Die Analyse von experimentellen Daten

von Dennis V. Lindley, Warwick; Übersetzung: Manfred Borovcnik, Klagenfurt

Kurzfassung: Ein klassisches Experiment über Verkosten von Tee wird benützt, um zu zeigen, daß viele Standardmethoden der Analyse von resultierenden Daten wenig zufriedenstellend sind. Ein ähnliches Experiment mit Wein soll zeigen, wie man eine vernünftigere Methode entwickelt.

1. Das Experiment mit Tee

An einem Nachmittag im Jahr 1920 in der Rothamsted Versuchsstation machte der Statistiker R.A. Fisher eine Tasse Tee für Muriel Bristol. Sie protestierte, als er den Tee in die Tasse goß, bevor er die Milch dazu gab und behauptete, daß sie unterscheiden könnte, ob die Milch zuerst oder als zweites dazu gegeben worden sei, und sie würde ersteres vorziehen. Daraufhin entwarf Fisher das klassische Experiment, das so schön erörtert wird im Kapitel 2 seines Buches (Fisher, 1935). Die Prinzipien, die dort entwickelt werden, sind heute weitverbreitet in der Auswertung vieler Typen von Experimenten. Weil das ursprüngliche Experiment zu technischen Schwierigkeiten in der Analyse führt, werden wir hier eine veränderte Version betrachten, die diese Schwierigkeiten vermeidet, obwohl sie alle wesentlichen Züge des Originals beibehält.

In der modifizierten Form wird der Dame ein Paar von Tassen gereicht; es wird ihr versichert, daß in einer die Milch zuerst eingegossen wurde, in der anderen wurde die Milch erst nach dem Tee eingeschenkt. Sie soll nun herausbekommen, welche Tasse welcher Prozedur unterzogen worden ist. Die beiden möglichen Ergebnisse sind *richtig*, bezeichnet mit R, und *falsch*, bezeichnet mit F. Das Experiment wird insgesamt mit 6 Paaren von Tassen wiederholt. Nehmen wir an, das Ergebnis sei RRR RRF, d.h. nur das letzte Paar ist falsch. Fishers Analyse läuft wie folgt.

Zuerst sei angenommen, die Dame wäre völlig unfähig, zu tun, was sie behauptet, sodaß sie einfach ratet, welche Tasse welche ist. Die Hypothese über ihre Unfähigkeit, die Aufgabe zu erfüllen, wird *Nullhypothese* genannt. Nach Fishers Ansicht ist der Zweck dieses und vieler anderer Experimente, eine

Original 'The Analysis of Experimental Data: The Appreciation of Tea and Wine' in *Teach. Stat.* **15**(1993), 22-25.
Stochastik in der Schule **14** (1994), Nr. 1, S. 29-39.

Gelegenheit zu bieten, die Nullhypothese, daß sie nur rät, als zweifelhaft erscheinen zu lassen. Dabei bedeutet die Nullhypothese, daß sich für jedes Paar mit je einer Wahrscheinlichkeit von ½ R bzw. F ergibt, unabhängig von den anderen. Das beobachtete Ergebnis hat eine Wahrscheinlichkeit von $(½)^6 = 1/64$.

Fisher machte dann geltend, daß entweder

(a) die Nullhypothese wahr ist und ein Ereignis mit kleiner Wahrscheinlichkeit ist eingetreten, oder

(b) die Nullhypothese ist falsch und die Dame hat die Fähigkeit, zu unterscheiden, wie eingegossen wurde.

In diesem Fall beträgt die kleine Wahrscheinlichkeit 1/64. Weil Ereignisse mit kleiner Wahrscheinlichkeit nur selten eintreten, könnten wir (b) vorziehen als die einleuchtendere Erklärung für das Ergebnis, alle Paare bis auf eines richtig einzustufen. Das Resultat heißt *signifikant* mit Wahrscheinlichkeit 1/64 und die Wahrscheinlichkeit ist das *Signifikanzniveau*. Die Schlüsselidee ist, daß die Nullhypothese als zweifelhaft erscheint, wenn etwas eintritt, was unter den Bedingungen der Nullhypothese ungewöhnlich ist. Heutzutage ist es üblich, einen Wert von 1/20 oder 5% als Grenzwert zu benützen und zu sagen, das Ergebnis ist *signifikant auf dem 5%-Niveau*, wenn die kleine Wahrscheinlichkeit kleiner oder gleich diesem Wert ist, wie dies auch mit unserem Ergebnis ist.

