Grafische Darstellung bivariater Daten: Handwerkszeug der deskriptiven Statistik

Thorsten Dickhaus

Beuth Hochschule für Technik

31. Oktober 2012

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 ...
```

Übersicht

Übersicht

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

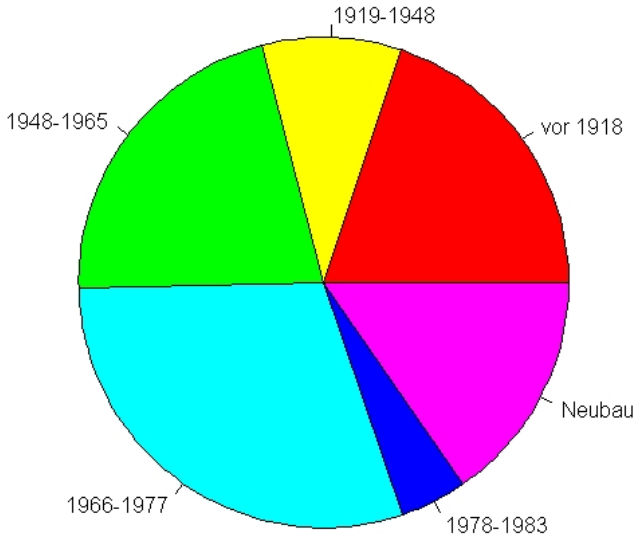
Randhäufigkeiten, marginale Verteilungen

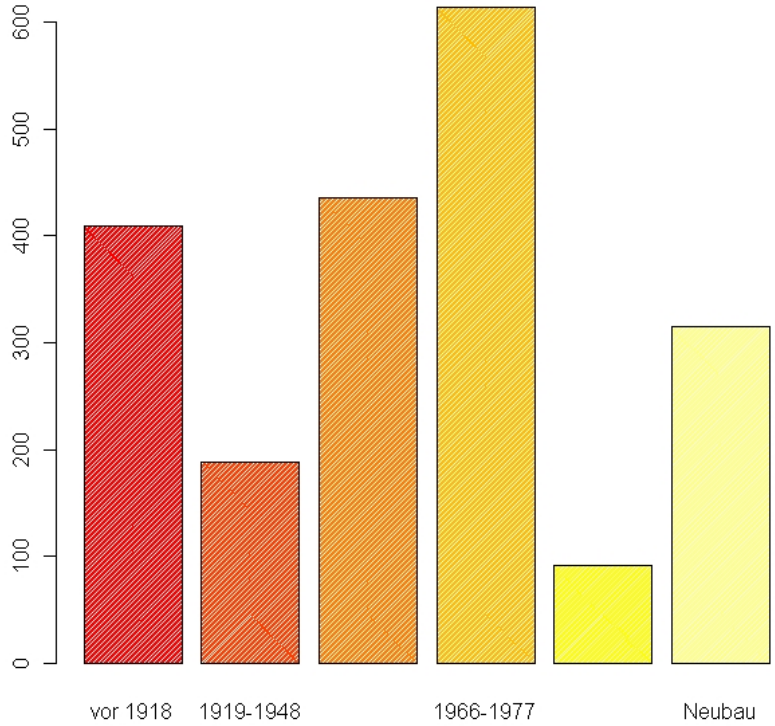
Der Vektor $\mathbf{n} = (n_{1.}, n_{2.}, \ldots, n_{k.}, n_{.1}, n_{.2}, \ldots, n_{.\ell}) \in \mathbb{N}^{k+\ell}$ heißt Vektor der (empirischen) Randhäufigkeiten.

