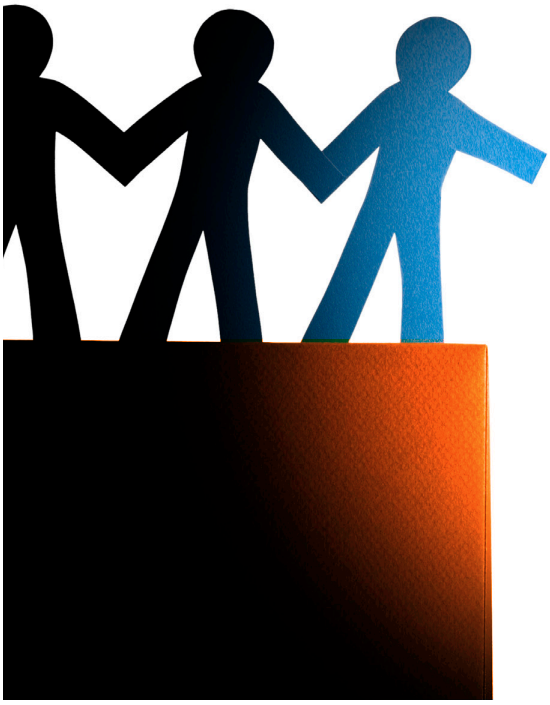
HEIDELBERGER
JAHRBÜCHER
ONLINE
Band 8 (2023)

Gesellschaft der Freunde
Universität Heidelberg e.V.



Krieg, Konflikt, Solidarität

Joachim Funke & Michael Wink (Hrsg.)

HEIDELBERG
UNIVERSITY PUBLISHING

Konflikte und Kontroversen in und um Statistik

CHRISTEL WEISS

Medizinische Fakultät Mannheim der Universität Heidelberg

Zusammenfassung

Die Statistik als eigenständige wissenschaftliche Disziplin und als angewandte Wissenschaft bietet Potenzial für vielerlei Kontroversen. Dieser Beitrag beleuchtet Konflikte, die wegen Statistik entstehen, auf mehreren Ebenen. Zunächst wird dargelegt, dass zwischen Statistikern interkollegiale Divergenzen und Meinungsverschiedenheiten wahrnehmbar sind – sei es wegen Streitigkeiten um Begriffe, wegen unterschiedlicher Vorstellungen bei der Durchführung analytischer Methoden oder bezüglich der Anwendungen von Statistik. Zudem wird darauf eingegangen, wie Statistiker bei der Datenanalyse Konflikte mit Kollegen anderer Fachrichtungen und mit sich selbst austragen – beginnend bei der Studienplanung über die Datenanalyse bis hin zur Interpretation der Ergebnisse. Schließlich wird anhand einiger konkreter Anwendungsbeispiele aufgezeigt, dass Statistik auch gesellschaftlicher Sprengstoff sein kann und Meinungsverschiedenheiten wegen Statistik nicht nur im akademischen Umfeld auftreten.

1 Einleitung

„Krieg, Konflikt und Solidarität“ in Assoziation mit Statistik – eine ungewöhnliche Kombination! Bei dem Wort „Statistik“ denkt man nicht direkt an kriegerische Auseinandersetzungen, bei denen sich verfeindete Staaten oder Gruppen mit militärischen Waffen bekämpfen, um wirtschaftliche, politische oder religiöse Ziele gewaltsam durchzusetzen. Zwar könnte man Statistiken zu Kriegen auflisten, die in

https://doi.org/10.17885/heiup.hdjbo.2023.1.24815

vergangenen Zeiten in zahlreichen Ländern dieser Welt durchgeführt wurden oder aktuell das Weltgeschehen beherrschen, indem man zähl- und messbare Elemente wie die Größe von Armeen, Militärausgaben, Kriegsdauern, Gewinne und Verluste, die Zahl der Opfer oder die vielfältigen Auswirkungen erfasst und quantifiziert. Dabei ließen sich statistische Methoden einsetzen, um diese Informationssammlungen zu verarbeiten, Strukturen zu erkennen oder neue Erkenntnisse zu gewinnen – ebenso wie dies bei politischen Umfragen, in empirischen Wissenschaften, bei der Wettervorhersage, in der Marktforschung oder in der Qualitätskontrolle geschieht. Die Statistik selbst ist jedoch keinesfalls als Ursache dieser Konflikte anzusehen und steht somit in keinem direkten Zusammenhang mit kriegerischen Auseinandersetzungen.

Allerdings bietet die Statistik als eigenständige Wissenschaft durchaus Potenzial – zwar nicht für Kriege, so doch für vielerlei Kontroversen. Obwohl fast jeder Mensch quasi permanent mit den Ergebnissen von statistischen Analysen konfrontiert wird – sei es im privaten Umfeld oder im Rahmen wissenschaftlicher Forschung –, begegnen die meisten Menschen der Statistik skeptisch oder gar ablehnend. Dies zeigt sich in seichten Witzen oder hämischen Bemerkungen wie beispielsweise die dem britischen Premierminister Winston Churchill zugeschriebene Aussage „Traue keiner Statistik, die du nicht selbst gefälscht hast."

Konflikte wegen oder mit Statistik werden auf unterschiedlichen Ebenen ausgetragen. Auf der untersten Ebene geht es um die Statistik als Wissenschaft und um Streitigkeiten unter Statistikern. Dies zeigt sich sowohl im akademischen Kontext (so existieren beispielsweise für den grundlegenden Begriff der Wahrscheinlichkeit mehrere, nicht kongruente Definitionen) als auch in der praktischen Anwendung von Statistik (etwa bezüglich der Durchführung eines statistischen Tests).

Auf einer höheren Ebene tragen Statistiker zuweilen Konflikte mit sich selbst aus. Gründe dafür sind dadurch gegeben, dass die Methodik, nach der eine komplexe Datenmenge analysiert wird, keineswegs stringent vorgegeben ist, und dass das Ergebnis einer statistischen Analyse nicht – wie gemeinhin üblich bei mathematischen oder physikalischen Fragestellungen – definitiv als richtig oder falsch angesehen werden kann.

Auf einer darüber liegenden Metaebene gehören interdisziplinäre Auseinandersetzungen mit Vertretern empirischer Wissenschaften wie beispielsweise der Medizin quasi zum Alltag eines praktisch tätigen Statistikers. Beide Seiten vertreten unterschiedliche Interessen und haben unterschiedliche Erwartungen von

einer effizienten Datenanalyse. So fällt es mitunter schwer, einen gemeinsamen Modus zu finden, um eine Studie zu planen, durchzuführen und deren Ergebnisse adäquat zu interpretieren. Kompromisse sind in aller Regel unausweichlich – beginnend bei der Studienplanung über die Datenanalyse bis hin zur Präsentation und Interpretation der Ergebnisse.

Schließlich sei noch erwähnt, dass Auseinandersetzungen wegen Statistik nicht nur im akademischen Umfeld gepflegt werden. Dies zeigt sich beispielsweise daran, dass sehr viele Zeitgenossen den Ergebnissen von statistischen Analysen (etwa Umfragen vor einer politischen Wahl) häufig mit Skepsis begegnen, oder dass statistische Kenngrößen falsch interpretiert werden und dann zu Irritationen oder unzulässigen Schlussfolgerungen verleiten.

Was sind die Ursachen für dieses Konfliktpotenzial? Welche Folgen hatten die intradisziplinären Auseinandersetzungen auf die Entwicklung der statistischen Wissenschaft und deren Anwendungen? Wie lassen sich die Bedürfnisse aller bei einer Studie beteiligten Wissenschaftler in Einklang bringen? Warum begegnen viele Menschen der Statistik und den Ergebnissen statistischer Analysen mit unverhohlener Skepsis oder gar Feindseligkeit? Diesen Fragen wird in diesem Beitrag nachgegangen.

2 Intradisziplinäre Querelen

Wie in jeder anderen wissenschaftlichen Disziplin sind auch in der Statistik interkollegiale Kontroversen an der Tagesordnung. Das manifestiert sich bei dem für diese Disziplin grundlegenden Begriff der Wahrscheinlichkeit ebenso wie bei der Frage, wie man mit Wahrscheinlichkeiten rechnet und wie das Ergebnis einer statistischen Analyse zu interpretieren ist.

2.1 Historische Entwicklungen

Lange Zeit galt das sichere Wissen, das seit den Zeiten des griechischen Universalgelehrten *Aristoteles* (384–322 v. Chr.) Leitvorstellung des Forschens gewesen war, als hehres Ideal. Erst die Reformation im 16. Jahrhundert und die darauffolgenden Auseinandersetzungen um Glaubensprinzipien sowie die Kulturepoche der Renaissance und die damit verbundene Rückbesinnung auf antikes Gedankengut führten zu einer Befreiung vom ehemals herrschenden religiösen Dogmatismus

und zu der Erkenntnis, dass all unser Wissen begrenzt ist, und dass absolute Sicherheit und zweifelsfreies Wissen unerreichbar sind.

