Quantitative Methoden

Elemente der Deskriptiven Statistik II

Bivariate Analysen

Prof. Dr. rer. nat.Thomas Wiebringhaus

Elemente der Deskriptiven Statistik I

Univariate Analysen (eine Variable)

- Merkmale + Skalenniveaus
- Absolute + relative Häufigkeiten
- Säulendiagramm , Empirische Verteilungsfunktion und Histogramm
- Lagemaße (MW, Median und Modus), Boxplots und Quantile
- Streumaße: Varianz und Standardabweichung (sd)

Elemente der Deskriptiven Statistik II

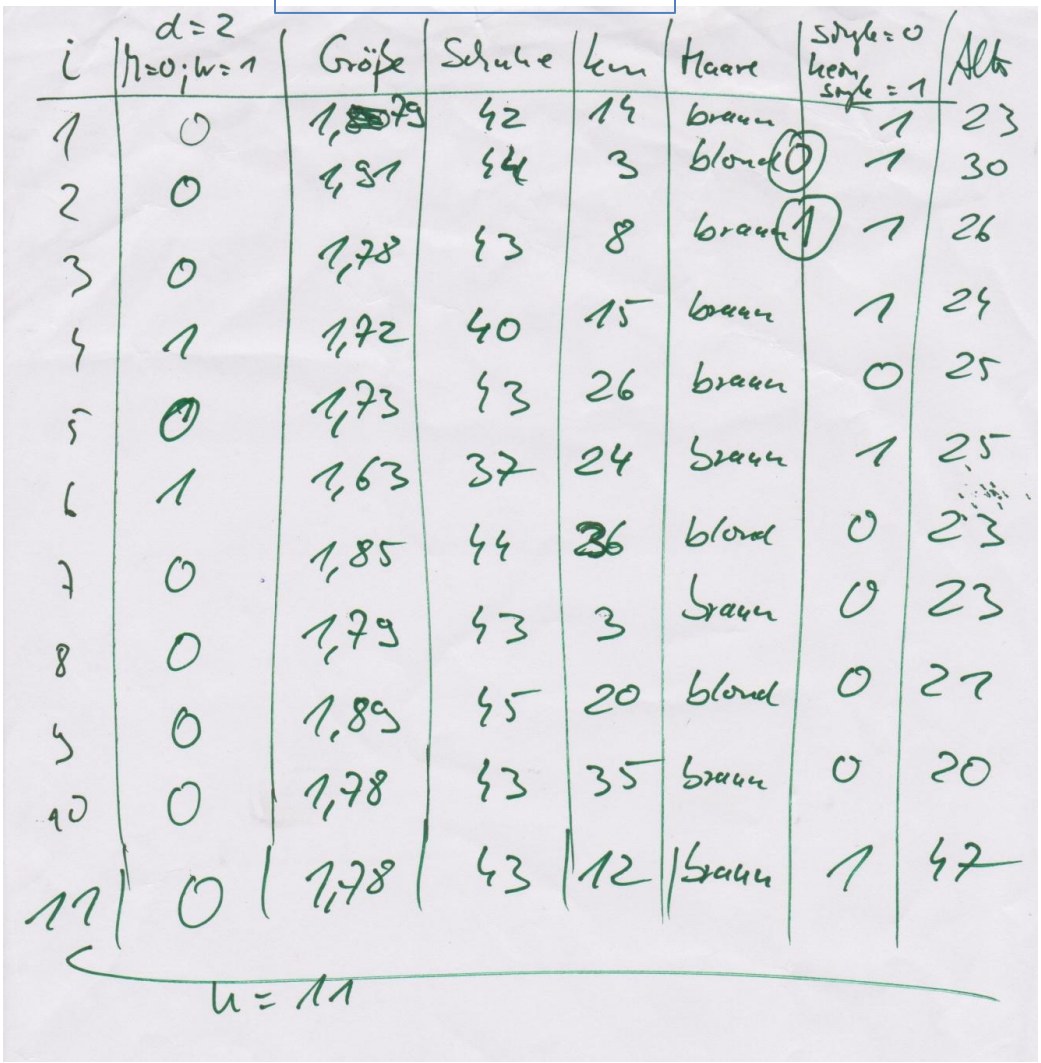
Bivariate Analysen (zwei Variablen)

- Kovarianz
- Korrelationskoeffizienten (Bravais-Pearson)
- Rangkorrelationskoeffizient (Spearman)
- Kontingenzkoeffizient (nominal) und Chi^2
- Lineare Regression

