Korrelation & Kausalität

Simon Schölzel, M.Sc.

(updated: 01.06.2022)

Agenda

Forschungsteam Berens

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

Agenda

Forschungsteam Berens

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:

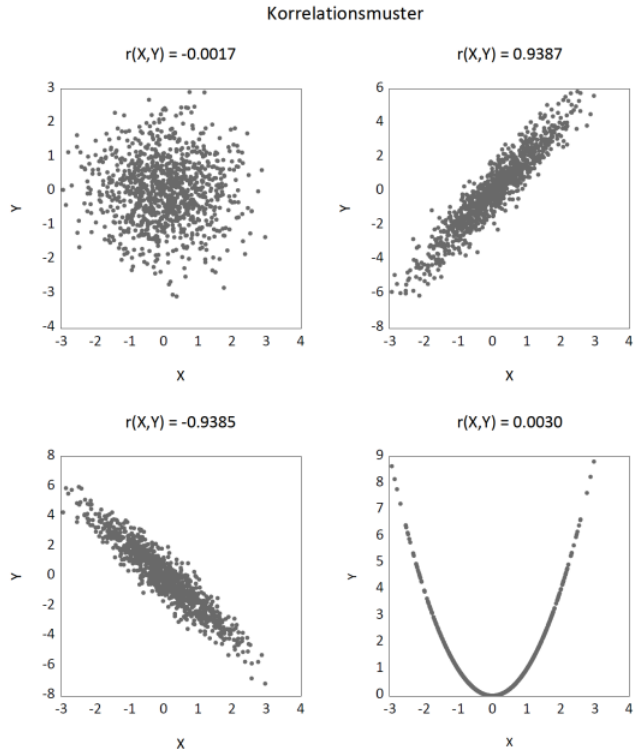
- Maß für den linearen Zusammenhang

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

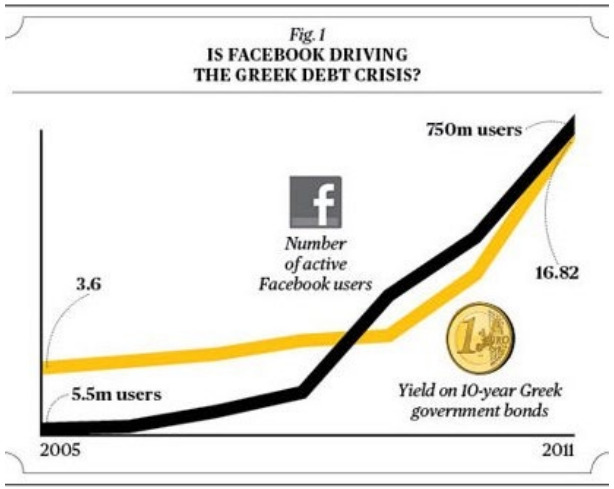
Rangkorrelationskoeffizient nach Spearman:

- Maß für monotone Zusammenhänge

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



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



Quelle: [Jeremy Bertomeu]

Quelle: [New England Journal of Medicine]

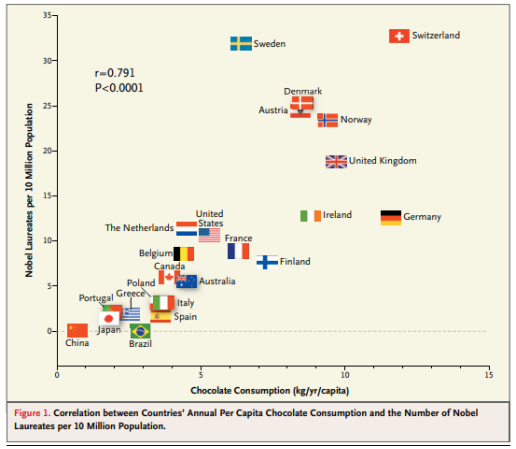
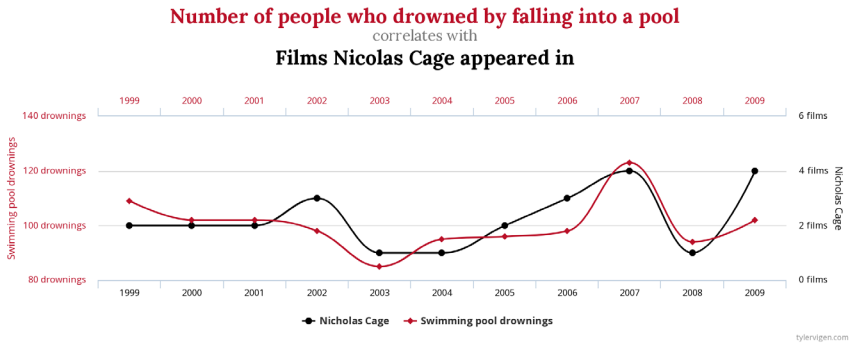
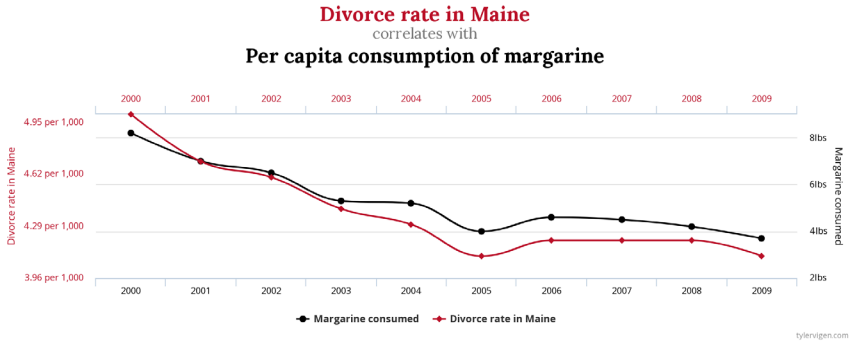
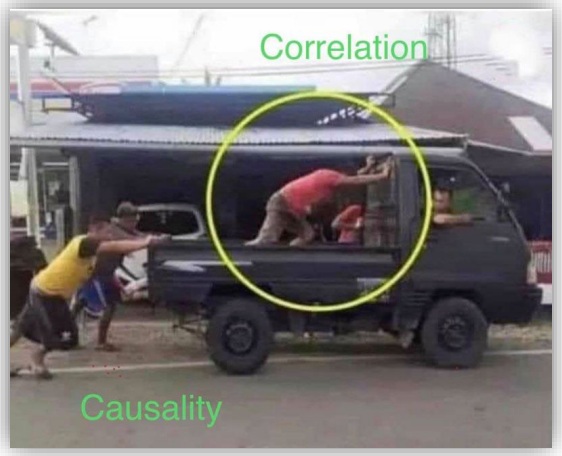


