3-4-2023

Eclass: σημειώσεις – R codes

G|G|1 simulation
Παραδείγματα
+ G|G|m

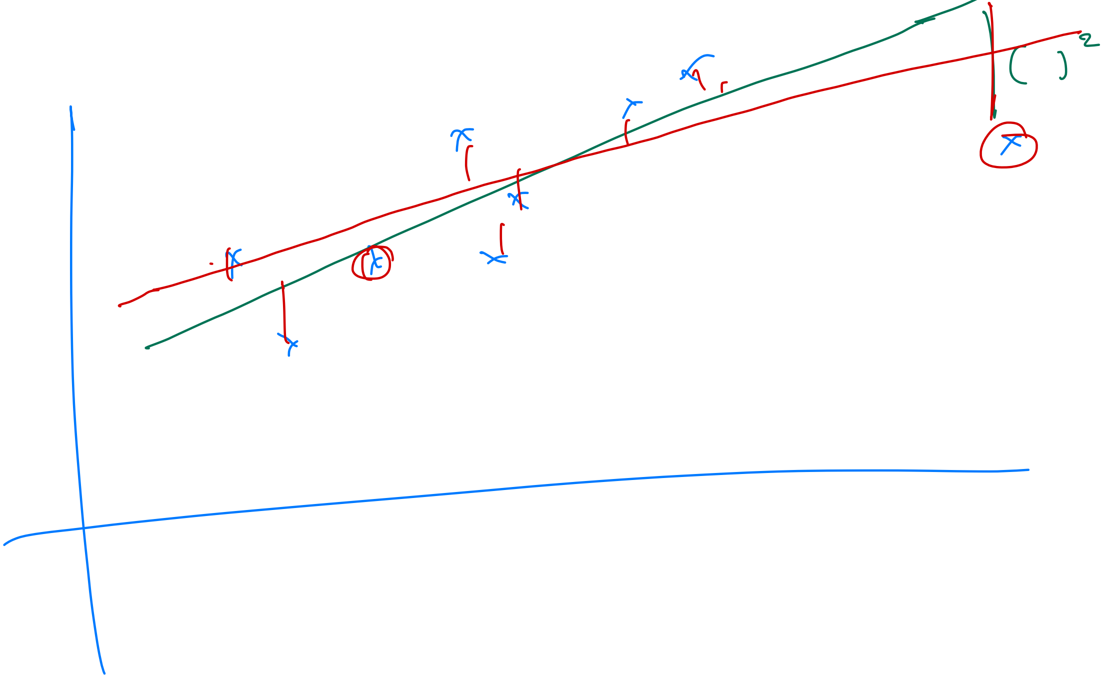
αρχείο ~~σε~~ σε σειρά ασκήσεων.

Στατιστική συμπερασματολογία / εκτιμήσεις από προσομοιωμένα δεδομένα.

1. Εκτίμηση MSE με αμερόληπτης εκτιμήτριας μέσω bootstrapping.

2. Μέθοδοι ελάττωσης διασποράς (variance reduction).

Bootstrap (Resampling methods)
(bootstrap - jacknife)



Bootstrap

$X_1, \dots, X_n$ iid $\sim F$

άγνωστη ποσότητα $\theta(F)$

Αν υπάρχει $Y = h(X)$: $E(Y) = \theta \Rightarrow$

$$(X_1, \dots, X_n) \to (Y_1, \dots, Y_n) \to \left.\begin{array}{l} \bar{Y}_n, \; S_{Y_n} \\ \downarrow \\ E(\bar{Y}_n) = \theta \end{array}\right\} \Rightarrow \text{ΔΕ}$$

$$Y_j = h(X_j)$$

π.χ. αν θ = ποσοστημόριο της F $= F^{-1}(\alpha)$

$$\theta = \frac{E(Y)}{E(X)}$$

ή $\theta = \frac{1}{E(X)}$ (π.χ. $X \sim Exp(\lambda)$, $\lambda = \frac{1}{E(X)} = \theta(F)$)

Έστω $\hat{\theta} = g(X_1, \dots, X_n)$

$E(\hat{\theta}) \neq \theta$: bias $(E(\hat{\theta}) - \theta)$

$\widehat{MSE}$ $\qquad MSE = E_F(\underset{\sim}{\hat{\theta}} - \theta)^2 = E_F\left(g(X_1, \dots, X_n) - \theta\right)^2$

