Die Schönheit der Statistik

CHRISTEL WEISS

Medizinische Fakultät Mannheim der Universität Heidelberg

Zusammenfassung

In diesem Beitrag wird den Gründen nachgegangen, weshalb Kenntnisse in Statistik in vielerlei Hinsicht von Nutzen sind und sogar zu begeistern vermögen. Anhand eines Beispiels (Mammographie-Screening) wird dargelegt, wie ein angemessener Umgang mit Wahrscheinlichkeiten hilft, Informationen sinnvoll einzuordnen und Risiken adäquat zu interpretieren. Außerdem wird erörtert, wie mit statistischen Kenngrößen komplexe Sachverhalte in genial einfacher Weise numerisch beschrieben werden, welche Aussagen in Schätzwerten verborgen sind und wie in empirischen Studien statistische Verfahren zu neuen Erkenntnissen führen können. Schließlich wird anhand der Beispiele einiger bekannter Persönlichkeiten aufgezeigt, wie diese die Schönheit und Eleganz statistischer Verfahren empfunden haben.

1 Einleitung

Statistik und Schönheit – viele Menschen sehen darin einen Widerspruch. Den Begriff „Schönheit“ verbindet man eher mit einem Kunstwerk, mit einem Musikstück oder einem menschlichen Antlitz als mit einer spröden Wissenschaft, die sich mit der Analyse von Zahlen befasst.

Statistik polarisiert, ähnlich wie Mathematik. Während jedoch die meisten Menschen in ihrem Alltag weitgehend ohne Mathematik auskommen, kann sich der Statistik kaum jemand entziehen. Beginnend bei der Zeitungslektüre am frühen Morgen bis zur Nachrichtensendung am späten Abend wird man quasi rund um

http://dx.doi.org/10.17885/heiup.hdjbo.2019.0.24017

die Uhr über diverse Kommunikationsmedien mit Wahrscheinlichkeiten, Durchschnittswerten, Zahlenwerken und Diagrammen konfrontiert. Die Reaktionen der Empfänger sind unterschiedlich: Manche nehmen diese Botschaften mit aufmerksamen Interesse wahr, weil sie glauben, auf diese Weise schnell, umfassend und objektiv über aktuelle Geschehnisse informiert zu werden. Andere fühlen sich unbehaglich angesichts der Fülle der auf sie einströmenden Informationen, weil sie befürchten, manipuliert zu werden. Aphorismen wie „Glaube keiner Statistik, die du nicht selbst gefälscht hast“ (Winston Churchill zugeschrieben) legen darüber ein beredtes Zeugnis ab.

In vielen empirischen Wissenschaften wie etwa in Bio- und Sozialwissenschaften, in der Psychologie, in Wirtschaftswissenschaften und im Versicherungswesen stellen statistische Analysemethoden eine unentbehrliche Hilfe dar. Die meisten Wissenschaftler, die statistische Analysen in ihrem Fachgebiet anwenden, sind zwar beeindruckt von den Möglichkeiten einer effizienten Datenanalyse, angefangen beim Strukturieren der Daten über das Aufdecken von Zusammenhängen bis hin zum Gewinnen neuer Erkenntnisse. Nicht alle Anwender dürften indessen Statistik mit den Begriffen „Schönheit“ oder „Ästhetik“ assoziieren – auch wenn sie diese Techniken nutzen und davon profitieren.

Derlei negative Empfindungen sind unter anderem darauf zurückzuführen, dass die meisten Menschen niemals systematisch gelernt haben, aus statistischen Darstellungen relevante Informationen zu extrahieren oder deren Ergebnisse adäquat zu interpretieren. Viele erinnern sich mit Grauen an ihren Mathematikunterricht in der Schule oder an die Statistikvorlesung, die sie gezwungenermaßen während ihres Studiums besuchen mussten.

Dieser Beitrag soll vermitteln, dass Statistik zu begeistern vermag – nicht nur Mathematiker (die statistische Methoden entwickeln), sondern auch Wissenschaftler anderer Fachrichtungen (die diese Methoden anwenden) und Menschen im ganz normalen Alltag (die mit den Ergebnissen statistischer Analysen konfrontiert werden). Besondere Beachtung finden hierbei statistische Anwendungen in der Medizin.

2 Historische Betrachtungen

Um der Frage nachzugehen, wie Statistik mit Schönheit assoziiert ist, lohnt es sich, deren historische Entwicklung zu betrachten. Die einfachste, seit Urzeiten an-

gewandte statistische Methode ist das Zählen – was Menschen seit jeher fasziniert. Bereits im Alten Testament wird im 4. Buch Mose (das den Namen „Numeri“ trägt) eine Volkszählung erwähnt, bei der alle Männer der zwölf Stämme Israels gezählt werden. Bis ins 18. Jahrhundert hinein war Statistik nahezu gleichbedeutend mit „Staatsbeschreibung“. So war es möglich, geografische, wirtschaftliche und politische Besonderheiten eines Landes quantitativ zu erfassen und darauf basierend objektive Vergleiche anzustellen. Daran hat sich bis heute prinzipiell nichts geändert.

Ein anderes Anwendungsgebiet der Statistik ergab sich, als man begann, alle Geburten und Todesfälle systematisch in Kirchenbüchern aufzuzeichnen. Der preußische Feldprediger **Johann Peter Süßmilch** (1707–1767) wertete derlei Eintragungen aus und entdeckte dabei zu seinem Erstaunen in der Gesamtheit aller unvorhersehbaren Einzelereignisse Regelmäßigkeiten, die er als Zeichen eines göttlichen Plans verstand. Daraufhin verfasste er das bahnbrechende Werk der deutschen Bevölkerungsstatistik, dem er den Titel „Die göttliche Ordnung in den Veränderungen des menschlichen Geschlechts“ gab [14]. Daneben bildete sich im 17. Jahrhundert eine Anwendergruppe mit gänzlich anderen Interessen: Ihnen ging es um die Berechnung von Gewinnchancen bei Glücksspielen. Als namhafte Vertreter dieser Richtung seien **Galileo Galilei** (1564–1642) und **Blaise Pascal** (1623–1662) genannt. Enthusiastisch grübelten sie über ihren Denksportaufgaben, woraus theoretische Abhandlungen resultierten, mit denen sie die Wahrscheinlichkeitsrechnung wesentlich bereichert haben. An dieser Stelle verdient auch der bekannte Mathematiker **Carl Friedrich Gauß** (1777–1855) Erwähnung. Gauß hat zahlreiche Beiträge zu diversen Wissenschaftsgebieten geleistet, unter anderem zur Statistik. Zweifelsohne konnte er sich an der Schönheit der Glockenkurve und der Eleganz der von ihm entwickelten Methoden erfreuen. Tatsächlich ist Schönheit für Mathematiker eine wichtige Kategorie.

Nach Auffassung des Schweizer Mathematikers **Daniel Bernoulli** (1700–1782) spiegeln Wahrscheinlichkeiten die Unvollkommenheit menschlichen Wissens wider. Nichtsdestotrotz gelang es ihm, Wahrscheinlichkeiten praktisch zu nutzen: Als in der Mitte des 18. Jahrhunderts die Pocken in Europa wüteten, stand eine Impfung zur Verfügung, die in der Türkei längst praktiziert wurde. Es bestand allerdings das Risiko 1:200, dass ein Geimpfter innerhalb eines Monats an den Folgen der Impfung verstarb. Trotzdem konnte Bernoulli rechnerisch nachweisen, dass durch die Impfung im Durchschnitt ein Zuwachs an Lebenserwartung zu erwarten

war. Er sah dies als eine Gelegenheit, die Öffentlichkeit objektiv vom Vorteil dieser Maßnahme zu überzeugen [5]. Auch der französische Mathematiker **Pierre Simon Marquis de Laplace** (1749–1827) stellte zu Beginn des 19. Jahrhunderts eine Verbindung zwischen Wahrscheinlichkeitsrechnung und Medizin her, indem er die Meinung vertrat, dass sich aus einer Vielzahl von Beobachtungen therapeutischer Erfolge oder Misserfolge Verallgemeinerungen ziehen ließen. Einige Jahre später erkannte der belgische Mathematiker **Adolphe Quetelet** (1796–1874) als einer der ersten Wissenschaftler die grundsätzliche Bedeutung der Statistik für die Beschreibung und Analyse von physischen Erscheinungen und sozialen Phänomen [12]. Freilich war dieser Ansatz zu seiner Zeit nicht unumstritten. Heute gilt Quetelet als Begründer der Sozialstatistik.

