Vorlesung vom 04.11.2012:

- Varianz und Standardabweichung
- Standardisierung als Spezialfall einer linearen Transformation
- Kovarianz und (empirische) Korrelation mehrdimensionaler Merkmale, genauer: von zwei Merkmalen

---

1) Varianz = mittlere quadratische Abweichung vom arithmetischen Mittelwert

$$\vec{x} = \underline{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n = \underbrace{\mathbb{R} \times \dots \times \mathbb{R}}_{n\text{-mal}}$$

$\vec{x}$ / $\underline{x}$: Datenvektor; $x_1, x_2, \dots, x_n$: Einzeldaten; $\mathbb{R}$: Menge der reellen Zahlen; $\times$: Kartesisches Produkt

Einzeldaten $x_i = X(\omega_i)$, $\omega_i \in \Omega$

Dann:

$$\boxed{\operatorname{var}(\vec{x}) = \operatorname{var}(\underline{x}) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{mit}$$

ist eine Zahl!!

$$\boxed{\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i} \quad \text{arithm. Mittelwert.}$$

Spezielle Formel für die Varianz (Herleitung):

$$\operatorname{var}(\vec{x}) = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \left\{ \sum_{i=1}^{n} x_i^2 - 2\bar{x} \cdot \underbrace{\sum_{i=1}^{n} x_i}_{=n\cdot\bar{x}} + \underbrace{\sum_{i=1}^{n} \bar{x}^2}_{=n\cdot\bar{x}^2} \right\}$$

$$(x_i - \bar{x})^2 = x_i^2 - 2x_i\bar{x} + \bar{x}^2$$

$$= (x_1^2 - 2x_1\bar{x} + \bar{x}^2) + (x_2^2 - 2x_2\bar{x} + \bar{x}^2) + \dots + (x_n^2 - 2x_n\bar{x} + \bar{x}^2)$$

$$= (x_1^2 + x_2^2 + \dots + x_n^2) - 2\bar{x}(x_1 + x_2 + \dots + x_n) + \underbrace{(\bar{x}^2 + \bar{x}^2 + \dots + \bar{x}^2)}_{n\text{-mal}}$$

$$= \frac{1}{n} \left\{ \sum_{i=1}^{n} x_i^2 - 2\bar{x} \cdot n \cdot \bar{x} + n \cdot \bar{x}^2 \right\}$$

$$= \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 \right) - 2\bar{x}^2 + \bar{x}^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 \right) - \bar{x}^2$$

---

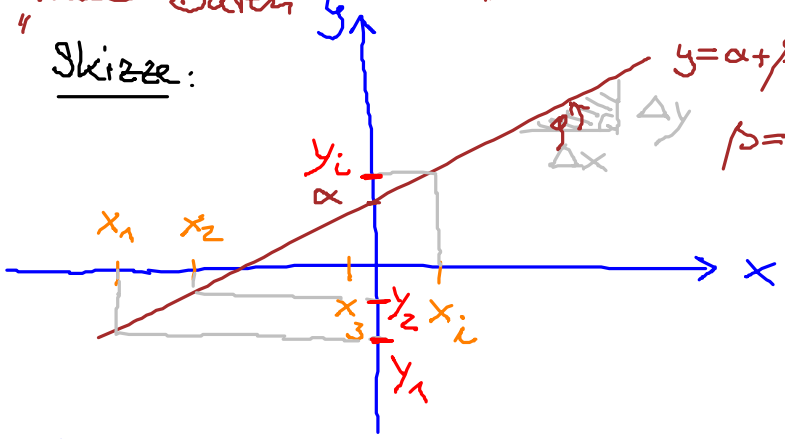
2) Lineare Transformation und Mittelwert/Varianz:

$y_i = \alpha + \beta x_i$ $(i = 1, \dots, n)$ Kurz: $\vec{y} = \underline{y} = \underline{\alpha} + \beta \underline{x} = \vec{\alpha} + \beta \vec{x}$

„neue" Daten — „alte" Daten — Datenvektor „neu" — Datenvektor „alt"

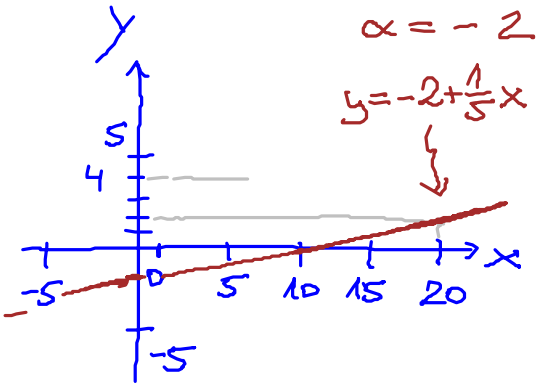
mit $\vec{\alpha} = \underline{\alpha} = \begin{bmatrix} \alpha \\ \alpha \\ \vdots \\ \alpha \end{bmatrix} = \alpha \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$

Skizze:



$\beta = \frac{\Delta y}{\Delta x} = \tan(\varphi)$

Bsp: $\underline{x} = (x_1, x_2, x_3, x_4, x_5) = (1, -1, 10, 5, 20)$



$\alpha = -2$, $\beta = \frac{1}{5}$: $y = -2 + \frac{1}{5}x$

$\underline{y} = \underline{\alpha} + \beta \underline{x} = -2 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \frac{1}{5} \cdot \begin{bmatrix} 1 \\ -1 \\ 10 \\ 5 \\ 20 \end{bmatrix} = \begin{bmatrix} -9/5 \\ -11/5 \\ 0 \\ -1 \\ 2 \end{bmatrix}$

