

Anmerkungen zur Schätztheorie

Jürgen Grieser

23.6.1998

1 Motivation

Bei unseren statistischen Berechnungen schätzen wir öfter als uns bewußt ist. So "berechnen" wir den Mittelwert von Zeitreihen und fassen diesen Stichprobenmittelwert als Mittelwert der zu untersuchenden Variable (Grundgesamtheit) auf. Man kann zeigen, daß dieses Momentenschätzverfahren in einem gewissen Sinn gut sein kann. Da es aber durchaus Fälle gibt, in denen es ganz extrem schlecht sein kann, sollte man sich bewußt machen, was man tut, wenn man berechnete Kenngrößen von Zeitreihen als Eigenschaften der zugrunde liegenden Variablen interpretiert. Dazu ein Beispiel:

Um zu testen, ob ein Residuum weitere Information enthält, testet man es auf Gaußverteilung. Dazu paßt man zunächst eine Gaußverteilung an, indem man die Stichprobenvarianz und den Stichprobenmittelwert als Parameter der anzupassenden Gaußverteilung verwendet. In einem zweiten Schritt, testet man (z.B. mit dem χ^2 - oder dem Kolmogoroff-Smirnoff-Test) ob die angepaßte Gaußverteilung gut paßt. Wenn dem so ist, ist man fertig. Nun könnte aber eine andere ähnliche Verteilung auch gut passen. Das wäre nicht schlimm, wenn nicht verschiedene (auch offensichtlich ähnliche) Verteilungen auf ganz unterschiedlichen Modellvorstellungen beruhen könnten, die zu ganz unterschiedlichen Interpretationen Anlaß geben. So sehen z.B. die Ableitung der Fermi-Funktion (FV) und die Cauchy-Verteilung (CV) der Gauß-Verteilung (GV) sehr ähnlich. Während aber die Gauß-Verteilung aus der Summe vieler Zufallsvariablen resultiert (zentraler Grenzwertsatz) und damit als reines additives Rauschen angesehen werden kann, ist die Cauchy-Verteilung das Verhältnis aus zwei Gauß-verteilten Variablen. Der Unterschied wird darin deutlich, daß der Mittelwert einer Realisation einer (normierten) Gauß-verteilten Variable gegen den Mittelwert der Grundgesamtheit der Variablen (und der ist 0) konvergiert, während der Mittelwert einer Cauchy-verteilten Variablen divergiert, d.h. mit zunehmender Stichprobenlänge immer größer wird. Der Erwartungswert der Cauchy-Verteilung ist unendlich. Die Fermi-Verteilung, die der Gauß-Verteilung sehr ähnlich sieht, hat den Vorteil, daß sie analytisch integrierbar ist. Durch diese Eigenschaft ist sie für viele praktische Anwendungen der Gauß-Verteilung überlegen. Als Fazit dieser Ausführungen bleibt festzuhalten, daß es sich lohnt, sich mit der Frage zu beschäftigen, wie man Parameter und damit Modellvorstellungen schätzt.

2 Schätzverfahren

2.1 Eigenschaften von Schätzern

Der Schätzer $\hat{\Theta}$ eines Parameters Θ wird nun als eine Funktion der beobachteten Stichprobe $\Theta(x_1, \dots, x_n)$ aufgefaßt. Man erwartet von Schätzern, daß sie möglichst *gut* sind. *Gut* muß aber irgendwie gemessen werden. Um dies zu tun werden die folgenden Eigenschaften von Schätzern definiert:

- Erwartungstreue (Unverzerrtheit)

Ein Schätzer ist erwartungstreu, falls dessen Erwartungswert $E(\hat{\Theta})$ gleich dem zu schätzenden Parameter ist. Das bedeutet, daß man im Mittel, wenn man viele Stichproben verwendet, den Parameter richtig schätzt. Als Bias (bzw. Verzerrtheit) bezeichnet man die Größe $B(\hat{\Theta}) = E(\hat{\Theta}) - \Theta$. So unterschätzt der Schätzer $\hat{\sigma}^2 = s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ die Varianz der der Stichprobe zugrunde liegenden Grundgesamtheit. $\hat{\sigma}^2 = \frac{n}{n-1} s^2$ ist demhingegen ein unverzerrter Schätzer der Grundgesamtheitsvarianz σ^2 aus der Stichprobenvarianz s^2 . Falls die Eigenschaft der Erwartungstreue nur im Grenzfall unendlich langer Stichproben gilt, spricht man von asymptotischer Erwartungstreue. So ist der Schätzer $\hat{\sigma}^2 = s^2$ zumindest asymptotisch erwartungstreu.

