4 Merkmalszusammenhänge

In diesem Kapitel geht es um die Beziehung zwischen zwei Zufallsvariablen. Wir sprechen dabei auch von **bivariaten** Daten[1], d.h. dass wir gleichzeitig *zwei* Merkmale untersuchen. Wir wollen dabei herausfinden, ob oder wie stark die beiden Zufallsvariablen einander beeinflussen. Wenn uns das gelingt, können wir für einige Phänomene der Wirklichkeit – zumindest statistisch – erklären, warum sie sich im Ergebnis unterscheiden, in manchen Fällen sogar in gewisser Weise Vorhersagen treffen. Nicht 100%ig perfekte Vorhersagen, aber immerhin. Umgekehrt können wir auch mitunter zeigen, dass an manchen scheinbaren Zusammenhängen nichts dran ist und nur einer getäuschten Intention (oder unserem Wunschdenken) entspricht. Gesucht sind letztlich Art und Stärke des Zusammenhangs.

Manchmal unterscheiden wir dabei in eine **Zielvariable** (auch: *interessierende* Variable, eng.: *response variable*) und eine **Einflussvariable** (auch: *erklärende* Variable, eng.: *explanatory variable*). Die Zielvariable lässt sich dabei aus der Einflussvariablen ableiten. Manchmal ist es aber auch so, dass es zwar einen Zusammenhang gibt, aber beide Variablen gleichberechtigt sind. Wir können also nicht immer genau sagen, welches die Ziel- und welches die Einflussvariable ist, oder ob sie nicht zum Beispiel beide von einer dritten beeinflusst werden.

Mathematisch geben wir die Beziehung zwischen zwei Zufallsvariablen an, indem wir eine Variable mehr oder weniger als Funktion der anderen darzustellen versuchen. *«Mehr oder weniger»* bedeutet dabei, dass es nicht um eine strenge Funktion im mathematischen Sinn geht (siehe Abb.4.1).

[1] vom. lat. *bis* = zweimal und *variare* = (sich) verändern. Bei der gleichzeitigen Betrachtung von mehr als zwei Zufallsvariablen sprechen wir von *multivariaten* Verfahren, betrachten wir jeweils nur einzelne Variable (wie in den Kapiteln bisher), von *univariaten*.

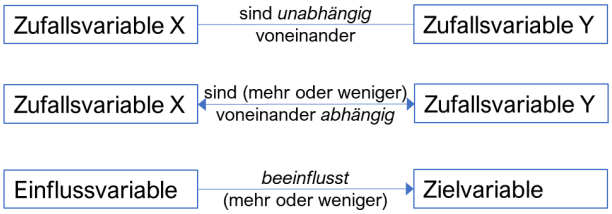

Abb. 4.1: Zufallsvariable können unterschiedlich zusammenhängen

4.1 Streu- und Bubblediagramme

Ein Beispiel: Sehen wir uns zunächst ein einfaches Beispiel an: Tabelle 4.1 zeigt das Ergebnis der Messung von Größe und Gewicht zwanzig zufällig ausgewählter Personen:

X Größe [cm]	Y Gewicht [kg]	X Größe [cm]	Y Gewicht [kg]
188	83	170	68
183	88	187	92
183	81	177	85
185	85	178	78
178	70	180	75
198	94	182	75
163	55	189	88
164	57	173	68
174	80	176	77
185	78	177	78

Tabelle 4.1: Größe und Gewicht 20 zufällig ausgewählter Personen

Wir können nun die beiden Zufallsgrößen Größe und Gewicht *gemeinsam* betrachten und in einem **Streudiagramm** (auch: *Punktdiagramm* oder *«Punktwolke»)* darstellen (Abb.4.2). Dazu stellen wir die beiden Variablen X und Y in einem Koordinatensystem dar und zeichnen für jeden Merkmalsträger einen Punkt an den Koordinaten (X,Y) ein.

Jeder Punkt im Streudiagramm repräsentiert somit Informationen über die Kombination aus zwei Merkmalen. Aus dem Diagramm können wir in weiterer Folge gut eventuelle «Muster» in unseren Daten visuell ablesen und Trends und augenscheinliche Zusammenhänge (und auch: Nicht-Zusammenhänge) erkennen. Möglich ist die Visualisierung in einem Punktdiagramm in der «klassischen» Form aber nur für metrische, unklassierte Daten.

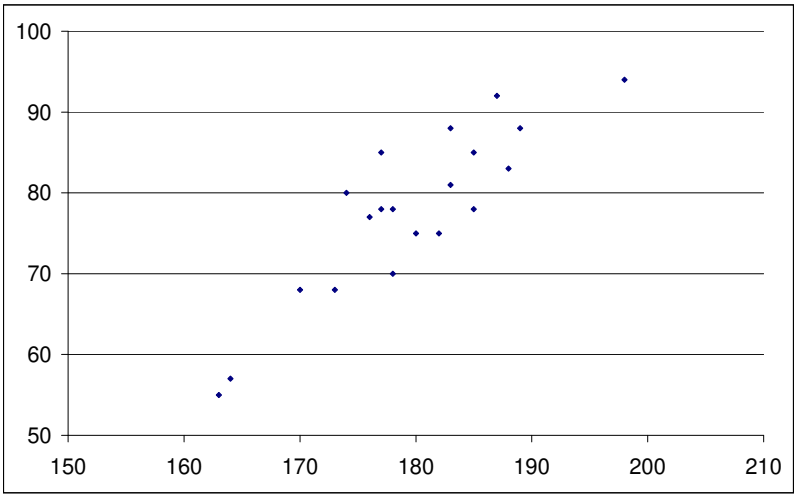

Abb. 4.2: Streudiagramm zu den Daten aus Tab.4.1

Bubble-Diagramme (*Blasendiagramme*) werden ähnlich wie Streudiagramme dafür verwendet, Zusammenhänge zwischen Zufallsgrößen zu visualisieren. Dabei werden zunächst zwei Merkmale in einem Streudiagramm eingezeichnet. Anstelle von einfachen, gleich großen Punkten verwendet man aber Punkte mit unterschiedlichen Durchmessern – dadurch werden aus den Punkten «Blasen» (eng. *bubble*).

Damit ist es möglich, noch eine dritte Variable und somit zusätzliche Information im Diagramm darzustellen. Verwendet man darüber hinaus auch noch unterschiedliche Farben für die Blasen, kann man auch noch ein viertes Merkmal in die grafische Darstellung hineinpacken.

