2.4.6 Konditionale Bayes-Inferenz: Begrifflicher Hintergrund und „Erinnerungen“

In Kapitel 1.6.1 wurde davon gesprochen, dass die Lösung/ Darstellung des datengestützten Entscheidungsproblems über das Auswertungsproblem und die damit verbunden Suche nach optimalen Entscheidungsfunktionen durchaus auch auf Kritik stösst. Diese stützt sich v.a. einerseits

- auf die immense computationale Komplexität

und andererseits ganz prinzipiell

- auf die Problematik kontrafaktischer Ereignisse bei der Bewertung von Entscheidungsfunktionen mittels der Risikofunktion.

Dies legt eine konditionale Sicht als mögliche Alternative nahe: Es werden optimale Lösungen für die konkret beobachtete Datenkonstellation $\{x\}$ gesucht („auf $\{x\}$ konditionierte Betrachtung“). Dies führt auf die „übliche Bayes-Inferenz“, die mit Hilfe der sogenannten Posteriori-Verteilung gegeben die Daten arbeitet, und in diesem Sinne hier zur Abgrenzung als „*konditionale Bayes-Inferenz*“ bezeichnet werde.
Zur Vorbereitung werde an das Theorem von Bayes in seiner allgemeinen Form erinnert:

Proposition 2.86 (Allgemeines Theorem von Bayes)

Seien X und U zwei Zufallsvariablen mit gemeinsamer Wahrscheinlichkeitsfunktion $f_{X,U}(\cdot)$ bzw. Dichte $f_{X,U}(\cdot)$ (bezüglich eines dominierenden σ-finiten Maßes $\nu \otimes \lambda$) und den bedingten Wahrscheinlichkeitsfunktionen bzw. bedingten Dichten $f_{X|U}(\cdot|u)$ und $f_{U|X}(\cdot|x)$ (bezüglich ν bzw. λ).
Dann gilt:

$$f_{U|X}(u|x) = \frac{f_{X|U}(x|u) \cdot f_U(u)}{f_X(x)} \tag{2.41}$$

mit

$$f_X(x) = \int f_{X|U}(x|u) \cdot f_U(u) \, d\nu(u). \tag{2.42}$$

Bem. 2.87

Bei stetigem U mit Dichte $f_U(u)$ erhält man also Proposition 2.86 mit

$$f_X(x) = \int f_{X|U}(x|u) \cdot f_U(u)du. \tag{2.43}$$

Im Fall von diskreten Zufallsvariablen X und U – mit $\mathcal{U}$ als Träger von U – ergibt sich

$$p(\{U = u\}|\{X = x\}) = \frac{p(\{X = x\}|\{U = u\}) \cdot p(\{U = u\})}{p(\{X = x\})} \tag{2.44}$$

mit

$$p(\{X = x\}) = \sum_{u \in \mathcal{U}} p(\{X = x\}|\{U = u\}) \cdot p(\{U = u\}) \tag{2.45}$$

Betrachtet werde im Folgenden immer einer dieser beiden Spezialfälle . Die allgemeinere Formulierung über beliebige Dichten bezüglich geeigneter dominierender Maße ist unproblematisch.

Bem. 2.88 (Normierungskonstante)

$f_X(x)$ aus (2.42) spielt die Rolle einer reinen Normierungskonstante, die nicht von u abhängt. Häufig reicht es daher, $f_{X|U}(x|u) \cdot f_U(u)$ zu berechnen. Da man weiß, dass sich insgesamt eine Wahrscheinlichkeitsdichte ergeben muss, kennt man implizit auch die Normierungskonstante.
Man schreibt dann mit $\propto$ als Symbol für „proportional zu“

$$f_{X|U}(x|u) \propto f_{X|U}(u|x) \cdot f_U(u).$$

Bem. 2.89 (Konditionale Bayes-Inferenz: Konzeptionelle Hintergr

Gegeben sei ein datengestütztes Entscheidungsproblem $((\mathbb{A}, \Theta, l(\cdot)); (\mathcal{X}, \sigma(\mathcal{X}), (p_\vartheta(\cdot))_{\vartheta\in\Theta})$, wobei $p_\vartheta(\cdot)$ die Wahrscheinlichkeitsfunktion bzw. Dichte $f_\vartheta(\cdot)$ besitze, und eine Priori-Verteilung auf einem geeigneten Maßraum $(\Theta, \sigma(\Theta))$ mit Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\cdot)$.
Sei, als gedankliche Hilfskonstruktion, U („Umwelt “, „Natur “) eine „Zufallsgrösse“ (Zufallsvariable/-element), die das Eintreten des Umweltzustands ϑ beschreibt und X eine Zufallsgrösse, die den Ausgang des Informationbeschaffungsexperiments beschreibt. Dann gelte für die Dichte bzw. Wahrscheinlichkeitsfunktion $f_{X,U}(x, \vartheta)$ der gemeinsamen Verteilung von X und U für alle $x \in \mathcal{X}$ und $\vartheta \in \Theta$:

$$f_{X,U}(x, \vartheta) = \pi(\vartheta) \cdot f_\vartheta(x)$$

Das heißt für alle $x \in \mathcal{X}$ und $\vartheta \in \Theta$ gilt

$$f_\vartheta(x) = f_{X|U}(x|\vartheta) \ ;$$

die Verteilung der Zufallsgrösse aus der Informationsstruktur wird als bedingte Verteilung von X gegeben U interpretiert (!!).
Dann ergibt sich (!!!) für jedes x aus dem Satz von Bayes gemäß Proposition 2.86 für (eine Version der bzw.) die Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\vartheta|x)$ der bedingten Verteilung von U gegeben $X = x$:

$$\pi(\vartheta|x) = c(x) \cdot f_\vartheta(x) \cdot \pi(\vartheta) \tag{2.46}$$

mit

$$\frac{1}{c(x)} = f_X(x) = \int f_{X|U}(x|\vartheta)\pi(\vartheta)d\vartheta \tag{2.47}$$

im Falle von stetigem X und U, und bei diskretem X und U

$$\frac{1}{c(x)} = p(\{X = x\}) = \sum_{j=1}^{m} p(\{X = x\}|\{U = \vartheta_j\}) \cdot \pi(\{U = \vartheta_j\}) \quad (2.48)$$

$$= \sum_{j=1}^{m} p(\{X = x\}|\{U = \vartheta_j\}) \cdot \pi(\vartheta_j).$$

Für jede Beobachtung $\in \mathcal{X}$ wird die bedingte Verteilung von U gegeben $X = x$ als *Posteriori-Verteilung des Parameters* ϑ *nach der Beobachtung* x bezeichnet. Die zugehörige Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\vartheta|x)$ heißt *Posteriori-Dichte nach der Beobachtung*, und $f_\vartheta(x)$ heißt *Likelihood*.

Die marginale Verteilung von X mit Dichte $f_X(x)$ aus (2.47) bzw. Wahrscheinlichkeitsfunktion $(p(\{X = x\}))_{x \in \mathcal{X}}$ aus (2.48) heißt *Priori-Prädiktive-Verteilung.*

Die Größen$f_X(x)$ und $p(\{X = x\})$ sind nicht zu verwechseln mit den als bedingte Verteilungen interpretierten $f_\vartheta(x)$ und $p_\vartheta(\{X = x\})$.

Analog gibt es auch eine *posteriori-prädiktive Verteilung*, wenn man in analoger Weise über die Posteriori-Verteilung herausintegriert bzw. -summiert. Dies ist dann die Wahrscheinlichkeitsverteilung der nächsten Beobachtung, basierend auf dem aktuellen Wissensstand.

Bem. 2.90 (Bayes Postulat (nicht entscheidungstheoretisch))

Nach der Beobachtung der Stichprobe enthält die (klassische) Posteriori-Verteilung die volle Information, d.h. sie beschreibt das Wissen über den unbekannten Parameter vollständig.
Alle statistischen Analysen haben sich ausschließlich auf die Posteriori zu stützen; darauf baut insbesondere auch die Konstruktion von

- Bayesschen-Punktschätzungen: *MPD-Schätzer (Maximum Posteriori Density-Schätzer)*

- Bayesschen-Intervallschätzungen: *HPD-Intervalle (Highest posterior density-Intervalle)*

- Bayes-Tests

- Suffiziente Statistik: Enthält volle Information der Sichprobe über den Parameter (allgemeine statistische Theorie). Jetzt zwei verschiedene Arten von „enthält volle Information“; Wie passt Suffizienz in diesen Zusammenhang?

- Posteriori-Verteilung enthält volle Information (vgl. Bem 2.90).

