

# Korrelation & Kausalität

Simon Schölzel, M.Sc.

(updated: 02.06.2022)

1

Einführung

2

Das Kontrollierte Experiment

3

Lösungsansatz A: Das Natürliche Experiment

4

Lösungsansatz B: Kontrollvariablen

5

Fazit

1	Einführung
2	Das Kontrollierte Experiment
3	Lösungsansatz A: Das Natürliche Experiment
4	Lösungsansatz B: Kontrollvariablen
5	Fazit

**Correlation:** a *relation* existing between phenomena or things or between mathematical or statistical variables which *tend to vary, be associated, or occur together* in a way *not expected on the basis of chance alone*.  
([Merriam-Webster](#))

### Korrelationskoeffizient nach Bravais-Pearson:

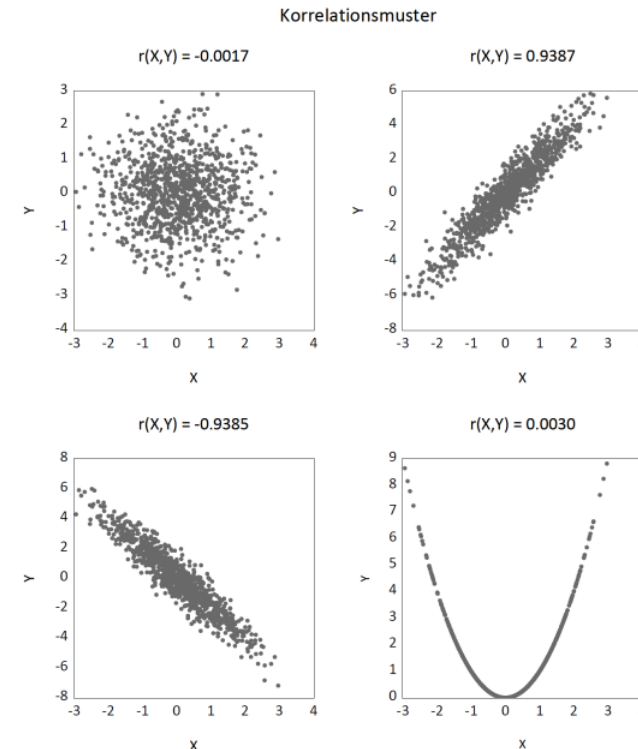
- Maß für den linearen Zusammenhang

$$r_{XY} = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2} \cdot \sqrt{\sum_{i=1}^n y_i^2 - n \cdot \bar{y}^2}}$$

### Rangkorrelationskoeffizient nach Spearman:

- Maß für monotone Zusammenhänge

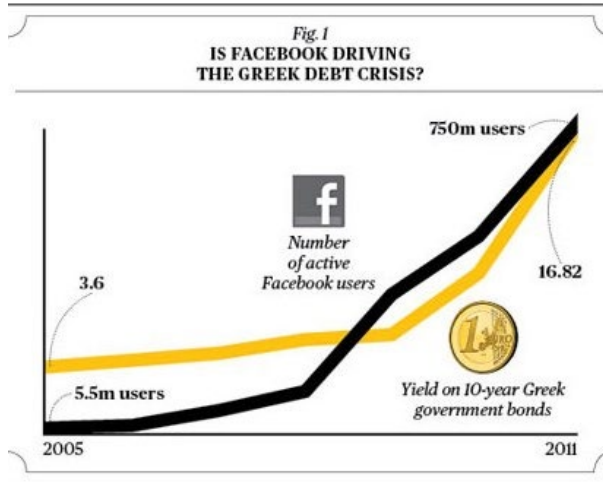
$$r_{XY}^R = \frac{\sum_{i=1}^n (R_X(x_i) - \overline{R_X}) \cdot (R_Y(y_i) - \overline{R_Y})}{\sqrt{\sum_{i=1}^n (R_X(x_i) - \overline{R_X})^2} \cdot \sqrt{\sum_{i=1}^n (R_Y(y_i) - \overline{R_Y})^2}}$$



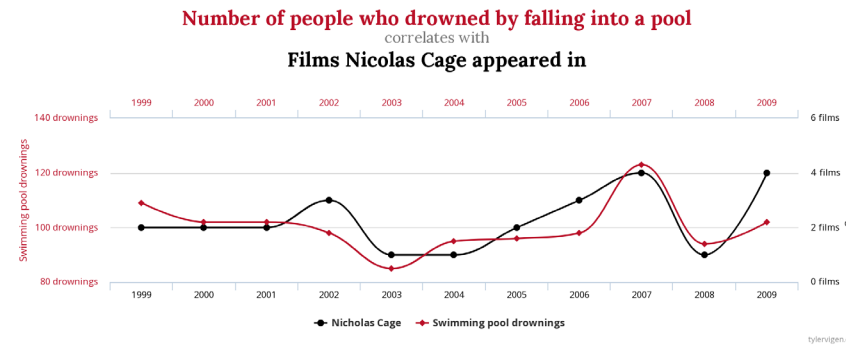
Quelle: Statistik I WS 2021/22, Bernd Wilfing  
vgl. auch: DLAK – Voresung 1, Folie 16

# 1 Einführung

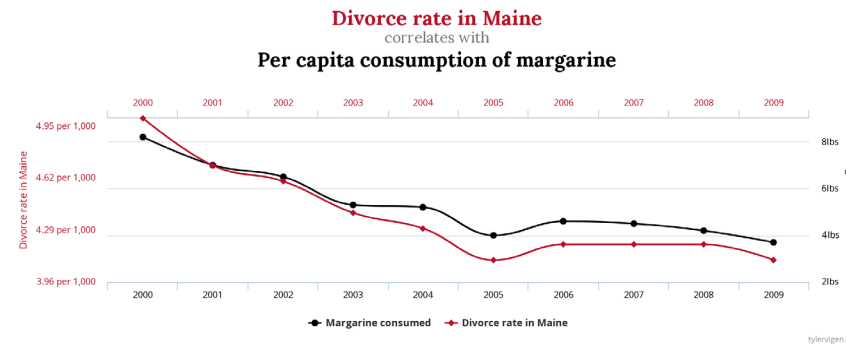
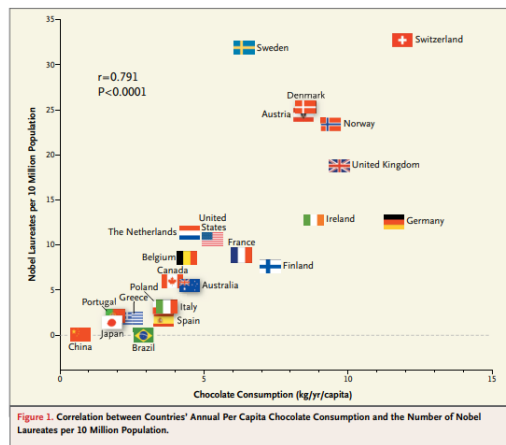
## 1.2 Korrelation oder Kausalität?



