탐색적 데이터 분석

고려대학교 석 준 희

목차

- 탐색적 데이터 분석 (Exploratory Data Analysis)
- 통계 분석 (Statistical Analysis)
- 관계 검정 (Relation Test)

탐색적 데이터 분석

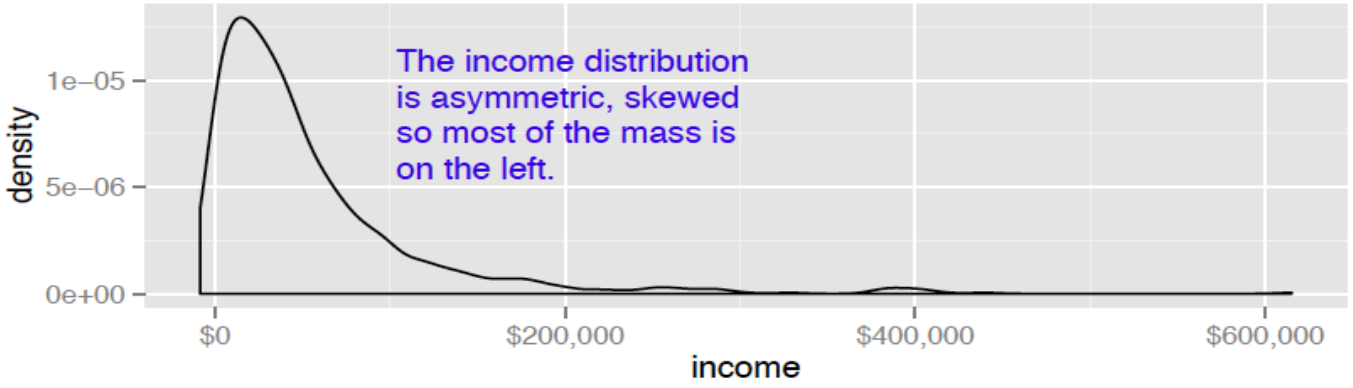
탐색적 데이터 분석 (EDA: Exploratory Data Analysis)

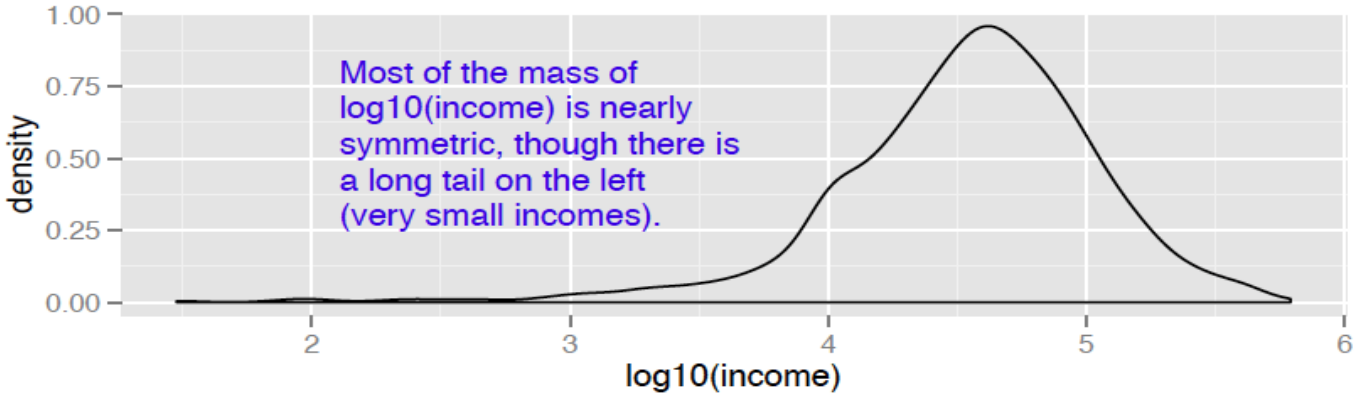
- 탐색적 데이터 분석
  - 실제 본격적인 분석 (CDA: Confirmatory Data Analysis)에 들어가기 전에 데이터를 살펴보는 과정
  - 주어진 데이터에 대한 감 혹은 일반적 이해를 목적으로 함
  - 주로 데이터 기반의 분석으로 다른 가정 없이 진행됨
  - 예를 들어 선형회귀는 선형성을 가정으로 진행되기 때문에 EDA에서는 사용되지 않음

- 탐색적 데이터 분석은 다음을 포함함
  - 데이터 요약 및 시각화
    - 차원 축소를 같이 사용하기도 함
  - 데이터 변형 (Feature engineering)
  - 이상치 탐색 (Outlier detection)
  - 결측치 처리 (Missing value handling)
  - 통계 분석 (Statistical analysis)

데이터 변형

- 수치형 데이터
  - 변환(Transformation): 정규분포에 가깝게 변형 (log, sqrt 등)
  - 표준화(Standardization): 특정 평균이나 분산을 같도록 변형
  - 정규화(Normalization): 값이 특정 범위내에 있도록 변형





