Bayes-Statistik mit *DERIVE*

STEFAN GÖTZ, WIEN

Zusammenfassung: *In dieser Arbeit wird die neben dem Schätzen von Parametern wohl wichtigste Thematik der beurteilenden Statistik, nämlich das Testen von Hypothesen, vom Bayesianischen Standpunkt besprochen. Konkret wird die Binomialverteilung als Versuchsverteilung herangezogen. Auch ein verteilungsfreies Testverfahren, der Vorzeichentest, wird diskutiert. Dabei treten in natürlicher Weise verschiedene Berechnungsprobleme auf, die mittels DERIVE behandelt werden können. Auch das Veranschaulichen von Funktionsgraphen gelingt mit DERIVE. Gemeinsam ist diesen Einsatzmöglichkeiten von DERIVE, daß sie aus der statistischen Fragestellung heraus motiviert werden und daß die von DERIVE gelieferten Ergebnisse interpretiert werden müssen. Dies unterstreicht den Werkzeugcharakter, der dem Computer(einsatz) im Mathematikunterricht hier zugedacht wird.*

Die in dieser Arbeit zu belegende These lautet also: DERIVE unterstützt durch Eröffnung von Variationsmöglichkeiten das Verstehen des Erkenntnisprozesses bei der Bayes-Statistik.

1 Vorbemerkung

Die grundlegende Idee der *klassischen* Testphilosophie ist folgende:

Wir machen eine Aussage über eine gewisse statistische Grundgesamtheit, indem wir z. B. einen konkreten Wert eines zur statistischen Beschreibung gehörenden Parameters (z. B. des Erwartungswertes μ) unterstellen. Diese Annahme nennen wir *Nullhypothese* H_0. Dann ziehen wir eine *Stichprobe* D vom Umfang n aus dieser Grundgesamtheit und untersuchen die Frage, ob diese Stichprobe mit unserer Annahme verträglich ist. Dazu berechnen wir die (bedingte) Wahrscheinlichkeit, dieses Stichprobenergebnis oder ein extremeres zu erhalten unter der Voraussetzung, daß H_0 gilt, also

$$\mathrm{P}(D|H_0).$$

Ist diese sehr klein (kleiner als ein vorgegebener Wert α z. B. von $0,05$), dann entschließen wir uns, H_0 abzulehnen, andernfalls beizubehalten (was nicht bedeuten muß, daß H_0 richtig ist).

Eine *Alternative* zu dieser Vorgehensweise finden wir darin, daß wir von der Tatsache ausgehen, die Daten D (die erhobene Stichprobe) sind passiert („Die Welt ist das, was der Fall ist!"), und wir daher der Hypothese H_0 aufgrund dieser Tatsache eine (bedingte) Wahrscheinlichkeit $\mathrm{P}(H_0|D)$ zuordnen. Dieser Ansatz liegt der *Bayesianischen* Sichtweise zugrunde, denn den Zusammenhang zwischen den genannten bedingten Wahrscheinlichkeiten liefert gerade das *Bayessche Theorem.*

Im folgenden werden nun für konkrete Versuchsverteilungen typische Beispiele vorgestellt, deren Bearbeitung gewisse *Berechnungsprobleme* in natürlicher Weise mit sich bringt, die mittels dem Computeralgebrasystem *DERIVE* gelöst werden können. Außerdem werden Funktionsgraphen (von Dichtefunktionen) mit *DERIVE* zur *Veranschaulichung* und zum tieferen Verständnis der (jeweiligen) statistischen Situation geplottet.

Als Konsequenz ergibt sich aus dem Einsatz von *DERIVE* ein hohes Maß an (numerischer) Flexibilität angesichts der geforderten Bayesianischen Behandlung (elementarer) statistischer Fragestellungen. Letztere wie auch die Interpretationen der u. a. von *DERIVE* gelieferten Ergebnisse müssen selbstverständlich nach wie vor durch Menschenhand geschehen. Dem Computer(einsatz) wird hier nur „Werkzeugcharakter" in seiner ursprünglichen Form zugedacht, also das bloße Abnehmen mechanischer Rechen- bzw. Zeichenarbeit, die im Prinzip auch von Hand aus gemacht werden könnte. Dadurch ist größtmögliche Transparenz in der Frage „Was macht der Computer jetzt eigentlich?" gewährleistet.

2 Einführendes Beispiel

BEISPIEL 1: Eine Maschine produziert Werkstücke mit einem gewissen Ausschußanteil p. A

priori wissen wir, daß *vier* Typen M_1, M_2, M_3 und M_4 unterschiedlicher Qualität dieser Maschine existieren:

$$\begin{aligned} p_1 &= 0,05\ ; \quad p_2 = 0,1\ ; \quad p_3 = 0,15\ ; \\ p_4 &= 0,2\ . \end{aligned}$$

Wie kann eine *Stichprobenentnahme* von produzierten Werkstücken helfen, einzuschätzen, um welchen Typ es sich konkret handelt, wenn dies aus gewissen Gründen nicht anders feststellbar ist?

Lösung: A priori (i. e. *vor* der Stichprobenerhebung) schätzen wir

$$\pi(M_1) = \pi(M_2) = \pi(M_3) = \pi(M_4) = \frac{1}{4}$$

ein, die Stichprobe vom Umfang $n = 100$ ergebe acht Stück Ausschuß, das sind die Daten D.

Unsere Aufgabe besteht nun darin, aus der A-priori-Einschätzung (die *vor* der Stichprobenerhebung bestanden hat) unter Berücksichtigung der erhobenen Daten eine A-posteriori-Einschätzung (also *nach* der Stichprobenerhebung entstehend) zu generieren. Dazu verwenden wir das *Bayessche Theorem,* welches in seiner einfachen Form für Ereignisse bekanntlich folgendermaßen lautet:

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B) \cdot \mathrm{P}(B)}{\mathrm{P}(A)}\ .$$

Zerfällt das Ereignis B in paarweise disjunkte Ereignisse $B_1, \ldots, B_m$, gilt also

$$B_j \cap B_k = \emptyset \quad \forall j \neq k$$

und

$$B_1 \cup \cdots \cup B_m = B,$$

so erhalten wir mit Hilfe des *Satzes von der vollständigen Wahrscheinlichkeit*

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}(A|B_1) \cdot \mathrm{P}(B_1) + \cdots + \\ &\quad + \mathrm{P}(A|B_m) \cdot \mathrm{P}(B_m)\ , \end{aligned}$$

und gelangen so zur *vollen* Bayesschen Formel für Ereignisse:

$$\begin{aligned} \mathrm{P}(B_i|A) &= \frac{\mathrm{P}(A|B_i) \cdot \mathrm{P}(B_i)}{\sum_{j=1}^{m} \mathrm{P}(A|B_j) \cdot \mathrm{P}(B_j)} \\ \forall i &= 1, \ldots, m\ . \end{aligned}$$

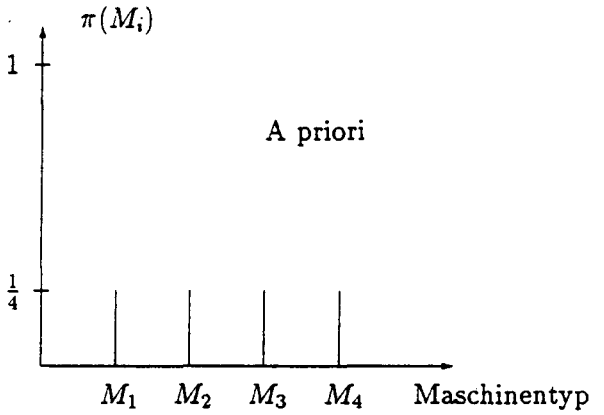
Auf diese Weise kommen wir *a posteriori* (i. e. *nach* der Stichprobenerhebung) mit Hilfe des *Bayesschen Theorems* und *DERIVE* zu

