Πιθανότητες και Στατιστική

Εκτιμήτριες

Κώστας Μανές

Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιώς

2021-2022

Δείγματα και στατιστικές συναρτήσεις

Συχνά καλούμαστε να βγάλουμε συμπεράσματα για κάποιο χαρακτηριστικό ενός πληθυσμού, βασιζόμενοι στις παρατηρήσεις που λαμβάνουμε από ένα (τυχαίο) δείγμα του πληθυσμού, το οποίο ορίζει κάποια ΤΜ X. Η ΤΜ X ακολουθεί κάποια (ίσως άγνωστη) κατανομή $F(x;\theta)$, όπου θ κάποια άγνωστη παράμετρος που συνήθως θέλουμε να εκτιμήσουμε (π.χ. η μέση τιμή ή η διακύμανση).

Ορισμός

Τυχαίο δείγμα μεγέθους n από την κατανομή F ονομάζεται μια n-άδα ανεξάρτητων ΤΜ $(X_1, X_2, \ldots, X_n)$ που ακολουθούν την κατανομή F (independent, identically distributed - iid), και γράφουμε $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$.

Στο εξής, όταν θα λέμε τυχαίο δείγμα, θα εννοούμε ότι $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$, εκτός αν αναφέρεται το αντίθετο.

Ορισμός

Η ΤΜ $T(X_1, X_2, \ldots, X_n)$, όπου $T/\mathbb{R}^n$ συνάρτηση n μεταβλητών, ονομάζεται **στατιστική συνάρτηση** (ή απλά **στατιστική**), εφόσον δεν εξαρτάται από άλλες παραμέτρους, παρά μόνο από τις τιμές των $X_1, \ldots, X_n$, και έχει πεδίο ορισμού το $S_{X_1} \times S_{X_2} \times \cdots \times S_{X_n}$.

Μερικές σημαντικές στατιστικές συναρτήσεις για ένα τυχαίο δείγμα $(X_1, X_2, \ldots, X_n)$ είναι:

- **Δειγματικός μέσος**: $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$.
- **Δειγματικό ποσοστό**: $\overline{P} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$.

  Αποτελεί ειδική περίπτωση δειγματικού μέσου, όταν οι X_i ακολουθούν κατανομή Bernoulli με κάποια παράμετρο p.

- **Δειγματική διακύμανση:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right).$$

Η τελευταία ισότητα προκύπτει από την επόμενη ιδιότητα:

Λήμμα

Για οποιουσδήποτε $x_1, x_2, \ldots, x_n \in \mathbb{R}$, $n \in \mathbb{N}^*$, ισχύουν οι σχέσεις

$$\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{και} \qquad \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$$

όπου $\overline{x} = (x_1 + \cdots + x_n)/n$.

Απόδειξη.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \overline{x} + \overline{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(\overline{x} - a)^2 + 2\sum_{i=1}^{n}(x_i - \overline{x})(\overline{x} - a) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(\overline{x} - a)^2 + 2\sum_{i=1}^{n}(x_i\overline{x} - ax_i - \overline{x}^2 + a\overline{x}) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(\overline{x} - a)^2 + 2n(\overline{x}^2 - a\overline{x} - \overline{x}^2 + a\overline{x}) \\
&= \sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(\overline{x} - a)^2
\end{aligned}$$

Η τελευταία παράσταση προφανώς ελαχιστοποιείται όταν $a = \overline{x}$.

Η δεύτερη σχέση προκύπτει θέτοντας $a = 0$ στην τελευταία. □

Κατανομές δειγματοληψίας

Οι στατιστικές συναρτήσεις είναι τυχαίες μεταβλητές που οι τιμές τους μεταβάλλονται από δείγμα σε δείγμα. Θεωρούμε ότι ακολουθούν κάποια κατανομή η οποία ονομάζεται **κατανομή δειγματοληψίας**. Στη συνέχεια θα δούμε τις κατανομές δειγματοληψίας για ορισμένες χρήσιμες στατιστικές συναρτήσεις.

