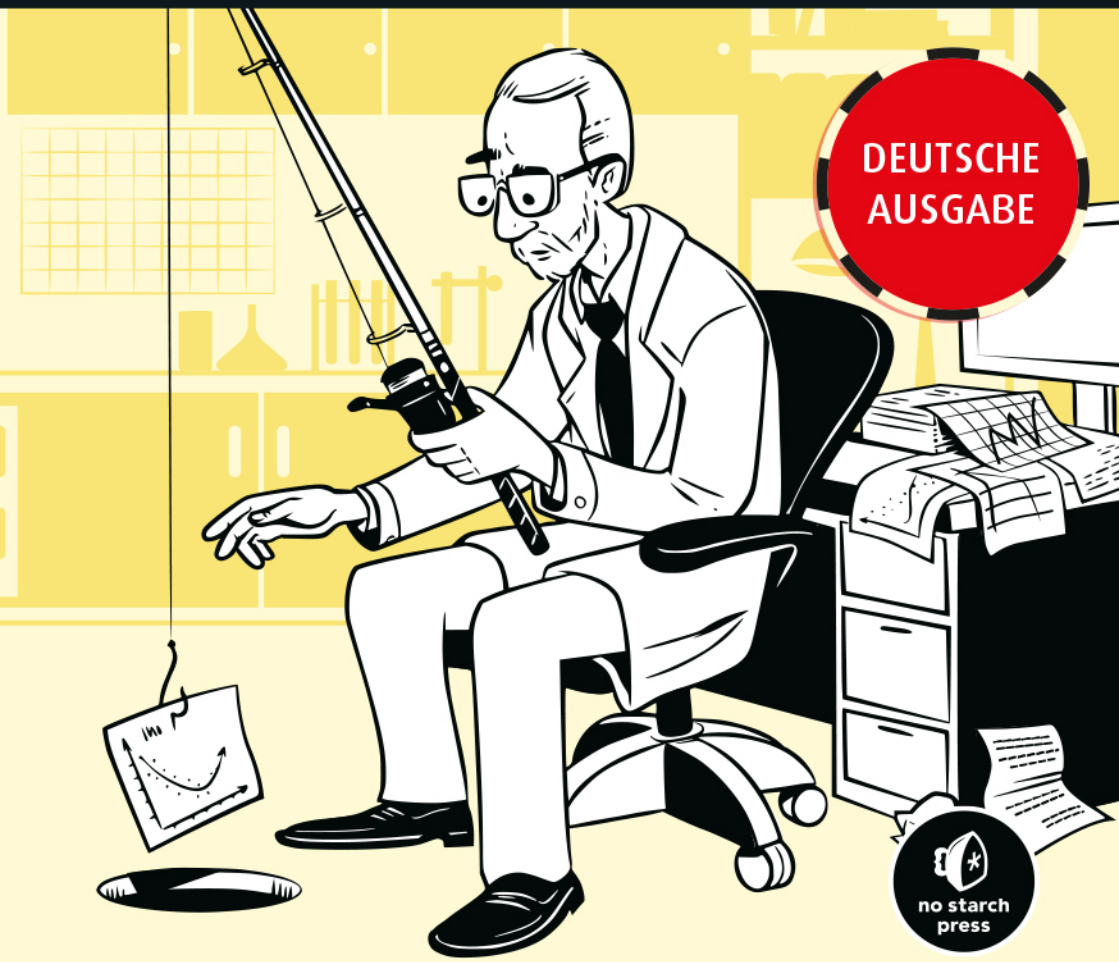
ALEX REINHART

mitp

STATISTICS DONE WRONG

STATISTIK RICHTIG ANWENDEN UND GÄNGIGE FEHLER VERMEIDEN



DEUTSCHE AUSGABE

no starch press

Inhaltsverzeichnis

Rezensionen 1

Über den Autor 10

Vorwort 11

Einleitung 16

1 **Eine Einführung in die statistische Signifikanz** 23
1.1 p-Werte 23
1.1.1 »Übernatürliche« Statistik 27
1.1.2 Neyman-Pearson-Tests 28
1.2 Intervallen vertrauen 31

2 **Teststärke und Stichprobenumfang** 35
2.1 Die Gütefunktion 35
2.2 Risiken unzureichender Teststärke 39
2.2.1 Weshalb unzureichende Teststärke? 41
2.2.2 Abbiegen bei Rot 43
2.3 Beeinflussung des Konfidenzintervalls 46
2.4 Überbewertung der Tatsachen 47
2.4.1 Kleine Extrema 51
2.5 Tipps 54

3 **Pseudoreplikation: Datenauswahl** 57
3.1 Pseudoreplikation in Aktion 58
3.2 Pseudoreplikation berücksichtigen 60
3.3 Biologie wie am Fließband 61
3.4 Synchronisierte Pseudoreplikation 63
3.5 Tipps 66

4 **p-Wert und Prävalenzfehler** 69
4.1 Prävalenzfehler 71
4.1.1 Ein Kurztest 72
4.1.2 Prävalenzfehler bei medizinischen Untersuchungen 73

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

4.1.3 Wie man mit Raucherstatistiken lügt 74
4.1.4 Maßnahmen zur Vermeidung von Prävalenzfehlern 77
4.2 Wenn es nicht sofort klappt, immer wieder probieren 80
4.3 Falsche Fährten bei Kernspintomografien des Gehirns 85
4.4 Erkennung fehlerhafter Zuordnungen 86
4.5 Tipps 88

5 Unzureichende Beurteilung der Signifikanz 89
5.1 Nicht signifikante Unterschiede der Signifikanz 89
5.2 Visueller Vergleich der Signifikanzen 94
5.3 Tipps 98

6 Doppelte Datennutzung 99
6.1 Zirkuläre Analysen 100
6.2 Regression zur Mitte 104
6.3 Abbruchregeln 107
6.4 Tipps 111

7 Fortbestehende Fehler 113
7.1 Überflüssige Zweiteilung 114
7.2 Statistische Verdunkelung 116
7.3 Verwirrende Störfaktoren 117
7.4 Tipps 120

8 Modellmissbrauch 121
8.1 Wassermelonen 123
8.2 Korrelation und Kausalität 128
8.3 Simpsons Paradoxon 130
8.4 Tipps 134

9 Forschungsfreiheit: Alles in Butter? 135
9.1 Ein wenig Spielraum kann gefährlich sein 137
9.2 Voreingenommenheit vermeiden 140
9.3 Tipps 144

10 Fehler begeht jeder 145
10.1 Unreproduzierbare Genetik 146
10.2 Reproduzierbarkeit erleichtern 149
10.3 Experimentieren, aufräumen, wiederholen 152
10.4 Tipps 154

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

11 Daten verbergen 155
11.1 Eingesperrte Daten 156
11.1.1 Hindernisse für das Teilen der Daten 157
11.1.2 Datenverfall 159
11.2 Einzelheiten weglassen 162
11.2.1 Bekannte Unbekannte 162
11.2.2 Voreingenommene Berichterstattung 164
11.3 Wissenschaft im Aktenschrank 167
11.3.1 Unveröffentlichte klinische Studien 167
11.3.2 Voreingenommene Berichterstattung erkennen 169
11.3.3 Erzwungene Offenlegung 171
11.4 Tipps 173

12 Was ist zu tun? 175
12.1 Statistikausbildung 178
12.2 Veröffentlichung wissenschaftlicher Arbeiten 182
12.3 Ihre Aufgaben 185

Anmerkungen 189

Stichwortverzeichnis 207

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Über den Autor

Alex Reinhart ist Doktorand an der Carnegie Mellon University in Pittsburgh, Pennsylvania, und unterrichtet dort Statistik. Seinen Bachelor of Science in Physik erlangte er an der University of Texas in Austin. Er befasst sich mit Forschungsarbeiten zum Aufspüren radioaktiver Objekte unter Zugrundelegung physikalischer und statistischer Gesetzmäßigkeiten.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Vorwort

Als ich vor einigen Jahren mein Hauptstudium im Fachbereich Physik an der University of Texas in Austin absolvierte, wurde ich im Rahmen eines Seminarkurses aufgefordert, mir ein Thema für einen individuellen 25-minütigen Vortrag auszusuchen, den alle Seminarteilnehmer halten sollten. »Ich würde gern etwas über Verschwörungstheorien machen«, sagte ich zu dem Dozenten Dr. Brent Iverson, der sich damit jedoch nicht zufriedengab. Das Thema sei zu weitreichend, meinte er, und ein engagierter Vortrag müsse fokussiert und ausführlich sein. Also sah ich mir die Liste der Themenvorschläge an. »Wie wäre es denn mit wissenschaftlichem Fehlverhalten?«, fragte er, und ich willigte ein.

