Kladroba, Andreas; von der Lippe, Peter; Westermann, Michael

Working Paper

Zwei Beiträge zur Interpretation von statistischen Tests: Replik auf den Artikel "Von der Wahrscheinlichkeit des Irrtums" von W. Weihe

Diskussionsbeitrag, No. 137

Provided in Cooperation with:
University of Duisburg-Essen, Institute of Business and Economic Studie (IBES)

Suggested Citation: Kladroba, Andreas; von der Lippe, Peter; Westermann, Michael (2004) : Zwei Beiträge zur Interpretation von statistischen Tests: Replik auf den Artikel "Von der Wahrscheinlichkeit des Irrtums" von W. Weihe, Diskussionsbeitrag, No. 137, Universität Duisburg-Essen, Fachbereich Wirtschaftswissenschaften, Essen

This Version is available at:
https://hdl.handle.net/10419/23142

Andreas Kladroba, Michael Westermann

Einmal zu viel um die Ecke gedacht

- Einige Anmerkungen zu Wolfgang Weihe –

Im „Deutschen Ärzteblatt“ 101 (2004), Heft 13, S. A834 – A838 versucht Wolfgang Weihe anhand eines fiktiven Gesprächs zwischen den Medizinern Sagredo, Simplicio (dessen Alkoholkonsum ein schlechtes Beispiel für seine Patienten darstellt) und Salviati einige Grundprinzipien klinischer Studien und der damit verbundenen statistischen Testtheorie kritisch zu hinterfragen. Er kommt dabei zu einigen verblüffenden Ergebnissen. Die angehenden Statistiker Gass, Bernalli und Poissin diskutieren im folgenden die Gedanken der drei Ärzte.

GASS: Seht mal, was ich hier gefunden habe. Eine Diskussion dreier Mediziner über statistische Methoden. Lest Euch das doch einmal durch.

(Bernalli und Poissin lesen)

BERNALLI: Klingt komisch.

POISSIN: Obwohl ich spontan nicht sagen kann, was komisch ist.

GASS: Schauen wir uns das doch einmal im Detail an. Als erstes haben wir ein Beispiel zweier fiktiver Studien bezüglich eines Medikamentes für Herzinfarktpatienten. In der ersten Studie, bei der eine Verum- genauso wie eine Placebogruppe nur aus je zwei Patienten bestehen, sterben beide Patienten der Placebogruppe, während die Patienten der Verumgruppe überleben. Bei der zweiten Studie (gleiche Gruppengrößen) ist es genau umgekehrt. Jetzt wird diskutiert, was das über die Wirksamkeit dieses Medikaments aussagt. Zunächst die ganz einfache die Frage an Euch: Was haben wir hier vorliegen?

BERNALLI: Den Fall eines Zwei-Stichproben-Tests auf Anteilswerte.

GASS: Wie lauten also die Hypothesen?

BERNALLI: Die Nullhypothese sagt, dass der Anteil der Erfolge in der Verumgruppe höchstens so groß ist wie der in der Placebo-Gruppe, die Gegenhypothese sagt entsprechend, dass der Anteil höher ist.

GASS: Und weiter?

BERNALLI: Wir haben hier natürlich das Problem einer extrem kleinen Stichprobe, so dass es nicht möglich ist, die hier vorliegende Binomialverteilung mit einer Normalverteilung zu approximieren.

POISSIN: Gehen wir doch einfach „zu Fuß“ vor. Nehmen wir als Prüfgröße eine Variable, die den Unterschied der Erfolge in der Verum- und der in der Placebogruppe anzeigt. Also, wenn von mir aus X die Anzahl der Erfolge in der Verum- und Y die Anzahl der Erfolge in der Patientengruppe ist, dann ist $Z = X – Y$ unsere Prüfgröße.

BERNALLI: Damit ist auch klar, dass Z größer als Null sein muss, damit ich die Nullhypothese ablehnen kann. Die Frage ist nur, muss $Z = 2$ sein, womit wir den Fall bekommen, den Simplicio beschreibt, nämlich dass in der Verumgruppe beide Pati-

enten überleben und beide Placebo-Patienten sterben oder reicht ein Z = 1 aus, also z.B. zwei Erfolge in der Verum- und ein Erfolg in der Placebogruppe oder ein Erfolg in der Verum- und kein Erfolg in der Placebogruppe um ein signifikantes Ergebnis zu bekommen?

POISSIN: Freundlicherweise werden uns ja die 16 möglichen Ergebnisse eines solchen Experiments von Simplicio aufgelistet. Berechnen wir daran doch einfach mal die Wahrscheinlichkeiten für die einzelnen Z-Werte. Sie sind in Tab. 1 zusammengefasst.

z_i	$P(Z = z_i)$	$P(Z \geq z_i)$
-2	0,0625	1
-1	0,25	0,75
0	0,375	0,375
1	0,25	0,125
2	0,0625	0,0625

Tab. 1

BERNALLI: Moment, das müsst Ihr mir erklären.

