Statistik

Corona - Notversion

M. Oettinger

13.05.2020

Bivariate Verteilungen

Bisher wurden lediglich Verteilungen eines einzelnen Merkmals $\{x_i\}$ (z.B. die Körpergröße bestimmter Personen) betrachtet - sog. univariate Verteilungen.

Bei Stichproben, die mehrere Merkmalswerte erfassen (die prinzipiell voneinander abhängig sein können), spricht man von multivariaten Verteilungen. Der Einfachheit halber werden wir uns hier auf Verteilungen zweier Merkmale, sog. bivariate Verteilungen $\{x_i, y_i\}$, beschränken. Die interessante Frage in der deskriptiven Statistik ist hier natürlich die nach der Beziehung der einzelnen Merkmale untereinander - sie können voneinander abhängig sein.

Bivariate Verteilungen

Eine Darstellungsform bivariater Verteilungen sind Kontingenz- oder Kreuztabellen. Die Daten werden zweidimensional für die beiden Merkmale in Zeilen bzw. Spalten dargestellt.

Merkmal Y	Merkmal X					
	x_1	...	x_j	...	x_k	$\sum$
y_1	h_{11}	...	h_{1j}	...	h_{1k}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_1	h_{i1}	...	h_{ij}	...	h_{ik}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_1	h_{m1}	...	h_{mj}	...	h_{mk}	$n_{m.}$
$\sum$	$n_{.1}$	...	$n_{.j}$	...	$n_{.k}$	n

Hier bedeuten x_i die i-te Merkmalsausprägung (bzw. Merkmalswert) des Merkmals X, h_{ij} die absoluten Häufigkeiten, mit denen die Ausprägungen x_j und y_i auftreten.

Bivariate Verteilungen

Berechnet man die relativen Häufigkeiten in den einzelnen Spalten (bzw. Zeilen), erhält man eine Kreuztabelle relativer Häufigkeiten, in der jede Spalte (Zeile) die Verteilung eines Merkmals bezogen auf das jeweils andere enthält.

Durch diese Art der Darstellung lassen sich Häufungen von Paaren feststellen. Falls die erhobenen Merkmale unabhängig voneinander sind, erwartet man bei einer solchen Auftragung von *relativen* Häufigkeiten dieselben Werte in den jeweiligen Spalten bzw. Zeilen (eben gerade, weil die *relativen* Häufigkeiten unabhängig vom jeweiligen zweiten Merkmal dieselben sind).

Bivariate Verteilungen

Beispiel: Abhängigkeit der Mathematiknote vom Vertiefungsfach.

Untersucht werden soll die Abhängigkeit der Mathematiknote vom jeweiligen Vertiefungsfach. Dazu wurden 25 Studenten zu ihrem Vertiefungsfach und zur im letzten Semester erzielten Mathematiknote befragt und die Zahl der Studenten mit Vertiefungsfach Y und Mathematiknote X notiert:

Mathenote	Vertiefungsfach			
Y	Hallen-Halma	Häkeln	Zitronenfalten	
2	1	6	3	10
3	4	3	3	10
4	0	1	4	5
$\sum$	5	10	10	25

Die relativen Häufigkeiten je Spalte liefern einen Hinweis auf die Abhängigkeit der Größen.

Beispiel: Abhängigkeit der Mathematiknote vom Vertiefungsfach.

Nach Berechnung der relativen Häufigkeiten ergibt sich das folgende Bild:

Mathenote Y	Vertiefungsfach Hallen-Halma	Häkeln	Zitronenfalten	
2	0,2	0,6	0,3	0,4
3	0,8	0,3	0,3	0,4
4	0	0,1	0,4	0,2
$\sum$	1	1	1	1

Tabelle: Mathenoten und Vertiefungsfach

Innerhalb der Zeilen der Verteilung relativer Häufigkeiten ergeben sich unterschiedliche Werte. Die Mathematiknote, die ein Student im letzten Semester erzielte, war also nicht von seinem Vertiefungsfach unabhängig.

