Leserbrief zum Beitrag „Hypothesentests und bedingte Wahrscheinlichkeit“ von Renate Motzer (2010)

Raphael Diepgen, Bochum

1 Einleitung

Insbesondere Gerd Gigerenzer (1993) vom Berliner Max-Planck-Institut für Bildungsforschung vertritt die plausible These, die Etablierung der heute üblichen Inferenzstatistik in Wissenschaften wie etwa der Psychologie sei vor allem der Tatsache zu verdanken, dass die Statistikanwender die unüberbrückbaren Unterschiede der verschiedenen inferenzstatistischen „Schulen“ – insbesondere verbunden mit den Namen Bayes, Fisher sowie Neyman und Pearson – überhaupt nicht verstanden haben und sich stattdessen an ein in sich widersprüchliches Gemenge daraus halten, von Gigerenzer abfällig Hybridstatistik genannt. Unklare Begrifflichkeiten sind wesentliches Element dieser irrationalen Hybridstatistik.

Es sei mit Bezug zu dem im Titel genannten SiS-Beitrag von Motzer (2010a) exemplarisch demonstriert, wie die entsprechenden Begriffverwirrungen auch das Denken von Mathematiklehrern und -didaktikern prägen. Exemplarisch (und damit vielleicht etwas unfair zu Lasten von Renate Motzer) – denn die demonstrierten Begriffsverwirrungen könnte man auch bei vielen (den meisten?) anderen Mathematiklehrern, Schulbuchautoren und selbst Stochastikdidaktikern zeigen. Die detaillierte Bezugnahme auf den Beitrag von Motzer (2010a) erfolgt jeweils abkürzend durch ein Tripel (Seite, Spalte, Zeile).

2 Bayes versus Fisher versus Neyman und Pearson

Bei *Bayes*[1] geht es darum, die Wahrscheinlichkeiten konkurrierender und erschöpfender Hypothesen im Lichte von Daten zu modifizieren. Es geht also darum, auf der Basis der apriori-Wahrscheinlichkeiten der Hypothesen und der bedingten bzw. hypothetischen Wahrscheinlichkeiten der beobachteten Daten unter den Hypothesen – den sogenannten Likelihoods – aposteriori-Wahrscheinlichkeiten für die Hypothesen zu gewinnen. Ergebnis einer Bayesschen Analyse sind also die sich – gemäß der berühmten Formel von Bayes aus den Daten – ergebenden Wahrscheinlichkeiten der konkurrierenden Hypothesen; irgendeine „Entscheidung“ zwischen diesen Hypothesen ist nicht (mehr) Bestandteil dieses Ansatzes. Insofern ist die Rede von einem „Bayes-Test“ bei Motzer (30, 2, 38) abwegig, und „Fehler“ kann es im Bayesschen Ansatz nicht geben, weil dort gar nicht gehandelt, nicht entschieden, nichts für wahr gehalten oder erkannt wird. Der Bayessche Ansatz setzt voraus, dass das Konzept der Wahrscheinlichkeit einer Hypothese sinnvoll ist – was von den „objektivistischen“ Statistikern wie Fisher und Neyman und Pearson für die typischen statistischen Fragestellungen bestritten wird; insbesondere haben für Fisher und Neyman und Pearson Hypothesen typischer Weise nicht den Charakter von – mehr oder minder wahrscheinlichen – Ereignissen. Insofern ist auch die Benennung von H_1 als „Gegenereignis“ zu H_0 im nichtbayesschen Kontext bei Motzer (31, 2, 11) falsch.

Bei *Fisher* geht es darum, die Plausibilität einer isolierten (Null-)Hypothese auf der Basis der Überschreitungswahrscheinlichkeit p der in einem Experiment erhobenen Daten zu bewerten: Ist diese zu klein – beispielsweise kleiner als 5 % –, wird dadurch die Hypothese unplausibel. Bei Fisher gibt es keinen expliziten Vergleich dieser Hypothese mit konkurrierenden Hypothesen, und Fisher sind die Konzepte von „Fehlern“ – schon gar der zweiten Art – fremd. Das Signifikanzniveau ist bei Fisher *nicht* als Limit für die Wahrscheinlichkeit konzipiert, sich gegen die (Null-)Hypothese zu „entscheiden“, obwohl sie zutrifft. Bei Fisher „entscheidet“ man sich nämlich gar nicht, sondern man „erkennt“ etwas, genauer: es „wird“ einem eine Hypothese unplausibel oder eben nicht.

