Frage: Ist die Reaktionszeit in einem bestimmten Experiment vom Geschlecht unabhängig?

Gegebene Objekte:

Gesunde Personen zwischen 30 und 40 Jahre.

Merkmale:

T = Reaktionszeit (ms)
S = Geschlecht (m, w)

Zeit: Stetig

Geschlecht: Diskret, nominal / skaliert

Stetig ist die Zeit

↓

Klassen

$K_1 = [100; 130)$

$K_2 = [130; 160)$

$K_3 = [160; 190)$

200 Personen

$$Z = (T, S)$$

P.P.1

$$z_i = (t_i, s_i)$$

Merkmalsausprägung eines

Probande i, $i = 1, \ldots, 200$

T \ S	0 m	1 w	Σ
[100, 130)	20	40	60
[130, 160)	50	60	110
[160, 190)	10	20	30
Σ	80	120	200

Absolute Zell-Häufigkeiten

T \ S	0 m	1 w	Σ
[100, 130)	0,1	0,2	0,3
[130, 160)	0,25	0,3	0,55
[160, 190)	0,05	0,1	0,15
Σ	0,4	0,6	1

Relative Zell-Häufigkeiten

Tab.: 3.1: Zweidimensionale Häufigkeitstabellen von Reaktionszeit T und Geschlecht S

Verteilungen in Gruppen, m und w bezogen auf Beobachtungszahlen aus der Gruppen:

T :	$[100, 130)$ K_1	$[130, 160)$ K_2	$[160, 190)$ K_3
m :	20/80	50/80	10/80
w :	40/120	60/120	20/120
Σ :	60/200	110/200	30/200

T:	K_1	K_2	K_3
u:	0,25	0,625	0,125
w:	0,33	0,5	0,17
Σ	0,3	0,55	0,15

Etwas Hilfe

? Mäßig ?

Zweidimensionale Häufigkeitsverteilungen

Sei $(X, Y) = Z$,

$x \in \mathbb{X}$, $y \in \mathbb{Y}$

ein 2-dim Merkmal

Fall 1: X, Y diskret

$\mathbb{X} = \{a_1, \ldots, a_k\}$

$\mathbb{Y} = \{b_1, \ldots, b_e\}$

$Z = \{(x, y)\}$

Wertebereich von Z:

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \{c_{ij} = (a_i, b_j) \mid a_i \in \mathcal{X},\ b_j \in \mathcal{Y}\}$

Stichprobe vom Umfang n:

$z_i = (x_i, y_i), \quad i = 1, \ldots, n$

.

Bezeichnung:

$C_{ij} = (a_i, b_j)$ – Zelle

$H_{ij} :=$ # Aller Paare (x_k, y_k) in der Stichprobe mit

$x_k = a_i$

$y_k = b_j$

$h_{ij} := \frac{H_{ij}}{n}$ Absolute Zellenhäufigkeit

relative Zellenhäufigkeit.

Fall 2: X stetig $\vee$ Y stetig.

Klassenteilung von X bzw. Y (bzw. X und Y).

K_i^X $i = 1, \dots, k$ Klassenteilung von X

K_j^Y $j = 1, \dots, \ell$ Klassenteilung von Y

(K_i^X, K_j^Y) – (Klassen-) Zelle

H_{ij} – Anzahl aller (x_ν, y_ν) mit $x_\nu \in K_i^X$, $y_\nu \in K_j^Y$

h_{ij} – Anteil aller (x_ν, y_ν) mit $x_\nu \in K_i^X$, $y_\nu \in K_j^Y$

Def: Die Gesamtheit aller absoluten bzw. relativen Häufigkeiten:

$$\left((h_{ij})_{i=1,\dots,k}^{j=1,\dots,\ell} \right)$$

bzw.

$$\left((h_{ij})_{i=1,\dots,k}^{j=1,\dots,\ell} \right)$$

heißt zweidimensionale absolute bzw. relative Häufigkeitsverteilung von Z.

