Kapitel 5
Deskriptive Statistik

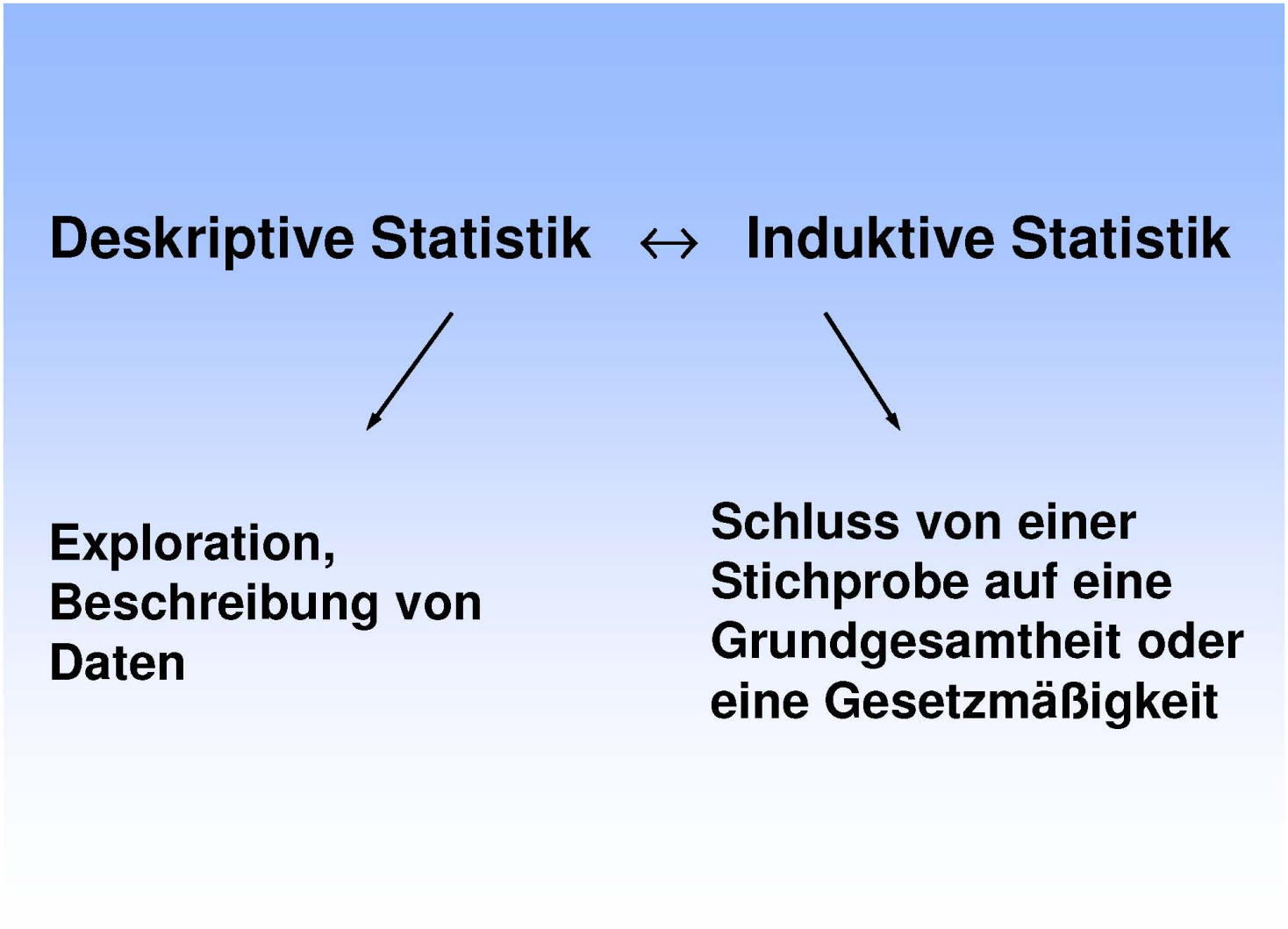
Thorsten Dickhaus

Universität Bremen
Institut für Statistik

Mathematik 3: Stochastik
Universität Bremen, Fachbereich 03, SoSe 2025

Übersicht

Übersicht


Deskriptive Statistik ↔ Induktive Statistik
Exploration, Beschreibung von Daten
Schluss von einer Stichprobe auf eine Grundgesamtheit oder eine Gesetzmäßigkeit

Deskriptive (beschreibende, explorative) Statistik

"Das Ziel der beschreibenden Statistik besteht darin, die bei Beobachtungen, Erhebungen und Experimenten anfallenden Daten so aufzubereiten, dass sie durchschaubar werden." (Lorenz, 1992).

Aufgaben

- Zusammenfassende und übersichtliche Darstellung quantitativer Aspekte beobachteter Sachverhalte
- Zusammenstellung von Ergebnissen in
  - ✓ Tabellen
  - ✓ Graphiken

Beobachtungseinheit (Versuchseinheit, Merkmalsträger, experimental unit, case, sampling unit)

ist die kleinste Einheit, an der die Beobachtungen durchgeführt werden.

Beispiele:

- Proband:innen
- Tiere
- Zellkulturplatten
- Gruppen von Individuen
  - ✓ Zwillingspaare
  - ✓ Nachkommen
- Bauteile

Merkmale

Merkmal (Variable)

ist messbare Eigenschaft einer Beobachtungseinheit,

z.B.

- ✓ Körpergröße
- ✓ Symptom
- ✓ Laborwerte
- ✓ Geschlecht

Merkmalsausprägung

Werte, die die Messung eines Merkmals ergeben

z.B.

- ✓ 181 cm bei Merkmal Körpergröße
- ✓ männlich/weiblich/divers bei Merkmal Geschlecht

Von Information zu Daten

Information

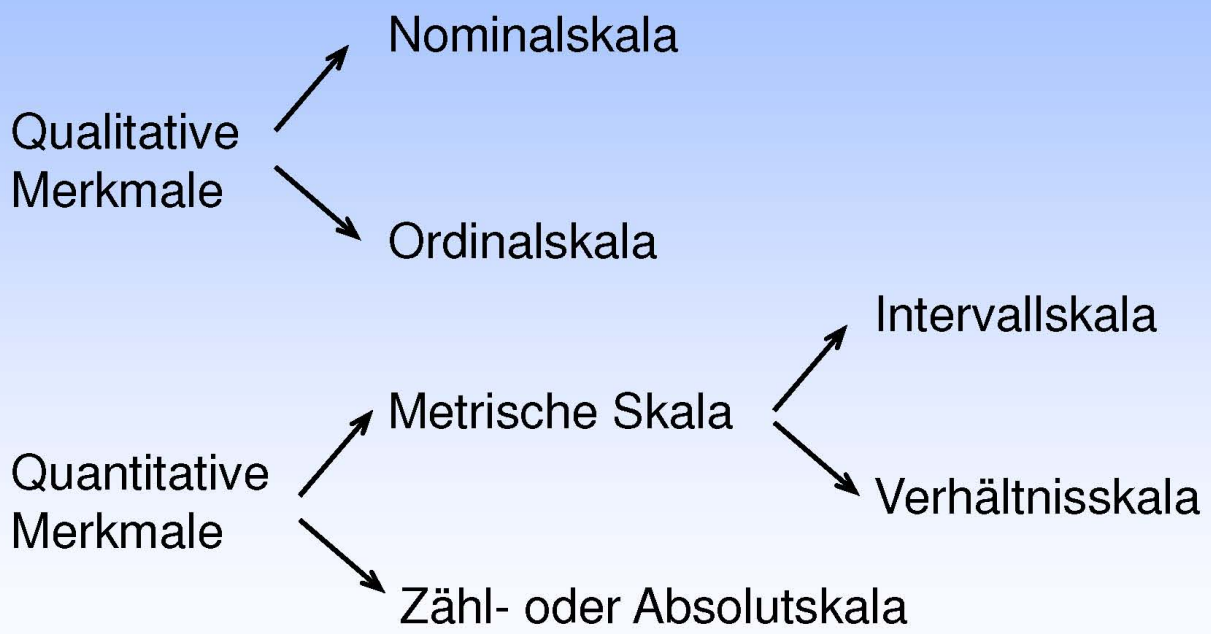
- Messwerte (Labor)
- Texte (Diagnosen)
- Bilder (EKG, CT...)
- Audio/Video
- .....