- Mediantreue

Eine Schätzfunktion heißt mediantreu, wenn sie den zu schätzenden Parameter in der Hälfte der Fälle unter- in der anderen Hälfte überschätzt. Die Wahrscheinlichkeit einer Überschätzung des Parameters ist dann ebenso groß wie die einer Unterschätzung.

- Effizienz

Ein erwartungstreuer Schätzer Θ_1 ist effizienter (wirksamer) als ein anderer erwartungstreuer Schätzer Θ_2 , wenn dessen Varianz $V(\hat{\Theta}_1)$ kleiner ist. D.h. falls $\eta = \frac{V(\hat{\Theta}_1)}{V(\hat{\Theta}_2)} = \frac{E[(\hat{\Theta}_1 - E(\hat{\Theta}_1))^2]}{E[(\hat{\Theta}_2 - E(\hat{\Theta}_2))^2]} < 1$ ist. Ein Schätzer heißt absolut effizient (am wirksamsten), falls kein anderer Schätzer effizienter ist.

- Konsistenz

Ein Schätzer heißt konsistent, wenn er für unendlich lange Stichproben nicht mehr vom wahren Wert abweicht, d.h. wenn gilt: $\lim_{n \rightarrow \infty} P(|\hat{\Theta}_n - \Theta| > \varepsilon) = 0$, für alle ε .

- Suffizienz

Eine Schätzfunktion ist suffizient (erschöpfend), wenn sie die maximal mögliche Information der Stichprobe nutzt.

- Normalität

Eine Schätzfunktion heißt normal, wenn sie Gauß-verteilt ist, d.h. wenn gilt $\frac{\hat{\Theta} - E(\hat{\Theta})}{\sqrt{V(\hat{\Theta})}} = N(0, 1)$, mit der Varianz des Schätzers $V(\hat{\Theta})$.

- Linearität

Eine Schätzfunktion ist linear, wenn gilt $\hat{\Theta} = a_0 + \sum_{i=1}^n a_i x_i$.

Diese einzelnen Eigenschaften können auch zusammengesetzt werden. So ist z.B. ein

- *gleichmäßig bester unverzerrter Schätzer* erwartungstreu und am effizientesten,
- *bester linearer unverzerrter Schätzer* erwartungstreu, linear und am effizientesten unter den linearen erwartungstreuen Schätzern (das ist der Blue-Schätzer, von Best Linear Unbiased Estimator),
- *bester asymptotisch normaler Schätzer* asymptotisch normalverteilt und besitzt die kleinst mögliche Varianz.

Mit den nun zur Verfügung stehenden Maßen können einige Schätzverfahren vorgestellt und verglichen werden.

2.2 Momentenverfahren

Bei den Momentenschätzern (gehen auf Pearson (1857 - 1936) zurück) werden die Grundgesamtheitsmomente gleich den Stichprobenmomenten gesetzt. Das bedeutet z.B. daß das Grundgesamtheitsmittel gleich dem Stichprobenmittel gesetzt wird und die Grundgesamtheitsvarianz gleich der Stichprobenvarianz. Alle Momentenschätzer sind

- immer konsistent,
- zumindest asymptotisch erwartungstreu,
- in der Regel asymptotisch normal,
- oft nicht wirksamst und
- oft nicht suffizient.

So ist die Momentenschätzung $\hat{\sigma}^2 = s^2$ für die Varianz der Grundgesamtheit aus der Stichprobenvarianz s^2 (also dem zweiten zentralen Moment) nur asymptotisch erwartungstreu.

