

Was bedeutet Signifikanz?

eine empirische Semesterarbeit

Abstract

Es wurde untersucht, wie akademische Psychologen statistische Signifikanz interpretieren. Es wurde ein Fragebogen verwendet, der verschiedene Aussagen über die Bedeutung eines signifikanten Ergebnisses enthielt. Die Probanden sollten entscheiden, welche dieser Aussagen sich aus der Prämisse, ein mit $p = 0.01$ signifikantes Ergebnis erhalten zu haben, logisch ableiten lassen. Keine der 6 Aussagen folgte aus den Prämissen, doch 92% der Befragten zeigte Fehlinterpretationen, indem sie mindestens eine davon bestätigten. Die Stichprobe setzte sich zum größten Teil aus Studenten und Dozenten der Freien Universität Berlin zusammen. Die vorliegende Studie ist die Replikation einer 1986 von Michael Oakes in den USA durchgeführten Erhebung.

Inhalt

2 Theorie	1
2.1 Die Geschichte der Signifikanztests	1
2.2 Fisher	2
2.3 Neyman & Pearson	2
2.4 Hybrid-Logik	2
2.5 Fehlinterpretationen	3
3 Die Originalstudie	3
3.1 Diskussion der einzelne Aussagen:	4
3.2 Exkurs: Umgekehrte Wahrscheinlichkeit (Bayes)	4
4 Erhebung	5
4.1 Teilnehmer:	5
4.2 Vorgehen:	5
5 Ergebnisse	6
6 Diskussion	7
6.1 Ausblick	8
Kontakt	8
Literaturverzeichnis	8
Anhang: Fragebogen	8

1 Einleitung

Nur 2 von 70 befragten akademischen Psychologen hatten in der 1986 von Michael Oakes in den USA durchgeführten Untersuchung keinen Fehler gemacht, als es darum ging, zu entscheiden, was ein signifikantes Ergebnis zu bedeuten habe. Die vorliegende Untersuchung, die durch Prof. Dr. Gigerenzer angeregt und betreut wurde, hat zum Ziel, den Stand der Kenntnis 13 Jahre später und in Deutschland zu erkunden.

Dazu soll zunächst die Geschichte der Signifikanztests angerissen, und die beiden widerstreitenden Ansätze vorgestellt werden, aus denen die heute weitläufig praktizierte Form des Nullhypothesentests hervorgegangen ist. In Abschnitt 3 wird die Originalstudie von Oakes und deren Ergebnisse dargestellt, sowie die einzelnen Aussagen diskutiert. Es folgen ein kurzer Abriß des eigenen Vorgehens, und in Abschnitt 5 die Ergebnisse der vorliegenden Studie. Im Anschluß daran (Abschnitt 6) sollen diese Ergebnisse diskutiert, und mögliche Interpretationen erörtert werden.

2 Theorie

Das Prinzip von Signifikanztests ist das folgende: um über das Zutreffen bestimmter Hypothesen entscheiden zu können, erhebt man entsprechende Daten, und berechnet dann die Wahrscheinlichkeit, solche Daten zu bekommen, falls die sogenannte Nullhypothese gilt. Ist diese Wahrscheinlichkeit sehr klein, verwirft man die Nullhypothese und entscheidet sich stattdessen für die Alternativhypothese. Man prüft also $p(D|H_0)$, die Wahrscheinlichkeit der Daten unter der Nullhypothese. Aus diesem Grund wird dieses Verfahren auch als „Nullhypothesentest“ bezeichnet. Auf den ersten Blick erscheint das sinnvoll, denn wenn es sehr unwahrscheinlich ist, solche Daten zu erhalten, falls die Null Hypothese gilt, dann wird diese wohl falsch sein, und damit die Alternativhypothese richtig. Bei näherem Hinsehen zeigt sich allerdings, daß dieser Schluß nicht gerechtfertigt ist, da man, wie wir sehen werden, nicht ohne weiteres von der Wahrscheinlichkeit der Daten auf die Wahrscheinlichkeit von Hypothesen schließen kann.

Sehen wir uns dazu zunächst die Entwicklungsgeschichte der heute gängigen Signifikanztests an.

2.1 Die Geschichte der Signifikanztests

Die Geschichte des Nullhypothesentests geht weit zurück. Bereits 1710 versuchte John Arbuthnot die Existenz eines allmächtigen Gottes nachzuweisen. Er stellte anhand des Londoner Geburtenregisters der Jahre 1629 - 1710 fest, daß in jedem dieser Jahre aufs neue mehr Jungen als Mädchen geboren wurden. Er berechnete die Wahrscheinlichkeit, daß während diesen 82 Jahren jedes Jahr „zufällig“ mehr Jungen zur Welt kommen. Diese kombinierte Wahrscheinlichkeit $(\frac{1}{2})^{82}$ ist verschwindend klein. Da es also kaum Zufall sein kann, argumentierte er, müsse es der allmächtige Gott sein, der den Menschen Jahr für Jahr mehr Männer schickt, weil doch bei den damaligen Lebensumständen mehr Männer starben als Frauen (Arbuthnot, 1710).

Ihren breiten Einzug in die Wissenschaft hielten die Signifikanztests allerdings erst im zwanzigsten Jahrhundert. So wurde vor hundert Jahren von Karl Pearson mit dem Chi-Quadrat Test der erste Signifikanztest entwickelt, der noch heute häufig Anwendung findet.

Es ist ein Test, der angibt, mit welcher Wahrscheinlichkeit die Abweichung einer beobachteten Häufigkeitsverteilung von einer erwarteten Verteilung durch Zufall entstanden ist.

Als den ersten großen Vertreter des Nullhypothesentestens kann man wohl getrost Ronald A. Fisher anführen.

2.2 Fisher

Sir Ronald Fisher, der sich bis dahin mit der Inferenzproblematik in der Agrarwissenschaft beschäftigte, veröffentlichte 1925 sein erstes Buch über statistische Methodenlehre („Statistical Methods for Research Workers“), und zehn Jahre später sein umfang- und einflußreiches Werk *The Design of Experiments* (1935), das systematische Anleitungen zum experimentellen Vorgehen enthielt.