Beispiel: Datenerhebung (Anatomie einer Urliste)

Spalten (columns) = Variablen

index [ ] | nominal | m e t r i s c h | nominal | nominal | metrisch



i	d=2 h=0; w=1	Größe	Schuhe	km	Haare	Style=0 kein Style=1	Alter
1	0	1,[illegible]79	42	14	braun	1	23
2	0	1,81	44	3	blond (0)	1	30
3	0	1,78	43	8	braun (1)	1	26
4	1	1,72	40	15	braun	1	24
5	0	1,73	43	26	braun	0	25
6	1	1,63	37	24	braun	1	25
7	0	1,85	44	36	blond	0	23
8	0	1,79	43	3	braun	0	23
9	0	1,89	45	20	blond	0	27
10	0	1,78	43	35	braun	0	20
11	0	1,78	43	12	braun	1	47

n = 11

← Zeilen (rows) = Beobachtungen

3 wiss. Kriterien:

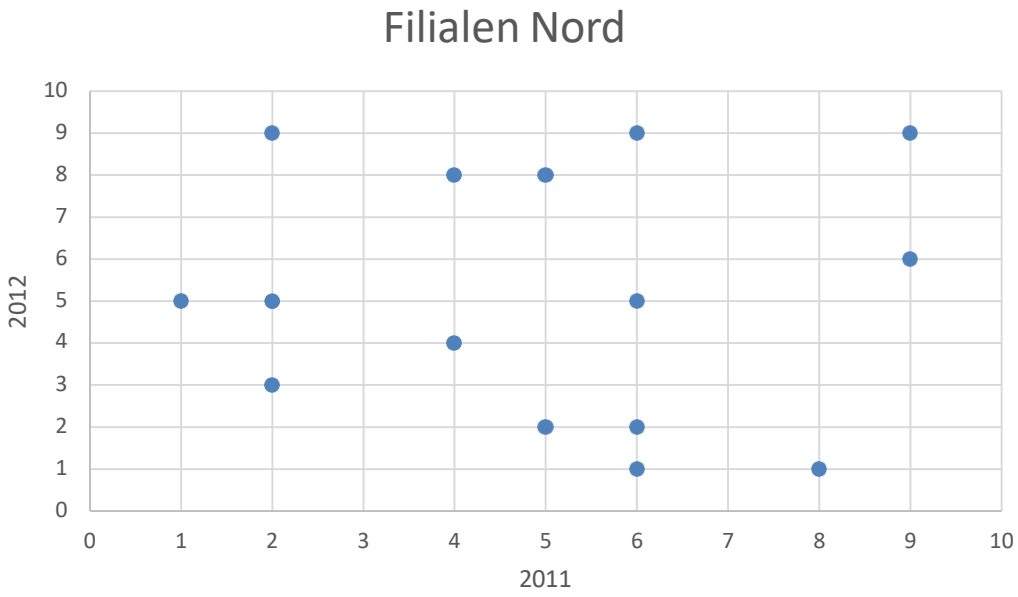
Objektiv: unabhängig vom Beobachter
Reliable: Wiederholung zuverlässig
Valide: das Gemessene ist gültig

