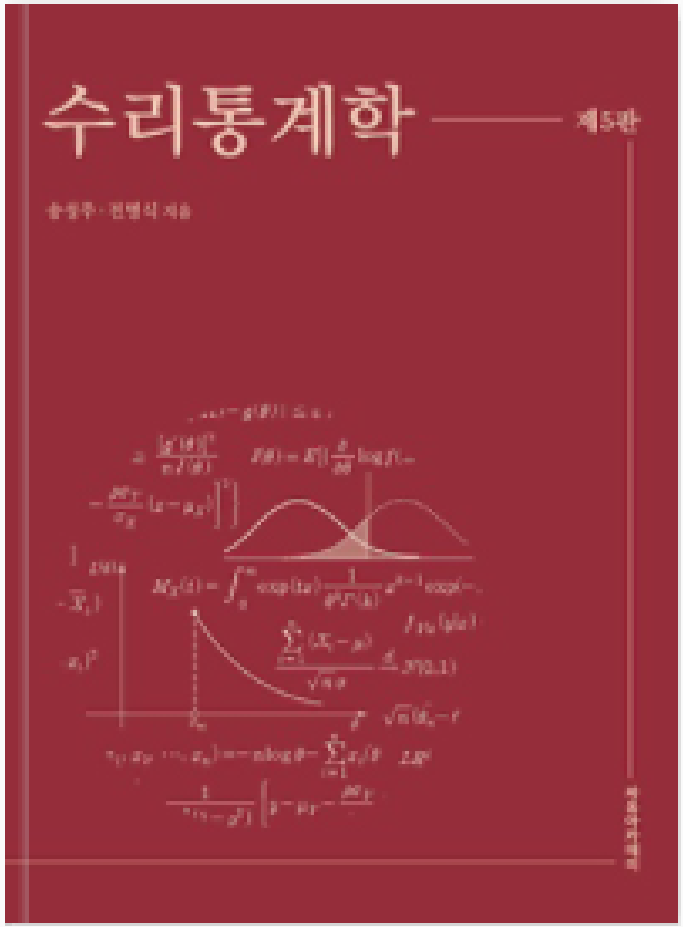
이론 통계학

2021년 봄학기

전북대학교 통계학과

교재 소개

수리통계학 R을 이용한 실습 (강기훈, 박진호)



수리통계학 (송성주, 전명식)

수리통계학 (Mathematical Statistics)

- 기초적인 통계학개론에서 출발하여 다양한 통계분석방법을 배우는 과정에서 필수적인 통계이론을 다루는 과목

- 일정부분 통계 이론을 이해하지 않고서는 좀 더 깊이 있는 자료분석 방법을 다루는 것이 쉽지 않음

- 통계학에서 사용되는 방법들의 이론적 타당성을 제공하고 서로 연결하는 기본적인 언어 역할

수리통계학 (Mathematical Statistics)

확률과 확률 분포

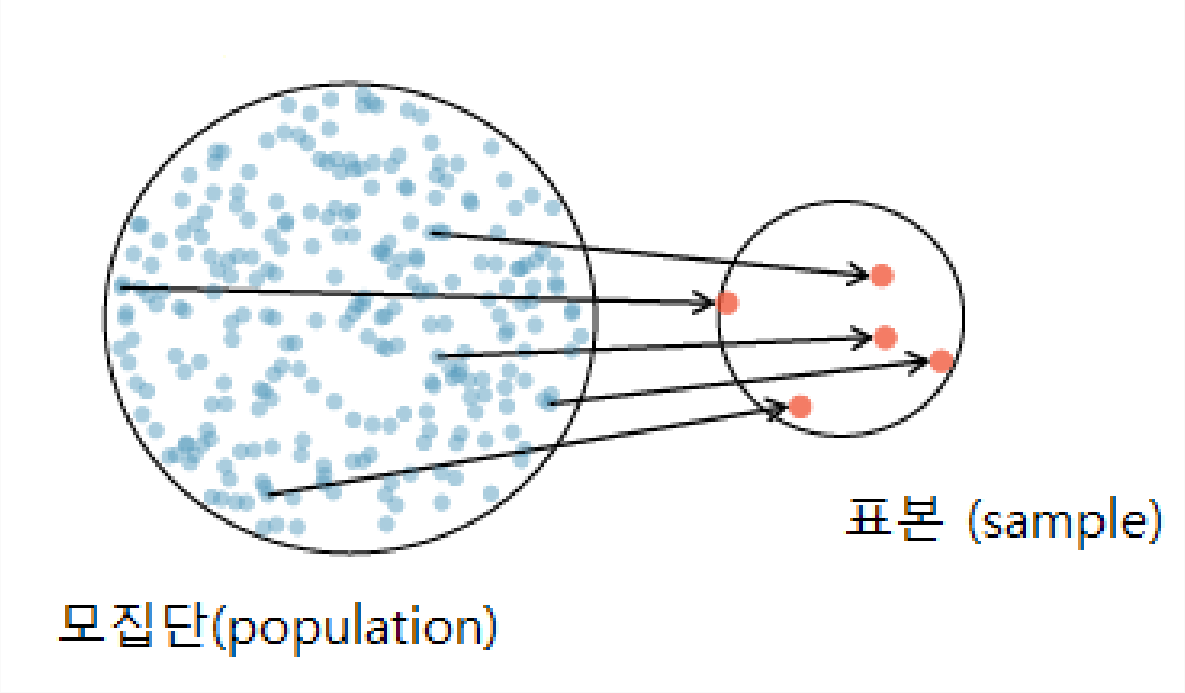
- 확률
- 확률변수와 확률분포
- 결합분포
- 기댓값
- 표본분포
- 확률변수의 극한

