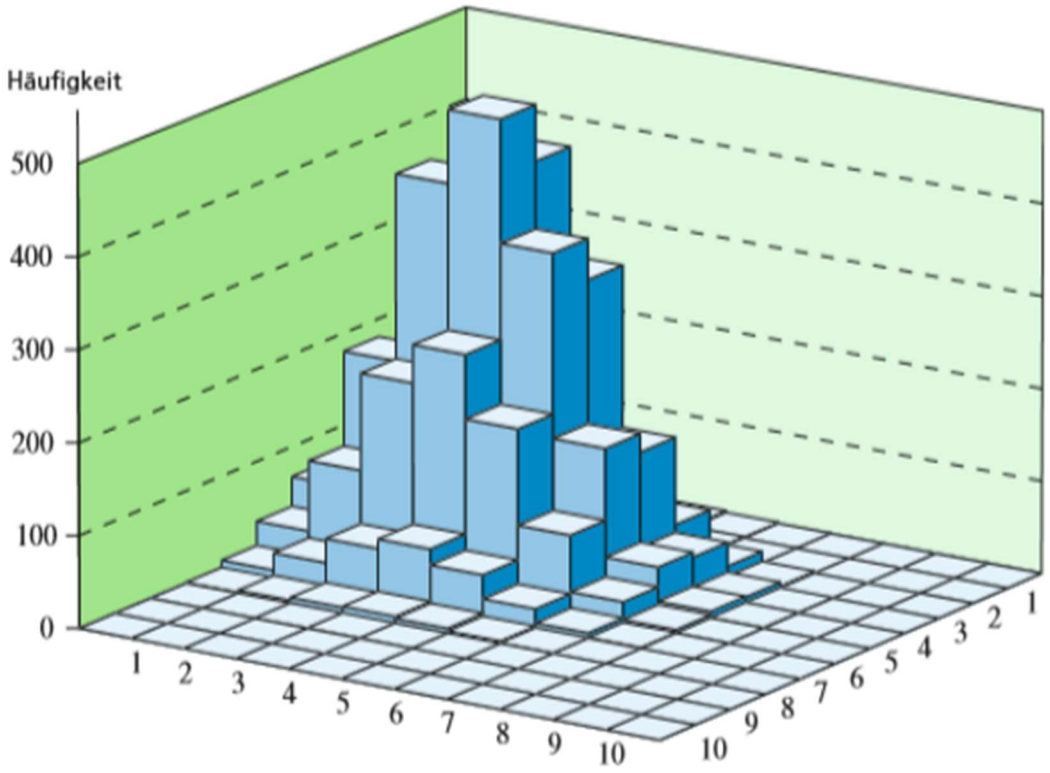
1.2 Bivariate Zusammenhänge



Wir betrachten zwei Merkmale $X: \Omega \to \mathbb{R}$ und $Y: \Omega \to \mathbb{R}$ auf derselben Grundgesamtheit Ω, und fassen zusammen $(X, Y): \Omega \to \mathbb{R} \times \mathbb{R}$ mit $(X, Y)(\omega) := \big(X(\omega), Y(\omega)\big) = (x, y)$
Eigentlich ist dies nichts neues, denn auch hier haben wir nur eine Abbildung $Z: \Omega \to \mathrm{M}$ speziell mit $M = \mathbb{R} \times \mathbb{R}$ definiert, also wieder ein Merkmal für die Grundgesamtheit Ω. Wir sprechen dann hier von einem **bivariaten** (bivariablen, zweidimensionalen) **Merkmal**. Auch hier können wir die Merkmalswerte $\big(X(\omega_i), Y(\omega_i)\big) = (x_i, y_i)$ nach ihrer Häufigkeit gruppieren und dies auch darstellen etwa mit einem Balkendiagramm wie oben (aus Arens, Kap. 36). Eine solche Häufigkeitsverteilung ist also dann eine **bivariate** (zweidimensionale) **Verteilung**. Man spricht auch von der gemeinsamen Verteilung der Verteilungen von X und Y.

