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

2019.6.5

目 次

4.2. F-分布

$\boldsymbol{\mu}$ の信頼区間

$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}, \Sigma)$ のとき, $\boldsymbol{\mu}$ の信頼領域を考える.
Σ が既知の場合：$\bar{\boldsymbol{X}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i$ とすると, $\bar{\boldsymbol{X}}$ は平均 $\boldsymbol{\mu}$ の推定量である. このとき

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \sim \chi_p^2$$

である. χ_p^2 の上側 5%点を $\chi_p^2(0.05)$ とすると

$$\mathrm{P}\Big(n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \le \chi_p^2(0.05)\Big) = 0.95$$

であるから,
$\boldsymbol{\mu} \in D := \{\boldsymbol{\mu} \,|\, n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \le \chi_p^2(0.05)\}$ が成り立つ確率が 95% となる.

区間 D は $\boldsymbol{\mu}$ の信頼係数 95%の信頼領域となる.

$\boldsymbol{\mu}$ の信頼領域 (続き)

Σ が未知の場合：Σ を, その推定量 $S = (n-1)^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$ で置き換えた領域を考える.

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' S^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

は n を大きくすると, χ^2_p に分布収束するので, n が大きければ,

$$\{\boldsymbol{\mu} \,|\, n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' S^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \chi^2_p(0.05)\}$$

は信頼係数がほぼ 95%の信頼領域となる.

n が大きくない場合に信頼係数が 95% となるような区間を求めるためには, $\chi^2_p(0.05)$ を T^2 の分布の上側 5%点で置き換える必要がある.

定義

F-分布

定義 X と Y が独立で, $X \sim \chi^2_m$, $Y \sim \chi^2_n$ であるとき

$$U = \frac{X/m}{Y/n}$$

の従う分布を自由度 m, n の F-分布といい, $U \sim F_{m,n}$ と書く.

※ $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \overset{i.i.d.}{\sim} N_p(\boldsymbol{\mu}, \Sigma)$ のとき

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' S^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})$$

とすると

$$\frac{n-p}{(n-1)p} T^2 \sim F_{p,n-p}$$

F-分布の確率密度関数

$U \sim F_{m,n}$ の確率密度関数は

$$
f(u;m,n) = \begin{cases} \dfrac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\dfrac{m}{n}\right)^{m/2} u^{m/2-1} \left(1+\dfrac{m}{n}u\right)^{-(m+n)/2} & (u>0) \\ 0 & (u \leq 0) \end{cases}
$$

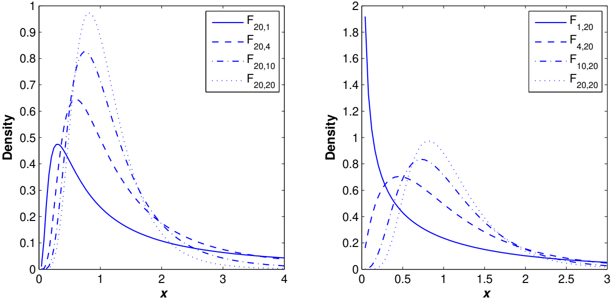
$$
(m, n \in \mathbb{N})
$$

注 4.2.1) この確率密度関数によって $F_{m,n}$ を定義してもよい．この場合, m, n は自然数でなくてもよい．

Memo

F-分布の特性

1. $F_{m,n}$ は 常に右に歪んだ分布であり, m が大きくなれば対称な分布に, n が大きくなれば 自由度 m のカイ 2 乗分布を m で割った分布に近づく.



分布の特性（続き）

2. $U \sim F_{m,n}$ のとき

$$\mathsf{E}[U] = \frac{n}{n-2} \ (n \geq 3), \ \mathsf{Var}[U] = \frac{2n^2(n+m-2)}{m(n-2)^2(n-4)} \ (n \geq 5)$$

$n \geq 2r+1$ ならば

$$\mathsf{E}[U^r] = \left(\frac{n}{m}\right)^r \frac{m(m+2)\cdots\{m+(2(r-1))\}}{(n-2)(n-4)\cdots(n-2r)}$$

3. $T \sim t_n$ のとき $T^2 \sim F_{1,n}$ である.

レポート問題 6

特性 2, 3 を示せ.

4.3. 一元配置分散分析

問題設定

データ: ある湖の 4 地点で取られた酸素溶解量の無作為標本

地点	酸素溶解量 (%)						平均
A_1	7.8	6.4	8.2	6.9			7.324
A_2	6.7	6.8	7.1	6.9	7.3		6.960
A_3	7.2	7.4	6.9	6.4	6.5		6.880
A_4	6.0	7.4	6.5	6.9	7.2	6.8	6.800

問題: 4 地点における酸素融解量に有意な差はあるか?