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



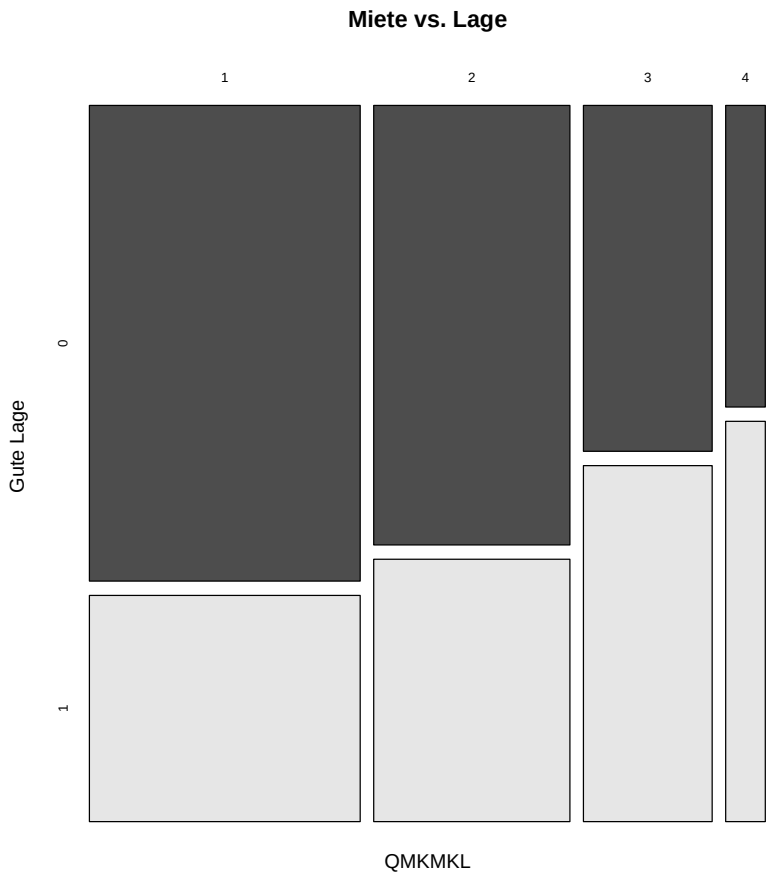


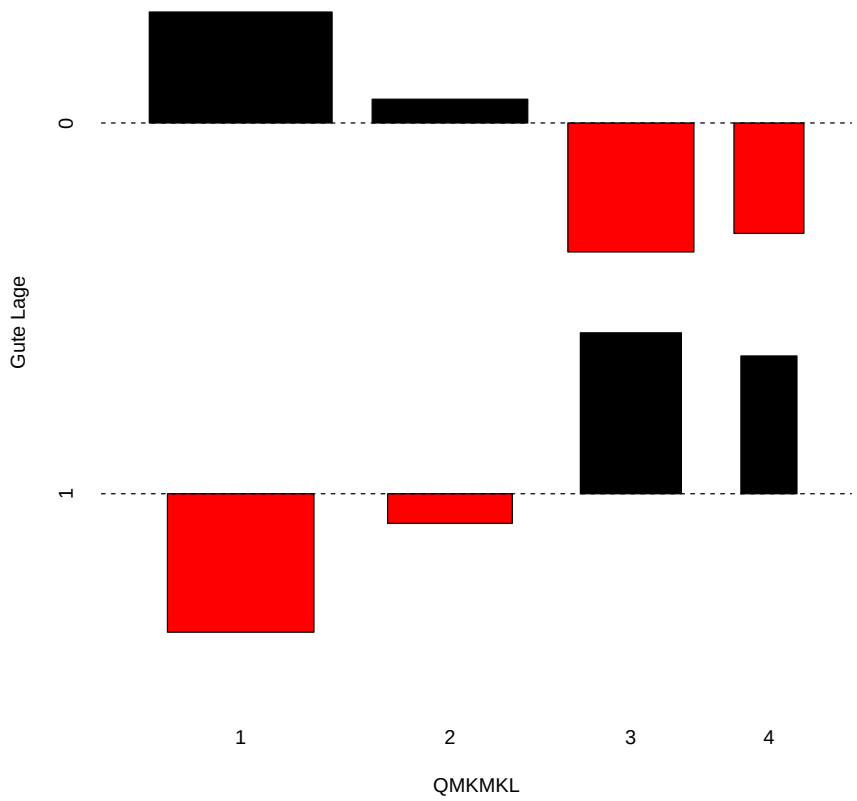
Grafische Darstellung bivariater diskreter Verteilungen

R Code: **mosaicplot** und assocplot

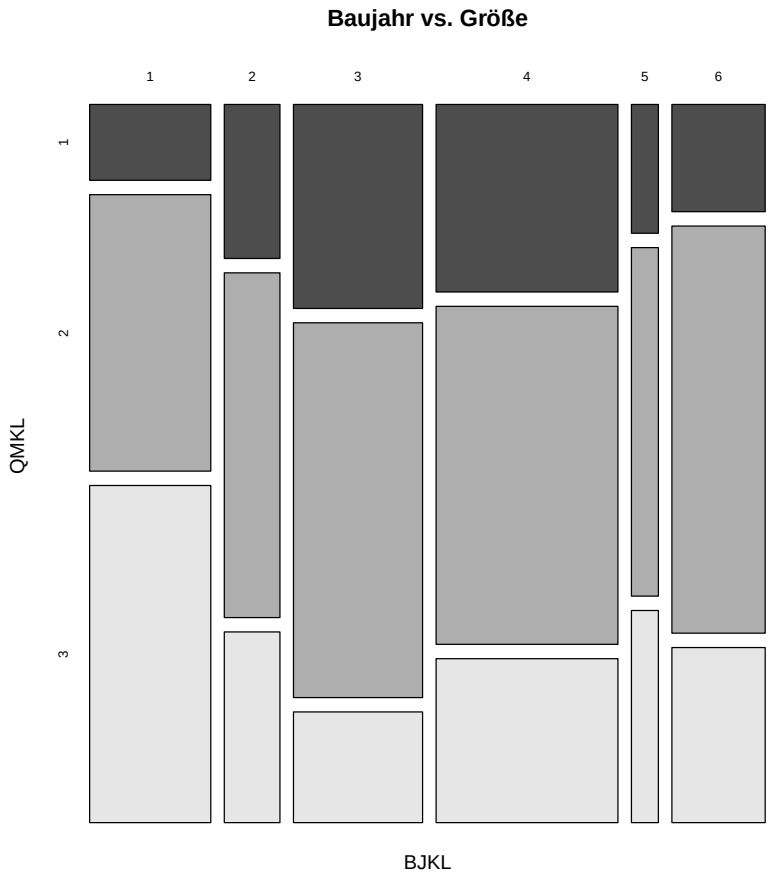
```
