Bayes-Statistik

Ivo W. Molenaar und Charles Lewis

Bei der Auswertung von Umfragedaten, Studien oder Experimenten werden üblicherweise statistische Methoden wie Parameterschätzung oder statistisches Hypothesentesten eingesetzt, um die Verläßlichkeit von aus den Daten gezogenen Schlußfolgerungen abschätzen zu können (vgl. die Kapitel von Diepgen, Klauer und Willmes, in diesem Band). Wahrscheinlichkeiten werden hierbei als asymptotische relative Häufigkeiten interpretiert, und die Kernfrage scheint zu sein, welche Ergebnisse resultieren würden, wenn man die vorliegende Untersuchung vielfach replizieren würde.

In diesem Kapitel wird die Bayes-Statistik vorgestellt, eine alternative statistische Methodenlehre, in der Wahrscheinlichkeiten auch Vorwissen bzw. Unsicherheit über die Populationsparameter, welche die Ergebnisse der Datenerhebung steuern, ausdrücken können. Abschnitt 1 gibt einen Überblick über verschiedene Quellen solchen Vorwissens. In Abschnitt 2 wird die Bayes-Methode anhand von zwei Beispielen veranschaulicht. Beispiel 1 behandelt das Problem der Schätzung von „wahren Werten" in Intelligenztests (vgl. Stumpf, in diesem Band), Beispiel 2 die Schätzung des Frauenanteils unter den Studierenden einer Universität.

Der erfolgreiche Einbezug von Vorwissen in die statistische Analyse setzt voraus, daß dieses Wissen valide und reliabel gemessen werden kann. In Abschnitt 3 wird argumentiert, daß dieses Problem ein nicht-triviales ist, dessen Lösung sowohl eine genaue Erhebungsmethode als auch ausreichendes Training der Person verlangt, deren Wissen in Form einer sogenannten Apriori-Verteilung formalisiert wird. Abschnitt 4 stellt eine Bayesianische Sicht der Ergebnisse von Parameterschätzungen und Hypothesentests vor. Im letzten Abschnitt werden Anwendungen von Bayes-Methoden in den Verhaltenswissenschaften vorgestellt und diskutiert.

1 Vorwissen und statistisches Urteilen

Den meisten Anwendern des üblichen, auf der Stichprobentheorie basierenden Zugangs zu Problemen der statistischen Inferenz (Parameterschätzung, Hypothesentesten) ist vermutlich nicht bewußt, daß es einen alternativen Ansatz zur Behandlung dieser Probleme gibt. Diese Alternative ist in einer Hinsicht radikal verschieden vom Standardverfahren, in einer anderen diesem jedoch wieder sehr ähnlich. Es ist das Ziel dieses Kapitels, sowohl die Unterschiede als auch die Ähnlichkeiten von Stichprobentheorie und dieser Alternative – der Bayesschen Inferenz – aufzuzeigen. Auf diese Weise hoffen wir, Leser über einen Zugang zur statistischen Inferenz in Kenntnis zu setzen, von dem wir meinen, daß er sehr nützlich ist. Zugleich wollen wir eine kritische Evaluation der allgemein gebräuchlichen Methoden der Parameterschätzung und des Hypothesentestens vornehmen.

https://doi.org/10.25521/HQM12

Wenn wir als Hauptziel statistischer Inferenz die Optimierung von Entscheidungen unter Unsicherheit sehen, dann impliziert dies, daß wir versuchen sollten, diese Unsicherheit dadurch zu reduzieren, daß wir alle verfügbaren Informationen so effektiv wie möglich nutzen. Während die Stichprobentheorie der Inferenz sich auf den Gebrauch von Informationen konzentriert, die in einem Datensatz verfügbar sind, erlaubt die Bayessche Inferenz überdies den Einbezug von Informationen aus anderen Quellen (üblicherweise als Vorwissen bezeichnet).

Die Bezeichnung „Bayessche Inferenz" nimmt Bezug auf Thomas Bayes, der in einem Artikel von 1763 erstmalig eine Methode der Kombination von Vorwissen mit Informationen aus Daten zum Zwecke statistischer Inferenz vorschlug. Hierzu leitete er ein wichtiges Resultat der mathematischen Wahrscheinlichkeitstheorie her, das man heute unter der Bezeichnung „Bayes-Theorem" kennt. Dieses Theorem beschreibt eine Beziehung zwischen bedingten Wahrscheinlichkeiten. Um es allerdings für das hier diskutierte Problemfeld nutzbar machen zu können, mußte Bayes die Annahme machen, daß es möglich ist, Vorwissen in Form von Wahrscheinlichkeiten zu beschreiben. Die Interpretation von Wahrscheinlichkeiten als asymptotische relative Häufigkeiten mußte also auf Maße der Sicherheit bzw. Unsicherheit erweitert werden.

Diese Ausweitung der Wahrscheinlichkeitsinterpretation ist Gegenstand beachtlicher Kritik am Bayesschen Ansatz statistischer Inferenz gewesen. Der Haupteinwand war, daß Vorwissen im Kontext einer objektiven wissenschaftlichen Inferenz unberücksichtigt bleiben sollte, da es von Individuum zu Individuum variieren kann. Vielmehr muß die Aufmerksamkeit auf Daten beschränkt bleiben, die im Kontext der jeweiligen Forschungsfragestellung erhoben wurden. Bei der Diskussion dieses Kritikpunktes ist es nützlich, zwischen verschiedenen potentiellen Quellen von Vorwissen zu unterscheiden. Am einfachsten ist es, wenn das Vorwissen Häufigkeitsangaben entstammt, so daß eine probabilistischen Beschreibung der Unsicherheit bereits verfügbar ist. Wir werden im nächsten Abschnitt ein solches Beispiel vorstellen.

