Sonderdruck aus

Computer Aided Sociological Research

Herausgegeben von
Klaus Troitzsch und Johannes Gladitz

Proceedings of a Workshop Organized by the Research Committee 33 of the International Sociological Association and the Acadeny of Sciences of the GDR, Held at Holzhau, GDR October 2nd to 6th, 1989

Edited by Johannes Gladitz and Klaus G. Troitzsch

Akademie-Verlag Berlin 1990

Akademie - Verlag Berlin

Kapitel 8

Johannes Gladitz, Berlin/DDR: Zur Anwendung Bayes'scher Methoden in den Sozialwissenschaften

Abstract

The most prejudices against the Bayesian methods are connected with the choice of the prior distribution. The crucial point here is that this choice affects the results of inference. The aim of the article is to present exemplarily three methods for an objectivated choice of a prior distribution: the Delphi method, the fiducial argument, and the empirical Bayes approach. The paper is mainly directed to social science methodologists.

8.1 Einleitung

Begünstigt durch die technischen Möglichkeiten, hat sich weltweit die Überzeugung durchgesetzt, daß man Daten jeglicher Art systematisieren, strukturieren, speichern und wiederauffindbar archivieren muß.

Auch im Bereich der Sozialwissenschaften sind große Datenbanken entstanden, die z.T. miteinander vernetzt und in internationalen Gesellschaften organisiert ihre Daten untereinander austauschen. Wenn sich die Investition großer Datenarchive auszahlen soll, müssen parallel dazu Anstrengungen zur breiteren Nutzung dieser Daten verstärkt werden.

Eine Möglichkeit, in früher oder andernorts erhobenen Daten gespeicherte Information geeignet quantifiziert in aktuellen Modellen oder Entscheidungen zu berücksichtigen, bietet der Bayessche Zugang. Die Anwender stehen diesem Vorgehen bisher eher skeptisch gegenüber. Ursachen dafür liegen im höheren Aufwand für Modellierung und Numerik sowie in gewissen Vorurteilen begründet. Die weitere Verbreitung matrixorientierter Hochsprachen wie APL oder GAUSS auf leistungsstarken Mikrorechnern hilft die objektiven Probleme zunehmend zu überwinden. Existierende Vorurteile vor allem im deutschen Sprachraum abzubauen, soll dieser Beitrag helfen.

8.2 Der Bayes'sche Zugang

"The thesis behind this talk is very simple: the only good statistics is Bayesian statistics. Bayesian statistics is not just another technique to be added to our repertoire alongside, for example, multivariate analysis; it is the only method that can produce sound inferences and decisions in multivariate, or any other branch of statistics." [Lin75]

Diesen Ausschließlichkeitsanspruch vertritt der Autor des vorliegenden Textes nicht. Er verwendet den Begriff „Bayes'sch" nicht für eine Heuristik, die die Wahrscheinlichkeit ausschließlich mit dem subjektiven Überzeugungsgrad eines rationalen Individuums identifiziert (siehe z.B. Kleiter [Kle81]). Ausgehend vom pragmatischen Standpunkt eines angewandten Statistikers, versteht er darunter eine Gruppe von Techniken, die mit mehreren heuristischen Grundprinzipien verbunden ist. Allen gemeinsam ist, daß zum Zusammenführen verschiedener Informationen die Bayes'sche Formel verwendet wird. Informationen sind hierbei im Sinne von Shannon [Sha48] Wahrscheinlichkeitsverteilungen über bestimmten Möglichkeitsfeldern. Sie können einen objektiv-frequentistischen oder einen subjektiven Ursprung haben, oder aus strukturlogischen Überlegungen herrühren. Wir werden im folgenden die einzelnen Möglichkeiten andeutungsweise vorstellen.

8.2.1 Die Bayes'sche Formel

Es sei $A_1, A_2, \ldots, A_n$ ein vollständiges System disjunkter zufälliger Ereignisse, d.h. A_i und A_j mit $i \neq j$ können nicht gemeinsam eintreten und eines der A_i tritt immer ein. Weiter sei B ein beliebiges zufälliges Ereignis. Dann kann man die Wahrscheinlichkeit von B folgendermaßen berechnen:

$$P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i) \tag{8.1}$$

(Satz von der totalen Wahrscheinlichkeit).

Hierbei bezeichnet $P(B|A_i)$ die (bedingte) Wahrscheinlichkeit von B unter der Bedingung, daß A_i eingetreten ist. Wenn außerdem $P(B) > 0$, so berechnet man die bedingte Wahrscheinlichkeit der A_i unter der Bedingung B nach

$$P(A_j \wedge B) = \frac{P(A_j|B)}{P(B)}$$

wobei

$$P(A_j \wedge B) = P(B|A_j) \cdot P(A_j)$$

Unter Verwendung von 8.1 erhalten wir

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{\sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)} \qquad j = 1, \ldots, n \tag{8.2}$$

Dieser Zusammenhang geht auf den englischen Geistlichen Thomas Bayes (1763) zurück und wird als Bayes'sche Formel bezeichnet. Sie bildet den Ausgangspunkt eines wichtigen Prinzips der induktiven Inferenz. Wir betrachten zwei Spezialfälle:

a. Diskreter Fall Angenommen, wir untersuchen zwei am selben Objekt gemessene diskrete Merkmale (X, Y), wobei X die Ausprägungen $k = 1, \ldots, r$ und Y die Ausprägungen $l = 1, \ldots, s$ habe. Dann kann man mit Hilfe von 8.2 z.B. die bedingte Verteilung von Y berechnen, wenn bekannt ist, daß $X = k$ vorliegt, nämlich durch

$$P(Y = l|X = k) = \frac{P(X = k|Y = l) \cdot P(Y = l)}{\sum_{j=1}^{s} P(X = k|Y = j) \cdot P(Y = j)} \qquad l = 1, \ldots, s$$