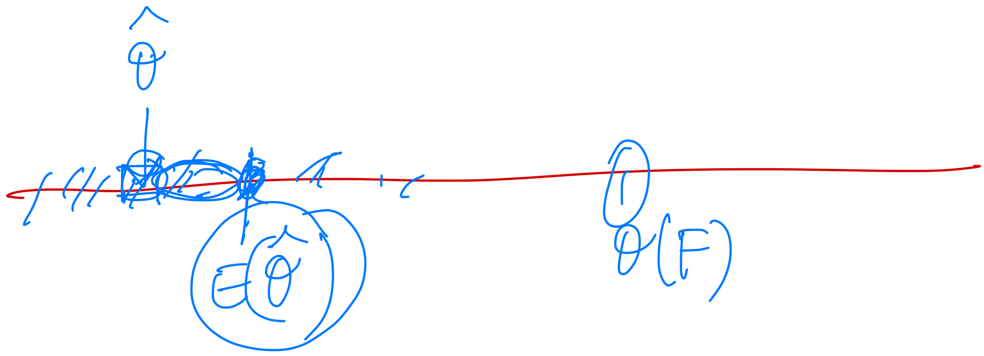
$\theta(\hat{F})$

$$= E\Big(\underbrace{\hat{\theta} - E(\hat{\theta})} + \underbrace{E(\hat{\theta}) - \theta}_{\text{bias}}\Big)^2 =$$

$$= E\left(\hat{\theta} - E(\hat{\theta})\right)^2 + \left(E(\hat{\theta}) - \theta\right)^2 +$$

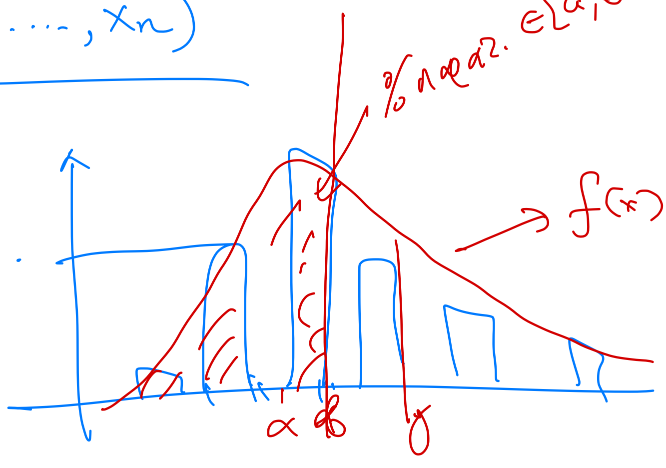
$$\cancel{2 \cdot (E(\hat{\theta}) - \theta) \cdot E}\underbrace{\left(\hat{\theta} - E(\hat{\theta})\right)}_{= 0}$$

$$\boxed{MSE(\theta) = Var(\hat{\theta}) + Bias(\hat{\theta})^2}$$



Δείγμα $(x_1, \dots, x_n)$

histogram
% συχνοτήτων



Από δείγμα → εκτίμηση της f η F $=\hat{F}$

Αν $\hat{F}_n \to F$

$\theta(F)$: "συνεχής"

