Anmerkungen zur Vorlesung Statistik vom 25.10.2016

Inhalt:

- Klausur SS2016: Aufgabe 1 (WDH univariate deskriptive Statistik)
- Zweidimensionale Daten:
  - Kontingenztabellen: (gemeinsame) Häufigkeiten
  - Korrelationsmaße: Bravais-Pearson-Korrelationskoeffizient, Spearman-Korrelationskoeffizient, Kontingenzkoeffizient
  - Darstellung zweidimensionaler Daten, insb. Mosaik-Plots

Aufgabensammlung Statistik:

18.10.2016
Hausaufgabe
Boxplot in A10
A14 - A20

25.10.2016
Hausaufgabe
A23-A31 ohne
Regressions-
aspekte

```
#####
# 25.10.2016: zweidimensionale Daten, (Mosaik-))Plots, Korellation, Regression
#####

setwd("C:/Users/winsanet/Dropbox/HSA/WS2016_17/Daten")
Umfrage<-read.csv2("Umfrage_HSA_2016_10.csv", header=T)
attach(Umfrage)


#gemeinsame Häufigkeiten mit R (nach F. 70)
table(Geschlecht, MatheZufr)
MatheZufr<-ordered(MatheZufr, levels=c("unzufrieden", "geht so", "zufrieden",
"sehr zufrieden") )
table(Geschlecht, MatheZufr) #gemeinsame Häufigkeiten
addmargins(table(Geschlecht, MatheZufr)) #...zzgl. Randhäufigkeiten

#Korrelation mit R (nach F. 81):
x<-c(2,4,3,9,7)
y<-c(4,3,6,7,8)
plot(x,y)
points(mean(x), mean(y), pch=17, col="red")
text(mean(x), mean(y)+0.25, "(x_mw|y_mw))")
#abline(a = mean(y), b = 0, col = "red", lty=2)
abline(h=mean(y), lty=2, col="red") #horizontal Linie
abline(v=mean(x), lty=2, col="red") #vertikale Linie

cor(x,y) #Bravais-Pearson-Korrelationskoeffizient
#cor(x,y,method="pearson")
cor(x,y,method="spearman") #Spearman-Korrelationskoeffizient


#Folie 90:
#tab = table(Farbe, Geschlecht)
#tab
#mosaicplot(t(tab), shade = TRUE, sort=2:1, main="")


#Regressionsanalyse mit R (nach F. 105):
Regression=lm(y~x)
Regression
summary(Regression)
#cor(x,y, method="pearson")^2
cor(x,y)^2

plot(x,y)
#bisschen netter:
plot(x, y, col="blue", pch=16, cex=1.5, xlab="X", ylab="Y", main="Regression")
grid()
points(mean(x), mean(y), pch=17, col="red", cex=1.5)
text(mean(x), mean(y)+0.25, "SP")
abline(h=mean(y), lty=2, col="red") #horizontal Linie
abline(v=mean(x), lty=2, col="red") #vertikale Linie
abline(Regression, col="green", lwd=2)
```

Auswertungsmethoden für zweidimensionale Daten

Zweidimensionale Urliste

Urliste vom Umfang n zu **zwei** Merkmalen X und Y:

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

Kontingenztabelle:

Sinnvoll bei wenigen Ausprägungen bzw. bei klassierten Daten.

	Ausprägungen von Y			
Ausprägungen von X	b_1	b_2	$\ldots$	b_l
a_1	h_{11}	h_{12}	$\ldots$	h_{1l}
a_2	h_{21}	h_{22}	$\ldots$	h_{2l}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
a_k	h_{k1}	h_{k2}	$\ldots$	h_{kl}

gemeinsame Häufigkeiten

Kontingenztabelle

Unterscheide:

- **Gemeinsame Häufigkeiten:**

$$h_{ij} = h(a_i, b_j)$$

- **Randhäufigkeiten:**

$$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \qquad \text{und} \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

- **Bedingte (relative) Häufigkeiten:**

$$f_1(a_i \mid b_j) = \frac{h_{ij}}{h_{\cdot j}} \qquad \text{und} \qquad f_2(b_j \mid a_i) = \frac{h_{ij}}{h_{i\cdot}}$$

Häufigkeiten

Beispiel: 400 unfallbeteiligte Autoinsassen:

	leicht verletzt $(= b_1)$	schwer verletzt $(= b_2)$	tot $(= b_3)$	
angegurtet $(= a_1)$	264 $(= h_{11})$	90 $(= h_{12})$	6 $(= h_{13})$	360 $(= h_{1.})$
nicht angegurtet $(= a_2)$	2 $(= h_{21})$	34 $(= h_{22})$	4 $(= h_{23})$	40 $(= h_{2.})$
	266 $(= h_{.1})$	124 $(= h_{.2})$	10 $(= h_{.3})$	400 $(= n)$

Randhäufigkeiten

$f_2(b_3 \mid a_2) = \frac{4}{40} = 0{,}1$ (10 % der nicht angegurteten starben.)

$f_1(a_2 \mid b_3) = \frac{4}{10} = 0{,}4$ (40 % der Todesopfer waren nicht angegurtet.)

bedingte relative Häufigkeiten

R

Streuungsdiagramm

Streuungsdiagramm sinnvoll bei vielen verschiedenen Ausprägungen (z.B. stetige Merkmale)

➠ Alle (x_i, y_i) sowie $(\bar{x}, \bar{y})$ in Koordinatensystem eintragen.

