分散
n で割るか $n-1$ で割るか

渡邉 俊夫

平均と分散

N 個のデータ x_i からなる母集団の平均（期待値）μ は

$$\mu = E(x_i) = \frac{1}{N}\sum_{i=1}^{N} x_i = \overline{x_i}$$

である。平均からのずれ（偏差）を

$$\epsilon_i = x_i - \mu$$

と表すと、母集団の分散 σ^2 は偏差の2乗平均であり、

$$\sigma^2 = V(x_i) = \frac{1}{N}\sum_{i=1}^{N} {\epsilon_i}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 = \overline{(x_i - \overline{x_i})^2}$$

である。

母集団のデータ数 N を母数と呼んではいけない。統計学において、母数とはパラメータの意味である。平均や分散が母数に相当する。

分散

分散 σ^2 は、次のように「2乗の平均」と「平均の2乗」の差で表される。

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} {\epsilon_i}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} ({x_i}^2 - 2x_i\mu + \mu^2) = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - 2\mu \cdot \frac{1}{N}\sum_{i=1}^{N} x_i + \mu^2 \cdot \frac{1}{N}\sum_{i=1}^{N} 1$$

$$= \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - 2\mu^2 + \mu^2 = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \mu^2$$

$$= \overline{{x_i}^2} - \mu^2 = \overline{{x_i}^2} - \overline{x_i}^2$$

分散

2乗の平均は

$$\overline{{x_i}^2} = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_N}^2}{N}$$

であり、平均の2乗は

$$\overline{x_i}^2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2 = \frac{1}{N^2}\left(\sum_{i=1}^{N} {x_i}^2 + 2\sum_{i=1}^{N}\sum_{i'>i}^{N} x_i x_{i'}\right)$$

$$= \frac{({x_1}^2 + {x_2}^2 + \cdots + {x_N}^2) + 2(x_1x_2 + x_1x_3 + \cdots + x_{N-1}x_N)}{N^2}$$

であるから

分散

分散 σ^2 は、次のようにも表せる。

$$\sigma^2 = \overline{x_i{}^2} - \bar{x}_i{}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i{}^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i{}^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i{}^2 + 2\sum_{i=1}^{N}\sum_{i'>i}^{N} x_i x_{i'}\right)$$

$$= \frac{1}{N^2}\left(\sum_{i=1}^{N}(N-1)x_i{}^2 - 2\sum_{i'>i}^{N} x_i x_{i'}\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'>i}^{N}(x_i - x_{i'})^2$$

分散

例えば、$N = 3$ のとき

$$\sigma^2 = \overline{x_i^2} - \overline{x_i}^2 = \frac{x_1^2 + x_2^2 + x_3^2}{3} - \left(\frac{x_1 + x_2 + x_3}{3}\right)^2$$

$$= \frac{x_1^2 + x_2^2 + x_3^2}{3} - \frac{(x_1^2 + x_2^2 + x_3^2) + 2(x_1x_2 + x_1x_3 + x_2x_3)}{9}$$

$$= \frac{2(x_1^2 + x_2^2 + x_3^2) - 2(x_1x_2 + x_1x_3 + x_2x_3)}{9}$$

$$= \frac{(x_1^2 - 2x_1x_2 + x_2^2) + (x_1^2 - 2x_1x_3 + x_3^2) + (x_2^2 - 2x_2x_3 + x_3^2)}{9}$$

$$= \frac{(x_1 - x_2)^2 + (x_1 - x_3)^2 + (x_2 - x_3)^2}{9}$$

サンプルの平均

母集団から n 個のデータを含むサンプルを取り出す（これをサンプリングという）。このサンプルの i 番目のデータを x_{si} とすると、その平均 μ_s は

$$\mu_s = \frac{1}{n}\sum_{i=1}^{n} x_{si} = \frac{1}{n}\sum_{i=1}^{n}(\mu + \epsilon_{si}) = \mu \cdot \frac{1}{n}\sum_{i=1}^{n} 1 + \frac{1}{n}\sum_{i=1}^{n} \epsilon_{si}$$

