Statistik

TEIL 7

Hans-Hermann Thulke
ba @ thulke-statistics.de
0172-3449934

Statistik

- Daten erheben, verstehen, werten
- Hypothesen prüfen
- Modellieren von Zusammenhängen

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Modellieren von Zusammenhängen

Multivariate Zufallsgrößen
Zusammenhangsmaße
Zusammenhangsmodelle
Trendbestimmung

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

„Multivariate Daten“

Paare von zufälligen Werten X,Y (z.B. Größe und Gewicht eines Menschen)

Wahrscheinlichkeitsfunktion

$f_{X;Y}(x; y) = WS(X = x; Y = y)$

Verteilungsfunktion

$F_{X;Y}(x; y) = WS(X \leq x; Y \leq y)$

X und Y werden identisch verteilt genannt, wenn

Für $f_X(x) = WS(X = x)$ $f_Y(y) = WS(Y = y)$ gilt

$f_X = f_Y$

X und Y werden unabhängig genannt, wenn

für alle (x, y) $f_{X,Y} = f_X * f_Y$

z.B. diskret $WS(X = x; Y = y) = WS(X = x) * WS(Y = y);$

Terminus technicus!! (auch mehr als 2)

Wichtig:

A: Verschiedene Merkmale an gleichen Objekten

B: Gleiche Merkmale an verschiedenen Objekten

Mittelwert (Erwartungswert); Varianz etc. ganz analog als Wertepaare berechnen

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Modellieren von Zusammenhängen

