

# Die wahre Bedeutung der statistischen Signifikanz

Hans-Hermann Dubben & Hans-Peter Beck-Bornholdt

© Dubben & Beck-Bornholdt

## **Zitierung:**

Dubben HH, Beck-Bornholdt HP: Die Bedeutung der statistischen Signifikanz. In: Diekmann, Andreas (Hg.), 2006: Methoden der Sozialforschung. Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Wiesbaden: VS-Verlag für Sozialwissenschaften. ISBN 3-531-14362-X

## **Korrespondenzanschrift**

PD Dr. rer. nat. Hans-Hermann Dubben  
Institut für Allgemeinmedizin  
Universitätsklinikum Hamburg-Eppendorf  
Martinistrasse 52  
20246 Hamburg  
dubben@uke.de

### Zusammenfassung

Zunächst die gute Nachricht: es besteht Hoffnung, dass in naher Zukunft die gewaltige und üblicherweise ungelesene Flut biomedizinischer Publikationen mit so genannten „statistisch signifikanten“ Ergebnissen sich zu einem übersichtlichen und lesenswerten Bach geschrumpft. Die schlechte Nachricht: die vorherrschende Interpretation von p-Werten ist unkorrekt und begünstigt die Produktion „statistisch signifikanter“, aber falscher Ergebnisse.

Stellen Sie sich eine perfekt durchgeführte klinische Studie vor. Die experimentelle Therapie zeigt eine höhere Überlebensrate als die Standardtherapie. Sie wissen, dass dieser Vorteil auch in methodisch perfekten Studien durch zufällige Schwankungen der Behandlungsergebnisse zustande gekommen sein kann. Deshalb stellen Sie einem Statistiker die klinisch relevante Frage: „Wie wahrscheinlich ist es, dass ich mich irre, wenn ich die experimentelle Therapie für besser halte?“ Nach ein paar Berechnungen kommt die Antwort: „Der beobachtete Unterschied ist statistisch signifikant ( $p=0,03$ ).“ Damit ist ein wichtiges Kriterium für eine Publikation erfüllt, aber: *Was bedeutet dieser Satz?*

Die Antwort des Statistikers bedeutet *nicht*, was die meisten glauben: „Wenn ich die experimentelle Therapie für besser als die Standardtherapie halte, beträgt die Irrtumswahrscheinlichkeit 3 Prozent.“ *Der p-Wert ist nicht die Antwort auf Ihre Frage!* Der p-Wert besagt lediglich: „Wenn beide Therapien in Wirklichkeit gleichwertig sind, dann kann die beobachtete (oder eine größere) Differenz mit 3 Prozent Wahrscheinlichkeit zufällig auftreten.“ Der Unterschied dieser beiden Sätze ist keine Haarspalterei, sondern möglicherweise der häufigste und schwerste Irrtum der biomedizinischen Forschung.

Der zugrunde liegende Irrtum wird an einem alltäglichen Beispiel und anhand „statistisch signifikanter“ Ergebnisse erläutert. Es wird gezeigt, dass p-Werte allein kein Maß für wissenschaftliche Evidenz sind.

**Stichworte:** Statistische Signifikanz; Signifikanztest; Hypothesentest; p-Wert; Irrtumswahrscheinlichkeit; Prädiktiver Wert; Bayesianische Statistik; Bayes

## **Einleitung**

Signifikanztests haben eine sehr große Bedeutung in der gegenwärtigen wissenschaftlichen Kultur. Kaum eine Publikation mit quantitativen Resultaten kommt ohne Statistik und ohne Signifikanztest aus. Sie werden als Ausdruck von Wissenschaftlichkeit angesehen und sind meist auch Eingangsvoraussetzung dafür, dass ein Manuskript zur Publikation überhaupt angenommen wird. Das Ziel statistischer Erwägungen ist es, wahre Ergebnisse möglichst zuverlässig von Zufallstreffern zu unterscheiden. Inwieweit dies tatsächlich auch erfüllt wird, soll in dieser Arbeit hinterfragt werden.

Alltägliche Beispiele werden zeigen, dass der Grundgedanke von Signifikanztests sehr einfach ist. Auch wird höhere Mathematik nicht von Nöten sein; vielleicht mal ein Dreisatz oder ein Blick in eine Abbildung oder eine Tabelle.

## **Fußball-Statistik**

Letzte Woche trafen der FC Aheim und der SC Beheim gegeneinander an. Die beiden Mannschaften sind wie zwei gleiche Klone. Exakt gleich gut, sagen die Kenner der Szene. Gestern hat der FC Aheim gewonnen. Niemanden hat es gewundert. Neulich beim Freundschaftsspiel war es andersrum.

Heute spielte der SC Beheim gegen TSV Versager. Der TSV ist unbestritten sehr viel schlechter als der SC, aber heute gab es immerhin ein Unentschieden. Das hat auch niemanden gewundert. Hätten sie länger gespielt, hätte der bessere sich sehr wahrscheinlich durchgesetzt.

Bemerkenswert ist, dass das Ergebnis eines einzelnen Spieles nicht unbedingt repräsentativ dafür ist, welche der beiden Mannschaften die bessere ist. Falls es überhaupt eine bessere gibt. Etwas anders sieht es bei einem sehr hohen Sieg von 6:0 oder extremer aus. Dann ist das Ergebnis sogar statistisch signifikant (Dubben & Beck-Bornholdt, 2006).

## **Was will der Verbraucher von Studienergebnissen wissen?**

In einer klinischen Studie hat ein Arzt die Wirksamkeit zweier Behandlungen, Therapie A und Therapie B, miteinander verglichen. In der Gruppe, die Therapie A erhalten hat, ist die 5-Jahres Überlebensrate um 10 Prozentpunkte höher als in der anderen Gruppe. Können wir nun sicher sein, dass Therapie A tatsächlich besser ist als Therapie B?

