



Einführung in die Statistik

Praktische Übung – Jürgen Hermes – IDH – SoSe 2023

Programm heute

- Dispersionsmaße: Varianz und Standardabweichung
- Korrelation
- Regression (nur initial)
- Zusammenfassung / Hausaufgabe



Diese und die folgenden Folien sind erstellt worden von Sascha Wolfer für seinen Kurs "Statistik mit R" an der Uni Basel. Ich nutze sie mit seiner freundlichen Genehmigung. DOI für die Materialien ist

[10.5281/zenodo.7431504](https://doi.org/10.5281/zenodo.7431504)

Korrelation

Korrelationen sind überall

- Speed-accuracy trade-off (Salthouse, 1979): Je schneller man etwas tut, desto ungenauer tut man es.
- Je mehr PS ein Auto hat, desto größer die Beschleunigung.
- Je häufiger ein Wort ist, desto kürzer ist es (Zipf, 1949).
- Je mehr Feuerwehrleute vor Ort sind, desto höher ist die Schadenssumme.
- Je mehr Störche in einem Gebiet leben, desto mehr Kinder werden dort geboren.

Korrelation

- Der **Korrelationskoeffizient** beschreibt den Zusammenhang zweier Variablen.
 - Pearson oder Spearman
- Der Korrelationskoeffizient kann Werte zwischen -1 und +1 annehmen.
 - +1: positiver Zusammenhang (je mehr x , desto mehr y)
 - -1: negativer Zusammenhang (je mehr x , desto weniger y)
 - 0: kein Zusammenhang

Korrelation

Table 1

Interpretation of the Pearson's and Spearman's correlation coefficients.

Correlation Coefficient	Dancey & Reidy (Psychology)	Quinnipiac University (Politics)	Chan YH (Medicine)
+1	–1	Perfect	Perfect
+0.9	–0.9	Strong	Very Strong
+0.8	–0.8	Strong	Very Strong
+0.7	–0.7	Strong	Moderate
+0.6	–0.6	Moderate	Moderate
+0.5	–0.5	Moderate	Fair
+0.4	–0.4	Moderate	Fair
+0.3	–0.3	Weak	Fair
+0.2	–0.2	Weak	Poor
+0.1	–0.1	Weak	Poor
0	0	Zero	None

Was als schwache, moderate und starke Korrelation gilt, variiert zwischen (und auch in) Disziplinen.

Taxifahren in London

- Datenlage: Anzahl der Dienstjahre von Personen, die in London Taxi fahren, korreliert positiv mit der Größe eines Teils des Hippocampus (zuständig u.a. für Orientierung).
- Untersucht wurden Personen mit unterschiedlicher Erfahrung.
- Folgerung: Taxifahren führt zur Vergrößerung des Hippocampus.



Interpretation von Korrelationen

- Nehmen wir an, zwei Variablen x und y korrelieren positiv (je mehr x , desto mehr y).
- Interpretation 1: Einseitige Steuerung x bewirkt y (oder andersherum)



Interpretation von Korrelationen

- Nehmen wir an, zwei Variablen x und y korrelieren positiv (je mehr x , desto mehr y).
- Interpretation 2: Gegenseitige Steuerung
 x wirkt auf y , y wirkt auf x zurück



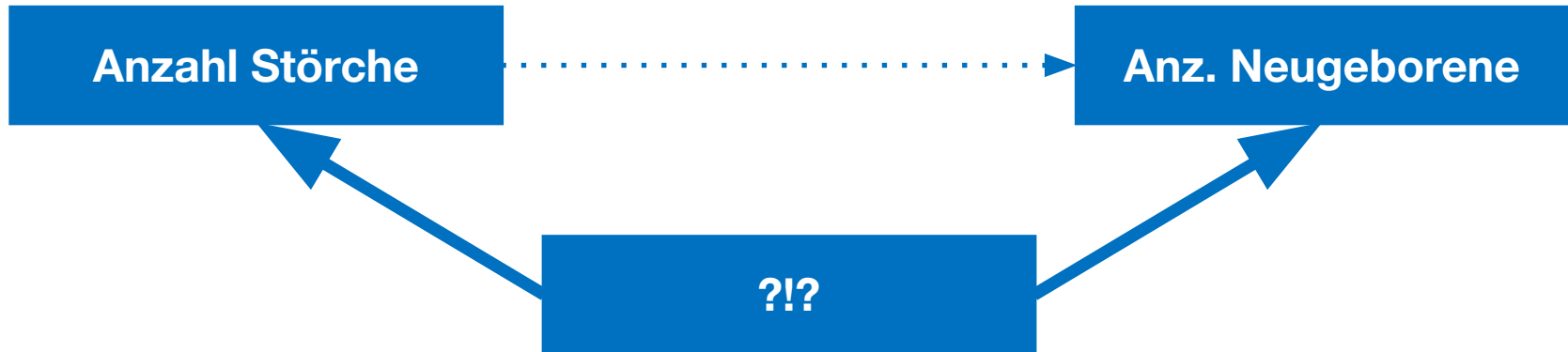
Interpretation von Korrelationen

- Nehmen wir an, zwei Variablen x und y korrelieren positiv (je mehr x , desto mehr y).
- Interpretation 3: Drittseitige Steuerung
 x und y hängen von einer dritten Variable z ab



