Bivariate Verteilungen

Bei der univariaten Datenanalyse wird jeweils eine Variable analysiert. Im Zuge der univariaten Datenanalyse werden zwar i.d.R. mehrere Variablen untersucht, aber jede Variablenverteilung isoliert von den anderen Variablen betrachtet.

Bei der bivariaten Datenanalyse wird die gemeinsame Verteilung zweier Variablen untersucht.

Der übliche Anwendungsfall der bivariaten Datenanalyse besteht in der Analyse von Zusammenhängen zwischen zwei Merkmalen.

Im Allgemeinen zeigt sich ein Zusammenhang zweier Merkmale darin, dass sich die Verteilung einer Variable je Ausprägung einer anderen Variable unterscheidet.

Beispiel: Wenn sich die Verteilung der Menge des konsumierten Alkohols zwischen Männern und Frauen unterscheidet, dann sprechen wir von einem Zusammenhang zwischen den Variablen Geschlecht und Alkoholkonsum.

Die Frage, wie dieser Zusammenhang zu interpretieren ist, ist jedoch i.d.R. auf theoretischer Basis zu entscheiden.

Ein Zusammenhang zwischen Variable A und B kann beispielsweise daraus resultieren,
dass Variable A Variable B beeinflusst,
dass Variable B Variable A beeinflusst, oder
dass es eine Variable C gibt, die Variable B und A beeinflusst.

Die hinter einem Zusammenhang stehenden Kausalstrukturen sind zumeist nur durch experimentelle Untersuchungsdesigns zu klären.

Tabellenanalyse

Weitere Bezeichungen:

- Zweidimensionale Tabellenanalyse
- Kreuztabellen
- Bivariate Auswertung
- Kontingenzanalyse

Grundelemente:

Kontingenztabelle

marginale / konditionale Verteilungen

Signifikanztest

Assoziationsmaße

Grundelement: 2-Dimensionale Tabelle (Kontingenztabelle)

		Variable 1		
		1	2	gesamt
Variable 2	1	n11	n12	n1.
	2	n21	n22	n2.
	gesamt	n.1	n.2	n.. (N)

marginale Verteilung Merkmal 2

Konditionale Verteilungen

marginale Verteilung Merkmal 1

Gesamtzahl (valide)

Tabelle der **Absoluten Häufigkeiten**

		Bierkonsum		
		oft	selten	gesamt
Geschlecht	weibl.	39	81	120
	männl.	119	58	177
	gesamt	158	139	297

Prozentuiert nach Gesamthäufigkeit

		Bierkonsum		
		oft	selten	gesamt
Geschlecht	weibl.	13,1	27,3	40,4
	männl.	40,1	19,5	59,6
	gesamt	53,2	46,8	100,0

Zeilenweise prozentuiert

		Bierkonsum		
		oft	selten	gesamt
Geschlecht	weibl.	32,5	67,5	100,0
	männl.	67,2	32,8	100,0
	gesamt	53,2	46,8	100,0

Spaltenweise prozentuiert

		Bierkonsum		
		oft	selten	gesamt
Geschlecht	weibl.	24,7	58,3	40,4
	männl.	75,3	41,7	59,6
	gesamt	100,0	100,0	100,0

Vorgehen bei der Tabellenanalyse

1. Formulierung einer Hypothese
2. Inspektion der absoluten Häufigkeiten
3. Prozentuierung
4. Vergleich der konditionalen Verteilungen
5. Berechnung eines Signifikanztests
6. Berechnung von Assoziationsmaßen

1. Formulierung einer Hypothese

Entscheidung über die Art des hypothetischen Zusammenhanges:

A $\longrightarrow$ B asymmetrischer Zusammenhang

H_0: Geschlecht $\not\longrightarrow$ Bierkonsum

H_1: Geschlecht $\longrightarrow$ Bierkonsum

Geschlecht=unabhängige Variable, Bierkonsum=abhängige Variable

Die Notation H_0 / H_1 entspricht der Testlogik beim statistischen Testen.

Hier: Es wird grundsätzlich davon ausgegangen, dass kein Zusammenhang besteht (H_0). Erst dann wenn sich starke Anhaltspunkte dafür zeigen, die für einen Zusammenhang sprechen, wird die Alternativhypothese (H_1) angenommen.

2. Inspektion der absoluten Häufigkeiten

		Bierkonsum		
		oft	selten	gesamt
Geschlecht	weibl.	39	81	120
	männl.	119	58	177
	gesamt	158	139	297

-Relation zwischen absoluten Häufigkeiten und den späteren Prozentwerten herstellen!
-Durch listenweise Falleliminierung kann die verbleibende Fallzahl drastisch reduziert werden!
-Bei vielen dünn besetzten Zellen müssen die Ausprägungen gegebenenfalls zusammengefasst werden (Umcodierung)

Beispiel für eine Tabelle mit zu dünn besetzten Zellen:

V111 ak_xtc * V113 ak_heroin Kreuztabelle

Anzahl

		V113 ak_heroin				
		1,00	2,00	3,00	4,00	Gesamt
V111 ak_xtc	1,00	175	1	0	0	176
	2,00	23	21	1	0	45
	3,00	5	3	3	0	11
	4,00	0	0	1	1	2
Gesamt		203	25	5	1	234

Diese Konstellation ist für eine Tabellenanalyse nicht sinnvoll.

