Die Interpretation statistischer Signifikanz +

Ian Birnbaum

übersetzt von Günter Fillbrunn

Das Signifikanzniveau von 5% bedeutet, daß im Durchschnitt in 5 von 100 Fällen, in denen wir die Nullhypothese ablehnen, falsch liegen.

Dieses schrieb einer meiner Schüler als Antwort auf die Frage nach der Interpretation des Signifikanzniveaus. Andere antworteten in einer ähnlichen, wenn auch weniger prägnanten Weise. Alle aus der Gruppe stimmten darin überein, daß die zitierte Antwort sehr vernünftig sei.
Sie ist natürlich falsch, aber ich frage mich, wieviele Schüler von ihrer Richtigkeit überzeugt sind. Die korrekte Antwort lautet:

Das Signifikanzniveau von 5% bedeutet, daß wir bei wahrer Nullhypothese im Durchschnitt in 5 von 100 Fällen diese ablehnen werden.

Diese Darstellung des Sachverhalts erweist sich jedoch nicht als sehr überzeugende Widerlegung der ersten Aussage. Sie erhellt auch nicht den Zusammenhang beider. Um dies zu erreichen, konstruieren wir ein Beispiel, aus dem sich alle wichtigen Punkte ergeben.
Und nun das Beispiel:

Ein Zauberer benutzt für einen Trick zwei Münzen. Die eine zeigt mit der Wahrscheinlichkeit 0,7 Wappen, während bei der anderen ebenfalls mit der Wahrscheinlichkeit 0,7 Zahl fällt. Unglücklicherweise hat er eine Münze verloren. Obgleich er nicht weiß welche, vermutet er aufgrund seiner Zauberer-Eingebung, daß es diejenige ist, die bevorzugt Zahl zeigt. Da er infolge des Verlusts den Trick ändern muß, möchte er wissen, welche Münze er noch hat. Dazu testet er die Nullhypothese

H_o: Die noch vorhandene Münze zeigt bevorzugt Wappen

gegen die Alternativhypothese

H_1: Die noch vorhandene Münze zeigt bevorzugt Zahl

mit einer Stichprobe vom Umfang 10. Er verwirft genau dann H_o, wenn er höchstens viermal Wappen wirft.

1. Wie groß ist das Signifikanzniveau α (also wie groß ist die Wahrscheinlichkeit, daß die Nullhypothese abgelehnt wird, obwohl sie wahr ist)? (0,04735)

+ Originaltitel in 'TEACHING STATISTICS' (1982) Heft 1, Band 4 'Interpreting Statistical Significance'

2. Wie groß ist die Wahrscheinlichkeit β für den Fehler 2. Art (d.h. wie groß ist die Wahrscheinlichkeit für die Annahme der Nullhypothese, vorausgesetzt, daß diese falsch ist)? (0,15027)
3. Wir stellen uns nun das in der Praxis höchst unwahrscheinliche Ereignis vor, daß 2000 Zauberer in der gleichen mißlichen Lage sind und daß die Zauberer-Eingebung nichts als ein reines Ratespiel ist, so daß in 1000 Fällen die Nullhypothese wahr und in den anderen 1000 falsch ist.
   a) Wieviele aus der ersten Gruppe (das ist die Gruppe mit wahrer Nullhypothese) werden die Nullhypothese im Durchschnitt ablehnen? ($1000 \cdot 0{,}04735 = 47{,}35$)
   b) Wieviele aus der zweiten Gruppe werden die Nullhypothese im Durchschnitt ablehnen? ($1000 \cdot (1-0{,}15027) = 849{,}73$)
   c) Benutze die Antworten zu a) und b), um zu ermitteln, wie hoch in den Fällen, in denen die Nullhypothese abgelehnt wird, der erwartete Anteil der falschen Entscheidungen ist. ($\frac{47{,}35}{47{,}35+849{,}73} = 0{,}053$)
   d) Wieviele aus der ersten Gruppe werden die Nullhypothese im Durchschnitt annehmen? ($1000 \cdot (1-0{,}04735) = 952{,}65$)
   e) Wieviele aus der zweiten Gruppe werden die Nullhypothese im Durchschnitt annehmen? ($1000 \cdot 0{,}15027 = 150{,}27$)
   f) Benutze die Antworten zu d),e), um zu bestimmen, wie hoch in den Fällen, in denen die Nullhypothese angenommen wird, der erwartete Anteil der falschen Entscheidungen ist. ($\frac{150{,}27}{952{,}65+150{,}27} = 0{,}14$)
4. Stelle dir nun vor, daß an der Zauberer-Eingebung etwas dran ist, so daß in 1950 Fällen die Nullhypothese wahr und in 50 Fällen falsch ist. Wiederhole nun die Aufgaben a)-f) von 3. (a) $1950 \cdot 0{,}04735 = 92{,}3325$; b) $50 \cdot (1-0{,}15027) = 42{,}4865$; c) $\frac{92{,}3325}{92{,}3325+42{,}4865} = 0{,}68$, also über $\frac{2}{3}$ aller Ablehnungen der Nullhypothese sind falsch; d) $1950 \cdot (1-0{,}04735) = 1857{,}6675$; e) $50 \cdot 0{,}15027 = 7{,}5135$; f) $\frac{7{,}5135}{1857{,}6675+7{,}5135} = 0{,}004$, also nur 40 von 10000 Annahmen der Nullhypothese sind falsch)
5. Gehe nun davon aus, daß die Zauberer-Eingebung erschreckend schlecht ist, so daß die Nullhypothese in 50 Fällen wahr und in 1950 Fällen falsch ist. Wiederhole abermals a)-f) von 3.

