통계 기초

머신러닝을 위한 통계 기초 개념 및 용어 알아보기

강명호

통계의 기본 개념

기술 통계

확률과 분포

추정과 가설검정

상관분석

개요

통계학은 데이터를 수집, 분석, 해석 및 표현하는 과학으로서, 이를 통해 불확실성을 다루고 데이터 기반 의사 결정을 지원하는 학문임

목적

- 데이터로부터 유의미한 정보 도출
- 현상 및 패턴 이해
- 연관성 파악
- 예측 및 추론
- 불확실성의 해소
- 의사결정 지원

중요성

- 데이터를 수치화하여 신뢰성을 부여함
- 의사 결정을 위한 근거 자료를 제시함
- 현상을 분석하여 실증 자료를 제시함

개요

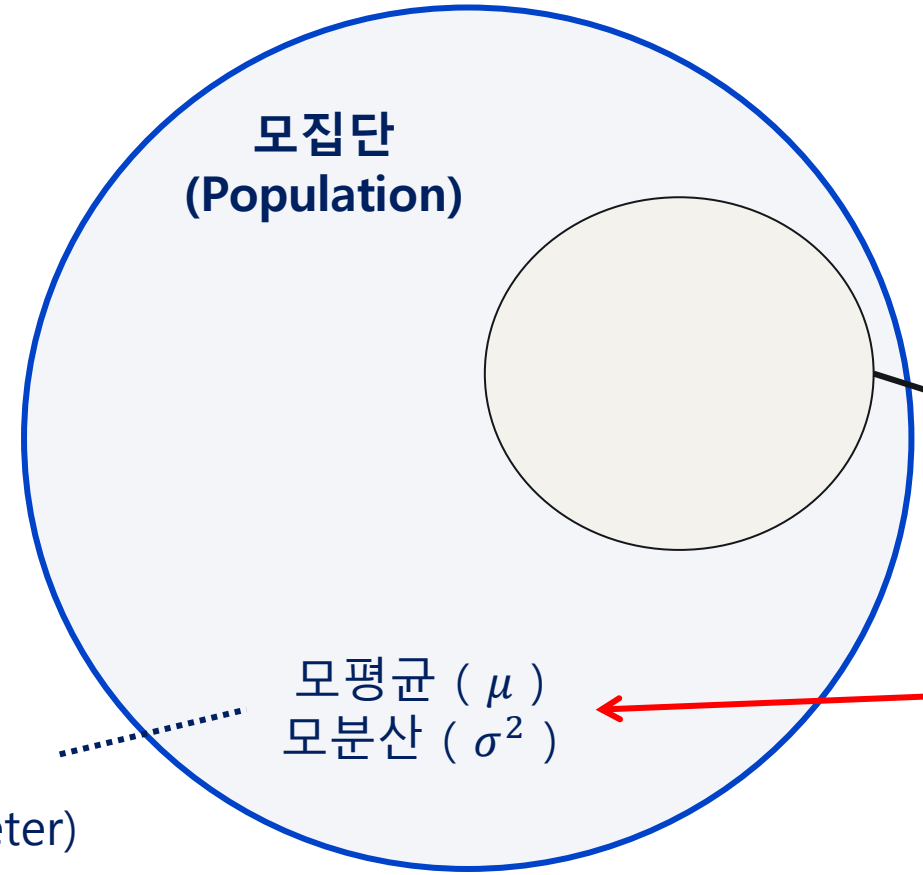
통계학은 데이터를 수집, 분석, 해석 및 표현하는 과학으로서, 이를 통해 불확실성을 다루고 데이터 기반 의사 결정을 지원하는 학문임

정의

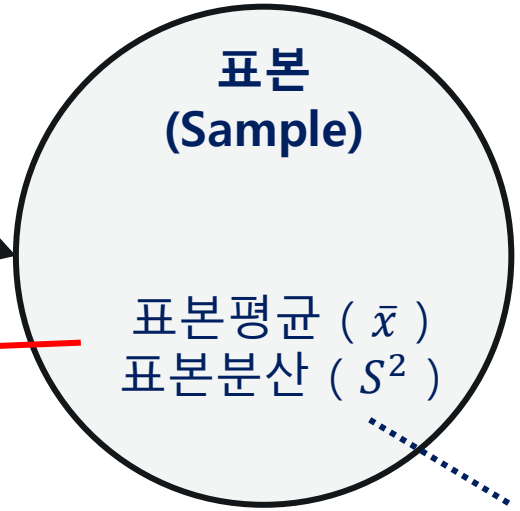
많은 현상과 사실들을 관찰하고 이를 수치화하여 비교하는 방법을 연구하는 학문

기술 통계	추론 통계
• 데이터에 대한 분석 결과인 수치들을 활용하여 데이터 집합의 특성을 설명	• 모집단에서 추출한 표본을 통해 모집단의 특성을 표현
• 평균값, 분산, 표준편차, 범위 • 히스토그램, 파이차트, 상자도표	• 지방선거 출구조사 • 제품 불량율 조사

용어 및 개념 : 모집단과 표본



표본 추출
(Sampling)



모집단 추정
(Estimation)

모수
(parameter)

: 모집단의 분포를 결정하는 정보

통계량
(statistic)

- 전수조사 : 모집단 전체를 조사하는 방법
- 표본조사 : 표본을 추출하여 현상을 관측, 자료 수집하는 방법

표본의 추출

- **확률적 표본 추출 (Probability sampling method)**
  - 동일한 확률 하에서 표본을 추출하는 방법
  - 무작위 표본 추출
    - 난수표 등을 따라 모집단에서 표본을 기계적으로 추출하는 방법
  - 체계적 표본 추출
    - 모집단에서 특정한 규칙으로 표본을 추출하는 방법
  - 층화 표본 추출
    - 모집단을 특정 특성에 따라 여러 하위 집단으로 구분한 후, 집단의 규모에 비례하도록 추출하는 방법
- **비확률적 표본 추출 (Non-probability sampling method)**
  - 조사자가 자의로 표본을 추출하거나 조사 대상이 자발적으로 표본에 참여하는 방법

통계의 기본 개념

기술 통계

확률과 분포

추정과 가설검정

상관분석

중심경향성 측정

- 중심경향성(Central Tendency) :
  - 표본 내의 원소들의 중심을 나타내는 지표.
  - 평균, 중앙값, 최빈값 등 다양한 방식으로 표현됨

- **평균(Mean)** : 모든 값을 더한 후 데이터의 개수로 나눈 값
  - 장점: 데이터를 통합하여 하나의 대표값으로 표현.
  - 단점: 이상치(outliers)에 민감하여 극단적인 값에 영향을 받음.
  - 연속적이고 정규 분포를 따르는 데이터에서 대표값을 구할 때 유용
  - 예시: {1, 2, 3, 4, 5}의 평균은 (1+2+3+4+5) / 5 = 3.