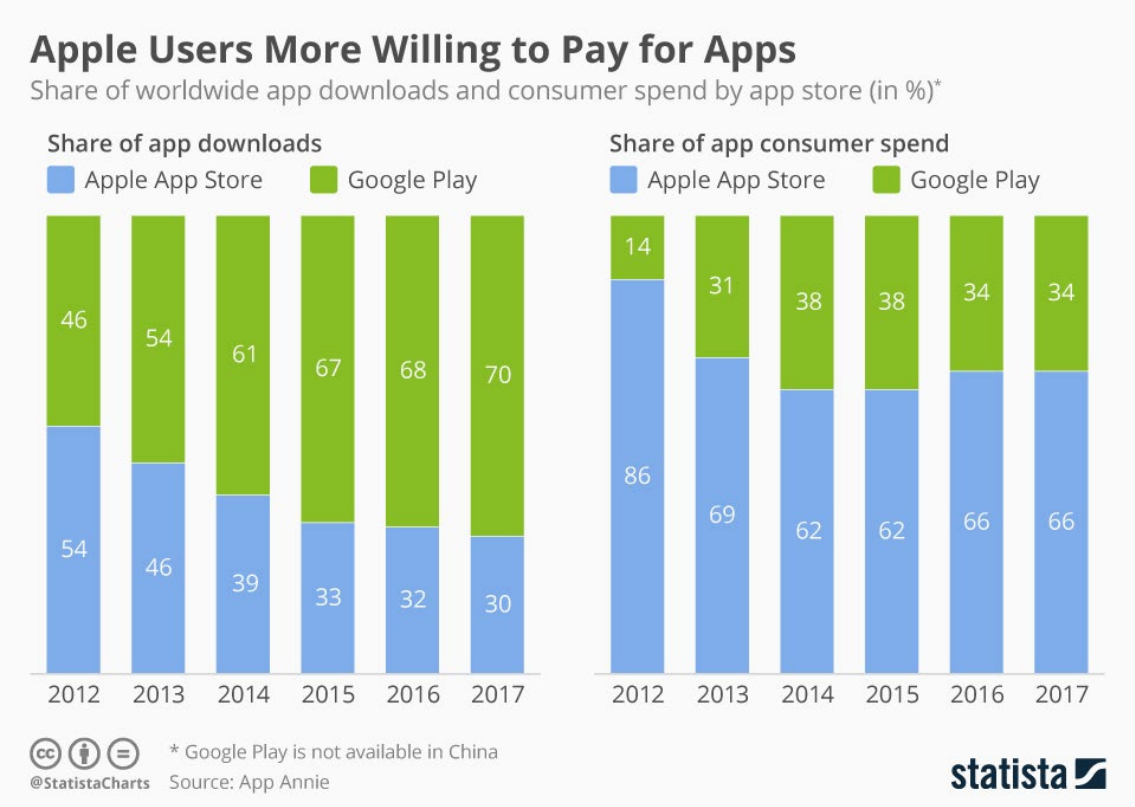
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.





Quelle: [Spurious-Correlations]





Unterstellter kausaler Zusammenhang:

Besitz eines iPhones → Zahlungsbereitschaft für Apps

- 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?

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 suppor
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

Quelle: Causal Inference: The Mixtape

Einige Daumenregeln:

- *„Wer suchet, 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? (Achtung: Häufig sind diese nicht trivial zu identifizieren!)
- Manchmal gibt es gar Fälle, da erwarten wir eine Korrelation, können aber keine beobachten…

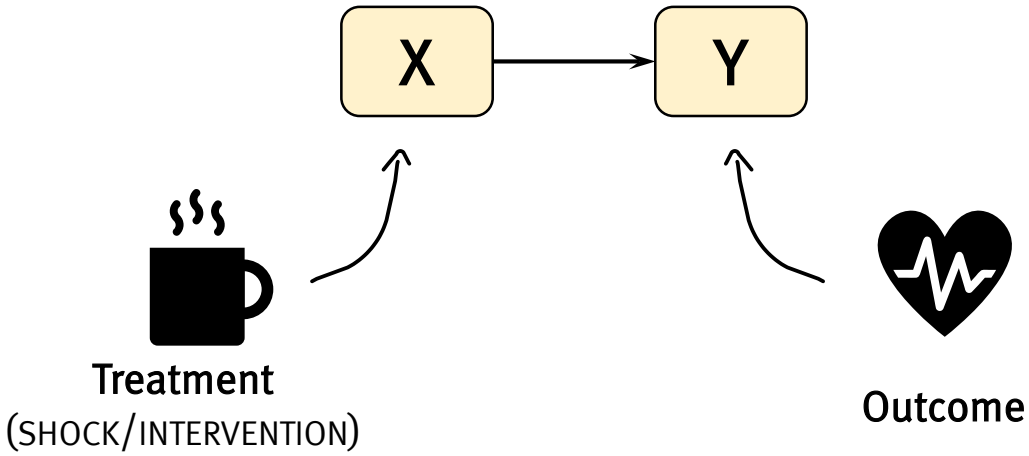
» Nur wenn alle alternativen Erklärungsansätze ausgeschlossen werden können, kann davon ausgegangen werden, dass die beobachtete Korrelation auch kausal ist!

