A 1995. évi Schweitzer verseny valószínűségfeladata,

illetve néhány mögötte levő érdekes a nagy eltérések elméletéhez tartozó kérdés.

Az 1995. évi Schweitzer verseny valószínűség feladata
Legyen $\xi_k = \eta_k + \vartheta$, ahol η_k, $k = 1, 2, \dots$, független valószínűségi változók ismert $F(x)$ eloszlással, ϑ pedig ismeretlen paraméter. Ha az $F(x)$ eloszlás által meghatározott mérték nem abszolut folytonos a Lebesgue mértékre nézve, akkor alkalmas $\alpha > 0$ szám esetén minden $\varepsilon > 0$-ra létezik olyan $T_n(x_1, \dots, x_n)$ függvény és $E \subset R^1$ halmaz, melyre $\lambda(E) < \varepsilon$, ahol λ jelöli a Lebesque mértéket, és

$$P(|T_n(\xi_1, \dots, \xi_n) - \vartheta| \notin E) < e^{-n\alpha}$$

minden elég nagy n-re.

Ha az $F(x)$ függvénynek létezik folytonos sűrűségfüggvénye, akkor minden $\alpha > 0$ számra létezik olyan $\varepsilon = \varepsilon(\alpha) > 0$ szám és $n(\varepsilon, \alpha)$ küszöbindex, amelyekre $n > n(\varepsilon, \alpha)$ és $\lambda(E) < \varepsilon$ esetén tetszőleges $T_n(x_1, \dots, x_n)$ függvényre

$$P(|T_n(\xi_1, \dots, \xi_n) - \vartheta| \notin E) > e^{-n\alpha} \ .$$

A feladat jobb megértése érdekében tekintsük a következő természetes jelöltet a ϑ becslésre: $T_n(\xi_1, \dots, \xi_n) = \dfrac{1}{n} \sum\limits_{k=1}^{n} (\xi_k - E\eta_k)$. Ez a becslés még a legszebb esetekben sem elégíti ki a feladat feltételeit. ($e^{-n\alpha}$ kicsiségű hiba nem lehetséges nagyon kis $\varepsilon > 0$ esetén.)

Legyen adva van két eloszlásfüggvény, F és G, és legyen $\xi_1, \dots, \xi_n$ n független egyforma eloszlású valószínűségi változó F eloszlásfüggvénnyel. Vezessük be az $F_n(x) = \dfrac{1}{n}\#\{p\colon 1 \le p \le n,\ \xi_p < x\}$ empirikus eloszlásfüggvényt. Be lehet bizonyítani, hogy $-\dfrac{1}{n} \log |F_n(x) - G(x)| < \varepsilon) \sim I(G\|F) = \int \log \dfrac{dG(u)}{dF(u)}\, dG(u)$ kis ε-ra és nagy n-re. Ez a híres Sanov tétel, mely informálisan megfogalmazva azt jelenti, hogy jellemezni tudjuk, milyen (exponenciálisan kicsi) valószínűséggel viselkedik egy F eloszlású minta úgy, mintha G eloszlású minta volna. Ennek a valószínűségnek az exponensét az $I(G\|F)$ úgynevezett I-divergencia adja meg. Ez azért érdekes a tárgyalt feladat szempontjából, mert a ϑ paraméterre akkor tudunk jó becslést adni, ha az F eloszlásfüggvény az előbbi értelemben jól megkülönböztethető eltoltjaitól. Ezt a megkülönböztethetőséget az I divergenciával lehet mérni. A feladat állítása pongyolán megfogalmazva azt jelenti, hogy egy eloszlásfüggvény akkor különböztethető meg jól eltoltjaitól, ha nem abszolut folytonos a Lebesgue mértékre nézve.

1.) Bizonyítsuk be, hogy a ϑ paraméter lokalizálható a következő értelemben: Minden $\alpha > 0$-hoz lehet definiálni olyan $A_n = A(\alpha, \xi_1, \dots, \xi_n, F)$, $B_n = B(\alpha, \xi_1, \dots, \xi_n, F)$ számokat úgy, hogy $P(A_n < \vartheta_n < B_n) > 1 - e^{-n\alpha}$, és $B_n - A_n < L$ egy determenisztikus (n-től független) L számmal.