Datenmatrix

Nr.	M_1	M_2	...	M_p
1				
2				
...				
n				

Übersicht

Skalenniveaus



Nominalskala

„Ein Merkmal heißt nominalskaliert, wenn die Ausprägungen Namen oder Kategorien sind, die den Einheiten zugeordnet werden, wie zum Beispiel Farben oder Religionszugehörigkeit.

Den Ausprägungen solcher qualitativen Merkmale werden häufig aus technischen Gründen Zahlen zugewiesen, die dementsprechend nur der Kodierung dienen und denselben Zweck wie Namen erfüllen. Diese zugeordneten Zahlen sind also nur Stellvertreter, deren numerischer Wert nicht als solcher benutzt werden sollte, d.h. übliche Rechenoperationen wie Addition oder Multiplikation lassen sich nicht mehr sinnvoll durchführen.“

(Seiten 15-16 in Fahrmeir et al. (2016): Statistik: Der Weg zur Datenanalyse, 8. Auflage, Springer-Verlag)

Ordinalskala

„Bei einem ordinalskalierten Merkmal können die Ausprägungen geordnet werden, aber ihre Abstände sind nicht interpretierbar.

Ein klassisches Beispiel für ein ordinalskaliertes Merkmal sind Schulnoten. Zwar weiß man, dass die Note 1 besser ist als die Note 2, aber der Abstand zwischen 1 und 2 lässt sich sicherlich nicht interpretieren oder vergleichen etwa mit demjenigen zwischen 4 und 5.“

(Seite 16 in Fahrmeir et al. (2016): Statistik: Der Weg zur Datenanalyse, 8. Auflage, Springer-Verlag)

Intervallskala

- Abstände zwischen den Ausprägungen, die Zahlen sind, können interpretiert werden.
- Intervallskalen haben im Gegensatz zu Verhältnisskalen keinen natürlichen, sondern einen willkürlich gesetzten Nullpunkt oder Mittelpunkt.
- Beispiel: Intelligenzquotient (IQ):
  Mittelpunkt willkürlich bei 100 gesetzt
- Man kann keine Aussagen über Verhältnisse machen:
  Eine Person mit einem IQ von 140 ist also nicht „doppelt so intelligent“ wie eine Person mit einem IQ von 70!

Verhältnisskala

Sind zusätzlich (zu den Gegebenheiten der Intervallskala) Quotienten von Ausprägungen interpretierbar, so heißt das Merkmal verhältnisskaliert.

Damit sind z. B. Nettomiete, Kredithöhe oder Semesteranzahl Beispiele für verhältnisskalierte Merkmale.

Häufig werden die Intervall- und die Verhältnisskala zur sogenannten Kardinalskala zusammengefasst. Ein kardinalskaliertes Merkmal wird zudem als metrisch bezeichnet.

(Seite 16 in Fahrmeir et al. (2016): Statistik: Der Weg zur Datenanalyse, 8. Auflage, Springer-Verlag)

Zähl- oder Absolutskala

- Abstände sind messbar.
- Nullpunkt und Einheit sind natürlich
  (also nicht willkürlich gewählt).
- Beispiele: Häufigkeiten oder alles, was sich zählen lässt.

	sinnvoll interpretierbare Berechnungen			
Skalenart	auszählen	ordnen	Differenzen bilden	Quotienten bilden
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
Intervall	ja	ja	ja	nein
Verhältnis	ja	ja	ja	ja

Tabelle 1.6: Sinnvolle Berechnungen für Daten verschiedener Skalen

(Seite 16 in Fahrmeir et al. (2016): Statistik: Der Weg zur Datenanalyse, 8. Auflage, Springer-Verlag)

Ein Merkmal heißt dichotom (oder binär), wenn es genau zwei Ausprägungen annehmen kann

Beispiele:

- ✓ Prüfung bestanden (ja/nein)
- ✓ Krankheit vorhanden/nicht vorhanden
- ✓ Münzwurfergebnis (Kopf/Zahl)

Beispiele

✓ Geschlecht	nominal
✓ Blutgruppe	nominal
✓ Symptome	nominal
✓ Schmerzintensität	ordinal
✓ Tumorstadium	ordinal
✓ Alter [in Jahren]	Verhältnis-skaliert
✓ Herzfrequenz	Verhältnis-, absolut skaliert
✓ Gewicht	Verhältnis-skaliert
✓ Größe	Verhältnis-skaliert
✓ Blutdruck	Verhältnis-skaliert
✓ Anzahl Thrombozyten	Verhältnis-, absolut skaliert

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
```

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

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

(Empirischer) Median:
$med(\mathbf{x}) = x_{[(n+1)/2]}$ bzw. $med(\mathbf{x}) = \left(x_{[n/2]} + x_{[(n+2)/2]}\right)/2$:

```
> 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) einer Zufallsvariable X:

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

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
```

Statistische Kenngrößen: Der Modus

Modus x_{mod}: Ausprägung mit größter Häufigkeit

Der Modus ist das wichtigste Lagemaß für kategoriale Merkmale und bereits auf Nominalskalenniveau sinnvoll.

In der Darstellung durch Stabdiagramme ist der Modus die Ausprägung mit dem höchsten Stab.

Der Modus ist eindeutig, falls die Häufigkeitsverteilung ein eindeutiges Maximum besitzt.

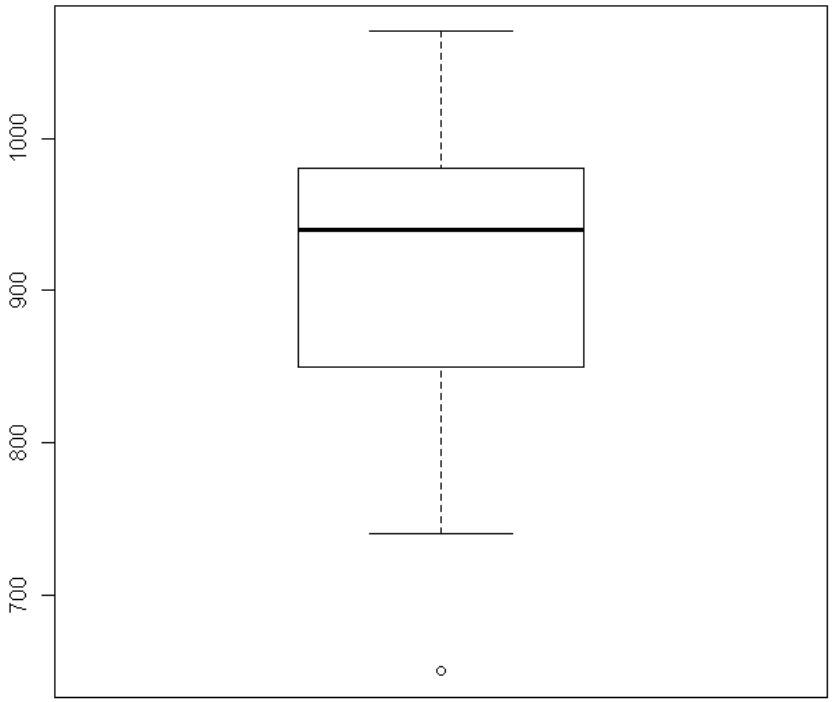
(Seite 53 in Fahrmeir et al. (2016): Statistik: Der Weg zur Datenanalyse, 8. Auflage, Springer-Verlag)

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)



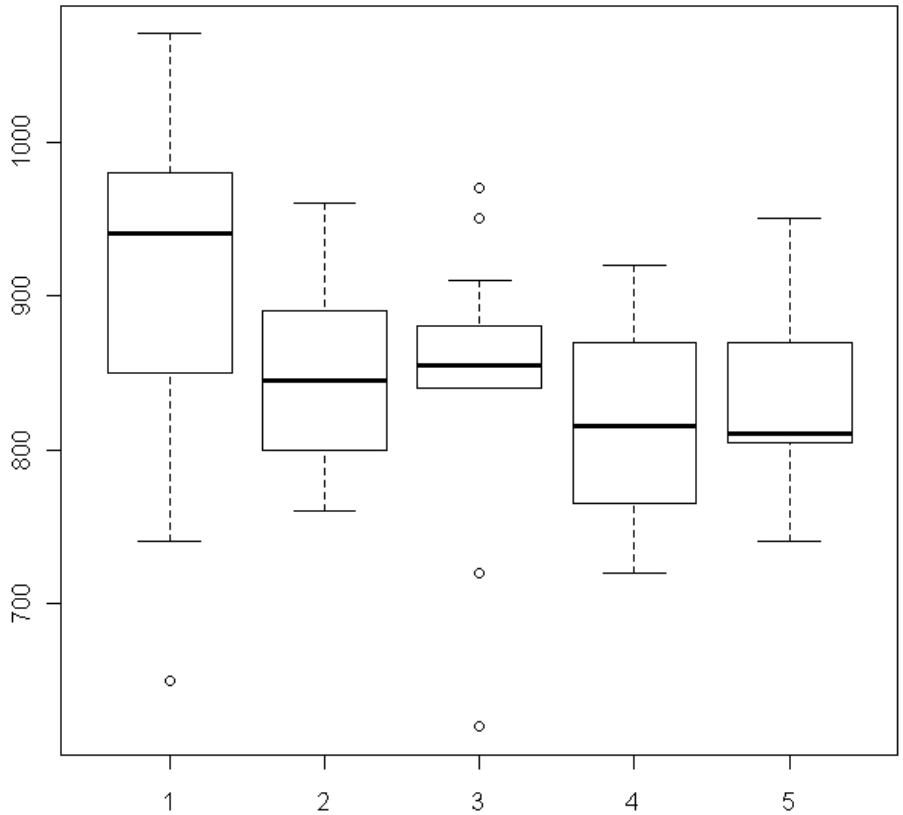
R Code: Herausnehmen von Ausreißern

