Statistik & Data Science - Zusammenfassung

Statistik

Grundbegriffe

Zufallsvariablen

Verteilungen

Definition 0.1 (Bernoulli-Verteilung): Eine Zufallsvariable X heißt bernoulliverteilt, falls sie Werte in $\{0,1\}$ annimmt und ihre Wahrscheinlichkeitsfunktion durch $p_X(x) = p^x(1-p)^{1-x}$ gegeben ist. In diesem Fall schreiben wir $X \sim \mathcal{B}(1,p)$.

Satz 0.1 (Erwartungwert & Varianz der Bernoulli-Verteilung): Sei $X \sim \mathcal{B}(1,p)$. Dann gilt $\mathbb{E}[X] = p$ und $\mathbb{V}[X] = p(1-p)$.

Beweis: ■

Definition 0.2 (Binomial-Verteilung): Eine Zufallsvariable X heißt binomialverteilt, falls sie Werte in $\{0,...,n\}$ annimmt und ihre Wahrscheinlichkeitsfunktion durch $p_X(x) = \binom{n}{x} p^x(1-p)^{n-x}$ gegeben ist. In diesem Fall schreiben wir $X \sim \mathcal{B}(n,p)$.

Satz 0.2 (Erwartungwert & Varianz der Binomial-Verteilung): Sei $X \sim \mathcal{B}(n,p)$. Dann gilt $\mathbb{E}[X] = np$ und $\mathbb{V}[X] = np(1-p)$.

Beweis: ■

Definition 0.3 (Poisson-Verteilung): Eine Zufallsvariable X heißt poissonverteilt mit Parameter λ, falls sie Werte in $\{0,1,...\}$ annimmt und ihre Wahrscheinlichkeitsfunktion durch $p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ gegeben ist. In diesem Fall schreiben wir $X \sim \mathcal{P}(\lambda)$.

Satz 0.3 (Erwartungwert & Varianz der Poisson-Verteilung): Sei $X \sim \mathcal{P}(\lambda)$. Dann gilt $\mathbb{E}[X] = \lambda = \mathbb{V}[X]$.

Beweis: ■

Definition 0.4 (Normal-Verteilung): Eine Zufallsvariable X heißt normalverteilt mit Parametern μ und σ^2, falls sie Werte in $\mathbb{R}$ annimmt und ihre Dichtefunktion durch $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ gegeben ist. In diesem Fall schreiben wir $X \sim \mathcal{N}(\mu, \sigma^2)$.

Satz 0.4 (Erwartungwert & Varianz der Normal-Verteilung): Sei $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt $\mathbb{E}[X] = \mu$ und $\mathbb{V}[X] = \sigma^2$.

Beweis: ■

Definition 0.5 (Uniforme Verteilung): Eine Zufallsvariable X heißt uniform auf $[a,b]$ verteilt, falls sie Werte in $\mathbb{R}$ annimmt und ihre Dichtefunktion durch $f_X(x) = \frac{1}{b-a}\mathbb{1}_{\{x \in [a,b]\}}(x)$ gegeben ist. In diesem Fall schreiben wir $X \sim \mathcal{U}[a,b]$.

Erwartungswert & Varianz

Unabhängigkeit

Ungleichungen & Grenzwertsätze

Satz 0.5 (Markow-Ungleichung): Sei $X \geq 0$, $a > 0$. Dann gilt: $\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$.

Beweis: ■

Satz 0.6 (Chebyshev-Ungleichung): $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \frac{\mathbb{V}[X]}{a^2}$.

Beweis: ■

Satz 0.7 (Schwaches Gesetz der großen Zahlen): Seien $X_1, ..., X_n \overset{\text{iid}}{\sim} X$. Dann gilt: $\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_i - \mathbb{E}[X]\right| \geq \varepsilon\right] \leq \frac{\mathbb{V}}{n\varepsilon^2}$.

Beweis: ■

Satz 0.8 (Zentraler Grenzwertsatz): Sei X eine Zufallsvariable mit $\mu := \mathbb{E}[X]$, $\sigma^2 := \mathbb{V}[X]$, sowie X_i iid mit $X_i \sim X$. Sei $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ und $S_n^* := \frac{\overline{X}_n - \mu}{\sqrt{n}\sigma}$. Dann gilt: $\lim_{n\to\infty} \mathbb{P}[a \leq S_n^* \leq b] = \Phi(b) - \Phi(a)$, wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

Beweis: ■

1 Parameterschätzung

Maximum likelihood estimation (MLE)

Asymptotische Normalität

Asymptotische Normalität ist in der mathematischen Statistik eine Eigenschaft von Statistiken bzw. Schätzern.
In diesem Abschnitt wollen wir untersuchen, unter welchen Voraussetzungen asymptotische Normalität für einen MLE gegeben ist.

Definition 0.6 (Konsistenz eines Schätzers): Ein Schätzer $\hat{\theta}^n$ heißt konsistent, falls $\hat{\theta}^n \xrightarrow{p} \theta$, also wenn $\hat{\theta}^n$ in Wahrscheinlichkeit gegen θ konvergiert, d.h.: $\forall \varepsilon > 0 : \mathbb{P}\left[\left|\hat{\theta}^n - \theta\right| > \varepsilon\right] \to 0$.

