Inferenzstatistik klassisch / bayesianisch — wann was?

DIETER WICKMANN, AACHEN

Zusammenfassung:

Zwei grundverschiedene Wege zur Lösung inferenzstatistischer Probleme stehen zur Verfügung: der sogenannte klassische, der im Signifikanztest konkretisiert ist, und der bayessche mit dem BAYES*schen Theorem als seinem „Herzstück". Die Schulbuchliteratur hat allein den Signifikanztest zum Gegenstand — was so seine Schwierigkeiten nach sich zieht. Neuere Einführungen in die Mathematik für das Lehramt betrachten auch den bayesschen Weg, allerdings mit dem Tenor, die Wahl des Weges sei letztlich von untergeordneter Bedeutung. Der Verfasser dieses Beitrags sieht die Dinge jedoch anders, was zu der Frage führt,* wann was *anzuwenden sei.*

Der Beitrag ist die ergänzte Fassung eines Vortrags, der auf der Tagung der Deutschen Arbeitsgemeinschaft Statistik (DAGStat) *im März 2007 in Bielefeld gehalten worden ist.*

1 Inferenzstatistik $\stackrel{?}{=}$ Beurteilende Statistik

Mit *Inferenzstatistik* oder *Schließende Statistik* ist ein Tätigkeitsfeld gemeint, das unglücklicherweise auch *Beurteilende Statistik* genannt wird. Wieso „unglücklicherweise"?

Zwei Hauptziele sind zu unterscheiden, die in folgenden beiden Fragen F1 und F2 zum Ausdruck kommen:

F1: Wie sicher kann man angesichts der vorliegenden Indizien (Daten) sein, dass die möglichen Hypothesen je der Fall sind?

Dabei sind die Hypothesen mögliche Parameterwerte, die die Daten als Stichprobenergebnisse beeinflussen. Die Parameter können auch in Intervallen zusammengefasst sein, in welchem Fall von zusammengesetzten Hypothesen die Rede ist. F1 fragt danach, wie im Rückschluss aus den Daten das Zutreffen einer jeden Hypothese zu *beurteilen* ist.

F2: Für welche der möglichen Handlungen sollte man sich angesichts der unsicheren Sachlage vernünftigerweise entscheiden?

Hinter F1 steht das Ziel einer Erkenntnis: Man möchte wissen, was der Fall ist. Wegen der Unerreichbarkeit des Ziels stellt man die bescheidenere Frage nach einem Sicherheitsgrad, mit dem eine Hypothese der Fall ist. Wir sagen, F1 sei *erkenntnisorientiert.* In F2 ist eine Handlungsentscheidung gesucht. Wir sagen, F2 sei *handlungsorientiert.*

Zwei unterschiedliche Denkansätze führen zu den beiden hier interessierenden Verfahren, nämlich dem klassischen Signifikanztest und der BAYES-Analyse, die die Fragen F1 und F2 auf grundverschiedene Art angehen.

Die BAYES-Analyse besteht aus zwei Etappen, deren erste das BAYES*sche Theorem* und deren zweite das BAYES*sche Prinzip* als Kern hat. Die erste Etappe ist erkenntnis- und die zweite handlungsorientiert. Demgegenüber ist der Signifikanztest ausschließlich handlungsorientiert. Er macht keine Aussagen über die Hypothesen selber, über deren Wahrheit oder Falschheit oder Wahrscheinlichkeit; er *beurteilt* sie nicht. Die Hypothesen spielen eine Zwischenrolle: Man handelt so, *als ob* sie wahr oder falsch seien – bei Signifikanz *als ob* sie falsch und bei Nichtsignifikanz *als ob* sie wahr seien. Das ist die Grundlage des NEYMAN-PEARSONschen Testkonzepts. Indem der Signifikanztest aber doch eine auf Stichprobendaten beruhende Handlungsentscheidung liefert, ist er ein inferenzstatistisches Verfahren, das eine Beurteilung des Parameterraums im Sinne der Frage F1 nicht zum Gegenstand hat. Eine inferenzstatistische Tätigkeit ist somit nicht notwendigerweise mit einer Beurteilung verbunden; so sollte die Bezeichnung „Beurteilende Statistik" mit Bedacht gebraucht werden.

2 Antwort auf die Titelfrage

Der vorige Abschnitt verlangt einige Erklärungen. Doch ergibt sich bereits eine Antwort auf die Titelfrage.

