

Unsinn in den Medien – Vom allzu sorglosen Umgang mit Daten: Statistisches Testen von Hypothesen

Ergebnis verzweifelt gesucht – egal welches

Viele Kritiker stoßen sich auch daran, dass der p -Wert »unsauberer Denken« fördert. Bestes Beispiel ist die Tatsache, dass er die Aufmerksamkeit von der tatsächlichen Größe des Effekts weglenkt.

Im Jahr 2013 ergab eine groß angelegte Studie mit mehr als 19000 Teilnehmern unter der Leitung von John T. Cacioppo von der University of Chicago, dass sich Eheleute, die sich online kennen gelernt haben, mit geringerer Wahrscheinlichkeit wieder scheiden lassen ($p < 0,002$). Außerdem waren die noch Verheirateten unter ihnen mit ihrer Ehe tendenziell zufriedener als jene, die auf traditionellem Weg zusammengekommen waren ($p < 0,001$). Diese Werte sehen eindrucksvoll aus, aber der eigentliche Effekt war winzig: Kennenlernen über das Internet drückte die Scheidungsrate von 7,67 auf 5,96 Prozent und hob die Zufriedenheit mit der Ehe von 5,48 auf 5,64 auf einer Sieben-Punkte-Skala.

Die Signifikanz eines Ergebnisses sage eben nichts über dessen praktische Bedeutung aus, erklärt Geoff Cumming, emeritierter Psychologieprofessor von der La Trobe University in Melbourne (Australien). »Wir sollten uns fragen: ›Wie groß ist der Effekt, mit dem wir es zu tun haben?‹ und nicht: ›Gibt es überhaupt einen Effekt?‹«

(gefunden in Spektrum der Wissenschaft SPEZIAL 3.17, S.34)

Kommentar: Da muss ich aber dem emeritierten Psychologieprofessor vehement widersprechen: Denn eine fehlende Relevanz von signifikanten Testergebnissen bedeutet tatsächlich nur eines mit Sicherheit – nämlich dass die Hypothesen falsch formuliert wurden. Warum wird denn getestet, ob die Scheidungsrate von 7,67 Prozent bei Internetbekanntschäften *in irgendeinem Ausmaß* unterschritten wird, wenn ein kleiner Unterschied gar nicht relevant ist? Wird z.B. erst eine Rate von unter 5 Prozent als praktisch bedeutsamer Effekt interpretiert, dann muss die Einshypothese natürlich $P < 5$ Prozent und nicht $P < 7,67$ Prozent lauten. Und schon ist die Testgröße in der Stichprobe (5,96 Prozent) nicht nur *nicht relevant*, sondern auch *nicht signifikant*.

Will man also testen, ob die Zufriedenheit mit der Ehe in dieser Gruppe der Bekanntschäften sich von der in den anderen Ehen praktisch bedeutsam unterscheidet, hat man sich wieder zuerst zu fragen, was ein *relevanter* Effekt ist, und dann genau *diesen* zu testen. So könnte die Einshypothese eben lauten: Mittelwert > 6 Punkte. Dann wäre ein solcher wie der gefundene (5,64) eben auch bei dieser Fragestellung nicht nur *nicht relevant*, sondern auch *nicht signifikant*. Ein *signifikantes* Ergebnis aber wäre dann auch *praktisch bedeutsam*. Denn wenn gleich die richtigen Hypothesen formuliert würden, also man das testet, was man testen möchte, dann sagt die Signifikanz eines Ergebnisses eben schon etwas über dessen praktische Bedeutung aus!

Wie man die nötigen Relevanzgrenzen zum Testen festlegen soll? – Diese müssten die Fachexpertinnen und -experten der Studie schon selbst festlegen können. Immerhin werden die tatsächlich gefundenen Effekte als nicht relevant eingestuft. Dann sollte sich doch eine Grenze zwischen den praktisch bedeutsamen und den nicht bedeutsamen Effekten ziehen lassen ...

(Für den Kommentar verantwortlich: Andreas Quatember, IFAS)