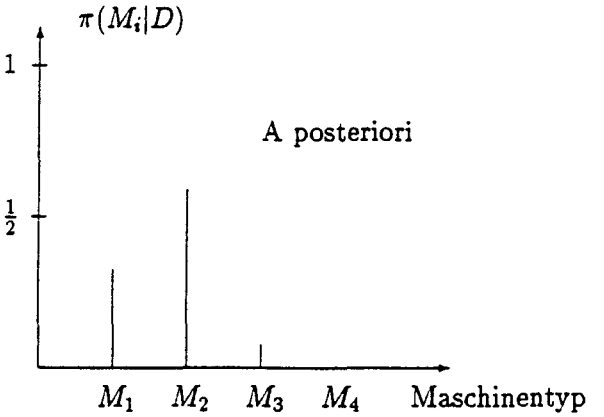
$$\begin{aligned} \pi(M_1|D) &= \frac{\mathrm{P}(D|M_1) \cdot \pi(M_1)}{\mathrm{P}(D)} = \\ &= \frac{\binom{100}{8} \cdot {p_1}^8 \cdot (1-p_1)^{92} \cdot \frac{1}{4}}{\sum_{k=1}^{4} \mathrm{P}(D|M_k) \cdot \pi(M_k)} = \\ &= \frac{\binom{100}{8} \cdot 0,05^8 \cdot 0,95^{92} \cdot \frac{1}{4}}{\frac{1}{4} \cdot \sum_{k=1}^{4} \binom{100}{8} \cdot {p_k}^8 \cdot (1-p_k)^{92}} = \\ &= 0,331679\ ; \end{aligned}$$

$$\begin{aligned} \pi(M_2|D) &= \frac{0,1^8 \cdot 0,9^{92}}{\sum_{k=1}^{4} {p_k}^8 \cdot (1-p_k)^{92}} = \\ &= 0,587081\ ; \end{aligned}$$

$$\begin{aligned} \pi(M_3|D) &= 0,0782814 \quad \text{und} \\ \pi(M_4|D) &= 0,0029586\ . \end{aligned}$$

Die *graphische Darstellung* der Änderung der Einschätzung zeigt eindrucksvoll den Erkenntniszuwachs, den die Stichprobenerhebung gebracht hat:





Beachten wir dabei (als eine mögliche Interpretation), daß

$$\pi(M_1|D) + \pi(M_2|D) \geq 0,9$$

gilt, das heißt mit hoher Wahrscheinlichkeit liegt ein Maschinentyp besserer Qualität vor!

Eine *neuerliche* Stichprobenerhebung vom Umfang $n = 100$ ergebe zehn Stück Ausschuß. Wir haben nun zwei Möglichkeiten zur Verarbeitung dieser zusätzlichen Information:

- Entweder setzen wir a priori

$$\begin{aligned} \pi(M_1) &= 0,331679\,; \\ \pi(M_2) &= 0,587081\,; \\ \pi(M_3) &= 0,0782814\,; \\ \pi(M_4) &= 0,0029586 \end{aligned}$$

  (das ist die A-posteriori-Verteilung von soeben) und berechnen daraus die A-posteriori-Verteilung wie gehabt.

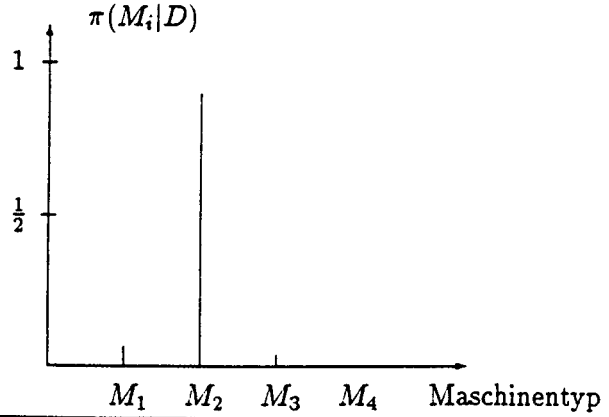
- Oder wir fassen die beiden Stichproben zusammen: $n = 200$ und 18 Stück Ausschuß. A priori ist wieder

$$\pi(M_1) = \frac{1}{4} = \pi(M_2) = \pi(M_3) = \pi(M_4)\,;$$

  a posteriori ergibt sich nun mittels *DERIVE*

$$\begin{aligned} \pi(M_1|D) &= 0,0641392\,; \\ \pi(M_2|D) &= 0,89558\,; \\ \pi(M_3|D) &= 0,0401656\,; \\ \pi(M_4|D) &= 1,152 \cdot 10^{-4}\,. \end{aligned}$$

Die graphische Darstellung spricht eine deutliche Sprache:



Bemerkung: Beide Methoden liefern (natürlich) dasselbe Ergebnis, wie man leicht nachrechnen kann, wir wollen das für die Maschine M_1 demonstrieren. Nach der ersten Methode ist a priori

$$\pi(M_1) = \frac{{p_1}^8 \cdot (1-p_1)^{92}}{\sum_{i=1}^{4} {p_i}^8 \cdot (1-p_i)^{92}}\,,$$

die hinzukommenden Daten D^* führen zur A-posteriori-Wahrscheinlichkeit für M_1:

$$\begin{aligned} &\pi(M_1|D^*) = \\ &= \frac{\mathrm{P}(D^*|M_1) \cdot \pi(M_1)}{\sum_{k=1}^{4} \mathrm{P}(D^*|M_k) \cdot \pi(M_k)} = \\ &= \frac{\binom{100}{10} \cdot {p_1}^{10} \cdot (1-p_1)^{90} \cdot \frac{{p_1}^8 (1-p_1)^{92}}{\sum_{i=1}^{4} {p_i}^8 (1-p_i)^{92}}}{\sum_{k=1}^{4} \binom{100}{10} \cdot {p_k}^{10} \cdot (1-p_k)^{90} \cdot \frac{{p_k}^8 (1-p_k)^{92}}{\sum_{i=1}^{4} {p_i}^8 (1-p_i)^{92}}} = \\ &= \frac{{p_1}^{18} \cdot (1-p_1)^{182}}{\sum_{k=1}^{4} {p_k}^{18} \cdot (1-p_k)^{182}}\,. \end{aligned}$$

Die zweite Methode kennen wir bereits vom eben gerechneten Beispiel, es ist jetzt nur $n = 200$ statt $n = 100$ und $k = 18$ statt $k = 8$ zu verzeichnen. Analog dazu erhalten wir also auch

$$\pi(M_1|D \cup D^*) = \frac{{p_1}^{18} \cdot (1-p_1)^{182}}{\sum_{k=1}^{4} {p_k}^{18} \cdot (1-p_k)^{182}}\,,$$

dabei meint $D \cup D^*$ die vereinigte, große Stichprobe vom Umfang $n = 200$ mit $k = 18$ Ausschußstücken.

Allgemein erkennen wir als Struktur dieser Probleme: Den (endlichen) Parameterraum $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$ schätzen wir a priori so ein: $\pi(\theta_1), \ldots, \pi(\theta_N)$. Eine Stichprobe mit den Daten D wird erhoben. Dies erlaubt uns, zu einer A-posteriori-Einschätzung zu kommen:

$$\pi(\theta_i|D) = \frac{\mathrm{P}(D|\theta_i) \cdot \pi(\theta_i)}{\sum_{j=1}^{N} \mathrm{P}(D|\theta_j) \cdot \pi(\theta_j)}\,;\ i = 1, \ldots, N\,.$$

Bei uns heißt das konkret: Die Stichprobe vom Umfang n enthalte k „Erfolge“ (Daten D). Der (diskrete) Parameterraum

$\Theta = \{p_1, \ldots, p_N\}$ mit der A-priori-Einschätzung $\pi(p_1), \ldots, \pi(p_N)$ wird a posteriori gemäß

$$\pi(p_i|D) = \frac{\binom{n}{k} \cdot {p_i}^k \cdot (1-p_i)^{n-k} \cdot \pi(p_i)}{\underbrace{\sum_{j=1}^{N} \binom{n}{k} \cdot {p_j}^k \cdot (1-p_j)^{n-k} \cdot \pi(p_j)}_{DERIVE}} ;$$
$$i = 1, \ldots, N$$

bewertet. Die Summenausdrücke im Nenner können dabei sehr leicht mittels *DERIVE* berechnet werden, da das (per Hand mühsame) Aufsummieren der einzelnen Summanden von *DERIVE* übernommen werden kann. Auch die Auswertung des Zählers und somit des ganzen Quotienten unterstützt *DERIVE*.