Eine andere, besondere Situation liegt vor, wenn kein Vorwissen verfügbar ist oder wenn wir versuchen, uns so zu verhalten, als sei kein Vorwissen vorhanden. Ein Forschungszweig der Bayes-Statistik hat sich mit der Entwicklung von Apriori-Wahrscheinlichkeitsverteilungen beschäftigt, die Unwissenheit, Indifferenz oder fehlendes Vorwissen repräsentieren sollen. Man kann zeigen, daß auf der Grundlage derartiger Apriori-Wahrscheinlichkeiten Ergebnisse resultieren, die den im Rahmen der Stichprobentheorie erzielten numerisch gleichen, wobei sich allerdings die Interpretationen der Ergebnisse unterscheiden. Auf diesen Punkt werden wir in Abschnitt 4 zurückkommen.

Schließlich können Apriori-Wahrscheinlichkeiten auf Angaben von Individuen beruhen, die ihre persönlichen Unsicherheiten bezüglich unbekannter Größen – üblicherweise Parameter in statistischen Modellen – reflektieren. Ein Beispiel für einen solchen Fall wird in Abschnitt 2 diskutiert. Die Unsicherheits-Angaben können natürlich von Individuum zu Individuum unterschiedlich ausfallen und dadurch zu unterschiedlichen Schlußfolgerungen bei identischem Datensatz führen. Man kann allerdings zeigen, daß die aus der Bayesschen Analyse resultierenden *Aposteriori*-Wahrscheinlichkeiten in einem solchen Fall besser übereinstimmen als die entsprechenden *Apriori*-Wahrscheinlichkeiten, welche die Unsicherheits-Angaben reflektie-

ren. Man kann somit sagen, daß Bayessche Inferenz einen Rahmen zur Reduktion bestehender Differenzen zwischen Forschern auf der Grundlage ihrer Forschungsbefunde bereitstellt. Einige Punkte, die bei der Erhebung von Unsicherheits-Angaben für diesen Fall eine Rolle spielen, werden in Abschnitt 3 besprochen.

2 Zwei einfache Beispiele

In der klassischen Testtheorie nimmt man an, daß sich der beobachtete Wert X einer Person additiv aus einem „wahren Wert" T und einem Fehlerwert E zusammensetzt (Stumpf, in diesem Band). Gehen wir einmal davon aus, daß ein Proband John in einem Test einen beobachteten IQ von $X = 130$ erzielt hat. Welche Aussage können wir über seinen wahren Wert T machen?

Nehmen wir an, wir wüßten, daß John einer Gruppe angehört, deren wahrer IQ normalverteilt ist mit einem Mittelwert von 100 und einer Varianz von 180, und daß für jeden Angehörigen dieser Gruppe die Fehlerwerte wiederum normalverteilt sind, unabhängig von T, mit einem Mittelwert von Null und einer Varianz von 45. Dann wäre für ein zufällig gewähltes Gruppenmitglied der beobachtete Testwert X $N(100, 225)$-verteilt (normalverteilt mit Mittelwert 100 und Varianz 225). Die Reliabilität des Tests wäre $\rho = \sigma_T^2/\sigma_X^2 = 180/225 = 0.80$. Bevor man mit dem Test den IQ von John erhebt, müßten wir den Gruppenmittelwert von 100 als einen – zugegebenermaßen sehr groben – Apriori-Schätzer seiner Intelligenz heranziehen. Man beachte, daß in diesem Beispiel das Wissen um die Häufigkeitsverteilung von Johns Bezugspopulation als Vorwissen herangezogen wird.

Die Stärke der Bayesschen Methode liegt nun darin, daß sie eine Regel zur Verfügung stellt, welche die Kombination solchen Vorwissens mit fehleranfälliger Evidenz aus Daten erlaubt. Tatsächlich kann man zeigen, daß im Falle eines beobachteten Wertes $X = 130$ die gewichtete Kombination $\rho \cdot 130 + (1 - \rho) \cdot 100 = 124$ die beste Aposteriori-Schätzung (in einem in Abschnitt 4 ausgeführten Sinne) für Johns wahren Wert ist. Ein formaler Beweis würde zuviel Platz beanspruchen (s. dazu z.B. Lord & Novick, 1968, Kap. 3.7), aber die zugrundeliegende Idee läßt sich leicht erklären. Der Fehlerwert E hat eine symmetrische Verteilung, in der $E = 5$ gerade genauso plausibel ist wie $E = -5$. Es gibt allerdings bei weitem mehr Personen mit $T = 125$ als mit $T = 135$, weil T $N(100, 180)$-verteilt ist. Also kann das Ergebnis $X = 130$ plausibler durch einen T-Wert unterhalb von 130, kombiniert mit einem positiven E (ein mäßig begabter Student, der Glück hat), erklärt werden als durch einem T-Wert oberhalb von 130, kombiniert mit einem negativen E (ein hochbegabter Student, der Pech hat).

