

Πανεπιστήμιο Κρήτης - Τμήμα Επιστήμης Υπολογιστών
ΗΥ-317: Εφαρμοσμένες Στοχαστικές Διαδικασίες-Εαρινό Εξάμηνο 2016
Διδάσκων: Π. Τσακαλίδης

3ο Φροντιστήριο

Επιμέλεια: Κωνσταντίνα Φωτιάδου

1. Θεωρία: Η De-Moivre Laplace Προσέγγιση της Διωνυμικής Κατανομής

Έχουμε δει ότι η διωνυμική τυχαία μεταβλητή S_n με παραμέτρους n και p εκφράζεται ως το άθροισμα n ανεξάρτητων τυχαίων μεταβλητών Bernoulli, $X_1, \dots, X_n$ με κοινή παράμετρο p , ως: $S_n = X_1 + \dots + X_n$. Γνωρίζουμε ότι η μέση τιμή και η διασπορά θα είναι:

$$\mu = E[X_i] = p$$
$$\sigma = \sqrt{\text{var}(X_i)} = \sqrt{p(1-p)}$$

Χρησιμοποιούμε το Κεντρικό Οριακό Θεώρημα για να προσεγγίσουμε την πιθανότητα του γεγονότος: $\{k \leq S_n \leq \ell\}$, όπου k και ℓ είναι αριθμοί.

Έχουμε ότι:

$$\{k \leq S_n \leq \ell\} \Leftrightarrow \frac{k - n \cdot p}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}}$$

Σύμφωνα με το Κεντρικό Οριακό Θεώρημα, η τ.μ $\frac{S_n - np}{\sqrt{np(1-p)}}$ ακολουθεί κανονική κατανομή.

Οπότε έχουμε:

$$\begin{aligned} P(\{k \leq S_n \leq \ell\}) &= P\left(\frac{k - n \cdot p}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{\ell - np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{\ell - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

όπου είναι το ολοκλήρωμα της standard PDF από το k στο ℓ .

Στην πράξη, για λόγους προσέγγισης της $P(S_n = k)$, ολοκληρώνουμε από το $k - 1/2$ ως το $\ell + 1/2$.

Οπότε έχουμε:

Αν η S_n είναι διωνυμική τυχαία μεταβλητή με παραμέτρους n, p , και το n είναι αρκετά μεγάλο, και k, ℓ μη-αρνητικές σταθερές, τότε:

$$P(\{k \leq S_n \leq \ell\}) = \Phi\left(\frac{\ell + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}}\right)$$

Παράδειγμα 1.1

Αν η S_n είναι διωνυμική τυχαία μεταβλητή με $n = 36$, και $p = 0.5$, υπολογίστε την πιθανότητα: $P(S_n \leq 21)$

- Ακριβής Υπολογισμός:

$$P(S_n \leq 21) = \sum_{k=0}^{21} \binom{36}{k} \cdot 0.5^k \cdot 0.5^{36-k} = \sum_{k=0}^{21} \binom{36}{k} \cdot 0.5^{36} = 0.8785$$

- De Moivre-Laplace Προσέγγιση:

$$P(S_n \leq 21) = \Phi\left(\frac{21 + 0.5 - 36 \cdot 0.5}{\sqrt{36 \cdot 0.5 \cdot 0.5}}\right) = \Phi\left(\frac{21.5 - 18}{3}\right) = \Phi(1.17) = 0.879$$

Άσκηση 1.

Σύμφωνα με το Κεντρικό Οριακό Θεώρημα, αν $Z = X_1 + X_2 + \dots + X_n$, όπου οι τ.μ $X_1, \dots, X_n$ είναι ανεξάρτητες μεταξύ τους και όμοια κατανομημένες με μέση τιμή μ και διασπορά σ^2 , η κατανομή της τ.μ Z προσεγγίζεται από την κανονική κατανομή με μέση τιμή $n\mu$ και διασπορά $n\sigma^2$. Δηλαδή η τ.μ $\frac{Z - n\mu}{\sigma \cdot \sqrt{n}}$ είναι μια τυπική κανονική κατανομή: $\frac{Z - n\mu}{\sigma \cdot \sqrt{n}} \sim N(0, 1)$.

