Vorlesung vom 08.11.2012:

Thema: Näheres zur Kovarianz und zur empirischen Korrelation

Formel zur Korrelation

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

Skalarprodukt

mit $$\operatorname{cov}(\underline{x},\underline{y}) = \frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y}) = \frac{1}{n}\cdot\begin{bmatrix} x_1-\bar{x} \\ \vdots \\ x_n-\bar{x}\end{bmatrix}\cdot\begin{bmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y}\end{bmatrix}$$

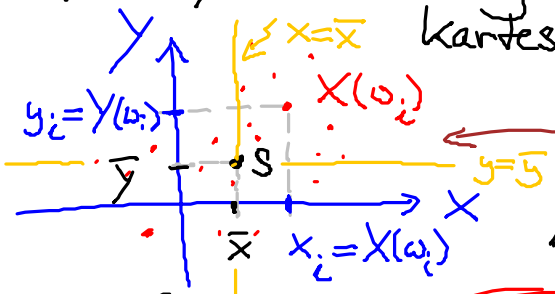
$= \underline{x}-\bar{x}\underline{1}$ $\quad$ $= \underline{y}-\bar{y}\cdot\underline{1}$

Gegeben dazu ist ein 2-dimensionales Merkmal (kardinales)

$\underline{X} = (X,Y)$ mit Ausprägungen $\underline{X}(\omega_i) = (X(\omega_i), Y(\omega_i)) = (x_i, y_i) \in \mathbb{R}^2$

(Untersuchungseinheit)

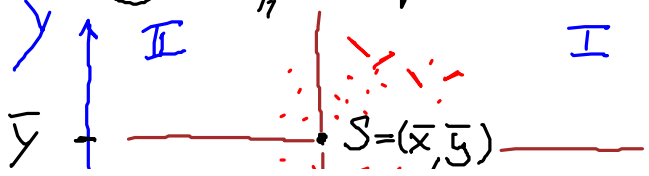
$\mathbb{R}^2 = \{(x,y) \mid x,y \in \mathbb{R}\}$: Menge aller Punkte der kartesischen Ebene.



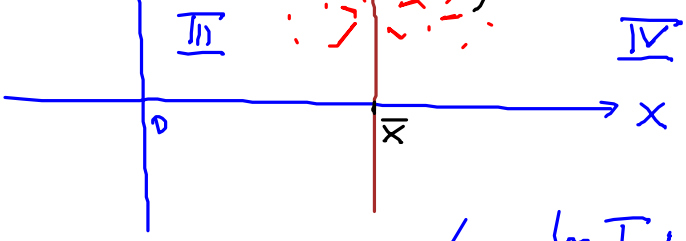
„Punktwolke"

„Schwerpunkt der Datenwolke": $S = (\bar{x},\bar{y})$ mit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Man untersucht die Datenwolke in Bezug zum „Schwerpunkt" $S = (\bar{x},\bar{y})$



Es entstehen durch das „Fadenkreuz" in Bezug zu S 4 Quadranten.



Punkte $\underline{X}(\omega_i) = (x_i, y_i)$ erfüllen bezüglich der Quadranten folgende Bedingungen:

In I: $x_i \geq \bar{x}$ und $y_i \geq \bar{y}$ $\Rightarrow (x_i - \bar{x}) \cdot (y_i - \bar{y}) \geq 0$ (≥ 0, ≥ 0)

In II: $x_i \leq \bar{x}$ — " — $y_i \geq \bar{y}$ $\Rightarrow (x_i - \bar{x}) \cdot (y_i - \bar{y}) \leq 0$ (≤ 0, ≥ 0)

In III: $x_i \leq \bar{x}$ — " — $y_i \leq \bar{y}$ $\Rightarrow (x_i - \bar{x}) \cdot (y_i - \bar{y}) \geq 0$ (≤ 0, ≤ 0)

In IV: $x_i \geq \bar{x}$ — " — $y_i \leq \bar{y}$ $\Rightarrow (x_i - \bar{x}) \cdot (y_i - \bar{y}) \leq 0$ (≥ 0, ≤ 0)

(+) · (−)

Diese Summanden werden in $\text{cov}(\underline{x}, \underline{y})$ aufsummiert !!

$\text{cov}(\underline{x}, \underline{y}) > 0$ zeigt ein „Übergewicht" in der Punktwolke für Quadrant I und III

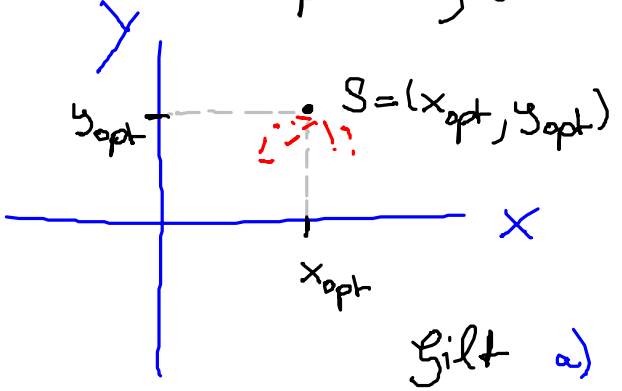
$\text{cov}(\underline{x}, \underline{y}) < 0$ — " — bezüglich Quadrant II und IV.

