KAPITEL 3

Zweidimensionale Verteilungen

Hängt die Dauer der Arbeitslosigkeit vom Ausbildungsstand ab, oder vom Alter und vom Geschlecht? Beeinflußt die Wachstumsrate der Geldmenge die Inflationsrate? Um wieviel verringert sich die Nachfrage, wenn Volkswagen seine Preise um 5% erhöht? Solche und ähnliche Fragestellungen erfordern die Untersuchung von Zusammenhängen und Abhängigkeiten zwischen *zwei* oder *mehreren Merkmalen*, die *gemeinsam* erhoben werden müssen. In diesem Kapitel wird ausgeführt, wie zweidimensionales Datenmaterial aufbereitet und dargestellt werden kann. Vor allem aber werden Verfahren und Maßzahlen vorgestellt, mit denen die Zusammenhänge und Abhängigkeiten aufgedeckt und gemessen werden können.

3.1 Streudiagramm und gemeinsame Verteilung

Jede statistische Einheit ω_i $(i = 1, \cdots, n)$ einer Grundgesamtheit Ω kann Träger **einer Vielzahl** von Merkmalen sein. Die univariate Statistik beachtet davon nur ein Merkmal bzw. nur eine Variable, die multivariate Statistik beobachtet von jedem Merkmalsträger ω_i mehrere Variablen

$$X_1(\omega_i),\ X_2(\omega_i),\ \cdots,\ X_m(\omega_i) \tag{3-1}$$

und analysiert die Beziehungen zwischen den Variablen. Der einfachste Fall einer mehrdimensionalen Statistik ist die zweidimensionale. Bei ihr sind zwei Variablen

$$X(\omega_i) \qquad \text{und} \qquad Y(\omega_i)$$

von Interesse. Das Ergebnis der Erhebung sind ***Wertepaare*** (x_i, y_i). Im ***Streudiagramm*** werden Wertepaare

$$\begin{aligned}
(x_1, y_1) &=: P_1 \\
(x_2, y_2) &=: P_2 \\
(x_3, y_3) &=: P_3 \\
&\ \ \vdots \\
(x_n, y_n) &=: P_n
\end{aligned}$$

als Koordinaten von Punkten P_i angesehen und in ein Koordinatensystem eingezeichnet:

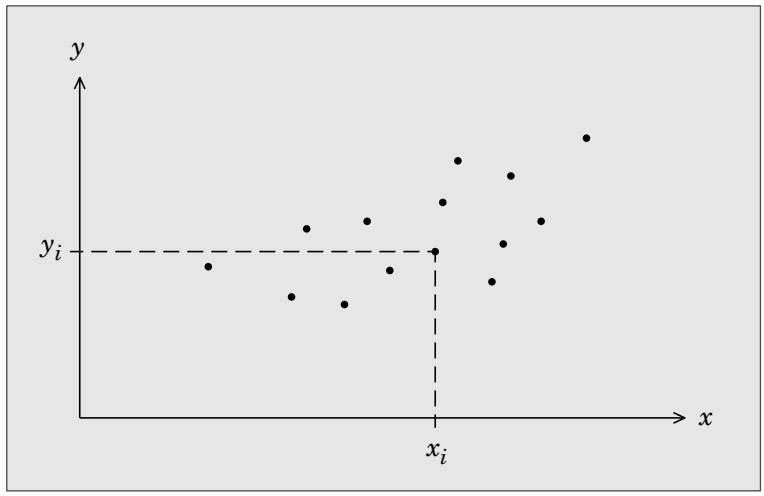

BILD 3.1 Punkte im Streudiagramm

Die ***Kontingenztabelle*** oder Korrelationstabelle

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_l	insge-samt
x_1	n_{11}	n_{12}		n_{1j}		n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2l}	$n_{2\bullet}$
$\vdots$				$\vdots$			$\vdots$
x_i	n_{i1}	n_{i2}		n_{ij}		n_{il}	$n_{i\bullet}$
$\vdots$				$\vdots$			$\vdots$
x_k	n_{k1}	n_{k2}		n_{kj}		n_{kl}	$n_{k\bullet}$
insge-samt	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet l}$	n

stellt die ***gemeinsame Verteilung*** der statistischen Variablen X und Y übersichtlich dar.[1] Dabei gehen wir davon aus, daß die Merkmale jeweils nur k respektive m Ausprägungen

[1] Bei der Behandlung der eindimensionalen statistischen Variablen wurde i als Laufindex und j als Summationsindex verwendet, und es konnte sorgfältig zwischen beiden unterschieden werden. In der Kontingenztabelle der zweidimensionalen statistischen Variablen (X, Y) bezeichnet i gleichzeitig den Summationsindex und Laufindex von X,

annehmen oder annehmen können. Sehr oft werden aber auch bei der Anfertigung von Kontingenztabellen **Größenklassen** gebildet. In der Tabelle bedeutet

$$n_{ij} = \text{absH}(X = x_i \cap Y = y_j), \tag{3-2}$$

für $i = 1, \cdots, k$ und $j = 1, \cdots, l$, die absolute Häufigkeit, mit der die Wertekombination (x_i, y_j), und

$$n_{i\bullet} = \sum_{j=1}^{l} n_{ij} \qquad \text{bzw.} \qquad n_{\bullet j} = \sum_{i=1}^{k} n_{ij} \tag{3-3}$$

die absolute Häufigkeit, mit der der Wert x_i bzw. y_j beobachtet wurde. Es gilt natürlich, daß die Summe der Zeilensummen gleich der Summe der Spaltensummen ist:

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = \sum_{i=1}^{k} n_{i\bullet} = n = \sum_{j=1}^{l} n_{\bullet j} = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij} .$$

Natürlich können in Kontingenztabellen auch die relativen Häufigkeiten oder Prozentwerte angegeben sein, was meist anschaulicher ist. Für die relativen Häufigkeiten $h_{ij} := n_{ij}/n$ gilt entsprechend:

$$\sum_{i=1}^{k}\sum_{j=1}^{l} h_{ij} = \sum_{i=1}^{k} h_{i\bullet} = 1 = \sum_{j=1}^{l} h_{\bullet j} = \sum_{j=1}^{l}\sum_{i=1}^{k} h_{ij} .$$

3.2 Randverteilungen

Natürlich kann man auch bei zwei- oder mehrdimensionalem Datenmaterial das Augenmerk nur auf das eine oder andere Merkmal richten und die Zusammenhänge zunächst unbeachtet lassen. Man wird dann diese eindimensionalen Merkmale getrennt behandeln und mit den Verfahren des vorigen Kapitels auswerten. Bildlich gesprochen bedeutet dies, daß man nur auf die *Ränder* der Kontingenztabelle schaut und das innere der Matrix nicht beachtet.

---

während j Summations- und Laufindex von Y ist. k bezeichnet die Anzahl der verschiedenen Ausprägungen von X und l die von Y, n ist die Anzahl der Beobachtungen bzw. Merkmalsträger.

