

Korrelation und Kausalität

Statistische Methoden: Korrelations- und Regressionstechniken

In der empirischen Bildungsforschung spielen Fragen nach den Zusammenhängen von Merkmalen eine gewichtige Rolle. Auf dieser empirischen Seite sollen die statistischen Methoden erläutert werden, um die publizierten Befunde besser nachvollziehen und bewerten zu können. Wegen der Gefahr von Fehlinterpretationen werden auch zu beachtende Randbedingungen und Aussagegrenzen deutlich gemacht.

Thomas Frein
Gerd Möller
Andreas Petermann
Michael Wilpicht

Ministerium für Schule und Weiterbildung
des Landes Nordrhein-Westfalen

Mithilfe von statistischen Methoden, den so genannten Korrelations- und Regressionsrechnungen, wird untersucht, ob verschiedene Merkmale (Variablen) des Untersuchungsfeldes in einem statistischen Zusammenhang zueinander stehen. Hierzu gibt es verschiedene Assoziationsmaße, die den statistischen Zusammenhang zweier Variablen darstellen. Häufig verwendete Maße sind die so genannten Kontingenz- bzw. Korrelationskoeffizienten nach Pearson, die Werte zwischen -1 und $+1$ annehmen können. Wobei der Wert 0 die statistische Unabhängigkeit beider Variablen und -1 bzw. $+1$ einen perfekten negativen bzw. positiven linearen Zusammenhang charakterisieren. Werte zwischen 0 und $-/+1$ deuten auf einen mehr oder weniger starken Zusammenhang je nach Ausprägung hin. Anhand der Grafik soll beispielhaft die prinzipielle Vorgehensweise der Korrelations- bzw. Regressionsrechnungen dargestellt werden. In dem Streudiagramm sind die Anteile der Risikoschüler, die im Schuljahr 1999/2000 am PISA-Test teilgenommen haben, den Anteilen der Schülerinnen und Schüler, die die allgemein bildenden Schulen ein Jahr später (2000/2001) ohne Hauptschulabschluss – so genannte Abbrecher – verlassen haben, zugeordnet.

Die eingezeichnete Gerade wurde so bestimmt, dass die Punkte des Streudiagramms einen möglichst geringen Abstand von der Gerade haben. Mathematisch ausgedrückt: Die Summe der quadratischen Ordinatenabstände der Punkte von der Geraden ist minimal. Der angegebene Korrelationskoeffizient r kann als Maß für die

Abweichung der Punktwolke von der Geraden interpretiert werden. Liegen alle Punkte auf der Geraden, dann ist der Wert bei einer steigenden Geraden $+1$ und bei einer fallenden Geraden -1 .

Bereits ohne Berücksichtigung des errechneten Korrelationskoeffizienten wird deutlich, dass es über die Länder (fast) keinen Zusammenhang zwischen der Risikoschüler- und Abbrecherquote gibt. So hat z. B. Bayern eine relativ geringe Quote von Risikoschülern, aber eine bemerkenswert hohe Quote von Abgängern der Sekundarstufe I ohne Hauptschulabschluss. In Nordrhein-Westfalen sind die Verhältnisse umgekehrt. Dies zeigt sich rechnerisch mit dem r -Wert ($r=+0,105$) von annähernd 0 . Das Ergebnis lässt sich so interpretieren, dass es in den Ländern offensichtlich unterschiedliche Kompetenzanforderungen bezüglich der Vergabe von Abschlüssen gibt. Mit der Einführung von Bildungsstandards sollten sich in Zukunft die unterschiedlichen Anforderungen allerdings angleichen.

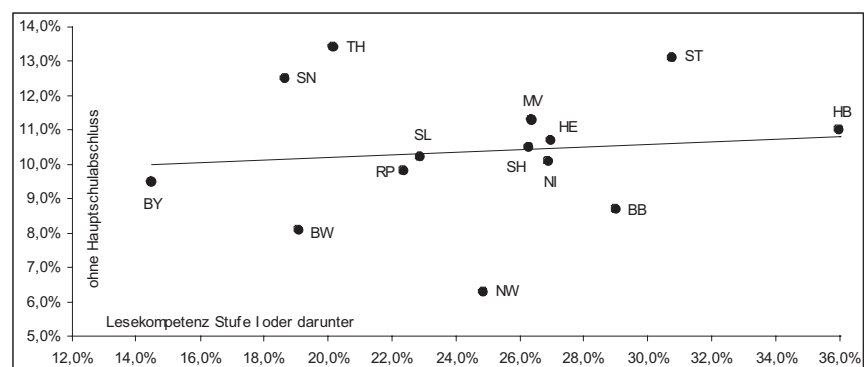
Es ist wichtig zu beachten, dass es Fälle geben kann, in denen zunächst kein rechnerischer Zusammenhang zwischen zwei Variablen festgestellt werden kann und erst bei Berücksichtigung weiterer Vari-

ablen ein korrelativer Zusammenhang sichtbar wird. Es kann also zwischen zwei Variablen Kausalbeziehungen geben, die durch Drittvariable unterdrückt werden und erst bei Kontrolle dieser Drittvariablen zum Vorschein kommen.

Umgekehrt ist zu beachten, dass ein statistischer Zusammenhang zwischen zwei Variablen noch nichts darüber aussagt, ob die eine Variable die andere beeinflusst oder umgekehrt: Ein starker korrelativer Zusammenhang muss keinesfalls einen kausalen Hintergrund anzeigen. Dies soll an einem klassischen Beispiel aus der statistischen Literatur näher erläutert werden: In den zwanziger Jahren des letzten Jahrhunderts wurde in Schweden ein hoher korrelativer Zusammenhang zwischen einer sinkenden Geburtenrate und einer Abnahme der Storchpopulation festgestellt. Es ist offensichtlich, dass es sich nicht um einen kausalen Zusammenhang handelt, denn die Kinder werden nun mal nicht von den Störchen gebracht. Die Ursache für den parallelen Rückgang der beiden Populationen dürfte hingegen in einem zunehmenden Trend zur Verstädterung zu finden sein.

Fazit

Korrelations- und Regressionstechniken stellen ein mächtiges Instrumentarium zum Aufdecken von Zusammenhängen dar. Ob die korrelativen Zusammenhänge auch kausaler Natur sind, kann aber nur eine begründete Theorie mit empirischer Absicherung klären. ■



Zusammenhang zwischen dem Anteil der Risikoschüler bezüglich der Lesekompetenz unter den 15-jährigen Deutschen in PISA 2000 und dem Anteil der Schülerinnen und Schüler ohne Hauptschulabschluss in den Ländern im Schuljahr 2001/2002