Peter von der Lippe

Irrtümer über die Irrtumswahrscheinlichkeit

Testtheorie nach T. Bayes in der medizinischen Statistik

Einen Artikel, wie der des Mediziners W. Weihe "Von der Wahrscheinlichkeit des Irrtums"[1] dürften wohl die meisten Statistiker ziemlich blamabel finden. Ich habe es jedenfalls so empfunden, ich kann mir aber auch gut vorstellen, dass umgekehrt Mediziner die Hände über den Kopf zusammenschlagen würden, wenn ich mich über medizinische Dinge äußern würde. Zugegeben, manchmal bringt der Mut des Nichtfachmanns Bewegung ins Geschehen, und er kann anregend und produktiv sein. Aber ein solcher Fall scheint hier nicht vorzuliegen. Weihes Ausführungen enthalten vielmehr einige katastrophale Mißverständnisse, und sie fordern vor allem in folgenden Punkten eine Erwiderung heraus:

1. Es wäre schön gewesen, wenn die angeblich so fehlerhafte Test-Logik klar herausgearbeitet worden wäre; weil dies jedoch nicht geschehen ist, wird in Abschnitt 1 versucht, dies nachzuholen, was um so wichtiger ist weil wir hierauf in den folgenden Abschnitten 2 und 3 aufbauen wollen.
2. Peinlich ist der belehrend vorgetragene Unsinn eines Unterschieds zwischen der Irrtumswahrscheinlichkeit und dem Signifikanzniveau. Dabei ist zu unterscheiden: Zwar ist die Kritik an der Unterschlagung nicht-signifikanter Ergebnisse in empirischen Forschungen durchaus berechtigt, aber die mysteriöse Veränderung der "Irrtumswahrscheinlichkeit" bei der Beurteilung der Wirksamkeit eines Medikaments durch die Existenz von anderen Forschern mit einem bessern oder schlechteren "Riecher" für wirksame Medikamente blanker Unsinn. Es ist schlicht und einfach Unsinn zu behaupten,

   "die Irrtumswahrscheinlichkeit hängt von einem völlig subjektiven und nicht berechenbaren Faktor ab, dem 'guten Händchen' oder dem 'Riecher' einer Forschungsgruppe" (Weihe),

   was uns Weihe und die von ihm als Kronzeugen bemühten Autoren Hans-Peter Beck-Bornholdt und Hans-Hermann Dubben[2] weismachen wollen.
3. Es ist Unsinn zumindest vom Standpunkt der "klassischen" Testtheorie von Neyman und Pearson, die mit recht die Statistik-Lehrbücher beherrscht, und die wir hier zu verteidigen versuchen. Es gibt zwar Beispiel, in denen die alternative Bayes'sche Betrachtungsweise, auf die sich Weihe sowie Beck-Bornholdt und Dubben berufen, Sinn macht, aber die hier zur Diskussion stehenden Überprüfungen der Wirksamkeit von Medikamenten gehören sicher nicht dazu. Wir versuchen dies in Abschn. 3 zu zeigen.

Die genannten und im folgenden kritisierten Arbeiten von Weihe, Beck-Bornholdt und Dubben sind vor allem deshalb ein Ärgernis, weil hier eine sehr spezielle Betrachtungsweise der Statistik, die sich nicht durchgesetzt hat, nämlich die Bayessche Testtheorie als allein gültig dargestellt wird und keine Gelegenheit versäumt wird, den Mainstream Statistiker als Dummkopf lächerlich zu machen. Noch dazu wird für

[1] Deutsches Ärzteblatt Jg. 101, Heft 13, 26. März 2004.
[2] Der Schein der Weisen, Irrtümer und Fehlurteile im täglichen Denken, Rowohlt Taschenbuch, Reinbek 2003.

einen ohnehin problematischen Punkt im Bayes'schen Ansatz, nämlich die Gewinnung von a priori Wahrscheinlichkeiten der höchst zweifelhafte Vorschlag gemacht, eine Statistik der Ergebnisse anderer Forschergruppen heranzuziehen.

Insbesondere zum ersten Punkt wurde im Text "Einmal zu viel um die Ecke gedacht" von M. Westermann und A. Kladroba bereits das Nötige gesagt. Es wurde dort deutlich gezeigt, wie die Hypothese der Wirksamkeit eines Medikaments bei Herzinfarkt Patienten statistisch getestet wird und wie die dabei unternommenen Verfahrensschritte zu interpretieren sind (und vor allem auch: wie nicht). Ich will hierauf nur kurz eingehen und mich dann darauf konzentrieren zu zeigen, wann eine Bayes'sche Verfahrenslogik (die allerdings in den genannten Texten eher mißverstanden ist) angebracht ist und wann nicht.