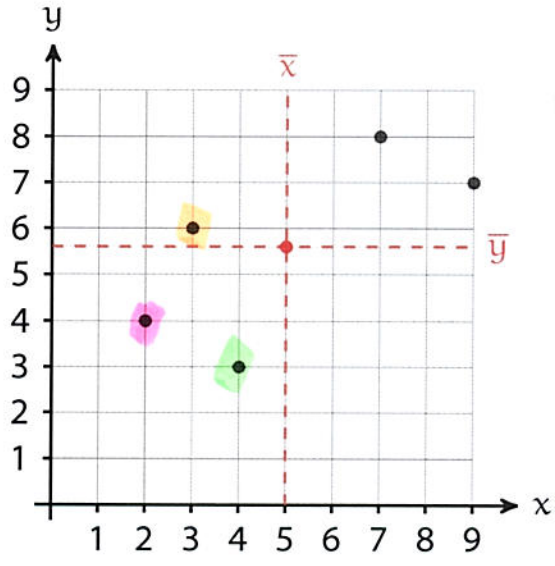
Beispiel:

i	1	2	3	4	5	$\sum$
x_i	2	4	3	9	7	25
y_i	4	3	6	7	8	28

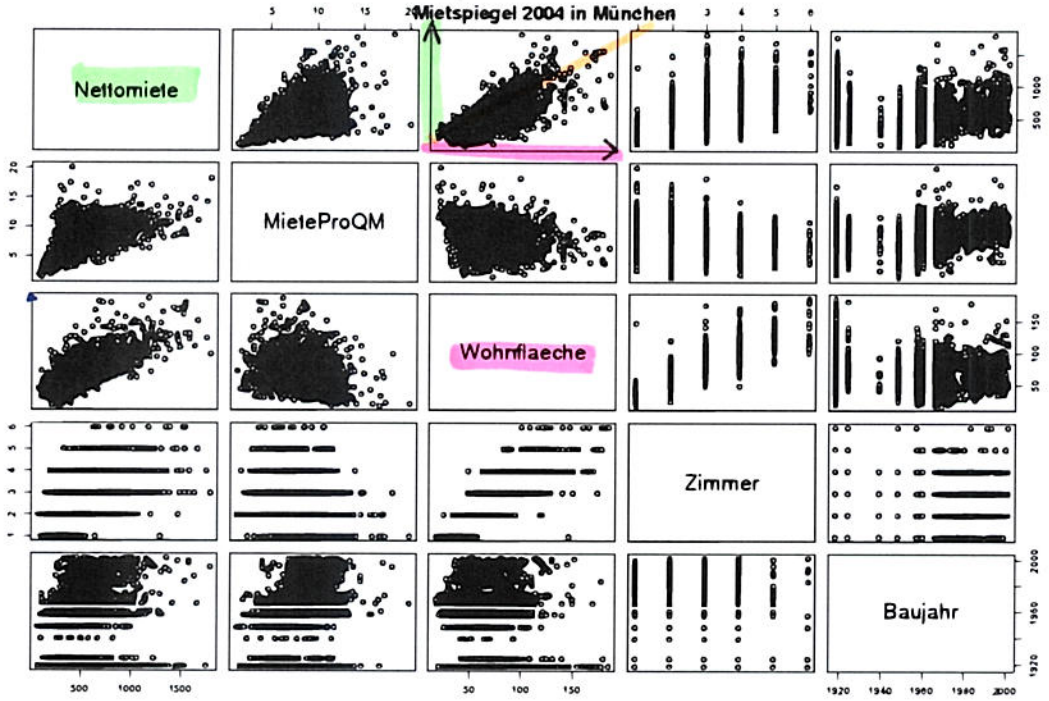
$\Rightarrow \bar{x} = \frac{25}{5} = 5$

$\bar{y} = \frac{28}{5} = 5{,}6$

„Schwerpunkt"



Beispiel Streuungsdiagramm

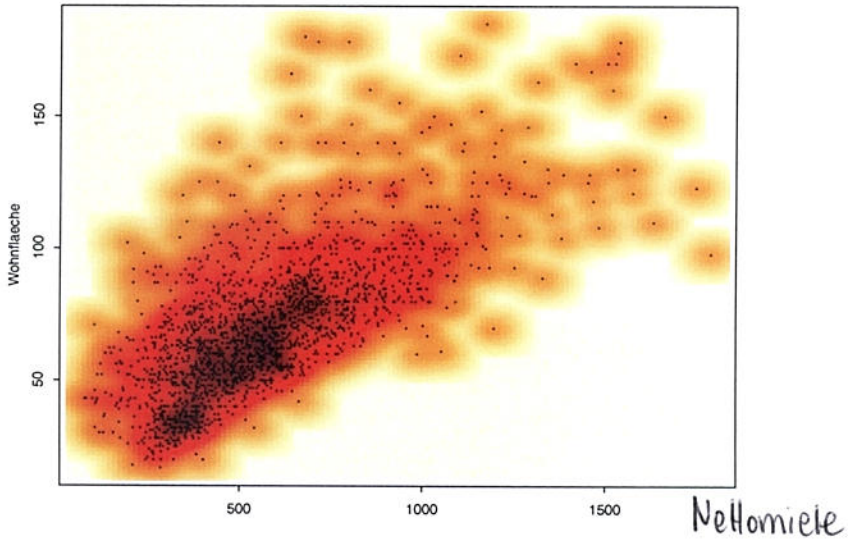


(Datenquelle: Fahrmeir u. a. (2009))

