

3. Οριακά θεωρήματα

Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ.)

Ένα από τα πιο συνηθισμένα προβλήματα που ανακύπτουν στη στατιστική είναι ο προσδιορισμός της κατανομής ενός μεγάλου αθροίσματος ανεξάρτητων τ.μ. Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες τ.μ. που ακολουθούν την ίδια κατανομή (ισόνομες τ.μ.) με

$$E(X_i) = \mu, \quad V(X_i) = \sigma^2, \quad i=1,2,\dots,n.$$

Είναι γνωστό ότι στην περίπτωση που $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$, τότε

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2).$$

Το ερώτημα που προκύπτει είναι το εξής: αν οι ανεξάρτητες και ισόνομες $X_1, X_2, \dots, X_n$ δεν ακολουθούν την κανονική κατανομή αλλά κάποια άλλη κατανομή (ίσως και άγνωστη), τότε ποια μπορεί να είναι προσεγγιστικά η κατανομή του αθροίσματος $X_1 + X_2 + \dots + X_n$ για μεγάλο n ;

Απάντηση σε αυτό το ερώτημα δίνει το περίφημο Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ.) που καταδεικνύει τη σπουδαιότητα της κανονικής κατανομής. Εκτός από το θεωρητικό του ενδιαφέρον και τη σπουδαιότητά του, το θεώρημα αυτό παρέχει μία απλή μέθοδο για τον προσεγγιστικό υπολογισμό πιθανοτήτων που αφορούν αθροίσματα από ανεξάρτητες τυχαίες μεταβλητές.

Θεώρημα 3.1. (Κ.Ο.Θ.) *Αν οι τ.μ. $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες και ισόνομες με $E(X_i) = \mu, V(X_i) = \sigma^2 < \infty, i=1,2,\dots,n$, τότε η τ.μ.*

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}},$$

ακολουθεί ασυμπτωτικά (για μεγάλο n , συνήθως μεγαλύτερο του 30) την τυπική κανονική κατανομή $N(0,1)$.

Λέγοντας παραπάνω ότι η κατανομή της τ.μ. Z ακολουθεί ασυμπτωτικά τυπική κανονική, εννοούμε ότι η σ.κ. F_Z της Z συγκλίνει στην σ.κ. Φ της τυπικής κανονικής (δηλ. $F_Z(x) \rightarrow \Phi(x), \forall x \in R$).

Παρατήρηση 3.1. Είμαστε τώρα σε θέση να απαντήσουμε στο ερώτημα που τέθηκε πριν το Κ.Ο.Θ. Συγκεκριμένα, το άθροισμα ανεξάρτητων και ισόνομων τ.μ. $X_1 + X_2 + \dots + X_n$ με $E(X_i) = \mu, V(X_i) = \sigma^2 < \infty$ και n αρκετά μεγάλο, θα ακολουθεί προσεγγιστικά κανονική κατανομή, άσχετα με το ποια είναι η κατανομή των X_i (κανονική η οποιαδήποτε άλλη). Πράγματι, από το Κ.Ο.Θ. ισχύει ότι, για μεγάλο n ,

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n\sigma^2}} \sim N(0,1),$$

και επομένως η τ.μ. $\sum_{i=1}^n X_i = \sigma\sqrt{n}Z + n\mu$ θα ακολουθεί και αυτή κανονική κατανομή με μέση τιμή και διασπορά

$$E\left(\sum_{i=1}^n X_i\right) = \sigma\sqrt{n}E(Z) + n\mu = n\mu, \quad V\left(\sum_{i=1}^n X_i\right) = V(\sigma\sqrt{n}Z + n\mu) = n\sigma^2 V(Z) = n\sigma^2.$$

Άρα τελικά για οποιεσδήποτε ανεξάρτητες και ισόνομες τ.μ. $X_1, X_2, \dots, X_n$ (που ακολουθούν οποιαδήποτε κατανομή),

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2) \text{ για μεγάλο } n.$$

Σύμφωνα λοιπόν με το Κ.Ο.Θ., κάθε φυσική ποσότητα η τιμή της οποίας μπορεί να θεωρηθεί ότι διαμορφώνεται από ένα μεγάλο αριθμό ανεξάρτητων παραγόντων (δηλαδή μπορεί να γραφεί σαν άθροισμα ενός μεγάλου αριθμού ανεξάρτητων και ισόνομων τ.μ.) ακολουθεί προσεγγιστικά την κανονική κατανομή.

