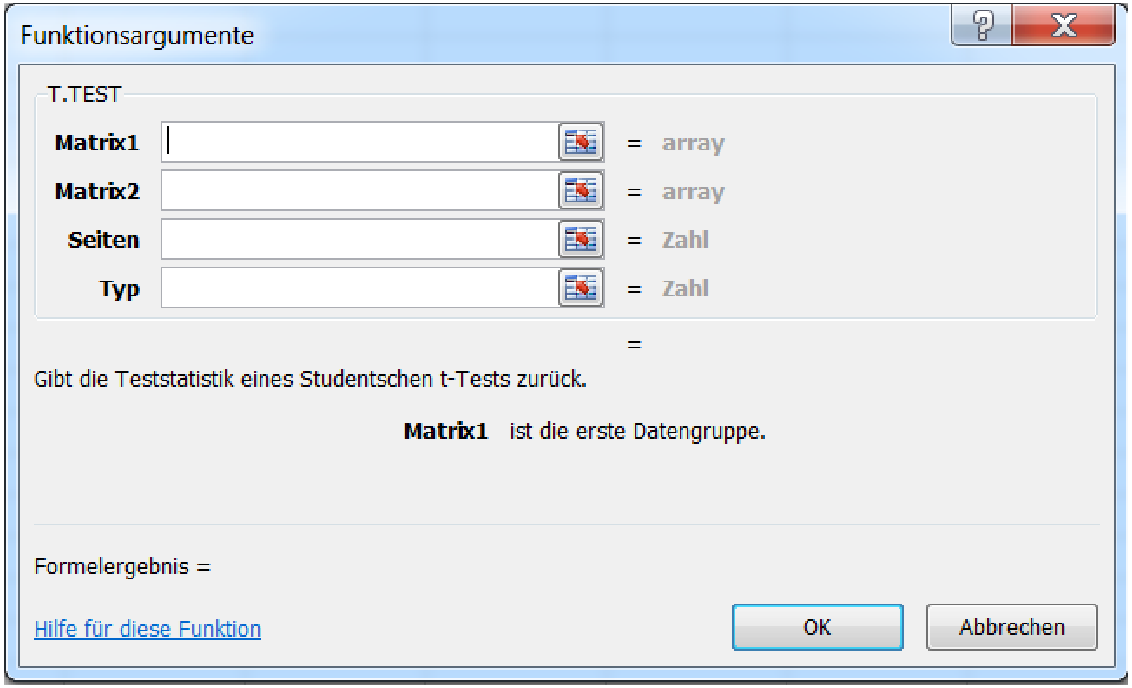
Allgemeine Schritte

1.: Frage -> Entscheidungsfrage (Ja/Nein)
H0: Nullhypothese oder Ansatz.
H0: Was wir als Daten bekommen werden,
gehört zu einer BEKANNTEn Verteilung.
2.: Festlegung des Signifikanzniveaus (Psign, Pkritisch,Signifikanz)
(Fürchten wir mehr falsche Annahme, oder ablehnen?)
3.: Experiment -> Daten Representativ, unverzerrt!
4.: Rechnen wir ein wenig ☺
Daten -> „ξ" : Parameter des Tests (ein Zahl)
„ ξ" -> P, die Wahrscheinlichkeit.
P ~ P(Daten | H0)
5.: Entscheidung
wenn P<Pkrit , H0 ablehnen (Fehler Typ I möglich)
wenn P>=Pkrit, H0 behalten (Fehler Typ II möglich)



=T.TEST( )

Matrix1 ist die erste Datengruppe.

Matrix2 ist die zweite Datengruppe.

Seiten bestimmt die Anzahl der Endflächen.

Typ bestimmt die Form des durchzuführenden t-Tests.

Parameter

Ist Typ gleich	Wird folgender Test ausgeführt
1	Gepaart ← **Einstichproben t-Test**
2	Zwei Stichproben, gleiche Varianz (homoskedastisch)
3	Zwei Stichproben, ungleiche Varianz (heteroskedastisch)

Test auf Varianzgleichheit: *F*-test

Nullhypothese: Die Varianzen sind gleich

Parameter: $F = \dfrac{s_1^2}{s_2^2}$; $s_1 > s_2$

Bei der Gültigkeit der Nullhypothese *F* folgt eine *F*-Verteilung mit n_1-1 und n_2-1 Freiheitsgrade

Bemerkung: Tabelle zum einseitigen Test
wir brauchen einen zweiseitigen Test

$$F = \frac{s_1^2}{s_2^2}$$

der kritische Wert (aus der Tabelle)

$$F < F_{n_1-1,n_2-1;5\%}$$

wir verwerfen die *Nullhypothese* nicht
d.h. die Varianzen sind gleich

$$F > F_{n_1-1,n_2-1;5\%}$$

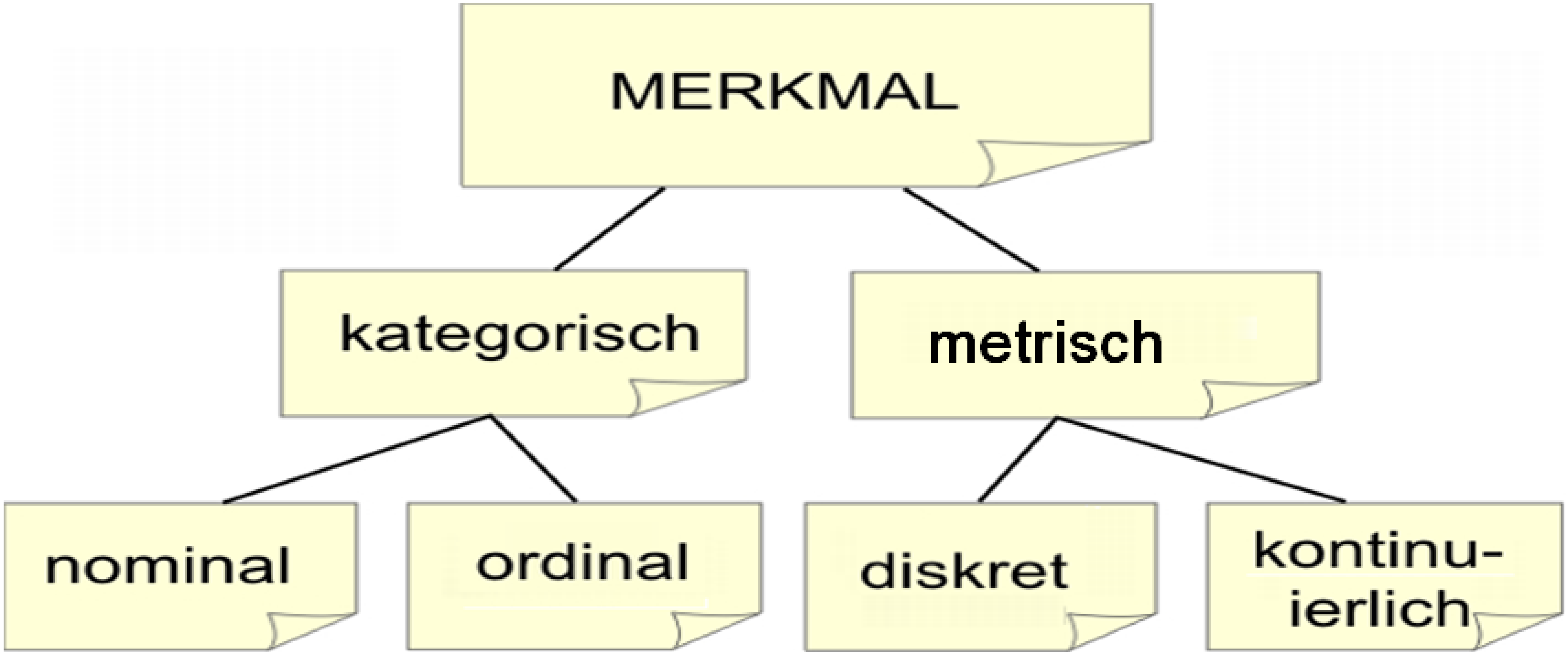
wir verwerfen die *Nullhypothese*
d.h. die Varianzen sind nicht gleich

Viel einfacher:

pF = F.TEST(daten1;daten2;2)

Wenn pF<5% dann ungleicher Var (also Typ=3 bei der T.TEST)
sonnst: gleicher Var. (Typ=2)

Klassifizierung der Merkmale



Übersicht der Testmethode

Verteilung / *Stichproben*	*normalverteilte Daten*	*die Verteilung der Daten ist unbekannt*
eine Stichprobe	Einstichproben t-Test	Wilcoxon Test
zwei Stichproben	Zweistichproben t-Test	Mann-Whittney U-Test
mehrere Stichproben	ANOVA (Varianzanalyse)	Kruskal-Wallis Test

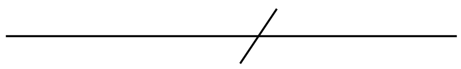
Nichtparametrische Methoden

- **Bedingungen der *t*-Tests**
- kontinuierliches Merkmal (z.B. Körperhöhe, Körpertemperatur...)
- die Daten müssen normalverteilt sein

Nichtparametrische Methoden

- nur ordinale Daten (Ordinalskala)
- keine Normalverteilung (auch bei unbekannter Verteilung möglich)

z. B. Schmerzmittel – wie es schmerzt?
Kann nur auf einer ordinalen Skala gemessen werden:

1, 2, 3, 4, 5 oder 

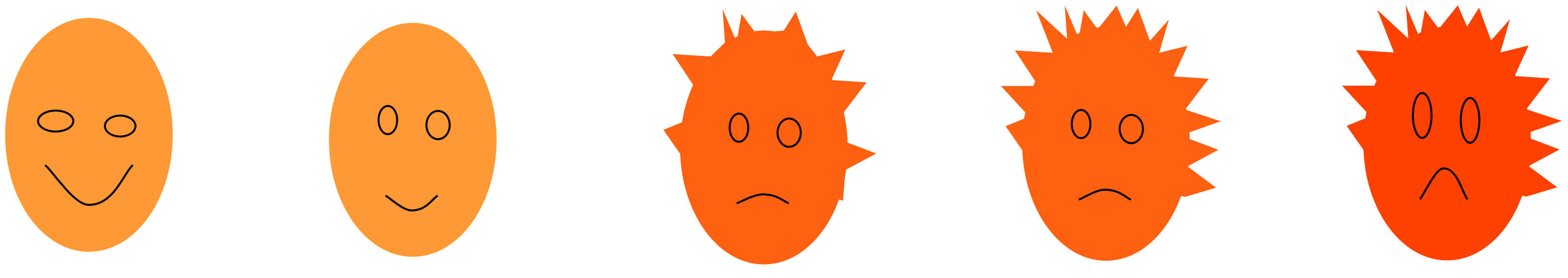
- **Vorteile:**
- Verteilungsunabhängigkeit
- Ordinal-, Intervall-, Verhältnisskalen

- **Nachteile:**
- Datenreduktion, Informationsverlust
- größere Wahrscheinlichkeit der Fehler 2. Art:
- nur größere Unterschiede können statistisch bewiesen werden

Prinzip der Rang-Tests

- Rang: Position eines Wertes innerhalb einer nach der Größe sortierten Wertereihe

z.B. Kopfschmerzen:



1 2 3 4 5

Mit Hilfe der Ränge führt man eine Gleichverteilung ein!

Rang Test Methode – Verbundene Ränge

Wenn zwei oder mehrere ursprüngliche Daten gleich sind:

originale Daten	3, 7, 1, 13, 13, 16
geordnete Daten	1, 3, 7, 13, 13, 16
Ränge	1, 2, 3, 4.5, 4.5, 6

Verbundene Ränge:

die bekommen den Durchschnittsrang

Durchschnitt der Ränge

• In steigende Reihe geordnete Daten:	x_1 , x_2 … $x_{(n-1)/2}$, $x_{(n+1)/2}$ …. x_{n-1} ,x_n
• Ränge:	1, 2 … (n-1)/2 (n+1)/2 … n-1, n

- (n ist ungerade)
- Durchschnitt der Ränge: $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$

- Durchschnittlicher Rang = Rang des Medians

- Wenn n ist gerade:
- Median= $(x_{n/2} + x_{n/2+1})/2$
- Durchschnittlicher Rang= (n+1)/2

Rangteste testen den Median!

Eine Stichprobe: Wilcoxon-Vorzeichen Rangtest

- eine Stichprobe (gepaarte Daten)
- ordinale Daten
- Ist der Median der Datenreihe gleich Null?
- (oder ein bestimmter Wert)?
- H_0: Der Median der Daten ist Null
  (oder ein bestimmter Wert).
- Die Ränge bekommen Vorzeihen.
- Der Durchschnitt der Ränge wird geprüft.
- Wenn die Nullhypothese gültig ist, es sind gleich viele und gleich große positive und negative Ränge, Durchschnitt der Ränge ist Null!

Wilcoxon-Vorzeichen Rangtest: Einführung mit einem Beispiel

- Überlebenszeit der Ratten:
- 168, 150, 280, 221, 230, 165, 179, 250, 195, 276
- Ist der **Median** der Überlebenszeiten unterschiedlich von 170 Tage?
- H_0: Der **Median** der Überlebenszeiten beträgt 170 Tage. **(H0: Median = REF.)**
- **Überlebenszeitenunterschiede der Ratten im Vergleich zur 170 Tage:**
- **-2, -20, +110, +51, +60, -5, +9, +80, +25, +106**
- **Geordnet nach Betrag der Änderung:**
- **-2, -5, +9, -20, +25, +51, +60, +80, +106, +110,**
- **Ränge (nach betrag der Änderung):**
- **1, 2, 3, 4, 5, 6, 7, 8, 9, 10**
- **Ränge mit Vorzeichen:**
- **-1, -2, +3, -4, +5, +6, +7, +8, +9, +10**

Durchschnitt: 4.10
Standardabw.: 4.91

Wilcoxon Vorzeichen Rangtest: Beispiel der Überlebenszeiten der Ratten

- Der Durchschnitt folgt einer Normalverteilung, wenn genug viele Daten sind (Zentraler Grenzwertsatz)
- Anwendung der t-Verteilung (Annäherung!):

$$t_{n-1} = \frac{\overline{R}}{\frac{s}{\sqrt{n}}}$$

$\overline{R}$: Durchschnitt der Ränge

s: Standardabweichung der Ränge

n: Anzahl der Daten

$n-1$: Freiheitsgrad

Entscheidung: wie beim Einstichproben ***t***-Test

Ränge mit Vorzeichen:

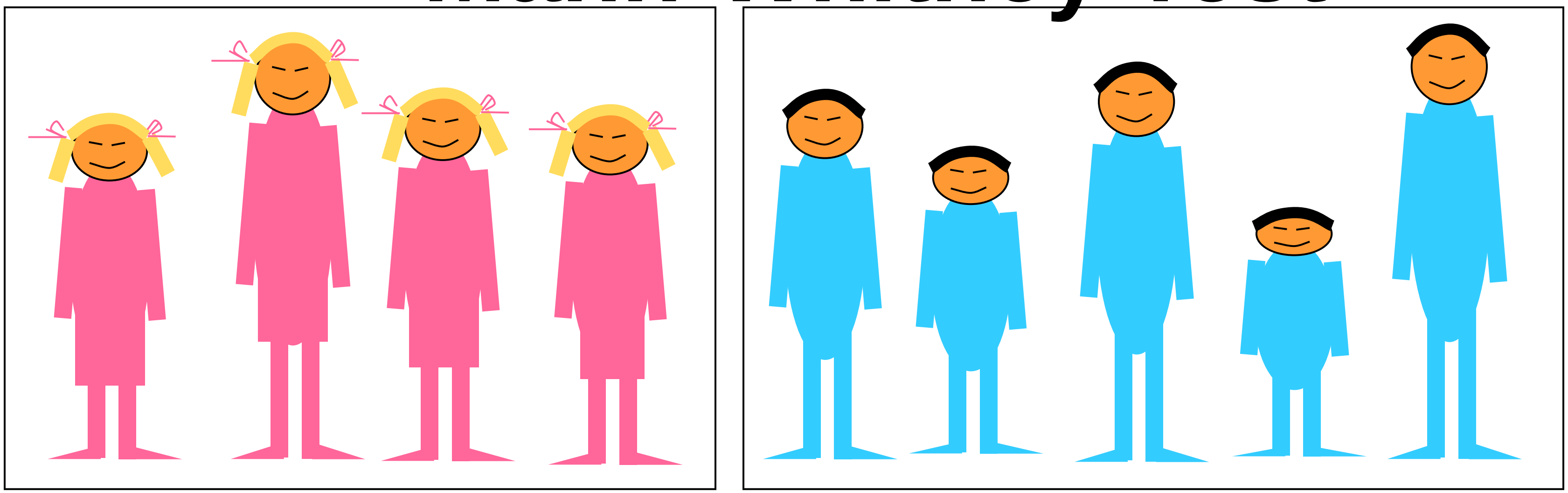
-1, -2, +3, -4, +5, +6, +7, +8, +9, +10 $\longrightarrow$ Durchschnitt: 4.10, Standardabw.: 4.91

$$t_9 = \frac{4{,}10}{4{,}91/\sqrt{10}} = 2{,}64 \quad \Rightarrow \quad t_9 > t_{9;5\%} \quad \Rightarrow \quad H_0 \text{ is abgelehnt}$$