2.3 Methode der kleinsten Quadrate

Diese auf Gauß und Laplace zurückgehende Methode sieht die Stichprobe als Summe einer Funktion $f(\vec{\Theta})$ des Parametervektors $\vec{\Theta}$ plus Rauschen an. Dabei muß der Stichprobenumfang größer sein, als die Dimension des Parametervektors. $\vec{\Theta}$ wird nun aus der Stichprobe so geschätzt, das das Rauschen minimiert wird. Dazu wird zunächst die Summe der Abstandsquadrate $S = \sum_{i=1}^n \left[y_i - f\left(\hat{\vec{\Theta}}\right) \right]^2$ gebildet und dann minimiert, indem man dessen Ableitung nach dem Schätzparametervektor $\hat{\vec{\Theta}}$ null setzt. Daraus folgt $\sum_{i=1}^n \left[y_i - f\left(\hat{\vec{\Theta}}\right) \right] \frac{\partial \hat{\vec{\Theta}}}{\partial \hat{\Theta}_i} = 0$ für alle $i = 1, \dots, m$, wenn der Parametervektor m -dimensional ist.

2.4 Maximum-Likelihood-Schätzer

Maximum-Likelihood-Schätzer (ML-Schätzer, nach Fisher, 1890 - 1962) basieren auf der Annahme, daß wahrscheinlich das Wahrscheinlichste wahr ist. Das klingt hoffentlich schon so, daß der kritische Leser, auf den Gedanken kommen könnte, daß die damit erzielten Ergebnisse auch total falsch sein können. Trotzdem neigt man auch im Alltag gerne dazu, nach diesem Prinzip zu schätzen. Weiß man von einem Koch z.B. daß er wesentlich wahrscheinlicher eine Suppe versalzt, wenn er verliebt ist, als wenn er es nicht ist, so wird man sich während dem Essen einer versalzenen Suppe denken: "wahrscheinlich ist der Koch (mal wieder) verliebt". Genau das ist eine ML-Schätzung.

Um ML-Schätzungen durchführen zu können, braucht man eine Likelihood-Funktion $L(\Theta; \vec{x})$, die angibt, welcher Wert des gesuchten Parameters Θ bei der vorgegebenen Stichprobe $\vec{x} = (x_1, x_2, \dots, x_n)$ wie wahrscheinlich ist. Das Maximum dieser Funktion in Abhängigkeit von Θ bei gegebener Stichprobe \vec{x} ist der Wert der ML-Schätzung. Um die Likelihood-Funktion zu erzeugen, entscheidet man sich aufgrund seiner Erfahrung (oder ähnlich überzeugender Argumente) dafür, daß die Stichprobe aus einem bestimmten Modell stammt. Das Modell kann dann Stichproben erzeugen. Die Likelihood-Funktion ist nun die Wahrscheinlichkeit für eine Realisation \vec{x} in Abhängigkeit von den m Parametern $\vec{\Theta} = \Theta_1, \Theta_2, \dots, \Theta_m$ des angenommenen Modells. Die Likelihood-Funktion braucht dann nur noch uminterpretiert zu werden. Man betrachtet nicht die Wahrscheinlichkeit für die x_i in Abhängigkeit von $\hat{\Theta}$, sondern $\hat{\Theta}$ als Variablen und die x_i als Parameter.

Ein Beispiel soll das verdeutlichen. Als Modell für die Stichprobe wird angenommen, sie stamme aus Gauß'schem weißen Rauschen. Dann ist jeder einzelne Wert von x_i unabhängig von den anderen Werten von x_j . Alle Werte sind Gauß-verteilt und stammen somit aus einer Verteilung mit den zwei Parametern μ und σ^2 . Wegen (und nur wegen) der Unabhängigkeit der Einzelereignisse kann die Likelihood-Funktion faktorisiert werden:

$$L(\mu, \sigma^2; x_1, x_2, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right). \quad (1)$$

Daraus folgt

$$L(\mu, \sigma^2; \vec{x}) = \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right). \quad (2)$$

Man könnte diese Funktion nun nach μ und nach σ^2 ableiten, die Ableitungen jeweils nullsetzen und damit die beiden Parameter bestimmen. Dies wäre recht aufwendig. Und da das nicht das einzige Beispiel ist, bei dem das recht aufwendig ist, geht man prinzipiell etwas anders vor. Man logarithmiert die Likelihood-Funktion vor dem ableiten und kommt somit zur Loglikelihood-Funktion $\ln L$:

$$\ln L(\mu, \sigma^2; \vec{x}) = \frac{-n}{2} \ln(2\pi) - n \ln \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2, \quad (3)$$

mit den partiellen Ableitungen

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) \\ \frac{\partial \ln L}{\partial \sigma^2} &= \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

Die Lösung dieses Gleichungssystems ist überraschenderweise der Momentenschätzer. Maximum-Likelihood-Schätzer sind

- konsistent
- zumindest asymptotisch erwartungstreu
- zumindest asymptotisch wirksamst
- suffizient
- bester asymptotisch normaler Schätzer.

