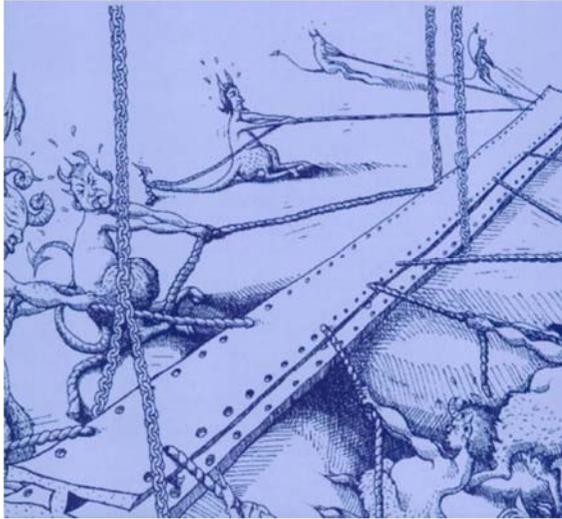


Regression und Korrelation

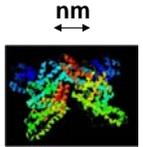


regression:
Zurückführung,
Rückschreiten

correlation:
Wechselbeziehung

KAD 2015.02.24

Praktische Annäherung (Beispiel1)



1 St. HSA Molekül

wieviele Eiweissmoleküle sind in dem Blutplasma?
(Stück, mol, g, ...)

wie gross ist die Eiweisskonzentration
des Blutplasmas? (St/L, mol/L, g/L)

bei Patienten in Nephrose (schwere Nierenkrankheit) nimmt der Wert stark ab

direkte Methode:

Bestimmung der Anzahl der Moleküle in einem Volumen(?)

indirekte Methode :

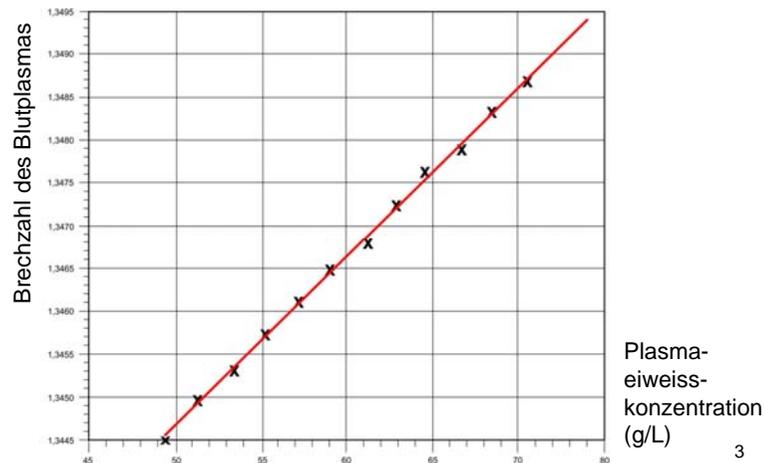
mit Hilfe einer (einfach) messbaren physikalischen Grösse, die
steht in streng monoton wachsendem Zusammenhang zu der
unbekannten Grösse

(die solche einfachste Funktion ist ...)

2

Bemerkung:

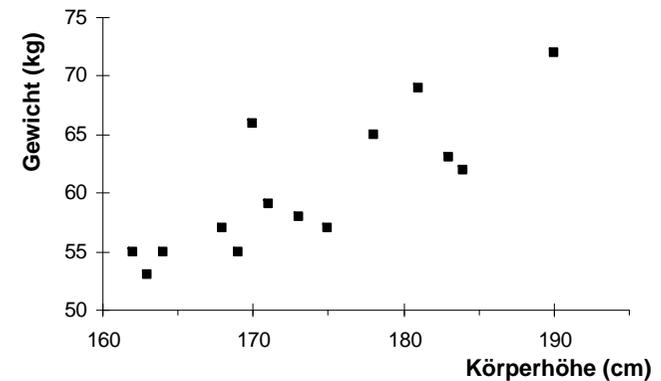
das Licht breitet sich in Blutplasma langsamer, wenn die
Plasmaeiweisskonzentration grösser ist, d.h. das Licht hat
grössere Brechzahl (deterministischer Zusammenhang, Messfehler)



3

(Beispiel2)

Daten aus einer Studentengruppe E2
(Sept. 1994) (zusammengehörige
Wertepaare)



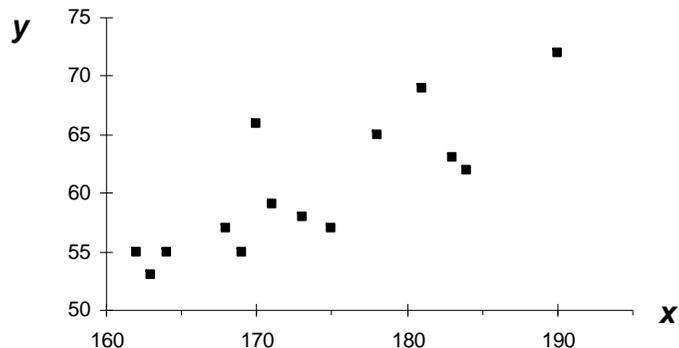
cm	kg
162	55
163	53
164	55
168	57
169	55
170	66
171	59
173	58
175	57
178	65
181	69
183	63
184	62
190	72

was für eine Tendenz kann man bemerken?

