Το κεντρικό Οριακό Θεώρημα

Πιθανότητες και Στατιστική

Τμήμα Πληροφορικής, Πανεπιστήμιο Πειραιώς

2023-2024

Ορισμός (σύγκλιση κατά κατανομή)

Έστω $(X_n)_{n\in\mathbb{N}^*}$ ακολουθία ΤΜ με κοινό σύνολο τιμών S και έστω $F_n(x) = P(X_n \leq x)$ η συνάρτηση κατανομής της X_n, $n \in \mathbb{N}$. Λέμε ότι η ακολουθία (X_n) συγκλίνει κατά κατανομή στην ΤΜ X που ακολουθεί μια κατανομή D και έχει συνάρτηση κατανομής $F(x) = P(X \leq x)$ και σύνολο τιμών S και γράφουμε $X_n \to D$, αν και μόνο αν

$$\lim_{n\to\infty} F_n(x) = F(x),$$

για κάθε $x \in S$ όπου F συνεχής.

Κεντρικό Οριακό Θεώρημα

Κεντρικό οριακό θεώρημα

Αν οι (διακριτές ή συνεχείς) ΤΜ $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες και ακολουθούν την ίδια κατανομή με μέση τιμή μ και διακύμανση σ^2, τότε

i) $S_n := X_1 + X_2 + \cdots + X_n \to N(n\mu, n\sigma^2)$,

ii) $\overline{X}_n := \dfrac{X_1 + X_2 + \cdots + X_n}{n} \to N(\mu, \sigma^2/n)$,

iii) $\overline{S}_n := \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \dfrac{S_n - n\mu}{\sigma\sqrt{n}} = \dfrac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n}(X_i - \mu) \to N(0,1)$,

iv) $\lim\limits_{n\to\infty} P(\overline{S}_n \le x) = \Phi(x)$, για κάθε $x \in \mathbb{R}$.

Το θεώρημα συνήθως εφαρμόζεται για τον προσεγγιστικό υπολογισμό πιθανοτήτων του εμπειρικού μέσου $\overline{X}_n$ που συνήθως προέρχεται από τις παρατηρήσεις σε ένα δείγμα μεγέθους n. Η ακρίβεια της προσέγγισης εξαρτάται κυρίως από το n και πολύ λιγότερο από την κατανομή των ΤΜ X_n, για μικρές τιμές του n. Στις εφαρμογές, συνήθως λαμβάνεται $n \ge 30$.

Κεντρικό Οριακό Θεώρημα

Προσέγγιση της διωνυμικής κατανομής από την κανονική:
Αν $X \sim \mathrm{Binom}(n, p)$, οπότε $\mu = np$ και $\sigma^2 = np(1-p)$, και το n είναι αρκετά μεγάλο, τότε η X ακολουθεί κατά προσέγγιση την $\mathcal{N}(\mu, \sigma^2)$, δηλαδή για κάθε ζεύγος ακεραίων a, b, με $0 \leq a \leq b \leq n$, ισχύει

$$P(a \leq X \leq b) = P(a - \frac{1}{2} \leq X \leq b + \frac{1}{2}) \simeq \Phi(\frac{b+1/2-\mu}{\sigma}) - \Phi(\frac{a-1/2-\mu}{\sigma}).$$

Η προσθήκη του $1/2$ στα δύο άκρα της ανισότητας ονομάζεται διόρθωση συνέχειας και γίνεται για τη βελτίωση της προσέγγισης, λόγω της μετάβασης από διακριτή σε συνεχή μεταβλητή.

Προσέγγιση της κατανομής Poisson από την κανονική:
Ομοίως, αν $X \sim \mathrm{Poisson}(\lambda)$, οπότε $\mu = \sigma^2 = \lambda$, και το λ είναι αρκετά μεγάλο ($\lambda > 20$), τότε η X ακολουθεί κατά προσέγγιση την $\mathcal{N}(\mu, \sigma^2)$ και εφαρμόζεται ο προηγούμενος προσεγγιστικός τύπος με διόρθωση συνέχειας.