bivariate Verteilungen

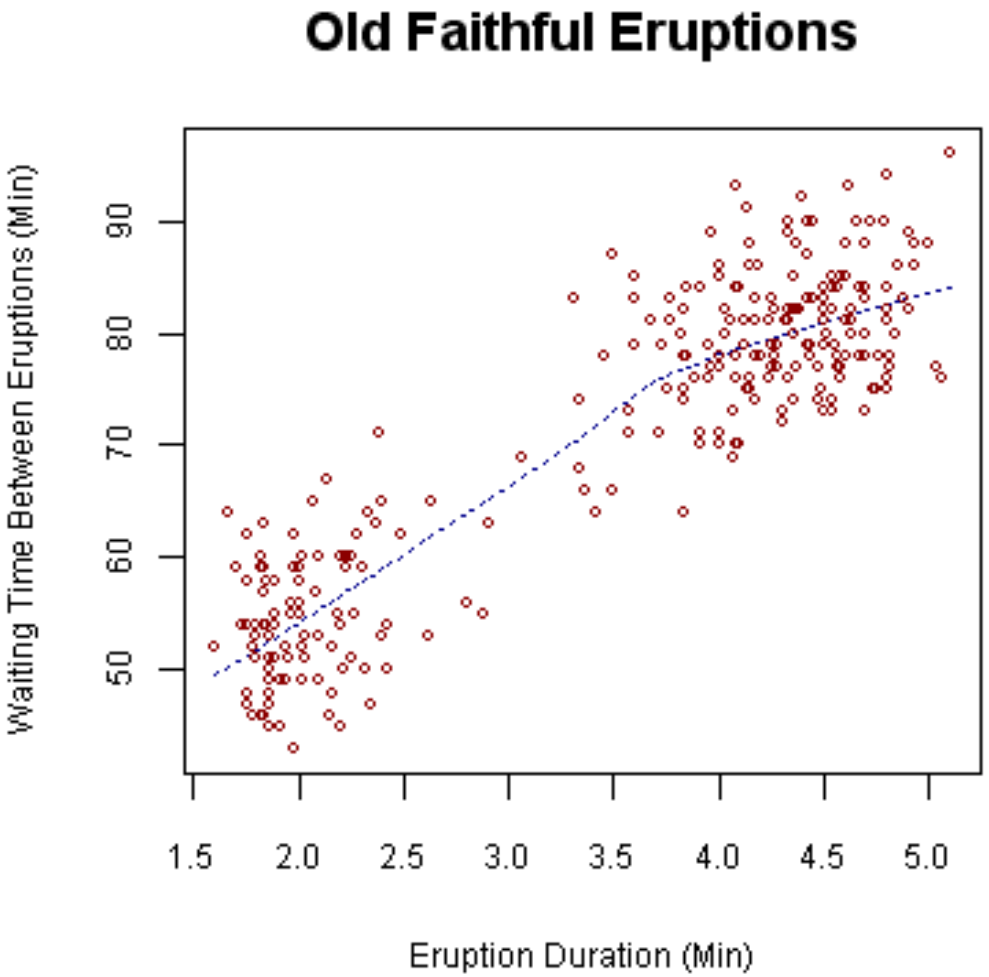
Bei vielen Beobachtungswerten stetiger Merkmale ist eine grafische Darstellung als Streudiagramm übersichtlicher. Merkmalswerte des Merkmals Y werden in einem Koordinatensystem über denen des Merkmals X aufgetragen.

In einer solchen Darstellung zeigt sich eine Abhängigkeit der Merkmale voneinander in einer Häufung von Datenpunkten entlang von Geraden (bei linearer Abhängigkeit) oder allgemeinerer Funktionen. Wir wollen natürlich versuchen, diese Funktionen aus dem vorliegenden Datenmaterial zu bestimmen. Dazu betrachten wir zunächst ein einfaches Beispiel einer bivariaten Stichprobe.

bivariate Verteilungen

Rechts abgebildet sind reale Messwerte des Geysirs 'Old faithful' zur Pause zwischen den Ausbrüchen und der Dauer der Ausbrüche [1].

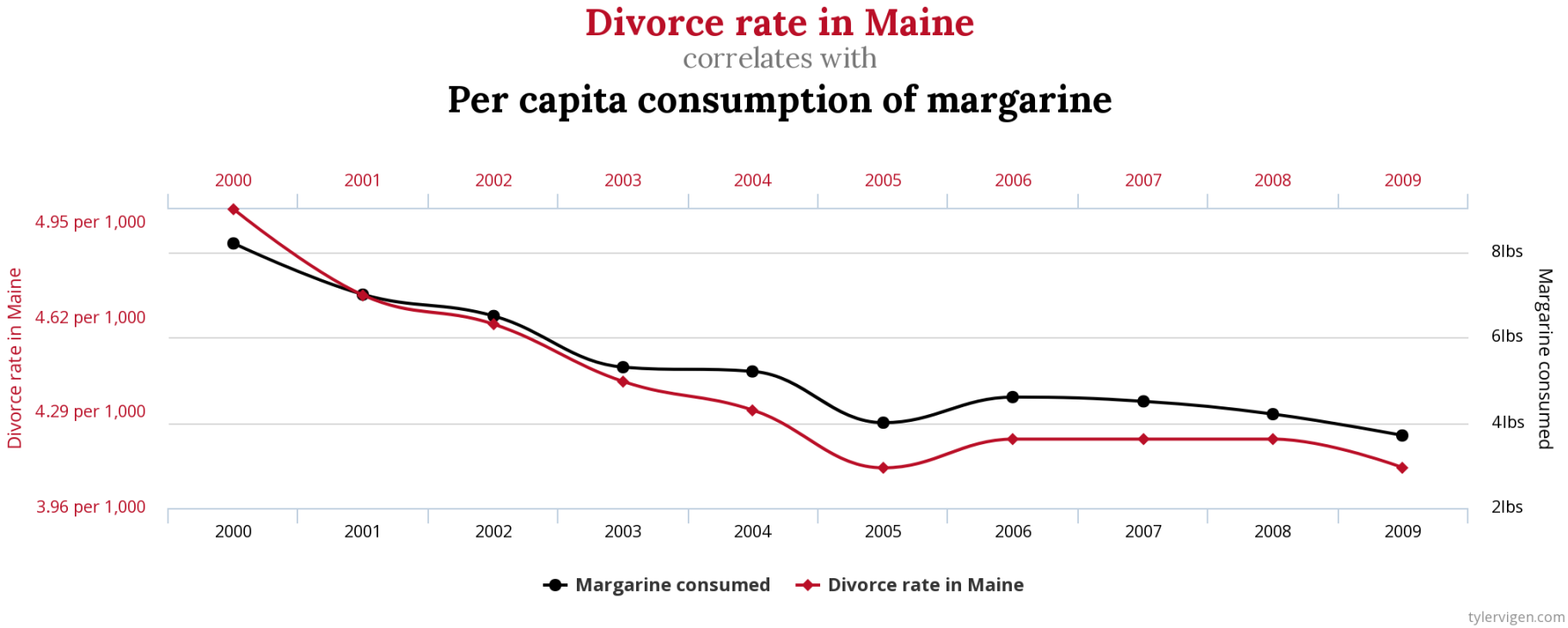
Die Verteilung der Daten legt einen Zusammenhang zwischen Dauer und Pause nahe.