데이터 변형

- 범주형 데이터
  - 작은 도수의 범주를 합쳐서 하나의 범주로 표현

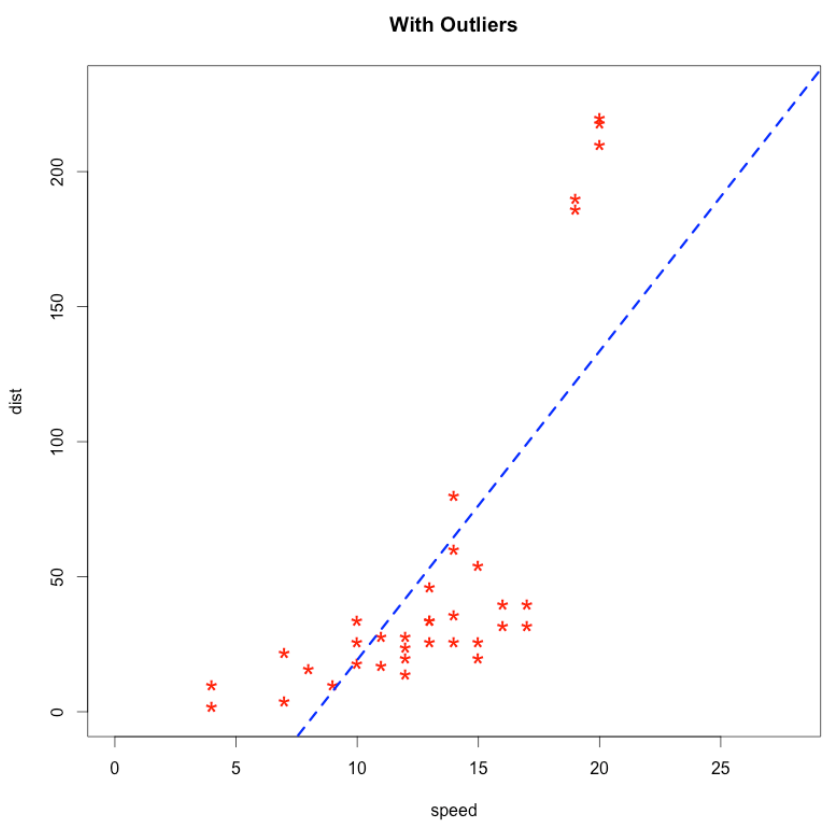
New variable

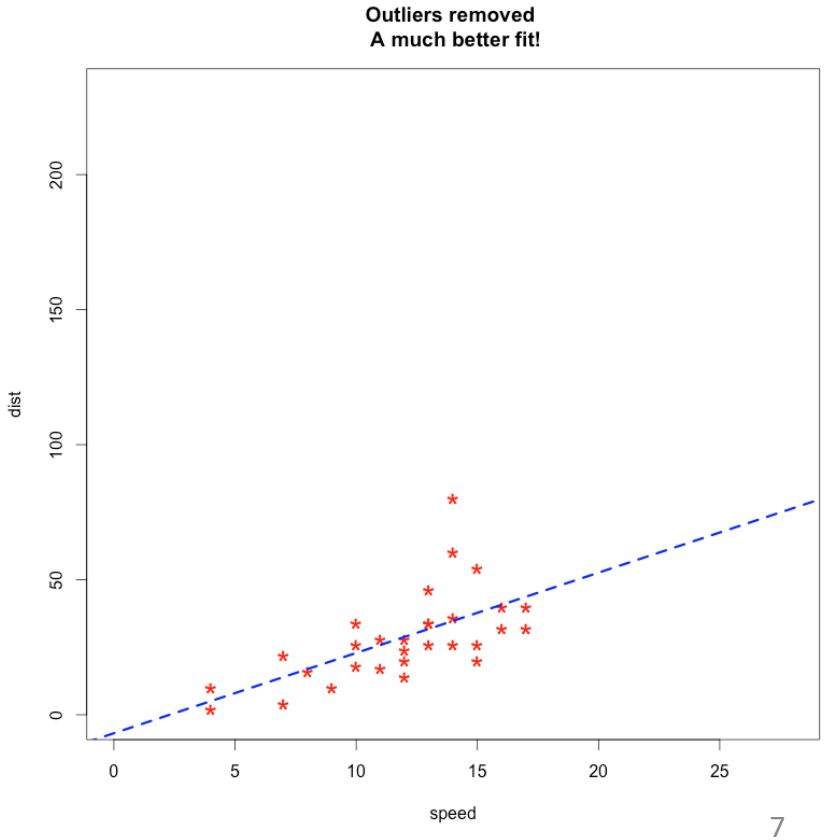
Original variable

		rec_nation Respondents' nationalities					
		1.00 Belgium	2.00 England	3.00 Spain	4.00 Sweden	5.00 Other	Total
nation Respondents' nationalities	1 Belgium	100	0	0	0	0	100
	2 England	0	201	0	0	0	201
	3 France	0	0	0	0	2	2
	13 Switzerland	0	0	0	0	1	1
	14 Ukraine	0	0	0	0	2	2
Total		100	201	120	160	20	601

이상치 검출 (Outlier Detection)

