Signifikanztest als Gerichtsverfahren

Jörn Loviscach
www.j3L7h.de
Stand: 2022-02-06

Weil Irrtumswahrscheinlichkeit, p-Werte & Co. derart missverständlich sind, habe ich versucht, das gängige Bild eines Gerichtsverfahrens nicht nur als Einstieg zu nehmen, sondern möglichst viele Begriffe und Probleme an diesem Bild zu veranschaulichen *(im Folgenden kursiv gesetzt)*. Es geht in diesem Text erstens um die grundsätzliche Idee des Signifikanztests, zweitens darum, warum p-Werte nicht genug sagen oder sogar irreführend sind, drittens um fragwürdige Praktiken und viertens um Philosophisches.

Grundsätzliche Idee

Der Nullhypothesen-Signifikanztest (NHST) startet mit einer Nullhypothese H_0 (normalerweise: kein Effekt, kein Zusammenhang) und einer Alternativhypothese H_1. Diese Alternativhypothese ist das genaue logische Gegenteil der Nullhypothese. In der „wahren Welt“ (zur Philosophie siehe weiter unten) trifft für ein bestimmtes Experiment entweder H_0 oder aber H_1 zu.

*H_0: Die*Der Angeklagte X ist unschuldig im Sinne der Anklage.*

*H_1: Die*Der Angeklagte X ist schuldig im Sinne der Anklage.*

Das Problem ist nun, dass wir niemals perfekt sicher wissen können, welche dieser beiden Hypothesen stimmt. Wir können nur Indizien sammeln und bewerten. *Vielleicht haben die Zeug*innen alle eine Sehschwäche, die Spurensicherung hat einen schlechten Tag gehabt oder die*der Angeklagte X erinnert sich sogar selbst falsch an die Tat. Die Aufgabe des Gerichts ist es, trotz aller solcher Unwägbarkeiten ein Urteil zu fällen. Wenn die Beweismittel genügend stark gegen X sprechen, ist sie*er zu verurteilen. Wenn die Beweismittel zu schwach sind, ist sie*er freizusprechen: „Im Zweifel für die*den Angeklagte*n“ (Unschuldsvermutung).*

Also kann man vier Fälle unterscheiden:

- *X ist unschuldig und die Beweismittel sind allenfalls schwach: X wird **richtigerweise** freigesprochen.*
- *X ist schuldig und die Beweismittel sprechen genügend stark für ihre*seine Schuld: X wird **richtigerweise** verurteilt.*
- *X ist unschuldig, aber die Beweismittel sprechen unglücklicherweise genügend stark für ihre*seine Schuld: X wird **fälschlicherweise** verurteilt.* (Fehler 1. Art, „false positive“)
- *X ist schuldig, aber die Beweismittel sind allenfalls schwach: X wird **fälschlicherweise** freigesprochen.* (Fehler 2. Art, „false negative“)

Die Herausforderung ist nun, möglichst selten Fehler (1. Art oder 2. Art) zu machen.

Den Fehler 2. Art verhindert man mit einem banalen Trick: *Man erklärt niemals jemanden für unschuldig, sondern sagt bei jedem Freispruch, dass dies nur ein Freispruch aus Mangel an Beweisen sei.* Die Nullhypothese H_0 wird also niemals akzeptiert, sondern nur „nicht abgelehnt“. In dieser Situation hat man zu ungenaue Daten, zu viel Zufall. Also sind die beiden Entscheidungen nun:

- *Wenn die Beweismittel hinreichend stark sind: Verurteilung.* Das Ergebnis des Tests ist „statistisch signifikant“. Wir lehnen die Nullhypothese H_0 ab und akzeptieren deshalb die Alternativhypothese H_1.

- *Wenn die Beweismittel zu schwach sind: Freispruch mangels Beweisen.* Das Ergebnis des Tests ist „statistisch nicht signifikant“. Wir können die Nullhypothese H_0 nicht ablehnen und deshalb die Alternativhypothese H_1 nicht akzeptieren.

Mit diesem Trick (also Nullhypothese H_0 niemals zu akzeptieren) wird es unmöglich, einen Fehler 2. Art zu machen. Es bleibt aber der Fehler 1. Art. Mit diesem geht man raffinierter um; hier kommen insbesondere die Begriffe „p-Wert“ (unten mehr dazu) und „Irrtumswahrscheinlichkeit“ ins Spiel.