Bei Fishers Paradigma geht man von einer Normalverteilung bestimmter Stichprobenkennwerte (z.B. Mittelwerte) aus, die theoretisch entstehen würde, zöge man bei zutreffender Nullhypothese beliebig viele Stichproben, da diese durch das Einwirken von Zufall normalverteilt um den wahren Populationsmittelwert streuen. Vergleicht man nun den experimentell erhaltenen Kennwert mit dieser Verteilung, dann kann man feststellen, wie wahrscheinlich es bei gültiger Nullhypothese ist, einen solchen Wert durch Zufall zu erhalten. Sinkt diese Wahrscheinlichkeit unter einen bestimmten kritischen Wert, entscheidet man sich, die Nullhypothese zu verwerfen. Diese Methode hat seine Wurzeln im Vorgehen beim Aussortieren von Ausreißerwerten bei der Messung stabiler Systeme (Gigerenzer, 1989, S. 80, S.95).

Fisher erklärte zwar, ein Signifikanztest sei nur ein „schwaches“ Argument, und nur in den Fällen anzuwenden, wo sehr wenig oder kein Wissen vorliege. Er sah es als die primitivste Form von Argumenten in einer Reihe möglicher statistischer Analysen. Er wies auch darauf hin, daß „kein isoliertes Experiment an sich, wie signifikant auch immer, genügen kann, um eine natürliches Phänomen experimentell aufzuzeigen...“ (Fisher, 1935, §7, aus Gigerenzer et al., 1989, Übers. d. Aut.). Andererseits spricht er eindeutig davon, Nullhypothesen per Signifikanztest widerlegen zu können. Beispielsweise schreibt er, die Nullhypothese könne nie bewiesen, aber sehr wohl widerlegt werden (Fisher 1935).

Anders ist dies bei der Theorie von Neyman und Pearson, die zwar von Fishers Ansatz ausgingen, diesen aber erweitert und modifiziert haben.

2.3 Neyman & Pearson

Im Paradigma von Neyman und Pearson prüft man nicht einfach gegen die Nullhypothese, sondern stellt eine konkrete Alternativhypothese auf, die es erlaubt, einen Erwartungswert vorherzusagen. Wir haben es mit zwei gleichberechtigt nebeneinander stehenden Hypothesen zu tun. Also auch mit zwei verschiedenen Kennwert-Verteilungen: Der von der Nullhypothese (H_0) vorhergesagten, und der von der Alternativhypothese (H_1) vorhergesagten. Dieser Ansatz macht es möglich, auch zwei verschiedene Arten von Fehlern zu spezifizieren.

Nämlich zum einen, die Nullhypothese zugunsten der H_1 zu verwerfen, obwohl sie richtig ist („Fehler erster Art, α -Fehler“), und zum anderen den Fehler, die H_0 beizubehalten, obwohl in Wirklichkeit H_1 zutrifft („Fehler zweiter Art, β -Fehler“).

Je nach Anwendungsfall kann man nun diese beiden möglichen Fehler gegeneinander abwägen, und das Entscheidungskriterium entsprechend festsetzen. Je nach dem, auf welcher Seite des Kriteriums der erhaltene Kennwert liegt, entscheidet man sich entweder für die H_1 oder die H_0 .

Man kann sich das Entscheidungskriterium als eine verschiebbare senkrechte Linie vorstellen, die von jeder der beiden Wahrscheinlichkeitsverteilungen eine Ecke abschneidet. Die Fläche unter der H_0 -Ecke, α genannt, entspricht (wie bei Fishers Vorgehen) der Wahrscheinlichkeit, einen Fehler erster Art zu begehen. Die von der H_1 -Verteilung abgeschnittene Ecke – und das ist das neue – entspricht der Wahrscheinlichkeit, einen Fehler zweiter Art zu begehen, und wird β genannt. So kann man durch das Verschieben des Entscheidungskriteriums die Fehlerrisiken in ein dem Gegenstand angemessenes Gleichgewicht bringen.

Den Rest der Fläche unter der H_1 -Kurve, also $1-\beta$, bezeichneten Neyman und Pearson als die statistische *Power* eines Tests (Teststärke). Die Power ist die Wahrscheinlichkeit, sich durch den Test für die Alternativhypothese zu entscheiden, wenn diese zutrifft. Anders ausgedrückt ist das die Wahrscheinlichkeit, einen vorhandenen Effekt auch zu finden.

Die Power eines Tests hängt von mehreren Faktoren ab: z.B. von der tatsächlichen Größe des Effektes, d.h. davon, wie groß der gesuchte Unterschied in der Population tatsächlich ist, von dem gewählten Entscheidungskriterium, und von der Größe der Stichprobe. Man kann also, wenn man eine Hypothese spezifiziert und ein Entscheidungskriterium festgelegt hat die Stichprobengröße errechnen, die notwendig ist, um sich mit diesem Test für die Alternativhypothese zu entscheiden, falls diese zutrifft.

Neyman und Pearson rückten davon ab, wie beim Nullhypothesentesten Hypothesen endgültig widerlegen zu wollen, und führten den Begriff des *induktiven Verhaltens* ein. Induktives Verhalten bedeutet nicht, daß man nach dem Test an eine bestimmte Hypothese glauben muß, oder dieses nicht mehr darf, sondern, daß man sich je nach Ausgang des Tests so *verhält, als ob* die eine oder die andere Hypothese zutrifft.

Dieses Vorgehen, das Konzept des β -Fehlers und das der Teststärke, bieten entscheidende Vorteile gegenüber Fishers Ansatz, der davon jedoch nichts wissen wollte.

Zwischen Neyman und Pearson auf der einen sowie Fisher auf der anderen Seite entbrannte ein heftiger persönlicher Streit, und Fisher wies die Ansätze der beiden bis zu seinem Tod 1962 hartnäckig zurück.

2.4 Hybrid-Logik

Was heute im Allgemeinen gelehrt wird, ist das Werk verschiedener Lehrbuchautoren und Verleger von Fachzeitschriften, die seit der Zeit des Zweiten Weltkrieges versuchten, diese beiden widerstreitenden Ansätze in

einem Topf weich zu kochen und mundgerecht zu präsentieren. Man nannte dies die Inferenz-Revolution. Gigerenzer et al. (1987, 1989, 1993) bezeichnen das, was dabei herauskam als Hybrid-Theorie, denn sie gibt bei näherem Hinsehen kein einheitliches Bild ab, und weder Fisher noch Neyman und Pearson hätten dieser Hybrid-Theorie zugestimmt.

