



STATISTIK

Eine signifikante Geschichte

von Christian Honey

Im Jahr 1908 legte ein Mitarbeiter der Guinness-Brauerei den Grundstein für einen Irrtum, der sich bis heute bemerkbar macht.

Am 11. Februar 2016 verkündeten die Astrophysiker des LIGO-Experiments ihre große Entdeckung: Zum ersten Mal hatten sie ein Signal aus den Tiefen des Alls aufgefangen, das perfekt mit der rund 100 Jahre alten Vorhersage von Gravitationswellen aus Einsteins Relativitätstheorie übereinstimmte. 5,1 Sigma stark war das Signal, hieß es in der Meldung, die alle Medienberichte aufgriffen. Damit sei das Ergebnis »signifikant«; die Existenz der Gravitationswellen unzweifelhaft bewiesen.

Solche Signifikanzwerte, wie sie sich mit Mitteln der Statistik berechnen lassen, werden heute in den Medien und der Forschung wie Grenzen der Wahrheit behandelt. Wer sich aber die Geschichte der Signifikanz anschaut, erkennt: Die größten Entdeckungen der Naturwissenschaft kamen zu Stande, auch ohne dass jemand ihre Signifikanz berechnete. Wie sich zeigt, ist Signifikanz alles andere als die Schwelle der Wahrheit. Und ihre vermeintlich fixen Grenzen beruhen auf reiner Intuition.

Wir schreiben den 20. Mai 1747. Das königliche britische Marineschiff *Salisbury* patrouilliert an der Biskaya, um eine Seeblockade gegen Frankreich durchzusetzen. In ihrem Bauch trägt sie 50 Kanonen, 300 Matrosen und hunderte Eichenfässer voller Proviant. In den »regulations« der königlichen Marine ist für jeden Seemann eine gigantische Ration vorgesehen: täglich 500 Gramm Schiffszwieback und 3,5 Liter Bier und jede Woche insgesamt 2 Kilogramm Rindfleisch, 1 Kilogramm Schweinefleisch, 1 Kilogramm Erbsen, 1,5 Liter Hafermehl für Brei, 170 Gramm Butter und 340 Gramm Käse.

Wissenschaft ohne Signifikanz

Doch nach mehr als acht Wochen ohne Landgang helfen auch diese 5000 Kalorien pro Tag und Mann nicht. »Zwölf kranke Männer hatte ich ins Vorschiff verlegt«, berichtet der Schiffsarzt James Lind. »Ihre Fälle hätten ähnlicher nicht sein können. Alle hatten verdorbenes Zahnfleisch, waren mit Flecken übersät, matt und schwach in den Knien.«

Für Lind war dies ein vorzüglicher Anblick, denn er wollte die Ursache des Skorbut finden. Je ähnlicher der Zustand der Matrosen war, desto

besser sollte sich an ihnen der Einfluss verschiedener Behandlungen zeigen. Allen zwölf Versuchsmatrosen verschrieb der Forscher das exakt gleiche tägliche Menü, mit nur einer kleinen Variation: Je zwei der Männer bekamen zusätzlich entweder Apfelwein, Apfelessig, etwas Schwefelsäure, Seewasser, Muskatnussmehl oder zwei Orangen und eine Zitrone.

Die Folgen waren erstaunlich deutlich: »Die schnellsten und offensichtlichsten guten Effekte hatten die Früchte«, schrieb er in seinem Bericht aus dem Jahr 1753. »Schon nach sechs Tagen war einer der beiden Männer wieder diensttauglich. Zurück in Plymouth, war auch der zweite genesen.«

Linds Experimente trugen maßgeblich dazu bei, dass ab 1795 auf allen Schiffen der britischen Marine der Grog mit Zitronensaft angereichert wurde und seither kaum ein Seemann mehr dem Skorbut erlag. Heute gilt Lind als der erste Forscher, der ein derart streng kontrolliertes klinisches Experiment durchführte. Sein wesentliches Verdienst war es zu zeigen, dass es sich lohnt, störende Faktoren zu minimieren; bei den Matrosen also individuelle Unterschiede in Alter, Größe, Schwere der Symptome oder auch in anderen Erkrankungen.

