Nichtparametrische Schätzung von Wahrscheinlichkeitsverteilungen

Wolfgang Wertz, TU Wien

1. Einleitung

1.1. Die Vermittlung der Grundlagen der Wahrscheinlichkeitsrechnung und noch mehr die der schließenden (analytischen) Statistik im Mathematikunterricht an höheren Schulen stößt auf eine Reihe grundsätzlicher Schwierigkeiten: Den rein mathematischen Inhalten gesellen sich Probleme der Modellbildung hinzu, auf die ich im Abschnitt 2 zurückkomme; vor allem aber erweist sich der Übergang vom deterministischen Denken zum Denken in der Kategorie der Zufälligkeit als sehr schwierig. Wenngleich Schüler vielleicht noch nicht so sehr dem kausalen Denken verhaftet sind wie Erwachsene und auch neuen Begriffsbildungen aufgeschlossener gegenüberstehen mögen, so stellt eine überzeugende und dabei sachlich richtige Darstellung doch eine bedeutende Herausforderung an die didaktischen Fähigkeiten der Lehrer dar.

1.2. Dabei sollten Fragen wie die nach der Existenz des Zufalls und der Zufälligkeit, etwa als eigene Kategorie oder selbständige Modalität, gar nicht angesprochen werden, da diese Fragen einerseits eher zur Philosophie gehören, andrerseits aber für den formal-mathematischen Umgang mit dem Phänomen „Zufall" nur von untergeordneter Bedeutung sind. Gewiß werden dann und wann diese Dinge von Schülern angesprochen, sei es aus echtem Interesse, sei es aus einem natürlichen Instinkt für Verzögerungstaktik heraus – es ist daher wichtig, daß die Lehrer über diese Fragen gut Bescheid wissen.

1.3. Eine unvermeidliche Schwierigkeit in der Wahrscheinlichkeitsrechnung bietet die Antwort auf die berechtigte Frage, warum denn bestimmte Modelle zugrunde gelegt werden, konkret die Annahme bestimmter Verteilungen wie Normal-, Poissonverteilung u.dgl. Läßt sich eine solche Annahme für einige wenige dieser Verteilungen (Gleichverteilung, Binomialverteilung und hypergeometrische Verteilung) leicht begründen, so erfordern etwa die zuerst genannten Verteilungen eine wesentlich aufwendigere Rechtfertigung, welche die Möglichkeiten des Mathematikunterrichtes beträchtlich überschreiten (immerhin sollten die Lehrer selbst über

einen profunden Einblick in die Problematik der wahrscheinlichkeitstheoretischen Modellbildung verfügen, sodaß sie prägnante Hinweise geben können, und die Schüler den erforderlichen Hinweis auf die Komplexität der Fragestellung nicht als Ausweichen verstehen). Besonders verwickelt wird die Lage, wenn die Rechtfertigung von Modellen Methoden der schließenden Statistik benützen müßte, die ja ihrerseits die Wahrscheinlichkeitsrechnung voraussetzt, sodaß der Eindruck eines logischen Zirkels entstehen mag. In **3.7** wird angedeutet, wie sich mit nichtparametrischen Verfahren manche Begriffe leichter und anschaulicher vermitteln lassen, weil der logische Umweg über parametrisierte Verteilungsklassen wegfällt. (Dies soll aber keinesfalls die große Bedeutung parametrischer Modelle in irgendeiner Weise in Frage stellen!)

1.4. Gerade die Behandlung nichtparametrischer Verfahren ermöglicht und erfordert einen engen Bezug zum Informatikunterricht, und manche dieser Probleme (z.B. Simulationen und andere EDV-aufwendige Projekte) können sinnvollerweise nur im Rahmen des Wahlpflichtfaches Informatik behandelt werden.

2. Die Schätzung von Wahrscheinlichkeiten

2.1. Die Begründung eines tragfähigen Begriffes von Wahrscheinlichkeit sollte bei der Beschäftigung mit der Wahrscheinlichkeitsrechnung an erster Stelle stehen. Um einen solchen wurde durch Jahrhunderte gerungen: Schon in der „Ars conjectandi" (1713) von Jakob Bernoulli finden sich, wenn auch vage, sowohl der „klassische" („Laplace'sche") Wahrscheinlichkeitsbegriff als auch der frequentistische. Nach einer Reihe mehr oder minder erfolgloser Versuche, die Wahrscheinlichkeitsrechnung auf eine mathematisch exakte Basis zu stellen (etwa die Kollektivtheorie von R.von Mises, *v.Mises (1919), (1931)*), gelang es erst A.N.Kolmogorow (*Kolmogorow (1933)*), eine befriedigende mathematische Theorie auf maßtheoretischer Grundlage zu entwickeln, die sich auch hinsichtlich ihrer Interpretation als äußerst flexibel erwiesen hat. Im Schulunterricht steht allerdings der „klassische" Wahrscheinlichkeitsbegriff immer noch sehr im Vordergrund; meine kritische Stellung dazu ergibt sich aus den in **2.3** angeführten Punkten (vgl. auch *Reichel, Hanisch, Müller (1987)*,2.Kap.)

2.2. Die „klassische Wahrscheinlichkeitsdefinition" läßt sich in heutiger Terminologie wie folgt fassen: Gegeben sei eine endliche Menge Ω als

Grundgesamtheit, die aus den möglichen Versuchsausgängen besteht. Jede Teilmenge $A \subseteq \Omega$ heißt *Ereignis*.

