

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
Τμήμα Επιστήμης Υπολογιστών

HY-317: Εφαρμοσμένες Στοχαστικές Διαδικασίες-Εαρινό Εξάμηνο 2023
Διδάσκων: Π. Τσακαλίδης

Λύσεις Δεύτερης Σειράς Ασκήσεων

Άσκηση 1

Τα μήλα μιας δένδροκαλλιέργειας 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 500 τυχαία επιλεγμένα μήλα. Μπορούμε να υπολογίσουμε κατά προσέγγιση ποιο ποσοστό των κιβωτίων περιέχει μήλα με μέσο βάρος:

- a) μεγαλύτερο των 50.46gr
- b) μικρότερο των 49gr;

Λύση

a) $P(\bar{X} > 50.46) = 1 - P(\bar{X} \leq 50.46) = 1 - P(Z \leq \frac{50.46-50}{\sqrt{(0.25)}}) = 1 - \Phi(0.92) = 0.178$.

Άρα, 0.25 το ποσοστό των κιβωτίων που περιέχουν μήλα με μέσο βάρος μεγαλύτερο των 51.25gr είναι κατά προσέγγιση 17.8%.

b) $P(\bar{X} < 49) = P(Z < \frac{49-50}{\sqrt{(0.25)}}) = \Phi(-2) = 1 - \Phi(2) = 0.0228$.

Άρα, το ποσοστό των κιβωτίων που περιέχουν μήλα με μέσο βάρος μικρότερο των 49gr είναι κατά προσέγγιση 2.28%.

Άσκηση 2

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 .
- c) το πλήθος των μερών που πρέπει να παίζει ώστε με πιθανότητα τουλάχιστον 95% το εισόδημά του να είναι κατά απόλυτη τιμή μικρότερο των 50 χιλιάδων ευρώ.

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)

$$\begin{aligned} P(S_{48} \geq 30) &= P\left(\frac{S_{48} - E[S_{48}]}{\sqrt{\text{Var}[S_{48}]}} \geq \frac{30 - 0}{\sqrt{48 \cdot 25/3}}\right) \\ &= P(Z \leq 30/20) \\ &= 1 - \Phi(1.5) \end{aligned}$$

b)

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

c)

$$\begin{aligned} P(|S_n| < 50) &\geq 0.95 \\ P\left(\frac{-50}{\sqrt{(25/3)n}} < \frac{S_n - E[S_n]}{\sqrt{\text{Var}[S_n]}} < \frac{50}{\sqrt{(25/3)n}}\right) &\geq 0.95 \\ P\left(\frac{-50}{\sqrt{(25/3)n}} < Z < \frac{50}{\sqrt{(25/3)n}}\right) &\geq 0.95 \\ \Phi\left(\frac{50}{\sqrt{(25/3)n}}\right) - \Phi\left(\frac{-50}{\sqrt{(25/3)n}}\right) &\geq 0.95 \\ n &\leq (10\sqrt{3}/1.96)^2 \end{aligned}$$

Άσκηση 3 Ένας στατιστικολόγος επιθυμεί να εκτιμήσει το μέσο ύψος h (σε m) ενός πληθυσμού, βασιζόμενος σε n ανεξάρτητα δείγματα $X_1, X_2, \dots, X_n$ τα οποία επιλέγονται ομοιόμορφα από ολόκληρο τον πληθυσμό. Για τον σκοπό αυτό, χρησιμοποιεί το δειγματικό μέσο $M_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ ως εκτιμητή του h , και διαλέγει τυπική απόκλιση ίση με 1 για τα δείγματα X_i .

- (i) Να υπολογίσετε το μέγεθος του δείγματος, n , έτσι ώστε η τυπική απόκλιση του M_n να είναι το πολύ $1cm$.
- (ii) Να υπολογίσετε το μέγεθος του δείγματος, n , έτσι ώστε η ανισότητα *Chebyshev* να εγγυάται ότι η εκτίμηση είναι εντός $5cm$ από το h με πιθανότητα τουλάχιστον 0.99 .

Λύση

- (i) Με βάση τη θεωρία, ξέρουμε ότι η τυπική απόκλιση του M_n δίνεται από τον εξής τύπο:

$$\sigma_{M_n} = \frac{1}{\sqrt{n}}$$

Με βάση την εκφώνηση της άσκησης, έχουμε:

$$\sigma_{M_n} \leq 0.01 \Rightarrow \frac{1}{\sqrt{n}} \leq 0.01 \Rightarrow \sqrt{n} \geq \frac{1}{0.01} \Rightarrow \sqrt{n} \geq 100 \Rightarrow n \geq 10000$$

- (ii) Με βάση την εκφώνηση της άσκησης, θέλουμε να ισχύει:

$$P(|M_n - h| \leq 0.05) \geq 0.99$$

Κάνοντας χρήση της ανισότητας *Chebyshev*, έχουμε:

$$\begin{aligned} P(|M_n - h| \leq 0.05) &\stackrel{h=E[M_n]}{=} P(|M_n - E[M_n]| \leq 0.05) \\ &= 1 - P(|M_n - E[M_n]| \geq 0.05) \\ &\stackrel{\substack{\sigma_{M_n}^2 = \frac{1}{n} \\ \geq c=0.05}}{=} 1 - \frac{\frac{1}{n}}{(0.05)^2} \end{aligned}$$

Άρα, τελικά, έχουμε:

$$\begin{aligned} P(|M_n - h| \leq 0.05) \geq 0.99 &\Rightarrow 1 - \frac{\frac{1}{n}}{(0.05)^2} \geq 0.99 \\ &\Rightarrow \frac{\frac{1}{n}}{(0.05)^2} \leq 0.01 \Rightarrow \frac{1}{n} \leq 0.000025 \\ &\Rightarrow n \geq \frac{1}{0.000025} \Rightarrow n \geq 40000 \end{aligned}$$

