https://www.gen-ethisches-netzwerk.de/node/3742

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 beru?cksichtigt werden. Foto: Austin Community College/flickr.com (CC BY 2.0)

Die wissenschaftliche Einordnung von gemessenen Ergebnissen als *signifikant* oder *nicht signifikant* wird in o?ffentlichen Debatten meist u?ber- nommen. Doch wissenschaftsintern wird u?ber die Methodik hinter dieser Einordnung viel diskutiert.

Bei der Zulassung von Arzneiund Pflanzenschutzmitteln spielt die Frage eine große Rolle, ob eine erwu?nschte Wirkung oder eine unerwu?nschte Nebenwirkung *signifikant* ist. Die am ha?ufigsten 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*-Artikel[1] , 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]

Zuna?chst zur Frage, was ein p-Wert eigentlich ist. Nehmen wir als Beispiel einen Fu?tterungsversuch mit Ma?usen, bei dem die Karzinogenita?t des Herbizids Glyphosat untersucht wird. Eine Gruppe von 100 zufa?llig ausgewa?hlten Ma?usen bekommt Futter, das mit Glyphosat versetzt ist, wa?hrend eine Kontrollgruppe aus ebenfalls 100 Ma?usen normales Futter ohne Glyphosat erha?lt. Nach einem festgelegten Zeitraum wird bestimmt, wie viele Ma?use in jeder Gruppe Krebs entwickelt haben. Die Ma?use stammen aus einer Population, die besonders krebsanfa?llig ist, was fu?r das Aufspu?ren von Karzinogenita?t hilfreich ist. Wegen der großen Anfa?lligkeit gibt es allerdings auch in der Kontrollgruppe Ma?use, die Krebs bekommen. Nehmen wir nun an, in der Kontrollgruppe erkranken acht Ma?use, in der Glyphosatgruppe siebzehn Ma?use. Ist damit jetzt nachgewiesen, dass Glyphosat krebserregend ist? Oder ko?nnte der beobachtete Unterschied auch auf Zufall beruhen? In anderen Worten, ko?nnte es bei einer Wiederholung des Versuchs genau so gut sein, dass die Kontrollgruppe eine ho?here Krebsrate zeigt? Hier hilft der p-Wert. Dazu nehmen wir zuna?chst an, dass die *Nullhypothese* gilt, also dass Glyphosat das Krebsrisiko nicht erho?ht. 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 Ma?usen rein zufallsbedingt.

Man kann nun fragen, wie wahrscheinlich es *bei Gu?ltigkeit der Nullhypothese* ist, rein zufa?llig eine A?nderung der beobachteten Krebsrate von acht Prozent auf siebzehn Prozent zu erhalten oder eine noch sta?rkere A?nderung. Diese Wahrscheinlichkeit ist der p-Wert. Es handelt sich also um eine *bedingte* Wahrscheinlichkeit, welche die Gu?ltigkeit der Nullhypothese voraussetzt. Ist nun die Wahrscheinlichkeit fu?r 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 betra?gt der p-Wert $p = 0,0428$. Ist das nun klein genug, um die Nullhypothese zu verwerfen und zu folgern, dass Verfu?tterung von Glyphosat eine erho?hte Krebsrate zur Folge hat? Die ga?ngige Antwort lautet „ja“, weil $p < 0,05$ ist.[4] Wenn fu?r $p < 0,05$ auf signifikante Unterschiede geschlossen wird, so betra?gt die *Irrtumswahrscheinlichkeit* fu?nf Prozent. Dies bedeutet, dass bei Gu?ltigkeit der Nullhypothese diese mit einer Wahrscheinlichkeit von fu?nf Prozent fa?lschlicherweise verworfen wird. An der Verwendung dieser starren Fu?nf-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 mo?chte zumindest auf die wichtigsten Faktoren in der Diskussion um die Aussagekraft des p-Wertes hinweisen.

Der p-Wert sagt nichts daru?ber, 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 fu?r Ma?use krebserregend ist. Dies ist vielleicht entta?uschend, 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 Einscha?tzung, wie wahrscheinlich es ist, dass Glyphosat unproblematisch ist, was wiederum Raum fu?r Diskussion und Subjektivita?t bietet.[5]

Trotzdem ist der p-Wert ein sinnvolles Kriterium, um die Plausibilita?t der Nullhypothese zu pru?fen. Es wird jedoch oft kritisiert, dass bei seiner Verwendung die Schwelle von fu?nf Prozent u?berbetont 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 fu?nf Prozent anlegen. Außerdem vermeidet dies die starre Dichotomisierung in „signifikant“ und „nicht signifikant“ und erlaubt, ein Ergebnis von $p = 0,0428$ a?hnlich 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 Scha?tzung, wie stark sich zwei Gruppen im Krebsrisiko unterscheiden. Dies wird auch als *Effektsta?rke* bezeichnet. Außerdem ist ein Maß dafu?r wichtig, wie genau die Effektsta?rke gescha?tzt wurde. Hierzu dienen zum Beispiel *Vertrauensintervalle*. Bei einem Vertrauensintervall muss zwar auch eine Irrtumswahrscheinlichkeit festgelegt werden (u?blicherweise ebenfalls fu?nf Prozent), aber das Intervall sagt, im Gegensatz zum p-Wert, wie groß der gescha?tzte Effekt ist und wie genau er gescha?tzt wurde.

