Μη Παραμετρική | Διάλεξη 1 (1)

Ύλη Μαθήματος

(1) Εμπειρική προσέγγιση στην εκτίμηση

σ.κ. $F \in \mathcal{F} \rightarrow$ χώρος σ.κ.

$X_1, X_2, \ldots, X_n$ τ.δ. από F

Πώς εκτιμάμε την F ? Λωρίδες Εμπιστοσύνης ?

(2) Εμπειρική Διαδικασία (Ανέλιξη)

Σύνδεση με στοχ. διαδ. + θέματα σύγκλισης

(3) Έλεγχος X^2

κατασκευή και διάφορες μορφές του ελέγχου

(4) Κλασικοί Απαραμ. Έλεγχοι Υποθ. (K.S. Mann-Whitney, Shapiro-Wilk...)

(5) Μη παραμετρική εκτίμηση πυκνότητας

ιστογράμματα + πυρήνες πυκνοτήτων

(6) Bootstrap

(7) Μη παραμετρική παλινδρόμηση ?

Βιβλιογραφία

(1) Wasserman, L. (2006) All of NonParametric Statistics

(2) Tsybakov A.B. Introduction to Nonparametric Estimation

+ άλλα + άρθρα.

(3) Σημειώσεις Καθηγητή

~~Εισαγωγικά~~.

Βαθμός:

Ασκήσεις 2/10, Εργασία 3/10, Τελική Εξέταση 5/10 + Συμμετοχή.

[πίνακας + προτζέκτορας (R) + e-class].

Εισαγωγικά

Γιατί μη παραμετρική προσέγγιση ;

(*) Αρχική προσέγγιση σε κάποιο πρόβλημα όταν δε γνωρίζουμε καθόλου τη φύση της άγνωστης κατανομής

(*) Ύπαρξη πυκνότητας αρκετά περίπλοκης μορφής (πολυκόρυφη κ.τ.λ.) οπότε δουλεύουμε με οικογ. πυκνοτήτων με κατάλ. Σ.Ο.

(*) Κατασκευή εργαλείων για να ελέγχουμε παραμετρικά μοντέλα σε ένα ευρύτερο πλαίσιο.

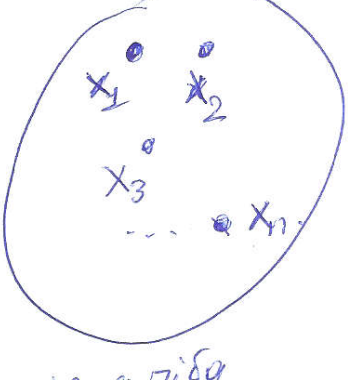
3.1. $x_1, x_2, \ldots, x_n$ δεδομένα

Υποθέτουμε ότι αποτελούν πραγματοποιήσεις τ.μ.

$$X_1, X_2, \ldots, X_n \sim F \quad + \text{ ανεξαρτησία}$$

Θέλουμε να "μάθουμε" για την $F \in \mathcal{F}$ (όλες οι σ.κ. πραγμ. τ.μ.)

Δημιουργούμε ένα διακριτό μηχανισμό εκμάθησης



κληρωτίδα

$$X_n^* = \begin{cases} x_1, & \text{με πιθ. } 1/n \\ x_2 & \text{"} \\ \vdots & \\ x_n & \text{"} \end{cases}$$

$$X_n^* \sim \text{discr. Unif}(\{x_1, x_2, \ldots, x_n\}).$$

Βασική Ιδέα

Η $X_n^* \approx F$ για μεγάλο n.
(ακολουθεί προσεγγιστικά).

Αν F_n^* είναι η σ.κ. της X_n^*, τότε

$$F_n^*(x) = P_{F}\left(X_n^* \leq x\right) = \sum_{k=1}^{n} \underbrace{P\left(X_n^* = x_k\right)}_{\frac{1}{n}} \cdot 1_{(-\infty, x]}(x_k)$$

$$= \frac{\sum_{k=1}^{n} 1_{(-\infty, x]}(x_k)}{n} \quad \left(= \text{ποσοστό των παρατηρήσεων } \{x_1, \dots, x_n\} \leq x \right).$$

↳ Τυχαιότητα

$$x_k \to X_k \Rightarrow F_n^* \to \mathbb{F}_n =: \underline{\mathbb{F}_{X_1, X_2, \dots, X_n}}.$$

Ορσ: Εμπειρική συνάρτηση κατανομής

$$\mathbb{F}_n(x) = \frac{\sum_{k=1}^{n} 1_{(-\infty, x]}(X_k)}{n}, \quad \forall x \in \mathbb{R}.$$

Προσοχή! Η ε.σ.κ. είναι τυχαία συνάρτηση του x που έχει πραγματοποιήσεις που είναι σ.κ.