Man stelle sich etwa vor, daß nur X direkt beobachtbar (manifest) ist. Nachdem X beobachtet wurde, möchte man eine Aussage über die zugehörige Realisierung von Y (latent) treffen, wobei eine *a priori*-Bewertung der Ausprägungen Y sowie die Wahrscheinlichkeiten für die einzelnen Realisierungen von X bei gegebenem Y bekannt seien. Die Entscheidung trifft man auf der Grundlage der sogenannten *a posteriori*-Wahrscheinlichkeiten $P(Y = l|X = k)$. Die Wahrscheinlichkeitsverteilung $P(Y = l), l = 1, \ldots, s$, heißt in diesem Zusammenhang *a priori*-Verteilung. Die *a priori*-Verteilung charakterisiert das Wissen über Y vor dem Experiment (z.B. vor der Befragung). Die *a posteriori*-Verteilung beschreibt unsere Information über Y nach dem Experiment. Die Differenz zwischen beiden Verteilungen (in einer entsprechenden Metrik) ist geeignet, den Informationszuwachs über die aktuelle Ausprägung des Merkmals Y durch das Experiment zu messen (s. Kullback [Kul59]). Hier würde man sich für dasjenige $Y = l^*$ mit der höchsten *a posteriori*-Wahrscheinlichkeit entscheiden, d.h. für den Modalwert der *a posteriori*-Verteilung. Natürlich ist auch eine randomisierte Entscheidung entsprechend der gegebenen *a posteriori*-Wahrscheinlichkeiten möglich oder aber die Wahl eines *a posteriori*-Konfidenzintervalls. Jegliche Inferenz basiert also hier auf der *a posteriori*-Verteilung.

b. Stetiger Fall Ganz analog ist das Vorgehen, wenn (X,Y) ein kontinuierlich verteiltes Merkmalspaar darstellt. Dann läßt sich eine entsprechende Formel für Dichten herleiten:

$$g(y|x) = \frac{f(x|y)g(y)}{\int_{\boldsymbol{R}^1} f(x|y)g(y)dy} \tag{8.3}$$

wobei $g(y)$ die Dichte zur Randverteilung von Y, $f(x|y)$ die Dichte zur bedingten Verteilung von X bei festgehaltenem $Y = y$ bedeuten. Auch hier wollen wir ein Beispiel betrachten:

Die Variable Y sei ein unbekannter Populations- oder Verteilungsparameter θ, über dessen Lage wir eine gewisse Vorstellung haben, den wir jedoch nicht genau kennen; d.h. θ erscheint uns als (subjektiv) zufällig. Unser Wissen über θ vor dem Experiment sei gegeben in Form einer Wahrscheinlichkeitsdichte $g(\theta)$, die eine *a priori*-Bewertung aller möglichen Parameterwerte darstellt. Weiter nehmen wir an, daß die Struktur von $f(x|\theta)$ für jedes mögliche θ bekannt sei, was oft angenommen wird. Mit 8.3 erhalten wir dann mit $g(\theta|x)$ eine neue Bewertung der möglichen θ unter Berücksichtigung des Ergebnisses des Experimentes $X = x$. Hierbei heißt $g(\theta|x)$ Dichte zur *a posteriori*-Verteilung von θ bei beobachtetem $X = x$. Als Punktschätzung für θ könnte man z.B. den Mittelwert bzgl. $g(\theta|x)$ empfehlen (*a posteriori*-Erwartungwert $E\langle\theta|x\rangle$), den *a posteriori*-Modalwert oder den *a posteriori*-Median. Als Bereichsschätzungen werden HPD-Intervalle oder HPD-Bereiche (high posterior density) verwendet, die ein Gebilde möglichst kleiner Fläche im Parameterraum darstellen, das eine vorgebene *a posteriori*-Wahrscheinlichkeit (z.B. 0.95) nicht unterschreitet. Auf Grund des bestehenden Zusammenhangs zwischen Konfidenzbereichen und Annahmebereichen statistischer Tests ist damit auch ein Bayes'scher Zugang zur Testtheorie beschrieben.

Die oben erwähnte subjektive Wahl der *a priori*-Verteilung $g(\theta)$ ist ein Grund für Vorbehalte vieler Praktiker. Wir werden später einige Möglichkeiten zur Objektivierung der Wahl von $g(\theta)$ kennenlernen. Die Beschränkung auf (eindimensionale) Zufallsvariablen X und Y erfolgte nur zur besseren Darstellung. Alle Aussagen gelten sinngemäß für Zufallsvektoren $\boldsymbol{X}$ und $\boldsymbol{Y}$.

Da die verwendeten Verteilungen (d.h. die Wahrscheinlichkeitsmodelle) meist nichts Natur- oder Gottgegebenes sind, sondern es in der Regel mehrere (meist unendlich viele) Verteilungen gibt, die etwa unser Vorwissen hinreichend adäquat abbilden, verwendet man bevorzugt solche Verteilungsklassen, die sich „gutartig“ zueinander verhalten. Das sind konkret sogenannte „konjugierte“ Verteilungsfamilien, die dergestalt sind, daß *a priori*- und *a posteriori*-Verteilung vom gleichen Verteilungstypus sind (siehe z.B. Raiffa und Schlaifer [RS61]).

8.3 Zur Wahl der *a priori*-Verteilung

Bevor wir auf einige exemplarische Anwendungen eingehen, wollen wir einige Möglichkeiten zur Spezifizierung der *a priori*-Verteilungen betrachten, die über Grundideen hier nicht hinausgehen können.

a. Subjektive *a priori*-Bewertungen Im wissenschaftstheoretischen Verständnis des Begriffes der Wahrscheinlichkeit kann man zwei Kulturkreise unterscheiden: die Objektivisten oder Frequentisten und die Subjektivisten. Die extremen Flügel beider Lager stehen sich bei Betonung des Alleinwahrheitsanspruches nahezu feindlich gegenüber. Für die Subjektivisten sind Wahrscheinlichkeiten identisch mit den Überzeugungsgraden eines rationalen Individuums, während für die Objektivisten Wahrscheinlichkeiten Grenzwerte relativer Häufigkeiten sind, die unabhängig von unserem Bewußtsein existieren und approximativ durch das Experiment empirisch bestimmt werden können.

