歩測の補足：統計学の初歩と誤差論

Katsuyoshi Matsushita

April 2019

1　母集団と推定

母集団 (population) とは要素からなる集合であり, それぞれの要素は測定できる値 x を持っているものとする. 母集団のもつ x の分布について情報をその部分集合 (標本, sample) から推定することが統計学の目的である. 歩測の場合は一回の歩測の試行がサンプルを構成する要素で, その集合の歩数 x(outcome) の分布から測定の目的とする距離を測ることになる. 分布の一つの特徴量として母平均 $\overline{x}$(population average) や母分散 σ_x^2(population variance) があり母数 (parameter) と呼ばれる. 母平均は母集団に属するすべての要素に対する平均値で i を母集団に属する要素の番号として i の集合を Ω で表すと,

$$\overline{x} = \frac{\sum_{i\in\Omega} x_i}{\sum_{i\in\Omega} 1} \tag{1}$$

である. 一方母分散は,

$$\sigma_x^2 = \frac{\sum_{i\in\Omega} (x_i - \overline{x})^2}{\sum_{i\in\Omega} 1} \tag{2}$$

である. これらを部分集合である標本から推定する問題を考える.

2　推定量の選び方

n 個の要素からなる標本, $X_1, X_2, X_3 \ldots x_n$, から母数を推定するとき, その推定するための量を推定量 (estimator) と呼ぶ. うまく母数を推定するためには推定量が幾つかの条件が成立していることが好ましい. その好ましいとされる条件を以下に挙げる. ここで n は標本サイズ (sample size) と呼ばれる量である.

母数を θ で表し, 標本サイズ n の標本の推定量を θ_n とする. このとき θ_n は以下の条件を満たすと良い. 一つ目は θ_n は標本に依存して値が異なり, あ

る分布 $P(\theta_n)$ を持つことに注意する．このとき, 以下で定式化される不偏性 (unbiasedness) という性質が一つ目の条件となる．

$$E(\theta_n) = \theta \tag{3}$$

ここで E(. . .) は期待値を意味し,

$$E(\theta_n) = \sum_{\theta_n} \theta_n P(\theta_n) \tag{4}$$

を表す．これは推定値の期待値が母数に一致することを意味する．

母平均に対してこの不偏性を満たす量は標本平均

$$\overline{X} = \frac{1}{N} \sum_{i}^{n} X_i \tag{5}$$

である．一方, 母分散に対しては不偏分散

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum (X_i - \overline{X})^2 \tag{6}$$

が普遍性を満たす．

次の条件は有効性 (efficiency) である．有効性とは不偏性を満たす推定量 (不偏推定量, unbiased estimator) に対する分散,

$$E((\theta_n - E(\theta_n))^2) \tag{7}$$

が最小になる量である．母平均については先に述べた標本平均が Cramér-Rao 下限から有効性を持つことが知られている．

さらに標本サイズが大きい時にその推定が正しくなる条件として一致性 (consistency) が存在する．いま母数 θ を基準にした差の分布として $P(\theta_n)$ を書き直した θ_n の分布を $P(\theta_n - \theta)$ と表すと,

$$\lim_{n \to \infty} P(\theta_n - \theta = 0) = 1 \tag{8}$$

で表される．この条件も先に述べた標本平均は満たしていることが知られ大数の法則と呼ばれる．母分散に対して一致性を持つ量として標本分散

$$\sigma_X^2 = \frac{1}{n} \sum_{i} (X_i - \overline{X})^2 \tag{9}$$

がある．したがって十分大きな標本サイズでは不偏分散の代わりに標本分散が使える．

3　統計量

母数の推定量としての標本の統計量 (statistics) を考える．統計量とは標本分布を特徴づける量であり, 標本平均などがある．統計量は先に述べた推定量として望ましい性質を持つものがあり, 母数の推定量として使われる．以下に統計量の例を挙げる．

1. モーメント (moment) として得られる基本統計量

   統計量の中には標本の測定量の確率密度からそのモーメントとして計算できる量がいくつかある．それを説明するため測定量 X が標本中に現れた回数を度数 (frequency) を考える．測定量 X に対する度数の関数 $h(X)$ をヒストグラム (histogram) と呼ぶ．$h(x)$ を標本サイズ n で割れば確率密度関数 $P(X)$ が得られる．いかでは $E(A(X))$ は X の関数 $A(X)$ の $P(X)$ での期待値

$$E(A(X)) = \sum_X XP(X) \tag{10}$$

   を指す．$P(X)$ の r 次のモーメントとは

$$\mu_r = E((X - E(X))^r) \tag{11}$$

   で定義される．

   - 平均値 (mean value) μ

$$\mu = E(X) \tag{12}$$

   - 分散 (variance) σ^2

$$\sigma^2 = \mu_2 \tag{13}$$

   - 標準偏差 (standard deviation) σ

$$\sigma = \sqrt{\sigma} \tag{14}$$

   - 歪度 (skewness) γ

$$\gamma = \frac{\mu_3}{\sigma^3} \tag{15}$$

   - 尖度 (kurtosis) β

$$\beta = \frac{\mu_4}{\sigma^4} \tag{16}$$

     もしくは

$$\beta = \frac{\mu_4}{\sigma^4} - 3 \tag{17}$$

2. 順序統計量 (order statistic)