Beispiel Streuungsdiagramm

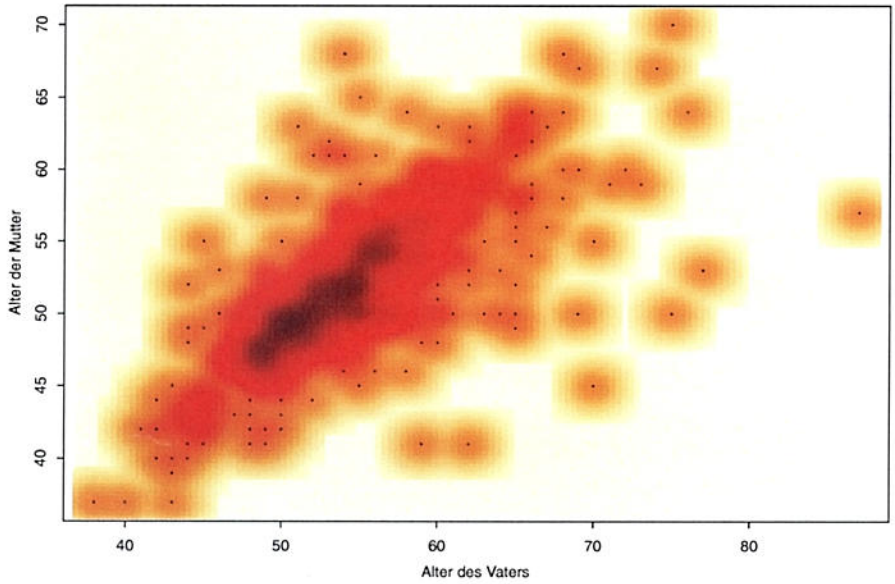

```
if (!require("RColorBrewer")) {
install.packages("RColorBrewer")
library(RColorBrewer)
}
mieten <- read.table('http://goo.gl/jhpJW4', header=TRUE, sep='\t',
                     check.names=TRUE, fill=TRUE, na.strings=c('',''))
x <- cbind(Nettomieten=mieten$nm, Wohnflaeche=mieten$wfl)

library("geneplotter")  ## from BioConductor
smoothScatter(x, nrpoints=Inf,
              colramp=colorRampPalette(brewer.pal(9,"YlOrRd")),
              bandwidth=c(30,3))
```



Beispiel Streuungsdiagramm


```
x = cbind("Alter des Vaters"=AlterV, "Alter der Mutter"=AlterM)
require("geneplotter") ## from BioConductor
smoothScatter(x, colramp=colorRampPalette(brewer.pal(9,"YlOrRd")) )
```




```
require(GGally)
ggpairs(MyData[, c("Alter", "AlterV", "AlterM", "Geschlecht")],
        upper = list(continuous = "density", combo = "box"),
       color='Geschlecht', alpha=0.5)
```

(Kern-) Dichte-Schätzer

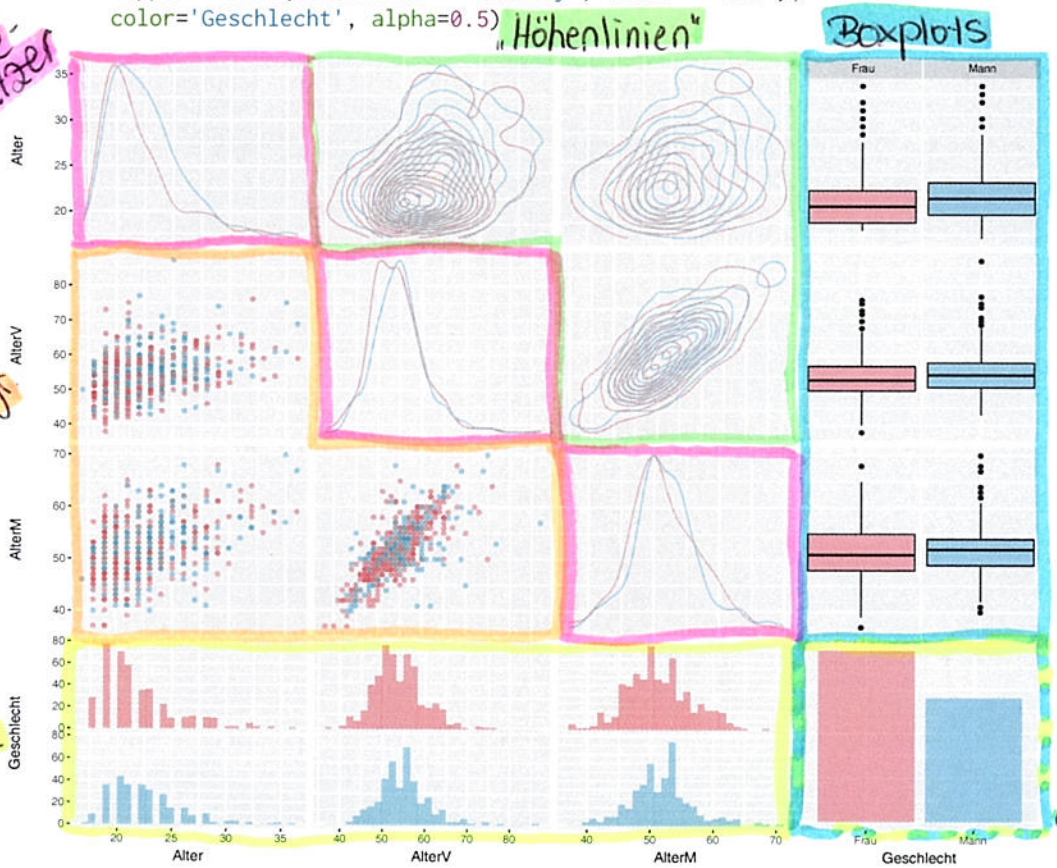
„Höhenlinien"

Boxplots

Streudiagr.