Κατανομή δειγματικού μέσου

- Αν $(X_1, X_2, \ldots, X_n)$ τυχαίο δείγμα ενός (άπειρου ή πεπερασμένου) πληθυσμού με μέσο μ και διακύμανση σ^2, όπου η δειγματοληψία γίνεται με επανατοποθέτηση ή γενικότερα όταν οι X_i θεωρούνται ανεξάρτητες, τότε

$$E(\overline{X}) = \frac{1}{n}E(X_1 + \cdots + X_n) = \mu \text{ και } V(\overline{X}) = \frac{1}{n^2}V(X_1 + \cdots + X_n) \overset{\text{iid}}{=} \frac{\sigma^2}{n}$$

- Αν $(X_1, X_2, \ldots, X_n)$ τυχαίο δείγμα από πληθυσμό μεγέθους N, με $0 < n < N$, με μέσο μ και διακύμανση σ^2, όπου η δειγματοληψία γίνεται χωρίς επανατοποθέτηση ή γενικότερα όταν οι X_i θεωρούνται εξαρτημένες, τότε

$$E(\overline{X}) = \mu \text{ και } V(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.$$

Παρατηρήστε ότι καθώς N τείνει στο $+\infty$ ο τύπος της διακύμανσης ταυτίζεται με αυτόν στην προηγούμενη περίπτωση.

Κατανομή δειγματικού μέσου

- Αν ένα τυχαίο δείγμα μεγέθους n προέρχεται από ένα πληθυσμό που ακολουθεί την $N(\mu, \sigma^2)$, τότε ο δειγματικός μέσος $\overline{X}$ ακολουθεί την $N(\mu, \frac{\sigma^2}{n})$, αλλιώς, σύμφωνα με το Κ.Ο.Θ., ισχύει η προσέγγιση $\overline{X} \to N(\mu, \frac{\sigma^2}{n})$.
- Αν οι μεταβλητές X_i ακολουθούν κατανομή $\mathrm{Bern}(p)$, με άγνωστη παράμετρο p, μετρώντας μια ιδιότητα που τα άτομα του πληθυσμού έχουν σε ποσοστό p, τότε $X_1 + X_2 + \cdots + X_n \sim \mathrm{Binom}(n, p)$, οπότε

$$E(\overline{P}) = \frac{np}{n} = p \qquad \text{και} \qquad V(\overline{P}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Κατανομή δειγματικού μέσου

- Αν $(X_1, X_2, \ldots, X_{n_1})$ και $(Y_1, Y_2, \ldots, Y_{n_2})$ δύο ανεξάρτητα τυχαία δείγματα δύο άπειρων πληθυσμών με μέσους μ_1, μ_2 και διακυμάνσεις σ_1, σ_2 αντίστοιχα, τότε

$$E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 \text{ και } V(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Ειδικά Αν $X_i \sim N(\mu_1, \sigma_1^2)$ και $Y_i \sim N(\mu_2, \sigma_2^2)$, τότε

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

όπου $\overline{X}, \overline{Y}$ είναι δειγματικοί μέσοι των δύο δειγμάτων αντίστοιχα.

Η κατανομή χ^2

Η κατανομή της δειγματικής διακύμανσης S^2 ενός δείγματος από κανονικό πληθυσμό σχετίζεται με την κατανομή χι τετράγωνο.

Ορισμός

Αν οι ΤΜ $X_1, X_2, \ldots, X_\nu$ είναι ανεξάρτητες και ακολουθούν την τυποποιημένη κανονική κατανομή $N(0,1)$ τότε η ΤΜ