중심경향성 측정

- 중심경향성(Central Tendency) :
  - 표본 내의 원소들의 중심을 나타내는 지표.
  - 평균, 중앙값, 최빈값 등 다양한 방식으로 표현됨

- **중앙값 (Median)** : 표본 내의 원소들을 크기 순서대로 나열했을 때 중앙에 위치한 값
  - 장점: 이상치의 영향을 덜 받음.
  - 단점: 표본의 크기가 클 경우 정렬에 많은 시간 소요.
  - 예시: {1, 2, 3, 4, 5}의 중앙값은 3. (짝수 개의 데이터인 경우, 두 중앙값의 평균을 사용)
  - 비대칭적이거나 이상치가 있을 때 사용

중심경향성 측정

- 중심경향성(Central Tendency) :
  - 표본 내의 원소들의 중심을 나타내는 지표.
  - 평균, 중앙값, 최빈값 등 다양한 방식으로 표현됨

- **최빈값 (Mode)** : 표본에서 가장 자주 나타나는 값.
  - 장점: 표본의 원소들의 빈도 분포를 잘 나타냄.
  - 단점: 표본 내에 최빈값이 존재하지 않을 수 있으며, 복수 개의 최빈값이 있을 수도 있음.
  - 예시: {1, 2, 2, 3, 4}의 최빈값은 2.
  - 범주형 데이터에서 빈도가 가장 높은 값을 찾을 때 유용

중심경향성 측정

- 중심경향성(Central Tendency) :
  - 표본 내의 원소들의 중심을 나타내는 지표.
  - 평균, 중앙값, 최빈값 등 다양한 방식으로 표현됨

- **가중평균 (Weighted Mean)** : 표본의 원소들의 값에 가중치를 부여하여 계산한 평균.
  - 모든 원소들의 값이 동일한 중요도를 가지지 않을 때 사용.
  - 예시: {1, 2, 3}와 각각의 가중치 {0.1, 0.3, 0.6}이 있을 때
    가중평균은 (1*0.1 + 2*0.3 + 3*0.6) / (0.1 + 0.3 + 0.6)
  - 예시: 5천원짜리 상품 8개와 9천원짜리 상품 2개를 구입했을 때 평균 구매값은
    (5000 * 8 + 9000 * 2) / (8 + 2) = 5800

중심경향성 측정

- 중심경향성(Central Tendency) :
  - 표본 내의 원소들의 중심을 나타내는 지표.
  - 평균, 중앙값, 최빈값 등 다양한 방식으로 표현됨

- **조화평균 (Harmonic Mean)** : 원소들의 값들의 역수의 평균을 다시 역수로 변환하여 계산.
  - 주로 비율이나 속도 데이터에서 사용.
  - 예시: {1, 2, 3}의 조화평균은 3 / (1/1 + 1/2 + 1/3) = 약 1.636.

$$\frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}} = 1.636\ldots$$

산포도 측정

- 산포도 (Measure of Dispersion) :
  - 통계에서 산포도를 측정하는 방법들은 데이터의 변동성을 파악하는 데 중요한 역할을 함

- **범위(Range)**
  - 표본 집합에서 가장 큰 값과 가장 작은 값의 차이.
  - 단순해서 많이 사용되지만, 극단값(outlier)에 민감함.

$$Range = max - min$$

산포도 측정

- 산포도 (Measure of Dispersion) :
  - 통계에서 산포도를 측정하는 방법들은 데이터의 변동성을 파악하는 데 중요한 역할을 함

- **분산 (Variance)**
  - 데이터 값들이 평균에서 얼마나 떨어져 있는지를 나타냄
  - 계산 방법: 각 데이터 값에서 평균을 뺀 값을 제곱한 후, 이 값들의 평균을 구함.

$$\sigma^2 = \frac{1}{N} \sum_{N}^{1} (x_i - \mu)^2$$

산포도 측정

- 산포도 (Measure of Dispersion) :
  - 통계에서 산포도를 측정하는 방법들은 데이터의 변동성을 파악하는 데 중요한 역할을 함

- **표준편차(Standard Deviation)**
  - 분산의 제곱근으로, 데이터가 평균에서 얼마나 떨어져 있는지에 대한 표준적인 거리를 나타냄.

$$\sigma = \sqrt{\frac{1}{N} \sum_{N}^{1} (x_i - \mu)^2}$$

산포도 측정

- 산포도 (Measure of Dispersion) :
  - 통계에서 산포도를 측정하는 방법들은 데이터의 변동성을 파악하는 데 중요한 역할을 함

- **사분위수 범위(Interquartile Range, IQR)**
  - 데이터의 중앙 50%가 포함된 범위로, Q3(3사분위수)와 Q1(1사분위수)의 차이
  - 극단값의 영향을 덜 받음

$$IQR = (Q3 - Q1)$$

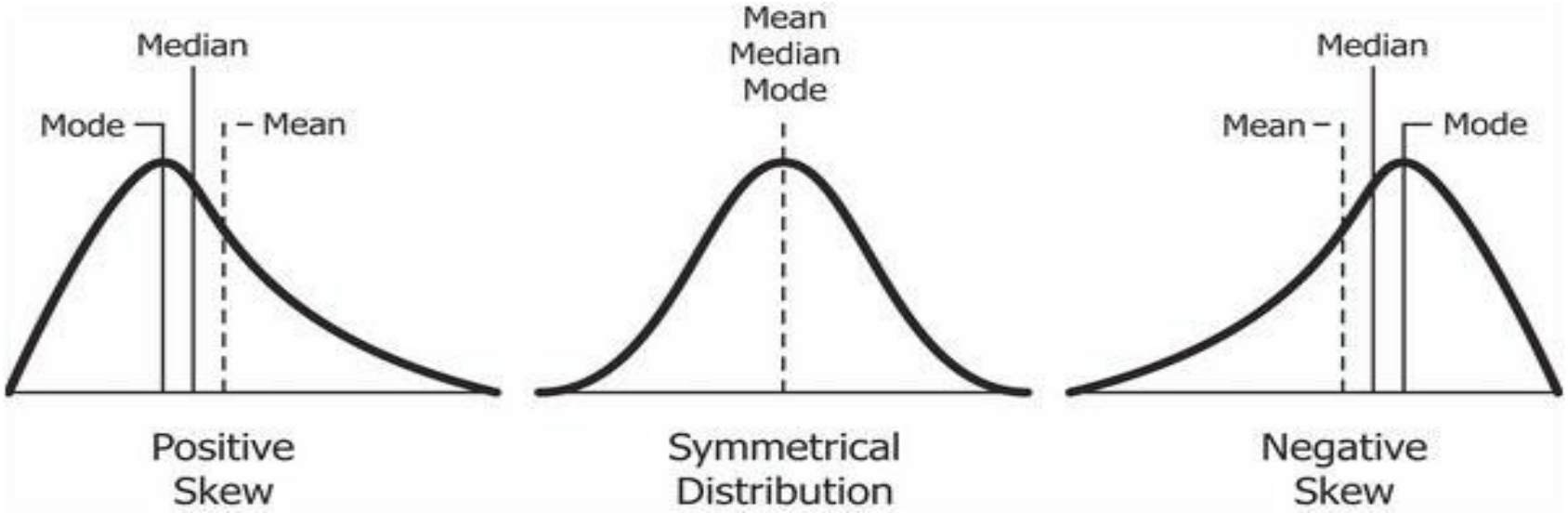
왜도와 첨도

- **왜도와 첨도** : 통계에서 데이터 분포의 특성을 설명하는 지표

  - 왜도 (Skewness) : 분포의 비대칭성을 측정하는 지표
  - 첨도 (Kurtosis) : 분포의 뾰족함과 꼬리의 두꺼움을 측정하는 지표