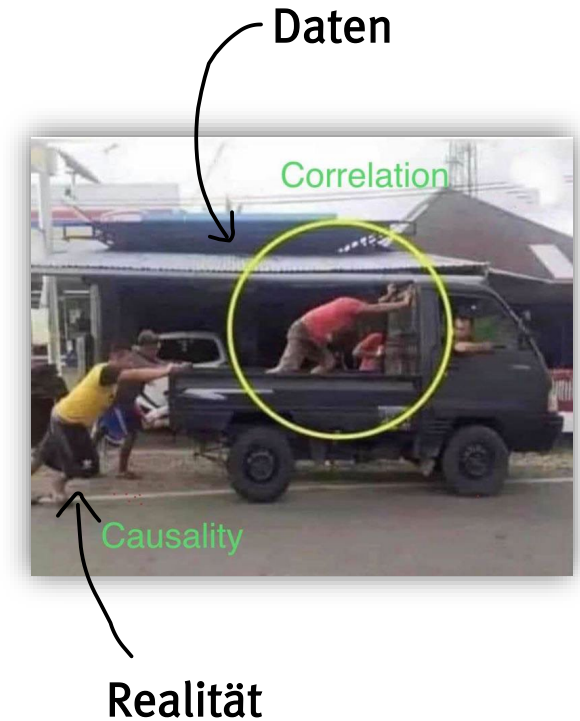
Quelle: [Jeremy Bertomeu](#)



Quelle: [New England Journal of Medicine](#)



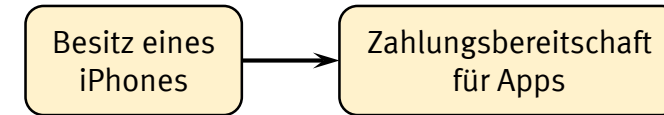
Quelle: [Spurious-Correlations](#)



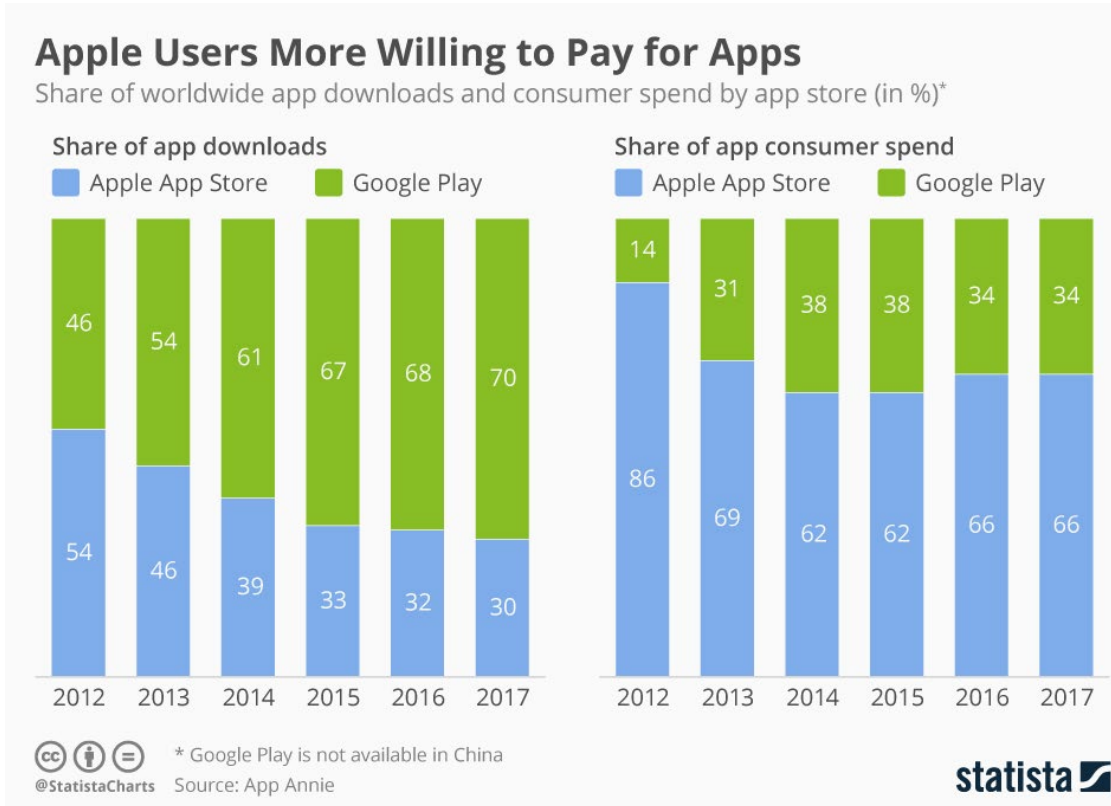
# 1 Einführung

## 1.2 Korrelation oder Kausalität?

### Unterstellter kausaler Zusammenhang:

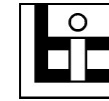


- Ist der durchschnittliche iPhone User wohlhabender?
- Hat der durchschnittliche iPhone User Zugriff auf bessere Apps?
- Sind iOS Apps durchschnittlich teurer?
- Gibt es insgesamt ein größeres Angebot an iOS Apps?



# 1 Einführung

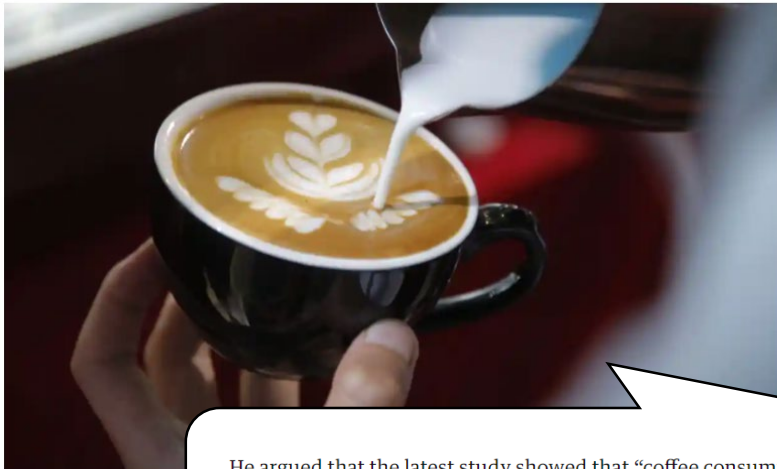
## 1.2 Korrelation oder Kausalität?



Forschungsteam  
Berens

### Three coffees a day linked to a range of health benefits

Research based on 200 previous studies worldwide says frequent drinkers less likely to get diabetes, heart disease, dementia and some cancers



The findings support  
Photograph: Wu Hong

He argued that the latest study showed that “coffee consumption seems generally safe”, but added: “Coffee is often consumed with products rich in refined sugars and unhealthy fats, and these may independently contribute to adverse health outcomes ...

Quelle: [The Guardian](#)

### Unrecht im Rechtsstaat

## Je teurer der Anwalt, umso geringer die Strafe



Unter Einsatz massiver Gewalt wurde Sven von der Polizei in Köln festgenommen. Zu Unrecht. Doch er war den Beamten hilflos ausgeliefert. © picture alliance / Fotostand

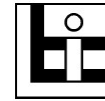
41:45 Minuten

Quelle: [Deutschlandfunk](#)



# 1 Einführung

## 1.2 Korrelation oder Kausalität?



Forschungsteam  
Berens



Quelle: [Causal Inference: The Mixtape](#)



### Einige Daumenregeln:

- „*Wer sucht, der findet!*“: Wer eine hinreichend große Anzahl an Variablen gegeneinander korrelieren lässt, wird mit hoher Wahrscheinlichkeit eine signifikante Korrelation finden (siehe  $p$ -Wert in Statistik 2).
- Häufig erscheinen sorgfältig ausgewählte Scheinkorrelationen auf den ersten Blick kausal, erst ein genauerer Blick enthüllt Ungereimtheiten in der Argumentation oder statistischen Auswertung.
- Gibt es alternative Erklärungsansätze für die beobachtete Korrelation? Können alternative Erklärungsansätze ausgeschlossen werden?
- Manchmal gibt es gar Fälle, da erwarten wir eine Korrelation, können aber keine beobachten...

