Methoden der Statistik
Kapitel 2: Deskriptive Statistik

Thorsten Dickhaus

Weierstraß-Institut für Angewandte Analysis und Stochastik Berlin

Wintersemester 2013/2014

Übersicht

Übersicht

Univariate Daten: Michelson's Lichtgeschwindigkeits-Daten

1	850	1	1
2	740	2	1
3	900	3	1
4	1070	4	1
5	930	5	1
6	850	6	1
7	950	7	1
8	980	8	1
9	980	9	1
10	880	10	1
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Interpretation der Daten

1	850	1	1
2	740	2	1
3	900	3	1
4	1070	4	1
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

Erste Spalte : Fortlaufende Nummer der Messungen (1-100)
Zweite Spalte : (Gemessene Geschwindigkeit - 299.000) in km/s
Dritte Spalte : Fortlaufende Nummer in der Messreihe (1-20)
Vierte Spalte : Nummer der Messreihe (1-5)

Einlesen der Daten

```
> l<-read.table("lightspeed.dat")
> str(l)
'data.frame':   100 obs. of  4 variables:
 $ V1: int  1 2 3 4 5 6 7 8 9 10 ...
 $ V2: int  850 740 900 1070 930 850 950 980 980 880 ...
 $ V3: int  1 2 3 4 5 6 7 8 9 10 ...
 $ V4: int  1 1 1 1 1 1 1 1 1 1 ...
> attributes(l)
> dim(l)
[1] 100   4
> is.matrix(l)
[1] FALSE
> is.list(l)
[1] TRUE
> mode(l)
[1] "list"
> speed<-l$V2
```

Variablen

```
> names(l)
[1] "V1" "V2" "V3" "V4"
> names(l)<-c("No","Speed","ExNo","Ex")
> attach(l)

    The following object(s) are masked from l ( position 3 ) :

     Ex ExNo No Speed

> ex1<-subset(l,Ex==1)
> s<-ex1$Speed
```

Statistische Kenngrößen

(Arithmetischer) Mittelwert $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$:

```
> mean(s) [1] 909
```

Standardabweichung $\sqrt{(1/(n-1)) \sum_i (x_i - \bar{x})^2}$:

```
> sd(s) [1] 104.9260
```

Median $med = x_{[(n+1)/2]}$:

```
> median(s) [1] 940
```

Median absoluter Abweichungen $MAD = n^{-1} \sum_i |x_i - med(\mathbf{x})|$:

```
> mad(s) [1] 88.956
```

Statistische Kenngrößen

Schiefe (skewness):

$$v(X) = \frac{\mathbb{E}\left[(X - \mathbb{E}X)^3\right]}{\mathbb{V}\mathrm{ar}(X)^{3/2}} \, .$$

Die Schiefe einer empirischen Verteilung:

$$v_e(\mathbf{x}) = \frac{n^{-1} \sum_i (x_i - \bar{x})^3}{\left(n^{-1} \sum_i (x_i - \bar{x})^2\right)^{3/2}}$$

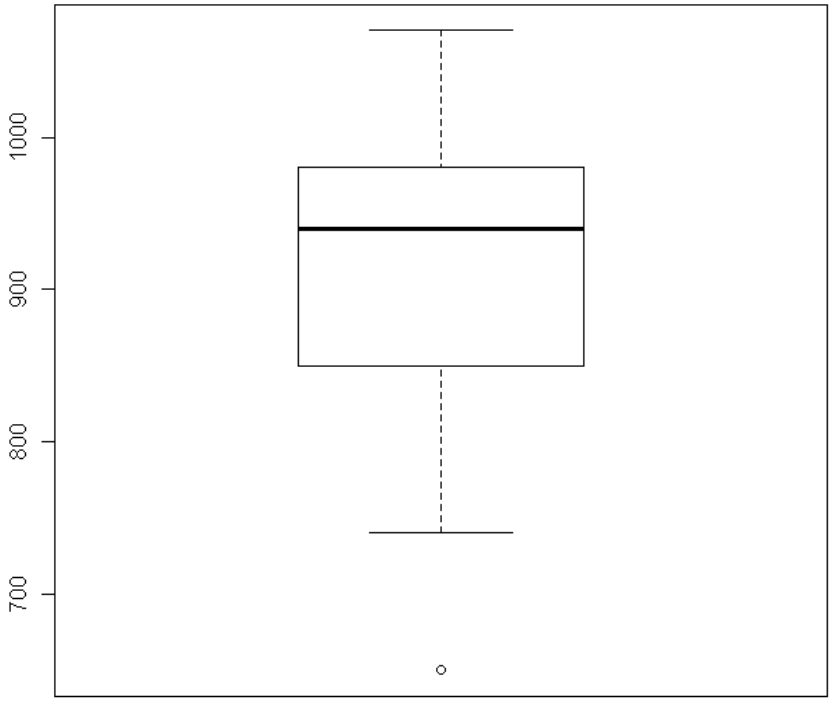
```
> skew<-function(x){
+ skewness <-  ((sqrt(length(x))*
+     sum((x-mean(x))^3))/(sum((x-mean(x))^2))^(3/2))
+ return(skewness)}
> skew(s)
[1] -0.890699
```

Die **summary**–Funktion

```
> summary(ex1)
       No              Speed            ExNo            Ex
 Min.   : 1.00   Min.   : 650   Min.   : 1.00   Min.   :1
 1st Qu.: 5.75   1st Qu.: 850   1st Qu.: 5.75   1st Qu.:1
 Median :10.50   Median : 940   Median :10.50   Median :1
 Mean   :10.50   Mean   : 909   Mean   :10.50   Mean   :1
 3rd Qu.:15.25   3rd Qu.: 980   3rd Qu.:15.25   3rd Qu.:1
 Max.   :20.00   Max.   :1070   Max.   :20.00   Max.   :1
```

Der Box–Whisker–Plot

>**boxplot**(s)



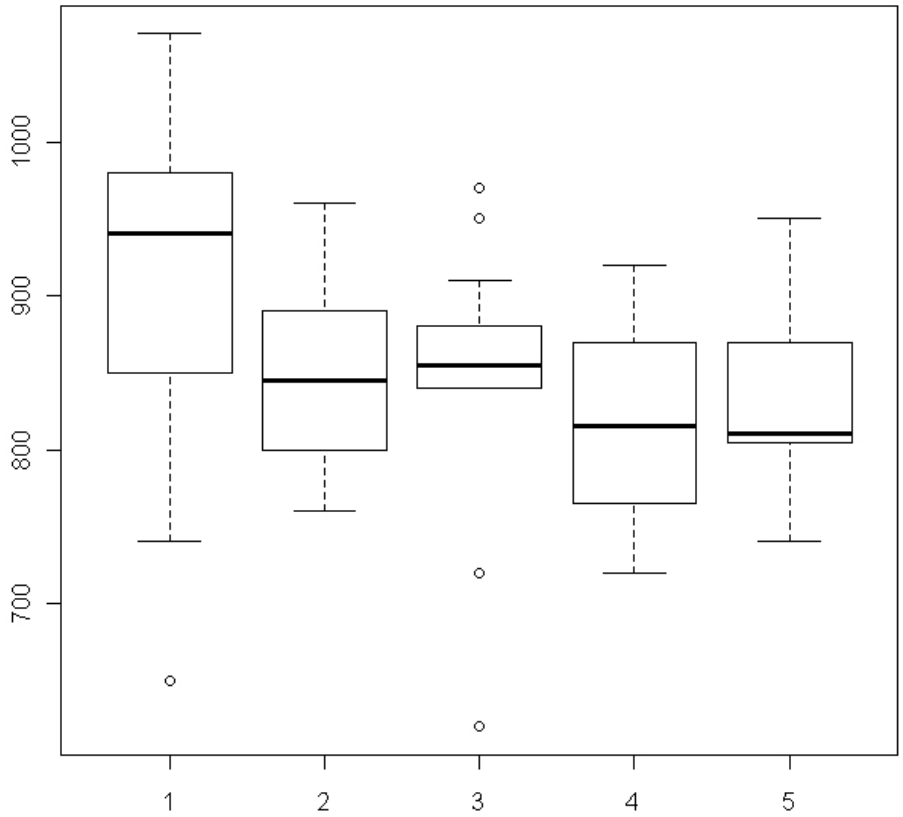
R Code: Herausnehmen von Ausreißern

```
> strim<-s[which(s>700) ]
> summary(strim)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
740.0	865.0	950.0	922.6	980.0	1070.0

Vergleich der Messreihen

