Einführung in die Bayes Statistik- Slides Begleitung zum Online Kurs

Einführung in die Wahrscheinlichkeit und Grundlagen

Martje Rave

Sommersemester 2025

ACHTUNG!

Diese Folien sind **NUR** Kurs begleitend und können noch viele Fehler beinhalten. Für die Klausurvorbereitung benutzt bitte den Online-Kurs und die Übungsblätter.
Bitte gebt mir Bescheid, wenn ihr Fehler findet.

Bayesianische Inferenz

- Wir beobachten n Daten x_i, die aus einem Zufallsprozess entstanden sind.
- Die x_i sind Realisierungen einer Zufallsvariable X_i.
- X_i hat eine Verteilung mit (Daten-)Dichte (Likelihood) $f(x)$.
- Wir kennen die Datendichte aber nur bis auf Parameter θ: $f(x|\theta)$
- θ ist unbekannt und wird durch eine Dichte beschrieben.
- Vor der Beobachtung (*a priori*): Priori-Dichte $p(\theta)$
- Nach der Beobachtung: Posteriori-Dichte $p(\theta|x)$

Bayesianische Inferenz (Forts.)

- Bisher: x und θ eindimensional
- Im Allgemeinen: auch mehrdimensional möglich

Aufgaben in der bayesianischen Inferenz

- Festlegung des statistischen Modells $f(x|\theta)$ (Likelihood)
- Festlegung der Priori-Dichte $p(\theta)$
- Berechnung der Posteriori $p(\theta|x)$

Bayes-Prinzip

Bayes-Formel

$$f(\theta|x) = \frac{f(x|\theta) \cdot f(\theta)}{\int f(x|\tilde{\theta}) f(\tilde{\theta}) d\tilde{\theta}}$$

Bayes-Prinzip

Alle Schlüsse werden **nur** aus der Posteriori-Verteilung gezogen.

Interpretation der Posteriori

- Posteriori enthält alle Information über θ nach der Beobachtung
- Je mehr Information in den Daten, desto kleiner die Varianz der Posteriori
- Die Dichte konzentriert sich stärker um wahrscheinliche Werte

Was machen wir mit der Posteriori?

- Möglichst vollständige Darstellung der Posteriori
- Bei mehrdimensionalem θ: Betrachtung marginaler Verteilungen $\theta_i|x$

Schlüsse aus der Posteriori

- **Punktschätzer**:
  - Posterior-Erwartungswert
  - Maximum-a-Posteriori (MAP)
  - Posterior-Median
- **Intervallschätzer**
- **Tests**
- **Modellvergleich**
- **Prädiktion**

Konjugierte Prioris

Wir hatten im Beispiel mit den Billardkugeln festgestellt, dass die Kombination von Binomialverteilung der Daten und Gleichverteilung als Priori gut zusammen passt: Wir erhalten eine bekannte Verteilung als Posteriori.

Allgemein **definieren** wir:

Eine Familie $\mathcal{F}$ von Verteilungen auf Θ heißt **konjugiert**, zu einer Dichte $f(x|\theta)$, wenn für jede Priori $p(\theta)$ auf $\mathcal{F}$ die Posteriori $p(\theta|x)$ ebenfalls zu F gehört

Beispiel nach Bayes

Im Beispiel der Billardkugeln hatten wir die Gleichverteilung als Spezialfall der Betaverteilung als Priori und die Betaverteilung als Posteriori.

- Anzahl der Kugeln rechts von der weißen Kugel: $X \sim B(n, \pi)$
- Priori-Annahme für π: $\pi \sim Beta(a, b)$ mit $a = b = 1$

Wir nennen dieses Modell **Beta-Binomial-Modell** (Beta-Priori und Binomial-Datenmodell). Im Folgenden benutzten wir ganz allgemein die Parameter a und b in der Beta-Priori und werden dann auch andere Werte für die *Priori-Parameter* zulassen.

Beispiel nach Bayes

Man erkennt die Konjugiertheit am ähnlichen Aufbau von Datendichte (eine Funktion in x) und Priori, bezogen auf den unbekannten Parameter π (wir betrachten hier nur jeweils den Kern der Dichte, lassen also Konstanten weg):

$$f(x|\pi) \propto \pi^x(1-\pi)^{n-x}.$$

$$p(\pi) \propto \pi^{a-1}(1-\pi)^{b-1}$$

Zusammen also

$$p(\pi|x) \propto f(x|\pi)p(\pi) \propto \pi^{x+a-1}(1-\pi)^{n-x+b-1}$$

Posteriori-Parameter

- Die Posteriori-Verteilung von $\pi|x$ ist also eine Beta$(\tilde{a}, \tilde{b})$-Verteilung mit $\tilde{a} = x + a$ und $\tilde{b} = n - x + b$. Die Parameter der Posterioriverteilung, also die *Posteriori-Parameter* setzten sich jeweils aus Informationen der Priori und der Datendichte zusammen.
- Allgemein fasst die Posteriori Information aus der Priori und der Datendichte zusammen.
- Hier ist $\tilde{a}$ die Summe des *Priori-Parameters* a und der Anzahl an Erfolgen x. Entsprechend ist $\tilde{b}$ die Summe des *Priori-Parameters* b und der Anzahl an Misserfolgen $n - x$.

Rückschluss auf die Priori-Parameter

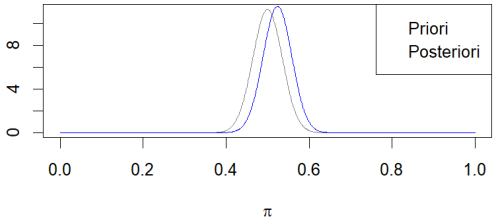
- Das heißt also, wenn wir a um eins erhöhen, ergibt sich für die Posteriori das selbe Ergebnis, wie wenn man die Anzahl der Erfolge um eins erhöht.
- a kann also in gewisser Weise also die *Priori-Anzahl an Erfolgen* interpretiert werden, entsprechend ist b die *Priori-Anzahl an Misserfolgen*.

Informative und subjektive Priori

Wir können also in die Priori-Verteilung "Information reinstecken". Und zwar theoretisch beliebig viel!

Zu viel Priori-Information

Nehmen wir im Billard-Beispiel als Priori $\pi \sim Beta(100, 100)$ und rollen dann zehn rote Kugeln, die alle links von der weißen Kugel zum liegen kommen. Die Posteriori ist dann $\pi|x \sim Beta(110, 100)$ und die Posteriori-Dichte sieht so aus:



Kritik an der Bayes-Inferenz

- Die Möglichkeit, mit der Priori die Posteriori – und damit das Ergebnis – weitgehend festzulegen, ist traditionell ein großer Kritikpunkt an der Bayes-Inferenz. $\rightarrow$ Schwierig.
- Als Ausweg daraus kann man **nicht-informative Prioris** verwenden. Dafür werden wir uns den **Begriff der Information** noch genauer ansehen müssen.
- Ein weiterer Anwendungsbereich der Bayes-Statistik liegt aber genau in der Nutzung von Vorwissen. Dieses kann zum Beispiel aus vorherigen Beobachtungen stammen – wir sprechen von **sequentiellem Lernen**, siehe dazu das Frosch-Beispiel – oder aus anderen Quellen, z.B. **Expertenwissen**.
- Ein dritter Ansatzpunkt ist, die Priori als Teil der Modellierung zu verwenden. Zum Beispiel in dem man Parameter absichtlich *Richtung Null drückt* oder Abhängigkeiten zwischen Parametern berücksichtigt. Insbesondere in hochdimensionalen, eventuell überparamtrisierten Modellen ist dies hilfreich. Dazu später mehr. (e.g. Lasso/Ridge Regression)

Nichtinformative Priori

Im Beispiel der Billardkugel hatten wir die Priori-Parameter a und b als *Priori-Erfolge* bzw. *Priori-Misserfolge* interpretiert. Intuitiv heißt keine Information, dass $a = 0$ und $b = 0$ ist.