Einige frühe Anwendungen der numerischen Methode im Bereich der Medizin stellten eindrucksvoll deren Nutzen unter Beweis: Der Landarzt **Edward Jenner** (1749–1823) verifizierte statistisch die prophylaktische Wirkung der Kuhpockenimpfung. Der Sozialreformer **Edwin Chadwick** (1800–1890) gab der Hygienebewegung wichtige Impulse. Seine Daten gründeten sich auf statistische Analysen von **William Farr** (1807–1883), der Berichte über Todesursachen in England publiziert hatte. Als der Begründer der klinischen Statistik gilt **Pierre Charles Alexandre Louis** (1787–1872), der eine naturwissenschaftlich orientierte Medizin vertrat [2]. Er überprüfte die Wirkung des Aderlasses und wies – nachdem diese Methode jahrhundertelang angewandt worden war – mittels statistischer Analysen nach, dass dieses Mittel nutzlos oder gar schädlich war [1]. **John Snow** (1813–1858) entdeckte, dass das Cholerarisiko in London mit der Qualität des Trinkwassers zusammenhing. Der Gynäkologe **Ignaz Philipp Semmelweis** (1818–1865) wies in einer Wiener Geburtsklinik nach, dass Kindbettfieber durch mangelnde Hygiene verursacht wurde. Die Mortalität unter den Wöchnerinnen sank drastisch, nachdem er hygienische Maßnahmen angeordnet hatte. Der Augustinermönch **Gregor Johann Mendel** (1822–1884) verifizierte die von ihm aufgestellten Vererbungsgesetze mit statistischen Methoden [14].

Trotz dieser beeindruckenden Erfolge reagierten viele der den Traditionen verhafteten Ärzte ablehnend oder verständnislos auf die Anwendung der numerischen Methode in der Medizin. Ihre Argumente waren nicht von der Hand zu weisen: Jeder Mensch ist ein Individuum; jeder Krankheitsverlauf ist einzigartig; der Vergleich eines Messwerts mit einem Mittelwert sei daher sinnlos; von einem Arzt erwarte man Sicherheit und keine Wahrscheinlichkeiten. Der Pariser Kliniker **Ar-**

mand Trousseau (1801–1867) argumentierte gar: „Diese Methode ist die Geißel der Intelligenz … Sie degradiert den Arzt zum Buchhalter.“ [13] Außerdem hielt man es vielfach für unmöglich, dass sich aufgrund von Wahrscheinlichkeiten oder Mittelwerten generalisierende Aussagen herleiten ließen. Schließlich besann man sich auf eine Methode zur Erkenntnisgewinnung, die bereits Jahrhunderte zuvor der englische Philosoph **Francis Bacon** (1561–1626) propagiert hatte: Sie beinhaltete die Beobachtung zahlreicher Einzelfälle, die Aufzeichnung der erhobenen Daten und deren rechnerische Auswertung. Dieses Vorgehen ist ein empirischer Ansatz, der objektive Erkenntnisse vermittelt, die jedoch vom Zufall beeinflusst sind. So begann allmählich die Statistik Einzug in die Medizin zu halten.

Tatsächlich sollte es noch bis weit ins 20. Jahrhundert dauern, ehe statistische Techniken ersonnen wurden, die es ermöglichen, aufgrund einer überschaubaren Stichprobe Aussagen bezüglich einer weitaus größeren Grundgesamtheit herzuleiten. Bekannte Wissenschaftler, die zu diesem Anwendungsgebiet beigetragen haben, sind **Karl Pearson** (1857–1936), der die Korrelations- und Regressionsanalyse vorantrieb, oder **Sir Ronald Aylmer Fisher** (1890–1962). Nicht zuletzt hat das Aufkommen leistungsfähiger Rechner und benutzerfreundlicher Software dazu beigetragen, dass komplexe statistische Methoden in den unterschiedlichsten Disziplinen breite Anwendung finden: Mediziner evaluieren damit die Wirkung und Sicherheit neu entwickelter Therapien, Versicherungsfachleute erarbeiten Risikoprofile, Meteorologen erstellen Prognosen. Die Anwendung der Statistik in ganz unterschiedlichen Disziplinen unterstreicht die Bedeutung dieses Fachs.

3 Der Umgang mit Wahrscheinlichkeiten und Risiken

„Statistisches Denken wird eines Tages für mündige Staatsbürger ebenso wichtig sein wie die Fähigkeit, zu lesen und zu schreiben.“ Diese Aussage des englischen Science-Fiction-Autors **Herbert George Wells** (1866–1946) drückt aus, dass das Vermitteln und das Anwenden statistischer Denkweisen nach wie vor auf große Schwierigkeiten stößt. Einerseits sind Daten und Informationen für jedermann quasi rund um die Uhr zugänglich. Andererseits wird das Bild einer modernen Gesellschaft immer mehr von Unsicherheiten und Risiken geprägt. Offensichtlich fällt es vielen Zeitgenossen schwer, sich angesichts der Masse der verfügbaren Informationen und der öffentlich geäußerten, teilweise widersprüchlichen Expertenmeinungen zurechtzufinden.

Der Zufall ist unser ständiger Begleiter. Dennoch ist selbst für Experten der Umgang mit Wahrscheinlichkeiten mitunter schwierig [4]. Dies soll am Beispiel des Mammographie-Screenings dargelegt werden. Diese Maßnahme hat zum Ziel, ein Karzinom möglichst früh zu erkennen, um durch rechtzeitige Therapie die Progression der Krankheit zu verhindern. Dies klingt so überzeugend, dass sich eine kritische Reflexion scheinbar erübrigt. Kritiker wenden dagegen ein, dass der Nutzen des Mammographie-Screenings wesentlich geringer sei als vielfach angenommen. Falsch positive Befunde, die keineswegs ausgeschlossen sind, gehen mit physischen und psychischen Belastungen für die betreffenden Frauen einher und führen zu weiteren diagnostischen und therapeutischen Maßnahmen, die unnötig sind, Risiken in sich bergen und das Gesundheitssystem belasten. Wie soll sich nun ein Laie in dieser kontroversen Diskussion eine eigene Meinung bilden? Ganz einfach: Lasst Zahlen sprechen!

In der Gruppe der Frauen, die sich einem Screening unterziehen, ist die Prävalenz gering: Nur vier von tausend haben ein Karzinom. Unter der Annahme einer Sensitivität von 90% und einer Spezifität von 95% erwartet man theoretisch die in Tabelle 1 (links) dargelegten Häufigkeiten. Demnach sind nur 36 von 534 positiven Befunden mit einer Erkrankung assoziiert; das entspricht einem recht geringen positiven Vorhersagewert von 6,7%. Der überwiegende Anteil der positiven Befunde ist also nicht korrekt. Bei Vorliegen eines negativen Befundes kann man sich dagegen mit einer sehr hohen Wahrscheinlichkeit von 99,96% (9462/9466) darauf verlassen, dass tatsächlich ein Karzinom ausgeschlossen werden kann.

Tabelle 1: Zu erwartende Häufigkeiten bei der Mammographie bei einer Prävalenz von 0,4% (links) und einer Prävalenz von 10%.

	Prävalenz 0,4%			Prävalenz 10%		
	mit Karzinom	ohne Karzinom	Summe	mit Karzinom	ohne Karzinom	Summe
Befund positiv	36	498	534	900	450	1.350
Befund negativ	4	9.462	9.466	100	8.550	8.650
Summe	40	9.960	10.000	1.000	9.000	10.000

Anders stellt sich die Situation dar, wenn die Mammographie als diagnostische Methode in einer Hochrisikogruppe mit einer Prävalenz von 10% angewandt wird. Mit analogen Rechnungen ergibt sich hier ein positiver Vorhersagewert von 66,7% und ein negativer Vorhersagewert von 98,8% (Tabelle 1, rechts). Aus diesen Überlegungen geht hervor: Die Vorhersagewerte sind offensichtlich von der Prävalenz abhängig. Bei geringer Prävalenz ist der positive Vorhersagewert mitunter extrem niedrig. Ein positiver Befund ist daher keinesfalls als sichere Diagnose anzusehen. Dessen sollte man sich bewusst sein, ehe man sich für oder gegen ein Screening entscheidet.