1. Interpretation des Zwei-Stichproben-Tests in der "klassischen" Testtheorie

Weihes Ausführungen machen nicht deutlich, wie der von ihm kritisierte Signifikanz-Test hinsichtlich Hypothesen, Prüfgröße und deren Stichprobenverteilung usw. aufgebaut ist. Es wird insbesondere nicht klar, dass ein Zwei-Stichproben-Test vorliegt, wie dies von Westermann und Kladroba richtiger weise festgestellt wurde. Bei der Fragestellung "wirkt Verum (V) oder besteht kein Unterschied zu einem Placebo (P)?" interessiert die Hypothese

H_0: $\pi_V = \pi_P$ oder gleichbedeutend $\pi_V - \pi_P = 0$ (daher auch der Ausdruck Nullhypothese), die gegen die Alternativhypothese

H_1: $\pi_V > \pi_P$ (was unsere eigentliche Vermutung ist)

getestet wird, wobei π jeweils die Wahrscheinlichkeit des "Erfolgs" (Überlebens[3]) bedeutet, die keineswegs 0,5 sein muß, sondern z.B. auch $\pi_V = \pi_P = 0{,}8$ sein kann, weil die Wahrscheinlichkeit, an der entsprechenden Krankheit zu sterben ohnehin nicht besonders groß ist (nur 0,2). Nur unter solchen Voraussetzungen macht es Sinn, zu fragen, wie wahrscheinlich es ist, dass alle n_V Personen (wie etwa im Beispiel[4] Weihes $n_V = 2$) der V-Gruppe überleben ($p_V = 1$) und dass keine der n_P Personen der P-Gruppe überlebt beiden. Bei hinreichend großem n_V und n_P ist die Prüfgröße

$$(1) \qquad t = \frac{p_V - p_P}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_V} + \frac{1}{n_P}\right)}} \text{ (mit } \bar{p} = \frac{n_V p_V + n_P p_P}{n_V + n_P}\text{) } t_{n-2} \text{ verteilt.}$$

Es ist wichtig sich klar zu machen, dass die folgenden Grundsätze (**G1** bis **G4**) für die Interpretation eines "klassischen" Tests gelten:

G1 (Stichprobenbeobachtung) Allein die Größen p_V und p_P schwanken zufällig, so dass man hier (und auch nur hier, also bei p_V und p_P) verschiedene Ergebnisse erzielen kann, je nachdem wie die Stichproben "ausfallen".

G2 (Parameter feste Größen) Die Größen π_V und π_P sind dagegen fest und völlig unveränderlich: es **ist** beispielsweise $\pi_V > \pi_P$ (und in diesem Sinne **ist** Verum

[3] natürlich stets bezogen auf einen bestimmten Zeitraum. Auch bei der Wahrscheinlichkeit zu sterben 1- π ("Fehlschlag" bei Weihe) geht es nicht darum, "irgendwann" an einer Krankheit zu sterben (was ja mit Sicherheit geschieht, wenn man den Zeitraum nur lange genug macht), sondern z.B. innerhalb des nächsten Jahres (so auch bei Weihe) zu sterben.

[4] Wir gehen absichtlich nicht von der unrealistischen Annahme aus, dass die Stichproben jeweils nur zwei Personen umfassen und präsentieren die allgemeine Vorgehensweise bei beliebigem (bzw. hinreichend großem) Umfang der V- und P-Gruppe.

"wirksam") und es kann dann nicht gleichzeitig $\pi_V = \pi_P$ sein (also V unwirksam sein), und es **ist** z.B. die Wahrscheinlichkeit $\pi_V = 0{,}7$ (als ein dem Medikament V inhärentes Maß der Wirksamkeit) und diese Wahrscheinlichkeit wird nicht dadurch kleiner oder größer weil irgend ein Forscher bei seiner Untersuchung $p_V = 0{,}6$ oder $p_V = 0{,}8$ findet.

G3 (Asymmetrie) Der Test dient der Widerlegung von H_0 und nur diese Hypothese H_0 ist genau spezifiziert, die Hypothese H_1 ist dagegen allgemein gehalten ($\pi_V > \pi_P$, nicht etwa $\pi_V - \pi_P = 0{,}2$ oder ähnlich): es geht um Annahme oder Ablehnung von H_0 weil das Stichprobenergebnis bei Geltung von H_0 zufällig erzeugt sein kann, oder aber so unwahrscheinlich (weniger wahrscheinlich als α)[5] ist, dass es durch Zufall nicht erklärt werden kann. Es geht nicht darum, welche Hypothese, H_0, H_1, H_2 usw. als die "glaubwürdigste" und in diesem Sinne die "wahrscheinlichste" gelten sollte, sondern nur darum, ein Urteil über H_0 abzugeben[6].

G4 (Tragweite der Entscheidung) Mit der Annahme (Ablehnung) von H_0 ist nicht gezeigt, dass H_0 richtig (bzw. falsch) ist, sondern nur, dass

a) es bei der konkret vorliegenden Stichprobe und **der** Grundgesamtheit aus der sie stammt gerechtfertigt erscheint H_0 anzunehmen[7] bzw. zu verwerfen, und zwar weil

b) das Stichprobenergebnis noch relativ wahrscheinlich (unwahrscheinlich) ist **wenn** H_0 richtig wäre (wie angenommen).

Aus der Sicht der "herrschenden Lehre" in der Statistik, d.h. der "klassischen Testtheorie" ist das alles Grundstudium im Fach Statistik, d.h. es ist ein Wissen, das eigentlich jeder Student in den ersten paar Semestern seines Studiums verstanden haben sollte.

