Mit eLearning #besser lernen

Statistik ohne Angst vor Formeln

Das Studienbuch für die Wirtschafts- und Sozialwissenschaften

7., aktualisierte Auflage

Andreas Quatember

Pearson

Zugangscode

Falls Sie beim Kauf Ihres eBooks keinen Zugangscode erhalten haben, kontaktieren Sie uns bitte über die folgende Seite und halten Sie Ihre Rechnung/Bestellbestätigung bereit:
https://www.pearson.de/ebook-zugangscode

im selben Land) einige wenige Superreiche sehr, sehr viel und auch alle anderen leben zumindest in einem angenehmen Wohlstand, dann ist der Ginikoeffizient des Gesamteinkommens in diesem zweiten Land größer als im ersten. Aber wo würden Sie lieber leben? Der Philosoph Harry Frankfurt schreibt darüber: „From the point of view of morality, it is not important everyone should have the *same*. What is morally important is that each should have *enough*." (Weiterführende Literatur findet sich etwa in: Zöfel (2003), ▶ *Kapitel 11*; vertiefende Literatur in: von der Lippe (2002), ▶ *Kapitel 6*).

Übungsaufgaben

Die Lösungen zu den nachfolgenden Übungsbeispielen finden Sie auf der Produktseite zum Buch auf www.pearson.de und im zugehörigen MyLab.

Für die mit * markierten Aufgaben verwenden Sie die auf der Produktseite zum Buch auf www.pearson.de und im MyLab ebenfalls bereitgestellte „Excel- Lerndatei".

Ü25*

Erstellen Sie unter Verwendung der Excel-Lerndatei bei Einkommen von fünf Personen in der Höhe von 7.000, 9.000, 10.000, 7.000 und nochmals 7.000 Euro in Excel die für das Zeichnen der Lorenzkurve der Konzentration der Merkmalssumme auf die einzelnen Erhebungseinheiten nötige Tabelle und zeichnen Sie diese anschließend der Anleitung folgend. Berechnen Sie ferner mit dem Taschenrechner den dazugehörigen Ginikoeffizienten.

Ü26

Laut der Lohnsteuerstatistik eines Landes (sehr großes N) verdienen die 25 Prozent der am schlechtesten verdienenden Arbeitnehmer*innen 3,9 Prozent des Gesamtbruttolohns, die 50 Prozent der mittleren Verdienenden 42,6 Prozent des Gesamtbruttolohns und die 25 Prozent der am besten Verdienenden insgesamt 53,5 Prozent des Gesamten. Skizzieren Sie die Lorenzkurve der Konzentration des Gesamtbruttolohns auf die drei Gruppen von Arbeitnehmer*innen und berechnen Sie den normierten Ginikoeffizienten.

1.3.4 Kennzahlen des statistischen Zusammenhanges

Im ▶ *Abschnitt 1.2.2* haben wir bereits die gemeinsame Häufigkeitsverteilung zweier Merkmale betrachtet: Es stellte sich bei dieser Betrachtung heraus, dass es vor allem die bedingten Verteilungen sind, die interessant sein können. Es wurde in ▶ *Beispiel 1.7* gezeigt, dass sich das Merkmal Studienrichtung aus ▶ *Beispiel 1.6* unter den Frauen und unter den Männern unterschiedlich verteilt hat. Das heißt anders formuliert, dass das Merkmal Geschlecht, das hier mit nur zwei Kategorien erfasst wurde, offenbar einen Zusammenhang mit dem Merkmal Studienrichtung aufgewiesen hat.

Dieser Zusammenhang ist ein statistischer (also in den Daten vorhandener), dessen Begründung nicht von den Daten mitgeliefert werden kann. Es bleibt also völlig offen, ob das Geschlecht der Befragten kausal direkt mit deren Studienrichtung oder indirekt über ein anderes Merkmal (oder mehrere andere) zusammenhängt. Ein diese Problematik sehr schön aufzeigendes Beispiel ist der **statistische Zusammenhang** zwischen den Umsätzen an Speiseeis in den Sommermonaten und den Versicherungsfallzahlen wegen Waldbränden. Das eine hat aber mit dem anderen direkt nichts zu tun. Beide Merkmale werden durch die Temperatur in diesen Monaten gesteuert. Eine Verringe-

rung des Speiseeiskonsums würde sich keineswegs auf die Versicherungsfälle wegen Waldbränden auswirken. Der Zusammenhang zwischen den beiden Merkmalen ist ein statistischer, aber keineswegs ein kausaler. Die Anwender*innen der statistischen Methoden zur Messung des Zusammenhangs zweier Merkmale müssen sich schon selbst Gedanken zur Begründung dieses Zusammenhangs machen. Das Ergebnis seiner Messung tut das nicht für sie!

