歩測の補足：統計学と誤差

Katsuyoshi Matsushita

June 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, $X_1, X_2, X_3 \ldots x_n$, からうまく母数を推定するためには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, 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普遍性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普遍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.

次の条件は有効性である. 有効性とは普遍性を満たす推定量 (普遍推定量) 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,

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

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 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

$$\lim_{n \to \infty} P(\overline{X}) = N(\bar{(}x), \frac{\sigma_x^2}{n}) \tag{10}$$

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

$$N(\mu, \sigma^2) = \frac{1}{2\pi\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right] \tag{11}$$

である. ここから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. 例えば 95%信頼できる区間として,

$$\overline{X} - 1.96\sqrt{\frac{\hat{\sigma}_X^2}{n}} < \overline{x} < \overline{X} + 1.96\sqrt{\frac{\hat{\sigma}_X^2}{n}} \tag{12}$$

などが良く使われる.