Nein, das können wir nicht. Dazu müssen wir nur an die Fußballergebnisse denken. Therapie A und Therapie B sind ein einziges Mal gegeneinander angetreten. Das Studienergebnis gibt nicht mit Sicherheit wieder, welche Therapie die bessere ist, falls es

überhaupt eine bessere gibt. Eine Studie kann man nicht so einfach wiederholen wie ein Fußballspiel und die nächsten Patienten sitzen auch bereits im Wartezimmer. Im Interesse seiner Patienten möchte unser Arzt daher sofort so viel wie möglich aus seiner Studie folgern. Er weiß, dass zufällige Schwankungen den Unterschied auch dort vortäuschen können, wo keiner ist. Der Arzt möchte daher wissen:

*"Wie groß ist die Wahrscheinlichkeit, dass ich mich irre, wenn ich Therapie A für besser als Therapie B halte?"*

Mit dieser Frage wendet er sich an einen Experten, beispielsweise an einen Biomathematiker oder Statistiker. Nach Durchführung üblicher, aber für Normalsterbliche undurchsichtige Berechnungen, erhält der Arzt folgende Antwort:

*„Therapie A ist statistisch signifikant besser als die Standardbehandlung ( $p=0,03$ ).“*

Dieser Satz wird wörtlich in die nun fällige Publikation übernommen und das Ergebnis mit dem Gütesiegel „ $p=0,03$ “ versehen.

Damit wurde einem üblichen, aber nicht unbedingt sinnvollen Ritual Folge geleistet. Die Frage des Arztes wurde damit nicht beantwortet. Der Satz *„Therapie A ist statistisch signifikant besser als die Standardbehandlung ( $p=0,03$ ).“* bedeutet nicht das, was die meisten glauben. Er bedeutet nicht: *„Wenn ich Therapie A für besser als Therapie B halte, dann irre ich mich mit einer Wahrscheinlichkeit von 3%“*. Letztere Interpretation ist schlicht falsch, auch wenn es in ansonsten grundsoliden Statistikbüchern so stehen mag. Häufig werden Signifikanzniveau und Irrtumswahrscheinlichkeit verwechselt.

Die wahre Bedeutung des fraglichen Satzes ist eine ganz andere:

*„Falls die beiden Therapien gleich gut sind, dann beträgt die Wahrscheinlichkeit 3%, dass die beobachteten oder noch extremere Ergebnisse auftreten“*.

Der Unterschied zwischen der falschen Interpretation und der wahren Bedeutung ist keine Haarspalterei, sondern möglicherweise der häufigste und folgenreichste Irrtum der modernen internationalen medizinischen Forschung. Woher weiß ich, ob die Therapien gleich gut sind? Und was ist, wenn sie nicht gleich gut sind (was ich ja vielleicht sogar hoffe)? Im Moment stehen wir noch im Dunkeln. Um ein erstes Licht in die Sache zu bringen, machen wir einen virtuellen Kurzurlaub im fernen Syldavien.

### **Der Fischttest: Angeln in Syldavien**

Syldavien lockt mit endlosen Wäldern und Seen. Einzig und allein in der Syldavischen Seenplatte kann man die schmackhaften Leckerellen angeln (Abb.1). Außer diesen gibt es hier nur noch eine einzige andere Fischart: die Ekelitzen. Die schmecken wie sie heißen und selbst hungrige Katzen greifen da lieber zum Dosenöffner. Damit Sie möglichst viele Leckerellen angeln, haben Sie sich einen zertifizierten und entsprechend teuren Köder beschafft. Dieser Köder verspricht, dass 90 Prozent der Leckerellen anbeißen, die sich dem Köder auf weniger als einen Meter genähert haben. Die meisten Ekelitzen hingegen verschmähen den Köder. Wenn sie sich dem Köder auf weniger als einen Meter genähert haben, beißen sie nur in 5% der Fälle an.

Sie genießen die Ruhe, den Blick in den wolken- und mückenlosen syldavischen Himmel gerichtet ... und plötzlich beißt ein Fisch an. Wie groß ist nun die Wahrscheinlichkeit, dass das eine Ekelitze ist? Bitte sehen Sie sich den Fisch nicht an und lesen Sie nicht weiter, ohne über die Frage nachgedacht zu haben!



Abb. 1: Angeln in Syldavien

Richtig! Die Frage ist so gar nicht beantwortbar. Die Wahrscheinlichkeit, dass eine Ekelitze am Haken hängt, hängt davon ab, wie groß der Leckerellen-Anteil am Fischbestand des Sees ist. Betrachten wir doch einfach einmal die Extreme. Wenn es keine Ekelitzen mehr gibt, dann zappelt immer eine Leckerelle am Haken. Das andere Extrem ist, dass es keine Leckerellen mehr gibt. Dann nützt Ihnen der beste Köder nichts und Sie haben nie eine Leckerelle gefangen, wenn es am Haken zieht. Nur mit den Anbißwahrscheinlichkeiten auf der Gebrauchsanleitung des Köders können sie Ihre Frage offensichtlich nicht beantworten. Sie benötigen noch eine dritte Größe: den Anteil der Leckerellen.

### **Testeigenschaften eines Feuermelders**

Etwas technischer, aber immer noch sehr alltäglich ist die Arbeitsweise eines Feuermelders. Sie ist in Tabelle 1 schematisch dargestellt. Ein Haus kann für diese Betrachtung genau zwei relevante Dinge tun: es brennt oder es brennt nicht. Der

Feuermelder selbst bietet auch keine große Vielfalt an Aktivitäten: entweder er schlägt Alarm oder er lässt es bleiben.

Wenn es brennt, wünschen wir uns, dass er Alarm gibt. Ein schlechter Feuermelder kann an dieser Stelle versagen, und trotz Feuer keinen Alarm geben. Das kann sehr gefährlich und sehr teuer werden. Eine hohe Ansprechwahrscheinlichkeit, die im Falle eines Brandes möglichst zuverlässig zu einem Alarm führt, ist also wünschenswert und zeichnet einen guten Feuermelder aus.

	Feuermelder	
	Alarm	Kein Alarm
Das Haus brennt	Richtiger Alarm ( <i>Ansprechwahrscheinlichkeit</i> )	falsch
Das Haus brennt nicht	Falscher Alarm ( <i>Fehlalarmwahrscheinlichkeit</i> )	richtig
Wahrscheinlichkeit, dass ein Alarm richtig ist:	Anzahl richtiger Alarme / Summe aller Alarme	

Tab.1: Testeigenschaften eines Feuermelders. Erläuterung siehe Text.

