統計学 2020 Lecture 7: 正規分布におけるパラメータの推定

北門　利英（東京海洋大学海洋生物資源学科）

2020 年 6 月 24 日

Attention:

- 授業 HP に授業に関する情報をアップデートしていきますので参照ください．(URL: https://toshihidekitakado.github.io/STAT2020/index.html)
- 今回およびこれまでの授業に関してわからないことがあれば，メールかリアルタイム接続時に遠慮なく質問してください（6 月 24 日も 13:00-14:00 とします）．

Point: 次の用語をしっかり理解すること．

- 正規分布におけるパラメータの推定法について理解する事
- 推定の不偏性について再確認する事

1　正規分布の復習

1.1　正規分布の定義

確率密度関数 $f(y)$ が確率分布を規定しており，特に以下の形の確率密度関数を持つとき，確率変数は正規分布にしたがうという．

連続型分布 1　確率変数 Y の確率密度関数が

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty \quad \left(\begin{array}{c} -\infty < \mu < \infty \\ 0 < \sigma < \infty \end{array} \right) \tag{1}$$

となるとき，Y は正規分布 (normal distribution) $N(\mu, \sigma^2)$ にしたがうという．

計算の詳細は割愛するが，確率変数 Y が正規分布 $N(\mu, \sigma^2)$ にしたがうとき，$E[Y] = \mu$ と $V[Y] = \sigma^2$ である．必ず覚えて下さい．

1.2　標準正規分布

性質 1　確率変数 Y が $N(\mu, \sigma^2)$ にしたがうとき，

$$Z = \frac{Y - \mu}{\sigma} \tag{2}$$

は $N(0,1)$ にしたがう．この変換を標準化 (standardization)，また $N(0,1)$ を標準正規分布 (standard normal distribution)という．

定理 1　確率変数 Y が $N(\mu,\sigma^2)$ にしたがうとき，$a+bY$ は $N(a+b\mu,b^2\sigma^2)$ にしたがう．

1.3　正規分布の再生性

正規分布にしたがう独立な確率変数の和もまた正規分布にしたがう．独立同一な n 個の正規分布にしたがう確率変数の和もまた正規分布にしたがう．これらの性質は区間推定と仮説検定のところで再度説明するので，今回は定理の存在だけ覚えておいてください．

定理 2　確率変数 Y_1,Y_2 が独立でそれぞれ $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$ にしたがうとき，aY_1+bY_2 は $N(a\mu_1+b\mu_2,a^2\sigma_1^2+b^2\sigma_2^2)$ にしたがう．

定理 3　確率変数 $Y_1,Y_2,\ldots,Y_n$ が独立同一に $N(\mu,\sigma^2)$ にしたがうとき，$\bar{Y}=\sum_{i=1}^{n}Y_i/n$ は $N(\mu,\sigma^2/n)$ にしたがう．

2　パラメータの推定

一般に，確率分布それ自身，および期待値や分散といった分布の性質は，確率分布に備わるパラメータによって決まることがわかる．観測データは確率的メカニズムを通して得られるものであり，そこに確率的要素が加わる以上，観測データから現象の真の姿やパラメータの真値を完全に言い当てることはできない．すなわち，確率的要素を起因とする不確実性を避けることはできない．しかし，推測の手段として精度の高い方法を模索し，同時に我々が行う推定の精度を客観的に推定結果に付与することで，科学的でかつ帰納的な推測となり，現実的に利用度の高い情報へと変えることができる．このように，データの情報を基にして真のパラメータやモデルに関する推測を行うことを統計的推測という．ここでは正規分布の 2 つのパラメータの推定について学ぶが，その前に統計的推定について改めて整理しておく．

パラメータの推定では，2 つの性質の重要性が浮き彫りになる． 1 つは推定量の期待値に関する性質であり，他方は推定量の散らばりに関する性質である．そこで，この推定の問題を少し一般的な形で表現してみよう．

定義 1　推定量と推定値　いま一般に，ある確率分布 $f(y|\theta)$ にしたがう確率変数のベクトル $Y=(Y_1,\ldots,Y_n)$ の値を観測することによって，確率分布に含まれるパラメータ θ を推定することを考える．簡略のためパラメータ θ は 1 次元（すなわち定数）とし，$\Theta\subset R^1$ 内を動くものとする．ここで，$R^1=(-\infty,\infty)$ である．

いま，観測値 $Y=y$ の値が得られれば推定値 $\hat{\theta}=\theta(y)$ が決まる方式を考える．このとき，θ を推定するために定義された Y の関数 $\hat{\theta}(Y)$ を θ の推定量 (estimator)，推定量の値に $Y=y$ の値を $\hat{\theta}(Y)$ に代入した数値 $\hat{\theta}=\hat{\theta}(y)$ を θ の推定値 (estimate) として区別する．この推定値がパラメータ真値 θ に近いことは確かに推定の性能としては良いことである．しかしながら，データである確率変数 Y が変動し，したがって確率変数 Y の関数である推定量も確率変数 Y の動きに連動する．このように，確率変数の関数である推定量もある種の確率分布にしたがうことになる．