Ο μηχανικός της υπηρεσίας συστημάτων και δικτύων του Υπολογιστικού Κέντρου του πανεπιστημίου Κρήτης, θέλει να εκτιμήσει το ποσοστό p , των πακέτων πληροφορίας που ταξιδεύουν μέσα από την οπτική ίνα που συνδέει την Κνωσσό με τις Βούτες, τα οποία πακέτα είναι ψηφιακού video (digital video disk (DVD)). Ο μηχανικός γράφει ένα script, για να ελέγξει $n = 1000$ πακέτα, μετράει το πλήθος S_n των DVD και χρησιμοποιεί τη σχέση: $\hat{p} = \frac{S_n}{1000}$ ως τον εκτιμητή του p . Κάθε πακέτο είναι ένα DVD πακέτο με πιθανότητα p , ανεξάρτητο από τα άλλα.

Να υπολογισθούν:

(α) Τι είδους τυχαία μεταβλητή είναι η S_n ;

(β) Βάσει του Κεντρικού Οριακού Θεωρήματος, ποια είναι η κανονική προσέγγιση για την τυχαία μεταβλητή $\hat{p} = \frac{S_n}{1000}$;

(γ) Ποια είναι η πιθανότητα ο εκτιμητής p να κάνει λάθος λιγότερο από 2%;

(δ) Βρείτε τον αριθμό δ για τον οποίο ισχύει: $P(\hat{p} - p < \delta) \approx 0.99$, όταν $p = 0.5$, και όταν $p = 0.1$. Δίνεται ότι: $\Phi(2.58) = 0.995$.

Λύση

(α) Η τυχαία μεταβλητή S_n είναι της μορφής: $S_n = X_1 + \dots + X_{1000}$, όπου η X_i είναι Bernoulli τυχαία μεταβλητή, με $X_i = 1$, όταν το i -οστό πακέτο είναι DVD πακέτο.

$$X_i = \begin{cases} 1, & p \\ 0, & 1 - p \end{cases}$$

Επομένως, η S_n είναι διωνυμική με παραμέτρους $n = 1000$ και p .

(β) Η τυχαία μεταβλητή S_n έχει διασπορά $n \cdot p(1 - p)$, $var(S_n) = n \cdot p(1 - p)$. Επομένως, η τ.μ $\hat{p} = \frac{S_n}{n}$ έχει διασπορά:

$$var(\hat{p}) = \frac{1}{n^2} var(S_n) = \frac{1}{n^2} \cdot n \cdot p \cdot (1 - p) = \frac{p(1 - p)}{n}$$

και μέση τιμή:

$$E[\hat{p}] = \frac{E[S_n]}{n} = \frac{n \cdot p}{n} = p$$

(γ) Η ζητούμενη πιθανότητα γίνεται:

$$\begin{aligned}
 P(|\hat{p} - p| \leq \delta) &= P(-\delta \leq \hat{p} - p \leq \delta) = P\left(\frac{-\delta}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\delta}{\sqrt{\frac{p(1-p)}{n}}}\right) \\
 &= P\left(\frac{-\delta \cdot \sqrt{n}}{\sqrt{p(1-p)}} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \frac{\delta \cdot \sqrt{n}}{\sqrt{p(1-p)}}\right) \\
 &\approx \Phi\left(\frac{\delta \cdot \sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{-\delta \cdot \sqrt{n}}{\sqrt{p(1-p)}}\right) \\
 &= \Phi\left(\frac{\delta \cdot \sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 + \Phi\left(\frac{\delta \cdot \sqrt{n}}{\sqrt{p(1-p)}}\right) \\
 &= 2 \cdot \Phi\left(\frac{\delta \cdot \sqrt{n}}{\sqrt{p(1-p)}}\right) - 1
 \end{aligned}$$

