応用統計学 2018 第2回

2018 年 10 月 10 日 (水) ver. 1
清 智也 sei@mist.i.u-tokyo.ac.jp
http://www.stat.t.u-tokyo.ac.jp/~sei/lec-j.html

2 推測統計の基礎，ブートストラップ法

推測統計の基本的な考え方を説明する。詳細は第 6 回以降の講義で扱うことにし，今回は計算機を使った強力な技法であるブートストラップ法をまず習得することを目指す。ブートストラップ法は，今後扱う多変量解析法においても有効な手法である。

2.1 母集団と標本

統計学の一つの役割は，母集団 (population) から標本 (sample) を抽出し，標本の特徴に基づいて母集団の特徴を推し測ることである。これが推測統計である。

母集団から標本を抽出するという手続きは，電話による世論調査のような状況では想像しやすいが，物理実験で得られる測定値のようなデータに対しては意味がはっきりしない。そのような場合，以下のように「確率モデル」を導入して数学的に扱った方が考えやすい。すなわち，データ $x_1, \ldots, x_n$ はある確率分布に従う独立な確率変数の実現値と考えるのである[1]。この確率分布を母集団分布 (population distribution) と呼ぶ。より正確には，測定値に対応する確率変数をそれぞれ $X_1, \ldots, X_n$ としたとき，各 X_i が区間 $(a_i, b_i] \subset \mathbb{R}$ に入る同時確率を

$$P(X_1 \in (a_1, b_1], \ldots, X_n \in (a_n, b_n]) = \prod_{i=1}^{n} (F(b_i) - F(a_i)) \tag{1}$$

と規定する。ここで $F(x)$ は母集団分布の累積分布関数（2.2 節）である。F から $x_1, \ldots, x_n$ が生成されることをデータ生成過程 (data generating process) ともいい，模式的に

$$F \quad \longrightarrow \quad x_1, \ldots, x_n$$

と表すこともある。世論調査のような場合でも，母集団の大きさが非常に大きいと考えれば，式 (1) の仮定で話を進めることができる。

このような母集団分布をわざわざ導入する最大の利点は，標本から得られる代表値の「誤差」について議論できることにある。それが 2.3 節で扱う標準誤差である。

[1] 簡単のため 1 変数を想定している。

2.2 確率論の復習

確率論の復習を簡単に行う。キーワードは確率空間，確率変数，期待値，独立性である。

まず確率空間 (probability space) とは (Ω, P) という 2 つ組である[2]。ここで Ω は標本空間と呼ばれる集合である。P は確率測度と呼ばれる写像で，Ω の各部分集合 A に対して $P(A) \in [0,1]$ という値を返す。確率測度が満たすべき公理は $P(\Omega) = 1$ という性質と，完全加法性と呼ばれる次の性質である：集合 $A_1, A_2, \ldots \subset \Omega$ が互いに排反ならば $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

確率変数 (random variable) とは，Ω から $\mathbb{R}$ への写像のことである。確率変数 X の累積分布関数は

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\})$$

によって定義される。F のことを単に分布と呼ぶこともある[3]。累積分布関数が $F(x) = \int_{-\infty}^{x} f(\xi) d\xi$ と表されるとき，$f(x)$ を X の確率密度関数と呼び，X は連続的であるという。また累積分布関数が $F(x) = \sum_{\xi \leq x} f(\xi)$ と表されるとき，$f(x)$ を X の確率関数と呼び，X は離散的であるという。

X の期待値 (expectation) は，累積分布関数 F を用いて

$$E[X] = \int_{-\infty}^{\infty} xF(dx)$$

と定義される。右辺の $F(dx)$ の意味は，リーマン積分の定義と同様に $\sum_{i=1}^{n} x_i(F(b_i) - F(a_i))$ の形の和の極限と考えればよい。連続分布の場合の期待値は $\int_{-\infty}^{\infty} xf(x)dx$，離散分布の場合は $\sum_x xf(x)$ と表される。期待値は推測統計の文脈では母平均 (population mean) とも呼ばれる。

次の定理は期待値の重要な性質を表している[4]。

定理 2.1. X の分布を F とし，h を $\mathbb{R}$ から $\mathbb{R}$ への関数とするとき，

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)F(dx)$$

が成り立つ。また 2 つの確率変数 X, Y に対して

$$E[X + Y] = E[X] + E[Y]$$

が成り立つ（期待値の線形性）。

確率変数 X と Y が独立 (independent) であるとは $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ が任意の $A, B \subset \mathbb{R}$ に対して成り立つことと定義する。ここまでの定義が式 (1) と整合的であることを確認されたい。次の性質はほぼ無意識に用いられる。