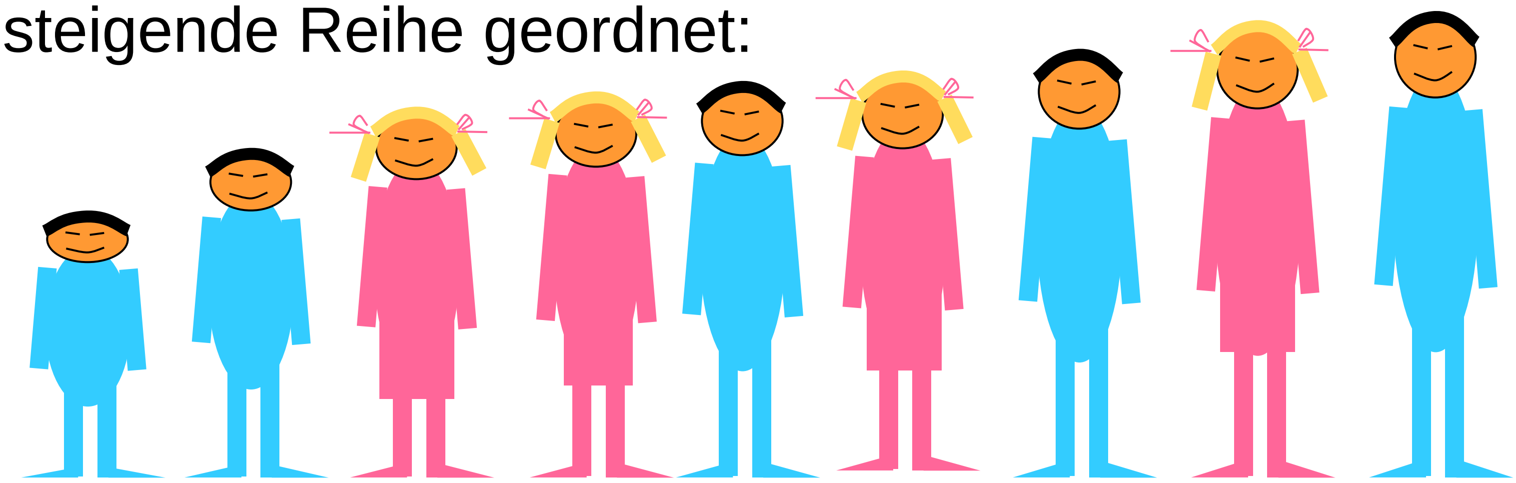
$t_{9;5\%} = 2{,}26$ (aus der Tabelle)

$p < 5\%$ (mit Excel)

Vergleich von zwei Stichproben: Mann-Whitney Test

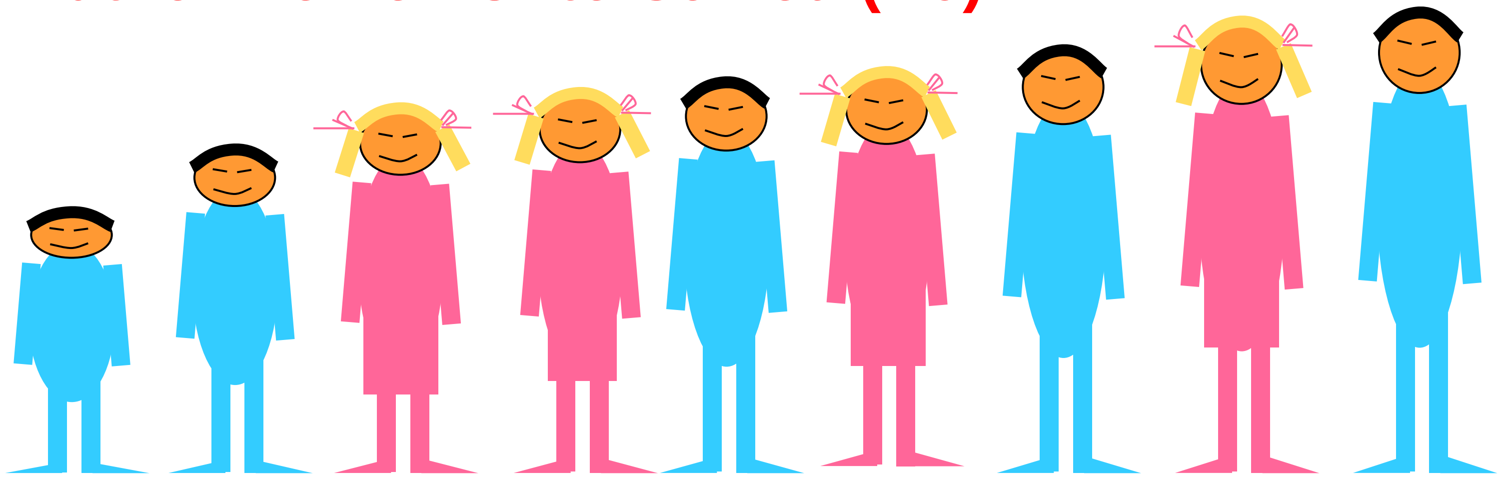


In steigende Reihe geordnet:

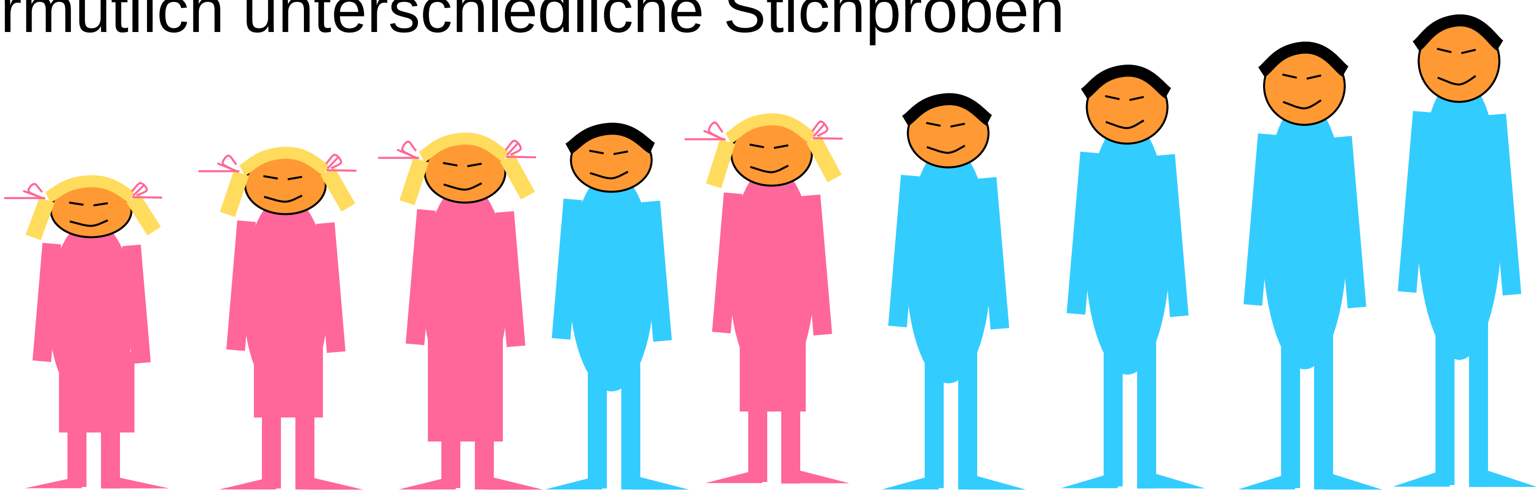


Ränge: 1 2 3 4 5 6 7 8 9

vermutlich keinen Unterschied (H0)



vermutlich unterschiedliche Stichproben



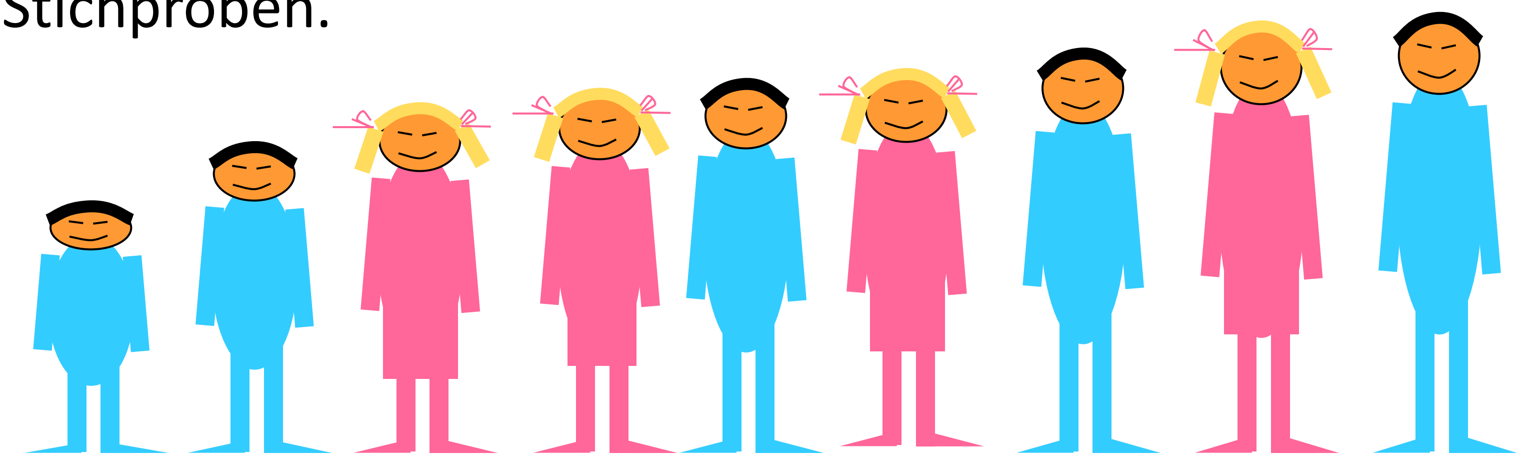
1 2 3 4 5 6 7 8 9

Wie unterschiedlich sollten die sein?? H1 ist nicht testbar…

Mann – Whitney *U* Test (Annäherung)

- (Auch als Wilcoxon Rank Summe Test genannt)
- Vergleich von zwei Stichproben (n_1, n_2)
- H_0: Die zwei Stichproben stammen aus der selben Grundgesamtheit

