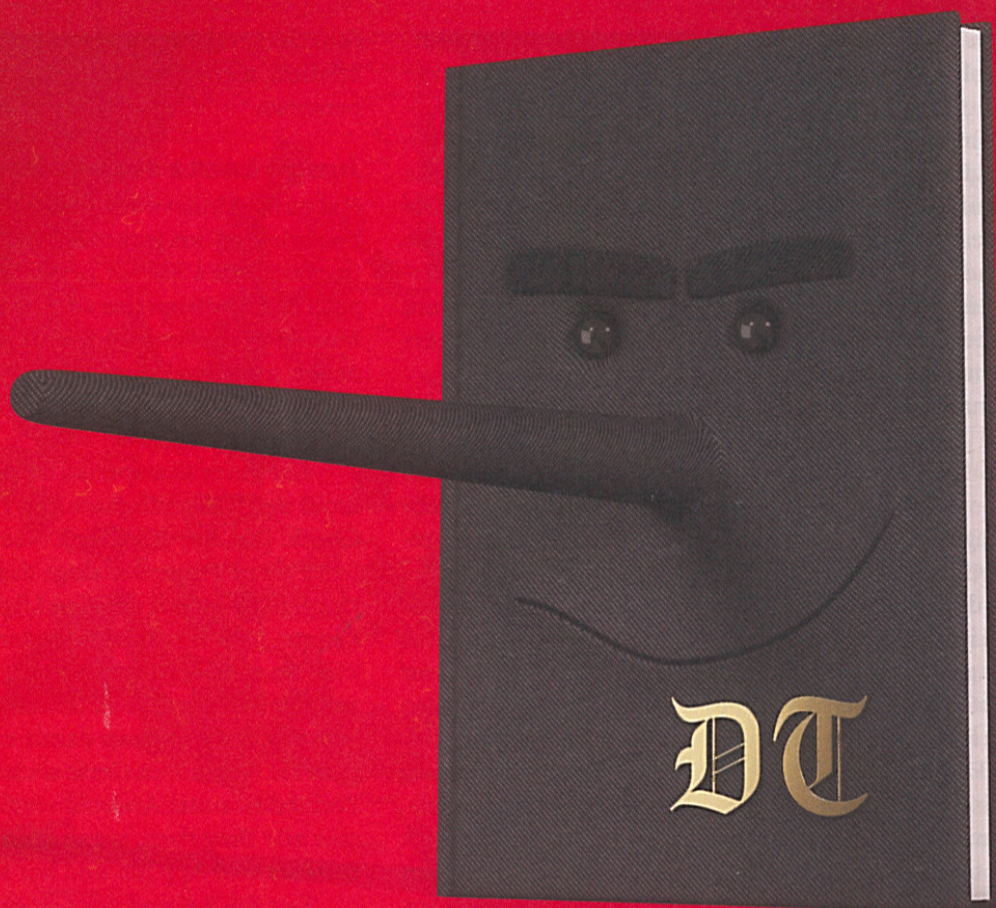




Schlechte Zahlen, böse Zahlen

Statistische Zerrbilder und wie man sie erkennt

Sogar aus korrekten Zahlen liefert manche Statistik ein Trugbild der Wirklichkeit. Wer es darauf anlegt, kann damit brillante Fake News konstruieren. Doch auch seriöse Forscher geraten mit der Statistik leicht auf den Holzweg. Dabei kann man viele Fallgruben im mathematischen Dickicht mit einfachen Faustregeln enttarnen.

Von Peter Schüler

Die Schlagzeile „Rauchen kommt wieder in Mode“ wird bei vielen Lesern auf Skepsis stoßen. Anders, wenn sie statistisch belegt wird: Laut BBC ist die Zahl der Raucher von 1980 bis 2012 weltweit um mehr als ein Drittel gestiegen, das kann man ja wohl nicht wegdiskutieren? Doch! Im selben Zeitraum ist nämlich die Weltbevölkerung um mehr als ein Drittel gewachsen, der Anteil der Raucher ist also tatsächlich nicht gestiegen, sondern gesunken (wie zumindest die BBC auch korrekt erläutert) [1].