In unserem zweiten Beispiel betrachten wir den Anteil π von Frauen unter den Studierenden der Universität von Groningen. Leserinnen und Leser mögen für einen Moment selber überlegen, welche Werte sie für diesen unbekannten Parameter π für plausibel halten. Man wird anhand dieses Beispiels feststellen, daß das Vorwissen aus mehr oder weniger vagen Erinnerungen an Geschlechterverteilungen in Untergruppen und/oder in ähnlichen Institutionen besteht. Anders als im ersten Beispiel gibt es für die Apriori-Wahrscheinlichkeiten keine Häufigkeitsinterpretation; vielmehr drücken sie den „Grad der Angemessenheit einer Meinung" aus. Als wir dieses Beispiel in unserer eigenen Forschungsarbeit verwendeten, sagte eine unserer Versuchspersonen

(Vpn), er sei sich ziemlich sicher, daß $0.15 < \pi < 0.50$ gilt, bei einem „*best guess*" von $\hat{\pi} = 0.32$ und einem besonders plausiblen Bereich $0.25 < \pi < 0.40$.

Wir haben einige Aspekte des Vorwissens einer Person über den unbekannten Parameter π skizziert. Wenn wir dies in die mathematische Form einer Wahrscheinlichkeitsdichteverteilung bringen wollen, ist es oft angemessen, eine bestimmte Familie solcher Dichten auszuwählen. Die Familie der Beta-Verteilungen ist auf der einen Seite ausreichend flexibel, um verschiedene Formen von Vorwissen darstellen zu können, und auf der anderen Seite leicht kombinierbar mit der binomialen *Likelihood*-Funktion einer Zufallsstichprobe. Ihre Dichtefunktion lautet

$$f(\pi) = \pi^{a-1}(1-\pi)^{b-1}/B(a,b), \tag{1}$$

wobei $B(a,b)$ die sogenannte Beta-Funktion der positiven Parameter a und b ist (vgl. auch Rost & Erdfelder, in diesem Band).

Mit Hilfe der Methoden, die im folgenden Abschnitt 3 besprochen werden, wurde gezeigt, daß eine Beta(a, b)-Dichte mit den Parametern $a = 16$ und $b = 34$ das Vorwissen unserer Vp angemessen beschrieb (s. Abbildung 1).

Würde man dieser Vp sagen, daß eine Zufallsstichprobe von $n = 30$ Studierenden

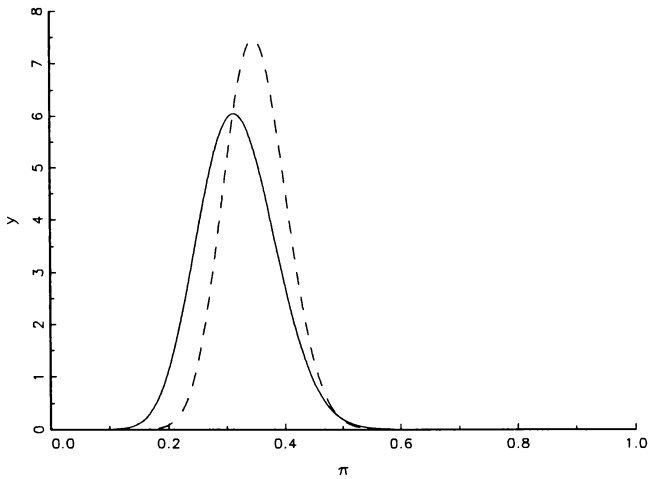

ABBILDUNG 1. **Die Beta(16,34)-Apriori-Verteilung des Studentinnen-Anteils (durchgezogene Kurve) und die entsprechende Beta(28,52)-Aposteriori-Verteilung (unterbrochene Kurve).**

gezogen wurde, in der $x = 12$ weiblich waren und $n - x = 18$ männlich, dann wäre nach dem Bayes-Theorem die Aposteriori-Dichte von π, bei gegebenem x

$$f(\pi|x = 12) = \pi^{16+12-1}(1-\pi)^{34+18-1}/B(28,52). \tag{2}$$

Exakt die gleiche Aposteriori-Verteilung würde im übrigen bei einer Stichprobe von 80 Studierenden, bestehend aus 28 Frauen und 52 Männern, resultieren, wenn von einer Apriori-Dichte ohne Information ausgegangen würde. Man kann somit sagen,

daß die Apriori-Verteilung unserer Vp soviel Informationswert besitzt wie eine Stichprobe von 50 Studierenden, darunter 16 Studentinnen. Es ist oft nützlich, Vorwissen in dieser Weise als „fiktive Zusatzstichprobe" zu betrachten. Zwar gibt es Meinungsverschiedenheiten darüber, ob „überhaupt keine Information" gleichbedeutend ist mit „keine Beobachtungen" (was auf eine degenerierte Beta(0, 0)-Apriori-Dichte hinausliefe) oder aber mit „Beobachtung eines Erfolgs und eines Mißerfolgs" (was einer Beta(1, 1)-Apriori-Dichte, d.h. einer Gleichverteilung im Intervall [0, 1], entspräche). Diese Feinheiten sollen uns hier jedoch nicht beschäftigen.