Παρατήρηση 3.2. Η τ.μ.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

καλείται δειγματικός μέσος του δείγματος $X_1, X_2, \dots, X_n$. Σύμφωνα με το Κ.Ο.Θ και ο δειγματικός μέσος (που προέρχεται από παρατηρήσεις που ακολουθούν οποιαδήποτε κατανομή) ακολουθεί ασυμπτωτικά κανονική κατανομή. Πράγματι, από το Κ.Ο.Θ., θα είναι

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{\sigma}{\sqrt{n}} Z + \mu \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

για μεγάλο n . Ισοδύναμα με το Κ.Ο.Θ. μπορούμε να γράψουμε ότι, η τ.μ.

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma} = \frac{n\bar{X} - n\mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \sim N(0,1),$$

για μεγάλο n .

Πόρισμα 3.1. Η διωνυμική κατανομή $B(n, p)$ προσεγγίζεται ικανοποιητικά από την κανονική κατανομή $N(np, np(1-p))$ όταν $np(1-p)$ μεγάλο (>10 ή καλύτερα >30)

Απόδειξη. Έστω n ανεξάρτητες και ισόνομες δίτιμες τ.μ. $X_1, X_2, \dots, X_n$ έτσι ώστε $P(X_i=1) = p$, $P(X_i=0) = 1-p$. Αν $W = X_1 + X_2 + \dots + X_n$ τότε είναι φανερό ότι η W μπορεί να θεωρηθεί ως το πλήθος των επιτυχιών σε n ανεξάρτητες δοκιμές η κάθε μία από τις οποίες έχει πιθανότητα επιτυχίας p και πιθανότητα αποτυχίας $1-p$ (θεωρούμε το 1 επιτυχία και το 0 αποτυχία). Επομένως, $W \sim B(n, p)$. Από το Κ.Ο.Θ. όμως θα ισχύει επίσης ότι, για μεγάλο n ,

$$W \sim N(nE(X_i), nV(X_i)),$$

και επειδή

$$E(X_i) = 0P(X_i=0) + 1P(X_i=1) = p, \quad E(X_i^2) = 0^2P(X_i=0) + 1^2P(X_i=1) = p,$$

$$V(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 = p(1-p),$$

έπεται ότι $W \sim N(np, np(1-p))$ και τελικά, θα είναι $B(n, p) \approx N(np, np(1-p))$.

Με παρόμοιο τρόπο αποδεικνύεται και το επόμενο πόρισμα.

Πόρισμα 3.2. Η κατανομή Poisson με μέση τιμή λ προσεγγίζεται ικανοποιητικά από την κανονική κατανομή $N(\lambda, \lambda)$ όταν λ μεγάλο (>10 ή καλύτερα >30).

Άσκηση 3.1. Ο χρόνος ζωής μιας ηλεκτρονικής λυχνίας που κατασκευάζει μία βιομηχανία είναι μία τ.μ. X με συνάρτηση κατανομής

$$F_X(x) = 1 - e^{-\lambda x}, \quad x > 0,$$

με $\lambda=0.005$.

i) Αν 49 λυχνίες χρησιμοποιηθούν διαδοχικά η μία μετά την άλλη, να βρεθεί η πιθανότητα ο συνολικός χρόνος λειτουργίας των λυχνιών αυτών να ξεπερνά τις 7700 ώρες.

ii) Ποια είναι η πιθανότητα ανάμεσα στις 49 αυτές λυχνίες το πολύ 14 να έχουν ζωή μικρότερη των 140 ωρών.

Λύση. Έστω $X_1, X_2, \dots, X_n$, $n=49$, οι τ.μ. που εκφράζουν τους χρόνους ζωής των 49 πρώτων λυχνιών. Οι τ.μ. $X_1, \dots, X_n$ είναι ανεξάρτητες και ισόνομες. Επειδή ακολουθούν εκθετική κατανομή με παράμετρο $\lambda=0.005=1/200$, έπεται ότι

$$E(X_i) = \frac{1}{\lambda} = 200, \quad V(X_i) = \frac{1}{\lambda^2} = 200^2.$$

Η τ.μ. $\sum_{i=1}^{49} X_i$ παριστά το συνολικό χρόνο λειτουργίας (σε ώρες) των 49 λυχνιών όταν αυτές χρησιμοποιηθούν διαδοχικά. Σύμφωνα με το Κ.Ο.Θ. θα ισχύει ότι ($n=49>30$), με μεγάλη προσέγγιση

$$Z = \frac{\sum_{i=1}^{49} X_i - 49 \cdot 200}{\sqrt{49 \cdot 200}} \sim N(0,1).$$