Kovarianz

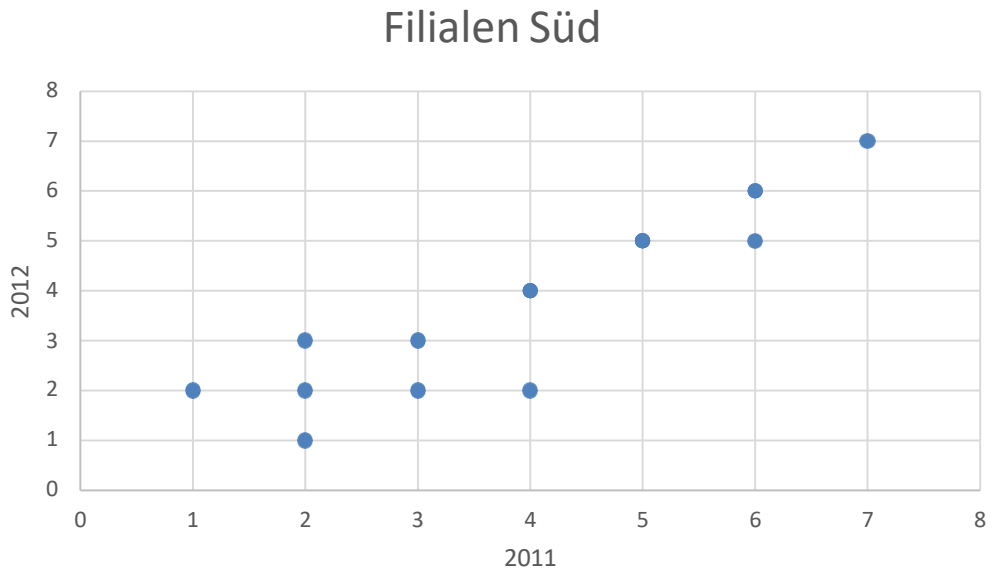
4.1 Kovarianz: ein Beispiel

Ihnen liegen Umsatzzahlen der Jahre 2011 und 2012 Ihrer Filialen Nord und Süd vor

No.	*2011*	*2012*
1	2	5
2	5	8
3	4	4
4	8	1
5	9	6
6	2	9
7	4	8
8	9	9
9	2	3
10	5	2
11	2	5
12	6	9
13	6	2
14	6	5
15	1	5
16	6	1



Streudiagramm
(scatter plot)

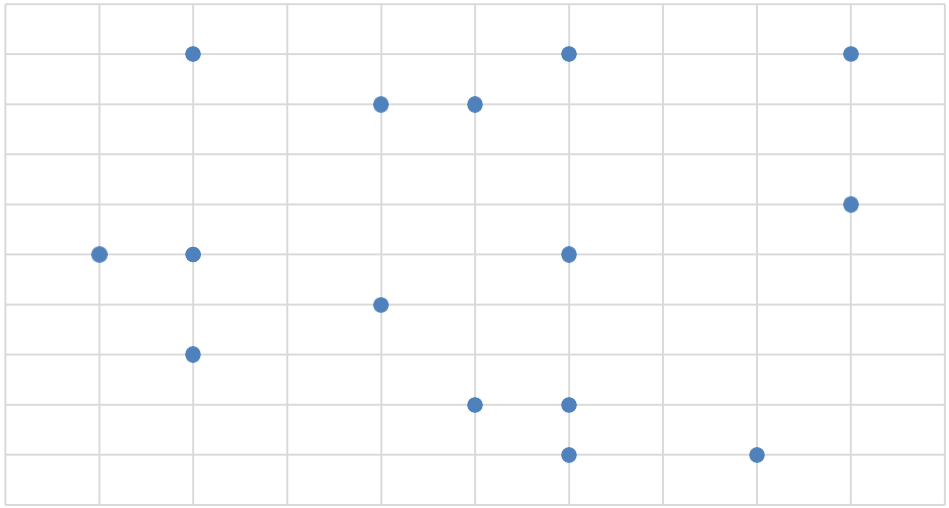


→ Lin. Regression

No.	*2011*	*2012*
1	1	2
2	2	1
3	3	3
4	2	2
5	3	2
6	2	3
7	4	2
8	4	4
9	5	5
10	4	4
11	5	5
12	6	6
13	5	5
14	6	6
15	6	5
16	7	7

