Michael Kühhirt

Design und Analyse von Quasi-Experimenten

Spezielle Erhebungsverfahren · Wintersemester 2015/16 · Universität zu Köln

zuletzt aktualisiert: 27. Oktober 2015

Inhaltsverzeichnis

Hinweise für Studierende

Dies ist der Veranstaltungsplan und das Skript zu den ersten drei Sitzungen des Kurses *Design und Analyse von Quasi-Experimenten* im Masterstudiengang *Soziologie und empirische Sozialforschung* der *Universität zu Köln*. Enthalten sind Informationen zum Inhalt und Ablauf der Veranstaltung sowie die fachlichen Grundlagen für die Themengebiete der ersten drei Vorlesungen. Letztere ermöglichen Ihnen eine eigenständige Vor- und Nachbereitung der jeweiligen Sitzung. In den darauf folgenden Sitzungen bilden verschiedene andere Texte die Grundlagenliteratur. Die sorgfältige Lektüre des Skripts und der weiteren Basistexte ist die Mindestvoraussetzung für eine erfolgreiche Teilnahme. Ich empfehle jedoch, zumindest punktuell auch die jeweils verwiesene weiterführende Literatur für die einzelnen Themen heranzuziehen. Die Sitzungen der Veranstaltung bestehen nicht in der Eins-zu-Eins-Wiedergabe der Basistexte; das wäre Zeitverschwendung. Vielmehr dienen sie dazu, durch die Behandlung von Beispielen die Inhalte weiter zu veranschaulichen und das Gelesene durch die Diskussion von Unklarheiten und Fragen zu festigen und zu vertiefen.

Inhalt der Veranstaltung

Sind westliche Medien in der Lage, die Zustimmung zu autoritäten Regimen zu untergraben (Kern und Hainmueller, 2009)? Fördern kleine Schulklassen den Lernerfolg (Angrist und Lavy, 1999)? Hat die Wohngegend einen Einfluss auf kriminelles Verhalten (Kirk, 2009)? Führen übergewichtige Freunde zu eigenem Übergewicht (Christakis und Fowler, 2007)? Möchten Sie solche oder ähnliche Fragen empirisch beantworten, können gleichzeitig aber kein randomisiertes Experiment durchführen, dann sind Sie auf quasi-experimentelle Untersuchungen angewiesen. Quasi-Experimente – im weitläufigen Sinne – sind empirische Studien des kausalen Einflusses einer Variable X, dem *Treatment* (z.B. Konsum westlicher Medien), auf eine Outcomevariable Y (z.B. Zustimmung zur Regierung). Was sie von experimentellen Untersuchungen unterscheidet, ist die fehlende Kontrolle des Forschers

über die genauen Ausprägungen des Treatments sowie über deren Zuweisung auf die Teilnehmer.

Der Versuch der Gewinnung von Wissen zu Kausalbeziehungen aus empirischen Daten allgemein wird als *kausale Inferenz* bezeichnet. Kausale Inferenz ist ein zentrales Ziel der empirischen Sozialforschung, aber nicht das einzige. Weitere wichtige Aufgaben der empirischen Sozialforschung sind die genaue Beschreibung von Populationen (z.B. Wie hoch ist die Zustimmung zur Regierung?) und sozialer Phänomene (z.B. Wie hat sich die Zustimmung zur Regierung über die Zeit geändert?) sowie die Erstellung statistischer Vorhersagemodelle, beispielsweise zu politischen Wahlen, zur Bevölkerungsentwicklung oder dem Auftreten von Verbrechen.

Anders als für diese Forschungsunterfangen, sind für die Beantwortung von Fragen nach Kausalbeziehungen empirische Daten und statistische Methoden alleine allerdings nicht ausreichend. Der Grund besteht darin, dass Kausalfragen stets eine *kontrafaktische* Komponente aufweisen, über die keine empirischen Daten vorliegen. So sind Kausalfragen stets „Was wäre, wenn? "-Fragen. Zur Beantwortung von Kausalfragen müssen Daten mit Kausalannahmen bzw. einer Theorie über die der Fragestellung zugrunde liegende Kausalstruktur, dem *Datengenerierungsprozess*, kombiniert werden. Die Validität der Untersuchungsergebnisse ist damit nicht nur abhängig von hochwertigen Daten sondern sie ist auch unmittelbar verknüpft mit der Qualität der theoretischen Überlegungen und des Vorwissens zum Datengenerierungsprozess.

Die Veranstaltung bietet eine anwendungsorientierte Einführung in grundlegende Logik und Probleme kausaler Inferenz, die aktuellsten quasi-experimentellen Untersuchungsdesigns sowie in für diese Designs angemessene Analysemethoden. Der erste Teil der Veranstaltung widmet sich zunächst einer klaren Unterscheidung zwischen Korrelation und Kausalität sowie der Einführung von graphischen Kausalmodellen (sog. *Directed Acyclic Graphs*, kurz: *DAGs*), mit deren Hilfe Kausalfragen in konkrete statistische Analysen übertragen werden können. Im Anschluss beschäftigt sich die Veranstaltung mit der Logik und der Analyse von *Instrumentalvariablendesigns*, *Regression Discontinuity Designs*, sowie *Mehrebenendesigns* (z.B. *Twin* und *Sibling Designs* oder *Paneldesigns*). Jedes Design wird zunächst anhand einer klassischen Forschungsfrage in seinen Grundzügen vorgestellt. Daran schließt sich die Analyse von aus diesen Designs gewonnenen Daten in *Stata* sowie eine Vertiefung durch die Besprechung aktueller empirischer Studien an. Der folgende Sitzungsplan gibt eine detaillierte Übersicht zum Verlauf der Veranstaltung sowie zur jeweiligen Lektüre.

Sitzungsplan und Lektüre

19 Okt: Korrelation und Kausalität

Basistext: Skript Kapitel 1

26 Okt: Identifikation kausaler Parameter

Basistext: Skript Kapitel 2

02 Nov: Schätzen kausaler Parameter

Basistext: Skript Kapitel 3

09 Nov: Intrumentalvariablendesigns (IV)

Basistext: Hernán und Robins (2006); Glymour (2006a)

16 Nov: IV: Analyse

Basistext: Swanson und Hernán (2013); Glymour et al. (2012)

23 Nov: IV: Diskussion und Erweiterungen

Basistext: Kern und Hainmueller (2009); Kirk (2009)

30 Nov: Regression Discontinuity Designs (RDD)

Basistext: Lee und Lemieux (2010, S. 281-299, S. 343-351); Steiner et al. (2015, S. 9-12)

07 Dez: RDD: Analyse

Basistext: Lee und Lemieux (2010, S. 307-327, S. 329-336); Calonico et al. (2014)

14 Dez: RDD: Diskussion und Erweiterungen

Basistext: Vardardottir (2013); Loeffler und Grunwald (2015)

11 Jan: Mehrebenendesigns

Basistext: Morgan und Winship (2015, S. 363-392); weiterer Text wird noch bekannt gegeben

18 Jan: Analyse von Mehrebenendesigns

Basistext: wird noch bekannt gegeben

25 Jan: Zusammenfassung, Diskussion, Fragen zur Klausur

Basistext: Musterklausur

1 Feb: Klausur

16:00-17:00Uhr, IBW, Hörsaal H 115

Grundlagenliteratur

Der Inhalt der Veranstaltung basiert auf den Fortschritten in der Formalisierung und dem Verständnis kausaler Inferenz über die letzten 40 Jahre durch Akteure in Statistik, Ökonometrie, Philosophie und künstlicher Intelligenz. Diese Forschritte bestehen weniger in der Entwicklung neuer statistischer Verfahren (für eine Ausnahme siehe z.B. Robins und Hernán, 2009), als vielmehr in einem besseren Verständnis der Stärken, vor allem aber auch der Probleme existierender Methoden. Die zentralen Bausteine moderner Kausalanalyse, und damit auch dieser Veranstaltung, sind das kontrafaktische Verständnis von Kausalität (Holland, 1986; Neyman, 1923; Rubin, 1974), (graphische) Modelle von Kausalstrukturen bzw. dem Datengenerierungsprozess (Heckman, 2005; Pearl, 2009b; Spirtes et al., 2001; Wright, 1921) sowie die Wiederentdeckung und Weiterentwicklung (quasi-)experimenteller Untersuchungsdesigns (Campbell und Stanley, 1963; Cook und Campbell, 1979; King et al., 1994; Rosenbaum, 2010).

In der jüngsten Vergangenheit ist zudem eine Reihe von Lehrbüchern erschienen, die aus jeweils unterschiedlichen fachlichen Richtungen eine Einführung in kausale Inferenz bieten. Das in meinen Augen didaktisch gelungenste und zugänglichste Grundlagenbuch ist *Causal Inference* von den Epidemiologen und Biostatistikern Miguel Hernán und James Robins (2016).[1] Weitere gute Einführungen aus soziologischer und ökonom(etr)ischer Perspetive bieten die zweite(!) Auflage von Morgan und Winship (2015) bzw. die beiden Bücher von Angrist und Pischke (2009, 2015). Imbens und Rubin (2015) ist ein Buch mit Fokus auf das kontrafaktische Modell, das graphische Kausalmodelle ablehnt. Aus diesem Grund ist es nur eingeschränkt zu empfehlen. Murnane und Willett (2010) sowie Shadish et al. (2002) konzentrieren sich wiederum auf Untersuchungsdesigns und Methoden, mit nur minimaler Beschäftigung mit den formalen Grundlagen kausaler Inferenz. Beide Bücher behandeln zudem keine graphischen Kausalmodelle. Mit dem *Handbook of Causal Analysis for Social Research* (Morgan, 2013) steht zudem noch eine Sammlung von Spezialartikeln zu den wichtigsten Themen im Bereich kausale Inferenz zur Verfügung.

Neben diesen ausführlicheren Veröffentlichungen sind auch mehrere Überblicksartikel zum Thema kausale Inferenz in verschiedenen Fachzeitschriften erschienen. Zu nennen sind hier u.a. Gangl (2010) für die Soziologie, Keele (2015) für die Politikwissenschaft und Imbens und Wooldridge (2009) für die VWL. Zusammenfassungen der zentralen Punkte der Arbeiten zu graphischen Kausalmodellen von Pearl (2009b) finden sich bei Elwert (2013), Glymour (2006b), Steiner et al. (2015) und Pearl (2009a, 2010).

[1] Das Buch ist noch nicht erschienen, die Kapitel 1-17 aber online verfügbar.

Was wir nicht (ausführlich) behandeln

In der Veranstaltung beschäftigen wir uns in erster Linie mit einem Teilbereich kausaler Inferenz, der als *causal prediction* bezeichnet wird. Hierbei geht es um die Frage, *ob und wie stark* ein Treatment das Outcome kausal beeinflusst. Die Frage, *warum und wie* dieser Effekt zustande kommt, werden wir nur am Rande betrachten, wenn wir über die Bedeutung kausaler Mechanismen sprechen. Diese sog. *causal explanation* oder *causal mediation analysis* ist selbstverständlich relevant und interessant, aber noch einmal deutlich anspruchsvoller als *causal prediction*. Für Interessierte empfehle ich die beiden ersten und gerade erst erschienenen Lehrbücher zu diesem Thema von VanderWeele (2015) und Hong (2015).

Bei der Besprechung der Untersuchungsdesigns und statistischer Schätzmethoden konzentrieren wir uns auf deren grundlegende Logik und Funktionsweise, nicht auf die mathematische Herleitung. Für mathematische (aber immer noch verständliche) Details zu den einzelnen Methoden verweise ich Sie auf das Buch von Wooldridge (2010).

Ebenso bietet der Kurs keine Einführung in die volle Pracht von Stata, sondern wir nutzen lediglich die für die hier genutzen Verfahren nötigen Befehle. Auch die Aufbereitung und Säuberung von Daten werden wir außen vor lassen. Die Daten, die Sie erhalten, sind bereits relativ übersichtlich aufbereitet. Das heißt aber nicht, dass die Datenaufbereitung unwichtig wäre. Im Gegenteil: hier können schon schwerwiegende Fehler unterlaufen, die jede weitere Analyse untergraben. Für einige zentrale Hinweise und Anleitungen für die Durchführung und Dokumentation empirischer Analysen sowie weiterführende Literatur zu diesem Thema schauen Sie bitte in das diesbezügliche *Protokoll* auf unserer Lehrstuhlhomepage.

Schließlich werden wir während des Kurses auch weitestgehend annehmen, dass wir die uns interessierenden theoretischen Konstrukte perfekt messen können (z.B. IQ-Testergebnis ist ein perfektes Maß für Intelligenz). Wir werden also das Problem von *Messfehlern* zumeist ausblenden. Grund dafür ist, dass Messfehler kein spezifisches Problem kausaler Inferenz oder quasi-experimenteller Designs sind, sondern alle Arten empirischer Untersuchungen gefährden. Andererseits werden wir sehen, dass kausale Inferenz auch unter der Annahme perfekter Messung schwierig genug ist. Eine exzellente Integration des Messfehlerproblems in graphische Kausalmodelle findet sich jedoch bei Hernán und Robins (2016, Kap. 9).

Voraussetzungen und Anforderungen

Von Studierenden im Masterstudiengang erwarte ich ein Mindestmaß an fachlichem Interesse sowie die Bereitschaft sich selbstständig (und in der Gruppe) mit den Veranstaltungsinhalten auseinanderzusetzten. Das bloße Erscheinen zur Veranstaltung ist in keinem Fall ausreichend. Neben diesen Qualitäten sollten Sie auch bestimmte methodische und statistische Kenntnisse mitbringen, um der Veranstal-

tung folgen zu können. Diese werden im Folgenden konkretisiert. Im Anschluss daran werden die einzelnen Prüfungsleistungen dargestellt.

Erforderliche Vorkenntnisse

Grundlagenkenntnisse im Bereich Methoden und Statistik auf dem Stand der Veranstaltung *Lineare Modelle* bzw. *Analysis of Cross-Sectional Data* sind für das Verständnis der Veranstaltungsinhalte zwingend erforderlich. Auch hier gehe ich davon aus, dass Sie diese nicht nur gehört haben, sondern beherrschen und auch anwenden können. Zudem empfehle ich den vorherigen Besuch der Veranstaltung *Kausalanalyse* bzw. *Analysis of Longitudinal Data*.

Prüfungsleistungen

Die Prüfungsleistung setzt sich aus drei Teilen zusammen: einer 60-minütigen Klausur, *sieben* Aufgaben zur Lektüre und *drei* Stata-Übungen. Insgesamt können 60 Punkte erreicht werden, davon 50 Punkte in der Klausur. Die restlichen 10 Punkte können durch eine beliebige Kombination von Bearbeitungen der Lektüreaufgaben (je 1 Punkt) und Stata-Übungen (je 3 Punkte) erworben werden. Mehr als 10 Punkte durch Lektüreaufgaben und Stata-Übungen zu erwerben ist *nicht* möglich.

Bitte beachten Sie: Für das Erreichen der 10 Punkte muss mindestens eine Stata-Übung *und* mindestens eine Lektüreaufgabe erfolgreich bearbeitet werden. Ohne jegliche Bearbeitung der Aufgaben können Sie die Veranstaltung maximal mit der Note 2,0 abschließen.

Aufgaben zur Lektüre: Für sieben ausgewählte Sitzungen erhalten Sie jeweils eine kurze Aufgabenstellung, die Sie zu Hause in Vorbereitung auf die Sitzung bearbeiten und fristgerecht über *Ilias* einreichen. Die adäquate Bearbeitung der Aufgabe wird mit jeweils 1 Punkt bewertet. Es handelt sich dabei stets um kurze Fragen zum Inhalt der Lektüre für die jeweilige Sitzung oder auch kleine Anwendungsprobleme desselben. Es ist erlaubt (und Sie werden ausdrücklich ermutigt!), die Aufgaben in der Gruppe zu besprechen. Dennoch muss jeder Teilnehmer eine eigenständige Bearbeitung einreichen. *Identische Einreichungen werden mit 0 Punkten bewertet!* Sofern Sie die Nichtbearbeitung einer Aufgabe entschuldigen können (z.B. wegen Krankheit), besteht die Möglichkeit, die Bearbeitung bis eine Woche nach Ablauf der Entschuldigung nachzureichen.

Für die Einreichung ist Folgendes zu beachten:

- ca. 0,5 Seite, Times, 12 pt, einfacher Zeilenabstand

- auf *Ilias* als PDF mit Ihrem Namen und der Aufgabennummer in der ersten Zeile
- Dateiname: [IhrNachname]_[Aufgabennummer].pdf
- jeweils bis Sonntag, 23:59Uhr, vor der Veranstaltung
- Abgabe per Email ist *nicht möglich!*

Bitte beachten Sie: Falls Sie Ihr Geschriebenes selbst nicht verstehen, werde ich es voraussichtlich auch nicht verstehen und somit keinen Punkt vergeben. Kontrollieren Sie Ihre Bearbeitung auch gründlich hinsichtlich Rechtschreibung und Grammatik! Wörtliche Übernahmen aus der Lektüre sind nicht zulässig. Ich möchte wissen, ob Sie das Gelesene verstanden haben und mit eigenen Worten wiedergeben können, nicht, ob Sie in der Lage sind, die passende Textpassage ausfindig zu machen und abzuschreiben.

Stata-Übungen: Die Analyse von Instrumentalvariablendesigns, Regression Discontinuity Designs und von Mehrebenendesigns in Stata wird durch drei kleine empirische Arbeiten eingeübt. Die Bearbeitung wird jeweils mit max. 3 Punkten bewertet. Zentral dabei ist, die korrekte Modellierung der Kausalstrukur in Form eines DAGs (1P), die entsprechende Umsetzung in Stata (1P) und die angemessene Interpretation der Ergebnisse (1P). Sie erhalten für jede Stata-Übung ein gesondertes Aufgabenblatt mit genaueren Anweisungen. Die Bearbeitung findet während der Veranstaltung statt, um Hilfestellung geben zu können, kann aber auch eigenständig erfolgen.

Für die Einreichung ist Folgendes zu beachten:

- auf *Ilias* als *eine* PDF-Datei mit schriftlicher Bearbeitung der Aufgaben, relevantem Stata-Code und -Output an geeigneter Stelle im Text
- Name und Übungsnummer in der ersten Zeile
- Text: Times, 12 pt, einfacher Zeilenabstand
- Code und Output: Courier New, 10 pt, einfacher Zeilenabstand
- Dateiname: [IhrNachname]_[Übungsnummer].pdf
- jeweils 14 Tage nach der Übungssitzung
- Abgabe per Email ist *nicht möglich!*
- Achten Sie auf die Lesbarkeit des Dokuments und passen Sie ggf. Schriftgröße und Seitenformat an

Klausur: Am Ende des Semesters findet eine 60-minütige Klausur statt. Dort wird sowohl vermitteltes Wissen abgefragt, als auch dessen korrekte Anwendung geprüft. In der Klausur sind maximal 50 Punkte zu erreichen. Datum und Ort der Klausur werden rechtzeitig bekannt gegeben. Es gibt keinen zweiten Klausurtermin. Für Aufbau der Klausur und Beispielaufgaben, schauen Sie sich bitte die Musterklausur auf *Ilias* an.

Bitte beachten Sie: Für die *Anmeldung* zur Klausur ist die Kursmitgliedschaft in *KLIPS* oder *Ilias nicht ausreichend*! Eine gesonderte Anmeldung beim Prüfungsamt ist zwingend erforderlich. Bitte informieren Sie sich eigenständig über die Anmeldefrist.

Benotung: Die Gesamtnote der Veranstaltung ergibt sich aus der Summe der max. 50 Punkte aus der Klausur und der max. 10 Punkte aus der Bearbeitung von Lektüreaufgaben und Stata-Übungen. Die Prüfung gilt als bestanden, wenn Sie mindestens 30 Punkte erreichen. Die Notenvergabe erfolgt nach dem folgenden Schema:

1,0: $60 \geq$ Punkte ≤ 58
1,3: $58 >$ Punkte ≤ 55
1,7: $55 >$ Punkte ≤ 51
2,0: $51 >$ Punkte ≤ 48
2,3: $48 >$ Punkte ≤ 45
2,7: $45 >$ Punkte ≤ 42
3,0: $42 >$ Punkte ≤ 39
3,3: $39 >$ Punkte ≤ 36
3,7: $36 >$ Punkte ≤ 33
4,0: $33 >$ Punkte ≤ 30
n.b.: $30 >$ Punkte ≤ 0

Organisatorisches

In den folgenden Abschnitten finden Sie noch einige abschließende Hinweise zu den Kursmaterialien sowie zu Ansprechpartnern bei Fragen und Problemen.

Kursmaterialien

Die Materialien zur Veranstaltung, also Skript, Literatur, Folien, Aufgabenblätter, do-files und Datensätze finden Sie allesamt auf *Ilias*. Dort wird Ihnen auch ausgewählte weiterführende Literatur (v.a. Artikel) zur Verfügung gestellt. Die meisten der oben erwähnten Fachbücher finden Sie in einem spezifischen Handapparat für die Veranstaltung in der Fachbibliothek Soziologie, Greinstaße 2.

Auf *Ilias* ist zudem ein Diskussionsforum angelegt, das dem Austausch über die Veranstaltungsinhalte, inkl. der Übungsaufhaben, dienen soll. Dort sind auch hilfreiche Seiten zum Thema kausale Inferenz aber auch zur Arbeit mit Stata verlinkt.

Bei Fragen zur Veranstaltung

Nutzen Sie das erwähnte *Ilias*-Forum bitte für alle allgemeinen Fragen an mich bzgl. Inhalt und Ablauf der Veranstaltung. Damit wird gewährleistet, dass andere Teilnehmer, die vielleicht gleiche oder ähnliche Fragen haben, meine Antwort auch

einsehen können. Eine Email an mich schreiben Sie bitte nur dann, wenn es sich um eine Frage handelt, die nur Sie persönlich betrifft, z.B. die Vereinbarung eines Sprechstundentermins oder Fragen zu meinem Feedback zu Ihren Aufgaben.

Bei Fragen zum Studium

Wenn Sie generell Fragen zu Ihrem Studium haben, z.B. zur Semesterplanung, zu ECTS-Punkten, zu Härtefallregelungen, wenden Sie sich bitte an die Geschäftsführung des Instituts in Person von Joël Binckli (binckli@wiso.uni-koeln.de). Geht es um allgemeine Auskünfte und Informationen zu Formalitäten ist das Sekretariat von Fr. Petra Altendorf (sekretariat.soziologie@wiso.uni-koeln.de) der richtige Anlaufpunkt. Machen Sie bitte bei allen Anfragen folgende Angaben: Name, Matrikelnr., ggf. Prüfungsnr., Studienrichtung, angestrebter Abschluss, ggf. Grund- oder Hauptstudium.

Teil I
Grundlagen kausaler Inferenz

Kapitel 1
„Korrelation“ und Kausalität

1.1 Empirische Daten und statistischer Zusammenhang

Um Kausalbeziehungen zu untersuchen, hat man letztlich keine andere Wahl als auf empirische Daten und auf aus ihnen zu errechnende „Korrelation“ – oder genauer gesagt: statistische (Un-)Abhängigkeit[1] – zurückzugreifen. Empirische Daten sind nichts anderes als Matrizen bestehend aus Merkmalen bzw. Variablen und Untersuchungseinheiten, die sich in bestimmter Weise über die möglichen Ausprägungen der Variablen verteilen. In der Veranstaltung werden alle Variablen als Zufallsvariablen behandelt und mit Großbuchstaben bezeichnet.[2] Kleinbuchstaben stehen für die Ausprägungen dieser Variablen für einzelne Untersuchungseinheiten. Das Subskript i bezeichnet die Untersuchungseinheiten. Damit steht der Ausdruck $Y_i = y$ für „Variable Y hat die Ausprägung y für Untersuchungseinheit i“. Die (marginale bzw. unbedingte) Verteilung dieser Variable bezeichen wir als $Pr(Y)$.

Beispiel 1.1. Einen Ausschnitt aus einer Datenmatrix zeigt Tabelle 1.1. Dargestellt sind ausgesuchte Merkmale früherer Teilnehmer dieser Veranstaltung, inklusive der Anwesenheitshäufigkeit (in Prozent aller Sitzungen) und der Punktzahl in der Klausur. Für Untersuchungsungseinheit 30 in Tabelle 1.1 schreiben wir damit $\text{kpoints}_{30} = 52$.

---

[1] Korrelation ist eine spezifische Form statistischer Abhängigkeit, die sich auf Abhängigkeit in Mittelwerten bzw. Erwartungswerten bezieht.

[2] Fürs Erste nehmen wir an, dass unsere Daten eine unendlich große Zufallsstichprobe aus einer ebenso unendlichen und klar definierten Zielpopulation (der statistischen Superpopulation) darstellen. Daher entsprechen die empirischen Häufigkeitsverteilungen der Variablen in den Daten ihren Wahrscheinlichkeitsverteilungen in der unbeobachteten Superpopulation. Auch Parameter wie das arithmetische Mittel in der Stichprobe sind damit equivalent zum Erwartungswert in der Superpopulation. Später (in Kapital 3) ersetzen wir diese Perspektive durch die herkömmliche Sichtweise: die Daten entsprechen einer endlichen Zufallsstichprobe aus einer genau definierten, aber meist ebenso endlichen Zielpopulation und es kommt zu Abweichungen zwischen empirischer Verteilung und Wahrscheinlichkeitsverteilung, die über den Standardfehler quantifiziert werden.