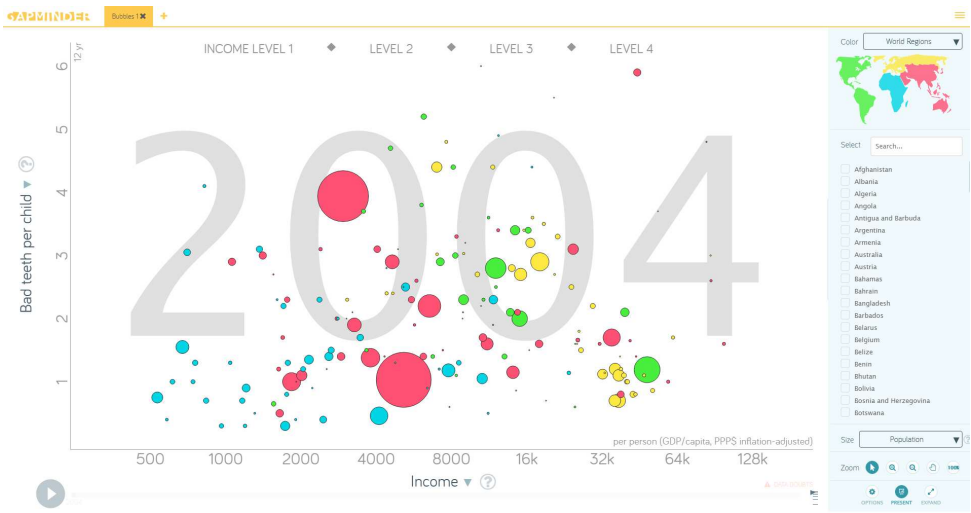

Abb. 4.3: Ein Bubblediagramm (Quelle: www.gapminder.org)

Abbildung 4.3 zeigt ein Beispiel dazu. Auf der x-Achse ist länderweise das «Bruttoinlandsprodukt pro Einwohner» eingezeichnet, auf der y-Achse der Mittelwert der «Anzahl an schlechten Zähnen», die ein zwölfjähriges Kind in diesen Ländern hat. Zusätzlich ist über die Größe der Blasen die Größe des jeweiligen Länder repräsentiert und über die Farbgebung die Zugehörigkeit zu einer bestimmten Region.

4.2 Regressionsrechnung

In der Physik beschreiben wir Zusammenhänge durch Formeln. Zum Beispiel ist beim Autofahren der Zusammenhang zwischen dem Anhalteweg s und der gefahrenen Geschwindigkeit v, der Reaktionszeit t und der Bremsverzögerung a gegeben durch:

$$s = t \cdot v + \frac{v^2}{2a}$$

So exakt die Formel auch aussehen mag: Nur wenn wir exakte Werte für t, a und v kennen, erhalten wir einen exakten Wert für s. Meist «schätzen» wir das Ergebnis, indem wir für $t = 0.8\,s$ Reaktionszeit und $a = 8.0\,m/s^2$ Bremsverzögerung einsetzen. Damit ist zum Beispiel bei einer Geschwindigkeit von $v = 50\,km/h$ der Anhalteweg $s = 23\,m$ lang, bei $v = 130\,km/h$ ist er $110\,m$ lang, etc.

Andere Zusammenhänge lassen sich zwar auch eindeutig abbilden, allerdings nicht immer durch eine in eine mathematische Formel gegossene Funktion. Zum Beispiel wird jedem Platz in einem Ski-Weltcuprennen ein eindeutiger Punktegewinn zugeschrieben. Abb. 4.4 zeigt diesen Zusammenhang.

Und dann gibt es Beispiele von Merkmalszusammenhängen, die sich eben nicht auf mathematische Funktionen oder eindeutige Zuordnungen zurückführen lassen, sondern eher *statistischer* Natur sind. Dazu betrachten wir noch einmal die beiden Abbildungen 4.2 und 4.3. Aus dem Bubblediagramm (Abb.4.3) lässt sich kein Zusammenhang zwischen den beiden Zufallsgrößen ableiten[2]. In Abb.4.2 hingegen können wir augenscheinlich feststellen, dass mit zunehmen- dem X auch die Variable Y tendenziell zunimmt. Das legt den Schluss nahe, dass sich das Körpergewicht aus der Körpergröße erklären lässt[3]. Dieser Zusammenhang ist natürlich kein streng *deterministischer*, d.h. es gibt kein naturwissenschaftliches Gesetz oder Funktion, nach dem man aus der Körpergröße das exakte

[2] Ob ein solcher überhaupt zu erwarten gewesen wäre, wollen wir hier nicht weiter erörtern...

[3] Zumindest teilweise. Wir wissen, dass die Größe nur eine Variable ist, die das Gewicht beeinflusst und noch andere Parameter eine Rolle spielen. Aber in dieser einfachen statistischen Untersuchung betrachten wir nur den Zusammenhang bivariater Daten.

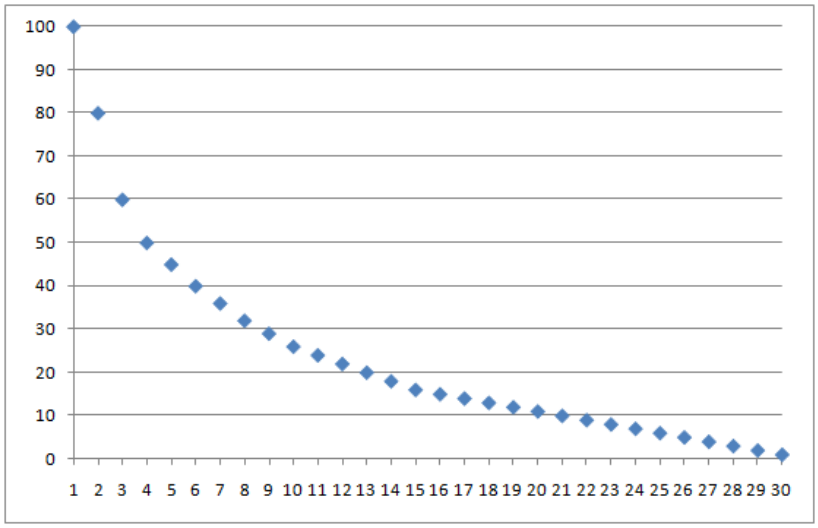

Abb. 4.4: Punktezuteilung zur erreichten Platzierung in einem Ski-Worldcuprennen

Gewicht errechnen kann. Es gibt aber einen *tendenziellen* Zusammenhang; wir nennen das auch einen *statistischen* bzw. einen *stochastischen* Zusammenhang. Den Beinamen «stochastisch» erhält er, weil wir ihn immer nur mit einer gewissen *Unschärfe* angeben können[4]. Aufgabe der **Regressionsrechnung** ist es, die Art des stochastischen Zusammenhangs zu beschreiben.

Zunächst einmal können wir in Abb.4.2 ein bestimmtes Muster erkennen, das von links unten nach rechts oben verläuft. Niedrigen Werten auf der x-Achse entsprechen niedrige Werte auf der y-Achse; steigt der x-Wert, dann steigt auch der y-Wert. Wir sprechen in diesem Fall von einem *positiven* Zusammenhang. Andernfalls – wenn das Muster also von links oben nach rechts unten läuft und niedrige Werte auf der x-Achse mit hohen Werten auf der y-Achse korrespondieren (und umgekehrt) – von einem *negativen*. Es kann natürlich auch sein, dass wir wirklich im wahrsten Sinn des Wortes einen Punkt-*Haufen* vor uns haben und zunächst einmal überhaupt kein nennenswerter Zusammenhang oder Muster erkennbar ist. Diese drei grundsätzlichen Möglichkeiten (positiver, negativer und kein Zusammenhang) sind in Abb.4.5 dargestellt.