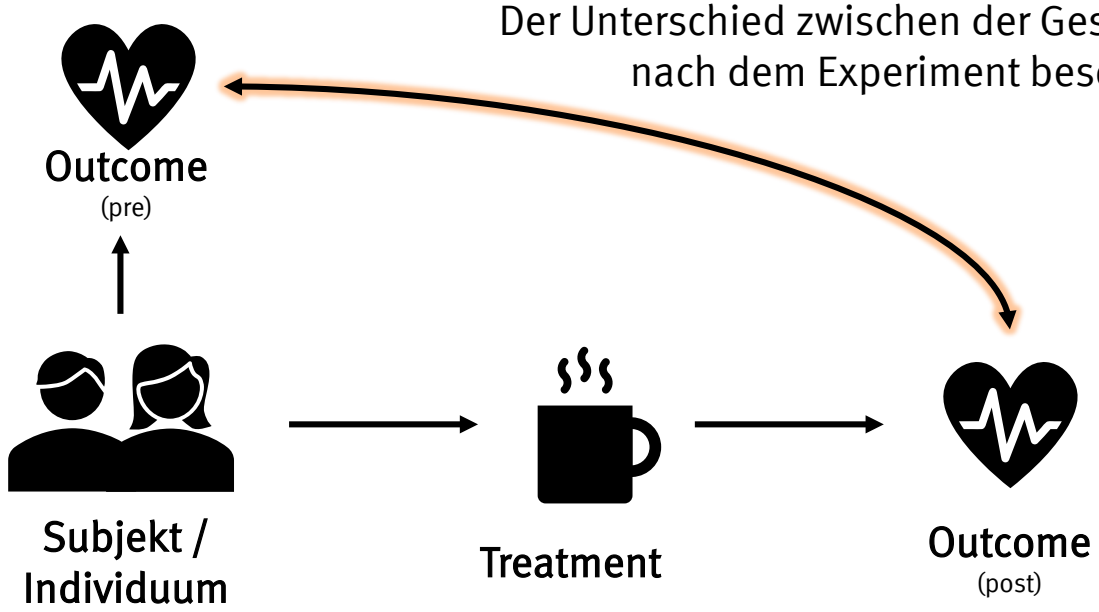
» Besser noch ist die Durchführung eines **Experiments** (RANDOMIZED CONTROL TRIAL / A/B TEST), in dem sichergestellt werden kann, dass es erst keine alternativen Erklärungsansätze gibt!

Agenda

Forschungsteam Berens

Ausgangspunkt: Was ist der Effekt von X auf Y?



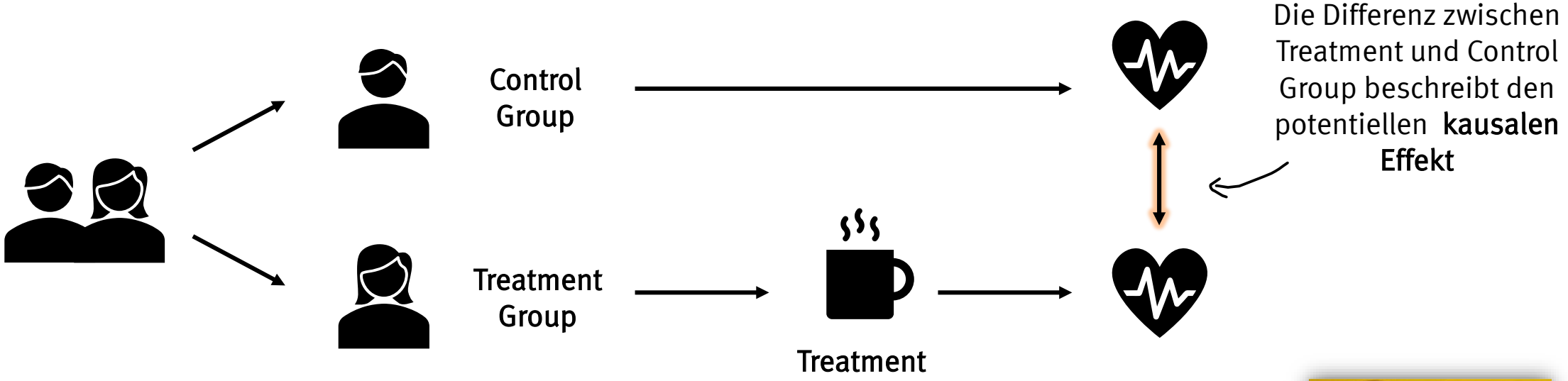


Der Unterschied zwischen der Gesundheit des Probanden vor und nach dem Experiment beschreibt den **beobachteten Effekt**

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

2 Das Kontrollierte Experiment

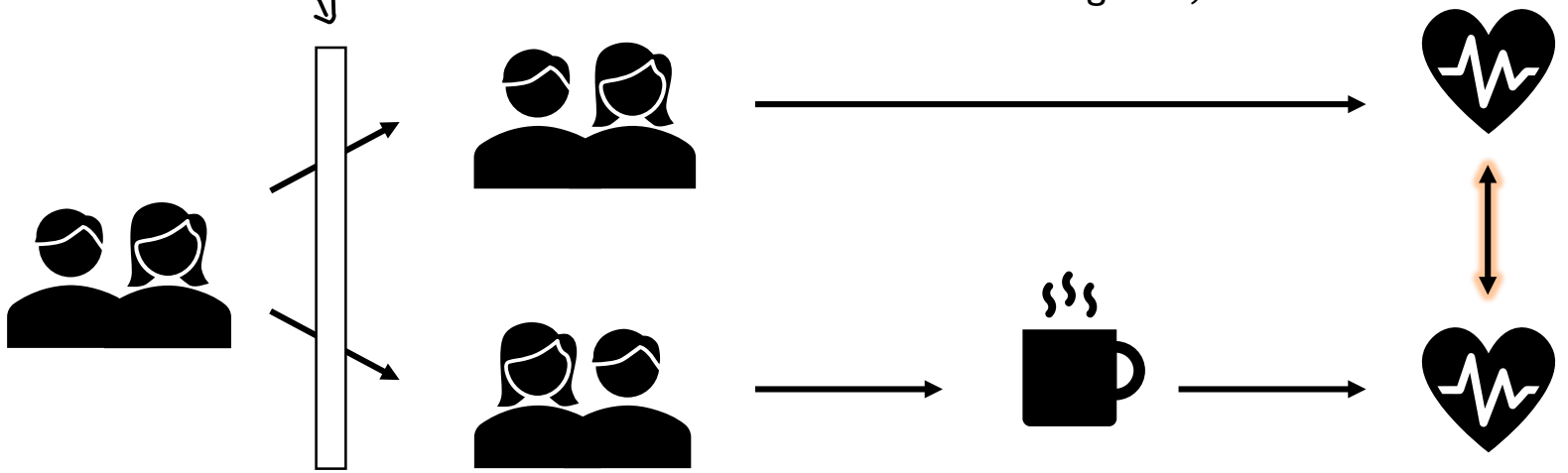
2.1 Beobachtung des kausalen Zusammenhangs