Wie erwa?hnt wird bei der Berechnung des p-Wertes die Gu?ltigkeit der Nullhypothese vorausgesetzt. Was aber, wenn die Alternativhypothese zutrifft, es also einen ech- ten 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 *Teststa?rke*. Sie ist wiederum eine Frage des tatsa?chlichen Gruppenunterschiedes und des Stichprobenumfanges. Je gro?ßer beide sind, um so eher lassen sich Unterschiede nachweisen. Leider wird oft kein ausreichender Stichprobenumfang verwendet, um relevante Unterschiede auch nachweisen zu ko?nnen. U?bliche Werte fu?r die Teststa?rke, die angestrebt werden, sind 80 Prozent oder gro?ßer. Nehmen wir an, es besteht eine krebserregende Wirkung von Glyphosat und der Stichprobenumfang ist so gewa?hlt, dass diese mit einer Teststa?rke von 80 Prozent nachgewiesen wird. Wie groß ist dann die Wahrscheinlichkeit, dass zwei unabha?ngige solche Studien beide ein signifikantes Ergebnis liefern, die erste Studie also durch die zweite reproduzierbar ist? Antwort: Nur 64 Prozent! Es ist also nichts Ungewo?hnliches, wenn die Reproduzierbarkeit nicht sehr groß ist. Grund zur Besorgnis ist allerdings, wenn viele Studien eine geringe Teststa?rke haben, weil der Stichprobenumfang zu gering ist. Solche Studien sind eine Verschwendung von Ressourcen, weil von vornherein abzusehen ist, das nicht viel dabei heraus kommen kann.

Ein nicht signifikanter Test beweist auch nicht die Gu?ltigkeit der Nullhypothese! Wenn also in einer Studie kein signifikant erho?htes Krebsrisiko durch Glyphosat gefun- den 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 u?berhaupt 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 erho?ht 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 Erho?hung des Krebsrisikos um den Betrag X auch tatsa?chlich 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 durchgefu?hrt. 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, fu?nf unabha?ngige Studien liefern jeweils einen p-Wert von $p = 0,10$. Kombinieren wir die fu?nf 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 fu?r sich

die Schwelle p < 0,05 verfehlt. Dies vielleicht u?berraschende Ergebnis liegt darin begru?ndet, dass bei Gu?ltigkeit der Nullhypothese jeder p-Wert zwischen 0 und 1 gleich wahrscheinlich ist, so dass eine Ha?ufung von fu?nf so kleinen (oder kleineren) p-Werten sehr unwahrschein- lich ist (eben p = 0,01). Um dieses Verfahren der Meta-Analyse anwenden zu ko?nnen, muss der exakte p-Wert in je- der einzelnen Studie angegeben sein. Noch besser ist es, die Scha?tzung der Effektsta?rke in jeder Studie anzugeben und diese Scha?tzungen 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 fu?hrt zu einer U?berscha?tzung von Effektsta?rken, 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 Gesamteinscha?tzung in einer Meta-Analyse leisten kann. Dieser wesentliche Kritikpunkt an der Verwendung von p-Werten bringt uns wieder an den Ausgangspunkt der Debatte zuru?ck, in dessen Folge manche Journale die Verwendung von p-Werten sogar ganz verbannt haben, zu Gunsten einer ausschließlichen Fokussierung auf Effektsta?rken und Vertrauensintervalle. Das schießt sicher u?ber das Ziel hinaus. Aber man sollte sich u?ber die Grenzen von p-Werten im Klaren sein und wenn mo?glich neben exakten p-Werten immer auch Effektsta?rken und Vertrauensintervalle angeben, allein schon deshalb, weil dies fu?r die Verwendung in einer aussagefa?higen 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 thres- holds and the crisis of unreplicable research. PeerJ 5:e3544, doi: 10.7717/peerj.3544.
- 4Viele solche Studien werden hier vorgestellt: Burtscher-Schaden H. 2017. Die Akte Glyphosat. Kremayr & Scheriau, Wien.
- 5Greenland S et al. 2016. Statistical tests, P values, confidence inter- vals, 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 A?quivalenz. GID 220, S. 28-29.
- 8Siehe Fußnote 1.

Informationen zur Veröffentlichung

Erschienen in:
GID Ausgabe 244 vom Februar 2018
Seite 14 - 16