Fisher erkannte sofort, daß dieses Argument versagt, weil jedes mögliche Ergebnis mit den 6 Paaren eine Wahrscheinlichkeit von $(½)^6 = 1/64$ hat, sodaß *jedes* Ergebnis signifikant bei 5% ist. Fisher umging diese Absurdität dadurch, daß er sagte, daß jeder Ausgang mit bloß 1 F und 5 Rs, ohne zu berücksichtigen, an welcher Stelle F auftritt, in gleicher Weise die Fähigkeit zu unterscheiden anzeigt und daher mitberücksichtigt werden sollte. Es gibt 6 solcher Ausgänge, sodaß die relevante Wahrscheinlichkeit für (a) von oben nun $6(½)^6 = 0{,}094$ beträgt; insgesamt bedeutet dies, daß das Ergebnis nun *nicht* signifikant ist bei 5%.

Fishers verbessertes Argument für eine allgemeine Situation ersetzt die Wahrscheinlichkeit für das Ergebnis unter der Nullhypothese durch die Wahrscheinlichkeit dieses und ähnlicher Ergebnisse; hier die Wahrscheinlichkeit von 1 Fehler in 6 Bewertungen. Fisher erkannte, daß auch das nicht funktionieren würde. Denn, was ist das wahrscheinlichste Ergebnis bei reinem Raten? Klarerweise die Hälfte der Paare richtig und die Hälfte falsch. Für 128

Paare von Tassen mit 64 R und 64 F beträgt die Wahrscheinlichkeit $\binom{128}{64}\cdot\left(\frac{1}{2}\right)^{128}$, was ungefähr 0,05 ist. Dies gilt für den wahrscheinlichsten Ausgang, alle anderen haben eine kleinere Wahrscheinlichkeit. Daher haben wir für die 128 Paare wieder die Schwierigkeit, daß jedes Ergebnis signifikant bei 5% ist. Um dies zu umgehen, behauptete Fisher treuherzig, daß, wenn 1 Fehler bei 6 Bewertungen signifikant ist, dann ist es sicherlich auch 0 Fehler oder 6 Rs. Mit anderen Worten, Fälle, die noch stärker auf die Fähigkeit zu unterscheiden hinweisen als im beobachteten Ergebnis, sollten auch berücksichtigt werden bei der Berechnung der Wahrscheinlichkeit, die mit den 5% zu vergleichen ist. Ergebnisse, die diese Fähigkeit gleich stark oder noch stärker andeuten als das beobachtete, werden ebenso *extrem* oder extremer genannt.

Das Fazit ist, daß Fishers einfache Alternative (a) oder (b) nun wie folgt zu verbessern ist: Entweder

(a) die Nullhypothese ist wahr und die Wahrscheinlichkeit von Ereignissen, die ebenso extrem oder extremer sind als das beobachtete Ergebnis, ist klein, oder

(b) die Nullhypothese ist falsch und die Dame hat die Fähigkeit zu unterscheiden, wie die Milch eingegossen worden ist.

Diese Form wird von den meisten Statistikern akzeptiert und die wissenschaftliche Literatur ist voll von 5%-Signifikanzen, wobei die 5% sich auf die Wahrscheinlichkeit aller Resultate beziehen, die *gleich oder noch extremer* als das beobachtete sind. Es sind die kursiven Worte, welche die akzeptierte Form von der ersten von Fisher unterscheiden. Beim Ergebnis RRR RRF mit der Wahrscheinlichkeit $(½)^6$ gibt es 5 andere, gleich extreme und 1 Ergebnis ohne Fehler, noch extremer, was insgesamt 7 Fälle und eine Gesamtwahrscheinlichkeit von $7(½)^6 = 0{,}109$ ergibt, was nicht signifikant bei 5% ist.