Die Aposteriori-Verteilung für $x = 12, n = 30$ ist ebenfalls in Abbildung 1 dargestellt. Man kann zeigen, daß ihr Mittelwert $28/80 = 0.35$ ist, was einem gewichteten Mittel $\frac{50}{80} \cdot 0.32 + \frac{30}{80} \cdot \frac{12}{30}$ des Apriori-Mittelwertes 0.32 und des Stichprobenmittelwerts 0.40 entspricht, analog zum o.g. Testwert-Beispiel. Man kann der Abbildung 1 entnehmen, daß die Aposteriori-Varianz kleiner ist als die Apriori-Varianz, da ersterer eine größere Stichprobe entspricht. Dies ist das normale Ergebnis einer Bayesschen Analyse. Wenn aber der Stichprobenmittelwert deutlich von dem der Apriori-Verteilung abweicht, kann es sein, daß die Aposteriori-Verteilung im Vergleich zur Apriori-Verteilung eine größere Varianz aufgeweist. Die Mathematik spiegelt in diesem Fall die Intuition wider, daß Beobachtungen, die nahe an der eigenen Erwartung liegen, zu einer erhöhten Sicherheit bezüglich des interessierenden Parameters führen können, während unerwartete Beobachtungen die Sicherheit reduzieren.

3 Erhebung des Vorwissens

Jeder Entscheidungsträger oder Forscher, der die Kombination von Vorwissen mit der Evidenz aus statistischen Daten modellieren möchte, muß ein valides und reliables Vorgehen finden, mit Hilfe dessen jede dieser Komponenten in eine mathematische Formel übersetzt werden kann. Für die Daten impliziert dies die Standardmodellierung von Beobachtungen in Form einer *Likelihood*-Funktion, die von geeignet gewählten Parametern abhängt. Hier ist das Vorgehen das gleiche wie in der Standard-Stichprobentheorie. Für beide Typen von Inferenz ist es gleichermaßen bedeutsam, daß die Verteilungsannahmen erfüllt sind. Im Beispiel „weibliche Studierende" würde das eine echte Zufallsstichprobe aus der gleichen Grundgesamtheit verlangen (konstantes π, unabhängige Ziehungen), was auf eine binomiale *Likelihood*-Funktion hinausläuft.

Typisch für das Bayessche Vorgehen ist die zusätzliche Berücksichtigung einer Apriori-Verteilung der unbekannten Parameter, die das Vorwissen einer Person widerspiegeln sollen. Man kann davon ausgehen, daß die meisten Forscher zumindest ein vage Vorstellung davon haben, welche Werte für die Parameter plausibel sind und welche nicht. Das Problem ist nun, wie sich solche eher diffusen Vorstellungen in eine formale Apriori-Verteilung „übersetzen" lassen.

Die Befragung von Experten und Novizen bezüglich ihres Vorwissens war Gegenstand zahlreicher Experimente. Verschiedene Befragungsmethoden sind dazu vorgeschlagen und verglichen worden; Überblicke finden sich zum Beispiel bei Hogarth (1980) und Schütt (1981).

Im Beispiel über den Frauen-Anteil π unter den Studierenden der Universität Groningen könnten mögliche Fragen sein:

„Nennen Sie den wahrscheinlichsten Wert für π."
„Geben Sie das untere Quartil an, also den Wert, bezüglich dessen Sie 3 zu 1 wetten würden, daß π größer ist als dieser Wert."
„Ordnen Sie die Intervalle $0 < \pi < 0.1$, $0.1 < \pi < 0.2$, $0.2 < \pi < 0.3, \ldots$ vom wahrscheinlichsten zum unwahrscheinlichsten."
„Wenn ich Ihnen sage, daß π höchstens 0.4 ist, ist dann $0.3 < \pi < 0.4$ plausibler als $0 < \pi < 0.3$?"

Es scheint, daß die Vertrautheit mit statistischen Methoden keineswegs logische, konsistente und stabile Antworten auf solche Befragungen garantiert. Einhorn und Hogarth (1981), Peterson und Beach (1967), Tversky (1974) sowie Tversky und Kahneman (1983) präsentieren ausreichende Evidenz dafür, daß „der Mensch als intuitiver Statistiker" oft schlecht abschneidet.

Da die valide und reliable Formalisierung von Vorwissen eine notwendige Bedingung für den erfolgreichen Vorwissens-Einbezug in eine Bayessche Analyse ist, wurden viele Versuche unternommen, Auswege aus dieser schwierigen Lage zu finden. Ein Ausweg ist die Kombination einer Reihe von verschiedenen Befragungsmethoden, gegebenenfalls gefolgt von Rückmeldung über mögliche Inkonsistenzen. Weil die Methoden unterschiedlich sensitiv für diverse Probleme und Artefakte sind, ist zu erwarten, daß ihre kombinierte Anwendung zu besseren Messungen des Vorwissens führt als die Anwendung jeder Methode einzeln. Interaktive Computerprogramme wie CADA (Novick & Jackson, 1974) und SPAT (Lourens, 1984; Terlouw, 1989) basieren auf diesem Prinzip. Der zweite Ausweg ist Übung. Es kann gezeigt werden, daß sich die Leistungen der Informanten verbessern, wenn sie regelmäßig über ein bestimmtes Wissensgebiet befragt werden und für jede Schätzung Rückmeldung über die Korrektheit der Schätzung erfolgt.

