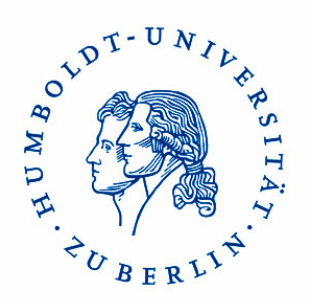
Stochastik–Praktikum
Deskriptive Statistik

Dirk Becherer

Humboldt-Universität zu Berlin

5. November 2014



Übersicht

Übersicht

Kenngrößen univariater Verteilungen

Endlichdimensionale Kenngrößen:

- Mittelwert $E[X]$
- Median $\frac{1}{2}(F^{-1}(0.5) + F^{-1}(0.5-))$
- Quantile $Q_\alpha = F^{-1}(a) = \inf\{x : F(x) \geq \alpha\}$ für $\alpha \in (0, 1)$
- Varianz und Standardabweichung $\sigma^2 = Var[X]$ bzw. $\sigma = \sqrt{Var[X]}$
- allgemeiner: un/zentrierte Momente $E[X^k]$ bzw. $E[(X - EX)^k]$ f?r $k \in \mathbb{N}$
- Skewness $E[(X - EX)^3]/\sigma^3$, Kurtosis $E[(X - EX)^4]/\sigma^4$,...

Kenngrößen univariater Verteilungen

die unendlich dimensionalen Kengrößen:

- (kumulative) Verteilungsfunktion $F(x) = P[X \leq x]$, $x \in \mathbb{R}$
- charakteristische Funktion $\phi(t) = E[\exp(itX)]$

beschreiben univariate Verteilungen vollst?ndig.

Kenngrößen für Stichproben

typischerweise auf eine von zwei Weisen berechnet

- wie zuvor bloßbezüglich der empirischen Verteilung $\frac{1}{n}\sum_1^n \delta_{x_i}$ der Stichprobe,
- ggf. modifiziert zu einem erwartungstreuem Schätzer der entsprechenden Kenngrö ßder zugrundeliegenden theoretischen Verteilung.

Kenngrößen multivariater Verteilungen

- Mittelwert und Median (Vektoren)
- gemischte (zentrierte) Momente
- Kovarianzmatrix (alle zentrierten Momente der Ordnung 2)
- gemeinsame charakteristische Funktion oder Verteilungsfunktion
- Randverteilungen und Copula der gemeinsamen Verteilung

Letztere beschreiben Verteilung eindeutig.

Übersicht

Visualisierungen

von uni- bzw. multivariaten Verteilungen/Stichproben in $\mathbb{R}^d$

- Dichtefunktion bzw Histogramm (skaliert auf PDF-Form)
- Regressionsgeraden, -polynome, oder ...
- Quantile/Quantile-Plots (QQ-Plots)
- Scatterplot
- Copula-Scatterplot

Visualisierungen

von uni- oder multivariaten kategorialen Datenstichproben

- Kuchendiagramm
- Balkendiagramme...
- Kontingenztafeln

Beispiele: Studienfächer, Nationalitäten, Geschlecht,...

Übersicht

Definition

X ist multivariat $\mathbf{N}(\mu, Q)$–verteilt in $\mathbf{R}^d$ mit Mittelwertvektor m und Kovarianzmatrix Q, falls für alle $\theta \in \mathbf{R}^d$ gilt

$$\theta^t X \sim \mathcal{N}(\theta^t m, \theta^t Q\, \theta)\text{-univariat normalverteilt}$$

Falls Matrix Q diagonal ist, werden Koordinaten von X unabhängig simuliert. Und anderenfalls ?