```
> 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)
```



Zum Begriff i.i.d.

Der englische Begriff i.i.d. bedeutet „(stochastically) independent and identically distributed“, also „(stochastisch) unabhängig und identisch verteilt“.

Ein i.i.d.-Modell ist ein mathematisches Modell für eine repräsentative Stichprobe aus einer homogenen Grundgesamtheit.

Die i.i.d.-Modelle bilden eine wichtige Grundlage der Inferenzstatistik.

Empirische Verteilungsfunktion

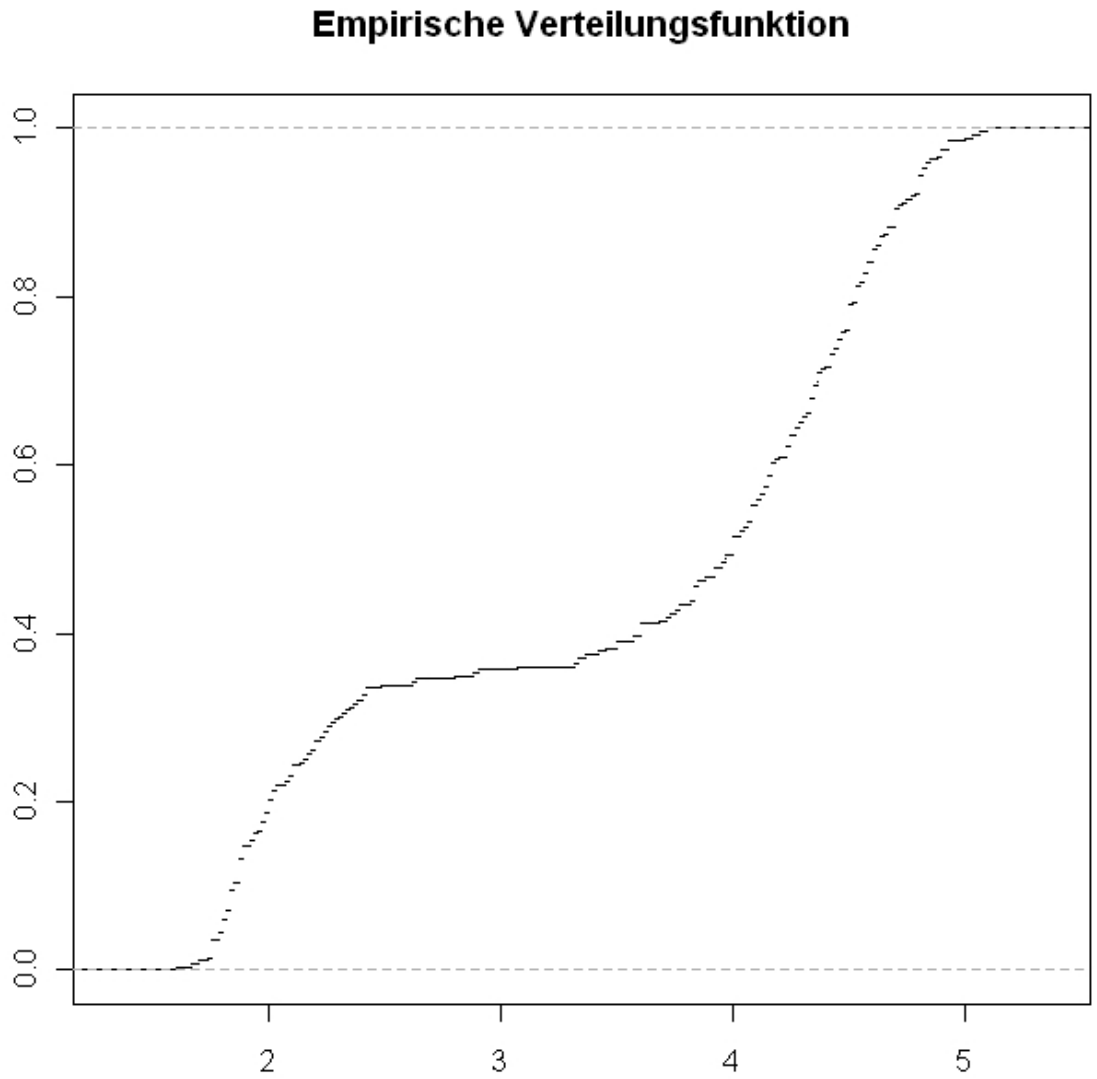
Seien $X_1, \ldots, X_n$ reellwertige i.i.d. Zufallsvariablen mit $X_1 \sim F$.

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

Satz von Glivenko–Cantelli liefert fast sichere gleichmäßige Konvergenz: Mit W'keit Eins gilt

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

```
> 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

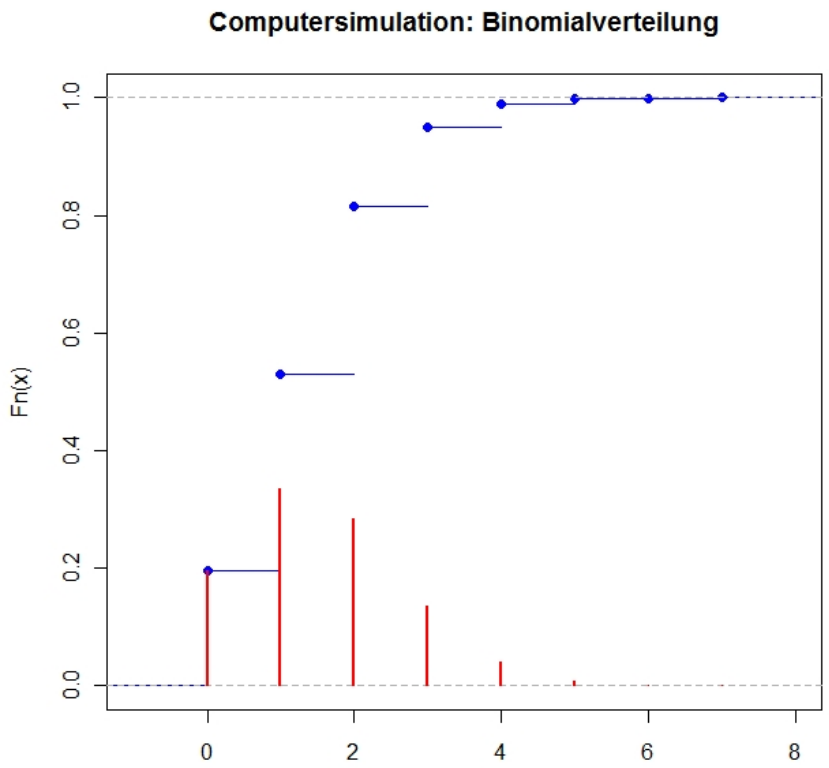
Diskrete Merkmale, Stabdiagramme

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 i.i.d. 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