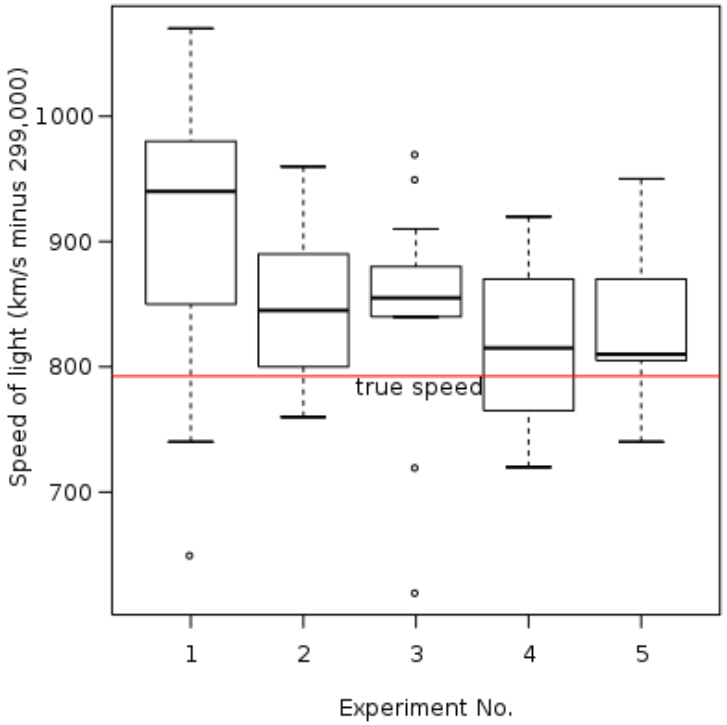
- 이상치 (Outlier): 이상하거나 극단적인 값
  - 오류, 실수 혹은 비정상적 사건에 의해 발생 가능
  - 올바른 분석을 방해할 수도 있지만, 새로운 발견일 가능성도 있음
  - 명확한 정의는 없음

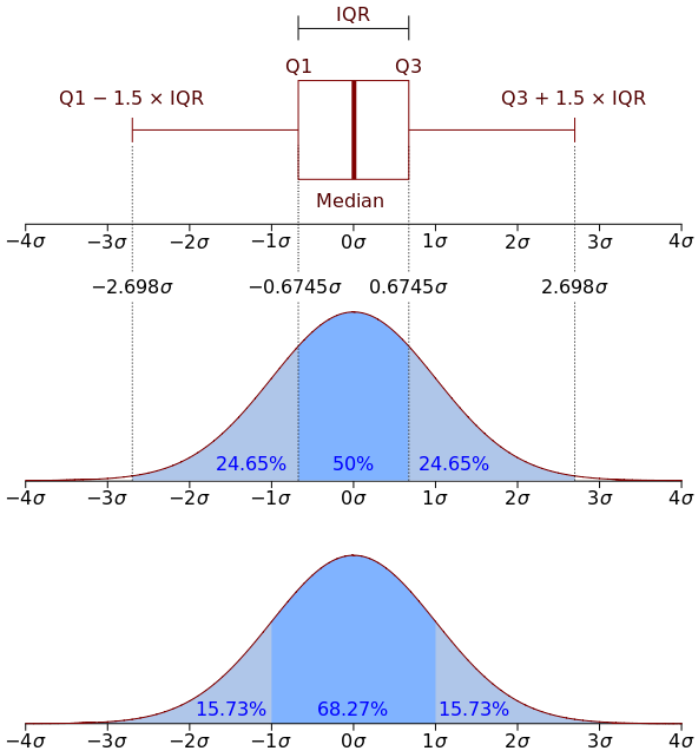




이상치 검출 (Outlier Detection)

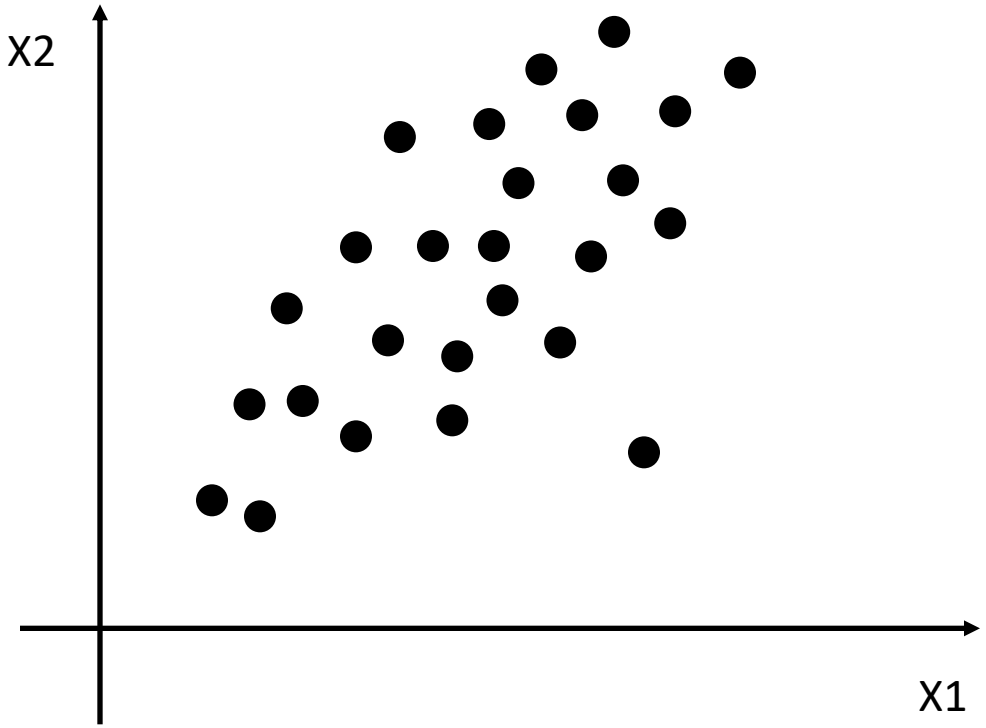
- 이상치 검출
  - 물리적으로 잘 못 된 값 (예: 음의 몸무게, 100도씨의 외부 온도 등)
  - 분포에서 나올 것 같지 않은 값
- 단변량 이상치 검출
  - 분포를 예상하여 해당 분포에서 확률적으로 나오기 어려운 값을 이상치로 검출
  - 박스플롯에서 이상치 검출, 정규화를 통한 검출 ($\mu \pm 3\sigma$)





