Vorlesung Statistik II

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sommersemester 2026

Beispiel: Bernoulli-Verteilung

Die Likelihood der Daten $x_1, \ldots, x_n$ lautet

$$L(p) = p^{x_1+\ldots+x_n} \cdot (1-p)^{n-(x_1+\ldots+x_n)} = p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n-\sum_{i=1}^n x_i}$$

Für die Log-Likelihood ergibt sich dann

$$\log L(p) = \log p \cdot \sum_{i=1}^n x_i + \log(1-p) \cdot \left(n - \sum_{i=1}^n x_i \right)$$

Für die Ableitung des Logarithmus gilt

$$\frac{d \log p}{dp} = \frac{1}{p} \quad \text{und} \quad \frac{d \log(1-p)}{dp} = \frac{1}{1-p} \cdot (-1)$$

Für die Ableitung der Log-Likelihood erhält man

$$\frac{d \log L(p)}{dp} = \frac{1}{p} \cdot \sum_{i=1}^n x_i + \frac{1}{1-p} \cdot (-1) \cdot \left(n - \sum_{i=1}^n x_i \right)$$

Beispiel: Bernoulli-Verteilung

Nullsetzen der Ableitung liefert für den Maximum-Likelihood-Schätzer $\hat{p}$ dann

$$\frac{1}{\hat{p}} \cdot \sum_{i=1}^{n} x_i - \frac{1}{1-\hat{p}} \cdot \left(n - \sum_{i=1}^{n} x_i \right) = 0$$

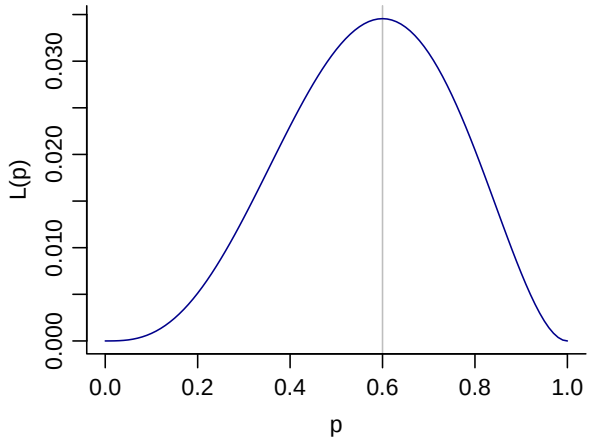
Daraus errechnet man durch elementare Umformungen

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

Das bedeutet, dass die relative Häufigkeit der Maximum-Likelihood-Schätzer für den Parameter p einer Bernoulli-Verteilung ist

Beispiel: Bernoulli-Verteilung

Für das obige Beispiel mit dem Stichprobenumfang $n = 5$ und den Realisierungen
$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0$
liegt das Maximum der Likelihood bei $\hat{p} = 3/5 = 0.6$



Beispiel: Normalverteilung

Seien $X_1, \ldots, X_n$ unabhängig identisch $N(\mu, \sigma^2)$-verteilte Zufallsvariablen

Dann lautet die Likelihood für die Realisierungen $x_1, \ldots, x_n$

$$
\begin{aligned}
L(\mu, \sigma) &= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right) \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right) \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right]
\end{aligned}
$$

Beispiel: Normalverteilung

Für die Log-Likelihood hat man dann

$$\begin{aligned}\log L(\mu,\sigma) &= \sum_{i=1}^{n}\left[\log\frac{1}{\sqrt{2\pi}\,\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right]\\ &= \sum_{i=1}^{n}\left[-\log\sqrt{2\pi} - \log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}\right]\end{aligned}$$

Da der Parameter $\vartheta = (\mu,\sigma)$ in diesem Fall ein Vektor ist, muss die Log-Likelihood sowohl nach μ als auch nach σ partiell abgeleitet werden

Beispiel: Normalverteilung

Partielles Differenzieren nach μ liefert

$$\begin{aligned}\frac{\partial \log L(\mu,\sigma)}{\partial \mu} &= \sum_{i=1}^{n} 2 \cdot \left(-\frac{x_i-\mu}{2\sigma^2}\right) \cdot (-1) \\ &= \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2}\end{aligned}$$

und damit erhält man

$$\sum_{i=1}^{n} \frac{x_i-\hat{\mu}}{\hat{\sigma}^2} = 0$$

Beispiel: Normalverteilung

Dann aber ist

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = \sum_{i=1}^{n} x_i - n \cdot \hat{\mu} = 0$$

