

Φροντιστήριο 2

Άσκηση 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) Ποια είναι η μέση τιμή και η διασπορά της τυχαίας μεταβλητής $\hat{p} = \frac{S_n}{1000}$;
- (γ) Ποια είναι η πιθανότητα ο εκτιμητής p να κάνει λάθος λιγότερο από 2%;
- (δ) Βρείτε τον αριθμό δ για τον οποίο ισχύει: $P(|\hat{p} - p| \leq \delta) \approx 0.99$, 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 $var(S_n) = np(1 - p)$. Επομένως, η τ.μ $\hat{p} = \frac{S_n}{n}$ έχει διασπορά:

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

και μέση τιμή:

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

- (γ) Η ζητούμενη πιθανότητα γίνεται:

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

Για $p = 0.5$, $\delta = 0.02$ 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:

$$2 \cdot \Phi\left(\frac{\delta\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 = 0.99$$

$$\Phi\left(\frac{\delta\sqrt{n}}{\sqrt{p(1-p)}}\right) = \frac{1.99}{2}$$

$$\Phi\left(\frac{\delta\sqrt{n}}{\sqrt{p(1-p)}}\right) = 0.995$$

$$\Phi\left(\frac{\delta\sqrt{n}}{\sqrt{p(1-p)}}\right) = \Phi(2.58)$$

$$\delta = 2.58 \cdot \sqrt{\frac{p(1-p)}{1000}}$$

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

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

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. Αρχεί 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.
Ορίζουμε Z_n ως:

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

Η ανισότητα του Chebyshev ορίζεται ως:

$$P(|Z_n - E[Z_n]| \geq \epsilon) \leq \frac{\text{var}(Z_n)}{\epsilon^2}.$$

Εφόσον $E[Z_n] = \frac{n \cdot E[X_i]}{n} = \frac{n \cdot p}{n} = p$ και $\text{var}(Z_n) = \frac{n \cdot \text{var}(X_i)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$, αφού οι X_i είναι ανεξάρτητες τ.μ. Bernoulli, έχουμε ότι:

$$P(|Z_n - p| \geq \epsilon) \leq \frac{\frac{p(1-p)}{n}}{\epsilon^2},$$

άρα

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

αφού

$$\lim_{n \rightarrow \infty} \frac{p(1-p)}{n \cdot \epsilon^2} = 0.$$

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

- (β) **Χρησιμοποιώντας την ανισότητα του Chebyshev:**

Εφόσον θέλουμε η Z_n , να βρίσκεται στη γειτονιά τιμών $p \pm 0.1$, με πιθανότητα 0.95, θέλουμε

$$P(|Z_n - p| \geq 0.1) = 1 - 0.95 = 0.05$$

Επίσης, εφαρμόζοντας την ανισότητα Chebyshev για $\epsilon = 0.1$ προκύπτει ότι:

$$P(|Z_n - p| \geq 0.1) \leq \frac{p(1-p)}{n \cdot 0.1^2} \leq \frac{1/4}{n \cdot 0.01}$$

αφού η διασπορά $\text{var}(X_i) = p(1-p)$ της X_i μεγιστοποιείται στο $p = \frac{1}{2}$, και άρα $p(1-p) \leq \frac{1}{4}$.
Άρα τελικά

$$0.05 \leq \frac{1/4}{n \cdot 0.01}$$

και

$$n \leq 500.$$

Επομένως, 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:

Καθώς $n \rightarrow \infty$, η $Z_n = \frac{X_1 + X_2 + \dots + X_n}{n}$, συμπεριφέρεται σαν κανονικά κατανομημένη με μέση τιμή $E[Z_n] = \frac{n \cdot p}{n} = p$ και διασπορά $var(Z_n) = \frac{1}{n^2} \cdot p \cdot (1 - p) \cdot n = \frac{p \cdot (1-p)}{n} = \frac{\sigma^2}{n}$.
Θέλουμε να βρούμε τέτοιο n ώστε:

$$P(|Z_n - p| \leq 0.1) = 0.95$$

$$P(-0.1 \leq Z_n - p \leq 0.1) = 0.95$$

$$P\left(-\frac{0.1}{\sigma/\sqrt{n}} \leq \frac{Z_n - p}{\sigma/\sqrt{n}} \leq \frac{0.1}{\sigma/\sqrt{n}}\right) = 0.95$$

$$\Phi\left(\frac{0.1}{\sigma/\sqrt{n}}\right) - \Phi\left(-\frac{0.1}{\sigma/\sqrt{n}}\right) = 2 \cdot \Phi\left(\frac{0.1}{\sigma/\sqrt{n}}\right) - 1 = 0.95$$

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

Έχουμε ότι $\sigma^2 = p(1-p) \leq \frac{1}{4}$.

