데이터 분석을 위한 기초 통계

with Dataset

Contents

1. 기술통계

2. 추론통계

3. 통계 분석 프로세스

1. 기술 통계

수집한 데이터를 분석하여 대상들의 속성을 파악하는 통계 방법

• 중심경향값: 전체 자료를 대표할 수 있는 수치들

• 분산도: 전체 자료가 얼마나 퍼져 있는지를 알 수 있는 수치들

• 상관계수 : 두 변수 간의 관계의 크기

• 회귀계수 : 독립변수(X)가 종속변수(y)에 미치는 영향의 크기

- 중심경향값: 전체 자료를 대표할 수 있는 수치들
 - 평균 : 전체 자료가 가치는 수치들의 총합을 전체 자료 수로 나눈 수치
 - 중앙값: 최대값과 최소값의 한가운데 수치
 - 최빈값 : 가장 많은 빈도를 가진 수치

- 분산도: 전체 자료가 얼마나 퍼져 있는지를 알 수 있는 수치들
 - 분산 : 각 자료가 평균으로 부터 떨어진 정도를 제곱한 수치들의 평균
 - 표준편차 : 분산의 제곱근

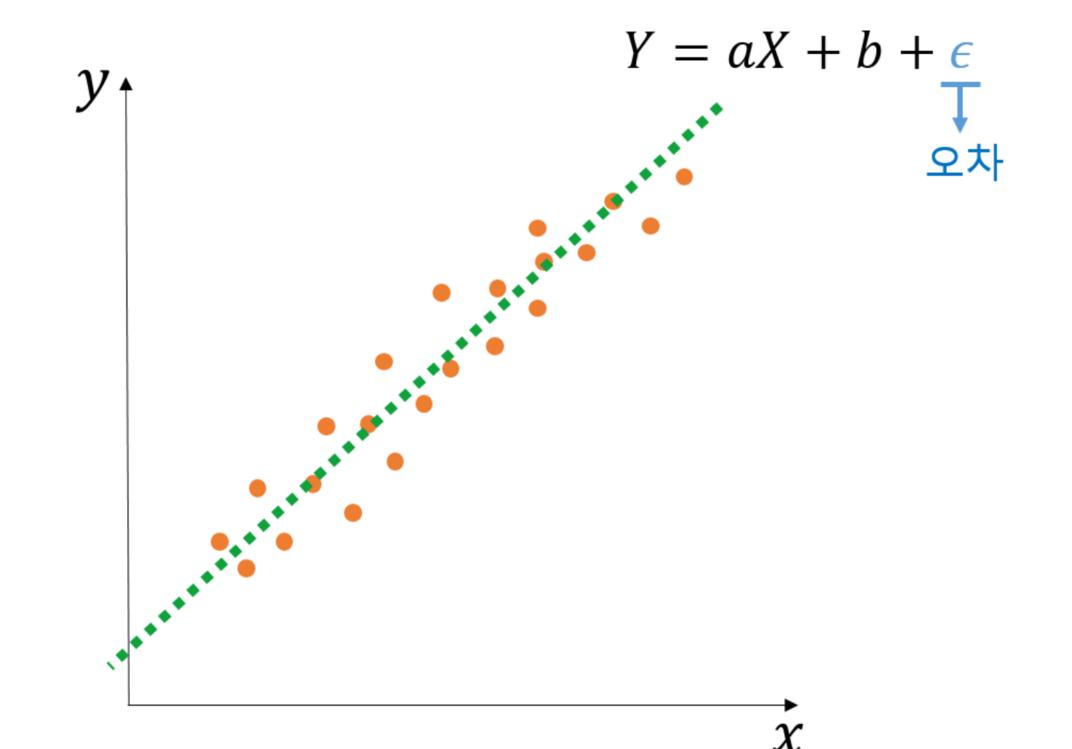
- 상관계수 : 두 변수 간의 관계의 크기
 - 공분산: 두 변수가 함께 각자의 평균으로부터 멀어지는 정도. 한 변수가 자신의 평균으로부터 멀어질 때, 다른 변수가 자신의 평균으로부터 멀어지는 정도를 의미.

$$cov(A, B) = \frac{\sum_{i=1}^{n} (A_i - \bar{A})(B_i - \bar{B})}{n}$$

- 상관계수 : 두 변수 간의 관계의 크기
 - 상관계수 : 두 변수 간의 관계로, 하나의 변수가 변화함에 따라 다른 변수가 변화하는 정도를 의미. [-1, 1]