이상치 검출 (Outlier Detection)

- 다변량 이상치 검출
  - 단별량 분석에서 이상치가 발견되지 않을 수 있음
  - 다변량 분포를 예상하여 유사하게 검출
- 고차원 데이터일수록 이상치를 검출하는 것이 어려움



결측치 처리 (Missing Value Handling)

- 이상치가 발견되었을 때….
  - 데이터 행렬에서 제거 or 결측으로 처리

- 결측치 (Missing Value)
  - (여러 이유에 의해) 단순히 측정되지 않음
    - 예: 값이 너무 낮아 센서가 탐지하지 못함, 실수로 누락됨,
  - 이상치로 탐지됨

- 결측치의 처리: 제거 or 채워넣음 (대치)

- 결측치 제거
  - 변수를 제거, 표본을 제거, 혹은 둘 다 제거
    - 어떻게 결정하나?
  - 결측치 제거의 문제점
    - 표본이나 변수의 부족으로 power가 떨어짐
    - 특정한 이유에 의해 결측이 발생한 경우 분석이 왜곡될 수 있음
    - 예: 부모의 학력, 성경험 여부

결측치 처리 (Missing Value Handling)

- 결측치 대치 (Missing Imputation)
  - 다른 정보에 기반하여 결측된 값이 어떤 값인지 예측하여 채워넣음
- 단순 대치법
  - 평균, 중간값, 0, 1, 정해진 적당한 값, 혹은 임의의 값으로 대치
  - 다른 표본의 값으로 대치 (표본 A의 X변수 값이 결측되었을 때, 표본 B의 X변수 값으로 대치)
- 모델 기반 대치법
  - 결측 변수를 다른 변수를 기반으로 모델링을 하고 예측하여 대치

- 예
  - 평균기반 대치: 173.5
  - 중간값 기반 대치: 169.5
  - 다른 표본 값으로 대치: 172
  - 모델 기반 대치: 173.7
    - 키의 제곱은 체중에 비례 $bmi = \frac{w}{h^2}$

$$\left(\frac{78}{1.75^2} + \frac{63}{1.67^2} + \frac{80}{1.80^2} + \frac{67}{1.72^2}\right) \times \frac{1}{4} = 23.85$$