Eine andere nützliche Darstellung ist der zweidimensionale Plot: Hier wird für jedes Objekt $\omega \in \Omega$ durch einen Punkt in der Ebene dargestellt mit den Koordinaten $\big(X(\omega), Y(\omega)\big) = (x, y) \in \mathbb{R} \times \mathbb{R}$. Dies ist vor allem dann hilfreich, wenn $X(\omega)$ und $Y(\omega)$ Messwerte auf einer kontinuierlichen Skala sind. Es entsteht dann eine Punktewolke wie in der folgenden Grafik (Arens, Kap. 36).

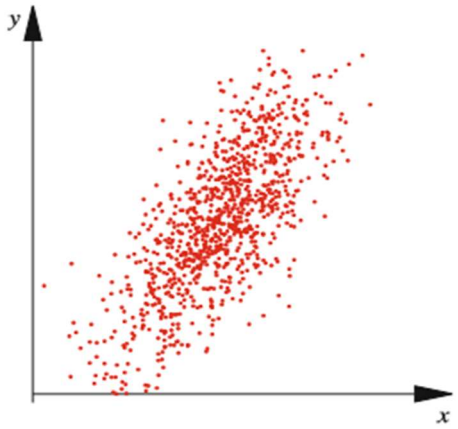


Abb. 36.35 Punktwolke mit Länge und Breite von Hühnereiern

Bei diesem Bild fällt die Tendenz auf, dass mit wachsendem x auch y wächst, also längere Eier „im Schnitt“ auch breiter sind. Als Indikatoren eines Zusammenhangs zweier Merkmale definiert man dazu die Größen Kovarianz und Korrelation:

Definition: Für $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$, $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ ist

$cov(\boldsymbol{x}, \boldsymbol{y}) := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ die (empirische) **Kovarianz** und

$r(\boldsymbol{x}, \boldsymbol{y}) := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$ der (empirische) **Korrelationskoeffizient** von $\boldsymbol{x}$ und $\boldsymbol{y}$.

Bemerkung: Gern wird die Kovarianz auch mit s_{xy} bezeichnet. Schreibt man dann die Standardabweichung als $s_x := \sqrt{var(\boldsymbol{x})}$, so kann man den Korrelationskoeffizienten auch definieren als

$r(\boldsymbol{x}, \boldsymbol{y}) = r := \frac{s_{xy}}{s_x s_y}$,

weil sich die Faktoren $\frac{1}{n}$ alle wegkürzen.

Eine weitere Darstellung ist (wie man leicht nachrechnet) $r(\boldsymbol{x}, \boldsymbol{y}) = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum {x_i}^2 - (\sum x_i)^2][n\sum {y_i}^2 - (\sum y_i)^2]}}$

Oder auch $r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\left[\frac{1}{n}\sum {x_i}^2 - \bar{x}^2\right]\left[\frac{1}{n}\sum {y_i}^2 - \bar{y}^2\right]}}$

Das folgende **Beispiel 1** ist wieder entnommen aus Arens, Kap. 36

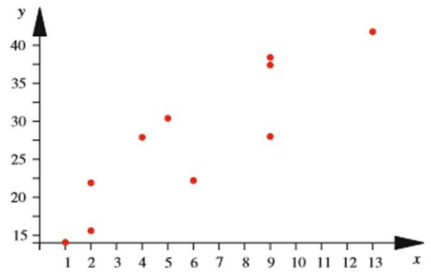
Beispiel Bei 10 Objekten seien jeweils die Länge X (in cm) und das Gewicht Y (in kg) gemessen worden. Die Messwerte und die Berechnung der Kovarianz zeigt die folgende Tabelle.

y_i	x_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
14.07	1.0	−5.0	−13.70	68.49
15.60	2.0	−4.0	−12.17	48.68
21.92	2.0	−4.0	−5.85	23.39
27.90	4.0	−2.0	0.13	−0.25
30.40	5.0	−1.0	2.62	−2.62
22.25	6.0	0.0	−5.52	0.00
37.43	9.0	3.0	9.66	28.97
38.40	9.0	3.0	10.63	31.90
27.95	9.0	3.0	0.18	0.53
41.79	13.0	7.0	14.02	98.13
277.71	60.0	0.0	0.0	297.21