3. Prozentuierung

in Richtung der unabhängigen Variablen

abhängige Variable lt. Hypothese

unabhängige Variable lt. Hypothese

		Bierkonsum		
		oft	selten	gesamt
Geschlecht	weibl.	32,5	67,5	100,0
	männl.	67,2	32,8	100,0
	gesamt	53,2	46,8	100,0

Prozentuierung

4. Vergleich der konditionalen Verteilungen

In Richtung der abhängigen Variable

		Bierkonsum		
		oft	selten	gesamt
Geschlecht	weibl.	32,5	67,5	100,0
	männl.	67,2	32,8	100,0
	gesamt	53,2	46,8	100,0

5. Berechnung eines Signifikanztests

Bei Zusammenhangsanalysen, welche lediglich die nominalen Eigenschaften von Merkmalen berücksichtigen: Chi-Quadrat (χ^2) Test

Beim χ^2-Test werden die beobachteten konditionalen Häufigkeiten mit den theoretischen Häufigkeiten verglichen, die auftreten würden, wenn die beiden Merkmale voneinander unabhängig wären (=Erwartungswerte bei statistischer Unabhängigkeit).

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(f_{b_{ij}} - f_{e_{ij}}\right)^2}{f_{e_{ij}}}$$

f_{bij}...beobachtete Häufigkeit in der i-ten Zeile und j-ten Spalte

f_{eij}...erwartete Häufigkeit in der i-ten Zeile und j-ten Spalte

Berechnung der Erwartungswerte

Erwartungswerte bei statistischer Unabhängigkeit: Jene konditionalen Häufigkeiten, die auftreten würden, wenn sich der Konsum zwischen den Geschlechtern nicht unterscheiden würde.

	oft	selten	gesamt
weibl.	$\frac{158}{297}*120=63{,}8$	$\frac{139}{297}*120=56{,}2$	120
männl.	$\frac{158}{297}*177=94{,}2$	$\frac{139}{297}*177=82{,}8$	177
gesamt	158	139	297

Kontrolle: Zeilenweises Prozentuieren bei Erwartungswerten

(Erwartete Häufigkeiten in Klammern)

	oft	selten	gesamt
weibl.	**53,2** (63,8)	**46,8** (56,2)	100,0 (120)
männl.	**53,2** (94,2)	**46,8** (82,8)	100,0 (177)
gesamt	**53,2** (158)	**46,8** (139)	(100,0) 297

Berechnung der χ2-Testgröße

Erwartungswerte in Klammern

	oft	selten	gesamt
weibl.	39 (63,8)	81 (56,2)	120
männl.	119 (94,2)	58 (82,8)	177
gesamt	158	139	297

$$\chi^2 = \frac{(39-63{,}8)^2}{63{,}8} + \frac{(81-56{,}2)^2}{56{,}2} + \frac{(119-94{,}2)^2}{94{,}2} + \frac{(58-82{,}8)^2}{82{,}8} = 34.6$$

df=(r-1)*(c-1) mit r, c Anzahl der Spalten, bzw. Zeilen =1

Vergleich der χ^2-Testgröße mit dem Tabellenwert der χ^2-Verteilung

Schlägt man in der Tabelle der χ^2-Verteilung nach, so ergibt sich bei fixierter Sicherheitswahrscheinlichkeit von 95% (Irrtumswahrscheinlichkeit $p=0.05$) und $df=1$ ein „kritischer“ χ^2-Wert von 3,84.

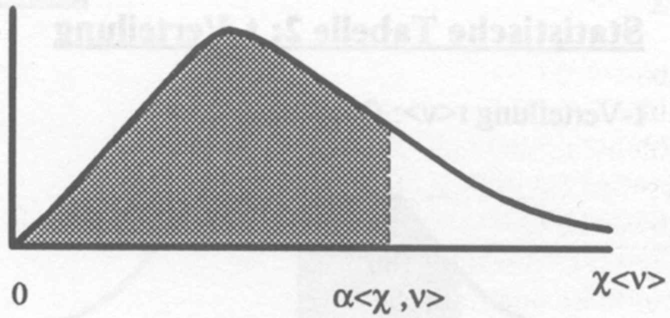
Der errechnete χ^2-Wert von 34,6 ist deutlich höher als der kritische Wert. Die H_o-Hypothese kann daher mit $p<0.05$ verworfen werden.

Anders formuliert: Ein χ^2-Wert von 34,6 bei $df=1$ tritt mit einer Wahrscheinlichkeit $p<0.0001$ auf, wenn die beiden Merkmale voneinander unabhängig sind.

Wir schließen daraus, dass wir mit einer Sicherheit von nahezu 100% davon ausgehen können, dass wir die H_0-Hypothese ablehnen können.

Statistische Tabelle 3: Chiquadrat-Verteilung

χ^2-Verteilung: Quantile $\alpha\langle\chi,\nu\rangle$



df ↓

ν	α 0,900	0,950	0,975	0,990	0,995
1	2,71	3,84	5,02	6,6	7,9
2	4,61	5,99	7,38	9,2	10,6
3	6,25	7,81	9,35	11,3	12,8
4	7,78	9,49	11,1	13,3	14,9
5	9,24	11,1	12,8	15,1	16,7
6	10,6	12,6	14,4	16,8	18,5
7	12,0	14,1	16,0	18,5	20,3
8	13,4	15,5	17,5	20,1	22,0
9	14,7	16,9	19,0	21,7	23,6
10	16,0	18,3	20,5	23,2	25,2

→ 34,6

„krit. Wert“

Einschränkung: Die erwarteten Häufigkeiten dürfen nicht zu klein sein.
Keine Erwartungswerte <1 und nicht mehr als 10% der Erwartungswerte <5
(manche Autoren „erlauben“ maximal 20% der Erwartungswerte <5)

Tabellenanalyse mit SPSS:

Beispiel: Ist die Akzeptanz von Cannabiskonsum vom Geschlecht abhängig?