[1] http://en.wikipedia.org/wiki/File:Oldfaithful3.png, public domain

bivariate Verteilungen

Dabei sollte stets klar sein, dass die Statistik nur Hinweise auf einen Zusammenhang zwischen Daten liefern kann. In diesem Beispiel ist das sehr !ahnliche Verhalten rein zufälliger Natur!



[2] http://www.tylervigen.com/spurious-correlations, Daten: National Vital Statistics Reports, U.S. Department of Agriculture

Beispiel: lineare Regression.

Ein Hersteller von Dingen produziert mehrere Modelle (nummeriert mit dem Index i) zum Preis von jeweils x_i. Um den Absatz zu optimieren, soll eine Preis-Absatz-Funktion ermittelt werden.

	Preis	Menge			
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	20	0	5	-5	-25
2	16	3	1	-2	-2
3	15	7	0	2	0
4	16	4	1	-1	-1
5	13	6	-2	1	-2
6	10	10	-5	5	-25
Summe	90	30	0	0	-55
Mittelwert	15	5			

Tabelle: Preise verschiedener Modelle von Dingen

Beispiel: lineare Regression.

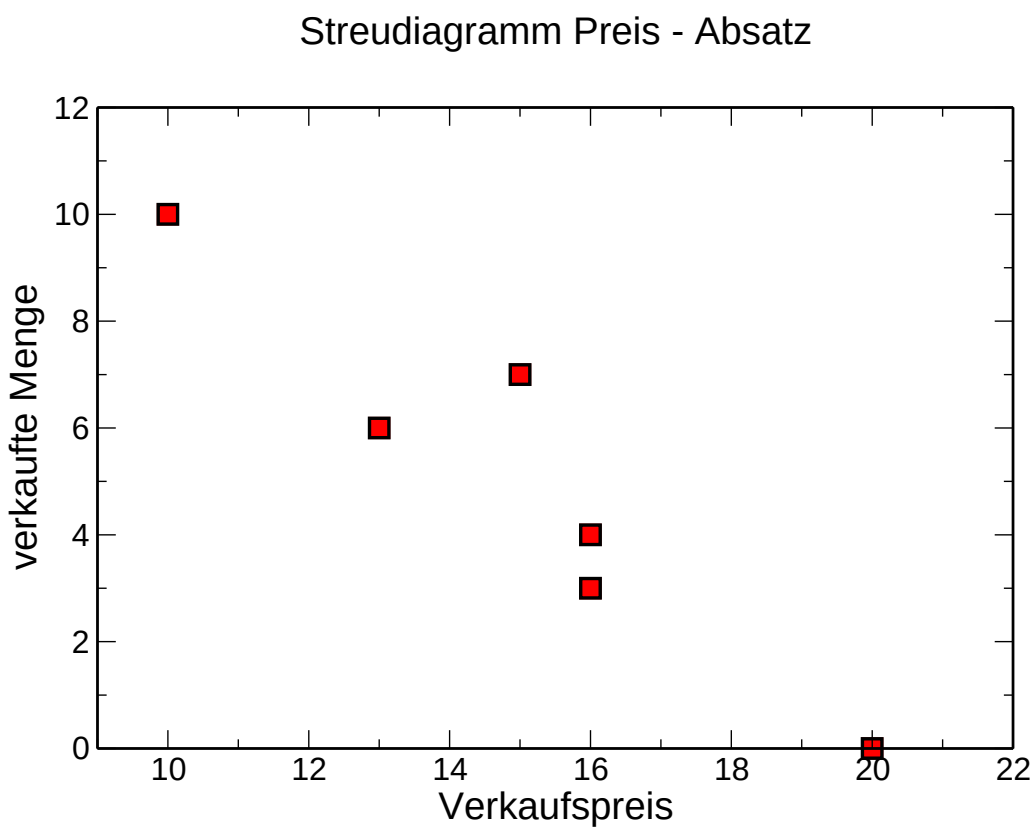


Abbildung: Merkmalswerte zum Beispiel des Herstellers von Dingen.