$\Longrightarrow$ Posteriori hängt tatsächlich nur von suffizienter Statistik ab.

Proposition 2.91 (Suffizienz und Posteriori-Verteilung)

Ist in der Situation von Bemerkung 2.89 T eine für ϑ suffiziente Statistik mit Wahrscheinlichkeitsfunktion bzw. Dichte $g_\vartheta(\cdot)$, so hängt die Posteriori $\pi(\vartheta|x)$ nur mehr über $t = T(x)$ von x ab. Es gilt [6]

$$\pi(\vartheta|x) \propto g_\vartheta(t) \cdot \pi(\vartheta)$$

Beweis:
Gemäß (2.46) ist

$$\pi(\vartheta|x) \propto f_\vartheta(x) \cdot \pi(\vartheta)$$

wobei wegen der Suffizienz von T sich $f_\vartheta(x)$ schreiben lässt als $f_\vartheta(x) = h_{X|T}(x) \cdot g_\vartheta(t)$. Einsetzen liefert die Behauptung.

[6] $\propto$ „proportional zu“, vgl. Bemerkung 2.88

Def. 2.92 (Vorbereitende Erinnerung: Exponentialfamilien)

Sei $(\mathcal{X}, \sigma(\mathcal{X}), (p_\vartheta)_{\vartheta\in\Theta})$ ein statistisches Modell mit $\Theta \subseteq \mathbb{R}^q$.

- $(p_\vartheta)_{\vartheta\in\Theta}$ bildet eine (oder p_ϑ ist für jedes $\vartheta \in \Theta$ ein Mitglied der) *q-parametrige(n) Exponentialfamilie* in $(T_1, ..., T_q)$ mit *natürlichem Parameter* $(c_1(\vartheta), ..., c_q(\vartheta))$, wenn sich die Dichte $f_\vartheta(x)$ bezüglich eines dominierenden σ-finiten Maßes (also insbesondere Dichte/Wahrscheinlichkeitsfunktion) in die folgende Form bringen läßt: Mit $t_1 := T_1(\vec{x}), ..., t_q := T_q(\vec{x})$ ist

$$f_\vartheta(x) = h(x) \cdot g(\vartheta) \cdot \exp\Big(\sum_{\ell=1}^{q} c_\ell(\vartheta) t_\ell\Big).$$

- Enthält Θ echt innere Punkte und sind $1, c_1(\vartheta), c_2(\vartheta), ..., c_q(\vartheta)$ und $1, T_1(x), T_2(x), ..., T_q(x)$ (f.-s.) jeweils linear unabhängig, so spricht man von einer *strikt* q-parametrigen Exponentialfamilie. (Der „natürliche Parameterraum“ hat wirklich die Dimension q.)

2.4.7 Konjugierte Verteilungen, Bayes-Lernen

a) Ein Motivationsbeispiel

Bsp. 2.93 (Beta-Binomialmodell)

Absolutes Standardbeispiel

- Stichprobenmodell: Bernoulliverteilung (allgemein: Binomialverteilung) zu Parameter ϑ

$$p_\vartheta\left(\{X_i = x_i\}\right) = \vartheta^{x_i}(1-\vartheta)^{1-x_i}$$

jetzt im Bayes Kontext als bedingte Verteilung schreiben (wieder mit „Hilfsvariable“ U:

$$p\left(\{X_i = x_i\} \mid \{U = \vartheta\}\right) = \vartheta^{x_i}(1-\vartheta)^{1-x_i}$$

- gebräuchliche Priori-Verteilung:
  Betaverteilung, gilt als sehr flexibel, zwei Parameter $a > 0$, $b > 0$
  hier als Priori verwendet, Bezeichnung $\pi(\cdot)$

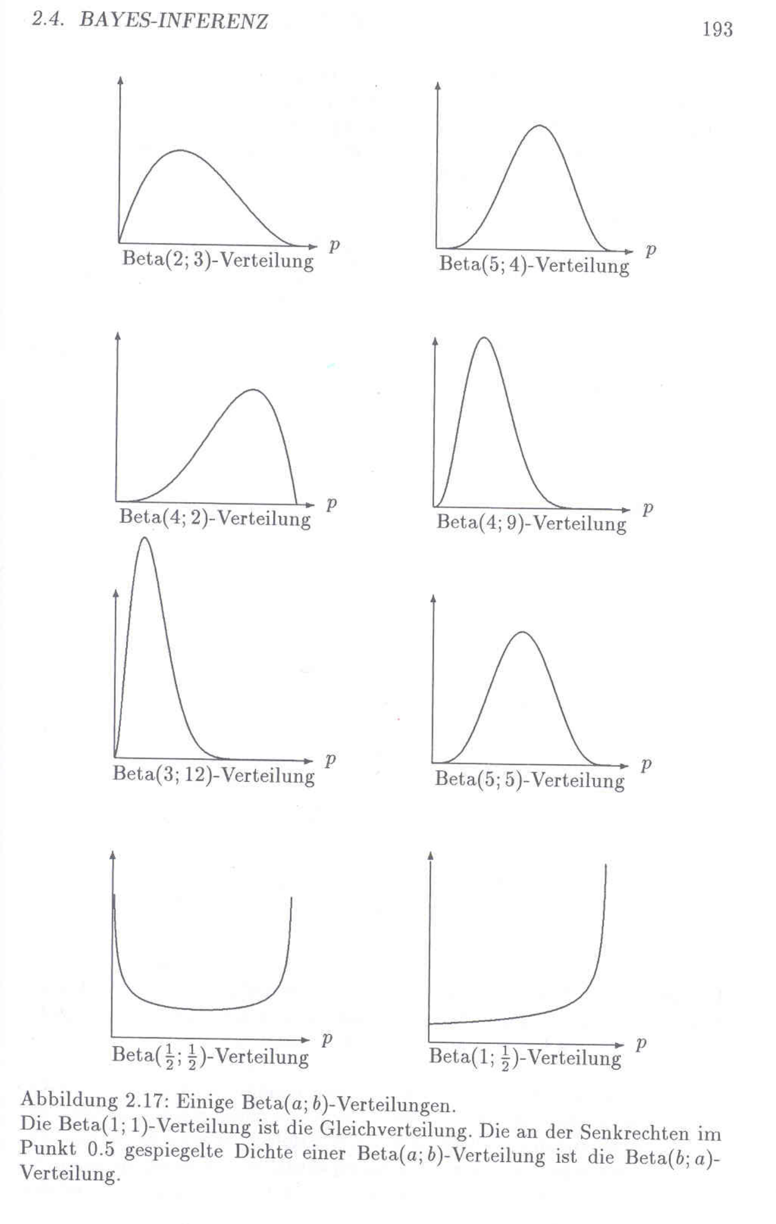
$$\pi(\vartheta) = \frac{\vartheta^{a-1}(1-\vartheta)^{b-1}}{B(a,b)} \cdot I_{[0;1]}(\vartheta) \tag{2.49}$$

  $B(a,b)$ ist eine reine Normierungskonstante.

Es gilt:

Erwartungswert: $\dfrac{a}{a+b}$ Modus: $\dfrac{a-1}{a+b-2}$, $a > 1, b > 1$

Abbildung 1: Rüger, (1999) Test- und Schätztheorie I, Seite 193



2.4. BAYES-INFERENZ 193

Beta(2; 3)-Verteilung p

Beta(5; 4)-Verteilung p

Beta(4; 2)-Verteilung p

Beta(4; 9)-Verteilung p

Beta(3; 12)-Verteilung p

Beta(5; 5)-Verteilung p

Beta($\frac{1}{2}$; $\frac{1}{2}$)-Verteilung p

Beta(1; $\frac{1}{2}$)-Verteilung p

Abbildung 2.17: Einige Beta(a; b)-Verteilungen.
Die Beta(1; 1)-Verteilung ist die Gleichverteilung. Die an der Senkrechten im Punkt 0.5 gespiegelte Dichte einer Beta(a; b)-Verteilung ist die Beta(b; a)-Verteilung.