$$= \mu + \frac{1}{n}\sum_{i=1}^{n} \epsilon_{si}$$

である。すなわち、第2項の分だけサンプルの平均 μ_s は母集団の平均 μ と異なる。

サンプルに含まれるデータの個数 n をサンプル数と呼んではいけない。
n はサンプルサイズと呼ぶのが正しい。

サンプルの分散

サンプルの分散 $\sigma_s{}^2$ は

$$\sigma_s{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu_s)^2 = \frac{1}{n}\sum_{i=1}^{n}\big((x_{si}-\mu)-(\mu_s-\mu)\big)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\big((x_{si}-\mu)^2 - 2(x_{si}-\mu)(\mu_s-\mu) + (\mu_s-\mu)^2\big)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu)^2 - 2(\mu_s-\mu)\cdot\frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu) + (\mu_s-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu)^2 - 2(\mu_s-\mu)\cdot\frac{1}{n}\sum_{i=1}^{n}x_{si} + 2(\mu_s-\mu)\mu + (\mu_s-\mu)^2$$

サンプルの分散

サンプルの分散 $\sigma_s{}^2$ は

$$\sigma_s{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu)^2 - 2(\mu_s-\mu)\cdot\frac{1}{n}\sum_{i=1}^{n}x_{si} + 2(\mu_s-\mu)\mu + (\mu_s-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu)^2 - 2(\mu_s-\mu)\cdot\mu_s + 2(\mu_s-\mu)\mu + (\mu_s-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu)^2 - 2(\mu_s-\mu)^2 + (\mu_s-\mu)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu)^2 - (\mu_s-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}\epsilon_{si}{}^2 - (\mu_s-\mu)^2$$

サンプルの分散

サンプルの分散 $\sigma_s{}^2$ は

$$\sigma_s{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{si}-\mu)^2 - (\mu_s-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}\epsilon_{si}{}^2 - (\mu_s-\mu)^2$$

である。第1項はサンプルに含まれる n 個のデータについて、母集団の平均 μ からの偏差 ϵ_{si} の2乗平均（＝母集団の平均 μ に対する分散）を表しており、第2項はサンプルの平均 μ_s と母集団の平均 μ との差の2乗を表している。

「サンプルの平均」の平均

母集団から、n 個のデータを含むサンプルを m 組取り出す。このとき、
「サンプルの平均 μ_j」の平均 $\overline{\mu_j}$ は

$$\overline{\mu_j} = \frac{1}{m}\sum_{j=1}^{m}\mu_j = \frac{1}{m}\sum_{j=1}^{m}\left(\mu + \frac{1}{n}\sum_{i=1}^{n}\epsilon_{ji}\right) = \mu + \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}$$

である。第2項は、サンプルの個数 m が大きいとき 0 に近づくから、
「サンプルの平均 μ_j」の平均 $\overline{\mu_j}$ は母集団の平均 μ に近づく。すなわち、

$$\overline{\mu_j} \to \mu$$

となる。

サンプルに含まれるデータの個数 n を
サンプル数と呼んではいけない。
n はサンプルサイズと呼ぶのが正しい。
サンプル数と言うと m のことになる。

「サンプルの平均」の分散

「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は

$$V(\mu_j) = \frac{1}{m}\sum_{j=1}^{m}\left(\mu_j - \overline{\mu_j}\right)^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\left(\mu_j - \mu\right) - \left(\overline{\mu_j} - \mu\right)\right)^2$$

$$= \frac{1}{m}\sum_{j=1}^{m}\left(\left(\mu_j - \mu\right)^2 - 2\left(\mu_j - \mu\right)\left(\overline{\mu_j} - \mu\right) + \left(\overline{\mu_j} - \mu\right)^2\right)$$

$$= \frac{1}{m}\sum_{j=1}^{m}\left(\mu_j - \mu\right)^2 - 2\left(\overline{\mu_j} - \mu\right)\cdot\frac{1}{m}\sum_{j=1}^{m}\left(\mu_j - \mu\right) + \left(\overline{\mu_j} - \mu\right)^2$$

