Einführung in die Statistik

Praktische Übung – Jürgen Hermes – IDH – SoSe 2024

Statistik (Stochastik) – Die Grundlagen

- Deskriptive Statistik
  - Mittelwerte
  - **Streuungsmaße**
  - **Korrelation**
  - Regression
- Inferenzstatistik

Dispersionsmaße in R

- Interquartilabstand: `IQR(x)`, Range: `range(x)`
- Perzentile: `quantile(x, <Wert>)` – Prozentwert zwischen 0 und 1
- `summary(x)` gibt 6 Werte aus:
  - Minimum & Maximum
  - 1. Quartil, 2. Quartil (Median), 3. Quartil
  - Mittelwert
- Varianz: `var(x)`
- Standardabweichung: `sd(x)` – *standard deviation*
- Achtung: `var(x)` und `sd(x)` werden mit $n - 1$ im Nenner berechnet (Schätzung von Populationsparametern).

Zusammenfassung

- In der deskriptiven Statistik beschreiben wir die **Stichprobe**, die wir gesammelt haben.
- Wir können unterscheiden zwischen Häufigkeitsverteilungen (**diskrete** Variablen) und Dichteverteilungen (**kontinuierliche** Variablen).
- Bei kontinuierlichen Variablen wollen wir wissen:
  - Wo ist die Mitte der Verteilung? ➜ **Maße der zentralen Tendenz**
  - Wie stark streut die Variable? ➜ **Streuungs-/Dispersionsmaße**

Zusammenfassung

- Maße der zentralen Tendenz:
  - **Modus / Modalwert**
  - **Median**
  - **Arithmetischer Mittelwert**
- Modalwert ist nur für diskrete Variablen sinnvoll.
- Mittelwert ist anfälliger für **Ausreisserwerte** als der Median.

Zusammenfassung

- Streuungsmaße:
  - **Interquartilabstand (IQR) / Spannweite**
  - **Varianz**
  - **Standardabweichung**
- **Quartile** teilen Daten in Viertel.
- Unter dem *x*ten **Perzentil** liegen *x* Prozent aller Datenpunkte.
  - Auch "**Quantil**" genannt.

Zusammenfassung Skalen / Mittelwerte / Dispersionsmaße

Niveau	Häufigkeit	Rangfolge	Abstand	Nullpunkt	Zentrale Tendenz	Dispersions-maß
Nominal	messbar				Modus	
Ordinal	messbar	messbar			Mod + Median	IQA / Spannweite
Intervall	messbar	messbar	messbar		Mod + Med + arithmetisches Mittel	IQA / Spannw. + Varianz / SA
Verhältnis	messbar	messbar	messbar	absolut	Mod + Med + arithmetisches Mittel	IQA / Spannw. + Varianz / SA

Begriffe

Deskriptive Statistik

Stichprobe

Population

Verteilungen

Maße der zentralen Tendenz

Dispersionsmaße

Modus / Modalwert

Median

Mittelwert

Interquartilabstand (IQR)

Spannweite / Range

Quartil

Perzentil

Boxplot

Varianz

Standardabweich.

Hausaufgabe (WH aus der letzten Sitzung)

- Laden Sie die Datei `Exp.csv` aus ILIAS in R. Dabei handelt es sich um Experimentdaten, in denen Reaktionszeiten unter verschiedenen Bedingungen gemessen wurden.
- Berechnen Sie Mittelwert und Median für die Spalte `RT`. (`NA` ist keine gültige Lösung!)
- Identifizieren Sie den Grund dafür, dass Median und Mittelwert so weit auseinanderliegen (schreiben Sie diesen in einen Kommentar).
- Erstellen Sie ein einfaches Diagramm der Daten in der Spalte RT: `plot(<Spalte>)`
- Berechnen Sie den Mittelwert für jede Bedingung (Spalte `Bedingung`)
- Berechnen Sie IQR, Spannweite, Varianz und Standardabweichung für die Spalte `RT`.
- Lassen Sie sich eine Summary der Spalte geben.
- Erstellen Sie einen Boxplot der Spalte.