Die nächste Frage, die wir uns stellen, ist: Von welchem Typ könnte eine Funktion sein, die wir in die Punktwolke hineinlegen können, und die als charakteristischer Repräsentant der Punktwolke gelten kann?

Prinzipiell unterscheiden wir dabei zwischen *linearen* und *nicht-linearen* Funktionen. Lineare Funktionen (z.B. Gerade) sind einfacher zu handhaben; nichtlineare Regressionszusammenhänge benötigen kompliziertere Funktionen. Wir

[4]Zum Wort *stochastisch* siehe Fußnote 6 auf Seite 55.

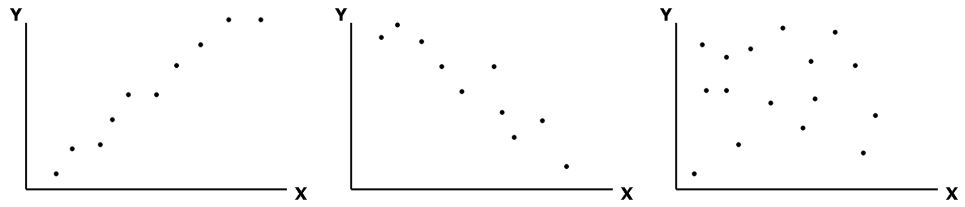

Abb. 4.5: Streudiagramme mit verschiedenen Mustern (positiv, negativ und «zusammenhangslos»)

werden uns in diesem Kurs auf lineare Zusammenhänge beschränken, also solche, die sich mit Geraden darstellen lassen – die so genannte *Regressionsgerade*. Das ist jene Gerade, die einen Punkthaufen wie jenen in Abb.4.2 «am besten» repräsentiert. Wie können wir die Parameter dieser Regressionsgeraden bestimmen?

Die Regressionsgerade

Eine Gerade (und ihre Gleichung) ist – wie wir aus der Mathematik wissen – durch zwei Parameter eindeutig bestimmt: den *Anstieg* der Geraden und den *Achsenabschnitt* auf der y-Achse (= die «Verschiebung» entlang der y-Achse relativ zum Ursprung des Koordinatensystems). Die Geradengleichung heißt dann:

$$y = kx + d \tag{4.1}$$

Für die Regressionsgerade gehen wir so vor: Zunächst berechnet man für jede Zufallsvariable den jeweiligen Mittelwert sowie die Varianz der Zufallsgröße X:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4.2}$$

und anschließend eine weitere Größe, die wir mit s_{xy} bezeichnen:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \tag{4.3}$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}\right) \tag{4.4}$$

Dann erhält man die Parameter der Regressionsgeraden aus

$$k = \frac{s_{xy}}{s_x^2} \qquad d = \bar{y} - k\bar{x} \tag{4.5}$$

Der Achsenabschnitt d der Regressionsgeraden wird auch als **Niveaufaktor** bezeichnet.

Der Anstieg k der Regressionsgeraden wird auch als **Regressionskoeffizient** bezeichnet. Er kann positiv oder negativ sein und dementsprechend sprechen wir von *positiver* bzw. *negativer linearer Regression*

In *MS Excel* und *LibreOffice Calc* können wir den Anstieg k der Regressionsgeraden mit dem Befehl `=STEIGUNG(Y-Werte; X-Werte)` berechnen, den Achsenabschnitt d mit `=ACHSENABSCHNITT(Y-Werte;X-Werte)`.
In *R* lautet der Befehl `lm(Y~X)`.

Beispiel 17 *Für unser Eingangsbeispiel erhalten wir:*

$$k = 1.08 \qquad d = -116.10$$

was wir auch gleich grafisch umsetzen können und in das Streudiagramm 4.2 die Regressionsgerade einzeichnen (Abb.4.6).

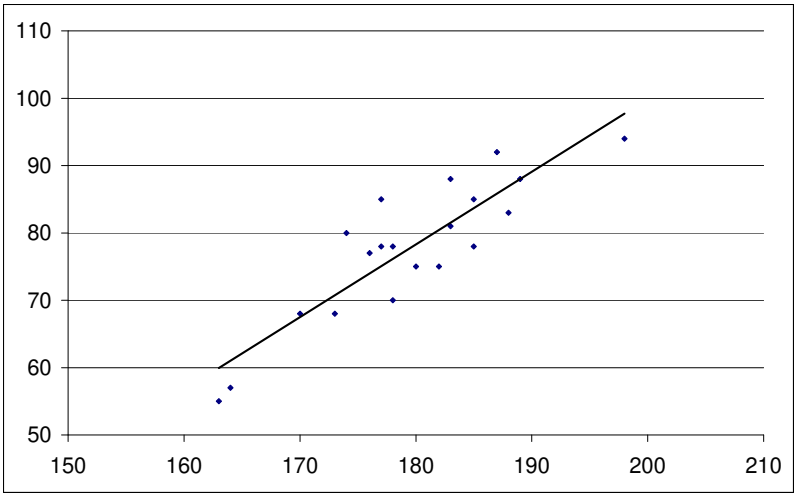

Abb. 4.6: Regressionsgerade zu den Daten aus Tab.4.1

Mit Hilfe der Regressionsgeraden sind durch einfaches Einsetzen nun auch Prognosen für nicht empirisch bestimmte Merkmalsausprägungen möglich. Wir können zum Beispiel angeben, welches Körpergewicht für einen Erwachsenen mit einer Körpergröße von 196 *cm* statistisch zu erwarten ist, nämlich:

$$y = kx + d = 1.08 \cdot 196 - 116.10 = 96\,kg$$

Achtung: Die «Vorhersage», die wir eben über eine 196 *cm* große Person getroffen haben, ist mathematisch gesehen eine **Interpolation**, d.h. wir haben für x einen Wert angegeben, der innerhalb des Wertebereiches liegt, mit dem wir die Parameter k und d berechnet haben. (Der kleinste Wert war 163, der größte 198). Dem gegenüber liegt eine **Extrapolation** vor, wenn wir für x einen Wert einsetzen, der außerhalb dieses Wertebereichs liegt.
Extrapolationen sind immer mit Vorsicht zu genießen. Setzt man in unserem Beispiel für $x = 100$ ein, käme für y ein negativer Wert heraus ($y = 1.08 \cdot 100 - 116.10 = -8.1$). Offensichtlich kann aber selbst ein Kind mit einer Körpergröße von 1 *m* kein negatives Körpergewicht haben. . . .

Aufgabe 10 In Tabelle 4.2 sind für sieben in der Vergangenheit in Wien abgehaltene Wahlen die Mittagstemperatur am jeweiligen Wahltag (x) sowie das Verhältnis der abgegebenen Stimmen zur Anzahl der Wahlberechtigten, also die Wahlbeteiligung (y) gegeben. Gib den Regressionskoeffizienten an.

