WuS - Lecture Notes Week 11

Ruben Schenk, ruben.schenk@inf.ethz.ch

July 4, 2022

1 Konfidenzintervalle

Die Grundidee ist wie folgt: Wie im vorigen Abschnitt suchen wir aus einer Familie $(\mathbb{P}_\theta)_{\theta\in\Theta}$ von Modellen eines, das zu unseren Daten $x_1, ..., x_n$ passt. Ein Schätzer für θ gibt uns dabei einen einzelnen zufälligen möglichen Parameterwert. Weil es schwirig ist, mit diesem einen Wert den richtigen Parameter zu treffen, suchen wir nun stattdessen eine **zufällige Teilmenge des Parameterbereichs,** die hoffentlich den wahren Parameter enthält.

1.1 Definitionen

Eir reichhaltig sind diese Schätzer? Werfen wir zum Beispiel eine Münze 100 mal, ohne die Wahrscheinlichkeit p von Kopf zu kennen. Falls wir 70 mal Kopf erhalten, ist der Maximum-Likelihood-Schätzer für p $T_{ML} = 0.7$. Wie weit ligt T_{ML} von dem wahren Wert p entfernt? Um diese Art von Fragen zu beantworten, führen wir den Begriff der Konfidenzintervalle ein.

Def: Sei $\alpha \in [0, 1]$. Ein **Konfidenzintervall für θ mit Niveau** $1-\alpha$ ist ein Zufallsintervall $I = [A, B]$, sodass gilt

$$\forall\theta \in \Theta \quad \mathbb{P}_\theta[A \leq \theta \leq B] \geq 1 - \alpha,$$

wobei A, B Zufallsvariablen der Form $A = a(X_1, ..., X_n)$, $B = b(X_1, ..., X_n)$ mit $a, b : \mathbb{R}^n \to \mathbb{R}$ sind.

Zu bemerken ist hier, dass θ in dieser Gleichung deterministisch und nicht zufällig ist. Nur die Schranken $A = a(X_1, ..., X_n)$ und $B = b(X_1, ..., X_n)$ sind Zufallsvariablen.

Beispiel (Konfidenzintervall für normales Modell mit bekannter Varianz): Seien $X_1, ..., X_n$ u.i.v. normalverteilte Zufallsvariablen mit Parametern m und $\sigma^2 = 1$. Wir betrachten somit ein stochastisches Modell mit bekannter Varianz ($\sigma^2 = 1$) aber unbekanntem Mittelwert μ ($X_1 \sim \mathcal{N}(\mu, 1)$). Mann kann zeigen, dass der Maximum-Likelihood Schätzer gegeben ist durch

$$T = T_{ML} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

mit $T_{ML} \sim \mathcal{N}(\mu, \frac{1}{n})$, also $Z = \sqrt{n}(T_{ML} - \mu) \sim \mathcal{N}(0, 1)$ (wir normalisieren T_{ML}). Wir suchen also für μ Konfidenzintervalle der Form

$$I = [T_{ML} - \frac{c}{\sqrt{n}}, T_{ML} + \frac{c}{\sqrt{n}}].$$

Zuerst betrachten wir

$$\mathbb{P}_\theta[T_{ML} - \frac{c}{\sqrt{n}} \leq mu \leq T_{ML} + \frac{c}{\sqrt{n}}] = \mathbb{P}_\theta[-c \leq Z \leq c].$$

Somit können wir die obige Wahrscheinlichkeit explizit bestimmen:

$$\mathbb{P}_\theta[-c \leq Z \leq c] = \mathbb{P}_\theta[Z \leq c] - \mathbb{P}_\theta[Z < -c] \,(= 1 - \mathbb{P}_\theta[Z \leq c]) = 2\Phi(c) - 1,$$

wobei $\Phi(c) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} \exp(\frac{-x^2}{2})\, dx$ ist und mittels einer Tabelle der Standardnormalverteilung nachgelesen werden kann. Somit ergibt sich $2\Phi(1.96) - 1 \geq 0.95$ und schliesslich:

$$I = [T_{ML} - \frac{1.96}{\sqrt{n}}, T + \frac{1.96}{\sqrt{n}}].$$

1.2 Verteilungsaussagen

Def: Eine stetige Z.V. X heisst χ^2**- Verteilung mit** m **Freiheitsgraden** falls ihre Dichte gegeben ist durch

$$f_X(y) = \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} e^{-\frac{1}{2}y} \quad \text{fuer } y \geq 0.$$