28.) Wie beurteilen Sie es persönlich, wenn Menschen folgende Substanzen konsumieren?

	völlig akzeptabel	eher akzeptabel	eher nicht akzeptabel	gar nicht akzeptabel	weiß nicht	
Zigaretten	①	②	③	④	○	
Wein, Bier, Sekt, etc.	①	②	③	④	○	
Hochprozentiges (Whisky, Wodka, etc.)	①	②	③	④	○	
Haschisch / Cannabis / Marihuana	①	②	③	④	○	V108

30.) Geschlecht: ○ Männlich ○ Weiblich V132

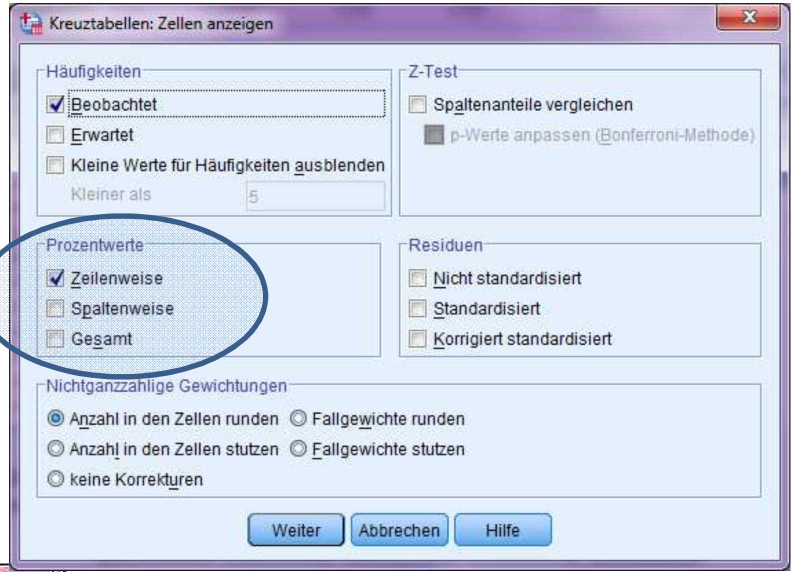
1 2

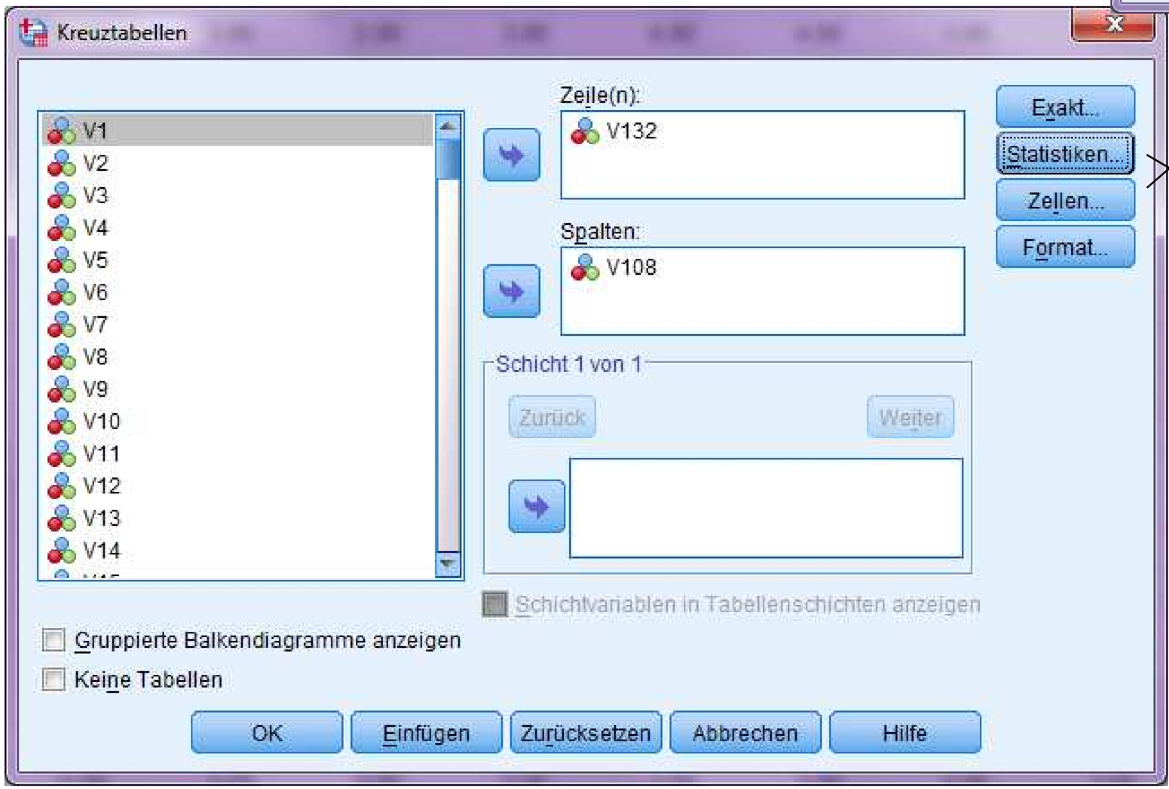
Für diese Analyse wird die Substanzakzeptanz dichotomisiert. Die Ausprägung "Weiß-nicht" wird als fehlender Wert codiert.