„Zahlen lügen nicht“ sagt das Sprichwort und suggeriert eindeutige Wahrheiten. Doch mit Statistik kann man aus Zahlen ein robustes Lügengebäude hochziehen, dessen Ursprung erfahrungsgemäß nur selten hinterfragt wird. Dabei muss eine Falschaussage gar nicht auf böser Absicht beruhen: Nicht Jeder ist mit den höheren Weihen der Statistik gesegnet, und mancher interpretiert die Zahlen unwissentlich falsch.

Statistik ist eine Technik, mit der man Sachverhalte zwar beschreiben, aber

normalerweise nicht erklären kann. Ohne Hintergrundwissen von einem gefundenen Zusammenhang auf eine kausale Beziehung zu schließen, ist immer Scharlatanerie. Allenfalls kann man einen kausalen Zusammenhang, den man sich hergeleitet hat, mit einem statistischen Test überprüfen. Dabei kommt dann heraus, dass die Ausgangshypothese mit einer bestimmten Wahrscheinlichkeit zutrifft oder dass sie durch zufällige Beobachtungen nur vorgetäuscht wird.

Deskriptive Statistik wie die Ermittlung von Durchschnittswerten und Streubereichen ist durchaus nützlich, etwa um auszudrücken, dass ein Liter Superbenzin zu einem bestimmten Zeitpunkt $1,34 \text{ Euro} \pm 0,16 \text{ Euro}$ kostet. Mit dieser Verkürzung aller berücksichtigten Tankstellenpreise auf zwei Zahlen wäre es ein Leichtes für eine Tankstelle, Verbraucher zu täuschen. Das könnte nur durchschauen, wer Zugriff auf alle Einzelwerte hat.

Vorsicht, Umfrage

Katharina Schüller durchleuchtet im Buch „Statistik und Intuition“ eine Beraterstudie über die wichtigsten Anforderungen von Jobsuchenden an ihren Arbeitsplatz [2]. Die Urheber der Studie vergleichen die Ergebnisse zweier Befragungen aus den Jahren 2008 und 2010 [3]. Demnach kam es im Jahr 2010 vor allem anderen aufs Betriebsklima und das Verhältnis zwischen Arbeit und Freizeit an. Das Gehalt zählte nur für 31 Prozent der Befragten zu den wichtigsten Kriterien und rangierte auf Platz vier der Rangliste. 2008 war es immerhin noch 78 Prozent der Befragten auch aufs Einkommen angekommen.

Allerdings waren die beiden Befragungen in vielerlei Hinsicht unterschiedlich aufgebaut, was man erst bei einem Blick auf die Details erkennt. Die einleitende Frage 2008 galt „Faktoren für die Arbeitgeberwahl“, während es 2010 zuerst um „Ziele und Werte im Leben“ und erst danach um Job- und Arbeitgebermerkmale ging. 2008 standen 16 Arbeitgeberereignisse zur Wahl, 2010 nur deren 13. In der 2008er-Befragung hätte man die dominierende Rangliste der 2010er-Umfrage gar nicht zum Ausdruck bringen können.

Auch in anderen Fällen empfiehlt sich eine Hintergrundrecherche, etwa wenn ein Milliardenmarkt für Medikamente auf dem Spiel steht: Nach heutigen Maßstäben haben in den USA fast doppelt so viele Patienten wie vor zehn Jahren erhöhte

Cholesterinwerte. Was nach einem Kollaps der Volksgesundheit aussieht, beruht aber nach der Erklärung des Arztes Gunter Frank in seinem Buch „Schlechte Medizin“ auf einem statistischen Artefakt [4].

Noch in den 1950er Jahren galten 260 mg Cholesterin pro Deziliter (dl) Blut als unbedenklich. US-Behörden orientieren sich seit mehr als einem Jahrzehnt an einem Schwellenwert von 240 mg/dl und dokumentieren nur geringe Veränderungen im Bevölkerungsdurchschnitt. Anders die Ärzteschaft: In der medizinischen Literatur ist der empfohlene Schwellenwert bis heute kontinuierlich auf 200 mg/dl gesunken. Frank hält das zwar für medizinisch unbegründet, doch die Absenkung des Schwellenwerts macht aus ursprünglich rund 50 Millionen US-Patienten nicht weniger als 92 Millionen – und damit ebenso viele potenzielle Pharma-Kunden.