Nun ist man leicht geneigt, sich bedingungslos dem Lager der Frequentisten anzuschließen. Dieser Entschluß wäre voreilig. Die meisten Entscheidungen in unserem Leben treffen wir aufgrund mehr oder weniger subjektiver Bewertungen erwarteter Bedingungen, Ereignisse oder Wirkungen. Wir versuchen unter Einbeziehung aller Informationen zum Sachverhalt — eingeschlossen Erfahrungen, das Wissen um Gesetzmäßigkeiten oder programmatische Zielvorstellungen — mit Logik und Intuition eine optimale Entscheidung zu finden. Trotz alledem wird die Entscheidung, von unterschiedlichen Personen getroffen, verschieden ausfallen. Auch die Präsenz kollektiver Beratungsorgane löst das Problem nicht vollständig, da bekannt ist, wie prägend dominante Persönlichkeiten Kollektivmeinungen beeinflussen können. Der Überwindung dieses Phänomens sollte ja bekanntlich die Entwicklung der Delphi-Technik dienen (siehe z.B. Helmer [Hel67]).

Andererseits gibt es gerade im Bereich der Sozialwissenschaften viele Ereignisse, deren Auftretenswahrscheinlichkeit einer frequentistischen Interpretation kaum zugänglich ist. Die Geschichte läuft einmalig und unwiederholbar ab. Selbst bei teilweiser Wiederholbarkeit oder Analogiebetrachtung ist eine Vorhersage gesellschaftlicher Ereignisse ausschließlich aufgrund der Abschätzung der relativen Häufigkeit bestimmter Ereigniskonstellationen nahezu unmöglich. Beim hohen Komplexitäts- und Verflechtungsgrad gesellschaftlicher Prozesse müßten hochdimensionale Merkmalsvektoren betrachtet werden, deren multivariate Verteilung abzuschätzen wäre. Das ist empirisch auch bei niedrigerdimensionalen Merkmalsvektoren aufgrund der mit der Dimension schnell wachsenden Zahl der zu identifizierenden Parameter kaum möglich. Eine ähnliche Situation hat man in der Zuverlässigkeitstheorie bei der Beurteilung der Auftretenswahrscheinlichkeit sehr seltener Ereignisse im Zusammenhang mit möglichen Schadensfällen bei komplexen technischen Systemen.

Die Techniken zum Aufstellen von subjektiven Wahrscheinlichkeitsverteilungen haben einerseits das Ziel, eine Expertenbewertung möglichst adäquat zu quantifizieren durch eine Abbildung der Vorstellungen zum betrachteten Sachverhalt in den Raum der Wahrscheinlichkeitsverteilungen und andererseits durch eine Zusammenfassung mehrerer subjektiver Verteilungen eine resultierende „objektivierte“ subjektive Verteilung zu finden.

Das Gesamtverfahren läßt sich nicht vollständig formalisieren. Einige Beispiele sind u.a. bei L.J. Savage [Sav54] und R.L. Winkler [Win67] beschrieben (siehe auch Sullivan und Claycombe [SC77]). Die Ermittlung der Wahrscheinlichkeiten erfolgt oft mit Hilfe eines Fragekataloges, der den Abbildungsprozeß auf eine quantitative Skala unterstützen soll. Der Experte selbst benötigt keine speziellen Kenntnisse auf dem Gebiet der Wahrscheinlichkeitsrechnung. Für die zu vergebenden Wahrscheinlichkeiten muß lediglich gelten

$$0 \leq P(E_i), i = 1, 2, \ldots, k, \qquad \text{und}$$
$$\sum_{i=1}^{k} P(E_i) = 1$$

wobei die E_i alle möglichen Ereignisse sind. Handelt es sich um ein metrisches Merkmal, dessen Eintreffen bewertet werden soll, (z.B. die Anzahl der Arbeitslosen in Osteuropa im Jahre 1990), so geht man von der kumulativen Verteilungsfunktion aus und fragt umgangssprachlich verbal verpackt bestimmte charakteristische Prozentpunkte (Quantile) der Verteilung ab, die als Stützstellen für eine anschließende Glättung dienen können (siehe Sullivan und Claycombe [SC77]). Oft findet in einem zweiten Schritt eine Rückkopplung zu den Ergebnissen der anderen Experten statt und eine anschließende Korrektur. Die Zusammenfassung aller Bewertungen zu einer gemeinsamen kann gleichberechtigt oder gewichtet etwa nach der Trefferquote früherer Bewertungen erfolgen.

Die Ideen zu derartigen Techniken sind heute über 40 Jahre alt und durchaus nicht unangefochten geblieben. Im Lichte der modernen Informations- und Kommunikationstechnologien stellen sich jedoch viele Fragen neu. Manches, was früher nur mit großem Aufwand realisierbar war und deshalb über Beispielanwendungen nicht hinauskam, ist heute bei einem vertretbaren Aufwand und kurzfristig machbar.

Derartige Expertenbewertungen (*a priori*-Bewertungen) können dann über die Bayes'sche Formel mit empirischen Befunden konfrontiert werden, die dann zu einer Neubewertung (*a posteriori*-Bewertung) des Sachverhaltes führen. Diese übernehmen dann im weiteren Verlauf die Rolle der *a priori*-Verteilung, die bei Vorliegen neuer Tatsachen über die Bayes'sche Formel weiter korrigiert werden. Damit ist das Bayes'sche Herangehen gut geeignet, Lernvorgänge bzw.

adaptive Steuerprozesse und Informationssysteme zu modellieren (siehe Repin und Tartakowski [RT77]).

b. Der Fiduzialzugang Während der subjektive Zugang ohne spezielle Verteilungsannahmen auskam, setzt der Fiduzialzugang — ähnlich dem Maximum-Likelikood-Prinzip — die Kenntnis der Struktur der Strichprobenverteilung oder gewisser anderer struktureller Zusammenhänge voraus. Derartige Annahmen ergeben sich oft aus struktur-logischen oder inhaltlich-fachlichen Überlegungen. Ausgehend von einem bekannten strukturellen Zusammenhang zwischen einer Zufallsgröße X (z.B. dem Stichprobenvektor, der die Messung eines Merkmals an n Individuen symbolisiert) und einem unbekannten Parameter θ wird nach Beobachtung von $X = x$ vermittels des Strukturzusammenhangs das Möglichkeitsfeld für θ mehr oder weniger eingeschränkt. Die sich ergebende Verteilung nennt man strukturelle Parameterverteilung oder Fiduzialverteilung (*fiducia* = Vertrauen). Die Grundidee dieses Zugangs stammt bereits von Ronald A. Fisher [Fis35a]. Wesentlich vorangebracht wurde diese Theorie in den sechziger Jahren von D.A.S. Fraser [Fra61, Fra68]. Einer breiten Anwendung standen bisher die auftretenden numerischen Schwierigkeiten entgegen.

