Eloszlásfüggvény becslése cenzorált minta segítségével.

A jó becslés megtalálása.

Az empirikus eloszlásfüggvény becslése cenzorált minta alapján a következő feladatot jelenti.

Legyen $X_1, \dots, X_n$ független, egyforma eloszlású valószínűségi változók sorozata F eloszlással, $Z_1, \dots, Z_n$ független, egyforma eloszlású valószínűségi változók sorozata G eloszlással, és legyen az $X_1, \dots, X_n$ és $Z_1, \dots, Z_n$ sorozat egymástól független.

Nem ismerjük sem az F sem a G eloszlásfüggvényt. Az F eloszlásfüggvényt kívánjuk megbecsülni, de csak az $Y_i = \min(X_i, Z_i)$ valószínűségi változókat, és a $\delta_i = I(X_i \le Z_i)$ eseményeket, $1 \le i \le n$, tudjuk megfigyelni, azaz azt, hogy a megfigyelt Y_i valószínűségi változó X_i-vel vagy Z_i-vel egyenlő-e.

Kaplan és Meyer a következő becslést javasolta az $S(u) = 1 - F(u)$ függvényre:

$$1 - F_n(u) = S_n(u) = \begin{cases} \displaystyle\prod_{i=1}^{n} \left(\frac{N(Y_i)}{N(Y_i)+1} \right)^{I(Y_i \le u, \delta_i = 1)} & \text{ha } u \le \max(Y_1, \dots, Y_n) \\ 0 & \text{ha } u \ge \max(Y_1, \dots, Y_n),\ \bar{\delta}_n = 1, \\ \text{nem definiált} & \text{ha } u \ge \max(Y_1, \dots, Y_n),\ \bar{\delta}_n = 0, \end{cases}$$

ahol

$$N(t) = \#\{Y_i,\ \ Y_i > t,\ 1 \le i \le n\} = \sum_{i=1}^{n} I(Y_i > t),$$

és $\bar{\delta}_n = \bar{\delta}_{i(n)}$, ahol az $i(n)$ indexet az $Y_{i(n)} = \max\limits_{1 \le i \le n} Y_i$ képlet definiálja. Az üres szorzat értékét (azaz az $S_n(u)$ kifejezés értékét abban az esetben, ha nincs az $Y_i \le u,\ \delta_i = 1$ feltételt kielégítő mintaelem) 1-nek definiáljuk.

Azt kívánjuk megmutatni, hogy ez a becslés maximum likelihood tipusú becslés a következő értelemben. Becsüljük meg az (F, G) eloszláspárt a következő módon. Legyenek $x_1, \dots, x_k$, $x_1 < x_2 < \dots < x_k$, azon Y_i valószínűségi változók értékei, amelyek valamelyik X_i változóval egyenlőek, azaz minden x_s, $1 \le s \le k$, ponthoz létezik olyan i index, amelyre $X_i = x_s$, és $\delta_i = 1$, és legyenek $z_1, \dots, z_l$, $z_1 < z_2 < \dots < z_l$, azon Y_i valószínűségi változók értékei, amelyek olyan Y_i változók értékei, amelyekre $Y_i = Z_i$, azaz $\delta_i = 0$. Feltesszük, hogy mind az F mind a G eloszlás atommentes, ezért nem jelenik meg ugyanaz a mintaelem kétszer. Ha $x_k > y_l$, akkor bevezetünk egy 'fiktív' $y_{l+1} > x_k$, ha $y_l > x_k$ akkor bevezetünk egy 'fiktív' $x_{k+1} > y_l$ elemet. Ennek értékét tetszőlegesen választjuk. Vegyük az összes olyan (F, G) eloszláspárt, amelyre F az x_i, $1 \le i \le k$, (vagy $1 \le i \le k+1$), G a z_j, $1 \le j \le l$, vagy $(1 \le j \le l+1)$ pontokba van koncentrálva. Számoljuk ki, hogy ezen eloszláspárok közül melyikre lesz a legnagyobb annak a valószínűsége, hogy az általunk megfigyelt sorozat jelenik meg. Azt állítom, hogy ez a feladat megoldható, és a megoldás az F eloszlásra a Kaplan és Meyer által javasolt képletet adja. A G eloszlásra egy hasonló képletet kapunk, de azt nem írom fel. A fiktív x_{k+1} vagy z_{l+1} értéket azért vezettük be, mert ha x_k a legnagyobb megfigyelt érték, akkor Z_1 pozitív valószínűséggel felvesz ennél nagyobb értékeket is, és hasonló argumentum érvényes akkor, ha y_l a legnagyobb megfigyelt érték.