μία τ.μ. $X : \Omega \to \mathbb{R}$ → πραγμ. τυχ. μεταβλ.

$\mathbb{F}_n : \Omega \to \mathcal{F}$ → σ.κ.

3.2 Σημειακές Ιδιότητες της ε.σ.κ.

Πρότ. 3.1 $\boxed{\forall x \in \mathbb{R}, \quad \mathbb{F}_n(x) \sim \frac{1}{n} Bin(n, F(x))}$

Αποδ. Έστω $x \in \mathbb{R}$.

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} 1_{(-\infty, x]}(X_k) = \frac{1}{n} \sum_{k=1}^{n} Y_k,$$

όπου $Y_k = \underbrace{1_{(-\infty, x]}(X_k)}_{g(X_k)}, \quad 1 \leq k \leq n$

$\{Y_k\}_{1 \leq k \leq n}$ είναι ανεξ + ισον. τ.μ., με

$Y_k \sim Be(p)$, όπου $p = P_F(Y_k = 1) = P_F(X_k \leq x) = F(x)$.

$$\Rightarrow \sum_{k=1}^{n} Y_k \underset{\text{ανεξ. } Be(F(x))}{\sim} Bin(n, F(x)).$$

Πόρισμα

$$E_F\left(\mathbb{F}_n(x)\right) = F(x), \quad \forall x \in \mathbb{R}$$

$$V_F\left(\mathbb{F}_n(x)\right) = \frac{F(x)\left(1-F(x)\right)}{n}, \quad \forall x \in \mathbb{R}.$$

Σχ. Η $\mathbb{F}_n(x)$ είναι α.ε. (αμερόληπτη εκτιμήτρια) του $F(x)$ ✓

και ~~...~~ συνεπής. Πράγματι, $V_F\left(\mathbb{F}_n(x)\right) \xrightarrow[n\to\infty]{} 0$.

[Μια α.ε. είναι συνεπής αν $V(\hat{\theta}_n) \longrightarrow 0$
Ικανή συνθήκη: $MT\Sigma(\hat{\theta}_n) \longrightarrow 0 \iff$ ασυμπτωτ. αμερολ. + $V(\hat{\theta}_n) \longrightarrow 0$.]

$$MT\Sigma(\hat{\theta}_n) = b^2(\hat{\theta}_n) + V(\hat{\theta}_n).$$

Πρότ:

$$\mathbb{F}_n(x) \xrightarrow{P_F - a.s.} F(x), \quad \forall x \in \mathbb{R}$$

Λέμε ότι η ε.σ.κ. είναι σημειακά ισχυρά συνεπής

Αποδ: Έστω $x \in \mathbb{R}$.

Έχουμε $\{Y_n\}_{n\geq 1}$ ~~...~~ ανεξ. α.ι.τ.μ. με $Y_n = \mathbb{1}_{(-\infty, x]}(X_k)$

$$Y_n \sim Be(F(x)), \quad n \geq 1$$

$$\mathbb{F}_n(x) = \frac{\sum_{k=1}^{n} \mathbb{1}_{(-\infty, x]}(X_k)}{n} = \frac{\sum_{k=1}^{n} Y_k}{n} = \overline{Y_n}.$$

Απο I.N.M.A.

$$\mathbb{F}_n(x) = \overline{Y_n} \xrightarrow{P_F - a.s.} E_F(Y_1) = F(x).$$

Πρότ.

$$\sqrt{n}\left(\mathbb{F}_n(x) - F(x)\right) \xrightarrow{d(F)} N\left(0, F(x)(1-F(x))\right)$$

Λέμε ότι η ε.σ.κ. είναι σημειακά ασυμπτωτικά κανονική

Αποδ.