Bekannt geworden ist der Fiduzialzugang im Zusammenhang mit dem Maximum-Likelihood-Prinzip. Es sei $X_1, X_2, \ldots, X_N$ eine Stichprobe zu einem gegebenen Merkmalsvektor. Zur besseren Darstellung nehmen wir an, die X_i seien stetig verteilt mit einer multivariaten Dichtefunktion $f(x_i|\theta)$, die von einem Parametervektor $\theta \in \Theta$ abhängen möge. Die X_i können auch diskret oder gemischt sein. Dann erhalten wir die Dichte zur Stichprobenverteilung für die Beobachtungs- oder Profilmatrix mit

$$\prod_{i=1}^{N} f(x_i|\theta) = L(\theta|x_1, \ldots, x_N)$$

die für beobachtetes $X_1 = x_1, \ldots, X_N = x_N$ eine Funktion des unbekannten Parameters $\theta \in \Theta$ darstellt, die sogenannte Likelihoodfunktion nämlich. Nun wählt man im Nachhinein den unbekannten Parameter $\theta \in \Theta$ so, daß den beobachteten $x_1, \ldots, x_N$ eine hohe Plausibiltät zukommt, d.h. man wählt θ so, daß $L(\theta|x_1, \ldots, x_N)$ maximal wird. Das ist die Heuristik des Maximum-Likelihood-Prinzips.

Durch eine ausschließliche Beschränkung auf das Maximum verschenkt man jedoch Informationen. Mit der Likelihoodfunktion wird ja eine Bewertung über den gesamten Parameterraum induziert, die die Paßfähigkeit aller möglichen Parameterkonstellationen in Bezug auf die vorliegende Beobachtungsmatrix mißt. Gelingt es, die Likelihoodfunktion zu normieren, d.h. ist

$$\int_\Theta L(\theta|x_1, \ldots, x_N) d\theta < \infty$$

so erhalten wir mit

$$g(\theta|x_1,\ldots,x_N) = \frac{L(\theta|x_1,\ldots,x_N)}{\int_\Theta L(\theta|x_1,\ldots,x_N)d\theta}$$

formal eine Wahrscheinlichkeitsdichte über dem Parameterraum Θ, die man als Fiduzial- oder strukturelle Parameterverteilung bezeichnet. Diese drückt unser quantifiziertens Wissen nach dem Experiment aus. Bei Fortsetzung der Versuchsreihe kann diese Verteilung als *a priori*-Verteilung verwendet werden. Oft kann man zeigen, daß die Fiduzialverteilung gerade die *a posteriori*-Verteilung zu einer nichtinformativen *a priori*-Verteilung ist (siehe z.B. Gladitz und Willuhn [GW83]). Nichtinformative *a priori*-Verteilungen sind solche, die jedem möglichen Wert im Parameterraum die gleiche Chance einräumen.

Schließlich sei darauf verwiesen, daß es einen engen natürlichen Zusammenhang zwischen Fiduzialzugang, Konfidenzbetrachtungen und statistischen Tests gibt. Einen gewissen Eindruck hierzu vermittelt die zuletzt erwähnte Arbeit.

c. Der empirisch Bayes'sche Zugang Abschließend betrachten wir einen Zugang, der auf dem Konzept der relativen Häufigkeit basiert, jedoch einige spezielle Voraussetzungen verlangt. Seit der Begründung dieser Technik durch Herbert Robbins [Rob55, Rob64] hat die empirisch Bayes'sche Methode viele interessante Anwendungen gefunden, so in der Zuverlässigkeitstheorie, in der Meteorologie, im Nachrichtenwesen, in der statistischen Qualitätskontrolle, im Versicherungswesen, in der Prozeßsteuerung, in der Psychologie, im Bildungswesen u.a.m. Eine gute methodische Übersicht ist in der Monographie von Johannes S. Maritz und T. Lwin [ML89] gegeben.

Im Unterschied zu den vorher beschriebenen Methoden zur Festlegung von *a priori*-Verteilungen wird hier angenommen, daß der zu identifizierende Parameter tatsächlich (frequentistisch) zufällig ist und z.B. gruppen-, chargen- oder individualitätsspezifische Ausprägungen annimmt.

Formell läßt sich das Problem folgendermaßen beschreiben:

Es sei

$$(X_1,\theta_1),(X_2,\theta_2),\ldots,(X_N,\theta_N),(X_{N+1},\theta_{N+1})$$

eine Folge von unabhängigen, identisch verteilten Zufallsgrößen. Die Parameter θ_i realisieren sich mit den X_i, bleiben jedoch unbekannt. Aufgrund der stochastischen Abhängigkeit zwischen X_i und θ_i enthalten die Beobachtungen $x_i, i = 1,\ldots,N$, auch (unscharfe) Informationen über die Realisierungen der θ_i, die verwendet werden können für eine Schätzung der *a priori*-Verteilung oder gewisser Parameter derselben. Bei der aktuellen Entscheidung über — sagen wir — θ_{N+1} wird nicht nur $X_{N+1} = x_{N+1}$ verwendet, sondern auch gewisse Informationen über die Verteilung der θ_i, die man aus $x_1,\ldots,x_N$

