2025/5/14

不偏分散－なぜ$n-1$で割るのか？

野津

平均μ，標準偏差σの母集団からn個の標本x_i $(i=1,\cdots,n)$を抽出し，$\bar{x}$を標本平均すなわち

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1}$$

とするとき

$$標本分散=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{2}$$

$$不偏分散=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{3}$$

である．分母がnであるか$n-1$であるかの違いであるが，このうち，母集団の分散（母分散）を推定するには不偏分散を用いる方が良いとされている．

母分散の推定に不偏分散が適しているとされるのは，不偏分散の期待値が母分散に一致するためである．すなわち，n個の標本の抽出を何度も行うことを考えた場合，標本から計算される不偏分散の値は毎回異なるけれども，繰り返し抽出を行ううちに，不偏分散の平均値は母分散となることが期待できる．ただし，前提として，$i\neq j$のときx_iとx_jは独立でなければならない．したがって，有限の要素からなる母集団から非復元抽出を行う場合には，この議論は当てはまらない．その意味で，別の項で紹介している標準誤差に比べ少し適用範囲が狭まっていると言える．

不偏分散の期待値が母分散に一致することは以下のように示すことができる．式(3)右辺の期待値をとると

$$\begin{aligned}
E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]&=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\mu+\mu-\bar{x})^2\right]\\
&=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)+\sum_{i=1}^{n}(\bar{x}-\mu)^2\right]\\
&=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\mu)^2-2n(\bar{x}-\mu)(\bar{x}-\mu)+n(\bar{x}-\mu)^2\right]\\
&=\frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2\right]\\
&=\frac{1}{n-1}\left[\sum_{i=1}^{n}E[(x_i-\mu)^2]-nE[(\bar{x}-\mu)^2]\right]
\end{aligned} \tag{4}$$

ここで，母分散の定義により

$$\sigma^2=E[(x_i-\mu)^2] \tag{5}$$

であるから式(4)右辺括弧内の第 1 項は$n\sigma^2$である．また，式(4)右辺括弧内の第 2 項の期待値部分は標本平均の分散である．すなわち

$$E[(\bar{x}-\mu)^2] = V[\bar{x}] = V\left[\frac{x_1+x_2+\cdots+x_n}{n}\right] = \frac{1}{n^2}V[x_1+x_2+\cdots+x_n] \tag{6}$$

である．ここで，先に述べたように$i \neq j$のときx_iとx_jは独立と仮定しているので，確率変数の和の分散に関する公式（**付録**参照）を適用でき

$$E[(\bar{x}-\mu)^2] = \frac{1}{n^2}\left[V[x_1]+V[x_2]+\cdots+V[x_n]\right] = \frac{1}{n}\sigma^2 \tag{7}$$

である．式(5)(7)を式(4)に代入すると

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{1}{n-1}[n\sigma^2-\sigma^2] = \sigma^2 \tag{8}$$

となる．すなわち不偏分散の期待値は母分散に一致する．

付録　互いに独立な確率変数の和の分散について

互いに独立な確率変数xとyの和の分散に関して次式が成立する．

$$V[x+y]=V[x]+V[y] \tag{A1}$$

これは次のように示すことができる．xとyの同時確率密度関数を$f(x,y)$とすると，定義により

$$V[x+y]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\left(x+y-\mu_x-\mu_y\right)^2dxdy$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)(x-\mu_x)^2dxdy+2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)(x-\mu_x)\left(y-\mu_y\right)dxdy+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)\left(y-\mu_y\right)^2dxdy$$

ここで, xとyは互いに独立であることから, xとyの同時確率密度関数$f(x,y)$は各々の確率密度関数$g(x)$と$h(y)$積として表すことができ,

$$\begin{aligned}V[x+y]&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x)h(y)(x-\mu_x)^2dxdy+2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x)h(y)(x-\mu_x)\left(y-\mu_y\right)dxdy\\&\quad+\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x)h(y)\left(y-\mu_y\right)^2dxdy\\&=\left[\int_{-\infty}^{\infty}g(x)(x-\mu_x)^2dx\right]\left[\int_{-\infty}^{\infty}h(y)dy\right]+2\left[\int_{-\infty}^{\infty}g(x)(x-\mu_x)dx\right]\left[\int_{-\infty}^{\infty}h(y)\left(y-\mu_y\right)dy\right]\\&\quad+\left[\int_{-\infty}^{\infty}g(x)dx\right]\left[\int_{-\infty}^{\infty}h(y)\left(y-\mu_y\right)^2dy\right]\\&=\int_{-\infty}^{\infty}g(x)(x-\mu_x)^2dx+\int_{-\infty}^{\infty}h(y)\left(y-\mu_y\right)^2dy\\&=V[x]+V[y]\end{aligned}$$

となる．すなわち式(A1)が成立することが示された．この性質を繰り返し用いれば，互いに独立な確率変数$x_1,x_2\cdots x_n$の和の分散に関して次式が成立することも言える．

$$V[x_1+x_2+\cdots+x_n]=V[x_1]+V[x_2]+\cdots+V[x_n] \tag{A2}$$