HEIDELBERGER
JAHRBÜCHER
ONLINE
Band 6 (2021)

Gesellschaft der Freunde
Universität Heidelberg e.V.

Intelligenz: Theoretische Grundlagen und praktische Anwendungen

Rainer M. Holm-Hadulla, Joachim Funke & Michael Wink (Hrsg.)

HEIDELBERG
UNIVERSITY PUBLISHING

Statistik und Intelligenz – eine wechselvolle Beziehung

CHRISTEL WEISS

Medizinische Fakultät Mannheim der Universität Heidelberg

Zusammenfassung

In diesem Beitrag wird das wechselvolle Beziehungsgeflecht „Statistik und Intelligenz“ untersucht. Es wird dargelegt, wie sich Statistik und Intelligenz gegenseitig beeinflussen. Einerseits werden statistische Methoden benötigt, um das Phänomen „Intelligenz“ messen zu können. Andererseits setzt die Entwicklung derartiger Methoden seitens Mathematikern oder Stochastikern Klugheit, Verstand und Scharfsinn voraus. Ferner wird aufgezeigt, welche Macht Daten inhärent ist, und welche Herausforderungen Statistiker und Fachvertreter (die Studien initiieren und durchführen) zu bewältigen haben, um den darin verborgenen Informationsgehalt aufzudecken, die Ergebnisse der Datenanalyse adäquat zu interpretieren und konsequent umzusetzen.

1 Einleitung

Statistik und Intelligenz – wie passt das zusammen? Während man mit „Intelligenz“ allseits anerkannte Eigenschaften wie beispielsweise eine breit gefächerte Allgemeinbildung, fundiertes Fachwissen, praktische Begabungen oder künstlerisches Talent assoziiert, weckt der Begriff „Statistik“ eher unangenehme Empfindungen wie Skepsis, Misstrauen oder Verständnislosigkeit.

Das Wort Intelligenz leitet sich ab vom lateinischen Verb „intellegere“. Dieses setzt sich zusammen aus den Bestandteilen „inter“ (zwischen, inmitten) und

https://doi.org/10.17885/heiup.hdjbo.2021.1.24391

„legere“ (lesen, auswählen) und lässt sich demnach frei übersetzen mit „zwischen den Zeilen lesen“. In einem sehr allgemeinen Sinne bezeichnet Intelligenz die kognitive Fähigkeit, neuen Anforderungen mit Denkleistungen zu begegnen und eigenständig Lösungen zu entwickeln.

Intelligenz gilt seit jeher als ein hohes Gut. Niemand wäre beglückt oder gar stolz, wenn man ihm nur einen mäßigen Intellekt attestieren würde. Dagegen gilt es in einigen Kreisen keineswegs als ehrenrührig, mit seinen bescheidenen Kenntnissen in Mathematik oder mangelhaften Fähigkeiten in Statistik zu prahlen. Es wird gar als Ausdruck kritischen Denkvermögens angesehen, wenn ein Zeitgenosse ohne stichhaltige Argumente kundtut, dass er die Aussagekraft eines Zahlenwerks bezweifelt – seien es Umfragewerte vor einer politischen Wahl oder in epidemiologischen Studien ermittelte Risiken. Häufig genügt die Erwähnung eines vermeintlichen Gegenbeispiels oder der Hinweis auf viel zitierte, aber wenig hinterfragte Sätze wie etwa „Mit Statistik kann man alles beweisen“, um auf breite Zustimmung zu stoßen.

Andererseits ist nachvollziehbar: Wer die Ergebnisse statistischer Analysen unvoreingenommen zu interpretieren weiß, wird weniger leicht manipuliert. Der Fortschritt in empirischen Wissenschaften wie beispielsweise der Medizin, der Soziologie oder der Meteorologie basiert in entscheidendem Maße auf den vielfältigen Möglichkeiten komplexer Analysetechniken. Um dies anzuerkennen, ist neben Offenheit und der Fähigkeit zu kritischem Denken ein gewisses Maß an Intelligenz nicht hinderlich.

Tatsächlich verbirgt sich in umfangreichem Datenmaterial ein gewaltiges Potential: seien es Register oder Merkmale, die im Rahmen von empirischen Studien erhoben werden, oder Messwerte, die bei der Patientenversorgung im klinischen Alltag routinemäßig erfasst werden. Diese Datensammlungen beinhalten wertvolle Informationen, auch wenn sich diese auf den ersten Blick nicht unbedingt erschließen. Ein Statistiker ist dann vor die herausfordernde Aufgabe gestellt, dieses Datenmaterial idealerweise gemeinsam mit einem Vertreter der jeweiligen Fachdisziplin so aufzubereiten, dass neue Erkenntnisse daraus hervorgehen. Deshalb sollten Forscher, die bei der Planung und Durchführung von empirischen Studien involviert sind, über einen gewissen Intellekt verfügen, um die Studie adäquat zu planen und aus den Ergebnissen der Datenanalyse sinnvolle Konsequenzen ziehen zu können und diese zum Wohle der Allgemeinheit umsetzen zu können.

Nicht zuletzt seien auch die Wissenschaftler erwähnt, die diese statistischen Methoden ersonnen haben. Meist handelt es sich dabei um Mathematiker oder Naturwissenschaftler, denen man ja im Allgemeinen zumindest eine Inselbegabung zugesteht. Als Beispiele seien *Carl Friedrich Gauß* (1777–1855), *Karl Pearson* (1857–1936) oder *Sir Ronald Aylmer Fisher* (1890–1962) genannt, deren Genialität wir bekannte und bis heute vielfach angewandte Konzepte der Statistik verdanken.

Schließlich ist es die Statistik selbst, die Methoden zur Verfügung stellt, die eine Messung der Intelligenz erst ermöglichen. Mit anderen Worten: Ohne Statistik wäre Intelligenz nur ein vages Konstrukt und es wäre reizlos, darüber zu fabulieren.

Diese vielfältigen und verwobenen Beziehungsgeflechte sollen in dem vorliegenden Beitrag näher beleuchtet werden.

2 Statistische Methoden zum Messen von Intelligenz

2.1 Historische Betrachtungen

Intelligente Menschen und elitäre Schichten erfreuten sich zu allen Zeiten hoher Wertschätzung. Dies erklärt, weshalb Individuen stets bemüht waren, ihre Fähigkeiten unter Beweis zu stellen und mit denen anderer Zeitgenossen zu vergleichen, etwa in Form von Wettbewerben. Olympische Spiele wurden bereits in der Antike durchgeführt. Auch im Mittelalter gab es Wettbewerbe, etwa um die musische Begabung von Minnesängern zu evaluieren.

Eine Auslese der Besten – der Elite – mag auch gesellschaftlich relevant sein. Ein Beispiel ist im Alten Testament nachzulesen: So wird im „Buch Richter“ des Alten Testaments in Kapitel 7 berichtet, wie der Richter Gideon eine Elite der 300 besten Soldaten zum Kampf gegen den Nomadenstamm der Midianiter auswählte. Der Begriff „Elite“ im Sinne einer Leistungs- oder Intelligenzelite tauchte erst wesentlich später auf, als sich zur Zeit der Industrialisierung gebildete Bürger von der breiten Masse der Ungebildeten abzuheben trachteten, um eine gesellschaftlich privilegierte Stellung oder politische Macht zu beanspruchen.

Der Begriff „Intelligenz“ war keineswegs präzise definiert (was sich bis heute nicht wesentlich geändert hat). Die Beurteilung der Intelligenz oder die Wahl einer Elite erfolgte in aller Regel intuitiv. Dabei sollten vor allem hochbegabte Menschen identifiziert werden; mittelmäßig oder gar minderbegabten Insdividuen schenkte man keine Beachtung.

