Wie artifiziell müssen Daten sein, damit sie für die Wirklichkeit relevant werden

MANFRED BOROVCNIK, KLAGENFURT

Unsere Gesellschaft ist eine durch und durch evidenz-basierte geworden. Argumente sind umso schwächer, je weniger man auf Daten zurückgreifen kann. Daten werden zum Angelpunkt für Entscheidungen, neue Erkenntnisse werden daraus geschöpft. In der Medizin, in der Politik, in der Wirtschaft, ja selbst in den Wissenschaften sind Daten unverzichtbar, was einerseits einen Paradigmenwechsel signalisiert, andererseits aber die „Verwundbarkeit" empirischer Wissenschaften anzeigt, wenn sie über die Evidenz aus Daten hinaus keinen theoretischen Rahmen entwickeln. In diesem Beitrag geht es weniger um Kompetenzen im Umgang mit Daten sondern um einen philosophischen Aspekt: Wie artifiziell müssen Daten sein, damit sie als Basis für Evidenz tauglich werden? Während in Borovcnik (2014) der Akzent ganz auf dem philosophischen Aspekt der Artifizialität von Daten liegt, stehen hier auch Fragen einer statistischen Allgemeinbildung (statistische Literalität) im Blickpunkt.

0. Vorbemerkungen

Die Statistik hat ihren Ursprung in der Beschreibung von Staaten und der Erfassung von Risiken bei Versicherungsgeschäften. Hier sind Daten ein zentraler Angelpunkt. In den Naturwissenschaften haben Daten seit der Renaissance eine zunehmende Rolle gespielt, allerdings eingebettet in eine Hintergrundtheorie. Diese Theorie hält ein kohärentes System von Beziehungen zusammen. Paradebeispiel ist natürlich die Mathematik, wo Daten dienend einem logischen System untergeordnet sind. Mit dem „Aufstieg" der Geisteswissenschaften verbunden ist ein Paradigmenwechsel. Wissen wird genuin durch Daten *erzeugt*. Neben die Begründung von Zusammenhängen durch (logische) Argumente tritt substantiell evidenz-basiertes Wissen; Wissen, das durch Erkennen von Mustern (oder Besonderheiten) in Daten gewonnen wurde. Diese Art von Wissen ersetzt mehr und mehr Argumente und logische Zusammenhänge. Daten gewinnen daher eine geradezu entscheidende Rolle, in allen Bereichen des Lebens und in den Wissenschaften. Die Medizin ist hierfür ein Musterbeispiel. Daten aus Experimenten entscheiden, ob eine bestimmte Behandlung anerkannt wird oder nicht.

Wenn wir feststellen, dass Rauchen Lungenkrebs verursacht, so beziehen wir uns keineswegs kausal auf die physiologischen Auswirkungen des Rauchens und Wissen darüber, wie Rauchen die Zellen derart schädigt, dass Dysplasien und schließlich invasiver Krebs entstehen. Wir beziehen uns auf Daten, welche zeigen, dass unter jenen, die an Lungenkrebs erkrankt sind, der Anteil an Rauchern größer ist als unter jenen, die keinen Lungenkrebs haben. Wie schwierig dieses evidenz-basierte Wissen zu werten ist, sieht man nicht nur am prolongierten gesellschaftlichen Streit darüber, ob Rauchen zu verbieten (oder wenigstens zu verbannen) ist. Nein, wir sehen das auch daran, dass wir einerseits immer wieder Leute antreffen, die, obwohl sie starke Raucher waren, ein hohes Alter erreicht haben, andererseits, dass wir einfach nicht ausschließen können, dass ein sogenannter Dritt-Faktor sowohl eine stärkere Neigung zum Rauchen provoziert als auch das Risiko, an Lungenkrebs zu erkranken, erhöht. Nennen wir diesen Faktor „Morbidität" (aufgefasst als Gegensatz zu Vitalität), so wäre die Spekulation, dass hohe Morbidität (geringe Vitalität) es diesen Menschen schwerer macht, mit Stress umzugehen (und die daher zu rauchen beginnen), und die anfälliger für verschiedene Erkrankungen (u.a. auch für Lungenkrebs) sind.

Nach dem modernen Standard müssten wir ein Experiment durchführen, das etwa so aussieht: in früher Jugend werden Personen zufällig einer der beiden Gruppen (Raucher, Nicht-Raucher) zugeordnet. Die „Raucher" müssten dann eine bestimmte Zahl an Zigaretten rauchen, die „Nicht-Raucher" dürften nie rauchen. So ein Experiment verbietet sich klarerweise aus ethischen Gründen. Es würde zu artifiziellen Daten führen im Vergleich zu den Daten, die wir einfach so bekommen haben; erst diese

Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), Heft 47, 2014, S 1 - 16

artifiziellen Daten brächten mehr Klarheit in die Zusammenhänge zwischen Rauchen und Lungenkrebs.

Artifiziell soll dabei nicht als Fachbegriff verstanden sein, sondern in „natürlicher“ Weise aufgefasst werden. Mit dem Begriff *artifizielle Daten* unterstellen wir, dass Daten keineswegs in einem natürlichen Prozess, also sozusagen in „freier Wildbahn“ entstehen. Daten werden einerseits immer erst durch Vorgaben erfassbar, andererseits werden sie unter Einschränkungen „produziert“. Wie artifiziell müssen Daten sein, damit sie für die Wirklichkeit relevant werden? Man sieht an Daten entweder Nichts oder Falsches. Daten sprechen nie für sich selbst. Immer müssen sie eingebettet werden in einen Rahmen. Dabei spielen der Kontext, aus dem die Daten stammen, die Eigenheiten der Werkzeuge und – gewissermaßen – die Artifizialität der Daten eine tragende Rolle. Bilder, d.h., Visualisierungen der Daten, erhöhen die Überzeugungskraft der Daten ungemein. Wir orientieren uns dabei an folgenden drei Thesen, die in den nächsten Abschnitten jeweils aufgegriffen werden:

1 Reale Daten erscheinen als nicht hinterfragbare Fakten.

2 Reale Daten werden erst durch einen artifiziellen Vergleich interpretierbar; sie gewinnen an Verbindlichkeit, wenn sie artifiziell entstehen.

3 Artifiziell erzeugte Daten illustrieren Modelle und verleihen ihnen wegen # 1 den Charakter von Fakten.

Im Grunde sind alle Daten künstlich. Das Wesen mathematischer Anwendung ist, dass sie auf Modellen beruht, welche ihrerseits auf Annahmen zurückgreifen müssen. Wir erfahren die Welt vermittelt. Je weiter wir von der eigenen Erfahrung und den eigenen Daten weggehen, umso stärker wird der Grad an Artifizialität. Die Chancen der statistischen Vorgangsweise liegen nun darin, dass die Daten über das subjektive Empfinden hinausweisen und verallgemeinerbare Züge erhalten, welche mehr Verbindlichkeit schaffen können.

1. Zur Bedeutung von Kontext und den statistischen Werkzeugen

Reale Daten erscheinen als nicht hinterfragbare Fakten (#1). Daten werden normalerweise aufbereitet (in Tabellen und insbesondere in Graphiken). Wenn jemand Daten anbietet, so werden diese bereits in bestimmter Absicht mitgeteilt, man zeigt, was man zeigen will (1.1). Andererseits ist man selbst nicht davor gefeit, dass man in den Daten (in ihrer Darstellung) nur das sieht, was man eigentlich sehen will (1.2); man sucht danach, sein eigenes „Vorurteil“ zu bestätigen. Damit man in Daten (bzw. in der Darstellung dieser) relevante Muster erkennen kann, muss man den Kontext, aus dem die Daten stammen, gut kennen (1.3). Man muss aber auch die statistischen Werkzeuge und deren Eigenheiten gut kennen (1.4 und 1.5), will man vermeiden, Artefakte der Methode als sachliche Erkenntnisse über das mit den Daten beschriebene Phänomen falsch zu interpretieren. Als weitere Gefahr stellt sich ein, dass man über bestimmte Variable keine Daten hat und daher die Zusammenhänge falsch interpretiert. Man muss diese potentiellen Zusammenhänge schon vor der Datenproduktion abklären, sodass man auch diese Daten beschafft, damit man die Daten im richtigen Licht sehen kann (1.6).

1.1 Man zeigt, was man zeigen will

Ein beliebtes Spiel in der Politik sowie in der begleitenden Darstellung in den Medien ist, das zu zeigen, was man will. Unabhängig davon, wie gut ein Programm der verbindlichen Anmeldung zur Krankenversicherung ist, kann man den aktuellen Stand der Anmeldungen so oder so sehen wollen. Befürworter sehen sich knapp vor dem Ziel, wenn man – trotz Schwierigkeiten bei der elektronischen Anmeldung – vier Tage vor Ende der Meldefrist fast 85% der Zielgröße erreicht hat. Gegner zeigen, dass das Programm gescheitert ist und stellen die Anmeldungen wie in Abb. 1 dar.