[2]本当は σ-加法族 $\mathcal{F}$ を含めた 3 つ組 $(\Omega, \mathcal{F}, P)$ であるが，今は $\mathcal{F}$ を意識しなくても問題ない。
[3]正確には，X によって $\mathbb{R}$ に誘導される確率測度のことを X の分布という。
[4]離散分布以外の場合は証明が厄介である。ルベーグ積分論を習うと見通しよく証明できる。

定理 2.2. X と Y が独立ならば，任意の関数 f, g に対して $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ が成り立つ。

確率変数 X の分散は

$$V[X] = E[(X - E[X])^2]$$

と定義される。推測統計の文脈ではこの量を母分散 (population variance) と呼ぶ。ここまでの定義や定理を用いれば，確率変数 X と Y が独立なとき，$V[X+Y] = V[X] + V[Y]$ が成り立つことが示される。

記述統計において代表値として定義された歪度や尖度などは，確率変数に対して次のように定義される：

$$\frac{E[(X - E[X])^3]}{\sigma^3}, \quad \frac{E[(X - E[X])^4]}{\sigma^4} - 3.$$

ただし $\sigma = \sqrt{V[X]}$ とおいた。逆に，このように期待値を使って定義される特性値 (characteristic) は，期待値を標本平均に置き換えることで 1 つの統計量となる。

確率論の復習事項としては，以上の他にモーメント母関数や大数の法則，中心極限定理，多次元確率分布などがあるが，これらは必要に応じて後日補足していく。

2.3 標準誤差

1 次元の量的データ $\boldsymbol{x} = (x_1, \ldots, x_n)'$ を考える。前回述べたように，データの代表値は

$$t(x_1, \ldots, x_n)$$

という形の関数で表される。たとえば平均値は $t(x_1, \ldots, x_n) = n^{-1} \sum_{i=1}^n x_i$ などと書ける。

$X_1, \ldots, X_n$ を独立同分布に従う確率変数の列とすると，統計量

$$T = t(X_1, \ldots, X_n)$$

も確率変数となる。この T が従う分布のことを標本分布 (sampling distribution) という。また T の標準偏差のことを標準誤差 (standard error) という。なぜこのような用語を新しく導入するかというと，統計量 T は母集団特性値を推定するために用いられるからである。具体的には標本平均は母平均を推定しようとしている。

T が標本平均の場合について，その標準誤差を求めてみよう。まず T の分散は

$$\begin{aligned} V[T] &= V\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] \\ &= \frac{1}{n^2} V\Big[\sum_{i=1}^{n} X_i\Big] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} V[X_i] \quad \text{（独立性）} \\ &= \frac{1}{n} V[X_1] \quad \text{（同分布性）} \end{aligned}$$

となる。したがって統計量 T の標準誤差は母標準偏差 σ を用いて

$$\mathrm{se}(T) = \frac{\sigma}{\sqrt{n}}$$

と表される。n が大きくなるとき標準誤差が $1/\sqrt{n}$ のオーダーで 0 に近づくことが分かる。これは重要な事実である。ところで，いまの標準誤差には σ が含まれており，これは母集団に依存する量である。そこで σ もデータから推定して

$$\widehat{\mathrm{se}}(T) := \frac{\hat{\sigma}}{\sqrt{n}} \tag{2}$$

とおく。ただし

$$\hat{\sigma} = \Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\Big)^{1/2}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

とする。この $\widehat{\mathrm{se}}(T)$ も標準誤差と呼ばれる[5]。以上を定理としてまとめておこう。

定理 2.3. 標本平均 $n^{-1}\sum_{i=1}^{n} X_i$ の標準誤差は式 (2) で与えられる。

例を示す。表 1 は喫煙の脈拍への影響を見るために，15 人の被験者における喫煙前後の 1 分間の脈拍を測定した結果である。いま喫煙前後の脈拍に有意な差が見られるかどうかに興味があるとする。このような問題は仮説検定と呼ばれ，詳しくは第 9 回の授業で扱う予定である。

表 1: 15 人の被験者の，喫煙の前後における 1 分間の脈拍。

(A) 喫煙前	70	69	72	74	66	68	69	70	71	69	73	72	68	72	67
(B) 喫煙後	69	72	71	74	68	67	72	72	72	70	75	73	71	72	69
(B) − (A)	−1	3	−1	0	2	−1	3	2	1	1	2	1	3	0	2

