確率統計基礎講義 A

若木宏文

wakaki@hiroshima-u.ac.jp
http://home.hiroshima-u.ac.jp/ wakaki/lecture/index.shtml

2019.4.10

目 次

0. 序論
何故分布論が必要か？
統計量の分布
期待値と特性値
特性関数

目的

確率分布&標本分布の基礎について勉強する.

- 連続型分布のみを取り扱う.
- 確率密度関数，特性関数，平均，分散，歪度，劣度，乱数発生法，確率変数を用いた定義等を扱う
- 標本分布としては，２次形式統計量の分布，一元配置分散分析での標本分布等を扱う.
- 特性関数の性質は周知の事実として扱う.
- 行列とベクトルの転置は，$'$ で表す.

0. 序論

何故分布論が必要か？

分布論は統計的推測のために必要
今，データ $x_1, \ldots, x_n$ に対して統計モデルを想定する.
$\Rightarrow$
データは確率変数 $X_1, \ldots, X_n$ の実現値

$$\mathrm{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$
$$= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, \ldots, u_n; \theta)) du_1 \ldots du_n$$
$\theta \in \Omega$ (f : 既知,θ: 未知)

次のような統計的推測を行うために，f を仮定して使用する.

推定

関数 $\hat{\theta} : (x_1, \dots, x_n) \in \mathbb{R}^n \mapsto \theta \in \Omega$

$\to$ **推定値**：$\hat{\theta}(x_1, \dots, x_n)$，**推定量**：$\hat{\theta}(X_1, \dots, X_n) = \hat{\theta}$

推定量の良さ（悪さ）の尺度

偏差： $\mathrm{E}[\hat{\theta}] - \theta$
分散： $\mathrm{Var}[\hat{\theta}]$
平均２乗誤差： $\mathrm{E}[(\hat{\theta} - \theta)^2]$
集中確率：$\mathrm{P}(\theta - \varepsilon \leq \hat{\theta} \leq \theta + \varepsilon)$

これらの尺度を求めるために，f を用いて $\hat{\theta}$ の標本分布を導出する必要がある．

標本分布が分かれば，θ の信頼区間を導出することができる．$\mathrm{P}(a(X_1, \dots, X_n) \leq \theta \leq b(X_1, \dots, X_n)) \geq 1 - \alpha$

仮設検定

仮説検定問題

帰無仮説 $H: \theta \in \Omega_0$ vs 対立仮説 $K: \theta \in \Omega_1 = \Omega - \Omega_0$

検定方法

$$T(x_1, \ldots, x_n) \geq c \quad \Rightarrow \quad H \text{ を棄却}$$

閾値 c の決め方：有意水準の条件を満たすように決める
T の良さ：検出力が大きいほど良い

$\Rightarrow$ 検定統計量 $T = T(X_1, \ldots, X_n)$ の標本分布が必要

仮説検定（続き）

有意水準の条件

指定された α (通常は $\alpha = 0.05$ or 0.01) に対して

$$\sup_{\theta \in \Omega_0} \mathsf{P}(T \geq c) \leq \alpha$$

検出力（$= 1 -$ 第２種の過誤確率 ）

$$\beta(\theta) = \mathsf{P}(T \geq c) \quad (\theta \in \Omega_1)$$

第１種の過誤

帰無仮説が正しいのに，棄却してしまうこと

第２種の過誤

帰無仮説が正しくないのに，棄却しないこと

統計量の分布

統計量：標本 $X_1, \ldots, X_n$ の関数として表される確率変数．
推定量や検定統計量など．
$T = T(X_1, \ldots, X_n)$ と書く．

T の分布：$\nu(A) = \mathrm{P}(T \in A)$ によって定義される確率測度．
分布関数や確率密度関数によって一意に定められる．

$$\mathrm{P}(T \leq x) = F(x) \quad (F(x)：\text{分布関数})$$
$$= \int_{-\infty}^{x} f(t)dt \quad (f(x)：\text{確率密度関数})$$

注. T の分布が，G であることを $T \sim G$ と書く．G には分布の名前，分布関数，確率密度関数などが入る．
$X_1, \ldots, X_n \overset{i.i.d.}{\sim} G$ と書くと，$X_1, \ldots, X_n$ が独立で，それぞれ分布 G に従うことを表す．

期待値

定義

$X \sim f(x)$ (確率密度関数) とする. 関数 g に対して

$$\mathsf{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$$

を $g(X)$ の**期待値**という.

