MEM-262 Παραμετρική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτη

Διδάσκων : Κώστας Σμαραγδάκης

Διάλεξη 6 : 19-10-2020

Παραμετρικό μοντέλο : Αναμενόμενη τιμή και διασπορά

$$\mathbf{X} = (\overbrace{X_1, X_2, \ldots, X_n}^{X})^T$$

$\mathcal{N}(\overbrace{\mu, \sigma^2}^{\theta})$ $d=2$

$\theta \in \Theta \subseteq \mathbb{R}^d$

- Συνάρτηση μάζας πιθανότητας (διακριτές τμ) : $p(\mathbf{x}; \theta)$
- Συνάρτηση πυκνότητας πιθανότητας (συνεχείς τμ) : $f(\mathbf{x}; \theta)$

Μέση τιμή $\mathbb{E}_\theta(g(\mathbf{X}))$

$$\mathbb{E}_\theta(g(\mathbf{X})) = \begin{cases} \sum_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x})p(\mathbf{x};\theta) & \text{δτμ} \\ \int_{\mathcal{X}} g(\mathbf{x})f(\mathbf{x};\theta)d\mathbf{x} & \text{στμ} \end{cases}$$

Διασπορά $\mathrm{Var}_\theta(g(\mathbf{X}))$

$$\mathrm{Var}_\theta(g(\mathbf{X})) = \mathbb{E}_\theta((g(\mathbf{X}) - \mathbb{E}_\theta(g(\mathbf{X})))^2)$$

Στατιστικές συναρτήσεις

$X \sim$ κατανομή με παράμετρο θ $\qquad \underset{\sim}{X} = (X_1, X_2, \dots, X_n)^T$

$\underset{\sim}{x} = (x_1, x_2, \dots, x_n)$

Ορισμός : Στατιστική συνάρτηση

Κάθε συνάρτηση μόνο του τυχαίου δείγματος καλείτε στατιστική συνάρτηση.

$$T : \mathbf{x} \in \mathcal{X} \to T(\mathbf{x}) = t \in \mathcal{T}$$

- Κάθε μέθοδος στατιστικής συμπερασματολογίας σχετίζεται με μια ή περισσότερες στατιστικές συναρτήσεις.

Παράδειγμα : Κέρμα με πιθανότητα $\theta \in [0,1]$ να λάβουμε κορώνα

$\theta = 1/2$

$\underset{\sim}{x} = (1, 0, 1, 1, 0)^T \qquad T(\underset{\sim}{X}) = \sum_{j=1}^{5} X_j$

$T(\underset{\sim}{x}) = \sum_{j=1}^{5} x_j = 3$

$\underset{\sim}{x} = (0, 0, 0, 0, 1)^T \qquad T(\underset{\sim}{x}) = 1$

θ άγνωστη

$T(\underset{\sim}{X}) = \bar{X}_5 = \frac{1}{5}(X_1 + X_2 + X_3 + X_4 + X_5)$

$T(\underset{\sim}{x}) = \frac{1}{5} 3 = 3/5 = 0.6$

Έστω ότι μια τυχαία μεταβλητή X ακολουθεί μια κατανομή με παράμετρο $\theta \in \Theta \subseteq \mathbb{R}^d$. Μια στατιστική συνάρτηση $\hat{\theta}$ την ονομάζουμε εκτιμήτρια της θ αν για αρκετά μεγάλο n έχουμε:

$$\hat{\theta} = \hat{\theta}(\mathbf{X}) = \hat{\theta}(X_1, X_2, \ldots, X_n) \approx \theta$$

Βέβαια πρέπει να καθοριστεί με ποιόν τρόπο η εκτιμήτρια πλησιάζει την πραγματική τιμή.

- Το θ μπορεί επίσης να εκφράζει οποιαδήποτε ποσότητα υπολογισμένη επί τη βάση της τυχαίας μεταβλητής X. Π.χ. $\mathbb{E}_\theta(X^m)$ για κάποιο $m \in \mathbb{Z}$.

