3. ZWEI KATEGORIALE MERKMALE (bivariate kategoriale Daten)

Beispiel: Gründe für Beliebtheit bei Klassenkameraden

478 neun- bis zwölfjährigen Schulkinder in Michigan, USA
warum ist man bei seinen Klassenkameraden beliebt

	Geschlecht		
Grund für Beliebtheit	weiblich	männlich	gesamt
gute Noten	55	39	94
gute Leistungen im Sport	38	127	165
gutes Aussehen	141	44	185
viel Taschengeld	17	17	34
gesamt	251	227	478

von insgesamt 251 Mädchen gaben 55 "Gute Noten" als wichtigsten Grund für Beliebheit an

Kontingenztafeln (Kreuztabellen, Kreuzklassifikation)

entstehen durch Aufteilen der Häufigkeiten einer Variable nach den Kategorien einer zweiten Variable

55 wird Zelle genannt

94
165
185
34

bzw 251 227 heissen Ränder (*margins*)

beschreiben Häufigkeiten jeweils nur einer Variable
(ergeben sich durch Summieren über Zeilen bzw. Spalten)

ARTEN VON INFORMATION

Beispiel: (Neben)Job und Fahrzeugbesitz bei Studenten

- Arbeiten Sie neben dem Studium ?
- Besitzen Sie ein Fahrzeug ?

Anordnung der Daten in einer Tabelle:

	Frage 1		
Frage 2	Ja	Nein	Gesamt
Ja	a	b	a+b
Nein	c	d	c+d
Gesamt	a+c	b+d	a+b+c+d =n

	Frage 1		
Frage 2	Ja	Nein	Gesamt
Ja			
Nein			
Gesamt			

- Was läßt sich mit der vorhandenen Information in der 2. Tabelle schon ausfüllen ?
- Welche Information benötigen wir zusätzlich, um den Rest auszufüllen?

WELCHE INFORMATION STECKT IN DER TABELLE:

- **Gemeinsame Information ("joint" information)**

Was sagt uns a / n ?
Was sagt uns c / n ?

wenn man gemeinsame Anteile von beide Variablen auf die Gesamtzahl beziehen will
("Gesamtprozent")

z.B. Wie hoch ist der Prozentsatz der Studenten die nicht arbeiten **und** ein Fahrzeug haben ?

- **Rand-Information ("marginal" information)**

Was sagt uns (a + c) / n ?
Was sagt uns (a + b) / n ?

(siehe Kapitel 2: Ein kategoriales Merkmal)

Bedingte Information ("conditional" information)

Was sagt uns a / (a + c) ?
Was sagt uns b / (b + d) ?

wenn man gemeinsame Anteile nur auf eine Kategorie einer der beiden Variable beziehen will
("Zeilenprozent" od. "Spaltenprozent")

z.B. Wie hoch ist der Prozentsatz der Studenten die ein Fahrzeug haben, **von denen**, die nicht arbeiten ?

DARSTELLUNG VON GEMEINSAMER KATEGORIALER INFORMATION

meist verwendet, wenn beide Variablen an einer Beobachtungseinheit erhoben wurden oder gleiche Kategorien haben

z.B. Beruf von Vater und Sohn, Sichtigkeit des linken und rechten Auges, Zustand vor und nach einer Therapie

Beispiel: Körpergröße bei 205 Ehepaaren

Häufigkeiten:

Anzahl

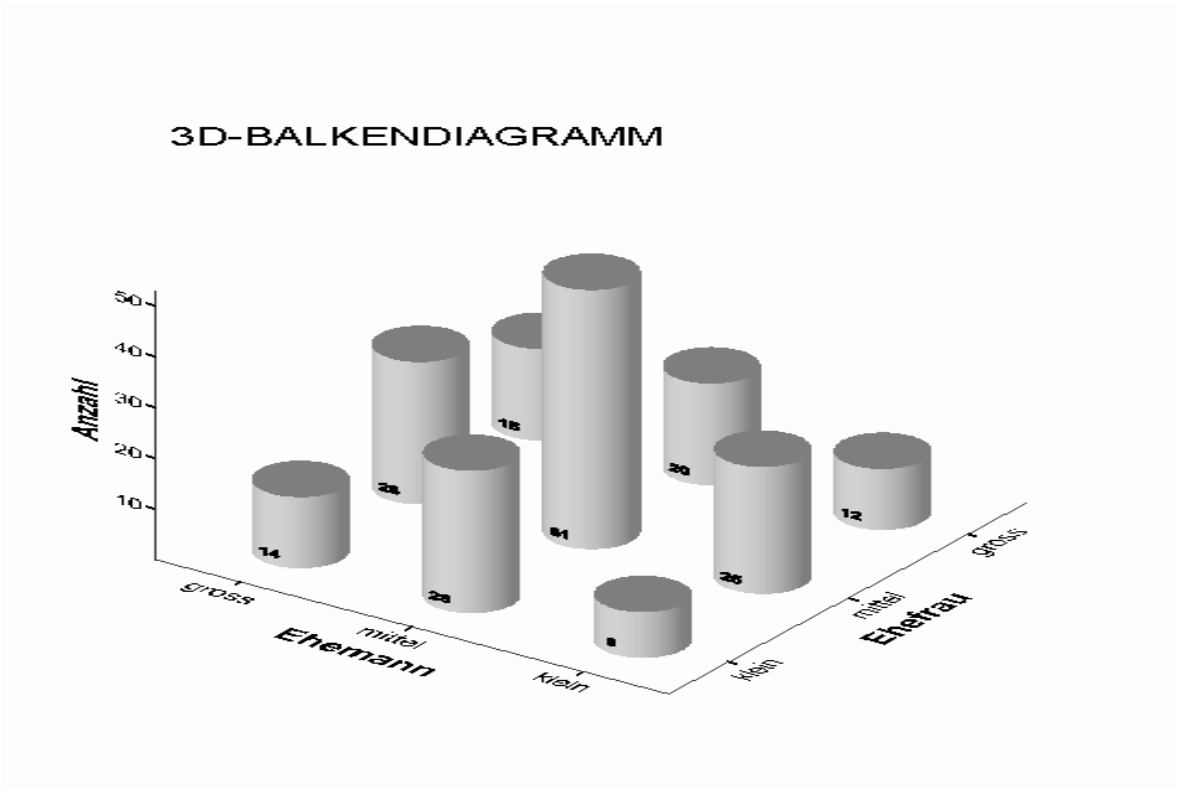
		Ehefrau			
		gross	mittel	klein	Gesamt
Ehemann	gross	18	28	14	60
	mittel	20	51	28	99
	klein	12	25	9	46
Gesamt		50	104	51	205