Insofern liegt es nahe, dass Wissenschaftler nach Methoden suchten, um Wahrscheinlichkeiten zu objektivieren und den Zufall unter Kontrolle zu bringen. Als Geburtsjahr der Wahrscheinlichkeitsrechnung wird das Jahr 1654 angesehen, in dem sich die französischen Mathematiker *Blaise Pascal* (1623–1662) und *Pierre Fermat* (1607–1665) in ihrem berühmt gewordenen Briefwechsel mit Gewinnchancen bei Glücksspielen befassten. Einige Jahre später philosophierte Pascal in seinen „Pensées“ über die Chancen, dass Gott existiert, und über die jeweiligen Folgen der Annahme oder Ablehnung dieser These. Seine Schriften zeigen, dass die Wahrscheinlichkeitstheorie zwei Wurzeln hat: Zum einen ging es darum, zufällige Prozesse wie beispielsweise Glücksspiele mathematisch quantitativ zu beschreiben. Zum anderen stand der Begriff „Wahrscheinlichkeit“ für eine neue Denkweise, bei der das bis dahin geltende und niemals hinterfragte Ideal der Sicherheit aufgegeben worden war und die auch Zweifel zuließ. Im Zentrum der neuen Rationalität wurde der Wahrscheinlichkeitsbegriff nicht mehr wie im Mittelalter im Sinne einer durch eine Autorität gestützte Meinung verwendet, sondern bezeichnete einen Grad an Zustimmung, der sich auf vorab erworbenes Wissen oder auf verfügbares Anschauungsmaterial stützte.

Dennoch war der Zufall in der Wissenschaft lange Zeit verpönt. Von einem Wissenschaftler forderte man Klarheit, nicht Unsicherheit. Im deterministisch-physikalischem Weltbild des 19. Jahrhunderts wurde der Zufall nur toleriert, wo es unvermeidbar erschien, etwa in der Thermodynamik, in Darwins Evolutionstheorie oder in der Quantentheorie, die bekanntlich nur Wahrscheinlichkeitsaussagen erlaubt. Die Entwicklung der Wahrscheinlichkeitstheorie erwies sich als sehr komplex und hat innerhalb der eigenen Disziplin und auch in der Zusammenarbeit mit anderen Disziplinen zu heftigen Kontroversen geführt.

2.2 Querelen um Wahrscheinlichkeiten

„Nichts in dieser Welt ist sicher, außer dem Tod und den Steuern.“ Dieses dem US-amerikanischem Staatsmann und Naturwissenschaftler *Benjamin Franklin* (1706–1790) zugeschriebene Diktum drückt aus, dass das allgemeine Lebensgefühl ebenso wie alltägliche Erfahrungen oder wissenschaftliche Erkenntnisse von Zufällen und Unsicherheiten geprägt sind. Ein Mensch, der nur zweifelsfrei wahre Behauptungen aufstellen wollte, wäre alsbald zum Schweigen verurteilt (es sei

denn, er bewegte sich ausschließlich in abstrakten Gefilden wie beispielsweise der reinen Mathematik).

Der Begriff „Wahrscheinlichkeit“ wird seit jeher verwendet, um den Grad subjektiver Gewissheit oder Zweifel über eine Vermutung auszudrücken. Es handelt sich dabei um eine vage Einschätzung, die intuitiv begründet ist und auf der Basis individueller Erfahrungen getroffen wird. Meist betrifft dies Aussagen über die Vergangenheit oder die Zukunft, deren Kausalitäten nicht oder nur unvollständig bekannt sind. Wie kann man sich mit Unsicherheiten arrangieren und Entscheidungen treffen? Im Alltag vertrauen viele Menschen – ohne lange nachzudenken – ihrem gesunden Menschenverstand oder Bauchgefühl, manche auch einem Aberglauben. Für wissenschaftliche Fragestellungen scheinen diese Ansätze ungeeignet zu sein, da die darauf basierenden Entscheidungen subjektiv und die Schlussfolgerungen nicht nachvollziehbar sind.

Lange Zeit verschwendete niemand einen Gedanken daran, diesen Begriff zu präzisieren. Erst ab dem 17. Jahrhundert wurden dafür diverse Konzepte entwickelt, und so entstanden im Laufe der Zeit mehrere Definitionen (Abbildung 1) und eine beachtliche Begriffsvielfalt (Tabelle 1).

Tabelle 1: Modelle zur Erfassung von Wahrscheinlichkeiten.

Modell	**Interpretation**	**zugrundeliegende Annahmen**	**Urheber**	**Anwendungen**
1	klassisch	Elementarereignisse mit gleichen Wahrscheinlichkeiten	Laplace	Glücksspiele
2	frequentistisch	Häufigkeiten	Mises	Demografie
3	epistemisch	Expertenwissen	Bayes	Gerichtspraxis
4	mathematisch	3 Axiome	Kolmogoroff	Berechnungen

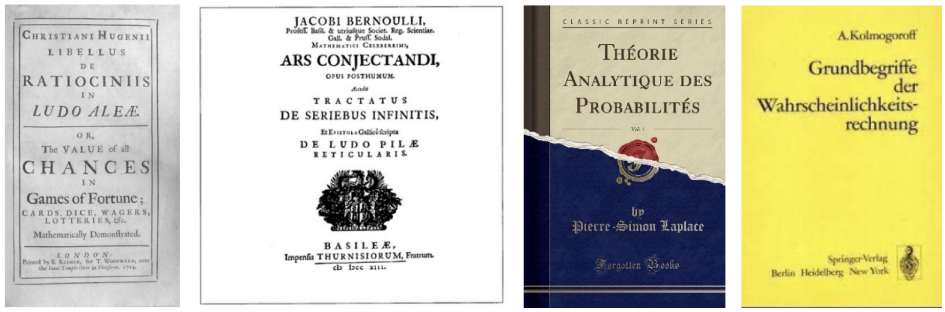


Abbildung 1: Standardwerke der Wahrscheinlichkeitsrechnung. Vlnr: De Ludo Aleae (Huygens, 1657), Ars Conjectandi (Bernoulli, 1713), Théorie Analytique des Probabilités (Laplace, 1812), Grundbegriffe der Wahrscheinlichkeitsrechnung (Kolmogoroff, 1933).

Modell 1: Klassische Wahrscheinlichkeit

Der Briefwechsel zwischen Pascal und Fermat wurde ausgelöst durch eine Wette des Chevaliers de Méré, der der Spielleidenschaft frönte und die beiden Mathematiker um Hilfe bat, seine Spielchancen zu optimieren. Dabei verwendeten Pascal und Fermat nicht den Begriff „Wahrscheinlichkeit“, sondern sie berechneten die zu erwartenden Gewinne bei bestimmten Konstellationen. Zu diesem Konzept haben auch der niederländische Physiker Christiaan Huygens (1629–1695) und der Schweizer Mathematiker Jakob Bernoulli (1655–1705) beigetragen, die sich in ihren Abhandlungen „De Ludo Aleae“ und „Ars Conjectandi“ (erschienen in den Jahren 1657 bzw. 1713) ebenfalls mit Gewinnstrategien bei Glücksspielen befassten [2]. Es sollten jedoch noch 100 Jahre vergehen, ehe der Mathematiker Pierre-Simon Laplace (1749–1827) in seinem Werk „Théorie Analytique des Probabilités“ die erste exakte Definition einer Wahrscheinlichkeit präsentierte: Demnach handelt es sich um den relativen Anteil günstiger Ereignisse bezogen auf die Gesamtzahl aller möglichen Ereignisse. – Der Briefwechsel zwischen Pascal und Fermat gilt trotz der Einschränkungen dieses Konzepts (gleiche Ausgangswahrscheinlichkeiten für alle Elementarereignisse) als der Ausgangspunkt für eine der größten intellektuellen Revolutionen. Mit der Einführung der klassischen Wahrscheinlichkeitsrechnung gelang nämlich der Nachweis, dass Unsicherheit tatsächlich objektivierbar ist.