Damit ergibt sich als ML-Schätzer für den Parameter μ

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

Das arithmetische Mittel ist also der ML-Schätzer für den Erwartungswert μ einer Normalverteilung

Beispiel: Normalverteilung

Partielles Differenzieren nach σ liefert

$$\begin{aligned}\frac{\partial \log L(\mu,\sigma)}{\partial \sigma} &= \sum_{i=1}^{n}\left(-\frac{1}{\sigma}-\frac{(-2)\cdot(x_i-\mu)^2}{2\sigma^3}\right)\\ &= \sum_{i=1}^{n}\left(-\frac{1}{\sigma}+\frac{(x_i-\mu)^2}{\sigma^3}\right)\end{aligned}$$

und damit erhält man

$$-\frac{n}{\hat{\sigma}}+\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\hat{\sigma}^3}=0$$

Beispiel: Normalverteilung

Daraus folgt

$$\sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{\hat{\sigma}^3} = \frac{n}{\hat{\sigma}}$$

$$\sum_{i=1}^{n} (x_i - \hat{\mu})^2 = n \cdot \hat{\sigma}^2$$

und als ML-Schätzer der Varianz einer Normalverteilung hat man

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

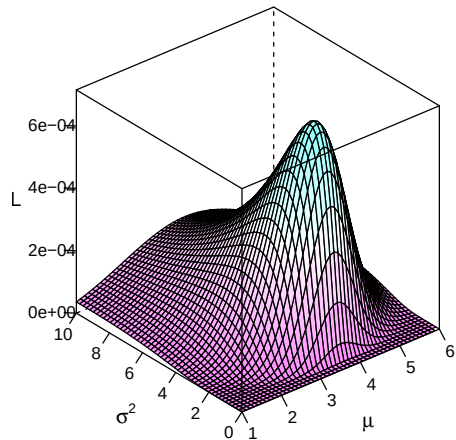
Der resultierende ML-Schätzer für die Varianz σ^2 zeigt, dass
Maximum-Likelihood-Schätzfunktionen nicht notwendig erwartungstreu sein müssen

Beispiel: Normalverteilung

Betrachten wir nun ein konkretes Beispiel mit den folgenden $n = 4$ Realisierungen unabhängig identisch normalverteilter Zufallsvariablen

$$x_1 = 2, x_2 = 4, x_3 = 3, x_4 = 6$$

Dann ergibt sich in Abhängigkeit von den Parametern μ und σ^2 folgende Likelihood, deren Maximum bei $\hat{\mu} = \bar{x} = 3.75$ und $\hat{\sigma}^2 = 2.19$ liegt



Beispiel: Normalverteilung

Erstellung der Grafik

```
## Dataframe mit Parameterwerten
g <- expand.grid( m = seq(1,  6, length.out = 50),   # mu
                  s2 = seq(0, 10, length.out = 50))  # sigma2

## Likelihoodfunktion und Beobachtungen
L <- function(x, mu, sigma2)
       prod(dnorm(x, mu, sqrt(sigma2)))
x <- c(2, 4, 3, 6)

## Abbildung
vL <- Vectorize(L, c("mu", "sigma2"))
lattice::wireframe(vL(x, m, s2) ~ m * s2, data = g, drape = TRUE,
  aspect = c(1, 1), scales = list(arrows = FALSE), screen = list(z = 40,
  x = -60), distance = 0.0, zoom = 0.92, colorkey = FALSE,
  xlab = expression(mu), ylab = expression(sigma^2), zlab = "L")
```

Beispiel: Alles-Oder-Nichts-Modell

Für ein einzelnes Item bestehen die Daten aus einer 0-1-Sequenz (x_t) über die Durchgänge $t = 1, \ldots, m$

Für n Items betrachtet man dann also die Sequenzen $(x_t)_i$ mit $i = 1, \ldots, n$

Wir führen folgende Zufallsvariablen ein

T_i	Gesamtzahl der Fehler beim i-ten Item
H_i	Position des letzten Fehlers beim i-ten Item

und

$$Z_i = \begin{cases} 1 & \text{falls } T_i = 0 \\ 0 & \text{falls } T_i > 0 \end{cases}$$