Histogramme

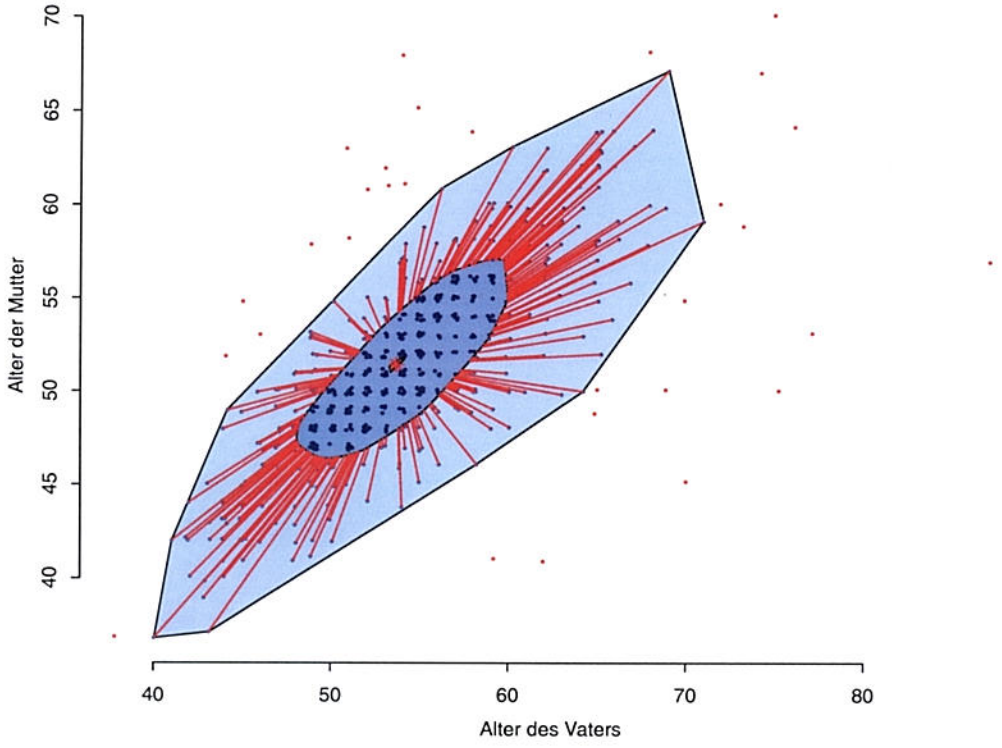
Balken-diagramm der kategoriellen Variable



Bagplot: Boxplot in 2 Dimensionen

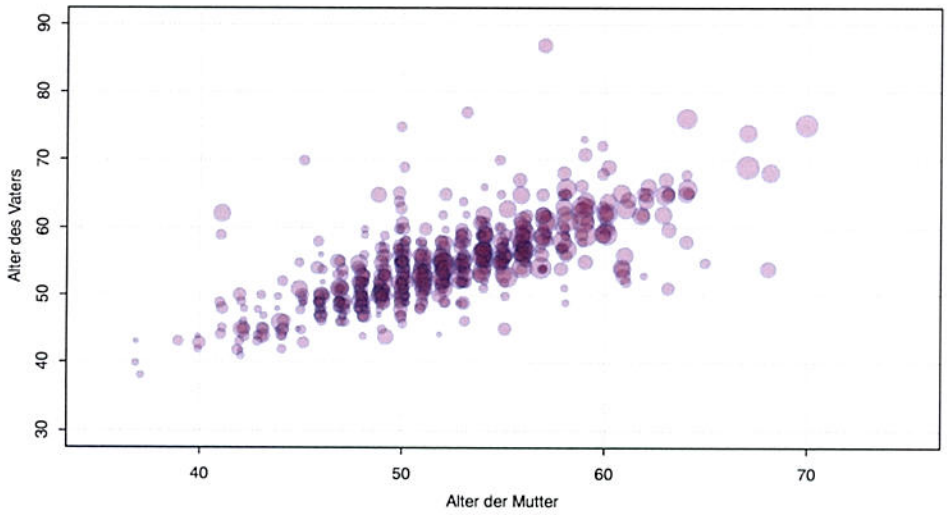
```
require(aplpack)
bagplot(jitter(AlterV), jitter(AlterM), xlab="Alter des Vaters", ylab="Alter der Mutter")
```

```
## [1] "Warning: NA elements have been exchanged by median values!!"
```



Bubbleplot: 3 metrische Variablen

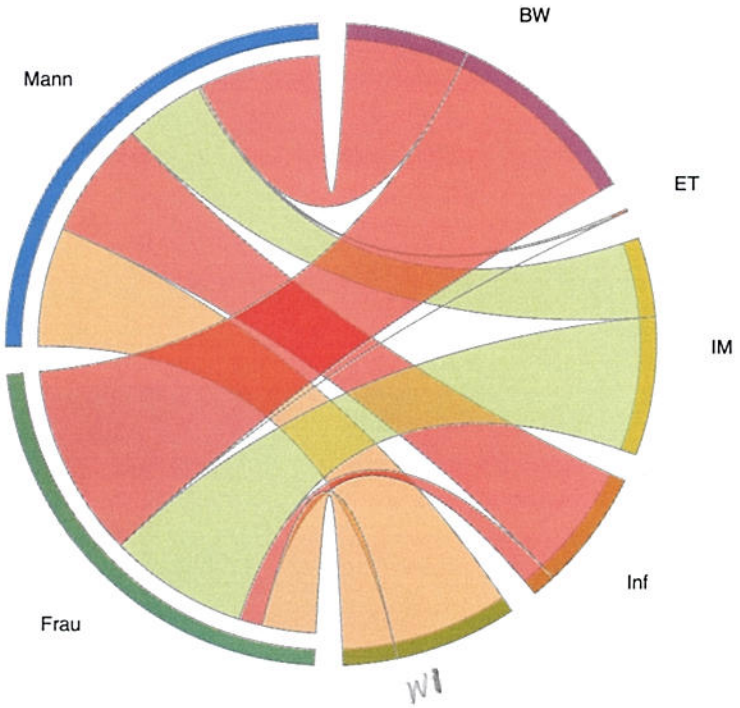
```
require(DescTools)
My.ohne.NA = na.exclude(MyData[,c("AlterM", "AlterV", "Alter")])
with(My.ohne.NA, {
  Alter.skaliert = (Alter-min(Alter))/(max(Alter)-min(Alter))
  PlotBubble(jitter(AlterM), jitter(AlterV), Alter.skaliert,
             col=SetAlpha("deeppink4",0.3),
             border=SetAlpha("darkblue",0.3),
             xlab="Alter der Mutter", ylab="Alter des Vaters",
             panel.first=grid(),
             main="")
})
```