Multivariate Zufallsgrößen

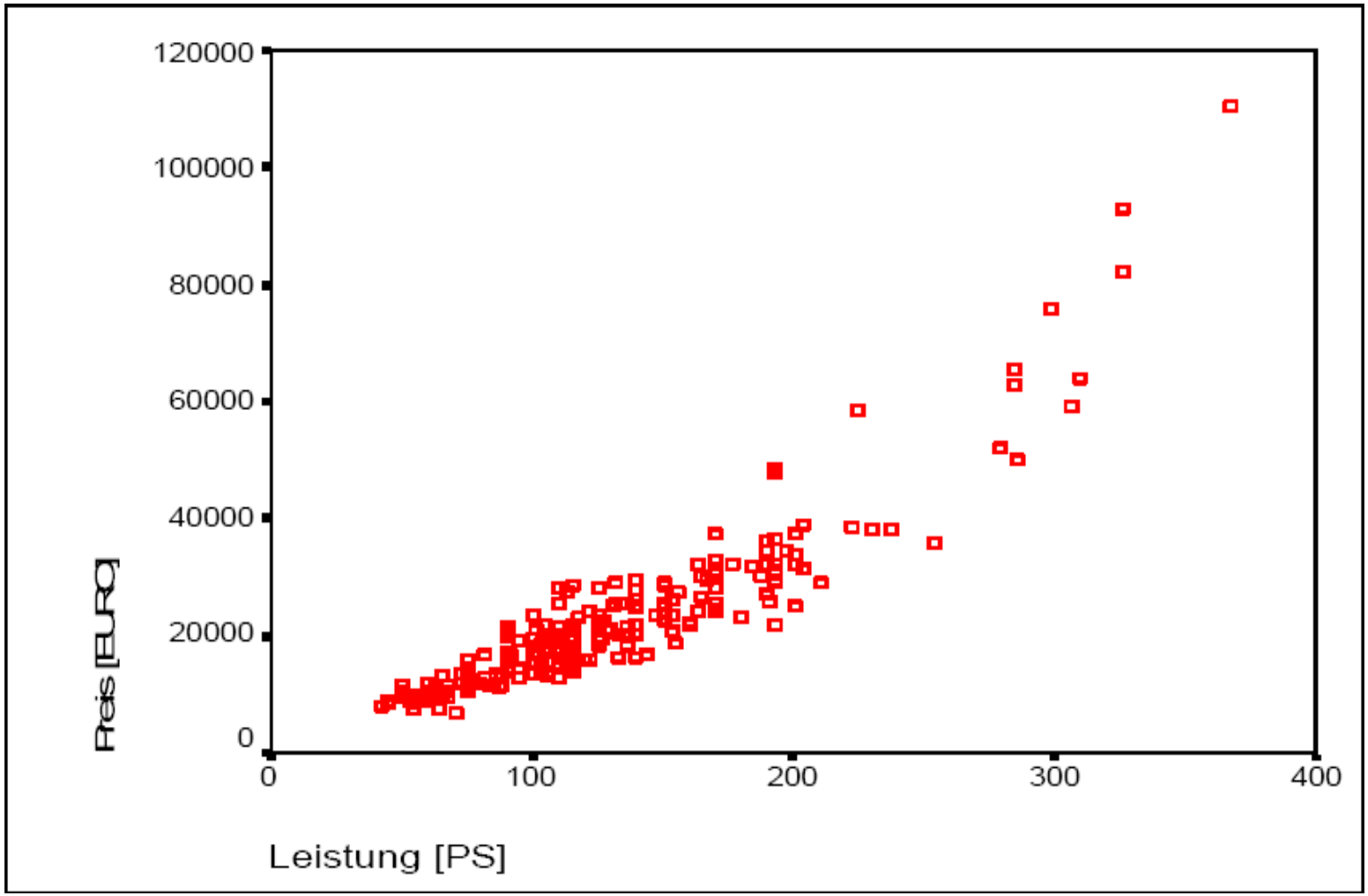
Zusammenhangsmaße

Zusammenhangsmodelle

Trendbestimmung

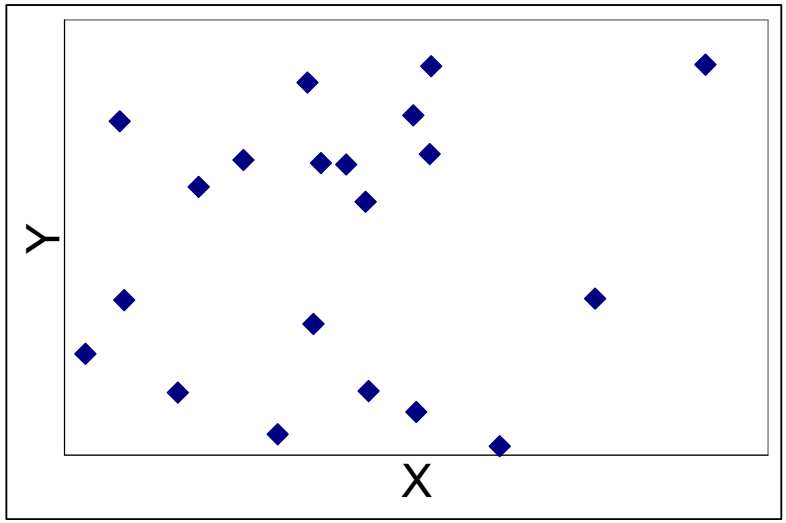
thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

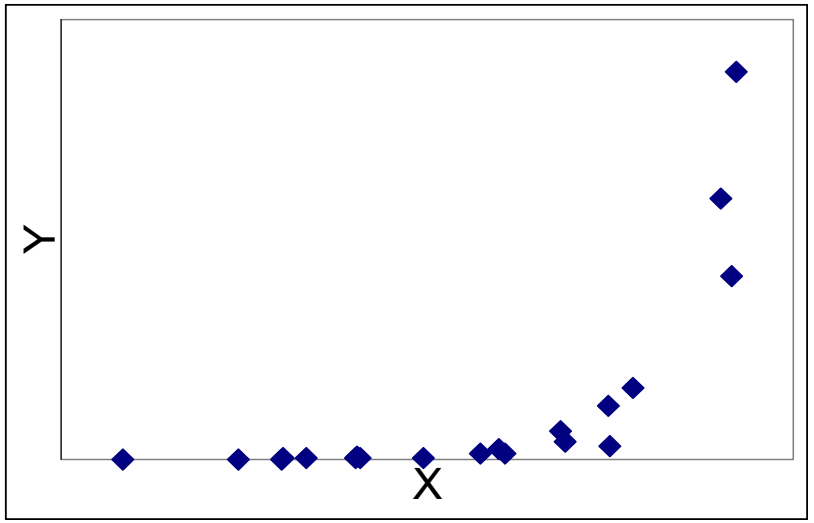
Zusammenhang - beschreiben



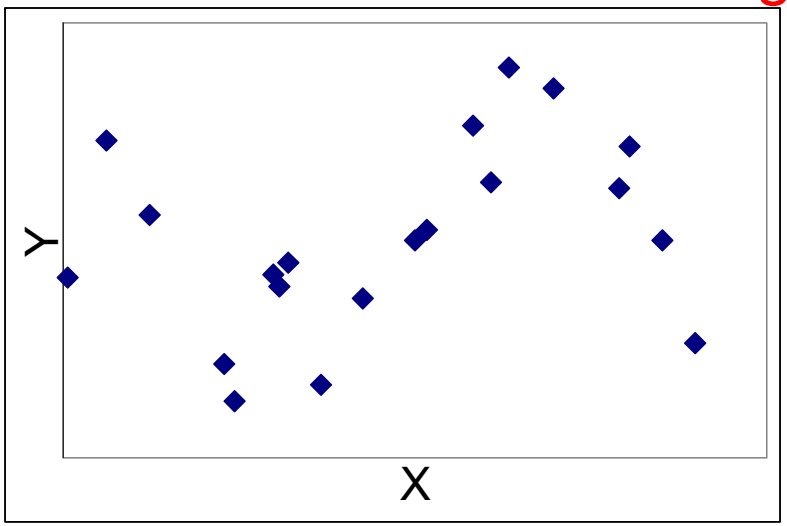
thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