2. Eine kritische Beurteilung

Viele Jahre lang blieb das Argument unwidersprochen und wurde durch alternative, mathematischere Ansätze von Neyman, Pearson und Wald gestützt. Kürzlich jedoch kamen Zweifel auf, die ursprünglich von Jeffreys vorgebracht wurden, und das Argument wird immer stärker angegriffen. Schauen wir uns einmal an, wie die Kritik für das Ergebnis RRR RRF ansetzt. Fisher muß beachten, welche Ergebnisse gleich extrem oder extremer sind, und, um dies zu tun, zieht er andere Möglichkeiten mit 6 Paaren von Tassen heran.

Aber warum die Zahl der Paare mit 6 festlegen? Der Wert 6 mag rein durch Zufall zustande gekommen sein. Vielleicht wollte Dr. Muriel Bristol nach dem Tee zu einer Besprechung und mußte daher nach 6 Paaren weggehen? Eine andere Form des Experiments wurde von J.B.S. Haldane im Zusammenhang mit Katzen und nicht mit Tee verkosten vorgeschlagen; man sollte so lange testen, bis der erste Fehler passiert. Das Ergebnis von Dr. Bristol ist mit dieser Art von Experiment verträglich. Daher sei das Fishersche Argument auf das Experiment von Haldane angewendet. Die Wahrscheinlichkeit der Folge RRR RRF ist noch immer $(½)^6$. Extremer sind jene Folgen, in welchen der erste Fehler nach dem 6. Paar auftaucht. Das sind beim 7. mit Wahrscheinlichkeit $(½)^7$, beim 8. mit $(½)^8$ usw. Die Wahrscheinlichkeit für das beobachtete und noch extremere Ergebnisse ist daher:

$$\left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 + \ldots = \left(\frac{1}{2}\right)^6/(1-\frac{1}{2}) = \left(\frac{1}{2}\right)^5 = 0{,}031$$

Vorher hatten wir 0,109, nun jedoch haben wir Signifikanz bei 5%. Das ist überraschend.

Überdenken wir nocheinmal, wo wir stehen. Wenn das Experiment darin besteht, 6 Paare von Tassen zu prüfen und das Ergebnis lautet RRR RRF, so ist die relevante Wahrscheinlichkeit 0,109. Wenn das Experiment darin besteht, Paare von Tassen so lange zu testen, bis der erste Fehler auftaucht, so ist, beim selben Ergebnis, die relevante Wahrscheinlichkeit 0,031, das ist weniger als ein Drittel der vorhergehenden Werts. Dazu noch wird aus fehlender Signifikanz beim ersten Fall nun Signifikanz beim zweiten Fall. Ist das nicht absurd? Da sind 6 Paare von Tassen, ehrlich geprüft mit dem Ergebnis RRR RRF; was macht es da aus, was noch passieren hätte können (z.B. RRR FRR im einen, RRR RRR RF im anderen Fall), aber gar nicht passiert ist? Wie würde die Wahrscheinlichkeit lauten, wenn Dr. Bristol wegen der Besprechung abgebrochen hätte?

Bringen wir doch die Schwierigkeit beim Fisherschen Argument auf den Punkt. Sie liegt in der Entscheidung, welche Ergebnisse gleich extrem oder extremer als das beobachtete Ergebnis sind. (Wir haben gesehen, daß extremere Ergebnisse miteinzuschließen sind, weil es sonst Experimente gibt, bei denen jedes Ergebnis ungewöhnlich ist.) Im Fall einer festgelegten Zahl, sei es 6, von Paaren von Tassen sind die extremeren Ergebnisse andere als im Fall der Fortsetzung der Prüfung bis zum ersten Fehler. Diese beiden Experimente seien im folgenden das *fixe* bzw. das *sequentielle* Experiment genannt. Man könnte behaupten, daß die Beurteilung davon abhängen sollte, ob das

fixe oder das sequentielle Experiment verwendet worden ist. Aber, wenn Sie dies meinen, so betrachten Sie doch folgendes Experiment.