Legyen $p_i = P(X_1 = x_i)$, $q_j = P(Z_1 = z_j)$, és számoljuk ki annak valószínűségét, hogy ilyen eloszlás esetén az általunk bevezetett esemény következik be. Vezessük be a $P_i = p_i + \cdots + p_k$, $1 \le i \le k$, számokat, ha nincs x_{k+1} elem, és a $P_i = p_i + \cdots + p_{k+1}$, $1 \le i \le k+1$, számokat, ha van x_{k+1} elem. Hasonlóan, legyen $Q_j = q_j + \cdots + q_l$, $1 \le j \le l$, ha nincs z_{l+1} elem, és a $Q_j = q_j + \cdots + q_{l+1}$, $1 \le j \le l+1$, ha van z_{l+1} elem. Teljesül a $P_1 = Q_1 = 1$ azonosság. Írjuk fel a minket érdeklő valószínűséget.

Annak a valószínűsége, hogy egy mérésnek x_i az eredménye $p_i Q_{r(i)}$, $1 \le i \le k$, és annak valószínűsége, hogy az z_j-vel egyenlő $q_j P_{s(j)}$, $1 \le j \le l$, ahol az $r(i)$ és $s(j)$ indexeket a $z_{r(i)-1} < x_j < z_{r(i)}$ és $x_{s(j)-1} < z_j < x_{s(j)}$ képletek határozzák meg. (Bevezetjük az $x_0 = z_0 = -\infty$ konvenciót.) Ennek alapján a keresett valószínűség

$$P = P(p_,p_2, \dots, q_1, q_2, \dots) = \prod_{i=1}^{k} p_i Q_{r(i)} \cdot \prod_{j=1}^{l} q_j P_{s(j)}.$$

Végezzük el a $p_i = P_i - P_{i+1}$, $1 \le i \le k$, és $q_j = Q_j - Q_{j+1}$, $1 \le j \le l$, helyettesítéseket a fenti képletben, ($P_{k+1} = 0$, ha nem létezik x_{k+1}, és $Q_{l+1} = 0$, ha nem étezik z_{l+1} elem), és rendezzük át a fenti egyenlet jobboldalát, mint két olyan kifejezés szorzatát, amelyek közül az első kifejezés egy olyan szorzat, amelyben csak P_i-től függő, a második kifejezés olyan szorzat, amelyben csak Q_j-től függő tagok vannak. Vezessük be a következő jelöléseket: $\alpha(i)$ a z_j pontok száma az (x_i, x_{i+1}) intervallumban, $\beta(j)$ az x_i pontok száma a (z_j, z_{j+1}) intervallumban. Ezzel a jelöléssel azt kapjuk, hogy

$$P = I_1 \cdot I_2,$$

ahol

$$I_1 = \prod_{i=1}^{k-1} (P_i - P_{i+1}) P_{i+1}^{\alpha(i)} \cdot P_k = \prod_{i=1}^{k-2} (P_i - P_{i+1}) P_{i+1}^{\alpha(i)} \cdot (P_{k-1} - P_k) P_k^{(\alpha(k-1)+1)},$$

ha x_{k+1} nem létezik, és

$$I_1 = \prod_{i=1}^{k} (P_i - P_{i+1}) P_{i+1}^{\alpha(i)},$$

ha x_{k+1} létezik. Hasonlóan

$$I_2 = \prod_{j=1}^{l-2} (Q_j - Q_{j+1}) Q_{j+1}^{\beta(j)} \cdot (Q_{l-1} - Q_l) Q_l^{(\beta(l)+1)},$$

ha z_{l+1} nem létezik, és

$$I_2 = \prod_{j=1}^{l} (Q_j - Q_{j+1}) Q_{j+1}^{\beta(j)},$$

ha z_{l+1} létezik.

Ennek alapján a keresett becslés megtalálásának érdekében a következő szélsőérték feladatot kell megoldanunk. Találjuk, meg milyen $(P_1, \dots, P_k)$ (vagy $(P_1, \dots, P_{k+1})$) vektorra veszi fel az I_1 kifejezés a maximumát az $1 = P_1 \geq P_2 \geq \dots \geq P_k \geq 0$ (illetve $1 = P_1 \geq P_2 \geq \dots \geq P_{k+1} \geq 0$) feltétel mellett.

A feladat megoldásában a következő lemmát fogjuk használni.

Lemma. *A $(P-x)x^\alpha$, $0 \leq x \leq P$, $\alpha \geq 1$, kifejezés az $x = \frac{\alpha}{\alpha+1}P$ pontban veszi fel a maximumát, és ott értéke $\frac{1}{\alpha+1}P^{\alpha+1}$.*

Bizonyítás. A $(P-x)x^\alpha$ függvény deriváltja $\alpha(P-x)x^{\alpha-1} - x^\alpha = x^{\alpha-1}[\alpha(P-x) - x]$. Ez az $x_0 = \frac{\alpha}{\alpha+1}$ pontban egyenlő nullával. Az x_0 pont a függvény maximumhelye, és a függvény értéke ebben a pontban $\frac{1}{\alpha+1}P^{\alpha+1}$.