Es ist danach vor allem **allein G2** mit dem "**objektiven**" (frequentistischen) **Wahrscheinlichkeitsbegriff** vereinbar. Es gilt entweder die Hypothese H_0 oder sie gilt nicht, es kann nicht beides gleichzeitig gelten oder mit einer bestimmten Wahrscheinlichkeit diese und einer bestimmten Wahrscheinlichkeit jene Hypothese gelten. Verum ist wirksam oder es ist nicht wirksam. Wirksamkeit ist eine dem Medikament inhärente Eigenschaft und nicht etwas, was mehr oder weniger zutrifft, je nachdem ob das wenige oder viele glauben, oder wie fest die Überzeugungen dieser Menschen sind. Objektive Wahrscheinlichkeit heißt, dass die beobachtbaren Konsequenzen unabhängig davon auftreten, was wir für möglich und wahrscheinlich halten, ganz genau so wie 1 + 1 = 2 ist und nicht 3 oder 5, und es dabei auch völlig egal ist, wer wie oft und mit welchem Ergebnis darüber nachgedacht, wie groß 1 + 1 ist.[8] Weil wir eine solche Situation für gegeben halten, also von **G2** ausgehen – aber auch wegen noch zu zeigender Widersprüche lehnen wir die alternative Betrachtungsweise von Weihe usw. ab.

---

[5] Die Orientierung erfolgt an der Wahrscheinlichkeit α. Die Wahrscheinlichkeit β kommt nicht ins Spiel und sie ist auch, solange nicht auch H_1 exakt spezifiziert wird, nicht berechenbar. Deshalb ist auch die Macht (power) $1-\beta$ nicht zu bestimmen.

[6] Die Asymmetrie zeigt sich daran, dass nur α nicht auch β numerisch vorgegeben (bekannt) sind und die Entscheidung lautet "H_0 annehmen" oder "H_0 ablehnen" (es gibt nie eine Entscheidung "H_1 annehmen" oder "H_1 ablehnen"

[7] "Annahme von H_0 heißt nicht, dass H_0 "wahr" ist, sondern nur "there is not enough evidence against H_0 in favor of H_1".

[8] Man kann keine subjektiven Wahrscheinlichkeiten bilden über derartige unbestreitbare Fakten, über etwas was zu allen Zeiten so gewesen ist und auch nie anders sein wird.

Für das Problem für das Weihe und wohl auch Beck-Bornholdt und Dubben ihren Bayes'schen Ansatz empfehlen wollen, halten wir diesen für ungeeignet. Er läuft darauf hinaus die obigen Grundsätze in Frage zu stellen. Bei der Überprüfung der Wirksamkeit eines Medikament liegt die Situation von **G3** vor, es ist nicht notwendig vor[9] und nach einem Stichprobenbefund Wahrscheinlichkeitsaussagen zu machen, sondern es ist (isoliert) ein bestimmtes Medikament zu beurteilen und im Falle eines Medikament dürfte **G2** zutreffend sein, nicht ein subjektiver (personalistischer) Wahrscheinlichkeitsbegriff (davon später mehr).

2. Wenn mehr als einmal ein statistischer Test durchgeführt wird: der magische Auftritt einer vom Signifikanzniveau verschiedenen "Irrtumswahrscheinlichkeit"

Es sind hier deutlich zwei Fälle zu unterscheiden, wobei es in einem Fall Sinn macht, im anderen dagegen überhaupt keinen Sinn macht, empirische Feststellungen zu problematisieren und den Begriff "Signifikanzniveau" zu relativieren:

a) gleiche Untersuchungen (gleiche Methodik, nicht notwendig gleiche Stichprobenumfänge) über **ein und das gleiche Medikament**, etwa Verum, nur dieses und kein anderes Medikament (in diesem Fall wird mit recht über die Unsitte geklagt, nur signifikante Ergebnisse zu publizieren und die nicht-signifikanten einfach unter den Tisch fallen zu lassen), und der Fall in dem

b) die Ergebnisse verschiedener Forschergruppen, die sich auf Untersuchungen über offenbar sehr unterschiedliche Medikamente beziehen, verglichen werden[10], und angeblich mit ins Kalkül zu ziehen sind, wenn es um die Irrtumswahrscheinlichkeit geht.

Die Argumentation von Weihe mag manchen Leser überzeugen, weil vielleicht nicht sofort erkennbar ist, dass bei ihr diese zwei sehr unterschiedlichen Fälle in einen Topf geworfen wurden.

a) Mehrere Untersuchung zu dem gleichen Medikament: was bedeutet die Unterschlagung nichtsignifikanter Ergebnisse?

Aus der Betrachtung der Prüfgröße t gem. Gl. 1 ergibt sich, dass beispielsweise bei den Anteilen $p_V = 0{,}5$ und $p_P = 0{,}4$ (mit $n_V = n_P = 50$) ein nicht-signifikanter Unterschied besteht, weil dann $t = 1{,}005$ ist während die berühmte 5 % Signifikanzschranke bei einem einseitigen Test und $n = 100$ den Wert 1,6449 annimmt[11]. Angenommen wir haben drei Forscher A1, A2 und B, die jeweils mit $n_V = n_P = 50$ Stichproben arbeiten und die Ergebnisse der Tab. 1 erzielt haben.

B hat als einziger ein signifikantes Ergebnis erzielt ($t = 2{,}0412 > 1{,}6449$). Unterschlägt man nun einfach die Ergebnisse von A1 und A2 so hat man eigentlich nicht das Ergebnis von B als gültigen Befund über V, sondern faktisch eine Stichprobe im Umfang von $n_V = 150$ Personen der V-Gruppe und $n_P = 150$ Personen der P-Gruppe gehabt. Man müßte also praktisch die drei Stichproben zu einer zusammenlegen und

---

[9] Es dürfte schwer sein, anders als pharmakologisch begründete subjektive Wahrscheinlichkeiten über neue Medikamente (im Vergleich zu unendlich vielen potenziellen anderen Medikamenten) zu bilden.