4

Die Korrelationsrechnung beschäftigt sich mit dem symmetrischen Zusammenhang zweier Zufallsgrößen

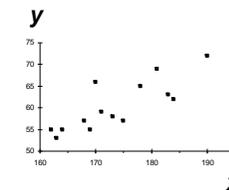
positive Korrelation: je mehr, desto mehr
negative Korrelation: je mehr, desto weniger



hier: positive Korrelation

Regressionsannäherung

Sucht man einen Funktionszusammenhang zwischen einer (oder mehreren) unabhängigen Variable (x) und einer abhängigen Variable (y)



Voraussetzungen: x und y numerische und stetige Merkmale, y Zufallsgröße (ihre Größe wird nicht nur von der unabhängigen Variable, sondern durch den Zufall beeinflusst)

(a: Steigung, b: Achsenabschnitt)

Regressionsmodell fixiert den Typ der Funktion:

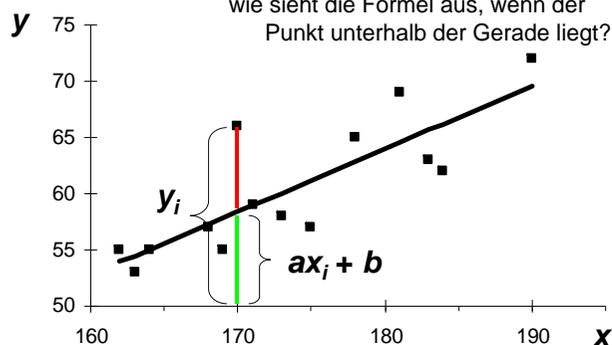
- lineare F. $y = (ax + b) + h$
- polinomiale F. $y = a + b_1x + b_2x^2 + \dots + b_nx^n + h$
- exponentiale F. $y = ab^x h$
- Potenzfunktion $y = ax^b h$

und wie wirkt der Zufall auf die abhängige Variable
additiver Fehler (+ h) oder multiplikativer Fehler (·h)

Das einfachste Regressionsmodell: lineare Regression

lineare Funktion: $y = (ax + b) + h$

$h_i = y_i - (ax_i + b)$ wenn der Punkt (x_i, y_i) oberhalb der Gerade liegt
wie sieht die Formel aus, wenn der Punkt unterhalb der Gerade liegt?



Beste Gerade: Summe der Fehlerquadrate ist minimal (Methode der kleinsten Quadraten)

	x_i	y_i
1	162	55
2	163	53
3	164	55
4	168	57
5	169	55
6	170	66
7	171	59
8	173	58
9	175	57
10	178	65
11	181	69
12	183	63
13	184	62
14	190	72

„Die beste“ Steigung:

$(y = ax + b)$

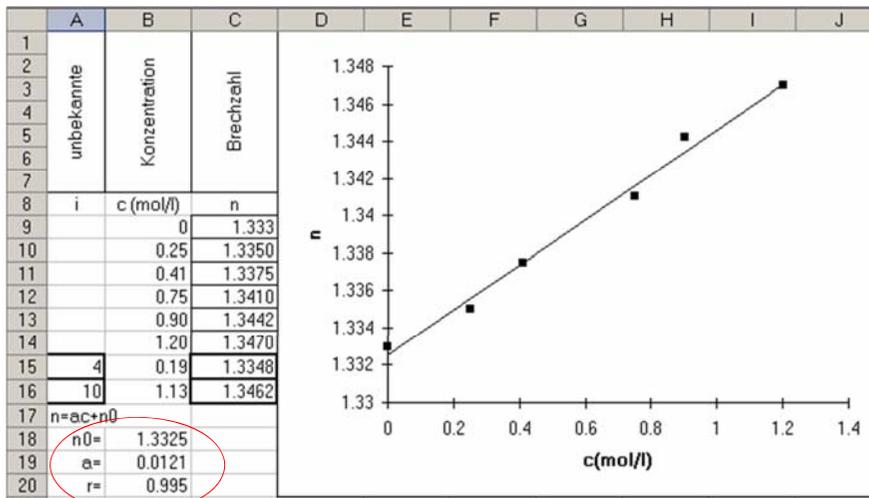
$$a^* = \frac{Q_{xy}}{Q_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{oder} \quad a^* = \frac{s_{xy}^2}{s_x^2}$$

„Der beste“ Achsenabschnitt:

$$b^* = \bar{y} - a^* \cdot \bar{x} = \frac{\sum_{i=1}^n y_i}{n} - a^* \frac{\sum_{i=1}^n x_i}{n}$$

wo $s_{xy}^2 = \frac{Q_{xy}}{n-1}$: Kovarianz

Beispiel: Refraktometrie



9

Wie gut passen die Messpunkte an die Regressionsgerade?

