Korrelationsmaße

Bsp 2: X – "Länge" von Eier

Y – "Breite" von Eier

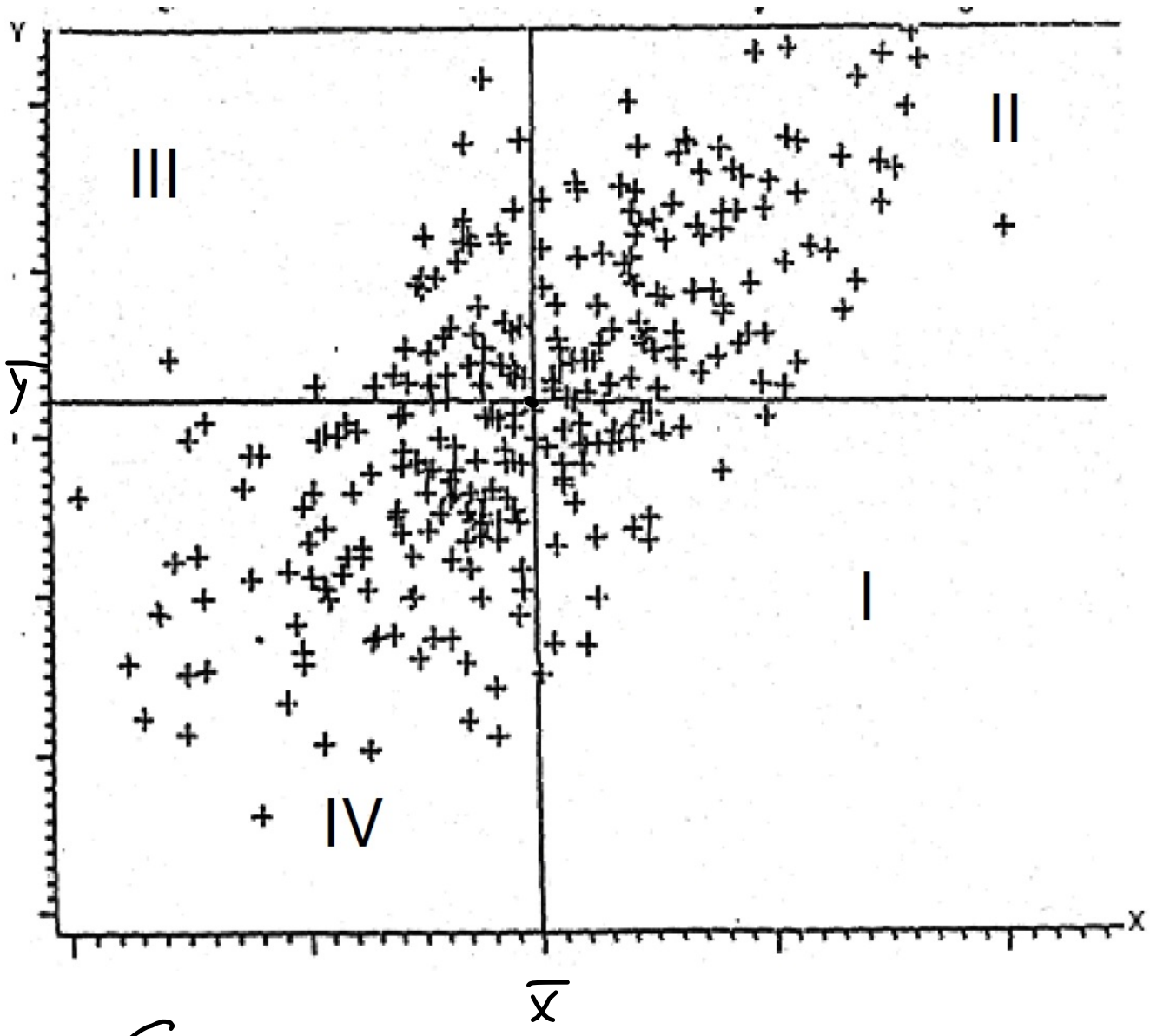
Frage: Hängen Länge und Breite zusammen?



Maß, um das
zu beurteilen?

Stichprobe, 1000 Eier

Paare (x_i, y_i), $i = 1, \ldots, 1000$



Scattergramm

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

Offensichtlich:

$$(x_i - \bar{x})(y_i - \bar{y}) < 0$$

$$\Downarrow$$

$(x_i; y_i)$ liegt in I. oder III. Quadrant

$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

$$\Downarrow$$

$(x_i; y_i)$ liegt in II. oder IV. Quadrant.

