통계학 치트시트

기술 통계 (Descriptive Statistics)

데이터의 특성을 요약하고 설명하는 방법.

- 중심 경향 (Central Tendency):
  - `평균 (Mean)`: 모든 값의 합을 값의 개수로 나눈 값. 이상치에 민감.
  - `중앙값 (Median)`: 데이터를 순서대로 나열했을 때 가운데에 위치하는 값. 이상치에 덜 민감.
  - `최빈값 (Mode)`: 가장 빈번하게 나타나는 값.
- 산포도 (Dispersion):
  - `범위 (Range)`: 최대값 - 최소값.
  - `분산 (Variance)`: 각 값이 평균으로부터 얼마나 떨어져 있는지 나타내는 지표. 편차 제곱의 평균.
  - `표준편차 (Standard Deviation)`: 분산의 제곱근. 데이터의 변동성을 직관적으로 이해하기 쉬움.
  - `사분위수 (Quartiles)`: 데이터를 4등분하는 값 (Q1, Q2, Q3). `IQR = Q3 - Q1`.

추론 통계 (Inferential Statistics)

표본 데이터를 사용하여 모집단에 대한 결론을 도출하는 방법.

- 가설 검정 (Hypothesis Testing):
  - `귀무가설 (Null Hypothesis, H0)`: 기존에 사실로 받아들여지는 주장. (예: 차이가 없다, 효과가 없다)
  - `대립가설 (Alternative Hypothesis, H1)`: 연구자가 입증하고자 하는 주장. (예: 차이가 있다, 효과가 있다)
  - `p-value`: 귀무가설이 맞다고 가정할 때, 관측된 데이터 또는 더 극단적인 데이터가 나타날 확률.
  - `유의수준 (Significance Level, α)`: p-value를 판단하는 기준. 보통 0.05, 0.01 등을 사용. `p < α` 이면 귀무가설을 기각.
- 신뢰 구간 (Confidence Interval): 모집단의 모수(예: 평균)가 포함될 것으로 신뢰하는 구간. (예: 95% 신뢰구간)

주요 확률 분포

- 정규 분포 (Normal Distribution): 평균을 중심으로 대칭적인 종 모양의 분포. 자연 및 사회 현상에서 흔히 나타남.
- 이항 분포 (Binomial Distribution): 성공/실패와 같이 두 가지 결과만 있는 베르누이 시행을 여러 번 반복했을 때의 성공 횟수 분포.
- 포아송 분포 (Poisson Distribution): 특정 시간 또는 공간 단위 내에서 어떤 사건이 발생하는 횟수에 대한 분포.
- t-분포 (Student's t-Distribution): 정규 분포와 유사하지만, 표본 크기가 작을 때 사용.
- 카이제곱 분포 (Chi-squared Distribution): 범주형 데이터 분석(교차 분석)에 사용.

주요 통계 검정 방법

- t-검정 (t-test): 두 집단의 평균을 비교.
  - `독립 표본 t-검정 (Independent Samples t-test)`: 서로 다른 두 집단 비교. (예: A반과 B반의 성적)
  - `대응 표본 t-검정 (Paired Samples t-test)`: 동일한 집단의 사전-사후 비교. (예: 약물 투여 전후의 혈압)
- 분산 분석 (ANOVA, Analysis of Variance): 셋 이상의 집단의 평균을 비교.
- 카이제곱 검정 (Chi-squared Test): 범주형 변수 간의 연관성(독립성)을 검정. (예: 흡연 여부와 폐암 발병률)
- 상관 분석 (Correlation Analysis): 두 연속형 변수 간의 선형 관계의 강도와 방향을 측정.
  - `피어슨 상관 계수 (Pearson Correlation Coefficient, r)`: −1에서 1 사이의 값을 가짐.

회귀 분석 (Regression Analysis)

하나 이상의 독립 변수(X)를 사용하여 종속 변수(Y)를 예측하는 모델을 만드는 방법.

- 단순 선형 회귀 (Simple Linear Regression): `Y = β0 + β1*X + ε`. 하나의 독립 변수.
- 다중 선형 회귀 (Multiple Linear Regression): `Y = β0 + β1*X1 + β2*X2 + ... + ε`. 여러 개의 독립 변수.
- 로지스틱 회귀 (Logistic Regression): 종속 변수가 범주형(예: 합격/불합격)일 때 사용.

통계적 오류

- 1종 오류 (Type I Error, α): 귀무가설이 사실인데 기각하는 오류. (False Positive)
- 2종 오류 (Type II Error, β): 귀무가설이 거짓인데 기각하지 못하는 오류. (False Negative)