Korrelationsrechnung beschreibt die lineare Beziehung zwischen zwei oder mehr statistischen Variablen

es beschreibt die Stärke der Korrelation
es gibt starke und schwache Korrelation

Korrelationskoeffizient (Pearson)

$$r = \frac{Q_{xy}}{\sqrt{Q_{xx} \cdot Q_{yy}}} = \frac{s_{xy}^2}{s_x s_y}$$

der Zähler ist gleich dem Zähler der Steigung der Regressionsgerade (der Nenner ist im beiden Fall positiv)

$$a^* = \frac{Q_{xy}}{Q_{xx}}$$



positive Steigung: r ist positive Zahl
negative Steigung: r ist negative Zahl

$$-1 \leq r \leq 1$$

10

weitere Bemerkungen:

$$-1 \leq r \leq 1$$

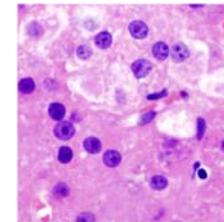
Korrelationskoeffizient (Pearson)

$$0 \leq r^2 \leq 1$$

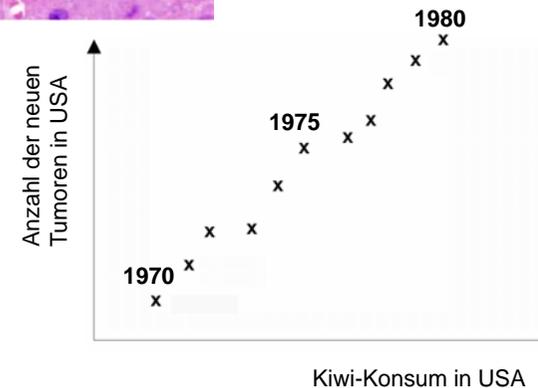
Bestimmtheitsmass (coefficient of determination)

Die Korrelation beschreibt nicht unbedingt eine Ursache-Wirkungs-Beziehung in die eine oder andere Richtung.

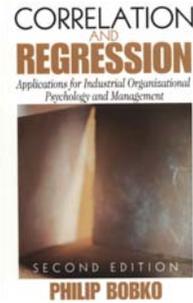
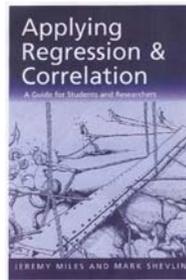
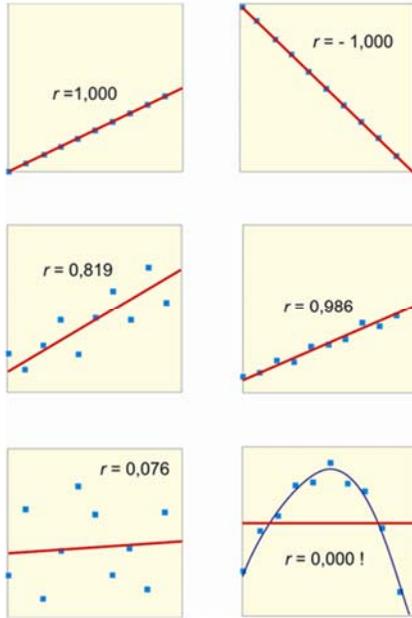
11



Korreliert heisst nicht notwendigerweise kausal verknüpft(!)