(a) In Problemen vom Typus F1 führe man die erste Etappe einer BAYES-Analyse aus, wende also das BAYESsche Theorem an. Solche Probleme sind erkenntnisorientiert; es besteht ein allgemeines oder theoretisches Interesse an gewissen Hypothesen um ihrer selbst willen. Dazu zwei Schulbuchbeispiele, die allerdings an den Fundstellen nicht bayesianisch betrachtet werden: Beispiel 1 (aus LAMBACHER-SCHWEITZER 2003): „Anika behauptet, dass sie nur am Geschmack erkennt, ob der Tee mit entkalktem oder nicht entkalktem Wasser hergestellt wurde. Bei 50 Versuchen stimmt ihre Angabe in 30 Fällen." Bei-

spiel 2 (aus STRICK: *Einführung in die Beurteilende Statistik*): „Ein Losverkäufer behauptet, dass 25% der Lose aus einer Lostrommel Gewinne seien. Man beobachtet, dass unter 64 verkauften Losen nur 10 Gewinne sind. Hat der Losverkäufer die Wahrheit gesagt?“ Anhand dieses zweiten Beispiels wird im nächsten Abschnitt gezeigt, in welche Schwierigkeit man gerät, wenn man derartige F1-Probleme mit dem Signifikanztest lösen will. Zunächst aber

(b) In Problemen vom Typus F2, in handlungsorientierten Problemen, kann man *dem Grunde nach* beides tun: den Signifikanztest anwenden oder nach der 1. auch die 2. Etappe einer BAYES-Analyse ausführen - aber nur dem Grunde nach, denn die Entscheidungen für eine Handlung kommen aufgrund verschiedener Sichtweisen und Kriterien zustande, so dass sie durchaus verschieden sein können.

3 Zum Signifikanztest

Vergegenwärtigen wir uns das Wesentliche des Signifikanztests. Ihm liegt der Begriff des unbeschränkter Wiederholung fähigen Zufallsversuchs zugrunde und mit diesem die Häufigkeitsinterpretation von Wahrscheinlichkeit, die als eine Eigenschaft des Zufallsversuchs eine *objektive* Wahrscheinlichkeit ist. Der Signifikanztest ist ein solcher Zufallsversuch. Man betrachtet eine Stichprobenfunktion, die zwei Werte annehmen kann, nämlich Signifikanz oder Nichtsignifikanz. Die Stichprobenfunktion ist bestimmt durch die Art des Versuchs, durch die Wahl der H_0 und der H_A sowie der Irrtumswahrscheinlichkeiten der 1. und 2. Art, wobei letztere durch die OC-Kurve beschrieben wird. Mit dem Testergebnis ist über einen Zwischenstopp eine Entscheidung verbunden. Der Zwischenstopp besteht aus einer *Verbalentscheidung*, wie ich es nennen möchte, die lautet: „Bei Eintritt des Signifikanzereignisses wird die H_0 zugunsten der H_A abgelehnt, andernfalls angenommen.“ Ablehnung und Annahme der H_0 bedeuten nicht, sie als falsch oder wahr zu erkennen, bedeuten also keinerlei Beurteilung der Hypothesen; vielmehr ziehen Ablehnung und Annahme der H_0 die eigentliche Entscheidung nach sich, die Entscheidung für eine *Handlung*, etwa ein angebotenes Warenlos zurückzuweisen oder anzunehmen. Man handelt so, *als ob* die H_0 falsch oder wahr sei. So ist es möglich, die Grundfrage zu beantworten:

(†) *Wie muss der Test konstruiert werden*, damit in Wiederholung seiner Ausführung die zu erwartenden Schäden durch Fehlentscheidungen der 1. und 2. Art möglichst gering werden?

Als Schwachstelle des Tests ist die notwendige Vorab-Wahl der H_0 anzusehen, die die Handlungsentscheidung beeinflusst. Gewöhnlich wird empfohlen, die Wahl nach einer mehr oder minder qualitativen Abwägung der Handlungs*folgen* festzulegen und diejenige Hypothese zur H_0 zu machen, deren fälschliche Rückweisung den größeren Schaden verursacht - und das ist eine reine Angelegenheit des entscheidungssuchenden Subjekts und führt außerdem zu gewisser Willkür, wenn die möglichen Schäden als gleichermaßen unerwünscht bewertet werden.

Aufs Ganze gesehen, zeigt die kurze Skizzierung, wie kompliziert die Struktur des Signifikanztests ist, sofern man ihn nur ernst nimmt. Sie zeigt aber auch, dass die Frequentisten ihr Ziel, ein handlungsorientiertes inferenzstatistisches Verfahren *ohne Rekurs auf subjektive Wahrscheinlichkeiten* zu entwickeln, mit dem Konstrukt des Signifikanztests erreicht haben.

