Serie „Studien verstehen“

Korrelation: Wie stark Merkmale zusammenhängen

Als übergeordneter Begriff dafür, dass zwei Merkmale miteinander zusammenhängen, hat sich der Begriff Korrelation eingebürgert. Das sagt allerdings noch nichts über eine Kausalität aus.

Roland Müller-Waldeck

Wenn zwei Merkmale miteinander korrelieren, kann daraus niemals zwingend geschlossen werden, dass ein Kausalzusammenhang zwischen ihnen besteht. Das scheint auf den ersten Blick unlogisch zu sein, wird aber schnell klar, wenn man verschiedene Typen von Kausalbeziehungen betrachtet:

Bei der einseitigen Abhängigkeit ist Merkmal 1 (M1) Ursache von Merkmal 2 (M2), bei einer wechselseitigen Abhängigkeit beeinflussen sich M1 und M2 gegenseitig. In diesen beiden Fällen besteht tatsächlich ein Kausalzusammenhang zwischen den beiden Merkmalen. Bei einer Abhängigkeit von einer gemeinsamen Ursache jedoch werden M1 und M2 von einem dritten Faktor beeinflusst und schließlich können M1 und M2 von einem komplizierten System mit vielen Regelkreisen geregelt werden. In solchen Fällen besteht keine Kausalbeziehung zwischen M1 und M2.

Darüber hinaus gibt es Schein- oder Nonsenskorrelationen, bei denen zwar ein zahlenmäßiger Zusammenhang vorliegt, aber ein biologischer Zusammenhang nicht auszumachen ist: So gibt es eine Korrelation zwischen der Anzahl der Störche, die in einem Ort nisten und der Anzahl der Kinder, die dort geboren werden. Der Grund ist der, dass in größeren Orten mit mehr Häusern mehr Kinder geboren werden, aber mehr Häuser auch mehr Störchen Nistmöglichkeiten bieten. Ob eine Korrelation biologisch-medizinisch sinnvoll ist, kann nur die fachliche Beurteilung ergeben. Stellt ein Wissenschaftler eine Korrelation fest, muss er danach trachten, mit Versuchen (kontrollierte randomisierte Studie) eine Kausalbeziehung nachzuweisen. Das gelingt nicht immer.

Metrische Daten: Maßkorrelationskoeffizient r

Der Korrelationskoeffizient r ist eine Maßzahl dafür, wie stark zwei Merkmale miteinander korrelieren. Er wird aus den in einer Studie ermittelten Wertepaaren errechnet und kann einen Wert von +1 bis −1 annehmen. Wenn die Messwerte der Merkmale auf einer exakten Geraden mit positiver Steigung liegen, ist $r = +1$, bei negativer Steigung ist $r = -1$.

Daraus kann jedoch nicht geschlossen werden, dass jede mathematisch beschreibbare Beziehung zwischen zwei Merkmalen zu einem Korrelationskoeffizienten mit dem Wert 1 führt: Die

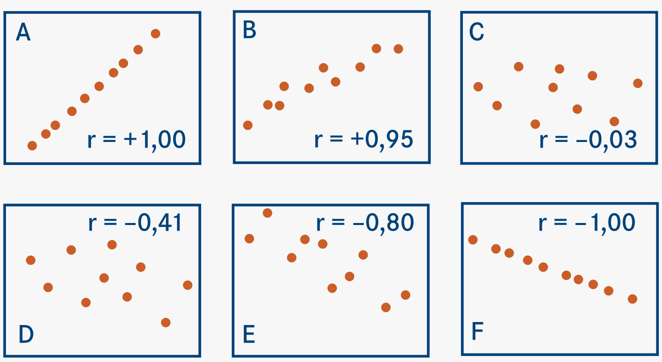

Abb. 1: In Abb. A und F liegen die Messwerte auf einer Geraden, daher beträgt r 1 oder −1. Bei B und E lässt sich noch gut mit dem Auge eine Gerade zwischen den Punkten ziehen und r weicht schon deutlich von 1 oder −1 ab. Bei C und D ist es schon schwer, mit dem Auge eine Gerade durch die Punkte zu legen und der reine Zahlenwert für r liegt schon unter 0,5. (Nach Lorenz)