In Abb. 1 wird durch den einfachen Trick, die zweite Achse nicht bei null beginnen zu lassen, der Eindruck erweckt, als ob man sich etwa auf einem Drittel des Ziels befindet und damit das Programm gescheitert ist. Der Eindruck wird noch verschärft dadurch, dass man die Beschriftung der Säulen *in* die Säulen verlagert hat, sodass die Flächen, die man visuell vergleicht noch ungleicher erscheinen. Fox News ist ja nicht irgendeine billige Sendung. Die Verantwortlichen können doch nicht allen Ernstes geglaubt haben, mit so einem billigen Trick durchzukommen. In diesem Fall hat sich der Sender doch – nach heftiger Kritik – zu einem Austausch der Abbildungen im Internet „herabgelassen". Jedoch, wer hat das gelesen? Und, der erste Eindruck bleibt.

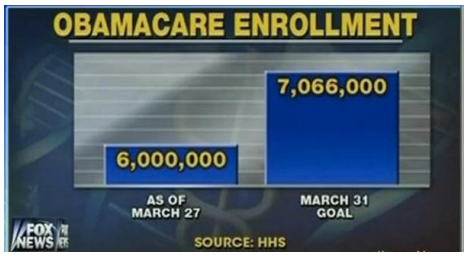

Abb. 1: Fox-News vom 27.3.2014 über den Stand der Anmeldungen für die neue medizinische Versicherung. Gezeigt werden die Anzahl der Anmeldungen und das erklärte Ziel. Es wird der Anschein erweckt, dass man von diesem weit weg ist.

Auch nach mehr als 30 Jahren Unterweisung in Statistik hat sich offenbar kaum statistische Literalität eingestellt. Es werden schöne Beispiele zur Manipulation mit statistischen Techniken bearbeitet, aber eine strukturierte Übersicht über Manipulationstechniken steht noch immer aus. So eine Übersicht über Typen der Manipulation ist generisch gemeint und würde auch „neue" Tricks miteinschließen. Jedenfalls ist der derzeitige Ansatz, Manipulationen einfach an Beispielen zu illustrieren, zu wenig effizient. Schon eine einfache Checkliste wäre hilfreich:

- Flächige oder dreidimensionale Darstellungen verzerren die Unterschiede, weil man Flächen oder Volumina sieht; dargestellt werden die Daten aber nach der Höhe oder dem Radius etc.
- Die dreidimensionale Darstellung der Daten (wobei die Daten eigentlich durch die Höhe abgebildet werden) wird räumlich angeordnet, wodurch perspektivische Eindrücke assoziiert werden: was im Hintergrund steht, erscheint wesentlich höher, weil unser räumliches Sehen die Entfernungen automatisch einbaut und korrigierend eingreift.
- Skalenausschnitte verzerren den Vergleich oder die Beurteilung einer Entwicklung.
- Die Zeitachse wird ungleichmäßig unterteilt: ungleiche Zeitabschnitte werden gleich dargestellt; eine logarithmische Skalierung verzerrt die Geschwindigkeit einer dargestellten Entwicklung.
- Zwischen Messzeitpunkten wird die Entwicklung interpoliert, auch wenn sich erst nach einem katastrophalen Ereignis etwas verändert: Meinung zu Kernenergie vor und nach Fukushima etwa.
- Messzeitpunkte werden willkürlich herausgegriffen, nur um den gewünschten Eindruck zu stützen.

1.2 Man sieht, was man sehen will

Wenn man im Bildungssystem verankert ist, möchte man natürlich sehen, dass sich Bildung auch in Berufschancen und Einkommen äußert. Wenn man als Staat ins Schulwesen investiert, so möchte man natürlich argumentieren, dass damit ein größeres Einkommen für jene erzielbar ist, welche die Schulen erfolgreich absolvieren.

Wenn dann Daten (nach Bezirken aggregiert) zeigen, dass die entsprechende Punktwolke ansteigt und die Korrelation mit 0,488 ausgewiesen wird, so möchte man das gerne als Beleg interpretieren für die Richtigkeit der Entscheidung, in Bildung zu investieren (als Staat bzw. als Einzelperson). Wir werden

gleich weiter unten sehen, dass die Dinge komplizierter stehen und sowohl der Kontext als auch die Kenntnis von Eigenheiten des verwendeten statistischen Werkzeugs die Interpretation erheblich beeinflussen (Abb. 2; Daten nach Freedman, Pisani und Purves, 2007).

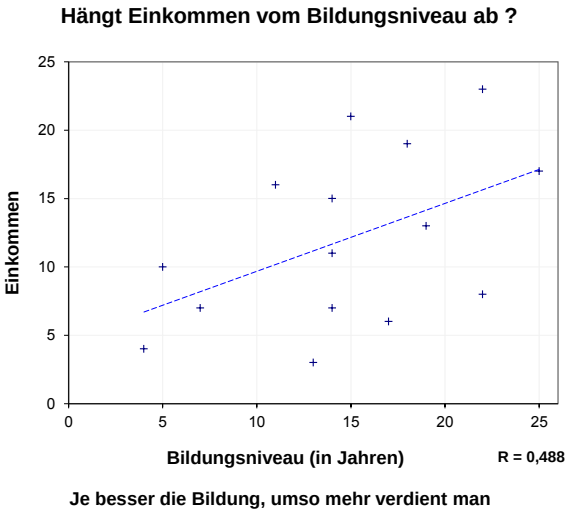

Abb. 2: Höhere Bildung verhilft zu höherem Einkommen. Die Korrelation beträgt 0,488 (auf Daten hinsichtlich des Bezirksniveaus).

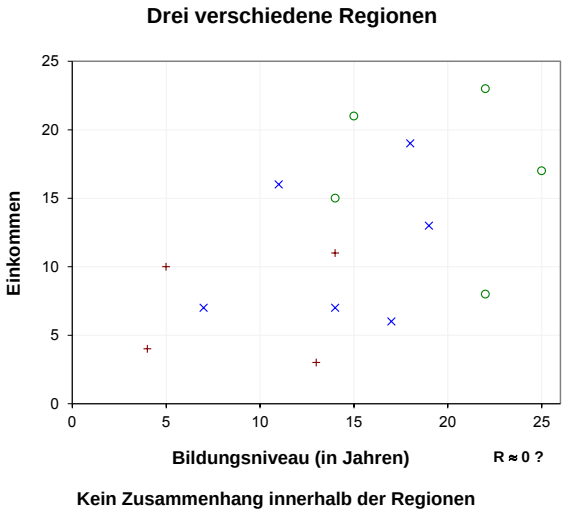

Abb. 3: Aufschlüsselung der Daten nach Regionen zeigt innerhalb der Regionen keinerlei Zusammenhang zwischen Bildungsniveau und Einkommen.

1.3 Man muss den Kontext *und* die weitläufigeren Zusammenhänge kennen, damit man sehen kann

In Abb. 3 zeigt sich, dass die Dritt-Variable Region den Zusammenhang, der zuerst beobachtet werden konnte, auf den Kopf stellt insofern, als es nun – in homogenen Regionen – keinen Zusammenhang zwischen Bildungsniveau und Einkommen gibt.

Es wird auch offenbar, dass man das eben beobachtete Phänomen gar nicht aufklären kann, wenn man keine Daten über die Dritt-Variable Region hat. Wenn man in der Phase der Systemanalyse die Zusammenhänge zwischen den zu untersuchenden Variablen vom Kontext her studiert, muss man die möglichen Zusammenhänge vorausahnen und entsprechend Daten dazu aufzeichnen. Die Daten, so wie sie stehen, müssen also sorgfältig geplant werden, will man solche Verzerrungen in den Griff bekommen und entsprechende Fehlschlüsse vermeiden.

1.4 Man muss das Werkzeug kennen, damit man sehen kann

Wie stark ist der Zusammenhang zwischen den dargestellten Merkmalen in folgenden Punktwolken (Abb. 4)?

Die meisten Personen in unseren privaten Befragungen (auch in Vorträgen vor Statistikern) ordneten die Punktwolken nach sehr großer Korrelation rechts unten und danach oben in der Mitte und sehr geringe Korrelation unten links. Sie waren höchst überrascht, als sie die Achsenteilung sahen und ihnen klar wurde, dass es sich um ein und denselben Datensatz handelt, verzerrt eben durch die Skalierung der Achsen.

Da Korrelation eine dimensionslose Größe ist, also die Eigenheit hat, maßstabsunabhängig zu sein, muss man die Darstellung normieren, wenn man sich visuell über die Größe der Korrelation orientieren möchte. Normierung bedeutet, dass die Punktwolke in etwa in ein Quadrat passt. Ansonsten ist es möglich, jede Größe des Korrelationskoeffizienten durch die Darstellung vorzugaukeln.