Definition: Die beiden eindimensionalen Verteilungen

$$h_{i\bullet} = \text{relH}(X = x_i) = \frac{n_{i\bullet}}{n}, \quad i = 1, \cdots, k \tag{3-4}$$

beziehungsweise

$$h_{\bullet j} = \text{relH}(Y = y_j) = \frac{n_{\bullet j}}{n}, \quad j = 1, \cdots, l \tag{3-5}$$

heißen ***Randverteilungen*** der statistischen Variablen *X* bzw. *Y*.

Betrachtet man aber nur die Ränder, geht die wesentliche Information einer zweidimensionalen Statistik, nämlich die über das gemeinsame Verhalten der Merkmale und deren Abhängigkeit oder Unabhängigkeit, leider verloren.

Berechnung von Mittelwert und Varianz

Die Randverteilungen geben die Verteilung *einer Variablen* an, ganz unabhängig davon, welchen Wert die *andere Variable* gerade hat. Mit der jeweiligen Randverteilung lassen sich Mittelwert und empirische Varianz für jede Variable einzeln berechnen als

$$\bar{x} = \sum_{i=1}^{k} h_{i\bullet} x_i \quad \text{beziehungsweise} \quad \bar{y} = \sum_{j=1}^{l} h_{\bullet j} y_j \tag{3-6}$$

und

$$s_X^2 = \sum_{i=1}^{k} h_{i\bullet} (x_i - \bar{x})^2 \quad \text{bzw.} \quad s_Y^2 = \sum_{j=1}^{l} h_{\bullet j} (y_j - \bar{y})^2 \,. \tag{3-7}$$

Beispiel [1] Abstraktes Rechenbeispiel für eine zweidimensionale Häufigkeitsverteilung.

Die Komponente X hat die $k = 4$ Merkmalsausprägungen

$$x_1 = 30,\ x_2 = 40,\ x_3 = 50,\ x_4 = 60.$$

Die Komponente Y hat die $l = 5$ Merkmalsausprägungen

$$y_1 = 1,\ y_2 = 2,\ y_3 = 4,\ y_4 = 5,\ y_5 = 8.$$

Die Anzahl der Merkmalsträger bzw. Wertepaare ist $n = 200$.
Die gemeinsame Verteilung sei gegeben durch die folgende Tabelle der absoluten Häufigkeiten:

X \ Y	1	2	4	5	8	insge-samt
30	4	8	8	0	0	20
40	4	8	16	20	12	60
50	12	10	16	28	14	80
60	0	4	10	16	10	40
insge-samt	20	30	50	64	36	$200 = n$

Die relativen Häufigkeiten erhält man durch Division aller Werte durch $n = 200$:

X \ Y	1	2	4	5	8	$h_{i\bullet}$
30	0.02	0.04	0.04	0	0	0.10
40	0.02	0.04	0.08	0.10	0.06	0.30
50	0.06	0.05	0.08	0.14	0.07	0.40
60	0	0.02	0.05	0.08	0.05	0.20
$h_{\bullet j}$	0.10	0.15	0.25	0.32	0.18	1

In der letzten Spalte und der untersten Zeile erkennt man die beiden Randverteilungen dieser gemeinsamen Verteilung, und zwar die für X

X	30	40	50	60
$h_{i\bullet}$	0.10	0.30	0.40	0.20

und die für die Komponente Y

Y	1	2	4	5	8
$h_{\bullet j}$	0.10	0.15	0.25	0.32	0.18

Mittelwerte und Varianzen werden mit den Randverteilungen berechnet.

Zunächst für X

$$\begin{aligned}\bar{x} = \sum_{i=1}^{4} h_{i\bullet} x_i &= 0.1 \cdot 30 + 0.3 \cdot 40 + 0.4 \cdot 50 + 0.2 \cdot 60 \\ &= 3 + 12 + 20 + 12 = 47\end{aligned}$$

$$\begin{aligned}s_X^2 = \sum_{i=1}^{4} h_{i\bullet}(x_i - \bar{x})^2 &= 0.1 \cdot (30-47)^2 + 0.3 \cdot (40-47)^2 \\ &\quad + 0.4 \cdot (50-47)^2 + 0.2 \cdot (60-47)^2 \\ &= 0.1 \cdot (-17)^2 + 0.3 \cdot (-7)^2 + 0.4 \cdot (3)^2 + 0.2 \cdot (13)^2 \\ &= 28.9 + 14.7 + 3.6 + 33.8 = 81\end{aligned}$$

$$s_X = \sqrt{81} = 9$$

und dann für Y

$$\begin{aligned}\bar{y} = \sum_{j=1}^{5} h_{\bullet j} y_j &= 0.10 \cdot 1 + 0.15 \cdot 2 + 0.25 \cdot 4 + 0.32 \cdot 5 + 0.18 \cdot 8 \\ &= 0.10 \cdot 1 + 0.15 \cdot 2 + 0.25 \cdot 4 + 0.32 \cdot 5 + 0.18 \cdot 8 \\ &= 0.10 + 0.30 + 1.00 + 1.60 + 1.44 = 4.44\end{aligned}$$

$$\begin{aligned}\sum_{j=1}^{5} h_{\bullet j} y_j^2 &= 0.10 \cdot 1^2 + 0.15 \cdot 2^2 + 0.25 \cdot 4^2 + 0.32 \cdot 5^2 + 0.18 \cdot 8^2 \\ &= 0.10 \cdot 1 + 0.15 \cdot 4 + 0.25 \cdot 16 + 0.32 \cdot 25 + 0.18 \cdot 64 \\ &= 0.10 + 0.60 + 4.00 + 8.00 + 11.52 = 24.22\end{aligned}$$

$$s_Y^2 = 24.22 - (4.44)^2 = 24.22 - 19.7136 = 4.5064$$

$$s_Y = \sqrt{4.5064} = 2.1228\,.$$

3.3 Bedingte Verteilungen und statistische Zusammenhänge

Besonders interessiert bei einer zweidimensionalen statistischen Variablen die Verteilung der relativen Häufigkeiten über einer Variablen, wenn (unter der Bedingung, daß) die andere auf einem bestimmten Wert festgehalten wird. Auf diese Weise erhält man einen wichtigen Einblick in die Art des Zusammenhangs zwischen beiden.

Definition: Die $j = 1, \cdots, l$ eindimensionalen Verteilungen

$$h_{i|y_j} = \text{relH}(X = x_i \,|\, Y = y_j), \quad i = 1, \cdots, k \tag{3-8}$$

und die $i = 1, \cdots, k$ eindimensionalen Verteilungen

$$h_{j|x_i} = \text{relH}(Y = y_j \,|\, X = x_i), \quad j = 1, \cdots, l \tag{3-9}$$

heißen ***bedingte Verteilungen***.