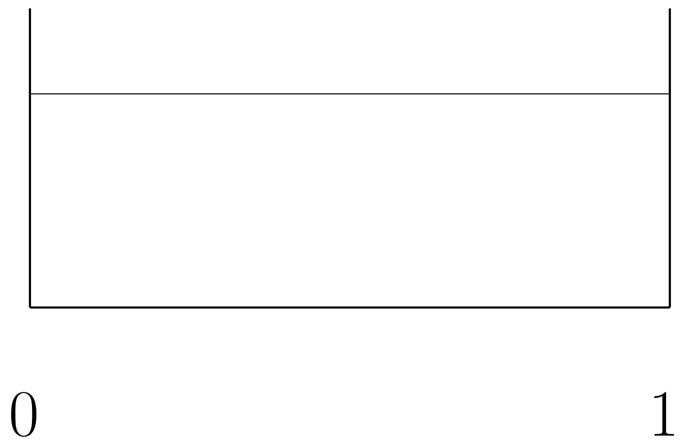
Jetzt Satz von Bayes anwenden: Posteriori nach einer Beobachtung berechnen.

$$\begin{aligned}
\pi(\vartheta|x_i) &= \frac{\vartheta^{x_i}(1-\vartheta)^{1-x_i} \cdot \vartheta^{a-1}(1-\vartheta)^{b-1}}{\underbrace{\text{Norm.} \cdot B(a,b)}_{\text{Normierung}}} \cdot I_{[0;1]}(\vartheta) \\
&\propto \vartheta^{x_i+a-1} \cdot (1-\vartheta)^{b-x_i} \cdot I_{[0;1]}(\vartheta) \\
&= \vartheta^{a'-1} \cdot (1-\vartheta)^{b'-1} \cdot I_{[0;1]}(\vartheta)
\end{aligned}$$

- Posteriori ist also wieder eine Betaverteilung, nun mit den Parametern

$$a' = a + x_i \quad \text{und} \quad b' = b - x_i + 1 = b + (1 - x_i).$$

Start z.B. mit $a^{(0)} = 1$, $b^{(0)} = 1$:



Gleichverteilung (als Nichtwissen verkaufbar?)

$\boxed{x_1 = 1}$ beobachtet $\Rightarrow$ $a^{(1)} = a^{(0)} + 1 = 2,\ b^{(1)} = b^{(0)} + 0 = 1$

$Beta(2,1)$-Verteilung

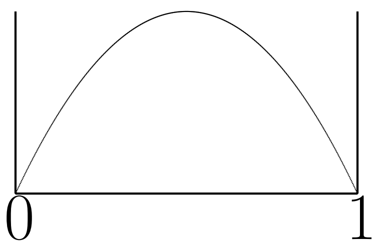
$\pi(\vartheta \mid x_1) \propto \vartheta I_{[0;1]}(\vartheta)$

- Jetzt weiteres Experiment:

  (neue) Priori = (alte Posteriori): Beta(2,1) Verteilung
  neue Stichprobe x_2
  neue Posteriori: $Beta(a^{(2)}, b^{(2)}) = Beta(a^{(1)} + x_i, b^{(1)} + (1 - x_i))$

  z.B. $\boxed{x_2 = 0} \rightarrow a^{(2)} = 2 + 0 = 2,\ b^{(2)} = 1 + 1 - 0 = 2$

  $$\pi_2(\vartheta|(x_1, x_2)') = \frac{\vartheta^{2-1} \cdot (1-\vartheta)^1}{\text{Norm.}} = \frac{\vartheta^1 \cdot (1-\vartheta)^1}{\text{Norm.}} = \frac{\vartheta - \vartheta^2}{\text{Norm.}} \quad \text{für} \quad \vartheta \in [0; 1]$$



$$\pi_2(0|x) = \pi_2(1|x) = 0$$

- Weitere Beobachtung $\boxed{x_3 = 1}$:

  neue Posteriori: $Beta(a^{(3)}, b^{(3)}) = Beta(a^{(2)} + x_i, b^{(2)} + (1 - x_i))$
  $a^{(3)} = 2 + 1,\ b^{(3)} = 2$

$$\pi_3(\vartheta|(x_1, x_2, x_3)^{'}) \propto \vartheta^2(1-\vartheta)^1 I_{[0;1]}(\vartheta) = \vartheta^2 - \vartheta^3 I_{[0;1]}(\vartheta)$$

- Weiter zusätzliche Beobachtung $\boxed{x_4 = 1}$:
  neue Posteriori: $Beta(a^{(4)}, b^{(4)}) = Beta(a^{(3)} + x_i, b^{(3)} + (1 - x_i))$
  $a^{(4)} = 3 + 1,\ b^{(4)} = 2$

$$\pi_4(\vartheta | (x_1, x_2, x_3, x_4)') \propto \vartheta^3 (1 - \vartheta)^1$$

- Allgemein gilt bei n unabhängigen Wiederholungen:

  Die Posteriori $\pi_n(\vartheta|(x_1, \ldots, x_n)')$ ist eine

  $$B\left(a^{(0)} + \sum_{i=1}^{n} x_i; b^{(0)} + n - \sum_{i=1}^{n} x_i\right) \text{ Verteilung.} \tag{2.50}$$

  Man kann zeigen: Dasselbe Ergebnis erhält man, wenn man $x_1, ..., x_n$ auf einmal verarbeitet.

- In diesem Beispiel gilt für die posteriori-prädiktive Verteilung

$$
\begin{aligned}
& p(X_{n+1} = 1 | X_1 = x_1, \ldots, X_n = x_n) \\
& = \mathbb{E}(\pi_n(\vartheta | x_1, \ldots, x_n)) \\
& = \frac{a + \displaystyle\sum_{i=1}^{n} x_i}{a + b + n}.
\end{aligned}
$$

Für die Gleichverteilung (vgl. oben) als Ausgangspriori ergibt sich wegen $a^{(0)} = b^{(0)} = 1$

$$p(X_{n+1} = 1 | X_n = x_n, \ldots, X_1 = x_1)$$

$$= \frac{\left(\sum_{i=1}^{n} x_i\right) + 1}{n + 2};$$

b) Konjugiertheit: Definition und klassische Ergebnisse (vgl. „Schätzen und Testen I“)

Def. 2.94 (Konjugiertheit)

Eine Verteilungsfamilie Π von Priori-Verteilungen heißt zu einer Menge $\mathcal{P}$ von Stichprobenverteilungen *konjugiert*, wenn für jede Priori-Verteilung $\pi(\cdot) \in \Pi$ und jedes $p(\cdot) \in \mathcal{P}$ die zugehörige Posteriori-Verteilung wieder ein Element von Π ist. Man sagt dann auch, dass jedes Element $\pi(\cdot) \in \Pi$ zu $\mathcal{P}$ konjugiert ist.

Proposition 2.95 (Beispiele für Konjugiertheit: Beta-Binomial/Dirichlet-Multinomial-Modell/Gamma-Poisson-Modell, Selbstkonjugiertheit der Normalverteilung)

a) Die Menge der Betaverteilungen als Priori ist zur Menge der Bernoulliverteilungen konjugiert (vgl. Bsp. 2.93).
Allgemeiner gilt:
Ist $\vec{X} = (X_1, \ldots, X_k)$ eine Stichprobe eines zum Parameter $\vec{\vartheta} = (\vartheta_1, \ldots, \vartheta_k)$ multinomial-verteilten Untersuchungsmerkmals, besitzt $\vec{X}$ also die Wahrscheinlichkeitsfunktion

$$f(x|\vec{\vartheta}) \propto \prod_{j=1}^{k} \vartheta_j^{x_j}$$

und wählt man die sog. *Dirichlet-Verteilung* zum Parameter $\vec{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$

$$\pi(\vec{\vartheta}) = \prod_{j=1}^{k} \vartheta_j^{(\alpha_j - 1)},$$

so ist die Posteriori-Verteilung eine Dirichlet-Verteilung mit dem Parameter $\alpha' = (\alpha'_1, \ldots, \alpha'_k)^T$, wobei

$$\alpha'_j = \alpha_j + x_j - 1, \quad j = 1, \ldots, k.$$

b) Ist $\vec{X} = (X_1, \ldots, X_n)$ eine i.i.d. Stichprobe eines zum Parameter λ Poisson verteilten Untersuchungsmerkmals, besitzt $\vec{X}$ also die Wahrscheinlichkeitsfunktion

$$f(x|\lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1!x_2!\ldots x_n!} e^{-n\lambda},$$

und wählt man als Priori-Verteilung eine Gamma-Verteilung mit Parametern a und b, d.h. eine Verteilung mit der Dichte

$$\pi(\lambda) = \frac{b^a}{\underbrace{\Gamma(a)}_{\text{Norm.konst.}}} \lambda^{a-1} e^{-b\lambda}, \tag{2.51}$$

so ist die Posteriori-Verteilung eine Gamma-Verteilung mit den

Parameretern

$$a + \sum_{i=1}^{n} x_i \quad \text{und} \quad b + n.$$

Bsp. 2.96 (Normalverteilung)

Ist $\vec{X} = (X_1, \ldots, X_n)$ eine i.i.d. Stichprobe eines mit den Parametern μ und σ^2 normalverteilten Untersuchungsmerkmals, so gilt:

(i) Ist σ^2 bekannt und wählt man als Priori-Verteilung für μ eine Normalverteilung mit den Parametern ν und ρ^2, so ist die a posteriori Verteilung $\pi(\mu|\vec{x})$ eine Normalverteilung mit den Parametern ν' und $\rho^{2'}$ mit

$$\nu' = \frac{\bar{x}\rho^2 + \nu\dfrac{\sigma^2}{n}}{\rho^2 + \dfrac{\sigma^2}{n}} \tag{2.52}$$

und

$$\rho^{2'} = \frac{\rho^2 \cdot \dfrac{\sigma^2}{n}}{\rho^2 + \dfrac{\sigma^2}{n}} \; . \tag{2.53}$$

(ii) Ist μ bekannt, aber σ^2 unbekannt, so erhält man die konjugierte Verteilung, indem man $\frac{1}{\sigma^2}$ als gammaverteilt annimmt. Man sagt dann, σ^2 sei *invers gammaverteilt*.