Problem: Treatment und Control Gruppe sind systematisch unterschiedlich voneinander!

2 Das Kontrollierte Experiment

2.1 Beobachtung des kausalen Zusammenhangs



Treatment und Control Gruppe unterscheiden sich nur durch das Treatment (d.h. sie sind CETERIS PARIBUS gleich)

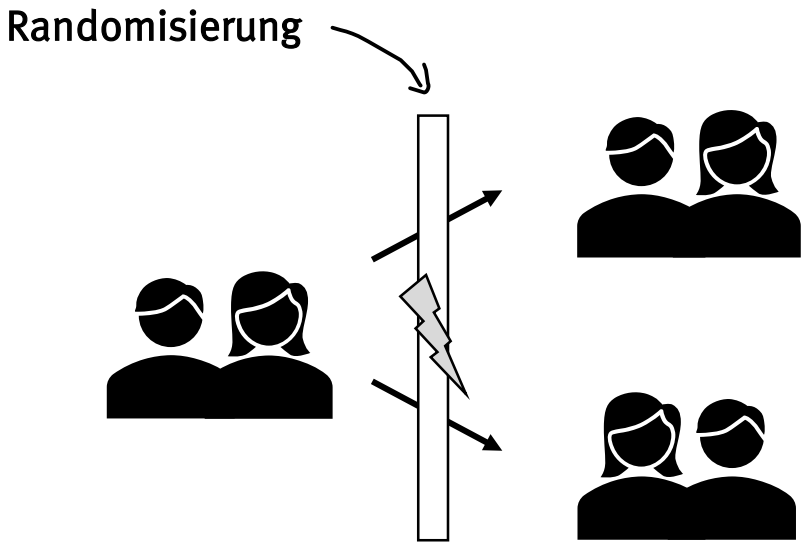
Die Differenz zwischen Treatment und Control Group beschreibt den **kausalen Effekt**

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

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 sind dadurch in der Regel systematisch unterschiedlich (SELF-SELECTION / SELECTION BIAS)



» **Lösungsansatz A:** Suche nach einer Randomisierung, die „natürlich“ auftritt (AS-IF RANDOM TREATMENT).

» **Lösungsansatz B:** Nachträgliche Korrektur für systematische Unterschiede (unter Verwendung sog. Kontrollvariablen).

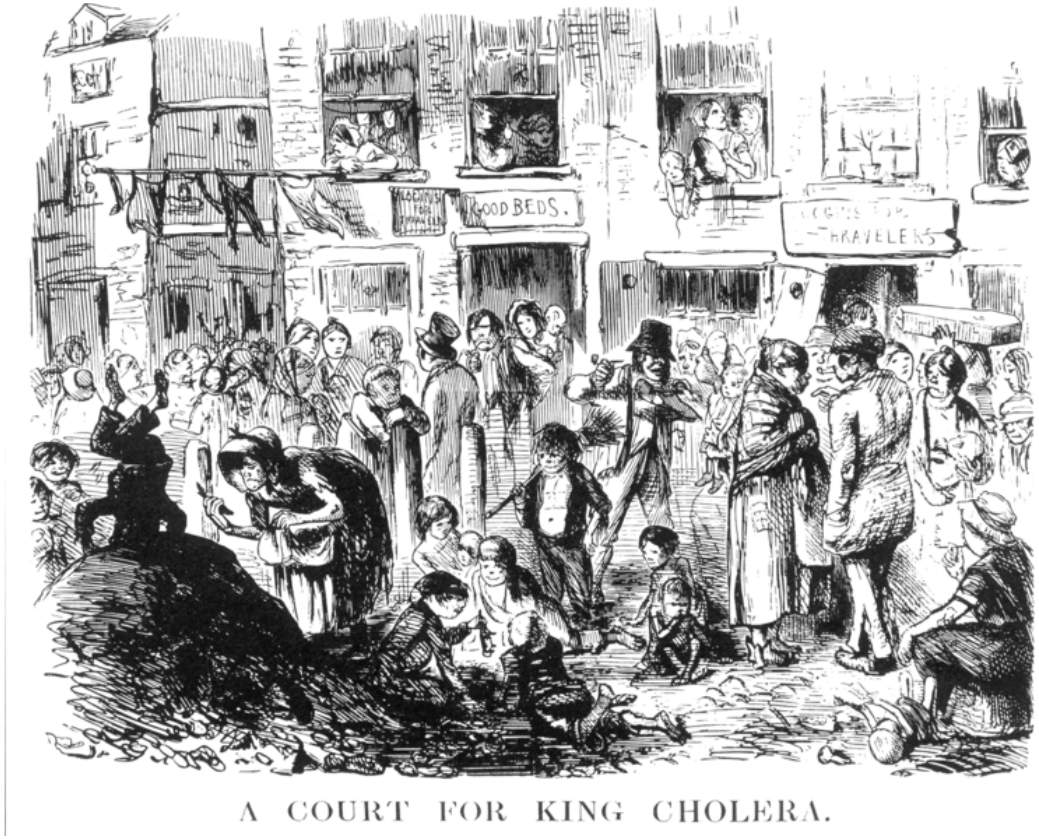
Agenda

Forschungsteam Berens

3 Lösungsansatz A: Das Natürliche Experiment

3.1 Ausbruch der 3. Cholera-Pandemie

Forschungsteam Berens



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).