Kontingenztabelle (k x l –Tabelle):

X \ Y	K_1^Y b_1	K_2^Y b_2	...	K_j^Y b_j	...	K_l^Y b_l	Zeilensumme Σ
K_1^X a_1	H_{11}	H_{12}	...	H_{1j}	...	H_{1l}	$H_{1\bullet}$
K_2^X a_2	H_{21}	H_{22}	...	H_{2j}	...	H_{2l}	$H_{2\bullet}$
⋮	⋮	⋮		⋮		⋮	⋮
K_i^Y a_i	H_{i1}	H_{i2}	...	H_{ij}	...	H_{il}	$H_{i\bullet}$
⋮	⋮	⋮		⋮		⋮	⋮
K_k^X a_k	H_{k1}	H_{k2}	...	H_{kj}	...	H_{kl}	$H_{k\bullet}$
Σ Spaltensumme	$H_{\bullet 1}$	$H_{\bullet 2}$	...	$H_{\bullet j}$	...	$H_{\bullet l}$	n

Tab. 3.2

$$H_{i\bullet} = \sum_{j=1}^{l} H_{ij}$$ – i-te Zeilensumme

$$H_{\bullet j} = \sum_{i=1}^{k} H_{ij}$$ – j-te Spaltensumme

$H_{i\bullet}, H_{\bullet j}$ – Randsummen

$$H_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{l} H_{ij} = n$$

– Stichprobenumfang

Def: Die Gesamtheit der Randsummen bilden die Häufigkeitsverteilung von X bzw. Y:

$\left((H_{i\cdot})_{i=1,\dots,k} \right)$ – Häufigkeitsverteilung von X

$\left((H_{\cdot j})^{j=1,\dots,\ell} \right)$ – Häufigkeitsverteilung von Y

Randverteilung

Bem.: Wenn wir statt

H_{ij} die relative Häuf.

h_{ij} betrachten, so erhalten

wir die Randsummen

$h_{i\cdot}$, $h_{\cdot j}$. Diese sind

relative Häufigkeitsverteilungen

von X bzw. Y.

1000 Eier

X = Länge (cm)

Y = Breite (cm)

Frage: Zusammenhang zwischen Länge und Breite?

Objekte: Eier

Merkmal: $Z = (X, Y)$ – 2-dim Merkmal

X, Y stetig → Klasseneinteilung für beide Merkmale

Häufigkeitstabelle der 1000 Beobachtungen:

Y / X	$2 < Y \leq 3$	$3 < Y \leq 4$	$4 < Y \leq 5$	$5 < Y \leq 6$	Summe
$4 < X \leq 5$	10	66	0	0	76
$5 < X \leq 6$	5	250	157	0	412
$6 < X \leq 7$	0	47	380	29	456
$7 < X \leq 8$	0	0	15	41	56
Summe	15	363	552	70	1000

Zweidimensionale Häufigkeitsverteilung für (X_g, Y_g)

