2

Was kann bei wissenschaftlichen Studien so alles passieren?

2.1 Korrelation ist nicht Kausalität!

Das ist das Grundproblem der Statistik: Sie liefert nur *Korrelationen*, eigentlich sind wir aber eher an *kausalen* Zusammenhängen interessiert. Leider scheint es zumindest in den Publikumsmedien eher die Regel als die Ausnahme zu sein – und ist auch im wissenschaftlichen Bereich nicht unbekannt –, dass zwischen Korrelation und Kausalität nicht sorgfältig unterschieden wird. Das merkt meist niemand, abgesehen von ein paar überkritischen Spielverderbern, und schnell ist wieder einmal eine weithin geglaubte Behauptung in der Welt, die mit der Realität nicht unbedingt viel zu tun hat.

In der Literatur finden Sie umfangreiche Besprechungen der beiden Begriffe und ihres Zusammenhangs.[1] Für unsere Zwecke genügt zu verstehen:

K. Weihe, *Fundiert entscheiden*,
https://doi.org/10.1007/978-3-662-54704-5_2

- Eine *Korrelation* zwischen zwei Indikatoren sagt etwas darüber aus, bis zu welchem Grad die beiden Indikatoren irgendwie parallel zueinander verlaufen.
- Eine *Kausalität* zwischen zwei Indikatoren ist eine gerichtete Wirkbeziehung: Der erste Indikator ist eine (Mit)ursache für den zweiten Indikator.

Für *Indikator* werden auch diverse Synonyme wie *Variable*, *Kenngröße* und *Meßgröße* verwendet. Ich verwende konsequent den Begriff *Indikator*, was ja soviel wie „Anzeiger" bedeutet, damit Sie von vornherein im Auge behalten, dass Kenngrößen oft nur *anzeigen*, was man eigentlich wissen möchte.

Sie denken, der Unterschied zwischen Korrelation und Kausalität betrifft Sie nicht persönlich? Dann lesen Sie das folgende Beispiel.

Fallbeispiel 22: Ihre persönliche Bonität

Manche Leute haben einen schlechten Schufa-Eintrag oder bekommen nur ungünstige Kreditkonditionen angeboten oder müssen hohe Versicherungsprämien zahlen oder bekommen keinen Handyvertrag, einfach weil irgendwelche Indikatoren bei ihnen hoch sind, die auch bei vielen Leuten mit schlechten Risiken hoch sind, zum Beispiel die Anzahl der Wohnungswechsel oder der Bankkonten.[2]

Tja, Pech gehabt, wenn man selbst eigentlich ein gutes Risiko ist. Die Agenturen richten sich nun einmal nach Korrelationen, nicht nach Kausalitäten.

Man kann es gut oder schlecht finden: Zumindest Kfz-Versicherungen machen inzwischen einen Schritt weg von der Korrelation hin zur Kausalität durch Überwachung des

Fahrstils.[3] An einem riskanten Fahrstil sind natürlich ganz allein Sie selbst schuld. □

Es gibt den Spruch: „Wer eine Korrelation findet, darf sich eine Kausalität dazu ausdenken." Wenn eine Korrelation zwischen zwei Indikatoren besteht, scheint häufig ein bestimmter kausaler Zusammenhang besonders plausibel zu sein. In vielen, vielen Artikeln in den Publikumsmedien und auch anderswo hält man sich dann auch gar nicht lang mit der Frage auf, ob diese doch so plausible Kausalität wirklich stimmt, sondern präsentiert sie als „bewiesene" Wahrheit. Aber der wahre kausale Zusammenhang kann auch ganz anders aussehen. In Abschn. 2.2 werden wir sehen, dass ein kausaler Zusammenhang sogar nicht einmal existieren muss.

Im Einzelfall kann es natürlich durchaus sein, dass ein kausaler Zusammenhang zwischen zwei Indikatoren A und B genau so besteht, wie er plausibel erscheint: A verursacht B. Es kann aber auch genau umgekehrt sein. Wann immer Sie von einem kausalen Zusammenhang lesen, drehen Sie die Kausalität einfach einmal um. Wenn der umgedrehte Zusammenhang, also B verursacht A, plausibler und weniger spektakulär ist, dann ist der vermutete Zusammenhang, A verursacht B, wahrscheinlich falsch (es sei denn, weitere Argumente stützen den vermuteten Zusammenhang).

Fallbeispiel 23: Optimismus und Gesundheit
Nach diversen Ratgeberbüchern sind gutgelaunte Optimisten gesünder als depressive und pessimistische Menschen. Die Anleitungen in diesen Büchern, wie man optimistischer und dadurch gesünder wird, scheinen viele Leser unter Druck

zu setzen, vor allem, wenn ihre Bemühungen um mehr Optimismus nicht so recht fruchten wollen.

Aber selbst wenn die Korrelation stimmen würde – woran es erhebliche Zweifel gibt[4] –, wäre die Schlussfolgerung, dass man eben optimistischer und glücklicher werden soll, um gesund zu bleiben, noch lange nicht gerechtfertigt. Die umgedrehte Kausalität – dass gesunde Menschen optimistischer sind, *weil* sie gesund sind und sich entsprechend gut fühlen – scheint deutlich plausibler zu sein und könnte die Korrelation zwischen Optimismus und Gesundheit weitgehend erklären. □

Manchmal ist es nur eine unklare Formulierungsweise, die aus einer Korrelation eine Kausalität macht. Sie lesen etwa Formulierungen der Art: „ein wichtiger Faktor ist". Das Wort „Faktor" suggeriert einen kausalen Zusammenhang, der nicht unbedingt gegeben sein muss. Noch subtiler ist das folgende Beispiel.

Fallbeispiel 24: Beeinflusst Lektüre die Persönlichkeit?
Ein Beispiel von Christensen und Christensen:[5] Frauen, die „Fifty Shades of Grey" gelesen haben, zeigen statistisch auffällige Persönlichkeitsmerkmale, die darauf hindeuten, dass Frauen ihre Persönlichkeit durch Lektüre dieses Buches verändern. In Wirklichkeit dürfte es wohl eher andersherum sein: Ein bestimmter Typ Frauen findet das Buch im Durchschnitt besonders interessant.

Die Umfrage bringt den Faktor Zeit in einer Form hinein, die eine bestimmte Kausalität nahelegt, denn eine Wirkung tritt ja immer zeitlich *nach* ihrer Ursache auf. Wenn die

Frauen nun gefragt worden sind, ob sie das Buch gelesen haben, dann sagen sie nur „Ja", *nachdem* sie das Buch gelesen haben. Sie haben also *zuerst* das Buch gelesen, und *danach* wird ihre Persönlichkeit eingeschätzt; über das Buchlesen wird in der Vergangenheitsform berichtet, über die Persönlichkeit in der Gegenwartsform. □

Aus Kostengründen sind wohl die meisten Studien als reine *Querschnittsstudien* angelegt, das heißt, man macht genau einmal eine Erhebung. Mit Querschnittsstudien kann man Korrelationen zwischen A und B finden, aber ohne weitere Informationen von anderswoher kann man grundsätzlich nicht klären, ob A von B oder B von A verursacht wird. *Längsschnittstudien* führen zu mehreren Zeitpunkten jeweils eine Erhebung mit möglichst identischer Stichprobe durch. Wenn nun eine Längsschnittstudie ergibt, dass die Entwicklung von A der Entwicklung von B eindeutig zeitlich hinterherhinkt, dann lässt sich immerhin mit einiger Sicherheit ausschließen, dass B durch A verursacht wird. Aber Längsschnittstudien sind teuer, und man braucht viel Geduld – nicht selten über Jahrzehnte hinweg –, bis belastbare Ergebnisse vorliegen.

Ein letzter Fall in diesem Abschnitt: Wenn die Indikatoren A und B miteinander korrelieren, kann es durchaus sein, dass sowohl A von B als auch B von A verursacht wird in Form eines Regelkreises. Sie kennen den Begriff „Teufelskreis" für den Fall, dass zwei Indikatoren sich gegenseitig zum Schlechteren treiben. Das folgende Fallbeispiel eines Regelkreises werden Sie als demokratisch gesinnter Staatsbürger sicherlich auch dann als einen Teufelskreis, also negativ bewerten, wenn nicht Sie selbst und Ihre politischen Freunde, sondern Ihre politischen *Gegner* davon betroffen sind.

Fallbeispiel 25: Die Schweigespirale[6]
Es ist nach momentanem Wissensstand nicht auszuschließen, dass Menschen sich in ihren Meinungsäußerungen – auch bei Umfragen – am allgemeinen Meinungsklima orientieren: Wer seine eigene Meinung als gesellschaftlich erwünscht einstuft, wird sie nach dieser Theorie tendenziell eher und öfter und auch dezidierter äußern als jemand, der seine eigene Meinung als gesellschaftlich unerwünscht empfindet. Dies würde dann bewirken, dass die vorherrschende Meinung sich weiter verfestigt. Ein weiter verfestigtes Meinungsklima wirkt dann noch stärker entmutigend auf die Anhänger der unerwünschten Meinung und macht die Verfechter der vorherrschenden Meinung noch selbstbewusster, und so geht es immer weiter. Die vorherrschende Meinung muss gar nicht die Mehrheitsmeinung sein, sondern kann durchaus von einer lautstarken Minderheit, die in Politik und Medien an den richtigen Positionen sitzt, entsprechend „gepusht" worden sein. □

Wir haben in diesem Abschnitt gesehen, dass eine Korrelation zwischen A und B bedeuten kann, dass A von B oder auch B von A verursacht ist, oder dass beide sich in einem Regelkreis gegenseitig bedingen. Der häufigste Fall von Fehlinterpretation von Korrelationen dürfte aber ein anderer sein, nämlich der Fall, dass A und B in einer allenfalls indirekten oder gar keinen kausalen Beziehung zueinander stehen. Man spricht auch von einer *Scheinkorrelation*, die Korrelation erweckt also nur den Anschein einer direkten kausalen Beziehung. Dieser Fall ist so wichtig und komplex, dass wir ihm einen eigenen Abschnitt widmen.

2.2 Scheinkorrelationen

Der Fall, dass zwei korrelierende Indikatoren nicht in einer direkten kausalen Beziehung zueinander stehen, wird leider häufig mit Beispielen eingeführt, die das Problem ins Lächerliche ziehen: zum Beispiel die durchaus starke Korrelation zwischen der Größe der Storchenpopulation und der Geburtenrate. Der Kausalschluss, dass der Storch die Kinder bringt, ist so offensichtlich falsch, dass man sich leicht über dieses Problem erhaben fühlen kann nach dem Motto: *Mir* kann ein solcher Fehlschluss *nicht* passieren.

Kann er doch. Machen Sie sich klar: Wenn zwei Indikatoren über denselben Zeitraum gemessen werden und jeder der beiden tendenziell steigend oder fallend ist, dann sind die beiden Indikatoren zwangsläufig signifikant stark miteinander korreliert. Problematisch wird es, wenn ein kausaler Zusammenhang zwischen zwei stark korrelierenden Indikatoren A und B durchaus sachlich plausibel erscheint, das ist eben beim Storchenbeispiel nicht der Fall. Auch bei plausibel erscheinenden Kausalitäten kann es immer noch sein, dass die Korrelation allein darauf beruht, dass A und B beide tendenziell eher wachsen oder sich sonstwie tendenziell ganz grob parallel zueinander entwickeln.

Fallbeispiel 26: Kühlschränke in Zeiten des Klimawandels

Mutmaßlich steigen die Temperaturen tendenziell seit vielen Jahren weltweit, andererseits nimmt die Anzahl von Kühlschränken weltweit zu. Beide Indikatoren sind daher positiv korreliert. Die Hypothese, dass der Klimawandel dazu führt,

dass mehr Kühlschränke verkauft werden, ist nicht von vornherein unplausibel. Aber in Wirklichkeit dürften beide Indikatoren wohl in keiner nennenswerten kausalen Beziehung zueinander stehen; dass sie beide mit der Zeit steigen, reicht für eine Korrelation. □

Das bisher in diesem Abschnitt Gesagte mag genügen für den Fall, dass es überhaupt keine Kausalität hinter einer Korrelation zwischen zwei Indikatoren A und B gibt. Aber selbst wenn es eine Kausalität dahinter gibt, muss weder A von B noch B von A verursacht sein, sondern es kann auch ein C dahinter stecken.

Fallbeispiel 27: Die Vorlesung bestimmt den Studienerfolg?

Studierende, die regelmäßig zur Vorlesung kommen, sind besser.[7] Das muss aber gar nicht an der Qualität der Vorlesung liegen: Zur Vorlesung gehen eher diejenigen Studierenden, die vor Ort leben. Zum Beispiel hat die TU Darmstadt, an der ich arbeite, einen großen Einzugsbereich, und viele Studierende von weiter her nehmen möglichst selten die Anreise auf sich. Diese Studierenden verbringen insgesamt weniger Zeit an der Uni, zum Beispiel mit Austausch unter Studierenden, und haben vielleicht teilweise auch den Kopf nicht so frei für das Studium wie andere. Das wären durchaus alternative Erklärungsmöglichkeiten für deren statistisch schlechteres Abschneiden.

Mit A, B und C gesprochen: Wenn A der Indikator ist, wie häufig ein Studierender die Vorlesung besucht, und B, wie gut er in der Prüfung abschneidet, dann kann die Ursache für die Korrelation zwischen A und B auch in dem Indikator

C liegen: wieviel Zeit der Studierende überhaupt seinem Studium widmet. □

Die Korrelation zwischen *A* und *B* kann ganz durch *C* entstanden sein oder auch nur teilweise.[8] Im letzten Beispiel ist ohne weitere Zusatzinformation nicht klar, ob *C* teilweise oder vollständig verantwortlich ist. Im nächsten Beispiel behauptet die zitierte Studie, dass der Indikator *C* (Bildung) den fraglichen Zusammenhang vollständig erklärt, im übernächsten Beispiel ist der Indikator *C* (Lebensalter) mit Sicherheit nur teilweise verantwortlich.

Fallbeispiel 28: Arbeitsethos nach Max Weber
Dieses Beispiel erhält seine Würze durch die jahrzehntelange starke Wirkung auf Wissenschaft und öffentliche Meinung: Zumindest in früheren Zeiten und zumindest in Deutschland waren Protestanten im Schnitt wirtschaftlich erfolgreicher als Katholiken. Der berühmte Soziologe Max Weber behauptete, dass das protestantische Arbeitsethos eine wesentliche Rolle spiele.

Eine Studie des ifo Institus aus dem Jahr 2008[9] legt hingegen nahe, dass der höhere durchschnittliche Bildungsgrad der Protestanten den Effekt schon völlig erklärt. Das heißt: Rechnet man den Indikator *C* (Bildung) mit den dafür üblichen statistischen Methoden heraus, dann bleibt zumindest in dieser Studie kein nennenswerter Zusammenhang zwischen Religion und wirtschaftlichem Erfolg mehr übrig. □

Wie oben angekündigt nun ein Fallbeispiel, in dem der Indikator *C* (Lebensalter) sicherlich nur teilweise das beobachtbare Phänomen erklärt.

Fallbeispiel 29: Wie ungerecht ist das Vermögen verteilt? Wenn, wie so häufig, einfach nur erhoben wird, wie viele Menschen in Deutschland nun wie viel Vermögen haben, dann kommt man schon auf eine recht starke Spreizung. Allerdings sind die Unterschiede zumindest zum Teil auf das Alter zurückzuführen: Menschen „in den besten Jahren" haben mehr Vermögen als ganz junge Menschen, zum Beispiel Studierende und Auszubildende. Das wird man wohl nicht unbedingt als ungerecht bewerten. Um zu ermessen, wie gerecht oder ungerecht es in Deutschland wirklich zugeht, muss man also das Alter und mutmaßlich noch weitere Faktoren herausrechnen. □

> Vergessen Sie jede Studie über Menschen und Kollektive, bei denen nicht die demographischen Faktoren herausgerechnet worden sind, also Alter, Geschlecht, Einkommen, Bildungsgrad, ethnische und religiöse Zugehörigkeit und so weiter.

Denn die Erfahrung lehrt, dass – wie in den letzten beiden Fallbeispielen – statistische Auffälligkeiten häufig stark mit einzelnen demographischen Faktoren korrelieren, was vermuten lässt, dass die wahre Ursache zumindest teilweise woanders liegt. Dazu könnten unzählige weitere Beispiele zitiert werden.[10]

Sehr häufig beruht eine Korrelation nicht auf einer verborgenen weiteren Wirkursache, sondern viel profaner auf einer vorsortierten Stichprobe, also einer unbeabsichtigten – manchmal vielleicht auch beabsichtigten – Vorauswahl.[11]

Fallbeispiel 30: Lehrevaluation

Irgendwann in der zweiten Hälfte des Semesters[12] werden die Studierenden in der Vorlesung gebeten, einen Fragebogen auszufüllen, auf dem sie verschiedene Aspekte der Vorlesung selbst und des begleitenden Übungsbetriebs bewerten können. Zumindest in Deutschland und einigen anderen Ländern herrscht in den meisten Vorlesungen keine Anwesenheitspflicht. Zumindest in Fächern wie meinen – Mathematik und Informatik –, in denen eher das geschriebene als das gesprochene Wort wichtig ist, bröckelt die Anwesenheit in den ersten Wochen des Semesters entsprechend, vor allem wenn das online zur Verfügung gestellte Lehrmaterial so umfassend ist, dass man die Prüfung auch ohne Vorlesungsbesuch bestehen kann. Nur ein harter Kern von Hörern bleibt übrig, die mehr oder weniger jede Woche kommen. Die anderen belegen den Übungsbetrieb weiter und legen am Ende auch die Prüfung ab, gehen aber eben nicht regelmäßig vorher in die Vorlesung. Sie kommen auch nicht unbedingt zum Termin der Lehrevaluation, selbst wenn dieser angekündigt wird.

Der Punkt ist: Es gibt überhaupt keinen Grund zur Annahme, dass dieser harte Kern, der dann an der Lehrevaluation teilnimmt, repräsentativ für die Gesamtheit antwortet. Wenn etwa ein Großteil der Hörer wegen mangelnder Qualität der Vorlesung fernbleibt, geht dieser Kritikpunkt nicht so stark in die Bewertung ein, wie er eigentlich sollte.

Leicht überspitzt gesagt: Will ein Dozent möglichst gute Evaluationsergebnisse bekommen, dann sollte er seine Vorlesung passgenau auf 10 % der Hörer ausrichten – diese 10 % sind hochzufrieden, die anderen 90 % hingegen bleiben bald fern und beeinflussen daher nicht das Evaluationsergebnis. □

2.3 Signifikanzniveau und statistische Signifikanz

Für jede statistische Studie setzen die Designer der Studie ein *Signifikanzniveau* fest, 5 % ist der gängige Wert. Wenn die Ergebnisse der Studie das dafür festgelegte Signifikanzniveau erreichen, dann bezeichnet man das Ergebnis als *statistisch signifikant*. Was heißt das?