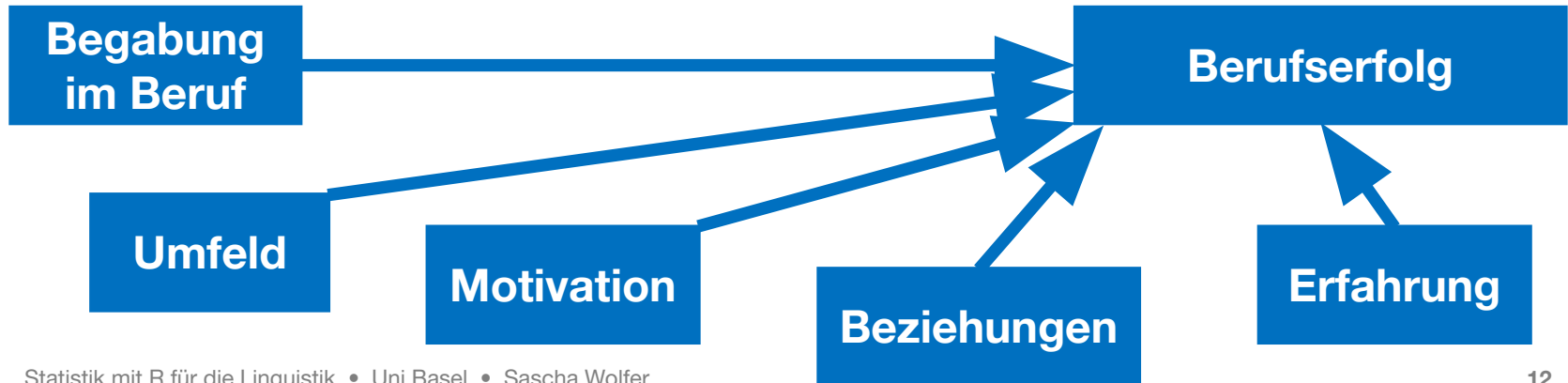
Interpretation von Korrelationen

- Nehmen wir an, zwei Variablen x und y korrelieren positiv (je mehr x , desto mehr y).
- Interpretation 3: Drittseitige Steuerung
 x und y hängen von einer dritten Variable z ab



Interpretation von Korrelationen

- Nehmen wir an, zwei Variablen x und y korrelieren positiv (je mehr x , desto mehr y).
- Interpretation 4: Komplexe Steuerung
Das Bedingungsgefüge ($a, b, c \dots x$) bewirkt y .

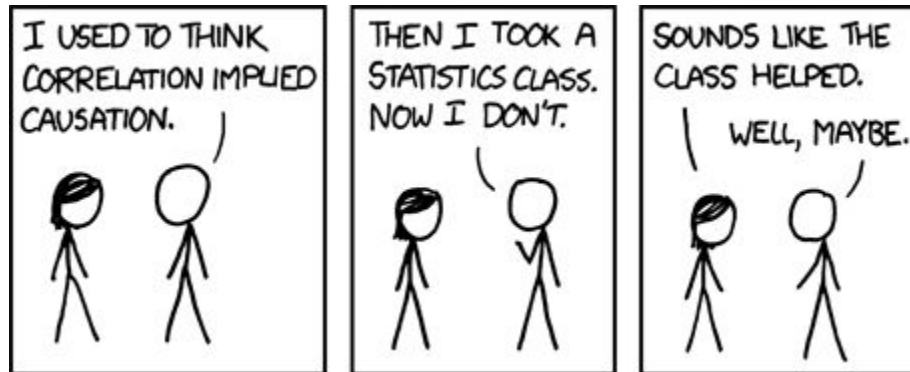


Interpretation von Korrelationen

- In der "echten Welt" haben wir es fast immer mit komplexen Steuerungen zu tun.
- Die Frage ist, wie wir mit den **Kovariaten** umgehen:
 - Konstant halten (wie im Labor)
 - Zufällig verteilen
 - Rechnerisch kontrollieren (also über statistische Modelle)
 - Vernachlässigen

Interpretation von Korrelationen

- Korrelation = gemeinsames Variieren von Variablen
- Vorsicht bei der **kausalen** Interpretation von Korrelationen!



