KORRELATION UND REGRESSION - EIN INHALTLICHER ZUGANG ZU DEN GRUNDLEGENDEN MATHEMATISCHEN KONZEPTEN

von Manfred Borovcnik in Klagenfurt

Kurzfassung: Regressions- und Korrelationsrechnung gilt als schwierig. In dieser Arbeit wird ein lehrgangsmäßiger Zugang dazu entwickelt, dessen Zielvorstellung es ist, mathematische Begriffe so einzuführen, daß sie inhaltlich, unmittelbar (ohne weitere Theorie) verständlich werden. Damit soll demonstriert werden, daß dieses Teilgebiet der Statistik auch in der Sekundarstufe II fruchtbringend unterrichtet werden kann.

A. Einleitung

1. Zur fachdidaktischen Einordnung dieser Arbeit

Begriffe und Methoden werden in einem ganz bestimmten Kontext zu ganz bestimmten Zwecken entwickelt. In späterer Folge werden sie oftmals davon abgelöst, sie werden logisch besser durchdrungen und gleichzeitig universeller anwendbar. Im Lernprozeß jedoch kann die allgemeine Darstellung den Zugang dazu, das Verständnis davon, erheblich beeinträchtigen. Für die Regressions- und Korrelationsrechnung trifft das auch in erheblichem Ausmaß zu. Der mathematische Kontext, in den die Konzepte von Regression und Korrelation heute eingebettet sind, ist schwierig zu verstehen und schwierig in Teilsequenzen so zu zerlegen, damit man Verständnis auf seiten des Lernenden ermöglicht.

Die mathematischen Voraussetzungen sind kompliziert (der unbedarfte Leser möge bitte die folgenden kursorischen Bemerkungen dazu einfach überlesen): Die Variablen müssen einer zweidimensionalen Normalverteilung folgen oder dem sogenannten linearen Modell genügen (siehe dazu jedes Statistik-Lehrbuch mit etwas stärker ausgeprägtem mathe-

matischen Touch). Ferner ist die Regressionsgerade eng mit orthogonalen Projektionen in Vektorräumen verwandt. Die übliche Ableitung über das Minimum eines Funktionals in mehreren Variablen (die Methode der kleinsten Quadrate) mit Hilfe der Analysis und die Auflösung der sogenannten Normalgleichungen ist aufwendig, auch was den formeltechnischen "Kram" anbelangt (auch wenn es jetzt doch gute Ansätze zur Elementarisierung gibt).

Die Art von Beziehungen zwischen Variablen, die mit Hilfe von Regression und Korrelation erfaßt wird, steht nahe zu einem Komplex der Ko-relation (des gemeinsamen Variierens von Merkmalen) und kausalen Beziehungen zwischen Merkmalen. Hier gibt es viele unscharfe Vorstellungen, die ja gerade erst durch die mathematischen Begriffe gestrafft werden sollen, und somit auch Fehldeutungen.

Klar, daß man in der fachdidaktischen Diskussion üblicherweise zu dem Schluß gelangt, daß dieses Teilgebiet der Statistik für den Unterricht auf dem Niveau der Sekundarstufe II nicht geeignet ist. Ich bin gegenteiliger Auffassung und möchte das durch die folgenden Ausführungen abstützen. Ich möchte dabei insbesondere folgende allgemeine Zielvorstellungen konkretisieren:

1) Zwischen inhaltlichen Vorstellungen und mathematischen Begriffen soll eine Brücke geschlagen werden.
2) Begriffe werden visuell eingeführt, numerische Berechnungen oder theoretische Ableitungen spielen eine untergeordnete Rolle.
3) Zentrale Ziele, zum Teil aus dem historischen Kontext, in dem die Konzepte entwickelt wurden, und häufig auftretende Mißverständnisse um die Deutung der Begriffe und Ergebnisse werden explizit gemacht.

Ich möchte mit diesem Artikel anregen, daß man das Thema im Unterricht aufgreift. Ich wäre dankbar für Rückmeldungen bzw. Berichte über Unterricht zum Thema, eventuell unter Einbau oder Weiterführung der im folgenden dargestellten Ideen.

2. Zum soziokulturellen Hintergrund der Entstehung der Ideen der Korrelations- und Regressionsrechnung

Ist Intelligenz erblich ? Sind andere Merkmale (Körpergröße, Haarfarbe) entscheidend durch Vererbung beeinflußt ? Heute haben wir durch unsere Fortschritte in der Gentechnologie sehr weitreichende Kenntnisse über die Struktur der Samenzellen und wie die in den Genen gespeicherte Information bei der Paarung weitergegeben wird. Für einzelne qualitative Merkmale wie Haarfarbe, Augenfarbe, aber auch für bestimmte genetische Schäden können wir den Vererbungsmechanismus gut durch stochastische Modelle nachvollziehen. Für ein so komplexes Phänomen wie Intelligenz reichen diese aber nicht aus. Schon Gregor Mendel (1822 - 1884) hat so kleine Informationsträger bestimmter Bauart vermutet, die nach eigenen, stochastischen Regeln den Typ der Tochterzelle bestimmen. Er hat ausführliche Experimente durchgeführt, um seine Spekulationen über die Vererbung zu belegen. Seine diesbezüglichen Veröffentlichungen (1865) sind aber über drei Jahrzehnte hinweg unbeachtet geblieben. So phantastisch unglaublich waren seine Ideen über Gene als Träger von Erbinformationen und doch sind sie 1930 mit der Entdeckung des Elektronenmikroskops bestätigt worden.

Die Idee der Vererbung hat besonders im vorigen Jahrhundert sehr viel Zuspruch gefunden. In der Ausformung der Nationen, eines Nationalbewußtseins, war dies gleichzeitig auch eine staatsbildende oder auch staatszerstörende Idee. Dieser beginnende Nationalismus, der in unserem Jahrhundert in eine Katastrophe gemündet hat, war auch dem victorianischen England um 1870 keineswegs fremd.
In der sogenannten eugenischen Bewegung war man um Verbesserung der eigenen Rasse bemüht, damit diese im Wettstreit mit den anderen Nationen bestehen könne. Eine kleine, aber einflußreiche Schicht in England hatte besonderes Interesse an den Ideen zur Vererbung. Es war dies eine Schicht, die sich im noch jungen Schulsystem bis über die universitäre Ebene hinaus durchgesetzt hat. Diese bildungsbewußte Schicht hatte jedoch gegenüber der alten Elite, dem reichen Adel, einen extremen Nachteil:

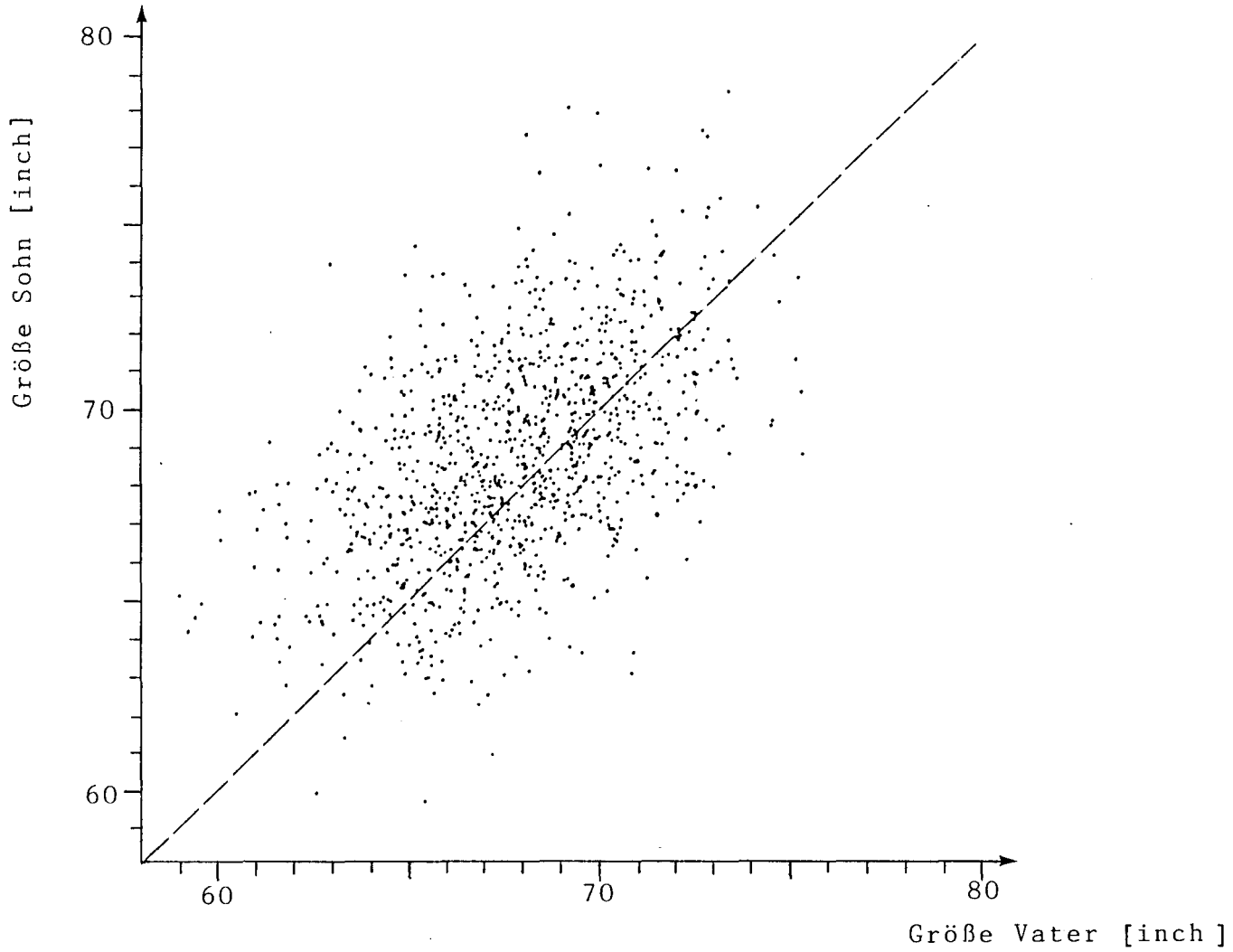
Sie hatte außer ihrer beruflichen Stellung keinen Einflußbereich und war zumeist arm. Sie hatte demgemäß enorme Probleme mit der Reproduktion ihrer Klasse, d.h. mit der Etablierung ihrer Kinder wiederum als Elite. Wenn nun Intelligenz vererbt ist, und da sie ihre Intelligenz durch ihr Durchsetzen im Bildungssystem nachgewiesen haben, sollten ihre Kinder privilegiert werden, zum Wohle der gesamten Gesellschaft wie natürlich auch zum Wohle dieser Bildungselite selbst. Das Problem aber war, daß keineswegs nachgewiesen war, daß Intelligenz vererbt ist, vielmehr gab es durchaus einen alten wissenschaftlichen Streit zwischen den Vererbungstheoretikern mit den Umwelttheoretikern, deren extreme Variante das junge Baby als tabula rasa (vollkommen reiner, leerer Tisch) ansahen, dessen Intelligenz erst durch Umwelteinflüsse nach und nach ausgeformt wurde. Da Intelligenz durch Bewährung im Schulsystem aber nicht in Zahlen erfaßt wurde, da ferner die Intelligenz der zu fördernden Kinder erst auf dem Prüfstand des Schulsystems geprüft wurde - der Intelligenztest kam im Rahmen ähnlicher Bestrebungen erst viel später - verlegte man die Bemühungen, den Nachweis der Gültigkeit der Vererbungstheorie zu erbringen, zunächst auf äußerliche, leicht zu messende Merkmale, wie Körpergröße etwa. Kinder sind ähnlich ihren Eltern. Große Väter haben z.B. große Söhne. Wie stark sind solche Ähnlichkeiten ? Wie kann man das messen? In diesem historischen Kontext wurden die Konzepte von Korrelation und Regression entwickelt.