- Για $p = 0.05$, $\delta = 0.02$ έχουμε:

$$P(|\hat{p} - 0.5| \leq 0.02) = 2 \cdot \Phi\left[\frac{0.02 \cdot \sqrt{1000}}{\sqrt{0.5 \cdot 0.5}}\right] - 1 = 2 \cdot \Phi(1.265) - 1 \approx 0.794 = 79.4\%$$

- Για $p = 0.1$, $\delta = 0.02$ έχουμε:

$$P(|\hat{p} - 0.1| \leq 0.02) = 2 \cdot \Phi\left[\frac{0.02 \cdot \sqrt{1000}}{\sqrt{0.1 \cdot 0.9}}\right] - 1 \approx 0.965 = 96.5\%$$

Παρατηρούμε ότι όσο πιο μικρό είναι το ποσοστό των πακέτων πληροφορίας, τόσο μεγαλύτερη είναι η πιθανότητα ο εκτιμητής να κάνει λάθος λιγότερο από 2%.

(δ) Έχουμε ότι:

$$\begin{aligned}
 2 \cdot \Phi\left(\frac{\delta \sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 &= 0.99 \\
 \Phi\left(\frac{\delta \sqrt{n}}{\sqrt{p(1-p)}}\right) &= \frac{1.99}{2} \\
 \Phi\left(\frac{\delta \sqrt{n}}{\sqrt{p(1-p)}}\right) &= 0.995 \\
 \Phi\left(\frac{\delta \sqrt{n}}{\sqrt{p(1-p)}}\right) &= \Phi(2.58) \\
 \delta &= 2.58 \cdot \sqrt{\frac{p(1-p)}{1000}}
 \end{aligned}$$

- Για $p = 0.5$, $\delta = 2.58 \cdot \sqrt{\frac{0.5 \cdot 0.5}{1000}} = 0.04$

- Για $p = 0.1$, $\delta = 2.58 \cdot \sqrt{\frac{0.1 \cdot 0.9}{1000}} = 0.0025$

Άσκηση 2.

Κατά τη διάρκεια της καλοκαιρινής πρακτικής άσκησης σας, εργάζεστε σ'ένα εργοστάσιο παραγωγής μικρο-επεξεργαστών. Ο επιβλέπων σας, ζητά να εκτιμήσετε την ποιότητα της γραμμής παραγωγής, δηλαδή την πιθανότητα p ένας μικρο-επεξεργαστής να μην είναι ελαττωματικός. Υποθέτουμε ότι οι μικρο-επεξεργαστές είναι καλοί με την ίδια πιθανότητα p , και ανεξάρτητοι ο ένας από τον άλλο.

(α) Υποθέτοντας ότι τεστάρετε ένα σύνολο από επιλεγμένους μικρο-επεξεργαστές, ορίστε μια εκτιμήτρια Z_n , για την παράμετρο p , ώστε η Z_n να συγκλίνει στο p ως προς την πιθανότητα.

(β) Ο επιβλέπων σας ζητά η εκτιμήτρια Z_n , να βρίσκεται στη γειτονιά τιμών $p \pm 0.1$, με πιθανότητα 0.95. Αρκεί να τεστάρετε 27 επιλεγμένους μικρο-επεξεργαστές για να ικανοποιήσετε αυτήν την απαίτηση;

Χρησιμοποιήστε τόσο την ανισότητα *Chebyshev*, όσο και το Κεντρικό Οριακό Θεώρημα, για να απαντήσετε στο ερώτημα.

Λύση

(α) Έστω X_i τ.μ. που χαρακτηρίζει την ποιότητα του i -οστού επεξεργαστή (0: για ελαττωματικός, 1: για μη-ελαττωματικός). Οι X_i , $i = 1, \dots, n$ είναι ανεξάρτητες *Bernoulli* τ.μ.

Η Z_n ορίζεται ως:

$$Z_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Η ανισότητα του *Chebyshev* ορίζεται ως:

$$P(|Z_n - p| \geq \epsilon) \leq \frac{\sigma^2}{n \cdot \epsilon^2}$$

Οπότε έχουμε:

$$\lim_{n \rightarrow \infty} P(|Z_n - p| \geq \epsilon) = 0,$$

Διότι το όριο:

$$\lim_{n \rightarrow \infty} \frac{\sigma^2}{n \cdot \epsilon^2} = 0$$

Οπότε, αυτό σημαίνει ότι η Z_n συγκλίνει ως προς την πιθανότητα.

