KAPITEL 4

Dichteschätzer

Es seien $X_1, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit Dichte f und Verteilungsfunktion F. Es sei $(x_1, \ldots, x_n)$ eine Realisierung von $(X_1, \ldots, X_n)$. In diesem Kapitel beschäftigen wir uns mit dem folgenden Problem: Man schätze die Dichte f anhand der Stichprobe $(x_1, \ldots, x_n)$.

Zunächst einmal kann man die folgende Idee ausprobieren. Wir können die Verteilungsfunktion F durch die empirische Verteilungsfunktion $\widehat{F}_n$ schätzen. Die Dichte f ist die Ableitung der Verteilungsfunktion F. Somit können wir versuchen, die Dichte f durch die Ableitung von $\widehat{F}_n$ zu schätzen. Diese Idee funktioniert allerdings nicht, da die Funktion $\widehat{F}_n$ nicht differenzierbar (und sogar nicht stetig) ist. Man muss also andere Methoden benutzen.

4.1. Histogramm

Wir wollen nun das Histogramm einführen, das als ein sehr primitiver Schätzer für die Dichte aufgefasst werden kann. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Sei $c_0, \ldots, c_k$ eine aufsteigende Folge reeller Zahlen mit der Eigenschaft, dass die komplette Stichprobe $x_1, \ldots, x_n$ im Intervall (c_0, c_k) liegt. Typischerweise wählt man die Zahlen c_i so, dass die Abstände zwischen den aufeinanderfolgenden Zahlen gleich sind. In diesem Fall nennt man $h := c_i - c_{i-1}$ die Bandbreite.

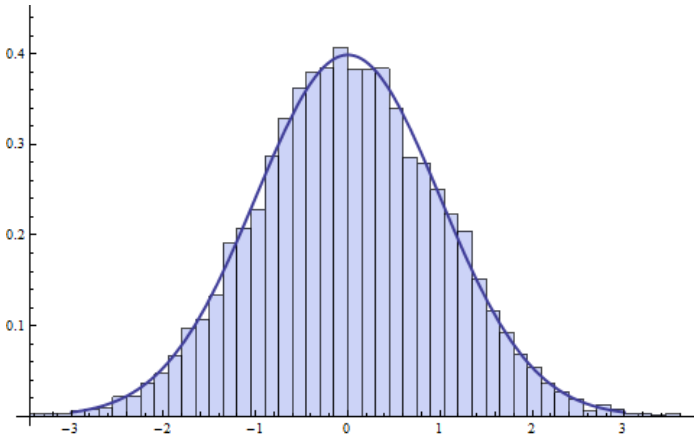

ABBILDUNG 1. Das Histogramm einer standardnormalverteilten Stichprobe vom Umfang $n = 10000$. Die glatte blaue Kurve ist die Dichte der Standardnormalverteilung.

