Bayessche Statistik

There are 46 656 varieties of Baysians![1]

U. Mortensen
.
25. 07. 2010, Korrektur 17. 07. 2015

Inhaltsverzeichnis

1 Einführung und Überblick **2**

1.1 Inferenzstatistische Ansätze 2

1.2 Das Likelihood-, das Suffizienz- und das Konditionalitätsprinzip . . 7

2 Bayessche Statistik **16**

2.1 Das Prinzip 16

2.1.1 Konfidenz und Kredibilität 19

2.1.2 Test von Hypothesen 20

2.2 Typen von a-priori-Verteilungen 22

2.2.1 Das Indifferenzprinzip 22

2.2.2 Likelihood, Scores und Information 24

2.2.3 Nichtinformative Priori-Verteilungen 28

2.2.4 Konjugierte Priors 29

2.2.5 Uneigentliche Priori-Verteilungen 33

2.2.6 Jeffreys' Prior: 34

2.2.7 Jaynes' Maximale Entropie und Transformationsgruppen . 38

2.2.8 Zur Geschichte der Theorie der Ereignisfolgen 45

2.3 Bayes-Asymptotik 46

3 Anhang **47**

3.1 Signifikanztests für Binomial- und inverse Binomialexperimente . . 47

[1] Good (1971)

1 Einführung und Überblick

1.1 Inferenzstatistische Ansätze

Die Schätzung unbekannter Parameter und der Test von Hypothesen sind zentrale Themen der Statistik. Es lassen sich vier Strategien insbesondere für den Test von Hypothesen angeben:

1. **Bayes-Ansatz:** Man geht von der Definition der bedingten Wahrscheinlichkeit aus, wie sie von Thomas Bayes[2] und unabhängig davon von Simon Laplace (1812) vorgelegt wurde: für zwei zufällige Ereignisse A und B ist die Bedingte Wahrscheinlichkeit des Ereignisses A, gegeben, dass B eingetreten ist, bzw. von B, gegeben dass A eingetreten ist:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}, \quad P(B|A) = \frac{P(A \wedge B)}{P(A)}; \tag{1}$$

   es folgt $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$, so dass man

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)} \text{ bzw. } P(B|A) = P(A|B)\frac{P(B)}{P(A)} \tag{2}$$

   folgern kann. Laplace wandte diese Definition auf physikalische Fragen (Himmelsmechanik) an. Setzt man etwa $A = H$, wobei H für eine Hypothese steht, und $B = D$, wobei D für 'Daten' steht, so erhält man

$$P(H|D) = P(D|H)\frac{P(H)}{P(D)}, \tag{3}$$

   mit

$$P(D) = P(D|H)P(H) + P(D|\neg H)P(\neg H); \tag{4}$$

   dies ist der 'Satz der Totalen Wahrscheinlichkeit', der sich sofort aus der Definition der bedingten Wahrscheinlichkeit ergibt. Dabei ist $P(H)$ die a-priori-Wahrscheinlichkeit der Hypothese H, und $P(H|D)$ ist die a-posteriori-Wahrscheinlichkeit für H, gegeben die Daten D. $P(D|H)$ ist die 'Likelihood' der Daten unter der Bedingung, dass H wahr ist[3].

2. **Fishers Ansatz** Im Laufe des 19-ten Jahrhunderts geriet der Bayessche Ansatz in die Kritik; der Mathematiker George Boole (1854) und der Philosoph John Venn (1866) kritisierten den Ansatz, die Hypothese über die

[2] Thomas Bayes (1702 – 1761), Mathematiker und presbyterianischen Pfarrer. Seine Arbeit *Essay Towards Solving a Problem in the Doctrine of Chances* erschien 1763 posthum.

[3] Der Begriff der Likelihood wurde 1935 von R.A. Fisher eingeführt.

bedingte Wahrscheinlichkeit $P(H|D)$ zu evaluieren, insbesondere wegen der Notwendigkeit, eine a-priori-Wahrscheinlichkeit für H definieren zu müssen, und fanden mit ihrer Kritik insbesondere bei Biologen Gehör, die sich mit Darwins Theorie beschäftigten (vergl. Jaynes (2003; 316)). Zu Beginn des zwanzigsten Jahrhunderts waren es wiederum insbesondere Biologen, die sich der Kritik an Bayes und Laplace anschlossen[4] und – wohl unter dem Einfluß Fishers – ihre Aufmerksamkeit auf die Likelihood $P(D|H)$ fokussierten. Sie bildeten eine, wie Jaynes es formuliert, 'extremely aggressive school', die die Entwicklung der Statistik dominierte und den theoretischen Rahmen schuf, der heute als 'orthodoxe Statistik' bezeichnet wird.

Es sei $\mathbf{X}$ irgendeine Strichprobe und $\mathbf{x}$ die spezielle Stichprobe, die sich bei einer Untersuchung ergeben hat. θ sei ein unbekannter Parameter, $\theta \in \Theta \subseteq \mathbb{R}$, Θ der Parameterraum. $f(\mathbf{X}|\theta)$ sein die Dichtefunktion (oder Wahrscheinlichkeitsfunktion) für $\mathbf{X}$, gegeben θ. Es soll die Hypothese $\theta = \theta_0$ getestet werden. Dazu wird eine Teststatistik $T(\mathbf{X})$ definiert und die Wahrscheinlichkeit

$$p = P(T(\mathbf{X}) > T(\mathbf{x})|H_0) = P_{\theta_0}(T(\mathbf{X}) > T(\mathbf{x})) \tag{5}$$

betrachtet. p heißt das *Signifikanzniveau*. Ist p (der p-Wert) klein, so ist der beobachtete Wert $T(\mathbf{x})$ "groß" relativ zur Gesamtmenge der Werte von T, wenn H_0 gilt. In diesem Sinne kann man sagen, dass die Daten $\mathbf{x}$ unter der Bedingung, dass H_0 gilt, eher unwahrscheinlich sind; je kleiner der Wert von p, desto geringer ist die Wahrscheinlichkeit der Daten unter H_0. Einer oft angewandten Regel entsprechend wird demnach H_0 "verworfen" oder "zurückgewiesen", wenn etwa $p \leq .05$; das Ergebnis der Untersuchung mit dem Resultat $\mathbf{x}$ ist dann *signifikant*.

Während also in (3) die *Wahrscheinlichkeit für die Hypothese* H angegeben wird, wird bei Fisher die *Wahrscheinlichkeit der Daten, gegeben eine Hypothese* (H_0) zur Basis der Evaluation der Hypothese gewählt.

Die in (5) ausgesprochene Regel, dass der p-Wert durch die Wahrscheinlichkeit, dass $T(X)$ größer als der tatsächlich beobachtete Wert $T(x)$ ist, gegeben die Hypothese H_0, ist nicht ohne eine gewisse Rätselhaftigkeit. Hacking (1965; 82) merkt an, dass Fisher nie eine Begründung für seine p-Definition gegeben hat. Lindley (1993) führt sie auf Fishers Begegnung mit Dr. Muriel Bristol zurück, der er eine Tasse Tee anbot und die protestierte, weil er erst den Tee und dann die Milch in die Tasse gab; erst die Milch und dann den Tee schmecke besser, argumentierte sie. Fisher führte den bekannten Test aus: er bot Dr. Bristol in zufälliger Folge 6 Paare von Tassen mit Tee + Milch an, wobei bei dreien erst die Milch und dann der

[4] Weil sie, so Jaynes, die Mathematik Laplaces nicht verstanden.

Tee, bei den drei übrigen erst der Tee und dann die Milch eingegossen worden waren. Dr. Bristol sollte bei jedem Paar entscheiden, bei welcher Tasse eines Paares zuerst die Milch und dann der Tee in die Tasse gegeben worden waren. Die Antwort ist dann entweder korrekt (R = richtig) oder falsch (F = falsch). Antwortet Dr. Bristol korrekt mit der Wahrscheinlichkeit θ und urteilt sie bei jedem Paar unabhängig von den anderen Paaren, so ist die Wahrscheinlichkeit einer bestimmten Folge von Antworten mit $X = k$ korrekten Antworten durch

$$P(X = k|\theta, n) = \theta^k(1 - \theta)^{n-k}$$

gegeben, wobei $n = 6$. Unter H_0 sind die Entscheidungen von Dr. Bristol "zufällig", d.h. es ist $\theta = 1/2$. Unter H_0 hat die resultierende Folge von Entscheidungen die Wahrscheinlichkeit $(1/2)^6 = 1/64 \approx .0156$. Nach Fisher ist nun entweder die Nullhypothese H_0 wahr und ein Ereignis mit kleiner Wahrscheinlichkeit ist aufgetreten, oder die Alternativhypothese H_1 ist wahr (Dr. Bristol kann die Reihenfolge von Tee und Milch unterscheiden). Allerdings hat unter H_0 *jede* Folge von R und F die gleiche Wahrscheinlichkeit von .0156. Fisher argumentierte nun, dass jede Folge mit nur einem F und fünf Rs gegen H_0 spräche, gleich, wo in der Folge das F auftauche, also an erster oder an zweiter oder ... oder an sechster Stelle. Die Wahrscheinlichkeit dass eine solche Folge auftritt, ist dann

$$\binom{6}{1}(1/2)^6 = 6/64 \approx .094.$$

Nun ist $.094 > .05$, und damit ist ein Ergebnis mit nur einem F *nicht signifikant*. Aber auch dieser Ansatz erscheint als nicht zufriedenstellend. Fisher argumentierte nun, dass, wenn ein F in 6 Paaren signifikant gegen H_0 spräche, dann erst recht kein F in sechs Paaren. Dies führt dann auf das in (5) formulierte Kriterium, dass die Wahrscheinlichkeit des gefundenen Werte plus der aller extremeren indikativ für die diskriminatorischen Fähigkeiten Dr. Bristols sein müßten. Dementsprechend ist $p = 6(1/2)^6 + (1/2)^6 = 7(1/2)^6 = .109$, und dieser Wert ist nicht signifikant.

Als rechte Begründung für (5) mag diese Argumentation nicht einleuchten, und auch Lindley akzeptiert sie nicht als solche; Hackings Anmerkung scheint berechtigt. Für viele Anfänger des Studiums statistischer Verfahren ergibt sich hier eine erste Schwierigkeit, denn es ist nicht unmittelbar einsichtig, warum H_0 nach dem Kriterium "$T(\mathbf{x})$ *oder größer*" beurteilt werden soll. Jeffreys (1939/1961, p. 385) hat den Sachverhalt auf knappe Weise auf den Punkt gebracht:

> "What the use of p implies, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure."

Eine Plausibilitätsbetrachtung zum Fisherschen Verfahren könnte darin bestehen, dass ein Teilbereich R des Wertebreichs von $T(X)$ definiert wird derart, dass

$$P(T(X) \in R|H_0) = \alpha.$$

Sicher ist

$$\int_R f_T(X)dX = \alpha.$$

Man wählt einen kleinen Wert für α, etwa $\alpha = .05$. Ist $T(x) \in R$, so ist

$$\int_{T(x)} f_T(X)dX = p \leq \alpha, \quad T(x) \geq \inf R.$$

Man hätte somit eine Motivation für den Fisherschen p-Wert, – aber diese Motivation wurde von Fisher nicht geliefert. Sie ist vielmehr ein Teil des Modells von Neyman & Pearson, das im folgenden Abschnitt geschildert wird.

3. **Neyman & Pearson** Entscheidet man sich, Fisher folgend, wegen eines kleinen p-Wertes gegen eine Hypothese, so ist diese Entscheidung mit einer gewissen Wahrscheinlichkeit α falsch: auch, wenn die Hypothese 'wahr' ist, können ja Daten x auftreten derart, dass $P(T > T(x)|H) = p$ ist. Behält man umgekehrt H bei, weil p hinreichend groß ist, so begeht man damit mit einer bestimmten Wahrscheinlichkeit β ebenfalls einen Fehler: obwohl H falsch ist, wird eine Stichprobe beobachtet, die mit ihr kompatibel erscheint, und die Alternativhypothese wird wegen der Akzeptanz von H_0 'verworfen'.

   Die tatsächliche Berechnung von β wird bei Fisher nicht vorgenommen, weil die Alternativhypothese nicht explizit formuliert wird. Dies kann man als Nachteil sehen. Neyman & Pearson fordern also, dass für einen vorgegebenen Wert von α die Wahrscheinlichkeit $1 - \beta$ (die Wahrscheinlichkeit, dass die Alternativhypothese H_1 irrtümlich abgelehnt wird) minimalisiert wird. Dies gelingt durch Annahme einer Effektgröße, mit der H_1 spezifiziert wird, über die der Stichprobenumfang bestimmt wird, die die geforderte Minimalisierung erlaubt.

4. **Likelihood-Inferenz** Bei dieser Strategie, Daten in Bezug auf die in Frage stehenden Hypothesen zu bewerten, wird nur von den Likelihoods der Daten, gegeben die Hypothesen, Gebrauch gemacht. Hat man die zwei alternativen Hypothesen H_1 und H_2, kann man den Likelihood-Quotienten

$$\Lambda(x) = \frac{f(x|H_1)}{f(x|H_2)} \tag{6}$$

betrachten. Für $\Lambda(x) > 1$ sind die Daten x wahrscheinlicher unter H_1, und für $\Lambda(x) < 1$ sind sie wahrscheinlicher unter H_2; für $\Lambda(x) = 1$ sind die Daten für beide Hypothesen gleichwahrscheinlich. Wie bei Fisher beurteilt man die Hypothesen nach Maßgabe der Wahrscheinlichkeit der Daten, gegeben eine Hypothese; natürlich kann man eine Schwelle $\lambda \neq 1$ einführen derart, dass man sich für H_1 entscheidet, wenn $\Lambda(x) > \lambda$, und für H_2, wenn $\Lambda(x) < \lambda$. λ kann die Kosten der jeweiligen Entscheidung reflektieren.

Der Vorteil des Bayesschen Ansatzes ist offenkundig: mit $P(H|D)$ erhält man eine direkte Aussage über die Wahrscheinlichkeit, mit der H korrekt ist. Bezieht sich H auf den Wert eines unbekannten Parameters θ, so liefert (3) eine Wahrscheinlichkeitsverteilung für θ, sweil jeder mögliche Wert von θ eine Hypothese repräsentiert. Allerdings muß eine Annahme über die Priori-Wahrscheinlichkeit gemacht werden, und darüber hinaus muß ein bestimmter Wahrscheinlichkeitsbegriff gewählt werden, denn $P(H)$ und $P(H|x)$ sind im Allgemeinen subjektive Größen. Die Frage, wie Wahrscheinlichkeiten als subjektive Größen zu definieren bzw. zu interpretieren sind, stellt sich weniger, wenn Fishers Ansatz gewählt wird. Hier muß nach der Wahrscheinlichkeitsverteilung der Statistik $T(X)$ gefragt werden. Sie wird zumindest für kleinere Stichprobenumfänge durch die Annahme über die Verteilung der Daten X bestimmt, und diese Annahme kann zumindest im Prinzip getestet werden. Die Problematik des Ansatzes liegt zunächst in der Definition des p-Wertes: p hängt nicht nur von den tatsächlich beobachteten Daten x ab, sondern von der Menge der nicht beobachteten Werte X, für die $T(X) > T(x)$) ist. Dies verdeutlicht noch einmal Jeffreys Anmerkung auf Seite 4. Darüber hinaus ist p natürlich ebenfalls eine zufällige Veränderliche, weil der Wert von p von den Daten x abhängt, die ja eine Zufallsstichprobe aus einer Population sind. Die Frage ist also, wie p etwa unter H_0 verteilt ist. Die Antwort auf diese Frage gibt Aufschluß, wie häufig man "hinreichend" kleine p-Werte bekommt, die zur Zurückweisung von H_0 führen, *obwohl* H_0 korrekt ist. Die intuitive Idee ist, dass dieser Fall eben nicht häufiger als etwa mit der Wahrscheinlichkeit .05 auftritt, wenn man im Falle $p < .05$ die Hyopthese H_0 verwirft. Leider ist diese Intuition nicht notwendig korrekt.