Tabelle 1.1 Matrix aus empirischen Daten mit drei Variablen und zehn Untersuchungseinheiten

```
. list sex pattend kpoints in 30/39, div sep(1)
```

	sex	pattend	kpoints
30.	maennlich	83	52
31.	weiblich	75	55
32.	maennlich	73	48
33.	weiblich	92	50
34.	weiblich	64	49
35.	weiblich	82	56
36.	maennlich	75	52
37.	maennlich	83	44
38.	weiblich	45	42
39.	weiblich	25	42

Quelle: `01kausal.do@2`

1.1.1 Definition statistischer Parameter

In *Populations*wissenschaften wie der Soziolgie, der Politikwissenschaft, der Ökonomie oder der Demographie führen wir, anders als Mediziner, Therapeuten oder Sozialpädagogen, keine Einzelfallbetrachtungen durch, sondern nutzen verschiedene *statistische* Parameter, um die Verteilungen $Pr(Y)$ zusammenzufassen und zu beschreiben. Über die Lage der Untersuchungseinheiten in Bezug auf die Messskala der Variablen geben Parameter wie Modus, Median ($Q_{0,5}[Y]$) oder arithmetisches Mittel ($E[Y]$) Auskunft. Die Streuung der Untersuchungseinheiten über die Ausprägungen wird durch Parameter wie Quantilsabstände oder Standardabweichung ($\sigma[Y]$) erfasst.

Beispiel 1.2. Die Verteilungen sowie ausgewählte Lage- und Streuungsparameter für Anwesenheitshäufigkeit und Punkte in der Klausur dieser Veranstaltung aus früheren Semestern sind in Abbildung 1.1 dargestellt. Wir sehen, dass die meisten Teilnehmer, mehr als 50% der Sitzungen besucht haben. Alle Teilnehmer (zumindest diejenigen, die die Klausur geschrieben haben) waren zumindest einmal anwesend. Der Median (`p50`) zeigt uns, dass die Hälfte der Teilnehmer 72% der Sitzungen oder mehr besucht haben. Die durchschnittliche Anwesenheitsquote (`mean`)

beträgt 66,81. Die mittlere Abweichung von diesem Mittel, also die Standardabweichung (`sd`), beträgt 22,32.

Für die Klausurpunkte zeigt sich, dass alle Teilnehmer mindestens 35 Punkte erreicht und damit bestanden haben. Die volle Punktzahl hat bisher jedoch niemand erreicht. Weiterhin hat die Hälfte der Teilnehmer 50 Punkte oder mehr erreicht. Die durchschnittliche Punktzahl beträgt ebenso rund 50 Punkte. Die Standardabweichung beträgt 6,11 Punkte.

Statistischer Zusammenhang

Auch die *gemeinsame* Verteilung von zwei Variablen, Y und X, kann untersucht werden. Dazu wird die Verteilung von Y nach den Ausprägungen von X dargestellt. Das heißt, man unterteilt die Gesamtmenge der Beobachtungen in Teilpopulationen mit gleichen Ausprägungen auf X und vergleicht dann die Verteilung von Y zwischen diesen Teilpopulationen.[3] Die resultierenden Verteilungen $Pr(Y|X)$ bezeichnet man auch als *bedingte* oder *konditionale* Verteilungen. Der „|“ steht dementsprechend für „unter der Bedingung“ bzw. „gegeben“. Der Vergleich bedingter Verteilungen untereinander bzw. mit der marginalen (d.h. unbedingten) Verteilung $Pr(Y)$ gibt Aufschluss darüber, ob ein statistischer Zusammenhang zwischen den Variablen vorliegt.

Definition 1.3. (*statistischer Zusammenhang*) Von einem statistischen Zusammenhang ganz allgemein spricht man dann, wenn sich die Verteilung einer Variable in irgendeiner Form für mindestens zwei Ausprägungen einer zweiten Variablen, also in mindestens zwei durch diese Variable definierten Teilpopulationen, unterscheidet. Anders gesagt, die *marginale* Verteilung einer Variable weicht ab von der durch die zweite Variable *bedingten* Verteilung:

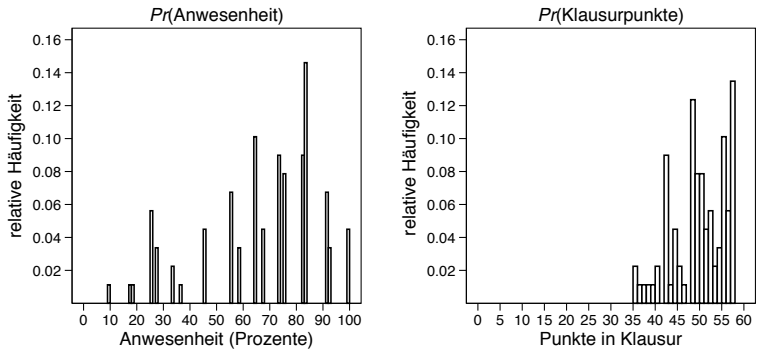
$$Pr(Y) \neq Pr(Y|X). \tag{1.1}$$

Definition 1.4. (*Statistische Unabhängigkeit*) Statistische Unabhängigkeit zwischen zwei Variablen ist folglich gegeben, wenn sich die Verteilung einer Variable *nicht* nach den Ausprägungen der zweiten Variable unterscheidet:

$$Pr(Y) = Pr(Y|X). \tag{1.2}$$

[3] Die Verteilung von Y kann auch nach *mehreren* Variablen $\mathbf{X}$ dargestellt werden. In diesem Fall unterteilt man die Beobachtungen in Gruppen definiert über all möglichen Kombinationen der Ausprägen von $\mathbf{X}$, im Fall von zwei $\mathbf{X}$-Variablen also in $x1 \times x2$ Gruppen.

```
. #delimit ;
delimiter now ;
. hist pattend, width(1) frac
>               color(white) lcolor(black)
>               xlabel(0(10)100, labsize(small))
>               ylabel(.02(.02).16, angle(0)
>               labsize(small) format(%3.2f))
>               xsize(5) ysize(5) ytitle("relative Häufigkeit")
>               title("{it:Pr}(Anwesenheit)", size(medsmall))
>               saving("${output}/01fig1a", replace) ;
(bin=91, start=9, width=1)
(file /Users/wmb222/Dropbox/Arbeit/Lehre/causality/2015WS/output/01fig1a
> .gph saved)
. hist kpoints, width(1) frac
>               color(white) lcolor(black)
>               xlabel(0(5)60, labsize(small))
>               ylabel(.02(.02).16, angle(0)
>               labsize(small) format(%3.2f))
>               xsize(5) ysize(5) ytitle(" relative Häufigkeit")
>               title("{it:Pr}(Klausurpunkte)", size(medsmall))
>               saving("${output}/01fig1b", replace) ;
(bin=23, start=35, width=1)
(file /Users/wmb222/Dropbox/Arbeit/Lehre/causality/2015WS/output/01fig1b
> .gph saved)
. graph combine "${output}/01fig1a" "${output}/01fig1b",
>               xsize(10) ysize(5) iscale(1.3) ;
. #delimit cr
delimiter now cr
```




```
. tabstat pattend kpoints, stat(med mean sd) format(%4.2f)
```

stats	pattend	kpoints
p50	73.00	50.00
mean	66.81	49.56
sd	22.32	6.11

Quelle: `01kausal.do@3`

Abb. 1.1 Verteilung von Anwesenheitshäufigkeit und Klausurpunkten für Teilnehmer früherer Semester

Zur Untersuchung statistischer Zusammenhänge können wiederum auch verschiedene Parameter der bedingten Verteilungen verglichen werden. Für Variablen mit metrischem Skalenniveau wird meist ausschließlich ein Vergleich der arithmitischen Mittel durchgeführt. Ein Zusammenhang liegt dann vor, wenn

$$E[Y|X = x] \neq E[Y|X = x'], \tag{1.3}$$

sich also das arithmetische Mittel von Y, $E[Y]$, für mindestens zwei verschiedene Ausprägungen von X, nämlich x und x' (lies: „nicht x“), unterscheidet.

Es können jedoch auch andere Parameter verglichen werden, beispielsweise Median oder Standardabweichung. Auch diese geben Aufschluss über statistische Zusammenhänge und sind für bestimmte Fragen aussagekräftiger als Mittelwertvergleiche.[4]

Für kategoriale Variablen, für die bekanntlicherweise kein arithmetisches Mittel berechnet werden kann, sind Unterschiede in der Häufigkeit einzelner Ausprägungen von Y nach Ausprägungen von X, also

$$Pr[Y = y|X = x] \neq Pr[Y = y|X = x'], \tag{1.4}$$

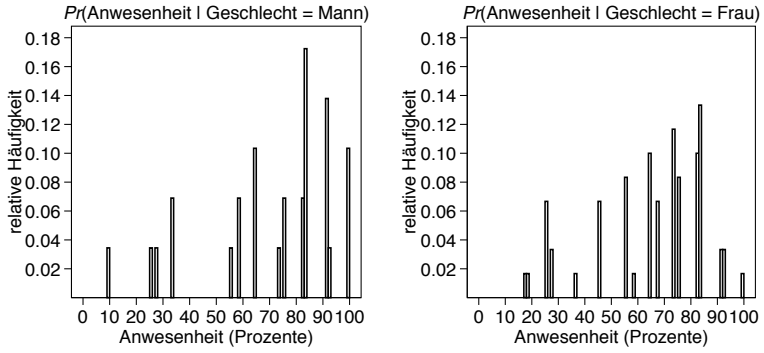
gleichbedeutend mit einem Zusammenhang zwischen beiden Variablen.

Beispiel 1.5. Abbildung 1.2 zeigt zur Veranschaulichung dieser Konzepte die Verteilung von Anwesenheit nach Geschlecht, also Pr(Anwesenheit | Geschlecht), zusammen mit den bedingten Medianen $Q_{0,50}$(Anwesenheit | Geschlecht), Mittelwerten E(Anwesenheit | Geschlecht) und Standardabweichungen σ(Anwesenheit | Geschlecht). Dazu werden zunächst alle Beobachtungen den Teilpopulationen „Männer“ und „Frauen“ zugeordnet und daraufhin die Verteilung und ihre Parameter in beiden Gruppen verglichen.

Dieser Vergleich weist auf einen statistischen Zusammenhang zwischen Anwesenheit und Geschlecht hin. So haben Frauen die Veranstaltung tendenziell seltener besucht als Männer. Sowohl Median als auch arithmetisches Mittel der Anwesenheit sind für Frauen geringer. Ein Blick auf die bedingten Verteilungen zeigt jedoch auch, dass die niedrigste Anwesenheit überhaupt für die Gruppe der männlichen Teilnehmer beobachtet wurde.

[4] Der Einfachheit halber konzentrieren wir uns jedoch in der Veranstaltung ebenso auf letztere.

```
. #delimit ;
delimiter now ;
. hist pattend if sex==0, width(1) frac
>               color(white) lcolor(black)
>               xlabel(0(10)100)
>               ylabel(.02(.02).18, angle(0) format(%3.2f))
>               xsize(5) ysize(5) ytitle("relative Häufigkeit")
>               title("{it:Pr}(Anwesenheit | Geschlecht = Mann)"
>             , size(medsmall))
>               saving("${output}/01fig2a", replace) ;
(bin=91, start=9, width=1)
(file /Users/wmb222/Dropbox/Arbeit/Lehre/causality/2015WS/output/01fig2a
> .gph saved)
. hist pattend if sex==1, width(1) frac
>               color(white) lcolor(black)
>               xlabel(0(10)100)
>               ylabel(.02(.02).18, angle(0) format(%3.2f))
>               xsize(5) ysize(5) ytitle("relative Häufigkeit")
>               title("{it:Pr}(Anwesenheit | Geschlecht = Frau)"
>             , size(medsmall))
>               saving("${output}/01fig2b", replace) ;
(bin=83, start=17, width=1)
(file /Users/wmb222/Dropbox/Arbeit/Lehre/causality/2015WS/output/01fig2b
> .gph saved)
. graph combine "${output}/01fig2a" "${output}/01fig2b",
>               xsize(10) ysize(5) iscale(1.3) ;
. #delimit cr
delimiter now cr
```




```
. tabstat pattend, by(sex) stat(med mean sd) format(%4.2f) notot
Summary for variables: pattend
     by categories of: sex (Geschlecht)
      sex |        p50       mean         sd
----------+------------------------------------
maennlich |      82.00      70.62      24.70
 weiblich |      73.00      64.97      21.04
----------------------------------------------
```

Quelle: `01kausal.do@4`

Abb. 1.2 Bedingte Verteilung von Anwesenheitshäufigkeit nach Geschlecht für Teilnehmer früherer Semester

Statistische Zusammenhangsmaße

Die Stärke eines Zusammenhangs kann auf verschiedene Weise quantifiziert werden. So lassen sich *Differenzen* zwischen Mittelwerten (oder anderen Parametern) bzw. Prozentsätzen berechnen:

$$\Delta E[Y|X] = E[Y|X=x] - E[Y|X=x'] \tag{1.5}$$

$$\Delta Pr[Y=y|X] = Pr[Y=y|X=x] - Pr[Y=y|X=x'] \tag{1.6}$$

Je größer die Differenz, desto stärker der Zusammenhang. Eine Differenz von Null ist gleichbedeutent mit der Abwesenheit eines statistischen Zusammenhangs dieser Parameter.

Unterschiede können jedoch auch in Form von *Verhältnissen* abgebildet werden:

$$\Phi E[Y|X] = \frac{E[Y|X=x]}{E[Y|X=x']} \tag{1.7}$$

$$\Phi Pr[Y=y|X] = \frac{Pr[Y=y|X=x]}{Pr[Y=y|X=x']} \tag{1.8}$$

Hier ist der Wert „1“ gleichbedeutend mit statistischer Unabhängigkeit dieser Parameter. Je weiter das Verhältnis nach oben oder nach unten von 1 abweicht, desto stärker der Zusammenhang.

Für kategoriale Variablen kann zudem noch das Odds Ratio berechnet werden:

$$\Psi Pr[Y=y|X] = \frac{Pr[Y=y|X=x]/Pr[Y=y'|X=x]}{Pr[Y=y|X=x']/Pr[Y=y'|X=x']} \tag{1.9}$$

Auch hier ist die 1 gleichbedeutend mit statistischer Unabhängigkeit. Beim Odds Ratio ist zudem darauf zu achten, dass es nicht als Wahrscheinlichkeitsverhältnis interpretiert wird, sondern als Chancenverhältnis.

Beispiel 1.6. Der statistische Zusammenhang zwischen Anwesenheit und Geschlecht gemessen durch die Mittelwertdifferenz beträgt

$$\Delta E[\text{Anwesenheit}|\text{Geschlecht}] = 70,62 - 64,97 = 5,65$$

Somit ist die Anwesenheitsquote von Männern im Mittel 5,65 Punkte höher als die von Frauen. Der statistische Zusammenhang ausgedrückt in Verhältnissen lautet folgendermaßen:

$$\Phi E[\text{Anwesenheit}|\text{Geschlecht}] = \frac{70,62}{64,97} = 1,09$$

In Worten: Die Anwesenheitsquote von Männern ist im Mittel um das 1,09-fache (oder auch: um 9%) höher als die von Frauen.

Die berechneten Parameter bezeichnet man als Zusammenhangsmaße. Es gibt noch eine ganze Reihe weiterer, komplexerer Zusammenhangsmaße. So können zur Berechnung *linearer* Zusammenhänge auf Basis von Mittelwerten (die eigentliche Korrelation) der Bravais-Pearsonsche Korrelationskoeffizient (r) oder die Koeffizienten einer OLS-Regression verwendet werden. Auch für kategoriale und ordinalskalierte Variablen gibt es komplexere Zusammenhangsmaße, beispielsweise χ^2-basierte Maße oder Spearmans Rangkorrelationskoeffizient sowie nichtlineare Regressionsmodelle (z.B. *Logit* oder *Probit*). Letztlich berechnen sämtliche statistische Analyseverfahren, egal wie aufwendig und kompliziert, „Korrelationen“. Allen Zusammenhangsmaßen ist gemein, dass sie Unterschiede (und deren Stärke) in der Verteilung einer Variable zwischen nach den Ausprägungen einer zweiten Variable definierten Teilpopulationen messen. Umgekehrt bedeutet dies: Unterscheidet sich die Verteilung einer Variablen nicht nach den Ausprägungen einer weiteren Variable (also: $Pr(Y) = Pr(Y|X)$), dann zeigen diese Maße keinen Zusammenhang an (z.B. $r = 0$).

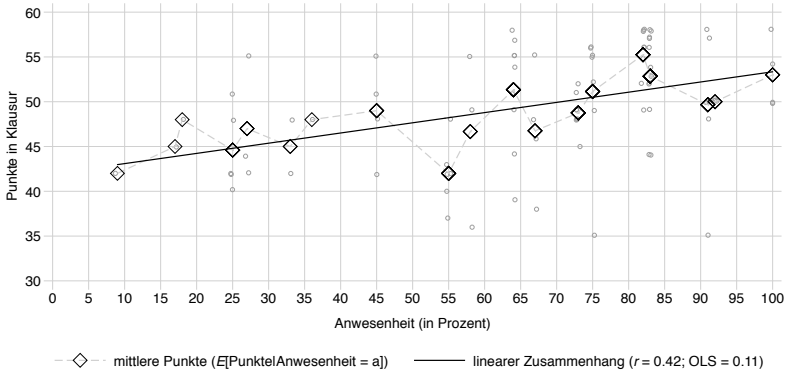
Beispiel 1.7. Abbildung 1.3 zeigt den Zusammenhang zwischen Anwesenheitshäufigkeit und Klausurpunkten für Teilnehmer der Veranstaltung in früheren Semestern. Da die Klausurpunkte metrisch skaliert sind, können wir einen Mittelwertvergleich nach Anwesenheitshäufigkeit vornehmen. Es zeigt sich, dass sich die mittlere Punktzahl nach den Ausprägungen von Anwesenheit unterscheidet. Damit liegt ein statistischer Zusammenhang zwischen beiden Variablen vor. Durch Augenmaß können wir zudem feststellen, dass sich auch die Streuung der Punktzahl für die Ausprägungen der Anwesenheit unterscheidet.

Für die Untersuchung der *Richtung* des Zusammenhangs (auf Basis der Mittelwerte) können wir komplexere Zusammenhangsmaße einsetzen. Der Bravais-Pearsonsche Korrelationskoeffizient deutet auf einen relativ starken positiv-linearen Zusammenhang der bedingten Mittelwerte hin ($r = 0,42$). Der Anstieg des eingezeichneten linearen Regressionsmodells sagt uns, dass die Punktzahl in der Klausur mit jedem zusätzlichen Punkt auf der Anwesenheit im Mittel um 0,11 Punkte zunimmt. Das bedeutet, zwei Teilnehmer unterscheiden sich um einen Punkt in der Klausur, wenn sich ihre Anwesenheitsquote um rund 9 Punkte unterscheidet. Für 10 Punkte Unterschied in der Klausur sind 90 Punkte Unterschied in der Anwesenheitsquote notwendig.

Bedingte statistische Zusammenhänge

Neben so genannten *marginalen* Zusammenhängen, also einfachen bivariaten Zusammenhängen, können auch *bedingte* (oder konditionale) Zusammenhänge zwischen zwei Variablen berechnet werden. Das bedeutet, der Zusammenhang wird nun getrennt nach den Ausprägungen einer dritten Variable berechnet. Hier sind also nicht mehr lediglich Unterschiede in der Verteilung nach einer Variablen von Interesse, sondern Unterschiede im Zusammenhang von zwei Variablen nach einer weiteren Variablen. Ein Unterschied im Zusammenhang liegt vor, wenn dieser

```
. #delimit ;
delimiter now ;
. twoway scatter   kpoints pattend, jitter(1) msize(small)
>                  msymbol(Oh) mcolor(gs10) mlwidth(vthin)     ||
>        connected m_kpoints pattend, lcolor(gs13) msymbol(Dh)
>                  mcolor(black) mlwidth(vthin) msize(large)   ||
>        lfit      m_kpoints pattend, lcolor(black) lpattern(solid)
>                , xlabel(0(5)100
>                , notick grid gmin glwidth(vthin) glcolor(gs13))
>                  ylabel(30(5)60
>                , notick angle(0) grid gmin
>                  glwidth(vthin) glcolor(gs13))
>        ytitle("Punkte in Klausur") plotregion(color(white))
>        xtitle("Anwesenheit (in Prozent)", margin(medsmall))
>        yscale(lpattern(blank)) xscale(lpattern(blank))
>        xsize(10) ysize(5)
> legend(label(2 "mittlere Punkte ({it:E}[Punkte|Anwesenheit = a])")
>        label(3 "linearer Zusammenhang ({it:r} = 0.42; OLS = 0.11)")
>        order(2 3) region(color(white))) ;
. #delimit cr
delimiter now cr
```



Quelle: `01kausal.do@5D`

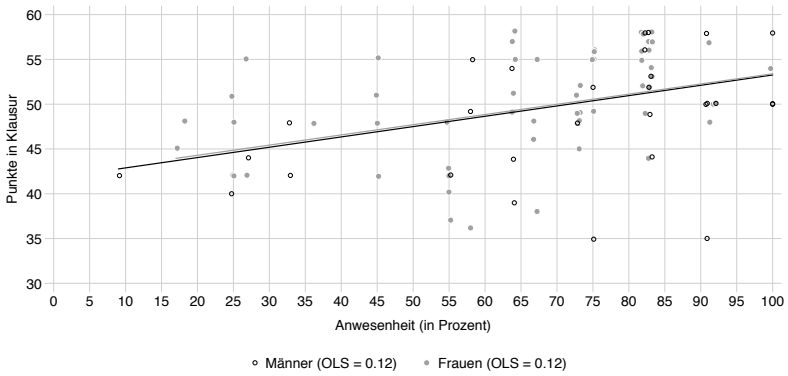
Abb. 1.3 Verteilung von Klausurpunkten nach Anwesenheit für Teilnehmer früherer Semester

sich für mindestens zwei Ausprägungen der dritten Variable unterscheidet. Für eine Mittelwertdifferenz beispielsweise also

$$\Delta E[Y|X, W = w] \neq \Delta E[Y|X, W = w']. \tag{1.10}$$

Für die Berechnung bedingter Zusammenhänge, also Zusammenhängen zwischen zwei Variablen „unter Kontrolle“ weiterer Variablen, werden meist Regressionsverfahren verwendet. Es gibt aber auch eine Reihe weiterer Techniken, die wir in Kapitel 3 kurz besprechen (z.B. einfache Stratifizierung, *Matching* und Gewichtungsverfahren).

```
. #delimit ;
delimiter now ;
. twoway scatter   kpoints pattend if sex==1, jitter(1) msize(medium)
>                  msymbol(O) mcolor(gs10)
>                  mlcolor(white) mlwidth(vthin)                    ||
>        scatter   kpoints pattend if sex==0, jitter(1) msize(small)
>                  msymbol(Oh) mcolor(black) mlwidth(vthin)         ||
>        lfit      m_kpoints pattend if sex==0
>                , lcolor(black) lpattern(solid)
>                , xlabel(0(5)100, notick)
>                  ylabel(30(5)60, notick angle(0))                 ||
>        lfit      m_kpoints pattend if sex==1
>                , lcolor(gs10) lpattern(solid)
>                , xlabel(0(5)100
>                , notick grid gmin glwidth(vthin) glcolor(gs13))
>                  ylabel(30(5)60
>                , notick angle(0) grid gmin
>                  glwidth(vthin) glcolor(gs13))
>        ytitle("Punkte in Klausur") plotregion(color(white))
>        xtitle("Anwesenheit (in Prozent)", margin(medsmall))
>        yscale(lpattern(blank)) xscale(lpattern(blank))
>        xsize(10) ysize(5)
> legend(label(2 "Männer (OLS = 0.12)")
>        label(1 "Frauen (OLS = 0.12)")
>        order(2 1) region(color(white))) ;
. #delimit cr
delimiter now cr
```



Quelle: `01kausal.do@6C`

Abb. 1.4 Zusammenhang zwischen Klausurpunkten und Anwesenheit nach Geschlecht der Teilnehmer

Beispiel 1.8. Für den Zusammenhang zwischen Klausurpunkten und Anwesenheit zeigt sich kein Unterschied nach Geschlecht, zumindest gemessen an Mittelwertunterschieden oder genauer: OLS-Regressionskoeffizienten (Abb. 1.4).

1.1.2 Interpretation statistischer Zusammenhänge

Was sagen uns nun aber diese und weitere statistische Zusammenhangsmaße inhaltlich? Zunächst ist festzuhalten, dass statistische Zusammenhänge im Speziellen und empirische Daten allgemein stets den beobachteten und damit *faktischen* Ist-Zustand *beschreiben*. Sie geben Auskunft darüber, wie sich die Ausprägungen bestimmter Merkmale zwischen verschiedenen existierenden Teilpopulationen (definiert nach den Ausprägungen weiterer Merkmale) unterscheiden. Empirische Daten enthalten damit keine Informationen über die Verteilung von Variablen unter alternativen oder *kontrafaktischen* Bedingungen wie sie durch eine *Änderung* bestimmter anderer Merkmale (und nur dieser Merkmale) entstehen würden, beispielsweise die Verabschiedung einer politischen Reform.