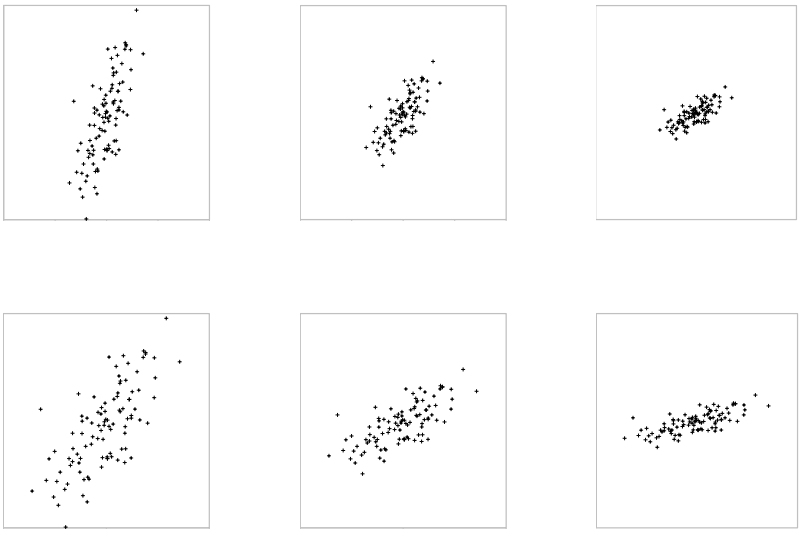

Abb. 4: Sechs verschiedene Punktwolken – wie groß schätzt man visuell die Größe des Korrelationskoeffizienten?

1.5 Artefakte: Aggregierte Daten, Simpson-Paradoxon und Regression zur Mitte

Wenn wir Zusammenhänge zwischen Merkmalen studieren, so sind die Methoden von Regression und Korrelation das Mittel der Wahl. Das genannte Beispiel zu Einkommen und Bildungsniveau kann auch – aggregiert auf Regionen – untersucht werden. Es ergibt sich hier eine Punktwolke mit nur drei Punkten und einer ganz hohen Korrelation von 0,998 (Abb. 5).

Während die drastische Reduktion auf drei Punkte weder Punktwolke noch Korrelation als sinnvoll erscheinen lässt, kann man dieses Beispiel dazu nutzen, um die Auswirkung der Aggregation von Daten auf die Methode der Korrelation zu illustrieren: Aggregierte Daten können eine wesentlich höhere Korrelation aufweisen als individuelle Daten.

Es wird also eine hohe Korrelation allenfalls durch Zusammenfassung von Daten erzeugt und damit zu einem Artefakt. So viel dazu, dass Daten Fakten sind. Aggregierte Daten werden in vielen Bereichen untersucht, etwa auch in der Epidemiologie, wenn die Verbreitung von Lungenkrebs mit der Anzahl der gerauchten Zigaretten in der jeweiligen Region untersucht wird.

Das nächste Beispiel (Abb. 6) zeigt einen Effekt, der im Kontext überraschend ist: je länger Personen studiert haben, umso höher war ihr Einstiegsgehalt in der ersten beruflichen Anstellung. Es deuten sich jedoch in der Punktwolke von Abb. 6 Untergruppen an. Wenn man keine Daten über jene Merkmale hat, welche diese Untergruppen charakterisieren, kann man das Phänomen – es wird Simpson-Paradoxon genannt – nicht näher erklären. Es verbleibt reine Spekulation, was die Untergruppen auszeichnet.

Ein weiteres Artefakt ist die so genannte *Regression zur Mitte*. Danach sind bei verbundenen Daten (etwa Körpergröße von Vätern und Söhnen) die zweiten Werte (Söhne) im Durchschnitt weniger extrem als die ersten (Väter). Dubben und Beck-Bornholdt (2010) oder Freedman, Pisani und Purves (2007) zeigen aber sehr schön, dass es sich hierbei um ein reines Artefakt handelt, es also Eigenheit der Methode und nicht ein interessantes Phänomen aus dem Kontext ist. Genetische Vererbung könnte wohl erklären, dass Söhne größer als der Durchschnitt aller sind, wenn deren Vater überdurchschnittlich groß ist. Wie aber sollte man erklären, dass es ein Rückschreiten zum Mittelwert gibt, wonach Söhne im Durchschnitt um einen bestimmten Faktor r ($0 < r < 1$) *weniger größer* sind als der Vater? (Für Väter, deren Größe unterhalb des Mittelwerts liegt, sind die Verhältnisse gegengleich.) Dieses r ist genau der Korrelationskoeffizient und wurde von Galton und Pearson als Reversionskoeffizient

benannt, in der Meinung, sie hätten ein wesentliches Charakteristikum der Vererbung entdeckt. Das Phänomen des Rückschreitens zur Mitte hat der Methode ihren Namen – Regression – verliehen. Es ist aber ein reines Artefakt. Selbst wenn man zwei *unabhängige*, identisch verteilte Zufallsvariable miteinander in Beziehung setzt, bekommt man eine Regression zur Mitte (für ein einfaches Zufallsexperiment dazu siehe Borovcnik und Schenk, 2011).

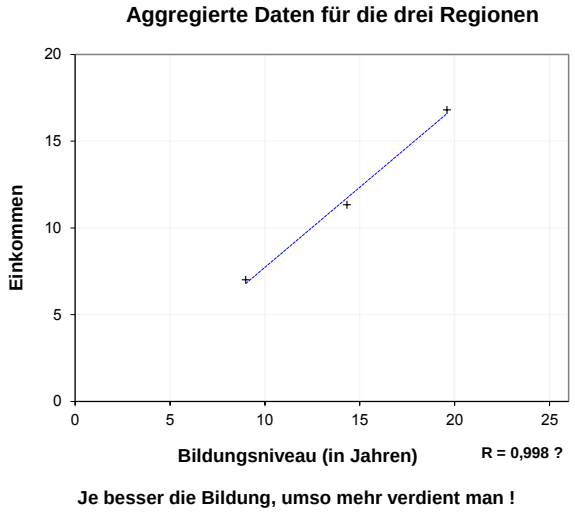

Abb. 5: Aggregation der Daten über Regionen erzeugt einen hohen Zusammenhang zwischen Bildungsniveau und Einkommen mit R = 0,998 als reinem Artefakt.

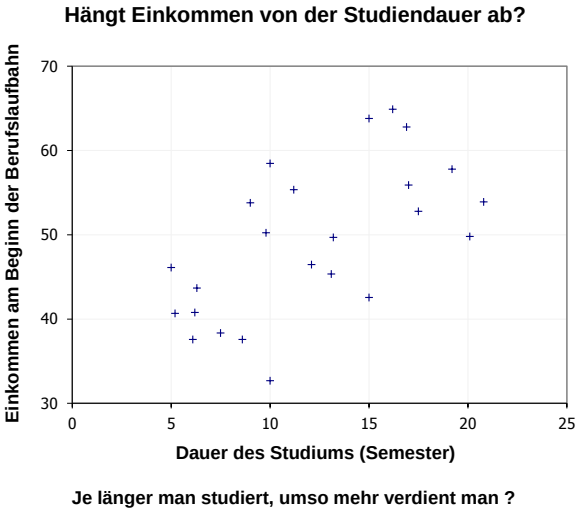

Abb. 6: Einstiegsgehalt im Beruf und Studiendauer. In den Subgruppen ergibt sich ein fallender Zusammenhang, während in der Population der Zusammenhang steigend ist – Simpson-Paradoxon.

1.6 Damit man richtig sieht, muss man die Zusammenhänge *vor* der Datenproduktion klären

Die Dritt-Variable *Studium* (mit den Ausprägungen Betriebswirtschaftslehre, Physik und Chemie) erklärt die Cluster und das Simpson-Paradoxon in Abb. 6. Besonders Chemie hat lange Studienzeiten (rechts oben), bietet aber im beruflichen Einstieg größere Gehälter als BWL (links unten). *Innerhalb* der Studienrichtungen ist aber die Dauer des Studiums durchaus ein Indikator für erfolgreiches Studieren, weil sie mit dem Einstiegsgehalt stark negativ korreliert.

Wenn man in der systemanalytischen Aufarbeitung der Fragestellung vor der Datenproduktion bestimmte Zusammenhänge außer Acht lässt, so sind später keine Daten darüber vorhanden, sodass die bestehenden Daten einen falschen Eindruck erzeugen. Daten sind das Produkt von Vorüberlegungen im Kontext und stellen keineswegs Fakten dar. Schon die Frage, worüber wir Daten haben, macht deutlich, dass wir keine Analysen anstellen können, wenn wir eben über einen Sachverhalt keine Daten haben. Dass über bestimmte Sachverhalte keine Daten verfügbar sind, bestimmt also mit, ob wir zu den einschlägigen Fragen vernünftig Stellung beziehen können oder nicht.

2. Die übliche Auffassung von Daten als Fakten entspringt einem prinzipiellen Missverständnis

Daten, noch mehr ihre Visualisierungen, erhalten den Charakter von Fakten, jedenfalls so, wie sie oft in der Argumentation verwendet werden oder wie sie bei den Rezipienten ankommen. Reale Daten werden erst durch einen artifiziellen Vergleich interpretierbar; sie gewinnen an Verbindlichkeit, wenn

sie artifiziell entstehen (#2). Gehen wir noch einen Schritt zurück und nehmen wir an, wir hätten über einen Sachverhalt keine Daten und könnten nur auf unsere persönlichen Erfahrungen zurückgreifen. Obwohl diese Erfahrungen ein allgemeines Problem andeuten könnten, können wir dies so nicht behaupten. Wir bräuchten erst Daten dazu.