```
> boxplot(l$Speed~l$Ex)
```



Empirische Verteilungsfunktion

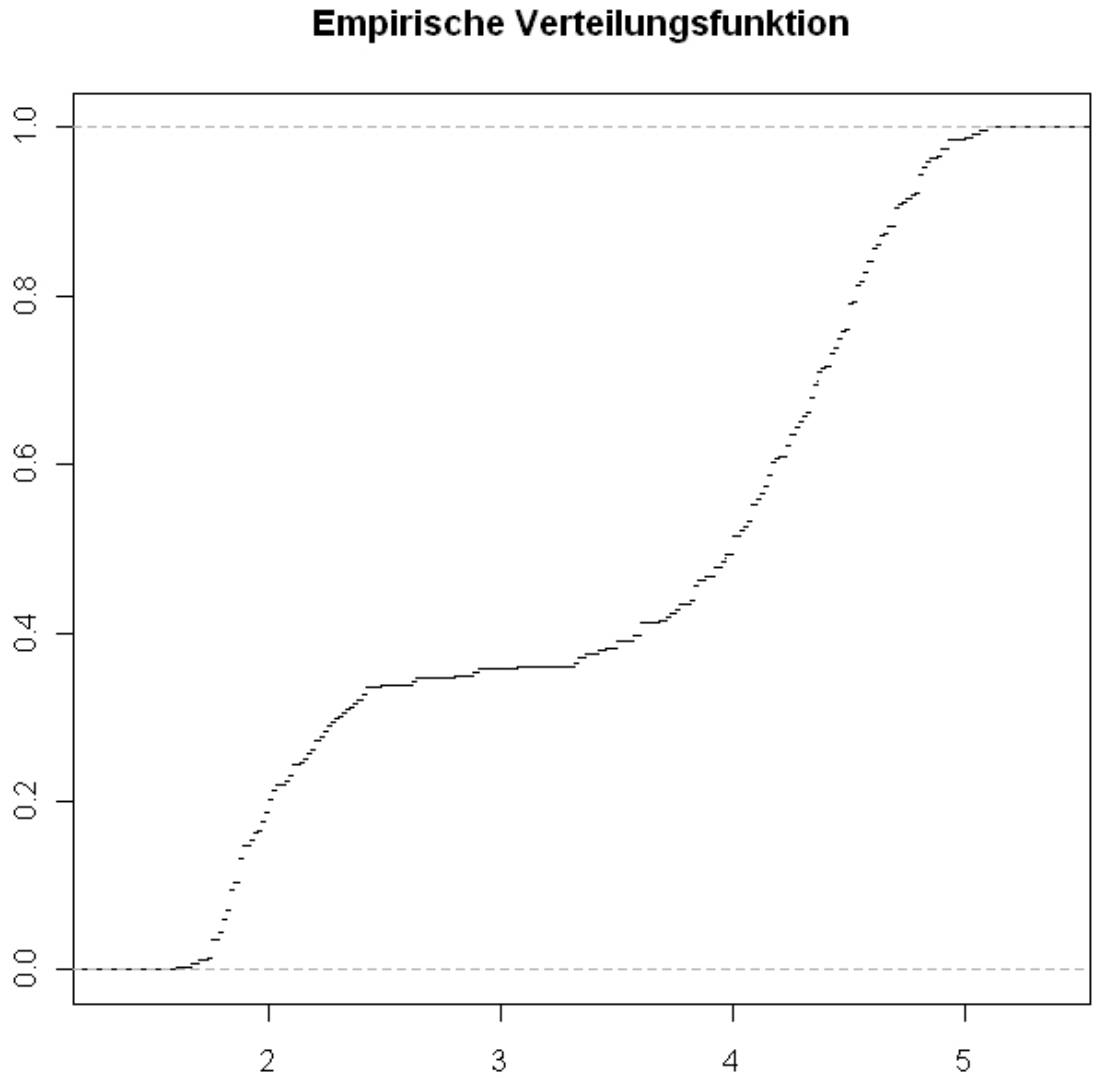
Seien $X_1, \ldots, X_n$ reellwertige iid. Zufallsvariablen mit $X_1 \sim F$ und $X = (X_1, \ldots, X_n)^t$.

$$\hat{F}_n(t) := \frac{\#\{x_i | x_i \leq t,\ i \in \{1, \ldots, n\}\}}{n} = \sum_{i=1}^{n} \frac{1}{n} \mathbb{1}_{(-\infty, t]}(x_i) \ .$$

Satz von Glivenko–Cantelli liefert fast sichere gleichmäßige Konvergenz:

$$\lim_{n \to \infty} \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F(t) \right| = 0 \quad \mathbb{P}_F - f.s.$$

```
> ecdf(er)
Empirical CDF
Call: ecdf(er)     #er: eruptions of a geysir
```


Empirische Verteilungsfunktion
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5

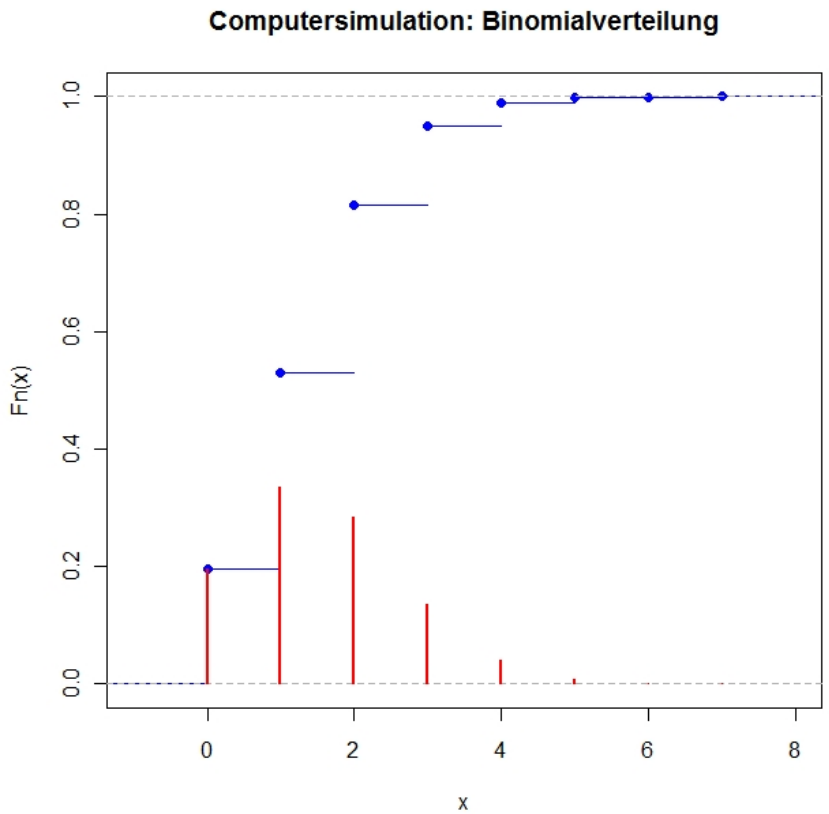
Diskrete Merkmale, Stabdiagramme

Die empirische Verteilungsfunktion ist eine rechtsseitig stetige, monoton wachsende Treppenfunktion, die an den Beobachtungspunkten springt.

Ist X_1 diskret verteilt, so ist $\mathcal{L}(X_1)$ festgelegt durch seine Wahrscheinlichkeitsfunktion, also durch die Angabe der Werte $\mathbb{P}_F(X_1 = k), k \in \text{supp}(X_1)$.

Auf der beschreibenden Ebene (empirisches Maß) führt das zu Stabdiagrammen der relativen Häufigkeiten der beobachteten Werte.

```
# Stabdiagramm und empirische Verteilungsfunktion
simanz = 5000; werte <- rbinom(n=simanz, size=10, prob=0.15)
plot(ecdf(werte), col='blue',
    main='Computersimulation: Binomialverteilung')
lines(sort(unique(werte)), table(werte)/simanz,
    type='h', col='red', lwd=2)
```



Stetiges Merkmal

Modellannahme:

$X_1, \ldots, X_n$ reellwertige iid. Zufallsvariablen, deren Verteilung die Dichte f bezüglich des Lebesgue–Maßes besitzt.

Datenbeispiel:

272 beobachtete Ausbrüche des “Old Faithful”–Geysirs im Yellowstone National Park mit Eruptionsdauer sowie der Wartezeit bis zum nächsten Ausbruch

```
> data(faithful)
> er<-faithful$eruptions
```

Histogramm–Schätzer

Das Histogramm ist ein stückweise konstanter Dichteschätzer.

Vorgehen: Wähle Intervalle („Klassen“, englisch: bins) I_k

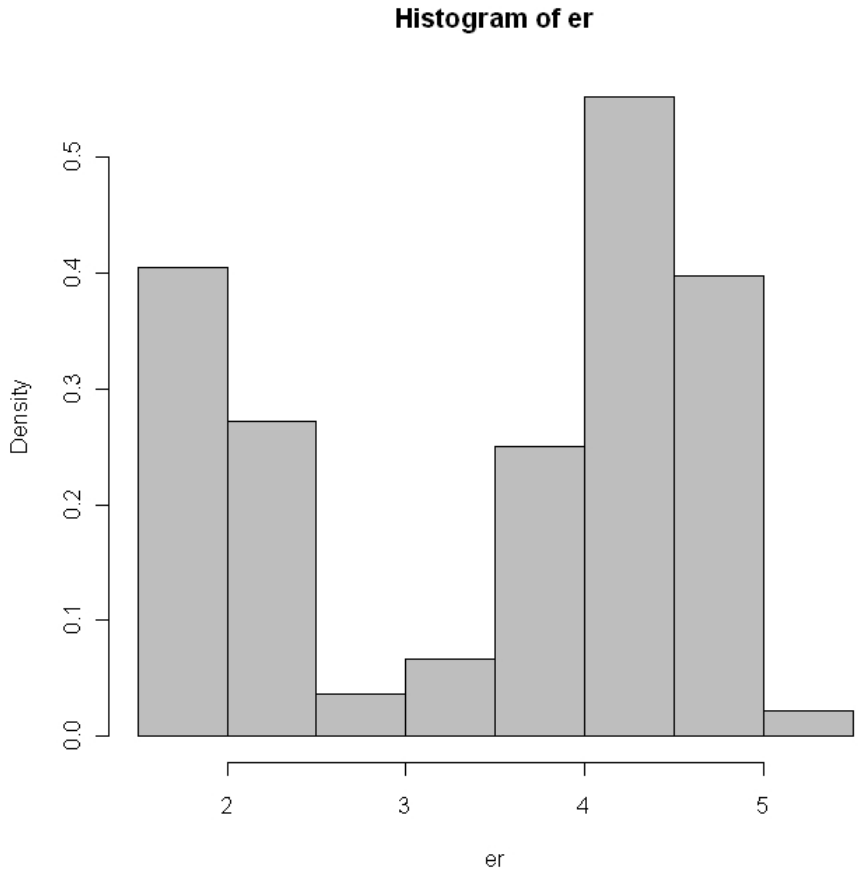
$$I_k = (a_{k-1}, a_k],\ k \in \{1, \ldots, K\}$$

$$n_k := \#\{x_i \in I_k,\ i \in \{1, \ldots, n\}\}$$

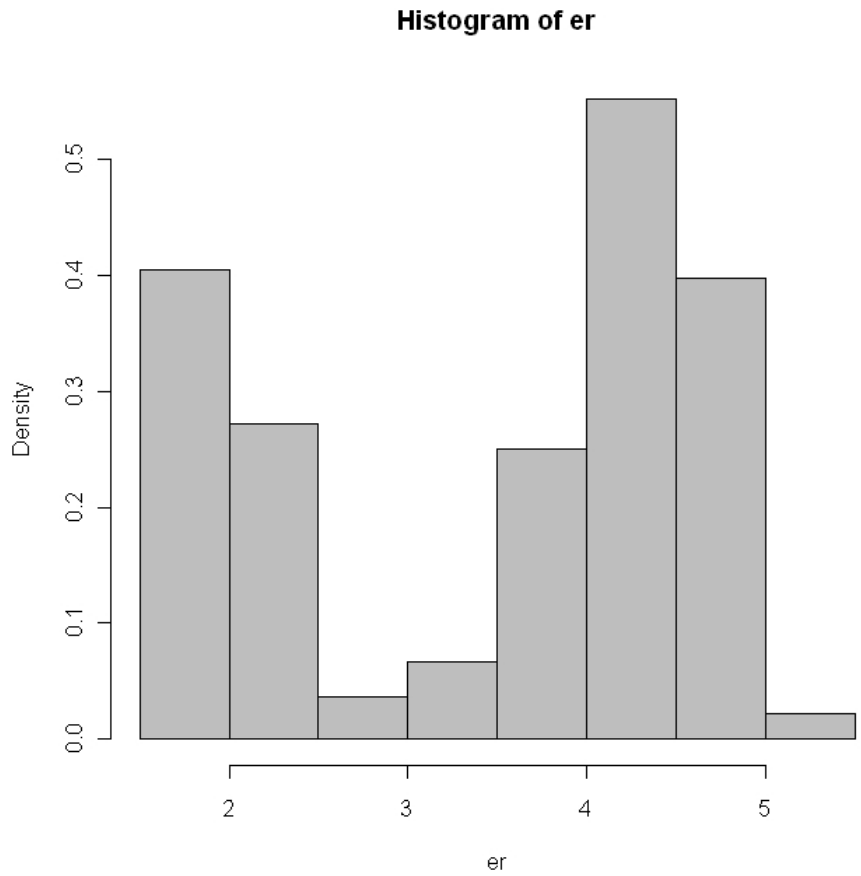
$$\hat{f}_{hist}(x) = \frac{n_k}{n} \frac{1}{a_k - a_{k-1}} \mathbb{1}_{\{I_k\}}(x)$$

Im Falle gleicher Intervalllängen mit
$a_k - a_{k-1} \equiv h\ \forall k \in \{1, \ldots, K\}$:

$$\hat{f}_{hist}(x) = \frac{n_k}{nh} \mathbb{1}_{\{I_k\}}(x)$$