Άσκηση 4 Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες τ.μ. ομοιόμορφα κατανομημένες στο διάστημα $[-1, 1]$. Να αποδείξετε ότι η ακολουθία τ.μ. $Y_1, Y_2, \dots, Y_n$ συγκλίνει κατά πιθανότητα στο 0 , για κάθε μία από τις παρακάτω περιπτώσεις:

- (i) $Y_n = \frac{X_n}{n}$
- (ii) $Y_n = X_1 \cdot X_2 \cdot \dots \cdot X_n$ (**Υπόδειξη:** Χρησιμοποιείτε κατάλληλα την ανισότητα *Chebyshev*)

Λύση

Εφόσον ζητείται να αποδειχθεί ότι η ακολουθία τ.μ. $Y_1, Y_2, \dots, Y_n$ συγκλίνει κατά πιθανότητα στο 0, θα πρέπει να αποδείξουμε ότι ισχύει η σχέση:

$$\lim_{n \rightarrow +\infty} P(|Y_n - 0| \geq \epsilon) = \lim_{n \rightarrow +\infty} P(|Y_n| \geq \epsilon) = 0 \Rightarrow P(|Y_n| \geq \epsilon) \rightarrow 0$$

(i) Για κάθε $\epsilon > 0$ έχουμε:

$$P(|Y_n - 0| \geq \epsilon) = P(|Y_n| \geq \epsilon) = 0$$

για όλα τα n με $\frac{1}{n} < \epsilon$. Άρα, τελικά είναι: $P(|Y_n| \geq \epsilon) \rightarrow 0$.

(ii) Εφόσον οι $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες τ.μ., έχουμε:

$$\begin{aligned} \bullet E[Y_n] &= E[X_1] \cdot E[X_2] \cdot \dots \cdot E[X_n] = 0 \cdot 0 \cdot \dots \cdot 0 = 0 \\ \bullet VAR[Y_n] &= E[Y_n^2] - (E[Y_n])^2 = E[Y_n^2] - 0^2 = E[Y_n^2] = E[X_1^2] \cdot E[X_2^2] \cdot \dots \cdot E[X_n^2] = \\ &= \frac{(1-(-1))^2}{12} \cdot \frac{(1-(-1))^2}{12} \cdot \dots \cdot \frac{(1-(-1))^2}{12} = \frac{4}{12} \cdot \frac{4}{12} \cdot \dots \cdot \frac{4}{12} = \left(\frac{4}{12}\right)^n = \left(\frac{1}{3}\right)^n \end{aligned}$$

Από τις παραπάνω σχέσεις, παρατηρούμε ότι ισχύει: $VAR[Y_n] \rightarrow 0$.

Εφόσον όλες οι Y_n έχουν κοινή μέση τιμή (ίση με 0), κάνοντας χρήση της ανισότητας Chebyshev, έχουμε:

$$P(|Y_n - \mu_{Y_n}| \geq c) \leq \frac{\sigma_{Y_n}^2}{c^2} \xrightarrow[\sigma_{Y_n}^2 = (\frac{1}{3})^n]{\mu_{Y_n} = 0} P(|Y_n - 0| \geq c) \leq \frac{(\frac{1}{3})^n}{c^2} \xrightarrow{c=\epsilon} P(|Y_n| \geq \epsilon) \leq \frac{(\frac{1}{3})^n}{\epsilon^2}$$

Άρα, τελικά είναι: $P(|Y_n| \geq \epsilon) \rightarrow 0$.

Άσκηση 5 Σε αυτή την άσκηση σας ζητείται να προσομοιώσετε μια εφαρμογή του Κεντρικού Οριακού Θεωρήματος, χρησιμοποιώντας το Matlab. Ακολουθήστε τα παρακάτω βήματα:

- Δημιουργείτε $N = 1000$ ψευδοτυχαία δείγματα μιας ομοιόμορφης τ.μ. X στο διάστημα $[0, 1]$ και μίας εκθετικής τ.μ. Y με παράμετρο $\lambda = 1$. (**Υπόδειξη:** Χρησιμοποιείτε την συνάρτηση `random()` με κατάλληλα ορίσματα για κάθε κατανομή.)
- Απεικονίστε τα ιστογράμματα των τ.μ. X και Y , για $n_bins = 20$. (**Υπόδειξη:** Χρησιμοποιείτε την συνάρτηση `histogram()` με κατάλληλα ορίσματα.)
- Δημιουργείτε $N = 1000$ ψευδοτυχαία δείγματα από $n = 20$ ομοιόμορφες τ.μ. $X_1, X_2, \dots, X_n$ στο διάστημα $[0, 1]$ και από $n = 20$ εκθετικές τ.μ. $Y_1, Y_2, \dots, Y_n$ με παράμετρο $\lambda = 1$. (**Υπόδειξη:** Χρησιμοποιείτε την συνάρτηση `random()` με κατάλληλα ορίσματα για κάθε κατανομή.)
- Θεωρείστε τους δειγματικούς μέσους όρους $M_{X_n} = \frac{X_1 + X_2 + \dots + X_n}{n}$ και $M_{Y_n} = \frac{Y_1 + Y_2 + \dots + Y_n}{n}$ και απεικονίστε τα ιστογράμμά τους για $n_bins = 20$. Τι παρατηρείτε? (**Υπόδειξη:** Χρησιμοποιείτε τις συναρτήσεις `mean()` και `histogram()` με κατάλληλα ορίσματα.)