Wie findet man solche konjugierten Paare?

Satz 2.97 (Zur Konjugiertheit in Exponentialfamilien)

Hat in der Situation von Def. 2.89 jedes Element der Menge $\mathcal{P}$ der Stichprobenverteilungen eine Dichte bzw. Wahrscheinlichkeitsfunktion $f(x|\vartheta)$ der Form

$$f(x|\vartheta) \propto h(\vartheta)\ \exp\big(T(x) \cdot b(\vartheta)\big) \tag{2.54}$$

und jedes Element der Menge Π, aus der die Priori-Verteilung stammt, eine Dichte bzw. Wahrscheinlichkeitsfunktion der Form

$$\pi(\vartheta) \propto [h(\vartheta)]^{\alpha} \exp\big(b(\vartheta) \cdot \beta\big), \tag{2.55}$$

so sind Π und $\mathcal{P}$ konjugiert. Es gilt dann

$$\pi(\vartheta|x) \propto [h(\vartheta)]^{\alpha+1} \cdot \exp\big((T(x) + \beta) \cdot b(\vartheta)\big). \tag{2.56}$$

Beweis:

(2.56) ergibt sich unmittelbar durch Anwenden der Formel für die Posteriori-Verteilung auf (2.54) und (2.55). Dann ist (2.56) mit $\alpha' := \alpha + 1$ und $\beta' := \beta + T(x)$ von der Form (2.55), also sind tatsächlich Π und $\mathcal{P}$ konjugiert.

Bem. 2.98 (zu Satz 2.97)

- Der Satz kann also direkt zur Konstruktion geeigneter, konjugierter Priori-Verteilungen verwendet werden, indem man die Stichprobenverteilung in die Form (2.54) bringt und dann eine Priori gemäß (2.55) wählt.

- $b(\vartheta)$ spielte in (2.54) und in (2.55) eine ganz unterschiedliche Rolle:
  In (2.54) ist $b(\vartheta)$ der natürliche Parameter der Exponentialfamilie, aus der die Likelihood / Stichprobenverteilung stammt.
  In (2.55) hingegen ist $b(\vartheta)$ die suffiziente Statistik für den natürlichen Parameter β der Exponentialfamilie, aus der die Priori stammt. (Bei der Priori ist ja der Wert von ϑ „zufällig".)

- Ähnliches gilt für $h(\vartheta)$.

Bsp. 2.99 (Beispiele zu Satz 2.97)

Man bestimme in folgenden Situationen unter Verwendung von Satz 2.97 jeweils eine konjugierte Priori-Verteilung:

a) $X_1, ...X_n$ ist i.i.d. normalverteilt mit unbekanntem μ und bekannter Varianz σ^2

b) X ist binomialverteilt zum unbekannten Parameter p

c) $X_1, ..., X_n$ ist i.i.d. Poisson-verteilt mit unbekanntem Parameter λ

2.4.8 (Reine) Bayes-Punktschätzung

Def. 2.100 (MPD-Schätzung)

Gegeben eine Beobachtung $\vec{x}$ und die Posteriori-Verteilung mit Dichte bzw. Wahrscheinlichkeitsfunktion $\pi(\vartheta|\vec{x})$ heißt $\widehat{\vartheta}$ mit

$$\pi(\hat{\vartheta}|\vec{x}) = \max_{\vartheta \in \Theta} \pi(\vartheta|\vec{x})$$

(reiner) Bayes-Schätzwert oder *Maximum (bzw Highest) Posteriori Density Schätzwert* (MPD- (bzw. HPD-) Schätzwert) oder *Posteriori-Modus-Schätzwert.* Die zugehörige Schätzfunktion $\hat{\vartheta}(\vec{X})$ heißt reine *Bayes-Schätzung* oder *MPD- (bzw. HPD-) Schätzung* bzw. *Posteriori-Modus-Schätzung.*

Bem. 2.101 (Zur MPD-Schätzung)

a) Ist die Posteriori-Verteilung unimodal, so ist $\widehat{\vartheta}$ der Modus der Posteriori.

b) Ist der Zustandsraum Θ beschränkt und liegt dem Schätzproblem als Priori-Verteilung eine Gleichverteilung zugrunde, so gilt

$$\pi(\vartheta|\vec{x}) \propto f(\vec{x}|\vartheta) \cdot \pi(\vartheta) = f(\vec{x}|\vartheta) \cdot \text{ Konstante}$$

D.h. der MPD-Schätzer ist dasjenige ϑ, das $f(x|\vartheta)$ maximiert, also der Maximum-Likelihood-Schätzwert.

c) Im Falle $\Theta = \mathbb{R}^+$ oder $\Theta = \mathbb{R}$ gibt es keine Gleichverteilung auf Θ, denn mit

$$f(x) = c \quad \text{ist} \qquad \int_0^\infty f(x)dx = \int_0^\infty c\, dx = \big[x\big]_0^\infty = \infty$$

unabhängig von $c > 0$.

Man kann aber zeigen, dass viele der zentralen Ergebnisse der Bayes-Theorie erhalten bleiben, wenn man auch nicht normierbare σ-finite Maße als Prioris zu lässt (z.B. Lebesque Maß $\lambda(\cdot)$; $\lambda([a, b]) := b - a$: „*improper priors*“)

Bsp. 2.102 (Beta-Binomialmodell)

$\pi(\vartheta|x_1,...,x_n)$ ist $B(a+\sum_{i=1}^n x_i; b+n-\sum_{i=1}^n x_i) =: B(a',b')$ -verteilt.
Hat man bei der Priori a=1=b gewählt, so ergibt sich mit $\dfrac{a'-1}{a'+b'-2}$ als Modus der Beta(a',b')-Verteilung der MPD-Schätzwert

$$\hat{\vartheta} = \frac{1+\sum_{i=1}^n x_i - 1}{1+\sum_{i=1}^n x_i + 1 + n - \sum_{i=1}^n x_i - 2} = \frac{1}{n}\sum_{i=1}^n x_i,$$

also in der Tat der ML-Schätzwert.

2.4.9 Der Hauptsatz der Bayes-Entscheidungstheorie

Def. 2.103 (Posteriori-Verlust-optimale Aktionen, konditionale Ba

Gegeben sei ein datenbasiertes Entscheidungsproblem $((\mathbb{A}, \Theta, l(\cdot)); (\mathcal{X}, A, (p_\vartheta)_{\vartheta\in\Theta}))$ und eine Priori-Verteilung $\pi(\cdot)$ über $(\Theta, \sigma(\Theta))$.
Eine Aktion $a_x^* \in \mathbb{A}$ heißt *Posteriori-Verlust optimal* zur *Beobachtung* $x \in \mathcal{X}$ oder *konditionale Bayes-Aktion zu* x *und der Priori-Verteilung* $\pi(\cdot)$, wenn gilt

$$\mathbb{E}_{\pi(\cdot|x)} l(a_x^*, \vartheta) \leq \mathbb{E}_{\pi(\cdot|x)} l(a, \vartheta) \quad \forall a \in \mathbb{A}.$$

a_x^* ist also sozusagen Bayes-Aktion zur Posteriori-Verteilung $\pi(\cdot|x)$ als „aufdatierter Priori-Verteilung“ $\pi(\cdot|x)$.