Παράδειγμα : $X \sim \mathcal{N}(\theta, 1)$

$\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)^T$ iid

$\hat{\theta}(\mathbf{X}) = \bar{X}_5$

$$\bar{X}_5 = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

$$\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)^T = (0.05,\ 0.5,\ -1,\ 0.7,\ -0.4)^T$$

$$\hat{\theta}(\mathbf{x}) = \bar{x}_5 = \frac{0.05 + 0.5 - 1 + 0.7 - 0.4}{5} = -0.03 \approx \theta$$

$$\bar{x}_3 = \frac{0.05 + 0.5 - 1}{3}$$

Ορισμός : Αμερόληπτη εκτιμήτρια

Έστω $\hat{\theta}(\mathbf{X})$ μια εκτιμήτρια της θ. Η $\hat{\theta}$ θα καλείται **αμερόληπτη** αν ισχύει:

$$\mathbb{E}_\theta(\hat{\theta}(\mathbf{X})) = \theta$$

Παράδειγμα : $X \sim \mathcal{N}(\theta, 1)$ **συνέχεια. . .** $f(x;\theta)$ $\quad \underset{\sim}{X} = (X_1, X_2, \dots, X_5)$ i.i.d

$$\hat{\theta}(\underset{\sim}{X}) = \bar{X}_5 = \frac{1}{5}\sum_{j=1}^{5} X_j \qquad \mathbb{E}_\theta(\hat{\theta}(\underset{\sim}{X})) = \mathbb{E}_\theta\left(\frac{1}{5}\sum_{j=1}^{5} X_j\right) = \frac{1}{5}\sum_{j=1}^{5}\mathbb{E}_\theta(X_j) = \frac{1}{\cancel{5}}\cdot\cancel{5}\cdot\theta = \theta$$

Παράδειγμα : $X \sim \mathcal{N}(1, \theta)$ $\quad \mathrm{Var}(X) = \mathbb{E}\left[\left(\overset{X_j}{X} - \underbrace{\mathbb{E}(X)}_{1}\right)^2\right]$

$$\hat{\theta}(\underset{\sim}{X}) = \frac{1}{n}\sum_{j=1}^{n}(X_j - 1)^2 \qquad \mathbb{E}(\hat{\theta}(X)) = \theta$$

$X \sim N(\theta_1, \theta_2)$

$$\hat{\theta}_2(\underset{\sim}{X}) = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$$

αμερ.

$$\hat{\theta}_{2b}(\underset{\sim}{X}) = \frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2$$

δεν είναι αμερ.

$$\mathbb{E}_\theta(\hat{\theta}_{2b}(\underset{\sim}{X})) = \frac{n-1}{n}\theta_2 \neq \theta_2$$

Στοιχεία θεωρίας αποφάσεων : Χώρος αποφάσεων

$\underset{\sim}{X} = (x_1, x_2, \ldots, x_n)$ $\underset{\sim}{x} \in \mathcal{X}$ $\theta \in \Theta$

Μέχρι τώρα έχουμε : δειγματικό χώρο $\mathcal{X}$, παραμετρικό χώρο Θ

Θα παρουσιάσουμε στη συνέχεια τα υπόλοιπα στοιχεία που συνθέτουν μαζί με τα προαναφερθέντα τη θεωρία αποφάσεων.

Χώρος αποφάσεων $\mathcal{A}$

Το σύνολο των διαθέσιμων αποφάσεων.

- Στη εκτιμητική: $\mathcal{A} = \Theta$ $\alpha \in \mathcal{A}$
  - Για την εκτίμηση του σ^2 μπορούμε να επιλέξουμε $\alpha = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x}_n)^2$
- Για ελέγχους υποθέσεων: $\mathcal{A} = \{0, 1\}$
  - $\alpha = 0$ δηλώνει ότι δεχόμαστε την υπόθεση $H_0 : \mu = 0$ (μηδενική υπόθεση)
  - $\alpha = 1$ δηλώνει ότι δεχόμαστε την υπόθεση $H_1 : \mu \neq 0$ (εναλλακτική υπόθεση)
- Για αποφάσεις μη σχετιζόμενες άμεσα με εκτίμηση του θ
  - $\alpha = 0$ δεχόμαστε την H_0 : απόρριψη δανείου για τον κύριο Αναξιόπιστο.
  - $\alpha = 1$ δεχόμαστε την H_1 : έγκριση δανείου για τον κύριο Αναξιόπιστο.