- (v) Ορίστε τις τ.μ. $Z_{Xn} = \frac{X_1+X_2+\dots+X_n-n\mu_X}{\sigma_X\sqrt{n}} = \frac{S_{Xn}-n\mu_X}{\sigma_X\sqrt{n}}$ και $Z_{Yn} = \frac{Y_1+Y_2+\dots+Y_n-n\mu_Y}{\sigma_Y\sqrt{n}} = \frac{S_{Yn}-n\mu_Y}{\sigma_Y\sqrt{n}}$ και δώστε τα ιστογράμμά τους για $n.bins = 20$. Άλλαξε κάτι στις κατανομές? (**Υπόδειξη**: Για μία ομοιόμορφη τ.μ. έχουμε $\mu = \frac{a+b}{2}$ και $\sigma^2 = \frac{1}{12}(a-b)^2$, όπου a, b τα όρια του δοθέντος διαστήματος. Για μία εκθετική τ.μ. έχουμε $\mu = \frac{1}{\lambda}$ και $\sigma^2 = \frac{1}{\lambda^2}$.)
- (vi) Εκτιμήστε τις συναρτήσεις πυκνότητας πιθανότητας (σππ) των δειγματικών μέσων όρων M_{Xn} και M_{Yn} , μέσω ιστογραμμάτων. (**Υπόδειξη**: Χρησιμοποιείτε την συνάρτηση `hist()` με κατάλληλα ορίσματα για κάθε κατανομή, ώστε να λάβετε τόσο τον αριθμό των στοιχείων σε κάθε bin (counts) όσο και τη θέση του κάθε bin στον άξονα x'x (centers). Ακολουθώντας, εκτιμήστε την "πυκνότητα-εμβαδόν" κάθε bin μέσω της σχέσης: $p = \frac{bincounts}{binlength \cdot N}$.)
- (vii) Εκτιμήστε τις συναρτήσεις πυκνότητας πιθανότητας (σππ) των δειγματικών μέσων όρων M_{Xn} και M_{Yn} , μέσω κανονικών (gaussian) κατανομών. (**Υπόδειξη**: Χρησιμοποιείτε την συνάρτηση `pdf()` με κατάλληλα ορίσματα για τις παραμέτρους των δειγματικών μέσων (M_{Xn} και M_{Yn}) κάθε κατανομής.)
- (viii) Απεικονίστε σε κοινά διαγράμματα της εκτιμήτριες των συναρτήσεων πυκνότητας πιθανότητας (σππ) των δειγματικών μέσων (M_{Xn} και M_{Yn}) που βρήκατε στα παραπάνω ερωτήματα. Τι παρατηρείτε? (**Υπόδειξη**: Χρησιμοποιείτε την συνάρτηση `plot()` με κατάλληλα ορίσματα.)
- (ix) "Παίξτε" με τις μεταβλητές N , $n.bins$ και n και καταγράψτε τις παρατηρήσεις σας σε κάθε περίπτωση. Τι αλλάζει στις κατανομές? Υπάρχουν αλλαγές στις διασπορές? Δώστε τα αντίστοιχα ιστογράμματα. (Π.χ. $N = 10000, n.bins = 100, n = 100$.)

Παραθέστε τον κώδικά σας, τα ζητούμενα γραφήματα και μία μικρή αναφορά με τις παρατηρήσεις σας για τα αποτελέσματα.

Λύση

Στο σχήμα απεικονίζεται το ιστόγραμμα της ομοιόμορφης τ.μ. X_1 , για $N = 1000$ δείγματα και 20 bins στο διάστημα $[0, 1]$. Στο σχήμα φαίνεται το ιστόγραμμα του δειγματικού μέσου όρου M_n , για $n = 20$ ανεξάρτητες ομοιόμορφες τ.μ. $X_1, X_2, \dots, X_n$ που μοιάζει πλέον, όπως περιμέναμε από τη θεωρία, να ακολουθεί κανονική κατανομή γύρω από τη μέση τιμή $\mu = \frac{a+b}{2} = \frac{0+1}{2} = 0.5$ των X_i . Το σχήμα αναπαριστά το ιστόγραμμα του μετασχηματισμού Z_n : Πρώτα έχουμε αφαιρέσει από την S_n , ώστε να προκύψει η τ.μ. $S_n - n\mu$ με μηδενική μέση τιμή και κατόπιν διαιρέσαμε με $\sigma\sqrt{n}$ ώστε η διασπορά να γίνει ίση με τη μονάδα. Τώρα, στο παραγόμενο ιστόγραμμα βλέπουμε στην πράξη το Κεντρικό Οριακό θεώρημα, που μας λέει ότι η ασυμπτωτική μορφή της Z_n τείνει οριακά στην Κανονική κατανομή, $Z_n \sim N(0, 1)$. Τέλος, στο σχήμα βλέπουμε τις εκτιμήσεις της pdf του δειγματικού μέσου όρου M_n με χρήση ιστογράμματος και κανονικής κατανομής, όπου γίνεται αντιληπτή η καλή προσέγγιση μεταξύ των δύο κατανομών.

Ανάλογα συμπεράσματα εξάγονται και για τις εκθετικές τ.μ. $Y_1, Y_2, \dots, Y_n$.

Αν αυξήσουμε τον αριθμό των μεταβλητών από $n = 20$ σε $n = 100$ και αφήσουμε τις υπόλοιπες παραμέτρους ως έχουν, παρατηρούμε στο παραγόμενο ιστόγραμμα ότι η διασπορά έχει μειωθεί αισθητά και η κατανομή της M_n έχει τον κύριο όγκο της πλέον πολύ κοντά στη μέση τιμή μ . Επιβεβαιώνεται έτσι η σχετική θεωρία που μας λέει ότι καθώς $n \Rightarrow \infty$, η διασπορά M_n τείνει στο $\emptyset$. Στην εικόνα, μετά το μετασχηματισμό, η διασπορά της Z_n έχει και πάλι επανέλθει στην τιμή 1 προσεγγίζοντας την κανονική κατανομή. Τέλος, στο σχήμα βλέπουμε τις εκτιμήσεις της pdf του δειγματικού μέσου όρου M_n με χρήση ιστογράμματος και κανονικής κατανομής, όπου γίνεται σαφές