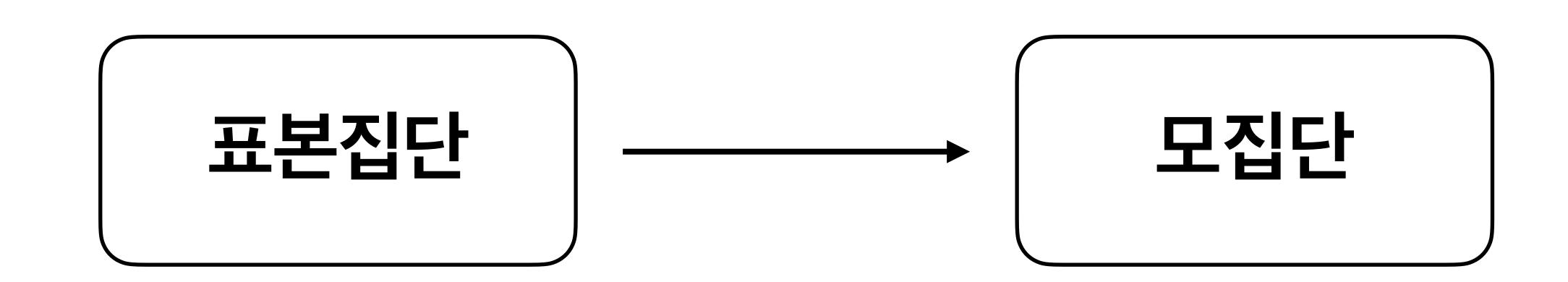
$$r_{AB} = \frac{Cov(A,B)}{s_a \times s_b}$$