$W(A) := \frac{|A|}{|\Omega|}$ heißt *Wahrscheinlichkeit des Ereignisses A*, wobei $g := |A|$ die Anzahl der Elemente von A bezeichnet und $m := |\Omega|$ die von Ω. g nennt man auch die *Anzahl der* (für das Eintreten von A) *günstigen* und m die *Anzahl der möglichen Fälle*. Die entscheidende Modellvoraussetzung ist dabei eine Symmetrieannahme, die gewährleistet, daß alle Versuchsausgänge „mit gleicher Wahrscheinlichkeit" auftreten.

2.3. Die klassische Wahrscheinlichkeitsdefinition kann eine Reihe falscher Eindrücke von Wahrscheinlichkeit wecken:

a. Sie erscheint zunächst als willkürlich; das Postulat gleichwahrscheinlicher Fälle hängt anfangs begrifflich in der Luft.

b. Es entsteht der Eindruck, daß sich Wahrscheinlichkeiten grundsätzlich berechnen lassen. Das wäre eine unzulässige Generalisierung eines Spezialfalles des Kolmogorow'schen Modells **2.6**.

c. Rein technische Überlegungen (Kombinatorik, „Baumdiagramme") stehen derart im Vordergrund, daß das Wesen „zufälligen Geschehens" verdeckt wird; das Abzählen von möglichen und günstigen Fällen fördert eher noch die Verhaftung im deterministischen Denken.

d. Die Definition von bedingter Wahrscheinlichkeit, Zufallsvariabler usw. gelingt nur im vorgesteckten engen Rahmen.

e. Die Definition des Erwartungswertes und von Momenten gerät sehr formal und schlecht interpretierbar.

f. Die Einführung stetiger Verteilungen und Zufallsvariabler gerät zu einem unmotivierten und logisch schlecht begründeten Kraftakt; gerade guten Schülern, die den Sinn von Modellbildung verstanden haben und kritisch Rechtfertigungen suchen, müssen die unvermeidlichen Verständnislücken stören.

2.4. Der frequentistischer Wahrscheinlichkeitsbegriff

V sei ein Versuch (Experiment), Ω die *Grundgesamtheit* von möglichen Versuchsausgängen einer beliebigen Menge, $\mathfrak{E}$ ein *Ereignisfeld*.

Formal ist ein Ereignisfeld über Ω durch folgende Axiome definiert:

$$A \subseteq \Omega \quad \forall A \in \mathfrak{E}$$

$$\varnothing \in \mathfrak{E}$$

$$A \in \mathfrak{E} \Rightarrow A^c \in \mathfrak{E}$$

$$A_1, A_2, \ldots, A_n, \ldots \in \mathfrak{E} \quad \Rightarrow \quad \bigcup_{n=1}^{\infty} A_n \in \mathfrak{E} \quad .$$

Gegeben sei ferner eine Folge $(V_n)_{n=1,2,\ldots}$ von Versuchen, die darin besteht, daß ein und derselbe Versuch V unabhängig und unter identischen Versuchsbedingungen durchgeführt wird. $\omega_n \in \Omega$ sei der Versuchsausgang bei V_n $(n \in \mathbb{N})$.

$h_n(A) := \#\{ i : 1 \le i \le n, \omega_i \in A \}$ sei die Anzahl der Versuche, bei denen A eingetreten ist und $r_n(A) := r_n(A;\omega_1,\ldots,\omega_n) := h_n(A)/n$ die *relative Häufigkeit* von A bei den ersten n Versuchen.

Die Erfahrung zeigt, daß für (praktisch) jede Folge (ω_n) die Folge $(r_n(A))$ ein konvergenzartiges Verhalten aufweist, also eine Art „Grenzwert" $W(A)$ besitzt, der außerdem von der Folge (ω_n) unabhängig ist.

2.5. Die relativen Häufigkeiten besitzen offensichtlich folgende drei Eigenschaften:

$$0 \le r_n(A) \le 1 \qquad \forall A \in \mathfrak{E} \qquad \ldots(2.1)$$

$$r_n(\Omega) = 1 \qquad \ldots(2.2)$$

(Ω heißt auch das *„sichere Ereignis"*)

Ist $(A_k)_{k=1,2,\ldots}$ eine Folge von einander ausschließenden Ereignissen $(A_i \cap A_j = \varnothing \text{ für } i \ne j)$, so folgt:

$$r_n\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} r_n(A_k) \qquad \ldots(2.3)$$

2.6. Das Kolmogorow'sche Modell

Das empirische (konvergenzartige) Verhalten der relativen Häufigkeiten legt nahe, *(2.1) – (2.3)* auf den „Grenzwert" W zu übertragen und zu definieren:

Eine Abbildung $W : \mathfrak{E} \to \mathbb{R}$ heißt *Wahrscheinlichkeitsverteilung*, wenn gilt:

$$0 \le W(A) \le 1 \qquad \forall A \in \mathfrak{E} \qquad \ldots(2.1')$$

$$W(\Omega) = 1 \qquad \ldots(2.2')$$

Ist $(A_k)_{k=1,2,\ldots}$ eine Folge von einander ausschließenden Ereignissen $(A_i \cap A_j = \varnothing \text{ für } i \ne j)$, so folgt:

$$W\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} W(A_k) \qquad \ldots(2.3')$$

$(\Omega,\mathfrak{E},W)$ heißt dann *Wahrscheinlichkeitsraum*, $W(A)$ heißt die *Wahrscheinlichkeit des Ereignisses A*. $(\Omega,\mathfrak{E},W)$ ist also ein Maßraum, der mit 1 normiert ist.
Von den zahlreichen Lehrbüchern über Wahrscheinlichkeitstheorie seien die folgenden besonders empfohlen: *Bauer* (1991), *Gnedenko* (1991), *Pfanzagl* (1988), *Rényi* (1962) und *Schürger* (1998).

