Didaktisches Seminar über Stochastik

Themen:

- Gemeinsame Verteilung von zwei Zufallsvariablen
- Lineare Regression
- Korrelation von zwei Zufallsvariablen

Michael Ralph Pape

Mai 1998

1 Gemeinsame Verteilung von zwei Zufallsvariablen

Bisher haben wir eine einzige Zufallsvariable untersucht. Nun wollen wir zwei Zufallsvariablen untersuchen, d.h. ihre Zustände n-mal beobachten.

Beispiel 1.1
Im Schuljahr 1978/79 besuchten insgesamt $n = 6801$ *Schüler die von der Gemeinde Freiburg getragenen Gymnasien. An jedem Schüler können wir die Zufallsvariable* "Name der besuchten Schule" *und die Zufallsvariable* "Klassestufe" *feststellen. In der folgenden Tabelle ist nun die Häufigkeitsverteilung der beiden Zufallsvariablen wiedergegeben.*

	Klassenstufe									
Gymnasien	*5.*	*6.*	*7.*	*8.*	*9.*	*10.*	*11.*	*12.*	*13.*	*insg.*
Droste-	*87*	*89*	*88*	*102*	*98*	*90*	*109*	*79*	*74*	*816*
Kepler-	*79*	*87*	*122*	*123*	*129*	*88*	*101*	*85*	*85*	*899*
Friedrich-	*79*	*61*	*71*	*65*	*59*	*62*	*51*	*56*	*65*	*569*
Bertholt-	*49*	*55*	*59*	*50*	*57*	*45*	*40*	*50*	*52*	*457*
Goethe-	*56*	*81*	*68*	*95*	*106*	*95*	*108*	*124*	*79*	*812*
Rotteck-	*100*	*105*	*91*	*100*	*129*	*83*	*75*	*81*	*59*	*823*
Wenzinger-	*156*	*164*	*149*	*156*	*186*	*176*	*115*	*94*	*64*	*1260*
Theodor-Heuss-	*98*	*123*	*145*	*130*	*88*	*79*	*49*	*0*	*0*	*712*
Deutsch-Franz.-	*62*	*61*	*84*	*91*	*73*	*48*	*34*	*0*	*0*	*453*
zusammen	*766*	*286*	*877*	*912*	*925*	*766*	*682*	*569*	*478*	*6801*

Addiert man die in den Spalten stehenden Häufigkeiten, so erhält man die Häufigkeitsverteilung der Zufallsvariable *"Klassenstufe"*.

Addiert man die in jeweils einer Zeile stehenden Häufigkeiten, so erhält man die Häufigkeitsverteilung der Zufallsvariable *"Name der besuchten Schule"*.

Diese Häufigkeitsverteilungen werden auch Randverteilung (Marginalverteilung) der Zufallsvariable *"Klassenstufe"* bzw. *"Name der besuchten Schule"* genannt. Die Randverteilungen sind also nichts anderes als die bisher schon bekannte Verteilung von einer Zufallsvariablen.

Man kann ihnen entnehmen, daß z.B. 569 Schüler das Friedrich-Gymnasium besuchten, oder z.B. daß es 912 Schüler in der 8. Klassenstufe gibt.

Die erste Zufallsvariable bezeichnen wir künftig mit X, die zweite mit Y. Im Allgemeinen wird der Wert der ersten Zufallsvariablen bei der i-ten Beobachtung mit "x_i" bezeichnet und entsprechend der Wert der zweiten Zufallsvariablen mit"y_i'.

So entsteht dann eine Urliste bestehend aus n Paaren (x_i, y_i) von Werten. Damit haben wir eine gemeinsame Häufigkeitsverteilung der beiden Zufallsvariablen "X" und "Y" gegeben.

Beispiel 1.2
Die von 5 zufällig ausgewählen Personen ermittelten Daten über die Zufallsvariable "Körpergröße" X und die Zufallsvariable "Gewicht" Y sind unten stehend abgebildet.

Person	*1*	*2*	*3*	*4*	*5*
x_i	*155*	*163*	*167*	*172*	*174*
y_i	*49*	*45*	*56*	*68*	*66*

Die beiden Meßwerte der i-ten Person stellen wir uns als Zahlenpaar (x_i, y_i) vor. Dabei bedeutet die Komponente x_i die "Körpergröße" und y_i das "Körpergewicht". Es entsteht eine Urliste in der Form

(155,49), (163,45), (167,56), (172,68), (174,66).

Diese Zahlenpaare tragen wir dann in ein kartesisches Koordinatensystem ein. Dabei wird der erste Wert in Richtung der x-Achse eingetragen und der zweite in Richtung der y-Achse.

Die folgende Abbildung 1 zeigt dies!