Analog definiert man eine *Posteriori-Nutzen-Optimalität.*

2 Arten, Bayes-Entscheidungstheorie zu betreiben

datengestütztes Entscheidungsproblem + Priori-Verteilung; Informationsbeschaffungsexperiment

Auswertungsproblem + Priori-Verteilung
komplexer Aktionenraum: alle Entscheidungsfunktionen

Bayes-optimale Entscheidungs**funktion** $d^* : \mathcal{X} \to \mathbb{A}$
($\to$ Test**funktion**, Schätz**funktion**)

konkrete Beobachtung x

Bayes optimale **Aktion**
$a^* = d^*(x)$

Priori-Verteilung

Stichprobenverteilung; Informationsbeschaffungsexperiment

konkrete Beobachtung

Posteriori-Verteilung

Bayes Postulat

Posteriori-Verlust optimale **Aktion** a^*_x,
z.B. reine (optimale) Bayes Schätzung $\hat{\vartheta}_x$
reiner/optimaler Bayes Test φ_x

?

„Priori-Risiko“ optimale Aktion

Satz 2.104 (Hauptsatz der Bayes-Entscheidungstheorie)

Gegeben sei ein datengestütztes Entscheidungsproblem $((\mathbb{A}, \Theta, \ell(\cdot)); (\mathcal{X}, \mathcal{A}, (p_\vartheta)_{\vartheta\in\Theta}))$, bestehend aus einem *datenfreien Entscheidungsproblem* $(\mathbb{A}, \Theta, \ell(\cdot))$ und einer Informationsstruktur $(\mathcal{X}, \mathcal{A}, (p_\vartheta)_{\vartheta\in\Theta})$ sowie eine Priori-Verteilung $\pi(\cdot)$ über $(\Theta, \sigma(\Theta))$.

Eine Entscheidungs**funktion**

$$\begin{aligned} d^* : \mathcal{X} &\longrightarrow \mathbb{A} \\ x &\longmapsto d^*(x) \end{aligned}$$

ist genau dann Bayes-optimal im zugehörigen Auswertungsproblem, wenn für jedes $x \in \mathcal{X}$ die zugehörige Aktion $d^*(x)$ Posteriori-Verlust optimal zur Beobachtung x ist.

<u>Beweis:</u> Für den diskreten Fall [7]

- Vorneweg eine Hilfsüberlegung: Suche die Lage des Minimums $\vec{z}_{min}$ einer Funktion $f(\vec{z})$ mit $\vec{z} = (z_1, ..., z_n)$, wobei $f(\vec{z}) = \sum_{i=1}^{n} c_i f_i(z_i)$, also die i-te Komponente von z nur im i-ten Summanden auftritt.

$$f(\vec{z}) = \sum_{i=1}^{n} c_i f_i(z_i) \to \min_{\vec{z}}$$

$$\Longleftrightarrow \; f_i(z_i) \longrightarrow \min_{z_i} \quad \text{für jedes i unabhängig von den anderen Summanden.}$$

- Der Deutlichkeit halber wird wieder eine Hilfsvariable U (vgl. Bem. 2.89) eingeführt und $p_\vartheta(\{X = x\})$ wird als $p(\{X = x\}|\{U = \vartheta\})$ geschrieben.

[7] für den allgemeinen Fall: siehe z.B. Rüger (1999, S. 283f.)

Angewendet auf Entscheidungsprobleme mit der Posteriori $\pi(\vartheta|x)$ ergibt sich mit dieser Notation

$$\pi(\vartheta|x) = \frac{p(\{X=x\}|\{U=\vartheta\}) \cdot \pi(\vartheta)}{p(\{X=x\})}. \tag{2.57}$$

Nun betrachte man Entscheidungsfunktionen $d(\cdot)$ im Auswertungsproblem:
Für die Risikofunktion

$$R(d,\vartheta) = \mathbb{E}_{p_\vartheta}(\ell(d(x),\vartheta)$$

gilt hier

$$R(d,\vartheta) = \sum_{x \in \mathcal{X}} \ell(d(x),\vartheta) \cdot p_\vartheta(\{X=x\}).$$

Die optimale Entscheidungsfunktion zur Priori $\pi(\cdot)$ minimiert unter allen d

$$\mathbb{E}_\pi(R(d,\vartheta)),$$

löst also

$$\sum_{\vartheta\in\Theta}\left(\sum_{x\in\mathcal{X}}\ell(d(x),\vartheta)\cdot p_\vartheta(\{X=x\})\right)\cdot\pi(\vartheta)\to\min_d$$

$$\iff \sum_{\vartheta\in\Theta} \textstyle\sum_{x\in\mathcal{X}} \left(\ell(d(x),\vartheta)\cdot \underbrace{p(\{X=x\}|U=\vartheta)\cdot\pi(\vartheta)}_{=\pi(\vartheta|x)\cdot p(\{X=x\})}\right) \to \min_d$$

$$\overset{(2.57)}{\iff} \underbrace{\sum_{x\in\mathcal{X}}}_{\triangleq\sum_{i=1}^n} \underbrace{\left(\sum_{\vartheta\in\Theta}\ell(d(x),\vartheta)\cdot\pi(\vartheta|x)\right)}_{\triangleq f_i(z_i)} \cdot \underbrace{p(\{X=x\})}_{\triangleq c_i;\ \text{priori-prädiktiv, marginal}} \to \min_d$$

- Wegen der Hilfsüberlegung ist dies äquivalent dazu, für jedes feste x

$$\sum_{\vartheta\in\Theta} l(d(x),\vartheta)\cdot\pi(\vartheta|x)$$

separat zu minimieren nach $a_x := d(x)$ für festes x.
Dies liefert jeweils genau die Posteriori-Verlust optimale Aktion, also die Bayes-Aktion zur Posteriori als neuer Priori.

Satz 2.105 (Bestimmung von Bayes-optimalen Entscheidungsfunktionen, z.B. Rüger (1999, Satz 2.20))

Gegeben sei das Schätzproblem als datengestütztes Entscheidungsproblem gemäß Kapitel 1.5 sowie eine Priori-Verteilung $\pi(\cdot)$.
Dann gilt:

i) Wählt man die quadratische bzw. absolute Verlustfunktion, so gilt für die Bayes-optimale Entscheidungsfunktion $d^*_{quad}(\cdot)$ bzw. $d^*_{abs}(\cdot)$:
Für jedes x ist $d^*_{quad}(\cdot)$ genau der Erwartungswert und $d^*_{abs}(\cdot)$ der Median der Posteriori-Verteilung $\pi(\vartheta|x)$.

ii) Die HPD-Schätzung ergibt sich näherungsweise für kleine ϵ, wenn man die sogenannte Toleranzverlustfunktion zum Grade ϵ verwendet:

$$\ell_\epsilon(\hat{\vartheta}, \vartheta) = \begin{cases} 1 & |\hat{\vartheta} - \vartheta| > \epsilon \\ 0 & |\hat{\vartheta} - \vartheta| \leq \epsilon \end{cases}$$

2.4.10 „Asymptotische Objektivität“ der konditionalen Bayes-Inferenz

Motivationsbeispiel: Betrachte (2.52) und (2.53) für $n \longrightarrow \infty$

$$\lim_{n\to\infty} \nu' = \frac{\bar{x}\rho^2 + 0}{\rho^2 + 0} = \bar{x} \tag{2.58}$$

$$\lim_{n\to\infty} \rho^{2'} = \frac{\rho^2 \cdot 0}{\rho^2 + 0} = 0 \tag{2.59}$$

1. Mal wieder kommt $\bar{X}$ raus.

2. Viel wichtiger: Die Grenzwerte (2.58) und (2.59) hängen nicht von ρ und ν, also von den Parametern der Priori-Verteilung ab. Hat man

eine sehr große Stichprobe, so „verschwindet der Einfluss der Priori-Verteilung“. Dies gilt ganz allgemein und wird oft als eine pragmatische Rechtfertigung dafür gesehen, bei großem Stichprobenumfängen „angenehme“, z.B. konjugierte, Prioris zu verwenden.