Eine faire Münze wird geworfen, fällt sie auf Kopf, so wird das fixe Experiment mit 6 Paaren durchgeführt, fällt sie auf Zahl, so wird das sequentielle angewendet. Das Ergebnis RRR RRF hat eine Wahrscheinlichkeit, die mit dem Mittelwert der beiden Experimente zusammenhängt, nämlich (0,109 + 0,031)/2 = 0,070. Wenn aber die Münze auf Zahl fällt, sollte man dann wirklich die 0,070 anführen, bloß weil die Münze auch auf Kopf hätte fallen können? Der Vorschlag scheint seltsam. Man hat Versuche unternommen, um festzulegen, was man unter extremer verstehen sollte, jedoch ohne Erfolg.

Daher müssen wir es aufgeben, extremere Ergebnisse zu verwenden. Das bringt uns auf den Punkt, wo wir nur die Wahrscheinlichkeit dessen, was passiert ist, benutzen können und wir haben schon gesehen, wie unbefriedigend dies ist, weil in einigen Experimenten alle Wahrscheinlichkeiten klein sind. Was sollten wir daher tun?

3. Eine alternative Analyse

Fishers Ansatz zieht nur Wahrscheinlichkeiten auf der Basis der Nullhypothese in Betracht. Er betrachtet keine Wahrscheinlichkeiten für den Fall, daß Dr. Muriel Bristol die Fähigkeit hat zu unterscheiden. Natürlich, hätte sie diese Fähigkeit in Vollendung, so würde R mit Wahrscheinlichkeit 1 auftreten und das sequentielle Experiment würde nie enden. Aber sogar der begeisterste Anhänger der These, daß die Milch zuerst in die Tasse gegossen werden muß, würde zugeben, daß er dies gelegentlich nicht merkt. Wir sahen, daß unter der Nullhypothese jedes R Wahrscheinlichkeit p hat, unabhängig von den anderen, und zwar mit p=½. Ein brauchbares Indiz für die Unterscheidungsfähigkeit würde einen Wert von p größer als ½ zulassen. Je höher der Wert von p, umso besser ist diese Fähigkeit der Dame ausgeprägt. Die Werte von p über ½ werden *alternative Hypothesen* genannt. Das Ergebnis RRR RRF hat Wahrscheinlichkeit $p^5(1-p)$; bei p=½ ergibt sich $(½)^6$ wie vorhin. Dieser Ausdruck ist die *Likelihood-Funktion* von p für das beobachtete Ergebnis, wobei p für die Wahrscheinlichkeit einer richtigen Klassifikation steht.

Im allgemeinen beschreibt die Likelihood-Funktion die Wahrscheinlichkeit des beobachteten Ergebnisses als eine Funktion von p. Moderne Arbeiten sagen aus, daß man diese Funktion braucht und nicht die Betrachtung von

extremeren Fällen. Die Wahrscheinlichkeit dessen, was tatsächlich passiert ist, wird unter verschiedenen Hypothesen betrachtet, anstatt daß man die Wahrscheinlichkeit von verschiedenen Ergebnissen einzig unter der Nullhypothese berechnet.

Man muß dann die Wahrscheinlichkeit unter der Nullhypothese mit jenen Wahrscheinlichkeiten für andere Werte von p vergleichen. Aber welche Werte für p sollte man nehmen? Um diese Frage zu beantworten, betrachten wir eine andere Dame.

4. Das Experiment mit Wein

Diese Dame nun ist eine Weinkennerin, bezeugt dadurch, daß sie sogar einen eigenen Titel dafür hat. Anstatt Tee zu verkosten, prüft sie Wein. Ihr werden 6 Paare von Gläsern präsentiert, nicht Tassen. Eines der Gläser enthält jeweils etwas französichen Bordeaux, das andere kalifornischen Cabernet Sauvignon. Mit anderen Worten, dieselbe Traubensorte, einmal aus Frankreich, das andere Mal aus Kalifornien. Sie muß erkennen, welcher Wein in welchem Glas ist. Das heißt, sie unterzieht sich demselben Experiment wie Dr. Bristol, aber mit Wein anstelle der verschiedenen Teezubereitung. Angenommen, sie erhält nun dasselbe Ergebnis RRR RRF und damit dieselbe Likelihood-Funktion $p^5(1-p)$, wobei p sich auf ihre Wahrscheinlichkeit bezieht, ein Paar von Gläsern korrekt zuzuordnen.

