

Regression einer Proportionalität

12 In den bisherigen Ausführungen wurde die Statistik einer **einzigsten Messgröße** betrachtet. In den meisten Untersuchungen will man aber Zusammenhänge zwischen **zwei oder mehreren Größen** erforschen. Entweder sucht man ein Gesetz (eine Formel), welches ja eine mathematische Verknüpfung mehrerer Größen ist, oder man will überhaupt erst herausfinden, ob zwischen zwei Größen ein Zusammenhang besteht.

13 Wir beginnen mit dem einfachsten Fall einer Verknüpfung und fragen, ob die Werte zweier Messreihen durch eine **Proportionalität** verknüpft werden können. Anhand des im Kasten beschriebenen Beispiels sollen drei wichtige Begriffe beschrieben werden: Das **Punktdiagramm (Streudiagramm, x-y-Diagramm)**, die **Regressionsgerade (Ausgleichsgerade)** und der **Korrelationskoeffizient**.

Einen ersten Eindruck über die Art des Zusammenhangs erhalten wir, indem wir die beiden Messwerte in einem Streudiagramm darstellen. Im **Beispiel** sehen wir: Die Werte liegen nicht exakt auf einer Geraden, aber doch ziemlich eindeutig in der Nähe einer Geraden.

14 Wir überlegen uns nun, wie wir jene Gerade berechnen können, die „am besten“ in die Menge der Messpunkte passt. Das Kriterium für die „**beste Gerade**“ (engl.: **best fit**) soll sein: Die Summe der quadrierten Abstände zwischen den Messwerten und der Geraden ist kleiner als bei jeder anderen Gerade.

Die Gerade, deren Steigung a wir noch nicht kennen, verläuft durch den **Koordinatenursprung**, da wir beim zu beschreibenden Zusammenhang von einer Proportionalität ausgehen.

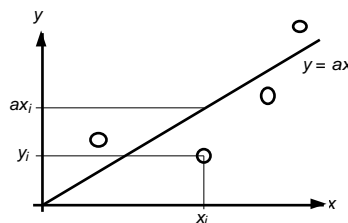
Für die Gerade gilt also:

$$(8) \quad y = a \cdot x$$

Nun bilden wir den Mittelwert aller quadrierten „vertikalen“ Entfernungen von der Geraden:

$$(9) \quad \frac{1}{n} \sum_{i=1}^n (a \cdot x_i - y_i)^2$$

Für die „beste“ Gerade (also für die passendste Steigung a) muss die Summe (9) minimal sein. Wir lösen also eine Extremalwertaufgabe bezüglich der Variablen a : Dazu leiten wir die Summe (9) nach der Variablen a ab und verlangen



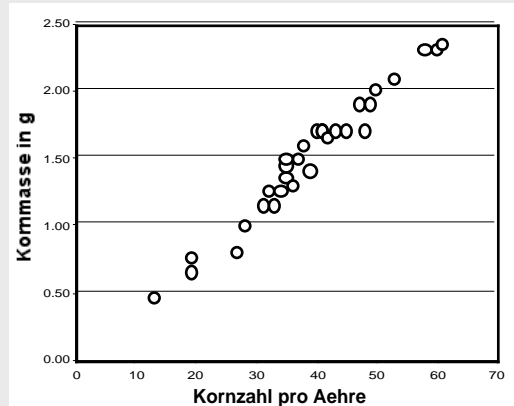
Beispiel: Die Vermutung ist naheliegend, dass eine Weizenähre mit grösserer Kornzahl auch eine grössere Gesamtkornmasse hat. Hätten wir es mit einem funktionalen Zusammenhang dieser beiden Größen zu tun, so müsste jedes Korn in der Ähre auch eine eindeutige Veränderung des Gesamtgewichts der Ähre bewirken. Wäre dieser Zusammenhang einmal mathematisch dargestellt, so könnte der Bauer auf ein Feld gehen, die Körner einer Ähre auszählen und dann (wenn er weiss, wie viele Ähren auf einer bestimmten Fläche wachsen) den Ertrag des gesamten Feldes genau berechnen.