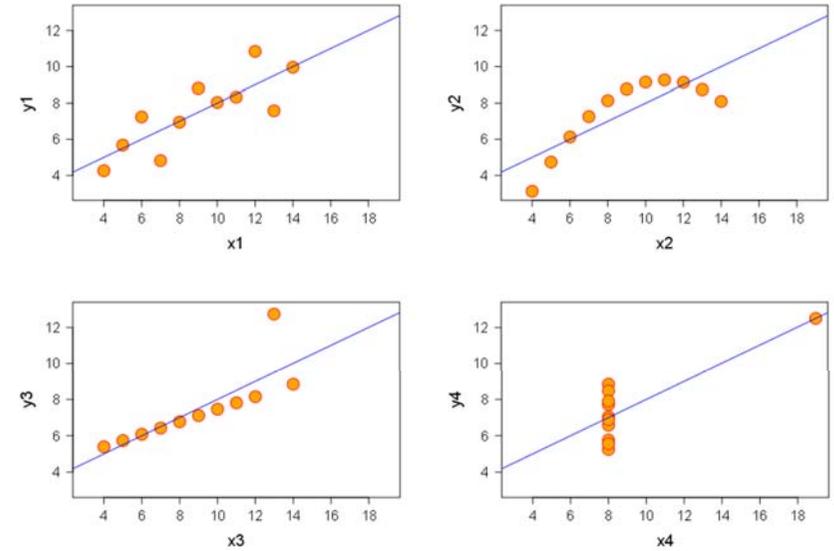


Beispiele:



Pr.Buch Abb. 15

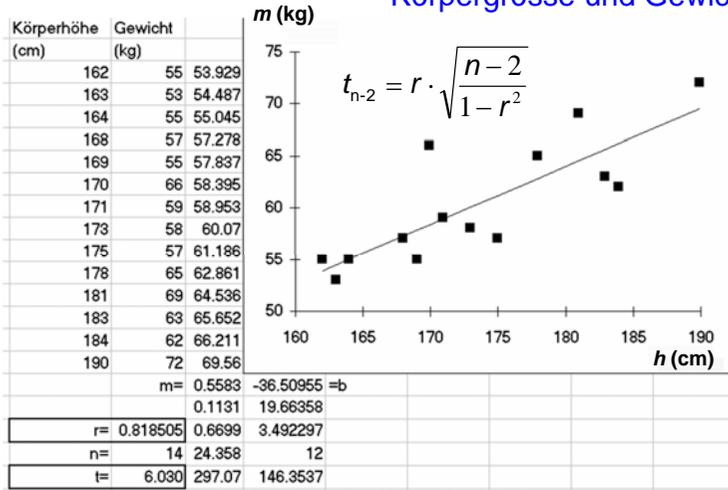
Extrembeispiel: $r=0.816, y = 3 + 0.5x$ (Anscombe's quartet)



http://en.wikipedia.org/wiki/Anscombe%27s_quartet

14

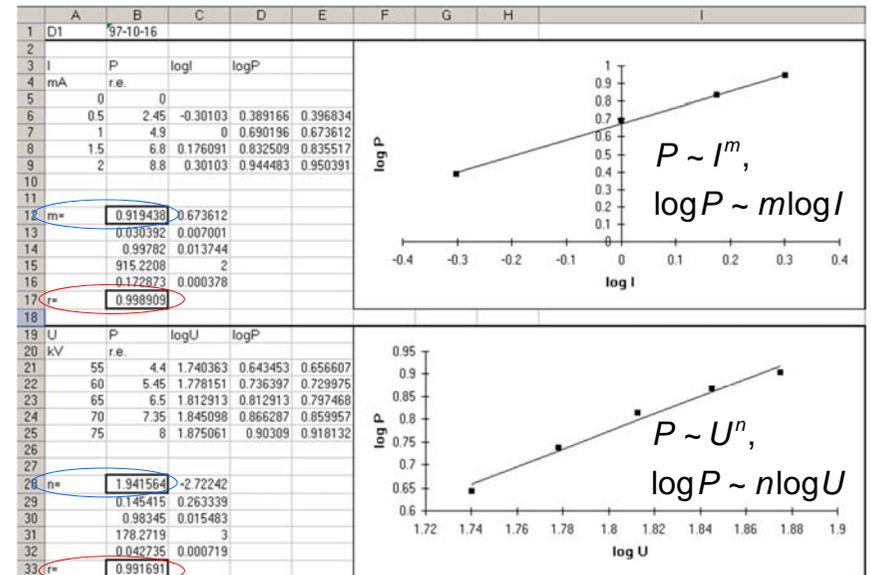
t-Test zur Korrelationsanalyse Gibt es eine Beziehung zw. der Körpergröße und Gewicht?



$|t| = 6.030 > t_{12, \text{krit}(0,05)} = 2.179 \Rightarrow H_0 \text{ ist falsch (} p < 0.05 \text{)}$
 $|t| = 6.030 > t_{12, \text{krit}(0,01)} = 3.055 \Rightarrow H_0 \text{ ist falsch (} p < 0.01 \text{)}$

15

Weiteres Beispiel: Leistung der Röntgen-Röhre



Kontingenztabellen. Chi-Quadrat-Test



Beispiel 1

	mit Brille	ohne Brille	Total
Frau	28	75	103
Mann	48	49	97
	76	124	200



?

