



Gen-ethischer Informationsdienst

Wie relevant ist ein signifikantes Ergebnis?

Wissenschaftliche Debatte um statistische Signifikanz

AutorIn

[Hans-Peter Piepho](#)



Auch statistisch „nicht signifikante“ Studien sollten in Meta-Analysen berücksichtigt werden. Foto: [Austin Community College/flickr.com](#) (CC BY 2.0)

Die wissenschaftliche Einordnung von gemessenen Ergebnissen als *signifikant* oder *nicht signifikant* wird in öffentlichen Debatten meist über- nommen. Doch wissenschaftsintern wird über die Methodik hinter dieser Einordnung viel diskutiert.

Bei der Zulassung von Arznei- und Pflanzenschutzmitteln spielt die Frage eine große Rolle, ob eine erwünschte Wirkung oder eine unerwünschte Nebenwirkung *signifikant* ist. Die am häufigsten verwendete Maßzahl ist hierbei der so genannte p-Wert und die Grenze von $p < 0,05$, um ein Ergebnis als

signifikant einzustufen. Zur Bedeutung von p-Werten gibt es immer wieder erhitzte Diskussionen, zuletzt befeuert durch einen *Nature*-Artikel¹, in dessen Folge sich die *American Statistical Association* zu einer offiziellen Stellungnahme veranlasst sah.² Die Debatte ist auch im Zusammenhang mit einer aktuell beklagten mangelnden Reproduzierbarkeit von Studienergebnissen zu sehen.³

Zunächst zur Frage, was ein p-Wert eigentlich ist. Nehmen wir als Beispiel einen Fütterungsversuch mit Mäusen, bei dem die Karzinogenität des Herbizids Glyphosat untersucht wird. Eine Gruppe von 100 zufällig ausgewählten Mäusen bekommt Futter, das mit Glyphosat versetzt ist, während eine Kontrollgruppe aus ebenfalls 100 Mäusen normales Futter ohne Glyphosat erhält. Nach einem festgelegten Zeitraum wird bestimmt, wie viele Mäuse in jeder Gruppe Krebs entwickelt haben. Die Mäuse stammen aus einer Population, die besonders krebsanfällig ist, was für das Aufspüren von Karzinogenität hilfreich ist. Wegen der großen Anfälligkeit gibt es allerdings auch in der Kontrollgruppe Mäuse, die Krebs bekommen. Nehmen wir nun an, in der Kontrollgruppe erkranken acht Mäuse, in der Glyphosatgruppe siebzehn Mäuse. Ist damit jetzt nachgewiesen, dass Glyphosat krebserregend ist? Oder könnte der beobachtete Unterschied auch auf Zufall beruhen? In anderen Worten, könnte es bei einer Wiederholung des Versuchs genau so gut sein, dass die Kontrollgruppe eine höhere Krebsrate zeigt? Hier hilft der p-Wert. Dazu nehmen wir zunächst an, dass die *Nullhypothese* gilt, also dass Glyphosat das Krebsrisiko nicht erhöht. In beiden Gruppen herrscht bei dieser Annahme die gleiche Wahrscheinlichkeit, an Krebs zu erkranken. Wenn das so ist, dann sind die beiden Gruppen austauschbar. Aufgrund von Zufallsschwankungen sind die beobachteten Anteile erkrankter Tiere aber meist nicht identisch. Wenn die Nullhypothese jedoch zutrifft, ist jeglicher numerische Unterschied in den beiden beobachteten Anzahlen von krebserkrankten Mäusen rein zufallsbedingt.

Man kann nun fragen, wie wahrscheinlich es bei *Gültigkeit der Nullhypothese* ist, rein zufällig eine Änderung der beobachteten Krebsrate von acht Prozent auf siebzehn Prozent zu erhalten oder eine noch stärkere Änderung. Diese Wahrscheinlichkeit ist der p-Wert. Es handelt sich also um eine *bedingte* Wahrscheinlichkeit, welche die Gültigkeit der Nullhypothese voraussetzt. Ist nun die Wahrscheinlichkeit für das beobachtete Ergebnis gering unter der Annahme der Nullhypothese, so kann man schließen, dass die Nullhypothese nicht plausibel ist und daher verworfen werden muss.