Aber könnte es nicht auch sein, dass mit steigender Kornzahl die Körner der Ähre kleiner werden und somit durchaus keine eindeutige Steigerung der Gesamtkornmasse eintritt?

Messwerte:

Kornzahl	Kornmasse in g
13	0.45
19	0.65
19	0.75
27	0.80
28	1.00
31	1.15
32	1.25
33	1.15
34	1.25
35	1.35
35	1.45
35	1.50
36	1.30
37	1.50
38	1.60

Kornzahl	Kornmasse in g
39	1.40
40	1.70
41	1.70
41	1.70
42	1.65
43	1.70
45	1.70
47	1.90
48	1.70
49	1.90
50	2.00
53	2.10
58	2.30
60	2.30
61	2.35



$$(10) \quad \frac{1}{n} \sum_{i=1}^n 2 \cdot x_i \cdot (a \cdot x_i - y_i) = 0 \quad .$$

Ausmultiplizieren und Division durch $2/n$ ergibt:


$$(11) \quad a \cdot \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i = 0$$

somit erhalten wir für die Steigung:

$$(12) \quad a = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

- 15 ? **Beispiel:** Sie können nun die Steigung aus den Tabellenwerten errechnen. Dividieren Sie die Summe aller Produkte der Messwerte durch die Summe aller Quadrate der Kornzahlen. Vielleicht ist Ihnen die Tabelle dabei nützlich.

Zeichnen Sie die Regressionsgerade korrekt in das Diagramm auf Seite 5 ein.

- 16  Nun benötigen wir noch eine Grösse, mit der wir die **Qualität der Anpassung** der Geraden beschreiben können. Diese Grösse sollte uns sagen, wie stark die Messwerte um die Gerade streuen. Die Aussage dieser Grösse ist also vergleichbar mit der Varianz (2) bei einer einzelnen Messreihe. Wir nennen diese Grösse **Korrelationskoeffizient** und im Falle der Proportionalität wird sie wie folgt definiert.

$$(13) \quad r_{xy} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$$

Wenn die Messpunkte genau auf der Geraden liegen, so wird $r_{xy} = 1$. Je mehr die Messwerte jedoch von der Geraden abweichen, desto mehr nähert sich r_{xy} dem Wert 0.

Die erste Behauptung kann man leicht überprüfen, indem man in (13) berücksichtigt, dass jeder Messwert y_i gleich $a x_i$ ist, wenn der Zusammenhang streng proportional ist.

- 17 ? **Beispiel:** Berechnen Sie nun den Korrelationskoeffizienten für unser Ähren-Beispiel.

Zudem können Sie auch (zur Repetition der Seiten 1 bis 4) die mittlere Kornzahl und die mittlere Masse einer Ähre und deren Standardabweichungen bestimmen

Anzahl Messungen: $n = 30$

x	y	xy	x ²	y ²
Kornzahl	Masse in g			
13	0.45			
19	0.65			
19	0.75			
27	0.80			
28	1.00			
31	1.15			
32	1.25			
33	1.15			
34	1.25			
35	1.35			
35	1.45			
35	1.50			
36	1.30			
37	1.50			
38	1.60			
39	1.40			
40	1.70			
41	1.70			
41	1.70			
42	1.65			
43	1.70			
45	1.70			
47	1.90			
48	1.70			
49	1.90			
50	2.00			
53	2.10			
58	2.30			
60	2.30			
61	2.35			
$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$

- Steigung der Regressionsgeraden _____
- Korrelationskoeffizient _____
- mittlere Kornzahl _____
- Standardabweichung der Kornzahl _____
- mittlere Ährenmasse _____
- Standardabweichung der Ährenmasse _____

Regressionsgerade und Korrelationskoeffizient

18 Im Allgemeinen verläuft eine **Regressionsgerade** nicht durch den Koordinatenursprung. Man legt sie stattdessen durch den „**Schwerpunkt des Punkthaufens**“, das ist der Punkt mit den Koordinaten \bar{x} und \bar{y} . Die Geradengleichung lautet:

$$(14) \quad y = ax + b$$

Und weil der Schwerpunkt auf der Geraden liegen soll, gilt auch:

$$(15) \quad \bar{y} = a\bar{x} + b$$

Der Koeffizient a berechnet sich nun aus

$$(16) \quad a = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2},$$

was hier ohne Beweis aufgeschrieben wird. n bedeutet darin die Zahl der Messwerte. Kennt man die Steigung a , so lässt sich mit Hilfe der Mittelwerte aus (15) auch der y -Achsenabschnitt b bestimmen.