Ein weiteres Beispiel für eine möglicherweise etwas voreilige Erklärung für einen statistischen Zusammenhang findet sich in der Zeitschrift „Gesund & Vital“ (Ausgabe Juli 2000) unter der Überschrift „Schwangerschaft und Zahnfleisch“: „Ärzte in den USA haben herausgefunden, dass schwangere Frauen mit Zahnfleischerkrankungen ein sieben- bis neunmal höheres Risiko für Fehlgeburten tragen. Rund 800 Frauen wurden untersucht. Der eindeutige Rat als Ergebnis der Studie: Die Zahn- und Zahnfleischuntersuchung soll selbstverständlicher Bestandteil eines jeden Vorsorgebesuches der Schwangeren beim Arzt sein.“ Das kann möglicherweise sogar stimmen. Doch sicher können wir uns nur auf Basis des gefundenen Zusammenhangs zwischen Zahnfleischerkrankungen und Fehlgeburten nicht sein. Genauso gut könnte es sein, dass in den USA die Intensität der ärztlichen Betreuung vom Einkommen der Familien abhängt und dass von Familien aus niedrigen Einkommensklassen demnach sowohl die dentale wie auch die pränatale Vorsorge weniger stark in Anspruch genommen wird als von reicheren Familien, die sich dies eher leisten können. Dadurch würden in beiden medizinischen Bereichen bei Ärmeren häufiger Probleme auftauchen als bei Reicheren. Dann wäre die Befolgung des Tipps, zur Zahn- und Zahnfleischuntersuchung zu gehen, nur gut für das Zahnfleisch und würde sich überhaupt nicht auf das Risiko von Fehlgeburten auswirken!

In diesem Abschnitt machen wir es uns zur Aufgabe, den Grad solcher *statistischer* Zusammenhänge durch eine Kennzahl zu messen. Es ist dabei evident, dass es – wie bei den Kennzahlen der Lage – für die verschiedenen Merkmalstypen wieder unterschiedliche Kennzahlen geben muss. Was aber tun, wenn der Zusammenhang zwischen zwei Merkmalen zu messen ist, die nicht dem gleichen Merkmalstyp angehören? Da man ein metrisches Merkmal wie ein ordinales beziehungsweise ein nominales Merkmal behandeln kann (dabei wird ein Informationsverlust in Kauf genommen) und ein ordinales wie ein nominales, jedoch nicht zum Beispiel ein nominales wie ein metrisches, gilt die Hierarchie: metrisch – ordinal – nominal. Es ist dann jene Kennzahl zu verwenden, die für den „niedrigeren Merkmalstyp“ der beiden Merkmale geeignet ist.

Nominale Merkmale

Wenn man wie in ▶ *Beispiel 1.6* zwei nominale Merkmale vorliegen hat (oder ein nominales und ein anderes in verschiedene Kategorien eingeteiltes Merkmal), dann gibt es – wie oben beschrieben – einen Zusammenhang zwischen den beiden Merkmalen, wenn die bedingten Verteilungen des einen Merkmals (zum Beispiel Studienrichtung) unter den durch die Merkmalsausprägungen des anderen Merkmals erzeugten Teilgesamtheiten (etwa unter den Frauen und den Männern) nicht gleich sind. Dies heißt ja, dass man aus der Kenntnis der Ausprägung einer Erhebungseinheit beim einen Merkmal eine Information über die Ausprägung beim anderen schöpfen kann. Wie könnte man aber den Grad der Stärke des statistischen Zusammenhangs zweier nominaler Merkmale durch eine Kennzahl darstellen? Betrachten wir zur Darstellung der Idee nochmals die Daten aus ▶ *Beispiel 1.6*.

Beispiel 1.13: Die Idee zur Messung des Zusammenhangs zweier nominaler Merkmale

Die Häufigkeiten der Merkmale Geschlecht und Studienrichtung betragen:

	Studienrichtung					
Geschlecht	BWL	Soz	VWL	Sowi	Stat	Summe
weiblich	110	120	20	30	20	300
männlich	90	60	30	10	10	200
Summe	200	180	50	40	30	500

Tabelle 1.13

In relativen Häufigkeiten ergibt sich folgendes Bild:

	Studienrichtung					
Geschlecht	BWL	Soz	VWL	Sowi	Stat	Summe
weiblich	0,22	0,24	0,04	0,06	0,04	0,60
männlich	0,18	0,12	0,06	0,02	0,02	0,40
Summe	0,40	0,36	0,10	0,08	0,06	1

Tabelle 1.14

Es wurde in der Erhebung also beobachtet, dass zum Beispiel 40 Prozent der befragten Studierenden BWL, 36 Prozent Soziologie und so weiter studieren. Wenn es nun keinerlei statistischen Zusammenhang zwischen den beiden Merkmalen Geschlecht und Studienrichtung gäbe, dann müssten doch auch in den Teilgesamtheiten der weiblichen und der männlichen Befragten jeweils 40 Prozent BWL, 36 Prozent Soziologie und so fort studieren, die bedingten Verteilungen der Studienrichtung unter den Frauen und unter den Männern gleich sein. Das heißt also, dass es genau dann keinen statistischen Zusammenhang zwischen den beiden Merkmalen gibt, wenn die bedingten Verteilungen der Studienrichtung unter den Frauen und unter den Männern der Randverteilung des Merkmals Studienrichtung unter allen Befragten entsprechen. Demnach müsste die Tabelle der gemeinsamen Häufigkeitsverteilung dieser beiden Merkmale, wenn sie keinen Zusammenhang aufweisen würden, so aussehen, dass sich die daraus ergebenden bedingten Verteilungen nicht unterscheiden und den jeweiligen Randverteilungen entsprechen. Unter den 300 befragten Frauen müssten sich also in unserem Beispiel genauso 40 Prozent für BWL entscheiden wie unter den 200 befragten Männern. Also müssten sich unter den Befragten, wenn es keinen Zusammenhang zwischen Geschlecht und Studienrichtung gibt, 120 weibliche und 80 männliche BWL-Studierende befinden. Dies heißt, dass die relative Häufigkeit der weiblichen BWL-Studierenden $120{:}500 = 0{,}24$ und die der männlichen $80{:}500 = 0{,}16$ betragen müsste. Diese relativen Häufigkeiten bei Fehlen eines Zusammenhangs erhält man auch ohne den Umweg über die Häufigkeiten aus den relativen Häufigkeiten (beziehungsweise Prozentzahlen), da doch von den 60 Prozent weiblichen Befragten 40 Prozent und von den 40 Prozent männlichen ebenfalls 40 Prozent BWL studieren müssten. In relativen Häufigkeiten ist dies ebenso $0{,}6 \cdot 0{,}4 = 0{,}24$ und $0{,}4 \cdot 0{,}4 = 0{,}16$. Die relativen Häufigkeiten der gemeinsamen Verteilung bei Fehlen eines Zusammen-