2.7. Das konvergenzartige Verhalten der relativen Häufigkeit bezieht sich stets auf bestimmte (feste) Ereignisse. Die Wahrscheinlichkeit $W(A)$ idealisiert das Verhalten der Folgen $(r_n(A;\omega_1,\ldots,\omega_n))$: $W(A)$ hängt weder von der konkreten Folge (ω_n) noch von n ab.

2.8. Im Modell **2.6** lassen sich Konvergenzbegriffe definieren (z.B. stochastische und fast sichere Konvergenz), in deren Sinn tatsächlich gilt:

$$\lim_{n\to\infty} r_n(A) = W(A) \qquad \forall A \in \mathfrak{E} \qquad \ldots(2.4)$$

(Dies ist eine Folge der Gesetze der großen Zahlen). Dies zeigt, daß sich der empirische Ausgangspunkt **2.4** im Modell beweisen läßt und dieses somit, zumindest in dieser Hinsicht, adäquat ist.
Die Schwierigkeit bei der Bestimmung von $W(A)$ gemäß *(2.4)* besteht darin, daß die (unendliche) Folge von Versuchen fiktiv ist, da ja nur endlich viele Versuche durchführbar sind und somit immer nur ein endlicher Abschnitt der Folge $(r_n(A))$ zur Verfügung steht, für die aber kein Bildungsgesetz vorliegt. In diesem Sinne läßt sich $W(A)$ also nur empirisch *schätzen*, aber im allgemeinen *nicht exakt bestimmen*.

2.9. Unter bestimmten Umständen lassen sich, wie in **1.3** erwähnt, bestimmte Verteilungen W als Modell wählen, in manchen Fällen aber nur approximativ angeben (vgl. Abschnitt **3** und **4**).
Andrerseits bietet *(2.4)* den Ansatz zur richtigen Interpretation des Modells $(\Omega,\mathfrak{E},W)$: Ist $A \in \mathfrak{E}$ beliebig gewählt, so ist zu erwarten, daß für (fast) jede Folge von künftigen Versuchsausgängen, so wie sie in **2.4** beschrieben ist, für „große" Werte von n, also viele Beobachtungen, annähernd gelten wird:

$$r_n(A;\omega_1,\ldots,\omega_n) \approx W(A)$$

Wie groß n gewählt werden muß, bleibt freilich offen.

2.10. Das Gesagte steht nicht in Widerspruch zu **2.2**. Weichen beispielsweise bei einer großen Zahl von Würfen mit einem Würfel die relativen Häufigkeiten der Augenzahlen stark von $\frac{1}{6}$ ab, so wird der Würfel nicht

homogen sein oder geometrisch nicht entsprechen; in diesem Falle wäre das Laplace'sche Modell eben ungeeignet. Es kann sich aber auch um eine wenig wahrscheinliche Folge von Versuchsausgängen handeln.

2.11. *(2.4)* gilt zwar für alle Ereignisse $A \in \mathfrak{E}$, die Konvergenzgeschwindigkeit für verschiedene A könnte aber unterschiedlich sein. Es fragt sich, ob es Teilfamilien $\mathfrak{K} \subseteq \mathfrak{E}$ gibt, sodaß in *(2.4)* Gleichmäßigkeit der Konvergenz in $A \in \mathfrak{K}$ vorliegt, d.h.:

$$\lim_{n\to\infty} \sup_{A\in\mathfrak{K}} \left| r_n(A) - W(A) \right| = 0 \qquad \ldots(2.5)$$

Falls $\mathfrak{E}$ endlich ist, folgt *(2.5)* mit $\mathfrak{K} = \mathfrak{E}$ unmittelbar aus *(2.4)*, für eine allgemeine Situationen geben wir in **3.5** aber auch **4.8** Ergebnisse an.

3. Die empirische Verteilungsfunktion

3.1. Es sei $(\Omega, \mathfrak{E}, W)$ ein gegebener Wahrscheinlichkeitsraum. Eine Abbildung $X : \Omega \to \mathbb{R}$ heißt *(reelle) Zufallsvariable* auf $(\Omega, \mathfrak{E}, W)$, wenn gilt:

$[a < X \le b] := \{\omega \in \Omega : a < X(\omega) \le b\} \in \mathfrak{E} \qquad \forall\, a,b : -\infty < a < b < \infty$

Gewöhnlich werden Zufallsvariable in einem wesentlich allgemeineren Rahmen definiert, vgl. dazu die in **2.6** zitierten Bücher. Für das Folgende reicht die hier gegebene Definition aber aus.

3.2. Durch $P((a,b]) := W([a < X \le b]) \qquad (-\infty < a < b < \infty)$ wird eine Wahrscheinlichkeitsverteilung auf einem geeigneten Ereignisfeld $\mathfrak{B}$ von Teilmengen von $\mathbb{R}$ *(Borel'sche Mengen)* festgelegt. P heißt die *Verteilung* von X.

