Prof. Dr. Timm Grams
Hochschule Fulda
Fachbereich Elektrotechnik und Informationstechnik
Marquardstraße 35
36039 Fulda

Tel (0661) 78157
(0661) 9640-550
Fax (0661) 9640-559
mailto:Timm.Grams@et.hs-fulda.de
www.hs-fulda.de/~grams

Mittwoch, 3. September 2008
(Ergänzung: 17.09.2008)

Bayes-Schätzung – Eine kritische Einführung

Übersicht

Die Formel von Bayes

Uns interessiert ein bestimmter Sachverhalt. Genauer: Wir wollen wissen, inwieweit eine bestimmte Hypothese H, diesen Sachverhalt betreffend, gilt. Einen ersten Schätzwert für die Wahrscheinlichkeit dieser Hypothese haben wir bereits: die A-priori-Wahrscheinlichkeit $p(H)$. Nun beobachten wir ein Ereignis E, das mehr oder weniger für die Hypothese spricht. Die Wahrscheinlichkeit der Hypothese steigt auf Grund der Beobachtung in demselben Verhältnis wie die Beobachtung durch die Hypothese wahrscheinlicher wird:

$$\frac{p(H \mid E)}{p(H)} = \frac{p(E \mid H)}{p(E)} .$$

Diese Formel ist Grundlage des plausiblen Schließens. Georg Pólya (1963) beschäftigt sich mit ihren Konsequenzen und mit ihren Anwendungen.

Wenn man in der obigen Formel $p(E)$ durch den dazu äquivalenten Ausdruck $p(E|H)\cdot p(H)+p(E|\neg H)\cdot p(\neg H)$ ersetzt, erhält man die *Formel von Bayes* (Sachs, 1992, S. 77 ff.).

Die A-posteriori-Wahrscheinlichkeit der Hypothese ist gegeben durch $p(H|E)$.

Die Harvard-Medical-School-Studie

Wir betrachten einen Test für eine Krankheit, die die Basisrate 1/1000 besitzt - also: Einer unter eintausend Menschen ist krank. Der Test liefert mit der Wahrscheinlichkeit von 5% ein falsches Ergebnis. Er hat also eine Falsch-positiv-Rate von 5% und eine Falsch-negativ-Rate von ebenfalls 5 %. Wie hoch ist die Wahrscheinlichkeit, dass eine Person mit einem positiven Testergebnis tatsächlich die Krankheit hat? (Tipp: Lesen Sie erst einmal nicht weiter. Notieren Sie zunächst Ihren Schätzwert auf einem Zettel. Folgen Sie Ihrem Bauchgefühl. Verzichten Sie auf Berechnungen.)

Für eine Analyse mit der Formel von Bayes wählen wir die Formelzeichen folgendermaßen:

H steht für das Vorliegen der Krankheit und

E für ein positives Testergebnis.

Wir haben die folgenden Werte: $p(H) = 0.001$, $p(E|H) = 0.95$, $p(E|\neg H) = 0.05$.

Daraus ergibt sich die Wahrscheinlichkeit eines positiven Testergebnisses zu $p(E) = 0.95 \cdot 0.001 + 0.05 \cdot 0.999 = 0.0509$ und die A-posteriori-Wahrscheinlichkeit ist gleich

$$p(H \mid E) = p(H) \cdot \frac{p(E \mid H)}{p(E)} = 0.001 \cdot \frac{0.95}{0.0509} = 0.018664$$

Das heißt, dass die tatsächliche Krankheitswahrscheinlichkeit bei positivem Test unter 2 % liegt.

Möglicherweise liegt der von Ihnen anfangs notierte Schätzwert deutlich über dem hier errechneten. Mit dieser Schätzung sind Sie nicht allein: Das häufigste Urteil von Professoren, Ärzten und Studenten im Rahmen der Harvard-Medical-School-Studie ist „95%“ (Hell, Fiedler, Gigerenzer, 1993, S. 107).

Wie am Beispiel gezeigt, eignet sich die Bayes-Schätzung bestens für die Bewertung der Effizienz diagnostischer Tests (Sachs, S. 84 ff.). Die im Folgenden deutlich werdende Kritik an gewissen Anwendung der Bayes-Schätzung trifft auf diagnostische Tests nicht zu: Anders als bei Parameterschätzungen sind hier die A-priori-Wahrscheinlichkeiten durch Statistiken meist gut belegt.

Schätzung einer kleinen Wahrscheinlichkeit

Wahrscheinlichkeit als Zufallsvariable

Wir stellen uns einen Betrieb vor, der ein bestimmtes Produkt fertigt, und zwar mit einer Fehlerwahrscheinlichkeit w. Die Fehlerwahrscheinlichkeit w kann man als Zufallsvariable auffassen. Dazu sollte man sich verschiedene Realisierungen der Fehlerwahrscheinlichkeit zumindest vorstellen können, beispielsweise so: Aufgrund der Unvollkommenheit des Produktionsprozesses liefert die Fabrik hin und wieder fehlerhafte Produkte aus. Die Fehlerwahrscheinlichkeit möge – in Abhängigkeit für die Produktionsbedingungen – für jedes Los eine feste Zahl sein. Diese Zahl kann aber von Los zu Los schwanken.