$$\sqrt{72/23.85} = 173.7$$

	Height	Weight
S1	175	78
S2	167	63
S3	NA	72
S4	180	80
S5	172	67

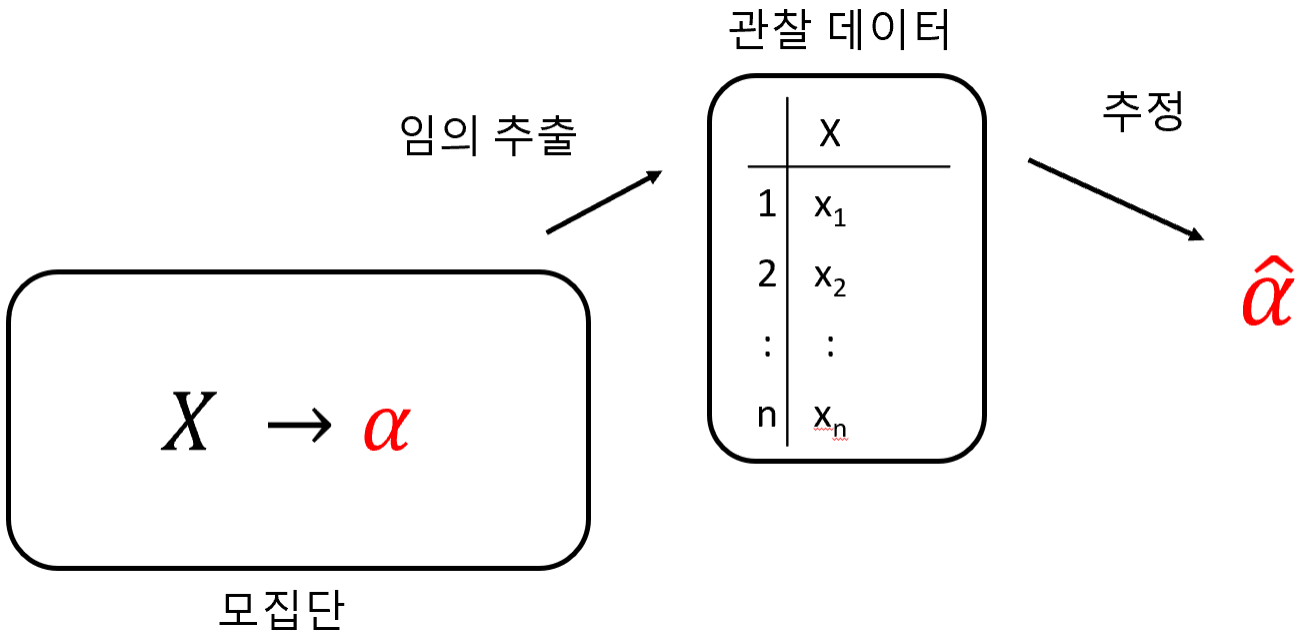
탐색적 데이터 분석

통계 분석

통계 분석 (Statistical Analysis)

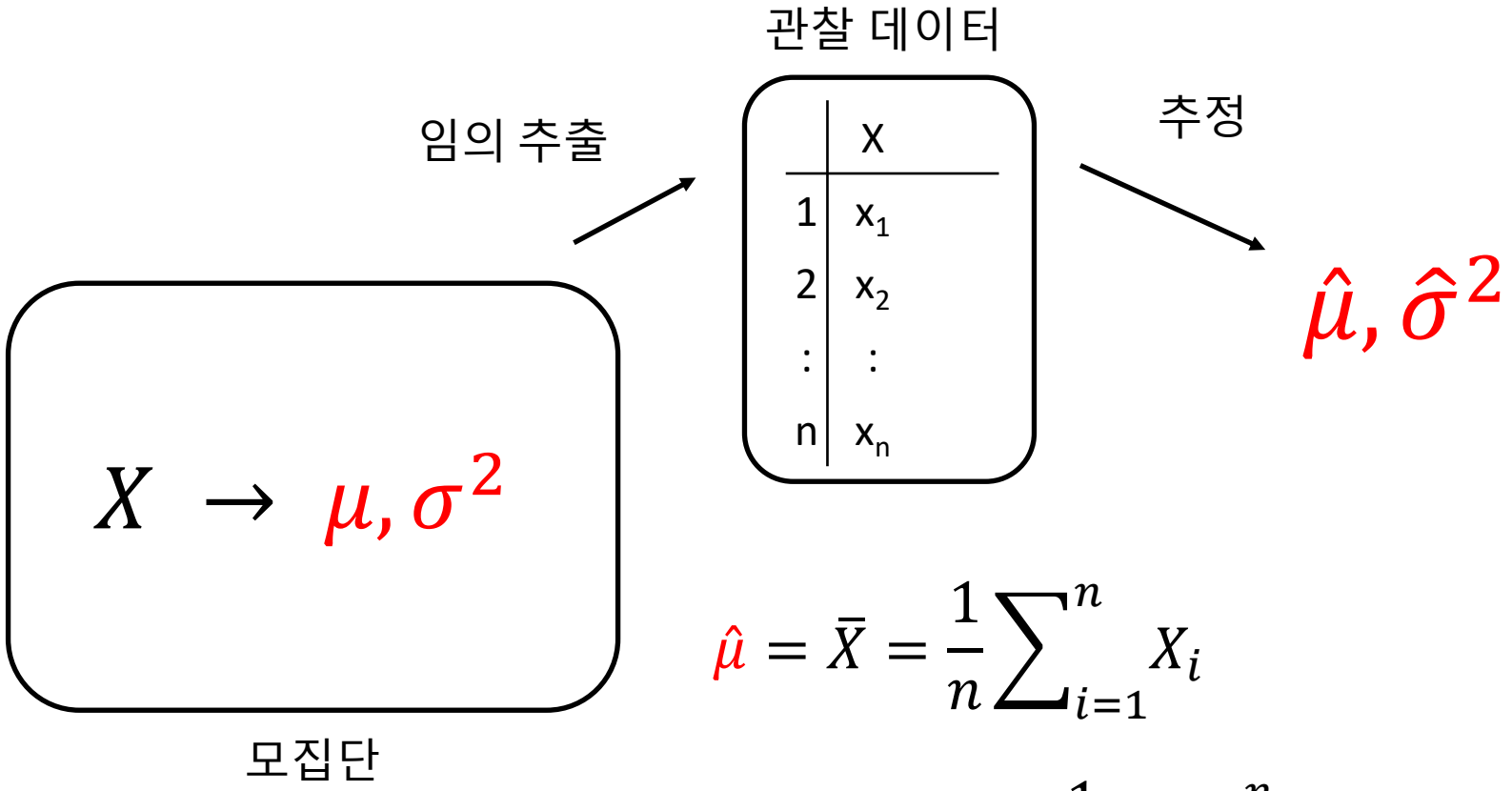
- 우리나라 성인 남성의 평균 키를 알고 싶다고 할 때…
  - 모든 성인 남성의 키를 측정하여 평균 (빅데이터적 접근법)
  - 무작위로 n명을 선택하여 평균을 취함 (통계적 접근법)
- 모집단 (Population): 관심의 대상이 되는 전체 집단
- 표본 (Sample): 모집단에서 임의로 추출되어 우리가 관측하는 데이터