Definition 0.7 (Asymptotische Normalität eines Schätzers): Ein Schätzer $\hat{\theta}^n$ heißt asymptotisch normal, falls ein σ_θ^2 existiert mit $\sqrt{n}\left(\hat{\theta}^n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_\theta^2\right)$, also der Schätzer in Verteilung gegen $\mathcal{N}\left(0, \sigma_\theta^2\right)$ konvergiert. *Zur Wiederholung: Konvergenz in Verteilung bedeutet, dass die Verteilungen der betrachteten Zufallsvariablen im Grenzwert identisch werden.*

Definition 0.8 (Kullback-Leibler Divergenz): Die Kullback-Leibler Divergenz zweier Verteilungen $f(x|\theta)$ und $f\left(x|\tilde{\theta}\right)$ ist gegeben durch:

$$D_{\mathrm{KL}}\left(f(x|\theta) \parallel f\left(x|\tilde{\theta}\right)\right) := \mathbb{E}_\theta\left[\log \frac{f(X|\theta)}{f\left(X|\tilde{\theta}\right)}\right] = \int f(x|\theta) \log \frac{f(X|\theta)}{f\left(X|\tilde{\theta}\right)} \, \mathrm{d}x.$$

Hierbei ist der Erwartungswert so zu verstehen, dass er von einer Funktion $g(X)$ der Zufallsvariablen $X \sim f(x|\theta)$ berechnet wird. Diese Funktion können wir explizit als $g(x) = \log \frac{f(X|\theta)}{f\left(X|\tilde{\theta}\right)}$ anschreiben.

In diesem Abschnitt wollen wir folgendes zeigen:

Satz 0.9: Unter gewissen Voraussetzungen gilt: $\widehat{\theta}^{(n)}_{\mathrm{MLE}}$ ist **konsistent** und **asymptotisch normal** mit $\sigma^2_\theta := \frac{1}{I(\theta)\sigma}$.

Hierbei bezeichnet $I(\theta)$ die *Fisher-Information*, welche wir in Kürze definieren werden.

Welche Voraussetzungen sind das? Um das zu präzisieren brauchen wir ein paar Resultate.

Zunächst stellen wir einen Zusammenhang zwischen MLE und Kullback-Leibler-Divergenz her. Das Maximierungsproblem für das Auffinden des MLE lautet bekanntlich $\max_{\tilde{\theta}} \ell\left(\tilde{\theta}\right)$ bzw. $\max_{\tilde{\theta}} \log\left(\ell\left(\tilde{\theta}\right)\right)$ mit likelihood-Funktion $l\left(\tilde{\theta}\right) = \prod_{i=1}^n f\left(x_i|\tilde{\theta}\right)$. Dies kann äquivalent wie folgt formuliert werden:

$$\max_{\tilde{\theta}} \log\left(\ell\left(\tilde{\theta}\right)\right) = \max \sum_{i=1}^n \log\left(f\left(x_i|\tilde{\theta}\right)\right) = -\min \sum_{i=1}^n \log\left(f\left(x_i|\tilde{\theta}\right)\right)$$

Das arg min dieses Problems ändert sich nicht wenn wir eine Konstante (d.h. einen Term unabhängig von $\tilde{\theta}$) dazuaddieren und mit einer Konstante $\frac{1}{n} > 0$ multiplizieren:

$$\begin{aligned}\log\left(\ell\left(\tilde{\theta}\right)\right) &= \frac{1}{n}\left(-\sum_{i=1}^n \log\left(f\left(x_i|\tilde{\theta}\right)\right) + \sum_{i=1}^n \log(f(x_i|\theta))\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^n \log \frac{f(x_i|\theta)}{f\left(x_i|\tilde{\theta}\right)}\right) =: R_n\left(\tilde{\theta}, \theta\right)\end{aligned}$$

Der Ausdruck $R_n\left(\tilde{\theta}, \theta\right)$ kann als *loss function* einer *Empirical Risk Minimization* aufgefasst werden. Wir fassen diese Beziehung nochmal zusammen:

$$\theta^n_{\mathrm{MLE}} = \arg\min_{\tilde{\theta}} R_n\left(\tilde{\theta}, \theta\right)$$

Wenn wir R_n als Sample-Mean $\frac{1}{n}\sum_{i=1}^n g(X_i)$ von Funktionen der Zufallsvariablen $X_1, ..., X_n$ interpretieren, bekommen wir für $n \to \infty$ mit dem Gesetz der großen Zahlen:

$$R_n\left(\tilde{\theta}, \theta\right) \to \mathbb{E}_\theta\left[\log \frac{f(X|\theta)}{f\left(X|\tilde{\theta}\right)}\right] = \int \log \frac{f(x|\theta)}{f\left(x|\tilde{\theta}\right)} f(x|\theta)\,\mathrm{d}x = D_{\mathrm{KL}}\left(f(x|\theta) \| f\left(x|\tilde{\theta}\right)\right) =: R\left(\tilde{\theta}, \theta\right)$$

womit wir einen Zusammenhang zur Kullback-Leibler-Divergenz gefunden haben.

Bevor wir unsere Annahmen präzisieren, beweisen wir noch eine elementare Eigenschaft der Kullback-Leibler-Divergenz:

Proposition 0.1 (Nicht-Negativität der KL-Divergenz): $D_{\mathrm{KL}}(X\|Y) \geq 0$ und $D_{\mathrm{KL}}(X\|Y) = 0 \leftrightarrow X = Y$ in Verteilung.

