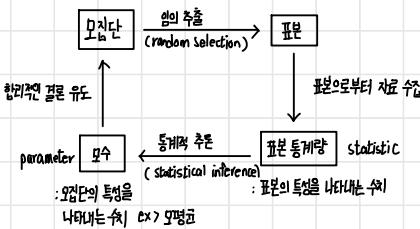


〈1-2장. 서론 및 표와 그림을 통한 자료의 요약〉

■ 용어

- **자료(Data)** : 연구나 조사의 바탕이 되는 사실
- **개체(Element)** : 자료를 가지고 있는 서로 구별되는 개인
- **변수(Variable)** : 관측되는 특성을 나타낸 것 → 보통 열
- **관측값(Observation)** : 한 개체로부터 얻을 수 있는 관측 결과를 모아놓은 집합 → 보통 행
- **모집단(Population)** : 알고자 하는 전체 대상. 모든 관측 가능한 주제들로 모든 집합
- **표본(sample)** : 모집단의 일부분에 해당하는 관측가들의 집합
- **전수조사(census)** : 모집단의 모든 대상에 대해서 자료 수집 과정 (\leftrightarrow 표본조사(Sample survey))



■ 자료의 요약

- 자료의 특성을 한 눈에 파악하기 위해 자료를 요약
- 자료를 요약하는 방법은 자료의 형태에 따라 달라진다.
- 자료의 형태별 분류
 - 범주형 자료(Categorical Data) · 질적 자료(Qualitative Data) : 관측값이 몇 개의 범주 또는 항목의 형태로 측정되는 자료
 - (명목형(Nominal)) : 범주에 순위가 의미 X ex) 혈액형
 - (순위형(ordinal)) : 범주에 순위가 의미 O ex) 선호도, 학점, 랭킹
 - 수치형 자료(Numerical Data) · 양적 자료(Quantitative Data) : 관측값이 수치로 측정되는 자료 ex) 중간고사 점수, 기·체중
 - (이산형(Discrete)) : 관측 가능한 값이 비연속적인 자료 ex) 사건 발생 건수, 동물원의 고기리 수
 - (연속형(Continuous)) : 관측 가능한 값이 연속적인 자료 ex) 귀·몸무게
- 변수(variable) : 관측되는 특성을 나타내는 것. 관측값들 사이에 이어갈 수 변동 있음. ex) 수치형 변수, 범주형 변수

■ 범주형 자료의 요약 : 도수분포표

- 전체 자료 중에서 각 범주에 속하는 자료의 횟수(frequency · 도수)를 요약하여 나타냄.
- 도수분포표(frequency table)
 - 도수(frequency) : 각 범주에 속하는 관측값의 개수
 - 상대도수(relative frequency) : 각 범주에 도수를 전체 도수로 나눈 값 (상대도수의 총합 = 1)
 - 각 범주에서 범주와 이에 대응하는 도수(상대도수)를 나열하여 표로 작성한 것
- Distinct (non-overlap) : 어떠한 관측값도 2개 이상의 범주에 포함 X
- Exhaustive : 모든 범주는 전체 데이터를 다 포함.

■ 통계학의 목표

- 자료의 수집 - 모집단에 대해 새로운 정보나 지식을 얻기 위해 자료를 효과적으로 수집 표본 추출의 과정과 범위 설정
- 자료의 요약 - 자료를 표·그림·수치(통계량)로 요약 정리 \Rightarrow 기술통계(Descriptive Statistics)
- 정보를 분석 - 불확실성 향시 배포 \Rightarrow 확률(Probability), 확률분포(Probability Distribution)
- 통계적 추론 - 추정·검정 \Rightarrow Statistical Inference

■ 원형그래프 (Pie Chart)

- 원을 각 범주간 상대도수에 비례하도록 중심각을 나누어 파이의 조각처럼 나타낸다.
- 각 범주가 전체에서 차지하는 비율을 파악하기 좋다.
- 각 범주간의 도수비교가 쉽지 않다.
- 범주의 수가 많은 경우에 그림기가 쉽지 않다.

■ 막대그래프 (Bar Chart)

- 각 범주에서 도수의 크기를 막대의 높이로 나타낸 그림
- 각 범주간 도수를 비교하기 좋다.
- 각 범주가 전체에서 차지하는 비율은 파악하기 쉽지 않다.
- 히스토그램과 다르게 범주별 구분을 위해 막대 사이에 공간 존재.
- 막대의 폭·넓이는 의미 X. (오직 높이)

■ 파레토그램 (Pareto Diagram)

- 막대그래프의 일정으로 상대도수가 큰 순으로 범주를 왼쪽부터 차례로 배열한 후, 누적 상대도수를 각 범주의 막대 위 중앙에 표시하고 그 점을 연결한 그림
- 여러 개의 범주 중에서 문제의 해석이나 해결에 도움을 주는 중요한 소수의 범주를 찾는데 도움을 준다.
- 각 범주들이 차지하는 비율과 상대도수가 증가하는 비율을 동시에 파악 → 어느 범주가 중요한지 쉽게 파악
- 순위형 (ordinal) 자료에는 의미 X

■ 이산형 자료의 요약

- 관측값이 종류가 적은 경우 : 범주형 자료를 요약하는 방법 사용 → 도수분포표, 원형그래프, 막대그래프
- 관측값의 종류가 많은 경우 : 연속형 자료를 요약하는 방법 사용

■ 연속형 자료의 요약 : 점도표

- 연속형 자료는 연속적인 값을 가지므로 범주형 자료처럼 몇 개의 범주로 나뉘어 있지 않음.
- 점도표 (Dot Diagram) 수평선 위에 각 관측값에 해당하는 위치에 점을 찍어 표시한 그림
관측값의 수가 적은 경우에 주로 사용 → 20~25개 이하 (너무 많으면 비효율적)
- 점도표는 자료의 불규칙성을 쉽게 파악할 수 있도록 한다.
- 자료의 수가 많은 경우에는 적절 X. → 자원을 몇 개의 그룹으로 나누어 표시 (도수분포표, 히스토그램)

■ 도수분포표 (Frequency Table)

- 관측값을 몇 개의 구간 (계급·class) 으로 나누고, 이 계급에 속하는 관측값의 수 (도수) 를 세어 작성
- 계급 구간 (class interval) : 각 계급에 포함되는 값의 범위
- 작성 방법
 - 1. 자료의 범위 (range) : 최대값 - 최소값
 - 2. 계급구간의 폭 or 계급의 수 : 계급의 수가 5~15 개가 되도록 자료의 범위 (or 계급의 수) 를 정함. ⇒ 계급의 폭을 결정
 - 3. 계급구간 : 관측값이 계급의 경계에 놓아지 않도록 계급구간 결정
 - 4. 각 계급에서 도수와 상대도수를 구함.
- 관측값이 경계에 오지 않고 최소값, 최대값이 각 계급의 중간에 오도록 시작값을 정한다
- 계급구간의 수 (or 계급구간의 폭) 를 정하는 법법
 - 계급의 수가 적으면 (계급 구간의 폭이 크면) 자료가 너무 간격히 요약 → 많은 정보를 잃게 된다
 - 계급의 수가 크면 (계급 구간의 폭이 작으면) 각 계급별로 어떤 경향을 가지는지 파악 어렵다
 - 자료 전체에 대한 분포 경향을 잘 나타내도록 계급의 수를 정한다 (5~15개)
- 도수분포표는 작성하는 방법에 따라 달라진다 (주관적)
- 자료의 특성에 따라서는 계급 구간의 폭을 다르게 할 수 있다. ↳ 소득 자료

■ 히스토그램(Histogram)

- 도수분포표를 바탕으로 각 계급에서 도수의 크기를 막대로 나타낸 그림 (이산형 자료의 막대그래프에 대응)
- 막대의 높이 = 상대도수 / 계급 구간의 폭
- 히스토그램의 전체 면적은 1
- 계급구간의 폭이 모두 같은 경우에는 막대의 높이를 이용해 비교 / 다른 경우에는 막대의 넓이를 이용해 비교
- 계급별 경계값을 공유해 막대 사이 공간 X . \rightarrow 어떤 구간의 도수가 0인 경우 공간 발생 가능
- 같은 자료라도 계급 구간의 수(폭), 시작값의 변화에 따라 히스토그램은 달라진다.

