Statistik

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2016

Veranstaltungen zur Statistik für BW/IM Sommersemester 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Vorlesung Statistik	Etschberger	Mi, 14.00-17.00	B2.14	16.03.2016
Vorlesung Statistik PLUS	Etschberger/Jansen	- Blocktermin -	?	?
Übung Statistik	Etschberger	Mi, 17.00-18.30	A1.10	30.03.2016
Übung Statistik	Jansen	Di, 11.30-13.00	W1.06	22.03.2016
Übung Statistik	Jansen	Di, 14.00-15.30	W2.14	22.03.2016
Übung Statistik	Jansen	Mi, 11.30-13.00	W2.11	30.03.2016
Übung Statistik	Jansen	Do, 14.00-15.30	W2.14	31.03.2016
Übung Statistik	Schneller	Do, 14.00-15.30	W3.03	31.03.2016
Übung Statistik	Schneller	Do, 15.30-17.00	W3.03	31.03.2016
Übung Statistik	Wins	Di, 14.00-15.30	J3.19	22.03.2016
Übung Statistik	Wins	Di, 15.30-17.00	J3.19	22.03.2016
Offener Statistikraum	Etschberger/Tutoren	Mo, 14.00-17.45	B3.05	04.04.
Veranstaltungen für Teilnehmer der WiMa-Klausur im Juli 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Tutorium Mathematik	Burkart	Do 13.30-15.00	W1.06	07.04.2016
Tutorium Mathematik	Burkart	Do 15.00-16.15	W1.06	07.04.2016
Offener Matheraum	Jansen/Tutoren	?	?	?

1	HSA Statistik SS 2016 Sessionlist		
2			
3	Datum	Statistik für IM/BW	Nr.
4	Mittwoch, 16. März 2016	Einführung, R Installation, Rstudio Einführung, Skalen	1
5	Mittwoch, 23. März 2016	univ. deskr. Stat., Quantile, Plots	2
6	Mittwoch, 30. März 2016	Streuung, Konzentrationsmaße	3
7	Mittwoch, 6. April 2016	Kontingenztabellen, Mosaikplots, Korrelation	4
8	Mittwoch, 13. April 2016	Preisindizes, lineare Regression	5
9	Mittwoch, 20. April 2016	Kombinatorik, Wahrscheinlichkeit	6
10	Mittwoch, 27. April 2016	Wahrscheinlichkeit, diskrete Zufallsvariablen	7
11	Mittwoch, 4. Mai 2016	Pyramid	
12	Mittwoch, 11. Mai 2016	Binomial-, Hypergeom.-, Poisson-Verteilung	8
13	Mittwoch, 18. Mai 2016	Stetige ZV, Gleichverteilung	9
14	Mittwoch, 25. Mai 2016	Normalverteilung, Verteilungsparameter	10
15	Mittwoch, 1. Juni 2016	Schätzfunktionen und Punktschätzer	11
16	Mittwoch, 8. Juni 2016	Konfidenzintervalle	12
17	Mittwoch, 15. Juni 2016	Tests	13
18	Mittwoch, 22. Juni 2016	Puffer, WH, Fragen zur Probekl.	14
19	Mittwoch, 29. Juni 2016	AW Prüfungswoche	

Prof. Dr. Stefan Etschberger
Hochschule Augsburg

Weitere Konzentrationsmaße

i: 1 2 3 4 5
p_i: p_1 p_2 p_3 p_4 p_5

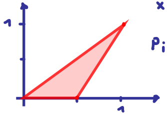
- **Konzentrationskoeffizient:**

$CR_2 = \sum_{i=5-2+1}^{5} p_i = \sum_{i=4}^{5} p_i = p_4 + p_5$

$$CR_g = \text{Anteil, der auf die } g \text{ größten entfällt} = \sum_{i=n-g+1}^{n} p_i = 1 - v_{n-g}$$