Diese Erkenntnis ist unproblematisch, solange sie bei der Interpretation empirischer Analysen berücksichtigt wird. Die bloße faktische Beschreibung von Populationen ist inhaltlich durchaus bedeutsam. So geben statistische Zusammenhänge (korrekte Messung und ggf. Modellierung vorausgesetzt) beispielsweise Auskunft über systematische Gruppenunterschiede in Bildung, Einkommen, Gesundheit, sozialer Teilhabe und weiteren Dimensionen sozialer Ungleichheit. Auch statistische Vorhersagen sind auf Basis von Zusammenhängen möglich. Mit Wissen über die Ausprägung einer Variable kann die Ausprägung der korrelierenden Variable ungefähr bestimmt werden. Mit Hilfe solcher Vorhersagen versucht man beispielsweise zukünftige Schwerpunktgebiete für Verbrechen zu lokalisieren oder ganz einfach den nächsten deutschen Fußballmeister zu bestimmen. Wirtschaftsunternehmen nutzen statistische Vorhersagemodelle für Produktempfehlungen oder Preisentwicklungen. Epidemiologen nutzen sie, um Risikogruppen für bestimmte Krankheiten zu ermitteln.

Die Probleme beginnen dann, wenn statistische Zusammenhänge ohne Weiteres als „Effekte“ interpretiert werden. Exemplarisch ist hier die traditionelle Lehrbuchinterpretation von Regressionen: „Ändert sich X um eine Einheit, steigt Y um β Einheiten.“ Diese impliziert, dass eine Vorhersage darüber möglich ist, wie die Verteilung von Y unter einem Alternativszenario nach Änderung von X aussehen würde. Eine Vorhersage ist jedoch lediglich dazu möglich, welchen Wert Y (im Mittel) annimmt, wenn für X eine bestimmte Ausprägung passiv *beobachtet* wird (und umgekehrt). Diese Vorhersage hält nur solange Stand, bis der Ist-Zustand (inklusive X) sich ändert bzw. geändert wird. Ein statistischer Zusammenhang an sich ist weder Evidenz für noch gegen einen Kausalzusammenhang.[5] Um empirische Daten für kausale Inferenz, zur Schätzung kausaler Effekte, zu nutzen, ist eine Reihe von Annahmen zum Datengenierungsprozess notwendig, die empirisch nicht getestet werden können. Aus diesem Grund ist es von fundamentaler Bedeutung diese Annahmen zu kennen und für jede Analyse explizit zu benen-

[5] Sogar folgendes Szenario ist möglich: Ein positiver Zusammenhang wird zwischen X und Y beobachtet. Daraufhin wird beschlossen X in der Population zu erhöhen. Voller Vorfreude wird Y gemessen, so dass der Schock groß ist, als festgestellt wird, dass Y im Mittel gesunken ist. Hier besteht ein negativer Kausalzusammenhang zwischen X und Y, obwohl ein positiver statistischer Zusammenhang beobachtet wurde.

nen. Nur so können diese Annahmen und damit kausale Inferenz insgesamt einer substantiellen Evaluation und Plausibilitätsprüfung im Rahmen des wissenschaftlichen Diskurses unterzogen werden. Ohne Beschäftigung mit diesen Annahmen ist die Nutzung von Kausalsprache für die Interpretation statistischer Analysen grob fahrlässig.

Beispiel 1.9. Auch auf Basis der Daten aus der Veranstaltung sind statistische Vorhersagen möglich. Wir haben gesehen, dass eine häufigere Anwesenheit mit einer besseren Klausurleistung einher geht. Sind Sie also eine Person, die häufig die Veranstaltung besucht, können Sie mit einer höheren Punktzahl rechnen als jemand, der seltener oder nie anwesend war. Die „Vorhersage“ funktioniert sogar rückwärts. Kenne ich die Punktzahl in der Klausur einer Person, kann ich daraus auf die Häufigkeit der Anwesenheit schließen.

Die Daten allein sagen allerdings nichts über die zu erwartende mittlere Punktzahl in der Klausur aus, wenn alle Teilnehmer zu 100% Teilnahme verpflichtet gewesen wären. Denn dieses Szenario war nicht gegeben. Folglich enthalten die Daten keine Informationen über die Verteilung der Klausurnote unter diesem Szenario. Gleichsam geben Ihnen die Daten keinen Hinweis darauf, ob Sie *durch* häufigere Anwesenheit mehr Punkte erzielen können. Genau dies ist aber von Interesse, wenn wir die Kausalbeziehung zwischen Klausurpunkten und Anwesenheit untersuchen wollen.

1.2 Kausalmodelle und kontrafaktische Kausalität

Bei der Untersuchung von Kausalzusammenhängen ist also die Kernfrage, ob eine Änderung in (der Verteilung) einer Kausalvariable, hier Treatment genannt[6], eine Änderung in (der Verteilung) einer Outcomevariable [7] hervorruft. Anders als bei der Berechnung statistischer Zusammenhänge, mit denen man die vorhandenen Daten lediglich passiv beschreibt, impliziert eine Kausalfrage also einen aktiven Eingriff, eine (zumindest hypothetische) Intervention in den Prozess der Datengenerierung.

Beispiel 1.10. Fragen wir nach dem kausalen Effekt der Anwesenheitshäufigkeit auf die Klausurleistung, interessiert uns nicht die unter dem *gegebenen* Teilnahmeverhalten vorliegende – und damit beobachtbare – Verteilung der Klausurpunkte. Stattdessen benötigen wir Informationen zu den Folgen einer spezifischen Änderung in der Verteilung der Anwesenheit für die Verteilung der Klausurnote. Beispielsweise könnte sich die Universität fragen, ob eine Anwesenheitspflicht, also die Festsetzung der Anwesenheit auf 100% oder zumindest 75% für alle Teilnehmer, den Lernerfolg und damit die Klausurleistung verbessert. Studierende wiederum sind vielleicht interessiert, ob sie auch ohne regelmäßige Teilnahme die gleiche

[6] Andere Bezeichnungen sind Ursache, Exposure, exogene Variable, unabhängige Variable.

[7] Weitere verwendete Begriffe sind Wirkung, Response, endogene Variable und abhängige Variable.

Note erzielen und die gewonnene Zeit lieber in Lektüre oder Freizeit investieren könnten. Hier ist somit die Anwesenheitshäufigkeit das Treatment und die Klausurpunkte sind das Outcome.

1.2.1 Graphische Kausalmodelle

Für die formale Behandlung und Definition kausaler Effekte bedarf es zunächst einer präzisen und gleichsam transparenten Form der Darstellung von den vermuteten Kausalbeziehungen zwischen den interessierenden Variablen. Graphische Kausalmodelle bieten eine solche Form der Darstellung. Mithilfe weniger Bausteine können sie das Wissen bzw. die Theorie des Forschers zu dem Prozess kommunizieren, der die Beobachtungseinheiten den Ausprägungen der für die Forschungsfrage relevanten Variablen zuweist. Dieser Prozess wird auch als Datengenerierungsprozess bezeichnet.

Grundelemente und -begriffe

Graphische Kausalmodelle bestehen aus *directed acyclic graphs*, kurz DAGs. Diese sind aufgebaut aus drei Elementen: (1) Knoten (2) gerichteten Kanten (Pfeilen) zwischen jeweils zwei Knoten und (3) *abwesenden* Kanten zwischen zwei Knoten. Knoten (○) stehen dabei für eine Variable oder für eine Gruppe von Variablen. Eine gerichtete (deswegen *directed* acyclic graph) Kante (→) zwischen zwei Knoten bedeutet, dass ein direkter Kausaleffekt in Richtung der Kante (für mindenstens ein Mitglied der Zielpopulation) vermutet wird. Die Richtung der Kante zeigt zudem die zeitliche Ordnung der Variablen an. Anders gesagt, der Variable, von der die Kante ausgeht, wird eine Ausprägung stets vor der Variable, auf die der Pfeil zeigt, zugewiesen. Aus diesem Grund sind kausale Zirkel der Form $A \rightarrow B \rightarrow A$ ausgeschlossen (deswegen directed *acyclic* graph). Denn die Zukunft kann die Vergangenheit nicht beeinflussen.[8] Das Vorhandensein einer Kante sagt jedoch an sich nichts über die Stärke oder die funktionale Form des Effekts aus. Das bedeutet auch, dass graphische Kausalmodelle von vornherein zulassen, dass Effekte über Individuen und Teilpopulationen variieren (siehe Kap. 1.2.2 zur Darstellung und Bedeutung von Effektheterogenität). Demgegenüber steht die Abwesenheit einer gerichteten Kante zwischen zwei Variablen. Dies bedeutet, dass zwischen diesen Variablen für *kein* Mitglied der Population ein direkter kausaler Effekt besteht, egal in welche Richtung (*sharp causal null*). Die Annahme einer fehlenden Kante ist somit ungleich stärker als die einer vorhandenen Kante. Denn erstere erlaubt *genau einen* möglichen Wert für den kausalen Effekt, nämlich null, während letz-

[8] Die in den Sozialwissenschaften häufig zitierte „gegenseitige Beeinflussung“ zweier Variablen scheint dem nur auf den ersten Blick zu widersprechen. Derartige vermeintliche Zirkel lassen sich mit etwas detaillierterer Notation ohne Weiteres als Kausal*ketten* darstellen, beispielsweise $\text{Einstellung}_{t=1} \rightarrow \text{Handeln}_{t=2} \rightarrow \text{Einstellung}_{t=3}$.

tere alle möglichen Werte zulässt, null eingeschlossen. Wie wir noch sehen werden, hängt das Gelingen kausaler Inferenz maßgeblich von fehlenden Kanten zwischen Variablen ab.

Beispiel 1.11. Ein aus didaktischen Gründen relativ einfaches Modell der Generierung der Daten zu Anwesenheit und Klausurpunkten ist in Abb. 1.5 dargestellt. Dieses zeigt, dass Studierende zunächst auf Basis ihrer Motivation die Auswahl ihrer Kurse treffen. Die Motivation entscheidet darüber hinaus auch über die Anwesenheitshäufigkeit, die Teilnahme an der Klausur und die in der Klausur erreichte Punktzahl. Auch die Anwesenheit wirkt sich auf die Klausurteilnahme aus, beispielsweise, da Studierende nach häufigerer Anwesenheit ein besseres Gefühl bezüglich der Erfolgswahrscheinlichkeit haben oder schlicht ihre Zeit nicht umsonst investiert haben wollen. Bezüglich der hier im Vordergrund stehenden Kausalbeziehung zwischen Anwesenheit und Klausurpunkten zeigt das Modell, dass ein kausaler Effekt von Anwesenheit auf Punktzahl vermutet wird, der zum Teil durch die Klärung von Verständnisfragen vermittelt wird. Aber das Modell enthält auch Informationen über nicht vorhandene Kausaleffekte. So haben beispielsweise weder die Kurswahl noch die Klausurteilnahme einen Einfluss auf irgendeine andere sich im Modell befindliche Variable. Ferner wird postuliert, dass sich die Motivation der Studierenden nicht direkt auf die Klärung von Fragen auswirkt.

Kausale und nicht kausale Pfade

Eine Kette von über Kanten verbundenen Knoten in DAGs, in der jeder enthaltende Knoten höchstens einmal durchlaufen wird, bezeichnet man auch als Pfad. Diese lassen sich in kausale und nichtkausale Pfade unterscheiden. In kausalen Pfaden deuten alle Kanten in diesselbe Richtung. Als nichtkausal werden alle anderen Pfade bezeichnet. Nichtkausale Pfade enthalten demzufolge mindestens eine Kante, die in eine andere Richtung deutet, als alle anderen. Kausale Pfade können noch unterschieden werden in direkte Kausaleffekte, bei denen zwei Variablen unmittelbar durch eine Kante verbunden sind, und indirekte Kausaleffekte, bei denen der kausale Pfad zwischen zwei Variablen über eine dritte Variable verläuft. Direkte

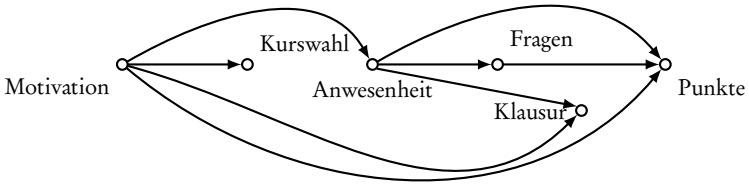


Abb. 1.5 Modell des Generierungsprozesses der Teilnehmerdaten

und indirekte Kausaleffekte zusammen (also die Summe aller Kausalpfade von einer Variable zu einer zweiten) ergeben den Gesamteffekt (*total causal effect*) einer Variable auf eine zweite.

Beispiel 1.12. Pfade in Abb. 1.5 sind Anwesenheit ← Motivation → Kurswahl oder auch Fragen → Punkte ← Anwesenheit → Klausur. Kein Pfad hingegen ist Motivation → Anweseheit → Klausur ← Motivation, da Motivation hier mehr als einmal auftritt. Beispiele für kausale Pfade sind Anwesenheit → Klausur oder Motivation → Anwesenheit → Fragen → Punkte. Die zuerst genannten Pfade hingegen sind gänzlich nichtkausal, da jeweils mindestens eine Kante auftritt, die nicht in dieselbe Richtung zeigt wie die restlichen Kanten. Die beiden Kausalpfade Anwesenheit → Punkte und Anwesenheit → Fragen → Punkte bilden zusammen den Gesamteffekt von Anwesenheit auf Punkte. Der erstgenannte Pfad wird dabei auch als direkter Effekt bezeichnet, der zweite Pfad als indirekter Effekt von Anwesenheit auf Punkte.

Verwandtschaftsbeziehungen zwischen Variablen und kausale DAGs

Das bestehen von Kausalpfaden zwischen mehreren Variablen impliziert bestimmte Verwandtschaftsbeziehungen. So bezeichnet man alle Variablen, die eine bestimmte Variable direkt oder indirekt verursachen, als deren *Vorfahren*. Die Untermenge von Vorfahren, die eine Variable direkt verursachen, werden *Eltern* genannt. Analog gelten alle direkten Folgen einer Variable als deren *Kinder*. Zu den *Nachkommen* einer Variable zählen neben ihren Kindern auch die Menge aller indirekten Folgen.

Beispiel 1.13. Die Vorfahren von Fragen in Abb. 1.5 sind Anwesenheit und Motivation. Die einzige Elternvariable von Fragen ist Anwesenheit. Zu den Nachkommen von Motivation zählen Kurswahl, Anwesenheit, Fragen, Klausur und Punkte. Kurswahl, Anwesenheit, Klausur und Punkte sind die Kinder von Motivation.

Damit ein DAG als *kausaler* DAG bezeichnet werden und damit als graphisches Kausalmodell genutzt werden kann, muss er sämtliche *gemeinsame* Vorfahren für jedes im DAG dargestellte Variablenpaar enthalten, unabhängig davon ob diese gemessen oder ungemessen sind. Variablen, die nicht im DAG dargestelt sind, sind damit nicht zwangsläufig ungemessen. Ihre Abwesenheit impliziert jedoch, dass sie nicht gleichzeitig zwei oder mehr der dargestellten Variablen verursachen. Auch etwaige Nachkommen der dargestellten Variablen müssen nicht im DAG enthalten sein, es sei denn, sie werden in der späteren statistischen Analyse „kontrolliert“ (siehe Kap. 2.1.4).

Wo kommt der DAG eigentlich her?

Ein (graphisches) Kausalmodell ist in erster Linie ein *theoretisches* Konstrukt. Das schließt jedoch nicht aus, dass konkretes empirisches Vorwissen in die Konstruk-

tion des Modells einfließen kann und sogar soll. Haben beispielsweise (möglichst viele) vorherige Studien (überwiegend) ergeben, dass zwischen zwei Variablen im Modell kein *Kausal*zusammenhang besteht, kann eine gerichtete Kante zwischen diesen Variablen mit gutem Grund weggelassen werden.[9] Weiß man zudem, dass eine Variable vor einer zweiten Variable im Modell gemessen wurde, kann eine von dieser zweiten Variable ausgehende gerichtete Kante auf die erste Variable ausgeschlossen werden. Auch Vorhersagen aus Theorien, die sich in der Vergangenheit empirisch gut bewährt haben, können bei der Modellbildung einfließen. Schließlich können Annahmen auch aus praktischen Gründen gemacht werden, nämlich dann, wenn ohne diese keine kausale Inferenz bezüglich der interessierenden Fragestellung möglich ist (siehe Kap. 2). Allerdings sollten diese Annahmen dem gegenwärtigen Kenntnisstand nicht zuwider laufen. Ansonsten ist die Glaubwürdigkeit der kausalen Schlussfolgerungen, die auf Grundlage des Modells getroffen werden, stark eingeschränkt. Aus diesem Grund schlagen Petersen und van der Laan (2014) für die Modellbildung eine strikte Trennung zwischen wissensbasierten Annahmen (*knowledge-based assumptions*) und ausschließlich praktisch begründeten Annahmen (*convenience-based assumptions*) vor. Die Qualität und Glaubwürdigkeit eines Kausalmodells hängt maßgeblich davon ab, wie plausibel die in ihnen vercodeten Kausalbeziehungen, und hier insbesondere die nicht vorhandenen Kausalbeziehungen, sind.

Beispiel 1.14. Die Erstellung eines Kausalmodells beginnt in der Regel mit den beiden Variablen, auf die sich die Kausalfrage bezieht. In unserem Beispiel sind das Anwesenheit und Punkzahl in der Klausur. Wir fragen, ob die Punkzahl durch die Anwesenheithäufigkeit kausal beeinflusst wird. Ausgangspunkt solcher Kausalfragen ist meist eine Idee (ein theoretisches Argument) darüber, *warum* das Treatment das Outcome beeinflussen sollte.[10] In unserem Kausalmodell in Abb. 1.5 ist mit der Möglichkeit zur Klärung von Fragen ein potentieller Faktor, über den Anwesenheit die Punktzahl beeinflussen könnte, explizit aufgeführt. Es sind jedoch noch weitere Prozesse vorstellbar, u.a. ein besseres Verständnis durch zusätzliche Beispiele oder konkrete Lernhinweise, die durch den Dozenten gegeben werden.[11] Diese und weitere Prozesse sind durch die direkte Kante von Anwesenheit auf Punktzahl repräsentiert. Da in einem kausalen DAG alle gemeinsamen Ursachen von den dargestellten Variablen enthalten sein müssen, ist auch die Motivation im Modell dargestellt. Denn die Vermutung liegt nahe, dass motivierte Studierende häufiger die Veranstaltung besuchen, gleichzeitig aber auch unabhängig von der Anwesenheit mehr Zeit in die Veranstaltung investieren und somit auch besser in der Klausur abschneiden. Problematisch am Modell ist, dass durchaus weitere gemeinsame Ursachen von Anwesenheit und Punkten in der Klausur denkbar

[9] Dies gilt jedoch nicht, sollten bisherige Studien lediglich keine Evidenz für einen statistischen Zusammenhang gefunden haben.

[10] Die postulierten Prozesse, über die der Einfluss konkret verläuft, werden auch als Kausal*mechanismen* bezeichnet.

[11] Kein Grund für einen Effekt ist, dass die Anwesenheit an sich in die Bewertung eingeht, wie von früheren Teilnehmern vermutet bzw. befürchtet.

sind, diese aber nicht im Modell enthalten sind. Ein Beispiel hierfür sind die Vorkenntnisse der Studierenden. Über die Annahme, dass Vorwissen entweder nur Anwesenheit oder nur die Punkte beeinflusst und deswegen nicht dargestellt werden muss, wird (später) zu diskutieren sein. Wenig umstritten dürften hingegen die fehlenden Kanten von Anwesenheit auf Kurswahl oder von Punktzahl auf Klausurteilnahme haben, da erstere Variablen den letzteren jeweils zeitlich nachgelagert sind.

DAGs als nichtparametrische Strukturgleichungsmodelle

Durch die explizite Berücksichtigung der die übrigen im DAG dargestellten Variablen nicht direkt beeinflussenden Vorfahren jeder Variablen (zusammengefasst als ε_V) wird schnell deutlich, dass DAGs als graphische Repräsentation nichtparametrischer Strukturgleichungsmodelle interpretiert werden können (siehe Abb. 1.6). Das heißt, die durch Kanten dargestellten Kausalstrukturen zwischen den Variablen können auch in Form mathematischer Gleichungen formuliert werden. Hierbei wird jede Variable V als (nichtparametrische) Funktion ihrer Eltern *und nur ihrer Eltern* $p(V)$ dargestellt:

$$V = f(p(V)) \tag{1.11}$$

Beispiel 1.15. Der in Abb. 1.6 dargestellte DAG kann somit auch durch das folgende System von Strukturgleichungen notiert werden:

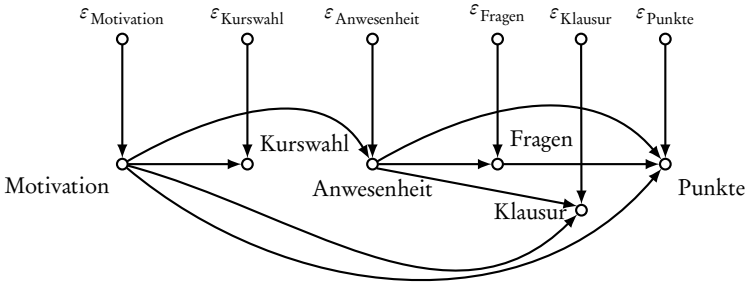


Abb. 1.6 Graphisches Kausalmodell aus Abb. 1.5 als nichtparametrisches Strukturgleichungsmodell

$$\begin{aligned}
\text{Motivation} &= f(\varepsilon_{\text{Motivation}}) && (1.12)\\
\text{Kurswahl} &= f(\text{Motivation}, \varepsilon_{\text{Kurswahl}}) && (1.13)\\
\text{Anwesenheit} &= f(\text{Motivation}, \varepsilon_{\text{Anwesenheit}}) && (1.14)\\
\text{Fragen} &= f(\text{Anwesenheit}, \varepsilon_{\text{Fragen}}) && (1.15)\\
\text{Klausur} &= f(\text{Motivation}, \text{Anwesenheit}, \varepsilon_{\text{Klausur}}) && (1.16)\\
\text{Punkte} &= f(\text{Motivation}, \text{Anwesenheit}, \text{Fragen}, \varepsilon_{\text{Punkte}}) && (1.17)
\end{aligned}$$

Jede Variable ist eine Funktion ihrer direkten Ursachen, also ihrer Eltern. So ergibt sich Anwesenheit aus Motivation und $\varepsilon_{\text{Anwesenheit}}$. Dass alle anderen Variablen auf der rechten Seite der Gleichung für Anwesenheit fehlen, bedeutet im Umkehrschluss, dass diese Variablen keinen direkten Einfluss auf Anwesenheit haben, im graphischen Modell ersichtlich durch das Fehlen einer auf Anwesenheit gerichteten Kante.[12]

Nachteil der Darstellung von Kausalmodellen als Systemen von Gleichungen ist, dass Gleichheitszeichen an sich keine Richtung implizieren. In mathematischen Gleichungen sind stets Umformungen möglich, durch die Terme auf der rechten Seite auch als Funktion von Termen der ursprünglich linken Seite dargestellt werden können.[13] Eine gerichtete Kante hingegen lässt intuitiv nur eine Interpretation zu: die Variable, von der die Kante ausgeht, beeinflusst die Variable, auf die die Kante zeigt. Graphische Kausalmodelle bieten damit auch für mathematische Laien die Möglichkeit, die im Modell beschriebenen Strukturen zu verstehen und auch beobachtbare Implikationen aus ihnen abzuleiten. Ein wichtiger Vorteil von Strukturgleichungen ist jedoch wiederum, dass die (vermutete) funktionale Form des Kausalzusammenhangs bei Bedarf auch explizit gemacht werden kann, indem auf der rechten Seite statt der nichtparametrischen Ausdrücke $f(p(V))$ konkrete Funktionen spezifiziert werden. Nimmt man also an, dass die Variable Punkte aus Abb. 1.6 linear und additiv von ihren Eltern abhängt, kann dies auch mit einer entsprechend spezifizierten Gleichung ausgedrückt werden:

$$\text{Punkte} = \theta_1 \text{Motivation} + \theta_2 \text{Anwesenheit} + \theta_3 \text{Fragen} + \varepsilon_{\text{Punkte}} \qquad (1.18)$$

Eine beliebige Anzahl an weiteren Spezifikationen ist möglich. Dazu gehören auch nichtlineare Zusammenhänge oder Interaktionen zwischen den Eltern. Noch zu betonen ist, dass Strukturgleichungen dieser Art nicht mit statistischen Modellen zu verwechseln sind, die beobachtete Zusammenhänge in den Daten beschreiben. Strukturgleichungen dagegen modellieren nicht direkt beobachtbare Kausalzusammenhänge. So können Strukturgleichungen anders als statistische Modelle auch Variablen enthalten, die in den jeweiligen Daten nicht gemessen wurden.

[12] Im Englischen bezeichnet man diese Modellannahme, die Variablen als Eltern einer zweiten Variable ausschließt, deswegen auch als *exclusion restriction*.