hangs ergeben sich also durch Multiplikation der jeweiligen relativen Randhäufigkeiten in der Tabelle.

Die vollständige Tabelle für den Fall, dass kein Zusammenhang zwischen Geschlecht und Studienrichtung vorliegt, hat demnach folgendermaßen auszusehen:

	Studienrichtung					
Geschlecht	BWL	Soz	VWL	Sowi	Stat	Summe
weiblich	0,24	0,216	0,06	0,048	0,036	0,60
männlich	0,16	0,144	0,04	0,032	0,024	0,40
Summe	0,40	0,36	0,10	0,08	0,06	1

Tabelle 1.15

Tatsächlich sind die beobachteten relativen Häufigkeiten (siehe ▶ *Tabelle 1.14*) somit geringfügig anders, nämlich zum Beispiel 0,22 und 0,18 und nicht 0,24 und 0,16. Da also die beobachtete Verteilung in der Erhebung der 500 Wahlberechtigten von dieser bei Fehlen eines Zusammenhangs zu erwartenden Verteilung (siehe ▶ *Tabelle 1.15*) abweicht, liegt hier *keine* statistische Unabhängigkeit der beiden Merkmale vor. Wie stark aber ist der Zusammenhang? Da die Abweichungen der tatsächlich auftretenden relativen Häufigkeiten von den bei Fehlen eines Zusammenhangs zu erwartenden gering ist, sollte man meinen, dass ein schwacher Zusammenhang existiert.

Die Idee zur Messung der Stärke des statistischen Zusammenhangs zweier nominaler Merkmale bedient sich genau dieser Tabellen der tatsächlich beobachteten und der bei Fehlen eines Zusammenhangs zwischen den beiden Merkmalen erwarteten relativen Häufigkeiten. Umso stärker der Zusammenhang ist, umso stärker müssen die beobachteten relativen Häufigkeiten von den bei Fehlen des Zusammenhangs zu erwartenden relativen Häufigkeiten abweichen. Wir bilden also zunächst die Abweichungen der einzelnen zweidimensionalen relativen Häufigkeiten von den bei Unabhängigkeit zu erwartenden, also $(0{,}22 - 0{,}24)$,$(0{,}24 - 0{,}216)$, $(0{,}04 - 0{,}06)$ und so weiter. Wenn man diese zehn Differenzen aus mathematischen Gründen auch noch quadriert, diese quadrierten Differenzen noch jeweils durch die dazugehörigen zu erwartenden relativen Häufigkeiten dividiert, die so erhaltenen Ergebnisse aufsummiert und diese Summe schließlich mit der Gesamtzahl der Erhebungseinheiten N multipliziert, dann erhalten wir eine häufig verwendete statistische Kennzahl. Es ist dies das Zusammenhangsmaß **Chiquadrat**χ^2 (χ … der griechische Buchstabe „Chi“).

Der Grund für die so komplexe Vorgehensweise liegt in der Möglichkeit, mit dieser auf diese Weise definierten Kennzahl den Zusammenhang zweier nominaler Merkmale in der schließenden Statistik testen zu können (siehe dazu ▶ *Abschnitt 3.6*).

In ▶ *Beispiel 1.13* erhalten wir als Zusammenhangsmaß:

$$\chi^2 = 500 \cdot \left[\frac{(0{,}22 - 0{,}24)^2}{0{,}24} + \frac{(0{,}24 - 0{,}216)^2}{0{,}216} + \frac{(0{,}04 - 0{,}06)^2}{0{,}06} + \ldots \right] = 18{,}06.$$

Bezeichnen wir mit p_{ij} die relativen Häufigkeiten der i-ten Zeile und j-ten Spalte einer solchen Tabelle (also ist zum Beispiel p_{11} die relative Häufigkeit, die in der ersten Zeile und ersten Spalte steht) und markieren wir die beobachteten relativen Häufigkeiten jeder Zelle der Tabelle zusätzlich mit dem Buchstaben b und die bei Unabhängigkeit der beiden Merkmale zu erwartenden relativen Häufigkeiten jeder Zelle mit e (sodass sich

etwa in unserem Beispiel ergibt: $p^b_{11} = 0{,}22$ und $p^e_{11} = 0{,}24$), dann wird die Vorgehensweise zur Berechnung von χ^2 formal darstellbar durch:

$$\chi^2 = N \cdot \sum \frac{\left(p^b_{ij} - p^e_{ij}\right)^2}{p^e_{ij}}. \tag{6}$$