Viele Menschen zeigen sich verwundert, wenn sie mit diesen Ergebnissen konfrontiert werden. Ähnliche Überlegungen lassen sich anstellen bei HIV-Tests, bei forensischen Methoden (die in der Rechtsprechung eingesetzt werden) oder bei Feuermeldern: Wenn die Prävalenz gering ist, sind die meisten Alarme falsch. Theoretisch liegt diesen Berechnungen der **Satz von Bayes** zugrunde. **Thomas Bayes** (1701–1761) war ein englischer Geistlicher, der den nach ihm benannten Satz entwickelte, mit dem bedingte Wahrscheinlichkeiten berechnet werden. Diese Formel muss man jedoch nicht explizit anwenden. Es genügt, wenn man – wie oben gezeigt – eine Vierfeldertafel erstellt und Grundrechenarten anwendet.

Die wichtigste Frage lautet indessen: „Wie häufig wird die Progression oder der Tod durch Brustkrebs tatsächlich verhindert, wenn eine Frau regelmäßig an einem Mammographie-Screening teilnimmt?“ In einer randomisierten epidemiologischen Studie, an der 500.000 Frauen teilnahmen, konnte nachgewiesen werden: Bei Frauen, die sich regelmäßig einem Screening unterziehen, beträgt die brustkrebsbedingte Mortalität bezogen auf einen Zeitraum von zehn Jahren 2,9 Promille. In der Kontrollgruppe liegt dieser Anteil bei 3,6 Promille [6]. Diese Zahlen zeigen zwar, dass bei Frauen, die regelmäßig gescreent werden, die Mortalität geringer ist. Andererseits decken sie auf, dass Screening den brustkrebsbedingten Tod nicht zuverlässig verhindern kann.

Bei vielen Studien der klinischen oder der epidemiologischen Forschung stehen Risiken wie die Inzidenz, die Mortalität oder die Nebenwirkungsrate im Fokus des Interesses. Risiken lassen sich auf unterschiedliche Weise darstellen. Im genannten Beispiel lässt sich das Ergebnis auch wie folgt beschreiben: Das Risiko, in einem Zeitraum von zehn Jahren an Brustkrebs zu sterben ist für Frauen, die *nicht gescreent* werden, um 24% höher als für Frauen, die *gescreent* werden (3,6/2,9). Oder: Eine Frau, die sich *nicht screenen* lässt, könnte ihr krankheitsbedingtes

Risiko um 19% senken (0,7/3,6). Diese Zahlen hören sich wesentlich dramatischer an als die Promilleanteile. Schließlich lässt sich noch die *Number Needed to Treat* (oder die *Number Needed to Screen*) angeben. Sie ergibt sich aus dem Kehrwert der absoluten Risikodifferenz $1/0{,}0007 \approx 1429$. Diese Zahl besagt: Wenn sich 1429 Frauen regelmäßig über die Dauer von zehn Jahren alle zwei Jahre *screenen* lassen, profitiert nur eine davon (dergestalt, dass sie aufgrund des Screenings vom Tod durch Brustkrebs bewahrt wird). Andererseits lässt sich nachrechnen: Eine Frau ohne Mammakarzinom muss (bei einer Spezifität von 95%) mit einer Wahrscheinlichkeit von etwa 23% ($1 - 0{,}95^5$) damit rechnen, dass sie bei fünf Untersuchungen mindestens einmal mit einem falsch positiven Befund belastet wird [10]. All diese Wahrscheinlichkeiten sprechen per se nicht für oder gegen das Mammographie-Screnning. Sie zeigen jedoch, dass ein Testbefund keine sichere Diagnose darstellt, dass Screening-Untersuchungen auch mit Unannehmlichkeiten verbunden sein können, und dass bei Weitem nicht alle Frauen profitieren.

Eine weitere Unsitte im Umgang mit Risiken besteht in sinnlosen Vergleichen. Dazu ein Beispiel: Zu Kaiser Wilhelms Zeiten starben in Deutschland etwa 5% der Menschen an Krebs, heute ist dies die Todesursache bei jedem vierten Einwohner. Das liegt jedoch weder am Monarchen noch an der vermeintlich geringeren Umweltverschmutzung zu Beginn des 20. Jahrhunderts (wie einem manche Leute weismachen wollen), sondern an den erfreulichen Tatsachen, dass typische Infektionskrankheiten der damaligen Zeit (z. B. Tuberkulose) verschwunden sind und dass die durchschnittliche Lebenserwartung massiv gestiegen ist [7].

Wieso sind dennoch viele Menschen bereit, Empfehlungen von Fachleuten blind zu vertrauen oder absurde Thesen kritiklos hinzunehmen? Wieso meinen sie, dass eine ärztliche Diagnose oder ein richterliches Urteil in jedem Fall korrekt ist, oder dass Vorsorgemaßnahmen das Entstehen einer schweren Erkrankung zuverlässig verhindern? Man glaubt an die Unfehlbarkeit der Apparatemedizin und gibt sich allzu gerne der Illusion einer trügerischen Gewissheit hin. Die erwähnten Beispiele zeigen jedoch: Ein unangemessener Umgang mit Wahrscheinlichkeiten kann fatale Konsequenzen nach sich ziehen. Adäquate Therapien und gerechte Urteile setzen voraus, dass Ärzte, Richter und andere Betroffene lernen, mit Wahrscheinlichkeiten umgehen. Besondere Vorsicht ist geboten bei relativen Risiken, die oft dramatischer scheinen als sie in Wirklichkeit sind. Kritisches Denken ist Pflicht! Es lohnt sich mitunter, altüberlieferte Dogmen in Frage zu stellen (Stichwort: Aderlass). Einfache statistische Überlegungen können helfen, gegen

mangelndes Zahlenverständnis anzugehen und manche polemische Diskussion oder apokalyptische Prognose zu versachlichen.

4 Reduktion auf das Wesentliche: Statistische Kenngrößen

Kaum ein Bereich der klinischen oder wissenschaftlichen Medizin ist ohne quantitative Messungen denkbar. Pathologische und physiologische Erscheinungen werden gemessen, Summenscores oder andere Parameter werden daraus abgeleitet, um Krankheitszustände zu erfassen, die Wirksamkeit einer Therapie zu überprüfen oder um eine Prognose zu stellen. Quantitative Messungen waren erst möglich geworden durch die Übernahme naturwissenschaftlicher Methoden in die Medizin seit Beginn der Renaissance. Vorher waren Ärzte bei der Behandlung ihrer Patienten auf ihre Sinnesorgane angewiesen, mit denen sie nur qualitative Befunde erheben konnten. Ihre Eindrücke waren naturgemäß subjektiv; die daraus abgeleiteten Schlussfolgerungen häufig spekulativ.

Die Einführung quantitativer Messmethoden erforderte ein Umdenken in der Ärzteschaft. Man unterschied fortan nicht mehr nur zwischen zwei Kategorien wie beispielsweise „krank“ und „gesund“, sondern erkannte, dass der Übergang von Gesundheit zu Krankheit kontinuierlich ist. Mediziner suchten nach präzisen Messverfahren, um physiologische Parameter, Laborwerte etc. zu erfassen (beispielsweise Fieberthermometer oder Blutdruckmessgeräte). Allerdings ergab sich nun ein Problem besonderer Art: Ein einzelner Messwert lässt sich wesentlich schwieriger beurteilen als ein qualitativer Befund. Um abzuwägen, ob und in welchem Grad ein Messwert „normal“ oder „abnormal“ ist, bedarf es Kenngrößen wie Durchschnittswerte, Minima und Maxima, mit denen Einzelwerte verglichen werden. Man benötigt eine „Theorie des Messfehlers“.

