I　正規分布とその派生

abstract　切断正規分布、混合正規分布、多変量正規分布の基本的な性質と重要な計算テクニックについて例題形式で学びます。

1　Introduction

正規分布から派生した分布に、切断正規分布、混合正規分布、多変量正規分布があります。これらの分布は統計学、機械学習の手法を学ぶうえで欠かせない知識です。例えば、切断正規分布はベイズ推定の事前分布やTobit回帰モデルへの応用が有名です。混合正規分布は、そのパラメータの推定がクラスタリングの代表的な手法になっています。多変量正規分布に至っては、判別分析・探索的因子分析・共分散構造分析をはじめとする人文科学分野の多変量解析手法を支えています。

これらの分布の性質は、その重要性と比例して、統計検定1級「統計応用」（人文科学）における頻出分野になっています。具体的には2016年度以降、このトピックが出題されなかったのは2021年度のみです。以下は出題履歴を表にまとまたものです。

年度	2016	2017	2018	2019	2021	2022	2023	2024
切断正規分布	✔	✔		✔				✔
混合正規分布			✔				✔	
多変量正規分布		✔				✔		✔

そこで第I章では、切断正規分布、混合正規分布、多変量正規分布の定義を紹介し、その特有の計算を解説します。これらの計算に慣れておくことは統計検定1級の勉強のみならず、普段の分析や統計・機械学習の手法を学ぶ際にもたいへん有益です。

2　切断正規分布

2.1　正規分布の復習

確率変数 X が**正規分布** $N(\mu, \sigma^2)$ に従うとは、X が $\mu, \sigma^2 > 0$ によって定まる以下の確率密度関数

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

を持つことと定義されます。特に期待値が 0 かつ分散が 1 の正規分布 $N(0,1^2)$ を**標準正規分布**といいます。

2.2節で大切になる公式を紹介します。

> **定理　標準正規分布の確率密度関数に関する公式**
>
> 標準正規分布の確率密度関数を $\phi(z)$ と表すとき、$\phi'(z) = -z\phi(z)$ が成り立つ。

証明

標準正規分布の確率密度関数は $\phi(z) = \dfrac{1}{\sqrt{2\pi}}\exp\left[-\dfrac{z^2}{2}\right]$ です。これを微分すると

$$\phi'(z) = -\frac{1}{\sqrt{2\pi}} z \exp\left[-\frac{z^2}{2}\right] = -z\phi(z)$$

が得られます。■

この他に重要な性質を二つ紹介します。証明は省略しますが、統計数理では証明を求められることがあります。必ず自力で証明できるようにしておきましょう。

- 期待値は $\mathbb{E}[X] = \mu$、分散は $\mathbb{V}[X] = \sigma^2$ になる。
- 再生性 : X_1, X_2 が独立に正規分布 $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$ に従うとき、$c_1X_1 + c_2X_2$ もまた正規分布 $N(c_1\mu_1 + c_2\mu_2, c_1^2\sigma_1^2 + c_2^2\sigma_2^2)$ に従う。

Remark　$\phi'(z) = -z\phi(z)$ はさまざまな場面で役立つ公式です。例えば、標準正規分布の裾確率を抑えるMills比の不等式を証明する際に現れ、ベイズ最適化などにつながっていきます。■

2.2　切断正規分布の定義と期待値・分散

確率変数 X が正規分布に従うとき、条件 $X \geq a$ の下での X の分布を**切断正規分布**といいます。切断正規分布の以下の性質は重要です。必ず計算できるようにしておきましょう。

定理　切断正規分布のpdfと期待値・分散

Z を標準正規分布に従う確率変数とし、累積分布関数を $\Phi(z)$、確率密度関数を $\phi(z)$ と表すことにする。このとき、以下のことが成り立つ。

[1]　条件 $Z \geq a$ の下での Z の確率密度関数は $f(z) = \dfrac{\phi(z)}{1-\Phi(a)}, (z \geq a)$ である。

[2]　[1]の確率密度関数で定義される分布の期待値と分散は、それぞれ以下のようになる。

$$\mathbb{E}[Z \mid Z \geq a] = \frac{\phi(a)}{1-\Phi(a)}$$
$$\mathbb{V}[Z \mid Z \geq a] = 1 + a\frac{\phi(a)}{1-\Phi(a)} - \left(\frac{\phi(a)}{1-\Phi(a)}\right)^2$$

証明

[1]　条件付き分布の定義から $f(z) = \dfrac{\phi(z)}{\mathbb{P}[Z \geq a]}, (z \geq a)$ です。この分母 $\mathbb{P}[Z \geq a]$ は累積分布関数の定義を用いて

$$\mathbb{P}[Z \geq a] = 1 - \mathbb{P}[Z < a] = 1 - \Phi(a)$$

と表せます。したがって、条件 $Z \geq a$ の下での Z の確率密度関数は $f(z) = \dfrac{\phi(z)}{1-\Phi(a)}$, $(z \geq a)$ です。