An dieser Stelle kann ich nur für mich selbst sprechen, aber ich hoffe, viele werden mir zustimmen. Es steht Ihnen jedoch frei, anderer Meinung und dennoch vernünftig zu sein. Ich glaube, daß spezielle Weinverkoster die kalifornische Imitation des französischen Originals unterscheiden können. Mathematisch drückt sich das durch $p>½$ aus. Dagegen denke ich, daß es zweifelhaft ist, ob es Damen gibt, welche die zwei Arten der Teeaufbereitung unterscheiden können. Es scheint ziemlich plausibel, $p=½$ zu nehmen, obwohl ich zugestehe, daß auch $p>½$ möglich ist. Was ich daher tun will, ist, etwas in die Analyse einzubringen, was meine Ansicht ausdrückt, daß Tee anders ist als Wein. Beachten Sie, daß die Likelihood für beides dieselbe ist, obwohl sich die Bedeutung von p unterscheidet.

Man kann dies durch Einführen von Wahrscheinlichkeitsverteilungen für p tun, die für Tee und für Wein entsprechend angepaßt sind. Lassen Sie mich meine Verteilungen zeigen, um die Idee zu verdeutlichen. Für Wein wählte ich den Ausdruck

$$48(1-p)(p-\tfrac{1}{2}), \quad \text{für } \tfrac{1}{2}<p<1$$

Man sieht die Gestalt in Abb. 1, sie trägt die Bezeichnung 'prior'. Diese Verteilung drückt die Tatsache aus, daß ich denke, die Dame kann den Wein unterscheiden, aber sie kann dabei auch Fehler machen. Der Wert 48 macht nur die gesamte Wahrscheinlichkeit zu 1. Für Tee nahm ich

$$0{,}8 \text{ für die Wahrscheinlichkeit, daß } p=\tfrac{1}{2} \text{ und } 1{,}6(1-p) \text{ für } p>\tfrac{1}{2} \qquad (1)$$

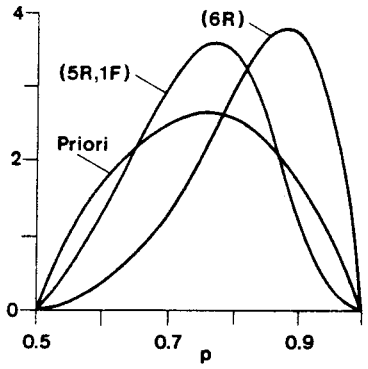

Abb. 1: Wein

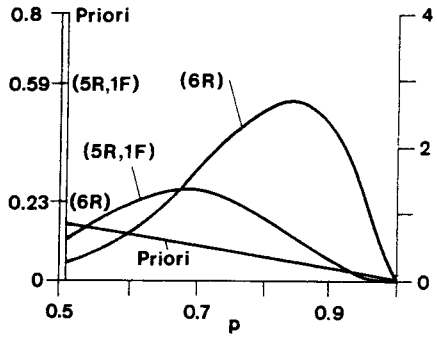

Abb. 2: Tee

Man sieht die Gestalt in Abb. 2, wieder mit 'prior' bezeichnet. Die Verteilung drückt meine persönliche Wahrscheinlichkeit von 0,8 aus, daß sie die Teezubereitung nicht unterscheiden kann. (Fisher mag einen solchen Wert gehabt haben, weil er seine Überraschung über Dr. Bristols Behauptung so ausdrückte: "Unsinn, natürlich macht es keinen Unterschied." Siehe Box, 1978.) Dies gibt eine Wahrscheinlichkeit von 0,2 dafür, daß sie es kann, wobei eine gute Unterscheidungsfähigkeit (p nahe bei 1) für weniger wahrscheinlich gehalten wird als eine mäßige (p nahe bei ½). Die Formeln drükken meine eigenen Ansichten aus, Sie mögen gerne Ihre eigenen einsetzen. Weitere Details findet man in Lindley (1984).