[10] Die von ihnen jeweils getroffene **Auswahl** ist **unterschiedlich**, sonst macht es ja auch keinen Sinn von der Forschergruppe mit und der ohne "Riecher" (für wirksame Medikamente) zu sprechen.

[11] bzw. etwa 1,98 wenn man die t-Verteilung betrachtet (statt wie bei hinreichend großem n die Normalverteilung).

hätte dann wie Tab. 1b zeigt ein nicht signifikantes Ergebnis gehabt. Das Ergebnis von B allein zeichnet also eigentlich kein zutreffendes Bild und die Unterschlagung der "negativen" Ergebnisse von A1 und A2 trägt zu dieser Irreführung bei.

Tabelle 1: a) Drei Forscher und zwei unterschlagene Ergebnisse (A1, A2)

	p_V	p_P	Prüfgröße
A1	0,5	0,4	$\frac{0,1}{\sqrt{0,45 \cdot 0,55 \cdot \frac{2}{50}}} = 1,0050$
A2	0,5	0,48	$\frac{0,02}{\sqrt{0,49 \cdot 0,51 \cdot \frac{2}{50}}} = 0,2004$
B	0,5	0,3	$\frac{0,2}{\sqrt{0,4 \cdot 0,6 \cdot \frac{2}{50}}} = 2,0412$

b) Alle drei Studien zusammengenommen

$p_V = 75/150 = 0,5$	$p_P = 59/150 = 0,3933$	$\frac{0,1067}{\sqrt{0,4467 \cdot 0,5533 \cdot \frac{2}{150}}} = 1,8581$

Die Sache mit der Unterschlagung ist allerdings nicht ganz so einfach, weil

- sich das Bild bei einem Ergebnis mit einer nicht ganz so geringen Prüfgröße (wie die von A2), das unterschlagen wurde schon anders darstellen kann, und
- weil mit wachsendem Stichprobenumfang wegen der Größen $1/n_V$ und $1/n_P$ in der Gl. 1 die Prüfgröße verändert und mit Zunahme von $n = n_V + n_P$ ein immer geringerer Unterschied zwischen p_V und p_P bereits "signifikant" wird.

Was den ersten Punkt betrifft, so denke man an den Fall der drei Forscher A1 und B (wie bisher) und A3 mit den Daten der Tab. 2.

Im Fall der Tab. 2 ist es nicht so klar, dass die alleinige Betrachtung von B ein falsches Bild erzeugt. Nimmt man auch die unterschlagenen Ergebnisse der Forscher A1 und A3 hinzu, so ändert sich an dem Bild (signifikanter Unterschied) wenig.

Tab. 3 zeigt schließlich, dass mit zunehmendem Wert von n ein immer kleinerer Unterschied zwischen p_P und $p_V = 0,5$ bereits signifikant auf dem 5% Niveau (mit der Schranke 1,6449) wird.

b) Untersuchung von verschiedenen Forschergruppen, (bei Geltung verschiedener Hypothesen); Berechnung der sog. "Irrtumswahrscheinlichkeit"

Nicht nur objektiv, auch für die subjektive Wahrscheinlichkeit leuchtet ein: Man fühlt sich sicherer in seinem Urteil, wenn man mehr Beobachtungen der **gleichen** Art (bezogen auf das gleiche Medikament) hat, also der Stichprobenumfang größer ist.

Aber was soll man demgegenüber daraus ableiten, dass verschiedene Forschergruppen offenbar **verschiedene** Medikamente untersuchen, etwa die weniger fähige

Gruppe A die Medikamente a, x, y und z und die fähigere Gruppe B Medikamente a, b, c und x, wobei a, b und c "wirksam" und x, y und z "unwirksam" sein sollen? Was kann man hieraus insbesondere herleiten über die "Irrtumswahrscheinlichkeit" bei der empirischen Untersuchung der Wirksamkeit von a? Aus Sicht der "klassischen" (Neyman-Pearsonschen) Testtheorie kann man **daraus nichts herleiten**, und das auch – wie wir meinen - aus sehr guten Gründen und es gibt überhaupt keinen Anlass, gegen diese klassische Testtheorie zu polemisieren.

Tabelle 2: a) Drei Forscher und zwei unterschlagene Ergebnisse (A1 und B wie bisher und A3)

	p_V	p_P	Prüfgröße
A3	0,5	0,38	$\frac{0{,}12}{\sqrt{0{,}44 \cdot 0{,}56 \cdot \frac{2}{50}}} = 1{,}2087$

b) Alle drei Studien zusammengenommen

A1,A3,B	$p_V = 0{,}5$	$p_P = 57/150 = 0{,}3933$	$\frac{0{,}12}{\sqrt{0{,}44 \cdot 0{,}56 \cdot \frac{2}{150}}} = 2{,}0936$

Tabelle 3: Werte von p_P bei denen der Unterschied zu $p_V = 0{,}5$ signifikant ist ($\alpha = 5\%$)

$n_V = n_P =$	$p_P =$
50	0,37279
100	0,40968
150	0,42616
200	0,43601
250	0,44274

$n_V = n_P =$	$p_P =$
300	0,44771
350	0,45158
400	0,45470
450	0,45729
500	0,45948