Größe der Blasen: Alter zwischen 0 (Jüngster) und 1 (Ältester)

Circular Plots: Assoziationen

```
require(DescTools)
with(MyData, {
PlotCirc(table(Studiengang, Geschlecht),
acol=c("dodgerblue","seagreen2","limegreen","olivedrab2","goldenrod2","tomato2"),
rcol=SetAlpha(c("red","orange","olivedrab1"), 0.5)
)})
```



Korrelationsrechnung

- Frage: Wie stark ist der Zusammenhang zwischen X und Y?
- Dazu: Korrelationskoeffizienten
- Verschiedene Varianten: Wahl abhängig vom Skalenniveau von X und Y:

misst linearen Zusammenhang

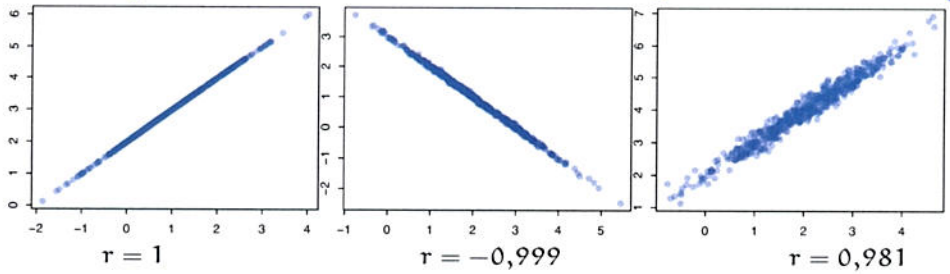
Skalierung von X	Skalierung von Y: kardinal	ordinal	nominal
kardinal	Bravais-Pearson-Korrelationskoeffizient*		
ordinal	misst monotonen Zusammenhang	Rangkorrelationskoeffizient von Spearman	
nominal	misst Abhängigkeit		Kontingenzkoeffizient

Korrelationskoeffizient von Bravais und Pearson

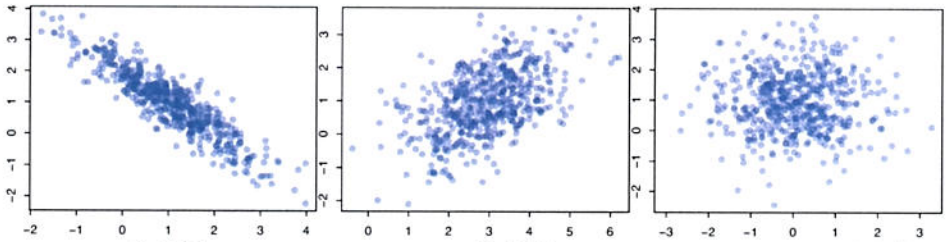
Bravais-Pearson-Korrelationskoeffizient

Voraussetzung: X, Y kardinalskaliert

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}} \in [-1;+1]$$



$r = 1$ $\quad r = -0{,}999$ $\quad r = 0{,}981$



$r = -0{,}903$ $\quad r = 0{,}516$ $\quad r = -0{,}014$

Bravais-Pearson-Korrelationskoeffizient

Im Beispiel:

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	4	4	16	8
2	4	3	16	9	12
3	3	6	9	36	18
4	9	7	81	49	63
5	7	8	49	64	56
$\sum$	25	28	159	174	157

$\Rightarrow \quad \bar{x} = 25/5 = 5$

$\bar{y} = 28/5 = 5{,}6$

$$r = \frac{157 - 5\cdot 5\cdot 5{,}6}{\sqrt{159 - 5\cdot 5^2}\sqrt{174 - 5\cdot 5{,}6^2}} = 0{,}703$$

(deutliche positive Korrelation)

nach F.81

Casio fx 991 DE X : Anleitung S 25ff

↝ Korrelationskoeffizient r

$r \simeq 0{,}7030$

nach F.84

Rangkorrelation nach Spearman

Beispiel: X: Inhalt Geldbeutel
Y: Reich?

// $i = 1,2,3$ reduziertes Beispiel
→ r_{SP} ohne Bind.

i	x_i	y_i	R_x	Rx	R_y	Ry
1	12,-	nicht reich	3	6	3	6
2	54,-	sehr reich	1	3	1	1,5 $= \frac{1+2}{2}$
3	27,-	reich	2	5	2	4 $= \frac{3+4+5}{3}$
4	112,-	reich		2		4 $= \frac{3+4+5}{3}$
5	134,-	sehr reich		1		1,5 $= \frac{1+2}{2}$
6	50,-	reich		4		4 $= \frac{3+4+5}{3}$

$i = 1,2,3$ (reduziertes Bsp): r_{SP} ohne Bindungen

// $r_{SP} = 1 - \frac{6 \cdot ((3-3)^2 + (1-1)^2 + (2-2)^2)}{2 \cdot 3 \cdot 4} = \ldots = 1$ //

$i = 1, \ldots, 6$: r_{SP} mit Bindungen

→ Berechnung mit TR
Bravais-Pearson - Korrelationskoeffizient r für Rangwerte
$r \simeq 0.7715$

Guess The Correlation

guessthecorrelation.com

Go for the Highscore!

