

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
Τμήμα Επιστήμης Υπολογιστών

HY-317: Εφαρμοσμένες Στοχαστικές Διαδικασίες - Εαρινό Εξάμηνο 2023

Φροντιστήριο 2

Άσκηση 1

Τα μήλα μιας δενδροκαλλιέργειας έχουν μέσο βάρος $\mu = 50\text{gr}$ με τυπική απόκλιση $\sigma = 10\text{gr}$. Τα μήλα συσκευάζονται σε κιβώτια και προωθούνται στην αγορά. Σε κάθε κιβώτιο τοποθετούνται 400 τυχαία επιλεγμένα μήλα. Μπορούμε να υπολογίσουμε κατά προσέγγιση ποιο ποσοστό των κιβωτίων περιέχει μήλα με μέσο βάρος: μεγαλύτερο των 51.25gr ;

Λύση

Τα μήλα κάθε κιβωτίου είναι ένα τυχαίο δείγμα μεγέθους $n = 400$ από τον πληθυσμό των μήλων της παραγωγής. Η μορφή της κατανομής των βαρών των μήλων δεν μας είναι γνωστή. Μπορεί να είναι οποιαδήποτε. Για την άγνωστη αυτή κατανομή γνωρίζουμε μόνο τη μέση τιμή της $\mu = 50\text{gr}$ και την τυπική απόκλιση της $\sigma = 10\text{gr}$. Τα ερωτήματά μας μπορούν να επαναδιατυπωθούν ως εξής: Ποιο κατά προσέγγιση ποσοστό των δειγματικών μέσων είναι μεγαλύτερο των 51.25gr .

Για να μπορέσουμε να απαντήσουμε στα ερωτήματα που θέσαμε πρέπει να γνωρίζουμε την κατανομή των δειγματικών μέσων. Το Κ.Ο.Θ. μας βεβαιώνει ότι, παρότι δε γνωρίζουμε την κατανομή των βαρών των μήλων, εντούτοις, γνωρίζουμε την κατανομή των δειγματικών μέσων αφού τα δείγματά μας έχουν μέγεθος αρκετά μεγάλο. Δηλαδή, αρκεί μόνο, το ότι γνωρίζουμε τη μέση τιμή και την τυπική απόκλιση της κατανομής των βαρών των μήλων.

Έτσι, αν συμβολίσουμε με X την τυχαία μεταβλητή που εκφράζει το βάρος ενός τυχαία επιλεγμένου μήλου της καλλιέργειας, και με $\bar{X}$ την τυχαία μεταβλητή που εκφράζει τους δειγματικούς μέσους, το Κ.Ο.Θ. μας βεβαιώνει ότι η $\bar{X}$ ακολουθεί κατά προσέγγιση κανονική κατανομή με μέση τιμή $\mu_{\bar{X}} = \mu = 50\text{gr}$ και διασπορά $\sigma_{\bar{X}}^2 = \sigma^2/n = 0.25\text{gr}^2$. Αφού, πλέον, γνωρίζουμε ότι $\bar{X} \sim N(50, 0.5^2)$, η απάντηση στα ερωτήματά μας είναι πολύ απλή. Ζητάμε τις πιθανότητες: $P(X > 51.25)$. Έτσι έχουμε:

$$P(\bar{X} > 51.25) = 1 - P(\bar{X} \leq 51.25) = 1 - P(Z \leq \frac{51.25 - 50}{\sqrt{0.25}}) = 1 - \Phi(2.5) = 0.0062.$$

Άρα, το ποσοστό των κιβωτίων που περιέχουν μήλα με μέσο βάρος μεγαλύτερο των 51.25gr είναι κατά προσέγγιση 0.62%.

Άσκηση 2

Έστω ότι ένας παίκτης casino ρίχνει το ζάρι και ανάλογα με τον αριθμό του ζαριού που πετυχαίνει του καταχωρείται συγκεκριμένο κέρδος. Συγκεκριμένα το κέδρος που αντιστοιχεί σε κάθε μια από τις ζαριές $\{1, 2, 3, 4, 5, 6\}$ είναι $\{-1, -2, -3, 1, 2, 3\}$, αντίστοιχα. Ποιός είναι ο προσεγγιστικός υπολογισμός πιθανότητας σε 42 ρίψεις να έχει συνολικό κέρδος τουλάχιστον 7;

Λύση

x_i :κέρδος στην i ρήψη

$S_{42} = \sum_{i=1}^{42} x_i$:συνολικό κέρδος

$$P(S_{42} \geq 7) = P(S_{42} \geq 6.5) = P\left(\frac{S_{42} - E[S_{42}]}{\sqrt{Var[S_{42}]}} \geq \frac{6.5 - E[S_{42}]}{\sqrt{Var[S_{42}]}}\right),$$

