3. Beschreibung von mehrdimensionalen (bivariate) Stichproben

3.1 Grafische Darstellungen einer Stichprobe von Zahlenpaaren

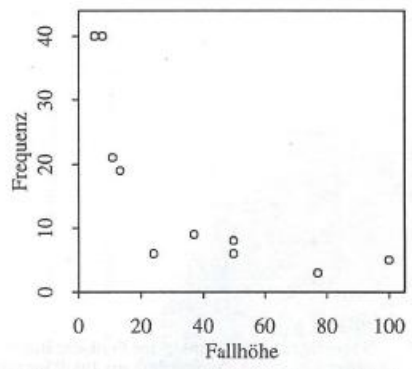
Werden bei einer Stichprobe zwei verschiedene Merkmale x_i und y_i beobachtet, so spricht man von einer **bivariaten Stichprobe**. Dabei stellt sich die Frage, ob es einen Zusammenhang, eine Abhängigkeit zwischen den beiden Grössen gibt. In diesem Fall sagt man, dass die beiden **Variablen korreliert** sind.

Beispiele:
Druck und Volumen bei einem physikalischen Experiment, Spraydosen-Treibgasverbrauch und mittlerer Ozongehalt der Atmosphäre in einem Jahr, Grösse des Vaters und des Sohnes bei einem Vater-Sohn-Paar.

Die Daten (x_i, y_i) können als Punkte in der Koordinatenebene dargestellt werden. Diese Darstellung heisst **Streudiagramm** (englisch: **Scatterplot**). Sie erleichtert es, bei Zahlenpaaren Trends, Muster, Ausreisser zu erkennen.

Beispiel:
Das Rauschen von Wasserfällen (Quelle: Science 164,1969, p. 1513-1514)

Bei verschiedenen Wasserfällen wurde die Höhe und die dominierende Frequenz im Spektrum der Bodenvibrationen gemessen. Im Streudiagramm ist ersichtlich, dass hohe Frequenzen mit niedrigen Höhen einhergehen und umgekehrt.



Name	x_i Höhe (m)	y_i Frequenz (Hz)
Lower Yellowstone	100	5
Yosemite	77	3
Canadian Niagara	50	6
American Niagara	50	8
Upper Yellowstone	37	9
Gullfoss (Lower)	24	6
Firehole	13.3	19
Godafoss	10.9	21
Gullfoss (Upper)	7.7	40
Fort Greeley	5.2	40

Bild 3.1.d Streudiagramm für Fallhöhe und Schwingungsfrequenz bei Wasserfällen

Sind mehr als zwei Variablen von Interesse, dann wird - wie im folgenden Beispiel - in der so genannten Scatterplot-Matrix jeweils für zwei der vier Aktien ein Streudiagramm abgebildet.

Beispiel:
Es sind die Streudiagramme der mittleren Monatsrenditen der vier Aktien: Volkswagen; BASF, Siemens, Münchner Rück und zusätzlich der Libor-Zins abgebildet. Der Libor-Zins (**L**ondon **I**nter**b**ank **O**ffered **r**ate) ist der Zinssatz, zu dem sich Geschäftsbanken gegenseitig Geld mit einer Laufzeit bis zu einem Jahr leihen. In der Matrix ist zu erkennen, dass die Höhe der Renditen für Aktien untereinander deutlich zusammenhängen, während kein deutlicher Zusammenhang zwischen Aktienrendite und Zins erkennbar ist.