Gerade die letzte Eigenschaft ist von besonderer Bedeutung. Sie erlaubt es nämlich nicht nur den wahrscheinlichsten Wert für den Schätzer anzugeben, sondern auch dessen Verteilung (zumindest asymptotisch). Kennt man aber die Verteilung des Schätzers, so kann man auch Gleichungen für andere Parameter als den Erwartungswert herleiten. So kann z.B. die Varianz des Erwartungswertes $V(\Theta)$ geschätzt werden. Damit lassen sich Konfidenzintervalle angeben, in denen der geschätzte Parameter mit einer vorgegebenen Wahrscheinlichkeit liegt.

Abschließend sei noch bemerkt, daß die Likelihood-Funktion keine Wahrscheinlichkeitsdichte ist. Das ist sie nur, wenn sie normiert wird. Dann müßte man meines Erachtens aus dieser Likelihood-Funktion die gleiche Information entnehmen können, wie aus einer Bayes'schen Vorgehensweise.

2.5 Bayes-Schätzer

Bayes-Schätzer basieren auf der Bayes'schen Formel (siehe Anhang). Dabei wird nicht nur der Erwartungswert eines Parameters geschätzt, sondern dieser als Zufallsvariable aufgefaßt und deren Verteilung geschätzt. Kennt man die Verteilung, so hat man alle verfügbare Information über den Parameter. Setzt man in Gleichung (21) für die Zufallsvariable X den Parameter Θ ein und für Y die n -dimensionale Stichprobe \vec{Y} , so folgt

$$f(\theta|\vec{y}) = \frac{f(\vec{y}|\theta) f(\theta)}{\int f(\vec{y}|\theta) f(\theta) d\theta}. \quad (4)$$

Mit dieser Gleichung kann man aus einem Prior-Schätzer $f(\theta)$ (der z.B. mehr oder weniger gut geraten ist) über eine sogenannte Likelihood-Funktion in Abhängigkeit von den Daten auf einen (womöglich aber nicht notwendigerweise) besseren Posterior-Schätzer $f(\theta|\vec{y})$ schließen. Dabei fällt auf, daß man dieses Verfahren sukzessive anwenden kann, in der Hoffnung, daß es möglichst rasch und gegen einen richtigen Posterior konvergiert. Andererseits muß man dann aber auch recht viele Integrale ausführen. Der Bayes-Schätzer für den gesuchten Parameter ist dann der Erwartungswert der Posterior-Verteilung und ist gegeben durch das Integral

$$\hat{\Theta} = E(\vec{\Theta}) = \int \vec{\theta} f(\vec{\theta}|\vec{y}) d\vec{\theta}. \quad (5)$$

Falls Θ ein Parametervektor ist, d.h. falls man mehrere Parameter zu schätzen hat, steht im Nenner von Gleichung (4) ein Mehrfachintegral, daß in der Regel nur noch durch Monte-Carlo-Integration gelöst werden kann.

Bayes-Schätzer können nicht nur dazu verwendet werden, um Parameterwertverteilungen bei vorgegebenen Daten und angenommener Modellvorstellung zu schätzen, sondern man kann auch Wahrscheinlichkeitsverteilungen für verschiedene Modelltypen untersuchen. Dies führt zu den Markov Chain Monte Carlo Methoden (MCMC). Dabei sind die verschiedenen (vorzugebenden) konkurrierenden Modelle die diskreten Werte des zu schätzenden Parameters (der Parameter ist in diesem Fall die Modellnummer). Da jedes Modell wieder eine gewisse Anzahl von Modellparametern hat, muß bei dieser Anwendung in verschiedenen (aber alternierenden) Stufen vorgegangen werden. So hat man die Modellstufe mit der Gleichung

$$f(\text{Modell} | \text{Daten}) = f(\text{Daten} | \text{Modell}) \frac{f(\text{Modell})}{f(\text{Daten})} \quad (6)$$

und die Parameterstufe mit

$$f(\text{Parameter} | \text{Daten}, \text{Modell}) = f(\text{Daten} | \text{Modell}, \text{Parameter}) \frac{f(\text{Parameter} | \text{Modell})}{f(\text{Daten})}. \quad (7)$$