Wir betrachten also die Fehlerwahrscheinlichkeit eines Loses als Realisierung einer Zufallsvariable w. Ähnliche Überlegungen lassen sich für die Versagenswahrscheinlichkeiten von Programmen anstellen.

Der zentrale Gedanke: Hypothesenwahrscheinlichkeit und Verteilungsfunktion

Was den Abnehmer interessiert, ist die Fehlerwahrscheinlichkeit w. Genauer: Er will den Nachweis, dass diese Fehlerwahrscheinlichkeit unterhalb eines bestimmten Grenzwerts p liegt. Und genau das ist die hier zu betrachtende Hypothese: $w < p$.

Die Wahrscheinlichkeit der Hypothese $P(w<p)$ ist gleich dem Wert der Verteilungsfunktion von w an der Stelle p: $P(H) = P(w<p) = F(p)$. Über die Verbesserung der Hypothesenschätzung aufgrund von Beobachtungen bekommt man – da p eine beliebig wählbare Zahl ist – zugleich eine Verbesserung der Verteilungsfunktion von w.

Das Apriori

Wir benötigen eine Anfangsbeschreibung der Produktionsverhältnisse, denn ansonsten könnten wir gar keine Wahrscheinlichkeiten für die Gültigkeit der Hypothese oder das Auftreten von bestimmten Stichprobenwerten angeben. Unsere Grundannahme ist, dass die Fabrik Lose liefert, deren Fehlerwahrscheinlichkeiten w im Intervall $[a, b)$ liegen und dort gleich verteilt sind. Für $a = 0$ und $b = 1$ haben wir den Sonderfall der vollständigen Unwissenheit. Für die folgende Rechnung wird $a \le p \le b$ vorausgesetzt.

Die Beobachtung E ist die fehlerfreie Stichprobe des Umfangs N, der negative Test also.

Die A-priori-Wahrscheinlichkeit der Hypothese ist gleich $P(H) = P(w < p) = \frac{p-a}{b-a}$. Daraus ergeben sich die Verteilungsfunktion F und die Verteilungsdichte f der Zufallsvariablen w im Intervall $[a, b]$ zu

$$F(x) = \frac{x-a}{b-a} \quad \text{und} \quad f(x) = \frac{1}{b-a}.$$

Die A-posteriori-Schätzung

Die Wahrscheinlichkeit einer fehlerfreien Stichprobe der Größe N ist unter den gegebenen Produktionsbedingungen gleich

$$P(E) = \int_a^b (1-x)^N \cdot \frac{1}{b-a} \cdot \mathrm{d}x = \frac{(1-a)^{N+1} - (1-b)^{N+1}}{(b-a)\cdot(N+1)}.$$

Bei Gültigkeit der Hypothese ist in den Formeln für F und f jeweils das b durch das p zu ersetzen. Die Wahrscheinlichkeit der Beobachtung ist dann gleich

$$P(E \mid H) = \int_a^p (1-x)^N \cdot \frac{1}{p-a} \cdot \mathrm{d}x = \frac{(1-a)^{N+1} - (1-p)^{N+1}}{(p-a)\cdot(N+1)}.$$

Die gesuchte A-posteriori-Wahrscheinlichkeit der Hypothese ist damit gegeben durch

$$P(H \mid E) = \frac{P(E \mid H)\cdot P(H)}{P(E)} = \frac{(1-a)^{N+1} - (1-p)^{N+1}}{(1-a)^{N+1} - (1-b)^{N+1}}.$$

Die nächste Verbesserung

Wir setzen nun die gemachte Beobachtung E voraus und haben für die Hypothese H, nämlich für $w < p$, die Wahrscheinlichkeit

$$P(H) = P(w < p) = \frac{(1-a)^{N+1} - (1-p)^{N+1}}{(1-a)^{N+1} - (1-b)^{N+1}}.$$

Das bedeutet nichts anderes, als dass die Fehlerwahrscheinlichkeit w jetzt die Verteilungsfunktion F hat, die auf dem Intervall $[a, b]$ gegeben ist durch

$$F(x) = \frac{(1-a)^{N+1} - (1-x)^{N+1}}{(1-a)^{N+1} - (1-b)^{N+1}}.$$

Die zugehörige Verteilungsdichte ist

$$f(x) = F'(x) = \frac{(N+1)\cdot(1-x)^N}{(1-a)^{N+1} - (1-b)^{N+1}} .$$

Nun wird ein Test mit der Stichprobengröße *M* durchgeführt. Er sei negativ. Das ist unsere neue Beobachtung *E*. Ihre Wahrscheinlichkeit ist bei der gegebenen Verteilung gleich

$$P(E) = \int_a^b (1-x)^M \cdot f(x) \cdot \mathrm{d}x = \frac{N+1}{M+N+1} \cdot \frac{(1-a)^{M+N+1} - (1-b)^{M+N+1}}{(1-a)^{N+1} - (1-b)^{N+1}} .$$