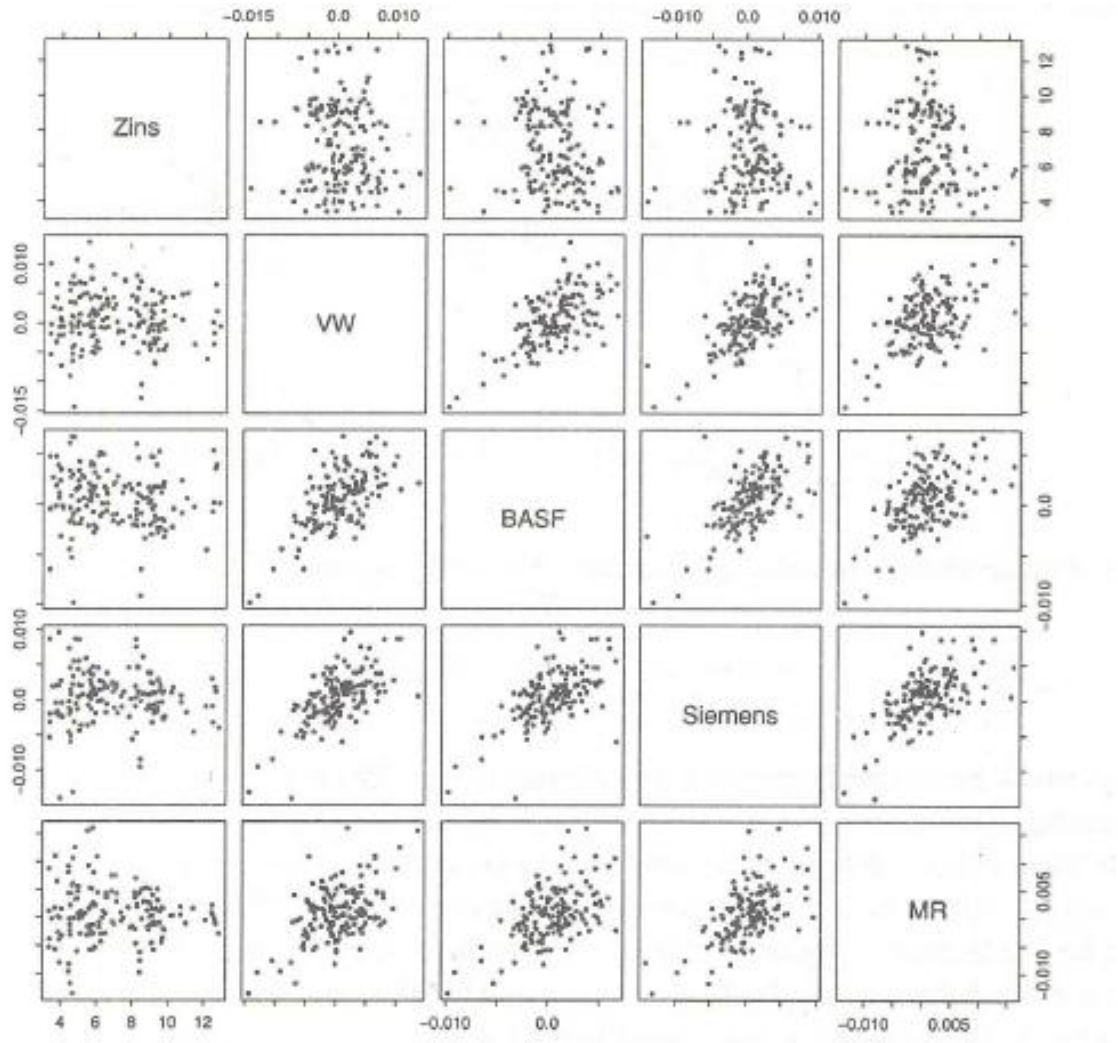


ABBILDUNG 3.10: Streudiagramm-Matrix der mittleren Monatsrenditen für Aktien und des Zinses in Form des Libors

Quelle: Fahrmeier: Statistik, Springer, 2001

3.2 Kennzahlen für bivariate Stichproben

Zunächst können die bekannten Kennzahlen für univariate Stichproben bestimmt werden:

Empirischer Mittelwert

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad 1)$$

Varianz s_x^2 bzw. Standardabweichung s_x

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \qquad 2)$$

bzw.

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \qquad 2´)$$

Als Gesamtmass für den Zusammenhang zwischen den beiden Stichproben wird die folgende sogenannte **Produkt-Momentenkorrelation r_{xy}** definiert:
(englisch: Pearson-Correlation) verwendet:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_I - \bar{y})^2}} \qquad 5)$$

Definiert man als Empirische Kovarianz s_{xy}

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i\, y_i - n\bar{x}\bar{y}\right)$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right) \qquad 6)$$

dann kann die **Produkt-Momentenkorrelation r_{xy}** mit 6) kurz folgendermassen geschrieben werden:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad 5')$$

Schreibt man die Definition 5) in der folgenden Form um, so ist zu erkennen, dass die Korrelation nicht vom Nullpunkt der Skalen von x bzw. y und von den entsprechenden Masseinheiten abhängt:

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

Führt man analog zum Vorgehen bei der Normalverteilung die folgenden neuen standardisierten Variablen ein

$$u_i = \frac{x_i - \bar{x}}{s_x}$$

bzw.