> miete$QMKMKL<-1*(QMKM<=8)+2*(QMKM>8)*(QMKM<=10)
  +3*(QMKM>10)*(QMKM<=12)+4*(QMKM>12);

> 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");
```

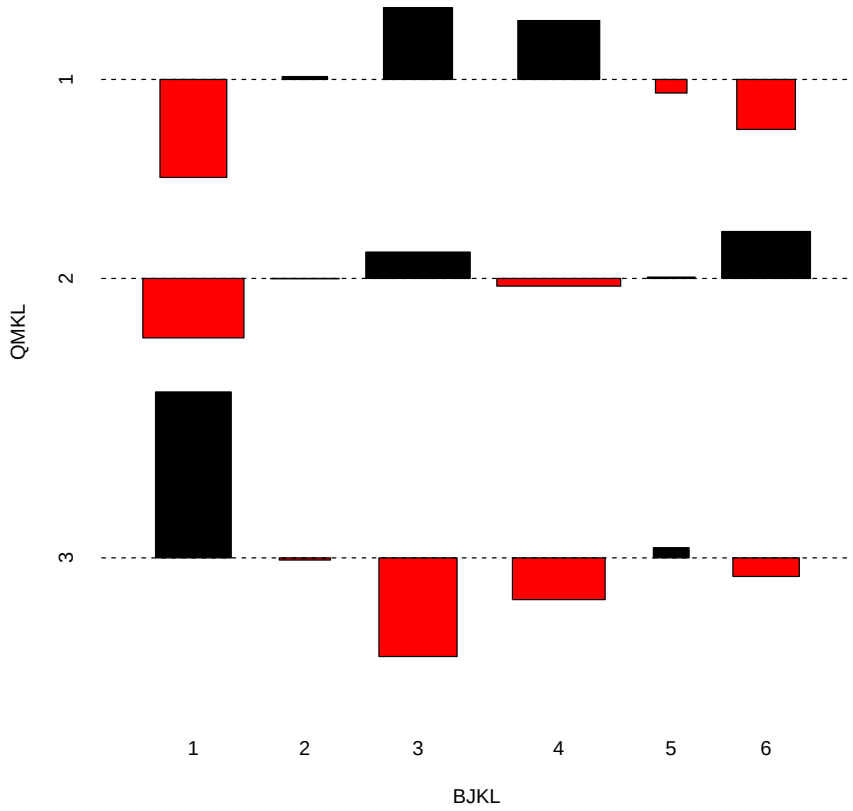
Miete versus Wohnlage





Baujahr versus Wohnungsgröße





Übersicht

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)