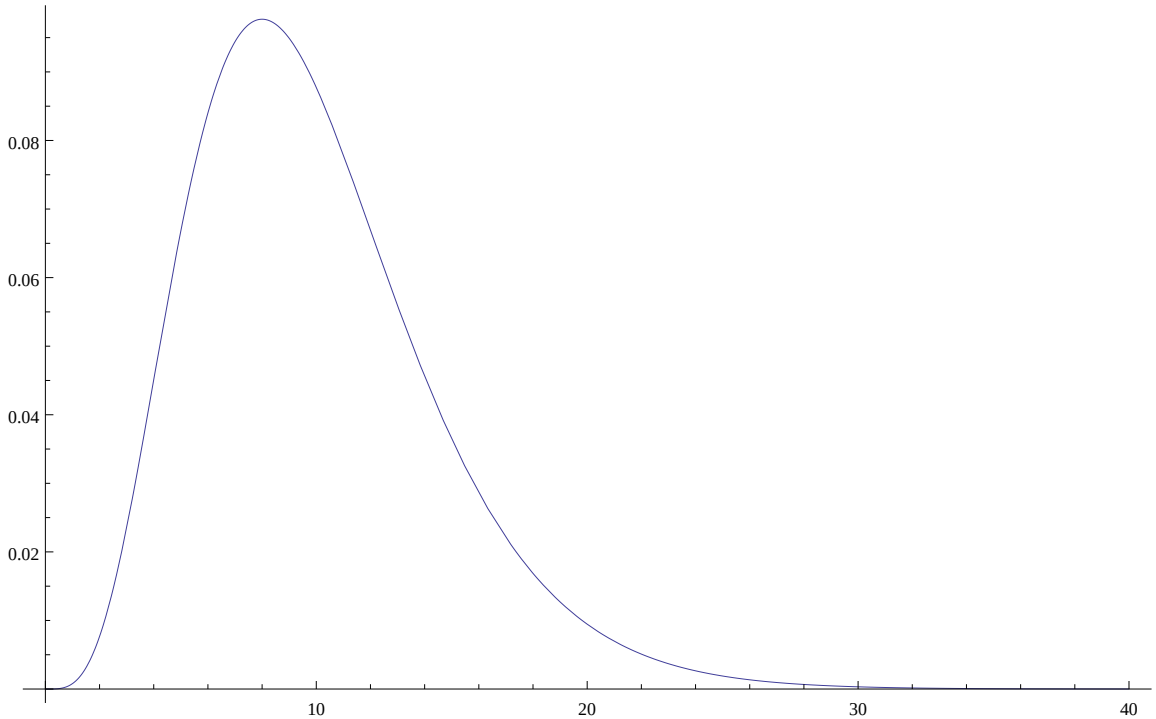
$$Y = X_1^2 + X_2^2 + \cdots + X_\nu^2$$

ακολουθεί την λεγόμενη **κατανομή χι τετράγωνο με ν βαθμούς ελευθερίας**, η οποία συμβολίζεται με χ^2_ν και έχει PDF

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma\left(\nu/2\right)} x^{\nu/2-1} e^{-x/2} & x > 0 \\ 0 & x < 0 \end{cases}$$

όπου $\Gamma(x)$ είναι η συνάρτηση Γάμμα.

Η κατανομή χ^2



Σχήμα: Η συνάρτηση πυκνότητας πιθανότητας της χ^2_ν για $\nu = 10$

Η κατανομή χ^2_ν είναι ειδική περίπτωση της κατανομής Γάμμα $\Gamma(a, b)$ με παραμέτρους $a = \nu/2$ και $\beta = 1/2$. Αποδεικνύεται εύκολα ότι αν $Y \sim \chi^2_\nu$ τότε

$$E(Y) = \nu \text{ και } V(Y) = 2\nu.$$

Η κατανομή χ^2

Για την χ^2_ν κατανομή υπάρχουν πίνακες όπου για δοσμένη πιθανότητα p και δοσμένους βαθμούς ελευθερίας $\nu \leq 30$ δίνουν την τιμή $\chi^2_{\nu,p}$, για την οποία $P(Y > \chi^2_{\nu,p}) = p$.
Για παράδειγμα, για $p = 0.05$, $\nu = 20$ είναι $\chi^2_{20,0.05} = 31.41$ το οποίο σημαίνει ότι για 20 βαθμούς ελευθερίας

$$P(\chi^2_{20} > \chi^2_{20,0.05}) = P(\chi^2_{20} > 31.41) = 0.05.$$

Η κατανομή χ^2

- Αν η ΤΜ X ακολουθεί την τυποποιημένη κανονική κατανομή $N(0,1)$ τότε η ΤΜ X^2 ακολουθεί την κατανομή χ^2_1 με ένα βαθμό ελευθερίας.
- Αν $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες ΤΜ που ακολουθούν την χ^2 κατανομή με $\nu_1, \nu_2, \ldots, \nu_n$ βαθμούς ελευθερίας αντίστοιχα, τότε η ΤΜ

$$Y = \sum_{i=1}^{n} X_i$$

ακολουθεί την κατανομή χ^2_m, όπου $m = \nu_1 + \nu_2 + \cdots + \nu_n$.