Der Neyman-Pearson-Ansatz scheint diese Problematik zu umgehen. Der Punkt ist aber, dass die Alternativhypothese H_1 nicht notwendig spezifiziert ist: zwei Bedingungen, unter denen Messwerte gewonnen werden, unterscheiden sich entweder nicht in ihrer Wirkung, oder eine von ihnen hat einen größeren Effekt, – aber um wieviel größer ist er? Eine Antwort findet man, wenn eine Effektstärke für H_1 formuliert wird, – aber damit nähert man sich einer Bayesschen Auffassung, denn die Effektstärke ist letzlich eine subjektive Größe, auch wenn sie unter Berufung auf bereits vorhandene Daten festgelegt wird. Darüber hinaus lassen sich Fälle angeben, bei denen der Neyman-Pearson-Ansatz zu Fehlentscheidungen führt, die über die α- und β-Fehler hinaus gehen.

Die Likelihood-Inferenz scheint von dieser Problematik frei zu sein, ebenso von der anderen Problematik, dass u. U. zu häufig gegen H_0 entschieden wird, wie schon Jeffreys argumentierte. Die Entscheidungen hängen nur von den Likelihoods ab. Andererseits ist die Wahl der Schwelle λ von subjektiven Entscheidungen abhängig. So kommt man in der einen oder anderen Form auf den Bayesschen Ansatz zurück.

Die wissenschaftstheoretische Diskussion über die Sinnhaftigkeit der verschiedenen Ansätze ist so ausgiebig geführt worden, dass wohl kaum einer ihrer verschiedenen Aspekte noch nicht behandelt wurde. Dementsprechend wird an dieser Stelle gar nicht erst der Versuch gemacht, diese Diskussion zu referieren. Statt dessen sollen einige der wesentlichen Aspekte des Bayesschen Ansatzes dargestellt werden.

1.2 Das Likelihood-, das Suffizienz- und das Konditionalitätsprinzip

Das Likelihood-Prinzip $x = (x_1, \ldots, x_n)$ sei eine Stichprobe mit $x_i \sim f(x|\theta)$, und θ sei unbekannt. Das Likelihood-Prinzip besteht in der Aussage, dass alle Information in der Stichprobe bezüglich des Parameters θ in der Likelihood $l(\theta|x)$ enthalten ist. Ist also $y = (y_1, \ldots, y_n)$ eine weitere Stichprobe mit $y_j \sim f(x|\theta)$ und sind die Likelihoods von x und y gleich, so enthalten beide die gleiche Information über θ.

In dieser Form hört sich das Prinzip harmlos an. Die Daten x und y können aber in verschiedenen Experimenten erhoben worden sein, z.B.

1. Der Wissenschaftler W-I führt ein Binomialexperiment durch: es werden n Bernoulli-Experimente durchgeführt. Es werden x "Erfolge" erzielt. Die Wahrscheinlichkeit für dieses Ergebnis ist

$$P(X = x|\theta, n) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{7}$$

2. Wissenschaftler W-II führt ein "Inverses Binomialexperiment" durch: es werden so viele Bernoulli-Versuche durchgeführt, bis x Erfolge verzeichnet wurden. Die Wahrscheinlichkeit, dass bei insgesamt n Versuchen gerade x Erfolge eingetreten sind, ist

$$P(N = n|x, \theta) = \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}. \tag{8}$$

Natürlich ist im Binomialexperiment die Likelihood-Funktion durch $L_1(\theta|x) = P(X = x|\theta, n)$ gegeben und im inversen Binomialexperiment ist sie durch $L_2(\theta|N) =$

$P(N = n|x, \theta)$ gegeben. Der von θ abhängende Term ist in beiden Likelihood-Funktionen identisch, nämlich

$$\theta^x(1-\theta)^{n-x}.$$

Schätzt man θ mit der ML-Methode, so kommt man zur gleichen Schätzung $\hat{\theta}$, da die jeweiligen nur von n und x abhängenden Faktoren beim Differenzieren verschwinden. Dem Likelihood-Prinzip zufolge kommt man im Falle gleicher x-Werte zu gleichen Schlußfolgerungen über θ.

In dieser eher abstrakten Präsentation erscheint das Likelihood-Prinzip plausibel zu sein. Es widerspricht aber z.B. dem Fisherschen Signifikanztest. Um dies zu sehen, werde das folgende Beispiel betrachtet:

Beispiel 1.1 Therapievergleich Man stelle sich vor, dass zwei Therapien T_1 und T_2 miteinander verglichen werden sollen. Ein (Bernoulli-)Versuch bestehe im Vergleich des Erfolges bei einem zufällig gebildeten Paar von Patienten (A_i, B_i), wobei A_i mit T_1 und B_i mit T_2 behandelt wird. Wirkt T_1 besser als T_2 ($T_1 \succ T_2$), so werde dies als "Misserfolg" notiert, wirkt T_2 besser als T_1 ($T_1 \prec T_2$), so sei dies ein "Erfolg". Gemäß H_0 sind beide Therapien gleich gut, so dass $\theta = P(T_1 \prec T_2) = P(T_1 \succ T_2) = 1 - \theta = 1/2$. Es werden die Paare (A_i, B_i), $i = 1, \ldots, n$ betrachtet; jedes Paar entspricht einem Bernoulli-Versuch. Dazu sollten die Paare unabhängig voneinander gebildet werden.

Es gibt zwei Möglichkeiten: entweder es wird der Wert von n, also die Anzahl der Paare, von vornherein festgelegt und dann die Anzahl x der "Erfolge" notiert; hier ist x eine zufällige Veränderliche. Oder es wird eine Anzahl von "Erfolgen" festgelegt und so lange getestet, bis man x Erfolge beobachtet hat. Bei diesem Vorgehen ist n eine zufällige Veränderliche.

Der Wissenschaftler W-I entscheidet sich, $n = 6$ Bernoulli-Versuche durchzuführen und findet $x = 1$ "Erfolg". Die Wahrscheinlichkeit dieses Ergebnisses ist unter H_0

$$P(X = 1|H_0) = \binom{6}{1}\left(\frac{1}{2}\right)^1\left(1-\frac{1}{2}\right)^{6-1} = 6\left(\frac{1}{2}\right)^6 = .094. \tag{9}$$

Nach Fisher muß nun die Wahrscheinlichkeit dieses oder eines extremeren Ergebnisses bestimmt werden, um zu prüfen, ob die Daten "signifikant" der Nullhypothese widersprechen. Dies ist die Wahrscheinlichkeit $P(X \leq 1|H_0)$ bzw. $P(Y \geq 5|H_0$, wenn Y die Anzahl der "Mißerfolge" bedeutet: offenbar ist

$$P(X \leq 1|H_0) = P(Y \geq 5|H_0).$$

Es ist

$$p = P(X \leq 1|H_0) = P(X = 0|H_0) + P(X = 1|H_0) = \left(\frac{1}{2}\right)^6 + 6\left(\frac{1}{2}\right)^6 = .11.$$

Erklärt man den p-Wert als 'signifikant', wenn $p < .05$, so muß man folgern, dass die Daten mit der Nullhypothese kompatibel sind.

Der Wissenschaftler W-II hat sich entschieden, $x = 1$ festzusetzen und nun den Wert von n zu bestimmen, er testet also so lange, bis sich der erste Erfolg eingestellt hat. Die Wahrscheinlichkeit des Ergebnisses ist

$$P(n = 6, x = 1|H_0) = \binom{5}{0}\left(\frac{1}{2}\right)^6 = 1\left(\frac{1}{2}\right)^6 = .016.$$

Die Frage ist, was es bedeutet, ein extremeres Ergebnis zu bekommen, d.h. es muß die Wahrscheinlichkeit $P(n > 6|H_0)$ bestimmt werden. Die Wahrscheinlichkeit hierfür ist (s. Anhang, Abschnitt 3.1),

$$p = P(n > 6|H_0) = .031. \tag{10}$$

Da $p < .05$, ist in diesem Experiment der Befund signifikant: θ ist anscheinend von $1/2$ verschieden.

Wissenschaftler W-I, der von vorn herein 6 Versuche geplant hat, hat also kein signifikantes Ergebnis erzielt, wohl aber Wissenschaftler W-II, der einfach so lange Versuche durchgeführt hat, bis sich ein "Erfolg" eingestellt hat. Die Frage ist nun, wer von beiden Recht hat. □

Das Problem entsteht, weil das Ergebnis – nicht signifikant oder signifikant – von der Planung der Untersuchung abhängt, die Daten aber identisch sind, nämlich 5 "Misserfolge" und ein "Erfolg". Findet man diese Daten, ohne dass aus einer Beschreibung hervor geht, nach welchem Plan sie erhoben wurden, und akzeptiert man den Fisherschen Signifikanztest, so darf man sie nicht auswerten, denn der unbekannte Versuchsplan und nicht die Daten entscheiden, ob H_0 beibehalten oder verworfen werden soll.

Geht man nach dem Likelihood-Prinzip vor, so entsteht dieses Problem nicht. Es wird keine Bewertung von H_0 nach Maßgabe der Wahrscheinlichkeit nicht beobachteter Werte ($n > 6$) vorgenommen, sondern es wird nur nach Maßgabe der Wahrscheinlichkeit der tatsächlich beobachteten Werte geurteilt. Die Likelihoods der Daten aus den beiden Untersuchungen unerscheiden sich nur durch einen nicht von θ abhängenden Faktor c:

$$\binom{6}{1}(1-\theta)^5\theta^1 = c\binom{5}{0}(1-\theta)^5\theta^1,$$

woraus

$$c = \binom{6}{1}/\binom{5}{0} = 6$$

folgt. Die "Evidenz" bezüglich H_0 in den Daten besteht nur aus dem Term $(1-\theta)^5\theta$, und in der Tat ist die ML-Schätzung für θ für beide Untersuchungen identisch, weil bei der Differentiation alle nicht von θ abhängenden Terme wegfallen.

Argumentiert man nach dem generell[5] akzeptierten Schlußverfahren des *tertium non datur*, so muß man den Fisher-Ansatz verwerfen. Diesem Schlußverfahren zufolge geht man von einer Annahme $\mathcal{A}$ aus und zeigt, dass sie auf einen Widerspruch führt. Daraufhin verwirft man $\mathcal{A}$ und schließt auf die Gültigkeit von $\neg\mathcal{A}$. Fishers Ansatz führt auf einen Widerspruch, also muß man in verwerfen. Das Likelihood-Prinzip liefert den Ansatz für Alternativen: entweder die Likelihood-Inferenz oder den Bayesschen Ansatz. Aber so einfach wird es nicht sein; die Menge der Arbeiten, in denen der Fisher-Ansatz gerechtfertigt wird, kann hier nicht diskutiert werden. Es kann nur an der anscheinenden Selbstverständlichkeit des p-Wert-Rituals gezweifelt werden.

Das Suffizienz-Prinzip Ist $x = (x_1, \ldots, x_n)$ eine Stichprobe, so lassen sich daraus Statistiken $T(x)$ berechnen: eine *Statistik* ist zunächst einmal irgendeine Funktion der Daten x. Insbesondere sind der Mittelwert $\bar{x}$, die Stichprobenvarianz s^2, der Median $Med(x)$ Statistiken. Statistiken können u. U. als Schätzer für unbekannte Parameter dienen: $\bar{x}$ für den Erwartungswert $\mu = \mathbb{E}(x)$, s^2 für die Varianz σ^2, etc. Ein Schätzer für einen Parameter wird um so besser sein, je mehr Information er über den Parameter in den Daten ausnutzt. Nutzt eine Statistik $T(x)$ die gesamte Information über einen Parameter θ aus, so heißt die Statistik *suffizient*. Die Frage ist, wie man feststellt, ob eine Statistik suffizient ist. Dazu hilft eine formalere Definition der Suffizienz:

Definition 1.1 *Die Statistik $T(x)$ heißt* suffizient *(für die Schätzung des Parameters θ), wenn die bedingte Verteilung der Stichprobe, gegeben $T(x)$, nicht von θ abhängt; in diesem Fall gilt*

$$f(x|T(x), \theta) = f(x|T(x)). \tag{11}$$

Die Gleichung (11) macht die Idee der Suffizienz explizit: Die Verteilung von x hängt zwar von θ ab, substituiert man aber $T(x)$ für θ, so ist sie durch Angabe des Schätzers für θ vollkommen spezifiert. Daraus ergibt sich die Frage, wie Verteilungen bzw. Dichten definiert sein müssen, damit suffiziente Statistiken existieren. Die Frage wird durch den Faktorisierungssatz beantwortet:

Satz 1.1 *Es sei $f(x|\theta)$ die Dichtefunktion für die Stichprobe x. Die Statistik $T(x)$ ist genau dann suffizient für den Parameter θ, wenn es Funktionen $g(T;\theta)$ und*

[5] In der von dem Mathematiker L.E.J. Brouwer (1881–1966) formulierten Philosophie der Mathematik (Intuitionismus) wird das Tertium-non-Datur nicht als allgemeines Schlußprinzip anerkannt.

$h(x)$ gibt derart, dass f in der Form

$$f(x|\theta) = g(T;\theta)h(x) \tag{12}$$

dargestellt werden kann. g hängt von x nur über die Statistik $T(x)$ ab und ist im Übrigen eine Funktion des Parameters θ, und h ist unabhängig von θ.

Beweis: Den Beweis findet man in den gängigen Lehrbüchern zur mathematischen Statistik. □

Der Satz soll an Beispielen illustriert werden.

Beispiel 1.2 Gegeben sei $x = (x_1, \ldots, x_n)$, wobei x_i die Anzahl der "Erfolge" in einem Bernoulli-Experiment ist, das insgesamt n-mal durchgeführt wurde; θ habe in allen Experimenten den gleichen Wert. Die Wahrscheinlichkeit der Daten ist dann durch

$$f(x|\theta) = \theta^T(1-\theta)^{n-T}, \quad T = T(x) = \sum_{i=1}^{n} x_i \tag{13}$$

gegeben. Offenbar ist die Kenntnis der einzelnen x_i-Werte nicht notwendig, um $f(x|\theta)$ zu spezifizieren, es genügt, $T(x)$ zu kennen. Also ist T eine suffiziente Statistik. □

Beispiel 1.3 Poisson-Verteilung Gegeben sei $x = (x_1, \ldots, x_n)$, wobei x_i gemäß einer Poisson-Verteilung mit dem Parameter $\theta = \lambda$ verteilt sei, d.h.

$$P(X_i = x_i|\lambda) = e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}$$

Für x ergibt sich

$$f(x|\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^T}{\prod_i x_i!}, \quad T = T(x) = \sum_{i=1}^{n} x_i. \tag{14}$$

Offenbar kann f in der Form $f = g(T,\lambda)h(x)$ dargestellt werden, mit

$$g(T,\lambda) = e^{-\lambda}, \quad h(x) = 1/\prod_{i=1}^{n} x_i!.$$

T ist eine sufiziente Statistik für λ.