Segítség: Legyen $A_n = \min\limits_{1\le k\le n} \xi_k$, $B_n = \max\limits_{1\le k\le n} \xi_k$, ha $B_n - A_n$ nem túl nagy.

Lássuk be a Schweitzer feladat pozitív állítását. Vezessük be a következő jelöléseket. Legyen az F eloszlás által meghatározott μ mérték dekompozíciója $\mu = \mu_c + \mu_s$, ahol μ_c a mérték abszolut folytonos, és μ_s a mérték szinguláris része. Legyen μ_s az $\mathbf{A}$ halmazra koncentrálva, azaz legyen $\mu_s(\mathbf{A}) = c > 0$, $\mu_s(R \setminus \mathbf{A}) = 0$, $\lambda(\mathbf{A}) = 0$. Mivel $\mu(\mathbf{A}) > 0$, az F eloszlású $\eta_1, \dots, \eta_n$ valószínűségi változók pozitív része esik az $\mathbf{A}$ halmazba. Ez akkor igaz a $(\xi_1, \dots, \xi_n) = (\eta_1 + \vartheta, \dots, \eta_n + \vartheta)$ változókra is, ha $\mu(\mathbf{A} - \vartheta) > 0$. Viszont majdnem minden ϑ-ra $\mu(\mathbf{A} - \vartheta) = 0$. E tény segítségével lehet kiszűrni a nem jó ϑ értékek nagy részét. Illetve érdemes az állítást folytonosítani, hogy ϑ-t nem tartalmazó kis intervallumokat is ki tudjuk szűrni.

2.) Legyen $\mathbf{K} = \{t\colon t \in R^1,\ \mu(\mathbf{A} - t) > 0\}$. Ekkor $\lambda(\mathbf{K}) = 0$. Létezik olyan kompakt $\mathbf{B} \subset \mathbf{A}$ halmaz, melyre $\mu(\mathbf{B}) \ge \dfrac{c}{2}$. Jelölje $\mathbf{B}^\delta = \{t\colon t \in R^1,\ \rho(t, \mathbf{B}) < \delta\}$ egy ilyen $\mathbf{B}$ halmaz δ sugarú környezetét. Lássuk be, hogy tetszőleges $\varepsilon > 0$-ra létezik olyan $\delta > 0$, hogy a $\mathbf{K}^\varepsilon = \{t\colon \mu(\mathbf{B}^\delta - t) > \varepsilon\}$ halmazra $\mu(\mathbf{K}^\varepsilon) < \varepsilon$, sőt ennek δ sugarú környezetére is $\mu\left((\mathbf{K}^e)^\delta\right) < \varepsilon$.

3.) Legyen $t \in \mathbf{K}^\varepsilon$, ahol $\mathbf{K}^\varepsilon$-t az előző feladatban definiáltuk, $t' < \left[t - \dfrac{\delta}{2}, t + \dfrac{\delta}{2}\right]$. Ekkor a $\nu_n(t') = \#\{l\colon 1 \le l \le n,\ \xi_k + t' \in \mathbf{B}^{\delta/2}\}$ változóra $P(\nu_n(t') > n|\log\varepsilon|^{-1}) < 2e^{-n/\varepsilon}$, mivel $\mu(\mathbf{B}^{\delta/2} - t') < \varepsilon$. Ha $t' \in \left[-\dfrac{\delta}{2}, \dfrac{\delta}{2}\right]$, akkor mivel $\mu\left(\mathbf{B}^\delta + t'\right) \ge \dfrac{c}{2}$ $P(\nu_n(t') < n|\log\varepsilon|^{-1}) \le e^{-n\alpha}$.

A fenti észrevételek segítségével oldjuk meg a Schweitzer feladat első felét.

Segítség: Azon a már lokalizált intervallumon, ahol ϑ nagy valószínűséggel található tekintsünk egy sűrű rácsot, és ha t eleme a rácsnak, nézzük meg, tartalmaz-e a $\mathbf{A}^\delta - t$ halmaz sok ξ_k pontot. Vegyük az első ilyen pontot.