Priori ohne Vorinformation

Setzen wir $a = 0$ und $b = 0$ in den Kern der Beta-Verteilung ein, erhalten wir die sogenannte **Haldane-Priori**:

$$p(\pi) \propto \pi^{-1}(1-\pi)^{-1}$$

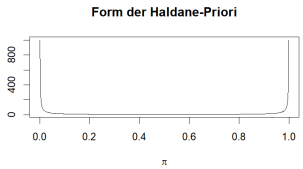


Figure: Haldane Prior

Allerdings: Das Integral $\int_0^1 \pi^{-1}(1-\pi)^{-1} = \infty$, existiert nicht. Da aber für jede Dichte gelten muss $\int p(\pi)d\pi = 1$, ist $p(\pi)$ hier keine Dichte ist!
Die *Haldane-Priori* kann man herleiten als Grenzfall einer Beta(a, b)-Verteilung mit $a \to 0$ und $b \to 0$.

Uneigentliche Verteilungen

Das die Haldane-Priori keine Dichte hat, ist aber (erstmal) kein Problem! Wir verwenden diese **Uneigentliche Verteilung** trotzdem.
Allgemein definieren wir eine **uneigentliche** oder **impropere** Verteilung mit Dichte $f(\theta)$ wie folgt:

- $f(\theta) \geq 0$ für alle θ (wie bei jeder Dichte)
- $\int f(\theta) d\theta = \infty$ ("eigentlich" müsste das Integral gleich 1 sein)

Aber warum ist das kein Problem?

Posteriori bei uneigentlicher Verteilung

Im Billard-Beispiel haben wir $\pi \sim$ "Beta$(0,0)$" - die Verteilung setzen wir hier in Anführungszeichen, denn die Priori ist *eigentlich* keine Betaverteilung. Trotzdem entspricht sie von der Form her der konjugierten Priori. Es gilt also für die Posteriori in diesem Fall:

$$
\begin{aligned}
p(\pi|x) &\propto p(\pi) \cdot f(x|\pi) && (1)\\
&\propto \pi^{-1}(1-\pi)^{-1} \cdot \pi^{x}(1-\pi)^{n-x} && (2)\\
&\propto \pi^{x-1}(1-\pi)^{n-x-1} && (3)
\end{aligned}
$$

Dies entspricht der Dichte einer Beta$(x, n-x)$-Verteilung, *wenn* $x > 0$ *und* $n - x > 0$, also wenn wir mindestens einen Erfolg und mindestens einen Misserfolg beobachtet haben.

Posteriori bei uneigentlicher Verteilung

- Aus einer *uneigentliche Posterioriverteilung* können wir keine Schlüsse ziehen: weder können wir eine Posterioriwahrscheinlichkeit berechnen noch einen Posteriori-Erwartungswert
- Die *uneigentliche Priori* führt aber im Regelfall zu einer *eigentlichen* oder *properen* Posterioriverteilung, aus der wir Schlüsse ziehen können.
- Nur in Ausnahmefällen kann eine uneigentliche Posterioriverteilung resultieren – dies muss man im Einzelfall überprüfen.

Laplace-Priori

- Bayes hatte in seinem Beispiel die Gleichverteilung benutzt. Laplace formulierte einige Jahre später das *Prinzip vom unzureichenden Grund* (Indifferenzprinzip):
- Wenn keine Gründe dafür bekannt sind, um eines von verschiedenen möglichen Ereignissen zu begünstigen, dann sind die Ereignisse als gleich wahrscheinlich anzusehen.
- Bekannt ist die Laplace-Wahrscheinlichkeit für Ergebnisse $\omega_i \in \Omega$:

$$P(\omega) = \frac{1}{|\Omega|}$$

mit $|\Omega|$ die Anzahl der Ergebnisse. Das entspricht einer *diskreten Gleichverteilung*. Analog spricht man auch bei der *stetigen Gleichverteilung* von der *Laplace-Verteilung*.

Laplace-Priori

Die Laplace-Priori ist also ganz allgemein

$$p(\theta) \propto \text{const.}$$

Sprich: die Dichte von θ ist proportional zu einer Konstante ("const." steht hierbei für eine beliebige Konstante).

Nebenbemerkung

Die stetige Gleichverteilung existiert nur für beschränkte θ. Wie wir bereits gesehen haben, können wir aber durchaus auch Prioris benutzen, die keine eigentlichen Verteilungen sind. Später verwenden wir also auch $p(\theta) \propto$ const. als "Gleichverteilung auf den reelen Zahlen".

Zusammenhang zur Likelihood-Inferenz

Für die Posteriori gilt:

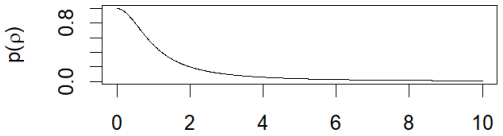
$$p(\theta|x) \propto p(\theta) \cdot f(x|\theta) \propto f(x|\theta).$$

Die Posteriori ist bei Laplace-Priori also proportional zur Datendichte bzw. Likelihood. Die Schlüsse werden dann nur aus der Likelihood gezogen. Die Bayes-Inferenz entspricht damit in diesem Fall weitgehend der Likelihood-Inferenz.

Informative Gleichverteilung

- Kommen wir zurück zum Billard-Beispiel. Wie wir oben gesehen haben, entspricht die Gleichverteilung der Vorinformation "ein Erfolg, ein Mißerfolg". Die Gleichverteilung ist in diesem Fall also *informativ*.
- Eine weitere Möglichkeit, sich das klar zu machen, ist, sich die Transformation des Parameters anzuschauen. Zum Beispiel können wir statt π den *odds* $\rho = \frac{\pi}{1-\pi}$ betrachten. Ist $\pi \sim U[0, 1]$, dann lässt sich mit dem Transformationssatz für Dichten zeigen:

$$p(\rho) = \frac{1}{(1+\rho)^2}$$



Schlüsse

Das heißt also

- Die Laplace-Priori bzw. Gleichverteilung führt dazu, dass die Posteriori der Likelihood entspricht.
- Die Laplace-Priori bzw. Gleichverteilung kann Information enthalten
- Benutzt man die Laplace-Priori auf einem transformierten Parameter, so kann dies zu einem anderen Posteriori-Ergebnis führen.