問題設定（続き）

モデル化: 要因 A について k 個の水準 $A_1, \ldots, A_k$ を考え，それぞれの水準における次の標本が得られるとする．

$$\begin{array}{ccc} \text{第 1 水準} & : & Y_{11}, \ldots, Y_{1n_1} \\ \vdots & & \vdots \\ \text{第 } k \text{ 水準} & : & Y_{k1}, \ldots, Y_{kn_k} \end{array},$$

統計モデル:

$$Y_{i1}, \ldots, Y_{in_i} \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2) \quad (i = 1, \ldots, k)$$

(水準が異なれば独立)

仮説検定問題:

$$H_0 : \mu_1 = \cdots = \mu_k \text{ (帰無仮説)}$$
$$H_1 : \mu_1, \ldots, \mu_k \text{ のうちどれかは異なる (対立仮説)}$$

検定統計量

郡内平方和：$\bar{Y}_i = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$ とおくと

$$W = \min_{(\mu_1,\ldots,\mu_k)\in\mathbb{R}^k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij}-\mu_i)^2 = \sum_{i=1}^{k} \min_{\mu_i\in\mathbb{R}} \sum_{j=1}^{n_i} (Y_{ij}-\mu_i)^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij}-\bar{Y}_i)^2$$ (郡内平方和と呼ぶ)

平方和：$\bar{Y} = n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$ $(n = \sum_{i=1}^{k} n_i)$ とおくと

$$S_0^2 = \min_{\mu\in\mathbb{R}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij}-\mu)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij}-\bar{Y})^2$$ (平方和)

検定統計量（続き）

平方和の分解：

$$B = S_0^2 - W = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\{(Y_{ij}-\bar{Y})^2 - (Y_{ij}-\bar{Y}_i)^2\}$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})(2Y_{ij} - \bar{Y} - \bar{Y}_i) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$$ (群間平方和と呼ぶ)

$$S_0^2 = B + W = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

が成り立つ.

検定統計量（続き）

定理 4.3.1 B と W は独立であり $W/\sigma^2 \sim \chi^2_{n-k}$. H_0 が正しいときは $B/\sigma^2 \sim \chi^2_{k-1}$ が成り立つ. 従って H_0 の下で

$$T = \frac{B/(k-1)}{W/(n-k)} \sim F_{k-1,n-k}$$

※ H_0 が正しいとき, $(k-1)^{-1}B$, $(n-k)^{-1}W$ はともに σ^2 の不偏推定量となるが, H_0 が正しくない (H_1 が正しい) ときは $\mathrm{E}[(k-1)^{-1}B] > \sigma^2$ であることがわかる.

$T > c$ のとき, H_0 を棄却する検定方法を考えることができる. c として, $F_{k-1,n-k}$ の上側 5%点を得れば，優位水準 5% の検定となる.

平方和の２次形式による表現

$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ $(i = 1, \ldots, k)$ とし,

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_k \end{pmatrix}, \quad D = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{n_k} \end{pmatrix}$$

とすると

$$(D'D)^{-1}D'\boldsymbol{Y} = (\bar{Y}_1, \ldots, \bar{Y}_k)'$$
$$D(D'D)^{-1}D'\boldsymbol{Y} = (\mathbf{1}'_{n_1}\bar{Y}_1, \ldots, \mathbf{1}'_{n_k}\bar{Y}_k)'$$

より

$$W = \boldsymbol{Y}'(I_n - D(D'D)^{-1}D')\boldsymbol{Y}$$

Memo

B と W の独立性

$Q = D(D'D)^{-1}D', P = \dfrac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ とおくと

$$PY = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\boldsymbol{Y} = (\mathbf{1}_{n_1}'\bar{Y}, \ldots, \mathbf{1}_{n_1}'\bar{Y})$$

$$B = (Q\boldsymbol{Y} - P\boldsymbol{Y})'(Q\boldsymbol{Y} - P\boldsymbol{Y}) = \boldsymbol{Y}'(Q-P)\boldsymbol{Y}$$

となる.

$\varepsilon_{ij} = Y_{ij} - \mu_i \ (i = 1, \ldots, k; \ j = 1, \ldots, n_i)$,

$\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{kn_k})$, $\boldsymbol{\mu} = (\mathbf{1}_{n_1}'\mu_1, \ldots, \mathbf{1}_{n_k}'\mu_k)'$

とおくと $\varepsilon_{11}, \ldots, \varepsilon_{kn_k} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ で,

$$B = (\boldsymbol{\varepsilon} + \boldsymbol{\mu})'(Q-P)(\boldsymbol{\varepsilon} + \boldsymbol{\mu}),$$

$$W = (\boldsymbol{\varepsilon} + \boldsymbol{\mu})'(I_n - Q)(\boldsymbol{\varepsilon} + \boldsymbol{\mu}) = \boldsymbol{\varepsilon}'(I_n - Q)\boldsymbol{\varepsilon}$$

となり, $(I_n - Q)(Q - P) = O_{n\times n}$ となるから,
B と W は独立となる.

Memo

Memo

検定結果

酸素溶解量のデータで仮説検定を行ってみると,
$n = 20$, $k = 4$ で, $F_{3,16}$ の上側 5%点は, 3.2389 である.

$$T = \frac{0.7180/3}{4.2675/16} = 0.8973$$

となるので, 有意水準 5%で仮説を棄却できない.