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

Korrelation

Korrelationen sind überall

Salthouse, T. A. (1979). Adult age and the speed-accuracy trade-off. *Ergonomics*, 22(7), 811–821.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.*

- 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

Akoglu, H. (2018). User's guide to correlation coefficients. Turkish Journal of Emergency Medicine, 18(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

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	Perfect
+0.9	−0.9	Strong	Very Strong	Very Strong
+0.8	−0.8	Strong	Very Strong	Very Strong
+0.7	−0.7	Strong	Very Strong	Moderate
+0.6	−0.6	Moderate	Strong	Moderate
+0.5	−0.5	Moderate	Strong	Fair
+0.4	−0.4	Moderate	Strong	Fair
+0.3	−0.3	Weak	Moderate	Fair
+0.2	−0.2	Weak	Weak	Poor
+0.1	−0.1	Weak	Negligible	Poor
0	0	Zero	None	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

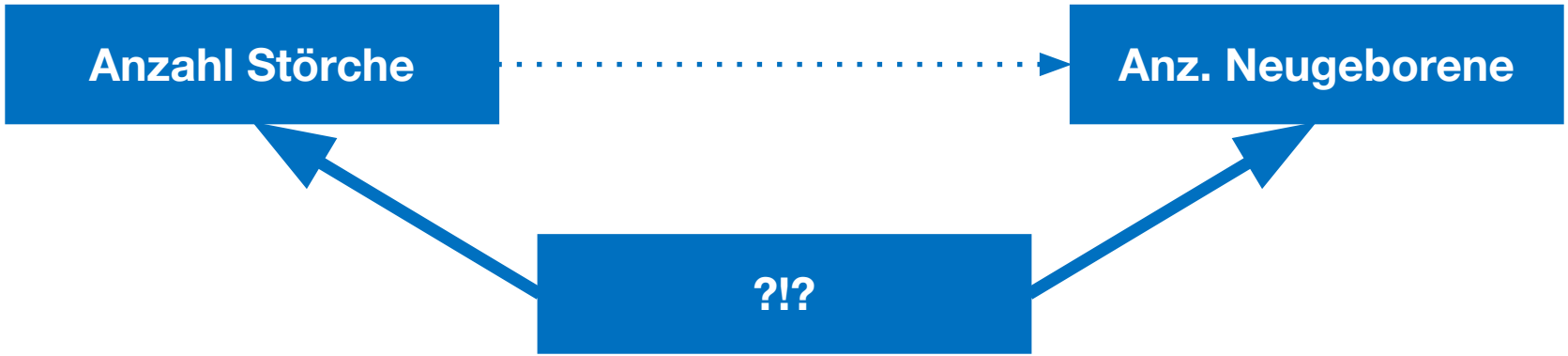
- Nehmen wir an, zwei Variablen *x* und *y* korrelieren positiv (je mehr *x*, desto mehr *y*).
- Interpretation 2: Gegenseitige Steuerung
  x wirkt auf *y*, *y* wirkt auf *x* zurück

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

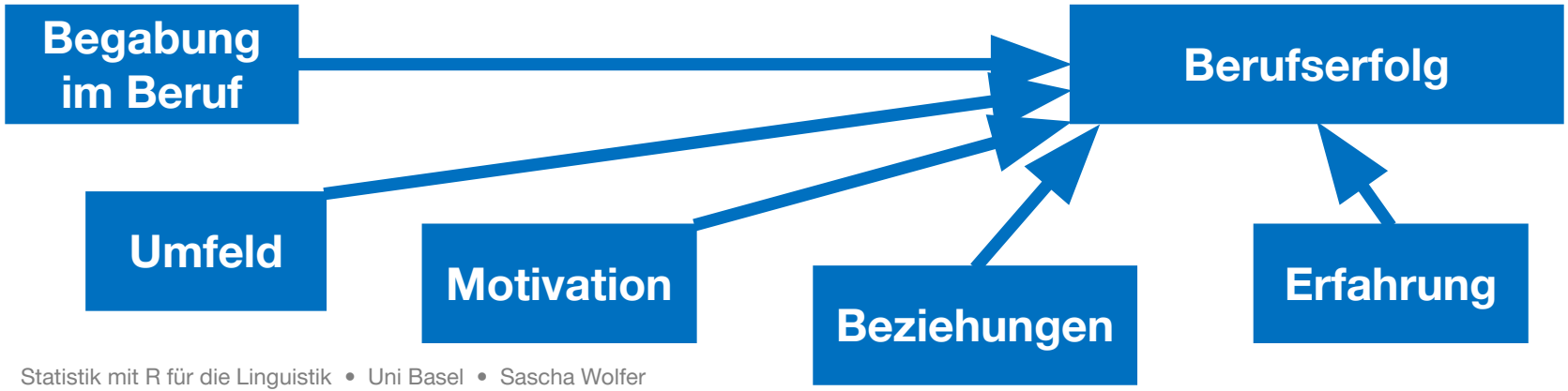
Interpretation von Korrelationen

- 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* ... *x*) bewirkt *y*.