Der Vergleich eines einzelnen Messwertes mit einem *Mittelwert* ist ungemein nützlich im Alltag, bei wissenschaftlichen Untersuchungen oder auch im Rahmen eines diagnostischen oder therapeutischen Prozesses. Er erlaubt eine grobe Abschätzung, ob ein Mess- oder Zählwert unterhalb oder oberhalb des Durchschnitts liegt. Dies hat bereits der oben erwähnte **Quetelet** erkannt, der die Meinung vertrat, dass für alle messbaren menschlichen Eigenschaften ein Durchschnittswert existiert, den er als Idealwert ansah. Diesem Herrn haben wir übrigens auch den Body Mass Index zu verdanken. Auch dies kann als Hinweis aufgefasst werden,

dass Quetelet von der Nützlichkeit quantitativer Messungen zutiefst überzeugt war, und dass seine Ideen sich letztlich durchgesetzt haben.

Obgleich diese Messtechniken spätestens seit Beginn des 20. Jahrhunderts im klinischen Alltag und auch in der experimentellen Forschung routinemäßig eingesetzt werden, gestaltete sich die Festlegung und Akzeptanz von Durchschnittswerten als schwierig [8]. Dieses Konzept setzte sich nur langsam durch. Allmählich wurden Durchschnitts- oder Mittelwerte als „Normwerte“ oder „Standardwerte“ angesehen und in umfangreichen Tabellenwerken dargelegt.

Eine Datenreihe wird jedoch nicht allein durch ihre zentrale Lage, sondern auch durch ihre Variabilität charakterisiert. Dies ist spätestens seit den Zeiten des englischen Statistikers **Sir Ronald Aylmer Fisher** bekannt. Eine dafür geeignete Maßzahl ist die *Standardabweichung s*. Dieses Streuungsmaß gestattet zusammen mit dem Mittelwert *m* eine präzisere Beurteilung eines Einzelwertes. Dazu muss man lediglich wissen: Bei normalverteilten Daten liegen etwa 2/3 aller Werte innerhalb des Intervalls $m \pm s$, etwa 95% liegen im Intervall $m \pm 2s$. Wenn nichts über die Verteilungsform bekannt ist, kann man zugrunde legen, dass 75% bzw. 8/9 aller Werte innerhalb der Intervalle $m \pm 2s$ bzw. $m \pm 3s$ liegen. Man braucht also nur zwei statistische Kenngrößen, um bei einem einzelnen Messwert einschätzen zu können, ob er innerhalb eines Normbereichs liegt oder ob es sich um einen extremen Wert handelt, der nicht zur Datenreihe zu passen scheint und den man genauer in Augenschein nehmen sollte.

Allerdings stößt der Mittelwert und dessen Interpretation häufig auf Irritationen. Dies hat mehrere Gründe. Der Mittelwert stimmt nicht immer mit einem realen Wert überein. So besagt etwa die Fertilitätsrate 1,4 (bezogen auf das Jahr 2015 in Deutschland), dass eine „Durchschnittsfrau“ im Laufe ihres Lebens 1,4 Kinder zu Welt bringt – obwohl dieser Wert für keine einzige Frau zutrifft. Ein anderes Manko: Bei ordinal skalierten Merkmalen könnte der Mittelwert zu unsinnigen Schlussfolgerungen verleiten. Wenn beispielsweise der Therapieerfolg im Rahmen einer klinischen Studie mit Werten zwischen 1 (vollständige Heilung) und 5 (Patient verstorben) beurteilt wird, widerstrebt es dem gesunden Menschenverstand, den „mittleren Therapieerfolg“ zu berechnen. Außerdem kann der Mittelwert insbesondere bei einem geringen Stichprobenumfang stark von Ausreißern beeinflusst werden; in diesen Fällen würde er ein verzerrtes Bild der tatsächlichen Verhältnisse wiedergeben.

Sinnvoller wäre in diesen Fällen die Angabe des *Medians*. Dieser ist zwar weniger bekannt als der Mittelwert, wenngleich nicht weniger informativ und einfach zu verstehen. Der Median teilt eine Stichprobe in zwei gleich große Subgruppen: Die Hälfte aller Daten ist maximal so groß wie der Median, die andere Hälfte mindestens so groß. Dieses Lagemaß ist unempfindlich gegen Ausreißer. Außerdem stimmt der Median (bei einem ungeraden Stichprobenumfang) mit einem tatsächlichen Wert überein. Bei einem geraden Stichprobenumfang wird der Median als der Durchschnitt der beiden mittleren Werte der Datenreihe berechnet. Ein Vergleich zwischen Mittelwert und Median gibt Auskunft über die Form der zugrunde liegenden Verteilung: Bei symmetrischen Verteilungen stimmen die beiden Lagemaße theoretisch überein. Wenn der Mittelwert wesentlich größer (oder kleiner) ist als der Median, ist dies ein Hinweis auf eine schiefe Verteilung. Ist es nicht genial, dass sich eine umfangreiche Datenreihe durch ein paar wenige, aussagekräftige Kenngrößen charakterisieren lässt?

Im Rahmen einer Studie werden jedoch nicht nur einzelne Merkmale beschrieben, sondern auch Zusammenhänge untersucht. Typische Fragen lauten beispielsweise: Ist die Wirkung eines blutdrucksenkenden Medikaments abhängig vom Alter der Patienten oder vom Basiswert, der vor Therapiebeginn gemessen wurde? Um die Stärke des Zusammenhangs zwischen zwei quantitativen Merkmalen zu quantifizieren, eignet sich der *Korrelationskoeffizient* nach Karl Pearson. Dieser hätte sich zu Beginn des 20. Jahrhunderts bestimmt nicht träumen lassen, dass der nach ihm benannte Koeffizient einmal so populär werden würde! Dies ist verständlich, wenn man sich dessen Eigenschaften vor Augen hält. Der Koeffizient ist skaleninvariant, eine dimensionslose Zahl zwischen -1 und +1 und einfach zu interpretieren: Ein positiver Koeffizient steht für einen gleichsinnigen, ein negativer für einen gegensinnigen Zusammenhang. Koeffizienten mit einem Zahlenwert nahe bei 0 zeigen, dass der Zusammenhang schwach ist. Je näher der Wert bei +1 oder -1, desto stärker ist der Zusammenhang.

Schließlich sei noch ein weiteres Beispiel für den Gebrauch eines statistischen Koeffizienten erwähnt: Wenn im klinischen Alltag ein neues Messverfahren eingeführt wird, muss vorab dessen Tauglichkeit evaluiert werden. Die wichtigsten Gütekriterien hierfür sind die Reliabilität und die Validität. Die Reliabilität bezeichnet die Verlässlichkeit des Messinstruments: Dieses gilt als reliabel, wenn mehrere Messungen an demselben Objekt, die unter unterschiedlichen Bedingungen durchgeführt werden, zum gleichen Ergebnis führen. So sollten etwa zwei

Ärzte, die unabhängig voneinander dieselben Patienten befunden, zu gleichen Resultaten gelangen. Die Validität steht für die Richtigkeit einer Messmethode. Sie wird ermittelt durch den Vergleich eines Messinstruments mit einem Goldstandard. Zur Bestimmung dieser Kriterien eignen sich *Intraklassenkorrelationskoeffizienten* und *Kappa-Koeffizienten*, die den Grad der Übereinstimmung bei quantitativen bzw. qualitativen Messverfahren quantifizieren. Ein Kappa-Koeffizient bzw. ein Intraklassenkorrelationskoeffizient mit dem Wert 1 steht für perfekte Übereinstimmung, der Wert 0 für eine unbrauchbare Messmethode.

Statistische Koeffizienten mit einem Wertebereich zwischen 0 und 1 sind außerordentlich beliebt, weil sie sich einfach und ohne Zusatzinformationen interpretieren lassen. Dazu bedarf es keinerlei statistischer oder sonstiger Spezialkenntnisse, was manche Anwender durchaus zu schätzen wissen (die mitunter komplizierten mathematischen Algorithmen, die sich dahinter verbergen, brauchen sie nicht unbedingt zu interessieren). Auch dieses Phänomen lässt sich mit Schönheit und Eleganz in Verbindung bringen: Dank findiger Mathematiker lässt sich ein hoch komplexer Sachverhalt wie beispielsweise die Reliabilität eines Messverfahrens in genial einfacher Weise durch eine Zahl zwischen 0 und 1 beschreiben.