Im Streudiagramm kann man einen linearen Zusammenhang (eine Gerade) erahnen.

Die Merkmale sind miteinander über eine lineare Beziehung verknüpft.

Die Kovarianz

Die empirische Varianz einer Häufigkeitsverteilung ist

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

Sie baut auf den quadrierten Abweichungen $(x_i - \bar{x})^2$ der Daten vom Schwerpunkt des Datensatzes, dem arithmetischen Mittel $\bar{x}$, auf und enthält damit Information über die Streuung der Daten.

Die Kovarianz basiert auf einem ähnlichen Prinzip, sie berechnet sich aber aus den Abständen der Beobachtungspaare (x_i, y_i) von den jeweiligen arithmetischen Mitteln. Statt der Abweichungsquadrate werden die Produkte der Abweichungen der einzelnen Merkmale vom arithmetischen Mittel betrachtet.

Die Kovarianz

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

ist die empirische *Kovarianz*, sie lässt sich für kardinal skalierte Merkmale berechnen.

Die Kovarianz ermöglicht das Erkennen eines tendenziellen Zusammenhangs zwischen zwei Merkmalen. In den Daten aus dem Beispiel hat man den Eindruck, eine fallende Tendenz zu erkennen, die sich rechnerisch nachvollziehen lässt.

Zeichnet man in das Streudiagramm die Parallelen zu den Achsen durch die arithmetische Mittel der betrachteten Merkmale, erhält man den Schwerpunkt der bivariaten Verteilung (den Schnittpunkt der beiden Geraden mit den Koordinaten $(\bar{x}, \bar{y})$).

Die Kovarianz

Die dadurch entstandenen vier Teilbereiche werden im Uhrzeigersinn - rechts oben beginnend - als I. bis IV. Quadrant bezeichnet.

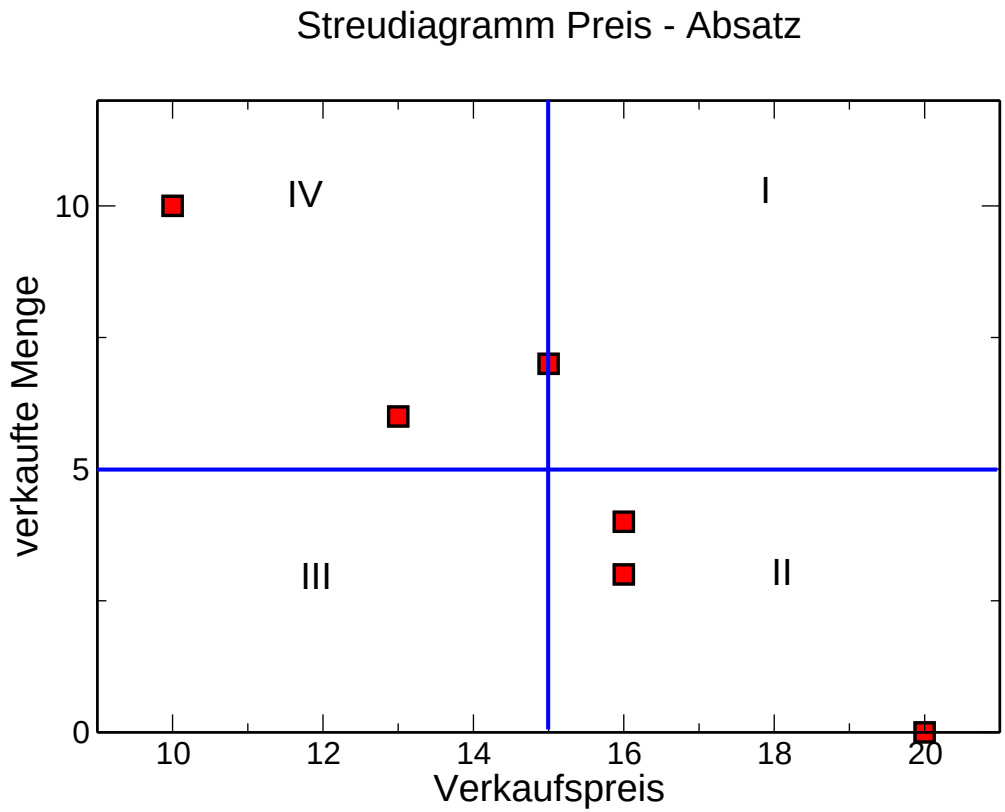


Abbildung: Vier-Quadranten-Schema am Beispiel des Herstellers von Dingen. Die Daten befinden sich in den Quadranten II und IV.