- Αν ο πληθυσμός ακολουθεί κανονική κατανομή $N(\mu, \sigma^2)$ και n το μέγεθος του δείγματος, τότε η στατιστική συνάρτηση

$$\frac{(n-1)S^2}{\sigma^2}$$

ακολουθεί την κατανομή χ^2_{n-1} και επιπλέον οι ΤΜ $\overline{X}$ και $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ είναι ανεξάρτητες.

Κατανομή t του Student

Η κατανομή t χρησιμοποιείται όταν η διακύμανση του σ^2 του πληθυσμού είναι άγνωστη.

Ορισμός

Αν X και Y είναι δύο ανεξάρτητες ΤΜ με $X \sim N(0,1)$ και $Y \sim \chi^2_\nu$ τότε λέμε ότι η μεταβλητή

$$T = \frac{X}{\sqrt{Y/\nu}}$$

ακολουθεί την **κατανομή** t **με** ν **βαθμούς ελευθερίας**, ή συμβολικά $T \sim t_\nu$, η οποία έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R}.$$

Όταν $\nu \to \infty$ τότε η T προσεγγίζει την $N(0,1)$ (Πρακτικά, αν $n > 30$).
Αποδεικνύεται ότι αν $T \sim t_\nu$ τότε

$$E(T) = 0 \text{ και } V(T) = \frac{\nu}{\nu - 2}.$$

Κατανομή t του Student

Η t κατανομή είναι συμμετρική και ισχύει ότι $F(t) + F(-t) = 1$ καθώς επίσης $t_{\nu,1-p} = -t_{\nu,p}$.
Υπάρχουν πίνακες που για δοσμένη δοσμένη πιθανότητα p και δοσμένους βαθμούς ελευθερίας ν δίνουν την τιμή $t_{\nu,p}$, όπου $P(t > t_{\nu,p}) = p$.
Για παράδειγμα, για $p = 0.05$, $\nu = 30$ είναι $t_{30,0.05} = 1.645$ το οποίο σημαίνει ότι για 30 βαθμούς ελευθερίας

$$P(t_\nu > t_{\nu,p}) = p \Leftrightarrow P(t_{30} > 1.645) = 0.05.$$

Κατανομή t του Student

Πρόταση

Αν $\overline{X}$ και S^2 είναι ο δειγματικός μέσος και η δειγματική διακύμανση ενός τυχαίου δείγματος μεγέθους n που έχει ληφθεί από κανονικό πληθυσμό με μέσο μ διακύμανση σ^2 τότε $T := \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

Απόδειξη.

Επειδή $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ έπεται ότι $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Επίσης, ισχύει ότι $Y = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Επιπλέον, επειδή οι $\overline{X}$ και S^2 είναι ανεξάρτητες, έπεται ότι και οι Z, Y είναι ανεξάρτητες. Επομένως, η ΤΜ $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} = \dfrac{Z}{\sqrt{Y/(n-1)}}$ ακολουθεί την κατανομή t με $n-1$ βαθμούς ελευθερίας. □

Κατανομή F του Fisher

Ορισμός

Αν X, Y είναι ανεξάρτητες ΤΜ με $X \sim \chi^2_{\nu_1}$ και $Y \sim \chi^2_{\nu_2}$ τότε η ΤΜ

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

ακολουθεί την **κατανομή F με ν_1, ν_2 βαθμούς ελευθερίας**, συμβολικά $F \sim F_{\nu_1,\nu_2}$, η οποία έχει PDF

$$f(x) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{\nu_1/2-1} (\nu_1 x + \nu_2)^{-(\nu_1+\nu_2)/2}.$$