```
> hist(er, freq=FALSE,col="grey")
```



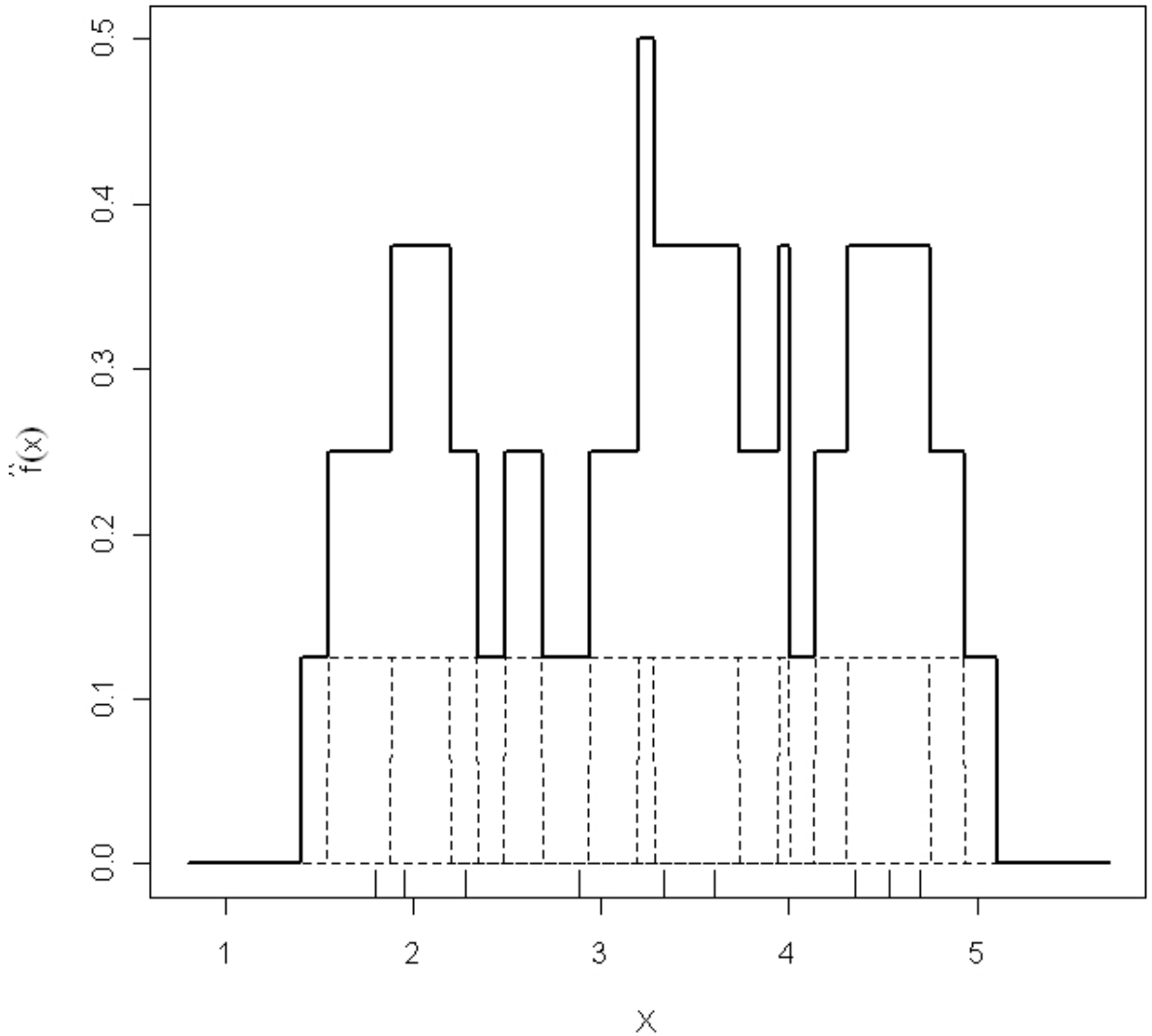
Nachteil des Histogramm–Schätzers:
Schätzer hängt von der Wahl der Klassen–Längen und des Startwertes a_0 ab!

Gleitendes Histogramm

Durch den gleitenden Histogramm–Schätzer

$$\hat{f}_{GH}(x) := \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h} = \frac{\#\{x_i | x_i \in (x-h, x+h]\}}{2hn}$$
$$= \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}_R\left(\frac{x - x_i}{h}\right) \quad \text{mit} \quad \mathcal{K}_R(t) = (1/2)\mathbb{1}_{[-1,1]}(t),$$

bei dem jede Beobachtung Mittelpunkt eines bins ist, lässt sich das Startwertproblem lösen.


f(x)
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
X

Kernfunktionen

Definition

Eine Funktion $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ heißt Kern, falls gilt:

1. $\int \mathcal{K}(x)dx = 1, \mathcal{K}(x) \geq 0 \quad \forall x \in \mathbb{R}, \mathcal{K}(x) = \mathcal{K}(-x)$

   Regularitätsbedingungen:

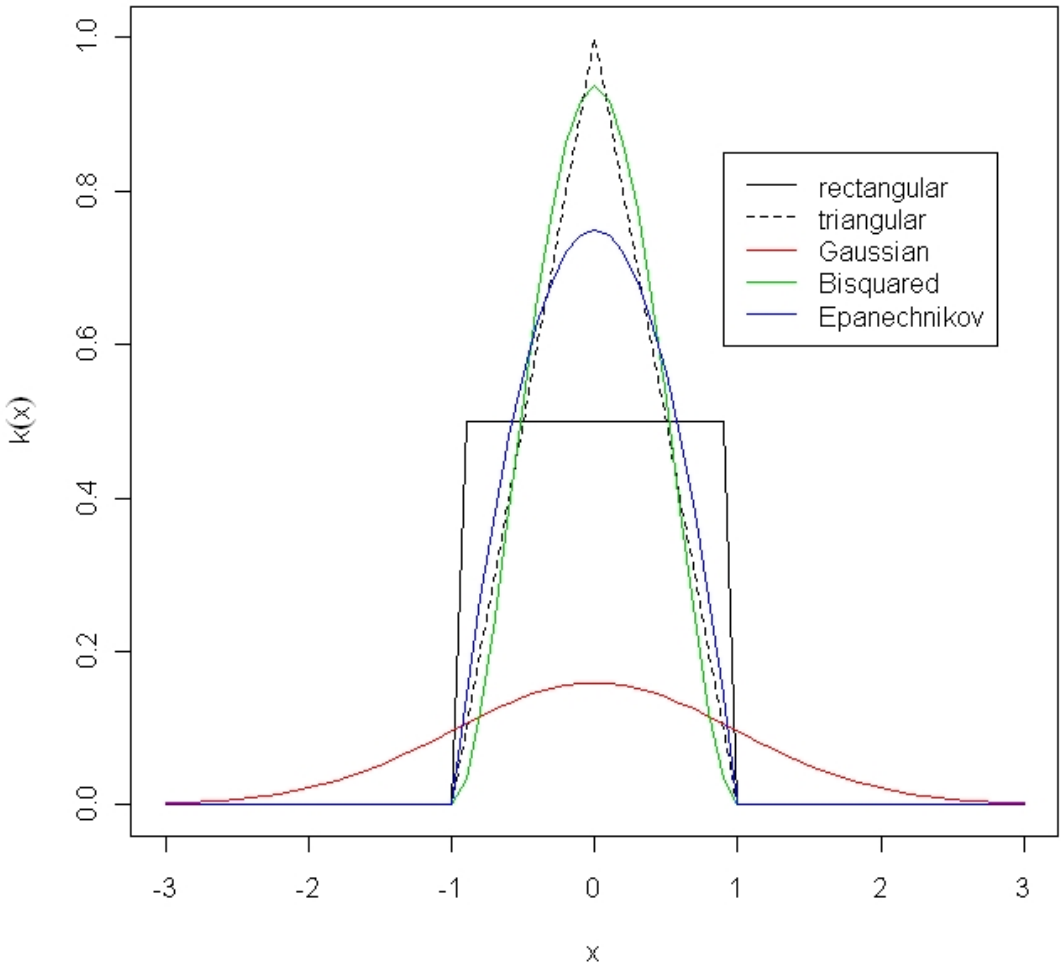
2. $\sup_{x \in \mathbb{R}} \mathcal{K}(x) = M < \infty$
3. $|x|\mathcal{K}(x) \to 0$ für $|x| \to 0$, $\int x^2 \mathcal{K}(x)dx =: k_2 < \infty$

Kernfunktionen: Beispiele

Beispiele für Kernfunktionen:

Rechteckskern	$\mathcal{K}(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x),$
Dreieckskern	$\mathcal{K}(x) = (1 - \lvert x\rvert)\mathbb{1}_{[-1,1]}(x),$
Gaußkern	$\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2),$
Bisquarekern	$\mathcal{K}(x) = \frac{15}{16}(1 - x^2)^2\mathbb{1}_{[-1,1]}(x),$
Epanechnikovkern	$\mathcal{K}(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x).$

Grafische Darstellung verschiedener Kernfunktionen



Univariater Kerndichteschätzer

Definition

Sei $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ ein Kern.

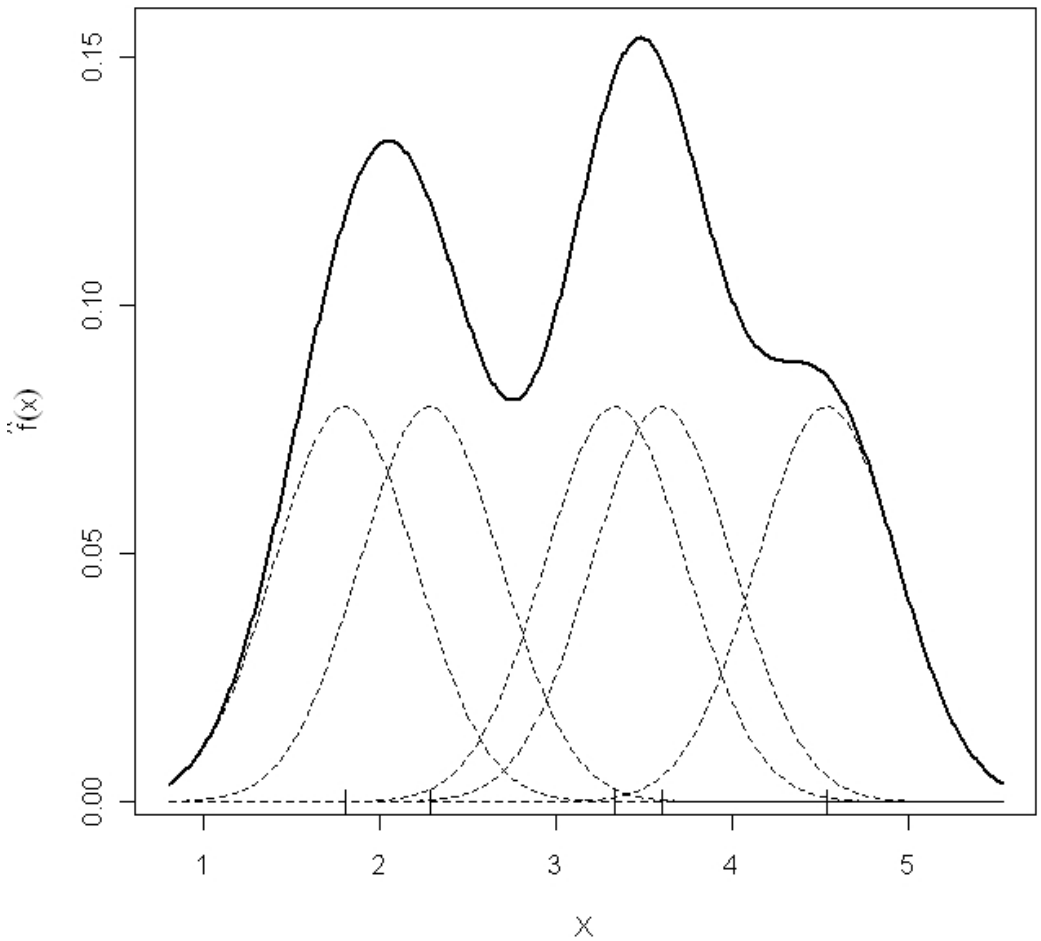
$$\hat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{t - x_i}{h}\right) = \int \frac{1}{h} \mathcal{K}\left(\frac{t - x}{h}\right) \hat{F}_n(dx)$$

heißt (univariater) Kerndichteschätzer mit Bandweite h und Kern $\mathcal{K}$.

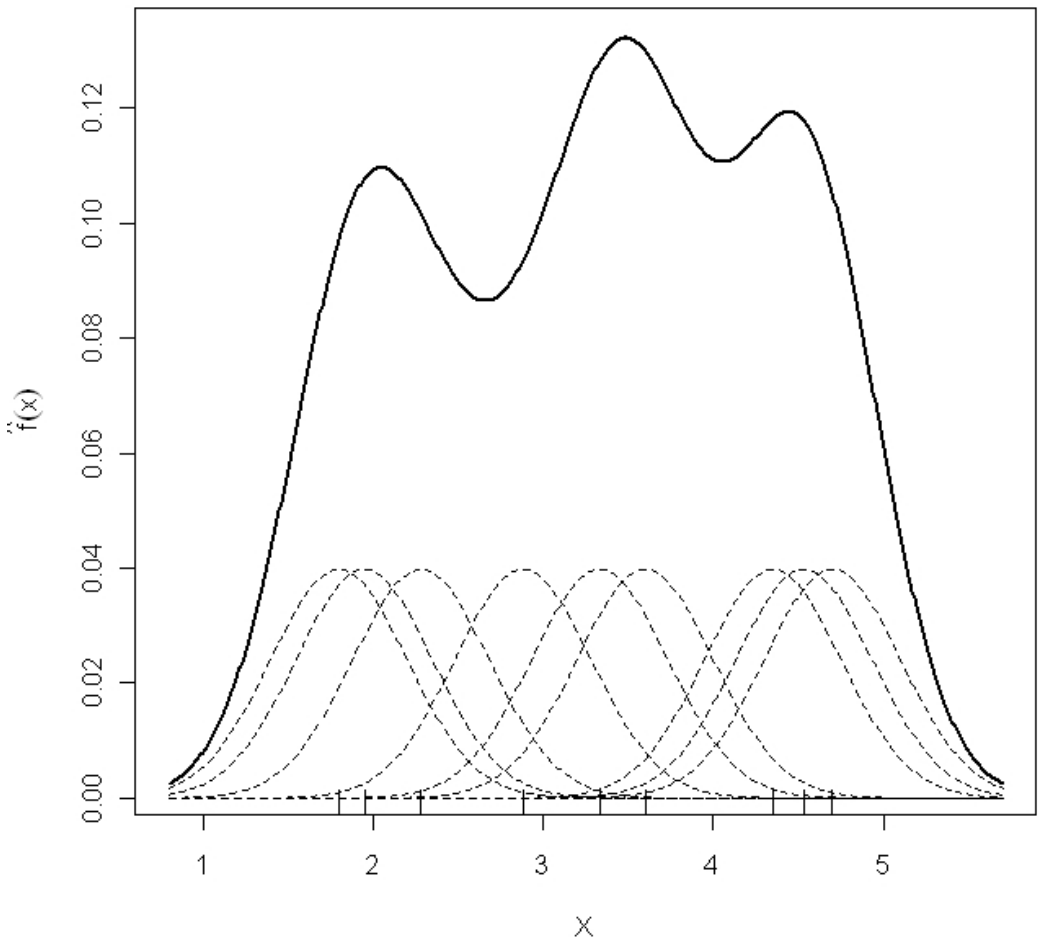
Mit $\tilde{\mathcal{K}}(t) := \int_{-\infty}^{t} \mathcal{K}(x)dx$ lässt sich auch $F(t)$ schätzen durch