Korrelationsanalyse zwischen kategorischen Merkmalen

Häufigkeitstabelle (Kontingenztable):
eine tabellarische Darstellung der gemeinsamen
Häufigkeitsverteilung zweier Variablen
X (z.B. Geschlecht) und Y (Brillenträgerschaft)

	mit Brille	ohne Brille	Total
Frau	a=28	b=75	103
Mann	c=48	d=49	97
	76	124	200

Frage: unterscheidet sich
die Häufigkeit eines
feststellbaren Merkmals
(Symptoms) in zwei
Populationen?

18

Aufstellung der Nullhypothese

H_0 : Geschlecht und Brillenträgerschaft
sind voneinander **unabhängig**
(es gibt keinen Unterschied)

$$\frac{a'}{b'} = \frac{c'}{d'} \quad \text{oder} \quad \frac{a'}{c'} = \frac{b'}{d'}$$

Wie gross wäre die **erwartete Häufigkeit**
(expected frequency) in der Zelle a' ,
wenn die Nullhypothese gültig ist?

Anzahl der Frauen:

$$a + b = 103$$

Anzahl der Personen mit Brille:

$$a + c = 76$$

Proportion der Frauen in der Stichprobe:

$$P(\text{Frau}) = (a + b)/n = 103/200$$

Proportion der Personen mit Brille:

$$P(\text{mit Brille}) = (a + c)/n = 76/200$$

	mit Brille	ohne Brille	Total
Frau	a'=?	b'=?	103
Mann	c'=?	d'=?	97
	76	124	200

erwartete (expected)
Kreuztabelle

19

Erwartete Häufigkeiten. Annahme: H_0 ist gültig \Rightarrow
Geschlecht und Brillenträgerschaft sind unabhängige Ereignisse

$$\text{erwartete Häufigkeit in der Zelle links oben: } a' = \frac{a+b}{n} \cdot \frac{a+c}{n} \cdot n = \frac{(a+b) \cdot (a+c)}{n}$$

$$\text{erwartete Häufigkeit in der Zelle rechts oben: } b' = \frac{a+b}{n} \cdot \frac{b+d}{n} \cdot n = \frac{(a+b) \cdot (b+d)}{n}$$

$$\text{erwartete Häufigkeit in der Zelle links unten: } c' = \frac{c+d}{n} \cdot \frac{a+c}{n} \cdot n = \frac{(c+d) \cdot (a+c)}{n}$$

$$\text{erwartete Häufigkeit in der Zelle rechts unten: } d' = \frac{c+d}{n} \cdot \frac{b+d}{n} \cdot n = \frac{(c+d) \cdot (b+d)}{n}$$

	mit	ohne	Total
F	a=28	b=75	103
M	c=48	d=49	97
	76	124	200

empirische (observierte,
observed) Kreuztabelle

	mit	ohne	Total
F	$103 \cdot 76 / 200$	$103 \cdot 124 / 200$	103
M	$97 \cdot 76 / 200$	$97 \cdot 124 / 200$	97
	76	124	200

erwartete (expected)
Kreuztabelle

20

Die erwartete Häufigkeiten aus der empirischen Häufigkeiten

	mit	ohne	Total		mit	ohne	Total
F	a=28	b=75	103	F	a'=39.14	b'=63.86	103
M	c=48	d=49	97	M	c'=36.86	d'=60.14	97
	76	124	200		76	124	200

empirische (observed)
Kreuztabelle

erwartete (expected)
Kreuztabelle

$$(\text{erwartete Häufigkeit}) = \frac{(\text{Spaltensumme}) \cdot (\text{Zeilensumme})}{(\text{Anzahl der Daten in der Stichprobe})}$$

21

Wenn die Nullhypothese ist gültig:

Die Werte in der entsprechenden Zellen der Kontingenztabelle mit empirischen und erwarteten Häufigkeiten sind ungefähr gleich.

Die folgende Prüfgrösse (gewichtete quadratische Summe) zeigt **Chi-quadrat Verteilung**:

Prüfgrösse

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

wobei

O_i die empirische (observed)

E_i die erwartete (expected) Häufigkeit in der i-ten Zelle sind.

Freiheitsgrad: (Anzahl der Zeilen - 1) * (Anzahl der Spalten - 1) für eindimensionalen Tabellen: $n-1$

z.B. 2*2 (vierfelder-) Tabelle: 1

22

Bedingungen der Durchführung

n (Stichprobenumfang) soll genügend gross sein

In der Kontingenztabelle der *erwarteten* Häufigkeiten sollen alle Zellenwerte grösser als 1 sein.

In der Kontingenztabelle der erwarteten Häufigkeiten soll die Anzahl der Zellen, in den der Wert zwischen 1 und 5 ist, weniger als 20 % der Stichprobenumfang sein.

