

UNIVERSITÄT LEIPZIG

Medizinische Fakultät

QSB1 – Tag 6

V 11 Statistisches Testen III – Power / Chi²-Test

Dr. rer. med. habil. Marita Ziepert /
Dr. rer. nat. Meinhard Mende,
ZKS & IMISE, Universität Leipzig

imise.

V 11

11.1 Einschub: Power von Tests

Lernziele

Nach Absolvierung dieser Vorlesung

- ist Ihnen der Terminus Power vertraut und Sie wissen, wie diese beeinflusst werden kann,
- verstehen Sie, wie der Chi²-Test funktioniert,
- können Sie den Chi²-Test rechnen, die Voraussetzung prüfen und das Testergebnis interpretieren.

Ausgänge beim Hypothesentest

– Optimale Tests?

Testentscheidung aufgrund Stichprobe	wahre Situation (Biologie)	
	H_0 richtig	H_A richtig
H_0 beibehalten	$1-\alpha$	β
H_0 verwerfen und H_A annehmen	α	$1-\beta$

Der Fehler II. Art ergibt sich aus den Daten (Szenarium,Variation, Effektstärke, Fallzahl)

→ Maximum

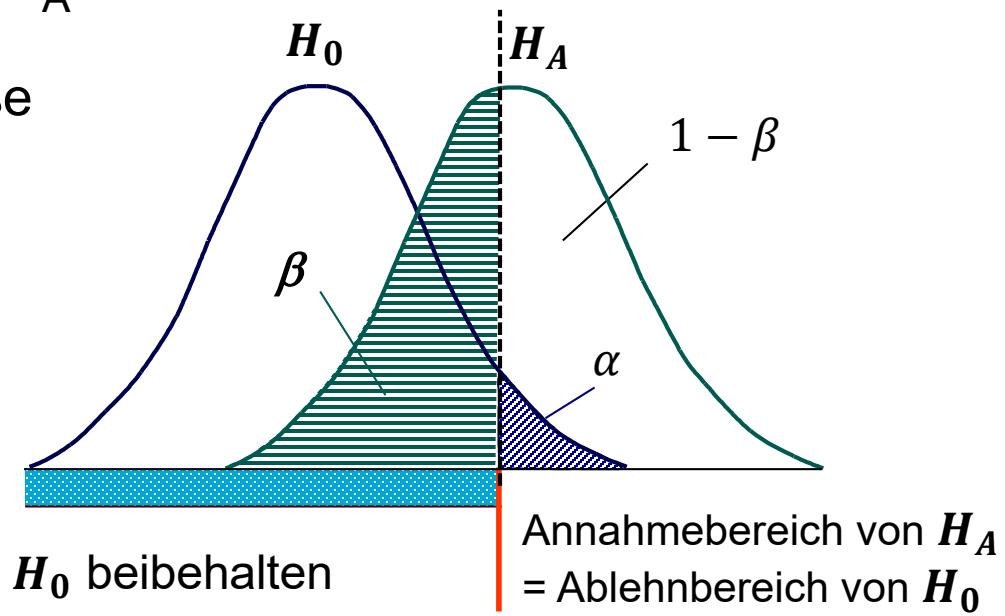
Der Test ist so konstruiert, dass das vorgegebene Signifikanz-niveau α eingehalten wird.

Die Power ergibt sich aus Algorithmus, Effektstärke, Variation und Fallzahl.

Optimale Tests?

- Null-Hypothese und eine speziell vorgegebene Alternativ-Hypothese legen die Verteilung der Prüfgröße fest
- hier: Punkthypothese für H_A