Ζητείται η πιθανότητα

$$\begin{aligned} P\left(\sum_{i=1}^{49} X_i > 7700\right) &= P\left(\frac{\sum_{i=1}^{49} X_i - 49 \cdot 200}{7 \cdot 200} > \frac{7700 - 49 \cdot 200}{7 \cdot 200}\right) = P(Z > -1.5) = 1 - P(Z \leq -1.5) \\ &\approx 1 - \Phi(-1.5) = \Phi(1.5) \approx 0.9332. \end{aligned}$$

ii) Αν Y είναι το πλήθος των λυχνιών (από τις 49) με ζωή μικρότερη των 140 ωρών τότε θα ισχύει ότι $Y \sim B(n=49, p=P(X_i \leq 140))$. Αυτό συμβαίνει διότι αν ως επιτυχία θεωρήσουμε το ενδεχόμενο $X_i \leq 140$ τότε η τ.μ. Y μπορεί να θεωρηθεί ως το πλήθος των επιτυχιών σε $n=49$ ανεξάρτητες και ισόνομες δοκιμές με πιθανότητα επιτυχίας σε κάθε δοκιμή ίση με $P(X_i \leq 140)$. Η πιθανότητα ανάμεσα στις 49 λυχνίες το πολύ 14 να έχουν ζωή μικρότερη των 140 ωρών είναι $P(Y \leq 14)$. Σε αυτό το σημείο θα μπορούσαμε να χρησιμοποιήσουμε τον ακριβή τύπο:

$$P(Y \leq 14) = \sum_{i=0}^{14} \binom{n}{i} p^i (1-p)^{n-i}, \quad \text{με } n=49, p=P(X_i \leq 140),$$

και $P(X_i \leq 140) = F_X(140) = 1 - e^{-0.005 \cdot 140} = 1 - e^{-0.7} \approx 0.5$. Παρατηρούμε όμως ότι είναι αρκετά δύσκολο να υπολογίσουμε το παραπάνω άθροισμα (ιδιαίτερα τους συνδυασμούς) και είναι προτιμότερο να χρησιμοποιήσουμε την προσέγγιση της διωνυμικής από την κανονική κατανομή. Συγκεκριμένα, από το Πόρισμα 2.1 γνωρίζουμε ότι $B(n, p) \approx N(np, np(1-p))$ ($np(1-p) = 49 \cdot 0.5^2 = 12.25 > 10$). Άρα, προσεγγιστικά,

$$Z = \frac{Y - 49 \cdot p}{\sqrt{49 p(1-p)}} = \frac{Y - 49 \cdot 0.5}{7 \cdot 0.5} = \frac{Y - 24.5}{3.5} \sim N(0,1),$$

και συνεπώς η ζητούμενη πιθανότητα προσεγγιστικά θα είναι

$$P(Y \leq 14) = P\left(\frac{Y - 24.5}{3.5} \leq \frac{14 - 24.5}{3.5}\right) = P(Z \leq -3) \approx \Phi(-3) = 1 - \Phi(3) = 1 - 0.9987 = 0.0013.$$

(ο παραπάνω ακριβής τύπος με την διωνυμική κατανομή δίνει $P(Y \leq 14) \approx 0.0019$).

Άσκηση 3.2. Έστω ότι k εστιατόρια συναγωνίζονται για τους n εργάτες ενός γειτονικού εργοστασίου. Αν κάθε εργάτης εκλέγει για το γεύμα του τυχαία ένα από τα εστιατόρια αυτά, να προσδιο-

ριστεί ο ελάχιστος αριθμός καθισμάτων που πρέπει να διαθέτει ένα εστιατόριο ώστε με πιθανότητα τουλάχιστο 0.95 να είναι σε θέση να εξυπηρετήσει όλους τους πελάτες που έρχονται σε αυτό ($n=500, k=5$).