Gesamtprozent:

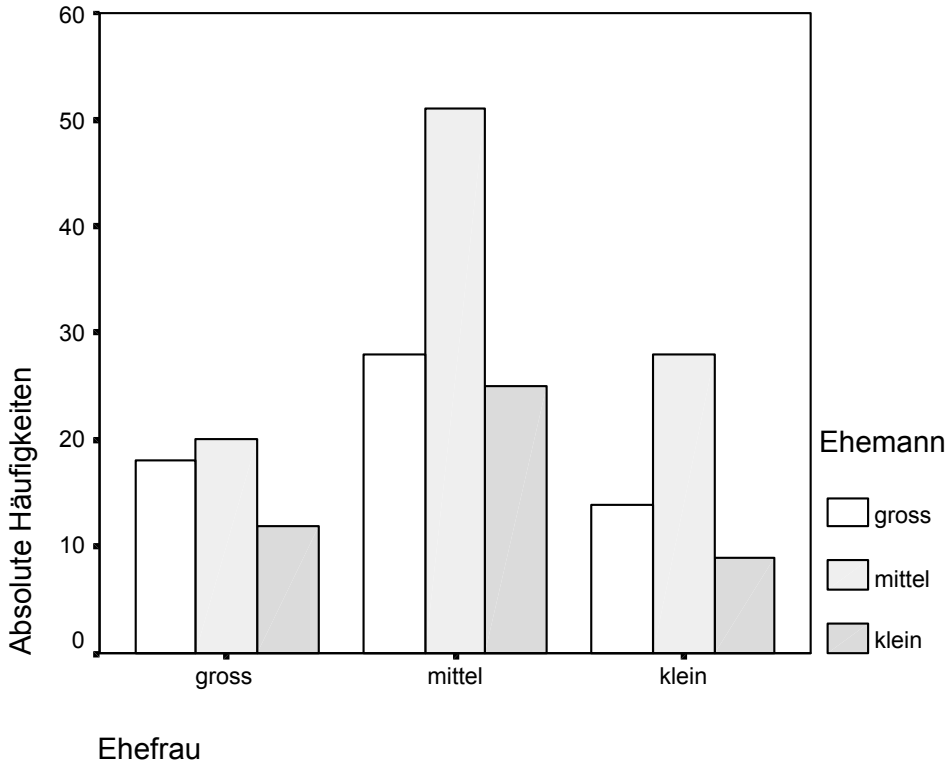
% der Gesamtzahl

		Ehefrau			
		gross	mittel	klein	Gesamt
Ehemann	gross	8,8%	13,7%	6,8%	29,3%
	mittel	9,8%	24,9%	13,7%	48,3%
	klein	5,9%	12,2%	4,4%	22,4%
Gesamt		24,4%	50,7%	24,9%	100,0%

GRAPHISCHE DARSTELLUNG VON GEMEINSAMER KATEGORIALER INFORMATION



GRUPPIERTES BALKENDIAGRAMM



DARSTELLUNG VON BEDINGTER KATEGORIALER INFORMATION

meist verwendet, wenn eine Variablen bei verschiedenen Gruppen beobachtet wurde

z.B. Einkommensklassen und Beruf, Sichtigkeit bei jungen und älteren Menschen, Zustand nach Therapie A oder B

Beispiel: Einschätzung des eigenen Gewichts bei Teenagern

Häufigkeiten

Anzahl

		GEWICHT			
		zu hoch	zu niedrig	gerade richtig	Gesamt
Geschlecht	männlich	42	65	224	331
	weiblich	121	12	212	345
Gesamt		163	77	436	676

Zeilenprozent

% von Geschlecht

		GEWICHT			
		zu hoch	zu niedrig	gerade richtig	Gesamt
Geschlecht	männlich	12,7%	19,6%	67,7%	100,0%
	weiblich	35,1%	3,5%	61,4%	100,0%
Gesamt		24,1%	11,4%	64,5%	100,0%