Die bedingten Verteilungen lassen sich leicht aus der Kontingenztabelle entnehmen; man braucht nur die Zeilen oder Spalten der Tabelle durch den ihnen entsprechenden Wert der Randverteilung zu dividieren:

$$h_{i|y_j} = \frac{h_{ij}}{h_{\bullet j}} = \frac{n_{ij}}{n_{\bullet j}} \qquad \text{und} \qquad h_{j|x_i} = \frac{h_{ij}}{h_{i\bullet}} = \frac{n_{ij}}{n_{i\bullet}}$$

Definition: Ist die gemeinsame Verteilung h_{ij} der statistischen Variablen X und Y gleich dem Produkt der beiden Randverteilungen

$$h_{ij} = h_{i\bullet} \cdot h_{\bullet j} \tag{3-10}$$

für $i = 1, \cdots, k$ und $j = 1, \cdots, l$, so heißen X und Y ***statistisch unabhängig***.

Bei unabhängigen statistischen Variablen sind die bedingten Verteilungen **identisch** und jeweils gleich der Randverteilung. Es gilt also für alle $j = 1, \cdots, l$ bedingten Verteilungen von X

$$h_{i|y_j} = \frac{h_{ij}}{h_{\bullet j}} = h_{i\bullet}\,, \quad i = 1, \cdots, k$$

und für alle $i = 1, \cdots, k$ bedingten Verteilungen

$$h_{j|x_i} = \frac{h_{ij}}{h_{i\bullet}} = h_{\bullet j} \ , \qquad j = 1, \cdots, l \ .$$

Beispiel [2] Für die gemeinsame Verteilung aus dem Zahlenbeispiel [1] gibt es fünf bedingte Verteilungen von *X* und eine Randverteilung von *X*:

X	$h_{i\|Y=1}$	$h_{i\|Y=2}$	$h_{i\|Y=4}$	$h_{i\|Y=5}$	$h_{i\|Y=8}$	$h_{i\bullet}$
30	0.2	0.267	0.160	0	0	0.10
40	0.2	0.267	0.320	0.313	0.333	0.30
50	0.6	0.333	0.320	0.437	0.389	0.40
60	0	0.133	0.200	0.250	0.278	0.20
	1	1	1	1	1	1

Alle diese fünf bedingten Verteilungen sind verschieden und keine ist gleich der Randverteilung. Die beiden Komponenten *X* und *Y* sind deshalb hier nicht unabhängig.

Es gibt vier bedingte Verteilungen von *Y* und eine Randverteilung von *Y*:

Y	1	2	4	5	8	
$h_{j\|X=30}$	0.200	0.400	0.400	0	0	1
$h_{j\|X=40}$	0.067	0.133	0.267	0.333	0.200	1
$h_{j\|X=50}$	0.150	0.125	0.200	0.350	0.175	1
$h_{j\|X=60}$	0	0.100	0.250	0.400	0.250	1
$h_{\bullet j}$	0.10	0.15	0.25	0.32	0.18	1

Zusammenfassende Maßzahlen

Die Elemente $\omega_i \ (i = 1, \cdots, n)$ einer statistischen Masse Ω vom Umfang n sind nach zwei Merkmalen untersucht, und die statistischen Variablen

$$x_i = X(\omega_i) \qquad \text{und} \qquad y_i = Y(\omega_i)$$

als Wertepaare erhoben worden. Von beiden Variablen seien sowohl Mittelwerte $\bar{x}$ und $\bar{y}$ als auch die Varianzen s_X^2 und s_Y^2 berechnet. Es gilt für die Summe $Z := X + Y$:

$$\bar{z} = \bar{x} + \bar{y}\,, \tag{3-11}$$

das heißt, *der Mittelwert einer Summe ist gleich der Summe der Mittelwerte*. Entsprechend ist *der Mittelwert einer Differenz gleich der Differenz der Mittelwerte*. Dies gilt ohne Ansehen der gemeinsamen Verteilung der beiden Variablen und ebenso für statistisch unabhängige wie für statistisch abhängige Variablen.

Beispiel [3] Das deutsche Einkommensteuergesetz kennt sieben Einkunftsarten. Viele Steuerpflichtige erzielen Einkünfte aus zwei oder mehreren Einkunftsarten. Seien X die von den Steuerpflichtigen erklärten Einkünfte aus nichtselbständiger Arbeit, Y die aus Kapitalvermögen und Z die Summe aus beiden. Dann gilt sicherlich für den Mittelwert der Summe $\bar{z} = \bar{x} + \bar{y}$. Aber wie ist es mit der Streuung?

Für die Varianz der Summe $Z = X + Y$ erhalten wir durch Anwenden der binomischen Formel

$$\begin{aligned} s_{X+Y}^2 &= \frac{1}{n}\sum_{j=1}^{n}[(x_j + y_j) - (\bar{x} + \bar{y})]^2 \\ &= \frac{1}{n}\sum[(x_j - \bar{x}) + (y_j - \bar{y})]^2 \\ &= \frac{1}{n}\sum[(x_j - \bar{x})^2 + (y_j - \bar{y})^2 + 2\cdot(x_j - \bar{x})(y_j - \bar{y})] \\ s_{X+Y}^2 &= s_X^2 + s_Y^2 + 2\cdot\frac{1}{n}\sum(x_j - \bar{x})(y_j - \bar{y}) \end{aligned} \tag{3-12}$$

und entsprechend für die Varianz der Differenz

$$s_{X-Y}^2 = s_X^2 + s_Y^2 - 2\cdot\frac{1}{n}\sum(x_j - \bar{x})(y_j - \bar{y})\,. \tag{3-13}$$

Nur für den Spezialfall, daß der letzte Term in (3-12) bzw. (3-13) verschwindet, wäre die Varianz einer Summe oder Differenz gleich der Summe der Einzelvarianzen:

$$s^2_{X\pm Y} = s^2_X + s^2_Y , \tag{3-14}$$

falls $\frac{1}{n}\sum(x_j - \bar{x})(y_j - \bar{y}) = 0$.

Ob nun dieser Term, der den linearen statistischen Zusammenhang beider Variablen widerspiegelt, verschwindet oder nicht, hängt von der gemeinsamen Verteilung von X und Y ab.

3.4 Kovarianz und Korrelationskoeffizient

Definition: Die aus den n Wertepaaren (x_i, y_i) berechnete Größe

$$c_{XY} := \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) \tag{3-15}$$

heißt ***empirische Kovarianz*** oder kurz die ***Kovarianz*** zwischen den statistischen Variablen X und Y.

Die Kovarianz ist nichts weiter als das arithmetische Mittel des Produkts der Abweichungen der einzelnen Beobachtungen von ihrem jeweiligen Mittel.

Ähnlich wie bei der Varianz gibt es auch bei der Kovarianz eine *vereinfachte Berechnung*. Statt die Abweichungsprodukte zu mitteln, kann man auch das Produkt der Werte selbst mitteln

$$c_{XY} = \frac{1}{n}\sum_{j=1}^{n} x_j\, y_j - \bar{x}\,\bar{y}$$

und anschließend das Produkt der beiden Mittelwerte abziehen. Die Kurzschreibweise

$$c_{XY} = \overline{xy} - \bar{x}\,\bar{y} \tag{3-15a}$$

drückt dies prägnant aus. Der Beweis ist leicht; man braucht nur die Abweichungsprodukte in (3-15) auszumultiplizieren und die vier Summanden getrennt zu mitteln.

