# Bayes-Verfahren zur Lösung von Testproblemen am Beispiel des Homogenitätstests zum Vergleich zweier Erfolgsraten

**Joachim Gerß[1], Martin Kappler[2]**

[1]Institute of Biostatistics and Clinical Research, University of Münster, Germany
[2]Novartis Pharma S.A.S., France

joachim.gerss@ukmuenster.de

## Einleitung

In der frequentistischen Statistik spielen Testprobleme und deren Lösung mit Hilfe von Signifikanztests eine zentrale Rolle. Die Bayes-Statistik wird dagegen üblicherweise weniger zur Lösung von Testproblemen eingesetzt, sondern ihre vorrangigen Anwendungsmöglichkeiten eher in der Schätztheorie gesehen. Im Rahmen der Bayes-Statistik lassen sich allerdings auch direkte Gegenstücke klassischer Signifikanztests entwickeln.

Auf dem vorliegenden Poster werden am Beispiel des Homogenitätstests zum Vergleich zweier Erfolgsraten verschiedene Bayes-Verfahren zur Lösung des Testproblems vorgestellt. Die vorgestellten Verfahren werden ggf. vereinheitlicht bzw. optimiert, so dass eine vergleichende Bewertung möglich ist. Dies geschieht mit Hilfe simulierter Daten nach üblichen Kriterien der frequentistischen Statistik.

## Methoden

Gegeben sei die Situation einer zweiarmigen Studie mit binärem Endpunkt. Um die Erfolgsraten zweier Therapien A und B miteinander zu vergleichen, wird ein einseitiges Testproblem aufgestellt, d.h. die Hypothesen

$$H_0: p_A=p_B \text{ und } H_1: p_A>p_B .$$

### Bayes‘sche Schätzung

Im einfachsten Bayes-Ansatz berechnet man auf der Grundlage einer nicht-informativen a-priori-Verteilung auf den Erfolgsraten $p_A$ und $p_B$ die a-posteriori-Wahrscheinlichkeit $Pr(p_A>p_B)$ bei gegebenen Daten [1]. Ist diese Wahrscheinlichkeit größer als eine festgelegte Grenze, so wird das Testergebnis als signifikant und die Überlegenheit der Therapie A als erwiesen angesehen. Dieser Ansatz kann allerdings nicht als echtes Gegenstück klassischer Signifikanztests betrachtet werden, sondern stellt eher einen Versuch dar, das Testproblem mit Methoden der Schätztheorie zu lösen.

### Bayes‘scher p-Wert

In einem zweiten Ansatz wird ein wirklicher „Bayes‘scher p-Wert“ hergeleitet, d.h. die a-posteriori-Wahrscheinlichkeit für eine Realisation der Teststatistik, die mindestens so stark im Widerspruch zur Nullhypothese steht wie die tatsächlich beobachtete [2].

| | Erfolgsrate | Anzahl Erfolge | Anzahl Misserfolge | Gesamt |
|---|---|---|---|---|
| Therapie A | $p_A$ | $y_{A1}$ | $y_{A0}$ | $n_A$ |
| Therapie B | $p_B$ | $y_{B1}$ | $y_{B0}$ | $n_B$ |
| Gesamt | p | $y_1$ | | |

Für alle folgenden Überlegungen wird vorausgesetzt, dass die Nullhypothese des aufgestellten Testproblems $H_0$: $p_A=p_B=p$ gilt. Die Gesamt-Erfolgsrate p stellt dabei einen *Nuisance Parameter* dar, der nicht von Interesse ist und in irgendeiner Weise eliminiert werden muss, um den p-Wert zu bestimmen. Im klassischen Test nach Fisher geschieht dies, indem die bedingte Überschreitungswahrscheinlichkeit bei gegebener Spaltensumme $y_1$ bestimmt wird, d.h. $Pr(Y_{A1}\geq y_{A1}|y_1)$. Im Rahmen eines Bayes-Ansatzes bietet es sich stattdessen an, den *Nuisance Parameter* durch Integration zu eliminieren.

$$m(y_{A1} \mid y_{B1}) = \int f(y_{A1}, p \mid y_{B1})\ dp = \int f(y_{A1} \mid y_{B1}, p)\cdot\pi(p \mid y_{B1})\ dp$$

$$p-Wert \ = \ \Pr^{m(y_{A1}|y_{B1})}(Y_{A1} \geq y_{A1})$$

Man bestimmt die Randverteilung der Daten $Y_{A1}$, indem aus der gemeinsamen Verteilung mit dem Parameter p der Parameter p herausintegriert wird. Die Integration geschieht bzgl. der *partiellen a-posteriori-Verteilung* $\pi(p|y_{B1})$. In der *partiellen a-posteriori-Verteilung* werden nicht die vollständigen Daten $(y_{A1},y_{B1})$ genutzt, sondern nur der Teil $y_{B1}$. Die partielle a-posteriori-Verteilung wird anstatt der vollständigen a-posteriori-Verteilung $\pi(p|y_{A1},y_{B1})$ verwendet, um eine doppelte Nutzung der Daten zu vermeiden, d.h. einerseits bei der Bildung der Randverteilung m und andererseits bei der Berechnung der Überschreitungswahrscheinlichkeit $Pr(Y_{A1}\geq y_{A1})$.