- 통계 (Statistics): 표본으로부터 모집단에 대한 특성을 추정(Estimation)하기 위한 도구



표본 평균과 표본 분산

- 표본 평균(Sample Mean)과 표본 분산(Sample Variance) 모두 진짜 평균과 분산 (모집단 평균/분산)을 추정하기 위한 통계값



$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

신뢰구간 (Confidence Interval)

- 모집단의 특성은 변하지 않지만 그것에 대한 추정 값은 데이터에 따라 달라짐
- 현재 추정 값이 모집단의 특성과 얼마나 비슷한지 어떻게 알 수 있을까?
- 예: 주사위 눈의 평균을 5번 던져서 추정하기
  - 5번을 던져 1, 4, 1, 2, 5 얻음 → 표본평균=2.6
  - 그러면 추정 값은 실제 값이랑 얼마나 비슷할까?

	X_1	X_2	X_3	X_4	X_5	$\bar{X}$
1	1	4	1	2	5	2.6
2	2	6	2	3	3	3.4
3	1	5	5	2	6	3.8
4	6	2	1	5	5	3.8
5	1	1	6	5	5	3.6

- 신뢰구간 (Confidence Interval)
  - 모집단의 실제 값이 특정 확률 (신뢰도)로 포함되어 있을 구간
  - 예) 위의 주사위 예제
    - 95% 신뢰구간 (1.0, 4.2) = 2.6 ± 1.6
    - 우리가 추정한 값은 2.6이지만 실제 값이 (1.0, 4.2) 사이에 있을 확률은 95%이다.

가설 검정 (Hypothesis Test)

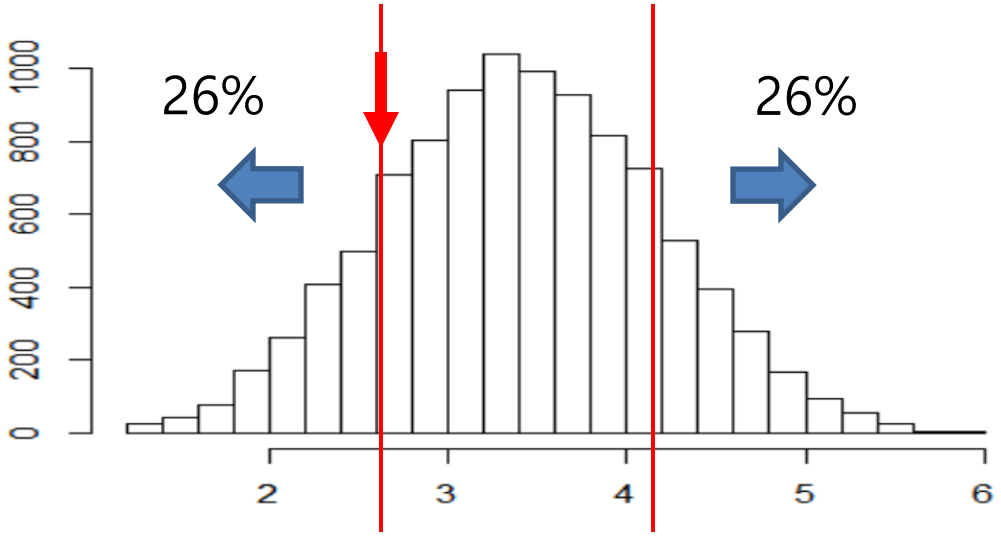
- 주사위 던지기 예제에서 모집단 평균은 3.5 표본 평균은 2.6
  - 지금 던지는 주사위가 정상적인 주사위인가? 진짜 평균이 3.5인가?
    - 가설 1 (귀무가설, null hypothesis): 정상 주사위이다 ($\mu = 3.5$)
    - 가설 2 (대립가설, alternative hypothesis): 비정상 주사위이다 ($\mu \neq 3.5$)

- 가설 검정 (hypothesis test)
  - 주어진 데이터로부터 가설이 성립하는지 확인
  - 보통 대립가설을 증명하기 위해 사용
  - 대립가설을 직접 확인하는 것이 아니라 귀무가설을 기각(reject)함으로써 이루어짐

- P 값: 귀무가설이 맞다고 가정했을 때 관측 값이 나타날 확률
  - 좀 더 정확히는 관측 값보다 더 극단적인 값
  - 큰 P: 귀무가설이 충분히 가능성있음 → 기각할 수 없음
  - 작은 P: 귀무가설이 맞을 가능성이 없음 → 기각
  - 얼마나 작은 것이 작은 것인가? 유의수준 (significance level)
    - 보통 5%나 1%를 사용

