

基本統計量：データの解析

よく知られた平均と分散、共分散の計算式は同値な変形式が

$$\text{平均: } \bar{X}_n = \frac{1}{n} \sum_{i=1}^n x_i = \binom{n}{2}^{-1} \sum_{i < j} \frac{x_i + x_j}{2}$$

$$\text{分散: } \text{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X}_n)^2 = \binom{n}{2}^{-1} \sum_{i < j} \frac{(x_i - x_j)^2}{2}$$

$$\text{共分散: } \text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X}_n)(y_i - \bar{Y}_n) = \binom{n}{2}^{-1} \sum_{i < j} \frac{(x_i - x_j)(y_i - y_j)}{2}$$

で与えられることを示す。

簡単のためにデータ数： $n = 5$ として

> 5

変数 x データ： $x = \{x1, x2, x3, x4, x5\}$ に対して、 x の平均： $\text{Mean}[x]$

> $\frac{1}{5}(x1 + x2 + x3 + x4 + x5)$

また変数 y データ： $y = \{y1, y2, y3, y4, y5\}$ とし、同様に y の平均： $\text{Mean}[y]$ を入力すると

> $\frac{1}{5}(y1 + y2 + y3 + y4 + y5)$

以上のようにデータを準備する。

まず x の不偏分散 ($n - 1$ で割る場合) $\text{Variance}[x]$ は、

> $\frac{1}{4}((x1 + \frac{1}{5}(-x1 - x2 - x3 - x4 - x5))^2 + \dots + (\frac{1}{5}(-x1 - x2 - x3 - x4 - x5) + x5)^2)$

データのリスト x から 2 番目を取り出すための方法： $x[[2]]$ と入力すれば、その結果 $x2$ が得られる。このようにリストからデータを取り出してそれらの和を計算する。まず変数 x では $\text{Sum}[x[[i]], \{i, 1, n\}]$ したがって結果は $x1 + x2 + x3 + x4 + x5$ となる。これを利用して次の計算を試みる。

いま $\frac{2}{n(n-1)} \sum_{i < j} (x_i - x_j)^2$ の計算のために

$$\text{RA} = \text{Sum}[(x[[i]] - x[[j]])^2/2, \{i, 1, n\}, \{j, i + 1, n\}] * 2 / (n(n - 1))$$

とおいて、展開すると

> $\frac{1}{10}(\frac{1}{2}(x1 - x2)^2 + \frac{1}{2}(x1 - x3)^2 + \frac{1}{2}(x1 - x4)^2 + \frac{1}{2}(x2 - x3)^2 + \dots + \frac{1}{2}(x4 - x5)^2)$

が得られる。一方不偏分散の定義 $\text{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X}_n)^2$ からは

$$\text{LA} = \text{Sum}[(x[[i]] - \text{Mean}[x])^2, \{i, 1, n\}] / (n - 1)$$

であるから、展開すれば、

> $\frac{1}{4}((x1 + \frac{1}{5}(-x1 - \dots - x5))^2 + (x2 + \frac{1}{5}(-x1 - \dots - x5))^2 + \dots + (\frac{1}{5}(-x1 - \dots - x5) + x5)^2)$

が得られるから、この両者が等しいことを示すために、差を計算して、ゼロになることを示す。 $\text{LA} - \text{RA}$ を計算していくと、

> $\frac{1}{10}(-\frac{1}{2}(x1 - x2)^2 - \frac{1}{2}(x1 - x3)^2 \dots - \frac{1}{2}(x4 - x5)^2) + \frac{1}{4}((x1 + \frac{1}{5}(-x1 - \dots - x5))^2 + \dots + (\frac{1}{5}(-x1 - \dots - x5) + x5)^2)$

差をとって展開する。Expand[LA - RA]

> 0

したがって両者は等しい。つまり結論として

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X}_n)^2 = \binom{n}{2}^{-1} \sum_{i < j} \frac{(x_i - x_j)^2}{2}$$

が得られた。ここで $\bar{X}_n$ は $\text{Mean}[x]$ とする。

こんどは共分散 $\frac{1}{n-1} \sum_i (x_i - \bar{X}_n)(y_i - \bar{Y}_n)$ について計算し、同様な形に変形できることを証明する。

$$\text{LB} = \text{Sum}[(x[[i]] - \text{Mean}[x])(y[[i]] - \text{Mean}[y]), \{i, 1, n\}]/(n-1)$$