Als *Variationsmöglichkeit* dieses Beispiels bietet sich eine Veränderung der A-priori-Einschätzung an:

$$\pi(M_1) = \pi(M_2) = \frac{1}{3} \; ; \; \pi(M_3) = \pi(M_4) = \frac{1}{6} \; .$$

Die Daten D bleiben gleich: $n = 100$ und $k = 8$. *DERIVE* liefert

$$\begin{aligned} \pi(M_1|D) &= \\ &= \frac{\binom{100}{8} \cdot 0,05^8 \cdot 0,95^{92} \cdot \frac{1}{3}}{\binom{100}{8} \cdot A} = \\ &= 0,345722 \end{aligned}$$

mit

$$\begin{aligned} A = \; & \frac{1}{3} \cdot (0,05^8 \cdot 0,95^{92} + 0,1^8 \cdot 0,9^{92}) + \\ & + \frac{1}{6} \cdot (0,15^8 \cdot 0,85^{92} + 0,2^8 \cdot 0,8^{92}) \; ; \end{aligned}$$

$$\begin{aligned} \pi(M_2|D) &= 0,611937 \; ; \\ \pi(M_3|D) &= 0,0407979 \quad \text{und} \\ \pi(M_4|D) &= 0,0015431 \end{aligned}$$

bzw. für $n = 200$ und $k = 18$:

$$\begin{aligned} \pi(M_1|D) &= 0,0654575 \; ; \\ \pi(M_2|D) &= 0,913988 \; ; \\ \pi(M_3|D) &= 0,0204956 \quad \text{und} \\ \pi(M_4|D) &= 5,89 \cdot 10^{-5} \; . \end{aligned}$$

Die Beträge der Differenzen der A-posteriori-Einschätzungen bezogen auf die beiden A-priori-Einschätzungen sind mit wachsendem Stichprobenumfang n *kleiner* geworden:

	$n = 100$	$n = 200$
M_1	0,014043	0,0013183
M_2	0,024856	0,018408
M_3	0,0374835	0,01967
M_4	$1,4155 \cdot 10^{-3}$	$5,63 \cdot 10^{-5}$

Wir sehen: mit wachsendem Stichprobenumfang wird der Einfluß der A-priori-Einschätzung zurückgedrängt. Intuitiv können wir das so verstehen, daß unabhängig vom Vorwissen (welches in die A-priori-Einschätzung einfließt und von Person zu Person durchaus verschieden sein kann) über die in Frage stehende Situation die immer größer werdende Datenmenge (die stets als solche wahrgenommen wird) eine Art intersubjektive Übereinstimmung in der Beurteilung eben dieser Situation erzeugt. Das Bayessche Theorem quantifiziert diesen (gemeinsamen) Lernprozeß. Ein mathematisches Argument erfolgt im nächsten Abschnitt an analoger Stelle.[1]

3 Die Binomialverteilung als Versuchsverteilung

BEISPIEL 2: Bei der letzten Wahl errang die Partei X 40% der Stimmen. Ein Jahr danach berichten die Meinungsforscher, daß von $n = 100$ Personen bei einer Wahl „am nächsten Sonntag" $k = 50$ für die Partei X votieren würden. [...] (Aus: [RE], S. 272.)
Wie ist das Umfrageergebnis zu bewerten?

Lösung: „Nichtwissen" a priori können wir durch

$$\pi(p) = 1 \qquad \forall p \in [0,1]$$

modellieren. In Gegensatz zum vorigen Beispiel (dort sind vier Ausschußanteile zur Wahl gestanden) ist nun der in Frage stehende Parameterraum $\Theta = [0,1]$ überabzählbar. Daher geben wir keine A-priori-Wahrscheinlichkeiten vor, sondern eine *A-priori-Dichte* für den Parameter p. Wir entscheiden uns für die gleichmäßige Verteilung auf $[0,1]$, weil wir (vorerst) keinen Grund haben, einen Parameterwert auszuzeichnen.

[1] Dieser Verweis ist ein schönes Beispiel dafür, um wie vieles einfacher die (mathematische) Welt wird, wenn sie kontinuierlich statt diskret betrachtet wird.

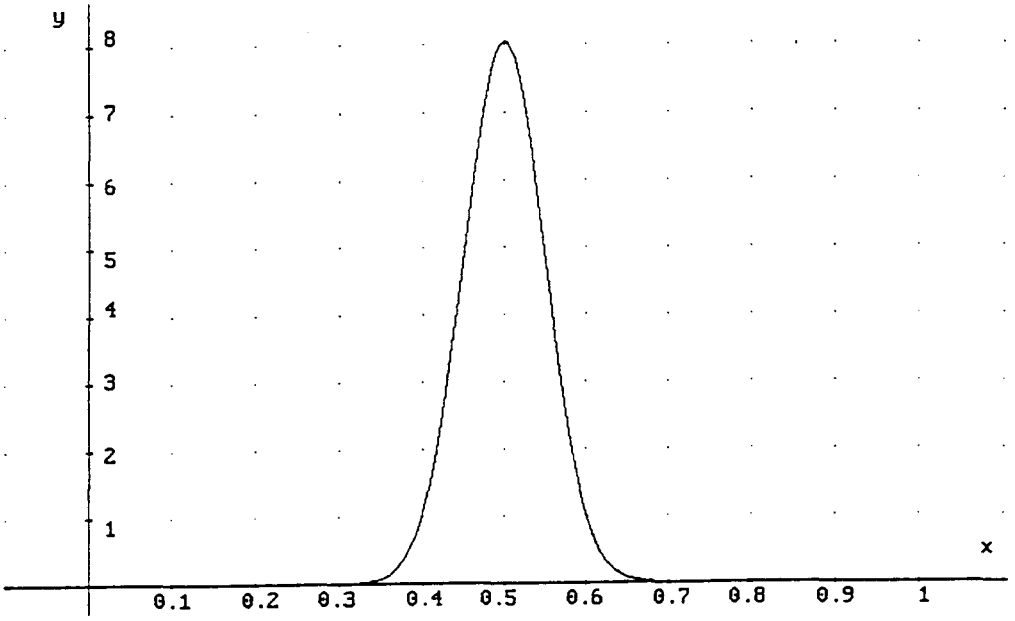


Abbildung 1: Graph der A-posteriori-Dichtefunktion für $n = 100$ und $k = 50$

Dementsprechend muß auch das Bayessche Theorem modifiziert werden, welches nun mittels der erhobenen Daten und der A-priori-Dichte die A-posteriori-Dichte erzeugt:

$$\pi(p|D) = \frac{\mathrm{P}(D|p)\cdot\pi(p)}{\mathrm{P}(D)} = \frac{\mathrm{P}(D|p)\cdot\pi(p)}{\int\limits_{\Theta}\mathrm{P}(D|p)\cdot\pi(p)\,dp}.$$