Francis Galton (1822-1911) steuerte ab 1865 Leitideen und gröbere mathematische Konzepte dazu bei, sein Schüler Karl Pearson (1857-1936) entwickelte in diesem Zusammenhang den Pearsonschen Korrelationskoeffizienten (1896). Der soziokulturelle Hintergrund in der zweiten Hälfte des 19. Jahrhunderts in England wird bei MacKenzie vortrefflich rekonstruiert.

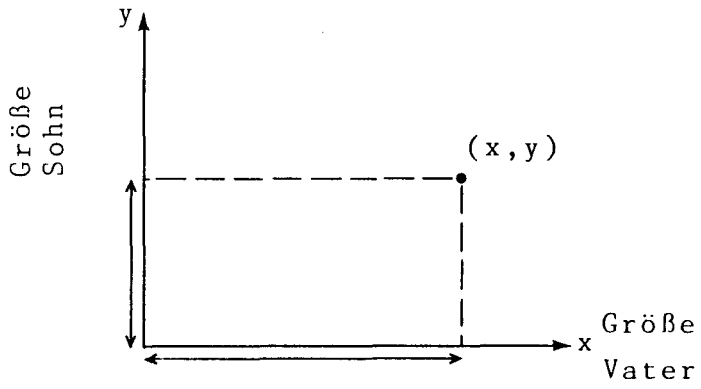
A. Korrelation

1. Punktwolke - Graphische Darstellung von zweidimensionalen Daten

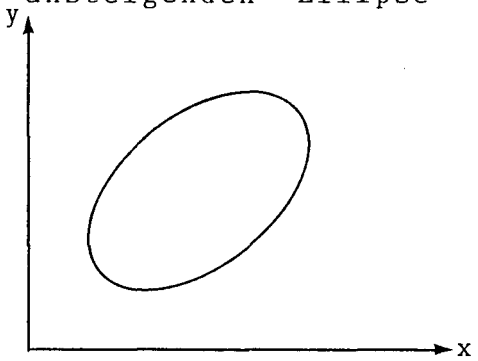
Beispiel: Körpergröße von Vätern und deren Söhnen. Daten von n = 1078 Vätern und Söhnen aus einer Studie von K. Pearson (1903).



Ein *einzelner Punkt* (x,y) stellt einen Vater und dessen Sohn dar

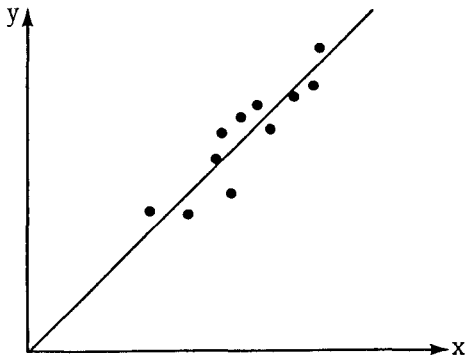


Die *Wolke* aller Punkte macht den Eindruck einer "ansteigenden" Ellipse

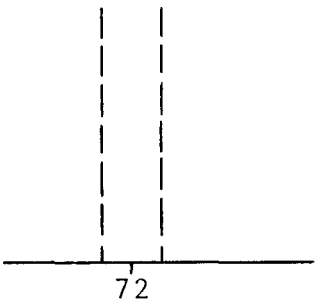


In der Abbildung der Punktwolke ist die Gerade y = x gestrichelt eingezeichnet. Sie entspricht jenen Paaren, bei denen der Sohn

gleich groß ist wie der Vater. Das sind nur ganz wenige. Wäre der Zusammenhang zwischen der Körpergröße von Vater und Sohn ein starker, wäre $y \approx x$, so sollte sich die Punktwolke recht eng um die Gerade $y = x$ anschmiegen, etwa so:



Würde man in so einem Fall von der Körpergröße $x = 72$ inch des Vaters die Körpergröße des Sohnes mit $y = 72$ voraussagen, so könnte man sich dieser Prognose sehr sicher sein. Die Punktwolke der Daten von Pearson ist jedoch viel breiter.
Legen Sie einen Streifen um $x = 72$ parallel zur 2. Achse, dann sehen Sie ganz deutlich wie sehr die Punkte (die Körpergrößen der Söhne) streuen.



Starker Zusammenhang bedeutet: *Leichter voraussagen* zu können den Wert y der *abhängigen* Variablen aus der Kenntnis des Werts x der *unabhängigen* Variablen.

Aufgabe:

a) Suchen Sie aus den Daten von Pearson den kleinsten Vater und dessen Sohn, die größten Väter und deren Söhne.

b) Väter mit $x = 72$. Welches sind darunter die kleinsten/größten Söhne.

c) Söhne mit $y = 76$. Wie groß sind die Väter ?

d) Schätzen Sie aus der Punktwolke:

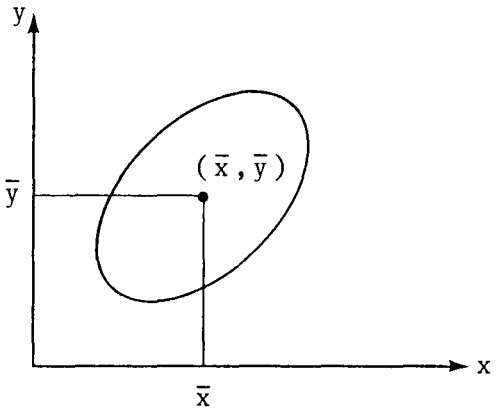
$\bar{x}$ eher 64, 68, 72 [inch] ?

s_x eher 3, 6, 9 [inch] ?

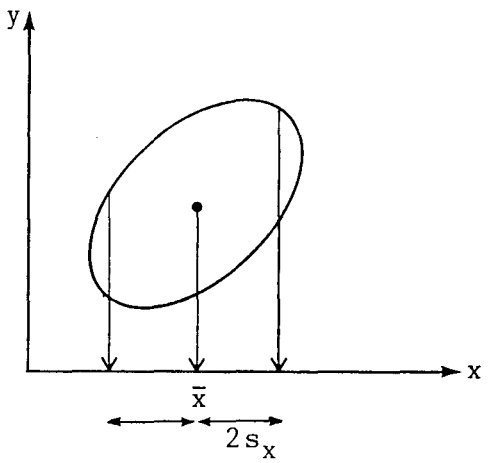
2. Quantitative Zusammenfassung der Punktwolke - der Korrelationskoeffizient

Wie bei gewöhnlichen Häufigkeitsverteilungen bzw. deren Histogrammen ist die zahlenmäßige Erfassung des Zentrums der Punktwolke sowie deren Ausbreitung interessant. Gesucht ist eine bündige Beschreibung durch wenig Zahlen.

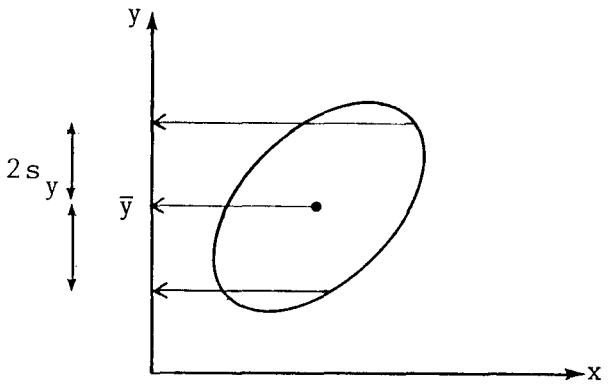
Zentrum = Punkt der Durchschnitte $(\bar{x},\bar{y})$



Breite in der 1. Achse



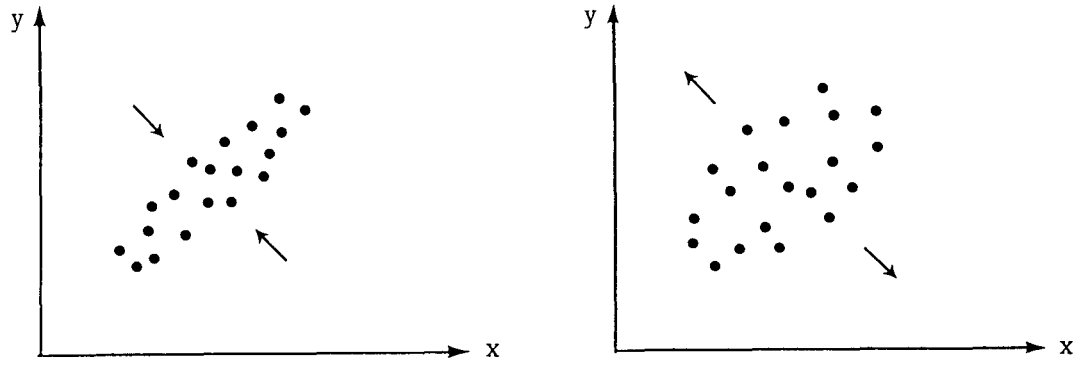
Breite in der 2. Achse



Wenn wir die Punkte längs der 2. Achse auf die erste Achse verschieben, erhalten wir die Verteilung der x-Daten alleine (ohne Differenzierung nach Werten von y). Die Breite dieser Häufigkeitsverteilung der x-Daten wird oft durch das Intervall

$(\bar{x}-2s_x, \bar{x}+2s_x)$ angegeben.