POISSIN: Die erste Spalte in Tabelle 1 gibt die Werte an, die Z annehmen kann. Die zweite Spalte beschreibt die Wahrscheinlichkeit, genau diesen Wert zu erhalten. Dabei haben wir vorausgesetzt, dass die Wahrscheinlichkeit einen Herzinfarkt zu überleben bei einer Nicht-Behandlung bei 50% liegt. Ich bin kein Arzt, ich weiß also nicht wie realistisch diese Zahl ist, aber Simplicio hat diese Annahme auch getroffen, obwohl ich mir nicht so ganz sicher bin, ob er es gemerkt hat. Die dritte Spalte ist etwas ungewöhnlich. Sie beschreibt die Wahrscheinlichkeit den entsprechenden Z-Wert oder einen höheren zu bekommen. Diese Wahrscheinlichkeit müsste unter unserer Signifikanzschranke liegen, damit die Null-Hypothese abgelehnt wird. Wir sehen, dass wir eine Signifikanz von 5% nirgendwo erreichen, aber ein Z = 2 führt zumindest zu einer Ablehnung bei einem Signifikanzniveau von 10%, was ja auch nicht so ungewöhnlich ist.

GASS: Was heißt das jetzt für die beiden hier zitierten Untersuchungen?

POISSIN: Völlig klar: Bei der ersten Untersuchung (alle Mitglieder der Verumgruppe überleben) wird die Nullhypothese abgelehnt (bei einem Niveau von 10%) und bei der zweiten Untersuchung mit dem umgekehrten Ergebnis wird sie angenommen. Einmal komme ich also zu einer positiven Beurteilung des Medikaments und einmal zu einer negativen. Das ist einzig und allein abhängig von der gezogenen Stichprobe und widerspricht sich in keinster Weise. Das nennt man in der Statistik den Stichprobenfehler.

BERNALLI: Das heißt?

POISSIN: Das heißt ganz einfach, dass ich das Pech haben kann eine Stichprobe zu ziehen, die so ungewöhnlich ist, dass ich zu einer völlig falschen Entscheidung komme. Und die beiden hier vorgestellten Stichproben sind ja auch wirklich sehr extrem.

BERNALLI: Aber eine der beiden Beurteilungen muss doch falsch sein.

POISSIN: Richtig, aber man kann leider nicht feststellen welche. Das einzige, was man sagen kann, ist, wie wahrscheinlich es ist, dass man im ersten Fall die Nullhypothese fälschlicherweise abgelehnt hat, nämlich hier 10%.

GASS: Diese Wahrscheinlichkeit entspricht nämlich dem Signifikanzniveau. Man bezeichnet das in der Statistik übrigens als den sogenannten α-Fehler oder auch Fehler 1. Art.

POISSIN: Und ich kann sagen, wie wahrscheinlich es ist, dass ich im zweiten Fall die Nullhypothese fälschlicherweise angenommen habe.

GASS: Der β-Fehler oder Fehler 2. Art. Diese Wahrscheinlichkeit ist allerdings relativ schwer zu berechnen und auch davon abhängig, wie man die Alternativhypothese quantifiziert. Wir sollten uns damit nicht aufhalten.

BERNALLI: Welches Fazit können wir bisher ziehen?

POISSIN: Wie ich gerade schon sagte: Wir haben hier zwar zwei unterschiedliche Aussagen über die Wirksamkeit des Medikaments. Die sind aber aus zwei sehr unterschiedlichen Stichproben hervorgegangen. Daher kommen wir zu dem Schluss, dass die Aussagen durchaus konsistent sind.

GASS: Kommen wir aber jetzt zum Kern der Salviati'schen Thesen, nämlich, dass Signifikanzniveau und Irrtumswahrscheinlichkeit nicht das gleiche sind. Salviati macht das an zwei Beispielen fest (Tabellen 3 und 4 im Text von Weihe). Zunächst sollten wir die Definitionen der beiden Begriffe, wie sie im Text vorgegeben sind, vergleichen.

BERNALLI: Also, das Signifikanzniveau wird definiert als die Wahrscheinlichkeit, dass die Studie ein positives Ergebnis zeigt, obwohl die untersuchten Behandlungen sich nicht unterscheiden. Ich versuche das gleich einmal in die Terminologie der Testtheorie zu übersetzen: Ein positives Ergebnis heißt, dass ein Medikament als wirksam eingestuft wird, also dass die Nullhypothese abgelehnt wird. Gleichzeitig soll es in Wirklichkeit aber keinen Unterschied geben. Die Nullhypothese wird also fälschlicherweise abgelehnt. Unser α-Fehler!

GASS: Richtig!

BERNALLI: Weiter geht es: Die Irrtumswahrscheinlichkeit wird definiert als die Wahrscheinlichkeit, „ein zufälliges Ergebnis für bare Münze zu nehmen". Hm, ein bisschen volkstümlich ausgedrückt. Was heißt das jetzt?

POISSIN: Im Rahmen der Testtheorie taucht der Zufall nur an einer Stelle auf, nämlich im Stichprobenergebnis. Das Stichprobenergebnis wird in den allermeisten Fällen von der Nullhypothese abweichen. Die Frage ist, ist diese Abweichung zufällig, also durch den Stichprobenfehler erklärbar oder ist die Wahrscheinlichkeit dafür so gering...