稲垣「数理統計学 改訂版」，問 10.1 より。

[5] $\mathrm{se}(T)$ と区別するために標本標準誤差と呼ぶ場合もある。

喫煙前後の脈拍の差を $\boldsymbol{x} = (x_1, \ldots, x_{15})$ とおく。その標本平均を求めると，

$$\bar{x} = \frac{x_1 + \cdots + x_{15}}{15} = 1.13$$

となる。この値が 0 でないからと言って直ちに「差がある」と結論付けることはできない（むしろ 0 になることの方が珍しい）。なぜならば，標本で差が見られても母集団では差がないかもしれないからである。そこで標準誤差を求めると，式 (2) から

$$\widehat{\mathrm{se}}(T) = \frac{1}{\sqrt{15}} \sqrt{\frac{1}{15} \sum_{i=1}^{15} (x_i - \bar{x})^2} = 0.36 \tag{3}$$

と算出される。0.36 は 1.13 に比べれば小さい値であるから，喫煙前後で有意な差が見られると言って良さそうである[6]。

さて，例えば標本歪度

$$T = \frac{n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^3}{\hat{\sigma}^3}$$

のような複雑な統計量の場合，その標準誤差はどのように計算すればよいだろうか？このような問題は統計学の分野では古くから研究されていたが，結果的に複雑な数式になってしまうケースが多かった。これに対しスタンフォード大学の Bradley Efron が 1970 年代後半に驚くべきアイデアを発表した。それが次節で述べるブートストラップ法である。

2.4 ブートストラップ法

実は，標準誤差（の近似値）は以下に述べるアルゴリズムでいとも簡単に計算できてしまう。ポイントは，計算機で乱数を生成することである。

ブートストラップ法 (bootstrap method) による標準誤差の計算

与えられたデータを $\boldsymbol{x} = (x_1, \ldots, x_n)$ とし，標準誤差を求めたい統計量を $t(\boldsymbol{x})$ とする。

1. $\{1, \ldots, n\}$ から独立に n 個の値を抽出し，それを $i_1, \ldots, i_n$ とおく（重複を許す）。
2. $\boldsymbol{x}^* = (x_{i_1}, \ldots, x_{i_n})$ とおく。
3. $t^* = t(\boldsymbol{x}^*)$ を計算する。
4. 以上の操作を B 回繰り返し，得られた値を $t_1^*, \ldots, t_B^*$ とおく。
5. $t_1^*, \ldots, t_B^*$ の標本標準偏差を求め，その値を標準誤差として出力する。

実用上は繰り返し回数を $B = 100$ 程度として計算する。

[6]後日扱う仮説検定の言葉を用いると，両側 t 検定における p 値は約 1%となる。

再び表 1 の例で考えよう。脈拍差の標本平均は $t(\boldsymbol{x}) = 1.13$ であった。これに対し，ブートストラップ法に従って $t_1^*, \ldots, t_{100}^*$ を計算したところ次のような結果が得られた（これは乱数に依存するので結果は毎回変わる）:

1.20,1.27,1.07,1.47,1.53,1.87,0.80,1.53,0.93,1.40,0.47,1.47,1.00,1.47,1.07,1.27,1.40,1.60,1.73,0.93,
1.13,1.07,0.60,1.00,1.00,1.00,1.87,1.73,0.40,0.73,0.47,1.47,0.87,1.00,1.27,1.07,0.67,1.40,0.73,0.93,
1.13,1.27,0.80,0.73,0.73,1.80,0.27,1.53,1.33,0.93,0.93,1.60,1.20,1.13,1.07,0.67,1.27,1.53,1.27,1.60,
1.33,1.27,1.40,1.60,1.40,1.53,1.07,0.40,1.13,1.53,1.20,0.93,0.93,1.53,0.87,1.60,0.87,1.40,1.07,1.07,
0.53,1.00,1.53,1.13,1.07,1.00,0.87,1.53,1.60,1.27,1.00,0.93,1.53,1.00,1.53,0.60,1.00,1.07,1.53,1.13

その標準偏差を求めると

$$\sqrt{\frac{1}{B}\sum_{b=1}^{B}(t_b^* - \overline{t^*})^2} = 0.35$$

となり，式 (3) で求めた値とほぼ同じとなる。以下に R プログラムの例を示す。

```
> x = c(-1, 3, -1, 0, 2, -1, 3, 2, 1, 1, 2, 1, 3, 0, 2)
> library(boot)
> my.mean = function(x, i) mean(x[i])
> boot(x, my.mean, 100)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = x, statistic = my.mean, R = 100)


Bootstrap Statistics :
    original     bias    std. error
t1* 1.133333 0.02333333   0.3523052
```

同じデータに対して，歪度の値を計算すると

$$t(\boldsymbol{x}) = \frac{1}{15\hat{\sigma}^3}\sum_{i=1}^{15}(x_i - \bar{x})^3 = -0.24$$

となる。これに対しブートストラップ標準誤差を見積もったところ 0.37 となった。これは 0.24 に比べて大きいので，母集団分布の歪度が 0 でないとは言えないことが結論される。