Αποδεικνύεται ότι αν $F \sim F_{\nu_1,\nu_2}$ τότε

$$E(F) = \frac{\nu_2}{\nu_2 - 2}, \text{ αν } \nu_2 > 2 \text{ και } V(F) = \frac{2\nu_2^2(\nu_1+\nu_2-4)}{\nu_1(\nu_2-2)^2(\nu_2-4)}, \text{ αν } \nu_2 > 4$$

Κατανομή F του Fisher

Υπάρχουν πίνακες που για δοσμένη δοσμένη πιθανότητα p και δοσμένους βαθμούς ελευθερίας ν_1, ν_2 δίνουν την τιμή $F_{\nu_1,\nu_2,p}$, όπου $P(X > F_{\nu_1,\nu_2,p}) = p$.
Για παράδειγμα, για $p = 0.05$, $\nu_1 = 20$, $\nu_2 = 21$ είναι $F_{20,21,0.05} = 2.09$ το οποίο σημαίνει ότι

$$P(F_{\nu_1,\nu_2} > F_{\nu_1,\nu_2}) = 0.05 \Leftrightarrow P(F_{20,21} > 2.09) = 0.05.$$

Οι τιμές $F_{\nu_1,\nu_2,p}$ και $F_{\nu_2,\nu_1,1-p}$ είναι αντίστροφοι αριθμοί.
Αν $X \sim F_{\nu_1,\nu_2}$ τότε η ΤΜ $Y = \dfrac{1}{X} \sim F_{\nu_2,\nu_1}$.

Κατανομή F του Fisher

Πρόταση

Αν S_1^2 και S_2^2 είναι οι δειγματικές διακυμάνσεις δύο τυχαίων δειγμάτων μεγέθους ν_1 και ν_2 αντίστοιχα που έχουν ληφθεί από κανονικούς πληθυσμούς με διακυμάνσεις σ_1^2 και σ_2^2 αντίστοιχα, τότε η στατιστική συνάρτηση $F = \dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ ακολουθεί την κατανομή F_{ν_1-1,ν_2-1}.

Απόδειξη.

Γνωρίζουμε ότι $X = \dfrac{(\nu_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{\nu_1-1}$ και $Y = \dfrac{(\nu_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{\nu_2-1}$.

Επειδή οι ΤΜ S_1^2 και S_2^2 είναι ανεξάρτητες, έπεται ότι και οι ΤΜ X, Y είναι επίσης ανεξάρτητες. Επομένως,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{X/(\nu_1-1)}{Y/(\nu_2-1)} \sim F_{\nu_1-1,\nu_2-1}.$$

□

Σύνοψη κατανομών δειγματοληψίας

Οι επόμενες προτάσεις συνοψίζουν τα προηγούμενα:

Πρόταση

Αν $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, τότε

$i)$ $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ $\quad$ $ii)$ $\dfrac{n-1}{\sigma^2} S^2 \sim \chi^2_{n-1}$

$iii)$ Οι ΤΜ $\overline{X}$ και S^2 είναι ανεξάρτητες. $\quad$ $iv)$ $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

Σύμφωνα με το Κ.Ο.Θ., ανεξάρτητα από το ποια είναι η F, όταν το n είναι πολύ μεγάλο, έχουμε ότι $\overline{X} \to \mathcal{N}(\mu, \sigma^2/n)$ και

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1), \quad T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \to \mathcal{N}(0,1), \quad \overline{P} \to \mathcal{N}(p, p(1-p)/n),$$

καθώς $n \to \infty$.