Die Struktur ähnelt der vollen Bayesschen Formel für Ereignisse, die Summe dort im Nenner wird – der Überabzählbarkeit von Θ Rechnung tragend – durch ein Integral hier ersetzt. Tatsächlich ist durch diesen Ausdruck $\pi(p|D)$ eine Dichtefunktion gegeben:

- Es ist $\pi(p|D) \geq 0 \quad \forall p \in \Theta$.
- Wegen $\pi(p|D) \equiv 0 \; \forall p \notin [0,1]$ ist $\lim\limits_{p\to\pm\infty} \pi(p|D) = 0$.
- Schließlich ist

$$\int\limits_{\Theta} \pi(p|D)\,dp = 1\,.$$

A posteriori ergibt sich daraus

$$\pi(p|D) = \frac{\mathrm{P}(D|p)\cdot\pi(p)}{\mathrm{P}(D)} = \frac{\binom{100}{50}\cdot p^{50}\cdot(1-p)^{50}\cdot 1}{\int\limits_0^1 \binom{100}{50}\cdot p^{50}\cdot(1-p)^{50}\cdot 1\,dp},$$

dabei ist p die Wahrscheinlichkeit, daß eine zufällig herausgegriffene Person für die Partei X stimmt und D bezeichnet die Daten, daß von 100 befragten Personen 50 für X gewesen sind.

Das Integral

$$\int\limits_0^1 p^{50}\cdot(1-p)^{50}\,dp$$

berechnen wir mittels *DERIVE* und sehen uns ebenfalls mit Hilfe von *DERIVE* $\pi(p|D)$ an: **Abbildung 1.**

Für $n = 50$ und $k = 25$ ergibt sich **Abbildung 2.** Deutlich sehen wir im Vergleich zur A-posteriori-Dichtefunktion, welche sich für $n = 100$ ergeben hat (und ebenfalls in Abbildung 2 zu sehen ist, vergleiche mit Abbildung 1), daß die für $n = 50$ wesentlich breiter ist. Der relative Anteil der X-Wähler und X-Wählerinnen ist wohl in beiden Fällen gleich zu bewerten, nicht jedoch die Sicherheit, mit der diese Aussagen getätigt werden können: diese wächst mit steigendem Stichprobenumfang, die A-posteriori-Einschätzung wird schärfer.

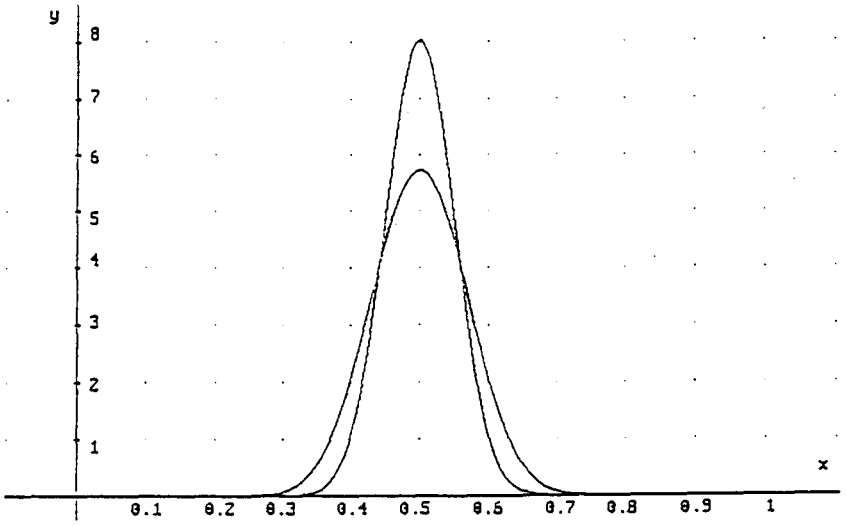


Abbildung 2: Graph der A-posteriori-Dichtefunktion für $n = 50$ und $k = 25$

Allgemein ist a priori

$$\pi(p) = 1 \qquad \forall p \in [0,1] = \Theta$$

und sind in $n \in \mathbb{N}$ Versuchen eines Bernoulli-Experiments $k \in \mathbb{N}$ „Erfolge" zu verzeichnen (Daten D), dann folgt a posteriori

$$\pi(p|D) = \frac{(n+1)!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k} \ .$$

(Dies ist die Dichte einer sogenannten *Beta-Verteilung.*)

Dazu berechnen wir für $n, m \in \mathbb{N}$ mittels *partieller Integration*

$$\begin{aligned}
&\int_0^1 \underbrace{x^n}_{u'} \cdot \underbrace{(1-x)^m}_{v}\, dx = \\
&= \underbrace{\left[\frac{x^{n+1}}{n+1} \cdot (1-x)^m\right]_0^1}_{=0} + \\
&\quad + \int_0^1 \frac{x^{n+1}}{n+1} \cdot m \cdot (1-x)^{m-1}\, dx = \\
&= \frac{m}{n+1} \cdot \int_0^1 x^{n+1} \cdot (1-x)^{m-1}\, dx = \\
&= \cdots = \\
&= \frac{m \cdot (m-1) \cdot \cdots \cdot 2}{(n+1) \cdot (n+2) \cdot \cdots \cdot (n+m-1)} \cdot \\
&\quad \cdot \int_0^1 x^{n+m-1} \cdot (1-x)\, dx = \\
&= \frac{m! \cdot n!}{(n+m-1)!} \cdot \\
&\quad \cdot \int_0^1 \left(x^{n+m-1} - x^{n+m}\right) dx = \\
&= \frac{m! \cdot n!}{(n+m-1)!} \cdot \\
&\quad \cdot \left[\frac{x^{n+m}}{n+m} - \frac{x^{n+m+1}}{n+m+1}\right]_0^1 = \\
&= \frac{m! \cdot n!}{(n+m-1)!} \cdot \\
&\quad \cdot \left(\frac{1}{n+m} - \frac{1}{n+m+1}\right) = \\
&= \frac{m! \cdot n!}{(n+m-1)!} \cdot \\
&\quad \cdot \frac{n+m+1-(n+m)}{(n+m) \cdot (n+m+1)} = \\
&= \frac{n! \cdot m!}{(n+m+1)!} \ ,
\end{aligned}$$

denn das Bayessche Theorem liefert ja

$$\pi(p|D) = \frac{\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \cdot 1}{\int_0^1 \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \cdot 1 \, dp} =$$

$$= \frac{p^k \cdot (1-p)^{n-k}}{\int_0^1 p^k \cdot (1-p)^{n-k} \, dp}$$

und das Integral im Nenner gilt es auszuwerten: es ist von der Form des Integrals, welches wir gerade berechnet haben. Damit sehen wir

$$\int_0^1 p^k \cdot (1-p)^{n-k} \, dp = \frac{k! \cdot (n-k)!}{[k+(n-k)+1]!} =$$

$$= \frac{k! \cdot (n-k)!}{(n+1)!} ,$$

woraus die obige Behauptung über die Gestalt von $\pi(p|D)$ folgt.

Bemerkungen: Die *Maximumsstelle* der Dichtefunktion dieser Beta-Verteilung liegt bei $\frac{k}{n}$, wie man sich leicht überzeugen kann.
Für $1 \leq k \leq n-1$ ist

$$\frac{d}{dp} \; p^k \cdot (1-p)^{n-k} =$$
$$= k \cdot p^{k-1} \cdot (1-p)^{n-k} -$$
$$- p^k \cdot (n-k) \cdot (1-p)^{n-k-1} .$$

Nullsetzen liefert

$$k \cdot (1-p) = p \cdot (n-k) ,$$

woraus wir

$$\frac{n-k}{k} = \frac{1-p}{p} \quad \text{bzw.} \quad \frac{n}{k} - 1 = \frac{1}{p} - 1$$

schließen, was uns letztlich zur behaupteten Maximumsstelle

$$p = \frac{k}{n}$$

führt.