Bei Vorliegen der Häufigkeiten anstelle der relativen Häufigkeiten müssen im Wesentlichen dieselben Rechenvorgänge durchgeführt werden, um zum gleichen Ergebnis zu kommen. Es sind dabei in (6) einfach die relativen Häufigkeiten p durch die Häufigkeiten h zu ersetzen. Jedoch ist am Schluss nicht mehr mit N zu multiplizieren, da dieser Umfang N der Grundgesamtheit in den Häufigkeiten schon enthalten ist.

χ^2 hat den Wert 0 bei Unabhängigkeit der Merkmale, denn dann ist ja die beobachtete Verteilung gleich mit der bei Unabhängigkeit zu erwartenden und die Häufigkeitsdifferenzen sind allesamt gleich null. Die Kennzahl χ^2 kann uns aber wenig Auskunft über die Stärke des statistischen Zusammenhangs geben, da sie noch nicht normiert, das heißt zwischen zwei Werten eingegrenzt ist. Dies kann erst das so genannte Cramersche Zusammenhangsmaß leisten, das häufig als **Cramers *V*** bezeichnet wird (und nicht mit dem Variationskoeffizienten v nach (5) verwechselt werden darf):

$$V = \sqrt[+]{\frac{\chi^2}{N \cdot (\min(s,t) - 1)}}. \tag{7}$$

(s,t … die Anzahlen der Merkmalsausprägungen der beiden Merkmale; min(s,t) … die kleinere der beiden Anzahlen.) Durch die Division von χ^2 durch N und das um eins verminderte Minimum der Anzahl der Merkmalsausprägungen s und t der beiden Merkmale erhalten wir eine Kennzahl, die zwischen 0 und 1 liegt und umso größer ist, umso stärker der Zusammenhang ist.

Der Wert von V beträgt in ▶ *Beispiel 1.13* wegen $s = 2$ (Anzahl der Ausprägungen des Merkmals Geschlecht) und $t = 5$ (Anzahl der Merkmalsausprägungen des Merkmals Studienrichtung) und somit $\min(s,t) = 2$:

$$V = \sqrt[+]{\frac{18{,}06}{500 \cdot (2-1)}} = 0{,}19.$$

Bei $V \approx 0$ schließen wir auf das Fehlen eines statistischen Zusammenhanges, da V nur dann 0 sein kann, wenn χ^2 null ist. Nun aber lässt sich endlich auch eine Aussage über die Stärke des statistischen Zusammenhangs machen. V liegt zwischen 0 und 1 und hat den Wert 1 nur bei einem **vollständigen Zusammenhang**. Dies käme zu Stande, wenn etwa alle weiblichen Befragten Soziologie und alle männlichen BWL studieren würden oder wenn die weiblichen nur die Ausprägungen Soziologie, Sozialwirtschaft oder Statistik und die männlichen nur BWL und VWL aufweisen würden, sodass man durch die Angabe des Geschlechts direkt auf die Ausprägungen des Merkmals Studienrichtung rückschließen könnte.

Je größer V ist, desto stärker ist der statistische Zusammenhang. Als willkürliche Faustregel zur verbalen Interpretation der Grade des Zusammenhangs sei angegeben, dass ein Wert von über 0 bis 0,2 auf einen **schwachen**, ein solcher zwischen 0,2 und 0,6 auf einen **mittleren** und ein Wert, der darüber und unter 1 liegt, auf einen **starken** statistischen Zusammenhang zwischen den beiden interessierenden Merkmalen schließen lässt. Diese Grenzen sind natürlich rein pragmatisch zu verstehen. An solche Festlegungen sind wir in unserem Leben gewohnt. Man denke etwa an die Grenzen für Alkohol am Steuer.

Eine knappe Unterschreitung bzw. Überschreitung dieser Promillegrenzen macht bezüglich der Fahrtüchtigkeit der betreffenden Person wohl kaum einen Unterschied, bei den Konsequenzen ist er jedoch immens (von Weiterfahren bis zur Entziehung der Lenkberechtigung). Irgendwo muss die Gesetzgebung wohl eine Grenze ziehen – eher willkürlich, aber eben pragmatisch. So sind auch die Grenzen für die verbale Interpretation von statistischen Zusammenhängen hier und im Folgenden zu verstehen.

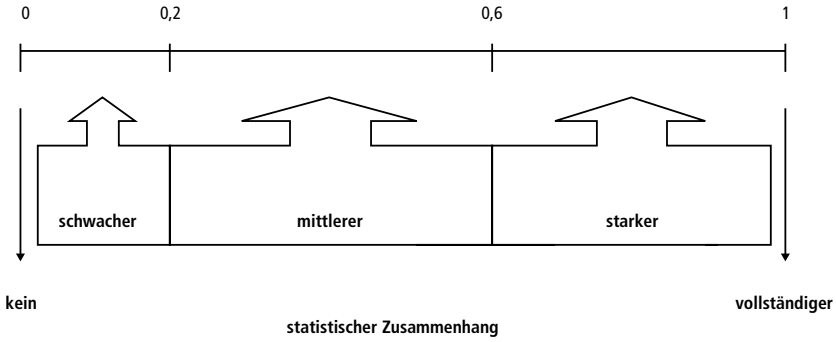

Abbildung 1.20 **Die Interpretation von Cramers *V* (Faustregeln)**

In ▶ *Beispiel 1.13* haben wir es demnach mit einem schwachen statistischen Zusammenhang zwischen den beiden Merkmalen zu tun.