Λύση. Ας κάνουμε τη μελέτη μας για ένα συγκεκριμένο εστιατόριο, έστω το 1^ο. Θεωρούμε τις τ.μ. $X_1, X_2, \dots, X_n$ έτσι ώστε $X_i=1$ αν ο i εργάτης εκλέγει το 1^ο εστιατόριο και $X_i=0$ αν ο i εργάτης δεν εκλέγει το 1^ο εστιατόριο. Οι δίτιμες τ.μ. $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες και ισόνομες με

$$P(X_i=1)=\frac{1}{k}, P(X_i=0)=1-\frac{1}{k}, E(X_i)=\frac{1}{k}, Var(X_i)=\frac{1}{k}\left(1-\frac{1}{k}\right).$$

Το πλήθος των πελατών που εκλέγουν το 1^ο εστιατόριο θα είναι ίσος με $\sum_{i=1}^n X_i$. Από το Κ.Ο.Θ. θα ισχύει όμως ότι

$$Z = \frac{\sum_{i=1}^n X_i - nE(X_i)}{\sqrt{n}\sqrt{V(X_i)}} = \frac{\sum_{i=1}^n X_i - n/k}{\sqrt{n}\sqrt{\frac{1}{k}(1-\frac{1}{k})}} \sim N(0,1).$$

Σύμφωνα με την εκφώνηση, αν $a_{n,k}$ είναι ο αριθμός καθισμάτων του εστιατορίου, θα πρέπει

$$P\left(\sum_{i=1}^n X_i \leq a_{n,k}\right) \geq 0.95 \Leftrightarrow P\left(\frac{\sum_{i=1}^n X_i - n/k}{\sqrt{n}\sqrt{\frac{1}{k}(1-\frac{1}{k})}} \leq \frac{a_{n,k} - n/k}{\sqrt{n}\sqrt{\frac{1}{k}(1-\frac{1}{k})}}\right) \geq 0.95,$$

ή ισοδύναμα,

$$P\left(Z \leq \frac{a_{n,k} - n/k}{\sqrt{\frac{n}{k}(1-\frac{1}{k})}}\right) \geq 0.95 = \Phi(1.645) \Leftrightarrow \Phi\left(\frac{a_{n,k} - n/k}{\sqrt{\frac{n}{k}(1-\frac{1}{k})}}\right) \geq \Phi(1.645),$$

και άρα θα πρέπει

$$\frac{a_{n,k} - n/k}{\sqrt{\frac{n}{k}(1-\frac{1}{k})}} \geq 1.645 \Leftrightarrow a_{n,k} \geq \frac{n}{k} + 1.645\sqrt{\frac{n}{k}\left(1-\frac{1}{k}\right)}.$$

Για $n=500, k=5$ θα είναι

$$a_{500,5} \geq \frac{500}{5} + 1.645\sqrt{\frac{500}{5}\left(1-\frac{1}{5}\right)} = 114.71,$$

και επομένως ο ελάχιστος αριθμός καθισμάτων που πρέπει να διαθέτει το 1^ο εστιατόριο είναι 115. Προφανώς το ίδιο ισχύει και για τα υπόλοιπα $k-1$ εστιατόρια

Άσκηση 3.3. Σε μία πόλη 4000 κατοίκων έστω ότι 10 άτομα την ημέρα κατά μέσο όρο χρειάζεται να εισαχθούν στο νοσοκομείο. Να υπολογιστεί (κατά προσέγγιση) ο μικρότερος αριθμός των ελεύθερων κρεβατιών που πρέπει να διαθέτει ημερησίως το νοσοκομείο για να είναι σε θέση να εξυπηρετεί την πόλη με πιθανότητα τουλάχιστον 95%.

Λύση. Έστω οι τ.μ. $X_1, X_2, \dots, X_n$ έτσι ώστε $X_i=1$ αν ο i κάτοικος χρειάζεται εισαγωγή (κάποια συγκεκριμένη ημέρα) και $X_i=0$ διαφορετικά ($n=4000$). Οι δίτιμες τ.μ. $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες και ισόνομες με

$$P(X_i=1)=\frac{10}{4000}=0.0025, P(X_i=0)=0.9975,$$

$$E(X_i)=0.0025, V(X_i)=0.0025 \cdot 0.9975 \approx 0.0025.$$

Το πλήθος των κατοίκων που χρειάζεται κρεβάτι θα είναι ίσο με $\sum_{i=1}^n X_i$. Από το Κ.Ο.Θ. θα ισχύει ότι

$$Z = \frac{\sum_{i=1}^n X_i - nE(X_i)}{\sqrt{n}\sqrt{V(X_i)}} = \frac{\sum_{i=1}^n X_i - 10}{\sqrt{10}} \sim N(0,1).$$