Wenn es nicht brennt, wollen wir keinen Alarm. So ein Fehlalarm kann ebenfalls gefährlich und teuer werden. Wenn die ganze Belegschaft eines Bürohauses durch das Treppenhaus auf die Straße stürmt, sind blaue Flecken und umgeknickte Füße nicht auszuschließen. Hinzu kommt der Arbeitsausfall und eventuell die Kosten für einen unnötigen Feuerwehreinsatz. Kurzum: Eine geringe Fehlalarmwahrscheinlichkeit ist wünschenswert und zeichnet ebenfalls einen guten Feuermelder aus.

Wichtig zu merken: Ein Feuermelder hat zwei Qualitätsmerkmale. Wenn es eine guter Feuermelder ist hat er 1.) eine geringe Fehlalarmwahrscheinlichkeit (d.h. selten Alarm wenn es nicht brennt) und 2.) eine hohe Ansprechwahrscheinlichkeit (d.h. sehr oft Alarm, wenn es tatsächlich brennt).

Der Gebrauchsanleitung ihres gerade erstendenden Feuermelders entnehmen sie folgendes: Die Fehlalarmwahrscheinlichkeit beträgt 5 Prozent. An 100 feuerfreien Tagen gibt es also 5 Fehlalarme (es gibt der Einfachheit halber maximal einen Alarm pro Tag). Die Ansprechwahrscheinlichkeit beträgt laut Gebrauchsanleitung 90 Prozent. Wie häufig werden Sie in hundert Tagen (es gibt der Einfachheit halber maximal einen Brand pro Tag) durch Fehlalarm aufgeschreckt und wie häufig ist es ein echter Alarm? Bitte denken Sie darüber nach!

Richtig, man muss wissen, wie häufig es dort, wo der Feuermelder steht, überhaupt brennt. Steht er völlig allein auf einem unbrennbaren Betonsockel im friesischen Watt

fernab etwaiger Öl- oder Erdgasfelder, dann ist es wohl immer ein Fehlalarm, wenn das Gerät Alarm schlägt. Je mehr die Umgebung jedoch feuergefährdet ist, umso wahrscheinlicher wird es ein echter Alarm sein. Um die obige Frage zu beantworten, werden wie bei den Leckerellen insgesamt drei Größen benötigt: die Fehlalarm- und die Ansprechwahrscheinlichkeit, die beide ausschließlich Eigenschaften des Feuermelders sind, sowie eine Aussage darüber, wie häufig es in der Umgebung des Feuermelders überhaupt brennt.

### **Signifikanztests in klinischen Studien**

Prof. McResult hat eine neue Therapie entwickelt (wofür oder wogegen spielt hier keine Rolle) und an 50 Patienten ausprobiert. Es wird ein etwas größerer Anteil Patienten geheilt als mit der Standardtherapie, die seit Jahren im ganzen Lande eingesetzt wird. Nur ein Patient hatte bei Prof. McResult eine Nebenwirkung. Das entspricht 2 Prozent. Die Standardtherapie, an tausenden von Patienten erprobt, hat ebenfalls 2 Prozent derselben Nebenwirkung. Auf einer eilig einberufenen Pressekonferenz verkündet Prof. McResult, dass die neue Therapie besser wirkt bei gleich bleibender Rate an Nebenwirkungen. Wer jetzt noch nach der alten Therapie verfähre, der beginge einen Kunstfehler und mache sich strafbar.

Vorsicht, Vorsicht! Zwei Dinge können Prof. McResult in seiner Euphorie entgangen sein. Erstens: die Ergebnisse seiner Therapie waren in dieser einen Studie nur zufällig besser. Wenn zwei gleich starke Fußballmannschaften gegeneinander spielen, ist das Ergebnis auch nicht immer unentschieden. Zweitens: Hätte nur einer mehr von McResults Patienten eine Nebenwirkung gehabt, dann hätte die Nebenwirkungsrate bereits 4 Prozent betragen und somit mehr (das Doppelte!) als in der Standardtherapie. Mit anderen Worten, es kann gut sein, dass Prof. McResult etwas wichtiges übersehen hat. Wenn zwei ungleiche Fußballmannschaften aufeinander treffen, kann es eben trotzdem ein „unentschieden“ geben. Mit welcher Wahrscheinlichkeit der eine oder der andere Fehler begangen werden, sind Fragen, deren Beantwortung Sache der Statistik ist. Damit sind wir beim Signifikanztest angelangt sind.

Erstaunlicherweise haben ein Signifikanztest in einer klinischen Studie und ein Feuermelder sehr viel gemeinsam (Tabellen 1 und 2). In einer Studie werden zwei Therapien A und B verglichen. Die beiden Therapien können in diesem Zusammenhang nur zwei relevante Eigenschaften haben: sie sind tatsächlich unterschiedlich oder sie sind es nicht. Der Signifikanztests selbst bietet wie der Feuermelder nur zwei Varianten: er gibt dem Ergebnis das Prädikat „statistisch signifikant“ oder er tut es nicht.

Wenn Therapie A und Therapie B tatsächlich unterschiedlich sind, möchten wir, dass der Signifikanztest uns dies mit dem Gütesiegel „statistisch signifikant“ möglichst zuverlässig anzeigt. Wie beim Feuermelder wünschen wir uns eine möglichst hohe Ansprechwahrscheinlichkeit. Sie hat in diesem Zusammenhang einen anderen Namen;

sie wird *Power* genannt. Wie fast alle Vergleiche hinkt auch dieser, allerdings ohne die Analogie zu zerstören. Die Ansprechwahrscheinlichkeit des Feuermelders hängt nicht unbedingt von der Größe des Brandes oder dem brennenden Material ab. Die *Power* einer Studie ist um so größer, je größer der Unterschied der Wirkung von Therapie A und Therapie B ist. Schließlich ist es ja auch viel wahrscheinlicher ein im Heuhaufen verlorenes Hufeisen wieder zu finden, als eine Stecknadel. Die *Power* steigt auch mit der Anzahl der Patienten, die in einer Studie ausgewertet werden. Dies sagt vielen bereits die Intuition. Einer Studie mit 1000 Patienten trauen wir es eher zu, einen Unterschied von fünf Prozentpunkten z.B. zwischen zwei Nebenwirkungsraten aufzudecken, als einer Studie mit nur 100 Patienten oder gar nur 10 Patienten.