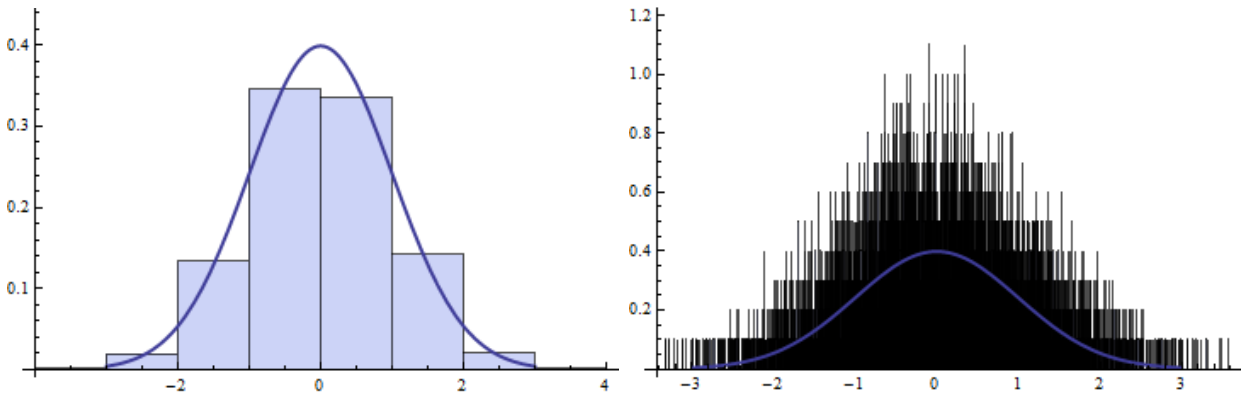

ABBILDUNG 2. Das Histogramm einer standardnormalverteilten Stichprobe vom Umfang 10000 mit einer schlecht gewählten Bandbreite $h = c_i - c_{i-1}$. Links: Die Bandbreite ist zu groß. Rechts: Die Bandbreite ist zu klein. In beiden Fällen zeigt die glatte blaue Kurve die Dichte der Standardnormalverteilung.

Die Anzahl der Stichprobenvariablen x_j im Intervall $(c_{i-1}, c_i]$ wird mit n_i bezeichnet, somit gilt

$$n_i = \sum_{j=1}^{n} \mathbb{1}_{x_j \in (c_{i-1}, c_i]}, \quad i = 1, \ldots, k.$$

Teilt man n_i durch den Stichprobenumfang n, so führt dies zur *relativen Häufigkeit*

$$f_i = \frac{n_i}{n}.$$

Als *Histogramm* wird die graphische Darstellung dieser relativen Häufigkeiten bezeichnet, siehe Abbildung 1. Man konstruiert nämlich über jedem Intervall $(c_{i-1}, c_i]$ ein Rechteck mit dem Flächeninhalt f_i. Das Histogramm ist dann die Vereinigung dieser Rechtecke. Es ist offensichtlich, dass die Summe der relativen Häufigkeiten 1 ergibt, d.h.

$$\sum_{i=1}^{k} f_i = 1.$$

Das bedeutet, dass der Flächeninhalt unter dem Histogramm gleich 1 ist. Außerdem gilt $f_i \geq 0$.

Das Histogramm hat den Nachteil, dass die Wahl der c_i's bzw. die Wahl der Bandbreite h willkürlich ist. Ist die Bandbreite zu klein oder zu groß gewählt, so kommt es zu Histogrammen, die die Dichte nur schlecht approximieren, siehe Abbildung 2. Außerdem ist das Histogramm eine lokal konstante, nicht stetige Funktion, obwohl die Dichte f meistens weder lokal konstant noch stetig ist. Im nächsten Abschnitt betrachten wir einen Dichteschätzer, der zumindest von diesem zweiten Nachteil frei ist.

4.2. Kerndichteschätzer

Wir werden nun eine bessere Methode zur Schätzung der Dichte betrachten, den Kerndichteschätzer.

DEFINITION 4.2.1. Ein *Kern* ist eine messbare Funktion $K : \mathbb{R} \to [0, \infty)$, so dass

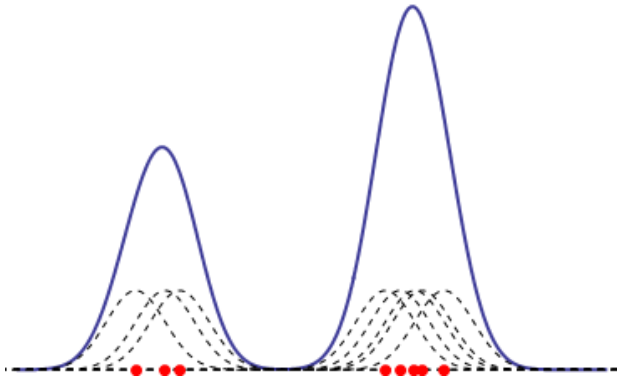

ABBILDUNG 3. Kerndichteschätzer.

(1) $K(x) \geq 0$ für alle $x \in \mathbb{R}$ und
(2) $\int_{\mathbb{R}} K(x)dx = 1$.

Die Bedingungen in der Definition eines Kerns sind somit die gleichen, wie in der Definition einer Dichte.

