Assoziation & Korrelation

Einleitung

- Bei Beobachtung von 2 Merkmalen stellt sich die Frage, ob es Zusammenhänge oder Abhängigkeiten zwischen den Merkmalen gibt.
- Für die Messung der quantitativen Stärke des Zusammenhangs dienen im Falle qualitativer Merkmale die sog. Assoziationsmaße im Falle quantitativer Merkmale spricht man von Korrelationsmaßen
- Bisher besprochene Assoziationsmaße:
  - Cross-product ratio
  - Assoziationskoeffizient nachYule

Maße der prädiktiven Assoziation

- Diese Maße basieren auf der proportionalen Fehlerreduktion, die sich bei der Vorhersage eines Merkmals bei Kenntnis des anderen Merkmals ergeben (Goodman-Kruskal λ)
- E0 ... Fehler bei Vorhersage von Merkmal Y ohne Kenntnis von X
- E1 ... Fehler bei Vorhersage von Merkmal Y bei Kenntnis von X
- PRE(X) = (E0-E1)/E0

Beispiel

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	327	306	141	774
SPD	198	300	216	714
FDP	49	109	41	199
Grüne	92	129	134	355
PDS	10	16	100	126
	676	860	632	2168

Quelle: Allbus 1996

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	48,4%	35,6%	22,3%	35,7%
SPD	29,3%	34,9%	34,2%	32,9%
FDP	7,2%	12,7%	6,5%	9,2%
Grüne	13,6%	15,0%	21,2%	16,4%
PDS	1,5%	1,9%	15,8%	5,8%
	100,0%	100,0%	100,0%	100,0%

Prognosefehler ohne Kenntnis des zweiten Merkmals

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	327	306	141	**774**
SPD	198	300	216	714
FDP	49	109	41	199
Grüne	92	129	134	355
PDS	10	16	100	126
	676	860	632	2168

E0=2168-774=1394

E0 ist der Vorhersagefehler für die Wahlabsicht ohne Kenntnis des Merkmals Konfession

Ohne Kenntnis der Konfession ist es am sinnvollsten auf CDU zu tippen (höchste Trefferquote)

Prognosefehler bei Kenntnis des zweiten Merkmals

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	327	306	141	**774**
SPD	198	300	216	714
FDP	49	109	41	199
Grüne	92	129	134	355
PDS	10	16	100	126
	676	860	632	2168

E1=(676-327) + (860-306) + (632-216) = 1319

E1 ist der Vorhersagefehler der Wahlabsicht bei Kenntnis des Merkmals Konfession

Bei Kenntnis der Konfession ist es am sinnvollsten bei den Ausprägungen katholisch und evangelisch auf CDU zu tippen (höchste Trefferquote) bei der Ausprägung keine auf SPD zu tippen

- Sei X das Merkmal Wahlabsicht und Y das Merkmal Konfession, so gilt für

  Pre(X) = 1- 1319/1394=0.054
- Demgemäß verbessert sich die Vorhersage der Wahlabsicht bei Kenntnis der Konfessionszugehörigkeit um 5,4%.
- Man beachte, dass dieses Maß gerichtet ist, d.h. dass es nicht symmetrisch in Bezug auf die Rollen der Variablen ist
- Die Vorhersage der Konfessionszugehörigkeit kann bei Kenntnis der Wahlabsicht um 8,4% gesteigert werden.

Vorhersage der Konfessionszugehörigkeit bei Kenntnis der Wahlabsicht

		Konfession			
	katholisch	evangelisch	keine	gesamt	
CDU	327	306	141	**774**	
SPD	198	300	216	714	
FDP	49	109	41	199	
Grüne	92	129	134	355	
PDS	10	16	100	126	
	676	860	632	2168	
	E0=2168-860=	1308			

		Konfession			
	katholisch	evangelisch	keine	gesamt	
CDU	327	306	141	**774**	447
SPD	198	300	216	714	414
FDP	49	109	41	199	90
Grüne	92	129	134	355	221
PDS	10	16	100	126	26
	676	860	632	2168	1198
					E1=1198
					8,4%