Das widerlegen der Nullhypothese sowie das spezifizieren einfacher Signifikanzniveaus als Entscheidungskriterium mit unspezifischen Alternativhypothesen nach Fishers Art, wurde mit Neyman-Pearson-Begriffen wie *Power* und *β-Fehler* vermischt, die jedoch nur bei spezifischen Hypothesen einen Sinn ergeben. In den meisten statistischen Lehrbüchern wurde die Hybrid-Theorie gelehrt, ohne darauf hinzuweisen, daß die Schulen, aus denen sie hervorgeht, bis zuletzt zerstritten waren und ihre Ansätze als unvereinbar sahen (Gigerenzer & Murray, 1987).

Das Verschleiern dieser Konflikte gab der Hybrid-Theorie ein „problemloses“ Gesicht, und da sie die Vorteile der beiden widerstreitenden Ansätze in sich zu vereinen scheint, fand sie sehr rasch weite Verbreitung.

2.5 Fehlinterpretationen

Das unkritische Verbreiten und Konsumieren dieser Hybrid-Theorie führte u. a. dazu, daß sich eine ganze Reihe von Fehlinterpretationen festsetzen konnten. Dies ging so weit, daß das Konzept der Signifikanz in weiten Teilen in ihrer Wichtigkeit über die Effektstärke (die Größe des Effektes) erhoben wurde. Dabei wurde scheinbar vergessen, daß die Signifikanz, wie hoch sie auch immer sein mag, nichts über das Ausmaß des gefundenen Effektes aussagt. In manchen Veröffentlichungen wurden gar nur noch Signifikanzwerte angegeben, ohne Effektgröße.

Diese Überbewertung der Signifikanz hatte auch vor den Verlagsetagen renommierter Zeitschriften nicht halt gemacht. So beschreibt zum Beispiel Melton (1962, S. 553f), der das „Journal of Experimental Psychology“ 12 Jahre lang verlegt hatte, „starke Zurückhaltung“ Ergebnisse zu publizieren, die nur auf dem 5% Niveau signifikant waren, während Ergebnisse, die das 1% Niveau erreichten einen Platz in der Zeitschrift verdienten. Die Anwendung des Nullhypothesentests hatte bereits damals ritualistische Züge, wovon selbst Fisher (1933, S. 46) gewarnt hatte.

Wie konnte es dazu kommen? Im Rahmen der Hybridisierung ist einige Verwirrung entstanden, und so war es möglich, daß der Signifikanz allerhand Aussagekraft zugeschrieben wurde, die sie nicht besitzt. Melton z.B. schreibt in seinem Editorial, daß das Signifikanzniveau den Grad des Vertrauens in die Wiederholbarkeit des Ergebnisses widerspiegeln, und daß es die Größe des Effektes spezifiziere (S. 553f, zit. aus Gigerenzer, 1987). Das ist natürlich falsch, denn die Größe des Effektes ergibt sich direkt aus den Daten (z.B. beim Vergleich zweier Mittelwerte), ohne die Anwendung irgendeines Signifikanztestes.

Dieses Mißverständnis ist jedoch bei weitem kein Einzelfall, denn die selben Fehler sind selbst in den statistischen Lehrbüchern zu finden. So ist z.B. bei

Nunally (1975) zu lesen, daß statistische Signifikanz auf dem 5% Niveau bedeute, „daß der Forscher mit einer Wahrscheinlichkeit von 95 aus 100 davon ausgehen kann, daß der beobachtete Unterschied sich in zukünftigen Untersuchungen wieder zeigen wird“ (S. 195, Übers. d. Aut.). Auch das ist falsch, denn die Signifikanz eines Ergebnisses sagt überhaupt nichts über dessen Replizierbarkeit aus.

Ein anderer sehr weit verbreiteter Fehler ist es, Signifikanz als die umgekehrte Wahrscheinlichkeit im Bayesschen Sinne zu interpretieren (Zum Ansatz von Bayes siehe 3.2): Wenn wir hypothesenprüfende Verfahren anwenden, suchen wir eine Antwort auf die Frage: „Wie wahrscheinlich ist es, daß diese oder jene Hypothese zutrifft?“, bzw. nach der Erhebung der Daten: „Wie wahrscheinlich ist es, daß die Hypothese zutrifft, wenn ich solche Daten erhalten habe?“ – gesucht ist $p(H|D)$, die Wahrscheinlichkeit der Hypothese, gegeben die Daten. Was uns der Signifikanztest liefert, ist aber $p(D|H)$, die Wahrscheinlichkeit der Daten, gegeben die (Null-)Hypothese. Hier wäre Bayes gefragt, doch ohne darauf einzugehen wird diese Wahrscheinlichkeit nur allzu gerne umgekehrt interpretiert und das Signifikanzniveau als die (sehr geringe) Wahrscheinlichkeit der Nullhypothese angesehen.

So geschehen u. a. im statistischen Anhang von Miller und Buckhouts (1973) Einführung in die Psychologie, wo der Autor (F. L. Brown) die Logik des statistischen Hypothesentests anhand eines Experimentes zu außersinnlichen Wahrnehmung erklärt:

Eine Versuchsperson versucht die Gedanken des Versuchsleiters zu lesen. Sie rät 69 von 100 Münzwürfen richtig. Die Wahrscheinlichkeit 69 oder mehr von 100 Münzwürfen richtig zu raten, ist ungefähr eins zu zehntausend, falls die Nullhypothese stimmt, die besagt, daß die Wahrscheinlichkeit einen Wurf richtig zu raten .5 ist. Soweit, so gut. Doch dann schreibt er, die „Wahrscheinlichkeit daß die Nullhypothese korrekt sein könnte“ sei „ungefähr 1 aus 10 000“ und die „Wahrscheinlichkeit daß die Nullhypothese falsch ist“ betrage „etwa 9 999 aus 10 000“ (Übers. d. Aut., zit. aus Gigerenzer und Murray, 1987). Hier wurde die Wahrscheinlichkeit der Nullhypothese nach einem solchen Ergebnis $p(H|D)$ gleichgesetzt mit $p(D|H)$, der Wahrscheinlichkeit des Ergebnisses unter der Nullhypothese.