Einer der ersten Wissenschaftler, der sich mit dem Phänomen „Intelligenz“ eingehend befasste, war der britische Naturforscher *Sir Francis Galton* (1822–1911), ein Cousin von *Charles Robert Darwin* (1809–1882). Von Darwins Vererbungslehre inspiriert, gründete er in London den ersten Lehrstuhl für Eugenik, jener wissenschaftlichen Disziplin, die sich zum Ziel setzte, basierend auf den Erkenntnissen der Humangenetik die Verbreitung günstiger Erbanlagen zu fördern. Galton war der Meinung, dass Intelligenz ausschließlich vererbt werde; Umweltfaktoren schenkte er keine Beachtung. Obgleich Persönlichkeitsmerkmale wie die Intelligenz allgemein als „inkommensurabel“ (nicht messbar) galten, war Galton der festen Überzeugung, dass dieses Merkmal (ebenso wie physiologische Parameter wie etwa die Körpergröße) quantifizierbar sei, auch wenn bis dahin noch keine Form der Operationalisierung bekannt war. Er nahm außerdem an, dass die Messwerte einer Population symmetrisch um einen Mittelwert schwanken und durch eine Normalverteilung beschrieben werden können. Auch Galtons Nachfolger, der Statistiker *Karl Pearson*, vertrat die Meinung, dass Intelligenz ebenso wie zahlreiche Krankheiten, Alkoholismus oder Kriminalität allein erblich bedingt sei. Erst Pearsons Nachfolger *Lionel Sharples Penrose* (1898–1972), der sich als Psychiater, Mathematiker und Genetiker einen Namen machte, äußerte die Vermutung, dass Intelligenz auch von zahlreichen Umweltparametern beeinflusst würde.

Die genannten Forscher befassten sich mit dem Phänomen „Intelligenz“ hauptsächlich im Kontext eugenischer Zielsetzungen. Das Erfassen von Intelligenz oder die Entwicklung einer geeigneten Messtechnik standen nicht im Fokus ihres Interesses. Allein Galton ersann ein Verfahren, um die Intelligenz der Masse zu bestimmen: Beim Besuch eines Volksfestes im Jahre 1906 durfte jeder Besucher das Gewicht eines Ochsen schätzen. Dabei zeigte sich: Der Durchschnittswert der fast 800 Einzelschätzungen wich nur knapp vom tatsächlichen Gewicht des Tieres ab – obwohl viele Schätzwerte weit über oder unter dem wahren Wert lagen. Dieses Ergebnis überraschte Galton: Eigentlich hatte er beabsichtigt, mit diesem Experiment die „Dummheit“ der Masse zu belegen. Dieses Phänomen, wonach größere Gruppen kognitive Aufgaben mitunter präziser lösen als einzelne Experten, wurde seither mehrfach bestätigt [8]. Für die Messung der Intelligenz von Individuen ist dieser Ansatz jedoch nicht brauchbar.

2.2 Intelligenz als komplexes Konstrukt

Das Merkmal „Intelligenz“ ist derart komplex, dass es sich einer direkten Beobachtung oder einer präzisen Messung mittels eines technischen Geräts entzieht. Subjektive Einschätzungen seitens des betroffenen Individuums oder durch eine andere Person etwa auf einer Visuellen Analogskala sind naturgemäß sehr vage und wenig objektiv.

Wie lässt sich ein solches Merkmal statistisch handhaben? Eine Möglichkeit besteht eventuell in der Verwendung einer einfach zu bestimmenden Ersatzgröße, eines sogenannten Surrogatmerkmals. So versuchte der Anatom *Franz Joseph Gall* (1758–1828) basierend auf seiner Schädellehre, nach der das Gehirn das Zentrum für alle mentalen Funktionen sei, die Intelligenz anhand des Schädelumfangs indirekt zu erfassen. Zu Beginn des 19. Jahrhunderts vertraten viele Forscher die Meinung, dass sich die Intelligenz im äußeren Erscheinungsbild widerspiegelte und dass Parameter wie Schädelumfang, Stirnbreite und -höhe Anhaltspunkte für besondere Begabungen lieferten. Andere Wissenschaftler wie der bereits erwähnte Galton oder der Anatom *Friedrich Tiedemann* (1781–1861) glaubten, dass man aus dem Schädelvolumen eines Menschen auf dessen intellektuelles Denkvermögen schließen könnte. Galton hatte die Prüfungsergebnisse von Studenten mit deren geschätztem Hirnvolumen korreliert (tatsächlich gilt Galton als einer der ersten Wissenschaftler, die den Korrelationskoeffizienten verwendeten); Tiedemann hatte zu diesem Zweck Totenschädel vermessen.

Bei einem geeigneten Surrogatmerkmal muss allerdings gewährleistet sein, dass es in engem Zusammenhang mit dem eigentlich zu messenden Merkmal steht und dass es verlässliche Informationen liefert. Bald zeigte sich, dass der Schädelumfang diese Kriterien nicht erfüllt. Zwar glaubten sowohl Tiedemann als auch Galton, eine Assoziation zwischen Intellekt und Hirnvolumen erkannt zu haben. Mittlerweile wurde jedoch nachgewiesen, dass das Hirnvolumen allenfalls sehr schwach mit der Intelligenz korreliert und daher als Surrogatmerkmal untauglich ist [16].

Ein konkreter Anlass zur Entwicklung eines Messverfahrens war gegeben, als im 19. Jahrhundert in Frankreich die allgemeine Schulpflicht eingeführt wurde. So entstand der Bedarf nach Selektionskriterien, um Schüler angemessen fördern zu können. Im Jahre 1904 erhielt der Psychologe *Alfred Binet* (1857–1911) von der französischen Regierung den Auftrag, ein Verfahren zu entwickeln, um die Begabung von Schülern nach objektiven Kriterien erfassen zu können. Zusammen

mit seinem Kollegen *Théodore Simon* (1873–1961) konstruierte er den ersten Intelligenztest. Dieses Konzept basierte auf der Vorstellung, dass das Phänomen „Intelligenz“ als ein Konstrukt anzusehen sei, das über mehrere Indikatoren (sogenannte Items) zu erfassen ist. Zu diesem Zweck ersannen Binet und Simon Testaufgaben unterschiedlichen Schwierigkeitsgrades, die von den Schulkindern zu lösen waren. Die Intelligenz wurde basierend auf der Anzahl korrekter Lösungen ermittelt.

In der Folgezeit wurde dieser Test mehrfach modifiziert und erweitert (worauf im nächsten Abschnitt eingegangen wird). Es fanden sich zahlreiche Befürworter, weil damit erstmals eine quantitative Methode zur Identifizierung lernschwacher Kinder gegeben war. Kritiker warfen dagegen ein, dass dieser Test nur einen Teilaspekt der Intelligenz erfassen würde. Dass jedoch Intelligenz ganz allgemein als ein gedankliches Konstrukt anzusehen ist, das nur mittelbar über zahlreiche Indikatoren erschlossen werden kann, ist nach wie vor unbestritten.

2.3 Intelligenztests

Binet und Simon gebührt das Verdienst, dass sie sich nicht nur theoretisch mit dem Phänomen „Intelligenz“ befassten, sondern in pragmatischer Weise ein Verfahren zu deren Operationalisierung ersannen. Dabei lag die Herausforderung weniger in der praktischen Durchführung des Tests, sondern vielmehr in der Wahl geeigneter Testaufgaben, die in ihrer Vielfalt möglichst trennscharf die Schüler in unterschiedliche Leistungsgruppen einteilen sollen.

Die Palette erstreckt sich von sehr einfachen Aufgaben, die ein Kleinkind bewältigen sollte (etwa das Verfolgen eines brennenden Streichholzes mit den Augen) über schwierigere (beispielsweise rückwärts zählen) bis hin zu sehr anspruchsvollen Aufgaben, die im Allgemeinen nur ältere Kinder lösen können (zum Beispiel, Reimworte zu finden). Für jede Altersstufe stehen in der Regel sechs Aufgaben zur Verfügung. Die höchste Altersstufe, deren Aufgaben ein Kind komplett bewältigt, ist das sogenannte Grundalter. Um das Intelligenzalter zu ermitteln, werden für jede zusätzliche Lösung einer höheren Altersgruppe zwei Monate zum Grundalter addiert. Im Durchschnitt (bezogen auf die Kinder einer bestimmten Altersgruppe) entspricht das Intelligenzalter dem Lebensalter. Ein Intelligenzalter, das deutlich unter dem Lebensalter liegt, weist darauf hin, dass gezielte Förderungsmaßnahmen für das betreffende Kind sinnvoll sind.

Dieser Test ermöglichte erstmals die Quantifizierung der Intelligenz und die Vergleichbarkeit zwischen Individuen derselben Altersgruppe. Subjektive Lehrerurteile bezüglich der Leistungsfähigkeit von Schulkindern wurden durch eine objektive und nachvollziehbare Messmethode ersetzt. Das Testergebnis ist ein numerischer Wert, der leicht zu ermitteln und einfach zu interpretieren ist. Allerdings ist der Binet-Simon-Test nicht uneingeschränkt anwendbar: Vergleiche zwischen verschiedenen Altersgruppen erweisen sich als problematisch, da der Unterschied zwischen Intelligenz- und Lebensalter bei jüngeren Kindern schwerer wiegt als bei älteren.