όπου

$$E[x_i] = \frac{1}{6}(-1) + \frac{1}{6}(-2) + \frac{1}{6}(-3) + \frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 = 0$$

$$E[S_{42}] = 42E[x_i] = 42 * 0 = 0,$$

$$Var[x_i] = E[x_i^2] - E[x_i]^2 = \frac{1}{6}(-1)^2 + \frac{1}{6}(-2)^2 + \frac{1}{6}(-3)^2 + \frac{1}{6}1^2 + \frac{1}{6}2^2 + \frac{1}{6}3^2 = 28/6$$

$$Var[S_{42}] = 42Var[x_i] = 4228/6 = 14^2$$

άρα

$$P(S_{42} \geq 7) \approx P(Z \geq \frac{6.5-0}{\sqrt{14^2}}) = P(Z \geq 0.46) = 1 - \Phi(0.46)$$

Άσκηση 3

Σε μια πόλη 4000 κατοίκων, κατά μέσο όρο 10 άτομα/μέρα χρειάζονται νοσηλεία. Ποιός είναι ο προσεγγιστικός υπολογισμός ελαχίστου αριθμού κρεβατιών, ώστε η πόλη να εξυπηρετείται χωρίς διακομιδές ασθενών σε άλλες πόλεις με πιθανότητα 95%;

Λύση

$I_k = 1$ εάν ο κάτοικος k χρειάζεται νοσηλεία μια συγκεκριμένη μέρα, $I_k = 0$ εάν όχι
 $S_{4000} = \sum_{k=1}^{4000} I_k$: πλήθος κατοίκων που χρειάζεται νοσηλεία μια συγκεκριμένη μέρα

$$\begin{aligned} P(S_{4000} \leq x) &\geq 0.95 \\ P(S_{4000} \leq x + 1/2) &\geq 0.95 \\ P\left(\frac{S_{4000} - E[S_{4000}]}{\sqrt{Var[S_{4000}]}} \leq \frac{x + 1/2 - E[S_{4000}]}{\sqrt{Var[S_{4000}]}}\right) &\geq 0.95 \\ P(Z \leq \frac{x + 1/2 - 1}{\sqrt{0.99}}) &\geq 0.95 \\ \Phi\left(\frac{x - 9.5}{\sqrt{9.9}}\right) &\geq 0.95 = \Phi(1.96) \\ x &\geq 0.5 + 1.96 * \sqrt{0.99} \end{aligned}$$

δεδομένου ότι

$$E[I_k] = 10/4000$$

$$E[S_{4000}] = 4000E[I_k]$$

$$\text{Var}[I_k] = E[I_k^2] - E[I_k]^2 = 399/400^2$$

$$\text{Var}[S_{4000}] = 4000 - 399/400^2 = 3.9.$$

Άσκηση 4

Η μέση τιμή και η τυπική απόκλιση του σιδηρού του αίματος του πληθυσμού ενός χωριού είναι 120mg/100ml και 15mg/100ml αντίστοιχα. Ποιά είναι η πιθανότητα ένα τυχαίο δείγμα 50 ατόμων από τον πληθυσμό αυτόν να έχει μέση τιμή σιδηρού του αίματος:

a) μεγαλύτερη από 125 mg/100ml

b) μεταξύ 115 και 125 mg/100ml;

Λύση

$$\begin{aligned} \text{a) } P(X_n > 125) &= P\left(\frac{X_n - 120}{2.12} > \frac{125 - 120}{2.12}\right) = P(Z > 2.36) \\ &= 1 - P(Z \leq 2.36) = 1 - 0.991 = 0.009 \end{aligned}$$

$$\begin{aligned} \text{b) } P(115 < X_n < 125) &= P\left(\frac{115 - 120}{2.12} \leq \frac{X_n - 120}{2.12} \leq \frac{125 - 120}{2.12}\right) \\ &= P(-2.36 \leq Z \leq 2.36) = 0.991 - 0.009 = 0.982 \end{aligned}$$

Άσκηση 5

Το ημερήσιο εισόδημα χαρτοπαίκτη ακολουθεί την κατανομή $\text{Uniform}([-5, 5])$ σε χιλιάδες ευρώ. Να υπολογίσετε προσεγγιστικά:

a) την πιθανότητα σε 48 μέρες να κερδίσει τουλάχιστον 30 χιλιάδες ευρώ

b) το ποσό s ώστε με πιθανότητα τουλάχιστον 95% το εισόδημα 48 μερών να είναι μικρότερο ή ίσο με s .