■ 도수 다각형(Frequency Polygon)

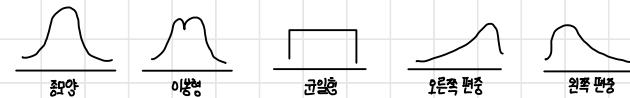
- 히스토그램의 각 계급구간의 막대 상단의 중간값을 연결한 그림
- 관측값의 변화에 따른 도수, 상대 도수의 변화, 자료의 중심 위치, 평균 정도 등 자료의 분포 특성을 히스토그램보다 쉽게 파악
- 하나의 차트에 여러 종류의 도수다각형을 나타내어 여러 자료를 비교하기 쉬운.

■ 줄기-잎 그림(Stem-and-leaf Plot)

- 히스토그램과 도수다각형은 자료의 분포를 쉽게 파악할 수 있지만 개개의 관측값에 대한 정보를 잊어버린다
- 줄기-잎 그림은 관측값을 앞단위(줄기)와 뒷단위(잎)로 나누어 나무의 줄기와 잎 모양으로 나타낸 그림
- 작성 방법
 1. 관측값을 앞단위와 뒷단위로 나눈다.
 2. 앞단위를 줄기로 하여 순서대로 세로로 배열, 그 옆에 수직선을 그린다.
 3. 뒷단위를 점으로 하여 오른쪽으로 가로로 배열
 4. 각 줄기에서 점 부분의 값을 작은 숫자가 원쪽으로 가도록 크기 순으로 재배열
- 히스토그램과 같이 자료의 분포 모양을 알 수 있으며, 관측값 개개의 정보를 얻을 수 있다.
- 자료의 수가 너무 많거나 흩어져 있는 경우 적용 X

■ 분포의 모양

- 대칭형 분포(symmetric distribution) : 가운데를 기준으로 양옆 대칭 \Rightarrow 중 모양 분포(정규 분포), 일봉형 분포
- 이봉형 분포(bimodal distribution) : 다른 2개의 집단의 가능성
- 균일형 분포(uniform distribution) : x축 값에 상관없이 높이 일정
- 편중된 분포(skewed distribution)
 - [오른쪽으로 편중] (skewed to the left) : 왼쪽으로 꼬리가 길
 - [왼쪽으로 편중] (skewed to the right) : 오른쪽으로 꼬리가 길



〈3장. 수치를 통한 연속형 자료의 요약〉

■ 모표, 그림, 수치

- 모표나 그림을 이용한 자료의 요약은 주관적이고 편관성이 부족 → 히스토그램은 계급구간의 폭(수), 시작값에 따라 달라짐.
- 수치를 이용한 요약은 상대적으로 객관적 → (중심위치, 페친정도) C 일반적인 위치의 측도

■ 중심 위치의 측도

• 평균 (mean)

$$- \text{표본평균 } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i / \text{ 모평균 (population mean)} : \mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- 중심위치의 측도로 가장 많이 사용되지만 극단적으로 폭거나 작은 값의 영향을 많이 받을 수 있다. (not robust)

• 중앙값 (median)

- 자료를 크기 순으로 배열할 때, 중앙에 위치하는 값

$$\left[\begin{array}{l} \text{자료의 수}(n) \text{가 짝수} : \frac{n+1}{2} \text{ 번째 관측값} \\ \text{자료의 수}(n) \text{가 짝수} : \frac{n}{2} \text{ 번째 관측값과 } (\frac{n}{2} + 1) \text{ 번째 관측값의 평균} \end{array} \right]$$

- 관측값의 50% 이상이 중앙값 이상이고, 관측값의 50% 이상이 중앙값 이하여야 한다.

- 평균과 달리 관측값의 변화에 민감하지 않고, 근/작은 관측값에 영향을 받지 않는다. (robust)

• 최빈값 (mode) : 관측값 중에서 빈도수가 가장 큰 값을 찾는다. 주로 이산형 범주형 자료에 대한 중심위치로 사용

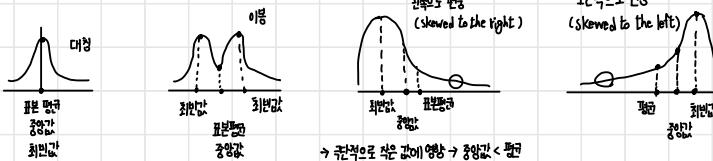
• 표본평균과 중앙값의 비교

(표본평균 : 이해하기 쉽고, 자료의 전체 값에 의해 결정. 자료에 대한 기본 통계 수치로 자주 이용)

(중앙값 : 자료의 중앙 부분에 대한 영향을 받음. 극단값 영향 X)

→ 편향된 자료 (skewed data)에 대해서는 중앙값이 표본평균보다 중심위치의 측도로 더 적합.

• 분포 모양에 따른 중심위치의 측도



■ 페친 정도의 측도

- 자료의 중심 위치에 대한 특징으로 자료의 분포 특성 파악 부족 → 같은 중심 위치라도 분포 형태가 다를 수 있음. → 자료가 중심 위치로부터 얼마나 퍼져있는지를 나타내는 측도

• 분산 (variance) 과 표준편차 (standard deviation)

- 편차 (deviation) : $x_i - \bar{x}$ (관측값 - 표본 평균) → (+), (-) 가능

$$- \text{표본분산} : S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$$

$$* \text{모분산} : \sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

$$\left(\begin{array}{l} \text{자료의 수} = n \\ \text{평자의 합} = \sum_{i=1}^n (x_i - \bar{x}) = 0 \end{array} \right)$$

평자제곱합의 자유도 (degrees of freedom) = $n-1$

• 표본분산의 성질

$$\left[\begin{array}{l} S^2 \geq 0 \\ S^2 = 0 \Leftrightarrow x_i = \bar{x} \text{ for all } i \text{ (매우 특수한 경우)} \end{array} \right]$$

S^2 이 크다 \Leftrightarrow 편차의 절대값이 크다 \Leftrightarrow 관측값이 평균으로부터 멀리 퍼져있다 (페친 정도가 크다)

- 표본 표준편차 : $S = \sqrt{S^2}$ → 표본 표준편차의 단위는 원 자료와 같다.

• 범위 (range) : 최대값 - 최소값

- 범위는 간단하게 구할 수 있지만, 자료의 양의 값에 의해서만 결정되고 중간 범위의 다른 관측값이 어떤 분포 특성을 가지는지 고려하지 못함

• 백분위수 (percentile) : 관측값을 작은 것부터 크기 순으로 배열했을 때, 자료의 $100 \times p\%$ 에 해당하는 값을 제 $100 \times p$ 백분위수라 함

- $100 \times p$ 백분위수 : 그 값보다 작거나 같은 자료의 수가 $n \times p$ 개 이상이고, 그 값보다 크거나 같은 자료의 수가 $n \times (1-p)$ 개 이상인 값

• 제 $100 \times p$ 백분위수 구하는 법

관측값을 크기 순으로 배열
※ 중앙값 = 제 50 백분위수

p 가 정수이면, $[np+1]$ 번째 값과 $(np+1)$ 번째 값의 평균

np 가 정수이면, $[np+1]$ 번째 값

- 위치의 측도: 백분위수, 세분위수

• 사분위수 (quartile) 와 사분위수 범위 (IQR)

- 사분위수 : 자료를 크기 순으로 배열할 때 사용하는 값

$$\begin{aligned} \text{제 1 사분위수 } (Q_1) &= \text{제 25 백분위수} \\ \text{제 2 사분위수 } (Q_2) &= \text{제 50 백분위수} = \text{중간값} \\ \text{제 3 사분위수 } (Q_3) &= \text{제 75 백분위수} \end{aligned}$$