$$v_i = \frac{y_i - \bar{y}}{s_y}$$

und bildet man mit diesen Werten die Vektoren $\vec{u}$ und $\vec{v}$, so vereinfacht sich wegen $|\vec{u}| = |\vec{v}| = \sqrt{n-1}$ die Formel und die Summe kann als Skalarprodukt aufgefasst werden.

$$r_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} u_i v_i = \frac{1}{n-1} \cdot (\vec{u} \cdot \vec{v}) \qquad 7)$$

Da die Definition der Korrelation nicht von der Wahl der linearen Skala abhängt, gilt zudem: $r_{xy} = r_{uv}$

Wenn beide Koordinaten das gleiche Vorzeichen haben (die zugehörigen Punkte liegen dann im verschobenen Koordinatensystem mit Zentrum im Schwerpunkt S $(\bar{x}, \bar{y})$ im ersten und dritten Quadranten), so deutet dies auf einen „positiven“ Zusammenhang zwischen den beiden Variablen, und entsprechend bedeuten Koordinaten mit ungleichem Vorzeichen auf einen „negativen“ Zusammenhang hin (die zugehörigen Punkte liegen im zweiten und vierten Quadranten).

Die aufgeführten Umformungen bei 1) bis 7) ermöglichen die Berechnung aller Kennzahlen aus geeigneten Summen. Allerdings besteht bei dieser Berechnungsart die Gefahr von Rundungsfehlern.

Beispiel Wasserfälle:

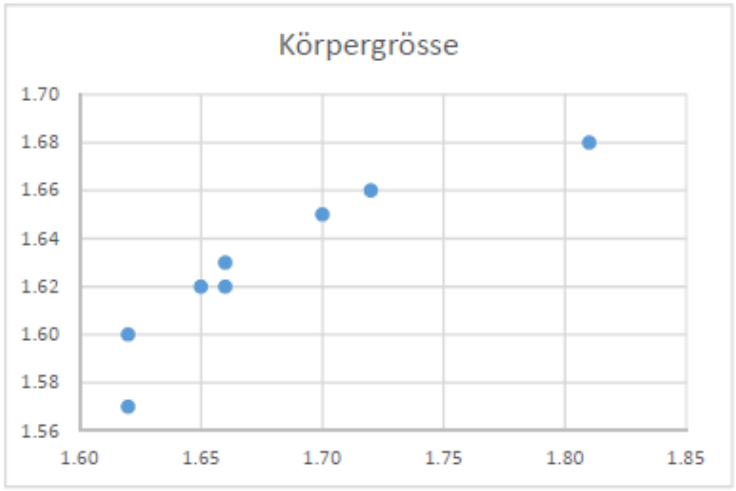
Hoehe [m]	Frequenz [Hz]							
x_i	y_i	x_i^2	y_i^2	$x_i y_i$		u_i	v_i	$u_i v_i$
100	5	10000.00	25	500.0		1.9559956	-0.75911772	-1.48483092
77	3	5929.00	9	231.0		1.23607403	-0.90100889	-1.11371369
50	6	2500.00	36	300.0		0.39094871	-0.68817214	-0.26904001
50	8	2500.00	64	400.0		0.39094871	-0.54628098	-0.21356784
37	9	1369.00	81	333.0		-0.01596348	-0.4753354	0.00758801
24	6	576.00	36	144.0		-0.42287567	-0.68817214	0.29101125
13.3	19	176.89	361	252.7		-0.7577957	0.23412042	-0.17741545
10.9	21	118.81	441	228.9		-0.83291795	0.37601158	-0.3131868
7.7	40	59.29	1600	308.0		-0.93308095	1.72397763	-1.60861069
5.2	40	27.04	1600	208.0		-1.0113333	1.72397763	-1.74351598
375.1	157	23256.03	4253	2905.6	Summe	0	0	-6.62528213
37.51					$\bar{x}$			
	15.7				$\bar{y}$			
		1020.67						
		31.95			s_x			
			198.68					
			14.10		s_y			
				-331.50	s_{xy}			
				-0.736	r_{xy}		r_{xy}	-0.736

Ein weiteres Beispiel:
In der Klasse 1C der Kantonsschule Zofingen wurden am 21.3.2003 die Körpergrössen der Paare Tochter- Mutter bzw. Sohn-Vater gemessen.