Σύμφωνα με την εκφώνηση, αν a είναι ο αριθμός των κλινών του νοσοκομείου, θα πρέπει

$$P\left(\sum_{i=1}^n X_i \leq a\right) \geq 0.95 \Leftrightarrow P\left(\frac{\sum_{i=1}^n X_i - 10}{\sqrt{10}} \leq \frac{a - 10}{\sqrt{10}}\right) \geq 0.95,$$

και

$$\Phi\left(\frac{a - 10}{\sqrt{10}}\right) \geq \Phi(1.645) \Leftrightarrow \frac{a - 10}{\sqrt{10}} \geq 1.645 \Leftrightarrow a \geq 10 + 1.645\sqrt{10} = 15.2.$$

Άρα ο μικρότερος αριθμός των ελεύθερων κρεβατιών που πρέπει να διαθέτει ημερησίως το νοσοκομείο είναι 16.

Άσκηση 3.4. Μία εταιρία κινητής τηλεφωνίας έχει σε μία πόλη 10000 συνδρομητές. Έχει βρεθεί ότι η πιθανότητα να χρησιμοποιήσει κάποιος το κινητό του τηλέφωνο (μία συγκεκριμένη ώρα της ημέρας και κάτω από κανονικές συνθήκες) είναι 3%. Να βρεθεί ο αριθμός των ελεύθερων γραμμών που πρέπει να διαθέτει το τηλεφωνικό κέντρο της εταιρίας ώστε το πολύ μία στις 100 κλήσεις να βρίσκει το δίκτυο κατειλημμένο.

Λύση. Έστω οι τ.μ. $X_1, X_2, \dots, X_n$ έτσι ώστε $X_i = 1$ αν ο i συνδρομητής χρησιμοποιεί το τηλέφωνό του (τη συγκεκριμένη ώρα) και $X_i = 0$ διαφορετικά ($n = 10000$). Οι δίτιμες τ.μ. $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες και ισόνομες με

$$P(X_i = 1) = 0.03, P(X_i = 0) = 0.97,$$

$$E(X_i) = 0.03, V(X_i) = 0.03 \cdot 0.97 \approx 0.0291.$$

Το πλήθος των συνδρομητών που χρησιμοποιεί το τηλέφωνό του θα είναι ίσος με $\sum_{i=1}^n X_i$. Από το Κ.Ο.Θ. θα ισχύει ότι

$$Z = \frac{\sum_{i=1}^n X_i - nE(X_i)}{\sqrt{n}\sqrt{V(X_i)}} = \frac{\sum_{i=1}^n X_i - 300}{\sqrt{291}} \sim N(0,1).$$

Έστω a ο αριθμός των ελεύθερων γραμμών που έχει το τηλεφωνικό κέντρο. Σύμφωνα με την εκφώνηση, θα πρέπει η πιθανότητα να είναι το δίκτυο κατειλημμένο να είναι $\leq 1\%$ ή ισοδύναμα,

$$P\left(\sum_{i=1}^n X_i > a\right) \leq 0.01 \Leftrightarrow P\left(\sum_{i=1}^n X_i \leq a\right) \geq 0.99 \Leftrightarrow P\left(\frac{\sum_{i=1}^n X_i - 300}{\sqrt{291}} \leq \frac{a - 300}{\sqrt{291}}\right) \geq 0.99,$$

και

$$\Phi\left(\frac{a - 300}{\sqrt{291}}\right) \geq \Phi(2.33) \Leftrightarrow \frac{a - 300}{\sqrt{291}} \geq 2.33 \Leftrightarrow a \geq 300 + 2.33\sqrt{291} = 339.74.$$

Άρα ο μικρότερος αριθμός ελεύθερων γραμμών που πρέπει να έχει το τηλεφωνικό κέντρο της εταιρίας είναι 340.

Άσκηση 3.5. Έστω ότι κάποιος θέλει να προσθέσει n πραγματικούς αριθμούς (με αρκετά δεκαδικά ψηφία ο καθένας). Για συντομία όμως δεν χρησιμοποιεί όλα τα ψηφία αλλά πριν κάνει την πρόσθεση, τους στρογγυλοποιεί προς τον πλησιέστερο ακέραιο. Ποια είναι η πιθανότητα το συνολικό σφάλμα (διαφορά πραγματικού αθροίσματος από το άθροισμα μετά τις στρογγυλοποιήσεις) να είναι μικρότερο του 10 εάν υποθέσουμε ότι τα σφάλματα των στρογγυλοποιήσεων είναι ανεξάρτητα και ακολουθούν την ομοιόμορφη στο $(-0.5, 0.5)$ κατανομή ($n = 1200$).