Die Gleichung (14) impliziert, dass

$$L(\lambda|x) \propto -n\lambda + T\log\lambda,$$

und

$$\frac{dL}{d\lambda} = -n + \frac{T}{\lambda}.$$

Dann wird $dL/d\lambda = 0$ für

$$\hat{\lambda} = \frac{1}{n}T. \tag{15}$$

Die ML-Schätzung für λ zeigt, in welchem Sinne T eine Schätzung für λ liefert. □

Beispiel 1.4 Gauß-Verteilung Wieder sei $x = (x_1, \ldots, x_n)$ gegeben, aber es sei $x_i \sim N(\mu, \sigma^2)$. Es ist, mit $\theta = \mu$ und bekanntem Wert von σ^2,

$$f(x|\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]. \tag{16}$$

Es ist

$$(x_i - \mu)^2 = (x_i - \bar{x} + \bar{x} - \mu)^2 = (x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2(x_i - \bar{x})(\bar{x} - \mu),$$

so dass

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,$$

denn

$$2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - \mu) = 2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \bar{x}) = 0,$$

da die Summe der Abweichungen vom Mittelwert stets gleich Null ist. Mithin hat man

$$f(x|\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]. \tag{17}$$

Offenbar kann f in der Form

$$f(x|\mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \exp\left[-\frac{1}{2\sigma^2}n(\bar{x} - \mu)^2\right] \tag{18}$$

geschrieben werden, also in der Form $f = gh$ mit

$$h(x) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - T(x))^2\right], \quad g(T, \mu) = \exp\left[-\frac{1}{2\sigma^2}n(\bar{x} - \mu)^2\right],$$

und $T(x) = \bar{x}$. T ist eine suffiziente Statistik für $\bar{x}$. □

Das Konditionalitätsprinzip Diesem Prinzip zufolge hängen die Schlußfolgerungen über einen Parameter nur vom tatsächlich durchgeführten Experiment ab, - und nicht von der Art und Weise, wie das Experiment zustande gekommen ist. Das mag auf den ersten Blick trivial erscheinen. Aber man denke an die Aufgabe, empirisch entscheiden zu sollen, welche von zwei Therapien die bessere ist. Wie bereits als Illustration des Likelihood-Prinzips diskutiert wurde, kann man entweder ein Binomialexperiment durchführen, beim dem die Anzahl n der Vergleiche vor Durchführung der Untersuchung festgelegt wird, oder man kann ein inverses Binomialexperiment planen, bei dem die Anzahl der Vergleiche eine zufällige Veränderliche ist, weil die Versuchsreihe gestoppt wird, wenn eine bestimmte Anzahl von "Erfolgen" (etwa: $T_1 \prec T_2$) erreicht wurde. Die Forschergruppe kann sich nicht entscheiden, welcher Typ von Untersuchung durchgeführt werden soll und überläßt schließlich die Entscheidung der Münze. Für "Kopf" wird das Binomialexperiment durchgeführt, für "Zahl" das inverse Binomialexperiment. Man spricht von einem Mischexperiment. Das Konditionalitätsprinzip postuliert, dass nur das Ergebnis betrachtet wird, dass bedobachtet wurde. Natürlich können mehr als zwei Experimentaltypen betrachtet werden. Stets soll es, dem Konditionalitätsprinzip zufolge, nur auf die tatsächlich erhaltenen Daten ankommen.

Man kan das Konditionalitätsprinzip allgemein formulieren (Helland, 1995):

> Es seien E_h, $h = 1, 2, \ldots, k$, k Experimente, und x_h bezeichne das Resultat des h-ten Experiments. Weiter seien $\pi_1, \ldots, \pi_k$ bestimmte Wahrscheinlichkeiten mit $\sum_h \pi_h = 1$. Nun werde das Experiment $E = E_h$ mit der Wahrscheinlichkeit π_h gewählt; das Ergebnis sei (h, x_h). Die über E gewonnene experimentelle Evidenz $Ev[E, (h, x_h)]$ ist gleich der, die aus dem Experiment E_h gewonnen wird: $Ev[E, (h, x_h)] = Ev[E_h, x_h]$.

Der Begriff 'experimentelle Evidenz' wurde von Birnbaum (1962) eingeführt, ebenso die Notation $Ev[E, (h, x_h)]$ bzw. $Ev[h, x_h]$, ohne dass allerdings eine spezifische Definition für diesen Begriff gegeben wurde. Die experimentelle Evidenz ist gewissermaßen die Information, die man den experimentellen Ergebnissen entnimmt oder entnehmen kann. Diese Evidenz kann z.B. "weiterverarbeitet" werden, indem man eine Statistik $T(x)$ definiert und Signifikanz konstatiert, wenn $T(x) \geq T(x_h)$ gefunden wird, – aber das ist nur eine Möglichkeit von mehreren, die Evidenz zu interpretieren.

Die allgemeine Formulierung des Prinzips mag auf den ersten Blick merkwürdig leer erscheinen; gleichwohl ist das Prinzip Gegenstand einer längeren Kontroverse unter Statistikern. Das heißt, es ist nicht so 'wahr', wie es auf den ersten Blick evident zu sein scheint. Zunächst fällt eine Ähnlichkeit zum Likelihood-

Prinzip auf: es soll das gelten, was tatsächlich beobachtet wurde, und nicht das, was hätte beobachtet werden können.

Ein Blick auf die Entscheidung zwischen Binomial- und inversem Binomialexperiment verdeutlicht aber, dass das Prinzip nicht trivial ist. Fragt man experimentell arbeitende Wissenschaftler, so antworten einige, das Binomialexperiment sei irgendwie valider als das inverse Binomialexperiment, wobei ihnen eine Begründung allerdings schwer fällt. In älteren Lehrbüchern (und in neueren, deren Texte durch Abschreiben von den älteren entstanden sind) wird oft die These vertreten, man müsse vor Beginn der Untersuchung u. A. den Stichprobenumfang und das Signifikanzniveau festlegen, andernfalls könne man ja aufhören, zu experimentieren, wenn einem das Ergebnis gerade paßt, oder so lange weiter experimentieren, bis es einem passt. Das inverse Binomialexperiment widerspricht dieser Forderung nicht, weil ja vor Beginn der Untersuchung festgelegt wird, wieviele "Erfolge" beobachtet werden müssen, damit die Folge von Untersuchungen abgebrochen werden kann. Die Evidenz bezüglich der Hypothese H_0 sollte für beide Experimente die gleiche sein: ob n festliegt und die Anzahl x der Erfolge als Ergebnis genommen wird oder ob x festliegt und die Gesamtzahl n der Versuche als Ergebnis genommen wird sollte keinen Einfluß auf die Bewertung von H_0 haben. Die Frage ist, wie dieses "sollte" bewiesen werden kann, ohne einfach auf nur intuitive Betrachtungen zu verweisen: diese führen, je nach dem mentalen Zustand desjenigen, der sie anstellt, mal auf die These und mal auf die Antithese. Nimmt man den Signifikanztest als Angelpunkt des Arguments, so erscheinen die Experimente nicht gleichwertig, d.h. man akzeptiert das Konditionalitätsprinzip *nicht*. Ist man andererseits vom Konditionalitätsprinzip – aus welchen Gründen auch immer – überzeugt, so muß man den Signifikanztest als geeignete Interpretation der Evidenz zurückweisen. Damit wird deutlich, warum es eine Kontroverse über dieses Prinzip gibt.

Es soll zunächst auf einen von Birnbaum (1962) und in verbesserter Form (1972) artikulierten Befund hingewiesen werden. Birnbaum bewies, dass das Likelihood-Prinzip aus dem Konditionalitäts- und dem Suffizienzprinzip logisch folgt. Im 1972-er Artikel führt Birnbaum *Axiome der statistischen Evidenz* ein, die allerdings einige Definitionen voraussetzen. Zunächst wird der Begriff des Modells eines Experiments eingeführt. Das Modell für statistische Evidenz wird durch das Tripel $E = (\Omega, S, f)$ spezifiziert, wobei Ω der Parameterraum ist, S ist der Ereignis- oder Stichprobenraum, und $f(x, \theta) = P(X = x|\theta)$, $\theta \in \Omega$. Mit $Ev(E, x)$ wird die statische Evidenz bezeichnet. Birnbaum (1972) führt den Begriff der *Mathematischen Evidenz* (M) ein:

$$\text{Wenn} \quad f(x,\theta) = f(x',\theta) \text{ für alle } \theta \in \Omega, \text{ dann } Ev(E,x) = Ev(E,x). \qquad (19)$$

Damit soll ein 'natürlicher' Begriff der statistischen Evidenz ausgedrückt werden: zwei Modelle statistischer Evidenz ist äqivalent, wenn sie sich nur in der Etiket-

tierung der Stichprobenwerte unterscheiden. Ein einfaches Beispiel illustriert den Begriff anhand zweier unabhängiger Bernoulli-Versuche: die Ereignisse $(0,1)$ und $(1,0)$ haben die gleiche Wahrscheinlichkeit $\theta(1-\theta)$, für jedes θ. Es sei entweder $\theta = .1$ oder $\theta = .5$. Dann ist

$$E = (f(x,\theta)) = \left(\begin{array}{c|cccc} & \theta\theta & \theta(1-\theta) & (1-\theta)\theta & (1-\theta)(1-\theta) \\ \hline \theta = .1 & .01 & .09 & .09 & .81 \\ \theta = .5 & .25 & .25 & .25 & .25 \end{array}\right) \qquad (20)$$

Das Prinzip (M) besagt, dass die Ergebnisse, die den Spalten 2 und 3 in (20) entsprechen, äquivalent sind, also $Ev(E,2) = Ev(E,3)$. In Birnbaum (1962) war (M) die Basis für die allgemeine Diskussion, ohne besonders fomalisiert zu werden. Birnbaum (1972) zeigt dann:

1. Das Konditionalitätsprinzip (C) und (M) implizieren zusammen das Likelihood-Prinzip (L),

2. Das Likelihood-Prinzip (L) impliziert das Suffizienz-Prinzip (S), dass das schwache Suffizienz-Prinzip (S') impliziert, das wiederum (M) impliziert.

   Schwaches Suffizienz-Prinzip (S'): Gilt, für irgend ein $c > 0$, $f(x,\theta) = cf(x*,\theta)$ für alle $\theta \in \Omega$, dann gilt auch $Ev(E,x) = Ev(E,x^*)$.

3. Korollar: Das Konditionalitätsprinzip (C) zusammen mit (M) implizieren das Suffizienz-Prinzip (S).

4. Jede der drei Implikationen

   $$(L) \leftarrow (S) \leftarrow (S') \leftarrow (M)$$

   ist falsch.

Birnbaums Arbeiten haben eine andauernde Diskussion über die Allgemeingültigkeit des Konditionalitätsprinzips und der dazu korrespondierenden Implikationen angeregt, die hier schon aus Platzgründen nicht wiedergegeben werden kann. Akaike (1982) argumentierte anhand des Binomial- und des inversen Binomialexperiments, dass die Begriffe der mathematischen Äquivalenz und des Likelihood-Prinzips Tautologien seien, die Birnbaums Beweis disqualifizieren, ohne dass deswegen der Bayes-Ansatz aufgegeben werden müsse, – Box & Tiao (1973) hätten die richtigen Anweisungen zum Gebrauch von geeigneten Priori-Verteilungen gerade für diese beiden Typen von Experimenten gegeben. Beispielhaft sei die Arbeit von Evans, Fraser & Monette (1986) (mit anschließender Diskussion, (Kalbfleisch, Berger, Dawid, Sprott (1986)) erwähnt; sie zeigten, dass zumindest für diskrete Stichprobenräume das Likelihood-Prinzip bereits aus dem Konditionalitätsprinzip folgt; das Suffizienzprinzip muß in diesem Fall gar nicht

mehr vorausgesetzt werden. Das Resultat kann auf allgemeinere Stichprobenräume verallgemeinert werden (Helland (1995)), so dass man von einer Äquivalenz des Likelihood- und des Konditionalitätsprinzips ausgehen kann. Gleichzeitig kritisiert Helland (1995) die Konzepte von einem klassischen Standpunkt aus, – um sofort von Lavine (1996) kritisiert zu werden. Diese kursorische Auflistung von Argumenten pro und contra das Konditionalitäts- und das Likelihood-Prinzip ist keineswegs vollständig, aber sie mag genügen, um zu zeigen, dass die Grundpositionen der Inferenzstatistik unklarer sind, als es der apodiktische Verweis auf auf den Signifikanztest in Lehrbüchern für Anwender vermuten läßt.

2 Bayessche Statistik

2.1 Das Prinzip

Viele Hypothesen beziehen sich auf Werte von unbekannen Parametern: man will den "wahren" Wert einer Person in einem Eignungstest abschätzen, man möchte wissen, ob sich zwei experimentelle Bedingungen im Mittel unterscheiden, d.h. man möchte den Wert der Differenz $\Delta\mu$ für die Verteilung der Differenzen schätzen, oder man möchte den Wert einer Korrelation zwischen zwei Variablen schätzen. In der klassischen oder orthodoxen Statistik wird postuliert, dass die Parameterwerte unbekannte Konstante sind. Wird der Bayessche Ansatz gewählt, so ist die Basis für jede Inferenz die Beziehung

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta)P(\theta) \tag{21}$$

(der Proportionalitätsfaktor ist $1/P(\mathbf{x})$. $P(\theta)$ ist die a-priori-Wahrscheinlichkeit für θ, im Folgenden kurz Prior genannt, und $P(\theta|\mathbf{x})$ ist die a-posteriori-Wahrscheinlichkeit für θ, gegeben die Daten $\mathbf{x}$, im Folgenden kurz Posterior genannt. Sowohl die Prior wie die Posterior sind auf einer Menge Θ definiert. Dieser Sachverhalt wird gelegentlich durch die Aussage, θ sei nun als zufällige Veränderliche konzipiert, gedeutet (Held (2008); 139). Die Frage, in welchem Sinn θ eine zufällige Veränderliche ist, ist allerdings nicht leicht zu beantworten. Einfacher ist es, den Wert eines Parameters nach wie vor als Konstante aufzufassen, deren Wert unbekannt ist. Die Wahrscheinlichkeit $P(\theta)$ ist eine epistemische Größe, die ausdrückt, für wie wahrscheinlich man den Wert θ für den unbekannten, aber doch konstanten Parameter hält. Man denke an das Ziegenproblem: der Preis – das schöne Auto – ist hinter einer von drei möglichen Türen, der (Lokations-)Parameter kann einen von drei Werten – 1, 2, oder 3, korrespondierend zu den Türen – haben. Der Spieler ordnet jeder Tür eine Wahrscheinlichkeit zu, die seinen Glauben an den Ort des Preises ausdrückt. Analog dazu kann man aber auch sagen, dass der Spielleiter zufällig zwischen den Werten 1, 2 und 3 gewählt hat, um den Preis hinter der entsprechenden Tür zu verbergen. Aber diese Interpretation überträgt

sich nicht notwendig auf den Wert etwa von Naturkonstanten; ist der Wert der Gravitationskonstante $g(\approx 9.81\ m/s^2)$ zufällig gewählt worden? Das kann man annehmen, der Wert von g ist dann eine Konsequenz der Tatsache, dass die Erde eine bestimmte Masse hat, und die Masse kann als Realisierung eines zufälligen Ereignisses gesehen werden. In ähnlicher Weise kann man den 'wahren' Wert einer Person in einem Eignungstest als Realisierung eines zufälligen Ereignisses sehen. Für bestimmte Autoren, etwa K.R. Popper, sind Hypothesen allerdings grundsätzlich keine zufälligen Ereignisse, ebensowenig, wie er den Begriff einer epistemischen Wahrscheinlichkeit schätzt; für ihn sind Wahrscheinlichkeiten objektive *Propensitäten*, die den beobachteten Prozessen inhärent sind. Es ist allerdings nicht klar, warum man sich Poppers Ansichten anschließen soll.

Geht man also der Einfachheit halber von epistemischen Wahrscheinlichkeiten aus, so ist die Posterior die Basis für die Inferenz bezüglich der zur Diskussion stehenden Hypothesen. Es ist sinnvoll, eine strengere Definition der Posterior einzuführen:

Definition 2.1 *Es sei $\boldsymbol{x}$ der beobachtete Wert der zufälligen Veränderlichen $\boldsymbol{X}$ (oder des zufälligen Vektors $\boldsymbol{X}$). $\boldsymbol{X}$ habe die Dichtefunktion $f(\boldsymbol{X}|\theta)$. Weiter habe θ die Prior-Dichte $f(\theta)$. Dann ist die Posterior-Dichte für θ durch*

$$f(\theta|\boldsymbol{X}) = \frac{f(\boldsymbol{X}|\theta)f(\theta)}{\int_\Theta f(\boldsymbol{X}|\theta)f(\theta)\,d\theta} \tag{22}$$

gegeben.