Körpergrösse 1C 21.03.2003

Tochter x	**Mutter y**
1.62	1.57
1.62	1.60
1.65	1.62
1.66	1.62
1.66	1.63
1.70	1.65
1.72	1.66
1.81	1.68

Korrelation 0.911

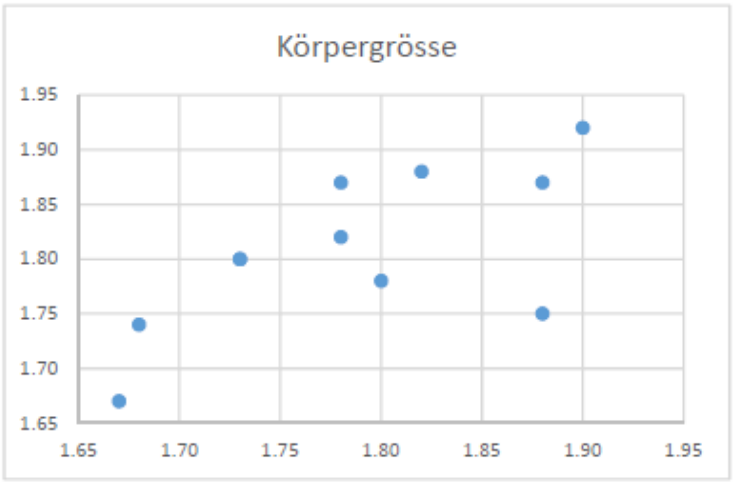


Teilresultate

$\bar{\bar{x}}$	1.680
sd_x	0.063
$\bar{\bar{y}}$	1.629
sd_y	0.035
Kovarianz	0.002
Korrelation	**0.911**

Sohn x **Vater y** **Korrelation 0.669**

1.67	1.67
1.68	1.74
1.73	1.80
1.73	1.80
1.78	1.87
1.78	1.82
1.80	1.78
1.82	1.88
1.88	1.75
1.88	1.87
1.90	1.92



Teilresultate

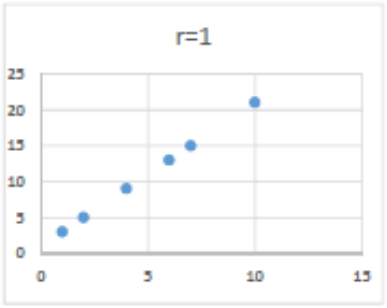
$\bar{\bar{x}}$	1.786
sd_x	0.079
$\bar{\bar{y}}$	1.809
sd_y	0.073
Kovarianz	0.004
Korrelation	**0.669**

Hypothese:
Die Körpergrössen des Paars Vater-Sohn korrelieren schwächer als beim Paar Mutter Tochter.
Im Kapitel „Schliessende Statistik“ wird die Frage behandelt, wie eine solche Hypothese überprüft werden kann.

In den folgenden Abbildungen sind die Scatterplots von acht Stichproben mit verschiedenen **Korrelationskoeffizienten** dargestellt. Zusätzlich sind die Werte der Standardabweichungen von x bzw. y und die Kovarianz angegeben, aus denen sich die Korrelation gemäss 5´) berechnen lässt.

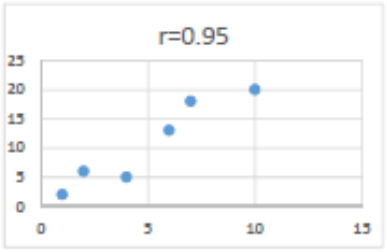
Scatterplot

i	x_i	y_i
1	1	3
2	2	5
3	4	9
4	6	13
5	7	15
6	10	21



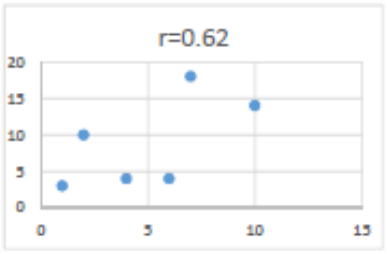
$\bar{x}$	5	$\bar{y}$	11.00
s_x	3.35	s_y	6.69
Kovarianz			22.40
Korrelation			**1.00**

i	x_i	y_i
1	1	2
2	2	6
3	4	5
4	6	13
5	7	18
6	10	20



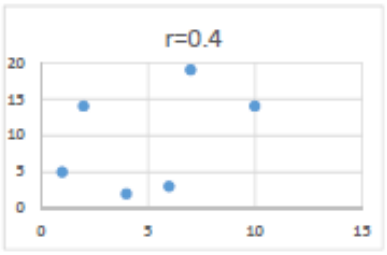
$\bar{x}$	5	$\bar{y}$	10.67
s_x	3.35	s_y	7.42
Kovarianz			23.60
Korrelation			**0.95**