Számoljuk ki I_1 maximumának az értékét először abban az esetben, ha x_{k+1} nem létezik. Ha a $P_1, \dots, P_{k-1}$ pontokat rögzítjük, és keressük azt a P_k számot, amelyre I_1 értéke maximális, azt kapjuk (a lemma segítségével), hogy $P_k = \frac{\alpha(k-1)+1}{\alpha(k-1)+2}P_{k-1}$, és

$$I_1 = \text{const.} \prod_{i=1}^{k-2} (P_i - P_{i+1}) P_{i+1}^{\alpha(i)} \cdot P_{k-1}^{\alpha(k-2)+\alpha(k-1)+2}$$

egy olyan konstanssal, amely csak az $\alpha(\cdot)$ számoktól függ, de nem függ a P_i számok választásától. Ezután a P_{k-1} értékét választjuk meg rögzített P_i, $1 \leq i \leq k-2$ értékek mellett úgy, hogy I_1 értéke maximális legyen. Azt kapjuk (a lemma segítségével), hogy $P_{k-1} = \frac{\alpha(k-2)+\alpha(k-1)+2}{\alpha(k-2)+\alpha(k-1)+3}P_{k-2}$, és

$$I_1 = \text{const.} \prod_{i=1}^{k-3} (P_i - P_{i+1}) P_{i+1}^{\alpha(i)} \cdot P_{k-2}^{\alpha(k-3)+\alpha(k-2)+\alpha(k-1)+3}.$$

Indukcióval kapjuk, hogy tetszőleges i indexre $2 \leq i \leq k$, az I_1 maximumát nyújtó vektor P_i koordinátájának a választása a következő.

$$P_i = \frac{\alpha(i-1) + \dots + \alpha(k-1) + k + 1 - i}{\alpha(i-1) + \dots + \alpha(k-1) + k + 2 - i} P_{i-1},$$

és

$$I_1 = \text{const.} \prod_{s=1}^{k-i-1} (P_s - P_{s+1}) P_{s+1}^{\alpha(s)} \cdot P_{i-1}^{\alpha(i-2)+\dots+\alpha(k-1)+k-i+2}.$$

Vegyük észre, hogy

$$\alpha(i-1) + \dots + \alpha(k-1) + k + 1 - i = N(x_{i-1})$$

(azzal az $N(\cdot)$ kifejezéssel, amelyik az $1 - F_n(u)$ becslőfüggvény definiciójában szerepel), mert $\alpha(i-1) + \dots + \alpha(k-1)$ darab x_{i-1}-nél nagyobb z_j és $k+1-i$ darab x_{i-1}-nél

nagyobb $x_{i'}$ szám van. Ezért a szélsőérték feladatot megoldó P_i számokra felírhatjuk, hogy

$$\frac{P_i}{P_{i-1}} = \frac{N(x_{i-1})}{N(x_{i-1})+1}, \quad 2 \le i \le k. \qquad (*)$$

Másrészt $P_1 = 1$. Innen azt kapjuk, hogy

$$P_i = \prod_{s=2}^{i} \frac{N(x_{s-1})}{N(x_{s-1})+1}, \quad 2 \le i \le k.$$

Ez tekinthető úgy, mint a Kaplan és Meyer által javasolt becslés felírása más alakban, mert $S_n(u) = 1 - F_n(u) = P_i$, ha $x_{i-1} \le u < x_i$. (Itt megint az $x_0 = -\infty$ konvenciót használjuk.)

Az az eset, amikor létezik x_{k+1} pont hasonlóan tárgyalható. De ilyenkor létezik a P_{k+1} valószínűség, és az $\alpha(k) \ge 1$ szám is. Az előző esethez hasonlóan számolva azt kapjuk, hogy $P_{k+1} = \frac{\alpha(k)}{\alpha(k)+1} P_k$, és az I_1 kifejezés is kiszámolható ezzel a P_{k+1} választással. Ezután indukcióval kiszámítjuk a P_{k-s} valószínűségek optimális választását és a hozzájuk tartozó I_1 kifejezés értékét. Azt kapjuk, hogy

$$P_i = \frac{\alpha(i-1) + \cdots + \alpha(k) + k + 1 - i}{\alpha(i-1) + \cdots + \alpha(k) + k + 2 - i} P_{i-1},$$

minden $2 \le i \le k$ indexre, (és az I_1 kifejezésre is felírhatunk az előző esethez hasonló formulát). Ezért a $(*)$ formula ekkor is érvényes. Következésképp most is a Kaplan és Meyer által javasolt formulát kapjuk, mint az itt tekintett maximumhely keresése feladat megoldását.