$$\int \tilde{\mathcal{K}}\left(\frac{t - x}{h}\right) \hat{F}_n(dx) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathcal{K}}\left(\frac{t - x_i}{h}\right).$$

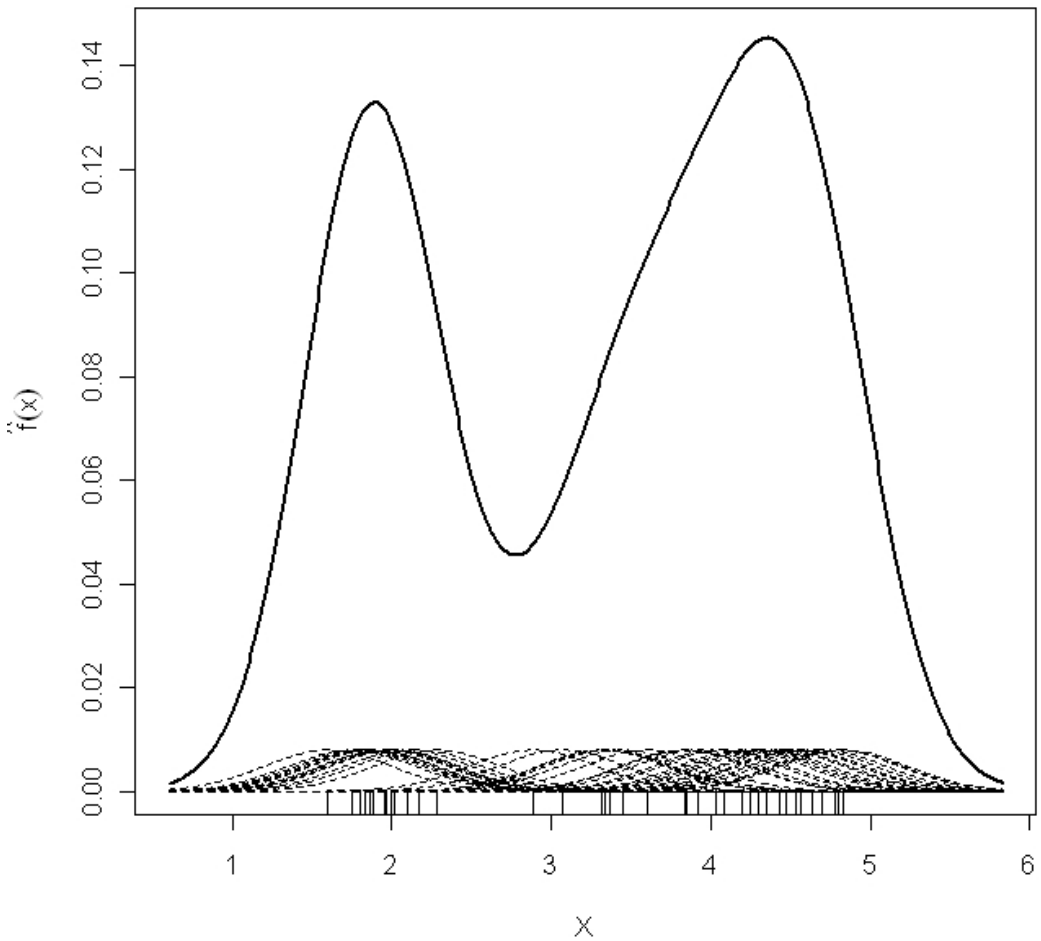
Gauß–Kernschätzer ($n = 5$)

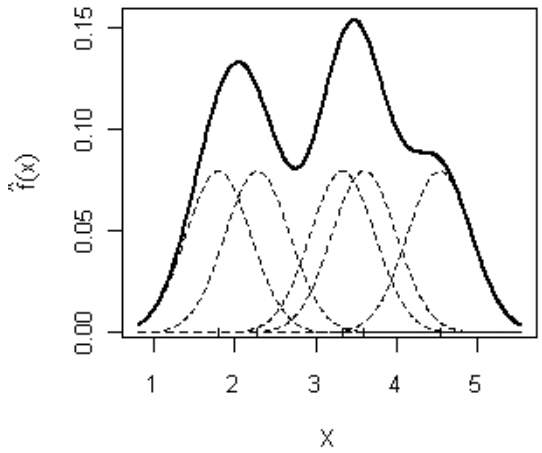


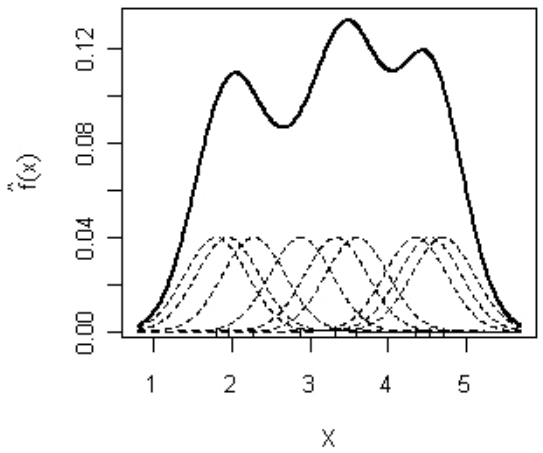
Gauß–Kernschätzer ($n = 9$)

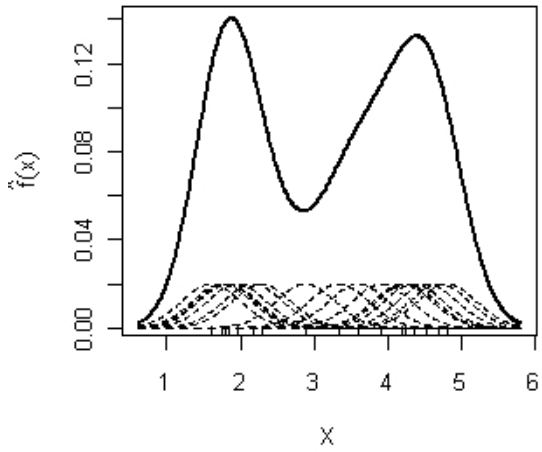


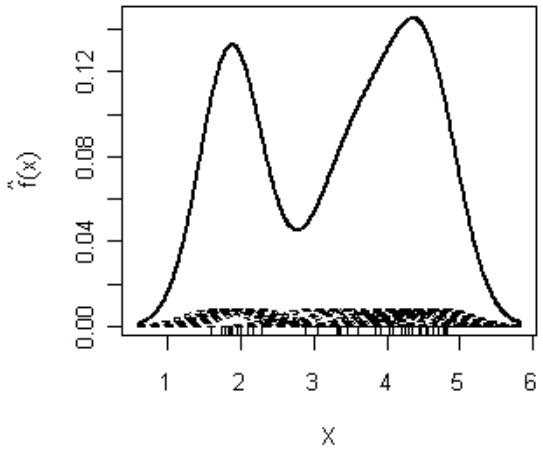
Gauß–Kernschätzer ($n = 50$)




f̂(x)
0.00
0.05
0.10
0.15
1
2
3
4
5
X


f̂(x)
0.00
0.04
0.08
0.12
1
2
3
4
5
X


f̂(x)
0.00
0.04
0.08
0.12
1
2
3
4
5
6
X


f̂(x)
0.00
0.04
0.08
0.12
1
2
3
4
5
6
X

Kernschätzung: Feinkalibrierung

Entscheidende Schwierigkeit: Wahl der Bandweite!

h zu groß $\longrightarrow$ *oversmoothing*
$\longrightarrow$ lokale Extrema werden nicht erkannt, zu glatt

h zu klein $\longrightarrow$ *undersmoothing*
$\longrightarrow$ lokale Moden, Schätzer ist „hairy“

Bias

Satz

Wenn $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ ein Kern ist, der die oben genannten Regularitätsbedingungen erfüllt, und $f \in \mathcal{C}^2(\mathbb{R})$, so gilt:

$$\mathbb{E}_f[\hat{f}_n(x)] - f(x) = \frac{h^2}{2} f''(x) \int x^2 \mathcal{K}(x) dx + o(h^2) \, .$$

Bias

Beweis über Taylor–Entwicklung von f:

$$
\begin{aligned}
f(x-ht) &= f(x) - htf'(x) + \frac{h^2t^2}{2}f''(x) + \mathfrak{o}(h^2t^2) \\
\mathbb{E}_f[\hat{f}_n(x)] - f(x) &= \int \frac{1}{h}\mathcal{K}\left(\frac{x-y}{h}\right) f(y)dy - f(x) \\
&= -hf'(x)\underbrace{\int y\mathcal{K}(y)dy}_{=0} + \\
&\quad \frac{h^2f''(x)}{2}\underbrace{\int y^2\mathcal{K}(y)dy}_{=k_2} + \mathfrak{o}(h^2t^2) \\
&= \frac{h^2f''(x)}{2}k_2 + \mathfrak{o}(h^2t^2).
\end{aligned}
$$

Varianz

Satz

Wenn $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ ein Kern ist, der die oben genannten Regularitätsbedingungen erfüllt, und $f \in \mathcal{C}(\mathbb{R})$, so gilt:

$$\mathbb{V}\mathrm{ar}_f\left(\hat{f}_n(x)\right) = \frac{1}{nh} f(x) \int \mathcal{K}^2(y)dy + \mathfrak{o}\left(n^{-1}h^{-1}\right) \ .$$

Varianz

