(인트로)
다변량 분석을 위한 확률표본

3.1 다변량 분포함수

성분들이 확률변수(random variable)로 이루어진 벡터를 **확률벡터(random vector)**,
성분들이 확률변수로 이루어진 행렬을 **확률행렬(random matrix)**.

확률변수 $X_1, X_2, \cdots, X_p$ 를 원소로 가진 $p \times 1$ 확률벡터 $\boldsymbol{X}$가 연속분포를 가지며 결합확률밀도함수 $f(x_1, x_{2,} \cdots, x_p)$ 를 가질 때,

■ **결합분포함수** F :

$$F(\boldsymbol{x}) = F(x_1, x_2, \cdots, x_p) = P(X_1 \leq x_1, \ X_2 \leq x_2, \cdots, \ X_p \leq x_p)$$

■ X_i의 **주변확률밀도함수**(marginal probability density function):

$$f_i(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_p) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_p$$

■ $X_1, X_2, \ldots, X_k$의 **주변확률밀도함수:**

$$f(x_1, x_2, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \cdots, x_p)\, dx_{k+1} dx_{k+2} \cdots dx_p$$

으로 $X_1, X_2, \ldots, X_k$를 제외한 변수들에 대해 적분한다.

3.2 확률벡터, 확률행렬의 기대값과 공분산행렬

확률행렬 $\boldsymbol{U} = \{U_{ij}\}_{n \times p}$, $\boldsymbol{V} = \{V_{ij}\}_{n \times p}$ 에 대하여 기대값은

■ $$E(\boldsymbol{U}) = \{E(U_{ij})\}_{n \times p}$$

여기서

$$E(U_{ij}) = \begin{cases} \int_{-\infty}^{\infty} u f_{ij}(u) du, & U_{ij}\text{가 연속} \\ \sum_{all \ u_{ij}} u_{ij} p_{ij}(u_{ij}), & U_{ij}\text{가 이산} \end{cases}$$

여기서 $f(u_{ij})$는 U_{ij}의 확률밀도함수(probability density function),

$p_{ij}(u_{ij}) = P(U_{ij} = u_{ij})$ 는 확률질량함수(probability mass function)

■ $p \times 1$ 확률벡터 $\boldsymbol{X}$에 대해

모평균벡터

$$E(\boldsymbol{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

모공분산행렬 $\boldsymbol{\Sigma} = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = Cov(\boldsymbol{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$

모상관행렬 $\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}$ 을 갖는다고 하자. 여기서 $\sigma_{jk} = Cov(X_j, X_k)$

$$\rho_{ik} = \frac{\sigma_{ik}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{kk}}}$$

3.3 다변량 확률표본의 기대값과 공분산행렬

모집단으로부터의 확률표본 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 에 대해 각 확률벡터 $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$, $i = 1, 2, \ldots, n$ 는 모집단 값으로 $\boldsymbol{\mu} = E(\boldsymbol{X})$, $\boldsymbol{\Sigma} = Cov(\boldsymbol{X})$ 를 갖는다.

■ 표본평균벡터:

$$\overline{\boldsymbol{X}} = \begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} X_{i1} \\ \frac{1}{n} \sum_{i=1}^{n} X_{i2} \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} X_{ip} \end{pmatrix}$$

■ 표본공분산행렬:

$$\begin{aligned} \boldsymbol{S} &= \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})' \\ &= \frac{1}{n-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' - n \overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}' \right) \\ &= \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ \vdots & \vdots & & \vdots \\ s_{1p} & \cdots & \cdots & s_{pp} \end{pmatrix} = \{ s_{jk} \} \end{aligned}$$

■ 표본상관행렬 :

$$\boldsymbol{R}= \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

여기서 $r_{jk} = \dfrac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$ 는 확률변수 X_j와 X_k의 표본상관계수

정리 3.1 $p \times 1$ 확률벡터 $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ 은 평균벡터 $\boldsymbol{\mu}$와 공분산행렬 $\boldsymbol{\Sigma}$ 를 가진 정규다변량 분포로부터의 확률표본(random sample)일 때,
$\overline{\boldsymbol{X}}$ 는 $\boldsymbol{\mu}$의 불편추정량(unbiased estimator)이며 $\boldsymbol{S}$는 $\boldsymbol{\Sigma}$ 의 불편추정량이다. 즉

$$E(\overline{\boldsymbol{X}}) = \boldsymbol{\mu}, \ E(\boldsymbol{S}) = \boldsymbol{\Sigma}$$

3.4 일반화분산과 총분산

3.4.1 일반화분산

전체 변동량에 대한 측도

일반화분산(generalized variance) = $|\boldsymbol{\Sigma}|$

일반화표본분산(generalized sample variance) = $|\boldsymbol{S}|$.

: 데이터의 변동으로 만들어내는 다변체 부피의 의미.

변동의 방향이나 각 변수의 변동의 폭에 대해 알려주지는 않는다.

$p=1$인 경우 일반화분산은 변수의 분산이 된다.

한 변수가 다른 변수들의 선형조합으로 완전히 표현될 경우: 일반화분산= 0

3.4.2 총분산

총분산(total variance) = $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = tr(\boldsymbol{\Sigma})$

총표본분산(total sample variance) = $s_{11} + s_{22} + \cdots + s_{pp} = tr(\boldsymbol{S})$

: 각 변수들의 분산의 합

변수들의 공분산은 고려하지 않는다.

모든 변수의 분산이 0 인 경우에만 총분산이 0 이 된다.

■ 일반화분산, 총분산 모두 변수들의 퍼짐이 클수록 큰 값을 나타낸다.