Παράδειγμα

Ρίχνουμε ένα αμερόληπτο κέρμα 1000 φορές και έστω X η ΤΜ που ισούται με τον αριθμό των εμφανίσεων της όψης ΓΡΑΜΜΑΤΑ.

i) Να βρεθούν οι πιθανότητες $P(490 \leq X \leq 510)$ και $P(485 \leq X \leq 515)$.

ii) Να βρεθεί ο ελάχιστος φυσικός N ώστε $P(X \leq N) \geq 0.99$.

Λύση

Έστω X_i, $i = 1, 2, \ldots, 1000$ οι ΤΜ με

$$X_i = \begin{cases} 1, & \text{αν στην } i\text{-οστή ρίψη έρθει η όψη ΓΡΑΜΜΑΤΑ} \\ 0, & \text{αλλιώς.} \end{cases}.$$

και $X = X_1 + X_2 + \cdots X_{1000}$.
Οι ΤΜ X_i ακολουθούν την κατανομή Bernoulli με παράμετρο $p = \frac{1}{2}$, άρα έχουν μέση τιμή $\mu = \frac{1}{2}$ και διακύμανση $\sigma^2 = p(1-p) = \frac{1}{2}(1-\frac{1}{2}) = \frac{1}{4}$.

Λύση (συνέχεια)

Επομένως, η X ακολουθεί την διωνυμική κατανομή με μέση τιμή $1000p = 500$ και διακύμανση $1000\frac{1}{4} = 250$.
Μπορούμε όμως, βάσει του κεντρικού οριακού θεωρήματος, να θεωρήσουμε ότι η X ακολουθεί ασυμπτωτικά την κανονική κατανομή $N(1000 \cdot \frac{1}{2}, 1000 \cdot \frac{1}{4}) = N(500, 250)$.

$$\begin{aligned} P(490 \leq X \leq 510) &= P(\frac{490-500}{\sqrt{250}} \leq \frac{X-500}{\sqrt{250}} \leq \frac{510-500}{\sqrt{250}}) \\ &\approx P(-\sqrt{2/5} \leq Z \leq \sqrt{2/5}) = 2\Phi(\sqrt{2/5}) - 1 \\ &= 2\Phi(0.6324) - 1 = 2 \cdot 0.7357 - 1 = 0.4714. \end{aligned}$$

$$\begin{aligned} P(485 \leq X \leq 515) &= P(\frac{485-500}{\sqrt{250}} \leq \frac{X-500}{\sqrt{250}} \leq \frac{515-500}{\sqrt{250}}) \\ &= P(-\frac{3}{\sqrt{10}} \leq Z \leq \frac{3}{\sqrt{10}}) \approx 2\Phi(0.95) - 1 = 0.6578. \end{aligned}$$

Λύση (συνέχεια)

Εισάγοντας τη διόρθωση συνέχειας, επειδή $\epsilon = \frac{1/2}{\sqrt{250}} = 0.0316$, έχουμε

$$P(490 \leq X \leq 510) = P(490 - 1/2 \leq X \leq 510 + 1/2)$$
$$= P(-\sqrt{2/5} - \epsilon \leq Z \leq \sqrt{2/5} + \epsilon) \approx 2\Phi(\sqrt{2/5} + \epsilon) - 1 = 0.49331$$

και ομοίως $P(485 \leq X \leq 515) = 2\Phi(0.95 + \epsilon) - 1 = 0.6737$.
Για να δούμε πόσο καλή είναι προσέγγιση της X από την κανονική κατανομή μπορούμε να υπολογίσουμε τις αντίστοιχες πιθανότητες με τους τύπους της διωνυμικής κατανομής.

$$P(490 \leq X \leq 510) = \sum_{i=490}^{510} \binom{1000}{i} (\frac{1}{2})^i (1 - \frac{1}{2})^{1000-i} = 0.49334.$$

και ομοίως $P(485 \leq X \leq 515) = 0.673063$. Δηλαδή, η προσέγγιση της X από την κανονική κατανομή είναι πολύ καλή.

Λύση (συνέχεια)

ii) Ψάχνουμε N ώστε $P(X \leq N) \geq 0.99$. Είναι

$$P(X \leq N) = P(\frac{X-500}{\sqrt{250}} \leq \frac{N-500}{\sqrt{250}}) = P(Z \leq \frac{N-500}{\sqrt{250}}) = \Phi(\frac{N-500}{\sqrt{250}}),$$

Άρα, $\frac{N-500}{\sqrt{250}} \geq 2.33 \Leftrightarrow N \geq 500 + 2.33 \cdot \sqrt{250} = 536.841$.