Aus der Tabelle lesen wir ab: $\bar{x} = \frac{60.0}{10} = 6$ und $\bar{y} = \frac{277.71}{10} = 27.771$. Es ist

$$\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{10} \cdot 297.21 = 29.72 .$$

Abbildung 36.39 zeigt den Plot von y gegen x. ◀



Angenommen, in diesem Beispiel wären die Merkmale X in Millimeter statt Zentimeter und Y in Gramm statt Kilogramm gemessen worden. Damit wäre jeder x-Wert 10-mal und jeder y-Wert 1000-mal größer. Die Kovarianz der neuen Werte wäre nun 297 200.

Für den Korrelationskoeffizienten hätten wir in beiden Fällen bekommen
$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{29{,}72}{\sqrt{13{,}8 \cdot 81{,}02784}} = \frac{297200}{\sqrt{1380 \cdot 81027840}} \approx 0{,}888$
Allgemein ist der Korrelationskoeffizient, im Gegensatz zur Kovarianz, **skaleninvariant**, wie wir gleich sehen werden.

Wir fassen kurz einige Rechenregeln zusammen. Dazu definieren wir, wie auch in der Linearen Algebra üblich, für $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$, $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$, $\boldsymbol{x} + \boldsymbol{y} \coloneqq (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$

Rechenregeln:

$cov(\boldsymbol{x}, \boldsymbol{y}) = cov(\boldsymbol{y}, \boldsymbol{x})$
$cov(\boldsymbol{x}, \boldsymbol{x}) = var(\boldsymbol{x})$
$cov(\boldsymbol{x}, \boldsymbol{y} + \boldsymbol{z}) = cov(\boldsymbol{x}, \boldsymbol{y}) + cov(\boldsymbol{x}, \boldsymbol{z})$
$var(\boldsymbol{x} + \boldsymbol{y}) = var(\boldsymbol{x}) + var(\boldsymbol{y}) + 2cov(\boldsymbol{x}, \boldsymbol{y})$
Ersetzen wir für $\alpha, \beta, \gamma, \delta \epsilon \mathbb{R}$
$\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ durch $\boldsymbol{x}' \coloneqq (\alpha x_1 + \gamma, \alpha x_2 + \gamma, \dots, \alpha x_n + \gamma)$ und
$\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ durch $\boldsymbol{y}' \coloneqq (\beta y_1 + \delta, \beta x_2 + \delta, \dots, \beta x_n + \delta)$, so gilt
$cov(\boldsymbol{x}', \boldsymbol{y}') = \alpha\beta \cdot cov(\boldsymbol{x}, \boldsymbol{y})$
und $r(\boldsymbol{x}', \boldsymbol{y}') = r(\boldsymbol{x}, \boldsymbol{y})$.

An der letzten Gleichung sieht man, dass der Korrelationskoeffizient invariant ist gegenüber linearen Transformationen, also insbesondere skaleninvariant, wie wir schon im vorigen Beispiel gesehen haben.

Regressionsgerade

Um die Frage zu klären, inwiefern Kovarianz und Korrelation einen Zusammenhang zweier Merkmale beschreiben können, stellen wir uns die folgende Aufgabe:

Gegeben ist eine „Punktewolke“ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ im $\mathbb{R}^2$
Gesucht sind zwei reelle Zahlen a und b so, dass die lineare Funktion $y = f(x) := ax + b$ im folgenden Sinne „möglichst gut“ zu der gegebenen Punktewolke passt: Dazu soll (nach einer Idee von Gauß) die Summe der Abstandsquadrate $[f(x_i) - y_i]^2$ minimal werden, also

$$q(a, b) := \sum [f(x_i) - y_i]^2 \to \min$$

Im „Idealfall“ geht der Graph der Funktion $y = f(x)$ genau durch die Punkte (x_i, y_i). Deshalb spricht man bei den Ausdrücken $[f(x_i) - y_i]^2$ auch von **Fehlerquadraten**.