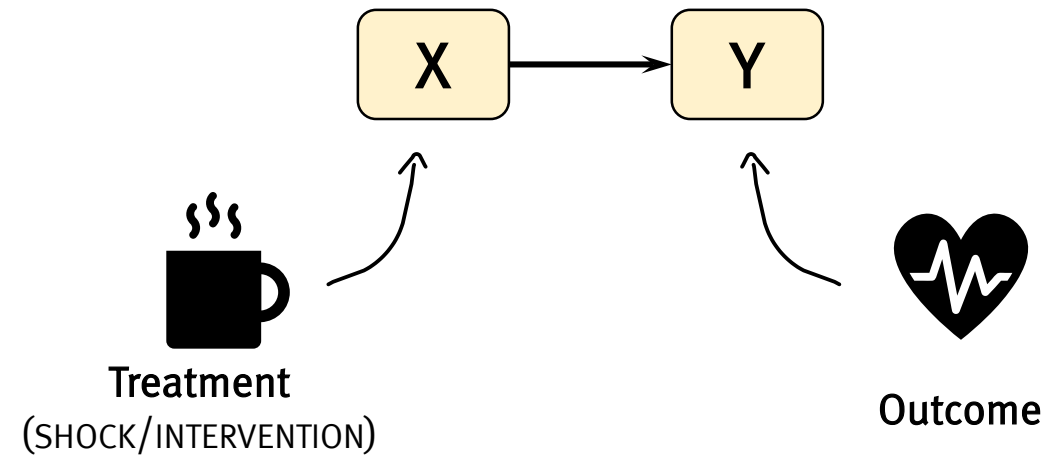
- 
- » Nur wenn alle alternativen Erklärungsansätze ausgeschlossen werden können, kann davon ausgegangen werden, dass die beobachtete Korrelation auch kausal ist!
  - » Besser noch ist die Durchführung eines **Experiments** (RANDOMIZED CONTROL TRIAL / A/B TEST), in dem sichergestellt werden kann, dass es keine alternativen Erklärungsansätze geben kann!

1	Einführung
2	<b>Das Kontrollierte Experiment</b>
3	Lösungsansatz A: Das Natürliche Experiment
4	Lösungsansatz B: Kontrollvariablen
5	Fazit