Die Kovarianz

Im Beispiel des Herstellers von Dingen liegt der überwiegende Teil der Punkte im zweiten und im vierten Quadranten. Dies ist offensichtlich Ausdruck der 'abwärtsgerichteten' Tendenz.
Allgemein gilt: liegen die Punkte des Streudiagramms hauptsächlich im

- I. und III. Quadranten, deutet dies auf einen positiven Zusammenhang zwischen den Merkmalen hin; mit steigenden Werten von x steigen auch die Werte von y. In diesem Fall ist die Kovarianz $s_{xy} > 0$.
- II. und IV. Quadranten, deutet dies auf einen negativen Zusammenhang zwischen den Merkmalen hin; mit steigenden Werten von x fallen die Werte von y. In diesem Fall ist $s_{xy} < 0$.
- Verteilen sich die Punkte in etwa gleichmäßig auf alle vier Quadranten, so deutet das daraufhin, dass kein Zusammenhang zwischen den Merkmalen besteht. Die Punktwolke hat eine diffuse Gestalt, es gilt $s_{xy} \approx 0$.

Die Kovarianz

Die Kovarianz gibt an, welche Tendenz der Zusammenhang zwischen den beiden Merkmalen besitzt. Sie ist aber allgemein dimensionsbehaftet und damit abhängig von der gewählten Skala, weshalb sie sich nicht vergleichen lässt.

Auch kann aus $s_{xy} = 0$ nicht geschlossen werden, dass kein Zusammenhang zwischen den beiden untersuchten Größen besteht. Einerseits hängt der Betrag der Kovarianz stark von der jeweiligen Skalierung der Merkmale ab, andererseits gibt es durchaus Zusammenhänge zwischen zwei Merkmalen, die sich nicht unbedingt durch die Kovarianz erfassen lassen (beispielsweise nichtlineare Zusammenhänge).

lineare Regression

Im Folgenden gehen wir davon aus, dass eine Datenreihe zweier Merkmale vorliegt, bei der ein Zusammenhang zwischen den Merkmalen existiert. Der Einfachheit halber wird ein linearer Zusammenhang angenommen:

$$y = a \cdot x + b$$

Es liegen also viele (n) unabhängige Angaben zum selben Sachverhalt vor (für jedes der n Paare):

$$y_i = a_i \cdot x + b_i$$

Da jedes einzelne Datenpaar mit einer (zufälligen) Abweichung vom realen Wert behaftet ist, kann für den Zusammenhang lediglich ein Schätzwert bestimmt werden. Gedanklich könnte jedes einzelne Paar von Punkten im Streudiagramm über eine Funktion (im einfachsten Fall eine Gerade) verbunden werden.

lineare Regression

Man erhält eine Anzahl von Funktionen, die sich wahrscheinlich ähnlich, aber nicht identisch sein werden - es existiert eine große Zahl von denkbaren Zusammenhängen (bei n Wertepaaren gibt es insgesamt $(n-1)!$ mögliche Geraden).

Welche der Geraden ist die Richtige?