Satz: Das obige Problem hat genau eine Lösung:

$a = \frac{s_{xy}}{s_x^2} = \frac{cov(\boldsymbol{x},\boldsymbol{y})}{var(\boldsymbol{x})} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
$b = \bar{y} - a\bar{x}$

Beispiel 2: Für die 5 Punkte $(x_i, y_i) = (1{,}5;1),(2{,}5;1{,}2),(3{,}5;2),(4{,}5;2{,}2),(5{,}5;3)$ ist
$\bar{x} = 3{,}5$; $\bar{y} = 1{,}88$; $cov(\boldsymbol{x}, \boldsymbol{y}) = 1$; $var(\boldsymbol{x}) = 2$; $a = 0{,}5$; $b = 1{,}88 - 0{,}5 \cdot 3{,}5 = 0{,}13$
Skizze → Tafel
In **Beispiel 1** war $\bar{x} = 6$; $\bar{y} = 27{,}771$; $cov(\boldsymbol{x}, \boldsymbol{y}) \approx 29{,}72$; $var(\boldsymbol{x}) \approx 13{,}8$; $a \approx 2{,}15$
$b \approx 27{,}771 - 2{,}15 \cdot 6 \approx 14{,}87$

Definition: Der Graph der Funktion $y = f(x) = ax + b$ heißt **Ausgleichsgerade** oder **Regressionsgerade**. Deren Steigung a heißt auch **empirischer Regressionskoeffizient** zu den gegebenen Daten (x_i, y_i).

Bemerkungen:

a) Den Beweis des obigen Satzes führt man am elegantesten, indem man die Summe der Fehlerquadrate (bei fest vorgegebenen Daten (x_i, y_i)) als stetig differenzierbare Funktion $q(a, b)$ der beiden Variablen a und b auffasst. Nullsetzen der beiden partiellen Ableitungen nach a und nach b liefert zwei lineare Gleichungen in a und b, deren Lösung dann gerade durch die obigen Formeln beschrieben wird. Ein elementarerer, aber mühsamerer Weg ginge über quadratische Ergänzung.

b) Bemerkenswert, aber auch intuitiv einleuchtend ist die zweite Formel, welche besagt, dass die Ausgleichsgerade immer genau durch den geometrischen **Schwerpunkt** $(\bar{x}, \bar{y})$ der Punkte (x_i, y_i) in der Ebene $\mathbb{R}^2$ geht, also $\bar{y} = a\bar{x} + b$.
(Hat man n Punkte im $\mathbb{R}^m$ mit den Ortsvektoren $\vec{p}_1, \vec{p}_2, \dots, \vec{p}_n$, so ist $\vec{s} = \frac{1}{n}\sum \vec{p}_i$ der Ortsvektor des Schwerpunktes dieser Punkte.)
Das sieht man unmittelbar: Für jedes fest gewählte a kann $q(a, b)$ nur dann minimal werden, wenn $b = \bar{y} - a\bar{x}$ ist. Denn $q(a, b)$ hängt dann nur noch von b ab, also
$q(b) = \sum[f(x_i) - y_i]^2 = \sum[ax_i + b - y_i]^2$, und aus $q'(b) = 2\sum[ax_i + b - y_i] = 0$ folgt unmittelbar $a\sum x_i + nb - \sum y_i = 0$. Division der Gleichung durch n liefert dann $a\bar{x} + b - \bar{y} = 0$, q.e.d. Im Prinzip haben wir hier bereits mit partiellem Ableiten argumentiert. Setzt man dieses Ergebnis dann in $q(a, b)$ ein, dann hängt $q(a, b)$ nur noch von a ab, und mit einer ähnlichen Rechnung bekommt man schließlich auch die obige Formel für a. Dies soll hier nicht mehr ausgeführt werden. Man kanns z.B. nachlesen in https://www.math.uni-bielefeld.de/~sek/funktion/leit03.pdf

⚠ Dort, wie in den meisten Lehrbüchern und Formelsammlungen, sind übrigens a und b in ihrer Bedeutung vertauscht, also $y = a + bx$, und nicht, wie hier, $y = ax + b$.