Fassen wir noch einmal zusammen:

Das Signifikanzniveau ist die Wahrscheinlichkeit der Daten unter der Nullhypothese und es besagt nichts über die Größe eines Effektes, die Replizierbarkeit des Ergebnisses oder die Wahrscheinlichkeit irgendeiner Hypothese.

3 Die Originalstudie

Im Rahmen einer Arbeit, die allerhand verbreitete Mißkonzeptionen und intuitive Fehlinterpretationen in der Statistik aufdeckte, führte Michael Oakes 1986 die folgende Studie an 70 akademischen Psychologen durch. Es soll hier zunächst der Originaltext referiert werden, darauf folgt die Übersetzung, die ich für meinen Fragebogen verwendet habe. Dieser enthält allerdings

zusätzlich noch einige nicht fachliche Fragen., und ist im Anhang einzusehen.

Fragebogen (original):

Suppose you have a treatment which you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means t test and your result is ($t = 2.7$, d.f. 18, $p = 0.01$) Please mark each of the statements below as 'true' or 'false'.

- (i) You have absolutely disproved the null hypothesis (that there is no difference between the population means).
- (ii) You have found the probability of the null hypothesis being true.
- (iii) You have absolutely proved your experimental hypothesis (that there is a difference between the population means).
- (iv) You can deduce the probability of the experimental hypothesis being true.
- (v) You know, if you decided to reject the null hypotheses, the probability that you are making the wrong decision.
- (vi) You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

(Ende des Fragebogens)

Deutsche Version des Fragebogens:

Stellen Sie sich vor, Sie haben ein Treatment, von dem Sie glauben, es könnte die Leistung bei einer bestimmten Aufgabe beeinflussen. Sie vergleichen die Mittelwerte ihrer Kontroll- und Experimentalgruppen von jeweils 20 Probanden. Nehmen wir weiterhin an, daß Sie einen einfachen t -Test für unabhängige Stichproben verwenden. Das Ergebnis ist: $t=2.7$, 18 Freiheitsgrade, $p = 0.01$. Der Unterschied zwischen den Gruppen ist also auf dem 1%-Niveau signifikant.

Bitte markieren Sie jede der folgenden Aussagen als „richtig“ oder „falsch“. „Falsch“ bedeutet, daß die Aussage nicht streng logisch aus den o. g. Prämissen folgt. Es können auch mehrere oder gar keine richtigen dabei sein!

- 1) Es ist eindeutig bewiesen, daß die Nullhypothese (daß zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist.
- 2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden.

3) Es ist eindeutig bewiesen, daß Ihre Alternativhypothese (daß es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist.

4) Man kann nun die Wahrscheinlichkeit ableiten, daß die Alternativhypothese richtig ist.

5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, daß diese Entscheidung falsch sein könnte.

6) Der experimentelle Befund ist reliabel in dem Sinne, daß man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde.

(Ende des Fragebogens)

3.1 Diskussion der einzelne Aussagen:

Die Aussagen 1 und 3 sind eindeutig falsch. In dem gegebenen Zusammenhang können überhaupt keine Aussagen *mit Sicherheit* gemacht werden. Nichts ist *eindeutig* bewiesen oder widerlegt.

Aussage 2 und Aussage 4 lassen sich ineinander überführen, denn wäre Wahrscheinlichkeit der Nullhypothese bekannt, könnte man auch die Wahrscheinlichkeit der Alternativhypothese ableiten (das Komplement).

Aussage 5 beschreibt die Wahrscheinlichkeit, einen Fehler erster Ordnung (α -Fehler) zu begehen, d.h. die Alternativhypothese anzunehmen, obwohl sie falsch ist. Da wir in unserem Fall ja annehmen, diese Entscheidung getroffen zu haben, ist sie genau dann falsch, wenn die Nullhypothese in Wirklichkeit zutrifft. Also entspricht die beschriebene Wahrscheinlichkeit genau der Wahrscheinlichkeit des Zutreffens der Nullhypothese. Damit ist Aussage 5 logisch gleichbedeutend mit den Aussagen 2 und 4. Die Aussagen 2, 4 und 5 sollten also konsistenterweise immer gleich beantwortet werden, und zwar als „falsch“ (siehe Abschn. 3.2, Bayes).

Aussage 6 spiegelt die Verwechslung der statistischen Signifikanz mit der statistischen Power wieder. Im Paradigma von Neyman und Pearson könnte man eine Aussage machen wie „Ich werde in 99% der Fälle zugunsten der H_1 entscheiden, falls sie wahr ist“. Wenn man das, was Signifikanz bedeutet ähnlich dieser Aussage frequentistisch formuliert, müsste man sagen: „Man bekäme in 1% der Fälle ein signifikantes Ergebnis, falls H_0 gilt“. In Aussage 6 jedoch, ist diese Annahme nicht enthalten – auch sie ist falsch.

3.2 Exkurs: das Konzept der umgekehrten Wahrscheinlichkeit – der Ansatz von T. Bayes

Der britische Theologe Thomas Bayes beschäftigte sich im achtzehnten Jahrhundert mit dem Konzept der umgekehrten Wahrscheinlichkeit: Ich habe eine bedingte Wahrscheinlichkeit $p(D|H)$, wie kann ich die umgekehrte Wahrscheinlichkeit $p(H|D)$ bestimmen? Wir benötigen zur Berechnung das Bayes-Theorem, eine Formel, die, basierend auf Bayes' Überlegungen, erst nach seinem

Tod veröffentlicht wurde, und sich recht einfach aus den Axiomen der Wahrscheinlichkeitstheorie ableiten läßt.

$$p(H|D) = \frac{p(H) \cdot p(D|H)}{p(D)} \quad \text{Das Bayes-Theorem}$$

Ich möchte das Konzept der umgekehrten Wahrscheinlichkeiten an einem Beispiel veranschaulichen:

Die Wahrscheinlichkeit AIDS zu haben, sofern man keiner der Risikogruppen (z.B. Fixer, Prostituierte, Homosexuelle) angehört, beträgt 0.01%. Die heutzutage verwendete Kombination von HIV-Tests hat eine Sensitivität von 99.9%, d.h. daß sie den Virus mit 99.9-prozentiger Wahrscheinlichkeit erkennt, falls er vorhanden ist. Die Falsch-Positiv-Rate ist 0.01%, das bedeutet, die Wahrscheinlichkeit ein positives Testergebnis zu bekommen, wenn man in Wirklichkeit keinen Virus hat, beträgt nur 0.01%. Der Test scheint also sehr sicher zu sein. Für einen Menschen, der gerade einen positiven AIDS-Test zurückbekommen hat, interessiert nun hauptsächlich die Frage, „Mit welcher Wahrscheinlichkeit bin ich wirklich infiziert, wenn ich ein positives Ergebnis bekommen habe?“. Die meisten Ärzte würden nun mit einem Blick auf die o. g. Daten intuitiv antworten „Mit sehr hoher Sicherheit!“

Stellen wir uns die Situation in natürlichen Häufigkeiten vor:

Gehen wir von 10 000 Personen aus. Davon hat eine AIDS (0.01%), der Rest (9 999) ist gesund (in Bezug auf AIDS). Die eine infizierte Person bekommt mit 99.9-prozentiger Wahrscheinlichkeit einen positiven Test. Von den restlichen 9 999 bekommt aufgrund der Falsch-Positiv-Rate von 0.01% ein weiterer ein positives Ergebnis, obwohl er den Virus nicht hat. Von 10 000 Personen bekommen also im Schnitt zwei einen positiven AIDS-Test, wovon nur einer den Virus wirklich hat. Die Chance, den Virus zu haben wenn man ein positives Ergebnis bekommen hat, stehen demnach 1:1 und nicht nahe 100%, wie man intuitiv hätte annehmen können.

Sehen wir dieses Beispiel als einen Hypothesentest, dann wäre es die Alternativhypothese, daß der Patient AIDS hat, und die Nullhypothese, daß er gesund ist. Bekommen wir nun vom Labor einen positiven Test zurück, dann ist dieses Ergebnis hochsignifikant, denn die Wahrscheinlichkeit ein solches Ergebnis zu erhalten, wenn der Patient in Wirklichkeit gesund ist (d.h. die H_0 gilt) beträgt ja nur 0.01% also $p = .0001$. Obwohl die Wahrscheinlichkeit der Daten unter der Nullhypothese $p(D|H_0) = .0001$, also verschwindend gering ist, beträgt die Wahrscheinlichkeit der Hypothese gegeben die Daten $p(H_0|D) = .5$. Es wäre fatal, dieses hochsignifikante Ergebnis als (fast) sichere Bestätigung zu interpretieren, daß der Patient den Virus trägt.

Die verwendeten Daten entsprechen der Wirklichkeit, und es gibt andere Beispiele, z.B. Mammographie, wo die Wahrscheinlichkeit, die Krankheit zu haben sogar nur 7,8% beträgt, wenn man ein positives Ergebnis bekommt (siehe dazu Gigerenzer & Hoffrage, 1995), wo es trotz positivem (signifikantem) Ergebnis um ein vielfaches wahrscheinlicher ist, die Krankheit *nicht* zu haben, wo also die H_0 trotz Signifikanz weitaus wahrscheinlicher ist als die H_1 .

Umgekehrt lassen sich ohne weiteres Fälle konstruieren, wo die H_1 plausibler ist, obwohl das Signifikanzniveau die üblichen kritischen Werten nicht erreicht.

Ich hoffe es ist hier deutlich geworden, daß sich das Signifikanzniveau keinesfalls als Wahrscheinlichkeit von Hypothesen auslegen läßt.

Um die Wahrscheinlichkeit einer Hypothese mit dem Bayes-Theorem bestimmen zu können, müßten wir die a-priori-Wahrscheinlichkeit der Hypothese kennen, d.h. die Grundwahrscheinlichkeit der Hypothese, bevor irgendwelche Daten erhoben wurden. Der Schwachpunkt des Bayesschen Ansatzes liegt in der Tatsache, daß man diese a-priori-Wahrscheinlichkeit allenfalls subjektiv schätzen kann; abgesehen von Versuchen der Objektivierung wie der Regel von Laplace (s. z.B. Keynes, 1943), die der Sache aber auch nicht gerecht werden.

Im übrigen war die Ablehnung des Bayesschen Ansatzes einer der wenigen Punkte in denen Neyman und Pearson mit Fisher übereinstimmten.

4 Erhebung

4.1 Teilnehmer:

Die verwendete Stichprobe von $n = 98$ setzt sich aus 45 Studenten der Psychologie sowie 53 Wissenschaftlern (Hauptsächlich der Freien Universität Berlin) zusammen.

Die untersuchten Personen sind allesamt als Forscher, Dozenten oder Studenten im psychologischen Bereich tätig.

Ausschlüsse

Es wurden drei Studenten von der Auswertung ausgeschlossen, da sie in der Vorlesung von Professor Gigerenzer, wo die Originalstudie (Oakes, 1986) diskutiert wurde, mit den richtigen Antworten konfrontiert worden waren. (Zwei von ihnen hatten den Fragebogen auch korrekt ausgefüllt.)

Ein weiterer Student mußte ausgeschlossen werden, da er den Fragebogen nicht sofort ausgefüllt hatte, sondern erst eine Woche nachdem er ihn bekommen hatte. Auch er hatte alle Fragen korrekt beantwortet.

4.2 Vorgehen:

Die Teilnehmer wurden angesprochen, und gefragt ob sie an einer kurzen anonymen Umfrage zum „statistischen Denken“ teilnehmen würden. Personen, die angaben, mit dem Fragebogen nichts anfangen zu können, wurden nicht weiter untersucht. Einige Ablehnungen wurden begründet mit Sätzen wie „Oh, da müßte ich ja richtig nachdenken...“, andere mit „...weil ich finde, daß Fragebogen kein geeignetes Instrument psychologischer Forschung sind.“ die Anzahl der Verweigerer schätze ich auf insgesamt etwa 40 % der Angesprochenen.