平均：$\mu = E[X]$ n
r 次のモーメント：$\mu_r = E[X^r]$
平均 (μ) 周りの r 次のモーメント：$\alpha_r = \mathsf{E}[(X - \mu)^r]$

分布の特性量

定義（分布の形状を表す統計量）

1. 平均：μ. 分布の重心，または中心を表す.
2. 分散：$\sigma^2 = \alpha_2$. 分布の散らばりを表す. $\sigma = \sqrt{\sigma^2}$ を標準偏差という
3. 歪度：$\kappa_3 = \alpha_3/\sigma^3$. 分布の歪みを表す.
   左右対称な分布 $\Rightarrow \kappa_3 = 0$
   右（左）に歪んだ分布，裾が右側（左側）に広がっている $\Rightarrow \kappa_3 > 0\ (< 0)$
4. 尖度：$\kappa_4 = \alpha_4/\sigma^4 - 3$. 標準偏差を 1 に固定したときの分布の裾の重さ，または尖り具合. 尖度が大きくなる程，裾が思い。$\kappa_4 \geq -2$ であり，等号は 1 点に退化した分布.

注. 正規分布では $\kappa_3 = 0, \kappa_4 = 0$ なので，歪度と尖度は正規分布からの逸脱度を表す指標としても用いられる.

特性関数

定義（特性関数）
$X \sim f(x)$（確率密度関数）とする.

$$\mathsf{E}[\exp(itX)] = \int_{-\infty}^{\infty} \cos(tx) f(x) dx + i \int_{-\infty}^{\infty} \sin(tx) f(x) dx$$

を X の特性関数といい，$\mathcal{C}(t)$ で表す。
（$\mathcal{C}_X(t)$ と書くこともある.）

一意性定理
X_1, X_2 の特性関数を $\mathcal{C}_1, \mathcal{C}_2$ とするとき

$\mathcal{C}_1 = \mathcal{C}_2 \quad \Leftrightarrow \quad X_1 \sim X_2$ （分布が同じということを表す）

モーメント公式

$$\mathsf{E}[|X|^r] < \infty \quad \Rightarrow \quad \mathsf{E}[X^r] = \frac{1}{i^r} \frac{d^r}{dt^r} \mathcal{C}(t) \bigg|_{t=0}$$

乱数生成

$F(x)$ を狭義の単調増加関数,
$U \sim U(0,1)$（区間 (0,1) 上の一様分布）とする.

$$\mathsf{P}(U \leq u) = \begin{cases} 1 & (u \geq 1) \\ u & (1 > u \geq 0) \\ 0 & (0 > u) \end{cases}$$

このとき $X = F^{-1}(U) \sim F(x)$ が成り立つ.
$\because$

$$\mathsf{P}(F^{-1}(U) \leq x) = \mathsf{P}(U \leq F(x)) = F(x)$$

注. $U(0,1)$ に従う確率変数（一様乱数）を計算機上で擬似的に生成すれば, F に従う確率変数を生成できる.

モンテカルロシミュレーション

$X \sim f(x)$（確率密度関数）とする.
$X_1, X_2, \ldots \overset{i.i.d.}{\sim} f(x)$ ならば $\mathsf{E}[g(X)] < \infty$ のとき

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} \mathsf{E}[g(X)] \text{ 大数の法則}$$

擬似的に，$X_1, X_2, \ldots$ を生成できれば $f(x)$ が分からなくても期待値の近似値が計算できる。

$$g(x) = \begin{cases} 1 & (x \leq u) \\ 0 & (x > u) \end{cases}$$

$$\Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} \mathsf{E}[g(X)] = \mathsf{P}(X \leq u) = F(u) \text{（分布関数）}$$