Statistik & Data Science - Zusammenfassung

Statistik

Grundbegriffe

Zufallsvariablen

Verteilungen

Definition 0.1 (Bernoulli-Verteilung): Eine Zufallsvariable X heißt <u>bernoulliverteilt</u>, 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 <u>binomialverteilt</u>, 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 <u>poissonverteilt</u> 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 <u>normalverteilt</u> 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 <u>uniform</u> auf $[a, b]$ <u>verteilt</u>, 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

Konsistenz: https://www.stat.cmu.edu/~larry/=stat705/Lecture15.pdf

Asymptotische Normalität: https://ocw.mit.edu/courses/18-443-statistics-for-applications-fall-2006/03b407da8a94b3fe22d987453807ca46_lecture3.pdf

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 $\sigma_\theta^2 := \frac{1}{I(\theta)\sigma}$.

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 $X_1, \ldots, X_n$ 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): $D_{\text{KL}}(X\|Y) \geq 0$ und $D_{\text{KL}}(X\|Y) = 0 \leftrightarrow X = Y$ in Verteilung.

Beweis:
1. $D_{\text{KL}}(X\|Y) = \int \log\left(\frac{f_X(x)}{f_Y(x)}\right) f_X(x) \, \mathrm{d}x = -\int \log\left(\frac{f_Y(x)}{f_X(x)}\right) f_X(x) \, \mathrm{d}x$. Bemerke: $g(x) := -\log(x)$ ist konvex. Daher kann Jensen-Ungleichung angewandt werden: … $= \mathbb{E}[g(h(x))] \geq g(\mathbb{E}[h(x)]) \geq -\log\left(\int \frac{f_Y(x)}{f_X(x)} f_X(x) \, \mathrm{d}x\right) = -\log(\int f_Y(x) \, \mathrm{d}x) = -\log(1) = 0$.
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:

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

   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):

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

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 $X_1, …, X_n$ 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

$$D_{\text{KL}}\left(f(x| \ \theta) \| f\left(x|\tilde{\theta}\right)\right) \geq \eta_\varepsilon$$

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 $n \to \infty$ 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 $n \to \infty$. Beachte, dass

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

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):

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

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

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

Lemma 0.3 (Bias-Variance-Zerlegung):

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

Beweis:

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

■

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, $\beta(\theta) := \mathbb{E}_\theta\left[\tilde{\theta}\right] - \theta$. Dann gilt:

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

Beweis: siehe Notizen ■

Bemerkung 0.3:
- $\beta(\theta)$ heißt *Bias* von $\tilde{\theta}$. Falls $\tilde{\theta}$ erwartungstreu, dann gilt $\beta(\theta) = 0$. Daraus folgt:

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

- 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:

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

  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

Konfidenzintervalle

Bayessche Schätzung

2 Hyptothesentests

Data Science

Lineare Regression

Statistical learning theory

Maschinelles Lernen

Perceptron Algorithmus

Logistische Regression

Kernelization

PCA

Johnson-Lindenstrauss Lemma

Graphen

Clustering

Spektrale Graphentheorie

Fourier Transformation