가설 검정 (Hypothesis Test)

- 주사위 예제: 5개의 표본으로부터 관측된 표본 평균 2.6
  - 귀무가설: 정상적인 주사위 (진짜 평균은 3.5이지만 우연에 의해 2.6을 관측)
  - 대립가설: 비정상적인 주사위 (진짜 평균이 3.5가 아니기 때문에 2.6을 관측)
  - 정상적인 주사위라면 (귀무가설이 맞다고 가정) 표본 평균의 분포는 아래와 같음
  - 이때 우연에 의해 2.6보다 더 평균에서 벗어날 값을 관측할 확률은 0.52 (p값)
  - 해석: 0.52는 충분히 발생할 수 있는 확률(유의수준 5%보다 큼)이기 때문에 귀무가설을 기각할 수 없음
    - 5% 유의수준에서 2.6은 3.5와 유의미하게 다르다고 할 수 없음



가설 검정 (Hypothesis Test)

- 또 다른 예제: 두 반의 시험 성적 차이
  - A반의 점수 평균 = 80, B반의 점수 평균 = 82
  - 관측된 평균 점수 차이 = 2, 실제 실력에 의한 점수 차이는?
  - 귀무가설: 두 반의 실력은 같지만 우연에 의해 차이가 발생 (실제 점수 차이 = 0)
  - 대립가설: 두 반의 실력이 달라 평균 점수 차이가 발생 (실제 점수 차이 ≠ 0)
  - 계산된 P값 = 0.03
  - 해석
    - 5% 유의수준에서는 귀무가설을 기각할 수 있음 or 두 반의 평균점수는 통계적으로 유의미한 차이가 있음
      - 두 반은 실제로 실력이 달라 평균 점수에서 차이가 남
    - 1% 유의수준에서는 귀무가설을 기각할 수 없음 or 두 반의 평균점수는 통계적으로 유의미한 차이가 없음
      - 두 반의 실력이 실제로 다르다는 근거가 없음

탐색적 데이터 분석

관계 검정

관계 검정 (Relation Test)

- 빅데이터에서의 변수
  - 관심 변수 Y와 함께 수집되는 설명 변수 X는 매우 많을 수 있음
  - 하지만, 대부분은 Y와 직접적인 관련이 없음
  - X의 개수를 줄여야지만 안정적인 데이터 분석이 가능

- 관계 검정 (Relation Test)
  - 본격적인 분석에 들어가기에 앞서, 관심 변수 Y와 관련이 있을 것으로 생각되는 중요 변수를 선택하기 위해 필요
  - 일반적으로 1:1의 관계에서 관련성이 있는지를 조사
  - 복합적인 관계를 조사하지 않음
  - 일반적으로 EDA 과정에서 1차적으로 수행됨

- 주어진 데이터를 바탕으로 X와 Y가 실제로 관련되어 있는지 확인
  - 주어진 데이터로부터 모집단에서의 관련성을 확인: 가설 검정
  - 귀무가설: 관련 없음, 대립가설: 관련 있음
  - 관련 검정의 결과 P값이 도출되고, P값을 해석하여 관련성을 평가
  - 관련성의 의미는? 변수에 따라 달라짐 → 변수 형태에 따라 서로 다른 관계 검정을 사용

변수의 형태에 따른 관계 검정

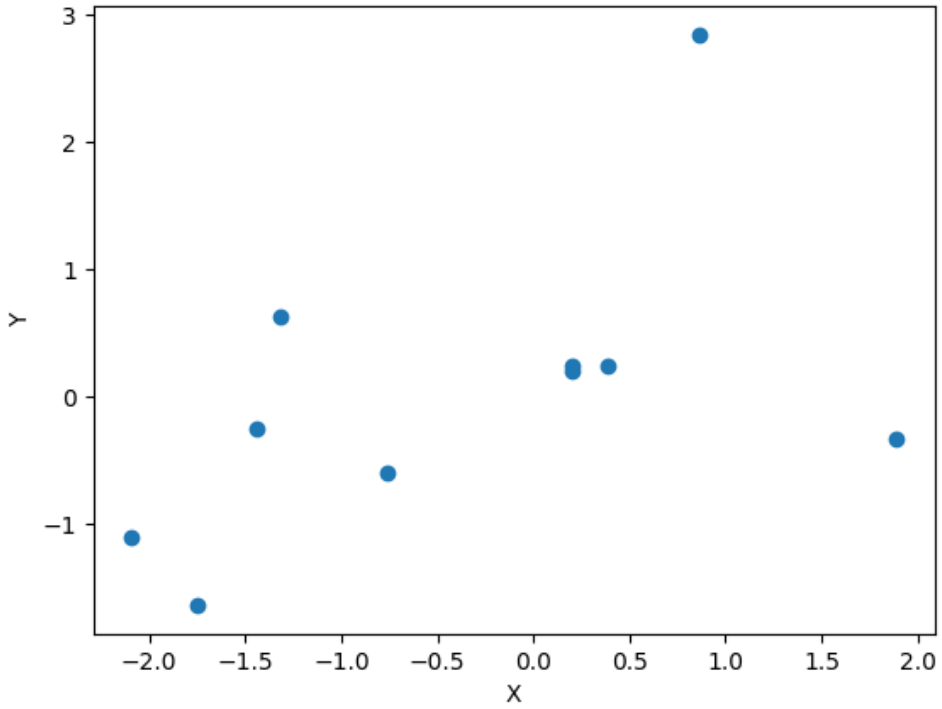
변수 형태	가설 검정	관련성	가설	특징	예
수치-수치	상관 검정 Correlation Test	두 변수의 상관계수가 0인지 아닌지로 판단	$H_0: \rho_{XY} = 0$ $H_A: \rho_{XY} \neq 0$	선형적 관계성만을 판단	키에 따른 소득의 차이
수치-범주	T검정 T-Test	두 범주 별 수치 변수의 평균이 같은지 다른지로 판단	$H_0: \bar{X}_{Y=1} = \bar{X}_{Y=2}$ $H_A: \bar{X}_{Y=1} \neq \bar{X}_{Y=2}$	분산이나 분포의 고려없이 평균만으로 판단	남녀 소득의 차이
범주-범주	카이제곱 검정 Chi-square Test	두 변수가 확률적으로 독립인지 아닌지로 판단	$H_0: X \perp Y$ $H_A: X \not\perp Y$	가장 근본적인 독립성으로 판단	지역별 종교 분포 차이

