다변량통계학-2025년 2학기

서울시립대학교 통계학과 이용희

2025-09-07

Table of contents

List of Figures

List of Tables

Preface

이 책은 2025년 다변량통계학에 대한 온라인 교재입니다.

> **ℹ 표기법**
>
> 이 책에서 사용된 기호, 표기법, 프로그램의 규칙과 쓰임은 다음과 같습니다.
>
> - 스칼라(scalar)와 일변량 확률변수는 일반적으로 보통 글씨체의 소문자로 표기한다. 특별한 이유가 있는 경우 대문자로 표시할 것이다.
> - 벡터, 행렬, 다변량 확률벡터는 **굵은 글씨체**로 표기한다.
> - 통계 프로그램은 `R`을 이용하였다. 각 예제에 사용된 `R` 프로그램은 코드 상자를 열면 나타난다.

1 다변량 자료의 표현과 분포

```r
library(tidyverse)
library(here)
library(knitr)
library(kableExtra)
library(flextable)
```

다변량 자료(multivariate data)는 두 개 이상의 변수를 측정한 자료를 말합니다. 예를 들어, 학생들의 키와 몸무게, 시험 점수와 공부 시간, 나이와 소득 등이 다변량 자료에 해당합니다. 다변량 자료는 변수들 간의 관계를 분석하고 이해하는 데 중요한 역할을 합니다. 다변량 자료를 효과적으로 표현하고 분석하기 위해 다양한 그래프와 통계 기법이 사용됩니다. 이 장에서는 다변량 자료의 표현 방법과 분포를 이해하는 데 필요한 기본 개념과 도구들을 소개합니다.

1.1 예제: 국민체력100

국민체력100은 국민의 체력증진과 건강증진을 위해 개발된 종합적인 체력측정 프로그램이다. 이 프로그램은 다양한 연령대와 성별에 맞춘 체력측정 항목을 포함하고 있으며, 이를 통해 개인의 체력 상태를 평가하고 개선할 수 있는 기회를 제공한다.

다음은 2024년에 청소년에 대한 국민체력100 측정 항목과 자료의 일부이다. 먼저 측정항목에 대한 설명에 대한 자료를 보자.

```r
load(here("data", "physical100_teen_2024.RData"))
ls()
```

```
[1] "selected_df"     "selected_var_df"
```

1.2 다변량 확률변수

1.2.1 일변량분포

일변량 확률변수 X가 확률밀도함수 $f(x)$를 가지는 분포를 따를때 기대값과 분산은 다음과 같이 정의된다.

$$E(X) = \int xf(x)dx = \mu, \quad V(X) = E[X - E(X)]^2 = \int (x-\mu)^2 f(x)dx = \sigma^2$$

새로운 확률변수 Y가 확률변수 X의 선형변환으로 표시된다면 (a와 b는 실수)

$$Y = aX + b$$

일변량 확률변수 X의 기대값(평균)과 분산은 다음과 같이 계산된다.

$$\begin{aligned} E(Y) &= E(aX+b) \\ &= \int (ax+b)f(x)dx \\ &= a\int xf(x)dx + b \\ &= aE(X) + b \\ &= a\mu + b \\ V(Y) &= Var(aX+b) \\ &= E[aX+b-E(aX+b)]^2 \\ &= E[a(X-\mu)]^2 \\ &= a^2E(X-\mu)^2 \\ &= a^2\sigma^2 \end{aligned}$$

1.3 확률벡터와 분포

확률벡터 $\boldsymbol{X}$가 p 차원의 다변량분포를 따른다고 하고 결합확률밀도함수 $f(\boldsymbol{x}) = f(x_1, x_2, ..., x_p)$를 를 가진다고 하자.

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ .. \\ X_p \end{bmatrix}$$

다변량 확률벡터의 기대값(평균벡터)과 공분산(행렬)은 다음과 같이 계산된다.

$$\boldsymbol{E}(\boldsymbol{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ E(X_3) \\ .. \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ .. \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

$$V(\boldsymbol{X}) = Cov(\boldsymbol{X}) = E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ & \cdots & \cdots & \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix} = \boldsymbol{\Sigma}$$

여기서 $\sigma_{ii} = V(X_i)$, $\sigma_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i)$이다. 따라서 공분산 행렬 $\boldsymbol{\Sigma}$는 대칭행렬(symmetric matrix)이다. 다음 공식은 유용한 공식이다.

$$\boldsymbol{\Sigma} = E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t = E(\boldsymbol{X}\boldsymbol{X}^t) - \boldsymbol{\mu}\boldsymbol{\mu}^t$$

두 확률변수의 상관계수 ρ_{ij}는 다음과 같이 정의된다.

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{V(X_i)V(X_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

새로운 확률벡터 $\boldsymbol{Y}$가 확률벡터 $\boldsymbol{X}$ 의 선형변환라고 하자.

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$$

단 여기서 $\boldsymbol{A} = \{a_{ij}\}$는 $p \times p$ 실수 행렬이고 $\boldsymbol{b} = (b_1 b_2 \dots b_p)^t$는 $p \times 1$ 실수 벡터이다.