$(Y_n)_{n\geq 1}$ όπως πριν. Από κλασικό Κ.Ο.Θ. (ισχύουν οι προϋποθ).

$$\sqrt{n}\left(\mathbb{F}_n(x) - F(x)\right) = \sqrt{n}\left(\overline{Y_n} - E_F(Y_1)\right) \xrightarrow{d(F)} N\left(0, V_F(Y_1)\right)$$
$$= N\left(0, F(x)(1-F(x))\right).$$

Διαστήματα Εμπιστοσύνης

Θέτουμε $p = F(x)$.

Η εύρεση Δ.Ε. για το $F(x)$ ανάγεται σε Δ.Ε. για την παράμετρο p μιας $Be(p)$ (παραμετρικό πρόβλημο).

Υπάρχουν πολλές μέθοδοι, βλ. Agresti & Coull (1998).

Θα δούμε 3 μεθόδους.

Εκτιμήτρια: $\hat{p}_n = \frac{\text{\# μονάδων}}{n}$ ποσοστό στο δείγμα

$\hat{p}_n = \bar{Y}_n$

(1) Wald (ασυμπτωτικό)

$$\boxed{I^{\infty,1}_{1-a} = \hat{p}_n \pm Z_{a/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}} \cap [0,1].$$

$$\sqrt{n}\,(\hat{p}_n - p) = \sqrt{n}\left(\bar{Y}_n - E(Y_1)\right) \xrightarrow[d]{\text{Κ.Ο.Θ.}} N(0, \underbrace{V(Y_1)}_{p(1-p)}) = N(0, p(1-p)).$$

$$\Rightarrow \underbrace{\sqrt{n}\,\frac{(\hat{p}_n - p)}{\sqrt{p(1-p)}}}\ \xrightarrow[n\to\infty]{d} N(0,1).$$

$$\overset{\text{Slutsky}}{\Rightarrow} \sqrt{n}\,\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{d} N(0,1).$$

$$\Rightarrow \lim_n P\left(\left|\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}\right| \le Z_{a/2}\right) = P(|\overset{N(0,1)}{\tilde{Z}}| \le Z_{a/2}) = 1-a.$$

Άρα για n μεγάλο,
$$P\left(\left|\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}\right| \le Z_{a/2}\right) \cong 1-a$$

και έτσι λύναμε
$$\left|\frac{\hat{p}_n - p}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}\right| \le Z_{a/2} \Leftrightarrow p \in I^{\infty,1}_{1-a}.$$

(2) Wilson (ασυμπτωτικό).

Είδαμε $\sqrt{n}\,\dfrac{\hat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1)$

παρόμοια $\left|\dfrac{\hat{p}_n - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq Z_{a/2}$ έχει πιθαν. $\to 1-a$

Κοιτάμε αν μπορεί να λυθεί ως προς p.

$$\Leftrightarrow \frac{(p-\hat{p}_n)^2}{\frac{p(1-p)}{n}} \leq Z_{a/2}^2 \Leftrightarrow (p-\hat{p}_n)^2 \leq \frac{Z_{a/2}^2}{n} p(1-p)$$

$$p^2 - 2\hat{p}_n p + (\hat{p}_n)^2 - \frac{Z_{a/2}^2}{n} p + \frac{Z_{a/2}^2}{n} p^2 \leq 0$$

$$\Leftrightarrow \boxed{\underbrace{\left(1+\frac{Z_{a/2}^2}{n}\right)}_{>0} p^2 - \left(2\hat{p}_n + \frac{Z_{a/2}^2}{n}\right) p + (\hat{p}_n)^2 \leq 0}$$

$$Ap^2 + Bp + \Gamma \leq 0.$$

$$p_{1,2} \quad \frac{2\hat{p}_n + \frac{Z_{a/2}^2}{n} \pm \sqrt{\left(2\hat{p}_n + \frac{Z_{a/2}^2}{n}\right)^2 - 4\left(1+\frac{Z_{a/2}^2}{n}\right)(\hat{p}_n)^2}}{2\left(1+\frac{Z_{a/2}^2}{n}\right)}$$

$$I_{1-a}^{\infty,2}$$

$\downarrow$

$$I_{1-a}^{Wilson} = \frac{\hat{p}_n + \frac{Z_{a/2}^2}{2n} \pm Z_{a/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n) + \frac{Z_{a/2}^2}{4n}}{n}}}{1+\frac{Z_{a/2}^2}{n}}$$

(3) $\boxed{\text{exact}}$ $\to$ ακριβές