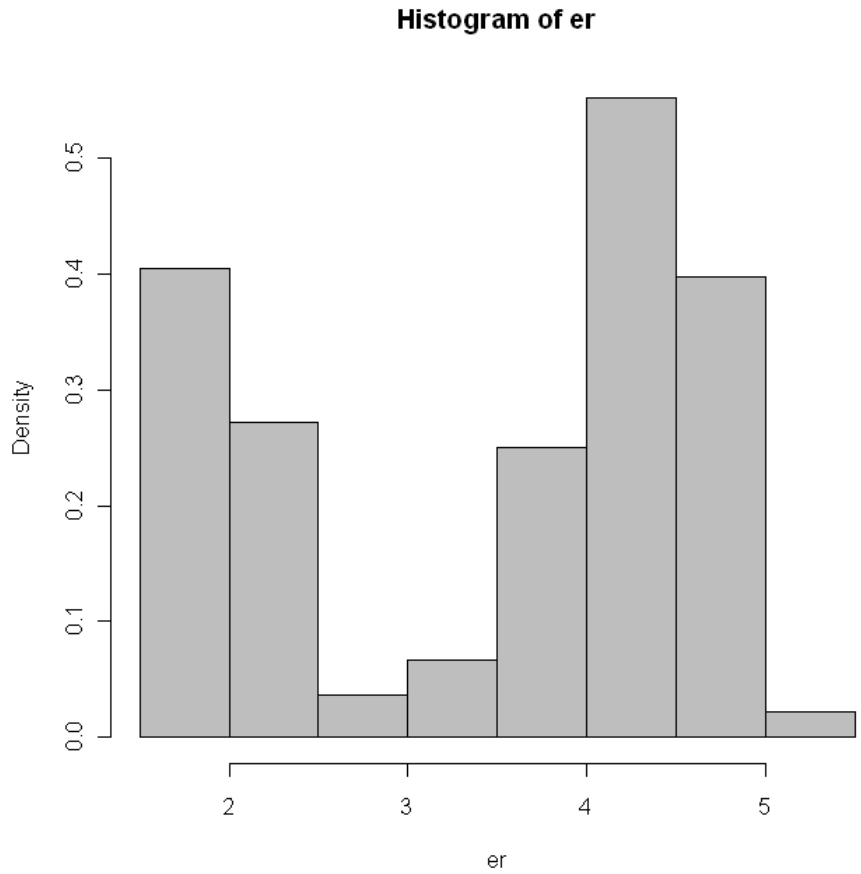
$$I_k = (a_{k-1}, a_k], \; k \in \{1, \ldots, K\}$$

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

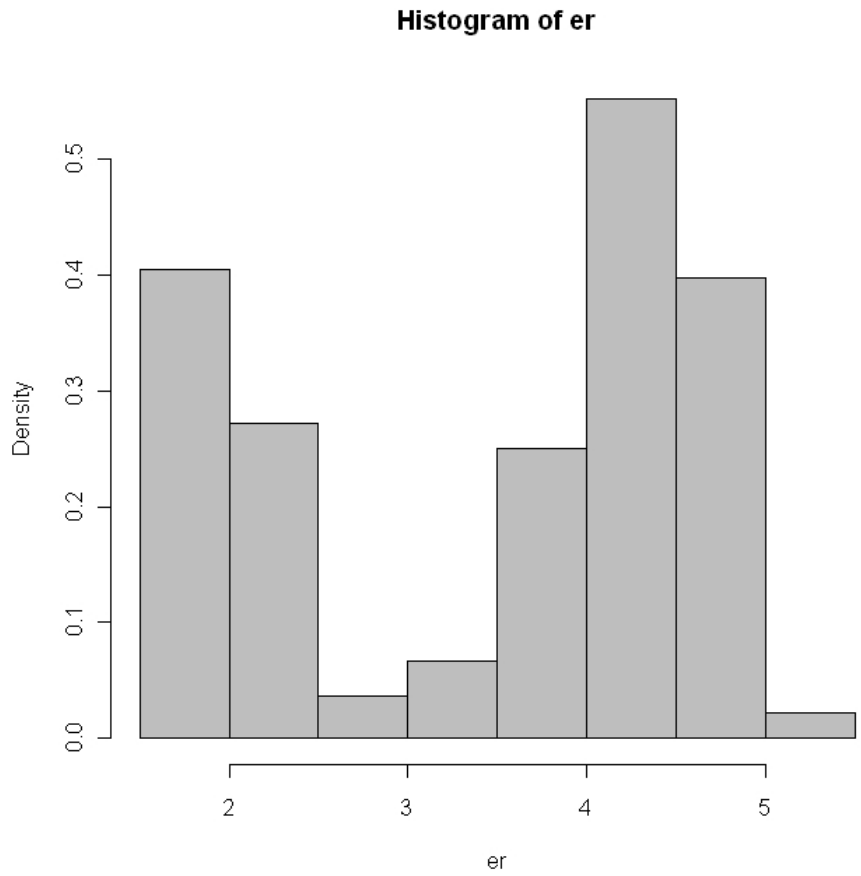
$$\hat{f}_{hist}(x) = \frac{n_k}{n} \frac{1}{a_k - a_{k-1}} \mathbb{1}_{\{I_k\}}(x)$$