Rangkorrelationskoeffizient von Spearman

- Voraussetzungen: X, Y (mindestens) ordinalskaliert, Ränge eindeutig (keine Doppelbelegung von Rängen)
- Vorgehensweise:
  1. Rangnummern R_i (X) bzw. R'_i (Y) mit $R_i^{(\prime)} = 1$ bei größtem Wert usw. "Beste" ≙ Rang 1 ... "Schlechteste" ≙ letzter Rang
  2. Berechne

r_{SP} ohne Bindungen:

$$r_{SP} = 1 - \frac{6 \sum_{i=1}^{n} (R_i - R'_i)^2}{(n-1)\,n\,(n+1)} \in [-1;+1]$$

- Hinweise:
  - $r_{SP} = +1$ wird erreicht bei $R_i = R'_i \quad \forall i = 1, \ldots, n$
  - $r_{SP} = -1$ wird erreicht bei $R_i = n + 1 - R'_i \quad \forall i = 1, \ldots, n$
- Falls Ränge nicht eindeutig: Bindungen, dann Berechnung von r_{SP} über Ränge und Formel des Korr.-Koeff. von Bravais-Pearson

r_{SP} mit Bindungen:

$$r_{SP} = \frac{\sum_{i=1}^{n} R_i \cdot R'_i - n \cdot \bar{R}^2}{\sqrt{\sum_{i=1}^{n} R_i^2 - n \cdot \bar{R}^2} \cdot \sqrt{\sum R_i'^2 - n \cdot \bar{R}^2}}$$

mit $\bar{R} = \frac{n+1}{2}$

Rangkorrelationskoeffizient von Spearman

Im Beispiel:

x_i	R_i	y_i	R'_i
2	5	4	4
4	3	3	5
3	4	6	3
9	1	7	2
7	2	8	1

$$r_{SP} = 1 - \frac{6 \cdot [(5-4)^2 + (3-5)^2 + (4-3)^2 + (1-2)^2 + (2-1)^2]}{(5-1) \cdot 5 \cdot (5+1)} = 0{,}6$$

Kontingenzkoeffizient

- Gegeben: Kontingenztabelle mit k Zeilen und l Spalten (vgl. hier)
- Vorgehensweise:
  1. Ergänze Randhäufigkeiten

$$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \qquad \text{und} \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

  2. Berechne theoretische Häufigkeiten

$$\tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}$$

  3. Berechne

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

χ^2 hängt von n ab! $(h_{ij} \mapsto 2 \cdot h_{ij} \Rightarrow \chi^2 \mapsto 2 \cdot \chi^2)$

nach F.85

Beispiel (vgl. F.70)

Gesucht: χ^2-Wert für folgende Häufigkeitstabelle

		y: Verletzung leicht	mittel	schwer	$h_{i\cdot}$
X: Gurt	ja	264 (239,4)	90 (111,6)	6 (9)	360
	nein	2 (26,6)	34 (12,4)	4 (1)	40
	$h_{\cdot j}$	266	124	10	400

(...) theoretische Häufigkeiten

$$\chi^2 = \frac{(264-239{,}4)^2}{239{,}4} + \frac{(90-111{,}6)^2}{111{,}6} + \ldots + \frac{(4-1)^2}{1} \simeq 4{,}1799$$

Aber: χ^2 ist vom STP-Umfang n abhängig!

nach F.86

$$K = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{4{,}1799}{4{,}1799+400}} \simeq 0{,}1017$$

Normierung zur Vergleichbarkeit (K ist durch K_{max} beschränkt)

$$K_* = \frac{K}{K_{max}} \quad \text{mit} \quad K_{max} = \sqrt{\frac{M-1}{M}} = \sqrt{\frac{1}{2}}$$

weil $M = \min\{k, l\}$, $k = 2$, $l = 3$

$$\rightsquigarrow K_* \simeq 0{,}1438$$

Kontingenzkoeffizient

④ Kontingenzkoeffizient:

$$K = \sqrt{\frac{\chi^2}{n+\chi^2}} \qquad \in [0; K_{max}]$$

wobei

$$K_{max} = \sqrt{\frac{M-1}{M}} \qquad \text{mit} \qquad M = \min\{k, l\}$$

⑤ Normierter Kontingenzkoeffizient:

$$K_* = \frac{K}{K_{max}} \qquad \in [0; 1]$$

$K_* = +1 \iff$
bei Kenntnis von x_i kann y_i erschlossen werden u.u.

Kontingenzkoeffizient

Beispiel

X : Staatsangehörigkeit (d,a)
Y : Geschlecht (m,w)

h_{ij}	m	w	$h_{i.}$
d	30	30	60
a	10	30	40
$h_{.j}$	40	60	100

$\Rightarrow$

$\tilde{h}_{ij}$	m	w
d	24	36
a	16	24

wobei $\tilde{h}_{11} = \frac{60 \cdot 40}{100} = 24$ usw.

$$\chi^2 = \frac{(30-24)^2}{24} + \frac{(30-36)^2}{36} + \frac{(10-16)^2}{16} + \frac{(30-24)^2}{24} = 6{,}25$$

$$K = \sqrt{\frac{6{,}25}{100+6{,}25}} = 0{,}2425; \qquad M = \min\{2,2\} = 2; \qquad K_{max} = \sqrt{\frac{2-1}{2}} = 0{,}7071$$

$$K_* = \frac{0{,}2425}{0{,}7071} = 0{,}3430$$

zu F. 88

Mosaikplots

Schritt 1: horizontaler Split

Diesen Plot nennt man Spine Plot. Er ist ähnlich einem Balkendiagramm, jedoch spiegelt nicht die Länge sondern die Breite die prozentualen Anteile der Merkmalsausprägungen pro Level wider
--> Also: Betrachte Breite der y-Achse

Schritt 2: zusätzlicher vertikaler Split
Jeder der so entstandenen „Balken" wird nun zusätzlich gemäß der anderen Merkmalsvariablen proportional zu den Häufigkeiten der Ausprägungen gesplittet.