Für eine nicht-informative Priori ist also eine sinnvolle Forderung, das die Transformation des Parameters erneut zu einer nicht-informativen Priori führt.

Jeffreys' Priori

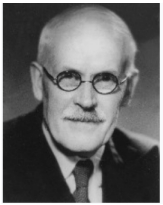


Figure: Harold Jeffrey's

- Die von *Harold Jeffreys* entwickelte und nach ihm benannte **Jeffreys' Priori** ist *invariant gegenüber Reparametrisierungen.*
- Wenn also $p^*(\theta)$ eine Jeffreys-Priori ist, und ϕ eine beliebige Funktion, dann ist $p^*(\phi(\theta))$ wieder eine Jeffreys-Priori für $\phi(\theta)$.
- Jeffreys' Priori ergibt sich aus der sogenannten **Fisher-Information**.
- Neben der Fisher-Information gibt es noch andere mathematische Informationsbegriffe wie die **Information nach Shannon**, die wir später kennen lernen.

Fisher-Information

- Die **Score-Funktion** ist definiert als Ableitung des Logarithmus der Likelihood bezüglich des Parameters:

$$S_\theta(x) := \frac{d}{d\theta} \ln f(x, \theta)$$

Setzt man die Score-Funktion gleich Null und löst nach θ auf, findet man das Maximum der Likelihood (also den Maximum-Likelihood-Schätzer).

- Die **Fisher-Information** (als Funktion in θ ist definiert als Varianz der Scorefunktion (die ja eine Funktion in x ist) in Abhängigkeit vom Parameter θ

$$I(\theta) := \text{Var}(S_\theta)$$

  - Die Fisher-Information kann damit als Krümmung der Log-Likelihood angesehen werden.
  - In der Nähe der Maximum-Likelihood-Schätzung zeigt eine niedrige Fisher-Information daher an, dass das Maximum relativ flach flach ist, d.h. es gibt viele nahe gelegene Werte mit einer ähnlichen Log-Likelihood.
  - Umgekehrt zeigt eine hohe Fisher-Information an, dass das Maximum herausticht ist.

Jeffreys' Priori

Jeffreys' Priori ist definiert als Wurzel aus der Fisherinformation:

$$p^*(\theta) \propto I^{1/2}(\theta)$$

Beispiel: Exponentialverteilung

- Sei $X \sim Exp(\lambda)$. Dann ist die Dichtefunktion $f(x) = \lambda \exp(-\lambda x)$
- Damit ist die Score-Funktion

$$S_\lambda(x) = \frac{d}{d\lambda} \ln f(x) = \frac{d}{d\lambda} (\ln(\lambda) - \lambda x) = \frac{1}{\lambda} - x$$

- Die Fisher-Information ergibt sich als

$$I(\lambda) = \text{Var}(S_\lambda) = \text{Var}\left(\frac{1}{\lambda} - x\right) = \text{Var}(x) = \frac{1}{\lambda^2}$$

und Jeffreys' Priori für den Parameter λ der Exponentialverteilung ist

$$p(\lambda) \propto \left(\frac{1}{\lambda^2}\right)^{1/2} = \lambda^{-1}$$

- Diese Priori lässt sich als Spezialfall der Gamma-Verteilung $Ga(a, b)$ mit $a \to 0$ und $b \to 0$ interpretieren.

Information

Jeffreys' Priori gilt als **nicht-informative Priori-Verteilung**

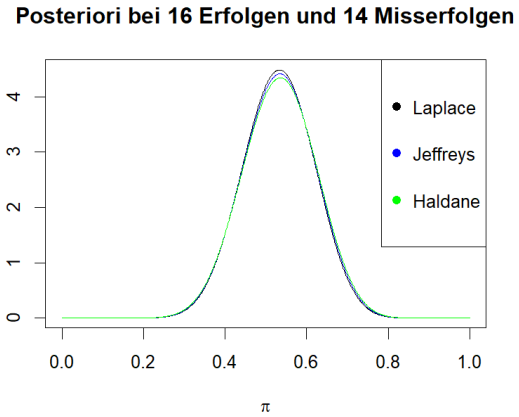
Nicht-informative Prioris

Wir haben in unserem Beispiel drei verschiedene Arten von *nicht-informativen Prioris* kennen gelernt:

- *Laplace-Priori*, also die Gleichverteilung
- *Haldane's Priori*, als Beispiel für eine uneigentliche Prioriverteilung
- *Jeffreys' Priori*, mit der Eigenschaft der Transformationsinvariabilität

Auswirkungen auf die Posteriori

Diese Prioris unterscheiden sich aber nur geringfügig. Schon bei wenigen Beobachtungen (in der Graphik $n = 30, x = 16$) ergeben sich kaum Unterschiede in der Posteriori



Wiederholung und Bemerkungen

- Die Gleichverteilung (Laplace-Priori) kann unter Umständen Information enthalten.
- Jeffreys' Priori entspricht nicht unbedingt der intuitiv nicht-informativen Priori (hier Haldane).
- Jeffreys' Priori kann uneigentlich sein, muss sie aber nicht sein.

Wiederholung und Bemerkungen

- Eine nichtinformative Priori kann auch die konjugierte sein, eventuell auch als Grenzfall der konjugierten Priori.
- Viele flache (nichtinformative) Prioris sind nicht proper. Dies ist unproblematisch, solange die Posteriori proper ist.
- Gerne benutzt man auch "relativ" *flache Prioris*, die proper sind (zum Beispiel statt der stetigen Gleichverteilung auf ganz $\mathbb{R}$), siehe Abschnitt Normalverteilung.

Jeffreys' Priori

- Es gibt verschiedene mathematische Definitionen von *Information*. Wir betrachten hier nur die *beobachtete* und die Fisher-Information (im nächsten Kapitel werden wir die Information nach *Shannon* kennen lernen).
- Sei wieder $f(x|\theta)$ die Datendichte der Zufallsvariable X gegeben dem (skalaren) Parameter θ. Diese Dichte bzw. Likelihood beschreibt uns den Zusammenhang der Daten mit dem Parameter, also wieviel Information über den Parameter in den Daten x vorliegt.

Fisher-Information

Als **Fisher-Information** definiert man den Erwartungswert des Quadrats der Ableitung der Log-Dichte:

$$I(\theta) = \mathrm{E}\left[\left(\frac{d}{d\theta}\log(f(X|\theta))\right)^2\right]$$

Alternativ kann man die **Fisher-Information** definieren als negativen Erwartungswert der zweiten Ableitung der Log-Dichte:

$$I(\theta) = \mathrm{E}\left[-\left(\frac{d^2}{d\theta^2}\log(f(X|\theta))\right)\right]$$

Hintergund

In der Likelihood-Inferenz bezeichnet man die Ableitung der Log-Likelihood als *Score-Funktion*. Setzt man die Score-Funktion gleich Null, kann man den *Maximum-Likelihood-Schätzer* herleiten.
Vor der Beobachtung ist die Scorefunktion aber eine Zufallsvariable

$$s(X) = \frac{d}{d\theta} \log(f(X|\theta))$$

Die *Fisher-Information* ist dann als Varianz der Score-Funktion definiert, was dem Erwartungswert des Quadrats der Score-Funktion entspricht (der Erwartungswert der Score-Funktion ist Null). Varianz ist hier ein Maß für Unsicherheit.
Steckt also viel Information in den Daten über den Parameter, so ist unsere Unsicherheit kleiner, damit die Varianz kleiner.