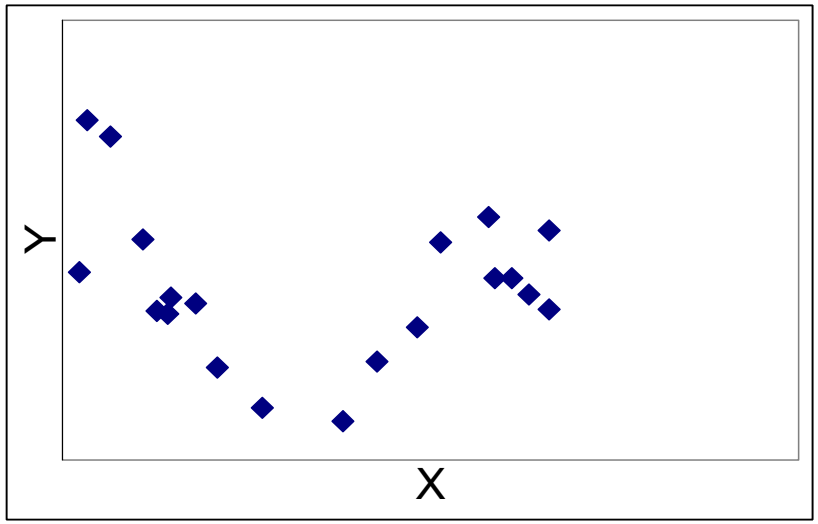
Zusammenhang - beschreiben





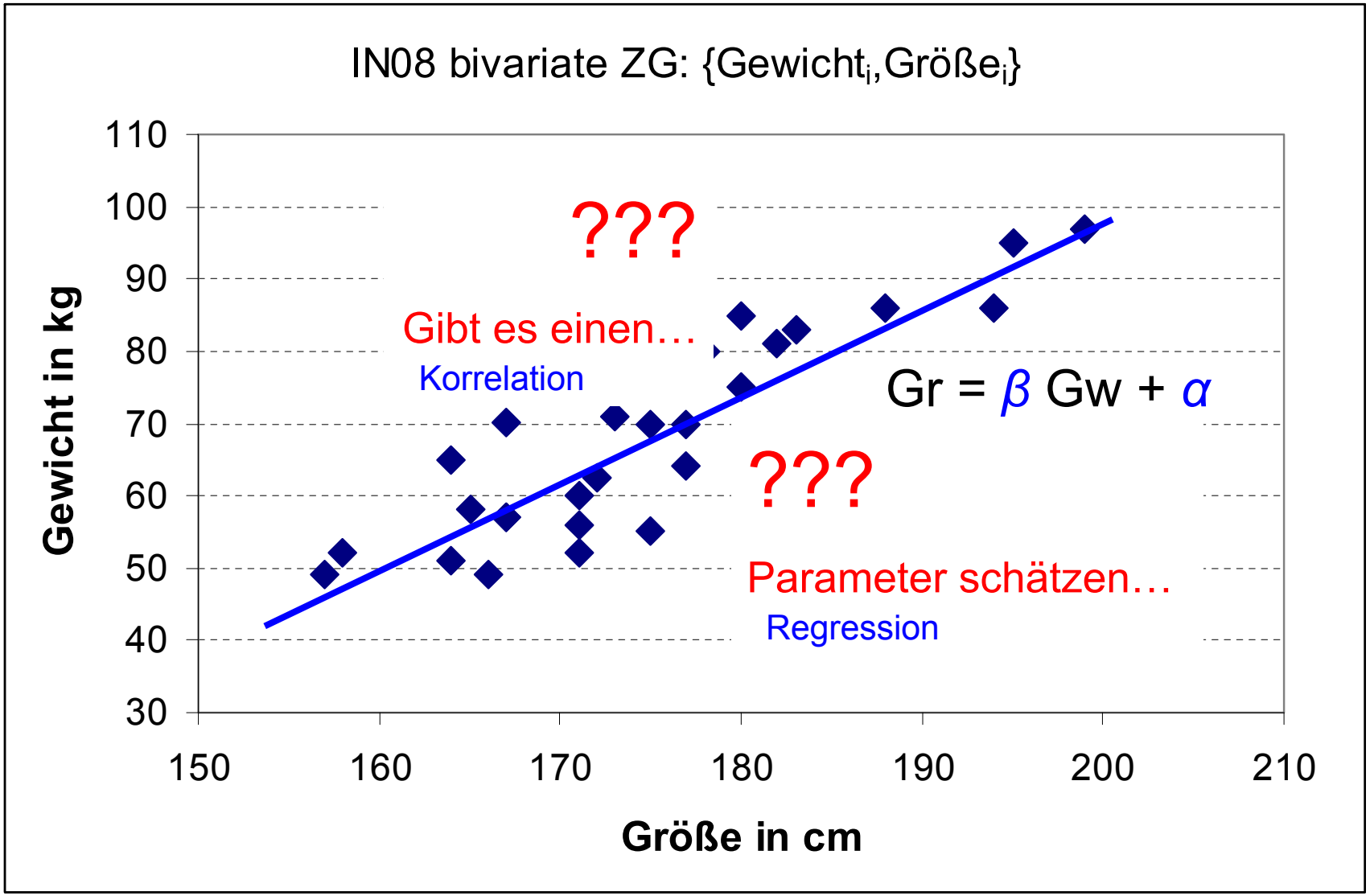
Form des Zusammenhang vermuten + ev. Linearisieren!!!





thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - beschreiben



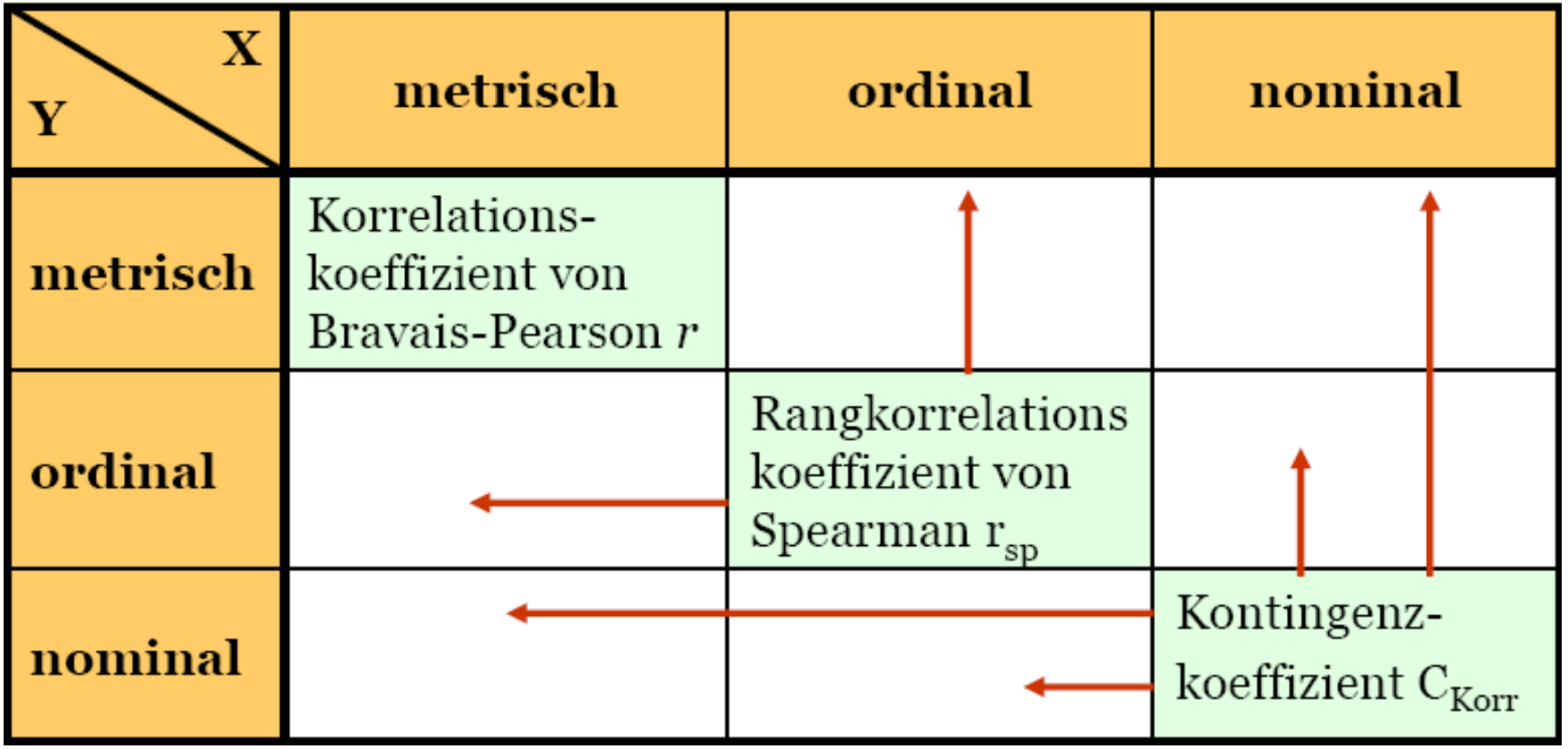
Unser Beispiel… Form? Linear!

