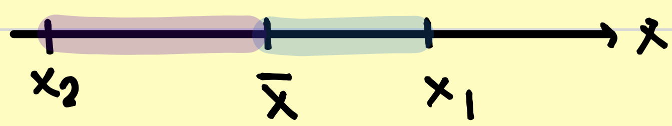
20240403_Data_Analytics_BU6

VARIANZ 1. VARIABEL $X = \{x_1, x_2, x_3, \dots, x_N\}$

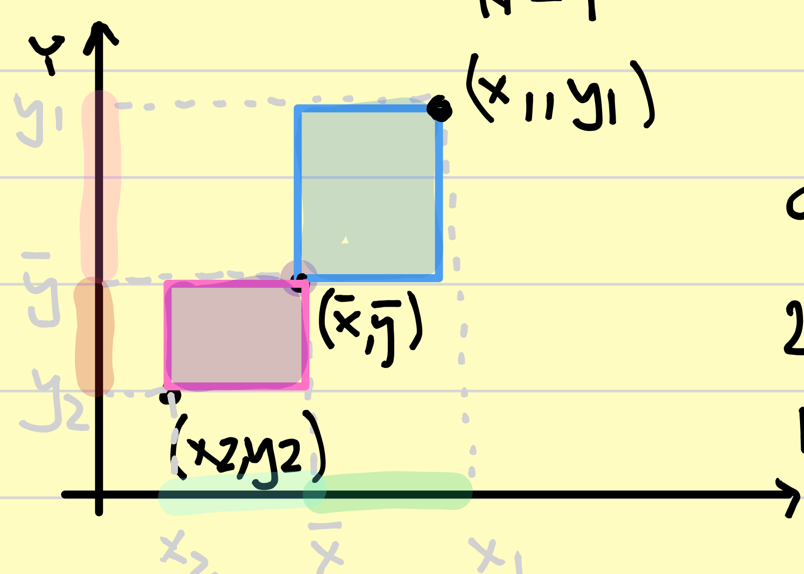
$$VAR = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N} = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\dots}{N}$$

Geometrisch bedeutet die Varianz die summe der Abstände zum Mittelwert quadrat.

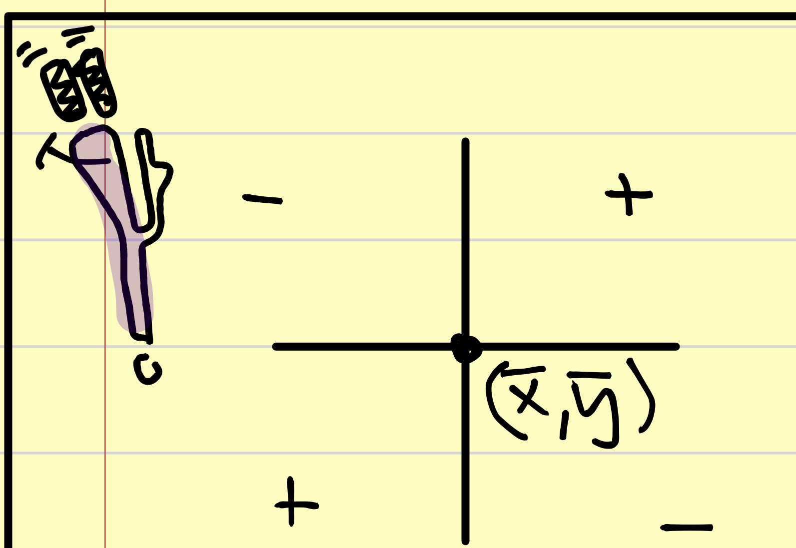


KOVARIANZ. 2. VARIABEL $X = \{x_1, x_2, \dots, x_N\}$ $Y = \{y_1, y_2, \dots, y_N\}$

$$cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N-1} = \frac{(x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \dots}{N-1}$$



Geometrisch ist die Kovarianz zwische 2 Variabeln, die Summe der Rechteckigen zwischen den Punkten und den Mittelwert.



cov(x,y) hat ein + oder −.

Interpretation der COVARIANZ:

$cov(x,y) \approx 0$: Die meisten Punkte sind um den Mittelwert.

$cov(x,y) \gg 0$ / $\ll 0$: Die meisten Punkte entfernen sich sehr vom Mittelwert.