19 Auch für den **Korrelationskoeffizienten** geben wir hier noch die allgemeine Form an:

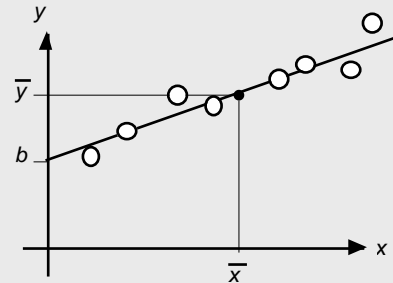
$$(17) \quad r_{xy} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - n\bar{x}^2\right) \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)}}$$

Der Korrelationskoeffizient nimmt einen Wert zwischen -1 und $+1$ an. Ist der Wert $+1$, so liegen die Messwerte exakt auf einer steigenden Geraden. Ist der Wert -1 , so liegen die Messwerte exakt auf einer fallenden Geraden. Je stärker die Werte um die Regressionsgerade streuen, desto näher ist der Korrelationskoeffizient bei Null.

Sowohl in (16) als auch in (17) erkennt man, dass die Formeln zur Proportionalität reproduziert werden, wenn der Schwerpunkt (\bar{x}, \bar{y}) durch den Koordinatenursprung $(0,0)$ ersetzt wird.

20 In Excel (und anderen **Tabellenkalkulationsprogrammen**) können Sie die Regressionsgerade (sie heisst dort **Trendlinie** (zu finden im Menu Diagramm) automatisch zeichnen und mit der Geradengleichung ausdrucken lassen. Vielleicht hat auch ihr Taschenrechner diese Möglichkeit einprogrammiert (z.B. Taste L.R.), falls es möglich ist, Wertepaare von Messreihen einzugeben.

**Zusammenfassung
Regressionsgerade
(Ausgleichsgerade)**



$$y = ax + b, \text{ mit}$$

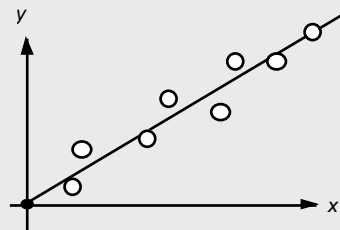
$$a = \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

$$b = \bar{y} - a\bar{x}$$

Korrelationskoeffizient

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

**Spezialfall:
Regression einer Proportionalität**




$$y = a \cdot x$$

$$a = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

$$r_{xy} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$$

Beispiel

- 21  Zwischen dem **BSP** (Bruttonsozialprodukt) pro Person und der **Lebenserwartung** sowie der **Zahl der Kinder** einer Frau besteht ein gewisser Zusammenhang. Untersuchen Sie diesen an Hand der kleinen Auswahl von Ländern, indem Sie für beide Zusammenhänge ein Streudiagramm herstellen, die Regressionsgerade und den Korrelationskoeffizienten berechnen.

Übersetzen Sie die Steigung der Regressionsgeraden in eine Aussage in Worten.

Ist die Kinderzahl oder die Lebenserwartung stärker an das BSP gebunden?

Weitere Kritik, weiterer Kommentar?

Die Daten stammen aus den Jahren 1993/94 und wurden beim Bundesamt für Statistik abgeholt.

	x	y_1	y_2	x^2	xy_1	xy_2	y_1^2	y_2^2
	BSP pro Einwohner in US\$	Lebenserwartung in Jahren	Kinder pro Frau					
CH	37'180	78	1.6					
D	25'580	76	1.3					
USA	25'860	76	2.1					
Japan	34'630	79	1.5					
Brasilien	3'370	67	2.8					
Kenia	340	59	6.5					
Indien	330	59	3.8					
Israel	14'410	77	2.8					
Mexiko	4'010	71	3.1					
Philippinen	960	67	3.9					
Vietnam	190	65	3.8					