Verteilungen der Prüfgröße



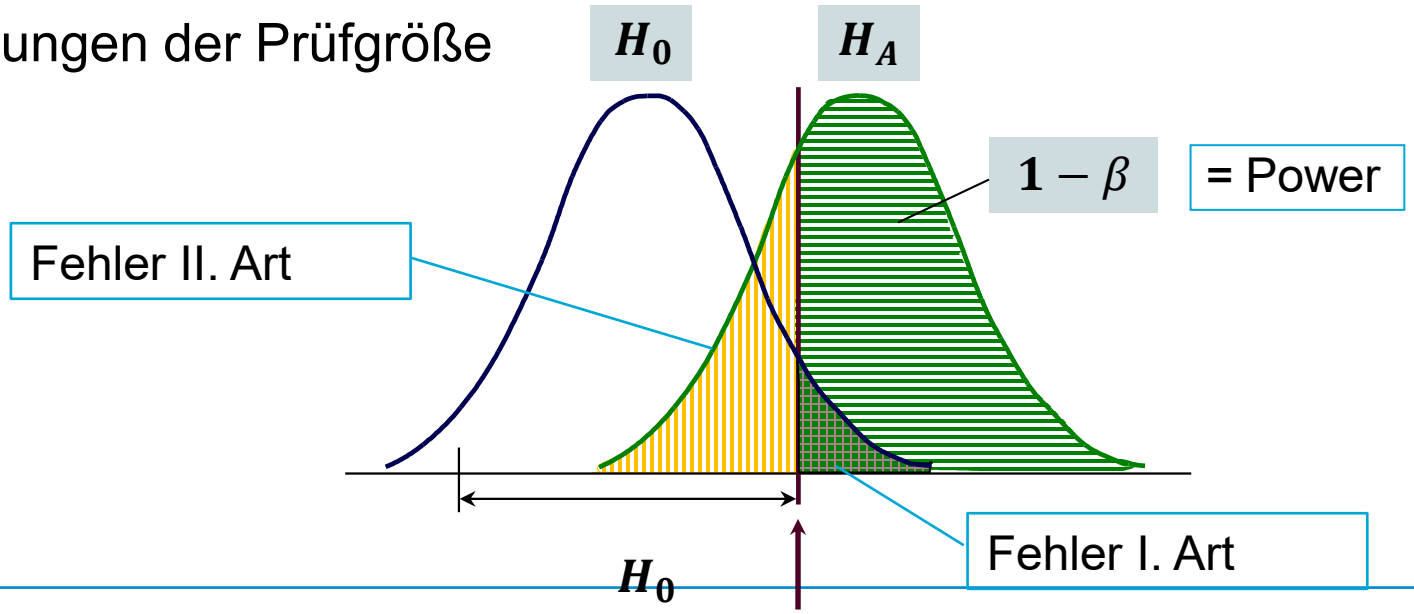
Power eines Tests I

– Power (Teststärke) des Hypothesentests:

Wahrscheinlichkeit, mit der H_0 abgelehnt wird, wenn H_A richtig ist:

$$1 - \beta = Power = P(H_0 \text{ ablehnen} \mid H_A \text{ richtig})$$

Verteilungen der Prüfgröße

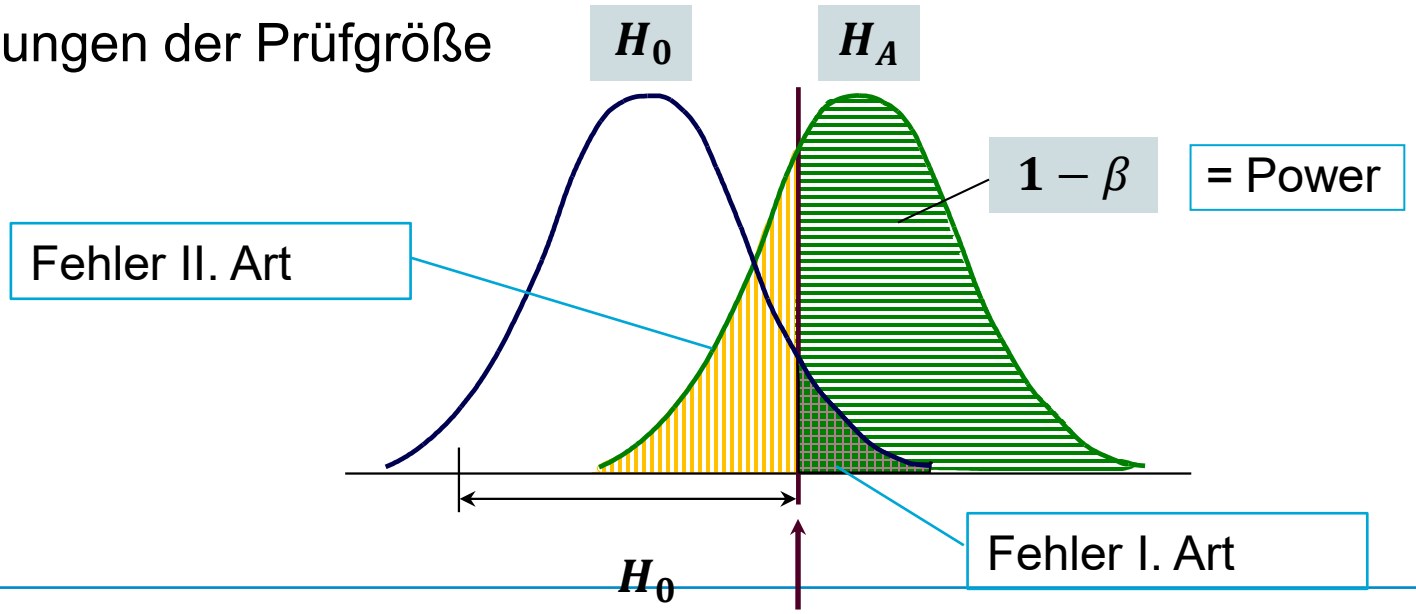


Power eines Tests II

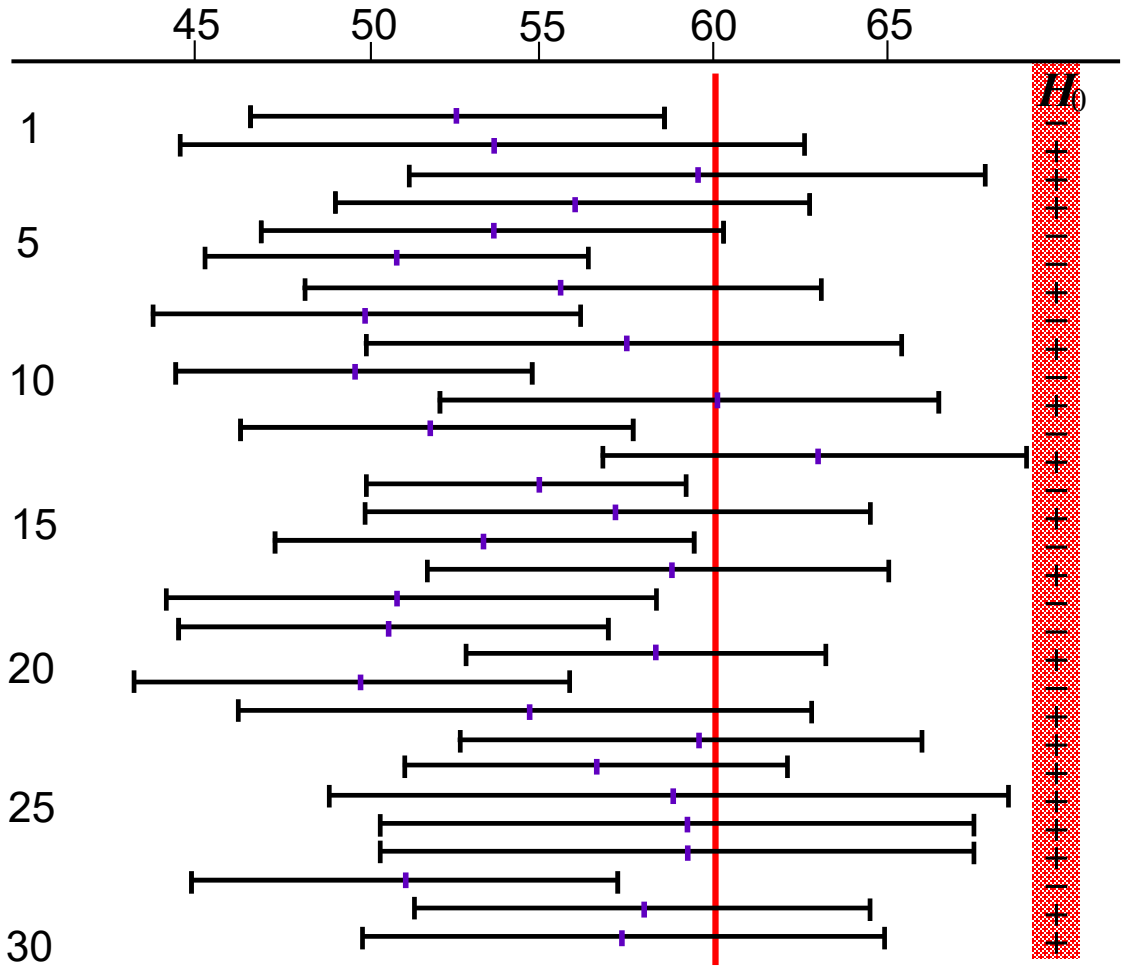
Ziel jedes Hypothesentests:

- ein ggf. vorhandener (Behandlungs-)Effekt soll mit maximaler Wsk. aufgedeckt werden:
- = maximale Power

Verteilungen der Prüfgröße



Power und Fallzahl I



Lipid-Screening:

HDL Frauen (n = 5529)

