

데이터 분석을 위한 기초 통계

with Dataset

Contents

1. 기술통계
2. 추론통계
3. 통계 분석 프로세스

1. 기술 통계

기술통계 (Descriptive Statistics)

수집한 데이터를 분석하여 대상들의 속성을 파악하는 통계 방법

- 중심경향값 : 전체 자료를 대표할 수 있는 수치들 (평균, 최빈값, ...)
- 분산도 : 전체 자료가 얼마나 퍼져 있는지를 알 수 있는 수치들 (분산)
- 상관계수 : 두 변수 간의 관계의 크기 (Pearson)
- 회귀계수 : 독립변수(X)가 종속변수(y)에 미치는 영향의 크기

$$y = \underset{\text{coefficient}}{w}X$$

기술통계 (Descriptive Statistics)

수집한 데이터를 분석하여 대상들의 속성을 파악하는 통계 방법

- 중심경향값 : 전체 자료를 대표할 수 있는 수치들 $\checkmark x_i : i\text{번째 데이터}$

- 평균 : 전체 자료가 가치는 수치들의 총합을 전체 자료 수로 나눈 수치

$$\frac{\sum_i^N x_i}{N}$$

- 중앙값 : 최대값과 최소값의 한가운데 수치 모든 데이터를 오름차순 정렬.

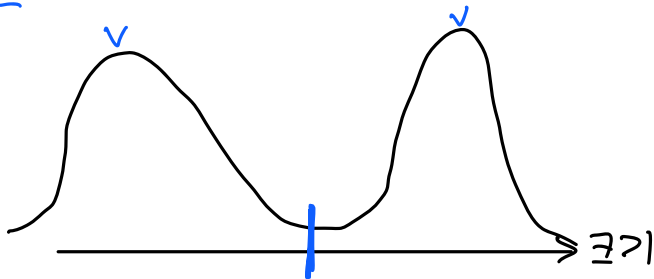
N 이 홀수 : 한가운데 값

- 최빈값 : 가장 많은 빈도를 가진 수치

N 이 짝수 : 가운데 2개의 평균

eg. $[1, 2, 4, 6, 10, 11]$

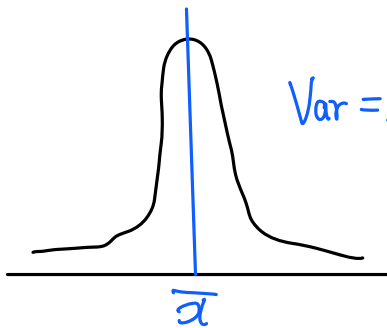
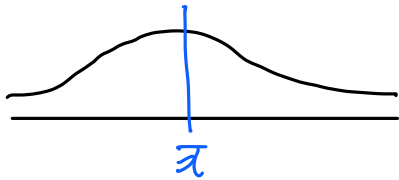
$$\frac{4+6}{2} = 5 : \text{중앙값}$$



기술통계 (Descriptive Statistics)

수집한 데이터를 분석하여 대상들의 속성을 파악하는 통계 방법

- 분산도 : 전체 자료가 얼마나 퍼져 있는지를 알 수 있는 수치들
 - 분산 : 각 자료가 평균으로 부터 떨어진 정도를 제공한 수치들의 평균 \bar{x}, μ
 - 표준편차 : 분산의 제곱근



$$\text{Var} = \frac{\sum_i^N (x_i - \bar{x})^2}{N}$$

$$\text{std} = \sqrt{\text{Var}}$$

기술통계 (Descriptive Statistics)

수집한 데이터를 분석하여 대상들의 속성을 파악하는 통계 방법

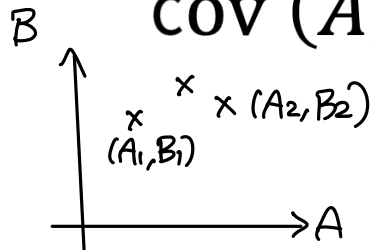
피어슨

- 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.

귀: x_1 , 몸무게: x_2

A의 평균 B의 평균

↓ ↓

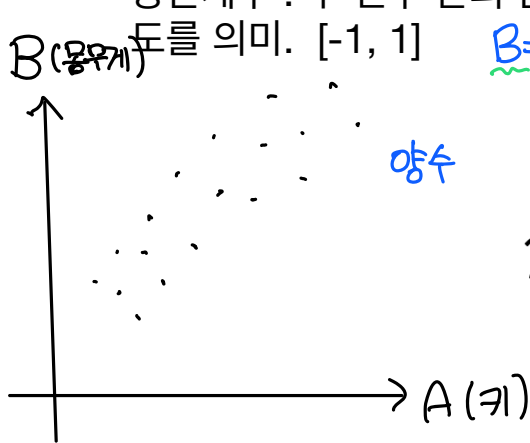
$$\text{cov}(A, B) = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{n}$$


기술통계 (Descriptive Statistics)