$\Rightarrow \boxed{\theta(\hat{F})} \to \theta(F)$

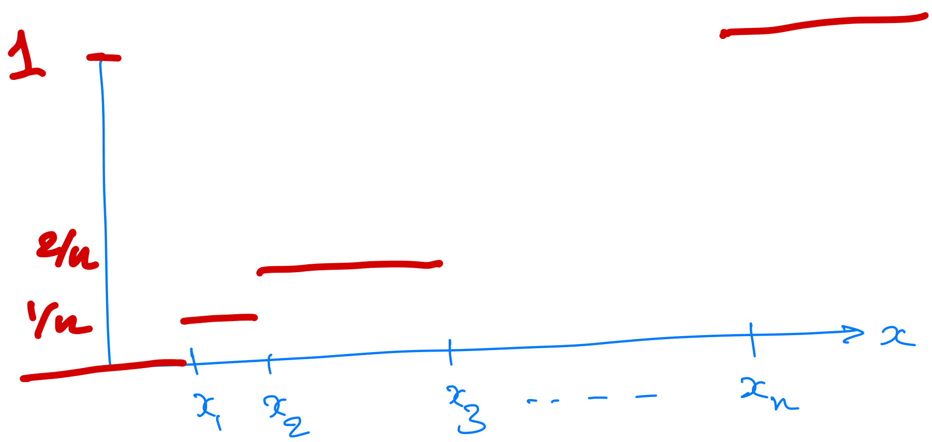
"προσομοίωση" μέσω bootstrap

Εμπειρική κατανομή δείγματος $F_e(x; x_1, \dots, x_n)$

$$\underbrace{F_e(x; x_1, \dots, x_n)}_{\text{τυχαία συνάρτηση}} = \frac{1}{n} \sum_{j=1}^{n} 1(x_j \le x)$$

$=$ % παρατ. του δείγματος $\le x$

$x \in \mathbb{R}$



Όταν $n \to \infty$ $F^e(x) \to F(x)$ ομοιόμορφα ως προς x με πιθ. 1

(Glivenko–Cantelli)

$\theta(F_e) \to \theta(F)$

π.χ. αν $n = 1000$

$\theta(F) = 95\%$ ποσοστό της F

$\theta(F_e) = ?$ 95% ποσοστό της F_e.

$n = 1000 \Rightarrow \boxed{\theta(F_e) = x_{950}}$

Επομένως μια εκτίμηση για $MSE(F) = E_F\left[(g - \theta(F))^2\right]$

είναι $MSE(F_e) = E_{F_e}\left[\left(\underbrace{g(x_1, \dots, x_n)}_{\hat{\theta}} - \theta(F_e)\right)^2\right]$

$\theta(F_e) = ?$ ← εκτίμηση μέσω προσομοίωσης

Έστω δείγμα $(x_1, \dots, x_n)$

$$F_e(x) = \begin{cases} 1/n, & x \le x_1 \\ 2/n, & x_1 < x \le x_2 \\ \vdots \\ 1, & x \ge x_n \end{cases}$$

← διακριτή ομοιόμορφη στο $\{x_1, \dots, x_n\}$

Ποια είναι η γεννήτρια?

Δημιουργία τυχαίου αρ. από $\{x_1, \dots, x_n\}$ 100 πιθανά

Εκτίμηση $\theta(F_e)$ από δείγμα $(x_1, \dots, x_n)$

$\vdots$

δημιουργώ νέο δείγμα μεγέθους N

N παρατηρήσεις από την διακρ. ομοιόμορφη $\{x_1, \dots, x_n\}$

(N παρατηρήσεις με επανάληψη από $\{x_1, \dots, x_n\}$)

Έστω $(y_1, \dots, y_N)$ το bootstrap sample.

$\downarrow$

$\theta(F_e)$

Παράδειγμα

Έστω μια αναγεννητική διαδικασία $\{X(t), t \geq 0\}$

με μήκος κύκλων αναγέννησης T

σε κάθε κύκλο υπολογίζουμε ένα κόστος κύκλου C

γενικά $$C_j = \int_{T_{j-1}}^{T_j} h(X(u))\,du$$

$(C_1, T_1), (C_2, T_2), \dots$ αιτμ. με από κοινού κατανομή F

Ανανεωτικό Θεώρημα: $\lim_{t\to\infty} \frac{1}{t}\int_0^t h(X(u))\,du = \frac{E(C)}{E(T)}$ μ.π. 1.

$$\theta(F) = \frac{E_F(C)}{E_F(T)} = \frac{E(\text{κόστος κύκλου})}{E(\text{μήκος κύκλου})}$$

$$\hat{\theta} = \frac{(C_1+\dots+C_n)/n}{(T_1+\dots+T_n)/n},$$ όπου $\begin{pmatrix} C_1, T_1 \\ \vdots \\ C_n, T_n \end{pmatrix}$ δείγμα από F

F_C, F_T

$$E(\hat{\theta}) = E\left(\frac{\bar{C}}{\bar{T}}\right) = E\left(\frac{\sum C_i}{\sum T_i}\right) \neq \frac{E(\bar{C})}{E(\bar{T})} = \frac{E(C)}{E(T)}$$