- 사분위수 범위 (IQR) : $Q_3 - Q_1 \rightarrow$ 상위 25% 및 하위 25%의 관측값을 제외한 중앙 부분 50%의 관측값의 범위

→ 극단값을 제외한 자료의 편진 정도를 나타내며, 한쪽으로 치우친 자료의 편진 정도를 나타낼 때 유용

• 표준편차, 범위, 사분위수 범위의 비교

- 표준편자는 평균과 같이 전체 자료의 값에 영향을 받는다
- 표준편자를 중심위치의 측도로 사용할 때, 표준편자는 편진 정도의 측도로 사용 → 중간값을 중심위치의 측도로 사용할 때, 사분위수 범위를 편진 정도의 측도
- 표준편자는 전체 자료의 편진 정도를 골고루 반영하는 반면, 극단값의 영향을 받음 (not robust)
- 사분위수 범위는 극단값의 영향을 덜 받지만 (robust), 전체 자료의 편진 정도를 나타내진 못함.
- 범위는 극단값에 큰 영향을 받고 전체 자료의 편진 정도도 나타내지 못함

• 변동계수 (coefficient of variation, CV)

- 두 자료의 편진 정도를 비교할 때, 자료의 단위에 영향을 끼울 수 있는 표준편자와 사분위수 범위는 적절 X

- 여러 자료의 편진 정도를 비교할 때는 자료의 단위에 영향을 끼치지 않는 상대적인 측도가 필요

$$\text{변동계수 (CV)} = \frac{\text{표준편자}}{\text{평균}} \times 100 \quad (\text{단위} = \%) \rightarrow \text{표본평균에 대한 상대적인 편진 정도를 백분율로 나타낸 값}$$

• 상자그림 (box plot)

- 최소값, 최대값, 사분위수 범위 등을 이용하여 자료의 중심위치, 편진 정도 등을 퀘트의 그림으로 나타낸 것

- 작성법 1. Q_1, Q_2, Q_3, IQR 을 구한다

2. Q_1, Q_3 를 상자로 연결. Q_2 에 수직선

3. 상자의 양 끝으로부터 $1.5 \times IQR$ 크기의 범위 경계. 이 범위 안에 포함되는 최소값과 최대값을 Q_1 과 Q_3 로 부터 선으로 연결

4. 양 경계로부터 벗어난 이상치를 *로 표시

• 표준화 (standardized) 된 자료

$$Z = \frac{x - \bar{x}}{s}$$

- 특정 자료값이 평균으로부터 표준편자의 몇 배만큼 떨어져 있는지를 측정하는 상대적 위치의 측도
- 두 종류 이상의 자료에서 한 자료의 특정 관측값과 다른 자료의 특정 관측값의 크기를 상대적으로 비교할 때 사용
- 총 표본의 대상 분포 성질

- 평균을 중심으로 자료의 분포를 표준편자를 이용하여 대체적으로 파악

- 자료의 분포가 대칭이고 종모양을 이를 때 경험적 사실로 다음의 성질이 있다

$$\left(\begin{array}{l} \bar{x} + s \text{ 구간에 } 68\% \\ \bar{x} + 2s \text{ 구간에 } 95\% \\ \bar{x} + 3s \text{ 구간에 } 99.7\% \end{array} \right) \text{ 의 자료가 포함}$$

< 4장. 두 변수 자료의 요약 >

■ 두 변수 자료의 요약

- 조사 대상의 각 개체 (sampling unit)로부터 두 개의 변수를 동시에 관측 → 주로 두 변수 사이의 연관성에 관심

■ 분할표 (contingency table)

- 두 변수가 모두 범주형인 경우에 표를 이용하여 요약하는 방법
- 한 변수에 대한 범주는 가로로, 다른 변수에 대한 범주는 세로로 하는 표를 만들어 각 셀마다 해당하는 도수 (상대도수)를 세어 나타낸 표
- Side-by-Side Bar Chart: 두 범주형 변수의 자료를 그림을 이용하여 요약 → column/row percentage 조건부 확률

■ 두 연속형 변수의 요약: 산점도 (Scatter Plot)

- 두 연속형 변수 X와 Y 사이의 어떤 연관성을 파악할 수 있는 그림 → 좌표평면 위에 (x_i, y_i) 에 해당하는 위치에 점을 찍어 나타내는 그림
- $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 쌍으로 구성 (상성 짝) → 모든 개체에 대해 두 차를 모두 있어야 함
- 두 변수 사이에 연관성이 있는가? 있다면 어떤 연관성 (증가 or 감소)이 있는가?
- 산점도를 통해 두 연속형 변수의 연관성을 기하학적으로 파악할 수 있지만, 산점도에 대한 해석은 주관적. → 단위에 따라 그레프도 달라짐.

■ 상관계수 (correlation coefficient)

- 두 연속형 변수 사이의 연관성을 확장적인 수치로 나타낼 수 있는 방법

• 표본상관계수 (sample correlation coefficient)

파이슨 (Pearson) 표본상관계수

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{\text{표본 공분산}}{\text{X의 표준편차} \times \text{Y의 표준편차}}$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} = \sum_{i=1}^n x_i y_i - \frac{1}{n} (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)$$

⇒ 표본 공분산 (sample covariance)

• 표본상관계수의 성질

- $-1 \leq r \leq 1$
- $r > 0 \Leftrightarrow X가 증가할 때 Y도 증가하는 경향이 있다. (양의 상관관계)$
- $r < 0 \Leftrightarrow X가 증가할 때 Y는 감소하는 경향이 있다. (음의 상관관계)$
- $r = 1 \Leftrightarrow 모든 관측값이 기울기가 양수인 직선 위에 있다. (y = ax + b, a > 0)$
- $r = -1 \Leftrightarrow 모든 관측값이 기울기가 음수인 직선 위에 있다.$
- $r = 0 \Leftrightarrow X와 Y 사이에 선형관계가 없다. (연관성이 없다는 의미)$
- $|r|이 1에 가까울수록 X와 Y 사이에 선형관계가 강하다.$
- $|r|이 1이 아예 가까울수록 X와 Y 사이에 선형관계가 약하다.$

- 상관계수는 두 변수 사이의 선형관계에 대한 측도로 r의 부호는 두 변수 사이의 연관성의 방향을, r의 절대값은 선형관계의 강도를 나타낸다.

- 상관계수는 단위가 없는 상대적인 값으로 여러 자료에 대한 상관계수를 이용하여 두 변수 사이의 연관성을 비교할 수 있다.

- 자료를 선형변환한 경우에도 상관계수의 절댓값 변화 X (단위도 상관 X)

- $r = 0$ 인 여러 가지 경우
 - 두 변수 사이에 특별한 관계 X

수직관계 (\parallel), 수평관계 ($\ldots\ldots$)

선형관계는 없지만 관련관계는 있다.

- 주의사항
 - 표본상관계수는 두 변수 사이의 선형관계를 측정하는 값 → 두 변수 사이에 선형 관계가 있는 경우에도 r은 0에 가깝게 될 수 있다.

▶ 예를 들어 두 변수 사이의 선형관계가 '크다' 혹은 '작다'를 판별할 수 없는 경우도 있다. (ex. 두 개의 그룹으로 된 경우 $r=0$ 이 될 수도)

• 상관관계와 인과관계

- 큰 상관계수의 값이 두 변수 사이의 인과관계를 의미 X

▶ 일부로 인과관계 예상 X

- 두 변수 X와 Y 사이의 관계를 제어 (control) 하는 또 다른 변수 (Z)가 존재 가능. ⇒ 잠재변수 (lurking variable)

- X와 Y 사이의 관계를 알기 위해서는 잠재변수 (Z)를 control 해야 함.