Zahlenrauschen

Selbst überprüfbar korrekte Zahlenwerte aus Naturwissenschaft, Technik und Ökonomie sind keineswegs ein sicheres Parkett. Viele Zahlenwerte lassen sich nur korrekt interpretieren, wenn man sie in den richtigen Kontext setzt. Zum Beispiel werden sich wohl die meisten Patienten für eine Therapie entscheiden, wenn ihr Arzt verspricht, dadurch sinke das Risiko eines Herzinfarkts statistisch um 28 Prozent, das Risiko, an der Therapie selbst zu sterben, liege aber nur bei 0,3 Prozent. Böse Falle: Die genannten 28 Prozent sind eine Relativangabe und sagen in Wirklichkeit, dass die Therapie das Infarkt-Risiko von 1,32 auf 0,95 Prozent vermindert, also um absolut 0,37 Prozent. Korrekterweise müsste der Arzt also die Risiken von 0,37 und 0,3 Prozent gegenüberstellen – das würde viele Patienten sicher zu einer anderen Entscheidung veranlassen.

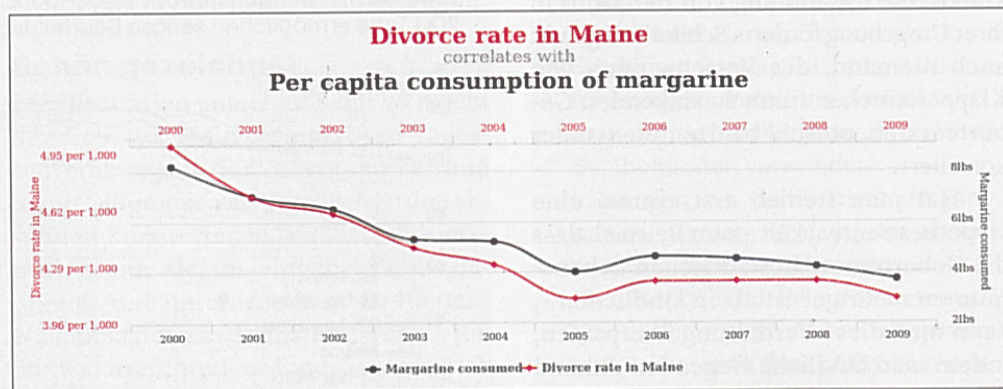
Aktuell verkündet mancher Werbespot etwa „96 Prozent der befragten Kunden sind begeistert von Produkt X.“ Die einzige echte Information darin ist, dass mindestens 25 Kunden befragt worden sind, sonst wäre der Wert von 96 Prozent nicht herleitbar. 25 Kunden wären für eine seriöse Aussage viel zu wenige, ganz abgesehen davon, dass nichts darüber gesagt wird, wie diese Kunden ausgewählt worden sind. Benutzer, denen zuvor eine Probe des Produkts geschenkt worden ist, wären vermutlich nicht repräsentativ. Die scheinbar statistisch belegte Behauptung ist daher leeres Gerede.

Auch die Präzision, die eine Zahl mit vielen Nachkommastellen ausstrahlt, zeugt mitunter nur von Unverstand. Zum Beispiel kommen nach Angaben des CIA World Factbook 2007 auf jeweils 1000 Bürger von Monaco 707,69 Pkw. In Liechtenstein sind es 705,88 Pkw, in Island 564,64 [4].

Die Angaben sind mit fünf gültigen Stellen überraschend genau. Rechnet man daraus aber mit den Bevölkerungszahlen aus derselben Quelle auf die absolute Zahl an Pkw in diesen Ländern zurück, gibt es dort offenbar genau 23.000, 24.000 und 169.000 Pkw. Der Schluss liegt nahe, dass die fünfstelligen pro-Kopf-Angaben auf Rohdaten beruhen, die gerade einmal auf zwei bis drei Stellen genau sind. Nur für größere Länder scheinen genauere Rohdaten in die Statistik eingeflossen zu sein.