$\hat{\theta}$ όχι αμερόληπτη $\quad \boxed{\hat{\theta} = \hat{\theta}((C_1,T_1),(C_2,T_2),\dots,(C_n,T_n))}$

Θέλουμε να εκτιμήσουμε το $\underline{MSE(\hat{\theta})}$

Έστω ένα δείγμα $\begin{pmatrix} (c_1,t_1) \\ \vdots \\ (c_n,t_n) \end{pmatrix}$ από F

F_e : διακρ. ομοιόμορφη στο $\{(c_1,t_1),\dots,(c_n,t_n)\}$

$$\theta(F_e) = \frac{E_{F_e}(C)}{E_{F_e}(T)} = \underline{\qquad\qquad}$$

Όμως κάτω από F_e $\quad C$: δ. ομοιόμορφη $\{c_1,\dots,c_n\}$

$$E_{F_e}(C) = \frac{1}{n} c_1 + \frac{1}{n} \cdot c_2 + \dots + \frac{1}{n} c_n = \bar{c}_n$$

$$E_{F_e}(T) = \bar{t}_n$$

$$\Rightarrow \boxed{\theta(F_e) = \frac{\bar{c}_n}{\bar{t}_n}}$$

αυτό όμως είναι αντιπροσωπευτικό

[Επειδή $\hat{\theta} = \frac{\bar{c}_n}{\bar{t}_n}$ από το συγκεκριμένο δείγμα.]

θα μπορούσε $\hat{\theta} = g(c_1, t_1, \dots, c_n, t_n)$ άλλη εκτιμήτρια.

$$MSE(\hat{\theta}) = E_F\left[(\hat{\theta} - \theta(F))^2\right]$$

$$\approx E_{F_e}\left[(\hat{\theta} - \theta(F_e))^2\right]$$

$$\boxed{\widehat{MSE}(\hat{\theta}) = E_{F_e}\left[\left(\hat{\theta} - \frac{\bar{c}_n}{\bar{t}_n}\right)^2\right]}$$

bootstrap estimate of $MSE(\hat{\theta})$.

$$\hat{\theta} = g(C_1, T_1, C_2, T_2, \dots, C_n, T_n)$$

όταν $(C_1, T_1), \dots, (C_n, T_n)$ iid από F_e (και όχι από F)

Για να εκτιμήσουμε το $\widehat{MSE}(\hat{\theta}) = E_{F_e}\left(\hat{\theta} - \frac{\bar{c}_n}{\bar{t}_n}\right)^2$

μέσω Monte Carlo:

$$\widehat{MSE} = E(Y)$$

όπου $\boxed{Y = \left(\dfrac{\overline{C}}{\overline{T}} - \dfrac{\overline{c}_n}{\overline{t}_n}\right)^2}$

όπου $\overline{C} = \dfrac{C_1 + \cdots + C_n}{n}$, $\overline{T} = \dfrac{T_1 + \cdots + T_n}{n}$

και $(C_1, T_1), \ldots, (C_n, T_n)$ διακρ. ομοι. από $\{(c_1, t_1), \ldots, (c_n, t_n)\}$

γεννήτρια της Y.

Για <u>μια παρατήρηση</u> Y_j : $j = 1, \ldots, N$

παίρνουμε n παρατ. από

διακρ. ομοιομ. $\{(c_1, t_1), (c_2, t_2), \ldots (c_n, t_n)\}$

έστω $\begin{pmatrix} C_{j1}, T_{j1} \\ C_{j2}, T_{j2} \\ \vdots \\ C_{jn}, T_{jn} \end{pmatrix} \Rightarrow Y_j = \left(\dfrac{\overline{C}_j}{\overline{T}_j} - \dfrac{\overline{c}_n}{\overline{t}_n}\right)^2$

$(j = 1, \ldots, N)$

$$\boxed{\widehat{MSE} = \frac{1}{N}\sum_{j=1}^{N} Y_j}$$

Όταν $N \to \infty$ $\quad \widehat{MSE}(\hat{\theta}) \Rightarrow E_{F_e}\left[(\hat{\theta} - \theta(F_e))^2\right]$

πόσο καλά είναι

στο $MSE = E_F\left[(\hat{\theta} - \theta(F))^2\right]$

Αν $n \to \infty \Rightarrow F_e \approx F \Rightarrow MSE_{F_e} \approx MSE_F$