# 2 Das Kontrollierte Experiment

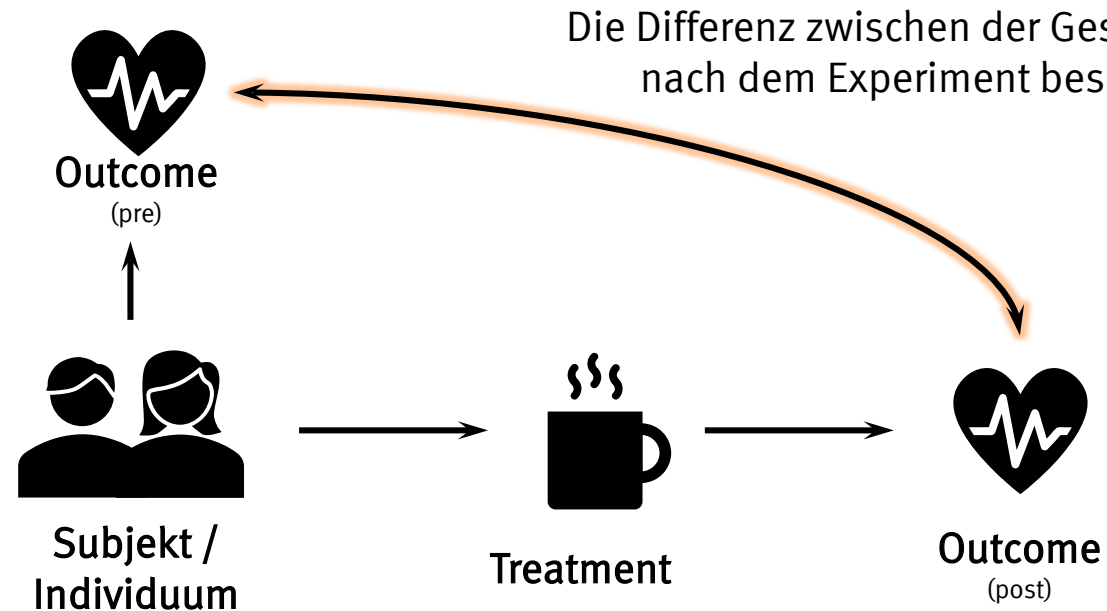
## 2.1 Terminologie

**Ausgangspunkt:**  
Was ist der Effekt von X auf Y?



# 2 Das Kontrollierte Experiment

## 2.2 Messung eines Zusammenhangs

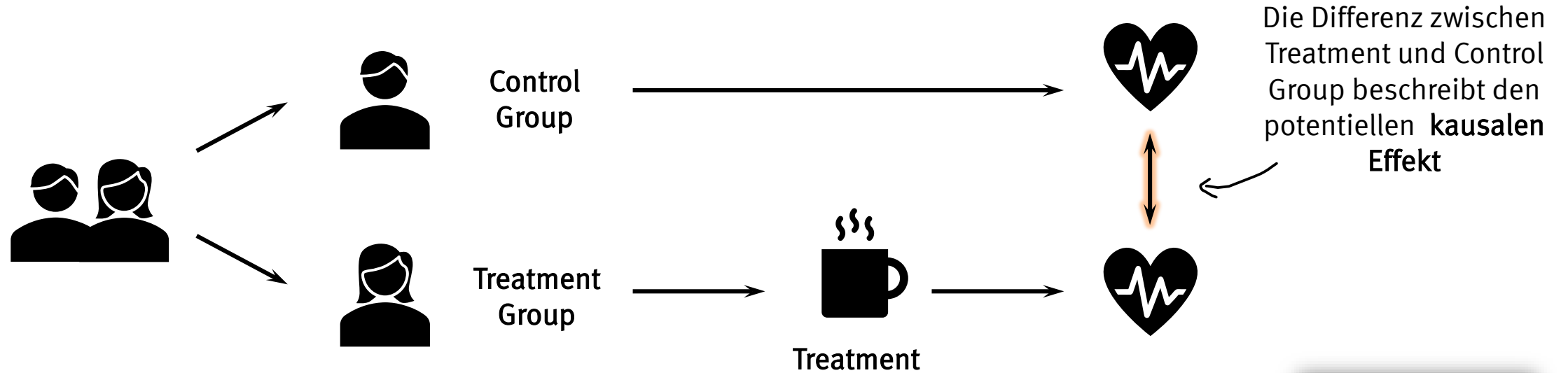


**Problem:** Wir wissen nicht, ob die Gesundheit der Probanden ohne Kaffeekonsum (COUNTERFACTUAL) ähnlich gut wäre!



## 2 Das Kontrollierte Experiment

### 2.3 Messung eines kausalen Zusammenhangs

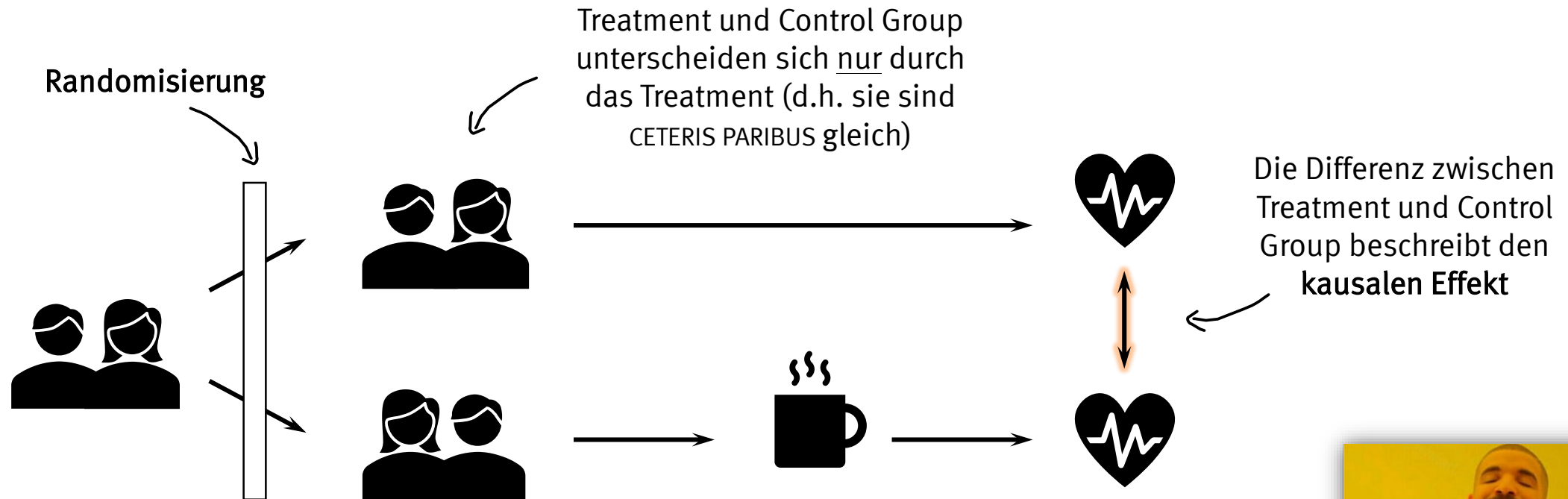


**Problem:** Treatment und Control Group sind systematisch unterschiedlich voneinander!



## 2 Das Kontrollierte Experiment

### 2.3 Messung eines kausalen Zusammenhangs



- » Sind Treatment und Control Group hinreichend groß, dann sorgt das Gesetz der großen Zahlen (LLN) dafür, dass beide Gruppen im Durchschnitt gleich sind.





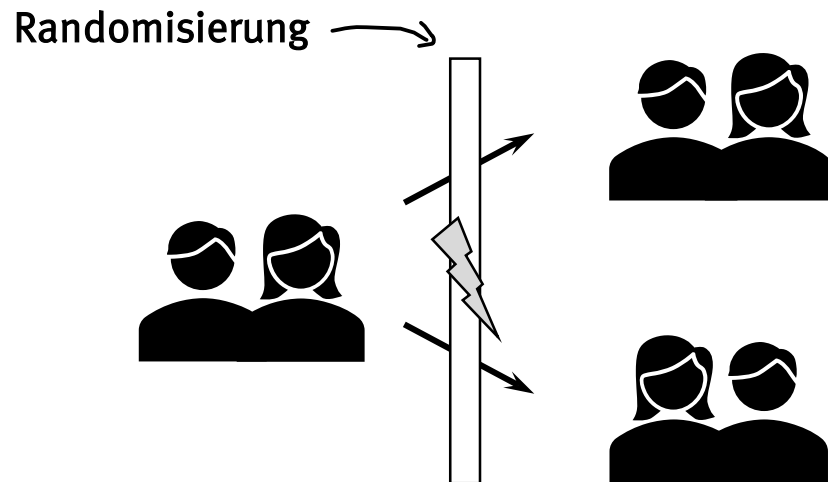
## 2 Das Kontrollierte Experiment

### 2.3 Messung eines kausalen Zusammenhangs

**Problem:** Häufig ist ein kontrolliertes Experiment aus ethischer, logistischer oder finanzieller Sicht unmöglich.

**Alternative:** Beobachtungsstudie (OBSERVATIONAL STUDY)

- Beobachtung der Subjekte ohne aktiv zu intervenieren, häufig retrospektiv
- Subjekte nehmen das Treatment eigenständig und unterliegen dadurch systematischen Unterschieden (SELF-SELECTION / SELECTION BIAS).



- » **Lösungsansatz A:** Suche nach einer Randomisierung, die „natürlich“ auftritt (AS-IF/QUASI RANDOM TREATMENT).
- » **Lösungsansatz B:** Nachträgliche Korrektur für systematische Unterschiede (unter Verwendung sog. Kontrollvariablen).

1	Einführung
2	Das Kontrollierte Experiment
3	Lösungsansatz A: Das Natürliche Experiment
4	Lösungsansatz B: Kontrollvariablen
5	Fazit

# 3 Lösungsansatz A: Das Natürliche Experiment

## 3.1 Ausbruch der [3. Cholera-Pandemie](#)



A COURT FOR KING CHOLERA.

### Miasma-Theorie:

- Üble Gerüche/Gestank als Auslöser der Krankheit.
- Durchaus plausibel, da Gerüche und Krankheiten häufig miteinander korrelieren.
- Vermutung, dass die Krankheit durch das Vertreiben des Gestanks ausgelöscht werden kann (z.B. Luftzirkulation, Blumen, Schießpulver).

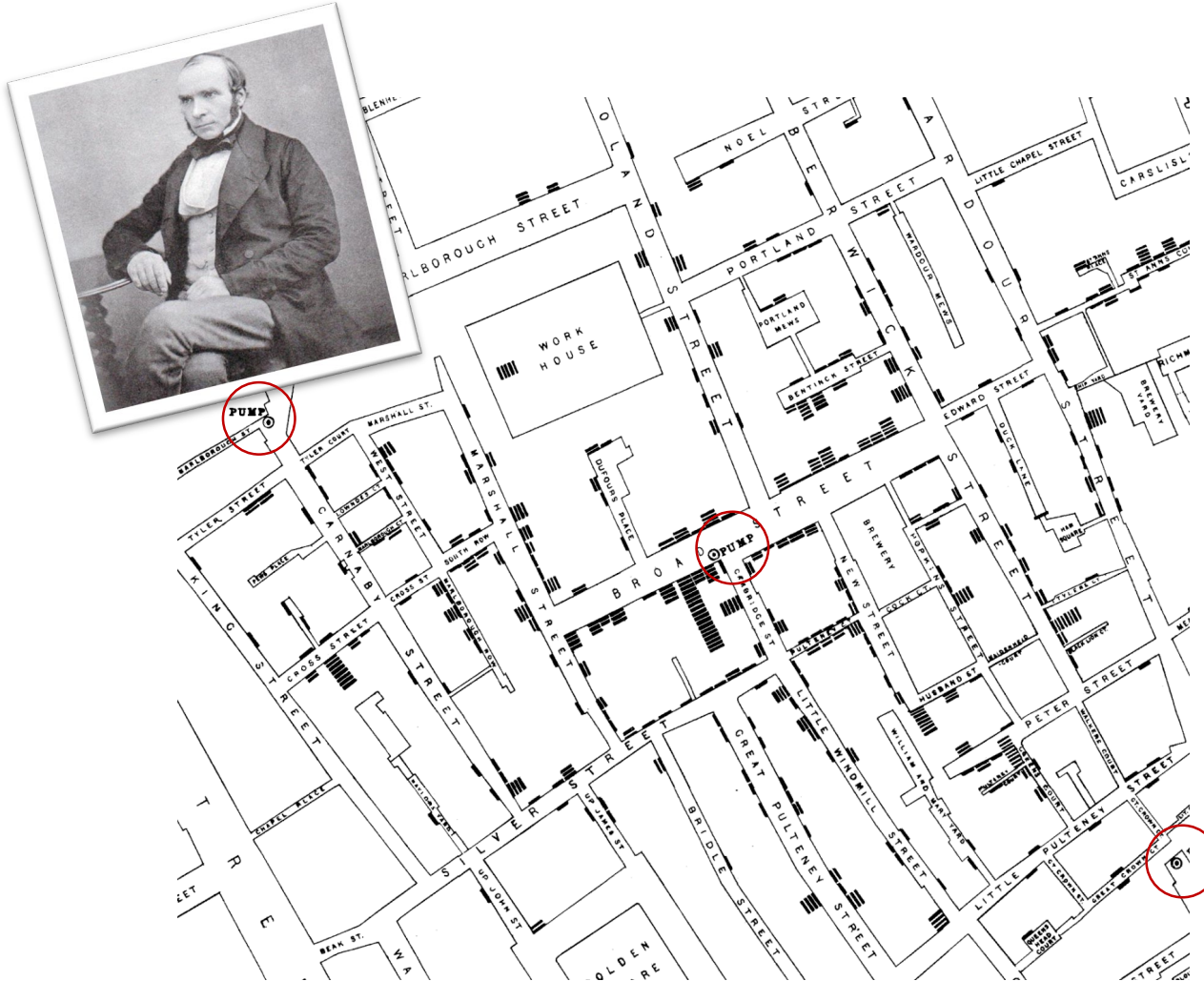
» Korrelation oder Kausalität?