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")
```



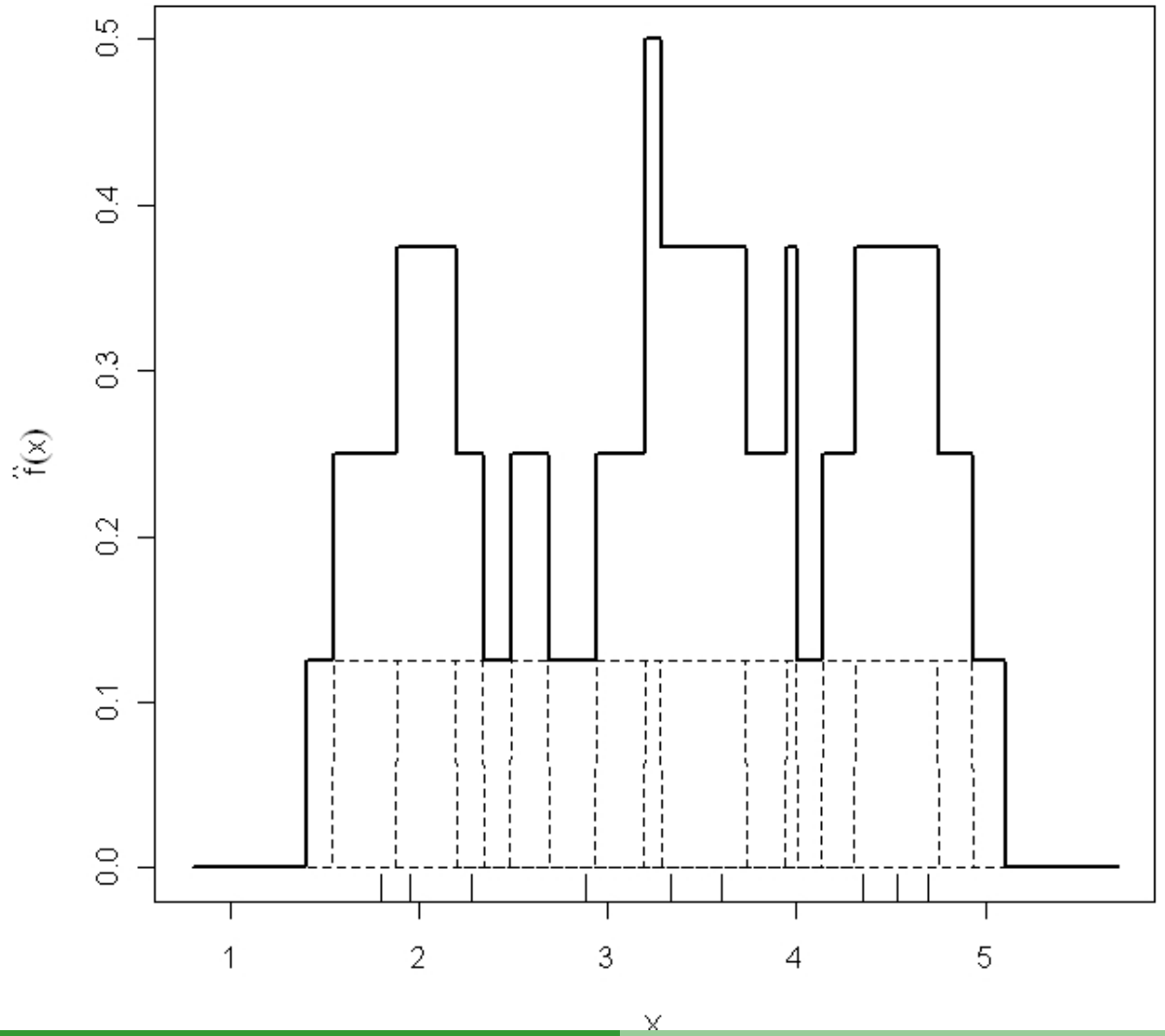
Nachteil des Histogramm–Schätzers:
Schätzer hängt von der Wahl der Klassen–Längen und des Startwertes a_0 ab!

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:

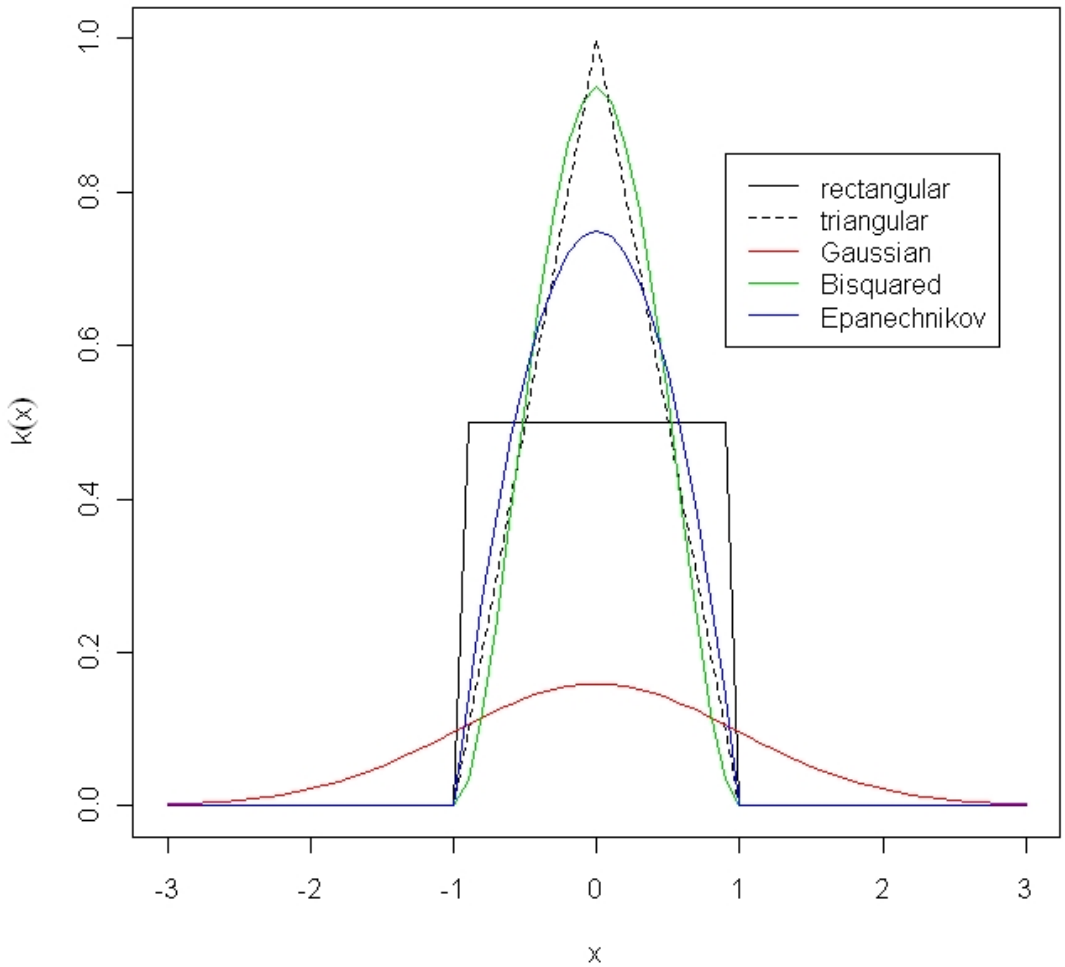
2. $\sup_{x \in \mathbb{R}} \mathcal{K}(x) = M < \infty$
3. $|x|\mathcal{K}(x) \to 0$ für $|x| \to \infty$, $\int x^2 \mathcal{K}(x)dx =: k_2 < \infty$

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.

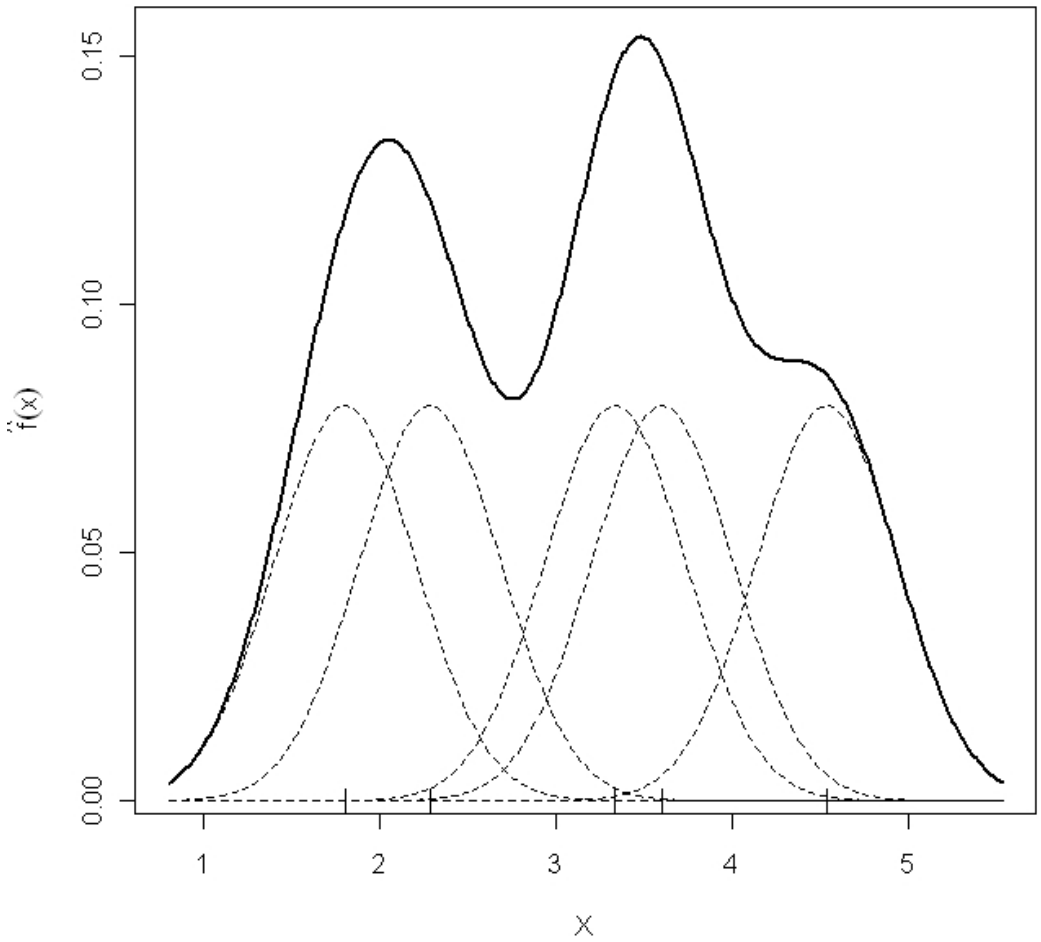
$$\hat{f}_n(t) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{t - x_i}{h}\right) = \int \frac{1}{h} \mathcal{K}\left(\frac{t - x}{h}\right) \hat{F}_n(dx)$$