Möchte man den statistischen Zusammenhang zwischen einem nominalen Merkmal wie Geschlecht und einem metrischen wie Einkommen messen, muss – wie eingangs zu diesem Abschnitt erwähnt – die Vorgehensweise für zwei nominale Merkmale gewählt werden. Um die dafür nötigen Tabellen der beobachteten und erwarteten relativen Häufigkeiten erstellen zu können, muss das metrische Merkmal in Kategorien eingeteilt werden. Auf diese Wiese könnten dann in ▶ *Beispiel 1.13* oben an Stelle der verschiedenen Studienrichtungen Einkommenskategorien wie „unter bis unter usf. stehen.

Metrische Merkmale

Bei metrischen Merkmalen ist die Situation grundlegend anders. Betrachten wir folgendes Beispiel, das uns die Idee und den sich daraus abgeleiteten Rechenvorgang bei der Messung des statistischen Zusammenhangs zweier metrischer Merkmale näher bringen soll.

Beispiel 1.14: Erhebung von zwei metrischen Merkmalen

In einem Betrieb arbeiten in einer Abteilung fünf Personen der gleichen Geschlechtskategorie (z. B. divers). An diesen wurden die Merkmale Alter (in vollendeten Lebensjahren) und Einkommen (in Euro) gemessen:

Person	Alter	Einkommen
A	21	1.850
B	46	2.500
C	55	2.560
D	35	2.230
E	28	1.800

Tabelle 1.16

Grafisch können diese Daten folgendermaßen dargestellt werden:

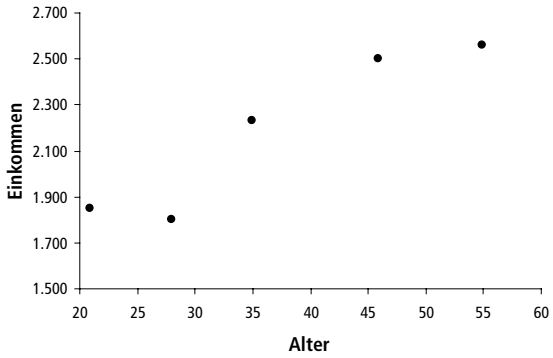

Abbildung 1.21 **Streudiagramm zweier metrischer Merkmale.** Grafische Darstellung der Daten aus ▶ *Beispiel 1.14*.

Diese Darstellung wird als **Streudiagramm** des zweidimensionalen Merkmals Alter und Einkommen bezeichnet. Betrachten wir das Diagramm, so gewinnt man den Eindruck, dass der Zusammenhang der beiden Merkmale solcherart ist, dass mit zunehmendem Alter auch das Einkommen steigt. Wie aber kann man dies durch eine Kennzahl zum Ausdruck bringen? Dazu betrachten wir die folgenden drei Streudiagramme:

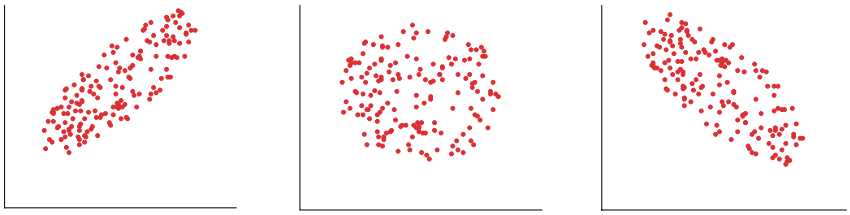

Abbildung 1.22 **Drei Streudiagramme für beliebige Merkmale *x* und *y*.** Richtung des statistischen Zusammenhangs zweier metrischer Merkmale an drei Beispielen.

Im linken Streudiagramm von ▶ *Abbildung 1.22* ist die Richtung des Zusammenhangs etwa so wie in ▶ *Abbildung 1.21*: Wächst x, so wächst tendenziell auch y. Einen solchen Zusammenhang nennt man gleichsinnig. In der Mitte ist gar keine Richtung feststellbar – x scheint mit y gar nicht zusammenzuhängen. Im rechten schließlich fällt y mit steigendem x. Dies ist ein gegensinniger Zusammenhang. Die Kennzahl, nach der wir suchen, soll uns diese Fälle unterscheiden helfen und auch Auskunft über die Stärke des Zusammenhangs geben!

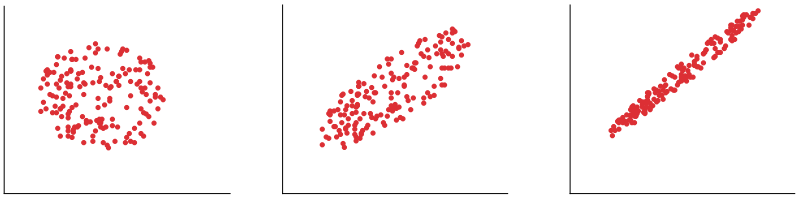

Abbildung 1.23 Drei Streudiagramme für beliebige Merkmale *x* und *y*. Stärke des statistischen Zusammenhangs zweier metrischer Merkmale an drei Beispielen.