Quelle Beispiel: [Data 8, UC Berkeley](#)

Quelle Bild: Punch (1852)

# 3 Lösungsansatz A: Das Natürliche Experiment

## 3.1 Ausbruch der [3. Cholera-Pandemie](#)



### John Snow-Theorie:

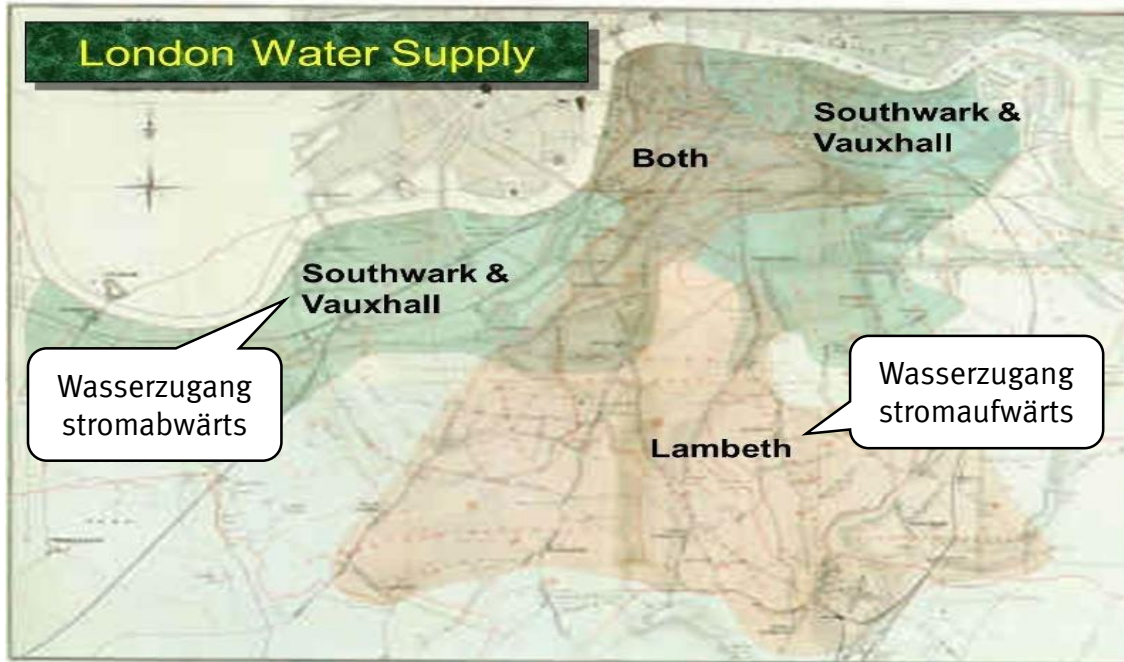
- Verunreinigtes Wasser als Auslöser der Krankheit.
- Selbst Bewohner:innen aus entfernteren Gebieten (mit eigenen Pumpen) kamen regelmäßig zur Broad Street Pump zur Wasserentnahme.
- Dicht besiedelte Blöcke, etwa das Work House oder die Brewery, hatten hingegen eigene Pumpen.

Quelle Beispiel: [Data 8, UC Berkeley](#)



# 3 Lösungsansatz A: Das Natürliche Experiment

## 3.1 Ausbruch der 3. Cholera-Pandemie



**Hypothese:** Verunreinigtes Wasser von S&V als Auslöser der Pandemie.

**Subjects:** Einwohner:innen von London

**Randomisierung:** In Gebieten, wo beide Wasserversorger tätig sind, wird angenommen, dass die Versorgung zufällig erfolgt (NATURAL EXPERIMENT). Es bestehen also keine systematischen Unterschiede zwischen den Einwohner:innen.

**Treatment Group:** Einwohner:innen, die Trinkwasser von S&V beziehen

**Control Group:** Einwohner:innen, die Trinkwasser von Lambeth beziehen

Die Differenz zwischen Treatment und Control Group beschreibt den **kausalen Effekt**

Supply Area	Number of houses	Cholera deaths	Deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

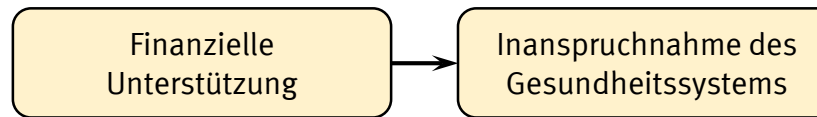
Quelle Beispiel: [Data 8, UC Berkeley](#)

# 3 Lösungsansatz A: Das Natürliche Experiment

## 3.2 Andere Beispiele

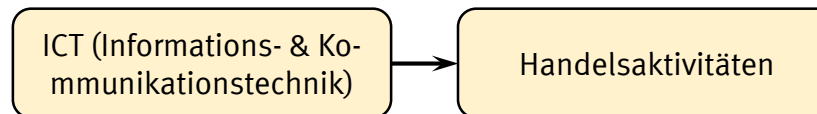
Häufig ist die Suche nach einer Randomisierung, die „natürlich“ auftritt (AS-IF RANDOM TREATMENT), ein Alleinstellungsmerkmal von guter und glaubwürdiger Forschung! Nachfolgend ein paar ausgewählte Beispiele:

### 1) Fragestellung:



- **Problem:** Menschen, die finanzielle Unterstützung erhalten, sind systematisch unterschiedlich.
- **Randomisierung:** Fixe Einkommensgrenzen determinieren, ob Bürger Anspruch auf finanzielle Hilfe haben oder nicht. Menschen, die kurz unter bzw. oberhalb der Grenze liegen, unterscheiden sich nicht systematisch.

### 2) Fragestellung:



- **Problem:** Firmen, die besseren Zugang zu ICTs haben, sind systematisch unterschiedlich.
- **Randomisierung:** Sukzessiver Rollout der Breitbandanbindung in Norwegen. Firmen, die besseren Internetzugang haben, unterscheiden sich potenziell nur noch durch ihren Firmensitz.