Rückblickend bin ich gar nicht mal sicher, inwieweit Falschangaben und fehlerhafte Schlussfolgerungen in der Wissenschaft ein begrenzteres Thema als Verschwörungstheorien sind, das spielte aber auch keine Rolle. Nach einigen Stunden doch ziemlich intensiver Recherche wurde mir klar, dass das wissenschaftliche Fehlverhalten an sich kein allzu interessantes Thema ist – jedenfalls nicht in Anbetracht all der Fehler, die Wissenschaftler *unabsichtlich* begehen.

Ungeachtet meiner beklagenswert mangelhaften Qualifikation in punkto Statistik suchte ich mir also einige Dutzend Forschungsarbeiten über die zahllosen statistischen Fehler zusammen, die Wissenschaftlern regelmäßig unterlaufen. Ich las sie, resümierte sie und arbeitete schließlich einen Vortrag aus, der Dr. Iverson zufriedenstellte. Und als angehender Wissenschaftler (und nun auch noch selbsternannter Statistikexperte) kam ich letztendlich zu der Überzeugung, dass ich vielleicht auch mal einen Statistikkurs besuchen sollte.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Gesagt, getan: Zwei Jahre und zwei Statistikkurse später immatrikulierte ich an der Carnegie Mellon University als Doktorand im Fachbereich Statistik – und bis heute bereitet es mir ein diebisches Vergnügen, Gelegenheiten zu finden, um mit statistischen Werten »herumzupfuschen«.

Dieses Buch ist ein Leitfaden, der den eher hanebüchenen statistischen Fehlern und Irrtümern gewidmet ist, die regelmäßig im Namen der Wissenschaft begangen werden. Da viele Wissenschaftler keine formale Statistikausbildung genießen – und weil ich meine Leserschaft nicht auf ohnehin an diesem Thema Interessierte beschränken möchte –, werden hier keine besonderen Vorkenntnisse in Sachen Statistik vorausgesetzt. Manch ein Leser mag daher möglicherweise versucht sein, das erste Kapitel zu überspringen, ich empfehle jedoch, es zumindest zu überfliegen, um sich mit meinem Erklärungsstil vertraut zu machen.

Mein Ziel besteht nicht nur darin, die oft begangenen Fehler beim Namen zu nennen und entsprechende Beispiele zu liefern, über die Sie schmunzeln können. Soweit es ohne ausführliche mathematische Exkurse möglich ist, erläutere ich außerdem, *warum* besagte statistische Fehler eigentlich Fehler sind und berufe mich in diesem Kontext auf Studien, die belegen, wie weit die meisten dieser Fehler tatsächlich verbreitet sind. Die Lektüre dieses Buches wird dadurch zwar ein wenig erschwert, ich bin jedoch der Ansicht, dass der hieraus resultierende Erkenntnisgewinn die Mühe durchaus wert ist – ein gründliches Verständnis von elementarer Statistik ist schließlich für jeden Wissenschaftler unverzichtbar.

Für Leser, die alltäglich statistische Analysen durchführen, finden sich am Ende der meisten Kapitel Tipps, die Aufschluss darüber geben, welche statistischen Verfahren verwendet werden können, um die häufigsten Fallstricke zu umgehen. Das vorliegende Werk ist allerdings kein Lehrbuch, daher werde ich auch nicht näher auf die Einzelheiten der genannten Verfahren eingehen. Ich hoffe jedoch, Sie auf die gängigsten Probleme aufmerksam machen und somit in die Lage versetzen zu können, das für eine gegebene Aufgabe am besten geeignete statistische Verfahren auszuwählen.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Für den Fall, dass ein oder mehrere der hier angesprochenen Themen Ihre Neugier wecken, finden Sie am Ende des Buches ein ausführliches Literaturverzeichnis. Außerdem sind alle statistischen Fehler und Irrtümer mit Quellverweisen versehen. Zugunsten eines grundsätzlichen Verständnisses habe ich sehr viel Mathematisches weggelassen. Sollten Sie jedoch an formal strengeren Abhandlungen interessiert sein, empfehle ich die Lektüre der diesem Buch zugrundeliegenden Originalarbeiten.

Bevor Sie dieses Buch lesen, hier noch ein kurzer Warnhinweis: Sobald man sich Wissen angeeignet hat, über das nur wenige andere Menschen verfügen, ist es verlockend, dies bei jeder sich bietenden Gelegenheit zur Schau zu stellen. Sollte dieses Buch auf wundersame Weise zum Bestseller werden, erwarte ich als Reaktion auf jegliche wissenschaftliche Meldungen in der Boulevardpresse das, was Paul Graham als »Ablehnung durch den Durchschnittsmenschen« bezeichnet – denn anstatt sich die Zeit zu nehmen, die interessanten Teile wissenschaftlicher Forschung zu verstehen, verwenden Möchtegern-Statistiker lieber die von irgendeiner übereifrigen Universitätspressestelle wiedergekäute ungenaue Beschreibung einer Studie, um sich auf solche Artikel einzuschießen und den statistischen Aufbau der Forschungsarbeit zu kritisieren.*

Auf den meisten Websites, die wissenschaftliche Meldungen diskutieren, geschieht dies bereits und es würde mich ungemein ärgern, wenn dieses Buch als Rechtfertigung dafür herhalten sollte. Die ersten Kommentare zu neuen Newsartikeln und Nachrichtenmeldungen sind fast immer Beschwerden wie »diese Variable wurde nicht kontrolliert« oder »die Stichprobe ist viel zu klein« – und in 9 von 10 Fällen hat der Kommentierende die wissenschaftliche Arbeit überhaupt nicht gelesen,

* Nebenbei bemerkt denke ich, dass Verschwörungstheorien aus eben diesem Grund so populär sind: Wenn man glaubt, etwas zu wissen, das niemand sonst weiß (die Regierung ist hinter uns her!), nimmt man jede Gelegenheit wahr, diese Erkenntnisse zur Schau zu stellen – und schließlich reagiert man auf alle Nachrichtenmeldungen damit, Gründe hervorzubringen, warum die Regierung sie fälscht. Das sollte Ihnen bei statistischen Fehlern jedoch bitte nicht passieren.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

denn sonst wüsste er, dass seinem Einwand im dritten Absatz begegnet wird.

Das Ganze ist absolut töricht: Einige wenige Statistikkenntnisse sind keine Entschuldigung dafür, die gesamte moderne Wissenschaft abzulehnen. Die statistischen Methoden einer Forschungsarbeit dürfen nur im Detail und im Kontext der anderen Methoden beurteilt werden: Aufbau der Studie, Messverfahren, Kostenbeschränkungen und Aufgabenstellung. Setzen Sie Ihre Statistikkenntnisse ein, um die Stärken, die Einschränkungen und die potenzielle Voreingenommenheit einer Untersuchung besser zu verstehen, nicht um jede Forschungsarbeit niederzumachen, die unsachgemäßen Gebrauch von einem *p*-Wert macht oder Ihren persönlichen Ansichten widerspricht. Bedenken Sie außerdem, dass auch eine Schlussfolgerung, die durch mangelhafte Statistiken belegt wird, dennoch richtig sein kann – statistische und logische Fehler führen nicht automatisch zu einem Fehlschluss, sie belegen ihn nur nicht.

Kurzum: Gehen Sie bitte verantwortungsvoll mit Statistiken um. Ich kann mir nur wünschen, dass Sie danach streben, die Wissenschaft zu verbessern, von der wir alle abhängig sind.

Danksagungen

Danke an James Scott, dessen Statistikkurse am Anfang meiner Laufbahn im Fachbereich Statistik standen und mir das zum Schreiben dieses Buches nötige Hintergrundwissen vermittelten. Weiterhin bedanke ich mich bei Raye Allen, der James' Hausaufgaben so viel spaßiger machte, bei Matthew Watson und Moriel Schottlender, die mir wertvolle Anregungen und unschätzbares Feedback zu meinen Textentwürfen gaben, bei meinen Eltern für ihre Denkanstöße und Rückmeldungen sowie bei Dr. Brent Iverson, dessen Seminar mich überhaupt erst dazu veranlasste, mich mit dem unsachgemäßen Gebrauch von Statistiken zu befassen. Und natürlich gilt mein Dank ebenso allen Wissenschaftlern und Statistikern, die gegen die Regeln verstoßen haben und mir somit einen Grund zum Schreiben lieferten.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Auch meine Freunde an der Carnegie Mellon University haben viele Ideen beigesteuert und jede Mengen Fragen beantwortet. Sie haben mir stets geduldig zugehört, wenn ich mal wieder versuchte, einen neuen statistischen Fehler zu erläutern. Das nötige Hintergrundwissen hierfür verdanke ich meinen Professoren, insbesondere Jing Lei, Valérie Ventura und Howard Seltman. Howard fungierte außerdem als fachlicher Redakteur und spürte in dieser Eigenschaft einige peinliche Fehler in meinem Buchentwurf auf. Sollten immer noch welche übrig sein, bin ich selbst dafür verantwortlich – allerdings werde ich frech behaupten, dass sie lediglich dem Titel des Buches Rechnung tragen.