ブートストラップ法の正当性を示すにはそれなりの準備が必要なので省略するが，導出のアイデアは「標準誤差の計算に現れる母集団分布関数 F を経験分布関数 $\hat{F}_n$ に置き換える」という極めてシンプルなものである。模式的に表せば，データ生成過程によって統計量が生成される様子を

$$F \quad \longrightarrow \quad \boldsymbol{x} \quad \longrightarrow \quad t(\boldsymbol{x})$$

と表すとき，ブートストラップ法では

$$\hat{F}_n \quad \longrightarrow \quad \boldsymbol{x}^* \quad \longrightarrow \quad t(\boldsymbol{x}^*) \tag{4}$$

の流れで擬似的な統計量 $t(\boldsymbol{x}^*)$ を生成している。

2.5　モンテカルロ法との関係

ブートストラップ法と関連する手法であるモンテカルロ法について言及しておく。

モンテカルロ法 (Monte Carlo method) とは，複雑な確率変数の期待値を計算する方法である。これは「既知」の確率分布に従う乱数をたくさん生成し，その標本平均を求めることで期待値を推定する，というものである。アルゴリズムとして書けば次のようになる。

モンテカルロ法

X を確率変数とする。期待値 $E[X]$ の近似値を計算したい。

1. X と同じ分布に従う乱数を独立に M 回生成し，$x_1, \ldots, x_M$ とおく。
2. $\bar{x} = (x_1 + \cdots + x_M)/M$ を出力する。同時に式 (2) に基づき標準誤差も出力する。

標準誤差は $1/\sqrt{M}$ に比例するから，M を増やせば増やすほど推定精度はよくなる。実用上は $M = 10^6$ 回程度の計算をすれば満足のいく結果が得られることが多い。

例えば，$U_1, \ldots, U_5$ を区間 $[1, 2]$ 上の一様分布に従う独立な確率変数として，

$$X = f(U_1, U_2, U_3, U_4, U_5) = \cfrac{1}{U_1 + \cfrac{1}{U_2 + \cfrac{1}{U_3 + \cfrac{1}{U_4 + \cfrac{1}{U_5}}}}}$$

の期待値 $E[X]$ を計算する，という問題を考える。これを定義通り計算しようとすると複雑な重積分を計算することになるが，モンテカルロ法を用いれば簡単に近似値を求めることができる。すなわち，$\{(u_{m1}, \ldots, u_{m5})\}_{m=1}^{M}$ を $(U_1, \ldots, U_5)$ と同じ分布に従う独立な乱数列として，それぞれについて

$$x_m = f(u_{m1}, \ldots, u_{m5})$$

を計算し，最後に標本平均

$$\frac{1}{M} \sum_{m=1}^{M} x_m$$

を計算すれば，それが $E[X]$ の推定値となる。さらにその標準誤差を求めることもできる。実際に $M = 10^4$ としてモンテカルロ法を実行したところ推定値は 0.509，標準誤差は 0.001 となった。

ブートストラップ法もモンテカルロ法の考え方を利用している。ただし，ブートストラップ法における「既知の確率分布」とは母集団分布ではなく，データの「経験分布」である（式 (4) 参照)。したがって，ブートストラップ法における繰り返し回数 B をいくら増やしても，経験分布以上の情報は得られない。そのため $B = 100$ くらいで計算してもあまり支障はない。

標語的に言えば，モンテカルロ法は「確率論における計算機集約型手法」，ブートストラップ法は「推測統計における計算機集約型手法」ということができる。

用語のまとめ

- 推測統計：母集団，標本，母集団分布，データ生成過程，母平均，母分散，特性値，標本分布，標準誤差，ブートストラップ法。
- 確率論：確率空間，確率変数，期待値，独立性，モンテカルロ法。

演習問題

問題 2.1. 表 1 のデータから求まる標本歪度とその標準誤差について，ブートストラップ法を用いて計算してみよ。また 2.4 節で示した結果と比較せよ。

問題 2.2. 表 1 のデータに対し，中央値の標準誤差をブートストラップ法で計算せよ。

問題 2.3. $0 \leq k \leq n$ とする。$\{1, \ldots, n\}$ から独立に n 個の値を復元抽出したとき，その n 個の中に 1 が k 個ある確率は

$$p_k = \frac{n!}{k!(n-k)!}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k},$$

となることを示せ。この値は $n \to \infty$ のときどんな確率分布に収束するか。

問題 2.4. モンテカルロ法を用いて，次の事象が起こる確率を推定し，またその標準誤差も推定せよ。

事象：「10 個のサイコロを同時に投げたとき，現れる目の最大度数が 4 となる」

また，モンテカルロ法を用いずに計算するプログラムも書き，結果を確認せよ。