Interpretation von Korrelationen

- Korrelationen sind zunächst einmal **Koinzidenzen**.
- Kausaler Zusammenhang kann – wenn überhaupt – nur angenommen werden, wenn eine Variable **systematisch variiert** wird.
 - Trainingszeit → Erfolg
 - Dosis → Wirkung
- Experimentelle Designs!
- Interpretation von Korrelationen bleibt immer **bidirektional**.

Pearson-Korrelation: Berechnung

Abweichungen jedes x-Werts
von seinem Mittelwert

Abweichungen jedes y-Werts
von seinem Mittelwert

Produkte werden
aufsummiert

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot s_x \cdot s_y}$$

Nenner: Produkt aus
Stichprobengröße und den
beiden Standardabweichungen

Abweichungen werden für
jedes Wertepaar multipliziert.

Pearson-Korrelation: Berechnung

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot s_x \cdot s_y}$$

- Was geschieht mit den Produkten im Zähler, wenn x_i und y_i beide nach oben von ihrem Mittelwert abweichen?
- Was geschieht mit den Produkten, wenn x_i nach unten und y_i nach oben vom Mittelwert abweicht?
- Was geschieht mit den Produkten, wenn x_i gleich seinem Mittelwert ist und y_i in eine Richtung abweicht?

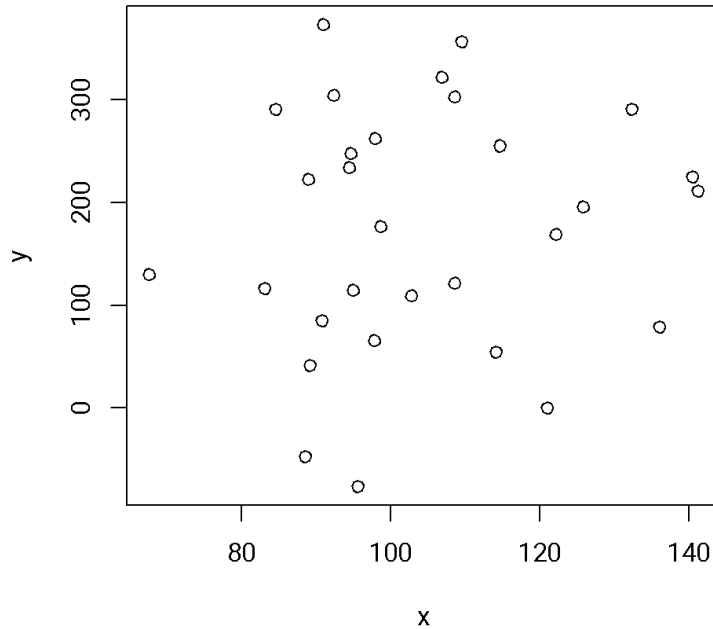
Pearson-Korrelation: Berechnung

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n \cdot s_x \cdot s_y}$$

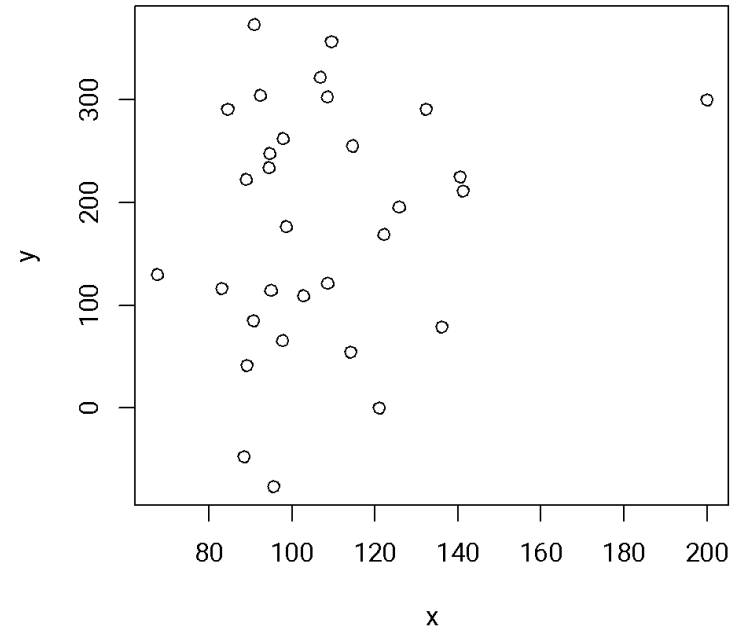
- Was geschieht mit r , wenn die Summe der Produkte sehr hoch positiv / negativ ist (der **Effekt**)?
- Was geschieht mit r , wenn eine der beiden Standardabweichungen s_x oder s_y sehr hoch ist (das **Rauschen**)?