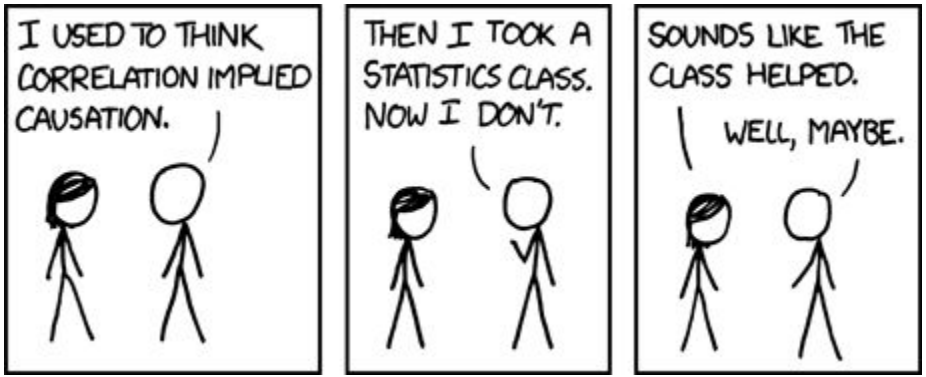
Interpretation von Korrelationen

- 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

https://xkcd.com/552/

Interpretation von Korrelationen

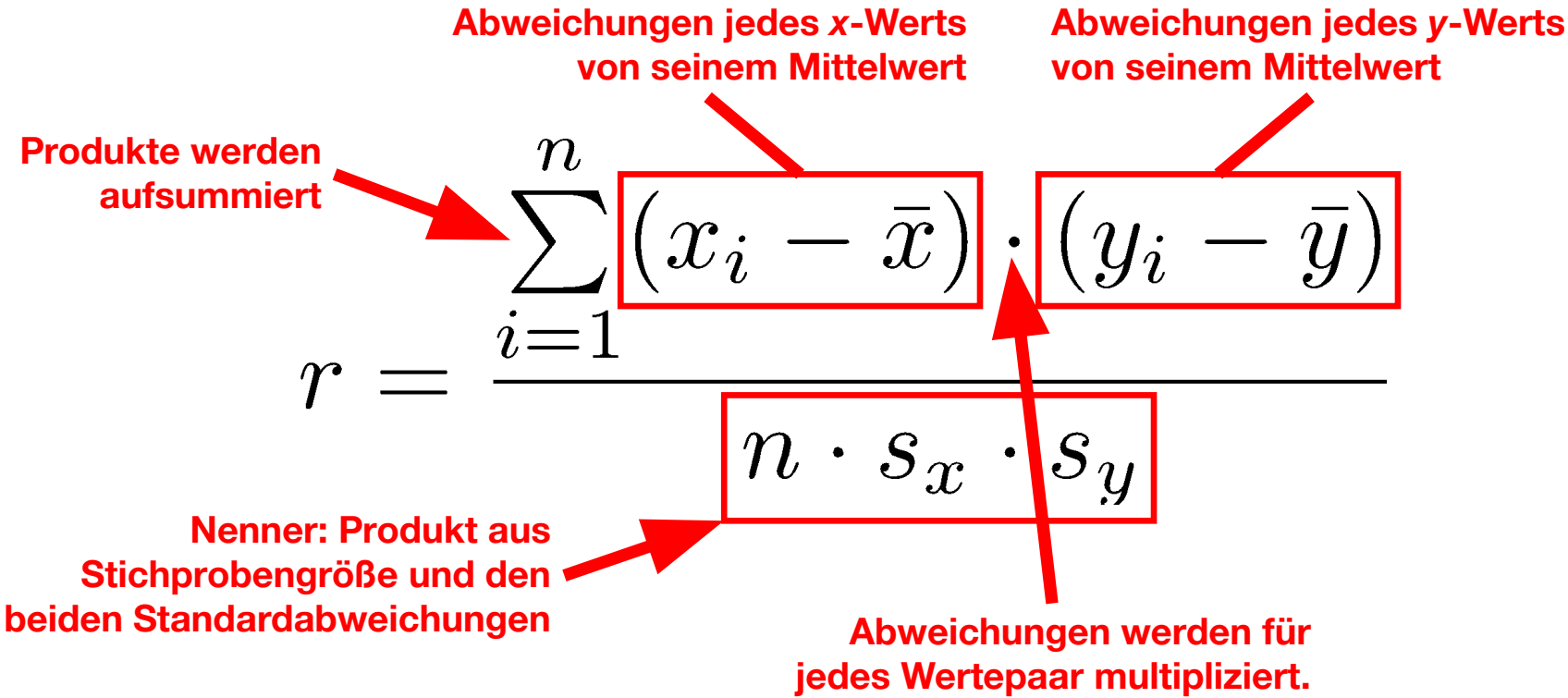
- Korrelation = gemeinsames Variieren von Variablen
- Vorsicht bei der **kausalen** Interpretation von Korrelationen!