Es fehlt jedoch noch ein wesentlicher Punkt der Beschreibung:



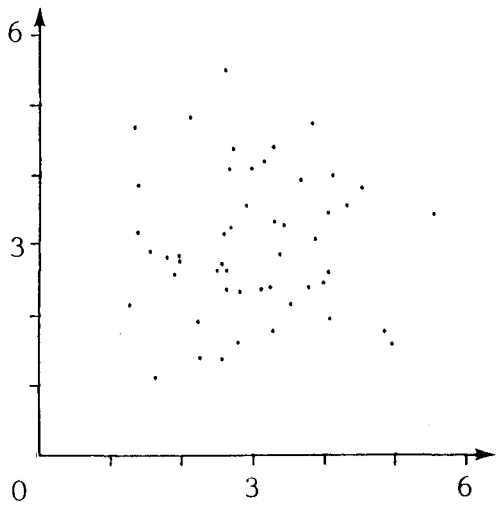
Diese Punktwolken sind offensichtlich unterschiedlich dick. Gerade das scheint im Hinblick auf die erwünschten Zusammenhänge, die untersucht werden sollen, eine kritische Größe zu sein. Die Dicke der Wolke mißt man indirekt mittels des sogenannten Korrelationskoeffizienten, in Zeichen r, den wir später genauer festlegen werden.

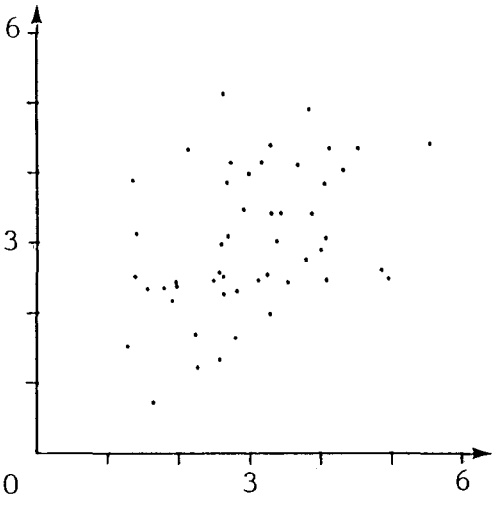
Beschreibung der Punktwolke: $\bar{x}$, $\bar{y}$; s_x, s_y; r.

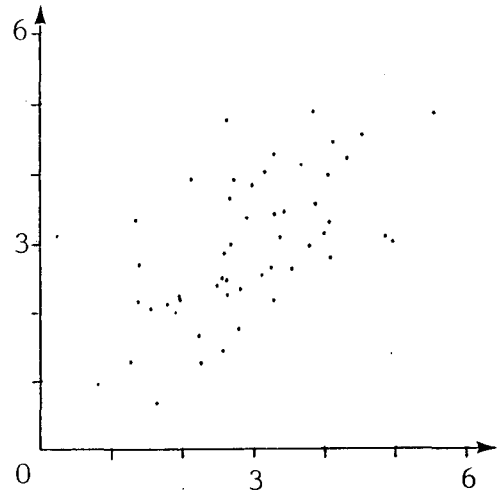
Wichtig ist nun der Zusammenhang der Gestalt der Punktwolke mit der konkreten Zahl, die r annimmt. Einen Eindruck vermittelt die folgende Serie von Punktwolken.

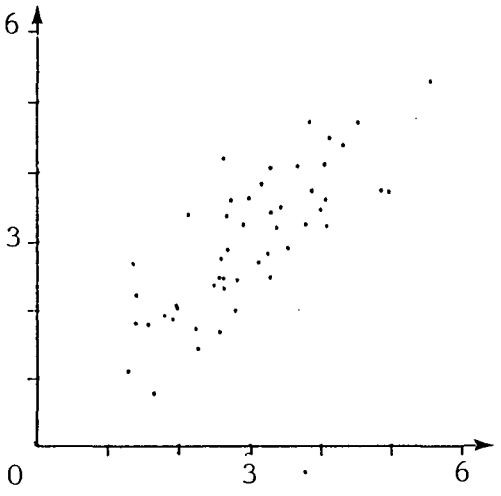
Aufgabe:

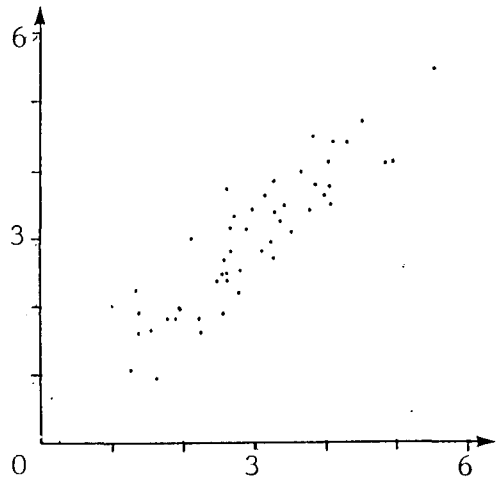
Auf der folgenden Seite sind mehrere Datensätze durch Punktwolken dargestellt. Versuchen Sie, selbst Werte für den Korrelationskoeffizienten r den einzelnen Punktwolken zuzuordnen. Der Korrelationskoeffizient soll Werte zwischen -1 und 1 annehmen. Die 1 würde man zuordnen, falls ein perfekter linearer Zusammenhang vorhanden wäre, d.h. alle Punkte liegen auf einer Geraden. Negative Werte für r würde man zuordnen, falls die Punktwolke "fällt".

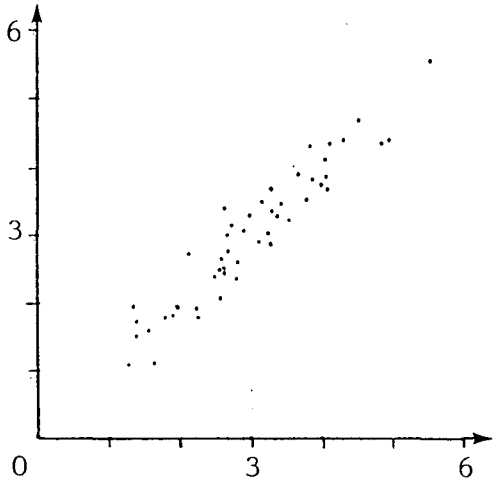

6
3
0
3
6


6
3
0
3
6


6
3
0
3
6


6
3
0
3
6

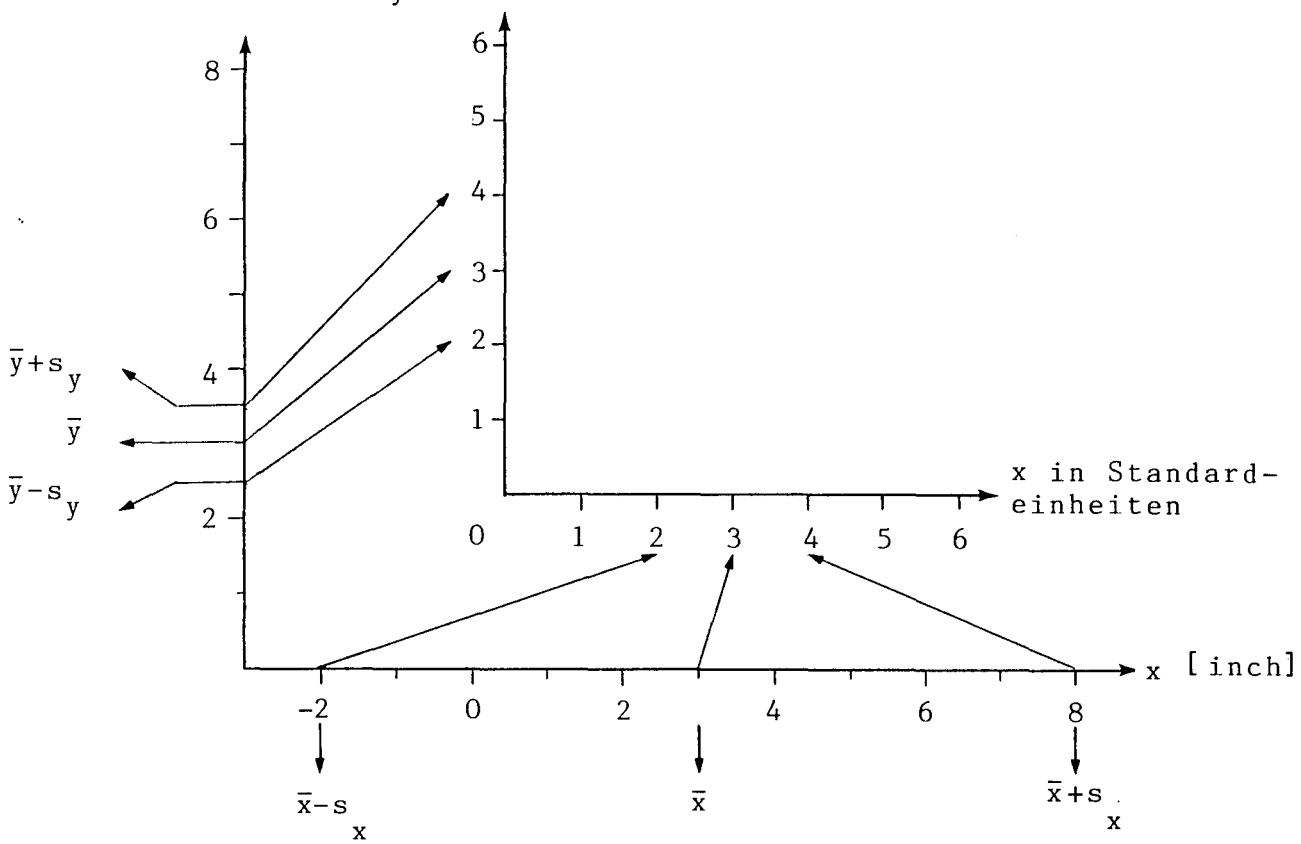

6
3
0
3
6


6
3
0
3
6

Die Monotonie der Zuordnung, daß r von links oben nach rechts unten größer werden soll, ist leicht einsichtig. Die konkreten Werte sind r = 0,00;r = 0,40;r = 0,60;r = 0,80;r = 0,90; und 0,95.

Beachten Sie bitte, daß die Punktwolken in der Schätzaufgabe für den Korrelationskoeffizienten r in folgendem Sinne *normiert* sind: Der Punkt der Mittelwerte ist immer in (3,3). Für die Standardabweichungen s_x und s_y wurden jeweils *eine* Einheit in der Zeichnung gewählt, ungeachtet ihres tatsächlichen numerischen Werts

Beispiel: $\bar{x} = 3$ $\bar{y} = 3$
$s_x = 5$ $s_y = 0,5$



In der 1. Achse wird die ursprüngliche Punktwolke gestaucht, in der 2. Achse gestreckt, damit man die normierte Punktwolke erhält. Wir werden später noch sehen, daß man keine Chance hat, einen vernünftigen Eindruck von der Streuung der Punktwolke zu erhalten, wenn man die Punktwolken nicht normiert. Eine einfache Fehlinterpretation des Korrelationskoeffizienten r sei gleich vorweggenommen: r = 0,80 bedeutet nicht, daß 80% der Punkte sehr eng an einer Geraden liegen und der Rest streut weiter. r = 0,80 bedeutet nicht,

daß die Punktwolke doppelt so eng ist wie jene bei r = 0,40. Die Dicke der Punktwolke wird durch r viel indirekter erfaßt.