1. Zuordnung der Ränge der in den zwei zusammengeordneten Stichproben.



Ränge: 1 2 3 4 5 6 7 8 9

2. Bestimmung die Summen der Ränge in eine Gruppe: T_1.

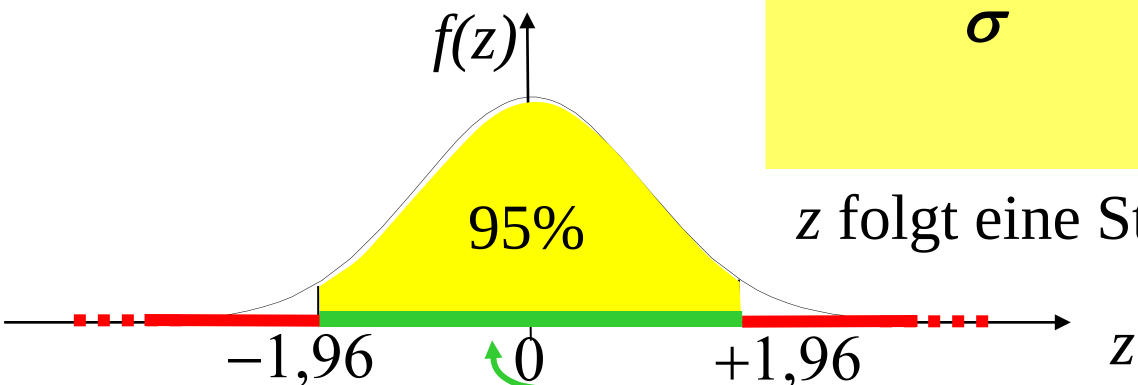
- T_1= 1+2+5+7+9=24

Mann – Whitney *U* Test: Annäherung

- Bei Gültigkeit der Nullhypothese folgen die Daten der Gruppe 1 eine Gleichverteilung, mit möglichen werten von $1 \ldots n_1+n_2$)
- Erwartungswert und die theoretische Streuung von T_1 können berechnet werden:

$$\mu = \frac{n_1(n_1+n_2+1)}{2} \qquad \sigma = \sqrt{n_1 n_2 (n_1+n_2+1)/12}$$

$$z = \frac{T_1 - \mu}{\sigma} = \frac{T_1 - \frac{n_1(n_1+n_2+1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}$$



z folgt eine Standard-Normalverteilung (wenn H_0 gültig ist)

z.B. $T_1=24$, $n_1=5$, $n_2=4$ => $\boxed{z = -0{,}245}$ => H_0 wird angenommen

Mit der erweiterten Excel Funktionen

WILCOXON_TEST(Matrix1;Matrix2;Seiten;Typ)

Wilcoxon-Vorzeichen-Rang-Test. Auf Grund exakter Rangsummen-, oder t-Verteilung gibt Wahrscheinlichkeiten zurück.

WILCOXON_TEST

Matrix1 =

Matrix2 =

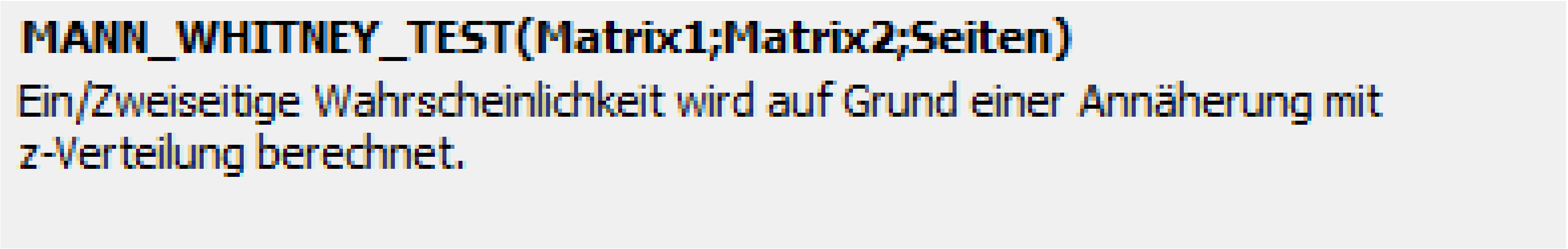
Seiten =

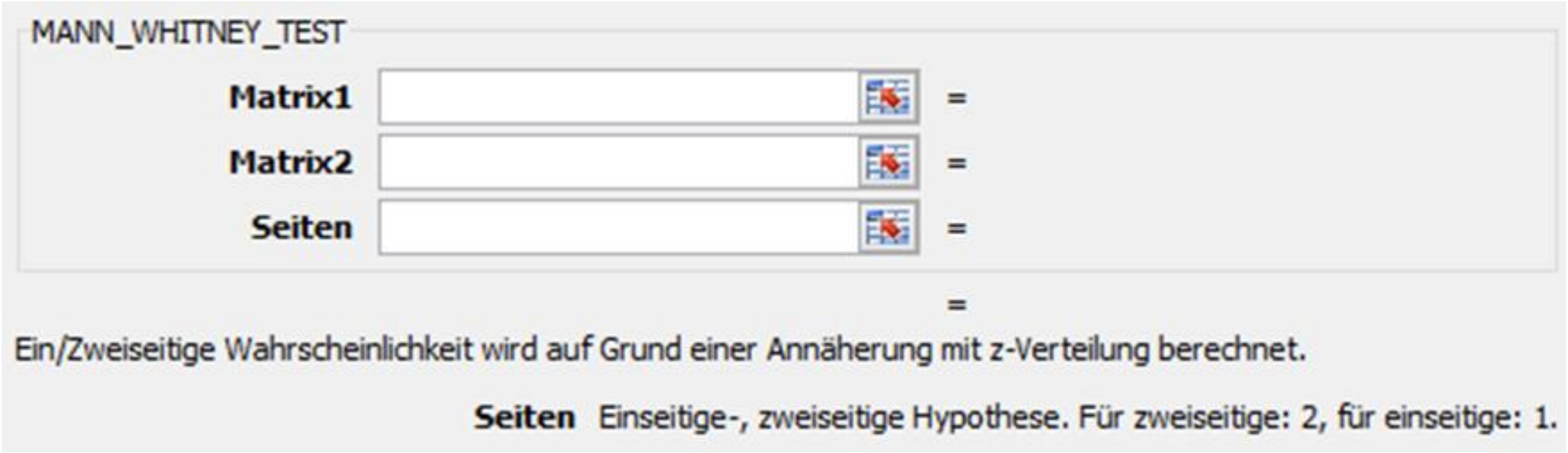
Typ =

=

Wilcoxon-Vorzeichen-Rang-Test. Auf Grund exakter Rangsummen-, oder t-Verteilung gibt Wahrscheinlichkeiten zurück.

Typ Exakt oder t-Verteilung; exakt:TRUE, t-Verteilung:FALSE.

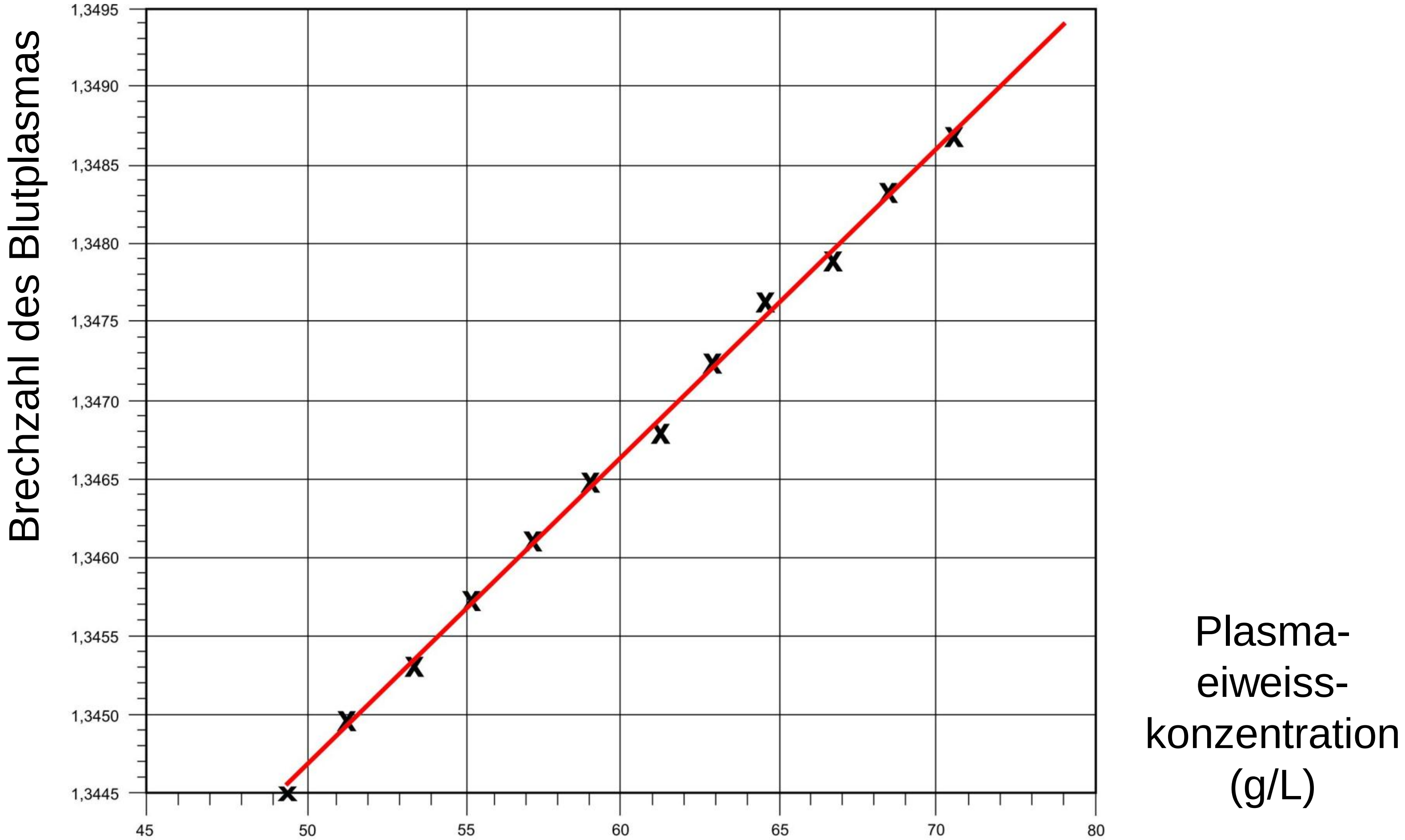




Die Funktion liefert direkt den P-Wert.