Allgemein bedeutet Signifikanzniveau *P* %: Sollte die *Nullhypothese* stimmen – sollte es also nichts Auffälliges zu berichten geben –, dann tritt mit höchstens *P*-prozentiger Wahrscheinlichkeit durch zufällige Fluktuationen eine so hohe Abweichung von der Nullhypothese auf, dass das Studienergebnis fälschlich als positiv akzeptiert, also fälschlich als statistisch signifikant angesehen wird. Mit dem Wert 5 % können wir es auch so formulieren: Von einhundert Studien mit Signifikanzniveau *P* %, die jeweils eine falsche Hypothese testen, ergeben im Schnitt höchstens fünf ein Ergebnis, das fälschlich als positiv akzeptiert wird. Wichtig ist aber, dass die Umkehrung nicht gilt: Signifikanzniveau 5 % bedeutet *nicht*, dass mit 95-prozentiger Wahrscheinlichkeit ein positives Ergebnis richtig ist.

Diese von mir gewählte Formulierung ist an die Fachsprache angelehnt, lässt sich aber gut anhand von Beispielen illustrieren. Einhundert Hypothesen könnten beispielsweise einhundert neu entwickelte Medikamente sein, die jeweils durch eine Studie auf Wirksamkeit getestet werden. Die erste Hypothese ist, dass das erste Medikament wirksam ist, die zweite Hypothese, dass das zweite Medikament wirksam ist, und so weiter. Ein falsch positives Ergebnis heißt, dass ein unwirksames Medikament durch statistische

Fluktuationen fälschlich als wirksam eingestuft wird. Einhundert Hypothesen könnten auch einhundert Patienten sein, die auf eine Krankheit getestet werden, etwa AIDS. Die erste Hypothese ist dann, dass der Patient Nr. 1 AIDS hat, und ein falsch positives Ergebnis wäre, dass der Patient fälschlich HIV-positiv getestet wird.

Um die praktischen Konsequenzen daraus zu beleuchten, bräuchten wir eigentlich noch einen Indikator, der oft gar nicht bekannt ist, nämlich wie viele der getesteten Hypothesen *tatsächlich* falsch sind, egal ob sie in der Studie positiv oder negativ getestet worden sind. Also beim Beispiel Medikamente von eben, wie viele getestete Medikamente tatsächlich unwirksam sind. Oder beim Beispiel HIV-Test, wie viele der getesteten Patienten tatsächlich *kein* AIDS haben. Nach mathematischem Brauch bezeichnen wir diese unbekannte Prozentzahl im Folgenden mit X.

Es geht uns hier nur um das Prinzip. Daher ist es ok, wenn wir die Realität ein bisschen vereinfachen, um die Beispielrechnung einfach zu halten:

Erste Annahme: Wie gesagt, Signifikanzniveau 5 % bedeutet, dass im Schnitt *höchstens* 5 % aller falschen Hypothesen fälschlich positiv getestet werden. Der Einfachheit halber gehen wir von dem Fall aus, dass es *genau* 5 % sind.

Zweite Annahme: Natürlich können umgekehrt auch zutreffende Hypothesen fälschlich negativ getestet werden. Diesen Fall schließen wir zur rechnerischen Vereinfachung aus.

Dritte Annahme: Schließlich gehen wir noch davon aus, dass die Fehler in den einzelnen Tests wirklich nur rein zufällige Fluktuationen sind, keine systematischen Verzerrungen.

Die erste Annahme besagt, dass $0,05 \cdot X$ von hundert Hypothesen falsch sind, aber zu Unrecht positiv getestet werden. Falls beispielsweise nur jede hundertste Hypothese zutrifft, also $X = 99$ ist, sind das $0,05 \cdot 99 = 4,95 \approx 5$ aller einhundert Hypothesen. Die zweite Annahme besagt nun, dass $100 - X$ Hypothesen wahr und positiv getestet sind, also eine der einhundert Hypothesen in diesem Fall. Gemäß der dritten Annahme beeinflussen sich diese beiden Fakten nicht gegenseitig. Man kann sie also einfach nebeneinander betrachten, so dass im Rechenbeispiel auf jedes *korrekt* positive Ergebnis fünf *falsch* positive Ergebnisse kommen.

Der Anteil der fälschlich positiv getesteten Hypothesen an allen positiv getesteten Hypothesen hängt von der Unbekannten X ab; je höher X, umso höher der Anteil falsch positiver Befunde; ist X nahe bei 100, dann sind *nahezu alle* positiven Befunde *falsch*!

Diese allgemeine Erkenntnis hat fundamentale Konsequenzen, die sich wieder an medizinischen Beispielen besonders gut verdeutlichen lassen.

Fallbeispiel 31: Sind Sie HIV-positiv, wenn der Test das sagt?[13]

Sie lassen sich auf eine Krankheit testen, die sehr selten vorkommt, zum Beispiel AIDS. Es gibt keinen konkreten Verdacht, dass Sie AIDS haben, und Sie gehören auch zu keiner Risikogruppe. Sie haben also kein erhöhtes Risiko.

Mit kleiner, aber nicht zu vernachlässigender Wahrscheinlichkeit produziert der HIV-Test ein falsch positives Resultat für Leute, die *nicht* infiziert sind. Da die allermeisten Menschen *kein* AIDS haben, also X nahe bei 100 ist, sind auch Sie höchstwahrscheinlich *nicht* infiziert, selbst wenn Ihr HIV-Test positiv ausfällt. □

Fallbeispiel 32: Wie viele als wirksam getestete Medikamente sind in Wirklichkeit unwirksam?[14]
Wir wissen das X natürlich nicht, also wie viel Prozent der Medikamente tatsächlich unwirksam sind. Als Rechenbeispiel nehmen wir hier einmal $X = 90$ und einmal $X = 20$ an. Im ersten Fall sind zehn von hundert neu entwickelten Medikamenten tatsächlich wirksam und neunzig unwirksam, im zweiten Fall sind achtzig von hundert wirksam und zwanzig unwirksam.

Gemäß zweiter Annahme oben gehen wir davon aus, dass alle wirksamen Medikamente positiv getestet sind, gut. Aber fünf Prozent der unwirksamen sind gemäß erster Annahme ebenfalls positiv getestet. Im ersten Fall, also $X = 90$, ergibt das zehn korrekterweise positiv getestete gegenüber viereinhalb fälschlich positiv getesteten Medikamenten. Somit wären $4,5/(10+4,5) \approx 31\,\%$ aller positiv getesteten Medikamente im ersten Fall unwirksam. Im zweiten Fall, $X = 20$, kommt hingegen nur ein falsch positives auf achtzig korrekt positive Ergebnisse.

Es spricht also einiges für möglichst strenge Zulassungsverfahren, eine einzelne Studie mit Signifikanzniveau 5 % wäre offenkundig kein ausreichend sicherer Nachweis. Nicht berücksichtigt in diesem Fallbeispiel ist ein weiteres Problem:

unzureichende Kontrolle der Studiendurchführung durch die Zulassungsbehörden.[15] □

Das Problem, dass ein statistisch signifikantes Ergebnis rein auf zufälligen Fluktuationen beruhen könnte, wirkt sich besonders stark aus, wenn die Rohdaten der Studie mehreren Tests unterworfen werden wie im nächsten Beispiel.

Fallbeispiel 33: Dutzende klinische Untersuchungen auf denselben Daten

Amerikanische Forscher haben sich 66 klinische Studien vorgenommen, die in angesehenen medizinischen Fachzeitschriften veröffentlicht wurden und in denen jeweils verschiedene statistische Tests auf den erhobenen Daten durchgeführt wurden, im Schnitt dreißig pro Studie.[16] Beim Standardwert 5 % für das Signifikanzniveau und einer solchen Anzahl von Tests ist die Wahrscheinlichkeit sehr hoch, dass darunter auch fälschlich positive Ergebnisse sind (40 % schon bei zehn, 64 % bei zwanzig Tests). Durch geeignete statistische Methoden[17] kann man eine Art gemeinsames Signifikanzniveau für mehrere Tests bestimmen. In 50 der 66 Arbeiten wurden statistisch signifikante Ergebnisse berichtet, die nach dieser Adjustierung nicht mehr statistisch signifikant waren. In den 51 Arbeiten, die ausgewählte statistisch signifikante Ergebnisse schon in der Zusammenfassung erwähnten, waren bei 40 Arbeiten auch solche Ergebnisse betroffen, bei 15 Arbeiten sogar alle in der Zusammenfassung erwähnten Ergebnisse. In keiner der betroffenen Arbeiten wurde dieses Problem thematisiert.

Man darf daher vermuten, dass in dem einen oder anderen Fall falsche Ergebnisse zur Veröffentlichung der Arbeit geführt haben. □

2.4 Rosinenpickerei und Survivorship Bias

Das Problem der zufälligen statistischen Fluktuationen kann zudem zum *Rosinenpicken* (engl. *cherry picking*) führen: Über die Forschungsfragen hinaus, die man mit einer Studie eigentlich beantworten wollte, kann man sich die Daten ja noch einmal intensiv von allen Seiten anschauen, ob sich nicht weitere Auffälligkeiten finden lassen. Findet man welche, kann man im Nachhinein weitere Forschungsfragen in die geplante Veröffentlichung aufnehmen, die dadurch scheinbar überzeugend beantwortet werden. Allerdings wird man aufgrund der statistischen Fluktuationen sehr häufig irgendwelche Auffälligkeiten finden, die überhaupt nichts besagen. Daher muss folgende goldene Regel bei Design und Auswertung einer Studie unbedingt eingehalten werden, das passiert aber nicht immer, oft wohl einfach aus Unwissenheit.

> Die Forschungsfragen, die mit der Studie beantwortet werden sollen, müssen vorab unverrückbar festgelegt sein, also insbesondere bevor die Ergebnisse der Studie bekannt sind. Und nur diese Forschungsfragen allein dürfen durch die Studie als valide beantwortet gelten!

Andere Ergebnisse der Studie dürfen ebenfalls gerne genannt und diskutiert werden, aber diese sind zunächst einmal rein spekulativ und müssen auch strikt so behandelt werden.

Eine Variation der Rosinenpickerei ist, den Grenzwert zwischen auffälligen und nicht auffälligen Ergebnissen erst

endgültig festzusetzen, nachdem das Studienergebnis vorliegt.[18] Eine andere Variante ist, eine Studie vorzeitig abzubrechen, sobald das gewünschte Ergebnis erscheint. Da die Zwischenergebnisse naturgemäß häufig auf und ab gehen – vor allem in den frühen Phasen der Studie, wenn der bis dahin ausgewertete Teil der Stichprobe noch nicht so groß ist –, ist die Wahrscheinlichkeit für ein falsch positives Ergebnis aufgrund von statistischen Fluktuationen hierbei natürlich besonders hoch.[19]

Der *Survivorship Bias* aus der Überschrift dieses Abschnitts wird im gleichnamigen (deutschsprachigen) Wikipedia-Artikel mit „Verzerrung zugunsten der Überlebenden" übersetzt. In gewisser Weise ist das eine Steigerung der Rosinenpickerei: Man erstellt überhaupt keine Statistik, sondern schaut sich nur ein paar Einzelfälle an, die die Behauptung unterstützen, übersieht aber die vielen Fälle, die *gegen* die Behauptung sprechen. In der Regel wird das wohl unabsichtlich passieren, weil die positive Art von Fällen im Licht steht, die negative Art im Schatten. Zum Beispiel sehen Sie erfolgreiche Sportler, Schauspieler und Gründer von Start-up-Unternehmen täglich in den Medien – aber die vielen erfolglosen, die genauso hart arbeiten und den Erfolg vielleicht genauso verdient hätten, sehen Sie nicht. Sie lesen in den Wallfahrtskirchen Dankschreiben für erhörte Gebete, aber zu den nicht erhörten Gebeten lesen Sie nichts. Sie finden enthusiastische Aussagen von Patienten, die von einem ganz besonderen Heiler von ihren Krankheiten befreit wurden, aber von den vielen Patienten, denen er nicht helfen konnte oder sogar geschadet hat, finden Sie nichts.

Einzelfälle ersetzen niemals eine seriöse Statistik, denn nur damit kann eingeschätzt werden, ob diese Einzelfälle repräsentativ oder vereinzelte Ausnahmen sind.

Der Survivorship Bias tritt besonders gerne im ökonomischen Bereich auf und ist sarkastisch gesprochen die methodische Grundlage für gewisse Bereiche der publikumswirksamen Beratungsliteratur im Bereich Ökonomie.

Fallbeispiel 34: Erfolgreiche Unternehmen
Wenn Sie sich wenige Firmen heraussuchen, die in letzter Zeit besonders erfolgreich waren, dann werden Sie schon irgendwelche Gemeinsamkeiten finden, die plausibel den Erfolg dieser Firmen erklären. In der Regel muss man aber davon ausgehen, dass es weit mehr Firmen gibt, die dieselben Gemeinsamkeiten aufweisen, aber ganz und gar nicht erfolgreich sind. Der Erfolg einer Firma scheint sehr viel mehr von glücklichen Zufällen und Fügungen abzuhängen, als man vielleicht denkt.[20]

Eine interessante Frage ist, ob vielleicht Gerd Gigerenzer in seinem Bestseller „Risiko" dem Survivorship Bias aufgesessen ist.[21] Über zwei Unternehmer, die an einer Podiumsdiskussion teilnahmen, schreibt er dort, offenbar zustimmend: „Sie hätten ihr Vermögen erworben, indem sie ihren Bauchgefühlen vertrauten, von denen sie selten getäuscht wurden". Nun ja, diejenigen, die im Vertrauen auf ihr Bauchgefühl Schiffbruch erlitten haben, sitzen natürlich nicht auf dem

Podium, sondern vielleicht eher auf der Straße oder beim Insolvenzberater. Weiter unten schreibt er: „Weniger defensive Entscheidungen würden Vorteile gegenüber Konkurrenten verschaffen".[22] Voraussetzung ist natürlich, dass das nicht schiefgeht; weniger defensive sind nun einmal per Definition riskante Entscheidungen. Wenn Sie etwa an Edzard Reuters „integrierten Technologiekonzern"[23] oder an Jürgen Schrempps „Hochzeit im Himmel" zwischen Daimler und Chrysler[24] denken – um nur beispielhaft zwei richtig große Schiffbrüche zu nennen –, dann sind sicherlich Zweifel daran erlaubt, dass „weniger defensive" Entscheidungen wirklich immer so vorteilhaft sind, wie es der Fokus auf ein paar positive Fälle – eben der Survivorship Bias – suggeriert. □

2.5 Signifikanz vs. *statistische* Signifikanz

Die in Abschn. 2.3 eingeführte *statistische Signifikanz* eines Ergebnisses kann man als ein Maß dafür interpretieren, wie wahrscheinlich der festgestellte Effekt tatsächlich real ist und nicht fälschlich durch statistische Fluktuationen hervorgerufen wurde. Sie ist *kein* Maß für die Stärke dieses Effekts.

Im allgemeinen Sprachgebrauch hingegen bedeutet Signifikanz aber genau das: wie stark, bedeutsam, gewichtig oder deutlich der Effekt ist. Selbst wenn seriöserweise das Attribut „statistisch" dabei steht, wird dieser Unterschied häufig übersehen: Man liest den Begriff „statistisch signifikant", realisiert aber gar nicht unbedingt, dass das etwas anderes ist als das, was man mit dem Begriff „signifkant" eigentlich verbindet.[25]

Fallbeispiel 35: Krebsrisiko von Wurst
Während ich die Vorlesungsreihe vorbereitet habe, aus der dieses Buch hervorgegangen ist, ging gerade durch die Medien, dass die Internationale Krebsforschungsagentur IARC den Verzehr von Wurst und verarbeitetem Fleisch bezüglich Krebsrisiko auf dieselbe Stufe stellt wie Tabak und Asbest. Wie das? Wurst so tödlich wie Asbest? Nein, was in den Medien weitgehend unterging, war dies: Die Stufen besagen *nicht*, wie gefährlich die Stoffe sind, sondern sie besagen, wie gesichert der Zusammenhang ist[26] – *statistische* Signifikanz eben und nicht *Signifikanz*. □

Insbesondere bei großen Studien ist Vorsicht angebracht: Wenn die Fallzahl nur groß genug ist, dann ist auch ein Verhältnis von 51:49 oder sogar ein Verhältnis von 50,1 zu 49,9 ein statistisch signifikanter Unterschied zur Nullhypothese 50:50. Das ist aber sicher nicht das, was man landläufig unter einem signifikanten Ergebnis verstehen würde.

2.6 Fehler mit Durchschnittswerten

Die Rohdaten müssen zu einzelnen Kennzahlen verdichtet werden, um sie zu verstehen und daraus Schlussfolgerungen ableiten zu können. In der Regel werden Durchschnittswerte gebildet. Das ist nicht falsch, aber tückisch.

Fallbeispiel 36: Ost-West oder Nord-Süd?
Immer noch werden gerne Vergleiche zwischen Ost und West in Deutschland angestellt, also zwischen der ehemaligen DDR und der „alten" Bundesrepublik. Es werden also Daten

in den einzelnen Bundesländern oder Landkreisen erhoben und zu zwei Zahlen zusammengefasst: Durchschnitt Ost und Durchschnitt West.

In manchen Fällen lohnt der Blick auf die Einzeldaten, weil das vermeintliche Ost-West-Gefälle hin und wieder in Wirklichkeit ein Nord-Süd-Gefälle ist, nämlich wenn die Werte im Norden der „alten" Bundesrepublik eher ähnlich denen in der ehemaligen DDR sind, die Werte im Süden hingegen tendenziell anders aussehen. Dies gilt beispielsweise für Arbeitslosenzahlen.[27] □

Nach diesem ersten, einstimmenden Fallbeispiel[28] betrachten wir drei klassische logische Fehler, die bei der Interpretation von Durchschnittswerten leicht passieren: das *Simpson-Paradoxon*, das *Will-Rogers-Paradoxon* und den *ökologischen Fehler*; zuerst das klassische Beispiel für das Simpson-Paradoxon.

Fallbeispiel 37: Der Berkeley-Fall

In den Siebzigern wurde die Universität Berkeley verklagt, weil 44 % der männlichen, aber nur 35 % der weiblichen Studienbewerber 1973 zugelassen wurden. Die Aufschlüsselung nach einzelnen Studiengängen zeigte allerdings ein sehr uneinheitliches, keinesfalls systematisch diskriminierendes Bild. Der Unterschied in den universitätsweiten Gesamtzulassungsquoten bei Männern und Frauen lässt sich damit erklären, dass die Frauen in großer Zahl Studiengänge mit sehr hohen Ablehnungsquoten – zum Beispiel Englisch – gewählt haben und dementsprechend häufiger abgelehnt wurden als die Männer, die unter anderem häufiger Maschinenbau wählten, wo die Ablehnungsquote eher gering war.[29] □

Das Simpson-Paradoxon, benannt nach dem britischen Statistiker Edward Hugh Simpson, kann immer dann auftreten, wenn verschiedene Daten in einen Topf geworfen werden, bei denen ein entscheidender Indikator im Hintergrund steht, nach dem die Ergebnisse eigentlich aufgeschlüsselt werden müssten, um paradoxe Gesamtergebnisse zu vermeiden.[30] Konkret im Berkeley-Beispiel war die unterschiedliche Studiengangswahl von Männern und Frauen der Indikator im Hintergrund, genauer: die unterschiedlich hohe durchschnittliche Ablehnungsquote in den von Männern beziehungsweise Frauen bevorzugten Fächern.