Sowohl die Anwendung von interaktiven Computerprogrammen, die verschiedene Befragungsmethoden kombinieren, als auch das Training der Informanten führten zu Verbesserungen der Qualität der Apriori-Verteilungen. Dennoch waren die Ergebnisse enttäuschend (Lourens, 1984; Terlouw, 1989). Die vollständige und sorgfältige Erhebung eines einzelnen Parameters (etwa einer Populationsproportion) dauerte normalerweise zwischen 5 und 15 Minuten. Einige Informanten hatten Schwierigkeiten mit der statistischer Terminologie, wie z.B. „unteres Quartil" oder „90 Prozent Wahrscheinlichkeit", und selbst nach ausgiebigem Training neigten viele Informanten immer noch dazu, die Genauigkeit ihres Vorwissens zu überschätzen. Wenn die Experimentatoren Almanachfragen verwendeten, für die sie die korrekten Antworten kannten – z.B. „Wieviel Prozent der holländischen Haushalte besitzt einen Videorecorder?" – zeigte sich, daß bei weit mehr als 10 Prozent der Urteile der wahre Wert nicht in dem vom Informanten erzeugten 90-Prozent-Konfidenzintervall enthalten war. Dieses Phänomen der „*overconfidence*" oder des „*lack of calibration*" läßt sich für viele Fragentypen und viele Klassen von Informanten finden.

Ein neueres Erhebungsinstrument, ELI, versucht diese Nachteile zu beseitigen. Es sollte eine schnelle und einfache Benutzung erlauben, das System sollte frei von statistischen Konzepten sein, anstelle von Zahleneingaben mit der Manipulation von graphischen Darstellungen arbeiten und sowohl Rückmeldungen als auch Vorausblicke erlauben. „Vorausblick" bezieht sich hierbei auf die Darbietung eines auf einer geeigneten Bewertungsregel basierenden Graphen, der es erlaubt, für jeden mögli-

chen Wert des unbekannten Parameters die Verluste oder Gewinne einzusehen, die resultieren würden, wenn dies der wahre Wert wäre. „Rückmeldung" heißt, daß das Instrument zu Trainingszwecken eingesetzt werden kann, wobei der Experimentator vor der Übung den wahren Wert für jedes Problem eingibt. Dieser Wert und der resultierende positive oder negative Punktestand können der Vp dargeboten werden, in der frühen Lernphase nach jedem einzelnen Problem, später dann kumulativ für eine Folge von Problemen. Die ersten Anwendungen von ELI verliefen recht zufriedenstellend (van Lenthe, 1993a, 1993b, 1993c, 1994). Leser können die Software zusammen mit dem Manual (van Lenthe & Molenaar, 1993) bestellen und in verschiedenen Kontexten ausprobieren.

4 Interpretation der Ergebnisse

Um das Rationale von Stichprobentheorie und Bayesscher Inferenz besser zu verstehen, ist es sinnvoll, beide hinsichtlich der Information, die sie für eine statistische Analyse liefern, zu vergleichen. Bezüglich des Problems, einen Punktschätzer für einen Parameter zu liefern, wie es in Abschnitt 2 dargestellt wurde, konzentriert sich die Stichprobentheorie auf die asymptotischen Eigenschaften solcher Schätzer. Diese sind letztlich Eigenschaften der Stichprobenverteilung dieser Schätzer. Zum Beispiel ist ein *unverzerrter* Schätzer (*unbiased estimate*) eines Parameters dadurch ausgezeichnet, daß der Mittelwert der Stichprobenverteilung mit dem wahren Wert des Parameters übereinstimmt (vgl. Klauer, in diesem Band).

In der Bayesschen Inferenz richtet sich die Aufmerksamkeit dagegen auf die Aposteriori-Verteilung des zu schätzenden Parameters. So könnte man den Mittelwert dieser Verteilung als Punktschätzer für den Parameter einsetzen, wie es in den Beispielen in Abschnitt 2 getan wurde. Für eine solche Bayessche Schätzung spielen asymptotische Eigenschaften des Schätzers (z.B. Unverzerrtheit) keine Rolle. Statt dessen dient die Schätzung als beste Beschreibung der Aposteriori-Verteilung durch einen einzelnen Wert und diese wiederum als vollständige Beschreibung des verfügbaren Wissens bezüglich des interessierenden Parameters, als Kombination des Vorwissens mit der Information aus den analysierten Daten.

Eigenschaften der Konfidenzintervalle basieren in der Stichprobentheorie ebenfalls auf Überlegungen bezüglich der Stichprobenverteilung. So ist ein 95% Konfidenzintervall für eine Populationsproportion π so konstruiert, daß über eine unendliche Folge von (hypothetischen) Replikationen hinweg die Wahrscheinlichkeit, daß der wahre Wert von π in einem auf diese Art konstruierten Intervall liegt, gerade .95 ist. Insbesondere kann also keine Wahrscheinlichkeitsaussage über das jeweilige konstruierte Intervall und seine Beziehung zum tatsächlichen π gemacht werden. (Im Falle des Frauen-Anteiles unter den Studierenden in Groningen reicht dieses Intervall von 0.23 bis 0.59.)