Spaltenprozent

% von GEWICHT

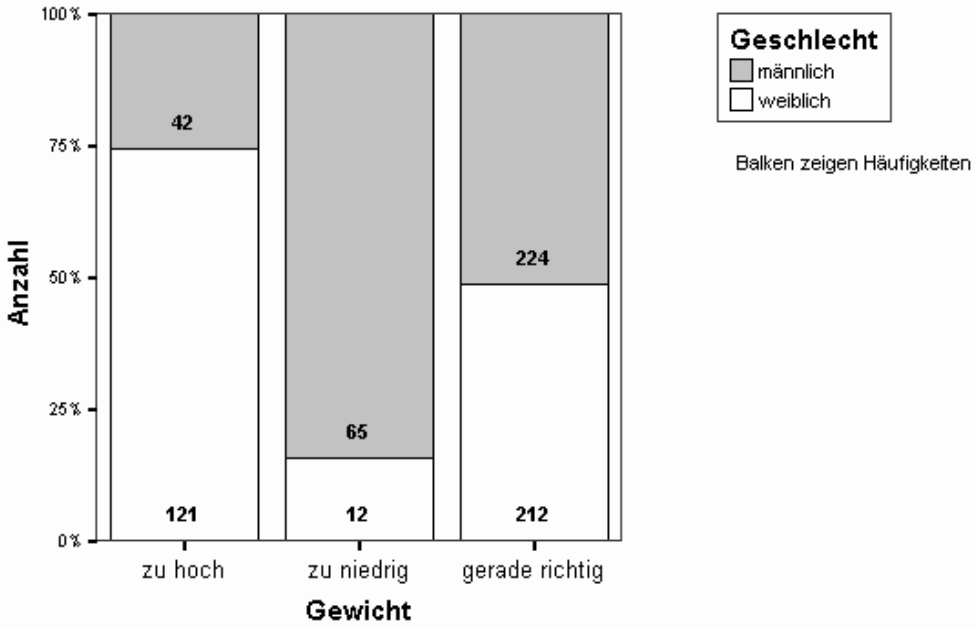
		GEWICHT			
		zu hoch	zu niedrig	gerade richtig	Gesamt
Geschlecht	männlich	25,8%	84,4%	51,4%	49,0%
	weiblich	74,2%	15,6%	48,6%	51,0%
Gesamt		100,0%	100,0%	100,0%	100,0%

Verwendung von Zeilen- oder Spaltenprozent je nach Fragestellung

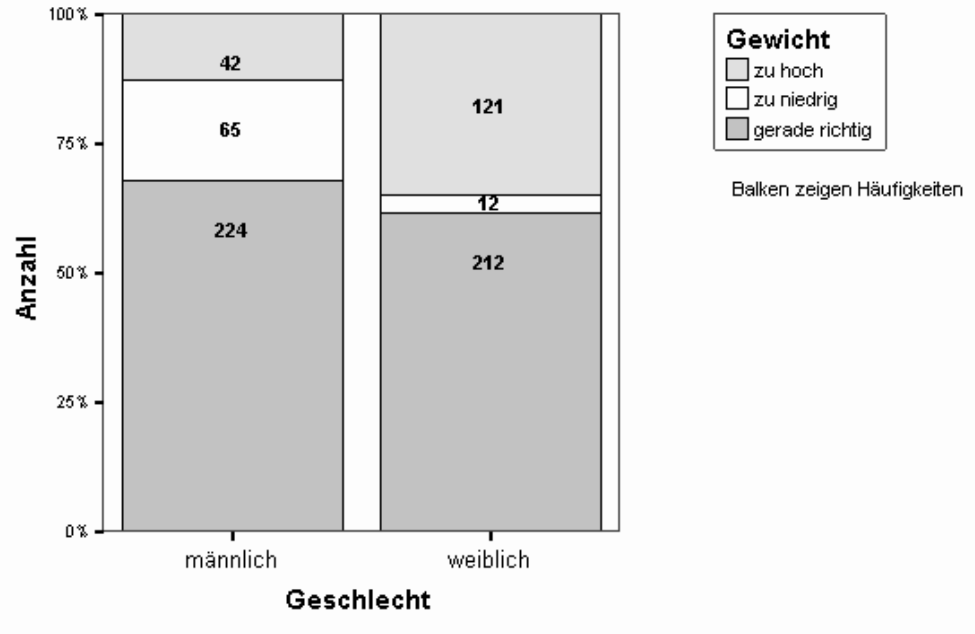
GRAPHISCHE DARSTELLUNG VON BEDINGTER KATEGORIALER INFORMATION

GESTAPELTE BALKENDIAGRAMME

nach Gewicht:



nach Geschlecht:



Beziehung zwischen zwei kategorialen Variablen

Zwei kategoriale Variablen stehen dann miteinander in Beziehung, wenn das Wissen über die Ausprägung einer Variable hilft, die Ausprägung der anderen Variable vorherzusagen

z.B. wenn ich weiß, daß ein bestimmter Student ein Fahrzeug besitzt, dann ist die Wahrscheinlichkeit höher, daß er (nebenbei) arbeitet

Welche der drei Informationsarten (Rand-, gemeinsame oder bedingte Information)sagt uns am meisten über eine mögliche Beziehung zwischen den beiden Variablen ?

am besten sieht man Beziehung an odds ratios

ODDS (dt."Chance")

ist das Verhältnis der Häufigkeiten von 2 Kategorien
("Chance" ist nicht das gleiche wie "Wahrscheinlichkeit" !)