확률벡터 $\boldsymbol{Y}$ 의 기대값(평균벡터)과 공분산은 다음과 같이 계산된다.

$$\begin{aligned} E(\boldsymbol{Y}) &= E(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}) \\ &= \boldsymbol{A}E(\boldsymbol{X}) + \boldsymbol{b} \\ &= \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \\ V(\boldsymbol{Y}) &= Var(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b}) \\ &= E[\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b} - E(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b})][\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b} - E(\boldsymbol{A}\boldsymbol{X}+\boldsymbol{b})]^t \\ &= E[\boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}\boldsymbol{\mu}][\boldsymbol{A}\boldsymbol{X} - \boldsymbol{A}\boldsymbol{\mu}]^t \\ &= E[\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})][\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu})]^t \\ &= \boldsymbol{A}E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^t]\boldsymbol{A}^t \\ &= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^t \end{aligned}$$

만약 표본 $\boldsymbol{X}_i, \boldsymbol{X}_2, \dots, \boldsymbol{X}_n$ 이 독립적으로 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 분포에서 추출되었다면 표본의 평균벡터 $\bar{\boldsymbol{X}}$ 는 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\frac{1}{n}\boldsymbol{\Sigma}$ 인 분포를 따른다.

$$\bar{\boldsymbol{X}} = \begin{bmatrix} \sum_{i=1}^{n} X_{i1}/n \\ \sum_{i=1}^{n} X_{i2}/n \\ \sum_{i=1}^{n} X_{i3}/n \\ .. \\ \sum_{i=1}^{n} X_{ip}/n \end{bmatrix}$$

여기서 X_{ij} 는 i 번째 표본벡터 $\boldsymbol{X}_i = (X_{i1}X_{i2} \dots X_{ip})^t$ 의 j 번째 확률변수이다.

1.4 다변량 정규분포

일변량 확률변수 X 가 평균이 μ 이고 분산이 σ^2 인 정규분포를 따른다면 다음과 같이 나타내고

$$X \sim N(\mu, \sigma^2)$$

확률밀도함수 $f(x)$ 는 다음과 같이 주어진다.

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

p-차원 확률벡터 $\boldsymbol{X}$ 가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 다변량 정규분포를 따른다면 다음과 같이 나타내고

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

확률밀도함수 $f(\boldsymbol{x})$ 는 다음과 같이 주어진다.

$$f(\boldsymbol{x}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^t}{2}\right)$$

다변량 정규분포 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 를 따르는 확률벡터 $\boldsymbol{X}$ 를 다음과 같이 두 부분으로 나누면

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \quad \boldsymbol{X}_1 = \begin{bmatrix} \boldsymbol{X}_{11} \\ \boldsymbol{X}_{12} \\ \vdots \\ \boldsymbol{X}_{1p} \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} \boldsymbol{X}_{21} \\ \boldsymbol{X}_{22} \\ \vdots \\ \boldsymbol{X}_{2q} \end{bmatrix}$$

각각 다변량 정규분포를 따르고 다음과 같이 나타낼 수 있다.

$$\begin{bmatrix} E(\boldsymbol{X}_1) \\ E(\boldsymbol{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \begin{bmatrix} V(\boldsymbol{X}_1) & Cov(\boldsymbol{X}_1, X_2) \\ Cov(\boldsymbol{X}_2 X_1) & V(\boldsymbol{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^t & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \Sigma_{12} \\ \boldsymbol{\Sigma}_{12}^t & \Sigma_{22} \end{bmatrix}\right)$$

확률벡터 $\boldsymbol{X}_2 = \boldsymbol{x}_2$가 주어진 경우 $\boldsymbol{X}_1$의 조건부 분포는 p-차원 다변량 정규분포를 따르고 평균과 공분산은 다음과 같다.

$$E(\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{x}_2), \quad V(\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t$$

예를 들어 2-차원 확률벡터 $\boldsymbol{X} = (X_1, X_2)^t$가 평균이 $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ 이고 공분산 $\boldsymbol{\Sigma}$가 다음과 같이 주어진

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

이변량 정규분포를 따른다면 확률밀도함수 $f(\boldsymbol{x})$에서 exp함수의 인자는 다음과 같이 주어진다.

$$(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^t = \\ -\frac{1}{2(1-\rho^2)}\left[\left(\frac{(x_1-\mu_1)^2}{\sigma_{11}}\right) + \left(\frac{(x_2-\mu_2)^2}{\sigma_{22}}\right) - 2\rho\left(\frac{(x_1-\mu_1)}{\sqrt{\sigma_{11}}}\right)\left(\frac{(x_2-\mu_2)}{\sqrt{\sigma_{22}}}\right)\right]$$

그리고 $p = 2$인 경우 확률밀도함수의 상수부분은 다음과 같이 주어진다.

$$(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}}$$

여기서 $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$

만약 $X_2 = x_2$가 주어졌을 때 X_1의 조건부 분포는 정규분포이고 평균과 분산은 다음과 같이 주어진다.

$$E(X_1|X_2 = x_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(\mu_2 - x_2) = \mu_1 + \rho\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(\mu_2 - x_2)$$

$$V(X_1|X_2 = x_2) = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} = \sigma_{11}(1-\rho^2)$$

다변량 정규분포에서 공분산이 0인 두 확률 변수는 독립이다.

$$\sigma_{ij} = 0 \leftrightarrow X_i \text{ and } X_j \text{ are independent}$$

1.5 표준정규분포로의 변환

일변량 확률변수 X가 평균이 μ 이고 분산이 σ^2 인 경우 다음과 같은 선형변환을 고려하면.

$$Z = \frac{X-\mu}{\sigma} = (\sigma^2)^{-1/2}(X-\mu)$$

확률변수 Z 는 평균이 0 이고 분산이 1인 분포를 따른다.

References