Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

Heinrich-Heine Universität Düsseldorf

3. Dezember 2009

ML-Schätzer für stetige Zufallsvariable, Wiederholung

- Es sei Θ eine Menge von Parameterwerten. Zu jedem Parameterwert $\theta \in \Theta$ gebe es eine stetige Verteilung P_θ mit Dichte f_θ.
- Von der Zufallsvariablen X sei bekannt, dass ihre Verteilung gleich einem der P_θ ist. Dieses θ soll geschätzt werden.
- Für jede mögliche Realisierung x von X bezeichnet man die Abbildung
$$L_x(\theta) = f_\theta(x)$$
als *Likelihood-Funktion* zu x.
- Gesucht: Derjenige Wert von θ, für den $L_x(\theta)$ maximal wird.

ML-Schätzer, Beispiel Exponentialverteilung

- Die Lebensdauer von Glühbirnen ist näherungsweise exponentialverteilt, besitzt also eine Dichte

$$f_\lambda(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

  Der Parameter λ ist zu schätzen.
- Dazu werden die Lebensdauern von n Glühbirnen bestimmt.

ML-Schätzer zur Exponentialverteilung, Fortsetzung

- Stochastisches Modell: $X_1, \dots, X_n$ unabhängige Zufallsvariable mit Dichte f_λ. Die Dichte der gemeinsamen Verteilung muss verwendet werden.
- Dann ist die Likelihood-Funktion zur Realisierung $(x_1, \dots, x_n)$ gegeben durch

$$\begin{aligned} g(\lambda) &= f_\lambda(x_1) \cdot f_\lambda(x_2) \cdots f_\lambda(x_n) \\ &= \lambda^n \cdot e^{-\lambda \cdot x_1} \cdots e^{-\lambda \cdot x_n} = \lambda^n \cdot exp(-\lambda \cdot (x_1 + \cdots + x_n)) \end{aligned}$$

- Es ist zweckmäßig, den Logarithmus von g zu betrachten

$$h(\lambda) = \ln(g(\lambda)) = n \cdot \ln(\lambda) - \lambda \cdot (x_1 + \cdots + x_n)$$

- h hat dieselben Maximalstellen wie g, lässt sich aber leichter nach λ differenzieren.

ML-Schätzer zur Exponentialverteilung, Fortsetzung

- $h(\lambda) = n \cdot \ln(\lambda) - \lambda \cdot (x_1 + \cdots + x_n)$, also

$$h'(\lambda) = \frac{n}{\lambda} - (x_1 + \cdots + x_n)$$

- Einzige kritische Stelle

$$\lambda_0 = \frac{n}{x_1 + \cdots + x_n}$$

- Dort muss das Maximum liegen.
- Begründung: Das ist die einzige kritische Stelle, und die Randwerte sind

$$\lim_{\lambda \to 0} h(\lambda) = -\infty \quad \text{und} \quad \lim_{\lambda \to \infty} h(\lambda) = -\infty$$

ML-Schätzer zur Exponentialverteilung, Zahlenbeispiel

Die folgenden Brenndauern sind gemessen worden (in Stunden)

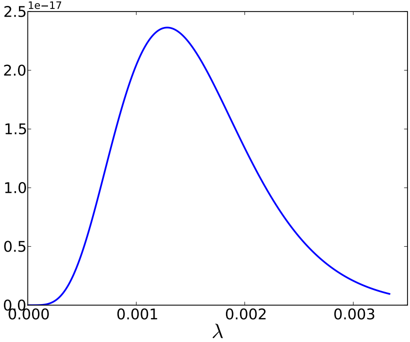
$$x_1 = 740, \quad x_2 = 800, \quad x_3 = 760, \quad x_4 = 600, \quad x_5 = 990$$

Dann ist der Maximum-Likelihood Schätzer für λ gleich

$$\hat{\lambda} = \frac{5}{740 + 800 + 760 + 600 + 990} = 0.001285$$

Die mittlere Brenndauer beträgt $1/0.001285 = 778$ Stunden.