Die Probanden füllten die Fragebogen unter Aufsicht aus, und konnten sie dann gefaltet in eine Schachtel zu den anderen werfen, um Anonymität zu gewährleisten.

Die Erhebung dauerte ca. vier Wochen. Nach den ersten Durchgängen wurde der Fragebogen noch einmal leicht geändert. Die letzte Instruktion, mit der Bitte, eine der Aussagen als die dem Probanden geläufige

Definition von Signifikanz zu markieren, oder diese zu formulieren, wurde am Anfang mündlich gegeben, doch nach dem dies einige Male vergessen worden war, wurde diese Instruktionen doch noch schriftlich mit auf den Bogen übernommen. Desweiteren wurde die Formulierung der zweiten Aussage leicht geändert, und „Die Wahrscheinlichkeit der Nullhypothese ist gefunden worden“ sinnerhaltend ergänzt zu „Die Wahrscheinlichkeit *des Zutreffens* der Nullhypothese ist gefunden worden“, da einige Probanden Schwierigkeiten mit der ersten Formulierung geäußert hatten.

Die zwei Versionen des verwendeten Fragebogens sind im Anhang beigefügt.

5 Ergebnisse

Da die Mehrzahl der Forscher auch in der Lehre tätig waren (42 von 53) und – bis auf einen – alle Lehrenden auch in der Forschung, habe ich der Übersichtlichkeit wegen Forscher und Lehrende als eine Gruppe zusammengefaßt, die ich nun alle als Forscher bezeichnen werde. Von diesen 53 Forschern gaben 87% an, Signifikanztests zu verwenden (oder verwendet zu haben).

Von allen Befragten machten nur acht Personen keinen Fehler, und erkannten alle sechs Aussagen als falsch. Alle Studenten (außer den ausgeschlossenen, s. Abschn. 4.1) hatten mindestens einen Fehler gemacht, und 45 der 53 Forscher auch (85 Prozent). Von den 14 Personen, die angaben, in der Methodenlehre tätig zu sein, hatten 10 Fehler gemacht.

Es hatten also 89 Personen (92%) mindestens einen Fehler gemacht.

Von diesen 89 machten 28 (32%) den Fehler, irgend eine Hypothese als absolut bewiesen bzw. absolut widerlegt zu sehen (Aussage 1 & 3). Die Hälfte von ihnen (44), meinten, die Wahrscheinlichkeit einer signifikanten Replikation gegeben zu haben (Aussage sechs).

Einundachtzig Personen (91% der Getäuschten) machten den Fehler, die Signifikanz als die umgekehrte Wahrscheinlichkeit zu interpretieren, und aus dem Signifikanzniveau die Wahrscheinlichkeit einer Hypothese abzuleiten (Aussagen 2, 4 oder 5). Doch bewerteten von diesen 81 nur 10 (12%) diese drei logisch gleichbedeutenden Aussagen konsistent. (Konsistent wäre, entweder alle drei als „richtig“ oder alle drei als „falsch“ zu bewerten.) Die Antworten der meisten Personen (71) widersprachen sich also sogar gegenseitig.

Selbst von jenen 46 Personen, die angaben, in der Forschung tätig zu sein und Signifikanztests gelernt und

angewendet zu haben, zeigten 38 (83%) Fehlinterpretationen.

Von welcher Personengruppe welche Fehler gemacht wurden ist zusammen mit der durchschnittlichen Anzahl der Fehler aus Tab. 1 zu entnehmen.

Welche Art der Fehler wie häufig gemacht wurde, ist in Abb. 1 noch einmal anschaulich gemacht. Dabei wurden die sechs Aussagen (wie oben) in drei Fehlergruppen zusammengefaßt:

- 1) den Fehler, etwas als *absolut* bewiesen oder widerlegt zu sehen (Aussage eins oder drei mit „richtig“ bewertet),
- 2) den Fehler, der *umgekehrten Wahrscheinlichkeit* (Aussagen 2, 4 oder 5) und
- 3) den Fehler mit der Annahme über Replizierbarkeit (Aussage 6). Zusätzlich ist der Prozentsatz der fehlerfreien Bewertungen gezeigt.

Leider kamen nur insgesamt 28 der befragten Personen der Aufforderung nach, eine der Aussagen als die ihnen geläufige Definition von Signifikanz zu markieren, oder diese schriftlich zu formulieren. Von den 77 Probanden, welche die Instruktion erhalten hatten, wurde sie von der Mehrzahl ignoriert. Viele gaben an, die Definition nicht zu wissen, oder diese „jetzt nicht aus dem Ärmel schütteln zu können“.

Sehen wir uns jedoch wenigstens die 28 erhaltenen Angaben an: in 20 Fällen wurde eine der zuvor als richtig bewerteten Aussagen als Signifikanzdefinition markiert. In acht Fällen wurden eigene „Definitionen“ formuliert. Drei davon waren falsch. Die restlichen fünf bezeichneten korrekterweise die Wahrscheinlichkeit, solche oder extremere Daten zu erhalten, unter der Voraussetzung, daß die Nullhypothese gilt. Zwei dieser fünf Personen, welche die richtige Definition geben konnten, hatten allerdings zusätzlich mindestens eine der falschen Aussagen als richtig bewertet.

Von den acht Personen, die keinen Fehler gemacht hatten, nannten also drei auch die korrekte Definition von Signifikanz, eine gab an, sie nicht zu kennen, und eine weitere lieferte eine Antwort, die sinngemäß gleichbedeutend mit Aussage fünf, und damit falsch war. Die übrigen drei „fehlerfreien“ Personen hatten die letzte Instruktion nicht erhalten.

Was noch auffiel: Ich hatte es zwar nicht mit erhoben, aber da mir doch einige Fragebogen gezeigt wurden konnte ich feststellen, daß mindestens 4 der 8 richtigen von Professoren stammten, obwohl die Professoren nur etwa 10% der Befragten ausmachten.