〈5장. 확률〉

■ 확률 (probability)

- 여러 가지 가능한 결과 중 하나가 일어나는 시험에서 그 총 일부가 일어날 가능성을 $0 \sim 1$ 사이의 값 ($0 \leq P(A) \leq 1$)으로 나타낸 것
- 확률에 대한 기본 용어

- 표본공간 (sample space : Ω) : 시험에서 일어날 수 있는 모든 결과들의 집합
- 근원사건 (elementary event : ω) : 시험에서 일어날 수 있는 개개인의 결과
- 사건 (event, A) : 어떤 특성을 가지고 있는 결과들의 합집 (표본공간의 부분집합)
- $P(A)$: 사건 A가 발생할 확률

■ 확률의 의미와 확률의 종류

- $P(A)$ [실현을 반복할 때 사건 A가 발생하는 비율의 극한]
 - 사건 A가 일어날 가능성에 대한 믿음을 수치화한 것
- 두 사건 A와 B에 대하여 $A \cap B = \emptyset$ 일 때, A와 B는 서로 배반 (disjoint · mutually exclusive)
- 확률의 공리
 - 모든 사건 A에 대하여 $0 \leq P(A) \leq 1 \rightarrow 0 \leq P(\cup A_i) \leq 1$
 - $P(\Omega) = 1 \rightarrow P(\{w_1, w_2, \dots, w_n\}) = P(w_1) + P(w_2) + \dots + P(w_n) = 1$
 - 사건 A_1, A_2, \dots 가 서로 배반일 때 (서로 다른 i, j 에 대하여 $A_i \cap A_j = \emptyset$) $P(A_1) + P(A_2) = P(A_1 \cup A_2)$, $P(\bigcup_{k=1}^n A_k) = \sum_{k=1}^n P(A_k)$

■ 확률의 계산

- 표본공간의 근원사건 수가 유한하고, 각 근원사건이 일어날 가능성이 동일할 때

$$P(A) = \frac{\text{A에 속하는 근원사건의 수}}{\text{표본공간에 속하는 근원사건의 수}} \quad (\text{수학적 확률})$$

• 실험을 N번 반복할 때

$$P(A) = \lim_{N \rightarrow \infty} \frac{\text{N번 중 A가 일어나는 횟수}}{N} \quad (\text{통계적 확률})$$

→ 서로 배반. 동시에 발생 X

- 확률의 법칙 : 한 사건의 확률 = 그 사건에 포함되는 근원사건들의 확률의 합

■ 경우의 수 (counting rule)

- 곱셈 법칙 (multiplication principle) : 여러 사건들이 서로 영향을 주지 않을 때, 그 사건들이 동시에 일어나는 경우의 수
- 순열 (permutations) : 전체 n개 중에서 r개를 뽑는 경우의 수 \rightarrow 뽑힌 r개의 순서 고려 ex) ${}_{10}P_3 = \frac{10!}{7!} = 10 \times 9 \times 8 = 720$
- 조합 (combinations) : 전체 n개 중에서 r개를 뽑는 경우의 수 \rightarrow 뽑힌 r개의 순서 고려 X ex) ${}_{10}C_3 = \frac{10!}{(10-3)!3!} = 120$
- 사건들의 기본적인 연산 \rightarrow 여사건 (A^c) / 합사건 ($A \cup B$) / 곱사건 ($A \cap B$) / 배반사건 ($A \cap B = \emptyset$)

■ 확률의 법칙

- 여사건의 확률법칙 : $P(A^c) = 1 - P(A)$, $P(A \cup A^c) = P(A) + P(A^c) = P(\Omega) = 1$
- 합사건의 확률법칙 : $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- 곱사건의 확률법칙 : 조건부확률
- 배반사건의 확률법칙 : 사건 A_1, A_2, \dots, A_n 가 서로 배반일 때, $A_i \cap A_j = \emptyset$ for $i \neq j$, $P(\bigcup_{k=1}^n A_k) = \sum_{k=1}^n P(A_k)$
- 조건부확률 (conditional probability)

 - 사건 B가 발생한다는 정보가 주어졌을 때, 사건 A가 발생할 확률 \rightarrow 조건부확률 ($P(A|B)$)
 - 일반적으로 아무런 정보가 없는 상태에서 사건 A가 발생할 확률 ($P(A)$)과 사건 B가 발생한다는 조건 하에서 사건 A가 발생할 조건부확률 ($P(A|B)$)는 다르다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

* $P(A \cap B) = P(B \cap A)$
 $P(A|B) = P(B|A)$
 $P(A|B) \neq P(B|A)$

- 조건부확률 $P(A|B)$ 은 사건 B 안에서 A가 차지하는 비율이라고 생각할 수 있다. 즉 조건부확률 $P(A|B)$ 를 구할 때는 표본공간이 B로 한정된다고 생각하고, 이 안에서 사건 A가 일어날 확률을 구하면 된다.

• 조건부 확률을 이용하면 $P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

〈 6장 확률분포 〉

■ 두 사건의 독립(independence)

• 두 사건 A와 B에 대하여 $P(A \cap B) = P(A)P(B)$ 이 성립

- $P(A|B) = P(A)$ 이 성립하면 $P(A \cap B) = P(A)P(B)$ 도 성립

$$\frac{P(A \cap B^c)}{P(B^c)} = \frac{P(A) - P(A \cap B)}{1 - P(B)} = \frac{P(A) - P(A|B)P(B)}{1 - P(B)} = \frac{P(A)(1 - P(B))}{1 - P(B)} = P(A)$$

- 사건 A가 발생할 확률은 사건 B가 발생하는 발생하지 않은 상관없이 같은 값

- 사건 B가 발생했다는 정보가 사건 A의 확률을 구하는 데 영향 X

- 사건 A와 사건 B는 서로 영향 X \Rightarrow 독립

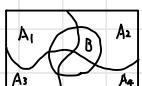
$$P(A|B) = P(A) \Leftrightarrow \frac{P(A \cap B)}{P(B)} = P(A) \Leftrightarrow P(A \cap B) = P(A)P(B)$$

• 두 사건 A와 B에 대하여 $P(A \cap B) = P(A)P(B)$ 가 성립할 때, 사건 A와 사건 B는 서로 독립(independent)

• 표본공간에서 사건 A가 차지하는 비율과 사건 B 안에서 A가 차지하는 비율이 같을 때, 사건 A와 B는 서로 독립.

■ 베이즈 정리(Bayes' rule)

• A_1, A_2, \dots, A_n 이 서로 배반이고, $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$ 일 때 A_1, A_2, \dots, A_n 은 표본공간의 분할(partition)



① 총 확률의 법칙 : $P(B) = \sum_{k=1}^n P(A_k)P(B|A_k) = \frac{n}{\sum_{k=1}^n P(A_k)} P(A_k \cap B)$ \rightarrow 기준으로 된 조건부확률

$$\textcircled{2} \text{ 베이즈 정리 : } P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{k=1}^n P(A_k)P(B|A_k)} = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_{k=1}^n P(A_k)P(B|A_k)} = \frac{P(A_k)P(B|A_k)}{\sum_{k=1}^n P(A_k)P(B|A_k)}$$

• 베이즈 정리는 기존의 경로와 새로운 정보를 이용하여 update 하는 과정에 활용될 수 있음

$P(A_1), P(A_2), \dots, P(A_n)$: n개의 사건에 대한 사전정보(prior information) $\rightarrow P(A_1|B), P(A_2|B), \dots, P(A_n|B)$: n개의 사건에 대한 사후정보
사건 B에 대한 발생 정보

■ 확률변수(random variable)

• 확률변수(random variable)란 표본공간에서의 각 결과(근원사건) \rightarrow 실수 값 대응시키는 함수 $\Rightarrow X, Y, Z$ 로 나타냄

- 기초적 표본공간 \xrightarrow{X} 새로운 표본공간(실수)

• 확률변수가 가질 수 있는 값에 따른 분류

① 이산확률변수(discrete random variable) : 확률변수가 가질 수 있는 값의 수가 유한개 혹은 무한개더라도 셀 수 있는 경우

② 연속확률변수(continuous random variable) : 확률변수가 어느 구간에 속하는 모든 값을 가질 수 있는 경우 \rightarrow 확률변수 반드시 무한

■ 이산확률변수(discrete random variable)

• 확률분포(probability distribution) : 확률변수가 가질 수 있는 값과 그에 대응하는 확률을 나타낸 것

• 이산확률변수의 경우에 X의 확률분포는 X가 가질 수 있는 값과 그 값을 가질 확률에 의해 확률분포 결정

■ 확률질량함수(probability mass function · PMF)

• 이산확률변수는 X가 가질 수 있는 값 x_1, x_2, \dots, x_n 에서의 확률 $P(X=x_i)$ 을 x_i 의 힘으로 확률분포를 나타냄.