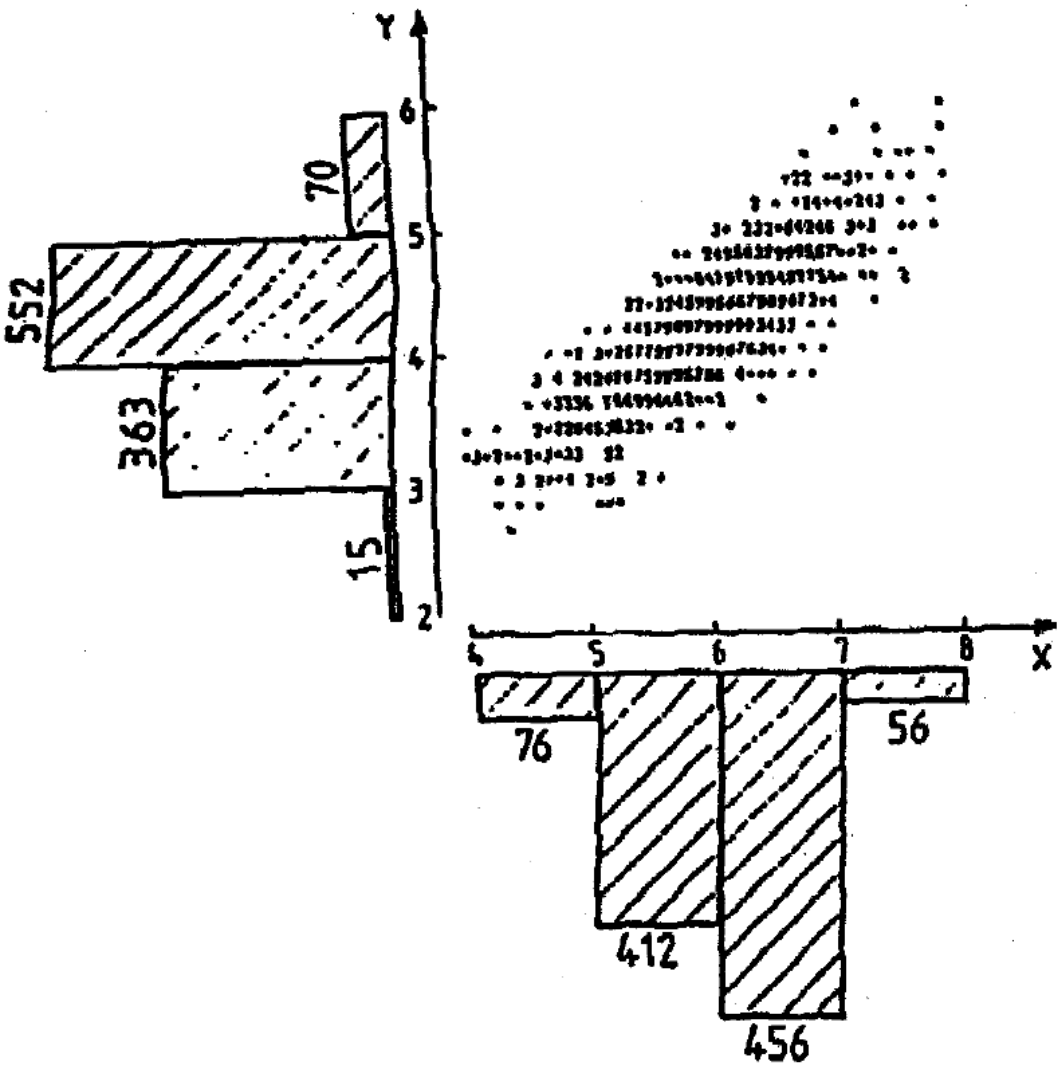
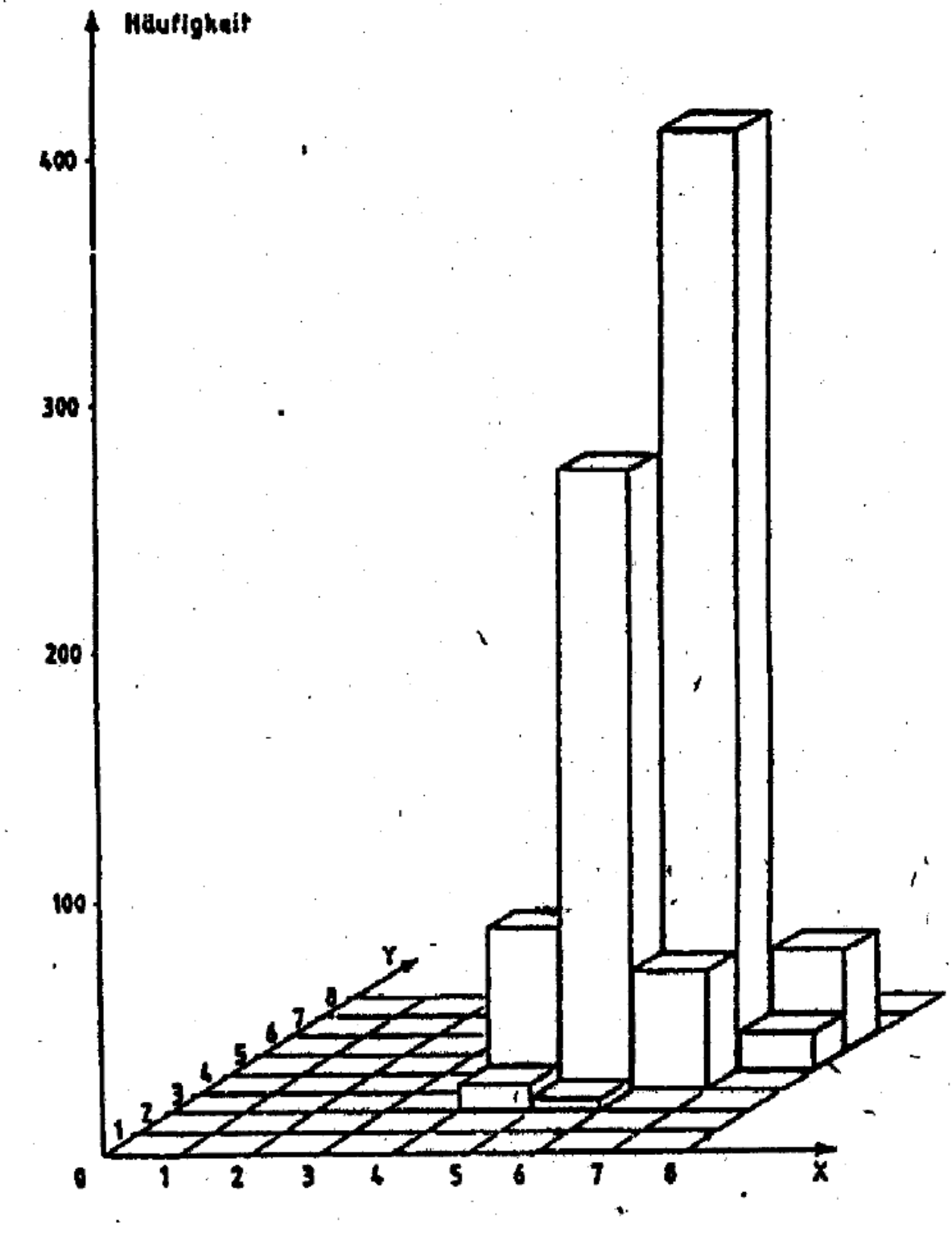