- **Herfindahl-Index:**

$$H = \sum_{i=1}^{n} p_i^2 \qquad (\in [\tfrac{1}{n}; 1])$$



$x = 0 \text{ Mio}, 1 \text{ Mio}$

$p_i = 0, 1$

$H = \sum p_i^2 = 0^2 + 1^2 = 1$

$E = \prod p_i^{p_i} = 0^0 \cdot 1^1 = 1$

Es gilt: $H = \frac{1}{n}(V^2 + 1)$ bzw. $V = \sqrt{n \cdot H - 1}$

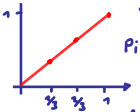
- **Exponentialindex:**

$0^0 \cdot 0.01^{0.01} \cdot 0.03^{0.03} \cdot 0.96^{0.96}$

$$E = \prod_{i=1}^{n} p_i^{p_i} \qquad (\in [\tfrac{1}{n}; 1]) \qquad \text{wobei} \qquad 0^0 = 1$$

Urliste: $x = 2 \text{ Mio}, 2 \text{ Mio}, 2 \text{ Mio}$

$p_i = \frac{1}{3}, \frac{1}{3}, \frac{1}{3}$



$H = \sum p_i^2 = (\frac{1}{3})^2 + (\frac{1}{3})^2 + (\frac{1}{3})^2 = 3 \cdot \frac{1}{9} = \frac{1}{3}$

- Im Beispiel mit $x = (1, 2, 2, 15)$:

$$CR_2 = \frac{17}{20} = 0{,}85$$

$$H = \left(\frac{1}{20}\right)^2 + \cdots + \left(\frac{15}{20}\right)^2 = 0{,}59$$

$$E = \left(\frac{1}{20}\right)^{\frac{1}{20}} \cdots \left(\frac{15}{20}\right)^{\frac{15}{20}} = 0{,}44$$

Beispiel: $x = \underbrace{1 \text{Mio}, 1 \text{Mio}, \ldots, 1 \text{Mio}}_{n\text{-mal}}$

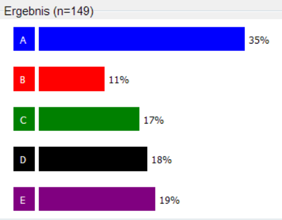
$p_i = \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}$

$\Rightarrow H = (\frac{1}{n})^2 + \cdots + (\frac{1}{n})^2 = n \cdot (\frac{1}{n})^2 = \frac{1}{n}$

```
> # Aufgabe zu Lorenzkurve und Gini-Koeffizient
> require(ineq) # Paket ist nötig
> set.seed(4)
> x = c(max(na.exclude(Umfrage$AusgKomm)), sample(Umfrage$AusgKomm, 4))
> x
[1]  4668 250  420  250  300
> # Aufgabe:
> # a) Zeichne Lorenzkurve (5 Minuten)
> # b) Berechne Gini-Koeffizient (3 Minuten)
>
> # Loesung
> plot(Lc(x))
> grid()
> sort(cumsum(x)/sum(x))
[1] 0.04245924 0.11379076 0.15625000 0.20720109 1.00000000
>
> # Gini von Hand
> pi = sort(x)/sum(x) # Sortieren nicht vergessen
> n = length(x)
> x.Gini = ((2*sum(pi*(1:n))) - 6 )/n
> x.Gini
[1] 0.6118207
>
> # Oder über Paket ineq
> Gini(x)
[1] 0.6118207
```

Umfrage: In der letzten Aufgabe hatte ich

A) Alles richtig
B) Alles bis auf die Zeichnung richtig
C) Einen Fehler in den Zahlen
D) Mehr als einen Fehler in den Zahlen
E) Ich wusste nicht, was zu tun ist



Mitschrift Statistik 6.4.2016

SE

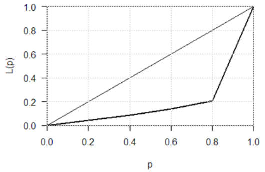
6 April 2016

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
# ---------------------------------------------------------------------
# Statistik SS 2016: R Mitschrift aus Vorlesung
# ---------------------------------------------------------------------