Definition

X ist multivariat $\mathbf{N}(\mu, Q)$–verteilt in $\mathbf{R}^d$ mit Mittelwertvektor m und Kovarianzmatrix Q, falls für alle $\theta \in \mathbf{R}^d$ gilt

$$\theta^t X \sim \mathcal{N}(\theta^t m, \theta^t Q\, \theta)\text{-univariat normalverteilt}$$

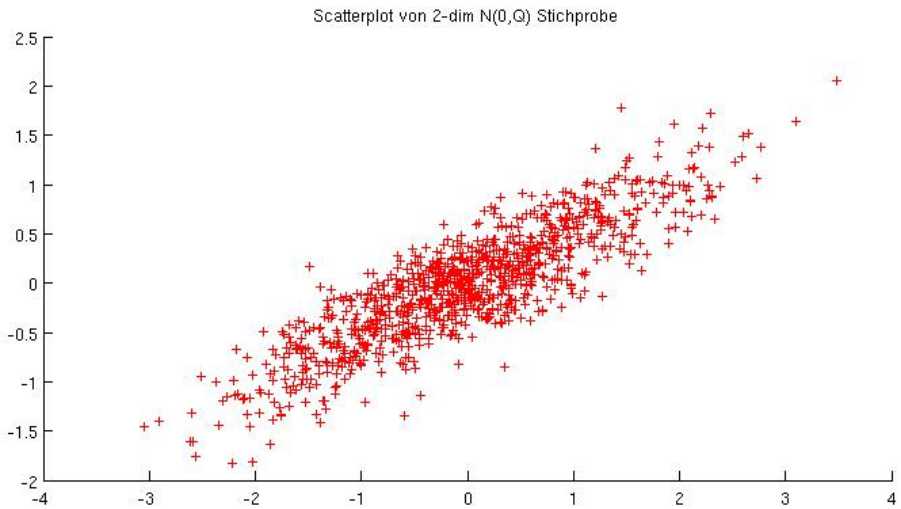
Für $Z \sim \mathcal{N}(0, I_d)$ standart multivariat normal ist

$$X := m + BZ \sim \mathcal{N}(m, BB^t) \quad m \in \mathbf{R}^d\,,\, B \in \mathbf{R}^{d\times d}$$

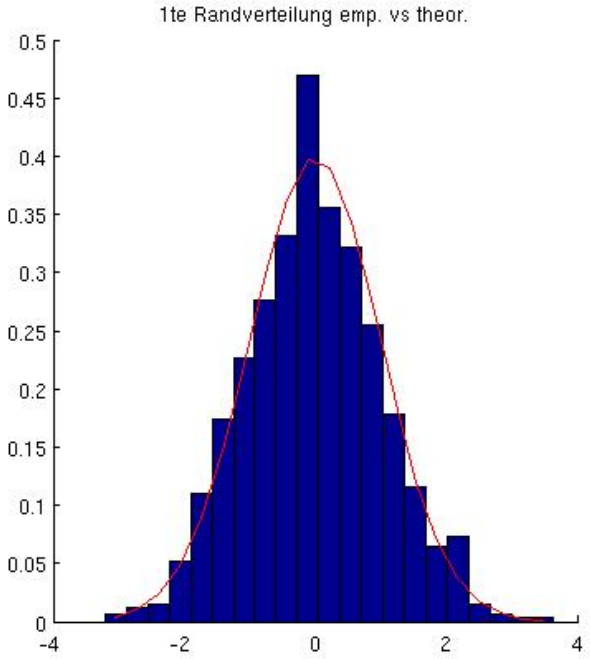
Für ggb. Kovarianz Q liefert Cholesky Zerlegung $Q = ADA^t$ ein $B := AD^{1/2}$ so dass $X := m + BZ \sim \mathcal{N}(m, Q)$-normalverteilt ist.

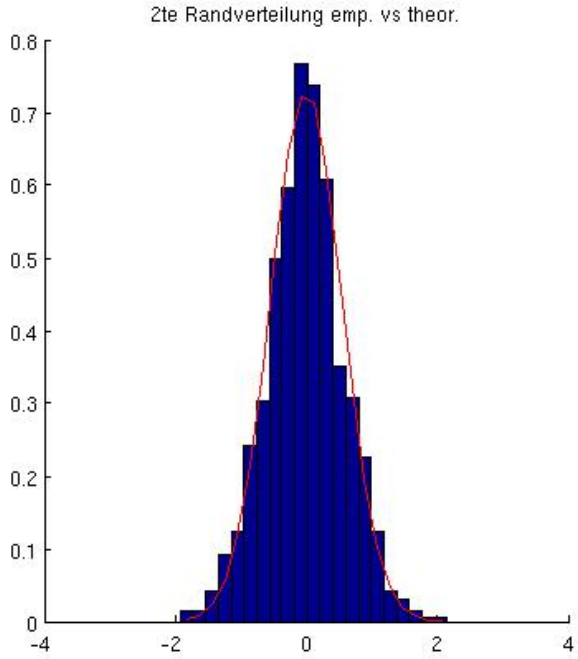
Scatterplot einer Stichprobe von 1000 Pseudozufallszahlen einer 2-dimensionalen $N(0, Q)$-Normalverteilung zu Kovarianz

$$Q = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.3 \end{pmatrix}$$



Dichte-Histogramme der empirischen Randverteilungen verglichen mit theoretischen Randverteilungsdichten:





Alternativ Übereinstimmung mit QQ-Plots prüfen

Übersicht

Experiment: Median vs Mean

Untersuche Verteilung von Mean und Median für Stichprobe mit $X \sim N(0,1)$, $Y = X^5$

```
for k=1:5
    disp('MC Experiment %i',k);
    X=randn(50,1);Y=X.^5;
    [mean(X) median(X);mean(Y) median(Y)]
end
```

Beobachtung? Erklärung??
Visualisieren Sie, z.B. mit hist, boxplot, qqplot...

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

```
> cd ../Daten/
> ls
michelson.dat
> [number speed run expt] = ...
textread('michelson.dat', '%d␣%d␣%d␣%d', 'headerlines' , 1);
> speed
speed =
          850
          740
          900
```

Auswahl der ersten Messreihe

```
> ind = (expt==1);
> s = speed(find(ind));
```

Statistische Kenngrößen

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

```
> mean(s)    909
```

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

```
> std(s)    104.9260
```

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

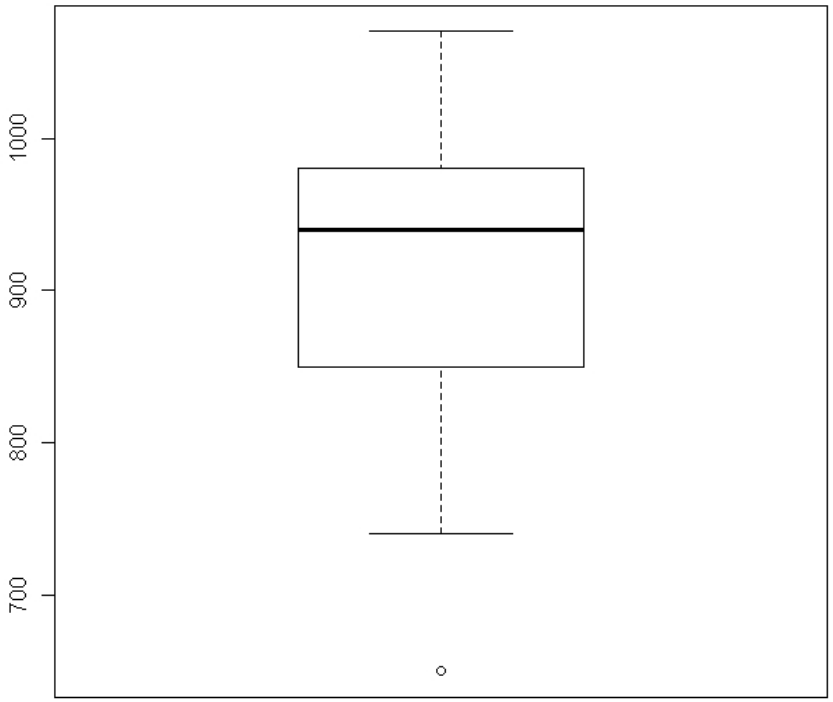
```
> median(s)    940
```

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

```
> mad(s)    88.956
```