Λύση. Έστω $X_1, X_2, \dots, X_n$ τα σφάλματα των στρογγυλοποιήσεων. Θα ισχύει ότι

$$E(X_i) = \frac{0.5 - 0.5}{2} = 0, V(X_i) = \frac{(0.5 + 0.5)^2}{12} = \frac{1}{12}.$$

Το συνολικό σφάλμα θα είναι ίσο με $\sum_{i=1}^n X_i$. Από το Κ.Ο.Θ. θα ισχύει ότι

$$Z = \frac{\sum_{i=1}^n X_i - nE(X_i)}{\sqrt{n} \sqrt{V(X_i)}} = \frac{\sum_{i=1}^n X_i - 0}{\sqrt{100}} = \frac{\sum_{i=1}^n X_i}{10} \sim N(0,1).$$

Ζητείται η πιθανότητα

$$P(-10 < \sum_{i=1}^n X_i < 10) = P(-1 < Z < 1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \approx 0.68.$$

Ανισότητες Markov, Chebyshev

Ας υποθέσουμε ότι γνωρίζουμε τη μέση τιμή $E(X)$ μιας τυχαίας μεταβλητής και ίσως ακόμη και τη διασπορά της $V(X)$. Μπορούμε σε αυτή την περίπτωση να εξάγουμε κάποια συμπεράσματα για την (άγνωστη) κατανομή της; Ας δούμε στη συνέχεια δύο αποτελέσματα που μας βοηθούν να βρούμε φράγματα πιθανοτήτων όταν είναι γνωστή μόνο η $E(X)$ ή οι $E(X)$, $V(X)$ της κατανομής.

Πρόταση 3.1. (Ανισότητα Markov). Αν X είναι μία θετική τ.μ. ($X \geq 0$), τότε για κάθε $a > 0$ ισχύει ότι

$$P(X \geq a) \leq \frac{E(X)}{a}, a > 0.$$

Απόδειξη. Θα δώσουμε μία απόδειξη για τη συνεχή περίπτωση. Έστω ότι η X έχει συνάρτηση πυκνότητας πιθανότητας f (επειδή $X \geq 0$, θα είναι $f(x) = 0$ για $x < 0$). Θα ισχύει ότι

$$E(X) = \int_0^\infty xf(x)dx = \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \geq \int_a^\infty xf(x)dx \geq \int_a^\infty a f(x)dx = a P(X \geq a),$$

από το οποίο προκύπτει το ζητούμενο. $\square$

Ως πόρισμα της παραπάνω ανισότητας έχουμε το ακόλουθο αποτέλεσμα.

Πρόταση 3.2. (Ανισότητα Chebyshev). Αν X είναι μία τ.μ. με μέση τιμή $\mu = E(X)$ και διασπορά $V(X)$ τότε, για κάθε $\varepsilon > 0$,

$$P(|X - \mu| \geq \varepsilon) \leq \frac{V(X)}{\varepsilon^2}.$$

Απόδειξη. Η τ.μ. $Y = (X - \mu)^2$ είναι θετική και επομένως από την ανισότητα Markov θα έχουμε ότι (θέτουμε $a = \varepsilon^2$),

$$P(Y \geq \varepsilon^2) \leq \frac{E(Y)}{\varepsilon^2} \Leftrightarrow P((X - \mu)^2 \geq \varepsilon^2) \leq \frac{E((X - \mu)^2)}{\varepsilon^2} = \frac{V(X)}{\varepsilon^2},$$

από όπου προκύπτει το ζητούμενο παρατηρώντας ότι $P((X - \mu)^2 \geq \varepsilon^2) = P(|X - \mu| \geq \varepsilon)$. $\square$

Οι παραπάνω ανισότητες είναι αρκετά χρήσιμες στην περίπτωση που επιθυμούμε να κατασκευάσουμε φράγματα για πιθανότητες μιας κατανομής όταν είναι γνωστή μόνο η μέση τιμή ή μόνο η μέση τιμή και η διασπορά της κατανομής αυτής. Είναι προφανές ότι, αν είναι γνωστή η κατανομή (π.χ. είναι κανονική) τότε οι εν λόγω πιθανότητες μπορούν να υπολογιστούν ακριβώς

και δεν υπάρχει ανάγκη χρήσης των παραπάνω ανισοτήτων. Ας δούμε τα παραπάνω μέσα από μία άσκηση.