Im Falle $k = 0$ nehmen wir das Randmaximum an der Stelle $p = 0 = \frac{k}{n} = \frac{0}{n}$ aus, ebenso ist für $k = n$ das Maximum von $\pi(p|D)$ am Rand des Definitionsbereichs $[0, 1]$ zu finden:
$p = 1 = \frac{k}{n} = \frac{n}{n}$.

Der *Erwartungswert* ergibt sich zu

$$\mathrm{E}(p) = \int_0^1 p \cdot \pi(p|D) \, dp = \frac{k+1}{n+2} ,$$

wie die folgende Rechnung zeigt:

$$\mathrm{E}(X) = \int_0^1 x \cdot \frac{(n+1)!}{k! \cdot (n-k)!} \cdot$$
$$\cdot x^k \cdot (1-x)^{n-k} \, dx =$$
$$= \frac{(n+1)!}{k! \cdot (n-k)!} \cdot$$
$$\cdot \int_0^1 x^{k+1} \cdot (1-x)^{n-k} \, dx =$$
$$= \frac{(n+1)!}{k! \cdot (n-k)!} \cdot$$
$$\cdot \frac{(k+1)! \cdot (n-k)!}{(n+2)!} = \frac{k+1}{n+2} .$$

Interessanterweise ist dieser Wert i. allg. *nicht* gleich der Stelle, an dem das Maximum der Dichte, nämlich $\frac{k}{n}$, angenommen wird (wie wir das z. B. von der *Normalverteilung* gewohnt sind). Gleichheit besteht dann und nur dann, wenn wegen $\frac{k}{n} = \frac{k+1}{n+2}$ und daraus folgend $n \cdot k + 2 \cdot k = n \cdot k + n$ endlich $k = \frac{n}{2}$ gilt. Genau in diesem Fall ist $\pi(p|D)$ symmetrisch um $p = \frac{1}{2} = \frac{k}{n}$, denn dann ist $\pi(\frac{1}{2} - p|D) = \pi(\frac{1}{2} + p|D) \; \forall p \in [0, \frac{1}{2}]$:

$$\left(\frac{1}{2} - p\right)^k \cdot \left(\frac{1}{2} + p\right)^{n-k} \propto \pi\left(\frac{1}{2} - p \middle| D\right) =$$

$$= \pi\left(\frac{1}{2} + p \middle| D\right) \propto \left(\frac{1}{2} + p\right)^k \cdot \left(\frac{1}{2} - p\right)^{n-k}$$

wegen $k = \frac{n}{2} \longleftrightarrow n - k = k$ und $\pi(p|D) \propto p^k \cdot (1-p)^{n-k}$.

Die *Varianz* $\mathrm{D}^2(X)$ berechnen wir mit Hilfe des *Verschiebungssatzes:*

$$\mathrm{D}^2(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 .$$

Es ist wieder

$$\mathrm{E}(X^2) = \int_0^1 x^2 \cdot \frac{(n+1)!}{k! \cdot (n-k)!} \cdot$$
$$\cdot x^k \cdot (1-x)^{n-k} \, dx =$$

$$= \frac{(n+1)!}{k! \cdot (n-k)!} \cdot \int_0^1 x^{k+2} \cdot (1-x)^{n-k}\, dx =$$

$$= \frac{(n+1)!}{k! \cdot (n-k)!} \cdot \frac{(k+2)! \cdot (n-k)!}{(n+3)!} =$$

$$= \frac{(k+1) \cdot (k+2)}{(n+2) \cdot (n+3)} ,$$

woraus schließlich

$$\begin{aligned} \mathrm{D}^2(X) &= \frac{(k+1) \cdot (k+2)}{(n+2) \cdot (n+3)} - \frac{(k+1)^2}{(n+2)^2} = \\ &= \frac{(k+1) \cdot (k+2) \cdot (n+2)}{(n+3) \cdot (n+2)^2} - \\ &\quad - \frac{(k+1)^2 \cdot (n+3)}{(n+3) \cdot (n+2)^2} = \\ &= \frac{(k+1) \cdot [(k+2) \cdot (n+2) -}{(n+3) \cdot (n+2)^2} \\ &\quad \frac{-(k+1) \cdot (n+3)]}{(n+3) \cdot (n+2)^2} = \\ &= \frac{(k+1) \cdot (k \cdot n + 2 \cdot k + 2 \cdot n +}{(n+3) \cdot (n+2)^2} \\ &\quad \frac{+4 - k \cdot n - n - 3 \cdot k - 3)}{(n+3) \cdot (n+2)^2} = \\ &= \frac{(k+1) \cdot (n-k+1)}{(n+3) \cdot (n+2)^2} \end{aligned}$$

folgt. Wie wir gesehen haben, werden die A-posteriori-Dichten mit wachsendem Stichprobenumfang schmäler, das heißt unsere Beurteilung der Situation wird präziser (Abbildung 2). Nun können wir uns davon überzeugen, daß dies so sein muß. Ein Maß für die Breite einer (A-posteriori-)Dichtefunktion ist die Varianz, die hier für steigendes n gegen Null geht:

$$\lim_{n \to \infty} \mathrm{D}^2(X) = \lim_{n \to \infty} \frac{(k+1) \cdot (n-k+1)}{(n+3) \cdot (n+2)^2} = 0 ,$$

da $0 \leq k \leq n$ gilt und der Grad des Nenners höher ist als der des Zählers. Ein altes Prinzip der Statistik erkennen wir also auch hier wieder: eine größere Datenerhebung bedeutet eine sicherere statistische Aussage.

DERIVE liefert Wahrscheinlichkeiten wie

$$\mathrm{P}(p \leq 0,4) = \int_0^{0,4} \pi(p|D)\, dp =$$

$$= \int_0^{0,4} \frac{101!}{50! \cdot 50!} \cdot p^{50} \cdot (1-p)^{50}\, dp =$$

$$= 0,0208966 ;$$

$$\mathrm{P}(0,4 \leq p \leq 0,6) = 0,958205 =$$

$$= \int_{0,4}^{0,6} \pi(p|D)\, dp \qquad \text{oder}$$

$$\mathrm{P}(p \geq 0,55) = \int_{0,55}^{1} \pi(p|D)\, dp =$$

$$0,156244 ,$$

welche eine quantitative Beurteilung des Stichprobenergebnisses (unter Berücksichtigung der – subjektiven – A-priori-Einschätzung) erlauben. Die erste Wahrscheinlichkeit zeigt z. B., daß es sehr unwahrscheinlich ist, daß sich die Partei X gegenüber der letzten Wahl in der Gunst ihrer Wähler(innen) nicht verbessert hat. Die zweite Wahrscheinlichkeit zeigt einen Bereich, in dem der Anteil der X-Stimmen mit hoher Wahrscheinlichkeit derzeit liegt usw.

Als *Variationsmöglichkeit* nehmen wir eine Änderung der A-priori-Einschätzung vor:

$$\pi(p) = \frac{11!}{4! \cdot 6!} \cdot p^4 \cdot (1-p)^6 \qquad \forall p \in [0,1]$$

(**Abbildung 3**). Diese A-priori-Dichte hat ihr Maximum an der Stelle 0,4. Offenbar liegen nun Informationen vor, daß der gesuchte Anteil p in der Nähe von 40% liegt (vielleicht wird das Ergebnis der letzten Wahl als Grundlage hergenommen). Natürlich gibt es auch noch andere Dichtefunktionen, die ihr Maximum an dieser Stelle haben, diese jedoch zeichnet sich dadurch aus, daß sie vom Typ her (als Dichte einer Beta-Verteilung) „gut" zur Binomialverteilung paßt, indem sie a posteriori wieder eine Dichte einer Beta-Verteilung liefert.[2] (Wir sagen: *Die Beta-Verteilung ist zur Binomialverteilung konjugiert.*)

[2] Natürlich gibt es unendlich viele andere Beta-Verteilungen, deren Dichten ebenfalls ihr Maximum an der Stelle 0,4 annehmen. Es muß ja nur $\frac{k}{n} = 0,4$ gelten.