Modell 2: Frequentistische Wahrscheinlichkeit

Diese Interpretation hatte ihren Ursprung in Listen von Todesfällen und Geburten, die seit dem 16. Jahrhundert in Kirchengemeinden geführt wurden. Der englische Kaufmann John Graunt (1620–1674) erstellte im Jahr 1662 die erste Sterbetafel, in der er für jedes Alter Überlebenswahrscheinlichkeiten basierend auf Einträgen in Kirchenbüchern schätzte. Später verwendete man demografische Daten zur Berechnung von Rentenpreisen und als Grundlage für Lebensversicherungen. 1693 veröffentlichte der Physiker Edmond Halley (1656–1742) die erste empirische Sterbetafel mit statistischen Lebenserwartungen. Um 1750 war die Mathematik der Sterblichkeit, vor allem in der Anwendung auf Rentenbestimmung, die Spitze der Forschung der Wahrscheinlichkeitstheorie [2]. Der Mathematiker Richard von Mises (1883–1953) präzisierte dieses auf Häufigkeiten basierende Konzept: Demnach wird eine Wahrscheinlichkeit für ein Ereignis, das einem zufälligen Prozess unterliegt, über die relative Häufigkeit dessen Auftretens quantifiziert. Dieses Prinzip folgt dem Gesetz der großen Zahlen, nach dem sich die relative Häufigkeit eines Zufallsereignisses um dessen Wahrscheinlichkeit stabilisiert, wenn das zugrunde liegende Zufallsexperiment hinreichend oft durchgeführt wird. Dieses Gesetz zeigt, dass der Zufall – obgleich Einzelereignisse nicht prognostizierbar sind – mathematischen Gesetzen unterliegt.

Modell 3: Epistemische Wahrscheinlichkeit

Das epistemische Verständnis kommt der ursprünglichen Bedeutung einer Wahrscheinlichkeit – nämlich als Grad subjektiver Gewissheit – recht nahe. Dieses Konzept geht zurück auf den englischen Geistlichen Thomas Bayes (1701–1761), der die Wahrscheinlichkeit eines Ereignisses als die Plausibilität für sein Eintreten auffasste. Bei dieser Definition hängt die Schätzung – im Gegensatz zum klassischen oder zum frequentistischen Modell – vom individuellen Kenntnisstand und dem Erfahrungsschatz des Betrachters ab. Sie bietet den Vorteil, dass sich auf Basis neuer Informationen Wahrscheinlichkeiten anpassen und Urteile revidieren lassen. So fand dieses Konzept Anwendung bei rechtlichen Fragen, wenn etwa ein Richter aufgrund von Zeugenaussagen und Beweismaterial ein Urteil fällen musste. Bayes hat eine Formel hinterlassen, die die Berechnung einer bedingten Wahrscheinlichkeit basierend auf einer gegebenen Evidenz erlaubt (wobei für diese a-priori-Wahrscheinlichkeit ein subjektiv geschätzter Wert angenommen werden kann).

Modell 4: Axiomatische Wahrscheinlichkeit

Der russische Mathematiker Andrei Kolmogoroff (1903–1987) definierte diesen Begriff mathematisch abstrakt, indem er anhand von drei Axiomen die notwendigen Eigenschaften einer Wahrscheinlichkeit festlegte. Dieses Konzept schließt die Definitionen der klassischen und der frequentistischen Wahrscheinlichkeit nach Laplace bzw. Mises ein. Das Axiomensystem gibt zwar Rechenregeln vor, vermeidet jedoch eine konkrete Aussage bezüglich des Wesens einer Wahrscheinlichkeit. Kolmogoroffs im Jahr 1933 in deutscher Sprache erschienenes Lehrbuch „Grundbegriffe der Wahrscheinlichkeitsrechnung“ bildet – nahezu 300 Jahre nach dem Briefwechsel zwischen Pascal und Fermat – den Abschluss der Entwicklung der Wahrscheinlichkeitstheorie.

Ein Begriff – vier Definitionen – viele Fragen! Welche Definition eignet sich für welche Fragestellung? Wie ist eine Wahrscheinlichkeit aufzufassen? Objektiv oder subjektiv? Pragmatisch oder mathematisch? Darüber wurde und wird bis heute trefflich gestritten. Klassische Wahrscheinlichkeitstheoretiker von Fermat bis Laplace (Abbildung 2) verwendeten diesen Begriff unbekümmert in mehreren Bedeutungen. In ihren Schriften führten sie aus, dass Wahrscheinlichkeiten objektiv messbar seien. Gleichzeitig stellten sie die Behauptung auf, dass Wahrscheinlichkeiten die Unvollkommenheit menschlichen Wissens reflektierten und dass es „echte“ Zufälle gar nicht gäbe. Insofern gelten Wahrscheinlichkeiten nur vorläufig, weil den Menschen, die sie ermittelten, die verborgenen Quellen und Gründe für bestimmte Entwicklungen unbekannt waren. Diese Ansicht vertraten auch Bevölkerungstheoretiker wie der preußische Feldprediger Johann Peter Süßmilch (1707–1767), der in Geburts-, Heirats- und Sterblichkeitsziffern „eine beständige, allgemeine, große, vollkommene und schöne Ordnung“ erkannte. Demnach gab es in der Frage nach dem Wesen einer Wahrscheinlichkeit eine große Toleranz auf allen Seiten, die es sogar gestattete, dass ein einzelner Wissenschaftler je nach Fragestellung sich die objektive oder die subjektive Betrachtung der Wahrscheinlichkeit zu eigen machte.

Nichtsdestotrotz waren auch Feindseligkeiten spürbar. Das Wort „subjektiv“ avancierte zum Schimpfwort der Frequentisten, die den Bayes’schen Ansatz kritisierten: Für Zeugenaussagen – so ihre Ansicht – dürfe man keine Wahrscheinlichkeiten angeben, weil die Quantifizierung von Unwägbarkeiten generell nicht möglich sei. Den Klassikern unter den Wahrscheinlichkeitstheoretikern warf man

eine „Verwirrung des Intellekts“ vor, weil sie mit ihren stark vereinfachenden Theorien den gesunden Menschenverstand aushebeln würden. Als der französische Mathematiker Siméon Denis Poisson (1781-1840) im Jahr 1837 behauptete, die Wahrscheinlichkeit für das Eintreten sehr selten auftretender Ereignisse berechnen zu können, wurde er von einigen Kollegen heftig attackiert. Der britische Philosoph John Stuart Mill (1806–1873) bezeichnete die Wahrscheinlichkeitsrechnung gar als „Schandfleck der Mathematik“ [3]. Viele Mathematiker schlossen sich dieser Meinung an. Sie konnten eine wissenschaftliche Disziplin, die sich mit Zufällen und unsicheren Ereignissen befasst, nicht in Einklang bringen mit dem Anspruch der Mathematik, eine Aussage unzweifelhaft als „wahr“ oder „falsch“ einzuordnen. Andere hatten Vorbehalte, weil sie den Ursprung der Wahrscheinlichkeitsrechnung – nämlich Glücksspiele – als unseriös empfanden.

Die unterschiedlichen Interpretationsmöglichkeiten einer Wahrscheinlichkeit stehen auch heute noch miteinander in Konflikt: Für einen Klassiker à la Laplace eignen sich Wahrscheinlichkeiten nur für Systeme oder Gegenstände, deren Struktur bekannt ist (etwa ein Würfel); Frequentisten benötigen hinreichend viele Daten, um Wahrscheinlichkeiten zu ermitteln (die dann aber nur sehr eingeschränkt gelten); ein Subjektivist à la Bayes kann dagegen für alle Forschungsgegenstände Wahrscheinlichkeiten mutmaßen (auch für Ereignisse, die noch niemals beobachtet wurden). Einen Mathematiker à la Kolmogoroff braucht dies nicht zu kümmern: Dessen Axiome sind universell anwendbar – unabhängig davon, wie die Wahrscheinlichkeiten ermittelt oder aufgefasst werden.

Abbildung 2: Pioniere der Wahrscheinlichkeitsrechnung mit unterschiedlichen Konzepten. Vlnr: Pierre-Simon Laplace (klassisch), Richard von Mises (frequentistisch), Thomas Bayes (epistemisch), Andrei Kolmogoroff (axiomatisch).