Fisher-Information

Über die Fisher-Information lässt sich zudem eine untere Schranke für die Varianz eines Parameterschätzers berechnen, die sogenannten "Cramér-Rao-Schranke":

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

für alle möglichen θ.
Die Fisher-Information ist identisch zur negativen zweite Ableitung der Log-Dichte, falls diese existiert, und wird oft auch so definiert.

Jeffreys' Priori

Aus der Fisher-Information lässt sich Jeffreys' Priori berechnen:

$$p^*(\theta) \propto I^{1/2}(\theta)$$

Jeffreys' Priori ist invariant gegenüber Reparametrisierungen. Wenn also $p^*(\theta)$ eine Jeffreys-Priori ist, dann ist $p^*(\phi(\theta))$ wieder eine Jeffreys-Priori für $\phi(\theta)$.

Beweis

Die Eigenschaft lässt sich (für bijektive Transformation $\phi(\theta)$) über den Transformationssatz für Dichten nachweisen. Sei $p(\theta)$ eine Dichte bezüglich θ und bezeichne $\theta(\phi)$ die Umkehrtransformation:

$$
\begin{aligned}
p^*(\phi) &= p^*(\theta) \left| \frac{d\theta(\phi)}{d\phi} \right| \\
&\propto \sqrt{E_X \left[\left(\frac{d \log(f(X|\theta(\phi)))}{d\theta(\phi)} \right)^2 \right]} \left| \frac{d\theta(\phi)}{d\phi} \right| \\
&= \sqrt{E_X \left[\left(\frac{d \log(f(X|\theta(\phi)))}{d\theta(\phi)} \frac{d\theta(\phi)}{d\phi} \right)^2 \right]} \\
&= \sqrt{E_X \left[\left(\frac{d \log(f(X|\theta(\phi)))}{d\phi} \right)^2 \right]} = \sqrt{I(\phi)}
\end{aligned}
$$

Binomialverteilung

Wir leiten Jeffreys' Priori für die Bernoulliverteilung her. Die Dichte ist

$$f(x|\pi) = \pi^x (1-\pi)^{1-x}$$

Die Log-Dichte ist

$$\log(f(x|\pi)) = x \log(\pi) + (1-x) \log(1-\pi)$$

Die Ableitung nach π ist

$$s(x) = \frac{d}{d\pi} \log(f(x|\pi)) = \frac{x}{\pi} - \frac{1-x}{1-\pi}$$

Vor der Beobachtung ersetzen wir x durch die Zufallsvariable X. Dann berechnen wir den Erwartungswert von $\left(s(X)^2\right)$:

$$\begin{aligned} I(\pi) &= \mathrm{E}\left[\left(s(X)^2\right)\right] \\ &= \pi \left(\frac{1}{\pi} - \frac{0}{1-\pi}\right)^2 + (1-\pi)\left(\frac{0}{\pi} - \frac{1}{1-\pi}\right)^2 \\ &= \frac{1}{\pi} + \frac{1}{1-\pi} = \frac{1}{\pi(1-\pi)} \end{aligned}$$

Jeffrey's Priori ist dann

$$p(\pi) \propto \sqrt{I(\pi)} = \pi^{-\frac{1}{2}} (1-\pi)^{-\frac{1}{2}} = \pi^{(\frac{1}{2}-1)} (1-\pi)^{(\frac{1}{2}-1)}$$

Dies entspricht der Dichte einer Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$-Verteilung.

Jeffrey Prior

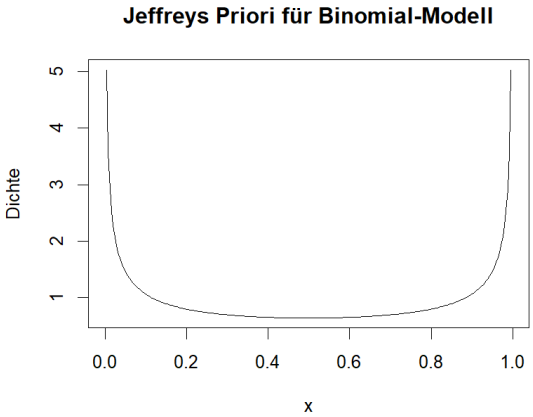


Figure: Jeffrey Prior Binomial Data

Subjektive Priori

- Wie zuvor besprochen, kann es unter Umständen sinnvoll sein, subjektive Priori-Informationen zu verwenden.
- Wir betrachten ein Beispiel nach Dupuis (1995).

Echsen-Beispiel

In einem biologischen Experiment werden Echsen markiert und später nochmals eingefangen.

- Das Einfangen ist ein Bernoulli-Experiment.
- Erfolg = markierte Echse wird wieder eingefangen.
- p_t sei die Einfangwahrscheinlichkeit nach t Zeiteinheiten.
- Biologen geben *a priori* Einschätzungen für p_t.

Einschätzungen der Biologen

Zeitpunkt	2	3	4	5
Mittelwert	0.3	0.4	0.5	0.2
95%-Intervall	[0.1, 0.5]	[0.2, 0.6]	[0.3, 0.7]	[0.05, 0.4]

Gewinnung Subjektiver Prioris

Subjektive Prioris basieren auf Expertenwissen, das nicht immer als Dichteform vorliegt.

- Für diskrete Parameter: Wahrscheinlichkeiten direkt angeben.
- Für stetige Parameter:
  - Diskretisierung in Intervalle
  - Parametrische Priori
  - Maximum Entropy Priori

Parametrische subjektive Priori

- Zuerst Verteilungsform vorgeben (z.B. konjugierte Beta-Verteilung)
- Dann Parameter mit Hilfe robuster Statistiken bestimmen (Median, Quantile, etc.)
- Konsistenz prüfen: Verteilung muss zu Einschätzung passen!

Echsenbeispiel – Zeitpunkt $t = 2$

- Mittelwert: 0.3
- 95%-Intervall: [0.1, 0.5]
- Erwartungswert der Beta-Verteilung: $\frac{a}{a+b} = 0.3 \Rightarrow b = \frac{7}{3}a$

Gesucht: a so, dass

$$P(0.1 < X < 0.5) = 0.95, \quad X \sim \text{Beta}(a, \tfrac{7}{3}a)$$

Bestimmung in R

```
fehler <- function(a) {
  wahrscheinlichkeit <- pbeta(0.5, a, 7*a/3) - pbeta(0.1, a, 7*a/3)
  return((wahrscheinlichkeit - 0.95)^2)
}
a <- optimize(fehler, c(0.01, 100))
print(a$minimum)
```

Ergebnis

- Funktion ‘fehler‘ misst quadratischen Abstand zu Zielwahrscheinlichkeit.
- ‘optimize‘ sucht optimalen a in Bereich $[0.01, 100]$.