Der Psychologe *William Louis Stern* (1871–1938) löste dieses Problem, indem er das Intelligenz- und das Lebensalter in ein Verhältnis setzte und dieses mit 100 multiplizierte. So entstand der Intelligenzquotient *IQ*:

$$IQ = \frac{Intelligenzalter}{Lebensalter} \times 100$$

Aufgrund dieser Definition ist gewährleistet, dass der *IQ* eines durchschnittlich intelligenten Kindes jedweder Altersgruppe 100 beträgt [14]. Ein 10 Jahre altes Kind mit einem Intelligenzalter von 11 Jahren hat demnach einen *IQ* von 110, während für ein 5-jähriges Kind mit einem Intelligenzalter von 6 Jahren ein *IQ* von 120 resultiert.

Aber auch dieser Intelligenzquotient hat Limitationen bezüglich seiner Anwendung: Da mit steigendem Lebensalter die Intelligenz langsamer zunimmt als im Kindesalter, würde nach Sterns Formel der *IQ* bei einem Individuum im Laufe seines Lebens stetig sinken. Für Erwachsene ist dieses Verfahren deshalb nicht sinnvoll.

Um diesem Problem zu begegnen, entwickelte der Psychologe *Lewis Madison Terman* (1877–1956) von der Universität Stanford mehrere altersspezifische Testvarianten für Kinder und Jugendliche sowie für normal und überdurchschnittlich begabte Erwachsene. Er normierte die Ergebnisse innerhalb jeder Variante – und zwar dergestalt, dass der *IQ* der jeweiligen Population einer Normalverteilung mit dem Mittelwert 100 und der Standardabweichung 15 unterliegt (vgl. Abb. 1). Der sogenannte Stanford-Binet-Test wurde bald zum Standardinstrument in der Psychologie und in der Schulberatung. Aus den Eigenschaften der Normalverteilung folgt, dass bei jeweils der Hälfte der Getesteten Werte über oder unter 100 gemessen werden; etwa zwei Drittel erzielen *IQ*-Werte zwischen 85 und 115 Punkten (das ist das Intervall „Mittelwert $\pm$ eine Standardabweichung“). Nur etwa

5% aller *IQ*-Werte liegen weiter als zwei Standardabweichungen vom Mittelwert entfernt. Demnach werden bei etwa 2,5% der Populationsmitglieder *IQ*-Werte unter 70 festgestellt, was mit einer Verminderung kognitiver Fähigkeiten einhergeht. Das andere Extrem – ein *IQ* ab 130 – wird ebenfalls nur von 2,5% erzielt. Diese Menschen gelten als hochbegabt.

Letzten Endes waren damit Galtons Vorstellungen Wirklichkeit geworden: Intelligenz konnte als ein quantitatives, messbares Merkmal angesehen werden, dessen Variabilität innerhalb einer Population durch das idealtypische Modell einer Normalverteilung beschrieben wird. Damit war eine Klassifizierung in durchschnittliche, unter- und überdurchschnittliche Subgruppen nach objektiven Kriterien möglich geworden.

Einen weiteren wichtigen Beitrag zur Intelligenzforschung leistete der britische Psychologe *Charles Edward Spearman* (1863–1945). Ihm war bei der Beobachtung von Schulkindern aufgefallen, dass die Ergebnisse verschiedener Leistungstests positiv korrelierten. Deshalb nahm er an, dass allen intellektuellen Leistungen eine einheitliche Form der Intelligenz zugrunde liegt. Diese wird in einem Allgemeinen Faktor, dem sogenannten *g*-Faktor („general intelligence"), erfasst [12]. Darüber hinaus – so postulierte er – gibt es Gruppenfaktoren (die so-

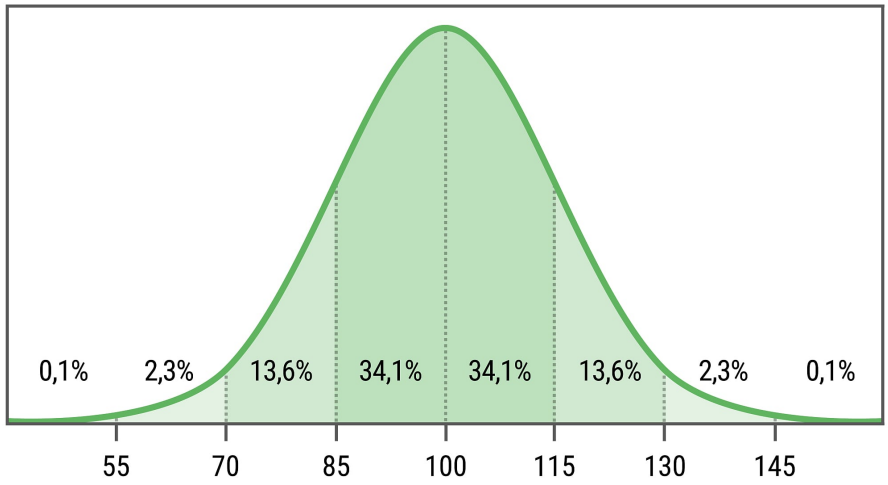

Abbildung 1: Gauß'sche Glockenkurve zur Beschreibung des Intelligenzquotienten mit dem Erwartungswert 100 und der Standardabweichung 15. Aus [16].

genannten *s*-Faktoren, etwa verbale Fähigkeiten oder räumliches Denkvermögen), von denen jeweils einer zur Lösung einer speziellen Problemstellung benötigt wird.

David Wechsler (1896–1981), ein Schüler Spearmans, war von der Sinnhaftigkeit der akademischen Definition von Intelligenz und der Aussagekraft der damals verwendeten *IQ*-Tests nicht überzeugt. Er verstand Intelligenz in einem umfassenderen Sinne, der auch nicht-intellektuelle Aspekte einschloss. Im Jahre 1939 stellte er eine Serie aus mehreren Einzeltests (Skalen) vor. Getestet werden u. a. die Fähigkeit zu logischem Denken, Arbeitsgedächtnis, Sprachverständnis, Verarbeitungsgeschwindigkeit und Konzentrationsvermögen. Wie beim Stanford-Binet-Test sind die Werte innerhalb einer Altersgruppe normalverteilt mit dem Erwartungswert 100 und der Standardabweichung 15.

Ursprünglich war der Test nach Wechsler für Erwachsene konzipiert und wurde vorwiegend in den USA verwendet. Seit 1956 liegen deutsche Testversionen vor, die unter der Bezeichnung „Hamburg-Wechsler-Intelligenztest“ bekannt sind. Diese Versionen wurden seither mehrfach revidiert und neuen Erkenntnissen der aktuellen Forschung angepasst [2].

Wechsler hielt an Spearmans Zweifaktorentheorie von einem allgemeinen und einem spezifischen Intelligenzfaktor fest. Bei jedem Wechsler-Intelligenztest wird der „Allgemeine Fähigkeitsindex“ ermittelt, vergleichbar Spearmans *g*-Faktor. *IQ*-Tests nach Wechsler zählen mittlerweile zu den weltweit am häufigsten verwendeten Instrumenten, um Talente zu entdecken, Kandidaten bei Bewerbungen einzuschätzen, geeignete Studienbewerber ausfindig zu machen, lernschwache Schüler oder Hochbegabte zu erkennen.

3 Der Einfluss von Intelligenz bei Datenanalysen

3.1 Intelligenz der Daten

Der Mensch ist von Natur aus neugierig. Deshalb werden Daten – in seinem sehr allgemeinen Sinne – seit Beginn der Menschheit erhoben: Mit den Sinnesorganen erfasst der Mensch seine Umwelt, die Sprache benutzt er zur Kommunikation, die Finger zum Zählen. So verwundert es nicht, dass sich das Ermitteln von Häufigkeiten als die älteste Form der statistischen Datenerfassung etabliert hat.