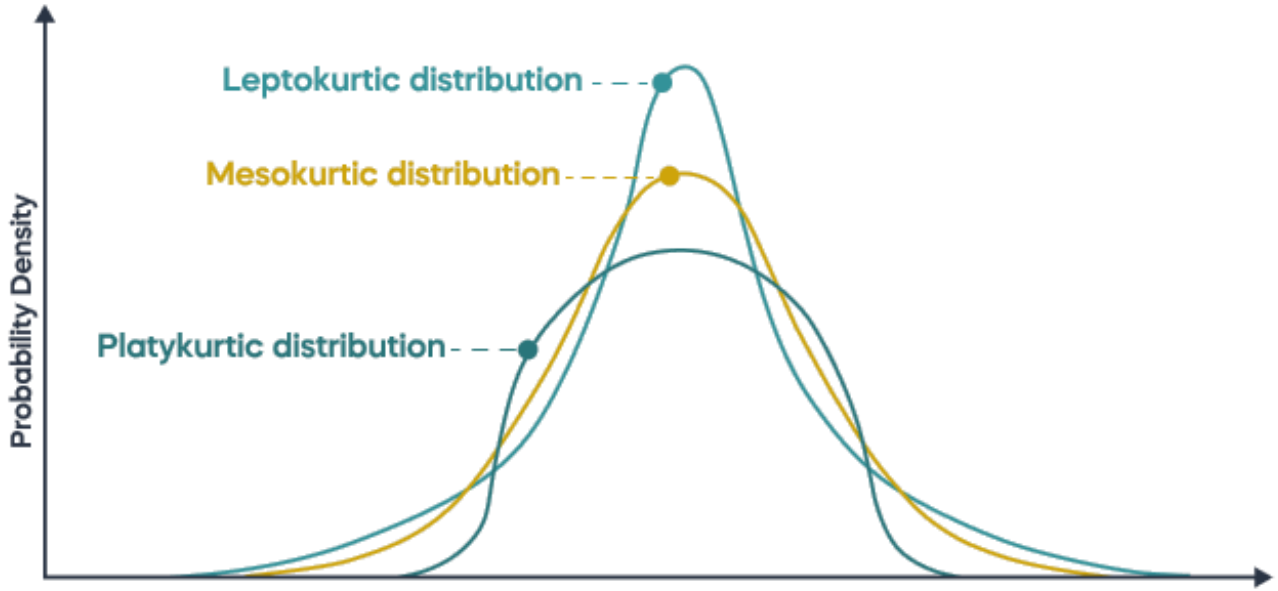
왜도와 첨도

- **왜도 (Skewness)** : 분포의 비대칭성을 측정하는 지표
  - 양의 왜도 (Positive Skewness):
    - 분포의 꼬리가 오른쪽(양의 방향)으로 길게 늘어져 있는 경우
    - 평균이 중앙값보다 큼. (예: 급여, 주택 가격 등)
  - 음의 왜도 (Negative Skewness):
    - 분포의 꼬리가 왼쪽(음의 방향)으로 길게 늘어져 있는 경우
    - 평균이 중앙값보다 작음. (예: 시험 점수 등)



왜도와 첨도

- **첨도 (Kurtosis)** : 분포의 뾰족함과 꼬리의 두꺼움을 측정하는 지표
  - 양의 첨도 (Leptokurtic): (Kurtosis > 3)
    - 정규 분포보다 뾰족하고 꼬리가 두꺼운 분포
    - 극단적인 값(이상치)이 더 자주 발생함
  - 음의 첨도 (Platykurtic): (Kurtosis < 3)
    - 정규 분포보다 평평하고 꼬리가 얇은 분포
    - 극단적인 값이 덜 발생



https://365datascience.com/calculators/kurtosis-calculator/

시각화의 필요성

시각화는 데이터의 패턴을 파악하고 데이터가 의미하는 내용을 효과적으로 전달하며, 신속하고 정확한 의사 결정을 위한, 기술 통계 방법의 일종임

데이터 이해

데이터의 시각화를 통해 데이터의 의미를 직관적으로 이해할 수 있으며, 패턴, 관계, 추세 등을 빠르게 파악할 수 있음

의사 소통

시각화를 통해 통계 분석 결과를 다른 사람에게 효과적으로 설명할 수 있으며, 데이터를 통한 의사 소통을 원활하게 함

패턴 파악

데이터 시각화는 데이터의 특이한 패턴을 파악하는데 도움을 주며, 수치 자료가 표현하지 못하는 이상치를 쉽게 파악하게 함

의사결정 지원

데이터의 의미와 패턴을 효과적으로 나타내는 시각화는 신속하고 정확한 의사결정을 위한 통찰을 제공함

통계의 기본 개념

기술 통계

확률과 분포

추정과 가설검정

상관분석

확률

- **표본 공간 (Sample Space)** : 실험을 통해 나타날 수 있는 모든 결과들의 집합
- **사건 (Event)** : 표본공간에 있는 일부 원소들로 이루어진 부분 집합
- **확률** : 특정 사건이 일어날 가능성을 나타내는 척도(사건이 발생할 가능성을 수치로 표현)
  - 0에서 1 사이의 값을 가지며, 0은 사건이 절대로 일어나지 않음을 의미하며,
    1은 사건이 반드시 일어남을 의미함
- **확률 함수 (Probability Function)** : 사건이 발생할 확률을 나타내는 함수
- **규칙**
  - 덧셈 법칙 : 두 사건 A, B 에 대하여, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - 곱셈 법칙 : 두 사건 A, B 가 독립 사건일 때, $P(A \cap B) = P(A) \times P(B)$

확률 변수와 확률 분포

- **확률 변수(Random Variable) :**
  - 정의 : 특정 값이 나타날 가능성이 확률적으로 주어지는 변수
    - 표본 공간의 각 표본점에 실수 값을 대응시키는 함수 역할을 함.
  - 이산 확률변수와 연속 확률변수로 구분