### Bayes‘scher Hypothesentest

Ein dritter Ansatz zur Lösung des Testproblems ergibt sich aus einer Weiterentwicklung eines Verfahrens für einarmige Studien, in denen die Erfolgsrate einer Therapie mit einem festen Wert verglichen wird [3]. Man stellt ein dreistufiges Modell auf, in dem die Wahrscheinlichkeit für die Gültigkeit der Nullhypothese $H_0$:RD:=$p_A$-$p_B$≤0 und der Alternative $H_1$:RD>0 ausdrücklich modelliert wird.

Datenmodell: Verteilung der Daten $y_{A1}$ bei gegebener Risikodifferenz RD (und Randhäufigkeit $y_1$)

A-priori-Verteilung: $\pi(RD|H_0) \sim I_{[-1,0]}$ , $\pi(RD|H_1) \sim I_{(0,1]}$

Hyper-prior: $Pr(H_0) = Pr(H_1) = 0.5$

Auf der Basis des erstellten Modells wird die a-posteriori-Wahrscheinlichkeit für die Gültigkeit der Nullhypothese und der Alternative berechnet. Die Nullhypothese wird abgelehnt, falls der Quotient beider Wahrscheinlichkeiten $Pr_{post}(H_1)$ / $Pr_{post}(H_0)$ eine festgelegte Grenze überschreitet.

## Ergebnisse

Die vorgestellten Verfahren werden so vereinheitlicht, dass in jedem Fall ein (einseitiger) Fehler 1. Art von 0.05 gilt. Dazu wird das Testergebnis jeweils als signifikant angesehen, wenn

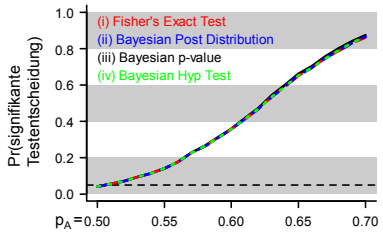
(i) im exakten Test nach Fisher gilt: $p_{Fisher} \leq 0.05$

(ii) im Rahmen der Bayes‘schen Schätzung für die a-posteriori-Wahrscheinlichkeit gilt $Pr(p_A>p_B) \geq 0.96$

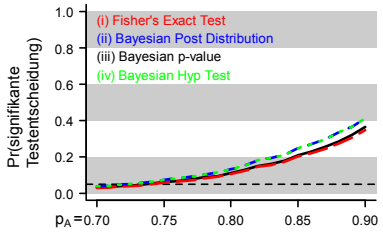
(iii) für den Bayes‘schen p-Wert gilt: $p_{PPOST} \leq 0.05$

(iv) im Bayes‘schen Hypothesentest gilt $Pr_{post}(H_1) / Pr_{post}(H_0) \geq 25$.

Gegeben seien Fallzahlen $n_A=n_B=100$ sowie eine feste Erfolgsrate $p_B=0.5$. Die folgende Abbildung zeigt die Powerkurve, die sich aus simulierten Daten mit 10000 Wiederholungen ergibt.




In einem zweiten Szenario seien die Fallzahlen $n_A=n_B=20$ sowie die Erfolgsrate $p_B=0.7$ gegeben.




## Schlussfolgerung

Die Powerkurven aller vier Verfahren stimmen in bestimmten Situationen offenbar nicht miteinander überein. Dieses Ergebnis gilt, obwohl die verschiedenen Ansätze insofern vereinheitlicht wurden. In den Bayes-Ansätzen wurde eine nicht-informative a-priori-Verteilung verwendet und die Verfahren wurden darüber hinaus auf die Einhaltung frequentistischer Gütekriterien getrimmt, d.h. die Kontrolle des Fehlers 1. Art. Dass die p-Wert-basierten Verfahren dennoch eine geringere Power aufweisen als die Bayes‘sche Schätzung und der Bayes‘sche Hypothesentest, steht in Zusammenhang zu deren besseren Ausschöpfung des Signifikanzniveaus.

## Literatur

1 Thompson WR (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika 25, 285-294.

2 Chen LS and Lin CY. Bayesian p-values for testing independence in 2x2 contingency tables (2009). Comm in Statistics - Theory and Methods 38, 1635-1648.

3 Johnson VE and Cook JD (2009). Bayesian design of single-arm phase II clinical trials with continuous monitoring. Clinical Trials 6, 217-226.