(a) $50 \cdot 0{,}04735 = 2{,}3675$; b) $1950 \cdot (1-0{,}15027) = 1656{,}9735$;

c) $\frac{2{,}3675}{2{,}3675+1656{,}9735} = 0{,}0014$, **somit sind lediglich 14 von** 10000 Ablehnungen der Nullhypothese falsch; d) $50 \cdot (1-0{,}04735) = 47{,}6325$; e) $1950 \cdot 0{,}15027 = 293{,}0265$; f) $\frac{293{,}0265}{47{,}6325+293{,}0265} = 0{,}86$, **was besagt, daß mehr als** $\frac{5}{6}$ aller Annahmen der Nullhypothese **falsch sind)**

Es dürfte für Schüler lehrreich sein, diese Fragen durchzuarbeiten. Die Lösungen mögen sie überraschen, aber sie sollten daraus nicht schließen, daß die Begriffe Signifikanzniveau und Risiko 2. Art schlecht sind. Die Information über die in Frage kommende Verteilung der wahren oder falschen Hypothesen, die für die Berechnungen der Anteile in c) und f) benötigt wird, sind in der Praxis selten erhältlich. Von Vorteil ist, daß zur Bestimmung der Wahrscheinlichkeiten α, β diese Kenntnis nicht erforderlich ist. Diese wurden zuerst eingeführt (von Neyman und Pearson), um eine statistische Folgerung in den Fällen zu erlauben, in denen ein 'gültiger' Bayes'scher Schluß unmöglich ist. (Bayesianer werden das Wort 'gültig' anzweifeln, aber vom Häufigkeitsstandpunkt ist Bayes'sches Schließen nur dann gültig, wenn eine wahre Verteilung der Hypothesen existiert, zu der die untersuchte gehört; subjektive Verteilungen sind unzulässig. Die obige Zaubereraufgabe ist ein Beispiel für eine gültige Bayes'sche Schlußweise.)
Das Standard-'Gedankenexperiment' für Signifikanztests ist das der wiederholten Stichprobenentnahme. So werden beim 5%-Signifikanzniveau bei wahrer Nullhypothese im Durchschnitt 5 von 100 Tests, die mit der gleichen Grundgesamtheit durchgeführt werden, eine Ablehnung dieser Hypothese zur Folge haben. Analog werden bei falscher Nullhypothese und bei einem Risiko 2. Art von 19% durchschnittlich 81 von 100 mit der gleichen Grundgesamtheit durchgeführten Tests zu einer Ablehnung der Nullhypothese führen. Entsprechend können wir sagen, daß bei 100 unabhängigen Tests, die mit verschiedenen Grundgesamtheiten durchgeführt werden und deren Nullhypothesen alle wahr sind, durchschnittlich 5 Ablehnungen von H_o vorkommen werden. Bei 100 unabhängigen Tests mit verschiedenen Grundgesamtheiten, deren Alternativhypothesen allesamt wahr sind, wird es bei β=19% im Schnitt zu 81 Ablehnungen von H_o kommen. Bei keiner dieser Feststellungen müssen wir irgend etwas über die Verteilung der wahren Hypothesen in dieser Testreihe wissen. Wenn wir jedoch weiter gehen können und die

einzelne Nullhypothese, die getestet wird, einer Klasse von Hypothesen zuordnen können (jede mit einer spezifischen Grundgesamtheit und mindestens im Prinzip testbar) und wenn wir den Anteil der Hypothesen, die wahr sind, kennen (oder zuverlässig schätzen können), dann und nur dann können wir die obigen Wahrscheinlichkeiten in c) und f) ausrechnen. Die Risiken α und β behalten natürlich ihren ursprünglichen Sinn als Wahrscheinlichkeiten, die unter der Annahme, daß die Nullhypothese wahr bzw. falsch ist, zutreffen. Sie stellen die Fehlerwahrscheinlichkeit dar, die von der Voraussetzung abhängt. Wir können aber auch die Fehlerwahrscheinlichkeit berechnen, die von der im Test fälligen Entscheidung abhängt, nämlich die Wahrscheinlichkeit in c), wenn wir die Nullhypothese ablehnen, und die in f), wenn wir sie annehmen. Dies sind die Wahrscheinlichkeiten, die meine Schüler für die Risiken α und β hielten. Obwohl das Zauberer-Beispiel künstlich ist, klärt es dieses Durcheinander.