Fallbeispiel 38: Raucherinnen leben länger als Nichtraucherinnen

Das Alter ist hier der entscheidende Indikator im Hintergrund: Der Anteil der Raucherinnen an allen Frauen wächst von Jahrgang zu Jahrgang tendenziell, weil im Laufe der Jahrzehnte immer mehr junge Frauen mit dem Rauchen angefangen hatten. Das heißt, unter den Jahrgängen mit der höchsten Wahrscheinlichkeit, noch mindestens zwanzig Jahre zu leben, gibt es besonders viele Raucherinnen.[31] Raucherinnen haben somit tatsächlich eine größere Chance als Nichtraucherinnen, die nächsten zwanzig Jahre zu überleben, aber dieser Effekt kehrt sich um, wenn man nur Frauen desselben Jahrgangs miteinander vergleicht. □

Auch das *Will-Rogers-Paradoxon* kann die Realität ins Gegenteil verkehren.[32] Das Zitat, das den amerikanischen Komiker Will Rogers zum Namensgeber gemacht hat, lautet sinngemäß ins Deutsche übersetzt: Als die Okies von Oklahoma nach Kalifornien zogen, hob das die durchschnittliche Intelligenz in beiden Bundesstaaten.

„Okies" bezeichnet dabei eine bestimmte Gruppe von Arbeitsmigranten, die nach der Großen Depression, zu der dann noch die Große Dürre hinzukam, aus dem mittleren Westen nach Kalifornien zogen – vor allem aus Oklahoma, daher „Okies".[33] Unter welcher Voraussetzung hätte Will Rogers mit seiner scheinbar paradoxen Aussage recht gehabt: wenn die durchschnittliche Intelligenz der „Okies" geringer als die der restlichen Einwohner von Oklahoma, aber höher als die der Einwohner von Kalifornien ist. Mr. Rogers hat seine spöttische Behauptung natürlich nicht auf eine solide wissenschaftliche Studienlage gegründet.

Nehmen wir im nächsten Fallbeispiel wieder die Medizin als ernsthaftes Beispiel.

Fallbeispiel 39: Therapiererfolg (vorher-nachher)
Betrachten Sie der Einfachheit halber eine Studie zu einer Therapie, bei der vor und nach der Therapie jeweils derselbe physiologische Indikator der Patienten bestimmt wird, zum Beispiel ein bestimmter Blutwert. In beiden Fällen – vorher und hinterher – werden die Probanden nach demselben Grenzwert in leichte Fälle (Gruppe L) und schwere Fälle (Gruppe S) unterteilt und in beiden Gruppen jeweils der Durchschnitt des Indikators berechnet.

Wenn die Therapie wirkt, dann werden einige Patienten den Grenzwert zwischen den beiden Gruppen unterschreiten, also von S nach L wechseln. In Gruppe S waren das wohl tendenziell eher leichtere Fälle, und wenn einige tendenziell leichtere Fälle nicht mehr in der Gruppe S sind, dann wird der Durchschnitt in S natürlich schlechter. In Gruppe L werden dieselben Patienten aber eher zu den schwereren Fällen gehören, das heißt, auch in L verschlechtert sich der

Durchschnitt. Insgesamt führt die Wirksamkeit der Therapie so zu einer Verschlechterung der Durchschnittswerte in beiden Gruppen, und es sieht so aus, als wäre die Therapie nicht nur unwirksam, sondern sogar schädlich! □

Zum Schluss dieses Abschnitts noch der ökologische Fehler. Der Name hat nichts mit Ökologie im üblichen Sinne zu tun. Der Fehler ist: Aus statistischen Zusammenhängen zwischen Indikatoren wird fälschlich geschlossen, dass ein entsprechender Zusammenhang bei den einzelnen Individuen besteht. Dieser Fehlschluss tritt beispielsweise häufig bei demographischen Daten auf, hier das klassische Beispiel von William S. Robinson, der den Begriff Ecological Fallacy geprägt hat.

Fallbeispiel 40: Robinsons Paradoxon
Die Alphabetisierungsrate in den einzelnen Bundesstaaten der USA war 1930 ziemlich stark positiv korreliert mit dem Prozentsatz der außerhalb der USA geborenen Bevölkerung. Der naive Schluss von der Statistik auf das Individuum – der ökologische Fehler eben – ist, dass die Lese- und Schreibfertigkeiten unter Migranten höher sind als bei der autochthonen Bevölkerung. Tatsächlich war er aber niedriger. Der wahre Zusammenhang ist wohl eher, dass Migranten sich vorzugsweise in Bundesstaaten niederlassen, wo auch die Alphabetisierungsrate seinerzeit relativ hoch war.[34] Der Zusammenhang ist eben rein statistischer Natur und muss für die Individuen hinter der Statistik überhaupt nicht bestehen – ja, kann wie in diesem Beispiel sogar entgegengesetzt sein, denn in der Tat war die Alphabetisierungsrate unter den Immigranten leicht niedriger als unter der autochthonen Bevölkerung. □

Nun ersetzen Sie die Alphabetisierungsrate durch die Kriminalitätsrate, und Sie haben ein aktuelles Beispiel: Aus der geographischen Verteilung von Ausländeranteil und Kriminalitätsrate lässt sich ohne Zusatzinformation erst einmal gar nichts schließen, denn beides ist in Städten höher als auf dem Land.

2.7 Temporale Fehlinterpretationen

Wann immer die zeitliche Entwicklung von Indikatoren eine Rolle spielt, ergeben sich daraus spezifische weitere Fehlerquellen. Die nächsten beiden Fallbeispiele zeigen, dass man bei zeitbehafteten Indikatoren nicht nur auf einzelne Zeitpunkte schauen darf, sondern auch den Kontext davor berücksichtigen muss. Denn was an einem Zeitpunkt so passiert, sieht manchmal spektakulärer aus, als es im zeitlichen Kontext gesehen tatsächlich ist.

Fallbeispiel 41: Seit wann überaltert unsere Gesellschaft?
Schon das gesamte zwanzigste Jahrhundert hindurch: So ist etwa der Jugendanteil an der Gesamtbevölkerung im letzten Jahrhundert von 44 auf 20 Prozent gefallen, und der Anteil Rentner ist um den Faktor 3 gestiegen.[35] □

Fallbeispiel 42: Selbstmordrate bei der französischen Telecom
Vor einigen Jahren wurde in den Medien weltweit über eine Selbstmordwelle bei der französischen Telecom berichtet, und als Ursache wurde durchgängig – soweit mir bekannt – die damalige Sparpolitik ausgemacht. Allerdings ist

die Suizidzahl gemessen an der Gesamtzahl Mitarbeiter im betreffenden Zeitraum eigentlich gar nicht allzu hoch.[36] Sie sieht nur so beeindruckend aus, weil das Unternehmen so viele Mitarbeiter hat. Da die Mitarbeiter eines Staatsunternehmens sicher alles andere als repräsentativ für die Gesamtbevölkerung sind, lässt sich schwer aus den allgemeinen Daten zu Suiziden bestimmen, welche Suizidrate bei der Telecom „normal" wäre. Hilfreich wären Daten darüber, wie sich die Suizidrate bei der Telecom zeitlich entwickelt hat. Nur wenn sie nach Einführung der Sparpolitik deutlich im Vergleich zur allgemeinen Suizidrate *angestiegen* ist, wird man wohl bei aller Vorsicht auf einen Zusammenhang schließen können. □

Auch wenn man den gesamten relevanten zeitlichen Kontext betrachtet, sind immer noch verschiedene Arten von Fehlschlüssen möglich. Das medizinische Fallbeispiel zum Will-Rogers-Paradoxon (Nummer 39 in Abschn. 2.6) ist auch ein Beispiel für *temporale* Fehlschlüsse, weil die Individuen eben im Laufe der Zeit von einer Kategorie in die andere wechseln. Neben dem Will-Rogers-Paradoxon kann bei solchen Kategorisierungen im Laufe der Zeit auch Folgendes passieren:

Minimale Schwankungen des Indikators bei einzelnen Individuen rund um den Grenzwert können dazu führen, dass eine größere Zahl von Individuen gerade so eben den Grenzwert von einer Kategorie A in die nächste Kategorie B überwindet oder umgekehrt. Wenn diese Fluktuationen nicht rein zufällig sind, sondern systematisch in die Richtung von A nach B verzerrt, dann ändern sich die Zahlen in den beiden Kategorien auf Dauer entsprechend stark. Ein solches Ergebnis muss dann natürlich korrekt interpretiert werden, nämlich als eine Tendenz in der Richtung von A nach B, die

zwar ausreichend viele Individuen betrifft, um bemerkbar zu sein, aber der Sprung jedes einzelnen Individuums kann trotzdem denkbar klein sein. Sieht man sich nur die Kategorien an, dann ist das Ergebnis dramatisch, aber die durchschnittliche Sprungweite der Individuen ist in diesem Exempel alles andere als dramatisch, siehe Abb. 2.1.

Bleiben wir noch bei Kategorien und Grenzwerten. Bei der Betrachtung von längeren Zeiträumen taucht immer wieder das Problem auf, dass die Definitionen sich im Laufe der Zeit ändern, wie in den folgenden beiden Beispielen.

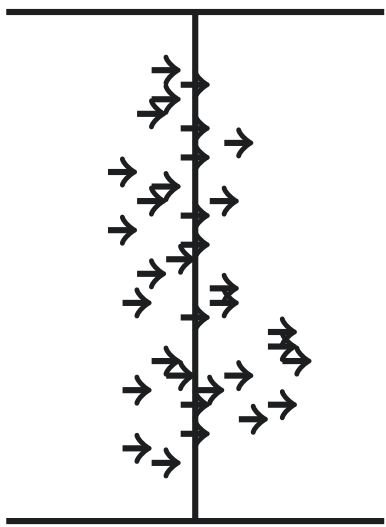


Abb. 2.1 Wenn viele Elemente der Stichprobe einen kleinen Schritt nach rechts machen, bleibt die Gesamtsituation nahezu gleich, aber die Anzahl von Elementen in den beiden Kategorien hat sich möglicherweise deutlich geändert

Fallbeispiel 43: Morbus Grenzwert
Wenn Sie aus Studien über Jahre oder Jahrzehnte hinweg den prozentualen Anteil von übergewichtigen Menschen oder von Menschen mit erhöhtem Blutdruck oder ähnlichem betrachten, dann werden Sie häufig eine bedenkliche Verschlechterung der Gesundheit in der Bevölkerung feststellen. Das liegt in manchen Fällen einfach daran, dass die Grenzwerte für diverse medizinische Indikatoren im Laufe der Zeit schrittweise eher enger gesetzt worden sind: Wer gestern noch zu den Gesunden gezählt wurde, zählt heute als gefährdet oder krank.[37] □

Wie das letzte Fallbeispiel zeigt, kann man also nicht einfach die Studienergebnisse der verschiedenen Zeitpunkte sammeln, sondern müsste eigentlich sämtliche Rohdaten nehmen und daraus nach identischem Schema die Ergebnisse neu berechnen. Das passiert selten, ist ja auch sehr aufwendig, und die Rohdaten sind auch gar nicht immer verfügbar.

Fallbeispiel 44: Die Bahn tut etwas gegen Verspätungen
Die Überschrift dieses Fallbeispiels stimmt natürlich. Das weiß ich deshalb so genau, weil ich im Rahmen meiner Forschung in die Bemühungen der Deutschen Bahn involviert bin. Aber manchmal sorgt einfach nur die Zählweise für einen „Erfolg" bei der Bekämpfung von Verspätungen. Bei der Schwedischen Bahn wurde 2012 der Grenzwert, ab wann eine Verspätung in die Statistik eingeht, von fünf auf fünfzehn Minuten erhöht, was natürlich zu einem einmaligen Sondererfolg in der Verspätungsstatistik führte.[38] □

Es muss nicht immer an einer Änderung von Grenzwerten liegen; auch Störfaktoren müssen über die Zeit hinweg nicht unbedingt konstant bleiben.

Fallbeispiel 45: Lokale Klimaveränderung
Wetterstationen, die bei ihrer Einrichtung noch außerhalb der Stadt waren – etwa an Flughäfen –, können über die Jahrzehnte hinweg von der wachsenden Stadt erreicht und umzingelt werden. In Städten sind die Temperaturen aber um ein paar Grad höher als auf dem freien Land. Oder die Stadt wird durch Zuzug rund um die Wetterstation noch dichter besiedelt, was die Temperatur ebenfalls steigen lässt. Der Fachbegriff lautet *Urban Heat Island* (*UHI*). Die daraus resultierenden Verfälschungen von Temperaturmessungen müssen bereinigt werden, um nicht solche Effekte, sondern nur reale Klimaveränderungen zu messen.[39] □

Als letzte Art von temporalem Fehlschluss betrachten wir jetzt noch die *Regression zur Mitte*.

Fallbeispiel 46: Bestrafen statt belohnen?
Wenn man die besten Leute belohnt und die schlechtesten bestraft, dann kann man nicht selten feststellen, dass die besten danach nicht mehr die besten und die schlechtesten nicht mehr die schlechtesten sind. Daraus wurde schon häufig gefolgert, dass Bestrafung besser als Belohnung funktioniere, ja, dass Belohnung eher kontraproduktiv sei.

Das ist aber ein Trugschluss, denn wenn die Leistungen auch nur zum Teil vom Zufall abhängen, zum Beispiel von der Tagesform, dann sind die besten und die schlechtesten Ergebnisse in der Regel seltene statistische Ausreißer. Auch ohne Belohnung und Bestrafung wäre wohl ungefähr dasselbe passiert: Die besten und schlechtesten Ergebnisse lassen sich nicht reproduzieren, die besten und die schlechtesten Probanden „regredieren“ also zur Mitte.[40] □

2.8 Sind die Ergebnisse überhaupt relevant?

Was gemessen wird und was man eigentlich herausfinden möchte, das sind erst einmal zwei verschiedene Dinge, die leider oft nicht unbedingt gut zusammenpassen. Der Unterschied zwischen beidem geht zudem häufig auf dem Weg in die breite Öffentlichkeit etwas unter.

Fallbeispiel 47: Wie misst man Werbeerfolg?
Zum Beispiel, indem man misst, wie viel Aufmerksamkeit erregt wird, ob das Image positiv ist oder ob die Werbung in Erinnerung bleibt. Inwieweit das mit dem eigentlichen Ziel von Werbung in Beziehung steht, nämlich dass das umworbene Produkt daraufhin dann auch *gekauft* wird, ist unklar.[41] □

Fallbeispiel 48: Was misst ein Intelligenztest?
Es gibt ein Bonmot: Ein Intelligenztest misst die Fähigkeit, schnell und erfolgreich Intelligenztests zu absolvieren. Diese Antwort ist zwar sicher richtig, aber nicht so recht hilfreich. Offensichtlich misst man ein ganzes Bündel von intellektuellen und mentalen, dauerhaft gleich bleibenden wie auch zeitlich sich ändernden Persönlichkeitsmerkmalen. Es liegt in der Natur der gängigen Tests, dass sie insbesondere Folgendes messen:

1. die Fähigkeit, *unterkomplexe* Aufgaben zu lösen, also Aufgaben, in denen (a) die Frage, die es zu beantworten gilt, präzise formuliert ist und nicht erst herausgearbeitet

werden muss, (b) alle infrage kommenden Optionen bekannt und ebenfalls präzise formuliert sind, (c) genau eine Option richtig ist, (d) alle notwendigen Informationen bekannt sind und (e) keine Abwägung vorgenommen werden muss;
2. die *Motivation*, diese Art von unterkomplexen Aufgaben zu lösen, denn je motivierter jemand für eine Aufgabe ist, umso besser wird er in der Regel abschneiden.

In der Psychologie soll Intelligenz „ein Sammelbegriff für die kognitive Leistungsfähigkeit des Menschen“[42] sein, was auch dem Alltagsverständnis entsprechen dürfte. Offensichtlich gibt es hier eine Diskrepanz. □

In der Medizin lautet der Fachbegriff *Surrogatmarker* für Indikatoren in klinischen Studien, mit denen man eigentlich etwas anderes messen will.[43]

Fallbeispiel 49: Body-Mass-Index (BMI)
Zur Berechnung des Body-Mass-Index eines Menschen wird das Körpergewicht in Kilogramm durch die Körpergröße in Metern und dann noch einmal durch die Körpergröße in Metern geteilt. Der BMI soll eigentlich medizinisch bedenkliche Formen von Übergewicht messen, kann aber nicht unterscheiden zwischen Fett, Muskeln und Knochen und auch nicht zwischen verschieden problematischen Fettzonen.[44] □

Fallbeispiel 50: Tierversuche
Die Frage, wie weit Versuchsergebnisse an Tieren sich auf den Menschen übertragen lassen, lässt sich nicht pauschal beantworten, sondern gut fundiert nur von Fall zu Fall und im Grunde auch erst im Nachhinein.

Freedman berichtet von einem Fall, bei dem die Versuchstiere selbst bei fünfhundertfacher Überdosierung nicht erkennbar geschädigt wurden, aber beim ersten Test an Menschen – natürlich mit normaler Dosis – wurden mehrere Testteilnehmer bedrohlich krank. In einer Fußnote sinniert er umgekehrt darüber, ob Penicillin in einem Verfahren nach heutigen Standards überhaupt zugelassen worden wäre, da Kaninchen und Meerschweinchen dieses Antibiotikum nicht vertragen.[45] □

Wenn keine Studien für den eigentlich interessierenden Kontext vorliegen, liegt der Gedanke nahe, Studienergebnisse aus einem anderen Kontext zu übertragen, beispielsweise von einem Land auf ein anderes.

Fallbeispiel 51: Bildungsforschung anderswo
Viele Studien über Hochschulen werden in den USA oder anderswo im angelsächsischen Bereich erstellt. Das System dort ist in vielerlei Hinsicht subtil, aber entscheidend anders. Ein Beispiel aus meiner eigenen Praxis ist Interaktion zwischen Dozent und Studierenden während der Vorlesungstermine, das heißt, die Studierenden hören nicht die ganze Zeit über passiv dem Monolog des Dozenten zu, sondern nehmen aktiv teil in Form von Gruppendiskussionen und Stillarbeitsphasen, in denen sie kleine Aufgaben bearbeiten.