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

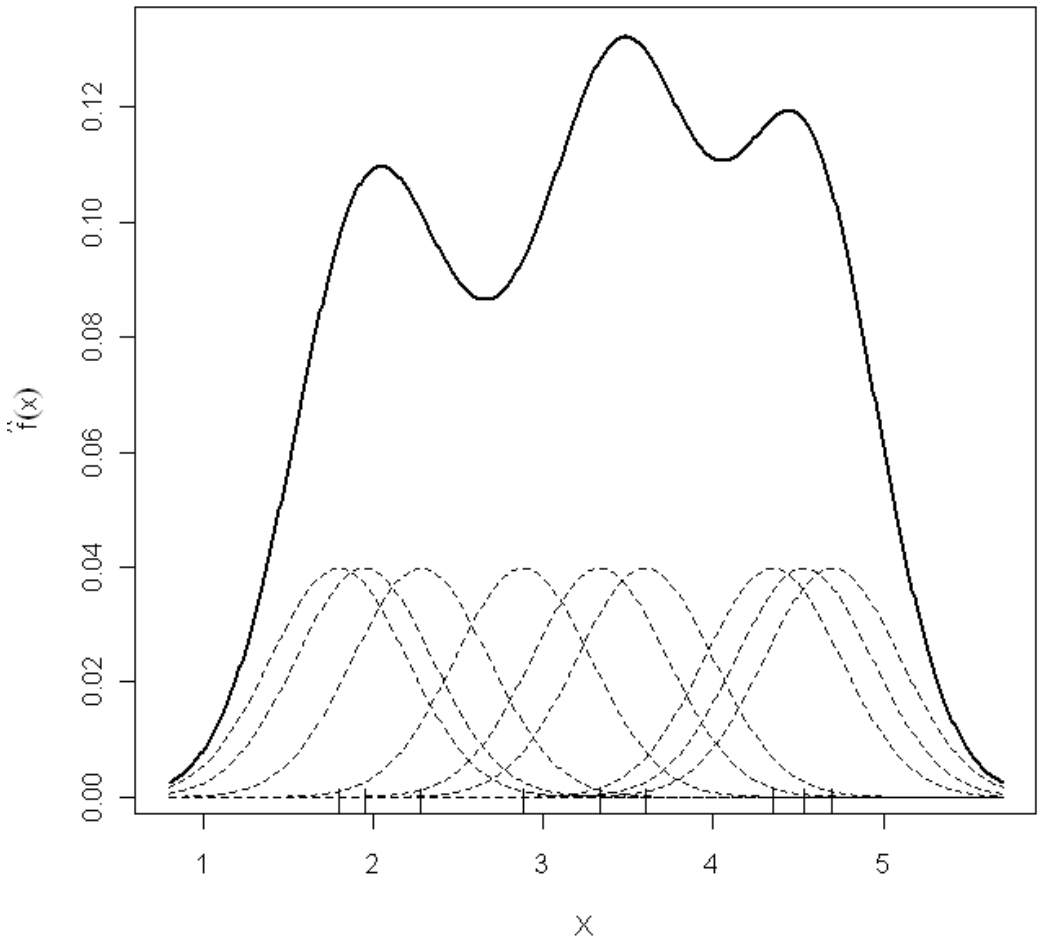
Mit $\tilde{\mathcal{K}}(t) := \int_{-\infty}^{t} \mathcal{K}(x)dx$ lässt sich auch $F(t)$ schätzen durch

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

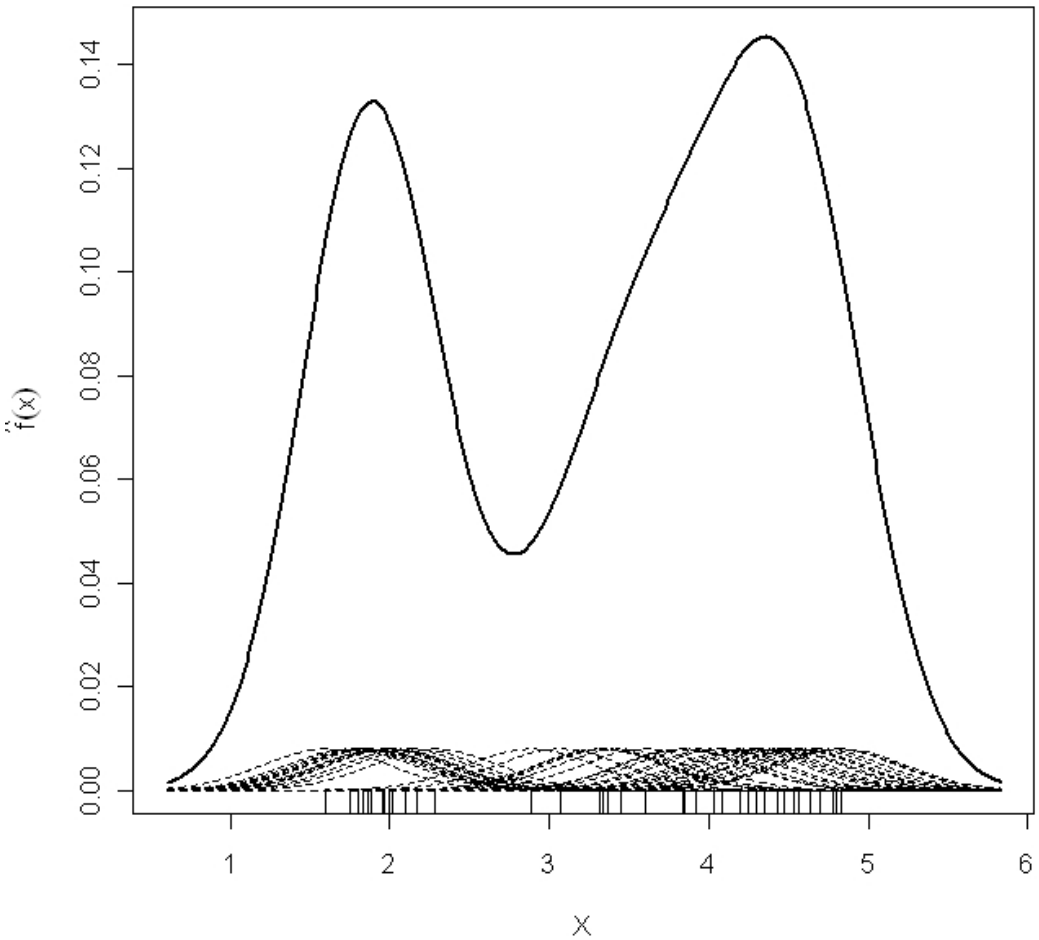
Gauß–Kernschätzer ($n = 5$)

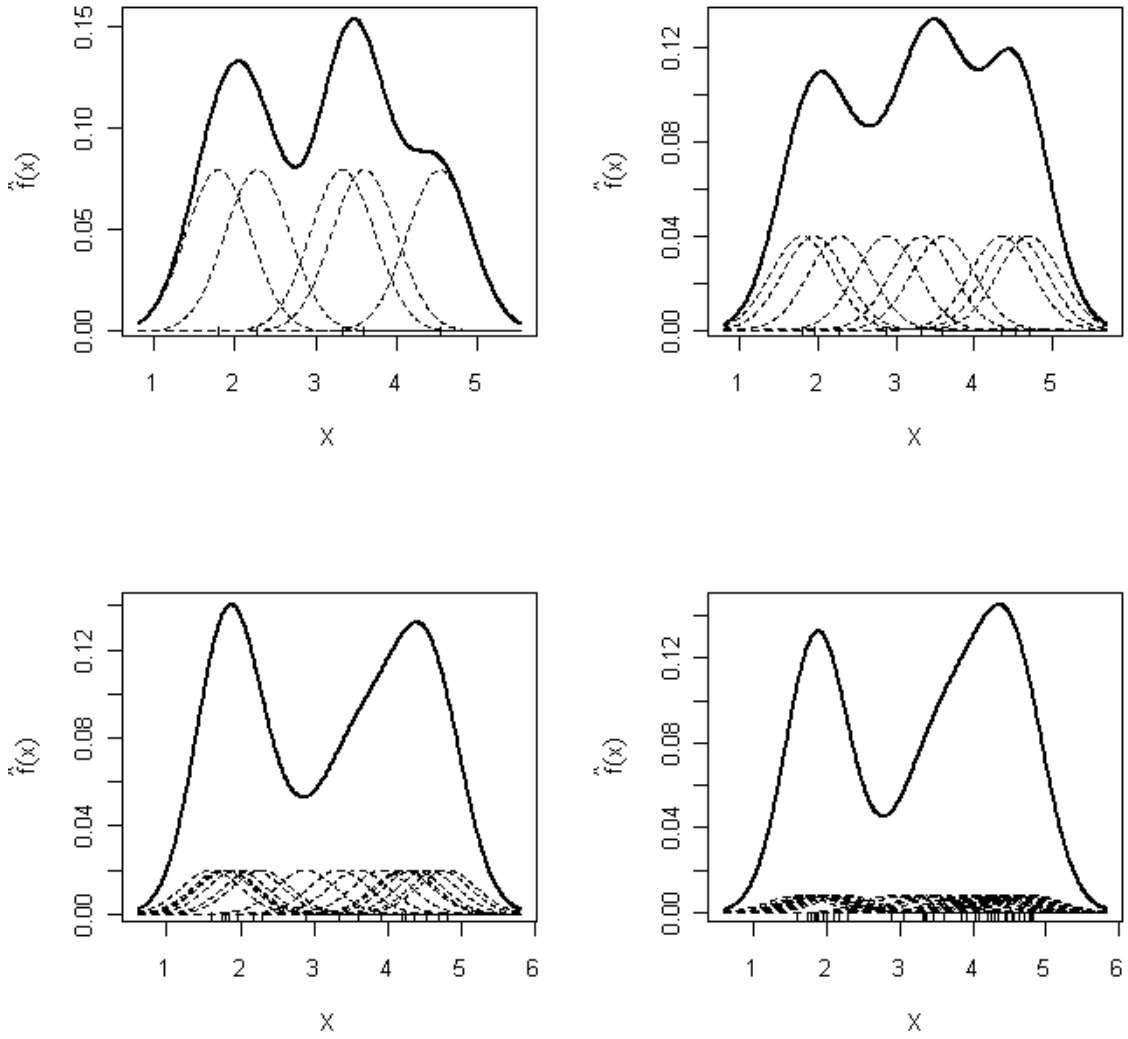


Gauß–Kernschätzer ($n = 9$)



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“

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 $\mathbf{x}$: 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