■ 일반화분산은 변수간의 다중공선성(multicollinearity)이 있을 때 작은 값을 나타내며 총분산은 각 변수들의 변동이 적을수록 작은 값을 나타낸다.

3.5 상호상관성 측도

변수들 전체에 대한 변수내 상관성에 대한 측도로 스칼라로 나타내고자한다.

상관행렬 $\boldsymbol{R}$ 의 고유값은 $\lambda_1, \cdots , \lambda_p$ 이고 $\lambda_1 > \cdots > \lambda_p$ 라고 할 때

변수내 상호상관(intercorrelation)을 나타내는 측도

(1) 조건수(condition number) $= \dfrac{\lambda_1}{\lambda_p}$

Mason, Gunst와 Webster(1975),

상관행렬 $\boldsymbol{R}$ 의 최대 고유값과 최소 고유값의 비

(2) $\displaystyle\sum_{j=1}^{p} \frac{1}{\lambda_j}$

Hoerl과 Kennard(1970)

≪예제 3.1≫ 다음은 A 과목을 수강한 학생들의 중간고사 점수(X_1)와 학기말고사 점수(X_2)

학생번호	X_1	X_2
1	90	80
2	80	90
3	75	80
4	70	70
5	65	80

(1) 표본평균벡터

$$\overline{X_1} = \frac{1}{n}\sum_{i=1}^{n} X_{i1} = \frac{1}{5}(90+80+75+70+65) = 76$$

$$\overline{X_2} = \frac{1}{n}\sum_{i=1}^{n} X_{i2} = \frac{1}{5}(80+90+80+70+80) = 80$$

표본평균벡터는 $\overline{\boldsymbol{X}} = \begin{pmatrix} \overline{X_1} \\ \overline{X_2} \end{pmatrix} = \begin{pmatrix} 76 \\ 80 \end{pmatrix}$

(2) 표본공분산행렬

$$s_{11}=\frac{1}{n-1}\sum_{i=1}^{n}(\ X_{i1}-\overline{X_1}\,)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_{i1}^2-n\overline{X_1^2}\right)$$

$$=\frac{1}{4}(90^2+80^2+75^2+70^2+65^2-5\cdot 76^2)=92.5$$

$$s_{22}=\frac{1}{n-1}\sum_{i=1}^{n}(\ X_{i2}-\overline{X_2}\,)^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_{i2}^2-n\overline{X_2^2}\right)$$

$$=\frac{1}{4}(80^2+90^2+80^2+70^2+80^2-5\cdot 80^2)=50.0$$

$$s_{12}=s_{21}=\frac{1}{n-1}\sum_{i=1}^{n}(X_{i1}-\overline{X_1})(X_{i2}-\overline{X_2})=\frac{1}{n-1}\left(\sum_{i=1}^{n}X_{i1}X_{i2}-n\cdot\overline{X_1}\,\overline{X_2}\right)$$

$$=\frac{1}{4}(90\cdot 80+80\cdot 90+75\cdot 80+70\cdot 70+65\cdot 80-5\cdot 76\cdot 80)=25.$$

공분산행렬 $\quad S=\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}=\begin{pmatrix} 92.5 & 25 \\ 25 & 50.0 \end{pmatrix}$.

(3) 표본상관행렬

X_1과 X_2 의 상관계수를 구하면

$$r_{12} = r_{21} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{25}{\sqrt{92.5}\sqrt{50.0}} = 0.368$$

상관행렬

$$\boldsymbol{R} = \begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.368 \\ 0.368 & 1 \end{pmatrix}$$

즉, 중간고사와 학기말고사 점수간의 상관계수가 0.368

(1) 일반화표본분산 $= |\boldsymbol{S}| = 92.5 \cdot 50.0 - 25^2 = 4000$

(2) 총표본분산 $= s_{11} + s_{22} = tr(\boldsymbol{S}) = 92.5 + 50 = 142.5$

(3) 변수들 상호상관 ($p=2$)

표본상관행렬 $\boldsymbol{R}=\begin{pmatrix} 1 & r_{12} \\ r_{21} & 1 \end{pmatrix}=\begin{pmatrix} 1 & 0.368 \\ 0.368 & 1 \end{pmatrix}$ 의 고유값 $\lambda_1=1.368$ $\lambda_2=0.632$

① 조건수 $=\dfrac{\lambda_1}{\lambda_2}=\dfrac{1.368}{0.632}=2.165$

② $\displaystyle\sum_{j=1}^{p}\frac{1}{\lambda_j}=\frac{1}{1.368}+\frac{1}{0.632}=2.313$

● R에서는 `det()` 함수로 일반화분산, `sum(diag() )` 함수를 사용하여 총분산을 구할 수 있다.

```
> x1=c(90,80,75,70,65)
> x2=c(80,90,80,70,80)
> a=cbind(x1,x2)
> a
     x1 x2
[1,] 90 80
[2,] 80 90
[3,] 75 80
[4,] 70 70
[5,] 65 80
> S=var(a)
> S
     x1 x2
x1 92.5 25
x2 25.0 50
[1] 142.5
```

```
> g_var= det(S)    # 일반화분산
> g_var
[1] 4000
> total_var=var(x1)+var(x2)  # 총분산
> total_var
[1] 142.5
> R=cor(a)
> R=round(R, digits=3)
> R
      x1    x2
x1 1.000 0.368
x2 0.368 1.000
```

```
> ea=eigen(R)
> ea
$values
[1] 1.368 0.632
$vectors
          [,1]       [,2]
[1,] 0.7071068 -0.7071068
[2,] 0.7071068  0.7071068
> cond_num = max(ea$values)/min(ea$values)
> cond_num
[1] 2.164557
```