Satz 2.106 („Asymptotische Objektivität von Bayes-Verfahren“, „Konsistenzsatz“)

Sei $\Theta = \{\vartheta_1, \ldots, \vartheta_m\}$ ein endlicher Parameterraum und $\vec{X} = (X_1, \ldots, X_n)$ eine i.i.d. Stichprobe eines beliebig verteilten (reellwertigen) Untersuchungsmerkmals mit Dichten $f(x_i|\vartheta_{wahr}),\ \vartheta_{wahr} \in \Theta$.
Sei $\pi(\vartheta)$ die Wahrscheinlichkeitsfunktion der Priori-Verteilung auf Θ mit $\pi(\vartheta) > 0$ für alle ϑ. Dann gilt für die Wahrscheinlichkeitsfunktion der nach n Beobachtungen gebildeten Posteriori-Verteilung $\pi_n(\vartheta|x)$

$$\lim_{n\to\infty} \pi_n(\vartheta|x) = \begin{cases} 1 & \text{falls } \vartheta = \vartheta_{wahr} \\ 0 & \text{falls } \vartheta \neq \vartheta_{wahr} \end{cases}$$

Bem. 2.107 (Erneute kritische Diskussion des Bayes-Ansatzes)

2.5 Einige alternative Regeln (im Kontext der klassischen Entscheidungstheorie

2.5.1 Die Laplace Regel

Foto:
http://www.mathematik.de/ger/information/landkarte/gebiete/wahrscheinlichkeitstheorie/wahrscheinlichkeitstheorie.html
[Stand: 25.06.13]

Def. 2.108 (Laplace-Regel)

Gegeben sei das datenfreie Entscheidungsproblem $(\mathbb{A}, \Theta, u(\cdot))$ mit endlichem $\Theta = \{\vartheta_1, \ldots, \vartheta_m\}$.
Die Kriteriumsfunktionen

$$\begin{aligned} \tilde{\Phi} : \quad \mathbb{A} &\longrightarrow \mathbb{R} \\ a &\longmapsto \sum_{j=1}^{m} u(a, \vartheta_j) \end{aligned} \tag{2.60}$$

und

$$\begin{aligned} \Phi : \quad \mathrm{I\!A} &\longrightarrow \mathbb{R} \\ a &\longmapsto \frac{1}{m} \sum_{j=1}^{m} u(a, \vartheta_j) \end{aligned} \tag{2.61}$$

heißen *Laplace-Regel*.

Bem. 2.109

Die beiden Kriteriumsfunktionen (2.60) und (2.61) liefern dieselbe Ordnung auf der Aktionenmenge.

a) Die Kriteriumsfunktion (2.61) entspricht einer Bayes-Regel mit Priori-Verteilung $\pi(\cdot) = (\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$ (mit $m = |\Theta|$), also einer Gleichverteilung auf Θ.
Damit geometrisch: Höhenlinien senkrecht auf Equilisator-Linie.

b) Viele Eigenschaften der optimalen Aktion können deshalb aus den in Kapitel 2.4 formulierten Sätzen über Bayes-Regeln abgeleitet werden. (Zulässigkeit, Entbehrlichkeit randomisierter Aktionen,...)

c) Rechtfertigung durch „Prinzip vom unzureichenden Grund“ (Laplace): Wenn nichts dafür spricht, dass eines der Elementarereignisse wahrscheinlicher ist als die anderen, dann sind sie gleichwahrscheinlich, also

$$\pi(\{\vartheta_1\}) = \pi(\{\vartheta_2\}) = \ldots = \pi(\{\vartheta_m\}).$$

Da

$$\pi(\{\vartheta_1\}) + \pi(\{\vartheta_2\}) + \ldots + \pi(\{\vartheta_m\}) = 1$$

ist zwangsläufig

$$\pi(\{\vartheta_j\}) = \frac{1}{m}\,, \qquad j = 1, \ldots, m.$$

d) Verallgemeinerung auf unendliches Θ:
Theorie der nichtinformativen Priori-Verteilungen, siehe Bemerkung 2.111.

Bem. 2.110 (Beispiel und Kritik)

Abwandlung von Beispiel aus Kapitel 1.3.2, Lotterie
Urnen bestehend aus einer unbekannten Anzahl von grünen, blauen und restlichen (rote, schwarze, violette) Kugeln.
Man kann entweder

- a_1 nicht spielen,
- a_2 zum Preis von $c_g = 60$€ auf grün setzen oder
- a_3 zum Preis von $c_b = 90$€ auf blau setzen.

Es wird eine Kugel zufällig gezogen. Man erhält 240€, wenn die Kugel, auf die man gesetzt hat, gezogen wird.

	$\{g\}$	$\{b\}$	$\{\text{rest}\}$		
a_1	0	0	0	0	
a_2	180	-60	-60	20	$\leftarrow$ optimale Aktion
a_3	-90	150	-90	-10	

Bem. 2.111 („Nichtinformative“ Priori-Verteilung und ihr Inform

In derkonditionalen Inferenz gibt es verschiedene Versuche, ähnlich der Laplace-Regel, „nichtinformative“ Priori-Verteilungen zu definieren und diese dann als Standardbewertungen heranzuziehen.

- z.B. die Gleichverteilung, diese ist aber nicht invariant gegenüber Transformationen des Parameters. Man hat dann also „keine Information“ über ϑ, aber eine informative Priori z.B. über eine bijektive, nichtlineare Transformation von ϑ.

- z.B. Verteilungen, die invariant bezüglich bijektiver Transformationen des Parameters sind (Jeffrey-Regel).

- z.B. Verteilungen, die die Entropie maximieren (Jaynes-Regel)

- Ganz neue Möglichkeiten ergeben sich beim Übergang zu Credalmengen (siehe Kapitel 3).

2.5.2 Die Minimax-Regret-Regel von L.J. Savage, auch Niehans-Savage-Regel genannt

fnewpage

Def. 2.112 (Minimax-Regret-Aktion)

Gegeben sei ein datenfreies Entscheidungsproblem $(\mathbb{A}, \Theta, u(\cdot))$ in Nutzenform bzw. $(\mathbb{A}, \Theta, l(\cdot))$ in Verlustform.
Seien $\mathbb{A}$ und Θ endlich, $\Theta = \{\vartheta_1, \ldots, \vartheta_m\}$, $\mathbb{A} = \{a_1, \ldots, a_n\}$.
Das datenfreie Entscheidungsproblem $(\mathbb{A}, \Theta, r(\cdot))$ in Verlustform mit

$$\begin{aligned} r : \mathbb{A} \times \Theta &\longrightarrow \mathbb{R} \\ (a_i, \vartheta_j) &\longmapsto r(a_i, \vartheta_j) \end{aligned}$$

und

$$r(a_i, \vartheta_j) = \max_{\ell=1,\dots,n} (u(a_\ell, \vartheta_j)) - u(a_i, \vartheta_j) \tag{2.62}$$

bzw.

$$r(a_i, \vartheta_j) = l(a_i, \vartheta_j) - \min_{\ell=1,\dots,n} (l(a_\ell, \vartheta_j)) \tag{2.63}$$

heißt **induziertes Regret Problem** bzw. **induzierte Regret-Tafel**.

Jedes $a^* \in \mathbb{A}$ mit

$$\max_{j=1,\dots,m} r(a^*, \vartheta_j) \le \max_{j=1,\dots,m} r(a, \vartheta_j) \quad \text{für alle } a \in \mathbb{A}$$

heißt **Minimax-Regret-Aktion**.

Bsp. 2.113 (Beispiel und Kritik)

Proposition 2.114 (Bayes-Aktionen in Regrettafeln)

Gegeben sei ein datenfreies Entscheidungsproblem $(\mathbb{A}, \Theta, u(\cdot))$ bzw. $(\mathbb{A}, \Theta, l(\cdot))$ mit $\mathbb{A} < \infty$ und $|\Theta| < \infty$ und die Priori-Bewertung $\pi(\cdot)$. Eine Aktion a^* ist genau dann Bayes-Aktion zu $\pi(\cdot)$, wenn a^* Bayes-Aktion zu $\pi(\cdot)$ im induzierten Regret-Problem ist.

2.5.3 Das Hurwicz-Kriterium

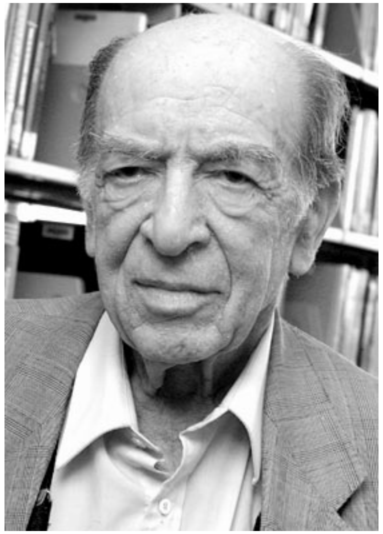


Foto: http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2007/hurwicz-facts.html
[Stand: 25.06.13]

Def. 2.115 (Hurwicz-Kriterium)

Gegeben sei ein datenfreies Entscheidungsproblem $(\mathbb{A}, \Theta, u(\cdot))$ mit $|\Theta| < \infty$, sowie $\alpha \in [0, 1]$.

$$\begin{aligned} \Phi : \mathbb{A} &\longrightarrow \mathbb{R} \\ a &\longmapsto \Phi(a) \end{aligned}$$

mit

$$\Phi(a) = \alpha \max_j u(a, \vartheta_j) + (1 - \alpha) \min_j u(a, \vartheta_j) \tag{2.64}$$

heißt *Hurwicz-Kriterium* zum *Optimismusparameter* α.