Im linken Streudiagramm von ▶ *Abbildung 1.23* sieht es wie im mittleren Streudiagramm von ▶ *Abbildung 1.22* danach aus, dass die Merkmale x und y nicht zusammenhängen. In den beiden anderen Streudiagrammen von ▶ *Abbildung 1.23* lässt sich ein gleichsinniger Zusammenhang feststellen. Hinsichtlich seiner Stärke wächst der Zusammenhang offensichtlich von links nach rechts an.

Betrachten wir für die zu suchende Kennzahl folgende Idee: Als Erstes berechnet man für jede Erhebungseinheit i folgendes Produkt: $(x_i - \bar{x}) \cdot (y_i - \bar{y}) \cdot x_i$ und y_i bezeichnen die Merkmalsausprägungen der beiden Merkmale x und y bei der i-ten Erhebungseinheit. Wir bilden also die Differenzen der Merkmalsausprägungen der beiden Merkmale zum jeweiligen Mittelwert und multiplizieren diese Differenzen. Zur grafischen Darstellung der Bedeutung dieses Produktes betrachten wir ▶ *Abbildung 1.24*.

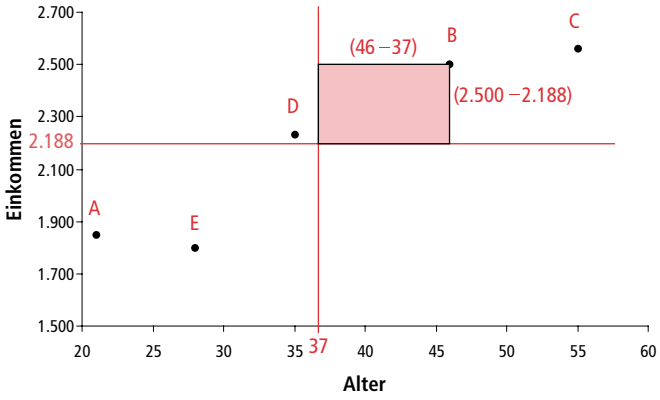

Abbildung 1.24 Grafische Darstellung der Idee zur Messung des Zusammenhangs zweier metrischer Merkmale. Verwendet werden die Daten aus ▶ *Beispiel 1.14*.

In ▶ *Abbildung 1.24* sind diese beiden Differenzen von den Mittelwerten 37 und 2.188 am Beispiel der Person *B* eingezeichnet. Multiplizieren wir diese Differenzen, so erhalten wir offenbar die Fläche des farbigen Rechtecks. Für Person *A* hat das Produkt dieser Differenzen ebenfalls ein positives Vorzeichen, da sowohl das Alter als auch das Einkommen unter dem Mittelwert liegen und das Produkt zweier negativer Zahlen positiv ist. Dies gilt ebenso für *E*. Für *C* gilt gleiches wie für *B*. Beide Differenzen sind positiv und somit auch das Produkt. Für Person *D* gilt aber, dass das Alter unter, das Einkommen aber über seinem Mittelwert liegt. Das Produkt der Differenzen zum jeweiligen Mittelwert ist somit negativ. Wenn solche Flächen sowohl positive als auch negative Werte aufweisen können, nennt man sie gerichtete Flächen.

Im nächsten Schritt addieren wir diese gerichteten Rechtecksflächen und dividieren sie durch die Anzahl. Die so berechnete Zahl nennt man die **Kovarianz** (*lat. cum = gemeinsam, variare = schwanken*) der Merkmale x und y und diese wird mit s_{xy} abgekürzt. Formal lässt sich das folgendermaßen darstellen:

$$s_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N}. \tag{8}$$

Vergleichen wir (8) mit Formel (2), so sehen wir, dass wir hier abermals einen Mittelwert berechnen, diesmal den der gerichteten Rechtecksflächen. Die Darstellungen (2a) und (2b) der Mittelwertsberechnung sind für die Kovarianzberechnung von untergeordneter Bedeutung, weil bei metrischen Merkmalen bestimmte Kombinationen von Ausprägungen der beiden betrachteten Merkmale oftmals nur einmal vorkommen und die Häufigkeit ihres gemeinsamen Auftretens somit gleich 1 und die relative Häufigkeit gleich $1/N$ ist.

Betrachten wir nun die drei Streudiagramme aus ▶ *Abbildung 1.22* hinsichtlich der dabei auftretenden Kovarianz: Denkt man sich die Mittelwerte von x und y wie in ▶ *Abbildung 1.24* eingezeichnet, so gilt für das erste Streudiagramm, dass bei der Berechnung der Kovarianz hauptsächlich positive Produkte (= positive gerichtete Rechtecksflächen) auftreten und die Kovarianz somit eine positive Zahl ist. Beim mittleren Streudiagramm werden sich die positiven und negativen Flächen ziemlich aufheben und die Kovarianz deshalb in der Nähe von null sein. Im dritten Streudiagramm schließlich werden die „negativen" Flächen überwiegen. Die Kovarianz wird deshalb negativ sein. Die Kovarianz ist somit eine zur Messung der Richtung des statistischen Zusammenhangs zweier metrischer Merkmale geeignete Kennzahl! Wenn sie einen negativen Wert aufweist, ist der Zusammenhang zwischen den Merkmalen gegensinnig, wenn sie einen positiven Wert aufweist gleichsinnig.