Άρα, στη χειρότερη περίπτωση έχουμε:

$$2 \cdot \Phi\left(\frac{0.1}{\sigma/\sqrt{n}}\right) - 1 = 0.95$$

$$2 \cdot \Phi\left(\frac{0.1 \cdot \sqrt{n}}{1/2}\right) - 1 = 0.95$$

$$2 \cdot \Phi(0.2 \cdot \sqrt{n}) - 1 = 0.95$$

$$\Phi(0.2 \cdot \sqrt{n}) = 0.9750$$

$$\Phi(0.2 \cdot \sqrt{n}) = \Phi(1.96)$$

$$0.2 \cdot \sqrt{n} = 1.96$$

$$\sqrt{n} = 9.8 \Leftrightarrow n = 96.04$$

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

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:

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

και

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

- (γ) Έχουμε ότι: $X + Y = 25.600$. 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:

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

Άσκηση 4

Ένα δίκαιο εξάεδρο ζάρι ρίχνεται n φορές. Έστω X_i , το αποτέλεσμα της i -οστής ρίψης. Υποθέτουμε ότι οι τ.μ X_i είναι ανεξάρτητες και ορίζουμε το άθροισμα των ρίψεων: $S_n = X_1 + X_2 + \dots + X_n$. Υπολογίστε ένα άνω όριο για την πιθανότητα $P(S_{100} \geq 400)$.

- (α) Χρησιμοποιώντας την ανισότητα Markov.
- (β) Χρησιμοποιώντας την ανισότητα Chebyshev.
- (γ) Χρησιμοποιώντας προσέγγιση με βάση το κεντρικό οριακό θεώρημα.

Τι παρατηρείτε;

Λύση

(α) Η μέση τιμή των τ.μ X_i γίνεται:

$$E[X_i] = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5$$

Σύμφωνα με την ανισότητα Markov έχουμε:

$$P[S_{100} \geq 400] \leq \frac{100 \cdot 3.5}{400}$$

(β) Υπολογίζουμε τη διασπορά των τ.μ X_i ως εξής:

$$\text{var}(X_i) = E[X_i^2] - (E[X_i])^2$$

Όμως, έχουμε ότι:

$$E[X_i^2] = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} = 15$$

Επομένως, η διασπορά γίνεται:

$$\text{var}(X_i) = E[X_i^2] - (E[X_i])^2 = 15 - 3.5^2 = 2.75$$

Σύμφωνα με την ανισότητα του Chebyshev έχουμε:

$$P[S_{100} \geq 400] = P[|S_{100} - 350| \geq 50] \leq \frac{100 \cdot 2.75}{50^2} = 0.11$$

(γ) Με βάση το Κεντρικό Οριακό Θεώρημα έχουμε:

$$\begin{aligned} P[S_{100} \geq 400] &= 1 - P[S_{100} \leq 400] = 1 - P\left[\frac{S_{100} - 350}{\sqrt{2.75 \cdot 100}} \leq \frac{400 - 350}{\sqrt{2.75 \cdot 100}}\right] \\ &\approx 1 - \Phi\left(\frac{400 - 350}{\sqrt{2.75 \cdot 100}}\right) \\ &\approx 1 - \Phi(3.015) \\ &\approx 0.0009 \end{aligned}$$

Άσκηση 5

Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες τ.μ. Bernoulli με πιθανότητα επιτυχίας $p = 0.2$.

(α) Έστω η τ.μ. $S_n = X_1 + X_2 + \dots + X_n$, για $n = 40$. Δημιουργείστε $N = 10000$ τυχαία δείγματα της τ.μ. S_n .

(β) Θεωρείστε το δειγματικό μέσο όρο τους $M_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{S_n}{n}$ και απεικονίστε το ιστόγραμμα του για 20 bins. Τι παρατηρείτε;

(γ) Ορίστε την τ.μ. $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ και δώστε το ιστόγραμμα της για 20 bins. Τι παρατηρείτε;

Λύση

(α) Η τ.μ. $S_n = X_1 + X_2 + \dots + X_n$ θα ακολουθεί τη Διωνυμική κατανομή με παραμέτρους $n = 40$ και $p = 0.2$.