i	x_i	y_i
1	1	3
2	2	10
3	4	4
4	6	4
5	7	18
6	10	14



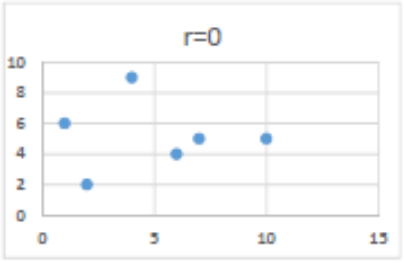
$\bar{x}$	5	$\bar{y}$	8.83
s_x	3.35	s_y	6.21
Kovarianz			12.80
Korrelation			**0.62**

i	x_i	y_i
1	1	5
2	2	14
3	4	2
4	6	3
5	7	19
6	10	14



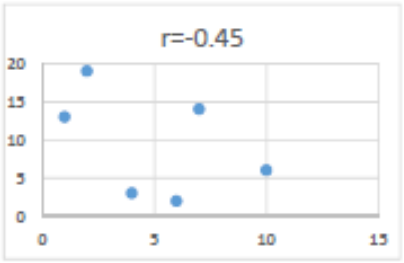
$\bar{x}$	5	$\bar{y}$	9.50
s_x	3.35	s_y	7.06
Kovarianz			9.40
Korrelation			**0.40**

i	x_i	y_i
1	1	6
2	2	2
3	4	9
4	6	4
5	7	5
6	10	5



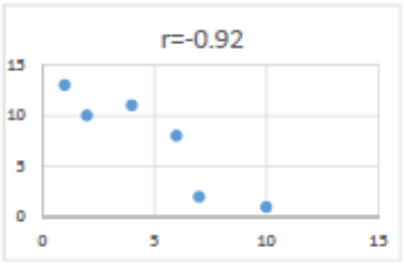
$\bar{x}$	5	$\bar{y}$	5.17
s_x	3.35	s_y	2.32
Kovarianz			0.00
Korrelation			**0.00**

i	x_i	y_i
1	1	13
2	2	19
3	4	3
4	6	2
5	7	14
6	10	6



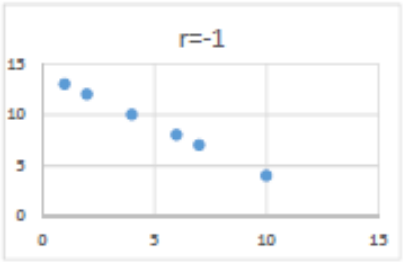
$\bar{x}$	5	$\bar{y}$	9.50
s_x	3.35	s_y	6.83
Kovarianz			-10.40
Korrelation			**-0.45**

i	x_i	y_i
1	1	13
2	2	10
3	4	11
4	6	8
5	7	2
6	10	1



$\bar{x}$	5	$\bar{y}$	7.50
s_x	3.35	s_y	4.93
Kovarianz			-15.20
Korrelation			**-0.92**

i	x_i	y_i
1	1	13
2	2	12
3	4	10
4	6	8
5	7	7
6	10	4



$\bar{x}$	5	$\bar{y}$	9.00
s_x	3.35	s_y	3.35
Kovarianz			-11.20
Korrelation			**-1.00**

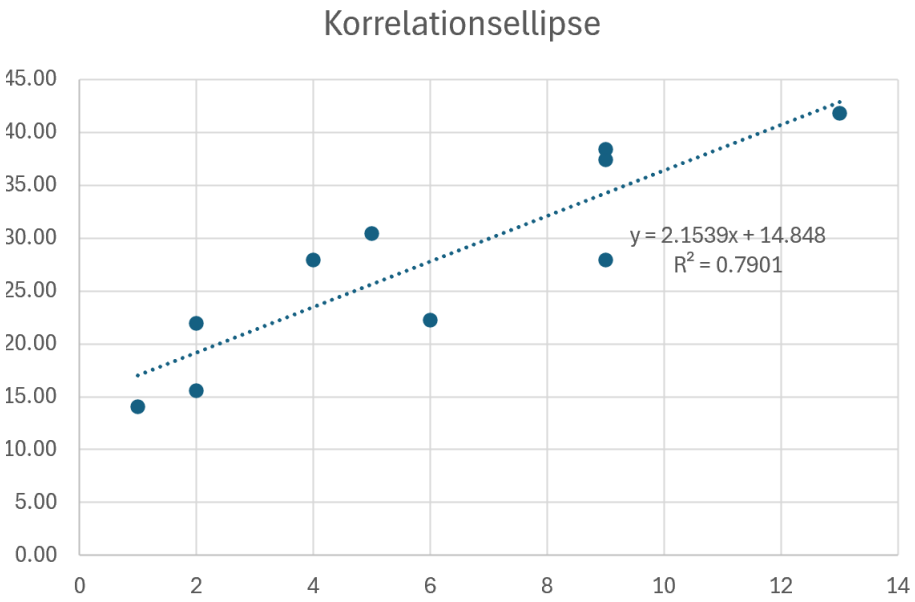
Die Konzentralellipse, Korrelationsellipse