Aufgabe:

Schätzen Sie die Korrelation in den Daten von Pearson zu den Körpergrößen von Vater und Sohn.

Bemerkung: Die Punktwolke zu den Daten von Pearson über die Körpergröße kann man prägnant durch folgende Zahlen beschreiben:
Väter: $\bar{x} = 67,7$ $\quad s_x = 2,74$
Söhne: $\bar{y} = 68,7$ $\quad s_y = 2,76$ in [inch] und $r \approx 0,508$.

3. Die s-Gerade

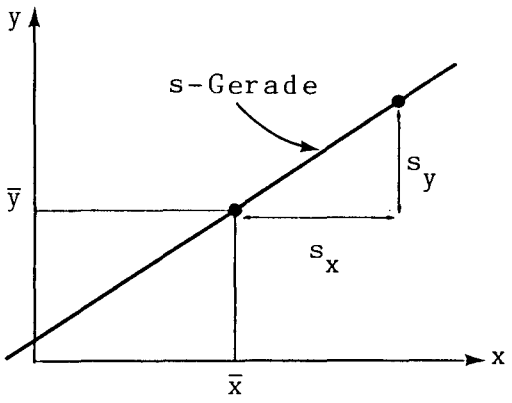
Je näher r = 1; desto enger gruppieren sich die Punkte der Punktwolke um eine Gerade, die sogenannte s-Gerade. Stellt man sich die Punktwolke als Ellipse vor, so geht die s-Gerade durch die Hauptachse dieser Ellipse. Auf ihr sind gerade jene Punkte (x,y), die für beide Variablen gleicherweise "extrem", genauer, gleich weit vom jeweiligen Mittelwert sind. Liegt x z.B. eine Standardeinheit s_x über dem Mittelwert $\bar{x}$, so liegt (x,y) auf der s-Gerade, falls y ebenfalls eine Standardeinheit, nun s_y, über dem Mittelwert $\bar{y}$ liegt.

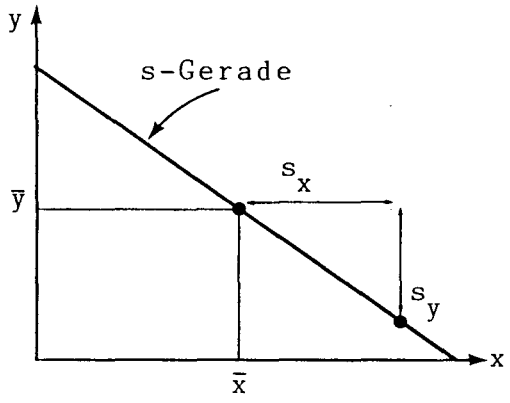
(s) $(\bar{x},\bar{y})$ liegt auf s-Gerade

Steigung: $\frac{s_y}{s_x}$ oder $-\frac{s_y}{s_x}$

$$\frac{x - \bar{x}}{s_x} = \frac{y - \bar{y}}{s_y}$$

Gleichung der s-Geraden





Beispiel: (Körpergröße, Körpergewicht) von männlichen College-Studenten. Beschreibung der Daten (1 inch = 2,54cm, 1 lib = 0,45kg):

$\bar{x} = 69$ [inch] $\qquad s_x = 3$ [inch]

$\bar{y} = 140$ [lib] $\qquad s_y = 20$ [lib]

$r = 0{,}60$.

Man bestimme für $x = 72$ jenes Körpergewicht y, sodaß (x,y) auf der s-Geraden liegt:

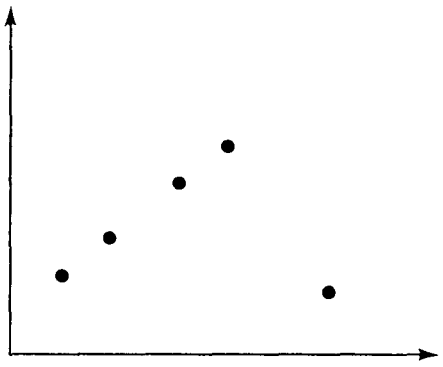
$$\frac{x_i - \bar{x}}{s_x} = \frac{72 - 69}{3} = \frac{3}{3} = 1, \quad \text{also soll gelten}$$

$$\frac{y_i - \bar{y}}{s_y} = 1 \text{ oder } y_i = \bar{y} + 1 \cdot s_y = 140 + 1 \cdot 20 = 160 \text{ [lib]}.$$

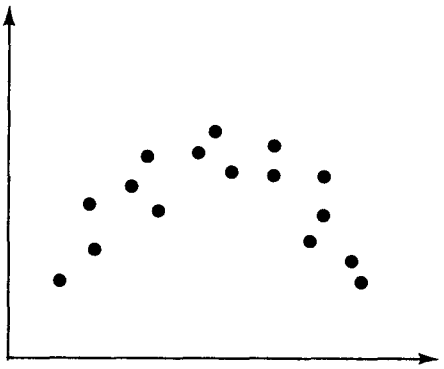
4. Ausnahmefälle, in denen der Korrelationskoeffizient versagt

Der Korrelationskoeffizient r gibt nicht immer eine gute Zusammenfassung der Stärke der Gruppierung:

Ausreißer



Nichtlinearität



In beiden Fällen hat die Punktwolke eine extrem von der elliptischen Form abweichende Gestalt. In beiden Fällen ist die zusammenfassende Beschreibung, insbesondere durch den Korrelationskoeffizienten, nicht zielführend.

5. Berechnung und Eigenschaften des Korrelationskoeffizienten

a) Normiere die Werte jeder Variablen:

$$\frac{x_i - \bar{x}}{s_x} \qquad \frac{y_i - \bar{y}}{s_y}$$

b) Bilde das Produkt entsprechender normierter Werte.

c) Summiere die Produkte über alle Datenpaare $i = 1, \ldots, n$.

d) Dividiere durch die Zahl der Datenpaare.

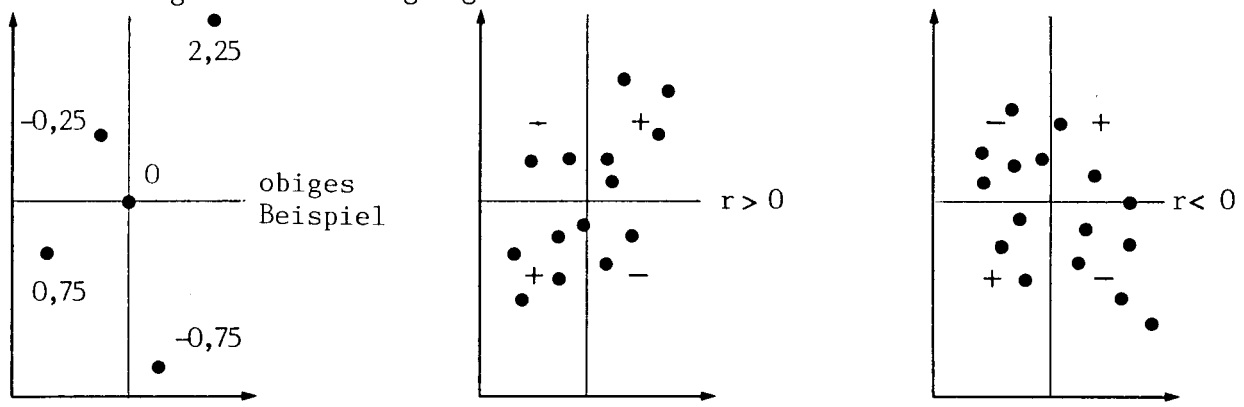
Beispiel: Berechnen Sie r für folgende hypothetische Daten:

x_i	y_i	x-Werte	y-Werte	Produkte
1	5	- 1,5	- 0,5	0,75
3	9	- 0,5	0,5	- 0,25
4	7	0,0	0,0	0,00
5	1	0,5	- 1,5	- 0,75
7	13	1,5	1,5	2,25

z.B. $\bar{x} = 4$, $s_x = 2$

Berechnung des normierten Werts zu x_1=1: $\frac{1-4}{2} = \frac{-3}{2} = -1,5$.
Der Mittelwert der Produkte (letzte Spalte) ergibt 0,40, das ist der gesuchte Wert für r.

Was bedeutet der implizit durch das obige Rechenverfahren festgelegte Korrelationskoeffizient? Eine einfache Deutung erhält man durch folgende Überlegung:



Ist die Punktwolke steigend, so sind die beiden Felder mit + stark besetzt, die beiden Felder mit - schwach, die Felder mit + sind gerade jene für die x und y in gleicher Richtung vom Mittelwert abweichen, das Produkt der normierten Werte ist daher positiv, bei Punkten, die in den Feldern - liegen, sind die entsprechenden Produkte negativ.
Ist die Punktwolke *steigend* wird demnach der Mittelwert der Produkte normierter Werte und somit der Korrelationskoeffizient *positiv* sein.

Welche mathematischen Eigenschaften hat der Korrelationskoeffizient?

Aufgabe:

Wie ändert sich der Korrelationskoeffizient im Beispiel mit den hypothetischen Daten, wenn man

a) die Variablen x und y vertauscht,
b) statt der Variablen x die Variable x' mit x' = x+3 nimmt,
c) statt der Variablen x, die Variable x" mit x" = 2x nimmt?

Der Korrelationskoeffizient ist eine *dimensionsfreie* Zahl, er ändert sich bei Maßstabsänderungen, das sind lineare Transformationen der Variablen, z.B.

$$x' = ax+b,$$

nicht, falls $a > 0$. Falls $a < 0$, ändert sich nur das Vorzeichen von r.