Bekanntlich wird der Signifikanztest aber auch angewandt, wenn kein Handlungsentscheid ansteht, wenn den Hypothesen selber das Interesse gilt, wie in den Beispielen mit Anika und dem Losverkäufer. Was soll nun die „Verbalentscheidung“, die H_0 abzulehnen oder anzunehmen, bedeuten? Die Lehrbuchliteratur, einschließlich der jüngsten Erscheinungen, äußert sich dazu in Formulierungen folgender Art: „Bei Signifikanz entscheide man sich, an die H_0 nicht zu glauben“, oder „Bei Signifikanz halte man die H_0 für falsch“, oder „Bei Signifikanz ist die H_0 mit den Daten nicht verträglich“.

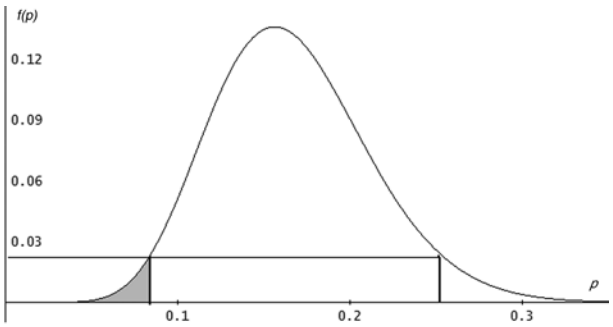
Nun kann man an etwas mehr oder minder stark glauben, kann aber nicht *sich entscheiden*, an etwas zu glauben, so wie man sich für eine Handlung entscheidet. An die H_0 nicht zu glauben oder, was dasselbe ist, sie für falsch zu halten, ist eine merkwürdig unlogische Aufforderung, weiß man doch gleichzeitig, dass eine H_0 zu Unrecht verworfen, also fälschlicherweise für falsch gehalten werden kann. So würde ja auch keine vernünftige Person die H_0 tatsächlich für falsch halten, sondern, sozusagen am dichotomen Testkonzept vorbei, sie angesichts der beobachteten Daten für *wahrscheinlich* falsch halten - was allerdings eine nicht-quantifizierte, subjektive Wahrscheinlichkeitsaussage wäre, mit der man hinter die Aussagekraft und Logik einer BAYES-Analyse zurückfiele.

Und was soll heißen „Die H_0 ist mit den Daten nicht verträglich“? Wenn das nicht einfach eine tautologische Redeweise für Signifikanz ist (und damit nichts

Neues besagt), was ist dann mit Unverträglichkeit gemeint? Wir haben es mit Zufallsereignissen zu tun; so kann von einer absoluten Unverträglichkeit im Sinne eines logischen Widerspruchs nicht die Rede sein – vielleicht aber von einer relativen? So dass etwa von zwei Hypothesen eine mit den Daten verträglicher als die andere wäre? Ich möchte zeigen, dass aus einer konstatierten Unverträglichkeit der H_0 mit dem Stichprobendatum nicht dessen Verträglichkeit mit der H_A folgt, und zwar anhand des Beispiels mit dem Losverkäufer.

Der Losverkäufer behauptet, in seiner Trommel befänden sich 25% Gewinne. Mit p für Gewinnwahrscheinlichkeit lautet die H_0: „$p = 0.25$“ gegen die H_A: „$p < 0.25$“. Mit dem Signifikanzniveau $\alpha = 0.05$ (dessen Deutung in erkenntnisorientierter Fragestellung im Übrigen schwer fallen dürfte) befindet sich das Stichprobendatum $x = 10$ im Ablehnungsbereich, was dann heißt: „Das Datum ist mit der H_0 nicht verträglich“ und damit „Das Datum ist mit der H_A verträglich“, ist also verträglich mit „Der Losverkäufer hat geschwindelt“.

Betrachten wir nun die Likelihoodfunktion, d. i. die Wahrscheinlichkeit für das Datum $x = 10$ in Abhängigkeit vom Parameter p:



Zunächst beträgt die Wahrscheinlichkeit für $x = 10$ bei Gültigkeit der H_0 0.026 (d. i. die Ordinate über $p = 0.25$).