2.5.4 Das Erfahrungskriterium von J.L. Hodges und E.L. Lehmann (1952)

Def. 2.116 (Erfahrungskriterium von Hodges & Lehmann)

Gegeben sei ein datenfreies Entscheidungsproblem $(\mathbb{A}, \Theta, u(\cdot))$, $|\Theta| < \infty$, $\mu \in [0, 1]$ und eine Priori-Bewertung $\pi(\cdot)$.

$$\begin{aligned} \Phi : \mathbb{A} &\longrightarrow \mathbb{R} \\ a &\longmapsto \Phi(a) \end{aligned}$$

mit

$$\Phi(a) = \mu \cdot \left(\sum_{j=1}^{m} u(a, \vartheta_j) \pi(\{\vartheta_j\}) \right) + (1 - \mu) \cdot \left(\min_j (u(a, \vartheta_j)) \right) \quad (2.65)$$

heißt *Erfahrungskriterium von Hodges und Lehmann zum Vertrauensparameter* μ.

2.6 Gleichmäßig beste Verfahren in der statistischen Entscheidungstheorie

Def. 2.117 (Gleichmäßig beste Verfahren)

Gegeben sei das auf eine Menge $\mathcal{D}_0$ eingeschränkte Auswertungsproblem $(\mathcal{D}_0, \Theta, R(\cdot))$ eines datengestütztes Entscheidungsproblem $((\mathbb{A}, \Theta, l(\cdot)); (\mathcal{X}, \sigma(\mathcal{X}), (p_\vartheta)_{\vartheta\in\Theta}))$

Eine Entscheidungsfunktion $d^* \in \mathcal{D}_0$ heißt *gleichmäßig bestes Verfahren* aus $\mathcal{D}_0$, wenn für die Risikofunktion gilt:

$$R(d^*, \vartheta) \leq R(d, \vartheta) \quad \text{für alle } \vartheta \in \Theta \text{ und alle } d \in \mathcal{D}_0. \tag{2.66}$$

Bem. 2.118 (zu Def. 2.117)

i) Man beachte, dass die Risikofunktion von d^* für *alle* $\vartheta \in \Theta$ nicht größer sein soll; d^* soll also im zugehörigen Auswertungsproblem $(\mathcal{D}_0, \Theta, R(\cdot))$ alle Elemente von $\mathcal{D}_0$ dominieren.

ii) Dies ist für „großes $\mathcal{D}_0$" eine extrem starke Forderung – insbesondere im Lichte der elementaren Beispiele aus Kapitel 1.3 – aber v.a. bei Exponentialfamilie und geeigneter Einschränkung der Menge der betrachteten Entscheidungsfunktionen möglich (UMVU-Schätzer, gleichmäßig bester Test, siehe später).

Satz 2.119 (Lehmann und Scheffé (1950))

Gegeben sei ein Schätzproblem im Sinne Beispiel 1.42. Ferner sei

- $(p_\vartheta)_{\vartheta \in \Theta}$ eine strikt q-parametrige Exponentialfamilie in $T(\vec{x}) = (T_1(\vec{x}), ..., T_q(\vec{x}))$.

- $\ell(\hat{\vartheta}, \vartheta)$ konvex in $\hat{\vartheta}$ für alle ϑ.

Betrachtet man die Schätzung einer Transformation $\gamma(\vartheta)$ von ϑ und die Klasse $\mathcal{D}_\gamma$ aller für $\gamma(\vartheta)$ erwartungstreuen Schätzer, so gilt:
Ist $\mathcal{D}_\gamma$ nicht leer, so gibt es eine nichtrandomisierte Schätzfunktion der Form $\eta(T(\vec{X}))$, die gleichmäßig beste in der Klasse $\mathcal{D}_\gamma$ ist.

Ist umgekehrt $\eta(T(\vec{X}))$ eine erwartungstreue Schätzfunktion für $\gamma(\vartheta)$, so ist $\eta(T(\vec{X}))$ gleichmäßig bestes Verfahren in D_γ.

Bem. 2.120 (Zur Interpretation des Satzes)

- „Informationsdeutung“:

- Konstruktive Anwendung:

Korollar 2.121

Gegeben sei eine strikt q-parametrige Exponentialfamilie in $T(\vec{x}) = (T_1(\vec{x}), ..., T_q(\vec{x}))$.
Dann ist für jede (messbare) Funktion $\eta(\cdot)$ der Schätzer $\eta(T_1, ..., T_q)$ UMVU-Schätzer für $\mathbb{E}_\vartheta(\eta(T_1, ..., T_q))$.

Bem. 2.122 (Zum Satz von Lehmann-Scheffé)

Die Beschränkung auf erwartungstreue Schätzer ist wesentlich.

Bem. 2.123

Korollar 2.121 wird typischerweise „andersherum“ angewendet. Will man eine Funktion $\gamma(\vartheta)$ schätzen, so sucht man eine Funktion $g(T)$, so dass $\mathbb{E}g(T) = \gamma(\vartheta)$. Gemäß Korollar 2.121 weiß man dann, dass $g(T)$ UMVU für $\gamma(\vartheta)$ ist.
Üblicherweise wird man versuchen einen Ansatz zu wählen, bei dem sich $g(\cdot)$ aus „einfachen Grundfunktionen“ zusammensetzt.

Bsp. 2.124

Gegeben sei eine i.i.d. Stichprobe eines normalverteilten Merkmals mit unbekanntem Mittelwert μ, aber bekannter Varianz σ^2. Man bestimme einen UMVU Schätzer
a) für μ und
b) für $\exp(\mu)$.

Satz 2.126 (Optimale Tests)

Betrachtet werde das Testproblem als Entscheidungsproblem gemäß Beispiel 1.44 mit

a_0 für H_0 entscheiden
a_1 für H_1 entscheiden

	ϑ	
a_0	0	1
a_1	1	0

$$l(a_0, \vartheta) = \begin{cases} 0 & \vartheta \in \Theta_0 \\ 1 & \vartheta \in \Theta_1 \end{cases}$$

$$l(a_1, \vartheta) = \begin{cases} 1 & \vartheta \in \Theta_0 \\ 0 & \vartheta \in \Theta_1 \end{cases}$$

Ferner sei $\Theta_0 = \{\vartheta | \vartheta \leq \vartheta_0\}, \Theta_1 = \{\vartheta | \vartheta \geq \vartheta_1\}, \vartheta_1 > \vartheta_0$, und ein Signifikanzniveau $\alpha \in (0,1)$ vorgegeben.
Bildet $(P_\vartheta^{\oplus n})$ eine strikt einparametrige Exponentialfamilie in T (mit dem natürlichen Parameter $c(\vartheta)$, der in eineindeutiger Beziehung zu ϑ stehe), so gibt es ein $\kappa \in \mathbb{R}$, so dass der Test

$$\varphi^*(\vec{x}) = \begin{cases} 1 & \mathrm{T}(\vec{x}) > \kappa \\ \gamma & \mathrm{T}(\vec{x}) = \kappa \\ 0 & \mathrm{T}(\vec{x}) < \kappa \end{cases} \tag{2.67}$$

mit

$$\mathbb{E}_{\vartheta_0}\varphi^\star = \alpha \tag{2.68}$$

gleichmäßig bester Test (UMP) ist, d.h es gilt

$$\mathbb{E}_\vartheta\varphi^\star \geq \mathbb{E}_\vartheta\varphi, \quad \forall\vartheta \in \Theta_1,$$

für alle φ mit $\sup\limits_{\vartheta\in\Theta_0} \mathbb{E}_{\vartheta_0}\varphi \leq \alpha$.

Bem. 2.127 (Zur Interpretation von Satz 2.126)

- Informationsdeutung:

- konstruktive Anwendung:

Bem. 2.128

- Die Aussage gilt allgemeiner für Verteilungen mit *monotonen Dichtequotienten*; bei denen also für alle $\vartheta_1 \in \Theta_1$, $\vartheta_0 \in \Theta_0$ und eine geeignete Funktion $T(\vec{X})$ der Quotient $\dfrac{f_{T||\vartheta_1}(t)}{f_{T||\vartheta_0}(t)}$ monoton in t ist.