# Daten aus dem Netz
Umfrage = read.csv("http://goo.gl/KCGF16", sep=";", dec=",")

# Aufgabe zu Lorenzkurve und Gini-Koeffizient
require(ineq) # Paket ist nötig

## Loading required package: ineq

## Warning: package 'ineq' was built under R version 3.1.3

set.seed(4)
x = c(max(na.exclude(Umfrage$AusgKomm)), sample(Umfrage$AusgKomm, 4))
x

## [1] 4668  250  420  250  300

# Aufgabe:
# a) Zeichne Lorenzkurve (5 Minuten)
```




```
cumsum(sort(x))/sum(x)

## [1] 0.04245924 0.08491848 0.13586957 0.20720109 1.00000000

# Gini von Hand
pi = sort(x)/sum(x) # Sortieren nicht vergessen
n = length(x)
x.Gini = ((2*sum(pi*(1:n))) - 6 )/n
pi

## [1] 0.04245924 0.04245924 0.05095109 0.07133152 0.79279891

x.Gini

## [1] 0.6118207

# Oder über Paket ineq
Gini(x)
```

Zweidimensionale Urliste

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

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

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}

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:

Rand-häufigkeit

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

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

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

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

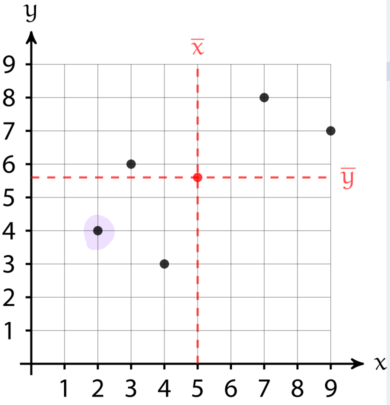
➠ Alle (x_i, y_i) sowie $(\bar{x}, \bar{y})$ in Koordinatensystem eintragen.

Beispiel:

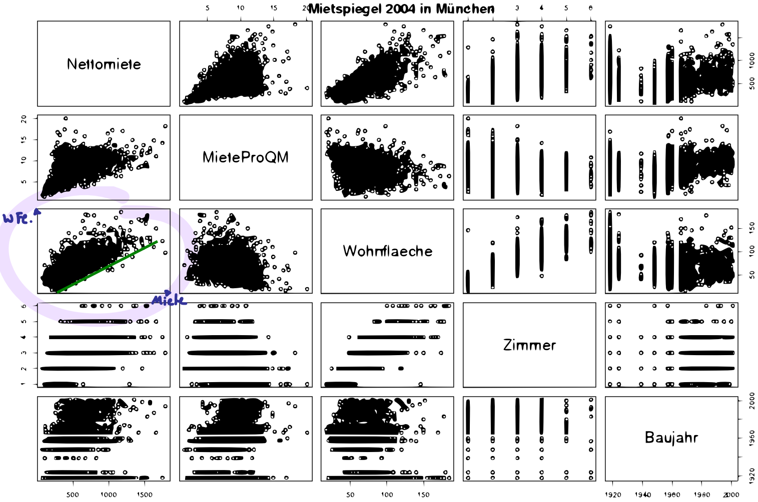
i	1	2	3	4	5	$\sum$
x_i	2	4	3	9	7	25
y_i	4	3	6	7	8	28

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

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



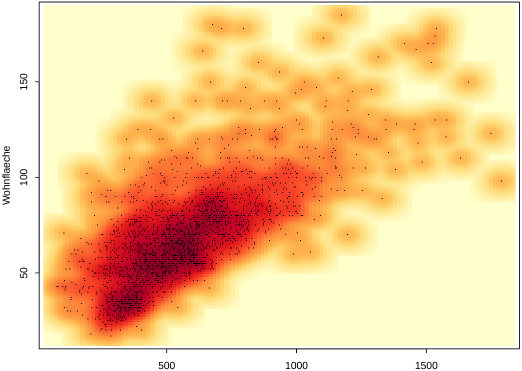
Beispiel Streuungsdiagramm



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

Beispiel Streuungsdiagramm

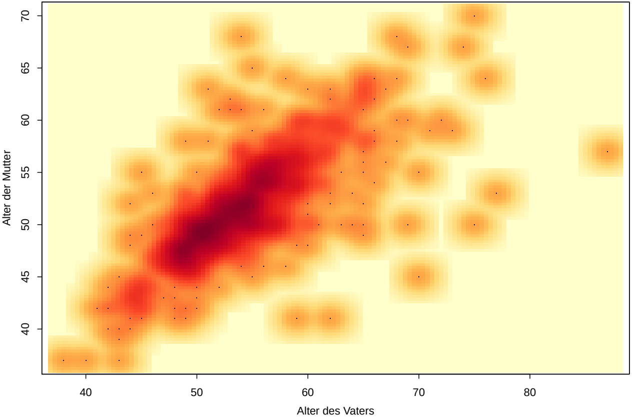
```r
if (!require("RColorBrewer")) {
install.packages("RColorBrewer")
library(RColorBrewer)
}
mieten <- read.table('http://goo.gl/jhpJW4', header=TRUE, sep='\t',
                     check.names=TRUE, fill=TRUE, na.strings=c('',''))
x <- cbind(Nettomieten=mieten$nm, Wohnflaeche=mieten$wfl)