Bsp:
die Chance, daß ein Student ein Fahrzeug besitzt ist

(a+b) / (c+d) bzw. (a+b) **:** (c+d)

die Chance, daß ein Student, der arbeitet, ein Fahrzeug hat ist

a / c bzw. a **:** c

die Chance, daß ein Student, der nicht arbeitet, ein Fahrzeug hat, ist

b / d bzw. b **:** d

ODDS RATIO (dt. „Verhältnis von Chancen“):

(ist nur für jeweils 2 Kategorien von 2 Variablen definiert)

Maß für Stärke der Beziehung zwischen 2 Variablen:

a	b
c	d

$$\text{odds ratio} = \frac{\frac{a}{c}}{\frac{b}{d}} \text{ bzw. } \frac{a \cdot d}{c \cdot b}$$

Bsp.:
die Chance, daß Studenten ein Fahrzeug haben
ist Mal höher für solche die arbeiten, als für
solche die nicht arbeiten

Eigenschaften von odds ratios:

wenn:

odds ratio ≈ 1: $\Rightarrow$ gleiche Chancen, oder
keine Beziehung zwischen den
Variablen

odds ratio ≠ 1: $\Rightarrow$ ungleiche Chance, bzw.
es könnte eine Beziehung zwischen
den beiden Variablen bestehen

Frage (wie bei X^2): ab wann ist odds ratio ≠ 1, d.h. wie sehr muß odds ratio in Stichprobe von 1 abweichen, so dass man auch für Population annimmt, dass es ungleich 1 ist (später)

ARTEN VON BEZIEHUNGEN ZWISCHEN 2 KATEGORIALEN MERKMALEN

- ***Homogenität***

bedeutet: die **Verteilung** der Häufigkeiten nach den Kategorien einer Variable ist bei 2 oder mehreren Gruppen **gleich**

wenn man eine kategoriale Variable an **2 oder mehreren Stichproben** erhoben hat und nach **Unterschieden** fragt

ähnlich wie bei Verwendung bedingter Information

Beispiel: unterscheiden sich Burschen und Mädchen in der Einschätzung ihres Gewichts

man unterscheidet die beiden Merkmale in:

WENN-Variable: erklärte (abhängige) Variable (Responsevariable)
DANN-Variable: erklärende (unabhängige) Variable

WENN	$\rightarrow$	DANN
erklärend	$\rightarrow$	erklärt
(unabhängig)	$\rightarrow$	(abhängig von WENN)

Was ist in folgenden Beispielen die abhängige (Y) und die unabhängige (X) Variable ?

- Das Verkehrsministerium möchte das Verhältnis zwischen *Straßenunebenheiten* und *Benzinverbrauch* untersuchen.

- Ein Händler, der seine Waren bei Fußballspielen verkauft, möchte die *Verkaufszahlen* auf die *Anzahl von Siegen* des Heimteams beziehen.

- Ein Psychologe möchte die *Einschätzung ihres Gewichts* auf Unterschiede zwischen *Burschen und Mädchen* untersuchen

Unabhängigkeit – Abhängigkeit

wenn man untersucht, ob **zwei Merkmale** voneinander **unabhängig** sind, bzw. ob zwei Variablen etwas miteinander zu tun haben (eine Beziehung besteht)

ohne WENN → DANN Beziehung oder wenn man die WENN → DANN Beziehung auch (sinnvollerweise) umdrehen kann

ähnlich wie Verwendung gemeinsamer Information

Beispiel: ist Körpergröße (klein / mittel / groß) bei Ehepartnern voneinander unabhängig

weitere Arten von Beziehungen
(werden bei metrischen Daten genauer behandelt)

- ***Assoziation (Zusammenhang)***

bei 2 Variablen mit geordneten Kategorien
Verwendung bei Fragestellungen mit: ***je - desto***
man möchte abschätzen, wie stark der Zusammenhang ist

Beispiel: je größer eine Ehefrau, desto größer ihr Ehemann ?

- ***Symmetrie***

meist bei wiederholten Messungen

Beispiel:
Personen werden nach Parteipräferenz gefragt, nach einem Monat wieder
Wechseln gleich viele Personen von Partei A zu Partei B wie umgekehrt von Partei B zu Partei A ?
Oder gibt es einen viel stärkeren Wechsel von A → B als von B → A

WICHTIGE FRAGESTELLUNGEN
(BEI ZWEI KATEGORIALEN MERKMALEN)

- Ist die Verteilung von Häufigkeiten in verschiedenen Gruppen gleich?

Diese Frage stellt man vor allem dann, wenn man wissen möchte, ob sich die Häufigkeiten in einzelnen Kategorien zwischen verschiedenen Gruppen unterscheiden

Beispiel: Showmaster-Werbung:
Wählen die Kinder, nachdem sie Showmaster-Werbung für CC gesehen haben, eher dieses Produkt, als wenn sie normale Werbung gesehen haben ?

- Sind zwei Variablen voneinander unabhängig ?

Diese Frage stellt man, wenn man wissen möchte, ob zwei Variablen voneinander abhängig sind.

Beispiel: Sind Verkaufsentwicklungen (Zu- / Abnahme) am Inlandsmarkt und am Auslandsmarkt voneinander unabhängig?

- In welchem Bereich kann man Chancen erwarten ? (Unterscheiden sich Chancen ?)

Diese Frage stellt man, wenn man abschätzen möchte, wie groß die Schwankungen von Chancen sein können, wenn man Daten nur aus Stichproben hat.