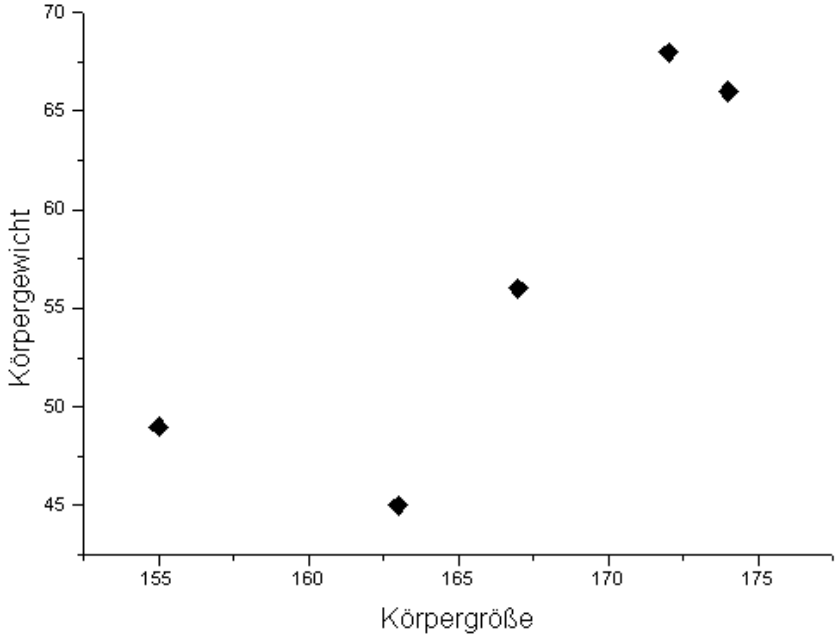

Abbildung 1: Punktediagramm von Beispiel 1.2

Wir haben natürlich auch noch andere Möglichkeiten unsere Wertepaare graphisch darzustellen, so z.B. in einem 3-dimensonalen Stabdiagramm. Allerdings wird man ein solches Diagramm nur dann benutzen, wenn man viele Wertepaare mit Mehrfachbelegung hat.

Bisher haben wir gelernt, wie wir die gemeinsame Verteilung von zwei Zufallsvariablen darstellen können.

Nun könnte jemand auf die Idee kommen, daß zwischen den beiden Zufallsvariablen ein Zusammenhang in irgendeiner Form besteht. Will man diesem Zusammenhang auf die Schliche kommen, so kennt die Mathematik das Mittel der linearen Regression.

2 Lineare Regression

Wir wollen nun versuchen einem solchen Zusammenhang auf die Schliche zu kommen. Dabei betrachten wir der Einfachkeit wegen das Beispiel 1.2. Wir könnten aber auch eine umfangreichere Urliste betrachten.

Beispiel 2.1

Hier wurden von 27 zufällig ausgewählen Personen die Daten "Körpergröße" (Zufallsvariable X) und "Gewicht" (Zufallsvariable Y) ermittelt. Sie sind unten stehend abgebildet.

Person	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
$x_i(cm)$	*155*	*155*	*157*	*159*	*159*	*162*	*163*	*163*	*163*	*163*
$y_i(kg)$	*47*	*49*	*47*	*50*	*52*	*50*	*45*	*47*	*50*	*55*
Person	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
$x_i(cm)$	*163*	*164*	*164*	*166*	*167*	*167*	*168*	*168*	*168*	*168*
$y_i(kg)$	*57*	*52*	*65*	*49*	*53*	*56*	*52*	*58*	*58*	*67*
Person	*21*	*22*	*23*	*24*	*25*	*26*	*27*			
$x_i(cm)$	*169*	*169*	*170*	*170*	*172*	*172*	*174*			
$y_i(kg)$	*55*	*64*	*53*	*61*	*61*	*68*	*66*			

Trägt man diese Daten auf, so ergibt sich folgendes Bild!