Erst bei *Neyman* und *Pearson* ist das Hypothesentesten explizit als ein – vor der Datenerhebung – festgelegtes Verfahren zur Entscheidung zwischen „two courses of action“ konzipiert, nämlich sich entweder so zu verhalten, als gelte H_1, oder so, als gelte H_0. Ein Test limitiert dabei die Wahrscheinlichkeit, das erste zu tun, obwohl H_0 gilt, und „minimiert“ dann die Wahrscheinlichkeit(en), das zweite zu tun, obwohl H_1 gilt. Hier geht es also um „induktives Verhalten“, während es bei Fisher um „induktives Schließen“ geht.

Gleich ob Bayes, Fisher oder Neyman und Pearson: Immer geht es nur um Wahrscheinlichkeiten, nicht um Sicherheiten – und damit scheidet eine Leitfrage wie „Ist sie“, die neue Therapiemethode, „*wirklich*

besser als die alte Methode?“, mit der Motzer in die Diskussion eines statistischen Tests einführt (32, 1, 6), von Vornherein aus: Und ganz typisch bearbeitet sie diese unbeantwortbare, allenfalls noch im „epistemiologischen“ Konzept von Fisher naheliegende Frage dann mit den damit inkompatiblen Konzepten aus der Tradition von Neyman und Pearson, also mit einem Fehler zweier Art riskierenden Entscheidungsverfahren, nicht etwa mit einem „Erkenntnisverfahren“.

3 Zur Analogie von statistischem Test und diagnostischem Test

Motzer geht es vor allem um die – didaktisch angeblich nutzbaren – Analogien zwischen dem Test statistischer, d. h. also auf Vielheiten bezogener Hypothesen einerseits, dem diagnostischen Test auf ein einzelnes Individuum bezogener Hypothesen andererseits, etwa in der Medizin, wo ein solcher Test die Hypothese H_1: „Der Patient hat die Krankheit“ gegen die Hypothese H_0: „Der Patient hat die Krankheit nicht“ testen mag und dabei ersichtlich Fehler – also Fehl*diagnosen* – zweier Art riskiert: Die falsch-positive Diagnose einer tatsächlich nicht vorhandenen Krankheit (Fehler 1. Art), und die falsch-negative Diagnose einer tatsächlich nicht vorhandenen Krankheitsfreiheit (Fehler 2. Art). Insofern man die individuenbezogenen Hypothesen H_1 und H_0 in diesem diagnostischen Kontext als Ereignisse mit positiver Wahrscheinlichkeit auffassen kann – etwa weil man den Patienten als Zufallsziehung aus einer Population betrachtet –, lässt sich hier sinnvoll von der bedingten Wahrscheinlichkeit der jeweiligen Fehldiagnose reden.

Ersichtlich könnte es hier allenfalls eine Analogie zum Konzept von Neyman und Pearson geben. Darauf will Motzer wohl hinaus, auch wenn sie laufend völlig verwirrend davon redet, „dass der Fehler 1. Art und der Fehler 2. Art beim Signifikanztest nur bedingte Wahrscheinlichkeiten sind“! (29, 2, 16, gleichlautend auch 30, 1, 14) Tatsächlich *sind* Fehler 1. und 2. Art keine bedingten Wahrscheinlichkeiten, sondern sie *haben* allenfalls solche. (Auch wenn die Verwechslung von „Haben oder Sein“ gesellschaftliches Zentralproblem sein mag – zumindest im Mathematikunterricht könnte man sie vermeiden.) Aber: Da für Neyman und Pearson ebenso wenig wie für Fisher Hypothesen Ereignisse sind – und schon gar nicht solche mit positiver Wahrscheinlichkeit –, widerspricht diese Begrifflichkeit der üblichen Begrifflichkeit, wie sie im traditionellen Stochastikunterricht eingeführt wird: Eine „bedingte“ Wahrscheinlichkeit von A unter der Bedingung B setzt dort nämlich voraus, dass die Bedingung B ein Ereignis ist – und dies auch noch mit positiver Wahrscheinlichkeit. Kurzum: Motzer übersieht, dass ihre Rede von den „bedingten Wahrscheinlichkeiten“ der Fehler 1. und 2. Art im Kontext statistischer Tests – anders als im Kontext diagnostischer Tests – dem formalen Begriff der bedingten Wahrscheinlichkeit widerspricht, den die Schüler vorher erworben haben dürften. Wenn also die Analogie zwischen diagnostischem und statistischem Test nicht mehr Verwirrung erzeugen denn vermeiden soll, müsste man vorher den Begriff der bedingten Wahrscheinlichkeit wohl etwas anders einführen (etwa als „Wahrscheinlichkeit für A, wenn man schon weiß (oder unterstellt), dass B der Fall ist“) oder aber hier einen anderen Begriff verwenden, etwa den der „hypothetischen Wahrscheinlichkeit“.