Im vorliegenden Fall beträgt der p-Wert $p = 0,0428$. Ist das nun klein genug, um die Nullhypothese zu verwerfen und zu folgern, dass Verfüttung von Glyphosat eine erhöhte Krebsrate zur Folge hat? Die gängige Antwort lautet „ja“, weil $p < 0,05$ ist.⁴ Wenn für $p < 0,05$ auf signifikante Unterschiede geschlossen wird, so beträgt die *Irrtumswahrscheinlichkeit* fünf Prozent. Dies bedeutet, dass bei Gültigkeit der Nullhypothese diese mit einer Wahrscheinlichkeit von fünf Prozent fälschlicherweise verworfen wird. An der Verwendung dieser starren Fünf-Prozent-Schwelle, die in vielen wissenschaftlichen Artikeln befolgt wird, scheiden sich allerdings die Geister. Diese Debatte kann hier nicht im Detail wiedergegeben werden, doch ich möchte zumindest auf die wichtigsten Faktoren in der Diskussion um die Aussagekraft des p-Wertes hinweisen.

Der p-Wert sagt nichts darüber, wie wahrscheinlich es ist, dass die Nullhypothese selbst zutrifft. Wie beschrieben ist er eine bedingte Wahrscheinlichkeit. So besagt der p-Wert von $p = 0,0428$ im obigen Beispiel *nicht*, dass die Nullhypothese nur mit einer Wahrscheinlichkeit von 4,28 Prozent zutrifft und Glyphosat mit einer Wahrscheinlichkeit von 95,72 Prozent für Mäuse krebserregend ist. Dies ist vielleicht enttäuschend, weil man genau diese Wahrscheinlichkeiten am liebsten wissen will. Aber klassische Signifikanztests erlauben solche Wahrscheinlichkeitsaussagen einfach nicht. Wer solche Aussagen machen will, muss sogenannte *Bayes*-Verfahren verwenden. Diese erfordern schon vor dem Versuch eine Einschätzung, wie wahrscheinlich es ist, dass Glyphosat unproblematisch ist, was wiederum Raum für Diskussion und Subjektivität bietet.⁵

Trotzdem ist der p-Wert ein sinnvolles Kriterium, um die Plausibilität der Nullhypothese zu prüfen. Es wird jedoch oft kritisiert, dass bei seiner Verwendung die Schwelle von fünf Prozent überbetont wird, bis dahin, dass nur noch angegeben wird, ob $p < 0,05$ ist oder nicht. Viel besser ist es, den p-Wert exakt

anzugeben. Dann kann sich jede Leserin selbst ein Bild davon machen wie signifikant das Ergebnis ist und auch eine andere Signifikanzschwelle als fünf Prozent anlegen. Außerdem vermeidet dies die starre Dichotomisierung in „signifikant“ und „nicht signifikant“ und erlaubt, ein Ergebnis von $p = 0,0428$ ähnlich zu werten wie ein Ergebnis von zum Beispiel $p = 0,0517$.

Signifikanz ist nicht dasselbe wie Relevanz. Ein p-Wert kann sehr klein sein, auch wenn der Gruppenunterschied so klein ist, dass er als irrelevant einzustufen ist, sofern gleichzeitig der Stichprobenumfang sehr groß ist. Mindestens genau so informativ wie ein p-Wert ist daher die Schätzung, wie stark sich zwei Gruppen im Krebsrisiko unterscheiden. Dies wird auch als *Effektstärke* bezeichnet. Außerdem ist ein Maß dafür wichtig, wie genau die Effektstärke geschätzt wurde. Hierzu dienen zum Beispiel *Vertrauensintervalle*. Bei einem Vertrauensintervall muss zwar auch eine Irrtumswahrscheinlichkeit festgelegt werden (üblicherweise ebenfalls fünf Prozent), aber das Intervall sagt, im Gegensatz zum p-Wert, wie groß der geschätzte Effekt ist und wie genau er geschätzt wurde.