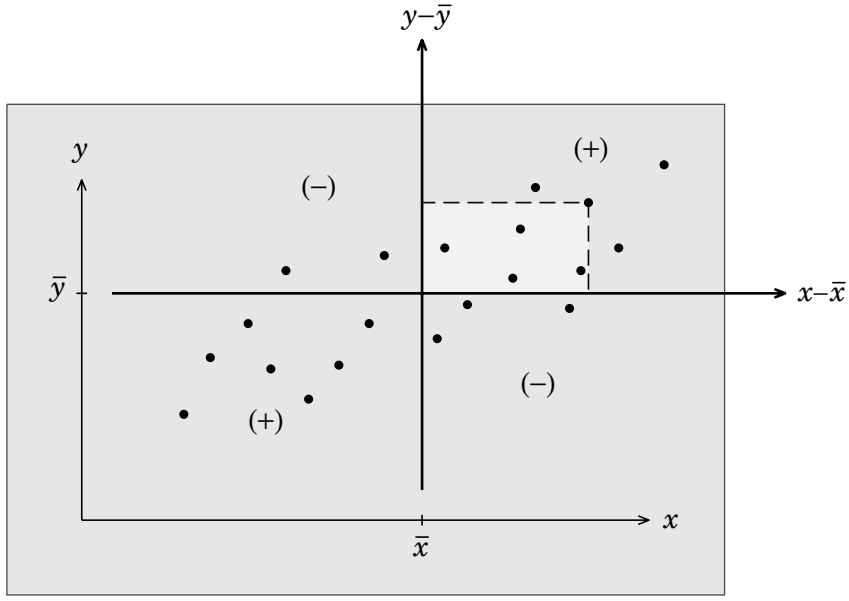

BILD 3.2 Illustration der Kovarianz

Zur Illustration der Kovarianz ist in BILD 3.2 ein Hilfs-Koordinatensystem eingezeichnet, das durch den Schwerpunkt $(\bar{x}, \bar{y})$ der Punktewolke geht. In diesem Koordinatensystem werden die Abweichungen der Beobachtungswerte von ihrem eigenen arithmetischen Mittel gemessen. Deshalb sind seine Achsen mit $x-\bar{x}$ und $y-\bar{y}$ bezeichnet. Die einzelnen Abweichungsprodukte $(x_i-\bar{x})(y_i-\bar{y})$ entsprechen den *Flächen der von den einzelnen Punkten aufgespannten Rechtecke*. Sind die Abweichungen groß, gibt es große, sind sie klein, gibt es kleine Rechtecke. Die Rechtecksflächen im I. und III. Quadranten entsprechen positiven Abweichungsprodukten. Im II. und IV. Quadranten haben die Abweichungen verschiedene Vorzeichen, was ein negatives Produkt ergibt. Überwiegen die positiven Abweichungsprodukte, *bleibt ihre Summe positiv*, überwiegen die Beobachtungswerte im II. und IV. Quadranten, *wird sie negativ*.

Eine positive Kovarianz beschreibt somit eine *gemeinsame Tendenz* der beobachteten Werte x_i und y_i: Relativ große Werte von X gehen im Durchschnitt der Beobachtungen mit relativ großen Werten von Y einher. Entsprechend zeigt eine negative Kovarianz an, daß die Beobachtungswerte im II. und IV. Quadranten überwiegen, das heißt große Werte der einen Variablen eher mit kleinen Werten der anderen einhergehen.

Die Kovarianz kann nur für Wertepaare berechnet werden, oder – was dasselbe ist – für zwei Variablen, die eine gemeinsame Verteilung besitzen. Unter Verwendung ihrer gemeinsamen Verteilung h_{ij} erhält die Definition die folgende Schreibweise:

$$c_{XY} := \sum_{i=1}^{k}\sum_{j=1}^{l} h_{ij}(x_i - \bar{x})(y_j - \bar{y}) \tag{3-15b}$$

Hierin wird deutlich, daß jedes in den Beobachtungen vorkommende Abweichungsprodukt mit seiner relativen Häufigkeit gewichtet berücksichtigt wird.

Sind zwei Variablen X und Y statistisch unabhängig, ist die Kovarianz zwischen ihnen Null.

Man beachte, daß dieser Satz nicht umkehrbar ist; aus der statistischen Unabhängigkeit folgt zwar das Verschwinden der Kovarianz, jedoch liegt keineswegs immer Unabhängigkeit vor, wenn die Kovarianz verschwindet. In der Tat mißt die Kovarianz nur den *linearen Anteil* der statistischen Abhängigkeit.

Definition: Der Quotient

$$r_{XY} := \frac{c_{XY}}{s_X \cdot s_Y} \tag{3-16}$$

heißt (empirischer) ***Korrelationskoeffizient*** zwischen X und Y.

Natürlich läßt sich dieser Quotient nur dann ausrechnen, wenn beide Standardabweichungen im Nenner größer als Null sind. Einige wichtige Eigenschaften des Korrelationskoeffizienten seien beachtet:

1. Mit der Division durch die beiden Standardabweichungen erhält man *ein normiertes* Maß für die Strenge des linearen statistischen Zusammenhanges. Denn ein großer Zahlenwert der Kovarianz kann auch daher rühren, daß die Streuung der beiden Komponenten für sich genommen schon groß ist, obwohl gar keine allzu große lineare Abhängigkeit zwischen ihnen besteht. Die Größe r_{XY} hat das gleiche Vorzeichen wie die Kovarianz, liegt aber stets zwischen –1 und +1, das heißt

$$-1 \le r_{XY} \le +1 .$$

2. Eine weitere Folge der Normierung ist, daß der Korrelationskoeffizient unverändert bleibt, wenn man eine oder beide Variablen linear transformiert, das heißt den Maßstab ändert. Es ist ihm egal, ob man in Dollar, Yen oder Euro rechnet. Um das zu zeigen, definieren wir zwei neue Variablen

$$U := a_1 + b_1 X, \quad \text{mit } b_1 \neq 0$$
$$V := a_2 + b_2 Y, \quad \text{mit } b_2 \neq 0$$

als lineare Transformation von X respektive Y und berechnen den Korrelationskoeffizienten zwischen ihnen. Wir erhalten unter Berücksichtigung der Rechenregel (2-18)

$$r_{UV} = \frac{c_{UV}}{s_U \cdot s_V} = \frac{b_1 \cdot b_2 \cdot c_{XY}}{|b_1|\ s_X \cdot |b_2|\ s_Y} = \frac{b_1 \cdot b_2}{|b_1| \cdot |b_2|} r_{XY},$$

daß sich der Korrelationskoeffizient nicht verändert, solange b_1 und b_2 beide positiv oder beide negativ sind. Andernfalls ändert sich lediglich das Vorzeichen von r, was ja nur plausibel ist.