Quelle Beispiel: [Plausibly Exogenous Galore](#)

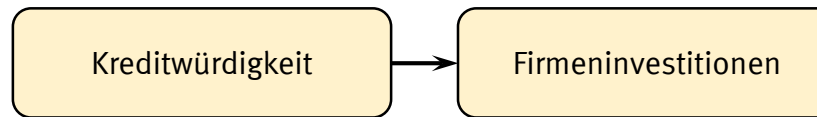


# 3 Lösungsansatz A: Das Natürliche Experiment

## 3.2 Andere Beispiele

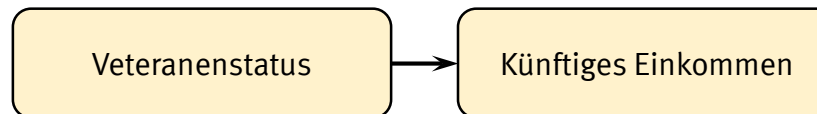
Häufig ist die Suche nach einer Randomisierung, die „natürlich“ auftritt (AS-IF RANDOM TREATMENT), ein Alleinstellungsmerkmal von guter und glaubwürdiger Forschung! Nachfolgend ein paar ausgewählte Beispiele:

### 3) Fragestellung:



- **Problem:** Firmen, mit besserer Kreditwürdigkeit, treffen systematisch andere Investitionsentscheidungen.
- **Randomisierung:** Methodik-Änderungen der Ratingagenturen. Unternehmen, die infolgedessen ein Upgrade oder Downgrade erhalten, werden (quasi) zufällig ausgewählt.

### 4) Fragestellung:



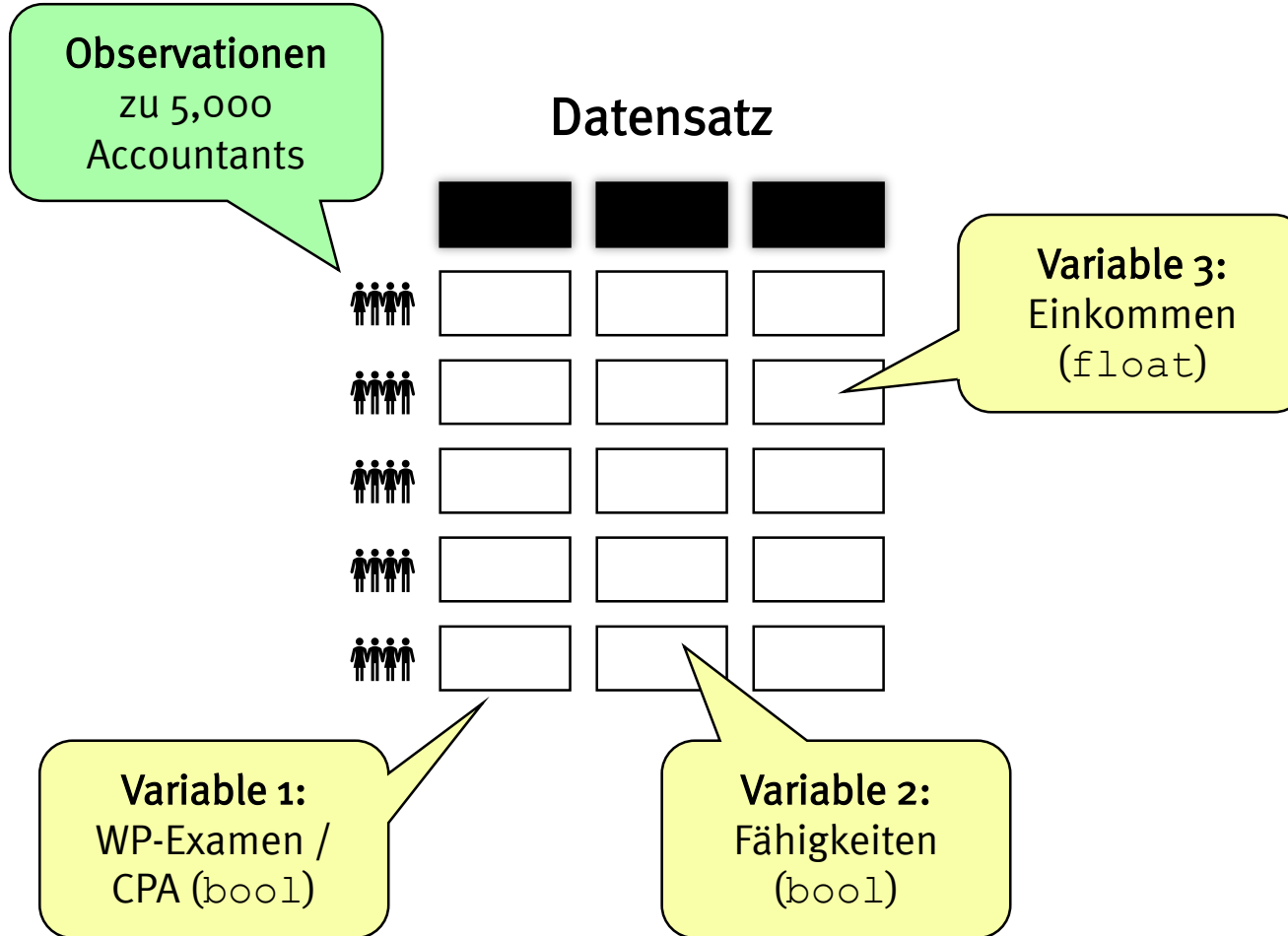
- **Problem:** Menschen, die zum Militär gehen, sind systematisch unterschiedlich.
- **Randomisierung:** Die „Vietnamkrieg Draft Lottery“ in den USA führt dazu, dass Rekruten und nicht-Rekruten nur zufällig unterschiedlich sind.

Quelle Beispiel: [Plausibly Exogenous Galore](#)

1	Einführung
2	Das Kontrollierte Experiment
3	Lösungsansatz A: Das Natürliche Experiment
4	Lösungsansatz B: Kontrollvariablen
5	Fazit

# 4 Lösungsansatz B: Kontrollvariablen

## 4.1 Datensatzbeschreibung



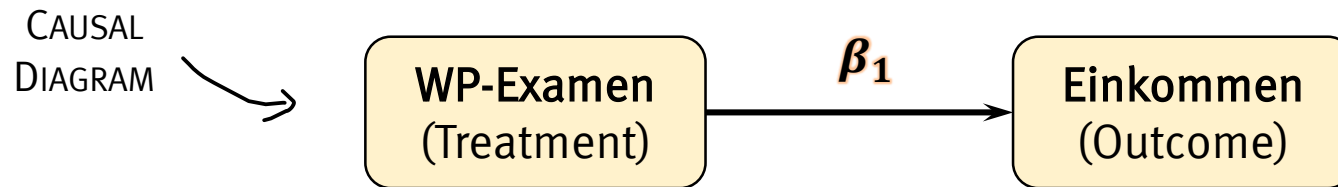
\* es handelt sich hier bei um einen  
simulierten (SYNTHETISCHEN) Datensatz

Quelle: [Whited et al. \(2022\)](#)