Wenn Daten da sind und als (Ersatz für ein) Argument gebraucht werden, verleiht die Interpretation von Daten als Fakten rein äußerlich Autorität: „Die Fakten sprechen für …“ Allerdings sind diese Daten ja nicht einfach so entstanden, sondern wurden erst durch eine bestimmte Art der Definition der Merkmale und die Erfassung der Daten bei den statistischen Einheiten ermöglicht (2.1), was nicht nur die Lesbarkeit der Daten sondern auch deren Relevanz verändert. Die Beurteilung von Daten wird erst durch einen Vergleichsmaßstab möglich (2.2); der Vergleichsmaßstab wird meist vorgegeben und kann nachhaltig die Interpretation der Daten beeinflussen. Es werden oft – anstelle der unübersichtlichen Fülle der Information in den Daten – weitere Kennziffern aus den Daten berechnet, welche als Kriterien zur Beurteilung einer Fragestellung dienen (2.3); diese Kriterien sind häufig so gewählt, dass die passenden Schlüsse aus den Daten nahe gelegt werden, und man übersieht gerne, dass es auch andere Kriterien gibt, welche ganz andere Schlüsse ziehen lassen.

Man behauptet ferner, die vorhandenen Daten sind repräsentativ für das untersuchte Geschehen, ein Zauberwort, das den Daten automatisch Relevanz zueignet (2.4); in der Statistik wird auch „gezeigt“, dass eine Zufallsstichprobe repräsentativ ist, also behauptet man, man hat eine Zufallsstichprobe und schon sind die Daten aufgewertet.

Die Interpretation von Daten (und abgeleiteten Kennziffern wie der Korrelation) erfordert eine Kombination von Wissen aus dem Kontext und über die statistischen Methoden (2.5); Statistik ist kein Ersatz für Sachwissen; Sachwissen allein reicht nicht aus, die Ergebnisse zu verstehen.

Es gibt ferner eine naive Vorstellung davon, was empirische Evidenz bedeutet (2.6); empirische Evidenz wird mit Beweis im Sinne der Logik verwechselt. Ein signifikantes Ergebnis aus Daten wird uminterpretiert in einen Beweis des dahinterstehenden Phänomens und das Attribut „statistisch“ wird gerne weggelassen. Findet man in Daten ein statistisch signifikantes Ergebnis durch langes Herumsuchen (Data Mining), so ist das keine „Bestätigung“ eines Phänomens (2.7); erst eine Replikation könnte mehr Klarheit bringen.

2.1 Definition und Entstehung der Daten

Wenn Daten durch eine Befragung entstehen, so können sie ganz einfach durch die Art der Fragestellung beeinflusst werden. Es spielt dann auch eine Rolle, wer wen befragt, insbesondere wenn es um Selbstauskünfte geht (etwa, wie haben Sie sich im Experimentalkurs gefühlt, anstatt die Leistung zu messen).

Daten hängen von der Definition ab, welche variieren kann. So etwa gibt es unterschiedliche Festlegungen von Arbeitslosigkeit zwischen Ländern; obendrein werden die Definitionen häufig geändert, so dass zwischen Ländern und über die Zeit hinweg kaum ein Vergleich möglich ist. Die Definition von Armut hingegen ist relativ zur Einkommensverteilung in einem Staat, sodass sich der Anteil an Armen nicht ändert, wenn alle um einen bestimmten Betrag mehr verdienen, oder alle eine prozentuelle Lohnerhöhung bekommen. Wie soll man dann Armut zwischen Staaten mit ganz anderem Lohnniveau vergleichen, oder in einem Staat die Entwicklung von Armut über lange Zeiten hinweg bewerten?

Vorhersagen, die auf speziellen Annahmen beruhen, werden als faktische Entwicklung präsentiert. So etwa werden Klimamodelle simuliert und die Daten als reale Entwicklung dargestellt. Die Annahmen gehen unter. Was, wenn sie nicht zutreffen?

Wenn es keine Daten zu einem Problem gibt, dann existiert es nicht. Persönliche Erfahrung wird als anekdotisch abgetan, weil das Ergebnis ja nicht aus einer gezielten Studie stammt.

2.2 Bewertung von Daten im Vergleich

Ohne Referenz sagen reale Daten wenig aus. Daten sind nur sinnvoll, wenn sie mit einem geeigneten Maßstab verglichen werden: Als Vergleich können persönliche Erwartungen, Vergleichswerte anderer Gruppen, Werte zu anderen Zeitpunkten etc. herangezogen werden. Dabei besteht eine große Offenheit und Beliebigkeit. Das so genannte Ankerphänomen lässt einen einmal vorgegebenen Wert zum Vergleich besonders nachhaltig werden. Beispiel: Wenn wir zuerst informiert werden, dass die Verspätung 3 Stunden dauert, so werden wir über eine Verspätung, die letztlich nur eine halbe Stunde beträgt, kaum ein Wort verlieren. In anderen Fällen hätten wir uns über eine Verspätung von einer halben Stunde sehr geärgert. Will heißen, ob Daten so oder so bewertet werden, hängt vom Vergleichsmaßstab ab, und der ist relativ.

Stellt man fest, dass die neue Wohnung 550,- Euro kostet (ohne Nebenkosten), so hat der eine das „Gefühl“, die ist teuer, der andere meint, das ist aber billig. Wir vergleichen etwa mit dem Mietspiegel als Maßstab (Fahrmeir et al., 2004):

- Statistische Analyse von vergleichbaren Wohnungen.
- Einflussfaktoren wie Größe, Lage, Ausstattung etc. werden untersucht.
- Eine Verteilung der Kosten vergleichbarer Wohnungen wird erstellt.
- Damit lassen sich die angegebenen Kosten bewerten.
- Sollte die besagte Wohnung im Top-1%-Bereich dieser Vergleichsverteilung liegen, so kann das etwa vor Gericht zu einer Kürzung der geforderten Miete führen.

Einzelne reale Daten werden erst in einem artifiziell konstruierten Vergleich sinnvoll. Dabei ist neben einem Richtwert auch eine Streuung der Werte maßgebend, welche die übliche (und tolerierbare) Variabilität modelliert.

Daten aus dem Zusammenhang zu reißen ist eine beliebte Manipulationsmethode. Vergleiche entsprechen einer Modellierung der Situation. Auch die Gewinnung der Daten muss modelliert werden. Das schließt ein so genanntes Design der Studie und die darauf aufbauende Einbringung des Zufalls (Zufällige Auswahl bzw. zufällige Zuordnung der Personen zu Versuchsgruppen) mit ein. Den Zufall ins Spiel zu bringen, erhöht paradoxerweise die Verbindlichkeit der entstehenden Daten.

2.3 Nicht die Daten selbst sondern Kriterien, die aus den Daten berechnet werden, dienen zur Beurteilung der Fragestellung

Die Kriterien, nach denen eine Maßnahme evaluiert wird, entscheiden das Ergebnis der Evaluation. So werden aus denselben Daten zum Überleben von Patienten nach einer bestimmten Therapie ganz verschiedene Schlüsse gezogen, je nachdem, welche Erfolgskriterien man anwendet. Lebenserwartung, Fünfjahres-Überlebensraten, die ganze Überlebenskurve, Qualität der Lebensverlängerung, erwartete Lebensdauerverlängerung verglichen mit alternativen Behandlungsmöglichkeiten (eventuell auch die Nicht-Behandlung). Die ganze Überlebenskurve ist komplizierter, speziell wenn man sie auch noch mit der Lebensqualität gewichtet, gibt aber viel bessere Auskunft als etwa die Fünfjahres-Überlebensraten.

Das Problem verschärft sich, wenn der Erfolg von Maßnahmen erst nach vielen Jahren sichtbar werden kann, etwa bei Präventivmaßnahmen im Gesundheitswesen. Ein jüngeres Beispiel ist die HPV-Impfung von jungen Mädchen vor deren geschlechtlicher Aktivität zur Vermeidung von Gebärmutterhalskrebs, der typischerweise nach 45 auftritt. Dann zieht man so genannte Surrogatvariable heran, das sind Merkmale, welche ex post mit dem Auftreten von Gebärmutterhalskrebs assoziiert sind. Auch

hier wieder ein Verwechseln von Assoziation mit Kausalität sowie das Vernachlässigen von plausiblen Dritt-Variablen wie zum Beispiel ein stabiles Immunsystem.

2.4 Der Mythos der repräsentativen Stichprobe

Man kann nicht allen Daten – unbeschadet ihrer Herkunft – trauen. „Trau keiner Statistik, die du nicht selbst gefälscht hast!“ Auch in den Geisteswissenschaften ist man nicht gefeit vor falschen Zuordnungen, wie das Zitat selbst belegt, denn es wurde vielmals – fälschlich (siehe Barke, 2004) – dem britischen Staatsmann Winston Churchill zugeschrieben.