- F-검정 (F-Test): 수치-범주에 사용되는 검정 방식으로 범주가 셋 이상일 때 사용

상관 검정 (Correlation Test)

- 두 수치 변수의 모집단의 진짜 상관계수가 0이 아니라면 관련 있다고 판단
  - 작은 P값은 관측된 상관계수가 실제로 0이 아니라는 것을 나타냄

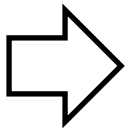
- 예: 상관 계수 = 0.52, P값 = 0.12
  - 실제 모집단의 상관계수는 0이지만 0.52보다 크게 관측될 확률이 0.12
  - 상관 계수(effect size)는 얼마나 두 변수가 연관되어 있는지를 나타내고, P값은 그 값이 얼마나 확실한지를 나타냄



T 검정 (T-Test)

- 두 개의 범주를 갖는 범주형 변수와 수치형 변수 사이의 관련성을 검사
- 두 범주별 수치형 변수의 평균이 다르다면 서로 관련있다고 판단
  - 작은 P값은 관측된 차이가 실제 차이가 0이 아니기 때문에 발생한다는 것을 의미
- 예: 키와 성별, 남성 키의 평균과 여성 키의 평균이 서로 같은지 판단
  - 남성 키 평균 = 171.6, 여성키 평균 = 157.3, 평균 차이 = 14.3, P값 = 0.036
  - 키와 성별은 관련 있는가? = 남성 키와 여성 키는 유의미하게 다른가?

Height	Sex
173	Men
165	Men
159	Women
181	Men
162	Women
151	Women
177	Men
162	Men



Group 1 Group 2

Men	Women
173	159
165	162
181	151
177	
162	

- F 검정 (F-Test): T검정과 유사하나 세 개이상의 범주를 갖는 범주형 변수에 사용
  - 여러 개의 범주 중 하나의 평균만 달라도 관련성이 있다고 판단

카이제곱 검정 (Chi-square Test)

- 두 범주형 변수가 확률적으로 독립이 아니라면 관련있다고 판단
  - X와 Y가 독립이라면 X를 이용해 Y를 예측하는 것이 불가능

- 예: 성별과 선호 음악 장르의 관련성
  - 관측 데이터 (크로스 테이블)

	락	발라드	힙합	계
남성	5	15	10	30
여성	5	10	5	20
계	10	25	15	50

  - 귀무가설(두 변수가 독립)이 맞다고 가정했을 때의 예상되는 데이터

	락	발라드	힙합	계
남성	6	15	9	30
여성	4	10	6	20
계	10	25	15	50

카이제곱 검정 (Chi-square Test)

- 예: 성별과 선호 음악 장르의 관련성
  - 독립이라면 남성-락은 6명이 관측되어야 하는데 실제로는 5명이 관측
  - 독립이지만 우연히 5명이 관측된 것인지, 독립이 아니기 때문에 5명이 관측된 것인지 판단
  - Q값의 계산

$$Q = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(5-6)^2}{6} + \frac{(10-9)^2}{9} + \frac{(5-4)^2}{4} + \frac{(5-6)^2}{6} = 0.69$$

  - 완전 독립이라면 Q는 0일 것으로 예상, 모집단에서 독립이지만 0이 아닌 것으로 나온 것인지 아니면 독립이 아니어서 0이 아닌지 판단
  - 독립이지만 Q가 0.69보다 크게 관측될 확률 = P값 = 0.71
  - 귀무가설을 기각할 수 없음 = 성별과 선호 음악 장르가 독립이 아니라고 판단할 근거가 부족 = 두 변수는 서로 관련있다고 판단하기 어려움

KOREA UNIVERSITY

감사합니다