3.3. Die durch

$$F(x) := W([X \le x]) = P((-\infty, x]) \qquad x \in \mathbb{R} \qquad \ldots(3.1)$$

definierte Funktion F: $\mathbb{R} \to \mathbb{R}$ heißt die *Verteilungsfunktion* von X bzw. von P. Umgekehrt wird durch F genau eine Verteilung P bestimmt. P ist durch

$$P((a,b]) := F(b) - F(a) \qquad \forall\, a,b : -\infty < a < b < \infty$$

festgelegt. Will man also die Wahrscheinlichkeitsverteilung P im Sinne von **2.8** schätzen, so genügt es, F zu schätzen.

3.4. Es seien $X_1,...,X_n$ unabhängige und nach einer Verteilung P mit Verteilungsfunktion F verteilte Zufallsvariable, $x_1,...,x_n$ seien die Werte, die für die Versuchsausgänge $\omega_1,...,\omega_n$ angenommen werden: $x_i := X(\omega_i)$. $(x_1,...,x_n)$ heißt dann *Stichprobe* (dieses Wort wird oft in unterschiedlichem Sinne gebraucht!). Ist in dieser Situation F unbekannt, so liegt folgendes Verfahren zur Schätzung von F nahe: Es sei

$$\hat{F}_n(x; x_1,...,x_n) := \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(x_i) \qquad \forall\, x \in \mathbb{R}; \qquad ...(3.2)$$

dabei bezeichnet $\mathbf{1}_{(-\infty,x]}(y) := 1$, falls $y \le x$, und $:= 0$ andernfalls.

$\hat{F}_n(\cdot; x_1,...,x_n)$ ist die Verteilungsfunktion der Verteilung $P_n = P_n(\cdot; x_1,...,x_n)$, die jedem x_i die Wahrscheinlichkeit $1/n$ zuordnet (gleiche Werte werden mehrfach gezählt). Aus *(3.2)* ergibt sich sofort, daß

$$P_n(A; x_1,...,x_n) = r_n(A; x_1,...,x_n)$$

gilt. $\hat{F}_n = \hat{F}_n(\cdot; x_1,...,x_n)$ heißt die *empirische Verteilungsfunktion* (zur Stichprobe $(x_1,...,x_n)$), $P_n = P_n(\cdot; x_1,...,x_n)$ die *empirische Verteilung*. In Bereichen, in denen viel „Wahrscheinlichkeitsmasse" von P konzentriert ist, also auch mit großer Wahrscheinlichkeit viele Werte x_i liegen werden, weist die Verteilungsfunktion F einen größeren Anstieg auf. Daher ist zu erwarten, daß $\hat{F}_n$ der Gestalt von F folgt (s. Abb. 1). $x_{(1)},...,x_{(n)}$ bezeichnet dort die Werte x_i, ihrer Größe nach geordnet, wobei gleiche Werte auch doppelt gezählt werden.

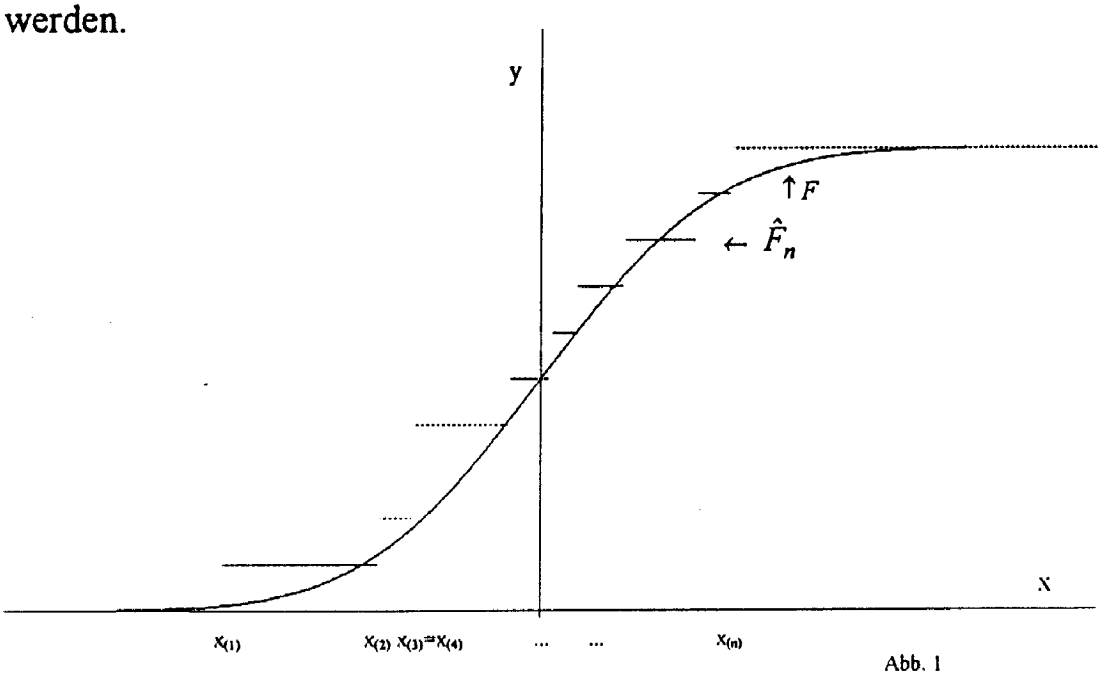

Abb. 1

3.5. Aus den Gesetzen der großen Zahlen folgt leicht:

$$W\left(\left[\lim_{n\to\infty} \hat{F}_n(x;X_1,\ldots,X_n) = F(x)\right]\right) = 1 \quad \textit{für jede Verteilungsfunktion } F.$$

Es gilt sogar

$$W\left(\left[\lim_{n\to\infty} \sup_{-\infty<x<\infty} \left|\hat{F}_n(x;X_1,\ldots,X_n) - F(x)\right| = 0\right]\right) = 1 \qquad \ldots(3.3)$$

für jede Verteilungsfunktion F.

