Logbuch Mathematik

Thilo Kuessner

Die Kritik am leichtfertigen Umgang mit p-Werten ist alt, notwendig und kann nicht oft genug wiederholt werden.

Joseph Kuhn

Unterschriften gegen die statistische Signifikanz

Jemand behauptet, die Ergebnisse von Münzwürfen durch Gedankenkraft so beeinflussen zu können, dass häufiger Zahl als Kopf kommt. Zum Test soll man 100-mal eine Münze werfen. Wieviele Male muss Zahl kommen, damit die Behauptung plausibel ist?

52-mal Zahl wäre kein wirklicher Beweis der telekinetischen Fähigkeiten, denn dafür beträgt die Wahrscheinlichkeit immerhin 38 Prozent. 70-mal Zahl wären sehr viel überzeugender, denn dafür beträgt die Wahrscheinlichkeit nur 0,004 Prozent.

Die Wahrscheinlichkeit, dass rein zufällig (mindestens) die beobachtete Anzahl von Zahlwürfen kommt, bezeichnet man als p -Wert. Und den p -Wert, ab dem man akzeptiert, dass jemand den Münzwurf beeinflussen kann, nennt man Signifikanzniveau. Dessen Wahl ist natürlich subjektiv. Klinische Studien gehen meist von 5 Prozent als Signifikanzniveau aus, weil eine normalverteilte Zufallsgröße mit einer Wahrscheinlichkeit von weniger als 5 Prozent einen sich vom Erwartungswert um mehr als die 1,96-fache Standardabweichung unterscheidenden Wert annimmt. In der Physik sind es Größenordnungen weniger, so wurde für den Nachweis des Higgs-Bosons ein Kriterium von 5 Standardabweichungen, entsprechend einem p -Wert von 1 in 3,5 Millionen, angesetzt.

Ein Mediziner würde also an die Beeinflussung des Münzwurfs glauben, wenn 59 oder mehr Zahlwürfe produziert wurden: der p -Wert des Resultats 59 liegt bei etwa 4 Prozent, also unter dem Signifikanzniveau. Bei 58-mal Zahl ist der p -Wert etwa 7 Prozent, also über der Signifikanzschranke. Damit ist bis zu 58-mal Zahl das Versuchsergebnis nicht statistisch signifikant. (Dieses Beispiel und die Berechnungen habe ich von Ulrich Berger.)

Im März veröffentlichte die Fachzeitschrift *Nature* einen von mehr als 800 Wissenschaftlern unterschriebenen Brandbrief gegen die Verwendung des Begriffs „Statistische Signifikanz“: *Scientists rise up against statistical significance* (www.nature.com/articles/d41586-019-00857-9).

Darin geht es eigentlich nicht gegen statistische Signifikanz, sondern gegen die Verwendung des Begriffes „statistisch nicht signifikant“, sobald irgendwelche Schwellen des p -Wertes überschritten werden.

Let's be clear about what must stop: we should never conclude there is 'no difference' or 'no association' just because a P value is larger than a

threshold such as 0.05 or, equivalently, because a confidence interval includes zero. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.

Die Autoren veranschaulichen das Problem mit dem Beispiel eines Forschungsprojekts zu den Nebenwirkungen entzündungshemmender Medikamente. Weil ihre Ergebnisse statistisch nicht signifikant waren, kam eine Gruppe von Forschern zu dem Schluss, dass die Medikamentenexposition nicht mit neu auftretendem Vorhofflimmern (der häufigsten Störung des Herzrhythmus) in Verbindung gebracht werden könne und dass die Ergebnisse ihrer Studie im Gegensatz zu den Ergebnissen einer früheren Studie ständen, die ein statistisch signifikantes Ergebnis erhalten hatte. Tatsächlich war das Risiko für Vorhofflimmern bei dem Medikament ausgesetzten Patienten um 20 Prozent höher gewesen. Die Forscher hatten ein Konfidenzniveau von 95 Prozent, worin alles von einem um 3 Prozent bis zu einem um 48 Prozent höheren Risiko enthalten war. Damit kamen sie dann auf einen p -Wert, der über der Schranke 0.05 lag. Die frühere Studie hatte das selbe um 20 Prozent erhöhte Risiko gefunden, sie hatte aber ein Konfidenzintervall von 9 Prozent bis 33 Prozent angesetzt und hatte einfach nur wegen dieser präziseren Annahme einen sehr viel kleineren p -Wert erhalten. Die neue Studie zeigte also nichts anderes als die ältere, sie kam nur wegen eines anderen Ansatzes für das Konfidenzintervall zu einer entgegengesetzten Beurteilung der statistischen Signifikanz. Das ist offensichtlich absurd.