Und dann: Die Analogie zwischen diagnostischem und statistischem Test hinkt – und verwirrt – offensichtlich, weil beim statistischen Test (auch in dem von Motzer diskutierten Beispiel der Wirksamkeit einer neuen Therapie) die Alternativhypothese typischerweise keine „punktförmige“ Hypothese etwa der Form $\pi = 0{,}5$ ist, sondern eine „zusammengesetzte“ Hypothese der Form $\pi > 0{,}5$ – und es deswegen gar nicht wie beim diagnostischen Test *die* – bedingte oder hypothetische – Wahrscheinlichkeit eines Fehlers 2. Art gibt, sondern allenfalls eine Funktion, die diese Wahrscheinlichkeit*en* in Abhängigkeit von dem Parameter π beschreibt. Dazu unten unter 5. mehr.

4 Fehlerwahrscheinlichkeit als Wahrscheinlichkeit von Daten?

„α und β … geben Auskunft über die Wahrscheinlichkeiten von Daten unter den Bedingungen H_0 und H_1.“ (31, 1, 30, ähnlich auch 32, 2, 11) Auch hier (verw)irrt Motzer: α und β sind eben keine Likelihoods, also keine – bedingten oder hypothetischen – Wahrscheinlichkeiten der beobachteten Daten, sondern – bedingte oder hypothetische – Wahrscheinlichkeiten dafür, dass eine aus den beobachteten Daten aggregierte (suffiziente) Statistik, gemeinhin Prüfgröße genannt, in ein Intervall fällt, gemeinhin Annahme- bzw. Ablehnungsbereich genannt, deren Ausdehnung sich allerdings nach dem Fundamentallemma von Neyman und Pearson aufgrund eines Likelihood-Verhältnisses ergibt. Oder in Fisherscher Begrifflichkeit: p ist nicht die Wahrscheinlichkeit der Daten (genauer: des sich aus diesen Daten ergebenden Wertes der Prüfgröße) selbst, sondern die damit verbundene *Überschreitungswahrscheinlichkeit*, also die Wahrscheinlichkeit der beobachteten *und aller noch extremeren denkbaren, tatsächlich aber nicht beobachteten* Daten bzw. Prüfgrößenwerte.

5 Nachträgliche Bayes-Analyse eines Hypothesentest?

Diese Verwirrung Motzers wirkt sich dann auch aus auf ihre Bayessche Analyse eines durchgeführten Neyman-Pearson-Tests (31, 1, 43), die ohnehin nur im Falle eines in der typischen Praxis so gut wie nie vorkommenden Alternativtestes funktioniert, also bei „punktförmiger" Alternativhypothese. Dort wird – bei vorgegebenen aprori-Wahrscheinlichkeiten für die (trotz ihrer Punktförmigkeit erschöpfenden!) konkurrierenden Hypothesen H_0 und H_1 – die Information verrechnet, dass ein durchgeführter Neyman-Pearson-Test ein signifikantes Ergebnis erbracht hat, dass also die empirisch ermittelte Prüfgröße in den Ablehnungsbereich gefallen ist. Motzer scheint – in logischer Konsequenz ihrer Verwechslung von Fehlerwahrscheinlichkeit mit Datenwahrscheinlichkeit –zu übersehen, dass diese Bayessche Analyse schon deshalb wenig sinnvoll ist, weil sie unnötig auf Information verzichtet, nämlich darauf, konkret welchen Wert die Prüfgröße angenommen hat. Verwertete man diese detaillierte Information und nicht nur die unpräzise Information, dass dieser Wert „irgendwo" im Ablehnungsbereich lag, resultierten gegebenenfalls ganz andere, insbesondere viel aussagekräftigere aposteriori-Wahrscheinlichkeiten.

