2-1 통계학 개론

제2장 통계분석

제1절 통계학 개론

통계 분석 개요

통계

- 특정집단을 대상으로 수행한 `조사`나 `실험`을 통해 나온 `결과`에 대한 `요약된 형태`
  - ex) 일기예보, 물가/실업률/GNP, 의식조사와 사회조사 분석 통계, 임상실험 통계
- `조사` 또는 `실험`을 통해 `데이터 확보`
- `조사 대상`에 따라 `총조사(census)`와 `표본조사`로 구분

통계자료의 획득 방법

- `총 조사/전수 조사(Census)`
  - 대상 `집단 모두`를 조사하는데 `많은 비용`과 `시간`이 소요되므로 특별한 경우를 제외하곤 사용하지 않음
    - ex) 인구주택 총 조사

- `표본조사`
  - 대부분의 설문조사가 표본조사로 진행
  - `모집단`에서 `샘플을 추출`하여 진행
  - `모집단(Population)` : 조사하고자 하는 `대상 집단 전체`
  - `원소(Element)` : `모집단을 구성`하는 `개체`
  - `표본(Sample)` : 조사하기 위해 `추출한 모집단의 일부 원소`

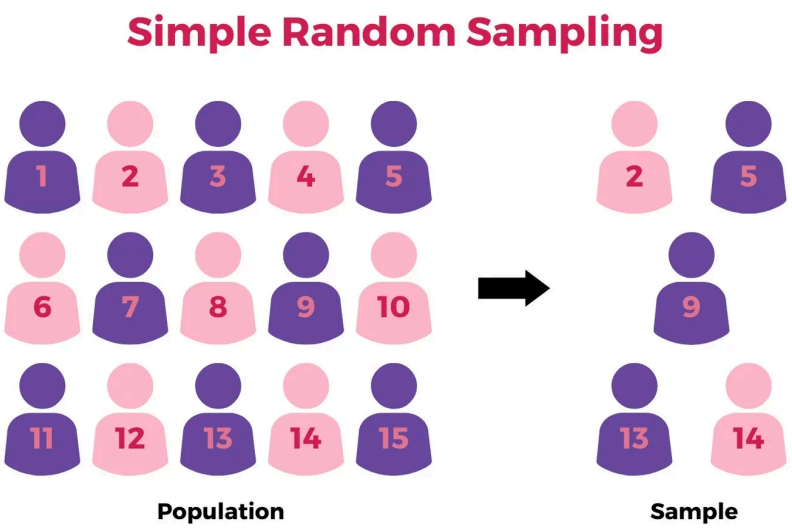
    - `모수(Parameter)` : 표본 관측에 의해 구하고자 하는 `모집단에 대한 정보(모집단의 특성)`
    - `모집단의 정의` , `표본 크기` , `조사 방법` , `조사기간` , `표본추출방법` 을 정확히 명시해야 함

- 표본오차 : 모집단의 일부인 표본에서 얻은 자료를 통해 모집단 전체의 특성을 추론함으로써 생기는 오차
    - 모집단을 대표할 수 있는 표본단위들이 조사대상으로 추출되지 못하면 발생
- 비표본오차 : 표본오차를 제외한 조사의 전체 과정에서 발생할 수 있는 모든 오차
- 표본편의 : 표본추출방법에서 기인하는 오차
    - 표본추출이 의도된 모집단의 일부 구성원이 다른 구성원보다 더 낮거나 더 높은 표본 추출 확률을 갖는 오차

표본 추출 방법 설명

1. 단순랜덤추출법 (Simple Random Sampling)