Beispiel: Alles-Oder-Nichts-Modell

Dann gilt für die Wahrscheinlichkeit für die Beobachtung der Sequenz $(x_t)_i$ beim i-ten Item

$$P((x_t)_i \mid \alpha) = (1-\alpha)^{H_i-1+Z_i}(1-g)^{T_i} g^{H_i-T_i+Z_i} b$$

Neben dem Lernparameter α und der Ratewahrscheinlichkeit g enthält die Formel auch die Wahrscheinlichkeit

$$b = \frac{\alpha}{1-(1-\alpha)\,g}$$

dass einer richtigen Antwort im Ratezustand keine falschen Antworten folgen

Beispiel: Alles-Oder-Nichts-Modell

Als Likelihood des gesamten Datensatzes ergibt sich

$$L(\alpha) = \prod_{i=1}^{n} P((x_t)_i \mid \alpha)$$

und mit

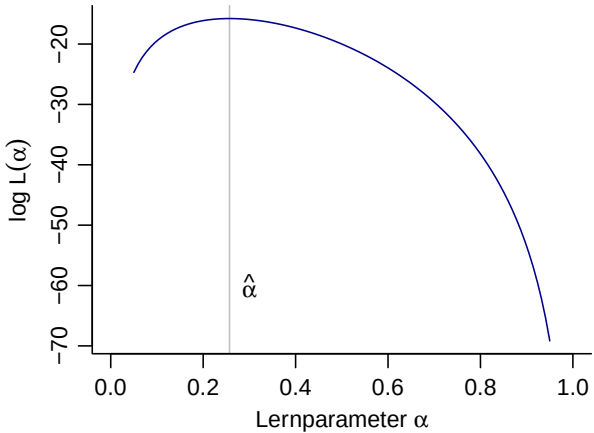
$$H = \sum_{i=1}^{n} H_i, \quad T = \sum_{i=1}^{n} T_i, \quad Z = \sum_{i=1}^{n} Z_i$$

kann man diese ausdrücken durch

$$L(\alpha) = (1-\alpha)^{H-n+Z}(1-g)^{T} g^{H-T+Z} b^{n}$$

Beispiel: Alles-Oder-Nichts-Modell

Grafische Darstellung der Log-Likelihood für die oben angegebenen Daten



Beispiel: Alles-Oder-Nichts-Modell

Ableitung der Log-Likelihood $\log L(\alpha)$ nach dem Lernparameter α und Nullsetzen der Ableitung liefert dann

$$-(\bar{H} - 1 + \bar{Z})\, g\, \hat{\alpha}^2 - (1 - g)\,(\bar{H} + \bar{Z})\, \hat{\alpha} + (1 - g) = 0$$

mit den Mittelwerten

$$\bar{H} = \frac{H}{n} = \frac{1}{n} \sum_{i=1}^{n} H_i, \quad \bar{Z} = \frac{Z}{n} = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

Beispiel: Alles-Oder-Nichts-Modell

Die einzige positive Lösung dieser quadratischen Gleichung ergibt als ML-Schätzer für den Lernparameter α im Alles-Oder-Nichts-Modell

$$\hat{\alpha} = \frac{\sqrt{(1-g)^2\,(\bar{H}+\bar{Z})^2 + 4\,g\,(\bar{H}-1+\bar{Z})\,(1-g)} - (1-g)\,(\bar{H}+\bar{Z})}{2\,g\,(\bar{H}-1+\bar{Z})}$$

Als ML-Schätzung für die oben angegebenen Daten berechnet man dann $\hat{\alpha} = 0.26$

Dieser Wert weicht numerisch von dem nach der Moment-Methode erhaltenen ab

Die Berechnung der Likelihood zeigt: Die Zufallsvariable T ist nicht suffizient für den Parameter α

Eigenschaften von Maximum-Likelihood-Schätzern

Die Herleitung des ML-Schätzers für die Varianz σ^2 der Normalverteilung zeigt, dass ML-Schätzfunktionen nicht notwendig erwartungstreu sind

Allgemein haben Maximum-Likelihood-Schätzfunktionen für einen Parameter ϑ aber folgende asymptotische Eigenschaften

- ML-Schätzer sind asymptotisch erwartungstreu, d. h. der Bias verschwindet mit wachsendem Stichprobenumfang $n \to \infty$
- ML-Schätzer sind asymptotisch effizient, d. h. für $n \to \infty$ gibt es keine erwartungstreue Schätzfunktion mit kleinerer Varianz
- ML-Schätzer sind asymptotisch normalverteilt, d. h. für $n \to \infty$ geht die Verteilung der Schätzfunktion gegen eine Normalverteilung mit Erwartungswert ϑ und einer Varianz, die durch die Fisher-Information bestimmt ist