Von den drei Größen *Power*, Unterschied und Anzahl der Patienten muss man zwei kennen, um die dritte berechnen zu können. Die für die Studienplanung wichtige Frage „Wie viele Patienten brauche ich für eine Studie mit 80 Prozent *Power*?“ scheint damit auf den ersten Blick unbeantwortbar, weil der Unterschied ja nicht bekannt ist. Würde ich ihn kennen, müsste und dürfte ich die Studie gar nicht durchführen. Entscheidend ist in dieser Situation nicht der tatsächlich vorhandene Unterschied, sondern wie groß der Unterschied sein müsste, damit er relevant ist und auch praktische Konsequenzen aus dem Ergebnis gezogen werden würden. Ein Hustensaft, der bei 6200 von 10000 Patienten wirkt, statt nur bei 6000 wie das Vergleichspräparat, hat rein mathematisch zunächst die Nase vorn. Aber wohl kaum ein Praktiker wird in diesem Fall die Verbesserung um zwei Prozentpunkte als relevant erachten. Man muss sich vor der Studie im klaren sein über den „minimalen klinisch relevanten Unterschied“, den man als eine Verbesserung ansehen würde.

Für die weiteren Belange dieser Abhandlung ist wichtig zu merken: Die *Power* ist die Wahrscheinlichkeit, den minimalen klinisch relevanten Unterschied aufzuzeigen, sofern er tatsächlich vorhanden ist. Wünschenswert ist natürlich eine möglichst hohe *Power*. In klinischen Studien wird meistens eine *Power* von 80 oder 90 Prozent angestrebt.

Eine *Power* von 80 Prozent bedeutet gleichzeitig, mit 20 Prozent Wahrscheinlichkeit (= 100 Prozent – *Power*) den vorhandenen Unterschied zu übersehen. Dies entspricht dem Versagen des Feuermelders bei einem Brand. Das Übersehen hat in der Statistik den Namen „Fehler zweiter Art“. Einen Unterschied zu übersehen, kann bedeuten, dass eine zum Greifen nahe liegende Therapieverbesserung doch nicht zum Einsatz kommt oder dass die Erhöhung einer Nebenwirkungsrate nicht bemerkt wird.

Tabelle 2 zeigt im Vergleich mit Tabelle 1 die Analogie zum Feuermelder. Bisher wurde der Fall erläutert, in dem Therapie A besser ist als Therapie B.

	Signifikanztest	
	Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	Richtig signifikantes Ergebnis ( <i>Power</i> )	Falsch (Fehler 2.Art)
Therapie A ist nicht besser als Therapie B	Falsch signifikantes Ergebnis oder Fehler 1.Art ( <i>p-Wert</i> )	Richtig
Positiver prädiktiver Wert:	Anzahl richtig signif. Ergebnisse / Summe aller signif. Ergebnisse	

Tab.2: Die Tabelle und darin enthaltene Begriffe werden nach und nach im Text erläutert.

Was ist, wenn die beiden Therapien sich tatsächlich nicht unterscheiden? Um diesen Fall richtig zu verstehen, ist es günstig, sich zunächst in ein Würfelspiel hineinzudenken.

Sie haben einen schwarzen und einen weißen Würfel. Sie werfen beide, schwarz zeigt eine „5“, weiß eine „3“ (Abb.2). Schwarz hat jetzt zwar gewonnen, aber würden Sie folgern, dass der schwarze Würfel tatsächlich besser (d.h. so präpariert ist, dass er systematisch höhere Zahlen zeigt) ist als der weiße? Wahrscheinlich nicht, denn ob ein tatsächlicher Unterschied zwischen den beiden Würfeln besteht, bekommen Sie nur heraus, indem sie mehrmals werfen. In der Forschung hieße dies, eine ganze Studie zu wiederholen, was natürlich weitaus schwieriger (u.U. sogar unmöglich) und zeitaufwändiger ist als einfach noch einmal zu würfeln.

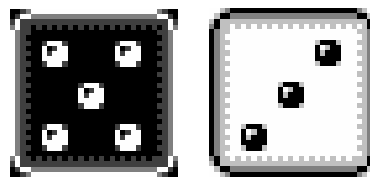


Abb. 2: Ergebnisse einer spielerischen Studie, in der ein schwarzer Würfel mit einem weißen Würfel verglichen wurde.

In Tabelle 3 sind alle Ergebnisse aufgelistet, die bei der Begegnung „Schwarzer Würfel gegen Weißer Würfel“ möglich sind. Wenn beide Würfel „gleichermaßen gut“ sind, also keiner irgendwie präpariert („gezinkt“) ist, dann sind all diese Ergebnisse gleich wahrscheinlich. Es gibt 36 Möglichkeiten. Sie sind hier nach der Differenz Schwarz – Weiß sortiert. In 10 von 36 Fällen, also  $10/36 = 0,28 = 28$  Prozent der Fälle hat schwarz zwei oder mehr Punkte Vorsprung vor weiß. Diese 28 Prozent sind der p-Wert des 5:3 Ergebnisses. Die Frage: „Wie wahrscheinlich ist es, dass schwarz zufällig 2 oder mehr

Punkte Vorsprung hat, obwohl beide Würfel gleich gut sind?“ wird durch diesen p-Wert beantwortet.