(β)

Χρησιμοποιώντας την ανισότητα του Chebyshev:

Κάνοντας χρήση της ανισότητας Chebyshev, μπορούμε να δείξουμε ότι πρέπει να τεστάρουμε περισσότερους από 500 μικρο-επεξεργαστές.

Εφόσον η διασπορά μιας *Bernoulli* τ.μ είναι: $p \cdot (1 - p)$, ισχύει ότι: $\sigma^2 \leq \frac{1}{4}$.

Επομένως, για $n \geq 500$ έχουμε:

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - p\right| \geq 0.1\right) \leq \frac{\sigma}{n \cdot 0.1^2} = \frac{1/4}{500 \cdot 0.01} = 0.05$$

Παρ' όλα αυτά, χρησιμοποιώντας την ανισότητα του *Chebyshev* δε μπορούμε να πούμε αν λιγότεροι από 500 μικρο-επεξεργαστές επαρκούν για το τεστ, μιας και δε γνωρίζουμε τη διασπορά. Αν για παράδειγμα η διασπορά είναι μικρή, το οποίο μπορεί να συμβεί όταν το p είναι αρκετά μικρό, 27 μικρο-επεξεργαστές θα μπορούσαν να επαρκούν στην πράξη. Συνεπώς, η απάντηση στο ερώτημα είναι: "εξαρτάται"!

Χρησιμοποιώντας το Κεντρικό Οριακό Θεώρημα έχουμε:

Καθώς $n \rightarrow \infty$, η $Z_n = \frac{X_1 + X_2 + \dots + X_n}{n}$, συμπεριφέρεται σαν κανονικά κατανοημένη με μέση τιμή:

$$E[Z_n] = \frac{n \cdot p}{n} = p$$

και διασπορά:

$$\text{var}(Z_n) = \frac{1}{n^2} \cdot p \cdot (1 - p) \cdot n = \frac{p \cdot (1 - p)}{n} = \frac{\sigma^2}{n}$$

Θέλουμε να βρούμε τέτοιο n ώστε:

$$\begin{aligned} P(p - 0.1 \leq Z_n \leq p + 0.1) &\geq 0.95 \Leftrightarrow \\ P(-0.1 \leq Z_n - p \leq 0.1) &\geq 0.95 \Leftrightarrow \\ P\left(\frac{-0.1}{\sigma/\sqrt{n}} \leq \frac{Z_n - p}{\sigma/\sqrt{n}} \leq \frac{0.1}{\sigma/\sqrt{n}}\right) &\geq 0.95 \approx \Phi\left(\frac{0.1}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-0.1}{\sigma/\sqrt{n}}\right) \\ &= 2 \cdot \Phi\left(\frac{0.1}{\sigma/\sqrt{n}}\right) - 1 \end{aligned}$$

Στη χειρότερη περίπτωση, η διασπορά της X_i είναι μεγάλη, οπότε απαιτούνται περισσότερα δείγματα n για να ικανοποιήσουμε το κριτήριο.

Μέγιστη διασπορά έχουμε για $p = \frac{1}{2} \Leftrightarrow \sigma^2 = p \cdot (1 - p) = \frac{1}{2} \cdot (1 - \frac{1}{2}) = \frac{1}{4}$

Άρα, στη χειρότερη περίπτωση έχουμε:

$$\begin{aligned} P(p - 0.1 \leq Z_n \leq p + 0.1) &\geq 0.95 \Leftrightarrow \\ 2 \cdot \Phi\left(\frac{0.1}{\sigma/\sqrt{n}}\right) - 1 &= 0.95 \\ 2 \cdot \Phi\left(\frac{0.1 \cdot \sqrt{n}}{1/2}\right) - 1 &= 0.95 \\ 2 \cdot \Phi(0.2 \cdot \sqrt{n}) - 1 &= 0.95 \\ \Phi(0.2 \cdot \sqrt{n}) &= 0.9750 \\ \Phi(0.2 \cdot \sqrt{n}) &= \Phi(1.96) \\ 0.2 \cdot \sqrt{n} &= 1.96 \\ \sqrt{n} &= 9.8 \Leftrightarrow n = 96.04 \end{aligned}$$

Επομένως, χρειαζόμαστε περισσότερες από 96 προσπάθειες. $n \geq 96$

Άσκηση 3.