Die Irrtumswahrscheinlichkeit bezieht sich üblicherweise nur auf den Fehler 1. Art, aber nicht auf den Fehler 2. Art. Mit der Irrtumswahrscheinlichkeit ist dann gemeint: *Man betrachtet nur die Angeklagten, die wirklich unschuldig sind, und fragt, wie viel Prozent von denen verurteilt werden – also zu Unrecht verurteilt werden.* Die Irrtumswahrscheinlichkeit sagt also: Wie oft sprechen die Daten gegen die Nullhypothese – und zwar in den Fällen, in denen die Nullhypothese gilt?

Um in einer einzigen Zahl auszudrücken, *wie stark die Beweismittel für die Schuld von X sprechen*, benutzt man eine Teststatistik. Eine Teststatistik ist eine Funktion, welche die erhobenen Daten (die „Stichprobe“) nimmt und daraus einen einzigen Zahlenwert berechnet. Ist dieser Zahlenwert nahe 0, spricht das bei den meisten Teststatistiken für die Nullhypothese. Ist er dagegen besonders groß oder (je nach Teststatistik) besonders negativ, spricht das gegen die Nullhypothese.

Es gibt für verschiedene Anforderungen und Randbedingungen einen großen Katalog an verschiedenen Teststatistiken mit Namen wie t, F und χ^2. *Eine (erfundene) Teststatistik für unser Gerichtsverfahren könnte sein: Wie viele Zeugen glauben, gesehen zu haben, dass X in das Haus des Opfers gegangen ist?*

Konzept der p-Werte

Die Werte der jeweils gewählten Teststatistik sind nicht leicht zu deuten. Was bedeutet es zum Beispiel, wenn aus unserer Teststatistik der Wert 2,0 herauskommt? Obendrein ist es schwer, Studien *(Gerichtsfälle)* zu vergleichen, in denen jeweils verschiedene Teststatistiken benutzt worden sind. Also versucht man, eine gemeinsame Einheit für alle Teststatistiken zu finden. Dazu verwendet man üblicherweise die Größe mit dem Namen „p-Wert“.

Der p-Wert gibt an, wie extrem das Ergebnis der Teststatistik ist, *wie stark es also von dem abweicht, was man für eine*n Unschuldige*n erwarten würde.* Allerdings schwankt das Ergebnis der Teststatistik auch für eine*n Unschuldige*n. Deshalb drückt man den Unterschied mit Hilfe einer Wahrscheinlichkeit aus: Der p-Wert sagt, mit welcher Wahrscheinlichkeit die gewählte Teststatistik den gerade ermittelten oder einen noch extremeren Wert hat, falls die Nullhypothese gilt.

Zwei Punkte sind an dieser Definition besonders zu beachten:

- Man nimmt die Nullhypothese als Basis der Berechnung. *Es ist nämlich schwer bis unmöglich, genauer zu sagen, wie sich die Teststatistik verhält, wenn die*der Verdächtige tatsächlich schuldig ist.*
- Man berechnet nicht die Wahrscheinlichkeit dafür, exakt den Wert wie im aktuellen Fall zu haben. Diese Wahrscheinlichkeit wäre wenig aussagekräftig und für die meisten Teststatistiken sogar schlichtweg null. (Entsprechung: Wie groß ist die Wahrscheinlichkeit, auf dem digitalen Thermometer eine Temperatur von genau 12,34 °C zu sehen? Winzig!) Statt nach dem einen exakten Wert schaut man nach allen Werten, die so sind oder extremer. (Entsprechung: Wie groß ist die Wahrscheinlichkeit, auf dem Thermometer eine Temperatur von 12,34 °C oder mehr zu sehen?)

Der p-Wert wird aus der Teststatistik berechnet; die Teststatistik wird vorher aus den Daten des Experiments berechnet. Diese Daten unterliegen dem Zufall. Also ist auch der p-Wert vom Zufall abhängig. Er ist keine feste Wahrscheinlichkeit. Ganz im Gegenteil: Wenn man alle Experimente betrachtet, in denen die Nullhypothese tatsächlich gilt, streuen die dabei ermittelten p-Werte gleichmäßig über den kompletten möglichen Bereich von 0 bis 1, also 0 % bis 100 %.