Anmerkung: $f(\theta|\mathbf{X})$ ist offenbar analog zum Satz von Bayes definiert; das Integral im Nenner entspricht dem Satz der Totalen Wahrscheinlichkeit. Kann θ nur Werte θ_i, $i = 1, 2, \ldots$ annehmen, wird das Integral durch eine Summe ersetzt. □

$f(\mathbf{X}|\theta)$ entspricht einer Likelihood $L(\theta)$. Oft wird für die Likelihood

$$l(\theta|\mathbf{X}) = f(\mathbf{X}|\theta) \tag{23}$$

geschrieben, und die Rede ist von einer *Likelihood-Funktion*. Für den Logarithmus der Likelihood-Funktion wird

$$L(\theta|\mathbf{X}) = \log l(\theta|\mathbf{X}) \tag{24}$$

geschrieben.

In (21) ist der Bayessche Satz in $\propto$-Form geschrieben worden; diese Form bleibt erhalten, wenn $l(\theta|\mathbf{X})$ mit einer beliebigen Konstanten multipliziert wird; dies ermöglicht, die Likelihood durch ein beliebiges Vielfaches von $l(\theta|\mathbf{X})$ zu definieren. Nun sei

$$\int_\Theta l(\theta|\mathbf{X})\,d\mathbf{X} < \infty.$$

Dann heißt

$$\frac{l(\theta|\mathbf{X})}{\int_\Theta l(\theta|\mathbf{X})\, d\mathbf{X}} \tag{25}$$

auch *standardisierte Likelihood.*

Es sei nun $\mathbf{Y}$ eine zweite Stichprobe (oder zufälliger Vektor), die unabhängig von $\mathbf{X}$ sei, und es sei

$$P(\theta|\mathbf{X}, \mathbf{Y}) \propto l(\theta|\mathbf{X}, \mathbf{Y})P(\theta).$$

Wegen der Unabhängigkeit hat man

$$P(\mathbf{X}, \mathbf{Y}|\theta) = P(\mathbf{X}|\theta)P(\mathbf{Y}|\theta),$$

so dass

$$l(\theta|\mathbf{X}; \mathbf{Y}) \propto l(\theta|\mathbf{X})l(\theta|\mathbf{Y}), \tag{26}$$

so dass

$$\begin{aligned} P(\theta|\mathbf{X}, \mathbf{Y}) &\propto l(\theta|\mathbf{X})l(\theta|\mathbf{Y})P(\theta) \\ &\propto P(\theta|\mathbf{X})l(\theta|\mathbf{Y}). \end{aligned} \tag{27}$$

Der folgende Begriff ist gelegentlich nützlich:

Definition 2.2 *Es sei*

$$P(\boldsymbol{X}) = \int_\Theta P(\boldsymbol{X}|\theta)P(\theta)d\theta \tag{28}$$

die Randverteilung für $\boldsymbol{X}$. $P(\boldsymbol{X})$ heißt auch die prädiktive Verteilung *von $\boldsymbol{X}$.*

Parameterschätzungen: Man kann nun erklären, was mit einer Bayesschen Parameterschätzung gemeint ist:

Definition 2.3 *Es sei*

$$\mathbb{E}(\theta|\boldsymbol{X}) = \int_\Theta \theta f(\theta|\boldsymbol{X})\, d\theta \tag{29}$$

der Posterior-Erwartungswert von θ, gegeben die Daten $\boldsymbol{X}$. $\mathbb{E}(\theta|\boldsymbol{X})$ heißt Bayesscher Punktschätzer für θ.

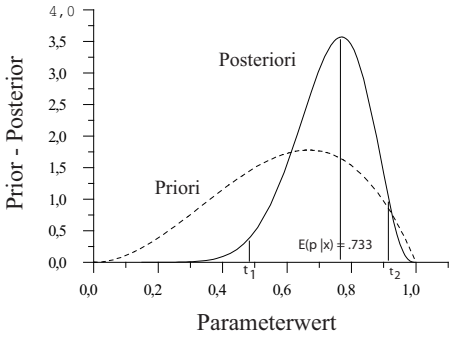
Anmerkung: Die Begriffe des Posteriori-Modus und des Posteriori-Medians sind analog definiert:

$$\text{Mod}(\theta|\mathbf{X}) = \max_\theta f(\theta|\mathbf{X}) \text{ Posteriori-Modus} \tag{30}$$

$$\text{Med}(\theta|\mathbf{X}) = m, \text{ Posteriori-Median, mit} \tag{31}$$

$$\int_{-\infty}^{m} f(\theta|\mathbf{X})d\mathbf{X} = \int_m^{\infty} f(\theta|\mathbf{X})d\mathbf{X} = \frac{1}{2}.$$

Abbildung 1: Priori Bet(3, 2) und Posteriori für Binomialverteilung Bin(10, 8)



2.1.1 Konfidenz und Kredibilität

In der orthodoxen Statistik ist es üblich, ein Konfidenzintervall für eine Parameterschätzung $\hat{\theta}$ anzugeben. Dies ist ein Intervall $(\hat{\theta}-a, \hat{\theta}+a)$, das mit einer Wahrscheinlichkeit von $1-\alpha$ den wahren Wert von θ enthält. In der Bayes-Statistik entspricht dem Konfidenzintervall das Kredibiltätsintervall:

Definition 2.4 *Es sei* $\alpha \in (0,1)$. *Das* Kredibilitätsintervall *ist das Intervall* (t_1, t_2), *über dem das Integral der Funktion* $f(\theta|\boldsymbol{X})$ *in Bezug auf* $\boldsymbol{X}$ *den Wert* $1-\alpha$ *hat:*

$$\int_{t_1}^{t_2} f(\theta|\boldsymbol{X})\, d\theta = 1-\alpha. \tag{32}$$

Die Wahrscheinlichkeit $1-\alpha$ *heißt* Kredibilitätsniveau.

Anders als das Konfidenzintervall[6] ist das Kredibilitätsintervall direkt zu interpretieren: der gesuchte Parameterwert liegt mit der Wahrscheinlichkeit $1-\alpha$ zwischen den Werten t_1 und t_2. Üblicherweise wählt man t_1 als das $\alpha/2$-Quantil der Posteriori-Verteilung, und t_2 als das $1-\alpha/2$-Quantil dieser Verteilung. Dies führt zu dem folgenden Begriff.

Definition 2.5 *Das Intervall* (θ_1, θ_2) *heißt HPD-Intervall (*highest posterior density interval*, wenn für alle* $\theta \in (\theta_1, \theta_2)$ *die Relation*

$$f(\theta|x) \geq f(\tilde{\theta}|x), \quad \textit{für alle } \tilde{\theta} \notin (\theta_1, \theta_2) \tag{33}$$

[6]Das Konfidenzintervall enthält mit der Wahrscheinlichkeit $1-\alpha$, d.h. in $(1-\alpha)100\%$ der Untersuchungen, die genau so wie die vorliegende durchgeführt wurden, den wahren Wert des geschätzten Parameters.

gilt.

Das Intervall (t_1, t_2) ist dann der Bereich, in dem die Posterior ihre maximalen Werte annimmt; die Rede ist dann auch von *highest density regions* (HDRs).

2.1.2 Test von Hypothesen

Es sei $H_0 : \theta \in \Theta_0$, $H_1 : \theta \in \Theta_1$. Die Posterioris sind dann

$$P_0 = P(\theta \in \Theta_0|\mathbf{x}), \quad P_1 = P(\theta \in \Theta_1|\mathbf{x}).$$

Dabei ist $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cap \Theta_1 = \Theta$, so dass

$$P_0 + P_1 = 1.$$

Die Priori- Wahrscheinlichkeiten seien $\pi_0 = P(H_0)$, $\pi_1 = P(H_1)$.

Definition 2.6 *Der* Bayes-Faktor *für H_0 versus H_1 ist durch*

$$B = \frac{P_0/P_1}{\pi_0/\pi_1} \tag{34}$$

definiert.

Der Bayes-Faktor kann gelegentlich direkt zur Bewertung von Hypothesen herangezogen werden. Aus (34) folgt sofort

$$\frac{P_0}{P_1} = \frac{\pi_0}{\pi_1} B,$$

und da $P_1 = 1 - P_0$ folgt

$$P_0 = (1 - P_0)\frac{\pi_0}{\pi_1} B, \text{ d.h. } P_0(1 + (\pi_0/\pi_1)B) = (\pi_0/\pi_1)B,$$

so dass, wegen $\pi_1 = 1 - \pi_0$,

$$P_0 = \frac{1}{1 + ((1 - \pi_0)/\pi_0)B^{-1}}. \tag{35}$$

Damit ist die Wahrscheinlichkeit von H_0 durch die Prior von H_0 und den Bayes-Faktor ausgedrückt.

Einfache Hypothesen: Es sei $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$. Dann ist

$$\frac{P_0}{P_1} = \frac{\pi_0}{\pi_1}\frac{P(\mathbf{x}|\theta_0)}{P(\mathbf{x}|\theta_1)}, \tag{36}$$

und der Bayes-Faktor ist in diesem Fall durch den Likelihood-Quotienten

$$B = \frac{P(\mathbf{x}|\theta_0)}{P(\mathbf{x}|\theta_1)}$$

gegeben.

Zusammengesetzte Hypothesen: Es sei

$$\rho_0(\theta) = \frac{P(\theta)}{\pi_0}, \text{ für } \theta \in \Theta_0 \tag{37}$$

$$\rho_1(\theta) = \frac{P(\theta)}{\pi_1}, \text{ für } \theta \in \Theta_1 \tag{38}$$

$P(\theta)$ ist, wie üblich, die Prior für H_0; dann ist ρ_0 die Einschränkung von $P(\theta)$ auf Θ_0, und ρ_1 ist die Einschränkung von $P(\theta)$ auf Θ_1. Dann folgt

$$\begin{aligned} P_0 &= P(\theta \in \Theta_0|\mathbf{x}) = \int_{\Theta_0} P(\theta|\mathbf{x})\, d\theta \\ &\propto \int_{\Theta_0} P(\theta)P(\mathbf{x}|\theta)\, d\theta = \pi_0 \int_{\Theta_0} P(\mathbf{x}|\theta)\rho_0\, d\theta \end{aligned} \tag{39}$$

Für die Posteriori P_1 findet man analog

$$P_1 \propto \pi_1 \int_{\Theta_1} P(\mathbf{x}|\theta)\rho_1(\theta)\, d\theta, \tag{40}$$

und der Bayes-Faktor ist

$$B = \frac{P_0/P_1}{\pi_0/\pi_1} = \frac{\int_{\Theta_0} P(\mathbf{x}|\theta)\rho_0(\theta)\, d\theta}{\int_{\Theta_1} P(\mathbf{x}|\theta)\rho_1(\theta)\, d\theta}. \tag{41}$$

Im Falle zusammengesetzter Hypothesen ist der Bayes-Faktor der Quotient gewichteter Likelihoods.

Wegen der "Gewichte" ρ_i, $i = 1, 2$ hängt B im Falle zusammengesetzter Hypothesen nicht nur von den Daten ab und ist insofern kein *allgemeines* Maß für die Stützung einer Hypothese, das nur von den Daten abhängt. In manchen Situationen hängt B kaum von den ρ_i ab, und dann kann B als Maß der relativen, nur von Daten abhängenden Stützung (support) für eine Hypothese gesehen werden.

Beispiel 2.1 s. Lee, p. 120 □

Konkrete Tests setzen die Wahl von Prior-Verteilungen voraus. Einige der mit dieser Wahl verbundenen Fragen werden im folgenden Abschnitt behandelt.

2.2 Typen von a-priori-Verteilungen

2.2.1 Das Indifferenzprinzip

Hat man spezielle Kenntnisse über die Wahrscheinlichkeiten der Hypothesen, kann man eine Verteilung finden, die diesen Kenntnissen entspricht. Häufig sind solche Kenntnisse nicht vorhanden. Es liegt dann nahe, alle Hypothesen – d.h. alle möglichen θ-Werte – als gleichwahrscheinlich zu betrachten; dies ist das Indifferenzprinzip, auch bekannt als (*Prinzip des Unzureichenden Grundes* (Bayes-Laplace)).

Die Annahme einer Gleichverteilung erscheint auf den ersten Blick durchaus sinnvoll zu sein. Soll man die Augenzahl raten, die ein Würfen nach einem Wurf "oben" zeigt, und geht man davon aus, dass der Würfel fair ist, so erscheinen alle Augenzahlen als gleichwahrscheinlich. Beobachtet man eine Folge von n Bernoulli-Versuchen, hat aber keine Information über den Parameter $\theta = p$ der Binomialverteilung, so liegt es nahe, θ als auf $(0, 1)$ gleichverteilt anzunehmen. Allgemein ist die Gleichverteilung auf einem Intervall (a, b) durch

$$F(x) = \frac{x-a}{b-a}, \quad f(x) = \frac{1}{b-a} \tag{42}$$

definiert. Für den Binomialparameter bedeutet dies

$$f(\theta) = \frac{1}{1-0} = 1 \text{ für alle } \theta.$$

Nun seien die beobachteten Ereignisse aber Poisson-verteilt, d.h.

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{43}$$

und $\lambda > 0$, d.h. $\lambda \in (0, \infty)$. Weiß man nichts über den Wert von λ und möchte deshalb eine Gleichverteilung als Priori annehmen, so findet man

$$f(\theta) = f(\lambda) = \frac{1}{\infty - 0} \equiv 0,$$

so dass die Posteriori stets gleich Null ist. Dies kann nicht sein, also kann die Annahme einer Gleichverteilung nicht korrekt sein. Man könnte fordern, statt der unbestimmten oberen Grenze ∞ einen endlichen Wert zu betrachten, – doch dann ergibt sich sofort die Frage, welchen Wert man denn annehmen möchte.

Aber selbst wenn der in Frage stehende Parameter nicht auf $[0, \infty)$ verteilt ist, kann die Annahme der Gleichverteilung auf unangenehme Fragen führen. So sei X irgendeine zufällige Veränderliche; insbesondere kann $X = \theta$ gelten. X sei auf $[a, b]$ gleichverteilt, so dass

$$F(x) = \frac{b-x}{b-a}.$$

Weiter sei $Y = 1/X$. Dann ist Y nicht auf einem geeignet gewählten Intervall gleichverteilt:

$$P(Y \leq y) = P(1/X \leq y) = P(1/y \leq X) = 1 - P(X \leq 1/y),$$

und $dP(Y \leq y)/dy$ liefert die Dichtefunktion von Y:

$$g(y) = \frac{dP(Y \leq y)}{dy} = \frac{b - 1/y}{b - a} \frac{1}{y^2}. \tag{44}$$

Die Gleichverteilung als Ausdruck des Indifferenzprinzips besagt, dass man keine spezifische Kenntnis über die möglichen Werte von X hat. Da $Y = 1/X$ eine deterministische Funktion von X ist, kann man auch keine spezifische Kenntnis über die möglichen Y-Werte haben. Aber dieser Folgerung widerspricht (44). Die folgenden Beispiele illustrieren diese Widersrpüchlichkeit.

1. **Das Geschwindigkeitsparadox:** Man möchte etwa die Geschwindigkeit, mit der bestimmte Bewegungen durchgeführt werden, bestimmen. Man hat zwei Möglichkeiten: (i) man mißt die Zeit, die benötigt wird, um eine Bewegung einer bestimmten Länge durchzuführen, oder (ii) man mißt die Strecke, die in einer vorgegebenen Zeit zurückgelegt wird. Man könnte als a-priori-Verteilung etwa eine Gleichverteilung für die Zeit annehmen, – aber ebenso gut auch für die Strecke. Nimmt man die Gleichverteilung für die Zeit an, so kann die Verteilung für die Strecken nicht gleichverteilt sein, und umgekehrt. Die zufälligen Veränderlichen X für die Zeit und Y für die Strecke sind reziprok zueinander.