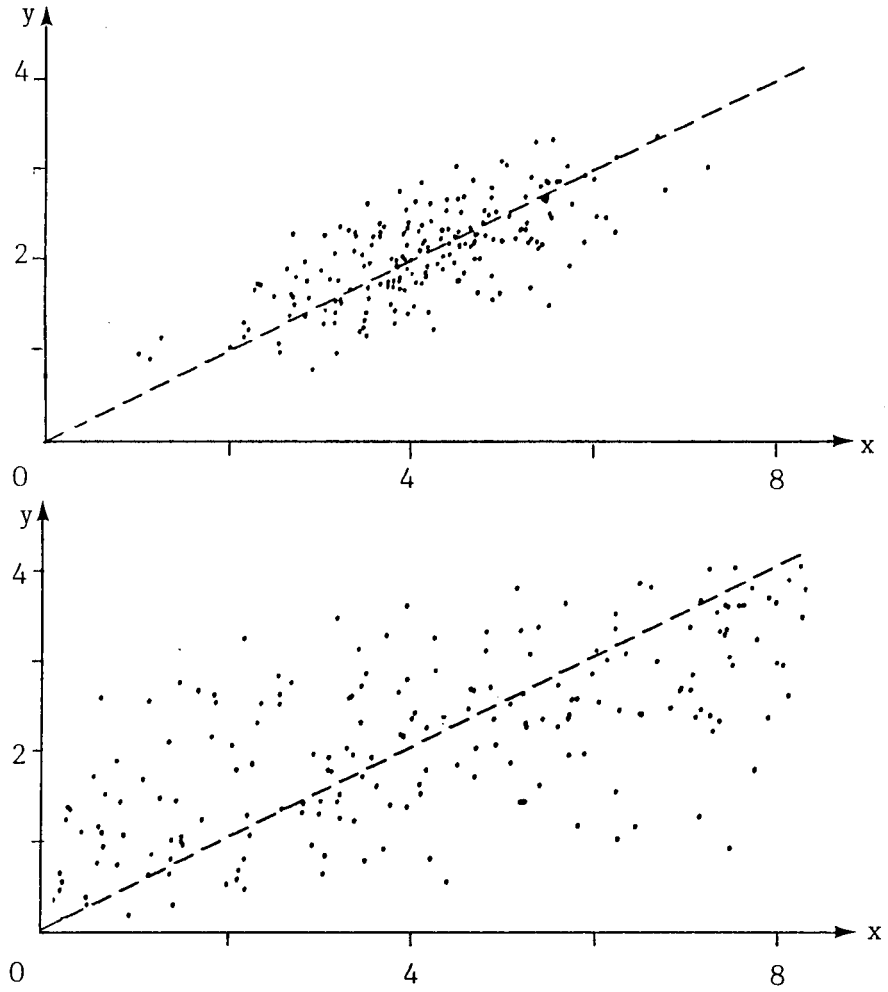
Das Berechnungsverfahren für den Korrelationskoeffizienten, wie es oben vorgeschlagen ist, ist nicht Standard. Die numerische Berechnung wird teilweise noch aufwendiger als mit der üblichen Formel:

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}=\frac{\sum xy-\frac{1}{n}(\sum x)(\sum y)}{\sqrt{[\sum x^2-\frac{1}{n}(\sum x)^2][\sum y^2-\frac{1}{n}(\sum y)^2]}}$$

Es geht mir hier mehr um das dadurch ermöglichte leichtere Verständnis der Formel: Schon bei den Punktwolken wurde zur Vereinheitlichung des Eindrucks aus dem Streubild "normiert", das findet seinen Niederschlag in der obigen Berechnungsprozedur. Für die tatsächliche Berechnung von Korrelationskoeffizienten wird man mit Vorteil Statistik-Taschenrechner verwenden, die ein eigenes Programm dafür haben. Zusammenfassend: Die obige Formel für den Korrelationskoeffizienten ist leichter einsichtig - die übliche Formel ist leichter numerisch auszuwerten. Ich greife auf die übliche Formel nicht zurück, weil ich der Meinung bin, daß man mit dem Taschenrechner um sie herum kommt.

6. Korrelationskoeffizient und Punktwolken

Schätzen Sie den Korrelationskoeffizienten aus den folgenden beiden Punktwolken.



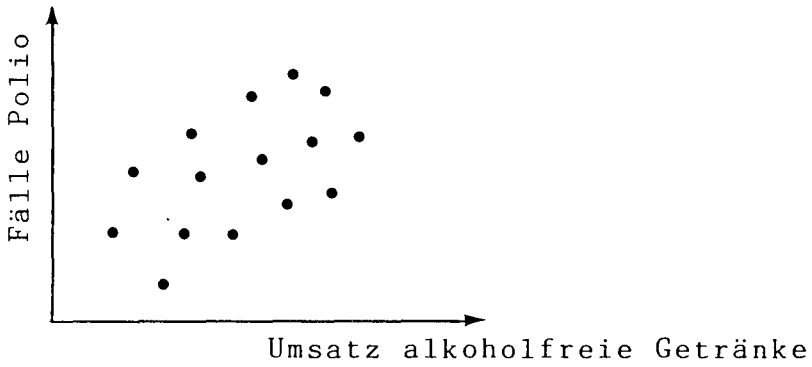
Der visuelle Eindruck aus der Punktwolke, wie eng sich diese um die s-Gerade gruppiert, hängt entscheidend von der Größe der Standardabweichungen s_x und s_y ab! Der Korrelationskoeffizient mißt nur relativ zu s_x und s_y. Wenn man in den beiden Bildern oben zu unterschiedlichen Schätzungen des Korrelationskoeffizienten kommt, dann liegt es daran, daß die Standardabweichungen unterschiedlich groß sind, der Korrelationskoeffizient r ist in beiden Fällen gleich 0,70! Wenn man aus Graphiken Korrelationskoeffizienten schätzen möchte, so muß man sich, um Täuschungen zu vermeiden, auf *normierte* Bilder beziehen.

B. Vorsicht mit der Interpretation von Zusammenhängen über den Korrelationskoeffizienten

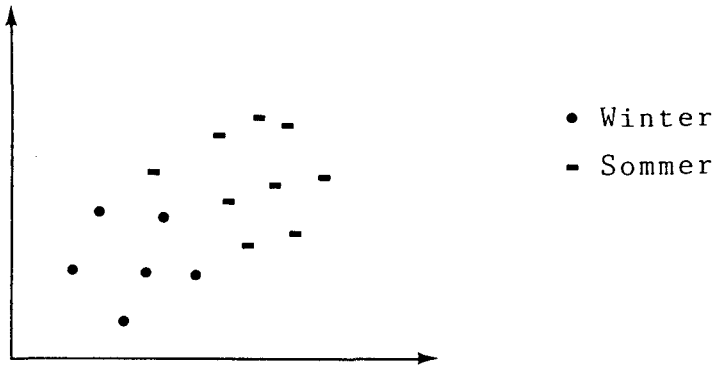
1. Statistischer Zusammenhang und Kausalität

Beispiel:

Vor der Einführung der Polioschutzimpfung (ca. 1954) hat man verschiedenste Einflüsse auf das Auftreten von Kinderlähmung untersucht. Ein interessantes Bild zeigt die Gegenüberstellung von Umsatz an alkoholfreien Getränken und der Zahl der Fälle von Polio pro Woche:



Die Punktwolke zeigt einen deutlichen positiven Zusammenhang, der Korrelationskoeffizient unterstreicht dies: r = 0,60.
Verursachen alkoholfreie Getränke Kinderlähmung? Das hat man zu keiner Zeit ernst genommen. Zu offensichtlich absurd ist die Unterstellung eines solchen Zusammenhangs. Was ist passiert? Trennt man die Daten nach einer dritten Variablen, nämlich der Jahreszeit (Sommer-Winter), so ergibt sich folgendes Bild:



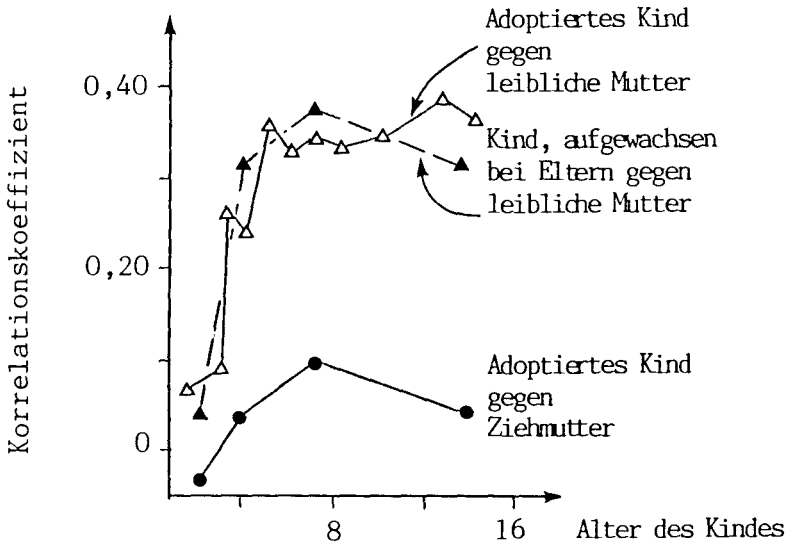
Für Winter und Sommer getrennt, ergibt sich keine nennenswert geformte Punktwolke, $r \approx 0{,}00$. Viele Korrelationsphänomene zerfließen in ähnlicher Weise, wenn man eine dritte Variable kontrolliert. Korrelation mißt nur die Stärke eines statistischen Zusammenhangs und zeigt nicht an, ob diese Beziehung kausal ist, zeigt also nicht, daß die unabhängige Variable die abhängige Variable kausal beeinflußt. Hat man Daten aus einer Beobachtungsstudie, man bekommt die Daten einfach so herein, ohne auf sie Einfluß nehmen zu können, so ist ein hoher Korrelationskoeffizient allein noch nicht ausreichend. Hat man hingegen eine experimentielle Studie, d.h. kann man das Niveau der x-Werte kontrollieren, und beobachtet dann den y-Wert, so kann man aus einem hohen Korrelationskoeffizienten viel eher kausale Beziehungen zwischen y und x in Betracht ziehen.

Aufgabe:

a) Vermuten Sie eine positive oder negative Korrelation zwischen Blutdruck und Alter?
b) Zwischen Einkommen und Alter?
c) Man hat eine positive Korrelation zwischen Blutdruck und Einkommen gefunden. Wie erklären Sie sich das?

2. Wirkungen, die die Korrelationsrechnung nicht erfaßt

Beispiel: Vererbungsstudie zwischen Müttern, Ziehmüttern, Kindern und Adoptivkindern (ca. 1935)



Der Verlauf der Korrelationskoeffizienten mit zunehmendem Alter der Kinder zeigt Zusammenhänge zwischen der Intelligenz der Kinder mit der ihrer leiblichen Mutter, obwohl sie bei einer Ziehmutter aufgewachsen sind, kaum jedoch Zusammenhänge mit ihrer Ziehmutter. Ist Intelligenz nur vererbt? Hat die Umwelt (Ziehmutter) keinen Einfluß?