Im folgenden Beispiel sind die Messwerte grafisch dargestellt. Dies führt zur Vermutung, dass die Punktewolke die Form einer Ellipse hat.

Beispiel Korrelation

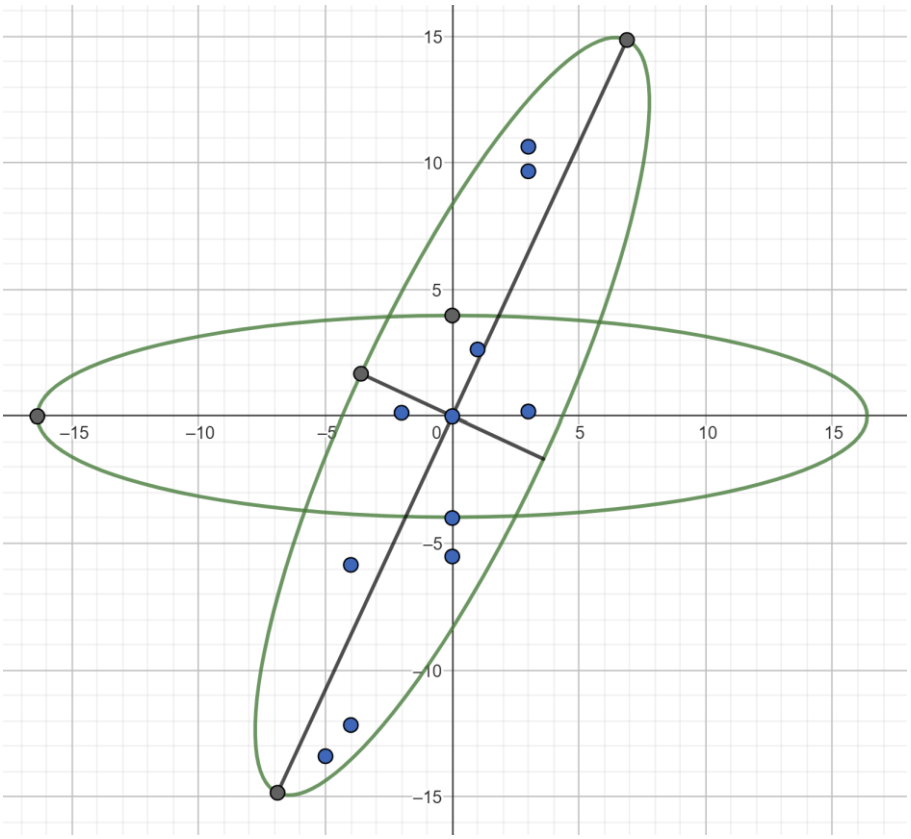
Bei 10 Objekten wurde die Länge X (in cm) un das Gewicht (in kg) gemessen
Die Messwerte und die Kovarianz sind in der folgenden Tabelle dargestellt