[13] Generell ist mathematische bzw. statistische *Standard*notation nicht ideal zur Beschreibung von Kausalzusammenhängen.

1.2.2 Kausale Effekte als Folgen von (hypothetischen) Interventionen im Datengenerierungsprozess

Kausalmodelle, ob in Form eines DAGs oder eines Systems von Strukturgleichungen, zeigen (die Theorie des Forschers), wie „die Natur" den interessierenden Variablen (in der Zielpopulation) die Ausprägungen zuweist. Sie sind Modelle der Funktionsweise der (sozialen) Welt. Kapitel 2 erläutert, wie aus diesem Modell beobachtbare statistische Zusammenhänge abgeleitet werden können. Graphische Kausalmodelle sind jedoch ebenso ausgezeichnet dazu geeignet, kausale Effekte auf nachvollziehbare Weise formal zu definieren und damit von statistischen Zusammenhängen abzugrenzen. So werden kausale Effekte als Ergebnis eines Eingriffs in den „natürlichen" Datengenerierungsprozess verstanden, durch den die interessierende Variable unabhängig von den weiteren an der Datengenerierung beteiligten Variablen auf eine bestimmte Art und Weise *geändert* wird. Der Forscher wird damit vom passiven Beobachter der „Natur", zumindest hypothetisch, zu ihrem aktiven Manipulator.

Hypothetische Interventionen und Pearls *Do*-Operator

Fragen nach dem Kausaleffekt einer Variable auf eine zweite lassen sich als (hypothetische) *Interventionen* im Datengenerierungsprozess konzeptionalisieren. Was würde mit (der Verteilung von) Y passieren, wenn X *unabhängig von allen anderen Variablen im Modell* für alle Mitglieder der Zielpopulation auf einen bestimmten Wert *festgesetzt* würde? Im graphischen Kausalmodell wird diese Intervention dadurch repräsentiert, dass alle gerichteten Kanten auf die geänderte Variable gelöscht werden. Dies verdeutlicht die *exogene* Festsetzung der Ausprägung der Variable auf einen bestimmten Wert. Anders als bei der Beobachtung verschiedener Ausprägungen einer Variable für verschiedene Individuen in empirischen Daten, wird hier die Ausprägung für alle Mitglieder der Population aktiv festgelegt. Für die Notation derartiger Interventionen auf eine Variable im Kausalmodell hat Judea Pearl (1995, 2009b) den sogenannten *do*-Operator vorgeschlagen. Eine Festsetzung der Variable X auf die Ausprägung x wird damit ausgedrückt als $do(X = x)$.[14]

Beispiel 1.16. Abb. 1.7 zeigt die graphische Repräsentation einer hypothetischen Intervention im Datengenerierungsprozess unseres Beispiels. Durch die exogene Festsetzung der Verteilung der Anwesenheit, beispielsweise auf 100% für alle Teilnehmer ($do(\text{Anwesenheit} = 100)$) wird der eigentlich vorhandene Einfluss der Motivation (und anderer möglicher Faktoren) auf die Anwesenheit ausgeschaltet. Damit kann der Pfeil von Motivation auf Anwesenheit gelöscht werden.

[14] Generell sind auch andere Interventionen als die Festsetzung einer für alle gleichen Ausprägung denkbar. Beispielsweise könnte man das arithmetische Mittel der Treatmentverteilung verschieben ($do(E[X] = x)$) oder deren Varianz verändern ($(do(\sigma[X] = s)$). Der Einfachheit wegen belassen wir es hier jedoch bei Interventionen, die allen Mitgliedern der Population pauschal die gleiche Ausprägung einer Variable zuweisen.

Kontrafaktische (Verteilung) von Outcomevariablen und kausale Effekte

Durch die Intervention bleiben die anderen Kausalzusammenhänge im Modell unberührt. Für die (vermuteten) Nachkommen der geänderten Variablen, darunter auch das interessierende Outcome, ergibt sich nun aber eine *neue Verteilung* unter der Kausalbedingung $do(X = x)$. Diese neue Verteilung, $Pr(Y|do(X = x))$ beschreibt nicht die in empirischen Daten beobachtbare bedingte Verteilung $Pr(Y|X = x)$, sondern die unter dem *kontrafaktischen* Szenario, dass alle Mitglieder der Zielpopulation die gleiche Ausprägung auf der Treatmentvariable zugewiesen bekommen.[15] Durch den Vergleich der kontrafaktischen Verteilungen von Y nach mindestens zwei alternativen Interventionen auf X kann nun untersucht werden, ob tatsächlich ein kausaler Effekt von X auf Y vorliegt, ob also Y ein Nachkomme von X ist.

Definition 1.17. (*Kausaler Effekt*) Ein kausaler Effekt[16] der Variable X auf die Variable Y lässt sich damit formal definieren als Unterschied in der kontrafaktischen Verteilungen von Y nach mindestens zwei alternativen (hypothetischen) Interventionen auf X:

$$Pr(Y|do(X = x)) \neq Pr(Y|do(X = x')). \tag{1.19}$$

Definition 1.18. (*Kausaler Nulleffekt*) Es liegt ein kausaler Nulleffekt (also kein Effekt) von X auf Y vor, wenn sich die Verteilung von Y für keine aller möglichen Interventionen auf X unterscheidet:

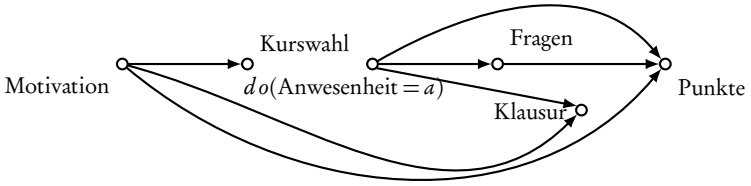


Abb. 1.7 Durch Intervention (do) auf Anwesenheit modifiziertes Modell des Datengenerierungsprozesses aus Abb. 1.5

[15] Neben der Schreibweise mit dem *do*-Operator finden sich in der Literatur noch weitere Notationen für das Outcome Y unter den Kausalbedingungen $X = x$. Am geläufigsten sind $Y(x)$, Y^x, $Y^{X=x}$ und Y_x. Zusätzlich werden für das Treatment häufig die Buchstaben A, T oder D anstelle von X verwendet. Die Notationsvielfalt kann beim Heranziehen verschiedener Literatur schnell zu Verwirrung führen. Umso wichtiger ist es, nicht die Notation auswendig zu lernen, sondern den Sinn dahinter zu verinnerlichen.

[16] Die Bezeichnung eines Effekts als *kausal* ist streng genommen redundant. Denn es gibt keine Effekte, außer kausale Effekte. Die nichtsdestotrotz häufige Nutzung des Wortes Effekt bei der Interpretation statistischer Analysen, ist ein weiteres Beispiel für das Durcheinander, das in Bezug auf kausale Inferenz in der angewandten Forschung herrscht. Die Nutzung des Begriffs Effekt für die Bezeichnung statistischer Zusammenhänge ist hochgradig irreführend und lädt geradezu ein zu Fehlinterpretationen.

$$Pr(Y|do(X=x)) = Pr(Y|do(X=x')), \text{ für alle } X=x. \tag{1.20}$$

Beispiel 1.19. Für die Untersuchung des Effekts der Anwesenheit auf die Punkte in der Klausur kann also neben der Festsetzung der Anwesenheit auf 100% eine zweite Intervention durchgeführt werden, bei der die Anwesenheit für alle Teilnehmer auf beispielsweise 75% gesetzt wird. Anschließend werden die Verteilungen der Klausurpunkte nach der jeweiligen Intervention verglichen. Unterscheiden sie sich, sagen wir, dass eine Änderung der Anwesenheit von 75% auf 100% die Klausurleistung beeinflusst. Sind beide Verteilungen gleich, bedeutet das, diese Änderung hat keine Auswirkung auf die erreichten Punkte. Um sagen zu können, dass eine Änderung der Anwesenheit generell keinen Einfluss auf die erreichten Punkte hat, dürften nach der Zuweisung aller denkbaren Ausprägungen zwischen keinen der resultierenden kontrafaktischen Punkteverteilungen ein Kontrast auftreten.

Der Vergleich der nach Intervention resultierenden Verteilungen der Outcomevariable zeigt dem Forscher somit, ob tatsächlich einer oder mehrere kausale Pfade von der Treatmentvariable auf die Outcomevariable verlaufen. Im Umkehrschluss bedeutet jeder Kausalpfad im Modell zwischen zwei beliebigen Variablen, dass eine Intervention auf der Variable, von der der Pfad ausgeht, eine Änderung in der Variable, in der der Pfad endet, zur Folge hat.[17] Besteht kein Kausalpfad zwischen zwei Variablen, folgt auf eine Intervention auf die erste Variable keine Änderung der zweiten Variable.

Maße für Kausaleffekte

Für die Quantifizierung kausaler Effekt lassen sich analog zu statistischen Zusammenhangsmaßen Differenzen und Verhältnisse nutzen. Die Differenz der Mittelwerte zweier kontrafaktischer Verteilungen,

$$\Delta E[Y|do(X)] = E[Y|do(X=x)] - E[Y|do(X=x')], \tag{1.21}$$

bezeichnet man als *durchschnittlichen Kausaleffekt*[18] von X auf Y. Die Differenz in der Häufigkeit spezifischer Ausprägungen zwischen zwei kontrafaktischen Verteilungen,

$$\Delta Pr[Y=y|do(X)] = Pr[Y=y|do(X=x)] - Pr[Y=y|do(X=x')] \tag{1.22}$$

wird *kausale Risikodifferenz* genannt. Dieser Unterschied wird häufig auch durch das Verhältnis

[17] Zu Ende gedacht bedeutet das: wären derartige Interventionen empirisch möglich, würden wir kein Kausalmodell benötigen, dass Kanten und abwesende Kanten zwischen den Variablen vorgibt bzw. annimmt. Wir könnten schlicht ausprobieren, welche Variablen auf eine Intervention reagieren, und aus dieser und weiteren Interventionen die Kausalbeziehungen zwischen den Variablen (so gut wie annahmefrei) ermitteln.

[18] In der englischsprachigen Literatur spricht man meist vom *average treatment effect* (ATE) oder *average causal effect* (ACE).

$$\Phi Pr[Y = y|do(X)] = \frac{Pr[Y = y|do(X = x)]}{Pr[Y = y|do(X = x')]} \tag{1.23}$$

gemessen, das als *kausales Risikoverhältnis* bezeichnet wird. Dies sind nur die am häufigsten genutzen Maße, die sich besonders dann eignen, wenn das Treatment nur wenige diskrete Ausprägungen hat, somit die Auswirkungen nur einer geringen Zahl hypothetischer Interventionen verglichen werden. Für metrische Treatments sind diese Maße zwar auch nutzbar. Für den Vergleich durch die genannten Zusammenhangsmaße muss man sich dann jedoch auf einige wenige Ausprägungen beschränken.

Metrische Treatments und *marginal structural models*

Die (nichtparametrische) Quantifizierung der Effekte metrischer Treatments über einfache Differenzen oder Verhältnisse ist zwar möglich, aber auch umständlich und unübersichtlich, denn theoretisch ist eine unendliche Anzahl von Kontrasten möglich. Man kann sich nun die interessantesten herauspicken oder den Effekt metrischer Treatments *modellieren*. Derartige Modelle werden als *marginal structural models* bezeichnet. Das Modell heißt *structural*, da hier Kausalstrukturen, keine beobachtbaren statistischen Zusammenhänge, modelliert werden. Das heißt, die „abhängige“ Variable auf der linken Seite der Modellgleichung ist die *kontrafaktische* Outcomevariable, die in Abhängigkeit der verschiedenen Ausprägungen des Treatment, also aller möglichen hypothetischen Inverventionen, dargestellt wird. Da lediglich dieser eine, unbedingte Zusammenhang modelliert wird, nennt man das Modell *marginal*. Die allgemeinste, nichtparametrische Formulierung eines marginal structural models lautet damit

$$Pr[Y|do(X)] = f(x, \varepsilon_y). \tag{1.24}$$

Man kann nun entscheiden, welcher Parameter der kontrafaktischen Verteilung modelliert werden soll: das arithmetische Mittel, der Median, die Varianz? Zudem kann eine funktionale Form des Kausalzusammenhangs zwischen X und Y spezifiziert werden, beispielsweise ein additiv linearer Zusammenhang. Entscheidet man sich für die Modellierung des arithmetischen Mittels mit dieser funkionalen Form resultiert das Modell

$$E[Y|do(X)] = \varepsilon_y + \theta x. \tag{1.25}$$

Das durchschnittliche kontrafaktische Outcome $E[Y|do(X = x)]$ ergibt sich als Funktion von den Treatmentausprägungen x und der Summe der Effekte aller weiteren Eltern des Outcomes, ε_y. Der Term ε_y steht dabei für den Wert, den das mittlere kontrafaktische Outcome annimmt, wenn das Treatment den Wert 0 annimmt (wenn also keine Änderung von X stattfindet). Es ist die Konstante des linearen Modells. θ gibt an, um welchen Wert sich $E[Y|do(X = x)]$ verschiebt,

wenn die Ausprägung des Treatments um eine Einheit erhöht wird, wenn also beispielsweise statt der Intervention $do(X = 1)$ die Intervention $do(X = 2)$ erfolgt.[19] Selbstverständlich können durch marginal structural models auch andere funktionale Formen als die lineare dargestellt werden. Auch für dichotome Treatments kann ein marginal structural model aufgestellt werden. Man spricht in diesem Fall von einem *saturierten* Modell, da für jede Ausprägung des Treatments ein Parameter vorhanden ist. ε_y steht weiterhin für das durchschnittliche kontrafaktische Outcome unter $x = 0$, also die (frei wählbare) Referenzausprägung. Die Addition von ε_y und θ_1 ergibt das durchschnittliche kontrafaktische Outcome unter dem alternativen Kausalzustand $x = 1$. θ steht damit für den durchschnittlichen Kausaleffekt $\Delta E[Y|do(X)]$ aus (1.21).

Effektheterogenität, Effektmoderation und bedingte kausale Effekte

Bisher haben wir ausschließlich (durchschnittliche) kausale Effekte für die gesamte Zielpopulation definiert. Diese Definition schließt jedoch nicht aus, dass kausale Effekte zwischen den Beobachtungen variieren. So ist durchaus möglich, dass ein Treatment für manche Beobachtungen eine positive, auf andere eine negative und auf wieder andere keine Wirkung hat. Auf diese Weise können sich auch systematische Unterschiede in der Stärke des Effekts zwischen durch bestimmte Variablen definierte Teilpopulationen ergeben. Die eine Gruppe profitiert eher vom Treatment, der anderen schadet es eher. Häufig interessieren wir uns explizit für derartige Fragen. Unterschiede im kausalen Effekt von X auf Y nach den Ausprägungen einer dritten Variable W (oder einer Merkmalskombination $\mathbf{W}$) bezeichnet man als Effekt*moderation* oder auch Effekt*modifikation*. Die Variable, nach deren Ausprägungen sich der Effekt unterscheidet, wird dementsprechend Effektmoderator bzw. Effektmodifikator genannt. Die für die durch den Moderator definierten Teilpopulationen spezifischen Effekte werden *bedingte* oder *konditionale* kausale Effekte genannt.

Definition 1.20. (*Effektmoderation*) Formal definieren lässt sich Effektmoderation damit als Unterschied in der kontrafaktischen Verteilung von Y, $Pr[Y|do(X = x)]$. nach mindestens zwei Ausprägungen einer dritten Variable W:

$$Pr(Y|do(X), W = w) \neq Pr(Y|do(X), W = w'). \tag{1.26}$$

Die Auswirkungen einer Intervention auf X für Y unterscheiden sich also für mindestens zwei Teilpopulationen, die sich an Hand des Merkmals W trennen lassen.

---

[19] Auch wenn dieses Modell einer linearen Regressionsgleichung der Form

$$E[Y|X] = \alpha + \beta x$$

trügerisch ähnlich sieht, ist es davon dennoch fundamental verschieden. Eine Regressionsgleichung modelliert beobachtete statistische Zusammenhänge in vorhandenen Daten. Ein marginal structural model modelliert unbeobachtbare Kausalzusammenhänge.

Beispiel 1.21. So ist das Geschlecht der Teilnehmer ein Effektmoderator, wenn sich die Änderung der Anwesenheitspflicht von 75% auf 100% für Männer anders auf die Verteilung der Punktzahl auswirken würde als für Frauen.

Effektmoderation in DAGs

Wie bereits in Kapital 1.2.1 beschrieben, können alle im DAG dargestellten Kausaleffekte individuell variieren. Effektmoderatoren können dabei generell alle Variablen sein, die selbst vom Treatment nicht beeinflusst werden, ihrerseits aber einen Effekt auf das interessierende Outcome haben. Das heißt, dass in 1.6 der Effekt der Anwesenheit auf Punkte nach Motivation variieren kann. Ebenso denkbar ist, dass sich der Effekt nach weiteren Ursachen der Punkte unterscheidet, die selbst nicht auf Anwesenheit wirken. Diese Information ist allerdings in einem Standard-DAG nicht explizit dargestellt und muss bei Bedarf verbal ergänzt werden. Steht Effektmodifikation bei einer Forschungsfrage im Zentrum, besteht die Möglichkeit die theoretisch vermutete Struktur dieser in separaten DAGs für einzelne Ausprägungen des Effektmodifikators darzustellen.

Beispiel 1.22. Wird vermutet, dass sich Anwesenheit nur für Männer auf die Punkte in der Klausur auswirkt, nicht aber für Frauen, können zwei separate DAGs erstellt werden. Der DAG für Männer enthält dabei einen oder mehrere Kausalpfade von Anwesenheit zu Punkte. Jener für Frauen enthält dementsprechend keine solchen Kausalpfade.

Alternativ besteht die Möglichkeit, Effektmoderation durch Strukturgleichungen darzustellen. In einem linearen marginal structural model kann dazu beispielsweise der Effektmoderator sowie das Produkt aus Effektmoderator und Treatment auf der rechten Seite der Gleichung aufgenommen werden:

$$E[Y|do(X)] = \varepsilon_y + \theta_1 x + \theta_2 w + \theta_3 x w. \tag{1.27}$$

Der Parameter θ_3 gibt hierbei Auskunft darüber, wie stark sich der Effekt von X auf Y nach den einzelnen Ausprägungen von W unterscheidet.

Effektmodifikation vs. Interaktion mehrerer Treatments

Von Effektherogenität und Effektmoderation ausdrücklich zu unterscheiden ist die *Interaktion* von zwei (oder mehr) expliziten Treatments. Einer solchen Interaktion liegt eine Kombination von zwei (oder mehr) hypothetischen Interventionen zu Grunde. In diesem Fall schreiben wir also

$$Pr(Y|do(X = x, A = a)) \neq Pr(Y|do(X = x, A = a')). \tag{1.28}$$

Im Unterschied zur Effektmoderation, die darüber Auskunft gibt, ob sich ein Kausalzusammenhang für bestimmte Ausprägungen einer Variable W unterscheidet,

besagt das Vorliegen einer kausalen Interaktion beispielsweise, dass ein Treatment $X = x$ erst dann Wirkung zeigt, wenn auch die Variable A auf einen bestimmten Wert *a festgesetzt* wird. Auf Grund der zusätzlichen Komplexität derartiger Inferenz, werden wir kausale Interaktionen in der Veranstaltung nicht behandeln.[20]

Beispiel 1.23. Es ist vorstellbar, dass eine Intervention auf die Anwesenheit erst dann vorteilhaft für die Klausurleistung ist, wenn gleichzeitig dafür gesorgt wird, dass die *Qualität* der Veranstaltung ein gewisses Mindestmaß nicht unterschreitet. Dies könnte beispielsweise durch eine Intervention auf die didaktischen Fähigkeiten der Dozenten erreicht werden.

1.2.3 Kausale Inferenz mit randomisierten Experimenten

Die Definition von Kausalität als Folgen *hypothetischer* Interventionen verdeutlicht, dass empirische Daten die Informationen, die wir für kausale Inferenz benötigen, nicht direkt enthalten. Denn eine Intervention, wie sie oben beschrieben wurde, ist nicht durchführbar. Erst recht nicht der Vergleich von zwei oder mehr alternativen Interventionen, da eine Population nicht zwei unterschiedlichen Kausalzuständen gleichzeitig ausgesetzt werden kann. Empirische Daten enthalten stets die Informationen über das Ergebnis des einen *Datengenerierungsprozesses*, der tatsächlich stattgefunden hat. Korrelation zwischen den interessierenden Variablen in diesen Daten sagt damit nichts über die interessierenden hypothetischen Szenarien aus. Allerdings ist es möglich, hypothetische Interventionen mehr oder weniger direkt mit empirischen Daten zu *simulieren*. Gelingt diese Simulation, können statistische Zusammenhänge als Kausalzusammenhänge *interpretiert* werden. Vor einer genaueren Erläuterung dieser Simulation, sei jedoch der Unterschied zwischen Kausalität und statistischem Zusammenhang noch einmal veranschaulicht.

Kausalität vs. Korrelation

Abb. 1.7 zeigt eine Population mit sieben Beobachtungseinheiten[21]. Für diese Population wird eine Variable beobachtet, für deren Effekt auf eine beliebige Variable Y wir uns interessieren. Der Übersichtlichkeit halber hat diese Variable lediglich zwei Ausprägungen, schwarz (s) und grau (g). Um den kausalen Effekt dieser Variable direkt zu beobachten, müssten wir *allen* Mitgliedern der Population die Ausprägung s und die Ausprägung g zuweisen und anschließend Unterschiede in der jeweils entstandenen Verteilung von Y untersuchen. Ein Vergleich der arith-

[20] Einführungen in kausale Interaktion finden sich bei Hernán und Robins (2016, Kap. 5) sowie in den Büchern von VanderWeele (2015) und Hong (2015).

[21] Die geringe Fallzahl dient der Übersichtlichkeit. Gerne können Sie sich vorstellen, dass jede Figur für 1000 Beobachtungen steht.

metischen Mittel dieser Verteilungen liefert den durchschnittlichen kausalen Effekt. Faktisch sind wir jedoch lediglich in der Lage, die *Teilgruppen* der Population zu vergleichen, für die unterschiedliche Ausprägungen des Treatments *beobachtet* wurden, also vier Einheiten mit der Ausprägung s und drei Einheiten mit der Ausprägung g. Wie deren Y unter der jeweils anderen Bedingung aussähe, wissen wir nicht. Anders als beim Vergleich der gesamten Population unter alternativen Kausalzuständen (bzw. nach alternativen hypothetischen Interventionen) besteht beim Vergleich der tatsächlichen Ausprägung des Treatments definierten Teilpopulationen die Möglichkeit, dass sich diese auch hinsichtlich weiterer Merkmale als dem Treatment unterscheiden. Unterschiede in der Verteilung des Outcomes zwischen den Gruppen können damit auch (oder zusätzlich) von diesen Merkmalen herrühren. Ohne weiteres Wissen oder Annahmen über das Zustandekommen der Daten, also wie die einzelnen Variablen ihre Ausprägungen erhalten haben, gibt es damit keinen Grund zu glauben, dass der beobachtete statistische Zusammenhang (also die Unterschiede in Y nach X) dem kausalen Effekt (also Unterschieden in Y nach $do(X)$) gleicht.

Randomisierte Experimente als Simulation hypothetischer Interventionen

Um statistische Zusammenhänge als kausale Effekte interpretieren zu können, ist es nötig, die nicht durchführbare hypothetische Intervention zumindest zu simulieren. Die beste und zugleich naheliegendste Simulation ist dabei eine tatsächlicher Eingriff in den Datengenerierungsprozesses durch ein kontrolliertes Experiment, bei dem einzelnen Ausprägungen des Treatments bestimmte Teilgruppen der Gesamtpopulation zugewiesen werden. Zentral dabei ist, dass bei Zuweisung in sämtlichen dieser Gruppen die Verteilungen aller Variablen (außer natürlich dem Treatment) mit jenen in der Gesamtpopulation identisch sind. Dies lässt sich am einfachsten durch eine Zufallsaufteilung der Beobachtungseinheiten auf die einzelnen

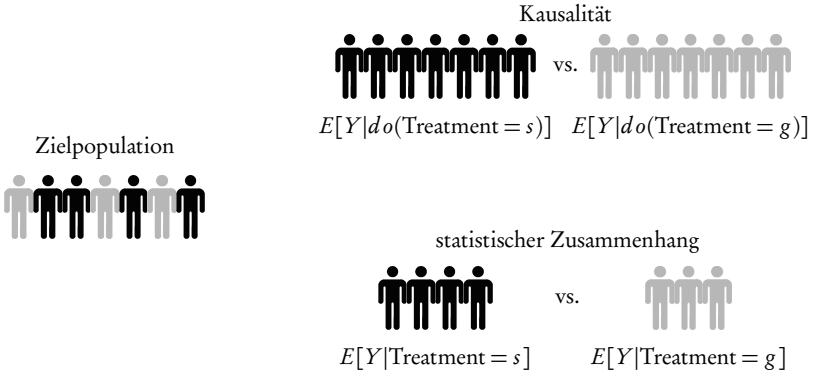


Abb. 1.8 Unterschied zwischen Kausalität und statistischem Zusammenhang (nach Hernán und Robins, 2016, Abb. 1.1.)