5 Das Geheimnis von Schätzwerten

Schätzungen bergen etwas Geheimnisvolles in sich. Die Kommunikationswissenschaftlerin und Wahlforscherin **Elisabeth Noelle-Neumann** (1916–2010) formulierte dies einst mit folgenden Worten: „Es ist mir noch heute rätselhaft, dass man herausbringt, was sechzig Millionen Menschen denken, wenn man zweitausend Menschen befragt. Erklären kann ich das nicht. Es ist eben so."

Zwar werden Politiker vor Wahlen nicht müde zu betonen, dass Wahlumfragen auf Schätzungen basieren und keineswegs das endgültige Wahlergebnis widerspiegeln. Das liegt jedoch nicht unbedingt an den Schätzverfahren, sondern daran, dass sich in den Tagen vor einer Wahl eventuell unvorhergesehene Geschehnisse ereignen, die das Wahlergebnis beeinflussen könnten. Am Wahlabend wundert sich kaum jemand, wenn bereits wenige Minuten nach Schließung der Wahllokale erste Prognosen über den Bildschirm flattern, die erstaunlich präzise das endgültige Wahlergebnis vorhersagen. Diese Schätzungen beruhen in der Regel auf repräsentativen Stichprobenerhebungen am Tag der Wahl und sind deshalb meist zuverlässig.

In vielen Situationen haben Schätzungen entscheidende Vorteile. Um beispielsweise den Alkoholgehalt im Blut eines Verkehrsteilnehmers zu bestimmen, reicht eine minimale „Stichprobe“ (im wahrsten Sinne des Wortes) aus. Um einen Eindruck von der wirtschaftlichen Situation einer Population zu gewinnen, reicht ein Mikrozensus, bei dem nur ein minimaler Anteil befragt wird. Um die Effektivität einer Therapie zu evaluieren, genügt eine Studie mit einer begrenzten Anzahl von Patienten (es ist keineswegs notwendig und ohnedies nicht durchführbar, *alle* Patienten, die von der Therapie profitieren könnten, zu untersuchen). Schätzungen sind allerdings nur sinnvoll, wenn die Stichprobe repräsentativ ist (das heißt: Jedes Element der Grundgesamt sollte die gleiche Chance haben, in die Stichprobe zu gelangen) und wenn die Fallzahl hinreichend hoch ist. Schätzungen aus nichtrepräsentativen Stichproben können komplett in die Irre führen. Ein Beispiel stellt die Wahl des US-amerikanischen Präsidenten im Jahr 1936 dar. Weil bei der Wahlumfrage hauptsächlich Bürger mit Telefonanschluss berücksichtigt wurden, wich das prognostizierte Wahlergebnis eklatant vom tatsächlichen Ergebnis ab.

Anwender aus empirischen Wissenschaften, die in der Regel auf Stichproben überschaubarer Größe zurückgreifen, ermitteln eine Kenngröße (z. B. einen Mittelwert) in der Hoffnung, einen zuverlässigen Schätzwert zu erhalten, der sich in der näheren Umgebung des „wahren“ Parameters befindet. Generell gilt: Je höher die zugrundeliegende Fallzahl, desto verlässlicher ist die Schätzung. Dies sagt einem der gesunde Menschenverstand; eine genauere Begründung liefert das *Gesetz der Großen Zahlen*. Indessen sieht man einem Mittelwert nicht an, ob er aus 3, 20 oder 100 Einzelwerten berechnet wurde. Deshalb sind Punktschätzungen in gewisser Weise unbefriedigend.

Ein Ausweg findet sich in der Konstruktion eines *Konfidenzintervalls*. Dieses wird nach einem mathematischen Algorithmus gebildet, der mit einer Wahrscheinlichkeit von üblicherweise 95% gewährleistet, dass das Intervall den wahren (aber unbekannten) Parameter einschließt. Etwas vereinfacht formuliert lässt sich sagen: Alle Werte innerhalb des Konfidenzintervalls erscheinen für den zu schätzenden Parameter plausibel zu sein. Dabei wird eine hohe Fallzahl belohnt: Damit ergibt sich ein schmaleres Intervall (was eine präzisere Schätzung bedeutet) als bei einer geringen Fallzahl.

Mit einer leistungsfähigen Statistiksoftware lassen sich Konfidenzintervalle für nahezu alle gängigen Kenngrößen wie Mittelwerte, Wahrscheinlichkeiten, Korrelationskoeffizienten, Odds Ratios etc. konstruieren. Für den Anwender ist es

nicht notwendig, die teilweise sehr komplexen mathematischen Grundlagen zu kennen, auf denen diese Berechnungen basieren. Er sollte allerdings in der Lage sein, ein Konfidenzintervall sinnvoll zu interpretieren. Das Intervall zeigt auf einen Blick, ob die Schätzung präzise oder unbrauchbar ist. Für den Anwender ist es in jedem Fall hilfreich, dies zu wissen.

6 Effizienz statistischer Verfahren

Der Fortschritt in einer empirischen Wissenschaft beruht im Wesentlichen auf Beobachtungen und Erfahrungen. Ein Kliniker ersinnt möglicherweise aufgrund der Kenntnis der Krankheitsmechanismen eine neue Therapie, von der er annimmt, dass sie der bislang verwendeten Standardtherapie überlegen ist, oder er glaubt nach der Beobachtung zahlreicher Patienten, eine Assoziation zwischen einer Krankheit und einem Risikofaktor entdeckt zu haben.

Anders als in der reinen Mathematik lassen sich derlei Vermutungen nicht exakt beweisen. Ein Wissenschaftler mag zwar von der Richtigkeit seiner Hypothesen überzeugt sein und Argumente dafür anführen – das allein kann jedoch kein objektives Kriterium darstellen. Außerdem ist es wegen der interindividuellen Variabilität von Patienten schier unmöglich, allgemein gültige Aussagen zu treffen, die in jedem Einzelfall zutreffen. Eine Therapie mag bei einem Patienten Wunder vollbringen, während sie bei einem anderen keine Wirkung zeigt. Auch wenn eine Assoziation etwa zwischen Rauchen und einer Erkrankung besteht, gibt es Raucher, die von der Krankheit verschont bleiben, und Nichtraucher, die erkranken.

Eine der größten Errungenschaften der induktiven Statistik besteht darin, dass sich Hypothesen, die in einer empirischen Wissenschaft aufgestellt werden, objektiv überprüfen und absichern lassen. Mit einem *Hypothesentest* lassen sich Unterschiede zwischen zwei oder mehreren Therapieformen nachweisen oder Zusammenhänge zwischen Merkmalen absichern – zwar nicht für jeden Einzelfall, aber für eine Population. Ein Anwendungsbeispiel mag dies verdeutlichen: Es werden zwei medikamentöse Therapien zur Senkung des systolischen Blutdrucks verglichen. Bei manchen Patienten ist eine starke Wirkung zu verzeichnen, bei anderen ist die Wirkung kaum nachweisbar; es gibt auch Patienten, bei denen der Blutdruck steigt. Wie ist diese Variabilität zu erklären? Liegt es an der Therapieform? Spielt eventuell das Alter der Patienten eine Rolle? Oder deren Geschlecht? Oder gibt es andere patientenspezifische Merkmale, die die Wirkung der Therapie

beeinflussen – etwa der Body Mass Index, bestimmte Verhaltensweisen (zum Beispiel Rauchen), Komorbiditäten oder genetische Dispositionen?