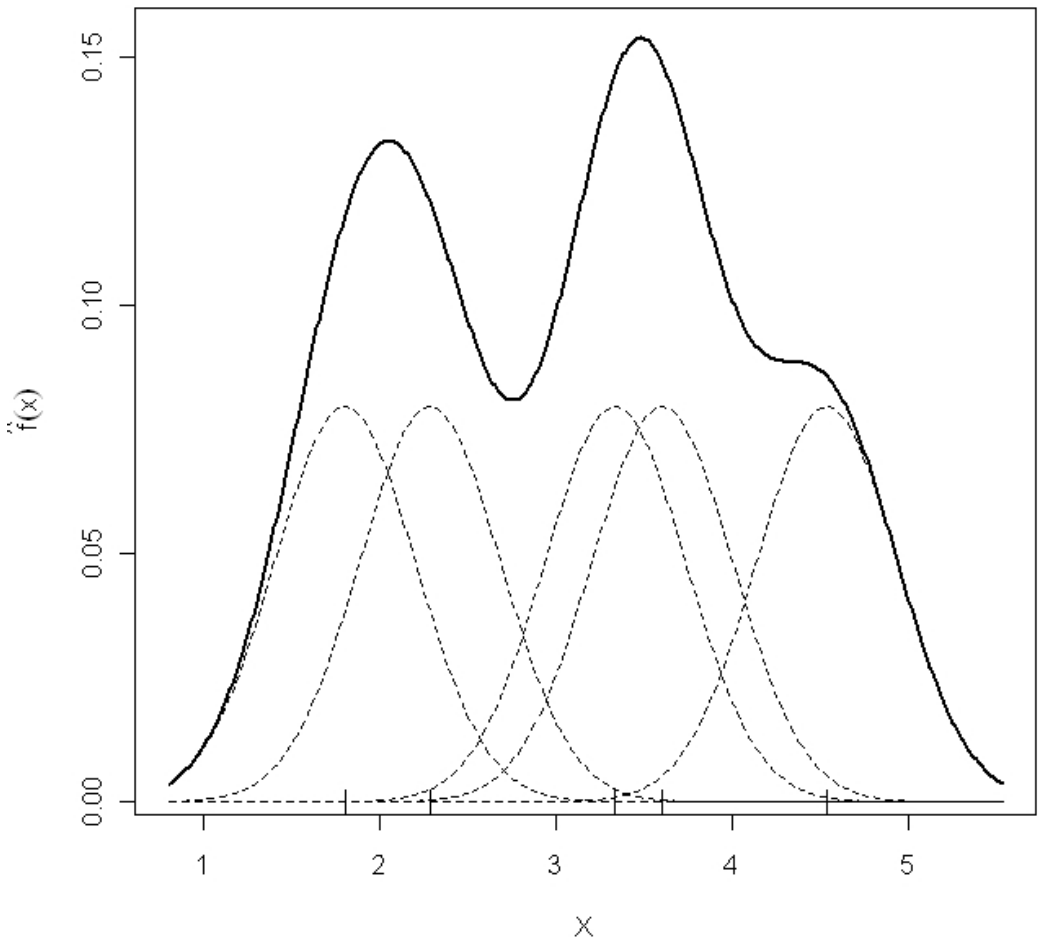
Ein Kern ist eine Funktion $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ mit

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

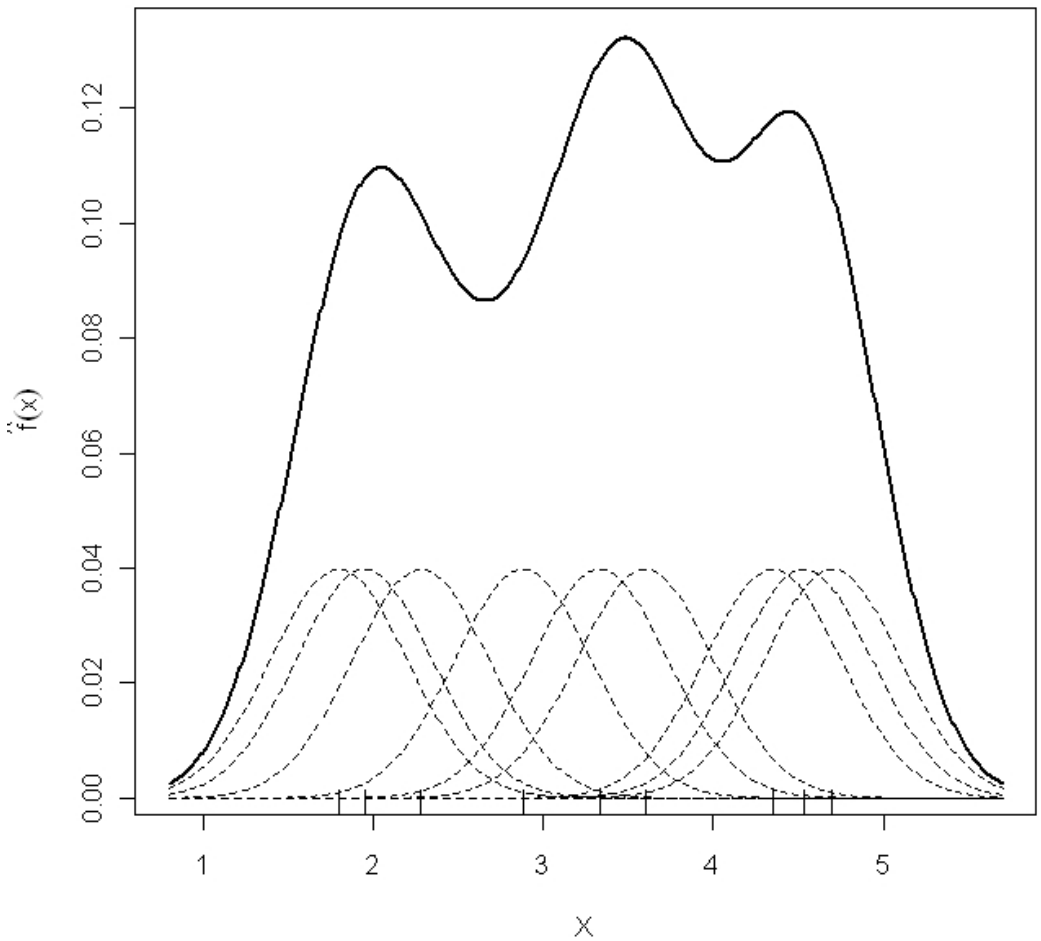
Regularitätsbedingungen:

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

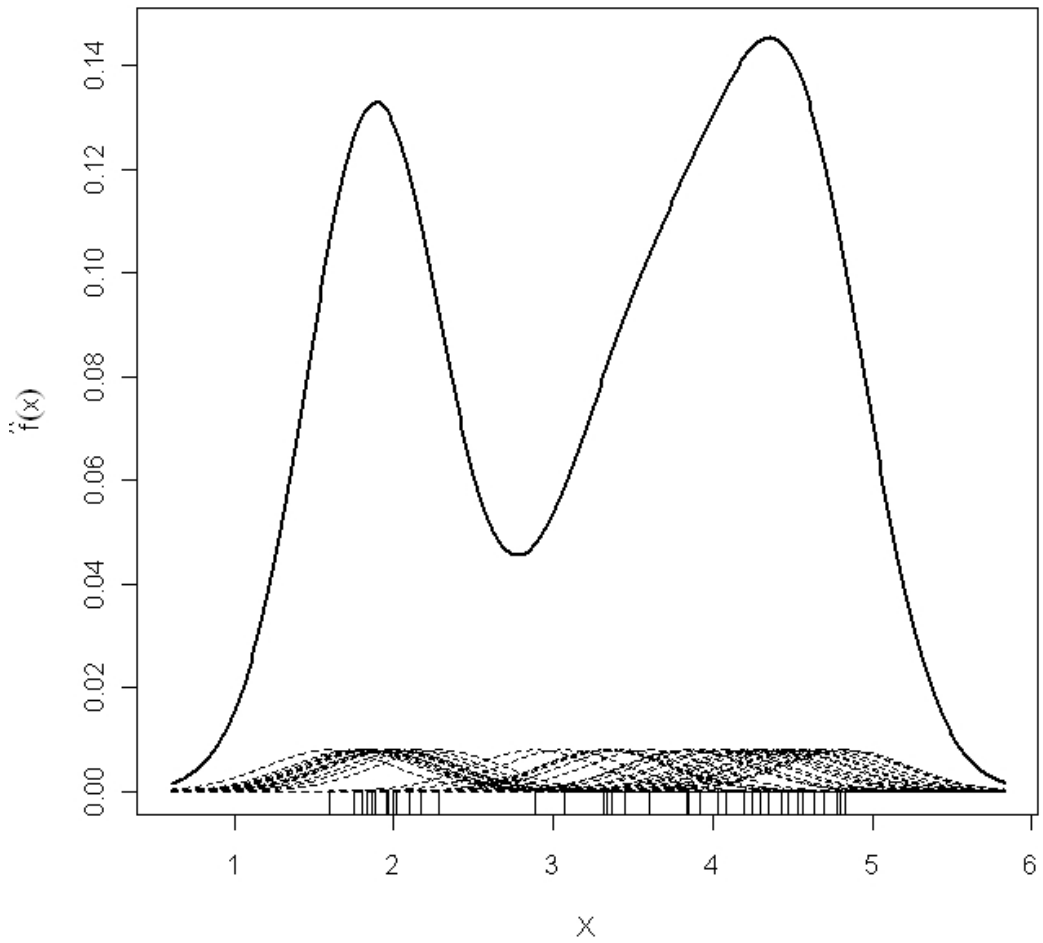
Gauß–Kernschätzer auf $\mathbb{R}$ ($n = 5$)

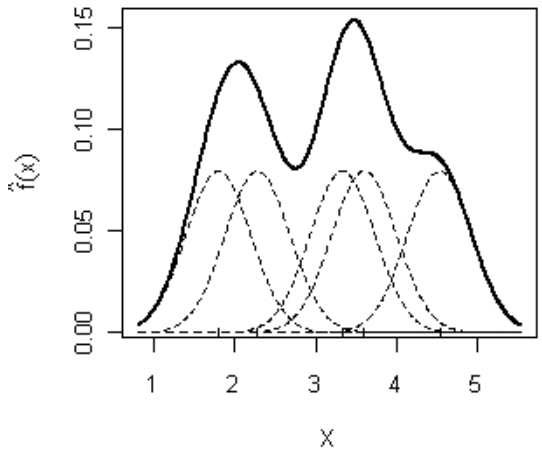


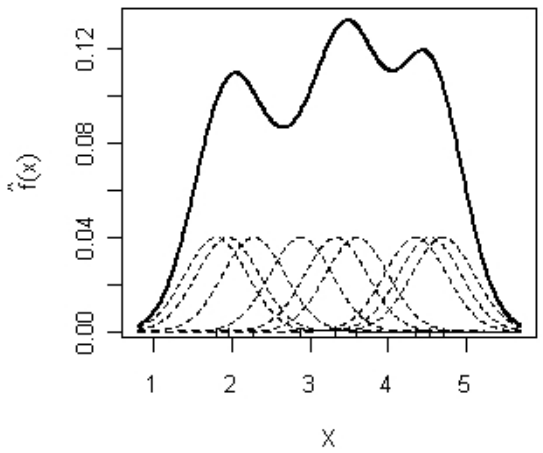
Gauß–Kernschätzer auf $\mathbb{R}$ ($n = 9$)

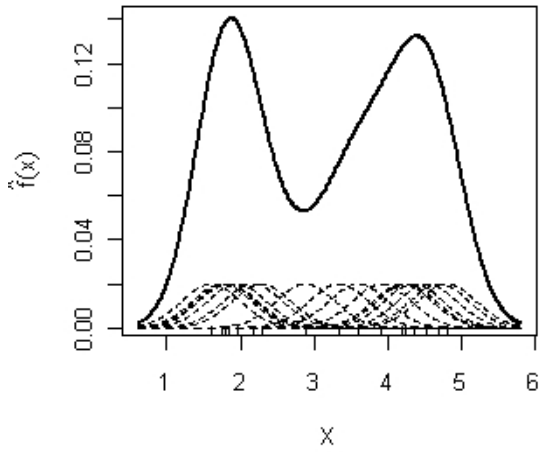