BERNALLI: Unter dem Signifikanzniveau!

GASS: Genau!

POISSIN: ..., dass die Abweichung nicht mehr zufällig ist, sondern systematisch.

BERNALLI: Ich übersetze wieder: Wir haben also ein zufälliges Ergebnis, sprich eine zufällige Abweichung zwischen Stichprobe und Nullhypothese. Das heißt die Nullhypothese gilt. Gleichzeitig nehmen wir die zufällige Abweichung als „bare

Münze“, lehnen die Nullhypothese also ab. Das ist doch das gleiche wie oben, nämlich der Fehler 1. Art!!

GASS: Das sehe ich auch so. Hinter beiden Begriffen versteckt sich die Wahrscheinlichkeit, dass ein Medikament positiv getestet wird, obwohl es völlig unwirksam ist. Das würde jeder Statistiker so auch unterschreiben. Betrachten wir aber jetzt das Beispiel von Simplicio. Er vergleicht zwei Forschergruppen, wobei Gruppe A jeden Wirkstoff zur Überprüfung annimmt und in 10% aller Fälle ein positives Ergebnis herausbekommt. Gruppe B stellt einige Vorüberlegungen an und nimmt eine Auswahl vor, was dazu führt, dass bei Gruppe B 40% der Medikamente positiv getestet werden.

BERNALLI: Moment! Das steht zwar so im Text, ist in den beiden Tabellen aber ganz anders umgesetzt. Die Tabellen sagen aus, dass 10% bzw. 40% der Medikamente in den beiden Gruppen wirksam sind, nicht dass sie so getestet wurden. Das ist etwas ganz anderes.

GASS: Da stimmt. Hier klaffen Text und Tabellen deutlich auseinander. Nehmen wir aber die Tabellen als weitere Arbeitsgrundlage, also nehmen wir an, dass 10% bzw. 40% wirklich wirksam sind. Was berechnet der Autor dann unter dem Stichwort der Irrtumswahrscheinlichkeit?

POISSIN: Ich zitiere: Die „Wahrscheinlichkeit, dass bei positivem Studienergebnis das neue Medikament wirklich besser ist“ Also, die Wahrscheinlichkeit, dass ein Medikament wirksam ist, wenn es positiv getestet wurde.

BERNALLI: Das ist ja etwas ganz anderes als oben in der Definition angegeben wurde!

GASS: Eben!

BERNALLI: Aber die Zahlen bleiben doch bestehen. Wenn ein Medikament von Gruppe A als positiv getestet wurde, beträgt die Wahrscheinlichkeit, dass es auch tatsächlich wirksam ist, nur 36%, während die Wahrscheinlichkeit bei Gruppe B bei 91% liegt. Das heißt doch, dass ich als Pharmaunternehmen lieber bei B testen lassen würde, weil die Wahrscheinlichkeit, dass mein Medikament nach einem positiven Test, tatsächlich wirksam ist, dann viel höher ist. Aber es bleibt doch das gleiche Medikament. Das kann doch gar nicht sein.

GASS: Kann es auch nicht. Es fehlt nämlich ein ganz entscheidender Teil in der Aussage: Damit B ein Medikament positiv testen kann, muss er es erst einmal zum Test annehmen. Das ganze ist nämlich eine bedingte Wahrscheinlichkeit: Wenn B ein Medikament annimmt und positiv testet, dann ist die Wahrscheinlichkeit für die Wirksamkeit 91%. Wir wissen aber überhaupt nichts über die Annahmequote des B. Bei A ist das kein Problem. Der nimmt alles an. Bei B hängen diese 40% (und die daraus folgenden 91%) aber ganz massiv von der Annahmequote ab. Nehmen wir einfach einmal an, B würde 70% aller Medikamente annehmen, dann wäre die „Irrtumswahrscheinlichkeit“ (mir fällt kein besseres Wort ein) $1-0{,}91\cdot 0{,}7=0{,}363$ und liegt somit im Bereich von A. Es ist sicherlich nicht zu gewagt, wenn man sagt, dass die Wahrscheinlichkeit für die Wirksamkeit eines Medikaments stark abnehmen würde, wenn B mehr Medikamente zum Test annehmen würde. Im Extremfall landet er bei den 10% von A.

BERNALLI: Welches Fazit ziehen wir also?

GASS: Zunächst das, dass die drei Herren hier einige Begriffe recht großzügig verwenden, aber das ist nicht das Problem. Als problematisch sehe ich an, dass das

Beispiel mit den beiden Forschungsgruppen nicht zu Ende gedacht wurde und das hier versucht wird Aussagen in die Testprozedur hinein zu interpretieren, die schlicht überhaupt nicht vorhanden sind. Ich denke, wir haben deutlich gemacht, was ein Test leisten kann und mehr sollte man auch nicht verlangen. Und was die Forschungs- und Veröffentlichungspraxis der Mediziner angeht, die hier zum Teil kritisiert wird: Das ist deren Problem. Damit müssen die zurechtkommen. Dafür kann man nicht die Statistik verantwortlich machen.