	28.09.08	07.06.09	10.10.10	29.09.13	25.05.14	11.10.15	24.04.16
x (Temperatur °C)	15	22	12	12	24	7	7
y (Wahlbeteiligung)	0.74	0.43	0.68	0.70	0.35	0.75	0.64

Tabelle 4.2: «Wahltemperatur» und Wahlbeteiligung in Wien

Aufgabe 11 (Fortsetzung zu Aufgabe 10): In obiger Tabelle ist die Bundespräsidentenwahl 2010 nicht enthalten. Die Mittagstemperatur am Wahltag (25.4.2010) betrug 18°. Welche Wahlbeteiligung war bei dieser Temperatur zu erwarten?

Der Weg zur Mittelmäßigkeit

An dieser Stelle noch ein weiterer Hinweis: Das Wort *Regression*[5] ist an sich keine sehr aussagekräftige Bezeichnung für diese Methode; sie wurde von ihrem Erfinder, *Francis Galton*[6], auf Grund eines einzigen Beispiels geprägt: Galton, ein Cousin von Charles Darwin, versuchte, die Evolutionstheorie seines Cousins

[5] vom lat. *regredior* = zurückgehen

[6] Sir Francis Galton, 1822-1911, englischer Arzt und Biologe. Er verfasste zahlreiche Arbeiten über Anthropologie und Vererbung und sammelte dazu Daten über verschiedene Merkmalsausprägungen der Menschen. Anschließend entwickelte er statistische Methoden zu ihrer Auswertung.

durch quantitative Beispiele zu untermauern. In einer großangelegten experimentellen Studie untersuchte er, ob es eine Beziehung zwischen der Körpergröße der Eltern und der ihrer Kinder gibt. Er fand heraus, dass zwar große Eltern tendenziell auch große Kinder haben und kleine Eltern kleine Kinder, allerdings in der Weise, dass die Kinder großer Eltern eher kleiner sind als ihre Eltern und umgekehrt. Eltern haben also meistens Kinder, deren Größe näher am Durchschnitt liegt als ihre eigene Größe. Er nannte diesen Zusammenhang «*regression* to mediocrity» – den «Rückschritt zum Mittelmaß»[7].

4.3 Korrelationsrechnung

Die Regressionsgerade beschreibt zwar die *Art* des statistischen Zusammenhangs, sagt aber nichts über seine *Stärke* aus. Wir werden umso ungenauere Prognosen abgeben, je geringer der statistische Zusammenhang der beiden Variablen ist. Eine Regressionsgerade lässt sich nach obigen Formeln ja in jedem Fall berechnen, auch wenn so gut wie kein Zusammenhang vorliegt. Die Frage ist aber, wie eng oder weit die Punktwolke um die erhaltene Regressionsgerade streut. Dies beantwortet die **Korrelationsrechnung**.

Dazu rufen wir uns zunächst die Größe in Erinnerung, die wir in Formel (4.4) auf Seite 80 verwendet haben. Es ist dies die

Kovarianz

Zwischen je zwei statistischen Variablen X und Y können wir einen Parameter für die «gemeinsame Streuung» angeben, genannt die «*Kovarianz* von X und Y». Sie lässt sich mit Hilfe der beiden Mittelwerte $\bar{x}$ und $\bar{y}$ ausrechnen:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) \tag{4.6}$$

Die Kovarianz ist also das *mittlere Abweichungsprodukt* und ist ein Maß für den wechselseitigen Zusammenhang der beiden Zufallsgrößen X und Y.

Ist die Kovarianz positiv, so sind die Zufallsgrößen X und Y tendenziell eher gleich, d.h. mit großer Wahrscheinlichkeit nimmt die eine zu, wenn auch die andere zunimmt, beziehungsweise ab, wenn die andere abnimmt.

[7] Galton, Francis. *Regression Towards Mediocrity in Hereditary Stature.* The Journal of the Anthropological Institute of Great Britain and Ireland 15 (1886): 246–63. archive.org/details/journalroyalant15irelgoog/page/244/mode/1up

Ist die Kovarianz hingegen negativ, verhalten sich die Zufallsgrößen tendenziell eher reziprok, d.h. mit großer Wahrscheinlichkeit nimmt die eine ab, wenn die andere zunimmt, beziehungsweise zu, wenn die andere abnimmt.

Zufallsgrößen, deren Kovarianz gleich Null ist, bezeichnen wir als *statistisch unabhängig* voneinander.

Der Korrelationskoeffizient

Der Wert der Kovarianz ist abhängig von der Dimension der beiden Zufallsgrößen X und Y. Beschreibt zum Beispiel X die Körpergröße und Y das Gewicht, so ist der Wert von s_{xy} unterschiedlich, je nachdem ob die Größe in *cm* oder *m* angegeben wird bzw. das Gewicht in *dag* oder *kg*. Das ist nicht besonders praktisch. Die Kovarianzen können aber *normiert* werden, indem sie durch die jeweiligen Standardabweichungen dividiert werden. Damit schafft man ein dimensionsloses Maß. Der entsprechende Quotient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \tag{4.7}$$

ist der **Korrelationskoeffizient**, genauer auch: der *Pearson-Korrelationskoeffizient* oder manchmal auch die *Produkt-Moment-Korrelation nach Bravais und Pearson*[8] genannt. Er ist ein Maß für den *linearen* statistischen Zusammenhang zwischen zwei Zufallsvariablen.

Zur Erinnerung: s_{xy} ist die Kovarianz – vgl. Formel (4.6), s_x die Standardabweichung der Variablen X und s_y die Standardabweichung der Variablen Y – beide werden mit der Formel (3.21) bzw. (3.22) ausgerechnet. Es gilt:

$$-1 \leq r \leq 1 \tag{4.8}$$

d.h. dass der Korrelationskoeffizient nie kleiner als minus Eins und nie größer als plus Eins werden kann, ganz egal, wie groß X oder Y sind.

Eine positive Korrelation bedeutet, dass eine Vergrößerung der Werte der einen Zufallsgröße auch eine Vergrößerung der Werte der anderen Zufallsgröße zur Folge hat, bzw. eine Verkleinerung der einen Zufallsgröße eine Verkleinerung der anderen Zufallsgröße. Eine negative Korrelation hingegen bedeutet, dass eine Vergrößerung der Werte der einen Zufallsgröße eine Verkleinerung der Werte der anderen Zufallsgröße bewirkt und vice versa.

[8] *Karl Pearson*, 1857 - 1936, britischer Mathematiker und Statistiker; wir haben ihn bereits beim Histogramm auf Seite 40 kennengelernt. *Auguste Bravais*, 1811 - 1863, französischer Astronom und Physiker, hatte schon eine Zeit vor Pearson grundlegende theoretische Überlegungen zur Korrelationsrechnung veröffentlicht.