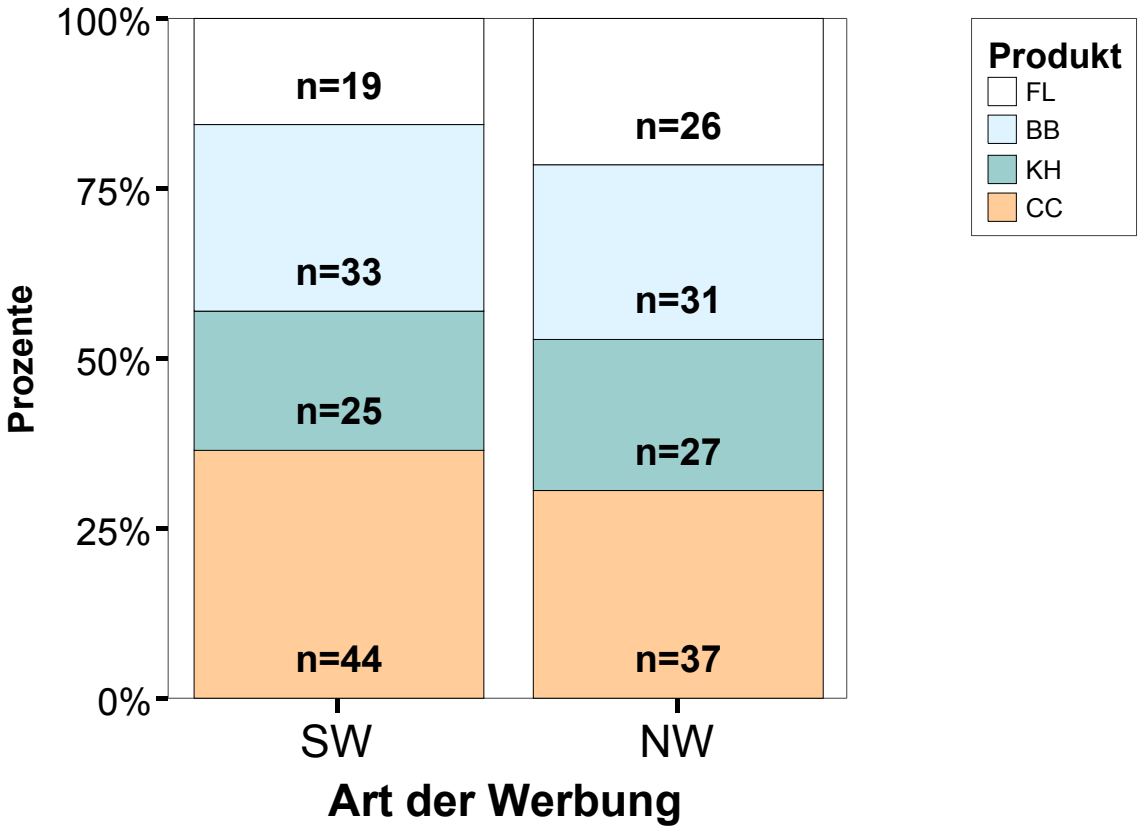
Beispiel: Ist die Chance höher mit Medikament A als mit Medikament B Schmerzen zu verringern?

FRAGESTELLUNG 1:

Ist die Verteilung von Häufigkeiten in verschiedenen Gruppen gleich ?

Bsp.: Präsentatorenwerbung und normale Werbung

Wird das Produkt CC häufiger dann gewählt, wenn ein Kind zuvor Showmasterwerbung gesehen hat, als wenn es normale Werbung gesehen hat (oder sind die Unterschiede in den beobachteten Häufigkeiten nur zufällig) ?



MASSZAHL FÜR DIE BEZIEHUNG ZWEIER KATEGORIALER VARIABLEN

Pearson´s X^2
auch andere Maße - X^2 ist das Wichtigste!

2 BEISPIELE:

politische Einstellung	positiv zum EURO	
	Ja	*Nein*
liberal	17	5
konservativ	8	23

→ starker Unterschied

	Schmerzlinderung	
	gering	stark
Medikament A	22	42
Medikament B	19	40

→ geringer Unterschied

wie lässt sich Maßzahl konstruieren, die die STÄRKE der BEZIEHUNG beschreibt?

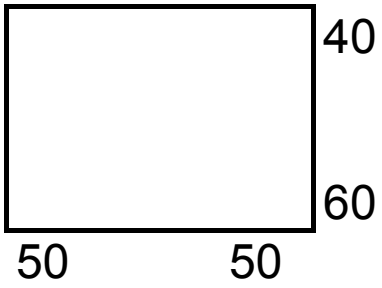
Konzept der beobachteten und erwarteten Häufigkeiten bei zwei kategorialen Merkmalen:

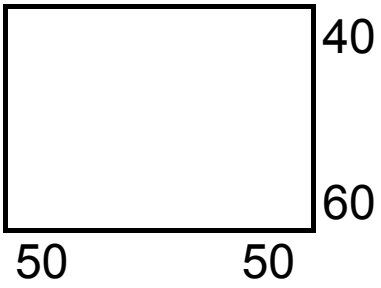
beobachtete Häufigkeiten:
jene Zahlen in der Tabelle, die beobachtet wurden

erwartete Häufigkeiten:
welche Zahlen, bei gegebenen Randsummen, würde ich erwarten, wenn keine Beziehung zwischen den beiden Variablen besteht

Beispiel:

Überlegen Sie sich bitte, wie die Zahlen in der Tabelle aussehen würden, wenn KEINE bzw. wenn eine STARKE BEZIEHUNG besteht





KEINE BEZIEHUNG STARKE BEZIEHUNG

Idee:
Maßzahl soll so konstruiert werden, dass sie umso größer wird, je stärker beobachtete und erwartete Häufigkeiten (wenn keine Beziehung besteht) voneinander abweichen.