Wenn man bei einer Wahlumfrage in den 1930ern nur jene erreicht, die ein Telefon und Auto haben, braucht man sich nur wenig zu wundern, dass die Wahlprognose völlig danebenliegt. Es gibt zwei wesentliche Gründe für den Flop: die verzerrte Ausgangsbasis (nur Telefon- und Autobesitzer) und der Effekt der Selbstselektion durch freiwillige Rückantwort. In gleicher Weise bekäme man zu einem Mann-Frau-Thema ein wenig repräsentatives Ergebnis, wenn man nur Frauen (oder nur Männer) befragte, aber die Daten auf die gesamte Population hochrechnete.

Damit Daten statistisch zuverlässig werden, müssen sie repräsentativ sein, was üblicherweise durch die Forderung, dass die Daten durch eine Zufallsstichprobe gewonnen werden, eingelöst wird. Wie schwer man sich bei der Einführung von Zufallsstichproben in den statistischen Büros getan hat und wie viel Überzeugungsarbeit geleistet werden musste, kann man in Kiaer (1899) nachlesen; zum eigenartigen Verhältnis von Quoten- und Zufallsstichproben siehe etwa Borovcnik (1992). In der empirischen Forschung ist der Goldstandard ein Experiment, in dem die Wirkung von etwas auf Versuchs- *und* Kontrollgruppe untersucht wird und die Probanden den Gruppen durch Zufall zugeordnet werden (siehe etwa Borovcnik, 2007a und b). Zufallsstichproben sind aber eher die Ausnahme als die Regel, und es wird der Mythos der Zufallsstichprobe wohl gepflegt, aber kaum eingehalten. Was das randomisierte kontrollierte Experiment anbelangt, so hat schon R.A. Fisher festgestellt, dass manche randomisierte Zuordnung so schlecht ist, dass man das eigentliche Experiment gar nicht erst beginnen muss, weil allein die Gruppenaufteilung die Ergebnisse vorweg nimmt. So etwa ist es bei randomisierter Zuordnung möglich, dass viele schwere Fälle in der einen Gruppe „landen“.

2.5 Interpretation erfordert eine Kombination von Wissen aus dem Kontext und über statistische Methoden

Adäquate Interpretation von Daten hängt von Wissen über den Kontext und Verstehen der statistischen Begriffe ab. So wird statistische Korrelation oft mit Kausalität und der Möglichkeit der Steuerung der Zielvariablen verwechselt: Bei der Klimaerwärmung geht man davon aus, dass man die mittlere Temperatur *steuern* kann, wenn man die „unabhängige“ Variable CO_2-Gehalt der Luft verringert. Aus einer *Ko*-Relation (!) wird so unzulässig eine Steuerungsrelation.

Noch schwieriger wird die Interpretation von Daten, wenn aus diesen Wahrscheinlichkeiten berechnet werden. So werden – bei bedingten Wahrscheinlichkeiten – die Bedingung und das bedingende Ereignis einfach ausgetauscht. Bei retrospektiven Studien untersucht man im Nachhinein die Kranken (etwa Hirntumor) und sieht nach, ob diese Personen Jahre zuvor verstärkt bestimmten Risiken ausgesetzt waren (starker Gebrauch von Mobiltelefonen). Die geschätzte Wahrscheinlichkeit, einem Risiko ausgesetzt gewesen zu sein, wird dann mit der Wahrscheinlichkeit gleich gesetzt, unter der Exposition des Risikos (Mobiltelefonieren) die Krankheit (Hirntumor) zu bekommen.

2.6 Empirische Evidenz und Beweis

Überall werden eigens Daten erhoben, um Erkenntnisse zu gewinnen und Entscheidungen zu begründen. In der Medizin hat sich das randomisierte, doppelblinde Experiment als Goldstandard etabliert;

dadurch dass man die beobachteten Unterschiede ausschließlich der Wirkung einer Therapie zuschreibt, kann man neues Wissen absichern. Doppelblind bedeutet dabei folgendes: Der Patient weiß nicht, welche Behandlung er tatsächlich erhält; das soll den so genannten Placebo-Effekt ausschließen, wonach allein die Erwartung einer Behandlung einen Heileffekt (kurzfristig) bewirken kann. Der behandelnde Arzt und jene Personen, welche die Daten aufzeichnen (oder analysieren), werden auch verblindet, d.h., sie wissen nicht, wer welche Behandlung erhält; damit sollen Verzerrungen der Behandlung selbst und der Aufzeichnung (sowie der Analyse) der Wirkung verhindert werden. Ein solcher Ansatz kann durchaus mit ethischen Werturteilen konfligieren. Für die Zulassung einer Behandlung ist er jedoch unverzichtbar. Es wird schwer, solche standardisierten Ergebnisse zu kritisieren.

Im Grunde verengt man die Alternativen zu stark. Die realen Daten sind eigentlich artifiziell und gelten nur unter Einschränkungen des verwendeten Modells (und dann nur mit einem „Restrisiko"). Wenn Daten nach einem besonderen Ritual (statistischer Signifikanztest) als signifikant bezeichnet werden, gilt das im üblichen Verständnis als Beweis, oft irreführend und euphemistisch als statistische Evidenz bezeichnet. Dabei wird die Forschungslogik auf den Kopf gestellt. Statistische Methoden dienen nicht zur Beweissicherung sondern vielmehr zur Trennung von Spreu und Weizen: Wenn etwa eine Korrelation zwischen Intelligenz von Eltern und Kindern als signifikant beurteilt wurde, ist dies nur ein Anlass, von der Substanzwissenschaft her darüber nachzudenken, wie Intelligenz weitergegeben werden kann und ob es plausible Erklärungen für die beobachteten Daten gibt (genetische Vererbung, die wiederum von der Theorie der Umwelteinflüsse in Frage gestellt wird). Die Daten selbst können keine Entscheidung zwischen den rivalisierenden Ansätzen liefern.

2.7 Der neue Mythos der explorativen Datenanalyse und des Data Mining

Explorative Datenanalyse erforscht die Daten (unabhängig von ihrer Herkunft als zufällige Stichprobe) nach allen „Regeln der Kunst". Verschiedene Darstellungen, oft graphisch unterstützt, sollen zwei- und dreidimensionale Variablen aus dem Datensatz auf Auffälligkeiten untersuchen, bis man endlich irgendwelche Muster findet. Diese Ergebnisse werden dann als Fakten präsentiert. Sie sind aber nicht mehr als Hypothesen, die sich erst bei einer Replikation der gesamten Studie bewähren können. Denn sie wurden implizit durch eine Vielzahl von statistischen Tests herausgearbeitet, was zu einer Inflation des alpha-Fehlers führt, das heißt auch, dass es ganz leicht wird, irgendwelche Abweichungen von „Nullhypothesen" (kein Effekt hinsichtlich des untersuchten Phänomens) zu finden (die Macht, neues – neue Muster bzw. Hypothesen – zu entdecken wird groß).

Das Phänomen des Zurechtbiegens ist bei der Aufteilung von Wahlkreisen seit langem bekannt. So etwa war kürzlich Ungarn im Kreuzfeuer der Kritik, weil eine neue Aufteilung der Wahlkreise der gegenwärtigen Regierung die Mehrheit sichern sollte (Axmann, 2014). Wenn man das Wahlverhalten in Wahlsprengeln genau studiert, so kann man – je nach Ausprägung des Mehrheitswahlrechts – die Wahlbezirke so einteilen, dass die favorisierte Partei auch tatsächlich gewinnt, auch wenn sie deutlich schlechter abschneiden würde als zuvor. Das Phänomen ist als Gerrymandering in den USA bekannt geworden (1812, Massachusetts; siehe Dubben und Beck-Bornholdt, 2010, 96ff). Genau das ist es, was beim Data Mining passiert. Man sucht nach Mustern in einem bestehenden Datensatz. Es ist aber unmöglich, Wissen aus Daten allein zu generieren, ohne auf Hypothesen zurückzugreifen.

3. Artifizielle Daten höheren Grades

Artifiziell erzeugte Daten illustrieren Modelle und verleihen ihnen wegen # 1 den Charakter von Fakten (#3). Da Wahrscheinlichkeitsmodelle virtuell und somit schwer nachvollziehbar sind, versucht man seit längerem mittels Simulation fiktive Daten zu erzeugen, welche das abstrakte Modell illustrieren, die Begriffe erklären und die Berechnungen von Wahrscheinlichkeiten umgehen. Frei nach dem

(ambivalenten) didaktischen Grundsatz: Ein konkreter Datensatz sagt mehr als 1.000 Erklärungen zu einem stochastischen Modell. Ambivalent deswegen, weil man, motiviert durch die Absicht, einen schwierigen Sachverhalt zu vereinfachen, leicht übersieht, wie durch die gewählte Vereinfachung der Charakter der Begriffe verändert wird. Wir sprechen von artifiziellen Daten höherer Ordnung. Dazu gehört die Berechnung von Erwartungswerten in komplexeren stochastischen Modellen, welche diese Modelle illustrieren und die komplizierten Berechnungen mit bedingten Wahrscheinlichkeiten umgehen sollen (3.1; natürliche Häufigkeiten; Gigerenzer, 2002). Dazu zählen ferner Daten, die mittels Simulation auf der Basis von Verteilungsannahmen erzeugt werden (3.2), und die Berechnung von Wahrscheinlichkeiten durch Anteile in den simulierten Daten ersetzen sollen. Ein interessanter Ansatz (3.3; Resampling) lässt Daten erzeugen, mit welchen man ohne stochastische Verteilungsannahmen statistische Inferenz „nachzuspielen“ versucht (Borovcnik, 2006).