Um diesen Fragen auf den Grund zu gehen, benötigt man relevante Daten und statistische Tests. Der anzuwendende Test ist abhängig von der Fragestellung und der Art der Daten: Zum Vergleich der Mittelwerte zweier Gruppen wird häufig ein t-Test verwendet, zum Vergleich von relativen Häufigkeiten wird ein Chi^2-Test oder Fishers exakter Test herangezogen. Aus jedem Test resultiert der sogenannte p-Wert, der als Irrtumswahrscheinlichkeit interpretiert werden kann. Dieser quantifiziert die Wahrscheinlichkeit dafür, dass der gefundene Effekt (Unterschied oder Zusammenhang) nur zufällig zustande kommt. Nach alter Konvention gilt ein Testergebnis als „statistisch signifikant", falls der p-Wert unter dem Signifikanzniveau von 0,05 liegt. Auf diese Weise wird der Zufall zwar nicht eliminiert, aber unter Kontrolle gebracht.

Kleine p-Werte bereiten vielen Anwendern großes Vergnügen: Mit einem signifikanten Testergebnis können sie sich rühmen, einen Effekt abgesichert zu haben und dürfen hoffen, dass dieses Ergebnis den Wert ihrer wissenschaftlichen Arbeit erhöht oder eine Publikation nach sich zieht.

Allerdings sollte bei aller Freude bedacht werden: Ein p-Wert (wie klein er auch sein mag) darf nicht als Beweis für die Richtigkeit einer Hypothese verstanden werden. Er beinhaltet nämlich keinerlei Information, wie ein Effekt zustande gekommen ist und welche Konsequenzen aus diesem Ergebnis zu ziehen sind. Außerdem ist zu berücksichtigen, dass der p-Wert sehr stark von der Fallzahl beeinflusst wird: Eine große Fallzahl führt leicht zu einem signifikanten Ergebnis, während mit einer geringen Fallzahl schwerlich ein Unterschied nachweisbar ist. Deshalb muss im Einzelfall immer beurteilt werden (am besten mit Hilfe eines Konfidenzintervalls), ob und inwieweit die Größenordnung des gefundenen Effekts relevant ist. Umgekehrt gilt auch: Ein p-Wert über 0,05 bedeutet nicht zwangsläufig, dass es generell keinen Unterschied (oder keinen Zusammenhang) gibt. Er besagt lediglich, dass anhand des vorhandenen Datenmaterials kein Effekt nachweisbar ist.

Da sich ein signifikantes Ergebnis leichter publizieren lässt als ein nicht signifikantes, tendieren manche Anwender dazu, einen kleinen p-Wert quasi gewaltsam herbeizuführen (für diese unlautere Vorgehensweise hat sich der englische Begriff „p-Hacking" etabliert). Dazu einige Vorschläge (die ausdrücklich *nicht* zur Nachahmung empfohlen werden): Unliebsame Daten werden ohne vernünftigen

Grund von der Analyse ausgeschlossen oder nachträglich in unzulässiger Weise korrigiert; man wendet für dieselbe Fragestellung mehrere Tests an und wählt dann denjenigen aus, der das „passendste“ Ergebnis liefert; es wird eine Vielzahl von Zielgrößen untersucht (dann steigt die Wahrscheinlichkeit, wenigstens ein signifikantes Ergebnis zu finden); in einer longitudinalen Studie mit mehreren Messzeitpunkten werden die Kenngrößen aller Zeitpunkte paarweise miteinander verglichen (dto.). Auch wenn mit derlei Tricks ein signifikantes Ergebnis oder gar eine Publikation zustande kommen mag: Die wissenschaftliche Arbeit wäre wertlos (auch wenn das nicht sofort für jedermann ersichtlich wird). Der Leser einer Publikation, der sich mit statistischen Analysemethoden ein wenig auskennt, hat zumindest den Vorteil, dass er kritische Fragen stellen kann und weiß, worauf er bei der Lektüre zu achten hat.

Nun ist eine Zielgröße (etwa die Wirkung einer Therapie) nicht nur von einer, sondern von diversen Einflussgrößen abhängig. Es bietet sich an, für jede potenzielle Einflussvariable einen Test durchzuführen und anhand der p-Werte zu beurteilen, welche Variable einen starken, einen schwachen oder keinen Einfluss auf die Zielgröße haben. Sinnvoll ist es darüber hinaus, mittels multipler Analysen mehrere Einflussgrößen simultan zu analysieren. Dank leistungsfähiger Rechner und benutzerfreundlicher Software sind diese komplexen Methoden sehr flexibel einsetzbar. Mit *Varianzanalysen* wird der Einfluss von einem oder mehreren qualitativen Faktoren auf eine quantitative Zielgröße geprüft; außerdem lassen sich Interaktionen (Wechselwirkungen zwischen zwei Faktoren) untersuchen. *Lineare Regressionsanalysen* setzen ebenfalls eine quantitative Zielgröße voraus, wobei diverse Einflussgrößen (quantitative und qualitative) berücksichtigt werden können. Bei binären Zielgrößen wird eine *logistische Regression* bevorzugt, bei Zähldaten bietet sich eine *Poisson-Regression* und bei Überlebenszeiten (die auch „zensiert“ sein dürfen, d.h. Zeiten sind nicht vollständig bekannt) eine *Cox-Regression* an. Bei der *logistischen Regression* wird eine Wahrscheinlichkeit modelliert (z. B. für das Auftreten einer Nebenwirkung), bei der Cox-Regression eine Hazard Ratio.

Aus einer Regressionsanalyse resultiert eine mathematische Gleichung, die die Zielgröße in mathematischer Formelsprache beschreibt. Sie beinhaltet in der Regel Einflussgrößen, die (in Kombination mit anderen Faktoren) mit der Zielgröße signifikant assoziiert sind. Damit kann man ausrechnen: Wie wird die Zielgröße beeinflusst, wenn sich der Wert einer oder mehrerer Einflussgrößen ändert? Was würde passieren, wenn derselbe Patient einer anderen Gruppe angehörte?

Multiple Modelle bieten sich in der klinischen und der epidemiologischen Forschung an, da ein Untersuchungsgegenstand im Allgemeinen nicht monokausal erklärt werden kann, sondern von diversen Faktoren beeinflusst ist (die sich wiederum gegenseitig beeinflussen oder zusammenwirken). Die Vorteile leuchten ein: Durch die Berücksichtigung mehrerer Variablen kann eine Diagnose oder eine Prognose präziser gestellt werden; individuelle Charakteristika der Studienteilnehmer lassen sich angemessen berücksichtigen. Außerdem kann nach möglichen *Confoundern* adjustiert werden (was insbesondere bei nicht randomisierten Studien immens wichtig ist).

Es soll nicht verschwiegen werden, dass die Konstruktion eines statistischen Modells den Anwender mitunter vor große Herausforderungen stellt. Zunächst muss er prüfen, ob die Voraussetzungen zumindest annährend erfüllt sind (ansonsten wäre die Modellierung wenig sinnvoll). Dann gilt es, geeignete Einflussgrößen für das finale Modell auszuwählen. Diese Wahl sollte unter inhaltlichen und statistischen Aspekten getroffen werden.

Freilich kann eine leistungsstarke Statistiksoftware bei der Wahl der Einflussvariablen für das finale Modell behilflich sein. Zum Vergleich mehrerer Modelle eignen sich *Gütemaße*. Bei linearen Regressionsanalysen bietet sich das Bestimmtheitsmaß R^2 an, das Werte zwischen 0 und 1 annimmt. R^2 quantifiziert, welcher Anteil der Variabilität der Zielgröße durch das statistische Modell erklärt wird. Bei logistischen Regressionen eignet sich die AUC („area under curve“); je näher dieser Wert bei 1 liegt, desto besser ist das Modell. Außerdem existieren spezielle Informationskriterien (etwa das Akaike-Informations-Kriterium AIC), die eine objektive Bewertung bezüglich der Güte eines Modells unter Berücksichtigung von dessen Komplexität gewährleisten. Durch diese Gütemaße wird einem Anwender auf sehr subtile Weise der Wert seiner Bemühungen vor Augen geführt!

Multiple Analysentechniken eröffnen ungeahnte Möglichkeiten. Allerdings muss der Anwender bereit sein, sich mit dieser komplexen und mathematisch anspruchsvollen Materie intensiv auseinanderzusetzen.