Varianz des Maximum-Likelihood-Schätzers

Die Varianz eines ML-Schätzers wird über die zweite Ableitung (Krümmung!) der Log-Likelihood bestimmt

$$H = \frac{\partial^2 \log L(\vartheta)}{\partial \vartheta^2}$$

Ist ϑ ein Vektor, so wird H eine Matrix von partiellen Ableitungen, die *Hesse-Matrix*

Die Fisher-Information ist dann der Erwartungswert der negativen Hesse-Matrix

$$I(\vartheta) = E(-H)$$

und die Varianz (bzw. Kovarianzmatrix) ist asymptotisch $I(\vartheta)^{-1}$

Die Varianz des ML-Schätzers verhält sich also invers zur Information: Je mehr Information bzgl. ϑ bereit steht, desto kleiner ist die Varianz von $\hat{\vartheta}$

Beispiel: Bernoulli-Verteilung

Für die zweite Ableitung der Log-Likelihood erhält man

$$H = \frac{d^2 \log L(p)}{dp^2} = -\frac{1}{p^2} \cdot \sum_{i=1}^{n} x_i - \frac{1}{(1-p)^2} \cdot \left(n - \sum_{i=1}^{n} x_i \right)$$

Damit ergibt sich für die Fisher-Information

$$\begin{aligned} I(p) &= E\left(\frac{1}{p^2} \cdot \sum_{i=1}^{n} x_i + \frac{1}{(1-p)^2} \cdot \left(n - \sum_{i=1}^{n} x_i \right) \right) \\ &= \frac{1}{p^2} \cdot \sum_{i=1}^{n} p + \frac{1}{(1-p)^2} \cdot \left(n - \sum_{i=1}^{n} p \right) \\ &= \frac{n}{p} + \frac{n}{1-p} \end{aligned}$$

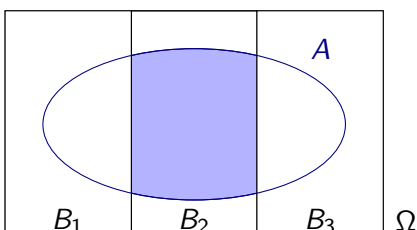
Mit $\hat{p} = \bar{x} = \frac{1}{n} \sum_i^n x_i$ ist schließlich die Varianz des Schätzers asymptotisch

$$Var(\hat{p}) = \frac{1}{\frac{n}{\bar{x}} + \frac{n}{1-\bar{x}}} = \frac{\bar{x} \cdot (1-\bar{x})}{n}$$

Bayessche Schätzverfahren

Rückblick: Die Bayessche Formel

- Bilden die Ereignisse $B_1, \ldots, B_m$ eine Ereignisdisjunktion, so gilt für alle Ereignisse A und $j = 1, \ldots, m$



$$
\begin{aligned}
P(B_j \mid A) &= \frac{P(A \mid B_j) \cdot P(B_j)}{P(A)} \\
&= \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{j=1}^{m} P(A \cap B_j)} \\
&= \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{j=1}^{m} P(A \mid B_j) \cdot P(B_j)}
\end{aligned}
$$

Beispielhafte Anwendung der Bayesschen Formel

Beispiel: Aufnahmeabteilung eines psychiatrischen Krankenhauses

Die Ereignisse B_1, B_2, B_3 sollen verschiedene Diagnosen kennzeichnen, z. B.

B_1	neurotisch
B_2	psychotisch
B_3	psychisch gesund

Die *A-priori-Wahrscheinlichkeiten* $P(B_1), P(B_2), P(B_3)$ kennzeichnen die (z. B. aus langjährigen Erfahrungen resultierende) Verteilung über die Diagnosen, noch vor einer Kenntnis der individuellen Symptomatik

Beispielhafte Anwendung der Bayesschen Formel

Das Ereignis A repräsentiere ein empirisches Datum, z. B. die Beobachtung eines Symptoms (wie etwa "hört Stimmen")

Die bedingten Wahrscheinlichkeiten $P(A \mid B_j)$, $j = 1, 2, 3$, kennzeichnen dann die Auftretenswahrscheinlichkeit des Symptoms A, gegeben die Diagnose B_j trifft zu

Die Bayessche Formel erlaubt dann die Berechnung der *A-posteriori-Wahrscheinlichkeit* $P(B_j \mid A)$ der Diagnose B_j gegeben das beobachtete Symptom A

