## Der Signifikanz-Mythos brĶckeltâ?¦

## **Description**

Vor Jahren traf ich mal den Altmeister der psychologischen Methodologie, den Harvard Psychologen Robert Rosenthal auf einem kleinen Symposion. Er pflegte zu sagen  $\hat{a}$ ? Surely, God loves p=.06 as he loves  $p=.05.\hat{a}$ ? Er wies damit auf die willk $\tilde{A}^1$ /arliche, ja fast unsinnige Fixierung fast aller Wissenschaftst $\tilde{A}$ mtigkeiten der quantitativen Wissenschaften, vor allem der Verhaltenswissenschaften und der Bio-Wissenschaften, auf ein konventionell festgelegtes Signifikanz-Niveau hin und predigte schon seit langer Zeit in seinen Beitr $\tilde{A}$ mgen, dass sich die Wissenschaft viel st $\tilde{A}$ mrker an anderen Kenngr $\tilde{A}$ 9. wie etwa Effektst $\tilde{A}$ mrken, Konvergenz von Befunden und methodischer  $\tilde{G}$ 1/4te von Studien orientieren sollte als an statistischen Signifikanz-Werten [1].

 $\hat{a}$ ??Surely, God loves p = .06 as he loves  $p = .05.\hat{a}$ ?•

Robert Rosenthal

Dass nun aber das Flaggschiff-Journal der Naturwissenschaften einen Aufruf bringt, die Signifikanz-Testerei zum alten Eisen zu legen, das kommt einer kleinen Revolution gleich [2]. Denn hier sprechen nicht die ohnedies oft nicht fù/4r ganz voll genommenen Verhaltens- oder Sozialwissenschaftler, sondern die Naturwissenschaftler aus dem Kern des Geschäftes: Valentin Amrhein ist Zoologe an der Uni Basel, die anderen beiden Autoren sind Statistiker. Als sie ihre Gedanken formuliert hatten, hatten sich innerhalb einer Woche 800 Wissenschaftler als Unterstù/4tzer des Aufrufs gefunden und kein geringeres Organ als *Nature* publizierte ihn.

Dahinter verbirgt sich eine sehr, sehr lange Debatte innerhalb der Wissenschaft, wie man am besten zu Entscheidungen kommt, ob es einen Sachverhalt, einen Effekt, einen Zusammenhang, eine Wirkung nun gibt oder nicht. Dass nicht einmal Wissenschaftler, die eine statistische Ausbildung genossen haben, die technischen ZusammenhĤnge hinter Signifikanztests immer korrekt verstehen, darauf haben schon viele Autoren immer wieder hingewiesen [3]. In der Statistik gibt es unterschiedliche AnsĤtze, um der Frage nachzugehen, ob ein vorliegender Unterschied oder Zusammenhang im Rahmen einer Zufallsschwankung liegt oder als genuiner Effekt angesehen werden soll.

Wann wird aus einer Ansammlung von Sandk $\tilde{A}$ ¶rnern ein Haufen? Wann wird aus einem Abendhimmel ein Sonnenuntergang? Wann wird aus mehreren Menschen eine Gruppe oder eine Versammlung? Wann werden verschiedene Fr $\tilde{A}^1$ 4chte zu einer Schale Obst? M $\tilde{A}^1$ 4ssen es zwei sein? Oder vier? Oder mehr als f $\tilde{A}^1$ 4nf? Man sieht an den einfachen Alltagsbeispielen: Gerade wenn es um kontinuierliche Zusammenh $\tilde{A}$ mge geht, ist es gar nicht leicht, eine dichotome Entscheidung f $\tilde{A}^1$ 4r oder gegen etwas zu treffen.

Wann wird aus einer Ansammlung von SandkA¶rnern ein Haufen?

Das wird aber in der Wissenschaft dauernd gemacht. Wir messen z.B. den Blutdruck eines Menschen und sagen dann: jetzt ist er erhĶht. Wir vergleichen den Depressionswert zweier unterschiedlicher Gruppen und sagen dann: Das Antidepressivum, mit dem die eine Gruppe behandelt wurde hat gewirkt (oder eben nicht).

Um solche dichotome Entscheidungen nach TatsĤchlichkeit eines Effekts, einer Wirkung, eines Zusammenhanges, treffen zu kĶnnen, mù⁄4ssen Zahlenwerte in einen Entscheidungsalgorithmus ù⁄4berfù⁄4hrt werden. Im Laufe der Zeit hat sich, vor allem in der Medizin, in den Biowissenschaften und den Sozialwissenschaften ein statistischer Ansatz durchgesetzt, der von Fisher 1935 in seinem bahnbrechenden Werk â??The Design of Experimentsâ?? vorgestellt wurde [4].