おおくの場合はデータ数 n でわるが、ここでは、先の不偏分散にならい、 $n-1$ で割る値を計算してみる。

$$> \frac{1}{4}((x1 + \frac{1}{5}(-x1 - \dots - x5))(y1 + \frac{1}{5}(-y1 - \dots - y5)) + \dots + (\frac{1}{5}(-x1 - \dots - x5) + x5)(\frac{1}{5}(-y1 - \dots - y5) + y5))$$

一方

$$\text{RB} = \text{Sum}[(x[[i]] - x[[j]])(y[[i]] - y[[j]])/2, \{i, 1, n\}, \{j, i+1, n\}] * 2/(n(n-1))$$

とおくことで比較をするから、

$$> \frac{1}{10}(\frac{1}{2}(x1 - x2)(y1 - y2) + \frac{1}{2}(x1 - x3)(y1 - y3) + \dots + \frac{1}{2}(x4 - x5)(y4 - y5))$$

この定義から、差を計算することで $\text{LB} - \text{RB}$ を展開して

$$> \frac{1}{10}(-\frac{1}{2}(x1 - x2)(y1 - y2) - \frac{1}{2}(x1 - x3)(y1 - y3) - \dots - \frac{1}{2}(x4 - x5)(y4 - y5)) + \frac{1}{4}((x1 + \frac{1}{5}(-x1 - \dots - x5))(y1 + \frac{1}{5}(-y1 - \dots - y5)) + (x2 + \frac{1}{5}(-x1 - \dots - x5))(y2 + \frac{1}{5}(-y1 - \dots - y5)) + \dots + (\frac{1}{5}(-x1 - \dots - x5) + x5)(\frac{1}{5}(-y1 - \dots - y5) + y5))$$

より、 $\text{Expand}[\text{LB} - \text{RB}]$ の出力はゼロ、つまり

> 0

となって等しいことを確認することができる。結論は

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X}_n)(y_i - \bar{Y}_n) = \binom{n}{2}^{-1} \sum_{i < j} \frac{(x_i - x_j)(y_i - y_j)}{2}$$

2 項係数の計算は $n=5$ であるから、 $\text{Binomial}[n, 2]$

> 10

となるが、これは和の条件 $i < j$ を満たす個数に他ならない。これを確認するため、次の計算

$$\binom{n}{2}^{-1} \sum_{i < j} \frac{x_i + x_j}{2}$$

を試みよう。

$$\text{RC} = (\text{Binomial}[n, 2])^{(-1)} \text{Sum}[(x[[i]] + x[[j]])/2, \{i, 1, n\}, \{j, i+1, n\}]$$

この式は、2 個のデータ x_i, x_j の中心 (重心、平均) をとったもので、

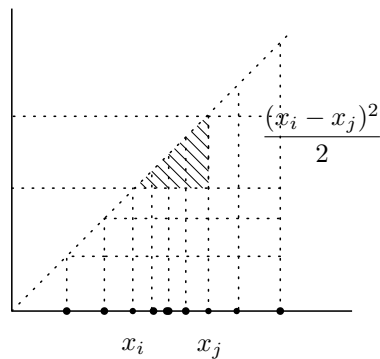
$$> \frac{1}{10}(\frac{x1+x2}{2} + \frac{x1+x3}{2} + \frac{x2+x3}{2} + \dots + \frac{x3+x5}{2} + \frac{x4+x5}{2})$$

となるから、つまり $\text{Expand}[\text{RC}]$

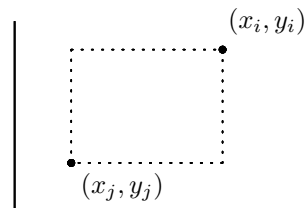
$$> \frac{x1}{5} + \frac{x2}{5} + \frac{x3}{5} + \frac{x4}{5} + \frac{x5}{5}$$

で、 n 個のデータ X の平均、 $\overline{X_n} = \frac{1}{n} \sum_{i=1}^n x_i$ である。

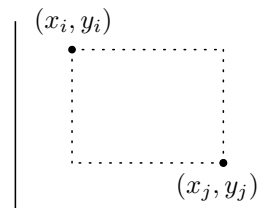
つぎの図は分散の値は、非負値しか取りえないこと、および 2 つのデータ間の離れ具合、つまり集積と離散の程度を、三角形の面積で評価し、これらの総和の値を個数で割った平均値とするものであることがわかる。したがって平均とも関連する。元の場合の式では平均との離れ具合を正方形の面積で評価し、それを平均したものと考えられる。



共分散はプラスの数とマイナスの数が両方あり得るから、正になったり、負になったりする。



$(x_i - x_j)(y_i - y_j)/2$ はプラス



$(x_i - x_j)(y_i - y_j)/2$ はマイナス