Pearson-Korrelation: Ausreisser

$r = 0,099$



$r = 0,202$



Rangkorrelation nach Spearman

- Verwendung von Rangplätzen anstatt tatsächlicher Werte

$$r_{Sp} = \frac{\sum_{i=1}^n (\text{rang}(x_i) - \overline{\text{rang}(x)}) (\text{rang}(y_i) - \overline{\text{rang}(y)})}{\sqrt{\sum_{i=1}^n (\text{rang}(x_i) - \overline{\text{rang}(x)})^2} \cdot \sqrt{\sum_{i=1}^n (\text{rang}(y_i) - \overline{\text{rang}(y)})^2}}$$

- Einige Vorteile:
 - Testet nicht nur auf linearen Zusammenhang
 - Macht keine Vorannahmen zur zugrundeliegenden Verteilung
 - Weniger anfällig für Effekte von Ausreisserwerten
- Rangkorrelation gilt gemeinhin als etwas konservativer.

Korrelation in R

- Funktion: `cor(x, y, method)`
- Argument `method`:
 - Default: "pearson"
 - Rangkorrelation: "spearman"
- `cor(x, y, method = "spearman")`
 - Benutzt den Rangkorrelationskoeffizienten von Spearman
- ~~Auch hier: Argument `na.rm`~~ Hier: Argument `use = "complete.obs"`, um NA-Werte auszuschließen

Zusammenhänge

- Datensatz zu mehreren Ländern der Erde mit den folgenden Variablen:
 - Geburtenrate (wie viele Menschen werden pro 1000 pro Jahr geboren)
 - Sterberate (wie viele Menschen von 1000 sterben pro Jahr)
 - Kindersterblichkeit (wie viele Kinder sterben pro 1000)
 - Lebenserwartung bei Geburt (wie alt werden heute Neugeborene)
 - Bevölkerungswachstum
- Wie können diese Variablen untereinander korrelativ zusammenhängen?
 - ++ / + / o / - / --

Zusammenhänge

	Geburtenrate	Sterberate	Kindersterblichkeit	Lebenserwartung	Bevölkerungswachstum
Geburtenrate	1				
Sterberate		1			
Kindersterblichkeit			1		
Lebenserwartung				1	
Bevölkerungswachstum					1

Visualisierung

- Stark ausgeprägte (+/-) Korrelationen lassen sich in einem Plot über beide Variablen gut erkennen.
 - `plot(X, Y)`
- Über einen Boxplot kann man nicht nur die Verteilung einer Variable visualisieren, sondern auch unterschiedliche Verteilungen in Abhängigkeit von einer anderen Variable:
 - `boxplot(X)`
 - `boxplot(X~Y)`

Zusammenfassung

- Korrelationskoeffizient gibt an, wie stark zwei Variablen **kovariieren** (= gemeinsam von ihrem jeweiligen Mittelwert abweichen).
- Vorsicht bei:
 - Interpretation von **Kausalität**
 - Interpretation der **Effektrichtung** (Korrelation ist immer bidirektional!)
 - Ausreißern
- Bei der **Rangkorrelation** wird mit dem Rang der Werte und nicht mit den Werten selbst gerechnet.

Begriffe

Korrelationskoeffizient

Pearson

Spearman

Einseitige Steuerung

Gegenseitige Steuerung

Drittseitige Steuerung

Komplexe Steuerung

Kovariaten

Korrelation & Kausalität

Koinzidenz

bidirektional

Effekt & Rauschen



Hausaufgabe

- Berechnen Sie die restlichen Pearson-Korrelationen zwischen den Variablen in der Tabelle auf S.23.
- Visualisieren Sie mindestens die Korrelation zwischen Geburtenrate und Bevölkerungswachstum. In diesem Plot sehen Sie zwei extreme Ausreißer. Ermitteln Sie die zugehörigen Länder dazu. Worauf ist das Verhalten dieser beiden Ausreißer wahrscheinlich zurückzuführen? Können Sie das anhand von anderen im `cia_factbook` vorliegenden Daten bestätigen?
- Wie hoch ist die Spearman-Korrelation zwischen Hubraum (`displacement`, `disp`) und PS (horse power, `hp`) im Datensatz `mtcars`? Bilden Sie vor der Berechnung eine Hypothese zur Richtung der Korrelation heraus.