Ist das Symptom informativ, dann werden sich die A-posteriori-Wahrscheinlichkeiten von den A-priori-Wahrscheinlichkeiten unterscheiden

Parameter als Zufallsvariable

Die im Beispiel deutlich gewordene Vorgehensweise wird nun auf die Parameterschätzung angewendet

Dabei wird von folgendem Grundverständnis ausgegangen

- Da der Wert eines Parameters zunächst nicht definitiv fest liegt, wird er als Zufallsvariable Θ betrachtet
- Ein Parameterwert ϑ wird damit als Realisierung der Zufallsvariablen Θ angesehen
- Die Unsicherheit über den Wert ϑ des Parameters wird durch eine Verteilung repräsentiert, die Wahrscheinlichkeitsverteilung der Zufallsvariablen Θ

A-priori-Verteilung

Bevor nun Daten aus einer Stichprobenziehung vorliegen, wird die Unsicherheit bezüglich des Parameterwertes durch die *A-priori-Verteilung* (prior distribution) beschrieben

- Diese A-priori-Verteilung kann Vorannahmen, Vorwissen oder Überzeugungen hinsichtlich des Parameterwertes repräsentieren oder jeden möglichen Parameterwert mit gleicher Wahrscheinlichkeit zulassen (nicht informative A-priori-Verteilung; uninformative prior)
- Es wird auch davon gesprochen, dass die Bayesschen Schätzverfahren auf einem *subjektiven Wahrscheinlichkeitsbegriff* basieren

  Dieser subjektive Wahrscheinlichkeitsbegriff wird einem *frequentistischen* gegenüber gestellt, der auf einer Interpretation von Wahrscheinlichkeiten als empirischen Häufigkeiten aufbaut

A-posteriori-Verteilung

Sobald nun Daten vorliegen, wird mit Hilfe einer geeigneten Version der Bayesschen Formel die *A-posteriori-Verteilung* (posterior distribution) berechnet

- Die A-posteriori-Verteilung der Zufallsvariablen Θ wird durch die Berücksichtigung der Daten mehr oder weniger deutlich von der A-priori-Verteilung abweichen (Bayesian learning)

Die A-posteriori-Verteilung bildet dann den Ausgangspunkt für die Bestimmung des (Punkt-)Schätzers für den unbekannten Parameter, z. B. durch deren

- Erwartungswert (im Mittel erwarteter Wert)
- Modus ("wahrscheinlichster Wert")

Notation

$f(x_1, \ldots, x_n \mid \vartheta)$ bezeichnet die Wahrscheinlichkeitsdichte (bzw. Wahrscheinlichkeitsfunktion) für die Daten (also für $X_1 = x_1, \ldots, X_n = x_n$), gegeben es gilt $\Theta = \vartheta$

$g(x_1, \ldots, x_n)$ bezeichnet die Randdichte (bzw. Randverteilung) der Daten

$h(\vartheta)$ bezeichnet die A-priori-Dichte (bzw. A-priori-Verteilung) von Θ (also die Randdichte von Θ)

$f(\vartheta \mid x_1, \ldots, x_n)$ bezeichnet die A-posteriori-Wahrscheinlichkeitsdichte (bzw. Wahrscheinlichkeitsfunktion) von Θ, gegeben die beobachteten Daten $x_1, \ldots, x_n$

Bayessche Inferenz

Die Formel von Bayes lautet nun allgemein

$$f(\vartheta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \vartheta) \cdot h(\vartheta)}{g(x_1, \ldots, x_n)}$$

Der Nenner hängt nur von den Daten, aber nicht vom Parameter ab!

Darüber hinaus ist

$$f(x_1, \ldots, x_n \mid \vartheta) = f(x_1 \mid \vartheta) \cdot \ldots \cdot f(x_n \mid \vartheta) = L(\vartheta)$$

die Likelihoodfunktion

Randdichte der Daten

Da die Zufallsvariable Θ im Allgemeinen stetig ist (z. B. wenn der Parameter Werte in einem reellen Intervall annehmen kann), erhalten wir die Randdichte $g(x_1, \ldots, x_n)$ durch Integration über alle möglichen Werte ϑ

$$g(x_1, \ldots, x_n) = \int f(x_1, \ldots, x_n \mid \vartheta) \cdot h(\vartheta) d\vartheta$$

Dieses Vorgehen wird auch so beschrieben

- Die Abhängigkeit vom unbekannten Parameterwert ϑ wird "herausintegriert"