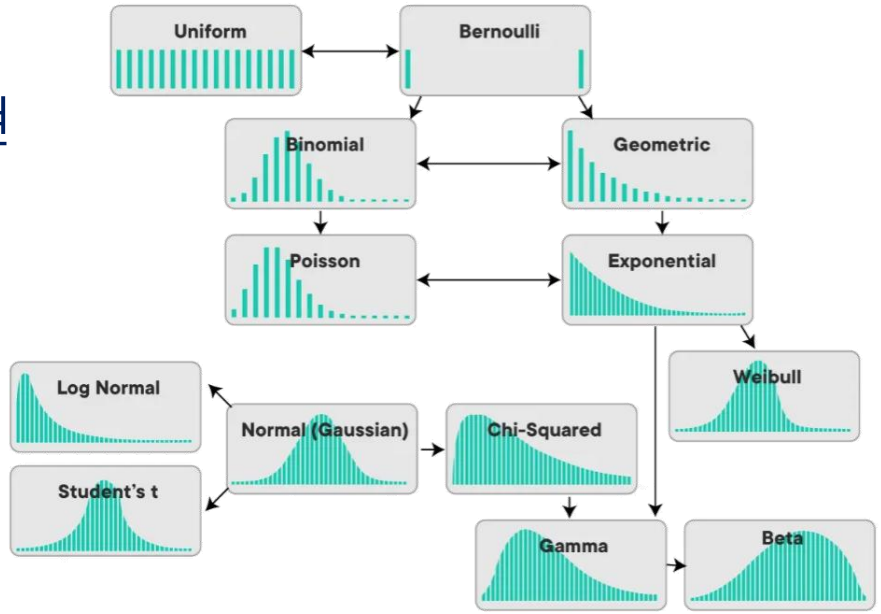
- 표본점 (Sample point) :
  - 모집단에서 무작위로 뽑은 하나의 표본
- 확률 표본 (Random sample) :
  - 모든 표본점(sample point)들이 동일한 확률로 추출된다는 조건 하에서 추출된 표본

확률 변수와 확률 분포

- **이산형 확률변수 (Discrete random variable)**
  - 0이 아닌 확률 값을 갖는 값이 셀 수 있는 경우의 확률 변수
  - 예: 주사위를 굴렸을 때 나오는 눈의 수, 동전 던지기의 결과.
  - **확률질량함수(Probability Mass Function)** : 이산형 확률변수의 확률함수
  - 이산형 확률변수에 의한 확률 분포 : 이항분포, 기하분포, 포아송분포 등
- **연속형 확률변수 (Continuous random variable)**
  - 가능한 값이 실수의 어느 특정 구간 전체에 해당하는 확률 변수
  - 연속적인 값을 가짐.
  - 예: 특정 시간 동안의 온도, 사람의 키
  - **확률밀도함수(Probability Density Function)** : 연속형 확률변수의 확률함수
  - 연속형 확률변수에 의한 확률 분포 : 정규분포, 균일분포, t-분포, 카이제곱분포 등

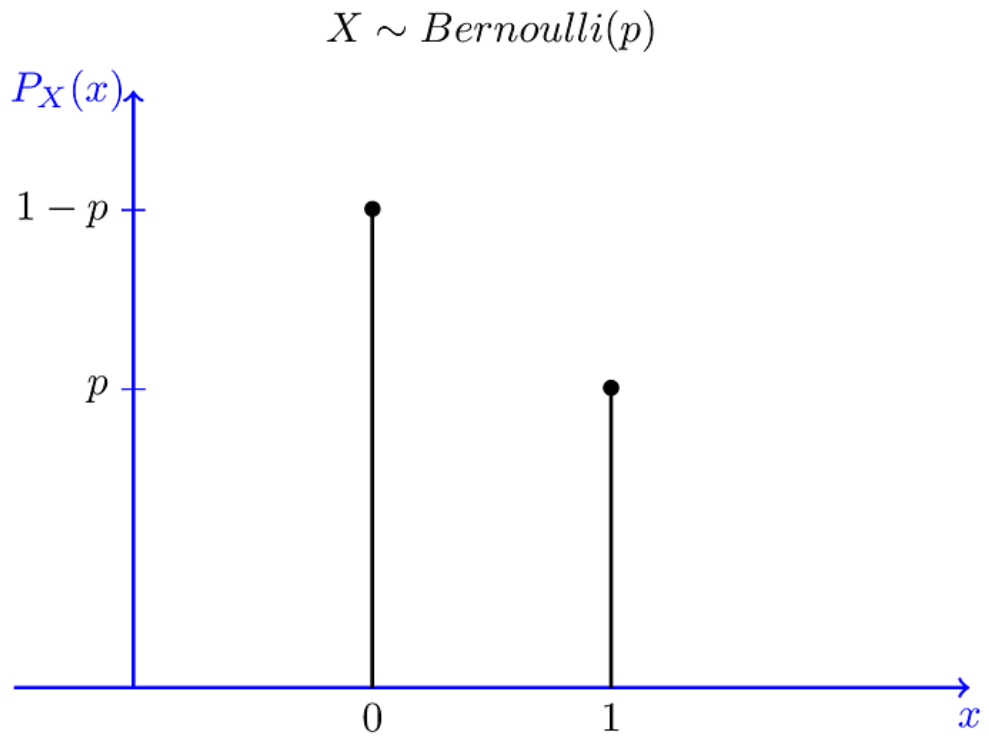
확률 분포

- **정의** : 확률 변수가 가질 수 있는 값들과 그 값들이 발생할 확률을 나타내는 함수, 표, 그래프
- **이산 확률 분포 (Discret Probability Distribution)**
  - 확률 변수가 취할 수 있는 값이 유한하거나 셀 수 있는 경우
  - 확률질량함수(Probability Mass Function : PMF)로 표현
  - 예 : 베르누이분포, 이항분포, 포아송분포
- **연속 확률 분포 (Continuous Probability Distribution)**
  - 확률 변수가 취할 수 있는 값이 연속적인 경우.
  - 확률밀도함수(Probability Density Function : PDF)로 표현
  - 예: 정규 분포, 지수 분포, 카이제곱 분포.



확률 분포 – 이산 확률 분포

- **베르누이 분포 (Bernoulli Distribution) :**
  - 두 가지 결과(성공 또는 실패)가 있는 단일 시행의 확률분포
  - 확률변수 X : 성공일때 1, 실패일때 0의 값
  - 확률질량함수 (PMF) : ($0 \leq p \leq 1$ 는 성공 확률)
    - $P(X = x) = \begin{cases} p, & if\ x = 1 \\ 1 - p, & if\ x = 0 \end{cases}$
  - 기대값 : $E(X) = p$
  - 예시 :
    - 동전 던지기(앞면/뒷면)
    - 제품 검수(불량/정상)