Und meine Herausgeber bei No Starch haben dem Manuskript schließlich den nötigen Feinschliff verpasst und es damit drastisch verbessert. Greg Poulos gab sich nach der sorgfältigen Lektüre der ersten Kapitel erst zufrieden, als er jedes einzelne Konzept verstanden hatte. Leslie Shen sorgte für die Glättung meiner Polemik in den letzten Kapiteln, und das gesamte Team hat maßgeblich dazu beigetragen, dass sich dieses Projekt insgesamt verblüffend einfach gestaltete.

Darüber hinaus bin ich vielen weiteren Leuten zu Dank verpflichtet, die mir Anregungen und Kommentare per E-Mail geschickt haben, nachdem der Leitfaden online gestellt wurde. Ich bedanke mich (in keiner bestimmten Reihenfolge) bei: Axel Boldt, Eric Franzosa, Robert O'Shea, Uri Bram, Dean Rowan, Jesse Weinstein, Peter Hozák, Chris Thorp, David Lovell, Harvey Chapman, Nathaniel Graham, Shaun Gallagher, Sara Alspaugh, Jordan Marsh, Nathan Gouwens, Arjen Noordzij, Kevin Pinto, Elizabeth Page-Gould und David Merfield, ohne deren Feedback meine Erläuterungen zweifelsohne weniger vollständig wären.

Vielleicht gesellen auch Sie sich bald zu dieser Liste hinzu, denn trotz meiner größten Bemühungen wird dieser Leitfaden auch jetzt unweigerlich noch einige Fehler und Versäumnisse aufweisen. Sollten Sie also eine Unstimmigkeit finden, eine Frage haben oder ein gängiges statistisches Fehlverhalten kennen, das ich übersehen habe, können Sie mich auf Englisch unter `alex@refsmmat.com` per E-Mail erreichen.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Einleitung

Im letzten Kapitel seines bekannten Buches *How to Lie with Statistics* (»Wie man mit Statistiken lügt«) schreibt Darrell Huff, dass »alles, was nach dem medizinischen Berufsstand riecht« oder von wissenschaftlichen Laboratorien und Universitäten mitgetragen wird, unser Vertrauen genießt – zwar kein bedingungsloses Vertrauen, aber sicherlich ein größeres als das den Medien oder Politikern entgegengebrachte. (Kein Wunder, denn schließlich enthüllt Huffs Buch eine Vielzahl der irreführenden statistischen Tricksereien, die in den Medien und der Politik Anwendung finden.) Doch nur die wenigsten Leute beklagen sich über die von ausgebildeten Wissenschaftlern erstellten Statistiken – denn schließlich streben besagte Akademiker ja nach Erkenntnisgewinn und nicht nach irgendwelchem Material, das sich gegen politische Gegner verwenden lässt.

Statistische Datenanalysen sind ein Grundpfeiler der Wissenschaft. Schlagen Sie irgendeine beliebige Seite in einem der Fachmagazine auf, die medizinische Forschungsergebnisse veröffentlichen, und Sie werden mit Statistiken überhäuft: *t*-Tests, *p*-Werte (Signifikanzwerte), Cox-Regressionen, Propensity-Score-Analysen (Verwirklichungstendenzen), logistische (Logit-Modell) oder lineare Regressionen mit der Methode der kleinsten Quadrate und Konfidenzintervalle (Erwartungsbereiche). Statistiker haben Wissenschaftlern außerordentlich mächtige Werkzeuge an die Hand gegeben, um Sinn und Ordnung in den kompliziertesten Datensammlungen aufzuspüren – und die Wissenschaftler haben sie mit Freuden entgegengenommen.

Sie haben sich allerdings weder einer *Statistikausbildung* unterzogen, noch ist in vielen wissenschaftlichen Studiengängen eine wie auch immer geartete Schulung für den Einsatz von Statistiken vorgesehen.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Seit den 1980er Jahren dokumentieren Forscher immer wieder statistische Irrtümer und Missverständnisse in der einschlägigen wissenschaftlichen Literatur und beanstanden, dass viele wissenschaftliche Publikationen (womöglich mehr als die Hälfte) diesen Fehlschlüssen zum Opfer fallen. Statistisches Unvermögen führt häufig dazu, dass die eigentlich gesuchten Erkenntnisse in vielen Studien gar nicht erst gefunden bzw. ermittelt werden können: Mehrfache Vergleiche und fehlinterpretierte Signifikanzwerte verursachen reihenweise falsche Positive, anpassungsfähige Datenanalysen begünstigen das Auffinden eigentlich gar nicht vorhandener Korrelationen, und die Auswahl ungeeigneter Modelle sorgt für die Verzerrung maßgeblicher Ergebnisse. Die meisten dieser Fehler werden von Fachkollegen und Redakteuren nicht bemerkt, weil sie oftmals nicht über eine spezielle statistische Ausbildung verfügen – denn nur wenige Fachmagazine beschäftigen »echte« Statistiker, um die eingereichten Beiträge zu überprüfen. Abgesehen davon liefern ohnehin nur die wenigsten Veröffentlichungen hinreichend genaue statistische Daten, um sie überhaupt korrekt bewerten zu können.

Das Problem besteht nicht etwa in betrügerischen Absichten, sondern vielmehr in mangelhaften statistischen Kenntnissen – und zwar derart mangelhaften, dass einige Wissenschaftler mutmaßen, dass die meisten veröffentlichten Forschungsergebnisse vermutlich falsch sind.[1] In den führenden Fachmagazinen finden sich regelmäßig Rezensionen und Leitartikel, die ein höheres statistisches Niveau und strengere Überprüfungen der Publikationen verlangen. Diesen Forderungen schenken allerdings nur wenige Wissenschaftler Gehör, und auch die von den Fachzeitschriften verfügten Standards werden häufig missachtet. Da die Anleitungen und Empfehlungen zum Umgang mit statistischen Daten auf oftmals irreführende Lehrbücher, Rezensionen in den verschiedensten Fachmagazinen und nur schwer verständliche Forschungsarbeiten zum Thema Statistik verteilt sind, haben die meisten Wissenschaftler größte Mühe, ihre Statistikkenntnisse zu verbessern.

Die Komplexität der Verfahrensweisen moderner Forschung bringt es mit sich, dass Wissenschaftler ohne umfassende Ausbildung im Bereich der Statistik womöglich einen Großteil der in ihren jeweiligen

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Fachbereichen veröffentlichten Forschungsarbeiten gar nicht wirklich umfassend begreifen können. Im Bereich Medizin beispielsweise hätte ein Arzt, der einen Standardeinführungskurs in Statistik absolviert hat, gerade mal ausreichende Kenntnisse, um lediglich ein Fünftel der im *New England Journal of Medicine* veröffentlichten Forschungsartikel komplett zu verstehen.[2] Die meisten Ärzte verfügen allerdings noch nicht einmal über dieses statistische Basiswissen – viele der in Ausbildung befindlichen Ärzte erwerben ihre dahingehenden Kenntnisse in Journal Clubs oder kurzen Lehrgängen statt in den eigentlich erforderlichen Kursen.[3] Zudem werden die den Medizinstudenten dort vermittelten Inhalte oft kaum wirklich verstanden – im Durchschnitt werden weniger als 50% der Fragen zu den in der Medizin gebräuchlichen statistischen Methoden korrekt beantwortet.[4] Selbst an Lehreinrichtungen mit medizinischem Fachbereich, die Schulungen für die Forschung anbieten, werden weniger als 75% richtige Antworten erreicht.

Die Situation ist tatsächlich so verfahren, dass sogar die Autoren von Studien über Statistikkenntnisse selbst nicht über das erforderliche Statistikwissen verfügen, um die Fragen korrekt zu formulieren – die soeben genannten Zahlen sind daher irreführend, denn die Befragung der in Ausbildung befindlichen Ärzte enthielt eine Frage, bei der es um die Definition eines *p*-Wertes ging, mit einer Multiple-Choice-Vorgabe, die allerdings vier fehlerhafte Antworten zur Auswahl stellte.[5] Wir können hier jedoch etwas Nachsicht mit den Autoren üben, weil viele einführende Lehrbücher zum Thema Statistik dieses grundlegende Konzept ebenfalls mangelhaft oder falsch definieren.