Άρα, πρέπει $N \geq 537$, οπότε επιλέγουμε, $N = 537$.

Μπορούμε να υπολογίσουμε με τους τύπους της διωνυμικής κατανομής τις πιθανότητες $P(X \leq 536)$ και $P(X \leq 537)$ για να δούμε αν η τιμή 537 είναι πράγματι η ελάχιστη τιμή N.

$$P(X \leq 536) = \sum_{i=0}^{536} P(X = i) = \sum_{i=0}^{536} \binom{1000}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{1000-i} = 0.989536$$

και $P(X \leq 537) = P(X \leq 536) + P(X = 537) = 0.991168$.

Δηλαδή, η τιμή 537 που βρήκαμε χρησιμοποιώντας την προσέγγιση της X από την κανονική κατανομή είναι πράγματι η ελάχιστη τιμή N.

Παράδειγμα

Έχει υπολογισθεί ότι το ποσό (σε ευρώ) που ξοδεύουν οι πελάτες ενός καταστήματος μια συγκεκριμένη ημέρα έχει μέση τιμή $\mu = 40$ και διακύμανση $\sigma^2 = 100$. Αν το κατάστημα δεχθεί 300 πελάτες, να βρεθεί η πιθανότητα τα συνολικά έσοδα να είναι τουλάχιστον 12000 ευρώ.

Λύση

Έστω X_i το ποσό που ξοδεύει ο i-οστός πελάτης, για κάθε $i = 1, 2, \ldots, 300$. Τα έσοδα του καταστήματος ισούνται με $X = X_1 + X_2 + \cdots + X_{300}$. Βάσει του κεντρικού οριακού θεωρήματος, η ΤΜ X ακολουθεί ασυμπτωτικά την κανονική κατανομή $N(300 \cdot 40, 300 \cdot 10^2) = N(12000, 3 \cdot 100^2)$. Επομένως,
$P(X \geq 12000) = P(\frac{X-12000}{100\sqrt{3}} \geq \frac{12000-12000}{100\sqrt{3}}) = P(Z \geq 0) \approx 1 - \Phi(0) = 0.5$,
όπου $Z = (X - \mu)/\sigma$.

Λυμένες ασκήσεις

Άσκηση

Το 40% των ψηφοφόρων μιας πόλης ευνοούν τον υποψήφιο A. Αν πάρουμε ένα τυχαίο δείγμα 80 ψηφοφόρων, ποια είναι η πιθανότητα να πλειοψηφούν στο δείγμα οι ευνοούντες τον A;

Λύση

Έστω X το πλήθος των ψηφοφόρων μεταξύ των 80 οι οποίοι ευνοούν τον A. Αυτοί πλειοψηφούν στο δείγμα ανν $X \geq 41$. Η X ακολουθεί την διωνυμική κατανομή με παραμέτρους $N = 80$ και $p = 0.4$, οπότε

$$P(X \geq 41) = 1 - P(X \leq 40) = 1 - \sum_{i=0}^{40} \binom{80}{i} 0.4^i 0.6^{80-i} = 0.0271236.$$

Από το ΚΟΘ, μπορούμε να προσεγγίσουμε την X από την κανονική κατανομή $N(80 \cdot 0.4, 80 \cdot 0.4 \cdot 0.6) = N(32, 19.2)$, οπότε

$$P(X \geq 41) = 1 - P(X \leq 40) = 1 - P\left(\frac{X - 32}{\sqrt{19.2}} \leq \frac{40 - 32}{\sqrt{19.2}}\right)$$

$$= 1 - P(Z \leq 1.82574) = 1 - \Phi(1.82574) = \Phi(-1.82574) = 0.0344.$$

Λυμένες ασκήσεις

Άσκηση

Ένας blogger θέλει να εμφανίζει ότι το site του έχει μεγάλη επισκεψιμότητα και χρησιμοποιεί ένα μετρητή επισκεψιμότητας, ο οποίος αυξάνει από τις πραγματικές επισκέψεις στο site αλλά και αυτόματα χρησιμοποιώντας μια γεννήτρια ψευδοτυχαίων αριθμών η οποία παράγει κάθε λεπτό, με ίση πιθανότητα, ένα φυσικό αριθμό από το 1 εως το 10, ο οποίος προστίθεται στον μετρητή επισκεψιμότητας.