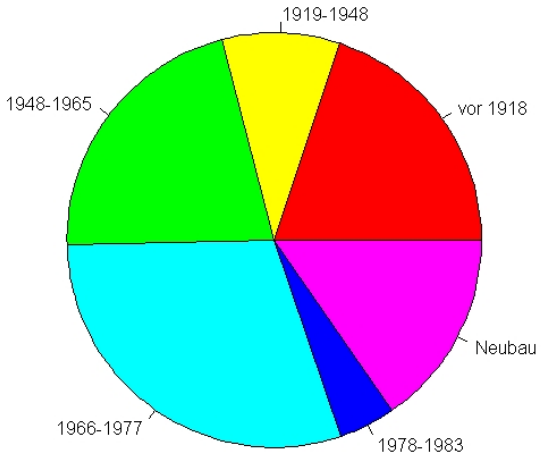
Randhäufigkeiten, marginale Verteilungen

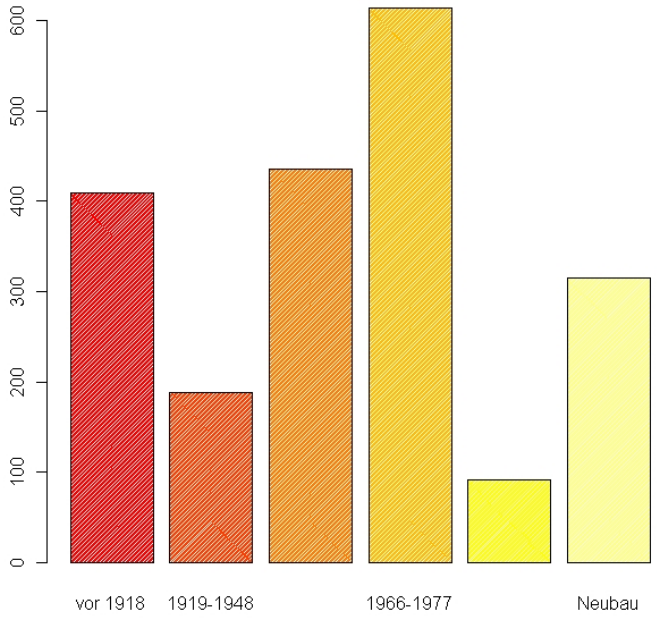
Der Vektor $\mathbf{n} = (n_{1.}, n_{2.}, \ldots, n_{k.}, n_{.1}, n_{.2}, \ldots, n_{.\ell}) \in \mathbb{N}^{k+\ell}$ heißt Vektor der (empirischen) Randhäufigkeiten.

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



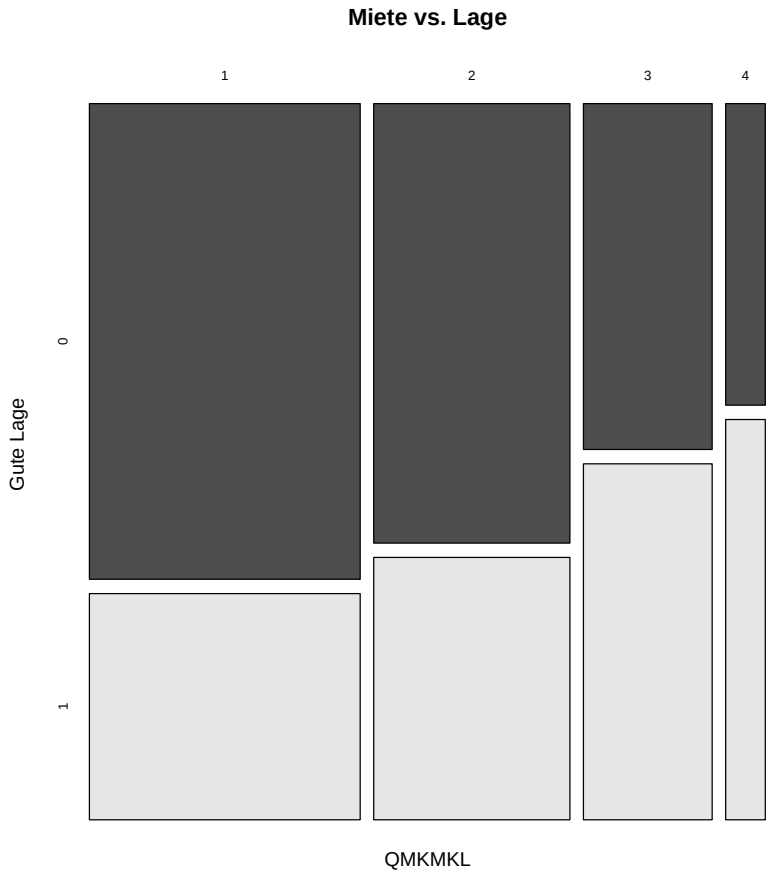


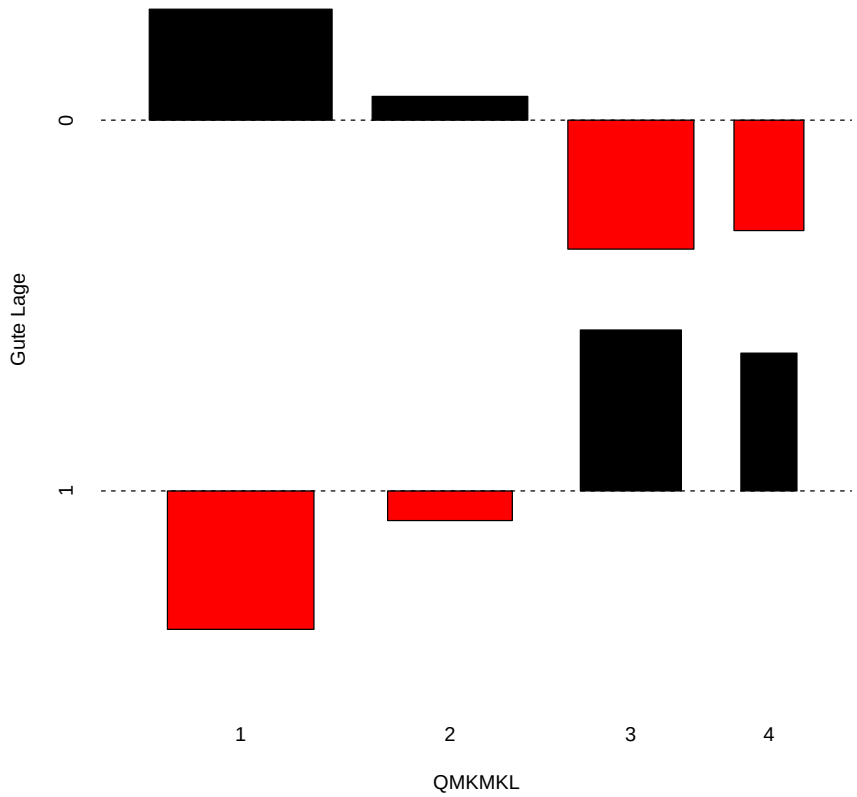
Grafische Darstellung bivariater diskreter Verteilungen

R Code: **mosaicplot** und assocplot

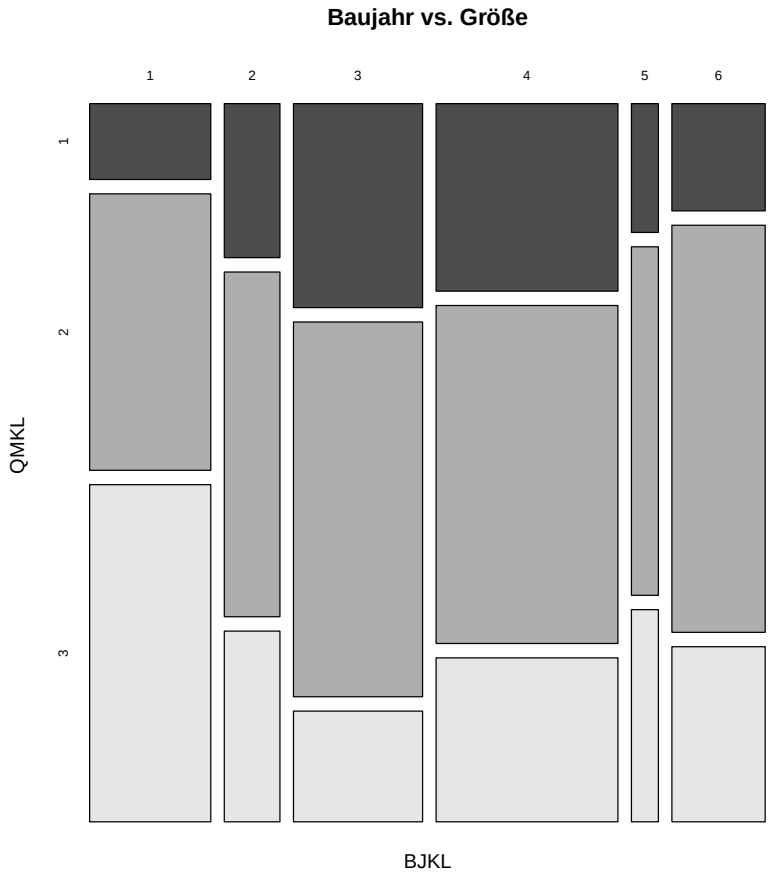
```
> miete$QMKMKL<-1*(QMKM<=8)+2*(QMKM>8)*(QMKM<=10)
  +3*(QMKM>10)*(QMKM<=12)+4*(QMKM>12);