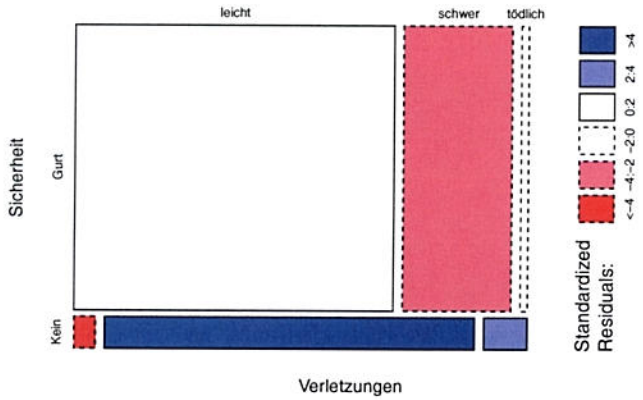
Interpretation des Mosaikplots: Beispiel Autounfälle
1. Erster Split nach der Variablen "Sicherheit/Gurt?" - diese Variable dominiert den Plot
2. Verteilung von "Verletzungen" gegeben einer Ausprägung von "Sicherheit/Gurt?" ebenfalls zu sehen
3. Verteilung von "Verletzung" ist nicht direkt ersichtlich

Die Farbgebung repräsentiert das Residual-Level der einzelnen Zellbereiche/Merkmalskombinationen (Legende rechts vom Plot).
Blau: Es gibt **mehr** Beobachtungen dieser Merkmalskombination als unter Unabhängigkeit der Merkmale zu erwarten wären (überrepräsentiert)
Rot: Es gibt **weniger** Beobachtungen dieser Merkmalskombination als unter Unabhängigkeit der Merkmale zu erwarten wären (unterrepräsentiert)
Blaue und rote Zellbereiche haben wesentlichen Einfluss auf die Höhe des Chi-Quadrat-Wertes und damit im weiteren dann auch auf die Aussage bzgl. der Merkmalsabhängigkeit.

Anmerkung: Reihenfolge des Splits wird über das Attribut sort= der Funktion mosaicplot( ) festgelegt (vgl. F. 90 vs. F. 92)

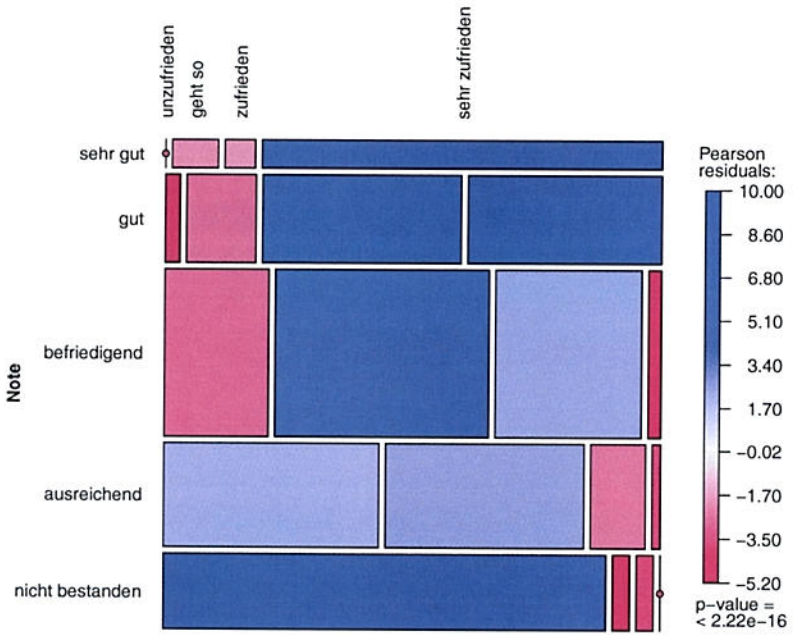
Beispiel Autounfälle

	Verletzung			
	leicht	schwer	tödlich	
angegurtet	264	90	6	360
nicht angegurtet	2	34	4	40
	266	124	10	400



Mosaikplot Autounfälle


```
Data.complete = na.omit(MyData[,c("MatheZufr", "NoteMathe")])
Noten.complete =
  ordered(cut(Data.complete$NoteMathe, breaks=c(0,1.5,2.5,3.5,4.1,5.0)),
          labels=c("sehr gut", "gut", "befriedigend", "ausreichend", "nicht bestanden"))
tab = table("Note"=Noten.complete, "Zufrieden mit Leistung"=Data.complete$MatheZufr)
require(vcd)
mosaic(tab, shade = TRUE, gp_args = list(interpolate = function(x) pmin(x/4, 1)), labeling_args =
   list(rot_labels = c(90,0,0,0), just_labels = c("left", "left", "right", "right"),
        offset_varnames = c(left = 5, top=5.5), offset_labels = c(right = 3)),
 margins = c(right = 1, bottom = 3, left=6, top=5))
```



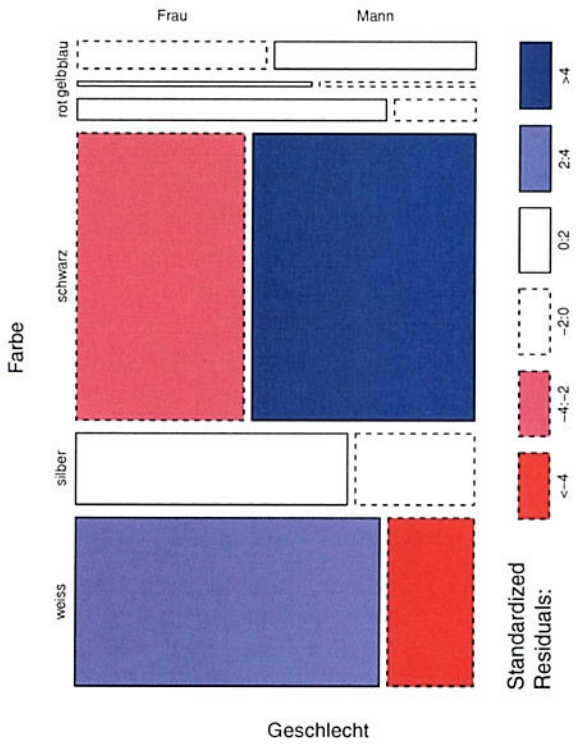
„Note in Matheklausur" gegen „Zufrieden mit Leistung"