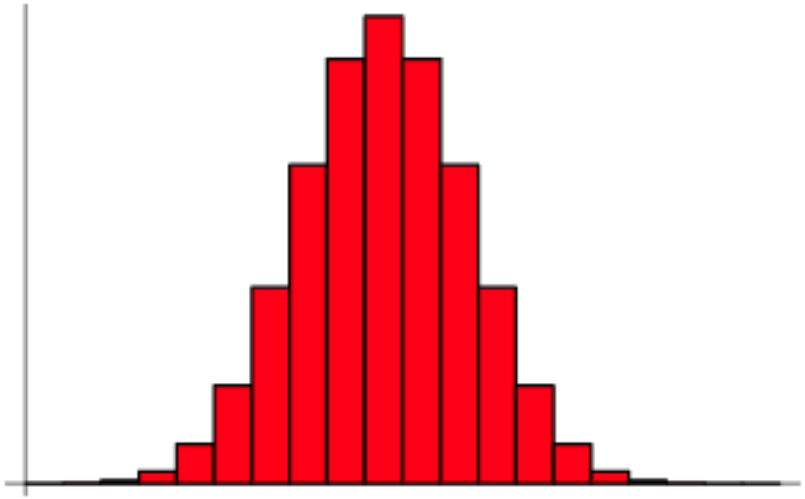
https://medium.com/mytake/understanding-different-types-of-distributions-you-will-encounter-as-a-data-scientist-27ea4c375eec

확률 분포 – 이산 확률 분포

- **이항 분포 (Binomial Distribution):**
  - 일정한 횟수의 독립적인 베르누이 시행에서 성공의 횟수를 나타내는 분포
  - 확률변수 X : n 번의 독립적인 시행 중 성공 횟수
  - 확률질량함수 (PMF) : k 번 발생할 확률
    - $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0,1,2,\dots,n$
  - 예시 :
    - 동전 여러 번 던지기 :

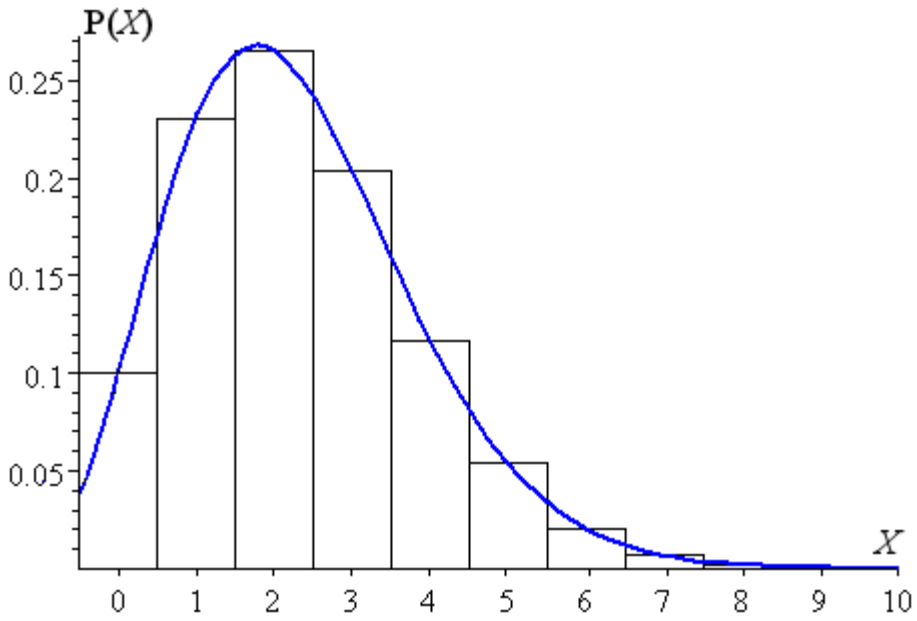
      동전을 20번 던져서 앞면이 나오는 횟수.

- 기대값 : $E(X) = np$
- 분산 : $Var(X) = np(1-p)$



확률 분포 – 이산 확률 분포

- **푸아송 분포 (Poisson Distribution):**
  - 단위 시간(또는 단위 공간)에서 특정 사건의 발생 빈도를 나타내는 분포
    - 드물게 발생하는 사건의 빈도를 모델링할 경우 사용됨
  - 확률변수 X : 일정한 시간에서 발생하는 사건의 횟수
  - 확률질량함수 (PMF) : (λ 는 단위 시간에서 사건이 발생하는 평균 횟수)
    - $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0,1,2, \ldots$
  - 예시 :
    - 콜센터 : 1시간 동안 걸려오는 전화의 수
    - 교통량 : 특정 도로에서 1시간 동안 발생하는 교통사고 수

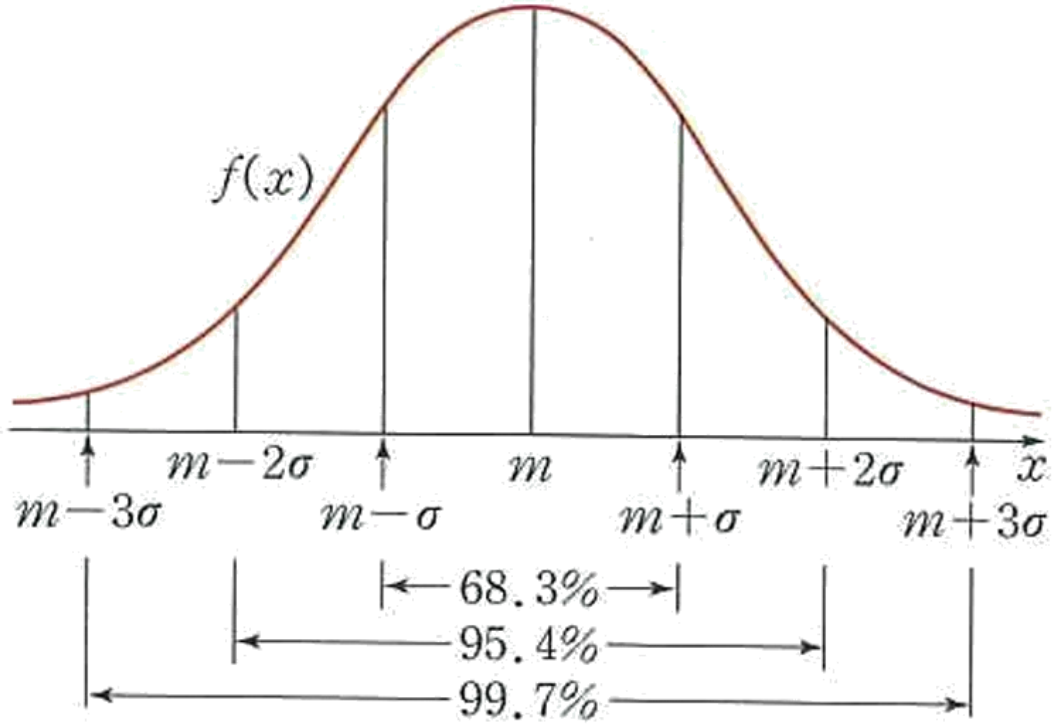


확률 분포 – 연속 확률 분포

- **정규 분포 (Normal Distribution) :**
  - 종 모양의 대칭 분포로, 자연 현상에서 많이 나타나는 일반적인 분포
  - 확률밀도함수 (PDF) : (μ 는 평균, σ 는 표준편차를 의미)

