기초사회과학통계

고려대 대학원 행정학과

2022 여름

최정호 University of Pennsylvania

chjho@upenn.edu

서론

- 통계학의 필요성
- 어떤 자료인가
- 데이터와 변수

서론

- 통계학의 필요성

서론

- 통계학의 필요성
- 어떤 자료인가

- cross-sectional data
- time-series data
- panel/longitudinal data

서론

- 통계학의 필요성
- 어떤 자료인가
- 데이터와 변수

범주형 변수 categorical, qualitative

- 명목척도 nominal
- 서열척도 ordinal

연속형 continuous, quantitative

- 등간척도 interval
- 비율척도 ratio

행정안전부 고객만족도조사 설문지

고객님께서는 앞에서 평가해 주신 ○○○과의 업무처리 내용, 서비스 전달과정, 결과(이미지) 측면들을 모두 고려할 때, ○○과의 업무처리에 대해 **전반적으로** 얼마나 만족하셨습니까?

(1) 매우 불만족 (2) 불만족 (3) 보통 (4) 만족 (5) 매우 불만족

데이터의 요약

- 데이터의 분포
- 데이터의 요약치

데이터의 요약

- 데이터의 분포

- 도수분포표
  1) 구간의 수 결정
  2) 구간의 크기 결정
  3) 경계값 설정
  4) 관측 데이터의 빈도수 계산

- 히스토그램 histogram

데이터의 요약

- 데이터의 분포
- 데이터의 요약치

- 중심경향도 central tendency
- 산포도 dispersion
- 비대칭도 skewness

데이터의 요약

- 데이터의 분포
- 데이터의 요약치

- 중심경향도 central tendency
  - 평균 mean

$$\frac{X_1 + X_2 + \cdots + X_n}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

  - 중앙치 median
  - 최빈치 mode

데이터의 요약

- 데이터의 분포
- 데이터의 요약치

- 산포도 dispersion
  - 분산 variance

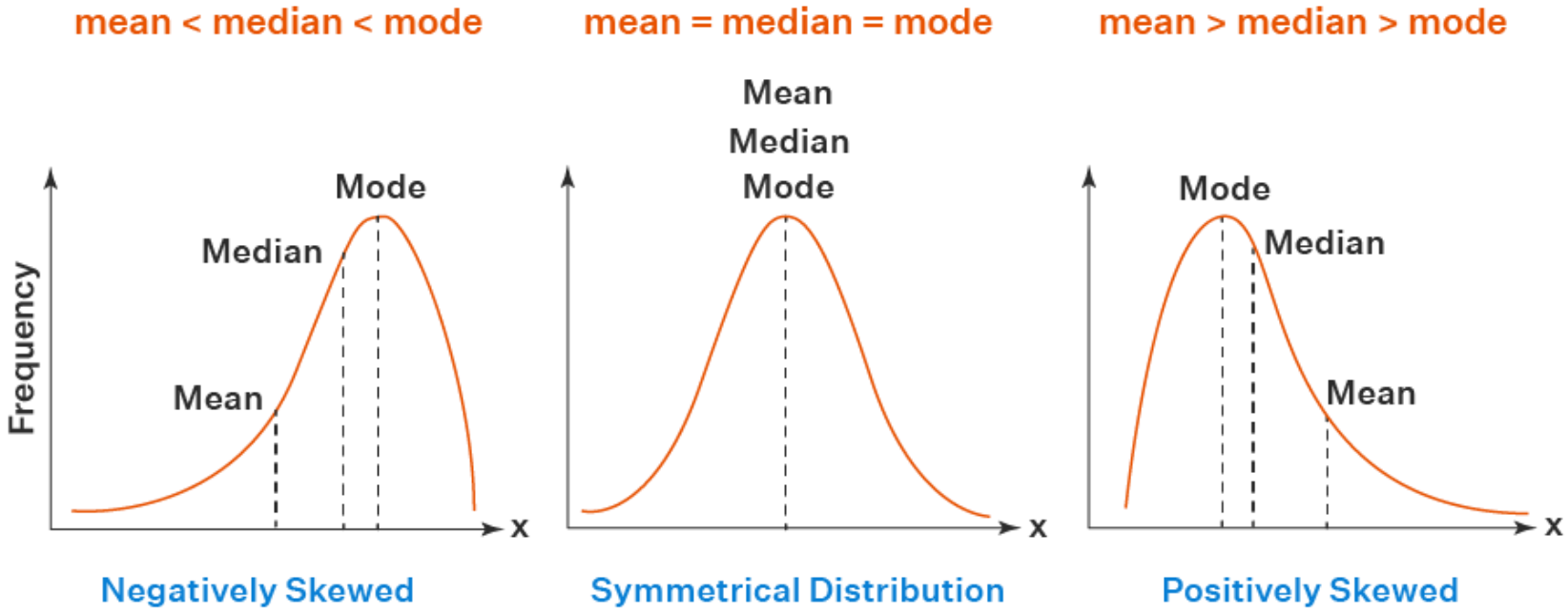
$$\text{분산}(\sigma^2) = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \qquad \text{표본분산}(S^2) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{(n-1)}$$

