

.....2.1.....

統計モデル

ランダムな試行を考える．試行の可能な結果の集まりを Ω と記し、これを標本空間と呼ぶ．標本空間上に確率ベクトル $\mathbf{X} = (X_1, X_2, \dots, X_n)$ を定義する¹⁾． $\omega (\omega \in \Omega)$ を試行の結果としたとき、 $\mathbf{X}(\omega)$ を観測値、実現値または標本と呼ぶ．観測できるのはランダム試行の結果である $\mathbf{X}(\omega)$ だけなので、 $\mathbf{X}(\omega)$ の確率分布を考える必要がある．この確率分布は $\mathbb{R}^n$ 上の確率分布のある族 $\mathcal{P}$ の属していると仮定する．このとき、 $\mathcal{P}$ を $\mathbf{X}(\omega)$ の**統計モデル** という．与えられた観測値は $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ 上のある（未知の）確率分布にしたがう確率変数 $\mathbf{X}$ の観測値とみなし、その観測値に基づきその確率分布についてのなんらかの判断を行うことを**統計的推測**という．

統計モデル $\mathcal{P}$ を記述するために、統計モデルに属する確率分布に添え字をつけることを考えよう．すなわち、添え字の空間 Θ から統計モデル $\mathcal{P}$ への写像 $\theta \rightarrow P_\theta (\theta \in \Theta)$ を考え、

$$\mathcal{P} = \{P_\theta \text{ は } \mathbf{X} \text{ の確率分布} : \theta \in \Theta\}$$

とする．このとき、 θ を母数と呼び、 Θ のことを母数空間と呼ぶ．

例 2.1 計測がスカラー値の n 個の観測 $x_1, x_2, \dots, x_n$ を独立同一に未知の確率分布関数 F に従う確率変数 $X_1, X_2, \dots, X_n$ の実現値としてモデル化することを考える．このことを

$$\mathbf{X} \equiv (X_1, X_2, \dots, X_n) \sim P_\theta, \quad \theta \in \Theta$$

と表記することにする．

(a) φ を標準正規分布の確率密度関数とする． $\mathbf{X}$ の確率分布 $P_\theta (\theta = (\mu, \sigma))$ は確率密度関数 $\sigma^{-n} \prod_{i=1}^n \varphi((x_i - \mu)/\sigma)$ を持ち、母数空間は $\Theta_1 = \mathbb{R} \times \mathbb{R}^+$ で $\mathcal{P}_1 = \{P_\theta : \theta \in \Theta_1\}$ とする．

(b) 母数空間を

$$\Theta_2 = \{(\mu, G) : \mu \in \mathbb{R}, G \text{ は確率密度関数 } g \text{ を持ち, } \int xg(x) dx = 0 \text{ をみたす.}\}$$

とし、確率分布 P_θ は累積分布関数 $\prod_{i=1}^n G(x_i - \mu)$ をもち、 $\mathcal{P}_2 = \{P_\theta : \theta \in \Theta_2\}$ とする．

特に、 $X_1, X_2, \dots, X_n$ が互いに独立に同一の確率分布 P にしたがうとき、 $X_1, X_2, \dots, X_n$ は分布 P からの**大きさ n のランダム標本**という．

例 2.1 の (a) のように母数空間がユークリッド空間の部分空間で、母数 θ がわかれば、その確率分布が完全に特定できる場合、その統計モデルをパラメトリックモデルと呼ぶ．(b) のようなモデルをノンパラメトリックモデルという．

¹⁾確率空間 $(\Omega, \mathcal{F}, \mathbb{P})$ 上の確率変数を $X_1, X_2, \dots, X_n$ とする．

(b) において，母数空間を

$$\Theta_3 = \{(\mu, G) : \mu \in \mathbb{R}, G \text{ は確率密度関数 } g \text{ を持つ.}\}$$

とする．このとき， $\mu_1 = 0, \mu_2 = 1$ とし， G_1, G_2 は確率密度関数 $\varphi(x), \varphi(x+1)$ を持つとすれば， $P_{(\mu_1, G_1)}$ と $P_{(\mu_2, G_2)}$ は標準正規分布となる．一般に， $\theta_1, \theta_2 \in \Theta$ に対し， $\theta_1 \neq \theta_2$ であるにも関わらず， $P_{\theta_1} = P_{\theta_2}$ であるとき，この母数化は**認定不可能**という．逆に， $\theta_1 \neq \theta_2$ ならば， $P_{\theta_1} \neq P_{\theta_2}$ であるとき，この母数化は**認定不可能**という．以後は認定可能な母数化のみを考える．