[2]　2.1節で紹介した公式 $\phi'(z) = -z\phi(z)$ が重要になります。期待値を計算します。

$$\mathbb{E}[Z \mid Z \geq a] = \frac{1}{1-\Phi(a)} \int_a^{\infty} z\phi(z)dz = \frac{1}{1-\Phi(a)} [-\phi(z)]_a^{\infty} = \frac{\phi(a)}{1-\Phi(a)}$$

分散を計算します。分散の公式

$$\mathbb{V}[Z \mid Z \geq a] = \mathbb{E}[Z^2 \mid Z \geq a] - \mathbb{E}[Z \mid Z \geq a]^2$$

に注目します。右辺第2項は期待値の計算から直ちに従うので、右辺第1項のみを計算します。

$\mathbb{E}[Z^2 \mid Z \geq a]$ を期待値の定義通りに書き下してみましょう。

$$\mathbb{E}[Z^2 \mid Z \geq a] = \frac{1}{1-\Phi(a)} \int_a^{\infty} z^2\phi(z)dz$$

この右辺を部分積分で計算します。

$$\int_a^{\infty} z^2\phi(z)dz = [-z\phi(z)]_a^{\infty} + \int_a^{\infty} \phi(z)dz = a\phi(a) + (1-\Phi(a))$$

従って $\mathbb{E}[Z^2 \mid Z \geq a] = 1 + a\dfrac{\phi(a)}{1-\Phi(a)}$ がわかり、分散の公式に代入すれば

$$\mathbb{V}[Z \mid Z \geq a] = 1 + a\frac{\phi(a)}{1-\Phi(a)} - \left(\frac{\phi(a)}{1-\Phi(a)}\right)^2$$

が得られます。■

特に、$a = 0$ の場合は $\mathbb{E}[Z \mid Z \geq 0] = \sqrt{\dfrac{2}{\pi}}$ かつ $\mathbb{V}[Z \mid Z \geq 0] = 1 - \dfrac{2}{\pi}$ になります。

2.3　切断正規分布の問題

例題1

ある大学のある学部には、数学の得点で上位 20% の成績を取った受験者を合格にする「特別合格」という制度がある。受験者の数学の得点 X が正規分布 $N(50, 10^2)$ に従うとみなして、以下の各問に答えよ。

[1]　特別合格者の数学の最低点はいくらか。

[2]　特別合格者の数学の得点の分布の確率密度関数を、標準正規分布の確率密度関数 $\phi(z)$ を用いて示せ。

[3]　特別合格者の数学の得点の期待値および分散はそれぞれいくらか。

3　混合正規分布

3.1　混合正規分布の定義

n 個の正規分布 $N(\mu_1, \sigma_1^2), \cdots, N(\mu_n, \sigma_n^2)$ を混合比 $\pi_1, \cdots, \pi_n$ で混ぜることでできる分布を**混合正規分布**といいます。厳密には、以下の確率密度関数によって定義される分布です。

$$f(x) = \sum_{i=1}^{n} \pi_i f_i(x; \mu_i, \sigma_i^2)$$

ここで、右辺に現れる関数 $f_i(x; \mu_i, \sigma_i^2)$ は正規分布 $N(\mu_i, \sigma_i^2)$ の確率密度関数とします。

3.2　混合正規分布の問題

例題2

ある大学のデータサイエンス学部では、入学時に統計学のリテラシーを測る試験を学生に実施している。入試にはA方式とB方式があり、どちらの方式で入学したかによって学生のリテラシーは異なると考えられている。

A方式で入学する学生の試験の得点の分布を正規分布 $N(60, 10^2)$、B方式で入学する学生の試験の得点の分布を正規分布 $N(70, 10^2)$ とみなし、A方式とB方式で学生の比率は $1:1$ であるとする。以下の各問に答えよ。

[1]　A方式とB方式をまとめた学生全体の試験の得点の分布の確率密度関数 $f(x)$ を示せ。

[2]　上問 [1] の分布の期待値と分散を求めよ。

[3]　確率密度関数 $f(x)$ の1階導関数 $f'(x)$ と2階導関数 $f''(x)$ を求めよ。

[4]　上問 [1] の分布が二峰性を示すかを答えよ。

4　多変量正規分布

4.1　多変量正規分布の定義と基本的な性質

n 個の確率変数の組 $X = (X_1, \cdots, X_n)^T$ が多変量正規分布 $N(\mu, \Sigma)$ に従うとは、ベクトル μ と正定値対称行列 Σ によって定まる確率密度関数

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

を持つことと定義されます。

重要な性質を三つ紹介します。

- 期待値は $\mathbb{E}[X] = \mu$、分散共分散行列は $\mathbb{V}[X] = \Sigma$ になる。
- 多変量正規分布の周辺分布もまた多変量正規分布になる。