   確率密度関数 $P(X)$ を X まで足した

$$F(X) = \sum_{Y \leq X} P(X) \tag{18}$$

   を累積分布確率とよぶ．また $P(X)$ を X 以上を足した関数も考える．

$$G(X) = \sum_{Y \geq X} P(X) \tag{19}$$

   これを使って順序統計量は以下のようなものである．

- 最大値 (maximum value) l

$$l = \max_{X, P(X)\neq 0}\{X\} \tag{20}$$

- 最小値 (minimum value) s

$$s = \min_{X, P(X)\neq 0}\{X\} \tag{21}$$

- 中央値 (median) $Q_{\frac{1}{2}}$

$$F(m) \leq \frac{1}{2} \ \text{ and } \ G(m) \geq \frac{1}{2} \tag{22}$$

- q 分位数 (q-quantile, q は 0 から 1 までの実数)

$$F(m) \leq q \ \text{ and } \ G(m) \geq 1 - q \tag{23}$$

3. その他の基本統計量

- 最頻値 (mode) m

$$m = \arg\max_{X} P(X) \tag{24}$$

推定量としての統計量に望ましい性質として十分性 (sufficiency) がある. 統計量が十分性を持つとは測定量が観測される確率はその統計量を決めた場合その測定量に寄らないことを指す. 実は特定の条件の下で十分性のある統計量の情報で不偏性を持つ推定値を与えることができる.

4 標本平均値の分布と信頼区間

標本平均 $\overline{X}$ は先に述べたように普遍推定量であり, 母平均を推定するのに使われる. そこで標本平均値がどのように分布するかが推定の良し悪しを考える上で有効になる. 標本平均 $\overline{X}$ は母集団の x の範囲の有界性などが成立した場合中心極限定理という定理が知られ $\overline{X}$ に対して

$$\lim_{n\to\infty} P(\overline{X}) = N(\overline{(}x), \frac{\sigma_x^2}{n}) \tag{25}$$

となる. ここで $N(\mu, \sigma^2)$ は平均 μ, 分散 σ^2 の正規分布

$$N(\mu, \sigma^2) = \frac{1}{2\pi\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right] \tag{26}$$

である. このように標本サイズが大きい時に正規分布に近づく性質を漸近正規性と呼び推定量に選ぶ基準とすることもある.

漸近正規性がある場合は正規分布を基に推定値がどの程度信頼できるかを推定することができる. 推定値を一つ与えるこれまで説明した方法を点推定と呼ぶのに対し, このような信頼できる領域を与える事を区間推定とよぶ. 例えば 95%の確率で $\overline{X}$ で得られる母数が入る区間もしくは優位水準 $\alpha = 0.05$ の信頼区間として,

$$\overline{X} - 1.96\sqrt{\frac{\hat{\sigma}_X^2}{n}} < \overline{x} < \overline{X} + 1.96\sqrt{\frac{\hat{\sigma}_X^2}{n}} \tag{27}$$

などが良く使われる.

5 最尤推定と最小二乗法

尤度関数とは母集団の分布が与えられたとき, その母数から標本 $X_i|i \in \Omega$ が得られる条件付き確率をその母数が得られる確率と見なした関数

$$P(\theta) = P(X_1, X_2, \ldots, X_n|\theta) \tag{28}$$

である. この確率関数が最大化されるときに母数 θ の尤 (もっと) もらしい推定値と考える. 通常対数関数の単調性から, 対数をとった対数尤度関数

$$F(\theta) = -\log P(\theta) \tag{29}$$

を最小化する θ_n を母数の推定値とする.

もし独立にとられた標本が母数に対して正規分布していると考えられると,

$$F(\theta) = -\sum_i^n \frac{(X_i - \theta)^2}{\sigma^2} + \text{const} \tag{30}$$

が対数尤度関数となる. ここから最小値を求めるために微分し 0 と置くと,

$$\frac{dF(\theta)}{d\theta} = -\frac{2}{\sigma^2}\sum_i^n (X_i - \theta) = 0 \tag{31}$$

となる. これは所謂点推定の最小二乗法である. 変形すると,

$$\theta = \frac{1}{N}\sum_i X_i \tag{32}$$

のように推定値として標本期待値が得られる.

一方で母数 θ の事前知識がある場合, 例えば母数の分布である事前確率 (prior probability)$P(\theta)$ が分かっている場合尤度ではなく次の事後確率 (posterior probability) $P(\theta|X_1, X_2, \ldots X_n)$ の最大化でより良い推定値が得られることがある.

$$P(\theta|X_1, X_2, \ldots X_n) = P(X_1, X_2, \ldots, X_n|\theta)P(\theta) \tag{33}$$

この推定はベイズ推定の一種で最大事後確率 (maximum a posteriori, MAP) 推定と呼ばれる. 事前確率が良く分かっていない場合は無情報事前確率として Jeffery 事前確率や共役事前分布 (conjugate prior) が用いられる. 尤度関数が正規分布で推定する母数が母平均の場合は, 前者は一様分布, 後者は母平均の正規分布となることが知られている.

6　測定量の関数の誤差と誤差伝搬の法則

ここまで測定値と知りたい量が一致するものとして議論してきた. 一般には知りたい量と測定量が一致せず, それが関数の場合がある. この場合知りたい量の推定値は測定量の推定値から得られる期待値とすることができるが, 誤差 (分散) については評価に注意が必要である. 何らかの測定値の関数として知りたい量 Y が得られた場合, その誤差を評価したい場合を考えよう. もし測定値の誤差 ΔX がその量の誤差 ΔY に与える影響を議論する. θ を測定値 X の関数として

$$Y = f(X) \tag{34}$$

と与えられたとする. X に対する ΔX の影響は,

$$Y \pm \Delta Y = f(X \pm \Delta X) \simeq f(X) \pm \frac{df(X)}{dX}\Delta X \tag{35}$$

のようにテーラー展開して

$$\Delta Y = \frac{df(X)}{dX}\Delta X \tag{36}$$

と評価できる.　このテーラ展開で決まる誤差の欲しい量 Y への伝播法則は誤差の伝播法則と呼ばれる.

7　2 標本の比較, 検定