Beweis:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(\hat{f}_n(x)\right) &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\mathbb{V}\mathrm{ar}\left(\mathcal{K}\left(\frac{x-x_i}{h}\right)\right) \\
&= \frac{1}{nh^2}\int \mathcal{K}^2\left(\frac{x-y}{h}\right) f(y)dy - \frac{1}{n}\left(\mathbb{E}[\hat{f}_n(x)]\right)^2 \\
&= \frac{1}{nh}\int \mathcal{K}^2(y) f(x-yh)dy - n^{-1}\left(\mathbb{E}[\hat{f}_n(x)]\right)^2 \\
&= \frac{1}{nh} f(x)\int \mathcal{K}^2(y)dy + \mathfrak{o}\left(n^{-1}h^{-1}\right).
\end{aligned}$$

Mean Squared Error: Bias–Varianz–Zerlegung

$$\mathbb{E}_f\left[\left(\hat{f}_n(x) - f(x)\right)^2\right] = \text{Bias}(\hat{f}_n(x|h))^2 + \mathbb{V}\text{ar}(\hat{f}_n(x|h))$$

$$= h^4\left(\frac{f''(x)}{2}k_2\right)^2 + \frac{1}{nh}f(x)\int \mathcal{K}^2(y)dy + \mathfrak{o}\left(h^4 + n^{-1}h^{-1}\right)$$

$\Rightarrow$ Trade–Off zwischen Bias und Varianz möglich!

Anmerkung:

Bias hängt nicht explizit vom Stichprobenumfang n ab.

Für Konsistenz muss indes $h \equiv h(n) \to 0$ und $nh \to \infty$ für $n \to \infty$ gelten!

Optimaler Kern

Minimierung des MISE bezüglich h ergibt optimale Bandweite:

$$h_{opt} = \frac{\left(\int \mathcal{K}^2(y)dy\right)^{1/5}}{n^{1/5}k_2^{1/5}\left(\int (f''(y))^2\,dy\right)^{1/5}} \,. \tag{1}$$

Setzt man h_{opt} in den MISE ein, erhält man

$$\text{MISE} \approx \frac{5}{4}\left(k_2^{2/5}\left(\int \mathcal{K}^2(y)dy\right)^{4/5}\right)\left(\int (f''(y))^2\,dy\right)^{1/5} \,.$$

Minimal für Epanechnikov–Kern:

$$\mathcal{K}_e(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbb{1}_{[-5,5]}(x) \,.$$

Optimaler Kern

Setzt man h_{opt} in den MISE ein, erhält man

$$\text{MISE} \approx \frac{5}{4}\left(k_2^{2/5}\left(\int \mathcal{K}^2(y)dy\right)^{4/5}\right)\left(\int \left(f''(y)\right)^2 dy\right)^{1/5} .$$

Minimal für Epanechnikov–Kern:

$$\mathcal{K}_e(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbb{1}_{[-5,5]}(x) .$$

Effizienz $eff(\mathcal{K})$ für $\mathcal{K} \neq \mathcal{K}_e$ und n gegeben:
Zahl $eff(\mathcal{K})$ löst Gleichung $\text{MISE}(n, \mathcal{K}) = \text{MISE}(n \cdot eff(\mathcal{K}), \mathcal{K}_e)$.

Gauß–Kern: Effizienz von ca. 0.95,
Rechteck–Kern: Effizienz von ca. 0.93.

Übersicht

Multivariate Daten: Mietspiegel–Daten