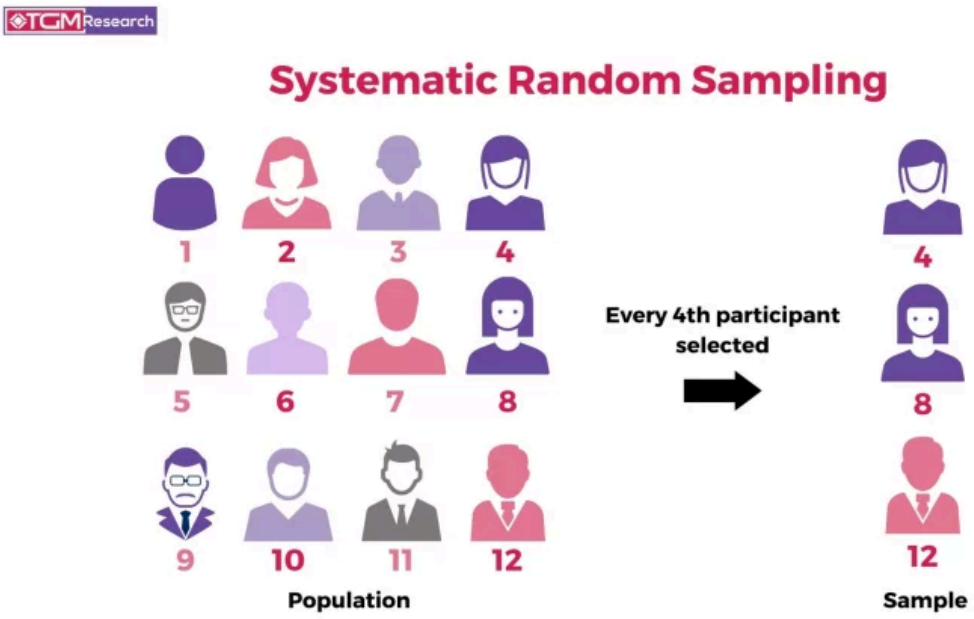
- 모집단의 모든 구성원에게 동일한 확률로 표본으로 선택될 기회를 부여하는 방식
    - 각 샘플에 번호를 부여해 임의의 n개를 추출하는 방법
    - 각 샘플은 선택될 확률이 동일
    - 비복원, 복원(추출한 원소를 다시 집어넣어 추출하는 경우) 추출
    - 예: 100명의 학생 중 무작위로 10명을 추첨기로 뽑는 방식
- 가장 기본적이고 공정하지만, 모집단이 크면 실행이 어려움.



출처 https://tgmresearch.com/simple-random-sampling.html

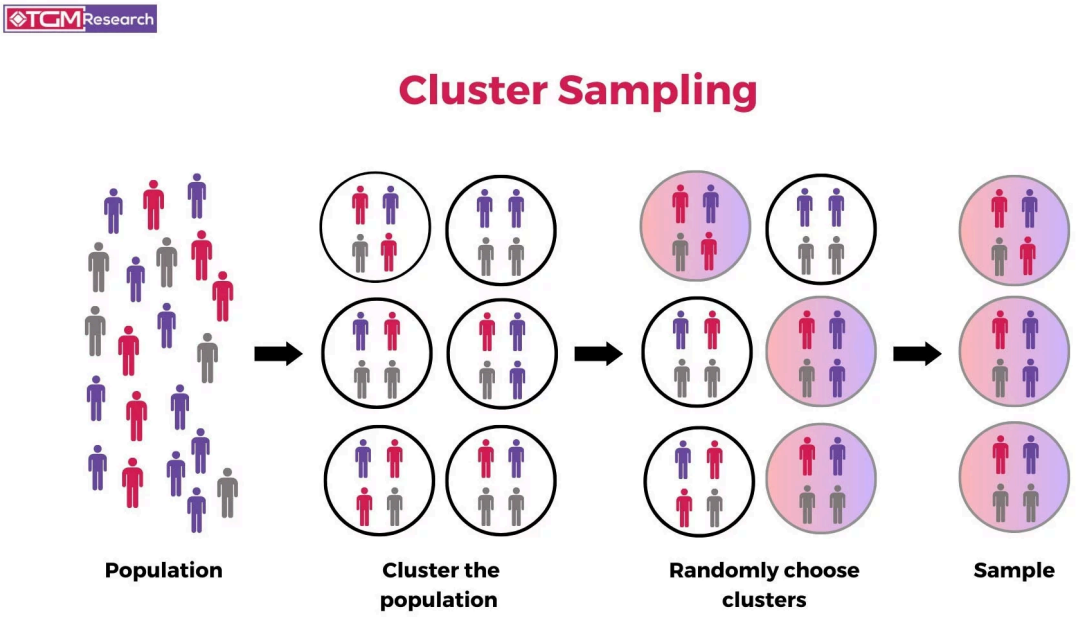
2. 계통추출법 (Systematic Sampling)

- 일정 간격으로 표본을 선택하는 방식, 단순랜덤추출법의 변형된 방식
  - 단순랜덤추출법의 변형된 방식
  - 번호를 부여한 샘플을 나열하여 K개씩 (K=N/n) n개의 구간으로 나누고 첫 구간(1, 2, … , K)에서 하나를 임의로 선택한 후에 K개씩 띄어서 n개의 표본 선택
  - 예: 1,000명 명단 중 첫 번째 사람을 무작위로 고르고 이후 10명마다 한 명씩 뽑음.
- 단순랜덤보다 효율적이지만, 주기성과 편향이 있는 경우 위험함