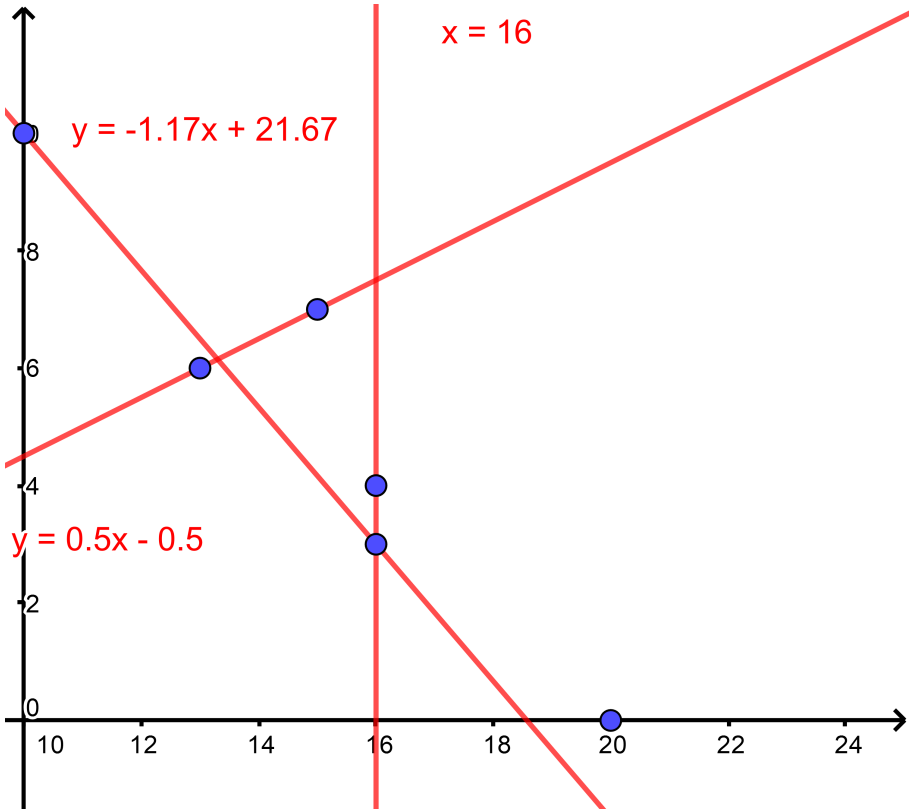


Abbildung: Einige mögliche Geraden $y = a_i x + b_i$

lineare Regression

Das Problem bei der Regression besteht darin, ein Kriterium für eine gut angepasste Gerade zu finden. Prinzipiell sind immer mehrere Möglichkeiten des 'Ausgleichens' solcher redundanten Messungen denkbar, beispielsweise könnte über die Steigungen a_i und Achsenabschnitte b_i einfach gemittelt werden.

Aus historischen Gründen (1801: Entdeckung von Ceres durch Giuseppe Piazzi) hat sich aber die *Methode der kleinsten Quadrate* durchgesetzt: in Abb. 1 sind Punkte in einem System der Variablen x, y gegeben. Die 'Punktwolke' soll durch eine Gerade möglichst gut repräsentiert werden. Gesucht sind nun die Parameter jener Geraden, die für diese Approximation 'am besten' geeignet ist. Wir nennen diese Gerade auch *Ausgleichsgerade*.

lineare Regression

Für die Regression nimmt man an, dass die wahren Werte der gemessenen Größen auf einer Geraden liegen, d.h. zwischen der Variablen x und den wahren Werten der beobachteten Größe y wird ein linearer Zusammenhang angenommen:

$$y = f(x) = a + b \cdot x$$

Mathematisch kann zur Bestimmung der beiden Geradenparameter Steigung a und Ordinatenabschnitt b ein Gleichungssystem aufgestellt werden, wobei jeder Punkt eine Gleichung beisteuert (dabei wird vorausgesetzt, dass mehr Datenpunkte vorliegen, als zur Bestimmung der Unbekannten notwendig sind) $\Rightarrow$ es handelt sich bei unserem Beispiel um ein Gleichungssystem mit zwei Unbekannten und 6 Gleichungen. Ein solches System wird als überbestimmt bezeichnet, wegen der leichten Abweichungen realer Messwerte kann es nicht eindeutig gelöst werden.

lineare Regression

Eine Gerade ist in der Ebene durch zwei Punkte definiert. Gibt es mehr als zwei Punkte, die auf einer Geraden liegen sollen, kann im Allgemeinen keine eindeutige Lösung angeben werden. Es muss also ein Kriterium dafür gefunden werden, welche Gerade 'möglichst gut' an die Punktwolke angepasst ist. Wir fordern dafür, dass die Abweichung der Funktionswerte einer unbekannten Funktion $y = f(x)$ von den gemessenen Werten möglichst klein ist.

Die Abweichung kann unterschiedlich definiert werden - plausibel erscheint eine Definition, in der ein größerer Fehler überproportional mehr wiegt als ein kleiner. Zudem darf das Fehlermaß nicht vorzeichenbehaftet sein (sonst würde ein negativer Fehler einen betragsgleichen positiven Fehler ausgleichen). Das Quadrat der Abweichung zwischen Mess- und Funktionswert erfüllt beide Forderungen.

lineare Regression

Als weitere Vereinfachung kann eines der Merkmale (meist x) als exakt angenommen werden, Abweichungen werden nur für das zweite Merkmal y betrachtet. Dann kann $(f(x) - y)^2$ als geeignetes Maß für die Abweichung der gemessenen Größe y betrachtet werden.

Gesucht wird nun eine Funktion, für die die Summe

$$S = \sum (f(x_i) - y_i)^2$$

der Quadrate der einzelnen Differenzen zwischen Funktions- und Messwerten minimal wird. Die notwendige Bedingung für ein Minimum der Summe ist, dass ihre erste Ableitungverschwindet. Für einen einen linearen Zusammenhang $f(x_i) = y_i = a + b \cdot x_i$, kann die Summe wie folgt umgeschrieben werden:

$$\sum_{i=1}^{n} (f(x_i) - y_i)^2 = \sum_{i=1}^{n} (y_i - (b + a \cdot x_i))^2 \longrightarrow \text{min!}$$

lineare Regression

Zur Erinnerung: die Werte x_i und $\bar{x}$ sind bekannt, die Summe kann als Funktion der Variablen a und b aufgefasst werden. Durch partielles Differenzieren und Nullsetzen der Ableitungen erster Ordnung erhält man ein System von Normalgleichungen. Die gesuchten Regressionskoeffizienten sind die Lösungen

