SpringerWienNewYork

Reinhard Viertl
Dietmar Hareter

Beschreibung und Analyse unscharfer Information

Statistische Methoden für unscharfe Daten

SpringerWienNewYork

o. Univ.-Prof. Dipl.-Ing. Dr. techn. Reinhard K. W. Viertl
Dipl.-Ing. Dr. techn. Dietmar Hareter
Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien
Wien, Österreich

Das Werk ist urheberrechtlich geschützt.
Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdruckes, der Entnahme von Abbildungen, der Funksendung, der Wiedergabe auf photomechanischem oder ähnlichem Wege und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten.
Produkthaftung: Sämtliche Angaben in diesem Fachbuch (wissenschaftlichen Werk) erfolgen trotz sorgfältiger Bearbeitung und Kontrolle ohne Gewähr. Insbesondere Angaben über Dosierungsanweisungen und Applikationsformen müssen vom jeweiligen Anwender im Einzelfall anhand anderer Literaturstellen auf ihre Richtigkeit überprüft werden. Eine Haftung des Autors oder des Verlages aus dem Inhalt dieses Werkes ist ausgeschlossen.
Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Buch berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürfen.
© 2006 Springer-Verlag/Wien
Printed in Austria

Reproduktionsfertige Vorlage von den Autoren
Druck: G. Grasl GmbH, 2540 Bad Vöslau, Österreich
Gedruckt auf säurefreiem, chlorfrei gebleichtem Papier – TCF
SPIN 1135346

Mit 66 Abbildungen

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISBN-10 3-211-23877-8 SpringerWienNewYork
ISBN-13 978-3-211-23877-6 SpringerWienNewYork

Vorwort

Beobachtungen und Messungen diverser Größen sind oft nicht einfach Zahlen oder Vektoren, sondern wichtige Fakten, auch Daten genannt. Das Wort kommt von Datum in der Bedeutung von Angabe und Tatsache, d.h. das, was aktuell gegeben ist. Bereits Nikolaus von Kues (Cusanus) stellte im 15. Jahrhundert die ,,grundsätzlich unvermeidbare Ungenauigkeit jeder quantitativen Messung'' fest. Der Mediziner Julius Robert von Mayer (1814–1878) bemerkte Jahrhunderte später: ,,Zahlen sind die gesuchten Fundamente einer exakten Naturforschung''. In der Tat trieb die numerische Beschreibung von realen Phänomenen die Naturwissenschaften in der Folge enorm voran. Die Galilei'sche Devise ,,Miss alles, was messbar ist, und das Nichtmessbare mache messbar'' ist nach wie vor grundlegend für die quantitative Wissenschaft.

Damit ist auch der Zusammenhang zwischen Messvorgängen und quantitativer Analyse angesprochen, der zentral für die folgenden Ausführungen ist.

Gemessene und beobachtete Daten stellen eine spezielle Art von *Information* dar. Als der Begriff ,,Information'' vor rund 70 Jahren Eingang in die Wissenschaft fand, ging es zunächst ausschließlich um die Übermittlung und Übertragung von Nachrichten. Seit dieser Zeit kristallisierte sich der Begriff als ähnlich grundlegend heraus, wie es beispielsweise die beiden Begriffe Energie und Materie als Basis jeder Naturwissenschaft geworden sind. Dabei steht der Begriff Information an der Grenze zwischen den Naturwissenschaften und den Geisteswissenschaften. Norbert Wiener (1894–1964), der Begründer der Kybernetik, formulierte dies mit den Worten: ,,Information ist Information, weder Materie noch Energie. Kein Materialismus, der dies nicht berücksichtigt, kann heute überleben''.

Die nähere Betrachtung und kritische Hinterfragung der Qualität von Informationen zeigt, dass sie häufig mit verschiedenen Arten von Ungewissheit behaftet sind. Vor allem in sprachlich übermittelten Informationen treten häufig linguistische Unsicherheiten auf. Beispielsweise sind die Aussagen ,,für

eine kurze Zeit“, „eine große Distanz“ oder „erhöhte Temperatur“ nicht Informationen im Sinne einer exakt bestimmbaren Zahl, sind aber trotzdem von hohem Informationsgehalt.

Die mathematische Beschreibung dieser Informationen, beispielsweise in der Modellierung des menschlichen Verhaltens, durch die Angabe von exakten reellen Zahlen als Bedeutung der einzelnen Aussagen, ist deshalb oft nicht adäquat. Selbst Informationen, die augenscheinlich als „exakt“ angesehen werden, wie beispielsweise die Ergebnisse von Messungen, sind bei näherer Betrachtung mit Ungewissheit behaftet. In Abschnitt 1.1 wird die Unschärfe von Messungen an einigen praktischen Beispielen erläutert. Grundsätzlich treten bei Bestimmung und Messung von eindimensionalen kontinuierlichen Größen mehrere Arten von Ungewissheit auf: Zufälligkeit, Unschärfe, Messfehler und Modellunsicherheiten.