Nebenfrage:
wie viele Zahlen in obigen Beispiel konnten Sie frei wählen, ohne dass dadurch die anderen festgelegt waren (Freiheitsgrade)

Berechnung der erwarteten Werte:

Beispiel:

- 0,3 ist Anteil der 1. Zeile an allen Beobachtungen (30/100)
- dieser teilt sich 40 zu 60 auf
- daher erwartete Werte für 1. Zeile: 40 x 0,3 = 12
  60 x 0,3 = 18
- usw. für Zeilen 2 und 3

oder:

z.B. 1. Zeile und 1. Spalte:

0,3 als Anteil ergibt sich aus 30/100

daher Berechnung: $\frac{40 \cdot 30}{100} = 12$

multiplizieren der Ränder und dividieren durch Gesamtanzahl

allgemein:

Tabelle der beobachteten Häufigkeiten

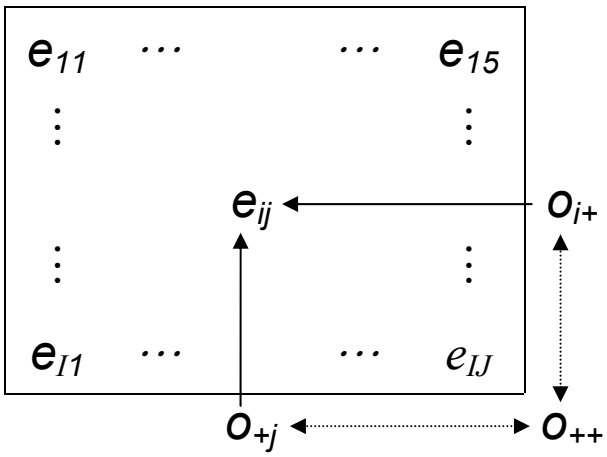
o_{11}	$\cdots$	o_{1j}	$\cdots$	o_{1J}	o_{1+}
$\vdots$		$\vdots$		$\vdots$	$\vdots$
o_{i1}	$\cdots$	o_{ij}	$\cdots$	o_{iJ}	o_{i+}
$\vdots$		$\vdots$		$\vdots$	$\vdots$
o_{I1}	$\cdots$	o_{Ij}	$\cdots$	o_{IJ}	o_{I+}
o_{+1}		o_{+j}		o_{+J}	o_{++}

o_{ij} ... beobachtete Häufigkeit für *i* - te Zeile und *j* - te Spalte

\+ ... Summe über Zeile bzw. Spalte

o_{++} ...Größe der Stichprobe ($o_{++} = n$)

Tabelle der erwarteten Häufigkeiten: e_{ij}



$$e_{ij} = \frac{o_{i+} \cdot o_{+j}}{o_{++}}$$

Pearson's X^2 für die Beziehung zwischen 2 kategorialen Variablen ergibt sich aus:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \quad \text{bzw.} \quad X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{o_{ij}^2}{e_{ij}} - o_{++}$$

Beobachtete Häufigkeiten

Anzahl

		Gewähltes Produkt				
		FL	BB	KH	CC	Gesamt
Art der Werbung	SW	19	33	25	44	121
	NW	26	31	27	37	121
Gesamt		45	64	52	81	242

Erwartete Häufigkeiten

Erwartete Anzahl

		Gewähltes Produkt				
		FL	BB	KH	CC	Gesamt
Art der Werbung	SW	22,5	32,0	26,0	40,5	121,0
	NW	22,5	32,0	26,0	40,5	121,0
Gesamt		45,0	64,0	52,0	81,0	242,0

z.B. die erwartete Häufigkeit für SW/FL ergibt sich aus:

$$\frac{o_{1+} \cdot o_{+1}}{n} = \frac{121 \cdot 45}{242} = 22{,}5$$

einsetzen in Formel für Pearson X^2 ergibt $X^2 = 1{,}833$
(je größer X^2, umso stärker ist die Beziehung)

Freiheitsgrade:

bei zweidimensionalen Kontingenztafeln:

$$\boxed{df = (I-1)\cdot(J-1)}$$

"Anzahl der Zeilen minus 1 mal Anzahl der Spalten minus 1"

(d.h. eine Zeile und eine Spalte streichen - Anzahl der Zellen, die übrigbleiben, sind Freiheitsgrade)

die Fragestellung lautete:

Wird das Produkt CC häufiger dann gewählt, wenn ein Kind zuvor Showmasterwerbung gesehen hat, als wenn es normale Werbung gesehen hat ?

ist eine Frage nach Homogenität:
Wenn (Art der Werbung) $\rightarrow$ Dann (gewähltes Produkt)

Nullhypothese (H_0):
Verteilung der gewählten Produkte ist für Showmastergruppe und Gruppe mit normaler Werbung gleich

Alternativhypothese (H_A):
Verteilung der gewählten Produkte ist für Showmastergruppe und Gruppe mit normaler Werbung nicht gleich

händische Berechnung (wieder mittels χ^2 –Tabelle):

1. suchen des kritischen χ^2 - Werts

- bestimmen der Freiheitsgrade *df (degrees of freedom)*
  in unserem Beispiel: df = (4 – 1)x(2 - 1) = 3