Diese wichtige Aussage, der *Satz von Gliwenko–Cantelli*, heißt auch *Zentralsatz der Statistik*. Er besagt, daß sich *jede* (eindimensionale) Verteilungsfunktion gleichmäßig mit Hilfe der empirischen Verteilungsfunktion approximieren läßt. Mittels einer (unendlichen) Versuchsreihe läßt sich also jede Verteilung bestimmen. (vgl. **2.8.**).

3.6. Aus *(3.3)* folgt unmittelbar:

$$W\left(\left[\lim_{n\to\infty} \sup_{A\in\mathfrak{K}} \left|P_n(A) - P(A)\right| = 0\right]\right) = 1 \quad \textit{für jede Verteilung } P \qquad \ldots(3.4)$$

wenn $\mathfrak{K} = \{(a,b] : -\infty < a < b < \infty\}$ gewählt wird. Für $\mathfrak{K} = \mathfrak{B}$ (vgl. **3.2**) wäre *(3.4)* aber falsch!

3.7. Parametrische Verfahren der Statistik setzen voraus, daß die mögliche Verteilung einer Zufallsvariablen einer Klasse $\{P_\gamma : \gamma \in \Gamma\}$ von Verteilungen angehört, sodaß nur der Parameter γ unbekannt bleibt. Das Schätzproblem besteht nun in der (annähernden) Bestimmung des Parameters γ aufgrund von beobachteten Werten $x_1,\ldots,x_n$. Bei vielen gängigen parametrisierten Verteilungsfamilien besitzen die Parameter meist eine leicht erfaßbare Bedeutung, im allgemeinen muß dies aber durchaus nicht der Fall sein. Ungeklärt bleibt jedenfalls, in welchem Sinne die unbekannte Verteilung selbst angenähert wird, wenn die Parameter gut geschätzt werden.
(3.4) hingegen macht von keiner Parametrisierung Gebrauch, sondern klärt das Problem ohne Umwege. Das Schätzen der Verteilungsfunktion mittels der empirischen Verteilungsfunktion ist also ein nichtparametrisches Verfahren. Die Bereiche parametrisch / nichtparametrisch lassen sich aber nicht scharf gegeneinander abgrenzen.

3.8. Die Aussagen *(3.3)* und *(3.4)* können durch Simulationsaufgaben im Informatikunterricht gut illustriert werden. Dabei lernen die Schüler

a. Simulation von Zufallszahlen nach vorgegebenen Verteilungen;

b. den Umgang mit relativ großen Datensätzen; realistische Aufgaben erfordern große Stichprobenumfänge n, die graphische Darstellung zahlreiche Stützpunkte x;

c. Fragestellungen, die die Konvergenzgeschwindigkeit betreffen, z.B. : Was ist die Größenordnung von

$$D_n := \sup_{-\infty<x<\infty} \left| \hat{F}_n(x; X_1,...,X_n) - F(x) \right| \quad ? \qquad ...(3.5)$$

(Im Eindimensionalen ist die Verteilung der Zufallsvariablen D_n übrigens von F unabhängig, sofern F nur als stetig vorausgesetzt wird.)

Im übrigen können die intuitiv einfachen Ein- und Zweistichprobentests von Kolmogorow und Smirnow (siehe z.B. *Büning, Trenkler (1994)*) einen guten Einstieg in die Testtheorie geben.

3.9. $\hat{F}_n(\cdot; x_1,...,x_n)$ enthält genau die gleiche Information über die unbekannte Verteilung wie die Stichprobe $(x_1,...,x_n)$ selbst, da diese aus $\hat{F}_n$ rekonstruierbar ist.

3.10. Für Modelle, in denen F als stetig vorausgesetzt wird, hat die empirische Verteilungsfunktion $\hat{F}_n$ den Nachteil, daß sie immer Unstetigkeitsstellen aufweist. Es liegt daher nahe, $\hat{F}_n$ zu glätten. Am einfachsten ist der Fall einer linearen Glättung (Abb. 2): Es seien alle x_i verschieden und $d_i := \min(x_{(i+1)} - x_{(i)}, x_{(i)} - x_{(i-1)})$ für $i = 1,...,n$, wobei $x_{(0)} := -\infty$, $x_{(n+1)} := +\infty$ gesetzt wird, $a \in (0, 1/2)$ eine beliebig gewählte Zahl und $\xi_i := x_{(i)} - ad_i$ und $\eta_i := x_{(i)} + ad_i$.

$$G_n(x) := \begin{cases} (2nad_i)^{-1}(x - x_{(i)}) + (2i-1)/(2n) & \text{falls } x \in [\xi_i, \eta_i] \quad (i = 1,...,n) \\ \hat{F}_n(x) & \text{sonst} \end{cases}$$

Offensichtlich ist a ein Glättungsparameter: ein großes a bedeutet stärkere, ein kleines a geringere Glättung. Die Wahl von a ist aber willkürlich.