Können Geleebohnen Akne verursachen?

Mit der Bedeutung von Signifikanz befasste sich im April 2011 der Comicautor Randall Munroe auf xkcd.com/882/:

“Jelly beans cause acne! Scientists! Investigate!”

“We found no link between jelly beans and acne ($p > 0.05$). That settles that.”

“I hear it's only a certain color that causes it. Scientists!”

Nach diesen Eingangssequenzen wird dann in 20 einzelnen Zeichnungen beschrieben, wie Testergebnisse zu

Geleebohnen in den verschiedenen Farben jeweils einzeln ausfallen: "We found no link between purple jelly beans and acne ($p > 0.05$). "We found no link between brown jelly beans and acne ($p > 0.05$). " usw. Eine der zwanzig Zeichnungen (die vierzehnte) unterscheidet sich (im Text, nicht in der Zeichnung) von den anderen: "We found a link between green jelly beans and acne ($p > 0.05$). " Die einundzwanzigste Zeichnung ist dann eine Zeitungsschlagzeile: "News! GREEN JELLY BEANS LINKED TO ACNE! 95 % coincidence ... only 5 % chance of coincidence ... "

Also, auch wenn empirische Untersuchungen mit 95prozentiger Sicherheit zum Ergebnis „Nein!“ kommen, sollte, wenn man dieselben Untersuchungen mit Geleebohnen in 20 verschiedenen Farben durchführt, den Gesetzen der Mathematik zufolge mit Wahrscheinlichkeit $20 \times 0,05 = 1$ bei einer der Farben das Ergebnis lauten: Ja!

In Wirklichkeit ist es nicht ganz so einfach, weil man eigentlich $1 - (0.95)^{20}$ rechnen muss. Aber auch das ist mit 0.64 immer noch recht groß. Und mit einer 26prozentigen Wahrscheinlichkeit gibt es zwei oder mehr Signifikanzen, womit der Erwartungswert der Treffer bei 20 Versuchen in die Nähe von 1 rückt.

Inkorrekte Bayes-Statistik

Manche meinen, man sollte besser Bayessche Statistik benutzen und bräuchte dann keine Signifikanzniveaus mehr. Im Oktober letzten Jahres hatte Randall Munroe auf xkcd.com/2059 einen Comic mit folgender Formel:

$$P(H|X) = P(H) \left(1 + P(C) \left(\frac{P(X|H)}{P(X)} - 1 \right) \right)$$

$P(C)$ ist hier die Wahrscheinlichkeit, dass man Bayessche Statistik korrekt nutzt: falls $P(C) = 1$ erhält man den korrekten Satz von Bayes, falls $P(C) = 0$ erhält man die falsche Formel $P(H|X) = P(H)$.

Bissfeste Wahrscheinlichkeitstheorie-Bücher

Die Satireseite *Der Postillon* berichtete im März 2015 über eine Studie des Didaktik-Professors Werner Zöll, derzufolge die Mehrheit der deutschen Grundschüler weder die Unterschiede zwischen Cauchy-Integral, Riemann-Integral, Stieltjes-Integral und Lebesgue-Integral kennen noch in der Lage sei, die Funktionsweise einer Taylorreihe zur Darstellung einer glatten Funktion zu erklären oder gar in einer Gleichung mathematisch herzuleiten. Erst- und Zweitklässler schnitten besonders schlecht ab. Offenbar verstünden viele der Kinder noch nicht einmal die Aufgabenstellung, weil in den Funktionen zum Teil Buchstaben vorkommen, die sie „noch gar nicht dranhatten“. (www.der-postillon.com/2015/03/peinlich-99-prozent-aller-grundschuler.html)

Wie seinerzeit der PISA-Schock scheint auch diese Studie Anlass für rasche Gegenmaßnahmen gewesen zu sein. Der US-amerikanische Kinderbuchverlag Sourcebooks Jabberwocky – und das ist jetzt keine Satire – hat mit der Herausgabe einer „Baby-Universität“ begonnen, von der in diesem Jahr im Loewe-Verlag die ersten Bände in deutscher Übersetzung erscheinen. Neben angewandten Themen wie Raketenwissenschaft sind auch Quantenphysik und Relativitätstheorie in bissfesten Pappbänden im Angebot. Die Mathematik war bisher zwar nicht vertreten, doch am 2. Juli ist *Bayesian Probability for Babies* erschienen:

Bayesian Probability for Babies is a colorfully simple introduction to the basic principles of probability.