Eine erste Form der systematischen Datenerhebung entstand ab dem 16. Jahrhundert, als die ersten Nationalstaaten aufblühten. Die erhobenen Daten dienten

vor allem dem Zweck, wesentliche Charakteristika einer Bevölkerung zu beschreiben und administrative Planungen zu ermöglichen. Seit dem Zeitalter der Renaissance erhoben Astronomen und Naturwissenschaftler wie beispielsweise *Johannes Kepler* (1571–1630) oder *Galileo Galilei* (1564–1641) Daten, um naturwissenschaftliche Gesetze herzuleiten. Einige Jahre später erstellten der Kaufmann *John Graunt* (1620–1674) in London und der preußische Feldprediger *Johann Peter Süßmilch* (1707–1767) in Berlin basierend auf Eintragungen in Kirchenbüchern Sterbetafeln und schätzten aus den so gewonnen Daten Lebenserwartungen. Seit Beginn des Industriezeitalters im 18. Jahrhundert stehen technische Geräte zur Verfügung, mit denen sich physikalische Größen messen lassen. Die ersten empirischen Studien im Bereich der epidemiologischen und klinischen Forschung wurden im 19. Jahrhundert durchgeführt: Dem englischen Arzt *John Snow* (1813–1858) gelang es im Jahr 1849, eine Cholera-Epidemie in London einzudämmen, nachdem er Daten von betroffenen Patienten und nicht erkrankten Kontrollen gesammelt und ausgewertet hatte. Der französische Arzt *Pierre Charles Alexandre Louis* (1787–1872) belegte anhand von geeignetem Datenmaterial, dass die bis dahin häufig angewandte Methode des Aderlasses bei den meisten Patienten nutzlos oder gar schädlich war. Der ungarische Gynäkologe *Philipp Ignaz Semmelweis* (1818–1865) konnte aufgrund systematischer Untersuchungen verbunden mit akribischer Datendokumentation nachweisen, dass das damals gefürchtete Kindbettfieber durch mangelnde Hygiene verursacht war.

Diese Beispiele zeigen, dass in Daten ein gewaltiges Potential steckt: Sowohl naturwissenschaftliche Gesetze als auch empirische Erkenntnisse in der Medizin oder den Sozialwissenschaften wurden und werden basierend auf einer adäquaten Datenmenge verifiziert. Daten sind das Fundament jeder statistischen Analyse. Mit ihrer Hilfe lassen sich Ereignisse erklären und Zusammenhänge erhellen.

Doch sind Daten intelligent? Sicher nicht im Sinne einer kognitiven Fähigkeit. Daten per se können keine eigenständigen Lösungen entwickeln. Sie sprechen nicht für sich selbst, sie geben von alleine nichts preis. Aus Daten kann zwar Information entstehen – jedoch nur für jemanden, der in der Lage ist, diese Information zu extrahieren. So besagt beispielsweise ein *Body Mass Index* (BMI) von 40 kg/m^2, dass die betreffende Person unter Adipositas leidet. Allerdings ist dies nur für jemanden erkennbar, der die Definition des BMI kennt und weiß, wie dieses Merkmal verteilt ist. Ein Konglomerat von Daten vermag eine Anekdote zu erzählen – jedoch nur jemandem, der weiß, wie die Daten zustande gekommen

sind und in welchem Kontext sie stehen. Nur mit diesem Hintergrundwissen ist man in der Lage, Einzeldaten zu einem sinnvollen Ganzen zusammenzufügen.

Jedes umfangreiche Datenmaterial ist wertlos, wenn es nicht adäquat aufbereitet wird. Dies sei am Beispiel der Wettervorhersage erläutert: Tausende Messstationen und Satelliten in allen Teilen dieser Welt liefern permanent Daten in einer schier unüberblickbaren Masse. Darin steckt das Potential, das Wetter des nächsten Tages in quasi allen Regionen dieser Erde präzise zu prognostizieren. Jedoch bedarf es dazu leistungsfähiger Computer, einer ausgefeilten Technik, eines professionellen Datenmanagements, komplexer mathematischer Algorithmen und nicht zuletzt fachlich versierter Meteorologen, die die Datenflut beherrschen und verarbeiten und die daraus hervorgehenden Ergebnisse zu deuten wissen.

Daten können qualitativ mehr oder weniger hochwertig sein. Dies betrifft die Art ihrer Erhebung, ihre Präzision, ihre Verlässlichkeit und ihre Vollständigkeit. Daten fallen nicht vom Himmel. Deshalb sollte in der Planungsphase einer Studie darauf geachtet werden, dass aussagekräftige und für die Fragestellung relevante Variablen ausgewählt werden, die die Realität valide abbilden, und dass die Daten vollständig und korrekt erhoben sowie sorgsam dokumentiert werden. Mag auch die Erfassung großer Datenmengen aus technischer Sicht kein nennenswertes Problem darstellen, so erfordern sie doch einen kompetenten Umgang, um das darin enthaltene Potential auszuschöpfen.

3.2 Intelligenz der Methoden

Sind statistische Methoden intelligent? Auch diese Frage ist zu verneinen. Methoden werden angewandt, um Daten effizient zu verarbeiten. Die darin verborgene Intelligenz sollte man deren Entwicklern attestieren.

Betrachten wir als einfaches Beispiel den arithmetischen Mittelwert. Eine der ersten bekannten Anwendungen wurde von dem Astronomen *Tycho Brahe* (1546–1601), einem Mitarbeiter von Johannes Kepler, ersonnen. Um den Ungenauigkeiten der damals vorhandenen Uhrwerke zu begegnen, verwendete Brahe mehrere Uhren gleichzeitig und berechnete aus den ermittelten Zeitdauern den Mittelwert. Ein genialer Gedanke im ausgehenden 16. Jahrhundert!

Zu Beginn des 20. Jahrhunderts machte die induktive Statistik rasante Fortschritte. Der Chemiker *Sealy Gosset* (1876–1937) ersann ausgehend von einem praktischen Problem (er wollte als Angestellter der Brauerei *Guinness* die Mittelwerte von Bieringredienzen schätzen) die *t*-Verteilung, die für die berühmten

t-Lagetests und die Konstruktion von Konfidenzintervallen verwendet wird. Im gleichen Zeitraum publizierten die bereits erwähnten Wissenschaftler Pearson und Spearman die nach ihnen benannten Korrelationskoeffizienten zur Quantifizierung der Stärke eines Zusammenhangs zwischen zwei Merkmalen. Die Bestimmung der Parameter einer Regressionsgeraden war bereits etwa hundert Jahre zuvor von dem bekannten Mathematiker Gauß basierend der Methode der kleinsten Quadrate hergeleitet worden. In den 1950er Jahren legte der britische Statistiker und Genetiker *Sir Ronald Aylmer Fisher* den Grundstein für Varianzanalysen. Seine Motivation: Er wollte den Einfluss verschiedener Bodenqualitäten auf das Pflanzenwachstum evaluieren.

Dies sind nur wenige Beispiele aus der Liste herausragender Wissenschaftler, die die Statistik maßgeblich bereichert haben. Sie grübelten über ihre Methoden nach, lange bevor Computer zur Verfügung standen, um die mitunter sehr komplexen Berechnungen durchzuführen. Wahrscheinlich hätten sie sich nicht träumen lassen, dass sich die von ihnen ersonnenen Techniken noch Jahrzehnte oder gar Jahrhunderte später größter Beliebtheit erfreuen.

Seit den 1970er Jahren haben leistungsfähige Rechner und benutzerfreundliche Software die weitere Entwicklung der Statistik stimuliert. Dies betrifft vor allem multivariate Verfahren, die explorative Datenanalyse, Monte-Carlo-Methoden oder Metaanalysen (um nur einige Beispiele zu nennen). Mit dem beschleunigten technologischen Fortschritt ergab sich der Bedarf nach speziellen statistischen Methoden. So sahen sich Mathematiker vor die Herausforderung gestellt, Methoden zu ersinnen, mit denen die Reliabilität (Verlässlichkeit) und die Validität einer Messmethode überprüft werden kann (etwa bei Messinstrumenten für Diagnostik und Therapie oder bei psychometrischen Messverfahren). Auf diese Weise entstanden der Kappa-Index nach Cohen [3] und der Intraklassenkorrelationskoeffizient als Übereinstimmungsmaße [11], Cronbachs Alpha zum Quantifizieren der internen Konsistenz eines Fragebogens [4] oder die Bland-Altman-Analyse [1], mit der sich der Grad der Übereinstimmung zweier quantitativer Messmethoden beurteilen lässt.

Alle in diesem Abschnitt genannten Wissenschaftler erwiesen sich in beeindruckender Weise als intelligent, indem sie neuen Anforderungen mit genialen Denkleistungen begegneten und eigenständig Lösungen entwickelten.

