Kapitel 3: Deskriptive und explorative Statistik

geg.: Messreihe (Stichprobe, Datensatz):

$x_1, \ldots, x_n$ (n=Stichprobenumfang)

Aufgabe der deskriptiven (beschreibenden) Statistik:

Übersichtliche Darstellung von Eigenschaften dieser Messreihe.

Kapitel 3: Deskriptive und explorative Statistik

geg.: Messreihe (Stichprobe, Datensatz):

$x_1, \ldots, x_n$ (n=Stichprobenumfang)

Aufgabe der deskriptiven (beschreibenden) Statistik:

Übersichtliche Darstellung von Eigenschaften dieser Messreihe.

Aufgabe der explorativen (erforschenden) Statistik:

Finden von (unbekannten) Strukturen.

Beispiel 1: Beschäftigungsquote der Männer zwischen 15 und 64 Jahren in 26 Ländern der europäischen Union im Jahr 2006 (Quelle: Eurostat):

67, 63.3, 73.6, 80.6, 72.5, 71.3, 77.3, 74.6, 76, 68.5, 71.1, 79.6, 68.2,
66.4, 63.9, 73.8, 80.8, 77, 60.2, 74, 65.2, 70.8, 66.9, 71.7, 75.5, 77

Beispiel 1: Beschäftigungsquote der Männer zwischen 15 und 64 Jahren in 26 Ländern der europäischen Union im Jahr 2006 (Quelle: Eurostat):

67, 63.3, 73.6, 80.6, 72.5, 71.3, 77.3, 74.6, 76, 68.5, 71.1, 79.6, 68.2,
66.4, 63.9, 73.8, 80.8, 77, 60.2, 74, 65.2, 70.8, 66.9, 71.7, 75.5, 77

Beispiel 2: Beschäftigungsquote der Frauen zwischen 15 und 64 Jahren in 26 Ländern der europäischen Union im Jahr 2006 (Quelle: Eurostat):

53.2, 55, 56.8, 73.2, 61.4, 66.4, 58.8, 47.5, 53.2, 57.7, 46.7, 59.8, 62.9,
61.1, 51.1, 34.6, 67.5, 63, 47.8, 62.4, 54.1, 63.3, 51.6, 68.1, 70.6, 65.8

Beispiel 3: Alter der ca. 32 Millionen männlichen Einwohner unter 95 Jahren im früheren Bundesgebiet der BRD im Jahr 2001 (Quelle: Statistisches Bundesamt, Angabe in Jahren):

79, 2, 34, . . .

Typen von Messgrößen (Merkmalen, Variablen):

Typen von Messgrößen (Merkmalen, Variablen):

1. mögliche Unterteilung:

- diskret: endlich oder abzählbar unendlich viele Ausprägungen
- stetig: alle Werte eines Intervalls sind Ausprägungen

2. mögliche Unterteilung:

	Abstandbegriff vorhanden ?	Ordnungsrelation vorhanden ?

2. mögliche Unterteilung:

	Abstandbegriff vorhanden ?	Ordnungsrelation vorhanden ?
reell	ja	ja
ordinal	nein	ja
zirkulär	ja	nein
nominal	nein	nein

3.1 Histogramme

Häufigkeitstabelle:

- Einteilung der Daten in k Klassen (z.B. $k \approx \sqrt{n}$ oder $k \approx 10 \cdot \log_{10} n$),

3.1 Histogramme

Häufigkeitstabelle:

- Einteilung der Daten in k Klassen (z.B. $k \approx \sqrt{n}$ oder $k \approx 10 \cdot \log_{10} n$),

- Ermittlung der Klassenhäufigkeiten n_i $(i = 1, \ldots, k)$,

3.1 Histogramme

Häufigkeitstabelle:

- Einteilung der Daten in k Klassen (z.B. $k \approx \sqrt{n}$ oder $k \approx 10 \cdot \log_{10} n$),

- Ermittlung der Klassenhäufigkeiten n_i $(i = 1, \ldots, k)$,

- Darstellung des Resultats in einer Tabelle.