3. Vertauscht man die Variablen X und Y, ändert sich dadurch nichts am Korrelationskoeffizienten, vielmehr ist

$$r_{XY} = r_{YX}.$$

Beide Merkmale werden in der Korrelationsrechnung symmetrisch behandelt, keines ist gegenüber dem anderen bevorzugt. Es wird zwar eine statistische Abhängigkeit konstatiert, ohne festzulegen, welche der beiden die abhänge oder die unabhängige Variable ist. Das ist in der Regressionsrechnung des folgenden Kapitels anders.

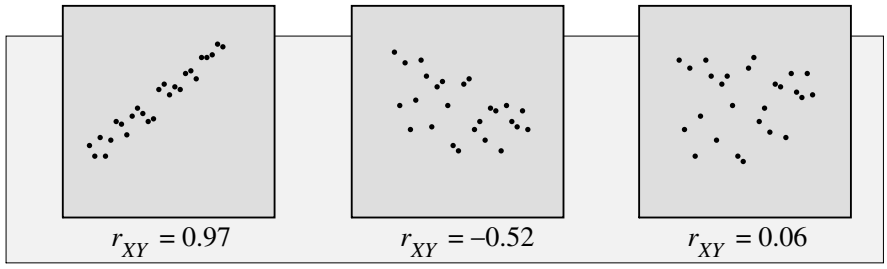

BILD 3.3 Punktewolken und Korrelationskoeffizienten

Beispiel [4] Für die gemeinsame Verteilung aus dem Zahlenbeispiel [1] erhält man für die Kovarianz

$$c_{XY} := \sum_{i=1}^{4}\sum_{j=1}^{5} h_{ij}(x_i - \bar{x})(y_j - \bar{y}) = \sum_{i=1}^{4}\sum_{j=1}^{5} h_{ij} x_i y_j - \bar{x}\,\bar{y}$$

über den *Umweg der vereinfachten Berechnung* zunächst

$$\begin{aligned}\sum_{i=1}^{4}\sum_{j=1}^{5} h_{ij}\,x_i\,y_j &= 0.02\cdot 30\cdot 1 + 0.04\cdot 30\cdot 2 + 0.04\cdot 30\cdot 4 + 0\cdot 30\cdot 5 + 0\cdot 30\cdot 8 \\ &\quad + 0.02\cdot 40\cdot 1 + 0.04\cdot 40\cdot 2 + 0.08\cdot 40\cdot 4 + 0.10\cdot 40\cdot 5 + 0.06\cdot 40\cdot 8 \\ &\quad + 0.06\cdot 50\cdot 1 + 0.05\cdot 50\cdot 2 + 0.08\cdot 50\cdot 4 + 0.14\cdot 50\cdot 5 + 0.07\cdot 50\cdot 8 \\ &\quad + 0\cdot 60\cdot 1 + 0.02\cdot 60\cdot 2 + 0.05\cdot 60\cdot 4 + 0.08\cdot 60\cdot 5 + 0.05\cdot 60\cdot 8 \\ &= 0.6 + 2.4 + 4.8 + 0.8 + 3.2 + 12.8 + 20.0 + 19.2 \\ &\quad + 3.0 + 5.0 + 16.0 + 35.0 + 28.0 + 2.4 + 12.0 + 24.0 + 24.0 \\ &= 7.8 + 56 + 87 + 62.4 = 213.2\end{aligned}$$

und dann

$$c_{XY} = 213.2 - 47\cdot 4.44 = 213.2 - 208.68 = 4.52\,.$$

Der Korrelationskoeffizient beträgt somit

$$r_{XY} = \frac{4.52}{9\cdot 2.1228} = +0.2366\,,$$

was eine schwache positive Korrelation bedeutet.

Es ist sehr wichtig zu betonen, daß Kovarianz und Korrelationskoeffizient nicht zwingend *eine kausale Beziehung* zwischen den Merkmalen bedeuten: Lediglich die gerade vorliegenden Beobachtungen zeigen eine *statistische Tendenz*, diese könnte aber auch rein zufällig sein. Je strenger die Korrelation allerdings ist, um so eher wird man geneigt sein, einen substantiellen Zusammenhang zu vermuten, der aber durch theoretische und sachliche Überlegungen sowie durch weitere empirische Forschungen gestützt werden müßte.

Bravais-Pearson und Spearman

Der oben definierte Korrelationskoeffizient wird oft als BRAVAIS[1]-PEARSON[2]-Korrelationskoeffizient oder PEARSON *r* bezeichnet. Denn es gibt noch einen anderen, nämlich den Korrelationskoeffizienten nach SPEARMAN[3] oder ***Rangkorrelationskoeffizienten***.

---

1 AUGUSTE BRAVAIS, 1811 – 1863) französischer Physiker, Professor an der École Polytechnique, Paris, berühmt durch die Entdeckung der Gitterstruktur der Kristalle (Bravais-Gitter). Wahrscheinlich hat er den Korrelationskoeffizienten „erfunden".

2 KARL PEARSON, 1857 – 1936, englischer Mathematiker und Antropologe am University College, London. Er ist einer der Begründer der modernen Statistik. Außerdem war er noch Rechtsanwalt, Poet und radikaler Politiker, aber nicht verwandt und nicht verschwägert mit dem Verlag, in dem dieses Lehrbuch erscheint.

3 CHARLES EDWARD SPEARMAN, 1863 – 1945, englischer Psychologe und wie PEARSON Professor am University College, London. Er schuf die Ansätze zur objektiven Messung von Intelligenz und anderen menschlichen Fähigkeiten.

Man verwendet ihn bei ordinal skalierten Merkmalen. Er ist nichts anderes als der *Korrelationskoeffizient zwischen den Rangplätzen* der Beobachtungen

$$r_{XY}^{\mathrm{Sp}} := r_{\mathrm{rg}(X),\,\mathrm{rg}(Y)}. \tag{3-17}$$

Die Formel zur Berechnung dieses Koeffizienten ist im Prinzip die gleiche, mit dem Unterschied, daß nicht mit den gemessenen Variablenwerten (x_i, y_i) selbst, sondern mit ihren Rangplätzen $[\mathrm{rg}(x_i), \mathrm{rg}(y_i)]$ gerechnet wird. Die Rangplätze sind die Indizes, nachdem die Beobachtungswerte der Größe nach sortiert worden sind. Es spielt dabei keine Rolle, ob man dem größten oder dem kleinsten Wert den Rangplatz 1 zuweist.

Treten dabei zwei oder mehrere gleich große Werte auf, so numeriert man zunächst einfach durch, ordnet aber dann den gleichen Werten das arithmetische Mittel ihrer Rangplätze zu.

Beispiel [5] Die folgende Tabelle zeigt die Ergebnisse der Abiturprüfungen von zehn Schülern in den Fächern Deutsch (Merkmal D) und Geschichte (Merkmal G). Die maximal erreichbare Punktzahl beträgt jeweils 15.