- nachsehen in Zeile mit Freiheitsgraden und in Spalte mit 0.950
  kritischer Wert ist in unserem Beispiel 7.815

2. Vergleich des X^2 - Werts mit dem kritischen χ^2 - Wert:

X^2Wert ist 1,833 bei df = 3, der kritische Wert ist χ^2_{krit} = 7,81

der X^2Wert ist kleiner als der kritische Wert (liegt noch unter der kritischen Schranke), daher Nullhypothese beibehalten

3. Interpretation (siehe nächste Seite)

Berechnung mit Statistikprogrammen:

aufsummieren über alle Zellen ergibt $X^2 = 1{,}833$ (bei df = 3)

SPSS

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	1,833	3	,608

R

```
Pearson's Chi-squared test

data: CEREAL and GROUP
X-squared = 1.8333, df = 3, p-value = 0.6077
```

Resultat:

- der p-Wert mit 0,608 ist deutlich größer als 0,05, daher wird die Nullhypothese beibehalten
- der Unterschied in der Verteilung der Produktwahl zwischen beiden Gruppen ist nicht signifikant
- Kinder die Showmaster Werbung gesehen haben, wählen nicht häufiger CC als jene, die normale Werbung gesehen haben

FRAGESTELLUNG 2:

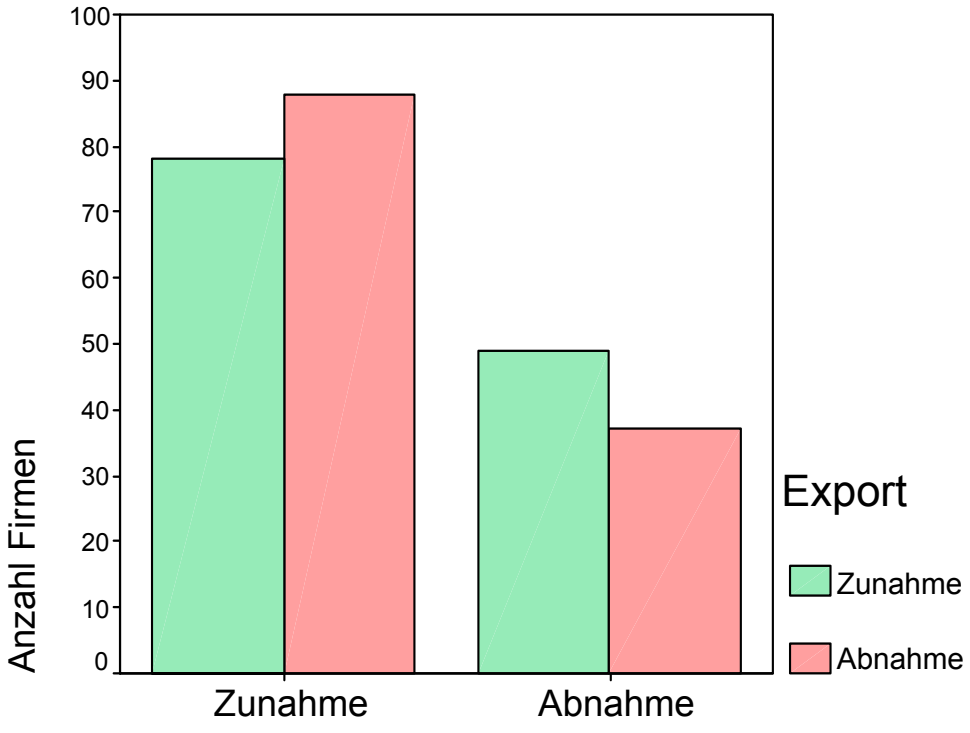
sind zwei kategoriale Merkmale voneinander unabhängig ?

Beispiel: Verkaufsentwicklungen im In- und Ausland

Sind Verkaufsentwicklungen am Inlandsmarkt und am Auslandsmarkt voneinander unabhängig ?

---

252 Firmen wurden über ihre Einschätzung bezüglich der Verkaufsentwicklungen am nationalen bzw. Export-Markt befragt. Sie sollten angeben, ob sie jeweils eine Zunahme oder Abnahme der Verkaufszahlen erwarten.



Inland

Berechnung:

Methoden sind gleich wie bei Fragestellung 1 $\Rightarrow$ χ^2-Test

Nullhypothese:
Inlands- und Auslandsverkaufsentwicklung sind voneinander unabhängig - „Inlandsoptimisten“ erwarten für den Export das Gleiche wie „Inlandspessimisten“

Alternativhypothese:
Inlands- und Auslandsverkaufsentwicklung sind nicht voneinander unabhängig - „Inlandsoptimisten“ erwarten für den Export etwas Anderes als „Inlandspessimisten“

SPSS

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (2-seitig)
Chi-Quadrat nach Pearson	2,261	1	,133

R

```
Pearson's Chi-squared test

data:  expsales
X-squared = 2.2611, df = 1, p-value = 0.1327
```

Resultat:

- der p-Wert mit 0,133 ist größer als 0,05, das Ergebnis ist nicht signifikant, daher wird die Nullhypothese beibehalten,
- Erwartungen über die Entwicklung von Inlands- und Auslandsverkäufen sind voneinander unabhängig

FRAGESTELLUNG 3:

In welchem Bereich kann man Chancen erwarten ? Unterscheiden sich Chancen ?