Nach Tabellen, statistischen Größen wie Mittel- oder Signifikanzwerten sucht man in Linds Bericht aber vergebens. Und noch 170 Jahre lang galten Ergebnisse als überzeugend, wenn das Experiment gut aufgebaut und die Daten klare Effekte zeigten. 170 Jahre, in denen der Sauerstoff, die elektrochemischen Grundlagen der Batterie und der Elektromagnetismus, die Radioaktivität und die Röntgenstrahlung entdeckt wurden. 170 Jahre, in denen sich herausstellte, dass alle Pflanzen aus Zellen aufgebaut sind, diese Zellen sich teilen, dass Arten durch natürliche Auslese ent-

JAMES LIND (1716–1794)

Der Schiffsarzt schottischer Abstammung fand durch gezieltes Experimentieren und genaues Beobachten eine Möglichkeit, den gefürchteten Skorbut zu verhindern. Auch in anderer Hinsicht erleichterte er den Seeleuten das Leben an Bord. So führte er etwa Hygienemaßnahmen wie regelmäßiges Lüften unter Deck ein.

»Keine große Entdeckung
in der Naturwissenschaft
hat allein auf Grund
kleiner Signifikanzwerte
stattgefunden«

[Stephen Ziliak]

stehen, es noch andere Galaxien als die Milchstraße gibt, das Universum expandiert und dass es deshalb wahrscheinlich einen Anfang hat.

»Sogar der erste empirische Nachweis der Krümmung des Lichts, eine Vorhersage aus Einsteins Relativitätstheorie durch Arthur Eddington im Jahr 1919, wurde ohne Signifikanzen geführt«, sagt Stephen Ziliak, Wirtschaftsprofessor an der University of Chicago. »Dafür fuhren Schiffe mit Teleskopen bestückt vor die Küsten Afrikas und Brasiliens, um die Position eines Sterns während einer Sonnenfinsternis zu vermessen. Die Daten waren so überzeugend, dass alle Experten sofort zustimmten: Newton liegt falsch, Einstein hat Recht.« Ziliak ist einer der schärfsten Kritiker unreflektierter Signifikanztesterei: »Keine große Entdeckung in der Naturwissenschaft hat allein auf Grund kleiner Signifikanzwerte stattgefunden«, sagt der Autor des Buchs »The Cult of Statistical Significance«.

Wie kam es zur Signifikanz?

Der Bedarf, die Signifikanz – oder anders ausgedrückt: die Zuverlässigkeit – eines Resultats zu quantifizieren, kam erst zur Wende des 20. Jahrhunderts auf. Seit dem Jahr 1899 war der englische Chemiker William Sealy Gosset bei der Guinness-Brauerei in Dublin angestellt. Seine Hauptaufgabe war es, die besten Rohstoffe (Hopfen, Gerste, Hefe) für den schwarzen Sud zu finden. Bei der Gerste aber fiel der Vergleich der Malzausbeute schwer: Viele Bauern aus verschiedenen Regionen kamen als Hauptlieferanten in Frage. Zudem konnte Gosset seine Experimente auf Grund hoher Kosten nur an kleinen Stichproben durchführen. Das aber könnte zur Folge haben, dass bei einem Vergleich zweier Lieferanten die mittlere Malzausbeute bei der ersten Stichprobe sehr unterschiedlich ausfiel, nur um bei der nächsten Stichprobe praktisch identisch zu sein. Welcher Lieferant war nun bessere? Wie sicher konnte man sich sein, dass ein beobachteter Unterschied in der Malzausbeute echt war und kein Kind von Zufallsschwankungen?

Gosset entwickelte ein mathematisches Verfahren, das den »wahrscheinlichen Fehler der Mittelwerte« bei kleinen Stichproben bestimmt. Je kleiner die Stichprobe, desto größer wird in seinem t-Test der wahrscheinliche Fehler und desto geringer die Signifikanz des Messwerts der Malzausbeute. Im Jahr 1908 veröffentlichte Gosset seinen Test – damals noch ein Geschäftsgeheimnis von Guinness – unter dem Pseudonym