Dagegen erlauben Bayessche Intervallschätzungen solche direkten Wahrscheinlichkeitsaussagen, und zwar in bezug auf die Aposteriori-Verteilung des Parameters. Ein Bayessches 95% Intervall für π ist darin nämlich enthalten. Genau wie im Falle der Bayesschen Punktschätzung kann man von einem solchen Intervall annehmen, daß es zusammenfassende Information über die relevante Aposteriori-Verteilung liefert. Das (*highest density*) 95% Aposteriori-Intervall für den Studentinnen-Anteil ist

(0.25, 0.45), was sich von dem o.g. Konfidenzintervall ziemlich unterscheidet. Das muß allerdings nicht immer der Fall sein. Benutzt man zum Beispiel in der Varianzanalyse eine Apriori-Verteilung, die vollständige Unwissenheit über Populationsmittelwerte und Populationsfehlervarianz repräsentiert, erhält man Intervalle, die denen aus der Stichprobentheorie numerisch identisch sind. Trotzdem verbleibt auch hier der Unterschied in der Interpretation dieser Intervalle im Rahmen beider Ansätze (s. Lewis, 1993, für weitere Informationen).

Die gebräuchlichsten Instrumente bei der Stichprobentheorie-Inferenz sind solche des statistischen Hypothesentestens (vgl. Willmes, in diesem Band). Insbesondere das Testen der Nullhypothese über den wahren Wert eines Parameters oder einer Gruppe von Parametern, zusammen mit den Konzepten der Typ-1- und Typ-2-Fehler, Signifikanzniveau und *power*, bilden den Kern fast jeder Statistikeinführung. Die Stichprobenverteilung der Teststatistiken spielen eine entscheidende Rolle. So ist das Signifikanzniveau eines Tests die Wahrscheinlichkeit (asymptotische relative Häufigkeit), mit welcher der Test zur Verwerfung der Nullhypothese führt, falls sie wahr ist. Es basiert auf der Stichprobenverteilung der Teststatistik bei Gültigkeit der Nullhypothese.

In seiner einfachsten Form benutzt Bayessches Hypothesentesten die Aposteriori-Verteilung des oder der betrachteten Parameter. Die Nullhypothese muß die Form eines Intervalls (oder Bereichs) von Parameterwerten annehmen. Wenn die Aposteriori-Wahrscheinlichkeit für diesen Bereich ausreichend klein ist, wird die Hypothese verworfen. Anders ausgedrückt: Die interessierende Größe ist die *Wahrscheinlichkeit der Gültigkeit der Nullhypothese bei gegebenen Daten.* Dies steht im krassen Gegensatz zur Stichprobentheorie, in der das Interesse auf die Wahrscheinlichkeit bestimmter Daten bei Gültigkeit der Nullhypothese gerichtet ist. Nun handelt es sich bei den Daten lediglich um die Resultate einer Studie, während sich die eigentlichen Forschungsfragen mit dem Zutreffen oder Nichtzutreffen der Nullhypothese befassen. Daher sollte es nicht überraschen zu erfahren, daß Autoren im Bereich der Bayes-Statistik das konventionelle Hypothesentesten als Beschäftigung mit der „falschen" bedingten Wahrscheinlichkeit charakterisiert haben.

Ein weiterer Unterschied der beiden Ansätze im Kontext des Hypothesentestens ist die Konzentration auf Punkthypothesen in der Stichprobentheorie und auf Intervallhypothesen in der Bayesschen Analyse. Obwohl eine Analyse von Punkthypothesen mit Bayesschen Methoden prinzipiell möglich ist, sind die den Punkthypothesen zugeordneten Apriori- und Aposteriori-Wahrscheinlichkeiten normalerweise gleich Null. Man könnte im Beispiel des Frauenanteils unter Groninger Studierenden im Rahmen eines konventionellen, stichprobentheoretischen Zugangs einen Test der Nullhypothese $\pi = 0.5$ gegen die Alternativhypothese $\pi \neq 0.5$ in Erwägung ziehen. Verwendet man als Apriori-Verteilung für π eine Beta-Dichte (oder irgendeine andere Dichtefunktion), so ist die Wahrscheinlichkeit, die einem Parameterwert von exakt 0.5 zugeordnet wird, gleich Null; unabhängig von dem Ergebnis der Datenerhebung wird diese Wahrscheinlichkeit in der Aposteriori-Verteilung Null bleiben. Folglich muß die Nullhypothese erweitert werden, um nichttrivial zu sein, z.B. auf den Bereich $0.40 < \pi < 0.60$. Mit anderen Worten: Anstelle der Frage, ob die Anteile der Geschlechterverteilung *genau* gleich sind – was sicherlich verneint werden muß – sollte man besser fragen, ob die Anteile *annähernd* gleich repräsentiert sind,

z.B. unter Verwendung einer Toleranzzone von ± 10%. Für die in Abschnitt 2 gegebenen Apriori- und Aposteriori-Verteilungen sind die Wahrscheinlichkeiten für das Zutreffen dieser erweiterten Hypothese 0.11 beziehungsweise 0.17. Hierbei spielt das Vorwissen eine entscheidende Rolle. Unsere Vp war sich ziemlich sicher, daß weniger als die Hälfte der Studentenschaft weiblich ist. Ein Apriori-Verteilung ohne Information, kombiniert mit den gleichen Daten, würde eine Aposteriori-Wahrscheinlichkeit von 0.48 für die Hypothese ergeben, ein eher mehrdeutiges Ergebnis.

Eine stichprobentheoretische Analyse der Daten (12 Frauen, 18 Männer) erlaubt keine Zurückweisung der Nullhypothese über Gleichheit auf dem .05 Niveau. Novick und Jackson (1974, S. 245) kritisierten ein solches Ergebnis als „eine irreführende Antwort auf eine Frage, die niemand stellt!"