Ergebnis:

$$p_2 \sim \text{Beta}(5.45, 12.72)$$

Weitere Zeitpunkte

Zeitpunkt	2	3	4	5	6
Mittelwert	0.3	0.4	0.5	0.2	0.2
95%-Intervall	[0.1, 0.5]	[0.2, 0.6]	[0.3, 0.7]	[0.05, 0.4]	[0.05, 0.4]

Abgeleitete Verteilungen

- $p_3 \sim \text{Beta}(8.58, 12.87)$
- $p_4 \sim \text{Beta}(11.26, 11.26)$
- $p_5 \sim \text{Beta}(3.50, 14.00)$
- $p_6 \sim \text{Beta}(3.50, 14.00)$

Maximum Entropy Priori

Ziel: Wähle eine Verteilung mit maximaler Entropie unter Berücksichtigung vorhandener Momente.

Shannon-Entropie

- Diskret: $H(\theta) = -\sum_\theta p(\theta) \log p(\theta)$
- Stetig: $H(\theta) \propto -\int p(\theta) \log p(\theta)\, d\theta$

Maximum-Entropie-Form

Sind Momente $E(g_k(X))$ bekannt, hat MEP folgende Form:

$$p_{\text{MEP}}(x) = c \cdot \exp\left(\sum_{k=1}^{K} \lambda_k g_k(x)\right)$$

Beispiel

Vorinformation:

- $E(\theta) = 1$
- $\text{Var}(\theta) = 4$

Aus dem Verschiebungssatz ergibt sich

$$Var(\theta) = E(\theta^2) + (E(\theta))^2 \tag{4}$$

$$\Rightarrow E(\theta^2) = Var(\theta) - (E(\theta))^2 = 4 - 1 = 3 \tag{5}$$

Hier sind also die Momente von $g_1(\theta) = \theta$ und $g_2(\theta) = \theta^2$ bekannt.

Dichteform

Bekannte Momente: $g_1(\theta) = \theta$, $g_2(\theta) = \theta^2$

$$p_{\text{MEP}} = c \cdot \exp(\lambda_1 \theta + \lambda_2 \theta^2)$$

Bestimme λ_1, λ_2, c so, dass:

- p ist eine gültige Dichte
- Momente stimmen

Resultat

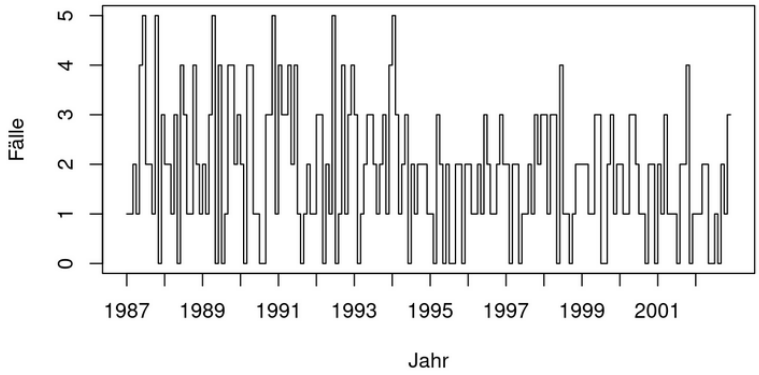
Für $\theta \in \mathbb{R}$ ergibt sich eine Normalverteilung:

$$p_{\text{MEP}} = \frac{1}{\sqrt{2\pi \cdot 4}} \exp\left(-\frac{1}{2 \cdot 4}(\theta - 1)^2\right)$$

Also: $\theta \sim \mathcal{N}(1, 4)$
Ist θ beschränkt, ergibt sich eine **trunkierte Normalverteilung**.
Für kompliziertere Fälle müssen die λ_k numerisch bestimmt werden.

Verkehrsunfälle in Linz

Die folgende Grafik zeigt die monatliche Anzahl schwerverletzter oder getöteter Kinder (6–10 Jahre) im Straßenverkehr in Linz von 1987 bis 2002.



Poissonverteilung

Für Zählvariablen wie diese eignet sich die **Poisson-Verteilung**:

$$X_t \sim \text{Po}(\lambda)$$

- λ ist Erwartungswert und Varianz von X_t
- Typisch für seltene Ereignisse
- Approximation der Binomialverteilung für kleine π

Datendichte

Die Wahrscheinlichkeitsdichte für einen Monat t:

$$f(x_t|\lambda) = \frac{\lambda^{x_t}}{x_t!}\exp(-\lambda)$$

Für unabhängige Daten $x = (x_1, \ldots, x_n)$ ergibt sich:

$$f(x|\lambda) = \prod_{t=1}^{n} \frac{\lambda^{x_t}}{x_t!}\exp(-\lambda) = \frac{\lambda^{\sum x_t}}{\prod x_t!}\exp(-n\lambda)$$

Fragestellungen

Was wollen wir aus den Daten lernen?

- Welchen Wert hat die Rate λ?
- Wie hoch ist die Unsicherheit der Schätzung?
- Wie viele Unfälle erwarten wir nächsten Monat?

Konjugierte Priori

Die **konjugierte Priori** zur Poisson-Verteilung ist die Gamma-Verteilung:

$$\lambda \sim \text{Ga}(a, b) \quad \text{mit} \quad p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}$$

- Dabei ist $\Gamma(a)$ die Gamma-Funktion mit Definition $\Gamma(a) = \int_0^\infty t^{a-1} \exp(t) dt$. Die Gamma-Funktion ist eine Verallgemeinerung der Fakultät, für natürlich Zahlen gilt: $\Gamma(a+1) = a!$.
- Erwartungswert: $\frac{a}{b}$, Varianz: $\frac{a}{b^2}$. Wir können die Priori-Parameter also zum einen so wählen, dass a/b der Wert ist, den wir für λ a priori annehmen, z.B. $a = b$, womit der *Priori-Erwartungswert* 1 wäre.
- Zum anderen können wir über die Priori-Varianz bestimmen, wie viel Information wir a priori geben. Große Varianz heißt wenig Information. Zum Beispiel wird oft $a = b = 0.001$ verwendet, so dass die Priori-Varianz gleich 1000 ist.

Konjugierte Posteriori

Aus Daten ($\bar{x} = \frac{1}{n}\sum x_t$) und Priori ergibt sich:

$$p(\lambda|x) \propto \lambda^{n\bar{x}+a-1} e^{-\lambda(b+n)}$$

$$\Rightarrow \lambda|x \sim \text{Ga}(a + n\bar{x}, b + n)$$

Jeffreys' Priori

Die **Jeffreys'-Priori** für die Poisson-Rate ist:

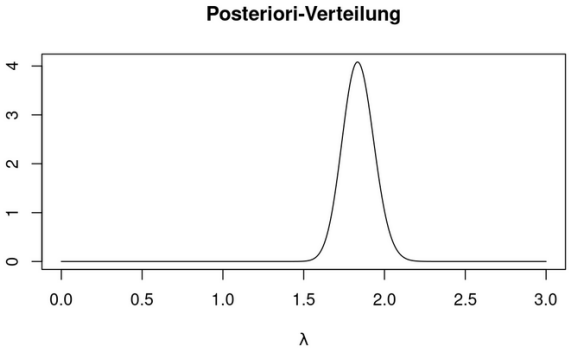
$$p(\lambda) \propto \lambda^{-1/2}$$

- Grenzfall der Gamma-Verteilung mit $a = 1/2$, $b \to 0$
- Uneigentliche, aber konjugierte Priori