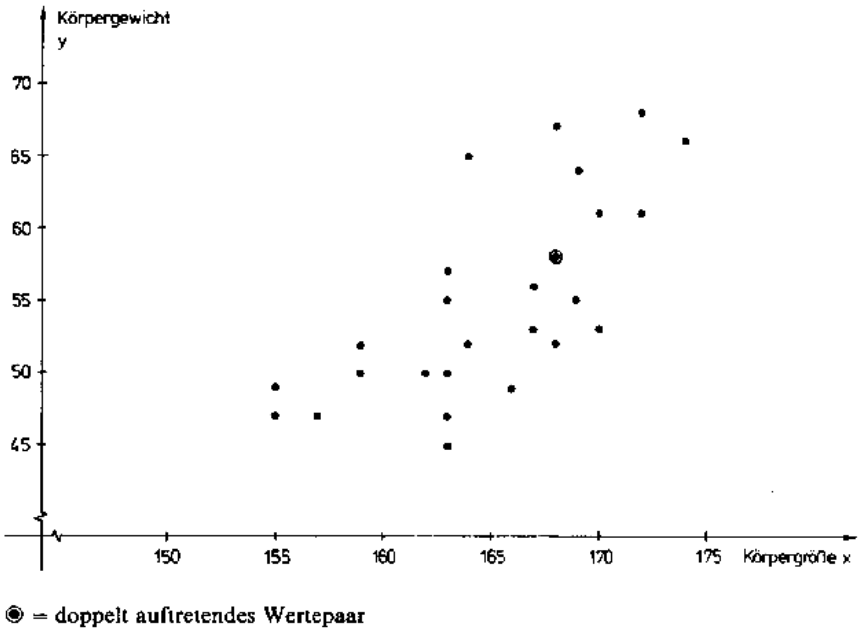

Abbildung 2: Punktediagramm von Beispiel 2.1

Bei genauer Betrachtung der Zahlenpaare oder der Punktwolke aus Beispiel 1.2 und Beispiel 2.1, und auch aus unserer persönlichen Erfahrung fällt auf, daß größere Personen “tendenziell” auch ein größeres Körpergewicht haben. Um Aussagen über diesen statistischen Zusammenhang zu treffen, liegt es nahe eine “Trendgerade”, später mit $\hat{y}(x)$ bezeichnet, möglichst gut durch diese Punktwolke zu legen. Möglichst gut bedeutet hier, daß möglichst viele Punkte auf der Gerade oder sehr nahe bei ihr liegen.

Wir suchen nun also eine Gerade $\hat{y} = bx + a$. Von dieser erwarten wir dann, daß die Summe der Abweichung der Ausprägungspaare (x_i, y_i), $i = 1, \ldots, 5$ von den Punkten $(x_i, \hat{y}(x_i))$ im Mittel möglichst klein ist. Diese Abweichung kann man am Besten durch das Quadrat $(y_i - \hat{y}(x_i))^2$ messen.

Dies läßt sich dabei folgendermaßen veranschaulichen.

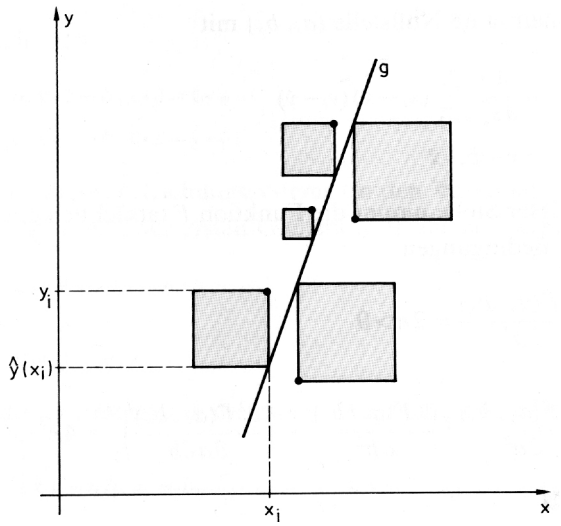

Abbildung 3: Ausgleichsgerade allgemein

Hier wollen wir also, daß diese Quadrate möglichst klein werden.

Die dafür notwendige Rechnung werden wir am Beispiel 1.2 durchführen.

Die Koeffizienten a und b der Gleichung $\hat{y}(x) = bx + a$ sind so zu bestimmen, daß die Funktion

$$(1) \quad F(a,b) := \sum_{i=1}^{5} (y_i - \hat{y}(x_i))^2 = \sum_{i=1}^{5} (y_i - bx_i - a)^2$$

minimal wird.

Dazu halten wir zunächst die Variable b fest und betrachten die Funktion

$$(*) \quad f(a) = \sum_{i=1}^{5} (y_i - bx_i - a)^2 \ .$$

[1.Argumentation]

Aus $f'(a) = -\ 2 \sum_{i=1}^{5} (y_i - bx_i - a) = 0$ erhält man $a = \bar{y} - b\bar{x}$. Dabei ist $\bar{x}$, $\bar{y}$ das arithmetische Mittel von X bzw. Y. Dieses Ergebnis setzt man in (1) ein. Es ergibt sich der Ausdruck

$$(2) \quad \sum_{i=1}^{5} (y_i - bx_i - \bar{y} + b\bar{x})^2 \ .$$

Dadurch wird dann eine Funktion

$$g(b) := \sum_{i=1}^{5} (y_i - \bar{y} - b(x_i - \bar{x}))^2$$

definiert. Für diese Funktion ist ebenfalls das Minimum zu berechnen. Dazu multiplizieren wir zunächst aus und sortieren nach den Potenzen von b. Es ergibt sich

$$g(b) = \underbrace{\sum_{i=1}^{5} (y_i - \bar{y})^2}_{=:S_{yy}} \ - \ 2b \ \underbrace{\sum_{i=1}^{5} (y_i - \bar{y}_i)(x_i - \bar{x})}_{=:S_{xy}} \ + \ b^2 \ \underbrace{\sum_{i=1}^{5} (x_i - \bar{x})^2}_{=:S_{xx}} \ .$$

Mit den Abkürzungen S_{yy}, S_{xy} und S_{xx} erhalten wir

$$g(b) = S_{yy} - 2bS_{xy} + b^2 S_{xx}.$$

Die Ableitung von $g(b)$ ist $g'(b) = -2S_{xy} + 2bS_{xx}$. Aus $g'(b) = 0$ erkennt man, daß das Minimum für $b = \frac{S_{xy}}{S_{xx}}$ angenommen wird.