Ein Korrelationskoeffizient von *exakt* $+1.0$ oder -1.0 bedeutet, dass nicht nur ein statistischer linearer Zusammenhang besteht, sondern die Punkte tatsächlich auch mathematisch auf einer Geraden liegen und die Veränderungen streng äquivalent erfolgen. Alle anderen Werte von r lassen auf einen mehr oder weniger starken linearen Zusammenhang schließen. Ganz grob könnten wir sagen: Ein Korrelationkoeffizent bis etwa 0.3 bedeutet einen kleinen statistischen Zusammenhang. bei einem (Absolut-)Wert des Korrelationskoeffzienten von mindestens 0.3 und maximal 0.5, sprechen wir von einer mittleren Korrelation und ab 0,5 von einem großen Zusammenhang. Etwas feiner granuliert ergibt sich eine Zuordnung wie sie zum Beispiel in Tabelle4.3 vorgeschlagen wird.

Korrelationskoeffizient	Bedeutung
$r = -1$	vollständige lineare Abhängigkeit
$-1 < r \leq -0.8$	starker negativer linearer Zusammenhang
$-0.8 < r \leq -0.6$	mäßig starker negativer linearer Zusammenhang
$-0.6 < r \leq -0.4$	mittlerer negativer linearer Zusammenhang
$-0.4 < r \leq -0.2$	geringer negativer linearer Zusammenhang
$-0.2 < r < 0$	sehr schwache Korrelation
$r = 0$	keine lineare statistische Abhängigkeit
$0 < r < 0.2$	sehr schwache Korrelation
$0.2 \leq r < 0.4$	geringer positiver linearer Zusammenhang
$0.4 \leq r < 0.6$	mittlerer positiver linearer Zusammenhang
$0.6 \leq r < 0.8$	mäßig starker positiver linearer Zusammenhang
$0.8 \leq r < 1$	starker positiver linearer Zusammenhang
$r = 1$	vollständige lineare Abhängigkeit

Tabelle 4.3: Aus dem Korrelationskoeffizienten lässt sich die Stärke des linearen Zusammenhangs ablesen.

In *MS Excel* und *LibreOffice Calc* erhalten wir den Korrelationskoeffizienten mit `=KORREL(Y-Werte; X-Werte)`, wobei es – im Gegensatz zu Steigung und Achsenabschnitt der Regressionsgeraden – nicht darauf ankommt, welche Zufallsgröße als Y-Werte und welche als X-Werte bezeichnet werden. In *R* lautet der Befehl zur Berechnung des Korrelationskoeffizienten `cor(Y,X)`.

Beispiel 18 *Berechne zu den Daten aus Tab. 4.1 den Korrelationskoeffizienten und interpretiere den erhaltenen Wert.*

Zur Berechnung verwenden wir MS Excel und erhalten: $r_{xy} = 0.88$.

Das ist ein positiver Wert, was darauf hindeutet, dass eine Vergrößerung der Werte

der einen Zufallsgröße tendenziell auch eine Vergrößerung der Werte der anderen Zufallsgröße zur Folge hat (Was nicht weiter überraschend ist: Je größer jemand ist, desto schwerer ist er oder sie im Allgemeinen auch . . .).
Der konkrete Wert von 0.88 ist zudem relativ groß und lässt auf einen starken linearen statistischen Zusammenhang zwischen Größe und Gewicht schließen.

Übrigens geben wir üblicherweise von einem Korrelationskoeffizienten nicht mehr als zwei Nachkommastellen an – natürlich unter Beachtung üblicher Rundungsregeln.

Aus Formel (4.7) kann man erkennen, dass für den Korrelationskoeffizienten – im Gegensatz zur Regression – eine Unterscheidung in eine *Ziel-* und eine *Einflussvariable* nicht möglich ist. Es spielt keine Rolle, was wir als X und was als Y bezeichnen – die Formel ist bezüglich X und Y symmetrisch. Genauer müssen wir daher sagen: Die Regression beschreibt die Abhängigkeit einer Zufallsvariablen von einer anderen, der Korrelationskoeffizient die Stärke der *wechselseitigen* (linearen) Abhängigkeit.

An dieser Stelle noch ein Hinweis auf die Berechnung des Korrelationskoeffizienten, wenn wir X und Y in Form von standardisierten Messwerten vorliegen haben, also die «z-Werte» z_x und z_y (jeweils berechnet nach Formel 3.26 auf Seite 74): Der Korrelationskoeffizient kann dann auch berechnet werden aus

$$r = \frac{\sum z_x \cdot z_y}{n-1} \tag{4.9}$$

Aufgabe 12 In welchem mathematischen Zusammenhang stehen der Korrelationskoeffizient r_{xy} und der Regressionskoeffizient k? (Hinweis: Gib eine mathematische Gleichung an, die sowohl r_{xy} als auch k enthält)

Aufgabe 13 Besteht zwischen der in Tab.4.2 gegebenen Wahlbeteiligung der Wienerinnen und Wiener und der am Wahltag vorherrschenden Temperatur eine hohe Korrelation?
Wie groß ist der Korrelationskoeffizient?

Am Beginn dieses Abschnitts haben wir geschrieben: *«Eine Regressionsgerade lässt sich [. . .] in jedem Fall berechnen, auch wenn so gut wie kein Zusammenhang vorliegt»*. Wir können nun umgekehrt näher ausführen: Nur wenn es sich über den Korrelationskoeffizienten zeigen lässt, dass ein linearer statistischer Zusammenhang vorliegt, macht es Sinn, eine Regressionsgerade aufzustellen.

Und: Auf Seite 68 haben wir darauf hingewiesen, dass der Mittelwert ziemlich anfällig auf Ausreißer reagiert. An dieser Stelle müssen wir ergänzen: Das

gilt auch für den Korrelationskoeffizienten, insbesondere für kleine Stichproben: Ausreißer können nicht vorhandene lineare Zusammenhänge vorgaukeln – oder vorhandene verschleiern.

Der Determinationskoeffizient

Wenn wir den Korrelationskoeffizienten r quadrieren, erhalten wir den so genannten **Determinationskoeffizienten** (auch: *Bestimmtheitsmaß*):

$$r_{xy}^2 = \left(\frac{s_{xy}}{s_x s_y} \right)^2 \tag{4.10}$$

Er wird in der Regel in Prozent angegeben (d.h. mit 100 multipliziert und mit einem %-Zeichen versehen) und man kann ihn folgendermaßen deuten:

Der Determinationskoeffizient gibt an, zu wieviel Prozent sich eine Änderung der einen Zufallsvariable durch eine Änderung der anderen Zufallsvariable erklären lässt, also zu wieviel Prozent die eine die andere Zufallsvariable «determiniert» (Daher auch der Name).

Beispiel 19 *Berechne zu den Daten aus Tab. 4.1 den Determinationskoeffizienten und interpretiere den erhaltenen Wert.*

Nachdem wir im letzten Beispiel bereits den Korrelationskoeffizienten berechnet haben, brauchen wir ihn jetzt nur noch quadrieren (wobei wir den auf vier Stellen gerundeten Wert 0.8809 *verwenden):*

$$r^2 = 0.8809^2 = 0.7760 \approx 78\%$$

D.h. zu etwa 78% lässt sich das Gewicht durch den linearen Zusammenhang zwischen Körpergröße und Körpergewicht ableiten.