Auch wenn eine Methode in den USA als erfolgreich evaluiert wurde, muss sie bei uns noch lange nicht erfolgreich sein, zum Beispiel aus dem Grund, dass Anwesenheit in den USA in der Regel viel verbindlicher ist als bei uns. Die Studierenden dort können einer aktivierenden und somit unbequemen Ausgestaltung der Vorlesung, auf die man

sich vielleicht sogar noch vorher vorbereiten müsste, nicht so leicht durch Abwesenheit ausweichen wie bei uns. Möglicherweise sind sie in den USA auch durch die hohen Studiengebühren motivierter als bei uns. □

Wenn es um demographische Daten geht, kommt immer wieder das Problem der Dunkelziffer ins Spiel, also die Diskrepanz zwischen der unbekannten realen Fallzahl und der oft weit niedrigeren Zahl registrierter Fälle.

Fallbeispiel 52: Wie belastbar sind Kriminalitätsstatistiken?

Kriminalitätsstatistiken enthalten natürlich nicht die realen Kriminalitätsfälle, sondern die offiziell registrierten. Der Unterschied – sprich: die Dunkelziffer – dürfte so hoch sein, dass die Statistik kaum aussagekräftig ist. Auch statistische Veränderungen von Jahr zu Jahr müssen nicht real sein, sondern können auch auf Veränderungen bei der Registrierung zurückzuführen sein. So sind Bürger, Polizisten und Staatsanwälte bei bestimmten Arten von Straftaten in den letzten Jahren sicherlich sensibler geworden und nun eher bereit als früher, konkrete Fälle anzuzeigen beziehungsweise zu verfolgen. Auch regionale Unterschiede in der Arbeit von Polizei und Staatsanwaltschaft sind insbesondere im föderalen deutschen System nicht auszuschließen, was Vergleiche zwischen verschiedenen Bundesländern problematisch macht.

Rainer Wendt, Bundesvorsitzender der Deutschen Polizeigewerkschaft, spricht von einer noch krasseren möglichen Ursache für Verzerrungen: „Es stimmt, manche Straftaten werden weniger, doch das kann man steuern. Wenn ich als Polizeichef will, dass in meiner Stadt die

Rauschgiftkriminalität sinkt, dann schicke ich die dafür zuständigen Kollegen in die Verkehrskontrolle. Dann verspreche ich Ihnen, dann sinkt die Rauschgiftkriminalität – zumindest statistisch".[46] □

Ich denke, gerade angesichts dieses Zitats ist die Frage erlaubt, ob die Kriminalitätsstatistik überhaupt irgendetwas Belastbares aussagt.[47] Entsprechendes gilt natürlich auch für andere Bereiche, etwa Registrierung von Krankheitsfällen.[48]

Im letzten Fallbeispiel, Kriminalitätsrate, weiß man aufgrund der Dunkelziffer nicht genau, wie die Realität aussieht; im nächsten Beispiel stellt sich eher die Frage, was man überhaupt als Realität ansehen sollte.

Fallbeispiel 53: Wie hoch ist die Arbeitslosenzahl?
„Zusätzliche amtliche Schätzungen, die für das vollständige Bild der Arbeitslosigkeit wichtig sind, [sind] ebenfalls ohne Probleme im Internet allgemein zugänglich".[49] Etwas überspitzt, heißt das also: Schön, wenn in den Medien wieder einmal nur kommentarlos die offizielle Arbeitslosenzahl genannt wird, dann müssen Sie als Leser nur in das Internet schauen und den traditionellen Job von Journalisten – Recherche – selbst erledigen, um „das vollständige Bild der Arbeitslosigkeit" zu erhalten. Dem Verfasser des einleitenden Zitats, Florian Diekmann, ist es sicherlich hoch anzurechnen, dass er in dem in Anmerkung 49 zitierten Artikel einmal auflistet, welche Personengruppen man durchaus noch zusätzlich in die Arbeitslosenstatistik aufnehmen könnte oder nach der offiziellen Definition im Sozialgesetzbuch vielleicht sogar sollte,[50] und dass er für Februar 2017 einmal vorrechnet, dass in diesem Unterschied mindestens eine Million Menschen stecken.

Allerdings bleibt noch das Problem aus Abschn. 2.7: Besonders interessant ist immer der Verlauf der Arbeitslosenzahl oder -quote über die Zeit hinweg, der Vergleich von heute mit früher. Da die offizielle Definition der Arbeitslosenzahl und auch die Erhebungsmethode im Laufe der Zeit durchaus geändert werden, muss man genau hinschauen, welche Zahl in Jahr X man mit welcher Zahl in Jahr Y vergleichen muss, um nicht Äpfel mit Birnen zu vergleichen. Über längere Zeiträume hinweg führt die *offizielle* Arbeitslosenzahl aufgrund dieser methodischen Änderungen jedenfalls in die Irre.[51] □

Auch in anderen Bereichen als der Demographie stellt sich häufig die Frage, wie aussagekräftig Zahlen sind.

Fallbeispiel 54: Interessiert Sie die offizielle Inflationsrate wirklich?
Machen Sie sich zunächst einmal klar, dass jeder Mensch seine eigene individuelle Inflationsrate hat, die natürlich durch sein Konsumverhalten bestimmt ist. Zum Beispiel: Je nachdem, wie viel Geld Sie anteilig für Lebensmittel oder auch für Heizöl und Benzin ausgeben, wirken sich die Lebensmittelpreise beziehungsweise der Ölpreis auf Ihre persönliche Inflationsrate aus. Die Unterschiede in persönlichen Inflationsraten können durchaus so groß sein, dass die offizielle Inflationsrate keine Aussagekraft für das Individuum mehr hat.[52]

Hinzu kommt ein grundsätzlich unlösbares Problem: Früher hatten Fernseher einen Röhrenbildschirm, heute einen Flachbildschirm; früher waren Filterkaffeemaschinen vorherrschend, heute Kapseln, Pads und Vollautomaten; früher war ein Handy einfach ein Telefon, heute ist es ein Smartphone, und so weiter. Produkte ändern sich – angeblich werden sie immer besser. In der Inflationsrate muss diese

Wertsteigerung natürlich berücksichtigt werden, dies nennt man die *hedonische Preisbereinigung*. Das Problem ist, dass diese Wertsteigerung nicht annähernd objektiv zu beziffern ist. Der Preis wäre der naheliegende Indikator für den Wert, aber das wäre hier offensichtlich witzlos, denn der Preisanstieg kann ja schlecht durch sich selbst bereinigt werden. □

Die Relevanz eines Studienergebnisses kann auch durch eine verzerrte Stichprobe eingeschränkt sein.

Fallbeispiel 55: Psychologische Studien
In der Psychologie nehmen in vielen Fällen überwiegend oder allein Studierende teil, die natürlich keineswegs repräsentativ für die Gesamtbevölkerung sind. Tatsächlich müssen Studierende der Psychologie vielerorts im Laufe ihres Studiums an Studien in der eigenen Hochschule teilnehmen, sonst bekommen sie ihren Abschluss nicht. □

Fallbeispiel 56: Medizinische Studien
Im medizinischen Bereich wird in vielen Fällen Geld an die freiwilligen Studienteilnehmer gezahlt, was vor allem finanziell schwächere Menschen anzieht. Die Stichprobe ist also auch in diesem Fall nicht unbedingt repräsentativ für die Gesamtbevölkerung. □

2.9 Unzulässiger Schluss auf den Einzelfall

Ihr Arzt, Ihr Finanzberater oder ein Verkäufer begründet seine Vorschläge, was Sie seiner Ansicht nach tun sollen, vielleicht mit Statistiken. Selbst wenn diese Statistiken seriös

sind, bleibt für Sie immer noch etwas Grundsätzliches zu bedenken: Häufig sind die einzelnen Fälle sehr unterschiedlich; der scheinbar typische Fall nah am fiktiven Durchschnittsfall ist dann eher selten, und eine starke Abweichung vom Durchschnitt ist der Normalfall.

> Aus Durchschnittswerten und ähnlichen Kennzahlen kann man ohne weitere Zusatzinformation *nichts* für den Einzelfall folgern!

Fallbeispiel 57: Medizinische Grenzwerte
Es mag ja vielleicht sogar sein, dass Indikatoren wie beispielsweise der Body-Mass-Index (vgl. Fallbeispiel 49 in Abschn. 2.8) im Durchschnitt eine gewisse Aussagekraft haben, aber für das Individuum gilt das nicht unbedingt. Gerne wird Arnold Schwarzenegger als Gegenbeispiel hergenommen, weil dessen Body-Mass-Index selbst in seinen besten Zeiten als hochgradig adipös einzustufen war, was wohl weniger an seinem Fettanteil lag.

Dasselbe gilt durchaus auch für Laborwerte: Nicht für jedes Individuum ist ein Laborwert außerhalb der Grenzwerte bedenklich beziehungsweise nicht für jedes Individuum ist ein Laborwert innerhalb der Grenzwerte unbedenklich.[53] □

Fallbeispiel 58: Studienzulassung
Der Studienerfolg ist für viele Studienrichtungen schon relativ stark mit der Abiturnote korreliert – hoch genug, dass man hoffen kann, durch Auswahl nach Abiturnote die Studienstatistik zu verbessern; aber bei Weitem nicht hoch

genug, dass für den *einzelnen* Studierenden eine belastbare Prognose für den Studienerfolg aus der Abiturnote ableitbar wäre. □

Fallbeispiel 59: Daran merkst du, dass du intelligenter bist …

…als achtzig Prozent der Bevölkerung. Das ist der Titel eines Artikels in der deutschen Huffington Post.[54] Im Artikel selbst werden dann zwölf verschiedene Merkmale benannt, zum Beispiel, dass man lustig ist, dass man groß ist oder dass man als Kind Musikunterricht hatte. Die Behauptung im Text ist dann, diese zwölf Merkmale seien „Anzeichen, dass du intelligenter als 80 Prozent der Menschen bist". Der Unterschied zwischen der aufmerksamkeitsheischenden Überschrift und der tatsächlichen, wohl weitaus weniger attraktiven Faktenlage ist genau das Thema dieses Abschnitts: Die Überschrift suggeriert – natürlich völlig zu Unrecht –, dass Sie aus diesen Merkmalen auf *Ihre persönliche* Inteligenz schließen können. Der Zusammenhang gilt aber bestenfalls statistisch. □

Fallbeispiel 60: Wirksamkeit von Therapien

Daraus, dass eine Therapie – etwa ein Medikament – im Durchschnitt eine positive Wirkung hat, kann man ohne weitere Informationen erst einmal gar nichts für den einzelnen Patienten folgern. Die Therapie könnte immer noch für einen größeren Anteil von Patienten völlig wirkungslos oder unterm Strich sogar schädlich sein. □

Der Fehler, von statistischen Ergebnissen auf den Einzelfall zu schließen, ist – leicht überspitzt – die methodische Grundlage für pauschale Ernährungs- und Diättipps in den Medien,

in denen ein und derselbe Ratschlag *allen* Menschen gleichermaßen gegeben wird: „Der richtige Weg zu einer gesunden Ernährung lässt sich … nicht für alle vereinheitlichen."[55]

Fallbeispiel 61: Frühstück
Zum Beispiel, dass für *alle* Menschen ein reichhaltiges Frühstück die wichtigste Mahlzeit des Tages sei und man abends am Besten gar nichts mehr essen solle. Die optimale Verteilung der Kalorienaufnahme über den Tag ist aber offensichtlich eine sehr individuelle Angelegenheit, abhängig vom persönlichen Tagesrhythmus. Und es scheint auch durchaus einiges *gegen* ein reichhaltiges Frühstück als Hauptmahlzeit des Tages zu sprechen.[56] □

So wie man nicht vom Durchschnitt auf den Einzelfall schließen darf, kann man aus dem statistischen Vergleich zweier Gruppen nicht darauf schließen, dass der Unterschied auch für die einzelnen Individuen zutrifft.

Fallbeispiel 62: Demographische Unterschiede
Zum Beispiel Geschlecht: Immer wieder wird über Studien berichtet, dass Männer und Frauen sich in diversen kognitiven oder emotionalen Aspekten unterscheiden. Das mag für den Durchschnitt so stimmen oder auch nicht. Aber die Streuung *innerhalb* beider Geschlechter ist so extrem hoch im Vergleich zu den konstatierten statistischen Unterschieden *zwischen* den Geschlechtern, dass der Vergleich der Durchschnittswerte eigentlich gar nichts über den einzelnen Mann oder die einzelne Frau aussagt. Sie können sich das vielleicht am einfachsten anhand von sichtbaren Indikatoren wie der Körpergröße klarmachen: Viele Frauen sind größer als der

durchschnittliche Mann, und viele Männer sind kleiner als die durchschnittliche Frau. □

Übrigens: Wenn man aus einem Unterschied der Durchschnittswerte folgert, dass die Mitglieder verschiedener demographischer Gruppen unterschiedlich gesetzlich oder moralisch behandelt werden sollen, beispielsweise verbindliche Festschreibung von Geschlechterrollen, dann begeht man noch einen weiteren Denkfehler, den sogenannten *naturalistischen Fehlschluss*.[57] Kurz und knapp gesprochen, ist das einfach die unreflektierte Schlussweise: So wie es ist, so ist es gut und richtig.

2.10 Rankings

Kaum etwas beeinflusst die öffentliche Wahrnehmung und auch die Politik so sehr wie Rankings.

Fallbeispiel 63: Bundesländer im Vergleich
Die drei Stadtstaaten schneiden bei Rankings zu diversen Themen immer wieder schlecht ab. Zum Beispiel ist die Häufigkeit von Straftaten je 100.000 Einwohner in Berlin, Hamburg und Bremen deutlich höher als in den Flächenländern.[58] Aber wenn man ausnahmsweise einmal die Daten für die einzelnen Kommunen zu sehen bekommt, stellt man häufig fest, dass in den Flächenländern die großen Städte schlechter abschneiden als die ländlichen Kommunen. Die unterschiedlichen Ergebnisse von Stadtstaaten und Flächenstaaten sind daher mutmaßlich eher durch unterschiedlichen

Urbanitätsgrad als durch unterschiedliche politische Ausrichtung der jeweiligen Landesregierungen verursacht, wie oft gemutmaßt wird. □

Insbesondere bei Rankings ganzer Nationen – beispielsweise bei Bildungstests wie PISA und TIMMS – findet die zeitliche Entwicklung besonderes Interesse, also um wie viele Plätze man sich von einer Runde zur nächsten verbessert oder verschlechtert hat. Wenn man das Ganze als eine Art sportlichen Wettbewerb der Nationen ansieht, dann sind die Platzierungen natürlich sehr interessant. Ist man hingegen an Verbesserungen in der Sache interessiert – und das sollte man ja wohl vorrangig sein –, dann sollte man sich vielleicht eher ansehen, ob die eigene *Punktzahl* sich verbessert oder verschlechtert hat. Denn wenn unser Land um ein paar Plätze abgefallen ist, könnte es trotzdem sein, dass wir uns verbessert haben, aber ein paar andere haben sich halt noch ein bisschen mehr verbessert. Schön für die anderen – wirklich ein Grund zur Sorge für uns?

Wie stark unterscheiden sich eigentlich die Punktzahlen im Ranking? Die Indikatoren, die in die Punktzahl eingeflossen sind, können auch Schwankungen unterliegen. Bei manchen Produkten etwa sind Preise, chemische Zusammensetzung und Verunreinigungen nicht unbedingt immer hundertprozentig konstant. Die Erhebung dieser Daten ist dann eine Momentaufnahme, die kurz davor oder kurz danach vielleicht etwas anders ausgesehen hätte, Wenn die Punktzahlen der einzelnen Produkte nicht deutlich unterschiedlich sind, kann die genaue Platzierung im Ranking in solchen Fällen vom Zeitpunkt der Momentaufnahme abhängen.

Als Nächstes zwei Beispiele dafür, dass man zu unterschiedlichen Ergebnissen kommen kann, je nachdem, wie man die Einzelfälle zusammenfasst.

Fallbeispiel 64: Die beliebtesten Freizeitaktivitäten der Deutschen

Das Statistische Bundesamt schlüsselt die verschiedenen Sportarten auf, und die beliebteste Sportart (Besuch im Fitnessstudio) landet so erst auf Platz 7. Wären alle Sportarten pauschal zusammengefasst worden, dann wäre die Freizeitaktivität „Sport" offensichtlich auf einem der ersten Plätze gelandet. Umgekehrt wäre Shopping wohl nicht auf Platz 2 gekommen, wenn hier ähnlich stark wie bei Sport differenziert worden wäre, etwa wenn zwischen der Suche nach modischer Bekleidung und Schuhen einerseits und Unterhaltungselektronik andererseits unterschieden worden wäre.[59] □

Im nächsten, zweiten Beispiel sind die Neugeborenen die Einzelfälle, und wir erhalten eine völlig irreführende Information, wenn wir sie nach Vornamen zusammenfassen.

Fallbeispiel 65: Mohammed häufigster Vorname bei Neugeborenen

Schlagzeilen mit diesem Tenor liest man immer wieder einmal.[60] Was eine solche Überschrift unterschweillig suggeriert, ist klar, aber was können Sie daraus *tatsächlich* schließen? Offenkundig gar nichts, außer dass der Vorname Mohammed bei muslimischen Eltern außerordentlich beliebt ist,

während es bei nicht muslimischen Eltern keinen „einsamen Spitzenreiter“ in der Namensgebung gibt. □

Ein weiterer Effekt tritt immer dann auf, wenn einige der im Ranking zu vergleichenden Entitäten stark zufällig streuen. Ein konkretes Beispiel sind Großunternehmen mit vielen Filialen von unterschiedlicher Größe: Die Filiale mit dem höchsten oder niedrigstem Umsatz pro Kunde oder die Filiale mit der höchsten oder niedrigsten Diebstahlquote dürfte überraschend häufig eine sehr kleine Filiale sein – nicht etwa, weil an diesen Filialen irgendetwas Besonderes wäre, sondern weil kleine Filialen aus rein statistischen Gründen stärker zufällig streuen als große. Subtile Variationen dieses Phänomens wie im folgenden Beispiel beeinflussen offenbar die Sicht vieler Menschen auf die Welt.

Fallbeispiel 66: Spielen Frauen schlechter Schach als Männer?

Die Spielstärke eines Schachspielers wird durch seine *Elo-Zahl* ausgedrückt, die seine Erfolge in Turnieren widerspiegelt. Daraus ergibt sich ein Ranking aller Turnierschachspieler dieser Welt: je höher die Elo-Zahl, umso besser der Spieler. Die erste Frau landet zurzeit auf Platz 66, und generell sind nur wenige Frauen in den oberen Rängen zu finden. Heißt das also, Frauen spielen schlechter Schach als Männer?