Der Box–Whisker–Plot

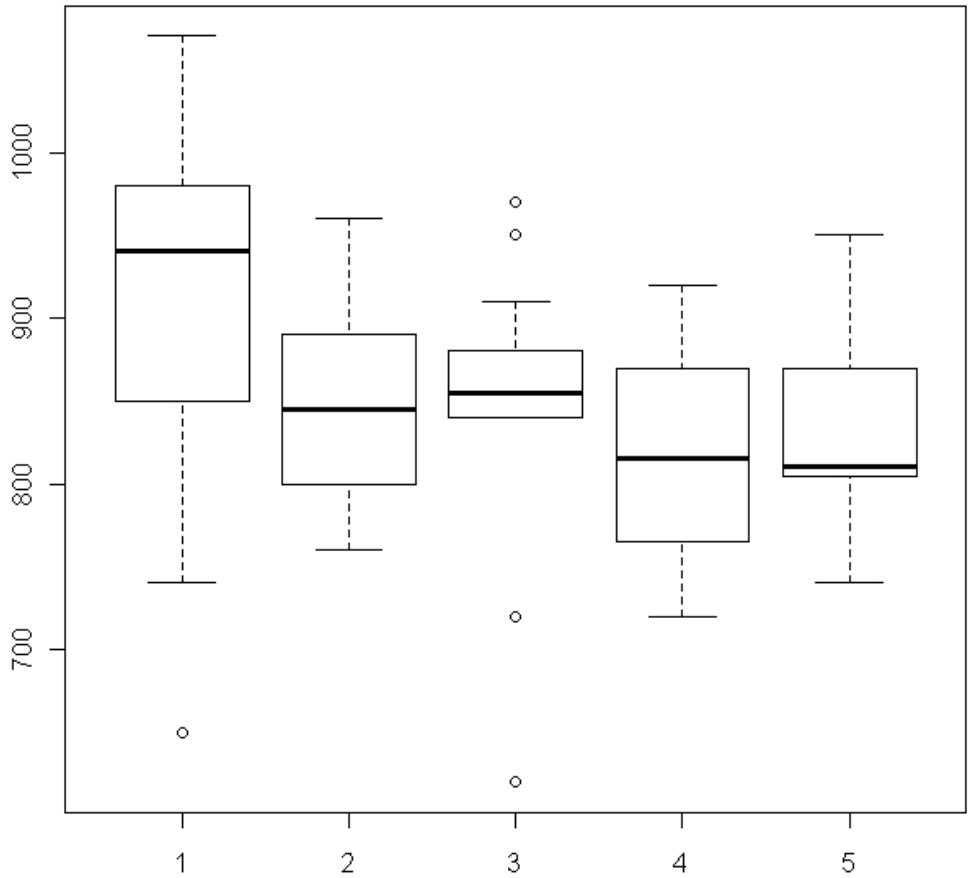
>**boxplot**(s)



Box–Plots der verschiedenen Meßreihen

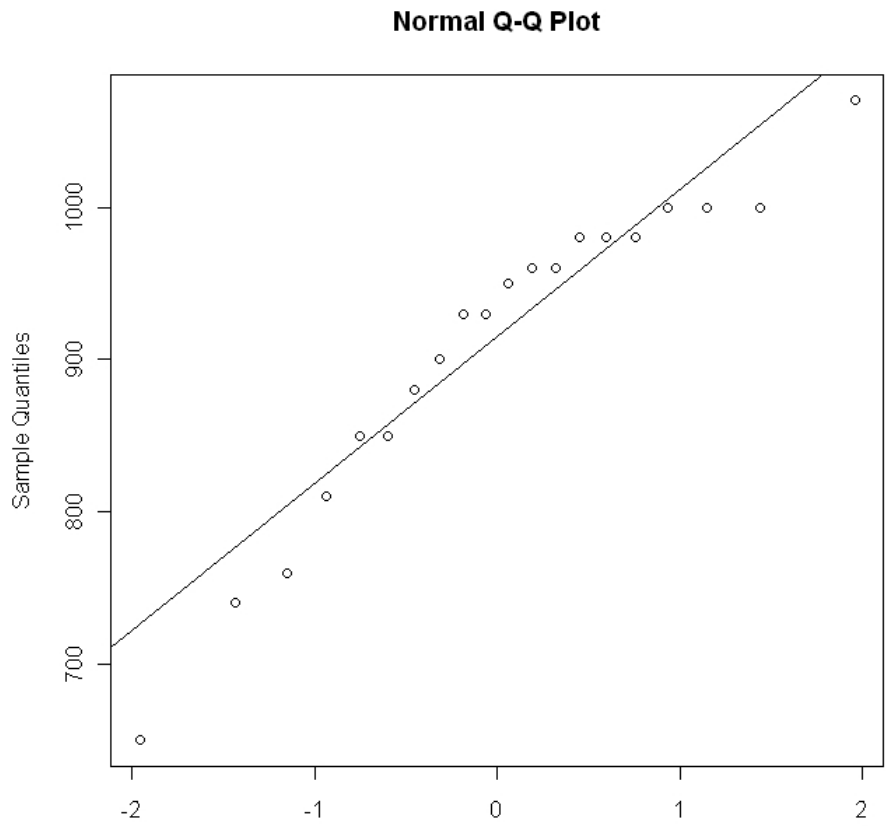
```
> figure('name','Boxplot Daten')
> boxplot(speed, expt)
> xlabel('Exp.Nummer')
> ylabel('Speed')
```