- Auch bei der Fragestellung $H_0 : \vartheta = \vartheta_0$ gegen $H_1 : \vartheta \neq \vartheta_0$, gibt es bei Exponentialfamilien einen gleichmäßig besten Test, wenn man sich auf *unverfälschte* Tests (d.h. Gütefunktion $\geq \alpha$ für alle Elemente der Alternative) beschränkt.
  Ähnliches gilt in der Situation $H_0 : \vartheta \in [\underline{\vartheta}_0, \bar{\vartheta}_0]$ gegen $H_1 : \vartheta \in [\underline{\vartheta}_1, \bar{\vartheta}_1]$, wenn man nur unverfälschte und ähnliche Tests (Gütefunktion am Rand der Nullhypothese $= \alpha$) betrachtet.

Bsp. 2.130 Anwendung von Satz 2.126 auf den Mittelwertstest bei der Normalverteilung und auf das Testen bei der Binomialverteilung

a) $X_1, \ldots, X_n \quad i.i.d. \qquad \sim N(\mu, \sigma^2)$

$H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu \geq \mu_1 > \mu_0$ σ^2 bekannt

$$f_{X_1,\ldots,X_{n||\mu}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2\sigma}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n}_{\substack{\text{konstant}\\ \text{(bei bek. } \sigma^2)}} \cdot \frac{1}{\sigma^n} \cdot \underbrace{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right)}_{\substack{\text{nur von } X \text{ abhängig}\\ \text{(bei bekanntem } \sigma^2)}} \cdot \underbrace{\exp\left(-\frac{1}{2\sigma^2} n\mu^2\right)}_{\substack{\text{nur von } \mu \text{ abh.}\\ \text{nicht von } x}} \cdot \exp\left(\frac{\mu}{\sigma^2} \cdot \sum_{i=1}^{n} x_i\right)$$

natürlicher Parameter $\dfrac{\mu}{\sigma^2}$

Suffiziente Statistik: $T = \displaystyle\sum_{i=1}^{n} X_i$

Die kritische Region $K(\kappa)$, d.h. der Bereich aller $\vec{x}$ mit $\varphi(\vec{x}) = 1$, ist wegen (3.5) von der Form

$$K(\kappa) := \{\vec{x} | T(\vec{x}) > \kappa\}$$

mit κ so, dass

$$P_{\mu_0}(K(\kappa)) \stackrel{!}{=} \alpha.$$

(Man weiß ja wegen (2.68), dass die maximale Fehlerwahrscheinlichkeit an der Grenze der Nullhypothese angenommen wird.)

Das heißt, es soll gelten

$$P_{\mu_0}(K(\kappa)) = P\mu_0(\{\vec{x}|T(\vec{x}) > \kappa\}) = P\mu_0(\{\vec{x}|\sum_{i=1}^{n} X_i > \kappa\}) = \alpha.$$

Zur Berechnung dieser Wahrscheinlichkeit benutzt man

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2),$$

d.h.

$$\frac{\sum_{i=1}^{n} X_i - n\mu_0}{\sigma\sqrt{n}} \sim N(0,1).$$

Also führt der Ansatz $P_{\mu_0}(K(\kappa)) \stackrel{!}{=} \alpha$, d.h.

$$P\left(\frac{\sum_{i=1}^n X_i - n\mu_0}{\sigma \cdot \sqrt{n}} \stackrel{!}{>} \underbrace{\frac{\kappa - n\mu}{\sigma \cdot \sqrt{n}}}_{\kappa' = \tau_\alpha}\right) \stackrel{!}{=} \alpha,$$

Fraktil der Normalverteilung

dazu, dass die kritische Region so zu wählen ist, dass

$$\begin{aligned} \frac{\sum_{i=1}^{n} X_i - n \cdot \mu}{\sigma \cdot \sqrt{n}} &> \tau_\alpha \\ \iff \frac{\bar{X} - \mu}{\sigma} \cdot \sqrt{n} &> \tau_\alpha \\ \iff \bar{X} > \mu_0 + \frac{\tau_\alpha \cdot \sigma}{\sqrt{n}}. \end{aligned}$$

Da ferner $\mu_1 > \mu_0$ bei dieser Konstruktion beliebig gewählt werden konnte, gilt die Aussage für *alle* $\mu > \mu_0$.

Beachte: es ergibt sich die übliche kritische Region des Gauss Tests, dieser ist also UMP.

b) Bei der Bernoulliverteilung sei zu einem konkreten Beispiel übergegangen.
Man testet $H_0 : p \leq 0.5$ gegen $H_1 : p \geq 0.6$, wobei bei der i.i.d. Stichprobe $X_1, \ldots X_n$ der Stichprobenumfang $n = 5$ sei und das Signifikanzniveau auf $\alpha = 0.1$ gesetzt sei.
Zur Anwendung von Satz 2.126 bringt man zunächst die Wahrscheinlichkeitsfunktion $f_p(\cdot)$ der Bernoulliverteilung auf

„Exponentialfamilien-Gestalt“.

$$\begin{aligned}
f_p(x_1,\ldots,x_n) &= \prod_{i=1}^{n} \cdot p^{x_i}\cdot(1-p)^{1-x_i} &&= \\
&= (1-p)^n\cdot\left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} &&= \\
&= (1-p)^n\cdot\exp\left(\ln\left(\left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i}\right)\right) &&= \\
&= (1-p)^n\cdot\exp\left(\sum_{i=1}^{n} x_i\cdot\ln\left(\frac{p}{1-p}\right)\right)
\end{aligned}$$

Man erhält $T = \sum\limits_{i=1}^{n} X_i$, und $\ln(\frac{p}{1-p})$ ist der natürliche Parameter.

Der Ansatz

$$K(\kappa) := \{\vec{x}|T(\vec{x}) > \kappa\}$$

für die kritische Region führt auf ein Problem. Da $T = \sum_{i=1}^{n} X_i$ binomialverteilt ist, ergibt sich [8]

$$P_{p_0}(K(\kappa)) = P_{p_0}(\{T > \kappa\}) = \sum_{\substack{j>\kappa \\ j\in\mathbb{N}}}^{n} \binom{5}{j} \underbrace{0.5^j \cdot 0.5^{5-j}}_{0.5^5}$$

[8] Wegen (2.68) setzt man hier wieder den oberen Randwert der Nullhypothese ein, als denjenigen Wert der Nullhypothese, der am schwersten von H_1 zu unterscheiden ist.

Allerdings gibt es kein κ, so dass diese Gleichung erfüllt ist:

$$\begin{aligned}
\kappa &\in (4,5] \quad P(K(\kappa)) = 0 \\
\kappa &= 4 \Rightarrow P(K(\kappa)) = \begin{pmatrix} 5 \\ 5 \end{pmatrix} \cdot (\tfrac{1}{2})^5 = \tfrac{1}{32} < 0.1 \\
\kappa &\in (3,4] \Rightarrow P(K(\kappa)) = P(K(4)) = \tfrac{1}{32} \\
\kappa &= 3 \Rightarrow P(K(\kappa)) = \tfrac{1}{32} + \begin{pmatrix} 5 \\ 4 \end{pmatrix} \cdot 0.5^5 = \tfrac{1}{32} + \tfrac{5}{32} = \tfrac{6}{32} > 0.1
\end{aligned}$$

Was tun? Klar ist

$\varphi(\vec{x}) = 1$, d.h. H_0 ablehnen, für alle $\vec{x}$ mit $T(\vec{x}) = 5$ und
$\varphi(\vec{x}) = 0$, d.h. H_0 nicht ablehnen, für alle $\vec{x}$ mit $T(\vec{x}) \leq 3$.

Ist $T(\vec{x}) = 4$, dann hat man so zu randomisieren, dass $\mathbb{E}_{0.5}\varphi = \alpha$. Man setzt hierzu $\varphi(\vec{x}) = \gamma$, falls $T(\vec{x}) = 4$, und erhält:

$$\begin{aligned}\mathbb{E}_{0.5}\varphi \;&=\; 1 \cdot P(\varphi(x) = 1) + \gamma \cdot P(\varphi(x) = \gamma) + \\ &\quad +0 \cdot P(\varphi(x) = 0) \overset{!}{=} \alpha\end{aligned}$$

also

$$\begin{aligned}\gamma \;&=\; \frac{\alpha - P(\varphi(x) = 1)}{P(\varphi(x) = \gamma)} = \\ &=\; \frac{\alpha - P(T = 5)}{P(T = 4)} = \frac{0.1 - \frac{1}{32}}{\frac{5}{32}} = 0.44\ .\end{aligned}$$