Beispiel

Aus den Daten:

- $n = 192$, $\bar{x} \approx 1.8385$
- Priori: $a = b = 0.001$
- Posteriori: $\lambda | x \sim \text{Ga}(352.993,\ 192.001)$



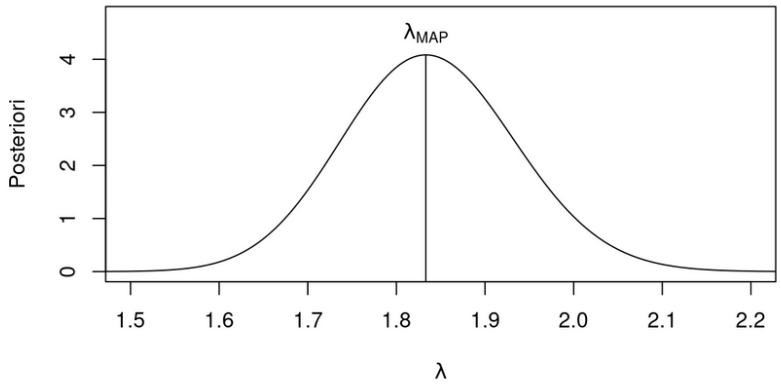
Posteriori-Erwartungswert

$$\hat{\lambda}_{\text{PE}} = \mathbb{E}[\lambda|x] = \frac{a + n\bar{x}}{b + n} = \frac{352.993}{192.001} \approx 1.8385$$

- Interpretation: Erwartungswert gegeben der Daten

Posteriori-Modus (MAP)

$$\hat{\lambda}_{\text{MAP}} = \frac{a - 1 + n\bar{x}}{b + n} = \frac{351.993}{192.001} \approx 1.8332$$



Posteriori-Median

$$\hat{\lambda}_{\text{med}} = \text{Median}(\lambda|x) \approx 1.8368$$

- Der mittlere Posteriori-Wert
- Robust gegenüber Ausreißern

Vergleich der Schätzer

- $\hat{\lambda}_{\text{PE}} = 1.8385$
- $\hat{\lambda}_{\text{MAP}} = 1.8332$
- $\hat{\lambda}_{\text{med}} = 1.8368$

Für symmetrische Verteilungen fallen diese fast zusammen.

Drei Schätzer

Nun haben wir drei verschiedene mögliche Punktschätzer:

- Posteriori-Erwartungswert
- Posteriori-Modus (MAP)
- Posteriori-Median

Welcher ist den nun der beste Schätzer?

Theoretisch

Welcher Schätzer der Beste ist, hängt davon ab, wie man "Bester" definiert. Das wiederum hängt davon ab, was man als "schlecht" definiert: Wie bewertet man die Abweichung zwischen geschätztem Wert und wahren Wert. In der Entscheidungstheorie benutzt man dazu eine Verlustfunktion. Es stellt sich heraus:

- Bei quadratischer Verlustfunktion ist der Posteriori-Erwartungswert der beste Schätzer
- Bei absoluter Verlustfunktion ist der Posteriori-Median der beste Schätzer
- Bei 0/1-Verlustfunktion ist der Posteriori-Modus der beste Schätzer

Praxisvergleich

Erwartungswert:

- Einfach für konjugierte Posteriori
- Schwierig bei nichtstandardisierten Verteilungen

Modus (MAP):

- Bei flacher Priori $\Rightarrow$ Maximum-Likelihood
- Sonst: penalisiertes ML

Median:

- Nur einfach, wenn Verteilungsfunktion explizit verfügbar
- Vorteilhaft bei Posteriori-Simulation

Weitere Hinweise

- Posteriori-Erwartungswert kann verzerrt sein
- MAP und PE nicht invariant unter monotonen Transformationen

Was ist ein Intervallschätzer?

Können wir nun einen Intervallschätzer für λ angeben? Das heißt, ein Intervall, in dem λ mit einer vorgegebenen Wahrscheinlichkeit α liegt.

Wir haben die Posteriori-Verteilung von λ gegeben x. Daraus können wir leicht für ein Intervall I die Wahrscheinlichkeit angeben:

$$P(\lambda \in I \mid x) = \alpha$$

Kredibilitätsintervall

Wir nennen ein solches Intervall **Kredibilitätsintervall** (auch: *Glaubwürdigkeitsintervall*).

Im Gegensatz zum klassischen *Konfidenzintervall*, bei dem α über viele Stichproben hinweg interpretiert wird, ist die Interpretation des Kredibilitätsintervalls viel intuitiver: es gibt die Wahrscheinlichkeit an, dass der wahre Parameter im Intervall liegt.

Highest Posterior Density Interval (HPD)

Es gibt unendlich viele Intervalle mit gleicher Wahrscheinlichkeit. Wir wählen daher:

Definition

Sei $I \subset \Theta$ mit $P(\theta \in I \mid x) = \alpha$ und $p(\theta) > p(\theta^*)$ für alle $\theta \in I$ und $\theta^* \notin I$.
Dann heißt I **HPD-Intervall** (Highest Posterior Density).

Zur Erinnerung: in unserem Beispiel war $\lambda \mid x \sim \text{Ga}(352.993, 192.001)$, das 95%-HPD-Intervall war:

$$I_{\text{HPD}} = [1.648,\ 2.032]$$

Eigenschaften des HPD-Intervalls

- Kürzestes Intervall bei gleicher Wahrscheinlichkeit
- Kann bei mehr-modalen Posterioris aus mehreren Intervallen bestehen
- Nicht invariant gegenüber Transformationen

Symmetrisches Kredibilitätsintervall

In der Praxis einfacher: symmetrisches Intervall $I = [u, o]$ mit

$$P(\theta < u) = P(\theta > o) = \frac{1 - \alpha}{2}$$

Im Beispiel: $I = [1.652, 2.036]$.