Die Gleichungen (6) und (7) werden dann in einem Markov-Ketten-Lauf alternierend verwendet, d.h. es wird versucht, das die Daten optimal beschreibende Modell mit der optimalen Parameterkombination zu finden. Eine solche Anwendung könnte man auch weniger

automatisiert (und meines Erachtens weniger aufwendig und klarer durchschaubar) realisieren, indem man einige in Frage kommenden Modellstrukturen auswählt (z.B. einige ARMA-Prozesse), diese alle optimal anpaßt, und dann die Güte miteinander vergleicht. Da dabei Modelle mit mehr Parametern eine höhere Freiheit bei der optimalen Anpassung haben, müssen diese beim Vergleich etwas benachteiligt werden. Dies geschieht unter Berücksichtigung eben dieser Freiheitsgrade mit Hilfe verschiedener Kriterien (z.B. Akaike-Informations-Kriterium, AIC).

3 Ein Beispiel zum Vergleich von ML- und Bayes-Schätzer

Gegeben sei ein Bernoulli-Experiment, d.h. ein Zufallsexperiment mit zwei möglichen Ausgängen. Solche Experimente, für die der Münzwurf Pate stehen muß, sind bei weitem nicht so unrealistisch, wie mancher hoffen mag. Die beiden Ausgänge des Experiments können durch 0 und 1 charakterisiert werden. Die 0 kann dann für das Nichteintreten eines Ereignisses, die 1 für dessen Eintreten stehen. Ereignisse können z.B. Treffer bei der Wettervorhersage sein, Vulkanausbrüche, Tornados, aber auch Über- bzw. Unterschreitungen von Schwellen bei beliebigen kontinuierlichen Variablen sein. Aus der Beobachtung solcher Variablen möchte man dann die Eintrittswahrscheinlichkeit p für das Ereignis 1 schätzen. Für die Wahrscheinlichkeitsdichten muß $f(0) + f(1) = 1$ gelten. Mit $f(1) = p$ folgt daraus sofort $f(0) = 1 - p$. Weiter gilt aber auch $1 = p^0 = (1 - p)^0$. Damit kann man die sogenannte Bernoulli-Verteilung definieren:

$$f(x) = p^x (1 - p)^{1-x}, \text{ für } x = 0, 1. \quad (8)$$

Da diese Verteilung von dem (einen) Parameter p abhängt, wird eine Realisation x_1 mit der Wahrscheinlichkeit $f(x|p)$ den Wert x haben. Nun betrachten wir eine Versuchsreihe mit $n = 10$ Realisationen: $(1, 0, 1, 1, 0, 1, 1, 1, 1, 0)$. Daraus soll der Parameter p möglichst *gut* geschätzt werden. Zunächst verwenden wir dazu den ML-Schätzer.

3.1 ML-Schätzer für p

Die Likelihood-Funktion $L(p, x_1, \dots, x_n)$ ist gegeben durch

$$L(p, \vec{x}) = \prod_{i=1}^n f(x_i, p) = \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i} = p^{\sum x_i} (1 - p)^{n - \sum x_i}. \quad (9)$$

Für das konkrete Zahlenbeispiel folgt damit $L = p^7 (1 - p)^3$. Würde man dies maximieren wollen, müßte man die Nullstellen der Ableitung dL/dp suchen. Im konkreten Fall also $dL/dp = 7p^6 (1 - p)^3 - 3p^7 (1 - p)^2 = 0$ lösen. Der Grad des Polynoms, dessen Nullstellen

man zu finden hat steigt linear mit der Anzahl der Beobachtungsdaten. Deshalb leitet man auch hier wieder die Loglikelihood-Funktion ab und sucht deren Nullstellen:

$$\begin{aligned}\frac{d \ln L}{dp} &= \frac{d}{dp} \left[(\ln p) \sum_{i=1}^n x_i + \ln(1-p) \left(n - \sum_{i=1}^n x_i \right) \right] \\ &= \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0.\end{aligned}\tag{10}$$