2.3 Querelen um die Durchführung statistische Tests

Geburten und Sterblichkeitstabellen waren nicht nur eine der ersten systematisch angelegten Datenbanken in Europa, sondern auch Grundlage für den ersten statistischen Test, durchgeführt vom schottischen Arzt John Arbuthnot (1667–1735) im Jahre 1710 [2]. Dieser hatte beobachtet, dass in jedem von 82 Jahren (auf deren Daten er Zugriff hatte) Knabengeburten häufiger verzeichnet wurden als Mädchengeburten. Für ein einzelnes Jahr müsste diese Wahrscheinlichkeit (falls nur der Zufall maßgebend wäre) 1⁄2 betragen. Da seine Beobachtung unter dieser Prämisse extrem unwahrscheinlich ist, schloss er reinen Zufall aus und sah seine Beobachtung als Beweis einer göttlichen Fügung an. Auch wenn wir als aufgeklärte Menschen im 21. Jahrhundert diese Schlussfolgerung nicht teilen, kann Arbuthnots Methodik als revolutionär bezeichnet werden: Um eine Hypothese zu belegen, verwendete er nicht (nur) rhetorische Mittel, sondern überprüfte sie auf der Basis empirischer Daten. Einen p-Wert verwendete er jedoch nicht, ebenso wenig wie 200 Jahre später der englische Statistiker William Sealy Gosset (1876–1937), der im Jahre 1908 unter dem Pseudonym „Student“ seinen t-Test zum Vergleich von Mittelwerten publizierte.

In den 1920er Jahren brach wegen statistischer Tests ein offener Konflikt zutage, an dem vier Statistiker beteiligt waren: Karl Pearson (1857–1936; bekannt durch den nach ihm benannten Korrelationskoeffizienten), dessen Sohn Egon Pearson (1895–1980), außerdem Jerzy Neyman (1894–1981) und Ronald Aylmer Fisher (1890–1962). Fisher führte den p-Wert als Maß der Evidenz gegen die Nullhypothese ein. Er erkannte die Bedeutung von Gossets Arbeit und erstellte Tabellen mit Quantilen für die Prüfgröße des t-Tests, anhand derer man entscheiden konnte, ob das Testergebnis signifikant auf dem 5%- oder dem 1%-Signifikanzniveau war. Fisher hatte jedoch keineswegs beabsichtigt, mit dem p-Wert eine Größe zu generieren, die die Relevanz des Ergebnisses einer statistischen Analyse quantifiziert. Den Begriff „signifikant“ verwendete er im Sinne von „auffällig“ oder „schwer mit dem Zufall vereinbar“.

Jerzy Neyman war Fishers stärkster Konkurrent (Abbildung 3). Dessen Teststrategie basierte auf einem a-priori festgelegten Signifikanzniveau α. Der Nullhypothese wird eine konkrete Alternativhypothese gegenübergestellt. Dies ermöglicht die Definition eines Akzeptanz- und eines kritischen Bereichs für die Nullhypothese, die Berechnung der Power und des α-Fehlers. Die Nullhypothese wurde abgelehnt, wenn die Prüfgröße in den kritischen Bereich fiel. Den p-Wert ließ

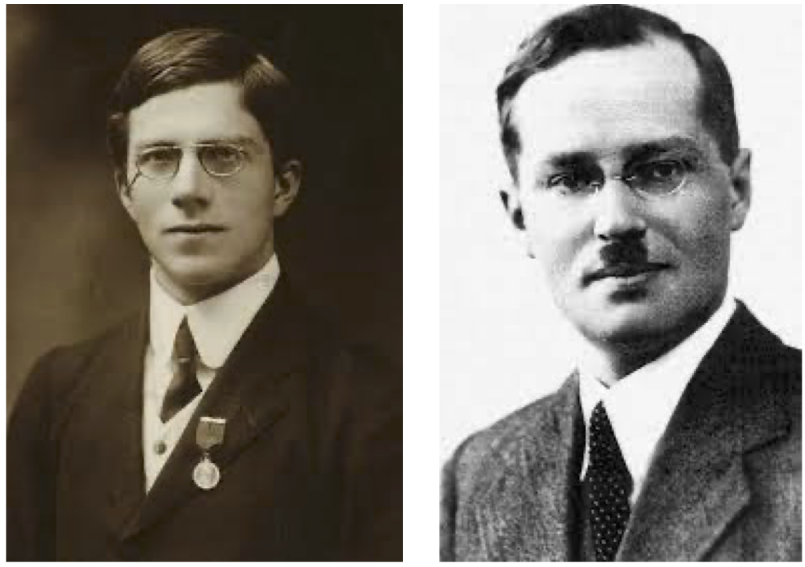


Abbildung 3: Kontrahenten wegen statistischer Tests: Ronald Aylmer Fisher und Jerzy Neyman.

Neyman aus dem Spiel [6]. Auch das Ergebnis eines statistischen Tests interpretierten Fisher und seine Konkurrenten in unterschiedlicher Weise: Fisher stellte an statistische Tests den Anspruch, neue Erkenntnisse über die Wahrheit einer Hypothese zu gewinnen (kognitivistischer Ansatz). Seine Kollegen hatten diesbezüglich eine pragmatischere Einstellung: Für sie diente das Testergebnis lediglich als eine subtile Entscheidungshilfe im Fall von Ungewissheit (dezisionistischer Ansatz).

Fisher und Neyman beschimpften sich auf unflätige Weise: Fisher bezeichnete Neymans Arbeit als „Graus für die abendländische Freiheit des Denkens“, Neyman nannte Fishers Ideen „schlimmer als nutzlos“. Auch mit den beiden Pearsons hatte Fisher heftige Auseinandersetzungen. Sie blockierten Publikationen von Fisher in der Zeitschrift „Biometrika“, deren Mitbegründer Karl Pearson war.

Wenn zwei sich streiten, freut sich der dritte: Einige Statistiker vermengten die beiden Ansätze, indem sie Fishers p-Wert und Neymans Signifikanzniveau nutzten, um eine eindeutige Testentscheidung zu treffen. So wurde der p-Wert auf ein Podest gehoben, auf dem er noch heute steht. Viele Anwender sehen den Schwellenwert $p = 0{,}05$ als eine Schranke, die Bedeutsames von Belanglosem trennt – was weder im Sinne Fishers noch im Sinne Neymans sein dürfte.

Bezüglich der Anwendung statistischer Tests gibt es bis heute noch einen anderen Kriegsschauplatz. Der oben beschriebene klassische Ansatz zum Testen von Hypothesen basiert auf empirisch gewonnenen Daten. Hin und wieder wird dafür plädiert, stattdessen die Bayes-Statistik zu verwenden. Dieser Ansatz berücksichtigt, was man a priori über die Fragestellung weiß oder annimmt. Dabei fließen also auch Informationen ein, die nicht aus dem Datensatz stammen – ganz im Sinne von Bayes, der Wahrscheinlichkeit als Plausibilität für das Eintreten eines Ereignisses ansah. Die zugrunde liegenden Wahrscheinlichkeiten lassen sich allerdings nur subjektiv schätzen (was von Vertretern der klassischen Statistik strikt abgelehnt wird). Andererseits gestattet die Bayes-Statistik, Hintergrundwissen und Erfahrungen einfließen zu lassen und Testentscheidungen zu revidieren.

Der klassische Ansatz wird im Allgemeinen in den Bio-, Sozial- und Wirtschaftswissenschaften bevorzugt [1,7,10], während der Ansatz nach Bayes eher in den Bereichen der Technik und der Künstlichen Intelligenz verwendet wird [8]. In der Fachliteratur werden diese beiden Denkweisen mitunter als verfeindete Glaubensrichtungen dargestellt. Befürworter des Bayes-Ansatzes verweisen gerne darauf, dass die getroffenen Entscheidungen näher an der Realität liegen. Skeptiker befürchten, dass mit Bayes Willkür in statistische Analysen Einzug halten könnte.

2.4 Querelen um Anwendungen

Im Jahre 1656 (also zwei Jahre nach dem Briefwechsel zwischen Pascal und Fermat) erschien das Buch „Teutscher Fürstenstaat“ des Historikers Veit Ludwig von Seckendorff (1626–1692). Im Jahre 1747, also fast 100 Jahre später, hielt der Historiker Gottfried Achenwall (1719–1772) eine Vorlesung zum Thema „Staatenkunde“, die er mit dem Titel „Statistik“ versah. Diese beiden Ereignisse gründeten die deutsche Universitätsstatistik, deren Aufgabe darin bestand, Informationen zur Beschreibung der Gesellschaft, der Politik, der Wirtschaft, der Verwaltung und geografischer Besonderheiten zu sammeln. Deren Beschreibung erfolgte – der Not gehorchend, weil Zahlenmaterial kaum zur Verfügung stand – überwiegend qualitativ [5]. Zeitgleich bediente man sich in England der Politischen Arithmetik, mit der basierend auf quantitativen Informationen aus Verwaltungsarchiven und Kirchenbüchern demografische Daten tabellarisch aufbereitet und sogar statistisch analysiert wurden. Berühmte Vertreter dieser Richtung waren die Abschnitt 2.2 erwähnten John Graunt und Edmond Halley. Achenwall und seine Anhänger verabscheuten diese Methoden und waren dabei in ihrer Wortwahl nicht zimperlich: Sie

verspotteten ihre britischen Kollegen als „Tabellenknechte“, die eine „gemeine“ Statistik betrieben. Zu diesen „Tabellenknechten“ zählte auch ihr deutscher Kollege Johann Peter Süßmilch, dessen bahnbrechendes Werk der Bevölkerungsstatistik 1741 erschien (Abbildung 4).