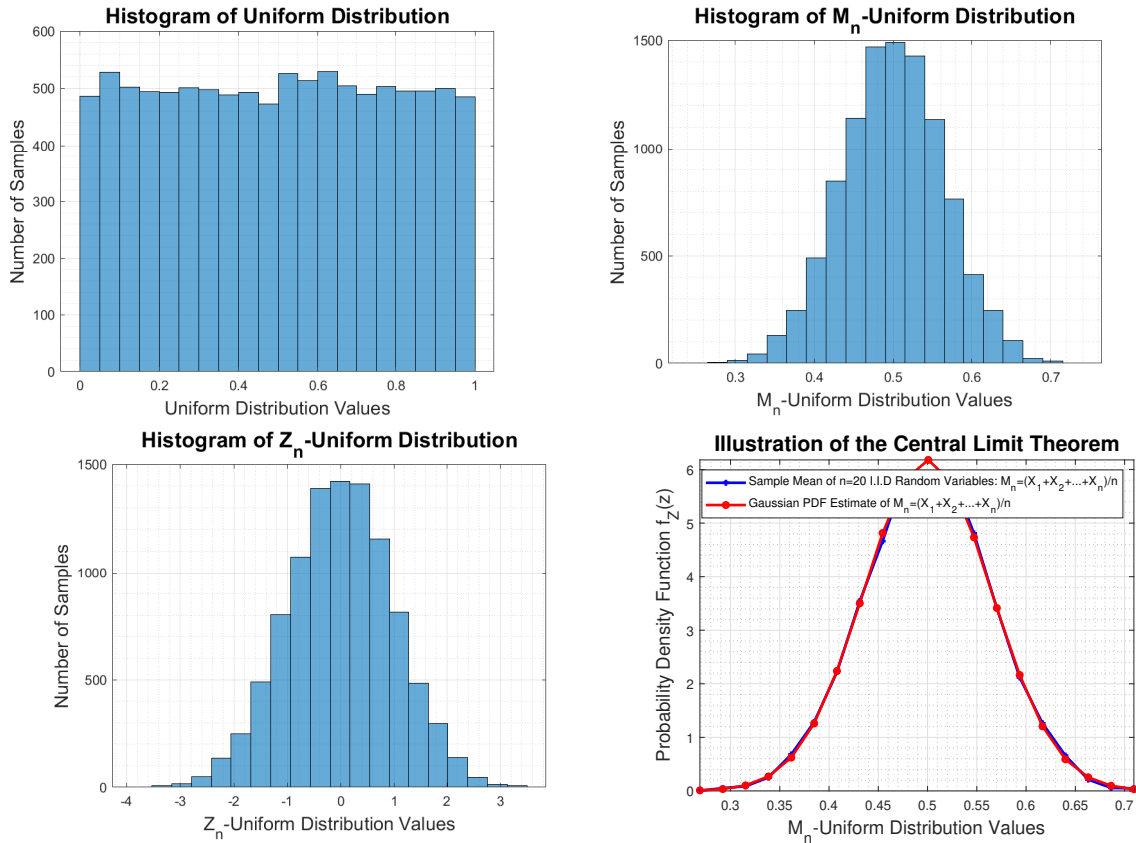


Figure 1: Τα ζητούμενα γραφήματα για την ομοιότητα T.M. X για $N = 1000$, $n = 20$, $n_{bins} = 20$.

ότι η διασπορά των κατανομών μειώνεται όσο ο αριθμός των τ.μ. αυξάνει, διατηρώντας παράλληλα την καλή προσέγγιση μεταξύ τους.

Ανάλογα συμπεράσματα εξάγονται και για τις εκθετικές τ.μ. $Y_1, Y_2, \dots, Y_n$, μέσω των αντίστοιχων σχημάτων.

Αν αυξήσουμε τον αριθμό των bins από 20 σε 100 και αφήσουμε τις υπόλοιπες παραμέτρους ως έχουν, η μόνη διαφορά που παρατηρούμε στα παραγόμενα ιστογράμματα είναι προφανώς μόνο η μείωση στη συχνότητα εμφάνισης στην εκάστοτε περιοχή τιμών (άξονας y). Τα ιστογράμματα ακολουθούν κατά τα άλλα τις ίδιες κατανομές με αυξημένο αριθμό bins, και δεν παρατηρείται καμία αλλαγή ως προς τη μέση τιμή ή τη διασπορά. Η αύξηση του αριθμού των bins από 20 σε 100 έχει ως αποτέλεσμα η αντίστοιχη εκτίμηση της pdf που προκύπτει να αποκλίνει περισσότερο από την κανονική. Τα σχετικά γραφήματα φαίνονται στα σχήματα 5 και 6 για την ομοιότητα και την εκθετική κατανομή αντίστοιχα.

Αν αυξήσουμε τον αριθμό των δειγμάτων από $N = 1000$ σε $N = 10000$ και αφήσουμε τις υπόλοιπες παραμέτρους ως έχουν, η μόνη διαφορά που παρατηρούμε στα παραγόμενα ιστογράμματα είναι προφανώς μόνο η αύξηση στη συχνότητα εμφάνισης στην εκάστοτε περιοχή τιμών (άξονας y). Δεν παρατηρείται καμία αλλαγή ως προς τη μέση τιμή ή τη διασπορά. Η αύξηση του αριθμού των δειγμάτων από $N = 1000$ σε $N = 10000$ έχει ως αποτέλεσμα η αντίστοιχη εκτίμηση της pdf που προκύπτει να συγκλίνει περισσότερο στην κανονική. Τα σχετικά γραφήματα φαίνονται στα σχήματα 7 και 8 για την ομοιότητα και την εκθετική κατανομή αντίστοιχα.