Wenn die Entwickler wissenschaftlicher Studien bei der Verwendung statistischer Werte nicht ausreichend Vorsicht walten lassen, werden womöglich jahrelange Anstrengungen und zig Tausende Euros in Forschungsprojekte gesteckt, die letztlich gar keine Antworten auf die Fragen liefern, zu deren Klärung sie eigentlich beitragen sollen. Der Psychologe Paul Meehl beschwerte sich darüber wie folgt:

> *Zwischenzeitlich hat unser bienenfleißiger Forscher, unbeeindruckt von wissenschaftlicher Logik und angesichts der »Exaktheit« der statistischen Überprüfung von Hypothesen glückselig, eine lange Liste von Veröffentlichungen produziert und wurde zum Professor*

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

> *berufen. Was seine Beiträge zu den nachhaltigen Erkenntnissen der Psychologie angeht: Er hat praktisch nichts geleistet. Tatsächlich stellt er einen zwar potenten, aber unfruchtbaren intellektuellen Lebemann dar, der auf seinem Lebensweg fröhlich reihenweise geschändete Jungfern hinter sich lässt, jedoch keinerlei brauchbaren wissenschaftlichen Nachwuchs zeugt.*[6]

Nun ist es vielleicht nicht ganz fair, einem Großteil der Wissenschaftler intellektuelle Unfruchtbarkeit zu unterstellen, denn schließlich hängen die meisten wissenschaftlichen Forschungsbereiche nicht von ein paar fehlinterpretierten *p*-Werten ab. Allerdings haben solche fehlerhaften Rückschlüsse im wahren Leben erhebliche Auswirkungen: Klinische medizinische Studien beeinflussen unser Gesundheitswesen und definieren, wie sicher wirksame verschreibungspflichtige Medikamente sind, Kriminologen werten auf der Grundlage statistischer Werte verschiedene Strategien zur Eindämmung von Verbrechen aus, Epidemiologen versuchen anhand entsprechender Daten die Ausbreitung neuer Krankheiten zu verhindern, und Vermarkter und Geschäftsleute suchen nach den besten Möglichkeiten, ihre Produkte zu verkaufen – letztendlich dreht sich alles um Statistiken. Um fehlerbehaftete Statistiken.

Wer sich schon mal über Ärzte beschwert hat, die sich keine Gedanken über das Wohlergehen der Patienten machen, kennt das Ausmaß dieses Problems. Mittlerweile überwiegt eine ablehnende Haltung gegenüber neuen Studien, die behaupten, dass uns ein bestimmtes Lebensmittel, eine bestimmte Ernährungsweise oder eine körperliche Ertüchtigungsform schadet – inzwischen warten wir lieber erst mal auf die einige Monate später unweigerlich erscheinende zweite Studie, die dann höchstwahrscheinlich das genaue Gegenteil zum Ergebnis hat. Ein prominenter Epidemiologe merkte dazu einmal an: »Wir werden zusehends zu einem gesellschaftlichen Ärgernis. Die Leute nehmen uns nicht mehr ernst – und wenn sie es doch tun, richten wir womöglich unabsichtlich mehr Schaden als Nutzen an.«[7] Unsere Instinkte funktionieren. In vielen Fachbereichen fallen die Ergebnisse der ursprünglichen und dann nachfolgend durchgeführten Untersuchungen zu ein und demselben Thema tendenziell widersprüchlich aus. Es scheint,

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

dass der Druck, möglichst schnell und häufig aufsehenerregende neue Erkenntnisse zu veröffentlichen, Vorrang über die Verantwortung gewonnen hat, sorgfältig überprüfte Resultate zu publizieren, die durch mehr als nur gerade mal hinreichende Fakten belegt sind.

Wir sollten jedoch nicht vorschnell urteilen. Manche statistischen Fehlschlüsse sind schlicht auf die mangelnde Bereitstellung finanzieller Mittel oder das Fehlen von Ressourcen zurückzuführen. Denken Sie beispielsweise an die Bestrebungen Mitte der 1970er Jahre zurück, den US-Autofahrern zu gestatten, an roten Ampeln rechts abzubiegen, um Zeit und Benzin zu sparen (ähnlich dem Grünpfeil in Deutschland). Die hier angewandte statistische Beweisführung zum Beleg dafür, dass dadurch keine zusätzlichen Unfälle verursacht werden würden, war, wie Sie später noch sehen werden, einfach unzureichend – und in der Konsequenz kostete die Einführung dieser neuen Verkehrsregel leider viele Menschenleben. In diesem Fall waren einzig und allein fehlende weiterführende Daten der Grund dafür, dass die Verkehrssicherheitsforscher nicht einschritten. Hätten sie über die finanziellen Mittel verfügt, um mehr Daten zu sammeln und weitere Untersuchungen durchzuführen (sowie über die nötige Zeit, um die Ergebnisse unabhängiger Forscher aus verschiedenen Bundesstaaten abzugleichen), wäre die traurige Wahrheit von vornherein offensichtlich gewesen.

Die als »Hanlon's Razor« (im Deutschen manchmal auch »Hanlons Rasiermesser« genannt) bekannte These besagt: »Schreibe nichts der Bösartigkeit zu, was durch Dummheit hinreichend erklärbar ist« – und in der Tat wurden einige Ergebnisse veröffentlicht, die der Kategorie »Lügen, verdammte Lügen und Statistiken« zuzuordnen sind. Für die pharmazeutische Industrie scheint es besonders verlockend zu sein, die Beweislage zu beeinflussen, indem Studien, die belegen, dass Medikamente wirkungslos sind, zurückgehalten bzw. gar nicht erst publik gemacht werden.* Wer dann die relevanten Veröffentlichungen zu

* An den statistischen Missgeschicken der Pharmaindustrie interessierten Lesern sei das Buch *Bad Pharma: How drug companies mislead doctors and harm patients* (zu Deutsch: *Die Pharma-Lüge: Wie Arzneimittelkonzerne Ärzte irreführen und Patienten schädigen*) von Ben Goldacre empfohlen, dessen Lektüre bei mir einen statistisch signifikanten Blutdruckanstieg auslöste.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

einem dieser Präparate nachliest, stellt vermutlich erfreut fest, dass 12 Studien auf die effektive Wirksamkeit eines Medikaments hinweisen – allerdings ohne sich der Existenz von 8 unveröffentlichten Studien bewusst zu sein, die seine Wirkungslosigkeit belegen. Andererseits würden die Fachmagazine solche Ergebnisse aber wahrscheinlich ohnehin nicht veröffentlichen, selbst wenn sie eingereicht würden. Die starke Voreingenommenheit gegenüber unspektakulären Resultaten bedeutet, dass Studien, die zu dem Ergebnis »Funktioniert nicht!« kommen, nie erscheinen und anderen Forschern vorenthalten bleiben. Fehlende Daten und von Befangenheit geprägte Publikationen sind die Plagen der Wissenschaft und verzerren die Wahrnehmung wichtiger Sachverhalte.

Aber selbst korrekt erstellten Statistiken ist nicht unbedingt zu trauen. Die Vielfalt der zur Verfügung stehenden Verfahren und Methoden lässt Forschern einen enormen Spielraum bei der Analyse ihrer Daten. Es ist kinderleicht, »die Daten zu foltern, bis sie gestehen«. Probieren Sie einfach der Reihe nach die verschiedenen Analysemethoden aus, die Ihre Statistiksoftware bietet, bis eine davon ein interessantes Ergebnis liefert – und behaupten Sie dann, dass es sich dabei um diejenige Methode handelt, die Sie ohnehin von Anfang an einsetzen wollten. Ohne übernatürliche Kräfte ist es fast unmöglich festzustellen, ob ein veröffentlichtes Ergebnis durch »Datenfolter« zustande gekommen ist oder nicht.

Es ist schwierig, Experimente für »weichere« Fachbereiche zu entwickeln, in denen Theorien weniger quantitativ und die Methoden weniger standardisiert sind. Diese zusätzliche Freiheit verursacht ebenfalls spürbare Verzerrungseffekte.[8]

Insbesondere US-amerikanische Forscher sind gezwungen, interessante Ergebnisse vorzuweisen und zu veröffentlichen, um ihre Karriere voranzutreiben. Aufgrund des intensiven Wettbewerbs um eine geringe Anzahl akademischer Stellen können es sich die Wissenschaftler nicht leisten, monate- oder jahrelang Daten zu sammeln und zu analysieren, um am Ende lediglich statistisch nicht signifikante Ergebnisse zu erzielen. Auch ohne böse Absicht neigen diese Wissenschaftler dazu, über-

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

triebene Ergebnisse zu liefern, die ihre Hypothesen stärker belegen, als es die Daten tatsächlich hergeben.