Mosaicplot Geschlecht, Wunschfarbe für Smartphone

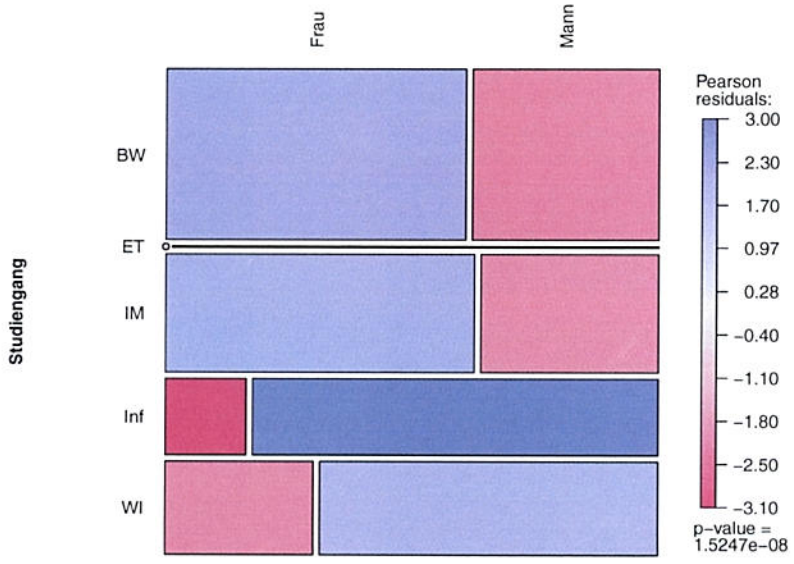

```
tab = table(Farbe, Geschlecht)
tab
```

```
##          Geschlecht
## Farbe      Frau Mann
##   blau       15   16
##   gelb        3    2
##   rot        19    5
##   schwarz   143  190
##   silber     57   25
##   weiss     152   43
```

```
mosaicplot(t(tab), shade = TRUE,
           sort=2:1, main="")
```




```
require(vcd)
Data.complete = na.omit(MyData[,c("Geschlecht", "Studiengang")])
with(Data.complete, {
  tab = table("Studiengang"=Studiengang, "Geschlecht"=Geschlecht)
mosaic(tab, shade = TRUE, gp_args = list(interpolate = function(x) pmin(x/4, 1)), labeling_args =
   list(rot_labels = c(90,0,0,0), just_labels = c("left", "left", "right", "right"),
        offset_varnames = c(left = 5, top=5.5), offset_labels = c(right = 3)),
 margins = c(right = 1, bottom = 3, left=6, top=5))
})
```



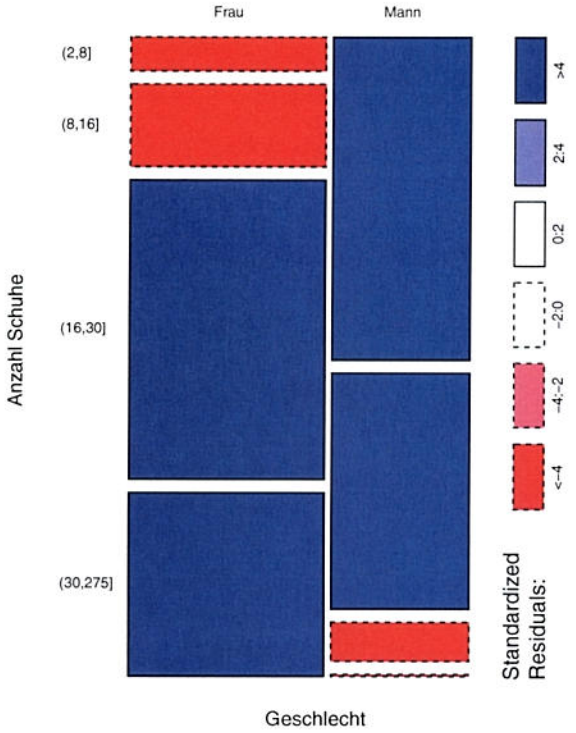
Studiengang vs. Geschlecht

Mosaicplot Geschlecht, Anzahl Schuhe


```
tab = table(
  "Anzahl Schuhe" =
  cut(AnzSchuhe,
      breaks =
        quantile(
          AnzSchuhe,
          probs = (0:4)/4
        )
  ),
  Geschlecht)
tab
```

```
##              Geschlecht
## Anzahl Schuhe Frau Mann
##       (2,8]     22   148
##       (8,16]    53   108
##       (16,30]  195    18
##       (30,275] 119     1
```

```
mosaicplot(t(tab), shade = TRUE,
           main="", las=1)
```



Preisindizes

- Preismesszahl: Misst Preisveränderung eines einzelnen Gutes:

$$\frac{\text{Preis zum Zeitpunkt } j}{\text{Preis zum Zeitpunkt } i}$$

dabei: j: Berichtsperiode, i: Basisperiode

- Preisindex: Misst Preisveränderung mehrerer Güter (Aggregation von Preismesszahlen durch Gewichtung)
- Notation:

$p_0(i)$: Preis des i-ten Gutes in Basisperiode 0
$p_t(i)$: Preis des i-ten Gutes in Berichtsperiode t
$q_0(i)$: Menge des i-ten Gutes in Basisperiode 0
$q_t(i)$: Menge des i-ten Gutes in Berichtsperiode t

⚠ Voress ohne F 93-97 (Preisindizes)