Klasse	Häufigkeit
1	n_1
2	n_2
$\vdots$	$\vdots$
k	n_k

In Beispiel 3 oben (Alter der ca. 32 Millionen männlichen Einwohner unter 95 Jahren im Jahr 2001, Quelle: Statistisches Bundesamt):

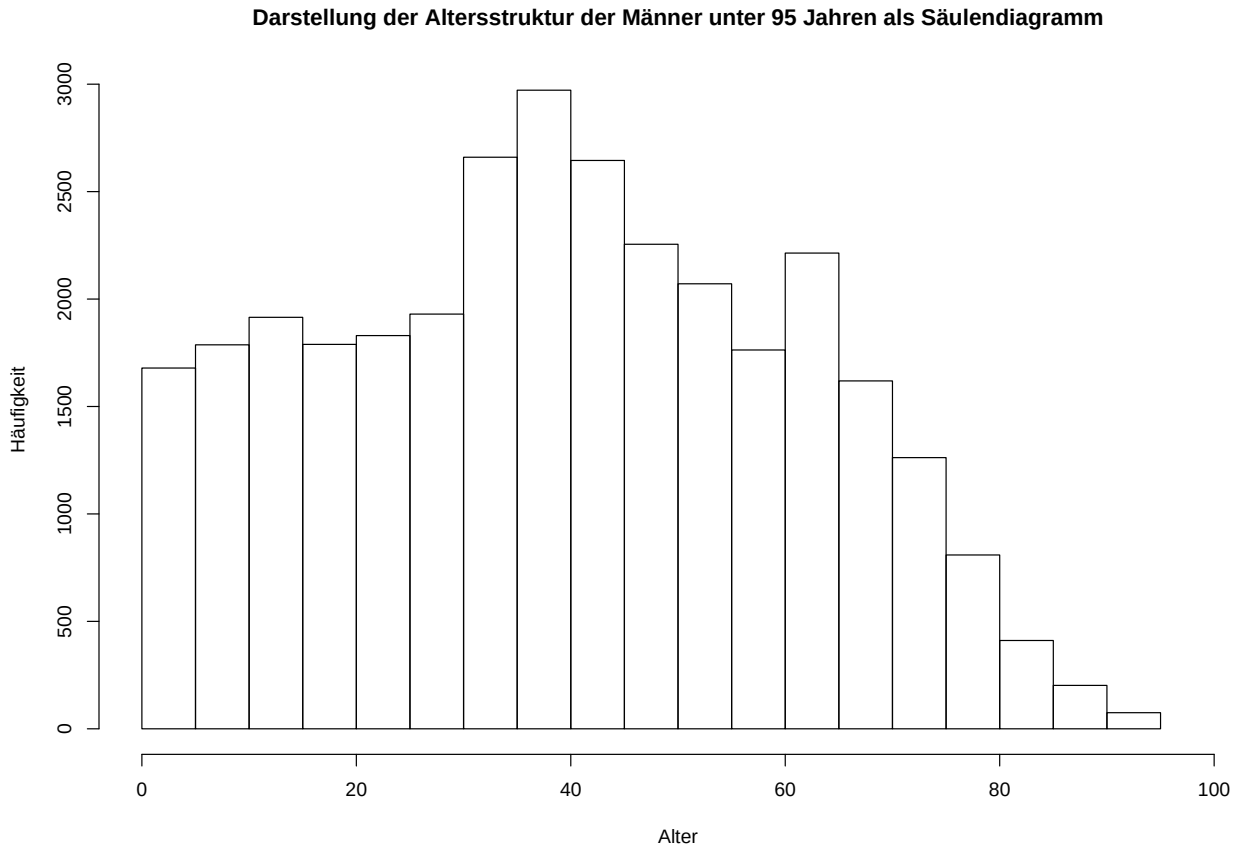
Unterteilung in 19 Klassen ergibt

In Beispiel 3 oben (Alter der ca. 32 Millionen männlichen Einwohner unter 95 Jahren im Jahr 2001, Quelle: Statistisches Bundesamt):