• $f(x_i) : x_1, x_2, x_3, \dots$

$$f(x_i) = P(X=x_i) \rightarrow \text{확률질량함수}$$

• 확률질량함수의 성질

$$\begin{cases} 0 \leq f(x_i) \leq 1 \\ \sum_i f(x_i) = 1 \end{cases}$$

$$P(a \leq X \leq b) = \sum_{x:a \leq x \leq b} f(x_i)$$

■ 확률변수의 기댓값(평균)

• 확률변수 X 는 여러 가지 값 중에서 하나의 값을 가지며, 각 값을 가질 확률은 서로 다를 수 있다. 이때 확률을 가중치로 한 X 의 중심위치(평균)를

X 의 기댓값(expected value), $E(X)$ 혹은 \bar{x} 라고 함.

• 이산확률변수의 기댓값 $E(X) = \sum_{x_i} x_i f(x_i) \rightarrow E(\bar{x})$ 는 확률을 가중치로 한 X 가 가질 수 있는 값의 가중 평균

$$\cdot E(a+bX) = a + bE(X)$$

• 확률변수의 평균과 표본평균

— n 개의 자료 x_1, x_2, \dots, x_n 의 표본평균은 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{n}{n} \cdot \frac{1}{n} \cdot x_i$

— n 개의 값 x_1, x_2, \dots, x_n 을 가질 확률이 $f(x_1), f(x_2), \dots, f(x_n)$ 인 이산확률변수 X 의 평균은 $E(X) = \sum_{x_i} x_i f(x_i)$

— 확률변수의 평균은 표본평균을 구하는 식에서 가중치인 $\frac{1}{n}$ 대신에 확률변수 X 의 값이 x_i 일 확률인 $f(x_i)$ 를 구한 것

↓ 기밀교사

〈 6장. 확률분포 (두부분) 〉

■ 확률변수의 분산과 표준편차

• 확률변수 X 의 기댓값(평균) → 표본 평균에 대응하는 모집단(확률 분포)의 평균 (X 의 확률 분포의 중심위치를 나타내는 속도) $M(E(X))$ vs \bar{x}

• 확률변수 X 의 분산(variance)은 표본 분산에 대응 → 모집단(확률 분포)의 편차 정도를 나타내는 속도 σ^2 ($Var(X)$) vs s^2

• 표본분산 ($s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$)은 자료에 대하여 표본평균으로부터의 편차의 제곱을 평균한 것

— 확률변수의 분산은 X 의 평균으로부터의 편차의 제곱에 대한 기댓값 → $Var(X)$ or σ^2

$$Var(X) = E[(X-M)^2] = \sum_{x_i} (x_i - M)^2 f(x_i)$$

$$E(h(x_i)) = \sum_{x_i} h(x_i) f(x_i) \rightarrow h(M) \cdot (X-M)^2$$

— 확률변수 X 의 표준편차(standard deviation)는 분산의 양의 제곱근 → $sd(X)$ or σ

— 분산의 간접 계산식

$$Var(X) = E(X^2) - (E(X))^2 = E(X^2) - M^2$$

■ 결합분포(joint distribution)

• 두 확률변수 X 와 Y 에 대하여 X 가 취하는 값과 Y 가 취하는 값 각 성에 대응하는 확률을 나타낸 것

$$\cdot$$
 이산형 결합분포(결합 확률질량함수·joint pmf) $f(x_i, y_j) = P(X=x_i, Y=y_j)$

■ 주변확률분포(marginal distribution)

• 두 확률변수 중에서 어느 한 확률변수의 확률분포

$$\cdot X$$
의 주변분포 : $f_x(x_i) = P(X=x_i) = \sum_j P(X=x_i, Y=y_j) = \sum_j f(x_i, y_j)$

$$\cdot Y$$
의 주변분포 : $f_y(y_j) = P(Y=y_j) = \sum_i P(X=x_i, Y=y_j) = \sum_i f(x_i, y_j)$

• 두 확률변수 X 와 Y 의 결합분포가 주어져 있을 때, X 의 기댓값이나 분산 등을 X 의 주변확률분포를 이용하여 계산 가능

$$\rightarrow E(X) = \sum x_i f_x, E(X^2) = \sum x_i^2 f_x, \text{Var}(X) = E(X^2) - (E(X))^2$$

$$\cdot$$
 일반적으로 $E(X+X_2+\dots+X_n) = E(X_1) + E(X_2) + \dots + E(X_n)$

■ 공분산(covariance)

• 두 확률변수 X 와 Y 의 결합분포를 사용하여 X 와 Y 의 선형 연관성을 나타낸 속도. $\rightarrow Cov(X, Y)$

$$Cov(X, Y) = E((X-M_X)(Y-M_Y)) = E(XY - XM_Y - YM_X + MM_Y)$$

$$= E(XY) - M_X E(Y) - M_Y E(X) + MM_Y = [E(XY) - M_X M_Y]$$

$$E(XY) = \sum_x \sum_y xy f(x, y)$$

$$E(XY) = E(X)E(Y)$$

• 공분산의 의미

— $Cov(X, Y) > 0$: X 와 Y 가 같은 방향으로 변화할 확률이 크다. $\rightarrow X$ 가 증가할 때 Y 도 증가하는 경향.

— $Cov(X, Y) < 0$: X 와 Y 가 다른 방향으로 변화할 확률이 크다. $\rightarrow X$ 가 증가할 때 Y 는 감소하는 경향.

• 공분산의 성질

$$\begin{aligned} \text{① } \text{Cov}(ax, by) &= E((ax - E(ax))(by - E(by))) = E(ab(X - E(X))(Y - E(Y))) \\ &= ab E((X - E(X))(Y - E(Y))) = ab \cdot \text{Cov}(X, Y) = ab(E(XY) - E(X)E(Y)) \end{aligned}$$

② 공분산은 X 와 Y 의 단위의 영향을 받음

③ 공분산의 값을 이용하여 두 확률변수의 연관성의 정도 측정 불가 → 상관계수 활용

■ 상관계수 (correlation coefficient)

• 두 확률변수 X 와 Y 의 단위에 영향을 받지 않는 확률변수 사이의 선형 연관성을 측정한 값 $\rightarrow \text{Corr}(X, Y)$, ρ

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}} = \rho$$

• 상관계수의 성질

$$\text{Corr}(ax, by) = \frac{\text{Cov}(ax, by)}{\sqrt{\text{Var}(ax)} \cdot \sqrt{\text{Var}(by)}} = \frac{ab \text{Cov}(X, Y)}{\sqrt{a^2 \text{Var}(X)} \cdot \sqrt{b^2 \text{Var}(Y)}} = \frac{ab \text{Cov}(X, Y)}{|ab| \sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} = \frac{ab}{|ab|} \text{Corr}(X, Y)$$

- 상관관계의 절대값은 단위 영향 X

- $-1 \leq \rho \leq 1$
- $\rho = 1 \Leftrightarrow P(Y = a + bX) = 1 \text{ for } b > 0$
- $\rho = -1 \Leftrightarrow P(Y = a + bX) = 1 \text{ for } b < 0$
- $\rho = 0 \Leftrightarrow \text{두 확률변수 사이에 선형관계 } X$
- $|\rho| \text{이 } 1 \text{에 가까울수록 } X \text{와 } Y \text{의 선형관계가 강하다}$

■ 확률변수의 독립(independence)

• 두 사건 A, B 에 대하여 $P(A \cap B) = P(A)P(B)$ 일 때, A 와 B 는 서로 독립

• 두 이산형 확률변수 X 와 Y 에 관하여, 모든 가능한 값 x_i, y_j 에서

$$P(X=x_i, Y=y_j) = P(X=x_i) \cdot P(Y=y_j) \Leftrightarrow f(x_i, y_j) = f(x_i) \cdot f(y_j)$$

일 때, X 와 Y 는 서로 독립

• X 와 Y 가 독립일 때

$$E(XY) = \sum_x \sum_y xy f(x, y) = \sum_x \sum_y xy f(x) f(y) = \sum_x x f(x) \cdot \sum_y y f(y) = E(X)E(Y)$$

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0 \Rightarrow \text{Cov}(X, Y) = 0$$

- 그러나 공분산(상관계수)이 0이라 하더라도, X 와 Y 가 독립이 아닐 수 있다.