「サンプルの平均」の分散

「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は

$$V(\mu_j) = \frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2 - 2(\overline{\mu_j} - \mu)\cdot\frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu) + (\overline{\mu_j} - \mu)^2$$

$$= \frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2 - 2(\overline{\mu_j} - \mu)\cdot\frac{1}{m}\sum_{j=1}^{m}\mu_j + 2(\overline{\mu_j} - \mu)\mu + (\overline{\mu_j} - \mu)^2$$

$$= \frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2 - 2(\overline{\mu_j} - \mu)\cdot\overline{\mu_j} + 2(\overline{\mu_j} - \mu)\mu + (\overline{\mu_j} - \mu)^2$$

「サンプルの平均」の分散

「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は

$$V(\mu_j) = \frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2 - 2(\overline{\mu_j} - \mu)\cdot\overline{\mu_j} + 2(\overline{\mu_j} - \mu)\mu + (\overline{\mu_j} - \mu)^2$$

$$= \frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2 - 2(\overline{\mu_j} - \mu)(\overline{\mu_j} - \mu) + (\overline{\mu_j} - \mu)^2$$

$$= \frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2 - (\overline{\mu_j} - \mu)^2$$

となる。

「サンプルの平均」の分散

「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は

$$V(\mu_j) = \frac{1}{m}\sum_{j=1}^{m}(\mu_j - \mu)^2 - (\overline{\mu_j} - \mu)^2$$

となる。第1項は m 個のサンプルについて、各サンプルの平均 μ_j と母集団の平均 μ との差の2乗平均（＝母集団の平均 μ に対する分散）を表しており、第2項は「サンプルの平均 μ_j」の平均 $\overline{\mu_j}$ と母集団の平均 μ との差の2乗を表している。

「サンプルの平均」の分散

ここで

$$\frac{1}{m}\sum_{j=1}^{m}(\mu_j-\mu)^2=\frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ji}\right)^2=\frac{1}{mn^2}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\epsilon_{ji}\right)^2$$

$$=\frac{1}{mn^2}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}{\epsilon_{ji}}^2+2\sum_{i=1}^{n}\sum_{i'>i}^{n}\epsilon_{ji}\epsilon_{ji'}\right)$$

$$=\frac{1}{n}\cdot\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}{\epsilon_{ji}}^2+\frac{2}{mn^2}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{i'>i}^{n}\epsilon_{ji}\epsilon_{ji'}$$

であり、

「サンプルの平均」の分散

また

$$\overline{\mu_j} - \mu = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}$$

だから、「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は

$$V(\mu_j) = \frac{1}{n}\left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}{\epsilon_{ji}}^2\right) + \frac{2}{mn^2}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{i'>i}^{n}\epsilon_{ji}\epsilon_{ji'} - \left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}\right)^2$$

となる。

「サンプルの平均」の分散

「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は

$$V(\mu_j) = \frac{1}{n}\left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}{\epsilon_{ji}}^2\right) + \frac{2}{mn^2}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{i'>i}^{n}\epsilon_{ji}\epsilon_{ji'} - \left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}\right)^2$$

となる。サンプルの個数 m が大きいとき、第1項の（　）内は σ^2 に近づき、第2項と第3項はいずれも 0 に近づくから、「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は σ^2/n に近づく。すなわち、

$$V(\mu_j) \to \frac{\sigma^2}{n}$$

である。

「サンプルの平均」の分散

「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は、サンプルの個数 m が大きいとき

$$V(\mu_j) \to \frac{\sigma^2}{n}$$

に近づく。これより、「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は、母集団の分散 σ^2 の $1/n$ であり、サンプルサイズ n が大きくなるほど、「サンプルの平均 μ_j」の分散 $V(\mu_j)$ は小さくなることがわかる。