Wie erwartet wird bei der Berechnung des p-Wertes die Gültigkeit der Nullhypothese vorausgesetzt. Was aber, wenn die Alternativhypothese zutrifft, es also einen echten Unterschied zwischen den verglichenen Gruppen (Kontrolle vs. Glyphosat) gibt? Mit welcher Wahrscheinlichkeit wird dies in einem Test bei $p < 0,05$ auch als signifikant erkannt? Diese Wahrscheinlichkeit heißt *Teststärke*. Sie ist wiederum eine Frage des tatsächlichen Gruppenunterschiedes und des Stichprobenumfanges. Je größer beide sind, um so eher lassen sich Unterschiede nachweisen. Leider wird oft kein ausreichender Stichprobenumfang verwendet, um relevante Unterschiede auch nachweisen zu können. Übliche Werte für die Teststärke, die angestrebt werden, sind 80 Prozent oder größer. Nehmen wir an, es besteht eine krebserregende Wirkung von Glyphosat und der Stichprobenumfang ist so gewählt, dass diese mit einer Teststärke von 80 Prozent nachgewiesen wird. Wie groß ist dann die Wahrscheinlichkeit, dass zwei unabhängige solche Studien beide ein signifikantes Ergebnis liefern, die erste Studie also durch die zweite reproduzierbar ist? Antwort: Nur 64 Prozent! Es ist also nichts Ungewöhnliches, wenn die Reproduzierbarkeit nicht sehr groß ist. Grund zur Besorgnis ist allerdings, wenn viele Studien eine geringe Teststärke haben, weil der Stichprobenumfang zu gering ist. Solche Studien sind eine Verschwendung von Ressourcen, weil von vornherein abzusehen ist, dass nicht viel dabei heraus kommen kann.

Ein nicht signifikanter Test beweist auch nicht die Gültigkeit der Nullhypothese! Wenn also in einer Studie kein signifikant erhöhtes Krebsrisiko durch Glyphosat gefunden wird, weil $p > 0,05$ ist, dann beweist das nicht, dass Glyphosat nicht krebserregend ist. Ein großer p-Wert, also ein nicht signifikantes Ergebnis, bedeutet streng genommen sogar, dass überhaupt keine Aussage getroffen werden kann („*Absence of evidence is not evidence of absence*“).⁶

Um nachzuweisen, dass Glyphosat unbedenklich ist, muss eine andere Nullhypothese definiert werden. Diese muss lauten: Glyphosat erhöht die Wahrscheinlichkeit, an Krebs zu erkranken um mindestens den Betrag X. Nur wenn diese Nullhypothese verworfen wird, besteht eine Basis, die Unbedenklichkeit zu reklamieren, wobei vorauszusetzen ist, dass eine Erhöhung des Krebsrisikos um den Betrag X auch tatsächlich als akzeptabel eingestuft wird.⁷ Oft wird aber unberechtigterweise ein nicht signifikanter Test der Nullhypothese „Glyphosat ist unbedenklich“ als Nachweis der Unbedenklichkeit eingestuft und dies meist auch so akzeptiert, obwohl das absoluter Unsinn ist.