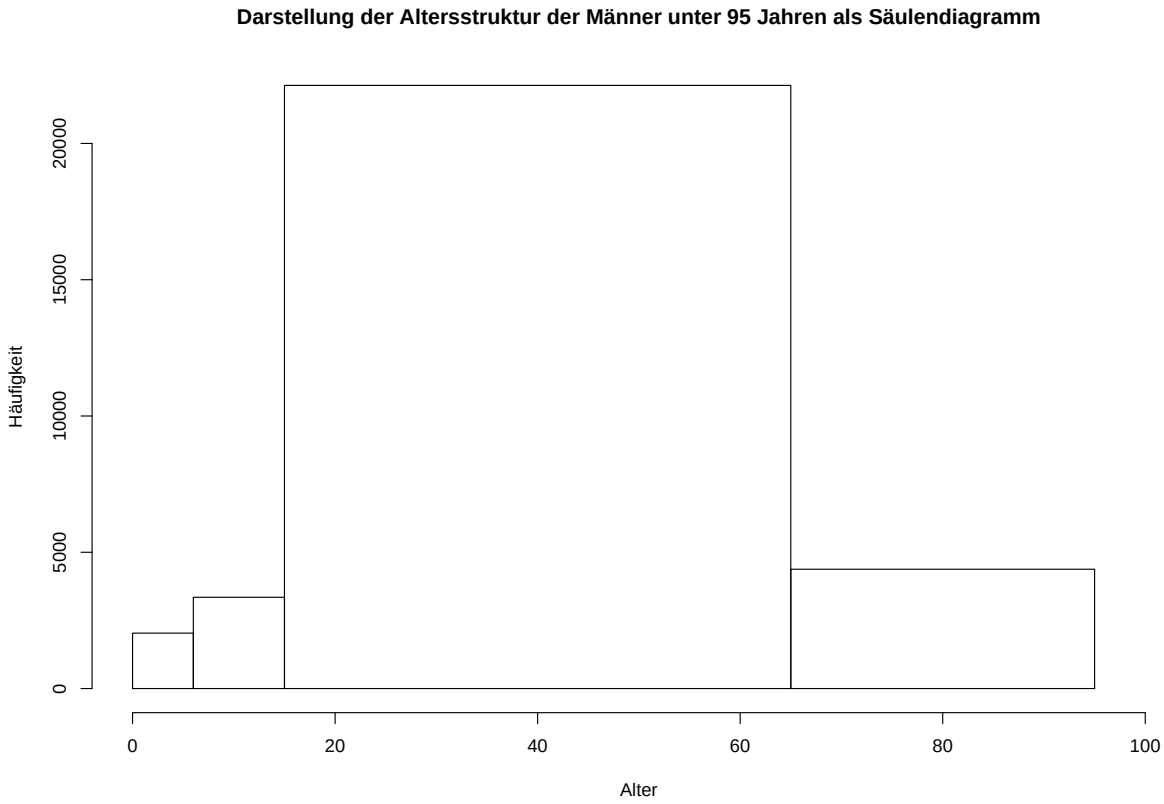
Unterteilung in 19 Klassen ergibt

Alter	Anzahl (in Tausenden)
$[0, 5)$	1679.3
$[5, 10)$	1787.2
$[10, 15)$	1913.2
$[15, 20)$	1788.7
$\vdots$	$\vdots$
$[65, 70)$	1618.4
$[70, 75)$	1262.2
$[75, 80)$	808.4
$[80, 85)$	411.9
$[85, 90)$	202.4
$[90, 95)$	73.9

Graphische Darstellung als Säulendiagramm:



Irreführend, falls die Klassen nicht alle gleich lang sind und die Klassenbreiten mit dargestellt werden:



Histogramm:

Im Gegensatz zum Säulendiagramm wird hier auch die Breite der Klassen mit berücksichtigt.