Teilgruppen erreichen, der sogenannten Randomisierung. Man erstellt also für jede Ausprägung des Treatments eine Zufallsstichprobe der Population. Dadurch kann es außer dem Treatment keine systematischen Unterschiede zwischen den Gruppen geben, denn alle Variablen, egal ob tatsächlich beobachtet oder unbeobachtet werden gleichmäßig über alle Gruppen verteilt. Folglich sind Unterschiede im Outcome zwischen den Vergleichsgruppen (bis auf Zufallsfehler) ausschließlich auf das Treatment zurückführbar. Man sagt deswegen auch, dass alle so erstellten Gruppen gegeneinander austauschbar sind. Es ist für das Ergebnis belanglos, welche Gruppe welcher Treatmentausprägung zugewiesen wird.

Abb. 1.9 zeigt die Kausalstruktur eines solchen randomisierten Experiments. Diese Struktur ist jener einer hypothetischen Intervention sehr ähnlich, aber nicht identisch. Denn es gibt weiterhin eine Ursache des Treatments, nämlich den gewählten Zufallsmechanismus der Zuweisung. Bei der hypothetischen Intervention hingegen wird die Ausprägung des Treatments ganz unzufällig auf einen bestimmten Wert gesetzt. Nichtsdestotrotz werden durch die Randomisierung die Einflüsse *aller anderen* „natürlichen" Ursachen des Treatments ausgeschaltet, repräsentiert durch die Abwesenheit zusätzlicher gerichteter Kanten auf Anwesenheit.

Funktioniert die Randomisierung, kann die kontrafaktische Outcomeverteilung als Funktion der beobachteten Outcomeverteilung dargestellt werden und der beobachtbare statistische Zusammenhang zwischen Treatment und Outcome entspricht tatsächlich dem kausalen Effekt:

$$Pr(Y|do(X=x)) = Pr(Y|X=x) \tag{1.29}$$

$$E[Y|do(X=x)] - E[Y|do(X=x')] = E[Y|X=x] - E[Y|X=x'] \tag{1.30}$$

Deswegen werden *ideale* randomisierte Experimente auch als der Goldstandard kausaler Inferenz bezeichnet. Einen einfacher Vergleich der Mittelwerte von Y nach den Ausprägungen von X liefert in Daten randomisierter Experimente den durchschnittlichen kausalen Effekt von X auf Y. Leider sind randomisierte Experimente *in der Praxis* mit spezifischen Problemen behaftet, die sie mehr oder weniger weit vom beschriebenen Ideal abweichen lassen und die Analyse und In-

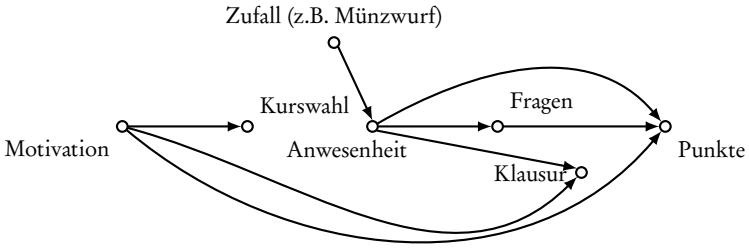


Abb. 1.9 Kausalstruktur eines idealen randomisierten Experiments mit Zuweisung der Anwesenheit

terpretation damit verkomplizieren (siehe Kapitel **??**). Zudem ist die Durchführung randomisierter Experimente meist mit großen logistischen Problemen und ethischen Bedenken verbunden. Obwohl sich die Zahl experimenteller Untersuchungen in den Sozialwissenschaften in den vergangenen Jahren stark erhöht hat und auch weiterhin ein Potential zum vermehrten Einsatz von Experimenten besteht, wird der Großteil sozialwissenschaftlicher Untersuchungen deswegen wohl auch in Zukunft auf nichtexperimentellen Daten oder, genauer gesagt, auf Designs ohne randomisierte Treatmentzuweisung beruhen. Da die Teilnehmer nichtexperimenteller Untersuchungen in der Regel zumindest teilweise die Ausprägung der Treatmentvariable selbst beeinflussen können, ist die Identifikation kausaler Effekte mit derartigen Daten komplizierter.

Bedingung für das Gelingen kausaler Inferenz bleibt dabei jedoch, mit den Daten eine hypothetische Intervention so gut wie möglich zu simulieren. Ohne die Möglichkeit randomisierter Treatmentzuweisung ist dies jedoch nur durch die Kombination hochwertiger Daten mit einer detaillierten Theorie bezüglich des Datengenerierungsprozesses machbar. An Hand dieser Theorie kann überprüft werden, welche zusätzlichen Variablen für die Analyse herangezogen werden müssen. Eine geläufige Strategie hierbei ist durch die statitistische Kontrolle von Drittvariablen eine *Quasi*-Randomisierung des Treatments (zumindest für eine Teilpopulation) herzustellen.

1.2.4 Ein „Fahrplan“ für kausale Inferenz ohne Randomisierung

Um kausale Inferenz ohne randomisierte Treatmentzuweisung zu systematisieren und die einzelnen Analyseschritte transparent zu gestalten, ist es ratsam, einem festen Algorithmus zu folgen. Ein überzeugender Vorschlag für einen solchen Algorithmus kommt von Petersen und van der Laan (2014). Dieser dient gleichzeitig als Orientierung für die weiteren Kapitel, in denen die wichtigsten Werkzeuge für die Umsetzung des Fahrplans genauer vorgestellt werden.

1. *Spezifikation des Datengenerierungsprozesses:* Zunächst wird das vorhandene Wissen zum Datengenerierungsprozess dargelegt und, wenn nötig, durch theoretische Annahmen ergänzt. Bei konkurrierenden Theorien ist es auch möglich, mehrere alternative Datengenerierungsprozesse darzustellen. Eine rigorose und gleichsam transparente Möglichkeit zur Darstellung von Datengenerierungsprozessen bieten die bereits vorgestellten graphischen Kausalmodelle.
2. *Kennzeichnung der in den empirischen Daten vorhandenen Variablen:* Zusätzlich zum Datengenerierungsprozess sollte verdeutlicht werden, welche der an der Datengenerierung beteiligten Variablen tatsächlich in den erhobenen bzw. zu erhebenden Daten enthalten sind und welche unbeobachtet bleiben.
3. *Definition des interessierenden Kausaleffekts inklusive Zielpopulation:* Die jeweilige Kausalfrage sollte in einen genau definierten Kausaleffekt übertragen werden. Dazu gehört, ob der Effekt eines einmaligen Treatments untersucht wird

oder ob sich das Treatment über die Zeit verändert. Auch die genaue hypothetische Intervention, die von Interesse ist, sollte explizit gemacht werden. Wird beispielsweise jeder Untersuchungseinheit die gleiche Treatmentausprägung zugewiesen oder wird das arithmetische Mittel der Treatmentverteilung verschoben. Zudem sollte geklärt werden, für welchen Parameter der kontrafaktischen Outcomeverteilung man sich genau interessiert; für das arithmetische Mittel, den Median, die Varianz? Schließlich sollte auch die Zielpopulation, für die der Effekt untersucht werden soll, genau definiert werden. Hierzu gehört auch, ob Unterschiede im kausalen Effekt zwischen verschiedenen Gruppen (also Effektmoderation) von Interesse sind. In der Veranstaltung selbst werden wir uns auf die Effekte einmaliger Treatments beschränken und lediglich Interventionen betrachten, bei denen allen Beobachtungseinheiten der Population die gleiche Treatmentausprägung zugewiesen wird.

4. *Überprüfung der Identifizierbarkeit dieses Effekts:* Mithilfe verhältnismäßig einfacher Regeln (siehe Kapitel 2) kann dann überprüft werden, ob der interessierende Kausaleffekt bei der gegebenen Kausalstruktur und den vorhandenen empirischen Daten identifiziert werden kann. Ist dies der Fall, kann die kausale Inferenz zum statistischen Teil übergehen. Wenn der interessierende Kausaleffekt nicht identifizierbar ist, gibt es vier Möglichkeiten: (a) Erhebung besserer Daten, (b) stärkere (und möglicherweise unplausible) Annahmen zum Datengenerierungsprozess, (c) Wahl eines anderen, besser identifizierbaren Kausaleffekts (beispielsweise nur für eine bestimmte Teilgruppe der eigentlichen Population) oder (d) Fortsetzung der Analyse als statistische Inferenz, deren Ergebnisse nicht mehr kausal interpretiert werden.[22]
5. *Spezifikation des statistischen Modells:* Erst jetzt beginnt der Teil kausaler Inferenz, für den die üblichen statistischen Verfahren von Bedeutung sind (siehe Kapitel 3). Ein identifizierter Kausaleffekt ist durch einen spezifischen beobachtbaren Zusammenhang schätzbar. Die Wahl der Schätzmethode hängt dabei in erster Linie ab vom Skalenniveau der Outcomevariable, aber auch von der Effizienz des resultierenden Schätzers. Es ist klar zu betonen, dass keine statistische Schätzmethode „kausaler“ ist als eine andere. Keine Methode, weder Regression noch Matching, machen aus einem statistischen Zusammenhang einen Kausaleffekt. Die Möglichkeit einer kausalen Interpretation liegt einzig und allein an der Qualität von Vorwissen und Annahmen zum Datengenerierungsprozess sowie dem Untersuchungsdesign und den daraus resultierenden Daten.
6. *Anwendung des Schätzmodells auf die empirischen Daten:* Nach Auswahl einer oder bestenfalls mehrerer alternativer Schätzmethoden wird in Kombination mit den empirischen Daten der interessierende Kausaleffekt geschätzt. Bei der

[22] Hierbei betritt man zügig eine Grauzone zwischen kausaler Inferenz und statistischer Inferenz. Dies äußert sich in der Forschung beispielsweise in der wiederholten Betonung, man schätze natürlich keine Kausaleffekte, nur um in den Schlussfolgerungen die Konsequenzen möglicher Interventionen aus den Ergebnissen abzuleiten. Ein solcher Bereich zwischen kausaler und statistischer Inferenz, in dem man auf fast magische Weise kausale Schlussfolgerungen aus deskriptiven Befunden ableiten kann, existiert jedoch ohne Zweifel nicht.

Analyse von Stichprobendaten kommt noch die Schätzung eines Standardfehlers und daraus sich ergebender Konfidenzintervalle hinzu (siehe Kapitel 3)

7. *Interpretation der Ergebnisse und Darlegung der notwendigen Annahmen:* Schließlich kann der geschätzte Effekt inhaltlich interpretiert werden. Hierbei sollte nicht die statistische Signifikanz im Vordergrund stehen, sondern die tatsächliche Stärke des Effekts. Bei der Interpretation sollten auch noch einmal die zur Identifikation notwendigen Annahmen zum Datengenerierungsprozess explizit gemacht werden. Diese sind danach zu unterscheiden, ob sie auf tatsächlichem Vorwissen beruhen oder ohne gesichertes Wissen getätigt wurden, um die Identifikation zu ermöglichen (*knowledge-based assumptions* vs. *convenience-based assumptions*). Annahmen, die *gegen* vorhandenes Wissen gemacht wurden, berechtigen zu starken Zweifeln an einer kausalen Interpretation der Ergebnisse.

1.3 Zum Weiterlesen

Zunächst ist anzumerken, dass die Einführungsliteratur zu kausaler Inferenz von erheblichen Abweichungen in der genutzten Notation geprägt ist. Dies kann anfangs zu einiger Verwirrung führen und fordert erhöhte Aufmerksamkeit bei der Lektüre, insbesondere, wenn gleichzeitig verschiedene Werke herangezogen werden.

Grundlagen zur Definition kausaler Effekte, zur Unterscheidung von Kausalität und Korrelation und zu graphischen Kausalmodellen finden sich beispielsweise in den entsprechenden Kapiteln bei Hernán und Robins (2016) oder Morgan und Winship (2015).

Eine gleichsam effiziente und verständliche Einführung in hypothetische Interventionen und den do-Operator sowie die besondere Rolle randomisierter Experimente bieten Kapitel 23 und 24 in Shalizi (2016). Ausführlichere Darstellungen dieses Materials, die aber gehobene Vorkenntnisse in mathematische Grundlagen erfordern, liefern die Originalarbeiten von Pearl (1995, 2009b).

Pearl (2009a) begründet überzeugend eine strikte Unterscheidung zwischen statistischen und kausalen Konzepten und die Notwendigkeit einer spezifischen Notation für kausale Zusammenhänge. Sobel (1998, 2000) formuliert eine vernichtende Kritik an der gedankenlosen Nutzung von „Kausalsprache“ in empirischen Analysen.

Zur Vertiefung der Konzepte von Effektheterogenität und Effektmoderation empfehlen sich Hernán und Robins (2016, Kap.4), Xie (2013) und Morgan und Winship (2012). VanderWeele und Robins (2007) sowie VanderWeele (2009) beschäftigen sich speziell mit Effektmoderation in DAGs sowie der Abgrenzung zu Interaktion. Das Thema Heterogenität kausaler Effekte wird auch in folgenden Kapiteln immer wieder aufgegriffen.

Kapitel 2
Identifikation kausaler Effekte

2.1 Ableitung beobachtbarer Zusammenhänge aus DAGs

Der durch den DAG modellierte Datengenerierungsprozess zeigt die *vermutete* Kausalstruktur zwischen den für die Analyse relevanten Variablen. Er ist ein Modell über die Funktionsweise (des interessierenden Ausschnitts) der Welt. Was der DAG *nicht direkt* zeigt, sind beobachtbare statistische Zusammenhänge zwischen den dargestellten Variablen. So kann es durchaus sein, dass für zwei Variablen, zwischen denen sich keine Kante befindet, ein statistischer Zusammenhang beobachtet werden kann. DAGs repräsentieren eben keine statistischen Modelle, sondern (theoretische) Kausalbeziehungen. Über eine kleine Anzahl einfacher Regeln können jedoch *Erwartungen* darüber *abgeleitet* werden, ob zwei Variablen in durch diesen DAG generierten Daten statistisch voneinander abhängig sind oder nicht. DAGs dienen damit als eine Art „Taschenrechner" (Pearl), über den theoretische Kausalstrukturen in beobachtbare statistische Zusammenhänge übersetzbar sind. DAGs sind also gleichzeitig Sprache zur Kommunikation von Kausalmodellen *und* Hypothesengenerator. Empirische Tests dieser Hypothesen können dann genutzt werden, um *neues* Wissen über unbekannte Aspekte dieses Modells zu gewinnen.

2.1.1 Unbedingte und bedingte statistische (Un-)Abhängigkeit

Bevor die Regeln zur Ableitung beobachtbarer Zusammenhänge aus einem Kausalmodell genauer dargestellt werden, seien zunächst zentrale Begriffe und Notation erläutert, die im weiteren Verlauf genutzt werden. Wie in Kapitel 1.1 bereits ausgeführt, sagt man, dass zwei Variablen statistisch voneinander abhängen, wenn sich die Verteilung der Variablen nach den Ausprägungen der jeweils anderen unterscheidet. Kennt man also die Ausprägung der einen Variable, kann man etwas dazu sagen, welche Ausprägung die zweite Variable wahrscheinlich annimmt (und umgekehrt). Ist das nicht der Fall, heißt es, die beiden Variablen sind statistisch

unabhängig. Statistische Unabhängigkeit zwischen den Variablen A und B kann folgendermaßen notiert werden (Dawid, 1979):

$$A \perp\!\!\!\perp B \tag{2.1}$$

Nun kann statistische (Un-)Abhängigkeit zwischen zwei Variablen auch bedingt nach dritten (und weiteren) Variablen untersucht werden. Dazu wird der Zusammenhang zwischen den Variablen getrennt nach den einzelnen Ausprägungen der dritten Variablen(n) betrachtet. Diesen Vorgang, Informationen von dritten Variablen für die Untersuchung statistischer Zusammenhänge heranzuziehen, bezeichnet man als *Konditionierung* (*conditioning*). In DAGs wird die Konditionierung auf einer Variable *hier* dadurch gekennzeichnet, dass um den jeweiligen Knoten ein Kasten (□) gezeichnet wird. Ein Beispiel für (parametrische) Konditionierung ist die Aufnahme von C als Kontrollvariable in eine Regression von B auf A. Ein weiteres Beispiel (diesmal für nichtparametrische Konditionierung) ist die getrennte Berechnung der Regression innerhalb von nach C definierten Subgruppen. Sind die beiden Variablen *gegeben* die dritte Variable voneinander abhängig, spricht man von *bedingter* oder auch *konditionaler* Abhängigkeit zwischen den Variablen. Gibt es *keinen* statistischen Zusammenhang zwischen beiden Variablen, gegeben die dritte Variable, spricht man von *bedingter* oder auch *konditionaler* Unabhängigkeit, die in Kurzform auf diese Weise notiert wird:

$$A \perp\!\!\!\perp B|C \tag{2.2}$$

Zur Abgrenzung spricht man bei *unbedingter* (Un-)Abhängigkeit auch von *marginaler* (Un-)Abhängigkeit.

2.1.2 Elementare Kausalstrukturen und statistische (Un-)Abhängigkeit

Aber wie kann (un)bedingte statistische (Un-)Abhängigkeit zwischen zwei Variablen überhaupt entstehen? In der Tat liegen die Gründe dafür in lediglich drei elementaren Kausalstrukturen, über die zwei Variablen A und B verbunden sein können:

1. Kausalität ($A \rightarrow B$ oder $A \leftarrow B$) oder auch Kausalkette (*causal chain*),
2. gemeinsame Vorfahren V ($A \leftarrow V \rightarrow B$) oder auch Gabel (*fork*),
3. gemeinsame Nachkommen N (also: $A \rightarrow N \leftarrow B$) oder auch Kollision (*collision*),

Statistische Abhängigkeit zwischen zwei Variablen kann nur entstehen, wenn diese auf irgendeine Weise über Pfade verbunden sind. Die Art und Weise, wie sich aus den drei genannten elementaren Pfaden *marginale* statistische (Un-)Abhängigkeit

ergibt, ist analog zu Verwandtschaftsbeziehungen in Familien.[1] So sind Kinder stets mit ihren Eltern verwandt und Eltern mit ihren Kindern. Auch sind die Kinder der gleichen Eltern stets miteinander verwandt. Aber ein gemeinsames Kind führt nicht zu Verwandtschaft zwischen zwei Eltern.

Kausalität

Die erste Kausalstruktur, über die eine statistische Abhängigkeit zwischen zwei Variablen enstehen kann, ist (natürlich!) ein kausaler Pfad zwischen beiden Variablen. Das heißt, entweder hat A einen Effekt auf B oder B einen Effekt auf A. Eine *marginale* statistische Abhängigkeit entsteht hier dadurch, dass eine Änderung der einen Variable eine Änderung der zweiten Variable zur Folge hat. Somit gehen mit unterschiedlichen Ausprägungen der einen Variable auch bestimmte Ausprägungen der anderen Variable typischerweise einher. Die Abhängigkeit entsteht dabei ungeachtet der Richtung, in welche die Kausalität verläuft. Aus diesem Grund gibt die statistische Abhängigkeit *allein* keine Auskunft über die Kausalrichtung. Ohne Wissen über die zeitliche Ordnung hilft daher nur eine Theorie, um zu entscheiden welche Variable die Ursache und welche die Wirkung ist. Marginale statistische Abhängigkeit zwischen A und B entsteht auch, wenn der Effekt über eine dritte Variable M vermittelt wird.

Beispiel 2.1. Zwischen Bildung und Einkommen wird in empirischen Daten eine marginale statistische Abhängigkeit bestehen, wenn Bildung einen kausalen Effekt auf das Einkommen hat (siehe Abb. 2.1a) (oder Einkommen auf die Bildung). Denn dieser kausale Effekt führt dazu, dass für Personen mit unterschiedlicher Bildung (im Mittel) auch ein unterschiedliches Einkommen beobachtet wird. Auch wenn der Effekt (wie beispielsweise von der Humankapitaltheorie postuliert) über die Variable Produktivität vermittelt wird (siehe Abb. 2.1b), besteht der Zusammenhang zwischen Bildung und Einkommen.

In diesem Szenario kann nun *bedingte* statistische *Un*abhängigkeit zwischen A und B erreicht werden, wenn auf die Variable, die sich auf dem Kausalpfad zwischen A und B befindet, konditioniert wird (und diese Variable die einzige ist, die den Effekt vermittelt). Denn außer über diese Variable, deren Ausprägungen nun konstant

Abb. 2.1 Kausalpfad von Bildung zu Einkommen (a. direkter Effekt, b. indirekter Effekt) produziert marginale statistische Abhängigkeit zwischen beiden

[1] Übrigens eine tolle Eselsbrücke für die gleich folgenden Regeln.

gehalten werden, kann A keine Änderung in der Verteilung von B hervorrufen (und umgekehrt). Innerhalb der Ausprägungen von M gibt es keine Abhängigkeit zwischen A und B.

Beispiel 2.2. Konditioniert man auf Produktivität, die einzige Variable im Modell (siehe Abb. 2.2), über die Bildung Einkommen beeinflust, besteht kein statistischer Zusammenhang mehr zwischen Bildung und Einkommen:

$$\text{Bildung} \perp\!\!\!\perp \text{Einkommen}|\text{Produktivität} \tag{2.3}$$

Kenne ich die Ausprägung von Produktivität, liefert mir Bildung keine zusätzlichen Informationen über die wahrscheinliche Ausprägung von Einkommen. Steht bei der gegebenen Kausalstruktur der (Gesamt-)Effekt von Bildung auf Einkommen im Mittelpunkt, sollte daher eine Konditionierung auf Produktivität, also deren „Kontrolle“ in der Analyse, unterbleiben.

Gemeinsamer Vorfahre (Gabel)

Statistische Abhängigkeit kann auch durch das Vorhandensein eines *nicht*kausalen Pfades zwischen zwei Variablen entstehen. Haben A und B einen *gemeinsamen* Vorfahren V, sind sie voneinander *marginal* abhängig, auch wenn *kein* Kausalpfad zwischen ihnen besteht.[2] Das liegt daran, dass eine Änderung von V sowohl zu einer Änderung in A als auch in B führt. Ohne V heranzuziehen (d.h. ohne Konditionierung auf V), enthält A damit Informationen über B und umgekehrt.

Beispiel 2.3. Abb. 2.3 zeigt eine hypothetische Kausalstruktur, in der Intelligenz sowohl die erreichte Bildung als auch das erzielte Einkommen verursacht. Hier hat Bildung nun keinen Effekt auf Einkommen (und Einkommen auch keinen Effekt auf Bildung). In empirischen Daten wäre unter diesem Szenario eine statistische Abhängigkeit zwischen Bildung und Einkommen beobachtbar. Führt beispielsweise höhere Intelligenz zu mehr Bildung und Einkommen, werden Personen mit

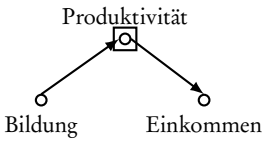


Abb. 2.2 Konditionierung auf Produktivität führt zu statistischer Unabhängigkeit zwischen Bildung und Einkommen

[2] Dieses Phänomen wird häufig als *Scheinkorrelation* bezeichnet. Der Begriff ist jedoch irreführend, da die Variablen nicht *scheinbar*, sondern *ganz real* korrelieren bzw. statistisch voneinander abhängen. Passender wäre Scheinkausalität. Aber auch dieser Begriff ist nicht ideal, da er impliziert, dass von vornherein davon ausgegangen wird, dass Korrelation eigentlich ein Hinweis für Kausalität sein sollte. Hier wird die Verwirrung deutlich, die kennzeichnend für die traditionelle Beschäftigung mit Kausalzusammenhängen ist. Aus diesem Grund werden wir den Begriff Scheinkorrelation im Folgenden nicht gebrauchen.

höherer Bildung (im Mittel) mehr verdienen, auch wenn die Bildung selbst keinen Effekt auf das Einkommen hat (eine Intervention auf Bildung also keine Änderung in Einkommen zur Folge hätte).

Eine Konditionierung auf den gemeinsamen Vorfahren V wird in diesem Szenario bedingte statistische Unabhängigkeit zwischen A und B herstellen. Ohne weiteren Pfad zwischen A und B liefert A innerhalb der Ausprägungen von V keine Informationen mehr zu B und umgekehrt. Denn alle Informationen, die A über B besitzt, sind bereits in V enthalten (Und? Umgekehrt!).

Beispiel 2.4. Gegeben der Intelligenz besteht unter der in Abb. 2.4 dargestellten Kausalstruktur kein statistischer Zusammenhang zwischen Bildung und Einkommen:

$$\text{Bildung} \perp\!\!\!\perp \text{Einkommen} | \text{Intelligenz} \tag{2.4}$$

Das Modell zeigt an, dass ohne eine Änderung in Intelligenz, kein Zusammenhang zwischen Bildung und Einkommen besteht. Ist man also am (in diesem Beispiel nicht vorhandenen) kausalen Effekt von Bildung auf Einkommen interessiert, sollte man Informationen über Intelligenz bei der Analyse heranziehen.