[2.Argumentation]

Aus der Gleichung (*) erhalten wir dann

$$\begin{aligned} f(a) &= (y_1 - bx_1)^2 - 2a(y_1 - bx_1) + a^2 + \ldots + (y_5 - bx_5)^2 - 2a(y_5 - bx_5) + a^2 \\ &= (y_1 - bx_1)^2 + \ldots + (y_5 - bx_5)^2 - 2a(y_1 - bx_1 + \ldots + y_5 - bx_5) + 5a^2 \\ &= (y_1 - bx_1)^2 + \ldots + (y_5 - bx_5)^2 - 2a(y_1 + \ldots + y_5 - b(x_1 + \ldots + x_5)) + 5a^2 \\ &= (y_1 - bx_1)^2 + \ldots + (y_5 - bx_5)^2 \underbrace{-2 \cdot 5a(\bar{y} - b\bar{x}) + 5a^2} \end{aligned}$$

d.h. mit quadratischer Ergaenzung ergibt sich

$$f(a) = 5(a - (\bar{y} - b\bar{x}))^2 - 5(\bar{y} - b\bar{x})^2 + (y_1 - bx_1)^2 + \ldots + (y_5 - bx_5)^2$$

Das Minimum wird also für $a = \bar{y} - b\bar{x}$ erreicht. Dies setzt man in die Funktion (1) ein und erhält dann eine Funktion $g(b)$.

$$(2) \quad g(b) = \sum_{i=1}^{5}(y_i - bx_i - \bar{y} + b\bar{x})^2 = \sum_{i=1}^{5}(y_i - \bar{y} - b(x_i - \bar{x}))^2 \ .$$

Für diese Funktion ist ebenfalls das Minimum zu berechnen. Dazu multiplizieren wir zunächst aus und sortieren nach den Potenzen von b. Es ergibt sich

$$g(b) = \underbrace{\sum_{i=1}^{5}(y_i - \bar{y})^2}_{=:S_{yy}} - 2b \underbrace{\sum_{i=1}^{5}(y_i - \bar{y}_i)(x_i - \bar{x})}_{=:S_{xy}} + b^2 \underbrace{\sum_{i=1}^{5}(x_i - \bar{x})^2}_{=:S_{xx}} \ .$$

Mit den Abkürzungen S_{yy}, S_{xy} und S_{xx} erhalten wir

$$\begin{aligned} g(b) &= S_{yy} - 2bS_{xy} + b^2 S_{xx} \\ &= S_{yy} + \underbrace{S_{xy}(b^2 - 2b\frac{S_{xy}}{S_{xx}})} \end{aligned}$$

Durch quadratische Ergänzung erhalten wir daraus die Gleichung

$$g(b) = S_{xx}(b - \frac{S_{xy}}{S_{xx}})^2 + S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Dies ist eine Parabel mit dem Scheitel $b = \frac{S_{xy}}{S_{xx}}$.

Die geschilderte Vorgehensweise läßt sich geometrisch leicht veranschaulichen. Die Gleichung

$$s(b) = \sum_{i=1}^{5}(y_i - bx_i - a)^2$$

stellt in einem (a,s)-Koordinatensystem eine Schar von Parabeln mit Scharparameter b dar. Die Rechnung zeigt uns, daß die Scheitel dieser Parabeln wieder auf einer Parabel liegen. Deren Scheitel ist nun das gesuchte Minimum von $\sum_{i=1}^{5}(y_i - bx_i - a)^2$, also $b = \frac{S_{xy}}{S_{xx}}$.
Dies verdeutlicht auch das folgende Bild.