Bsp.: Chancen bei verschiedenen Prüfern

	Prüfer			
Prüfungsergebnis	A	B	C	gesamt
positiv	45	32	21	98
negativ	16	11	16	43
gesamt	61	43	37	141

Ist die Chance positiv abzuschneiden bei Prüfer A und B höher als bei Prüfer C ?

odds ratios: für Durchkommen

A im Vergleich zu C: $OR_{A/C} = (45/16)/(21/16) = 2,15$
B im Vergleich zu C: $OR_{B/C} = (32/11)/(21/16) = 2,22$
A im Vergleich zu B: $OR_{A/B} = (45/16)/(32/11) = 0,97$

Berechnung beruht auf Stichprobe von 141 Studenten
bei anderen Stichproben - andere Ergebnisse

Daten aus Stichprobe schwanken zufällig um tatsächliche Werte in Population (gleiche Idee wie bei Konfidenzintervallen für Anteile)

wie groß können diese Schwankungen sein ?

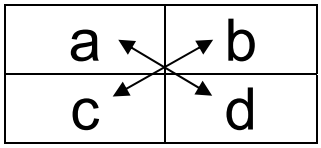
Schwankungsbreiten sind für logarithmierte odds ratios (kurz "log odds ratios") definiert

Bestimmung von "Schwankungsbreiten" für log odds ratios:

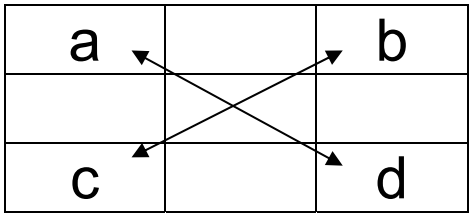
$$c = 1{,}96 \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

bei kleinen a,b,c oder d verwendet man (a+0.5), (b+0.5), ...

bei 2 x 2 Tafeln



bei r x c Tafeln:
(mehrere ORs möglich)
z.B.:



für Vergleich der Prüfer A mit C:

$$c = 1{,}96 \cdot \sqrt{\frac{1}{45{,}5} + \frac{1}{21{,}5} + \frac{1}{16{,}5} + \frac{1}{16{,}5}} = 1{,}96 \cdot 0{,}434 = 0{,}851$$

Konfidenzintervall für log odds ratios:

wieder: Berechung von zwei Grenzen so, dass es sehr plausibel ist, dass tatsächliches odds ratio in der Grundgesamtheit innerhalb dieser Grenzen liegt

$$\text{KI:} \quad \left[\ln OR - c \;\; ; \;\; \ln OR + c \right]$$

Bsp: für Prüfer A im Vergleich zu C:

[ln 2,15 - 0,85 ; ln 2,15 + 0,85] bzw. [-0,085 ; 1,616]

damit man wieder zur nicht-logarithmischen Skala zurückkehrt müssen die Intervallgrenzen exponentiert werden, d.h.

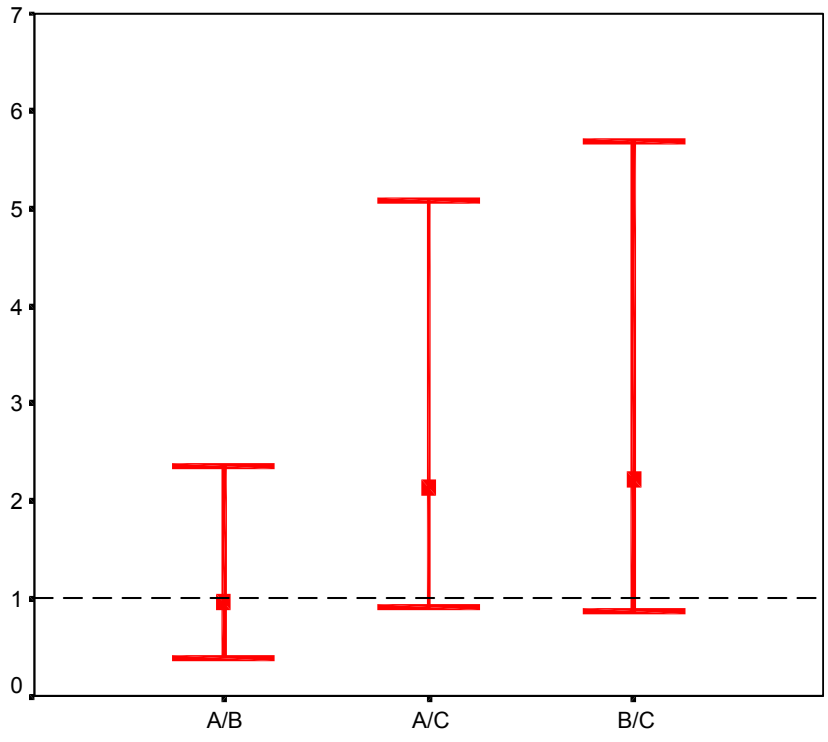
untere Grenze: $e^{-0,085} = 0,918$

obere Grenze: $e^{1,616} = 5,033$

Resultat: die Chance bei Prüfer A besser abzuschneiden ist zwischen 0,9 und 5 Mal höher als bei Prüfer C (mit 95% Sicherheit)

Anmerkung: da 0,9 kleiner und 5 größer als 1 ist, kann man nicht behaupten, dass man bei Prüfer A besser abschneidet als bei C, da 1 (gleiche Chance) im KI enthalten ist

Resultat für alle Prüfer:



KI für odds ratios:
A/B: [0,40 ; 2,36]
A/C: [0,92 ; 5,03]
B/C: [0,86 ; 5,70]