Δημιουργούμε $N = 10000$ τυχαία δείγματα της Διωνυμικής τ.μ. S_n , χρησιμοποιώντας τη συνάρτηση `binornd` της Matlab, όπως φαίνεται στον κώδικα παρακάτω.

```
% Initialize variables
```

```
N = 10000;
```

```
n = 40;
```

```
p = 0.2;
```

```
n_bins = 20;
```

```
% a) Binomial random variable Sn = X1+X2+...+Xn
```

```
Sn = binornd(n,p,N,1);
```

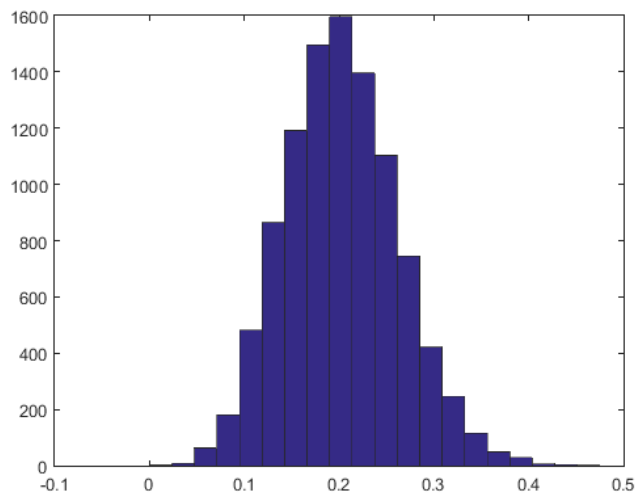
(β) Υπολογίζουμε το δειγματικό μέσο όρο $M_n = \frac{S_n}{n}$ των δειγμάτων που δημιουργήσαμε, όπως φαίνεται στον παρακάτω κώδικα.

```
% b) Mn
```

```
Mn = Sn/n;
```

```
figure; hist(Mn,n_bins);
```

Στο σχήμα 1 φαίνεται το ιστόγραμμα του για 20 bins. Όπως περιμέναμε από τη θεωρία, η M_n ακολουθεί κανονική κατανομή γύρω από τη μέση τιμή $\mu = p = 0.2$ των X_i .

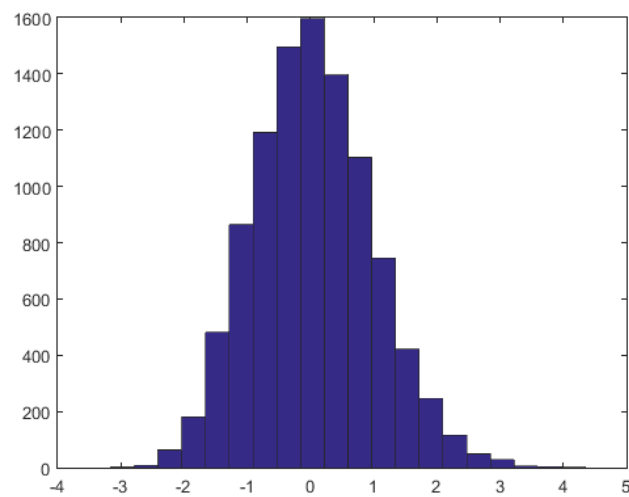


Σχήμα 1: Ιστόγραμμα του δειγματικού μέσου όρου M_n .

(γ) Υπολογίζουμε την τ.μ. $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$, με μέση τιμή $\mu = p$ και διασπορά $\sigma^2 = p(1-p)$ των τ.μ. Bernoulli X_i , όπως φαίνεται και στον παρακάτω κώδικα.

```
% c) Random Variable Zn = (Sn-n*mu)/(sigma*sqrt(n))
mu = p;
sigma = sqrt(p*(1-p));
Zn = (Sn-n*mu)/(sigma*sqrt(n));
figure; hist(Zn,n_bins);
```

Στο σχήμα 2 φαίνεται το ιστόγραμμα του για 20 bins. Σύμφωνα και με το θεώρημα κεντρικού ορίου, παρατηρούμε ότι η Z_n συγκλίνει στην τυπική κανονική κατανομή με μέση τιμή 0 και διασπορά 1.



Σχήμα 2: Ιστόγραμμα της τ.μ. Z_n .