통계적 추론

- 모수의 추정
- 가설검정
- 통계적 결정이론과 베이즈추론
- 비모수적 추정 및 검정 (*)

통계학

- 자료를 분석하여 모집단에 대한 유용한 정보를 이끌어내는 것을 목표로 함.



- 일부분의 자료를 이용하여 모집단에 대한 결론을 얻어내는 과정에는 추측 단계를 거침
- 통계학은 이러한 추측과정을 체계적으로 할 수 있도록 하여 추측과정에서 발생하는 오류의 가능성 낮추려고 함
- 다양한 자료분석방법 - 회귀분석, 시계열분석, 범주형자료분석, 다변량자료분석, 생존분석 등

모형화 (modelling)

- 모형화 - 모집단에 대하여 가정
- 자료의 특징이나 성격에 따라 자료로부터 유용한 정보를 얻는 한 가지 방법
- 모형화 과정을 통해 우리가 알고자 하는 대상인 모집단을 그 특성을 결정지어 주는 몇 개의 값으로 나타낼 수 있음

모수 (parameter)
모집단의 특성을 결정하는 값

- 통계학 : '자료로부터 유용한 정보를 이끌어 내는 것'

  모형화를 통하여

  : $\Rightarrow$ '자료로부터 모집단의 특성을 나타내는 모수에 대한 추측을 하는 것'

모형화 (modelling)의 예

ex1) 모집단의 특징을 파악하기 위해 모집단의 분포이 정규분포를 따른다고 가정

ex2) 두 변수사이의 연관성을 분석하기 위해 두 변수 사이에 선형관계를 가정

모집단이 정규분포를 따른다는 가정의 예

- 정규분포는 평균과 분산에 의해 결정됨.

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

- 이 가정 하에서는 평균과 분산에 의하여 모집단의 특성이 결정됨

  $\rightarrow$ 모수 : 평균, 분산

- 모수를 어떻게 추측?
- 추측한 모수의 값의 성질은?

위 질문에 대한 답을 찾는 방법을 다루는 것이 수리통계학,

즉 자료분석에서 사용되는 기본적인 통계학적 언어와 방법론을 다루는 분야.

모형화 과정과 모수 추측의 예 1

우리나라 성인 남성의 콜레스테롤 수준을 알아보기위해 n명의 성인 남성을 대상으로 조사하였다.

- 관측한 n명의 콜레스테롤 수준: $X_1, X_2, \ldots, X_n$
- 모집단 : 우리나라 전체 성인 남성의 콜레스테롤 수준
- 모집단에 대한 가정 즉 모형화없이 주어진 n개의 자료로부터 모집단에 대한 유용한 추측 어려움
- 모집단의 확률분포 가정 - 모형화의 한 가지 방법
- 성인 남성의 콜레스테롤 수준 $\sim N(\mu, \sigma^2)$ (가정의 타당성 검토 필요)

모형화 과정과 모수 추측의 예 1

- 자료로부터 모수 μ, σ^2 추측
- 모수추측값: $\hat{\mu}, \hat{\sigma}^2$
- 성인 남성의 콜레스테롤 수준 $\sim N(\hat{\mu}, \hat{\sigma}^2)$라고 추측
- $\hat{\mu}, \hat{\sigma}^2$을 어떻게 구할까?
- 자료의 표본평균과 표본분산으로 추측하면 좋을까?
- 더 좋은 방법은 없을까?

모형화 과정과 모수 추측의 예 2

두 변수 사이의 연관성 규명하고자 하는 경우

- 학생들의 TV 시청시간이 시험점수에 영향을 준다고 생각되어 두 변수의 연관성을 분석하고자 한다.
- 관측한 학생 n명의 TV시청시간과 시험점수는 다음과 같다고 하자.

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

- TV 시청시간과 시험점수 사이에 선형관계 가정를 가정하자.

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

모형화 과정과 모수 추측의 예 2

학생들의 TV 시청시간이 시험점수에 영향을 준다고 생각되어 두 변수의 연관성을 분석하고자 한다.

- 관측한 학생 n명의 TV시청시간과 시험점수:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

- TV 시청시간과 시험점수 사이에 선형관계 가정

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \Leftrightarrow Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- Y의 확률분포, 모집단의 분포를 결정짓는 모수: α, β, σ^2
- 추정 : TV 시청시간과 시험점수의 연관성 규명하기 위한 모형에서 중요한 모수 β
- 가설검정 : $\beta = 0$? $\beta < 0$?