5 Übersicht über Anwendungen

Diskutiert man die Anwendungen Bayesscher Überlegungen auf psychologische Fragestellungen, so ist es nützlich, zwei generelle Kategorien zu unterscheiden: inhaltliche und methodologische Fragestellungen. Betrachten wir zunächst die erste Kategorie der inhaltlichen Fragestellungen. Das Bayes-Theorem und die Idee, daß Wahrscheinlichkeitsmaße unsicheres Wissen repräsentieren können, haben viele Forscher im Bereich der kognitiven Psychologie inspiriert. Einige der spezifischen Forschungsgebiete umfassen Informationsverarbeitung (Edwards, 1968), Entscheidungsprozesse (Slovic & Lichtenstein, 1971) und Einstellungsänderung (Fishbein & Ajzen, 1975). Das Problem der Spezifikation von Vorwissen, das ebenfalls hierzu gehört, wurde bereits in Abschnitt 3 besprochen.

Methodologische Anwendungen, die in engerem Bezug zum statistischen Fokus dieses Kapitels stehen, haben auch nicht gefehlt, insbesondere im Bereich psychologischen Testens. Das in Abschnitt 2 geschilderte Beispiel der Abschätzung individueller „wahrer" IQ-Werte ist ein klassischer Anwendungsfall des Bayes-Theorems, der zuerst von Kelley (1927) vorgeschlagen wurde. 1969 beschrieb Birnbaum die Bayessche Analyse eines individuellen Werts auf einem *latent trait* im Kontext der *Item-Response*-Theorie (IRT, vgl. dazu auch das Kapitel über *Latent-Trait*-Modelle von Roskam, in diesem Band). Van der Ven (1974) und Molenaar (1977) betrachteten das Problem der Ratekorrektur für einen *Multiple-Choice*-Test aus dem Blickwinkel der Bayesschen Theorie.

Eine Reihe von Forschern haben im Anschluß an Cronbach und Gleser (1965; vgl. dazu auch Kubinger, in diesem Band) Bayessche Methoden im Kontext statistischer Entscheidungstheorie angewendet, um Selektions- und Klassifikationsproblemen von Individuen in Ausbildungs-, klinischen oder Einstellungssituationen zu begegnen (Lewis & Sheehan, 1990).

Ferner gibt es ökonometrische oder psychometrische Modelle, in denen stichprobentheoretisch begründete Parameterschätzungen manchmal nicht auf zulässige und stabile Schätzungen hin konvergieren, während Bayes-Schätzungen dies eher leisten.

Die Heranziehung Bayesscher Statistik zum Einbezug persönlichen Vorwissens in die psychologische Forschung wurde erstmals von Edwards, Lindeman und Savage (1963) vorgeschlagen, fand aber wenig oder keine Zustimmung bei den Forschern auf diesem Gebiet.

Greenwald (1975) beschrieb eine Bayessche Varianzanalyse von Daten über sogenannte extrasensorische Wahrnehmungen, die demonstriert, wie Experimente kombiniert und so empirische Evidenz zugunsten der Nullhypothese akkumuliert werden kann.

Es ist eine unserer Hoffnungen, daß einführende Darstellungen wie diese (s. auch den Abschnitt über weiterführende Literatur) Psychologen mit dem Bayesschen Ansatz zur statistischen Inferenz besser vertraut machen und so zu einer Verbreitung des Ansatzes beitragen.

Für den Fall, daß die Apriori-Wahrscheinlichkeiten eine Häufigkeitsinterpretation zulassen, ist die Anwendbarkeit der Bayes-Methode nicht umstritten und hat zur Entwicklung einer schnell anwachsenden Zahl von Techniken und Anwendungen geführt. Zufallseffekte in Varianzanalysen haben inzwischen eine Geschichte von 50 Jahren. In den letzten Jahrzehnten wurden außer Gruppenmittelwerten auch noch viele andere Parameter als Zufallsvariablen behandelt, so etwa Regressionsgewichte und additive Konstanten in Regressionsproblemen (z.B. Lindley & Smith, 1972 oder Bryk & Raudenbush, 1992) oder Personenparameter, die eine Fähigkeit oder Einstellung im Rahmen der *Item-Response*-Theorie messen. Man kann tatsächlich argumentieren, daß jegliche Anwendung von *Random-Effects*-ANOVA, *Multilevel*-Methoden, empirischen Bayesschen Methoden, *Latent-Class*-Modellen (vgl. Langeheine & Rost, in diesem Band), des EM Algorithmus, des *Gibbs sampler*, von Mischverteilungsmodellen (vgl. Rost & Erdfelder, in diesem Band) oder *Marginal-Maximum-Likelihood*-Methoden darauf beruht, daß bestimmte Parameter als Realisationen von Zufallsvariablen aufgefaßt werden, so als wären sie aus einer Verteilung durch Zufallsstichprobenziehung gewonnen worden, statt sie als feste, unbekannte Größen zu behandeln.

In diesem Sinn hat in jüngster Zeit eine enorme gegenseitige Befruchtung von Bayesschen und klassischen Methoden stattgefunden, was auf dem Hintergrund der in der Vergangenheit nicht unüblichen bitteren Auseinandersetzungen zwischen Bayesianern und Frequentisten sehr erfreulich ist (z.B. Savage, 1954 und van Dantzig, 1957).