- $X = (X_1, \cdots, X_n)^T$ が多変量正規分布に従うとき、線形和 $c_1X_1 + \cdots + c_nX_n$ は正規分布に従う。

Remark　行列 Σ について、最初は「正定値」という言葉は気にせず、各変数の分散と共分散を正方形状に並べた行列だと思うだけで構いません。

4.2　多変量正規分布の基本的な問題

例題3

S大学では入学後、数学が苦手な学生のための補講を実施している。学生全員に数学のテストを2回実施し、これらの得点を参考に補講に呼ぶ学生を決めている。学生の1回目のテストの得点を X、2回目のテストの結果を Y としたとき、$\begin{pmatrix} X \\ Y \end{pmatrix}$ は2変量正規分布

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 50 \\ 50 \end{pmatrix}, \begin{pmatrix} 10^2 & 50 \\ 50 & 10^2 \end{pmatrix} \right)$$

に従うとみなして、以下の各問に答えよ。

[1]　I先生は2回のテストの合計得点 $Z = X + Y$ を参考に補講に呼ぶ生徒を決めようと提案した。Z が従う分布を求めよ。

[2]　U先生は2回のテストのうち高い方の得点 $M = \max(X, Y)$ を参考に補講に呼ぶ生徒を決めようと提案した。M の確率密度関数 $f(m)$ を標準正規分布の確率密度関数 $\phi(z)$ と累積分布関数 $\Phi(z)$ を用いて表せ。

Remark　[2] は計算が重たいので注意してください。

4.3　多変量正規分布の条件付き分布の公式

多変量正規分布の**条件付き分布の公式**は、発展的な統計学の話題を学ぶときに、とても使用頻度の高い重要な公式です。例えば、ワンルームマンションの面積を X_1、家賃を X_2 として、これらが2変量正規分布に従っているとしましょう。このとき、面積が具体的に $X_1 = x_1$ 平米の物件に限って、家賃 $Y \mid X_1 = x_1$ の分布を知りたい。このような問題に役立つのが条件付き分布の公式です。

定理　多変量正規分布の条件付き分布の公式

確率変数の組 X を $X = (X_1, X_2)^T$ と表すことにする。特に、期待値が $\mathbb{E}[X_1] = \mu_1, \mathbb{E}[X_2] = \mu_2$、分散が $\mathbb{V}[X_1] = \Sigma_{11}, \mathbb{V}[X_2] = \Sigma_{22}$、共分散が $Cov[X_1, X_2] = \Sigma_{12}$ の多変量正規分布とする。つまり、

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \right)$$

です。このとき、条件 $X_1 = x_1$ の下での X_2 の分布は以下のようになる。

$$X_2 \mid X_1 = x_1 \sim N\left(\mu_2 + \Sigma_{12}^T\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}\right)$$

この公式は一目、難しく見えるかもしれません。しかし、期待値の部分を面積 X_1 で家賃 X_2 を予測する線形回帰と解釈するとわかりやすいのではないでしょうか。つまり、面積が $X_1 = x_1$

のワンルームの家賃は、面積 x_1 が全体平均 μ_1 に比べて広ければ広いほど高くなる

$$\mathbb{E}[\text{家賃}X_2 \mid \text{面積}X_1 = x_1] = \text{家賃の全体平均}\mu_2 + \Sigma_{12}^T\Sigma_{11}^{-1}(\text{面積}x_1 - \text{面積の全体平均}\mu_1)$$

と考えるのです。

分散の部分も同様に、面積 X_1 の値を絞った物件のみを考慮すれば、家賃 X_2 の値のばらつきも小さくなる

- ワンルーム全体の家賃の分散 : Σ_{22}
- 面積 $X_1 = x_1$ のワンルームに絞った場合の家賃の分散 : $\Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}$

という様子を表していると解釈すれば、わかりやすいでしょう。

Remark　条件付き期待値に現れる係数 $\Sigma_{12}^T\Sigma_{11}^{-1}$ もわかりやすく解釈することができます。

4.4　条件付き分布の公式を応用する問題

例題4

S大学では入学後、学生全員に数学のテストを2回実施し、これらの得点を参考に補講に呼ぶ学生を決めている。学生の1回目のテストの得点を X、2回目のテストの結果を Y としたとき、$(X, Y)^T$ は2変量正規分布

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 50 \\ 50 \end{pmatrix}, \begin{pmatrix} 10^2 & 50 \\ 50 & 10^2 \end{pmatrix}\right)$$

に従うとみなして、以下の各問に答えよ。

[1]　$X = 70$ である学生の Y の分布を求めよ。

[2]　$X \geq 50$ である学生の X の条件付き期待値 $E[X \mid X \geq 50]$ を求めよ。

[3]　$X \geq 50$ である学生の Y の条件付き期待値 $E[Y \mid X \geq 50]$ を求めよ。

[4]　上問 [2] と上問 [3] の値を比較し、このような現象を何というか答えよ。