$\rightarrow X$ 와 Y 의 공분산이 0이라는 것은 단지 두 확률변수 사이에 선형 관계가 없다는 것을 의미

■ 합과 차의 분산

• 합의 분산

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

• 차의 분산

$$\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$$

$$\left. \begin{aligned} &\text{X와 } Y \text{가 독립} \rightarrow \text{Cov}(X, Y) = 0 \\ &\therefore \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) \end{aligned} \right)$$

〈7장. 이항분포와 그에 관련된 분포들〉

■ 베르누이 시행 (Bernoulli trial)

- 시행 또는 실험의 결과가 두 가지 총 하나인 경우
 - 각 시행의 결과를 성공(s)과 실패(f) 중 하나로 분류
 - 각 시행에서 성공의 확률은 p (항상 일정, constant)
 - 각 시행은 서로 독립

ex) 동전 던지기 (앞면-성공 / 뒷면-실패) $\rightarrow p=0.5$ 고정. 시행이 서로 독립

■ 이항분포 (binomial distribution)

- $X =$ 성공의 확률이 p인 베르누이 시행을 n번 반복할 때 성공의 수
 - n번 조건이 동일한 실험 반복 (Identical trials)
 - 2개의 결과만 존재 (s or f)
 - constant p
 - 각 시행은 독립
- X 가 가질 수 있는 값: $X = 0, 1, 2, \dots, n \rightarrow X \sim B(n, p)$: 이항 확률변수 (binomial random variable)

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (x=0, 1, \dots, n)$$

• 정수 a와 b에 대하여

$$- P(X=a) = P(X \leq a) - P(X < a) = P(X \leq a) - P(X \neq a-1)$$

$$- P(X \geq a) = 1 - P(X < a) = 1 - P(X \leq a-1)$$

$$- P(a \leq X \leq b) = P(X \leq b) - P(X < a) = P(X \leq b) - P(X \leq a-1)$$

$$- P(a < X < b) = P(X < b) - P(X \leq a) = P(X \leq b-1) - P(X \leq a)$$

• R function

$$- P(X \leq c) = pbinom(c, n, p)$$

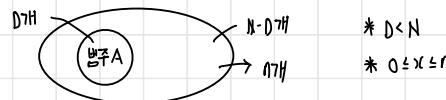
$$\begin{aligned} - P(X \leq 2) &= pbisom(2, 5, 0.3) = 0.827 \\ - P(X \geq 2) &= dbinom(2, 5, 0.3) = 0.3087 \end{aligned}$$

• 기대값 / 분산

$$E(x) = np \quad / \quad \text{Var}(x) = np(1-p) \quad / \quad E(x(x-1)) = n(n-1)p^2$$

■ 초가능분포 (hypergeometric distribution)

• $X = N$ 개의 원소로 이루어진 모집단에서 D개는 범주 A에 속할 때, 모집단에서 임의로 추출한 k개 중에서 범주 A에 속하는 원소의 수



• X 가 가질 수 있는 값: $\max(0, n-(N-D)), \dots, \min(n, D)$

$$P(X=x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}}$$

• 기대값과 분산

$$E(x) = \sum_{\substack{\text{모집단} \\ \text{선택}} \text{경우}} x f(x) = \sum_{\substack{\text{모집단} \\ \text{선택}} \text{경우}} x \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}} = n \cdot \frac{D}{N} = np \quad (p = \frac{D}{N})$$

$\rightarrow np(1-p) \cdot \frac{N-n}{N-1}$

$$\text{Var}(x) = E(x^2) - [E(x)]^2 = E(x(x-1)) + E(x) - [E(x)]^2 = n \cdot \frac{D}{N} \left(1 - \frac{D}{N}\right) \cdot \frac{N-n}{N-1}$$

■ 포아송분포 (Poisson distribution)

• X : 일정한 구간에서 다음 사건을 만족하는 특정한 사건의 발생 횟수

- 아주 짧은 구간에서는 사건이 2회 이상 발생할 확률은 0에 가깝다 (사건 0 or 1회 발생)

- 아주 짧은 구간에서 사건이 발생할 확률은 구간의 길이에 비례

- 서로 겹쳐지 않는 두 구간에서 발생하는 사건의 수는 서로 독립

• X 가 가질 수 있는 값: $X = 0, 1, 2, \dots \Rightarrow X \sim \text{Poisson}(m)$ • 포아송 확률변수

$$P(X=x) = \frac{m^x e^{-m}}{x!} \quad (x=0, 1, 2, \dots) \rightarrow m: \text{시간당 평균 발생 횟수}$$

- $P(X \leq c) \rightarrow \text{ppois}(c, m) / P(X=c) \rightarrow \text{dpois}(c, m)$

• 포아송분포의 기댓값과 분산

$$E(X) = \sum_{x=0}^{\infty} x \cdot f(x) = \sum_{x=0}^{\infty} x \cdot \frac{m^x e^{-m}}{x!} = \sum_{x=1}^{\infty} \frac{m \cdot m^{x-1} e^{-m}}{(x-1)!} = m \cdot \sum_{j=0}^{\infty} \frac{m^j e^{-m}}{j!} = [m]$$

$$\text{Var}(X) = E(X^2) - (E(X))^2 = E(X(X-1)) + E(X) - (E(X))^2 = \sum_{x=0}^{\infty} x(x-1) \cdot \frac{m^x e^{-m}}{x!} + m - m^2 = [m]$$

• 포아송분포와 이항분포 사이의 관계

- X : 단위 시간당 발생하는 사건의 수 $\sim \text{Poisson}(m)$ 일 때 구간을 나누면하고, n 이 충분히 크리고 하자.

- n 이 충분히 크면 한 구간에서 발생하는 사건의 수는 0또는 1이고, 각 구간에서 사건이 발생할 확률은 $\frac{m}{n}$ 이므로, 근사적으로

$$X \sim B(n, \frac{m}{n})$$

• 이항분포의 포아송근사

- $X \sim B(n, \frac{m}{n})$ 일 때, n 이 충분히 크면 $P(X=x) \approx \frac{m^x e^{-m}}{x!}$

- $X \sim B(n, p)$ 일 때, n 이 충분히 크고 p 가 매우 작으면 $P(X=x) \approx \frac{m^x e^{-m}}{x!} (m=np)$

〈 8. 정규분포 〉

■ 정규분포 (Normal distribution)

• 종 모양 (대칭)의 확률분포이며, 여러 종류의 자료를 설명할 수 있는 확률분포

• 실수 영역에서 값이 가지는 연속확률변수에 대한 분포

- 연속확률변수 (continuous random variable): 어느 구간에 속하는 모든 값을 가질 수 있는 확률변수 ex) 버스를 기다리는 시간, 제품의 수명 등

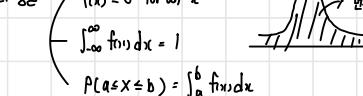
- 이산확률변수의 확률 분포는 확률변수가 가지는 각 값에서의 확률에 의하여 정해지지만, 연속확률변수는 각 구간에서의 확률에 의하여 확률분포가 정해짐

→ 연속 확률변수는 각 값을 가질 확률이 언제나 0 $\Rightarrow P(X=x) = 0 \text{ for all } 0 \leq x \leq 1$ (구간이 아닌 점에 대한 확률은 0)