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

  - 특성 :
    - 평균 μ 를 중심으로 대칭
    - 확률변수의 68%가 $\mu \pm \sigma$, 95%가 $\mu \pm 2\sigma$, 99.7%가 $\mu \pm 3\sigma$ 범위 내에 위치함
  - 예시 :
    - 키 : 성인의 키 분포
    - 시험정수 : 대규모 시험의 성적 분포

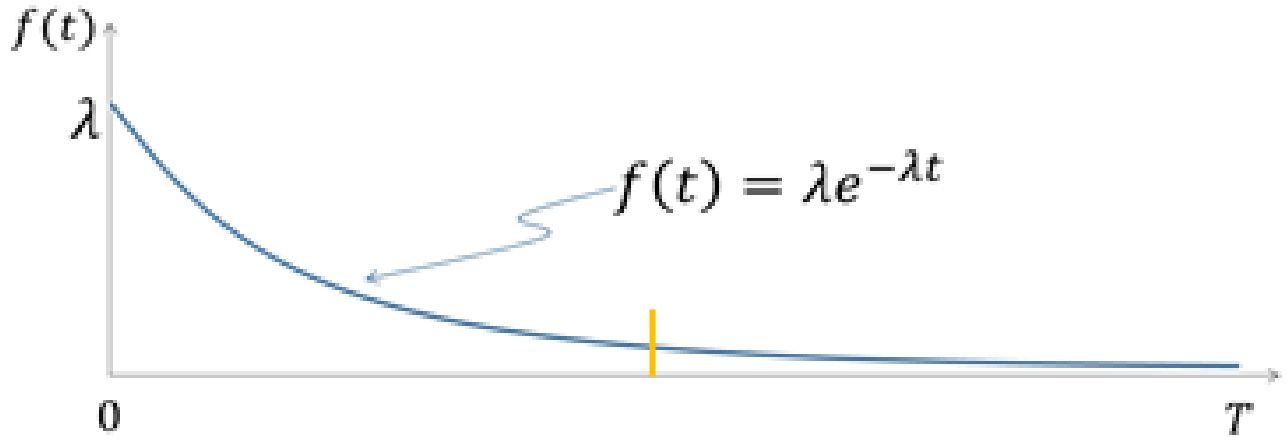


확률 분포 – 연속 확률 분포

- **지수 분포 (Exponential Distribution) :**
  - 사건 간의 시간 간격에 대한 분포. 주로 포아송 과정에서 시간 간격을 모델링하는 데 사용
  - 확률밀도함수 (PDF) : (λ 는 사건 발생률)

$$f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0$$

  - 예시 :
    - 콜센터 대기 시간 : 다음 전화가 걸려올 때까지의 시간.

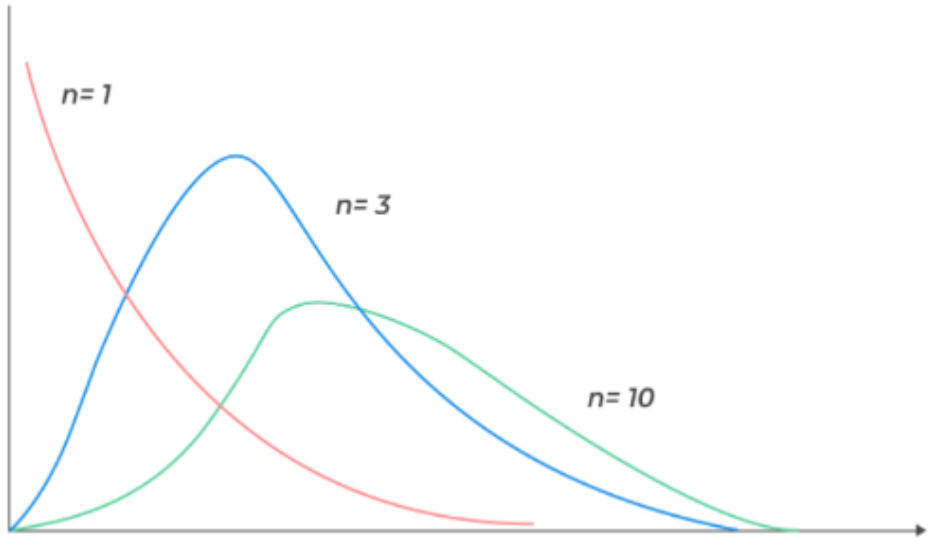


확률 분포 – 연속 확률 분포

- **카이제곱 분포 (Chi-Square Distribution)**
  - 정규 분포를 따르는 독립적인 변수들의 제곱의 합의 분포
  - 모집단의 **분산을 추정**할 때 주로 사용됨
  - 확률밀도함수 (PDF) : (k 는 자유도, Γ 는 감마함수)

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, \qquad x \geq 0$$

  - 예시 :
    - **적합도 검정** : 관찰된 데이터가 기대 분포와 일치하는지 검정
    - 분산 분석 : 여러 집단의 분산을 비교



통계의 기본 개념

기술 통계

확률과 분포

추정과 가설검정

상관분석

추론

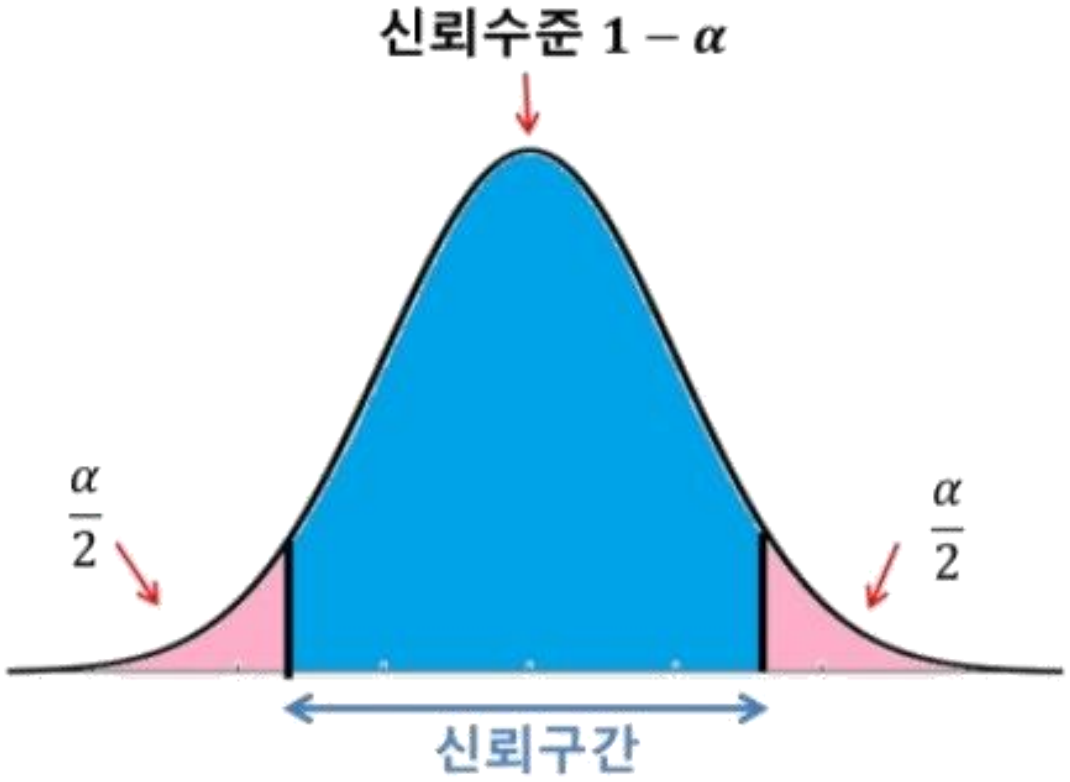
- **추론**
  - 표본을 활용하여 모집단의 특성을 추측함
  - 모집단 전체를 조사할 수 없을 경우 주로 사용됨
- **추론의 방법**
  - 모수의 **추정 (Estimation) :**
    - 미지수인, 모집단의 모수에 대한 추측 또는 추측값을 정확도와 함께 제시함
  - 모수에 대한 **가설검정 (Hypothesis Testing) :**
    - 모집단의 모수에 대한 여러 가설들이 적합한지 여부를 표본으로부터 판단함
  - 예시 : (직장인들의 연말 상여금 조사 예시)
    - 표본조사를 통해 전국 직장인 상여금의 평균을 하나의 값으로 추정
    - 표본조사를 통해 전국 직장인 상여금의 평균이 포함될만한 구간을 정함
    - 표본조사를 통해 전국 직장인 상여금 평균이 5년전 평균값과 얼마다 다른지 판단함