30 Stichproben zu je 20 Probanden,

H_0: $\mu = \mu_0 = 60$

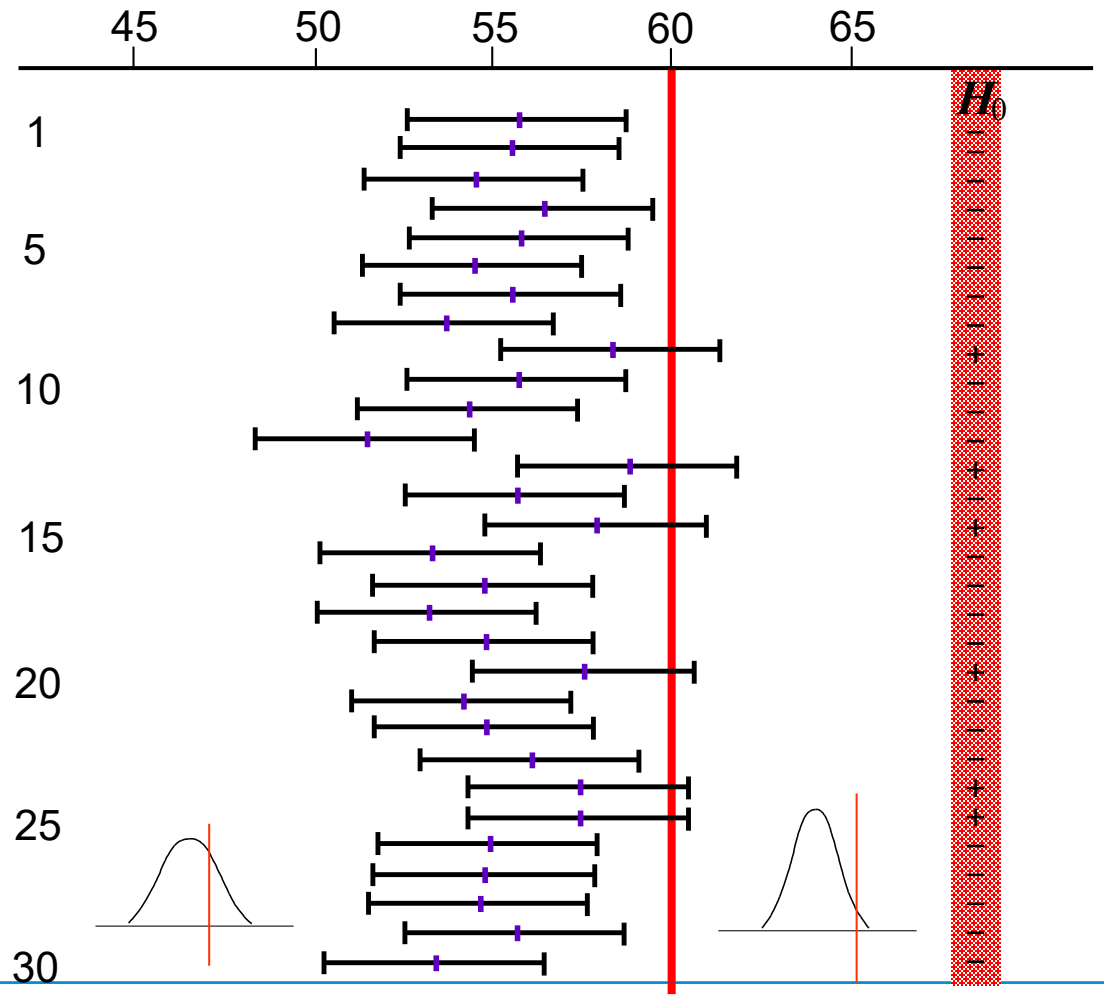
H_A: $\mu \neq 60$

$\alpha = 0{,}05$

12 / 30 Stichproben kommen zu einer korrekten Ablehnung von H_0

$\widehat{power} = 40\%$

Power und Fallzahl II



Lipid-Screening:

HDL Frauen (n = 5529)

30 Stichproben zu je 100 Probanden,

H_0: $\mu = \mu_0 = 60$

H_A: $\mu \neq 60$

$\alpha = 0{,}05$

24 / 30 Stichproben kommen zu einer korrekten Ablehnung von H_0

$\widehat{power} = 80\%$

Fallzahlplanung

Die Power eines Tests ist nur quantifizierbar,

- wenn man eine konkrete Alternativ-Hypothese spezifiziert.

- In der Versuchsplanung wird zu jeder Nullhypothese eine Alternativ-Hypothese $\boldsymbol{H_A}$ gewählt, um zwischen beiden zu entscheiden.
- Diese hat direkte Auswirkung auf die notwendige Fallzahl.
- Für jedes Szenario gibt es andere Formeln zur Fallzahlberechnung (sample size calculation).

Schaden durch Fehler beim Testen

Beurteilen Sie, welcher Fehler (I. oder II. Art) in folgenden Szenarien aus der Onkologie schlimmere Auswirkungen hat:

- Eine neue Substanz zur Krebsbekämpfung wird in toxikologischen Tests auf Nebenwirkungen untersucht (H_0: keine NW).

- Die Behandlung mit einer neuen Substanz zur Krebsbekämpfung wird mit Standardbehandlung verglichen (H_0: kein Unterschied im rezidivfreien Überleben).

Interventionseffekt - Power

Beispiele aus Studienprotokollen

- Alter-G-Studie: Indikation Knie- oder Knöchel-OP
  - Alter-G-Gerät vs. Standard-Care: Primärer EP KOOS-Score
  - $\Delta = 15\ \text{Punkte}$ bei einer Standardabweichung (SD) von 18 Punkten
  - $\alpha = 5\%, 1 - \beta = 80\%$
- HCT vs. CT in Elderly AML: primärer EP 5y survival
  - CT: 25%, HCT: 45%
  - $\alpha = 5\%, 1 - \beta = 90\%$
- CAT-PVC: primärer EP relative Abnahme ventrikulärer Extrasystolen (VES)
  - Amiodarone: 55%, Ablation: 75% bei SD = 25%
  - $\alpha = 5\%, 1 - \beta = 90\%$