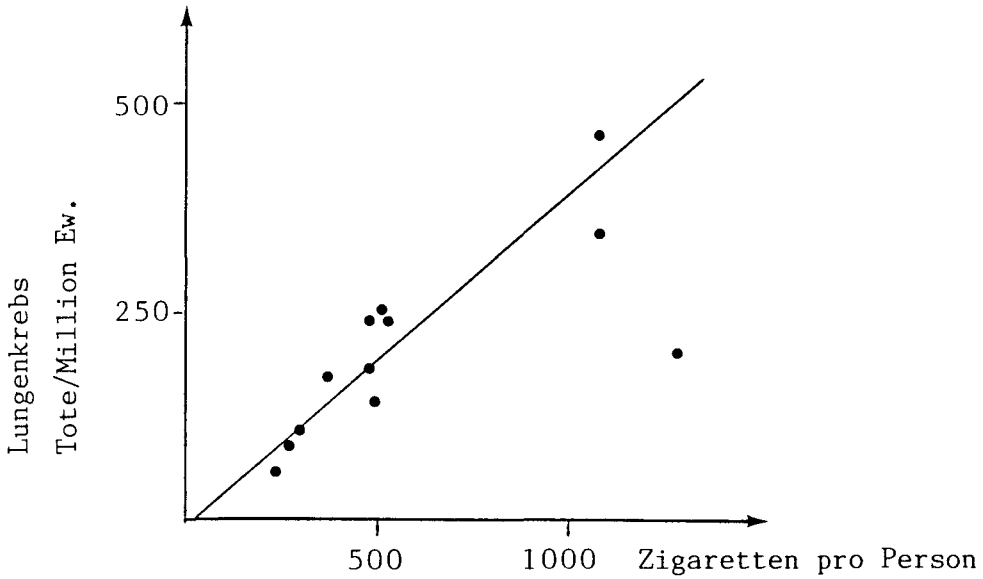
Eine nähere Inspektion zeigt jedoch für die mittleren Intelligenzquotienten (IQ) der leiblichen Mütter 86, für deren Kinder, die bei Ziehmüttern aufgewachsen sind, jedoch 106! Die Kinder haben im Durchschnitt 20 IQ-Punkte, und das alle ziemlich gleichmäßig, dazugewonnen. Dieser Einfluß der Umwelt ergibt nur einen "Shift" der Punktwolke nach oben, ohne deren "Dicke" zu beeinflussen, d.h. eine solche Wirkung kann mittels des Korrelationskoeffizienten gar nicht gemessen werden.

3. Korrelation von Mittelwerten

Beispiel: Studie: Rauchen und Lungenkrebs (1955)

In Ermangelung von Daten für Einzelpersonen bezüglich des Rauch-

verhaltens und dem Auftreten von Lungenkrebs hat man den Nachweis für die Schädlichkeit des Rauchens so zu führen versucht:

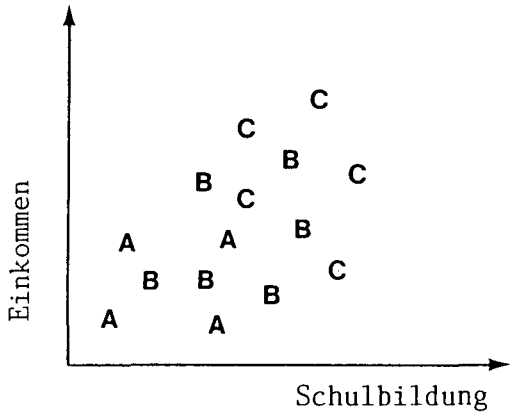


Der Punkt ganz rechts außen z.B. entspricht den USA mit dem Konsum von 1300 Zigaretten pro Einwohner im Jahre 1930 und 200 Toten an Lungenkrebs pro eine Million Einwohner im Jahre 1950. Die eingezeichnete Gerade könnte nicht besser passen, sieht man von den USA ab. Ist somit der Nachweis erbracht, daß Rauchen Lungenkrebs verursacht?

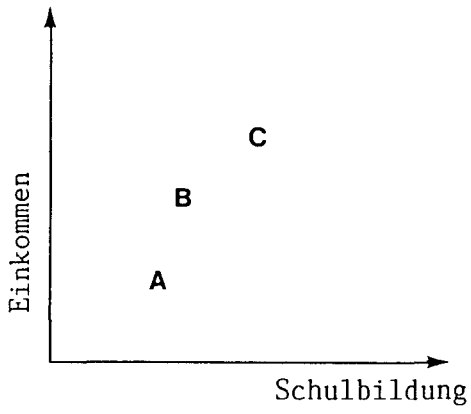
Das folgende Beispiel soll demonstrieren, daß man mit diesbezüglichen Schlüssen sehr vorsichtig sein muß.

Beispiel: Schulbildung - Einkommen in den Regionen A, B, C.

Punktwolke für einzelne Personen



Punktwolke für die Mittelwerte für die Regionen



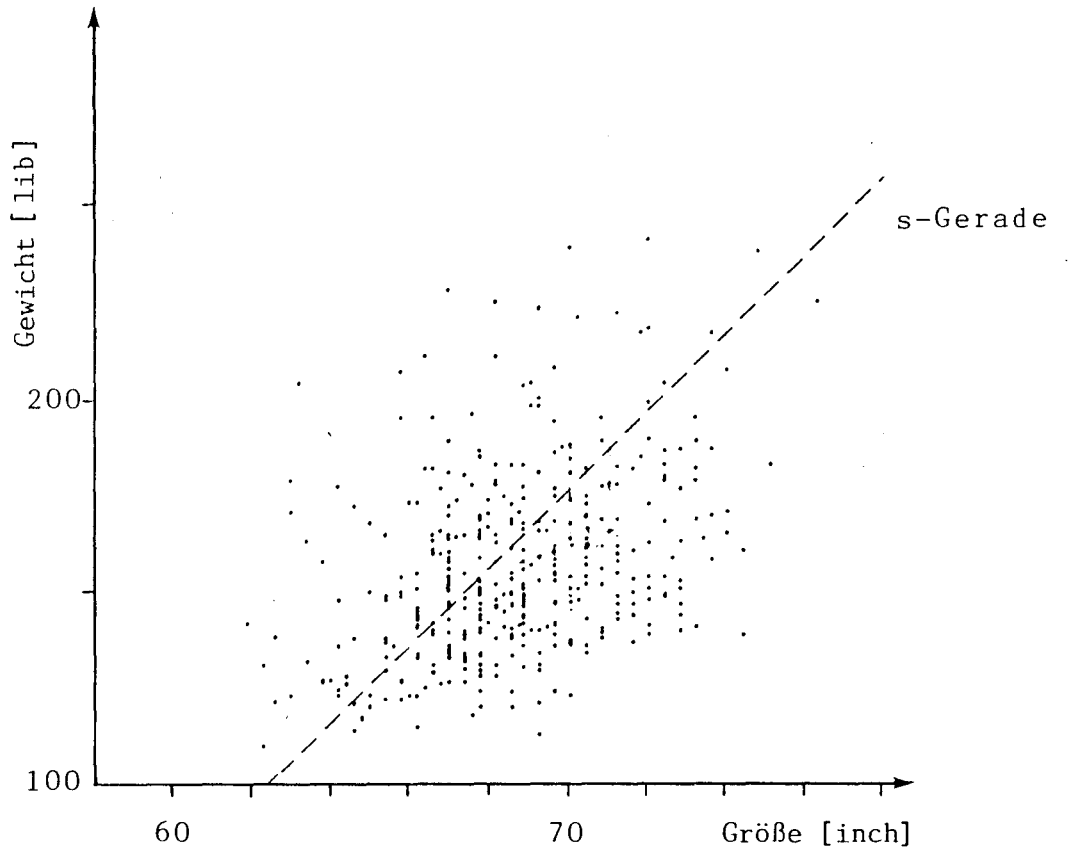
Die Mittelbildung erbringt eine viel engere Gruppierung der Punkte um die s-Gerade - der Korrelationskoeffizient aus den gemittelten Daten ist viel zu groß, gemessen am eigentlichen Streuverhalten der Daten.

C. Regression

1. Graph der Durchschnitte

Beispiel:

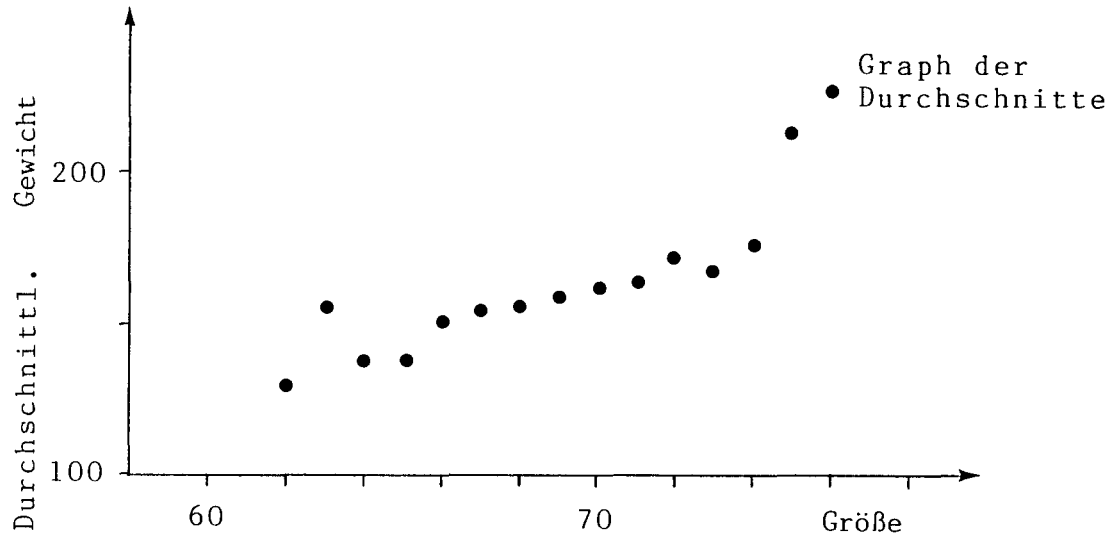
Aus einer US-Gesundheitsstudie entnimmt man u. a. folgende Daten für Körpergröße und Körpergewicht für 18-24jährige Männer (n = 411)



Mittlere Größe $\bar{x}$ = 68 [inch] , mittleres Gewicht 158 [lib]. Im Bild eingezeichnet ist auch die s-Gerade, um die sich die Punktwolke "symmetrisch" gruppiert. Aus der Punktwolke erkennt man einen (losen) positiven Zusammenhang zwischen Größe und Gewicht. Männer, die z. B. ein inch über dem Mittelwert lagen,

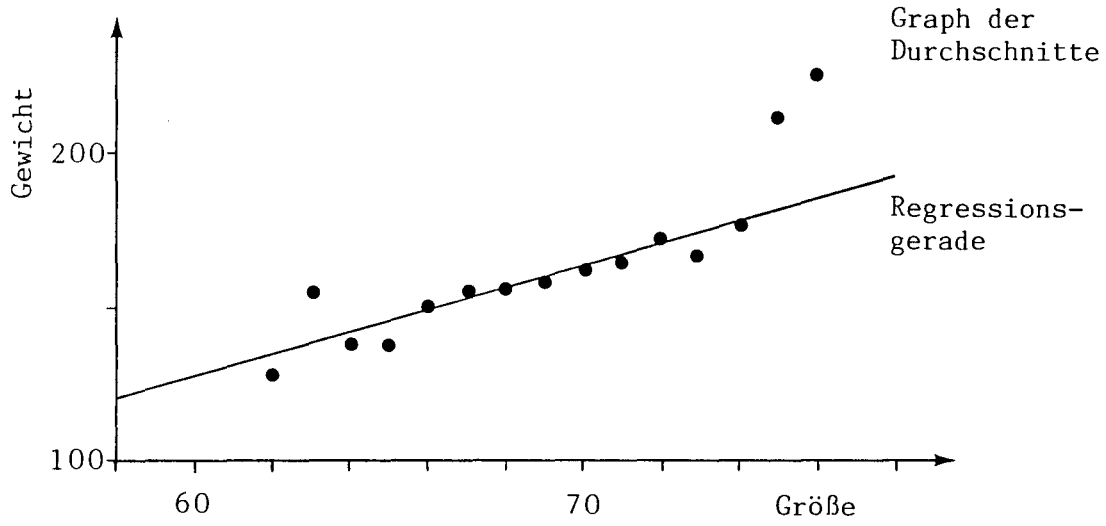
d. h. x = 69, hatten auch etwas mehr Gewicht als der Mittelwert $\bar{y}$ = 158, im Durchschnitt. Das trifft in stärkerem Ausmaß auf jene zu, die 2 inch über dem Mittelwert bezüglich der Größe lagen. Im Durchschnitt: Wieviel Zuwachs an Gewicht hängt mit einem Zuwachs einer Einheit Größe zusammen?