Histogramm:

Im Gegensatz zum Säulendiagramm wird hier auch die Breite der Klassen mit berücksichtigt.

Vorgehen:

- Unterteile Wertebereich der (reellen) Messgröße in k Intervalle $I_1, \ldots I_k$.

Histogramm:

Im Gegensatz zum Säulendiagramm wird hier auch die Breite der Klassen mit berücksichtigt.

Vorgehen:

- Unterteile Wertebereich der (reellen) Messgröße in k Intervalle $I_1, \ldots I_k$.
- Bestimme für jedes Intervall I_j die Anzahl n_j der Datenpunkte in diesem Intervall.

Histogramm:

Im Gegensatz zum Säulendiagramm wird hier auch die Breite der Klassen mit berücksichtigt.

Vorgehen:

- Unterteile Wertebereich der (reellen) Messgröße in k Intervalle $I_1, \ldots I_k$.
- Bestimme für jedes Intervall I_j die Anzahl n_j der Datenpunkte in diesem Intervall.
- Trage über I_j den Wert

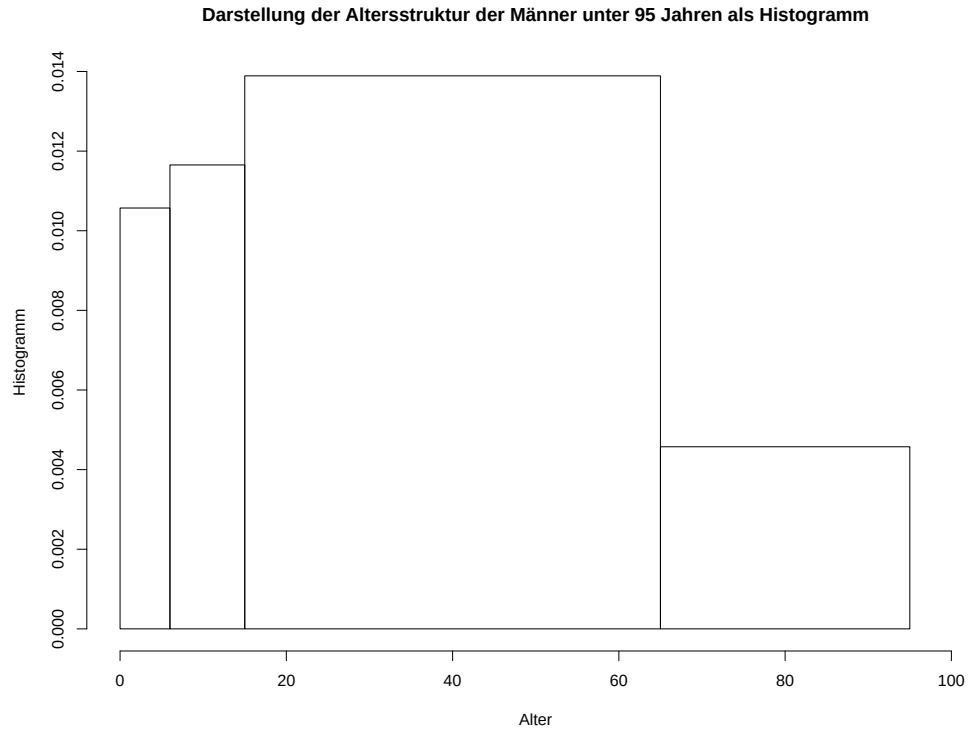
$$\frac{n_j}{n \cdot \lambda(I_j)}$$

  auf, wobei $\lambda(I_j) =$ Länge von I_j.

Bemerkung: Flächeninhalt eines Rechtecks ist gleich dem prozentualen Anteil der Datenpunkte im zugrunde liegenden Intervall.

Bemerkung: Flächeninhalt eines Rechtecks ist gleich dem prozentualen Anteil der Datenpunkte im zugrunde liegenden Intervall.

