ΜΕΜ-262 Παραμετρική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτη

Διδάσκων : Κώστας Σμαραγδάκης

Διάλεξη 5 : 16-10-2020

Έστω ακολουθία τυχαίων μεταβλητών $X_1, X_2, \ldots$ στον ίδιο χώρο πιθανότητας και g συνεχής συνάρτηση τότε:

$$X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$$

$$X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$$

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$

Σύγκλιση της δειγματικής διασποράς

$\sigma^2 = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$

Θα δείξουμε ότι

$S_n^2 = \dfrac{\sum_{k=1}^{n}(X_k - \bar{X}_n)^2}{n-1} \xrightarrow{p} \sigma^2$

$S_n \xrightarrow{p} \sigma$

$X_1, X_2, \dots$ iid $\quad \mathbb{E}\{X_k\} = \mu$

$\mathrm{Var}\{X_k\} = \sigma^2$

$\bar{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ Δειγματικός μέσος

Δειγματική διασπορά

$S_n^2 = \dfrac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$

$S_n^2 = \dfrac{n}{n-1}\left[\dfrac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2\right]$

$\dfrac{1}{n}\sum_{k=1}^{n} X_k = \bar{X}_n \xrightarrow{p} \mu$

$\dfrac{1}{n}\sum_{k}(X_k - \bar{X}_n)^2 = \dfrac{1}{n}\sum_{k}\left[(X_k-\mu) - (\bar{X}_n - \mu)\right]^2 = \dfrac{1}{n}\sum_{k}(X_k-\mu)^2 - 2(\bar{X}_n-\mu)\dfrac{1}{n}\sum_{k}(X_k-\mu) + (\bar{X}_n-\mu)^2 =$

$= \dfrac{1}{n}\sum_{k=1}^{n}(X_k-\mu)^2 - (\bar{X}_n-\mu)^2$

$\xrightarrow{p}\sigma^2 \qquad \xrightarrow{p} 0$

$2(\bar{X}_n-\mu)$

ΝΜΑ: $\bar{X}_n \xrightarrow{p} \mu$ Επίσης $g(x) = (x-\mu)^2$

$g(\bar{X}_n) = (\bar{X}_n - \mu)^2 \qquad g(\mu) = 0$

$Y_1 = (X_1-\mu)^2, \; Y_2 = (X_2-\mu)^2, \dots, Y_n = (X_n-\mu)^2$ iid

ΝΜΑ: $\dfrac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow{p} \mathbb{E}(Y_1) = \mathbb{E}\left[(X_1-\mu)^2\right] = \mathrm{Var}(X_1) = \sigma^2$

$\dfrac{1}{n}\sum_{k=1}^{n}(X_k-\mu)^2 \xrightarrow{p} \sigma^2$

$Z_n = \dfrac{n}{n-1}$

$\dfrac{n}{n-1} \xrightarrow{p} 1$

$P(|Z_n - 1| > \varepsilon) \xrightarrow{n\to\infty} 0$

$$\left.\begin{array}{l} X_n \xrightarrow{d} X \\ Y_n \xrightarrow{p} c \end{array}\right\} \quad X_n Y_n \xrightarrow{d} cX$$

$$Z_n = \frac{1}{n}\sum_{k=1}^{n}(X_k-\mu)^2 \xrightarrow{d} \sigma^2$$

$$S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k-\bar{X}_n)^2 \qquad Var(X) = \mathbb{E}\left((X-\mathbb{E}(X))^2\right)$$

$$S_n^2 \xrightarrow{p} \sigma^2$$

δηλαδή $\mathbb{E}(S_n^2) = \sigma^2$

Αν $X_n \xrightarrow{d} c$ σταθερά τότε $X_n \xrightarrow{p} c$

$$\mathbb{E}\left(\sum_k (X_k-\bar{X}_n)^2\right) = \mathbb{E}\left(\sum_k X_k^2 - 2\bar{X}_n\sum_{k=1}^{n}X_k + n\bar{X}_n^2\right) = \sum_{k=1}^{n}\mathbb{E}(X_k^2) - \left(\mathbb{E}(n\bar{X}_n^2)\right) =$$