2. **Das Wein-Wasser-Paradox:** Dieses Paradoxon geht auf R. von Mises zurück (v. Mises (1981), p. 77). Es sei ein Krug mit einer Mischung von Wasser und Wein gegeben. Das genaue Verhältnis von Wein und Wasser ist nicht bekannt, aber man weiß, dass der Anteil einer der beiden Substanzen höchstens dreimal so groß wie der der anderen ist. Ist also X das Verhältnis von Wein zu Wasser, so muß $X \geq 1/3$ sein, andernfalls wäre der Anteil von Wasser mehr als dreimal so groß wie der des Weins. Ebenso muß $X \leq 3$ gelten, sonst wäre der Anteil des Weins mehr als dreimal so groß wie der des Wassers. Also muß gelten

   $$\frac{1}{3} \leq X \leq 3, \quad \frac{1}{3} \leq Y \leq 3. \tag{45}$$

   mit $Y = 1/X$, und der rechte Ausdruck ergibt sich durch eine analoge Argumentation. Weiß man nichts über das tatsächliche Verhältnis von Wein und Wasser, außer den Bedingungen (45), so führt das Prinzip der Indifferenz auf eine Gleichverteilung für X auf $[1/3, 3]$. Aber dann ist Y nach

(44) nicht gleichverteilt. Andererseits kann man ebenso gut annehmen, Y sei auf $[1/3, 3]$ gleichverteilt. Aber dann kann X nicht mehr gleichverteilt sein. In der üblichen Formulierung des Paradoxons wird gezeigt, dass die Annahme der Gleichverteilung sowohl für X als auch für Y auf widersprüchliche Ergebnisse führt, was nach den vorangegangenen Überlegungen nicht verwunderlich ist: so werde etwa nach der Wahrscheinlichkeit $P(X \leq 2)$ gefragt. Es ist

$$P(X \leq 2) = P(1/Y \leq 2) = P(1/2 \leq Y). \tag{46}$$

Nimmt man nun sowohl für X als auch für Y eine Gleichverteilung an, so erhält man einerseits

$$P(X \leq 2) = \frac{2 - 1/3}{3 - 1/3} = \frac{5}{8},$$

und andererseits

$$P(Y \geq 1/2) = \frac{3 - 1/2}{3 - 1/3} = \frac{15}{16},$$

also $P(X \leq 2) \neq P(Y \geq 1/2)$, in Widerspruch zu (44).

Der Widerspruch zwischen $P(X \leq 2) = P(Y \geq 1/2)$ einerseits und $P(X \leq 2) \neq P(Y \geq 1/2)$ andererseits wird im Allgemeinen dem Indifferenzprinzip angelastet. Keynes versuchte, den Widerspruch zu überwinden, indem er forderte, es dürfe nur endlich viele, nicht weiter teilbare Alternativen geben; es läßt sich aber zeigen, dass dieses Postulat nicht aufrechtzuerhalten ist. Van Fraassen (1989) hält das Wein-Wasser-Paradoxon für "the ultimate defeat" des Indifferenzprinzips, Gillies (2000a) spricht von einem "tödlichen" Argument gegen dieses Prinzip, und Oakes (1986) folgert aus dem Paradoxon, dass dieses die klassische Konzeption der Wahrscheinlichkeit überhaupt ins Wanken bringe.

Deakin diskutiert mögliche Lösungen für das Paradox.

Die Beispiele implizieren nicht, dass die Gleichverteilung grundsätzlich nicht als Priori gewählt werden kann. Sie zeigen nur, dass es nicht notwendig klar ist, welche Variable als gleichverteilt anzunehmen ist. Jedenfalls führte die Problematik der Gleichverteiltung zu Versuchen, das Indifferenzprinzip durch andere Verteilungen zu repräsentieren. Diese Versuche führen auf den Begriff der nichtinformativen Verteilungen. Bevor auf diese Verteilungen eingegangen werden kann, müssen einige Begriffe eingeführt werden.

2.2.2 Likelihood, Scores und Information

Es sei $x = (x_1, \ldots, x_n)$ eine Stichprobe mit $x_i \sim f(x|\theta)$ für alle i.

Likelihood Die Likelihood ist dann

$$l(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta). \tag{47}$$

Log-Likelihood Die Log-Likelihood ist

$$L(\theta|x) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta). \tag{48}$$

Score-Funktion Die *Score-Funktion* ist durch

$$S(\theta) = \frac{dL(\theta|x)}{d\theta} = \sum_{i=1}^{n} \frac{1}{f(x_i|\theta)} \frac{\partial f(x_i|\theta)}{\partial \theta.} \tag{49}$$

definiert; die Maximum-Likelihood(ML)-Schätzung $\hat{\theta}$ von θ ist die Lösung der Gleichung $S(\theta) = 0$.

Fisher-Information Die Fisher-Information ist durch

$$I(\theta|x) = -\frac{dS(\theta)}{d\theta} = -\frac{\partial^2 l(\theta|x)}{\partial \theta^2} \tag{50}$$

definiert. Der Erwartungswert bezüglich x liefert die *erwartete Fisher-Information*

$$\mathbb{E}_x(I(\theta|x)) = -\mathbb{E}\left(\frac{\partial^2 l(\theta|x)}{\partial \theta^2}\right) \tag{51}$$

Die Bedeutung der Fisher-Information bzw. der erwarteten Fisher-Information wird anhand der folgenden Beispiele erläutert.

Beispiel 2.2 Binomial-Verteilung Die Stichprobe $x = (x_1, \ldots, x_m)$ bestehe aus Häufigkeiten aus einem Bernoulli-Experiment, d.h. $x_i \sim B(\theta, n)$. Die Likelihood der Daten ist

$$l(\theta|x) = \prod_{i=1}^{m} \binom{n}{x_i} \theta^{x_i} (1-\theta)^{n-x_i}$$

Die Log-Likelihood ist

$$L(\theta|x) = \sum_{i=1}^{m} x_i \log \theta + (n - x_i) \log(1-\theta) - \log \binom{n}{x_i},$$

und die Score-Funktion ergibt sich als

$$S(\theta) = \sum_{i=1}^{m} \left(\frac{x_i}{\theta} - \frac{n - x_i}{1-\theta}\right).$$

Die ML-Schätzung für θ ist die Lösung von

$$\frac{1}{\theta}\sum_{i=1}^{m} x_i = \frac{1}{1-\theta}\sum_{i=1}^{m}(n - x_i),$$

woraus sich

$$\hat{\theta} = \frac{1}{m \cdot n}\sum_{i=1}^{m} x_i \tag{52}$$

ergibt. Für $m = 1$ erhält man die übliche relative Häufigkeit x/n als Schätzung für θ.

Die Fisher-Information ist

$$I(\theta|x) = -\frac{dS(\theta)}{d\theta} = -\sum_{i=1}^{m}\left(-\frac{x_i}{\theta^2} + \frac{n - x_i}{(1-\theta)^2}\right) = \sum_{i=1}^{m}\left(\frac{x_i}{\theta^2} - \frac{n - x_i}{(1-\theta)^2}\right).$$

Der Erwartungswert von x_i ist $\mathbb{E}(x_i) = n\theta$ für alle i. Für die erwartete Fischer-Information erhält man demnach

$$J(\theta) = \frac{m \cdot n\theta}{\theta^2} + \frac{m \cdot n}{(1-\theta)^2} - \frac{m \cdot n\theta}{(1-\theta)^2},$$

woraus sich

$$J(\theta) = \frac{m \cdot n}{\theta(1-\theta)} \tag{53}$$

ergibt.

Es sei $m = 1$. Es sei ξ_j die Bernoulli-Variable beim j-ten Versuch, und $x = \sum_j \xi_j$.

x ist dann die Anzahl der "Erfolge" (Summe der Bernoulli-Variablen), und $\hat{\theta} = x/n = \bar{\xi}$; d.h. $\hat{\theta}$ entspricht dem arithmetischen Mittel der Bernoulli-Variablen ξ_j. Es ist $\mathbb{V}(\xi) = \theta(1-\theta)$, $\mathbb{V}(x) = n\theta(1-\theta)$ und

$$\mathbb{V}(\bar{\xi}) = \mathbb{V}\left(\frac{\sum_j \xi_j}{n}\right) = \frac{1}{n^2} n\theta(1-\theta),$$

d.h.

$$\mathbb{V}(\bar{\xi}) = \frac{\theta(1-\theta)}{n} = \frac{1}{J(\theta)}. \tag{54}$$

Im Fall $m = 1$ ist die erwartete Fisher-Information gerade gleich der reziproken Varianz der mittleren Bernoulli-Variablen, also der Schätzung $\hat{\theta}$ des unbekannten Parameters θ.

□

Beispiel 2.3 Es sei $X \sim N(\mu, \sigma^2)$, σ^2 sei bekannt, und $\theta = \mu$. Dann ist die Likelihood durch

$$L(\mu) = \frac{1}{\sigma^n (2\pi^{n/2}} \exp\left[-\frac{1}{2}\sum_i (x_i - \mu)^2\right]$$

gegeben, und die Score-Funktion durch

$$V = \frac{\partial}{\partial \mu}\left[\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 - \log(\sigma^n (2\pi)^{n/2}\right] = -\frac{1}{\sigma^2}\sum_i (x_i - \mu).$$

Dass $\mathbb{E}(V) = 0$, ist sofort einsichtig. Die Fisher-Information ist dann durch

$$\mathcal{I}(\theta) = \mathbb{V}(V) = \mathbb{E}\left[\left(\frac{\partial \log L}{\partial \mu}\right)^2\right] = \mathbb{E}\left[\frac{\partial^2 \log L}{\partial \mu^2}\right]$$

gegeben. Es ist

$$\mathbb{E}\left[\left(\frac{\partial \log L}{\partial \mu}\right)^2\right] = \frac{1}{\sigma^4}\mathbb{E}\left[\left(\sum_i (x_i - \mu)\right)^2\right],$$

und

$$\left(\sum_i (x_i - \mu)\right)^2 = \sum_i (x_i - \mu)^2,$$

wegen der Unshängigkeit der x_i, und mithin

$$\frac{1}{\sigma^4}\mathbb{E}\left[\left(\sum_i (x_i - \mu)\right)^2\right] = \frac{1}{\sigma^4}\mathbb{E}\left[\sum_i (x_i - \mu)^2\right] = \frac{1}{\sigma^4}\sum_i \mathbb{E}(x_i - \mu)^2 = \frac{n\sigma^2}{\sigma^4},$$

d.h.

$$\mathcal{I}(\theta) = \frac{n}{\sigma^2}. \tag{55}$$

Ebenso folgt

$$\mathbb{E}\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] = \mathbb{E}\left[\frac{\partial}{\partial \mu}\left(-\frac{1}{\sigma^2}\sum_i x_i + \frac{n\mu}{\sigma^2}\right)\right] = \mathbb{E}\left[\frac{n}{\sigma^2}\right] = \frac{n}{\sigma^2},$$

also natürlich das gleiche Resultat wie (55): Die Information in der Stichprobe bezüglich des Parameters μ ist proportional zum Stichprobenumfang n und umgekehrt proportional zur Varianz σ^2 der Daten. □

2.2.3 Nichtinformative Priori-Verteilungen

Nichtinformative Prioris sollen mangelndes Vorwissen ausdrücken – die Gleichverteilung selbst ist eine typische nichtinformative Priori.

Der Übergang zu $Y = 1/X$ kann als eine Transformation von X gesehen werden. Andere Transformationen sind denkbar, etwa Skalentransformationen $Y = bX + a$. Welche Priori sollte etwa für den Parameter σ^2 einer Gauß-Verteilung gewählt werden, wenn man keine spezifischen Kenntnisse über die Varianz der betrachteten zufälligen Veränderlichen – etwa X – hat? Für $Y = bX + a$ geht σ_x^2 in $\sigma_y^2 = b^2\sigma_x^2$ über. Eine nichtinformative Priori für σ^2 sollte in Bezug auf derartige Transformationen ihre Nichtinformativität behalten.

Es sei nun ganz allgemein $\varphi = h(\theta)$ eine umkehrbar eindeutige Transformation der Variablen θ. Dann ist

$$P(h(\theta) \leq \varphi_0) = \begin{cases} P(\theta \leq h^{-1}(\varphi_0)) & h \text{ monoton wachsend} \\ \\ P(\theta > h^{-1}(\varphi_0), & h \text{ monoton fallend} \end{cases}$$

Für die Dichte folgt dann

$$f(\varphi) = f(h^{-1}(\varphi)) \left| \frac{dh^{-1}(\varphi)}{d\varphi} \right|. \tag{56}$$

Ist die Dichte für θ eine Konstante, so ergibt sich für die Transformation nur dann eine Konstante, wenn h linear ist. Ist h nichtlinear, so ist $f(\varphi)$ nicht mehr konstant. Damit gerät man in einen Widerspruch, denn dann hätte man ja für θ gleich eine nicht-konstante Priori wählen können.

Dieser Widerspruch scheint auf ein Problem mit der Gleichverteilung zu verweisen. Andererseits hat man eine verwandte Situation mit jeder Prior-Verteilung. Denn es sei $F(x) = P(X \leq x)$ eine Verteilungsfunktion, deren Dichte $f(x) = dF(x)/dx$ als Priori-Dichte verwendet werden soll. Weiter sei $Y = \phi(X)$ eine invertierbare Funktion von X, d.h. ϕ^{-1} soll für alle X existieren; insbesondere sei ϕ streng monoton wachsend. Dann ist

$$G(y) = P(Y \leq y) = P(\phi(X) \leq y) = P(X \leq \phi^{-1}(y) = F(\phi^{-1}(y)).$$

Im Allgemeinen ist $F(y) \neq G(y) = F(\phi^{-1}(y))$. Die Information, die in der Wahl von F bezüglich der Verteilung von X ausgedrückt wird, ist einerseits wegen der deterministischen Kopplung ϕ von Y and X zunächst die Gleiche, die für die Y-Werte zur Verfügung steht, andererseits unterscheiden sich die Verteilungsfunktion G von F.

2.2.4 Konjugierte Priors

Prioris dieser Klasse haben die Eigenschaft, dass die zugehörige Posteriori wieder zur gleichen Klasse gehört. Die beiden Verteilungen heißen *konjugiert bezüglich der Likelihood-Funktion*

Beispiel 2.4 Binomialverteilung: Die zufällige Veränderliche X sei binomialverteilt, d.h. $X \sim B(n, \theta)$. Dann ist

$$P(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \propto \theta^x(1-\theta)^{n-x}. \tag{57}$$

Als Prior werde die Beta-Funktion angenommen:

$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{58}$$

Der Erwartungswert ist

$$\mathbb{E}(\theta) = \frac{a}{a+b}. \tag{59}$$

Dann ist die Posterior durch

$$P(\theta|x) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}, \tag{60}$$

d.h. die Posterior ist, wie die Prior, eine Binomialverteilung $\theta|x \sim \text{Bin}(n, \theta)$. Der Erwartungswert für $\theta|x$ ist dann

$$\mathbb{E}(\theta|x) = \frac{a+x}{a+b+n}, \tag{61}$$

un der Modus ist

$$\text{Mod}(\theta|x) = \frac{a+x-1}{a+b+n-2}. \tag{62}$$

Der Erwartungswert läßt sich in der folgenden Weise darstellen:

$$\mathbb{E}(\theta|x) = \frac{a+b}{a+b+n} \cdot \frac{a}{a+b} + \frac{n}{a+b+n} \cdot \frac{x}{n} \tag{63}$$

Hierin ist $\hat{\theta} = x/n = \bar{x}$ der ML-Schätzer für θ, und nach (63) läßt sich $\mathbb{E}(\theta|x)$ als gewogenes Mittel des a-priori-Erwartungswertes $a/(a+b)$ und des ML-Schätzers $\bar{x}$ darstellen. Dies ist ein Spezialfall eines allgemeineren Befundes. Man sieht leicht, dass

$$\lim_{n\to\infty} \frac{a+b}{a+b+n} = 0, \quad \lim_{n\to\infty} = \frac{n}{a+b+n} = \lim_{n\to\infty} \frac{1}{(a+b)/n+1} = 1,$$

d.h. je größer der Wert von n, desto größer ist das Gewicht des ML-Schätzer $\bar{x}$ in $\mathbb{E}(\theta|x)$.