Ein symmetrisches Assoziationsmaß: Cramer`s V

- Basiert auf dem Vergleich von beobachteten und unter Unabhängigkeit erwarteten Häufigkeiten
- Für eine Tabelle mit I Zeilen und J Spalten und N Beobachtungen wird wie folgt definiert:

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{Beobachtete Häufigkeit} - \text{Erwartete Häufigkeit})^2}{\text{Erwartete Häufigkeit}_i}$$

$$V = \sqrt{\frac{\chi^2}{N \min(I-1, J-1)}}$$

Beispiel:

Beobachtete Häufigkeiten

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	327	306	141	774
SPD	198	300	216	714
FDP	49	109	41	199
Grüne	92	129	134	355
PDS	10	16	100	126
	676	860	632	2168

Quelle: Allbus 1996

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	48,4%	35,6%	22,3%	35,7%
SPD	29,3%	34,9%	34,2%	32,9%
FDP	7,2%	12,7%	6,5%	9,2%
Grüne	13,6%	15,0%	21,2%	16,4%
PDS	1,5%	1,9%	15,8%	5,8%
	100,0%	100,0%	100,0%	100,0%

Erwartete Häufigkeiten

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	241,3	307,0	225,6	774
SPD	222,6	283,2	208,1	714
FDP	62,0	78,9	58,0	199
Grüne	110,7	140,8	103,5	355
PDS	39,3	50,0	36,7	126
	676	860	632	2168

	Konfession			
	katholisch	evangelisch	keine	gesamt
CDU	35,7%	35,7%	35,7%	35,7%
SPD	32,9%	32,9%	32,9%	32,9%
FDP	9,2%	9,2%	9,2%	9,2%
Grüne	16,4%	16,4%	16,4%	16,4%
PDS	5,8%	5,8%	5,8%	5,8%
	100,0%	100,0%	100,0%	100,0%

	Konfession			
	katholisch	evangelisch	keine	
CDU	30,4	0,0	31,7	
SPD	2,7	1,0	0,3	
FDP	2,7	11,4	5,0	
Grüne	3,2	1,0	9,0	
PDS	21,8	23,1	109,0	
				252,4 Chi²-Wert
n=2168	I=5	J=3		
				0,241 Cramer`s V

Interpretation: 0,1 < V < 0,2 ... geringer ZH

0,2 < V < 0,4 ... mäßiger ZH

V > 0,4 ..... Starker ZH

Kovarianz

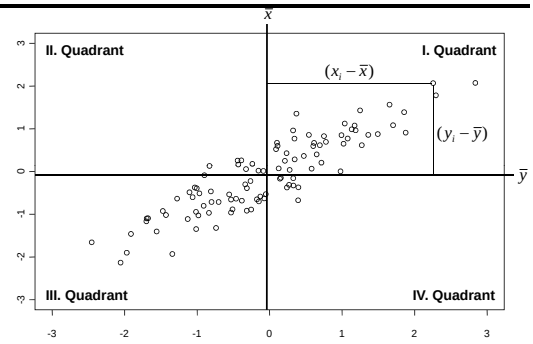
Kovarianz: Zusammenhangsmaß bei intervallskalierten Merkmalen, das sich unmittelbar aus der Varianz ableitet

$$s_{XX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i x_i - n\bar{x}\bar{x}\right)$$

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

Nachteil: keine Normierung

Korrelationskoeffizient



Korrelationskoeffizient

Der Korrelationskoeffizient ist ein Maß für den <u>linearen</u> Zusammenhang zwischen zwei Variablen X und Y.

Er ist durch folgende Formel charakterisiert:

$$r_{xy} = corr_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} =$$

$$= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - \left(\sum x_i\right)^2} \cdot \sqrt{n \sum y_i^2 - \left(\sum y_i\right)^2}}$$

Korrelationskoeffizient

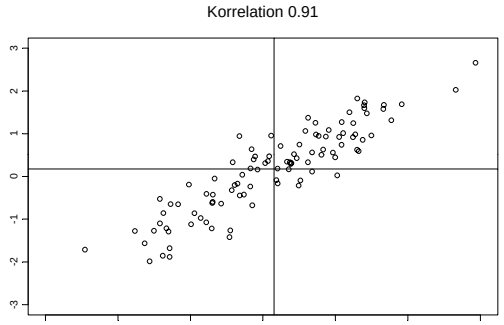
Der Korrelationskoeffizient liegt stets zwischen -1 und +1.