Wie sieht demgegenüber die Testtheorie von Beck-Bornholdt und Dubben und den sich diesen Autoren gegenüber offenbar als Sprachrohr betätigenden Weihe aus? Wie wird dort insbesondere die magische Irrtumswahrscheinlichkeit berechnet? Betrachtet man das Beispiel von Weihe (Tabelle 4) und ähnliche Überlegungen in dem Buch von Beck-Bornholdt und Dubben, so werden Berechnungen einer angeblichen Irrtumswahrscheinlichkeit (im Unterschied zum Signifikanzniveau) nach Art von Tabelle 4 (die sich so bei Weihe findet) vorgeführt[12]:

Die Irrtumswahrscheinlichkeit wird dann berechnet zu (1 – 8/12,5)100 = 100 – 64 = 36%. Bei einer anderen Forschergruppe (B) mit dem "besseren Riecher" gelten die Prozentsätze $h_1 = 40\%$ und $h_0 = 60\%$ und man erhält dann bei gleichem Signifikanzniveau von $\alpha = 5\%$ (= 0,05) eine wesentlich geringere Irrtumswahrscheinlichkeit von nur 9% (statt 36%). Vom Standpunkt der "klassischen" Testtheorie ist das alles sehr abenteuerlich:

[12] Das Muster beruht – wie gleich gezeigt wird – offenbar auf der Berechnung von a posteriori Wahrscheinlichkeiten nach Bayes.

- Gleiches Signifikanzniveau und trotzdem unterschiedliche Irrtumswahrscheinlichkeit je nachdem ob der Forscher einen guten oder keinen guten Riecher hat: Wie soll das möglich sein? Wird ein Irrtum eines Forscher wahrscheinlicher wenn es auch einen anderen Forscher gibt?

Tab. 4: Berechnung der Irrtumswahrscheinlichkeit von Forschergruppe A

	Anteil der Studien	Anteil positiver Studienergebnisse	Anteil negativer Studienergebnisse
Das neue Medikament ist tatsächlich wirksam (gute Idee gehabt)	$h_1 = 10\%$	10 mal 0,8 = 8% wegen Macht (1-β) von 80%	10 mal 0,2 = 2%
Das neue Medikament ist tatsächl. nicht wirksam (schlechte Idee)	$h_0 = 90\%$	90 mal 0,05 = 4,5% wegen Signifikanzniveau $\alpha = 5\%$	90 mal 0,95 = 85,5%
Summe	100%	12,5%	87,5%

- Es kommt anscheinend nicht auf die absolute Anzahl der Untersuchungen eines Medikaments an, sondern nur auf die Prozentsätze 10 und 90% oder 40 und 60%, ob also hinter den 10% Untersuchungen 3 oder 80 Untersuchungen stehen ist nicht relevant (obgleich es ja doch wohl ein Unterschied ausmachen dürfte, ob etwas 3 mal oder 80 mal untersucht worden ist).
- Es kommt auch nicht auf die Stichprobenumfänge bei den Untersuchungen an, ob also bei den (allen?) Untersuchungen mit der guten (oder der schlechten) Idee 50 oder 500 Personen in der Experiment- und in der Kontrollgruppe betrachtet worden sind.
- Die Prozentsätze 10% und 90% in Tab. 4 dienen als Schätzung der a priori Wahrscheinlichkeiten der Hypothesen H_0 und H_1 (eine merkwürdige Vorstellung aus Sicht des Grundsatzes G2): dabei spielt es offenbar keine Rolle, wie viele Beobachtungen (Versuchspersonen) der Einschätzung zugrunde liegen (!!).
- Nach "klassischer" Vorstellung ist eine Hypothese entweder zutreffend oder nicht zutreffend. Der Umstand dass man nicht **weiß** was zutrifft, H_0 oder H_1 heißt nicht, dass die Realität nicht so **oder** so **ist** (dass nämlich entweder H_0 oder H_1 gilt, nicht beides). Nach Grundsatz **G2** kann es keine Wahrscheinlichkeiten für H_0 und H_1 geben (anders wenn das Bayessche Theorem angewendet wird auf eine Zufallsvariable, wie unter Abschn. 3a).
- Selbst wenn H_0 und H_1 nicht Aussagen über feste Parameter wären, sondern Zufallsvariablen oder Ereignisse (wie z.B. dass es morgen regnen wird, R) wären, dann wäre es immer noch nicht klar, ob und wann es legitim ist, Feststellungen über Häufigkeiten als Basis für die Einschätzung einer subjektiven Wahrscheinlichkeit heranzuziehen. Ist es legitim, aus den Erfahrungen zweier Forschergruppen, die diese offenbar mit *verschiedenen* Medikamenten gewonnen haben auf subjektive Wahrscheinlichkeiten zu schließen?[13]
- Was bedeutet es, bei einem Medikament die subjektive Wahrscheinlichkeit für dessen Wirksamkeit (um *wieviel* ist dabei π_V größer als π_P?) oder Unwirksam-

[13] Tut man dies, wie weihe und Co. dann ist es noch nicht einmal klar, *wessen* subjektive Vorstellungen über das Zutreffen von H_0 und H_1 damit erfaßt werden.

keit zu entwickeln? Kann man Vorstellungen über eine solche Wahrscheinlichkeit gewinnen? Wenn ja, wie?[14]

- Wer als erster (und damit zunächst als einziger) ein wirksames neues Medikament mit den Ergebnis testet dass es wirksam ist, hat nach Weihe eine Irrtumswahrscheinlichkeit von nur 0%; kommen danach andere Forscher, die ebenfalls das Medikament untersuchen, dann kann es nur schlechter werden mit der Irrtumswahrscheinlichkeit[15]. Wird nur eine Untersuchung angestellt, bei der sich dann die Wirksamkeit des Medikaments herausstellt, dann gilt