V 11

11.2 Vergleich von Häufigkeiten

Ablaufschema für statistische Tests I

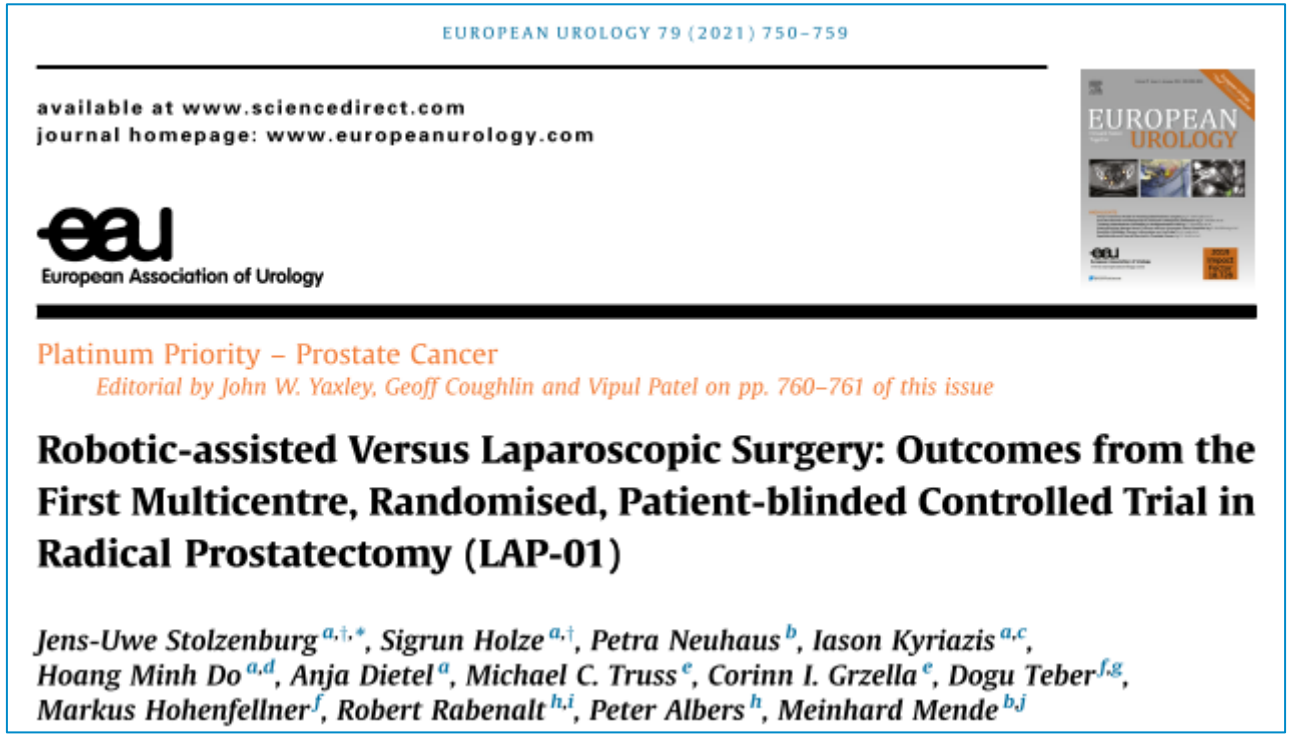
1. Festlegung der Zielsetzung (Was will man wissen?)
   - Fragestellung
   - Zielgröße
2. Stichproben (Versuchsplanung): Anzahl und Paarung
3. Formulierung der Hypothesen $\boldsymbol{H_0}$ und $\boldsymbol{H_A}$
4. Wahl des vorzugebenden Signifikanzniveaus α
5. Auswahl der Prüfgröße und der dazugehörigen Testverteilung
6. Prüfung von Testvoraussetzungen (z. B. Verteilungsannahmen
   - Anwendbarkeit (z.B. Varianzhomogenität, Asymptotik usw.)
   - ggf. Modifikation von 5.

Ablaufschema für statistische Tests II

7. Ermittlung des Ablehn- (bzw. Akzeptanz-)bereiches
8. Berechnung der Prüfgröße aus der Stichprobe
9. Testentscheidung
   - Vergleich der berechneten Prüfgröße mit dem Quantil der Prüfverteilung, oder
   - Angabe des *p*-Wertes: Vergleich mit α

   → H_0 ablehnen / H_A annehmen oder H_0 beibehalten
10. Interpretation

Beispiel 11.1: die LAP-01-Studie (→ B 3.1)



EUROPEAN UROLOGY 79 (2021) 750–759

available at www.sciencedirect.com
journal homepage: www.europeanurology.com

European Association of Urology

Platinum Priority – Prostate Cancer

Editorial by John W. Yaxley, Geoff Coughlin and Vipul Patel on pp. 760–761 of this issue