Tab. 1: Häufigkeiten und Prozentwerte der falschen Antworten in den Gruppen

Aussage	Anzahl	% der Stud.	% der Forscher
1) H_0 ist eindeutig widerlegt	23	34%	15%
2) Wahrscheinlichkeit von H_0	27	32%	23%
3) H_1 eindeutig bewiesen	16	20%	13%
4) Wahrscheinlichkeit von H_1	45	59%	34%
5) Wahrscheinlichkeit Fehler erster Art zu begehen	66	68%	64%
6) Ergebnis reliabel / replizierbar	46	41%	49%
Durchschnittliche Fehlerzahl	2,22	2,54	1,98

6 Diskussion

Die vorliegende Studie ist bestimmt in einigen Punkten angreifbar. Da wäre zunächst die Art der Stichprobe zu bemängeln, die aufs Geratewohl gezogen wurde, und möglicherweise die Verhältnisse in der Population nicht exakt abbildet. Was verbirgt sich hinter der hohen Zahl der Verweigerer? Da die meisten, die die Befragung abgelehnten, dies mit einem Kommentar taten wie: „Das ist schon so lange her“ oder „Sowas kann ich nicht mehr“, gehe ich davon aus, daß sich unter ihnen allenfalls sehr wenige befanden, die den Fragebogen richtig gelöst hätten. Eine Einbeziehung aller nicht Befragten hätte also wahrscheinlich noch zur Verdeutlichung der gefundenen Tendenz beigetragen.

Das Argument, daß der hohe Fehleranteil hauptsächlich auf „ignorante“ Studenten zurückzuführen sei, ist durch die getrennte Auswertung von Studenten und Forschern wohl entkräftet.

Was sicherlich einige Personen zu einem falschen Kreuz verleitet hat, ist die – zugegebenermaßen ein wenig gemeine – Tatsache, daß unter den zu bewertenden Aussagen keine richtige dabei war. Sicherlich hätten einige Probanden die korrekte Interpretation wiedererkannt, und keine falsche Aussage angekreuzt. „Eine muß doch richtig sein!“ – wird sich der Eine oder Andere gedacht haben. Obwohl die Instruktionen ausdrücklich darauf hinwies, daß auch alle oder gar keine der Aussagen richtig sein könnten, und die Instruktion lautete, jede Aussage einzeln zu bewerten, und nicht „die richtige herauszufinden“, hatten das wohl einige so aufgefaßt. Der Fragebogen war also gewissermaßen eine besonders strenge Prüfung. Doch wer wirklich weiß, was Signifikanz bedeutet, der wird auch keine der falschen Aussagen angekreuzt haben.

Eine Person mutmaßte: „Da soll sicher raus kommen, wie wenig Ahnung die kritischen Psychologen von

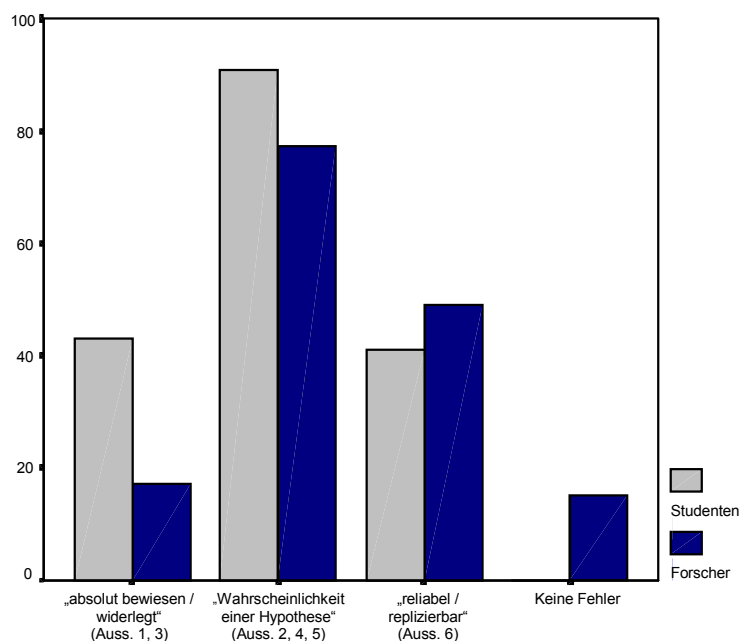


Abb. 1 Die Prozentuale Häufigkeit mit der Fehler einer bestimmten Art gemacht wurden, getrennt für Studenten und Forscher.

Signifikanztests haben.“ Und fügte hinzu: „Signifikanztests sind aber kein Gegenstand kritischer Wissenschaft!“. Eine andere Person entschuldigte ihr Nichtwissen damit, daß sie zur Zeit keine derartigen Methoden anwende, versicherte aber, daß sie sich wieder genau kundig machen würde, wenn es dazukommen sollte. Ja, wer keine Signifikanztests anwendet, und sich auch nicht mit Forschung auseinandersetzt, wo solche zur Anwendung kamen, der braucht tatsächlich nicht zu wissen, was Signifikanz bedeutet. Allerdings bedeutet dies meiner Ansicht nach, sich dem Großteil, der heute praktizierten Forschung, und damit sicherlich auch einigen interessanten Erkenntnissen zu verschließen. Im übrigen ist es meines Erachtens für eine differenzierte Kritik notwendig, ein gesundes Verständnis der zugrundeliegenden Konzepte zu haben. Es kann auch für „kritische“ Psychologen nichts schaden, ein wenig über die von ihnen (oft zu Recht) kritisierten „quantitativen“ Methoden zu wissen.

Noch bedenklicher allerdings ist, daß sich gezeigt hat, daß selbst jene Forscher, die quantitative Methoden anwenden, und wohl auch deren Ergebnisse konsumieren, in der großen Mehrzahl Fehlinterpretationen zeigen, und scheinbar kein fundiertes Verständnis des Prinzips haben, welches den heute gängigen hypothesenprüfenden Verfahren zugrunde liegt.

Die Tatsache, daß viele Personen die drei gleichbedeutenden Aussagen nicht konsistent angekreuzt haben, kann man nicht auf Fehlinformation zurückführen, da es mit einem minimalen Verständnis dessen, wie sich Null- und Alternativhypothese zueinander verhalten, möglich gewesen wäre, die Aussagen konsistent zu beantworten. Sie zeigt meiner Meinung nach vielmehr, daß den meisten Personen die Problematik nicht geläufig ist.