Σύνοψη κατανομών δειγματοληψίας

Πρόταση

Αν $X_1, X_2, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ και $Y_1, Y_2, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, τότε

i) $Z = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1)$ *ii)* $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$

iii) Αν $\sigma_1 = \sigma_2$, τότε $T_{n_1+n_2-2} = \dfrac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$, όπου

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}$$

Σημειακές εκτιμήτριες

Στην ενότητα αυτή θα δούμε τεχνικές για την εκτίμηση μιας άγνωστης παραμέτρου θ της κατανομής $F(x;\theta)$ ενός πληθυσμού, με βάση τις τιμές ενός δείγματος $(X_1, X_2, \ldots, X_n)$. Μια στατιστική συνάρτηση $T = T(X_1, X_2, \ldots, X_n)$ ονομάζεται (σημειακή) **εκτιμήτρια της παραμέτρου** θ, αν δεν εξαρτάται από την θ. Η ποσότητα

$$\mathrm{bias}(T) := E(T) - \theta$$

καλείται **μεροληψία** της εκτιμήτριας T. Η ποσότητα

$$\mathrm{mse(T)} := E((T-\theta)^2) = V(T) + \mathrm{bias}^2(T)$$

καλείται **μέσο τετραγωνικό σφάλμα** της εκτιμήτριας T από την θ. Η δεύτερη ισότητα προκύπτει, θέτοντας $c = E(T)$, ως εξής:

$$\begin{aligned} E((T-\theta)^2) &= E((T-c+c-\theta)^2) \\ &= E((T-c)^2 + 2(T-c)(c-\theta) + (c-\theta)^2) \\ &= E((T-c)^2) + 2E(cT - \theta T - c^2 + c\theta) + E((c-\theta)^2) \\ &= V(T) + \mathrm{bias}^2(T) \end{aligned}$$

Σημειακές εκτιμήτριες

Υπάρχουν 4 βασικά χαρακτηριστικά που επιθυμούμε να έχει μια εκτιμήτρια:

- **Αμεροληψία**: Μια εκτιμήτρια T μιας παραμέτρου θ ονομάζεται **αμερόληπτη** ανν $E(T) = \theta$.
- **Αποτελεσματικότητα**: Η εκτιμήτρια T_1 είναι αποτελεσματικότερη από την T_2, ανν $\mathrm{mse}(\mathrm{T}_1) < \mathrm{mse}(\mathrm{T}_2)$. Ειδικότερα, αν είναι αμερόληπτες, τότε $V(T_1) < V(T_2)$.
- **Επάρκεια**: Μια εκτιμήτρια T της παραμέτρου θ ονομάζεται **επαρκής** αν χρησιμοποιεί όλες τις πληροφορίες του δείγματος (η τιμή της εξαρτάται από όλες τις ΤΜ $X_1, \ldots, X_n$).
- **Συνέπεια**: Η ακολουθία εκτιμητριών (T_n) είναι **συνεπής**, αν τείνει κατά πιθανότητα στην θ (γράφουμε $T_n \xrightarrow{p} \theta$), δηλαδή αν $\lim_{n\to\infty} P(|T_n - \theta| \geq \epsilon) = 0$, για κάθε $\epsilon > 0$.

Ένα απλό κριτήριο για τη συνέπεια είναι το ακόλουθο:

Πρόταση

Αν για την ακολουθία εκτιμητριών (T_n) ισχύουν $\lim_{n\to\infty} E(T_n) = \theta$ και $\lim_{n\to\infty} V(T_n) = 0$, τότε η (T_n) είναι συνεπής.

Απόδειξη.

Έστω $\epsilon > 0$. Αφού $\lim_{n\to\infty} E(T_n) = \theta$, έπεται ότι υπάρχει $n_0 \in \mathbb{N}^*$ τέτοιο ώστε $n \geq n_0 \Rightarrow |E(T_n) - \theta| < \epsilon \Rightarrow \epsilon - |E(T_n) - \theta| > 0$. Επομένως, για $n \geq n_0$, βάσει της ανισότητας Chebyshev, είναι

$$\begin{aligned} P(|T_n - \theta| \geq \epsilon) &= P(|T_n - E(T_n) + (E(T_n) - \theta)| \geq \epsilon) \\ &\leq P(|T_n - E(T_n)| + |E(T_n) - \theta| \geq \epsilon) \\ &= P(|T_n - E(T_n)| \geq \epsilon - |E(T_n) - \theta|) \leq \frac{V(T_n)}{(\epsilon - |E(T_n) - \theta|)^2} \end{aligned}$$

