

Beispiel 1.4 (Displayangebot - Fortsetzung)

Für die Displayangebots-Daten, siehe das Beispiel 1.1, ergeben sich die in der Abbildung 1.2 dargestellten Diagramme. Das Stabdiagramm macht zumindest deutlich, dass die Beobachtungen im unteren Bereich stark konzentriert sind. Es gibt zudem nicht viele gleiche Werte. Größere Werte kommen nur vereinzelt vor.

R-Code 1.2 (Einfache Diagramme für univariate Daten)

```
stem(y)           # Stem-and-Leaf-Diagramm
par(mfrow=c(1,2)) # Halbieren des Grafik-Fensters
plot(table(y))    # Stabdiagramm
hist(y,breaks=19) # Histogramm
```

Stem-and-Leaf-Diagramm, Stabdiagramm und Histogramm lassen sich in R sehr leicht erstellen. Die Displaydaten sind dabei in der Variablen `y` gespeichert. `stem` zählt nicht zu den eigentlichen Grafikfunktionen; die Ausgabe erfolgt auch im Konsolen- und nicht im Grafikfenster.

Für das Stabdiagramm wird mit dem dritten Aufruf zunächst mit `table(y)` eine Häufigkeitstabelle erstellt. Die grafische Darstellung mittels `plot` ergibt das gewünschte. Beim Histogramm wird über `breaks=19` die Anzahl der Klassen festgelegt. Auch die explizite Angabe der Klassengrenzen wäre möglich. Die Grafik ist in der Abbildung 1.2 wiedergegeben.

Das # dient als Kommentarzeichen; alles, was auf einer Zeile hinter dem #-Zeichen steht, wird ignoriert.

Das Histogramm ist unstetig und hängt sehr stark von der Wahl der Klassengrenzen ab. Als naheliegende Verbesserung bietet sich daher an, die Klassen über den Bereich der Beobachtungen 'gleiten' zu lassen. Das führt bei einer festen Klassenbreite h zu

$$\hat{p}(y) = \frac{1}{h \cdot n} \sum_{i=1}^n I_{(y-h/2, y+h/2]}(y_i) = \frac{1}{h \cdot n} \sum_{i=1}^n I_{(-1/2, 1/2]} \left(\frac{y - y_i}{h} \right).$$

Das Resultat ist i.d.R. von sehr unruhiger Gestalt. Eine Verbesserung im Sinne eines glatteren Funktionsverlaufes erhält man durch die Ersetzung der Indikatorfunktion durch eine stetige Funktion, einen sogenannten *Kern* $K(u)$:

$$\hat{p}(y) = \frac{1}{h \cdot n} \sum_{i=1}^n K \left(\frac{y - y_i}{h} \right). \quad (1.5)$$

Dabei muss $K(u)$ die gleichen Eigenschaften wie $I_{(-1/2, 1/2]}(u)$ haben, nämlich $K(u) \geq 0$ und $\int_{-\infty}^{\infty} K(u) du = 1$. Diese beiden Eigenschaften zeichnen gerade Dichtefunktionen aus. Für $K(u)$ wird daher oft die Dichte der Standardnormalverteilung genommen. Eine andere Möglichkeit ist der Epanechnikov-Kern. Das Resultat wird als *Kerndichteschätzung* bezeichnet. Die Wahl der Bandbreite h ist ein ähnlich kritischer Punkt wie die Klassenbreite beim Histogramm. Hier gibt Silverman (1986, S.43ff) eine gute Diskussion.

Beispiel 1.5 (Displayangebot - Fortsetzung)

In der Abbildung 1.3 ist der Normalverteilungskern mit der Kerndichteschätzung der Displaydaten aus dem Beispiel 1.1 dargestellt. Die Bandbreite wurde automatisch bestimmt.

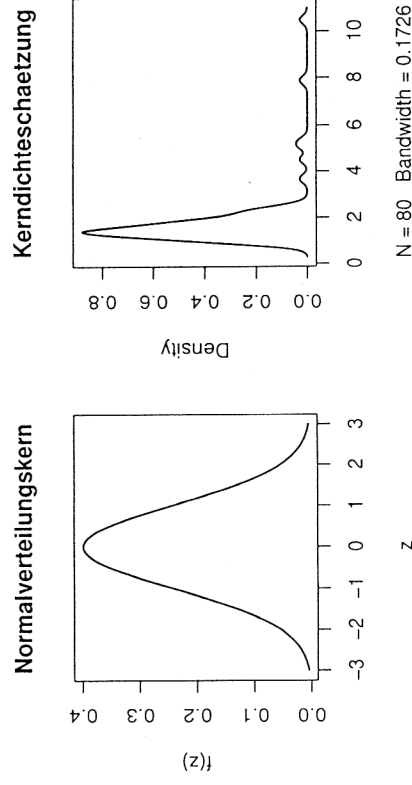


Abbildung 1.3: Displayangebot

R-Code 1.3 (Kerndichteschätzung)

```
d<-density(y, kernel="gaussian")
z<-seq(-3,3,.1)
f<-dnorm(z,mean=0,sd=1)
par(mfrow=c(1,2))
plot(z,f,type="l",lwd=1.5,ylab="f(z)",main="Normalverteilungskern")
plot(d,lwd=1.5,main="Kerndichteschätzung")
```

Die Displaydaten sind in der Variablen `y` gespeichert. In der ersten Zeile wird die Kerndichteschätzung durchgeführt. Die Bandbreite wird nach einem geeigneten Kriterium automatisch bestimmt. Für die Darstellung werden die Punkte auf der Abszisse in der zweiten Zeile erzeugt und der Variablen `z` zugewiesen. Die Funktion `dnorm` bestimmt dann die zugehörigen Werte der Normalverteilungsdichte mit Erwartungswert 0 und Standardabweichung 1. Der Parameter `type="l"` verlangt Verbindungslinien. Die Punkte auf der Abszisse werden die Zeichenkette kenntlich. Das Resultat ist die Abbildung 1.3.

Bedeutsam ist weiterhin die grafische Darstellung der empirischen Verteilungsfunktion $\hat{F}(y)$. Sie gibt die relative Häufigkeit der Beobachtungen, die kleiner oder gleich