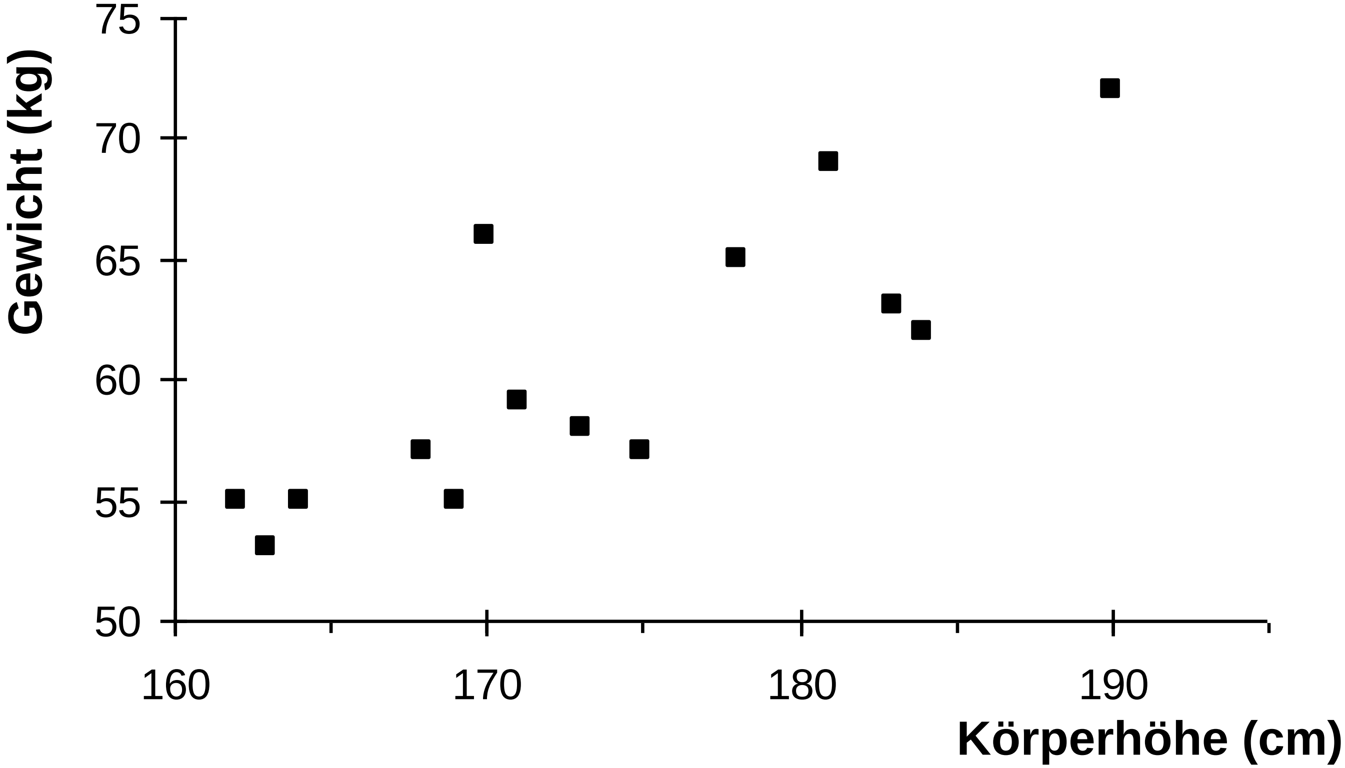
Korrelation

Gibt es ein Zusammenhang?



Brechzahl des Blutplasmas
1,3495
1,3490
1,3485
1,3480
1,3475
1,3470
1,3465
1,3460
1,3455
1,3450
1,3445
45
50
55
60
65
70
80
Plasma-
eiweiss-
konzentration
(g/L)

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



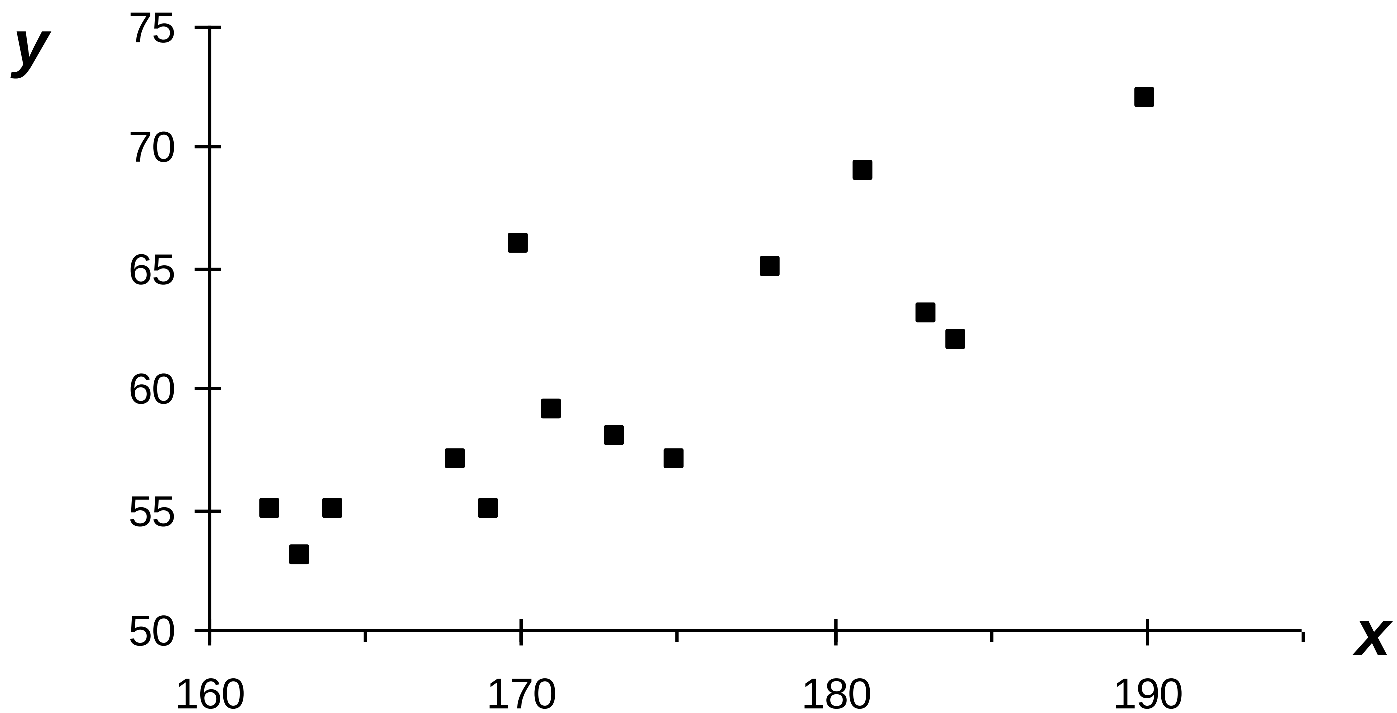
cm	kg
162	55
163	53
164	55
168	57
169	55
170	66
171	59
173	58
175	57
178	65
181	69
183	63
184	62
190	72

was für eine Tendenz kann man bemerken?