gewinnt. Dabei wird im Entscheidungsprozeß die zusätzliche Unschärfe, die aus der Schätzung der unbekannten *a priori*-Verteilung resultiert, quantitativ berücksichtigt. Hierbei ist die Schätzung der in der Regel multivariaten *a priori*-Verteilung kein einfaches Problem. Es gibt jedoch viele Beispiele, wo lediglich die Schätzung einiger Parameter der unbekannten *a priori*-Verteilung erforderlich ist, die man aus Parametern der Randverteilung der X_i gewinnen kann, so z.B. bei einem sogenannten Linear Bayes'schen Ansatz (siehe Hartigan [Har69], Bunke und Gladitz [BG74]), wo nur eine Identifikation der ersten beiden Momente der *a priori*-Verteilung nötig ist. Diese Methoden haben eine besonders breite Anwendung gefunden (siehe Hoffmann [Hof86], Norberg [Nor80]). Sie verbessern die Identifizierbarkeit aktueller Parameter und wirken sich günstig auf die numerische Stabilität von Berechnungsalgorithmen aus. Entscheidungsverbessernd wirken diese Verfahren vor allem dann, wenn die verfügbare Information aus $X_{N+1} = x_{N+1}$ nicht ausreicht, θ_{N+1} zu identifizieren bzw. eine Identifikation nur mit großer Unschärfe möglich ist.

Einschränkend muß man erwähnen, daß die Anwendung empirisch Bayes'scher Methoden eine Stabilität der *a priori*-Verteilung in der Zeit voraussetzt. Auch hier sind jedoch Verallgemeinerungen möglich, indem man etwa den älteren Daten einen Vergessensprozeß überlagert und die $X_i = x_i, i = 1, 2, \ldots,$ gewichtet für die Schätzung der *a priori*-Parameter verwendet.

8.4 Einige Beispiele mit sozialwissenschaftlicher Relevanz

Die folgenden Beispiele sollen einige Denkanstöße vermitteln, die die vielfältigen Einsatzmöglichkeiten Bayes'scher Techniken plastisch machen sollen. Auch hier müssen wir uns auf Grundideen beschränken.

8.4.1 Vorhersage von seltenen Ereignissen mit hoher Konsequenz

Nehmen wir an, wir wollen den Zeitpunkt des Eintreffens eines seltenen Ereignisses vorhersagen. Genauer gesagt wollen wir die Prädiktionsverteilung (oder Parameter derselben) für den Auftretenszeitpunkt bestimmen. Unter Verwendung der Delphi-Technik wird eine Gruppe von unabhängigen Experten mit allen verfügbaren Informationen versorgt und in der üblichen Weise in mehreren Runden befragt. Nach jeder Runde wird jeder einzelne (unabhängig voneinander) mit dem aggregierten Ergebnis konfrontiert. So tastet man sich schrittweise an eine Wahrscheinlichkeitsverteilung für das Auftreten des seltenen Ereignisses heran, die das aggregierte Wissen der Expertengruppe zum betrachteten Problem beinhaltet. Wird die Zeit in diskreten Einheiten gemessen (z.B. Monate oder Jahre), so erhält man also eine Verteilung $P(t_i) = p_i, i = 1, \ldots, k$. Nachdem dieser Prozeß abgeschlossen ist, gelangt man an neues Wissen, z.B. Informationen über „Beinahestörfälle", Ereignisse, die nach der Expertenbefragung auftreten oder bekannt werden und einen Bezug zum interessierenden seltenen Ereignis haben, empirische Daten aus vergleichbaren Situationen o. a. Angenommen, es handelt sich um ein Ereignis A. Nun bestimmt man die Wahrscheinlichkeit für das Auftreten von A, wenn das interessierende Ereignis tatsächlich zum Zeitpunkt t_i auftreten würde, und zwar für alle $i = 1, \ldots, k$. Auch hierzu kann man die Meinung von Experten (möglichst anderer) einholen.

Man erhält $P(A|t_i), i = 1, \ldots, k$. Mit der Bayes'schen Formel ergibt sich nun eine Neubewertung für die Auftretenswahrscheinlichkeiten

$$P(t_i|A) = \frac{P(A|t_i)P(t_i)}{\sum_{i=1}^{k} P(A|t_i)P(t_i)}$$

die als *a posteriori*-Wahrscheinlichkeiten unter Berücksichtigung des Ereignisses A betrachtet werden können. Damit kann man nun eine Punktvorhersage etwa unter Verwendung des wahrscheinlichsten Zeitpunktes oder des *a posteriori*-Erwartungswertes

$$\hat{t} = \sum_{i=1}^{k} t_i P(t_i|A)$$

oder aber ein Prädiktionsintervall z.B. als 95 %-HPD-Intervall konstruieren. Eine detaillierte Diskussion dieses Vorgehens, angewendet auf die Vorhersage eines Flugzeugunglückes über einem dicht besiedelten Gebiet, findet man im Buch von Sullivan und Claycombe [SC77].

8.4.2 Beurteilung der Präzision von Hochrechnungen

Hochrechnungen gehören zum klassischen Handwerk der Sozialwissenschaften. Brauchbar sind sie dann, wenn etwa die geschätzten Prozentwerte ergänzt werden durch ihre Präzision z.B. in Form von Konfidenzintervallen. Bei Berücksichtigung von Zusatzinformationen kann oft diese Präzision verbessert werden. Wir betrachten ein Beispiel von Stange [Sta77]:

Angenommen, wir beabsichtigen, einen Vorschlag einem Gremium von 100 Mitgliedern zur Entscheidung vorzulegen. Der Vorschlag wird akzeptiert, wenn mindestens 51 Mitglieder dafür stimmen. Zur Sicherheit wählt man zufällig 10 Personen aus und befragt sie nach ihrer Meinung. Es seien 6 dafür und 4 dagegen. Gibt uns diese Stichprobe genügend Sicherheit, um mit gutem Gewissen in die Abstimmung zu gehen?