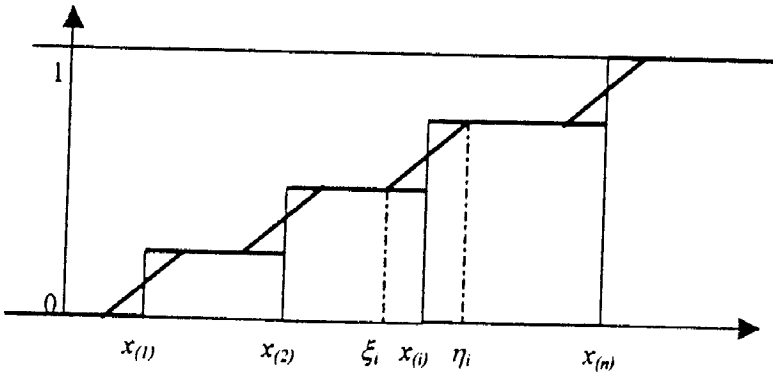

Abb. 2

3.11. Die Verallgemeinerung auf höhere Dimensionen ist möglich, Verteilungsfunktionen erweisen sich in diesem Rahmen aber als sehr umständlich. Überdies gelten manche Aussagen über Verteilungsfreiheit (vgl. **3.8.**) dann nicht mehr so wie im eindimensionalen Fall.

4. Dichteschätzung

4.1. Das Problem der Dichteschätzung wurde erst in den Fünfzigerjahren untersucht, zunächst Histogramme und Nächste-Nachbar-Schätzer. Die hier dargestellten *Kernschätzer* werden meist nach *Rosenblatt (1956)* und *Parzen (1962)* benannt, tatsächlich hat sie jedoch bereits *Akaike (1954)* definiert. Eine Einführung in das Gebiet bietet *Wertz (1978)*, einen ausführlichen Überblick *Prakasa Rao (1983)*. *Devroye, Györfi (1985)* stellen (mathematisch anspruchsvoll) sehr grundlegende Ergebnisse dar. Die Literatur über dieses Thema ist sehr umfangreich geworden, und Dichteschätzungen spielen für die meisten Anwendungsgebiete der Statistik eine wichtige Rolle. In der Folge wird der Einfachheit halber nur der eindimensionale Fall behandelt.

4.2. Eine (eindimensionale) Wahrscheinlichkeitsverteilung P mit Verteilungsfunktion F besitzt eine Dichte f, wenn gilt

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \qquad \forall\, x \in \mathbb{R}$$

mit einer nichtnegativen, integrierbaren Funktion f. An Stetigkeitspunkten x von f gilt:

$$f(x) = F'(x) \qquad \ldots(4.1)$$

Es gilt außerdem für jedes Ereignis $A \in \mathfrak{B}$ (siehe **3.2**):

$$P(A) = \int_{A} f(x)\,dx \qquad \ldots(4.2)$$

4.3. Wenn vorausgesetzt wird, daß F eine Dichte besitzt, liegt es nahe, diese direkt zu schätzen. Jede (meßbare) Funktion $\hat{f}_n : \mathbb{R}^{n+1} \to \mathbb{R}$ heißt *Dichteschätzer*. Wenn $(x_1,\ldots,x_n)$ eine Stichprobe ist (**3.5**), so dient $\hat{f}_n(x; x_1,\ldots,x_n)$ als Schätzwert für $f(x)$.
Unter der Annahme, daß $x \mapsto \hat{f}_n(x; x_1,\ldots,x_n)$ selbst eine Dichte ist, läßt sich $P(A)$ durch

$$\hat{P}_n(A) = \hat{P}_n(A; x_1,\ldots,x_n) = \int_A \hat{f}_n(x; x_1,\ldots,x_n)\,dx \qquad (A \in \mathfrak{B})$$

schätzen.

4.4. Es seien f und g Wahrscheinlichkeitsdichten und P, Q die zugehörigen Wahrscheinlichkeitsverteilungen, also es gilt *(4.2)* und

$$Q(A) = \int_A g(x)\,dx \qquad \forall A \in \mathfrak{B}.$$

Dann stellt $d(f,g) := \int_{-\infty}^{\infty} |f(x) - g(x)|\,dx$ die Fläche zwischen den beiden Dichten dar. (Abb. 3)
Es gilt:

$$\sup_{A \in \mathfrak{B}} |P(A) - Q(A)| = \tfrac{1}{2} \cdot \int |f(x) - g(x)|\,dx = \tfrac{1}{2} \cdot d(f,g) \quad . \qquad \ldots(4.3)$$

Daher stellt $d(f,g)$ ein Abstandsmaß zwischen f und g dar, das eine unmittelbare stochastische Bedeutung besitzt und übrigens gegenüber streng monotone Maßstabtransformationen invariant ist (siehe *Devroye (1979)* und vgl. *Wertz (1992)*).
Daneben ist auch eine Reihe anderer Abstandsmaße zwischen Dichten gebräuchlich; besonders bequem und häufig verwendet, aber schlecht interpretierbar ist $\int_{-\infty}^{\infty} [f(x) - g(x)]^2\,dx$.

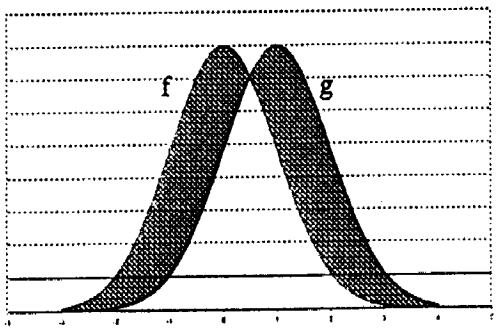

Abb. 3

4.5. Wendet man **4.4** mit $g = \hat{f}_n(\cdot; x_1,\ldots,x_n)$ an, so wird $Q = \hat{P}_n$ und *(4.3)* ergibt:

$$\sup_{A \in \mathfrak{B}} \left| \hat{P}_n(A) - P(A) \right| = \tfrac{1}{2} \cdot d(\hat{f}_n, f) \qquad \ldots(4.4)$$

Um die Verteilung P gut schätzen zu können, muß man also $d(\hat{f}_n, f)$ klein machen.