- 회귀계수: 독립변수(X)가 종속변수(y)에 미치는 영향의 크기
 - 오차(Error)를 최소화하는 방향으로 직선이 생성됨

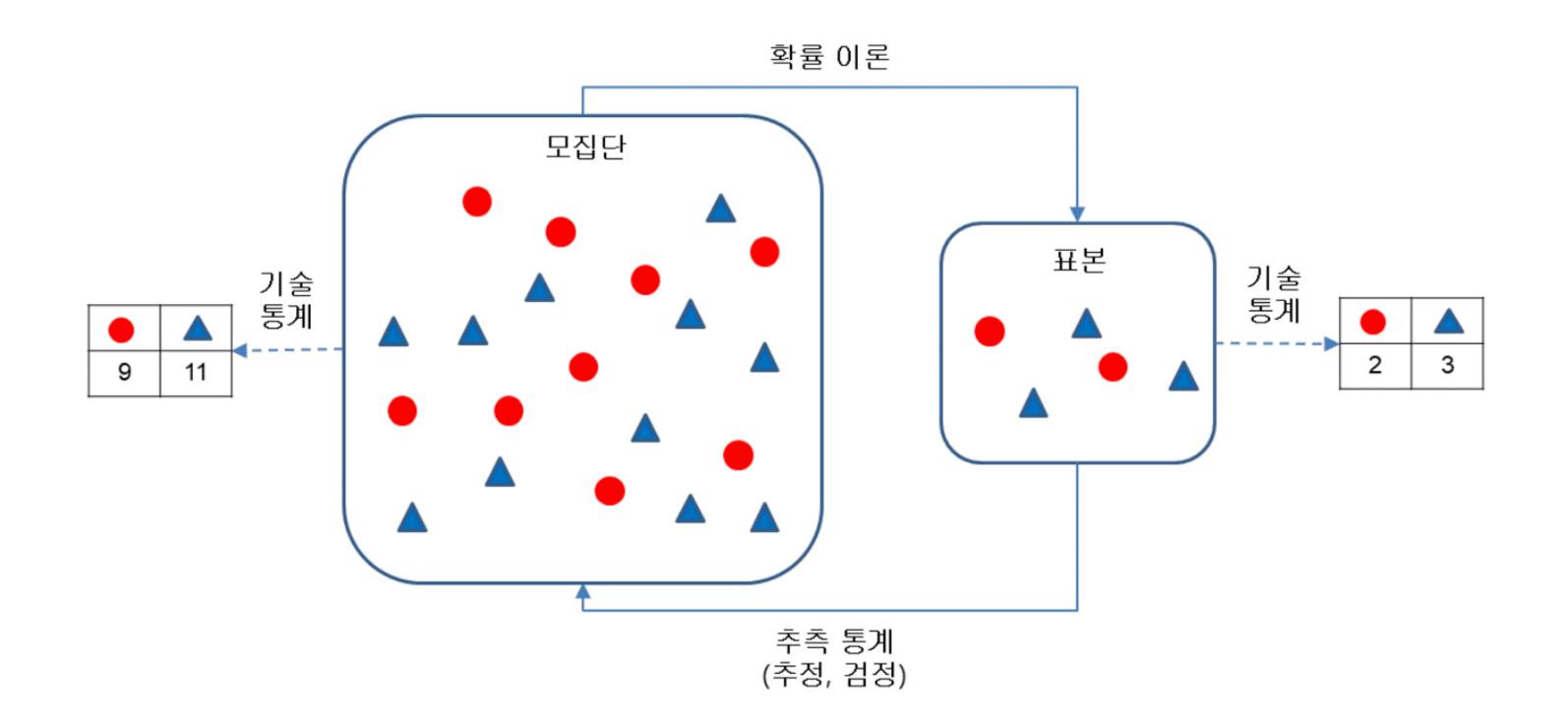


2. 주론 통계

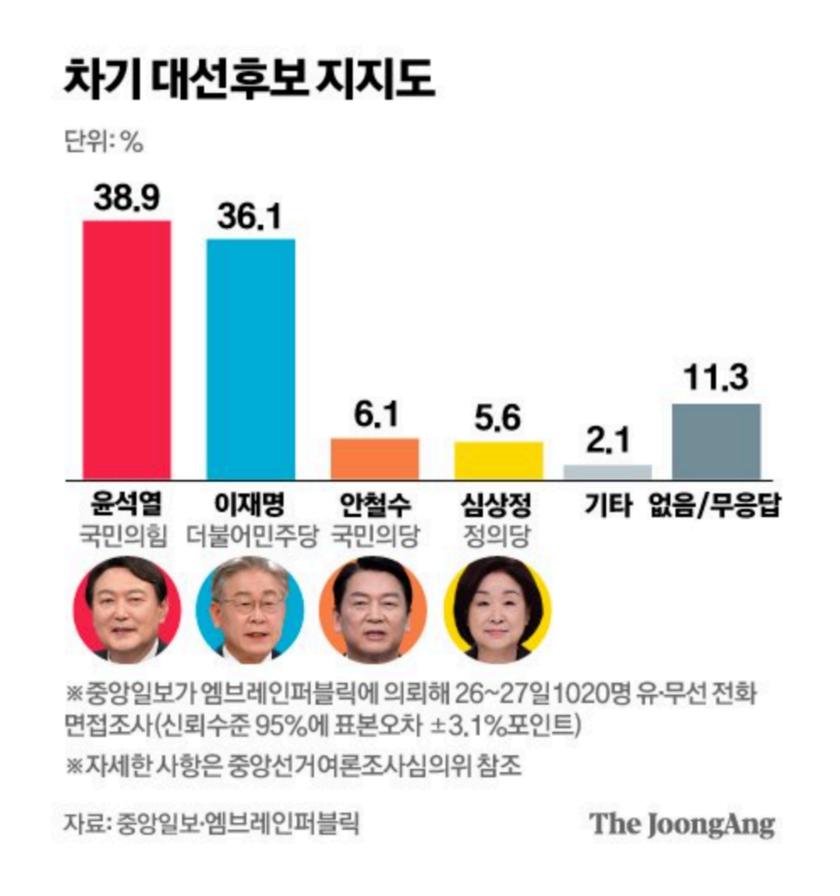
수집한 데이터를 확인하고 데이터의 표본이 되는 값을 찾아서, 기술통계를 이용하여 모집단의 속성들을 유추하는 통계 기법



수집한 데이터를 확인하고 데이터의 표본이 되는 값을 찾아서, 기술통계를 이용하여 모집단의 속성들을 유추하는 통계 기법



수집한 데이터를 확인하고 데이터의 표본이 되는 값을 찾아서, 기술통계를 이용하여 모집단의 속성들을 유추하는 통계 기법



Source: https://www.joongang.co.kr/article/25027541#home

차기 대선후보 지지도. 그래픽=김영옥 기자 yesok@joongang.co.kr



데일리안이 여론조사공정㈜에 의뢰해 최종 응답자 3004명 규모로 대선 여론조사를 실시한 결과, 당선 가능성 예측에서 국민의힘 윤석열 후보가 47.8%, 더불어민주당 이재명 후보가 44.8%로 나타났다. 이번 설문의 오차범위는 95% 신뢰수준에서 ±1.8%p였다. ⓒ데일리안 박진희 그래픽디자이너

Source: https://m.dailian.co.kr/news/view/1088304/

수집한 데이터를 확인하고 데이터의 표본이 되는 값을 찾아서, 기술통계를 이용하여 모집단의 속성들을 유추하는 통계 기법

• 배터리 공정과정에서 일부 표본을 뽑아서, 불량률 예측하기

• 임의로 선정한 20대를 대상으로 선호도 분석하기

• 설문지를 바탕으로 박물관 방문 동기 분석하기

•

3. 통계 분석 프로세스

수집한 데이터를 확인하고 데이터의 표본이 되는 값을 찾아서, 기술통계를 이용하여 모집단의 속성들을 유추하는 통계 기법

Statistical Analysis Workflow

| 과제 정의 | 데이터 수집 | 데이터 전처리 | | 추론 모델 개발 |
|-----------------|--------------|--------------|--------------|------------|
| Module 1 | Module 2 | Module 3 | Module 4 | Module 5 |
| 통계 분석 문제 정의 | 데이터 수집 및 정제 | 탐색적 데이터 분석 | 데이터 전처리 | 추론 모델 개발 |
| ● 이슈 파악 & 문제 도출 | ● 데이터 마트 생성 | ● 결측치 처리 | ● 차원의 저주 | ● 추론 모델 수립 |
| ● 분석 데이터 정의 | ● 데이터 정합성 평가 | ● 클래스 불균형 문제 | ● 상관관계 분석 | ● 추론 모델 검정 |
| | ● 데이터 취합 | ● 데이터 시각화 | ● 피처 추출 | ● 추론 모델 평가 |
| | ● 데이터 포맷 통일 | ● 통계량 분석 | ● 범주형 변수 인코딩 | |

Questions?