Derartige Fragen sind folgendem Versuchsaufbau äquivalent:

Wir haben eine Urne mit N Kugeln, von denen eine unbekannte Anzahl N_1 schwarz, der Rest $N - N_1$ weiß sei. Zufällig entnehmen wir der Urne eine Stichprobe von n Kugeln mit n_1 schwarzen und $n - n_1$ weißen. Aus kombinatorischen Überlegungen berechnet man die Wahrscheinlichkeit für n_1 nach der bekannten hypergeometrischen Verteilung

$$P_n^N(n_1|N_1) = \frac{\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}{\binom{N}{n}} \tag{8.4}$$

deren Anwendung allerdings die Kenntnis von N_1 voraussetzen würde. Intuitiv könnte man als Schätzung für N_1 die Größe $\hat{N}_1 = \frac{n_1}{n}N$ verwenden; d.h. im obigen Beispiel würden wir mit 60 Befürwortern rechnen. Wie zuverlässig ist jedoch dieser Wert?

Dazu betrachten wir 8.4 als Funktion von N_1. Nach der Ziehung ist n_1 bekannt und man könnte sich interessieren für die Wahrscheinlichkeit von n_1, wenn $N_1 = i$ wäre für verschiedene Werte von i. Nun nimmt man weiter an, daß, wenn man n_1 beobachtet hat, dieses nicht ganz unwahrscheinlich sein kann, d.h. man ordnet entsprechend dem Likelihoodprinzip denjenigen Werten von N_1 die zu einer höheren Wahrscheinlichkeit von n_1 führen, eine höhere Plausibilität zu. Die Funktion 8.4 heißt für festgehaltenes n_1 dann Likelihoodfunktion

$$L_n^N(N_1|n_1) := P_n^N(N_1|n_1)$$

Sie ist nicht notwendig normiert, beschreibt aber unterschiedliche Plausibilitätsgrade für N_1 bei beobachtetem n_1. Normierung von L_n führt dann zur Fiduzialverteilung von N_1:

$$F_n^N(N_1|n_1) = \frac{L_n^N(N_1|n_1)}{\sum_{N_1=n_1}^{N-n+n_1} L_n^N(N_1|n_1)} = \frac{\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}{\binom{N+1}{n+1}}$$

die formell eine Wahrscheinlichkeitsverteilung darstellt und sich aus der Struktur der Likelihoodfunktion ergibt.

Für unser Beispiel nimmt sie die in Abb. 8.1 dargestellte Form an.

Damit bekommt man einen optischen Eindruck vom Wissen über N_1 nach dem Experiment. Die Fiduzialwahrscheinlichkeit dafür, daß $N_1 \leq 50$ ist, erhält man als

$$\sum_{N_1=6}^{50} F_{10}^{100}(N_1|n_1 = 6) = 0.274 \quad \hat{=} \quad 27.4\ \%$$

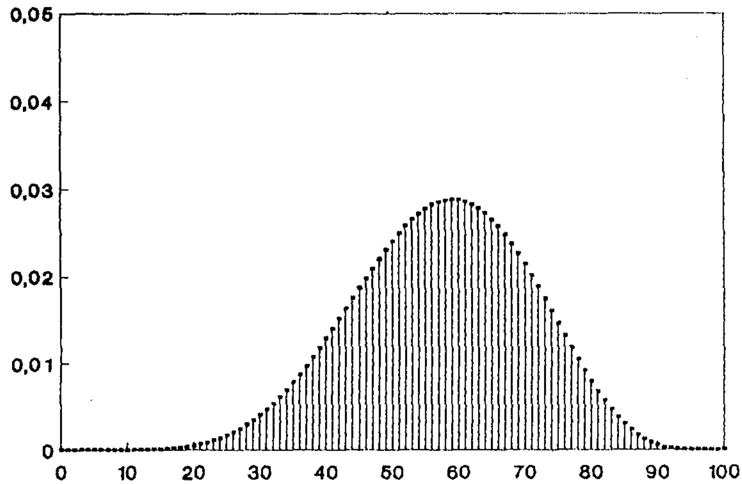

Abbildung 8.1: Die Fiduzialverteilung $F_{10}^{100}(N_1|n_1 = 6)$

Damit können wir unsere Unsicherheit quantifizieren. Wie oben erwähnt, ergibt sich die Fiduzialverteilung auch als *a posteriori*-Verteilung in einem Bayes'schen Zugang bei nichtinformativer *a priori*-Verteilung $P(N_1) = 1/(N+1)$. Vom Typ her ist die Fiduzialverteilung eine negative hypergeometrische Verteilung, die allgemein darstellbar ist durch

$$P_\theta(N_1) = \frac{\binom{N_1+a-1}{a-1}\binom{N-N_1+b-1}{b-1}}{\binom{N+a+b-1}{a+b-1}}, \qquad \theta = (a, b, N)$$

wobei a und b natürliche Zahlen größer als 0 seien. Im Falle der Fiduzialverteilung sind also $a := 1+n_1$, $b := 1+n-n_1$, $N := N-n$ und $N_1 := N_1-n_1$. Auch die erwähnte nichtinformative Verteilung ergibt sich als negativ hypergeometrische mit $a = b = 1$. Allgemein kann man sogar zeigen, daß hypergeometrische und negativ hypergeometrische Verteilung ein konjugiertes Verteilungspaar sind, d.h. *a priori*-und a posteriori Verteilung sind vom gleichen Verteilungstyp (s. Gladitz und Spieß [GS79]).