Die Nullstellen dieser Gleichung lassen sich sehr einfach finden und es folgt

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n}.\tag{11}$$

Für das gewählte Zahlenbeispiel folgt daraus ein Wert von

$$\hat{p}_{ML} = \frac{7}{10}.\tag{12}$$

3.2 Bayes-Schätzer für p

Um den Bayes-Schätzer für p angeben zu können, braucht man zunächst einen Prior. Da man über p keine Information voraussetzen kann, nimmt man (nicht zuletzt auch der Einfachheit halber) $f(p) = 1$ für alle Werte von p zwischen 0 und 1 an. Für den Posterior folgt dann nach Gleichung (4)

$$f(p|\vec{x}) = \frac{f(\vec{x}|p) f(p)}{\int f(\vec{x}|p) f(p) dp}.\tag{13}$$

Dabei ist

$$f(\vec{x}|p) = \prod_{i=1}^n f(x_i|p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{n-\sum x_i} = p^m (1-p)^{n-m},\tag{14}$$

mit $m = \sum x_i$ die Likelihood-Funktion. Für das Integral im Nenner von Gleichung (13) folgt dann

$$\int_0^1 f(\vec{x}|p) f(p) dp = \int_0^1 p^m (1-p)^{n-m} dp = \frac{m!(n-m)!}{(m+n-m+1)!}$$

und damit folgt durch Einsetzen in Gleichung (13)

$$f(p|\vec{X}) = \frac{p^m (1-p)^{n-m} (n+1)!}{m!(n+m)!}.$$

Das ist nun der Posterior, der aus dem Prior $f(p) = 1$ und der Likelihood-Funktion folgt. Er gibt eine Wahrscheinlichkeitsverteilung für den zu bestimmenden Koeffizienten an. Man

sieht, daß man durch eine Normierung der Likelihood-Funktion die gleiche Wahrscheinlichkeitsdichte auch bei der ML-Anwendung erhalten hätte. Das ist aber nur in den seltenen Fällen so, in denen der Prior eine Konstante ist. Nun müssen wir nur noch den Erwartungswert $E(f(p|\vec{x}))$ berechnen, der dann der Bayes-Schätzer für den Parameter ist. (Man beachte, daß man, da man die Wahrscheinlichkeitsdichte kennt, auch jedes beliebige Quantil (also auch Konfidenzintervalle) für den Parameter berechnen kann.)

$$\begin{aligned}
 E(f(p|\vec{x})) &= \int_0^1 p f(p|\vec{x}) dp \\
 &= \frac{(n+1)!}{m!(n-m)!} \int_0^1 p^{m+1} (1-p)^{n-m} dp \\
 &= \frac{(n+1)!}{m!(n-m)!} \frac{(m+1)!(n-m)!}{(m+1+n-m+1)!} \\
 &= \frac{m+1}{n+2} = \frac{1 + \sum_{i=1}^n x_i}{n+2}.
 \end{aligned} \tag{15}$$

Für das gewählte Zahlenbeispiel folgt daraus der Bayes-Schätzer

$$\hat{p}_B = \frac{8}{12} = \frac{20}{30} < \frac{21}{30} = \frac{7}{10} = \hat{p}_{ML}.$$

Dem aufmerksamen Leser müßte jetzt nicht nur aufgefallen sein, *daß* sich diese beiden Schätzer unterscheiden (der Unterschied wird für große n allerdings beliebig klein), sondern auch *warum* sie sich unterscheiden. Während der ML-Schätzer den wahrscheinlichsten Wert, d.h. den mit der höchsten Wahrscheinlichkeitsdichte, also den Modalwert der Wahrscheinlichkeitsverteilung für den Parameterwert auswählt, wählt der Bayes-Schätzer den Erwartungswert.

Anhang: Bayes'sche Formel

Die Bayes'sche Formel (der Bayes'sche Satz, das Bayes'sche Theorem) hat Thomas Bayes (1701-1761 und Schüler von de Moivre) im Jahre 1750 entdeckt. Sie wurde trotzdem erst nach seinem Tod veröffentlicht. Da Bayes sich nicht mehr gegen die Kritik daran wehren konnte, schlimmer noch, nicht für seine Formel werben konnte, wurde sie nicht weiter beachtet, bis sie gegen Ende des 20. Jh im Zusammenhang mit der Parameterschätzung interessant wurde. Seitdem scheinen die Statistiker in zwei Lager gespalten zu sein: den Bayesianern und den Anti-Bayesianern (siehe z.B. Szekely, 1990). Frei von der sich immer weiter vertiefenden Kluft (Szekely, 1990) zwischen diesen beiden Lagern scheinen die deutschen Statistiker zu sein, die zwar in ihren Lehrbüchern auf die Existenz der Bayes'schen Formel aufmerksam machen, weiter aber nichts damit zu tun haben. Da mit Hilfe moderner Großrechenanlagen sehr aufwendige Anwendungen der Bayes'schen Formel möglich werden und da diese auch in der statistischen Klimatologie Anwendung finden, sollte man sich wenigstens grob bewußt machen, wovon Bayesianer reden.