Punkte einer Parabel ergeben zum Beispiel r = 0, obwohl sie perfekt zusammenhängen, jedoch mit einer nichtlinearen Funktion. Der Korrelationskoeffizient kann eine nichtlineare Beziehung nicht erkennen. Daraus, dass r keine Korrelation ergibt, kann also nicht geschlossen werden, dass die Messwerte nicht miteinander zusammenhängen, umgekehrt aber korrelieren Messwerte, die nicht voneinander abhängen, nicht miteinander. Daher ist es immer empfehlenswert, sich die Messpunkte zunächst in einem Koordinatensystem anzuschauen, um sich einen Überblick zu verschaffen. In Abb. 1 sind verschiedene Punktewolken mit ihren Werten für r dargestellt. Das 95% Konfidenzintervall (KI) für r schließt den Bereich ein, der mit 95%iger Wahrscheinlichkeit den Korrelationskoeffizienten der Gesamtpopulation enthält. Dabei ist das 95% KI nicht symmetrisch, weil r nur Werte zwischen +1 und −1 annehmen kann. Um den p-Wert für r = x zu interpretieren, muss man sich die Nullhypothese klar machen. Sie lautet: Es gibt keine Korrelation zwischen M1 und M2 in der gesamten Population. Der p-Wert gibt wieder, wie groß in diesem Fall die Wahrscheinlichkeit ist, dass die ausgewertete Anzahl von unabhängigen Messungen zu einem Wert x für r führt, der größer als +x oder kleiner als −x ist. Liegt diese Wahrscheinlichkeit unter 5%, ist das Ergebnis signifikant.

Je mehr Störche, desto mehr Kinder – mag sein, aber hier besteht sicher kein kausaler Zusammenhang.

Häufige Fehler

Eine Korrelation beweist keinen kausalen Zusammenhang: Oben wurde kurz angerissen, dass Werte korrelieren können, auch wenn sie nicht voneinander abhängen. Hinzu kommt, dass der Wert für r zufällig größer sein kann als 0, obwohl r in der Gesamtpopulation 0 ist. Der p-Wert hilft dabei, einen solchen Fall zu erkennen. Welche oben angeführte Form eines Kausalzusammenhangs vorliegt, kann nur mit Experimenten ermittelt werden, in denen nur ein Enflussfaktor verändert und die daraus resultierende Auswirkung gemessen wird.

Fokussierung auf den p-Wert: Wie bei allen andere statistischen Auswertungen gilt auch hier: Auch wenn der p-Wert sehr niedrig sein kann, entscheidet doch die Größe des gemessenen Effekts darüber, ob er Bedeutung besitzt oder nicht. Ist r niedrig, korrelieren x und y nur schwach miteinander. Daran ändert auch der niedrige p-Wert nichts.

Man sollte r und den p-Wert für r nicht interpretieren, bevor man einen Blick auf die Punktewolke geworfen hat: Abb. 2 zeigt künstliche Daten von Anscomb, auf die zutrifft: r = 0,816 und p = 0,022. Nur der erste Graph erfüllt die Voraussetzungen für eine Korrelation.

Voraussetzungen für eine Korrelation

- Zufällige Messwerte
- Jeder Messwert besteht aus x- und y-Wert
- Die Messwerte entstammen der selben Population
- Es handelt sich um unabhängige Beobachtungen, bei denen ein x nur ein y beeinflusst
- y-Werte werden nicht aus x-Werten errechnet
- x-Werte werden nicht wie in einem Experiment kontrolliert
- x und y folgen einer Gaußschen Normalverteilung
- Der Zusammenhang zwischen x und y ist linear
- Es gibt keine Ausreißer; sie können r dramatisch beeinflussen.

Literatur:

Lorenz RJ. Grundbegriffe der Biometrie. 2. Auflage, Fischer, 1988

Motulsky H. Intuitive Biostatistics. Oxford University Press, 4. Auflage, 2018

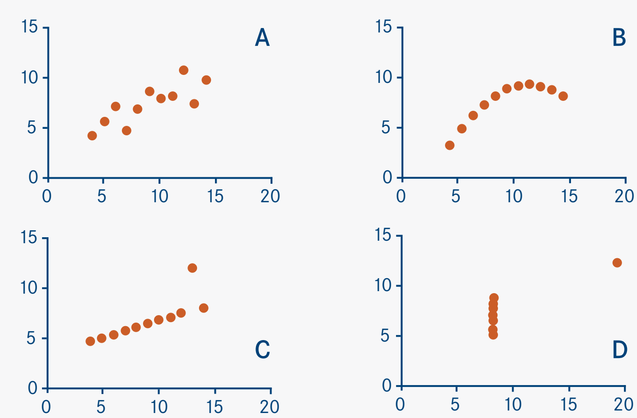

Abb. 2: Vier Punktewolken, von denen nur A die Voraussetzungen für eine Korrelation erfüllt: In B liegt keine lineare Beziehung vor, in C gibt es einen Ausreißer, in D gibt es einen Ausreißer und der x-Wert bei allen Messungen ist gleich, unterliegt also keiner Gaußschen Normalverteilung. (Nach Motulsky)