Abb.: Zweidimensionaler Plot


Häufigkeit
400
300
200
100
0
1
2
3
4
5
6
7
8
X
Y

Messung der Unabhängigkeit zweier Merkmale:

χ^2 - Kontingenzkoeffizient

$$h_n(X = a_i) = \frac{H_{i\cdot}}{n}$$

- relative Häufigkeit von "$X = a_i$"

$$h_n(X = a_i \mid Y = b_j) = \frac{H_{ij}}{H_{\cdot j}}$$

- bedingte relative Häufigkeit von "$X = a_i$" unter der Bedingung "$Y = b_j$".

Def.: Die Gesamtheit der bedingten relativen Häufigkeiten $\left(\frac{H_{ij}}{H_{\cdot j}}, i = 1, \ldots, k\right)$ heißt **bedingte relative Häufigkeitsverteilung unter der Bedingung** $Y = \delta_j$ (bzw. $Y \in K_j^*$).

Für das Bsp 1
(Mit Reaktionszeit (Geschlecht))

(0,25; 0,625; 0,125) –
bedingte Verteilung der
Reaktionszeit unter Bedingung,
dass das Geschlecht männlich ist

(0,33; 0.5; 0,17) – bedingte
Vert. der Reaktionszeit unter
der Bed., dass das Geschlecht
weiblich ist.