Robotic-assisted Versus Laparoscopic Surgery: Outcomes from the First Multicentre, Randomised, Patient-blinded Controlled Trial in Radical Prostatectomy (LAP-01)

Jens-Uwe Stolzenburg[a,†,], Sigrun Holze[a,†], Petra Neuhaus[b], Iason Kyriazis[a,c], Hoang Minh Do[a,d], Anja Dietel[a], Michael C. Truss[e], Corinn I. Grzella[e], Dogu Teber[f,g], Markus Hohenfellner[f], Robert Rabenalt[h,i], Peter Albers[h], Meinhard Mende[b,j]*

– erste multizentrische randomisierte Studie, welche laparoskopische radikale Prostatektomie ohne / mit Roboterunterstützung verglich

Beispiel 11.1

- Endpunkt: Kontinenz nach 12 Wochen (0 pads or 1 safety pad)
- (Erfassung durch Tagebuch über benötigte Pads)

		R-ARP	LRP	Total (ITT)
Continence 12 weeks	No	250 (45,7%)	93 (54,4%)	343
	Yes	297 (54,3%)	78 (45,6%)	375
Total		547	171	718

- Frage: Kann der Unterschied von 8,7% noch mit dem Zufall erklärt werden?
- Signifikanzniveau 5%

Szenario

- Studienarm 2 Ausprägungen (R-ARP / LRP)
- Endpunkt Kontinenz 2 Ausprägungen (nein / ja)

Rückblende (VL 8):

- Absolute und relative Häufigkeiten in einer $k \times m$- Kontingenztafel:

		Merkmal 2				Randsumme
		1	2	…	s	
Merkmal 1	1	b_{11}	b_{12}	…	b_{1s}	$b_{1.}$
	2	b_{21}	b_{22}	…	b_{2s}	$b_{2.}$
	…	…	…	…	…	…
	z	b_{z1}	b_{z2}	…	b_{zs}	$b_{z.}$
Randsumme		$b_{.1}$	$b_{.2}$	…	$b_{.s}$	n

Chi²-Test I

Hypothesen

a) Unabhängigkeit

- H_0: Studienarm und Erreichen der Kontinenz nach 12 Wochen sind unabhängig. voneinander.
- H_A: Das erreichen der Kontinenz hängt vom Studienarm ab.

b) Äquivalent dazu: Gleichheit von Wahrscheinlichkeiten:

- H_0: Der Anteil Patienten mit Kontinenz ist nach 12 Wo. in beiden Studienarmen gleich.
- H_A: Beide Arme unterscheiden sich im Anteil nach 12 Wochen erreichter Kontinenz.

Beide Hypothesenpaare führen auf mathematisch äquivalente Tests.

Chi²-Test II

Grundgedanke:

- Jede Merkmalskombination bedeutet eine UND-Verknüpfung zweier Ausprägungen beider Merkmale X und Y, z. B. im Beispiel LAP-01:

$$P(\text{Kontinenz im LRP-Arm}) = P(\text{Kontinenz} \cap \text{LRP-Arm})$$

- Unter H_0 (angenommener Unabhängigkeit) sind erwartete und beobachtete Häufigkeit gleich:
- Dann gilt die Produktregel und es lassen sich die Wahrscheinlichkeiten für Kontinenz bzw. Armzugehörigkeit multiplizieren.
- Je mehr beobachtete und erwartete Häufigkeiten voneinander abweichen, desto mehr spricht dies gegen H_0.

Chi²-Test III

		Merkmal 2		Randsumme
		1	2	
Merkmal 1	1	b_{11}	b_{12}	$b_{1.}$
	2	b_{21}	b_{22}	$b_{2.}$
Randsumme		$b_{.1}$	$b_{.2}$	n

- Bei Unabhängigkeit (H_0) ergibt sich die erwartete Häufigkeit in Zeile i / Spalte j:
- $e_{ij} = \frac{b_{i.}}{n} \cdot \frac{b_{.j}}{n} \cdot n$ bzw. $e_{ij} = \frac{b_{i.} \cdot b_{.j}}{n}$

z. B. Zelle (2, 2):

- $e_{22} = 0{,}238 \cdot 0{,}523 \cdot 718 = 89{,}4$

		R-ARP	LRP	Total (ITT)
Continence 12 weeks	No	250	93	343 (47,7%)
	Yes	297	78	375 (52,3%)
Total		76,2%	23,8%	718

- D. h., im LRP-Arm würden 89 Patienten mit erreichter Kontinenz erwartet.