Ein vehementer Befürworter und Anwender der politischen Arithmetik war der belgische Statistiker Adolphe Quetelet (1796–1874). Er versuchte, spezielle Charakteristika des sozialen Lebens zu ergründen, indem er zahlreiche Eigenschaften von Individuen quantitativ erfasste. Er beschränkte sich dabei nicht auf einfach zu messende physische Größen wie etwa die Körpergröße, sondern untersuchte auch moralische und charakterliche Besonderheiten wie etwa den Hang zur Kriminalität. Dabei entdeckte er, dass viele dieser Merkmale normalverteilt sind. Mit den gesammelten Daten entwickelte er den „statistischen Durchschnittsmenschen“ als abstraktes Wesen. Während wirkliche Einzelmenschen zu zahlreich und zu verschieden sind und in ihrem Wesen zu komplex erscheinen, als dass sich mit ihnen spezielle Eigenschaften und Besonderheiten der menschlichen Spezies und der Gesellschaft im Ganzen erschließen ließen, ist der Durchschnittsmensch leicht zu verstehen. Dieser sollte dazu dienen, einen individuellen Messwert einzuordnen, ohne die jeweilige Person sozial oder moralisch zu werten. In England befasste sich der Naturforscher Francis Galton (1822–1911), angeregt durch seinen Cousin Charles Darwin, mit den Grundlagen der Vererbungslehre und entwickelte dabei Methoden zur Messung von menschlichen Eigenschaften. Er bereicherte die Statistik, indem er beispielsweise den Median und die Regressionsanalyse einführte. Anders als Quetelet sprach er jedoch von „mittelmäßigen“ Menschen; unterdurchschnittlich war für ihn gleichbedeutend mit „minderwertig“.

Abbildung 4: Anwender mit unterschiedlicher Methodik und ihre Werke. Vlnr: Gottfried Achenwall (Staatskundler, qualitative Beschreibung); Johann Peter Süßmilch (Demograf, politische Arithmetik).

Karl Pearson war von Galtons Untersuchungen beeindruckt und setzte in den folgenden Jahren alles daran, mittels statistischer Methoden das Prinzip der Vererbung und das menschliche Verhalten zu erforschen. Dabei wurde er unterstützt von Walter Frank Raphael Weldon (1860–1906), einem Zoologen, der die Mendelschen Regeln mit statistischen Methoden zu widerlegen versuchte. In der Royal Society bildeten sich nun zwei verfeindete Lager: Auf der einen Seite gab es die Anhänger Galtons und Weldons, für die nicht nur Gene, sondern auch das Umfeld den Phänotyp bestimmen; auf der anderen Seite standen die Mendelianer, die es streng ablehnten, mit statistischen Methoden Aussagen über biologische Sachverhalte herzuleiten. Die Debatten wurden erbittert geführt. Dies führte schließlich zur Gründung der Zeitschrift „Biometrika“ durch Galton, Karl Pearson und Weldon.

3 Konflikte mit anderen

3.1 Interdisziplinäre Auseinandersetzungen

Statistische Analysen werden in aller Regel nicht um ihrer selbst willen betrieben. Wissenschaftliche Zeitschriften berichten gerne über aufsehenerregende Studien oder neue Erkenntnisse, die mit Begriffen wie „statistisch signifikant“ belegt werden. Nicht immer sind sich die Herausgeber von Zeitschriften, die Autoren dieser Publikationen, deren Begutachter oder Leser über die Bedeutung dieses Begriffs im Klaren. „Statistisch signifikant“ bedeutet lediglich, dass das Ergebnis eines statistischen Tests schwerlich allein durch den Zufall zu erklären ist. Dies ist keineswegs gleichbedeutend mit „wissenschaftlich brisant“ oder „praktisch relevant“. Anders formuliert: Es gibt auch Ergebnisse mit einem p-Wert über 0,05, die Beachtung verdienen, weil sie möglicherweise eine neue Erkenntnis in sich bergen, und ebenso signifikante Ergebnisse, die rein zufällig zustande gekommen sind oder aus anderen Gründen irrelevant sind.

Nichtsdestotrotz bietet ein statistisch signifikantes Ergebnis eine weit größere Chance zur Veröffentlichung als ein nicht-signifikantes. Wissenschaftler in empirischen Fächern wie beispielsweise der Medizin benötigen in der Regel zahlreiche Publikationen, um als erfolgreich zu gelten. Es gibt einige Tricks, um ein statistisch signifikantes Ergebnis zu erhalten: Man führt eine Vielzahl von statistischen Tests durch, indem man zahlreiche Variablen untersucht, zahlreiche Subgruppen oder Messungen zu unterschiedlichen Zeitpunkten miteinander vergleicht. Derlei wissenschaftliches Fehlverhalten erstreckt sich von „p-Hacking“ (dem Jagen nach

Signifikanz) bis hin zu schweren Betrugsfällen. Eine weitere Möglichkeit, ein signifikantes Ergebnis zu erzielen, ist das sogenannte „HARKing“ (Hypothesizing After the Results are Known): In diesen Fällen wird – nachdem alle Ergebnisse vorliegen und die Analyse der primären Zielgröße nicht das erhoffte Ergebnis geliefert hat – die Forschungsfrage umformuliert. Diese Tricks führen letztlich dazu, dass signifikante Ergebnisse in der Literatur überrepräsentiert sind. Der Epidemiologe John Ioannidis (*1965) argumentiert gar, dass die meisten publizierten Ergebnisse der biomedizinischen Forschung falsch seien [4].

Aus diesen Gründen geraten Statistiker bei der Zusammenarbeit mit Vertretern einer anderen Wissenschaft leicht in einen Zielkonflikt: Sie sind in aller Regel bestrebt, die Daten effizient zu analysieren, während die Empiriker in erster Linie an einem signifikanten Ergebnis interessiert sind. Hin und wieder ergeben sich persönliche Konflikte, etwa dann, wenn Statistiker als unliebsame Konkurrenten wahrgenommen werden, oder wenn sie lediglich als Dienstleister angesehen werden. Ferner können Beurteilungskonflikte hinzukommen: Statistiker schätzen die Bedeutung eines p-Werts anders ein als Studienleiter, die ihre Hypothese bestätigt sehen möchten.

Um derartige Konflikte zu vermeiden, ist es sinnvoll, bei der Durchführung einer Studie wie beispielsweise in der Medizin einen Statistiker bereits in der frühen Planungsphase zu involvieren. Grundsätzlich sollte zunächst konkretisiert werden: Welche Fragestellung liegt der Studie zugrunde? Welche Rahmenbedingungen sind zu beachten? Welches Studiendesign ist vorgesehen (retrospektiv oder prospektiv, randomisiert oder beobachtend)? Bei einer prospektiven Studie sollte vor der Datenerhebung geklärt werden: Welche primäre Zielgröße ist vorgesehen? Welcher Effekt wird erwartet oder als relevant erachtet? Welche Merkmale sollten erhoben werden? Bei einer retrospektiven Studie ist basierend auf dem bereits vorhandenen Datenmaterial zu klären, ob und mit welcher Power ein Effekt nachweisbar ist. Für den Statistiker sind eine hohe Fallzahl, möglichst homogene Stichproben und ein randomisiertes Studiendesign ideal; praktisch sind diese Bedingungen aber häufig nicht oder nur eingeschränkt realisierbar. Daher sind Kompromisse auf beiden Seiten unausweichlich. Diese sollten thematisiert werden: So werden bereits im Vorfeld Erwartungen konkretisiert, Möglichkeiten ausgelotet und Grenzen aufgezeigt.

Statistiker führen zuweilen auch Kämpfe mit sich selbst. Die Güte eines statistischen Modells wird nicht nur von der Expertise des Statistikers, sondern auch von

den gegebenen Rahmenbedingungen wie etwa der praktischen Durchführbarkeit einer Studie oder der Qualität der zur Verfügung stehenden Daten bestimmt. Das Ergebnis hat aber letzten Endes der Statistiker zu verantworten. So muss er sich fragen: Welche Analysemethoden sind am ehesten geeignet? Sind deren Voraussetzungen erfüllt? Wie lassen sich die Ergebnisse angemessen präsentieren? Wie werden Probleme (etwa fehlende Daten) gehandhabt? Könnte das Ergebnis durch einen Bias verzerrt sein? Ist das Ergebnis kausal bedingt, oder ist es möglicherweise durch Confounding verursacht? Welche Kompromisse sind vertretbar? Freilich wird jeder Statistiker bemüht sein, eine Hypothese – falls möglich – zu bestätigen. Es würde jedoch gegen sein Berufsethos verstoßen, nicht adäquate Methoden anzuwenden oder gar Daten zu manipulieren. Also muss ein Statistiker auch mit sich selbst ringen, ehe er ein finales Modell zu seiner eigenen Zufriedenheit entwickelt hat. Am Ende profitieren alle Beteiligten, wenn die Ergebnisse der Analyse gemeinsam aus verschiedenen Blickwinkeln begutachtet und interpretiert werden.