수집한 데이터를 분석하여 대상들의 속성을 파악하는 통계 방법

- 상관계수 : 두 변수 간의 관계의 크기

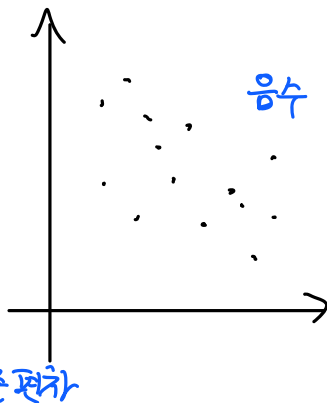
- 상관계수 : 두 변수 간의 관계로, 하나의 변수가 변화함에 따라 다른 변수가 변화하는 정도를 의미. $[-1, 1]$



$B = wA : 1$, $B = -wA : -1$
A와 B의 공분산

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

A의 표준편차 B의 표준편차



기술통계 (Descriptive Statistics)

수집한 데이터를 분석하여 대상들의 속성을 파악하는 통계 방법

- 회귀계수 : 독립변수(X)가 종속변수(y)에 미치는 영향의 크기

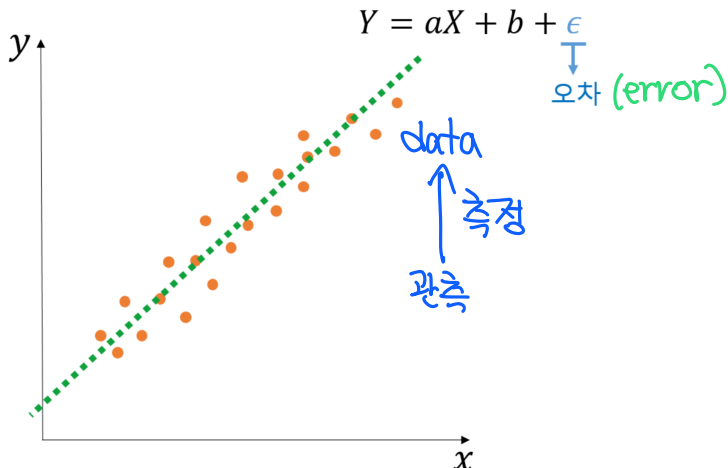
내일 비트코인 가격

4노닥 거주

$$y = aX + b + \epsilon$$

- 오차(Error)를 최소화하는 방향으로 직선이 생성됨

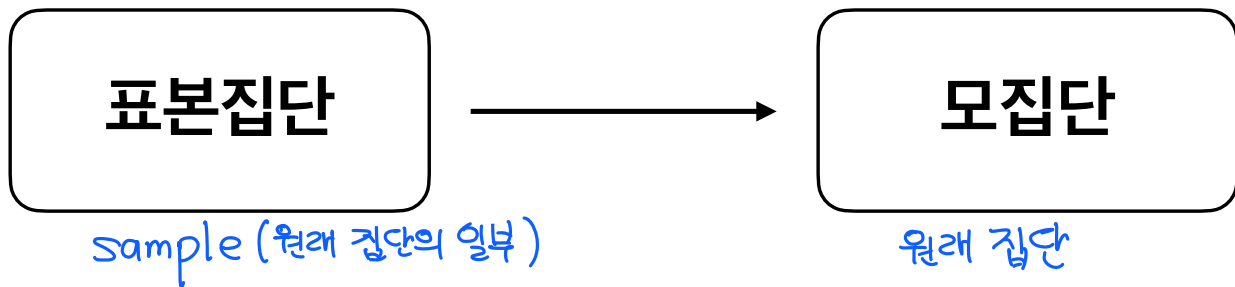
덜 틀리자!



2. 추론 통계

추론통계 (Inferential Statistics)

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

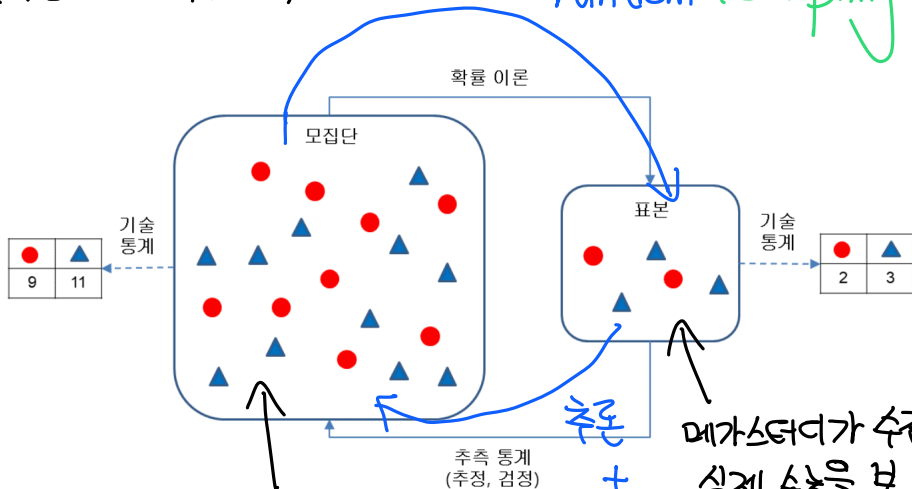


추론통계 (Inferential Statistics)