Die H_A ist eine zusammengesetzte Hypothese bestehend aus allen $p \in [0, 0.25)$. Man sieht nun, dass mit abnehmendem p die Wahrscheinlichkeit für $x = 10$ zunächst wächst und dann abfällt, und zwar im grau markierten Bereich sogar unter den Wert 0.026, was heißt, dass das Datum mit den p-Werten in diesem Bereich der H_A noch weniger verträglich als mit der H_0 ist, was dann auch besagt, dass aus der Unverträglichkeit des Datums mit der H_0 nicht notwendigerweise die Verträglichkeit mit der H_A folgt.

Kurzum, der Signifikanztest ist das falsche Verfahren, wenn das Zutreffen von Hypothesen zu beurteilen ist. Die genannten, in der Lehrbuchliteratur vorgeschlagenen Deutungsvarianten des Signifikanzereignisses stehen in vermeintlicher Analogie zum handlungsorientierten, dichotomen Entscheidungskonzept des Tests und lassen so Erkenntnisse erwarten, die in ihm nicht angelegt sind. Gesucht ist eben $P(H|x)$ und nicht $P(x|H)$.

4 Zur Bayes-Analyse

Im Unterschied zum Signifikanztest gehört die subjektive Wahrscheinlichkeit, auch *Kenntniswahrscheinlichkeit* genannt (im Gegensatz zur objektiven *Voraussagewahrscheinlichkeit*), zu den Kernbegriffen einer BAYES-Analyse. Subjektive Wahrscheinlichkeit wird als ein notwendiger Begriff verstanden, der im Wechselspiel zwischen erkennendem Subjekt und dem Erkenntnisobjekt Grade von Erkenntnissicherheit beschreibt. So kann einer unbekannten objektiven Ereigniswahrscheinlichkeit (z.B. der unbekannten Gewinnwahrscheinlichkeit p beim Losverkäufer) eine subjektive Erkenntniswahrscheinlichkeit zugeordnet werden, was das Anwendungsfeld der Inferenzstatistik vergrößert und flexibilisiert. Die BAYESsche Theorie hat mitsamt ihrer Anwendung in den Natur- und Gesellschaftswissenschaften in den letzten Jahrzehnten enorme Weiterungen erfahren, und das frühe Gedankengut wird neu gesehen. Dazu gehört vor allem der Begriff der subjektiven Wahrscheinlichkeit, dem der Ruch des Ideosynkratischen, des persönlich Eigenartigen, des wissenschaftlich Unbrauchbaren anhaftet. Nichtsdestoweniger ist der Vorgang des Beurteilens ein menschlicher Akt, d. h. *notwendigerweise* subjektiv. Man kann sich Klarheit verschaffen, indem man die Grenzen zwischen *objektiv*, *subjektiv* und *intersubjektiv* ein wenig anders als im Sprachgebrauch üblich zieht (was in [1] näher beleuchtet ist). Historisch ist anzumerken, dass der einflussreiche Sir R. A. FISHER – der Objektivist und Frequentist, dem die Entwicklung der Statistik so viel zu verdanken hat – der subjektiven Wahrscheinlichkeit und mit ihr dem BAYESschen Theorem reserviert gegenüber gestanden hat. Mit einem starken Argument hat er der BAYESschen Theorie einen empfindlichen Stoß versetzt, der heute noch mehr oder minder offen nachwirkt. FISHER wollte mit seinem *Transformationsargument*, wie man's heute nennt, zeigen, dass die Darstellung von unsicherem Wissen, insbesondere von gänzlichem Nichtwissen, nicht eindeutig und damit beliebig ist. Die schlagende Wirkung des Arguments

hat die objektivistische Sicht der Dinge entscheidend begünstigt. Es ist merkwürdig, dass im Entwicklungsschub der BAYES-Theorie die FISHERsche Kritik erst spät entkräftet worden ist; erst Ende der Achtzigerjahre hat sich F. SCHREIBER, ein Theoretiker der Nachrichtentechnik, der Sache angenommen. Im Zusammenhang mit der Analyse störungsbehafteter Signalübertragung, zu der er bayessche Methoden verwendete, konnte er zeigen, dass das FISHERsche Argument nicht stichhaltig ist. SCHREIBERs Entgegnung befreit die BAYES-Theorie von einem ihr angehängten Makel, und deshalb halte ich es für wichtig, besonders darauf hinzuweisen. (Mehr dazu und Literatur in [1], S. 81, Fußnote 45.)