3.2 Statistik als gesellschaftlicher Sprengstoff

Die Statistik entwickelte sich langsamer als andere mathematische Disziplinen, da sie mit schwerwiegenden Problemen zu kämpfen hatte. Diese sind teilweise durch die Eigentümlichkeiten und die schwere Erfassbarkeit der Begriffe „Zufall“ und „Wahrscheinlichkeit“ bedingt. Darüber hinaus sind sie auch durch Vorbehalte aus anderen Wissenschaften wie Theologie oder Philosophie begründet, in denen eher von Schicksal oder Fügung gesprochen wird. Selbst von Seiten der Mathematik werden Vorbehalte geäußert, weil sich das Phänomen des Zufalls nur sehr schwer quantifizieren lässt und das Befassen mit Unsicherheiten manchen Mathematikern widerstrebt. Andere Wissenschaftler argumentieren, dass Zufall nur ein Synonym für menschliche Unvollkommenheit sei, weil der Mensch nicht in der Lage ist, komplexe physikalische Ereignisse in ihrer Gänze zu erfassen. Der Physiker Albert Einstein (1879–1954) formulierte dies treffend mit dem Satz „Gott würfelt nicht“.

Nach wie vor spielt Statistik eine große Rolle als Mittel der Kommunikation und der Argumentation. Doch auch aus der Gesellschaft schlägt der Statistik häufig große Skepsis entgegen. Argumente, die auf statistischen Analysen basieren, stoßen mitunter auf großes Misstrauen, das meist unverhohlen zum Ausdruck gebracht wird. Dazu einige Beispiele.

Beispiel 1: Wahlumfragen

Den Ergebnissen von Wahlumfragen begegnen viele Zeitgenossen mit großer Skepsis. Grund dafür bieten Wahlumfragen aus der Vergangenheit, deren Prognosen sich im Nachhinein als unzutreffend erwiesen hatten. Außerdem erscheint es vielen Menschen unverständlich, wie es möglich sein sollte, eine präzise Prognose zu wagen, wenn nur ein Bruchteil der Wahlberechtigten befragt wird. Sind die falschen Prognosen allein auf Unzulänglichkeiten der statistischen Methoden zurückzuführen? Grundsätzlich ist zu bedenken: Jede Umfrage ist eine Momentaufnahme. Diese spiegelt die Realität nur dann präzise wider, wenn 1. die Anzahl der Befragten hoch genug ist, wenn 2. deren Auswahl repräsentativ ist für die Population aller wahlberechtigten Bürger und wenn 3. die Befragten ehrlich antworten. Doch auch wenn diese Bedingungen erfüllt sind, können sich kurz vor der Wahl unerwartete Dinge ereignen, die das Wahlergebnis in eine andere Richtung lenken.

Ferner ist nicht auszuschließen, dass manche Wahlberechtigte aufgrund einer veröffentlichten Prognose ihr Wahlverhalten ändern – sei es, dass sie eine andere Partei wählen als ursprünglich beabsichtigt oder dass sie glauben, ihre Stimmabgabe sei nicht nötig. Hinzu kommen Faktoren, deren Einfluss unberechenbar ist, wie beispielsweise das Wetter am Wahltag. Fazit: Wenn eine Prognose abweicht vom endgültigen Wahlergebnis, ist das nicht in jedem Fall der Statistik anzulasten.

Beispiel 2: Mittelwerte

„Ein Mensch, der von Statistik hört, denkt dabei nur an Mittelwert“. So beginnt ein Gedicht von Paul Heinz List (1924–1998), ehemals Professor für Pharmazeutische Technologie an der Philipps-Universität Marburg. Das Gedicht handelt von einem Jäger, der zweimal auf eine Ente schießt, einmal eine Handbreit vor und einmal eine Handbreit nach. „Statistisch ist die Ente tot“, lautet des Jägers Fazit. Tatsächlich orientieren sich Menschen gerne an einem Durchschnittswert. Verbraucher verlassen sich beispielsweise beim Kauf eines technischen Geräts auf die mittlere Lebensdauer und sind enttäuscht, wenn ihr Gerät vorher kaputtgeht. Alte Menschen weisen zuweilen sarkastisch darauf hin, dass sie „statistisch“ tot sein müssten, weil sie die durchschnittliche Lebenserwartung überschritten haben. Diese Denkweisen sind jedoch nicht korrekt. Ein Mittelwert ist ein Maß der zentralen Tendenz und besagt nichts über einen konkreten Einzelfall. Der

Vergleich eines Einzelwerts mit dem Mittelwert gestattet jedoch – im Sinne von Quetelets Konzept des Durchschnittsmenschen – Zuordnungen wie über- oder unterdurchschnittlich und kann insofern auf Abnormalitäten oder pathologische Besonderheiten hinwiesen.

Beispiel 3: Relative Risiken

Auch relative Risiken werden häufig falsch oder unangemessen interpretiert [12]. Als Beispiel sei eine Mitteilung des britischen Komitees für Arzneimittelsicherheit aus dem Jahre 1995 erwähnt, wonach eine neue Sorte einer Anti-Baby-Pille das Thromboserisiko um 100 Prozent erhöhte. Diese Meldung versetzte viele Frauen, die diese Pille nahmen, in Panik. Sie glaubten irrtümlicherweise, dass sie nun mit 100%iger Sicherheit eine Thrombose erleiden würden und setzten das Präparat voreilig ab. In der Folge kam es zu zahlreichen ungewollten Schwangerschaften und 13.000 zusätzlichen Abtreibungen. Die absoluten Erkrankungsrisiken waren jedoch mit 2/7000 und 1/7000 bei dem neuen bzw. dem alten Präparat sehr gering. Mathematisch sind die Angaben „Steigung des Risikos um 100%“ oder „Verdoppelung des Risikos“ korrekt. Das Problem besteht jedoch darin, dass derlei Angaben zwar beeindruckend klingen, doch ohne die Angabe der absoluten Risiken wenig informativ sind.

Beispiel 4: Sinnlose Vergleiche

Eine weitere Unsitte bei der Interpretation von Risiken besteht in sinnlosen Vergleichen. Hierzu ein Beispiel: Unter Kaiser Wilhelm II. starben 5% der Untertanen an Krebs. Heute ist dies die Todesursache bei jedem vierten Deutschen. Liegt die geringe krebsbedingte Mortalität zu Beginn des 20. Jahrhunderts am Monarchen? Oder stieg die Mortalität aufgrund dramatischer Umweltverschmutzung auf ein derart hohes Niveau? Beide Fragen sind mit „nein“ zu beantworten. Die höhere krebsbedingte Mortalität ist den erfreulichen Tatsachen zu verdanken, dass typische Infektionskrankheiten wie beispielsweise die Tuberkulose kaum noch auftreten, und dass die durchschnittliche Lebenserwartung enorm gestiegen ist.