Schüler i	Deutsch D	Geschichte G	rg(D)	rg(G)
1	13	15	4	1
2	14	8	2.5 (2)	4 (3)
3	8	1	9	10
4	10	7	7	6.5 (6)
5	15	9	1	2
6	1	5	10	9
7	14	8	2.5 (3)	4 (4)
8	12	7	5	6.5 (7)
9	9	6	8	8
10	11	8	6	4 (5)

Sind die Noten korreliert? Gehen gute Leistungen in Deutsch mit guten Geschichtskenntnissen einher? Zuerst werden für jeden Schüler in jedem der beiden Fächer die Rangplätze bestimmt. Dazu ordnen wir die Schüler nach den von ihnen erzielten Ergebnissen in den Fächern an. Schülern mit gleichem Ergebnis wird das arithmetische Mittel derjenigen Rangplätze zugeordnet, die sie bei willkürlicher Anordnung erhalten hätten (sind in Klammern jeweils angegeben). So kann es zu Rangplätzen 2.5 oder 6.5 kommen. Dann berechnen wir Varianzen, Standardabweichungen und die Kovarianz der Rangplätze und erhalten mit

$$r_{DG}^{\mathrm{Sp}} = \frac{6.95}{2.8636 \cdot 2.8284} = 0.8581$$

eine recht positive Korrelation, was zu erwarten war.

Man wird im Einzelfall auch dann lieber die Rangkorrelation berechnen, wenn man der Qualität der Skala eines oder beider Merkmale nicht recht traut, also nicht weiß, ob sie abstandstreu ist. Bei Examensnoten etwa werden die meisten zustimmen, daß eine 1 wohl besser ist als eine 2, aber ob die Differenz zwischen der 1 und der 2 genau so viel bedeutet wie die zwischen der 2 und der 3, ist fraglich.

Während der BRAVAIS-PEARSON-Korrelationskoeffizient den linearen statistischen Zusammenhang angibt, mißt der SPEARMANsche Rangkorrelationskoeffizient nur den *monotonen* Anteil des statistischen Zusammenhangs der beiden Variablen. *Streng monotone Transformationen* der beiden Variablen verändern ihn nicht, denn sie lassen die Rangplätze unverändert. Die linearen Transformationen gehören natürlich zu den monotonen Transformationen, aber auch etwa das Logarithmieren wäre eine monotone Transformation. Auch hier ändert der Korrelationskoeffizient allenfalls das Vorzeichen, nämlich genau dann, wenn die eine Transformation streng monoton fallend war und die andere steigend.

3.5 Kontingenzkoeffizient

Die Berechnung und sinnvolle Interpretation der Kovarianz und des Korrelationskoeffizienten setzt voraus, daß die statistischen Variablen eine metrische Meßbarkeit haben. Für den Rangkorrelationskoeffizienten reicht eine ordinale Meßbarkeit aus, aber wie mißt man den statistischen Zusammenhang, wenn nur nominalskalierte Merkmale vorliegen?

Ausgangspunkt für die Überlegungen ist der Begriff der *statistischen Unabhängigkeit*. Nach der der Definition (3-10) würden zwei Komponenten X und Y als statistisch unabhängig bezeichnet werden, wenn sich ihre gemeinsame Verteilung aus dem Produkt der beiden Randverteilungen

$$h_{ij} = h_{i\bullet} \cdot h_{\bullet j} \tag{3-18}$$

für $i = 1, \cdots, k$ und $j = 1, \cdots, l$ berechnen ließe. In *absoluten* Häufigkeiten ausgedrückt würde das Unabhängigkeitskriterium

$$E_{ij} := n h_{i\bullet} \cdot h_{\bullet j} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \tag{3-19}$$

lauten. Dabei ist zu beachten, daß die Zahlen E_{ij} eben hypothetische Werte sind, die auch keineswegs ganzzahlig zu sein brauchen. Um das Ausmaß der Abhängigkeit zu quantifizieren, wird man auf die Abweichungen

$$n_{ij} - E_{ij}$$

schauen. Im allgemeinen aber sind empirische gemeinsame Verteilungen nicht un-

abhängig, sondern es gibt mehr oder weniger große Abweichungen. Je stärker die tatsächlichen Häufigkeiten von den hypothetischen abweichen, um so größer wird der statistische Zusammenhang sein. Um eine Maßzahl zu gewinnen, quadriert man die Abweichungen, teilt sie durch den hypothetischen Wert und summiert über alle Felder der Kontingenztabelle auf.

Definition: Die Summe der relativen quadratischen Abweichungen

$$QK := \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \tag{3-20}$$

heißt ***quadratische Kontingenz*** oder ***Chi-Quadrat-Koeffizient***.

Die quadratische Kontingenz wäre im Falle vollkommener Unabhängigkeit natürlich Null, in allen anderen Fällen positiv, und sie kann, wenn Abhängigkeit vorliegt, für große n sehr groß werden. Deswegen ist sie als Zusammenhangsmaß nicht besonders geeignet. Man würde ein normiertes Maß vorziehen. Der ***Kontingenzkoeffizient***

$$KK := \sqrt{\frac{QK}{QK + n}}$$

ist ebenfalls Null, wenn die quadratische Kontingenz Null ist. Für großes QK wird auch KK größer, erreicht den Wert Eins aber nicht ganz, sondern maximal den Wert $KK_{\max}$

$$0 \le KK \le KK_{\max} = \sqrt{\frac{m-1}{m}} < 1 ,$$

der von der Größe der Kontingenztabelle abhängt, das heißt von ihrer Zeilenzahl k und Spaltenzahl l, wobei m die kleinere von beiden ist. Unter Berücksichtigung dieses Sachverhalts korrigiert man den Kontingenzkoeffizienten in einem zweiten Normierungsschritt.

Definition: Die Größe

$$\boxed{KK^* := \frac{KK}{KK_{\max}} = \sqrt{\frac{QK \cdot m}{(QK + n)(m-1)}}} \tag{3-21}$$

heißt ***korrigierter Kontingenzkoeffizient***.

Es ist nun $0 \le KK^* \le 1$, und man kann damit auch die Stärke des Zusammenhangs von verschiedenen Kontingenztabellen eher vergleichen als mit KK.