Box–Plots der verschiedenen Meßreihen



Vergleich verschiedener Meßreihen

QQ-Plot von Quantilen 1ter Meßreihe gegen Quantile 2ter Meßreihe



Multivariate Daten: Mietspiegel–Daten

Einlesen der Daten

```
> cd /Verzeichnis/zur/Datei
>  [GKM QMKM QM Zi BJ B L best WW ZH BK BA KUE] =  ...
>    textread('miete03p.csv','%f␣%f␣%d␣%d␣%f␣%d␣%d␣%d␣%d␣%d␣%d␣%d␣%d',
```

Abgeleitete Variablen

Hier: Klassfizierung von Baujahr und Quadratmeterzahl

```
> %Einsortieren in BJ-Klasse 1-6
> BJKL = 1 + (BJ>1918)+(BJ>1948)+(BJ>1965)+(BJ>1977)+(BJ>1983);
> %Einsortieren in QM-Klasse 1-3
> QMKL = 1 + (QM > 50) + (QM > 80) ;
```

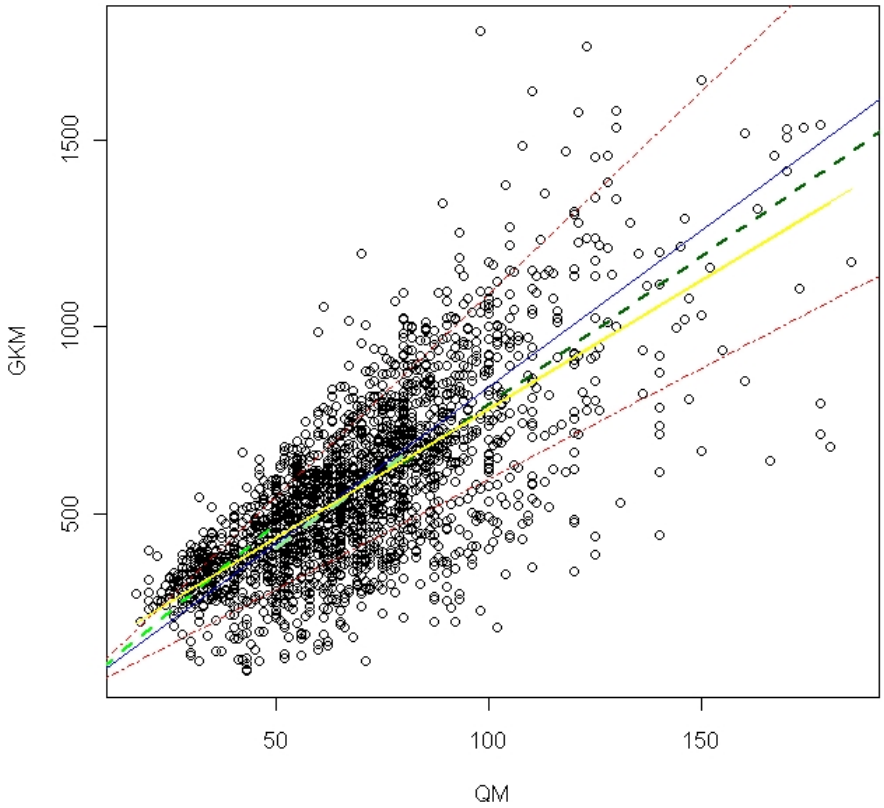
Zusammenhänge zwischen Variablen

Darstellung linearer Zusammengänge : Regressionsgraden

```
 %Regressionsgerade
>     p = polyfit(QM, GKM, 1);
>     x = [0 200];
>     y = polyval(p,x);
>     plot(x,y,'-.r','LineWidth',2)
```

lineare Zusammenhänge

Regressionsgrade, (mittlere QM-Miete±1Std)*QM, QMKL-weise...