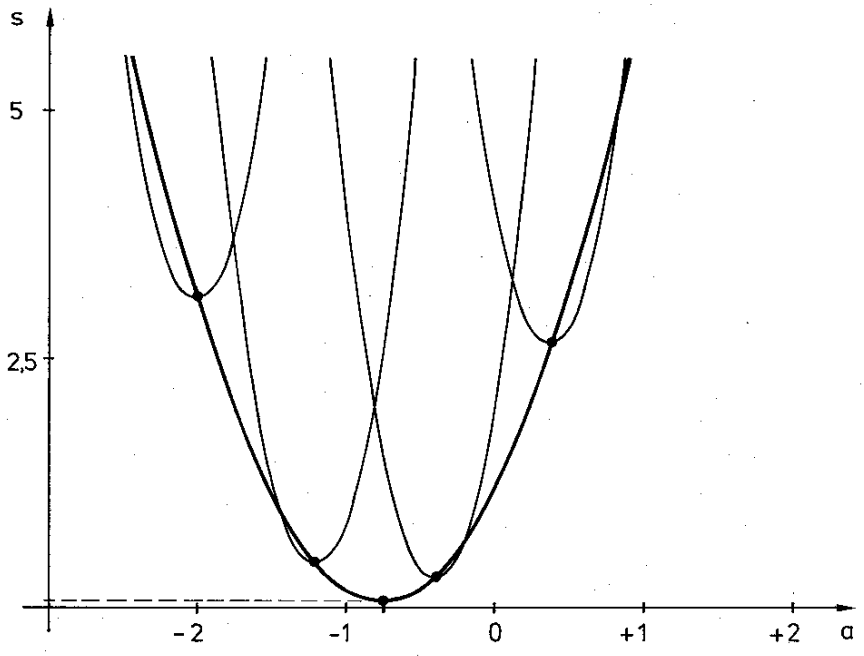

Abbildung 4: Schematische Darstellung der Parabelschar

Die nun eindeutig bestimmte Funktion $\hat{y}(x) = bx + a$ kann man nun benutzen, um von einem nicht notwendigerweise beobachteten Wert von X den Funktionswert $\hat{y}(x)$ zu bestimmen.

Dieses Verfahren heißt “Lineare Regression”, die dabei verwendete Gerade heißt “Regressionsgerade”.

Satz 2.1

Die Regressionsgerade für die Beschreibung der Abhängingkeit der Zufallsvariable Y von der Zufallsvariable X hat die Gleichung

$$\hat{y}(x) = bx + a$$

mit $b = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{5}(x_i-\bar{x})^2}$ *und* $a = \bar{y} - \frac{S_{xy}}{S_{xx}}\,\bar{x} = \bar{y} - b\bar{x}$.

Nun wenden wir auf das Beispiel 1.2 das folgende Schema an.

i	x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
1	155	49	-11,2	-7,8	87,36	125,44	60,84
.	163	45	-3,2	-11,8	37,76	10,24	139,24
.	167	56	0,8	-0,8	-0,64	0,64	0,64
.	172	68	5,8	11,2	64,96	33,64	125,44
5	174	66	7,8	9,2	71,76	60,84	84,64
	$\bar{x} =$	$\bar{y} =$			$S_{xy} =$	$S_{xx} =$	$S_{yy} =$
	166,2	56,8			261,2	230,8	410,8

Abbildung 5: Rechenschema angewand auf das Beispiel 1.2

Wir erhalten für $b = \frac{S_{xy}}{S_{xx}} = 1,13$ und für $a = \bar{y} - b\bar{x} = -131,01$. Und somit ergibt sich unsere Regressionsgerade zu $\hat{y}(x) = 1,13x - 131,01$.

Ihr Verlauf ist nachstehend abgebildet.

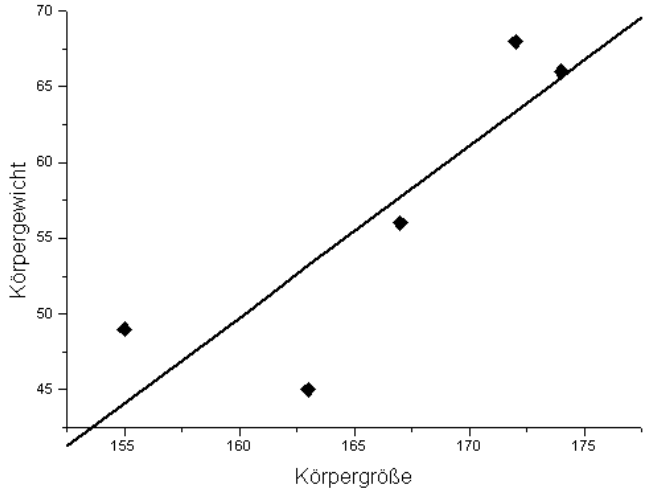

Abbildung 6: Verlauf der Regressionsgeraden zu Beispiel 1.2

Bisher haben wir nun kennengelernt wie man zu jedem Wert der Zufallsvariable X im Punktdiagramm einen “Schätzwert” für die Zufallsvariable Y angeben kann.

Es ist aber oft auch interessant, umgekehrt, also von dem Wert der Zufallsvariable Y auf den Wert der Zufallsvariable X zu schließen. Will man dies tun, so muß man eine neue Regressionsgerade berechnen, also die Gerade für die die Summe der *horizontalen Abweichungsquadrate* minimal ist.