Die Grundüberlegung ist einfach. Man nimmt eine Stichprobe, in Fishers Fall waren das Ensembles von Pflanzen, die auf bestimmten Böden unter bestimmten Bedingungen wuchsen, behandelt sie oder auch nicht; vergleicht statistische Kennwerte, etwa mittlere Länge, mittleren Ertrag, etc. eines kleinen Stück Landes, das mit einer bestimmten Sorte Mais oder Weizen bebaut ist, und vergleicht es mit einer anderen. Wenn man jetzt noch ein paar statistische Ã?berlegungen hinzu nimmt â?? etwa die Ã?berlegung, dass man ja Mais zufällig aus einem groÃ?en Sack von Körnern zur Aussaat verwendet hat, oder die Verteilungsannahme, dass sich natürliche GröÃ?en wie Wachstum oder Ertrag im Sinne einer Gaussâ??schen Glockenkurve normal verteilen â?? dann kann man von der empirischen Situation auf eine ideale Situation rückschlieÃ?en und entscheiden, ob eine gegebene empirische Differenz, etwa von Ertrags- oder WachstumsmaÃ?en, sich von der Erwartung unterscheidet, dass zwischen beiden Gruppen eben kein Unterschied sein sollte.

Diese Erwartung von â??keinem Unterschiedâ??, oder â??keinem Zusammenhangâ??, ist die sog. â??Nullhypotheseâ??. Das statistische Rationale der Hypothesentestung erlaubt nun zu quantifizieren, mit welcher Wahrscheinlichkeit die Erwartung, dass es keinen solchen Unterschied gibt â?? oder die Nullhypothese â?? verletzt wird. Der Kennwert ist die Irrtumswahrscheinlichkeit p, oder der â??Signifikanzwertâ??. Dieser wurde irgendwann mal, eigentlich ziemlich willkürlich, auf p = .05 festgesetzt, also auf 5%. Das heisst: wir nehmen in Kauf, dass wir in 5% der Fälle, in denen wir behaupten, es gäbe einen Zusammenhang, oder einen Unterschied, einen Fehler machen, weil tatsächlich keiner vorhanden ist. (Ob es sich im Ã?brigen um einen Unterschied oder einen Zusammenhang handelt, ist für die rein statistische Betrachtung egal; das wird durch die konkrete Fragestellung definiert. Deswegen sage ich hier immer â??Unterschied oder Zusammenhangâ??.)

Der  $\hat{a}$ ??Signifikanzwert $\hat{a}$ ?? wurde irgendwann mal willk $\tilde{A}$ 1/4rlich auf p = .05 festgesetzt, also auf 5%.

Diese Art des statistischen Denkens ist nun so tief in den Köpfen und Herzen der Forscher verwurzelt, dass die Willkürlichkeit, die Konvention, ja auch die Reduktion einer kontinuierlichen Information auf eine dichotome Entscheidung kaum mehr wahrgenommen wird. Wir haben erst vor kurzem versucht eine Studie zu publizieren, bei der wir Signifikanzwerte von p = .07 gefunden haben, also knapp nicht signifikante Werte. Der zuständige Editor schrieb uns daraufhin: Die Studie sei nicht interessant für die Leser, weil ja kein Effekt gefunden worden sei und es sei eine Politik des Journals, solche Daten nicht mit Priorität zu behandeln.

Fishers Statistik wurde vom Hilfsmittel für Entscheidungen zur Keule, mit denen Daten und Forscher mürbe geschlagen wurden. Was kleiner oder gleich .05 ist, ist es wert betrachtet zu werden, was gröÃ?er ist, wirft man auf den Mist. Das ist etwa so, wie wenn einer sagen würde: nur eine Schale mit 31 Kirschen ist eine Kirschenschale. Sind es 30, werfen wir sie weg. Seit Generationen liefen nachdenkliche Forscher Sturm gegen dieses Art des Umgangs mit Daten und Denkens. Man wies darauf hin, dass die Tatsache, ob ein Effekt signifikant ist oder nicht, weniger vom Effekt selber, als von der GröÃ?e der Studie und damit der statistischen Mächtigkeit zusammenhängt [5].

Fishers Statistik wurde vom Hilfsmittel fÃ1/4r Entscheidungen zur Keule, mit denen Daten und

Forscher mürbe geschlagen wurden.