Die Wahrscheinlichkeit der Beobachtung unter der Bedingung, dass die Hypothese $w < p$ gilt, ist gegeben durch

$$P(E \mid H) = \int_a^p (1-x)^M \cdot \frac{f(x)}{F(p)} \cdot \mathrm{d}x = \frac{1}{F(p)} \cdot \frac{N+1}{M+N+1} \cdot \frac{(1-a)^{N+N+1} - (1-p)^{M+N+1}}{(1-a)^{N+1} - (1-b)^{N+1}} .$$

Für die A-posteriori-Wahrscheinlichkeit der Hypothese $w < p$ gilt dann

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} = \frac{(1-a)^{M+N+1} - (1-p)^{M+N+1}}{(1-a)^{M+N+1} - (1-b)^{M+N+1}} .$$

Dieselbe A-posteriori-Wahrscheinlichkeit hätte sich ergeben, wenn man nicht in zwei Stufen vorgegangen wäre, sondern wenn man gleich eine Stichprobe des Umfangs *M*+*N* gewählt hätte. Diese Eigenschaft der Bayes-Schätzung, nämlich dass ein zweistufiger Lernvorgang aufgrund zweier Beobachtungen zu demselben Ergebnis führt wie ein einstufiger aufgrund der Kombination beider Beobachtungen, wird *Konsistenzbedingung* genannt (Rüger, 1999, S. 190).

Anfangsschätzung nach dem Indifferenzprinzip

Wir setzen totale Unwissenheit voraus. Wir benutzen also das *Indifferenzprinzip*. So hat der berühmte Volkswirtschaftler und Autor des Buches *A Treatise on Probability* John Maynard Keynes das „Prinzip vom mangelnden zureichenden Grunde“ genannt: „Wenn keine Gründe dafür bekannt sind, um eines von verschiedenen möglichen Ereignissen zu begünstigen, dann sind die Ereignisse als gleich wahrscheinlich anzusehen“ (zitiert nach Carnap/Stegmüller, 1958, S. 3).

Damit fehlt uns die anfängliche Einschränkung der Wahrscheinlichkeiten und wir setzen a=0 und b=1. Bei dieser Anfangsschätzung und unter der Bedingung, dass ein Test der Stichprobengröße *N* bestanden wird, ergibt sich die Formel

$$P(w < p) = 1 - (1-p)^{N+1} .$$

Allgemeine Schätzung einer Wahrscheinlichkeit

Das Apriori und die Beobachtung

Wir gehen von einer Anfangsschätzung (A-priori-Wahrscheinlichkeit) nach dem Indifferenzprinzip aus. Allerdings lassen wir jetzt auch Fehler in der Stichprobe zu. Die Beobachtung *E* sein gegeben durch *k* Fehler in einer Stichprobe des Umfangs *N*. Wir fragen nach der Verteilung der Wahrscheinlichkeit aufgrund dieser Beobachtung (A-posteriori-Wahrscheinlichkeit).

Die Beta-Verteilung

Zur einfacheren Beschreibung der Zusammenhänge führen wir die Beta-Verteilung mit den Parametern *a* und *b* ein (Fisz, 1976, S. 186 f.). Wir setzen diese Zahlen als ganz und positiv voraus: $0<a$, $0<b$. (Achtung: In diesem Kapitel haben die Variablenbezeichner *a* und *b* eine andere Bedeutung als in den vorhergehenden Abschnitten.)

Auf dem Intervall [0, 1] ist die Verteilungsdichte der Beta-Verteilung gegeben durch

$$f_{a,b}(x) = x^{a-1}(1-x)^{b-1}/B_{a,\,b}.$$

Außerhalb des Intervalls ist sie gleich null. Die Konstante $B_{a,\,b}$ ist durch die Forderung definiert, dass für die Verteilungsdichte das Integral gleich eins sein muss:

$$B_{a,b} = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x\,.$$

Für $B_{a,\,b}$ gibt es eine explizite Darstellung. Für die grafischen Darstellungen reicht es aus, den Wert mittels numerischer Integration zu bestimmen.

Für den Erwartungswert μ ergibt sich daraus die Formel $\mu = B_{a+1,b}/B_{a,b}$. Partielle Integration und einige einfache Umformungen liefern

$$\mu = a/(a+b).$$

Anfangsschätzung und Verbesserung

Die anfängliche Verteilungsdichte der Zufallsvariablen w ist nach dem Indifferenzprinzip gegeben durch $f_{1,1}(x)=1$. Die Hypothese H sei $w < p$. Die A-priori-Wahrscheinlichkeit der Hypothese ist demnach gleich

$$P(H) = P(w < p) = p\,.$$

Die Wahrscheinlichkeit für k Fehler in einer Stichprobe der Größe N (das ist die Beobachtung E) ist unter den gegebenen Produktionsbedingungen gleich

$$P(E) = \int_0^1 \binom{N}{k} x^k (1-x)^{N-k}\,\mathrm{d}x = \binom{N}{k} \cdot B_{k+1,N-k+1}\,.$$

Die Wahrscheinlichkeit der Beobachtung unter der Bedingung, dass die Hypothese wahr ist, ist gleich

$$P(E \mid H) = \int_0^p \binom{N}{k} x^k (1-x)^{N-k} \cdot \frac{1}{p} \cdot \mathrm{d}x = \binom{N}{k} \cdot \frac{B_{k+1,N-k+1}}{p} \cdot \int_0^p f_{k+1,N-k+1}(x)\,\mathrm{d}x\,.$$