5. Die Formel von Bayes

Der nächste Schritt ist, diese persönlichen Ansichten zu verbinden mit der Beweiskraft der Daten, die durch die Likelihood-Funktion ausgedrückt wird. Der Kalkül mit Wahrscheinlichkeiten gibt uns Aufschluß, wie man das tut,

nämlich durch Multiplikation der ursprünglichen Wahrscheinlichkeit mit der Likelihood-Funktion. Für die Dame, die den Tee verkostet, erhalten wir

$$48(1-p)(p-\tfrac{1}{2})p^5(1-p) \quad \text{für} \quad \tfrac{1}{2}<p<1$$

Abgesehen vom Umstand, daß die gesamte Wahrscheinlichkeit nun nicht 1 beträgt, ist der letzte Ausdruck eine Wahrscheinlichkeitsverteilung. Einfache, aber umständliche Berechnungen lassen uns eine Konstante K finden, sodaß

$$K(1-p)^2p^5(p-\tfrac{1}{2}) \quad \text{für} \quad \tfrac{1}{2}<p<1 \qquad (2)$$

tatsächlich eine Wahrscheinlichkeitsverteilung ist, für welche das Integral über ½ bis 1 den Wert 1 hat. Die Wahrscheinlichkeitsverteilung (1) von vorher heißt *priori* (mit priori ist gemeint, *vor* den Daten). Die eine, die wir gerade erhalten haben, (2), wird *posteriori* genannt (*nach* den Daten). Die Formel besagt

$$\text{posteriori} = K \times \text{priori} \times \text{Likelihood},$$

wobei K eine Konstante ist, die das Integral der rechten Seite gerade zu 1 normiert. Sie heißt *Bayes-Formel* und die Methode heißt *Bayesianisch*. (Für Details kann man auch in Wickmann, 1990 lesen.) Die einzige Komplikation in der Berechnung ist die Bestimmung von K.

Abb. 1 zeigt für die Weinprobe (i) die priori Verteilung (1), (ii) die posteriori Verteilung (2) bei 6 Gläsern mit 1 Fehler, und (iii) die posteriori bei 0 Fehlern. Ursprünglich dachte ich, p=¾ wäre der wahrscheinlichste Wert, aber es gab eine große Unsicherheit darüber, was durch die große Streuung der priori um diesen Wert zum Ausdruck kommt. Bei 1 Fehler gibt es kaum eine Verschiebung des wahrscheinlichsten Wertes, aber ich bin etwas sicherer, daß p nahe bei ¾ liegt, was sich durch die kleinere Ausbreitung der posteriori ausdrückt. Um diese Breite besser zu verstehen, betrachten wir einmal die Fläche unter diesen Kurven zwischen sagen wir 0,6 und 0,9, das ist ± 0,15 von p = ¾. Die Fläche und damit die Wahrscheinlichkeit ist ein wenig größer für die posteriori als für die priori. Gibt es 0 Fehler bei der Zuordnung, so verändert sich die Situation und der wahrscheinlichste Wert steigt auf rund 0,87 an, wobei die Streuung wesentlich geringer wird. Z.B. beträgt die Wahrscheinlichkeit, daß p kleiner als 0,75 ist, ungefähr 0,2, wohingegen sie ursprünglich 0,5 betragen hat.

Die Situation beim Tee ist heikler, weil ich ursprünglich eine Wahrscheinlichkeit von 0,8 hatte, daß die Dame die Zubereitung nicht unterscheiden kann, d.h. p=½, was beim Wein nicht angenommen wurde. Die analogen Graphen