In *MS Excel* und *LibreOffice Calc* erhalten wir den Determinationskoeffizienten mit `=BESTIMMTHEITSMASS(Y-Werte; X-Werte)`
In *R* gibt es keinen Befehl für den Determinationskoeffizienten alleine. Mit dem Befehl `summary(lm(Y X))` werden aber alle möglichen Regressionsergebnisse angezeigt, darunter auch der Wert `Multiple R-squared`. Das ist das Bestimmtheitsmaß.

Formel 4.6 – uns somit Formel 4.7 – lassen sich nur anwenden, wenn sich auch die arithmetischen Mittelwerte $\bar{x}$ und $\bar{y}$ berechnen lassen. Das geht aber laut Tab.3.6 nur für metrische Merkmalswerte auf einer Intervall- oder Rationalskala. Wenn wir eine Korrelation zwischen zwei Zufallsgrößen angeben wollen, von denen eine oder beide «nur» ordinalskaliert sind, müssen wir die **Rangkorrelation** verwenden.

Für die Untersuchung des Zusammenhangs zweier Rangmerkmale müssen wir zunächst schauen, ob es auch «ex aequo-Plätze» gibt, ob es also Werte gibt, die mehrfach auftreten. Ist das nicht der Fall, ist die Berechnung sehr einfach: In Formel 4.6 werden einfach für x_i und y_i (und $\bar{x}$ und $\bar{y}$) die Rangplätze eingesetzt und dann nach 4.7 der Korrelationskoeffizient berechnet. Wir können diese einfache Vorgangsweise in der Praxis auch anwenden, wenn es nur einige wenige Mehrfachvorkommen gibt. (Sie hat nämlich den großen Vorteil, dass es eine Excel-Funktion gibt, in die sich direkt einsetzen lässt).

Gibt es hingegen mehr als «einige wenige» ex aequo-Ränge, dann müssen wir anstelle von Formel 4.7 den **Rangkorrelationskoeffizient nach Spearman**[9] verwenden:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n^3 - n} \tag{4.11}$$

wobei die D_i die Differenzen zwischen den beiden Rangzahlen des i-ten Elements sind.

Ist eine Variable ordinalskaliert, die andere aber rationalskaliert, muss vor der Berechnung des Korrelationskoeffizienten auch die rationalskalierte Variable auf eine Ordinalskala «herabskaliert» werden.

Beispiel 20 *Die folgende Tabelle zeigt das Körpergewicht und die Platzierung von 20 Teilnehmern an einem Laufbewerb. Gibt es zwischen diesen beiden Größen einen statistischen Zusammenhang? Gib den entsprechenden Korrelationskoeffizienten an:*

[9] Charles Edward Spearman, 1863 - 1945, britischer Psychologe

X Gewicht [kg]	Y Platzierung	X Gewicht [kg]	Y Platzierung
58	2	55	6
84	15	85	12
92	14	95	20
70	7	58	4
86	16	63	1
81	17	87	19
63	3	61	10
90	13	64	5
97	18	71	8
89	11	63	9

Wenn wir die «normale» Pearson-Korrelation nach 4.7 ausrechnen, erhalten wir einen Korrelationskoeffizient von 0.86. Allerdings haben wir dabei nicht beachtet, dass das Gewicht eine rationalskalierte Zufallsgröße ist und die Platzierung ordinalskaliert. Formel 4.7 darf daher nicht unmittelbar angewandt werden; die Zufallsgröße «Gewicht» muss zuvor ebenfalls ordinalskaliert werden. Daraus ergibt sich in diesem Beispiel:

X Gewichts-Rang	Y Lauf-Rang	X Gewichts-Rang	Y Lauf-Rang
2	2	1	6
12	15	13	12
18	14	19	20
9	7	2	4
14	16	5	1
11	17	15	19
5	3	4	10
17	13	8	5
20	18	10	8
16	11	5	9

und daraus für die Rangkorrelation, nach 4.7 ein Korrelationskoeffizient von 0.82.

Es gibt allerdings in den Originaldaten auch ex aequo-Plätze: Zwei Personen haben je 58 kg, drei Personen 63 kg. Der Pearson-Korrelationskoeffizient ist für die Rangkorrelation daher nur ein Näherungswert. Für die exakte Berechnung müssen wir den Spearman-Koeffizienten verwenden und dazu die Differenzen zwischen den jeweiligen Rangzahlen (und in weiterer Folge deren Quadratsumme) angeben:

X Gewichts-Rang	*Y Lauf-Rang*	D_i	D_i^2
2	*2*	*0*	*0*
12	*15*	*-3*	*9*
18	*14*	*4*	*16*
9	*7*	*2*	*4*
14	*16*	*-2*	*4*
11	*17*	*-6*	*36*
5	*3*	*2*	*4*
17	*13*	*4*	*16*
20	*18*	*2*	*4*
16	*11*	*5*	*25*
1	*6*	*-5*	*25*
13	*12*	*1*	*1*
19	*20*	*-1*	*1*
2	*4*	*-2*	*4*
5	*1*	*4*	*16*
15	*19*	*-4*	*16*
4	*10*	*-6*	*36*
8	*5*	*3*	*9*
10	*8*	*2*	*4*
5	*9*	*-4*	*16*
			$\Sigma = 246$

Daraus ergibt sich der Spearman-Rangkorrelationskoeffizient:

$$r_s = 1 - \frac{6 \cdot 246}{20^3 - 20} = 1 - \frac{1476}{7980} = \underline{\underline{0.82}}$$

Für das Beispiel 20 haben wir für die Berechnung des «Gewichts-Ranges» die *Excel*-Funktion `=RANG.GLEICH(Zahl;Bezug;1)` verwendet. In *LibreOffice Calc* lautet der entsprechende Funktionsaufruf ebenfalls `=RANG.GLEICH(Zahl;Bezug;1)`.

Aufgabe 14 Welche Werte kann ein Rangkorrelationskoeffizient annehmen?

4.4 Zusammenhänge kategorischer Merkmale

Bei kategorischen Merkmalen, zum Beispiel beim Vorliegen qualitativer Nominaldaten (z.B. Geschlecht, Beruf, Herkunftsland,...) aber auch quantitativen metrischen Daten, die in Klassen eingeteilt wurden (z.B. nach Altersgruppen), können die bisherigen Methoden dieses Kapitels nicht so einfach angewandt werden. Mit Kategorien oder Klassenintervallen können wir ja nicht wirklich gut einen Korrelationskoeffizienten berechnen. Für Nominaldaten können wir keine Rangkorrelation angeben, sie können auch in keinem Streudiagramm dargestellt werden ☹.