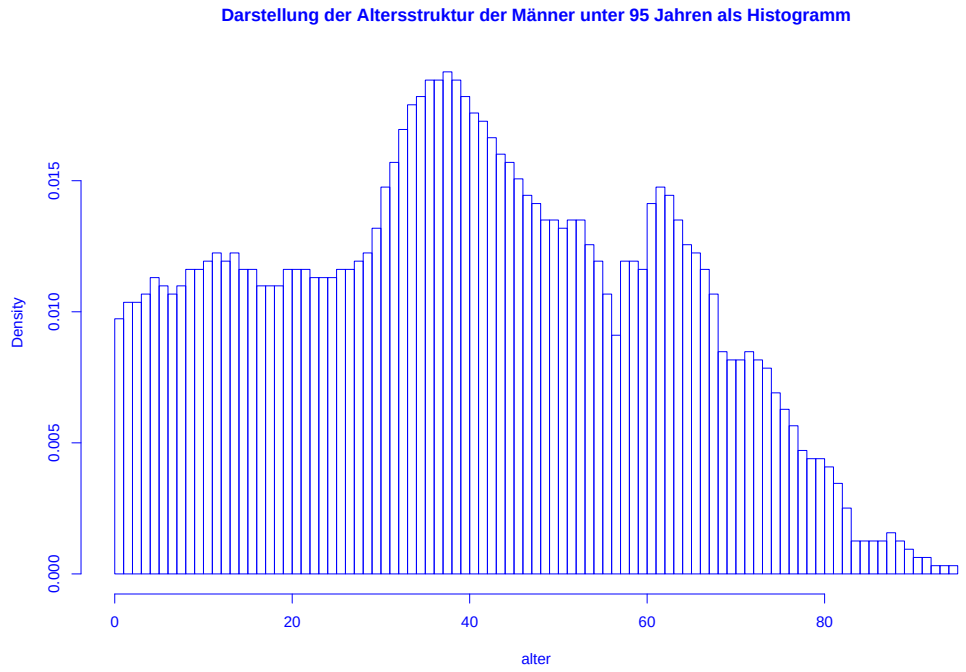
In Beispiel 3 oben erhält man



3.2 Dichteschätzung

Nachteil des Histogramms:

Unstetigkeit erschwert Interpretation zugrunde liegender Strukturen.



Ausweg:

Beschreibe Lage der Daten durch "glatte" Funktion.

Ausweg:

Beschreibe Lage der Daten durch "glatte" Funktion.

Wie bisher soll gelten:

- Funktionswerte nichtnegativ.
- Flächeninhalt Eins.
- Fläche über Intervall ungefähr proportional zur Anzahl Datenpunkte in dem Intervall.

Definition: *Eine Funktion* $f : \mathbb{R} \to \mathbb{R}$ *mit*

$$f(x) \geq 0 \quad \textit{für alle } x \in \mathbb{R}$$

und

$$\int_{\mathbb{R}} f(x)\, dx = 1$$

heißt Dichte.

Definition: *Eine Funktion* $f : \mathbb{R} \to \mathbb{R}$ *mit*

$$f(x) \geq 0 \quad \textit{für alle } x \in \mathbb{R}$$

und

$$\int_{\mathbb{R}} f(x)\, dx = 1$$

heißt Dichte.

Ziel: Beschreibe Lage der Daten durch glatte Dichtefunktion.

Anpassung einer Dichtefunktion an Daten:

1. Schritt: Gleitendes Histogramm.

Anpassung einer Dichtefunktion an Daten:

1. Schritt: Gleitendes Histogramm.

$$f_h(x) \quad = \quad \frac{\frac{1}{n} \cdot \text{Anzahl Datenpunkte } x_i \text{ in } [x-h, x+h]}{2h}$$

Anpassung einer Dichtefunktion an Daten:

1. Schritt: Gleitendes Histogramm.

$$
\begin{aligned}
f_h(x) &= \frac{\frac{1}{n} \cdot \text{Anzahl Datenpunkte } x_i \text{ in } [x-h, x+h]}{2h} \\
&= \frac{1}{n \cdot h} \sum_{i=1}^{n} \frac{1}{2} \cdot 1_{[x-h,x+h]}(x_i).
\end{aligned}
$$