$$= \underbrace{\sum_{k=1}^{n}\mathbb{E}(X_k^2)}_{=} - \underbrace{n\,\mathbb{E}(\bar{X}_n^2)}_{=} = n\sigma^2 - \sigma^2 = (n-1)\sigma^2$$

$$Var(X) = \mathbb{E}(X^2) - \left(\mathbb{E}(X)\right)^2 \Rightarrow \mathbb{E}(X^2) = Var(X) + \left(\mathbb{E}(X)\right)^2 \quad n\left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$\mathbb{E}(\bar{X}_n^2) = Var(\bar{X}_n) + \left(\mathbb{E}(\bar{X}_n)\right)^2$$

$$\mathbb{E}\left(S_n^2\right) = \sigma^2.$$

Κεντρικό οριακό θεώρημα

Έστω $X_1, X_2, \dots$ $i.i.d$ και $\mu = \mathbb{E}(X_1),\ \sigma^2 = \mathrm{Var}(X_1) < \infty$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \sim \mathcal{N}(0,1)$$

όπου Z τυχαία μεταβλητή η οποία ακολουθεί την τυπική κανονική κατανομή $\mathcal{N}(0,1)$.

Αποτέλεσμα για άγνωστη διασπορά σ^2

$$\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} Z \sim N(0,1) \quad ?$$

$$\frac{\sqrt{n}\,(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{\sigma} \cdot \frac{\sigma}{S_n} \xrightarrow{d} Z \sim N(0,1)$$

- Στη στατιστική συμπερασματολογία παρατηρήσιμα δεδομένα μοντελοποιούνται ως πραγματοποιήσεις τυχαίων μεταβλητών.
- Αξιοποιώντας αυτά τα δεδομένα μελετούμε το μηχανισμό (στατιστική κατανομή) για τη δημιουργία τους.

Βασικό κομμάτι λοιπόν της στατιστικής συμπερασματολογίας αποτελεί η επεξεργασία και η ανάλυση παρατηρήσεων.

Τα παρατηρήσιμα δεδομένα θα αναπαριστώνται ως **x** με

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$$

- Κάθε x_i περιγράφει μια παρατήρηση που όπως είπαμε θεωρείται πραγματοποίηση μιας τυχαίας μεταβλητής την οποία θα συμβολίζουμε ως X_i.
- Τα x_i μπορούν να είναι πέρα από αριθμούς και διανύσματα. Σε αυτή τη περίπτωση το **x** θα είναι πίνακας.

Σύμφωνα με τα παραπάνω μπορούμε να δούμε το **x** σαν μια πραγματοποίηση μιας διανυσματικής τυχαίας μεταβλητής **X** όπου:

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

Η **X** εν γένει θα λαμβάνει τιμές στο $\mathbb{R}^n$ (ή θα περιγράφεται από πραγματικούς πίνακες εάν x_i είναι διανύσματα). **Γενικά τον χώρο στον οποίο λαμβάνει τιμές η X τον συμβολίζουμε με $\mathcal{X}$ και θα τον καλούμε δειγματικό χώρο της X.**

Παράδειγμα (ref: Liero-Zwanzig)

Ένας κατασκευαστής στυλό πραγματοποιεί τυχαία δειγματοληψία 400 μονάδων ανά ημέρα από τα 40000 στυλό της ημερήσιας παραγωγής. Οι παρατηρήσεις είναι ο αριθμός των ελαττωματικών προϊόντων ανά ημέρα. Για παράδειγμα:

Mon	Tue	Wed	Thu	Fri	Sat	Sun
8	5	9	4	6	8	10

$0-400$ $0-400$ $0-400$ - - - -

Θα περιγράψουμε τα δεδομένα, την τυχαία μεταβλητή που θεωρήται υπέυθυνη για την παραγωγή των δεδομένων και τον δειγματικό της χώρο.