  - 표준편차 standard deviation

데이터의 요약

- 데이터의 분포
- 데이터의 요약치

- 비대칭도 skewness



확률

- 확률의 주요개념
- 확률변수의 의미
- 기대값과 분산
- 공분산과 상관계수

확률

- 확률의 주요개념

- 확률 probability
  - 조건 1)
  - 조건 2)
- 실험 experiment
- 사건 event
- 표본공간 sample space

확률

- 확률의 주요개념
- 확률변수의 의미

- 상수와 변수
- 확률변수 random variable
  - 역할: 확률 분포의 수치화
  - 이산확률변수와 연속확률변수

확률

- 확률의 주요개념
- 확률변수의 의미
- 기대값과 분산

- 기대값 expected value

$$E(aX+b)=aE(X)+b$$

$$E(X+Y)=E(X)+E(Y)$$

- 분산 variance

$$Var(aX+b)=a^2Var(X)$$

$$Var(X+Y)=Var(X)+Var(Y)+2cov(X,Y)$$

$$Var(X+Y)=Var(X)+Var(Y) \quad \text{if } X \text{ and } Y \text{ are independent.}$$

$$Var(X-Y)=Var(X)+Var(Y)-2cov(X,Y)$$

$$Var(X-Y)=Var(X)+Var(Y) \quad \text{if } X \text{ and } Y \text{ are independent.}$$

확률

- 확률의 주요개념
- 확률변수의 의미
- 기대값과 분산
- 공분산과 상관계수

- 공분산 covariance

$$Cov(X,Y) = \sigma_{XY} = E(x-\mu_X)(y-\mu_Y)$$

여기서, > 0 : 양의 선형관계

< 0 : 음의 선형관계

$= 0$: 선형관계가 없음 x_i y_i μ_x μ_y

- 상관계수 correlation coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho_{XY} \leq +1$$

확률분포

- 확률분포의 결정
- 균등분포
- 정규분포
- 표준정규분포
- 이항분포

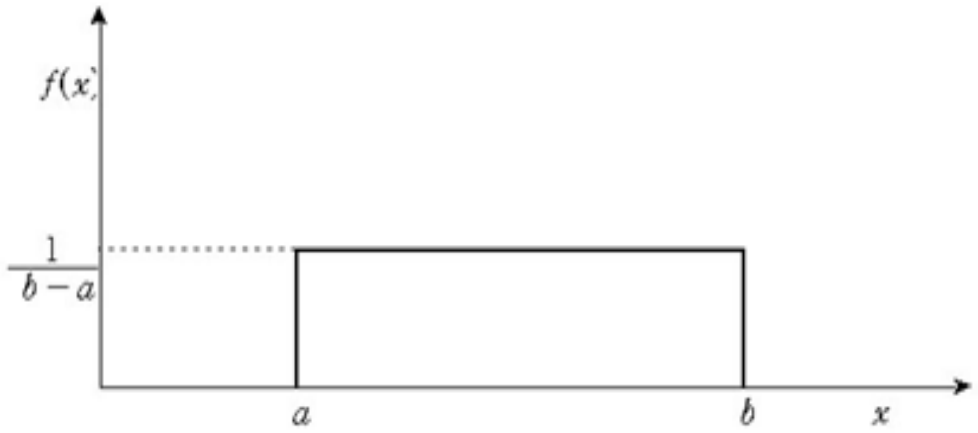
확률분포

- 확률분포의 결정

- 경험의 양 + 경험의 내용 (모수 parameter)