Wir ziehen Streifen von 1 inch Breite parallel zur zweiten Achse um jeden ganzzahligen Wert für die Körpergröße, z. B. für x = 64 und betrachten alle Punkte, die in diesem Streifen liegen, das sind allejene Männer, die Körpergröße x = 64 inch haben. Der Durchschnitt des Gewichts dieser Männer ist $\bar{y}_{64} \approx 138$ [lib]. Der Graph der Durchschnitte zeigt den Verlauf der so berechneten Mittelwerte mit zunehmender Größe und gibt ein gutes Bild der Zusammenhänge zwischen Körpergröße und Körpergewicht.

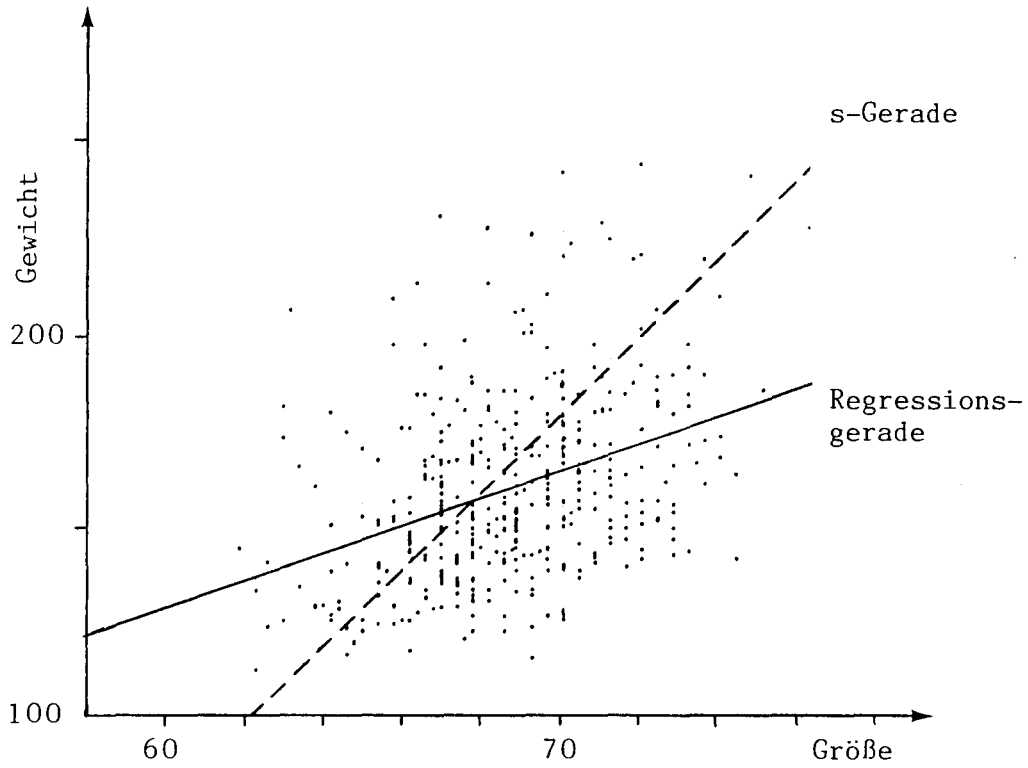


2. Regressionsgerade zur Beschreibung des Zusammenhangs zwischen unabhängiger und abhängiger Variablen

Der Graph der Durchschnitte gibt vielleicht Fluktuationen wieder, die nur der Stichprobe zuzuschreiben sind, die aber nicht *generalisierbar* sind. Man kann versuchen, eine Gerade anzupassen:



Nimmt man die ursprüngliche Punktwolke, so ersieht man daraus, wie diese Gerade, die sogenannte Regressionsgerade (die geglättete Version des Graphs der Durchschnitte) von der s-Geraden abweicht und wie sie die Punktwolke zusammenfaßt und so den Zusammenhang zwischen Größe und Gewicht beschreibt.



Die Datenzusammenfassung ergibt

Größe:	$\bar{x}$	$\approx$ 68	$s_x \approx$ 2,5	[inch]
Gewicht:	$\bar{y}$	$\approx$158	$s_y \approx$ 25	[lib]
	r	$\approx$ 0,36		

s_x und s_y wurden in der Darstellung durch dieselbe Einheit dargestellt (genormtes Bild), die s-Gerade hat daher eine Steigung von 45°. Die Größe der Streuung der Punktwolke um die s-Gerade wird durch r≈0,36 entsprechend numerisch ausgewiesen.

Der Graph der Durchschnitte ist schon systematisch von der s-Geraden abgewichen. Die Regressionsgerade erlaubt, eine Schätzung für die Mittelwerte der zweiten Variablen zu ermitteln, wenn man sich auf jene Männer bezieht, die eine ganz bestimmte Größe haben. Wie geht man vor?

Bestimme Regressionsschätzung für jene Männer, deren Größe $1 \cdot s_x$ über dem Mittelwert $\bar{x}$ bezüglich der Größe liegt (x=70,5). Ist ein solcher Mann darunter, der auch $1.s_y$ über $\bar{y}$ bezüglich des Gewichts liegt, so liegt sein Punkt auf der s-Geraden. Bei der Ermittlung des Graphen der Durchschnitte haben wir schon gesehen, daß viel mehr Punkte *unter* der s-Geraden als über ihr liegen, der Mittelwert $\bar{y}_{70,5}$ wird daher kleiner als der Punkt auf der s-Geraden sein müssen. Um welchen Faktor kleiner? Ungefähr um den Faktor r!

Größe

x $1.s_x$ über $\bar{x}$

$x = \bar{x} + 1.s_x =$

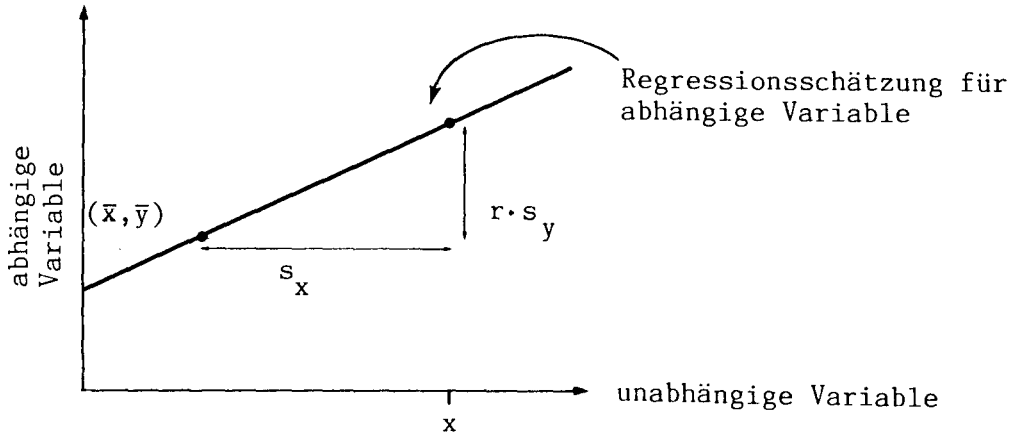
$= 68 + 2,5 = 70,5$ [inch]

Gewicht - Regressionschätzung

y $r.s_y$ über $\bar{y}$

$y = \bar{y} + r.s_y =$

$= 158 + 0,36 \cdot 25 = 167$ [lbs]

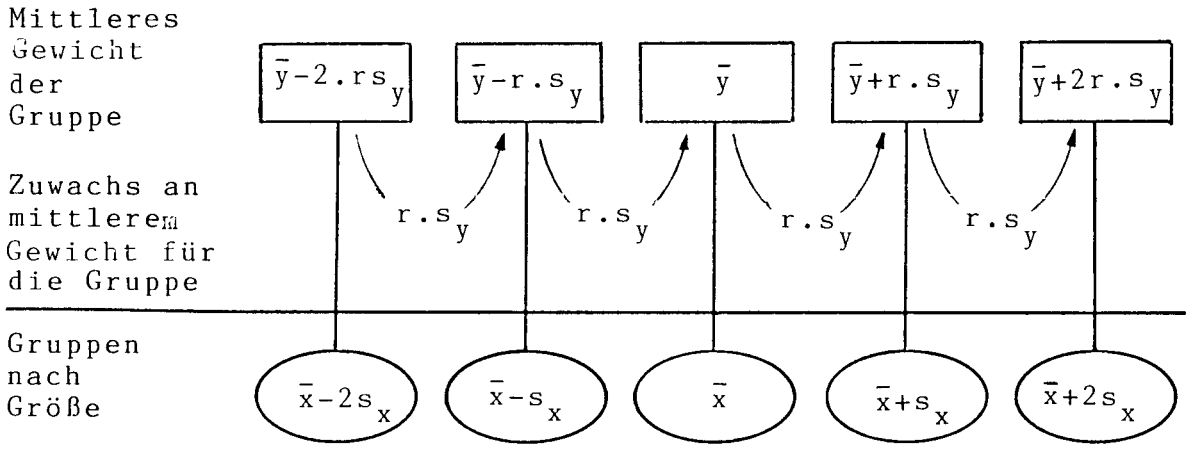


Gleichung der Regressionsgeraden

$$(r) \qquad \frac{y - \bar{y}}{s_y} = r \cdot \frac{x - \bar{x}}{s_x}$$

Warum taucht hier r auf? Das ist erst mit mehr Mathematik zu begründen (ist eine Konsequenz der sogenannten Methode der kleinsten Quadrate). Plausibel sind die Fälle r = 0 bzw. r = 1 (r = -1).

Folgende Vorstellung hilft, die Auswirkung der Regressionsschätzung zu verdeutlichen:
Man hat die Männer nach Größe gruppiert und geht die Gruppen der Größe nach durch. Wie verändert sich das mittlere Gewicht für die Gruppe, wenn man fortschreitet?



Beispiel:
Bestimmen Sie die Regressionsschätzung für den Blutdruck, für Männer mit Körpergröße x = 72 [inch].