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



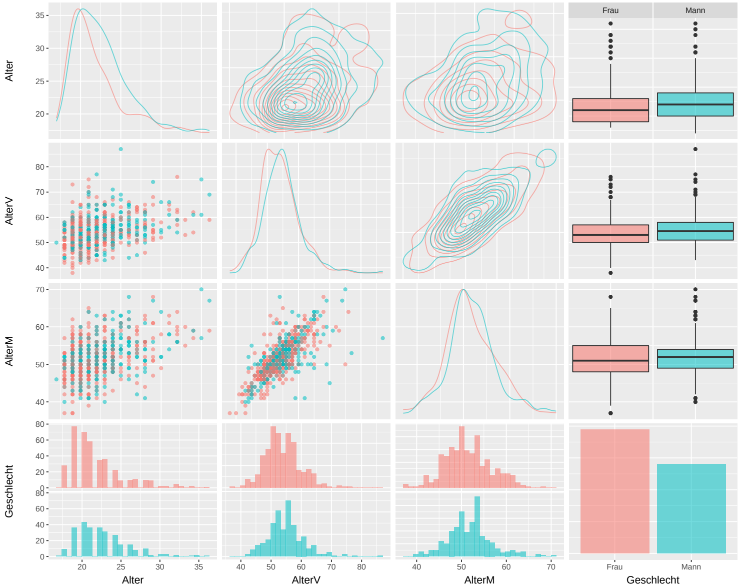
Nettomiete

Beispiel Streuungsdiagramm

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

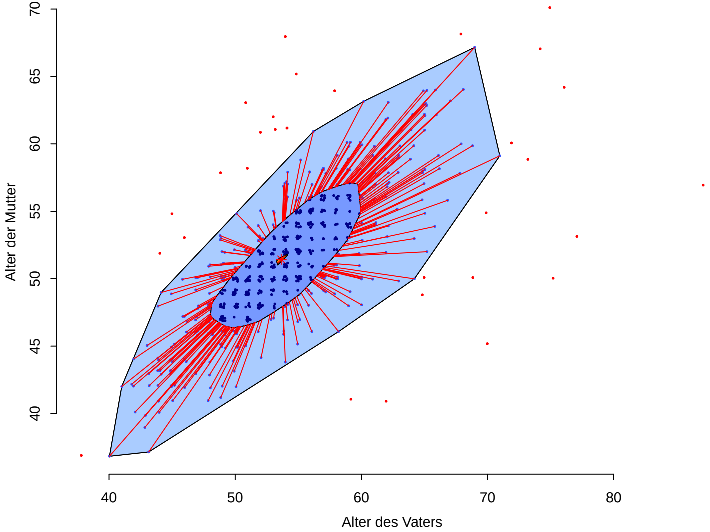


```r
require(GGally)
ggpairs(MyData[, c("Alter", "AlterV", "AlterM", "Geschlecht")],
        upper = list(continuous = "density", combo = "box"),
       color='Geschlecht', alpha=0.5)
```