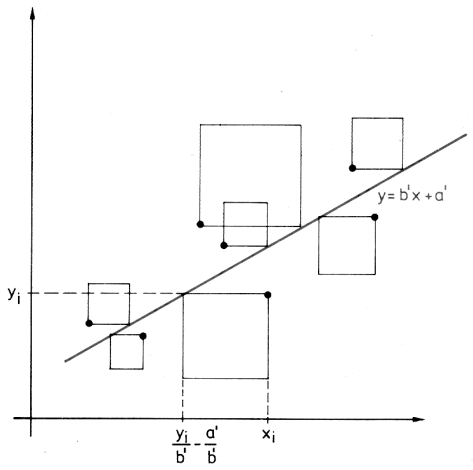

Abbildung 7: Schematische Darstellung der Regressionsgerade Y nach X

Um die Abhängigkeit der Zufallsvariable X von der Zufallsvariablen Y zu bestimmen, vertauscht man einfach die Rollen von X und Y. Dann erhält man die Regressionsgerade mit der Gleichung

$$\hat{x}(y) = b_y y + a_y$$

mit

$$b_y = \frac{S_{xy}}{S_{yy}}, \text{ und } a_y = \bar{x} - b_y \bar{y}$$

Die Werte für S_{xy}, S_{yy}, a, $\bar{x}$ sowie $\bar{y}$ kann man aus der Abbildung 5 einfach ablesen. Daraus lassen sich leicht die Werte für a_y und b_y bestimmen.

Als Regressionsgerade ergibt sich dann

$$x(y) = 0,56y + 134,28 \ .$$

Im Bild sieht das dann so aus.

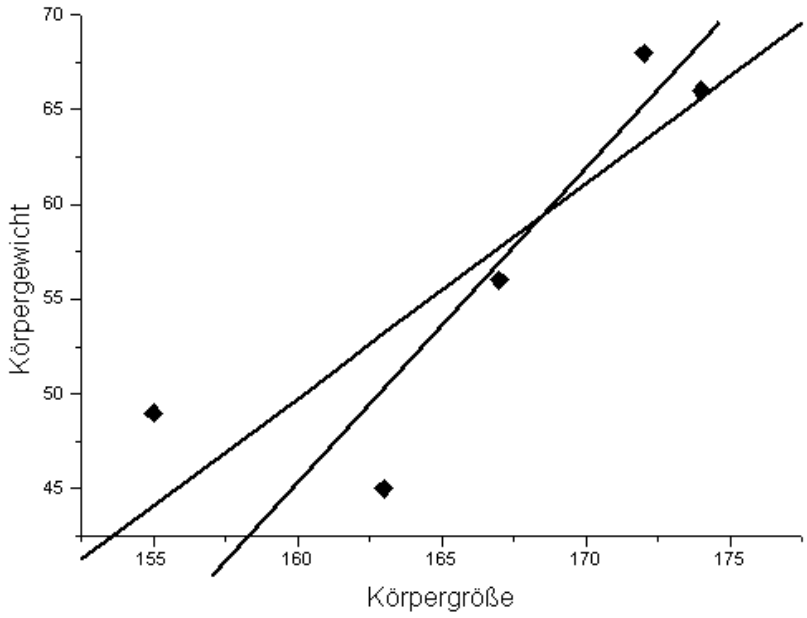

Abbildung 8: Darstellung der Regressionsgeraden Y nach X und X nach Y

Es ergeben sich also i.A. zu einer Häufigkeitsverteilung von zwei Zufallsvariablen zwei Regressionsgeraden, d.h. man muß bei der Verwendung von Regressionsgeraden auf die verschiedenen Definitionen und Bedeutungen achten. Will man für einen vorgegebenen x-Wert auf den zugehörigen y-Wert schließen so wird man die erste Regressionsgerade verwenden, weil bei dieser die Summe der Abweichungsquadrate in y-Richtung minimiert wird. Bei dem Schluß von der Zufallsvariable Y auf die Zufallsvariable X wird man dann die zweite Regressionsgerade verwenden.

Wir wissen nun wie und wozu wir Regressionsgeraden berechnen können.

Nun könnten wir uns noch die Frage stellen, ob vielleicht zwischen den beiden Regressionsgeraden ein Zusammenhang, genauer, ein linearer Zusammenhang besteht?

Ja, es besteht ein solcher Zusammenhang.

3 Korrelation

Was ist den nun der “linearen Zusammenhangs” zweier Zufallsvariablen?

Der “linearen Zusammenhangs” zweier Zufallsvariablen ist eine Kennzahl r, genau wie zum Beispiel der Erwartungswert, die Varianz oder die Standardabweichung. Diese Kennzahl heißt auch noch Korrelationskoeffizient.

Nun sehen wir uns folgende Bildserien an. Was fällt auf?