και το δεύτερο μέλος τείνει στο 0 καθώς $n \to \infty$. □

Βάσει της επόμενης πρότασης, μπορούμε να κατασκευάσουμε εκτιμήτριες για τη μετασχηματισμένη παράμετρο $g(\theta)$:

Πρόταση

Αν για τις ακολουθίες εκτιμητριών $(T_n), (W_n)$ ισχύουν $T_n \xrightarrow{p} \theta_1$ και $W_n \xrightarrow{p} \theta_2$, και g συνεχής συνάρτηση στο θ_1, τότε

$$g(T_n) \xrightarrow{p} g(\theta_1), \quad T_n + W_n \xrightarrow{p} \theta_1 + \theta_2, \quad T_n - W_n \xrightarrow{p} \theta_1 - \theta_2,$$

$$T_n W_n \xrightarrow{p} \theta_1 \theta_2, \quad T_n / W_n \xrightarrow{p} \theta_1 / \theta_2.$$

Παράδειγμα

Έστω ένας πληθυσμός με μέση τιμή μ και διακύμανση σ^2. Να δειχθεί ότι:

i) Ο δειγματικός μέσος $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ είναι αμερόληπτη και συνεπής εκτιμήτρια της μέσης τιμής μ.

ii) Η στατιστική συνάρτηση $T = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ είναι αμερόληπτη εκτιμήτρια για την διακύμανση σ^2.

iii) Η στατιστική συνάρτηση $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ είναι αμερόληπτη εκτιμήτρια για την διακύμανση σ^2.

iv) Η στατιστική συνάρτηση $T = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ δεν είναι αμερόληπτη εκτιμήτρια για την διακύμανση σ^2.

Λύση.

i) Έχει ήδη δειχθεί ότι $E(\overline{X}) = \mu$ και $V(\overline{X}) = \dfrac{\sigma^2}{n}$, άρα η $\overline{X}$ είναι αμερόληπτη και συνεπής.

ii) $$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}E((X_i-\mu)^2) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2.$$

iii) $$E((n-1)S^2) = E\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) = \sum_{i=1}^{n}E(X_i^2) - nE(\overline{X}^2)$$

$$= \sum_{i=1}^{n}(\sigma^2+\mu^2) - n(V(\overline{X})+\mu^2) = n\sigma^2 - \frac{n}{n^2}n\sigma^2 = (n-1)\sigma^2$$

άρα $E(S^2) = \sigma^2$.

iv) $E(nT/(n-1)) = E(S^2) = \sigma^2$, άρα $E(T) = \dfrac{n-1}{n}\sigma^2 \neq \sigma^2$. □

Μέθοδος μέγιστης πιθανοφάνειας

Η πιο γνωστή μέθοδος εύρεσης εκτιμητριών είναι η μέθοδος μέγιστης πιθανοφάνειας.
Έστω $(X_1, X_2, \ldots, X_n)$ τυχαίο δείγμα ενός πληθυσμού με PDF (ή PMF) $f(x;\theta)$, όπου θ είναι μια άγνωστη παράμετρος που θέλουμε να εκτιμήσουμε. Τότε επειδή οι $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες ΤΜ από την ίδια κατανομή, η από κοινού PDF (ή PMF) είναι η

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Αν γνωρίζουμε τα $x_1, x_2, \ldots, x_n$ (τις τιμές του δείγματος), τότε το γινόμενο αυτό είναι μια συνάρτηση του θ που συμβολίζεται με

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

και ονομάζεται **συνάρτηση πιθανοφάνειας** για το δείγμα $(X_1, X_2, \ldots, X_n)$.

Μέθοδος μέγιστης πιθανοφάνειας

Αν η $L(\theta)$ έχει μέγιστο για κάποιο $\theta = \bar{\theta}$, ή γενικότερα

$$L(\bar{\theta}) = \sup_{\theta} L(\theta),$$

τότε το $\bar{\theta}$ αυτό είναι προφανώς συνάρτηση των $(X_1, X_2, \ldots, X_n)$, άρα είναι μια εκτιμήτρια του θ, και ονομάζεται εκτιμήτρια μέγιστης πιθανοφάνειας (Maximum Likelihood Estimator - MLE).