Auf den nachfolgenden Seiten möchte ich Ihnen solche gängigen Verfehlungen, Missinterpretationen und vieles andere vorstellen. Viele dieser Fehler sind in der unüberschaubaren Menge der Veröffentlichungen weit verbreitet und ziehen die Ergebnisse Tausender Arbeiten in Zweifel.

In den vergangenen Jahren wurden Stimmen laut, die eine Reform der Statistiklandschaft fordern. Naturgemäß gibt es verschiedene Ansichten zu der Frage, wie die Probleme am besten in Angriff genommen werden sollten. Manche bestehen darauf, dass *p*-Werte, die, wie ich noch aufzeigen werde, oft irreführend und verwirrend sind, komplett abgeschafft werden sollten. Andere befürworten eine »neue Statistik«, die auf Konfidenzintervallen beruht. Und es gibt andere, die den Wechsel zu Bayes'schen Methoden fordern, die besser interpretierbare Ergebnisse liefern. Eine weitere Fraktion ist der Meinung, dass die derzeit gelehrte Statistik völlig in Ordnung ist, jedoch mangelhaft angewendet wird. All diese Positionen haben ihre Berechtigung und ich werde in diesem Buch keine davon bevorzugen. Ich möchte mich vielmehr auf die Statistik an sich konzentrieren, wie sie gegenwärtig von praktisch tätigen Wissenschaftlern genutzt wird.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

1

Eine Einführung in die statistische Signifikanz

Ein Großteil der experimentellen Wissenschaft beruht letzten Endes auf der Messung von Unterschieden. Wirkt ein Medikament besser als ein anderes? Produzieren Zellen mit einer bestimmten Variante eines Gens mehr Enzyme als Zellen mit einer anderen Variante des Gens? Kann ein Algorithmus zur Signalverarbeitung Pulsare besser erkennen als ein anderer? Ist ein Katalysator bei der Beschleunigung einer chemischen Reaktion effektiver als ein anderer?

Wir nutzen Statistiken, um derartige Unterschiede beurteilen zu können. Nun lassen sich im Rahmen von Vergleichen allerdings immer *irgendwelche* Unterschiede feststellen, sei es durch Zufall oder aufgrund unbedeutender Schwankungen, daher sprechen Statistiker von *statistisch signifikanten* Unterschieden, wenn die Abweichungen größer sind als durch rein zufällige Fluktuationen zu erwarten wäre. Wir müssen also zunächst einmal verstehen, wie diese Entscheidung getroffen werden kann.

1.1 p-Werte

Nehmen wir an, Sie testen Erkältungsmittel. Ihr neues Medikament soll die Dauer der Erkältungssymptome angeblich um einen Tag verkürzen. Um das zu beweisen, suchen Sie nach 20 erkälteten Patienten

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

und verabreichen der einen Hälfte das neue Mittel, während die andere Hälfte ein Placebo erhält. Dann verfolgen Sie den Verlauf der Erkältung, um herauszufinden, wie lange sie durchschnittlich mit und ohne Einnahme der Medizin dauert.

Nun laufen aber nicht alle Erkältungen auf dieselbe Weise ab. Eine normale Erkältung dauert vielleicht eine Woche, manche jedoch nur ein paar Tage. Andere ziehen sich wiederum über zwei Wochen oder länger hin. Es wäre daher denkbar, dass die Gruppe der 10 Patienten, die tatsächlich die Medizin erhalten haben, an einer ziemlich kurzlebigen Erkältung erkrankt waren. Wie können Sie also beweisen, dass die Medizin wirkt, statt nur zu belegen, dass einige Patienten schlicht und einfach Glück hatten?

Die Antwort liefern statistische Tests zur Überprüfung einer Hypothese. Wenn Ihnen die Verteilung typischer Erkältungsfälle bekannt ist, also wie viele Patienten ungefähr an Erkältungen von kurzer, langer und durchschnittlicher Dauer erkranken, können Sie die Wahrscheinlichkeit dafür ermitteln, dass alle stichprobenartig ausgewählten Patienten einer Studiengruppe an einer Erkältung leiden, die länger oder kürzer als eine durchschnittliche Erkältung dauert. Die Durchführung eines *Hypothesentests* (der auch als *Signifikanztest* bezeichnet wird) beantwortet die Frage: »Angenommen, mein Medikament ist vollkommen wirkungslos, wie groß ist dann die Wahrscheinlichkeit, dass mein Experiment das beobachtete Ergebnis liefert?«

Wenn Sie Ihr Medikament nur an einer einzigen Person testen, ist es nicht allzu verwunderlich, wenn die Erkältung etwas früher als gewöhnlich vorübergeht – nur die wenigsten Erkältungen entsprechen exakt dem Durchschnitt. Wenn Sie die Medizin allerdings an 10 Millionen Patienten testen, ist es äußerst unwahrscheinlich, dass all diese Patienten zufällig an einer kurzen Erkältung erkrankt waren. In diesem Fall ist es wahrscheinlicher, dass Ihr Mittel tatsächlich wirkt.

Dieses anschauliche Verhalten wird von Wissenschaftlern durch den *p-Wert* quantifiziert, für den auch die Bezeichnung *Signifikanzwert* verwendet wird. Ausgehend von der Annahme, dass ein echter Effekt oder ein wirklicher Unterschied nicht vorhanden ist, gibt der *p*-Wert die

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Wahrscheinlichkeit dafür an, dass die gesammelten Daten mindestens dem tatsächlich beobachteten Wert entsprechen oder einen extremeren Wert annehmen.

Wenn Sie Ihr Medikament also 100 Patienten verabreichen und feststellen, dass die Erkältungen im Durchschnitt einen Tag kürzer dauern, dann beschreibt der *p*-Wert dieses Ergebnisses die Wahrscheinlichkeit dafür, dass die durchschnittliche Erkältungsdauer, sofern Ihr Medikament überhaupt nicht gewirkt hat, rein zufällig einen Tag weniger beträgt als die Erkrankungszeit der Kontrollgruppe. Wie Sie sicherlich schon vermutet haben, ist der *p*-Wert einerseits von der Größe des Effekts abhängig – vier Tage kürzere Erkältungen sind definitiv seltener als nur einen Tag kürzere –, und zum anderen spielt in diesem Kontext natürlich auch die Anzahl der Patienten eine Rolle, an denen das Erkältungsmittel getestet wird.

Der *p*-Wert ist also **kein** Maßstab dafür, wie richtig (oder falsch) Sie mit Ihrer Hypothese liegen oder wie bedeutsam ein Unterschied im Endeffekt ist. Betrachten Sie ihn lieber als einen Indikator dafür, wie unerwartet oder überraschend ein Ergebnis ausfällt. Wenn Sie davon ausgehen, dass Ihr Mittel wirkungslos ist und das Auftreten von Unterschieden zwischen den beiden Testgruppen lediglich Zufall wäre, bedeutet ein kleinerer *p*-Wert, dass Abweichungen umso überraschender und unerwarteter zu verzeichnen sind – oder aber die zugrunde liegende Hypothese ist falsch und die Medizin hilft wirklich.

Und wie kann man anhand des *p*-Wertes beurteilen, ob es tatsächlich einen Unterschied zwischen beiden Gruppen gibt? Als Faustregel gilt, dass ein Unterschied dann signifikant ist, wenn $p < 0{,}05$ ist. Für den Wert 0,05 gibt es keinen speziellen logischen oder statistischen Grund, er ist vielmehr durch jahrzehntelangen Gebrauch zu einer Konvention geworden.

Beachten Sie bitte, dass der *p*-Wert auf der Annahme beruht, dass es keinen Unterschied zwischen den am Experiment teilnehmenden Gruppen gibt. Diese Eigenschaft von Signifikanztests ist nicht gerade intuitiv: Wenn Sie beweisen möchten, dass Ihr Medikament tatsächlich wirkt, müssen Sie belegen, dass die beobachteten Daten *nicht* den

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Rückschluss zulassen, dass die Medizin *ohne* Wirkung ist. Dementsprechend können Sie *p*-Werte jederzeit dazu verwenden, Hypothesen zu widerlegen, die sich mathematisch formulieren lassen.