Doch kehren wir zu unserem Beispiel zurück. Angenommen, die Unsicherheit von 27,4 % erscheint uns noch zu hoch, so müßten wir weitere zufällig herausgegriffene Mitglieder des o.g. Gremiums nach ihrer Meinung zur erwähnten Vorlage befragen. Dazu ziehen wir eine zweite Stichprobe vom Umfang 10 und erhalten diesmal 7 Befürworter. Wie ist jetzt die verbleibende Unsicherheit zu bewerten? Dazu verwenden wir die nach der ersten Probe erhaltene Fiduzialverteilung als *a priori*-Verteilung für den 2. Schritt und erhalten als *a*

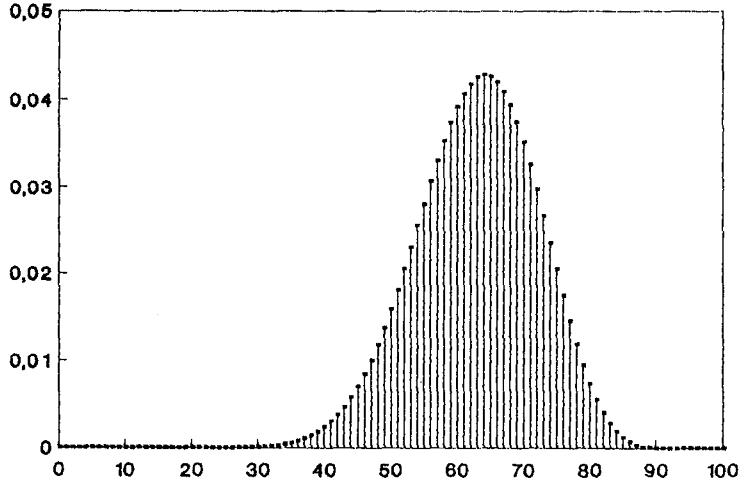

Abbildung 8.2: Die *a posteriori*-Verteilung $P_\theta(N_1|n_1 = 7)$

posteriori-Verteilung bei einem Bayes'schen Vorgehen (vgl. Abb. 8.2)

$$P_\theta(N_1|n_1) = \frac{P_{10}^{90}(7|N_1)F_{10}^{100}(N_1|n_1 = 6)}{\sum_{N_1=7}^{87} P_{10}^{90}(7|N_1)F_{10}^{100}(N_1|n_1 = 6)}$$

$$= \frac{\begin{pmatrix} N_1+6 \\ 13 \end{pmatrix}\begin{pmatrix} 94-N_1 \\ 7 \end{pmatrix}}{\begin{pmatrix} 101 \\ 21 \end{pmatrix}} \qquad 7 \le N_1 \le 87$$

Jetzt beträgt die Wahrscheinlichkeit für eine Ablehnung der Vorlage

$$\sum_{N_1=7}^{50} \frac{\begin{pmatrix} N_1+6 \\ 13 \end{pmatrix}\begin{pmatrix} 94 \\ 7 \end{pmatrix}}{\begin{pmatrix} 101 \\ 21 \end{pmatrix}} = 0.09$$

Das gleiche Ergebnis erhalten wir bei einer Vernachlässigung des schrittweisen Vorgehens, wenn wir von einer Stichprobe mit $n = 20$ und $n_1 = 13$ ausgehen und die zugehörige Fiduzialverteilung berechnen, d.h.

$$P_\theta(N_1|n_1 = 7) = F_{20}^{100}(N_1|n_1 = 13)$$

Man kann auch sequentiell vorgehen und die *a posteriori*-Verteilung nach jeder Einzelbefragung berechnen. Als *a priori*-Verteilung wird die Vorgänger-*a posteriori*-Verteilung verwendet. Man erhält so das selbe Endergebnis. Gleichzeitig sieht man bei Verwendung eines Graphik-Displays, wie jede einzelne Befragung unser Wissen verändert und damit zu einer meßbaren Abnahme der Entropie führt.

8.4.3 Empirisch Linear Bayes'sche Modelle in der Ökonometrie

Ein schwer zu lösendes technisches Problem bei der Anwendung frequentistischer Modelle in mehrdimensionalen Situationen ist die Spezifikation multivariater *a priori*-Verteilungen. Das ist keine Besonderheit des Bayes'schen Zugangs. Die Identifikation multivariater Verteilungen erfordert allgemein einen sehr hohen Aufwand an empirischen Daten und Informationen, insbesondere dann, wenn die Struktur der Verteilung noch nicht aus sachlogischen Gesichtspunkten als vorgegeben betrachtet werden kann. Eine Verbesserung der Parameteridentifikation gegenüber klassischen Verfahren ist aber nur dann zu erwarten, wenn die geschätzte *a priori*-Verteilung in der Nähe der „wahren" Parameterverteilung liegt. Dies berücksichtigend, ist im Sinne eines globalen Risikos ein geeigneter Kompromiß zu finden zwischen der Grobheit der Modellierung und der in das Modell durch die Schätzung einer möglicherweise hohen Anzahl unbekannter Parameter zusätzlich hereingetragenen Unschärfe. In diesem Zusammenhang ist die häufige Anwendung linearer, quasilinearer oder linearisierter Modelle zu sehen.

Wir wollen hier kurz auf die Modelle mit zufälligen Koeffizienten (RCR-Modelle) eingehen, die in vielen Wissenschaftsdisziplinen Anwendungen gefunden haben, ihren Ursprung jedoch in der Biologie und Ökonomie hatten. Ausgangspunkt ist eine Problemstellung der folgenden Art:

Für mehrere ähnliche Objekte (Individuen, Gruppen) soll die funktionelle Abhängigkeit einer zufälligen Zielgröße Y von einem steuerbaren oder exakt meßbaren Einflußvektor $x \in \mathbb{R}^k$ untersucht und modelliert werden.

Fassen wir die n_i Messungen am i-ten Objekt ($i = 1, 2, \ldots, N$) in einem n_i-Beobachtungsvektor $\boldsymbol{Y}_i$ zusammen und die zugehörigen Einstellungen in einer $(n_i \times k)$-Matrix $\boldsymbol{X}_i$, so betrachten wir für das i-te Objekt folgenden linearen Ansatz:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$

wobei $\{\boldsymbol{\varepsilon}_i, \boldsymbol{\beta}_i | i = 1, \ldots, N\}$ eine Menge stochastisch unabhängiger zufälliger Vektoren sei mit gewissen Verteilungsannahmen. Die ε_i können als zufällige Fehlervektoren und die β_i als nichtbeobachtbare zufällige Parameter gedeutet werden, die in jedem Modell unabhängig ihre objektspezifische Ausprägung (Realisierung) finden.