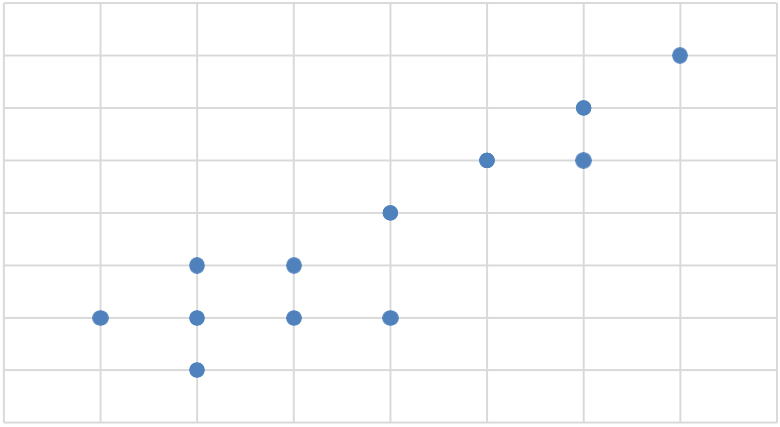
4.1 Kovarianz: Beispiele

Bivariat (zweidimensional): 2 Merkmalsausprägungen

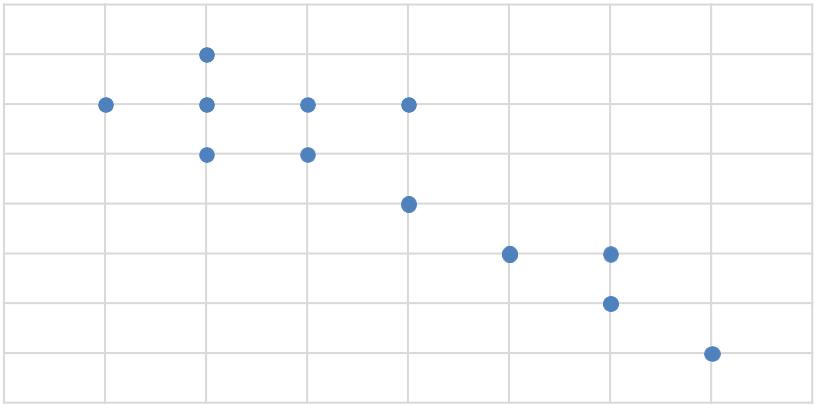


→ Punktwolke; Kein Zusammenhang
→ Kovarianz um 0

Anforderung an die Kovarianz:
pos/ neg trend
Streuung



→ Kovarianz positiv

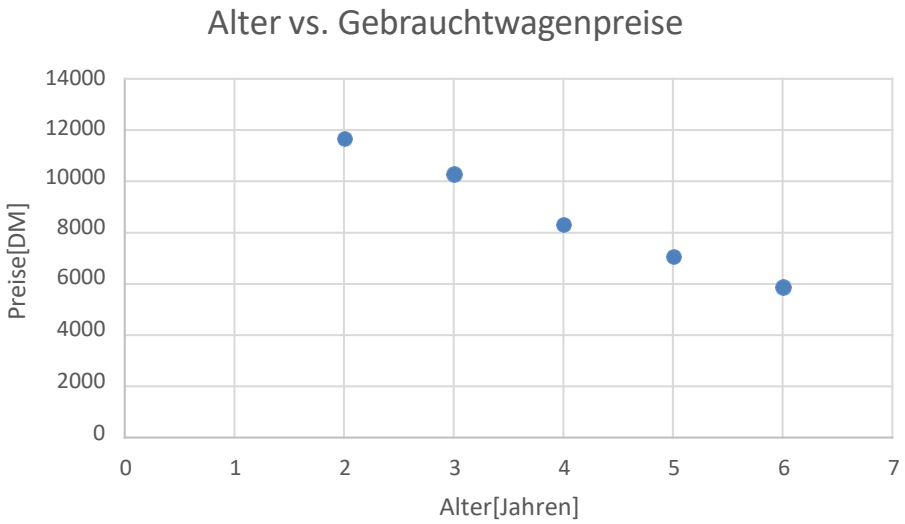


→ Kovarianz negativ

4.1 Kovarianz: weiteres Beispiel

Im Jahre 1984 galten für gebrauchte PKW eines speziellen Typs folgende Händlerverkaufspreise:

i	Alter x_i	Preis y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
1	2	11700	-2	3030	-6060
2	3	10300	-1	1630	-1630
3	4	8350	0	-320	0
4	5	7100	1	-1570	-1570
5	6	5900	2	-2770	-5540
Summe $\sum$	**20**	**43350**	**0**	**0**	**-14800**
MW	**$\bar{x}$= 4**	**$\bar{y}$= 8670**			



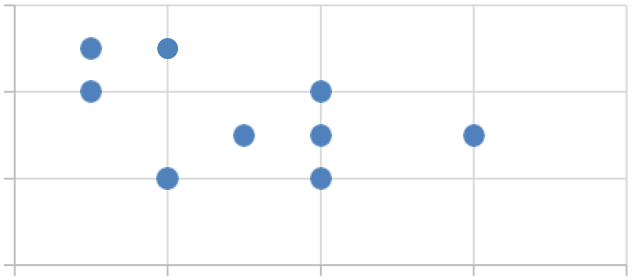
$$s_{xy} = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{5}(-14800) = -2960$$

$$\text{Varianz } s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{Kovarianz } s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Quelle: Schlittgen S. 94

4.1 Kovarianz: noch ein Beispiel



Das Streudiagramm läßt einen negativen Zusammenhang erwarten

No.	x_i	y_i	x_i-x	y_i-y	$(x_i$ -$x)(y_i$ -$y)$
1	1	5	-1,9	1,4	-2,66
2	2	5	-0,9	1,4	-1,26
3	3	3	0,1	-0,6	-0,06
4	2	2	-0,9	-1,6	1,44
5	1	4	-1,9	0,4	-0,76
6	6	3	3,1	-0,6	-1,86
7	4	2	1,1	-1,6	-1,76
8	4	4	1,1	0,4	0,44
9	2	5	-0,9	1,4	-1,26
10	4	3	1,1	-0,6	-0,66
Summe	29	36	0	0	-8,4
MW	2,9	3,6			**-0,84**