i) Να βρεθεί μέσος αριθμός των εικονικών επισκέψεων στο site του blogger μέσα σε 1 μέρα.

ii) Αν σε 1 μέρα το site είχε 9000 επισκέψεις (πραγματικές και εικονικές), πόσες από αυτές είναι πραγματικές με πιθανότητα τουλάχιστον 99%;

Λύση

Έστω X_i ο αριθμός των εικονικών επισκέψεων στο i-στό λεπτό.
Η X_i ακολουθεί την (διακριτή) ομοιόμορφη κατανομή με μέση τιμή $E(X_i) = \frac{10+1}{2} = 5.5$ και διακύμανση $\frac{10^2-1}{12} = 8.25$.
Μια μέρα έχει $60 * 24 = 1440$ λεπτά, επομένως ο αριθμός X των εικονικών επισκέψεων στο site είναι ίσος με $X = X_1 + X_2 + \cdots + X_{1440}$.
i) Η μέση τιμή της X είναι

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_{1440}) = 1440 \cdot 5.5 = 7920.$$

Επιπλέον, η διακύμανση της X είναι, λόγω ανεξαρτησίας,

$$V(X) = V(X_1) + V(X_2) + \cdots + V(X_{1440}) = 1440 \cdot 8.25 = 11880.$$

Επομένως, ο μέσος αριθμός εικονικών επισκέψεων στο site του blogger σε μια μέρα είναι 7920.

Λύση (συνέχεια)

ii) Μπορούμε να προσεγγίσουμε την ΤΜ X από την κανονική κατανομή με μέση τιμή $\mu = 7920$ και διακύμανση $\sigma^2 = 11880$. Γνωρίζουμε ότι, για μια ΤΜ $Z \sim N(0,1)$, είναι
$P(|Z| \leq z) = P(-z \leq Z \leq z) = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1$, οπότε

$$P(|Z| \leq z) = 0.99 \Leftrightarrow 2\Phi(z) - 1 = 0.99 \Leftrightarrow \Phi(z) = 0.995 \Leftrightarrow z \approx 2.576.$$

Επομένως, βάσει του ΚΟΘ, είναι

$$\begin{aligned} 0.99 &\approx P(-z \leq \frac{X-\mu}{\sigma} \leq z) = P(\mu - z\sigma \leq X \leq \mu + z\sigma) \\ &= P(\mu - 280.77 \leq X \leq \mu + 280.77) = P(7639.23 \leq X \leq 8200.77), \end{aligned}$$

δηλαδή, με πιθανότητα 99% οι εικονικές επισκέψεις είναι από 7639 έως 8201, οπότε οι πραγματικές είναι από 799 έως 1361.

Λυμένες ασκήσεις

Άσκηση

Υπολογίστε την πιθανότητα ο συνολικός χρόνος ολοκλήρωσης 10 ανεξάρτητων εργασιών να είναι μεγαλύτερος του 70, όταν ο χρόνος της κάθε μιας ακολουθεί

i) εκθετική κατανομή με μέση τιμή 6,

ii) ομοιόμορφη κατανομή στο διάστημα $[0, 12]$

Λύση

Η εργασία $i \in [10]$ απαιτεί χρόνο X_i και ο συνολικός χρόνος είναι $X = X_1 + X_2 + \cdots + X_{10}$. Σύμφωνα με το ΚΟΘ, θα είναι $X \to N(10\mu, 10\sigma^2)$, όπου μ και σ^2 η μέση τιμή και διακύμανση των X_i, και $P(X > 70) \approx 1 - \Phi(\frac{70 - 10\mu}{\sigma\sqrt{10}})$.

i) Είναι $\mu = \sigma = 6$, οπότε $P(X > 70) \approx 0.3$.

ii) Είναι $\mu = 6$ και $\sigma = \sqrt{12}$, οπότε $P(X > 70) \approx 0.18$.