Die gesuchte A-posteriori-Wahrscheinlichkeit der Hypothese ist damit gegeben durch

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} = \int_0^p f_{k+1,N-k+1}(x) \cdot \mathrm{d}x\,.$$

Die Verteilungsdichte der Wahrscheinlichkeit p aufgrund der Beobachtung ist demnach gleich der Verteilungsdichte der Betaverteilung mit den Parametern k+1 und N-k+1. Die Verteilungsdichte hat sich durch die Beobachtung folgendermaßen verändert:

$$f_{1,1}(x) \rightarrow f_{1+k,1+N-k}. \qquad (*)$$

Für k=0 ergibt sich die Verteilungdichte $f_{1,1+N}(x) = (1-x)^N/B_{1,\,N+1}$ und die Verteilungsfunktion F wird zu

$$F(p) = \frac{1}{B_{1,N+1}} \int_0^p (1-x)^N\,\mathrm{d}x = \frac{1}{B_{1,N+1} \cdot (N+1)} (1-(1-p)^{N+1}) = 1-(1-p)^{N+1}\,.$$

Damit haben wir die bereits oben hergeleitete Formel $P(w < p) = 1-(1-p)^{N+1}$.

Die nächste Verbesserung

Die Verteilungsdichte der Zufallsvariablen p sei anfangs gegeben durch $f_{a,b}(x)$. Die Hypothese H ist $w < p$. Die A-priori-Wahrscheinlichkeit der Hypothese ist demnach gleich

$$P(H) = P(w < p) = \int_0^p f_{a,b}(x) \cdot \mathrm{d}x \,.$$

Die Wahrscheinlichkeit für k Fehler in einer Stichprobe der Größe N (das ist die Beobachtung E) ist unter den gegebenen Produktionsbedingungen gleich

$$P(E) = \int_0^1 \binom{N}{k} x^k (1-x)^{N-k} f_{a,b}(x) \cdot \mathrm{d}x = \int_0^1 \binom{N}{k} x^k (1-x)^{N-k} \cdot \frac{1}{B_{a,b}} x^{a-1} (1-x)^{b-1} \cdot \mathrm{d}x = \binom{N}{k} \cdot \frac{B_{a+k,b+N-k}}{B_{a,b}} \,.$$

Die Wahrscheinlichkeit der Beobachtung unter der Bedingung, dass die Hypothese wahr ist, ist gleich

$$P(E \mid H) = \frac{P(EH)}{P(H)} = \frac{1}{P(H)} \int_0^p \binom{N}{k} x^k (1-x)^{N-k} \cdot f_{a,b}(x) \cdot \mathrm{d}x = \binom{N}{k} \cdot \frac{B_{a+k,b+N-k}}{P(H) \cdot B_{a,b}} \cdot \int_0^p f_{a+k,b+N-k}(x)\, \mathrm{d}x \,.$$

Die gesuchte A-posteriori-Wahrscheinlichkeit der Hypothese ist damit gegeben durch

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} = \int_0^p f_{a+k,b+N-k}(x) \cdot \mathrm{d}x \,.$$

Die Verteilung der Wahrscheinlichkeit w aufgrund der Beobachtung ist gegeben durch die Betaverteilung mit den Parametern $a+k$ und $b+N-k$. Die Verteilungsdichte hat sich durch die Beobachtung folgendermaßen verändert:

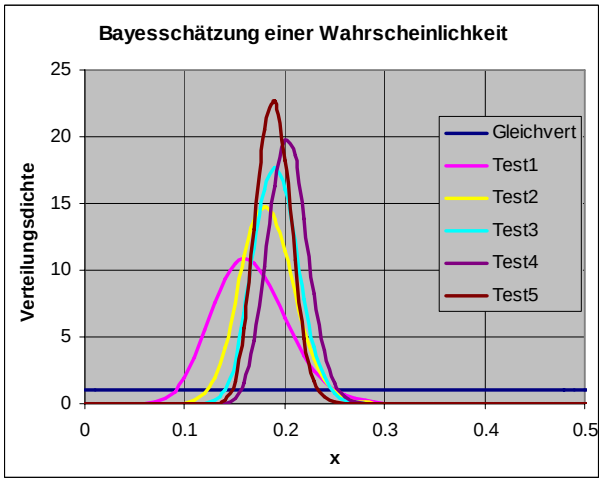
$$f_{a,b}(x) \rightarrow f_{a+k,b+N-k}.$$

Das ist die Verallgemeinerung der Formel (*) des vorigen Abschnitts (Székely, S. 125 f.).

Ein Simulationsexperiment

Ein Produktionsprozess mit konstanter Fehlerwahrscheinlichkeit

Wir simulieren einen Produktionsprozess, dessen Fehlerwahrscheinlichkeit auf w=0.2 festgelegt ist. Diese Tatsache ist dem Abnehmer des Produkts unbekannt. Mittels Stichproben wird er versuchen, die ihm unbekannte Verteilung der Fehlerwahrscheinlichkeit (hier eine „Einpunktverteilung“) herauszufinden.