Hier tritt ein häufiges statistisches Phänomen zutage: Da weitaus mehr Männer als Frauen Schach spielen, ist zu erwarten, dass unter den Extremen – auf beiden Seiten der Elo-Skala – viel mehr Männer als Frauen zu finden sind. Man kann berechnen, wie groß dieser Effekt ungefähr wäre, wenn die Frauen, die Turnierschach spielen, exakt genauso gut

wären wie die Männer, die Turnierschach spielen. Die Verteilung von Frauen und Männern im realen Elo-Ranking ist gar nicht so weit von dieser fiktiven Größe entfernt, das heißt, die Hypothese, dass Frauen schlechter als Männer Schach spielen, lässt sich mit dem Elo-Ranking *nicht* begründen.[61]

Warum so viel weniger Frauen als Männer Turnierschach spielen, muss mangels belastbarer Daten offen bleiben. □

Viele Rankings basieren nicht nur auf einem einzelnen Kriterium, sondern auf einer Auswahl von mehreren Kriterien, die mit unterschiedlichen Gewichtungen in die Gesamtbewertung eingehen. Zwangsläufig – man kann es gar nicht verhindern – ergibt sich ein großer Spielraum, den man irgendwie ausgestalten *muss*. Je nach Ausgestaltung kommen durchaus unterschiedliche Ergebnisse heraus. Damit will ich niemandem Absicht unterstellen, aber dieser Spielraum muss ja nun einmal irgendwie ausgenutzt werden, die Berechnungsformel für das Ranking muss irgendwie im Detail festgelegt werden. Unvermeidlich kommt damit ein Element der Willkür hinein.

Das Problem fängt schon bei der exakten Definition der einzelnen Kriterien an, nach denen bewertet wird. Zum Beispiel ist nichts dagegen zu sagen, wenn etwa gute Handhabbarkeit und Umweltverträglichkeit Kriterien für Produkte sind. Aber dafür, wie man solche Kriterien nun operationalisiert, gibt es meist keine vorgegebene Standardisierung. Entsprechend häufig gibt es sachbezogene Kritik von Anbietern, die im Ranking schlecht weggekommen sind. Das gilt sinngemäß natürlich für alle Arten von Rankings, nicht nur für Produktrankings.

Wenn mehrere Kriterien zusammen ein Rankingergebnis ergeben sollen, dann stellt sich unvermeidlich die Frage, wie

die einzelnen Kriterien relativ zueinander gewichtet werden sollen. Die gewählte Gewichtung muss nicht Ihren Präferenzen entsprechen!

Fallbeispiel 67: Service sehr gut – aber kaum erreichbar
Die Hotline eines Mobilfunkunternehmens erhielt in einem Test die Note *sehr gut* für Freundlichkeit und Kompetenz, aber ein *mangelhaft* für Erreichbarkeit. Gesamtergebnis war daher *befriedigend*.[62] Ich vermute, auch noch so guten Service wird nicht jeder Leser als befriedigend empfinden, wenn er den Service gar nicht erst erreicht; so mancher Leser dürfte für sich die Frage, ob er überhaupt durchkommt, sicherlich sehr viel höher gewichten und daher ein Gesamturteil „befriedigend" vielleicht falsch interpretieren. □

Dass diese Gewichtungsfaktoren zumindest ein Stück weit willkürlich sind, sieht man schon daran, dass sie häufig Vielfache von 5 % oder gar von 10 % sind, was sich sicher nicht durch Sachargumente rechtfertigen lässt. Sie meinen, ob 5 % oder 4 % oder 6 % – das spielt keine Rolle? Dann lesen Sie:

Fallbeispiel 68: Testergebnisse im Test
Anhand konkreter Beispiele von Stiftung Warentest und diversen anderen Medien konnten wir aufzeigen, dass schon kleine Änderungen an den Gewichtungsfaktoren teilweise zu großen Änderungen im Ergebnis führen: Einzelne Produkte machen große Sprünge nach oben oder nach unten, und in manchen Rankings bewegt sich jedes einzelne Produkt im Durchschnitt um einen Platz.[63] □

Häufig werden Grenzwerte gesetzt. Zum Beispiel werden Regeln angewendet von der Art: Wenn ein Produkt nicht umweltverträglich ist, dann kann es keine gute oder sehr

gute Note mehr bekommen. Die Grenzwerte in offiziellen DIN-Normen und ähnlichem sind zuweilen zu konservativ definiert, um aussagekräftig zu sein, und in vielen Fällen müssen sie ja sogar von allen Produkten gesetzlich erfüllt sein. Also muss ein eigener Grenzwert definiert werden.

Fallbeispiel 69: Arsen im Mineralwasser
Im Mineralwassertest von Ökotest haben zwei Produkte die *Hälfte* der gesetzlichen Höchstmenge bei Arsen überschritten und konnten deswegen keine bessere Note als „befriedigend" erhalten.[64] Man hätte die Grenze auch auf den gesetzlichen Höchstwert oder auf ein Zehntel oder Hundertstel davon setzen können, und der Gesamttest wäre jeweils entsprechend anders ausgefallen. □

Natürlich gibt es immer auch eine sachbezogene Begründung, warum der Grenzwert gerade so und nicht anders festgelegt wurde. Aber überzeugende sachbezogene Begründungen finden sich auch für andere mögliche Grenzwerte.

2.11 Umfragen

Das Grundproblem bei Umfragen ist erst einmal, wie repräsentativ die Umfrageteilnehmer eigentlich für die jeweilige Grundgesamtheit sind. Selbst wenn die Umfrageteilnehmer rein zufällig ausgewählt werden, ist durchaus zu hinterfragen, ob die Auswahl der Befragten wirklich immer repräsentativ ist.[65] Ein Beispiel: Noch vor wenigen Jahren wurden telefonische Umfragen überwiegend durch Anrufe auf Festnetztelefone durchgeführt. Aber schon damals waren

sehr viele Leute eigentlich nur noch per Handy erreichbar. Und offensichtlich sind diejenigen, die gut per Festnetz erreichbar sind, nicht unbedingt repräsentativ für diejenigen, die eher nur per Handy erreichbar sind. So ergab eine Studie von 2007, dass sich schon beim damaligen geringen Anteil an reinen Handynutzern, die keinen Festnetzanschluss mehr haben, Unterschiede zeigen, die in knappen Fällen durchaus zu unterschiedlichen Ergebnissen führen können.[66]

Immer wieder werden Umfragen als „repräsentativ" bezeichnet, nur weil die Befragten repräsentativ *ausgewählt* wurden. Aber: Die Rücklaufquote ist im Allgemeinen sehr gering. Laut Wikipedia gilt eine Rücklaufquote von mehr als 15 % zumindest bei schriftlichen Befragungen schon als „bemerkenswert hoch".[67] Das heißt, selbst wenn die Auswahl der Befragten repräsentativ ist – was, wie wir schon gesehen haben, durchaus anzweifelbar ist –, gibt es erst einmal keinen Grund zur Annahme, dass die kleine Minderheit, die geantwortet hat, tatsächlich repräsentativ ist für die große Mehrheit, die *nicht* geantwortet hat.

Häufig werden die Rohdaten deswegen so gewichtet, dass zumindest bei einigen grundlegenden demographischen Faktoren wie Alter und Geschlecht die Diskrepanz zwischen der Gesamtbevölkerung und den Antwortenden bereinigt wird. Aber auch dann gibt es per se erst einmal keinen Grund zur Annahme, dass die Antwortenden in *allen* relevanten Faktoren ausreichend repräsentativ für die Grundgesamtheit sind, damit das Umfrageergebnis wenigstens näherungsweise vertrauenswürdig ist.

Wenn die Teilnehmerzahl zu einer Umfrage veröffentlicht wird, wird sehr häufig nicht dazu gesagt, ob diejenigen gezählt wurden, die angefragt wurden, oder diejenigen, die

tatsächlich geantwortet haben. Im ersteren Fall wäre die Gesamtzahl der Antworten entsprechend nur ein Bruchteil der publizierten Teilnehmerzahl.

Eine subtile Problematik ergibt sich, wenn die Anzahl der Befragten zwar ausreichend groß für summarische Schlussfolgerungen ist, dann aber auch über Ergebnisse bei einzelnen Teilgruppen berichtet wird, die nicht mehr groß genug sind. Christensen und Christensen formulieren das so: „…dass sich im Bereich der Marktforschung hartnäckig eine zauberhafte Zahl hält, die sich auf die Anzahl der Befragten bezieht: Werden (gut) 1000 Menschen befragt und dabei mindestens die Verteilung von Geschlecht, Alter und Region in der Stichprobe in Bezug auf die Grundgesamtheit kontrolliert, wird eine Befragung als ‚repräsentativ' bezeichnet … Wenn nun aber regional differenzierte Analysen durchgeführt werden, ist schnell ersichtlich, dass die Fallzahlen pro Region extrem klein werden."[68]

Der Fachbegriff für das Problem, dass die antwortende Minderheit nicht repräsentativ für die anderen ist, lautet *Schweigeverzerrung*. Oft kann man wenigstens plausibel, wenn auch nicht belastbar schlussfolgern, in welche Richtung die Schweigeverzerrung mutmaßlich gehen dürfte. Dazu lege ich Ihnen im Folgenden zwei Beispiele vor.

Fallbeispiel 70: Umfragen unter Absolventen Jahre nach Ende des Studiums

Man muss vermuten, dass eher die Absolventen antworten werden, die es zu etwas gebracht haben; erstens, weil erfolgreiche Menschen leichter auch nach Jahren auffindbar sein dürften; zweitens werden Menschen, die Erfolge vorzuweisen haben, wahrscheinlich bereitwilliger antworten. Zum

Beispiel bei der gerne gestellten und für Studieninteressierte überaus interessanten Frage nach dem Einkommen wird sich daraus wohl eine Schweigeverzerrung nach oben ergeben.[69] □

Fallbeispiel 71: Scharia
Umfragen unter Muslimen in Westeuropa bringen ans Licht, dass mehr als die Hälfte dieser Gruppe religiöse über weltliche Gesetze stellt.[70] Laut Originalpublikation[71] wurden die Rohdaten der Umfrage durch das Institut um Geschlecht, Alter, erste/zweite/dritte Generation und Aspekte der Anrufstrategie bereinigt. Man darf vermuten, dass diejenigen Muslime, die sich von der Mehrheitsgesellschaft in Westeuropa eher *abwenden*, auch unter dieser Korrektur nicht überrepräsentiert sind unter denen, die bereit waren, an der Umfrage teilzunehmen. Das heißt, der Anteil derjenigen unter den westeuropäischen Muslimen, die religiöse über weltliche Gesetze stellen, dürfte zumindest nicht kleiner sein als der, den das Studienergebnis behauptet.

Dazu muss ich allerdings eine wichtige Anmerkung machen: Bevor man der Mehrheit der Muslime in Westeuropa aufgrund des Umfrageergebnisses tatsächlich Verfassungsfeindlichkeit unterstellt, müsste erst geklärt werden, wie genau die Frage auf dem Fragebogen – in den verschiedenen Sprachen, in die er übersetzt wurde – formuliert war und wie sie in ihren Nuancen von den Adressaten mutmaßlich verstanden worden ist. Dass man die religiösen über die weltlichen Gesetze stellt, *kann* eine politische Aussage sein, dann ist sie sicherlich verfassungsfeindlich. Sie kann aber auch einfach als „Herzensaussage" gemeint sein von Mitbürgern, die eben mit dem Herzen an ihrem Glauben hängen, die

Gesetze ihres westeuropäischen Heimatlandes aber dennoch voll und ganz respektieren. So schreibt der Jurist und Islamwissenschaftler Mathias Rohe: „Nach einer 1000 Jahre alten, verbreitet angenommenen Lehre müssen Muslime, die sicher in nicht-islamischen Ländern leben, *auch aus religiösen Gründen* die dort geltenden Gesetze achten" (Hervorhebung von mir).[72] □

Übrigens: Haben Sie es bemerkt? Bei der Überschrift des letzten Fallbeispiels habe ich genau das gemacht, was ich in Abschn. 1.4 in Bezug auf Medien für die breite Öffentlichkeit diskutiert hatte: irreführende Teaser. Denn das Wort „Scharia" kam weder in der zitierten Pressemitteilung noch im Originalartikel überhaupt vor.

So manche Umfrage ist von vornherein wertlos, weil die Frage oder die Antwortoptionen missverständlich formuliert sind, was bei den meisten Themen auch bei größter Sorgfalt kaum vermeidbar, manchmal vielleicht sogar gewollt ist.[73]

Fallbeispiel 72: Sind Sie gewaltbereit?
Wenn Sie mit „Ja" antworten auf die Frage: „Ich würde selbst nie körperliche Gewalt anwenden. Aber ich finde es gut, wenn es Leute gibt, die auf diese Weise für Ordnung sorgen", dann sind Sie das nach verbreiteter medialer Lesart, auch wenn Sie vielleicht – naheliegenderweise – an Polizei und nicht an Lynchjustiz dachten.[74] □

Fallbeispiel 73: Stuttgart 21
Im Vorfeld der Volksabstimmung zu *Stuttgart 21* gab es Befürchtungen, dass die – aus juristischen Gründen wohl nicht anders formulierbare – Fragestellung auf dem

Abstimmungszettel irreführend sein könnte: Wer *für* Stuttgart 21 war, musste mit „Nein" stimmen, wer *dagegen* war, mit „Ja".[75] Leider ist mir keine Untersuchung darüber bekannt, ob dieser Effekt sich letztlich auf das Abstimmungsergebnis ausgewirkt hat oder nicht. □

Fallbeispiel 74: Können Sie auch manchmal barsch sein? Die *Big Five* Persönlichkeitsmerkmale sind Offenheit, Gewissenhaftigkeit, Extraversion, Verträglichkeit und Neurotizismus.[76] In einem Fragebogen zur Einschätzung dieser Merkmale fand ich die Aussage: „Ich kann manchmal auch barsch sein." Dazu war „Ja" oder „Nein" anzukreuzen, man sollte also angeben, ob man „manchmal barsch sein kann" oder nicht.

Was heißt das denn, wenn Sie dieser Aussage zustimmen? Heißt es, Sie haben sich manchmal nicht im Griff und werden barsch, obwohl Sie das eigentlich gar nicht wollen? Oder heißt es, Barschheit gehört zu Ihrem Verhaltensrepertoire, das Sie gezielt und rational einsetzen können, um Ihre Ziele zu erreichen? Beide Interpretationen sind möglich, zeigen aber völlig unterschiedliche Charaktermerkmale an. □

Fallbeispiel 75: Finden auch Sie Vergewaltigung ok? Während ich den ersten Entwurf dieses Buches Korrektur lese, geht ein erschreckendes Umfrageergebnis durch die Medien: Jeder vierte Deutsche findet Vergewaltigung unter gewissen Umständen ok!

Was war die Frage, die von einem Viertel der Deutschen mit „Ja" beantwortet wurde: „Es gibt Personen, die finden, dass Geschlechtsverkehr ohne Einwilligung unter bestimmten Umständen gerechtfertigt ist. Glauben Sie, dass dies auf

folgende Situationen zutrifft?“[77] Die Autoren beider Quellen, die ich in Anmerkung 77 dazu zitiere, scheinen ein „Ja“ auf diese Frage tatsächlich als Beleg zu sehen, dass der Befragte Vergewaltigung unter gewissen Umständen ok findet. Sehen Sie das auch so?

Also, ich persönlich finde, etwas selbst ok zu finden ist etwas völlig anderes als eine Einschätzung darüber abzugeben, was irgendwelche „Personen“ ok finden. Wir wissen nicht, ob die Antwortenden die Frage so wie die Autoren der zitierten Artikel aufgefasst oder die wahre Sprachlogik dieser arg verklausulierten Frage verstanden und dementsprechend geantwortet haben. Daher sehe ich für die Interpretation, dass so viele Leute Vergewaltigung ok finden, keine Basis.

Nebenbei bemerkt, wäre auch hier wieder die Rücklaufquote sehr aufschlussreich. Ich zum Beispiel könnte mir nicht vorstellen, überhaupt bei einer Umfrage mit solchen Fragen mitzumachen. Spätestens bei der hier diskutierten Frage würde ich aus der Befragung aussteigen. Vielleicht bin ich damit ja nicht allein. □

Wann immer die Umfrageteilnehmer Auskunft über sich selbst geben sollen, sind die Antworten subjektiv und daher potentiell systematisch verzerrt. Wenn die Fragen eigentlich auf unparteiische Antworten abzielen, ist das natürlich ein Problem. Diese systematische Verzerrung wird besonders prägnant durch den *Lake-Wobegon-Effekt* demonstriert, der besagt, dass die meisten Menschen sich für überdurchschnittlich befähigt halten, zum Beispiel als Autofahrer oder Liebhaber.[78]

Zuweilen sind die Fragen so gestellt, dass man leicht raten kann, mit welcher Antwort man ein positives Bild von

sich selbst vermittelt, und mit welcher Antwort ein eher negatives Bild. Eine Verzerrung hin zur ersten Antwortoption ist nur menschlich. Selbst bei bestem Willen machen Menschen systematisch zu hohe Angaben beispielsweise bezüglich ihrer Arbeitslast[79] oder auch zu niedrige Angaben, wie viel sie so essen, und wie viel davon ungesund ist (den süßen Snack zwischendurch vergisst man leicht).

Ob Interviewer bei persönlichen Interviews unter vier Augen – am Telefon oder vor Ort – immer so ganz unparteiisch sind und weder Einfluss auf die Befragten ausüben noch die Angaben der Befragten verfälscht eintragen, muss hier dahingestellt bleiben.

Bei einem Fragebogen zum Ankreuzen können die Antworten natürlich nur im Rahmen der vorgegebenen Antwortoptionen sein. Das kann ein falsches Bild liefern, beispielsweise wenn man „Nein" ankreuzen möchte, aber nur verschiedene Varianten von „Ja" zur Auswahl hat.

> Glauben Sie keinem Umfrageergebnis, bei dem Sie die exakte Formulierung der Frage und der Antwortoptionen nicht kennen!

Oft hat man Zahlenwerte anzukreuzen, zum Beispiel von 1 bis 10 oder von −2 bis +2 mit der Bedeutung: −2 ist sehr schlecht, −1 etwas schlecht, 0 neutral, +1 ist etwas gut und +2 sehr gut. Daraus dann einen Durchschnittswert zu bilden, wie man es oft sieht, ist ein elementarer handwerklicher Fehler. Denn diese Zahlen stehen ja nur symbolisch für rein qualitative, nicht numerische Antwortoptionen. Fachlich gesprochen: Die Antwortoptionen bilden nur

eine *Ordinalskala*, keine *Kardinalskala*.[80] Man kann es auch so formulieren: Nur wenn alle Befragten exakt dieselbe intuitive Vorstellung davon hätten, wie viel mehr ein sehr gut beziehungsweise sehr schlecht gegenüber einem etwas gut beziehungsweise etwas schlecht wiegt, dürfte man mit diesen relativen Gewichtungen einen gewichteten Durchschnitt bilden. Davon kann man natürlich in der Regel nicht ausgehen.