□

Beispiel 2.5 Poisson-Verteilung Die zufällige Veränderliche X sei Poisson-verteilt, d.h.

$$P(X = x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad \lambda > 0.$$

Es liege eine Stichprobe $(x_1, \ldots, x_n)$ vor; die x_i sind Häufigkeiten. Die Likelihood der Stichprobe ist dann

$$L(\lambda|x) = \prod_{i=1}^{n} e^{\cdot\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}, \tag{64}$$

und die Log-Likelihood-Funktion ist

$$l(\lambda|x) = -n\lambda + \log\lambda\sum_{i=1}^{n} x_i - \log\prod_{i=1}^{n} x_i! \tag{65}$$

Die Score-Funktion ist

$$S(\lambda) = \frac{dl(\lambda|x)}{d\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i, \tag{66}$$

woraus sich sofort die ML-Schätzung

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{67}$$

ergibt.

Gesucht ist nun die relativ zur Likelihood-Funktion konjugiert Prior-Verteilung. Es zeigt sich, dass die Gammaverteilung

$$f(\lambda) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}, \quad \Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt \tag{68}$$

die konjugierte Priori ist; a und b sind hier die *Hyperparameter*. Schreibt man

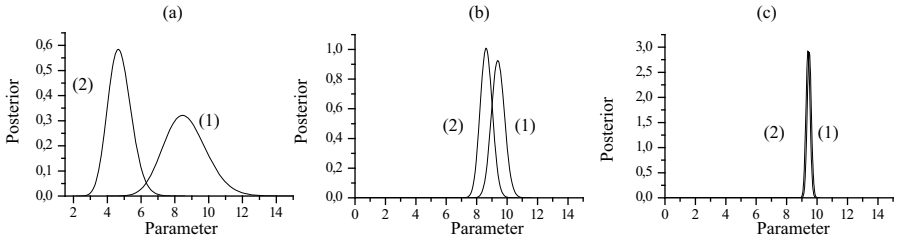
$$L(\lambda) \propto e^{-n\lambda}\lambda^N, \quad N = \sum_{i=1}^{n} x_i,$$

so erhält man für die Posteriori

$$P(\lambda|x) \propto e^{-2n\lambda + a\log b}\lambda^{N+a-1}, \quad N = \sum_{i=1}^{n} x_i, \tag{69}$$

und dies ist, bis auf die Normierungskonstante, eine Poisson-Verteilung. Gelegentlich

Abbildung 2: Posterioris für die Poisson-Verteilung: Prior (1) χ^2-Verteilung; (a) $n = 5$, (b) $n = 50$, (c) $n = 500$, (2) Gamma-Verteilung, $a = 1, b = 5$, . "Wahrer" λ-Wert $\lambda = 9$.



wird die χ^2-Verteilung als konjugierte Priori für die Poisson-Verteilung angegeben (etwa in Lee, 1997). Die χ^2-Verteilung ist aber ein Spezialfall der Gamma-Verteilung. Dazu betrachte man die Dichte der χ^2-Verteilung: es ist $\chi^2 = z_1^2 + \cdots + z_n^2$ und mit $\chi^2 = x$ hat man

$$f(x) = \frac{1}{2^n \Gamma(n/2)} x^{n/2-1} e^{-x/2}. \tag{70}$$

Setzt man hierin $x = \lambda$, $a = n/2$ und $b = 1/2$, so wird (70) zu (68). Die Posteriori nimmt dann die Form

$$P(\lambda|x) \propto e^{-(n+1/2)\lambda} \lambda^{N+n/2-1} \tag{71}$$

an. Abbildung 2 zeigt die Posteriori-Verteilungen für Priori (1): χ^2-Verteilung ($a = n/2$ und $b = 1/2$), und (2): Gamma-Verteilung mit $a = 1$ und $b = 5$. Der "wahre" λ-Wert ist $\lambda = 9$. Die zur Priori-χ^2-Verteilung korrespondierende Posteri hat einen Modus, der schon für (a) $n = 5$ nahe am wahren Wert liegt; dies rechtfertigt den Fokus auf den Spezialfall der χ^2-Verteilung. Die Werte $a = 1$ und $b = 5$ der Gamma-Verteilung bewirken eine Verschiebung auf der λ-Skala, d.h. man kann über diese Parameter Annahmen über die Lokation der λ-Verteilung eingehen lassen. Mit wachsendem Wert von n rückt die Gamma-Verteilung aber an die χ^2-Verteilung heran, bis die Verteilungen bei einem hinreichend großen Wert nahezu identisch werden. Dies illusriert noch einmal den Sachverhalt, dass die empirische Evidenz am Ende die Schätzungen immer gegen die wahren Wert konvergieren lassen, auch wenn die ersten Priori-Annahmen weit von diesem abweichen. □

Beispiel 2.6 (**Gauss-Prior**) Es sei nun $x \sim N(\theta, \sigma^2)$, d.h.

$$h(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \tag{72}$$

Für die Prior gelte

$$f(\theta) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(\theta-\theta_0)^2}{2\sigma_0^2}\right) \tag{73}$$

Weiter sei $\mathbf{X} = (x_1, \ldots, x_n)$ eine Stichprobe unabhängiger Messwerte. Die Likelihood ist dann

$$\begin{aligned} l(\theta) &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2\right)\right) \\ &\propto \exp\left(-\frac{n}{2\sigma^2}(\theta-\bar{x})^2\right) \end{aligned} \tag{74}$$

Für die Posteriori ergibt sich nun

$$\begin{aligned} P(\theta|\mathbf{X}) &\propto L(\theta)f(\theta) \\ &= \exp\left(-\frac{1}{2}\left(\frac{n}{\sigma^2}(\theta-\bar{x})^2 + \frac{1}{\sigma_0^2}\right)(\theta-\theta_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left[\theta - \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{n\bar{x}}{\sigma^2}+\frac{\theta_0}{\sigma_0^2}\right)\right]^2\right)\right) \end{aligned} \tag{75}$$

Also folgt

$$\theta|\mathbf{X} \sim N\left(\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{n\bar{x}}{\sigma^2}+\frac{\theta}{\sigma_0^2}\right), \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\right) \tag{76}$$

Der Erwartungswert ist also ein gewogenes Mittel aus Prior-Erwartungswert und der Maximum-Likelihood-Schätzung $\bar{x}$. Je größer der Wert von n (Stichprobenumfang), desto größer ist das Gewicht des ML-Schätzers $\bar{x}$. Die Posteriori-Varianz $1/(n/\sigma^2 + 1/\sigma_0^2)$ wird mit wachsendem n kleiner.

Man kann den Begriff der *Präzision* einbringen:

$$\kappa = \frac{1}{\sigma^2}, \quad \lambda = \frac{1}{\sigma_0^2}. \tag{77}$$

Dann kann $\theta|\mathbf{X}$ in der Form

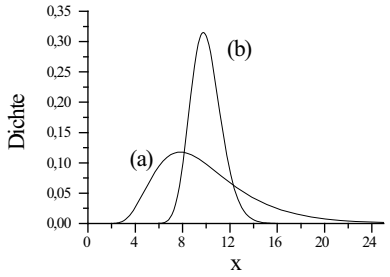
$$\theta|\mathbf{X} \sim N\left(\frac{n\kappa\bar{x}}{n\kappa+\lambda}, \frac{1}{(n\kappa+\lambda}\right) \tag{78}$$

angeschrieben werden. □

Tabelle 1: Konjugierte Verteilungen zu bestimmten Likelihood-Funktionen

Likelihood	Konjugierte Verteilungsklasse
$\mathrm{Bin}(n, \pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$
$\mathrm{Geom}(\pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$
$\mathrm{Poisson}(\lambda)$	$\lambda \sim \mathrm{G}(\alpha, \beta)$
$\mathrm{Exp}(\lambda)$	$\lambda \sim \mathrm{G}(\alpha, \beta)$
$\mathrm{N}(\mu, \sigma^2)$, σ^2 bekannt	$\mu \sim N(\nu, \tau^2)$
$\mathrm{N}(\mu, \sigma^2)$, μ bekannt	$\sigma^2 \sim \mathrm{IG}(\alpha, \beta^2)$

Abbildung 3: Inverse Gauß-Verteilungen: $\mu = 10$, (a) $\lambda = 60$, (b) $\lambda = 600$.



Die inverse Gauß-Verteilung ist durch

$$f(x; \mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right), \quad 0 < x < \infty \tag{79}$$

gegeben; Erwartungswert und Varianz sind

$$\mathbb{E}(X) = \mu, \quad \mathbb{V}(X) = \frac{\mu^3}{\lambda}, \tag{80}$$

vergl. Abbildung 3.

2.2.5 Uneigentliche Priori-Verteilungen

Wenn wenig über den Parameter bekannt ist, möchte man i. A. den Einfluß der Priori auf die Posteriori so gering wie möglich halten. Die Gleichverteilung als Priori hat gelegentlich ihre Tücken. Andererseits könnte man eine Gauß-Verteilung mit sehr großer Varianz wählen. Im Extremfall führt dies zu Priori-

Tabelle 2: Faustregel zur Bewertung von Bayes-Faktoren (nach Held (2008), p. 218)

Stufe	Bayes-Faktor B_{12}	Beweiskraft für M_1 gegen M_2
1	1 bis 3	kaum der Rede wert
2	3 bis 20	positiv
3	20 bis 150	stark
4	ab 150	stark

Funktionen, die nicht mehr integrierbar sind (man erinnere sich: das Integral einer Wahrscheinlichkeitsdichte ist stets gleich 1).

So sei (73) die Priori-Verteilung. Damit für $\theta = \mu$ der gesamte Bereich $\mathbb{R}$ zur Verfügung steht, werde $\theta = 0$ gesetzt (positive wie negative Werte treten dann mit der Wahrscheinlichkeit 1/2 auf). Für $\sigma_0^2 \to \infty$ wird dann aus der Dichte eine Konstante über $(-\infty, \infty)$.

Definition 2.7 *Die Dichte $f(\theta) > 0$ heißt* uneigentliche Priori[7]*, wenn*

$$\int_\Theta f(\theta)\, d\theta = \infty, \quad bzw. \sum_{\theta \in \Theta} f(\theta) = \infty. \tag{81}$$

Beispiel 2.7 (**Haldane-Priori**) Für die Binomialverteilung konnte die Beta-Funktion $B(\alpha, \beta)$ als Priori-Verteilung gewählt werden, die die Gleichverteilung als Spezialfall zuläßt. Für $\alpha = \beta = 0$ erhält man die *Haldane-Priori*

$$f(\theta) \propto \theta^{-1}(1-\theta)^{-1}, \tag{82}$$

nach J.B.S. Haldane[8], der sie zuerst für die Binomialverteilung betrachtete. Sie liefert als Posteriori-Verteilung eine Bet$(k, n-k)$-Verteilung. Wie Abb. 4 zeigt, approximiert die Haldane-Prior über einen großen Teil des Intervalls (0, 1) eine Gleichverteilung; den Werten an den Endpunkten des Intervalls wird aber großes Gewicht beigemessen. □

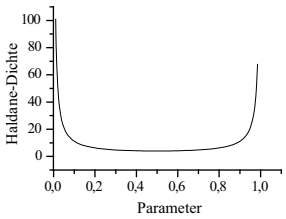
2.2.6 Jeffreys' Prior:

Jeffreys (1939/1961/2003) hat das Problem der nicht-informativen Prioris bereits ausgiebig diskutiert. Er ging davon aus, dass Prioris invariant gegenüber

[7] improper prior

[8] John Burdon Sanderson Haldane (1892 – 1964), theoretischer Biologe und Genetiker, Mitbegründer der Populationsgenetik; Sohn von Scott Haldane, Physiologe

Abbildung 4: Haldane-Priori



umkehrbar eindeutigen Transformation von θ sein sollten.

Es sei $L(\mathbf{X}|\theta)$ die Likelihood für $\mathbf{X}$. Bekanntlich ist nach der Cramér-Rao-Ungleichung

$$\mathbb{V}(\hat{\theta}) \geq \frac{1}{I(\theta)}, \tag{83}$$

wobei

$$J(\theta) = \mathbb{E}\left[\left(\frac{\partial \log L}{\partial \theta}\right)^2\right] \tag{84}$$

die *Fisher-Information* ist. Jeffreys (2003) hat dann gezeigt

Satz 2.1 *Die Priori sei durch*

$$f(\theta) \propto \sqrt{J(\theta)} \tag{85}$$

definiert. Dann ist f invariant gegenüber umkehrbar eindeutigen Transformationen φ von θ.

Prioris, die wie in (85) definiert sind, heißen auch *Jeffreys-Prioris.*

Beispiel 2.8 Die zufällige Veränderliche X sei Poisson-verteilt, d.h.

$$P(X = k|\theta) = e^{-\theta}\frac{\theta^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{86}$$

Gesucht ist eine nicht-informative Priori für θ. Wie schon angemerkt, ist die Annahme einer Gleichverteilung für θ auf $(0, \infty)$ nicht besonders sinnvoll. Gesucht ist dann eine Transformation für θ derart, dass Satz 2.1 gilt. Weiter sei $\mathbf{X} = (k_1, k_2, \ldots, k_n)$, k_i die beobachteten Häufigkeiten, $i = 1, \ldots, n$. Die Likelihood ist dann

$$L(\theta) = e^{-n\theta}\prod_{i=1}^{n}\frac{\theta^{k_i}}{k_i!}, \tag{87}$$

und wegen

$$\log L(\theta) = l(\theta) = -n\theta + \log\theta \sum_{i=1}^{n} k_i - \sum_{i=1}^{n} \log k_i!,$$

und wegen $\sum_i k_i = n$,

$$S(\theta) = \frac{dl(\theta)}{d\theta} = -n + \frac{n}{\theta}, \tag{88}$$

so dass die Information durch

$$I(\theta) = -\frac{dS(\theta)}{d\theta} = \frac{n}{\theta^2} \tag{89}$$

gegeben ist. Da $I(\theta)$ nicht von den x-Werten abhängt, ist die Information auch gleich der erwarteten Information $J(\theta)$, und man hat

$$f(\theta) \propto \sqrt{J(\theta)} = \sqrt{\theta}. \tag{90}$$

□

Beispiel 2.9 Gauß-Verteilung: Es sei wieder $\mathbf{X} = (x_1, \ldots, x_n)$ mit $x_i \sim N(\mu, \sigma^2)$, μ und bekannt, σ^2 bekannt. Man findet $J(\mu) = n/\sigma^2$ hängt nich tvon μ ab. Also ist Jeffreys Priori konstant über $\mathbb{R}$, d.h. $f(\mu) =$ Konstante. Für die Posteriori-Verteilung erhält man dann

$$\mu|\mathbf{X} \sim N\left(\bar{x}, \sigma^2/n\right). \tag{91}$$

Nun sei μ bekannt, aber σ^2 unbekannt. Dann ist

$$J(\sigma^2) = \frac{n}{2\sigma^2}, \tag{92}$$

und man erhält für die Posteriori

$$f(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{93}$$

Die Likelihood ist

$$L(\sigma^2) = \frac{1}{(\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right] \tag{94}$$

und

$$f(\sigma^2|\mathbf{X}) \propto (\sigma^2)^{-(1+n/2)} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right]. \tag{95}$$