Beweis:

1. $D_{\mathrm{KL}}(X\|Y) = \int \log\Big(\frac{f_X(x)}{f_Y(x)}\Big) f_X(x)\,\mathrm{d}x = -\int \log\Big(\frac{f_Y(x)}{f_X(x)}\Big) f_X(x)\,\mathrm{d}x$. Bemerke: $g(x) := -\log(x)$ ist konvex. Daher kann Jensen-Ungleichung angewandt werden: $\ldots = \mathbb{E}[g(h(x))] \geq g(\mathbb{E}[h(x)]) \geq -\log\Big(\int \frac{f_Y(x)}{f_X(x)} f_X(x)\,\mathrm{d}x\Big) = -\log(\int f_Y(x)\,\mathrm{d}x) = -\log(1) = 0$.
2. $\leftarrow$ trivial, $\rightarrow$ folgt aus Gleichheitsbedingung der Jensen-Ungleichung: Wenn g strikt konvex dann gilt Gleichheit genau dann wenn h(x) konstant [Skript Beiglböck], d.h. $f_Y(x) = c \cdot f_X(x) \rightarrow$ X=Y in Verteilung.

■

Bemerkung 0.1: $X = Y$ in Verteilung bedeutet $F_X = F_Y$, d.h. die Verteilungsfunktionen stimmen überein.

Wir präzisieren nun unsere Annahmen:

1. **Starke Identifizierbarkeit:** Eine grundlegende Voraussetzung zur Konsistenz von Schätzern ist die sogenannte Identifizierbarkeit. Diese besagt, dass unterschiedliche Werte der Parameter zu unterschiedlichen Wahrscheinlichkeitsverteilungen der beobachtbaren Zufallsvariablen führen müssen. Mathematisch ausgedrückt:

   $$\theta_1 \neq \theta_2 \Rightarrow f_X(x|\theta_1) \neq f_X(x|\theta_2)$$

   Um unser Resultat zu beweisen brauchen wir eine etwas stärkere Aussage:

   $$\forall \varepsilon > 0 : \inf_{|\tilde{\theta}-\theta|>\varepsilon} D_{\mathrm{KL}}\Big(f(.|\theta)\|f\Big(.|\tilde{\theta}\Big)\Big) = \eta_\varepsilon > 0$$

   Diese Bedingung ist im Wesentlichen dieselbe wie die normale Identifizierbarkeit, außer dass sie verhindert, dass der Unterschied zwischen den beiden Verteilungen verschwindend klein wird (er bleibt immer mindestens so groß wie ihre KL-Divergenz. Aufrgund der vorher gezeigten Nicht-Negativität der KL-Divergenz ist der Ausdruck garantiert > 0). Die beiden Bedingungen sind äquivalent, wenn θ auf eine kompakte Menge beschränkt wird.

2. **Uniformes Gesetz der großen Zahlen:** Das normale Gesetz der großen Zahlen ist als "punktweise" im Hinblick auf die Punkte $\tilde{\theta}$ zu verstehen. Die uniforme (gleichmäßige) Version sieht wie folgt aus (vergleiche mit dem Begriff der gleichmäßigen Konvergenz):

   $$\forall \varepsilon > 0 : \mathbb{P}\left[\sup_{\tilde{\theta}} \Big|R_n\Big(\tilde{\theta}, \theta\Big) - R\Big(\tilde{\theta}, \theta\Big)\Big| > \varepsilon\right] \rightarrow 0$$

Nun sind wir gewappnet das erste fundamentale Resultat dieses Abschnitts zu beweisen, und zwar die Konsistenz des MLE:

Satz 0.10 (Konsistenz des MLE): Sei $X_1, \ldots, X_n$ eine iid-Folge von Zufallsvariablen mit $X_i \sim f(x_i|\theta)$. Für die Folge gelte starke Identifizierbarkeit und uniformes GgZ. Dann ist θ^n_{MLE} konsistent.

Beweis: Fixiere ein $\varepsilon > 0$. Unter Verwendung der starken Identifizierbarkeit sehen wir, dass für jedes $\varepsilon > 0$ ein $\eta_\varepsilon > 0$ existiert, sodass

$$D_{\mathrm{KL}}\Big(f(x|\ \theta)\|f\Big(x|\tilde{\theta}\Big)\Big) \geq \eta_\varepsilon$$

wenn $|\tilde{\theta} - \theta| \geq \varepsilon$. Wir werden zeigen, dass für θ^n_{MLE} gilt, dass $D_{\mathrm{KL}}(f(x|\theta)\|f(x|\theta^n_{\mathrm{MLE}})) \leq \eta_\varepsilon$ für $n \rightarrow \infty$ in Wahrscheinlichkeit. Dies impliziert wiederum, dass $|\theta^n_{\mathrm{MLE}} - \theta| \leq \varepsilon$, was wiederum impliziert, dass $\theta^n_{\mathrm{MLE}} \xrightarrow{p} \theta$. Es bleibt zu zeigen, dass $D_{\mathrm{KL}}(f(x|\theta)\|f(x|\theta^n_{\mathrm{MLE}})) \leq \eta_\varepsilon$, wenn $n \rightarrow \infty$. Beachte, dass

$$D_{\mathrm{KL}}(f(x|\theta)\|f(x|\theta^n_{\mathrm{MLE}})) = R(\theta^n_{\mathrm{MLE}},\theta) = R(\theta^n_{\mathrm{MLE}},\theta) - R_n(\theta^n_{\mathrm{MLE}},\theta) + R_n(\theta^n_{\mathrm{MLE}},\theta)$$
$$\xrightarrow{p} R(\theta^n_{\mathrm{MLE}},\theta) - R_n(\theta^n_{\mathrm{MLE}},\theta) \xrightarrow{p} 0$$

wobei die finale Konvergenz das uniforme GgZ verwendet. Die zu zeigende Ungleichung folgt, da

$$R_n(\theta^n_{\mathrm{MLE}},\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i|\theta)}{f(X_i|\theta^n_{\mathrm{MLE}})} \leq 0.$$

■

Nun wollen wir uns der asymptotischen Normalität widmen.