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

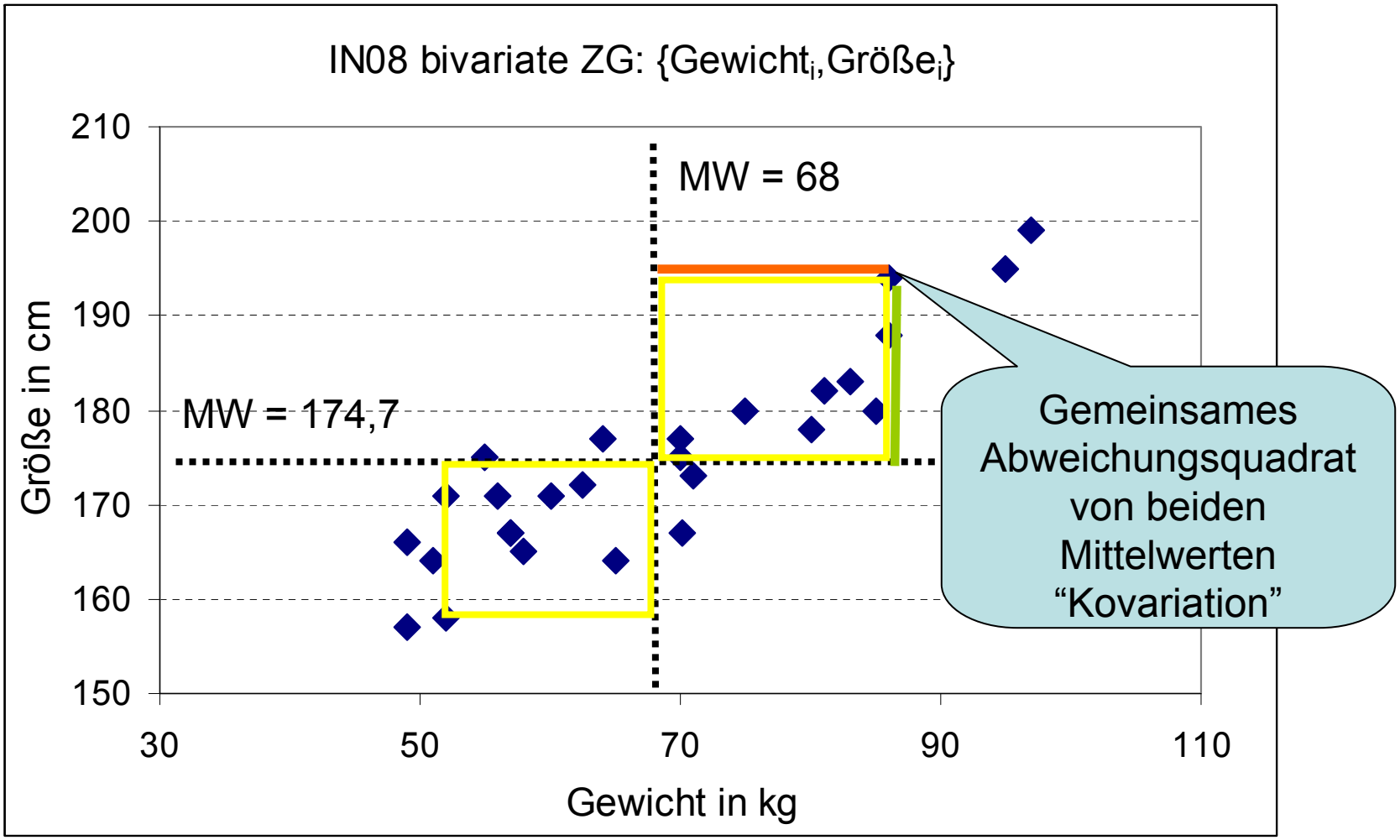
Korrelationsanalyse bestimmt Stärke des statistischen Zusammenhangs

Skalenniveau der Merkmale bestimmt welcher Korrelationskoeffizienten.



thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

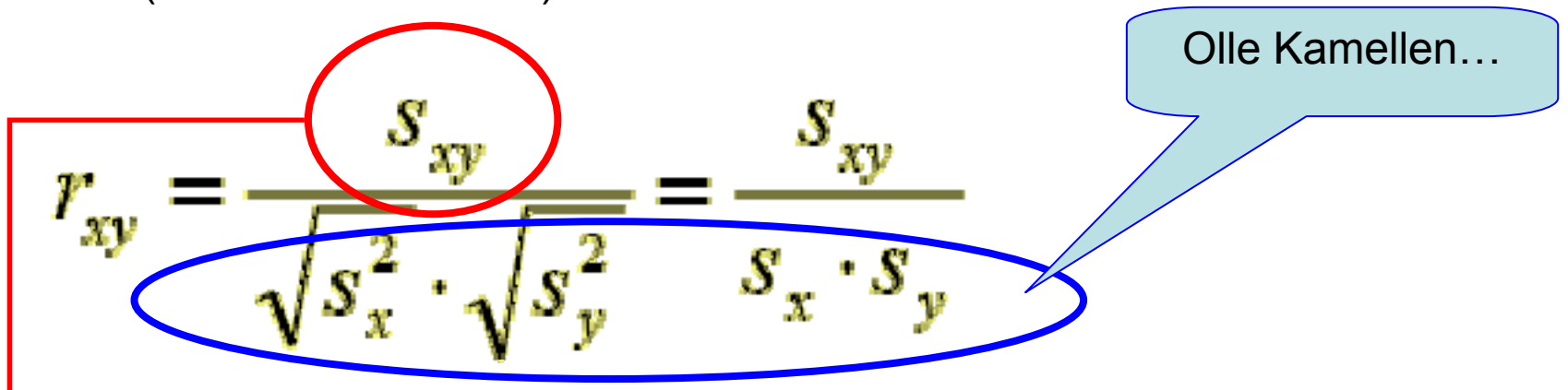
Zusammenhang - Korrelation



thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Pearson (**metrisch-metrisch**)



$$-1 \leq r_{xy} \leq 1$$

→ Kovarianz (kovariieren)

$$Cov(X,Y) = s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

$$Cov(X,X) = s_{xx} = s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = Varianz(X)$$

Modellieren von Zusammenhängen

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Pearson (**metrisch-metrisch**)

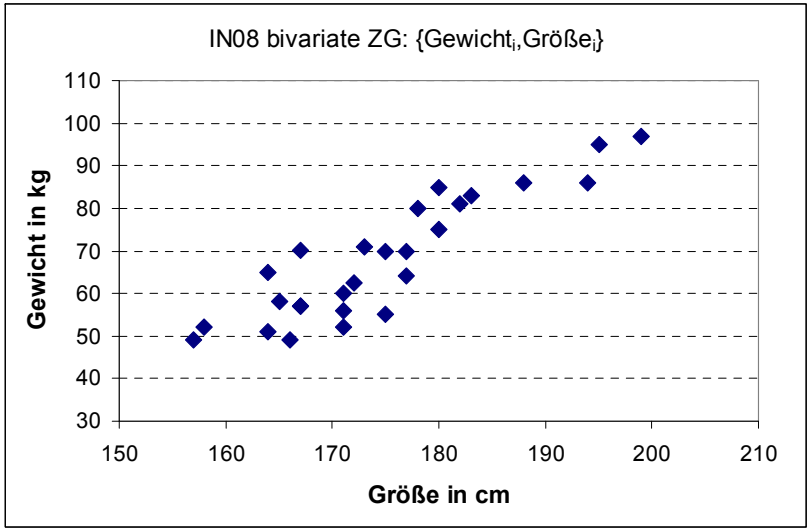
$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x} \cdot \bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

Effektivste Handrechnung / EXCEL → PEARSON (X;Y)

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Beispiel

Besteht ein Zusammenhang zwischen Wuchshöhe und Gewicht?