(z.B. Vierfeldertabelle: alle Elemente sollen grösser als 5 sein)

23

Speziellfall für Vierfeldertabelle (Praktikumsbuch 2.b.30)

Vierfeldertest

	das untersuchte Merkmal		insgesamt
	ist vorhanden	ist nicht vorhanden	
Kollektiv A	a	b	$a+b$
Kollektiv B	c	d	$c+d$
insgesamt	$a+c$	$b+d$	n

$$\chi_M^2 = \frac{n \cdot (ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

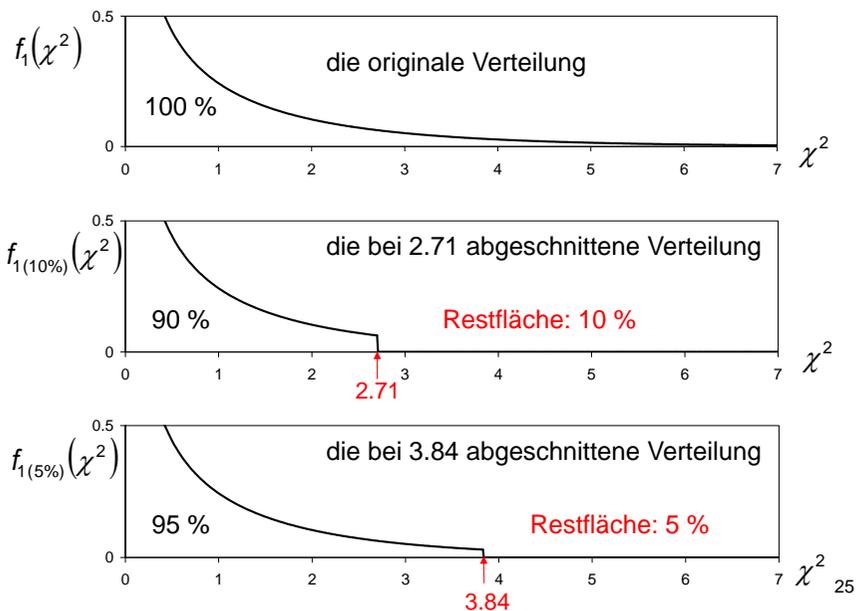
Die Bedingung der Durchführung:

das Produkt der zwei kleinsten Teilsummen soll grösser sein als $5n$

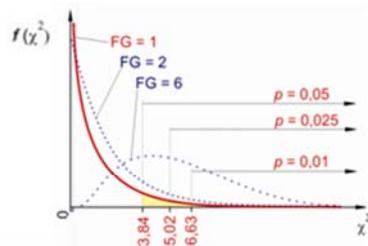
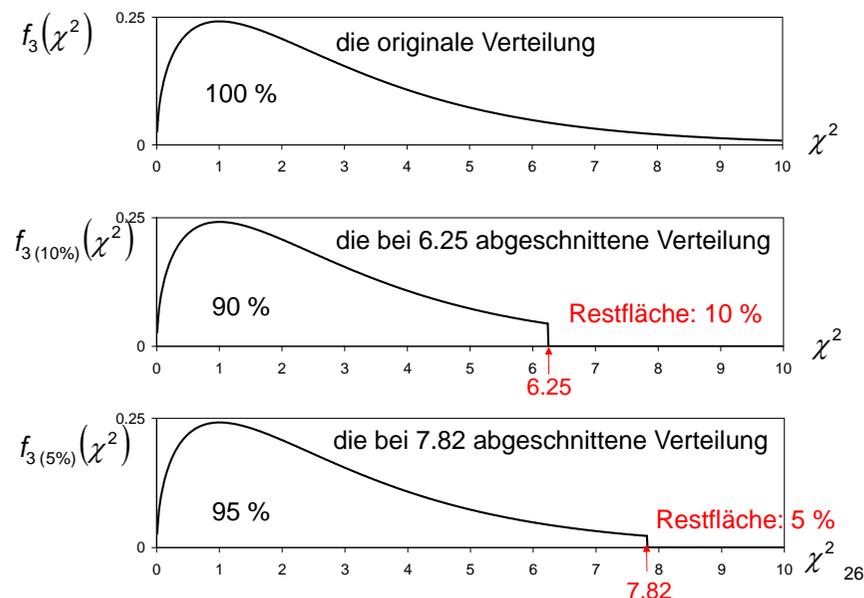
$$\frac{a}{b} = \frac{c}{d} \Leftrightarrow ad = bc$$