	Schwarz	Weiß	Differenz		Schwarz	Weiß	Differenz
1	6	1	5	19	3	3	0
2	6	2	4	20	2	2	0
3	5	1	4	21	1	1	0
4	6	3	3	22	5	6	-1
5	5	2	3	23	4	5	-1
6	4	1	3	24	3	4	-1
7	6	4	2	25	2	3	-1
8	5	3	2	26	1	2	-1
9	4	2	2	27	4	6	-2
10	3	1	2	28	3	5	-2
11	6	5	1	29	2	4	-2
12	5	4	1	30	1	3	-2
13	4	3	1	31	3	6	-3
14	3	2	1	32	2	5	-3
15	2	1	1	33	1	4	-3
16	6	6	0	34	2	6	-4
17	5	5	0	35	1	5	-4
18	4	4	0	36	1	6	-5

Tabelle 3: Mögliche Ergebnisse beim Werfen von zwei Würfeln

Übertragen auf eine klinische Studie lautet die Frage: „Wie wahrscheinlich ist es, dass die Ergebnisse von Therapie A und Therapie B sich so wie beobachtet oder extremer unterscheiden, obwohl beide Therapien gleichermaßen wirksam sind?“ Die Antwort ist der p-Wert. Dessen Berechnung ist allerdings selten so übersichtlich wie bei unserem Würfelspiel. Der grundlegende Gedankengang ist derselbe und nur um den soll es hier gehen.

Wann ist denn das Ergebnis nun statistisch signifikant? Die Antwort ist kurz und einfach: wenn der p-Wert weniger als 5 Prozent beträgt. Dieses so genannte 5-Prozent-Signifikanzniveau fällt vom Himmel. Es ist reine Konvention. Ob diese 5 Prozent sinnvoll sind, werden wir später näher betrachten. Aber wir sollten es hier schon einmal auf die Würfel anwenden. Hätte schwarz mit einer „6“ gegen eine weiße „1“ gewonnen, hätten wir einen p-Wert von  $1/36 = 0,028 = 2,8 \text{ Prozent}^1$  erhalten, denn nur in einem von 36 Fällen ist schwarz um 5 Punkte (oder mehr) besser als weiß. In der medizinischen Forschung würde man nun nach einem einzigen Wurf sagen können, dass der schwarze Würfel „statistisch signifikant“ besser ist als der weiße und meistens daraus folgern, dass er tatsächlich und systematisch besser ist.

Nun kehren wir zurück zu der Frage, was ist, wenn die beiden Therapien sich tatsächlich nicht unterscheiden. Mit der gerade wiedergegebenen Definition der statistischen

<sup>1</sup> Für Insider: Dies gilt für einen einseitigen Signifikanztest, der voraussetzt, dass schwarz nur besser, aber nicht schlechter als weiß sein kann.

Signifikanz und des p-Wertes werden mit 5 Prozent Wahrscheinlichkeit die Ergebnisse von Therapie A und Therapie B sich soweit unterscheiden, dass sie als statistisch signifikant gelten. Mit 5 Prozent Wahrscheinlichkeit wird daher ein Fehllalarm ausgelöst. In der Statistik wird der Fehllalarm „Fehler erster Art“ genannt.

Eine Studie zeichnet sich also durch zwei Angaben aus. Erstens, die *Power*, die angibt mit welcher Wahrscheinlichkeit ein tatsächlich vorhandener Unterschied aufgezeigt wird (tatsächlich vorhandene Leckerellen zum Beißen verführt werden oder bei einem tatsächlichen Brand der Feuermelder Alarm schlägt) und, zweitens, das Signifikanzniveau, das angibt, mit welcher Wahrscheinlichkeit bei tatsächlich nicht vorhandenem Unterschied Zufallsergebnisse für bare Münze genommen werden (eine Ekelitze am Haken hängt oder ein Fehllalarm ausgelöst wird).

Um Tabelle 2 mit Leben zu füllen, benötigen wir so etwas wie die Häufigkeit der Leckerellen oder der Brände. Sonst können wir nichts rechnen. Die Wahrscheinlichkeit, dass in einer Studie tatsächlich etwas untersucht wird, das besser ist als die Standardbehandlung („A ist besser als B“) soll im folgenden als „Wahrscheinlichkeit guter Ideen“ bezeichnet werden. In der nächsten Tabelle beträgt sie 10 Prozent. Die Zahl fällt vom Himmel, aber mit irgendetwas muss man anfangen. Dieser Punkt wird später noch diskutiert werden.

Unter insgesamt 1000 Studien (Tabelle 4) ist Therapie A entsprechend einer Wahrscheinlichkeit guter Ideen von 10 Prozent in 100 Studien tatsächlich besser als in Therapie B. Aufgrund der Power von 80 Prozent werden davon 80 richtig erkannt und erhalten das Prädikat „statistisch signifikant“. In 20 Fällen wird der Unterschied zwischen A und B übersehen. In 900 Studien ist Therapie A besser als B. Entsprechend der Definition des Signifikanzniveaus sind die Unterschiede in den Ergebnissen in 5 Prozent der Fälle zufällig so groß, dass sie fälschlicherweise als „statistisch signifikant“ bezeichnet werden. In insgesamt 45 Studien fallen wir also auf ein Zufallsergebnis herein.

Wir haben nun 125 Studien mit signifikantem Ergebnis. Davon gehören 80 zu einem tatsächlichen „A ist besser als B“. Die Wahrscheinlichkeit, dass ein signifikantes Ergebnis ein tatsächliches „A ist besser als B“ signalisiert, beträgt somit  $80/125 = 64$  Prozent. Dies ist der so genannte „positive prädiktive Wert“ der Studie. In  $100\% - 64\% = 36\%$  der Fälle irren wir uns jedoch.

Wenn wir unseren Forschern unterstellen, dass sie in jeder zehnten Studie eine wirklich gute Idee verfolgen, dass sie eine methodisch perfekte Studie durchführten und dass diese Studie eine Power von 80% bei einem Signifikanzniveau von 5% hatte, dann ist die Aussage etwa einer von drei derartiger Studien trotzdem ein Irrtum!

	Anzahl Studien	Signifikanztest	
		Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	100	80% richtig signifikant: 80 Studien	Falsch nicht signifikant: 20 Studien
Therapie A ist nicht besser als Therapie B	900	5% falsch signifikant: 45 Studien	Richtig nicht signifikant: 855 Studien
Summe:	1000	125	
Positiver prädiktiver Wert:		80/125 = 64%	
Irrtumswahrscheinlichkeit:		100% - 64% = 36%	

Tab. 4: Positiver prädiktiver Wert einer einwandfreien klinischen Studie. Die Wahrscheinlichkeit vor der Studie, dass Therapie A besser ist als Therapie B, beträgt 10 Prozent („Wahrscheinlichkeit guter Ideen“). Die Power beträgt 80 Prozent, das Signifikanzniveau 5 Prozent.