Zur Abschätzung der Fehlerwahrscheinlichkeit werden in einer stochastischen Simulation mehrere Stichproben des Umfangs N=100 gezogen. Für jeden dieser Tests wird die Anzahl k der fehlerhaften Exemplare ermittelt. Ausgehend von der Gleichverteilung ergeben sich daraus Schritt für Schritt neue Bayes-Schätzungen der Verteilungsdichte des Fehlers. Das Diagramm zeigt das Ergebnis einer solchen Simulation mit dem Tabellenkalkulationsblatt Bayes.xls.

Dieses Experiment widerspricht insofern dem bayesschen Ansatz, als der Parameter w in diesem Computerexperiment eigentlich gar nicht zufällig ist. Wir haben den Parameter für jede Stichprobe gleich gewählt. Aber dieser deterministische Sonderfall kann noch als „Einpunktverteilung“ durchgehen. In der Tat wird sich die Beta-Verteilung aufgrund weiterer Tests auf den Punkt w „zusammenziehen“.

Ein paradoxes Ergebnis: Mehr Daten führen zu schlechteren Folgerungen

Das paradoxe an der Angelegenheit ist, dass das Schätzverfahren gerade für Fälle nicht funktioniert, für die es gedacht ist. Die Bayes-Schätzung kann unter Umständen mit jeder weiteren Beobachtung sogar schlechter werden anstatt besser. Um das einzusehen, ändern wir das Simulationsmodell. Vorbild ist das Münzwurf-Problem von Papoulis (1965, S. 78-81).

Der Produktionsprozess möge nun Lose mit unterschiedlichen Fehlerwahrscheinlichkeiten liefern. Und dieser Parameter sei für jedes Los die Realisierung einer Zufallsvariablen, die auf dem Intervall [0, 0.3] gleichverteilt ist. Die Tests werden für verschiedene Lose durchgeführt.

Auch in diesem Fall wird die Verteilungsdichte mit jedem weiteren Test und mit jeder weiteren Beobachtung immer schmaler. Die Verteilung wird sich schließlich auf eine kleine Umgebung des Wertes 0.15 konzentrieren. Jedenfalls hat die Schätzung immer weniger mit der eigentlich zu findenden Gleichverteilung auf dem Intervall [0, 0.3] zu tun. Es ist so, wie bereits in ähnlich gelagerten Fällen festgestellt wurde: Mehr Daten führen zu schlechteren Folgerungen (Székely, 1990, S. 126).

Man könnte einwenden, dass man die Bayes-Schätzung nur anwenden darf, wenn der Parameter p eigentlich als konstant anzusehen ist, und dass die Annahme einer Verteilung für p nur unser Unwissen über den Parameter beschreiben soll.

Ein derartiger „Gewährleistungsausschluss“ ist jedoch eine Forderung, die sich nur von den erzielten Resultaten her und nicht etwa aus der Herleitung des Verfahrens begründen lässt: „Mit der Angabe einer Verteilung des Parameters ist hier keine Behauptung oder Annahme über die ‚Natur’ des Parameters verbunden, etwa darüber, ob der Parameter vom Zufall abhängt (eine Zufallsgröße ist) oder eine deterministische Größe darstellt, ja nicht einmal darüber, ob der Parameter überhaupt verschiedene Werte annehmen, also ‚schwanken’ kann, oder einen eindeutig feststehenden, unbekannten Wert besitzt, ein Standpunkt, der innerhalb der Bayes-Inferenz vorherrscht.“ (Rüger, 1999, S. 188)

Ein allgemeiner Ansatz für Parameterschätzungen

Sei $f_\alpha(x)$ die vom Parameter α abhängige Verteilungsdichte einer Zufallsvariablen X. Für den Parameter haben wir eine Anfangsverteilung, ein Apriori, mit der Verteilungsdichte $g(\alpha)$. Sei x ein gemessener Wert, eine Realisierung der Zufallsvariablen X. Dann ergibt sich die A-posteriori-Verteilungsdichte $h(\alpha)$ zu

$$h(\alpha) = \frac{f_\alpha(x) \cdot g(\alpha)}{\int f_\alpha(x) \cdot g(\alpha) \cdot \mathrm{d}\alpha} = \frac{f_\alpha(x) \cdot g(\alpha)}{f(x)}.$$

Um den Zusammenhang mit der bayesschen Formel besser erkennen zu können, werden die Dichten mit den Differenzialen multipliziert und diese Ausdrücke den Hypothesen- und Beobachtungswahrscheinlichkeiten zugeordnet:

$P(E) = f(x)\cdot \mathrm{d}x$ ist die Wahrscheinlichkeit dafür, dass der beobachtete Wert in ein (kleines) Intervall der Länge $\mathrm{d}x$ fällt, das den Wert x enthält. Das ist die anfängliche Wahrscheinlichkeit der Beobachtung.

$P(E|H) = f_\alpha(x)\cdot \mathrm{d}x$ ist die Wahrscheinlichkeit, dass der beobachtete Wert x in das eben beschriebene Intervall fällt unter der Bedingung, dass der Parameter (in etwa) gleich α ist.