Def!

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

heißt Stichprobenkovarianz von X und Y bzgl. der Stichprobe (x_i, y_i), $i = 1, \ldots, n$

Bem! Liegen die Daten hauptsächlich in

I. und III. Quadrant $\Rightarrow S_{xy} < 0$

II. und IV. Quadrant $\Rightarrow S_{xy} > 0$

Sind X und Y unabhängig:

⇓

gleich viele in I., II., III.,
und IV. Quadranten

⇓⇓

$S_{xy} = 0$

Problem: Beurteilung von
~~S_{xy}~~ S_{xy} → wie weit von 0?

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)

(5. Vorlesung vom 18.06.)

⇓

$s_{xy} = 1,2$

Ist $s_{xy} = 1,2$ groß oder klein?

⇒ s kann man nicht verwenden, erst normieren!

$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\downarrow\downarrow$$

$$S_{xx} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) =$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = S_x^2$$

Strenuy (Stichprobenvarianz)
von X

Analog !

$$S_{yy} = S_y^2$$

Def:

Das Maß

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

heißt

Pearson'scher Korrelations-koeffizient von X und Y

bzgl. der Stichprobe (x_i, y_i), $i = 1, \ldots, n$

Satz: Es gilt:

(1) $-1 \leq r \leq 1$

(r normiert)

(2) $r = \begin{cases} 1, & y_i = a \cdot x_i + b, \\ & a > 0 \\ -1, & y_i = a \cdot x_i + b, \\ & a < 0 \end{cases}$

Der Pearson'sche Korrelationskoeffizient r misst den Grad des linearen Zusammenhangs zwischen zwei Merkmalen X und Y

Standard (Bezeichnung und Auswertung):

$r < -0,2 \Rightarrow$ Daten sind negativ korreliert

(großes x $\Leftrightarrow$ kleines y)

(I., IV. Quadrant)

$r > 0{,}2 \Rightarrow$ Daten sind positiv korreliert

(große x $\Leftrightarrow$ große y)

II., IV. Quadrant

"Feinabstimmung"

1.	$0 \leq \lvert r \rvert \leq 0{,}2$	unkorreliert
2.	$0{,}2 < \lvert r \rvert \leq 0{,}5$	schwach korreliert
3.	$0{,}5 < \lvert r \rvert \leq 0{,}8$	korreliert
4.	$0{,}8 < \lvert r \rvert$	hoch korreliert

Für unser Beispiel
mit Eier:

$$r = 0,83$$

$\Rightarrow$ Länge und Breite
sind positiv hoch korreliert

Typische „Punktwolke“ für Korrelationskoeffizienten r:

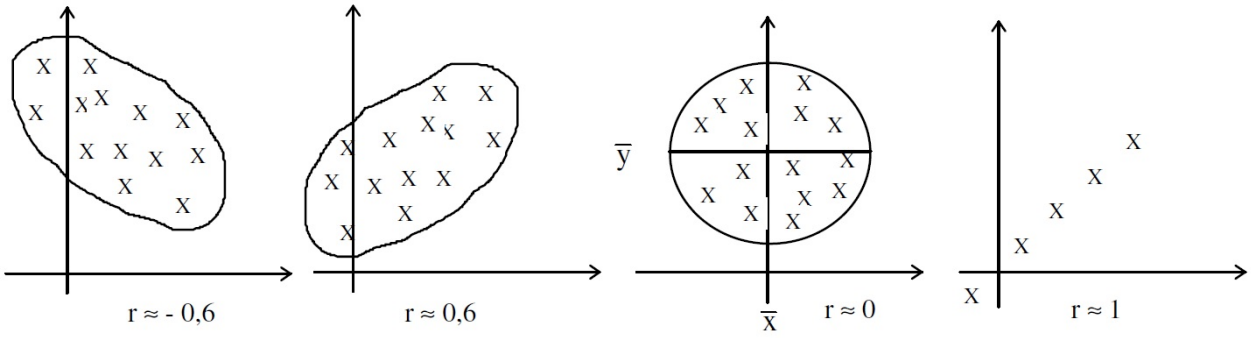


Abb. 3.3

Bem:

1) Voraussetzung: X, Y stetig (mindestens intervallskaliert)

2) $Y = X^2$: $y_i = x_i^2$

$\Rightarrow r = 0$

"r=0" bedeutet nicht

"Unabhängigkeit", sondern

"(linear) unkorreliert"

$\Longrightarrow$ Anderes Maß !