Wie bei den Leckerellen und dem Feuermelder benötigen wir drei Größen, um den positiven prädiktiven Wert unserer Schlussfolgerung „A ist besser als B“ abschätzen zu können. Zwei Größen, die Eigenschaften des Tests sind (p-Wert, power), und eine, die das Gewässer charakterisiert, in dem wir fischen. Letztere, die Wahrscheinlichkeit für gute Ideen, ist unbekannt, da kein verlässliches Verfahren bekannt ist, mit dem man sie abschätzen könnte. Die Eingangs gestellte Frage „*Wie groß ist die Wahrscheinlichkeit, dass ich mich irre, wenn ich Therapie A für besser als Therapie B halte?*“ kann daher nicht beantwortet werden. Erst recht wird sie nicht beantwortet, wenn man zusätzlich noch die Power unberücksichtigt lässt und nur den p-Wert bzw. das Signifikanzniveau kennt.

### **Anzahl der Patienten und Power**

Sogar bei den in der methodischen Qualität hoch angesiedelten randomisierten kontrollierten Studien erreicht nur ein kleiner Teil tatsächlich eine Power von 80% (Freiman et al. 1992). Im Mittel (Median) haben real existierende Studien eine Power von lediglich 25 Prozent (hier gerechnet für einen Unterschied von 20 Prozentpunkten zwischen den verglichenen Therapien). Der Grund dafür ist einfach, dass die Studien viel zu wenig Patienten enthalten. Oft hat man das Patientenaufkommen überschätzt und daher nicht genug Patienten zusammenbekommen und einfach aufgehört oder man hat sich überhaupt gar keine Gedanken darüber gemacht.

**Signifikanzniveau: es gibt viele Arten, die 5-Prozent-Hürde zu nehmen**

Auch mit dem Signifikanzniveau wird es in der klinischen Forschung nicht so genau genommen. Mit Absicht oder aus Unwissenheit, das sei dahingestellt. Wenn wir in unserem Würfelspiel nicht einmal gewürfelt hätten, sondern zweimal, dann wäre das Ergebnis nicht mehr statistisch signifikant, denn die Wahrscheinlichkeit für einen „6:1-Sieg“ beim ersten oder zweiten Wurf ist natürlich viel größer als in ein „6:1-Sieg“ auf Anhieb. In der klinischen Forschung ist es nicht unüblich, zunächst einmal viele Daten zu sammeln, und dann zu schauen, ob nicht irgendetwas signifikantes dabei ist. Es werden dutzendweise Einflussparameter und Endpunkte auf Signifikanz getestet. Definitionsgemäß ist im Mittel jedes zwanzigste Ergebnis statistisch signifikant, entsprechend 5 Prozent. Und entsprechend vielfältig und widersprüchlich fallen die damit verbundenen klinischen Empfehlungen aus. Ein weiteres probates Mittel ist es, eine Studie zu starten, alle zwei Wochen einen Signifikanztest zu machen und die Studie in dem Augenblick für beendet zu erklären, wenn die favorisierte Therapie signifikant vorne liegt. Zum Vergleich: beim Fußball steht von vornherein fest, wie lange gespielt wird. Der Trainer darf das Spiel nicht abpfeifen, sowie seine Mannschaft den Führungstreffer erzielt hat. Pocock et al. (1987) fanden, dass in randomisierten Studien im Mittel über 6 Endpunkte berichtet und 4 Signifikanztests gemacht werden. Damit steigt die Wahrscheinlichkeit für einen Fehler erster Art um das vierfache. Das tatsächliche Signifikanzniveau beträgt jetzt nicht mehr 5 Prozent sondern 20 Prozent<sup>2</sup>. Pococks Feststellung stammt aus dem Jahre 1987. Die mittlere Anzahl an Signifikanztests pro Studie dürfte bis zum Jahre 2004 ganz enorm angestiegen sein, weil mittlerweile verfügbare Statistikprogramme es immer einfacher machen, innerhalb von Minuten Dutzende von Variablen und Subgruppen auf Signifikanz zu testen.

---

<sup>2</sup> Genau genommen beträgt es  $1 - 0,95 \times 0,95 \times 0,95 \times 0,95 = 1 - 0,81 = 0,19 = 19\%$

### Eine schon etwas realistischere Studie

Wir nehmen den realistischen Fall an, dass die Power einer Studie nur 25 Prozent beträgt. Ferner werden vier Signifikanztests statt einem gemacht, wodurch das effektive Signifikanzniveau 20 Prozent beträgt. Die einzelnen Zahlenwerte sind in Tabelle 5 eingetragen. Der positive prädiktive Wert beträgt in diesem Fall nur noch 12%. Nur jede achte Studie mit einem signifikanten Ergebnis zeigt eine tatsächliche Verbesserung an. Von 8 Publikationen ist nur eine brauchbar – aber welche?

	Anzahl Studien	Signifikanztest	
		Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	100	25% richtig signifikant: 25 Studien	Falsch nicht signifikant: 75 Studien
Therapie A ist nicht besser als Therapie B	900	20% falsch signifikant: 180 Studien	Richtig nicht signifikant: 720 Studien
Summe:	1000	205	
Positiver prädiktiver Wert:		25/205 = 12%	
Irrtumswahrscheinlichkeit:		100% - 12% = 88%	

Tabelle 5: Positiver prädiktiver Wert einer ansonsten einwandfreien klinischen Studie mit einer Power von 25 Prozent. Diese geringe Power ist der Normalfall in klinischen Studien. Es wurden vier Signifikanztests durchgeführt. Das effektive Signifikanzniveau beträgt daher 20%. Wahrscheinlichkeit vor der Studie, dass Therapie A besser ist als Therapie B: 10 Prozent.