3.3 Intelligenz der Anwender

Statistische Maßzahlen und Analysemethoden können Dinge auf den Punkt bringen. Die wichtigsten Eigenschaften einer Datenreihe oder die Stärke eines Zusammenhangs lassen sich mittels geeigneter Lage-, Streuungs- oder Assoziationsmaße quantifizieren; der Unterschied zwischen zwei Gruppen kann durch einen statistischen Test nachgewiesen und die Abhängigkeit einer Zielgröße von mehreren Einflussgrößen durch ein multiples Modell erklärt werden. Dazu bedarf es kompetenter Statistiker, die diese Verfahren auswählen.

Sind die Anwender der Statistik intelligent? Das wäre jedenfalls wünschenswert! Statistiker haben gemeinhin einen zweifelhaften Ruf, die Statistik als Wissenschaft gilt als rätselhaft und wenig attraktiv. Statistik ist zwar ein Teilgebiet der Mathematik, also einer exakten Wissenschaft. Sie orientiert sich an Daten, also an harten Fakten. Dennoch sind die Ergebnisse ihrer Methoden vom Zufall beeinflusst. Der p-Wert (die Irrtumswahrscheinlichkeit) mag einerseits als ein exaktes Evidenzmaß für die Plausibilität der Nullhypothese angesehen werden. Andererseits spiegelt dieser Wert die Unsicherheit wider, mit der das Ergebnis eines statistischen Tests behaftet ist.

Wie lassen sich unter dieser Prämisse Daten effizient analysieren? Um aussagekräftige Daten zu generieren, adäquate statistische Methoden auszuwählen und ein statistisches Modell zu erstellen, das basierend auf den erhobenen Daten die Realität bestmöglich beschreibt, bedarf es außer methodisch-fachlichen Wissens und empirischer Erfahrung einer feinen Beobachtungsgabe sowie eines vertieften wissenschaftlichen Verständnisses der zugrunde liegenden Fragestellung, Aufgeschlossenheit und Neugierde – denn die Ergebnisse der Analyse werden in aller Regel für ein anderes Fachgebiet erarbeitet.

Die optimale Datenanalyse ist keineswegs vorgegeben. Es reicht nicht aus und es ist nicht immer sinnvoll, sich strikt an Vorgaben zu halten; vielmehr ist in jeder Studienphase selbstständiges und kritisches Denken gefragt. Bei jeder Fragestellung erscheinen mehrere Herangehensweisen denkbar (obgleich jede statistische Methode an gewisse Voraussetzungen geknüpft ist, die vorab zu prüfen sind). Deshalb ist vor der Datenaufbereitung zu bedenken: Welches Studiendesign liegt vor? Sind die Daten korrekt? Wie lässt sich dies überprüfen? Wie behandelt man fehlende Werte? Wie verfährt man mit Ausreißern? Welche Analysemethoden bieten sich an? Was ist bei der Generierung eines multiplen Modells zu beachten?

Auch wenn allgemeiner Konsens darüber besteht, dass das Ergebnis einer statistischen Analyse als signifikant erachtet wird, wenn der p-Wert unter 0,05 liegt, lassen scheinbar exakte numerische Ergebnisse diverse Interpretationen zu. Deshalb sollte man hinterfragen: Wie kam ein Ergebnis zustande? Wie stark ist der nachgewiesene Zusammenhang? Ist er kausal? Sind die Ergebnisse eventuell mit einem Bias behaftet? Wie präzise ist die Schätzung? Und schließlich: Ist das Ergebnis unter fachlichen Aspekten relevant? Welche inhaltlichen Konsequenzen sind daraus zu ziehen?

Um diese Fragen zu klären, sollten sowohl Statistiker als auch Fachvertreter, die die betreffende Studie initiiert haben, bereit sein, sich Grundkenntnisse über das jeweils andere Fachgebiet anzueignen. Es erweist sich als produktiv, wenn Kollegen mehrerer Fachrichtungen im interdisziplinären Austausch das Ergebnis der Datenanalyse aus unterschiedlichen Perspektiven, in einer Kombination aus methodisch-statistischer und empirisch-fachwissenschaftlicher Sicht begutachten, kritisch hinterfragen, Anwendungsmöglichkeiten und Limitationen erörtern. Dabei ist neben Toleranz und Bedachtsamkeit auch ein gewisses Maß an gesundem Menschenverstand von Vorteil.

In diesen Fähigkeiten offenbart sich die Intelligenz aller an einer Studie beteiligten Wissenschaftler und insbesondere der Statistiker. Deren berufliche Tätigkeit zeichnet sich ja gerade dadurch aus, dass sie an der Planung und Durchführung von Studien aus diversen Fachgebieten beteiligt sind. Dazu müssen sie bereit sein, sich intellektuellen Herausforderungen zu stellen, sich in fachfremde Gebiete einzuarbeiten, eigenständig Lösungen zu entwickeln und über ihren Horizont hinauszusehen.

3.4 Intelligenz der Konsumenten

Datenanalysen dienen nicht nur dem wissenschaftlichen Fortschritt, sondern auch dem gesellschaftlichen Nutzen. Wenn beispielsweise bei einer Risikostudie die Assoziation zwischen einem ätiologischen Faktor und einer Krankheit aufgezeigt wird oder wenn bei einer Therapiestudie die Wirksamkeit eines neuen Medikaments nachgewiesen wird, sind diese Erkenntnisse nur dann zielführend, wenn möglichst viele Menschen davon profitieren. Wenn vor einer politischen Wahl eine repräsentative Umfrage durchgeführt wird, sind die ermittelten Werte von allgemeinem Interesse. Deshalb müssen die aus einer Datenanalyse hervorgehenden Ergebnisse der Öffentlichkeit zugänglich gemacht werden.

Tatsächlich kann sich kaum jemand der Statistik entziehen – weder im gewöhnlichen Alltag noch im wissenschaftlichen Diskurs. Diverse Medien sorgen dafür, dass man quasi rund um die Uhr mit Prozentzahlen, Tabellen, Diagrammen und Durchschnittswerten über aktuelle Ereignisse auf dem Laufenden gehalten wird. Jeder Arzt hat die Möglichkeit, permanent über Datenbanken auf weltweit verfügbares Wissen zugreifen zu können. Dies gewährleistet (zumindest theoretisch), dass jede seiner Entscheidungen etwa bezüglich Diagnosestellung oder Therapie eines Patienten auf dem neuesten Stand der Forschung basiert.

Manche Zeitgenossen fühlen sich jedoch unbehaglich angesichts der Fülle an Informationen, die auf sie einströmt. Sie befürchten, manipuliert zu werden, weil ihnen vermittelt wird, dass sie ihrem eigenen Verstand und ihren individuellen Erfahrungen kaum noch trauen können. Andere sind skeptisch, weil sie glauben, dass der Fortschritt (etwa in der Medizin) längst nicht so rasant verläuft, wie es die zahlreichen Publikationen in internationalen Journals suggerieren. „Statistisches Denken wird eines Tages für mündige Bürger ebenso wichtig sein wie die Fähigkeit, zu lesen und zu schreiben." Diese Aussage, die dem englischen Schriftsteller *Herbert George Wells* (1866–1946) zugeschrieben wird, zeigt sehr deutlich, dass statistisches Denken bereits zu seiner Zeit auf Schwierigkeiten stieß. Leider hat sich daran bis heute wenig geändert.

Grafische Darstellungen eignen sich hervorragend, um belanglose Effekte zu dramatisieren oder um unerwünschte Effekte zu kaschieren. Auch Zahlen sind mitunter weniger objektiv als es scheint. Insbesondere relative Risiken stiften leicht Verwirrung. Dazu ein Beispiel: In den 1990er Jahren wurden in britischen Medien alarmierende Informationen bezüglich der Antibabypille verbreitet. Die Einnahme eines bestimmten Präparats erhöhte angeblich das Risiko einer Thromboembolie um sage und schreibe 100%! Angesichts dieser Zahl setzten viele Frauen die Pille sofort panikartig ab, weil sie annahmen, dass diese lebensgefährliche Nebenwirkung mit Sicherheit eines Tages bei ihnen auftreten würde. Tausende unerwünschter Schwangerschaften und Abtreibungen waren die Folge dieses Missverständnisses. Die 100% quantifizierten nur eine relative Risikoerhöhung; die tatsächlichen Risiken betrugen lediglich 1/7000 bzw. 2/7000. – Auch bei der Anwendung von diagnostischen Tests und Screenings ist ein kritischer Umgang mit Wahrscheinlichkeiten vonnöten (und zwar sowohl bei Ärzten als auch bei Patienten), damit der Testbefund in Abhängigkeit von der Prävalenz angemessen interpretiert wird [15].