Es werden meist mehrere Studien zur selben Fragestellung durchgeführt. Wichtiger als die Betrachtung einer einzigen Studie ist es, die Ergebnisse mehrerer Studien zusammenzufassen. Eine solche *Meta-Analyse* kann man auch mit p-Werten machen. Ein einfaches hypothetisches Beispiel zeigt, wie wertvoll dies ist. Nehmen wir an, fünf unabhängige Studien liefern jeweils einen p-Wert von $p = 0,10$. Kombinieren wir die fünf p-Werte jedoch in einer Meta-Analyse, so ergibt sich ein p-Wert von $p = 0,01$. Das Gesamtergebnis ist also signifikant, wenn die Schwelle $p < 0,05$ angelegt wird, und das, obwohl jede einzelne Studie für sich die Schwelle $p < 0,05$ verfehlt. Dies vielleicht überraschende Ergebnis liegt darin begründet, dass bei Gültigkeit der Nullhypothese jeder p-Wert zwischen 0 und 1 gleich wahrscheinlich ist, so dass eine Häufung von fünf so kleinen (oder kleineren) p-Werten sehr unwahrscheinlich ist (eben $p = 0,01$). Um

dieses Verfahren der Meta-Analyse anzuwenden, muss der exakte p-Wert in jeder einzelnen Studie angegeben sein. Noch besser ist es, die Schätzung der Effektstärke in jeder Studie anzugeben und diese Schätzungen dann in einer Meta-Analyse zu kombinieren.

Ein Hauptproblem bei der Verwendung von festen Signifikanzschwellen ist, dass viele Journale vorzugsweise solche Ergebnisse publizieren, die bei $p < 0,05$ signifikant sind. Und viele Autoren trauen sich erst gar nicht, ihre Studie einzureichen, wenn nicht $p < 0,05$ ist. Besonders problematisch ist es, wenn in einer großen Menge von Merkmalen so lange gesucht wird, bis ein signifikantes Ergebnis gefunden wird und dann nur dies publiziert wird. Dieses weitverbreitete und oft unbewusste Fehlverhalten wird auch als *p-hacking* bezeichnet.⁸ All dies führt zu einer Überschätzung von Effektstärken, leider auch in Meta-Analysen. Zu fordern ist daher unbedingt, dass auch „nicht signifikante“ Ergebnisse publiziert werden, weil nur so Verzerrungen vermieden werden, und weil jede Studie, auch die „nicht signifikanten“, einen wichtigen Beitrag zu einer Gesamtschätzung in einer Meta-Analyse leisten kann. Dieser wesentliche Kritikpunkt an der Verwendung von p-Werten bringt uns wieder an den Ausgangspunkt der Debatte zurück, in dessen Folge manche Journale die Verwendung von p-Werten sogar ganz verbannt haben, zu Gunsten einer ausschließlichen Fokussierung auf Effektstärken und Vertrauensintervalle. Das schießt sicher über das Ziel hinaus. Aber man sollte sich über die Grenzen von p-Werten im Klaren sein und wenn möglich neben exakten p-Werten immer auch Effektstärken und Vertrauensintervalle angeben, allein schon deshalb, weil dies für die Verwendung in einer aussagefähigen Meta-Analyse notwendig ist.

- ¹Nuzzo R. 2014. Scientific method: statistical errors. Nature 506:150- 152, doi: 10.1038/506150a.
- ²Wasserstein RL, Lazar NA. 2016. The ASA's statement on p-values: context, process, and purpose. The American Statistician 70:129-133, doi: 10.1080/00031305.2016.1154108.
- ³Amrhein V et al. 2017. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. PeerJ 5:e3544, doi: 10.7717/peerj.3544.
- ⁴Viele solche Studien werden hier vorgestellt: Burtcher-Schaden H. 2017. Die Akte Glyphosat. Kremayr & Scheriau, Wien.
- ⁵Greenland S et al. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31:337-350, doi: 10.1007/s10654-016-0149-3.
- ⁶Altman DG, Bland JM. 1995. Absence of evidence is not evidence of absence. British Medical Journal 311:485, doi: 10.1136/bmj.311. 7003.485.
- ⁷Piepho HP. 2013. Sicherheitsforschung: Signifikanz und Äquivalenz. GID 220, S. 28-29.
- ⁸Siehe Fußnote 1.

Informationen zur Veröffentlichung

Erschienen in:

GID Ausgabe 244 vom Februar 2018

Seite 14 - 16