Ich selber habe <u>in einem meiner Methodenblogs genau diesen Zusammenhang erklĤrt</u>, eben weil er so wichtig ist fýr das VerstĤndnis statistischen Testens; ich hatte mir dieses VerstĤndnis ýbrigens selber angeeignet. Zu meiner Studienzeit war das noch nicht Gegenstand des methodischen Curriculums. Denn die Wahrscheinlichkeit, einen Effekt einer bestimmten GröÃ?e zu entdecken, die sog. statistische Mächtigkeit oder Power, ist abhängig von der GröÃ?e der Studie. Um groÃ?e Effekte zu zeigen benötigt man lediglich wenige Fälle, fýr kleine Effekte sehr viele Fälle. Daher ist die Aussage, eine Studie hätte belegt, dass xyz-Intervention bei abc-Krankheit statistisch signifikant wirksam ist im besten Falle eine kluge Verschleierungstaktik, im schlimmsten Falle sogar Betrug.

Die groÄ?en Lipidsenkerstudien haben z.B. meistens gezeigt, dass Lipidsenker den Cholesterinwert senken; statistisch signifikant. Aber der klinische Wert ist immer noch umstritten. <u>Denn die Effekte sind so klein, dass</u> man sehr viele Menschen behandeln muss, damit bei einem Menschen der Effekt positiv auftritt.

Daher ist das Konzept der Signifikanz problematisch. Denn es fokussiert eine Forschungsfrage unzulĤssigerweise auf eine vereinfachte Sichtweise. In der Psychologie ist diese Debatte, auch in der deutschen Psychologie, wie gesagt schon einige Dekaden alt [6]. In der Physik ist man schon lange vom Signifikanztestritual abgegangen und verlangt, dass Effekte, sollen sie ernstgenommen werden, mindestens 5 Standardabweichungen vom Erwartungswert oder einer Kontrollmessung entfernt sein sollen, nimmt man eine Reihe von Messserien zusammen. Dass die Medizin weit weg von solchen Effekten ist, hat vor Kurzem der Editor von *Lancet* beklagt [7].

Dass die Medizin weit weg von solchen Effekten ist, hat vor Kurzem der Editor von Lancet beklagt.

Die Tatsache, dass Forscher, Herausgeber und Gutachter auf das leidige Signifikanzniveau fixiert sind, l $\tilde{A}$  $^{\mu}$ dt f $\tilde{A}$  $^{\mu}$ rmlich zu Missbrauch ein und hat u.a. auch zur Replikationskrise beigetragen, wie sie derzeit die Psychologie, aber auch die Medizin heimsucht [8, 9]. Denn mit ein bisschen Datenmassage  $\hat{a}$ ?? p-hacking wird das genannt  $\hat{a}$ ?? kann man rasch ein paar Outcomes vertauschen und einen Datensatz als  $\hat{a}$ ??signifikant $\hat{a}$ ?? beschreiben, der vielleicht alles andere als  $\tilde{A}$  $^{\mu}$ berzeugend ist und erzeugt damit falsch positive Daten, die andere nicht mehr replizieren k $\tilde{A}$  $^{\mu}$ nnen [10]. Weil aber nicht-replizierte Daten f $\tilde{A}$  $^{\mu}$ r die meisten Journals uninteressant sind und daher oftmals nicht, oder nur sehr obskur publiziert sind, entsteht ein Zerrbild unseres Wissensbestandes.

Der hier von Amrhein und Kollegen publizierte Aufruf sich vom p-Wert-Ritual zu verabschieden ist der radikalste Aufruf dieser Art, den ich bis jetzt wahrgenommen habe. Er fasst beinahe eine halbe Dekade Reflexion und Unzufriedenheit ýber die herrschenden Praktiken zusammen und bringt sie auf den Punkt. Die Argumente sind alle nicht neu: Wenn eine Studie, die einen behaupteten Effekt zu replizieren versucht an der Signifikanzgrenze scheitert, aber eine EffektschĤtzung vorlegt, die der bereits bekannten sehr Ĥhnlich ist, ist dann nicht der Effekt damit repliziert? Klar wĤre er das, wenn wir bereit wĤren, unser Denken etwas flexibler einzustellen und nicht nur auf statistische Signifikanz schielen.

Mit ein bisschen p-hacking kann man rasch ein paar Outcomes vertauschen und einen Datensatz als â??signifikantâ?? beschreiben.

Wenn ein unplausibel gro�er Effekt, der einmal als signifikant publiziert wurde, von anderen, auch gröÃ?eren Studien, nicht repliziert werden kann â?? ist dann der ursprünglich publizierte Effekt immer noch als gültig anzusehen? Eigentlich nicht. Auch wenn die Wahrscheinlichkeit klein ist, mit einer kleineren Studie einen Effekt statistisch signifikant zu bekommen, so ist sie doch nie Null. Laut statistischer Theorie kann es immer auch zufällige Ausreisser geben. Es muss also immer der Gesamtbestand der Daten, die Konsistenz der Befunde, die Konvergenz mit theoretischen Ã?berlegungen, die Robustheit des Effektes über verschiedene Bedingungen hinweg in Rechnung gestellt werden, bevor man einen Effekt als gegeben oder als nicht gegeben akzeptiert.