$P(H)=g(\alpha)\cdot \mathrm{d}\alpha$ ist die A-priori-Wahrscheinlichkeit dafür, dass der Parameter in ein (kleines) Intervall der Länge $\mathrm{d}\alpha$ fällt, das den Wert α enthält.

$P(H|E)=h(\alpha)\cdot \mathrm{d}\alpha$ ist die entsprechende A-posteriori-Wahrscheinlichkeit.

Die obige Formel folgt dann direkt aus der Formel von Bayes.

Dass die A-posteriori-Wahrscheinlichkeit als eine gegenüber dem Apriori verbesserte Schätzung anzusehen ist, formuliert Rüger (1999) als *drittes Bayes-Postulat:* „Das nach der Beobachtung *x* vorhandene Wissen über den Parameter wird durch die nach der Bayes-Formel bestimmte bedingte Verteilung des Parameters unter der Bedingung *X*=*x*, der sogenannten *a posteriori* Verteilung … wiedergegeben." (S. 186)

„Das dritte Postulat ist *pragmatisch* orientiert". Es beschreibt die Bayes-Schätzung als ein „*Dazulernen aus Beobachtungen"* (S. 188).

Anwendung: Zuverlässigkeitsschätzung

Schätzen der Verteilungsfunktion

Untersuchungsgegenstand ist ein System mit einer exponentialverteilten Zeit bis zum Versagen. Diese Versagensrate λ ist der zu bestimmende Parameter. In der Zuverlässigkeitstheorie wird dieser Fall üblicherweise so abgehandelt, dass man den Parameter als zwar unbekannt aber immerhin fest annimmt (Grams, 2001 ff., Abschnitt 4.2). Hier wird ein anderer Standpunkt vertreten: Die Versagensrate ist eine Zufallsvariable, deren Verteilungsfunktion möglichst genau zu ermitteln ist.

Trotz des unterschiedlichen Standpunkts werden wir auf denselben Apparat von Formeln und Verteilungen treffen wie in der Zuverlässigkeitstheorie. Das ist ein schönes Beispiel dafür, dass es nicht allein auf die Formeln ankommt, sondern auch auf den Bezugsrahmen. Nur so gelingt es, die Formeln richtig zu interpretieren.

Für die Verteilung dieser Versagensrate wählen wir, ausgehend von der Funktion

$$L_n(x) = 1 - \sum_{i=0}^{n-1} \frac{x^i}{i!} e^{-x} = \int_0^x \frac{u^{n-1}}{(n-1)!} \cdot e^{-u} \mathrm{d}u$$

als Apriori die Dichte

$$g_{a,n}(\lambda) = \mathrm{d}L_n(a\lambda)/\mathrm{d}\lambda = \frac{a^n}{(n-1)!} \lambda^{n-1} e^{-a\lambda} . \qquad (**)$$

Das ist eine Gammaverteilung mit den Parametern *a* (reellwertig) und *n* (ganzzahlig). Dieser Sonderfall der Gammaverteilung ist auch als Erlangverteilung mit dem Parameter *a* und *n* bekannt. Sie tritt im Zusammenhang mit der Beschreibung von Poisson-Strömen auf.

Wir setzen $0<a$ und $0<n$ voraus. Der Mittelwert μ der Zufallsvariablen dieser Verteilung ist gegeben durch

$$\mu = \frac{n}{a}.$$

Die Verteilungsdichte f_λ der Zufallsvariablen (Zeit bis zum Versagen) ist gegeben durch

$$f_\lambda(t) = \lambda e^{-\lambda t}.$$

Nun wird nach der Zeit *T* ein Versagen beobachtet. Wie lässt sich die Verteilung des Parameters λ aufgrund dieser Beobachtung verbessern? Die Formel des vorigen Abschnitts lässt sich für den hier betrachteten Fall so umformulieren:

$$h(\lambda) = \frac{f_\lambda(T) \cdot g_{a,n}(\lambda)}{\int_0^\infty f_\lambda(T) \cdot g_{a,n}(\lambda) \cdot \mathrm{d}\lambda} = \frac{f_\lambda(T) \cdot g_{a,n}(\lambda)}{f(T)}.$$

Der Zähler ist gleich

$$f_\lambda(T)\cdot g_{a,n}(\lambda)=\frac{a^n}{(n-1)!}\lambda^n e^{-(a+T)\lambda}$$

und für den Nenner ergibt sich der Wert

$$f(T)=n\frac{a^n}{(a+T)^{n+1}}.$$

Also ist die A-posteriori-Verteilungsdichte des Parameters λ gleich

$$h(\lambda)=\frac{(a+T)^{n+1}}{n!}\lambda^n e^{-(a+T)\lambda}=g_{a+T,n+1}(\lambda).$$

Wir erhalten also wieder die Verteilungsdichte einer Gammaverteilung. Die Parameter der Verteilung und die Mittelwerte verändern sich durch die Beobachtung folgendermaßen:

$$(a,\ n)\rightarrow(a+T,\ n+1).$$

$$\frac{n}{a}\rightarrow\frac{n+1}{a+T}.$$

Wir betrachten nun die Zeit bis zum n-ten Versagen. Die Versagensabstände mögen nacheinander t_1, t_2, t_3, …, t_n sein. Die Zeit bis zum n-ten Versagen, die akkumulierten Versagensabstände, nennen wir t. Es ist also $t = t_1+t_2+\ldots+t_n$. Als Apriori wählen wir die Gammaverteilung mit den Parametern t_1 und 1. Die anderen Werte dienen der Verbesserung. Damit erhalten wir schließlich für die Verteilungdichte des Parameters λ eine Gammaverteilung mit den Parametern t und n, also: $g_{t,n}(\lambda)$. Die A-posteriori-Verteilungsfunktion des Parameters λ ist gleich $L_n(\lambda t)$.