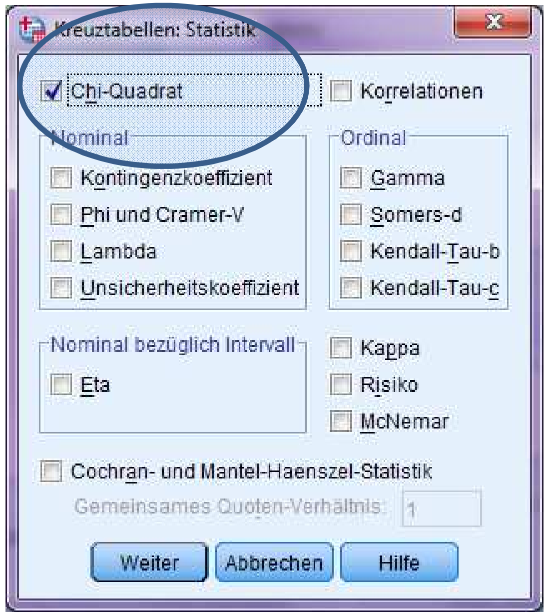
recode v108(1,2=1)(3,4=2)(5=sysmis).

Tabellenanalyse unter "Analysieren" -> "Deskriptive Statistiken" -> "Kreuztabelle" anfordern.

Empfehlenswert: Unabhängige Variable=Zeilenvariable, abhängige Variable=Spaltenvariable.







Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
V132 geschlecht * V108 ak_cannabis	234	97,5%	6	2,5%	240	100,0%

234 von 240 Fällen bleiben für die Analyse

V132 geschlecht * V108 ak_cannabis Kreuztabelle

			V108 ak_cannabis		Gesamt
			1,00	2,00	
V132 geschlecht	1,00	Anzahl	60	63	123
		% innerhalb von V132 geschlecht	48,8%	51,2%	100,0%
	2,00	Anzahl	34	77	111
		% innerhalb von V132 geschlecht	30,6%	69,4%	100,0%
Gesamt		Anzahl	94	140	234
		% innerhalb von V132 geschlecht	40,2%	59,8%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)	Exakte Signifikanz (2-seitig)	Exakte Signifikanz (1-seitig)
Chi-Quadrat nach Pearson	7,997[a]	1	,005		
Kontinuitätskorrektur[b]	7,260	1	,007		
Likelihood-Quotient	8,073	1	,004		
Exakter Test nach Fisher				,005	,003
Zusammenhang linear-mit-linear	7,963	1	,005		
Anzahl der gültigen Fälle	234				

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 44,59.

b. Wird nur für eine 2x2-Tabelle berechnet

χ2-Wert von 7,997 ist bei df=1 mit p=0,005 signifikant

keine Probleme mit kleinen Erwartungswerten

Berechnung von Assoziationsmaßen

Neben der Signifikanz ist auch die Stärke des Zusammenhanges relevant.

Es gibt zwar einen Zusammenhang zwischen der Stärke des Zusammenhanges und dem $\chi 2$-Wert, da dieser umso größer wird, je stärker sich die konditionalen Verteilungen unterscheiden.

Allerdings ist der $\chi 2$-Wert nicht normiert (hängt von der Tabellengröße und der Stichprobengröße ab).

Je größer die Fallzahl, umso relevanter ist ein Stärkemaß gegenüber dem Signifikanzurteil!

Als Stärkemaße werden Assoziationsmaße (umgangssprachlich Korrelationskoeffizienten) berechnet.

Assoziationsmaße sind üblicherweise normiert auf den Bereich zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang).

Es gibt je nach Messniveau bzw. Anzahl der Ausprägungen unterschiedliche Assoziationsmaße.

Assoziationsmaße für ordinales und metrisches Meßniveau besitzen ein Vorzeichen (-1 bis +1).

Assoziationsmaß für zwei dichotome Merkmale (2x2-Tabelle):

Der Phi-Koeffizient ist ein Stärkemaß für den Zusammenhang zweier dichotomer Merkmale. Er wird aus der $\chi 2$-Testgröße berechnet, indem diese mit der Stichprobengröße normiert wird.

Eine Normierung der Tabellengröße entfällt, da diese bei dichotomen Merkmalen fixiert ist.

$$\Phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{7{,}997}{234}} = 0{,}1848$$

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,185	,005
	Cramer-V	,185	,005
Anzahl der gültigen Fälle		234	

Die Signifikanz des Phi-Koeffizienten ist die Signifikanz des $\chi 2$-Tests.

Assoziationsmaße für polytome nominale Merkmale:

Kontingenzkoeffizient $$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Ungünstig: Maximalwert abhängig von Anzahl der Zeilen bzw. Spalten.

$$c_{\max} = \sqrt{(r-1)/r}$$

χ^2... Chi-Quadrat Testgröße
n... Anzahl der Fälle
r... min (Zeilenanzahl, Spaltenanzahl)

Cramers V $$\mathrm{V} = \sqrt{\frac{\chi^2}{n * (\mathrm{r} - 1)}}$$

χ^2... Chi-Quadrat Testgröße
n... Anzahl der Fälle
r... min (Zeilenanzahl, Spaltenanzahl)

Signifikanztest für beide Koeffizienten ist der χ2-Test.

Ist die Akzeptanz des Cannabis-Konsums abhängig von der Fakultat, an der die Befragten studieren?