• 확률질량함수 (pmf): 이산확률변수 X 에 대하여 $P(a \leq X \leq b) = \sum_{x=a}^b f(x)$ 일 때 $f(x)$

• 확률밀도함수 (pdf): 연속확률변수 X 에 대하여 $P(a \leq X \leq b) = \int_a^b f(x) dx$ 일 때 $f(x)$

- 확률밀도함수의 성질



$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

→ 이산확률변수와 달리 둘로 나뉨 ×

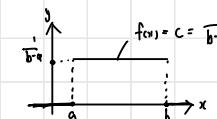
- 연속확률변수는 각 값을 가질 확률이 0이므로 $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X \leq b)$

→ 즉, 모든 x 에 대하여 $P(X=x) = 0$

■ 균일분포 (Uniform distribution)

• 균일분포의 확률밀도함수

$$f(x) = \begin{cases} \frac{1}{b-a} & (a \leq x \leq b) \\ 0 & (x < a \text{ or } x > b) \end{cases}$$



• 기호: $X \sim \text{Unif}(a, b)$, $U(a, b)$ → 균일분포는 양 끝부분 (a, b) 에 의하여 확률분포 결정

• $P(X \leq c) = P(X < c) = \text{밀연 } X \text{ 높이} = (c-a) \times \frac{1}{b-a} = \frac{c-a}{b-a}$

• $P(c \leq X \leq d) = P(c < X < d) = \text{밀연 } X \text{ 높이} = (d-c) \times \frac{1}{b-a} = \frac{d-c}{b-a}$

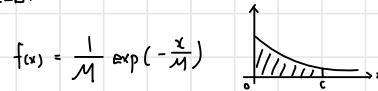


• 정규분포의 평균과 분산

$$M = E(x) = \frac{a+b}{2} / \sigma^2 = Var(x) = \frac{(b-a)^2}{12}$$

■ 지수분포(exponential distribution)

• 지수분포의 확률밀도함수



• 기호: $X \sim Exp(M) \rightarrow$ 지수분포는 평균 M 에 의하여 확률분포가 정해짐

$$\cdot P(X \leq c) = P(X < c) = \int_0^c f(x) dx = \int_0^c \frac{1}{M} \exp(-\frac{x}{M}) dx = 1 - \exp(-\frac{c}{M}) \rightarrow \text{누적분포함수(CDF)}$$

• 지수분포의 평균과 분산

$$M = E(x) = M / \sigma^2 = Var(x) = M^2$$

■ 정규분포(normal distribution)

• 정규분포의 확률밀도함수

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

• 기호: $X \sim N(M, \sigma^2) \rightarrow$ 정규분포는 평균 M 과 분산 σ^2 (표준편차 σ)에 의하여 확률분포가 정해짐.

• 정규분포의 특징

— 평균 M 을 중심으로 좌우 대칭이며, 서에서 확률밀도함수의 값이 가장 큽니다.

— 평균 M 에서 멀어질수록 확률밀도함수의 값은 점차 작아집니다.

— 분산 σ^2 이 클수록 확률밀도함수의 폭이 커지며, 두께가� 줄어들게 됩니다.

• 통계적 사실 $\left[\begin{array}{l} \text{평균 } M \text{을 중심으로 } \pm 1\sigma \text{ 안에 들어갈 확률 } 0.6827 \rightarrow P(M-\sigma \leq X \leq M+\sigma) = 0.6827 \\ \text{평균 } M \text{을 중심으로 } \pm 2\sigma \text{ 안에 들어갈 확률 } 0.9545 \rightarrow P(M-2\sigma \leq X \leq M+2\sigma) = 0.9545 \\ \text{평균 } M \text{을 중심으로 } \pm 3\sigma \text{ 안에 들어갈 확률 } 0.9973 \rightarrow P(M-3\sigma \leq X \leq M+3\sigma) = 0.9973 \end{array} \right]$

→ 정규분포의 확률밀도함수를 직접 계산하여 확률을 구하는 것은 힘듭니다.

■ 표준정규분포(standard normal distribution)

• 평균이 0이고 분산이 1인 정규분포 $\rightarrow Z, P(Z \leq z) : \text{누적확률}$

• Z 의 확률밀도함수는 0에 대해 대칭이므로 $P(Z \geq z) = 1 - P(Z \leq z) = P(Z \leq -z) / P(Z \geq 0) = P(Z \leq 0) = 0.5$

• 정규분포의 표준화

— $X \sim N(M, \sigma^2)$ 일 때, $aX + b \sim N(aM + b, a^2\sigma^2)$

— 정규분포를 따른다는 확률변수를 선형 변환하면 그 확률변수는 다시 정규분포를 따릅니다

$$X \sim N(M, \sigma^2) \text{ 일 때, } \frac{X-M}{\sigma} \sim N(0, 1) \rightarrow X \sim N(M, \sigma^2) \text{ 일 때, } \frac{X-M}{\sigma} \text{는 표준정규분포를 따른다}$$

■ 이항분포의 정규근사

• 이항분포 $B(n, p)$ 에서 노이 매우 큰 경우는 직접 확률을 구하는 게 어려움. \rightarrow 노이 매우 크고 p 가 매우 작으면 표준화를 이용하여 확률을 구함.

• 이항분포 $B(n, p)$ 에서 노이 매우 크고 p 가 0이나 1에 가깝지 않아서 (0.5 에 가까워서) np 와 $n(1-p)$ 모두 충분히 큰 경우 (보통 $np \geq 10, n(1-p) \geq 10$) 이항분포는 정규분포에 가까워짐. \rightarrow 노이 퀴질수록 종 모양의 정규분포에 가까워짐.

• 중심극한정리(Central Limit Theorem) : $X \sim B(n, p)$ 이고 $np \geq 10, n(1-p) \geq 10$ 모두 충분히 큰 때, X 는 근사적으로 평균이 np , 분산이 $np(1-p)$ 인 정규분포를 따릅니다.

$$X \sim N(np, np(1-p)) , Z = \frac{X-np}{\sqrt{np(1-p)}} \sim N(0, 1)$$

■ 연속성 수정 (continuity correction)

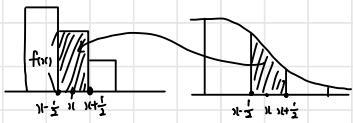
- 이항분포의 정규근사에서 이항분포는 이산형 확률분포이지만 정규분포는 연속형 확률분포

- 이항분포에서 정수 x 에 대하여

$$P(X=x) = P\left(x - \frac{1}{2} \leq X \leq x + \frac{1}{2}\right) \rightarrow \text{연속을 가정한 정규분포}$$

- 이항분포의 히스토그램에서 x 를 중심으로 밸런스의 굴이가 1인 구간 ($x - \frac{1}{2}, x + \frac{1}{2}$)과 높이 $f_{(x)} = P(X=x)$ 로 만들어지는 주사파형의 넓이를 균등으로

정규분포에서 $P\left(x - \frac{1}{2} \leq X \leq x + \frac{1}{2}\right)$ 에 포함된 확률과 같다.