Testidee

– Beobachtete und …

		R-ARP	LRP	Total (ITT)
Continence 12 weeks	No	250	93	343
	Yes	297	78	375
Total		547	171	718

erwartete Häufigkeiten

		R-ARP	LRP	Total (ITT)
Continence 12 weeks	No	261	82	343 (47,7%)
	Yes	287	89	375 (52,3%)
Total		76,2%	23,8%	718

– Prüfgröße = Abstand der beobachteten von den erwarteten Häufigkeiten:

$$X^2 = \sum_{i,j} \frac{(b_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(z-1)\cdot(s-1)}$$

(asymptotisch chi²-verteilt mit f Freiheitsgraden:

$$f = (z-1) \cdot (s-1))$$

Spezialfall: Vierfeldertafel

		Merkmal 2		Randsumme
		1	2	
Merkmal 1	1	b_{11}	b_{12}	$b_{1.}$
	2	b_{21}	b_{22}	$b_{2.}$
Randsumme		$b_{.1}$	$b_{.2}$	n

Vereinfachte Formel:

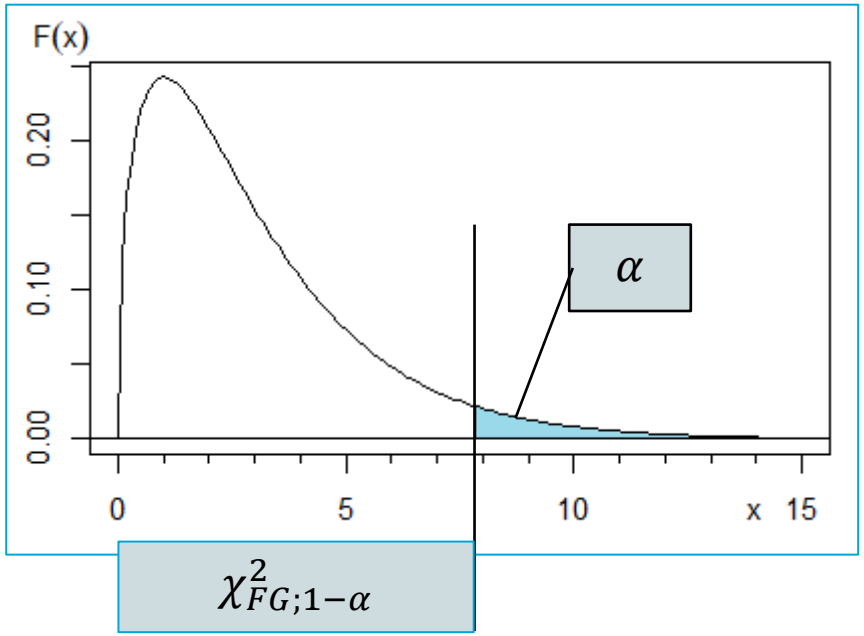
$$X^2 = \frac{n \cdot (a \cdot d - b \cdot c)^2}{(a+c) \cdot (b+d) \cdot (a+b) \cdot (c+d)} \sim \chi_1^2$$

hier: $X^2 = 3{,}94$

		Merkmal 2		Randsumme
		1	2	
Merkmal 1	1	a	b	a + b
	2	c	d	c + d
Randsumme		a + c	b + d	n

Quantile der χ^2-Verteilung

- Die χ²-Verteilung ist schief und hat nur nichtnegative Werte.
- Warum?
- Deshalb vergleicht man beim Chi²-Test zweiseitig mit dem $(1-\alpha)$-Quantil.



Testentscheidung

- Quantil: s. Skriptenheft, Tab. 11.6, S. 181:
- $\chi^2_{1;0,95} = 3{,}84$

Tabelle 11.6: *Quantile der χ^2-Verteilung*

FG	5 %	1 %	0,1 %	FG	5
1	3,84	6,63	10,83	51	68,
2	5,99	9,21	13,82	52	69,
3	7,81	11,34	16,27	53	70,
4	9,49	13,28	18,47	54	72,

Es gilt: $X^2 = 3{,}94 > 3{,}84 = \chi^2_{1;0,95}$

- Also …

Interpretation:

- Im R-ARP-Arm erreichen 8,7% mehr Patienten nach 12 Wochen Kontinenz,
- das ist …

Übungsaufgabe

- Wie viele Freiheitsgrade hat eine X^2-Statistik für eine 4 x 2-Tabelle?

- Lösung:

Chi²-Test – Voraussetzung I

Anwendbarkeit der Pearson'schen Teststatistik (nach Sachs, 1992):

Die Approximation der Teststatistik an die χ^2- Verteilung ist passabel,

- wenn n groß genug ist ($n \geq 40$) und alle Zeilen und Spalten ausreichend besetzt sind.
- Als *Faustregel* gilt für die erwarteten Häufigkeiten in den Feldern *(i, j)*:
  - Vierfeldertafel: $alle\ e_{ij} \geq 5$
  - Sechsfeldertafel: $alle\ e_{ij} \geq 3$
  - ≥ Achtfeldertafel: $alle\ e_{ij} \geq 1$
- Es gibt etwas liberalere Regeln (→ SPSS).