Das ist überraschend, aber doch zu verstehen: Angenommen, dass die Nullhypothese gilt. Wenn der p-Wert gleich 0,23 ist, ist die Teststatistik in 23 % der Fälle mindestens so extrem wie ihr aktueller Wert. Wenn der p-Wert gleich 0,24 wäre *(eine etwas schwächere Beweislage)*, wäre die Teststatistik in 24 % der Fälle mindestens so extrem wie ein Wert, der ein wenig schwächer ist als der aktuelle Wert. Diese 24 % der Fälle mit ihrer etwas schwächeren Bedingung umfassen die gesamten ersten 23 % der Fälle. Nun kann man sich das 1 % der Fälle ansehen, die zwar zu den 24 % gehören, aber nicht zu den 23 %. Genau in diesem 1 % der Fälle liegt der p-Wert also zwischen 0,23 und 0,24. Entsprechend für andere Werte, zum Beispiel zwischen 0,75 und 0,76. Jedes Prozent der p-Werte kommt also in einem Prozent der Fälle vor (wenn die Nullhypothese gilt): eine Gleichverteilung (wenn die Nullhypothese gilt).

Die Gleichverteilung der p-Werte, wenn die Nullhypothese gilt, lässt sich benutzen, um seltener einen Fehler 1. Art zu begehen. Man legt dazu ein „Signifikanzniveau“ fest, traditionell 5 %. Wenn der p-Wert unterhalb dieses Signifikanzniveaus von 0,05 = 5 % liegt, lehnt man die Nullhypothese ab, sonst nicht. Dann macht man nur in 5 % der Fälle, in denen die Nullhypothese gilt, einen Fehler 1. Art. Denn, wenn die Nullhypothese gilt, ist der p-Wert gleichverteilt, liegt also in 5 % der Fälle zwischen 0,00 und 0,05, gilt als in diesen 5 % der Fälle als „signifikant“. Das Signifikanzniveau gibt damit automatisch die Irrtumswahrscheinlichkeit an.

Wenn allerdings die Nullhypothese nicht gilt, verschiebt sich die Verteilung der Werte der Teststatistik typischerweise[1] hin zu extremeren Werten, so dass sich die Verteilung der p-Werte zu kleineren p-Werten verschiebt – auch wenn man oft nichts über die Details dieser Verschiebung sagen kann.

Wenn man als Entscheidungsregel nicht „Lehne die Nullhypothese ab, wenn der p-Wert kleiner als 0,05 ist“ nehmen würde, sondern „Lehne die Nullhypothese ab, wenn der p-Wert zwischen 0,70 und 0,75 liegt“, würde die Irrtumswahrscheinlichkeit ebenfalls 5 % betragen. Dass man den Bereich von 0,00 bis 0,05 statt etwa den von 0,70 bis 0,75 benutzt, hat also nicht mit der Nullhypothese H_0 zu tun, sondern mit der Alternativhypothese H_1, unter der sich die Verteilung der p-Werte typischerweise nach unten verschiebt (siehe den vorigen Absatz).

Irreführung durch p-Werte

Statische Signifikanz und praktische Relevanz sind oft nicht gleichbedeutend:

- Ein statistisch signifikantes Ergebnis kann praktisch irrelevant sein. *Die Videoüberwachung zeigt mit hoher Sicherheit, wie Y in der Gemeinschaftsküche ein Plätzchen stiehlt.*
- Umgekehrt kann ein statistisch nicht signifikantes Ergebnis praktisch höchst relevant sein. *Person Z ist auf unklare Art in einen Mordfall verwickelt, wird freigesprochen; es stellt sich*

1 Wenn die Nullhypothese nicht stimmt und obendrein weitere, technische Bedingungen verletzt sind (zum Beispiel, wenn beim t-Test keine Normalverteilung vorliegt), ist schwer zu sagen, was mit der Verteilung der Teststatistik und folglich mit der Verteilung der p-Werte passiert. Die Frage ist, wie robust („robust“ ist der Fachbegriff dafür) die jeweilige Teststatistik gegenüber solchen Abweichungen ist. Die besagten „weiteren, technischen Bedingungen“ kann man auch als Bestandteil der Nullhypothese interpretieren („Die Verteilung hat den Erwartungswert 0 *und* ist normal.“), aber dann weiß man gar nichts mehr über den Fall, dass die Nullhypothese nicht gilt.

aber die Frage, ob es nicht zu gefährlich ist, Z mit drei anderen Leuten auf eine Reise zum Mars zu schicken.