Visualisierung der Intervalle

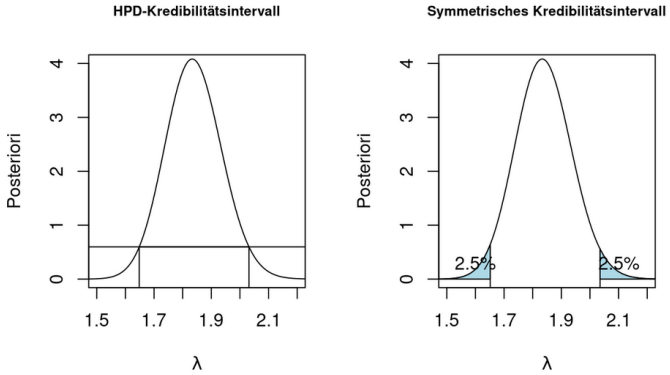


Figure: Highest Posterior Density

Posteriori-Unsicherheit

Statt eines Intervalls kann man auch die Streuung der Posteriori angeben, z.B. durch Varianz oder Standardabweichung:

$$\text{Var}(\lambda \mid x) = \frac{352.993}{192.001^2} \approx 0.009575$$

$$\text{sd}(\lambda \mid x) \approx \sqrt{0.009575} \approx 0.09785$$

Sensitivitätsanalyse

Verschiedene Prioris führen zu verschiedenen Schätzern. Die folgende Tabelle zeigt den Posteriori-Erwartungswert und das HPD-Intervall:

a / b	0.01	0.1	1	10
0.01	1.839 (1.648, 2.032)	1.839 (1.649, 2.032)	1.844 (1.653, 2.037)	1.890 (1.698, 2.086)
0.1	1.838 (1.648, 2.031)	1.838 (1.648, 2.031)	1.843 (1.652, 2.036)	1.890 (1.697, 2.085)
1	1.829 (1.640, 2.021)	1.829 (1.640, 2.022)	1.834 (1.645, 2.026)	1.881 (1.689, 2.076)
10	1.748 (1.567, 1.931)	1.748 (1.567, 1.932)	1.752 (1.571, 1.936)	1.797 (1.614, 1.983)

Sensitivitätsanalyse: Vergleich verschiedener Priori-Parameter zeigt, wie robust die Schätzung ist.

Entscheidungstheoretische Grundlage der Bayes-Schätzer

Im vorherigen Abschnitt haben wir drei verschiedene bayesianische Punktschätzer kennengelernt:

- Posteriori-Erwartungswert
- Posteriori-Modus
- Posteriori-Median

Doch welcher dieser Schätzer ist der beste"?

Das hängt davon ab, was man unter beste" versteht. Oder anders gefragt: Für welchen Schätzer sollte man sich entscheiden?

Die Entscheidungstheorie liefert darauf eine fundierte Antwort.

Das Sankt-Petersburg-Paradoxon

Hintergrund: Daniel Bernoulli stellte das folgende berühmte Gedankenexperiment vor:

Spielregeln

In einem Glücksspiel wird eine faire Münze so lange geworfen, bis zum ersten Mal Kopf“ fällt. Dies beendet das Spiel.
Der Gewinn richtet sich nach der Anzahl der Würfe insgesamt:

- 1. Wurf: 1 €
- 2. Wurf: 2 €
- 3. Wurf: 4 €
- ...
- Allgemein: 2^{k-1}€, wenn beim k-ten Wurf das erste Mal Kopf“ erscheint.

Frage

Welchen Geldbetrag sollte man maximal für die Teilnahme an diesem Spiel zahlen?

Antwortidee

$$P(X = k) = (1/2)^k$$

- Klassisch: Der *erwartete Gewinn* ist unendlich:

$$\sum_{k=1}^{\infty} \left(\frac{1}{2^k} \cdot 2^{k-1} \right) = \sum_{k=1}^{\infty} \frac{1}{2} = \infty$$

- Aber: Kaum jemand wäre bereit, mehr als z.B. 10€ für die Teilnahme zu zahlen.
- **Warum?** $\rightarrow$ Begrenzte Risikobereitschaft, abnehmender Grenznutzen.

Fazit: Die *Entscheidungstheorie* und *Nutzenfunktionen* helfen, rationale Entscheidungen zu modellieren.

Entscheidungen

- Das Ziel der statistischen Inferenz ist in der Regel, einen Parameter θ zu schätzen (Punktschätzer oder Intervallschätzer) oder eine Hypothese zu testen. Im weiteren Sinn gilt es eine Entscheidung zu treffen.
- In der **Entscheidungstheorie** wird die Inferenz direkt für die Entscheidung betrieben. Dazu definieren wir uns einen Entscheidungsraum D möglicher Entscheidungen. Für die Zufallsvariable X wollen wir für jedes mögliche Stichprobenergebnis x aus dem Stichprobenraum X eine Entscheidung $d(x)$ treffen.

Entscheidungsfunktion

Eine Entscheidungsfunktion ist eine Abbildung vom Stichprobenraum X in den Entscheidungsraum D

$$d \ : \ \chi \to D; \ x \to d(x)$$

Die Zufallsvariable X hänge von Parametern $\Theta \in \theta$ ab. Im Fall des Punktschätzers ist $D = \Theta$ (der Parameterraum); wir entscheiden wir uns für einen Wert als Punktschätzer. Beim Testen entscheidet man sich für eine Hypothese.

Verlustfunktion

Ziel: Wie bewerten wir mögliche Entscheidungen? Wann ist eine Entscheidung gut oder schlecht?

Verlustfunktion

Die Verlustfunktion (engl. *loss function*) ordnet jeder Entscheidung einen Verlust zu:

$$L : D \times \Theta \to \mathbb{R}, \quad (d, \theta) \mapsto L(d, \theta)$$

Die Verlustfunktion hängt sowohl von der Entscheidung d, als auch vom unbekannten Parameter θ ab.

Wichtigste Verlustfunktionen

Klassische Verlustfunktionen:

- $L_1(d, \theta) = |d - \theta|$ (absoluter Verlust)
- $L_2(d, \theta) = (d - \theta)^2$ (quadratischer Verlust)
- $L_p(d, \theta) = |d - \theta|^p$ (L_p-Verlust)
- $L_\varepsilon(d, \theta) = \begin{cases} 1 & \text{falls } |d - \theta| > \varepsilon \\ 0 & \text{falls } |d - \theta| \leq \varepsilon \end{cases}$ (0-1-Verlust)

Schätzen als Entscheidungsproblem

Sei $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ mit bekanntem σ^2 und unbekanntem μ.

Der Entscheidungsraum D ist der Raum der möglichen Schätzungen, d.h. $d = \hat{\mu} \in \mathbb{R}$.

Quadratischer Verlust:

$$L(d(x), \mu) = (d(x) - \mu)^2$$

$\Rightarrow$ optimaler Schätzer (Kleinstquadrat-Schätzer):

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

Absoluter Verlust:

$$L(d(x), \mu) = |d(x) - \mu|$$

$\Rightarrow$ robusterer Schätzer (Posteriori-Median)

Gewinnfunktion

Gelegentlich wird statt einer Verlustfunktion auch eine Gewinnfunktion (engl. *utility function*) verwendet.

Die Wahl von Verlust- oder Gewinnfunktion ist **subjektiv** und hängt von der spezifischen Entscheidungssituation ab.

Risiko

Nehmen wir an, eine Verlustfunktion L wurde gewählt:

- Vor der Beobachtung von X ist die Entscheidung $d(X)$ eine Zufallsvariable.
- Damit ist $L(d(X), \theta)$ eine Zufallsvariable.
- Der Verlust hängt zudem vom unbekannten Parameter θ ab.

Risiko

Der mittlere Verlust bei Entscheidung d und wahrem Parameter θ ist:

$$R(d, \theta) = \mathbb{E}_X \left[L(d(X), \theta) \right] = \int_X L(d(x), \theta) f(x|\theta) dx$$

Das Risiko bewertet also die Entscheidungsfunktion d, hängt jedoch vom wahren θ ab.

Zulässigkeit

Ziel: Entscheidung mit minimalem Risiko für alle möglichen θ finden.

