推測統計への導入（３）付録

伊庭幸人

大学共同利用機関法人 情報・システム研究機構
統計数理研究所

付録（用語や記号）

一部は「補助資料」と重複します

確率の記号など(1)

* 確率変数 X が離散値（整数値）をとる場合
  確率 $P(X = x)$ を $p(x)$ と書く

* 集合 A について確率 $P(X \in A)$ を $P(A)$ と書く
  集合 A を事象 A と呼ぶことがある

確率の記号など(2)

* 確率変数Xの分布が$p(x)$であることを
  $X \sim p(x)$のように書く

* 慣習では，確率変数を大文字 $X, Y, Z \cdots$
  分布の引数を同じ文字の小文字 $x, y, z \cdots$で書く

  統計関係では，文字が増えてくると
  守られないことも多い（全部小文字にするなど）

* 別の関数でも全部p, Pで書く慣習がある
  （引数で見分ける；混乱したら記号を換える）

記号の慣用（参考）

	データ（標本）	確率分布（母集団）	正規分布などのパラメータ
期待値/平均	標本平均 $\bar{X}$	期待値 $\mathbf{E}[X]$	平均 μ
分散	不偏分散 s^2 標本分散 $\hat{\sigma}^2$	分散 $\mathbf{V}[X]$	分散 σ^2
相関係数	標本相関係数 r	相関係数 $\rho = \dfrac{\mathbf{cov}[X,Y]}{\sqrt{\mathbf{V}[X]\mathbf{V}[Y]}}$	

* いろいろ流儀があるが本講座での使い分けを示す
* 教科書を書くのでなければ，記号の使い分けに神経質になる必要はない（中身の理解が重要！）
* 「標本分散」は$N-1$でなくNで割ったものをさす
  （分散の最尤推定値と一致するので，上ではその記号を使用）

付録（不偏分散について）

$N-1$で割る理由

不偏分散 $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ の意味

誤差の基準点であるはずの平均が
標本に依存する標本平均 $\bar{X}$ で置き換えられ
しかも $\bar{X}$ には X_i が含まれている

$(N \to N-1)$で「小さい方にずれる」傾向を補正

$\mathrm{E}[s^2] = \sigma^2$（不偏性）が成り立つ

導出(1)

準備として以下の共分散を計算しておく

$$\mathbf{cov}(X_i, \bar{X}) = \mathbf{cov}\left(X_i, \frac{1}{N}\sum_{j=1}^{N} X_j\right)$$

$$= \frac{1}{N}\sum_{j=1}^{N} \mathbf{cov}(X_i, X_j) = \frac{1}{N}\mathbf{cov}(X_i, X_i) = \frac{1}{N}\sigma^2$$

ここで，定義からすぐ示せる関係式

$$\mathbf{cov}(W, U+V) = \mathbf{cov}(W, U) + \mathbf{cov}(W, V)$$

$$\mathbf{cov}(W, W) = \mathbf{var}(W)$$

及び，iの異なるX_iの間の独立性を利用した

導出(2)

$$\mathbf{E}\left[(X_i - \bar{X})^2\right]$$

$$= \mathbf{V}[X_i - \bar{X}]$$

$\because \mathbf{E}[X_i - \bar{X}] = \mathbf{E}[X_i] - \mathbf{E}[\bar{X}] = 0$

$$= \mathbf{V}[X_i] - 2 \times \mathbf{cov}[X_i, \bar{X}] + \mathbf{V}[\bar{X}]$$

$$= \sigma^2 - \frac{2\sigma^2}{N} + \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2$$

$\because$導出 (1)の結果と誤差の$\sqrt{N}$則

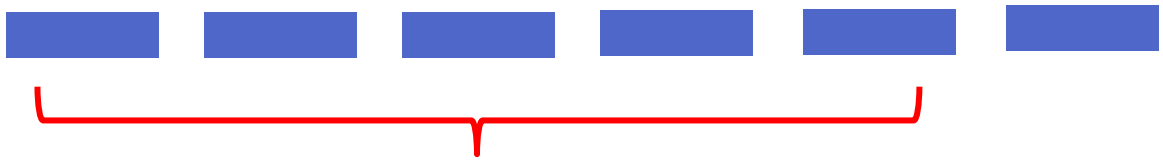
したがって

$$\mathbf{E}\left[\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2\right] = \frac{N}{N-1}\frac{N-1}{N}\sigma^2 = \sigma^2$$

不偏性と一致性の違い

$\mathrm{E}[\theta^*(X)]$ =母集団のθ

推定量$\theta^*(X)$は不偏性を持つという（Nは有限）



有限サイズの標本から推定→多数平均すると真の値

サンプルサイズ$N \to \infty$で, $\theta^*(X) \to$ 母集団のθ

推定量$\theta^*(X)$は一致性を持つという

１個の大きな標本から推定すると真の値に近くなる

共分散の不偏推定量

$$\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})$$

導出のあらすじ

- 「不偏推定量の和や差」は「和や差の不偏推定量」
  これは「和の期待値は期待値の和」から明らか
- 以下の関係式が成り立つ

$$\mathbf{V}[X+Y]=\mathbf{V}[X]+2\mathbf{cov}[X,Y]+\mathbf{V}[Y]$$

$$\Rightarrow\ \mathbf{cov}[X,Y]=\frac{1}{2}\{\mathbf{V}[X+Y]-\mathbf{V}[X]-\mathbf{V}[Y]\}$$

- 上の式の右辺の各項に各々の不偏推定量を代入して計算

不偏性は不可欠か？

例）「標準偏差の不偏推定量」は
「分散の不偏推定量の平方根」ではない

しかし「標準偏差の不偏推定量」を使うことは
まずない（形が複雑で母集団の分布に依存）

例）通常使われる標本相関係数は不偏ではない

不偏分散は形がシンプルなので普通に使われて
いて，それに異論を唱える必要はないが
極端に不偏性にこだわる必要はないと思われる

不偏性が重要になりうる場合

* 有限のNが本質的な場合
  本講座では導出まで触れないが，情報量規準AICの導出では不偏性が重要な役割を演じる
* 時系列のような相関のあるデータ
  この場合，分散などへの補正は$(N \Rightarrow N-1)$よりもずっと大きくなるかもしれない
* 集計したデータへの曲線のあてはめや再集計