Viele Studien, in denen der p-Wert angewendet wird, untersuchen bloß die Nullhypothese H_0, nie die Alternativhypothese H_1. Das ist bequem, weil die Nullhypothese meist viel leichter mit Hilfe von Wahrscheinlichkeiten zu beschreiben ist. Nicht die Alternativhypothese zu betrachten, verzerrt aber die Wahrnehmung. *Angenommen, wir haben 100 Angeklagte, davon 50 tatsächlich Schuldige und 50 tatsächlich Unschuldige. Bei einer Irrtumswahrscheinlichkeit von 5 % werden etwa zwei bis drei Unschuldige verurteilt – und hoffentlich viel mehr Schuldige verurteilt, aber deren Zahl können wir mit diesen Angaben nicht schätzen. Krasser angenommen, wir haben wieder 100 Angeklagte, davon aber nun 10 tatsächlich Schuldige und 90 tatsächlich Unschuldige. Bei einer Irrtumswahrscheinlichkeit von 5 % werden etwa vier bis fünf Unschuldige verurteilt – und maximal alle zehn Schuldigen, vielleicht auch nur fünf Schuldige. Mit anderen Worten: Es könnte sein, dass genauso viele Schuldige wie Unschuldige verurteilt werden.* Diesen dramatischen Effekt müsste man berücksichtigen. Das passiert allerdings in viel zu wenigen Studien. (Stichwörter für Interessierte: Teststärke, A-priori-Verteilung)

p-Werte beantworten die oft gar nicht interessante Frage: Wie oft sieht man solche oder extremere Daten, wenn die Nullhypothese gilt? *Wie häufig sprechen die Beweismittel zufällig so schlimm oder schlimmer gegen eine*n unschuldige*n Angeklagte*n?* Die üblicherweise viel interessantere, aber vom p-Wert nicht beantwortete Frage lautet dagegen: Wie wahrscheinlich ist es bei diesen Daten, dass die Alternativhypothese gilt? *Wie wahrscheinlich ist es angesichts der vorgelegten Beweismittel, dass X schuldig ist?* Die letztere Frage ist allerdings nicht leicht zu beantworten; obendrein setzt sie typischerweise eine andere Philosophie voraus (Bayes, siehe unten). Die Arbeit mit p-Werten erinnert deshalb an die*den Betrunkene*n, die*der den Haustürschlüssel nachts im Gebüsch verloren hat, ihn aber unter der Straßenlaterne sucht, weil er sich dort besser finden lässt.

Zumindest in den Sozialwissenschaften kann man davon ausgehen, dass die Nullhypothese sogar nie gilt, denn irgendeinen – wenn auch winzigen – Einfluss wird es immer geben. Man lernt also durch den Hypothesentest eigentlich nur, ob die Modellvorstellung, die man hier testet, gut genug für die Datenlage ist.

Fragwürdige Praktiken

Die augenfälligste fragwürdige Praxis ist die, dass vorwiegend Studien publiziert werden, die ein statistisch signifikantes Ergebnis zeigen, in denen also die Nullhypothese abgelehnt wird. Dies ist eine wesentliche Ursache für das „Publication Bias“: *Jemand wird wegen ihrer*seiner grünen Haare sehr oft verdächtigt. Sie*Er steht in 20 Fällen vor Gericht, wird fast immer freigesprochen, aber einmal versehentlich zu Unrecht verurteilt. Nur dieses eine Urteil steht in der Zeitung. Die Öffentlichkeit hält die*ihn deshalb für schuldig.*

Inzwischen gibt es mehr und mehr Aktivitäten, um das Publication Bias zu verringern. Wesentlich dazu ist, dass auch die Studien, deren Testergebnisse nicht signifikant sind, öffentlich bekannt werden. Besser noch ist, wenn eine Veröffentlichung nicht davon abhängt, ob das Testergebnis signifikant ist oder nicht. Denn dann besteht auch weniger Versuchung, die p-Werte in Richtung Signifikanz zu tricksen (absichtlich oder unabsichtlich). Im massiven Umfang über das Fehlen von statistischer Signifikanz zu berichten, widerspricht allerdings der traditionellen Vorstellung von Veröffentlichungen als Berichten über sensationelle Entdeckungen.