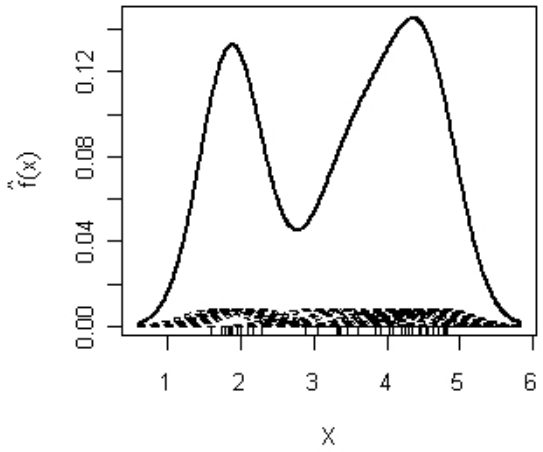
Gauß–Kernschätzer auf $\mathbb{R}$ ($n = 50$)




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

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

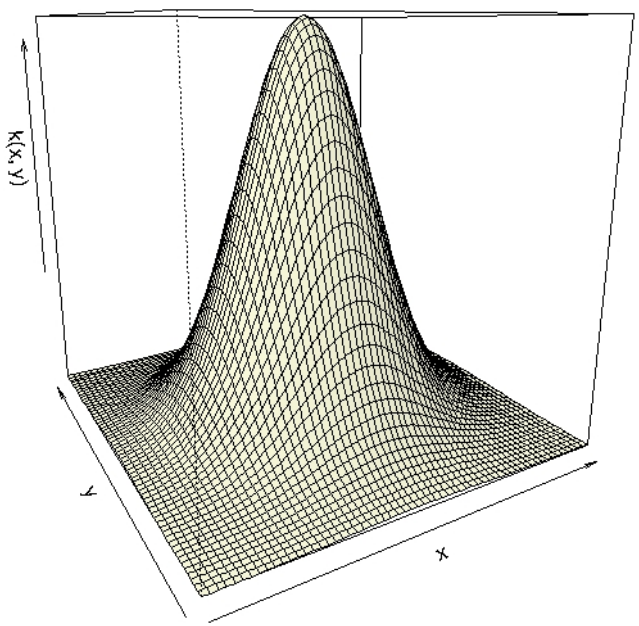
Sei $\mathcal{K} : \mathbb{R}^p \to \mathbb{R}$ ein Kern.