In der Regel nicht möglich $\Rightarrow$ Wir beschränken uns auf zulässige Entscheidungen.

Definition (Zulässigkeit)

Die Entscheidungsfunktion d ist zulässig, wenn es kein d^* gibt, sodass

$$R(d^*, \theta) < R(d, \theta) \quad \text{für alle } \theta \in \Theta.$$

Zulässig heißt: Es gibt keine andere Entscheidung mit durchgängig geringerem Risiko.

Minimax-Entscheidung

Wenn das Risiko nicht für alle θ minimiert werden kann, minimieren wir das maximale Risiko.

Definition (Minimax)

Die Minimax-Entscheidung ist die Entscheidung d, die das Maximum der Risikofunktion minimiert:

$$d^*_{\text{minimax}} = \arg\min_d \left(\max_\theta R(d, \theta) \right)$$

Diese Strategie ist konservativ: Sie schützt gegen den schlimmstmöglichen Fall.

Minimax-Entscheidung

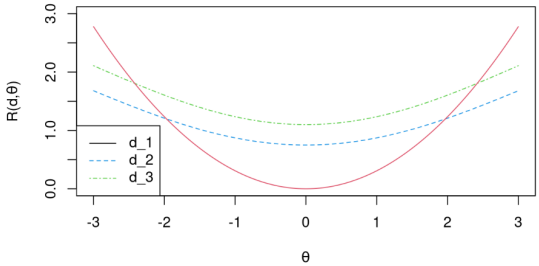


Figure: Minimax- Beispiel

Was aber, wenn wir es für wahrscheinlich halten, dass in diesem Beispiel θ nahe Null liegt? Dann würden wir vermutlich eher zu d_1 tendieren.

Bayes-Risiko

Eventuell hat das maximale Risiko eine sehr geringe Wahrscheinlichkeit. Wenn wir Prior-Information $p(\theta|x)$ haben:

A posteriori erwarteter Verlust

$$r(d, p|x) = \mathbb{E}_\theta[L(d(X), \theta)|x] = \int_\Theta L(d(x), \theta)p(\theta|x)d\theta$$

Es gilt weiter:

$$\begin{aligned} r(d, p|x) &= \mathbb{E}_\theta\left(R(d, \theta)\right) = \mathbb{E}_\theta\left(\mathbb{E}_X[L(d(X), \theta)]\right) \\ &= \int_\Theta \left[\int_X L(d(x), \theta)f(x|\theta)dx\right] p(\theta)d\theta \end{aligned}$$

Bayes-Risiko im Bayes-Sinn

Minimieren wir den a posteriori erwarteten Verlust, erhalten wir:

Definition (Bayes-Risiko)

Für eine Verlustfunktion L und eine Priorverteilung p ist jede Entscheidung d^*, welche den a posteriori erwarteten Verlust $r(d, p|x)$ minimiert, Bayes-optimal.
Der Wert

$$r^*(p) = r(d^*, p|x)$$

heißt dann Bayes-Risiko.

Eigenschaften

- Bayes-optimale Entscheidungen sind immer zulässig.
- Das Bayes-Risiko ist immer kleiner oder gleich dem Minimax-Risiko.
- Wenn d_0 eine Bayes-optimale Entscheidung ist und $R(d, \theta) \leq r^*(p_0)$ für alle θ im Träger von p_0,
  dann ist d_0 die Minimax-Entscheidung und p_0 die ungünstigste Priorverteilung.

Bayesianische Punktschätzer

Für das statistische Entscheidungsproblem $d = \hat{\theta}$ gilt:

- Bei quadratischer Verlustfunktion ist der Posteriori-Erwartungswert Bayes-optimal.
- Bei absoluter Verlustfunktion ist der Posteriori-Median Bayes-optimal.
- Bei 0-1-Verlustfunktion ist der Posteriori-Modus (MAP) Bayes-optimal.
- Bei 0-1-Verlustfunktion und flacher Prior ($p(\theta) \propto 1$) ist der Maximum-Likelihood-Schätzer Bayes-optimal.

Fazit

Fazit

Welchen Bayesianischen Punkt-Schätzer wir nehmen, hängt also davon ab, welche Verlustfunktion wir wählen.

Prädiktion

- Wir wollen eine Vorhersage für einen neuen Wert der Zufallsvariable X, basierend auf $X = x$.
- Bayesscher Ansatz: Verwende die prädiktive Verteilung.
- Die prädiktive Verteilung berücksichtigt die Unsicherheit über den Parameter θ, basierend auf der Verteilung $\pi(\theta \mid x)$.

Prädiktive Verteilung

$$p(x_0 \mid x) = \int p(x_0 \mid \theta)\, \pi(\theta \mid x)\, d\theta$$

Prädiktive Posterior-Verteilung

Allgemeiner Fall

$$p(x_0 \mid x) = \int p(x_0 \mid \theta)\, \pi(\theta \mid x)\, d\theta$$

- Hierbei ist $p(x_0 \mid \theta)$ die Likelihood eines neuen Werts x_0 gegeben θ.
- $\pi(\theta \mid x)$ ist die posterior-Verteilung basierend auf den Daten x.

Prädiktive Posterior-Verteilung

Posterior-Prädiktive Verteilung

$$f(\tilde{x}|\lambda) = \frac{f(\tilde{x}, \lambda|\mathbf{x})}{p(\lambda|\mathbf{x})} \Leftrightarrow f(\tilde{x}, \lambda|\mathbf{x}) = f(\tilde{x}|\lambda)p(\lambda|\mathbf{x}) \tag{6}$$

$$f(\tilde{x}|\mathbf{x}) = \int f(\tilde{x}, \lambda|\mathbf{x})d\lambda = \int f(\tilde{x}|\lambda)p(\lambda|\mathbf{x})d\lambda \tag{7}$$

- Erwartungswert:

$$\mathbb{E}[x_0 \mid x] = \mathbb{E}_{\theta|x}[\mathbb{E}[x_0 \mid \theta]]$$

Prädiktive Dichte

- Likelihood: $X_i \sim Poi(\lambda)$
- Prior: $\lambda \sim Ga(a, b)$
- Posterior: $\lambda \mid \mathbf{x} \sim Ga(\tilde{a}, \tilde{b})$

$$
\begin{aligned}
f(x_z, \lambda \mid x) &= f(x_z \mid \lambda) p(\lambda \mid x) \\
&= \frac{\lambda^{x_z}}{x_z!} \exp(-\lambda) \cdot \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})} \lambda^{\tilde{a}-1} \exp(-\tilde{b}\lambda) \\
&= \frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})\Gamma(x_z+1)} \lambda^{\tilde{a}+x_z-1} \exp\left(-(\tilde{b}+1)\lambda\right)
\end{aligned}
$$

- $f(x_z, \lambda \mid x)$ ist die gemeinsame Dichte von x_z und λ.
- Uns interessiert aber die *prädiktive Dichte* $f(x_z \mid x)$, also ohne λ.
- Diese erhalten wir, indem wir die gemeinsame Dichte über λ integrieren:

$$
f(x_z \mid x) = \int f(x_z, \lambda \mid x)\, d\lambda
$$