Fiduzial- und Konfidenzintervalle

Wir wollen ein Intervall bestimmen, das den zufälligen Parameter λ mit der Wahrscheinlichkeit von, sagen wir, 95% einschließt. Dieses Intervall ergibt sich aus der A-posteriori-Verteilungsfunktion folgendermaßen: Wir bestimmen die Zahlen u und o aus den Formeln $L_n(u) = 2.5\%$ und $L_n(o) = 97.5\%$. Ein Intervall mit der gesuchten Eigenschaft ist $[u/t,\ o/t]$. Das so bestimmte Intervall wird *Fiduzialintervall* genannt (Gladitz, 1994).

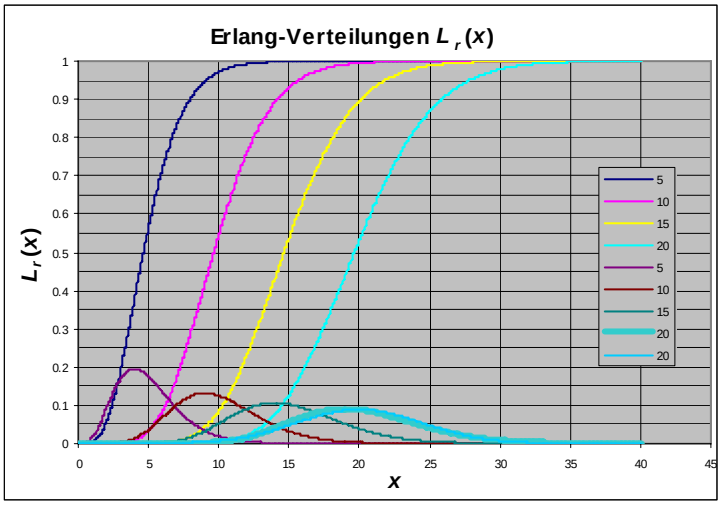
Damit schließt sich die bayessche Betrachtungsweise nahtlos an die Zuverlässigkeitsschätzung mit Konfidenzintervallen an (Grams, 2001 ff., S. 27 ff.). Nach denselben Formeln, die zur Berechnung des Fiduzialintervalls dienen, wird dort das *Konfidenzintervall* (Vertrauensintervall) ermittelt. In beiden Fällen ist eine bestimmte Aussagesicherheit vorauszusetzen, hier beträgt sie beispielsweise 95%.

Die Intervalle sind Zufallsergebnisse, und sie werden nach denselben Formeln errechnet. Dennoch sind die Unterschiede zu beachten. Sie betreffen die Interpretation: Ein Fiduzialintervall schließt den als *zufällig* gedachten Parameter mit der vorgegebenen Sicherheit ein. Dagegen schließen die Vertrauensintervalle den zwar unbekannten aber *festen* Parameter mit der vorgegebenen Sicherheit ein.

Zahlenbeispiel

Bei einem Programm wurden – bei konstantem Operationsprofil – bisher folgende fünfzehn Versagensabstände (in Sekunden) gemessen: 3929, 1441, 591, 351, 64, 150, 743, 705, 33, 601, 1096, 4008, 423, 273, 74. Die akkumulierten Versagensabstände ergeben sich zu t = 14482 s ≈ 4 h.

Gesucht ist für die Versagensrate λ das Fiduzialintervall zur Sicherheit von 90%.



Die Grafik zeigt die normierten Verteilungsfunktionen $L_r(x)$ der Erlang-Verteilung zu den Parametern r = 5, 10, 15, 20. Wir setzen $L_{15}(u)$ = 5% und $L_{15}(o)$ = 95% und entnehmen der Grafik die Werte $u \approx 9$ und $o \approx 22$. Folglich gilt für die Versagensrate näherungsweise die Wahrscheinlichkeitsaussage $P(2.25/\text{h} \leq \lambda \leq 5.5/\text{h}) = 90\%$.

Mängel der Bayes-Schätzung von Parametern

Die folgende Kritik betrifft nicht die die Bewertung der Effizienz diagnostischer Tests mittels Bayes-Schätzung sondern nur die Parameterschätzung nach dem Bayes-Verfahren.

Schwachpunkt Apriori

Die Unterschiede in der Interpretation von Fiduzial- und Konfidenzintervallen scheinen nur eine Frage der „Philosophie“ zu sein: Im ersten Fall wird der Parameter als zufällig und im zweiten als fest angesehen. Im Falle der Zuverlässigkeitsschätzung ergeben sich auf beiden Wegen dieselben numerischen Ergebnisse.