```
> miete <- read.table(file="miete03.dat", header=TRUE)
> str(miete)
'data.frame':   2053 obs. of  13 variables:
 $ GKM       : num  741 716 528 554 698 ...
 $ QM        : int  68 65 63 65 100 81 55 79 52 77 ...
 $ QMKM      : num  10.9 11.01 8.38 8.52 6.98 ...
 $ Rooms     : int  2 2 3 3 4 4 2 3 1 3 ...
 $ BJ        : num  1918 1995 1918 1983 1995 ...
 $ lage_gut  : int  1 1 1 0 1 0 0 0 0 0 ...

 $ bez       : int  2 2 2 16 16 16 6 6 6 6 ...
 $ wohnbest  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ww0       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ zh0       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ badkach0  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ badextra  : int  0 0 0 1 1 0 1 0 0 0 ...
 $ kueche    : int  0 0 0 0 1 0 0 0 0 0 ...
```

Abgeleitete Variablen

Hier: Klassierung von Baujahr und Quadratmeterzahl

```
> miete$BJKL<-1*(BJ<=1918)+2*(BJ<=1948)*(BJ>1919)+3*(BJ<=1965)
                *(BJ>1948)+4*(BJ<=1977)*(BJ>1965)+5*(BJ<=1983)
                *(BJ>1977)+6*(BJ>1983)

> miete$QMKL<-1*(QM<=50)+2*(QM>50)*(QM<=80)+3*(QM>80)
```

Zwei diskrete Merkmale: Kontingenztafeln

Mögliche Werte für Merkmal 1: $a_1, a_2, \ldots, a_k$
Mögliche Werte für Merkmal 2: $b_1, b_2, \ldots, b_\ell$

Beobachtung x: Matrix der absoluten Häufigkeiten aller Kombinationen (a_i, b_j), $1 \leq i \leq k$, $1 \leq j \leq \ell$ in der Stichprobe vom Umfang n

Darstellung als Kontingenztafel (auch: $(k \times \ell)$-Feldertafel):

	b_1	b_2	$\ldots$	b_ℓ	$\sum$
a_1	x_{11}	x_{12}	$\ldots$	$x_{1\ell}$	$n_{1.}$
a_2	x_{21}	x_{22}	$\ldots$	$x_{2\ell}$	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
a_k	x_{k1}	x_{k2}	$\ldots$	$x_{k\ell}$	$n_{k.}$
$\sum$	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.\ell}$	n

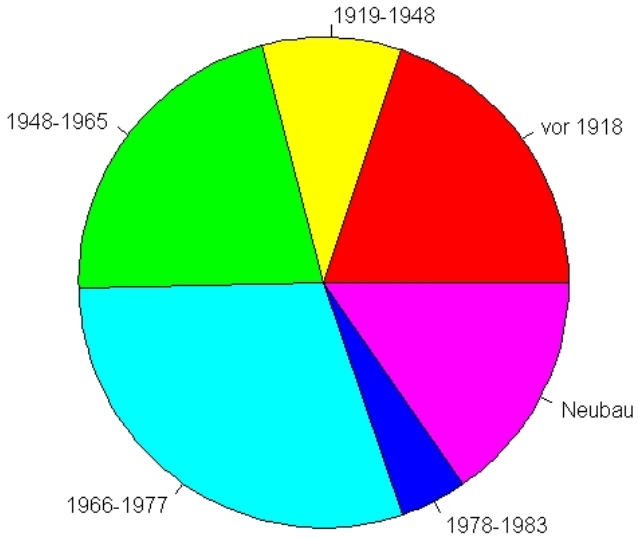
Randhäufigkeiten, marginale Verteilungen

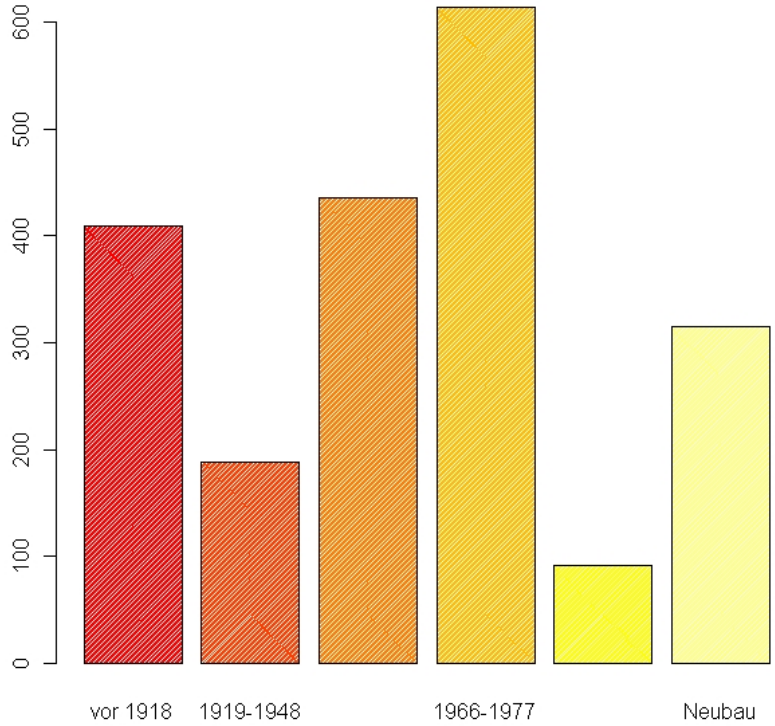
Der Vektor $\mathbf{n} = (n_{1.}, n_{2.}, \ldots, n_{k.}, n_{.1}, n_{.2}, \ldots, n_{.\ell}) \in \mathbb{N}^{k+\ell}$ heißt Vektor der (empirischen) Randhäufigkeiten.

Die (emprirische) diskrete Verteilung, die durch die Randhäufigkeiten eines Merkmals gegeben ist, bezeichnet man als Randverteilung oder auch marginale Verteilung dieses Merkmals.

```
> h<-numeric(6)
> for(i in 1:6){
+ h[i]<-length(which(BJKL==i))}
> names(h)<-c("vor 1918","1919-1948","1948-1965","1966-1977",
+ "1978-1983","Neubau")

> pie(h,col=rainbow(6))
> barplot(h,col=heat.colors(6),density=100)
```

Grafische Darstellung von Randverteilungen



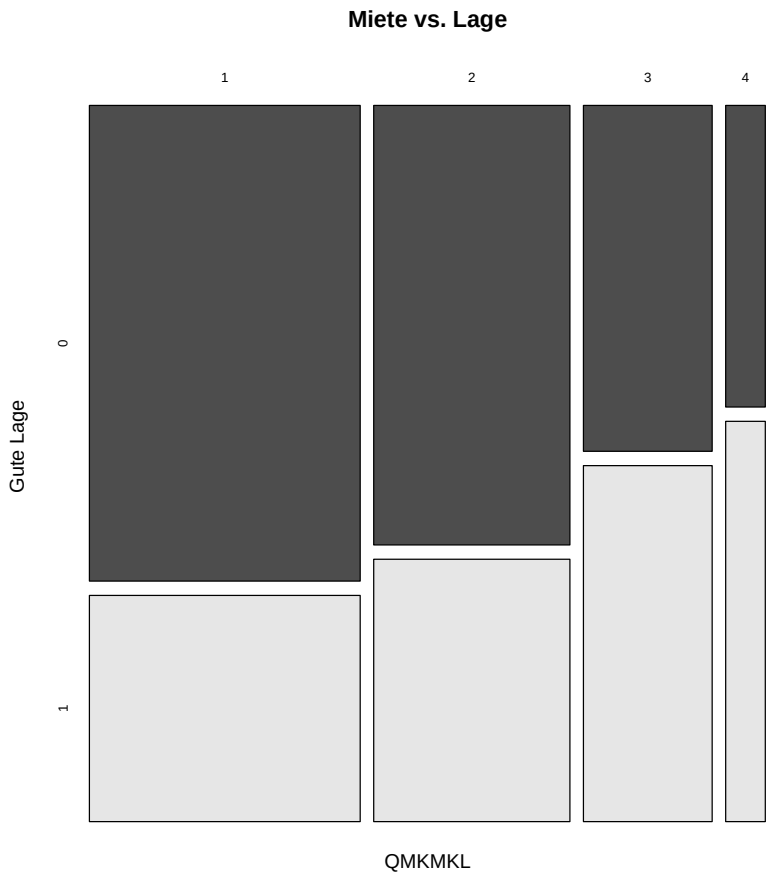


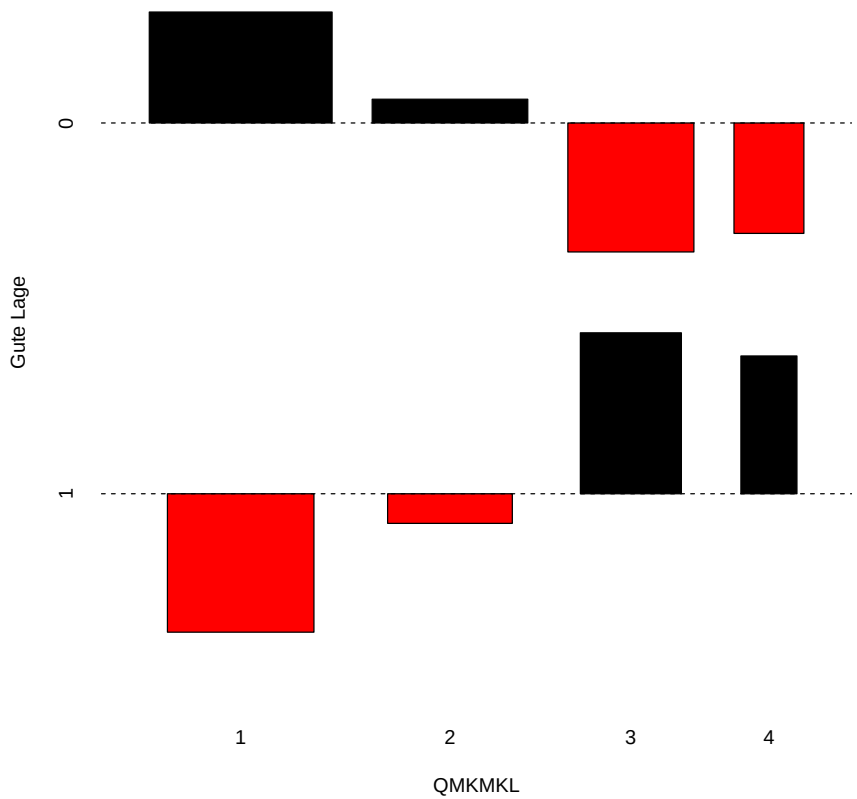
Grafische Darstellung bivariater diskreter Verteilungen

R Code: **mosaicplot** und assocplot

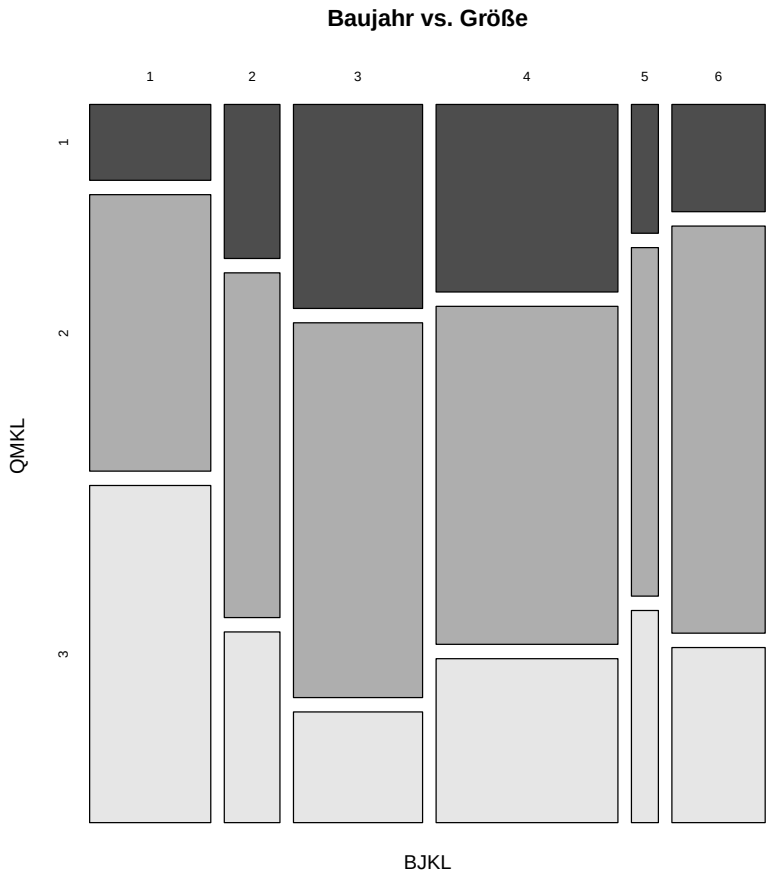
```
> miete$QMKMKL<-1*(QMKM<=8)+2*(QMKM>8)*(QMKM<=10)
  +3*(QMKM>10)*(QMKM<=12)+4*(QMKM>12);