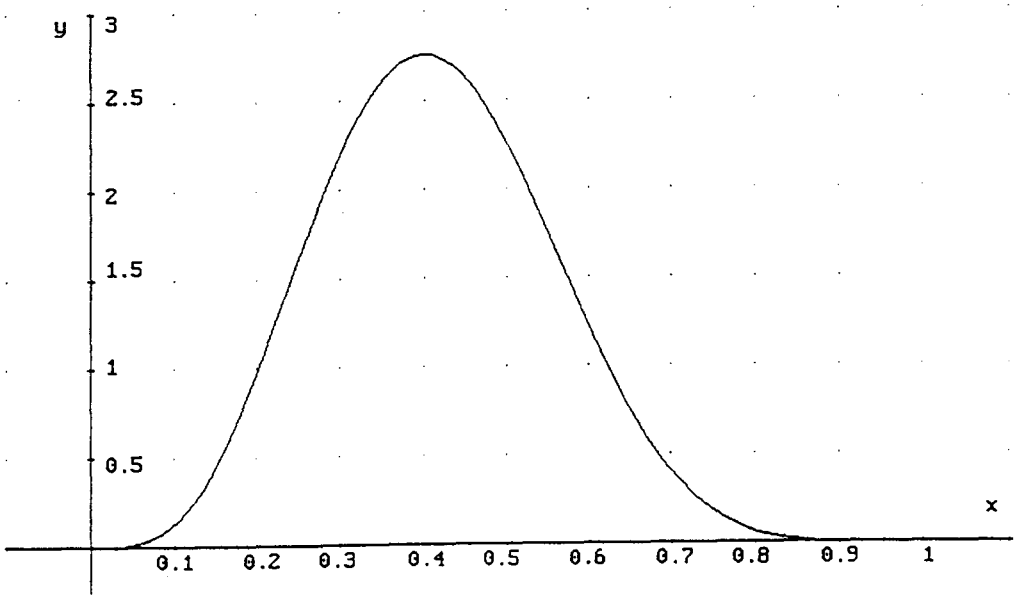


Abbildung 3: Graph der A-priori-Dichtefunktion

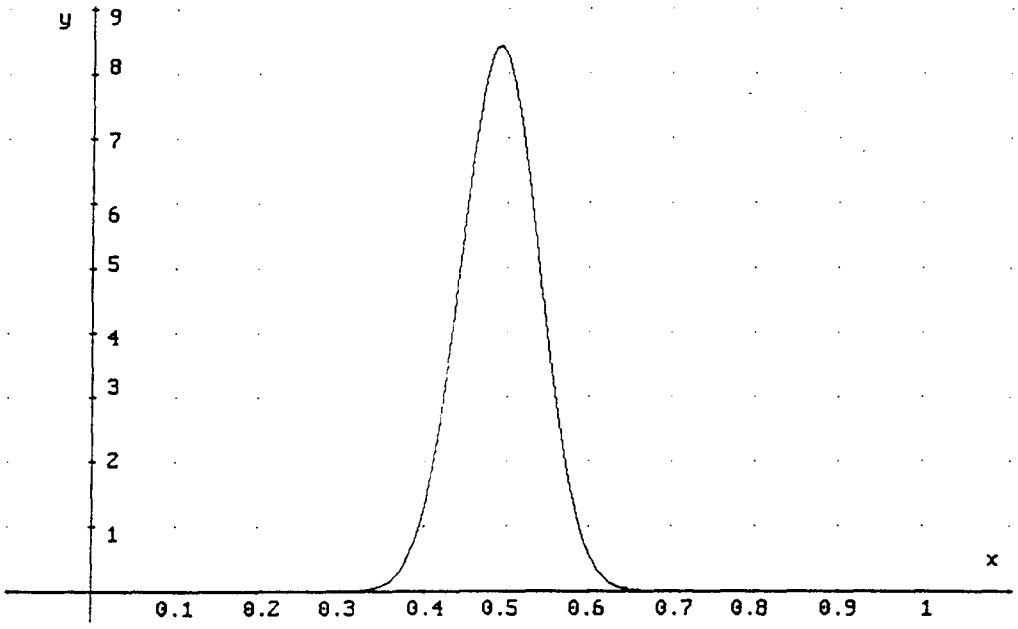


Abbildung 4: Graph der A-posteriori-Dichtefunktion

A posteriori ist nämlich (mit den unveränderten Daten D: $n = 100$ und $k = 50$)

$$\pi(p|D) = \frac{\binom{100}{50} \cdot p^{50} \cdot (1-p)^{50} \cdot}{\int_0^1 \binom{100}{50} \cdot p^{50} \cdot (1-p)^{50} \cdot} \cdot \frac{\frac{11!}{4!\cdot 6!} \cdot p^4 \cdot (1-p)^6}{\frac{11!}{4!\cdot 6!} \cdot p^4 \cdot (1-p)^6 \, dp} =$$

$$= \frac{p^{54} \cdot (1-p)^{56}}{\int_0^1 p^{54} \cdot (1-p)^{56} \, dp} =$$

$$= \frac{111!}{54! \cdot 56!} \cdot p^{54} \cdot (1-p)^{56}$$

(**Abbildung 4** mittels *DERIVE*). Wir sehen, daß sich aufgrund des Stichprobenergebnisses unsere Einschätzung von 40% in Richtung 50% für p bewegt hat.

Beachten wir dabei, daß ausgehend von

$$\pi(p) = 1 \qquad \forall p \in [0,1]$$

eine Stichprobe mit

$$n = 110 = 100 + 10 \quad \text{und} \quad k = 54 = 50 + 4$$

auch diese A-posteriori-Dichte $\pi(p|D)$ liefern würde.

Wir schließen daraus:
Die Daten bestimmen im zunehmenden Maße (i. e. mit wachsendem Stichprobenumfang) die A-posteriori-Einschätzung, der Einfluß der A-priori-Einschätzung wird dabei kleiner.

Jetzt berechnen wir mit *DERIVE*

$$\begin{aligned} \mathrm{P}(p \le 0,4) &= 0,0260024\,; \\ \mathrm{P}(0,4 \le p \le 0,6) &= 0,963926 \\ \text{und} \quad \mathrm{P}(p \ge 0,55) &= 0,105938\,, \end{aligned}$$

die Werte differieren von den vorher bestimmten, die aus der ursprünglichen A-priori-Einschätzung folgen. Dies zeigt eben, daß andere (Vor-)Informationen, die im mathematischen Modell stecken, zu anderen Resultaten führen. Das ist zugleich Stärke und Schwäche dieser Methode: Es ist ja an sich positiv, Ansätze zu kennen, die ein Einfließen von (subjektiven) Einschätzungen, Informationen etc. zulassen, der negative Aspekt der sich daraus ergebenden Beliebigkeit des Ergebnisses a posteriori wird durch die eben festgestellte Tatsache, daß die Daten mit steigender Menge die A-posteriori-Einschätzung immer mehr bestimmen, gemildert.

4 Der Vorzeichentest

BEISPIEL 3: Eine bestimmte sportliche Leistung wird bei $n = 20$ Personen *vor* und *nach* Absolvierung eines speziellen Ernährungsprogramms gemessen. Dabei ergibt sich in $k = 14$ Fällen eine Leistungssteigerung, in sechs Fällen dagegen eine Minderung dieser sportlichen Leistung. Wie ist dieses Resultat zu beurteilen?