Λύση

$P(S_n) \leq \geq x \geq y$, όπου S_n, x, y είναι οι μέρες, το ποσό και το επίπεδο βεβαιότητας αντίστοιχα

X_i είναι το εισόδημα την μέρα i

$$E[X_i] = \int_{-5}^5 x \frac{1}{b-a} dx = 0$$

$$E[S_n] = 0n = 0$$

$$\text{Var}[X_i] = E[X_i^2] - E[X_i]^2 = \int_{-5}^5 x^2 \frac{1}{b-a} dx = \frac{1}{10} \left[\frac{x^3}{3} \right]_{x=-5}^5 = 25/3$$

$$\text{Var}[S_n] = n25/3$$

a)

$$\begin{aligned}
 P(S_{48} \geq 30) &= P\left(\frac{S_{48} - E[S_{48}]}{\sqrt{\text{Var}[S_{48}]}} \geq \frac{30 - 0}{\sqrt{4825/3}}\right) \\
 &= P(Z \leq 30/20) \\
 &= 1 - \Phi(1.5)
 \end{aligned}$$

b)

$$\begin{aligned}
 P(|S_{48}| \leq s) &\geq 0.95 \\
 P(-s \leq S_{48} \leq s) &\geq 0.95 \\
 P\left(\frac{-s - E[S_{48}]}{\sqrt{\text{Var}[S_{48}]}} \leq \frac{S_{48} - E[S_{48}]}{\sqrt{\text{Var}[S_{48}]}} \leq \frac{-s - E[S_{48}]}{\sqrt{\text{Var}[S_{48}]}}\right) &\geq 0.95 \\
 P(-s/20 \leq Z \leq s/20) &\geq 0.95 \\
 \Phi(s/20) - \Phi(-s/20) &\geq 0.95 \\
 2\Phi(s/20) - 1 &\geq 0.95 \\
 \Phi(s/20) &\geq 0.975 = \Phi(1.96) \\
 s &\geq 20 * 1.96
 \end{aligned}$$

Άσκηση 6

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.
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 κατανοημένη με μέση τιμή $E[Z_n] = \frac{n \cdot p}{n} = p$ και διασπορά $\text{var}(Z_n) = \frac{1}{n^2} \cdot p \cdot (1-p) \cdot n = \frac{p(1-p)}{n} = \frac{\sigma^2}{n}$. 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 7

Ένα δίκαιο εξαέδρο 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 8

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 ιστόγραμμά 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 .

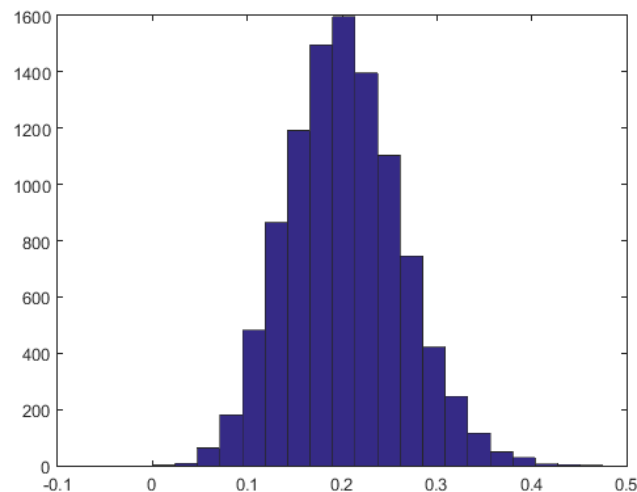


Figure 1: Ιστόγραμμα του δειγματικού μέσου όρου M_n .

(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.

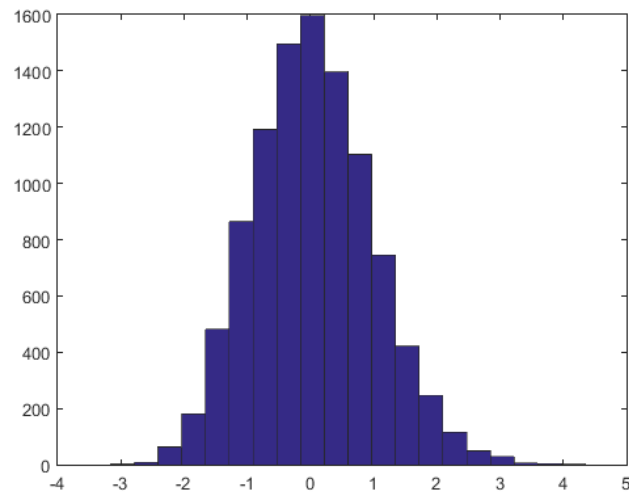


Figure 2: Ιστόγραμμα της τ.μ. Z_n .