

Mathematik II für Biologen  
Beschreibende Statistik  
Zweidimensionale (bivariate) Daten

Stefan Keppeler

26. April 2013



## Prolog

## Zweidimensionale Stichproben

### Graphisch: Streudiagramm

- Lineare Regression
- Transformationen

### Numerisch: Korrelationen

- Produktmomenten-Korrelation
- Rangkorrelation
- Warnung



Schlagzeilen | Hilfe | RSS | Newsletter | Mobil | Wetter | TV-Programm

# SPIEGEL ONLINE WISSENSCHAFT

NACHRICHTEN VIDEO THEMEN FORUM ENGLISH DER SPIEGEL SPIEGEL TV ABO SHOP

Home Politik Wirtschaft Panorama Sport Kultur Netzwerk Wissenschaft Gesundheit einestages Karriere Uni Schule Reise Auto

Nachrichten > Wissenschaft > Natur > Artenschutz > Biodiversität in Europa: Artensterben geschieht mit Zeitverzögerung [Login](#) | [Registrierung](#)

## Biodiversität in Europa: Artensterben vollzieht sich mit Zeitverzögerung



Fotos ▶

Wolfgang Fabitsch

Das Artensterben ist bedrohlich genug, doch laut einer neuen Studie sieht es noch düsterer aus: Die Effekte von Zersiedlung und Umweltverschmutzung schlagen sich bei vielen Spezies erst Jahrzehnte später nieder. Die Zahl der bedrohten Arten müsste deutlich nach oben korrigiert werden.

<http://www.spiegel.de/wissenschaft/natur/a-894478.html>

“Die Forscher um Stefan Dullinger von der Universität Wien verglichen die **aktuellen Nationalen Roten Listen** der gefährdeten Arten mit der **wirtschaftlichen Entwicklung am Anfang, in der Mitte und gegen Ende des 20. Jahrhunderts**. Dabei beachteten sie Bevölkerungsdichte, Intensität der Landnutzung und Bruttoinlandsprodukt pro Kopf.

“Bei Pflanzen, Libellen und Heuschrecken **erklärte der Entwicklungsstand des Landes um das Jahr 1900 den derzeitigen Bedrohungsstatus am besten**. Bei Säugetieren und Reptilien waren die Daten von 1900 und 1950 ähnlich aussagekräftig. Nur der Status der Fischbestände ließ sich am besten von aktuellen Daten ableiten.”



Stichprobe  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  von Paaren von Zahlen.  
Oft:

- ▶  $x$ : Ausgangsgröße, “unabhängige” Variable
- ▶  $y$ : Zielgröße, Idee  $y = f(x)$ , “abhängige” Variable

**Beispiel 1:** Grille (vgl. Mathematik I, Aufgabe 60)

- ▶  $x_i$ : Temperatur [ $^{\circ}\text{C}$ ]
- ▶  $y_i$ : Zirpfrequenz (Tonhöhe) [ $1/\text{s}$ ]

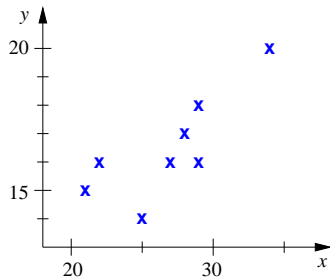
$x_i$	21	22	25	27	28	29	29	34
$y_i$	15	16	14	16	17	16	18	20



$x_i$	21	22	25	27	28	29	29	34
$y_i$	15	16	14	16	17	16	18	20

## Beschreibungsmöglichkeiten

- ▶ Wende Methoden für eindimensionale Stichproben getrennt auf  $x$  und  $y$  an.  
**Nachteil:** Zusammenhang zwischen  $x$  und  $y$  geht verloren.
- ▶ Graphisch: **Streudiagramm** (scatter plot)



Falls Streudiagramm eine Gerade suggeriert: **Lineare Regression**  
(siehe Mathematik I, Vorlesung 14)

$$y(x) = mx + b + \text{“kleiner Fehler”}$$

Wähle  $m$  und  $b$  so, dass

$$\sum_{i=1}^n (y_i - (mx_i + b))^2$$

minimal. Ergebnis:

