Institut für Angewandte Mathematik
Sommersemester 2022
Andreas Eberle, Luis Aniello La Rocca

11. Übungsblatt „Einführung in die Statistik”

Abgabe bis Mittwoch 29.6., 16 Uhr.

---

1. (Stochastische Dominanz) Seien μ und ν zwei Wahrscheinlichkeitsverteilungen auf $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, und seien $G_\mu(u) = \inf\{c \in \mathbb{R} : F_\mu(c) \geq u\}$ und $G_\nu(u)$ die verallgemeinerten Inversen der Verteilungsfunktionen von μ und ν.

a) Zeigen Sie, dass die folgenden Aussagen alle äquivalent sind:

(i) μ wird von ν *stochastisch dominiert* ($\mu \preceq \nu$), d.h.
$$\mu[(c,\infty)] \;\leq\; \nu[(c,\infty)] \quad \text{für alle } c \in \mathbb{R}.$$

(ii) Für alle $c \in \mathbb{R}$ gilt $F_\mu(c) \geq F_\nu(c)$.

(iii) Für alle $u \in (0,1)$ gilt $G_\mu(u) \leq G_\nu(u)$.

(iv) Es gibt Zufallsvariablen X und Y mit Verteilungen μ und ν, welche auf einem gemeinsamen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ definiert sind, so dass
$$\mathbb{P}[X \leq Y] \;=\; 1\,.$$

(v) Für jede monoton wachsende Funktion $h : \mathbb{R} \to [0,\infty)$ gilt
$$\mathbb{E}_\mu[h] \leq \mathbb{E}_\nu[h].$$

b) Zeigen Sie für $n, m \in \mathbb{N}$ und $p, q \in [0,1]$:

(i) $\mathrm{Bin}(n,p) \preceq \mathrm{Bin}(n,q)$ genau dann, wenn $p \leq q$.

(ii) Für $n \leq m$ gilt $\mathrm{Bin}(n,p) \preceq \mathrm{Bin}(m,p)$.

2. (Mann-Whitney U-Test a.k.a. Wilkoxon Rangsummentest) Wir beobachten die Realisierungen von $n = k + l$ unabhängigen reellen Zufallsvariablen $X_1, \ldots, X_{k+l}$ mit unbekannten Verteilungen
$$X_1, \ldots X_k \sim \mu \quad \text{und} \quad X_{k+1}, \ldots, X_{k+l} \sim \nu.$$
Wir nehmen an, dass die Verteilungsfunktionen stetig sind, so dass die Stichproben mit Wahrscheinlichkeit 1 alle verschieden sind. Im Mann-Whitney U-Test für das Testproblem
$$H_0 : \;\mu = \nu \quad \text{vs.} \quad H_1 : \;\mu \prec \nu$$
verwirft man die Nullhypothese, falls der Wert der *U-Statistik*
$$U \;=\; \sum_{i=1}^{k} \sum_{j=k+1}^{k+l} 1_{\{X_i > X_j\}}$$
strikt unterhalb eines vorgegebenen Schwellenwerts c liegt. Hierbei bedeutet $\mu \prec \nu$, dass $\mu \preceq \mu$ und $\nu \neq \mu$, siehe Aufgabe 1.

a) Zeigen Sie, dass sich die U-Statistik auch darstellen lässt als

$$U = W - \frac{k(k+1)}{2} \quad \text{mit} \quad W = \sum_{i=1}^{k} R_i \quad \text{(Wilcoxons Rangsumme).}$$

Hierbei bezeichnet R_i den (fast sicher eindeutigen) Rang von X_i unter $X_1, \dots X_{k+l}$.

b) Zeigen Sie $\mathbb{E}[U] = kl/2$ und $\mathrm{Var}[U] = kl(k+l+1)/12$.

c) Man kann beweisen, dass die U-Statistik asymptotisch normalverteilt ist, das heißt $(U - \mathbb{E}[U]) / \sqrt{\mathrm{Var}[U]}$ konvergiert für $k, l \to \infty$ in Verteilung gegen die Standardnormalverteilung, siehe z.B. Satz 11.27 im Buch von Georgii. Leiten Sie daraus eine Wahl für den Schwellenwert c her, so dass der Test asymptotisch das Niveau α hat.

d) In einer medizinischen Studie wurde ein physiologischer Parameter bei zwölf Diabetikern und Nichtdiabetikern gemessen. Die Frage ist, ob es einen systematischen Unterschied zwischen beiden Gruppen in Bezug auf diesen Parameter gibt. Ermitteln Sie einen approximativen p-Wert.

Diab.	11,5	12,1	16,1	17,8	24,0	28,8	33,9	40,7	51,3	56,2	61,7	69,2
Nichtdiab.	4,1	6,3	7,8	8,5	8,9	10,4	11,5	12,0	13,8	17,6	24,3	37,2

3. (Informationsungleichung, effiziente Schätzer und exponentielle Familien)
Seien $X_1, \dots, X_n$ unabhängige Stichproben von einer Wahrscheinlichkeitsverteilung μ_θ mit Dichte $g_\theta(x)$, wobei $\theta \in \mathbb{R}$ ein unbekannter Parameter ist.

a) Formulieren Sie in diesem Modell die Informationsungleichung.

b) In Aufgabe 1 von Blatt 3 haben wir für $\mu_\theta = \mathrm{Unif}(0, \theta)$ einen erwartungstreuen Schätzer mit mittlerem quadratischen Fehler von der Ordnung $O(n^{-2})$ konstruiert. Warum ist dies kein Widerspruch zur Informationsungleichung?

c) Zeigen Sie, dass für die Doppelexponentialverteilung mit Dichte $g_\theta(x) = \frac{1}{2}e^{-|x-\theta|}$ der Median ein asymptotisch effizienter Schätzer ist.

d) Beweisen Sie unter geeigneten Voraussetzungen, dass ein Schätzer $T(X)$ für θ genau dann effizient ist, wenn sich die Likelihood-Funktion in der Form

$$L(\theta; x) = e^{c(\theta)T(x) + d(\theta) + U(x)}$$

mit reellwertigen Funktionen c, d und U darstellen lässt.