Παρατήρηση: Γνωρίζουμε ότι όταν $X_i \sim E(\theta)$, τότε $X \sim \Gamma(n, \theta)$, $n = 10$, οπότε, για το i) (εδώ είναι $\theta = 1/6$), μπορούμε να βρούμε ακριβώς ότι

$$P(X > 70) = 0.2727.$$

Σύμφωνα με το ΚΟΘ, η τιμή που βρήκαμε ήταν 0.299 και η απόκλιση αυτή οφείλεται στο ότι το $n = 10$ ήταν σχετικά μικρό.
Όταν $X_i \sim U(0, 12)$, τότε η X ΔΕΝ είναι ομοιόμορφη και δεν μπορούμε να υπολογίσουμε απευθείας την πιθανότητα $P(X > 70)$ εύκολα.

Λυμένες ασκήσεις

Άσκηση

Στο δημοψήφισμα του 2015, το ποσοστό του ΟΧΙ ήταν $p = 61.8\%$. Μια εταιρεία δημοσκοπήσεων έκανε δημοσκόπηση σε δείγμα $N = 1325$ ατόμων και προέβλεψε ότι το ποσοστό θα ήταν 49.8% με σφάλμα $\pm 2.7\%$.

i) Ποια είναι η πιθανότητα η δημοσκόπηση να πέσει έξω κατά $\pm 2.7\%$;

ii) Ποια είναι η πιθανότητα η δημοσκόπηση να πέσει τόσο έξω, ώστε να δώσει αποτέλεσμα μικρότερο ή ίσο του $49.8\% + 2.7\% = 52.5\%$;

Λύση

Η ψήφος X_i του i-οστού ατόμου, $i \in [N]$, είναι μια ΤΜ *Bernoulli* (1 για το ΟΧΙ, 0 για το ΝΑΙ) με παράμετρο p. Θεωρώντας για απλότητα, ότι η δειγματοληψία γίνεται με επανατοποθέτηση (κάτι που επηρεάζει ελάχιστα την ακρίβεια γιατί ο πληθυσμός είναι πολύ μεγάλος), τότε η ΤΜ $X = X_1 + \cdots + X_N$ ακολουθεί την $\mathrm{Binom}(N, p)$ και, σύμφωνα με το ΚΟΘ, είναι $\overline{X} := X/n \to N(p, p(1-p)/n)$. Η $\overline{X}$ είναι το δειγματικό ποσοστό, το οποίο αποτελεί μια εκτίμηση του ποσοστού p.

Λύση (συνέχεια)

i) Για $\sigma = \sqrt{p(1-p)/n}$ και $\epsilon = 0.027$, έχουμε ότι

$$P(|\overline{X} - p| > \epsilon) = P(|(\overline{X} - p)/\sigma| > \epsilon/\sigma) \approx 1 - \Phi(\epsilon/\sigma) + \Phi(-\epsilon/\sigma)$$
$$= 2 - 2\Phi(\epsilon/\sigma) = 0.0431.$$

Αυτή είναι η πιθανότητα να πέσει έξω η δημοσκόπηση κατά $\pm\epsilon$. Είναι τόσο μικρή, διότι το N είναι πολύ μεγάλο.
ii) Για $q = 0.525$, η πιθανότητα τόσο μεγάλης απόκλισης είναι

$$P(\overline{X} \leq q) = P\left(\frac{\overline{X} - p}{\sigma} \leq \frac{q-p}{\sigma}\right) \approx \Phi\left(\frac{q-p}{\sigma}\right) = 1.615 \cdot 10^{-12}.$$

Αν χρησιμοποιήσουμε ότι $X \sim \mathrm{Binom}(N, p)$, έχουμε ότι
$P(\overline{X} \leq q) = P(X \leq nq) = 2.75755 \cdot 10^{-12}$.
Αν δεν έχουμε επανατοποθέτηση, τότε $X \sim \mathrm{HGeom}(M, \lfloor Mp \rfloor, N)$, όπου $M = 8 \cdot 10^6$ το πλήθος των ψηφοφόρων, έχουμε ότι
$P(X \leq nq) = 2.746 \cdot 10^{-12}$.