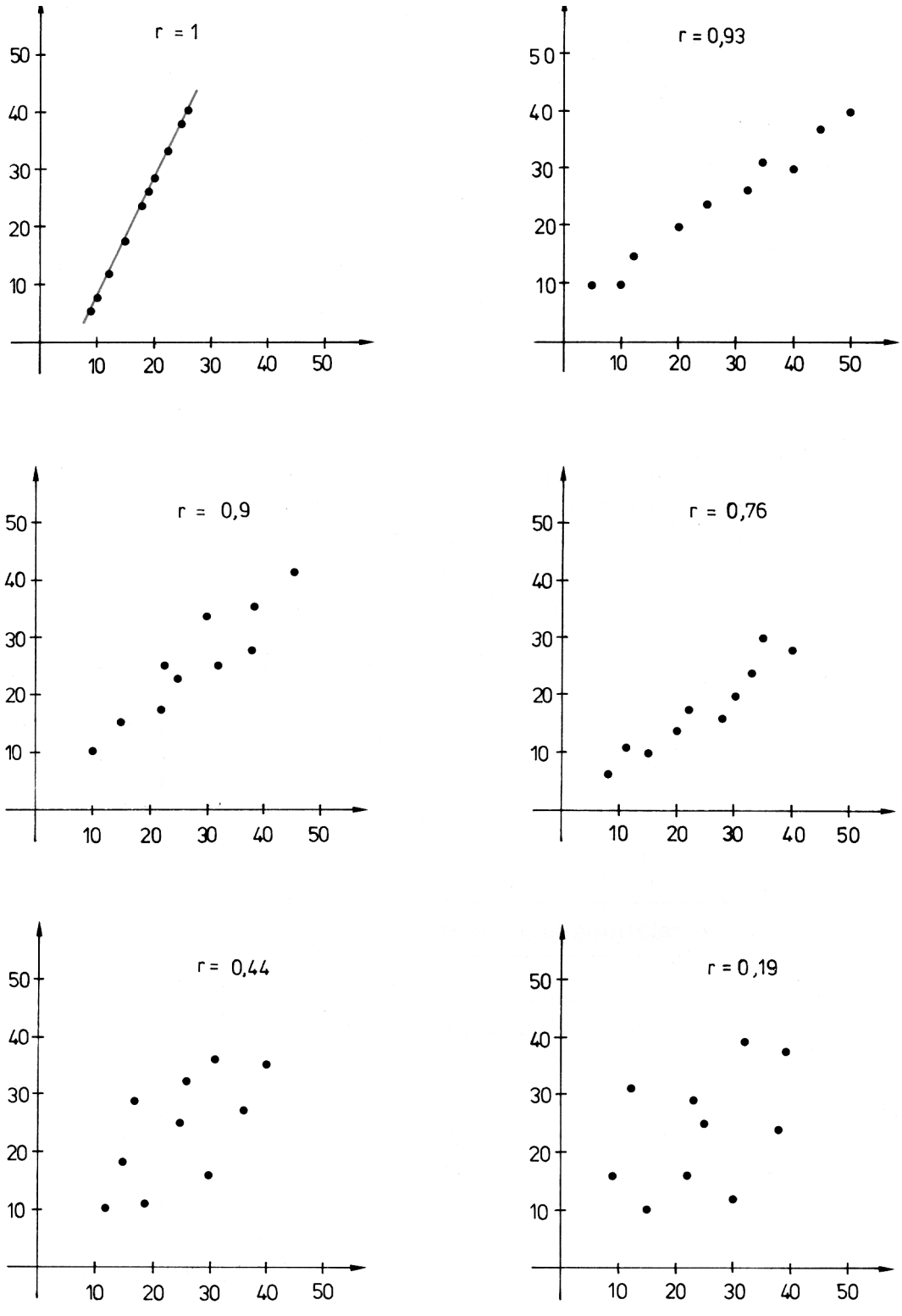

Abbildung 9: Serie 1

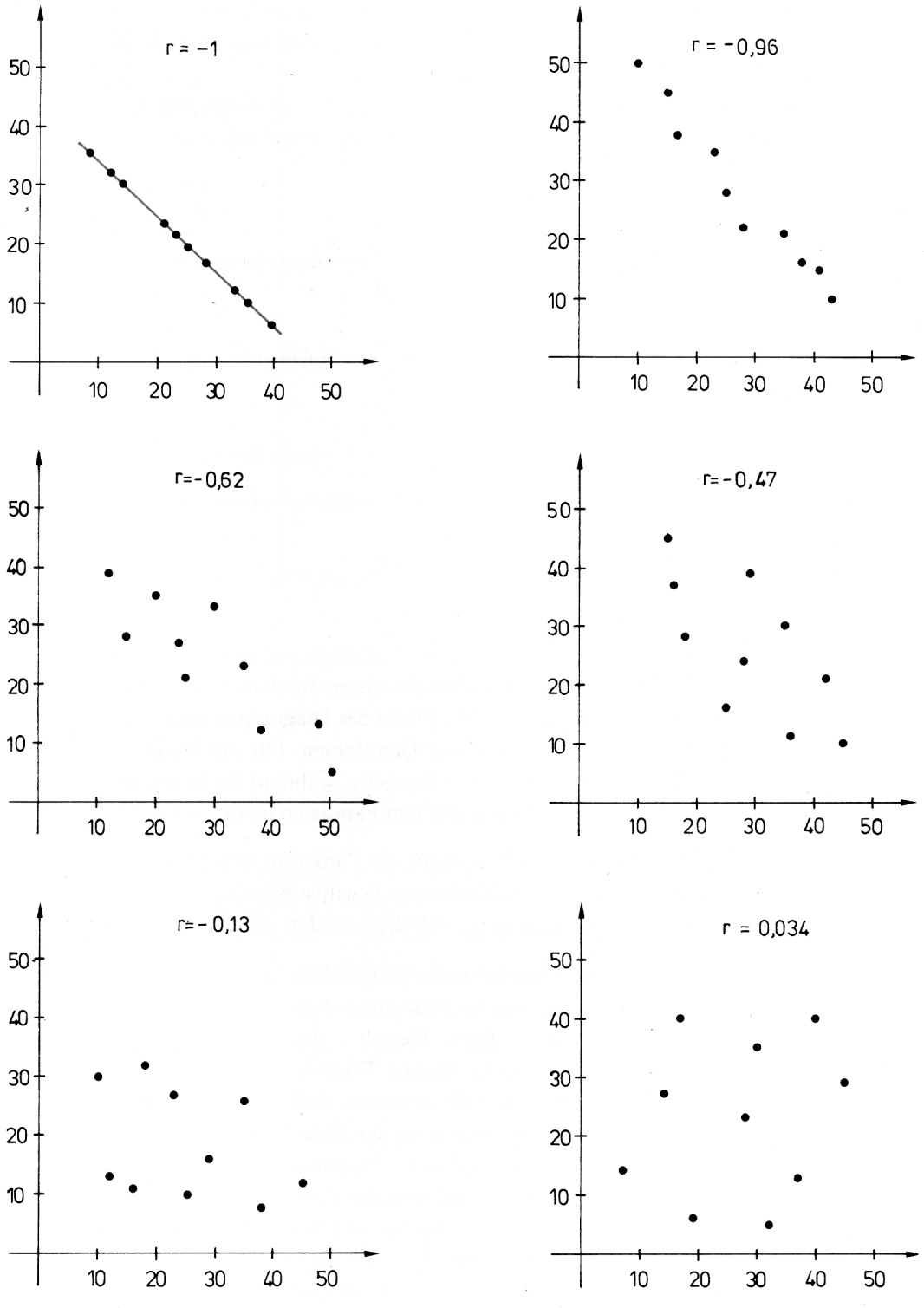

Abbildung 10: Serie 2

Aufgrund der Bilderfolgen wird man nun vermuten, daß der Wert von $|r|$ bei Diagrammen mit zunehmend ausgeprägter linearer Tendenz steigt und bei Diagrammen, bei denen kaum eine lineare Tendenz zu erkennen ist, nahe bei 0 liegt.

Wir wollen nun den Korrelationskoeffizienten für unser Beispiel 1.2 ausrechnen.

Definition 3.1
Die Zahl

$$r := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

heißt Korrelationskoeffizient der gemeinsamen Häufigkeitsverteilung der Zufallsvariablen X und Y. Dabei ist vorausgesetzt, daß der Nenner positiv ist.

Mit der vorhin berechneten Tabelle ergibt sich der Korrelationskoeffizient aus $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ zu $r = 0,85$.

Aus den Bilderfolgen kann man auch erkennen, daß $|r| \leq 1$ gilt und daß die Werte $r = 1$ und $r = -1$ in den Ausnahmefällen in denen die Punkte auf einer Geraden liegen, angenommen werden.

Satz 3.1
Für den Korrelationskoeffizienten r von n Wertepaaren (x_i, y_i) mit $S_{xx} \cdot S_{yy} \neq 0$ gilt

$$-1 \leq r \leq 1 \, .$$

Die Werte -1 und 1 werden dann angenommen, wenn die Punkte (x_i, y_i) auf einer Geraden liegen.

Liegt der Wert der Korrelationskoeffizienten r in der Nähe von +1 bzw. -1, so liegen die Werte fast alle auf einer Geraden, d.h. der lineare Zusammenhang ist stark ausgeprägt. Liegt der Wert von r in der Nähe von Null, so besteht kein linearer Zusammenhang. Wir sehen lediglich ein Punktwolke.