4.6. Zur Konstruktion von Schätzern $\hat{f}_n$ geht man etwa von *(4.1)* aus. $\hat{F}_n$ schätzt die Verteilungsfunktion F; da aber $\hat{F}_n$ nicht differenzierbar ist, approximiert man $f(x)=F'(x)$ durch einen Differenzenquotienten $\frac{F(x+b)-F(x-b)}{2b}$ (mit $b>0$; je kleiner b, desto besser die Approximation) und schätzt in diesem F durch $\hat{F}_n$. Es gilt also:

$$f(x) \approx \frac{F(x+b)-F(x-b)}{2b} \approx \frac{\hat{F}_n(x+b)-\hat{F}_n(x-b)}{2b}$$

Der Ausdruck rechts läßt sich in der Gestalt

$$\hat{f}_n(x;x_1,\ldots,x_n) := \frac{1}{nb}\sum_{i=1}^{n} K\left(\frac{x-x_i}{b}\right) \qquad \ldots(4.5)$$

mit $K(y)=\frac{1}{2}$ für $-1 \le y < 1$ und $=0$ sonst schreiben.

4.7. Ein Schätzer der Gestalt *(4.5)*, wo K eine Wahrscheinlichkeitsdichte (meist mit $K(t)=K(-t)$) und $b>0$ ist, heißt *Kernschätzer (Akaike-Rosenblatt-Parzen-Schätzer)*.
K heißt dabei *Kern* und b *Bandbreite* von $\hat{f}_n$. Da b in Abhängigkeit von n gewählt wird, nimmt der Schätzer diese Gestalt an:

$$\hat{f}_n(x;x_1,\ldots,x_n) := \frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{b_n}\right) \qquad \ldots(4.6)$$

4.8. *Devroye, Györfi (1985)* haben die Äquivalenz folgender Bedingungen gezeigt:

$$\lim_{n\to\infty} b_n = 0, \quad \lim_{n\to\infty} nb_n = \infty \qquad \ldots(4.7)$$

und

$$\lim_{n\to\infty} d(\hat{f}_n, f) = 0 \qquad \textit{für jede Wahrscheinlichkeitsdichte } f \qquad \ldots(4.8)$$

Damit erhält man mit Hilfe von *(4.4)* die gleichmäßige Konvergenz von $\hat{P}_n$ gegen die unbekannte Verteilung P mit Dichte f. (Also wird in dieser Situation **3.6** wesentlich verschärft.)

4.9. Die Wahl der Bandbreiten b_n hat wesentlichen Einfluß auf die Qualität von $\hat{f}_n$: Zu groß gewähltes b_n führt zu übermäßiger Glättung (Abb. 4c) und zu kleines b_n zu starken Fluktuationen (Abb. 4a).Bei Abb. 4b ist b_n optimal gewählt. Unter zusätzlichen Voraussetzungen über die Dichte f (Differenzierbarkeit und Streuungsverhalten) lassen sich (asymptotisch) opti-

male Werte für b_n bestimmen. Die Wahl des Kernes beeinflußt die Qualität der Schätzer viel weniger. Die (asymptotisch) optimale Wahl stellt den Kern

$$K_0(y) := \begin{cases} (3/4)(1-y^2) & \text{für } |y| \le 1 \\ 0 & \text{für } |y| > 1 \end{cases}$$

dar (*Epanechnikow (1969)*).
Läßt man zu, daß die Bandbreite b_n auch von den Beobachtungen $x_1,\ldots,x_n$ abhängen kann, so kann man erreichen, daß diese Dichteschätzer (*automatische Kernschätzer*) invariant gegenüber Maßstabtransformationen der Daten werden. Für die Schätzer *(4.6)* gilt nur Translationsinvarianz. Für diesen Problemkreis vgl. *Wertz (1992)* und die dort angeführte Literatur.

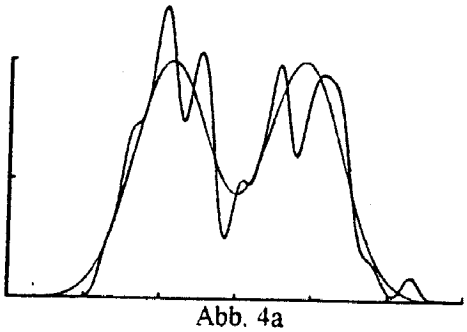
Abb. 4a

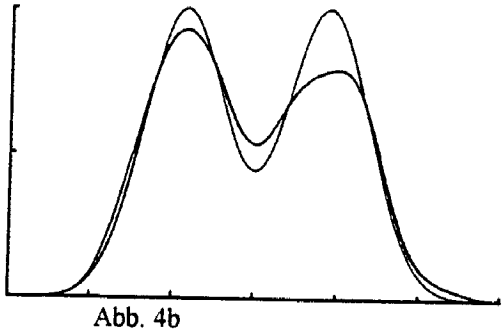
Abb. 4b

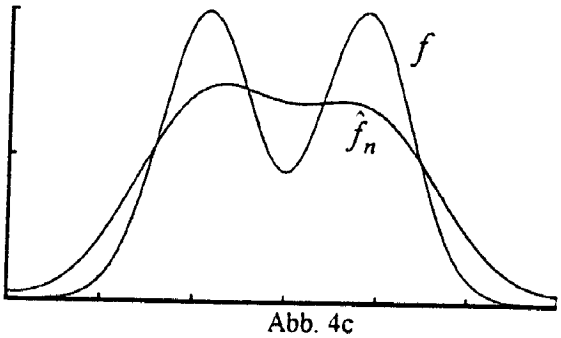

Abb. 4c

4.10. Die Schätzung von Dichten hat gegenüber der Schätzung der Verteilungsfunktion eine Reihe von Vorteilen, von denen hier einige angeführt werden.