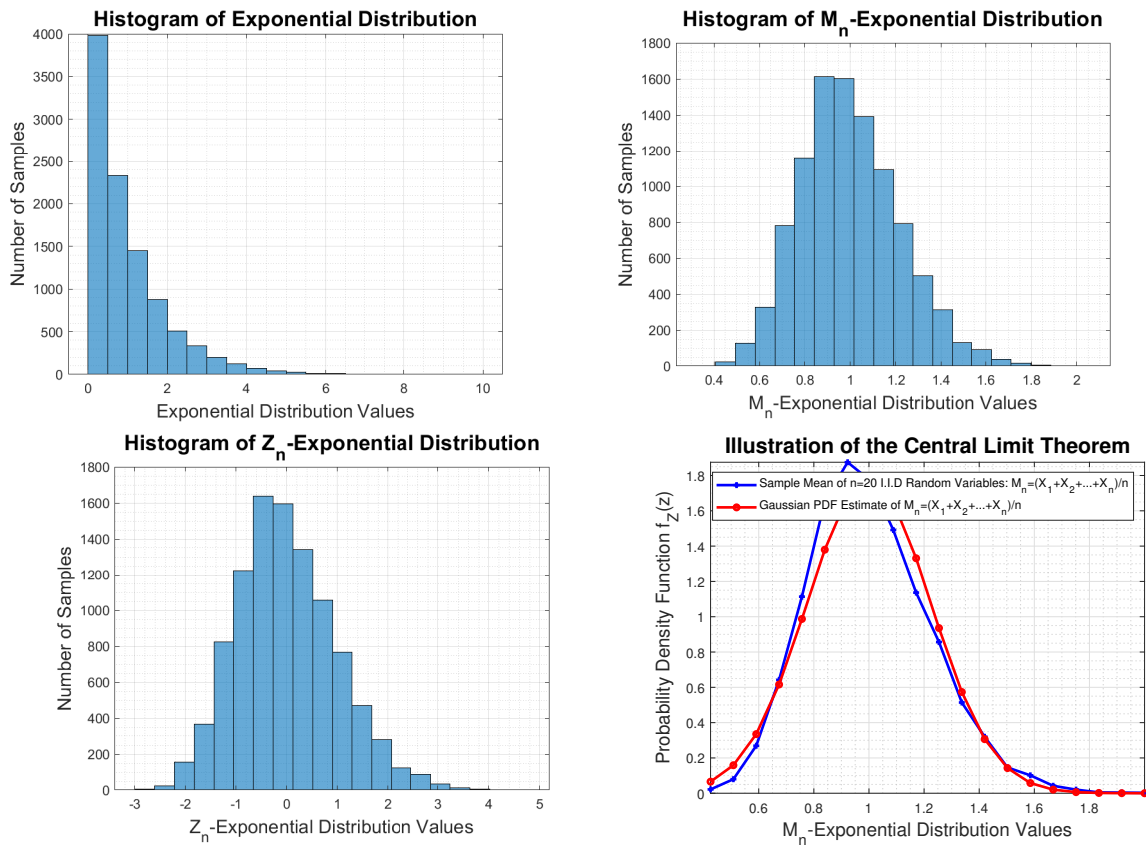


Figure 2: Τα ζητούμενα γραφήματα για την εκθετική T.M. Y για $N = 1000$, $n = 20$, $n_{bins} = 20$.

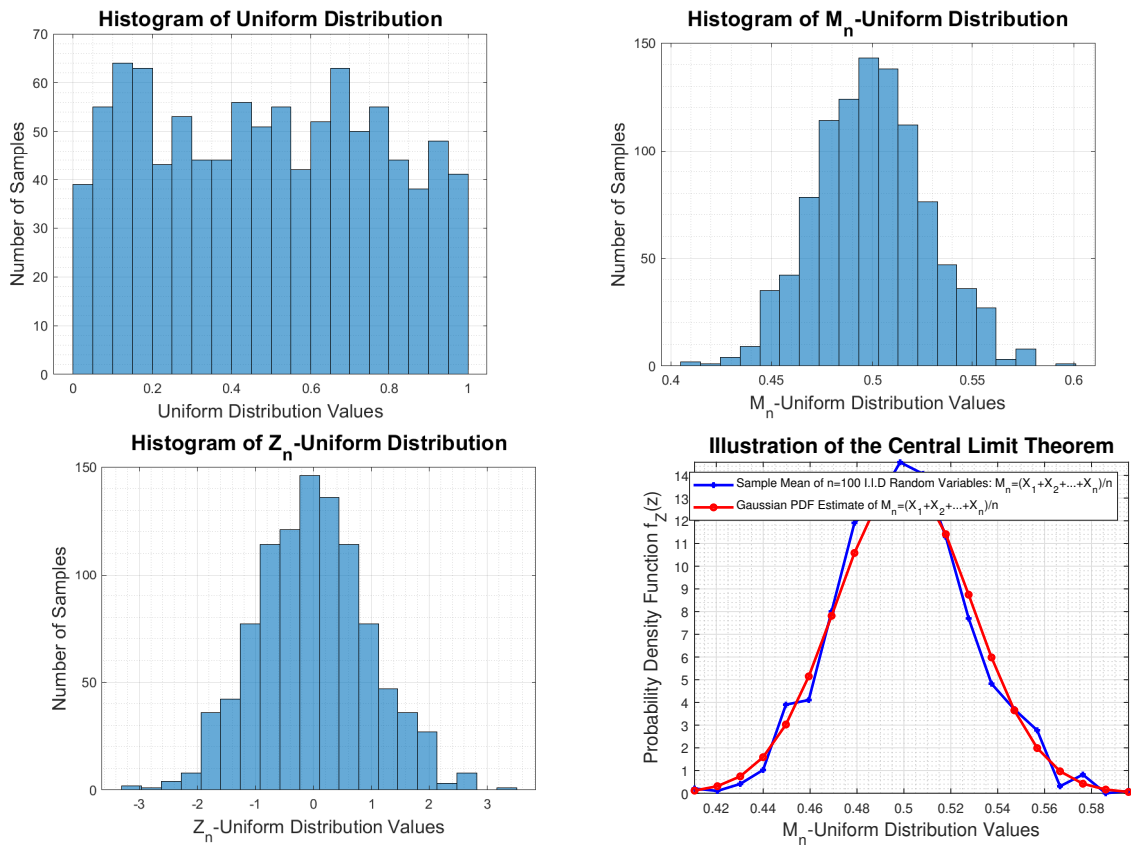


Figure 3: Τα ζητούμενα γραφήματα για την ομοιόμορφη Τ.Μ. X για $N = 1000$, $n = 100$, $n_{bins} = 20$.