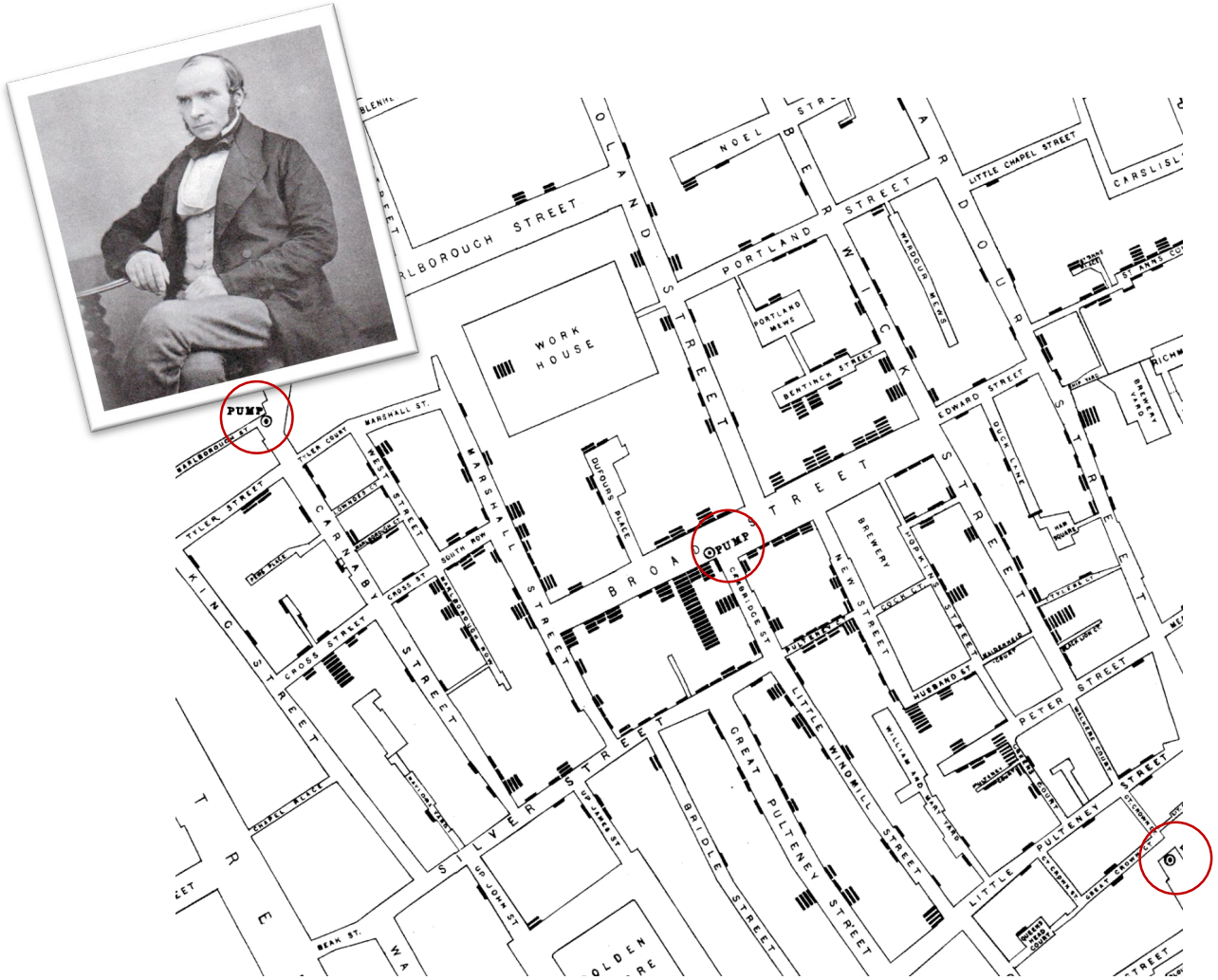
» **Korrelation oder Kausalität?**

Quelle Beispiel: Data 8, UC Berkeley
Quelle Bild: Punch (1852)

3 Lösungsansatz A: Das Natürliche Experiment

3.1 Ausbruch der 3. Cholera-Pandemie

Forschungsteam Berens



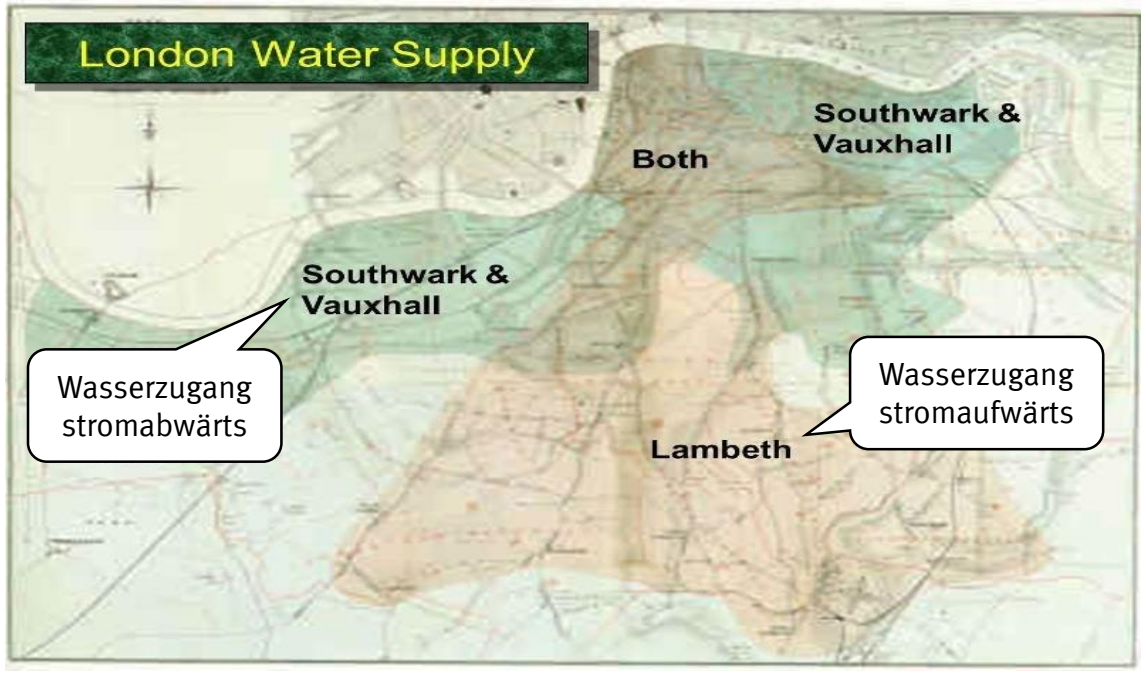
John Snow-Theorie:

- Verunreinigtes Wasser als Auslöser der Krankheit.
- Selbst BewohnerInnen 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 interne Pumpen.

