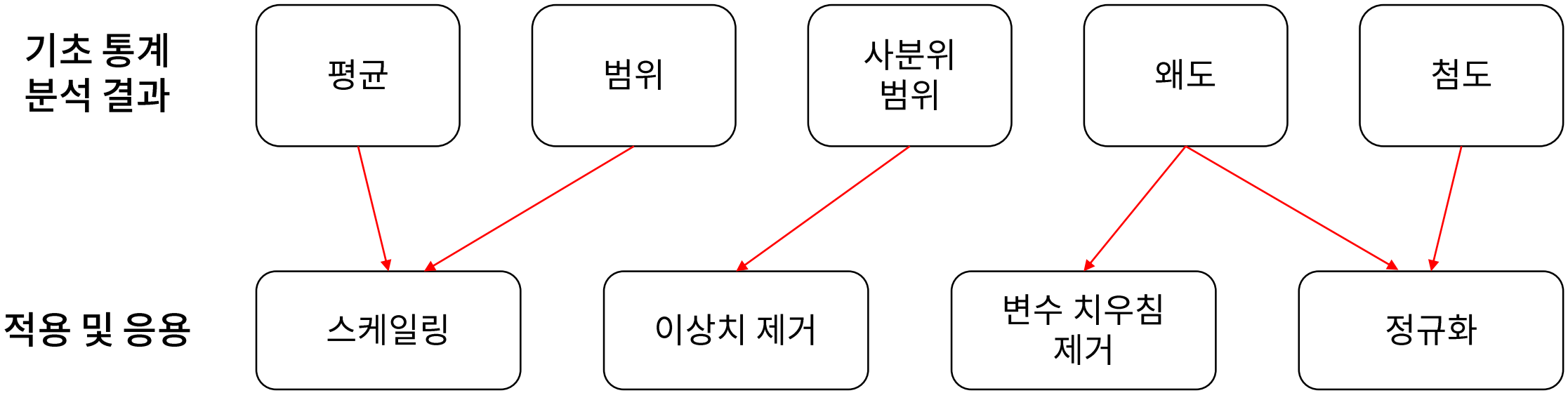
머신러닝에서의 기초 통계 분석

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

변수 분포 문제 확인

- 머신러닝에서 각 **변수를 이해**하고, 특별한 분포 **문제가 없는지 확인**하기 위해 기초 통계 분석을 수행함



Chapter.

변수가 어떻게 생겼나: 기초 통계 분석

감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승