Viel grundsätzlicher ist aber Folgendes: Wenn man davon ausgeht, dass die Hypothesen eines im Sinne von Fisher oder Neyman und Pearson bereits durchgeführten Tests Wahrscheinlichkeiten haben, dann hat man den konzeptionellen Rahmen von Fisher sowie Neymann und Pearson verlassen, und es gibt dann überhaupt keinen rationalen Grund mehr, sich im Nachhinein aus Bayesscher Perspektive mit dem Ergebnis eines Nichtbayesschen Tests zu beschäftigen, der grundsätzlich nur auf die Likelihoods rekurriert und notwendiger Weise die apriori-Wahrscheinlichkeiten ignoriert. (Dies gilt auch, wenn man die apriori-Wahrscheinlichkeit wie Motzer (31, 1, 38) in eine fiktive Annahme über den „Anteil der Studien" kleidet, bei denen H_1 gilt – eine völlig unrealistische „frequentistische" Verkleidung eines tatsächlich „subjektiven" Wahrscheinlichkeitsbegriffs, die schon daran leidet, dass man dann im Ernst eine konkrete Studie als Zufallsziehung aus einer Population von Studien interpretieren müsste.) Pointiert formuliert: Wer einen Test im Sinne von Fisher oder Neyman und Pearson durchführt, tut dies rationaler Weise wohl *nur* deshalb, weil er nicht davon ausgeht, dass seine Hypothesen Wahrscheinlichkeiten hätten. Würde er diese Prämisse nicht machen, gäbe es wohl keinen Grund mehr, einen Test im Sinne von Fisher oder Neyman und Pearson zu machen, also die vorhandenen Hypothesenwahrscheinlichkeiten zu ignorieren. Es ist kaum eine Situation vorstellbar, wo eine solche Ignoranz gegenüber Hypothesenwahrscheinlichkeiten (in der Fachsprache: „Vorinformation") – sofern vorhanden – irgendeinen Sinn machen könnte. Genau darauf läuft aber die Argumentation von Motzer hinaus, indem sie zunächst einen nur für den Fall fehlender Hypothesenwahrscheinlichkeiten gedachten Test im Sinne von Neyman und Pearson durchführt und dann auf der Basis des Ergebnisses dieses Tests (Entscheidung gegen H_0) im Nachhinein fragt, was *daraus* – also nicht etwa aus den erhobenen Daten selbst und ihren Likelihoods – im Bayesschen Kontext folgt, wenn man nun plötzlich doch Hypothesenwahrscheinlichkeiten unterstellt. Ich kann mir kaum vorstellen, dass so etwas nicht zur weiteren Verwirrung von Schülern beiträgt.

6 Fazit

Es gibt noch weitere, eher lustige Verwirrungen. So behauptet Motzer (31, 2, 18): „Liegt ein Testergebnis vor, kann nur noch einer der beiden Fehler gemacht werden." Dumm nur: Wenn das Ergebnis eines Hypothesentests vorliegt, der Test also gemacht, die Entscheidung also gefallen ist, dann kann man gar keinen Fehler mehr machen, sondern man *hat* allenfalls einen solchen bereits gemacht.

Und nachdem Motzer in ihrem hier kritisierten Beitrag ausführlichst mit den nur aus der Tradition von Neyman und Pearson stammenden Konzepten von Fehlern der 1. und der 2. Art und ihren (bedingten) Wahrscheinlichkeiten beim Hypothesentesten argumentiert hat, operiert sie (2010b) im sofort daran anschließenden Beitrag „13 – eine Pechzahl beim Lotto?" ausschließlich mit den Konzepten von Fisher, insbesondere mit dem Argument, ein Untersuchungsergebnis sei „auffallend" (also wohl „signifikant"), „wenn es um mindestens 2 Standardabweichungen vom Erwartungswert abweicht", also letztlich mit dem Konzept der hinreichend geringen Überschreitungswahrscheinlichkeit. (Und da ist es dann auch möglich, dass ein Test nur „fast signifikant" wird, etwas, was im theoretischen Rahmen von Neyman und Pearson völlig bedeutungslos wäre.) Hier ist plötzlich keine Rede mehr von Fehlerwahrscheinlichkeiten, und zwar nicht nur nicht von denen der 2. Art (die mit den simplen 2 σ–Umgebungsüberlegungen Motzers auch gar nicht modellierbar wären), sondern auch nicht von denen der 1. Art. Dabei unterscheidet sich der Test der Nullhypothese, dass bei einer Lottoziehung die 13 mit der Wahrscheinlichkeit 6/49 auftaucht, wohl in nichts von dem in dem oben kritisierten Beitrag (2010a) diskutierten Test der Nullhypothese, dass die Heilungswahrscheinlichkeit einer neuen Therapie sich nicht von der der alten Therapie unterscheidet.[2] Wie soll

das ein Schüler verstehen, wenn eine „Unterrichtseinheit … am Ende einer Sequenz über Hypothesentesten“ (Motzer 2010b, S. 34) gerade die von Motzer (2010a) zuvor als für das Hypothesentesten wesentlich charakterisierten – und durch Analogie zum diagnostischen Test vermeintlich besonders verständlich gemachten – Konzepte von Fehlern zweier Art und ihren (bedingten) Wahrscheinlichkeiten überhaupt nicht anwendet, sondern es bei ganz anderen Konzepten wie „auffällig“ oder „signifikant“ (im Sinne von Überzufälligkeit oder geringer Überschreitungswahrscheinlichkeit oder geringem p-Wert) belässt? Oder könnte es sein, dass hier der Lehrer selbst gar nicht verstanden hat, dass es sich um gänzlich verschiedene Konzeptionen vom Hypothesentesten handelt – hie Fisher, dort Neyman und Pearson?