Allerdings haben *p*-Werte auch ihre Grenzen. Wie Sie wissen, ist *p* ein Maß für den »Überraschungsfaktor«, wobei ein kleinerer Wert ein umso überraschenderes Ergebnis bedeutet. Der *p*-Wert ist jedoch kein Indikator für die Größe des Effekts: Sie können bei einem enormen Effekt einen sehr kleinen *p*-Wert erhalten (»Dank dieser Medizin werden die Leute vier Mal älter«) oder einen winzigen Effekt mit großer Sicherheit messen. Und da jede Behandlung oder Verabreichung von Medikamenten für gewöhnlich *irgendeinen* Effekt hat, kann man immer statistisch signifikante Auswirkungen feststellen, wenn man so viele Daten sammelt, dass sich äußerst kleine, aber relativ unbedeutende Effekte finden lassen. Bruce Thompson schrieb dazu:

> *Statistische Signifikanztests sind manchmal mit einer tautologischen Vorgehensweise verbunden: Erschöpfte Forscher, die Daten über Hunderte von Studienobjekten gesammelt haben, führen erst danach statistische Tests durch, um zu überprüfen, ob sich unter diesen Objekten nicht bereits viele bekannte befinden – denn schließlich haben sie deren Daten schon erfasst und ihnen ist klar, dass sie erschöpft sind. Diese Tautologie hat hinsichtlich des Zusammentragens von Wissen beträchtlichen Schaden angerichtet.*[1]

Zusammengefasst heißt das: Eine statistische Signifikanz bedeutet *nicht*, dass ein Ergebnis *in der Praxis* irgendeine Signifikanz besitzt. Und statistisch *nicht* signifikante Resultate besagen nicht viel. Bei statistisch nicht signifikanten Unterschieden könnte es sich um rein zufällige Schwankungen (»Rauschen«) handeln. Sollte es tatsächlich einen Effekt geben, sind jedoch mehr Daten erforderlich, um ihn zu belegen.

Es gibt kein mathematisches Instrument, mit dem Sie feststellen können, ob eine Hypothese richtig oder falsch ist – Sie können lediglich überprüfen, ob sie mit den Daten konsistent ist. Wenn allerdings nur wenige oder unklare Daten vorliegen, werden auch die daraus gezogenen Schlussfolgerungen unsicher sein.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

1.1.1 »Übernatürliche« Statistik

Neben den genannten Einschränkungen ergeben sich zudem noch einige subtilere Schwierigkeiten mit den *p*-Werten. Der Berechnung eines *p*-Wertes liegt die Annahme zugrunde, dass im Rahmen des Experiments ausschließlich der Zufall eine Rolle spielt – und nicht die infrage stehende Behandlung oder das Medikament. Der *p*-Wert ist definiert als die Wahrscheinlichkeit, mindestens das tatsächlich beobachtete *oder ein noch stärker abweichendes* Ergebnis zu erzielen. *p*-Werte zwingen Sie also dazu, über Ergebnisse nachzudenken, die niemals eingetreten sind, nämlich diejenigen, die noch extremer als Ihres sind. Die Wahrscheinlichkeit, solch ein Ergebnis zu erzielen, hängt dabei von der Art des Experiments ab – und das lässt *p*-Werte fast schon »übernatürlich« erscheinen: Zwei unterschiedlich angelegte Experimente können identische Daten, aber verschiedene *p*-Werte liefern, weil sich die *nicht beobachteten* Daten unterscheiden.

Nehmen wir einmal an, ich würde Ihnen 12 Fragen über statistische Schlussfolgerungen stellen, die mit richtig oder falsch zu beantworten sind, und Sie beantworten 9 davon korrekt. Und nun möchte ich die Hypothese testen, dass Sie die Fragen durch reines Raten beantwortet haben. Zu diesem Zweck muss ich die Wahrscheinlichkeit berechnen, dass Sie *mindestens* 9 richtige Antworten erzielen, wenn Sie bei jeder Frage zufällig richtig oder falsch auswählen. Wenn ich davon ausgehe, dass Sie sich mit derselben Wahrscheinlichkeit für richtig oder falsch entscheiden, erhalte ich $p = 0{,}073$.* Es gilt also $p > 0{,}05$ – und damit ist es plausibel, dass Sie geraten haben. Wenn Sie raten, werden Sie in 7,3% der Fälle 9 oder mehr richtige Antworten geben.[2]

Aber vielleicht hatte ich ursprünglich gar nicht vor, Ihnen nur 12 Fragen zu stellen. Ich könnte doch auch einen Computer nutzen, der eine

* Ich habe dieser Berechnung eine Binomialverteilung zugrunde gelegt. Im nächsten Absatz berechne ich *p* mit einer anderen Verteilung, die als *negative Binomialverteilung* oder *Pascal-Verteilung* bezeichnet wird. Eine ausführliche Erläuterung von Wahrscheinlichkeitsverteilungen würde den Rahmen dieses Buches sprengen – an dieser Stelle interessiert nur, wie *p*-Werte zu interpretieren sind, nicht jedoch ihre Berechnung.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

endlose Zahl von Fragen erzeugt und Sie so lange damit malträtieren, bis Sie 3 falsche Antworten gegeben haben. In diesem Fall muss ich die Wahrscheinlichkeit berechnen, dass Sie nach 15, 20 oder 47 gestellten Fragen 3 davon falsch beantworten. Ich muss nun sogar die entfernte Möglichkeit berücksichtigen, dass Sie 175.231 Fragen schaffen, bevor Sie 3 Fragen nicht korrekt beantwortet haben. Wenn ich das ausrechne, erhalte ich $p = 0{,}033$. Damit gilt $p < 0{,}05$, also schlussfolgere ich, das es unwahrscheinlich ist, dieses Ergebnis durch reines Raten zu erzielen.

Das ist beunruhigend: Zwei Experimente liefern dieselben Daten, die jedoch zu verschiedenen Schlussfolgerungen führen. Auf irgendeine Art und Weise erkennt der *p*-Wert Ihr Vorhaben.

1.1.2 Neyman-Pearson-Tests

Um die Schwierigkeiten mit dem *p*-Wert besser zu verstehen, sind Kenntnisse der historischen Entwicklung der Statistik hilfreich. In Bezug auf statistische Signifikanztests gibt es zwei grundsätzliche Thesen. Die erste dieser Denkweisen wurde in den 1920er Jahren durch Ronald Aylmer Fisher bekannt gemacht. Fisher betrachtete *p* als praktische, zwanglose Methode, um zu beschreiben, wie überraschend eine Datenmenge sein kann, und nicht als Teil eines streng formalen Verfahrens für Hypothesentests. Kombiniert man den *p*-Wert mit der vorausgehenden Erfahrung des Experimentators und etwas Fachwissen, kann er sich im Hinblick auf die Entscheidung, wie neue Daten zu interpretieren sind, als hilfreich erweisen.

Nach der Veröffentlichung von Fishers Arbeiten nahmen Jerzy Neyman und Egon Pearson einige noch offene Fragen in Angriff. Bei dem Erkältungsmittelexperiment können Sie die beiden Gruppen beispielsweise anhand des Medianwertes, des Durchschnittswertes oder irgendeiner anderen Formel miteinander vergleichen, sofern Sie für den Vergleich einen *p*-Wert herleiten können. Woher aber soll man wissen, welcher Wert am besten geeignet ist? Was bedeutet »am besten« bei einem Hypothesentest überhaupt?

In der Wissenschaft sollten zwei Arten von Fehlern soweit wie möglich ausgeschlossen werden:

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

- *Falsche Positive* – Es wird geschlussfolgert, dass es einen Effekt gibt, der tatsächlich nicht vorhanden ist. Man spricht hier auch von einem »Fehler 1. Art«.
- *Falsche Negative* – Ein tatsächlich vorhandener Effekt wird übersehen (»Fehler 2. Art«).

In gewisser Hinsicht sind falsche Positive und falsche Negative zwei Seiten derselben Medaille: Wenn wir allzu bereit sind, auf einen Effekt zu schließen, laufen wir Gefahr, falsche Positive zu erhalten – sind wir hingegen zu zurückhaltend, begehen wir den Fehler, einen Effekt nicht zu erkennen und erhalten falsche Negative.