24

Chi-Quadrat-Verteilung mit dem Freiheitsgrad 1



Chi-Quadrat-Verteilung mit dem Freiheitsgrad 3



χ^2 (CHI-QUADRAT)-VERTEILUNG

Freiheitsgrad (FG)	p (Irrtumswahrscheinlichkeit)						
	0,99	0,975	0,95	0,05	0,025	0,01	0,001
1	0,0000157	0,0000982	0,000393	3,84	5,02	6,63	10,83
2	0,0201	0,0506	0,103	5,99	7,88	9,21	13,82
3	0,115	0,216	0,352	7,81	9,35	11,34	16,27
4	0,297	0,484	0,711	9,49	11,14	13,28	18,47
5	0,554	0,831	1,15	11,07	12,83	15,09	20,51
6	0,872	1,24	1,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	14,07	16,01	18,47	24,32
8	1,65	2,18	2,73	15,51	17,53	20,09	26,13

Beispiel 1 Die Bedingung des Tests:
das Produkt der zwei kleinsten
Teilsommen soll grösser sein als $5n$

	mit Brille	ohne Brille	Total
Frau	a=28	b=75	103
Mann	c=48	d=49	97
	76	124	200

$$76 \cdot 97 = 7372 > 5 \cdot 200 = 1000$$

Man darf den Chi-
Quadrat-Test anwenden

Es gibt einen
Zusammenhang zw.
dem Geschlecht
und der
Brillenträgerschaft
(Männer tragen
Brille öfter)

$$\chi^2_M = \frac{200 \cdot (28 \cdot 49 - 48 \cdot 75)^2}{76 \cdot 124 \cdot 103 \cdot 97} = 10.54$$

$$10.54 > \chi^2_{\text{krit}} = 3.84 \quad H_0 \text{ ist falsch}$$

Freiheitsgrad (FG)	p (Irrtumswahrscheinlichkeit)						
	0,99	0,975	0,95	0,05	0,025	0,01	0,001
1	0,0000157	0,0000982	0,000393	3,84	5,02	6,63	10,83

$$\chi_M^2 = \frac{200 \cdot (28 \cdot 49 - 48 \cdot 75)^2}{76 \cdot 124 \cdot 103 \cdot 97} = 10.54$$

$10.54 > \chi_{krit}^2 = 3.84$ H_0 ist falsch

$10.54 > \chi_{krit}^2 = 6.63$ H_0 ist falsch

mit einem Signifikanzniveau: < 0.01

29

	A	B	C	D
1	Empirische Werte			
2		mit Brille	ohne Brille	
3	Frau	28	75	=SUMME(B3:C3)
4	Mann	48	49	=SUMME(B4:C4)
5		=SUMME(B3:B4)	=SUMME(C3:C4)	=SUMME(B5:C5)
6				
7	Ewartete Werte			
8		mit Brille	ohne Brille	
9	Frau	=D3*B\$5/\$D\$5	=D3*C\$5/\$D\$5	=SUMME(B9:C9)
10	Mann	=D4*B\$5/\$D\$5	=D4*C\$5/\$D\$5	=SUMME(B10:C10)
11		=SUMME(B9:B10)	=SUMME(C9:C10)	=SUMME(B11:C11)
12				
13			Signifikanzniveau:	=CHITEST(B3:C4,B9:C10)
14			Chi ² -Wert:	=CHIINV(D13,1)

	A	B	C	D
1	Empirische Werte			
2		mit Brille	ohne Brille	
3	Frau	28	75	103
4	Mann	48	49	97
5		76	124	200
6				
7	Ewartete Werte			
8		mit Brille	ohne Brille	
9	Frau	39.140	63.860	103
10	Mann	36.860	60.140	97
11		76	124	200
12				
13			Signifikanzniveau:	0.0012
14			Chi ² -Wert:	10.5442606

Kalkulation mit Excel

30

Beispiel 2



	mit Brille	ohne Brille	Total
Frau	1	3	4
Mann	5	3	8
	6	6	12



$$4 \cdot 6 = 24 < 5 \cdot 12 = 60$$

Dürfen wir in diesem Fall den Chi-Quadrat-Test nicht anwenden.



Erhöhung des Umfanges der Stichprobe



	mit	ohne	Total
F	1	3	4
M	5	3	8
	6	6	12

12 → 200

$$\frac{n_{mit}}{n_{ohne}} = \frac{1}{3} = 0.33$$

Frauen

$$\frac{n_{mit}}{n_{ohne}} = \frac{28}{75} = 0.37$$

$$\frac{n_{mit}}{n_{ohne}} = \frac{5}{3} = 1.67$$

Männer

$$\frac{n_{mit}}{n_{ohne}} = \frac{48}{49} = 0.98$$

es gibt eine Vermutung, aber der Nachweis geht nicht

n vergrößert sich (12 → 200): der Nachweis geht

Beispiel 3 H_0 : die Häufigkeit von Lungenkrebs bei Rauchern und Nichtrauchern ist identisch, d.h. $\chi^2 = 0$.

