WuS - Lecture Notes Week 9

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

June 24, 2022

1 Statistische Grundideen

Wir befassen uns im Folgenden mit der **induktiven Statistik.** Die Grundidee dabei ist wie folgt: Man fasst die Daten $x_1, ..., x_n$ auf als Realisierung / realisierte Werte $X_1(\omega), ..., X_n(\omega)$ von Z.V. $X_1, ..., X_n$, und sucht dann Aussagen über die Verteilung von $X_1, ..., X_n$.

Wichtig: Man muss immer sauber unterscheiden zwischen den *Daten* $x_1, ..., xn$ und dem generierenden Mechanismus $X_1, ..., X_n$ (bezeichnet mit grossen Buchstaben, sind Z.V., also Funktionen auf einem Ω).

Terminologie: Die Gesamtheit der Beobachtungen $x_1, ..., x_n$ oder Z.V. $X_1, ..., X_n$ nennt man of eine **Stichprobe**, die Anzahl n heisst dann der **Stichprobenumfang.**

2 Schätzer

Setup:

- Parameterraum $\Theta \subset \mathbb{R}$
- Grundraum Ω
- sigma-Algebra $\mathcal{F}$
- $(\mathbb{P}_\theta)_{\theta \in \Theta}$ Familie von Wahrscheinlichkeitsmasse auf (Ω, $\mathcal{F}$)
- $X_1, ..., X_n$ Zufallsvariablen auf (Ω, $\mathcal{F}$)

2.1 Grundbegriffe

Wir suchen für den Parameter θ einen Schätzer T aufgrund unserer Stichprobe $(X_1, ..., X_n)$.

Def: Ein **Schätzer** ist eine Zufallsvariable $T : \Omega \to \mathbb{R}$ der Form

$$T = t(X_1, ..., X_n),$$

wobei $t : \mathbb{R}^n \to \mathbb{R}$.

Einsetzen von Daten $x_i = X_i(\omega)$, $i = 1, ..., n$ liefert dann **Schätzwerte** $T(\omega) = t(x_1, ..., x_n)$ für θ.

Beispiel: Jemand behauptet zu schmecken, ob in einer Tasse Tee zuerst die Milch oder der Tee eingegossen worden ist. Wie kann man überprüfen, ob das stimmen kann?
Wie geben der Person an n Tagen je zwei Tassen, von welchen Sie sagen soll, in welcher zuerst die Milch und in welcher zuerst der Tee eingegossen wurde. Wir notieren uns dabei die Ergebnisse $x_1, ..., x_n \in \{0, 1\}$ und fassen wie üblich diese Daten als Realisation von Z.V. $X_1, ..., X_n$ auf. Dann ist $S_n = \sum_{i=1}^{n} X_i$ die zufällige Anzahl der korrekt klassifizierten Tassenpaare, und $s_n = \sum_{i=1}^{n} x_i$ die beobachtete Anzahl von Erfolgen.
Als Modell nehmen wir nun an, dass die X_i unter $\mathbb{P}_\theta$ i.i.d. $\sim \text{Ber}(\theta)$ mit $\theta \in \Theta = [0, 1]$ sind. Dann ist natürlich $S_n \sim \text{Bin}(n, \theta)$ unter $\mathbb{P}_\theta$, d.h. im Modell $\mathbb{P}_\theta$, das zu θ gehört, ist die Anzahl S_n der Erfolge binomialverteilt mit Parametern n und θ.
Weil wir den Parameter θ nicht kennen, liegt es nahe, zuerst einmal dafür einen Schätzer zu suchen. Eine erste Möglichkeit wäre, einfach das letzte Ergebnis zu nehmen. Unser erster Schätzer $\hat{T}$ für θ wäre also $\hat{T} = X_n$. Ein zweiter naheliegender Schätzer wäre die durchschnittliche Anzahl der Erfolge bei den n Versuchen. Unser zweiter Schätzer wäre also $T = \bar{X_n} = \frac{1}{n} S_n$. Für gegebene Daten $x_1, ..., x_n$ gibt uns das dann zwei Schätzwerte $\hat{t}(x_1, ..., x_n) = x_n$ und $t(x_1, ..., x_n) = \bar{x_n} = \frac{1}{n} \sum_{i=1}^{n} x_i$, die wir konkret berechnen können.

2.2 Bias

Def: Ein Schätzer T heiss **erwartungstreu (unbiased)** für θ, falls für alle $\theta \in \Theta$ gilt:

$$\mathbb{E}_\theta[T] = \theta$$

Die Interpretation dazu ist wie folgt: Im Mittel (über alle denkbaren Realisationen ω) schätzt T also richtig, und zwar unabhängig davon, welches Modell $\mathbb{P}_\theta$ zu Grunde liegt.

Def: Sei $\theta \in \Theta$ und T ein Schätzer. Der **Bias** (oder erwartete Schätzfehler) von T im Modell $\mathbb{P}_\theta$ ist definiert als

$$\mathbb{E}_\theta[T] - \theta.$$

Der mittlere quadratische Schätzfehler (mean squared error, MSE) von T im Modell $\mathbb{P}_\theta$ ist definiert als

$$\text{MSE}_\theta[T] := \mathbb{E}_\theta[(T - \theta)^2].$$

Bemerkung: Mann kann den MSE zerlegen als

$$\text{MSE}_\theta[T] = \mathbb{E}_\theta[(T - \theta)^2] = \text{Var}_\theta[T] + (\mathbb{E}_\theta[T] - \theta)^2,$$

also in die Summe aus der Varianz des Schätzers T un dem Quadrat des Bias.