Kommt Ihnen dieser Fehler nicht irgendwie aus Schule, Ausbildung oder Studium bekannt vor? Darum geht es im nächsten Beispiel.

Fallbeispiel 76: Notendurchschnitte

Die Berechnung von Notendurchschnitten bei Prüfungen ist ebenfalls eine unzulässige Operation, denn auch Noten sind eigentlich keine numerischen Werte, wie man etwa an der amerikanischen Skala von A bis F sofort sieht. Um die Problematik von Notendurchschnitten noch einmal zu verdeutlichen, nehmen Sie an, dass in einer schriftlichen Prüfung insgesamt 100 Punkte erreicht werden können, ein „sehr gut" ab 95 Punkten, ein „gut" ab 85, ein „befriedigend" ab 70 und ein „ausreichend" ab 50 Punkten. Wenn man nun wie üblich den Notendurchschnitt bildet, dann gewichtet man die Intervalle für die einzelnen Noten gleich, obwohl sie unterschiedlich groß sind. Natürlich hindert einen niemand daran, trotzdem den Notendurchschnitt zu bilden, man muss sich aber bewusst sein, dass beispielsweise der Vergleich der Notendurchschnitte zweier Prüfungen nur bedingt aussagefähig ist.

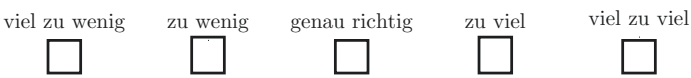
Er kann sogar irreführend sein. Betrachten Sie dazu folgendes einfaches Zahlenbeispiel: Zwei Kandidaten, X und Y, nehmen an zwei Prüfungen, A und B, teil, und neben X

und Y gibt es keine weiteren Teilnehmer. Kandidat X erreicht 95 Punkte in A und 98 Punkte in B, Kandidat Y schafft 85 Punkte in A und 84 Punkte in B. Dann hat Prüfung A einen Notendurchschnitt von 1,5 und B nur einen von 2,0, obwohl in A durchschnittlich 90 Punkte, in B aber 91 Punkte erreicht wurden. □

Auf vielen Fragebögen haben Sie eines von mehreren nebeneinander stehenden Kästchen anzukreuzen, um eine Frage zu beantworten. Die Antwortoptionen können von links nach rechts immer negativer werden, und die uneingeschränkt positive Antwort ist dann ganz links:

Es kann natürlich auch genau umgekehrt sein:

Die uneingeschränkt positive Antwort kann aber auch in der Mitte sein, im Sinne von „gerade richtig“. Dann besagen die Antwortoptionen links vielleicht „zu wenig“, „zu langsam“ oder ähnlich, und die Antwortoptionen rechts besagen dann „zu viel“ oder „zu schnell“:



Das Problem ist: Wenn das Schema von einer Frage zur nächsten wechselt, passiert es nicht wenigen Menschen, dass sie die zweite Frage blindlings nach dem Schema der ersten beantworten und damit ihre eigene Antwort völlig verfälschen.

Im Rest dieses Abschnitts betrachten wir speziell Wahlumfragen. Wenn Wahlprognosen stark danebenliegen, dann wird das in den Publikumsmedien hin und wieder durchaus thematisiert. Nicht ganz selten haben sich *alle* Umfrageinstitute mehr oder weniger stark vertan. Denken Sie etwa an die Prognosen zum „Brexit", dem britischen Referendum 2016 zum Austritt aus der EU, oder an die Wahl des US-Präsidenten 2016.[81] Dieselben Medien präsentieren nicht selten ein paar Wochen später wieder völlig kommentarlos die Ergebnisse der letzten Sonntagsfrage – also was wäre, wenn nächsten Sonntag gewählt würde – und analysieren die politischen Implikationen ganz so, als wären diese Zahlen die unzweifelhafte Realität.

Die ersten Hochrechnungen, die unmittelbar nach Schließung der Wahllokale oder sogar schon vorher zirkulieren, sind immer schon sehr nah am amtlichen Endergebnis. Die Situation am Wahltag ist allerdings sehr speziell, und die Meinungsforschungsinstitute investieren am Wahltag sehr viel mehr Mühe als für die monatliche Sonntagsfrage und andere Umfragen. Es werden viel mehr Leute interviewt. Zudem werden die Leute unmittelbar nach dem Verlassen des Wahllokals interviewt. Die Wahrscheinlichkeit, dass die Befragten etwas Falsches sagen, ist sehr viel geringer als bei Telefoninterviews. Die Ergebnisse von Sonntagsfragen werden außerdem dadurch verwässert, dass viele noch kurz vor der Wahl unentschlossen sind, ob und wen sie wählen sollen. Dieser Faktor

fällt bei der Befragung am Wahltag vor den Wahllokalen ebenfalls weg.

Die Umfrageergebnisse werden übrigens gar nicht unbedingt eins-zu-eins veröffentlicht, sondern die Teilergebnisse für die einzelnen demographischen Gruppen werden mit irgendwelchen Gewichtungen versehen. In diese Gewichtungen fließen die Erfahrungen aus früheren Jahren und Jahrzehnten ein, wie Umfrageergebnisse sich von Wahlergebnissen unterscheiden, zum Beispiel weil einige Befragte im persönlichen Gespräch – selbst am Telefon – eher die mutmaßlich sozial erwünschte Meinung als ihre eigene Meinung angeben. Das nennt man den *Bradley-Effekt*.

Die politische Landschaft hat sich allerdings massiv geändert und ändert sich weiter, denken Sie etwa an den gestiegenen Anteil der Wechselwähler und der Nichtwähler. Daher sind Fortschreibungen früherer Gegebenheiten durchaus fragwürdig.

2.12 Prognosen und Simulationen

Wissenschaftsbasierte Prognosen sind das Ergebnis einer speziellen Art von Simulation: Man simuliert den Fall, dass die aktuelle Situation und die aktuellen Entwicklungen in die Zukunft fortgeschrieben werden können beziehungsweise dass gewisse Annahmen über die Zukunft zutreffen. Daher behandle ich beides – Simulationen und Prognosen – in diesem Abschnitt zusammen.

Es gibt mehrere grundlegende Probleme mit Simulationen und Prognosen: Known Unknowns, Unknown Unknowns und der berühmte schwarze Schwan. Hinzu kommen die

Begrenztheit der Computerhardware und die Gefahr der Überanpassung.

Fallbeispiel 77: Klimasimulationen und -prognosen
Alle paar Jahre stellt man wieder einmal fest, dass die Klimamodelle die Wirklichkeit nicht korrekt beschreiben, sondern nachjustiert werden müssen.[82] Das liegt einerseits an Known Unknowns: Einflussfaktoren, die zwar bekannt sind, von denen man bislang aber nicht weiß, ob und wie stark sie sich tatsächlich auf das Klima auswirken. Dazu gehört beispielsweise die Sonnenfleckenaktivität, denn sie lässt sich nicht präzise vorhersagen, und ihr Einfluss auf das Klima lässt sich auch nur grob abschätzen.[83]

Die Unknown Unknowns hingegen sind Einflussfaktoren, die die Klimaforscher überhaupt nicht auf dem Radar haben. Die Bindungsfähigkeit der Weltmeere für CO_2 gehörte lange Zeit dazu.

Bei praktisch allen naturwissenschaftlichen Simulationen – so auch hier – ist die Begrenztheit der Computerhardware ein massives Problem: Computer sind zwar schnell, aber nie schnell genug, so dass man Raum und Zeit gröber modellieren muss als eigentlich sinnvoll wäre.[84] Und für ganz korrekte Rechnungen müssten Zahlen eigentlich mit unendlicher Genauigkeit gespeichert und verarbeitet werden, was natürlich nicht geht. Leider gilt: kleine Ursache – große Wirkung, das heißt, kleine Ungenauigkeiten, die durch ein zu grobes Raum-Zeit-Raster und endliche Zahldarstellung unvermeidlich sind, schaukeln sich in den numerischen Berechnungen immer weiter auf. Der Fehler wächst über die aberbillionen Rechenschritte hinweg so stark, dass das Rechenergebnis mit der Realität nichts mehr zu tun hat. Dieses

Problem wird man naturgemäß nie hundertprozentig in den Griff bekommen.

Das zweite methodische Problem, Überanpassung, ergibt sich aus dem Bestreben, die kleinen Freiheitsgrade im Modell und in der Berechnungsweise so festzulegen, dass die Rechenergebnisse für reale Klimadaten aus der Vergangenheit stimmen, denn das ist die einzig mögliche Nagelprobe. Leider sind die Klimadaten aus der Vergangenheit nur bis zu einem gewissen Grad repräsentativ für die Zukunft, so dass diese Freiheitsgrade für korrekte Zukunftsprognosen vielleicht doch besser anders festgelegt werden sollten. Aber man weiß nicht, wie. Dummerweise gilt auch hier wieder: kleine Ursache – große Wirkung. □

Vier Probleme aus der Aufzählung zu Beginn des Abschnitts haben wir im letzten Fallbeispiel gesehen: Known Unknowns und Unknown Unknowns, Begrenztheit der Computerhardware und Gefahr der Überanpassung. Das fünfte Problem wurde noch nicht thematisiert: der *schwarze Schwan*, der durch Nassim Nicholas Talebs gleichnamiges Buch populär wurde.[85] So nennt man ein unvorhergesehenes Ereignis beziehungsweise eine unvorhergesehene Erkenntnis, die so einige bisherige Vorstellungen, Erkenntnisse oder Pläne völlig umwirft. Namensgeber ist die unvorhergesehene Erkenntnis bei der Entdeckung Australiens, dass Schwäne nicht grundsätzlich weiß sind; manche Arten in bis dahin unentdeckten Welten sind schwarz.

Fallbeispiel 78: Wenn es plötzlich talwärts geht

Massive Einbrüche bei Börsenkursen und andere Talfahrten werden regelmäßig nicht vorhergesehen, fließen daher

nicht in ökonomische Prognosen ein und durchkreuzen jede Planung, die auf diesen Prognosen basiert.[86] □

Fallbeispiel 79: Projektplanung – und täglich grüßt der schwarze Schwan

Bekanntlich geht bei Projekten immer einiges schief; der schwarze Schwan ist praktisch alltäglich. Leider führen schwarze Schwäne höchst selten zu einer Verkürzung der Projektdauer oder zu einer Verringerung der Ausgaben; in der Regel dauert das Projekt länger und kostet mehr. Wenn es dumm läuft, gilt auch hier wieder: kleine Ursache – große Wirkung.

Man kann nur versuchen, aus der Vergangenheit für die Zukunft zu lernen. Das läuft darauf hinaus, dass die bisherigen schwarzen Schwäne in Form von Wahrscheinlichkeitsverteilungen für die Dauer der einzelnen Arbeitsschritte des Projektes quantifiziert werden. Beispielsweise heißt das für einen Arbeitsschritt, der ohne unvorhergesehene Verzögerungen vielleicht zehn Tage dauern und zehntausend Euro kosten würde: Man versucht auf Basis bisheriger Erfahrungen einzugrenzen, wie wahrscheinlich der unverzögerte Ablauf ist, wie wahrscheinlich eine Verzögerung um einen, zwei, drei Tage und so weiter wäre, und wie teuer das jeweils käme. Mit solchen Quantifizierungen kann man dann Simulationen durchführen und erhält belastbare Schätzwerte dafür, in welchem Rahmen Projektdauer und -kosten etwa mit 80 %, 90 % oder 95 % Wahrscheinlichkeit bleiben werden.

Wenn die einzelnen Arbeitsschritte im Projekt Routine sind – wie etwa in Standardbauvorhaben –, dann ist das durchaus ein gangbarer Weg. Übrig bleiben als Unsicherheitsfaktoren dann die ganz großen schwarzen Schwäne wie etwa

Streiks, die man nicht sinnvoll quantifizieren und daher auch nicht in die Simulation einbeziehen kann. In größeren, ambitionierten Softwareentwicklungsprojekten hingegen wird man bei etlichen einzelnen Arbeitspaketen von Routine überhaupt nicht reden können, hier kann man sich daher auch weniger auf Schätzwerte verlassen. □

Fallbeispiel 80: Demographische Prognosen
Ein paar Known Unknowns machen jede Prognose von vornherein problematisch: die zukünftige Entwicklung von Geburtenrate, statistischer Lebenserwartung, Einwanderung, Auswanderung und so weiter. Diese werden potentiell beeinflusst durch Unknown Unknowns, etwa erfolgreiche geburtenfördernde Maßnahmen einer zukünftigen Bundesregierung oder unvorhergesehene dramatische Entwicklungen in der einen oder anderen bevölkerungsreichen Weltregion, die zu einer erhöhten Migration nach Europa führen.

Seriöse Arbeiten legen deshalb auch nicht nur eine einzige Modellrechnung vor, sondern sehr viele, bei denen verschiedene Werte für die einzelnen Known Unknowns eingesetzt werden. Naturgemäß gehen die Prognosen aus den einzelnen Modellrechnungen so weit auseinander, dass man nur noch wenig daraus ableiten kann. Leider findet oft nur eine einzige Prognose den Weg in die Medien und das öffentliche Bewusstsein, und aufgrund der Sachzwänge im medialen Bereich ist das in der Regel eine extreme und ziemlich alarmierende. □

Fallbeispiel 81: Marktprognosen
Prognosen, wie sich die Märkte für bestimmte Produkte oder Produktarten in Zukunft entwickeln werden, sind noch

anfälliger als etwa Demographie für Unknown Unknowns und schwarze Schwäne, denn im Gegensatz zu Bevölkerungszahlen können Verkaufszahlen praktisch von einem Tag auf den anderen massiv auf neue Sachverhalte reagieren und daher beliebig stark und auf unvorhersehbare Weise von der Prognose abweichen.[87] □

Was sagt uns das alles nun? Sollen wir Prognosen und Simulationen also schlicht und einfach vergessen?

Jörg Hinze vom Hamburgischen Welt-Wirtschafts-Archiv schreibt speziell mit Blick auf die Konjunkturprognosen 2001 bis 2004:[88] „Bei rein quantitativer Gegenüberstellung der Prognosen wichtiger Größen wie dem Wirtschaftswachstum mit den realisierten Werten trifft der Eindruck zweifelsohne zu, daß die Jahre 2001 bis 2003 Fehlprognosen waren, 2004 hingegen ein ‚Volltreffer' …Wissenschaftliche Prognosen sind bedingte Wahrscheinlichkeitsaussagen, das heißt, sie basieren auf einer Reihe von Annahmen und Setzungen. Ändern sich diese während des Prognosezeitraums, muß sich das *konsequenterweise* [Hervorhebung im Original] in Änderungen gegenüber den ursprünglichen Prognosewerten und – soweit die Prognose korrekt abgeleitet war – in entsprechenden Abweichungen zu den realisierten Werten niederschlagen." Sein Resümee: „Prognosen können zwar die Unsicherheit über die Zukunft nicht beseitigen, sie können sie aber, insbesondere wenn die Rahmenbedingungen und Risikoabwägungen beachtet werden, reduzieren und dadurch helfen, rationale Entscheidungen zu treffen."

Meines Erachtens ist das konform mit der Haltung, die ich hier vermitteln möchte: Man darf sich nicht blindlings auf das Prognoseergebnis verlassen, sondern sollte genau hinschauen, auf welchen Annahmen es beruht. Nur wenn

die Unknown Unknowns und schwarzen Schwäne nicht dazwischenpfuschen, wird das Prognoseergebnis zutreffen. Und nur unter dieser Prämisse dürften seriöserweise Schlussfolgerungen aus der Prognose gezogen werden, also in der Form: Wenn *alle* Rahmenbedingungen so-und-so sind, *dann* stimmt die Prognose.

In der praktischen Umsetzung heißt das, wenn man auf Basis von Prognosen agieren will, dann muss man vorsichtig und Schritt für Schritt vorwärtsgehen, dabei immer schauen, ob die Rahmenbedingungen sich geändert haben, und falls ja, die nächsten Schritte entsprechend nachjustieren.

2.13 Die falsche Zahl

Täglich werden Sie mit Zahlen konfrontiert – in den Nachrichten, in der Werbung, beim Thema Gesundheit, beinahe überall. Aber die Zahlen, die Sie zu sehen bekommen, sind nicht unbedingt die eigentlich relevanten.

Fallbeispiel 82: Das kleine Wörtchen „ab“
Preisen in der Werbung ist häufig – deutlich kleiner gedruckt – das Wort „ab“ vorangestellt. Nur unter sehr eingegrenzten Umständen wird wirklich dieser Preis angeboten; der Preis, den *Sie* zu zahlen haben, dürfte in der Regel erheblich höher sein.

Der Punkt ist: Dort, wo der geringste „Ab-Preis“ angeboten wird, muss nicht der beste Preis *speziell für Sie* herausspringen. Es ist durchaus nicht unwahrscheinlich, dass Sie bei einem Anbieter mit einem eher hohen „Ab-Preis“ am besten wegkommen. Denn wenn ein Anbieter für eine bestimmte

Zielgruppe sehr gute Konditionen anbietet, um mit einem attraktiven „Ab-Preis“ werben zu können, dann muss das Geld ja irgendwo anders wieder hereinkommen. □

Zuweilen werden alle relevanten Informationen durchaus genannt, aber in eher unauffälliger, leicht zu übersehender Weise (das berühmte Kleingedruckte) oder erst sehr spät. Zu Letzterem gehören beispielsweise Abschlussgebühren bei Internetkäufen auf diversen Onlineportalen[89] oder versteckte Kosten beim Hausbau und -kauf[90] oder auch das folgende Beispiel.

Fallbeispiel 83: Wie teuer wird der Kredit wirklich?
Eine Beispielklasse für sich sind Zusatzpakete, die nicht zum angebotenen Paket gehören, deren Kosten also nicht im Angebot von vornherein einberechnet werden müssen, die dann aber doch verbindlich oder zumindest sehr empfehlenswert sind.

Die Kreditausfallversicherung beziehungsweise Restschuldversicherung, die Ihnen vom Bankberater mit eindringlichen Worten ans Herz gelegt wird, ist offiziell nicht unbedingt verbindlich, denn sonst müsste sie in den angebotenen effektiven Jahreszins einberechnet werden.[91] Aber es mag Ihnen so vorkommen, als wäre sie es.[92] Und vielleicht wird es auch schwieriger mit dem Kredit, wenn Sie das unverbindliche Angebot ablehnen.[93]

In Summe können sich Kosten ergeben, die einem deutlichen höheren Zinssatz entsprechen würden. Eigentlich müsste ein Interessent jetzt noch einmal einen Schritt zurückgehen und noch einmal alle Angebote vergleichen, diesmal

inklusive der Versicherung, denn das beste Angebot *inklusive* Versicherung muss ganz und gar nicht das Angebot sein, das zuvor, also *ohne* Versicherung am besten erschien. Aber wie viele machen das schon, und wie viele bleiben stattdessen bei diesem Anbieter, auf den Sie sich aufgrund des attraktiv erscheinenden effektiven Jahreszinses festgelegt hatten?[94] □

Die nächsten beiden Fallbeispiele zeigen, dass Sie auch bei jedem Bericht in Medien und anderswo immer noch einmal kurz darüber nachdenken sollten, ob die präsentierten Zahlen wirklich aussagekräftig sind oder sie nicht besser andere Zahlen in Erfahrung bringen sollten, um den Sachverhalt richtig einzuschätzen.