Anpassung einer Dichtefunktion an Daten:

1. Schritt: Gleitendes Histogramm.

$$\begin{aligned} f_h(x) &= \frac{\frac{1}{n} \cdot \text{Anzahl Datenpunkte } x_i \text{ in } [x-h, x+h]}{2h} \\ &= \frac{1}{n \cdot h} \sum_{i=1}^{n} \frac{1}{2} \cdot 1_{[x-h,x+h]}(x_i). \end{aligned}$$

Mit

$$1_{[x-h,x+h]}(x_i) = 1 \Leftrightarrow x - h \leq x_i \leq x + h \Leftrightarrow -1 \leq \frac{x - x_i}{h} \leq 1$$

erhält man

$$f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

mit Dichte

$$K(u) = \frac{1}{2} \cdot 1_{[-1,1]}(u).$$

erhält man

$$f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

mit Dichte

$$K(u) = \frac{1}{2} \cdot 1_{[-1,1]}(u).$$

Deutung: Mittelung von Dichtefunktionen, die um die einzelnen Datenpunkte konzentriert sind.

2. Schritt: Verallgemeinerung.

2. Schritt: Verallgemeinerung.

$$f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

mit $h > 0$ (sog. **Bandbreite**) und beliebiger Dichte $K : \mathbb{R} \to \mathbb{R}$ (sog. **Kernfunktion**) heißt **Kern-Dichteschätzer**.

2. Schritt: Verallgemeinerung.

$$f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

mit $h > 0$ (sog. **Bandbreite**) und beliebiger Dichte $K : \mathbb{R} \to \mathbb{R}$ (sog. **Kernfunktion**) heißt **Kern-Dichteschätzer**.

Z.B. Epanechnikov-Kern:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{für} \quad -1 \leq u \leq 1, \\ 0 & \text{für} \quad u < -1 \text{ oder } u > 1, \end{cases}$$

2. Schritt: Verallgemeinerung.

$$f_h(x) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

mit $h > 0$ (sog. **Bandbreite**) und beliebiger Dichte $K : \mathbb{R} \to \mathbb{R}$ (sog. **Kernfunktion**) heißt **Kern-Dichteschätzer**.

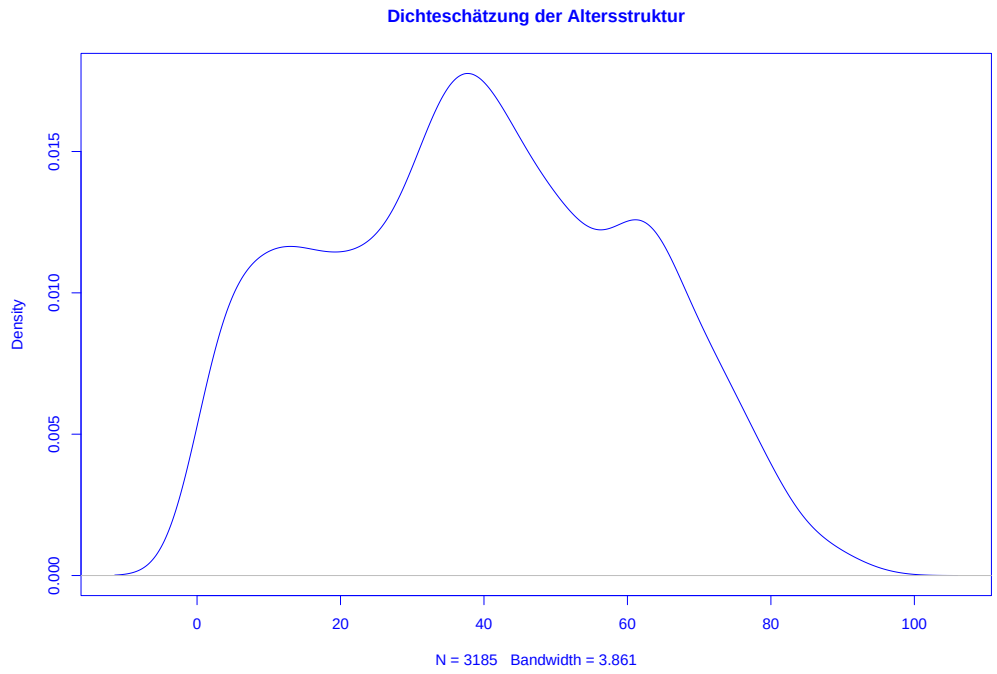
Z.B. Epanechnikov-Kern:

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & \text{für} \quad -1 \leq u \leq 1, \\ 0 & \text{für} \quad u < -1 \text{ oder } u > 1, \end{cases}$$

oder Gauss-Kern: $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right)$.

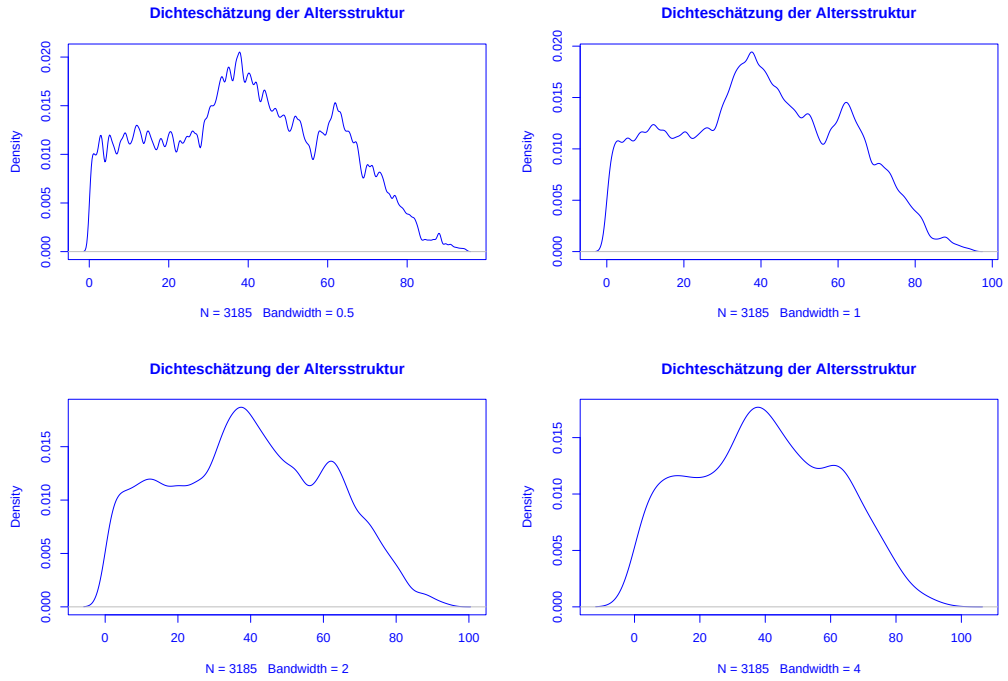
In Beispiel 3 (Altersverteilung der männlichen Einwohner unter 95 Jahren im früheren Bundesgebiet der BRD (ohne Berlin-West) im Jahr 2001) erhält man als Schätzung der Dichte:

In Beispiel 3 (Altersverteilung der männlichen Einwohner unter 95 Jahren im früheren Bundesgebiet der BRD (ohne Berlin-West) im Jahr 2001) erhält man als Schätzung der Dichte:



Mittels h lässt sich die "Glattheit" des Kern-Dichteschätzers $f_h(x)$ kontrollieren:

Mittels h lässt sich die "Glattheit" des Kern-Dichteschätzers $f_h(x)$ kontrollieren:



Ist h sehr klein, so wird $f_h(x)$ als Funktion von x sehr stark schwanken, ist dagegen h groß, so variiert $f_h(x)$ als Funktion von x kaum noch.