$\Downarrow$

Spearman'scher Korrelations-koeffizient

$\longrightarrow$ für ordinal skalierte Daten $\longrightarrow$ auch mittelwert, varianz

X, Y zwei Zufallsgrößen,
(mindestens ordinal skaliert)

Rangplätze R[x:]

Sei $x_1, \ldots, x_n$ eine Stichprobe
vom Umfang n, und

$x_{[1]} \leq x_{[2]} \leq \ldots\ldots \leq x_{[n]}$

die geordnete Stichprobe
(eindeutige Zuordnung
$j \leftrightarrow [i]$)

Unter dem Rangplatz R[xj]

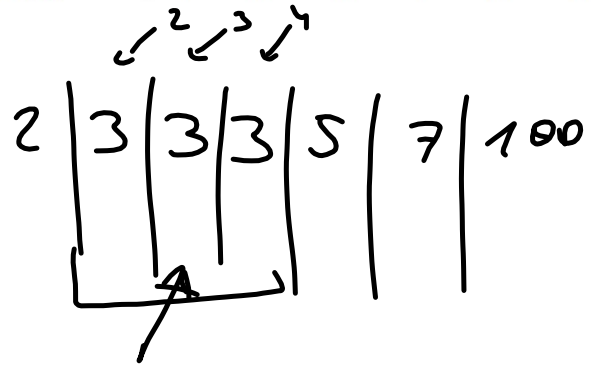
Versteht man den Platz
[i] von x_j in der geordneten
Stichprobe.

Man ordnet die Stichprobe
so, daß der in der
Ordnung an der ersten
Stelle stehende Merkmals-
wert "der beste" ist,
dann "zweitbeste", u.s.w.

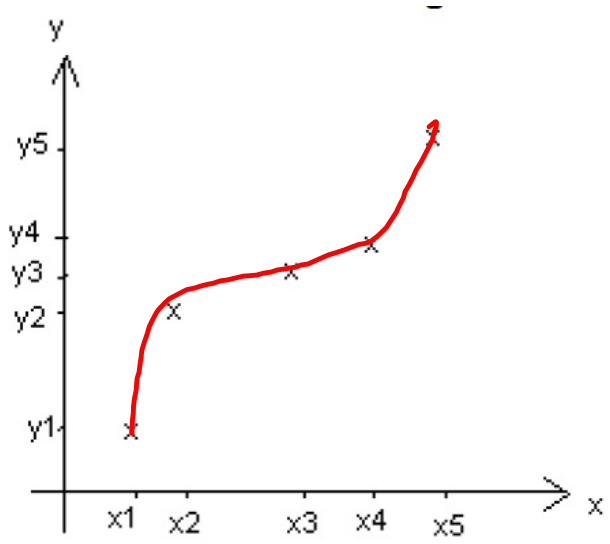
Wenn mehrere Daten x_i den gleichen Wert haben, so wird ihnen der mittlere Rangplatz zugewiesen:

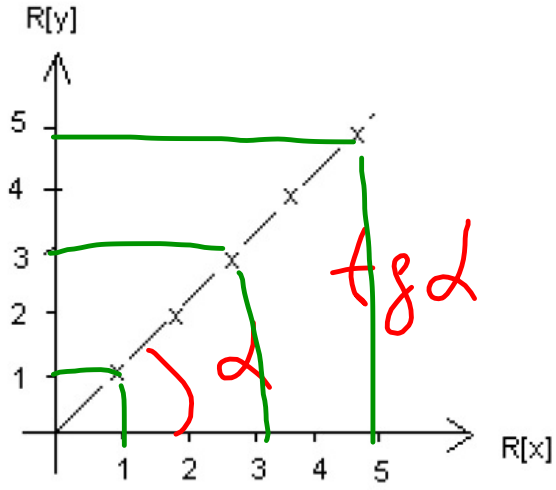
Ausprägung (Qualitätsklasse)	A	B	C	D	E	F	G
durch Skalierung zugeordnete Daten: x_i	2	3	4	5	6	7	100

x_i	2	3	7	3	5	3	100
Platz in geordneter Stichprobe	1	2	6	3	5	4	7
$R[x_i]$	1	3	6	3	5	3	7



$$R[3] = \frac{2+3+4}{3} = 3$$





Wenn (x_i, y_i), $i = 1, \dots, n$ auf einer streng monoton wachsenden Funktion liegen, so liegen die Rangpaare

$(R[x_i], R[y_i])$, $i = 1, \dots, n$

auf einer Geraden mit positiver Anstieg.