i	x_i	yi	xi-xquer	yi-yquer	varianz x	Varianz y		xi*yi
1	1	14.07	-5	-13.70	25	187.72	68.51	14.07
2	2	15.60	-4	-12.17	16	148.13	48.68	31.20
3	2	21.92	-4	-5.85	16	34.23	23.40	43.84
4	4	27.90	-2	0.13	4	0.02	-0.26	111.60
5	5	30.40	-1	2.63	1	6.91	-2.63	152.00
6	6	22.25	0	-5.52	0	30.48	0.00	133.50
7	9	37.43	3	9.66	9	93.30	28.98	336.87
8	9	38.40	3	10.63	9	112.98	31.89	345.60
9	9	27.95	3	0.18	9	0.03	0.54	251.55
10	13	41.79	7	14.02	49	196.53	98.13	543.27
summe	60	277.71	0	-0	138	810.3	297.24	196.35
								-166.63
	x quer	y quer	stabws(x)		3.715	cov(x,y)	29.724	
Mittelwerte	6	27.771	stabws(y)		9.002	**kovarianz**	29.724	29.724
						excel		
							masstababhängig	
Kovarianz	29.72		297.24			**bravaisperson**		0.8889 r
stabw_x	3.715		3.715			Korrelationskoeff		
stabw_y	9.002		9.002			Bestimmheitsmass		0.7901 r^2



Wegen des grossen Korrelationskoeffizienten scheinen die Punkte auch einer Gerade zu folgen mit der Steigung 2.1539 und dem y-Achsenabschnitt 14.484. Die rechnerische Bestimmung dieser Werte wird im folgenden Kapitel «lineare Regression» hergeleitet. In der Abbildung ist ausserdem das sogenannte Bestimmungsmass R^2 angegeben. Ebenfalls im nächsten Kapitel wird gezeigt, dass R mit der Bravais-Pearson übereinstimmt. Dort wird sich ergeben, dass die weiteren Parameter (Ellipsenachsen a ≈ 16.363 und b ≈ 3.968, Ellipsenmittelpunkt$M(\bar{x} \approx 6, \bar{y} \approx 27.771)$, Winkel der gedrehten Ellipse $\alpha \approx 65.1°$) durch den Korrelationskoeffizienzen abhängen- Verweis: «Weitere Themen»→Lineare Algebra→Hauptachsentransformation.

In dieser Abbildung sind die Punkte mit Zentrum (0,0) dargestellt.



Eigenschaften der Produkt-Momenten-Korrelation

a)
Die Korrelation r_{xy} ist +1 für eine Gerade mit positiver Steigung, -1 für eine mit negativer Steigung. Für Werte von r_{xy} nahe bei +1 oder -1 liegen die Punkte eng um eine Gerade. Die Korrelation ist also ein Mass für die Stärke und Richtung eines linearen Zusammenhangs. Ist r_{xy} = 0, so besteht kein linearer Zusammenhang.

b)
Es gilt $|r_{xy}| \leq 1$

{
Ein einfacher Beweis ergibt sich nach der sogenannten **Cauchy-Schwarzschen Ungleichung**, die von Ebene und Raum her bereits bekannt ist:

Zum Beweis wählt man für die Berechnung der Korrelation wie in 3.2 die standardisierten Variablen

$u_i = \frac{x_i - \bar{x}}{s_x}$ bzw. $v_i = \frac{y_i - \bar{y}}{s_y}$

und bildet mit diesen Werten die Vektoren $\vec{u}$ und $\vec{v}$. gilt

Wegen der Standardisierung folgt daraus für die Varianz nach 2)

$$s_u^2 = \frac{1}{n-1}|\vec{u}|^2$$

$$s_v^2 = \frac{1}{n-1}|\vec{v}|^2$$

und wegen

$$\left(\sum_{i=1}^{n} u_i\right) = \left(\sum_{i=1}^{n} v_i\right) = 0$$

nach 6) für die Kovarianz

$$s_{uv} = \frac{1}{n-1} \cdot (\vec{u} \cdot \vec{v})$$

und schliesslich für die Produkt-Momenten-Korrelation nach 5´) und Cauchy-Schwarz

$$r_{xy}^2 = r_{uv}^2 = \left(\frac{s_{uv}}{s_u s_v}\right)^2 = \frac{\frac{(\vec{u} \cdot \vec{v})^2}{(n-1)^2}}{\frac{1}{n-1}|\vec{u}|^2 \cdot \frac{1}{n-1}|\vec{v}|^2} \frac{(\vec{u} \cdot \vec{v})^2}{(|\vec{u}| \cdot |\vec{v}|)^2} \leq 1$$

}

c)

Die folgenden Streudiagramme stellen Stichproben mit gleichem Korrelationskoeffizienten 0.7 dar (Quelle: Chambers, Cleveland, Kleiner and Tukey 1983).

Es ist der grosse Einfluss von allfälligen Ausreissern zu erkennen. In Anwendungen sollte deshalb nicht nur die Korrelation allein, sondern auch das zugehörige Streudiagramm beachtet werden.