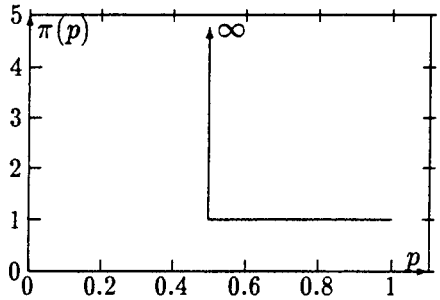
Bemerkung: Es handelt sich hier um einen Test für *rangskalierte* Daten.

Lösung: Die ZV X zähle, wie oft das Ergebnis nachher besser ist als vorher, X ist daher binomialverteilt mit den Parametern n und p.

Zeigt die „Behandlung" keine Auswirkung, so ist $p = \frac{1}{2}$: *Hypothese* H_0.
Andernfalls ist $p \neq \frac{1}{2}$ bzw. $p > \frac{1}{2}$: *Hypothese* H_1.

A priori bewerten wir diese beiden Hypothesen gleich, daraus ergibt sich folgende „A-priori-Dichte" für p:

Wegen $\mathrm{P}(H_0) = \mathrm{P}(H_1) = \frac{1}{2}$ ergibt sich für den Bereich von H_1, nämlich $p \in (\frac{1}{2}, 1]$, die gleichmäßige Dichte (wiederum wollen wir keinen p-Wert a priori auszeichnen) auf der Höhe 1 (der Flächeninhalt unter der Dichtekurve ist dann $\frac{1}{2} \cdot 1 = \frac{1}{2}$ und entspricht somit der A-priori-Wahrscheinlichkeit für H_1). H_0 dagegen „lebt" nur auf dem Punkt $p = \frac{1}{2}$, um ihr ebenfalls das Gewicht $\frac{1}{2}$ zukommen zu lassen, muß der Funktionswert an dieser Stelle $\frac{1}{2}$ über alle Maßen wachsen. (Wir können uns dafür eine „halbe" Normalverteilung vorstellen, deren Varianz gegen Null geht.) Die folgende Abbildung gibt unsere A-priori-Einschätzung wieder.



A posteriori ist

$$\mathrm{P}(H_i|k) = \frac{\mathrm{P}(k|H_i) \cdot \mathrm{P}(H_i)}{\mathrm{P}(k)} \quad (i = 0, 1)\,,$$

ausgeschrieben bedeutet dies

$$P(H_0|k) = \frac{\binom{n}{k}\cdot\left(\frac{1}{2}\right)^n\cdot\frac{1}{2}}{P(k)} = \frac{\binom{n}{k}\cdot\left(\frac{1}{2}\right)^{n+1}}{P(k)} = \frac{Z_0}{P(k)}$$

und

$$P(H_1|k) = \frac{\int_{\frac{1}{2}}^{1}\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}\cdot 1\,dp\cdot\frac{1}{2}}{P(k)} = \frac{Z_1}{P(k)}\,.$$

Also ist konkret mittels *DERIVE*

$$Z_0 = \binom{20}{14}\cdot\left(\frac{1}{2}\right)^{21} = 0,0184822 \quad \text{und}$$

$$Z_1 = \binom{20}{14}\cdot\frac{1}{2}\cdot\int_{\frac{1}{2}}^{1} p^{14}\cdot(1-p)^6\,dp = 0,0228767\,.$$

Wegen

$$P(H_0|k) + P(H_1|k) = 1$$

ist $P(k) = Z_0 + Z_1 = 0,0413589$ und schließlich

$$P(H_0|k) \approx 0,45 \quad \text{und} \quad P(H_1|k) \approx 0,55\,.$$

Als *Variationsmöglichkeit* bietet sich an, andere Stichprobenergebnisse bei konstantem Stichprobenumfang (hier: $n = 20$) zu untersuchen. Dabei erhalten wir mit Hilfe von *DERIVE* folgende Ergebnisse:

	$P(H_0\|k)$	$P(H_1\|k)$
$k = 12$	0,76	0,24
$k = 13$	0,63	0,37
$k = 15$	0,24	0,76
$k = 16$	0,09	0,91

In analoger Weise wie soeben vorgeführt wird hier für verschiedene Stichprobenergebnisse $k \in \{12, 13, 15, 16\}$ die A-posteriori-Verteilung berechnet. Dabei fällt auf, wie sehr diese Bewertung (bei dem eher kleinen Stichprobenumfang) von den erhobenen Daten abhängt: Für $k = 14$ ist das Vertrauensverhältnis in die beiden Hypothesen noch mehr oder weniger ausgewogen, schon ein Fall positiver Leistungssteigerung mehr ($k = 15$) läßt uns H_1 gegenüber H_0 mit einer Quote von ungefähr 3 : 1 bevorzugen.

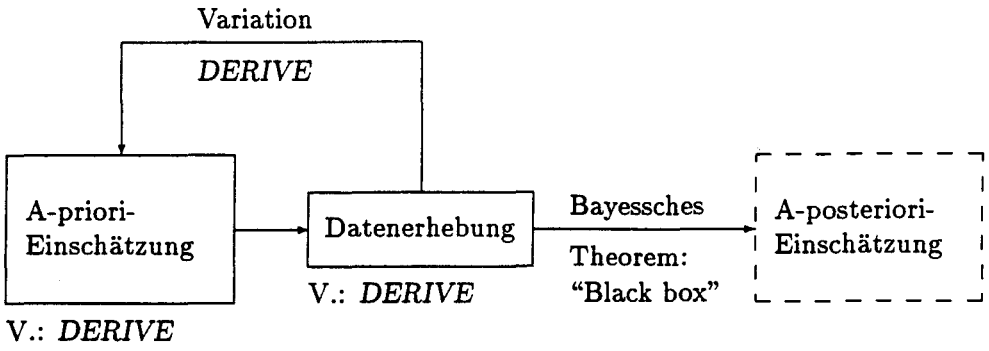
5 Resümee

Was zeichnet die *Bayesianische Methode* auf diesem Niveau aus? Im Mittelpunkt steht sicher die Berechnung von P[Hypothese (über Parameter θ)|Daten D], also die A-posteriori-Einschätzung von θ. Dazu benötigen wir eine A-priori-Einschätzung von θ, eine Stichprobenerhebung, welche die Daten D liefert, die Versuchsverteilung, um Wahrscheinlichkeiten $P(D|\theta)$ hinschreiben zu können und natürlich das Bayessche Theorem.

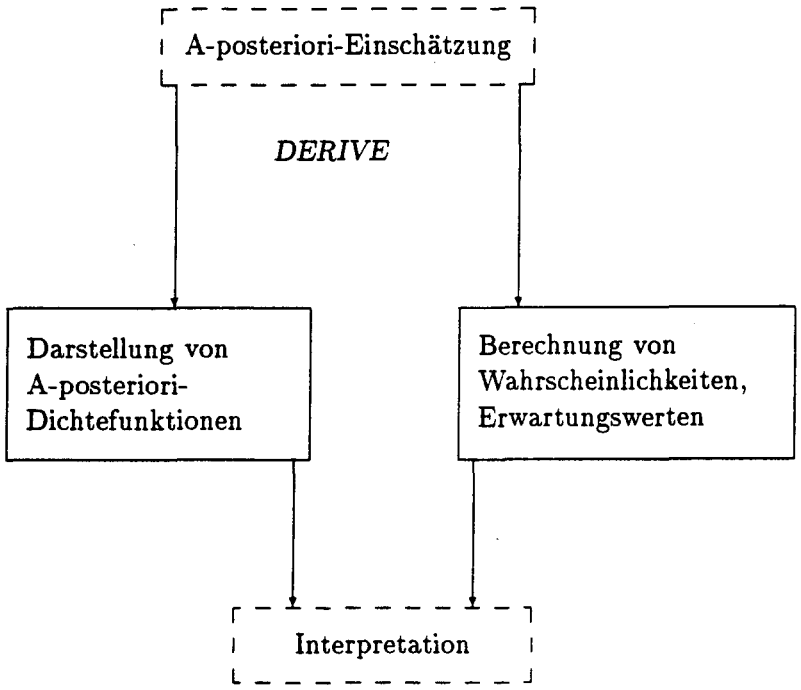
Welche *Berechnungsprobleme* treten dabei auf? Ist der Parameterraum $\Theta \ni \theta$ endlich, dann sind Summen zu bearbeiten (siehe das einführende Beispiel), ist dagegen Θ überabzählbar, müssen wir Integrale auswerten (die anderen beiden Beispiele). Den mathematischen Kern stellt natürlich die Anwendung des *Bayesschen Theorems* dar, die Auswertung der darin vorkommenden Ausdrücke (Integrale) kann zugunsten eines bloßen Mitteilens des Ergebnisses entfallen, ohne die eigentliche Idee der Bayes-Statistik zu verwässern. Für das Hintergrundwissen der Lehrenden ist es allerdings meines Erachtens unverzichtbar, die entsprechenden Rechnungen wenigstens im Prinzip zu kennen, daher sind sie auch hier ausgeführt worden. Ist der Stichprobenumfang n groß, bekommen wir auf natürliche Weise (z. B. durch Fakultäten) große Zahlen zur Verarbeitung. In allen Fällen kann *DERIVE* weiterhelfen.