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



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?

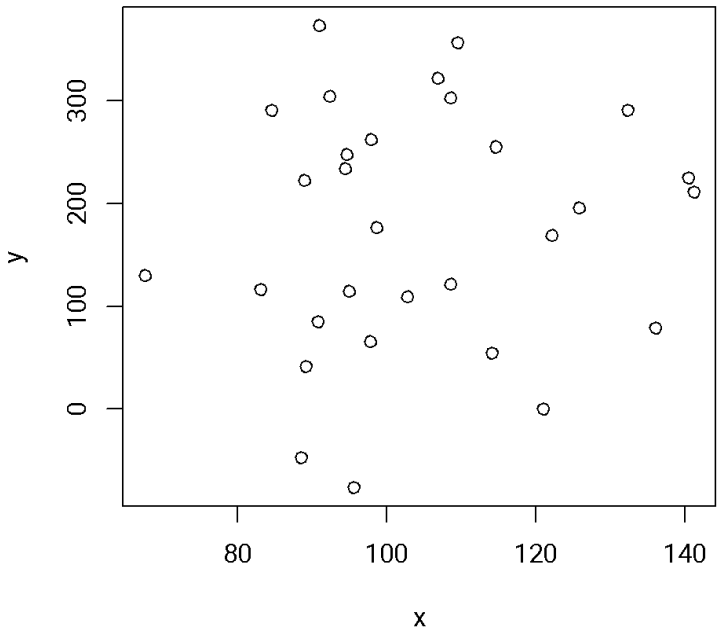
Pearson-Korrelation: Berechnung

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

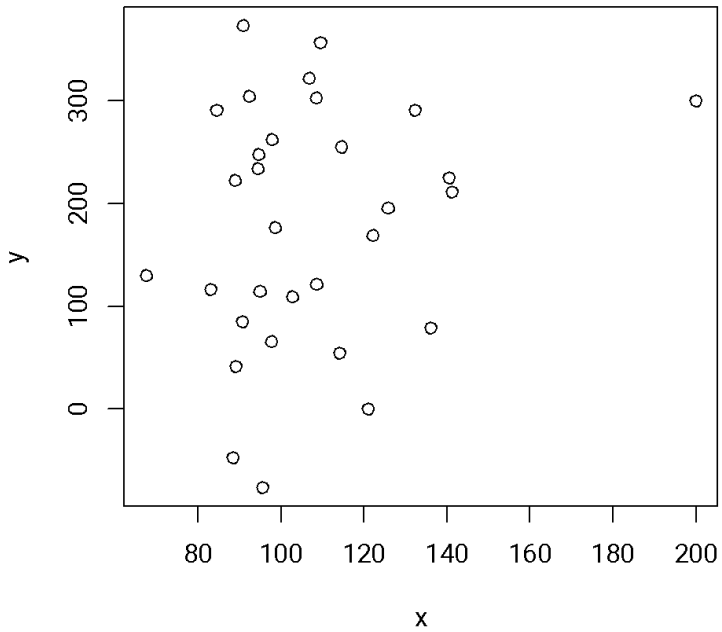
- Was geschieht mit *r*, wenn die Summe der Produkte sehr hoch positiv / negativ ist (der **Effekt**)?
- Was geschieht mit *r*, wenn eine der beiden Standardabweichungen s_x oder s_y sehr hoch ist (das **Rauschen**)?

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
- Hier nicht “`na.rm=T`” sondern: 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

	Geburten-rate	Sterberate	Kinder-sterblichkeit	Lebens-erwartung	Bevölkerungs-wachstum
Geburtenrate	1				
Sterberate		1			
Kinder-sterblichkeit			1		
Lebens-erwartung				1	
Bevölkerungs-wachstum					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)

 Völlig neue Folie (Jürgen Hermes)

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 der Folie “Zusammenhänge”. Die entsprechenden Werte finden Sie im `openintro::cia_factbook`
- 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 `openintro::mtcars`? Bilden Sie vor der Berechnung eine Hypothese zur Richtung der Korrelation heraus.