「サンプルの分散」の平均

「サンプルの分散 σ_j^2」の平均 $\overline{\sigma_j^2}$ は

$$\overline{\sigma_j^2} = \frac{1}{m}\sum_{j=1}^{m}\sigma_j^2 = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ji}^2 - \left(\mu_j - \mu\right)^2\right)$$

$$= \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}^2 - \frac{1}{m}\sum_{j=1}^{m}\left(\mu_j - \mu\right)^2$$

$$= \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}^2 - \left(V(\mu_j) + \left(\overline{\mu_j} - \mu\right)^2\right)$$

「サンプルの分散」の平均

「サンプルの分散 $\sigma_j{}^2$」の平均 $\overline{\sigma_j{}^2}$ は

$$\overline{\sigma_j{}^2} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}{}^2 - \left(V(\mu_j) + \left(\overline{\mu_j} - \mu\right)^2\right)$$

$$= \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}{}^2 - V(\mu_j) - \left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}\right)^2$$

となる。

「サンプルの分散」の平均

「サンプルの分散 $\sigma_j{}^2$」の平均 $\overline{\sigma_j{}^2}$ は

$$\overline{\sigma_j{}^2} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}{}^2 - V(\mu_j) - \left(\frac{1}{mn}\sum_{j=1}^{m}\sum_{i=1}^{n}\epsilon_{ji}\right)^2$$

となる。サンプルの個数 m が大きいとき、第1項は σ^2 に、第2項は σ^2/n に、第3項は 0 にそれぞれ近づくから

$$\overline{\sigma_j{}^2} \to \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

である。

「サンプルの分散」の平均

「サンプルの分散 $\sigma_j{}^2$」の平均 $\overline{\sigma_j{}^2}$ は、サンプルの個数 m が大きいとき

$$\overline{\sigma_j{}^2} \to \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

に近づく。これは、サンプルの各データの「サンプルの平均 μ_j からのずれ」の2乗平均（＝サンプルの分散 $\sigma_j{}^2$）は、母集団の各データの「平均からのずれ」の2乗平均（＝母集団の分散 σ^2）から、サンプルの平均 μ_j と母集団の平均 μ の差の2乗の期待値（＝「サンプルの平均 μ_j」の分散 $V(\mu_j)$）を引いたものに等しいことを示している。したがって、「サンプルの分散 $\sigma_j{}^2$」の期待値は、母集団の分散 σ^2 と一致しない。

不偏分散

「サンプルの分散 $\sigma_j{}^2$」の平均は、サンプルの個数 m が大きいとき

$$\overline{\sigma_j{}^2} \to \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

に近づく。したがって、「サンプルの分散 $\sigma_j{}^2$」の期待値は、母集団の分散 σ^2 と一致しない。しかし、$\sigma_j{}^2$ の代わりに

$$s_j{}^2 = \frac{n}{n-1}\sigma_j{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ji} - \mu_j\right)^2 \to \sigma^2$$

を用いると、$s_j{}^2$ の期待値は母集団の分散 σ^2 と一致する。$s_j{}^2$ を不偏分散という。

まとめ

N 個のデータからなる母集団の分散 σ^2 を求めるときは

$$\sigma^2 = V(x_i) = \frac{1}{N}\sum_{i=1}^{N} {\epsilon_i}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

であり、全データ数 N で割る。

母集団から n 個のデータを含むサンプルを取り出し、それから母集団の分散 σ^2 を推定するときは

$${s_j}^2 = \frac{n}{n-1}{\sigma_j}^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{ji} - \mu_j\right)^2 \rightarrow \sigma^2$$

であり、データ数 n から 1 を引いて $n-1$ で割る。

補足

表計算ソフトウェアのExcel（Microsoft）で、分散

$$\sigma^2 = V(x_i) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x_i})^2$$

を計算するときは、関数 VAR.P （Excel 2007以前では VARP）を用いる。

いっぽう、不偏分散

$$s_j{}^2 = \frac{n}{n-1}\sigma_j{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ji} - \mu_j\right)^2 \to \sigma^2$$

を計算するときは、関数 VAR.S （Excel 2007以前では VAR）を用いる。