Das derzeit vorherrschende Konzept zur Beschreibung der Unsicherheit in den Daten ist die Verwendung von auf Wahrscheinlichkeiten basierenden stochastischen Modellen. Diese stochastischen Modelle beziehen allerdings nur die zufällige Variabilität in die Modellbildung ein, während andere Formen der Ungewissheit ignoriert werden. Speziell die Unschärfe der Daten, deren Ursache häufig im Bestimmungsprozess selbst liegt und die sich auf die Darstellung und Beschreibung *einer* Beobachtung bezieht, ist nicht stochastischer Art.

Datenqualität, Genauigkeit oder Ungenauigkeit von Daten und anderen Informationen ist ein grundlegender Aspekt von Messungen und Beobachtungen, der quantitativ beschrieben werden muss, um unrealistische Resultate von Analysen zu vermeiden. In vielen praktischen Anwendungen erscheint die Angabe reeller Zahlen als vorliegende Datenelemente fragwürdig. Oftmals können die einzelnen Werte bestenfalls durch eine obere und eine untere Schranke abgegrenzt werden, wobei diese so genannten Intervallzahlen Spezialfälle so genannter *unscharfer Zahlen* sind. Die Verwendung von unscharfen Zahlen ermöglicht es, die Unschärfe in die Modellbildung mit einzubeziehen, und erlaubt somit eine realistischere Beschreibung der Daten. Die Unschärfe der Daten darf dabei nicht als Ersatz zur Wahrscheinlichkeitstheorie aufgefasst, sondern muss vielmehr als ein Konzept zur mathematischen Beschreibung und Behandlung nichtstochastischer Ungewissheiten angesehen werden. Die Kombination von statistischen Modellen für die Analyse mehrfacher Informationsangaben derselben Größe, z.B. durch mehrmalige Messung, mit der Beschreibung der Einzelmessungen mittels unscharfer Zahlen oder unscharfer Vektoren bildet den geeigneten Rahmen für die Analyse unscharfer Daten. Dies ist ein hybrider Ansatz, der zwei verschiedene Arten von Ungewissheit vereint.

Die statistischen Methoden dieses Bandes sind für Leser geschrieben, die

mit elementaren stochastischen Modellen und statistischen Verfahren vertraut sind. Das notwendige Vorwissen entspricht dem einer Einführung in die Stochastik, z.B [Vi03a] des Literaturverzeichnisses.

Ziel der Ausführungen ist es, Methoden der quantitativen Beschreibung unscharfer Beobachtungen stochastischer Größen vorzustellen und in die Grundlagen der statistischen Analyse solcher Daten einzuführen. Der praktische Umgang mit den vorgestellten Theorien und Methoden wird dem Leser anhand zahlreicher Übungsaufgaben nähergebracht.

Reinhard Viertl
Wien, September 2005 *Dietmar Hareter*

Inhaltsverzeichnis

Symbolverzeichnis

$\oplus$	Summation zweier unscharfer Zahlen
$\odot$	Multiplikation zweier unscharfer Zahlen
$\oslash$	Division zweier unscharfer Zahlen
$\ominus$	Differenzbildung zweier unscharfer Zahlen
$\emptyset$	leere Menge
$\oint$	verallgemeinerte Integration
$\oint\!\!\!\!/$	Integration zur Berechnung unscharfer Wahrscheinlichkeiten
$\times$	cartesisches Produkt
$\cup$	mengentheoretische Vereinigung
$\cap$	mengentheoretischer Durchschnitt
$\sim$	verteilt nach
$\langle m, s, l, r \rangle_{LR}$	LR-Darstellung einer unscharfen Zahl
$\#$	Anzahl der Elemente in einer Menge
A	klassische Teilmenge oder Annahmebereich bei Tests
A^c	Komplement der Menge A
$A^\star$	unscharfe Menge
$C_\delta(x^\star)$	δ-Schnitt der unscharfen Zahl $x^\star$
$C_\delta(\boldsymbol{x}^\star)$	δ-Schnitt des unscharfen Vektors $\boldsymbol{x}^\star$
$co\,[\cdot]$	konvexe Hülle
$\mathcal{D}$	Menge der möglichen Entscheidungen
$\vartheta(\cdot, \ldots, \cdot)$	Schätzfunktion für einen Parameter
$d^\star(m, l, r)$	dreieckförmige unscharfe Zahl