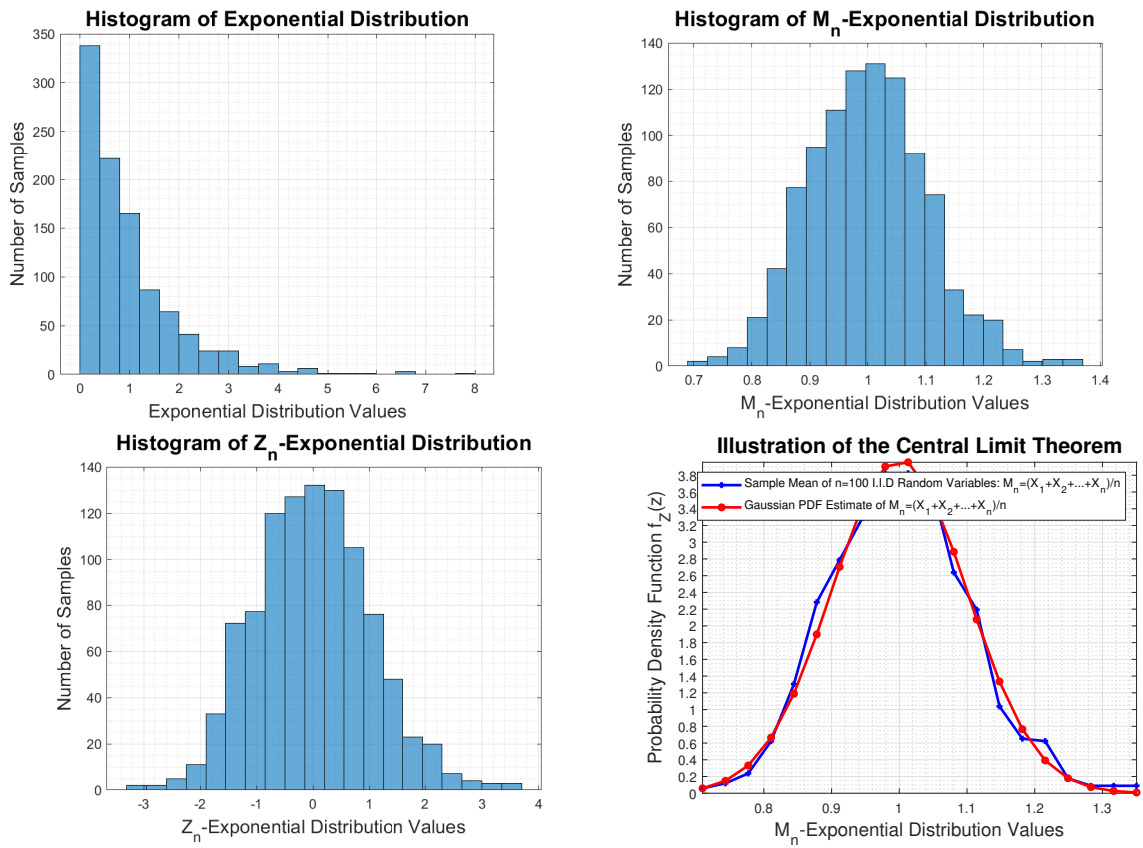


Figure 4: Τα ζητούμενα γραφήματα για την εκθετική T.M. Y για $N = 1000$, $n = 100$, $n_{bins} = 20$.