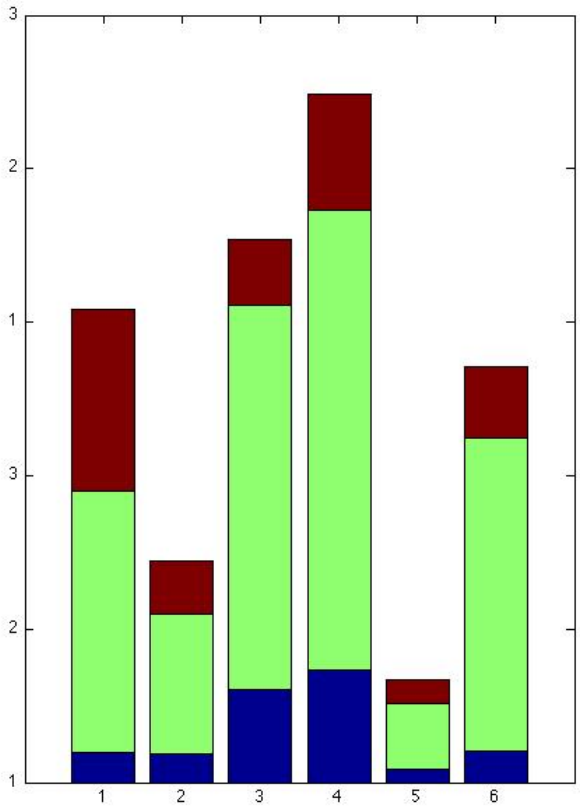


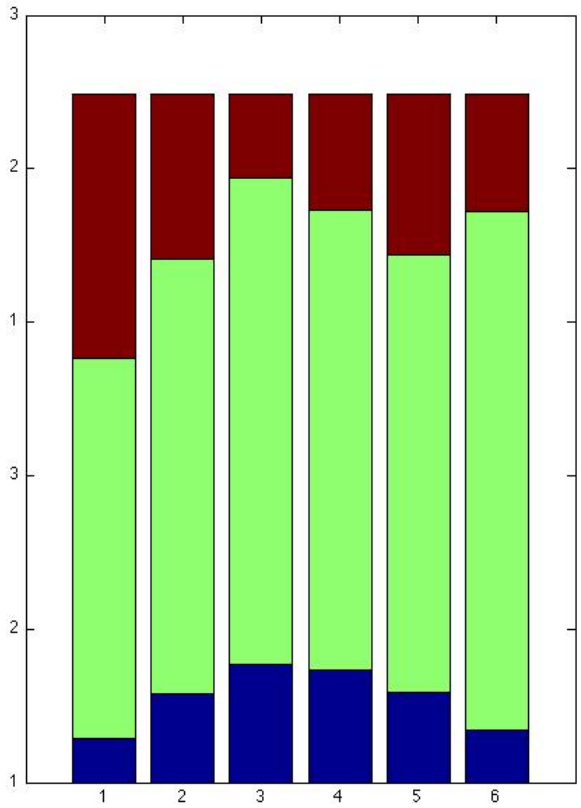
Visualisierung von Kontingenztafeln

Kontingenztafel der QM- u BJ-Klassen bestimmen und mit bar(**data**,'stack') darstellen

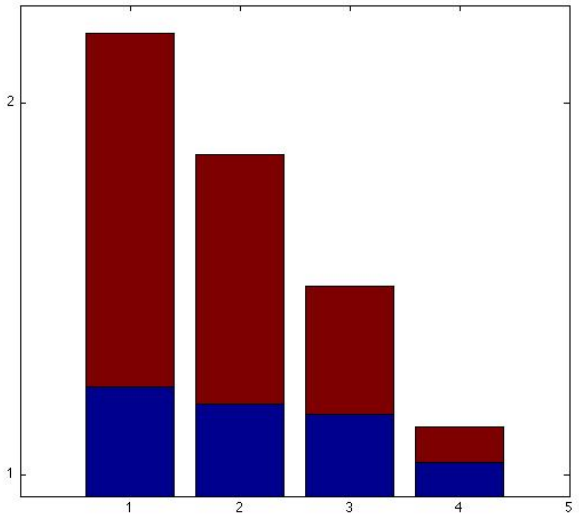
```
> for i=1:6
>         for k=1:3
>             data(i,k) = sum (find(BJKL(find(QMKL==k))==i));
>         end
>     end
>     bar(data,'stack')
```