Definition 0.9 (Fisher-Information): Sei $f(x|\theta)$ die von einem Parameter θ abhängige Verteilung der Zufallsvariable X. Die Fisher-Information $I(\theta)$ ist gegeben durch

$$I(\theta) := -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log(f(X|\theta))\right].$$

Lemma 0.1:
- $\frac{\partial^2}{\partial\theta^2} D_{\mathrm{KL}}\left(f(.|\theta)\|f\left(.|\tilde{\theta}\right)\right)|_{\theta=\tilde{\theta}} = I(\theta)$
- $\frac{\partial}{\partial\theta} D_{\mathrm{KL}}\left(f(x|\theta)\|f\left(x|\tilde{\theta}\right)\right)\ |_{\theta=\tilde{\theta}} = 0$

Beweis: Übung. ■

Lemma 0.2: $I(\theta) = \mathbb{E}\left[\left(\left(\frac{\partial}{\partial\theta}\right)\log f(X|\theta)\right)^2\right]$

Beweis:

$$I(\theta) \overset{\text{def}}{=} -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log(f(X|\theta))\right] = -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\right] \overset{\text{Quot.regel}}{=} \mathbb{E}_\theta\left[\frac{\left(\frac{\partial^2}{\partial\theta^2}f(X|\theta)\right)f(X|\theta) - \left(\frac{\partial}{\partial\theta}f(X|\theta)\right)^2}{f(X|\theta)^2}\right]$$

$$= \mathbb{E}_\theta\left[\frac{\left(\frac{\partial^2}{\partial\theta^2}f(X|\theta)\right)f(X|\theta)}{f(X|\theta)^2}\right] - \mathbb{E}_\theta\left[\frac{\left(\frac{\partial}{\partial\theta}f(X|\theta)\right)^2}{f(X|\theta)^2}\right] = \underbrace{\mathbb{E}_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X|\theta)}{f(X|\theta)}\right]}_{=:A} - \underbrace{\mathbb{E}_\theta\left[\left(\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\right)^2\right]}_{=:B}.$$

Betrachte A und B:

$$A = \int \frac{\frac{\partial^2}{\partial\theta^2}f(x|\theta)}{f(x|\theta)} f(x|\theta)\,\mathrm{d}x = \int \frac{\partial^2}{\partial\theta^2} f(x|\theta)\,\mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\underbrace{\int f(x|\theta)\,\mathrm{d}x}_{=1} = 0$$

$$B = \int \left(\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right)^2 f(x|\theta)\,\mathrm{d}x = \int \left(\frac{\partial}{\partial\theta}f(x|\theta)\right)^2 f(x|\theta)\,\mathrm{d}x = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}f(x|\theta)\right)^2\right]$$

■

Satz 0.11 (Asymptotische Normalität des MLE): Wenn θ^n_{MLE} konsistent, dann ist θ^n_{MLE} asymptotisch normal mit $\sigma^2_\theta = \frac{1}{I(\theta)}$, d.h. es gilt

$$\sqrt{n}(\theta^n_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right).$$

Beweis: siehe Notizen ■

Bias-Variance Zerlegung

Definition 0.10 (MSE):

$$\text{MSE}\big(\tilde{\theta}, \theta\big) := \mathbb{E}\Big[\big|\tilde{\theta} - \theta\big|^2\Big]$$

Bemerkung 0.2 (Anwendung der Markow-Ungleichung): Eine mit dieser neuen Notation zur **Markow-Ungleichung** äquivalentes Resultat lautet wie folgt:

$$\mathbb{P}\big[\big|\tilde{\theta} - \theta\big| > \varepsilon\big] \leq \frac{\text{MSE}\big(\tilde{\theta}, \theta\big)}{\varepsilon^2}$$

Lemma 0.3 (Bias-Variance-Zerlegung):

$$\text{MSE}\big(\tilde{\theta}, \theta\big) = \big|\mathbb{E}\big[\tilde{\theta}\big] - \theta\big|^2 + \mathbb{V}\big[\tilde{\theta}\big]$$