Bei Unabhängigkeit von X und Y:

Verteilung von X hängt nicht von Y ab

$\Updownarrow$ bzw.

die bedingte Verteilung von X $\approx$ der unbedingten Verteilung von X

$\Updownarrow$ bzw.

$$h_{i\cdot}\left(X=a_i \mid Y=b_j\right) \approx h_{i\cdot}\left(X=a_i\right)$$

$\Updownarrow$ bzw.

$$\frac{H_{ij}}{H_{\cdot j}} \approx \frac{H_{i\cdot}}{n}$$

$\Updownarrow$

$$H_{ij} \approx \frac{H_{i\cdot} \cdot H_{\cdot j}}{n}$$

$\forall\, i = 1, \ldots, k;$

$\forall\, j = 1, \ldots, \ell$

Bezeichnung:

$$H_{ij}^{(E)} := \frac{H_{i\cdot} \cdot H_{\cdot j}}{n}$$

"erwartete Zell-Häufigkeit bei Unabhängigkeit"

$$H_{ij}^{(B)} := H_{ij}$$ "beobachtete Zell-Häufigkeit".

Methode:

Wir stellen $H_{ij}^{(E)}$ und $H_{ij}^{(B)}$ gegenüber;

falls $H_{ij}^{(E)} \approx H_{ij}^{(B)}$ H_{ij}

$\Downarrow$

X, Y unabhängig.

Zurück zum Bsp 1:

T \ S	m		w		Σ
[100, 130)	$\frac{20}{24}$	(4)	$\frac{40}{36}$	(4)	60
[130, 160)	$\frac{50}{44}$	(6)	$\frac{60}{66}$	(6)	110
[160, 190)	$\frac{10}{12}$	(2)	$\frac{20}{18}$	(2)	30
Σ	80		120		200

$$H_{11}^{(E)} = \frac{H_{1\cdot} \cdot H_{\cdot 1}}{n} = \frac{80 \cdot 60}{200} =$$

$$= 24$$

$$H_{12}^{(E)} = \frac{H_{1\cdot} \cdot H_{\cdot 2}}{n} = \frac{60 \cdot 120}{200} =$$

$$= 36$$

$$H_{21}^{(E)} = \frac{H_{2\cdot} \cdot H_{\cdot 1}}{n} = \frac{110 \cdot 80}{200} =$$

..... für alle 6 Zelle.

Frage!

Wie sind die Differenzen (Quotienten) ehr zu schätzen?

groß?

klein?

(!)

Statistische Meßzahl?!

Def.:

$$\chi^2 = \mathrm{CHI}^2 = \sum_{j=1}^{l}\sum_{i=1}^{k}\frac{\left(H_{ij}^{(E)} - H_{ij}^{(B)}\right)^2}{H_{ij}^{(E)}}$$

heißt *χ^2-Kontingenzkoeffizient*

Bei Unabhängigkeit gilt:

$$\chi^2 \approx 0$$

Im Bsp 1:

$$\chi^2 = \frac{4^2}{24} + \frac{4^2}{36} + \frac{6^2}{44} +$$

$$+ \frac{6^2}{66} + \frac{2^2}{12} + \frac{2^2}{18} = 3{,}03$$

$\Rightarrow$ Groß? Klein?

Normiertes Maß!?

Def.:

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

Eigenschaften: $C = 0$ bei Unabhängigkeit

aber: $C < 1$ ⇒ C eignet sich nicht als Maß

Def.: $C_{korr} = C \cdot \sqrt{\frac{q^*}{q^* - 1}}$ mit $q^* = \min(l, k)$ heißt *Korrigierter Kontingenzkoeffizient*

Wir verwenden C_{koor}, um die Unabhängigkeit zweier Merkmale zu ermitteln!