Größe	$\bar{x} \approx 68$	$s_x \approx 2,5$	[inch]
Blutdruck	$\bar{y} \approx 120$	$s_y \approx 15$	[mm]
Korrelationskoeffizient		r	$-0,2$	

Regressionsschätzung $\hat{y}$?

$$\frac{72 - 68}{2,5} = \frac{4}{2,5} = 1,6 \qquad \hat{y} = \bar{y} + r \cdot 1,6 s_y = 120 - 0,2 \cdot 1,6 \cdot 15 = 115,2$$

3. Die Regressionsfalle

Beispiel:

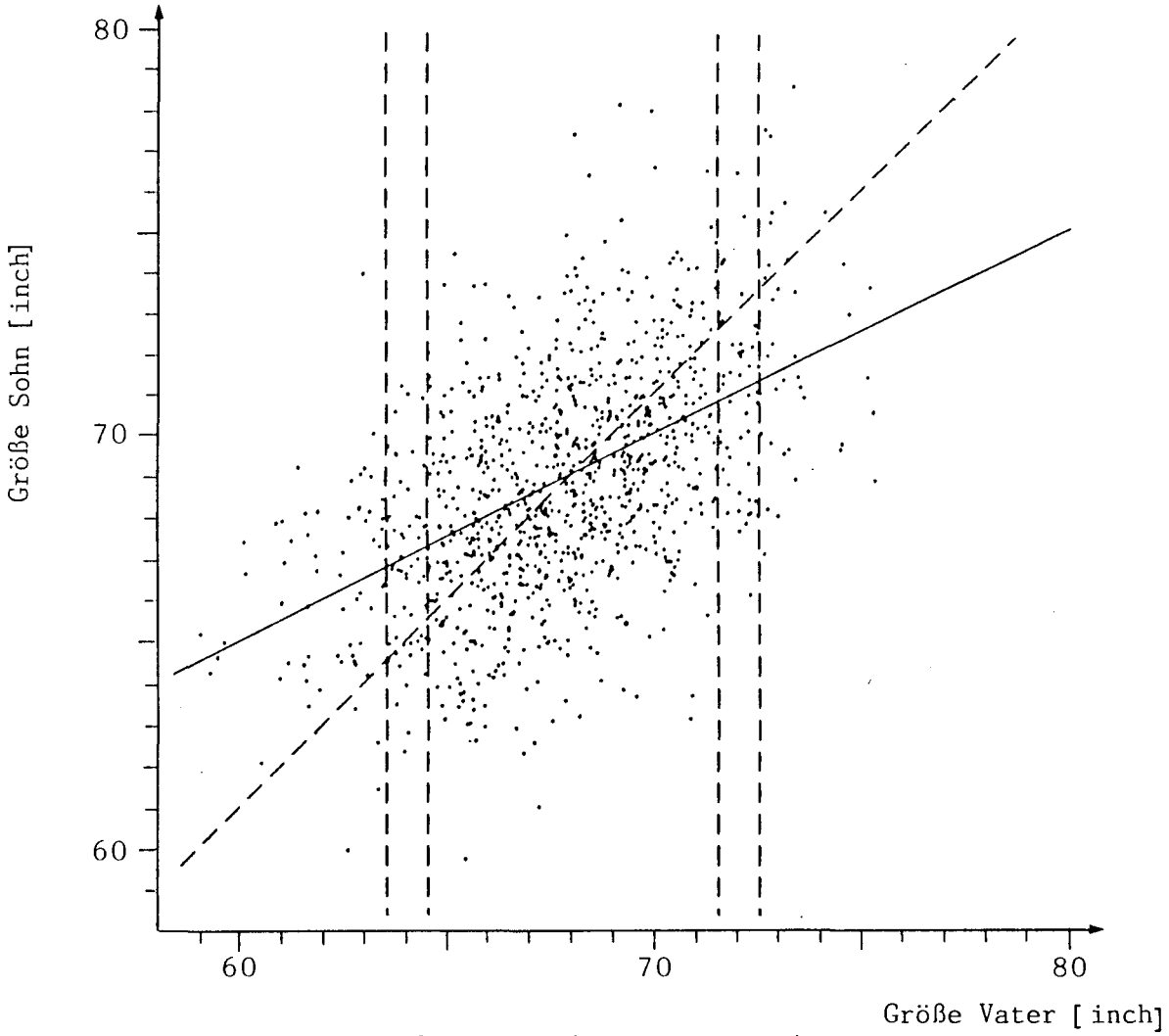
Ein Vorschulprogramm zur Erhöhung des Intelligenzquotienten soll auf Wirksamkeit geprüft werden. Dazu wird ein Vortest (vor dem Programm) und ein Nachtest (nachher) vorgenommen.

Vortest: $\bar{x} \approx 100$ $s_x \approx 15$
Nachtest: $\bar{y} \approx 100$ $s_y \approx 15$ r klein

Das Programm hat, gemessen am kleinen Korrelationskoeffizienten, keine Wirkung.

Aber:
Die Kinder, die im Vortest unter dem Mittelwert $\bar{x}$ lagen, hatten im Nachtest einen durchschnittlichen Zuwachs von 5 Punkten zu verzeichnen, jene, die über $\bar{x}$ lagen, eine Abnahme von durchschnittlich 5 Punkten. Macht das Programm die Kinder gleicher?

Beispiel:
Körpergrößen von Vätern und Söhnen (Fortsetzung)- Daten von K. Pearson (1903). Die summarische Beschreibung der Daten durch fünf Zahlen sei nocheinmal wiedergegeben:
Väter: $\bar{x} \approx 68$ $s_x \approx 2,7$
Söhne: $\bar{y} \approx 69$ $s_y \approx 2,7$ $r \approx 0,50$
Anhand dieser Daten soll das im vorhergehenden Beispiel enthaltene Phänomen, eine Art "Rückschreiten zum Mittelwert", näher beleuchtet werden.



Voraussage mittels s-Geraden (strichliert)

Vater $x_i = 72$ → Sohn $y_i = 73$
Vater $x_i = 64$ → Sohn $y_i = 65$

Im Streifen x = 72 sind jedoch die meisten Punkte *unter* der s-Geraden, der Mittelwert dieser Punkte ist nicht 73 sondern 71. Im Streifen x = 64 : Nicht 65 sondern 67 für die Söhne im Durchschnitt. Die Regressionsgerade, die den Graphen der Durchschnitte glätten soll, hat einen flacheren Anstieg als die s-Gerade.

Große Väter (gutes Ergebnis im Vortest) haben im Durchschnitt kleinere Söhne (nicht ganz so gutes Ergebnis im Nachtest).
Die Punktwolke ist symmetrisch um die s-Gerade wie eine Ellipse

zu ihrer Hauptachse, die Punktwolke ist in Streifen, die parallel zur 2. Achse gezogen sind, nicht symmetrisch zur s-Geraden, sondern, in gewissem Sinne, symmetrisch zur Regressionsgeraden. Schon Galton hat dieses Phänomen herausgearbeitet und es im Zusammenhang mit seinen Studien zur Vererblichkeit von körperlichen Merkmalen (u.a. auch die Körpergröße) als "Regression to mediocrity", als Rückbildung hin zum Mittelwert, bezeichnet (1877,1885). Dies ist kein besonderes Phänomen, sondern ein künstliches Artefakt der Regressionsrechnung. Die Regressionsfalle besteht nun darin, dieses Artefakt den Sachumständen zuzuschreiben. Etwa im Vortest-Nachtest-Beispiel der Wirkung des Schulprogramms.

4. Verbesserte Prognosen durch die Korrelationsrechnung

Sei x die unabhängige Variable, y die abhängige. Weiß man nun nichts über den Wert von x, so kann man die Verteilung der y-Daten prägnant mittels

$$\bar{y} \text{ und } s_y$$

beschreiben. Hat man den y-Wert einer "Person" vorauszusagen, so gibt man am besten $\bar{y}$ an, wobei die Unsicherheit der Voraussage mit s_y "gemessen" werden kann:
Das Intervall

$$[\bar{y} - 2s_y, \bar{y} + 2s_y]$$

enthält ca. 95 % (bei symmetrischen und eingipfeligen Verteilungen, die Normalverteilung ist ein Prototyp davon) der y-Daten.

Kennt man nun den x-Wert der Person, so kann man die Voraussage verbessern, umso präziser, je größer der Korrelationskoeffizient ist. Man sagt nun nicht mehr den allgemeinen Mittelwert $\bar{y}$ voraus, sondern ersetzt diesen durch die Regressionsschätzung $\hat{y}$, die sich aus der Regressionsgleichung ergibt:

$$\frac{\hat{y} - \bar{y}}{s_y} = r \cdot \frac{x - \bar{x}}{s_x} .$$

Die Unsicherheit der Voraussage kann nun mit $\sqrt{1 - r^2} \cdot s_y$ angegeben werden, d.h. $[\bar{y}-2\sqrt{1-r^2}\, s_y, \bar{y}+2\sqrt{1-r^2}\, s_y]$, ist nun ein Intervall, das ca. 95 % aller Punkte der Punktwolke, innerhalb eines vertikalen Streifens um x, enthält. Auch das kann erst innerhalb eines enger gesteckten mathematischen Modells erfaßt und bewiesen werden.

5. Schlußbemerkungen

Die Methoden der Korrelationsrechnung und Regressionsrechnung wurden gleich in einem mathematischen Gewand entwickelt. Galton war zwar primär am Vererbungskontext interessiert, aber er benötigte Methoden zum "strengen Nachweis" der Erblichkeit wesentlicher Merkmale. Dazu engagierte er einen Mathematiker. Obwohl die numerischen Korrelationskoeffizienten alle nur in der Größenordnung von 0,50 lagen, war das ein schlagendes Argument in der Meinung der Vererbungstheoretiker. Die Methoden wurden verfeinert und modifiziert. Der Kontingenz-Koeffizient für qualitative Variable z.B. ist nach dem Vorbild des Korrelationskoeffizienten entstanden. Heute stellt die Regressions- und Korrelationsrechnung ein vielfach eingesetztes, aber auch häufig mißverstandenes Gebiet der Statistik dar.

L i t e r a t u r

BOROVCNIK, M. und FISCHER, R.: D. A. Mackenzie: Statistics in Britain - 1865-1930 - The Social Construction of Scientific Knowledge. In: Educational Studies in Mathematics 14 (1983), 101-104.

FREEDMAN , D., PISANI, R. und PURVES, S.: Statistics. London: W.W. Norton 1978.

MACKENZIE, D. A.: Statistics in Britain - 1865-1930 - The Social Construction of Scientific Knowledge. Edingburgh: Edingburgh University Press 1981.

Leitende Ideen und Beispiele habe ich aus dem Buch von Freedman e. a. übernommen. Dieses Lehrbuch ist für Anwender gedacht und wirklich lesenswert.