c) Das Auffinden einer Ausgleichsgerade spielt nicht nur in der Statistik eine Rolle. Zum Beispiel könnte aus der Theorie bekannt sein, dass zwischen zwei physikalischen Größen ein linearer Zusammenhang $y = ax + b$ besteht. Man bestimmt dann a und b mittels n Messungen. Weil diese fehlerbehaftet sind, liegen die (x_i, y_i) nicht, wie es eigentlich sein müsste, auf einer Geraden (sonst würden ja zwei Wertepaare genügen). Die Ausgleichsgerade soll also dann die Fehler optimal „**ausgleichen**".

Zur Begriffsklärung „**Regression**" aus H. Weber: Einführung in die Wahrscheinlichkeitsrechnung und Statistik für Ingenieure, Stuttgart, 1992

In der Mathematik heißt eine derartige Gerade Ausgleichsgerade. Den Begriff Regression führte Galton 1889 in Zusammenhang mit seinen Studien zur Vererbungslehre ein. Galton behauptete, daß Eltern, die hinsichtlich eines Merkmals vom Durchschnitt abweichen, Kinder haben, die bezüglich dieses Merkmals zwar in der gleichen Richtung vom Mittel abweichen, jedoch in geringerem Maße als ihre Eltern. Galton nannte diesen Effekt Regression. K. Pearson, ein Freund Galtons, untersuchte 1903 an 1078 Vater-Sohn-Paaren den Zusammenhang zwischen den Merkmalen X = Körpergröße des Vaters und Y = Körpergröße des Sohnes und erhielt als Gerade, die sich dem von ihm erhaltenen Streuungsdiagramm am besten anpaßte

$$y = 85{,}674 + 0{,}516.x \quad [\text{cm}],$$

die er Regressionsgerade nannte. Dieses spezielle Beispiel bestimmte die Namensgebung für ein Teilgebiet der Statistik.

d) Man kann zeigen, dass für den Korrelationskoeffizienten $r = \frac{s_{xy}}{s_x s_y} = a \cdot \frac{s_x}{s_y}$
(dabei haben r, der Regressionskoeffizient, also die Steigung a und die Kovarianz s_{xy} das gleiche Vorzeichen) immer gilt

$$-1 \leq r \leq 1$$

Das ist im Wesentlichen die Cauchy-Schwarzsche Ungleichung aus der Linearen Algebra. Zwischen dem Korrelationskoeffizienten r und der Regressionsgeraden besteht nun folgende Beziehung, welche unsere ursprüngliche Frage „Inwiefern können Kovarianz und Korrelation einen Zusammenhang zweier Merkmale beschreiben?" beantwortet:

1) $r = -1$: Alle Punkte der Punktewolke (auch Streuungsdiagramm genannt) liegen auf einer Geraden mit negativer Steigung (vollständige negative Korrelation).

2) $r = 1$: Alle Punkte der Punktewolke des Streuungsdiagramms liegen auf einer Geraden mit positiver Steigung (vollständige positive Korrelation).

3) $r = 0$: Es besteht keine Korrelation zwischen den Merkmalen X und Y (Genaugenommen heißt das nur: Es besteht kein **Linearer** Zusammenhang).

4) $r < 0$: Es besteht eine negative Korrelation, d.h. größeren Werten des Merkmals X entsprechen im Mittel kleinere Werte des Merkmals Y, d.h. die Regressionsgerade hat eine negative Steigung.

5) $r > 0$: Es besteht eine positive Korrelation, d.h. größeren Werten des Merkmals X entsprechen im Mittel auch größere Werte des Merkmals Y, d.h. die Regressionsgerade hat eine positive Steigung.

6) Je näher $|r|$ an 1 ist, desto „mehr" besteht ein linearer Zusammenhang. Nicht immer lässt sich aber daraus auf einen kausalen Zusammenhang oder eine Beeinflussung zwischen den Merkmalen schließen. Aufgabe der beurteilenden (induktiven) Statistik ist es auch, **Scheinkorrelationen** aufzudecken.