7 Schlussfolgerungen

Worin liegt nun die Schönheit der Statistik? Für **Theoretiker** wie Carl Friedrich Gauß, Karl Pearson oder Abraham Wald (1902–1950), stellt(e) sich diese Frage nicht. Für diese Wissenschaftler ist Statistik als Teilgebiet der Mathematik eine

Kunst. Es bereitet ihnen Freude, über mathematischen Problemen zu tüfteln. Dazu benutzen sie hauptsächlich ihren Verstand; aufwendige technische Hilfsmittel sind in der Regel gar nicht notwendig. Gauß hat die Normalverteilung beschrieben und deren Bedeutung für die angewandte Statistik erkannt; außerdem hat er basierend auf der Methode der kleinsten Quadrate die Formeln für die Berechnung einer Regressionsgeraden hergeleitet. Pearson haben wir den nach ihm benannten Koeffizienten zu verdanken; dank Wald lassen sich Konfidenzintervalle für diverse Kenngrößen konstruieren. Diese Auswahl namhafter Statistiker ist freilich bei weitem nicht vollständig; man könnte noch viele Namen hinzufügen.

Zu allen Zeiten gab und gibt es Mathematiker, die sich aktueller Problemstellungen annehmen und Lösungen suchen. Diese Menschen bekommen leuchtende Augen, wenn sie den Beweis eines Theorems nachvollziehen können oder wenn sie eine neue Analysemethode ersonnen haben. Für sie stehen die Eleganz und Effizienz der Verfahren im Vordergrund – wenngleich sich die Probleme, mit denen sie sich befassen, aus der Praxis ergeben.

So hat beispielsweise der britische Chemiker und Hobbystatistiker **Sealy Gosset** (1876–1937) die t-Verteilung entwickelt, weil er vor das Problem gestellt war, anhand kleiner Stichproben Konfidenzintervalle für Mittelwerte zu bilden. Ein Beispiel aus der neueren Zeit: In der Therapieforschung gelten randomisierte Studien als Goldstandard zum Vergleich mehrerer Therapieformen, da allein dieses Studiendesign die gleichmäßige Verteilung aller bekannten und unbekannten Einflussfaktoren auf die zu vergleichenden Gruppen gewährleistet. Allerdings ist dieses Design in der klinischen Forschung häufig nicht praktikabel. Ein weiterer Nachteil liegt in der meist unzureichenden externen Validität: Wirkungsnachweise aus randomisierten Studien lassen sich nicht ohne Weiteres auf die klinische Praxis oder auf die Praxis niedergelassener Ärzte übertragen. Eine nicht-randomisierte Beobachtungsstudie (‚Pragmatic Trial') würde zwar die Gegebenheiten im klinischen Alltag realistischer abbilden; sie könnte jedoch wegen fehlender Strukturgleichheit zu verzerrten Ergebnissen führen und zu unzulässigen Schlussfolgerungen verleiten. Diese Problematik inspirierte die Mathematiker **Paul Rosenbaum** und **Donald Rubin**, eine statistische Methode zu entwickeln (bekannt unter der Bezeichnung „Propensity Score Method"), die einen validen Therapievergleich auch bei nicht-randomisierten Studien ermöglicht [11] und seither breite Anwendung findet [3, 9].

Anders stellt sich die Situation für die Gruppe der **Anwender** dar. Diese interessieren sich in der Regel weniger für die Eleganz statistischer Methoden, sondern fragen stattdessen: Welchem Zweck dienen diese Verfahren? Schönheit ist subjektiv. Ein Anwender wird beispielsweise den Zusammenhang zwischen zwei quantitativen Variablen emotionslos durch einen Korrelationskoeffizienten und eine Regressionsgerade beschreiben – ohne die genialen Ideen von Pearson und Gauß zu würdigen, denen diese Techniken zu verdanken sind. Er wird sich freuen, wenn er mit einem t-Test zum Vergleich zweier Mittelwerte einen p-Wert unter 0,05 erhält – ohne auf die Genialität der von Sealy Gosset erdachten Prüfgröße zu achten. Das ist bedauerlich.

Daten und Techniken – die Grundlagen jeder Datenanalyse – vermögen den Reiz der Statistik jedenfalls nicht zu erklären. Das eigentliche statistische Denken besteht eben nicht nur im routinemäßigen Berechnen von Kenngrößen oder im bloßen Anwenden von einfachen statistischen Tests (in der Hoffnung, ein signifikantes Ergebnis zu erhalten), sondern vielmehr im Bemühen, Licht ins Dunkel der Daten zu bringen und der Wahrheit ein Stück näher auf die Spur zu kommen. Wenn es einem Anwender gelingt, anhand von vorhandenem Datenmaterial ein adäquates statistisches Modell zu entwickeln und eine mathematische Gleichung zu generieren, die die Zusammenhänge zwischen diversen Merkmalen statistisch und inhaltlich in optimaler Weise darstellt, hat er das Gefühl, etwas Eigenes kreiert zu haben. Auch wenn die verwendeten Algorithmen in aller Regel etabliert sind, werden Anwender aufgrund ihrer persönlichen Erfahrung verschiedene Fragestellungen formulieren, diese individuell bearbeiten und am Ende zu unterschiedlichen Resultaten gelangen. So gesehen bietet die Statistik dem Anwender intellektuelle Freiräume („researcher degrees of freedom“), die er nach eigenem Empfinden nutzen kann. Die Schönheit ergibt sich aus der Erkenntnis, dass sich das ursprüngliche Datenchaos in ein harmonisches Ganzes fügt, dass aus scheinbar zusammenhanglosen Einzelfällen Regelmäßigkeiten ableitbar sind, aus denen sich überraschende Erkenntnisse ergeben, und dass der Analytiker das triumphierende Gefühl auskosten darf, den Zufall gezähmt zu haben – auch wenn die ästhetischen Kriterien eher abstrakter als visueller Natur sind. Für diesen Personenkreis ist Statistik gleichsam Kunst und Wissenschaft.

Nichtsdestotrotz hegen Anwender aus einer empirischen Wissenschaft (etwa Kliniker) zuweilen eine gewisse Skepsis gegenüber statistischen Methoden. Eine mögliche Erklärung mag darin liegen, dass die damit gewonnenen Erkenntnisse

ausschließlich auf Daten basieren – naturwissenschaftliche oder medizinische Fachkenntnisse scheinen für diese Analysen nicht erforderlich zu sein. Dies könnte das unangenehme Gefühl erzeugen, dass individuelle Erfahrungen des Klinikers nur eine untergeordnete Rolle spielen, während die Ergebnisse der Studie und die sich daraus ergebenden Konsequenzen in erster Linie von mathematischen Algorithmen bestimmt werden. Dem ist entgegenzuhalten: Zahlreiche Beispiele aus dem Bereich der medizinischen Forschung (insbesondere die Untersuchungen von Semmelweis) belegen, dass sich klinische Erfahrung und Datenanalyse erstaunlich gut ergänzen. Mit statistischen Methoden lassen sich Erkentnisse absichern – auch dann, wenn sich die Wirkmechanismen (noch) nicht auf molekularer oder zellulärer Ebene erklären lassen.

Der Einsatz statistischer Verfahren in einer empirischen Wissenschaft ist allerdings nur dann nützlich, wenn der Anwender in der Lage ist, aus der Fülle der ihm zur Verfügung stehenden Analysetechniken die geeigneten auszuwählen. Dies kann jedoch nicht ohne Weiteres angenommen werden. In der Regel haben diese Fachvertreter zwar Grundlagen der Statistik und einfache Tests in einer Vorlesung kennengelernt. Vielfach fehlen ihnen jedoch die Zeit oder das mathematische Fachwissen, die erforderlich wären, um sich Kenntnisse bezüglich komplexer statistischer Techniken anzueignen und diese effizient einzusetzen. So geben sie sich häufig mit einem signifikanten Testergebnis zufrieden, wenn sie glauben, dass ihre Hypothese dadurch bestätigt wird – zwar erleichtert, aber ohne große Begeisterung und ohne so recht vom Sinn ihres Tuns überzeugt zu sein (und ohne darüber nachzudenken, welche faszinierenden Methoden hinter ihren Ergebnissen schlummern).