### Ein pragmatischer Verbesserungsvorschlag

Offenbar führt der gegenwärtige Umgang mit Signifikanztests zu einer unvermeidbar hohen Fehlerquote. Glücklicherweise gibt es einen sehr einfachen und wirksamen Verbesserungsvorschlag (Sterne & Smith 2001): eine Power von 90 Prozent und ein Signifikanzniveau von 0,1%. Im Vergleich zum bislang angestrebten (90 Prozent Power und 5% Signifikanzniveau) erfordert dieser Vorschlag etwa doppelt so viele Patienten in einer Studie. Was es uns einbringt, können wir mit Tabelle 6 ausrechnen. Damit wir bei ganzen Zahlen bleiben können, wird jetzt von 10000 Studien ausgegangen. 10 Prozent davon, also 1000, untersuchen eine tatsächliche Verbesserung. Diese wird aufgrund der Power von 90 Prozent in 900 Studien auch festgestellt, indem das Ergebnis das Gütesiegel „statistisch signifikant“ erhält. Mit dem neuen Signifikanzniveau bringen die 9000 Studien, in denen Therapie A und Therapie B gleichwertig sind, 9 Fehlalarme ein. Der resultierende positive prädiktive Wert beträgt somit  $900/909 = 0,99 = 99$  Prozent.

	Anzahl Studien	Signifikanztest	
		Statistisch signifikant	Nicht stat. signifikant
Therapie A ist besser als Therapie B	1.000	90% richtig signifikant: 900 Studien	Falsch nicht signifikant: 100 Studien
Therapie A ist nicht besser als Therapie B	9.000	0,1% falsch signifikant: 9 Studien	Richtig nicht signifikant: 8991 Studien
Summe:	10.000	909	
Positiver prädiktiver Wert:		900/909 = 99%	
Irrtumswahrscheinlichkeit:		100% - 99% = 1%	

Tab.6: Positiver prädiktiver Wert von Studien mit 90% Power und einem Signifikanzniveau von 0,1%. Wahrscheinlichkeit vor der Studie, dass Therapie A besser ist als Therapie B: 10 Prozent („Wahrscheinlichkeit guter Ideen“).

Da macht das Lesen von Publikationen wieder Spaß, wenn man mit einer so hohen Gewissheit etwas dauerhaftes dazu lernen kann. Auch wenn die Wahrscheinlichkeit guter Ideen geringer ausfallen sollte als hier angenommen, ist der prädiktive Wert noch immer recht gut (Abb. 3). Mit 1 Prozent guten Ideen beträgt er 90 Prozent. Sollte nur jede tausendste Idee gut sein, sinkt der positive prädiktive Wert auf 40 Prozent. Immerhin wäre noch ungefähr jede zweite Studie es wert, sie zu lesen. Die gegenwärtige medizinische Forschung spielt sich deutlich unterhalb der gestrichelten Kurve in Abb.3 ab. Der prädiktive Wert ist entsprechend gering.

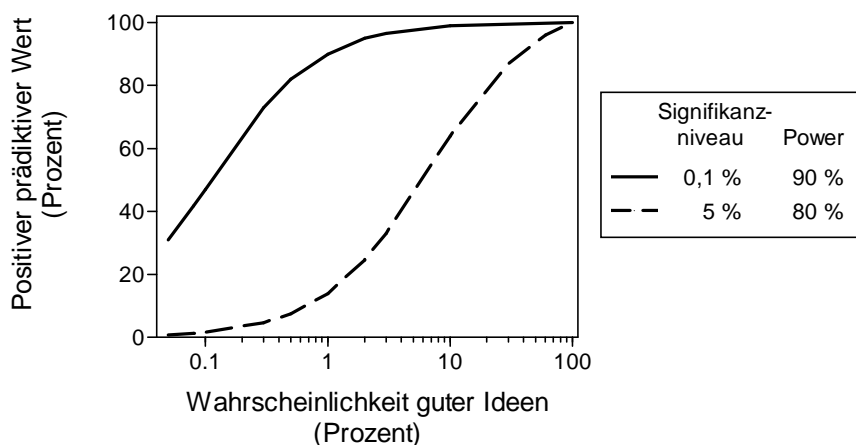


Abb. 3: Positiver prädiktiver Wert von Studien mit unterschiedlicher Wahrscheinlichkeit guter Ideen (Power und Signifikanzniveau wie angegeben). Der positive prädiktive Wert gibt an, mit welcher Wahrscheinlichkeit das Ergebnis einer Studie richtig ist. Die gegenwärtige medizinische Forschung spielt sich weit unterhalb der gestrichelten Kurve ab.

Sehr viel geringer fällt der positive prädiktive Wert für eine Power von 80 Prozent bei dem üblichen Signifikanzniveau von 5 Prozent aus. Da die genannte Power und das Signifikanzniveau wie oben besprochen in den wenigsten Studien tatsächlich erreicht wird, spielt sich die gegenwärtige medizinische Forschung weit unterhalb der gestrichelten Kurve ab.

### **Ein Gedanke zur Wahrscheinlichkeit guter Ideen.**

Zehn Prozent gute Ideen? Das ist doch viel zu wenig, oder? Greifen wir zu den Sternen: Hundert Prozent gute Ideen sind es ganz gewiss nicht. Dann könnte man sich gleich alle Studien sparen und müsste Wissenschaftler nur noch wie ein Orakel befragen. Wie wäre es mit 50 Prozent? Das bedeutet, dass ein Experte gleichfalls 50% Irrtumswahrscheinlichkeit hätte. Wenn sich fünf Experten einig sind über eine neue Behandlung, dann beträgt die Irrtumswahrscheinlichkeit nur noch  $0,5 \times 0,5 \times 0,5 \times 0,5 \times 0,5 = 0,03125 = 3,125\%$ . Das ist weniger als 5 Prozent. Eine derartige Studie wäre also bereits statistisch signifikant, bevor sie überhaupt begonnen hat.

### **Geschichtliches: ein alter Hut**

Die in diesem Artikel dargestellten Gedanken basieren im Wesentlichen auf den Überlegungen des englischen Geistlichen Thomas Bayes (1702 -1761). Sein wichtigstes Werk wurde erst nach seinem Tode von John Canton in den Jahren 1763 und 1764 in den *Philosophical Transactions of the Royal Society of London* (Band 53, Seiten 376-399 und Band 54, Seiten 298-310) unter dem Titel „An Essay Towards Solving a Problem in the Doctrine of Chances“ veröffentlicht. Diese Arbeit markiert einen wichtigen Wendepunkt in der Entwicklung der Wahrscheinlichkeitsrechnung.