# 4 Lösungsansatz B: Kontrollvariablen

## 4.2 Der Effekt des Wirtschaftsprüfer Examens auf das Einkommen

**Frage:** Was ist der Effekt eines WP-Examens auf das Einkommen?



1) Vergleich des durchschnittlichen Einkommens von WPs mit nicht WPs

oder

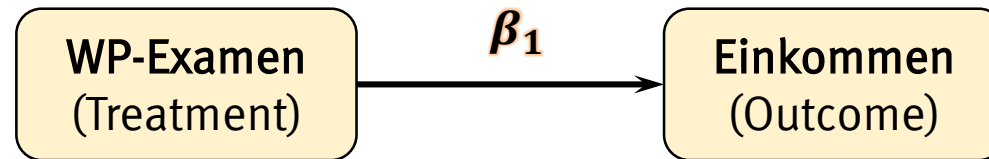
2) Berechnung eines linearen Regressionsmodells:

$$Einkommen = \beta_0 + \beta_1 \times WP + \varepsilon$$

# 4 Lösungsansatz B: Kontrollvariablen

## 4.2 Der Effekt des Wirtschaftsprüfer Examens auf das Einkommen

**Frage:** Was ist der Effekt eines WP-Examens auf das Einkommen?



**Problem:** Accountants mit einem abgeschlossenen WP-Examen haben systematisch andere Fähigkeiten als solche, die keines abgelegt haben (SELF-SELECTION).

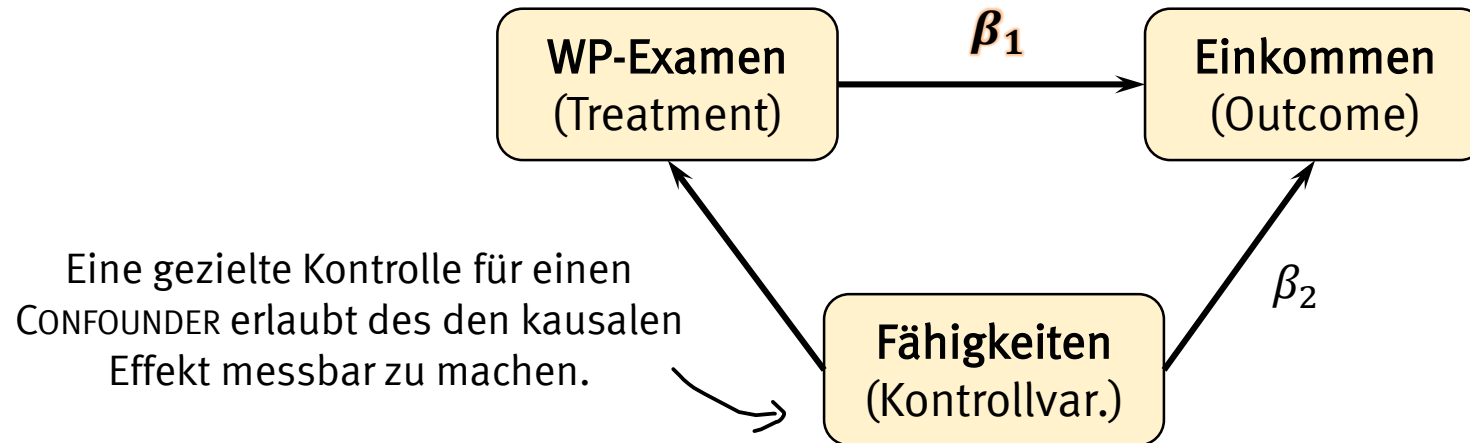
**Lösung:** Verwendung einer **Kontrollvariable** für individuelle Fähigkeiten (COUNFOUNDER)

↳ Kontrollvariablen kontrollieren für potenziell systematische Unterschiede;  
sie halten eine auftretende Veränderung (z.B. Fähigkeiten) konstant  
(d.h. ermöglichen einen Vergleich zwischen WPs und Nicht-WPs mit ansonsten gleichen Fähigkeiten)

# 4 Lösungsansatz B: Kontrollvariablen

## 4.2 Der Effekt des Wirtschaftsprüfer Examens auf das Einkommen

**Frage:** Was ist der kausale Effekt eines WP-Examens auf das Einkommen?



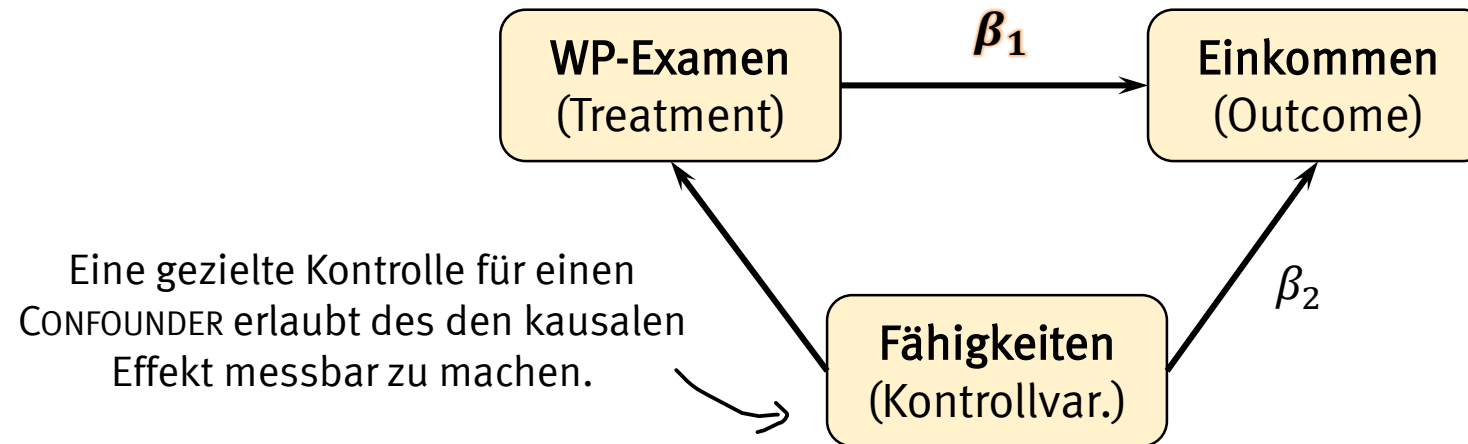
- 1) Vergleich des durchschnittlichen Einkommens von WPs mit nicht WPs, bei gleichen Fähigkeiten (c.p.). Der kausale Effekt ergibt sich dann als eine gewichtete Summe der Mittelwert-Differenzen.



# 4 Lösungsansatz B: Kontrollvariablen

## 4.2 Der Effekt des Wirtschaftsprüfer Examens auf das Einkommen

**Frage:** Was ist der kausale Effekt eines WP-Examens auf das Einkommen?



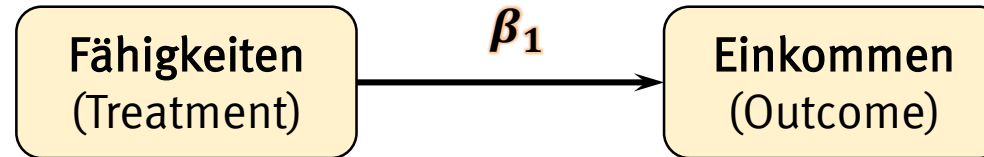
2) Berechnung eines **multiplen** linearen Regressionsmodells mit Kontrollvariable:

$$\text{Einkommen} = \beta_0 + \beta_1 \times \text{WP} + \beta_2 \times \text{Fähigkeiten} + \varepsilon$$

# 4 Lösungsansatz B: Kontrollvariablen

## 4.3 Der Effekt der individuellen Fähigkeiten auf das Einkommen

**Frage:** Was ist der Effekt der individuellen Fähigkeiten auf das Einkommen?



1) Vergleich des durchschnittlichen Einkommens von Accountants mit hohen und geringen Fähigkeiten  
oder