Beweis:

$$\begin{aligned}
\text{MSE}\big(\tilde{\theta}, \theta\big) = \mathbb{E}\Big[\big|\tilde{\theta} - \theta\big|^2\Big] &= \mathbb{E}\left[\left(\tilde{\theta} - \underbrace{\mathbb{E}\tilde{\theta} + \mathbb{E}\tilde{\theta}}_{0} - \theta\right)^2\right] \\
&= \underbrace{\mathbb{E}\Big[\big(\tilde{\theta} - \mathbb{E}\tilde{\theta}\big)^2\Big]}_{\mathbb{V}\tilde{\theta}} + 2\mathbb{E}\left[\big(\tilde{\theta} - \mathbb{E}\tilde{\theta}\big)\underbrace{\big(\mathbb{E}\tilde{\theta} - \theta\big)}_{\text{konstant}}\right] + \mathbb{E}\left[\underbrace{\big(\mathbb{E}\tilde{\theta} - \theta\big)^2}_{\text{konstant}}\right] \\
&= \mathbb{V}\tilde{\theta} + 2\big(\mathbb{E}\tilde{\theta} - \theta\big)\underbrace{\mathbb{E}\big[\tilde{\theta} - \mathbb{E}\tilde{\theta}\big]}_{0} + \big|\mathbb{E}\tilde{\theta} - \theta\big|^2 \\
&= \mathbb{V}\tilde{\theta} + \big|\mathbb{E}\tilde{\theta} - \theta\big|^2
\end{aligned}$$

■

Satz 0.12 (Cramer-Rao): Seien $X_1, ..., X_n \overset{\text{iid}}{\sim} f(x|\theta)$, sei $\tilde{\theta} : \mathbb{R} \to \mathbb{R}$ eine beliebige Statistik, $\beta(\theta) := \mathbb{E}_\theta\big[\tilde{\theta}\big] - \theta$. Dann gilt:

$$\mathbb{V}\big[\tilde{\theta}\big] \geq \frac{\big(\beta'(\theta) + 1\big)^2}{n \cdot I(\theta)}.$$

Beweis: siehe Notizen ■

Bemerkung 0.3:

- $\beta(\theta)$ heißt *Bias* von $\tilde{\theta}$. Falls $\tilde{\theta}$ erwartungstreu, dann gilt $\beta(\theta) = 0$. Daraus folgt:

$$\mathrm{MSE}\big(\tilde{\theta}, \theta\big) = \mathbb{V}\tilde{\theta} \geq \frac{1}{n \cdot I(\theta)}.$$

- Falls θ^n_{MLE} erwartungstreu, dann gilt:

$$\lim_{n\to\infty} \frac{\mathbb{V}[\theta^n_{\mathrm{MLE}}]}{\frac{1}{n\cdot I(\theta)}} = 1,$$

  da aufgrund der asymptotischen Normalität des MLE gilt $\mathbb{V}[\theta^n_{\mathrm{MLE}}] \to \frac{1}{n\cdot I(\theta)}$ für $n \to \infty$. Dies bezeichnet man auch als *asymptotische Optimalität.*

Suffiziente Statistiken

Ang. wir haben Daten $X_1, ..., X_n \overset{\text{iid}}{\sim} f(x|\theta)$ und möchten θ schätzen. Können wir die Daten reduzieren, ohne Information über θ zu verlieren?

Definition 0.11 (Suffizienz einer Statistik): Eine Statistik $T : \mathbb{R}^n \to \mathbb{R}$ heißt suffizient für θ, falls für alle $(x_1, ..., x_n) \in \mathbb{R}^n$, $t \in \mathbb{R}$:

$$\mathbb{P}[X_1 \leq x_1, ..., X_n \leq x_n | T = t] \text{ ist unabhängig von } \theta.$$

Bemerkung 0.4:

- Die Definition ist so zu interpretieren, dass falls wir $t = T(x_1, ..., x_n)$ kennen, dann hat weitere Kenntniss von $X_1, ..., X_n$ keinen zusätzlichen Informationsgehalt mehr bzgl. θ.
- Man kann zeigen:

$$\mathrm{T} \text{ suffizient} \Leftrightarrow I(\theta, T(X_1, ..., X_n)) = I(\theta, X_1, ..., X_n),$$

  wobei I in diesem Falls als eine "Mutual Information" zu deuten ist.

Satz 0.13 (Fisher-Neyman-Faktorisierung): Für $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}$ gilt:

$$T \text{ ist suffizient für } \theta \Leftrightarrow f(x_1, ..., x_n|\theta) = g(T(x_1, ..., x_n)|\theta) \cdot h(x_1, ..., x_n).$$

In Worten: T ist genau dann suffizient, wenn eine Faktorisierung in $g \cdot h$ existiert, wobei g von T abhängen kann, h aber nur von der Stichprobe $x_1, ..., x_n$.

Beweis: siehe Notizen ■

Bedingte Erwartung

Definition 0.12 (Bedingte Wahrscheinlichkeit, bedingte Erwartung): Seien X, Y diskrete Zufallsvariablen.

$$\mathbb{P}[Y = j, X = i] := \frac{\mathbb{P}[Y = j, X = i]}{\mathbb{P}[X = i]},$$

$$\mathbb{E}[Y|X = i] := \sum_j j \cdot \mathbb{P}[Y = j|X = i] = \sum_j j \cdot \frac{\mathbb{P}[Y = j, X = i]}{\mathbb{P}[X = i]},$$

$$\mathbb{E}[Y|X] := \sum_i \mathbb{E}[Y|X = i] \cdot \mathbb{1}_{\{X=i\}}.$$

$\mathbb{E}[Y|X = i]$ ist eine reelle Zahl, $\mathbb{E}[Y|X]$ hingegen eine Zufallsvariable welche den Wert $\mathbb{E}[Y|X = i]$ annimmt genau dann wenn $X(\omega) = i$.