Quelle Beispiel: Data 8, UC Berkeley

3 Lösungsansatz A: Das Natürliche Experiment

3.1 Ausbruch der 3. Cholera-Pandemie



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

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

Subjects: Einwohner 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 Einwohnern.

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

Control Group: Einwohner, die Trinkwasser von Lambeth beziehen

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

Quelle Beispiel: Data 8, UC Berkeley

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: Finanzielle Unterstützung → Inanspruchnahme des Gesundheitssystems

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

2) Fragestellung: ICT (Informations- & Kommunikationstechnik) → Handelsaktivitäten

- **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

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: Kreditwürdigkeit → Firmeninvestitionen

- **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: Veteranenstatus → Künftiges Einkommen

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

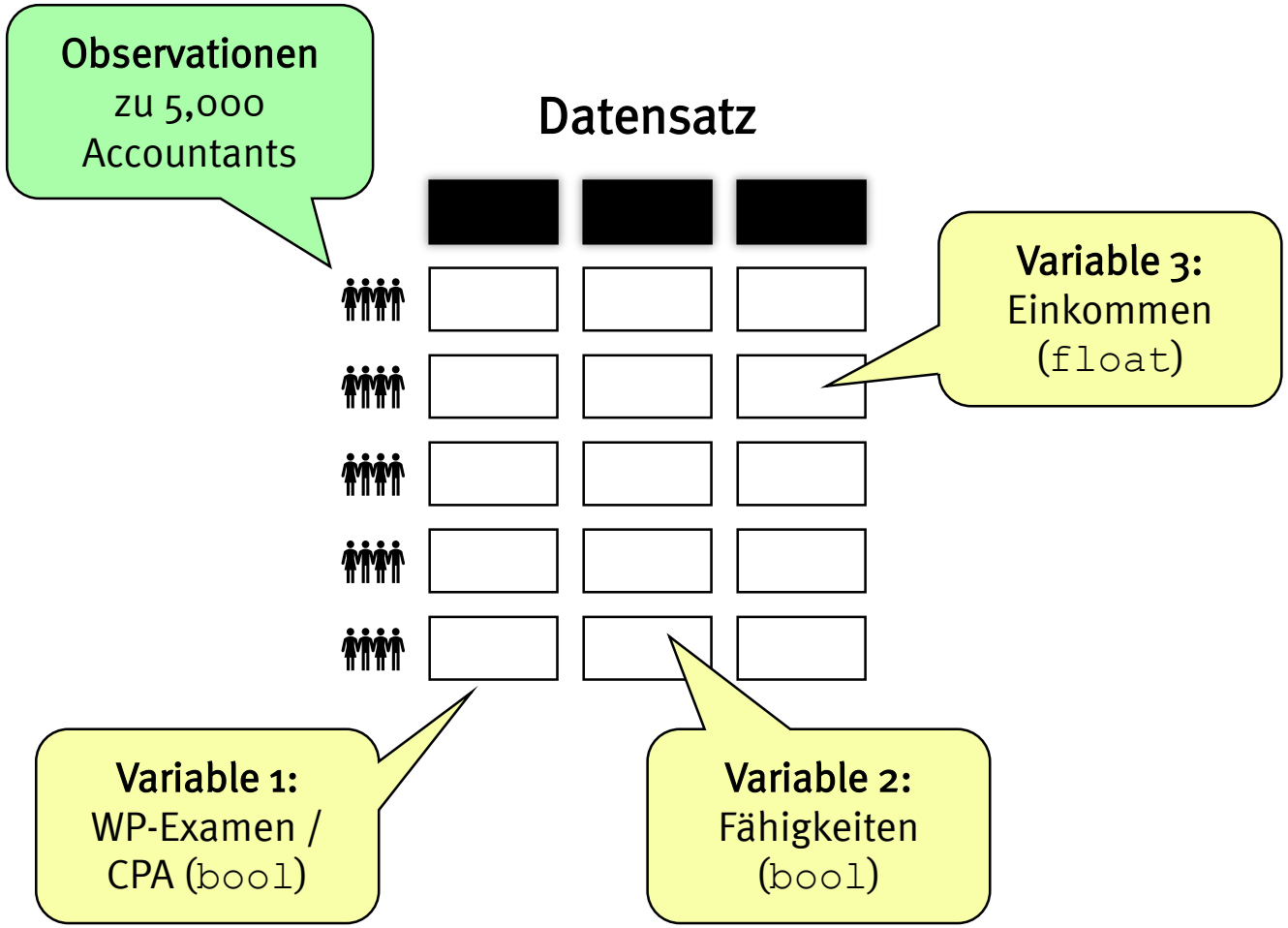
Quelle Beispiel: Plausibly Exogenous Galore

Agenda

Forschungsteam Berens

4 Lösungsansatz B: Kontrollvariablen

4.1 Datensatzbeschreibung

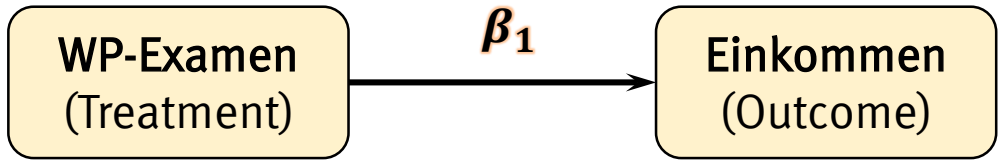


* es handelt sich hier bei um einen simulierten (SYNTHETISCHEN) Datensatz

Quelle: Whited, R. L., Swanquist, Q. T., Shipman, J. E., & Moon, J. R. (2022): Out of Control: The (Over) Use of Controls in Accounting Research. *The Accounting Review*.