Die Charakterisierungen der Dichten $f(x_1, \ldots, x_n \mid \vartheta)$ und $g(x_1, \ldots, x_n)$ lassen sich in die Formel von Bayes einsetzen

Bayessche Inferenz

Insgesamt erhält man also folgende alternative Darstellungen der *Bayesschen Inferenz*

$$\begin{aligned} f(\vartheta \mid x_1, \ldots, x_n) &= \frac{f(x_1, \ldots, x_n \mid \vartheta) \cdot h(\vartheta)}{\int f(x_1, \ldots, x_n \mid \vartheta) \cdot h(\vartheta) d\vartheta} \\ &= \frac{f(x_1 \mid \vartheta) \cdot \ldots \cdot f(x_n \mid \vartheta) \cdot h(\vartheta)}{\int f(x_1 \mid \vartheta) \cdot \ldots \cdot f(x_n \mid \vartheta) \cdot h(\vartheta) d\vartheta} \\ &= \frac{L(\vartheta) \cdot h(\vartheta)}{\int L(\vartheta) \cdot h(\vartheta) d\vartheta} \end{aligned}$$

EAP-Schätzer

Wählt man als Punktschätzer für den unbekannten Parameter den Erwartungswert der A-posteriori-Verteilung, dann bildet man

$$\hat{\vartheta}_p = E(\vartheta \mid x_1, \ldots, x_n) = \int \vartheta \cdot f(\vartheta \mid x_1, \ldots, x_n) d\vartheta$$

Zur expliziten Berechnung der A-posteriori-Verteilung muss das Integral $\int L(\vartheta) \cdot h(\vartheta) d\vartheta$ berechnet werden

Diese Berechnung kann unter Umständen kompliziert werden und ist möglicherweise analytisch nicht mehr durchführbar

In diesem Fall werden numerische Integrationsverfahren eingesetzt oder Simulationstechniken benutzt

MAP-Schätzer

Diese Problematik umgeht man, wenn man als Punktschätzer für den unbekannten Parameter den Modus der A-posteriori-Verteilung verwendet, den *Maximum-a-posteriori-Schätzer* (MAP)

Da der Nenner ja nicht mehr von ϑ abhängt, kann der Modus durch Maximieren des Zählerausdrucks $L(\vartheta) \cdot h(\vartheta)$ (bzw. des Logarithmus $\log L(\vartheta) + \log h(\vartheta)$) erfolgen

Man wählt also den Parameterwert $\hat{\vartheta}_{MAP}$ so, dass gilt

$$L(\hat{\vartheta}_{MAP}) \cdot h(\hat{\vartheta}_{MAP}) = \max_{\vartheta} L(\vartheta) \cdot h(\vartheta)$$

Damit wird auch die Beziehung zur Maximum-Likelihood-Schätzung deutlich

- Anstatt der Likelihood $L(\vartheta)$ wird hier die mit der A-priori-Verteilung $h(\vartheta)$ gewichtete Funktion $L(\vartheta) \cdot h(\vartheta)$ maximiert

Beziehung zur Maximum-Likelihood-Schätzung

Ist die A-priori-Dichte $h(\vartheta)$ sehr flach, so drückt dies wenig Vorwissen bezüglich des Wertes ϑ aus

Im Extremfall ist $h(\vartheta)$ beispielsweise eine Gleichverteilung über einem Intervall $[-a, a]$ mit einer sehr großen reellen Zahl a

Dann sind der ML-Schätzer $\hat{\vartheta}$ und der MAP-Schätzer $\hat{\vartheta}_{MAP}$ kaum unterscheidbar und stimmen für den Grenzfall $a \to \infty$ sogar völlig überein

Beispiel

Wir betrachten die Präsentation eines Reizpaares (x, s) im Kontext der Konstanzmethode

Zur Bestimmung der Wahrscheinlichkeit p, mit der eine Versuchsperson den Vergleichsreiz x über den Standardreiz s bevorzugt, erheben wir über eine Stichprobe binärer Zufallsvariablen $X_1, \ldots, X_n$ die Daten $x_1, \ldots, x_n$

Die Likelihoodfunktion ist dann durch die Bernoulli-Verteilung gegeben

$$L(p) = p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n-\sum_{i=1}^{n} x_i}$$

Beispiel

Wir nehmen an, dass die Versuchsperson in 20 von 30 Darbietungen den Vergleichsreiz als intensiver als den Standardreiz beurteilt