Lemma 0.4:
- $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$
- $\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$
- $\mathbb{E}[g(Y)|X]$ ist Funktion von X
- X, Y unabhängig $\Rightarrow \mathbb{E}[g(Y)|X] = \mathbb{E}[g(Y)]$

Beweis:

■

Wir widmen uns nun einem Resultat, mit dem man aus einem zunächst beliebigem (bzw. beliebig schlechtem) Schätzer durch eine Transformation einen in gewissem Sinne "optimalen" Schätzer konstruieren kann.

Satz 0.14 (Rao-Blackwell): Sei $X_1, ..., X_n \overset{\text{iid}}{\sim} f(X|\theta)$, $\tilde{\theta}$ ein beliebiger erwartungstreuer Schätzer für θ und T eine suffiziente Statistik für θ. Sei $\theta^* := \mathbb{E}\left[\tilde{\theta}|T\right]$. Dann gilt:
- θ^* ist erwartungstreuer Schätzer
- $\text{MSE}(\theta^*, \theta) \leq \text{MSE}\left(\tilde{\theta}, \theta\right)$.

Die Transformation von $\tilde{\theta}$ durch den Ausdruck θ^* wird oft "Rao-Blackwellization" genannt. Das Resultat garantiert, das der MSE nach der Transformation kleiner wird - durch dieses Verfahren erhalten wir also einen besseren Schätzer!

Beweis:
1. θ^* ist erwartungstreuer Schätzer: $\mathbb{E}[\theta^*] = \mathbb{E}\left[\mathbb{E}\left[\tilde{\theta}|T\right]\right] = \mathbb{E}\left[\tilde{\theta}\right] = \theta$, womit Erwartungstreue gilt. Wir müssen noch überprüfen, ob θ^* tatsächlich ein Schätzer ist. Eine Statistik ist Schätzer eines Parameters θ, wenn sie unabhängig von θ ist. Wir überprüfen:
$$\theta^* \overset{\text{def}}{=} \mathbb{E}\left[\tilde{\theta}|T\right] = \int \tilde{\theta}(x) \cdot \underbrace{f(x|T)}_{\text{unabh. von } \theta} \, \mathrm{d}x.$$
2. Da $\tilde{\theta}$ bzw. θ^* erwartungstreu sind, erhalten wir mittels Bias-Variance-Zerlegung $\text{MSE}\left(\tilde{\theta}, \theta\right) = \mathbb{V}\left[\tilde{\theta}\right]$ bzw. $\text{MSE}(\theta^*, \theta) = \mathbb{V}[\theta^*]$. Nun gilt nach vorherigen Lemma:
$$\begin{aligned} \mathbb{V}\left[\tilde{\theta}\right] &= \mathbb{V}\left[\underbrace{\mathbb{E}\left[\tilde{\theta}|T\right]}_{\theta^*}\right] + \mathbb{E}\left[\mathbb{V}\left[\tilde{\theta}|T\right]\right] \\ \Leftrightarrow \mathbb{V}[\theta^*] &= \mathbb{V}\left[\tilde{\theta}\right] - \mathbb{E}\left[\mathbb{V}\left[\tilde{\theta}|T\right]\right] \leq \mathbb{V}\left[\tilde{\theta}\right], \end{aligned}$$
wobei die Ungleichung dadurch gerechtfertigt ist, dass die (bedingte) Varianz und somit ihr Erwartungswert stets nichtnegativ ist.

■

Beispiel 0.1:
1. Für nicht-suffiziente Statistiken T ist liefert das Resultat nicht unbedingt einen Schätzer: Seien $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$, $\tilde{\theta} := \frac{1}{2}(X_1 + X_2)$ und $T(X) := X_1$ eine nicht-suffiziente Statistik. T ist nicht suffizient, da für die gemeinsame Verteilung $f(x_1, ..., x_n|\theta)$ von $X_1, ..., X_N$ gilt:

$$f(x_1,...,x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(x_i-\theta)^2}{2}\right) = (2\pi)^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$
$$= (2\pi)^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i^2 - 2x_i\theta + \theta^2)\right)$$
$$= (2\pi)^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right) \cdot \exp\left(\theta\sum_{i=1}^{n} x_i\right) \cdot \exp\left(-\frac{1}{2}n\theta^2\right)$$

Wir finden für unser T keine gültigen $g(T(X_1,...,X_n,\theta))$, $h(X_1,...,X_n)$ mit $f = g \cdot h$.
(Bemerke: Für $U(X_1,...,X_n) \coloneqq \sum_{i=1}^{n} X_i$ gäbe es eine solche Faktorisierung, und zwar:

$$f = \underbrace{(2\pi)^{-\frac{n}{2}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right)}_{h(X_1,...,X_n)} \cdot \underbrace{\exp\left(\theta\sum_{i=1}^{n} x_i\right) \cdot \exp\left(-\frac{1}{2}n\theta^2\right)}_{g(U,\theta)},$$

d.h. U wäre suffizient).
Wir wollen also versuchen, $\tilde{\theta}$ mit Hilfe von T zu verbessern. Wir erhalten durch Rao-Blackwellization:

$$\theta^* = \mathbb{E}\left[\tilde{\theta}|T\right] = \mathbb{E}\left[\frac{1}{2}(X_1+X_2)|T\right] = \frac{1}{2}\left(\underbrace{\mathbb{E}[X_1|X_1]}_{\substack{=X_1 \\ \text{siehe} \\ \text{Def bed. Erw.}}} + \underbrace{\mathbb{E}[X_2|X_1]}_{\substack{=\mathbb{E}[X_2] \\ \text{weil } X_1,X_2 \\ \text{unabh.}}}\right) = \frac{1}{2}X_1 + \frac{1}{2}\theta.$$