> par(mfrow=c(1, 2));
> mosaicplot(table(miete$QMKMKL, miete$lage_gut),col=TRUE,
+              xlab="QMKMKL", ylab="Gute Lage");
> assocplot(table(miete$QMKMKL, miete$lage_gut),
+             xlab="QMKMKL", ylab="Gute Lage");

> par(mfrow=c(1, 2));
> mosaicplot(table(miete$BJKL, miete$QMKL),col=TRUE,
+              xlab="BJKL", ylab="QMKL");
> assocplot(table(miete$BJKL, miete$QMKL),
+              xlab="BJKL", ylab="QMKL");
```

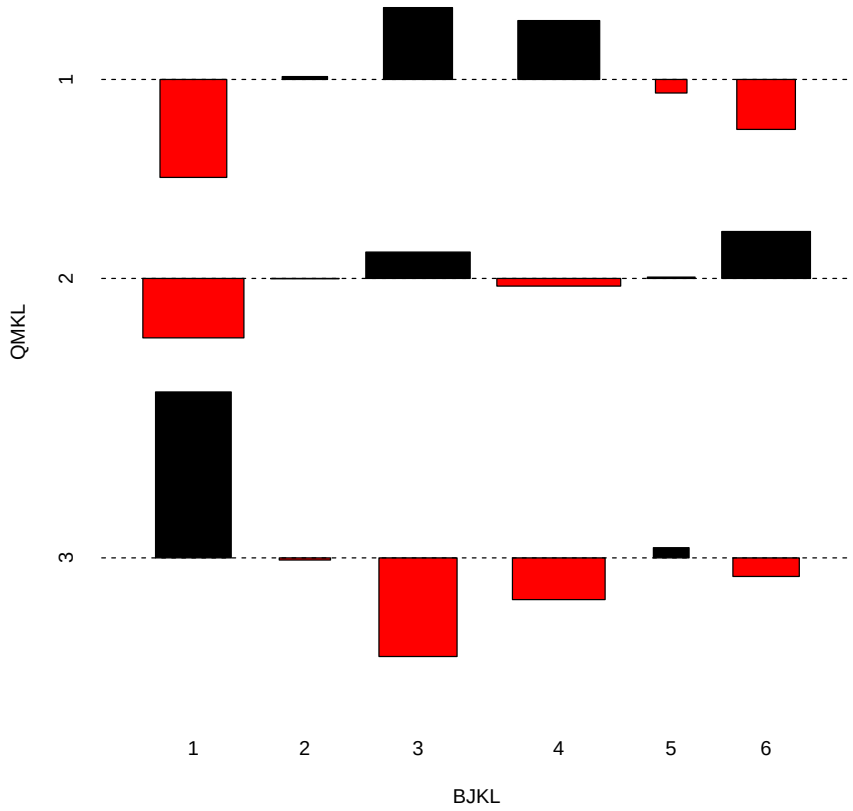
Miete versus Wohnlage





Baujahr versus Wohnungsgröße





Multivariate stetige Verteilungen

Modell: $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ i. i. d. $\sim f$, f Verteilungsdichte auf $\mathbb{R}^p$.

Definition (p-dimensionaler Kern)

Ein Kern ist eine Funktion $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ mit

$$\int_{\mathbb{R}^p} \mathcal{K}(\mathbf{y}) d\mathbf{y} = 1 \quad \text{und}$$

Regularitätsbedingungen:

- $\mathcal{K}$ ist radialsymmetrische Wahrscheinlichkeitsdichte
- Existierendes zweites Moment $\mu_2(\mathcal{K})$

p-dim. Kernfunktionen, Kerndichteschätzer

Beispiele:

uniformer Kern $\quad \mathcal{K}(\mathbf{x}) = \frac{1}{v_p}$ für $\mathbf{x}^T\mathbf{x} \leq 1$,

Gaußkern $\quad \mathcal{K}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right)$,

Epanechnikovkern $\quad \mathcal{K}(\mathbf{x}) = \frac{1+p/2}{v_p}(1 - \mathbf{x}^T\mathbf{x})$, $\mathbf{x}^T\mathbf{x} \leq 1$.

Definition

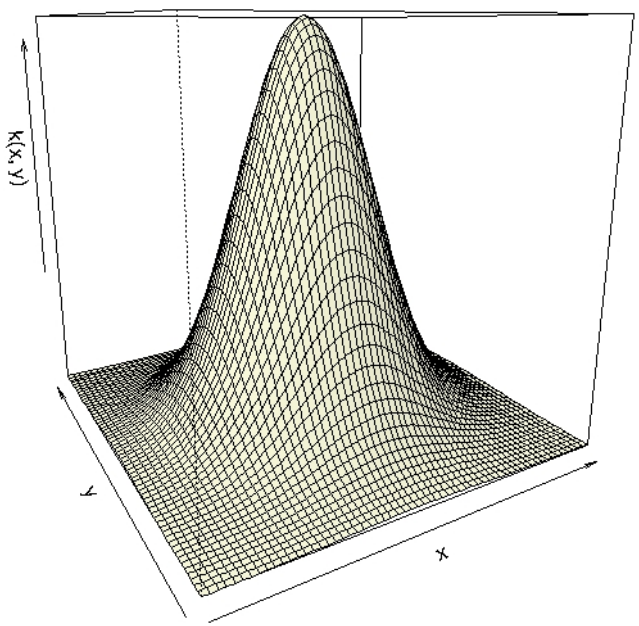
Sei $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ ein Kern.

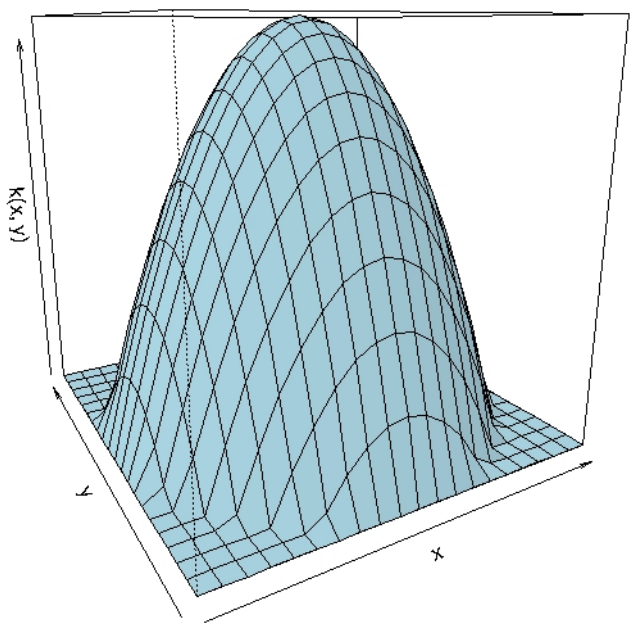
$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right), \ \mathbf{x} \in \mathbb{R}^p$$

heißt multivariater Kerndichteschätzer mit Bandweite h und Kern $\mathcal{K}$.

Darstellung zweidimensionaler Kernfunktionen

Gaußkern und Epanechnikovkern mit $p = 2$





Bandweitenwahl

Bias und Varianz:

$$Bias_h\left(\hat{f}_n(\mathbf{x})\right) = \frac{h^2}{2}\mathbb{H}_f(\mathbf{x})\mu_2(\mathcal{K}) + \mathcal{o}(h^2)\ ,$$

$$\mathbb{V}\mathrm{ar}_h\left(\hat{f}_n(\mathbf{x})\right) = \frac{1}{nh^p}||\mathcal{K}||_2^2 f(\mathbf{x}) + \mathcal{o}\left(n^{-1}h^{-p}\right)\ .$$

Minimierung des MISE:

$$(h_{opt})^{p+4} = \frac{p}{n}\frac{||\mathcal{K}||_2^2}{\mu_2^2(\mathcal{K})\int_{\mathbb{R}^p}\mathbb{H}_f^2(\mathbf{y})d\mathbf{y}} \Rightarrow h_{opt} \sim n^{-1/(p+4)}.$$

Verschiedene Bandweiten in unterschiedliche Richtungen

Allgemeiner als in obiger Definition kann man den multivariaten Kerndichteschätzer mit einer Bandweitenmatrix H definieren:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n|H|} \sum_{i=1}^{n} \mathcal{K}\left(H^{-1}(\mathbf{x} - \mathbf{X}_i)\right), \mathbf{x} \in \mathbb{R}^p.$$

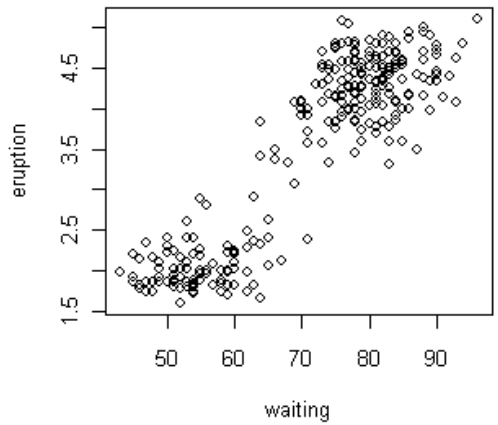
Zuvor: $H = h\mathbb{1}_p$, wobei $\mathbb{1}_p$ die p–dimensionale Einheitsmatrix bezeichnet

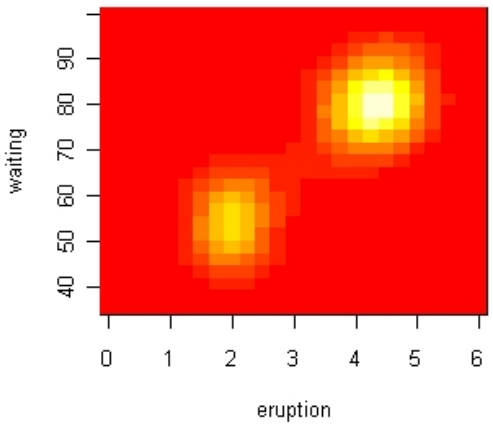
In R: Diagonalmatrix H angebbar.

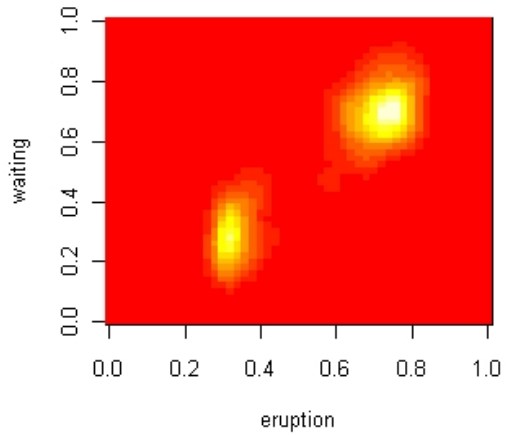
R Code: Zweidimensionale Kerndichteschätzung

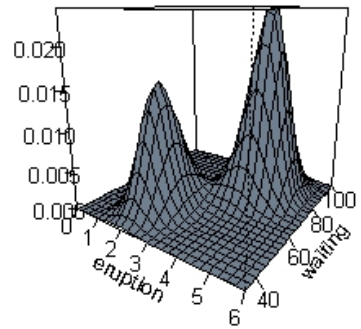
```
> library(MASS)
> library(KernSmooth)
> data(faithful)
> x<-faithful$eruptions
> y<-faithful$waiting