Likelihood-Funktion im Glühlampenbeispiel



Graph von

$$h(\lambda) = \lambda^5 e^{-\lambda(740+800+760+600+990)}$$

Abschnitt 3.2

Konfidenzintervalle

Idee

- Parameterschätzungen geben keine Auskunft über ihre Genauigkeit.
- Jeder Schätzer ist selbst eine Zufallsvariable und hätte bei einem anderen Versuch unter denselben Bedingungen einen anderen Wert angenommen.
- Bei der Intervallschätzung wird im Gegensatz zur Parameterschätzung ein Intervall geschätzt, das den korrekten Wert mit vorgeschriebener Wahrscheinlichkeit enthält.
- Dieses Intervall heißt *Konfidenzintervall.*

Definition

- Es sei Θ eine Menge von Parameterwerten. Zu jedem Parameterwert $\theta \in \Theta$ gebe es eine Verteilung P_θ.
- Von der Zufallsvariablen X sei bekannt, dass ihre Verteilung gleich einem der P_θ ist. Für dieses θ soll ein Konfidenzintervall geschätzt werden.
- Ein Intervall $[G_u, G_o]$ mit der Eigenschaft

$$P_\theta(G_u \leq \theta \leq G_o) = 1 - \alpha$$

  heißt *Konfidenzintervall* für den Parameter θ zum Konfidenzniveau $1 - \alpha$.
- Die Ungleichung gilt dabei unabhängig von θ.
- Übliche Konfidenzniveaus sind 90%, 95% und 99%.
- G_u und G_o sind Zufallsvariable. Man bezeichnet sie als untere und obere *Vertrauensgrenze.*

Schätzung eines Konfidenzintervalls für den Erwartungswert bei bekannter Varianz mittels der 3σ-Regel

- $X_1, \ldots, X_n$ seien normalverteilt gemäß $N(\mu, \sigma^2)$ für bekanntes σ und unbekanntes μ
- Für μ soll ein Konfidenzintervall zum Konfidenzniveau 95% geschätzt werden.
- Verwende dazu das arithmetische Mittel

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

- 3σ-Regel

$$P\left(\mu - \frac{2\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

- Beachte die Gleichwertigkeit der Ungleichungen

$$\overline{X} \leq \mu + \frac{2\sigma}{\sqrt{n}} \quad \text{und} \quad \mu \geq \overline{X} - \frac{2\sigma}{\sqrt{n}}$$

Schätzung eines Konfidenzintervalls für den Erwartungswert bei bekannter Varianz, Fortsetzung

- Damit sieht die 3σ-Regel so aus

$$P\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.95$$

- Also ist folgendes Intervall ein Konfidenzintervall zum Konfidenzniveau 95%

$$\left[\overline{X} - \frac{2\sigma}{\sqrt{n}}, \quad \overline{X} + \frac{2\sigma}{\sqrt{n}}\right]$$

Beispiel: Roggenpflanzen

- Gesunde Roggenpflanzen einer bestimmten Art sind im Mittel 102.5 *cm* lang, wobei die Länge um 7 *cm* streut. Die Länge sei normalverteilt.
- Einige Pflanzen wurden geschädigt. Dadurch nimmt die mittlere Länge ab, ohne dass sich die Streuung ändert.
- Die folgenden Längen wurden gemessen

96.62	94.91	85.05	101.61	109.55
93.05	97.86	96.66	95.08	98.87

- Arithmetisches Mittel der Daten

$$\overline{x} = 96.93$$

Roggenpflanzen, Fortsetzung

- $\frac{2\sigma}{\sqrt{n}} = \frac{14}{\sqrt{10}} = 4.43$
- Also ergibt sich das folgende Konfidenzintervall zum Konfidenzniveau 95%

$$[92.50, 101.36]$$

- Mit 95-prozentiger Wahrscheinlichkeit liegt die mittlere Länge der geschädigten Pflanzen zwischen 92.50 *cm* und 101.36 *cm*
- Ersetzt man 2σ durch 3σ, so findet man mittels der 3σ-Regel das Konfidenzintervall zum Konfidenzniveau 99%

$$[90.29, 103.57]$$

- Zum Konfidenzniveau 99% kann nicht ausgeschlossen werden, dass die geschädigten Pflanzen sogar länger sind.