$y_1 = -2 + \frac{1}{5} x_1 = -2 + \frac{1}{5} \cdot 1 = -\frac{9}{5}$

$y_2 = -2 + \frac{1}{5} x_2 = -2 + \frac{1}{5}(-1) = -\frac{11}{5}$ $\left(= -1/5\right)$

$y_3 = -2 + \frac{1}{5} \cdot 10 = 0$ $\quad (= x_3)$

Satz: Ist $\underline{y} = \underline{\alpha} + \beta \underline{x}$ die zugrunde gelegte lineare Transformation

$y_i = \alpha + \beta x_i$ $(i = 1, \dots n)$, so gilt für den Mittelwert auch die Varianz der neuen Daten!

$$\underline{y} = \alpha + \beta \underline{x} \quad , \quad \mathrm{var}(\underline{y}) = \beta^2 \mathrm{var}(\underline{x})$$

neuer TM — alter TM — neue Varianz — alte Varianz

Beweis:

1) $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha+\beta x_i) = \frac{1}{n}\left\{\alpha \cdot \underbrace{\sum_{i=1}^{n} 1}_{=n\cdot\alpha} + \beta \cdot \sum_{i=1}^{n} x_i\right\}$$

$$= \frac{\alpha n}{n} + \beta \cdot \frac{1}{n}\cdot \underbrace{\sum_{i=1}^{n} x_i}_{=\bar{x}} = \alpha + \beta\bar{x}$$

2) $$\mathrm{var}(\underline{y}) = \frac{1}{n}\cdot\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}\{(\alpha+\beta x_i) - (\alpha+\beta\bar{x})\}^2$$

(y_i, $\bar{y}$)

$$= \frac{1}{n}\cdot\sum_{i=1}^{n}\{\beta(x_i-\bar{x})\}^2 = \beta^2\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2 = \beta^2\cdot\mathrm{var}(\underline{x})$$

($=\beta^2\cdot(x_i-\bar{x})^2$, $=\mathrm{var}(\underline{x})$)

Standardisierung von Daten:

$$x_i^* := \frac{x_i - \bar{x}}{s} = \underbrace{-\frac{\bar{x}}{s}}_{=\alpha} + \underbrace{\frac{1}{s}}_{=\beta}\cdot x_i = \alpha + \beta x_i \quad \text{mit } \alpha = -\frac{\bar{x}}{s},\ \beta = \frac{1}{s}$$

mit $s = \sqrt{\mathrm{var}(\underline{x})}$ Standardabweichung!

Dann: $\bar{x}^* = \alpha + \beta\bar{x} = -\frac{\bar{x}}{s} + \frac{1}{s}\bar{x} = 0$

$$\mathrm{var}(\underline{x}^*) = \beta^2\cdot\mathrm{var}(\underline{x}) = \frac{1}{s^2}\cdot\mathrm{var}(\underline{x}) = \frac{\mathrm{var}(\underline{x})}{\mathrm{var}(\underline{x})} = 1$$

---

3) Kovarianz, Korrelation:

$\underline{X} = (X, Y)$ sei zweidimensionales Merkmal

z.B. $(X, Y) = (\text{Gewicht}, \text{Körpergröße})$

Ein einzelnes Datum $\underline{X}(\omega) = (X(\omega), Y(\omega))$

bzw. $\underline{x}_i = (x_i, y_i)$ für ein individuellen $\underline{x}_i$

Dann definiert man den Wert

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

Datenvektor Gewicht

Datenvektor Körpergröße

als Kovarianz von X und Y.

Empirischer Korrelationskoeffizient:

$$r(\underline{x}, \underline{y}) := \operatorname{cov}(\underline{x}^*, \underline{y}^*)$$

standardisierte Daten / Ausprägungen!!

Bemerkung: Verwendet man die spezielle lineare Transformation

$$x_i^* = -\frac{\bar{x}}{s_x} + \frac{1}{s_x} x_i \quad \text{und} \quad y_i^* = -\frac{\bar{y}}{s_y} + \frac{1}{s_y} y_i \quad (i = 1, \dots, n),$$

Standardabweichung bezüglich Merkmal X

Standardabweichung bezüglich Merkmal Y

dann erhält man als alternative Formel für den empirischen Korrelationskoeffizienten:

$$\operatorname{cov}(\underline{x}^*, \underline{y}^*) = \frac{1}{n} \sum_{i=1}^{n} x_i^* \cdot y_i^* = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$

$\bar{x}^* = \bar{y}^* = 0$

x_i^* $\qquad$ y_i^*

$$= \frac{1}{s_x \cdot s_y} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{\operatorname{cov}(\underline{x}, \underline{y})}{\sqrt{\operatorname{var}(\underline{x})} \cdot \sqrt{\operatorname{var}(\underline{y})}}$$

$= \operatorname{var}(\underline{x})$ $\qquad$ $= \sqrt{\operatorname{var}(\underline{y})}$ $\qquad$ $= \operatorname{cov}(\underline{x}, \underline{y})$

ENDE der Vorlesung !!