Fakultät: 1=Re, 2=SOWI, 3=TNF

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
fakultät * V108 ak_cannabis	234	97,5%	6	2,5%	240	100,0%

fakultät * V108 ak_cannabis Kreuztabelle

			V108 ak_cannabis				Gesamt
			1,00	2,00	3,00	4,00	
fakultät	1,00	Anzahl	9	16	35	25	85
		% innerhalb von fakultät	10,6%	18,8%	41,2%	29,4%	100,0%
	2,00	Anzahl	11	35	32	31	109
		% innerhalb von fakultät	10,1%	32,1%	29,4%	28,4%	100,0%
	3,00	Anzahl	7	16	7	10	40
		% innerhalb von fakultät	17,5%	40,0%	17,5%	25,0%	100,0%
Gesamt		Anzahl	27	67	74	66	234
		% innerhalb von fakultät	11,5%	28,6%	31,6%	28,2%	100,0%

Nicht signifikant (p>0.05)

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	11,963[a]	6	,063
Likelihood-Quotient	12,269	6	,056
Zusammenhang linear-mit-linear	4,088	1	,043
Anzahl der gültigen Fälle	234		

a. 1 Zellen (8,3%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 4,62.

Stärke: 0,16 (Cramers V) bzw. 0,22 (Kontingenzkoeffizient).
Phi darf nicht verwendet werden!

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,226	,063
	Cramer-V	,160	,063
	Kontingenzkoeffizient	,221	,063
Anzahl der gültigen Fälle		234	

Berechnung mit

comp v108II=v108.
recode v108II(1,2=1)(3,4=2).

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Proz
fakultät * v108II	234	97,5%	6	2,5%	240	100,

fakultät * v108II Kreuztabelle

			v108II		Gesamt
			1,00	2,00	
fakultät	1,00	Anzahl	25	60	85
		% innerhalb von fakultät	29,4%	70,6%	100,0%
	2,00	Anzahl	46	63	109
		% innerhalb von fakultät	42,2%	57,8%	100,0%
	3,00	Anzahl	23	17	40
		% innerhalb von fakultät	57,5%	42,5%	100,0%
Gesamt		Anzahl	94	140	234
		% innerhalb von fakultät	40,2%	59,8%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	9,279[a]	2	,010
Likelihood-Quotient	9,313	2	,009
Zusammenhang linear-mit-linear	9,204	1	,002
Anzahl der gültigen Fälle	234		

a. 0 Zellen (0,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist 16,07.

Symmetrische Maße

		Wert	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,199	,010
	Cramer-V	,199	,010
	Kontingenzkoeffizient	,195	,010
Anzahl der gültigen Fälle		234	

Zusammenhangsanalysen mit ordinalen Variablen

Beispiel: Ist die Akzeptanz von Cannabiskonsum höher, wenn die Befragten eine höhere Akzeptanz von Nikotinkonsum haben?

Der χ2-Test ist zur Beantwortung der Frage nicht zielführend.

Beide Variablen besitzen mind. ordinales Messniveau. Aufgrund der Hierarchie der Messniveaus kann zwar ein χ2-Test berechnet werden.

Die Hypothese bezieht sich aber auf die ordinale und nicht auf die nominale Information!

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
V110 ak_zig * V108 ak_cannabis	234	97,5%	6	2,5%	240	100,0%

V110 ak_zig * V108 ak_cannabis Kreuztabelle

			V108 ak_cannabis				Gesamt
			1,00	2,00	3,00	4,00	
V110 ak_zig	1,00	Anzahl	21	26	32	14	93
		% innerhalb von V110 ak_zig	22,6%	28,0%	34,4%	15,1%	100,0%
	2,00	Anzahl	5	32	33	34	104
		% innerhalb von V110 ak_zig	4,8%	30,8%	31,7%	32,7%	100,0%
	3,00	Anzahl	1	8	8	16	33
		% innerhalb von V110 ak_zig	3,0%	24,2%	24,2%	48,5%	100,0%
	4,00	Anzahl	0	1	1	2	4
		% innerhalb von V110 ak_zig	0,0%	25,0%	25,0%	50,0%	100,0%
Gesamt		Anzahl	27	67	74	66	234
		% innerhalb von V110 ak_zig	11,5%	28,6%	31,6%	28,2%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	29,647[a]	9	,001
Likelihood-Quotient	30,186	9	,000
Zusammenhang linear-mit-linear	19,406	1	,000
Anzahl der gültigen Fälle	234		

a. 5 Zellen (31,2%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,46.

Illustration des Problems:

Der χ2-Test ist ein Test, der die nominale Information berücksichtigt.

Die Ordnung der Ausprägungen ist daher irrelevant:

Wir "vertauschen" die Ausprägungen von V110:

recode v110(1=3)(2=4)(3=1)(4=2).

Ergebnis:
Der χ2-Wert bleibt unverändert!

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
V110 ak_zig * V108 ak_cannabis	234	97,5%	6	2,5%	240	100,0%

V110 ak_zig * V108 ak_cannabis Kreuztabelle

			V108 ak_cannabis				Gesamt
			1,00	2,00	3,00	4,00	
V110 ak_zig	1,00 eher nicht akzeptabel	Anzahl	1	8	8	16	33
		% innerhalb von V110 ak_zig	3,0%	24,2%	24,2%	48,5%	100,0%
	2,00 gar nicht akzeptabel	Anzahl	0	1	1	2	4
		% innerhalb von V110 ak_zig	0,0%	25,0%	25,0%	50,0%	100,0%
	3,00 völlig akzeptabel	Anzahl	21	26	32	14	93
		% innerhalb von V110 ak_zig	22,6%	28,0%	34,4%	15,1%	100,0%
	4,00 eher akzeptabel	Anzahl	5	32	33	34	104
		% innerhalb von V110 ak_zig	4,8%	30,8%	31,7%	32,7%	100,0%
Gesamt		Anzahl	27	67	74	66	234
		% innerhalb von V110 ak_zig	11,5%	28,6%	31,6%	28,2%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	29,647[a]	9	,001
Likelihood-Quotient	30,186	9	,000
Zusammenhang linear-mit-linear	,735	1	,391
Anzahl der gültigen Fälle	234		

a. 5 Zellen (31,2%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häufigkeit ist ,46.