Quantile

- Die Zahl q_α mit $\Phi(q_\alpha) = \alpha$ heißt α-*Quantil* der Standardnormalverteilung. Die wichtigsten Quantile der Standardnormalverteilung sind tabelliert

$\Phi(u)$	70%	80%	90%	95%	97.5%	99%	99.5%
u	0.524	0.842	1.282	1.645	1.960	2.326	2.576

- Z. B. ist das 0.975-Quantil gleich 1.960, also näherungsweise gleich 2
- Umrechnungsformel
$$q_\alpha = -q_{1-\alpha}$$
- Beispiel
$$q_{0.05} = -q_{0.95} = -1.645$$
denn $\Phi(-1.645) = 1 - \Phi(1.645) = 1 - 0.95 = 0.05$

Quantile, Fortsetzung

- Das 50%-Quantil ist der Median, das 25%-Quantil das erste Quartil und das 75%-Quantil das dritte Quartil
- Die Quantile werden benötigt, wenn die 3σ-Regel nicht ausreicht.

Einseitige Konfidenzintervalle

- Bei einigen Fragestellungen ist statt eines Intervall nur eine obere oder untere Schranke für den Erwartungswert von Interesse.
- Ich variiere jetzt das Beispiel der Roggenpflanzen.
- Ich will nur noch wissen, ob die geschädigten Pflanzen im Mittel tatsächlich kürzer sind, d. h. ich suche eine obere Grenze für die mittlere Länge.
- Mit 99%-iger Wahrscheinlichkeit liegt die mittlere Halmlänge zwischen 90.29 cm und 103.57 cm.
- In der Fehlerrate von 1% ist die Unterschreitung der unteren Grenze berücksichtigt.

Schätzung eines einseitigen Konfidenzintervalls für den Erwartungswert bei bekannter Varianz

- $X_1, \ldots, X_n$ seien normalverteilt gemäß $N(\mu, \sigma^2)$ für bekanntes σ und unbekanntes μ
- Für μ soll ein einseitiges Konfidenzintervall $-\infty < \mu \leq G_o$ zum Konfidenzniveau $1 - \alpha$ geschätzt werden.
- Verwende dazu das arithmetische Mittel

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

- Seine Standardisierung ist

$$Y = \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu)$$

Schätzung eines einseitigen Konfidenzintervalls für den Erwartungswert bei bekannter Varianz, Fortsetzung

- $q_{1-\alpha}$ sei das $1-\alpha$-Quantil
- Y ist standardnormalverteilt, also

$$P(Y \geq -q_{1-\alpha}) = 1 - \Phi(-q_{1-\alpha}) = \Phi(q_{1-\alpha}) = 1 - \alpha$$

- Für $\overline{X}$ heißt das

$$P\left(\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu) \geq -q_{1-\alpha}\right) = 1 - \alpha$$

- Man stellt μ frei

$$P\left(\mu \leq \overline{X} + \frac{\sigma \cdot q_{1-\alpha}}{\sqrt{n}}\right) = 1 - \alpha$$

- Das Konfidenzintervall ist $-\infty < \mu \leq \overline{X} + \dfrac{\sigma \cdot q_{1-\alpha}}{\sqrt{n}}$

Beispiel: Roggenpflanzen

- Zurück zu den Roggenpflanzen: Dort $\overline{x} = 96.93$ und $\sigma = 7$, und es wurden 10 Messungen unternommen.
- Das Konfidenzniveau sei $1 - \alpha = 0.99$. Das Quantil ist

$$q_{0.99} = 2.326$$

- Damit erhält man als obere Schranke für das einseitige Konfidenzintervall

$$\overline{x} + \frac{\sigma \cdot q_{1-\alpha}}{\sqrt{n}} = 96.93 + \frac{7 \cdot 2.326}{\sqrt{10}} = 102.1$$

- Zum Vergleich: Die obere Schranke des zweiseitigen Konfidenzintervalls zum selben Konfidenzniveau unter Benutzung der 3σ-Regel war 103.57

Tabelle: Konfidenzintervall für den Erwartungswert bei bekannter Varianz

zweiseitig zum Konfidenzniveau $1-\alpha$

$$\overline{X} - \frac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{\sigma \cdot q_{1-\alpha/2}}{\sqrt{n}}$$

einseitig, obere Schranke zum Konfidenzniveau $1-\alpha$

$$-\infty < \mu \leq \overline{X} + \frac{\sigma \cdot q_{1-\alpha}}{\sqrt{n}}$$

einseitig, untere Schranke zum Konfidenzniveau $1-\alpha$

$$\overline{X} - \frac{\sigma \cdot q_{1-\alpha}}{\sqrt{n}} \leq \mu < \infty$$