Berechne
Summe xi² = 826700
Summe yi² = 130342
Summe xi*yi = 324387

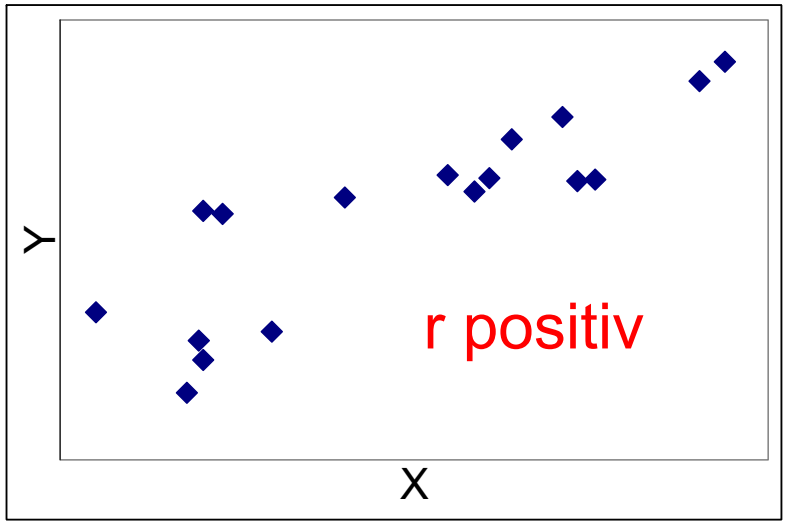
Formel oben = 3594
Unten = 54,5 * 73,6

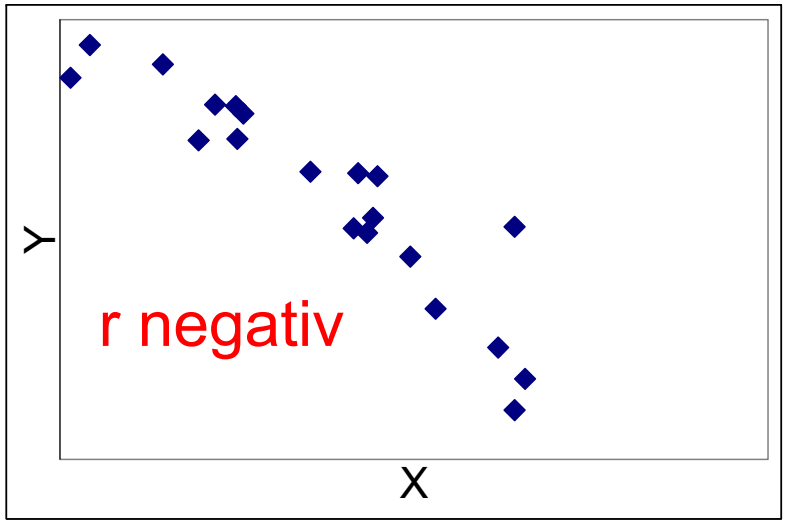
r = 0,896

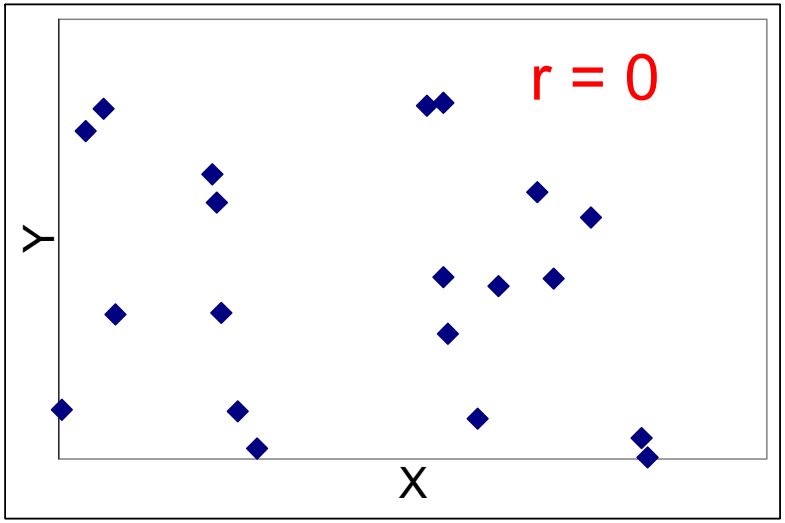
Es besteht ein starker positiver Zusammenhang zwischen den Merkmalen Wuchshöhe und Gewicht!

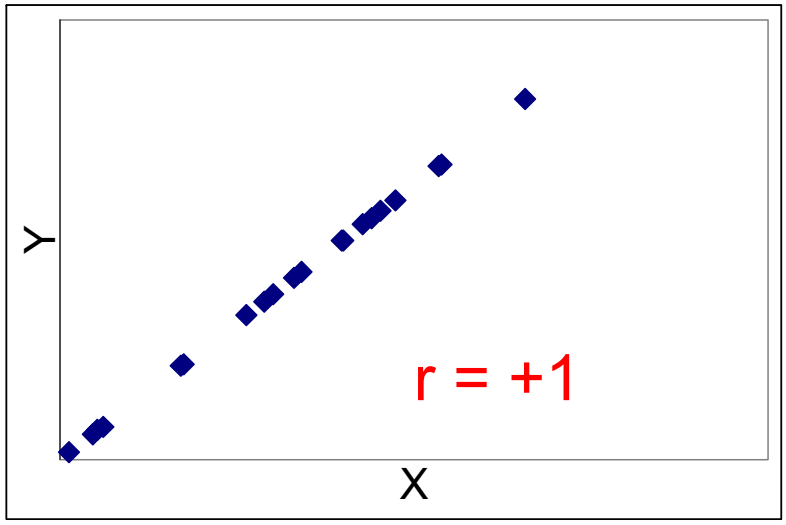
Weight	Height
64	177
70	177
80	178
75	180
85	180
81	182
83	183
86	188
86	194
95	195
97	199
49	157
52	158
51	164
65	164
58	165
49	166
57	167
57	167
70,1	167
52	171
56	171
60	171
62,5	172
71	173
55	175
70	175

Zusammenhang - Korrelation









thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Pearson (**metrisch-metrisch**)

Korrelationskoeffizienten ist eine **geschätzte** statistische Kenngröße, daher Konfidenzintervall!!!

r geschätzt, dann (mit Fisher-Transformation erreicht man Normalverteilung)

$$z_{1,2} = 0{,}5 \ln\left(\frac{1+r}{1-r}\right) \quad +/- \quad \frac{z_{1-\alpha/2}}{\sqrt{n-3}}$$

Retransformation:

$$r_1 = (e^{2z_1} - 1)/(e^{2z_1} + 1)$$
$$r_2 = (e^{2z_2} - 1)/(e^{2z_2} + 1)$$

Konfidenzintervall für den Korrelationskoeffizienten: $r_1 \leq \varrho \leq r_2$.

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Korrelationskoeffizienten ist eine **geschätzte** statistische Kenngröße!!!
Oder… prüfen ob das berechnete r nur zufällig von 0 verschieden sein kann!

t-Test anwenden (gilt nur wenn X & Y normalverteilt mit gleichem σ)
H0: r = 0 (SOLL abgelehnt werden)
HA: r ≠ 0

Prüfgröße $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ t-verteilt mit n-2 Freiheitsgraden

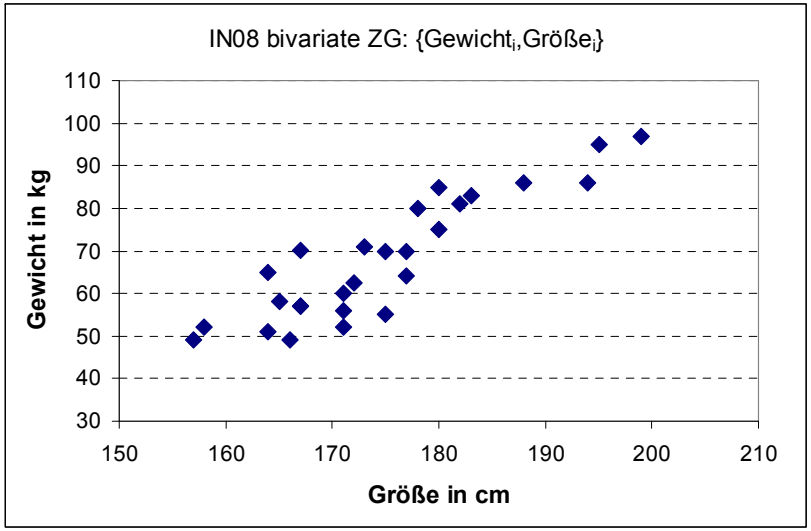
synonym Teststatistik

Ablehnung wie immer: $|t| \geq t_{n-2;\ 1-\alpha/2}$

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Beispiel

Besteht ein Zusammenhang zwischen Wuchshöhe und Gewicht?