Tab. 5: Berechnung der Irrtumswahrscheinlichkeit bei einmaliger Untersuchung eines Medikaments

	Anteil der Studien	Anteil positiver Ergebnisse	Anteil negativer Ergebnisse
Medikament wirksam	h_1 = 100%	100 mal 0,8 = 80%	20%
Medikament nicht wirks.	h_0 = 0%	0 mal 0,05 = 0%	0 mal 0,95 = 0%
Summe	100%	80%	20%

was dann nach der oben angegebenen Berechnungsweise (1 – 80/80)100 = 0% als sog. Irrtumswahrscheinlichkeit. Es kann durch weitere Untersuchungen nur schlimmer (größer) werden mit der Irrtumswahrscheinlichkeit[16]. Spätestens jetzt kommt der Verdacht auf, dass etwas nicht ganz stimmen kann bei dem Rechenschema für die sog. Irrtumswahrscheinlichkeit. Dieses Schema sieht wie folgt aus (Tab. 6):

Tab. 6: Interpretation des Rechenschema für die Irrtumswahrscheinlichkeit von Beck-Bornholdt und Dubben nach Art des Bayes'schen Theorems

Zustand	Anteil der Studien	E_1 = positive Studienergebnisse = Entscheidung für H_1	E_0 = negative Studienergebnisse = Entscheidung für H_0
H_1 gilt	a priori Wahrscheinlichkeit $P(H_1)$, $h_1/100$	1- β (H_0 verworfen, richtiges H_1 erkannt) $P(E_1 \mid H_1)$	β (H_0 angenommen, obgleich H_0 falsch ist): $P(E_0 \mid H_1)$
H_0 gilt	a priori Wahrscheinlichkeit $P(H_0)$, $h_0/100$	α (H_0 verworfen obgleich H_0 richtig) $P(E_1 \mid H_0)$	1- α richtige Entscheidung für H_0 $P(E_0 \mid H_0)$

Als "Irrtumswahrscheinlichkeit" (sie sei $\phi = P(H_0 \mid E_1)$ genannt) wird nun berechnet

$$(2) \qquad \phi = 1 - \frac{P(H_1)\cdot(1-\beta)}{P(H_0)\cdot\alpha + P(H_1)\cdot(1-\beta)} = \frac{P(H_0)\cdot\alpha}{P(H_0)\cdot\alpha + P(H_1)\cdot(1-\beta)}$$

[14] Man könnte z.B. daran denken, dass man Wahrscheinlichkeiten aufgrund der chemischen Zusammensetzung eines Medikaments bilden könnte, aber sicherlich nicht aufgrund der Herangehensweise sehr verschiedener Forschergruppen

[15] Allein diese Konsequenz sollte den Verdacht aufkommen lassen, dass mit den Konstruktionen von Bayesianern wie Weihe usw. etwas nicht ganz stimmen kann.

[16] Umgekehrt ist es bei *einer* Untersuchung eines unwirksamen Mittels. Die Irrtumswahrscheinlichkeit ist 100% und sie kann durch weitere Untersuchungen nur besser (geringer) werden. Es gibt offenbar eine geheimnisvolle Belohnung (Bestrafung) der guten (schlechten) Forscher: je besser (schlechter) der "Riecher" desto geringer (größer) die Irrtumswahrscheinlichkeit.

oder wegen $P(H_1) = 1- P(H_0)$

(2a) $$\phi = \frac{\alpha}{\alpha + \beta - 1 + \frac{1-\beta}{P(H_0)}}$$

und in Symbolen, die eher zum Bayes'schen Theorem (Bayes' rule) passen

(2b) $$\phi = P(H_0|E_1) = \frac{P(E_1|H_0)\cdot P(H_0)}{P(E_1|H_0)\cdot P(H_0) + P(E_1|H_1)\cdot P(H_1)} =$$

$$\frac{P(E_1 \cap H_0)}{P(E_1 \cap H_0) + P(E_1 \cap H_1)} = \frac{P(E_1 \cap H_0)}{P(E_1)}.$$

Wie man sieht hängt die sog. "Irrtumswahrscheinlichkeit" ϕ ab von den a priori Wahrscheinlichkeiten der Hypothesen, von dem Signifikanzniveau α und der Macht $1- \beta$ des statistischen Tests, wobei letztere auch von dem konkreten Wert für die Differenz $\pi_V - \pi_P$ abhängt, ganz abgesehen vom Stichprobenumfang, der ansonsten gar nicht explizit nicht in den Gleichungen 2 bis 2b vorkommt.

3. Die richtige Anwendung der Bayesschen Schätz- und Testtheorie und Argumente für und gegen diese Theorie im Vergleich zur "klassischen" Schätz- und Testtheorie

Es ist nachgerade klar, dass sich Beck-Bornholdt und Dubben - und in ihrem Schlepptau auch Weihe - auf die Bayesschen Schätz- und Testtheorie berufen. Die bereits erfolgten kritischen Hinweise sollten unseren Standpunkt untermauern, dass sie dies zu unrecht tun und offenbar diese Testtheorie nicht richtig verstanden haben. Wir wollen deshalb abschließend ein paar Hinweise dazu geben, wie demgegenüber die Fälle gelagert sind, in denen man sich durchaus zu recht der Bayesschen Theorie bedienen mag.

a) Statt H_0 und H_1 ein (zufälliges) Ereignis R bzw. Nicht-R

Die zweifellos legitime Anwendung des Bayesschen Theorems ist der Fall eines Ereignisses, über dessen Auftreten ein ebenfalls zufällig bestimmter Indikator (Anzeige E_0 und E_1 als Ereignisse) Auskunft geben mag. Beispiel: gefragt ist nach der Wahrscheinlichkeit, dass es *morgen* regnet R oder nicht regnet, was als Zufallsereignis aufgefaßt werden kann (anders als die Frage der Wirksamkeit eines Medikaments oder die Frage ob es *gestern* geregnet hat), so daß man von einer Wahrscheinlichkeit für dieses Ereignis sprechen kann.