> par(mfrow=c(2,2),pty="m")
> plot(y,x,ylab="eruption",
    xlab="waiting") #Scatterplot, Streubild

> z<-kde2d(x,y,lims=c(0,6,35,100))
> zz<-bkde2D(faithful,range.x=list(c(0,6),c(35,100)),
+ bandwidth=c(bw.SJ(x),bw.SJ(y)))                    #b: binned
> image(z,xlab="eruption",ylab="waiting")
> image(zz$fhat,xlab="eruption",ylab="waiting") #Heat-Maps

> persp(z,col="slategrey",theta=35,xlim=c(0,6),ylim=c(35,100),
+ ticktype="detailed",xlab="eruption",ylab="waiting",zlab="")
```

Zweidimensionale Kerndichteschätzung



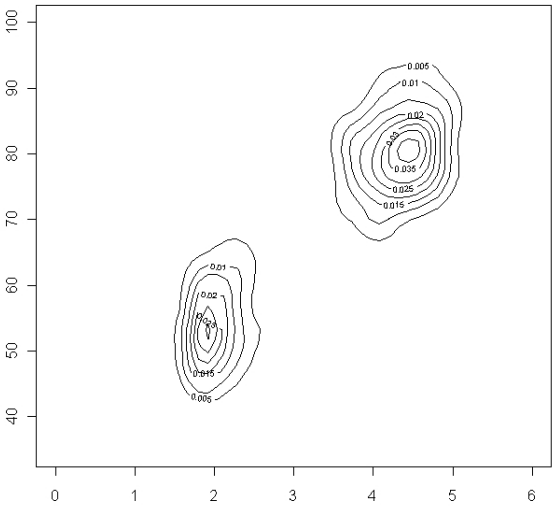


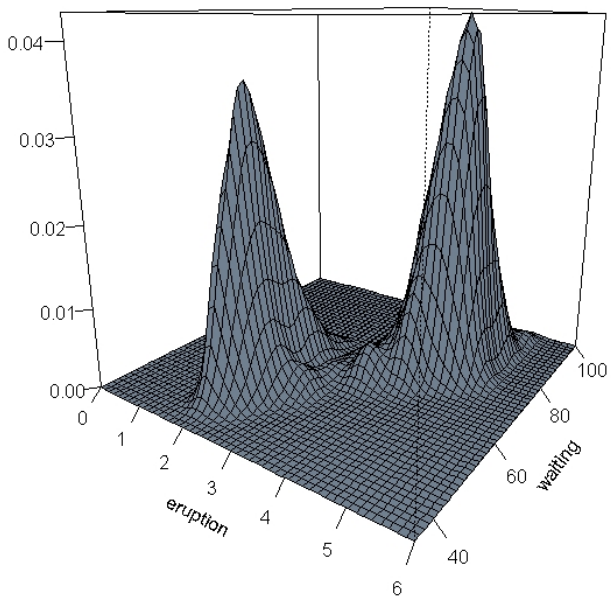




Kontur-Plots, 3D-Plots

```
> contour(zz$x1,zz$x2,zz$fhat)

> persp(zz$x1,zz$x2,zz$fhat,col="slategrey",
+ theta=35,xlim=c(0,6),ylim=c(35,100),
+ ticktype="detailed",xlab="eruption",
+ ylab="waiting",zlab="")
```

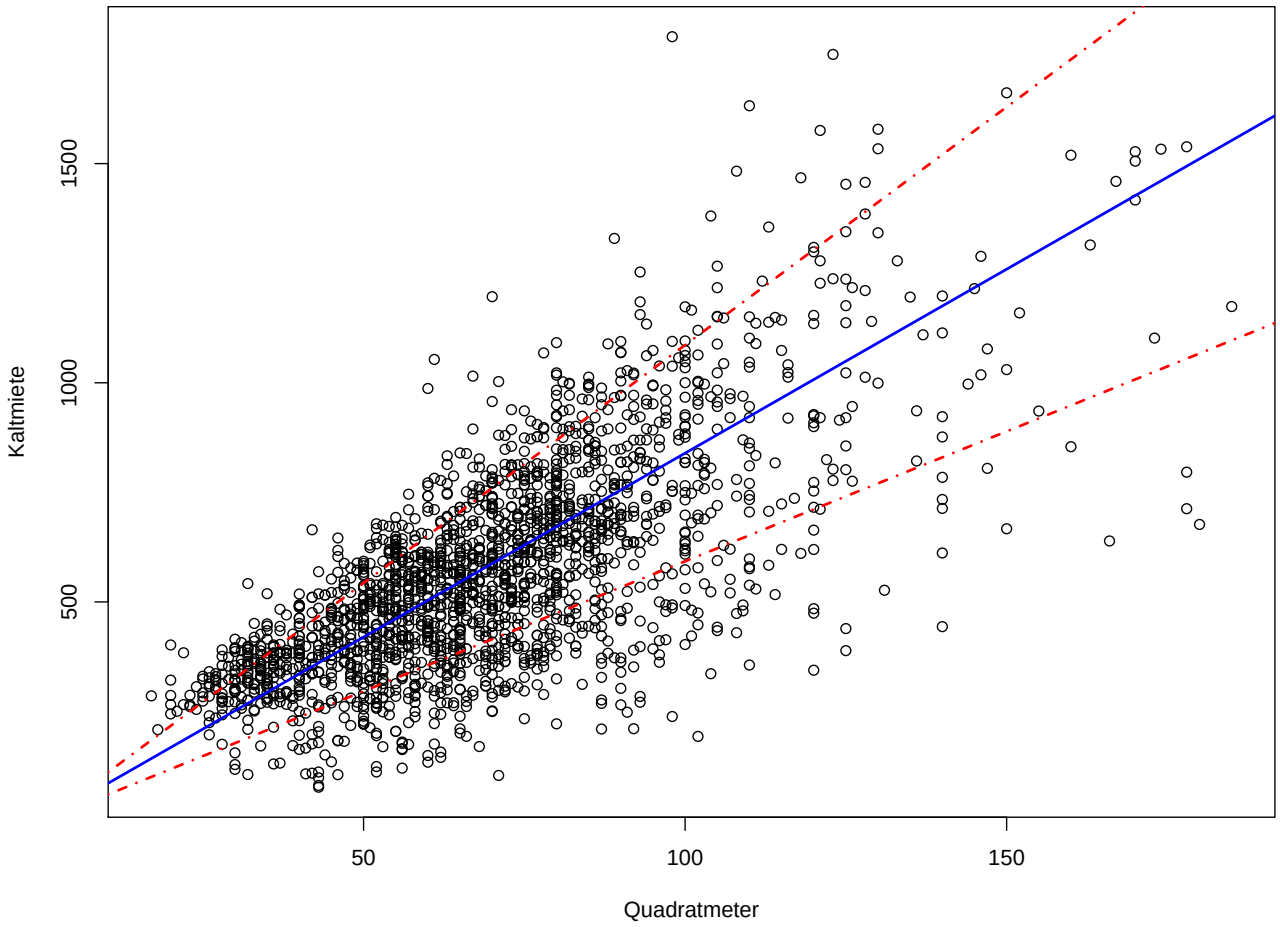




Zusammenhänge zwischen stetigen Variablen

Drei Ursprungsgeraden zur Beschreibung des Zusammenhangs zwischen den stetigen Merkmalen „Quadratmeterzahl“ und „Gesamtkaltmiete“:

```
plot(miete$QM, miete$GKM, xlab="Quadratmeter",
                          ylab="Kaltmiete");
abline(0,mean(QMKM),col="blue", lwd=2);
abline(0,mean(QMKM)+sd(QMKM),col="red",lty=4, lwd=2);
abline(0,mean(QMKM)-sd(QMKM),col="red",lty=4, lwd=2);
```


Kaltmiete
500
1000
1500
50
100
150
Quadratmeter