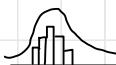
- 이항확률분포를 정규분포로 근사할 때, $P(a \leq X \leq b) = P(a - \frac{1}{2} \leq X \leq b + \frac{1}{2})$ 와 같이 $\frac{1}{2}$ 씩 가감하여 확률의 근사값을 구할 수 있음

- $nPC(-p)$ ($=$ 분산)이 충분히 크면 이항분포의 정규근사에 $\sqrt{nPC(-p)}$ ($=$ 표준편차)가 분포하기 때문에 연속성 수정에 큰 영향을 받지 않으므로 굳이 연속성 수정 X

■ 정규분포에 대한 가정 확인

- 자료를 분석할 때 모집단이 정규분포를 따른다고 가정하는 경우 많음 → 관측한 자료가 정규분포를 따른지 확인하는 방법

- 자료에 대한 히스토그램의 정규분포에 기대값이 확인



→ 그려는 방법에 따라 해석이 달라질 수 있어 주관적 판단

② 정량적 사용 $\rightarrow X \sim N(\mu, \sigma^2)$ 일 때

$$\begin{aligned} P(\mu - 6 \leq X \leq \mu + 6) &= 0.6827 (\bar{x} \pm 1\sigma) \\ P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &= 0.9545 (\bar{x} \pm 2\sigma) \\ P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &= 0.9973 (\bar{x} \pm 3\sigma) \end{aligned}$$

관측값 중에서 $(\bar{x}-3\sigma, \bar{x}+3\sigma), \dots, (\bar{x}-3S, \bar{x}+3S)$ 에 속하는 비율이 유사한지 확인

■ 정규확률그림 (normal probability plot)

- 자료의 백분위 수와 정규분포의 백분위수를 그림을 통해 비교하여 자료가 정규분포를 따른지 확인

- n개의 자료: $X_{(1)} < X_{(2)} < X_{(3)} < \dots < X_{(n)}$ 크기순 배열 \rightarrow 0과 1 사이의 확률을 균등하게 ($n+1$) 등분하는 백분위수들

- n개의 자료가 정규분포를 따른 모집단 $N(\mu, \sigma^2)$ 에서 얻어진 자료라면 모집단 $N(\mu, \sigma^2)$ 을 균등하게 ($n+1$) 등분하는 n개의 점 $a_{(1)}, a_{(2)}, \dots, a_{(n)}$, 즉 $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ 은 유사할 것

- $X \sim N(\mu, \sigma^2)$ 일 때 $P(X \leq a_{(n)}) = P(a_{(1)} \leq X \leq a_{(n)}) = P(a_{(1)} \leq X \leq a_{(2)}) \dots = \frac{1}{n+1}$

\Rightarrow $\frac{1}{n+1}, N(\mu, \sigma^2)$ 을 ($n+1$) 등분하는 n개의 점 $a_{(1)}, a_{(2)}, \dots, a_{(n)}$ 이 대체로 $X_{(1)} \approx a_{(1)}, X_{(2)} \approx a_{(2)}, \dots, X_{(n)} \approx a_{(n)}$ 일 것이다

- 문제는 모집단의 평균 μ 와 분산 σ^2 을 알지 못해 $a_{(n)}$ 를 직접 구할 수 \rightarrow 모평균 \bar{x} 대신 \bar{x} , 모분산 σ^2 대신 S^2 사용 가능

- 대안) 표준정규분포 $N(0,1)$ 을 ($n+1$) 등분하는 n개의 점 $Z_{(1)}, Z_{(2)}, \dots, Z_{(n)}$ → 정규 점수 (normal score)라고 할 때

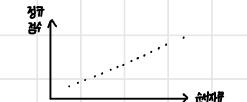
$a_{(n)} \approx \bar{x} + \sigma Z_{(n)}$ 가 심플히므로 $(X_{(1)}, Z_{(1)}), (X_{(2)}, Z_{(2)}), \dots, (X_{(n)}, Z_{(n)})$ 은 좌표평면에 점으로 나타내면 직선이 기대워야 함

• 정규확률그림 그리는법

- 자료: X_1, X_2, \dots, X_n 을 크기 순으로 배열 ($X_{(1)} < X_{(2)} < \dots < X_{(n)}$)

- 표준정규분포 ($n+1$) 등분하는 n개의 점 $Z_{(1)} < Z_{(2)} < \dots < Z_{(n)}$

- $(X_{(1)}, Z_{(1)}), (X_{(2)}, Z_{(2)}), \dots, (X_{(n)}, Z_{(n)})$ 을 좌표평면에 점으로 나타남



정규확률이 직선에 기대워면
 \Rightarrow 모집단이 정규분포를 따른다고 할 수 있다.

- 원자료가 정규분포를 따르지 않더라도 변환을 하면 정규분포를 따른 수 있음 $\rightarrow x, x^2, \sqrt{x}, \sqrt[4]{x} (\cdot x^{\frac{1}{4}}), \log x, \frac{1}{x}$ 등

9장. 표집분포

• 표본평균의 평균(기대값)은 모평균 M 과 일치하지만, 표본평균의 분산은 크기 n 에 반비례하여 감소

$$\rightarrow \bar{x} \text{의 표준편차} = \frac{\sigma}{\sqrt{n}}$$

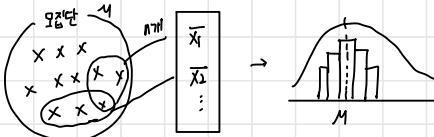
\Rightarrow 모집단이 정규분포 $N(M, \sigma^2)$ 을 따를 때, \bar{x} 의 추출한 표본의 평균 \bar{x} 는 $N(M, \frac{\sigma^2}{n})$ 을 따른다.

통계학 (statistics)

- 목표: 관측한 자료(모집단의 일부)를 이용하여 모집단에 대한 추측을 하는 것
- 모수(parameter): 모집단의 특성을 나타내는 수치 값 \rightarrow 모평균(M), 모분산(σ^2), 모표준편차(σ), 모비율(p)
- 추론(inference): 자료 표본을 이용하여 모집단의 모수에 대한 추측을 하는 과정
- 통계량(statistic): 표본(관측된 자료)에 의해서 결정되는 값 \rightarrow 표본평균(\bar{x}), 표본분산(S^2), 표본표준편차(S), 표본비율(p)
 - 통계량을 이용하여 모수를 추측할 때 통계량이 모수와 일치한다는 보장 X ($\text{ex: } \bar{x} \neq M$)
 - 통계량은 추출된 표본에 따라 달라짐 \rightarrow 통계량은 확률변수로서 확률분포를 가짐.
 - 표집분포(sampling distribution): 통계량의 확률분포 \rightarrow 모집단의 분포 / 표본의 크기에 영향을 끂음.

표본평균의 분포

- 임의 표본(random sample) X_1, X_2, \dots, X_n : 평균이 M 이고 분산이 σ^2 인 모집단에서 임은 임의표본 (iid: independent & identically distributed)
- 표본 평균 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- 표본 평균의 기대값 $E(\bar{X})$: $E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n} \{E(X_1) + E(X_2) + \dots + E(X_n)\} = \frac{1}{n} \times nM = M$
- 표본 평균의 분산 $\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{1}{n^2} \text{Var}(X_1 + X_2 + \dots + X_n) = \frac{1}{n^2} \{ \text{Var}(X_1) + \dots + \text{Var}(X_n) \} = \frac{1}{n} \times n\sigma^2 = \frac{\sigma^2}{n}$
- 표본 평균의 표준편차 $s_d(\bar{X}) = \sqrt{\text{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}$
- \bar{X} 의 확률 분포는 모집단의 확률 분포 (X_1, X_2 각각의 확률분포) 와 비교할 때 평균은 같고, 분산은 작다(중앙에 더 몰려있다)



(x)
① 모집단의 분포가 정규분포라면, \bar{x} 의 분포는 정규분포

중심극한정리

② 모집단의 분포를 몰라도 샘플을 무작위로 뽑았고, $n \geq 30$ 이면 \bar{x} 의 분포는 정규분포

중심극한정리 (central limit theorem, C.L.T.)

- 평균이 M 이고 분산이 σ^2 인 모집단에서 임의추출한 표본의 크기가 충분히 크면(보통 $n \geq 30$), 표본평균은 극사적으로 정규분포를 따름.

$$Z = \frac{\bar{X} - M}{\sigma/\sqrt{n}} \sim N(0, 1)$$

이항분포의 정규근사

거의 0.5

- $X \sim B(n, p)$ 에서 ① n 이 충분히 크고, ② $p \neq 0, p \neq 1 \rightarrow np \geq 10, n(1-p) \geq 10$ 이면 정규분포에 가까움
- 이 때 각각의 X_1, \dots, X_n 은 $X_i \sim B(1, p)$ 이라고 생각할 수 있으며 $E(X_i) = p$, $\text{Var}(X_i) = p(1-p)$ 이므로

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\sum X_i - np}{\sqrt{np(1-p)}} \sim N(0, 1)$$

- 표본의 크기가 커질수록 정규분포에 가까워짐.