よりよい推定量に求められる基準をここでは 2 つ挙げる．

定義 2　推定量の不偏性　推定量 $\theta(Y)$ が

$$E[\hat{\theta}(Y)] = \theta \quad \text{for all } \theta \in \Theta \tag{3}$$

を満たすとき，推定量 $\theta(Y)$ は θ に対して不偏 (unbiased)であるという．また，不偏である推定量を不偏推定量 (unbiased estimator)とよぶ．

定義 3　推定量の分散　推定量 $\theta(Y)$ が不偏であるならば，$\theta(Y)$ は真のパラメータ値 θ を中心にばらつく．この時，なるべくばらつきが小さい方が精度よく θ を推定できるから，$\theta(Y)$ の分散を最小化することが望ましい．

$$V\left[\hat{\theta}(Y)\right] \to \min \tag{4}$$

3　正規分布におけるパラメータの推定

3.1　観測値に対する仮定

ここでは独立かつ同一な確率変数とし正規性を仮定すると，一般に

$$Y_1, Y_2, \ldots, Y_n \sim (iid)N(\mu, \sigma^2) \tag{5}$$

と表現できる．

3.2　平均値 μ に対する推定

パラメータ μ に対して不偏でかつ分散が最小となる推定量は

$$\hat{\mu}(Y) = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{6}$$

である．

再生性のところで確認したように，標本平均 $\bar{Y}$ は

$$\bar{Y} \sim N(\mu, \sigma^2/n) \tag{7}$$

のように正規分布にしたがうから，推定量の分散は

$$V[\hat{\mu}(Y)] = \frac{\sigma^2}{n} \tag{8}$$

となる．したがって，観測値の数 n が大きいほど，分散の小さく，精度よい推定ができることになる．また推定量の標準誤差（推定量の標準偏差の評価値）は

$$SE[\hat{\mu}] = \frac{\hat{\sigma}}{\sqrt{n}} \tag{9}$$

で求めることができる．

3.3 分散 σ^2 に対する推定

次に，（理論）分散 σ^2 の推定量はどうであろうか．理論分散と区別するために，観測値から計算される分散のことを標本分散とよび，

$$\hat{\sigma^2}(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{Y})^2 \tag{10}$$

のように定義する．ここで分母が n ではなく $n-1$ となっているのは，証明は省略するが，

$$E\left[\sum_{i=1}^{n}(Y_i - \hat{Y})^2\right] = (n-1)\sigma^2 \tag{11}$$

成り立ち，したがって

$$E[\hat{\sigma}^2(Y)] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{Y})^2\right] = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2 \tag{12}$$

となり，$\hat{\sigma}^2$ を不偏するためである事が分かる．

3.4 例題

例 1　ある海域に生息するマグロ種の 1 歳魚を 9 個体ランダムにサンプリングし，その体長を測定したところ 42, 40, 50, 39, 40, 43, 42, 48, 43 (cm) であった．これらが独立同一に正規分布 $N(\mu, \sigma^2)$ にしたがうとするとき，期待値 μ と分散 σ^2 の推定値を求めよ．また，μ の推定値の推定誤差を求めよ．

$$Y_1, Y_2, \ldots, Y_n \sim (iid)N(\mu, \sigma^2)$$

と仮定し $(n = 9)$,

$$\hat{\mu} = \bar{y} = \frac{1}{9}(42 + 40 + 50 + 39 + 40 + 43 + 42 + 48 + 43) = 43$$

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y})^2 \\ &= \frac{1}{8}\left[(42-43)^2 + (40-43)^2 + \cdots + (43-43)^2\right] = 13.75\end{aligned}$$

よって，$\hat{\mu}$ の標準誤差は以下のように求まる．

$$SE[\hat{\mu}] = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{\sqrt{13.75}}{\sqrt{9}} = 1.236$$

(注意) 推定量に具体的な観測値の値を入れた時には $\hat{\mu}, \hat{\sigma}^2$ のように y を省略して書いています．

4　提出課題

前回の Lecture 6 と今回の Lecture 7 の課題と併せて一つの課題とします．計算式と答えを書いて，必ずまとめて 1 つのファイルで提出してください．ワードで数式の書けない人は，手書きの計算結果を写真に撮ってワードに張り付けて下さっても構いません．答だけの回答は大きく減点します．提出期限は 7 月 3 日 23:59 とします．

Lecture06-HW1 (再掲) あるチョコレート工場で 100g の板チョコを生産しているが，製品によってばらつきが生じ，正規分布 $N(102, 2^2)$ にしたがうとされている．100g 未満のチョコレートは出荷できないとき，生産したチョコレートの何%が不良品となるか？また，不良品率を 1 パーセント以下にしたいとき，板チョコの重さの平均値をいくらにするように生産工程を変えればよいか？(平均値を変えてもばらつき，すなわち分散は変わらないとする)

Lecture07-HW1 ある航空会社の XX 路線の予定飛行時間は 100 分とされているが，過去 10 日間の記録をみると 105, 98, 110, 99, 95, 112, 120, 93, 102, 106 (分) であった．これらが独立同一に正規分布 $N(\mu, \sigma^2)$ にしたがうとするとき，期待値 μ と分散 σ^2 の推定値を求めよ．また，μ の推定値の推定誤差を求めよ．さらに，推定した μ, σ^2 の値を基にして，この路線の飛行機が予定飛行時間よりも遅れる確率も求めよ（前回の標準正規分布表を用いてください）．