Gemeinsamer Nachkomme (Kollision)

Die dritte elementare Kausalstruktur, über die zwei Variablen A und B verbunden sein können, ist die Existenz eines gemeinsamen Nachkommen, N. Über *diesen* nichtkausalen Pfad zwischen A und B kommt allerdings *keine* marginale statistische Abhängigkeit der beiden zustande. Zwar enthält der gemeinsame Nachkomme Informationen von A und B. Aber die Eltern enthalten keine Informationen voneinander. Denn eine Änderung in einem Elternteil geht zwar mit einer Änderung im Nachkommen, aber nicht mit einer Änderung im anderen Elternteil einher. Der gemeinsame Nachkomme unterbricht quasi den Fluss statistischer Abhängigkeit. Variablen, an denen zwei gerichtete Kanten aufeinanderprallen, werden auch als *collider* bezeichnet.

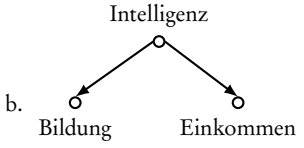


Abb. 2.3 Gemeinsamer Vorfahre von Bildung und Einkommen, Intelligenz, produziert marginale statistische Abhängigkeit

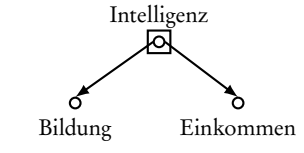


Abb. 2.4 Konditionierung auf Intelligenz führt zu statistischer Unabhängigkeit zwischen Bildung und Einkommen

Beispiel 2.5. So ist denkbar, dass sowohl Bildung als auch Einkommen, die Teilnahme an sozialwissenschaftlichen Umfragen beeinflussen. In Abwesenheit anderer Pfade zwischen beiden Variablen (siehe Abb. 2.5) bedeutet dieser Umstand jedoch nicht, dass eine statistische Abhängigkeit zwischen Bildung und Einkommen besteht. Im Gegenteil: Bildung und Einkommen wären in durch diese Kausalstruktur generierten Daten marginal unabhängig:

$$\text{Bildung} \perp \text{Einkommen} \tag{2.5}$$

Wird nun aber (freiwillig oder unfreiwillig) auf den gemeinsamen Nachfahren konditioniert, kommt es zu *bedingter* statistischer Abhängigkeit zwischen A und B. Dies mag auf den ersten Blick verwundern, ist aber an einem konkreten Beispiel leicht nachvollziehbar.

Beispiel 2.6. Nimmt man die Kausalstruktur aus Abb. 2.5 zum Ausgangspunkt und konditioniert in durch diese Struktur generierten Daten auf Surveyteilnahme wie in Abb. 2.6 (was stets bereits schon dadurch passiert, dass Daten lediglich für die Gruppe der Surveyteilnehmer erhoben werden), wird eine statistische Abhängigkeit zwischen Bildung und Einkommen entstehen, auch wenn anderweitig keine Verbindung zwischen den beiden Variablen vorliegt. Nehmen wir an, dass die Bereitschaft zur Teilnahme mit zunehmender Bildung steigt, aber mit zunehmendem Einkommen fällt. Daraus folgt, dass für eine beliebige Person in der Gruppe der Surveyteilnehmer mit hoher Bildung ein eher hohes Einkommen einhergeht. Somit wird man einen positiven Zusammenhang zwischen Bildung und Einkommen innerhalb dieser Gruppe beobachten können, obwohl es diesen in der Gesamtbevölkerung nicht gibt und obwohl es keinen kausalen Zusammenhang zwischen Bildung und Einkommen gibt. Ist der Test auf einen (in diesem Beispiel ebenso nicht vorhandenen) Kausaleffekt von Bildung auf Einkommen von Interesse, sollte die Konditionierung auf Surveyteilnahme vermieden werden (was natürlich in dieser Frage in der Realität schwierig ist, da niemand zur Teilnahme gezwungen werden kann).

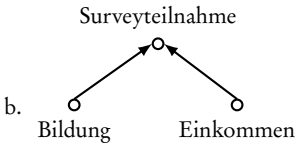


Abb. 2.5 Gemeinsamer Nachkomme Surveyteilnahme sorgt für marginale statistische Unabhängigkeit zwischen Bildung und Einkommen

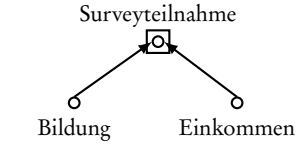


Abb. 2.6 Konditionierung auf dem gemeinsamen Nachkommen von Bildung und Einkommen, Surveyteilnahme, produziert bedingte statistische Abhängigkeit

Beispiel 2.7. Zum besseren Verständnis ein weiteres Beispiel: Abb. 2.7 zeigt Daten, in denen Motivation und Leistung in der Population statistisch unabhängig sind, da sie einander weder kausal beeinflussen noch gemeinsame Vorfahren haben. Beide beeinflussen aber den Besuch eines Elite-Colleges in den USA positiv. Denn um in ein solches College aufgenommen zu werden, muss man entweder sehr motiviert, sehr leistungsstark oder auf beiden zumindest etwas überdurchschnittlich sein. Damit sind Personen auf einem Elite-College, die sehr motiviert sind, im Mittel eher weniger leistungsstark. Dies äußert sich in einem negativen statistischen Zusammenhang zwischen beiden Variablen, wenn man diesen ausschließlich für Personen auf einem Elite-College untersucht (also auf der Variable Besuch eines Elite-College konditioniert), obwohl zwischen beiden Variablen weder in der Gesamtpopulation noch in einer der beiden Teilgruppen ein Kausalzusammenhang besteht.

2.1.3 D-separation

Jeder DAG besteht aus einer Kombination der oben beschriebenen elementaren Kausalstrukturen. Aus den geschilderten Gründen für statistische (Un-)Abhängigkeit, ergeben sich nun allgemeine Regeln, *unter welchen Bedingungen* zwei beliebige Variablen im Kausalmodell voneinander statistisch (un)abhängig sind. Kern dieser Regeln ist die Feststellung, dass *jegliche* statistische Abhängigkeit über Pfade zwischen Variablen transportiert wird, dass aber *nicht alle* Pfade statistische Abhängigkeit leiten. Und zwar transportiert ein gegebener Pfad zwischen zwei Variablen genau dann *keine* statistische Abhängigkeit, wenn dieser Pfad

1. mindestens einen *Nicht-collider* enthält, auf den *konditioniert* wird, oder

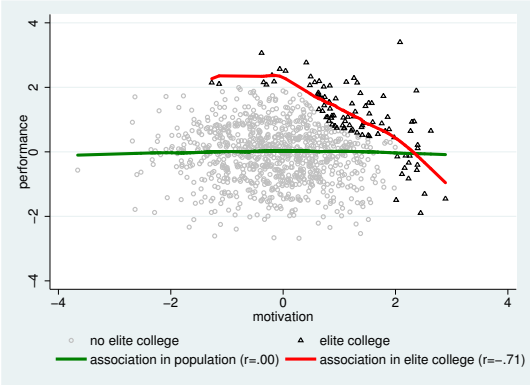


Abb. 2.7 Konditionierung auf Elite-College führt zu statistischer Abhängigkeit zwischen Motivation und Leistung (nach Morgan und Winship, 2015, S.108)

2. mindestens einen *collider* enthält und *weder* auf diesen *collider noch* auf einen seiner Nachkommen konditioniert wird.[3]

Transportiert ein Pfad keine statistische Abhängigkeit sagt man, der Pfad ist *d-separated*[4], blockiert oder geschlossen.

Als *d-connected*, unblockiert oder offen wird hingegen ein Pfad bezeichnet, der

1. weder *collider* noch Nicht-*collider* enthält (also zwei Variablen direkt verbindet),
2. ausschließlich Nicht-*collider* enthält, auf die nicht konditioniert wird,
3. ausschließlich Nicht-*collider*, auf die nicht konditioniert wird, und *collider* enthält, auf die selbst oder auf deren Nachkommen konditioniert wird, oder
4. ausschließlich *collider* enthält, auf die selbst oder auf deren Nachkommen konditioniert wird.

Ein solcher Pfad überträgt statistische Abhängigkeit.

Bestehen zwischen zwei Variablen in einem DAG ausschließlich geschlossene Pfade, sagt man diese Variablen sind *d-separated*. Diese Variablen sind statistisch unabhängig. Gibt es *mindestens einen Pfad* zwischen zwei Variablen, der offen ist, sind die beiden Variablen *d-connected*. Zwischen diesen Variablen besteht statistische Abhängigkeit. Alle Variablen, deren Konditionierung für *d-separation* bzw. *d-connection* zwischen zwei Variablen sorgen, sind die Bedingungen statistischer Unabhängigkeit bzw. Abhängigkeit zwischen diesen Variablen.

Beispiel 2.8. Mithilfe dieser Regeln können wir nun ableiten, unter welchen Bedingungen die Variablen des Modells in Abb. 1.5 in durch dieses Modell generierten Daten voneinander statistisch (un)abhängig sind. Die aus diesem Modell ableitbaren, bedingten statistischen Unabhängigkeiten sind:

$$\text{Anwesenheit} \perp \text{Kurswahl} \mid \text{Motivation} \tag{2.6}$$
$$\text{Punkte} \perp \text{Klausur} \mid \text{Anwesenheit, Motivation} \tag{2.7}$$
$$\text{Punkte} \perp \text{Kurswahl} \mid \text{Motivation} \tag{2.8}$$
$$\text{Fragen} \perp \text{Klausur} \mid \text{Anwesenheit} \tag{2.9}$$
$$\text{Fragen} \perp \text{Kurswahl} \mid \text{Motivation} \tag{2.10}$$
$$\text{Fragen} \perp \text{Kurswahl} \mid \text{Anwesenheit} \tag{2.11}$$
$$\text{Fragen} \perp \text{Motivation} \mid \text{Anwesenheit} \tag{2.12}$$
$$\text{Klausur} \perp \text{Kurswahl} \mid \text{Motivation} \tag{2.13}$$

[3] Die Kontrolle des Nachkommens eines *colliders* produziert statistische Abhängigkeit zwischen den Eltern des *colliders*, da er auch Informationen über beide Elternteile enthält. Damit ein Pfad keine statistische Abhängigkeit überträgt, darf also auch auf keinem Nachkommen eines *colliders* auf diesem Pfad konditioniert werden. Der Nachkomme agiert hier quasi als Proxy des eigentlichen *colliders*.

[4] Das *d* steht hier für *directed*. *D-separation* bedeutet nichts anderes als der Umstand, dass zwei Knoten in DAGs (gegeben dritter Variablen) voneinander abgeschnitten, also separiert, sind und somit keine statistische Abhängigkeit zwischen ihnen übertragen wird.

Alle anderen Variablen im Modell sind unter jeder der jeweils weiteren möglichen Bedingungen voneinander statistisch abhängig.

Durch die der d-separation zugrunde liegenden Regeln, können aus den im DAG dargestellten *theoretischen* Kausalbeziehungen, beobachtbare statistische Zusammenhänge abgeleitet werden. Tests dieser Implikation anhand empirischer Daten können dann Aufschluss über die Validität des Modells geben.

Es muss noch erwähnt werden, dass die Ableitung beobachtbarer Zusammenhänge aus Kausalstrukturen über d-separation auf drei Annahmen beruht:

1. der kausalen Markow-Bedingung,
2. *faithfulness*,
3. keine Zufallsfehler

Die kausale Markow-Bedingung (KMB) besagt, dass jede Variable, *nach Konditionierung auf ihren Eltern*, von allen Variablen im Modell unabhängig ist, außer von ihren eigenen Nachfahren. Die obigen Regeln ergeben sich quasi als direkte Folge der KMB. Die bereits in Kapitel 1 getätigte Forderung, dass ein kausaler DAG alle gemeinsamen Vorfahren jedes bereits dargestellten Variablenpaares enthalten muss, resultiert ebenso direkt aus der KMB.

Faithfulness hingegen beschreibt die Annahme, sich zwei kausale Effekte im Modell niemals gegeneinander aufheben dürfen. Sie ist notwendig, damit aus einem DAG auch statistische Abhängigkeit und nicht nur Unabhängigkeit vorhergesagt werden kann. Denn ist es beispielsweise möglich, dass sich die Effekte eines gemeinsamen Vorfahren auf zwei ihrer Kinder gegenseitig aufheben, wird man in Daten keinen statistischen Zusammenhang beobachten, obwohl die beiden Kinder einen gemeinsamen Vorfahren haben. Das gleiche gilt für eine Struktur wie sie in Abb. 2.1b dargestellt ist. Wiegen sich der Effekt von Bildung auf Produktivität und der von Produktivität auf Einkommen perfekt gegeneinander auf, wäre auch ohne Konditionierung auf Produktivität bereits keine statistische Abhängigkeit zwischen Bildung und Einkommen mehr zu beobachten. *Faithfulness* legt per Annahme fest, dass solche Szenarien der perfekten gegenseitigen Aufhebung nicht vorkommen.

Schließlich muss (per Annahme) ausgeschlossen werden, dass zwei Variablen in empirischen Daten aus purem Zufall voneinander abhängig sind. Das passiert hier (vorläufig) dadurch, dass wir annehmen, über Daten mit unendlich großer Fallzahl zu verfügen, in denen der Zufall keine nennenswerte Rolle mehr spielt. In Kapitel 3 geben wir dies dann auf und nutzen die üblichen Techniken *statistischer* Inferenz (also Standardfehler, Konfidenzintervalle), um die Wahrscheinlichkeit des Auftretens von Zufallsfehlern zu quantifizieren.

2.1.4 Empirische Modelltests und konkurrierende DAGs

Durch die Regeln der *d-separation* kann ermittelt werden, unter welchen Bedingungen ein beliebiges Variablenpaar des gegebenen Modells in durch dieses Mo-

dell generierten Daten statistisch (un)abgängig sein sollten. Verfügt man über Daten, die sämtliche Variablen des Modells enthalten, können die tatsächlichen (Un)abhängigkeiten berechnet und das Modell somit getestet werden. Ein Test kann der jeweiligen Vorhersage dabei auf zwei verschiedene Arten widersprechen:

1. Eine Abhängigkeit zwischen Variablen wird gefunden, die nicht vorhergesagt wurde.
2. Eine vorhergesagte Abhängigkeit wird nicht gefunden.

Beide Widersprüche bedeuten, dass das Kausalmodell falsch sein muss. Denn ein wahres Modell kann keine falschen Vorhersagen machen. Dabei bezeichnet man den ersten Widerspruch als starken Widerspruch, da dieser auch unter Aufgabe von *faithfulness* noch bedeutet, dass das Modell auf jeden Fall falsch ist. Gibt man *faithfulness* auf, ist Widerspruch 2 hingegen weiterhin konsistent mit dem Modell.

Beispiel 2.9. Aus dem DAG in Abb. 1.5 lässt sich eine statistische Unabhängigkeit von Anwesenheit und Kurswahl gegeben Motivation ableiten. Findet man allerdings statt dieser Unabhängigkeit eine Abhängigkeit, bedeudet das, das Modell ist falsch. Denn die statistische Abhängigkeit impliziert offene Pfade zwischen Anwesenheit und Kurswahl, die nicht über Motivation verlaufen. Zwischen Anwesenheit und Punkte wird hingegen (auch gegeben allen anderen Variablen im Modell) eine Abhängigkeit vorhergesagt. Sind beide stattdessen in den Daten unabhängig, muss das Modell unter *faithfulness* verworfen werden. Gibt man *faithfulness* auf, ist das Modell jedoch weiterhin konsistent mit den Daten, da es möglich ist, dass der direkte Effekt von Anwesenheit auf Punkte durch den indirekten Effekt aufgehoben wird. Da man das Modell dadurch retten kann, ist der zweite Test schwächer als der erste.

Nun könnte man auf die Idee kommen, dass das Zutreffen einer Vohersage (also das Auffinden einer vorhergesagten (Un)Abhängigkeit) bedeutet, das getestete Modell ist wahr. Dieser Schluss ist jedoch nicht möglich. Zwar lassen sich aus einem gegebenen Modell eindeutige Vorhersagen zu statistischen Abhängigkeiten machen. Eine bestimmte (Un-)Abhängigkeit ist jedoch stets mit mehreren Modellen vereinbar. Das heißt, verschiedene Kausalmodelle sagen mitunter die gleichen statistischen Abhängigkeiten voraus (auch unter *faithfulness*).

Beispiel 2.10. So sagen das bekannte Modell aus Abb. 1.5 aber auch das Modell in Abb. 2.8 voraus, dass

$$\text{Anwesenheit} \perp \text{Kurswahl} | \text{Motivation}. \tag{2.14}$$

Auf Basis dieser Unabhängigkeit kann ich also nicht entscheiden, welches Modell wahr ist, oder besser gesagt: näher an der wirklichen Kausalstruktur.

Dies hat nun verschiedene Konsequenzen. Die offensichtlichste ist, induktive Modellbildung ohne vorherige Annahmen über die Kausalstruktur ist nicht möglich.[5]

[5] Unter *faithfullness* ist es jedoch bereits möglich, von statistischen Unabhängigkeiten auf die Kausalstruktur zu schließen. Ein zentraler Ausgangspunkt hierfür ist, dass zwei gegebene Variablen,

Ferner kann ein Test immer nur das Modell als ganzes testen, nicht den Zusammenhang zwischen den beiden Variablen, auf den getestet wird. Für diesen Zusammenhang gilt der Test nur unter der Voraussetzung, dass der Rest des Modells korrekt ist. Schließlich können zwei konkurrierende Modelle nur an Hand von Vorhersagen unterschieden werden, die sich zwischen den Modellen unterscheiden.[6]

Beispiel 2.11. Wie wir bereits gesehen haben, können die Modelle in Abb. 1.5 und Abb. 2.8 nicht anhand der bedingten Unabhängigkeit in (2.14) unterschieden werden. Ein Test auf

$$\text{Anwesenheit} \perp \text{Punkte} | \text{Motivation} \tag{2.15}$$

würde beide Modelle jedoch gegeneinander testen. Das Modell, dass die richtige Vorhersage macht, würde im Anschluss (vorerst) beibehalten werden, das andere verworfen.

2.2 Identifikationsstrategien für den Gesamteffekt

Die Regeln der *d-separation* sind auch die Grundlage bei der Untersuchung von Kausalzusammenhängen zwischen zwei spezifischen Variablen, also für kausale Inferenz. Im einfachsten Fall werden dabei mindestens zwei Modelle gegeneinander getestet: eines mit und eines ohne Kausalpfade zwischen den beiden interessieren-

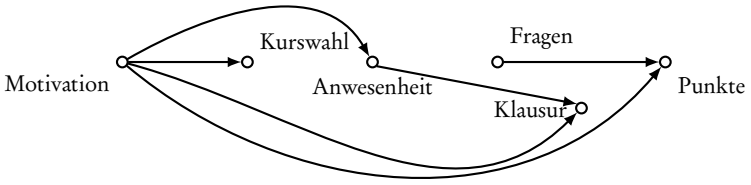


Abb. 2.8 Alternatives Modell zu Abb.1.5, das auch Zusammenhang zwischen Anwesenheit und Punkten gg. Motivation vorhersagt.

---

die marginal unabhängig sind, konditional auf eine dritte Variable nur dann abhängig sind, wenn diese Variable ein gemeinsamer Nachkomme ist. Somit kann auf Basis der beobachteten marginalen Unabhängigkeit und bedingten Abhängigkeit jeweils eine Kante von beiden Variablen auf die konditionierte Variable gezeichnet werden. Diese und weitere Strategien zur Entdeckung von Kausalstruktur sind gerade Gegenstand intensiver Forschung (Spirtes et al., 2001; Shalizi, 2016, Kap. 26). Allerdings sind diese mit mehr als einer Hand voll Variablen nicht mehr ohne automatisiertes *machine learning* umsetzbar, und damit nicht Teil der gegenwärtig standardmäßigen Methodenausbildung in den Sozialwissenschaften.

[6] Kurze Analogie: eine korrekte Uhr geht niemals falsch. Aber eine falsche Uhr zeigt zweimal am Tag die richtige Zeit an, macht also korrekte Vorhersagen. Habe ich nur Daten über diese zwei Zeitpunkte, ist es unmöglich beide Uhren zu unterscheiden.

den Variablen. Damit der Test funktioniert, muss es möglich sein, alle nichtkausalen Pfade zwischen den beiden Variablen zu schließen, während alle kausalen Pfade geöffnet bleiben. Man trennt also nichtkausale Abhängigkeit von dem Teil der Abhängigkeit, der auf den kausalen Effekt zurück geht. Dieser verbleibende Teil der Abhängigkeit kann dann, *gegeben der Richtigkeit des Restmodells*, als kausaler Gesamteffekt interpretiert werden. Ist dies möglich, spricht man davon, dass der kausale Effekt von X auf Y *identifizierbar* ist und dass der auf diese Weise bedingte statistische Zusammenhang zwischen X und Y den kausalen Effekt von X auf Y *identifiziert*.

Die Trennung kausaler Pfade von nichtkausalen Pfaden ist im Übrigen genau das, was eine hypothetische Intervention im DAG vollzieht. Durch die Simulation dieser Intervention über randomisierte Treatmentzuweisung geschieht das gleiche. Konditionierung auf der Menge an Variablen S, die alle nichtkausalen Pfade zwischen X und Y schließt, alle kausalen Pfade aber offen lässt, ist damit eine weitere mögliche Simulation einer hypothetischen Intervention. Der auf S konditionierte Zusammenhang zwischen X und Y kann damit genutzt werden, um den Effekt der interessierenden hypothetischen Intervention zu bestimmen:

$$Pr(Y|do(X=x)) = \sum_s Pr(Y|X=x, S=s)Pr(S=s).^{7} \tag{2.16}$$

Verknüpfung der empirischen Daten mit dem Kausalmodell

Bevor jedoch eine Kausalhypothese (oder jede andere aus einem DAG abgeleitete Hypothese) empirisch getestet werden kann, müssen die tatsächlich vorhandenen Daten mit dem Kausalmodell verknüpft werden. Denn meist sind nicht alle in die Datengenerierung involvierten Variablen in den Daten selbst gemessen. Außerdem kann es sein, dass während der Datenerhebung bereits (unfreiwillig) auf bestimmte Variablen konditioniert wird. Diese Sachverhalte können in den ursprünglichen DAG integriert werden. So werden wir im Folgenden in den vorliegenden Daten *gemessene* Variablen durch geschlossene Knoten (●) darstellen. Offene Knoten (○), stehen daher nur noch für ungemessene Variablen. Eine in den Daten automatisch vorhandene Konditionierung einer Variablen wird durch den bereits bekannten Kasten um diese Variable verdeutlicht. Automatisch konditioniert wird beispielsweise dadurch, dass nur für eine Untergruppe der eigentlich interessierenden Population Daten erhoben wurden, beispielsweise durch Verweigerung der Teilnah-

[7] Das heißt, interessiere ich mich für den durchschnittlichen kausalen Effekt eines dichotomen Treatments auf ein metrisches Outcome, kann ich diesen als Differenz der nach X und S bedingten und nach S gewichteten Mittelwerte von Y berechnen:

$$E[Y|do(X=1)] - E[Y|do(X=0)]$$

$$= \sum_s E[Y|X=1, S=s)]Pr(S=s) - \sum_s E[Y|(X=0), S=s]Pr(S=s).$$

me an einer Befragung oder den Ausfall von Beobachtungseinheiten über die Zeit (auch: Panelmortalität, Attrition).[8]

Beispiel 2.12. Abb. 2.9 zeigt unser ursprüngliches Modell zur Generierung der Teilnehmerdaten unter Berücksichtigung der tatsächlich vorhandenen Daten. Da keine Informationen zur Motivation der Teilnehmer und zum Frageverhalten vorliegen, sind diese Variablen weiterhin durch offene Knoten gekennzeichnet. Ferner wird in den Daten sowohl auf die Kurswahl als auch auf die Teilnahme an der Klausur konditioniert. Denn die Daten sind ausschließlich in einer konkreten Veranstaltung (in zwei verschiedenen Semestern) erhoben und Informationen über die Punkte sind nur für Teilnehmer vorhanden, die auch die Klausur geschrieben haben.

Zielpopulation, Effektheterogenität und Generalisierbarkeit

Ein weiterer Aspekt, der vor der eigentlichen Identifikation kausaler Effekte geklärt werden sollte, ist die genaue Zielpopulation, auf die die Analyseergebnisse bezogen werden sollen. Diese Frage ist häufig komplizierter zu beantworten, als es auf den ersten Blick scheint. So ist die offensichtliche Zielpopulation diejenige, aus der die Beobachtungseinheiten in den Daten stammen. Aber diese Daten und die in Ihnen enthaltenen Beobachtungseinheiten entstammen einem bestimmten Raum und einer bestimmten Zeit. Bei der Untersuchung kausaler Zusammenhänge will man jedoch meist Schlussfolgerungen ziehen, die möglichst von Raum und Zeit losgelöst, also möglichst universell, gelten.

Beispiel 2.13. Eigentlich interessiert bei der Frage nach dem Effekt der Anwesenheit auf die Klausurleistung doch, ob dies über die in den Daten enthaltenen Teilnehmer hinausgeht. So würde ich doch wenigstens gerne Aussagen über alle Master-Studierenden in Soziologie in Köln machen, aber am liebsten über alle Kölner

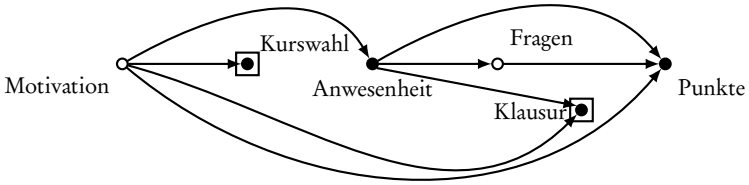


Abb. 2.9 Modell des Generierungsprozesses der Teilnehmerdaten unter Berücksichtigung der tatsächlich vorhandenen Daten

[8] An dieser Stelle können im Prinzip auch vermutete Messfehler gekennzeichnet werden. Diese blenden wir jedoch wie gesagt der Einfachheit halber aus.