> 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");
```

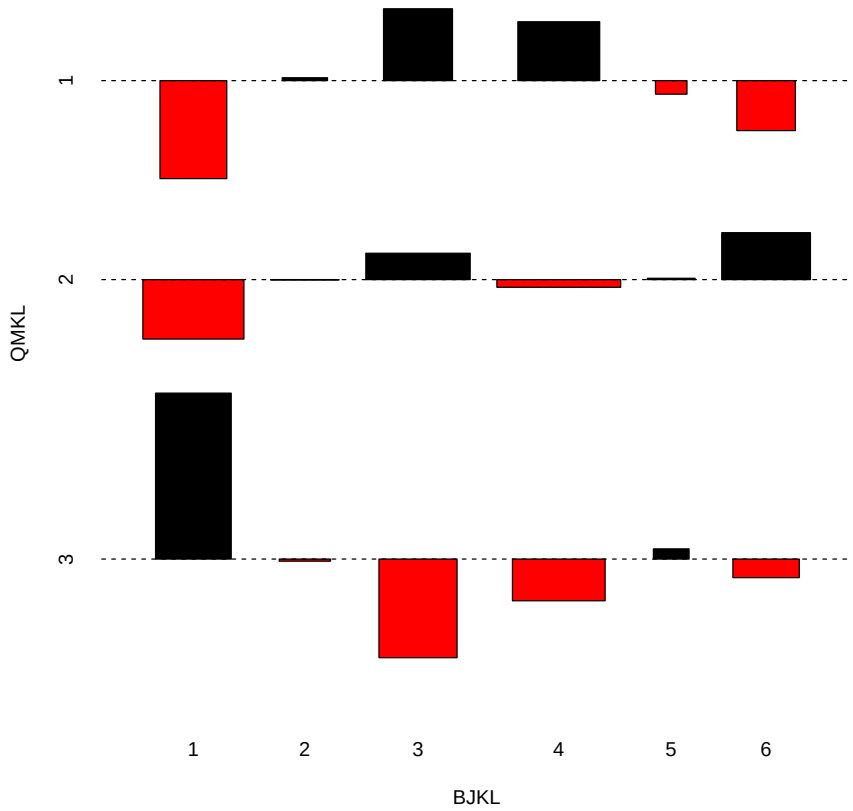
Miete versus Wohnlage





Baujahr versus Wohnungsgröße





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);
```

Scatter-Plot (Streubild) und Ausgleichsgerade