Wir schreiben dann $X \sim \chi^2_m$. Dabei ist die sogenannte Gamma-Funktion für $v \geq 0$ gegeben durch

$$\Gamma(v) := \int_0^\infty t^{v-1} e^{-t}\, dt.$$

Es gilt $\Gamma(n) = (n-1)!$ für $v = n \in \mathbb{N}$.

Bemerkung: Die χ^2 Verteilung mit m Freiheitsgraden ist der Spezialfall einer $Ga(\alpha, \lambda)$-Verteilung mit $\alpha = \frac{m}{2}$ und $\lambda = \frac{1}{2}$. Für $m = 2$ ergibt sich eine Exponentialverteilung mit $X \sim Exp(\frac{1}{2})$.

Satz: Sind die Z.V. $X_1, ..., X_m$ u.i.v. $\sim \mathcal{N}(0, 1)$, so ist die Summe $Y := \sum_{i=1}^{m} X_i^2 \sim \chi^2_m$.

Def: Eine stetige Z.V. X heisst t**verteilt mit** m **Freiheitsgraden** falls ihre Dichte gegeben ist durch

$$f_X(x) = \frac{\Gamma(\frac{m+1}{2})}{\sqrt{m\pi}\Gamma(\frac{m}{2})} \Big(1 + \frac{x^2}{m}\Big)^{-\frac{m+1}{2}} \quad \text{fuer } x \in \mathbb{R}.$$

Wir schreiben dann $X \sim t_m$.

Bemerkung: Für $m = 1$ ist das eine Cauchy-Verteilung, und für $m \to \infty$ erhält man asymptotisch eine $\mathcal{N}(0, 1)$-Verteilung. Wie die $\mathcal{N}(0, 1)$-Verteilung ist die t-Verteilung symmetrisch um 0, sie ist aber langschwänziger (d.h. ihre Dichte geht langsamer gegen 0, wenn das Argument gegen $\pm\infty$ geht), und zwar umso mehr, je kleiner m ist.

Satz: Sind X und Y unabhängig mit $X \sim \mathcal{N}(0,\,1)$ und $Y \sim \chi^2_m$, so ist der Quotient

$$Z := \frac{X}{\sqrt{\frac{1}{m}Y}}.$$

1.3 Normalverteilung mit σ und m unbekannt

Wir erinnern an die Notationen

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2$$

für das Stichprobenmittel und die Stichprobenvarianz.

Satz: Seien $X_1, ..., X_n$ i.i.d. $\sim \mathcal{N}(\mu,\,\sigma^2)$. Dann sind $\bar{X_n}$ und S^2 unabhängig.

In diesem Fall sind $\bar{X_n}$ und S^2 unsere Schätzer für μ und σ^2. Es gilt zudem $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Beispiel: Betrachten wir nun $A = \frac{\bar{X_n}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,\,1)$ und $B = \sqrt{\frac{\frac{n-1}{\sigma^2}S^2}{n-1}}$, wobei $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$, dann gilt

$$Z := \frac{A}{B} = \frac{\bar{X_n} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}.$$

Damit können wir ein Konfidenzintervall für μ der Form

$$[A,\,B] = \Big[\bar{X_n} - c \cdot \sqrt{\frac{S^2}{n}},\, \bar{X_n} + c \cdot \sqrt{\frac{S^2}{n}}\Big],$$

wobei wir dazu wieder die Tabelle für die t_{n-1}-Verteilung benutzen, um einen geigneten Wert für c zu finden.