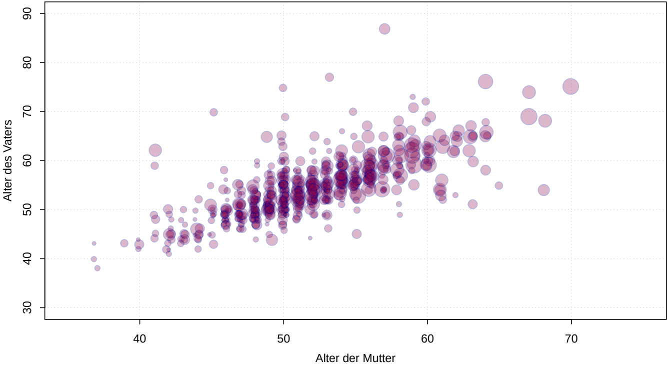
Bagplot: Boxplot in 2 Dimensionen

```r
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

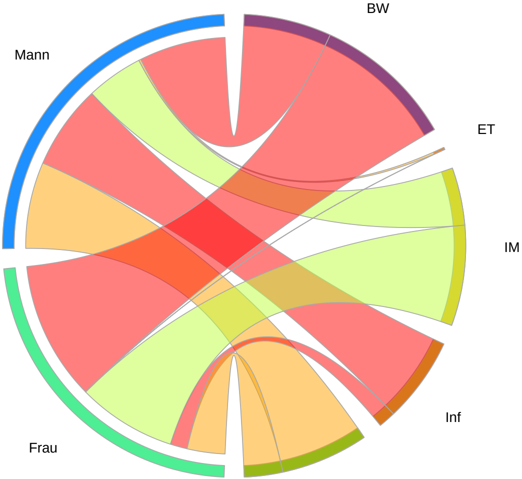
```
require(DescTools)
My.ohne.NA = na.exclude(MyData[,c("AlterM", "AlterV", "Alter")])
with(My.ohne.NA, {
  Alter.skaliert = (Alter-min(Alter))/(max(Alter)-min(Alter))
  PlotBubble(jitter(AlterM), jitter(AlterV), Alter.skaliert,
             col=SetAlpha("deeppink4",0.3),
             border=SetAlpha("darkblue",0.3),
             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


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



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

	Skalierung von Y		
Skalierung von X	kardinal	ordinal	nominal
kardinal	Bravais-Pearson-Korrelations-koeffizient		
ordinal		Rangkorrelations-koeffizient von Spearman	
nominal			Kontingenz-koeffizient

Korrelationskoeffizient von Bravais und Pearson

Bravais-Pearson-Korrelationskoeffizient

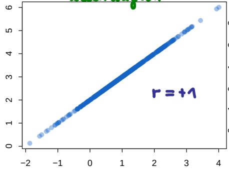
Voraussetzung: X, Y kardinalskaliert

Kovarianz von X und Y

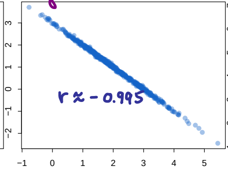
$$r = \frac{\frac{1}{n}\cdot\sum\limits_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2\cdot\frac{1}{n}\sum\limits_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum\limits_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum\limits_{i=1}^{n}x_i^2-n\bar{x}^2}\sqrt{\sum\limits_{i=1}^{n}y_i^2-n\bar{y}^2}}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