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

eg. 수능의 평균점 예측

random (sampling method)



수능을 실제로 응시한 인원의
수리 점수

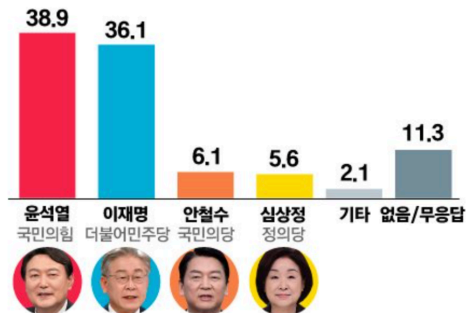
메가스터디가 수집한
실제 수능을 본 인원 수리 점수
오차 (error)
&
신뢰도

추론통계 (Inferential Statistics)

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

차기 대선후보 지지도

단위: %



※ 중앙일보가 엠브레인퍼블릭에 의뢰해 26~27일 1020명 유·무선 전화 면접조사(신뢰수준 95%에 표본오차 $\pm 3.1\%$ 포인트)

※ 자세한 사항은 중앙선거여론조사심의위 참조

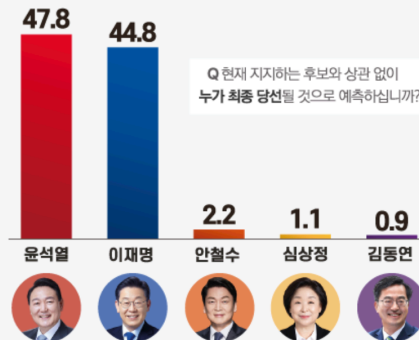
자료: 중앙일보-엠브레인퍼블릭

The JoongAng

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

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

당선 예측 (단위 %)



Q 현재 지지하는 후보와 상관 없이 누가 최종 당선될 것으로 예측하십니까?

※ 조사기간: 2022년 2월 25일(금)~27일(일) ※ 조사대상: 전국 만18세 이상 남녀
※ 조사방법: 휴대전화RDD 100% 자동응답전화조사 ※ 응답자수: 3,004명 ※ 응답률: 9.0%
※ 표본오차: 95%신뢰수준 $\pm 1.8\%$ ※ 의뢰자: 데일리안 ※ 조사기관: 여론조사공정위

데일리안이 여론조사공정위에 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 설문은 오차범위는 95% 신뢰수준에서 $\pm 1.8\%$ p였다. ©데일리안 박진희 그래픽디자인

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

추론통계 (Inferential Statistics)

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

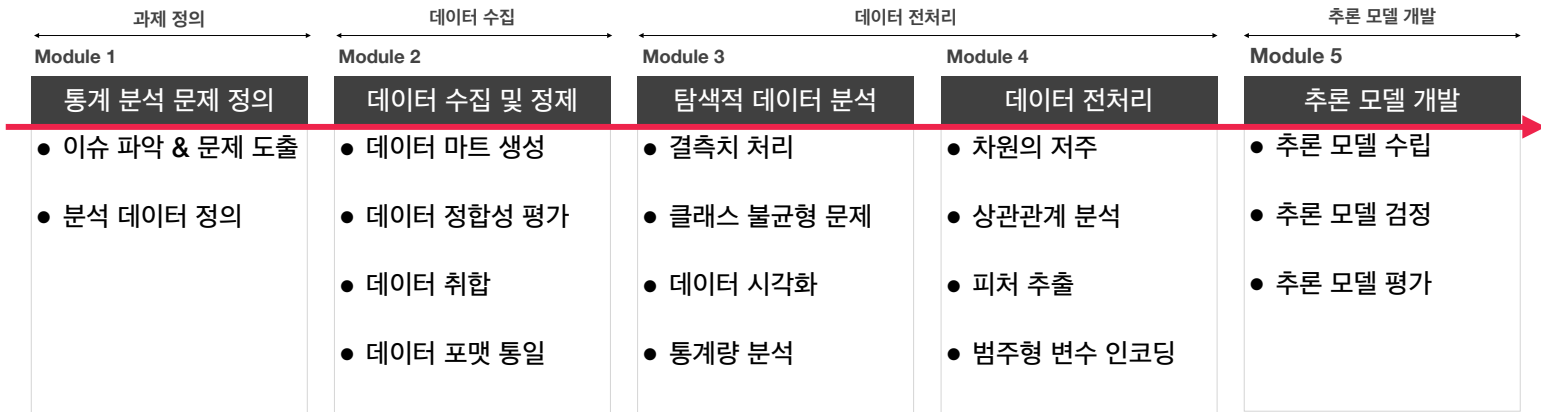
- 배터리 공정과정에서 일부 표본을 뽑아서, 불량률 예측하기
- 임의로 선정한 20대를 대상으로 선호도 분석하기
- 설문지를 바탕으로 박물관 방문 동기 분석하기
-

3. 통계 분석 프로세스

추론통계 (Inferential Statistics)

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

Statistical Analysis Workflow



Questions?