추정

- **점추정 (Point Estimation)**
  - 정의 : 모집단의 모수를 단일 값(점)으로 추정하는 방법
  - 대푯값 : 일반적으로 표본평균, 표본분산 등이 사용됨
  - 예시 : 표본평균이 50이라면 모집단 평균도 50으로 추정함
- **구간추정 (Interval Estimation)**
  - 정의 : 모집단의 모수를 포함할 것으로 예상되는 구간을 제시하는 방법
  - 구성 : 신뢰구간(Confidence Interval)과 신뢰수준(Confidence Level)으로 구성
    - **신뢰구간** : 모수가 포함될 것으로 예상되는 범위
    - **신뢰수준** : 모수가 해당 구간에 포함될 확률
  - 예시 : 95% 신뢰수준에서 신뢰구간이 [25, 55]라면 관심있는 모수가 25~55 사이에 있을 확률이 95%라는 의미

신뢰구간, 신뢰수준

- **신뢰구간 (Confidence Interval)** : 모수가 신뢰구간 안에 포함될 것으로 예상되는 범위
  - 표본을 사용하여 계산하며 구간이 넓을 수록 모수를 포함할 확률이 높아짐
- **신뢰수준 (Confidence Level)** : 모수가 신뢰구간에 포함될 확률
  - $P(a \leq \mu \leq b) = 1 - \alpha$
  - 90%, 95%, 99%가 자주 사용됨
  - **유의수준** α :
    - 모수가 신뢰구간에 포함되지 않을 확률



가설 검정

- **가설 검정** : 어떤 추측이나 주장, 가설에 대해 타당성을 조사하는 작업
  - 표본 통계량으로 모수 추정 시, 추정한 모수값이나 확률분포 등이 타당한지 평가하는 통계적 추론 방법
  - **귀무가설** (Null Hypothesis, H0) : 버릴 것으로 예상하는 가설
  - **대립가설** (Alternative Hypothesis, H1) : 실제 주장 또는 증명하려는 가설

- **가설 검정 단계**

가설 검정 예시

(진통제 개발의 예시)

- **귀무가설/대립가설 수립** :
  - 귀무가설 H0 : 새로운 진통제의 효과가 기존 진통제와 차이가 없다.
  - 대립가설 H1 : 새로운 진통제가 기존 진통제보다 더 효과적이다.

- **유의수준 결정 :**
  - 유의수준을 0.05로 설정

    (5%의 확률로 귀무가설이 참인데도 불구하고 기각할 가능성을 허용함)

가설 검정 예시

(진통제 개발의 예시)

- **검정 통계량 계산** :
  - 임상 실험을 통해 두 그룹의 환자들을 각각 새로운 진통제와 기존 진통제 투약
  - 각 그룹에서 진통 정도를 수치화한 데이터 수집
  - 두 그룹의 평균 진통 수치를 비교하여 검정 통계량(예: t-검정) 계산
- **기각/채택 결정 :**
  - P-value 계산 : 검정 통계량을 바탕으로 P-value를 계산
    - **P-value** : 귀무가설 하에서 관측된 데이터가 발생할 확률 (H0을 지지하는 값)
  - 결정 :
    - P-value가 유의수준 0.05보다 작으면 귀무가설 기각 (효과 있다)
    - 그렇지 않으면 귀무가설 채택 (효과 없다)

가설 검정 - 오류

- **1종 오류 (Type I Error)** :
  - 정의 : 귀무가설이 참인데도 불구하고 이를 기각하는 오류
  - 예시 : 새로운 진통제와 기존 진통제의 효과에 차이가 없는데, 새로운 진통제가 더 효과적이라고 결론 내리는 경우
  - 결과 : **잘못된 긍정(False Positive)** 결과 초래
- **2종 오류 (Type II Error)**
  - 정의 : 대립가설이 참인데도 불구하고 귀무가설을 기각하지 않는 오류
  - 예시 : 새로운 진통제가 기존 진통제보다 효과가 있음에도 불구하고, 두 진통제의 효과 차이가 없다고 결론 내리는 경우
  - 결과 : **잘못된 부정(False Negative)** 결과 초래

실제값

	Positive	Negative
Positive	TP	**FP**
Negative	**FN**	TN

추론 결과

가설 검정 방법

검정 방법	목적	비교 대상	예시
t-검정 (t-test)	두 그룹의 평균 비교	두 그룹 또는 한 그룹의 두 상황	두 학급 간 평균 성적 비교
카이제곱 검정 (Chi-square test)	데이터 간 독립성 또는 적합도 검정	두 범주형 변수 또는 한 변수	성별과 흡연여부 간의 독립성 확인
분산 분석 (ANOVA)	세 그룹 이상의 평균 비교	세 그룹 이상	여러 학급 간 평균 성적 비교

가설 검정 방법 – t 검정

- **t-검정(t-test)** :
  - 두 그룹 간의 평균의 차이가 유의미한지 확인하고자 할 때 주로 사용

검정 방법	목적	비교 대상	예시
단일 표본 t-검정 (One-sample t-test)	특정 값과의 비교	한 그룹	한 학급의 평균 성적이 70점인지 확인
독립 표본 t-검정 (Independent two-sample t-test)	두 그룹의 평균 비교	두 그룹	남학생과 여학생의 평균 성적 비교
대응 표본 t-검정 (Paired sample t-test)	같은 그룹의 두 상황 비교	한 그룹의 두 상황	다이어트 전후의 체중 비교