Hierbei interessieren u. a. folgende Fragestellungen:

- Welche Aussagen kann man über die meist unbekannten Parameter der gemeinsamen Verteilung von $(\boldsymbol{\varepsilon}_i, \boldsymbol{\beta}_i)$ (Populationsparameter) auf der Basis der beobachtbaren (manifesten) Vektoren $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$ treffen?

- Wie kann man die nichtbeobachtbaren (latenten) objektspezifischen Realisierungen der $\boldsymbol{\beta}_i$ schätzen (vorhersagen) unter Berücksichtigung der

Erfahrungen bei der Untersuchung anderer Objekte?

- Was kann man über den Fehler bei der Identifikation des Individualparameters aussagen, der durch die Schätzung der Populationsparameter impliziert wird?

Derartige Fragen wurden für verschiedene Modellansätze in Gladitz [Gla81] behandelt. Es werden empirisch linear Bayes'sche Schätzer (ELBE) vorgeschlagen, die nur von den ersten zwei Momenten der *a priori*-Verteilung abhängen und konsistent geschätzt werden können. Für die ELBE selbst wurde asymptotische Optimalität im Sinne von Robbins [Rob64] nachgewiesen.

Abschließend wollen wir als Beispiel die Kombination von Zeitreihen und Querschnittsdaten in der Ökonometrie betrachten:

Auch ökonomische Prozesse (z.B. Wachstumsprozesse, Investitionsverhalten, Sparverhalten, Konsumtionsverhalten usw.) werden für gewisse Zeitabschnitte durch lineare Modelle approximiert. Denkt man dabei an reale Planungsmodelle etwa für einen Industriezweig, so müssen große Systeme identifiziert werden, die möglicherweise von vielen unbekannte Parametern abhängen. Da aber ökonomische Kennziffern fast immer nur jahresweise sinnvoll verglichen werden können (bei halbjährlicher oder quartalsweiser Betrachtung treten oft Saisoneffekte auf), ist die Anzahl der anfallenden Primärdaten für eine Schätzung der unbekannten Systemparameter meist unzureichend. Neben anderen Zusatzinformationen ist die Verwendung von Zeitreihen aus ähnlichen Bereichen, Zweigen, Bezirken, Betrieben der nationalen Wirtschaft oder aus anderen Ländern mit vergleichbarer Wirtschaftsstruktur (Querschnittsdaten) für die Systemidentifikation notwendig. In günstigen Fällen kann man aus diesen Sekundärdaten eine *a priori*-Verteilung für eine Teilmenge der unbekannten Parameter oder für gewisse Charakteristika derselben schätzen. Ein anschließendes Bayes'sches Vorgehen liefert dann eine sinnvolle Kopplung der Information aus der Primärzeitreihe und den Querschnittsdaten.

Als einfachstes Beispiel wurde in der Literatur (z.B. Swamy [Swa71]) das Keynes'sche Konsumtionsmodell

$$C_t^{(i)} = \alpha^{(i)} + \beta^{(i)} x_t^{(i)} + \varepsilon_t^{(i)} \qquad i = 1, \ldots, N; \quad t = 1, 2, \ldots, T_i$$

betrachtet, das für den i-ten Sektor den Zusammenhang zwischen der Konsumtion C und dem Realeinkommen x im Jahre t widerspiegeln soll (ε_t zufälliger Fehlerterm).

Unter geeigneten Stabilitätsannahmen kann die Information aus den Querschnittsdaten $(C_t^{(j)}, x_t^{(j)}), j \neq i, t = 1, \ldots, T$ bei einer Entscheidung über die Realisierung von $(\alpha^{(i)}, \beta^{(i)})$ mit verwendet werden.

Ein ähnliches Vorgehen bietet sich an bei dem in der Ökonomie häufig verwendeten Ansatz von Produktionsfunktionen. Zur Veranschaulichung modifizieren wir ein von K.M.S. Humak [Hum77] angegebenes Beispiel:

Für einen Sektor i soll das Bruttosozialprodukt $y_t^{(i)}$ im Jahre t in Abhängigkeit vom Grundfondsbestand $x_t^{(i)}$ und der Arbeitskräftezahl $z_t^{(i)}$ dargestellt werden. Als Ansatz wird oft eine sogenannte Cobb-Douglas-Produktionsfunktion

$$y_t^{(i)} = \left[x_t^{(i)}\right]^{\alpha_i} \left[z_t^{(i)}\right]^{\beta_i} \varepsilon_t^{(i)}$$

gewählt, die — logarithmiert — auf ein lineares Modell führt. Auch hier nehmen wir an, wir hätten bereits Daten für einige Jahre im Sektor i sowie aus anderen ähnlichen Sektoren, d.h. $i = 1, \ldots, N; t = 1, \ldots, T_i$. Das Problem ist die Bestimmung der unbekannten Parameter α_i, β_i mit dem Ziel der Vorhersage des Bruttosozialprodukts $y_{T+1}^{(i)}$ für ein zukünftiges Jahr $T + 1$ bei $x_{T+1}^{(i)}$ und $z_{T+1}^{(i)}$. Wie oben kann für eine Entscheidung unter gewissen Bedingungen die Information aus den Querschnittsdaten mit verwendet werden.

Weitere interessante Beispiele hat Donald B. Rubin [Rub84] angegeben.

8.5 Abschließende Bemerkungen

Der vorliegende Beitrag hatte nicht die Vorstellung neuer Originalresultate zum Ziel. Das Anliegen bestand vielmehr darin, das Bayes'sche Informationsprinzip einem sozialwissenschaftlichen Leserkreis nahezubringen, einige grundlegende Ideen und Anregungen zu vermitteln, sowie bestehende Vorurteile abzubauen. Weitere Anstrengungen sind nötig, um dem Bayes'schen Zugang den ihm gebührenden Stellenwert in sozialwissenschaftlichen Anwendungen einzuräumen.

Die Literaturliste existiert nur als Gesamtliste am Ende des Tagungsbandes. Wer diese haben will, findet diese entweder im Tagungsband 'Computer Aided Sociological Research' oder schickt eine Email an j.gladitz@statistik-service.de.