Eine andere Unsitte im Umgang mit Wahrscheinlichkeiten besteht in sinnlosen Vergleichen. Dazu eine Anekdote: Der an einem Prostatakarzinom erkrankte Rudy Giuliani, ehemaliger Bürgermeister von New York, verkündete einst, dass er sich glücklich schätzte, in den USA behandelt zu werden. Dort sei die Überlebensrate bei diesem Krankheitsbild nahezu doppelt so hoch wie in England (82% versus 44%). Dieser Unterschied ist zweifellos beachtlich; der Vergleich führt dennoch in die Irre. In den USA werden – anders als in England – flächendeckende Screenings angeboten. Deshalb werden bei US-amerikanischen Männern wesentlich mehr Prostatakarzinome diagnostiziert als bei Briten. Die Karzinome in den USA werden in der Regel zu einem früheren Zeitpunkt entdeckt und sind weniger aggressiv als Karzinome, die sich erst später durch klinische Symptome bemerkbar machen. Fair wäre dagegen ein Vergleich der krankheitsspezifischen Mortalitätsraten bezogen auf alle erwachsenen Männer des jeweiligen Landes: Diese betragen 26 bzw. 27 pro 100.000 Männer pro Jahr. Weitere Beispiele zur unsachgemäßen Interpretation von Wahrscheinlichkeiten finden sich in [7].

Daten haben die Macht, Entwicklungen zu erklären und Ereignisse zu prognostizieren. Sie können aber auch, wenn sie unsachgemäß analysiert werden, Ergebnisse verzerren und in die Irre führen. Daten lügen nicht, wohl aber derjenige, der Daten manipuliert, nur bestimmte Befunde aufzeigt oder unerwünschte Resultate verschleiert. Die verständliche Kommunikation von Wahrscheinlichkeiten und Unsicherheiten stellt eine Herausforderung dar. Deshalb sollten sich Forscher oder Journalisten, die die Ergebnisse statistischer Analysen aufbereiten und verbreiten, bei der Darstellung um Objektivität bemühen, möglichst aussagekräftige Ergebnisse präsentieren und die Limitationen ihrer Untersuchungen offen diskutieren. Die Leser einer Publikation sollten sich von kleinen p-Werten nicht blenden lassen, sondern das Studiendesign kritisch beurteilen und abschließend versuchen, eine Antwort auf die Frage zu finden: Für welche Personengruppe oder für welchen Forschungsgegenstand sind die beschriebenen Ergebnisse von Nutzen?

Diese Ausführungen verdeutlichen: Auch die Konsumenten der Statistik benötigen intellektuelle Fähigkeiten, um aus Zahlen und Diagrammen relevante Information zu extrahieren – sowohl in den hohen Sphären der Wissenschaft als auch in den Niederungen des Alltags.

4 Schlussfolgerungen

4.1 Der Einfluss der Intelligenzforschung auf die Statistik

Die Statistik ermöglichte die Intelligenzforschung, weil mit ihrer Hilfe aussagekräftiger Messinstrumente entwickelt wurden. Andererseits beeinflusste die Intelligenzforschung die Statistik massiv und nachhaltig. Deren Pioniere – Francis Galton und Karl Pearson – gelten als die Erfinder der Korrelation und der Regression. Diese Techniken wurden in den Folgejahren für unzählige Anwendungsbereiche aufgegriffen, insbesondere in den Biowissenschaften, aber auch in der Psychologie und der Ökonomie. Galton war auf Messungen versessen und von der Normalverteilung fasziniert. Dieses Konzept übernahm er von seinem Kollegen, dem Statistiker *Adolphe Quetelet* (1796–1874), der bereits einige Jahre zuvor die Bedeutung dieser Verteilung zur Beschreibung menschlicher Merkmale erkannt hatte. Im Gegensatz zu Quetelet orientierte sich Galton jedoch nicht an einem Durchschnittsmenschen, sondern richtete seine Aufmerksamkeit auf die Variabilität der Individuen. Während Quetelet größere Abweichungen eines Messwerts vom Mittelwert als fehlerhaft oder abnorm ansah (tatsächlich nannte man die Normalverteilung zu Quetelets Zeiten „la loi des erreurs“), nutzte Galton dieses Konzept, um eine Klassifizierung zu ermöglichen. Zu diesem Zweck führte er neue Lagemaße ein, nämlich den Median und die Quartile, die eine Population in zwei bzw. vier gleichgewichtige Subgruppen zerlegen. Als Streuungsmaß bevorzugte er den Quartilsabstand als die Breite des Intervalls, in dem die mittleren 50% der Messwerte liegen.

Die von Galton und Pearson ersonnenen Verfahren haben sich als Standardwerkzeuge der Statistik etabliert. Pearson gilt als Wegbereiter der Biometrie, jener wissenschaftlichen Disziplin, die sich mit der Messung von physiologischen Merkmalen an Lebewesen und den dazu gehörenden Analysemethoden befasst. Galton kommt das Verdienst zu, Fragestellungen der Intelligenzforschung erstmals thematisiert und die Bedeutung der Normalverteilung zur Beschreibung einer Population erkannt zu haben.

Auch Spearman beeinflusste die statistische Wissenschaft maßgeblich. Seine Beiträge zur Intelligenzforschung lieferten nicht nur den nach ihm benannten Rang-Korrelationskoeffizienten, sondern führte auch zur Entwicklung der Faktorenanalyse [13]. Damit gründete er die Psychometrie – eine Disziplin, die sich der Entwicklung von Messinstrumenten zum Erfassen von Persönlichkeitseigenschaf-

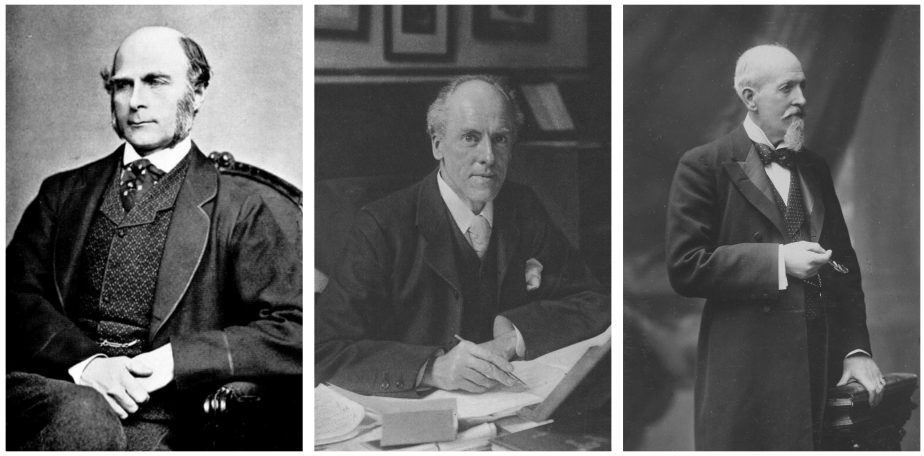

Abbildung 2: Pioniere der Intelligenzforschung: Francis Galton (Naturforscher), Karl Pearson (Mathematiker) und Charles Spearman (Psychologe). Bildquelle Galton: https://de.wikipedia.org/wiki/Francis_Galton, Bildquelle Pearson: https://de.wikipedia.org/wiki/Karl_Pearson; Bildquelle Spearman: https://de.wikipedia.org/wiki/Charles_Spearman

ten widmet. Es gelang ihm, den allgemeinen Gruppenfaktor *g* zu extrahieren und die Ergebnisse der Intelligenztests auf wenige Variablen zu reduzieren.

Während Spearmans Modell darauf basiert, dass jede kognitive Leistung sich aus einem allgemeinen und einem davon unabhängigen spezifischen Faktor zusammensetzt und dass all diese Faktoren voneinander unabhängig sind, wählte Spearmans Kollege *Louis Leon Thurstone* (1887–1955) einen etwas komplexeren Ansatz. Er identifizierte sieben Primärfaktoren, die die Grundlage der menschlichen Intelligenz darstellen sollten und nicht notwendigerweise als unabhängig zu betrachten wären: Zahlenrechnen, Sprachverständnis, räumliches Vorstellungsvermögen, Gedächtnis, schlussfolgerndes Denken, Wortflüssigkeit und Auffassungsgeschwindigkeit. Damit verallgemeinerte er Spearmans Methode auf eine multivariate Faktorenanalyse. Diese Technik gilt heute als ein universell verwendbares Tool, mit dessen Hilfe es möglich ist, von messbaren Erscheinungen auf deren nicht direkt sichtbare Ursachen zu schließen.