Es erscheint daher sinnvoll, dass ein Statistiker und ein Vertreter der jeweiligen Fachdisziplin im Sinne einer umfassenden und effizienten Datenanalyse interdisziplinär zusammenarbeiten – beginnend von der Studienplanung und Fallzahlschätzung über die Datenanalyse bis hin zum Darstellen und Interpretieren der Ergebnisse und dem Aufzeigen von Konsequenzen. Eine solche Kooperation setzt freilich voraus, dass alle Beteiligten ihre eigenen Fertigkeiten einbringen, der anderen Disziplin vorurteilsfrei begegnen sowie die spezifischen Fähigkeiten und Erfahrungen ihres Kollegen anerkennen. Nur dann ist gewährleistet, dass beide Seiten in optimaler Weise profitieren: Der Empiriker, weil er neue Erkenntnisse gewinnt; der Statistiker, weil er das befriedigende Gefühl hat, ein optimales Mo-

dell erschaffen zu haben; und beide, weil sie gemeinsam zum wissenschaftlichen Fortschritt beigetragen haben.

Letzten Endes sind alle Wissenschaftler die Nutznießer einer statistischen Analyse, auch diejenigen, die nicht an der Durchführung einer Studie und der Datenanalyse unmittelbar beteiligt sind (z. B. die behandelnden Ärzte). Mithilfe der Statistik lassen sich nämlich, trotz der Unvorhersehbarkeit im Einzelfall, allgemein gültige Aussagen herleiten. Diese bilden die Basis für jedes daraus abgeleitete ärztliche Handeln. Mit dieser Vorgehensweise ist zwar nicht sichergestellt, dass eine getroffene Entscheidung in jedem Fall zum gewünschten Ergebnis führt. Die Entscheidung ist aber nachvollziehbar, und das Risiko einer Fehlentscheidung ist minimiert. Insofern wird ein Arzt von den Ergebnissen statistischer Analysen in keiner Weise bevormundet, sondern vielmehr bei seinen Entscheidungen im klinischen Alltag unterstützt. Ähnliche Überlegungen lassen sich anstellen für die Potenziale der „Artificial Intelligence“, die speziell in der Röntgendiagnostik hochspezialisierte Untersuchungen in kürzester Zeit zu leisten vermögen. Sie können den Diagnoseprozess beschleunigen, den behandelnden Arzt unterstützen und die Chancen einer Heilung beim Patienten verbessern.

Als **Konsument** kann man sich auch im alltäglichen Leben an der Schönheit der Statistik erfreuen. Eine Person, die mit Wahrscheinlichkeiten umzugehen weiß, kann weniger leicht in die Irre geführt werden. Sie wird weder an Wunder noch an höhere Mächte glauben, wenn sie mit sensationellen Versprechungen konfrontiert wird; sie wird nicht in Panik verfallen, wenn ihr apokalyptische Prognosen zu Ohren kommen; sie wird auch keine Wissenschaftsdogmen akzeptieren, ohne sie kritisch zu hinterfragen. Sie weiß, dass Wahrscheinlichkeiten nichts aussagen über Kausalitäten und wird deshalb keine vorschnellen Schlüsse ziehen (auch wenn diese im aktuellen Trend liegen sollten). Statistische Überlegungen gepaart mit kühlem Sachverstand helfen, Informationen kritisch zu reflektieren, Risiken realistisch einzuordnen und somit gelassener zu werden.

Zusammenfassend lässt sich schlussfolgern: Jeder Mensch sollte der Statistik unbefangen begegnen und sich der Schönheit dieser Disziplin öffnen – sei es als Wissenschaftler, der Studien durchführt; als Nutznießer, der die Ergebnisse praktisch umsetzt; oder als ein Konsument, der dazu angehalten ist, die auf ihn einströmenden Informationen zu bewerten. Dann wird er oder sie erfahren, welche Souveränität statistische Kenntnisse zu verleihen imstande sind. Am ehesten gelingt dies, wenn man (im Sinne von Herbert George Wells) Kinder bereits

im Grundschulalter dafür sensibilisiert und ihnen (zusammen mit Lesen und Schreiben) Verständnis für Zahlen und die statistische Denkweise nahebringt.

Literatur

1. Armitage P: Trials and errors: The emergence of clinical statistics. J R Stat Soc 1983; 146: 321–334. DOI: 10.2307/2981451
2. Bollet AJ: Pierre Louis: The numerical method and the foundation of quantitative medicine. Am J Med Sci 1973; 266: 23–101. DOI: 10.1097/00000441-197308000-00002
3. Ferdinand D, Otto M, Weiss C: Get the most from your data. A propensity score model comparison in real life-data. Int J Gen Med 2016; 9: 123–131. DOI: 10.2147/IJGM.S104313
4. Gigerenzer Gerd: Das Einmaleins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. Berlin Verlag. 3. Auflage 2003
5. Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L: Klassische Wahrscheinlichkeiten: 1660–1840. In: Das Reich des Zufalls. Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Unschärfen. Spektrum Akademischer Verlag 1999
6. Kerlikowske K: Efficacy of screening mammography among women aged 40 to 49 years and 50 to 50 years: Comparison of relative and absolute benefit. J Natl Cancer Inst Mono 1997; 22: 79–86. DOI: 10.1093/jncimono/1997.22.79
7. Krämer W: Konkurrierende Risiken. In: Statistik für die Westentasche; Seite 56 ff., Piper-Verlag München 2002
8. Martin M, Fangerau H: Claude Bernard und der „europäische Durchschnittsharn". Urologe 2010; 49: 855–860. DOI: 10.1007/s00120-010-2277-9
9. Porzsolt F, Rocah NG, Toledo-Arruda AC, Thomaz TG, Moraes C, Bessa-Guerra TR, Leao M, Migowski A, Araujo da Silva AR, Weiss C: Efficacy and effectiveness trials have different goals, use different tools, and generate different messages. Pragmat Obs Res 2015; 6: 47–54. DOI: 10.2147/POR.S89946
10. Porzsolt F, Wambo GOK, Rösch MC, Weiß C: Prävention muss effizienter werden. Dtsch Med Wochenschr 2016; 141: 651–653. DOI: 10.1055/s-0042-101669
11. Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41–55. DOI: 10.1093/biomet/70.1.41
12. Strick HK: Adolphe Quetelet (1796–1874): Der mittlere Mensch. In: Spektrum.de. 1. Februar 2011

13. Tröhler U: Die therapeutische „Erfahrung“ – Geschichte ihrer Bewertung zwischen subjektiv sicherem Wissen und objektiv wahrscheinlichen Kenntnissen. In: Köbberling J: Die Wissenschaft in der Medizin. Selbstverständnis und Stellenwert in der Gesellschaft. Schattauer-Verlag 1992
14. Weiss C: Entwicklung der Medizinischen Statistik in Deutschland. Der lange Weg dahin. GMDS Med Inform Biom Epidemiol; 2005

Über die Autorin

Christel Weiß ist Professorin für Biomathematik und Epidemiologie an der Medizinischen Fakultät Mannheim der Universität Heidelberg. Zu ihrem Tätigkeitsbereich zählen Lehrveranstaltungen für Studierende der Medizin sowie die statistische Beratung von Ärzten, wissenschaftlichen Mitarbeitern und Doktoranden bei der Planung und Durchführung von klinischen und epidemiologischen Studien. Frau Weiß ist Autorin des Lehrbuchs „Basiswissen Medizinische Statistik“ (erschienen im Springer-Verlag, demnächst 7. Auflage), des Ratgebers „Promotion. Die medizinische Doktorarbeit – von der Themensuche bis zur Dissertation“ (zusammen mit Prof. Dr. Axel Bauer, erschienen im Thieme-Verlag, 4. Auflage) sowie Coautorin zahlreicher wissenschaftlicher Papers.

Korrespondenz:
Prof. Dr. Christel Weiß
Medizinische Fakultät Mannheim der Universität Heidelberg
Abteilung für Medizinische Statistik und Biomathematik
Theodor-Kutzer-Ufer 1
68167 Mannheim, Germany
E-Mail: Christel.Weiss@medma.uni-heidelberg.de