θ^* ist erwartungstreu, da:

$$\mathbb{E}[\theta^*] = \mathbb{E}\left[\frac{1}{2}X_1 + \frac{1}{2}\theta\right] = \frac{1}{2}\mathbb{E}[X_1] + \frac{1}{2}\theta = \frac{1}{2}\theta + \frac{1}{2}\theta = \theta.$$

Außerdem gilt:

$$\mathbb{V}[\theta^*] = \mathbb{V}\left[\frac{1}{2}X_1 + \frac{1}{2}\theta\right] = \mathbb{V}\left[\frac{1}{2}X_1\right] = \frac{1}{4} < \frac{1}{2} = \frac{1}{4}(1+1) = \mathbb{V}\left[\frac{1}{2}(X_1+X_2)\right] = \mathbb{V}\left[\tilde{\theta}\right].$$

Aber: θ^* ist <u>kein</u> Schätzer, da der Ausdruck nicht unabhängig von θ ist.

Konfidenzintervalle
In diesem Abschnitt wollen wir Konfidenzintervalle konstruieren. Dafür suchen wir für eine gegebene Verfeilungsfunktion

$$F_X : \mathbb{R} \longrightarrow [0,1]$$
$$x \longmapsto \mathbb{P}[X \leq x] =: F_X(x)$$

eine Art Inverse.

Definition 0.13 (Quantilsfunktion):

$$\begin{aligned} F_X^- : (0,1) &\longrightarrow \mathbb{R} \\ \alpha &\longmapsto \inf\{t \in \mathbb{R} \mid F_X(t) \geq \alpha\} =: F_X^-(\alpha) \end{aligned}$$

heißt Quantilsfunktion von X,

$$q_\alpha := F_X^-(\alpha)$$

heißt α-Quantil.

Dies ist also quasi eine Umkehrfunktion von F_X. Aber warum so umständlich? Es wäre doch naheliegender, die Quantilsfunktion als $F_X^{-1}(\{\alpha\})$ im Sinne des Urbildes oder noch besser gleich als $F_X^{-1}(\alpha)$ im Sinne einer Umkehrfunktion zu definieren? Nun, dass eine Funktion nicht unbedingt eine Umkehrfunktion besitzt ist einleuchtend, also wäre die zweite Option wohl zu optimistisch. Die erste Variante, bei der wir vom Urbild von F_X Gebrauch machen, ist faktisch unsere Definition der Quantilsfunktion, nur garantieren wir mit dieser Definition zusätzlich, dass wir einen eindeutigen Wert erhalten. In der VO haben wir gesehen, dass die Quantilsfunktion der Bernoulliverteilung sonst nicht eindeutig wäre, da sie stückweise konstant ist.

Für die Definition eines Konfidenzintervalls brauchen wir den Begriff des **Intervallschätzers**. Im Gegensatz zu den uns bereits bekannten *Punktschätzern* geben diese Intervalle $[a(X), b(X)]$ aus, in denen der Wert θ mit hoher Wahrscheinlichkeit liegen soll.

Definition 0.14 (Intervallschätzer): Ein Intervallschätzer ist eine Funktion

$$\begin{aligned} \mathcal{X}^n &\longrightarrow \mathcal{I}(\mathbb{R}) \\ X &\longmapsto [a(X), b(X)], \end{aligned}$$

wobei $\mathcal{X}^n$ den Wertebereich der Zufallsvariablen $X = (X_1, ..., X_n)$ und $\mathcal{I}(\mathbb{R})$ die Menge der Intervalle aus $\mathbb{R}$ bezeichnet.

Definition 0.15 (Konfidenzintervall): Sei $X \longmapsto [a(X), b(X)]$ ein Intervallschätzer, $X \sim f(x|\theta)$ einer von θ abhängigen Verteilung folgend. Falls

$$\mathbb{P}_\theta[a(X) \leq \theta \leq b(X)] = \gamma$$

für alle $\theta \in \Theta$, so heißt $[a(.), b(.)]$ ein $100\gamma\%$-iges Konfidenzintervall.

Beispiel 0.2: Seien $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ^2 bekannt, μ unbekannt. Bestimme ein $100\gamma\%$-iges Konfidenzintervall für μ.

Betrachte

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Durch Standardisierung erhalten wir

$$\frac{\left(\overline{X} - \mu\right)\sqrt{n}}{\sigma} \sim \mathcal{N}(0,1),$$

wofür wir mittels Technologieeinsatz die Quantilen $z_{\frac{1-\gamma}{2}}$ und $z_{\gamma+\frac{1-\gamma}{2}} = z_{\frac{1+\gamma}{2}}$ errechnen können (siehe VO Notizen für Skizze). Somit gilt:

$$\mathbb{P}\left[z_{\frac{1-\gamma}{2}} \leq \frac{\left(\overline{X}-\mu\right)\sqrt{n}}{\sigma} \leq z_{\frac{1+\gamma}{2}}\right] = \gamma$$

$$\Leftrightarrow \mathbb{P}\left[\frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1-\gamma}{2}} - \overline{X} \leq -\mu \leq \frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1+\gamma}{2}} - \overline{X}\right] = \gamma$$

$$\Leftrightarrow \mathbb{P}\left[-\frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1+\gamma}{2}} + \overline{X} \leq \mu \leq -\frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1-\gamma}{2}} + \overline{X}\right] = \gamma$$

$$\Leftrightarrow \mathbb{P}\left[-\frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1+\gamma}{2}} + \overline{X} \leq \mu \leq \frac{\sigma}{\sqrt{n}} \cdot z_{\frac{1+\gamma}{2}} + \overline{X}\right] = \gamma,$$

wobei wir in der letzten Zeile verwendet haben, dass $z_{\frac{1-\gamma}{2}} = -z_{\frac{1+\gamma}{2}}$. Somit haben wir eine explizite Formel zur Berechnung des Konfidenzintervalls!