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

CAUSAL DIAGRAM



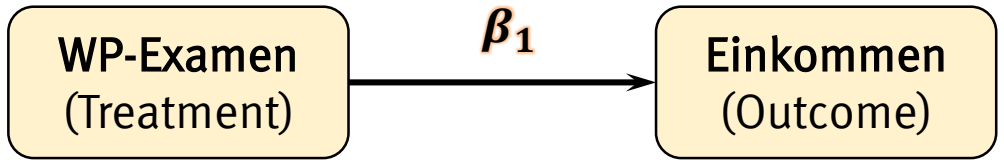
1) Vergleich des durchschnittlichen Einkommens von WPs mit nicht WPs

oder

2) Berechnung eines linearen Regressionsmodells:

$$Einkommen = \beta_0 + \boldsymbol{\beta_1} \times WP + \varepsilon$$

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

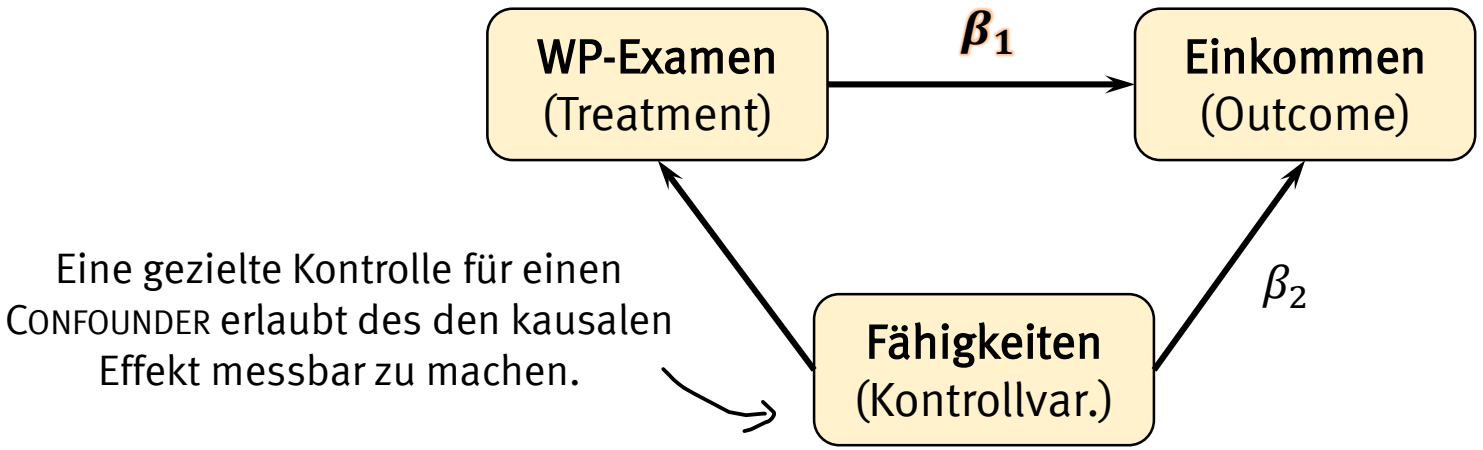


Problem: Accountants mit einem abgeschlossenen WP-Examen haben systematisch andere Fähigkeiten als solche, die keines abgelegt haben (SELF-SELECTION).

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

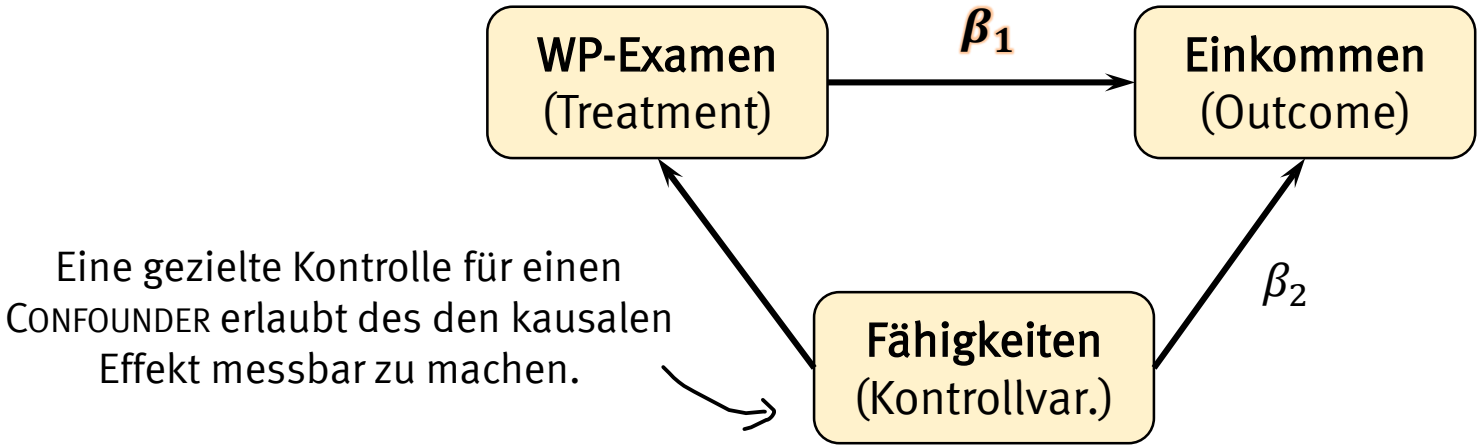
↳ Kontrollvariablen kontrollieren für potenziell systematische Unterschiede;
sie halten eine auftretende Veränderung (z.B. Fähigkeiten) konstant
(d.h. ermöglich einen Vergleich zwischen WPs und Nicht-WPs mit ansonsten gleichen Fähigkeiten)

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 gibt sich dann als eine gewichtete Summe der Mittelwert-Differenzen.