Scheidungsgrund Margarine?

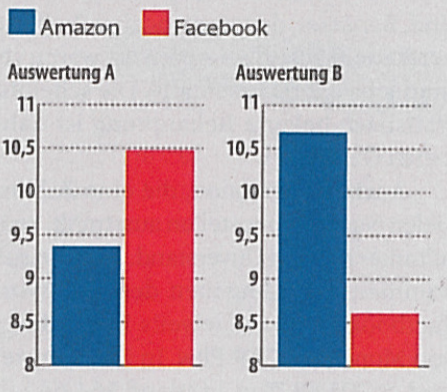
Der US-Analyst Tyler Vigen hat zum Spaß eine Software entwickelt, die beliebige große Datenbestände nach überraschenden Zusammenhängen durchforstet [5]. Aus seinen Befunden kann man zum Beispiel lernen, dass die Scheidungsrate im US-Bundesstaat Maine von 2000 bis 2009 im selben Maß zurückgegangen ist wie



Die zeitlichen Verläufe des Margarineverbrauchs pro US-Bürger und der Scheidungsrate im Bundesstaat Maine sind sich verblüffend ähnlich.

Prozentualer Wertzuwachs

Für ein und denselben Zeitraum lassen sich die Kurszuwächse von Facebook- und Amazon-Aktien ganz unterschiedlich angeben. Beide Auswertungen basieren auf den authentischen Kursdaten.



der Pro-Kopf-Verbrauch von Margarine in den ganzen USA. Bedrohliche Erklärungen etwa zur entfremdenden Wirkung des Margarinekonsums sind uns nicht bekannt, wären aber auch fragwürdig. Wenn es dessen noch bedürfte, wäre ein Anlass zur Skepsis bereits die Vermengung von Daten aus den ganzen USA mit solchen aus einem einzigen Bundesstaat. Tyler hat mehr als 24.000 solcher Korrelationen entdeckt, bis zum Zusammenhang zwischen dem durchschnittlichen Honig-Ertrag je Bienenvolk in den USA und der Zahl von Rechtsanwälten in Amerikanisch-Samoa.

Die Suche nach ähnlichen Mustern kann durchaus Hinweise auf mögliche Zusammenhänge liefern, etwa wenn man in der Nähe von Kernkraftwerken besonders viele Fälle von Blutkrebs beobachtet. Die Übereinstimmung der Beobachtungen allein beweist aber mitnichten, dass Kernkraftwerke die Bildung von Blutkrebs in ihrer Umgebung fördern. Schließlich glaubt auch niemand, das Verschwinden von Klapperstörchen führe zu sinkenden Geburtenraten, obwohl beides miteinander korreliert.

Hat man freilich erst einmal eine Hypothese entwickelt – zum Beispiel, dass die Geburtenrate in städtischen Lebensräumen niedriger ist als in ländlichen –, kann man diese Vermutung überprüfen, indem man ländliche Gegenden anhand der Dichte von Storchhorsten ausmacht. Tatsächlich hängen Storch- und Geburtenzahl begründbar von diesem gemein-

samen Faktor ab. Dieses Phänomen bezeichnen Statistiker als Scheinkorrelation, auch wenn der Begriff Scheinkausalität vielleicht mehr Sinn ergäbe.

Die Guten ins Töpfchen ...

Viele Statistiker sehen sich in Entscheidungsnöten, wenn sie in einer Datengsamtheit einige Werte finden, die nicht ins Bild passen. Wenn man das Durchschnittsalter der Mitglieder eines Sportvereins ermittelt, ist sofort klar, dass man Datensätze mit der Altersangabe 0 ignorieren muss. In diesen Einträgen ist offenbar etwas schiefgegangen – vielleicht hat das Mitglied das zugehörige Datenfeld im Aufnahmeantrag schlicht übersehen.

