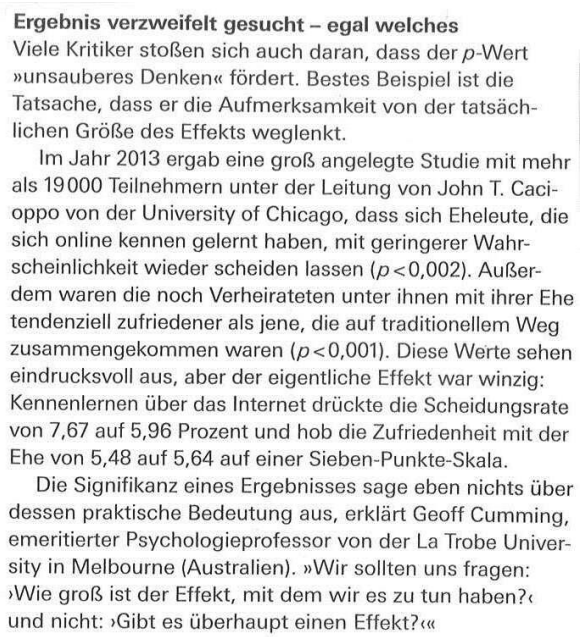
Unsinn in den Medien – Vom allzu sorglosen Umgang mit Daten:
Statistisches Testen von Hypothesen



Ergebnis verzweifelt gesucht – egal welches

Viele Kritiker stoßen sich auch daran, dass der *p*-Wert »unsauberes 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 sich die Scheidungsraten von Internetbekanntschaften *in irgendeinem Ausmaß* von jenen, die sich nicht Online kennengelernt haben, unterscheiden, wenn ein kleiner Unterschied gar nicht relevant ist? Wird z.B. erst ein Unterschied D von zwei Prozentpunkten als praktisch bedeutsamer „Effekt" interpretiert, dann muss die Einshypothese natürlich $|D| > 2$ Prozentpunkte und nicht $|D| > 0$ lauten. Und schon ist die Testgröße in der Stichprobe (1,71 Prozentpunkte) nicht nur *nicht relevant*, sondern auch *nicht signifikant*.

Will man testen, ob die Zufriedenheit mit der Ehe in dieser Gruppe der Bekanntschaften sich von der in den anderen Ehen praktisch bedeutsam unterscheidet, hat man sich wieder zuerst zu fragen, was ein *relevanter* Unterschied ist, und dann genau *diesen* zu testen. So könnte die Einshypothese etwa lauten: Betrag der Mittelwertsdifferenz $D > 0{,}5$ Punkte. Dann wäre eine solche wie der gefundene (0,16) 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 werden, also man das testet, was man tatsächlich testen möchte, dann sagt die Signifikanz eines Ergebnisses auch etwas über dessen praktische Bedeutung aus, sonst nicht!

Wie man die nötigen Relevanzgrenzen zum Testen festlegen soll? – Diese müssten die Fachexpert*innen der Studie schon festlegen können. Immerhin werden die tatsächlich gefundenen Effekte von ihnen als nicht relevant eingestuft. Dann sollten sie doch eine Grenze zwischen den praktisch bedeutsamen und den nicht so bedeutsamen Effekten ziehen können …

(Für den Kommentar verantwortlich: Andreas Quatember, IFAS)