Standardabweichung von X

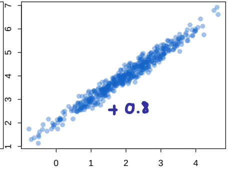
Std. abw. von Y



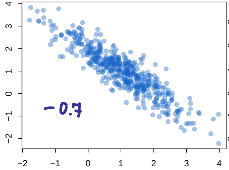
$r = 1$



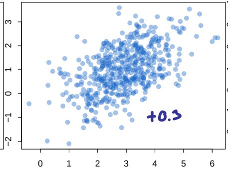
$r = -0{,}999$



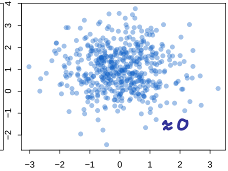
$r = 0{,}981$



$r = -0{,}903$



$r = 0{,}516$

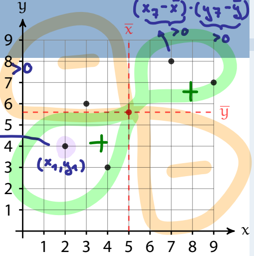


$r = -0{,}014$

Bravais-Pearson-Korrelationskoeffizient

$$\underbrace{(x_1-\bar{x})}_{<0}\cdot\underbrace{(y_1-\bar{y})}_{<0}>0$$

$$\underbrace{(x_7-\bar{x})}_{>0}\cdot\underbrace{(y_7-\bar{y})}_{>0}>0$$



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

$$\Rightarrow \quad \bar{x} = 25/5 = 5 \qquad \bar{y} = 28/5 = 5{,}6$$

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

(deutliche positive Korrelation)

Bravais-Pearson - Korrelationskoeffizient mit dem TR

x_i	y_i
2	4
4	3
3	6
9	7
7	8

- Mode → STAT → A+Bx
- Daten eingeben ...
- $\boxed{\text{AC}}$; Shift → STAT → REG

(0.703) r

Rangkorr. nach Spearman

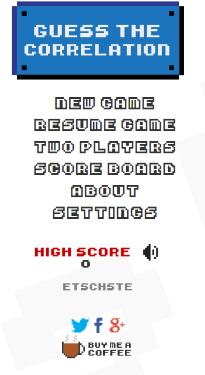
Beispiel : x : Inhalt Geldbeutel
Y : Wie reich ?

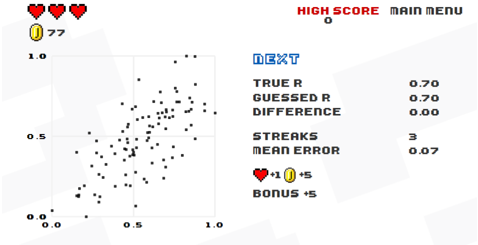
i	x_i	Y_i	Rx		Ry
1	5,23	nicht	6	6	6
2	25,83	schon	4	3	4
3	134,17	sehr	1	1	1,5
4	27,80	sehr	3	2	1,5
5	12,17	schon	5	4	4
6	34,50	schon	2	5	4

Rangkorrelationskoeffizient nach Spearman entspricht dem Korr. Koeff. von Bravais-Pearson auf Basis der Ränge beider Merkmale

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.
  2. Berechne

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

- 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

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
  $$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \quad \text{und} \quad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$
  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! $(h_{ij} \mapsto 2 \cdot h_{ij} \Rightarrow \chi^2 \mapsto 2 \cdot \chi^2)$

	Verletzung			
	leicht	schwer	tödlich	
angegurtet	264 239,4	90 111,6	6 9	360
nicht angegurtet	2 26,6	34 12,4	4 1	40
	266	124	10	400

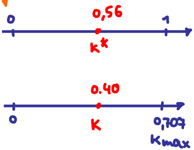
$$\frac{360 \cdot 124}{400}$$

$$\chi^2 = \frac{(264-239{,}4)^2}{239{,}4} + \frac{(90-111{,}6)^2}{111{,}6} + \frac{(6-9)^2}{9}$$

$$+ \frac{(2-26{,}6)^2}{26{,}6} + \frac{(34-12{,}4)^2}{12{,}4} + \frac{(4-1)^2}{1}$$

chi-Quadrat

„keine Korr.“ „perfekte Korrelation“



④ Kontingenzkoeffizient:

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

Unfall bsp: $\chi^2 \approx 77,\ldots$, $n = 400$

$\Rightarrow K \approx 0{,}40$

3 Spalten in Kont. tabelle

wobei

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

2 Zeilen

$\Rightarrow M = \min\{2, 3\} = 2$

$\Rightarrow K_{max} = \sqrt{\frac{2-1}{2}} = \sqrt{0.5} \approx 0{,}707$

⑤ Normierter Kontingenzkoeffizient:

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

$K_* \approx \frac{0{,}40}{0{,}707} \approx 0.56$

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

Beispiel

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

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

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

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

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

6.4.2016