Anforderung an die Kovarianz:
pos/ neg trend !
Streuung ? (nicht leicht interpretierbar)

Korrelation nach Pearson

4.2 Korrelationskoeffizient (Pearson)

$$Korrelationskoeffizient = \frac{Kovarianz}{std_x \cdot std_y}$$

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

- Der Korrelationskoeffizient r_{xy} normiert die Kovarianz mit den Standardabweichungen

- Dadurch bewegt sich der r_{xy} zwischen
  -1 (maximal negativer Zusammenhang) und
  +1 (maximal positiver Zusammenhang)
  Werte um 0 sind (zumindest nicht linear) korreliert

Korrelation (Beträge)	
0:	keine
0 - 0,5:	schwache
0,5 – 0,8 :	mittlere
0,8 – 1:	starke
1:	perfekte

4.2 Korrelationskoeffizient (Pearson)

i	Alter x_i	Preis y_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	2	11700	-2	4	3030	9180900	-6060
2	3	10300	-1	1	1630	2656900	-1630
3	4	8350	0	0	-320	102400	0
4	5	7100	1	1	-1570	2464900	-1570
5	6	5900	2	4	-2770	7672900	-5540
Summe	**20**	**43350**	**0**	**10**	**0**	**22078000**	**-14800**
MW	4	8670		2		4415600	-2960

$$s_{xy} = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{5}(-14800) = -2960$$

$$r_{xy} = \frac{-2960}{\sqrt{2 \cdot 4415600}} = -0{,}9960516$$

4.2 Korrelationskoeffizient (Pearson): Beispiel

i	x_i	y_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$y_i-\bar{y}$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	1	5	-1,9	3,61	1,4	1,96	-2,66
2	2	5	-0,9	0,81	1,4	1,96	-1,26
3	3	3	0,1	0,01	-0,6	0,36	-0,06
4	2	2	-0,9	0,81	-1,6	2,56	1,44
5	1	4	-1,9	3,61	0,4	0,16	-0,76
6	6	3	3,1	9,61	-0,6	0,36	-1,86
7	4	2	1,1	1,21	-1,6	2,56	-1,76
8	4	4	1,1	1,21	0,4	0,16	0,44
9	2	5	-0,9	0,81	1,4	1,96	-1,26
10	4	3	1,1	1,21	-0,6	0,36	-0,66
Summe	**29**	**36**	**0**	**22,9**	**0**	**12,4**	**-8,4**
MW	2,9	3,6	0	2,29	0	1,24	-0,84

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

$$r_{xy} = \frac{-0{,}84}{\sqrt{2{,}29 \cdot 1{,}24}} = -0{,}4984834$$

Rangkorrelation nach Spearman

4.3 Rangkorrelation (Spearman)

- Robust gegenüber Ausreißer
- anzuwenden bei nicht normalverteilten Daten (folgt), oder ordinale Daten
- Daten sortieren, die Position steht für den Rang
- Bei gleichen Werten (Bindungen/ ties) wird der MW (der Ränge) gebildet

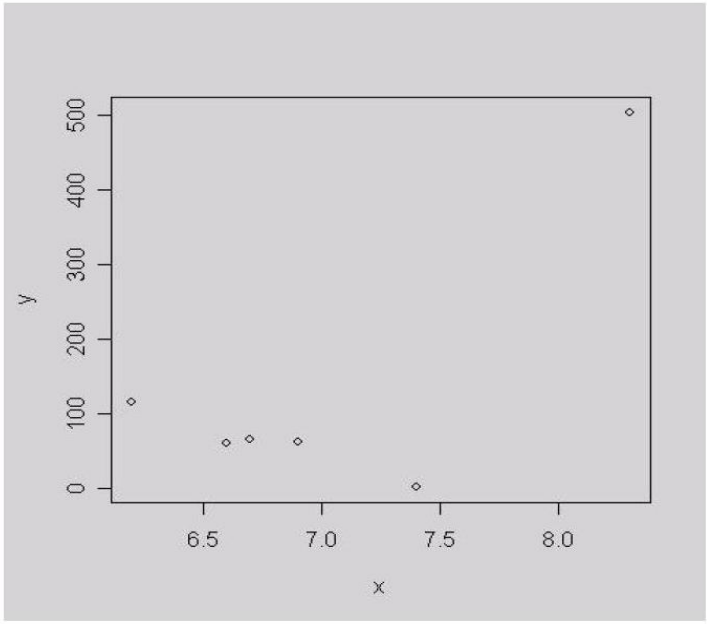
Korrelation (Beträge)	
0:	keine
0 - 0,5:	schwache
0,5 – 0,8 :	mittlere
0,8 – 1:	starke
1:	perfekte

$$\text{Spearman } r_{SP} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \; mit \; d_i = rg(x_i) - rg(y_i)$$