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad b = \bar{y} - m\bar{x}$$

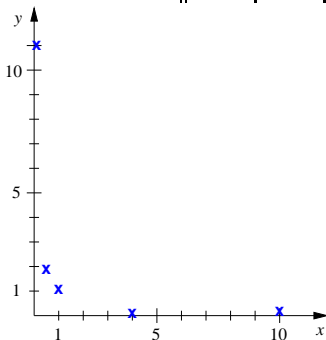


Manchmal erinnert das Streudiagramm erst nach **Transformation(en)** an eine Gerade,

$$x_i \mapsto g(x_i), \quad y_i \mapsto f(y_i).$$

**Beispiel 2:** Andere Stichprobe

$x_i$	0,1	0,5	1,0	4,0	10
$y_i$	11	1,9	1,1	0,15	0,2



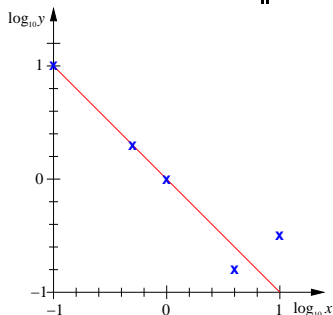
Sieht nicht nach Gerade aus...  
 Vielleicht Potenzgesetz?





$x_i$	0,1	0,5	1,0	4,0	10
$y_i$	11	1,9	1,1	0,15	0,2

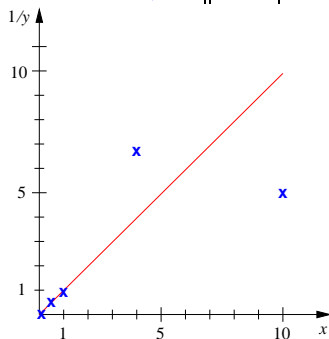
$\log_{10} x_i$	-1,0	-0,3	0,0	0,6	1,0
$\log_{10} y_i$	1,0	0,3	0,0	-0,8	-0,7



Ungefähr Gerade mit Steigung  $-1$ .  
Also wäre auch  $y_i \mapsto 1/y_i$  gut gewesen...

$x_i$	0,1	0,5	1,0	4,0	10
$y_i$	11	1,9	1,1	0,15	0,2

$x_i$	0,1	0,5	1,0	4,0	10
$1/y_i$	0,1	0,5	0,9	6,7	5,0



Gerade mit Steigung 1?



Die **Produktmomenten-Korrelation**  $r_{xy}$  nach **Pearson** misst die Stärke eines **linearen** Zusammenhangs zwischen  $x$  und  $y$ ,

$$r_{xy} := \frac{s_{xy}}{s_x s_y},$$

wobei:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{Stichprobenkovarianz,}$$

$$s_x, s_y \quad \text{Standardabweichungen.}$$

Für den Wert gilt immer:  $-1 \leq r_{xy} \leq 1$ , denn...

**Interpretation:** Kosinus des Winkels zwischen den Vektoren

$$\vec{a} = (x_1 - \bar{x}, \dots, x_n - \bar{x})$$
$$\vec{b} = (y_1 - \bar{y}, \dots, y_n - \bar{y}) \quad , \quad r_{xy} = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|} = \cos \angle(\vec{a}, \vec{b})$$



Für den Wert gilt immer:  $-1 \leq r_{xy} \leq 1$ .

Je näher  $|r_{xy}|$  bei 1, desto stärker ist der lineare Zusammenhang zwischen  $x$  und  $y$ .

$|r_{xy}| = 1$       perfekter linearer Zusammenhang

$r_{xy} \approx 0$       kein linearer Zusammenhang

Vorzeichen (VZ):

VZ von  $r_{xy} =$  VZ der Steigung  $m$  der Regressionsgeraden

### Beispiele:

- ▶ "Grille":  $r_{xy} = 0,8$
- ▶ Beispiel 2:  $r_{xy} = -0,5$
- ▶ weitere qualitativ...