Fazit: Der χ2-Test darf natürlich auch bei ordinalem Messniveau berechnet werden.

In der Regel wird die zugrunde liegende Zusammenhangshypothese jedoch die ordinale Information berücksichtigen ("**je-desto-Hypothese**").

Wir fragen nicht, ob die Ausprägungen der Variable Cannabis in Abhängigkeit der Ausprägungen der Variable Nikotin unterschiedlich verteilt sind.

Wir fragen, ob ein Zusammenhang in der Form beobachtbar ist, dass eine höhere Ausprägung bei Cannabis mit einer höheren Ausprägung bei Nikotin korrespondiert!

Je-desto-Hypothese: Je akzeptabler Nikotinkonsum, desto akzeptabler Cannabiskonsum,
bzw. umgekehrt: je weniger Nikotinkonsum akzeptiert wird, umso weniger wird Cannabiskonsum akzeptiert.

Zur Beantwortung dieser Hypothese ist ein Test nötig, welcher mindestens ordinale Information berücksichtigt!

Statistische Tests / Assoziationsmaße für ordinales Messniveau

Basieren auf der Logik der Paarvergleiche – Vergleich der Zeilen- und Spaltenwerte bei je zwei Fällen der Stichprobe.

Mögliche Ergebnisse eines paarweisen Vergleichs:

konkordant: Person 1 hat in beiden Variablen niedrigere Ausprägungen als Person2

diskordant: Person 1 hat in einer Variablen eine niedrigere Ausprägungen als Person2, in der anderen Variable eine höhere Ausprägung als Person 2.

X-verbunden: Beide Personen haben dieselbe Ausprägungen bei Variable X

Y-verbunden: Beide Personen haben dieselbe Ausprägungen bei Variable Y

X,Y-verbunden: Beide Personen haben bei beiden Variablen dieselbe Ausprägung.

Ziel ist es zu ermitteln, welche Tendenz insgesamt überwiegt.

Beispiel:

Befr.	x	y
A	1	1
B	3	2
C	4	1
D	3	2

A – B $x_A < x_B, y_A < y_B$ => konkordant

A – C $x_A < x_C, y_A = y_C$ => gebunden in y

A – D $x_A < x_D, y_A < y_D$ => konkordant

B – C $x_B < x_C, y_B > y_C$ => diskordant

B – D $x_B = x_D, y_B = y_D$ => gebunden in XY

C – D $x_C > x_D, y_C < y_D$ => diskordant

Anzahl mögl. Paare:

n*(n-1)/2

4*3/2 = 6

Ordinale Assoziationsmaße

Aus diesen Informationen werden unterschiedliche Assoziationsmaße berechnet

$$\text{Gamma} = \frac{N_k - N_d}{N_k + N_d}$$

Es wird nur die Anzahl der konkordanten und der diskordanten Paare berücksichtigt.

Da die Bindungen unberücksichtigt sind, nimmt dieser Koeffizient häufig sehr hohe Werte an.

$$\text{Tau}_b = \frac{N_k - N_d}{\sqrt{(N_k + N_d + T_x) * (N_k + N_d + T_y)}}$$

In dieser Formel werden die Bindungen berücksichtigt (T_x u. T_Y)

Falls Bindungen vorliegen, ist dieser Koeffizient kleiner

Überwiegen die Konkordanten Paare, resultiert ein positiver Wert.
Überwiegen die Diskordanten Paare, resultiert ein negativer Wert.

Ordinale Assoziationsmaße besitzen einen Wertebereich zwischen -1 und +1

Negative Werte bedeuten, dass

… eine höhere Ausprägung der einen Variable tendenziell mit einer niedrigeren Ausprägung der anderen Variable einhergeht (indirekt proportionaler Zusammenhang)

positive Werte bedeuten, dass

… eine höhere Ausprägung der einen Variable tendenziell mit einer höheren Ausprägung der anderen Variable einhergeht (direkt proportionaler Zusammenhang)

Null bedeutet kein Zusammenhang.

Die Signifikanz ordinaler Zusammenhänge wird jeweils anhand eines statistischen Test des ordinalen Assoziationsmaßes überprüft (H_0: Assoziationsmaß =0; H_1: Assoziationsmaß <> 0)

Die Analyse bestätigt einen direkt proportionalen Zusammenhang.

Der Zusammenhang ist mit $p<0.05$ signifikant.