Da der ML-Schätzer die relative Häufigkeit ist, erhält man dann den Schätzwert

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{20}{30} = \frac{2}{3}$$

Der ML-Schätzer nutzt keinerlei Vorinformation, die z. B. darin bestehen kann, dass bekannt ist, dass der Vergleichsreiz physikalisch intensiver (oder nicht intensiver) als der Standardreiz ist

Beispiel

Das Vorwissen über den Wert des Parameters p beschreiben wir durch die Dichtefunktion

$$h(p) = \frac{p^{c-1} \cdot (1-p)^{c-1}}{B(c)}$$

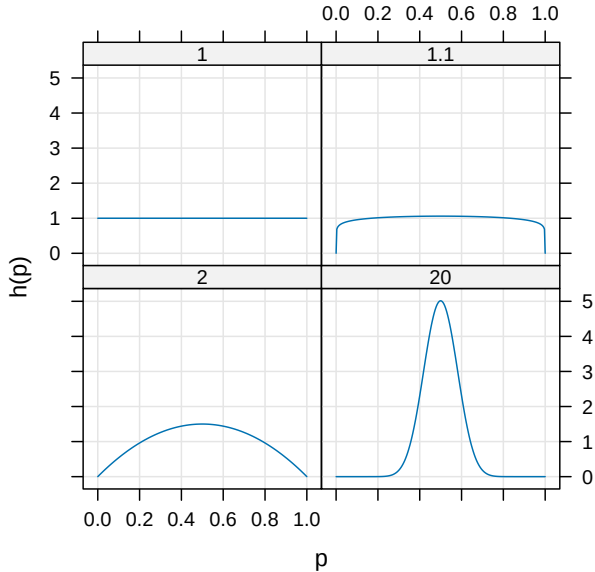
mit einem "Hyperparameter" $c \geq 1$ und einer geeignet gewählten Konstante $B(c)$, die nicht von p abhängt (Spezialfall einer Beta-Verteilung)

Mit dem Hyperparameter c steuern wir die Spezifität der Vorannahmen über den Parameter p

- Für $c = 1$ gilt $h(p) = 1/B(1)$, sodass in diesem Fall alle Parameterwerte gleich wahrscheinlich sind (uninformative prior), während sich mit steigendem c-Wert die Dichtefunktion zunehmend um den Wert 0.5 konzentriert

Beispiel

A-priori-Verteilungen $h(p)$
für $c \in \{1, 1.1, 2, 20\}$



Beispiel

Die Konzentration der Wahrscheinlichkeitsmasse um 0.5 kann z. B. durch das Vorwissen motiviert sein, dass der Vergleichsreiz hinsichtlich der physikalischen Reizintensität relativ nahe beim Standardreiz liegt

Zur Bestimmung des MAP-Schätzers $\hat{p}_{MAP}$ betrachten wir dann die Funktion

$$\begin{aligned} L(p) \cdot h(p) &= p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n-\sum_{i=1}^{n} x_i} \cdot \frac{p^{c-1} \cdot (1-p)^{c-1}}{B(c)} \\ &= \frac{p^{\sum_{i=1}^{n} x_i + c - 1} \cdot (1-p)^{n-\sum_{i=1}^{n} x_i + c - 1}}{B(c)} \end{aligned}$$

Beispiel

Für die Log-Likelihood gilt

$$\log L(p) + \log h(p) = \log p \left(\sum_{i=1}^{n} x_i + c - 1 \right) + \log(1-p) \left(n - \sum_{i=1}^{n} x_i + c - 1 \right) - \log B(c)$$

Ableiten nach p und Nullsetzen liefert dann den MAP-Schätzer

$$\hat{p}_{MAP} = \frac{\sum_{i=1}^{n} x_i + c - 1}{n + 2\,c - 2}$$

Beispiel

Der Hyperparameter c steuert in dieser Schätzgleichung die Gewichtung des Einflusses der Daten bzw. des Vorwissens