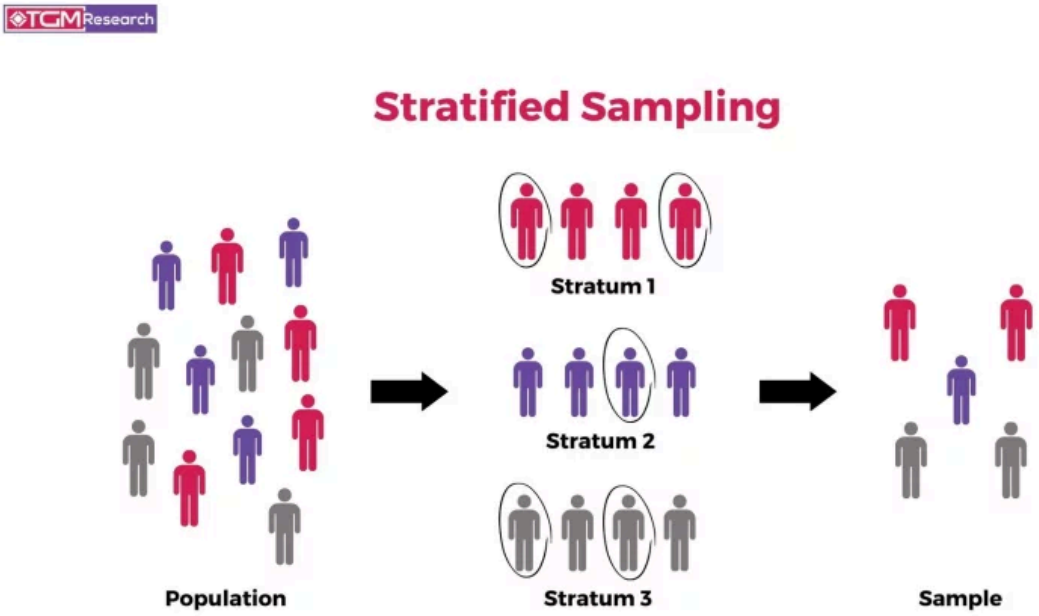
3. 집락추출법 (Cluster Sampling)

- 모집단을 여러 '집락(Cluster)'으로 나누고, 그 중 일부 집락을 랜덤으로 선택하여 전체를 조사하는 방식
  - 예: 학교를 집락으로 보고 몇 개 학교만 골라서 전체 학생을 조사
- 시간과 비용은 절약되지만, 집락 간 동질성이 낮으면 대표성이 떨어질 수 있음



4. 층화추출법 (Stratified Sampling)

- 모집단을 동질적인 층(Stratum)으로 나누고, 각 층에서 랜덤 추출하는 방식
  - 이질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법
  - 유사한 원소끼리 몇 개의 층(Stratum)으로 나누어 각 층에서 랜덤 추출
  - 예: 성별, 학년별로 나누고 각각에서 표본을 무작위로 추출
- 계층 간의 대표성을 확보하기 좋아서 정확도가 높음



표본추출 방법 비교

추출법	특징	장점	단점
단순랜덤	모든 구성원이 같은 확률	가장 공정	모집단 클 경우 어렵다
계통추출	주기적으로 선택	단순하고 빠름	주기적 패턴이 있으면 위험
집락추출	집단(특성이 유사) 단위로 선택	비용 절감	대표성 떨어질 수 있음
층화추출	층(특성이 다름)별로 나누고 선택	정밀도 높음	층 구분이 어려울 수 있음

자료의 종류

측정(Measurement)

표본조사나 실험을 실시하는 과정에서 추출된 원소들이나 실험 단위로부터 주어진 목적에 적합하도록 관측해 자료를 얻는 것

측정 방법

구분	구분	설명	유형 구분
질적척도(범주형 또는 이산형)	**명목척도**	측정 대상이 어느 집단에 속하는지 분류할 때 사용(성별, 출생지 구분)	범주형 자료, 숫자들의 크기 차이가 계산되지 않는 척도
질적척도(범주형 또는 이산형)	**순서척도**	측정 대상의 서열관계를 관측하는 척도(만족도, 우호도, 학년, 신용등급)	
양적척도(수치형 또는 연속형)	**구간척도**(등간척도)	측정 대상이 갖고 있는 속성의 양을 측정구간이나 구간 사이의 간격이 의미가 있는 자료(온도, 지수)	수치형 자료, 숫자들의 크기와 차이를 계산할 수 있는 척도
양적척도(수치형 또는 연속형)	**비율척도**	간격(차이)에 대한 비율의 의미를 가지는 자료절대적 기준인 0의 존재, 사칙연산 가능하며 제일 많은 정보를 가지는 척도(무게, 나이, 시간, 거리)	