Παρατήρηση

Επειδή η $L(\theta)$ είναι γινόμενο μη αρνητικών όρων, αντί να μεγιστοποίησουμε την $L(\theta)$ συνήθως μεγιστοποιούμε την $\ell(\theta) := \ln L(\theta)$.

Μέθοδος μέγιστης πιθανοφάνειας

Παράδειγμα

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την κατανομή Bernoulli με παράμετρο p. Να βρεθεί εκτιμήτρια για το p με την μέθοδο της μέγιστης πιθανοφάνειας.

Λύση

Για κάθε $i \in [n]$, η συνάρτηση πιθανότητας της X_i είναι η

$$f(x_i; p) = p^{x_i}(1-p)^{1-x_i}, \qquad x_i \in \{0, 1\}.$$

Επομένως,

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

και, θέτοντας $a = \sum_{i=1}^{n} x_i$, προκύπτει ότι

Λύση (συνέχεια)

$$\ell(p) = \ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$
$$= a \ln p + (n-a) \ln(1-p).$$

Παραγωγίζοντας ως προς p,

$$\ell'(p) = \frac{a}{p} - \frac{n-a}{1-p} = \frac{a(1-p) - (n-a)p}{p(1-p)} = \frac{a - np}{p(1-p)} \geq 0 \Leftrightarrow p \leq \frac{a}{n}$$

Επομένως, η $L(p)$ μεγιστοποιείται όταν $p = \frac{a}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$ και η MLE για την παράμετρο p είναι το δειγματικό ποσοστό $\overline{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Μέθοδος μέγιστης πιθανοφάνειας

Παράδειγμα

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την κατανομή Poisson με παράμετρο λ. Να βρεθεί εκτιμήτρια για το λ με την μέθοδο της μέγιστης πιθανοφάνειας.

Λύση

Ομοίως με το προηγούμενο παράδειγμα,

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdots x_n!}$$

$$\ell(\lambda) = -n\lambda + (x_1 + \cdots + x_n) \ln \lambda - c, \qquad c = \ln(x_1! \cdots x_n!)$$

Θέτοντας $a = x_1 + \cdots + x_n$ και παραγωγίζοντας ως προς λ, προκύπτει ότι $\ell'(\lambda) = -n + \frac{a}{\lambda} \geq 0 \Leftrightarrow \lambda \leq \frac{a}{n}$. Επομένως, η $L(\lambda)$ μεγιστοποιείται όταν $\lambda = a/n$ και η MLE για την παράμετρο λ είναι ο δειγματικός μέσος $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Μέθοδος μέγιστης πιθανοφάνειας

Παράδειγμα

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$. Να βρεθούν εκτιμήτριες για τις μ και σ^2 με την μέθοδο της μέγιστης πιθανοφάνειας.

Λύση

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right)$$

$$\ell(\mu, \sigma^2) = \ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}$$

Παραγωγίζοντας ως προς μ,

$$\frac{\partial \ell}{\partial \mu} = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(x_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

Λύση (συνέχεια)

Παραγωγίζοντας ως προς σ^2 και θέτοντας $a = \sum_{i=1}^{n}(x_i - \mu)^2$,

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{a}{2(\sigma^2)^2} = \frac{a - n\sigma^2}{2\sigma^4} = 0 \Leftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Άρα, οι δύο μερικές παράγωγοι γίνονται ίσες με 0 στο σημείο $(\hat{\mu}, \hat{\sigma}^2)$, όπου $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ και $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$, στο οποίο η L μεγιστοποιείται και οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι αντίστοιχα η $\overline{X}$ για το μ και η $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ για το σ^2.

Μέθοδος μέγιστης πιθανοφάνειας

Άσκηση

Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την εκθετική κατανομή με παράμετρο θ. Να βρεθεί εκτιμήτρια για το θ με την μέθοδο της μέγιστης πιθανοφάνειας.