Analog ist die
Folge streng monoton
fallend, so liegt Ranghöhe
auf einer Geraden
mit negativem Anstieg

Streng monoton wachsend

$\Downarrow$

die Werte wiederholen sich nicht

$$\begin{cases} R[x_i] = i; \\ R[y_i] = i; \end{cases} \quad i = 1, \dots, n$$

$$(R[x_i], R[y_i]) = (i, i)$$

$\Rightarrow$ Winkelhalbierende Gerade!

Anstieg : 1

Strong monoton fallend:

Anstieg : -1

Abweichung in der

Monotonie der Daten

↓

Abweichung von der linearen

Abhängigkeit der Rangplätze.

Def.: Der Koeffizient

$$\rho := r_{R[x],R[y]} = \frac{s_{R[x],R[y]}}{\sqrt{s^2_{R[x]} \cdot s^2_{R[y]}}},$$

heißt Spearman'scher Korrelationskoeffizient

(ρ := "Pearson'scher Korrelationskoeffizient für die Rangplatzzahlen")

("Rangplatzkorrelationskoeffizient")

Satz: Es gilt:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

mit

$$d_i = R[x_i] - R[y_i]$$

Berechnung: (Spearman)

(1) Jeder x_i und jeder y_i entsp. Rangplätze $R[x_i]$ und $R[y_i]$ werden zugeordnet.

(2) Man bildet die Rangplatzdifferenzen $d_i = R[x_i] - R[y_i]$

d_i^2, $\sum_{i=1} d_i^2$

(3) Man berechnet ρ.

x	x_1	x_i	x_n	
y	y_1	y_i	y_n	
R[x]	$R[x_1]$	$R[x_i]$	$R[x_n]$	
R[y]	$R[y_1]$	$R[y_i]$	$R[y_n]$	
d_i	d_1	d_i	d_n	$\sum$
d_i^2	d_1^2	d_i^2	d_n^2	$\sum d_i^2$

Tab. 3.6

$$-1 \leq \rho \leq 1$$

Auswertung:

$\lvert\rho\rvert \leq 0{,}2$	$\Rightarrow$ kein Korrelation
$0{,}2 < \lvert\rho\rvert \leq 0{,}5$	$\Rightarrow$ schwache Korrelation
$0{,}5 < \lvert\rho\rvert \leq 0{,}8$	$\Rightarrow$ korrelieren die Daten
$0{,}8 < \lvert\rho\rvert \leq 1$	$\Rightarrow$ hoch korreliert!

positive und negative Korrelation

Person	1	2	3	4	5	6	
Mathe	1 1	5 4	4 3	6 4	2 2	3 3	Noten
Englisch	6 10	5 14	1 19	3 16	4 14	2 19	Punkte
$R_{[Mathe]}$	1	5,5	3,5	5,5	2	3,5	
$R_{[Englisch]}$	6	4,5	1,5	3	4,5	1,5	
| di |	5	1	2	2,5	2,5	2	Σ
d_i^2	25	1	4	6,25	6,25	4	**46,5**

$$\rho = 1 - \frac{6 \cdot 46{,}5}{6 \cdot 35} = 1 - \frac{46{,}5}{35} = \underline{\underline{-0{,}32}}$$

Läufer	1	2	3	4	5	6	7	8	9	
A-Note	5,32	5,61	5,07	5,33	4,98	4,69	5,34	5,06	5,25	
B-Note	5,43	5,42	5,15	5,24	5,07	4,59	5,51	4,89	5,16	
Rangplatz A	3	1	6,5	3	8	9	3	6,5	5	
Rangplatz B	2,5	2,5	5,5	4	7	9	1	8	5,5	
$\lvert d_i \rvert$	0,5	1,5	1	1	1	0	2	1,5	0,5	Σ
d_i^2	0,25	2,25	1	1	1	0	4	2,25	0,25	12

$$\rho = 1 - \frac{6 \cdot 12}{9 \cdot 80} = 1 - \frac{1}{10} = 0{,}9$$

A und B sind hoch positiv korreliert!