Anders als bei H_0 und H_1 kann man bei den Ereignissen R und nicht-R (Symbol $\overline{R}$) durchaus von zufälligen Erscheinungen sprechen (d.h. es gilt *nicht* Grundsatz **G2**) und man kann sie auch symmetrisch behandeln (es gilt *nicht* **G3**)

(3) $$P(R|E_1) = \frac{P(E_1|R)\cdot P(R)}{P(E_1|R)\cdot P(R) + P(E_1|\overline{R})\cdot P(\overline{R})}$$

Mit $P(E_1|R)$ oder $P(E_1|\overline{R})$ wird ein zufälliges Ereignis E_1 (etwa die Feststellung, dass das Barometer steigt statt sinkt, was dann E_0 wäre) in Verbindung mit Regen oder Nicht-Regen gebracht. Es muß also ein Ereignis E_1 (oder B, also Barometer steigt) und E_0 (oder $\overline{B}$) geben und es darf nicht unabhängig sein von R, also nicht

$P(B|R) = P(B|\overline{R}) = P(B)$ gelten. Nur dann kann man "aus Erfahrung lernen". Wir haben in den kritisierten Schriften die folgende (nach unserer Meinung) an den Haaren herbeigezogene Analogie zu der sinnvollen Betrachtung des Beispiels mit dem Regen und dem Barometer:

Tab. 7: Bayes' Theorem und dessen Interpretation im medizinischen Beispiel

Theorem von Bayes	im Fall des Hypothesentests
a priori Wahrscheinlichkeiten $P(R), P(\overline{R})$	$P(H_0), P(H_1)$
Likelihoods $P(\overline{B}\|R) = \alpha$, $P(B\|R) = 1-\alpha$ $P(\overline{B}\|\overline{R}) = 1-\beta$, $P(B\|\overline{R}) = \beta$	$P(E_1\|H_0) = \alpha$, $P(E_0\|H_0) = 1-\alpha$ $P(E_1\|H_1) = 1-\beta$, $P(E_0\|H_1) = \beta$

Ob man nun die analog zu $P(H_0|E_1)$ gebildete Größe $P(R|\overline{B}) = \frac{P(\overline{B}|R)P(R)}{P(\overline{B})}$ als Irrtumswahrscheinlichkeit bezeichnet (worin besteht der Irrtum?) oder $P(\overline{R}|B)$ in Analogie zu $P(H_1|E_0)$ ist wohl eher eine Geschmackssache.

Wichtig ist, dass ein beobachtbares Ereignis mit R in einem empirischen Zusammenhang steht und nicht etwa durch Konvention festgelegt wird. Es genügt nicht, mit R eine Zufallsvariable zu haben (was im Falle von H_0 ja wohl nicht der Fall ist), es muß auch Likelihoods geben. Man kann z.B. sinnvoll subjektive a priori Wahrscheinlichkeiten darüber bilden ob die unsichtbare nicht unbeträchtliche "schwarze" Materie des Weltalls aus WIMPs (weakly interactive massive particles) oder aus MACHOs (massive compact halo objects) besteht also P(W) und P(M) beispielsweise mit 0,3 un 0,7 beziffern, aber es gibt keine Beobachtung (kein Stichprobenergebnis), das als Indikator für das Auftreten von W oder M dienen kann, so dass es auch nichts aus der Erfahrung zu lernen gibt.

Man beachte, dass wir auch bisher noch nicht von einer Stichprobe und einem Stichprobenumfang n gesprochen haben, nur von der Wahrscheinlichkeit des Auftretens eines Ereignisses wie B (oder in "Analogie" H_0) und es ist auch merkwürdig, dass bei der Berechnung der "Irrtumswahrscheinlichkeit" bei Weihe & Co dieses n auch nicht irgendwo auftritt.

b) Variablen statt Ereignisse, Schätztheorie nach Bayes

Man kann die Bayessche Analyse auch anwenden auf eine zu schätzende oder zu testende numerische (metrisch skalierte) Variable (etwa μ oder θ) statt auf Ereignisse.