Mit artifiziellen Daten höherer Ordnung kann man – mindestens ebenso gut – manipulieren wie mit realen Daten. Denn leicht übersieht man die Einschränkungen (die Annahmen hinter den Hypothesen), unter welchen sie erzeugt wurden. Die Faszination von Daten als – nicht hinterfragbare – Fakten verleitet zu einer Überinterpretation der Ergebnisse bzw. der verwendeten Methoden. Leichtfertige Illustration von Modellen durch simulierte Daten kann daher die abstrakten Begriffe stark verzerren. Auch die statistische Beurteilung durch Resampling hat einen gänzlich anderen Charakter als die induktive Logik der statistischen Inferenz.

3.1 Umsetzen von Wahrscheinlichkeiten in Erwartungswerte

Speziell bei Aufgaben mit der Bayes-Formel bietet die Umsetzung von (bedingten) Wahrscheinlichkeiten in Erwartungswerte eine drastische Vereinfachung der Berechnungen und der Interpretation. Dabei werden Wahrscheinlichkeiten durch ihre Erwartungswerte bei einer Gesamtheit von 100 (1.000) Einheiten veranschaulicht. Allerdings wird dadurch der Charakter von Wahrscheinlichkeitsaussagen und Modellen (die eben passen oder auch nicht) verändert und der Umstand verschleiert, dass wir allenfalls ungenaue Schätzungen über die zugrunde gelegten Wahrscheinlichkeiten haben.

Statistische Dörfer und natürliche Häufigkeiten

Wahrscheinlichkeiten sind indirekte Aussagen. „Was passiert, wenn die Welt ein Dorf mit 100 Menschen wäre, ...“ verdeutlicht solche Aussagen. Eine Wahrscheinlichkeit von ½ wird damit zu 50:50; d.h., wenn wir eine faire Münze 100 Mal werfen, so sollten 50 auf Wappen und 50 auf Zahl enden. Wenn die Welt ein Dorf mit 1.000 Leuten wäre, so wären 60 Nordamerikaner, 80 Südamerikaner, 564 Asiaten, 86 Afrikaner, 210 Europäer, ...; davon ausgehend kann man weitere Eigenheiten darstellen (Glaube, Geschlecht etc.).

Kein Vorteil ist aber ohne Nachteil: Wir verschieben den Charakter von einer Modellaussage (die vielleicht ungenau ist, deren Wert etwa aus anderen Daten geschätzt wurde) hin zu einer faktischen Aussage: 0,01 wird zu einer von hundert. Hier fehlt Spielraum für Abweichungen und schwierig wird es, diese Zahl als Modellgröße statt als Faktum aufzufassen.

In der Jurisprudenz sowie in der Medizin, aber auch in Medien tauchen immer wieder Bayes-Probleme auf, in denen anhand von Indizien ein Sachverhalt neu beurteilt wird. Man kann über damit zusammenhängende Fehlvorstellungen sehr viel lesen (etwa Gigerenzer, 2002). Das folgende Beispiel ist aus Borovcnik (2014).

Ist Mammographie-Screening zur Diagnose bzw. zur Prävention von Brustkrebs geeignet? Von 40-50jährigen Frauen bekommen 0,8% Brustkrebs (Prävalenz). Über die Zuverlässigkeit der Diagnose gehen wir von folgenden Zahlen aus: Hat eine Frau Brustkrebs, so zeigt das ein Mammogramm mit einer Zuverlässigkeit von 90% an. Wenn sie in Wirklichkeit gesund ist, besteht ein Risiko von 7% für ein positives Mammogramm. (Die Angaben bezüglich des Risikos falsch-positiver Diagnosen sind

sehr unterschiedlich. Darüber gibt es kaum zielgerichtete Studien.) Wie groß ist die Wahrscheinlichkeit, dass eine Frau tatsächlich Brustkrebs hat, wenn der Mammographie-Befund positiv ist? Selbst Gynäkologen schätzen diese Wahrscheinlichkeit als sehr hoch ein (viele liegen über 90%; jedenfalls in einem Workshop mit 36 Primarärzten, darunter einigen Gynäkologen). Die mit der Bayes-Formel berechnete Wahrscheinlichkeit von ca. 9% überzeugt kaum einen; die Ärzte bleiben bei ihrer ersten Einschätzung.

Alle Wahrscheinlichkeiten werden zuerst in erwartete Anzahlen in einem Dorf von 1.000 umgerechnet. Diese artifiziellen Daten können zur Veranschaulichung in einer Vierfeldertafel oder in einem zweistufigen Baumdiagramm angeordnet werden. Im statistischen Dorf mit 1.000 Frauen erhalten wir 8 mit Brustkrebs, davon 7 (wir runden die 7,2 ab) mit einem positiven Mammogramm. Ferner ist zu erwarten, dass sich von den 992 Frauen ohne Brustkrebs 70 mit einem falsch-positiven Mammogramm einstellen (7% von 992, gerundet).

Von den 77 positiven Befunden entfallen also 7 auf Brustkrebs. Mit dem Argument zufälliger Auswahl (zunächst aus allen Personen, dann nur mehr aus jenen, welche einen positiven Befund haben) erhält man 7/77, also rund 9% für die gesuchte Wahrscheinlichkeit. Durchschnittlich einer von insgesamt 11 positiven Befunden (im Dorf) entfällt auf Brustkrebs. Die Darstellung mit natürlichen Häufigkeiten überzeugt (fast) alle. Die artifiziellen natürlichen Häufigkeiten stellen aber die Zusammenhänge als Fakten dar: Es gibt (!) 8, es gibt (!) 11 positive Befunde usw. Eine von 11 Frauen mit positivem Befund hat Brustkrebs. Gibt es denn 1.000 so wie ich? Auf wen bezieht sich diese Rechnung?

Hinterfragen der Basisdaten, die ins Modell eingehen

Die Darstellung verwischt, dass die Daten reine Modellgrößen sind; die genauen ganzen Zahlen erzeugen überdies die Illusion von Exaktheit und verschleiern, dass man auf unzuverlässige Schätzungen (oder gar auf mehr oder weniger plausible Szenarien) der relevanten Wahrscheinlichkeiten zugreifen musste. Besonders schwierig wird es, wenn kleine Wahrscheinlichkeiten mit im Spiel sind, wie die folgende Überlegung aus Borovcnik (2014) zeigt.

Die Grundvoraussetzung für die Schätzung von Wahrscheinlichkeiten ist die Annahme einer zufälligen Stichprobe. Die Daten erfüllen solche Forderungen selten, speziell für die Sicherheiten der medizinischen Diagnoseverfahren oder die Prävalenz kann man die Werte eigentlich nur als plausible Szenario-Größen verstehen, welche heuristisch überprüft worden sind. Und hätten wir Zufallsstichproben, ist die Variation sehr groß. Eine Schätzung von Wahrscheinlichkeiten wird besonders fragwürdig, wenn es sich um eine kleine Zahl handelt. Wir illustrieren den Sachverhalt, wie schlecht wir eine Wahrscheinlichkeit von $p = 10^{-4}$ schätzen können, selbst wenn 10.000 Daten aus einer Stichprobe vorliegen: Die 1 als „natürliche Häufigkeit im Dorf“ in Tabelle 1 könnte genauso gut eine 4 oder eben eine 0 sein. Hat man 1, wird diese Variabilität ausgeblendet: 1 von 10.000!

Tabelle 1: Statistisches Dorf für einen Zufallsversuch mit kleiner Wahrscheinlichkeit

Personen mit X im Dorf mit 10.000 Leuten	Wahrscheinlichkeit	Schätzung von p
0	36,8 %	0,0000
1	36,8 %	0,0001
2	18,4 %	0,0002
3	6,1 %	0,0003
4 und mehr	1,9 %	0,0004

3.2 Simulierte Daten

Jede noch so komplizierte Modellierung mit Wahrscheinlichkeiten kann durch Simulation nachgespielt werden. Die gesuchte Wahrscheinlichkeit wird dabei durch Wiederholung des Simulationsszenarios aus relativen Häufigkeiten geschätzt. Weil die materielle Simulation zu aufwändig wäre, bedient man sich mathematischer Algorithmen zur Erzeugung von Zufallszahlen. Simulierte Daten stellen in doppeltem Sinne artifizielle Daten dar. Die Methode der Simulation dient aber nicht nur der approximativen Lösung von stochastischen Problemen, sie kann auch zur Illustration von Wahrscheinlichkeitsaussagen und von Eigenschaften von Methoden der beurteilenden Statistik verwendet werden.