Berechne
Summe xi² = 826700
Summe yi² = 130342
Summe xi*yi = 324387

Formel oben = 3594
Unten = 54,5 * 73,6

r = 0,896

Weight	Height
64	177
70	177
80	178
75	180
85	180
81	182
83	183
86	188
86	194
95	195
97	199
49	157
52	158
51	164
65	164
58	165
49	166
57	167
57	167
70,1	167
52	171
56	171
60	171
62,5	172
71	173
55	175
70	175

Es besteht ein starker positiver Zusammenhang zwischen
Den Merkmalen Wuchshöhe und Gewicht?

Test:
H0 : $r_{tatsächlich} = 0$
HA: $r_{tatsächlich} \neq 0$

p-Wert < 0,0000000003
Verwende EXCEL TVERT(10,7;FG;2)

Teststatistik bzw. Prüfgröße berechnen ergibt: 10,1
Quantil: 2,06 → **H0 ist abzulehnen**

PG >> Quantil

Modellieren von Zusammenhängen

Zusammenhang - Korrelation

Normalverteilung oder Linearität liegt NICHT vor

Spearmanscher Rangkorrelationskoeffizient (**ordinal-ordinal**)

Vorgehen:

→ $x_1; \ldots ; x_n$ und $y_1; \ldots ; y_n$ jeweils der Größe nach ordnen

→ Messwerte $x_i; y_i$ durch ihre Ränge $r_{x;i}; r_{y;i}$ ersetzen

→ Gleiche Werte bekommen die mittlere Position die sie besetzten als Rang

$$r_S := 1 - \frac{6 \sum_{i=1}^{n} (r_{x,i} - r_{y,i})^2}{n(n^2 - 1)} \in [-1; 1]$$

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Merkmal nur nominal skaliert d.h. nur auszählen aber kein ordnen möglich

Kontingenzkoeffizient (**nominal-beliebig**)

Mehr- bzw. Vierfeldertafeln (Merkmale dichotomisiert)

	Y ja	Y nein	
X ja	a	b	Σ X ja
X nein	c	d	Σ X nein
	Σ Y ja	Σ Y nein	gesamt

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Kontingenzkoeffizient (**nominal-beliebig**)

Mehr- bzw. Vierfeldertafeln (Merkmale dichotom)

	Y ja	Y nein	
X ja	a a=● * ═	b	Σ X ja
X nein	c	d	Σ X nein
	Σ Y ja	Σ Y nein	gesamt

Wenn kein Zusammenhang… müsste gelten

Vergleich a und a: $\frac{(a - a)^2}{a}$ Standardisierte Abweichung „Beobachtet“ vs. „Erwartet“

Summieren über alle Felder ergibt die sog. χ^2 Prüfgröße

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Kontingenzkoeffizient (**nominal-beliebig**)

Mehrfeldertafeln

$$n_{jk}^{*} = \frac{n_{j\cdot} \cdot n_{\cdot k}}{n}$$

$$\chi^2 = \sum_{j=1}^{m} \sum_{k=1}^{r} \frac{(n_{jk} - n_{jk}^{*})^2}{n_{jk}^{*}}$$

a a a

Chi² Test

H0: Das Merkmal X ist vom Merkmal Y stochastisch unabhängig d.h. es besteht kein Zusammenhang
HA: Zusammenhang

*H*0 abgelehnt, wenn $\chi^2 > \chi^2 (1-\alpha; (m-1)(r-1))$,

dem (1-*α*)-Quantil der χ^2-Verteilung mit (*m*-1)(*r*-1) Freiheitsgraden ist.

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation

Kontingenzkoeffizient (**nominal-beliebig**)

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} \cdot \sqrt{\frac{1 + \min\{s-1, r-1\}}{\min\{s-1, r-1\}}}$$

$$0 \leq C \leq 1$$

Korrektur der Chi² Größe auf Werte zwischen 0 und 1 (“Korrelation”)

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Anmerkung

Der Chi²-Test ist ein sehr wirkungsvolles Werkzeug, um verschiedene Merkmale zu vergleichen.

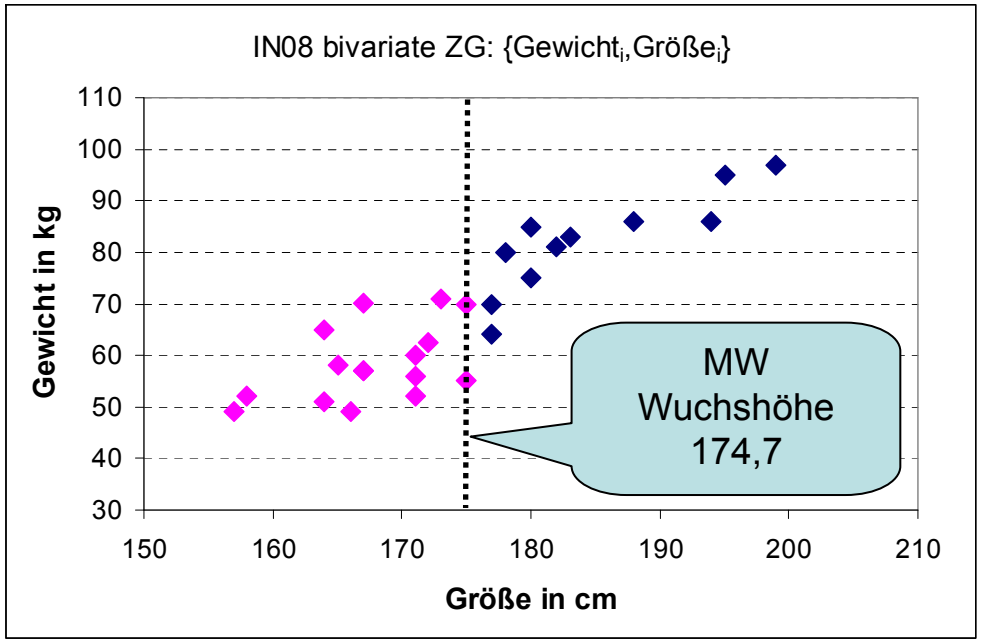
Denn, jedes Merkmal lässt sich in z.B. ein nominales überführen…

Siehe folgendes Beispiel….