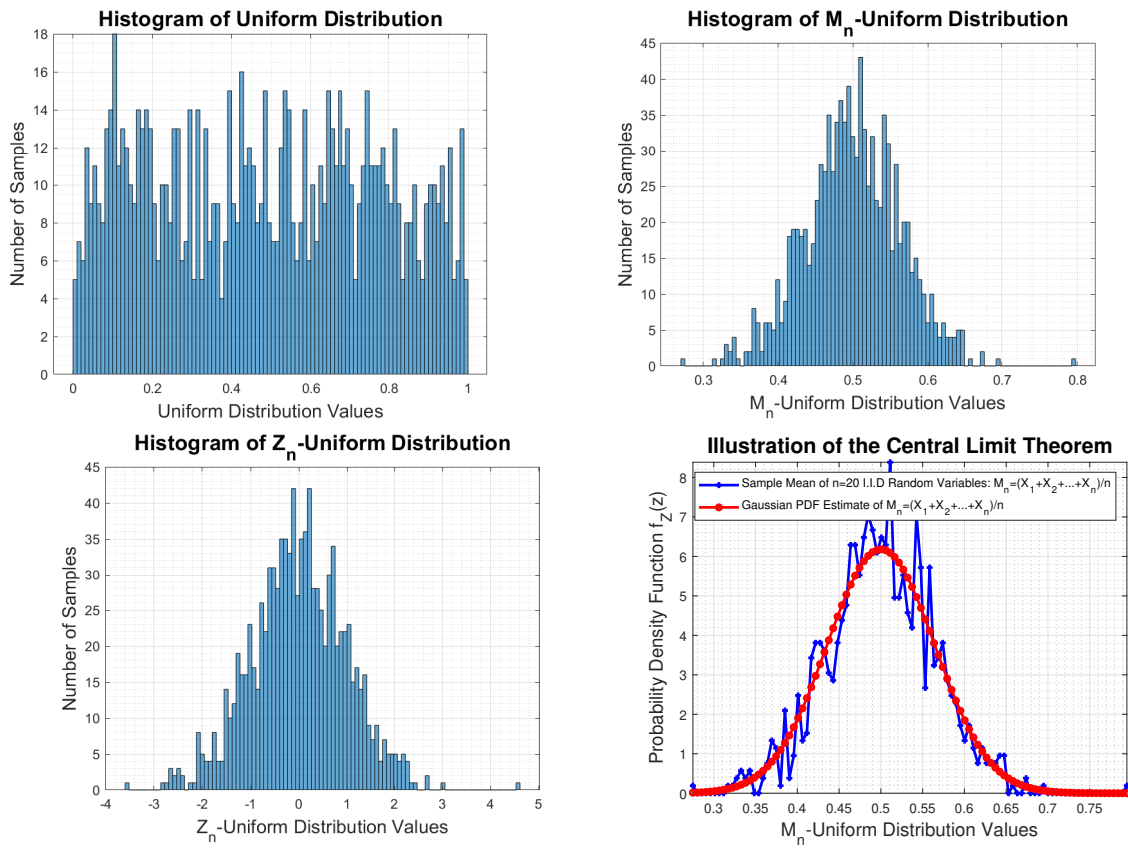


Figure 5: Τα ζητούμενα γραφήματα για την ομοιόμορφη T.M. X για $N = 1000$, $n = 20$, $n_{bins} = 20$.

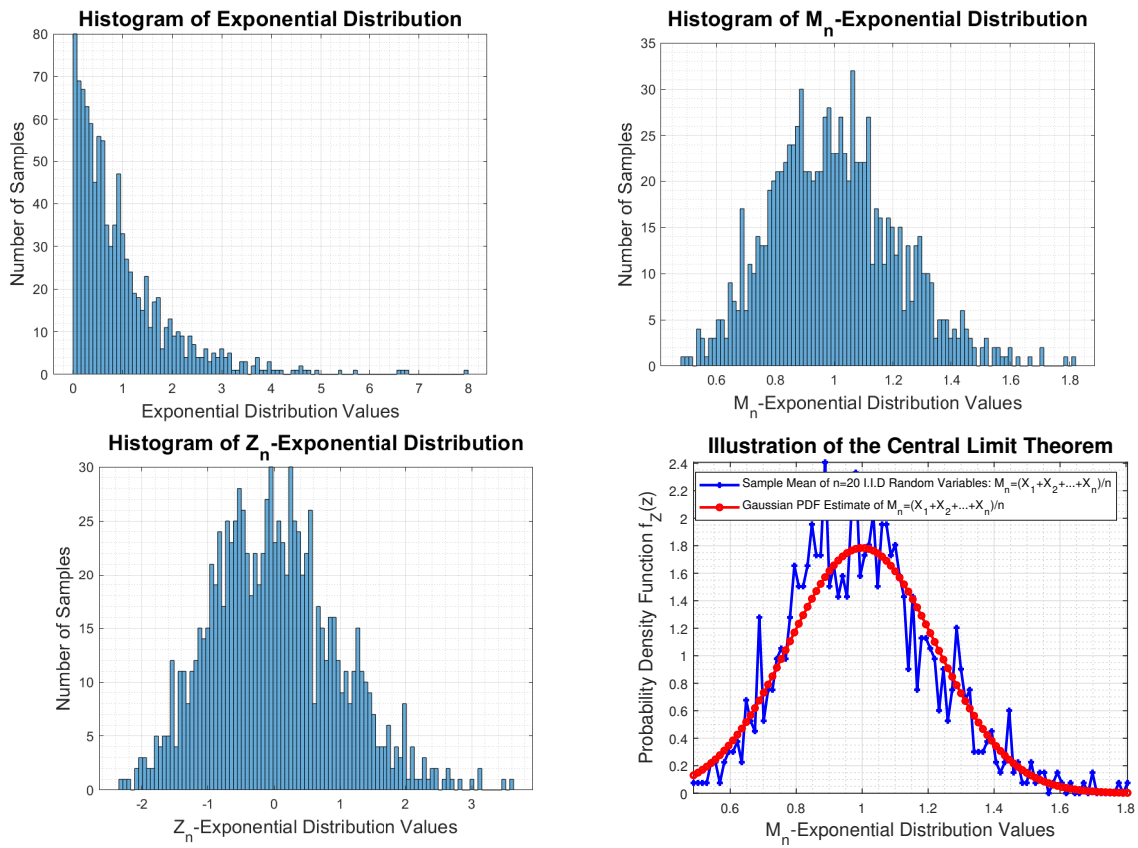


Figure 6: Τα ζητούμενα γραφήματα για την εκθετική T.M. Υ για $N = 1000$, $n = 20$, $n_{bins} = 20$.