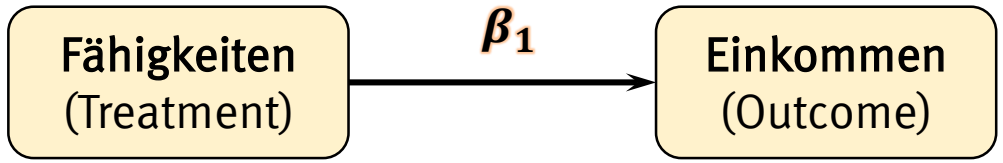
Frage: Was ist der kausale Effekt eines WP-Examens auf das Einkommen?



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

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

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



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

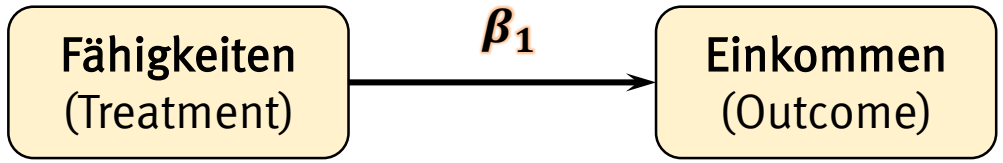
oder

2) Berechnung eines linearen Regressionsmodells:

$$Einkommen = \beta_0 + \boldsymbol{\beta_1} \times Fähigkeiten + \varepsilon$$

Live Session:
5 Minuten

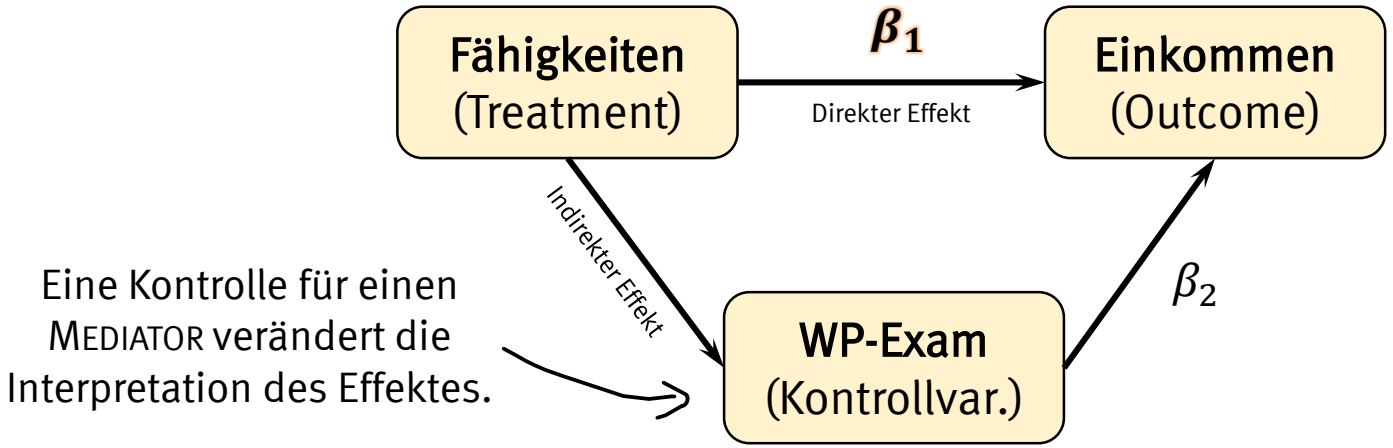
Frage: Was ist der direkte Effekt der individuellen Fähigkeiten auf das Einkommen?



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

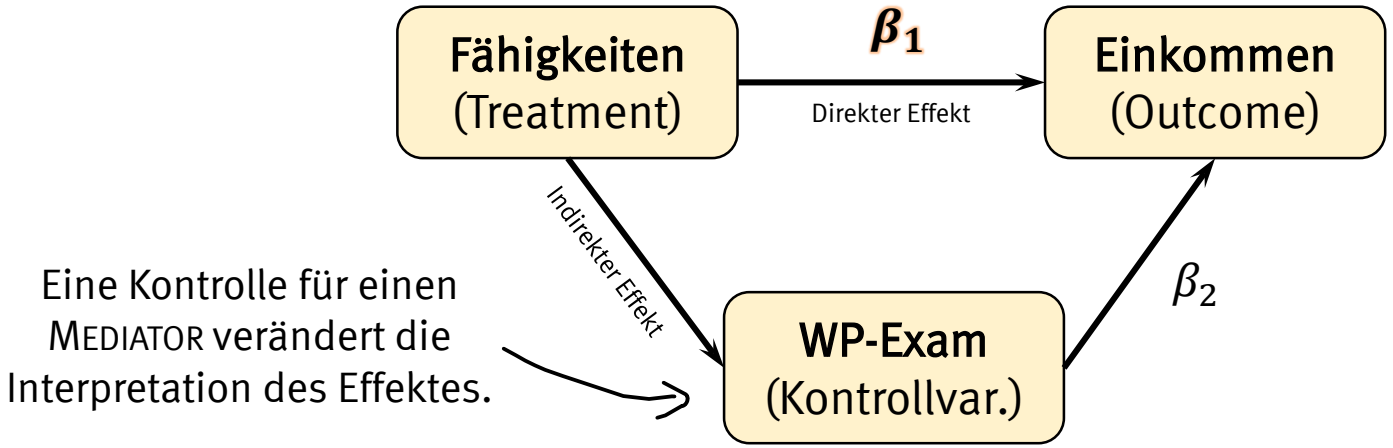
Lösung: Verwendung einer **Kontrollvariable** für WP-Exam
(d.h. ermöglich einen Vergleich zwischen Accountants mit unterschiedlichen Fähigkeiten bei ansonsten gleicher Ausbildung)

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



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

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



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

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

Live Session:
5 Minuten

Agenda

Forschungsteam Berens

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 diverse 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.