Studierenden oder sogar alle Studierenden in Deutschland. Die Daten beruhen aber lediglich auf Teilnehmern einer einzigen Veranstaltung im Master Soziologie an der Universität zu Köln in zwei Semestern im Jahr 2013/14. Wenn ich mich dann schon in meinen Schlussfolgerungen auf typische Teilnehmer genau dieser Veranstaltung beschränke, dann möchte ich wenigstens auch etwas über den untersuchten Zusammenhang in zukünftigen Veranstaltungen sagen können. Kurz: Ausschließlich Aussagen über die tatsächlichen Teilnehmer zu treffen, wird selten mein Anspruch sein.

In Abgrenzung zur Zielpopulation spricht man deswegen von der Population, aus der die Beobachtungseinheiten direkt stammen, als Studienpopulation. Basiert die Studie auf einer Vollerhebung oder einer Zufallsstichprobe der Zielpopulation, dann (und nur dann) sind Studienpopulation und Zielpopulation identisch. Selbst wenn die Daten einer Studie genutzt werden sollen, um etwas über diesselben Beobachtungseinheiten ein Jahr später auszusagen, entspricht die Zielpopulation nicht mehr der ursprünglichen Studienpopulation.

Ob die Ergebnisse für die Studienpopulation auch über diese hinaus Geltung haben, ist unmittelbar mit dem Vorliegen von Heterogenität des interessierenden Kausaleffekts verküpft. Generalisierbarkeit von der Studienpopulation auf eine davon verschiedene Zielpopulation ist nur dann möglich, wenn sich beide hinsichtlich der Verteilung von Effektmoderatoren *nicht* unterscheiden, oder, wenn die Unterschiede bekannt sind und damit in diese Analyse einfließen können. Generalisierungen ohne Kenntnis der Effektmoderatoren und ihrer Verteilung in Studien- und Zielpopulation beruhen damit stets auf der Annahme, dass sich der kausale Effekt in beiden (im Mittel) nicht unterscheidet. Man nennt diese Annahme deswegen auch Effekthomogenität.

Beispiel 2.14. Nehmen wir an, wir möchten unsere Analyse auf alle Masterstudierenden in Soziologie in Köln verallgemeinern. Unser DAG zum Zusammenhang zwischen Anwesenheit und Motivation bietet nicht direkt Hinweise darauf, ob sich Studien- und Zielpopulation unterscheiden, da Effektheterogenität und Effektmoderatoren nicht explizit gemacht sind. Wir sehen jedoch, dass die Teilnahme an der Veranstaltung und damit die Teilnahme an der Analyse von der Motivation beeinflusst wird. Ist Motivation nun ein Moderator des Effekts von Anwesenheit auf Punkte, impliziert dies Effektheterogenität zwischen Studien- und Zielpopulation. Denn beide unterscheiden sich systematisch auf Motivation und Motivation wiederum sorgt für systematische Unterschiede im Effekt von Anwesenheit auf Punkte. In diesem Fall können ohne weiteres Wissen über die genauen Motivationsunterschiede in Studien- und Zielpopulation die Ergebnisse der Untersuchung nicht auf alle Masterstudierenden in Soziologie verallgemeinert werden (und schon gar nicht auf alle Veranstaltungen in diesem Studiengang).

2.2.1 Identifikation durch einfache Konditionierung

Kommen wir nun aber zurück zur eigentlichen Identifikation des Gesamteffekts von X auf Y. Unter welchen Bedingungen ist diese durch einfache Konditionierung möglich? Die Antwort darauf lässt sich in zwei Punkte unterteilen:

1. Es muss eine Menge an Variablen S geben, eingeschlossen der leeren Menge, deren Konditionierung sämtliche nichtkausale Pfade zwischen X und Y schließt und gleichzeitg alle kausalen Pfade zwischen ihnen offen belässt.
2. Sämtliche Variablen S (und natürlich X und Y) müssen in den Daten, mit denen die kausale Inferenz durchgeführt werden soll, gemessen sein, denn auf ungemessene Variablen kann nicht konditioniert werden.

Umgekehrt kann die Identifikation des Gesamteffekts von X auf Y generell aus *drei* Gründen scheitern, die in der Literatur häufig als eine bestimmte Art von Vezerrung (*bias*) bezeichnet werden:

1. ausbleibende Konditionierung eines Nicht-*colliders* auf einem offenen nichtkausalen Pfad zwischen X und Y (auch: Konfundierung oder *confounding bias*)[9]
2. Öffnung eines geschlossenen nichtkausalen Pfades zwischen X und Y durch Konditionierung eines *colliders* oder einer seiner Nachkommen (neuerdings: *endogenous selection bias*)[10].
3. Schließung eines Kausalpfades zwischen X und Y durch Konditionierung auf einer sich auf diesem Pfad befindlichen Variable (auch: Überkontrolle oder *overcontrol bias*)

Die Regeln zur Identifikation und die mit ihr verbundenen Probleme machen deutlich, dass es einen Unterschied macht, auf welche Variablen konditioniert und damit in der Analyse kontrolliert wird. Nicht nur kann die *Nichtkontrolle* bestimmter Variablen zu Problemen führen. Auch die *Kontrolle* der „falschen" Variablen lässt kausale Inferenz mitunter scheitern. Herkömmliche Strategien zur Auswahl von „Kontrollvariablen", die noch immer häufig ihren Weg in Lehrbücher und Hörsäle finden, sind unvollständig und führen häufig gar zum Scheitern kausaler Inferenz. Besonders negative Beispiele hier sind: „Kontrolliere alles, was du kannst" (*kitchen sink approach*) oder „Kontrolliere alle Variablen, die sowohl mit

[9] Andere Bezeichnungen in der Literatur sind *self-selection*, *selection on unobservables*, *unobserved heterogeneity* oder auch einfach *selection bias*. *Selection* bezieht sich dabei stets auf das Treatment, nicht auf die Stichprobe. Wir werden insbesondere diese Ausdrücke hier nicht verwenden, um Verwirrung mit der nächsten Art der Verzerrung zu vermeiden.

[10] In den letzten Jahren konnte eindrucksvoll gezeigt werden, dass viele bisher disparat betrachtete Verzerrungen in der empirischen Sozialforschung *Spezialfälle* der „Kontrolle einer *collider*-Variable" sind. Darunter fallen *sample selection bias*, *attrition bias*, *survivor bias*, „Selektion auf der abhängigen Variable", *item nonresponse bias* oder auch *homophily bias* in der Analyse sozialer Netzwerke, um nur einige zu nennen (Elwert und Winship, 2014; Glymour, 2006b; Hernán et al., 2004; Shalizi und Thomas, 2011)

Treatment als auch mit Outcome *korrelieren*". Letzteres führt fast zwangsläufig zu *endogenous selection bias* und *overcontrol bias*.[11]

Ebenso zeigt sich, dass die Auswahl der Kontrollvariablen stets von einem Kausalmodell geleitet werden muss. Gänzlich ohne Annahmen über die zugrundeliegenden Kausalstrukturen ist keine kausale Inferenz möglich. Denn eine Interpretation als Kausaleffekt funktioniert nur auf Basis eines (möglichst wahren) Kausalmodells. Für Analysen, für die keinerlei Kausalmodell (ob als DAG oder einer alternativen Notation wie Strukturgleichungen) postuliert wird, können die Ergebnisse niemals kausal interpretiert werden. Gleichzeitig sind kausale Interpretation stets abhängig von der Richtigkeit des zugrunde gelegten Strukturmodells. Kausale Effekte sind also nie absolut, sondern stets provisorisch bis zur Widerlegung (oder schon Anzweiflung) des Modells. So kann es durchaus zu Kontroversen zwischen (Gruppen von) Wissenschaftlern kommen, die sich uneinig darüber sind, ob ein kausaler Effekt mit den gegebenen Daten identifiziert werden kann. In DAG-Terminologie ausgedrückt heißt das beispielsweise, man ist sich uneinig darüber, ob alle nichtkausalen Pfade geschlossen werden können. Eine Gruppe für die das plausibel ist, interpretiert ihre Ergebnisse kausal, die Kritiker bezweifeln diese Interpretation auf Basis ihres (plausiblen) Alternativmodells. Dadurch, dass beide ihre Modelle aber explizit machen, ist es gut möglich, dass die Modelle in Zukunft (mit besseren Daten) gegeneinander getestet werden können und der Streit entschieden werden kann. Bis zum nächsten plausiblen Alternativmodell. Das Kausalmodell erlaubt also eine systematische, wissenschaftliche Auseinandersetzung. Die mehr oder weniger durch Zufall geleitete Variablenauswahl hingegen gleicht Stochern im Nebel. So meinte schon der berühmte britische Statistiker George Box (1979, S. 2):

> The great advantage of the model-based over the ad hoc approach, it seems to me, is that at any given time we know what we are doing.

Einfache Konditionierung lässt sich beispielsweise über die im folgenden Algorithmus beschriebenen Schritte in die Praxis umsetzen:

1. Gibt es in meinem mit den vorhandenen Daten verknüpften Modell *offene nicht*kausale Pfade zwischen X und Y?
   Wenn nein, gehe zu 5.
2. Wenn ja, liegen auf diesen Pfaden Nicht-*collider*?
   Wenn nein, gehe zu 6.
3. Wenn ja, sind diese Nicht-*collider* in meinen Daten gemessen?
   Wenn nein, gehe zu 6.
4. Wenn ja, bleiben durch die Konditionierung alle nichtkausalen Pfade geschlossen und alle kausalen Pfade geöffnet?
   Wenn nein, gehe zu 6.
5. Wenn ja, Gesamteffekt von X auf Y identifiziert
   Wenn nein, gehe zu 6.

[11] Mildere Beispiele für solche Faustregeln sind „Kontrolliere alle Variablen, die dem Treatment zeitlich vorausgehen" oder „Kontrolliere alle Variablen, die das Treatment beeinflussen".

6. Es gibt nun folgende Alternativen:
   a. erhebe Daten mit Variablen, für die 2.-4. positiv beantwortet werden können,
   b. wähle eine alternative Identifikationsstrategie (siehe Kap. 2.2.2 und 2.3 sowie die restliche Veranstaltung),
   c. passe den DAG durch möglichst wenige plausible Annahmen so an, dass 1.-5. möglich sind, oder
   d. interpretiere die Ergebnisse ausschließlich als statistische Zusammenhänge (d.h. keine Empfehlungen für politische Interventionen, keine Sprache, die Kausalzusammenhang impliziert etc.)

Zentral hierbei ist, dass die laut Kausalmodell notwendigen Konditionierungsvariablen in den gegebenen Daten gemessen sein müssen. Ist dies nicht der Fall, fällt einfache Konditionierung als Strategie zur Identifkation kausaler Effekte (in der eigentlichen Zielpopulation) aus. Die gegenwärtig in der Praxis leider allzu häufig gewählte „Antwort" auf offene nichtkausale Pfade ist eine Abwandlung von 6c. oben, bei der nämlich die gewählte Methode (z.B. Regression) die Annahmen implizit (und häufig ohne Wissen der Anwender) vorgibt, die Ergebnisse dann, trotz äußerst geringer Plausibilität dieser Annahmen, zumindest indirekt (z.B. in Form von Politikempfehlungen) kausal interpretiert werden („X führt zu Y"...). Der Beitrag zu wirklicher kausaler Inferenz solcher Untersuchungen ist, freundlich ausgedrückt, zweifelhaft. Die Mindestanforderung an kausale Inferenz ist, dass deren Konsument mithilfe eines transparenten Kausalmodells, das den Ergebnissen zur Seite gestellt ist, für sich selbst die Plausibilität dieses Modells und damit die Plausibilität einer kausalen Interpretation der Ergebnisse prüfen kann.

Beispiel 2.15. Versuchen wir nun den kausalen Effekt von Anwesenheit auf Punkte mit unseren Daten durch einfache Konditionierung zu identifizieren (siehe Abb. 2.9):

1. Gibt es in meinem mit den vorhandenen Daten verknüpften Modell *offene nicht*kausale Pfade zwischen Anwesenheit und Punkte?
   Ja:
   - Anwesenheit ← Motivation → Punkte
   - Anwesenheit → [Klausur] ← Motivation → Punkte
2. Liegen auf diesen Pfaden Nicht-*collider*?
   Ja: Motivation
3. Sind diese Nicht-*collider* in meinen Daten gemessen?
   Nein
4. Es gibt nun folgenden Alternativen:
   a. Erhebe Daten, in denen Motivation gemessen ist.[12]

[12] Daten, in denen lediglich auf Klausur *nicht* konditioniert wird, wären nicht ausreichend, da weiterhin der nichtkausale Pfad Anwesenheit ← Motivation → Punkte bestehen würde. Gleichzeitig

b. Wähle eine alternative Identifikationsstrategie, die zumindest keine Messung von Motivation erfordert (siehe Kap. 2.2.2 und 2.3 sowie die restliche Veranstaltung),
c. treffe die Annahme, dass Motivation nicht auf Anwesenheit wirkt, oder
d. interpretiere den Zusammenhang zwischen Anwesenheit und Motivation ausschließlich als Unterschied in der beobachteten Punktzahl nach Anwesenheitshäufigkeit: „Teilnehmer, die häufiger anwesent waren, haben mehr Punkte bekommen (möglicherweise, weil diese Teilnehmer generell motivierter waren)."

Einfache Konditionierung funktioniert als Strategie zur Identifikation kausaler Effekte nur dann funktioniert, wenn auf allen offenen nichtkausalen Pfaden mindestens eine Variable liegt, die in den vorliegenden Daten gemessen ist und damit für eine Konditionierung herangezogen werden kann. Muss man also kausale Inferenz aufgeben, wenn es offene nichtkausale Pfade mit ausschließlich ungemessenen Variablen gibt? Nicht zwangsläufig. Denn es gibt eine Reihe alternativer Identifikationsstrategien, die auch funktionieren, wenn nicht alle offenen nichtkausalen Pfade zwischen X und Y geschlossen werden können.

2.2.2 Identifikation durch frontdoor-Konditionierung

Eine Alternative zur Identifikation des Gesamteffekts von X auf Y ist die sogenannte *frontdoor*-Konditionierung über Variablen M, die auf kausalen Pfaden von X nach Y liegen. *Frontdoor*-Konditionierung besteht aus einer Sequenz einfacher Konditionierungen, über die zunächst die Effekte von X auf M identifiziert werden und im Anschluss die Effekte von M auf Y. Die oben angegebenen Bedingungen zum Funktionieren der einfachen Konditionierung müssen damit für alle diese Effekte zutreffen. Der Gesamteffekt von X auf Y ergibt sich dann aus der Summe aller indirekten Effekte über M. Dazu muss jedoch auf jedem Pfad natürlich mindestens eine Variable M in den Daten gemessen sein.

Beispiel 2.16. Abb. 2.10 zeigt ein Szenario, in dem die Identifikation des Gesamteffekts von Anwesenheit auf Punkte durch *frontdoor*-Konditionierung auf Lernhinweise und Fragen möglich ist, obwohl einfache Konditionierung nicht funktioniert, da Motivation ungemessen ist. Diese Strategie funktioniert, weil

1. mindestens eine Variable auf jedem kausalen Pfad von Anwesenheit zu Punkten gemessen ist,
2. es weder offene nichtkausale Pfade zwischen Anwesenheit und Lernhinweise noch zwischen Anwesenheit und Fragen gibt (Punkte ist jeweils *collider*)

---

ist es auch nicht schlimm, wenn in den neuen Daten weiterhin auf Klausur konditioniert wird, denn der so geöffnete Pfad kann durch die Konditionierung auf Motivation wieder geschlossen werden.

3. es (gegeben Anwesenheit) weder offene nichtkausale Pfade zwischen Lernhinweise und Punkte noch zwischen Fragen und Punkte gibt.

2.3 Identifikation weniger umfassender Kausaleffekte

Sind auch die Bedingungen für Identifikation durch *frontdoor*-Konditionierung nicht erfüllt, muss man sich (außer über Zusatzannahmen) entweder von der Identifikation des Gesamteffekts, des Gesamteffekts für die ursprüngliche Zielpopulation oder des genauen Gesamteffekts verabschieden. Ist man dazu bereit bzw. gezwungen, gibt es verschiedene weitere Strategien „engere“ Effekte als den genauen Gesamteffekt für die Zielpopulation zu identifizieren. Dies sind indirekte Effekte, bedingte oder auch lokale Effekte, die nur für einen mehr oder weniger genau definierten Teil der Zielpopulation gelten, und partiell identifizierte Effekte, die keine genaue Angabe zur Größe des interessierenden Effekts machen. Die im weiteren Verlauf der Veranstaltung behandelten Untersuchungsdesigns sind besonders dazu geeignet, lokale kausale Effekte zu identifizieren.

2.4 Zum Weiterlesen

Eher untechnische und trotzdem sehr klare Einführungen in die Identifikation kausaler Effekte mithilfe von DAGs bieten Glymour (2006b) und Elwert (2013). Weiterführende (auch technisch-mathematische) Grundlagen finden sich bei Pearl (1995), Greenland et al. (1999) und Robins (2001) in Artikelform und bei Glymour (2001), Spirtes et al. (2001) und Pearl (2009b) in Buchform. Hernán et al. (2004) sowie Elwert und Winship (2014) erläutern *endogenous selection* detailliert und an Hand zahlreicher Beispiele. Die technischen Grundlagen von Identifikation durch einfache Konditionierung finden sich z. B. bei Shpitser et al. (2010). Eine verständliche Einführung in *frontdoor*-Konditionierung bieten Knight und Winship (2013),

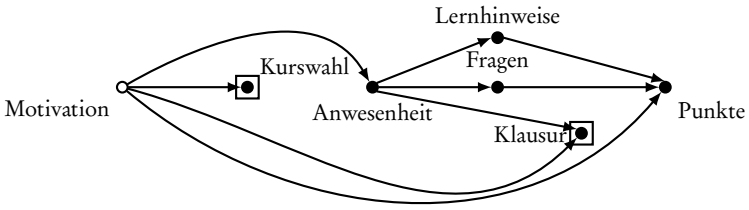


Abb. 2.10 Modell des Generierungsprozesses der Teilnehmerdaten unter Berücksichtigung der tatsächlich vorhandenen Daten

die Grundlagen hingegen Pearl (1995, 2009b). Die Identifikation indirekter (und direkter) Effekte behandeln Robins und Greenland (1992) und Pearl (2005). Auch die Bücher von VanderWeele (2015) und Hong (2015) befassen sich ausführlich damit. Morgan und Winship (2015, Kap. 12) geben einen sehr guten Überblick zu partieller Identifikation und systematischer Sensitivitätsanalyse.

Kapitel 3
Schätzen kausaler Effekte

3.1 Nichtparametrische Schätzung

Sind kausale Effekte durch einfache Konditionierung identifizierbar, lässt sich die kontrafaktische Verteilung von Y nach Festsetzung von X auf x als Funktion der beobachteten Verteilung von Y bedingt durch Treatment X und die Konditionierungsvariablen S darstellen:

$$Pr(Y|do(X=x)) = \sum_s Pr(Y|X=x, S=s)Pr(S=s) \tag{3.1}$$

Der kausale Effekt von X auf Y entspricht damit (gegeben der Korrektheit des jeweiligen DAG) dem durch S bedingten statistischen Zusammenhang zwischen X und Y. Liegen ein dichotomes Treatment und ein metrisches Outcome vor, lässt sich der durchschnittliche kausale Effekt von X auf Y damit folgendermaßen berechnen:[1]

$$\begin{aligned} &E[Y|do(X=1)] - E[Y|do(X=0)] \\ &= \sum_s E[Y|X=1, S=s)]Pr(S=s) - \sum_s E[Y|(X=0), S=s]Pr(S=s) \end{aligned} \tag{3.2}$$

Für die Berechnung gibt es verschiedene Vorgehensweisen, die jedoch alle zum gleichen Ergebnis führen und nur unter der Bedingung funktionieren, dass Konditionierung auf S tatsächlich sämtliche nichtkausalen Pfade zwischen X und Y schließt und alle kausalen Pfade offen bleiben. Diese Berechnungsmethoden werden als nichtparametrische Schätzer bezeichnet, da sie keine Vorgaben bzw. An-

[1] Der Schätzer für das kausale Risikoverhältnis im Falle eines dichotomen Outcomes sieht entsprechend so aus:

$$\frac{Pr(Y=1|do(X=1))}{Pr(Y=1|do(X=0))} = \frac{\sum_s Pr(Y=1|X=1, S=s)Pr(S=s)}{\sum_s Pr(Y=1|X=0, S=s)Pr(S=s)}$$

Tabelle 3.1 Vorgehen bei Standardisierung

	$X=1$		$X=0$	
	$S=1$	$S=0$	$S=1$	$S=0$
$E[Y\|X=x,S=s]$	5	2	7	3
$Pr(S=s)$	0.2	0.8	0.2	0.8
$E[Y\|X=x,S=s]Pr(S=s)$	1	1.6	1.4	2.4
$\sum_s E[Y\|X=x,S=s]Pr(S=s)$	2.6		3.8	

nahmen über die funktionale Form des berechneten Zusammenhangs und damit des durch diesen Zusammenhang identifizierten kausalen Effekts machen.

3.1.1 Standardisierung

Eine erste Schätzmethode wird als Standardisierung bezeichnet. Hierbei werden wie in Gleichung 3.2. notiert einfach die nach S bedingten und gewichteten Mittelwerte von Y innerhalb der (beiden) Ausprägungen des Treatments berechnet und anschließend deren Differenz gebildet. Diese Differenz ist ein nichtparametrischer Schätzer des durchschnittlichen kausalen Effekts von X auf Y. Vorstellen kann man sich diese Methode wie eine mehrdimensionale Kreuztabelle, die das mittlere Y für alle Merkmalskombinationen von S innerhalb der Ausprägungen von X enthält (siehe Tabelle 3.1). Das heißt, gibt es mehrere Variablen S muss für jede Kombination eine eigene Spalte erstellt werden. Haben wir beispielsweise zwei dichotome S müssten innerhalb jeder Ausprägung von X vier Spalten eingerichtet werden. Bei mehr als zwei Ausprägungen des Treatments vergrößert sich die Tabelle entsprechend ebenso. Hierbei wird schon die praktische Schwierigkeit, nichtparametrischer Methoden klar: Was, wenn in den Daten nicht jede Merkmalskombination (in ausreichender Fallzahl) beobachtet wird (siehe Kap. 3.1.4)?

3.1.2 Matching

Matching, oder besser gesagt: exaktes Matching, ist eine alternative Methode der nichtparametrischen Konditionierung. Hierbei geht man nach folgendem Algorithmus vor:

1. Trenne alle Beobachtungen in zwei Gruppen $x=1$ und $x=0$.
2. Nimm erste Beobachtung aus $x=1$ und notiere deren Ausprägung auf S, s.
3. Suche Beobachtung mit gleichem s aus Gruppe $x=0$ (statistischer Zwilling).
4. Berechne Differenz auf Y.
5. Wiederhole 2.-4. für alle Beobachtungen $x=1$
6. Wiederhole 2.-4. für alle Beobachtungen $x=0$, suche Match aus $x=1$.

7. Berechne arithmetisches Mittel über alle Differenzen.

Die errechnete Differenz ist der Matchingschätzer des durchschnittlichen kausalen Effekts von X auf Y. Auch hier vergrößert sich die Anzahl der Gruppen mit den Ausprägungen von X. Bei mehreren S ist darauf zu achten, dass der jeweilige Match der vorliegenden Beobachtung auf den Ausprägungen jeder S gleicht.

3.1.3 Gewichtung mit der inversen Treatmentwahrscheinlichkeit

Eine weitere Methode, die bisher vorallem in der Epidemiologie bekannt ist, ist die Gewichtung mit der inversen Treatmentwahrscheinlichkeit (*inverse probability of treatment weighting*). Ausgangspunkt hierbei ist die Tatsache, dass bei offenen nichtkausalen Pfaden über S die Treatmentvariable ungleich über die Ausprägungen von S verteilt ist. Das heißt, die Wahrscheinlichkeit eine bestimmte Ausprägung des Treatments zu besitzen, unterscheidet sich über die Ausprägungen von S. Indem man jede Beobachtung mit der inversen Wahrscheinlichkeit der für die jeweilige Beobachtung vorliegenden Ausprägung des Treatments gewichtet, erzielt man eine Gleichverteilung der Treatmentausprägungen über die Ausprägungen von S. In der auf diese Weise gewichteten Population, auch Pseudo-Population genannt, entspricht nun der marginale statistische Zusammenhang dem kausalen Effekt von X auf Y. Nach Gewichtung muss man also nicht mehr aktiv auf S konditionieren. Die Pseudo-Population stellt die Simulation eines Szenarios dar, bei dem alle Beobachtungen gleichzeitig $x = 1$ und $x = 0$ erhalten, so dass jede Beobachtung als ihre eigene Kontrollbeobachtung fungiert (siehe Hernán und Robins (2016, Kap. 2.4)).