Anders liegt der Fall beim Body-Mass-Index (BMI) der Mitglieder. Je mehr Mitglieder man betrachtet, desto besser lässt sich zwar abschätzen, ob eine Angabe noch im Bereich des Üblichen liegt. Mitunter hat sich ein Mitglied bei der Angabe von Größe oder Gewicht vertan oder war vielleicht mit der Berechnung des BMI überfordert. Solche Datensätze muss man bei der Berechnung ausklammern – nur sind diese Kandidaten nicht immer ganz leicht auszumachen.

Die Methode nach Lehrbuch sieht vor, dass man für jeden möglichen BMI-Wert aufträgt, bei wie vielen Mitgliedern er auftaucht. Bei kontinuierlichen Merkmalen teilt man den Gesamtbereich möglicher Werte in gleich große Intervalle und zählt, wie viele Einzelwerte in jedes Intervall fallen. Daraus erhält man ein sogenanntes Histogramm. In diesem Fall sollten sich die Häufigkeiten der gemessenen Werte als

Normalverteilung beschreiben lassen und das Histogramm sollte etwa die Silhouette einer Glocke wiedergeben. Diese Hypothese lässt sich mit mathematischen Tests überprüfen. Abhängig von den Einzelwerten und deren Anzahl erfährt man daraus, mit welcher Wahrscheinlichkeit tatsächlich eine Normalverteilung vorliegt.

Ist das bestätigt, kann man für jeden ungewöhnlichen Wert ausrechnen, mit welcher Wahrscheinlichkeit er einen Ausreißer darstellt und ignoriert werden darf. Die Prozedur ist mühselig und erfordert einigen statistischen Sachverstand. Viele Forscher verzichten auf diese Ochsentour und identifizieren Ausreißer aus dem Bauch heraus. Wenn sie zudem ein bestimmtes Ergebnis erwarten, können sie damit unbewusst auch die scheinbar harte Zahlenrechnung beachtlich aufweichen.

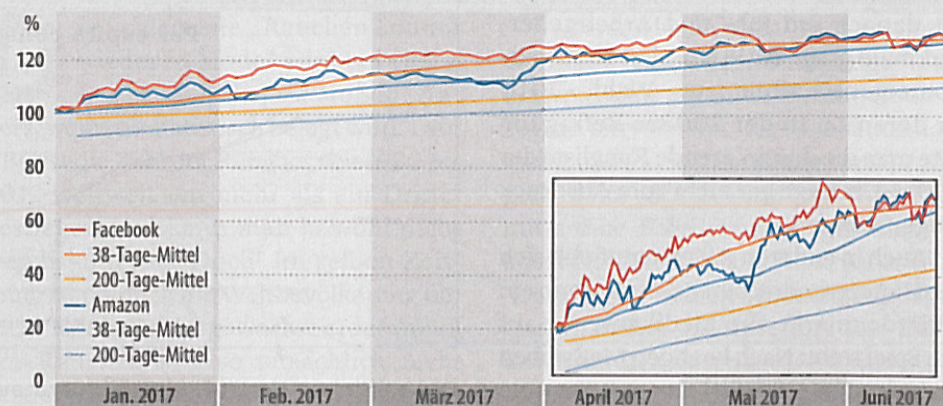
Fischgeruch

Histogramme verraten Einiges über die Entstehung einer statistischen Aussage. So überprüfen Industriebetriebe, die riesige Serien von Zulieferteilen beziehen, die Fertigungsgenauigkeit dieser Teile mit Histogrammen von Vorserien. Entsprechen die Prüfmaße dieser Teile nicht der Normalverteilung, liegt nahe, dass der Zulieferer seinen Herstellungsprozess nicht gut genug beherrscht und die Teile mit unzulässigen Abweichungen nachträglich von Hand aussortiert hat. Das könnte den Kunden davon abhalten, die Serienfertigung freizugeben, obwohl statistisch 100 Prozent aller Teile den Anforderungen entsprechen.