Die Korrelationsrechnung beschäftigt sich mit dem symmetrischen Zusammenhang zweier Zufallsgrössen

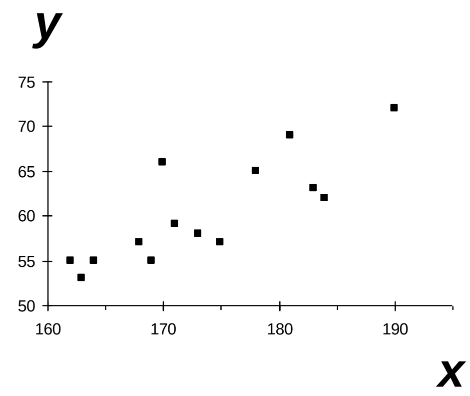
positive Korrelation: je mehr, desto mehr

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. $y = a + b_1x + b_2x^2 + \ldots + b_nx^n + h$

exponentiale F. $y = ab^x h$

Potenzfunktion $y = ax^b h$

und wie wirkt der Zufall auf die abhängige Variable

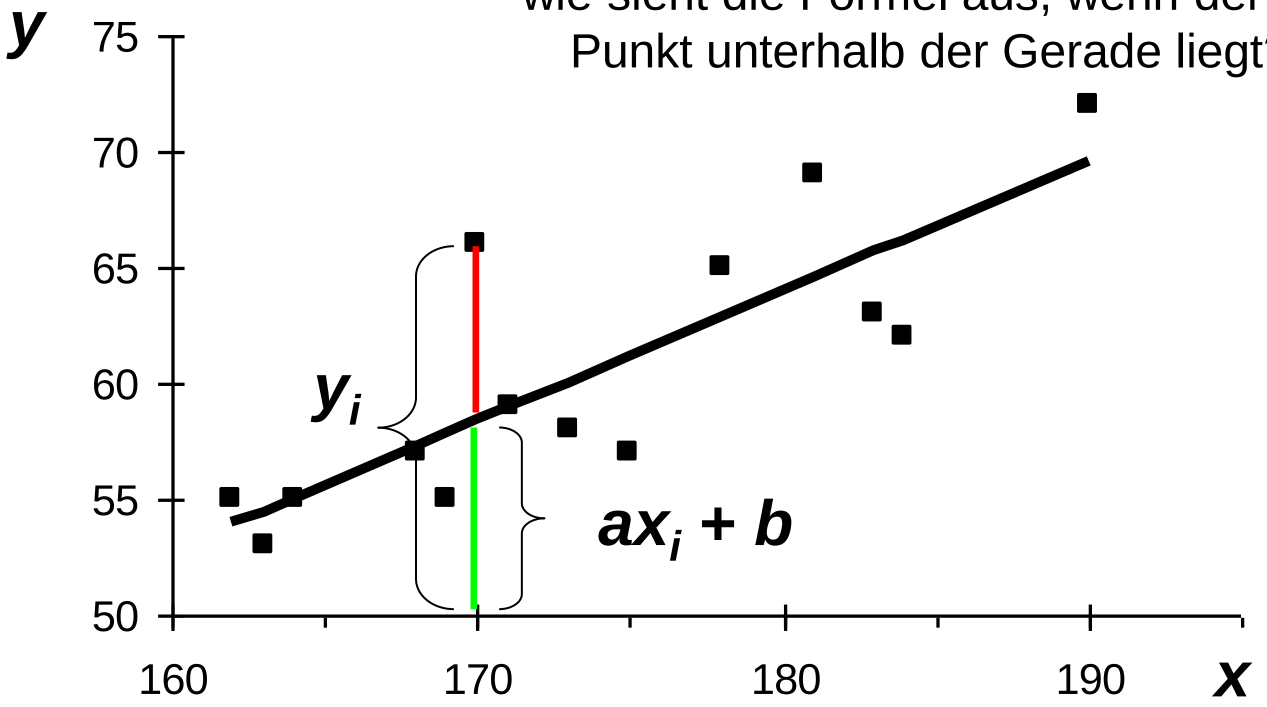
additiver Fehler ($+ h$) oder multiplikativer Fehler ($\cdot h$)

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?



	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

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

„Die beste” Steigung: $(y = ax + b)$

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

oder $a^* = \frac{s_{xy}^2}{s_x^2}$

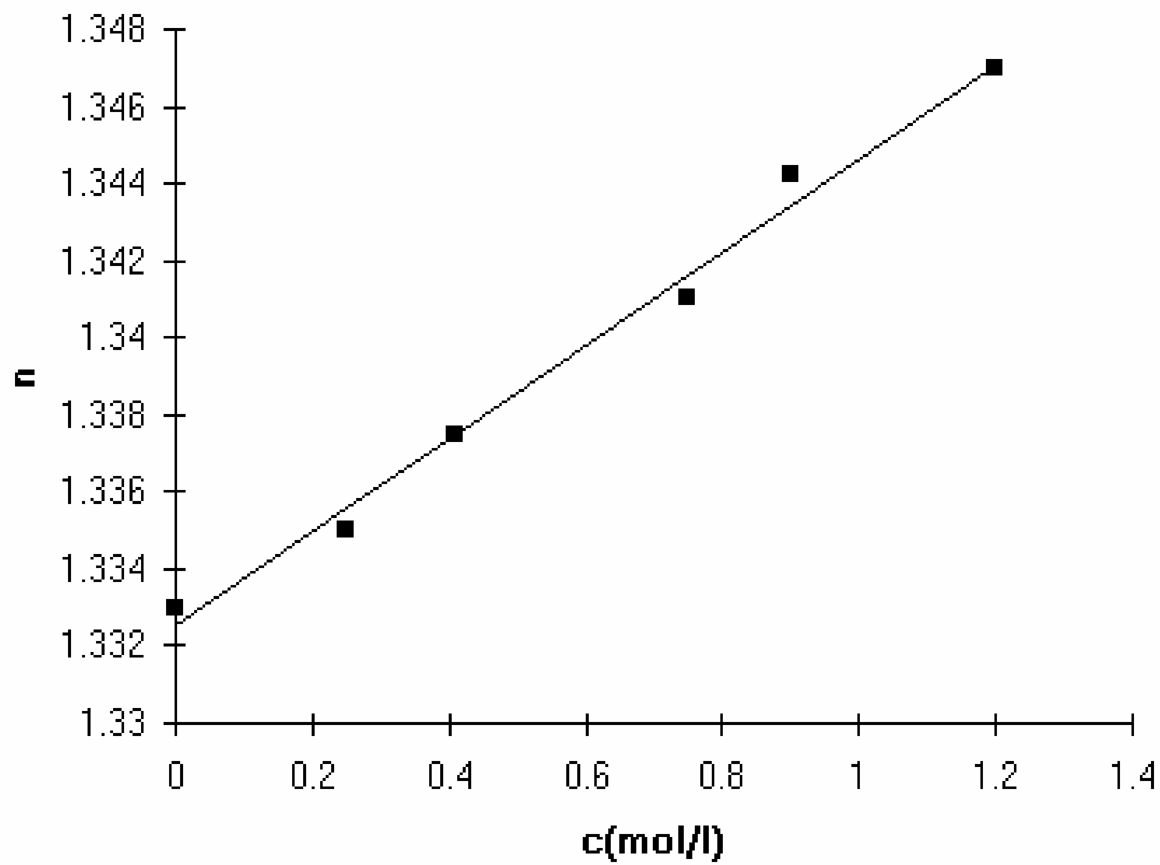
„Der beste” Achsenabschnitt:

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

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

Beispiel: Refraktometrie

	A	B	C
1–7	unbekannte	Konzentration	Brechzahl
8	i	c (mol/l)	n
9		0	1.333
10		0.25	1.3350
11		0.41	1.3375
12		0.75	1.3410
13		0.90	1.3442
14		1.20	1.3470
15	4	0.19	1.3348
16	10	1.13	1.3462
17	n=ac+n0		
18	n0=	1.3325	
19	a=	0.0121	
20	r=	0.995	



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)

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

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

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

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

$$-1 \le r \le 1$$

weitere Bemerkungen:

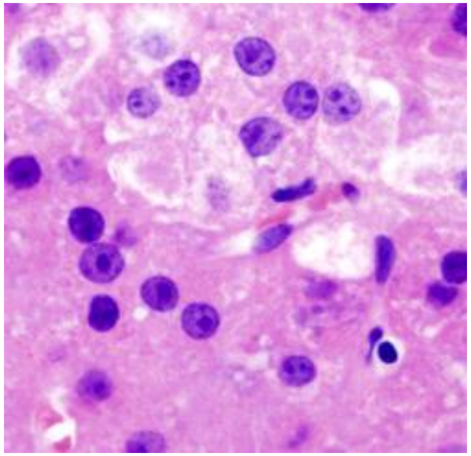
$$-1 \leq r \leq 1$$

Korrelationskoeffizient
(Pearson)