Simulation von Wahrscheinlichkeitsmodellen

Der Widerspruch „Zufall kann prinzipiell nicht vorhergesagt werden“ mit „die Zufallszahlen werden durch einen mathematischen Algorithmus berechnet“ wird pragmatisch gelöst: Die Algorithmen werden ausgetestet und liefern Zahlen, die sich äußerlich wie „echte“ Zufallszahlen verhalten. Damit diese artifiziellen Daten eine Wahrscheinlichkeit gut schätzen lassen, benötigt man viele Simulationsläufe: bei 1.000 Simulationen erhält man Schätzungen auf rund ±3 Prozentpunkte genau (bei einer statistischen „Sicherheit“ von 95%). Meist sind daher umfangreiche Simulationsszenarien erforderlich.

Beispiel (Schlafverlängerung *A* bzw. *B* gegenüber einer Baseline durch *Alphasomnium* und *Betasomnium*; aus Borovcnik, 2014): Beide Medikamente werden an denselben Personen geprüft (verbundene Stichproben). Das Ereignis $B > A$ bedeutet, *Betasomnium* ist besser als das Vergleichspräparat. Die *Hypothese*, wonach *beide gleich wirksam* sind, wird in natürlicher Weise durch $P(B > A) = 0{,}5$ abgebildet. D.h., wir modellieren die Auswirkungen der *Null*hypothese durch ein mathematisches Modell, das sich als „Werfen einer fairen Münze“ beschreiben lässt. Dabei werden alle Daten (Probanden) entfernt, bei denen beide Medikamente zur selben Schlafverlängerung geführt haben.

Die Ablehngrenzen für den statistischen Test werden auf der Basis des Nullmodells (Binomialverteilung mit $p = 0{,}5$) berechnet. Eine Studie habe 39 gültige Daten mit 32, bei denen Medikament *Betasomnium* besser als *Alphasomnium* wirkte, erzeugt. Die Wahrscheinlichkeit – unter der Nullhypothese (kein Unterschied) – für so ein deutliches Ergebnis zugunsten *Betasomnium* beträgt dann 0,00004. Wie viel überzeugender wirkt die Aussage, dass ein so gutes Ergebnis bei 3.000 simulierten Studien (mit artifiziellen Daten) gar nie beobachtet wurde und nur einmal in 30.000 Studien zu erwarten ist.

Rerandomisierung und Resampling

Normalerweise wird ein statistischer Test von Hypothesen (hier der unterschiedlichen Wirkung) dadurch möglich, dass wir für jede Versuchsbedingung ein *statistisches Modell* (eine Verteilung) *unterstellen*. Die realisierten Daten werden in dieses Modell eingepasst und *innerhalb* dieses Modells wird eine Testentscheidung „berechnet“. Rerandomisierung und Resampling gehen wohl von der Annahme einer Stichprobe aus, setzen aber keinerlei spezielle Verteilung voraus. Aus den vorhandenen Daten werden durch Simulation neue, artifizielle Daten höherer Ordnung erzeugt. Mit diesen „Daten“ vergleicht man die Ausgangsdaten und stellt damit fest, ob die Annahme gleicher Wirkung abgelehnt werden kann oder nicht.

Wir unterstellen dieselbe Situation wie vorhin mit dem Vergleich von Schlafmitteln mit dem Unterschied, dass nun die beiden Medikamente an verschiedenen Personen ausprobiert werden: *Betasomnium* wird in der Versuchsgruppe, *Alphasomnium* in der Kontrollgruppe verabreicht. Die Probanden werden zuvor den Gruppen durch *Randomisierung* zugeordnet. Folgende Daten haben sich für die Schlafverlängerung (in Stunden) im Vergleich zur üblichen Schlafdauer ergeben: *A*: 0,7; –1,6; –0,2; –1; 0,1; 3,4; 3,7; 0,8; 0,0; 2,0 bzw. *B*: 2,2; 0,8; 1,1; 0,1; –0,4; 4,4; 5,5; 1,6; 4,6; 3,4.

Die Nullhypothese, dass kein Effekt der Behandlung existiert, kann nun probabilistisch so umgesetzt werden: Wenn es keinerlei Zusammenhang der Schlafverlängerung mit dem verabreichten Medika-

ment gibt, so ist *jede* Aufteilung der insgesamt 20 Daten auf zwei Gruppen (mit hier je 10 Werten) *gleichberechtigt* und erhält damit *dieselbe* Wahrscheinlichkeit; diese Modellierung nennt man Rerandomisierung: Man bildet alle möglichen Aufteilungen (es sind fast 200.000 Kombinationen) und bestimmt jeweils die Mittelwerte jeder Gruppe und deren Differenz $\bar{b} - \bar{a}$, so erhält man Daten, welche (mit je gleicher Wahrscheinlichkeit) die Nullhypothese repräsentieren. Das Ergebnis der bestehenden Aufteilung wird dann in diese artifizielle Verteilung eingeordnet und bewertet.

Diese artifizielle Verteilung unter der Nullhypothese wird am besten durch Simulation der Zuordnung erschlossen. Ein Szenario – wieder mit 3.000 Wiederholungen und simulierten Daten – gibt dann rasch Aufschluss darüber, wo die Grenzen der Ablehnung der Nullhypothese festzulegen sind und ob die ausgangs beobachtete Stichprobe darüber liegt. Bei der fiktiven Neuzuordnung zu Versuchs- und Kontrollgruppe kann man die Originaldaten durch ihre *Ränge* ersetzen. Man bringt dazu alle (20) Daten in eine Rangfolge und bestimmt die Rangsumme für *B* (und damit für *A*). Liegt die Rangsumme der realen Ausgangsdaten unter den 5% extremsten, so führt ein Signifikanztest (zum Niveau 5%) zur Ablehnung der Null-(Effekt-)Hypothese. Das ist der Wilcoxon-Test (siehe Fahrmeir et al., 2004).

Ein neuerer Ansatz, artifizielle Daten zu erzeugen, um der Frage nachzugehen, ob zwischen den Behandlungsmöglichkeiten Unterschiede bestehen oder nicht, ist Resampling oder Bootstrap (didaktische Erörterungen zu diesem Ansatz sind zu finden in Borovcnik, 2006, oder in Engel und Grübel, 2008). Hier werden die beiden Datensätze der Versuchs- und der Kontrollgruppe jeweils für sich als Schätzung der entsprechenden Modellverteilung herangezogen. Dann werden aus diesen beiden Verteilungen jeweils (mit Zurücklegen) neue Datensätze wiederholt simuliert. Mit Hilfe dieser artifiziellen Daten wird das Ausgangsproblem beurteilt.

4. Daten und Fakten

4.1 Wirksamkeit von HPV-Impfungen

In den letzten zehn Jahren wurde intensiv an der Wirksamkeit von HPV-Impfungen gegen Cervixkarzinom geforscht (siehe Sprenger, 2013). Dahinter stehen reale Daten aus retrospektiven Untersuchungen, wonach bei fast allen Frauen mit einem Karzinom am Muttermund auch ein Befall mit (bestimmten Typen von) HPV-Viren (humane Papilloma-Viren) festgestellt wurde. Diese Viren sind weit verbreitet und fast drei Viertel der Frauen sind – im Laufe ihres Lebens – davon betroffen. Allerdings kann ein gesundes Immunsystem damit gut fertig werden und der Befall geht entsprechend auch wieder zurück. Die Datenlage, dass bei Cervixkarzinom fast durchwegs auch HPV vorliegt, ist unumstritten. Aber die Relationen werden umgedreht und so gedeutet, als ob diese Viren das Karzinom *verursachen*. Und das wird als Faktum ausgewiesen. Der Erfolg der Impfung auf die Surrogatvariable „HPV-Befall“ wird daher als Erfolg in der Krebsvorsorge interpretiert. Die Schwierigkeiten, die sich aus der Analyse von Surrogat- anstelle der eigentlichen Zielvariablen ergeben, sind etwa in Dubben und Beck-Bornholdt (2006) oder in Goldacre (2008) dargestellt.

4.2 Innere und äußere Zusammenhänge – Drittvariable

Zusammenhänge zwischen Merkmalen werden je nach Mess-Skala durch Korrelation und Assoziation erfasst. Diese Konstrukte untersuchen Zusammenhänge nur vom äußeren Phänomen her; dagegen sagen sie über innere (etwa kausale) Zusammenhänge nichts aus. Daten werden aber häufig wie in der Diskussion von HPV interpretiert. Es ist wenig erforscht, wie diese Viren das Wachstum von Tumorzellen verursachen; man kann aber einen Dritt-Faktor schlecht wegdiskutieren: Ein gesundes Immunsystem wird mit Infektionen durch HPV-Viren fertig *und* es wird auch Zellschädigungen im Vorstadium von Krebs (so genannte Zelldysplasien) entsprechend ausscheiden. Zellveränderungen sind reversibel; es ist zellphysiologisch ungeklärt, warum in dem einen oder anderen Fall das Tumorwachstum

invasiv wird und ein Cervixkarzinom entsteht (Sprenger, 2013). Es könnte daher viel sinnvoller sein, das Immunsystem durch entsprechende Maßnahmen zu stärken, anstatt sich auf eine Impfung zu verlassen, ohne den Lebensstil anzupassen. Medizinische Entscheidungen sind immer schwierig, weil die unterschiedlichen „Stakeholder" verschiedene Interessen haben. Man kann von den anderen Beteiligten allein aufgrund ihrer Rolle kaum einen objektiven Rat erfragen; dazu sind noch die Beteiligten von den Folgen der Entscheidungen ganz anders betroffen (der Patient hat die Folgen am eigenen Leib zu tragen; der Arzt ist u.a. juristisch haftbar); ferner stellen sich die Folgen oft erst langfristig ein oder es ist schwierig, sie – kausal – zuzuordnen (siehe Borovcnik und Kapadia, 2011a; b).