Verarbeitete Fälle

	Fälle					
	Gültig		Fehlend		Gesamt	
	N	Prozent	N	Prozent	N	Prozent
V110 ak_zig * V108 ak_cannabis	234	97,5%	6	2,5%	240	100,0%

V110 ak_zig * V108 ak_cannabis Kreuztabelle

			V108 ak_cannabis				Gesamt
			1,00	2,00	3,00	4,00	
V110 ak_zig	1,00	Anzahl	21	26	32	14	93
		% innerhalb von V110 ak_zig	22,6%	28,0%	34,4%	15,1%	100,0%
	2,00	Anzahl	5	32	33	34	104
		% innerhalb von V110 ak_zig	4,8%	30,8%	31,7%	32,7%	100,0%
	3,00	Anzahl	1	8	8	16	33
		% innerhalb von V110 ak_zig	3,0%	24,2%	24,2%	48,5%	100,0%
	4,00	Anzahl	0	1	1	2	4
		% innerhalb von V110 ak_zig	0,0%	25,0%	25,0%	50,0%	100,0%
Gesamt		Anzahl	27	67	74	66	234
		% innerhalb von V110 ak_zig	11,5%	28,6%	31,6%	28,2%	100,0%

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Nominal- bzgl. Nominalmaß	Phi	,356			,001
	Cramer-V	,206			,001
	Kontingenzkoeffizient	,335			,001
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-b	,254	,054	4,633	,000
	Gamma	,371	,077	4,633	,000
Anzahl der gültigen Fälle		234			

Zusammenhänge zwischen Metrischen Variablen:
Produkt-Moment Korrelationskoeffizienten
(Korrelation nach Pearson)

$$r = \frac{\mathrm{cov}(xy)}{\sqrt{s_x^{\ 2}} \cdot \sqrt{s_y^{\ 2}}} = \frac{\mathrm{cov}(xy)}{s_x \cdot s_y}$$

Korrelation nach Pearson

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

Kovarianz

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

Varianz
s_x=Standardabweichung

Logik der Korrelation

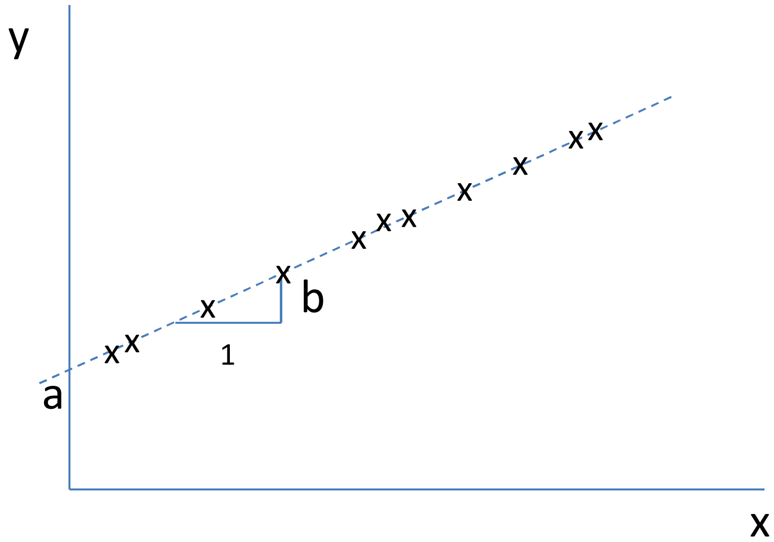
Der Korrelationskoeffizient setzt die gemeinsame Streuung zweier Merkmale im Verhältnis zu den Einzelstreuungen der Merkmale und kann zwischen 0 und ±1 variieren

Wenn die gemeinsame Streuung den Einzelstreuungen entspricht, ist das Verhältnis gleich 1. Das ist genau dann der Fall, wenn alle xy-Paare auf einer Geraden liegen.

d.h. $y=a+b*x$

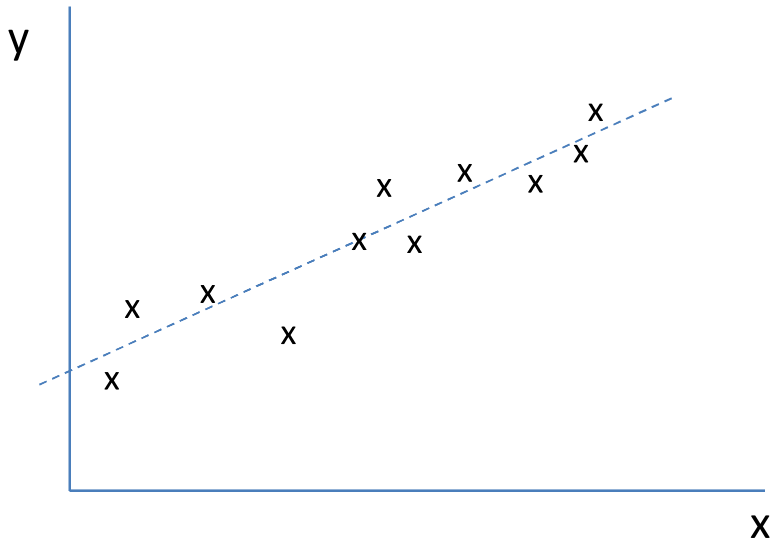
Je stärker die xy-Paare von einer Gerade abweichen umso größer wird die Differenz zwischen gemeinsamer Streuung und Einzelstreuung und das Verhältnis nähert sich dem Wert 0

-Wenn b ein positiver Wert ist, dann hat der Korrelationskoeffizient ein positives Vorzeichen (höherer Wert von x ⇔ höherer Wert von y),
-wenn b ein negativer Wert ist, dann hat der Korrelationskoeffizient ein negatives Vorzeichen (höherer Wert von x ⇔ niedriger Wert von y),