Aber Vorsicht: Beide Parameter-Schätzungen arbeiten mit Annahmen. Eine Annahme ist ihnen gemeinsam: Die Zeit bis zum Versagen wird als exponentialverteilt angesehen. Bei der Schätzung des Fiduzialintervalls nach dem Bayes-Verfahren sind darüber hinaus weitere Annahmen erforderlich. Die A-posteriori-Verteilung des Parameters hängt von einer *willkürlich angenommenen Anfangsverteilung* (**) ab. Eine andere Wahl des Apriori hätte zu anderen Ergebnissen geführt. Auch Anfangsschätzungen nach dem Indifferenzprinzip sind unter Statistikern äußerst umstritten (Székely, 1990, S. 109).

Im Werk von Lyu finden wir das Zuverlässigkeitswachstumsmodell von Littlewood, das auf Bayes-Schätzungen beruht. Auch in diesem Zusammenhang wird das Apriori als kritischer Punkt erkannt: „Die A-priori-Verteilung, die die Modellparameter aus Sicht der Daten der Vergangenheit widerspiegelt, ist ein zentraler Teil dieser Methode. Sie bringt den Standpunkt zum Ausdruck, dass man die Informationen der Vergangenheit, unter anderem aus vergleichbaren Projekten, in die Schätzung der gegenwärtigen und der zukünftigen Zuverlässigkeitsdaten einbringen sollte. Diese Verteilung ist gleichzeitig eine der Stärken und eine der Schwächen der bayesschen Methode. Man sollte die Vergangenheit einbringen, aber *wie*, das ist die Frage.“ (Lyu, 1996, S. 104)

Praxisfremd

“Wie wir gesehen haben, hängt das Ergebnis von der [A-priori-Verteilungsdichte] ab, die im Allgemeinen unbekannt ist. Aber selbst wenn wir sie als bekannt voraussetzen, könnte das Ergebnis für uns wertlos sein… Da es uns eigentlich nicht interessiert, was geschähe, wenn

wir mit mehreren Münzen experimentierten, dürfen wir [die Wahrscheinlichkeit *p* des Ereignisses ‚Kopf'] nicht als Zufallsvariable betrachten, sondern als einen unbekannten Parameter. Die Schlussfolgerungen [aus dem Theorem von Bayes] sind deshalb für uns wertlos." (Papoulis, 1965, S. 112-114)

Paradoxe Ergebnisse: Mehr Daten – schlechtere Folgerungen

Beim Bayes-Verfahren bleibt jedenfalls die Frage offen, wie genau die A-posteriori-Verteilung die tatsächliche Verteilung darstellt. An einem Beispiel wurde gezeigt, dass die geschätzte Verteilung sich mit jedem Korrekturschritt sogar immer weiter von der tatsächlichen Verteilung entfernen kann. Wem das Beispiel nicht reicht, der findet im Buch von Székely (1990, S. 76-79, 108-112, 125-127) mehr davon. Eine Analyse der Ursachen enthält meine Sammlung der *Denkfallen und Paradoxa* unter dem Stichwort Bayes-Schätzung.

Literaturverzeichnis

Carnap, Rudolf; Stegmüller, Wolfgang: Induktive Logik und Wahrscheinlichkeit. Springer, Wien 1959

Fisz, Marek: Wahrscheinlichkeitsrechnung und mathematische Statistik. DVW Berlin 1976

Gladitz, Johannes: Fiduzialintervalle für den Parameter der Binomialverteilung mit SPSS 6.0 für Windows. RZ-Mitteilungen der Humboldt-Universität zu Berlin Nr. 9 (Dezember 1994), S. 21-26 (Erreichbar über den edoc-Server der Humboldt-Universtät zu Berlin. Artikel aus dem cms-journal)

Grams, Timm: Denkfallen und Paradoxa. (http://www2.hs-fulda.de/~grams/dnkfln.htm)

Grams, Timm: Grundlagen des Qualitäts- und Risikomanagements. Zuverlässigkeit, Sicherheit, Bedienbarkeit. 2001 ff. (http://www2.hs-fulda.de/~grams/Q&R/Q&R-v4.pdf)

Hell, Wolfgang; Fiedler, Klaus; Gigerenzer, Gerd (Hrsg.): Kognitive Täuschungen. Spektrum Akademischer Verlag, Heidelberg, Berlin, Oxford 1993

Lyu, Michael R. (edt.): Handbook of Software Reliability Engineering. McGraw-Hill, New York 1996

Papoulis, Athanasios: Probability, Random Variables, and Stochastic Processes. McGraw-Hill, New York 1965

Pólya, Georg: Mathematik und plausibles Schließen. Band 2: Typen und Strukturen plausibler Folgerung. Birkhäuser, Basel 1963

Rüger, Bernhard: Test- und Schätztheorie. Band I: Grundlagen. Oldenbourg, München 1999

Sachs, Lothar: Angewandte Statistik. Anwendung statistischer Methoden. Springer, Berlin, Heidelberg 1992

Székely, Gábor J.: Paradoxa. Klassische und neue Überraschungen aus Wahrscheinlichkeitsrechnung und mathematischer Statistik. Harri Deutsch, Thun, Frankfurt am Main, 1990