Die fadenscheinigen Tricks, um (scheinbar) signifikante p-Werte zu erhalten („p-hacking“), laufen unter der Bezeichnung „fragwürdige Forschungspraktiken“ („questionable research practices“). Hier sind einige davon:

- *Verhafte irgendjemanden und befrage die Zeug*innen von 100 ungeklärten Straffällen, ob sie sie*ihn wiederzuerkennen glauben. Das wird wohl versehentlich bei mindestens einem jener Straffälle gelingen. Verurteile sie*ihn für diesen einen Fall und erzähle niemanden etwas von den 99 anderen Fällen* („selective reporting of dependent variables"). Wenn man 100 voneinander unabhängige Tests mit einer Irrtumswahrscheinlichkeit von jeweils 5 % macht, beträgt die Wahrscheinlichkeit dafür, dass mindestens einer dieser 100 Tests zufällig signifikant ist, obwohl für alle die Nullhypothese gilt, satte 99,4 %. (Stichwort: Bonferroni-Korrektur)
- *Wenn die*der Zeug*in Z die*den Verdächtige*n belastet, aber die Zeug*innen A bis Y die*den Verdächtige*n entlasten, sorge dafür, dass nur Z gehört wird* („only report the experimental condition that worked").
- *Verhafte irgendjemanden auf der Straße und prüfe, ob sie*er für jeden Zeitpunkt belegen kann, dass sie*er dort etwas rechtlich Einwandfreies getan hat.* Statt vorher feste Hypothesen aufzustellen, kann man kann sich von den Ergebnissen seiner Beobachtungen oder Umfragen inspirieren lassen und nachträglich passende Hypothesen aufstellen, ohne zu verraten, dass diese im Nachhinein entstanden sind („HARKing = hypothesizing after the results are known"). Damit lässt sich Abstruses scheinbar belegen: „Alle in geraden Monaten geborenen Menschen, deren Geburtsdatum eine durch 7 teilbare Quersumme hat, fühlen sich glücklicher." (Oder welche Kombination von Zahlen gerade mit den Daten hinkommt.)
- *Höre keine Zeug*innen mehr an, sobald die bisher gehörten Zeug*innen die*den Angeklagte*n stark belastet haben* („data peeking").
- Probiere andere Teststatistiken aus; lege dich insbesondere nicht vorher auf eine bestimmte fest. Probiere andere Auswertungsmethoden aus, zum Beispiel zum Aussortieren von (nur vermeintlichen?) Ausreißern. Irgendetwas wird schon zufällig eine Signifikanz zeigen („garden of forking paths").
- Verwende Teststatistiken, deren Modellannahmen verletzt sind, so dass sie im jeweiligen Fall eigentlich nicht angewendet werden dürfen. Zum Beispiel verlangt der t-Test – das wohl gängigste Verfahren – normalverteilte Daten oder aber eine hinreichende Größe der Stichprobe.
- *Wenn die Beweismittel gegen X nicht ausreichen, schaue noch genauer nach. Zumindest einen Verstoß gegen die Straßenverkehrsordnung wird man X schon anhängen können.* Die exakte Nullhypothese ist nur selten wirklich wahr. Man muss meist nur noch mehr Versuchspersonen befragen oder Experimente machen, um doch eine Signifikanz zu sehen.

Philosophisches

Eine grundlegende Frage ist, ob man annehmen will, dass es eine einzige „wahre Welt" gibt. (Stichwort: Realismus)

Wahrscheinlichkeiten kann man als relative Häufigkeiten verstehen („frequentistisch"). Beispiel: In 0,3 % der Spiele gewinne ich. Aber man kann Wahrscheinlichkeiten alternativ auch als Maß dafür verstehen, wie sicher man sich ist („Bayes"). Beispiel: Ich bin mir zu 80 % sicher, dass ich diese Klausur schaffe. In der Bayesschen Denkweise ist es kein Problem, über die Wahrscheinlichkeit zum Beispiel der Nullhypothese nachzudenken. Der übliche Nullhypothesen-Signifikanztest ist allerdings frequentistisch.

Vielen Dank an Martin Hovekamp für seine Korrekturvorschläge.