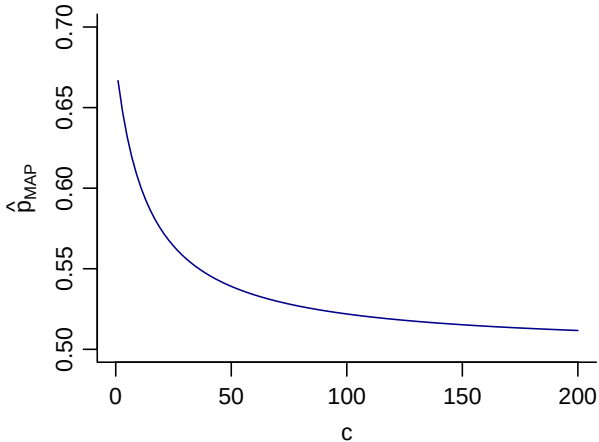
- Für $c = 1$ ist der MAP-Schätzer identisch mit dem ML-Schätzer, also der relativen Häufigkeit, was mit dem Fehlen jeglichen Vorwissens korrespondiert ("flache" A-priori-Verteilung)
- Für steigende c-Werte wird der Quotient zunehmend durch das Verhältnis von c im Zähler und $2\,c$ im Nenner dominiert und erreicht im Grenzfall $c \to \infty$ weitgehend unabhängig von den Daten den Wert 0.5

Je spezifischer also die Vorannahmen sind, desto mehr Einfluss gewinnen sie auf den Wert des MAP-Schätzers

Andererseits je informativer die Daten sind, desto weniger Einfluss haben die Vorannahmen

Beispiel

Beurteilt die Versuchsperson den Vergleichsreiz bei 30 Darbietungen insgesamt 20-mal als intensiver als den Standardreiz, dann ergeben sich in Abhängigkeit vom Hyperparameter c folgende Werte für den MAP-Schätzer



Verfahren zur Parameterschätzung

Zielsetzung

- Gewinnung von Informationen über den Wert der Parameter eines unterliegenden statistischen Modells

Methoden

- Vorgabe von Prinzipien zur Konstruktion von Schätzfunktionen, mit denen die Übereinstimmung zwischen den beobachteten Daten und der Vorhersage des statistischen Modells optimiert wird

Die verschiedenen Verfahren unterscheiden sich bezüglich

- des verwendeten Optimalitätskriteriums
- der Güte der resultierenden Schätzfunktionen

Zusammenfassung

Häufigste Verfahren in der psychologischen Forschung

- Maximum-Likelihood-Schätzung
- Bayes-Schätzung

Weitere häufige Verfahren

- Minimum-Chi-Quadrat-Schätzung
  Die hier minimierte Abweichung zwischen beobachteten Daten und der Vorhersage des Modells (beobachtete versus erwartete Häufigkeiten) wird erst später eingeführt

Moment-Methode

Parameter werden durch Momente der unterliegende Verteilung ausgedrückt

Theoretische Momente (z. B. Erwartungswert) werden durch korrespondierende Stichprobenmomente (z. B. arithmetisches Mittel) ersetzt, was zu Schätzfunktionen führt

Einfach in der Anwendung

Eigenschaften der resultierenden Schätzfunktion häufig unklar oder suboptimal

Methode der kleinsten Quadrate

Minimiert die Abweichungsquadrate zwischen

- den beobachteten Daten und
- den durch das unterliegende statistische Modell (z. B. lineare Regression) vorhergesagten Werten

Anwendung vorwiegend im Kontext der linearen und nicht-linearen Regressionsanalyse

- In diesem Zusammenhang liefert das Verfahren optimale Schätzfunktionen
- Diese werden als “best linear unbiased estimates” (BLUE) bezeichnet

Maximum-Likelihood-Schätzung

Maximiert die “Wahrscheinlichkeit” der Daten gegeben die Parameter des unterliegenden statistischen Modells

ML-Schätzer haben viele wünschenswerte Eigenschaften

- asymptotisch erwartungstreu
- asymptotisch effizient
- asymptotisch normalverteilt
- wenn es überhaupt eine suffiziente Schätzfunktion für den Parameter gibt, dann liefert ML-Schätzung eine solche

Verfahren wird häufig eingesetzt (z. B. in der Testtheorie)

Bayessche Schätzung

Kennzeichnet die Unsicherheit über den Wert eines Parameters durch entsprechende Verteilungen

- Die A-priori-Verteilung ermöglicht die Einbeziehung von Vorwissen bzw. subjektiven Einschätzungen in die Parameterschätzung
- Die A-posteriori-Verteilung kennzeichnet die Unsicherheit nach Auswertung der in den Daten verfügbaren Information und bildet die Grundlage für die Parameterschätzung

MAP-Schätzung (Maximum-a-posteriori) ist im Grenzfall fehlenden Vorwissens äquivalent zur ML-Schätzung

Verfahren wird in der psychologischen Forschung zunehmend eingesetzt (z. B. Testtheorie, Wahrnehmungspsychologie), insbesondere auch im Rahmen einer rekursiven Anwendung (adaptive Verfahren)