Korrelationskoeffizient nahe -1:
Die Mehrzahl der Datenpunkte konzentrieren sich um eine Gerade mit negativer Steigung.

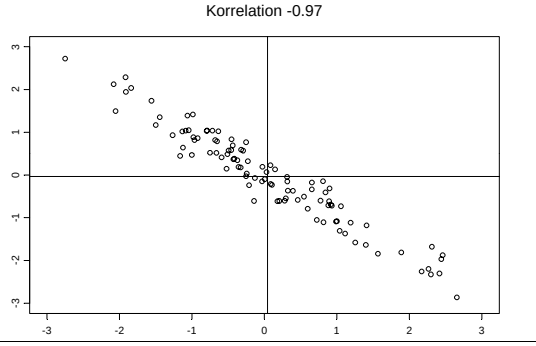
Korrelationskoeffizient ungefähr 0:
Die Datenpunkte sind entweder auf alle vier Quadranten ungefähr gleichmäßig verteilt oder sie liegen um eine Gerade die parallel zu einer Achse verläuft.

Korrelationskoeffizient nahe +1:
Die Mehrzahl der Datenpunkte konzentrieren sich um eine Gerade mit positiver Steigung.

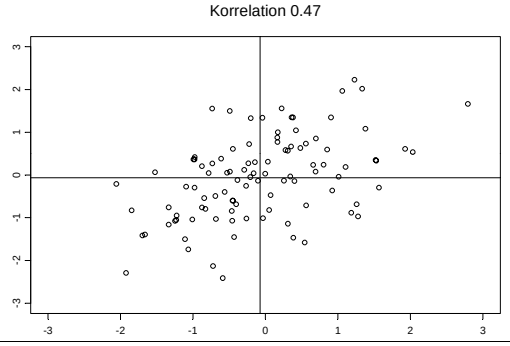
Hohe positive Korrelation



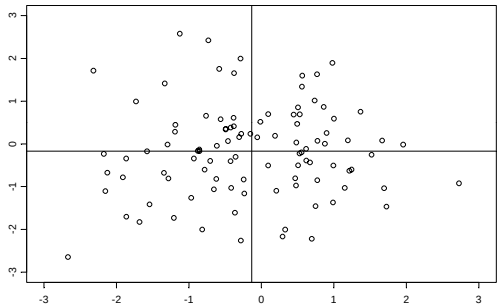
Hohe negative Korrelation



Mittlere positive Korrelation

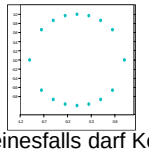


Korrelation nahe 0



Unabhängigkeit und Kausalität

- Sind zwei Variablen unabhängig, so folgt daraus, dass der Korrelationskoeffizient den Wert 0 annimmt.
- Umgekehrt kann aus einer Korrelation nicht auf Unabhängigkeit geschlossen werden, da die Korrelation nur den linearen Zusammenhang misst.



Die Punkte im linken Beispiel haben Korrelation null!

- Keinesfalls darf Korrelation mit Kausalität gleichgesetzt werden. Problem: Scheinkorrelation

Beispiel: X Gewicht des Vaters, Y Gewicht des Sohnes

i	X	Y	X²	XY	Y²
1	65	68	4225	4420	4624
2	63	66	3969	4158	4356
3	67	68	4489	4556	4624
4	64	65	4096	4160	4225
5	68	69	4624	4692	4761
6	62	66	3844	4092	4356
7	70	68	4900	4760	4624
8	66	65	4356	4290	4225
9	68	71	4624	4828	5041
10	67	67	4489	4489	4489
11	69	68	4761	4692	4624
12	71	70	5041	4970	4900
Summe	800	811	53418	54107	54849
Mittel	66,7	67,6	4451,5	4508,9	4570,8

Kovarianz	Sxy	3,361	3,361	
Varianz X	Sxx	7,056	7,056	
Varianz Y	Syy	3,243	3,243	
Korrelation	Rxy	0,70		

Excel-Funktionen:
- Kovar
- Varianzen
- Korrel