Die Zufallsvariable $\frac{(\overline{X}-\mu)\sqrt{n}}{\sigma}$ war in diesem Beispiel maßgeblich, um das Konfidenzintervall explizit anschreiben zu können. Diese Zufallsvariable ist ein Beispiel für eine sogenannte *Pivot-Zufallsvariable*.

Definition 0.16 (Pivot): Seien $X_1, ..., X_n \sim$ iid $f(x|\theta), x_i \in \mathcal{X}, \theta \in \Theta$. Eine Funktion $g : \mathcal{X}^n \times \Theta \longrightarrow \mathbb{R}$ heißt **Pivot**, falls:
1. $\theta \mapsto g(x_1, ..., x_n, \theta)$ stetig ist für alle $(x_1, ..., x_n) \in \mathcal{X}^n$ und
2. Die Verteilung von $g(X_1, ..., X_n, \theta)$ nicht von θ abhängt.

Unsere obige Zufallsvariable ist sicherlich stetig in μ und ihre Verteilung $\mathcal{N}(0,1)$ hängt offensichtlich nicht von μ ab, also ist sie ein Pivot.

Mit diesem Begriff lässt sich eine **generelle Vorgehensweise** zum Auffinden von Konfidenzregionen erklären:
1. Finde Pivot-Zufallsvariable $g(X_1, ..., X_n, \theta)$
2. Bestimme Quantile q_1 und q_2 mit $\mathbb{P}[q_1 \leq g(X_1, ..., X_n, \theta) \leq q_2] = \gamma$
3. $C := \{\theta : g(X_1, ..., X_n, \theta) \in [q_1, q_2]\}$ ist $100\gamma\%$-ige Konfidenzregion. Falls g affin, dann ist C ein Intervall.

Der letzte Schritt entspricht in der Praxis dem Umformen der Ungleichung $q_1 \leq g(X_1, ..., X_n, \theta) \leq q_2$ nach θ, so wie wir es in obigem Beispiel gemacht haben.

Proposition 0.2 (Student-t-Verteilung): Seien $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, σ^2 unbekannt, μ unbekannt. Dann ist

$$T := \left(\overline{X} - \mu\right)\frac{\sqrt{n}}{S} \sim t_{n-1}$$

eine Pivot-Zufallsvariable, die einer Student-t-Verteilung mit $n-1$ Freiheitsgraden folgt. $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ bezeichnet die Stichprobenvarianz.

Im Allgemeinen existieren Pivots nicht in praktikable Form. Daher bedienen wir uns einer asymptotischen Variante:

Definition 0.17 (Asymptotischer Pivot): Seien $X_1, ..., X_n \overset{\text{iid}}{\sim} f_\theta$. $g : \mathcal{X}^n \times \Theta \longrightarrow \mathbb{R}$ heißt asymptotischer Pivot, falls:
1. $\theta \mapsto g(x_1, ..., x_n, \theta)$ stetig ist für alle $n \in \mathbb{N}$, $(x_1, ..., x_n) \in \mathcal{X}^n$ und
2. $g(X_1, ..., X_n, \theta)$ konvergiert in Verteilung für $n \to \infty$ gegen eine Verteilung, die nicht von θ abhängt.

Nun lautet hier die Vorgehensweise:
1. Bestimme Quantile q_1, q_2 mit $\lim_{n\to\infty} \mathbb{P}[q_1 \leq g(X_1, ..., X_n, \theta) \leq q_2] = \gamma$
2. $C_n := \{\theta : g(X_1, ..., X_n, \theta) \in [q_1, q_2]\} \Rightarrow \lim_{n\to\infty} \mathbb{P}[\theta \in C_n] = \gamma$

Bayessche Schätzung
tba (siehe Notizen Tauböck 10.4.)

Hoeffding-Ungleichung
Das Resultat dieses kurzen Abschnitts, die *Hoeffding-Ungleichung*, hat gar nicht so viel mit Statistik bzw. dem nächsten Kapitel *Hypothesentests* zu tun, aber ist generell relevant und wird für uns später noch von Bedeutung sein. Bevor wir die eigentliche Ungleichung beweisen, zeigen wir noch ein Hilfsresultat:

Lemma 0.5 (Hoeffding): Sei $a \leq X \leq b$, $\mathbb{E}[X] = 0$. Dann gilt für alle $\lambda > 0$:

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}$$

wichtig für uns

Satz 0.15 (Hoeffding-Ungleichung):

2 Hyptothesentests
tba

Data Science

Lineare Regression

Statistical learning theory

Maschinelles Lernen

Perceptron Algorithmus

Logistische Regression

Kernelization

PCA

Johnson-Lindenstrauss Lemma

Graphen

Clustering

Spektrale Graphentheorie

Fourier Transformation