Ein Lehrbuch-Beispiel ist die Schätzung eines Konfidenzintervalls für $\mu = E(X)$ wenn X normalverteilt ist und neben einer Stichprobe vom Umfang n mit der Schätzfunktion $\overline{X}$ auch eine a priori Vermutung über $\mu = \mu_0$ existiert. Hat μ eine a priori Verteilung (prior distribution) mit $N(\mu_0, \sigma_0)$, also eine Normalverteilung und ist $\overline{X}$ nach dem zentralen Grenzwertsatz ebenfalls normalverteilt mit $N(\mu, \frac{\sigma}{\sqrt{n}})$, dann gilt für die a posteriori Verteilung von μ

(4) $$N\left(\frac{w_1\bar{x}+w_2\mu_0}{w_1+w_2},\sqrt{\frac{1}{w_1+w_2}}\right) \text{ mit } w_1=\frac{1}{\sigma^2/n} \text{ und } w_2=\frac{1}{\sigma_0^2}.$$

Wie man sofort sieht, gilt bei $w_2 = 0$ eine Normalverteilung mit Erwartungswert $\bar{X}$ und einer Varianz von σ^2/n und das Konfidenzintervall bei einer Irrtumswahrscheinlichkeit (im üblichen Sinne) von α ist gegeben (als Spezialfall) mit

(5) $$P\left(\bar{x}-z_\alpha\frac{\sigma}{\sqrt{n}}<\mu\le\bar{x}+z_\alpha\frac{\sigma}{\sqrt{n}}\right)=1-\alpha.$$

Das folgende Zahlenbeispiel mag die Zusammenhänge veranschaulichen. Gegeben sei eine Stichprobe mit n = 100 und $\bar{x}$ = 300, ferner sei σ = 50. Man erhält dann nach Gl. 5 für das 95% zweiseitige Konfidenzintervall ($z\alpha$ = 1,96) in der üblichen ("klassischen") Weise also ohne Berücksichtigung von a priori "Wissen" die Grenzen:

$\bar{x}\pm z_\alpha\frac{\sigma}{\sqrt{n}}=300\pm1{,}96\cdot\frac{50}{10}$, es liegt also zwischen 290,2 und 309,8.

Verbindet man dies jedoch mit der a priori Verteilung mit μ_0 = 290 und σ_0 = 40, so erhält man wegen $w_1=\frac{1}{\sigma^2/n}=\frac{100}{50^2}=\frac{100}{2500}=\frac{1}{25}$ und $w_2=\frac{1}{40^2}=\frac{1}{1600}$ für Gl. 4

$$N\left(\frac{300\cdot\frac{1}{25}+290\cdot\frac{1}{1600}}{1/25+1/1600}=299{,}846,\sqrt{\frac{1}{1/25+1/1600}}=\sqrt{24{,}615}=4{,}9614\right)$$ und damit für die

Grenzen des Konfidenzintervalls sind 299,8 ± 1,96 4,96 = 299,8 ± 9,724 und damit 290,12 und 309,57 statt 300 ± 10. Die Berücksichtigung von a priori Wissen (oder besser: subjektiver Annahmen) hat also eine geringere Breite des Konfidenzintervalls (19,45 statt 20) bewirkt. Bei großen Stichprobenumfängen ist der Unterschied noch geringer.

c) Testtheorie nach Bayes und Vergleich mit dem klassischem "significance approach"

Die testtheoretischen Ansätze nach Bayes sollen nicht so ausführlich dargestellt werden wie die Schätztheorie. Eine zentrale Rolle spielt dabei

1. die Verlustfunktion (loss function) Λ_{ah} also der Verlust der bei Entscheidung (action) a eintritt wenn die Hypothese h gilt, bzw. die regret function r_h (Λ abzügl. des bei dieser Hypothese maximalen Verlusts),
2. die Likelihoodfunktion L(**x**, h) – die Wahrscheinlichkeit des Stichprobenbefunds (Vektor **x**) und der daraus abgeleiteten Aktion a (Entscheidung) in Abhängigkeit von der Hypothese h und
3. die Wahrscheinlichkeit für die Hypothese h (etwa das θ_1 oder θ_0 gilt), also $p(\theta_0)$ und $p(\theta_1)$

Man entscheidet sich für θ_0 und gegen θ_1 wenn gilt

(6) $$\frac{P(x|\theta_1)}{P(x|\theta_1)}<\frac{r_0P(\theta_0)}{r_1P(\theta_1)},$$

so dass die "likelihood ratio" $\frac{L(\theta_1)}{L(\theta_0)} = \frac{P(x|\theta_1)}{P(x|\theta_1)}$ sehr entscheidend ist. Man spricht deshalb auch vom likelihood ratio criterion. Unter den Verlustfunktionen spielt die quadratische Verlustfunktion eine wichtige Rolle. Sie ist im Falle der Schätzung von beispielsweise θ durch $\bar{x}$ proportional zu $(\bar{x} - \theta)^2$. Die Bevorzugung des arithmetischen Mittels wird verständlich, wenn man davon ausgeht, dass die Verlustfunktion üblicherweise minimiert wird.

Abschließend noch eine tabellarische Übersicht über die in der Literatur zu findenden Argument für und gegen desn Bayes – und den significance approach

Tab. 8: Pro und contra der beiden Ansätze

klassischer "significance approach "	Bayes approach
Vorteile	
Gut wenn nur eine Hypothese zur Diskussion steht und es schwierig wäre für **alle** Alternativen Wahrscheinlichkeiten zu finden (difficult to formulate alternatives and to assign probabilities to them)	Man kann auch über die Stichprobe hinausgehende Informationen verarbeiten; bei kleinen Stichproben evtl. Konfidenzintervall schmaler + (inferences to other populations not sampled)
Nachteile	
Es gibt keine Möglichkeit a priori Wissen "einzubauen" und man kann Hypothesen nicht nach dem Grad ihrer Glaubwürdigkeit ordnen (no formal mechanism for utilizing information on possible alternative hypotheses)	Subjektive Einschätzungen haben Einfluß; Schwierigkeiten a priori Wahrsch. zu definieren (problem to quantify prior beliefs + requirement that an exhaustive set of alternatives must be defined). Auch loss function unbekannt.