Ihre Häufigkeitsverteilung können wir aber in einer Art «gemeinsame Häufigkeitstabelle» darstellen, genannt **Kontingenztabelle** (auch: *Kreuztabelle*). Das sehen wir uns am besten an Hand eines Beispiels an:

Beispiel 21 *In einem Unternehmen gibt es drei «Verwendungsgruppen» für die Mitarbeiterinnen und Mitarbeiter:*

- *A : Arbeitnehmer:innen, die qualifizierte Tätigkeiten aufgrund ihrer Kenntnisse und Erfahrungen im Rahmen an sie erteilter Aufträge weitgehend selbstständig erledigen*
- *B : Arbeitnehmer:innen, die verantwortungsvolle Expert:innen-Tätigkeiten mit entsprechendem Entscheidungsspielraum selbstständig verrichten*
- *C : Arbeitnehmer:innen mit erhöhtem Verantwortungsbereich in leitenden Stellungen, inkl. Mitarbeiter:innenführung*

Im konkreten Fall der Firma ABC gibt es $n = 51$ Mitarbeiter:innen, darunter in der Verwendungsgruppe A 25 weibliche und 5 männliche Angestellte, in der Verwendungsgruppe B 5 weibliche und 7 männliche Angestellte und in Verwendungsgruppe C 4 Frauen und 5 Männer.

Das können wir auch in einer Tabelle darstellen:

	Geschlecht	
Verwendungsgruppe	*weiblich*	*männlich*
A	*25*	*5*
B	*5*	*7*
C	*4*	*5*

Für eine Kontingenztabelle ergänzen wir nun sowohl die Spalten als auch die Zeilen in obiger Tabelle um eine Summenspalte bzw. -zeile:

Beispiel 21 *(Fortsetzung)*

Verwendungsgruppe	*Geschlecht* *weiblich*	*männlich*	Σ
A	*25*	*5*	*30*
B	*5*	*7*	*12*
C	*4*	*5*	*9*
Σ	*34*	*17*	*51*

Anschließend bestimmen wir die so genannten **Randverteilungen**. Dazu geben wir zunächst die relativen Häufigkeiten (in Prozent) an, d.h. wir dividieren einfach jeden Wert in der Tabelle durch n (in unserem Beispiel: durch 51)

Beispiel 21 *(Fortsetzung)*

Verwendungsgruppe	*Geschlecht* *weiblich*	*männlich*	*RV*
A	*49.0%*	*9.8%*	***58.8%***
B	*9.8%*	*13.7%*	***23.5%***
C	*7.8%*	*9.8%*	***17.6%***
Σ	***66.7%***	***33.3%***	*100.0%*

Die Randverteilungsspalte *bedeutet: 58.8% der Mitarbeiter:innen sind in Verwendungsgruppe A angestellt, 23.5% in Verwendungsgruppe B und 17.6% in C. Und die* Randverteilungszeile*: 66.7% der Angestellten sind weiblich und 33.3% männlich.*

Unser Ziel ist, eventuelle *Zusammenhänge* zwischen den auftretenden Zufallsvariablen zu untersuchen. In unserem Beispiel können wir hinterfragen, ob es einen Zusammenhang zwischen dem Geschlecht und der Verwendungsgruppe gibt, oder ob diese beiden Merkmale unabhängig voneinander sind.

Die weitere Vorgangsweise schaut zugegebenermaßen auf den ersten Blick etwas kompliziert aus, tatsächlich lässt sie sich aber z.B. in EXCEL ziemlich einfach bewerkstelligen. Zunächst einmal überlegen wir, welche Häufigkeitsverteilung wir in den einzelnen Verwendungsgruppen erwarten würden, wenn es eine vom Geschlecht unabhängige Verteilung gäbe. Offensichtlich wären das sowohl unter den 34 Frauen als auch unter den 17 Männern jeweils 58.8% in Gruppe A, 23.5% in Gruppe B und 17.6% in Gruppe C.

Beispiel 21 *(Fortsetzung) Die* ***erwartete*** *Verteilung sieht so aus:*

Verwendungsgruppe	*Geschlecht*	
	weiblich	*männlich*
A	*20*	*10*
B	*8*	*4*
C	*6*	*3*

Wir sehen: Zwischen Realität und erwarteter Verteilung besteht ein Unterschied. Diese Differenzen rechnen wir zunächst aus und quadrieren sie anschließend. Und dann dividieren wir noch alle quadratischen Differenzen durch die erwarteten Werte:

Beispiel 21 *(Fortsetzung)*

$$\frac{(25-20)^2}{20} = 1.25 \quad \frac{(5-10)^2}{10} = 2.5$$
$$\frac{(5-8)^2}{8} = 1.13 \quad \frac{(7-4)^2}{4} = 2.25$$
$$\frac{(4-6)^2}{6} = 0.67 \quad \frac{(5-3)^2}{3} = 1.33$$

Wenn wir jetzt die **Summe** der eben berechneten Werte bilden, sind wir schon fast am Ziel:

Beispiel 21 *(Fortsetzung)*

$$\chi^2 = 1.25 + 2.5 + 1.13 + 2.25 + 0.67 + 1.33 = 9.13$$

Das Formelzeichen, das wir dafür verwendet haben, ist übrigens ein griechisches *Chi* (bzw. ein Chi zum Quadrat, ausgesprochen «Ki quadrat»); der Wert selbst ist ein Maß für die *Quadratische Kontingenz*. Letztendlich können wir aus dem χ^2 dann den **korrigierten Kontingenzkoeffizienten** ausrechnen:

$$C_{korr} = \sqrt{\frac{k}{k-1} \cdot \frac{\chi^2}{\chi^2+n}}$$
mit:
$k = \min(i; j)$ also die kleinere der beiden Zahlen i und j
$j =$ Anzahl der unterschiedlichen Kategorien der Zufallsvariablen X
$i =$ Anzahl der unterschiedlichen Kategorien der Zufallsvariablen Y (4.12)

Beispiel 21 *(Fortsetzung)*

In unserem Beispiel gibt es $j = 2$ *Geschlechter und* $i = 3$ *Verwendungsgruppen, die kleinere der beiden Zahlen ist* 2, *daher* $k = \min(3; 2) = 2$ *und*

$$C_{korr} = \sqrt{\frac{2}{2-1} \cdot \frac{9.13}{9.13+51}} = \underline{\underline{0.55}}$$

Der Wert deutet darauf hin, dass die beiden Merkmale nicht unabhängig voneinander sind.

Ob es sich dabei um eine signifikante *Abhängigkeit handelt, darauf werden wir in einem späteren Kapitel noch einmal zurückkommen.*

Zusammenfassend können wir sagen: Welches Maß wir für die Angabe des statistischen Zusammenhangs angeben können, ist abhängig vom Skalenniveau der Zufallsvariablen:

- Für *metrische* Daten können wir den **Bravais-Pearson Korrelationskoeffizienten** verwenden,
- für *Ordinaldaten* den **Rangkorrelationskoeffizienten nach Spearman**; und
- für *Nominaldaten* den **korrigierten Kontingenzkoeffizienten**.

Sind die beiden Zufallsvariablen auf unterschiedlichem Niveau, dann können wir nur jenes Merkmalsmaß verwenden, das für die Zufallsvariable auf niedrigerem Niveau möglich ist. Für den Zusammenhang zwischen einer metrischen Variable und einer rangskalierten, kann «nur» der Spearman-Koeffzient angegeben werden, nicht aber ein Bravais-Pearson-Koeffizient; für den Zusammenhang zwischen einer metrischen Variable und einer kategorialen nur ein Kontingenzkoeffizient.