Dies ist die *inverse Gamma-Verteilung* mit den Parametern $n/2$ und $\sum_i (x_i - \mu)^2/2$. Für den Erwartungswert, gegeben die Daten X, erhält man

$$\mathbb{E}(\sigma^2|\mathbf{X}) = \frac{\sum_i (x_i - \mu)^2}{n-2}. \tag{96}$$

Der Maximum-Likelihood-Schätzer dagegen ist

$$\hat{\sigma}^2_{ML} = \frac{1}{n} \sum_i (x_i - \mu)^2. \tag{97}$$

Für hinreichend großen Wert von n verschwindet der Unterschied zum Bayes-Schätzer. □

Man kann statt des Parameters σ^2 auch die Streuung σ betrachten. Es ist $\sigma = \sqrt{\sigma^2}$ und der Transformationssatz für Dichten liefert dann für die Posterior-Verteilung

$$f(\sigma) \propto \frac{1}{\sigma}, \tag{98}$$

und für die Präzision $\kappa = 1/\sigma^2$ findet man

$$f(\kappa) \propto \frac{1}{\kappa}. \tag{99}$$

Jeffreys' Priori ist für den jeweiligen Parameter stets proportional zum Reziprokwert! Die folgende Tabelle enthält die Jeffreys-Prioris für eine Reihe von Verteilungen:

Tabelle 3: Jeffreys' Prioris

Likelihood	Jeffreys' Priori $f(\theta) \propto$
$B(\theta)$	$17\sqrt{[\theta(1-\theta)]}$
$\text{Geom}(\theta)$	$1/(\theta\sqrt{1-\theta})$
$\text{Po}(\theta)$	$1/\sqrt{\theta}$
$\exp(\theta)$	$1/\theta$
$N(\mu, \sigma^2)$ (σ^2 bekannt)	1
$N(\mu, \sigma^2)$ (μ bekannt)	$1/\sigma^2$

Beispiel 2.10 Korrelation Es sei ρ der Produkt-Moment-Korrelationskoeffizient. Die Transformation

$$h(\rho) = \tanh^{-1}(\rho) = \frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right) \tag{100}$$

Dann ist

$$\frac{dh(\rho)}{d\rho} = \frac{1}{1-\rho^2}, \tag{101}$$

woraus sich die nicht-informative Priori-Verteilung

$$f(\rho) \propto \frac{1}{1-\rho^2} \tag{102}$$

ergibt. □

2.2.7 Jaynes' Maximale Entropie und Transformationsgruppen

Jaynes (1968) betrachtet a-priori-Verteilungen vom Standpunkt der Informationstheorie aus. Für eine Verteilung über n Zuständen mit den Wahrscheinlichkeiten p_i läßt sich die Entropie

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \tag{103}$$

definieren; diese soll dann unter eventuell gegebenen Nebenbedingungen – außer der, dass $\sum_i p_i = 1$ – maximiert werden.

Es werde angenommen, dass $\mathbf{x} = (x_1, \ldots, x_n)$ gegeben sei, wobei n endlich oder zumindest abzählbar unendlich sei. Die x_i dürfen beliebig sein. Die verfügbare Information I möge eine Reihe von Randbedingungen für die Wahrscheinlichkeitsverteilung $p(x_i|I)$ spezifizieren. Eine mögliche Spezifikation ist, dass diese Randbedingungen die Mittelwerte der Funktionen $\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$ festlegen, $m < n$. Dann soll

$$H = -\sum_{i=1}^{n} p(x_i|I) \log p(x_i|I)$$

unter den Bedingungen

$$\sum_i p(x_i|I) = 1, \quad \sum_i p(x_i|I) f_k(x_i) = F_k, \;\; k = 1, \ldots, m \tag{104}$$

maximiert werden. Die F_k sind festgelegte Mittelwerte. Eine allgemeine Lösung ist (Jaynes (1968, p. 7), (2003, p. 355))

$$p(x_i|I) = \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp(\lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)). \tag{105}$$

$Z(\lambda_1, \ldots, \lambda_m)$ ist eine *partition function*

$$Z(\lambda_1, \ldots, \lambda_m) = \sum_i \exp(\lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)), \quad \lambda_k \in \mathbb{R},$$

wobei die λ_k gemäß den Randbedingungen (104) bestimmt werden, die sich in der Form

$$F_k = \frac{\partial}{\partial \lambda_k} \log Z(\lambda_1, \ldots, \lambda_m) \tag{106}$$

darstellen lassen. Wie Jaynes es formuliert: die Verteilung (105) ist diejenige, die unter den gegebenen Randbedingungen so weit es irgend geht "ausgespreizt" ist. Damit erfüllt sie eine Bedingung für a-priori-Verteilungen: maximale Ungewißheit unter gegebenen Randbedingungen zu repräsentieren und damit maximale Freiheit für die Entscheidungen über die Parameter zu lassen.

x werde nun durch ein Zufallsexperiment bestimmt: Das Experiment werde M-mal wiederholt. Die Frage ist, was über die x_i ausgesagt werden kann, wenn $M \to \infty$; dieser Fall wird bei der frequentistischen Interpretation der Wahrscheinlichkeit angenommen. Es gibt n^M mögliche Resultate. Die M Wiederholungen liefern m_1 x_1-Werte, m_2 x_2-Werte, etc,und $\sum_i m_i = M$. Die festgelegten Mittelwerte mögen gefunden worden sein, so dass es die Bedingungen

$$\sum_{i=1}^{n} m_i f_k(x_i) = M F_k, \quad k = 1, 2, \ldots, m \tag{107}$$

gibt. Die Frage ist, wie viele der n^M Resultate nun mit den gefundenen Zahlen $\{m_1, m_2, \ldots, m_n\}$ kompatibel sind. Die Anzahl ist durch den Multinomialkoeffizienten

$$W = \frac{M!}{m_1! m_2! \cdots m_n!} = \frac{M!}{(M f_1)! \cdots (M F_m)!} \tag{108}$$

gegeben. Damit ist die Menge der Häufigkeiten $\{f_1, f_2, \ldots,_n\}$, die in der größtmöglichen Variation erzeugt werden kann, diejenige, die (108) maximiert relativ zu den Bedingungen (107). Gleichwertig dazu kann man jede monoton wachsende Funktion von W maximieren, etwa $M^{-1} \log W$, und über die Stirling-Approximation erhält man für $M \to \infty$

$$M^{-1} \log W \to -\sum_{i=1}^{n} f_i \log f_i = H_f. \tag{109}$$

Hat man also *testbare* Information, so ist die Wahrscheinlichkeitsverteilung, die die Entropie maximiert, identisch mit der Häufigkeitsverteilung, die in der größmöglichen Anzahl realisiert werden kann.

Beispiel 2.11 Herleitung der Binomialverteilung (Jaynes (1968, p. 11)) Es wird ein Experiment durchgeführt, das dem Experimentator eine "Mitteilung" liefert: das Alphabet bestehe aus dem möglichen Resultat eines Versuchsdurchganges, und in jedem Versuchsdurchgang wird ein "Buchstabe" der Mitteilung

geliefert. Insbesondere handele es sich um Bernoulli-Versuche, mit den zufälligen Veränderlichen

$$y_i = \begin{cases} 1, & \text{"Erfolg"} \\ 0, & \text{"kein Erfolg"} \end{cases}$$

Nach n Wiederholungen bzw. Versuchsdurchgängen hat man die Mitteilung

$$M \equiv \{y_1, y_2, \ldots, y_n\}.$$

Die Gesamtzahl der "Erfolge" ist dann

$$r(M) = \sum_i y_i.$$

Es werde nun $\mathbb{E}(r) = np$ angenommen. Gesucht ist nun Wahrscheinlichkeit, r Erfolge in n Versuchen zu erhalten. Dazu wird das Maximum-Entropie-Prinzip angewendet. Gesucht ist die Wahrscheinlichkeit

$$P_M \equiv p\{y_0 y_1 \cdots y_n\},$$

auf dem 2^n-Stichprobenraum aller möglichen Mitteilungen; gesucht ist also die Verteilung P_M, die die Entropie

$$H = -\sum_M P_M \log P_M$$

maximiert unter der Nebenbedingung $\mathbb{E}(r) = np$. (105) liefert

$$P_M = \frac{1}{Z(\lambda)} \exp(\lambda r(M)), \tag{110}$$

mit

$$Z(\lambda) = \sum_M \exp(\lambda r(M)) = (e^\lambda + 1)^n.$$

Die Lösung ergibt sich durch Anwendung von (106):

$$\mathbb{E}(r) = \frac{\partial}{\partial \lambda} \frac{n}{\exp(-\lambda) + 1},$$

woraus sich

$$\lambda = \log \frac{\mathbb{E}(r)}{n - \mathbb{E}(r)} = \log \frac{p}{1-p}$$

ergibt. Hieraus und aus (110) folgt dann

$$P_M = p^r (1-p)^{n-r}. \tag{111}$$

P_M ist die Wahrscheinlichkeit, eine bestimmte Mitteilung zu erhalten, mit "Erfolgen" in bestimmten Versuchen. Kommt es auf die Positionen der Erfolge nicht an, muß noch mit $\binom{n}{r}$ multpliziert werden, und man erhält die Binomialverteilung

$$p(r|n) = \binom{n}{r} p^r (1-p)^{n-r}. \tag{112}$$

□

Stetige Verteilungen Die Verallgemeinerung auf den Fall stetiger Verteilungen ist schwierig, weil die Verallgemeinerung der Definition der Entropie nicht einfach in der Form

$$H = -\int p(x) \log p(x) dx$$

angeschrieben werden kann; dieser Ausdruck ist nicht invariant unter Variablentransformationen $x \to y(x)$. Die gleiche Aussage gilt für das Bayessche Theorem mit der Konsequenz, dass man nicht sagen kann, welche Parametrisierung in Bezug auf das Indifferenzprinzip gewählt werden muß[9]. Eine ausführliche Darstellung findet man in Jaynes (2003, Kap. 11).

Jaynes schlägt vor, das Problem über geeignete Transformationsgruppen zu lösen, und illustriert den Gedanken zunächst an einem Beispiel.

Es wird eine Stichprobe $\mathbf{x}$ gebildet, wobei die Population durch eine 2-Parameterverteilung

$$p(dx|\mu, \sigma) = h\left(\frac{x-\mu}{\sigma}\right)\frac{dx}{\sigma} \tag{113}$$

definiert sei. Gegeben $\mathbf{x} = (x_1, \ldots, x_n)$ ist die Aufgabe, μ und σ zu schätzen. So lange keine a-priori-Verteilung $f(\mu, \sigma)d\mu d\sigma$ erklärt ist, ist das Schätzproblem nicht definiert. Wenn man nun nach dem Indifferenzprinzip vollständige Unwissenheit postuliert, ist nicht klar, welche Funktion f gewählt werden muß.

Andererseits ist μ ein Lokationsparameter und σ ein Skalenparameter, und die Funktion h in (113) ist bekannt. Nimmt man nun komplettes Unwissen über μ und σ an, so heißt das, dass ein Wechsel des Skalenparameters und des Lokationsparameters den Zustand vollständigen Unwissens nicht ändert. Man betrachte die Transformationen

$$\mu' = \mu + b, \quad \sigma' = a\sigma, \quad x' - u' = a(x-\mu), \tag{114}$$

mit $0 < a < \infty$, $(-\infty < b < \infty)$. Die Verteilung (113) geht über in

$$p(dx'|\mu', \sigma') = h\left(\frac{x'-\mu'}{\sigma'}\right)\frac{dx'}{\sigma'}, \tag{115}$$

[9]Wie Jaynes (1968, p. 16) anmerkt, ist das hier entstehende Problem nicht typisch für die Bayes-Statistik; es existiert auch für erwartungstreue und effiziente Schätzer, kleinste Konfidenzinervalle, etc, der "orthodoxen" Statistik.

d.h. d.h. h bleibt unverändert. Die a-priori-Verteilung geht aber über in

$$g(\mu', \sigma') = \frac{1}{a} f(\mu, \sigma). \tag{116}$$

Man habe nun eine zweite Stichprobe $\mathbf{x}' = (x'_1, \ldots, x'_n)$ und soll μ' und σ' schätzen. Man sei wieder vollständig unwissend bezüglich der Werte dieser Parameter. Die Fragestellung ist vollständig symmetrisch zu der gerade behandelten, so dass aus Konsistenzgründen die a-priori-Vereilungen identisch sein müssen, d.h. es muß

$$f(\mu, \sigma) = g(\mu, \sigma) \tag{117}$$

gelten, unabhängig von den Werten von a und b in (114). Allerdings ist nun die Form von f bzw. g festgelegt, denn wegen (116) muß nun die Funktionalgleichung

$$f(\mu, \sigma) = a f(\mu + b, a\sigma) \tag{118}$$

gelten. Diese Gleichung hat die Lösung

$$f(\mu, \sigma) = \frac{\text{Konstante}}{\sigma}. \tag{119}$$

Diese a-priori-Verteilung wurde bereits von Jeffreys betrachtet und hat deshalb den Namen *Jeffreys-Regel*.

Beispiel 2.12 Poisson-Parameter Die Wahrscheinlichkeit, dass genau n Ereignisse im Zeitintervall t beobachtet werden, sei durch die Poisson-Verteilung

$$p(n|\lambda, t) = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \tag{120}$$

gegeben. Aus der Anzahl der beobachteten Ereignisse soll nun der Parameter λ geschätzt werden. Man stelle sich zwei Beobachter vor, deren Uhren auf verschiedenen Zeitskalen laufen. Die Messungen eines bestimmten Intervalls sind dann durch die Beziehung $t = qt'$ aufeinander bezogen. Da beide das gleichen physikalische Experiment beobachten, sind die Raten λ und λ' durch $\lambda t = \lambda' t'$, also durch $\lambda' = q\lambda$ aufeinander bezogen. Die Beobachter wählen nun die a-priori-Verteilungen

$$p(d\lambda|\mathbf{X}) = f(\lambda) d\lambda, \quad p(d\lambda'|\mathbf{X}') = g(\lambda') d\lambda'. \tag{121}$$

Die Verteilungen müssen wechselseitig konsistent sein, so dass $f(\lambda)d\lambda = g(\lambda')d\lambda'$ gelten muß. Nun seien die beiden Beobachter vollständig unwissend bezüglich λ bzw. λ'. Also muß $f = g$ gelten, d.h. es muß gelten

$$f(\lambda) = q g(q\lambda), \text{ d.h. } p(d\lambda'|X') = \frac{d\lambda}{\lambda}. \tag{122}$$

Jede andere Wahl der a-priori-Verteilung bedeutet, dass eine Änderung der Zeitskala die Form der Verteilung änderte und damit ein anderes Maß an Wissen über λ ausdrücken würde. Aber die Annahme vollständiger Unwissenheit, entsprechend dem Indifferenzprinzip, legt die Form der a-priori-Verteilung durch (122) fest, wenn sie invariant gegenüber Skalentransformationen sein soll. □

Die Transformationen, die hier betrachtet wurden, waren linear. Linearität ist aber eine weder notwendige noch hinreichende Bedingung, wie das folgende Beispiel zeigt.

Beispiel 2.13 Unbekannter Bernoulli-Parameter Es werden n Bernoulli-Versuche mit unbekannter Erfolgswahrscheinlichkeit θ betrachtet; die Wahrscheinlichkeit von r Erfolgen sei also durch

$$P(r|n,\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}. \tag{123}$$

Gesucht ist nun die a-priori-Verteilung, die vollständiges Uniwssen über θ ausdrückt.

Die "natürliche" Annahme ist, jedem Wert für θ zwischen 0 und 1 die gleiche Wahrscheinlichkeit zuzuordnen, so dass man die a-priori-Verteilung $f(\theta) = 1$ erhält. Aber die *rule of succession* zeigt merkwürdige Implikationen dieser Annahme. Der Ansatz, sich der Frage nach der a-prriori-Verteilung über Transformationsgruppen zu nähern, kann hier aber nicht so einfach wie im Falle der Parameter μ und σ durchgeführt werden, denn eine lineare Transformation von θ führt leicht aus dem Intervall $[0, 1]$ heraus. Gesucht ist eine Transformation, die θ wieder auf das Intervall $[0, 1]$ abbildet.