Tekintsük a feladat negatív részét, annak bizonyítását, hogy amennyiben az $F(x)$ eloszlásfüggvénynek van folytonos deriváltja (folytonos sűrűségfüggvénye), akkor nem lehet olyan jó becslést adni ϑ-ra, mint a nem abszolut folytonos esetben. Azt kell megfogalmazni és bebizonyítani, hogy ebben az esetben az F eloszlásfüggvényből, illetve annak kis eltoltjaiból származó minták nem különböztethetőek meg nagyon.

4.) Bizonyítsuk be, hogy ha F-nek van folytonos $F'(x)$ sűrűségfüggvénye, akkor tetszőleges $\delta > 0$-ra van olyan $L = L(\delta)$ szám, melyre a $\mathbf{B}(L) = \left\{t\colon |t| < L, F'(t) \ge \dfrac{1}{L}\right\}$ halmaz F eloszlás szerint meghatározott μ_F mértéke nagyobb mint $1 - \delta$. Legyen $\mathbf{A} \subset R^n$ olyan halmaz, amelyikre $\mathbf{A} \subset \mathbf{B}(L) \times \cdots \times \mathbf{B}(L)$. Jelölje $\mu_{F_\vartheta^{(n)}}$ az $F(x - \vartheta)$ eloszlás n-szeres direkt szorzata által meghatározott mértéket. Bizonyítsuk be, hogy $\mu_{F_\vartheta^{(n)}}(\mathbf{A}) \ge e^{-n\delta}\mu_{F_0^{(n)}}(\mathbf{A})$.

5.) Bizonyítsuk be az előző feladat segítségével a Schweitzer feladat negatív felének következő gyengített változatát: Ha az F eloszlásnak létezik folytonos sűrűség-

függvénye, akkor rögzített $\alpha > 0$-ra és nagyon kicsi $\varepsilon = \varepsilon(\alpha) > 0$-ra nincs olyan $T_n(x_1, \dots, x_n)$ becslés a ϑ paraméterre, melyre

$$P(|T_n(\xi_1, \dots, \xi_n) - \vartheta| < \varepsilon) > e^{-n\alpha}.$$

(Azaz az $E = (-\varepsilon, \varepsilon)$ választással nem lehet a feladat állítását teljesíteni.)

Az előző feladat bizonyítása azon alapult, hogy elég kis ε-ra nem lehet két diszjunkt halmazt találni ($\{(x_1, \dots, x_n)\colon |T_n(x_1, \dots, x_n)| < \varepsilon\}$ és $\{(x_1, \dots, x_n)\colon |T_n(x_1, \dots, x_n) - 2\varepsilon| < \varepsilon\}$ halmazok) az R^n térben, melyeknek az $\mu_{F_0^{(n)}}$ illetve a hozzá közel levő $\mu_{F_\varepsilon^{(n)}}$ szerinti mértéke nagy. A feladat bizonyításához ezt az érvet finomítani kell. Azt kell kihasználni, hogy ha az E halmaz mértéke kicsi, akkor a $\{|T_n(\xi_1, \dots, \xi_n) - \vartheta| \in E\}$ halmazoknak nem lehet nagy metszete minden különböző ϑ_1, ϑ_2 párra. Ezt az állítást fogjuk pontosabban megfogalmazni.

6.) Definiáljunk egy ν mértéket a számegyenesen, mint a következő feltételes eloszlást.

$$\nu(\mathbf{A}) = \mu_{F_0^{(n)}}\left(\{(x_1, \dots, x_n)\colon T_n(x_1, \dots, x_n) \in \mathbf{A}\} \,|\, x_j \in \mathbf{B}(L),\ 1 \le L \le n\right).$$

Ekkor $\nu(E) > \dfrac{1}{2}$ elég nagy n-re, ahol E a Schweitzer feladatban szereplő E halmaz, ha $\varepsilon > 0$ elég kicsi, az L konstans pedig elég nagy a $\mathbf{B}(L)$ halmaz definiciójában.

Tetszőleges $\eta > 0$-ra és $\varepsilon = \varepsilon(\eta) > 0$ küszöbindexre igaz, hogy ha ν olyan mérték a számegyenesen, melyre $\nu(E) > \dfrac{1}{2}$, és $\lambda(E) < \varepsilon$, akkor létezik $0 \le t \le \eta$, melyre $\nu(E \setminus (E + t)) > \dfrac{1}{4}$.

6.) Bizonyítsuk be a Schweitzer verseny állításának negatív felét is.