4.3 Rangkorrelation (Spearman): Beispiel

In der folgenden Tabelle finden Sie die Stärke eines Erdbebens x_i und die Anzahl y_i der Personen, die bei diesem Erdbeben starben:

i	x_i	y_i	r_i	s_i	d_i
1	6,6	60	2	2	0
2	8,3	503	6	6	0
3	6,2	115	1	5	4
4	6,7	65	3	4	1
5	6,9	62	4	3	-1
6	7,4	1	5	1	-4



$$r_s = 1 - \frac{6 \cdot [0^2 + 0^2 + 4^2 + 1^2 + (-1)^2 + (-4)^2]}{6 \cdot (6^2 - 1)} = 0,029$$

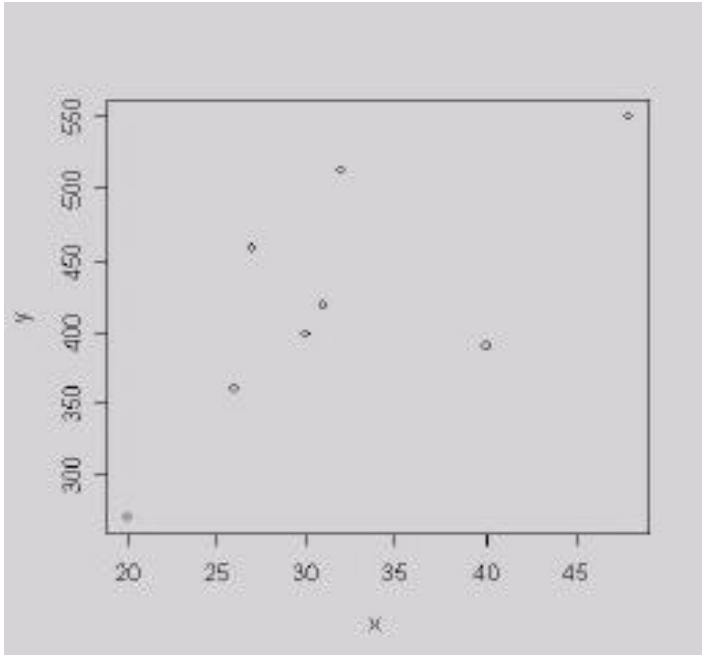
$$\text{Spearman } r_{SP} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)} \; mit \; d_i = rg(x_i) - rg(y_i)$$

4.3 Rangkorrelation (Spearman): Beispiel

Es sind zu acht Wohnungen die Fläche x_i und die Kaltmiete y_i gegeben:

(c) Bestimmen und interpretieren Sie den Rangkorrelationskoeffizienten von Spearman.

i	x_i	y_i	r_i	s_i	d_i
1	20	270	1	1	0
2	27	460	3	6	3
3	32	512	6	7	1
4	48	550	8	8	0
5	26	360	2	2	0
6	30	399	4	4	0
7	31	419	5	5	0
8	40	390	7	3	-4



$$r_s = 1 - \frac{6 \cdot [0^2 + 3^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2 + (-4)^2]}{8 \cdot (8^2 - 1)} = 0,69$$

Der Wert i.H.v. $0,69$ spricht für einen recht starken, positiv monotonen Zusammenhang zwischen der Fläche einer Wohnung und der Miete, d.h. mit steigender Fläche einer Wohnung steigt auch die Miete.

Pearson:

$$r_{x,y} = \frac{13815,38 - 31,75 \cdot 420}{\sqrt{(1074,25 - 1008,06)} \cdot \sqrt{(183200,8 - 176400)}} = 0,716$$

Dieser spricht für einen starken, positiv linearen Zusammenhang zwischen der Größe einer Wohnung und der Miete.