4.5 Statistische und kausale Zusammenhänge

Die Korrelation beschreibt per se zunächst *statistische* und nicht unbedingt *kausale* Zusammenhänge. Das heißt selbst ein sehr, sehr hoher Wert des Korrelations- oder Kontingenzkoeffizienten (nahe ± 1) sagt nichts darüber aus, dass das eine Merkmal die *Ursache* für die Größe des anderen Merkmals ist. Natürlich *kann* eine kausale Beziehung bestehen, das muss aber nicht der Fall sein. Hier muss man unterscheiden, ob die Daten, die wir ausgewertet haben, aus einer reinen *Beobachtung* stammen oder aus einem gezielten *Experiment*.

Der Unterschied sei am Beispiel des «Mozarteffekts» erläutert:

Bei einer reinen Beobachtung fragen wir Studierende, wie oft und wie lange sie während des Lernens Musik von Mozart hören. Das vergleichen wir – individuell – mit der Anzahl der Punkte, die diese Studierenden auf die Tests bekommen haben, für die sie (mit oder ohne Mozart) gelernt haben. Das ergibt zwar statistische Zusammenhänge (vielleicht), aber keine kausalen.

Frances Rauscher, Gordon Shaw und Katherine Ky von der Universität von Irvine, Kalifornien, berichteten 1993 im Wissenschaftsjournal *Nature*, dass Studierende nach dem Anhören von Mozarts Sonate für zwei Klaviere, KV 448, in einem anschließenden Test über ihre Fähigkeiten zum räumlichen Denken signifikant höhere Leistungen erzielt hatten als ihre Kollegen, die entweder Entspannungsmusik zu hören bekamen oder überhaupt in aller Ruhe den Test absolvierten. (Rauscher Frances, Gordon Shaw und Katherine Ky. 1993. «Music and spatial task performance». In: *Nature* Vol.365, Oktober 1993, S.611). Nachdem das Thema von diversen Medien mit Begeisterung aufgenommen und verbreitet wurde, ließ sich ein geschäftstüchtiger Autor den Begriff «Mozart Effect» schützen und verdiente gut mit einem Buch und Vorträgen, in denen er der Macht Mozarts Musik gleich auch die Linderung von körperlichen Beschwerden und heilende Effekte im Fall von Aids, diversen Allergien und Diabetes versprach. (Mozart selbst war übrigens von Kindheit an immer wieder kränklich und starb 1791 mit nur 36 Jahren. Er hätte öfter seine eigene Musik hören sollen).
Hast du die Mozartsonate angehört (zum Beispiel unter t1p.de/mozartkv448, aber es hat nicht mit dem Zuwachs räumlicher Intelligenz oder der Verbesserung deiner Gesundheit geklappt, kannst du sie auch anderweitig verwenden: Ein Milchbauer aus der Nähe von Madrid beschallt seine 700 Kühe jeden Tag mit Mozart. «Es klappt nur mit Mozart», schwört Nicolas Siebert. Die Kühe seien nicht nur ausgeglichener und einfacher im Umgang, jede einzelne produziere auch ein bis sechs Liter mehr Milch pro Tag.

(Quellen: Swartz, Luke. 2000. *The Mozart Effect: Does Mozart Make You Smarter?* http://xenon.stanford.edu/~lswartz/mozarteffect.pdf. Sowie: Driessen, Barbara. 2008. *Mozart-Sonaten beruhigen Kinder und Kühe.* WELT ONLINE 28.2.2008. http://www.welt.de/wissenschaft/article1735411/)

Der *Mozart-Effekt:* Statistischer oder kausaler Zusammenhang?

In diesem Zusammenhang spricht man auch oft von einer **Scheinkorrelation**.

Als Statistiker:innen wissen wir, dass Zusammenhänge wie beim Mozart-Effekt zwar vielleicht tatsächlich aufzeigbar sind, dass es sich dabei aber eben um *statistische* Zusammenhänge handelt und nicht um *kausale*. Es kann zum Beispiel sein, dass Menschen, die intelligenter sind, auch eher klassische Musik hören, als Menschen mit einem niedrigen Intelligenzquotienten. Daraus kann aber nicht abgeleitet werden, dass ein wenig Mozart-Hören praktisch ohne sonstigen Aufwand die Intelligenz steigert. Auch zwischen dem gesundheitlichen Wohlbefinden und der Vorliebe für bestimmte Musik *kann* ein Zusammenhang bestehen, aber auch hier ist – zumindest mit der statistischen Methode der Korrelationsrechnung – keine Kausalitätsrichtung auszumachen.

Bei reiner Beobachtung gilt: Es lässt sich nichts über den kausalen Zusammenhang sagen.

Es gibt daher drei Möglichkeiten: Das Hören von Mozart führt zu besseren Lernergebnissen. Oder: Wer gut lernt, hört auch gerne Mozart. Oder: Es gibt eine dritte, uns unbekannte Variable, die zu einer Verbesserung der Lernergebnisse bei denen, die gerne Klassik hören, geführt hat. Diese dritte Variable nennen wir auch *Störfaktor* (engl.: *confounder*)[10].

Bei einem Experiment hingegen ginge das so: Zufällig ausgewählte Menschen bekommen Mozart vorgespielt (während des Lernens); eine andere Gruppe bekommt keine Musik. Wenn die beiden Gruppen wirklich zufällig ausgewählt wurden, finden sich in beiden Gruppen ungefähr gleich viele Klassikliebhaber wie solche, denen diese Musik nicht besonders gefällt. Wenn jetzt trotzdem die eine Gruppe einen besseren Lernerfolg zeigt, zeigt das einen kausalen Zusammenhang.

Korrelation bedeutet nicht Kausalität. Ob ein kausaler Zusammenhang besteht, ist nur aus der Art der Datenerhebung ableitbar: Aus einer Beobachtung alleine ist keine Kausalität ableitbar, aus einem Experiment hingegen kann eine Kausalität vermutet werden.

Aufgabe 15 In einem bestimmten Jahrgang wurden bei einer Analyse der Test- und Prüfungsergebnisse aus MAT101 (Mathematik) und MAT102 (Statistik) u.a. folgende Zusammenhänge beobachtet:

Korrelation zwischen den Gesamtpunkten aus MT122 und den im Vorsemester erreichten Punkten aus MAT101: $r = 0.37$ (Determinationskoeffizient: $r^2 = 14\%$).
Korrelation zwischen den aus den Online-Tests in MT122 erreichten Punkten und der Zeit, die im Durchschnitt für die Bearbeitung der Online-Tests aufgewandt wurde: $r = -0.08$ (Determinationskoeffizient: $r^2 = 1\%$).

Was lässt sich daraus über den Zusammenhang zwischen Mathematik- und Statistik-Kenntnissen bzw. den Zeitaufwand, den Studierende für die Online-Tests aufwenden, sagen?

[10] Wobei die deutsche Bezeichnung *Störfaktor* ein wenig unglücklich ist, weil ja keine Störung im Sinne eines *Störenfrieds* vorliegt, der das Ergebnis des Experiments verunmöglicht, sondern lediglich ein Faktor von außen einen Einfluss nimmt, an den wir nicht gedacht haben.