Dazu Aufgabe

„U13) $\underline{X} = (X, Y)$ mit X: Wassergehalt Boden, Y: Ernteertrag

$\underline{X}(\omega_i) = (X(\omega_i), Y(\omega_i))$ mit $x_i = X(\omega_i)$, $y_i = Y(\omega_i)$

Sei $x_{opt} = \bar{x}$, $\bar{y}$: mittlerer Ernteertrag. Schwerpunkt $S = (x_{opt}, \bar{y})$



Zu erwarten ist ein kausaler Zusammenhang zwischen X und Y.

Wie steht es um die Korrelation $\text{cov}(\underline{x}, \underline{y})$

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

Gilt a) $X \leq x_{opt} \Rightarrow \rho(X,Y) \geq 0$ und $X \geq x_{opt} \Rightarrow \rho(X,Y) \leq 0$

oder b) $X \leq x_{opt} \Rightarrow \rho(X,Y) \leq 0$ und $X \geq x_{opt} \Rightarrow \rho(X,Y) \geq 0$

oder weder a) noch b) ?

Wir untersuchen den Ausdruck $(X - x_{opt})(Y - y_{opt})$ bzw.

$$\text{cov}(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - x_{opt})(y_i - y_{opt})$$

$X < x_{opt}$, d.h. $x_i < x_{opt} \Rightarrow y_i < y_{opt} \Rightarrow (x_i - x_{opt})(y_i - y_{opt}) > 0$ ((−) · (−))

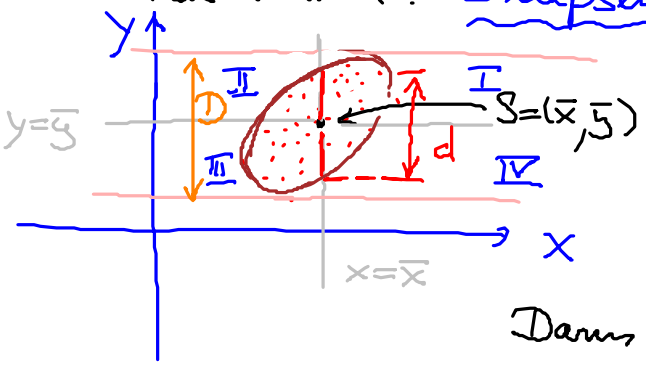
$X > x_{opt}$, d.h. $x_i > x_{opt} \Rightarrow y_i < y_{opt} \Rightarrow (x_i - x_{opt})(y_i - y_{opt}) < 0$ ((+) · (−))

$\Rightarrow \rho(X,Y) \geq 0$ für $X \leq x_{opt}$

$\rho(X,Y) \leq 0$ für $X \geq x_{opt}$

} Aussage (a) ist richtig !

Es gibt die Möglichkeit, den Korrelationskoeffizienten annähernd zeichnerisch zu ermitteln: Ellipsenregel!



Zeichne eine „elliptische Hülle" mit Mittelpunkt $S=(\bar{x},\bar{y})$ um die Punktwolke

vertikaler Gesamtdurchmesser: $D>0$

vertikaler Durchmesser an der Stelle $x=\bar{x}$: d

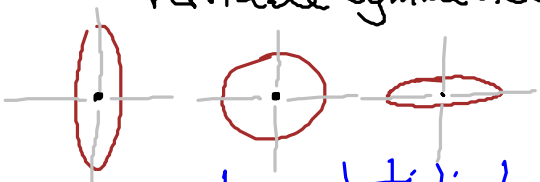
Dann gilt: $\boxed{r(x,y)^2 \approx 1-\left(\frac{d}{D}\right)^2}$ $\quad d \geq 0$

„ungefähr gleich"

Das heißt:

$d=0$, d.h. Ellipse = „Strich" $\Rightarrow r(x,y) \approx \pm 1$

$d \approx D$, d.h. Ellipse besitzt eine vertikale Symmetrieachse $\Rightarrow \frac{d}{D} \approx 1 \Rightarrow \boxed{r(x,y) \approx 0}$



$r(x,y) \approx 0 \Rightarrow$ kein statistischer Zusammenhang zwischen X und Y

$\Rightarrow$ –"– kausaler –"– zwischen X und Y.

$r(x,y) \approx \pm 1 \Rightarrow$ Es besteht ein enger statistischer Zusammenhang zwischen X und Y, der nicht unbedingt kausal begründet ist.

ENDE des Teils „Deskriptive Statistik" und der heutigen Vorlesung!