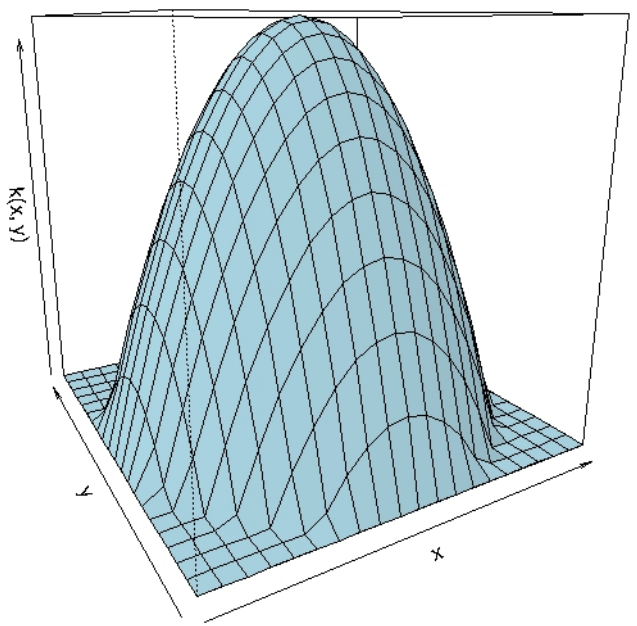
$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right), \ \mathbf{x} \in \mathbb{R}^p$$

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

Darstellung zweidimensionaler Kernfunktionen

Gaußkern und Epanechnikovkern mit $p = 2$





Bandweitenwahl

Bias und Varianz:

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

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

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:

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

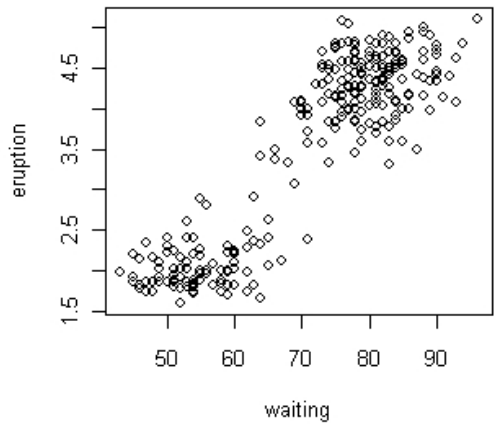
Zuvor: $H = h\mathbb{1}_p$, wobei $\mathbb{1}_p$ die p–dimensionale Einheitsmatrix bezeichnet