Άσκηση 3.6. Γνωρίζουμε ότι το πλήθος X των μονάδων που κατασκευάζει ένα εργοστάσιο κατά τη διάρκεια μίας εβδομάδας είναι μία τ.μ. με μέση τιμή 500.

(α) Τι μπορούμε να πούμε για την πιθανότητα αυτή την εβδομάδα η παραγωγή να είναι τουλάχιστον 1000 μονάδες;

(β) Αν επιπλέον γνωρίζουμε ότι η διασπορά της εβδομαδιαίας παραγωγής X είναι 100, τότε τι μπορούμε να πούμε για την πιθανότητα αυτή την εβδομάδα η παραγωγή να είναι μεταξύ 400 και 600 μονάδων.

Λύση. (α) Από την ανισότητα Markov προκύπτει ότι

$$P(X \geq 1000) \leq \frac{E(X)}{1000} = \frac{500}{1000} = 50\%.$$

(β) Από την ανισότητα Chebyshev θα είναι

$$P(400 < X < 600) = 1 - P(|X - 500| \geq 100) \geq 1 - \frac{V(X)}{100^2} = 1 - \frac{1}{100} = 99\%.$$

Η χρησιμότητα όμως των παραπάνω ανισοτήτων δεν εξαντλείται μόνο σε περιπτώσεις όπως αυτή της παραπάνω άσκησης. Με τη βοήθεια της ανισότητας Chebyshev μπορούμε να αποδείξουμε ένα πολύ σημαντικό οριακό θεώρημα, το γνωστό ως νόμο των μεγάλων αριθμών. Το νόμο αυτό τον είχαμε επικαλεστεί και στο παρελθόν χωρίς να τον έχουμε διατυπώσει αυστηρά. Πριν προχωρήσουμε, ας θέσουμε έναν προβληματισμό σχετικά με την έννοια της «πιθανότητας» ενός ενδεχομένου ή της «μέση τιμής» μιας τ.μ. Τι καταλαβαίνουμε διαισθητικά ή τι εννοούμε πρακτικά όταν λέμε ότι η πιθανότητα ενός ενδεχομένου A είναι π.χ. 30% ή όταν λέμε ότι μία τ.μ. X έχει μέση τιμή π.χ. 10; Συνήθως αυτό που έχουμε στο μυαλό μας είναι ότι αν εκτελεστεί το ίδιο πείραμα (που αφορά το A) πάρα πολλές φορές, τότε το ενδεχόμενο A θα πραγματοποιηθεί στο 30% των περιπτώσεων. Αυτός όμως είναι στην ουσία ο ορισμός κατά τον Von Mises της πιθανότητας ως οριακής σχετικής συχνότητας (βλ. Κεφ I, σημειώσεις Στατ. II). Όμως, η θεωρία την οποία αναπτύσσουμε δεν βασίζεται στον ορισμό του Von Mises αλλά στα αξιώματα Kolmogorov τα οποία δεν κάνουν κανένα λόγο για συχνότητα εμφάνισης ενδεχομένων.

Αυτό που πραγματικά συμβαίνει είναι ότι ο κανόνας υπολογισμού της πιθανότητας ως οριακής σχετικής συχνότητας από τον Von Mises αποτελεί (και αυτός, όπως και ο κατά Laplace ορισμός της πιθανότητας) πόρισμα των αξιωμάτων Kolmogorov. Το αξιοσημείωτο αυτό γεγονός καθώς και άλλα συμπεράσματα που θα εξετάσουμε στη συνέχεια προκύπτουν το «νόμο των μεγάλων αριθμών».