Der Statistiker Sergeij Schpilkin will an den Auszählungsdaten erkannt haben,

Verrauschte Aktienkurse

Unterschiede der Kursentwicklung von Facebook- und Amazon-Aktien sind winzig im Vergleich zu den Kurswerten selbst. Das Bild im Bild zeigt, wie sehr der Eindruck durch Tagesschwankungen getrübt wird. Höchstens die gleitenden Mittelwerte über 38 oder 200 Tage ermöglichen seriöse Beurteilungen mit den jeweiligen Zeit-Auflösungen.



dass die russische Duma-Wahl 2008 manipuliert worden ist. Das Wahlergebnis begründete die Ernennung von Dmitri Medwedew zum Nachfolger von Wladimir Putin als Staatspräsident.

Auch wenn die genaue Vorgehensweise nicht publiziert ist, hat Schpilkin offenbar für jede Partei die Häufigkeit von Stimmanteilen über der Zahl der Stimmbezirke aufgetragen. Bei vielen internationalen Wahlen, auch russischen Duma-Wahlen, fand er für alle Parteien ähnliche Verteilungskurven, bis auf Putins und Medwedews Partei „Einiges Russland“ in der besagten Duma-Wahl. Deren Stimmverteilung war so verzerrt, dass die Partei in unerwartet vielen Stimmbezirken – vor allem in solchen mit annähernd 100-prozentiger Wahlbeteiligung – überdurchschnittlich gut abgeschnitten hat.

Das Beispiel ist symptomatisch für viele statistische Argumentationen: Die Beobachtung begründet einen massiven Verdacht auf Wahlfälschung, taugt aber nicht als Beweis. Diesen könnte man nur mit detektivischen Nachforschungen erbringen, zu denen es freilich kaum kommen wird. Folglich bleibt man als Beobachter darauf angewiesen, sich anhand der Statistik ein persönliches, nicht objektivierbares Urteil zu bilden.

Facebook oder Amazon?

Zwei Bankberater empfehlen, Geld in Internet-Aktien zu investieren. Berater Hinz plädiert für den Kauf von Facebook-Aktien mit dem Argument, die hätten im Zeitraum Februar bis März 2017 stärker an Wert zugelegt als die alternativ vorgeschlagenen Anteile von Amazon. Dieselbe Entwicklung könne man auch für die unmittelbare Zukunft erwarten. In einer anderen Bank zeigt Berater Kunz ein ganz anderes Bild: Er legt Zahlen für denselben Zeitraum vor, doch denen zufolge ist Amazon fast um ein Drittel stärker gewachsen als Facebook.

Beide Banker waren ganz ehrlich – die Zahlen sind authentisch, beziehen sich auf dieselben Wertpapiere und Handelsplätze. Der einzige Unterschied ist: Herr Hinz hat den Zeitraum vom 31. Januar bis zum 31. März betrachtet, Herr Kunz rechnet vom 3. Februar bis zum 3. April.

Für die absehbaren Börsentrends hat dieser Unterschied keinerlei Bedeutung: Die detaillierten, im Maßstab angeglichenen Kursverläufe im Diagramm unten auf Seite 84 zeichnen ein realistisches Bild der Tendenzen mit erheblichen

Kursschwankungen von Tag zu Tag. Die gezeigten Balkendiagramme beruhen dagegen für jede Aktie nur auf zwei herausgegriffenen Einzelwerten. Daraus errechnet sich nur eine zufällige Kursdifferenz, aber keine repräsentative.

Für Prognosen mit einer besseren Treffsicherheit sind viel mehr Ausgangsdaten erforderlich. Im konkreten Fall ließe sich das übrigens recht einfach bewerkstelligen, wenn man statt der verrauchten Einzelwerte Mittelwerte aus mehreren aufeinanderfolgenden Kurswerten verwendete. Kurven solcher gleitender Mittelwerte aus jeweils 38 und 200 Werten, wie sie ebenfalls im unteren Bild eingezeichnet sind, dienen Börsianern häufig als Entscheidungshilfen.