Beispiel 5: Bedingte Wahrscheinlichkeiten

Bedingte Wahrscheinlichkeiten werden häufig falsch verstanden. Das Paradebeispiel hierzu ist ein diagnostischer Test, der angewandt wird, um Informationen

bezüglich des Krankheitsstatus einer Person zu gewinnen. Man erwartet einen positiven Befund bei einer erkrankten und einen negativen Befund bei einer nicht-erkrankten Person. Diese beiden Wahrscheinlichkeiten werden als Sensitivität bzw. Spezifität bezeichnet. Bei einem HIV-Test beträgt die Sensitivität 99,9%; die Spezifität liegt bei 99,99%. Demnach sind falsche Befunde zwar nicht ausgeschlossen, aber sehr unwahrscheinlich. Nur bei einer von 1.000 infizierten Personen wird die Infektion nicht erkannt; nur eine von 10.000 nicht infizierten Personen erhält fälschlicherweise einen positiven Befund. Wie ist nun ein positiver Befund zu werten, wenn der Test an einer Person durchgeführt wird, die keiner Risikogruppe angehört (bei dieser Population ist nur eine von 10.000 Personen infiziert)? Die Antwort stößt allenthalben auf Unglauben: Nur eines von zwei positiven Testbefunden ist mit einer Infektion assoziiert; mit anderen Worten: Der positive Vorhersagewert beträgt nur 50% [12]. Der Grund ist durch die geringe Prävalenz und die hohe Spezifität gegeben, die nur ein richtig positives und ein falsch positives bei 10.000 Testpersonen erwarten lassen. Anders sieht die Situation bei einer Hochrisikogruppe mit einer Prävalenz von 10% aus: In diesem Fall liegt der positive Vorhersagewert bei 99,9%! Auch wenn die Intuition im Vertrauen auf die hohen Werte von Sensitivität und Spezifität etwas anderes suggeriert: Die Vorhersagewerte sind bedingte Wahrscheinlichkeiten, die sich mit der Formel von Bayes explizit berechnen lassen, wobei für die a-priori-Wahrscheinlichkeit die Prävalenz einzusetzen ist. Wer dies nicht glaubt (und die Formel nicht kennt), sollte eine Vierfeldertafel erstellen und die zu erwartenden Häufigkeiten berechnen. – Diese Problematik betrifft keineswegs nur HIV-Tests. Bei allen Szenarien mit niedriger Prävalenz (wie sie etwa bei Screening-Untersuchungen gegeben ist) ist der positive Vorhersagewert meist gering. Dies ist keineswegs nur ein statistisches Problem. Falsch positive Befunde verunsichern die getesteten Personen; sie ziehen eine Reihe weiterer Untersuchungen nach sich, die unnötig und für die betreffende Person sehr belastend sind.

Beispiel 6: Kausalitäten

Bereits Karl Pearson hat darauf hingewiesen, dass eine statistisch nachgewiesene Korrelation keineswegs als Beleg für eine Kausalität interpretiert werden sollte. Nun sind Menschen und insbesondere Wissenschaftler gemeinhin daran interessiert, Ursachen zu finden. So streben zum Beispiel Mediziner danach, ätiologische Faktoren zu finden, die mit dem Entstehen einer Krankheit kausal assoziiert sind.

Nur mit diesem Wissen können sie Maßnahmen ergreifen, um deren Auftreten zu verhindern. Ein statistisch abgesicherter Zusammenhang könnte jedoch durch einen Confounder bedingt sein. Dabei handelt es sich um einen zumeist unbekannten Faktor, der sowohl mit der Krankheit als auch mit dem untersuchten Faktor assoziiert ist und deshalb einen Zusammenhang vortäuscht, der zwar statistisch nachweisbar, aber nicht kausal bedingt ist. Dazu ein Beispiel: In den 1950er und 1960er Jahren beobachtete man, dass die meisten neugeborenen Kinder mit Trisomie 21 mehrere ältere Geschwister hatten. Das führte zu der irrigen Annahme, dass zahlreiche Schwangerschaften zu Veränderungen im Körper einer Frau führen, die das Risiko erhöhen und kausal mit der Erkrankung assoziiert sind. Später erkannte man, dass das Alter der Mutter die eigentliche Ursache war. – Wie lässt sich ein Confounder vehindern oder erkennen? Zum einen durch das Studiendesign: Durch eine Randomisierung (also eine zufällige Zuteilung der Studienteilnehmer auf eine Gruppe) wird Confounding vermieden. Wenn beispielsweise im Rahmen einer randomisierten Therapiestudie nachgewiesen wird, dass ein Medikament bezüglich seiner Wirkung einem Vergleichsmedikament überlegen ist, darf angenommen werden, dass diese Assoziation kausal bedingt ist. Die Randomisierung ist jedoch nur eingeschränkt anwendbar. So lassen sich beispielsweise Risikostudien nur als Beobachtungsstudien durchführen. In diesen Fällen eignen sich die Bradford-Hill-Kriterien als argumentative Hilfen bei der Beurteilung, ob ein statistisch abgesicherter Zusammenhang kausal ist [11].

Beispiel 7: Das Ziegenproblem

Vor einigen Jahren sorgte das Ziegenproblem für Furore (Abbildung 5). Dabei ging es um folgende Fragestellung: Ein Kandidat darf in einem Quiz eine von drei Türen wählen. Hinter einer Tür verbirgt sich ein Auto, hinter den anderen beiden Türen wartet jeweils eine Ziege (also eine Niete). Nachdem der Kandidat eine Tür gewählt hat, öffnet der Quizmaster von den verbleibenden Türen eine Tür mit einer Ziege und fragt den Kandidaten, ob er seine Entscheidung revidieren möchte. Sollte er dies tun? Klar ist: Die Wahrscheinlichkeit, dass er die Tür mit dem Auto gewählt hat, beträgt nach Laplace 1/3. Die meisten Menschen glauben nun intuitiv, dass die Wahrscheinlichkeit nach dem Öffnen der einen Tür von 1/3 auf 1/2 gestiegen ist (da ja nur noch zwei Türen für den Gewinn in Frage kommen). Daher sei es unerheblich, ob der Kandidat seine Entscheidung revidiert. Aber sie irren sich! Tatsächlich würde eine Entscheidung für die andere

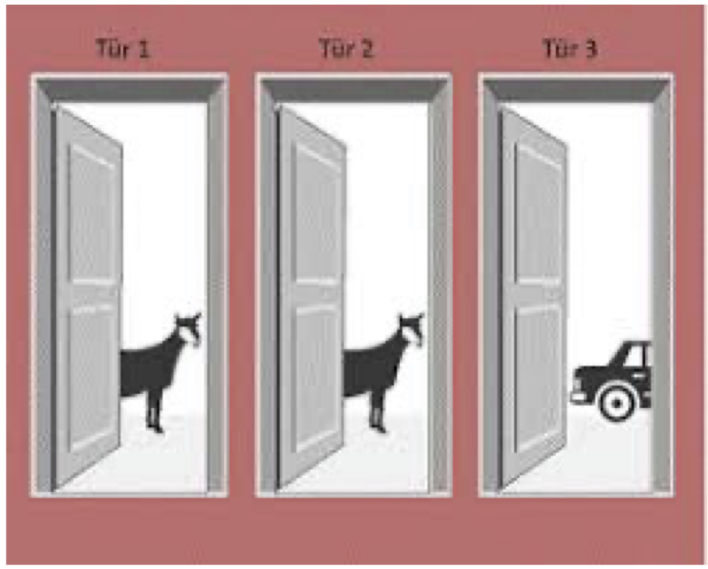


Abbildung 5: Illustration des Ziegenproblems.

Tür die Gewinnwahrscheinlichkeit auf 2/3 erhöhen! Wie ist das zu erklären? Die Wahrscheinlichkeit, dass das Auto hinter einer der nicht gewählten Türen steht, beträgt von Anfang an 2/3. Dieser Wert ändert sich keineswegs, nachdem der Quizmaster den Blick hinter eine der beiden verschlossenen Türen gestattet hat. Da die Wahrscheinlichkeit, dass das Auto hinter der geöffneten Tür steht, 0 beträgt, verbleibt eine Wahrscheinlichkeit von 2/3 für die nicht geöffnete Tür. – Als die Schriftstellerin Marilyn vos Savants (*1946) im Jahre 1990 diese knifflige Aufgabe (auch als Monty-Hall-Dilemma bezeichnet) und deren Lösung publizierte, löste sie heftige Debatten aus [9]. Sie musste polemische Anfeindungen über sich ergehen lassen, einige sogar von Mathematikprofessoren, die ihr fehlende Grundkenntnisse in Wahrscheinlichkeitstheorie oder mangelnde mathematische Fertigkeiten vorwarfen.

Jedes dieser Beispiele hat zu heftigen Kontroversen geführt und auf seine Weise dazu beigetragen, die Statistik als Wissenschaft in Verruf zu bringen. Dabei ließen sich die oben dargelegten Missverständnisse auf einfache Weise klären, etwa durch korrekte Interpretationen eines Mittelwerts, eines Korrelationskoeffizienten oder eines relativen Risikos oder durch den adäquaten Umgang mit Wahrschein-

lichkeiten. Freilich bereiten einige Beispiele Verständnisschwierigkeiten oder erfordern eine fundierte fachliche Expertise. Dies erklärt jedoch nicht, weshalb viele Menschen der Statistik ablehnend bis feindselig begegnen. Es drängt sich der Eindruck auf, dass es manchen Leuten Freude bereitet, vermeintliche Schwächen der Statistik aufzuzeigen und sich darüber zu mokieren (ohne ernsthaft darüber nachzudenken).