Beispiel [6] Streben männliche und weibliche Jugendliche in Deutschland in die gleichen Berufe? Die folgende Kontingenztabelle zeigt die gemeinsamen Häufigkeiten der beiden Merkmale Geschlecht und Ausbildungsbereich in Deutschland im Jahr 1999 (in tausend Personen):

TABELLE 3.1a Azubis in Deutschland

Ausbildungs-bereich	männlich	weiblich	gesamt
Industrie und Handel	471.5	361.5	833.0
Handwerk	485.5	131.4	616.9
öffentlicher Dienst	17.6	29.9	47.5
	974.6	522.8	1497.4

Quelle: Deutschland in Zahlen 2001, Institut der deutschen Wirtschaft

Wäre die Berufswahl *unabhängig vom Geschlecht*, müßte die gemeinsame Verteilung etwa so aussehen:

TABELLE 3.1b Verteilung der Azubis bei Unabhängigkeit

Ausbildungs-bereich	männlich	weiblich	gesamt
Industrie und Handel	*542.17*	*290.83*	833.0
Handwerk	*401.52*	*215.38*	616.9
öffentlicher Dienst	*30.92*	*16.58*	47.5
	974.6	522.8	1497.4

Mit Hilfe der folgenden Arbeitstabelle berechnen wir zuerst die quadratische Kontingenz:

i	j	$n_{i\bullet}$	$n_{\bullet j}$	n_{ij}	$E_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$	$\frac{(n_{ij} - E_{ij})^2}{E_{ij}}$
1	1	833.0	974.6	471.5	*542.1676*	9.2114
2	1	616.9	974.6	485.5	*401.5165*	17.5665
3	1	47.5	974.6	17.6	*30.9159*	5.7354
1	2	833.0	522.8	361.5	*290.8324*	17.1711
2	2	616.9	522.8	131.4	*215.3835*	32.7473
3	2	47.5	522.8	29.9	*16.5841*	10.6918
						QK = 93.1231

Der korrigierte Kontingenzkoeffizient

$$KK^* = \sqrt{\frac{93.1231 \cdot 2}{(93.1231 + 1497.4) \cdot (2-1)}} = 0.3422$$

zeigt, daß die Berufswahl auch heute durchaus nicht unabhängig vom Geschlecht ist.

Der Kontingenzkoeffizient kann natürlich auch für *ordinale* und sogar *metrische Merkmale* berechnet und sinnvoll interpretiert werden. Jedoch ist zu beachten, daß er nur angibt, *wie stark* der Zusammenhang ist, aber nichts über die Richtung des Zusammenhanges aussagt, wie es etwa der Korrelationskoeffizient tut. Man kann aufgrund eines großen *KK* eben nicht sagen, daß große Werte der einen Variablen tendenziell mit großen Werten der anderen einhergehen. Das liegt daran, daß eben bei der Berechnung der *QK* nur das Nominalskalenniveau beachtet wird. Größen und Abstände der Merkmalswerte werden nicht berücksichtigt, sie kommen in den Formeln gar nicht vor. Auch beliebige Umstellungen von Spalten oder Zeilen in der Kontingenztabelle verändern nichts an den Kontingenzmaßen.

Beispiel [7] Die folgenden Verteilungen haben alle die gleichen korrigierten Kontingenzkoeffizienten, aber verschiedene Korrelationskoeffizienten:

		Y		
		1	3	5
X	4	12		
	5		8	
	6		3	10

$r = 0.9438$

$KK^* = 0.9560$

		Y		
		1	2	6
X	10	12		
	20			8
	30		10	3

$r = 0.3679$

$KK^* = 0.9560$

		Y		
		10	15	20
X	1	10	3	
	2			12
	3		8	

$r = 0.4895$

$KK^* = 0.9560$

Der korrigierte Kontingenzkoeffizient einer Verteilung ist genau dann eins, wenn in jeder Zeile höchstens eine Spalte und jeder Spalte höchstens eine Zeile mit Häufigkeiten besetzt ist und somit *vollkommene Abhängigkeit* besteht.

Kontrollfragen

1 Was ist der Unterschied zwischen univariater und multivariater Statistik? Überlegen Sie sich ein Beispiel der bivariaten Statistik!

2 Welchen Aufbau und welche Funktion haben Kontingenztabellen? Gibt es auch Kontingenztabellen für mehr als zwei Merkmale?

3 Wie viele Randverteilungen hat eine 3-dimensionale statistische Verteilung?

4 Wann ist die Varianz einer Summe kleiner als die Summe der Varianzen?

5 Was ist statistische Unabhängigkeit? In welchem Zusammenhang steht hierbei die Kovarianz?

6 Was sagt der Korrelationskoeffizient aus? Bedeutet ein empirischer Korrelationskoeffizient von 0, daß es keinen sachlichen Zusammenhang zwischen den betrachteten Merkmalen gibt?

7 Was ist eine Rangkorrelation? Womit mißt man sie?

8 Warum ist die quadratische Kontingenz nicht von den Variablenwerten abhängig?

ERGÄNZENDE LITERATUR

Everitt, B. S.: *The analysis of contingency tables*, 2. Auflage, Boca-Raton: Chapman & Hall, 2000

Fahrmeir, L.; Künstler, R.; Pigeot, I.; Tutz, G.: *Statistik: Der Weg zur Datenanalyse*, 4. Aufl., Berlin, Heidelberg, New York: Springer, 2002, Kapitel 3

Hartung, J.; Elpelt, B.: *Multivariate Statistik*, 6. Auflage, München, Wien: Oldenbourg, 1999

Kendall, M. G.; Gibbons J. D.: *Rank correlation methods*, 5. Auflage, New York: Oxford University Press, 1990

Kotz, S.; Drouet, M. D.: *Correlation and Dependence*, London: Imperial College Press, 2001

Wickens, Th. D.: *Multiway contingency tables analysis for the social sciences*, Hillsdale: Lawrence Erlbaum Associates, 1989

PRAXIS

Zahlt sich ein Studium aus?

Häufig ist die Frage gestellt worden, ob sich ein Studium überhaupt lohnt. Wird man im späteren Leben ein höheres Einkommen erzielen, wenn man besser ausgebildet ist, einen Master oder gar einen Doktortitel hat? Um die Frage zu klären, werden die Erhebungen der Einkommens- und Verbrauchsstichprobe (EVS) herangezogen. Die EVS wird vom STATISTISCHEN BUNDESAMT seit 1962 in der Regel alle fünf Jahre erstellt und erfaßt 0.2% aller privaten Haushalte in Deutschland. Aus den Daten von 1993 errechnen wir die

absolute Häufigkeitsverteilung des jährlichen Bruttoeinkommens des Haushaltsvorstandes in Abhängigkeit vom Ausbildungsabschluß und erhalten:

TABELLE 3.2 Verteilung des Bruttoeinkommens

Bruttoeinkommen in tsd DM	Berufs-fachschule	Meister/ Techniker	FH	Uni	Σ
bis 30	1 336	311	196	318	2 161
30 – 50	2 958	539	394	542	4 433
50 – 70	3 565	831	770	654	5 820
70 – 90	2 185	688	852	995	4 720
90 – 110	1 295	456	578	774	3 103
110 – 130	626	270	331	517	1 744
130 – 150	334	130	257	357	1 078
150 – 170	157	70	127	245	599
170 – 190	80	44	61	120	305
über 190	69	32	57	157	315
Σ	12 605	3 371	3 623	4 679	24 278