sieht man in Abb. 2. Der priori Wert von p=½ war 0,8, er sinkt auf 0,59 bei 1 Fehler bei 6 Paaren, und auf 0,23 bei 0 Fehlern. Genau diese Werte können mit den Signifikanzniveaus verglichen werden, das sind die Wahrscheinlichkeiten für gleich extreme oder extremere Ergebnisse unter der Nullhypothese. Die letzteren waren 0,109 bzw. 0,016. Beachten Sie in beiden Fällen, daß die Signifikanzen wesentlich kleiner sind als die posteriori Wahrscheinlichkeiten. Der Grund liegt teilweise in der hohen priori von 0,8. Aber die Feststellung trifft auch dann noch zu, wenn man denkt, daß es gleich wahrscheinlich ist, daß die Dame die Teezubereitung unterscheiden kann oder eben nicht kann. Z.B. bei 1 Fehler bei 6 Zuordnungen ergibt die posteriori 0,26, was zu vergleichen ist mit einem Signifikanzniveau von 0,109. Es ist typisch, daß die posteriori Wahrscheinlichkeit der Nullhypothese das Signifikanzniveau überschreitet, obwohl es keine logische Verbindung zwischen diesen beiden Werten gibt. Das Verhalten der Kurven für p>½ ist ähnlich dem beim Wein.

Die soeben präsentierte Analyse hängt stark von meiner Einschätzung über die Fähigkeiten der Damen ab. Ihre Einschätzung mag davon verschieden sein. Bei der dürftigen Beweiskraft von 12 Tassen oder Gläsern ist es nicht überraschend, daß unsere Einschätzungen sich unterscheiden, gerade so wie sich gegenwärtig Wissenschafter über den Glashauseffekt nicht einig sind, weil das Beweismaterial nicht ausreicht. Aber hätten wir Beweismaterial aus 1200 Tassen, vielleicht mit 100 Damen, so würden die unterschiedlichen anfänglichen Einschätzungen von der Beweiskraft der Daten überschwemmt werden und wir würden im wesentlichen übereinstimmen. Technisch gesprochen dominiert die Likelihood-Funktion die priori Verteilung bei einer großen Stichprobe. Das kommt in den Wissenschaften schon vor. Vor 20 Jahren etwa waren viele von uns sehr argwöhnisch gegen Behauptungen, daß Blei Intelligenz beeinflußt. Die Belege nun erdrücken die ursprünglichen Auffassungen. Alles, was Beweismittel tun, ist, die Einschätzungen verändern: sie erschaffen keine Einschätzungen.

6. Schlußfolgerungen

Es gibt vier Lehren, die man aus dieser Analyse ziehen kann:

(a) Weil das Signifikanzniveau typischerweise niedriger ist als die posteriori Wahrscheinlichkeit der Nullhypothese und weil eine kleiner Wert der Signifikanz wie 5% Zweifel auf die Nullhypothese wirft, folgt daraus, daß Nullhypothesen nach der Fisherschen Methode leichter beeinträchtigt werden als beim Bayesschen Ansatz. Wenn man bedenkt, daß eine typische Nullhypo-

these lautet, ein Medikament oder eine Behandlung ist nicht wirksam, wird man sehen, daß die Überfülle von Signifikanztests heute trügerisches Vertrauen in die Wirksamkeit von Medikamenten oder Behandlungen ermutigt. Wann immer Sie lesen, daß eine Wirkung entdeckt worden ist, bedenken Sie, daß sich dies wahrscheinlich auf Signifikanzen bezieht, welche allzu leicht einen Effekt andeuten, wo gar keiner besteht.

(b) Die Bayesianische Analyse stellt dem Wissenschafter alles bereit, was er benötigt. Er ist interessiert daran, ob die Nullhypothese wahr ist (wie beim Tee) oder welche Größenordnung der untersuchte Effekt hat (wie beim Wein). Er benötigt ein Maß des Vertrauens, Wahrscheinlichkeit bietet ein solches Maß. Für die Nullhypothese direkt; für die Größe des Effekts, in unserem Beispiel ausgedrückt durch p, durch eine Wahrscheinlichkeitsverteilung, wie sie durch die Kurven in den Abbildungen dargestellt ist. Dies steht in krassem Gegensatz zum Signifikanzniveau, welches eine Wahrscheinlichkeit bereitstellt für etwas, was nicht passiert ist auf der Basis einer Hypothese, die vielleicht nicht zutrifft.