Kontingenz und Häufigkeiten (Chi^2)

4.4 Kontingenzkoeffizient: Beispiel

Bei einer Befragung wurden 25 Personen nach ihrem Geschlecht befragt Außerdem mussten die Personen den folgenden Satz ergänzen
Zu Risiken und Nebenwirkungen...
Von den 13 Frauen haben 7 und von den Männern 3 den Satz richtig ergänzt.

(a) Stellen Sie die Kontingenztabelle auf.

Satz \ Geschlecht	weiblich	männlich	
richtig	7	3	
falsch			
	13		25

Kontingenztabelle

(c) Bestimmen Sie den Wert des korrigierten Kontingenzkoeffizientens

4.4 Kontingenzkoeffizient: Beispiel

Bei einer Befragung wurden 25 Personen nach ihrem Geschlecht befragt Außerdem mussten die Personen den folgenden Satz ergänzen
Zu Risiken und Nebenwirkungen...
Von den 13 Frauen haben 7 und von den Männern 3 den Satz richtig ergänzt.

(a) Stellen Sie die Kontingenztabelle auf.

Satz \ Geschlecht	weiblich	männlich	
richtig	7	3	10
falsch	6	9	15
	13	12	25

Kontingenztabelle

4.4 Kontingenzkoeffizient: Beispiel

Satz \ Geschlecht	weiblich	männlich	
richtig	7	3	10
falsch	6	9	15
	13	12	25

1. Tabelle der erwarteten Häufigkeiten:

Satz \ Geschlecht	weiblich	m‰nnlich	
richtig	5,2	4,8	10
falsch	7,8	7,2	15
	13	12	25

(10*13)/25 = 5,2

2. χ^2:

Differenzen quadrieren und normieren

$$\chi^2 = \frac{(7-5,2)^2}{5,2} + \frac{(3-4,8)^2}{4,8} + \frac{(6-7,8)^2}{7,8} + \frac{(9-7,2)^2}{7,2}$$
$$= 2,17$$

3. k:

$$k = \sqrt{\frac{2,17}{2,17+25}} = 0,2826$$

Korrelation (Beträge)

0: keine

0 - 0,5: schwache

0,5 – 0,8 : mittlere

0,8 – 1: starke

1: perfekte

4. korrigierter Kontingenzkoeffizient k^*:

$$M = min\{2,2\} = 2 \;\rightarrow\; k_{max} = \sqrt{\frac{2-1}{2}} = \sqrt{0,5} \;\rightarrow\; k^* = \frac{0,2826}{\sqrt{0,5}} = 0,399$$

Formel auf der nächsten Folie

4.4 Kontingenzkoeffizient

	Frauen	Männer	SUMME
Ja	19	18	37
Nein	43	20	63
SUMME	62	38	100

Berechnung des χ^2-Koeffizienten:

$$\frac{(19-\frac{37*62}{100})^2}{\frac{37*62}{100}}+\frac{(18-\frac{37*38}{100})^2}{\frac{37*38}{100}}+\frac{(43-\frac{63*62}{100})^2}{\frac{63*62}{100}}+\frac{(20-\frac{63*38}{100})^2}{\frac{63*38}{100}}=2{,}83$$

Kontingenzkoeffizient *k*:

$$k=\sqrt{\frac{\chi^2}{\chi^2+n}} \qquad k=\sqrt{\frac{2{,}83}{2{,}83+100}}=0{,}1659$$

Korrigierter Kontingenzkoeffizient *k**:

$$k^*=\frac{k}{k_{\max}}=\frac{0{,}1659}{\sqrt{0{,}5}}=0{,}234$$

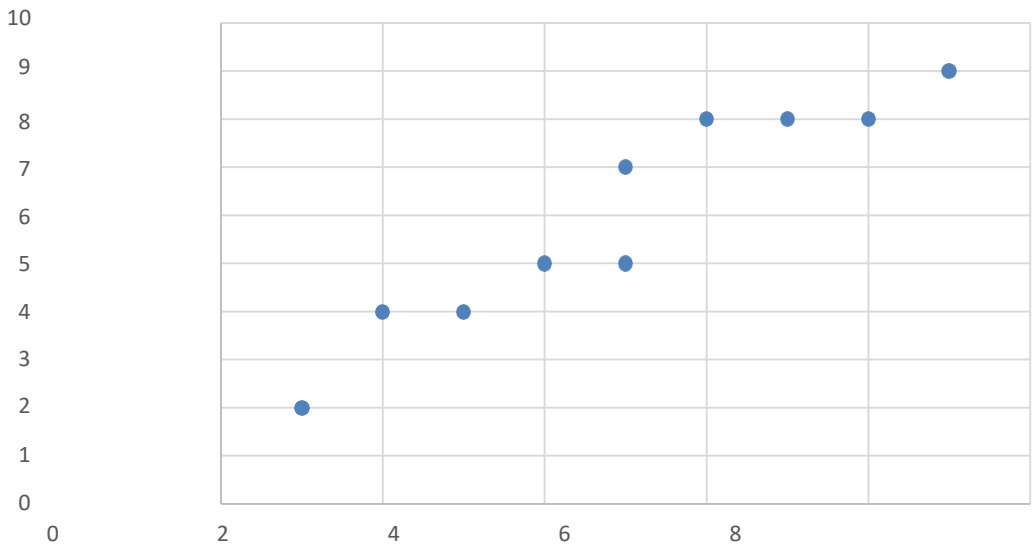
$$k_{max}=\sqrt{\frac{M-1}{M}} \quad \text{mit} \quad M=\min(\mathrm{I};\mathrm{J})$$

$$k_{max}=\sqrt{\frac{2-1}{2}}=\sqrt{0{,}5} \qquad M=\min(2;2)=2$$

Lineare Regression

4.5 Lineare Regression

i	x_i	y_i
1	1	2
2	2	4
3	3	4
4	5	5
5	4	5
6	7	8
7	8	8
8	9	9
9	5	7
10	6	8



Lineare Funktion:
y=mx + b (Normalform)
m =Steigung
b = y- Achsenabschnitt

Lineare Regression (Ausgleichsgrade):

$$\hat{y} = a + bx$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{cov_{xy}}{var_x}$$

$$a = \bar{y} - b\bar{x}$$

i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	1	2	-4	-4	16	16
2	2	4	-3	-2	9	6
3	3	4	-2	-2	4	4
4	5	5	0	-1	0	0
5	4	5	-1	-1	1	1
6	7	8	2	2	4	4
7	8	8	3	2	9	6
8	9	9	4	3	16	12
9	5	7	0	1	0	0
10	6	8	1	2	1	2
Summe	50	60	0	0	60	51
MW	5	6	0	0	6	5,1

$$\hat{y} = a + bx$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{cov_{xy}}{var_x}$$

$$a = \bar{y} - b\bar{x}$$

$$a = \bar{y} - bx = 6 - 0{,}85 * 5 = 1{,}75$$

$$b = \frac{\sigma_{xy}}{\sigma_x} = \frac{5{,}1}{6} = 0{,}85$$

$$\hat{y} = a + bx$$

$$\hat{y} = 1{,}75 + 0{,}85x$$

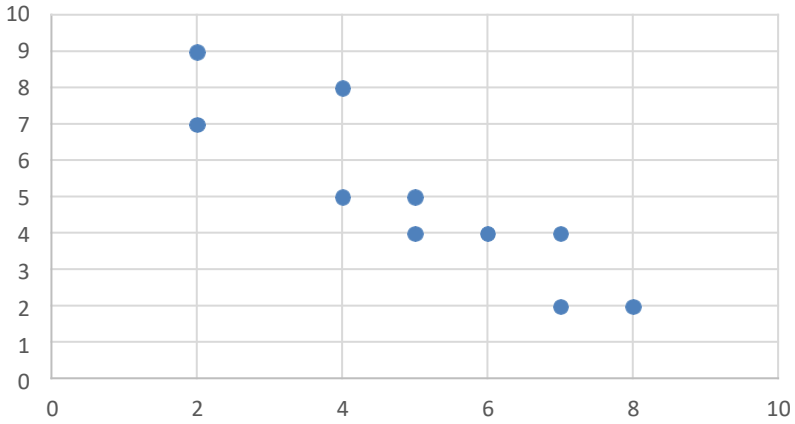
i	x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	8	2	3	-3	9	-9
2	4	5	-1	0	1	0
3	6	4	1	-1	1	-1
4	7	4	2	-1	4	-2
5	7	2	2	-3	4	-6
6	5	5	0	0	0	0
7	5	4	0	-1	0	0
8	2	7	-3	2	9	-6
9	2	9	-3	4	9	-12
10	4	8	-1	3	1	-3
Summe	50	50	0	0	38	-39
MW	5	5	0	0	3,8	-3,9

$$\hat{y} = a + bx$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{cov_{xy}}{var_x}$$

$$a = \bar{y} - b\bar{x}$$

-1,0263158	b
10,1315789	a



$\hat{y}$=10,13-1,026x