Böse mit System

Die krasse Abweichung zwischen den beiden Prognosen ist zwar nachweislich möglich, aber äußerst atypisch. Das Risiko, im geschilderten Szenario so absurde Ergebnisse zu erhalten, ist im Alltag vernachlässigbar – nur kann man den Grafiken das nicht ansehen.

Um zum gezeigten Ergebnis zu kommen, haben wir uns auf die dunkle Seite der Statistik begeben: Wir haben den üppigen Vorrat an Datenpunkten gezielt nach Mustern durchsucht, die zum gewünschten Ergebnis führen. Erwartungsgemäß sind wir dabei mühelos fündig geworden; irgendwelche passenden Zusammenhänge lassen sich in einem großen Datenbestand fast immer entdecken.

Außerdem haben wir vergleichsweise kleine Differenzen künstlich aufgebläht: Sowohl grafisch als auch in Zahlen wirkt der Unterschied zwischen 8,5 und 10,7 Prozent Wertzuwachs viel bedeutender, als wenn wir statt der Differenz die gesteigerten Werte selbst miteinander verglichen hätten: dann hätten nämlich 108,5 und 110,7 Prozent nebeneinander gestanden.

Schön gezeichnet

Die Illustration unten auf Seite 84 macht deutlich, wie klein die nervösen Tagesschwankungen der Börsenkurse und deren Differenzen im Vergleich zu den absoluten Kurswerten in Wirklichkeit sind – auch wenn sie im eingebetteten Diagramm mit unterdrücktem Ordinaten-Nullpunkt noch so dramatisch wirken. Ein ähnlich irreführender Eindruck lässt sich mit Symbolen in Grafiken erzielen: Dann markiert man etwa die Datenpunkte einer Auto-Verkaufsstatistik mit symbolischen

Autobildern und „verdeutlicht“ eine zehnprozentige Steigerung durch eine Verdopplung der Symbolgröße.

Unfallfrei durchs Minenfeld

Die obigen Beispiele zeigen auf, mit wie unterschiedlichen Fußangeln eine statistische Argumentation gespickt sein kann. Um darüber nicht ins Stolpern zu geraten, hilft schon das Problembewusstsein, dass die Zahlenwerte einer Statistik aus vertrauenswürdiger Quelle nicht automatisch die unbestreitbare Wahrheit verkünden. Außerdem gibt der Kasten unten Hinweise, wie man in vielen Fällen zwischen Erleuchtung und Unfug unterscheiden kann. (hps@ct.de) **ct**

Literatur

- [1] BBC über Raucher-Trends: www.bbc.com/news/health-25635121
- [2] Katharina Schüller, Statistik und Intuition
- [3] www.pressebox.de/inaktiv/kienbaum-consultants-international-gmbh/Arbeitgeberwahl-Geld-ist-nicht-mehr-so-wichtig/boxid/325138
- [4] Dr. med. Gunter Frank, „Schlechte Medizin: Ein Wutbuch“
- [5] Statistik über Pkw-Besitz: www.welt-in-zahlen.de/laendervergleich.phtml
- [6] Tyler Vigen, Spurious Correlations

Mit Vorsicht genießen

Bevor man einer veröffentlichten Statistik traut, sollte man sich routinemäßig ein paar Fragen stellen, um deren Plausibilität zu prüfen.

- Passen die gezeigten Daten genau zur konkret formulierten These oder werden zum Beispiel absolute Zahlen als Stellvertreter für prozentuale Anteile genannt?
- Sind die zeitlichen und logischen Grenzen des Datenbereichs erklärlich oder sind womöglich unpassende Zeiträume ausgeblendet worden?
- Gibt es Quellenangaben, mit denen man die Echtheit der Daten überprüfen könnte?
- Suggestiert eine Grafik falsche Schlussfolgerungen, indem sie irregulär skalierte Achsen oder unpassend variierte Symbolgrößen verwendet?
- Lässt sich die Genauigkeit der Angaben mit dem Umfang der untersuchten Daten rechtfertigen?
- (Schwierig) Sind Teile des Datenbestands durch nicht erwähnte Faktoren verfälscht, etwa, weil gleichlautende Antworten aus unterschiedliche Umfragen verglichen werden?