Die *Darstellung* des Graphen der A-posteriori-Dichtefunktion mittels *DERIVE* unterstützt die *Interpretation* der A-posteriori-Einschätzung für ein gegebenes Problem. Diese kann u. a. durch *Variation* der A-priori-Einschätzung oder durch Variation der Daten verändert werden, die jeweilige Auswirkung auf die A-posteriori-Einschätzung muß wieder erkannt und erklärt werden. Schließlich stellt die *Berechnung* von Wahrscheinlichkeiten, Erwartungswerten und Varianzen eine weitere Möglichkeit dar,

1. Etappe:



2. Etappe:



die A-posteriori-Einschätzung zu bewerten. Dabei fällt aber auf, daß diese Berechnungen zum Teil „zu Fuß" erfolgen, um hier – nach Meinung des Autors – das nötige Hintergrundwissen für die Lehrenden nicht aus den Augen zu verlieren. Wo die Grenze zu ziehen ist zwischen dem reinen zur Kenntnis Nehmen eines (vom Computer gelieferten) Ergebnisses und dem – in allgemeinerem Rahmen – (manuellen) Nachvollziehen desselben, kann generell so nicht beantwortet werden. Gleichwohl stellt die Beantwortung dieser Frage (von Fall zu Fall) meines Erachtens die Herausforderung der Lehrenden im (computerbegleiteten) Unterricht bzw. in der Unterrichtsvorbereitung dar. Ein bewußter Umgang mit diesem Problem ist ein Kennzeichen für die Güte eines solchen Unterrichts.

(Dazu paßt auch der mittels *DERIVE* gewonnene Eindruck, daß das Erhöhen des Stichprobenumfangs die Einschätzung der Lage a posteriori schärfer macht, denn die A-posteriori-Dichte wird schmäler. Dieser wird durch den nicht mit Hilfe des Computers geführten Nachweis, daß die Varianz als Maß für die Breite einer Verteilung mit steigendem Stichprobenumfang immer kleiner wird, bestätigt, siehe dazu auch [HU].)

In jedem Fall ist der Werkzeugcharakter des Computereinsatzes unverkennbar: Die anstehende Rechnung ergibt sich in natürlicher Weise aus der Aufgabenstellung.

Zwei Etappen des Bearbeitens dieser Art von Aufgaben können wir ausmachen: Erstens ist der Weg von der A-priori- zur A-posteriori-

Einschätzung zurückzulegen. Zweitens kann die A-posteriori-Einschätzung in vielerlei Hinsicht ausgewertet werden. Das vorstehende Diagramm soll dies zusammenfassen: Die strichlierten Boxen entziehen sich weitgehend einer Behandlung mit *DERIVE*, die übrigen können durch den Einsatz von *DERIVE* wesentlich angereichert werden.

Die abschließende **These** lautet:

> *DERIVE unterstützt durch Eröffnung von Variationsmöglichkeiten das Verstehen des Erkenntnisprozesses bei der Bayes-Statistik.*

Ziel dieser Arbeit ist es, durch möglichst einfache (damit die zugrundeliegende Struktur sichtbar wird), konkrete Beispiele diese These zu belegen (siehe dazu auch [ME]). Keineswegs geht es darum, einen vollständigen Bayesianischen Stochastiklehrgang zu entwickeln, noch einen (kritischen) Vergleich zum klassischen Ansatz der (beurteilenden) Statistik zu wagen. Hiezu sei auf [WI1] und [WI2], aber auch auf [GÖ1] und [GÖ2] verwiesen. Dagegen ist es sehr wohl ein Anliegen dieser Arbeit, die technischen Anforderungen, die der Unterricht der Bayes-Statistik auf diesem Niveau mit sich bringt, aufzuzeigen und zu ihrer Bewältigung beizutragen. Dies geschieht eben zum Teil mit *DERIVE*, andererseits durch theoretische Überlegungen, die nicht unbedingt Themen des konkreten Mathematikunterrichts sein müssen (Black-Box-Prinzip). Diese konkreten Vorschläge, die sich unmittelbar in der Schule umsetzen lassen (die Angaben der vorkommenden Beispiele stammen entweder direkt aus Schulbüchern oder sind sinngemäß sehr nahe an den tatsächlich – meist wohl klassisch – im Stochastikunterricht behandelten Aufgaben), können vielleicht dazu beitragen, daß die Bayesianische Sichtweise in der beurteilenden Statistik neben der klassischen den anwendungsorientierten Mathematikunterricht bereichert.

Schlußbemerkung

Ein entsprechender Vortrag ist vom Verfasser anläßlich des 8. Internationalen Symposiums zur Didaktik der Mathematik zum Thema „Mathematische Bildung und neue Technologien“ an der Universität Klagenfurt (28.09. – 02.10.1998) ebendort gehalten worden.

Literatur

[GÖ1] Götz, S.: Bayes-Statistik – ein alternativer Zugang zur beurteilenden Statistik in der siebenten und achten Klasse AHS. Dissertation an der Universität Wien, 1997.

[GÖ2] Götz, S.: Klassische und Bayesianische Behandlung von Stochastikaufgaben aus österreichischen Schulbüchern. Preprint (35 pp.), 1999.

[HU] Humenberger, H.: Erwartungswerte und Gewinnwahrscheinlichkeiten bei einem Würfelbudenspiel – ein neuer Beitrag zu einem alten Thema. Erscheint in: PM.

[ME] Meyer, M. u. D.: Hypothesentests nach Bayes. Entscheidungen für Hypothesen mit der Bayes-Formel. Erschienen in: MU **44** (1998), Heft 1 (S. 50–61).

[RE] Reichel, H.-C., Müller, R., Hanisch, G. u. Laub, J.: Lehrbuch der Mathematik 7. Verlag öbv&hpt, Wien 1999^3.

[WI1] Wickmann, D.: Bayes-Statistik. Einsicht gewinnen und entscheiden bei Unsicherheit. Mathematische Texte, Band 4 (herausgegeben von N. Knoche und H. Scheid). BI Wissenschaftsverlag, Mannheim/Wien/Zürich 1990.

[WI2] Wickmann, D.: Zur Begriffsbildung im Stochastikunterricht. Erschienen in: JMD **19** (1998), Heft 1 (S. 46–80).

Adressen des Autors

Mag. Dr. Stefan Götz
Institut für Mathematik
Universität Wien
Strudlhofgasse 4

A-1090 Wien

e-mail:Stefan.Goetz@univie.ac.at

und

Akademisches Gymnasium Wien
Beethovenplatz 1

A-1010 Wien