(c) Die Baysianische Analyse unterscheidet zwischen Tee und Wein. Fishers Analyse benutzt nur Wahrscheinlichkeiten unter der Annahme des Ratens, und Raten bedeutet dasselbe für beides, wie das Wort 'Raten' es auch besagt. Die Baysianische Sicht berücksichtigt, daß jemandes Meinung über das Verkosten der zwei Flüssigkeiten verschieden sein kann oder daß die beiden Damen unterschiedlich ausgeprägte Fähigkeiten haben können.

(d) Dies ist bei weitem der wichtigste Punkt unter den vieren. Die Bayesianische Methode ist *vergleichend.* Sie vergleicht Wahrscheinlichkeiten des beobachteten Ergebnisses auf der Basis der Nullhypothese und der alternativen Hypothesen dazu. In dieser Hinsicht ist sie grundverschieden vom Fisherschen Ansatz, welcher absolut ist in dem Sinne, daß er eine einzige Hypothese, nämlich die Nullhypothese, betrachtet. Alle unsere Entscheidungen bei Unsicherheit sollten vergleichend sein, es gibt keine absoluten Entscheidungen hier. Ein schlagendes Beispiel dafür tritt bei Gericht auf. Wenn ein Beweisstück B vorliegt, das über Schuld S oder Unschuld U eines Angeklagten entscheiden soll, so ist es nicht genug, nur die Wahrscheinlichkeit von B unter S zu betrachten; man muß auch die Wahrscheinlichkeit von B unter U erwägen. Tatsächlich ist die relevante Größe der Quotient dieser beiden Wahrscheinlichkeiten. Allgemein, wenn ein Beweismaterial vorliegt, das irgendeine Behauptung stützen soll, so muß man auch die Glaubwürdigkeit des Beweismaterials betrachten für den Fall, daß die Behauptung falsch ist. Wann

immer eine Entscheidungsfindung überdacht wird, sind es nicht die Vor- und Nachteile jeder einzelnen Entscheidung, sondern einzig und allein der Vergleich dieser Qualitäten mit anderen Entscheidungen, was zählt.

7. Zusammenfassung

Die Hauptpunkte werden nun zusammengefaßt. Fisher argumentierte in der Form einer Dichotomie: entweder (a) ein Ereignis mit einer kleinen Wahrscheinlichkeit unter der Nullhypothese ist eingetreten, oder (b) die Nullhypothese ist falsch. Dies funktionierte nicht und die Wahrscheinlichkeit hatte sich auch auf Ereignisse zu erstrecken, die nicht eingetreten sind, die gleich extrem oder extremer sind. Dies wiederum funktionierte nicht, weil der Begriff 'gleich oder extremer' mehrdeutig ist. Der Ausweg aus dieser Schwierigkeit ist, die Wahrscheinlichkeiten dessen, was sich ereignet hat, unter der Nullhypothese und unter Alternativen dazu miteinander zu vergleichen. Weil es für gewöhnlich verschiedene Alternativen gibt, muß man sie gewichten. Dies wird getan durch persönliche Einschätzungen über die Situation vor dem Experiment. Diese priori Einschätzung in der Form von Wahrscheinlichkeiten wird durch die experimentellen Daten verändert; aus der Bayes-Formel werden die posteriori Einschätzungen berechnet. Man vergleicht die verschiedenen Erklärungen für das, was eingetreten ist, und man vergleicht seine posteriori Einschätzungen mit denen, die man anfänglich hatte. Die ganze Analyse ist vergleichend.

Literatur

Box, J.F. (1978): *R.A. Fisher, the Life of a Scientist,* New York: Wiley.

Fisher, R.A. (1935): *The Design of Experiments,* Edinburgh: Oliver and Boyd.

Lindley, D.V. (1984): A Bayesian Lady Tasting Tea, in David, H.A. und David, H.T. (Hrsg.), *Statistics: an Appraisal,* Ames: Iowa State University Press, 455-485 (mit Diskussion).

Wickmann, D. (1990): *Bayes-Statistik. Einsicht gewinnen und entscheiden bei Unsicherheit,* Mannheim: Bibliographisches Institut.

Prof. Dr. Dennis V. Lindley, University of Warwick, Warwick, England