확률분포

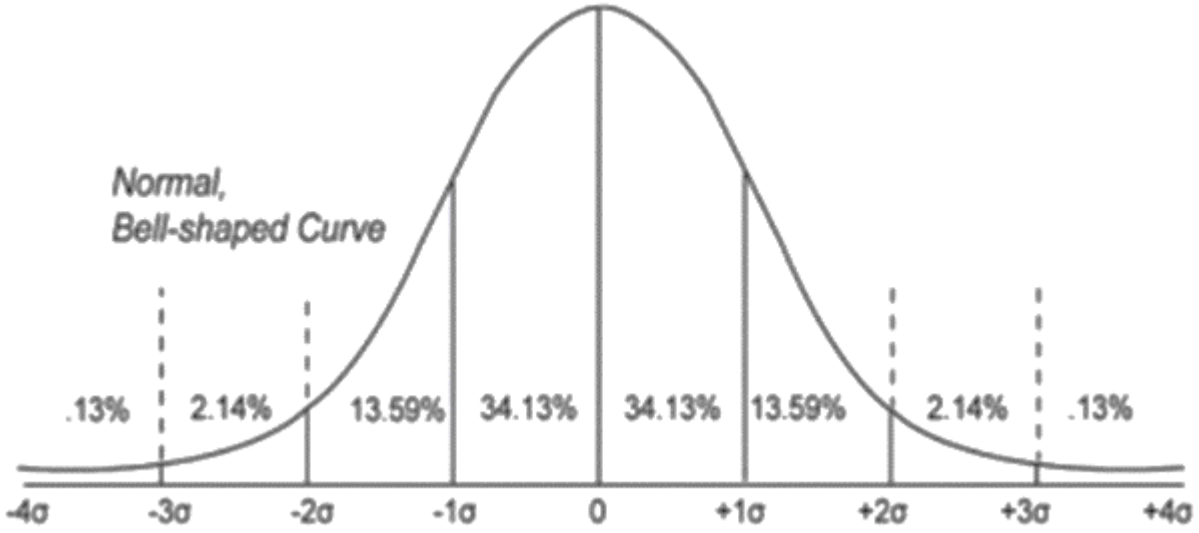
- 확률분포의 결정
- 균등분포



확률분포

- 확률분포의 결정
- 균등분포
- 정규분포

- Bell shape(경험적 법칙).
- 그래프 아래의 전체 면적의 합은 1
- 곡선의 최고 높이(즉 가장 높은 확률)는 평균
- 평균을 중심으로 좌우 대칭인 분포



확률분포

- 확률분포의 결정
- 균등분포
- 정규분포

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$X \sim N(\mu, \sigma^2)$$

확률분포

- 확률분포의 결정
- 균등분포
- 정규분포
- 표준정규분포

- 일반정규분포에서 표준정규분포로

$$Z \sim N(0, 1^2) \ \ where\ Z = \frac{X-\mu}{\sigma}$$

확률분포

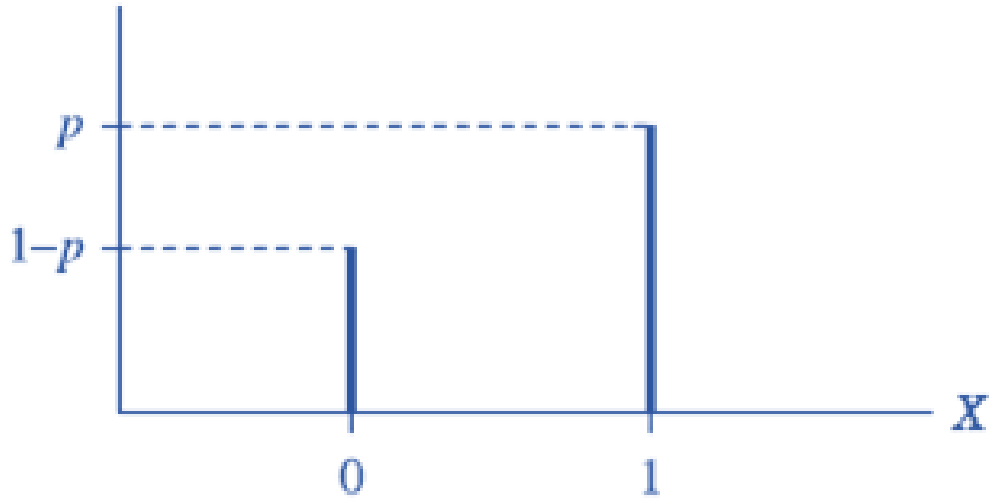
- 확률분포의 결정
- 균등분포
- 정규분포
- 표준정규분포
- 이항분포

- 베르누이 시행

조건 1: 두 개의 상호 배타적 원소로 구성된 실험의 시행

조건 2: 성공확률 p (실패 확률 1–p)는 시행횟수에 관계없이 일정

- 베르누이 분포



확률분포

- 확률분포의 결정
- 균등분포
- 정규분포
- 표준정규분포
- 이항분포

- 이항분포

$$f_X(x) = {}_nC\ p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \cdots, n$$

- 이항분포를 위한 실험은 n번의 베르누이 시행으로 구성
- 성공확률 p: 시행회수에 관계없이 항상 일정 (복원 추출)
- 각 시행이 통계적으로 독립적

확률분포

- 확률분포의 결정
- 균등분포
- 정규분포
- 표준정규분포
- 이항분포