Das Signifikanz-Niveau ist bestenfalls einer unter vielen Kennwerten, der ber $\tilde{A}^{1}$ 4cksichtigt werden muss. Viel wichtiger sind die deskriptiven Werte: Wie gro $\tilde{A}$ ? ist der Effekt? Ist er unter gegebenen Umst $\tilde{A}$  $^{\mu}$ nden interessant oder wertvoll, oder nicht? Wie sieht  $\hat{a}$ ?? im Fall einer therapeutischen Fragestellung  $\hat{a}$ ?? das Verh $\tilde{A}$  $^{\mu}$ ltnis zwischen einem Effekt, seiner Gr $\tilde{A}$  $^{\mu}$ 6, den finanziellen und sozialen Kosten aus, etwa in Form von Risiken oder Nebenwirkungen. All diese Fragen kann die Feststellung einer Signifikanz nicht beantworten, sondern sie erfordert komplexe Evaluationen. Also, auf gut Deutsch,  $\tilde{A}$ ?berlegungen. Anstrengungen des Denkens, soziale Kommunikation.

Es wird Gruppen von Menschen geben, denen das gar nicht behagt. Zulassungsbehörden etwa. Die hatten es bis jetzt vergleichsweise leicht. Sie zählten die Anzahl der signifikanten Studien, und wenn diese gröÃ?er gleich zwei war, dann erteilten sie einer Substanz in aller Regel die Zulassung. Wenn man den Vorschlag von Amrhein und Kollegen ernst nimmt, wird es damit vorbei sein. Dann muss man GröÃ?e des Effekts, Kosten, Leichtigkeit, mit dem sich der Effekt nachweisen lässt, und Risiken in einem komplexen Kalkù⁄4l zusammen verarbeiten, bevor man zu einer Aussage kommen kann. Dann sind Kochbuch-Entscheidungen, ob in der Wissenschaft oder bei Behörden, auf jeden Fall schwerer. Aber warum sollen es Wissenschaftler und Behörden leichter haben in einer Welt, die eben komplex ist?

All diese Fragen kann die Feststellung einer Signifikanz nicht beantworten, sondern sie erfordert komplexe Evaluationen.

Man kann dem Aufruf von Amrhein und Kollegen nur w $\tilde{A}^{1}$ 4nschen, dass er  $\hat{a}$ ?? endlich  $\hat{a}$ ?? nachhaltige Konsequenzen hat. Das w $\tilde{A}^{1}$ 4rde dazu f $\tilde{A}^{1}$ 4hren, dass Autoren, Herausgeber und Gutachter sorgf $\tilde{A}^{m}$ 1tiger argumentieren und berichten; dass weniger k $\tilde{A}^{1}$ 4nstlich-dichotome Entscheidungen gef $\tilde{A}^{m}$ 1lt werden, sondern dass, wenn Entscheidungen gef $\tilde{A}^{m}$ 1lt werden m $\tilde{A}^{1}$ 4ssen, wie etwa bei einer Zulassung, diese auf komplexen, nachvollziehbaren und sozial verhandelten Begr $\tilde{A}^{m}$ 4ndungen aufruhen, von denen die statistische Signifikanz vielleicht eine S $\tilde{A}^{m}$ 2ule unter vielen sein wird.

## Quellen und Literatur

- [1] Rosnow, R. L., & Rosenthal, R. (2005). *Beginning Behavioral Research : A Conceptual Primer* (5th Edition ed.). Upper Saddle River, NJ: Pearson/ Prentice Hall.
- [2] Amrhein, V., Greenland, S., & MCShane, B. (2019). Retire statistical significance. Nature, 567, 305-307.
- [3] Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, *33*, 587-606. https://www.sciencedirect.com/science/article/abs/pii/S1053535704000927?via%3Dihub
- [4] Fisher, R. A. (1971 (orig. 1935)). The Design of Experiments. New York: Hafner.
- [5] Cohen, J. (1977). Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press. Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49, 997-1003.
- [6] Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: ErgĤnzungen und Alternativen. *Methods of Psychological Reserach Online*, *1*(4). <a href="https://www.dgps.de/fachgruppen/methoden/mpronline/issue1/art3/article.html">https://www.dgps.de/fachgruppen/methoden/mpronline/issue1/art3/article.html</a>

[7] Horton, R. (2015). Offline: What is medicineâ??s 5 sigma? *Lancet*, 385, 1380. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)60696-1/fulltext

[8] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349* (6251), aac4716. <a href="http://science.sciencemag.org/content/349/6251/aac4716">http://science.sciencemag.org/content/349/6251/aac4716</a>

[9] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

[10] Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359-1366. https://journals.sagepub.com/doi/full/10.1177/0956797611417632

## **Date Created**

27.03.2019