Neyman und Pearson argumentierten, dass es zwar unmöglich ist, falsche Positive und falsche Negative völlig auszuschließen, aber sehr wohl möglich, ein formales Entscheidungsverfahren zu entwickeln, das gewährleistet, dass falsche Positive nur gemäß einer vorgegebenen Rate auftreten. Diese Rate bezeichneten sie mit α. Nach ihrer Theorie sollen Experimentatoren anhand ihrer Erfahrung und Erwartung einen Wert für α festlegen. Wenn wir also beispielsweise bereit sind, 10% falsche Positive in Kauf zu nehmen, setzen wir $\alpha = 0{,}1$. Wenn wir bei der Beurteilung konservativer sein müssen, setzen wir $\alpha = 0{,}01$ oder noch niedriger. Um das beste Testverfahren zu ermitteln, überprüfen wir, welches die niedrigste Anzahl falscher Positiver bei vorgegebenen α liefert.

Und wie funktioniert das in der Praxis? Nach dem Neyman-Pearson-System definieren wir eine *Nullhypothese*, die besagt, dass es keinen Effekt gibt, sowie eine *Alternativhypothese* wie beispielsweise »Der Effekt ist größer als null«. Dann erstellen wir einen Test, der diese beiden Hypothesen miteinander vergleicht, und berechnen, welches Ergebnis wir erwarten würden, falls die Nullhypothese zuträfe. Wir verwenden den p-Wert, um das Neyman-Pearson-Testverfahren zu implementieren, indem wir die Nullhypothese zurückweisen, sofern $p < \alpha$ gilt. Im Gegensatz zu Fishers Verfahren berücksichtigt diese Methode bewusst nicht die Beweiskraft irgendeines bestimmten Experiments. Hier sind wir nur daran interessiert, ob die Hypothese abzulehnen ist oder nicht. Der p-Wert wird nicht zum Vergleich von Experimenten

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

oder für andere Schlussfolgerungen als »Die Nullhypothese ist zurückzuweisen« genutzt. Neyman und Pearson schreiben:

> *Was eine bestimmte Hypothese betrifft, neigen wir zu der Auffassung, dass kein auf der Wahrscheinlichkeitstheorie basierender Test allein einen brauchbaren Beweis für die Richtigkeit oder Falschheit dieser Hypothese liefern kann.*
>
> *Wir können den Zweck von Tests jedoch auch aus einer anderen Perspektive betrachten: Statt darauf zu hoffen, erkennen zu können, ob jede einzelne Hypothese wahr oder falsch ist, können wir nach Regeln suchen, die unsere Beurteilung der Hypothesen bestimmen. Wenn wir diesen Regeln folgen, sollten wir der langfristigen Erfahrung nach nicht allzu oft zu einem falschen Ergebnis kommen.*[3]

Der Ansatz von Neyman und Pearson unterscheidet sich konzeptionell zwar von Fishers Vorgehensweise, dennoch werden beide Verfahren von praktisch tätigen Wissenschaftlern häufig kombiniert.[4,5,6] Mit einem vorab gewählten Schwellenwert für p, der langfristig eine bestimmte Falsch-Positiv-Rate gewährleistet, liefert der Neyman-Pearson-Ansatz »statistische Signifikanz«. Nehmen Sie nun aber an, dass Sie ein Experiment durchführen und $p = 0{,}032$ erhalten. Wenn Sie als Schwellenwert das konventionelle $p < 0{,}05$ verwenden, ist der Wert statistisch signifikant. Er wäre allerdings ebenfalls statistisch signifikant, wenn Ihr Schwellenwert $p < 0{,}033$ gewesen wäre. Es ist daher verlockend – und eine häufige Fehlinterpretation – zu sagen: »Meine Falsch-Positiv-Rate beträgt 3,2%.«

Diese Aussage ergibt jedoch keinen Sinn: Ein einzelnes Experiment besitzt keine Falsch-Positiv-Rate. Diese Rate wird durch Ihr *Verfahren* bestimmt, nicht durch das Ergebnis irgendeines einzelnen Experiments. Sie können nicht behaupten, dass jedes einzelne Experiment eine Falsch-Positiv-Rate von exakt p besitzt (der eigentliche Wert spielt hier keine Rolle), wenn Sie ein bestimmtes Verfahren einsetzen, um langfristig eine Falsch-Positiv-Rate von α zu erzielen.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

1.2 Intervallen vertrauen

Seitdem der Begriff »statistisch signifikant« in jedem Lexikon zu finden ist, erhalten Signifikanztests tendenziell viel Aufmerksamkeit. Insbesondere Forschungsergebnisse aus der Biologie und den Sozialwissenschaften enthalten oft *p*-Werte. *p* ist jedoch keineswegs die einzige Möglichkeit, die Beweiskraft zu beurteilen. Mit *Konfidenzintervallen* (die auch als *Erwartungsbereiche* bezeichnet werden) lassen sich dieselben Fragen beantworten wie mit *p*-Werten, sie besitzen jedoch den Vorteil, dass sie mehr Informationen liefern und einfacher zu interpretieren sind.

Ein Konfidenzintervall umfasst die Schätzung eines Wertes sowie die Unsicherheit dieser Schätzung. Sie könnten beispielsweise sagen, dass Ihr neues experimentelles Medikament die durchschnittliche Dauer einer Erkältung um 36 Stunden verkürzt und ein 95%-Konfidenzintervall von 24 bis 48 Stunden angeben. (Das Konfidenzintervall bezieht sich auf die *durchschnittliche* Dauer; bei einzelnen Patienten kann die Dauer stark variieren.) Wenn Sie dasselbe Experiment 100 Mal durchführen, wird der Wert, den Sie messen möchten, etwa 95 Mal im Konfidenzintervall enthalten sein.

Das Konfidenzintervall quantifiziert die Unsicherheit Ihrer Schlussfolgerungen und stellt erheblich mehr Informationen als ein *p*-Wert zur Verfügung, der nichts über die Größe des Effekts aussagt. Wenn Sie testen möchten, ob sich ein Effekt signifikant von null unterscheidet, können Sie ein 95%-Konfidenzintervall verwenden und überprüfen, ob es null enthält. Bei diesem Vorgang erfahren Sie außerdem, wie präzise die Schätzung ist. Sollte das Konfidenzintervall zu breit sein, benötigen Sie weitere Daten.

Nehmen wir beispielsweise an, Sie würden eine klinische Studie durchführen und es ergäbe sich ein Konfidenzintervall, das darauf hinweist, dass Ihr Medikament die Symptome zwischen 15 und 25% reduziert. Der Effekt wäre damit statistisch signifikant, denn null ist nicht im Konfidenzintervall enthalten. Nun können Sie die Bedeutung dieses Unterschieds anhand Ihres klinischen Wissens über die fragliche

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Krankheit beurteilen. Dieser Schritt ist bei der Verwendung von *p*-Werten wichtig – Sie sollten das Ergebnis besser nicht als große Entdeckung herausposaunen, ohne es im richtigen Kontext bewertet zu haben. Wenn das Symptom ohnehin ziemlich harmlos ist, wäre eine Verbesserung von 15 bis 25% kaum von Bedeutung. Würde es sich hingegen um ein Symptom wie die spontane menschliche Selbstentzündung handeln, dürften Sie sich in diesem Fall über jede noch so kleine Verbesserung freuen.

Wenn ein Ergebnis als Konfidenzintervall statt als *p*-Wert darstellbar ist, sollten Sie dies auch nutzen.[7] Konfidenzintervalle umgehen die meisten mit *p*-Werten einhergehenden Schwierigkeiten bei der Interpretation und machen die Ergebnisse dadurch verständlicher. Warum aber sind Konfidenzintervalle dann so unbeliebt? In Fachmagazinen zur experimentellen Psychologie enthalten 97% der Arbeiten Signifikanztests, aber nur 10% geben Konfidenzintervalle an – und selbst diese Arbeiten nutzen die Intervalle kaum zu dem Zweck, die Schlussfolgerungen zu belegen, sondern verwenden dafür wiederum Signifikanztests.[8] Selbst das renommierte Magazin *Nature* bleibt hier zurück: 89% der Artikel enthalten *p*-Werte ohne Konfidenzintervalle oder Angaben zur Größe der Effekte, was es unmöglich macht, die Ergebnisse im Kontext zu interpretieren.[9] Einer der Redakteure merkte hierzu an »*p*-Werte sind wie Moskitos«, denn »sie belegen eine evolutionäre Nische, und hier hilft (leider) alles Kratzen, Zerquetschen und Sprühen nichts, um sie zu vertreiben«.[10]