Um die Bayes'sche Formel abzuleiten gehen wir von zwei Zufallsvariablen X und Y aus. Diese können jeweils Werte aus einem bestimmten Wertebereich annehmen. Es existieren dann eindimensionale Wahrscheinlichkeitsdichten dafür, daß X den Wert x annimmt ($p(X = x) = f(x)$) und dafür, daß Y den Wert y annimmt ($p(Y = y) = f(y)$). Zusätzlich kann man die gemeinsame (zweidimensionale) Wahrscheinlichkeitsdichte $p(X = x, Y = y) = f(x, y)$ definieren. Falls es nun zwischen den Zufallsvariablen X und Y keinen Zusammenhang gibt, gilt

$$f(x, y) = f(x) f(y).$$

Andernfalls gilt

$$f(x, y) = f(x) f(y|x) = f(y) f(x|y). \quad (16)$$

In dieser nicht auf Anhieb einsichtigen Gleichung, stellen die Funktionen $f(x|y)$ und $f(y|x)$ bedingte Wahrscheinlichkeiten dar. Dabei bedeutet $f(x|y)$ die Wahrscheinlichkeit dafür, daß die Variable X den Wert x annimmt, wenn die Variable Y den Wert y angenommen hat. $f(y|x)$ ist demnach genau umgekehrt definiert. Also sagt Gleichung (16) aus, daß die Wahrscheinlichkeit dafür, daß X den Wert x und Y den Wert y annimmt ($f(x, y)$), gegeben ist durch die Wahrscheinlichkeit, mit der x realisiert wird ($f(x)$) mal der Wahrscheinlichkeit, daß y realisiert wird, wenn x realisiert wird ($f(y|x)$). Das Gleiche gilt natürlich auch, wenn man x und y vertauscht. Nun kann Gleichung (16) dazu verwendet werden bedingte Wahrscheinlichkeiten zu invertieren, denn es folgt sofort

$$f(x|y) = f(y|x) \frac{f(x)}{f(y)}. \quad (17)$$

$f(y)$ kann noch etwas weiter verändert werden. Im Fall von diskreten Variablen X und Y summiert man den rechten Teil von Gleichung (16) über alle möglichen Werte von x und erhält wegen $\sum_{\forall i} f(y, x_i) = f(y)$

$$f(y) = \sum_{\forall i} f(y|x_i) f(x_i). \quad (18)$$

Diese Gleichung besagt, daß die Wahrscheinlichkeit, daß die Zufallsvariable Y den Wert y annimmt, aus der Summe der bedingten Wahrscheinlichkeiten für $Y = y$ unter der Bedingung $X = x_i$ folgt. Dabei muß die Summe über alle möglichen Realisierungen von X (d.h. alle x_i) laufen. Da Gleichung (18) die vollständige Wahrscheinlichkeit als Summe von Einzelwahrscheinlichkeiten beschreibt, wird sie in der Literatur oft als Satz von der vollständigen Wahrscheinlichkeit eingeführt.

Setzt man nun Gleichung (18) in Gleichung (17) ein, so erhält man die Bayes'sche Formel für diskrete Zufallsvariable:

$$f(x_i|y) = \frac{f(y|x_i) f(x_i)}{\sum_{\forall j} f(y|x_j) f(x_j)}. \quad (19)$$

In analoger Weise kann man für kontinuierliche Wertebereiche den Satz von der vollständigen Wahrscheinlichkeit formulieren

$$f(y) = \int f(y|x) f(x) dx \quad (20)$$

und daraus die Bayes'sche Formel für kontinuierliche Zufallsvariable

$$f(x|y) = \frac{f(y|x) f(x)}{\int f(y|x) f(x) dx}. \quad (21)$$

Manche Autoren (z.B.) Hsu (1996) nennen Gleichung (17) Bayes'sche Regel und die Gleichungen (19) und (21) Bayes'sches Theorem.