If you took a bite out of a cookie and that bite has no candy in it, what is the probability that bite came from a candy cookie or a cookie with no candy? You and baby will find out the probability and discover it through different types of distribution.

Yet another book of simple explanations of complex ideas for your future genius!

Bonsummen und ihre Endziffern

Eine Aufgabe zur Wahrscheinlichkeitstheorie gab es im letzten Heft der *Mitteilungen* auf der Kinderseite: „Wie wahrscheinlich ist es, beim Einkaufen im Supermarkt einen glatten Eurobetrag zu bezahlen? Hängt es auch von der Anzahl der gekauften Artikel ab?“

Bei einer Gleichverteilung der Nachkommastellen wäre die Antwort natürlich 1 Prozent. In den Supermärkten, die ich frequenziere, enden aber die allermeisten Preise auf eine 9 als letzte Ziffer. Einige Ausnahmen enden auf eine 5 und bei Zeitschriften endet der Preis auf eine 0. Obst und Gemüse muss man abwiegen, da dürfte die letzte Ziffer wirklich gleichverteilt sein.

Tatsächlich gibt es zu dieser Frage eine 2015 veröffentlichte Münzgeldstudie der Deutschen Bank mit der in der Tabelle wiedergegebenen Auflistung der häufigsten Preise.

Zweck der Studie war die Folgen einer Abschaffung der 1- und 2-Cent-Stücke zu prognostizieren. Deshalb wurden nur die Endziffern der Bonsummen – also nicht der Preise einzelner Artikel, sondern der an der Kasse tatsächlich gezahlten Beträge – erfasst. Diese verteilen sich wie folgt.

Verteilung der Endziffern von Bonsummen (Quelle: EHI Retail Institute. Aus: Frank Horst, *Münzgeldstudie. Folgenabschätzung einer Rundungsregel im Einzelhandel*, S. 17. Deutsche Bundesbank. Frankfurt 2015)

Man kann sicher davon ausgehen, dass die vorletzte Ziffer deutlich gleichverteilter ist als die letzte. Die Wahrscheinlichkeit eines glatten Eurobetrags dürfte also an der Kasse nicht wesentlich unter 1,27 Prozent liegen.

Verteilung der 20 häufigsten Preise (Quelle: EHI Retail Institute, zitiert nach Frank Horst, *Münzgeldstudie. Folgenabschätzung einer Rundungsregel im Einzelhandel*, S. 28. Deutsche Bundesbank. Frankfurt 2015)

Rang	Preis (€)	Anzahl	Anteil-Artikel (%)	Anteil-Artikel-kum (%)	Wert (€)	Rangwert	Anteil-Wert (%)	Anteil-Wert-kum (%)
1.	0,99	39.768	6,3	6,3	39.370,32	2	3,0	3,0
2.	1,99	32.409	5,1	11,4	64.493,91	1	4,9	7,9
3.	0,49	19.349	3,1	14,5	9.481,01	29	0,7	8,6
4.	0,79	19.216	3,0	17,5	15.180,64	17	1,2	9,8
5.	1,49	18.927	3,0	20,5	28.201,23	4	2,2	12,0
6.	1,29	18.284	2,9	23,4	23.586,36	8	1,8	13,8
7.	0,89	18.249	2,9	26,3	16.241,61	14	1,2	15,0
8.	0,59	16.136	2,6	28,9	9.520,24	28	0,7	15,7
9.	1,79	15.102	2,4	31,3	27.032,58	6	2,1	17,8
10.	0,69	13.549	2,1	33,4	9.348,81	31	0,7	18,5
11.	1,19	13.475	2,1	35,5	16.035,25	15	1,2	19,7
12.	2,99	12.168	1,9	37,5	36.382,32	3	2,8	22,5
13.	1,59	11.343	1,8	39,3	18.035,37	11	1,4	23,9
14.	1,39	10.921	1,7	41,0	15.180,19	18	1,2	25,0
15.	1,69	10.355	1,6	42,6	17.499,95	12	1,3	26,4
16.	0,45	9.849	1,6	44,2	4.432,05	56	0,3	26,7
17.	1,89	9.217	1,5	45,6	17.420,13	13	1,3	28,0
18.	0,85	9.160	1,5	47,1	7.786,00	34	0,6	28,6
19.	0,39	9.141	1,4	48,5	3.564,99	70	0,3	28,9
20.	0,29	8.968	1,4	50,0	2.600,72	91	0,2	29,1