Jaynes (2003, p. 384) findet die Transformation über einen interessanten Umweg. $f(\theta)$ beschreibe nicht das Wissen einer Person, sondern die Verteilung der θ-Werte in einer Population von Individuen, bei der jede Person durch einen zunächst festen θ-Wert gekennzeichnet ist. Der Begriff der vollständigen Unwissenheit über θ muß nun auf die Population angewendet werden: f soll den Zustand vollständiger Konfusion bezüglich θ in der Population beschreiben. Jedes Mitglied der Population verändere aber sein Wissen nach Maßgabe des Bayesschen Satzes. Vor Beginn des eigentlichen Experimentes werde nun jedes Mitglied der Population mit der gleichen Evidenz E bezüglich θ versehen. Herr A habe davor die Ansicht gehabt, die Wahrscheinlichkeit eines Erfolges sei durch $\theta = p(S|A)$ gegeben. Diese Wert wird durch E in

$$\theta' = P(S|EA) = \frac{P(E|SA)P(S|A)}{P(E|SA)P(S|A) + P(E|FA)P(F|A)}$$

verwandelt. Dabei ist $P(F|X) = 1 - P(S|A)$, S steht für Erfolg (Success), und F für Mißerfolg (Failure). Der neue Wert θ' und der alte Wert θ sind demnach

durch die Beziehung

$$\theta' = \frac{a\theta}{1-\theta+a\theta} \tag{124}$$

aufeinander bezogen, mit $a = P(E|SA)/P(E|FA)$.

Die Population als Ganzes habe durch die neue Evidenz E nichts gelernt, durch konfligierende Propaganda sei sie in einen Zustand totaler Konfusion bzw. vollständiger Unwissenheit versetzt worden. Dies soll bedeuten, dass nach der Transformation (124) der Anteil der Personen mit $\theta_1 < \theta < \theta_2$ der gleiche ist wie vor der Gabe von E. Ist die a-priori-Verteilung vor der Transformation f und nach der Transformation g, so soll demnach

$$f(\theta)d\theta = g(\theta')d\theta', \tag{125}$$

und wenn die Population durch E nichts gelernt hat, muß darüber hinaus gelten

$$f(\theta) = g(\theta). \tag{126}$$

Kombiniert man nun (124), (125) und (126), so ergibt sich die Funktionalgleichung

$$af\left(\frac{a\theta}{1-\theta-a\theta}\right) = (1-\theta-a\theta)^2 f(\theta). \tag{127}$$

Die Gleichung läßt sich lösen durch Elimination von a über (124) und (127)), oder durch Differentiation nach a und anschließender Setzung $a = 1$. Es ergibt sich die Differentialgleichung

$$\theta(1\theta)f'(\theta) = (2\theta-1)f(\theta) \tag{128}$$

mit der Lösung

$$f(\theta) = \frac{\text{Konstante}}{\theta(1-\theta)}. \tag{129}$$

Die vielen Leute aus der Population können nun wieder zu einer Person zusammengefasst werden. Hat man r Erfolge in n Versuchen beobachtet, so erhält man aus (123) und (129) die a-posteriori-Verteilung

$$P(d\theta|r,n) = \frac{(n-1)!}{(r-1)!(n-r-1)!}\theta^{r-1}(1-\theta)^{n-r-1}d\theta, \tag{130}$$

und diese Verteilung hat den Ewartungwert und die Varianz

$$\mathbb{E}(\theta) = \frac{r}{n}, \quad \mathbb{V}(\theta) = \frac{r(n-r)}{n^2(n+1)}. \tag{131}$$

Also ist die beste Schätzung für θ durch r/n gegeben, und r/n ist auch die Wahrscheinlichkeit des Erfolgs im folgenden Versuch(sdurchgang), wie es sich für eine Folge von Bernoulli-Versuchen gehört, im Unterschied zu der Vorhersage durch die Laplacesche Folgeregel, die $(r+1)/(n+2)$ behaupten würde und die sich aus der Annahme der Gleichverteilung für θ auf $[0,1]$ als a-priori-Verteilung ergibt. □

2.2.8 Zur Geschichte der Theorie der Ereignisfolgen

Keynes (1921/2008) verweist auf einige Arbeiten, die die Mühsal reflektieren, die das Verständnis von Folgen von Ereignissen machte, bevor sich Kolmogoroffs Axiomatik durchsetzte, die ebenfalls das Verständnis von 'Bernoullis Theorem' erleichterte. Dies besagt, nach Keynes (p. 338)

> "... in its simplest form ... If the probability of an event's occurrence under certain conditions is p, then, if these conditions are present on m occasions, the most probable number of the event's occurrences is mp (or the nearest integer to this, i.e. the most probable *proportion* of its occurrences to the total number of occasions is p; further, the propability that the proportion p by less than a given amount b, increases as m increases, the value of this probability being calculable by a process of approximation."

Das ist also die Aussage, dass bei einer binomialverteilten Variablen X bei m Beobachtungen der Erwartungswert durch $\mathbb{E}(X) = mp$ und die Varianz durch $\mathbb{V}(X) = mp(1-p)$ gegeben ist; der Anteil der 'Erfolge' ist dann $\mathbb{E}(X)/m = p$; ist k die Anzahl der Erfolge und $\hat{p} = k/m$, so gilt bekanntlich

$$\lim_{m\to\infty} P(|\hat{p} - p|) = 0,$$

d.h. kleinere Abweichungen haben bei wachsemdem m eine größere Wahrscheinlichkeit als größere Abweichungen; dies sieht man leicht ein, wegen (i) $\mathbb{E}(\hat{p}) = p$, also $\hat{p} \to p$, und $\mathbb{V}(\hat{p}) = p(1-p)/m$, also $\lim_{m\to\infty} \mathbb{V}(\hat{p}) \to 0$. Keynes berichtet, dass Simon Laplace der Ansicht war, die hier genannten Sachverhalte seien Ausdruck eines allgemeinen Naturgesetzes seien. Sein berühmtes Werk *Essai philosophique sur la probabilité* aus dem Jahre 1812 war ursprünglich Napoleon (*A Napoléon-le-Grand*) gewidmet. In der Neuauflage aus dem Jahr 1814 hat er diese Widmung ersetzt durch eine Deutung des Bernouliischen Theorems. Es bringe zum Ausdruck, dass jede große Kraft, die, trunken an ihrer Liebe zur Eroberung und universeller Herrschaft, am Ende zum Niedergang gezwungen werde[10]. Allgemein nahm man an, dass Bernoullis Theorem auf alle "Korrekt" berechneten Wahrscheinlichkeiten anzuwenden sei. So hat man längliche Auszählungen von Folgen von Ereignissen beim Roulette vorgenommen (Beispiele bei Keynes, Seite 363). Besonderes Interesse erregten die Untersuchungen von Dr. Karl Marbe (der später Professor und Begründer des Würzburger Instituts für Psychologie wurde). Der untersuchte die Folgen von 80 000 Würfen beim Roulette in Monaco und ähnlichen Anstalten und untersuchte insbesondere das Auftreten bestimmter

[10] "C'est encore un résultat du calcul des probabilités, confirmé par des nombreuses ert funestes expériences."

Folgen von Ereignissen. Seine Ergebnisse bestätigten sich in seiner Ansicht, dass die Welt so strukturiert sei, dass lange "runs" nicht nur unwahrscheinlich seien, sondern überhaupt nicht vorkämen. Er versuchte also, eine metaphysische Aussage über das Universum über das Roulette zu bestätigen (Marbe (1899, 1916)). Keynes führt aus, dass Marbes (1899) Buch vor allem in Deutschland diskutiert worden sei, – aber nicht wegen der absurden (preposterous) Art und Weise, ein allgemeines Gesetz konstituieren zu wollen, also aus Freude am Bizarren, sondern weil man es ernst nahm. Man mag spekulieren, welch philosophischer Zeitgeist in Deutschland dazu beigetragen hat, dass die Marbeschen Ideen hier mehr verfingen als in anderen Ländern; schließlich geht es nur um die richtige Verwendung des Begriffs unabhängiger Ereignisse.

2.3 Bayes-Asymptotik

Eine zentrale Frage für jede Parameterschätzung ist die nach der *Konsistenz*, d.h. ob die Schätzungen für $N \to \infty$ gegen den wahren Wert des Parameters streben.

Es wird zunächst ein Distanz- bzw. Diskrepanzbegriff für zwei Wahrscheinlichkeitsverteilungen $P(x)$ und $Q(x)$ eingeführt:

Definition 2.8 *Mit $D(P,Q)$ werde ein Abstandsmaß für dieVerteilungen P und Q bezeichnet; insbesondere heißt*

$$D(P,Q) = \begin{cases} \sum P(x)\log_2 P(x))/Q(x), & \textit{diskrete Verteilungen} \\ \int p(x)\log_2 p(x)/q(x)\,dx, & \textit{stetige Verteilungen} \end{cases} \tag{132}$$

Kullback-Leibler-Distanz *der Verteilungen P und Q (diskret) bzw. p und q (stetig).*

Man kann nun $p(x) = f(x|\theta_w)$, $q(x) = f(x|\theta)$ annehmen, wobei θ_w der wahre Parameterwert ist, und θ ein beliebiger anderer Wert. Dann hat man für diese beiden Verteilungen die Kullback-Leibler-Distanz

$$D(\theta_w,\theta) = \int p(x|\theta_w)\log_2 \frac{p(x|\theta_w)}{q(x|\theta)}\,d\theta. \tag{133}$$

Jetzt kann der folgende Satz bewiesen werden

Satz 2.2 *$D(\theta_w,\theta)$ sei wie in (133) definiert. Dann folgt*

$$\lim_{n\to\infty} f(\theta_w|x) = 1, \quad \lim_{n\to\infty} f(\theta|x) = 0, \quad \theta_w \neq \theta. \tag{134}$$

Dies bedeutet, dass – offenbar unabhängig von der zunächst gewählten Prior – die Posteriori-Wahrscheinlichkeit für den wahren Wert des Parameters θ gegen 1 geht, und korrespondierend dazu geht die Posteriori-Wahrscheinlichkeit für einen falschen Wert geht gegen Null.

Es ist noch von Interesse, gegen welche Verteilung die Posteriori-Verteilung konvergiert, wenn der Stichprobenumfang n groß wird, – es könnte ja sein, dass die Grenzverteilung von der Priori-Verteilung abhängt.

Satz 2.3 *Es sei n der Stichprobenumfang. Dann gilt für hinreichend großes n*

$$\theta|\boldsymbol{x} \overset{a}{\sim} N(\hat{\theta}_n, I(\hat{\theta}_n)^{-1}, \tag{135}$$

wobei $\hat{\theta}_n$ der ML-Schätzer für θ ist ($\hat{\theta}_n$ kann ein Vektor sein) und die Kovarianzmatrix durch $I(\hat{\theta}_n)$ gegeben ist.

3 Anhang

3.1 Signifikanztests für Binomial- und inverse Binomialexperimente

$$P(n > 6|H_0) = \theta \sum_{k=7}^{\infty} (1-\theta)^{k-1}$$

Nun ist

$$\sum_{k=1}^{\infty} (1-\theta)^{k-1} = \sum_{k=1}^{6} (1-\theta)^{k-1} + \sum_{k=7}^{\infty} (1-\theta)^{k-1}.$$

Bekanntlich gilt, für $1-\theta = a < 1$,

$$\sum_{k=1}^{\infty} a^{k-1} = \frac{1}{1-a}, \quad \sum_{k=1}^{r} a^{k-1} = \frac{1-a^r}{1-a},$$

so dass

$$\sum_{r+1}^{\infty} a^{k-1} = \frac{1}{1-a} - \frac{1-a^r}{1-a} = \frac{1}{1-a}(1-(1-a^r) = \frac{a^r}{1-a} = \frac{(1-\theta)^r}{\theta},$$

und schließlich, wenn man wieder $1-\theta$ mit $\theta = 1/2$ für a einsetzt,

$$P(n > 6|H_0) = \theta \sum_{k=7}^{\infty} (1-\theta)^{k-1} = \left(\frac{1}{2}\right)^6,$$

und

$$p = P(n = 6|H_0) + P(n > 6|H_0) = \left(\frac{1}{2}\right)^6 + \left(\frac{1}{2}\right)^6 = 2\left(\frac{1}{2}\right)^6 = .031.$$

Literatur

[1] Akaike, H. (1982) On the fallacy of the likelihood principle. *Statistics & Proability Letters*, 1, 75 – 78

[2] Birnbaum, A. (1962) On the foundations of statistical inference. *Journal of the American Statistical Association*, 67, 269 – 306

[3] Birnbaum, A. (1972) More concepts of statistical evidence. *Journal of the American Statistica Association*, 67 (340), 858– 861

[4] Box, G. E.P., Tiao, G.C.: Bayesian inference in statistical analysis. Addison-Wesley Publication Company, Reading (Mass.) 1973

[5] Deakin, M.A.B. The Wine/Water Paradox: background, provenance and proposed resolutions.
http://www.austms.org.au/Gazette/2006/Jul06/mdeakin.pdf

[6] Evans, M.J., Fraser, D.A.S., Monette, G. (1986) On principles and arguments to likelihood. *The Canadian Journal of Statistics*, 14 (3), 181 – 199

[7] Good, I.J. (1971), 46656 varieties of Bayesians. In: Hamdan, M.A., Pratt, J.W., Gottlieb, P. Good, I.J., Hamdan, A., Carmer,S. G., Walker,W. M., Valentine,T. J., D'Agostino, R.B., Kshirsagar, A. M., Gwyn Evans, I., Kabe, D. G., Cureton, E. E., Rutherford, J. R., Sharma, J. K., Harvey, J. R.: Letters to the editor. *The American Statistician*, 25 (5), 56 – 63

[8] Hacking, I.: Logic of statistical inference. Cambridge, Cambridge University Press 1965/2009

[9] Helland, I.S. (1995) Simple counterexamples against the conditionality principle. *American Statistician*, 49 (4), 351 – 356

[10] Jaynes, E. T.: Probability Theory. The Logic of Science. Cambridge University Press, Cambridge 2003

[11] Jeffreys, H.: Theory of probability. Oxford 1961/2003

[12] Kalbfleisch, J.D., Berger, J., Sprott, D.A. (1986) [On principles and arguments to likelihood] Discussion. *The Canadian Journal of Statistics*, 14 (3), 194 – 199

[13] Lavine, M. (1996) Conditionality alive and well. Unpublished Manuscript (http://www.math.umass.edu/ lavine/whatisbayes.pdf)

[14] Lindley, D. V. (1993) The Analysis of Experimental Data: the appreciation of tea and wine. *Teaching Statistics*, 15 (1), 22 – 25

[15] von Mises, R.: Probability, statistics and truth. New York 1957/1981

Index

Bayes-Faktor, 20
Bernoullis Theorem, 45

Evidenz
    mathematische, 14
    statistische, 14

Fisher-Information, 35

Haldane-Priori, 34
HDR = highest density region, 20
HPD-Intervall, 19
Hyperparameter, 30
Hypothesentests, 20

Indifferenzprinzip, 22

Jeffreys-Priori, 35
Jeffreys-Regel, 42

konjugiert
    bezüglich d. Likelihood-Funktion, 29
Konsistenz, 46
Kredibilitätsintervall, 19
Kullback-Leibler-Distanz, 46

Likelihood
    -Funktion, 17
    standardisierte, 18

Marbe, Dr.Karl, 45
Maximum-Entropie-Prinzip, 38

partition function, 38
prädiktive Verteilung, 18
Präzision, 32, 37
Prinzip des Unzureichenden Grundes, 22
Priori
    uneigentliche, 34
Propensität, 17

signifikant, 3
Signifikanzniveau, 3

Wein-Wasser Paradox, 23