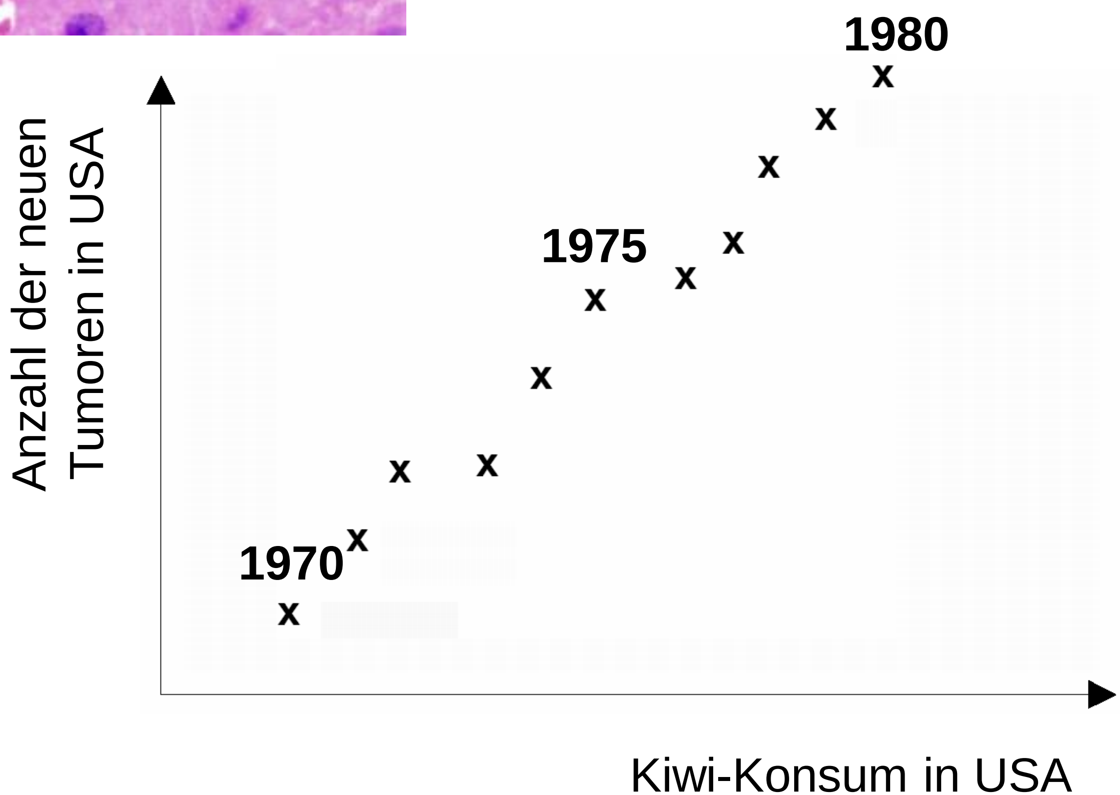
$$0 \leq r^2 \leq 1$$

Bestimmtheitsmass
(coefficient of determination)

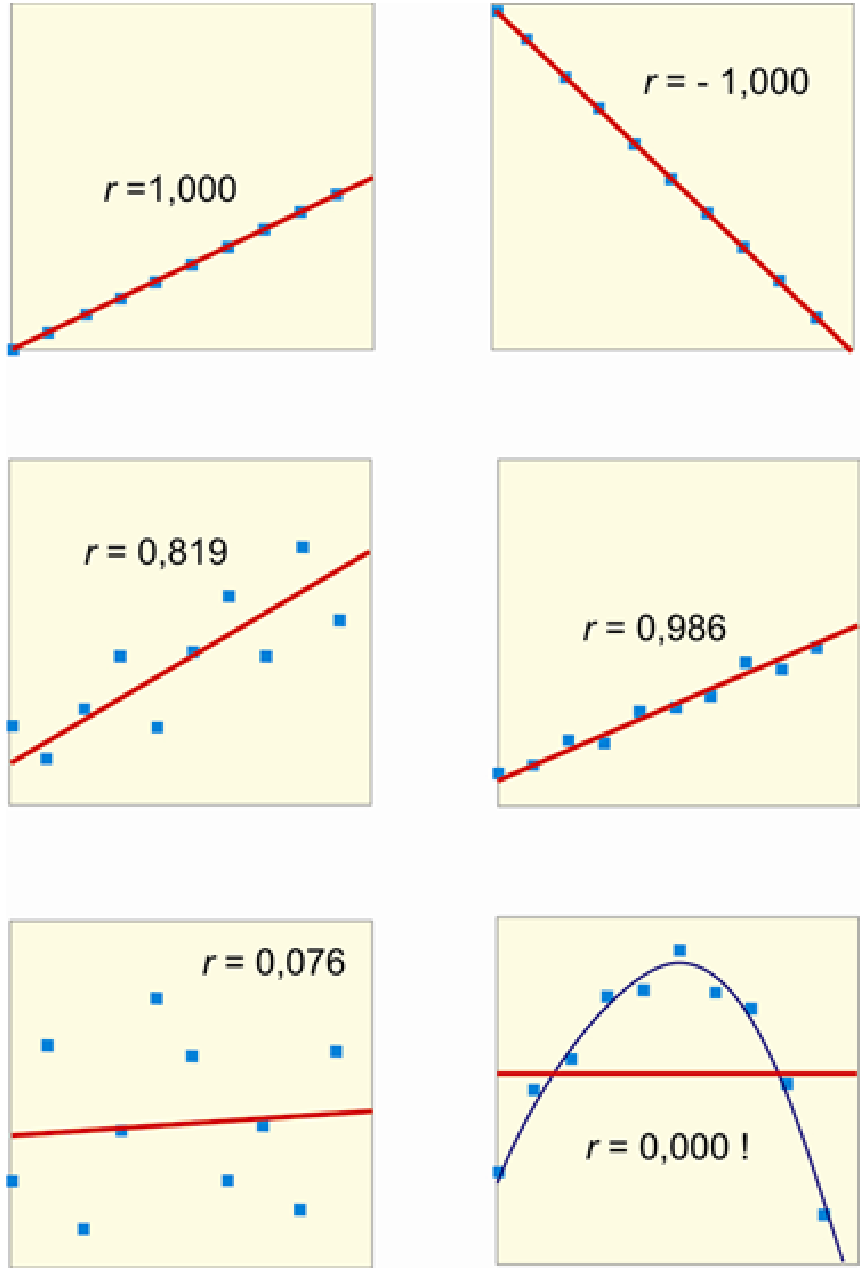
Die Korrelation beschreibt nicht unbedingt eine Ursache-Wirkungs-Beziehung in die eine oder andere Richtung.

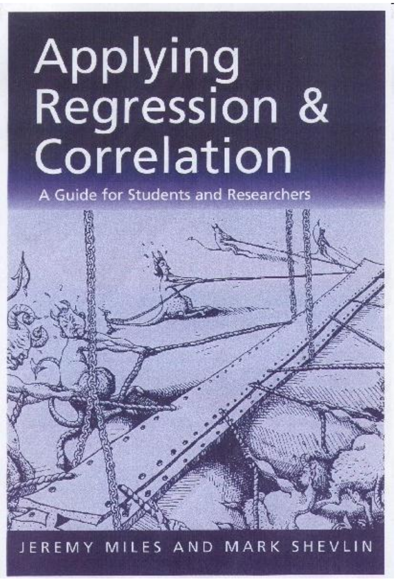


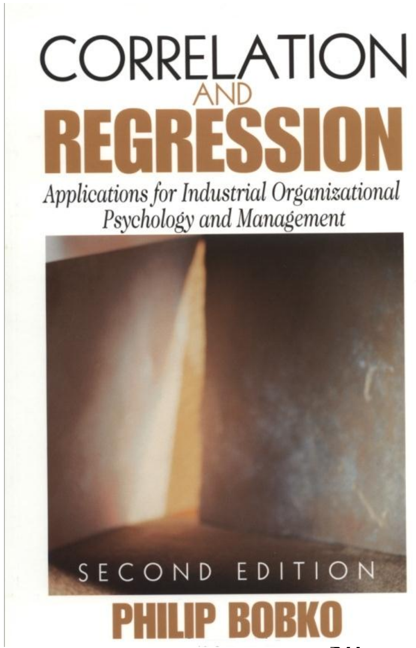
Korreliert heisst nicht notwendigerweise kausal verknüpft(!)



Beispiele:

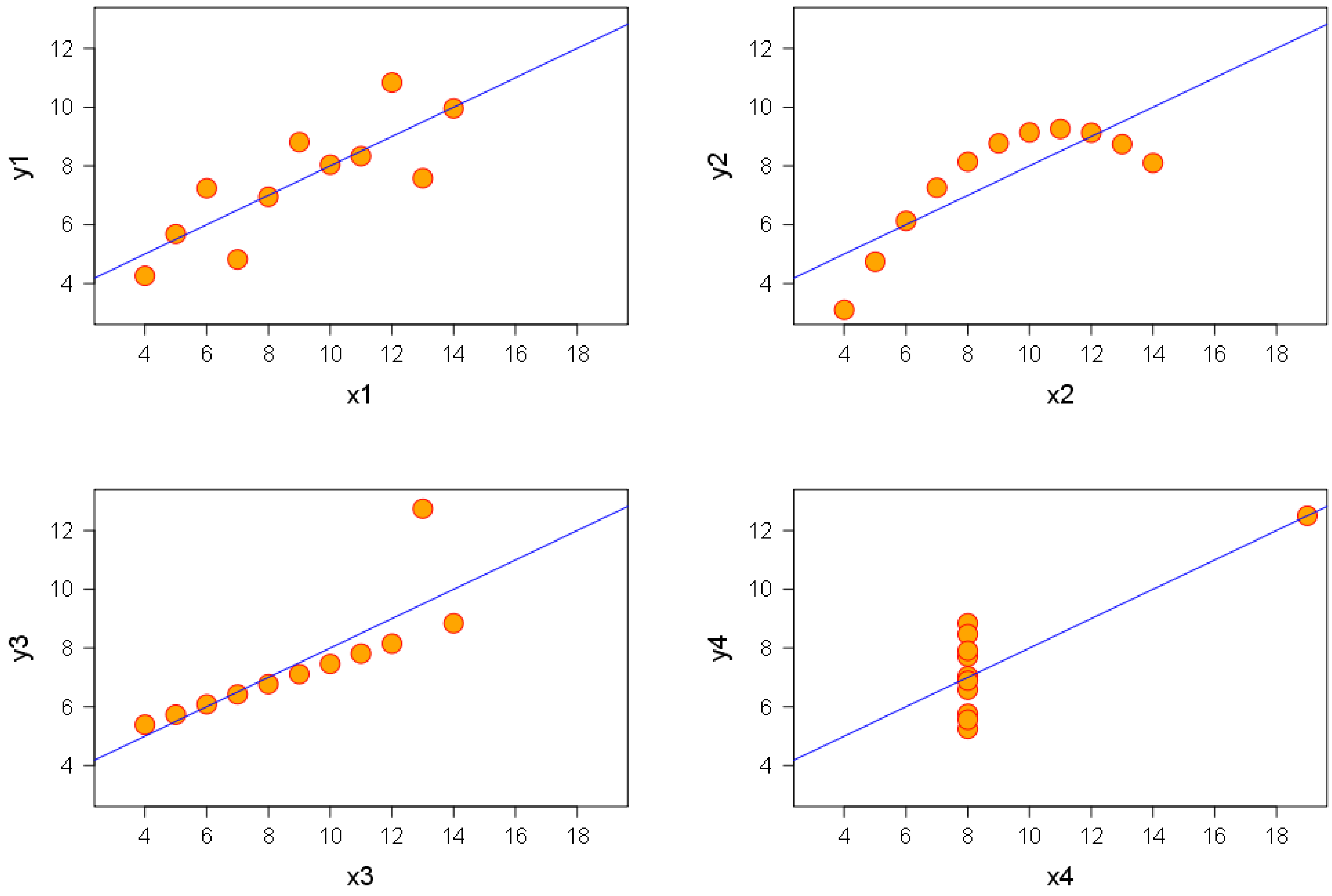






Pr.Buch Abb. 15

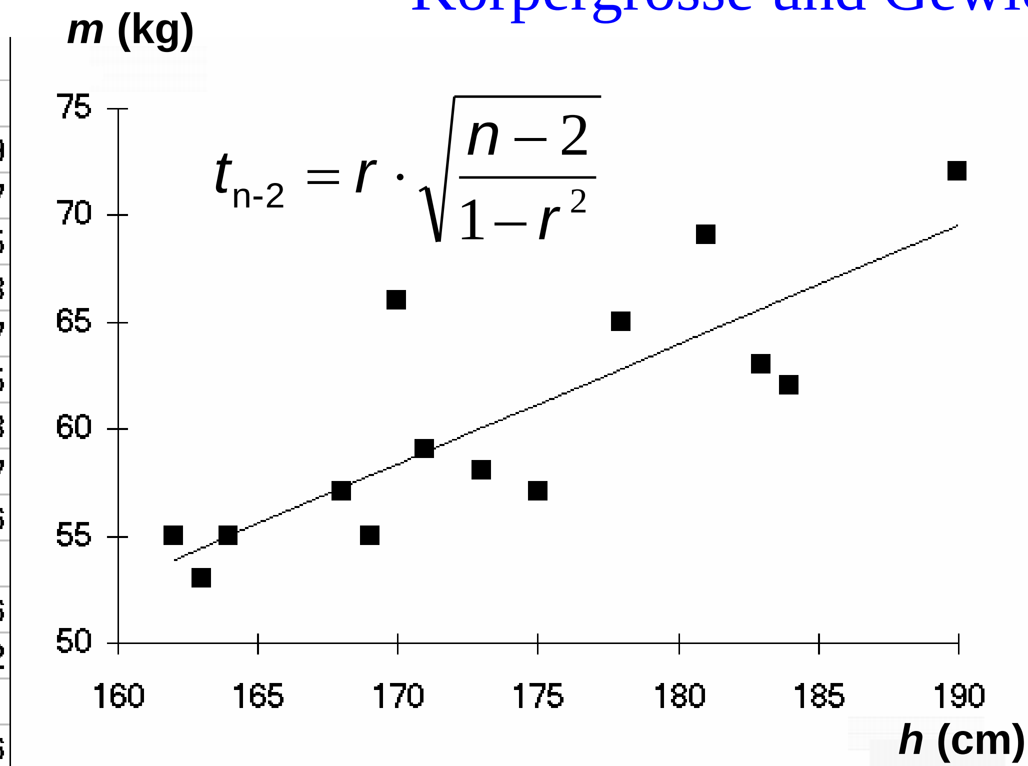
Extrembeispiel: r=0.816, y = 3 + 0.5x (Anscombe's quartet)



t-Test zur Korrelationsanalyse

Gibt es eine Beziehung zw. der Körpergrösse und Gewicht?

Körperhöhe (cm)	Gewicht (kg)			
162	55	53.929		
163	53	54.487		
164	55	55.045		
168	57	57.278		
169	55	57.837		
170	66	58.395		
171	59	58.953		
173	58	60.07		
175	57	61.186		
178	65	62.861		
181	69	64.536		
183	63	65.652		
184	62	66.211		
190	72	69.56		
	m=	0.5583	-36.50955	=b
		0.1131	19.66358	
r=	0.818505	0.6699	3.492297	
n=	14	24.358	12	
t=	6.030	297.07	146.3537	



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

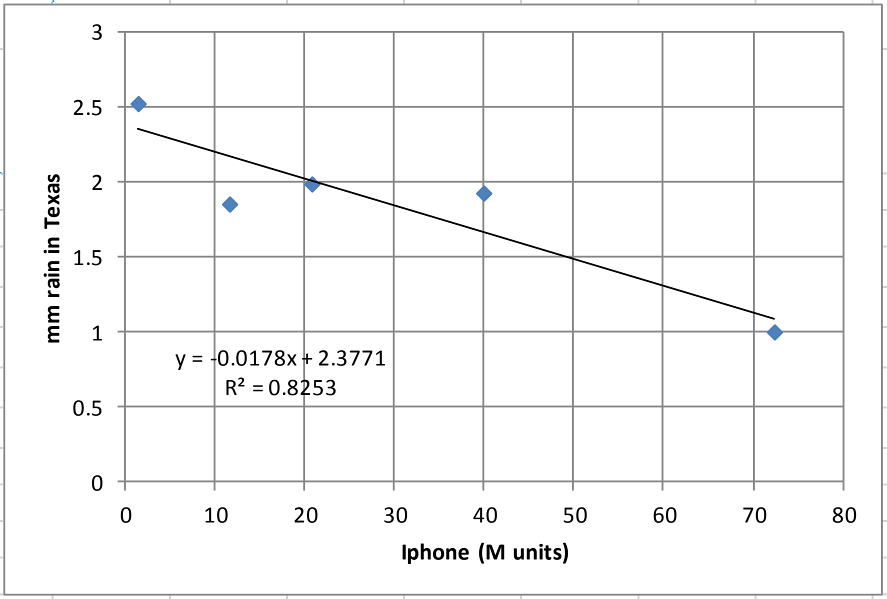
$$|t| = 6.030 > t_{12,\,\mathrm{krit}(0{,}01)} = 3.055 \Rightarrow H_0 \text{ ist falsch } (p<0.01)$$

Korrelation heisst noch lange nicht Uhrsache!!!

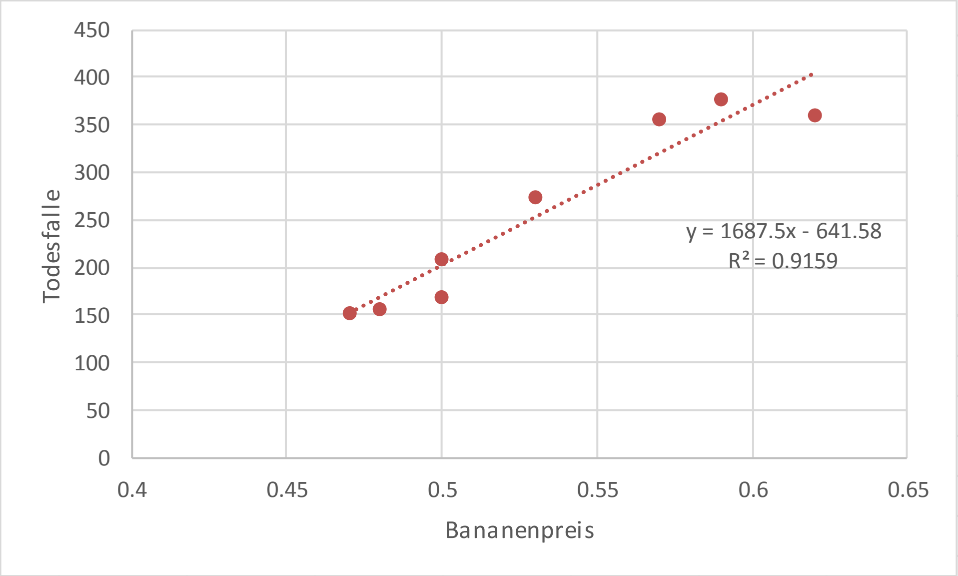
	2007	2008	2009	2010	2011
Apple iPhone sales *Millions of units ()*	1.39	11.63	20.73	39.99	72.29
Precipitation in Texas *Avg Daily Precipitation (mm) (CDC)*	2.52	1.85	1.99	1.93	1

1.39	2.52
11.63	1.85
20.73	1.99
39.99	1.93
72.29	1

R	-0.90846
n	5
t	-3.76471
P	0.032784
	p<0.05



	1999	2000	2001	2002	2003	2004	2005	2006
Cost of bananas (unadjusted) *Dollars per pound (Bureau of Labor)*	0.5	0.47	0.48	0.5	0.53	0.62	0.57	0.59
People who died by falling out of their wheelchair *Deaths (US) (CDC)*	169	154	157	209	274	360	356	377



http://tylervigen.com/discover