4.2 Anwendungen und Limitationen von Intelligenztests

Mit dem Test von Binet und Simon stand erstmals ein Messinstrument zur Erfassung der Intelligenz zur Verfügung. Dieses Konzept war zu Beginn des 20. Jahrhunderts eine große Errungenschaft, das in der Folgezeit mehrfach überarbeitet und neuen Erkenntnissen angepasst wurde. Es findet nach wie vor breite Anwendung – ungeachtet der Tatsache, dass sich Wissenschaftler bislang nicht auf eine einheitliche, umfassende Definition von Intelligenz einigen konnten. Deren Quantifizierung mittels einer normalverteilten Variablen hat wesentlich dazu beigetragen, dass sich die Intelligenzforschung als ein Teilbereich der Psychologie etabliert hat, das Ursachen und Auswirkungen der verschiedenen Formen von Intelligenz erforscht. Ein bedeutender Vertreter ist der bereits erwähnte Penrose, der in seiner „Colchester-Studie" die Ursachen von Lernschwäche erstmals systematisch untersuchte.

Dennoch sind diese Tests nicht unumstritten. Ein häufig geäußerter Kritikpunkt lautet, dass ein Test die Intelligenz in ihrer Komplexität nicht vollständig erfassen könne. Zuweilen wird gar infrage gestellt, ob ein solcher Test das misst, was er zu messen vorgibt, und ob man sich auf das Ergebnis verlassen kann. Das leicht abgewandelte Zitat des amerikanischen Psychologen *Edwin Boring* (1886–1968) „Intelligenz ist, was ein Intelligenztest misst" vermag keine befriedigende Antwort auf diese Fragen zu geben.

Fakt ist andererseits: Keiner dieser Tests erhebt den Anspruch, die Intelligenz eines Menschen in allen Facetten zu erfassen. Der *IQ* ist – wie jedes statistische Modell – eine vereinfachte und damit unvollkommene Abbildung der Wirklichkeit. Das Testergebnis informiert darüber, inwieweit die getestete Person in der Lage ist, bestimmte Aufgaben zu lösen, und ist insofern durchaus aussagekräftig. Darüber hinaus lassen sich sehr pragmatische Gründe für die Beliebtheit dieser Tests anführen. Wie sollten individuelle Leistungseinschätzungen oder Intelligenzforschung möglich sein, wenn es kein Instrument gäbe, das objektive und vergleichbare Ergebnisse lieferte?

Um die Tauglichkeit eines Intelligenztests zu bewerten, orientiert man sich üblicherweise an zwei Qualitätskriterien: der Reliabilität und der Validität. Die *Reliabilität* bezeichnet die Verlässlichkeit. Sie gibt an, inwieweit die Testergebnisse reproduzierbar sind, wenn der Test unter ähnlichen Bedingungen wiederholt wird. In mehreren Studien konnte gezeigt werden, dass die Testergebnisse stark korrelieren, wenn eine Gruppe von Individuen zwei Tests mit einem gewissen zeitlichen

Abstand absolvieren (mit Korrelationskoeffizienten zwischen 0,7 und 0,9 [10]). Die *Validität* ist dagegen schwieriger zu bewerten: Dieses Kriterium gibt an, ob der Test wirklich das misst, was er zu messen vorgibt. Immerhin konnte gezeigt werden, dass Spearmans *g*-Faktor mit einer Reihe von Persönlichkeitsmerkmalen positiv korreliert: der Höhe des Schulabschlusses, der Abschlussnote des Studiums, dem beruflichen Erfolg, dem Einkommen, mit der Lebenserwartung und dem allgemeinen Wohlbefinden. Dies spricht für die Validität der Intelligenztests, da ja gemeinhin diese Eigenschaften mit Intelligenz assoziiert werden.

Allerdings ist einschränkend hinzuzufügen, dass der Grad dieser Korrelationen gering bis mittelmäßig ist. Der *IQ* ist ein theoretisches Konstrukt, das nichts darüber aussagt, wie der jeweilige Mensch seine Fähigkeiten einsetzt. Daher hat im Einzelfall das Ergebnis eines *IQ*-Tests bei einem Schulkind nur eine geringe prognostische Relevanz für das spätere Lebensglück. Die Lebenserfahrung lehrt: Intelligenz und gute Bildung sind eben nur ein Teil dessen, was den beruflichen Erfolg und die soziale Anerkennung bestimmt. Wichtige Komponenten sind darüber hinaus andere kognitive Fähigkeiten, die in den üblichen Intelligenztests ausgeblendet werden: Kreativität, Fantasie, Neugier, Disziplin, Beharrlichkeit sowie emotionale und praktische Intelligenz. Die emotionale Intelligenz bezeichnet die Fähigkeit, eigene Emotionen und die Emotionen seiner Mitmenschen wahrzunehmen und angemessen darauf zu reagieren. Bei der praktischen Intelligenz geht es darum, theoretisches Wissen praktisch umsetzen zu können. Diese Eigenschaften sind für ein erfolgreiches Leben ebenso wichtig wie die akademische Intelligenz. Allerdings gibt es bisher kaum Tests, die dies valide überprüfen könnten. Möglicherweise ist darin ein Grund zu sehen, weshalb die Relevanz von Intelligenztests zuweilen überbewertet wird.

4.3 Intelligenz und Umwelt

Es soll nicht unerwähnt bleiben, dass die Theorien von Galton und Pearson auch extrem negative Auswirkungen hatten. In der irrigen Annahme, dass Umweltfaktoren allenfalls eine sehr geringe Auswirkung auf die Persönlichkeit haben, schwebte den beiden Wissenschaftlern vor, die angeborenen Eigenschaften der „menschlichen Rasse“ durch Auslöschen von unerwünschtem Erbgut zu optimieren. Diese Idee verbreitete sich rasch und fand weltweit Anhänger. Unter den Nationalsozialisten erfuhr diese „wissenschaftlich fundierte“ Eugenik schließlich

ihre perfide Perfektion. Mittlerweile sind Galtons und Pearsons Theorien längst widerlegt.

Im Rahmen der Zwillingsforschung wurde nachgewiesen, dass die *IQ*-Werte von eineiigen Zwillingen (die bekanntlich die gleiche DNA aufweisen) enger beieinander liegen als die von zweieiigen Zwillingen, und dass diese wiederum ähnlicher sind als die Ergebnisse *IQ*-Werte von normalen Geschwistern. Diese Ergebnisse offenbaren, dass Intelligenz sowohl erblich bedingt ist als auch von Umwelteinflüssen bestimmt wird. Damit wurde die Theorie Galtons, der mutmaßte, Intelligenz sei allein erblich bedingt, mit seinen eigenen Methoden (nämlich der Anwendung von Korrelationskoeffizienten) widerlegt. Heute herrscht allgemeiner Konsens darüber, dass die Intelligenz und andere komplexe Persönlichkeitsmerkmale von einer Vielzahl von Genen und diversen Umweltfaktoren beeinflusst werden, und dass gerade diese Vielfalt für das Überleben einer Population die Grundvoraussetzung bildet.

Statistisch lässt sich dieses Zusammenwirken folgendermaßen beschreiben: Die Varianz der Intelligenz (das Quadrat der Standardabweichung) setzt sich additiv zusammen aus der genetisch bedingten und der umweltbedingten Varianz. Die Erblichkeit ist definiert als der Anteil der genetisch bedingten Varianz an der Gesamtvarianz. Sie bezieht sich demnach auf die Variabilität *innerhalb* einer Population, nicht jedoch auf Intelligenzunterschiede *zwischen* mehreren Subgruppen. So besagt eine Erblichkeit von 50% (wie sie für Mitteleuropa nachgewiesen wurde), dass Intelligenzunterschiede innerhalb einer Population zur Hälfte genetisch bedingt sind. Dieser Wert bedeutet jedoch keineswegs, dass ein Individuum jeweils die Hälfte seiner Intelligenz seinen Genen bzw. der Umwelt zu verdanken hat.