### **Was wird für gute Studienplanung benötigt?**

Ein Signifikanzniveau von 0,1 Prozent statt der üblichen 5 Prozent ist wünschenswert. Das liefert erheblich weniger falsch positive Ergebnisse und damit auch weniger Literatur über falsch positive Ergebnisse. Die derzeit gigantische Verschwendung von forscherschem Enthusiasmus und materieller Ressourcen für sinnlose Zufallsergebnisse wird dadurch entscheidend reduziert.

Eine Studie sollte mindestens eine Power von 90 Prozent für den untersuchten Endpunkt haben, um relevante Unterschiede überhaupt aufdecken zu können und weil eine hohe Power maßgeblich zum prädiktiven Wert einer Studie beiträgt. Je weniger Patienten in einer Studie sind, um so geringer ist deren Power und um so geringer ist der prädiktive Wert der Studie. Und das auch bei positivem Ergebnis! Es ist für die Planung also unerlässlich, die benötigte Patientenzahl vor Beginn einer Studie sachlich abzuschätzen und mit dem tatsächlich vorhandenen Patientenaufkommen zu vergleichen.

Die Planer einer klinischen Studie sollten Experten auf dem medizinischen Gebiet sein, auf dem mit der Studie geforscht wird. Das erhöht die Wahrscheinlichkeit, dass in der Studie tatsächlich eine gute Idee verfolgt wird. Es ist unbedingt ratsam, bereits bei der Planung einen Statistiker mit einzubeziehen.

Der Einsatz korrekter Forschungsmethodik ist unerlässlich. Hier konnten nur wenige Punkte erwähnt werden. Man kann ganze Bücher darüber schreiben (z.B. Dubben & Beck-Bornholdt, 2005, 2006; Beck-Bornholdt & Dubben, 2003). Die zu messenden Endpunkte (Heilungsrate oder Blutdruck oder Überlebenszeit oder ...) müssen eindeutig und vor Beginn der Studie definiert werden. Damit dies für Außenstehende nachvollziehbar ist, sollte es ein Studienprotokoll geben, das nach Möglichkeit bei neutraler Stelle vor Studienbeginn hinterlegt wird.

### **Konsequenz für Studieninterpretation**

Dies ist der Abschnitt für alle, die Publikationen zu klinischen Studien lesen (müssen). Ein Großteil dessen, was in „wissenschaftlichen“ Journalen gedruckt wird, ist nicht valide. Die Kunst des Lesens besteht daher maßgeblich darin, möglichst frühzeitig zu erkennen, ob die Publikation, mit der man sich gerade beschäftigt, es überhaupt wert ist.

Wichtig ist, dass der p-Wert *nicht* die Irrtumswahrscheinlichkeit ist. In vielen Statistikbüchern steht dies anders. Die gegenwärtig normale Praxis, ausschließlich den p-Wert zu berücksichtigen, ist nicht geeignet, wahre Ergebnisse von Zufallstreffern hinreichend zuverlässig zu unterscheiden. Sie ist aber geeignet, zu „beweisen“, dass der Papst ein Außerirdischer ist (Beck-Bornholdt & Dubben, 1996).

Vorsicht ist geboten bei Studien mit geringer Power, sprich: bei Studien mit nur wenigen Patienten. Der Erkenntnisgewinn durch solche Studien ist gering. Auch bei positivem Ergebnis ist der prädiktive Wert sehr wahrscheinlich sehr gering. Dies wird von Statistikern häufig heruntergespielt oder gar verneint. Studien mit zu geringer Power haben geringe Aussagekraft.

Achten Sie darauf, dass ein primärer Endpunkt *a priori* festgelegt wurde. Bei mehreren und bei unklar definierten Endpunkten ist der wahre p-Wert eines Ergebnisses u.U. sehr viel größer als der angegebene p-Wert bzw. das angegebene Signifikanzniveau. Das Ergebnis ist dann nicht interpretierbar.

Last not least sollte man sich fragen, ob man das, was in einer Studie untersucht wurde, überhaupt für eine gute Idee hält. Und noch ein last-not-least: es gibt unermesslich viele weiterer Fehler und Irrtümer, mit denen die Aussagekraft teurer und aufwendiger Studien zunichte gemacht werden kann. Falls Sie ein paar davon kennen lernen möchten, und auch erfahren möchten, welches Unwesen sie im täglichen Leben treiben, empfehlen wir unsere Bücher.

**Literaturhinweise**

- Beck-Bornholdt HP* und *Dubben HH*, 2003: Der Schein der Weisen - Irrtümer und Fehltritte im täglichen Denken. Rowohlt Verlag.
- Beck-Bornholdt HP* und *Dubben HH*, 1996: Is the pope an alien? *Nature* 381: 730.
- Dubben HH* und *Beck-Bornholdt HP*, 2006: Der Hund, der Eier legt - Erkennen von Fehlinformation durch Querdenken. Rowohlt Verlag.
- Dubben HH*, *Beck-Bornholdt HP*, 2005: Mit an Wahrscheinlichkeit grenzender Sicherheit – Logisches Denken und Zufall. Rowohlt Verlag.
- Freiman JA*, *Chalmers TC*, *Smith H* und *Kuebler RR*, 1992: The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. S.357-373 in: J.C. Bailar III, Mosteller, F. (Hg.), *Medical Uses of Statistics*. New England Journal of Medicine Books, Boston, MA, USA.
- Ioannidis JPA*, 2005: Why Most Published Research Findings Are False. *PLoS Med* 2(8): e124.
- Pocock SJ*, *Hughes MD* und *Lee RJ*, 1987: Statistical problems in the reporting of clinical trials. A survey of three medical journals. *New England Journal of Medicine* 317:426-32.
- Sterne JAC* und *Smith GD*, 2001: Sifting the evidence — what’s wrong with significance tests? *British Medical Journal* 322, 226-231.