

なぜ不偏分散は標本数ではなく標本数-1 で割るのか

或る母集団からサンプリング(標本抽出)して得た n 個の実現値(実際に計測した値) $x_1, x_2, \dots, x_n$ を考えます。 $x_1, x_2, \dots, x_n$ という個々の数値は、概念的にはそれぞれ $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ という確率変数の実現値です。実際には同じ母集団から標本抽出されているので n 個の確率変数 $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ は、全て同一の確率分布に従う、と考えます。大きな母集団(大抵は計測対象に限りが無い、無限母集団)では、抽出した実現値は互いに影響を与えないという仮定が科学的に妥当です。つまり、実現値は、**互いに独立で同一の確率分布に従う i.i.d(independent identically distributed)な確率変数**であると考えます。

i.i.d な確率変数 $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ で表される量を統計量といいます。標本確率変数から

1. 標本平均... $\bar{X} = \frac{1}{n} \sum_{l=1}^n X^{(l)}$ 確率変数の重心を表す
2. 標本分散... $V^2 = \frac{1}{n} \sum_{l=1}^n (X^{(l)} - \bar{X})^2$ 確率変数のバラツキを表す

の2つの統計量が導入できます。 $\bar{X}$ も V^2 も確率変数であり、確率変数 $X^{(1)}, X^{(2)}, \dots, X^{(n)}$ の代わりにその実現値 $x_1, x_2, \dots, x_n$ を使って計算した数値は、それぞれ標本平均値、標本分散値と呼びます。

確率変数 $X^{(l)}$ の母集団分布に関する期待値 $E[X]$ を母平均といい μ (真の平均) で表し、分散の期待値 $E[(X - E[X])^2]$ を母分散といい σ^2 (真の分散) で表します。 μ 、 σ^2 と $\bar{X}$ と V^2 との関係は、

$$E[\bar{X}] = \frac{1}{n} \{E[X^{(1)}] + E[X^{(2)}] + \dots + E[X^{(n)}]\} = \frac{1}{n} \cdot n\mu = \mu \quad \text{となり、標本平均の期待値は母平均に一致する}$$

ことが分かります。一方、

$$\begin{aligned} E[V^2] &= \frac{1}{n} \sum_{l=1}^n E[(X^{(l)} - \bar{X})^2] = \frac{1}{n} \sum_{l=1}^n E[\{(X^{(l)} - \mu) - (\bar{X} - \mu)\}^2] \\ &= \frac{1}{n} \sum_{l=1}^n E[(X^{(l)} - \mu)^2] + \frac{1}{n} \sum_{l=1}^n E[(\bar{X} - \mu)^2] - 2 \cdot \frac{1}{n} \sum_{l=1}^n E[(X^{(l)} - \mu)(\bar{X} - \mu)] \end{aligned}$$

上式の第一項は、 $\frac{1}{n} \sum_{l=1}^n E[(X^{(l)} - \mu)^2] = \frac{1}{n} \cdot n\sigma^2 = \sigma^2$ であり、上式の第三項は、

$$-2 \cdot \frac{1}{n} \sum_{l=1}^n E[(X^{(l)} - \mu)(\bar{X} - \mu)] = -2 \cdot E\left[\left(\frac{1}{n} \sum_{l=1}^n X^{(l)} - \frac{1}{n} \sum_{l=1}^n \mu\right)(\bar{X} - \mu)\right] = -2E\left[\left(\bar{X} - \frac{1}{n} \cdot n\mu\right)(\bar{X} - \mu)\right] = -2E[(\bar{X} - \mu)^2]$$

となるので、結局、 $E[V^2] = \sigma^2 - E[(\bar{X} - \mu)^2] = \frac{n-1}{n} \sigma^2$ となります。

つまり標本分散の期待値は母分散に一致しません。そこで、 S^2 という統計量を導入し、

$$S^2 = \frac{n}{n-1} V^2 = \frac{1}{n-1} \sum_{l=1}^n (X^{(l)} - \bar{X})^2 \quad \text{とすると、} E[S^2] = \sigma^2 \text{ であり、} S^2 \text{ の期待値は母分散に等しく}$$

なります。この S^2 を不偏分散といいます。よって不偏分散値 s^2 の計算には $n-1$ を使用するのです。