DEFINITION 4.2.2. Sei $(x_1, \ldots, x_n) \in \mathbb{R}^n$ eine Stichprobe. Sei K ein Kern und $h > 0$ ein Parameter, der die *Bandbreite* heißt. Der *Kerndichteschätzer* ist definiert durch

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad x \in \mathbb{R}.$$

BEMERKUNG 4.2.3. Jedem Punkt x_i in der Stichprobe wird in dieser Formel ein "Beitrag" der Form

$$\frac{1}{nh} K\left(\frac{x - x_i}{h}\right)$$

zugeordnet. Der Kerndichteschätzer $\hat{f}_n$ ist die Summe der einzelnen Beiträge. Das Integral jedes einzelnen Beitrags ist gleich $1/n$, denn

$$\int_{\mathbb{R}} \frac{1}{hn} K\left(\frac{x - x_i}{h}\right) dx = \frac{1}{n} \int_{\mathbb{R}} K(y)dy = \frac{1}{n}.$$

Um das Integral zu berechnen, haben wir dabei die Variable $y := \frac{x - x_i}{h}$ mit $dy = \frac{dx}{h}$ eingeführt. Somit ist das Integral von $\hat{f}_n$ gleich 1:

$$\int_{\mathbb{R}} \hat{f}_n(x)dx = 1.$$

Es ist außerdem klar, dass $f_n(x) \geq 0$ für alle $x \in \mathbb{R}$. Somit ist $\hat{f}_n$ tatsächlich eine Dichte.

BEMERKUNG 4.2.4. Die Idee hinter dem Kerndichteschätzer zeigt Abbildung 3. Auf dieser Abbildung ist der Kerndichteschätzer der Stichprobe

$$(-4, -3, -2.5, 4.5, 5.0, 5.5, 5.75, 6.5)$$

zu sehen. Die Zahlen aus der Stichprobe werden durch rote Kreise auf der x-Achse dargestellt. Die gestrichelten Kurven zeigen die Beiträge der einzelnen Punkte. In diesem Fall benutzen

wir den Gauß–Kern, der unten eingeführt wird. Die Summe der einzelnen Beiträge ist der Kerndichteschätzer $\hat{f}_n$, der durch die blaue Kurve dargestellt wird.

In der Definition des Kerndichteschätzers kommen zwei noch zu wählende Parameter vor: Der Kern K und die Bandbreite h. Für die Wahl des Kerns gibt es z.B. die folgenden Möglichkeiten.

Beispiel 4.2.5. Der *Rechteckskern* ist definiert durch

$$K(x) = \frac{1}{2}\mathbb{1}_{x\in[-1,1]}.$$

Der mit dem Rechteckskern assoziierte Kerndichteschätzer ist somit gegeben durch

$$\hat{f}_n(x) = \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{1}_{x_i\in[x-h,x+h]}$$

und wird auch als *gleitendes Histogramm* bezeichnet. Ein Nachteil des Rechteckskerns ist, dass er nicht stetig ist.

Beispiel 4.2.6. Der *Gauß-Kern* ist nichts Anderes, als die Dichte der Standardnormalverteilung:

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad x \in \mathbb{R}.$$

Es gilt dann

$$\frac{1}{h}K\left(\frac{x-x_i}{h}\right) = \frac{1}{\sqrt{2\pi}h}\exp\left(-\frac{(x-x_i)^2}{2h^2}\right),$$

was der Dichte der Normalverteilung $\mathrm{N}(x_i, h^2)$ entspricht. Der Kerndichteschätzer $\hat{f}_n$ ist dass das arithmetische Mittel solcher Dichten.

Beispiel 4.2.7. Der *Epanechnikov-Kern* ist definiert durch

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2), & \text{falls } x \in (-1,1), \\ 0, & \text{sonst.} \end{cases}$$

Dieser Kern verschwindet außerhalb des Intervalls $(-1,1)$, hat also einen kompakten Träger.

Beispiel 4.2.8. Der *Bisquare-Kern* ist gegeben durch

$$K(x) = \begin{cases} \frac{15}{16}(1-x^2)^2, & \text{falls } x \in (-1,1), \\ 0, & \text{sonst.} \end{cases}$$

Dieser Kern besitzt ebenfalls einen kompakten Träger und ist glatter als der Epanechnikov-Kern.

Die optimale Wahl der Bandbreite h ist ein nichttriviales Problem, mit dem wir uns in dieser Vorlesung nicht beschäftigen werden.