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Beispiel

Besteht ein Zusammenhang zwischen dem Merkmal Sex und Wuchshöhe?



Nominal vs. Metrisch???

z.B. Dichotomisieren des metrischen Merkmals:

„<“ od. „>“ MW

H0: m & w unterscheiden sich nicht bzgl ihrer Lage zum gemeinsamen MW

HA. Es besteht (irgendein!!!) Zusammenhang zwischen m/w und Wuchshöhe

Height	Sex
177	m
177	m
178	m
180	m
180	m
182	m
183	m
188	m
194	m
195	m
199	m
157	w
158	w
164	w
164	w
165	w
166	w
167	w
167	w
167	w
171	w
171	w
171	w
172	w
173	w
175	w
175	w

	+	-	
m	11	0	11
w	2	14	16
	13	14	**27**
	5,30	5,70	
	7,70	8,30	
	6,14	5,70	
	4,22	3,92	
			19,99	0,000008

Randsummen

Auszählen

Erwartete Anzahl: Jede Zeilenanzahl wird gemäß des Verhältnisses der Spaltenanteile aufgeteilt

(Tatsächlich – Erwartet)² / Erwartet

p-Wert <<< 0,05 d.h. H= ablehnen

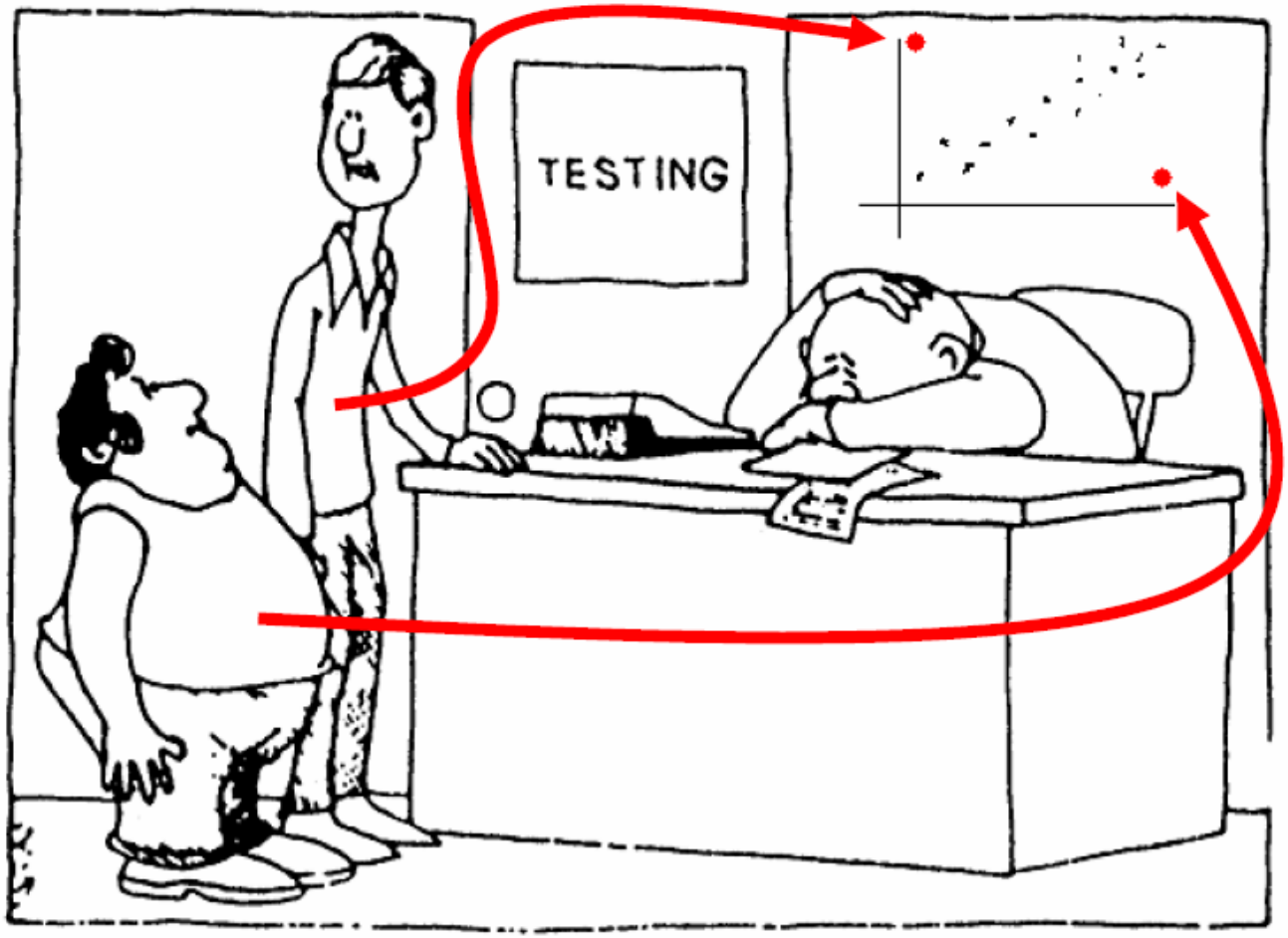
Summe über alle 4 ist Wert der PG

95% Quantil der Chi² Vert. 3,84
→ **PG >> Quantil d.h. H0 ablehnen**

Modellieren von Zusammenhängen

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Zusammenhang - Korrelation



»Er sagt, wir ruinieren seine ganze schöne Korrelation zwischen Größe und Gewicht.«

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de

Modellieren von Zusammenhängen

Multivariate Zufallsgrößen

Zusammenhangsmaße

Zusammenhangsmodelle

Trendbestimmung

thulke.statistics – leipzig – 0172-3449934 – ba@thulke-statistics.de