$\mathbb{E}\,X$	Erwartungswert der stochastischen Größe X
$\mathcal{F}(\mathbb{R})$	Menge der unscharfen Zahlen
$\mathcal{F}(\mathbb{R}^n)$	Menge der n-dimensionalen unscharfen Vektoren
$\mathcal{F}_c(\mathbb{R}^n)$	Menge der n-dimensionalen unscharfen Vektoren mit konvexen δ-Schnitten
$f(\cdot \mid \cdot)$	bedingte oder Prädiktivdichte
$f^\star(\cdot)$	unscharfe Funktion
$\overline{f}_\delta(\cdot)$	obere δ-Niveaukurve der unscharfen Funktion $f^\star(\cdot)$
$\underline{f}_\delta(\cdot)$	untere δ-Niveaukurve der unscharfen Funktion $f^\star(\cdot)$
$F(\cdot)$	Verteilungsfunktion
$F^{-1}(\cdot)$	verallgemeinerte Inverse der Verteilungsfunktion $F(\cdot)$
$\widehat{F}_n(\cdot)$	empirische Verteilungsfunktion
$\widehat{F}_n^{-1}(\cdot)$	verallgemeinerte Inverse der empirischen Verteilungsfunktion $\widehat{F}_n(\cdot)$
$\widehat{F}_n^\star(\cdot)$	geglättete empirische Verteilungsfunktion oder unscharfe empirische Verteilungsfunktion
$\widehat{F}_{\mathrm{O},\delta}(\cdot)$	obere δ-Niveaukurve der unscharfen empirischen Verteilungsfunktion $\widehat{F}_n^\star(\cdot)$
$\widehat{F}_{\mathrm{U},\delta}(\cdot)$	untere δ-Niveaukurve der unscharfen empirischen Verteilungsfunktion $\widehat{F}_n^\star(\cdot)$
$h_n^\star(\cdot)$	unscharfe relative Häufigkeit
$\overline{h}_{n,\delta}(\cdot)$	obere Grenze des δ-Schnittes der unscharfen relativen Häufigkeit $h_n^\star(\cdot)$
$\underline{h}_{n,\delta}(\cdot)$	untere Grenze des δ-Schnittes der unscharfen relativen Häufigkeit $h_n^\star(\cdot)$
$H_n^\star(\cdot)$	unscharfe absolute Häufigkeit
$\overline{H}_{n,\delta}(\cdot)$	obere Grenze des δ-Schnittes der unscharfen absoluten Häufigkeit $H_n^\star(\cdot)$
$\underline{H}_{n,\delta}(\cdot)$	untere Grenze des δ-Schnittes der unscharfen absoluten Häufigkeit $H_n^\star(\cdot)$
$I_A(\cdot)$	Indikatorfunktion der Menge A
$i^\star(m,s)$	Intervallzahl
$K_n(\cdot,\ldots,\cdot)$	n-dimensionale Kombinationsregel

$\kappa(\cdot,\ldots,\cdot)$	Konfidenzfunktion
$L(\cdot)$	linke Begrenzungsfunktion in der LR-Darstellung
$L(\cdot,\cdot)$	Verlustfunktion
$l(\cdot\,;\ldots)$	Likelihood- oder Plausibilitätsfunktion
$l^\star(\cdot\,;\ldots)$	unscharfe Likelihood- oder Plausibilitätsfunktion
$\overline{l}_\delta(\cdot\,;\ldots)$	obere δ-Niveaukurve der unscharfen Likelihoodfunktion $l^\star(\cdot\,;\ldots)$
$\underline{l}_\delta(\cdot\,;\ldots)$	untere δ-Niveaukurve der unscharfen Likelihoodfunktion $l^\star(\cdot\,;\ldots)$
M	Merkmalraum
M_X	Merkmalraum der stochastischen Größe X
M_X^n	Stichprobenraum der stochastischen Größe X
m^k	k-tes Moment
$P(\cdot)$	klassische Wahrscheinlichkeitsverteilung
$P^\star(\cdot)$	unscharfe Wahrscheinlichkeitsverteilung
$\overline{P}_\delta(\cdot)$	obere Grenze des δ-Schnittes der unscharfen Wahrscheinlichkeit $P^\star(\cdot)$
$\underline{P}_\delta(\cdot)$	untere Grenze des δ-Schnittes der unscharfen Wahrscheinlichkeit $P^\star(\cdot)$
$\pi(\cdot)$	A-priori-Dichte
$\pi(\cdot\,\vert\,\cdot)$	A-posteriori-Dichte
$\pi^\star(\cdot)$	unscharfe Dichtefunktion
$\overline{\pi}_\delta(\cdot)$	obere δ-Niveaukurve der unscharfen Dichte $\pi^\star(\cdot)$
$\underline{\pi}_\delta(\cdot)$	untere δ-Niveaukurve der unscharfen Dichte $\pi^\star(\cdot)$
$\propto$	proportional
$R(\cdot)$	rechte Begrenzungsfunktion in der LR-Darstellung
s_n^2	Stichprobenvarianz
$\left(s_n^2\right)^\star$	unscharfe Stichprobenvarianz
$S_n^\star(\cdot)$	Summenkurve
θ	Parameter
$\widetilde{\theta}$	stochastischer Parameter
Θ	Parameterraum
$\mathcal{T}(\cdot,\ldots,\cdot)$	Statistik
$T(\cdot,\cdot)$	t-Norm