- 서열척도는 명목척도와 달리 매겨진 숫자의 크기를 의미있게 활용 가능
  - ex) 1등이 2등보다 성적이 높다
- 구간척도는 절대적 크기를 측정할 수 없기 때문에 사칙연산중 더하기/빼기는 가능하나 비율처럼 곱하기/나누기는 불가능

통계분석

특정한 집단이나 불확실한 현상을 대상으로 자료를 수집해 대상 집단에 대한 정보를 구하고, 적절한 통계분석 방법을 이용해 의사결정을 하는 과정

기술통계(Descriptive Statistic)

주어진 자료로부터 어떠한 판단이나 예측과 같은 주관이 섞일 수 있는 과정을 배제하여 통계 집단들의 여러 특성을 수량화하여 객관적인 데이터로 나타내는 통계분석 방법론

- Sample에 대한 특성인 평균, 표준편차, 중위수, 최빈값, 그래프, 왜도, 첨도 등을 구하는 것

통계적 추론(추측통계, Inference Statistics)

수집된 자료를 이용해 대상 집단(모집단)에 대한 의사결정을 하는 것

- Sample을 통해 모집단을 추정하는 것

모수추정()

표본집단으로부터 모집단의 특성인 모수(평균/분산 등)를 분석하여 모집단 추론

가설검정

대상집단에 대해 특정한 가설을 설정한 후

- 그 가설이 옳은지 그른지에 대한 채택여부를 결정하는 방법론

예측

미래의 불확실성을 해결해 효율적인 의사결정을 하기 위해 활용

- 회귀분석, 시계열분석 등

추정과 가설검정

- 확률표본(Random Sample): 확률분포는 분포를 결정하는 평균, 분산 등의 모수(Parameter)를 가짐
  - 특정한 확률분포로부터 독립적으로 반복해 표본을 추출
  - 각 관찰값들은 서로 독립적이며 동일한 분포를 가짐
- 추정: 표본으로부터 미지의 모수를 추측하는 것
  - 점추정과 구간추정으로 구분
- 점추정(Point Estimation): 모수가 특정한 값일 것이라고 추정
  - 표본의 평균, 중위수, 최빈값 등을 사용

표본평균과 표본 분산

- 표본평균과 표본분산은 통계학의 기본적인 개념으로, 모집단이 아닌 일부 표본으로부터 평균과 분산을 추정할 때 사용
- 표본평균은 데이터를 모두 더한 뒤 표본의 개수로 나눈 값이고, 표본분산은 각 값이 평균에서 얼마나 떨어져 있는지를 제곱한 값을 평균한 값
- 표본분산 계산 시 'n'이 아닌 'n-1'로 나누는 이유는 편향을 보정하기 위함

표본평균과 표본분산 정의 및 수식

표본평균 (Sample Mean)

- 표본평균은 주어진 표본 데이터의 중심값을 나타냄.
- 수식:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - 여기서 xi는 각 표본값, n은 표본의 개수
  - 예: 5개의 시험 점수의 평균을 구할 때 각각의 점수를 더해 5로 나눔

표본분산 (Sample Variance)

- 표본분산은 데이터의 흩어짐(산포도)을 측정함.
- 수식:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

구간 추정(Interval Estimation)

점추정의 정확성 보완을 위해 확률로 표현된 믿음의 정도 하에서 모수가 특정한 구간에 있을 것이라고 선언하는 방법

- 항상 추정량의 분포에 대한 전제가 주어져야 하고, 구해진 구간 안에 모수가 있을 가능성의 크기(신뢰수준)이 주어져야 함
- n : 전체 표본수

가설 검정(Hypothesis Testing)