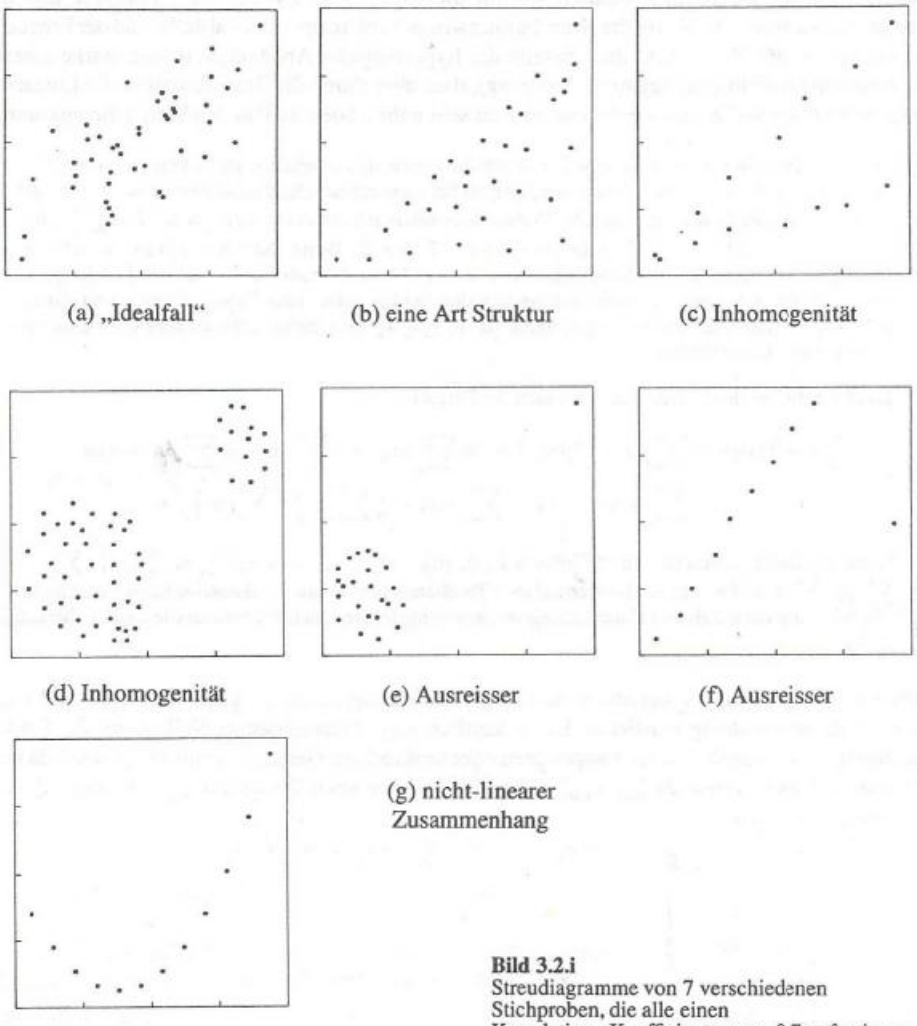


Bild 3.2.i
Streudiagramme von 7 verschiedenen Stichproben, die alle einen Korrelations-Koeffizienten von 0.7 aufweisen

d)

Korrelationen sollten vorsichtig interpretiert werden. Selbst ein offensichtlicher Zusammenhang braucht keinen ursächlichen Zusammenhang zu bedeuten. Es kann nämlich sein, dass beide Grössen von einer dritten Grösse abhängen.

Dass die Anzahl der bei einem Brand eingesetzten Feuerwehrleute hoch korreliert ist mit dem angerichteten Schaden, bedeutet nicht, dass die Feuerwehrleute besser daheim bleiben würden.
Heikel sind insbesondere Zeitreihen mit linearer Komponente. Im bekannten Beispiel ist die hohe Korrelation zwischen der Zahl der Störche und der Zahl der Geburten zwischen 1900 und 1970 kein Beweis, dass der Klapperstorch die Babys bringt.
Es besteht z.B. eine positive Korrelation zwischen Körpergewicht und manueller Geschicklichkeit bei Schulkindern, denn Kinder werden mit steigendem Alter im Mittel schwerer und gleichzeitig geschickter.
Bei Männern besteht zwischen dem Einkommen und der Zahl der Haare auf dem Kopf eine negative Korrelation. Die erklärende Variable ist auch in diesem Fall das Alter.