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

und (die Regressionsgerade geht immer durch den Punkt $(\bar{x}, \bar{y})$)

$$\bar{y} = a \cdot \bar{x} + b$$
$$\Rightarrow b = \bar{y} - b\bar{x}$$

Beispiel: lineare Regression

Für die Daten des Herstellers von Dingen

	Preis	Menge				
i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	20	0	5	-5	-25	25
2	16	3	1	-2	-2	1
3	15	7	0	2	0	0
4	16	4	1	-1	-1	1
5	13	6	-2	1	-2	4
6	10	10	-5	5	-25	25
Summe	90	30	0	0	-55	56
Mittelwert	15	5				

Tabelle: lineare Regression am Beispiel von Dingen

Beispiel: lineare Regression

ergeben sich für die beiden gesuchten Regressionsparameter die Werte

$$a = \frac{S_{xy}}{S_{xx}} = \frac{-55}{56} = -0,98$$

und

$$b = \bar{y} - a \cdot \bar{x} = 5 + 0,98 \cdot 15 = 19,73$$

Die jeweilige verkaufte Menge an Produkten y hängt also mit dem Preis x angenähert wie

$$y = a \cdot x + b = -0,98x + 19,73$$

zusammen. Mit einer Preiserhöhung um eine Einheit sinkt der Absatz um etwa eins.

Beispiel: lineare Regression

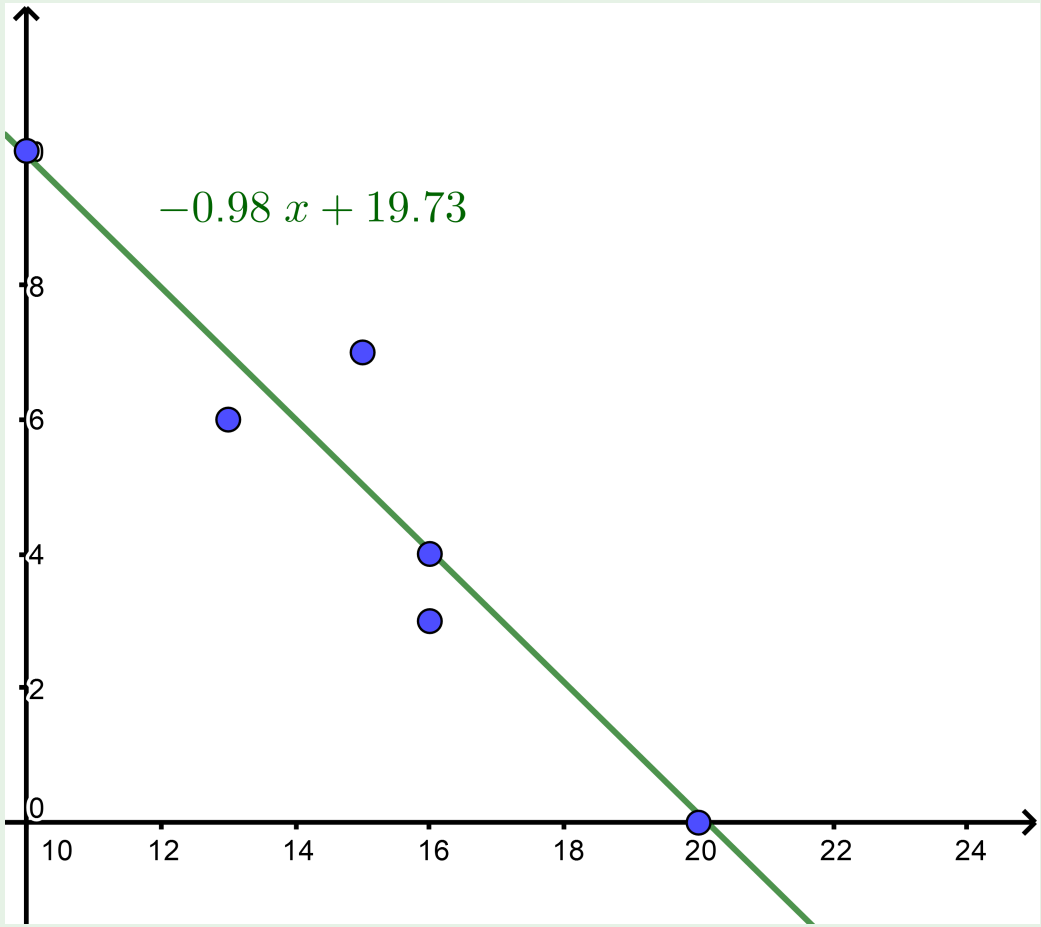


Abbildung: Daten zum Beispiel des Herstellers von Dingen mit Ausgleichsgerade.