가설 검정 방법 – 카이제곱 검정

- **카이제곱 검정(Chi-square test)** :
  - 범주형 데이터에서 기대 빈도와 관찰된 빈도 간의 차이를 확인할 때 사용
  - 두 범주형 데이터에서 변수 간 독립성 여부를 검정할 때 사용

검정 방법	목적	비교 대상	예시
적합도 검정 (Goodness of fit test)	관찰 빈도와 기대 빈도의 일치 여부	하나의 범주형 변수	주사위 굴리기 결과의 기대와 일치 확인
독립성 검정 (test of independence)	두 변수 간의 독립성 확인	두 범주형 변수	성별과 흡연 여부간의 독립성 확인
동질성 검정 (test of homogeneity)	여러 표본이 동일한 분포를 따르는지 확인	두개 이상의 범주형 변수	여러 지역의 질병 발생률 비교

가설 검정 방법 – 분산 분석

- **분산 분석(ANOVA : Analysis Of VAriance)** :
  - 3개 이상의 집단에 대한 평균 차이를 검증하는 분석 방법

검정 방법	목적	비교 대상	예시
일원 분산 분석 (One-way ANOVA)	한 요인에 대한 평균 비교	여러 그룹	여러 학급 간 평균 성적 비교
이원 분산 분석 (Two-way ANOVA)	두 요인에 대한 평균 비교	여러 그룹	학급과 성별에 따른 평균 성적 비교

통계의 기본 개념

기술 통계

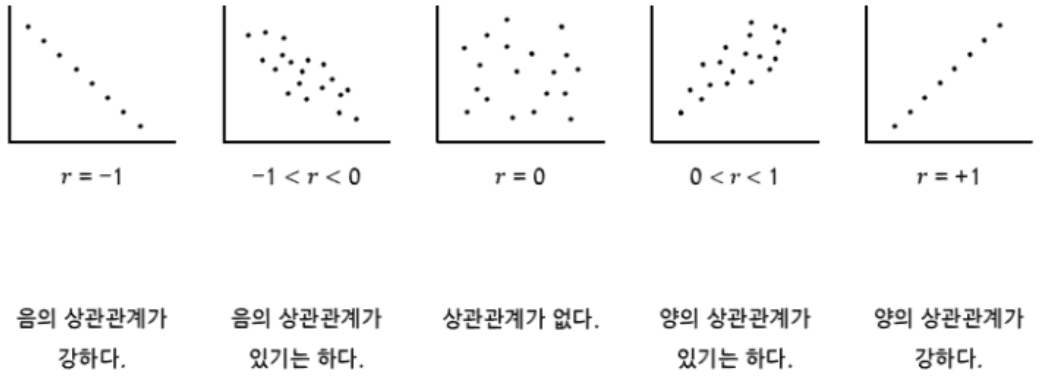
확률과 분포

추정과 가설검정

상관분석

상관분석 - 개요

- **상관분석**
  - 두 변수 간의 관계를 정략적으로 평가하는 통계 기법
  - 하나의 변수가 변할 때 다른 변수가 어떻게 변하는지 파악할 수 있음.
- **상관계수 (Correlation Coefficient)**
  - 두 변수 간의 관계의 강도와 방향을 나타내는 수치
- **상관관계의 방향**
  - 양의 상관관계 : 한 변수가 증가할 때 다른 변수도 증가하는 경향이 있는 경우
  - 음의 상관관계 : 한 변수가 증가할 때 다른 변수는 감소하는 경향이 있는 경우
- **상관관계의 강도**
  - 강한 상관관계 : 상관계수의 절댓값이 0.7 이상
  - 중간 상관관계 : 상관계수의 절댓값이 0.3~0.7 사이
  - 약한 상관관계 : 상관계수의 절댓값이 0.3 미만



상관분석

- **피어슨 상관계수 (Pearson Correlation Coefficient)**
  - 측정 대상 : 연속형 변수 간의 선형 관계를 측정
  - 계산 방법 : 두 변수의 공분산을 각 변수의 표준편차로 나누어 계산
    - 즉, 두 변수가 얼마나 함께 변하는지를 표준편차를 통해 정규화한 값
  - 해석
    - $r = 1$: 완전한 양의 상관관계
    - $r = -1$: 완전한 음의 상관관계
    - $r = 0$: 상관관계 없음
  - 장점 : 데이터의 선형 관계를 직접적으로 반영하며, 해석이 직관적
  - 단점 : 두 변수 간의 관계가 비선형 관계일 경우 이를 제대로 반영하지 못함
  - 예시: 키와 몸무게 간의 관계.

    (일반적으로 키가 큰 사람이 몸무게도 더 나가는 경향이 있을 때 사용)

상관분석

- **스피어만 상관계수 (Spearman's Rank Correlation Coefficient)**
  - 측정 대상 : 순위형 변수 간의 관계 측정 (비선형 관계일 경우에도 사용 가능)
  - 계산 방법 : 각 변수의 순위를 매긴 후, 그 순위들 간의 상관계수를 계산함
    (순위 차이에 따라 값을 계산하므로, 변수 간의 비선형 관계 반영 가능)
  - 해석 :
    - $r_s = 1$: 완전한 양의 상관관계
    - $r_s = -1$: 완전한 음의 상관관계
    - $r_s = 0$: 상관관계 없음
  - 장점 : 비선형 관계와 이상치에 민감하지 않음
  - 단점 : 데이터의 순위 정보만 사용하여 정보 손실이 발생할 수 있음 (원래 데이터의 크기 정보 등)
  - 예시: 학생 성적 순위와 스포츠 성적 순위 간의 관계.
    (학생들의 성적과 스포트 성적 간의 관계 분석 시 점수를 순위로 변경하여 상관관계 분석)

상관분석

- **켄달의 타우 (Kendall's Tau, Kendall's Rank Correlation Coefficient)**
  - 측정 대상 : 변수의 순서 간의 상관성을 측정 (작은 데이터셋에 적합)
  - 계산 방법 : 순위 쌍 간의 일치와 불일치를 비교하여 상관관계를 계산
  - 해석 : 피어슨 상관계수와 동일하게 해석함 (τ)
  - 장점 : 순위 정보에 기반하여 비선형 관계도 잘 반영
  - 단점 : 계산이 복잡하며, 해석이 어려울 수 있음 (특히 데이터셋이 클 경우 계산 부담 증가)
  - 예시 : 직무 수행 순위와 승진 순위 간의 관계

	사람 A	사람 B	사람 C	사람 D	사람 E
키	1	2	3	4	5
몸무게	3	4	1	2	5

$\tau = 0.2$

	사람 A	사람 B	사람 C	사람 D	사람 E
키	1	2	3	4	5
몸무게	1	2	3	4	5

$\tau = 1$

	사람 A	사람 B	사람 C	사람 D	사람 E
키	1	2	3	4	5
몸무게	5	4	3	2	1

$\tau = -1$

https://medium.com/@leejukyung/%EC%BC%84%EB%8B%AC%ED%83%80%EC%9A%B0-kendalltau-18fb90ba4e7