Fallbeispiel 84: Wie ändert sich die Beschäftigungssituation?
Wenn Sie sich ein Bild machen wollen, wie die Beschäftigungssituation in Deutschland sich im Laufe der Jahre und Jahrzehnte geändert hat, können Sie natürlich auf die Arbeitslosenraten schauen. Aber wesentlich aussagekräftiger dürfte das Gesamtvolumen sein, also wie viele Stunden alle Beschäftigten zusammen gearbeitet haben. In dieser Zahl sind dann auch Entwicklungen in Richtung mehr Teilzeitarbeit berücksichtigt. Nicht unwichtig dürfte auch ein bestimmter Teil davon sein: das Gesamtvolumen aller *sozialversicherungspflichtig* geleisteten Arbeitsstunden. □

Fallbeispiel 85: Ist es wirklich Ihre Mortalität, die Sie wissen wollen?
Grob gesprochen, ist die *Mortalität* einer potentiell tödlichen Krankheit oder anderen tödlichen Gefahr der Anteil aller

Menschen, die daran sterben, mit anderen Worten: die Wahrscheinlichkeit, dass Sie daran sterben, wenn Sie nicht zu einer speziellen Hoch- oder Niedrigrisikogruppe gehören. Die verschiedenen Risiken für Leib und Leben unterscheiden sich allerdings nicht nur in der Mortalität, sondern auch darin, wie früh im Leben sie statistisch daran sterben. Je früher der Tod eintritt, umso mehr Lebensjahre haben Sie verloren. Die Anzahl Lebensjahre, die Sie statistisch verlieren (*Reduktion der Lebenserwartung*), ist aber vielleicht die Größe, die Sie interessiert. Bei riskantem Verhalten im Straßenverkehr etwa kommt da schon ein recht hoher Wert heraus, auch wenn die Mortalität wesentlich geringer als beispielsweise die von Krebs ist. □

Wenn die einzelnen Fälle in einer Statistik einfach aufsummiert werden, dann bietet es sich an zu überlegen, ob die einzelnen Fälle nicht unterschiedliches Gewicht haben und ob die Summe dieser Gewichte beziehungsweise das Durchschnittsgewicht nicht vielleicht die wichtigere Zahl sein könnte.

Was heißt das? Nun, in Fallbeispiel 84 waren die Beschäftigten die einzelnen Fälle in diesem Sinne, und die Gewichtung eines Beschäftigten ist seine Anzahl Arbeitsstunden; in Fallbeispiel 85 waren die Todesopfer die einzelnen Fälle, und der individuelle Verlust an Lebensjahren ist die Gewichtung. In Fallbeispiel 84 waren wir somit an der Summe der Gewichte (also der Gesamtstundenzahl über alle Beschäftigten) interessiert, im Fallbeispiel 85 am Durchschnitt (also

der verminderten statistischen Lebenserwartung). In Fallbeispiel 84 wäre der Durchschnitt über alle Beschäftigten nicht aussagekräftig für die Gesamtbeschäftigungssituation gewesen, da neben der Stundenzahl pro Beschäftigtem auch die *Anzahl* der Beschäftigten wichtig ist, etwa wenn ein Vollzeitjob in mehrere geringfügige Beschäftigungsverhältnisse aufgeteilt wird: Dann nimmt die durchschnittliche Stundenzahl ab, aber die unverändert gebliebene Gesamtstundenzahl sagt offensichtlich mehr aus.

Fallbeispiel 86: Ist Sport wirklich gesund?
Diese Frage ist in gewisser Weise das Leitmotiv des Lexikons der Fitnessirrtümer: Man darf nicht nur die positiven Wirkungen des Sports sehen, sondern man muss auch negative Konsequenzen – wie beispielsweise Sportverletzungen und Verschleisserscheinungen – in die Rechnung einbeziehen. Die Gesamtbilanz ist meines Wissens unklar, zudem hat jeder Sportler seine eigene individuelle Gesamtbilanz. Aber so uneingeschränkt positiv wie oft suggeriert wird, muss sie nicht unbedingt sein.[95] □

> Nicht der Nutzen allein, sondern das *Verhältnis* von Kosten und Nutzen beziehungsweise Risiko und Nutzen ist entscheidend. Fragen Sie also *immer* nach den Kosten und Risiken, wenn nur der Nutzen thematisiert wird (und umgekehrt).

Häufig wird Ihnen der Erwartungswert als Ergebnis einer Studie präsentiert, aber das ist gar nicht unbedingt immer ein aussagekräftiger Indikator.

Fallbeispiel 87: Erwartungswert vs. Quantile
Der Erwartungswert ist *nicht* der Wert, den man erwarten kann!

Der Begriff „Erwartungswert" ist aus der Alltagssprache entlehnt, ist aber zunächst einmal nur eine bestimmte mathematischer Größe, die einen relativ schwachen Bezug dazu hat, wie man „Erwartung" im Alltag verstehen würde. Für unsere Zwecke genügt es, den Erwartungswert als durchschnittliches Ergebnis zu definieren, also beispielsweise als durchschnittlichen Profit, durchschnittliche Kosten oder durchschnittliches Risiko. Im Grunde ist der Erwartungswert nur dann aussagekäftig, wenn eine große Zahl ähnlich gelagerter Fälle auftritt. Ein konkretes Beispiel, in dem der Erwartungswert sinnvoll ist, ist die Schadensfallstatistik einer Versicherung: Stark vereinfacht gesprochen, basiert die Kalkulation der Prämien auf dem Erwartungswert, wie viel Schaden der einzelne Versicherungsnehmer im nächsten Jahr durchschnittlich produzieren wird. Siehe Abb. 2.2.

Häufig ist die Streuung so groß, dass der Erwartungswert nichts, aber auch gar nichts über den Einzelfall aussagt. Wesentlich aussagekräftiger sind die *Quantile*. Was heißt das? Allgemein gesprochen, ist das X-%-Quantil der Wert, der mit X% Wahrscheinlichkeit nicht überschritten wird. Wenn Sie eher vorsichtig sind, wird Sie das 80-%-Quantil für den Verlust interessieren, also welcher Verlustbetrag mit 80 % Wahrscheinlichkeit nicht überschritten wird; wenn Sie ein sehr hohes Sicherheitsbedürfnis haben, sogar eher das 90-%- oder 95-%-Quantil.

Der *Median* ist das 50-%-Quantil, also der Wert, der mit 50 % Wahrscheinlichkeit nicht überschritten wird. Der Mensch scheint so gestrickt zu sein, dass er sich unter dem

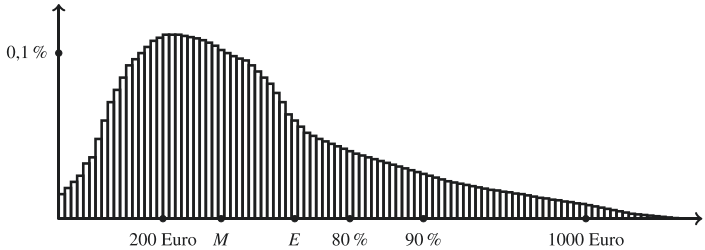


Abb. 2.2 Eine fiktive, aber nicht ganz untypische Verteilung in vielen Situationen, beispielsweise bei Schadensfällen einer Versicherung: Die Höhe eines Balkens gibt an, wie wahrscheinlich ein Versicherungsnehmer einen Schaden in der Höhe verursacht, wie auf der *X*-Achse angegeben ist. Der Erwartungswert *E* beträgt in diesem fiktiven Beispiel ca. 450 Euro, das ist die durchschnittliche Schadenshöhe eines Versicherungsnehmers, der einen Schaden verursacht. Der Median *M*, also das 50-%-Quantil, liegt bei 308 Euro: In der Hälfte aller Fälle ist der Schaden höchstens 308 Euro, in der anderen Hälfte ist er größer. Das 80-%- beziehungsweise 90-%-Quantil sind weiter rechts zu finden, weil 80 % beziehungsweise 90% aller Schadensfälle geringer sind als dieser Wert

Durchschnitts- oder Mittelwert eher den Median vorstellt. Bisher habe ich in Publikationen für das breite Publikum aber meist nur den Erwartungswert gefunden. Wichtig ist, dass der Erwartungswert und der Median beliebig weit auseinanderliegen können. Das ist bei sogenannten *schiefen Verteilungen* die Regel, zum Beispiel bei der Lebenserwartung: Weitaus mehr als 50 % aller Menschen sterben *nach* ihrer statistischen Lebenserwartung, weil der Median eben ein paar Jahre darüber liegt.[96]

Zumindest für eher konservative Anleger ist auch bei Geldanlagen der Erwartungswert für die Rendite sicher nicht der interessante Wert, auch nicht der Median, sondern wie oben eher das 80-%-, 90-%- oder 95-%-Quantil – je nachdem,

wie konservativ sie Ihre Ersparnisse anlegen möchten. Diese Werte sind in der Regel weitaus weniger attraktiv als der Erwartungswert. □

Ein immer wiederkehrender Fall ist die unüberlegte – manchmal vielleicht auch überlegte – Verwechslung von relativen und absoluten Größen. Häufig lesen Sie Schlagzeilen, die mit „immer mehr" oder „immer häufiger" beginnen. Selbst wenn das stimmt, was auch nicht immer der Fall ist: Machen Sie sich klar, dass „immer mehr" noch lange nicht „viel" und „immer häufiger" noch lange nicht „häufig" bedeuten muss.

Genauso lesen Sie immer wieder Horrormeldungen in den Medien: Wenn Sie nicht auf dieses oder jenes Lebensmittel weitgehend verzichten, dann steigt Ihr Risiko für eine bestimmte Krebsart oder ein anderes übles Gesundheitsrisiko um einen dreistelligen Prozentsatz, also um ein Mehrfaches. Ob ein solcher Befund – immer vorausgesetzt, dass er überhaupt stimmt – wirklich so dramatisch ist, lässt sich so pauschal gar nicht sagen. Sehr häufig geht es um hochgradig unwahrscheinliche Krankheiten. Dann ist diese Wahrscheinlichkeit auch mit einer solchen Steigerung immer noch recht gering, und es besteht kein Grund zur Panik.

Fallbeispiel 88: Was bringt ein neues Medikament an Gesamtnutzen?

Wenn beispielsweise ein Medikament bewirkt, dass die Anzahl der Menschen, die irgendwann an Krankheit X erkranken und auch sterben, von $0,04\,\%$ auf $0,03\,\%$ sinkt, dann sinkt die Mortalität dieser Krankheit um $25\,\%$. Das klingt sicher beeindruckend. Aber nur ein Mensch von zehntausend

profitiert von dem neuen Medikament, das klingt sicherlich deutlich weniger beeindruckend. □

Um drastisch falsche Einschätzungen der Wirksamkeit von Medikamenten, Vorsorgeuntersuchungen und Ähnlichem zu vermeiden, hat sich ein Indikator als besonders sinnvoll erwiesen, die Number Needed To Treat, kurz NNT, also die Anzahl an Patienten, die behandelt werden müssen, damit *ein einziger* Patient einen Nutzen hat. Ein vereinfachtes Beispiel dazu:[97] Nehmen wir eine fiktive Vorsorgeuntersuchung für ein Leiden an, an dem 2 % der Untersuchten tatsächlich unerkannt leiden. Nehmen wir weiter an, nur bei 30 % davon wird das Leiden durch die Untersuchung tatsächlich erkannt, und die Therapie wirkt nur bei 60 % der daraufhin behandelten Patienten. Dann haben nur $2\,\% \cdot 30\,\% \cdot 60\,\% = 0,36\,\%$ aller untersuchten Menschen am Ende einen Nutzen. Anders herum gesagt: Damit *ein einziger* Mensch einen Nutzen hat, müssen durchschnittlich 1/0, 36 % Menschen untersucht werden, das sind etwa 277.

> Wann immer Ihnen eine beeindruckende relative Steigerung oder Verminderung eines Indikators präsentiert wird, fragen Sie sich, ob der Absolutwert des Indikators den Eindruck nicht doch stark relativiert.

Wenn eine Steigerung oder Verminderung sich nicht auf einen Absolutwert, sondern auf einen (prozentualen) Anteil bezieht, wird die Situation noch einmal unübersichtlicher. Um Verwirrung zu vermeiden, hat sich der Begriff Prozent*punkt* etabliert – beziehungsweise leider nicht wirklich etabliert.

Fallbeispiel 89: Wie viel hat die Partei verloren?
Wenn etwa eine Partei bei der letzten Wahl 32 % der Stimmen errungen und bei dieser Wahl nur noch 24 % erreicht hat, dann hat sie *nicht* acht Prozent verloren, wie häufig in einem solchen Fall geschrieben wird. Sie hat acht Prozent*punkte* verloren, das sind in diesem Rechenbeispiel aber 25 % Verlust an Wählerstimmen! □

Zum Abschluss dieses Abschnitts noch ein Beispiel, bei dem die penible Unterscheidung zwischen *relativ* und *absolut* wohl so einige Menschen vor dem Bankrott bewahrt hätte.

Fallbeispiel 90: Nach zehnmal Kopf kommt ganz sicher Zahl
Viele Leute scheinen davon auszugehen, dass es so eine Art ausgleichender Gerechtigkeit bei Münzwurf, Roulette, Lotto, Würfelspiel und Ähnlichem gibt: Wenn ein Ergebnis – bei einem Münzwurf wären das Kopf oder Zahl – eine Zeitlang gar nicht oder selten vorkam, habe es beim nächsten Versuch eine entsprechend höhere Wahrscheinlichkeit.

Der Denkfehler ist: Je öfter die Münze geworfen wird, umso näher kommt zwar das *relative Verhältnis* von Kopf und Zahl dem Gleichstand, fifty-fifty. Das folgt aus dem mathematischen *Gesetz der großen Zahlen*.[98] Aber die beiden Anzahlen, wie häufig Kopf beziehungsweise Zahl oben liegt, nähern sich *nicht* einander an, sondern können sogar beliebig stark auseinandergehen. Das heißt, diese Anzahlen werden durch *keine* höhere Macht zueinander hingezogen. Wie geht beides zugleich?

Nehmen wir an, Sie haben die Münze tausendmal geworfen und 520-mal Kopf, also 480-mal Zahl erhalten. Die

Differenz ist dann 520 – 480 = 40, und das Verhältnis ist 520/480. Wenn Sie beim eintausendundersten Wurf Kopf werfen, nimmt die absolute Differenz um 1 zu, und wenn Sie Zahl werfen, nimmt sie um 1 ab. Das relative Verhältnis ändert sich entweder zu 521/480 oder zu 520/481. Der Punkt ist: Die Änderung von 520/480 zu 520/481 ist ein klitzekleines bisschen größer als die Änderung von 520/480 zu 521/480 (rechnen Sie nach!).

Das heißt, wenn Kopf und Zahl exakt gleich wahrscheinlich sind, ist zwar gleich wahrscheinlich, ob das Verhältnis sich durch den Münzwurf an 50 % annähert oder sich noch weiter davon entfernt, aber der *Betrag* ist im Falle der Annäherung ein wenig größer als im anderen Fall, 520/480 – 520/481 ist (ein wenig) größer als 521/480 – 520/480. Bei sehr vielen Würfen ist es praktisch unvermeidlich, dass das Verhältnis immer näher an 50 % herankommt, egal wie die Differenz der Anzahlen sich weiterentwickelt.

Allerdings könnte zehnmal Kopf hintereinander auch darauf hinweisen, dass die Münze „gezinkt" ist, so dass Sie beim nächsten Wurf vielleicht besser nicht auf ausgleichende Gerechtigkeit, sondern weiter auf Kopf wetten sollten… □

Anmerkungen

1 Siehe etwa Wikipedia-Artikel „Korrelation", insbesondere Abschnitt „Korrelation und Kausalzusammenhang" (Version: 24.12.2016 um 02:53)
2 Wikipedia-Artikel „Schufa", Prämbel des Abschnitts „Scoring" (Version: 24.11.2016 um 22:20)

3 Siehe beispielsweise „Überwachtes Fahrverhalten: Revolution der Kfz-Versicherung“ von Christian Siedenbiedel, FAZ online vom 13.1.2014 (zugegriffen: 31.12.2016)
4 Pollmer U, Frank G, Warmuth S (2006) Lexikon der Fitnessirrtümer. Eichborn, Frankfurt/Main, Abschnitt „Optimisten leben länger“, S 283 ff
5 Christensen B, Christensen S (2015) Achtung: Statistik – 150 Kolumnen zum Nachdenken und Schmunzeln. Springer, Heidelberg, S 126 ff
6 Wikipedia-Artikel „Schweigespirale“ (Version: 5.1.2017 um 17:20)
7 „Anwesenheitspflicht: Wer nicht kommt, verliert“ von Jan-Martin Wiarda, ZEIT online vom 26.11.2015 (zugegriffen: 27.1.2017); die Originalarbeit ist „Abwesenheit von Lehrveranstaltungen – Ein nur scheinbar triviales Problem“ von Rolf Schulmeister auf campus-innovation.de (zugegriffen: 13.1.2017), siehe hierin insbesondere S 15 ff.
8 Genaueres finden Sie in den Wikipedia-Artikeln „Moderatorvariable“ (Version: 12.4.2014 um 8:44) und „Intervenierende Variable“ (Version: 23.9.2016 um 18:47).
9 Siehe „Cuius regio eius religio“, ifo Standpunkt Nr. 91 vom 12.2.2008, online verfügbar (zugegriffen: 31.12.2016)
10 Weitere Beispiele finden Sie in Krämer W (2006) So lügt man mit Statistik. Piper, München (8. Auflage) in Kap 2 und Kap 14 sowie in Bosbach G, Korff J J (2012) Lügen mit Zahlen – Wie wir mit Statistiken manipuliert werden. Heyne, München (3. Auflage), Kap 3, S 63 ff
11 Zu vorsortierten Stichproben siehe auch Bosbach G, Korff J J (2012) Lügen mit Zahlen – Wie wir mit Statistiken manipuliert werden. Heyne, München (3. Auflage), Kap 6, S 95 ff. und Krämer W (2006) So lügt man mit Statistik. Piper, München (8. Auflage), Kap 8