6 Weiterführende Literatur

Alse erste, sehr anschauliche Einführung in die Bayes-Statistik empfehlen wir das Buch von Iversen (1984). Winkler (1993) gibt einen hilfreichen Überblick über das Thema. Novick und Jackson (1974) weiten dies aus, indem sie einerseites mehr Details über andere univariate Modelle ausführen und andererseits zur zentralen Rolle des Vorwissens sowie zu Anwendungen in pädagogischer und psychologischer Forschung Stellung nehmen.

Raiffa (1968) gibt eine gute Einführung in Entscheidungsanalyse, Phillips (1973) eine relativ vollständige Einführung in das Gebiet der Statistik und der Messung für Sozialwissenschaftler aus Bayesscher Perspektive, und Pollard (1986) liefert eine sehr sorgfältige Einführung (allerdings mit einigen unsauberen Formeln) für eine begrenzte Anzahl von Modellen. Für diejenigen, die sich hauptsächlich für Varianzanalyse interessieren, ist Lewis (1993) eine gute Informationsquelle.

Alle genannten Bücher können mit Grundkenntnissen in Statistik, Wahrscheinlichkeitstheorie und Differentialrechnung gelesen werden.

Literaturverzeichnis

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418. (Wiederabgedruckt in *Biometrika, 45*, 1958, 293–315).

Birnbaum, A. (1969). Statistical theory for logistic mental test models with prior distribution of ability. *Journal of Mathematical Psychology, 6*, 258–276.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park: Sage.

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological tests and personal decisions.* Urbana: University of Illinois Press.

van Dantzig, D. (1957). Statistical priesthood (Savage on personal probabilities). *Statistica Neerlandica, 11*, 1–16.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgement* (pp. 17–53). New York: Wiley.

Edwards, W., Lindeman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193–242.

Einhorn, T. J. & Hogarth, R. M. (1981). Behavioral decision theory. *Annual Review of Psychology, 32*, 53–88.

Fishbein, M. & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading: Addison-Wesley.

Greenwald, A. G. (1975). Significance, nonsignificance, and interpretation of an ESP experiment. *Journal of Experimental Social Psychology, 11*, 180–191.

Hogarth, R. M. (1980). *Judgement and choice. The psychology of decision.* New York: Wiley.

Iversen, G. R. (1984). *Bayesian statistical inference.* Beverly Hills: Sage.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: World Book Company.

van Lenthe, J. (1993a). *ELI. The use of proper scoring rules for eliciting subjective probability distributions* (unpublished dissertation). Groningen: Rijksuniversiteit Groningen.

van Lenthe, J. (1993b). ELI. An interactive eliciation technique for subjective probability distributions. *Organizational Behavior and Human Decision Processes, 55*, 379–413.

van Lenthe, J. (1993c). A blueprint of ELI: A new method for eliciting subjective probability distributions. *Behavior Research Methods, Instruments, & Computers, 25*, 425–433.

van Lenthe, J. & Molenaar, I. W. (1993). *ELI. ELIcitation of uncertain knowledge. Preliminary Manual.* Groningen: iec ProGAMMA.

van Lenthe, J. (1994). Scoring-rule feedforward and the eliciting of subjective probability distributions. *Organizational Behavior and Human Decision Processes, 59*, 188–209.

Lewis, C. (1993). Bayesian methods for analysis of variance. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Volume II - Statistical issues* (pp. 233–256). Hillsdale: Erlbaum.

Lewis, C. & Sheehan, K. M. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386.

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. *The Journal of the Royal Statistical Society, B 34*, 1–41.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Lourens, P. F. (1984). *The formalization of knowledge by specification of subjective probability distributions: An experimental approach* (unpublished dissertation). Groningen: Rijksuniversiteit Groningen.

Molenaar, I. W. (1977). On Bayesian formula scores for random guessing in multiple choice tests. *British Journal of Mathematical and Statistical Psychology, 30*, 79–89.

Novick, M. R. & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Peterson, C. R. & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29–46.

Phillips, L. D. (1973). *Bayesian statistics for social scientists.* London: Whitefriars Press.

Pollard, W. E. (1986). *Bayesian statistics for evaluation research. An introduction.* Beverly Hills: Sage.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty.* Reading: Addison-Wesley.

Savage, L. J. (1954). *The foundation of statistics.* New York: Wiley.

Schütt, K. P. (1981). *Wahrscheinlichkeitsschätzungen im Computer-Dialog: Theorie, Methoden und eine experimentelle Studie zur Schätzung von subjektiven Wahrscheinlichkeiten.* Stuttgart: Pöschel.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior and Human Performance, 6*, 649–744.

Terlouw, P. (1989). *Subjective probability distributions, a psychometric approach* (unpublished dissertation). Groningen: Rijksuniversiteit Groningen.

Tversky, A. (1974). Assessing uncertainty. *Journal of the Royal Statistical Society, B 36*, 148–159.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review, 90*, 293–315.

van der Ven, A. H. G. S. (1974). A Bayesian formula score for the simple knowledge or random guessing model. *Nederlands Tijdschrift voor de Psychologie, 29*, 409–414.

Winkler, R. L. (1993). Bayesian statistics: An overview. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Volume II - Statistical issues* (pp. 201–232). Hillsdale: Erlbaum.