Eine Anforderung an eine Kennzahl zur Messung des Zusammenhanges ist aber auch, dass wir damit auch dessen Stärke bestimmen können. In einem Streudiagramm wie dem ersten in ▶ *Abbildung 1.23* werden sich (wie beim mittleren in ▶ *Abbildung 1.22*) die gerichteten Rechtecksflächen ziemlich aufheben und die Kovarianz nahe bei null liegen. Im daneben befindlichen Streudiagramm werden die positiven Rechtecksflächen die negativen überwiegen (wie im linken Streudiagramm von ▶ *Abbildung 1.22*) und die Kovarianz wird positiv sein. Im Streudiagramm ganz rechts schließlich werden die positiven Flächen die negativen noch deutlicher überwiegen und die Kovarianz wird deshalb größer sein als bei der Verteilung im mittleren Streudiagramm. Umso größer der Wert der Kovarianz bei gleichsinnigen Zusammenhängen also ist, desto größer ist der statistische Zusammenhang zwischen den beiden Merkmalen. Bei gegensinnigen statistischen Zusammenhängen überwiegen die negativen Rechtecksflächen und das eben Beschriebene gilt somit analog für negative Werte der Kovarianz. Die Kovarianz ist jedoch – ähnlich wie das Zusammenhangsmaß χ^2 aus Abschnitt „*Nominale Merkmale*" – nicht nach oben beziehungsweise unten beschränkt, sodass man aus ihr nicht sofort ablesen kann, wie stark der Zusammenhang ist.

Zur konkreten Bestimmung der Stärke des Zusammenhangs müssen wir die Kovarianz deshalb (wie das auch bei χ^2 der Fall war) noch normieren. Dies gelingt, wenn man sie durch das Produkt der beiden Standardabweichungen von x und y – wir bezeichnen sie nun zu ihrer Unterscheidung mit s_x und s_y – dividiert. Auf diese Weise erhält man

den **Korrelationskoeffizienten** von Bravais und Pearson, den wir mit dem Buchstaben r kennzeichnen. Formal lässt er sich also folgendermaßen darstellen:

$$r = \frac{s_{xy}}{s_x \cdot s_y}. \tag{9}$$

Der mögliche Wertebereich des Korrelationskoeffizienten umfasst das Intervall $[-1;+1]$. Diese Kennzahl besitzt (wie Cramers V) bei Unabhängigkeit der beiden Merkmale den Wert 0, weil dann die Kovarianz null ist. Das Vorzeichen von r wird durch das Vorzeichen der Kovarianz bestimmt, weil die Standardabweichungen jedenfalls positive Zahlen sind. Somit gibt uns, wie bei der Kovarianz, das Vorzeichen des Korrelationskoeffizienten die Richtung des Zusammenhanges an. Ein positives Vorzeichen bedeutet, dass der Zusammenhang **gleichsinnig** ist (wenn das Merkmal x zunimmt, dann auch das Merkmal y). Ist r negativ, so ist der Zusammenhang **gegensinnig** (wenn x zunimmt, dann nimmt y ab und umgekehrt).

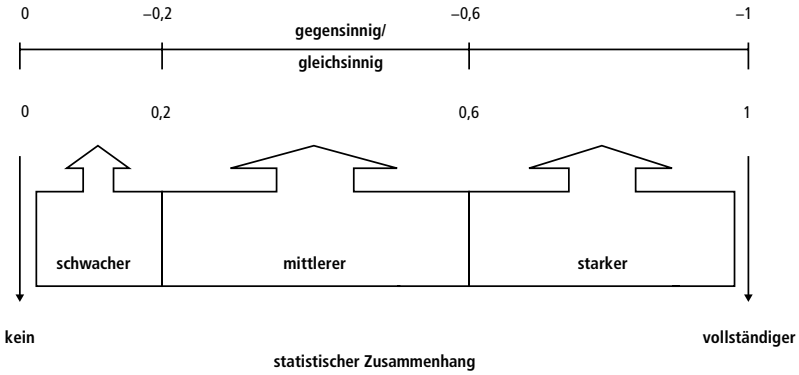

Abbildung 1.25 **Die Interpretation des Korrelationskoeffizienten (Faustregeln)**

Die Stärke des Zusammenhangs ist an der Entfernung des Wertes r von der Zahl null abzulesen. Umso größer der Betrag von r ist, umso stärker ist der Zusammenhang. Wie bei den willkürlichen Faustregeln zur verbalen Interpretation von Cramers V kann man bei einem Betrag von r bis etwa 0,2 von einem schwachen, bei einem solchen zwischen 0,2 und 0,6 von einem mittleren und bei einem über 0,6 von einem starken Zusammenhang sprechen. Ist $r = 1$ oder $r = -1$, so ist der Zusammenhang vollständig und das heißt hier linear, das heißt dass im Streudiagramm alle Punkte auf einer Geraden liegen. Denn genau genommen misst man mit dem Korrelationskoeffizienten natürlich nur den **linearen** statistischen Zusammenhang zweier Merkmale. Wenn mit Zunahme des Merkmals x das Merkmal y zuerst auch steigt, sich dies an einem gewissen Punkt jedoch umdreht und bei weiterer Zunahme von x das Merkmal y wieder sinkt, dann besitzen die beiden Merkmale natürlich auch einen Zusammenhang. Der Korrelationskoeffizient kann dabei jedoch durchaus null sein, weil kein linearer Zusammenhang vorliegt. Ein Beispiel für einen solchen Zusammenhang ist jener zwischen Düngermitteleinsatz und Ernteertrag in der Landwirtschaft.