Quelle: Einkommens- und Verbrauchsstichprobe 1993

Bruttoeinkommen in tsd DM	Berufs-fachschule	Meister/ Techniker	FH	Uni	alle
bis 30	10.6	9.2	5.4	6.8	8.9
30 – 50	23.5	16.0	10.9	11.6	18.3
50 – 70	**28.3**	**24.7**	21.3	14.0	24.0
70 – 90	17.3	20.4	**23.5**	**21.3**	19.4
90 – 110	10.3	13.5	16.0	16.5	12.8
110 – 130	5.0	8.0	9.1	11.0	7.2
130 – 150	2.6	3.9	7.1	7.6	4.4
150 – 170	1.2	2.1	3.5	5.2	2.5
170 – 190	0.6	1.3	1.7	2.6	1.3
über 190	0.5	0.9	1.6	3.4	1.3
Σ	100	100	100	100	100
arithm. Mittel	66.715	75.411	86.179	92.444	75.786
Standard-abweichung	34.139	38.091	41.006	46.196	39.761
Median	59.820	70.100	79467	86730	68.963
3. Quartil	83.907	95.905	106.833	118.661	95.852
90%-Quantil	110.141	123.178	139.568	153.187	128.041

Wie wir sehen, kann man auch mit geringer Ausbildung hohe Einkommen erzielen und umgekehrt. Zur Beantwortung der Frage nach der statistischen Abhängigkeit oder Unabhängigkeit von individueller Ausbildung und Einkommen schauen wir auf die bedingten Verteilungen, aber auch auf die durchschnittlichen Einkommen und die Medianeinkommen. Die Durchschnittseinkommen sind größer als die Mediane, die

Verteilungen sind also rechtsschief, was typisch ist für Einkommensverteilungen. Der Kontingenzkoeffizient *KK* beträgt 0.281.

Fazit: Es ist also eine deutliche Abhängigkeit der Bruttoeinkommen vom Ausbildungsniveau erkennbar. Gleichwohl sind die absoluten Einkommensunterschiede wenig dramatisch, auch verglichen mit den Standardabweichungen innerhalb einer Gruppe. Außerdem wird durch den progressiven Einkommensteuertarif, die Sozialgesetzgebung und die öffentlichen Leistungen eine weitere Nivellierung der Nettoeinkommen erreicht. Um die eingangs gestellte Frage zu beantworten, wären noch die Kosten eines Studiums zu bedenken, die Alternativkosten des entgangenen Einkommens bei einer anderen Beschäftigung, aber auch weitere, sich nicht in Geldeinkommen ausdrückende Erträge.

AUFGABEN

3.1 **Zwillingsforschung.** Der bekannte Psychologe A. Skinner mißt den Intelligenzquotienten *IQ* von sieben eineiigen Zwillingen, die nach der Geburt voneinander getrennt worden waren. In der folgenden Tabelle stehen in der ersten Zeile (*X*) die *IQ*s der im Elternhaus aufgewachsenen, in der zweiten Zeile (*Y*) die der bei Pflegeeltern aufgewachsenen Testpersonen. Untereinander stehen jeweils die *IQ*s eines Zwillingspaares:

X :	98	100	104	104	102	102	104
Y :	94	94	103	105	99	102	103

Untersuchen Sie den Zusammenhang in dieser Statistik

a) indem Sie den möglichen statistischen Zusammenhang geeignet graphisch darstellen und erläutern und

b) den Zusammenhang rechnerisch ermitteln und interpretieren.

3.2 Berechnen Sie für die statistischen Reihen in Aufgabe **2.11** die Kovarianzen und Korrelationskoeffizienten

a) c_{XY}, c_{YZ} b) c_{ZU}, c_{VT}, c_{UV}

c) r_{UW}, r_{ZU}, r_{VT}

3.3 Gegeben ist die statistische Reihe *X*. Sie hat den Mittelwert 240 und die Varianz 81. Die statistische Reihe *Y* errechnet sich aus *X*, indem man jedes Element der Reihe *X* mit dem konstanten Faktor $b > 0$ multipliziert, also

$$y_i := b\,x_i \quad \text{für } i = 1, \cdots, n\,.$$

a) Berechnen Sie dir Kovarianz zwischen *X* und *Y* und den Korrelationskoeffizienten.

b) Welchen Wert hat r_{XY}, wenn der Faktor *b* negativ ist?

3.4 **Refa.** Gehen Sie von dem Sachverhalt und dem statistischen Material der Aufgabe **2.9** aus.

a) Zeichnen Sie ein sorgfältiges Streudiagramm. Berechnen Sie den Korrelationskoeffizienten.

b) Welcher der folgenden Aussagen:

(1) Die Korrelation zwischen den Arbeitszeiten für den 1. und 2. Arbeitsgang ist positiv

(2) Die Korrelation ist stark negativ

(3) Die Korrelation ist schwach negativ

(4) Es gibt keinen linearen statistischen Zusammenhang

würden Sie zustimmen?

c) Versuchen Sie, das statistische Ergebnis der Erhebung bezüglich Geschicklichkeit und Sorgfalt verbal zu interpretieren.

3.5 **Erwerbstätige.** In der amtlichen Statistik finden Sie folgende Verteilung der Erwerbstätigen in der Bundesrepublik Deutschland für April 1990 (in 1000 Personen):

Altersgruppe von ... bis unter ...	Selbständige und mithelfende Familienangehörige	abhängig Beschäftigte
15 – 25	99	5002
25 – 35	531	7009
35 – 45	1243	5731
45 – 55	937	6051
55 – 65	595	2284
65 – 75	160	63
75 – 95	42	16

Quelle: Statistisches Jahrbuch 1992

a) Was sind die statistischen Einheiten, Grundgesamtheiten und Merkmale?

b) Zeichnen Sie ein Histogramm der Randverteilung und der beiden bedingten Verteilungen des Merkmals Alter.

c) Zeichnen Sie beide bedingten Verteilungsfunktionen in ein Koordinatensystem. Geben Sie die bedingten Mediane an.

d) Berechnen Sie die beiden bedingten Mittelwerte.

e) Müssen die Selbständigen länger arbeiten? Welcher Anteil der Selbständigen und welcher Anteil der Unselbständigen ist 55 Jahre und älter? Welcher Anteil der über 65jährigen Erwerbstätigen ist selbständig? Kann man aus diesen Daten die durchschnittliche „Lebensarbeitszeit" berechnen?

Hinweis: Gehen Sie von der Annahme einer gleichmäßigen Verteilung innerhalb der Altersgruppen aus.

3.6 **Der Verschiebungssatz** für die empirische Kovarianz lautet:

$$c_{XY} = \overline{(x-a)(y-b)} - \overline{(x-a)} \cdot \overline{(y-b)},$$

wobei a und b konstanten Größen sind. Beweisen Sie diesen Satz.

LÖSUNGEN

3.1 b) 0.936

3.2 a 0; 0 b) 0; 0; –
c) 0.09167; – ; 0

3.3 a) $81b$; 1 b) –1

3.4 a) –3.3333; – 0.5556

3.5 c) 46; 39
d) 46.2; 39.4
e) 22.1%; 9.0%
71.9%; 28.1%