Diese Erkenntnis hat weitreichende Konsequenzen. Unterschiede zwischen sozialen Schichten bezüglich des durchschnittlichen *IQs* lassen sich nicht allein durch genetische Einflüsse erklären (wie Galton vermutete). Vielmehr ist anzunehmen, dass dafür unterschiedliche Lebensbedingungen, die in Bildung, Gesundheit oder Wohlstand offensichtlich werden, ausschlaggebend sind.

Dass Umweltfaktoren eine große Bedeutung zukommt, wird eindrucksvoll durch den sogenannten Flynn-Effekt bestätigt (benannt nach seinem Entdecker, dem Politologen James Robert Flynn, 1934-2020). Dieser Effekt beschreibt, dass sich in Industrienationen der durchschnittliche *IQ* seit Beginn der Messungen innerhalb von einer Generation um 5 bis 25 Punkte erhöht hat [6]. Nun ist kaum anzunehmen, dass heute lebende Menschen ihren Vorfahren bezüglich ihrer geis-

tigen Fähigkeiten haushoch überlegen sind. Eine plausiblere Erklärung wäre, dass – neben verlängerten Schulzeiten, besserer Ernährung und medizinischer Versorgung – die spezifischen Fertigkeiten, die in *IQ*-Tests gefragt sind (etwa bezüglich Sprachverständnis, logisch-analytischer Denkweise oder Arbeiten unter Zeitdruck), heutzutage besser antrainiert werden als in früheren Zeiten. Seit einigen Jahren scheint dieser Effekt jedoch zu stagnieren [9]. Die Ursachen für dieses Phänomen sind nicht abschließend geklärt. Es ist denkbar, dass dies nicht alle kognitiven Fähigkeiten betrifft. Vielleicht ist das Stagnieren auch darauf zurückzuführen, dass bezüglich Umweltfaktoren in weiten Teilen der Bevölkerung mittlerweile ein Optimum erreicht ist.

Des Weiteren konnte dargelegt werden, dass der *IQ* im Laufe eines Menschenlebens einigermaßen stabil bleibt. Beim Vergleich der Testergebnisse, die bei denselben Personen als Schulkinder und als Erwachsene ermittelt wurden, ergaben sich Korrelationskoeffizienten von bis zu 0,7 [5]. Andererseits wird deutlich, dass der *IQ* kein absolutes Persönlichkeitsmerkmal ist. Tatsächlich wurde in zahlreichen Studien gezeigt, dass Umweltfaktoren wie Erziehung, Bildung, soziales Umfeld, kulturelle Einflüsse oder gezielte Förderung den *IQ* passiv beeinflussen können – in negativer und in positiver Richtung. Ferner kann jedes Individuum durch sportliche Betätigung, gesunde Ernährung oder dem Erwerb einer neuen FähigkeitFähigkeit (z. B. dem Erlernen einer Fremdsprache) aktiv der Verminderung seines *IQs* entgegenwirken.

4.4 Abschließende Bemerkungen

Es ist sicherlich kein Zufall, dass sich viele Statistiker mit dem Messen von Intelligenz befassten. Einerseits ließen sie sich von der Statistik inspirieren; andererseits haben sie der Statistik enorme Impulse verliehen.

In letzter Zeit wird zunehmend häufiger über Schlagworte wie *Big Data*, *Data Science*, *Künstliche Intelligenz* (KI) oder *Machine Learning*, die vielfach synonym verwendet werden, debattiert. Unter Big Data versteht man – wie der Name vermuten lässt – eine immens große Datenmenge, die häufig aus mehreren Datenquellen zusammengefügt ist und in kurzer Zeit zu analysieren ist. Künstliche Intelligenz ist eine leistungsstarke Software, deren Ziel darin besteht, Muster in der Datenvielfalt zu erkennen. „Machine Learning“ ist ein Teilgebiet der KI, das darauf abzielt, selbstständig Lösungen zu entwickeln und Entscheidungen zu treffen. Nichtsdestotrotz basieren alle diese Techniken auf von Menschen ent-

worfenen Algorithmen, und für die Interpretation der aus statistischen Analysen hervorgehenden Ergebnisse ist allein der Mensch verantwortlich.

Es bleibt festzuhalten: Eine immense Datenmenge, extrem schnelle Hochleistungsrechner, ein optimales Datenmanagement, präzise Mustererkennung und hocheffiziente statistische Analysemethoden sind kein Ersatz für menschliche Intelligenz. Um zu neuen Erkenntnissen zu gelangen, bedarf es nicht nur Intelligenz, sondern einer Menge an zusätzlichen Fähigkeiten: einer soliden Bildung, einer hinlänglichen Erfahrung, der Bereitschaft sich ständig weiterzubilden, ferner Intuition, Scharfsinn, Begeisterung, Teamfähigkeit und des unbedingten Willens, Dingen auf die Spur zu kommen. Dies gilt für alle Forscher, die empirische Studien planen oder durchführen, die Daten analysieren und deren Ergebnisse interpretieren. Dies gilt in gleicher Weise für Journalisten und Autoren wissenschaftlicher Papers, die die daraus hervorgegangenen Erkenntnisse publizieren, sowie für diejenigen, die diese Ergebnisse praktisch umsetzen. Nicht zuletzt sind bei den Konsumenten der Statistik ein klarer Verstand und eine gesunde Skepsis gefragt, damit sie von Zahlen oder Diagrammen nicht geblendet, sondern stattdessen unvoreingenommen informiert werden.

Referenzen

[1] Bland JM, Altman DF: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8); 307–310; 1986.

[2] Boake C: From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. J Clin Exp Neuropsychol 24(3): 383–405; 2002.

[3] Cohen J: A coefficient of agreement for nominal scales. Educ Physiol Meas 20: 37–46; 1960.

[4] Cronbach LJ: Coefficient alpha and the internal structure of tests. Psychometrika 16: 297–332; 1951.

[5] Deary IJ: The stability of intelligence from childhood to old age. Curr Dir Psychol Sci 23(4): 239–245; 2014.

[6] Flynn JR: Massive IQ gains in 14 nations: What IQ tests really measure. Psychol Bul, 101(2): 171–191; 1987.

[7] Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S: Helping doctors and patients make sens of health statistics. Psychocol Sci Public Interest 8(2): 53–96; 2007.

[8] Hesse C: Schwarmintelligenz. In: Warum Mathematik glücklich macht, Kapitel 122. Verlag C.H. Beck, 4. Auflage; 2013.

[9] Schaarschmidt T: Flynn-Effekt. Warum die Intelligenz nicht weiter steigt. In: Spektrum Psychologie, 2; 2019.

[10] Schuerger JM, Witt AC: The temporal stability of individually tested intelligence. J Clin Psychol 45(2): 294–302; 1989.

[11] Shrout PE, Fleiss JL: Intraclass Correlation: Uses in assessing rater reliability. Psychol Bull 1: 30–46; 1973.

[12] Spearman C: „General Intelligence“, objectively determined and measured. Am J Physiol 15: 201–293; 1904.

[13] Spearman C: The theory of two factors, Psychol Rev 21: 101–115; 1914.

[14] Stern W: Die psychologischen Methoden der Intelligenzprüfung und deren Anwendung an Schulkindern. Verlag von Johann Ambrosius Barth, Leipzig; 1912.

[15] Weiß C: Statistik kritisch beleuchtet. In: Kupka MS: Reproduktionsmedizin. Zahlen und Fakten für die Beratung, Kapitel 14. Verlag Urban & Fischer Verlag, München; 2021.

[16] Wolf C: Mehr Hirn, mehr Köpfchen. Spektrum.de https://www.spektrum.de/news/mehr-hirn-mehr-koepfchen/1618626 (abgerufen am 10. April 2021).

Über die Autorin

Christel Weiß ist Professorin für Biomathematik und Epidemiologie an der Medizinischen Fakultät Mannheim der Universität Heidelberg. In ihren Verantwortungsbereich fallen Lehrveranstaltungen für Studierende der Medizin und Masterkurs-Absolventen, Seminare sowie die 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, 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 Koautorin zahlreicher Papers und Buchbeiträge.

Korrespondenzadresse:
Prof. Dr. Christel Weiß
Medizinische Fakultät Mannheim der Universität Heidelberg
Abteilung für Medizinische Statistik und Biomathematik
Theodor-Kutzer-Ufer 1
68167 Mannheim

E-Mail: christel.weiss@medma.uni-heidelberg.de
Homepage: https://www.umm.uni-heidelberg.de/inst/biom/