Σύνολο κανόνων αποφάσεων $\mathcal{D}$

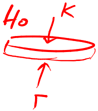
Ένα στοιχείο $d : \mathcal{X} \to \mathcal{A}$ του $\mathcal{D}$ αντιστοιχίζει κάθε $\mathbf{x}$ με μια συγκεκριμένη απόφαση $d(\mathbf{x}) \in \mathcal{A}$. Το d ονομάζεται **κανόνας ή συνάρτηση αποφάσεως**.

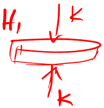
Παράδειγμα : Εκτίμηση της διασποράς του πληθυσμού

$X \sim N(\mu, \sigma^2)$

$$d(\mathbf{X}) = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2 \in \mathcal{A} = \Theta$$

Παράδειγμα : Νόμισμα με 2 κορώνες; Έλεγχος υποθέσεων





$\underset{\sim}{X} = (X_1, X_2, \dots, X_n)$

$X_j = \begin{cases} 0, & \Gamma \\ 1, & K \end{cases}$

$$d(\mathbf{X}) = \begin{cases} 0, & \sum_{k=1}^{n} X_k < n \\ 1, & \sum_{k=1}^{n} X_k = n \quad ✓ \end{cases}$$

$$\alpha = 0 \leftrightarrow H_0 : \{1 \text{ κορώνα, } 1 \text{ γράμματα}\}, \quad \alpha = 1 \leftrightarrow H_1 : \{2 \text{ κορώνες}\}$$

Συνάρτηση απώλειας $L(\vartheta, \alpha)$

Η συνάρτηση $L : \Theta \times \mathcal{A} \to [0, \infty)$ αντιστοιχίζει μια τιμή σε κάθε συνδυασμό απόφασης $\alpha \in \mathcal{A}$ και τιμής της παραμέτρου $\theta \in \Theta$.

- Όσο μικρότερη είναι η τιμή της συνάρτησης απώλειας τόσο πιο ενδεδειγμένη θεωρείται η απόφαση σε σχέση με την συγκεκριμένη τιμή της παραμέτρου.

Έστω $\theta' \in \Theta,\ \alpha_1, \alpha_2 \in \mathcal{A}$ και επιπλέον

$$L(\theta', \alpha_1) < L(\theta', \alpha_2)$$

- Στη παραπάνω περίπτωση η α_1 αποτελεί καλύτερη απόφαση από την α_2 εάν η τιμή της παραμέτρου είναι θ'.

Συνάρτηση κινδύνου

$L(\theta, \alpha)$

$d(X) : \mathcal{X} \to \mathcal{A}$

$X = (X_1, X_2, \dots, X_n)$

$$R(\theta, d) = \mathbb{E}_\theta(L(\theta, d(\mathbf{X})))$$

- Διακριτές τυχαίες μεταβλητές:

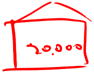
$$R(\theta, d) = \sum_{\mathbf{x} \in \mathcal{X}} L(\theta, d(\mathbf{x}))p(\mathbf{x}; \theta)$$

- Συνεχείς τυχαίες μεταβλητές:

$$R(\theta, d) = \int_{\mathcal{X}} L(\theta, d(\mathbf{x}))f(\mathbf{x}; \theta)d\mathbf{x}$$

Συνάρτηση πραγματικής ή τεχνητής απώλειας;

20000 − m



20000 − 9.999 = 10.001   R = 10.001

- d_1 ► Προσφορά 1 : 9 999 € με πιθανότητα 1
- d_2 ► Προσφορά 2 : 20 000 € με πιθανότητα 1/2 < 20.000 € / 0 €