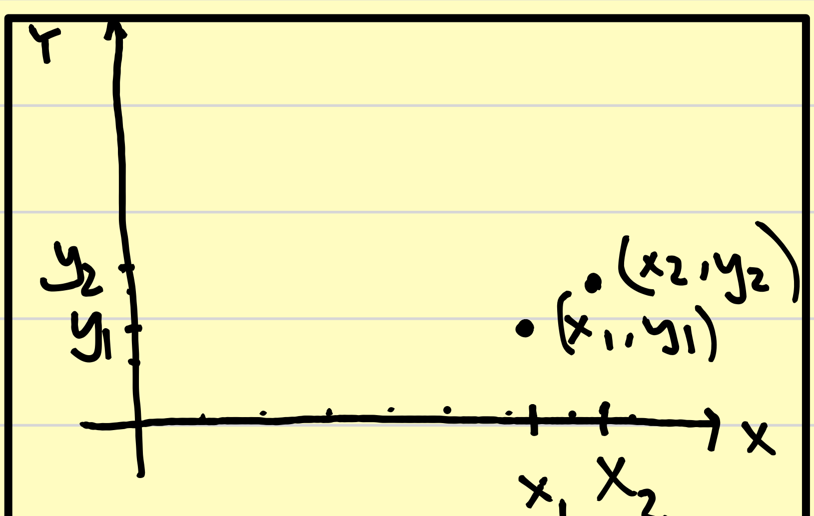
Beispiel: Bitte berechnen Sie die Kovarianz von 2 Variabeln X, Y

$$X = \{\overset{x_1}{6{,}3},\ \overset{x_2}{7{,}5},\ 8{,}7,\ 4{,}2\} \qquad Y = \{\overset{y_1}{1{,}5},\ \overset{y_2}{2{,}3},\ 0{,}2,\ 3{,}7\}$$

$$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{4} (x_i - \bar{x})(y_i - \bar{y})}{4-1} = (*)$$

$$\bar{x} = \frac{6{,}3 + 7{,}5 + 8{,}7 + 4{,}2}{4} \ ; \ \bar{y} = \frac{1{,}5 + 2{,}3 + 0{,}2 + 3{,}7}{4}$$

$$(*) = \frac{(6{,}3 - \bar{x})(1{,}5 - \bar{y}) + (7{,}5 - \bar{x})(2{,}3 - \bar{y}) + (8{,}7 - \bar{x})(0{,}2 - \bar{y}) + (4{,}2 - \bar{x})(3{,}7 - \bar{y})}{4-1}$$



Punkte sind Paaren! (2 Dimensionen)

KOVARIANZ MATRIX. 3 oder mehr VARIABELN

für 3 VARABIABELN:

$X = \{x_1, x_2, \dots, x_N\}$

$Y = \{y_1, y_2, \dots, y_N\}$

$Z = \{z_1, z_2, \dots, z_N\}$

$$A = \begin{bmatrix} \mathrm{VAR}(X) & \operatorname{cov}(X,Y) & \operatorname{cov}(X,Z) \\ \operatorname{cov}(Y,X) & \mathrm{VAR}(Y) & \operatorname{cov}(Y,Z) \\ \operatorname{cov}(Z,X) & \operatorname{cov}(Z,Y) & \mathrm{VAR}(Z) \end{bmatrix}$$

Die Kovarianzmatrix ist symmetrisch $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$

$$cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N-1} = \frac{\sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})}{N-1} = cov(Y,X)$$

Beispiel: wir haben einen Kennzahlensystem in einem Werk mit 3 KPIs: DLZ (Durchlaufzeit), output $\left(\frac{\text{Stück}}{\text{Zeit}}\right)$, Qualität (ppm). Bitte ermitteln Sie die Kovarianzmatrix vom System.

$$DLZ = \{20, 18, 17, 16, 14, 13\} \equiv X$$

$$output = \{200, 250, 380, 410, 390, 440\} \equiv Y$$

$$Q = \{3300, 2500, 1800, 1700, 1500, 1300\} \equiv Z$$

$$\bar{x} = \frac{20+18+17+16+14+13}{6}; \quad \bar{y} = \frac{200+250+380+410+390+440}{6}$$

$$\bar{z} = \frac{3300+2500+1800+1700+1500+1300}{6}$$

$$VAR(X) = \frac{\sum_{i=1}^{6}(x_i-\bar{x})^2}{6} = \frac{(20-\bar{x})^2+(18-\bar{x})^2+(17-\bar{x})^2+(16-\bar{x})^2+(14-\bar{x})^2+(13-\bar{x})^2}{6}$$