6.1 Ausblick

Das Ergebnis dieser Studie spiegelt meines Erachtens den Stellenwert wieder, den die Statistik in den Köpfen vieler Psychologen einnimmt: als notwendiges Übel oder als Mittel, um einer Untersuchung ein objektives und wissenschaftliches Gesicht zu geben. Nur allzu gerne verzichten wir darauf, uns mit den scheinbar unwichtigen Feinheiten der Statistik abzugeben.

Literaturverzeichnis

- Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society*, 27, 186-190. (Zit. aus Gigerenzer & Murray, 1987)
- Fisher, R. A. (1933). The contributions of Rothamsted to the development of the science of statistics. *Annual Report of the Rothamsted Station*, 43-50 (Reprinted in *Collected papers*, Vol. 3, 84-91) (Zit. aus Gigerenzer & Murray, 1987)
- Fisher, R. A. (1935). *The design of experiments* (5th ed., 1951; 7th ed., 1960; 8th ed. 1966). Edinburgh: Oliver & Boyd. (Zit. aus Gigerenzer, 1989)
- Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, New Jersey: Lawrence Erlbaum Ass..
- Gigerenzer, G. (Hrsg.) (1989). *The empire of chance: how probability changed science and everyday life*. Cambridge: Cambridge University Press.
Inzwischen auch auf Deutsch erschienen: (Gigerenzer, 1999)
- Gigerenzer, G. (1993). The Superego, the Ego, and the Id, in Statistical Reasoning. In G. Keren & Ch. Lewis (Hrsg.): *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* Kapitel 11. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Gigerenzer, G. et. al. (1999). *Das Reich des Zufalls: Wissen zwischen Wahrscheinlichkeiten, Häufigkeiten und Unschärfen*. Heidelberg: Spektrum Akademischer Verlag.
- Gigerenzer, G. & Hoffrage, U. (1995). How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. In *Psychological Review*, Vol. 102, No. 4. 684-704.
- Keynes, 1943 (Zit. aus Gigerenzer, 1987)
- Melton, A. W. (1962). Editorial. *Journal of Experimental Psychology*, 64, 553-557. (Zit. aus Gigerenzer & Murray, 1987)
- Miller, G. A., & Buckhout, R. (1973). *Psychology: The science of a mental life* (Appendix: Introduction to statistical methods in psychology by F. L. Brown). New York: Harper & Row. (Zit. aus Gigerenzer & Murray, 1987)
- Nunally, J. C. (1975). *Introduction to statistics for psychology and education*. New York: McGraw-Hill.
- Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences* Kapitel 3. N.Y.: Wiley.

Anhang: Fragebogen

Es folgt der Fragebogen in der ersten und zweiten Fassung. Es wurde vorwiegend die zweite Version verwendet (s. 4.2 - Vorgehen).

Was wir brauchen sind keine statistischen Rituale, die schematisch auf alle möglichen Problemstellungen angewendet werden, sondern einen fundierten, breiten Methoden-Schatz, um für verschiedene Forschungsfragen die bestmögliche Methode parat zu haben.

Es gibt Alternativen zum klassischen Nullhypotesentesten. Zum Beispiel könnte man, wenn man die Alternativhypothese spezifisch formuliert, also eine Effektstärke vorhersagt, den Ansatz von Neyman und Pearson konsistent anwenden.

Auch sollte man sich vielleicht öfters fragen, ob es für gewisse Fragestellungen nicht fruchtbarer wäre, eine qualitative oder Bayesianische Herangehensweise zu wählen.

Auf jeden Fall sollte man im Auge behalten, was die Signifikanz wirklich aussagt, und was nicht, und ihr keine allzu große Entscheidungsmacht geben.

Kontakt

Fragen, Kritik, Diskussion:

www.heikohaller.de/kontakt.html

Stellen Sie sich vor, Sie haben ein Treatment, von dem Sie glauben, es könnte die Leistung bei einer bestimmten Aufgabe beeinflussen. Sie vergleichen die Mittelwerte ihrer Kontroll- und Experimentalgruppen von jeweils 20 Probanden. Nehmen wir weiterhin an, daß Sie einen einfachen t-Test für unabhängige Stichproben verwenden. Das Ergebnis ist: $t=2.7$, 18 Freiheitsgrade, $p = 0.01$. Der Unterschied zwischen den Gruppen ist also auf dem 1%-Niveau signifikant.

Bitte markieren Sie jede der folgenden Aussagen als "richtig" oder "falsch".
"Falsch" bedeutet, daß die Aussage nicht streng logisch aus den o. g. Prämissen folgt.
Es können auch mehrere oder gar keine richtigen dabei sein!

- 1) Es ist eindeutig bewiesen, daß die Nullhypothese (daß zwischen den Populationsmittelwerten kein Unterschied besteht) falsch ist. [richtig / falsch
- 2) Die Wahrscheinlichkeit des Zutreffens der Nullhypothese ist gefunden worden. [richtig / falsch
- 3) Es ist eindeutig bewiesen, daß Ihre Alternativhypothese (daß es einen Unterschied zwischen den Populationsmittelwerten gibt) wahr ist. [richtig / falsch
- 4) Man kann nun die Wahrscheinlichkeit ableiten, daß die Alternativhypothese richtig ist. [richtig / falsch
- 5) Entscheidet man sich nun, die Nullhypothese zu verwerfen, dann weiß man jetzt die Wahrscheinlichkeit, daß diese Entscheidung falsch sein könnte. [richtig / falsch
- 6) Der experimentelle Befund ist reliabel in dem Sinne, daß man in 99% der Fälle ein signifikantes Ergebnis bekäme, wenn man das Experiment sehr oft wiederholen würde. [richtig / falsch

Haben Sie die Anwendung von Signifikanztests gelernt? [ja / nein

Verwende(te)n Sie Signifikanztests
(auch in der Vergangenheit oder im Ex-Prak)? [ja / nein

Sind Sie in der Lehre tätig? [ja / nein

(Wenn ja) Lehren Sie Methodenlehre? [ja / nein

Sind Sie in der Forschung tätig? [ja / nein

Kennen Sie das Bayes-Theorem? [ja / nein

Entspricht eine der obigen Aussagen der Ihnen geläufigen Definition von Signifikanz?
Wenn ja: bitte markieren,
wenn nein: können Sie diese Formulieren? – Bitte aufschreiben.

vielen Dank!