$y=a+b*x$

$r=1$



$0 < r < 1$

V110 ak_zig * V108 ak_cannabis Kreuztabelle

			V108 ak_cannabis					
			1,00	2,00	3,00	4,00	5,00	Gesamt
V110 ak_zig	1,00	Anzahl	21	26	32	14	1	94
		% innerhalb von V110 ak_zig	22,3%	27,7%	34,0%	14,9%	1,1%	100,0%
	2,00	Anzahl	5	32	33	34	3	107
		% innerhalb von V110 ak_zig	4,7%	29,9%	30,8%	31,8%	2,8%	100,0%
	3,00	Anzahl	1	8	8	16	0	33
		% innerhalb von V110 ak_zig	3,0%	24,2%	24,2%	48,5%	0,0%	100,0%
	4,00	Anzahl	0	1	1	2	1	5
		% innerhalb von V110 ak_zig	0,0%	20,0%	20,0%	40,0%	20,0%	100,0%
Gesamt		Anzahl	27	67	74	66	5	239
		% innerhalb von V110 ak_zig	11,3%	28,0%	31,0%	27,6%	2,1%	100,0%

Symmetrische Maße

		Wert	Asymptotischer Standardfehler[a]	Näherungsweises T[b]	Näherungsweise Signifikanz
Ordinal- bzgl. Ordinalmaß	Kendall-Tau-b	,254	,053	4,728	,000
	Korrelation nach Spearman	,290	,060	4,670	,000[c]
Intervall- bzgl. Intervallmaß	Pearson-R	,293	,059	4,725	,000[c]
Anzahl der gültigen Fälle		239			

Korrelationsmatrix

Korrelationen und Tau-b können auch ohne Kreuztabellen mittels Korrelationsmatrix berechnet werden.
"Analysieren" -> "Korrelation" -> "Bivariat"

Korrelationen

		V108 ak_cannabis	V110 ak_zig	V5 Unterstützungspersonen	V6 Kummerpersonen	V14 Freizeitpersonen
V108 ak_cannabis	Korrelation nach Pearson	1	,293**	-,018	,034	-,096
	Signifikanz (2-seitig)		,000	,781	,598	,138
	N	239	239	239	239	239
V110 ak_zig	Korrelation nach Pearson	,293**	1	-,068	,015	,087
	Signifikanz (2-seitig)	,000		,296	,816	,181
	N	239	240	240	240	240
V5 Unterstützungspersonen	Korrelation nach Pearson	-,018	-,068	1	,341**	,344**
	Signifikanz (2-seitig)	,781	,296		,000	,000
	N	239	240	240	240	240
V6 Kummerpersonen	Korrelation nach Pearson	,034	,015	,341**	1	,247**
	Signifikanz (2-seitig)	,598	,816	,000		,000
	N	239	240	240	240	240
V14 Freizeitpersonen	Korrelation nach Pearson	-,096	,087	,344**	,247**	1
	Signifikanz (2-seitig)	,138	,181	,000	,000	
	N	239	240	240	240	240

**. Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.

Insbesondere, wenn Variablen viele unterschiedliche Ausprägungen haben, ist die Kreuztabelle nicht informativ. Ordinale und metrische Zusammenhangsmaße können dennoch berechnet werden. Aber Vorsicht bei Ausreißern!

Zur Stärkeinterpretation von Zusammenhangsmaßen

In der Literatur gibt es unterschiedliche (willkürliche) Einteilungen zur Stärke-Interpretation des Korrelationskoeffizienten.

z.B.: Kühnel & Krebs, 2007:

$0{,}00 < \lvert r \rvert < 0{,}05$	kein Zusammenhang
$0{,}05 < \lvert r \rvert < 0{,}20$	geringer Zusammenhang
$0{,}20 < \lvert r \rvert < 0{,}50$	mittlerer Zusammenhang
$0{,}50 < \lvert r \rvert < 0{,}70$	hoher Zusammenhang
$\lvert r \rvert > 0{,}70$	sehr hoher Zusammenhang

Brosius, 1999:

$0{,}00 < \lvert r \rvert < 0{,}20$	sehr schwach
$0{,}20 < \lvert r \rvert < 0{,}40$	schwach
$0{,}40 < \lvert r \rvert < 0{,}60$	mittel
$0{,}60 < \lvert r \rvert < 0{,}80$	stark
$\lvert r \rvert > 0{,}8$	sehr stark

Wenngleich die Einteilung von Kühnel & Krebs für sozialwissenschaftliche Verhältnisse realitätsnaher erscheint, ist festzuhalten, dass eine allgemeingültige Einteilung (ohne Berücksichtigung des Kontexts) nicht möglich ist.

Wenn beispielsweise auf Basis messtheoretischer Überlegungen angenommen wird, dass zwei Items das selbe Konstrukt messen, dann wäre eine Korrelation in Höhe von 0,3 als schwacher Zusammenhang zu interpretieren.

Wenn aufgrund theoretischer Überlegungen angenommen wird, dass der Gesundheitszustand einer Person von ihrem sozialen Kapital oder der Bildung abhängt, so wäre eine Korrelation von 0,3 als starker Zusammenhang zu interpretieren.

einige Gründe dafür:

- Wir erwarten hier keinesfalls einen perfekten Zusammenhang
- Das Phänomen Gesundheit ist multikausal bedingt
- Die beteiligten Variablen können nur mit eingeschränkter Reliabilität gemessen werden
- Heterogene Untersuchungspopulationen

Ergänzende Literaturempfehlung:

Benninghaus, Hans (2001). Einführung in die Sozialwissenschaftliche Datenanalyse. München, Oldenbourg, 6.Auflage oder höher.