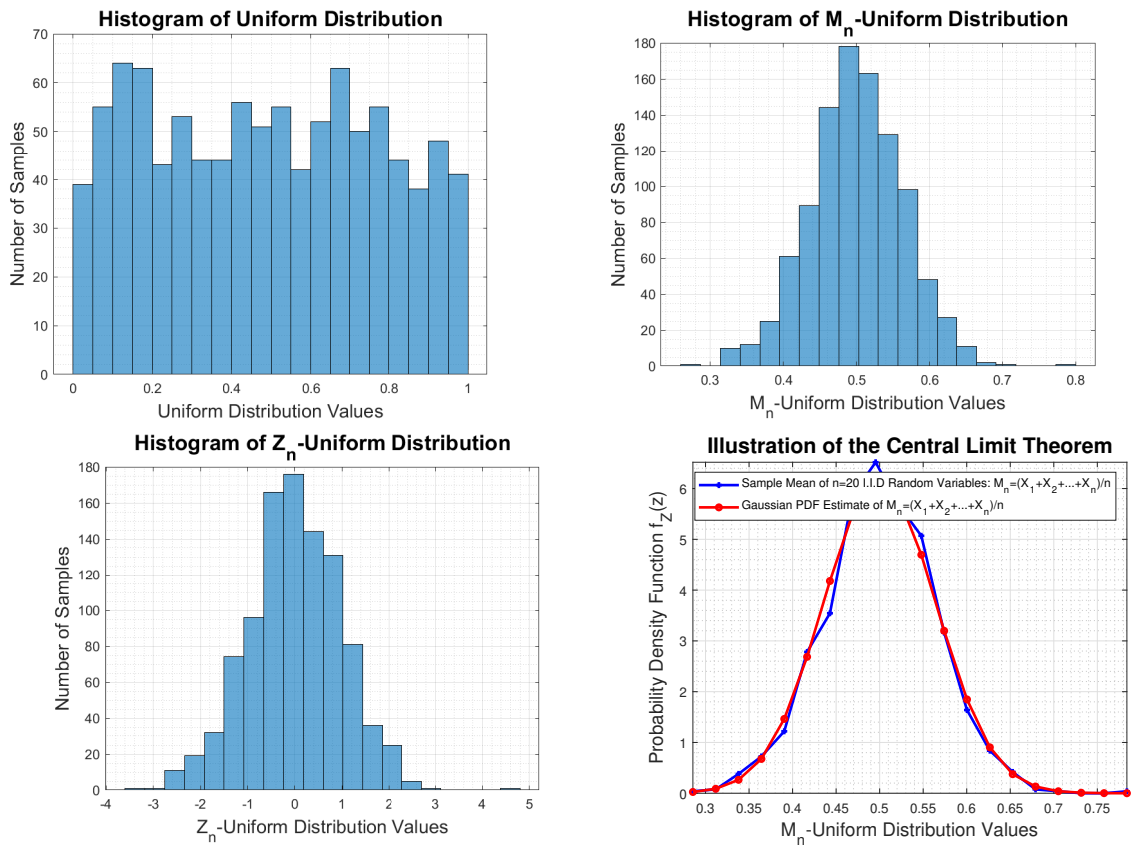


Figure 7: Τα ζητούμενα γραφήματα για την ομοιόμορφη Τ.Μ. X για $N = 10000$, $n = 20$, $n_{bins} = 20$.

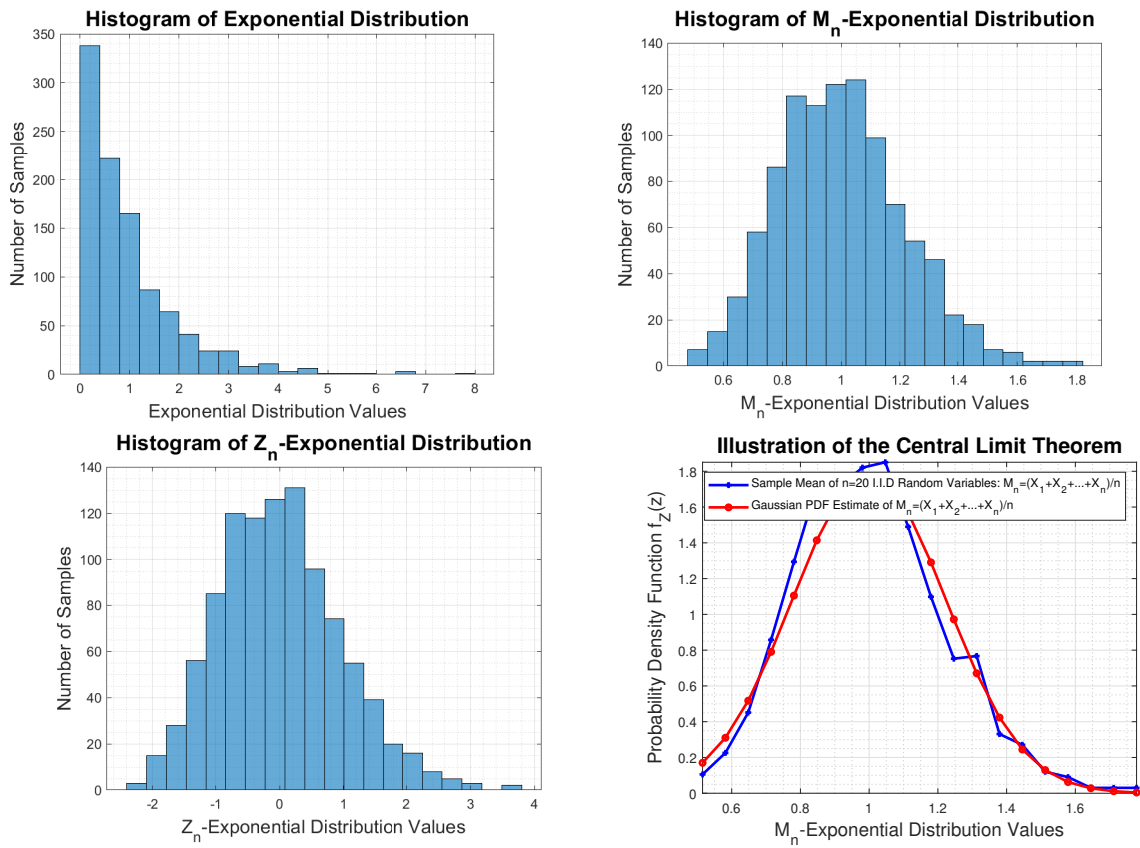


Figure 8: Τα ζητούμενα γραφήματα για την εκθετική T.M. Y για $N = 10000$, $n = 20$, $n_{bins} = 20$.