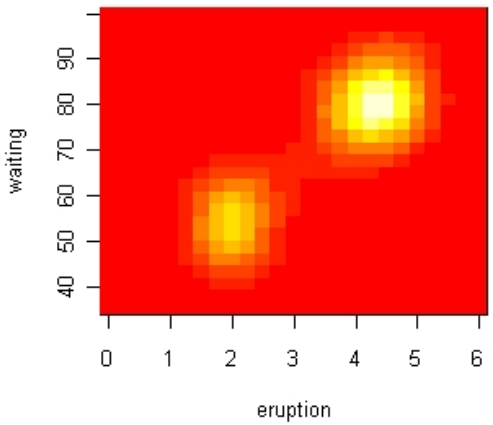
In R: Diagonalmatrix H angebbar.

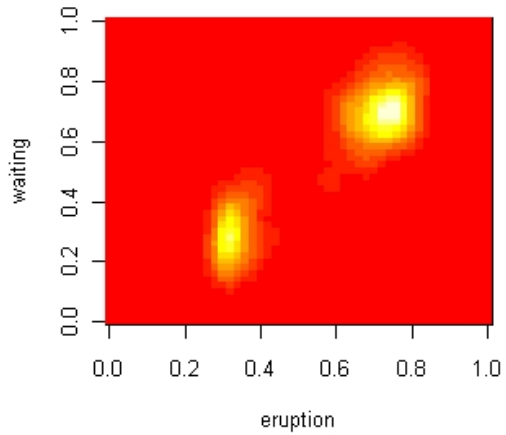
R Code: Zweidimensionale Kerndichteschätzung

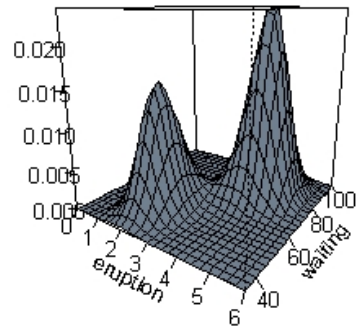
```
> library(MASS)
> library(KernSmooth)
> data(faithful)
> x<-faithful$eruptions
> y<-faithful$waiting

> 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



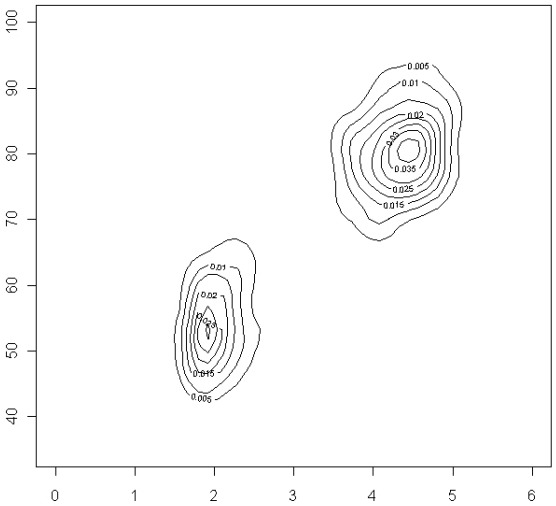


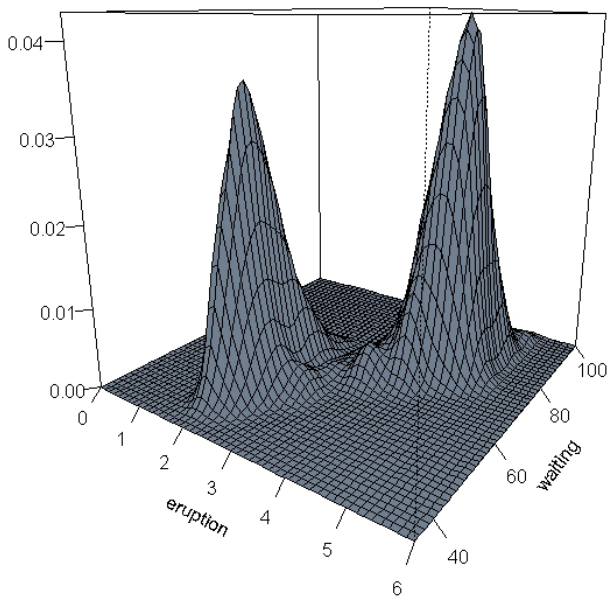




Kontur-Plots, 3D-Plots

```
> 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="")
```

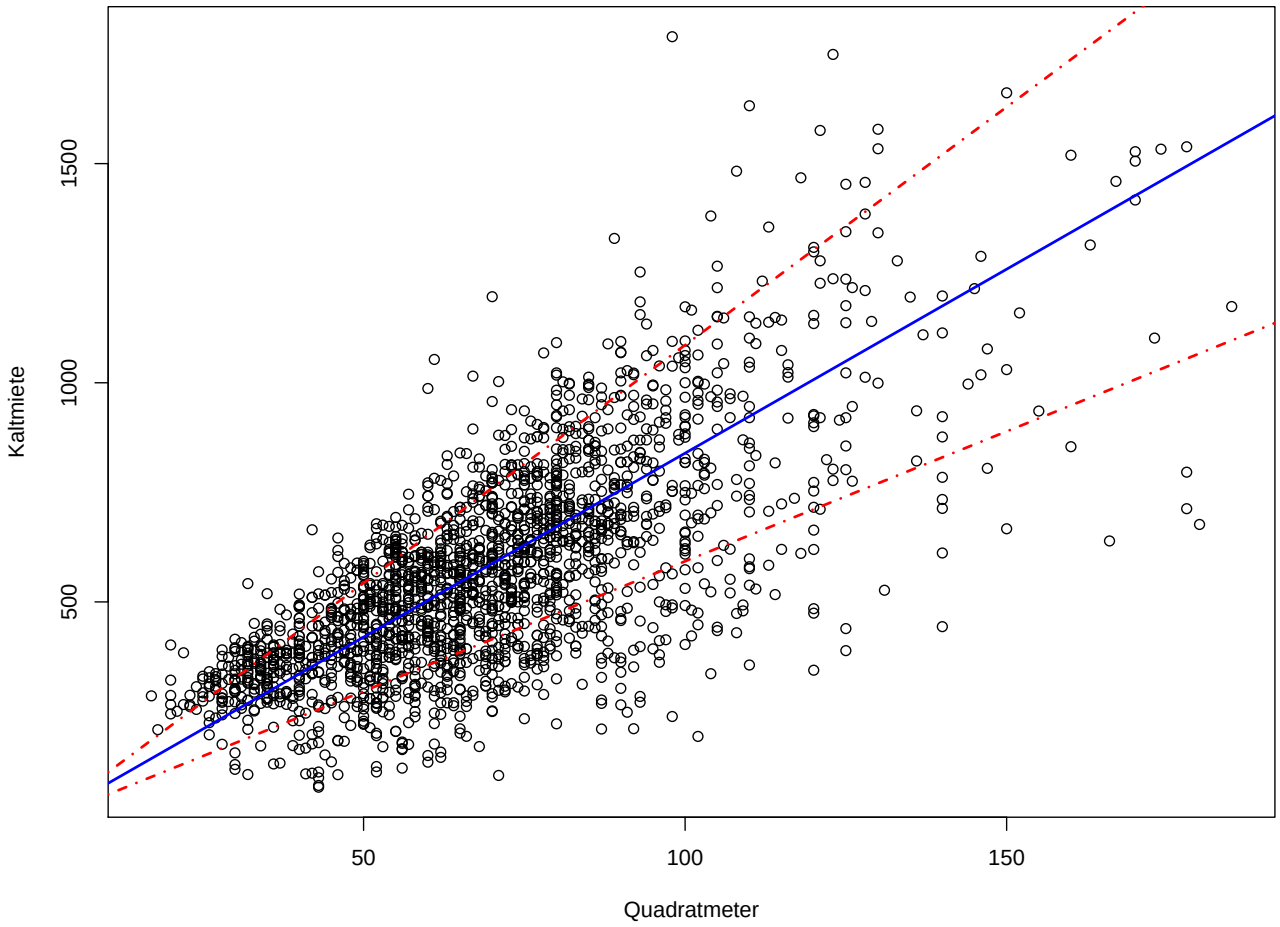




Zusammenhänge zwischen stetigen Variablen

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