[교과목 소개] 다변량 자료분석

- **교과목 개요:**

  본 교과목 다변량 자료분석(Multivariate Data Analysis)에서는 다양한 고급 통계분석 방법들 중 여러 개의 변수로 관측된 (다변량) 자료의 통계분석 방법들을 중점으로 다음의 내용을 다룹니다.

  다변량 자료의 공분산 구조를 활용하여 방대하게 수집되는 **데이터의 차원 축소**(주성분분석, 요인분석), **시각화**(t-SNE), 개체들의 유사성에 의해 **개체의 분류**(판별분석, 군집분석) 등 보다 응용된 통계분석 방법들의 이론을 학습합니다. 배운 이론을 바탕으로 R 통계 패키지를 이용하여 실제 자료를 분석하고 분석 결과의 해석을 실습합니다.

- **선수과목:** 수학1, 선형대수학, 확률 및 통계 1 & 2 (확률 및 통계 1만 수강해도 가능)
- **학습평가 방법:** 중간시험(30%), 기말보고서(30%), 과제(20%), 출석 및 기타(20%)
- **수업시간 및 장소:** 화F/목E, 팔달관 311호
- **교재 및 참고자료:** 강의노트 위주로 수업이 진행되며, 다음 참고자료를 활용합니다.

  Applied Multivariate Statistical Analysis - Johnson and Wichern (2008)

  Introduction to Applied Multivariate Statistical Analysis with R - Brian S. Everitt and Torsten Hothorn (2014)
- **문의:** 담당교수(안수현, T: 2560 E-mail: shahn@ajou.ac.kr)

<다변량 자료분석 방법론 소개>

Q. 단변량 자료분석 vs 다변량 자료분석

어떤 개체의 성질을 규명하기 위해 한 변수로 관찰하고 분석하는 것을 단변량 자료분석이라 할 수 있다. 하지만 실제 우리 주변에는 어떤 한 변수로 설명하기 어려운 경우가 많다. 따라서 각 개체들을 좀 더 다양한 각도와 측면에서 수집한 자료를 동시에 분석할 필요가 있다.

통계적으로 변수들의 연관성을 고려하여 여러 변수들을 동시에, 다차원적으로 분석하는 것을 다변량 자료분석이라 한다.

Q. 다변량 자료분석의 목적

빅데이터 시대에서 변수 개수와 개체 수가 많은 대용량이고 복잡한 구조의 다차원 자료를 분석하여 각 변수마다 단변량 분석하는 것보다 개선된 결과를 얻는다.

다변량 자료의 공분산 구조(변수들의 상관관계)를 활용하여 통계적추론, 방대하게 수집되는 데이터의 차원 축소(주성분분석, 요인분석), 개체들의 유사성에 의해 개체들을 분류(판별분석, 군집분석)하는 등 보다 응용된 통계분석 방법이다.

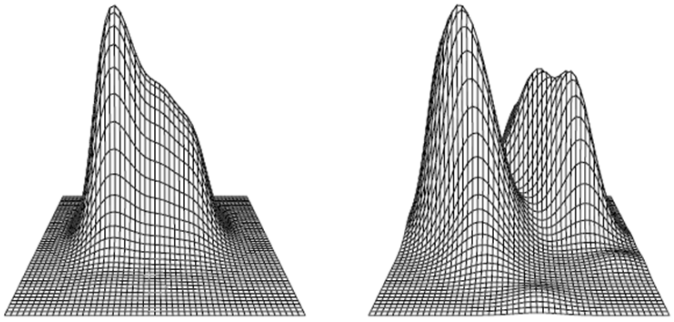


그림 1. 이차원 자료의 확률밀도함수. 두 변수의 상관관계를 고려하지 않은 두 변수의 단변량 확률밀도함수 곱(왼쪽)과 두 변수의 상관관계를 고려한 결합확률밀도함수(오른쪽).

Q. 본 교과목에서 다루게 될 다변량 자료분석 방법들

- **두 집단 다차원 비교 검정 (두 평균 벡터 검정)**

  : 단변량 정규분포 가정 하에 t 통계량을 이용하여 두 집단의 평균을 비교 검정하는 것과 유사하게 **다변량 정규분포** 가정 하에 Hotelling's T^2 통계량을 이용하여 두 집단의 평균 벡터를 비교 검정하는 방법이다.

- **주성분분석 (Principle Component Analysis, PCA)**

  : 최소한의 정보손실과 함께 가능한 적은 변수(주성분)를 구해 원 자료의 차원을 축소하는 방법이다. 추후 개체들의 분류나 회귀 분석 시 발생하는 다중공선성 문제 해결 등에 활용된다.

- **요인분석 (Factor Analysis, FA)**

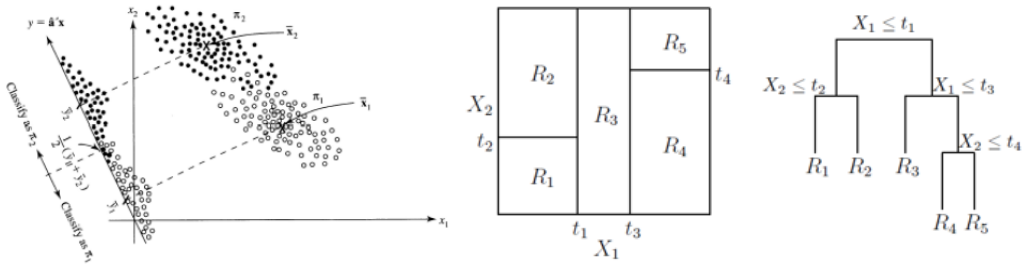
  : 원 변수를 설명하는 내재변인(요인)에 의해 원 변수를 그룹화하는 방법이다.

  Ex) 원 변수: 수학, 과학, 영어, 국어

  -> 요인: 수리능력 (수학, 과학), 언어능력 (영어, 국어)

- **판별/분류 (Discrimination or Classification)**

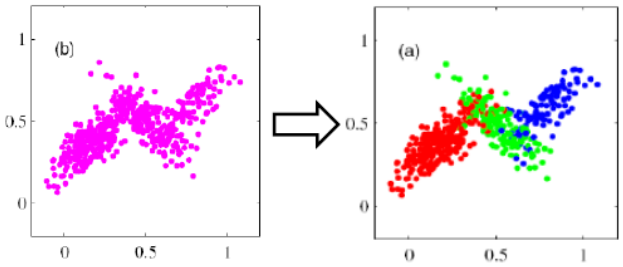
  : 두 개 이상의 그룹으로 나뉜 개체들로부터 분류에 영향을 미칠 것 같은 변수(특성)를 측정하고 이를 이용하여 판별식을 구하는 등 새로운 개체를 분류하는 방법이다.



- **군집분석 (Clustering)**

  : 판별 분석과 개체들을 그룹화한다는 면에서 유사하나 판별 분석과 달리 분석을 통해 적절한 그룹 개수

를 결정하고 개체를 분류하는 방법이다.



Q. 수업진행방식

위 다양한 다변량 자료분석 방법들의 **이론을 학습**하고, **R 통계 프로그램을** 이용하여 실제 자료 분석 및 결과 해석을 **실습**한다.