Es gilt: $0 \leq C_{korr} \leq 1$

$C_{korr} \approx 0$ falls Unabhängigkeit

Auswertung von C_{korr}

$0 \leq C_{korr} \leq 0{,}2 \rightarrow$ Unabhängigkeit

$0{,}2 \leq C_{korr} \leq 0{,}5 \rightarrow$ schwach abhängigkeit

$0{,}5 < 0{,}5 \leq 1 \rightarrow$ abhängigkeit.

$$C = \sqrt{\frac{3,03}{203,03}}$$

$$C_{korr} = \sqrt{\frac{3,03}{203,03}}\sqrt{2} =$$

$$= 0,173$$

$$\Downarrow$$

T unabhängig von S

Beantworten Sie unter Benutzung nachfolgender Häufigkeitstabelle die Frage, ob das Wahlverhalten der Saarländer abhängig vom Geschlecht ist?
(Verwenden Sie als Maßzahl den korrigierten Kontingenzkoeffizienten!)

Häufigkeitsverteilung des Wahlverhaltens und des Geschlechts von 100 Saarländern:

Wahlverhalten/ Geschlecht	**männlich**	**weiblich**	
CDU/CSU	25	10	35
SPD	15	10	25
FDP	5	5	10
B90/Grüne	10	10	20
Sonstige	5	5	10
	60	40	100

$$H_{11}^{(\hat{E})} = \frac{H_{1.} \cdot H_{.1}}{100} = \frac{35 \cdot 60}{100} = 21$$

$$H_{12}^{(\hat{E})} = \frac{H_{1.} \cdot H_{.2}}{100} = \frac{35 \cdot 40}{100} = 14$$

$$H_{21}^{(\hat{E})} = \frac{25 \cdot 60}{100} = 15$$

$$H_{22}^{(\hat{E})} = \frac{25 \cdot 40}{100} = 10$$

$$H_{31}^{(E)} = \frac{10 \cdot 60}{100} = 6$$

$$H_{32}^{(E)} = \frac{10 \cdot 40}{100} = 4$$

$$H_{321}^{(E)} = \frac{60 \cdot 20}{100} = 12$$

$$H_{42}^{(E)} = \frac{40 \cdot 20}{100} = 8$$

$$H_{51}^{(E)} = \frac{10 \cdot 60}{100} = 6$$

$$H_{52}^{(E)} = 4$$

21	4	14	4
15	0	10	0
6	1	4	1
12	2	8	2
6	1	4	1
$H_{i1}^{(E)}$	$\lvert H_{i1}^{(E)} - H_{i1}^{(B)} \rvert$	$H_{i2}^{(E)}$	$\lvert H_{i2}^{(E)} - H_{i2}^{(B)} \rvert$

$$\chi^2 = \sum_i \sum_j \frac{\left(H_{ij}^{(t)} - H_{ij}^{(s)}\right)^2}{H_{ij}^{e}} =$$

$$= \frac{4^2}{21} + \frac{4^2}{14} + \frac{0^2}{15} + \frac{0^2}{10} +$$

$$+ \frac{1^2}{6} + \frac{1^2}{4} + \frac{2^2}{12} + \frac{2^2}{8} +$$

$$+ \frac{1^2}{6} + \frac{1^2}{4} =$$

$$= \frac{16}{21} + \frac{16}{14} + \frac{2}{6} + \frac{2}{4} + \frac{4}{12} + \frac{4}{8} =$$

$$= \frac{16}{21} + \frac{8}{7} + \frac{1}{3} + \frac{1}{2} + \frac{1}{3} + \frac{1}{2} =$$

$$= \frac{16}{21} + \frac{24}{21} + \frac{14}{21} + \frac{21}{21} =$$

$$= \frac{75}{21} = \frac{25}{7} \approx 3{,}57$$

$$q^* = 2 = \min\{5, 2\}$$

$$C_{Korr} = C \cdot \sqrt{\frac{q^*}{q^* - 1}} =$$

$$= \sqrt{2} \sqrt{\frac{3,57}{103,57}} =$$

$$= \sqrt{\frac{7,14}{103,57}} = 0,26$$

schwache Abhängigkeit