Eine mögliche Erklärung für die fehlenden Konfidenzintervalle ist die Tatsache, dass sie oft beschämend breit sind.[11] Ein weiteres Argument könnte sein, dass der Konformitätsdruck auf die von Fachkollegen begutachtete Wissenschaft zu groß ist. Am besten nutzt man Statistiken so wie alle anderen auch, weil die Gutachter eine Veröffentlichung sonst womöglich ablehnen. Vielleicht verstellen auch die weit verbreiteten Irrtümer über *p*-Werte den Blick auf die Vorteile von Konfidenzintervallen. Oder die übertriebene Bedeutung, die den Hypothesentests in Statistikkursen beigemessen wird, führt dazu, dass die meisten Wissenschaftler gar nicht wissen, wie man Konfidenzintervalle berechnet und verwendet.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Die Herausgeber von Fachmagazinen haben gelegentlich versucht, die Angabe von Konfidenzintervallen zu erzwingen. Kenneth Rothman, Mitherausgeber des *American Journal of Public Health*, wies Mitte der 1980er Jahre eingereichte Arbeiten mit unmissverständlich formulierten Briefen zurück:

> *Alle Verweise auf statistische Hypothesentests und statistische Signifikanz sollten aus der Arbeit entfernt werden. Ich bitte darum, sowohl p-Werte als auch Kommentare über die statistische Signifikanz zu löschen. Wenn Sie meinen Forderungen (hinsichtlich der Unangemessenheit von Signifikanztests) nicht nachkommen, steht es Ihnen frei, darüber zu diskutieren oder das, was Sie für meine fehlgeleitete Ansicht halten, schlicht und einfach zu ignorieren, indem Sie Ihre Arbeit woanders veröffentlichen.*[12]

Während Rothmans dreijähriger Amtszeit als Mitherausgeber sank die Anzahl der Arbeiten, die ausschließlich *p*-Werte verwendeten, radikal. Nach seinem Fortgang traten die Signifikanztests wieder verstärkt in Erscheinung, allerdings gelang es den ihm nachfolgenden Herausgebern, die Forscher dazu bewegen, auch Konfidenzintervalle anzugeben. Leider wurden sie jedoch nur von wenigen Artikelautoren auch erörtert oder für Schlussfolgerungen verwendet – stattdessen bevorzugten sie es, Konfidenzintervalle lediglich wie Signifikanztests zu behandeln.[12]

Rothman gründete dann das Magazin *Epidemiology*, das strenge Richtlinien für die Verwendung von Statistiken vorgab. Anfänglich zogen es mit Signifikanztests vertraute Autoren vor, neben den Konfidenzintervallen auch *p*-Werte anzugeben. 10 Jahre später hatte sich die Einstellung dazu jedoch gewandelt und es wurde gängige Praxis, nur noch Konfidenzintervalle anzugeben.[12]

Vielleicht folgen mutige (und geduldige) Herausgeber Rothmans Beispiel und wirken auch in ihren Fachbereichen auf ein Umdenken hinsichtlich der Verwendungsweise von Statistiken hin.

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

Stichwortverzeichnis

A
Abbruchregel 107, 110
Akaikes Informationskriterium 126
Aktivitätsmuster 103
Alphafehler-Kumulierung 42
Alternativhypothese 29
Amgen 153
ANOVA 119
Autokorrelation 59

B
Baggerly, Keith 146
Basisrate 70
Basisratenfehler 71
Bayer 153
Bayessches Informationskriterium 126
Bem, Daryl 170
Benjamini, Yoav 87
Benjamini-Hochberg-Verfahren 88
Binomialverteilung 27
Blindanalyse 141
Blindstudie 141
Body-Mass-Index (BMI) 116
Bonferroni-Korrektur 84, 98

C
Cholesterinspiegel 93
Cochrane Collaboration 164
Codeoffenlegung 151
Cohen, Jacob 41
CONSORT-Prüfliste 165
Coombes, Kevin 146
CRAPL 151

D
Daten
  unvollständige 165
Datenfolter 21
Datenformat 159
Datenverfall 160
Dichotomisierung 114, 117
Digital Object Identifier (DOI) 160
Doppelte Datennutzung 102
Dryad 160
Dunkelziffer 78
Dunnington, Frank 141

E
Effektgröße 32, 38, 50, 90
Einflussvariable 118
Elektrode 100
eLife 184
Empirismus 175
Erkrankungsrate 52
Erwartungsbereich 31
Europäische Arzneimittel-Agentur (EMA) 157

F
Falsche Negative 29
Falsche Positive 29, 84
Fehler 1. Art 29
Fehler 2. Art 29
Fehlerbalken 96
Figshare 160
Fisher, Ronald Aylmer 28
Food and Drug Administration (FDA) 168
Force Concept Inventory (FCI) 179
Francis, Gregory 170

G
Galton, Francis 105
Gewichtung 52
Gießkannenprinzip 102
Gruppe 113
Gütefunktion 37

H
Hanlon's Razor 20
Hanlons Rasiermesser 20
Higgs-Boson 71
Hirnhautentzündung 133
Hirnregion 102
Hochberg, Yosef 87
Homosexualität 92
Hotelling, Harold 107
Huff, Darrell 16, 74
Hypothese
  statistische 139
Hypothesentest 24, 180
  Teststärke 37

I
Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) 168
Ioannidis, John 175
IPython-Notebook 149
IQ-Test 39
Irrtumswahrscheinlichkeit 71

K
Kanazawa, Satoshi 49
Kausalität 157
Konfidenzintervall 31
  Beeinflussung 46
  Überschneidung 95
Korrelation 128
Kreuzvalidierung 127

L
Lancet 162
Leave-One-Out-Kreuzvalidierung 127
Look-elsewhere-Effekt 83

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252

M

Mammografie 73
Manipulierte Münze 37
McClintock, Martha 64
Medianwertaufteilung 115
Meehl, Paul 18
Messfehler 38
Metaanalyse 177
Microarray 61
Might, Matt 151
Mittelwert
  gewichteter 52
Münzwurf 36

N

National Cancer Institute (NCI) 147
Nebenwirkung 40
Neyman, Jerzy 28
Neymann-Pearson-Test 29
Nicht beobachtete Daten 27
Nullhypothese 29

O

OpenIntro Statistics 181

P

Pascal-Verteilung 27
Pearson, Egon 28
Pheromone 64
Placebo 69
Plausibilität 27
Potti, Anil 147
Power 35
Prävalenzfehler 71
Prüfplan 110
Pseudoreplikation 58
Psi-Effekt 170
Publikationsbias 169, 170
p-Wert 23, 24, 72, 90
  Definition 25
  Konvention 25
Python 149

R

R (Programmiersprache) 149
Randomisierung 57
Raten 28
Rauschen 26
Reboxetin 168
Rechtsabbiegen bei Rot 44
Registrierte Studie 111
Regression 121
  schrittweise 124
Regression zur Mitte 105
Regressionsanalyse 117
Regressionsmodellierung 113
Reifegrad 123
Replikationsstudie 91
Reproduzierbarkeit 148
Richtlinien zur Reproduzierbarkeit 150
Rohdaten 146
Rothman, Kenneth 33
Rückwärtseliminierung 126

S

Schätzung 31
Schekman, Randy 184
Schoenfeld, Jonathan 175
Schrumpfen 52
Schubladenproblem 169
Schusswaffenbesitz 77
Secrist, Horace 106
Sicherheitsgrad 46
Signal/Rauschen 39
Signifikanz 44, 69
Signifikanztest 24, 31, 99
Signifikanzunterschied 89
Signifikanzwert 24
Simpsons Paradoxon 130
Sphygmomanometer 58
Standardabweichung 95
Standardfehler 95
Statin 93
Statistikausbildung 178
Statistikberatung 187
Statistiksoftware 149
Statistische Signifikanz 23
Sterblichkeitsrate 76
Stichprobenumfang 38
Stimuli 101, 103
Störfaktor 63, 118, 122, 128
Streuung 95
Sweave (Software) 149
Synchronisierung 64

T

Teststärke 35
  Berechnung 42
  Stichprobenumfang 42
  unzureichende 39
Thompson, Bruce 26
Tierversuche 41
TP53-Protein 167
Trivers-Willard-Hypothese 49
Truth Inflation 48
t-Test 96

U

Überanpassung 123
Überbewertung der Tatsachen 48, 126
Unterschiede messen 23
Unveröffentlichte Studien 167

V

Varianzanalyse 118
Verblindung 141
Veröffentlichungspflicht 172
Vogelrufe 59
Voreingenommene Berichterstattung 164
Vorwärtsauswahl 126
Voxel 85

W

Weggelassene Informationen 162
Wicherts, Jelte 156

Z

Zielvariable 118
Zirkuläre Analyse 100
Zulassungsverfahren 110
Zweiteilung 114

© des Titels »Statistics Done Wrong« (ISBN 9783958452527) 2016 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/252