20000 − 20.000 = 0   R = 10.000

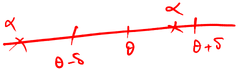
20000 − 0 = 20.000

0 € ⟷ 9999 €

9.999 € ⟷ 20.000 €

Διάφορες συναρτήσεις απώλειας

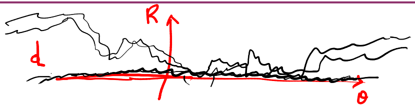
- ~~Μέσο~~ απόλυτο σφάλμα : $L(\theta, \alpha) = |\theta - \alpha|$
- ~~Μέσο~~ τετραγωνικό σφάλμα : $L(\theta, \alpha) = (\theta - \alpha)^2$
- $L(\theta, \alpha) = \begin{cases} 0, & |\theta - \alpha| \leq \delta \\ 1, & |\theta - \alpha| > \delta \end{cases}$



- $L(\theta, \alpha) = \begin{cases} 1, & \theta \in H_0 \text{ και } \alpha = 1 \\ 1, & \underline{\theta \in H_1} \text{ και } \underline{\alpha = 0} \\ 0, & \text{διαφορετικά} \end{cases}$

Εύρεση κατάλληλων συναρτήσεων αποφάσεως

$D = \{d\}$



$X \sim N(\theta, 1)$

- Ιδανικά, αναζητούμε συνάρτηση αποφάσεως για την οποία η συνάρτηση κινδύνου λαμβάνει ομοιόμορφα μικρές τιμές για όλες τις τιμές του θ.
- Δυστυχώς στις περισσότερες πραγματικές εφαρμογές δεν μπορούμε να βρούμε μια συνάρτηση αποφάσεως με τόσο καλά χαρακτηριστικά.
- Εναλλακτικά, θεωρούμε ένα σύνολο από κριτήρια για τον περιορισμό των υπό εξέταση συναρτήσεων αποφάσεως.

Ορισμός

Για δύο συναρτήσεις αποφάσεως d και d' λέμε ότι η d είναι αυστηρά επικρατέστερη της d', εάν $R(\theta, d) \leq R(\theta, d')$ για κάθε $\theta \in \Theta$ και $R(\theta, d) < R(\theta, d')$ για τουλάχιστον ένα $\theta \in \Theta$.

Ορισμός : Παραδεκτή συνάρτηση αποφάσεως

Εάν για μια d δεν υπάρχει συνάρτηση αποφάσεως που να είναι αυστηρά επικρατέστερη αυτής, τότε θα λέμε ότι η d είναι **παραδεκτή συνάρτηση αποφάσεως**.

- Στη πραγματικότητα δεν είναι πάντα εύκολο να δείξουμε ότι μια d είναι παραδεκτή.
- Μερικές φορές συναρτήσεις απόφασης που μοιάζουν ως λογικές επιλογές είναι μη παραδεκτές.

Ορισμός : Μέγιστος κίνδυνος

Ο **μέγιστος κίνδυνος** για μια συνάρτηση αποφάσεως $d \in \mathcal{D}$ ορίζεται ώς:

$$\mathrm{MR}(d) = \sup_{\theta \in \Theta} R(\theta, d)$$

Ορισμός : Minimax συνάρτηση αποφάσεως

Μια συνάρτηση αποφάσεως d είναι **minimax** εάν: $\mathrm{MR}(d) \leq \mathrm{MR}(d')$ για κάθε $d' \in \mathcal{D}$

Εναλλακτικά για να είναι το d minimax πρέπει να ικανοποιεί:

$$\sup_{\theta \in \Theta} R(\theta, d) = \inf_{d' \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, d')$$

Στις περισσότερες περιπτώσεις ισχύει:

$$\max_{\theta \in \Theta} R(\theta, d) = \min_{d' \in \mathcal{D}} \max_{\theta \in \Theta} R(\theta, d')$$

Ορισμός : Αμερόληπτη συνάρτηση αποφάσεως

Μια συνάρτηση d λέγεται **αμερόληπτη** αν

$$\mathbb{E}_\theta(L(\theta', d(\mathbf{X}))) \geq \mathbb{E}_\theta(L(\theta, d(\mathbf{X}))) \text{ για κάθε } \theta, \theta' \in \Theta$$

Εφαρμογή : Μέσο τετραγωνικό σφάλμα

Ως συνάρτηση απώλειας θεωρούμε το μέσο τετραγωνικό σφάλμα

$$L(\theta, d(\mathbf{X})) = (\theta - d(\mathbf{X}))^2$$