4.3 Genetisches Material zur Beweisführung vor Gericht

Genetische Befunde sind sehr zuverlässig, denn sie identifizieren eine Person mit extrem geringen Fehlerwahrscheinlichkeiten. Wenn DNA einer Person an einer „blutigen" Tatwaffe gefunden wurde, wird dies als Faktum so interpretiert, dass diese Person mit an Sicherheit grenzender Wahrscheinlichkeit die Tat begangen hat. Aber auch kleine Fehlerquoten haben gelegentlich große Auswirkungen. Abhängig vom Zustand der Probe gibt es eine Fehlerquote zwischen 10^{-6} und 10^{-8}. Im schlechtesten Fall bei 10^{-6} hätte man dann für Deutschland 80 Personen zu erwarten, die mit der gefundenen DNA übereinstimmen (siehe Dubben und Beck-Bornholdt, 2010, 50-53). Der genetische Befund selbst lässt die Frage, *wie* die DNA an die Tatwaffe gekommen ist, völlig offen. Das wird jedoch gerne übersehen. Es ist ein leichtes, sich die DNA einer Zielperson zu beschaffen und sie irgendwo aufzubringen.

Weitere Beispiele gibt es zuhauf. Man denke etwa an den höchst artifiziellen Charakter der PISA-Daten, welche als reale Daten der Leistung von Ländern missinterpretiert werden und als Motor ständiger Reformen dienen. Wie kompliziert die Verfahren sind, zeigt sich daran, dass die offiziellen Ergebnisse für Österreich 2000 aufgrund von Kritik von außerhalb (Neuwirth, Ponocny und Grossmann, 2006) stark abgeändert werden mussten. Für eine kritische Hinterfragung des Ansatzes hinter PISA sei auf Bender (2005) verwiesen.

5. Fazit

Reale Daten haben einen hohen Charakter an Glaubwürdigkeit. Ziel einer aufgeklärten statistischen Unterweisung ist es, einen sensiblen Umgang damit anzustreben. Auf dem Weg von den Statistikern zum Zielpublikum werden die Annahmen immer weniger verstanden und auch weggelassen. Die Annahmen werden nicht übersehen sondern bewusst weggelassen. Das kann unterschiedliche Gründe haben. Zum einen will man die Adressaten nicht verschrecken, weil sie die Annahmen kaum verstehen und daher vielleicht überhaupt „weghören". Zum anderen kann es auch Manipulation sein, die verschleiern soll, wie die Daten (Fakten?) eigentlich – nämlich nur unter sehr starken Einschränkungen – zu interpretieren sind. Dann würden die Ergebnisse vielleicht doch viel weniger relevant erscheinen.

Weil die Annahmen weggelassen oder verschwiegen werden, steigt der scheinbar faktische Gehalt der Daten. Dies wird insbesondere durch – populäre – Visualisierungen und eine allgemeine Erwartung von Präzision mitbedingt. Wir sollten Ergebnissen von statistischen Studien eine Art Beipackzettel beilegen, wie bei Medikamenten auch, damit sich wenigstens die interessierten Rezipienten über die Relevanz der Schlussfolgerungen orientieren können. Einige didaktische Kniffe verwischen die Grenze zwischen artifiziellen Daten und Fakten. Ergebnisse und eingesetzte Modelle gewinnen durch die *artifiziellen, aber konkreten* (!) Daten – zu Unrecht – an Glaubwürdigkeit. Statistik als Handwerk in einem Erkenntnisprozess zu lernen, mag helfen, den Grenzgang zwischen realen und artifiziellen Daten zu schaffen (Borovcnik, 2013). Die Dualität von Daten und probabilistischen Modellen zeigt, dass ein realen Daten innewohnendes Signal erst auf artifiziellem Weg zu erkennen ist. „*Traue keiner Statistik, die du nicht selbst gefälscht hast!*"

Der Autor dankt Herrn Reinhard Winkler für die kritische Durchsicht des Beitrags, welche die Lesbarkeit der Ideen erheblich verbessert hat.

Literatur

Axmann, M. (2014): Parlamentswahlen in Ungarn. *Politischer Hintergrundbericht*. München: Hanns-Seidel-Stiftung. Online: www.hss.de/fileadmin/media/downloads/Berichte/140416_Ungarn_PHB.pdf.

Barke, W. (2004): „Ich glaube nur der Statistik, die ich selbst gefälscht habe ...". *Statistisches Monatsheft Baden-Württemberg* 11/2004, 50-53.

Bender, P. (2005): Die etwas andere Sicht auf PISA, TIMSS und IGLU. *Der Mathematikunterricht* 51 (2/3), 36-57.

Borovcnik, M. (1992): *Stochastik im Wechselspiel von Intuitionen und Mathematik*. Mannheim: Bibliographisches Institut.

Borovcnik, M. (2006): Daten – Zufall – Resampling. In: J. Meyer (Hrsg.): *Anregungen zum Stochastikunterricht* Bd. 3. Hildesheim, Berlin: Franzbecker, 143-158.

Borovcnik, M. (2007a): Goldstandard: Das randomisierte, doppelblinde, Placebo-kontrollierte Experiment. *Stochastik in der Schule*, 27(2), 26.

Borovcnik, M. (2007b): Kann man ethisch vertreten, dass eine Gruppe ein Placebo statt einer zielgerichteten Behandlung erhält? *Stochastik in der Schule*, 27(2), 24.

Borovcnik, M. (2012): Interaktive Statistik. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft*, 45, 1-18.

Borovcnik, M. (2013): Forschungsprozess und probabilistische Modellbildung – Stochastische Denkweisen. In: J. Maaß und S. Siller (Hrsg.): *Materialien für einen realitätsbezogenen Mathematikunterricht*. Hildesheim: Franzbecker.

Borovcnik, M. (2014): Vom Nutzen artifizieller Daten. In U. Sproesser, S. Wessolowski, C. Wörn (Hrsg.): *Daten, Zufall und der Rest der Welt*. Wiesbaden: Springer Fachmedien, 27-43.

Borovcnik, M.; Kapadia, R. (2011a): Risk in health: more information and more uncertainty. *IASE Satellite Conference on „Statistics Education and Outreach"*. Voorburg: ISI (6 S.).

Borovcnik, M.; Kapadia, R. (2011b): Determinants of decision-making in risky situations. *58th World Statistics Congress*. Voorburg: ISI (6 S.).

Borovcnik, M.; Schenk, M. (2011). Simulationen im Stochastik-Unterricht. *Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft*, 44, 1-16.

Dubben, H.-H.; H.-P. Beck-Bornholdt (2010): *Mit an Sicherheit grenzender Wahrscheinlichkeit. Logisches Denken und Zufall*. Reinbek: Rowohlt.

Dubben, H.-H.; H.-P. Beck-Bornholdt (2006): *Der Hund, der Eier legt. Erkennen von Fehlinformation durch Querdenken*. Reinbek: Rowohlt.

Engel, J.; Grübel, R. (2008): Bootstrap – oder die Kunst, sich selbst aus dem Sumpf zu ziehen. *Mathematische Semesterberichte*, 55, 113-130.

Fahrmeir, L.; Künstler, R.; Pigeot, I.; Tutz, G. (2004): Statistik. 5. Auflage. Berlin: Springer.

Freedman, D.; Pisani, R.; Purves, R. (2007): *Statistics*. New York: W.W. Norton & Company.

Gigerenzer, G. (2002): *Das Einmaleins der Skepsis. Über den Umgang mit Zahlen und Risiken*. Berlin: Berlin Verlag.

Goldacre, B. (2008): *Badscience*. London: Fourth Estate.

Kiaer, A.N. (1899): Die repräsentative Untersuchungsmethode. *Allgemeines Statistisches Archiv*, 5, 1-22.

Neuwirth, E.; Ponocny, I.; Grossmann, W. (2006): *PISA 2000 und PISA 2003. Vertiefende Analysen und Beiträge zur Methodik*. Graz: Leykam.

Sprenger, M. (2013): Update HPV-Impfung. *Österreichische Krankenhauszeitung*, 54(1-2), 10-13.

Verfasser

Manfred Borovcnik
Alpen-Adria-Universität Klagenfurt, Institut für Statistik
Universitätsstraße 65, 9020 Klagenfurt
manfred.borovcnik@uni-klu.ac.at