a. Verteilungsfunktionen sind in höheren Dimensionen sehr umständlich, und es gehen wichtige Eigenschaften der empirischen Verteilungsfunktion verloren (vgl. auch die Bemerkung in **3.8**).

b. Dichten können für Wahrscheinlichkeitsverteilungen auf sehr allgemeinen Strukturen (z.B. Kugeloberfläche) definiert werden, wo kein Gegenstück zu Verteilungsfunktionen existiert.

c. Wenn eine Dichte existiert, ist es wünschenswert, als Schätzer eine Verteilung mit Dichte zu erhalten. Glättungen der empirischen Ver-

teilungsfunktion, sodaß diese differenzierbar wird, erfordern auch die Wahl eines geeigneten Glättungsparameters (vgl.**3.10**).
d. Die Wahrscheinlichkeitsdichte ist anschaulicher als die Verteilungsfunktion; Lage der Verteilung, lokale Maxima und dgl. sind leicht erkennbar.
e. Die Optimierung vieler statistischer Verfahren hängt unmittelbar von der zugrundeliegenden Dichte ab. Ist diese unbekannt, so muß sie geschätzt werden.
f. Die Verwerfungsmethode zur Erzeugung von Zufallszahlen (vgl. *Devroye, Györfi (1989)*, Kap. 8) benützt die Dichte der vorgegebenen Verteilung. Für manche Simulationsstudien ist diese zunächst aus Daten zu schätzen.

4.11. Für Dichteschätzer gilt das in **3.8** Gesagte in noch höherem Ausmaß. Das Studium der Qualität der Dichteschätzer in Abhängigkeit von der Bandbreite b_n bietet interessante Möglichkeiten, erfordert aber längere Rechenzeiten und Geschick im rationellen Umgang mit den vorhandenen EDV-Gegebenheiten. Natürlich ist auch ein genaueres Studium der theoretischen Grundlagen Voraussetzung für eine sinnvolle Arbeit im (spezialisierten) Unterricht. Da die einschlägige Fachliteratur größtenteils in englischer Sprache verfaßt ist, ergibt sich die Notwendigkeit des Einarbeitens in diese Fachterminologie.

Schrifttum

Akaike, H. (1954): An approximation to the density function.
Ann.Inst.Statist.Math., Tokyo 6, 127-132

Bauer, H. (1991): Wahrscheinlichkeitstheorie, 4. Auflage.
W. de Gruyter, Berlin

Büning, H.; Trenkler, G. (1994): Nichtparametrische Statistische Methoden.
W. de Gruyter, Berlin

Devroye, L.; Györfi, L. (1985): Nonparametric Density Estimation – The L_1 View.
J.Wiley&Sons, New York

Epanechnikov, V.A. (1969): Nonparametric estimates of a multivariate density .
Teor.Verojatnost. i Primenen. 14, 156-162

Gnedenko, B.W. (1991): Einführung in die Wahrscheinlichkeitstheorie, 4.Auflage.
Akademie Verlag, Berlin

Kolmogorow, A.N. (1933): Grundbegriffe der Wahrscheinlichkeitsrechnung.
Springer, Berlin

Parzen, E. (1962): On estimation of a probability density function and mode.
Ann.Math.Statist. 33, 1065-1076

Pfanzagl, J. (1988): Elementare Wahrscheinlichkeitsrechnung.
W. de Gruyter, Berlin

Prakasa Rao, B.L.S. (1983): Nonparametric Functional Estimation.
Academic Press, Orlando

Reichel, H.C.; Hanisch, G.; Müller, R. (1987): Wahrscheinlichkeitsrechnung und Statistik.
Hölder-Pichler-Tempsky, Wien

Rényi, A. (1962): Wahrscheinlichkeitstheorie mit einem Anhang über Informationstheorie.
VEB Deutscher Verlag der Wissenschaften, Berlin

Rosenblatt, M. (1956): Remarks on some nonparametric estimates of a density function.
Ann.Math.Statist. 27, 832-837

Schürger, K. (1998): Wahrscheinlichkeitstheorie.
R. Oldenbourg, München

von Mises, R. (1919): Grundbegriffe der Wahrscheinlichkeitsrechnung.
Math.Z. 5, 52-99

von Mises, R. (1931): Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und Theoretischen Physik, Bd.I.
Deuticke, Leipzig

Wertz, W. (1978): Statistical Density Estimation – A Survey.
Vandenhoeck & Ruprecht, Göttingen

Wertz, W. (1992): On quasi-invariant density estimation.
J.Statist.Plann.Inference 32, 271-280

Anschrift des Verfassers:
Wolfgang WERTZ
Institut für Statistik, Wahrscheinlichkeitstheorie
und Versicherungsmathematik
TU Wien
Wiedner Hauptstraße 8-10/107
A-1040 WIEN