Die **Rangkorrelation** nach **Spearman** misst die Stärke eines **monotonen** Zusammenhangs, zwischen  $x$  und  $y$ ,

$$r_{xy}^{(\text{SP})} = r_{\text{Rang}(x) \text{Rang}(y)}$$


**In Beispiel 2:**

$x_i$	0,1	0,5	1,0	4,0	10
$y_i$	11	1,9	1,1	0,15	0,2

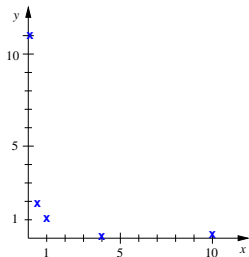
Rang $x_i$	1	2	3	4	5
Rang $y_i$	5	4	3	1	2

$$r_{xy}^{(\text{SP})} = -0,9, \text{ aber } r_{xy} = -0,5:$$

Monotoner Zusammenhang, aber nicht linear.

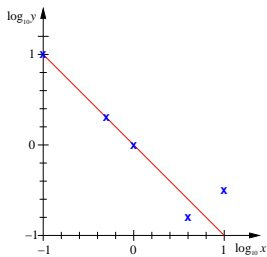
**Übrigens:**  $r_{xy}^{(\text{SP})}$  robust,  $r_{xy}$  nicht. 





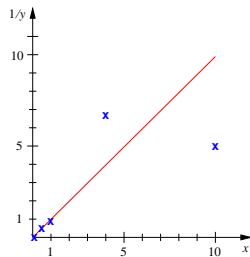
$$r_{xy} = -0,5$$

$$r_{xy}^{(SP)} = -0,9$$



$$r = -0,97$$

$$r^{(SP)} = -0,9$$



$$r = 0,74$$

$$r^{(SP)} = 0,9$$

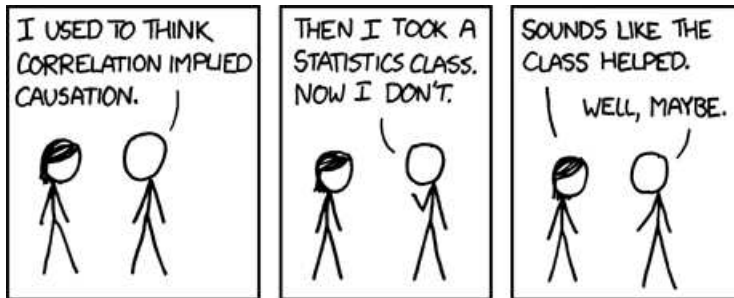
$|r_{xy}^{(SP)}|$  ändert sich nicht bei monotoner Transformation.



## Vorsicht: Interpretation von Korrelationen nicht einfach!

- ▶  $r$  (mit kleinem Betrag) kann rein zufällig von Null verschieden sein. Ob zufällig oder nicht: Schließende Statistik (später)
- ▶ Eine Korrelation  $r \neq 0$  sagt nichts über einen ursächlichen Zusammenhang. Viele Möglichkeiten:
  - ▶  $x$  beeinflusst  $y$ .
  - ▶  $y$  beeinflusst  $x$ .
  - ▶  $x$  und  $y$  haben eine gemeinsame Ursache  $z$ .
  - ▶ Schein-Korrelationen, z.B.: Seien  $x$ ,  $y$ ,  $z$  unkorreliert. Dann sind  $x/z$  und  $y/z$  automatisch korreliert.
  - ▶ V.a. bei Zeitreihen: Unabhängige lineare Trends in  $x$  und  $y$  führen zu “Unsinn-Korrelationen”. Beispiel:  
 $x_i = \# \text{ Störche im Jahr } 1900 + i$   
 $y_i = \# \text{ Geburtenrate im Jahr } 1900 + i$   
 $r_{xy}$  deutlich von Null verschieden  $\Rightarrow$  ???





<http://xkcd.com/552>