Kommen wir zu ▶ *Beispiel 1.14* zurück: Die Kovarianz berechnet sich nun durch (8) mit

$$s_{xy} = \frac{(21-37)\cdot(1.850-2.188)+\ldots+(28-37)\cdot(1.800-2.188)}{5} = 3.664.$$

Die durchschnittliche gerichtete Rechtecksfläche hat den Wert +3664. Der Zusammenhang ist demnach gleichsinnig, wie schon ein Blick auf das Streudiagramm in ▶ *Abbildung 1.21* bestätigt.

Berechnen wir nun noch den Korrelationskoeffizienten nach (9). Dafür müssen wir noch die beiden Varianzen der Merkmale Alter und Einkommen mit (3) berechnen. Wir erhalten $s_x^2 = 149{,}2$ und $s_y^2 = 100.456$. Und somit ist

$$r = \frac{3.664}{\sqrt{149{,}2} \cdot \sqrt{100.456}} = 0{,}946.$$

Dies zeigt an, dass zwischen den beiden Merkmalen Alter und Einkommen aus ▶ *Beispiel 1.14* ein sehr starker, gleichsinniger (linearer) statistischer Zusammenhang existiert.

Bei der Interpretation des Ergebnisses eines Korrelationskoeffizienten ist wiederum zu beachten, dass man den statistischen Zusammenhang nicht automatisch als *kausal* bezeichnen darf. Das Alter selbst bestimmt natürlich nicht das Einkommen. Oftmals sind es etwa die Dienstjahre, die sowohl mit dem Alter als auch mit dem Einkommen zusammenhängen, wodurch auch Alter und Einkommen positiv korrelieren.

Im Januar 1987 sorgte eine Meldung für Aufsehen, die unter anderem auch in der oberösterreichischen Zeitung „Neues Volksblatt“ am 17.1.1987 erschienen ist: „Steirischer Arzt warnt: ‚Kat fördert AIDS‘.“ In dem Aufsatz wird berichtet, dass der steirische Mediziner Dr. Fritz Lautner in der Zeitschrift der Österreichischen Ärztekammer Meldungen aufgegriffen hatte, „wonach Katalysator-Autos möglicherweise die Verbreitung von Herpes, AIDS und bestimmter Krebsformen begünstigen könnten ... Man kann eine gewisse Korrelation zwischen der Einführung der Katalysatortechnik und den gehäuften AIDSFällen zum Beispiel in Los Angeles nicht von der Hand weisen!“

Und mit Letzterem hatte er völlig Recht! Die Merkmale Anzahl der monatlich neu registrierten AIDS-Fälle und Anzahl der monatlich neu produzierten Katalysator-Autos korrelierten in den Achtziger-Jahren wohl leicht positiv. In diesem Zeitraum ist sowohl die Anzahl der Kat-Autos wie auch die Anzahl der AIDS-Fälle ständig gestiegen. Deshalb gibt es, wenn man diese Zahlenreihen in Verbindung setzt, eine gleichsinnige Korrelation zwischen diesen beiden Merkmalen. Aber diese Korrelation liefert um Himmels willen noch keine Begründung! Auch die Anzahl der verkauften CDs oder PCs wuchs in diesem Zeitraum. Also fördern auch CDs und PCs genauso AIDS wie die Kat-Autos? Die Anzahl an Schallplatten ging zurück, ebenso die Bestzeit im 10.000-Meter-Lauf der Herren, und das Joggen erlebte einen Aufschwung. Die Anzahl der AIDS-Fälle korreliert somit gegensinnig mit der jährlichen Schallplattenproduktion und mit der Laufzeit der besten Leichtathleten sowie gleichsinnig mit den gelaufenen Jogging-Kilometern der Menschheit. Also sofort wieder Schallplatten produzieren (beziehungsweise langsamer laufen und weniger joggen)?

Der Korrelationskoeffizient liefert Auskunft über den Zusammenhang der Zahlen. Deswegen wird der Zusammenhang auch als *statistischer* Zusammenhang bezeichnet. Ob dieser auch ein *kausaler* ist, das muss vom jeweiligen Untersuchenden selbst eingeschätzt werden. Das gibt uns der Korrelationskoeffizient nicht an!

Mit dem **kausalen Zusammenhang** zwischen metrischen Merkmalen beschäftigt sich die **Regressionsrechnung**. Dabei soll über eine mathematische Funktion aus den Werten eines oder mehrerer unabhängiger Merkmale, der **Regressoren**, der Wert eines von

Copyright