Ψηφοφόροι κάποιας περιφέρειας ψήφισαν για έναν από τους υποψηφίους Α και Β. Η αρχική καταμέτρηση έδειξε ότι ο υποψήφιος Β προηγείται του Α, κατά 537 ψήφους. Όμως, υπάρχουν 25.600 ψηφοδέλτια τα οποία εκκρεμούν στο εκλοδικείο και δεν έχουν καταχωρηθεί σε κανέναν από τους δύο υποψηφίους. Υποθέστε ότι τελικά και τα 25.600 ψηφοδέλτια, θα βρεθούν έγκυρα και θα αποδοθούν ισοπίθانا στους δύο υποψηφίους, κάθε ένα ανεξάρτητα από τα υπόλοιπα. Έστω X και Y το πλήθος των ψηφοδελτίων (από τα 25.600) που θα αποδοθούν στους Α και Β αντίστοιχα.

(α) Τι είδους μεταβλητή είναι η X ;

(β) Να βρεθούν η μέση τιμή και η διασπορά της τ.μ X .

(γ) Είναι οι τ.μ X και Y ανεξάρτητες;

(δ) Από τη στιγμή που και οι 25.600 ψήφοι προσμετρηθούν στο τελικό αποτέλεσμα, ποια είναι η μέση τιμή και η διασπορά της τ.μ $Z = X + Y$, του καθαρού κέρδους του υποψηφίου Α στην ψηφοφορία;

(ε) Από τη στιγμή που και οι 25.600 ψήφοι προσμετρηθούν στο τελικό αποτέλεσμα, χρησιμοποιήστε το Κεντρικό Οριακό Θεώρημα για να βρείτε την πιθανότητα: "Να κερδίσει ο Β την εκλογή".

Λύση

(α) Η τυχαία μεταβλητή X είναι διωνυμική με παραμέτρους: $(n, p) = (25.600, 1/2)$

(β) Η μέση τιμή και η διασπορά της τ.μ X γίνονται:

$$E[X] = n \cdot p = 25.600 \cdot \frac{1}{2} = 12.800$$

και:

$$\text{var}(X) = n \cdot p^2 = 25.600 \cdot \left(\frac{1}{2}\right)^2 = 6.400$$

(γ) Έχουμε ότι: $X + Y = 25.600 \Rightarrow$ Άρα, οι τ.μ X, Y δεν μπορούν να είναι ανεξάρτητες.

(δ) Έχουμε ότι: $Z = X + Y$. Όμως: $X + Y = 25.600 \Rightarrow Y = 25.600 - X$.

Η μέση τιμή της τ.μ Z γίνεται:

$$E[Z] = E[X] - E[Y] = E[X] - E[25.600 - X] = 2 \cdot E[X] - 25.600 = 2 \cdot 12.800 - 25.600 = 0$$

Για τη διασπορά της τ.μ Z έχουμε:

$$\text{var}(Z) = \text{var}(X - Y) = \text{var}(X - 25.600 + X) = \text{var}(2X - 25.600) = 4 \cdot \text{var}(X) = 4 \cdot 6.400 = 25.600$$

(ε) Η Z μπορεί να προσεγγιστεί από μια κανονική τ.μ με $Z \sim N(0, 25.600)$. Σύμφωνα με το Κεντρικό Οριακό Θεώρημα, η πιθανότητα γίνεται:

$$P(\text{"Ο Β κερδίζει"}) = P(Z \leq 536) \approx \Phi\left(\frac{536 - 0}{\sqrt{25.600}}\right) = \Phi(3.35) \approx 0.9996$$