모집단에 대한 어떤 가설을 설정한 뒤에 표본관찰을 통해 그 가설의 채택여부를 결정하는 분석방법

- 표본관찰 또는 실험을 통해 귀무 가설(Null Hypothesis, $H0$)과 대립 가설(Alternative Hypothesis, $H1$) 중 하나를 선택하는 과정
- 귀무가설이 옳다는 전제하에 검정통계량 값을 구한 후에 이 값이 나타날 가능성의 크기에 의해 귀무가설 채택여부 결정

주요 용어

- 귀무가설(Null Hypothesis, $H0$)
  : 비교하는 값과 차이가 없다, 동일하다를 기본 개념으로 하는 가설
- 대립가설(Alternative Hypothesis, $H1$)
  : 뚜렷한 증거가 있을 때 주장하는 가설
- 검정통계량(Test Statistics)
  : 관찰된 표본으로부터 구하는 통계량으로 검정 시 가설의 진위 판단 기준
- 유의수준(Significance Level, α)
  : 귀무가설을 기각하게 되는 확률의 크기로 귀무가설이 옳은데도 이를 기각하는 확률의 크기
- 기각역(Critical Region, C)
  : 귀무가설이 옳다는 전제 하에서 구한 검정통계량의 분포에서 확률이 유의수준 α인 부분
  - 반대 : 채택역(Acceptance Region)
- 유의확률(p-value)
  : 귀무가설이 맞다고 가정할 때 얻을 수 있는 결과보다 실제값이 더 극단에 위치할 확률

- 검정력(Statistical Power)
  : 대립가설이 사실일 때, 대립가설을 채택하는 옳은 결정을 할 확률

1. 기각역 (Critical Region, C)

- 정의: **귀무가설(H_0)을 기각하는 구간**
- 예: z-test에서 유의수준 0.05일 때, z < -1.96 또는 z > 1.96이 기각역
- 기준: 유의수준(α)에 의해 **사전에 정해짐**

2. p-value (유의확률)

- 정의: **귀무가설이 참일 때, 관측된 통계량보다 극단적인 값이 나올 확률**
- 역할: 실제 관측값이 얼마나 "놀라운지"를 수치로 표현함
- 해석: p-value < α이면 H_0 기각

3. 유의수준 (Significance Level, α)

- 정의: **기각역의 경계 기준**. 즉, H_0가 맞는데도 기각할 확률
- 보통 0.05, 0.01, 0.10을 사용함
- 실험 전에 **미리 설정**함

4. 신뢰구간 (Confidence Interval)

- 정의: **모수(parameter)가 포함될 것으로 기대되는 구간**
- 예: 평균에 대한 95% 신뢰구간은 "진짜 평균이 이 구간 안에 있을 확률이 95%"라는 뜻
- 귀무가설에 대응되는 값이 **신뢰구간 밖에 있으면** H_0를 기각할 수도 있음

올바른 관계 요약

개념	요약 설명	서로의 관계
기각역	통계량이 이 안에 있으면 H_0 기각	α 기준으로 미리 정함
p-value	실제 데이터에서 H_0가 얼마나 말이 안 되는지를 수치로 표현	p < α → 기각
유의수준(α)	기각 기준값 (예: 0.05)	기각역을 정하는 기준

개념	요약 설명	서로의 관계
신뢰구간	모수가 포함될 것으로 예상되는 범위	H_0의 값이 포함되지 않으면 기각 가능

귀무가설(歸無假說) 이해

- '귀무가설(歸無假說)'에서 **'귀무(歸無)'는 문자 그대로 "없음으로 돌아간다"**는 뜻
  - 즉, 어떤 변화나 효과가 **'없다'는 전제로 출발**하는 기본 가설
- '무(無)'를 기본값(default)으로 간주하고, 특별한 증거가 없는 한 받아들이는 보수적 입장이라 보면 된다.

'귀무(歸無)'의 의미

한자 풀이

- **'귀(歸)'**: 돌아갈 귀 → "돌아간다"
- **'무(無)'**: 없을 무 → "없다"

  → **'귀무(歸無)'란? '없음으로 돌아간다'** → 즉, "특별한 것이 없다는 쪽으로 회귀한다"는 뜻

통계학적 해석

- 통계에서는 **현상이나 효과가 '없다'는 가정을 먼저 설정**함
  - 예: "약이 효과 없다", "성별에 따라 차이 없다"
- 이를 **기본 상태(baseline)**로 삼고,

  → 표본 데이터를 통해 **이 가설을 뒤집을 수 있는가**를 검정함

왜 '귀무'를 먼저 세우는가?