$Tr(x^\star)$	Träger der unscharfen Zahl $x^\star$
$Tr(\boldsymbol{x}^\star)$	Träger des unscharfen Vektors $\boldsymbol{x}^\star$
$t^\star(m,s,l,r)$	trapezförmige unscharfe Zahl
$U(\cdot,\cdot)$	Nutzenfunktion
$U^\star(\cdot,\cdot)$	unscharfe Nutzenfunktion
V	Verwerfungsbereich bei Tests
X	stochastische Größe
$x_1,\ldots,x_n$	reellwertige Stichprobe
$x_{(1)},\ldots,x_{(n)}$	geordnete reellwertige Stichprobe
$x^\star$	unscharfe Zahl, unscharfe Beobachtung
$x_1^\star,\ldots,x_n^\star$	unscharfe Stichprobe
$\boldsymbol{x}^\star$	unscharfer Vektor oder unscharfer kombinierter Vektor
$\overline{x}_n$	Mittelwert einer Stichprobe
$\xi_{x^\star}(\cdot)$	charakterisierende Funktion der unscharfen Zahl $x^\star$
$\xi_{\boldsymbol{x}^\star}(\cdot,\ldots,\cdot)$	vektorcharakterisierende Funktion des unscharfen Vektors $\boldsymbol{x}^\star$
$\mathcal{Z}_n(\mathbb{R}^n)$	Menge der n-dimensionalen Zugehörigkeitsfunktionen

1

Ungewissheit und Information

1.1 Unscharfe Information und unscharfe Daten

Vieles im Leben ist ungewiss. Dies beginnt bei der Lebensdauer von Menschen, geht über Preisentwicklungen und über technische Gegebenheiten bis hin zu zukünftigen Umweltgegebenheiten. Um fundierte Entscheidungen treffen zu können, ist die adäquate, möglichst gute, quantitative Beschreibung der betrachteten Größen notwendig. Neben Messungen und Beobachtungen sind häufig auch auf Erfahrungen gegründete Einschätzungen von Ungewissheiten durch Experten, also quantitative Beschreibungen von so genannter A-priori-Information in Entscheidungsanalysen wesentlich.

Die adäquate mathematische Beschreibung von realen Informationen ist oft nicht durch exakte Zahlen bzw. Vektoren möglich. Dies gilt vor allem, wenn für die entsprechende Größe bzw. Menge nur eine vage Beschreibung oder Charakterisierung vorliegt, wie es häufig bei linguistischen Aussagen der Fall ist. Beispielsweise ist die vage definierte Information ,,der Patient hat erhöhte Temperatur'' als Teilmenge von $\mathbb{R}$ nicht eindeutig festgelegt. Einerseits ist die quantitative Beschreibung der Information bzw. der Menge ,,erhöhte Temperatur'' für die Modellierung von medizinischem Wissen notwendig, andererseits ist diese Festlegung mit vielen praktischen Schwierigkeiten verbunden: Ist eine Temperatur von 37.5 °C erhöht? Bei welchen Temperaturen soll die obere und untere Grenze der Menge festgelegt werden? Viel entscheidender ist die Frage, ob eine strikte Festlegung der Grenzen für die Beschreibung dieser Menge überhaupt sinnvoll ist.

Ein ähnliches Problem stellt die Anteilschätzung dar. Die Frage nach dem Anteil von Rauchern in einer Firma kann beispielsweise nicht immer eindeutig beantwortet werden, da die Definition eines ,,Rauchers'' ein zu allgemeiner Begriff ist. Sind Gelegenheitsraucher den Rauchern zuzuordnen? Wie werden Mitarbeiter eingeordnet, die sich gerade das Rauchen abgewöhnen wollen?