Anmerkungen

1 Die Namen Bayes, Fisher sowie Neyman und Pearson stehen hier nur als Chiffren. Ob und inwieweit diese Personen selbst die mit ihren „Schulen“ gemeinhin assoziierten Ideen, die hier ohnehin nur ganz grob und vereinfacht skizziert werden, geteilt haben, sei dahingestellt.

2 Zumindest wird ein Unterschied zwischen diesen beiden stochastischen Situationen von Motzer nicht benannt oder auf den Begriff gebracht. Natürlich könnte man dies versuchen, indem man für die Frage nach der Wirksamkeit einer neuen Therapie auf die offensichtlichen praktischen Konsequenzen verweist – und damit (eher) auf den Ansatz von Neyman und Pearson – und für die Frage nach der Wahrscheinlichkeit der 13 beim Lotto auf deren rein „theoretischen“ Charakter – und damit (eher) auf den Ansatz von Fisher. Was die von Motzer aufgrund eines replizierten signifikanten Ergebnisses als Pechzahl qualifizierte 13 angeht, erscheint allerdings zweifelhaft, dass dieses zweimal signifikante Testergebnis irgendjemand ernsthaft dazu bewegen könnte daran zu glauben, dass die 13 tatsächlich eine geringere Ziehungswahrscheinlichkeit hat als die anderen Zahlen, und kritische Schüler würden dies wohl auch sofort einwenden – und damit gleich einen klassischen Einwand der Bayesianer gegen den Signifikanztest entdecken: Die Blindheit gegenüber relevanter Vorinformation. Alle unsere Vorkenntnisse über die Physik der Lottoziehung schließen es nämlich so gut wie sicher aus, dass irgendeine Zahl mit höherer Wahrscheinlichkeit als andere Zahlen gezogen wird. Und noch eine andere Schwäche des Signifikanztestes würden kritische Schüler hier gleich mitentdecken, nämlich die logische Inkonsistenz der Hypothesen, die man „abgelehnt“ oder „nicht abgelehnt“ bzw. für die man sich „entschieden“ hat: Wenn man nur solche Nullhypothesen für unplausibel halten soll (Fisher) oder dem eigenen Verhalten nicht zugrunde legen soll (Neyman und Pearson), gegen die man ein (repliziertes) signifikantes Ergebnis gefunden hat, dann hat man beim Lotto ein Problem, weil sich ein solches – nach Motzer – nur für die 13 fand: Man müsste daran glauben bzw. in seinem Verhalten davon ausgehen, dass zwar die Ziehungswahrscheinlichkeit der 13 (pro Einzelziehung) ungleich 1/49 ist, die aller anderen Zahlen aber gleich 1/49. Dies wäre aber ein logischer Widerspruch, denn die Ziehungswahrscheinlichkeit der 13 kann nun mal nur dann von 1/49 abweichen, wenn auch andere Zahlen nicht mit der Wahrscheinlichkeit 1/49 gezogen werden.

Literatur

Gigerenzer, G. (1993): The Superego, the Ego, and the Id in statistical reasoning. In: G. Keren & C. Lewis (Hrsg.). *A handbook for data analysis in the behavioral sciences: Methodological issues.* Hillsdale, NJ: Erlbaum, S. 311–339. http://www.mpib-berlin.mpg.de/en/institut/dok/full/gg/ggstehfda/ggstehfda.html. (Zugriff: 30.11.2010)

Motzer, R. (2010a): Hypothesentests und bedingte Wahrscheinlichkeit. In: *Stochastik in der Schule* 30 (3), S. 29–32.

Motzer, R. (2010b): 13 – eine Pechzahl beim Lotto? In: *Stochastik in der Schule* 30 (3), S. 33–35.

Anschrift des Verfassers

Raphael Diepgen
Fakultät für Psychologie
Ruhr-Universität Bochum
Universitätsstr. 150
44780 Bochum
raphael.diepgen@rub.de