$x = (8,5,9,4,6,8,10)^T$ $\quad X = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$

$\mathcal{X} = \{0,1,2,\dots,400\}^7$

Παράδειγμα (ref: Liero-Zwanzig)

Ο παρακάτω πίνακας παρουσιάζει την ημερήσια παραγωγή γαλακτός (kg/day) και τη αντίστοιχη παραγωγή πρωτεΐνη γαλακτός (kg/day) για 5 (Friesian) αγελάδες.

Milk production	Milk protein
42.7	1.20
40.2	1.16
38.2	1.07
37.6	1.13
32.2	0.96

$X = (X_1, \dots, X_5)$

$X_1 = (42.7, 1.2)$

$X_2 = (40.2, 1.16)$

$\begin{bmatrix} 42.7 & 1.2 \\ 40.2 & 1.16 \\ \vdots & \vdots \\ 32.2 & 0.96 \end{bmatrix}$

$X = \mathbb{R}^{2,5}$

Θα περιγράψουμε τα δεδομένα, την τυχαία μεταβλητή που θεωρήται υπέυθυνη για την παραγωγή των δεδομένων και τον δειγματικό της χώρο.

- Δειγματικός χώρος $\mathcal{X}$ της τυχαίας μεταβλητής $\mathbf{X}$ που περιγράφει τις παρατηρήσεις.
- Παραμετρικός χώρος Θ επί τη βάση του οποίου ορίζεται η οικογένεια των πιθανών μέτρων πιθανότητας $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$

Συνήθως θεωρούμε επαναλαμβανόμενα πειράματα που διεξάγονται υπό τις ίδιες συνθήκες και στα οποία η κάθε πραγματοποίηση δεν επηρεάζει τις υπόλοιπες.

- Τυχαίο δείγμα πειράματος:

$$X_1, X_2, \ldots, X_n \ i.i.d$$

- Μέτρο πιθανότητας:

$$P_\theta = P_{1,\theta} \otimes P_{1,\theta} \otimes \cdots \otimes P_{1,\theta}$$

- Συνάρτηση μάζας (διακριτές τμ) ή συνάρτηση πυκνότητας (συνεχείς τμ) πιθανότητας

$$p(\mathbf{x}; \theta) \quad \text{ή} \quad f(\mathbf{x}; \theta)$$

Θα ασχοληθούμε με παραμετρικά μοντέλα για τα οποία ισχύει:

Για $\theta_1, \theta_2 \in \Theta$ με $\theta_1 \neq \theta_2$ οι κατανομές P_{θ_1} και P_{θ_2} είναι διαφορετικές.

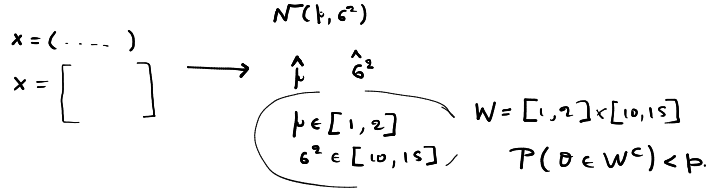
Έστω ένα ενδεχόμενο A του $\mathcal{X}$

- Για $\mathbf{X}$ διακριτή τυχαία μεταβλητή έχουμε:

$$P_\theta(A) = \sum_{\mathbf{x} \in A} p(\mathbf{x}; \theta)$$

- Για $\mathbf{X}$ συνεχή τυχαία μεταβλητή έχουμε:

$$P_\theta(A) = \int_A f(\mathbf{x}; \theta) dx$$



1. Σημειακή εκτίμηση : $\mathbf{x} \to \hat{\theta}(\mathbf{x}) \approx \theta$ $\quad p = 0.05$
2. Περιοχές εμπιστοσύνης : $\mathbf{x} \to W \subset \Theta$ τω $P(\theta \notin W) < p$ για p κοντά στο 0.
3. Έλεγχοι υποθέσεων : Λέμε ότι ισχύει μια υπόθεση H_0 (μηδενική) για το θ εκτός και εάν μπορεί να καταρριφθεί από το $\mathbf{x}$ για χάρη της εναλλακτικής υπόθεσης H_1.

$H_0 : \mu = 2$

$H_1 : \mu \neq 2$