Chi²-Test – Voraussetzung II

Sind diese Voraussetzungen nicht erfüllt, kann man:

- Spalten und / oder Zeilen zusammenfassen oder
- den *Exakten Test von Fisher* anwenden:

- Dieser Test ist weder exakter noch besser als der Chi²-Test. (Der p-Wert ergibt sich als Summe der Wahrscheinlichkeiten der beobachteten Tabelle und aller unter H_0 extremeren.)
- Weil die p-Werte diskret verteilt sind, schöpft man meist das Signifikanz-niveau α oft nicht völlig aus; der Test ist (sehr) konservativ.
- Deshalb lehnen manche Autoren Fishers exakten Test generell ab (Power verschenken).

Beispiel 11.2: Diast-CHF

Unterscheidet sich die NYHA-Klasse zwischen Patienten mit / ohne Z.n. akutem Myokardinfarkt?

Signifikanzniveau: 1%

		Akuter Myokardinfarkt		
		nein	ja	Total
NYHA-Klasse	I	59	12	71
	II	107	27	134
	III	45	12	57
	IV	0	1	1
Total		211	52	263

SPSS-Ausgaben

- optional: erwartete Anzahlen
- Spaltenprozente

- 25% Zellen mit erwarteten Werten <5 → exakten Test nach Fisher auswerten:
- $p = 0{,}26 > 0{,}01$
- Unterschiede in den Anteilen beider Gruppen können zum 1%-Niveau nicht nachgewiesen werden.

NYHA-Klasse * Z. n. akutem MI Kreuztabelle

			Z. n. akutem MI nein	Z. n. akutem MI ja	Gesamt
NYHA-Klasse	I	Anzahl	59	12	71
		Erwartete Anzahl	57,0	14,0	71,0
		% von Z. n. akutem MI	28,0%	23,1%	27,0%
	II	Anzahl	107	27	134
		Erwartete Anzahl	107,5	26,5	134,0
		% von Z. n. akutem MI	50,7%	51,9%	51,0%
	III	Anzahl	45	12	57
		Erwartete Anzahl	45,7	11,3	57,0
		% von Z. n. akutem MI	21,3%	23,1%	21,7%
	IV	Anzahl	0	1	1
		Erwartete Anzahl	,8	,2	1,0
		% von Z. n. akutem MI	0,0%	1,9%	0,4%
Gesamt		Anzahl	211	52	263
		Erwartete Anzahl	211,0	52,0	263,0
		% von Z. n. akutem MI	100,0%	100,0%	100,0%

Chi-Quadrat-Tests

	Wert	df	Asymptotische Signifikanz (zweiseitig)	Exakte Sig. (zweiseitig)
Pearson-Chi-Quadrat	4,497[a]	3	,213	,263
Likelihood-Quotient	3,695	3	,296	,318
Exakter Test nach Fisher-Freeman-Halton	3,670			,328
Zusammenhang linear-mit-linear	,913[b]	1	,339	,382
Anzahl der gültigen Fälle	263			

a. 2 Zellen (25,0%) haben eine erwartete Häufigkeit kleiner 5. Die minimale erwartete Häu

b. Die standardisierte Statistik ist ,956.

McNemar-Test I

http://www.festartikelhaus.de/Deko-und-Gartenfiguren/Menschen/Oma-mit-Fernglas-K672::100684.html



Werden von denselben Probanden zwei Merkmale erhoben,

- sind die Beobachtungen *nicht unabhängig*.
- Dann ist der Chi² - Test nach Pearson nicht anwendbar.
- Für Vierfeldertafeln gibt es dafür den McNemar – Test

Beispiele:

- Schmerzen vor Operation (j / n) vs. Schmerzen nach Operation (j / n)
- Beurteilung histologischer Befunde (malign vs. non-malign) von zwei Gutachtern usw.

McNemar-Test II

Beispiel: Zwei Pathologen beurteilen 100 histologische Präparate auf Malignität (Signifikanzniveau 5%):

Geben die Daten Hinweise (*evidence*) darauf, dass die Pathologen die Präparate unterschiedlich beurteilen?

	Beurteiler 1		
Beurteiler 2	malign	benign	Summe
malign	58	5	63
benign	7	30	37
Summe	65	35	100

- H_0: Die Beurteilungen sind identisch.
- H_A: Die Beurteilungen unterscheiden sich.
- $X^2_{McNemar} = \frac{(b-c)^2}{b+c} \sim \chi^2_1$
- $X^2_{McNemar} = \frac{(5-7)^2}{5+7} = \frac{4}{12} = 0{,}333 < 3{,}84$

Zum 5%-Niveau kein Hinweis auf unterschiedliche Bewertung.

Zusammenfassung

- Häufigkeiten zweier Gruppen stellt man in einer Kontingenztafel dar.
- Mit dem Chi²-Test prüft man die Gleichheit von Proportionen (H_0) bzw. die Unabhängigkeit zweier kategorialer Größen.
- Die X^2-Statistik ist angenähert Chi²-verteilt mit $(z-1)\cdot(s-1)$ Freiheitsgraden.
- Sind die Bedingungen für den asymptotischen Test nicht erfüllt, bietet Fisher's exakter Test eine Alternative.