4 Diskussion und Schlussfolgerungen

Meinungsverschiedenheiten innerhalb einer Wissenschaft oder zwischen Vertretern unterschiedlicher Disziplinen sind Bestandteil des wissenschaftlichen Diskurses. Wie in den obigen Kapiteln dargelegt wurde, waren und sind Auseinandersetzungen zwischen und mit Statistikern teilweise recht heftig bis feindselig. Zwar gab es keine kriegerischen Auseinandersetzungen, aber durchaus nationale und internationale Konflikte, die die Statistik als Wissenschaft und die Anwendung ihrer Methoden maßgeblich beeinflusst haben.

Bergen die unterschiedlichen Konzepte zur Definition einer Wahrscheinlichkeit tatsächlich Konfliktpotenzial in sich? Oder bieten sie vielmehr eine Möglichkeit, diesen Begriff in unterschiedlicher Weise zu handhaben und dabei anzuerkennen, dass die Definitionen der Herren Laplace, von Mises und Bayes sowie das Axiomensystem von Kolmogoroff zu ihrer Zeit geniale Leistungen waren?

Warum kritisierten deutsche Staatenkundler die methodisch orientierte Entwicklung der Statistik in England und verhinderten sogar die Einführung quantitativer Methoden in Deutschland? Sachliche Kommunikation und akademischer Erfahrungsaustausch wären zielführender gewesen.

Mit einem zeitlichen Abstand von nahezu einhundert Jahren stellt sich die Frage: War der offen ausgetragene Konflikt zwischen Fisher und seinen Kontrahenten Pearson und Neyman unausweichlich? Ging es wirklich um philosophische Feinheiten, oder wurden eher persönliche Animositäten ausgetragen? Aus heutiger Sicht wirken die Streitigkeiten unverständlich. Es ist seit Langem gängige Praxis, vor der Durchführung eines statistischen Tests ein Signifikanzniveau festzulegen (nach Neyman), einen p-Wert zu berechnen (nach Fisher), das Resultat zu diskutieren und Konsequenzen aufzuzeigen. So sind beide Ansätze harmonisch vereint. In einer anderen Frage waren sich Fisher und Karl Pearson dagegen einig. Sie interessierten sich für die Forschungsarbeiten Galtons auf dem Gebiet der Euge-

nik und gelangten beide zu der Überzeugung, dass es sinnvoll sei, die gesunden Mitglieder einer Gesellschaft den schwachen vorzuziehen. Langfristig hatte dies fatale Auswirkungen: Derlei Argumentationen dienten in späteren Jahren den Nationalsozialisten als Begründung für ihre Euthanasieprogramme. Es ist tragisch, dass sich hochintelligente Wissenschaftler wie Galton, Fisher oder Pearson von derart menschenverachtenden Ideologien vereinnahmen und sich dazu verführen ließen, ihre eigenen Ideen und statistischen Konzepte dafür zu missbrauchen.

Des Weiteren ist sachlich festzuhalten: Der klassische Ansatz und der Ansatz von Bayes zum Testen von Hypothesen stehen in keinem Widerspruch. Subjektive Wahrscheinlichkeiten, die im Bayes-Ansatz verwendet werden, sind nicht gleichzusetzen mit willkürlichen Festlegungen (wie von manchen Anhängern der klassischen Statistik gemutmaßt wird). Keiner dieser Ansätze kann grundsätzlich abgelehnt oder befürwortet werden.

Von manchen Vertretern empirischer Wissenschaften wird Statistik als notwendiges Übel angesehen; Konflikte mit dem Statistiker sind quasi vorprogrammiert. Für alle Beteiligten sollte es möglich sein, Verständnis für andere Sichtweisen aufzubringen, Grenzen des eigenen Fachgebiets anzuerkennen und eigene Denkweisen zu korrigieren, um eine für alle Beteiligten akzeptable Lösung zu finden. Bei jeder wissenschaftlichen Studie sollten die Autoren die verwendeten Verfahren offen darlegen und diskutieren. Die Reviewer und die Leser einer Publikation sollten – ehe sie zum Angriff übergehen – bedenken, dass eine statistische Analyse die Realität niemals exakt beschreiben kann und dass die Durchführung einer Studie von allen Beteiligten Kompromisse verlangt.

Die Attacken, denen sich Statistiker zuweilen ausgesetzt sehen, sind häufig auf Unkenntnis zurückzuführen oder darauf, dass viele Menschen ihrer Intuition blind vertrauen (und nicht bereit sind, ihre Position in Frage zu stellen). Es kommt hinzu, dass es in manchen gesellschaftlichen Kreisen keineswegs als verwerflich gilt, mit seinen geringen Kenntnissen in Mathematik oder Statistik zu prahlen; im Gegenteil: Man darf mit Verständnis und Zustimmung rechnen! Eigentlich ist es ein Paradox: Kaum jemand kann sich der Statistik entziehen. Im Alltag wird man quasi rund um die Uhr über diverse Kommunikationsmedien mit Zahlenwerken, Tabellen, Diagrammen und Prognosen versorgt. Wissenschaftler aus Bio-, Sozial- und Naturwissenschaften sind auf statistische Analysen angewiesen, um Fakten aufzufinden und zu neuen Erkenntnissen zu gelangen. Dennoch wird die Statistik

wie kaum eine andere Wissenschaft skeptisch beurteilt und teilweise sogar infrage gestellt oder bekämpft.

Schließlich sei noch darauf hingewiesen, dass die Geschehnisse während des Zweiten Weltkriegs die Entwicklung der Medizinischen Statistik und damit auch der klinischen Forschung in Deutschland lange Zeit verzögerten. Eingedenk der menschenverachtenden Experimente, die während der Nazizeit durchgeführt worden waren, lehnte man hierzulande Therapiestudien bis in die 1970er Jahre ab. In England wurden dagegen bereits im Jahre 1948 randomisierte Therapiestudien durchgeführt [13]. Diesbezüglich hatte der Krieg also unmittelbare Auswirkungen auf die Forschung in Deutschland.

Diese Ausführungen belegen: Verbale Attacken oder persönliche Angriffe sind in keinem Fall zweckdienlich. Weit sinnvoller sind offene und faire Diskurse gepaart mit gegenseitigem Respekt, was freilich die Fähigkeit zur kritischen Selbstreflexion voraussetzt. Der Vergleich mit aktuellen Kriegsgeschehen mag verwegen erscheinen. Doch eingedenk der bitteren Erkenntnisse nach zahllosen Kriegen, die in der Vergangenheit geführt wurden, und eingedenk der aktuellen Situation in Europa muss man zu dem Fazit gelangen, dass achtungsvolle Begegnungen, gegenseitige Toleranz und Solidarität im menschlichen Miteinander weitaus erfolgversprechender und angenehmer sind als Kriege und offen ausgetragene Konflikte – sei es in der Politik, in der Wissenschaft oder im gesellschaftlichen Miteinander.

Referenzen

[1] Döring N, Bortz J: Forschungsmethoden und Evaluation in den Human- und Sozialwissenschaften. 5. Auflage, Springer-Verlag Heidelberg (2016)

[2] Gigerenzer G: Die Evolution des statistischen Denkens. Unterrichtswissenschaft 32(1), 4–22; 2004

[3] Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L: Das Reich des Zufalls. Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Unschärfen. Spektrum Akademischer Verlag, Seite 56 (1998)

[4] Ioannidis JPA: Why most published research findings are false. In: PLoS Medicine 2, e124 (2005)

[5] Klein I: Grundlagenstreit in der Statistik. Diskussionspapier 60/2004, Friedrich Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie. https://www.econstor.eu/handle/10419/29608 (2004)

[6] Lehmann E: The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association 88(424): 1242–1249 (1993)

[7] Rößler I, Ungerer A: Statistik für Wirtschaftswissenschaftler. Eine anwendungsorientierte Einführung. 6. Auflage, Springer-Gabler-Verlag Berlin (2019)

[8] Russell S, Norvig P: Künstliche Intelligenz. Ein moderner Ansatz. 3. Auflage, Pearson-Verlag München (2012)

[9] vos Savant M: Brainpower. Die Kraft des logischen Denkens. Rowohlt Taschenbuch Verlag, Seiten 28–42 (2000)

[10] Weiß C: Basiswissen medizinische Statistik. 7. Auflage, Springer-Verlag Heidelberg (2019)

[11] Weiß C: Nachweis einer Kausalität. In: Basiswissen medizinische Statistik. 7. Auflage, Springer-Verlag Heidelberg, Seite 233 f. (2019)

[12] Weiß C: Statistik kritisch beleuchtet. In: Reproduktionsmedizin. Zahlen und Fakten für die Beratung. Herausgegeben von Kupka MS. Urban & Fischer in Elsevier, S. 113–120 (2021)

[13] 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. 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 Autorin oder 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/