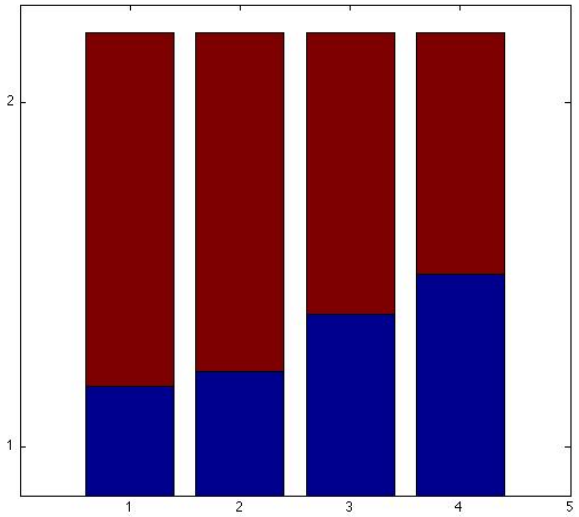
Baujahr $\leftrightarrow$ Wohnungsgr?ße





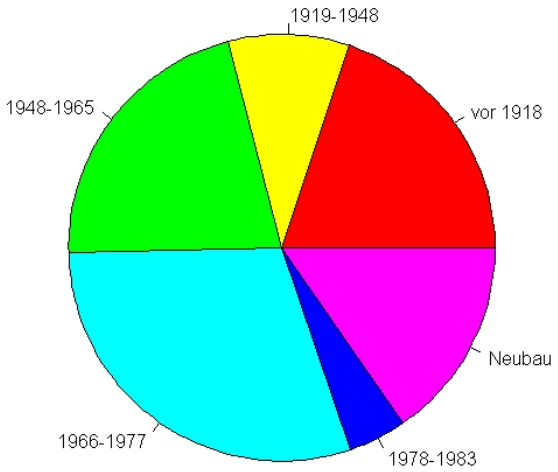
Miete $\leftrightarrow$ Wohnlage

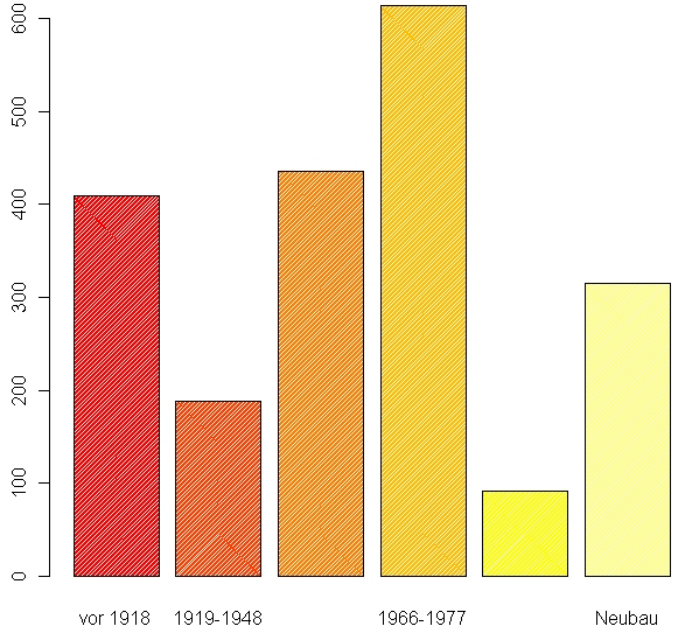




m-Code: Häufigkeitsdarstellungen

```
> 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)
```


1919-1948
vor 1918
1948-1965
Neubau
1966-1977
1978-1983


0
100
200
300
400
500
600
vor 1918
1919-1948
1966-1977
Neubau