2) Berechnung eines linearen Regressionsmodells:  
$$\text{Einkommen} = \beta_0 + \beta_1 \times \text{Fähigkeiten} + \varepsilon$$



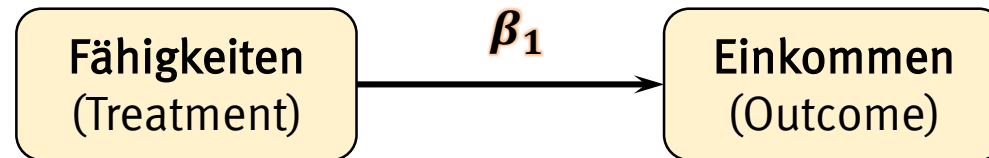
Live Session:  
5 Minuten



# 4 Lösungsansatz B: Kontrollvariablen

## 4.3 Der Effekt der individuellen Fähigkeiten auf das Einkommen

**Frage:** Was ist der direkte Effekt der individuellen Fähigkeiten auf das Einkommen?



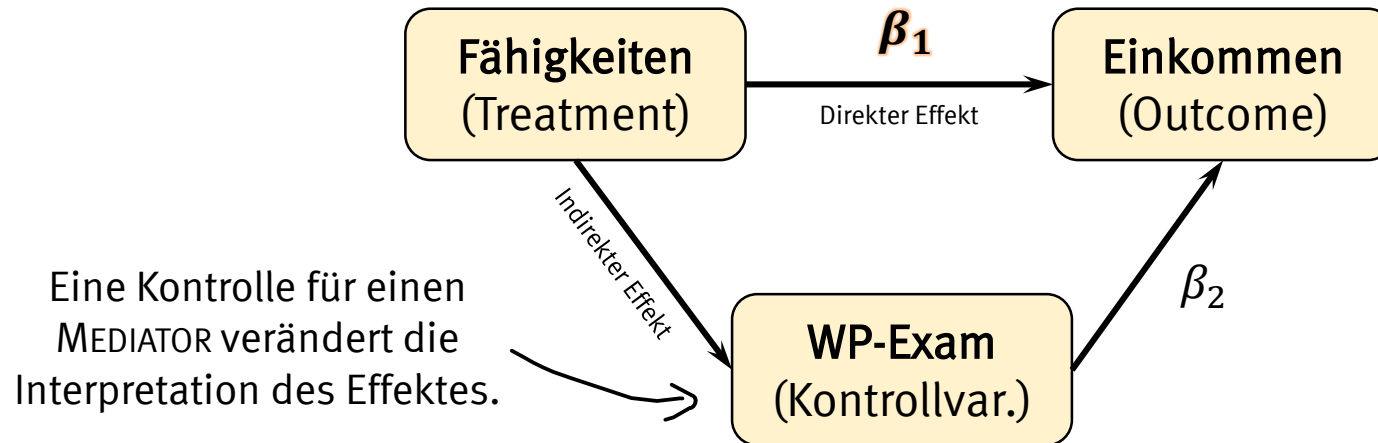
**Problem:** Die individuellen Fähigkeiten haben einen indirekten Einfluss auf das Einkommen, nämlich über den Umstand, ob ein WP-Exam abgelegt wurde oder nicht.

**Lösung:** Verwendung einer **Kontrollvariable** für WP-Exam  
(d.h. ermöglicht einen Vergleich zwischen Accountants mit unterschiedlichen Fähigkeiten bei ansonsten gleicher Ausbildung)

# 4 Lösungsansatz B: Kontrollvariablen

## 4.3 Der Effekt der individuellen Fähigkeiten auf das Einkommen

**Frage:** Was ist der direkte Effekt der individuellen Fähigkeiten auf das Einkommen?

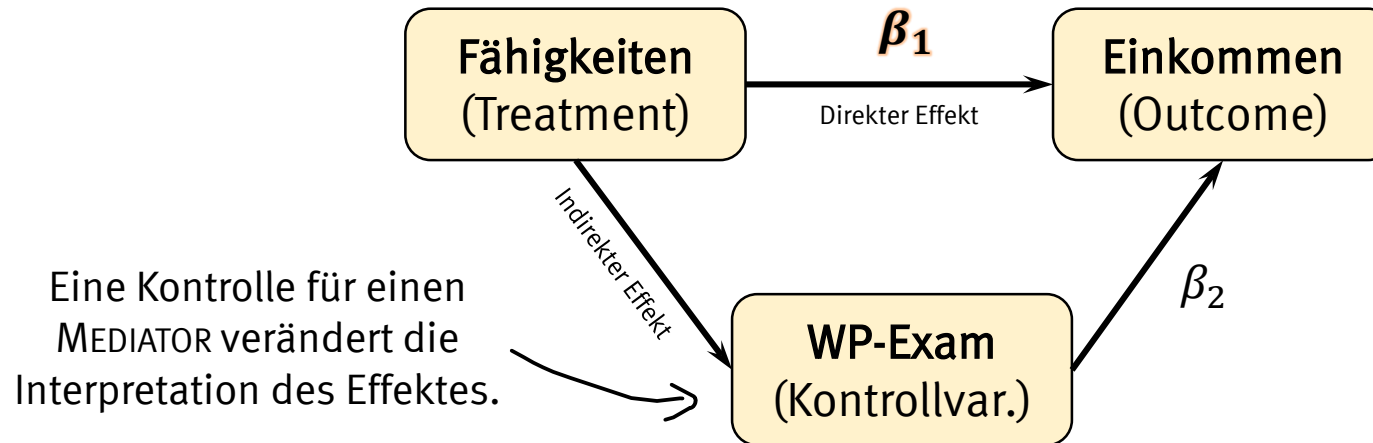


- 1) Vergleich des durchschnittlichen Einkommens von Accountants mit hohen und geringen Fähigkeiten, bei gleichem Abschluss. Der direkte Effekt ergibt sich als eine gewichtete Summe der Mittelwert-Differenzen.

# 4 Lösungsansatz B: Kontrollvariablen

## 4.3 Der Effekt der individuellen Fähigkeiten auf das Einkommen

**Frage:** Was ist der direkte Effekt der individuellen Fähigkeiten auf das Einkommen?



2) Berechnung eines **multiplen** linearen Regressionsmodells mit Kontrollvariable:

$$\text{Einkommen} = \beta_0 + \beta_1 \times \text{Fähigkeiten} + \beta_2 \times \text{WP} + \varepsilon$$



Live Session:  
5 Minuten



1	Einführung
2	Das Kontrollierte Experiment
3	Lösungsansatz A: Das Natürliche Experiment
4	Lösungsansatz B: Kontrollvariablen
5	Fazit



- » Kausale Zusammenhänge herauszufinden ist schwer! Das ist insbesondere der Fall, wenn kein kontrolliertes Experiment durchgeführt werden kann.
- » Häufig lässt sich nur mit einem genaueren Blick auf die Datengrundlage und die statistische Analyse erkennen, ob kausale Zusammenhänge oder (Schein)Korrelationen identifiziert wurden.
- » Glücklicherweise existieren Phänomene in der realen Welt, die dazu führen, dass es zu einer quasi-zufälligen Randomisierung von Subjekten kommt (QUASI-RANDOM EXPERIMENTS). Alternativ kann für systematische Unterschiede zwischen Untersuchungssubjekten mittel Kontrollvariablen kontrolliert werden.