Θεώρημα 3.2. (Νόμος μεγάλων αριθμών) Έστω $X_1, X_2, \dots$ μία ακολουθία ανεξάρτητων και ισόνομων τ.μ. με $E(X_i) = \mu$. Τότε, με πιθανότητα 1,

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mu \text{ όταν } n \rightarrow \infty.$$

Απόδειξη. (Για απλότητα θα αποδείξουμε το παραπάνω χρησιμοποιώντας τη λεγόμενη «ασθενή» σύγκλιση και θεωρώντας επιπλέον ότι $V(X_i) = \sigma^2 < \infty$). Από την ανισότητα Chebyshev για την τ.μ. $\bar{X}_n$ θα ισχύει ότι

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{V(\bar{X}_n)}{\varepsilon^2} \text{ για κάθε } \varepsilon > 0,$$

αλλά,

$$V(\bar{X}_n) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n},$$

και άρα τελικά,

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \xrightarrow{n \rightarrow \infty} 0 \quad \text{για κάθε } \varepsilon > 0.$$

Άρα τελικά $P(\mu - \varepsilon < \bar{X}_n < \mu + \varepsilon) \rightarrow_{n \rightarrow \infty} 1$ για κάθε $\varepsilon > 0$. Με άλλα λόγια, οσοδήποτε μικρή περιοχή γύρω από το μ και αν θεωρήσουμε, το $\bar{X}_n$ θα ανήκει σε αυτήν για n μεγάλο, με πιθανότητα «σχεδόν» 1. $\square$

Με απλά λόγια, αν εκτελέσουμε το ίδιο τυχαίο πείραμα n φορές και X_i είναι η τ.μ. που εκφράζει το αποτέλεσμα του i -πειράματος, τότε, σύμφωνα με το νόμο των μεγάλων αριθμών, ο μέσος όρος $\bar{X}_n$ των X_i συγκλίνει στη μέση τιμή $E(X_i)$ όταν $n \rightarrow \infty$. Έτσι, όταν λέμε ότι μία κατανομή με σ.κ. F έχει μέση τιμή μ , υπονοούμε ότι ο μέσος όρος ενός μεγάλου δείγματος από την κατανομή αυτή (μέσος όρος ανεξάρτητων τ.μ. $X_1, X_2, \dots, X_n \sim F$), θα συγκλίνει στο μ .

Έστω τώρα ότι έχουμε έναν μεγάλο (θεωρητικά άπειρο) πληθυσμό (του οποίου εξετάζουμε ένα χαρακτηριστικό) και επιλέγουμε από αυτόν τυχαία n άτομα με χαρακτηριστικά $X_1, X_2, \dots, X_n$. Σύμφωνα με το νόμο των μεγάλων αριθμών, ο μέσος $\bar{X}_n$ του δείγματος θα συγκλίνει (για μεγάλο n) στη μέση τιμή $\mu = E(X_i)$. Θεωρητικά, αν επιλέξουμε όλο τον πληθυσμό (δηλ. $n \rightarrow \infty$) τότε ο μέσος του πληθυσμού $\bar{X}_n$ θα είναι ίσος με $\mu = E(X_i)$. Για το λόγο αυτό η μέση τιμή $E(X_i)$ μερικές φορές καλείται και *πληθυσμιακή μέση τιμή* (η μέση τιμή του χαρακτηριστικού σε «ολόκληρο» τον πληθυσμό). Αντίθετα, η $\bar{X}_n$ καλείται *δειγματική μέση τιμή* (η μέση τιμή του χαρακτηριστικού στο τυχαίο δείγμα).

Ας δούμε τώρα πως προκύπτει ο ορισμός κατά Von Mises της πιθανότητας από το νόμο των μεγάλων αριθμών. Έστω Ω ο δειγματικός χώρος ενός πειράματος τύχης και A ένα ενδεχόμενο, υποσύνολο του Ω . Αν θεωρήσουμε ότι το πείραμα αυτό επαναλαμβάνεται n φορές (ανεξάρτητες μεταξύ τους) και θέσουμε $X_i = 1$ ή 0 ανάλογα με το αν στο i -πείραμα πραγματοποιηθεί το A ή όχι, τότε, από το νόμο των μεγάλων αριθμών θα ισχύει ότι

$$\frac{1}{n} \sum_{i=1}^n X_i \rightarrow E(X_i) \quad \text{για } n \rightarrow \infty.$$

Αλλά το $\frac{1}{n} \sum_{i=1}^n X_i$ εκφράζει το πλήθος των πραγματοποιήσεων του A στα n πειράματα, δια n . Συνεπώς, εκφράζει τη σχετική συχνότητα εμφάνισης του A στα n πειράματα. Επίσης, ισχύει ότι

$$E(X_i) = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = P(X_i = 1) = P(A),$$

και επομένως, η παραπάνω σχέση στην ουσία εκφράζει το γεγονός ότι η οριακή σχετική συχνότητα εμφάνισης του A συγκλίνει στην πιθανότητα του A .