12 Korrekterweise müsste ich hier „Vorlesungszeit“ statt „Semester“ schreiben, denn nach offizieller Definition werden vielerorts auch die vorlesungsfreien Zeiten in die Semester gerechnet. Aber meines Erachtens wäre es eher verwirrend, hier bei den Begrifflichkeiten hundertprozentig korrekt zu sein.
13 Dieses Beispiel finden Sie in verschiedenen Variationen in etlichen Büchern, zum Beispiel Gigerenzer G (2011) Das Einmaleins der Skepsis – Über den richtigen Umgang mit Zahlen und Risiken. Berlin Verlag (7. Auflage), Abschnitt II.7
14 In etwas weniger vereinfachter Form finden Sie dieses Beispiel auch in der Einleitung zu Kapitel 4 von Reinhart A (2016) Statistics done wrong – Statistik richtig anwenden und gängige Fehler vermeiden. mitp, Frechen
15 „Medikamenten-Zulassung: Kontrolle in der Kritik“ von Katja Riedel, Süddeutsche online vom 23.1.2015 (zugegriffen: 1.5.2017)
16 „Impact of multiple comparisons in randomized clinical trials“ von David Gary Smith et al., The American Journal of Medicine (1987) 83:545-50
17 Für die mathematischen Hintergründe siehe Wikipedia-Artikel „Alphafehler-Kumulierung“ (Version: 12.12.2016 um 05:58)
18 Mehr dazu in Abschnitt 7.1 von Reinhart A (2016) Statistics done wrong – Statistik richtig anwenden und gängige Fehler vermeiden. mitp, Frechen
19 Ebenda, Abschnitt 6.3
20 Freedman diskutiert dieses Phänomen sehr ausführlich in Freedman D H (2010) Falsch! – Warum uns Experten täuschen und wie wir erkennen, wann wir ihnen nicht trauen sollten. Riemann, München, Kap 6, S 188 ff

21 Gigerenzer G (2013) Risiko – Wie man die richtigen Entscheidungen trifft. Bertelsmann, München (5. Auflage), Abschnitt II.6, S 147 ff
22 Ebenda, S 157
23 Wikipedia-Artikel „Edzard Reuter", Abschnitt „Diversifizierung bei Daimler" (Version: 12. Mai 2016 um 12:31)
24 „Hochzeit des Grauens", Süddeutsche online vom 17.5.2010 (zugegriffen: 31.12.2016)
25 Siehe auch Christensen B, Christensen S (2015) Achtung: Statistik – 150 Kolumnen zum Nachdenken und Schmunzeln. Springer, Heidelberg, S 111 ff
26 „Verarbeitete Fleischprodukte, rotes Fleisch: Risiko für Krebs?" Krebsinformationsdienst des Deutschen Krebsforschungszentrums (DKFZ), 26.10.2015, online verfügbar (zugegriffen: 31.12.2016)
27 „Nord-Süd-Gefälle löst Ost-West-Gegensatz ab" von Stefan von Borstel, WELT online vom 1.10.2015 (zugegriffen: 3.11.2016)
28 Weitere allgemeine Ausführungen und Beispiele finden Sie in Abschnitt 4 – „Unvergleichliche Mittelwerte" – von Quatember A (2015) Statistischer Unsinn – Wenn Medien an der Prozenthürde scheitern. Springer, Heidelberg
29 Wikipedia-Artikel „Simpson-Paradoxon" (Version: 24.7.2016 um 13:48); Originalstudie: „Sex bias in graduate admissions: data from Berkeley" von Peter J. Bickel et al, Science 187(4175):398–404, 1975, online verfügbar (zugegriffen: 6.5.2017)
30 Mehr zum Simpson-Paradoxon finden Sie in Bosbach G, Korff J J (2012) Lügen mit Zahlen – Wie wir mit Statistiken manipuliert werden. Heyne, München (3. Auflage), in Kap 9, S 151 ff, Abschnitt 8.3 von Reinhart A (2016) Statistics done wrong – Statistik richtig anwenden und gängige

Fehler vermeiden. mitp, Frechen sowie in Kap 12 von Dubben H-H, Beck-Bornholdt H-P (2010) Mit an Wahrscheinlichkeit grenzender Sicherheit – Logisches Denken und Zufall. Rowohlt, Reinbek (5. Auflage)

31 Dubben H-H, Beck-Bornholdt H-P (2010) Mit an Wahrscheinlichkeit grenzender Sicherheit – Logisches Denken und Zufall. Rowohlt, Reinbek (5. Auflage), Kap 12, S 145 ff

32 Mehr zum Will-Rogers-Paradoxon finden Sie in Bosbach G, Korff J J (2012) Lügen mit Zahlen – Wie wir mit Statistiken manipuliert werden. Heyne, München (3. Auflage), Kap 9, S 146 ff

33 Wikipedia-Artikel „Will-Rogers-Phänomen“ (Version: 11.11.2016 um 09:44)

34 Artikel „Ecological fallacy“ Abschnitt „Robinson's paradox“ in der englischsprachigen Wikipedia (Version: 20.9.2016 um 04:57)

35 Bosbach G, Korff J J (2012) Lügen mit Zahlen – Wie wir mit Statistiken manipuliert werden. Heyne, München (3. Auflage), Kap 1, S 23

36 Wikipedia-Artikel „Orange (Unternehmen)“, Abschnitt „Sonstiges“ (Version: 7.1.2017 um 15:41); siehe auch „Flucht in den Tod“ von Gero von Randow, ZEIT online vom 8.10.2009 (zugegriffen: 31.12.2016)

37 „Morbus Grenzwert – Wie Gesunde zu Patienten gemacht werden“ von Peggy Fuhrmann, SWR2 online vom 1.12.2015 (zugegriffen: 16.1.2017)

38 Christensen B, Christensen S (2015) Achtung: Statistik – 150 Kolumnen zum Nachdenken und Schmunzeln. Springer, Heidelberg, S 98 ff

39 Artikel „Urban heat island“ in der englischsprachigen Wikipedia, Abschnitt „Global warming“ (Version: 27.12.2016 um 12:04)

40 „Die Regressionsfalle“ von Klaus Fiedler, Forschungsmagazin Ruperto Carola der Univ. Heidelberg, Ausgabe 2/2000, online verfügbar (zugegriffen: 16.1.2017)
41 Wikipedia-Artikel „Werbeerfolgskontrolle“ (Version: 10.2.2017 um 10:36)
42 Präambel des Wikipedia-Artikels „Intelligenz“ (Version: 12.1.2017 um 10:52)
43 Der Wikipedia-Artikel „Surrogatmarker“, Abschnitt „Beispiele falscher Surrogat-Marker“ listet eine Reihe weiterer Fälle auf, bei denen Surrogatmarker zu falschen Ergebnissen kommen (Version: 12.10.2016 um 07:29).
44 So scheint *Viszeralfett*, also Fett in der Bauchhöhle, problematischer zu sein als Fett in anderen Zonen. Siehe Wikipedia-Artikel „Viszeralfett“, Abschnitt „Ursachen und Auswirkungen“ (Version: 25.11.2016 um 22:51)
45 Freedman D H (2010) Falsch! – Warum uns Experten täuschen und wie wir erkennen, wann wir ihnen nicht trauen sollten. Riemann, München, Kap 2, S 70 ff
46 „Meine CDU-Mitgliedschaft beruht auf heimlichem Irrtum“ von Marcel Leubecher, WELT online vom 6.10.2016 (zugegriffen: 31.12.2016)
47 Siehe auch „Was zählt die Statistik der Polizei?“ von Thomas Fischer, ZEIT online vom 12.5.2016 (zugegriffen: 2.11.2016).
48 „Bin ich wirklich krank?“ von Corinna Schöps, ZEIT online vom 26.1.2016 (zugegriffen: 2.11.2016)
49 „So wird die Arbeitslosigkeit schöngerechnet“ von Florian Diekmann, SPIEGEL vom 1.3.2017 (zugegriffen: 2.3.2017)
50 Für eine detaillierte Aufarbeitung der statistischen Methodik siehe „Methodenbericht Umfassende Arbeitsmarktstatistik: Arbeitslosigkeit und Unterbeschäftigung“

der Bundesagentur für Arbeit vom Mai 2009 (Autor: Michael Hartmann), online verfügbar über statistik.arbeitsagentur.de (zugegriffen: 2.3.2017). Siehe insbesondere Abschnitt 2.1.2 für die in diesem Fallbeispiel aufgeworfene Problematik: „Mit Wirkung vom 1. Januar 2004 wurde der §16 SGB III um einen zweiten Absatz erweitert. Der neue Absatz 2 hat folgenden Wortlaut: ‚Teilnehmer an Maßnahmen der aktiven Arbeitsmarktpolitik gelten als nicht arbeitslos.'" (Im Original ist der letzte Satz unterstrichen.)

51 Ebenda, Abschnitt 2.2

52 Über destatis.de bietet das Statistische Bundesamt eine Seite „Persönlicher Inflationsrechner" an.

53 „Medizinische Grenzwerte: Krank gesund" von Josephina Mayer, ZEIT online vom 19.6.2014 (zugegriffen: 16.1.2017)

54 „Daran merkst du, dass du intelligenter bist als 80 Prozent der Bevölkerung" von Lisa Mayerhofer, Huffington Post vom 3.11.2016 (zugegriffen: selber Tag)

55 „Der Allesfresser – Über Mythen und Wahrheiten der menschlichen Ernährung" von Yurdagül Zopf, Forschung & Lehre, Juli 2016, online verfügbar (zugegriffen: 31.12.2016)

56 „Dickmacher Frühstück" von Claudia Füßler, Süddeutsche online vom 20.1.2011 (zugegriffen: 16.1.2017) oder „Und dann noch ein Müsli" von Georg Rüschemeyer, FAZ online vom 30.1.2017 (zugegriffen: selber Tag)

57 Wikipedia-Artikel „Naturalistischer Fehlschluss" (Version: 4.1.2017 um 17:32)

58 „Häufigkeitszahl von Straftaten (Straftaten pro 100.000 Einwohner) nach Bundesländern von 2010 bis 2015", Statistisches Bundesamt, online verfügbar über statista.com (zugegriffen: 7.5.2017)

59 Statistik „Beliebteste Freizeitaktivitäten, Hobbies und Sportarten in Deutschland nach häufiger Ausübung in den Jahren 2015 und 2016“, Statistisches Bundesamt, online verfügbar über statista.com (zugegriffen: 8.2.2017)
60 Nur ein Beispiel von vielen: „Mohammed beliebter als Harry“, Süddeutsche online vom 28.10.2010 (zugegriffen: 25.3.2017)
61 „Können Frauen oder Männer besser Schachspielen?“ von Fanny Jiménez, WELT online vom 4.3.2017 (zugegriffen: 5.3.2017)
62 „38 von 50 Anrufe unbeantwortet: Hotline-Test zeigt, warum O2 nicht erreichbar ist“, FOCUS online vom 9.3.2017 (zugegriffen: selber Tag)
63 „Sind Rankings inhärent willkürlich?“ von Dominik Rohn und Karsten Weihe, Forschung & Lehre Nr. 9/2013, S 740–741, online verfügbar über www.wissenschaftsmanagement-online.de
64 Siehe „Ökotest: Schwermetalle im Mineralwasser“ von Christiane Fux, FOCUS online vom 31.7.2006 (zugegriffen: 3.11.2016); für den weiteren Fortgang siehe auch „Schwermetalle im Mineralwasser: Wunderbare Wasser-Wandlung“ von Markus C. Schulte von Drach, Süddeutsche online vom 22.5.2010 (zugegriffen: 25.1.2017)
65 Weitere Aspekte der Problematik, inwieweit Umfragen repräsentativ sind, finden Sie in Abschnitt 6 – „Die Repräsentativitätslüge“ – von Quatember A (2015) Statistischer Unsinn – Wenn Medien an der Prozenthürde scheitern. Springer, Heidelberg.
66 „Die Kombination von Mobilfunk- und Festnetzstichproben“ von Stefan Hunsicker und Yvonne Schroth, Methoden-Daten-Analysen 2007, Heft 2, S 161–182, online verfügbar (zugegriffen: 3.11.2016)

67 Wikipedia-Artikel „Ausschöpfungsquote“, Abschnitt „Bedeutung“ (Version: 19.9.2016 um 02:27)
68 Christensen B, Christensen S (2015) Achtung: Statistik – 150 Kolumnen zum Nachdenken und Schmunzeln. Springer, Heidelberg, S 176
69 Huff D (1956) Wie lügt man mit Statistik. Sanssouci, Zürich, Kap „Die Kunst der statistischen Befragung“, S 9 ff
70 „Islamischer religiöser Fundamentalismus ist weit verbreitet“, Online-Pressemitteilung des Wissenschaftszentrums Berlin für Sozialforschung vom 9.12.2013 (zugegriffen: 31.12.2016)
71 https://bibliothek.wzb.eu/pdf/2014/vi14-101.pdf (zugegriffen: 31.12.2016)
72 „Verhalten der Menschen prüfen“ von Mathias Rohe, Forschung & Lehre 11/16, S 958–960, online verfügbar über wissenschaftsmanagement-online.de (zugegriffen: 25.1.2017)
73 Siehe auch „Daten und Umfragen: Prozente, die nichts bedeuten“ von Thomas Perry, Cicero online vom 7.3.2017 (zugegriffen: selber Tag)
74 „Die enthemmten Wissenschaftler“ von Jasper von Altenbockum, FAZ online vom 17.6.2016 (zugegriffen: 3.11.2016)
75 „Volksabstimmung zu Stuttgart 21: Irreführende Formulierung“, Stuttgarter Zeitung online vom 1.10.2011 (zugegriffen: 3.11.2016)
76 Wikipedia-Artikel „Big Five (Psychologie)“ (Version: 4.1.2017 um 16:22)
77 Siehe etwa „Jeder vierte Deutsche findet Vergewaltigungen okay - manchmal“, WELT online vom 28.11.2016, oder „Jeder vierte Deutsche findet Vergewaltigungen manchmal gerechtfertigt“, STERN online vom 27.11.2016 (beide zugegriffen: 13.12.2016)

78 Wikipedia-Artikel „Garrison Keillor", Abschnitt „Lake Wobegon" (Version: 21.11.2016 um 17:27)
79 Siehe beispielsweise „Erschöpft vom Bummeln" von Manfred Dworschak, SPIEGEL online vom 20.9.2010 (zugegriffen: 19.2.2017)
80 Im Wikipedia-Artikel „Skalenniveau", Abschnitt „Systematik der Skalen" (Version: 28.11.2016 um 15:29) finden Sie eine Auflistung der verschiedenen Skalenarten, und welche mathematischen Operationen jeweils erlaubt sind.
81 „US-Wahl: Warum lagen die Umfragen falsch?" von Lars Fischer, Spektrum der Wissenschaft online vom 9.11.2016 (zugegriffen: 31.12.2016)
82 Ein bei Abfassung dieses Buches aktuelles Beispiel für die andauernde Methodendiskussion finden Sie in „Eisschwund in der Arktis – Schuld ist nicht nur der Mensch" von Christoph Seidler, SPIEGEL online vom 14.3.2017 (zugegriffen: 15.3.2017). Der Autor liefert meines Erachtens interessante Einsichten in die Problematik, zum Beispiel im Absatz zur Kritik an der Methodik der vorgestellten Studie.
83 Ein weiteres, bei Abfassung dieses Buches brandaktuelles Beispiel für ein Known Unknown in Klimamodellen – wie sich globale Erwärmung auf den Golfstrom auswirken würde – finden Sie in „Forscher warnen vor Kollaps des Golfstroms" von Christopher Schrader, SPIEGEL online vom 23.1.2017 (zugegriffen: selber Tag).
84 Der in Anmerkung 83 zitierte Artikel spricht auf S 4 auch die Problematik der zu groben Modellierung an: „Sonst braucht man zu viel Rechenzeit."
85 Taleb N N (2010) Der schwarze Schwan – Die Macht höchst unwahrscheinlicher Ereignisse. dtv, München

86 Nur ein Beispiel von vielen: „Ökonomen: Zielsicher daneben“ von Matthias Auer, Die Presse online vom 28.6.2014 (zugegriffen: 31.12.2016)
87 Siehe beispielsweise „Das Scheitern der Marktforscher“ von Ingo Pakalski, Golem News vom 1.12.2016 (zugegriffen: 7.12.2016)
88 „Konjunkturprognosen: Falsche Erwartungen an Treffgenauigkeit“ von Jörg Hinze, Wirtschaftsdienst, Heft 2, S 117–123, online verfügbar (zugegriffen: 23.9.2016)
89 Zum Beispiel „Bis zu 100 Euro extra: Das sind versteckte Kosten beim Onlineshopping“, Focus online vom 19.6.2015 (zugegriffen: 8.2.2017)
90 Siehe „Vorsicht vor Lockvogelangeboten: Bei diesen Verträgen zahlen Bauherren drauf“ von Tatjana Grassl, Focus online vom 26.10.2016 (zugegriffen: 27.10.2016) oder „Die neue Falle beim Kauf der eigenen vier Wände“ von Richard Haimann auf WELT online vom 28.10.2016 (zugegriffen: 31.12.2016)
91 „Restschuldversicherungen sind teuer und oft überflüssig“ von Britta Beate Schön, Finanztip vom 1.2.2017, online verfügbar (zugegriffen: 5.2.2017)
92 „Restkreditversicherungen: Policen mit schlechtem Ruf“ von Philipp Krohn, FAZ online vom 29.11.2016 (zugegriffen: selber Tag), Zitat: „Verbraucherschützer hätten regelmäßig mit Konsumenten zu tun, die sich nicht im Klaren darüber waren, dass der Vertrag für sie nur optional und nicht verpflichtend war.“
93 Ebenda: „Oft werden sie sanft unter Druck gesetzt, nach dem Motto: sie wollen doch den Kredit.“
94 Hartmut Walz diskutiert dieses Verbleiben bei einer einmal getroffenen Entscheidung ausführlicher unter dem Titel „Gefrorene Entscheidung – Wer A sagt, muss nicht B sagen“ in Walz H (2015) Einfach genial entscheiden – die 55

wichtigsten Erkenntnisse für Ihren Erfolg. Haufe, Freiburg (2. Auflage), S 24 ff
95 Pollmer U, Frank G, Warmuth S (2006) Lexikon der Fitnessirrtümer. Eichborn, Frankfurt/Main
96 „Die Tücken des Durchschnitts" von Gerhard Schwarz, Onlineportal von Avenir Suisse vom 2.2.2015 (zugegriffen: 26.1.2017)
97 Siehe beispielsweise den Wikipedia-Artikel „Anzahl der notwendigen Behandlungen" (Version: 28.10. 2016 um 12:39) für die allgemeine Definition der NNT, die sich auf den *Vergleich* von *zwei* Therapien bezieht (wovon eine natürlich auch eine Placebobehandlung oder Nichtbehandlung sein kann)
98 Wikipedia-Artikel „Gesetz der großen Zahlen" (Version: 11.10.2016 um 14:40)

http://www.springer.com/978-3-662-54703-8

Fundiert entscheiden
Ein kleines Handbuch für alle Lebenslagen
Weihe, K.
2018, XVI, 290 S. 14 Abb. Book + eBook., Softcover
ISBN: 978-3-662-54703-8