Der Korrelationskoeffizient

Der Pearson-Korrelationskoeffizient oder Korrelationswert[1] ist ein dimensionsloses Maß, das den Grad eines linearen Zusammenhangs zwischen zwei kardinalen Merkmalen angibt. Er eignet sich zur Untersuchung der Kreuzkorrelation (Korrelation zeitgleicher Messwerte zweier Merkmale) und der Autokorrelation (Korrelation zeitlich verschiedener Messwerte eines einzelnen Merkmals).

Der Korrelationswert kann Werte zwischen -1 und $+1$ annehmen, bei einem Wert von $+1$ (bzw. -1) besteht ein vollständig positiver (negativer) linearer Zusammenhang zwischen den Merkmalen. Weist der Korrelationskoeffizient den Wert 0 auf, sind die betrachteten Merkmale nicht linear voneinander abhängig (sie können aber in nicht-linearer Weise voneinander abhängen - der Korrelationskoeffizient ist kein geeignetes Maß für die allgemeine stochastische Abhängigkeit von Merkmalen).

[1] nach Bravais und Pearson

Der Korrelationskoeffizient

Der Pearson-Korrelationskoeffizient ist für zwei Zufallsvariablen X, Y mit positiver (von Null verschiedener) Standardabweichung $\varsigma(X), \varsigma(Y)$ und der Kovarianz $\mathrm{COV}(X, Y)$ definiert durch

$$\rho_{xy} = \frac{\mathrm{COV}(X, Y)}{\varsigma(X)\varsigma(Y)}$$

Wir kennen bereits die empirische Standardabweichung und die empirische Kovarianz für eine Messreihe gepaarter Merkmalsausprägungen $(x_1; y_1), (x_2; y_2), \ldots (x_n; y_n)$ - das erlaubt uns natürlich die Berechnung eines empirischen (geschätzten) Pearson-Korrelationskoeffizienten.

Der Korrelationskoeffizient

Definition: empirischer Korrelationskoeffizient

Bezeichnen $\bar{x}$ und $\bar{y}$ die arithmetischen Mittel der Merkmale X und Y in einer Messreihe gepaarter Merkmalsausprägungen $(x_1; y_1), (x_2; y_2), \ldots (x_n; y_n)$, so bezeichnet man

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

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

als den Korrelationskoeffizienten der Messreihe.

Der Korrelationskoeffizient

Interpretation des Korrelationskoeffizienten:

Aus der Definition ist der Wert des Koeffizienten für perfekte Korrelation sofort ersichtlich. Der perfekte lineare Zusammenhang ist sicher gegeben, wenn man eine Variable als abhängig von sich selbst betrachtet, also:

$$\begin{aligned} r_{xx} :&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1. \end{aligned}$$

Für Wertepaare, die auf einer perfekten Geraden liegen (perfekte lineare Abhängigkeit), ist der Korrelationskoeffizient offensichtlich 1, falls die Gerade ansteigt (sonst -1).

Der Korrelationskoeffizient

Je mehr die Gestalt der durch die Wertepaare gebildeten Punktwolke von einer Geraden abweicht, desto kleiner wird der Wert r_{xy}, bei $r_{xy} = 0$ kann der Zusammenhang zwischen den Merkmalen nicht mehr durch eine eindeutig steigende oder fallende Gerade dargestellt werden. Dies bedeutet, dass die Werte nicht mehr verlässlich an eine Gerade angepasst werden können (oder natürlich, dass eines der beiden Merkmale konstant ist - auch dann ist kein Zusammenhang gegeben).

Auch der Korrelationskoeffizient kann als eine Prozentzahl gelesen werden, die die Qualität des linearen Zusammenhangs angibt.

Der Korrelationskoeffizient

Beispiel:

Für unser Beispiel des Herstellers von Dingen sind die Werte zur Berechnung der empirischen Kovarianz und der Standardabweichung s_x des Merkmals X bereits bekannt.
Wir benötigen lediglich die Werte für die Standardabweichung s_y:

	Preis	Menge				
x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
20	0	5	-5	-25	25	25
16	3	1	-2	-2	1	4
15	7	0	2	0	0	4
16	4	1	-1	-1	1	1
13	6	-2	1	-2	4	1
10	10	-5	5	-25	25	25
90	30	0	0	-55	56	60
Mittelwert	15	5				

Tabelle: Bestimmung des Pearson-Korrelationskoeffizienten

Der Korrelationskoeffizient

Beispiel:

Der Korrelationskoeffizient nach Pearson und Bravais kann jetzt einfach berechnet werden.

Für das Beispiel ist

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{-55}{\sqrt{56} \cdot \sqrt{60}} = -0,95$$

Der Wert nahe -1 legt einen relativ starken linearen Zusammenhang nahe, das negative Vorzeichen bedeutet, dass die Gerade abfällt.