H_1 : die beiden Häufigkeiten unterscheiden sich, also ist $\chi^2 \neq 0$.

In der Tabelle sind die Häufigkeiten der zwei Kollektive aus der Stichprobe einer Lungenfürsorge dargestellt.

	Lungenkrebs	kein Lungenkrebs	
Raucher	14	13	27
Nichtraucher	9	25	34
	23	38	61

Da $23 \cdot 27 = 621 > 9 \cdot 34 = 306$, kann der Test durchgeführt werden.

$$\chi_M^2 = \frac{61 \cdot (14 \cdot 25 - 9 \cdot 13)^2}{23 \cdot 38 \cdot 34 \cdot 27} = 4.13$$

Es ist zu sehen, dass $\chi_M^2 \neq 0$ ist, aber ist der Unterschied auch signifikant (oder nur zufällig)?

Sei das Signifikanzniveau: 5%.
Der Freiheitsgrad (2x2 Tabelle) ist: 1.

$$4.13 > \chi_{krit}^2 = 3.84 \Rightarrow H_0 \text{ ist falsch}$$

Danach ist der Unterschied in der Häufigkeit von Lungenkrebs bei Rauchern und Nicht-rauchern signifikant (bei einem Signifikanzniveau von 5%).

33

Pr.Buch 2.b.29 (Bemerkung 15)

Beispiel 4 (Pr. Buch, R.103.) Über eine erfolgreiche operative Korrektur einer bestimmten Augenkrankheit (ischaemische optische Neuropathie vom nicht-arterialen Typ) wurde im Jahre 1989 eine Veröffentlichung ausgegeben. Da in dieser Krankheit früher keinerlei wirksame Behandlungsmethode bekannt war, wurde dieser Eingriff verbreitet angewendet. Kürzlich erschienen jedoch Berichte auch von erfolglosen Eingriffen, daher hat man 244 solche Kranken in 25 klinischen Zentren statistisch erfasst, von denen bei 119 Personen die Operation durchgeführt wurde, bei 125 Kranken jedoch nicht. Die Beobachtungen in tabellarischer Form:

empirische Häufigkeiten

	operiert	nicht op.	insg.
verbessert	39	53	92
nicht verbessert	52	56	108
verschlechtert	28	16	44
insgesamt	119	125	244

erwartete Häufigkeiten

	operiert	nicht op.	insg.
verbessert	45	47	92
nicht verbessert	53	55	108
verschlechtert	21	23	44
insgesamt	119	125	244

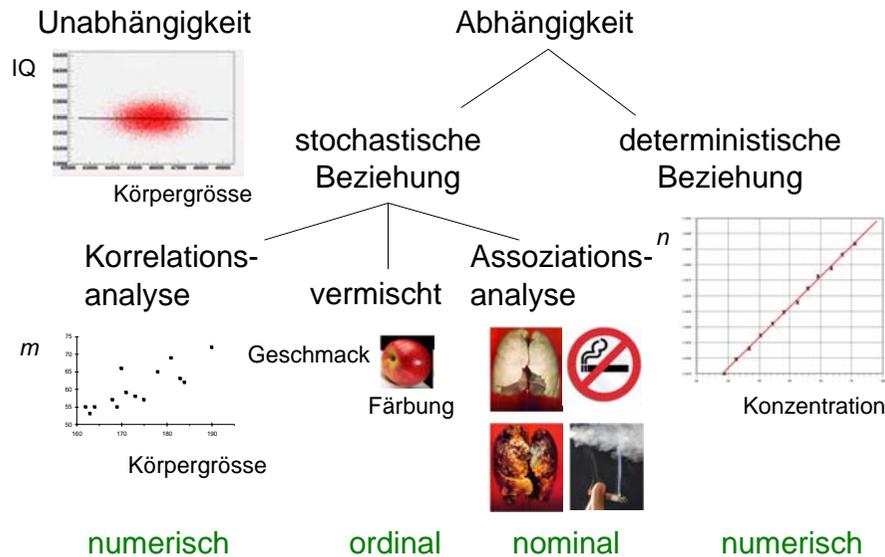
Es ist mit statistischen Methoden zu prüfen, ob die Anzahl der Besserungen ohne Operation tatsächlich höher war? H_0 : keine Differenz

$$\chi^2 = \frac{(39-44.87)^2}{44.87} + \frac{(53-47.13)^2}{47.13} + \frac{(52-52.67)^2}{52.67} + \frac{(56-55.33)^2}{55.33} + \frac{(28-21.46)^2}{21.46} + \frac{(16-22.54)^2}{22.54} = 5.407$$

Weil $5.407 < 5.991 = \chi_{krit, FG=2}^2$, ablehnen wir die H_0 nicht.

34

Arten von Abhängigkeitsbeziehungen



35