보수적 사고 기반

- 과학적 검증은 기본적으로 **회의주의적 태도**에서 출발함
  - "특별한 증거 없으면, 변화나 차이는 없다"는 관점

- 즉, **변화나 효과를 주장하려면 강한 증거가 필요하다**는 것

논리적 구조

- 귀무가설이 기각되면 → **대립가설이 받아들여짐**
- 반대로, 귀무가설을 기각하지 못하면 → **그냥 그대로 유지**

가설검정의 오류 판단

정확한 사실	**가설검정 결과:** **H_o가 사실이라고 판단**	**가설검정 결과:** **H_o가 사실이 아니라고 판단**
귀무가설 H_o가 사실임	✅ 옳은 결정: (H_o가 옳은데 그대로 둠) → 옳은 결정	❌ **제1종 오류 (α)**
귀무가설 H_o가 사실이 아님	❌ **제2종 오류 (β):** (H_o가 틀렸는데 그대로 둠) → ❌ 오류	✅ 옳은 결정

가설검정결과 / 정확한사실	H0가 사실이라고 판정	H0가 사실이 아니라고 판정
H0가 사실임	옳은 결정	제 1종 오류(α)
H0가 사실이 아님	제2종 오류(β)	옳은 결정

- **제1종 오류(Type I Error)** : 귀무가설 H_o이 옳은데도 귀무가설을 기각하게 되는 오류
- **제2종 오류(Type II Error)** : 귀무가설 H_o가 옳지 않은데도 귀무가설을 채택하게 되는 오류
- 두 가지 오류는 서로 **상충관계**
    - 일반적으로 가설검정에서는 제1종 오류 α의 크기를 0.1, 0.05, 0.01 등으로 고정시킨 후 제2종 오류 β가 최소가 되도록 기각역 설정

오류 종류	통계적 정의	은유적 표현	실제 문제 상황
1종 오류 (α)	참인 H0을 기각함: 귀무가설 H0이 **참인데**, 잘못 판단해서 **기각한 경우**	**괜히 소리 질렀다**	오경보, 오진단, 과잉 반응
2종 오류 (β)	거짓인 H0을 채택함: 귀무가설 H0이 **거짓인데**, 잘못 판단해서 **기각하지 못한 경우**	**소리 안 질러서 당했다**	위험 방치, 병 놓침

상황	귀무가설	대립가설	1종 오류	2종 오류
법정	피고는 무죄	피고는 유죄	무죄인 사람을 유죄로 판단	유죄인 사람을 무죄로 판단
병원	병이 없다	병이 있다	건강한 사람을 병자라 진단	아픈 사람을 건강하다고 진단

비모수 검정

모수적 방법(Parametric method))

- 검정하고자 하는 모집단의 분포에 대한 가정을 하고,
  - 그 가정하에서 검정통계량과 검정통계량의 분포를 유도해 검정을 실시

비모수적 방법(Non-parametric method)

- 자료가 추출된 모집단의 분포에 대한 아무 제약을 가하지 않고 검정을 실시
  - 관측된 자료가 특정분포를 따른다고 가정할 수 없는 경우에 이용
  - 관측된 자료의 수가 적거나(30개 미만) 자료가 개체 간의 서열관계를 나타내는 경우에 이용

- “모수적 방법은 고급 승용차야 — 빠르고 정교하지만 도로가 평탄해야 해.
- 비모수적 방법은 SUV야 — 험한 길에서도 걷지만 속도는 좀 느리지.
- 너의 데이터 지형에 따라 골라 타면 되는 거다~ 🚗🚙”

항목	**모수적 방법**	**비모수적 방법**
분포 가정	정규분포 등 특정 분포를 가정함	분포에 대한 가정 없음
필요 정보	평균, 분산 등 **모수(parameter)** 필요	순위, 중앙값 등 **비모수(statistic)** 활용
검정력	적절한 조건에서는 더 강력	조건이 맞지 않으면 약해질 수 있음
표본 크기	작아도 효율적	일반적으로 더 많은 표본 필요
대표 예시	t-검정, z-검정, 분산분석(ANOVA), 회귀분석	윌콕슨 부호 순위합검정, 윌콕슨 순위합검정, 부호검정, 만-위트니의 U검정, 런검정, 스피어만의 순위상관계수, 크루스칼-왈리스 검정
민감도	이상치에 민감함	이상치에 덜 민감함
사용 예시	정규성 확인된 데이터 분석	순위 데이터나 정규성 위반된 데이터