Im übergreifenden Vergleich des Signifikanztests mit der BAYES-Analyse ist deren logisch klarere Struktur zu verzeichnen, die besonders dem Neuling in der Inferenz-Statistik das Verständnis erleichtert. Die erste Etappe liefert die Antwort auf die Frage: „Wie wahrscheinlich sind die hypothetischen Zustände je der Fall?“ In dieser Etappe wird nichts entschieden, und die BAYES-Analyse endet hier, wenn allein das Zutreffen der möglichen Hypothesen bewertet werden soll. Die zweite Etappe liefert die Antwort auf die Frage: „Welche ist die beste Handlung im Sinne des BAYESschen Prinzips?“ Hier werden die Handlungs*folgen* in Gestalt der Nutzenfunktion quantitativ bewertet. Jede Stelle im quasi modularen Analyseablauf lässt klar die (inter-)subjektiven und die objektiven Einflüsse auf die Handlungsentscheidung erkennen, und jeder Schritt wird quantitativ vollzogen.

Zusammengefasst: Der Blick des Signifikanztests ist *vorwärts* gerichtet: Man stellt die Frage (†) in Abschnitt 3. Es ist der Blick der *Fließbandsituation*, deren Standardbeispiel die Qualitätskontrolle ist. Demgegenüber ist der Blick der BAYES-Analyse *rückwärts* gerichtet: *Nicht* stellt man die Frage, mit welcher Wahrscheinlichkeit gewisse Stichprobendaten bei Ausführung eines Zufallsversuchs zu erwarten sind; vielmehr geht man von den Stichprobendaten (Indizien) aus und fragt umgekehrt nach der Wahrscheinlichkeit, mit der gewisse, die Daten möglicherweise beeinflussende Ursachen (Zustände, Parameter) je der Fall sind. — Schaut man sich die unsichere Welt daraufhin an, scheinen die Fragen mit bayesschem Rückwärtsblick deutlich in der Überzahl zu sein.

5 Empfehlung

So möchte man die Lehrbuchautoren auffordern, dem Geist der BAYES-Theorie ein größeres, wenn nicht das größere, Gewicht zu verleihen. Tatsächlich sind in einigen Neuerscheinungen Ansätze in dieser Richtung unternommen worden, sie sind jedoch zu zaghaft. Die BAYESsche Methodik als eine interessante Alternative zum klassischen Signifikanztest zu charakterisieren, ist zu blass — allein angesichts der genannten Deutungsschwierigkeit, die mit dem Hypothesentesten im erkenntnisorientierten Sinn verbunden ist.

Die Kenntniswahrscheinlichkeit, deren formales Paradigma die faire Wette ist, ist ein Kernbegriff der BAYES-Theorie. Sie als Maß der Urteilssicherheit zu entwickeln, das von der physikalischen Objektivität einer Voraussagewahrscheinlichkeit semantisch zu unterscheiden ist, sollte bereits Anliegen eines SI-Stochastikkurses sein, weit bevor kompliziertere Dinge (wie etwa das BAYESsche Theorem) in Angriff genommen werden. (Zur didaktisch geschickten Vermittlung des BAYESschen Theorems s. Literatur in [1].)

6 Anmerkungen zum Rechnerprogramm VisualBayes

Bleibt noch zu erwähnen, dass der BAYES-Analyse ein technisches Hindernis entgegengestanden hat, nämlich das des relativ hohen Rechenaufwandes, sobald die Vorverteilung[a] nicht von „ganz einfachem“ Typus ist. So ist dem Titel [1] ein (auf DERIVE aufgebautes) Programmpaket auf einer CD beigefügt, mit dem der Schüler oder Student ohne Rechenaufwand Inferenzprobleme auf bayessche Art lösen können soll. Eine Besonderheit des Programms besteht in der Möglichkeit, mit der Maus die Vorverteilung auf dem Bildschirm frei zu zeichnen, so dass beliebige Vorkenntnisse vom Parameterraum in den Kalkül eingehen können, ohne auf die früher notwendigen *konjugierten Funktionsfamilien* rekurrieren zu müssen.

[a] *A-priori-Verteilung* und *A-posteriori-Verteilung* sind bekanntlich in der BAYES-Methodik sehr häufig verwendete Begriffe. Warum nicht diese schwerfälligen Ausdrücke einfach durch die Bezeichnungen *Vorverteilung* bzw. *Nachverteilung* ersetzen? (Was in [1] bereits geschehen ist.)

Literatur

[1] WICKMANN D.: *VisualBayes. Ein Rechnerprogramm zur Einführung in die Bayes-Statistik.* Verlag Franzbecker, Hildesheim 2006. 89 Seiten plus Programm-CD.

Kontakt
dieter.wickmann@post.rwth-aachen.de