$$VAR(Y) = \frac{\sum_{i=1}^{6}(y_i-\bar{y})^2}{6} = \frac{(200-\bar{y})^2+(250-\bar{y})^2+(380-\bar{y})^2+(410-\bar{y})^2+(390-\bar{y})^2+(440-\bar{y})^2}{6}$$

$$VAR(Z) = \frac{\sum_{i=1}^{6}(z_i-\bar{z})^2}{6} = \frac{(3300-\bar{z})^2+(2500-\bar{z})^2+(1800-\bar{z})^2+(1700-\bar{z})^2+(1500-\bar{z})^2+(1300-\bar{z})^2}{}$$

$$cov(X,Y) = \frac{\sum_{i=1}^{6}(x_i-\bar{x})(y_i-\bar{y})}{6-1} = \frac{(20-\bar{x})(200-\bar{y})+(18-\bar{x})(250-\bar{y})+(17-\bar{x})(380-\bar{y})+(16-\bar{x})(410-\bar{y})+(14-\bar{x})(390-\bar{y})+(13-\bar{x})(440-\bar{y})}{6-1} =$$

$$\| $$

$$cov(Y,X)$$

$$\mathrm{cov}(X,Z) = \frac{\sum_{i=1}^{6}(x_i-\bar{x})(z_i-\bar{z})}{6-1} = \frac{(20-\bar{x})(3300-\bar{z})+(18-\bar{x})(2500-\bar{z})+(17-\bar{x})(1800-\bar{z})+(16-\bar{x})(1700-\bar{z})+(14-\bar{x})(1500-\bar{z})+(13-\bar{x})(1300-\bar{z})}{6-1}$$

$$\shortparallel$$

$$\mathrm{cov}(Z,X)$$

$$\mathrm{cov}(Y,Z) = \frac{\sum_{i=1}^{6}(y_i-\bar{y})(z_i-\bar{z})}{6-1} = \ldots$$

$$\boxed{A = \text{KOVARIANZ MATRIX} = \begin{bmatrix} \mathrm{VAR}(X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z) \\ \mathrm{cov}(Y,X) & \mathrm{VAR}(Y) & \mathrm{cov}(Y,Z) \\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{VAR}(Z) \end{bmatrix}}$$

Normierung von Daten

$\bar{x}$ Normalverteilt mit mittelwert μ und std abweichung σ

$\bar{x} \equiv N(\mu,\sigma)$ Kann ich „normieren" in dem Alle Datensätze „dem Mittelwert substrahiere und durch die std Abweichung teile.

Somit wird $\bar{x} \equiv N(\mu,\sigma)$ zu einer x^* variable mit mittelwert „0" und std Abweidung „1"

$x^* \equiv N(0,1)$.

$$x = \{x_1, x_2, x_3, \cdots, x_N\}$$

$$x^* = \left\{\frac{x_1-\bar{x}}{\sigma}, \frac{x_2-\bar{x}}{\sigma}, \frac{x_3-\bar{x}}{\sigma}, \ldots, \frac{x_N-\bar{x}}{\sigma}\right\} = \{x_1^*, x_2^*, x_3^*, \cdots, x_N^*\}$$

Nach Normierung der $\bar{X}, Y, Z$ variabeln, die Kovarianz matrix sieht folgendermaßen aus:

$$\operatorname{cov}(X^*, Y^*, Z^*) = \begin{bmatrix} 1 & \operatorname{cov}(X^*, Y^*) & \operatorname{cov}(X^*, Z^*) \\ & 1 & \operatorname{cov}(Y^*, Z^*) \\ & & 1 \end{bmatrix}$$

Übung: nehme bitte die Daten vom Werk oben, normiere sie in 3 neue Variable (X^*, Y^*, Z^*) und berechne die neue Kovarianz matrix der neuen normierten Variabeln.

$$DLZ^* = \left\{ \frac{20-\bar{x}}{\sqrt{VAR\,X}}, \frac{18-\bar{x}}{\sqrt{VAR\,X}}, \frac{17-\bar{x}}{\sqrt{VAR\,X}}, \frac{16-\bar{x}}{\sqrt{VAR\,X}}, \frac{14-\bar{x}}{\sqrt{VAR\,X}}, \frac{13-\bar{x}}{\sqrt{VAR\,X}} \right\}$$

$Output^* =$

$Qualität^* =$