3.1.4 Positivität und der Fluch der Mehrdimensionalität

Voraussetzung für das Funktionieren dieser Methoden ist, wie wir wissen, dass alle nichtkausalen Pfade zwischen X und Y durch Konditionierung auf S geschlossen werden können. Das heißt, ohne Identifikation, berechnet keine dieser Methoden den wahren Kausaleffekt von X auf Y. Für alle Methoden ist daher zwingend, dass alle für die Konditionierung benötigten S auch gemessen sind. Ist auch nur eine S nicht gemessen, kann weder eine vollständige Kreuztabelle erstellt noch auf allen S gemachted werden oder die Treatmentwahrscheinlichkeit innerhalb von allen S ermittelt werden.

Damit der Gesamteffekt von X auf Y für die gesamte interessierende Population berechnet werden kann, muss jedoch noch eine weitere Bedingung erfüllt sein. Diese Bedingung wird als Positivität bezeichnet[2] und besagt, dass alle Ausprägungen des Treatments in jeder vorkommenden Kombination der Ausprägungen von

[2] Weitere in der Literatur genutzte Begriffe sind *common support* und *overlap*.

S vetreten sein müssen. Anders ausgedrückt: Die Wahrscheinlichkeit des Auftretens jeder Ausprägung von X innerhalb aller existierenden Kombinationen von s muss größer als Null sein:

$$Pr(X = x|S = s) > 0, \text{ für alle s mit } Pr(S = s) \neq 0. \tag{3.3}$$

Ist dies nicht gegeben, gibt es *leere* Zellen in der $x \times s$-Matrix. Dann können für die Standardisierung nicht mehr alle benötigten Mittelwerte berechnet werden. Beim Matching treten Fälle auf, für die kein Zwilling gefunden werden kann. Und beim Gewichten mit der inversen Treatmentwahrscheinlichkeit kann nicht für alle Ausprägungen x innerhalb der Ausprägungen von S die Wahrscheinlichkeit berechnet werden bzw. für manche Ausprägungen beträgt sie null. Noch anders ausgedrückt: Es finden sich nicht für alle Beobachtungen bzw. Gruppen Vergleichsbeobachtungen bzw. -gruppen mit jeweils anderer Treatmentausprägung x.

Eine Verletzung von Positivität kann systematisch auftreten oder zufällig sein. Ein nicht sonderlich elegantes aber anschauliches Beispiel für eine systematische Verletzung von Positivität ergibt sich beim Treatment Schwangerschaft und der Kovariate Geschlecht. Innerhalb der Gruppe der Männer kann das Treatment Schwangerschaft ausschließlich die Ausprägung null annehmen. Für Frauen hingegen sind mehrere Ausprägungen möglich. Beobachtungen mit Ausprägungen auf S, für die Positivität systematisch verletzt ist, sollten stets von der Analyse ausgeschlossen werden.

Die Wahrscheinlichkeit zufälliger Verletzungen der Positivität wird durch drei Faktoren beeinflusst:

1. die Zahl der Ausprägungen des Treatments
2. die Zahl der Konditionierungsvariablen und deren Ausprägungen
3. die Zahl der Beobachtungen.

Je weniger Fälle sich auf immer mehr Zellen der Matrix $x \times s$, dem Variablenraum, verteilen müssen, desto wahrscheinlicher wird, dass nicht mehr alle Zellen gefüllt werden können. Das ist gleichbedeutend mit der Verletzung von Positivität. In realen Analysen, die meist eine große Zahl an Variablen beinhalten aber nur begrenzte Fallzahlen zur Verfügung haben, sind Verletzungen der Positivität damit kaum zu vermeiden. Meist gibt es mehr mögliche Kombinationen der Ausprägungen der Variablen als Fälle. Dies ist nicht schwer vorstellbar, führt man sich vor Augen, dass eine Kreuztabelle wie in Tabelle 3.1 für 20 Variablen mit jeweils 2 Ausprägungen bereits $2^{20} = 1.048.576$ Spalten umfasst. Man spricht in diesem Zusammenhang auch häufig vom Fluch der Mehrdimensionalität des Variablenraums.

Auch bei derartigen zufälligen Verletzungen der Positivität besteht die Möglichkeit, die betroffenen Variablenkombinationen von der Analyse auszuschließen. Dies kann jedoch insbesondere bei geringer Fallzahl dazu führen, dass ein großer Teil der Beobachtungen verloren geht. Die Standardantwort auf zufällige Positivitätsverletzungen ist deswegen die *statistische* Modellierung.

3.2 Parametrische Schätzung

Ist Positivität *zufällig* verletzt, besteht die Möglichkeit, die leer gebliebenen Zellen über parametrische Annahmen zur funktionalen Form des jeweiligen statistischen Zusammenhangs zwischen den Variablen zu füllen. Wie das funktioniert, lässt sich bereits an Hand eines bivariaten Zusammenhangs, wie in Abb. 1.3 dargestellt, demonstrieren. Dort sind bereits nicht alle möglichen Zellen mit Beobachtungen gefüllt. Beispielsweise gibt es keine Beobachtungen mit Anwesenheit=40. Für diese Ausprägung kann also mit diesen Daten die mittlere Punkzahl (nichtparametrisch) nicht berechnet werden. Nutzt man jedoch ein statistisches Modell, wie das dargestellte lineare Regressionsmodell, kann eine Aussage über die mittlere Punktzahl bei Anwesenheit=40 gemacht werden. Genauso funktioniert die Modellierung im Fall von mehreren Variablen. Stets wird auf Basis der vorhandenen Daten zunächst das ideale Modell mit der vorgebenen funktionalen Form gefunden und dieses dann benutzt, um die fehlenden Informationen mithilfe der Modellannahmen zu interpolieren. Sobald mehr als zwei Variablen für die Modellierung herangezogen werden, gehört zur funktionalen Form nicht mehr nur der Verlauf des Zusammenhangs (also linear, quadratisch, etc.), sondern auch, ob sich ein bestimmter Zusammenhang nach den Ausprägungen der anderen Variablen unterscheidet.

Modelliert werden können verschiedene Aspekte der gemeinsamen Verteilung von S, X und Y. Bei Modellen, bei denen nur Zusammenhänge zwischen S und X modelliert werden, spricht man von Treatmentmodellen. Beispiele hierfür sind Propensity Score Matching oder Inverse Probability of Treatment Weighting of Marginal Structural Models. Werden die Zusammenhänge von S und X mit Y modelliert, spricht man von Outcomemodellen. Die Regression in allen ihren Standard-Variationen ist ein Outcomemodell.

An dieser Stelle ist zu betonen, dass derartige statistische Modelle lediglich ein *technisches* Hilfsmittel sind, Datenknappheit und resultierende Positivitätsverletzungen abzumildern. Sie können jedoch nicht ungemessene Variablen auf irgendeine Art herbeizaubern. Sind kausale Effekte wegen offenen nichtkausalen Pfaden zwischen X und Y nicht identifizierbar, wird keine einfache Konditionierungsmethode den gewünschten Kausaleffekt berechnen, sondern einen statistischen Zusammenhang der mehr oder weniger weit davon entfernt liegt. Die Komplexität der Methode macht hier keinen Unterschied. Sind alle S gemessen und liegen ausreichend Beobachtungen vor, liefert eine Kreuztabelle einen ebenso guten Schätzer für den kausalen Effekt von Interesse wie eine Regression oder ein Propensity Score Matching. Keine statistische Methode ist „kausaler“ als eine andere und keine Methode macht aus statistischen Zusammenhängen kausale Effekte. Die kausale Interpretationen hängt stets von der Plausibilität der *theoretischen* Kausalstrukturen ab, die im DAG dargestellt sind. Umgekehrt kann es höchstens passieren, dass bei fehlerhaften parametrischen Annahmen (nichtparametrisch) identifizierte kausale Effekt verzerrt geschätzt werden. Wird beispielsweise ein linearer Zusammenhang spezifiziert, der tatsächliche Zusammenhang folgt aber einer andersartigen

Kurve, kann es zu mehr oder weniger großen Verzerrungen kommen, je nachdem wie stark die Abweichung des Modells vom wahren Zusammenhang ist.

Die parametrische Modellierung sei kurz am Beispiel der linearen Regression verdeutlicht. Generell liefert jedes lineare Regressionsmodell Schätzwerte für bedingte Mittelwerte von Y. Somit kann sie auch verwendet werden, um diejenigen Mittelwerte zu schätzen, die jeweils die Mittelwerte der kontrafaktischen Verteilung identifizieren. Damit dies jedoch tatsächlich funktioniert, sind einige Abweichungen vom Standardvorgehen vorzunehmen. So sind zunächst sämtliche Variablen S auf ihrem arithmetischen Mittel zu zentrieren. Die Konstante der Regression schätzt damit nicht mehr das mittlere Y, wenn alle S sowie X den Wert Null annehmen, sondern das mittlere Y, wenn alle S ihren Mittelwert annehmen und X den Wert Null. Anschließend sind Produktterme zwischen allen S und X zu bilden und das folgende Modell zu schätzen:

$$E[Y|X = x, S = s] = \alpha + \beta_1 x + \beta_2 s + \beta_3 xs \tag{3.4}$$

Der Regressionskoeffizient β_1 ist in diesem Modell ein unverzerrter Schätzer für einen durchschnittlichen (linearen) Kausaleffekt von X auf Y. Die Konstante α ist ein Schätzer des mittleren kontrafaktischen Y unter der Kausalbedingung $X = 0$, $E[Y|do(X = 0]$. Die Summe aus α und β_1 ist damit ein Schätzer für $E[Y|do(X = 1]$. Die Aufnahme der Produktterme aus S und X ist notwendig, da nicht (ohne Zusatzannahmen) auszuschließen ist, dass der Effekt von X nach S variiert, dass also Effektmoderation nach S vorliegt. Durch das Standardregressionsmodell der Form

$$E[Y|X = x, S = s] = \alpha + \beta_1 x + \beta_2 s \tag{3.5}$$

wäre per funktionaler Form festgelegt, dass der Effekt von X nicht nach S variiert. Ist diese Annahme aber falsch, da S den Effekt von X auf Y moderiert, ist β_1 kein unverzerrter Schätzer des Kausaleffekts mehr. Falsche parametrische Annahmen können also dazu führen, dass ein identifizierter Effekt nicht unverzerrt geschätzt werden kann. Bei der Modellierung kommt es daher stets zu einem Trade-off zwischen Verletzung der Positivität und der Verletzung parametrischer Annahmen. Es sollte somit stets eine Modellierung gewählt werden, die ausreicht, um Positivität zu gewährleisten, darüber hinaus aber nicht zu viele parametrische Annahmen macht. Eine Regression kann beispielsweise dadurch flexibilisiert werden, für Variablen mit wenigen Ausprägungen Indikatoren (*dummies*) aufzunehmen, statt die Variable in metrischer Form ins Modell aufzunehmen aufzunehmen. Weitere Möglichkeiten bietet die Aufnahme von Polynomen oder Splines.

3.3 Zufallsfehler und statistische Inferenz

Neben dem Problem der Verletzung von Positivität ist die Schätzung kausaler Effekte mit realen Daten mit dem allgegenwärtigen Problem des Zufallsfehlers und der Verallgemeinerung von Stichprobendaten auf die Zielpopulation konfrontiert.

Ebenso wie Positivitätsverletzungen nimmt der Einfluss von Zufallsfehlern mit sinkender Fallzahl zu. Dies kann man sich leicht vergegenwärtigen, wenn man sich vorstellt, dass eine einzige (durch Zufall) abweichende Beobachtung viel größeren Einfluss hat, wenn insgesamt nur sehr wenige Beobachtungen vorliegen, die diese Abweichung wieder „aufwiegen“ können. In Stichproben kann dies dazu führen, dass der den Kausaleffekt identifizierende bedingte Zusammenhang nicht dem kausalen Effekt in der Zielpopulation entspricht, sondern durch Zufall davon abweicht. Hierbei handelt es sich um das übliche Problem statistischer Inferenz, dass man nicht vollständig lösen kann, durch den Einsatz von Konfindenzintervallen aber quantifizieren kann.

3.3.1 Konfidenzintervalle für Kausaleffekte

Die statistische Inferenz, also das Schließen von Stichprobendaten auf unbekannte Populationsdaten hat an sich nichts mit kausaler Inferenz zu tun. Liegen aber wie in den meisten Fällen keine Populationsdaten vor, kann der statistische Zusammenhang, der einen Kausaleffekt identifiziert, nicht direkt in der Population berechnet werden, sondern muss aus Stichprobendaten geschätzt werden. Wieder gilt, identifiziert der Zusammenhang in der Population bereits den Kausaleffekt nicht, wird auch der Zusammenhang in der Stichprobe dies nicht tun, egal wie hoch oder niedrig die statistische Signifikanz des Schätzwerts oder die Breite des dazugehörigen Konfidenzintervalls. So gibt beispielsweise das 95%- Konfidenzintervall an, wie weit der Stichprobenschätzwert für einen statistischen Zusammenhang in 95 von 100 Stichproben der Fallzahl n streuen würde, wenn diese erhoben würden. Identifiziert dieser Zusammenhang aber den interessierenden Kausaleffekt *nicht*, überdeckt dieses Konfidenzintervall den Kausaleffekt mit einer Wahrscheinlichkeit kleiner als 95%. Diese kann sogar Null werden, wenn eine starke Verzerrung vorliegt, wenn es also besonders viele oder besonders starke offene nichtkausale Pfade zwischen X und Y gibt.

3.3.2 Statistische vs. inhaltliche Signifikanz

Auch die statistische Signifikanz eines geschätzten Zusammenhangs sagt nichts darüber aus, ob dieser einen kausalen Effekt identifiziert und schon gar nicht, wie stark dieser ist. Die statistische Signifikanz in Form des p-Werts gibt lediglich die Wahrscheinlichkeit an, mit der in einer Stichprobe ein statistischer Zusammenhang gefunden wird, wenn der wahre statistische Zusammenhang in der Population gleich Null ist. Somit wird das Risiko quantifiziert, einen statistischen Zusammenhang zu finden, der in der Population gar nicht existiert. Wenn der statistische Zusammenhang in der Population ungleich Null ist, ist dieser Standardtest vollkommen ohne Aussagekraft. Die Stärke eines Zusammenhangs wiederum ist

ausschließlich an Hand der Größe des Schätzwerts inhaltlich zu bewerten. Diese Bewertung hängt selbstverständlich von der Einheit ab, in der das Outcome (und somit auch der Zusammenhang) gemessen ist. Selbst statistische Zusammenhänge mit verschwindend geringem p-Wert können von geringer inhaltlicher Signifikanz sein. Das liegt darin, dass bei ausreichend großer Fallzahl, jede kleinste Abweichung eines Zusammenhangs von Null statistisch signifikant wird. Umgekehrt können inhaltlich signifikante, also starke Zusammenhänge, bei relativ geringer Fallzahl nicht die geläufigen Kriterien für statistische Signifikanz erfüllen.

3.4 Schätzung vs. Identifikation

In der kausalen Inferenz ist die Identifikation kausaler Effekte klar von deren Schätzung zu trennen. Identifikation geht dabei stets der Schätzung logisch voraus. Denn es muss zunächst geprüft werden, ob es überhaupt möglich ist, mit dem gegebenen Kausalmodell und den vorhandenen Daten den Kausaleffekt (nichtparametrisch) zur berechnen. Die Frage ist dabei stets, ob es durch die Konditionierung auf in den Daten vorhandenen Variablen möglich ist, alle nichtkausalen Pfade zwischen X und Y zu schließen und dabei alle kausalen Pfade offen zu lassen. Ist dies nicht möglich kann durch einfache Konditionierung keine Methode einen unverzerrten Schätzer des Kausaleffekts liefern, selbst wenn Populationsdaten mit unendlich großen Fallzahlen vorlägen.

Bei der Schätzung kommen dann zwei weitere Probleme hinzu. Einerseits die mögliche Verletzung der Positivität und andererseits die Unsicherheit bei der Schätzung mit Stichprobendaten. Begegnet man Verletzungen der Positivität durch statistische Modellierung, kann es passieren, dass durch fehlerhafte parametrische Annahmen über die funktionale Form der zu schätzenden Zusammenhänge nichtparametrisch identifizierte Effekte verzerrt geschätzt werden. Durch Zufallsfehler kann es besonders in kleinen Stichproben dazu kommen, dass ein statistischer Zusammenhang, der mit Populationsdaten den interessierenden Kausaleffekt identifizieren würde, durch Stichprobendaten geschätzt, weit vom wahren Kausaleffekt entfernt liegt.

Der fundamentale Unterschied dieser Schätzprobleme mit Identifikationsproblemen ist, dass Schätzprobleme im Prinzip durch eine Vergrößerung der Fallzahl gelöst werden können. Ist ein Effekt jedoch nicht identifiziert, helfen auch unendlich viele Beobachtungen nichts. Dann benötigt man schlicht Daten über andere bzw. mehr Variablen. Hoffnungen, dass Probleme kausaler Inferenz alleine durch größere Datensätze, wie sie beispielsweise die viel zitierten *big data* darstellen, gelöst werden, sind somit absolut unbegründet. Wenn in diesen Daten offene nichtkausale Pfade nicht durch Konditionierung geschlossen werden können, ist es auch mit Abermillionen von Beobachtungen nicht möglich, kausale Effekte (ohne unplausible Adhoc-Annahmen) zu schätzen.

3.5 Zum Weiterlesen

Eine sehr anschauliche und klare Behandlung statistischer Modellierung und statistischer Inferenz im Rahmen kausaler Inferenz bieten die Kapitel 11 und 10 in Hernán und Robins (2016). Dort und auch in weiteren Einführungsbüchern zu kausaler Inferenz finden sich Darstellungen zahlreicher parametrischer Schätzmethoden mit ihren jeweils spezifischen Stärken und Schwächen. Die Annahme der Effekthomogenität in Standardregressionen diskutieren Elwert und Winship (2010) sowie Morgan und Todd (2008). Petersen und van der Laan (2014) liefern eine überzeugende Anleitung zu der Verzahnung von Identifikation und Schätzung kausaler Effekte. Bei Krämer (2011) findet sich eine Abhandlung zu häufigen Fehlinterpretationen statistischer Signifikanz sowie ihrer korrekten Verwendung.

Literaturverzeichnis

Angrist, J. D. und Lavy, V. (1999). Using Maimonides' Rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, 114(2), 533–575.

Angrist, J. D. und Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.

Angrist, J. D. und Pischke, J.-S. (2015). *Mastering 'Metrics. The Path from Cause to Effect*. Princeton University Press, Princeton, NJ.

Box, G. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, 74(365), 1–4.

Calonico, S., Cattaneo, M. D., und Titiunik, R. (2014). Robust data-driven inference in the regression-Discontinuity design. *Stata Journal*, 15(2), 1–36.

Campbell, D. T. und Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Rand McNally, Chicago.

Christakis, N. A. und Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4), 370–379.

Cook, T. D. und Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Rand McNally, Chicago.

Dawid, A. P. (1979). Conditional Independence in statistical theory. *Journal of the Royal Statistical Society Series B-Methodological*, 41(1), 1–31.

Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan, Hrsg., "Handbook of Causal Analysis for Social Research," S. 245–272. Springer, New York.

Elwert, F. und Winship, C. (2010). Effect Heterogeneity and Bias in Main-Effects-Only Regression Models. In R. Dechter, H. Geffner, und J. Y. Halpern, Hrsg., "Heuristics, Probability and Causality: A Tribute to Judea Pearl," S. 327–336. College Publications, England.

Elwert, F. und Winship, C. (2014). Endogeneous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, 40, 31–53.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology*, 36, 21–47.

Glymour, C. (2001). *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. MIT, Cambridge, MA.

Glymour, M. M. (2006a). Natural experiments and instrumental variable analyses in social epidemiology. In J. M. Oakes und J. S. Kaufman, Hrsg., "Methods in Social Epidemiology," S. 393–428. Jossey-Bass, San Francisco, CA.

Glymour, M. M. (2006b). Using causal diagrams to understand common problems in social epidemiology. In J. M. Oakes und J. S. Kaufman, Hrsg., "Methods in Social Epidemiology," S. 393–428. Jossey-Bass, San Francisco, CA.

Glymour, M. M., Tchetgen Tchetgen, E. J., und Robins, J. M. (2012). Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions. *American Journal of Epidemiology*, 175(4), 332–339.

Greenland, S., Pearl, J., und Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1), 37–48.

Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, 35, 1–97.

Hernán, M. A., Hernández-Diaz, S., und Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5), 615–625.

Hernán, M. A. und Robins, J. M. (2006). Instruments for causal inference. An epidemiologist's dream? *Epidemiology*, 17(4), 360–372.

Hernán, M. A. und Robins, J. M. (2016). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396), 945–960.

Hong, G. (2015). *Causality in a Social World: Moderation, Mediation, and Spill-Over*. Wiley-Blackwell, West Sussex, UK.

Imbens, G. W. und Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. und Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1), 5–86.

Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis*, 23(3), 313–335.

Kern, H. L. und Hainmueller, J. (2009). Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes. *Political Analysis*, 17(4), 377–399.

King, G., Keone, R., und Verba, S. (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press, Princeton, NJ.

Kirk, D. S. (2009). A Natural Experiment on Residential Change and Recidivism: Lessons from Hurricane Katrina. *American Sociological Review*, 74(3), 484–505.

Knight, C. und Winship, C. (2013). The causal implications of mechanistic thinking: Identification using directed acyclic graphs (DAGs). In S. L. Morgan, Hrsg., "Handbook of Causal Analysis for Social Research," Handbooks of Sociology and Social Research. Springer, Dordrecht u.a.

Krämer, W. (2011). The cult of statistical significance – What economists should and should not do to make their data talk. *Schmollers Jahrbuch*, 131, 455–468.

Lee, D. S. und Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2), 281–355.

Loeffler, C. E. und Grunwald, B. (2015). Processed as an adult: A regression discontinuity estimate of the crime effects of charging nontransfer juveniles as adults. *Journal of Research in Crime and Delinquency*, OnlineFirst.

Morgan, S. L. (2013). *Handbook of Causal Analysis for Social Research*. Handbooks of Sociology and Social Research. Springer, Dordrecht u.a.

Morgan, S. L. und Todd, J. J. (2008). A diagnostic routine for the detection of consequential Heterogeneity of causal effects. *Sociological Methodology*, 38, 231–281.

Morgan, S. L. und Winship, C. (2012). Bringing context and variability back into causal analysis. In H. Kincaid, Hrsg., "Oxford Handbook of the Philosophy of the Social Sciences.", S. 319–354. Oxford University Press.

Morgan, S. L. und Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research. Second Edition*. Cambridge University Press, Cambridge, UK.

Murnane, R. J. und Willett, J. B. (2010). *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Translated in Statistical Science*, 5(4), 465–472.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, 82(4), 669–710.

Pearl, J. (2005). Direct and indirect effects. *Proceedings of the American Statistical Association Joint Statistical Meetings*, S. 1572–1581.

Pearl, J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, 3, 96–146.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference. Second Edition*. Cambridge University Press, Cambridge.

Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, 40(1), 75–149.

Petersen, M. L. und van der Laan, M. J. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, 25(3), 418–426.

Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, 12(3), 313–320.

Robins, J. M. und Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2), 143–155.

Robins, J. M. und Hernán, M. A. (2009). Estimation of the Causal Effects of Time-Varying Exposures. In F. G, M. Davidian, G. Verbeke, und G. Molenberghs, Hrsg., "Longitudinal Data Analysis. Handbooks of Modern Statistical Methods," Chapman & Hall/CRC, Boca Raton (Florida).

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer, New York.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5), 688–701.

Shadish, W., Cook, T., und Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Cengage Learning, Belmont, CA.

Shalizi, C. R. (2016). *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, New York. URL http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

Shalizi, C. R. und Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, 40(2), 211–239.

Shpitser, I., VanderWeele, T. J., und Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In "Proceedings of the 26th conference on Uncertainty and Artificail Intelligence," S. 527–536. AUAI Press, Corvallis.

Sobel, M. E. (1998). Causal inference in statistical models of the process of socioeconomic achievement. *Sociological Methods & Research*, 27(2), 318–348.

Sobel, M. E. (2000). Causal inference in the social sciences. *Journal of the American Statistical Association*, 95(450), 647–651.

Spirtes, P., Glymour, C., und Scheines, R. (2001). *Causation, Prediction, and Search. Second Edition*. MIT Press, Cambridge, MA.

Steiner, P. M., Kim, Y., Hall, C. E., und Su, D. (2015). Graphical models for quasi-experimental Designs. *Sociological Methods & Research*, Advance Access, May 14, 2015.

Swanson, S. A. und Hernán, M. A. (2013). How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, 24(3), 370–374.

VanderWeele, T. J. (2009). On the distiction between interaction and effect modification. *Epidemiology*, 20(6), 863–871.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York.

VanderWeele, T. J. und Robins, J. M. (2007). Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs. *Epidemiology*, 18(5), 561–568.

Vardardottir (2013). Peer effects and academic achievement: A regression discontinuity approach. *Economics of Education Review*, 36, 108–121.

Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data. Second Edition*. MIT Press Books.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(7), 557–585.

Xie, Y. (2013). Population heterogeneity and causal inference. *Proceedings of the National Academy of Sciences of the United States of America*, 110(16), 6262–6268.