Κεφάλαιο 2

Πιθανότητες και Τυχαίες Μεταβλητές

Μπορούμε να καταλάβουμε την έννοια της πιθανότητας από τη σχετική συχνότητα εμφάνισης n_i κάποιας τιμής x_i μιας διακριτής τ.μ. X. Αν είχαμε τη δυνατότητα να συλλέξουμε αυθαίρετα πολλές n παρατηρήσεις ($n \to \infty$), τότε το όριο της σχετικής συχνότητας είναι η **πιθανότητα** η τ.μ. X να πάρει την τιμή x_i

$$\mathrm{P}(x_i) \equiv \mathrm{P}(X = x_i) = \lim_{n \longrightarrow \infty} \frac{n_i}{n} \tag{2.1}$$

(το σύμβολο ≡ σημαίνει ισοδυναμία συμβολισμού). Για να είναι έγκυρος αυτός ο ορισμός πρέπει επίσης να υποθέσουμε ότι οι συνθήκες για την τ.μ. X σε κάθε επανάληψη της παρατήρησης παραμένουν οι ίδιες, και αυτή η ιδιότητα ονομάζεται *στατιστική ομαλότητα* (statistical regularity). Για παράδειγμα, η πιθανότητα βροχής σε μια περιοχή (σε μια τυχαία μέρα του χρόνου ή ενός συγκεκριμένου μήνα) μπορεί να έχει αλλάξει τα τελευταία χρόνια λόγω του φαινομένου του θερμοκηπίου.

Για συνεχή τ.μ. X δεν έχει νόημα να μιλάμε για την πιθανότητα η X να πάρει μια συγκεκριμένη τιμή αλλά για την πιθανότητα η X να ανήκει σε ένα διάστημα τιμών $\mathrm{d}x$. Ποια είναι η πιθανότητα να έχει κάποιος συμφοιτητής σας ένα συγκεκριμένο ύψος που ορίζεται με ακρίβεια πολλών (άπειρων) δεκαδικών, π.χ. 1.80123256538634255; Μπορείτε όμως να προσδώσετε μη μηδενική πιθανότητα για το γεγονός ότι ένας συμφοιτητής σας έχει ύψος στα 1.80 μέτρα (όπου με βάση τη στρογγυλοποίηση του εκατοστού έχουμε $\mathrm{d}x = [1.795, 1.805)$).

2.1 Κατανομή πιθανότητας

2.1.1 Κατανομή πιθανότητας μιας τ.μ.

Η πιθανότητα η τ.μ. X να πάρει κάποια τιμή x_i, αν είναι διακριτή, ή να βρίσκεται σε ένα διάστημα τιμών $\mathrm{d}x$, αν είναι συνεχής, μπορεί να μεταβάλλεται στο σύνολο των διακεκριμένων τιμών ή σε διαφορετικά διαστήματα και δίνεται ως συνάρτηση της τ.μ. X. Για διακριτή τ.μ. X που παίρνει τις τιμές $x_1, x_2, \ldots, x_m$, η συνάρτηση αυτή λέγεται **συνάρτηση μάζας πιθανότητας, σμπ** (probability mass function), ορίζεται ως $f_X(x_i) = \mathrm{P}(X = x_i)$ και ικανοποιεί τις συνθήκες

$$f_X(x_i) \geq 0 \qquad \text{και} \qquad \sum_{i=1}^{m} f_X(x_i) = 1. \tag{2.2}$$

Αντίστοιχα, για συνεχή τ.μ. X ($X \in \mathbf{R}$) ορίζεται η **συνάρτηση πυκνότητας πιθανότητας, σππ** (probability density function) $f_X(x)$ που ικανοποιεί τις συνθήκες

$$f_X(x) \geq 0 \qquad \text{και} \qquad \int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1. \tag{2.3}$$

Η *κατανομή πιθανότητας* (probability distribution) της τ.μ. X ορίζεται επίσης από την **αθροιστική συνάρτηση κατανομής, ασκ** (cumulative distribution function) $F_X(x)$, που δηλώνει την πιθανότητα η τ.μ. X να πάρει τιμές μικρότερες ή ίσες από κάποια τιμή x. Για διακριτή τ.μ. X είναι

$$F_X(x_i) = \mathrm{P}(X \leq x_i) = \sum_{x \leq x_i} f_X(x) \tag{2.4}$$

και για συνεχή τ.μ. X

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u. \tag{2.5}$$

Σημειώνεται ότι μια συνεχής μεταβλητή μπορεί να μετατραπεί σε διακριτή με κατάλληλη διαμέριση του πεδίου τιμών της. Αν η συνεχής τ.μ. X ορίζεται στο διάστημα $[a, b]$ μια διαμέριση Σ σε m κελιά δίνεται ως

$$\Sigma = \{a = r_0, r_1, \ldots, r_{m-1}, r_m = b\}, \quad \text{όπου} \quad r_0 < r_1 < \ldots < r_m.$$

Αντιστοιχίζοντας διακεκριμένες τιμές x_i, $i = 1, \ldots, m$, σε κάθε κελί (διάστημα) $[r_{i-1}, r_i)$, η πιθανότητα εμφάνισης μιας τιμής x_i της διακριτικοποιημένης τ.μ. X', $f_{X'}(x_i) = \mathrm{P}(X' = x_i)$, δίνεται από την πιθανότητα η συνεχής τ.μ. X να παίρνει τιμές στο διάστημα $[r_{i-1}, r_i)$, $\mathrm{P}(r_{i-1} \leq X \leq r_i) = F_X(r_i) - F_X(r_{i-1})$.

2.1.2 Από κοινού πιθανότητα δύο τ.μ.

Σε πολλά προβλήματα χρειάζεται να ορίσουμε την συνδυασμένη μεταβλητότητα δύο τ.μ. X και Y, δηλαδή την από κοινού κατανομή πιθανότητας τους. Έστω η διακριτή τ.μ. X με δυνατές διακεκριμένες τιμές $x_1, x_2, \ldots, x_n$ και Y με δυνατές διακεκριμένες τιμές $y_1, y_2, \ldots, y_m$, αντίστοιχα. Η **από κοινού συνάρτηση μάζας πιθανότητας** (joint probability mass function) $f_{XY}(x, y)$ ορίζεται για κάθε ζεύγος δυνατών τιμών (x_i, y_i) ως

$$f_{XY}(x_i, y_i) = \mathrm{P}(X = x_i, Y = y_i) \tag{2.6}$$

και η **από κοινού αθροιστική συνάρτηση κατανομής** (joint cumulative density function) ορίζεται ως

$$F_{XY}(x_i, y_i) = \mathrm{P}(X \leq x_i, Y \leq y_i) = \sum_{x \leq x_i} \sum_{y \leq y_i} f_{XY}(x_i, y_i). \tag{2.7}$$

Η **από κοινού συνάρτηση πυκνότητας πιθανότητας** (joint probability density function) $f_{XY}(x, y)$ για δύο συνεχείς τ.μ. X και Y θα πρέπει να ικανοποιεί τις συνθήκες

$$f_{XY}(x, y) \geq 0 \qquad \text{και} \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) \mathrm{d}y \mathrm{d}x = 1. \tag{2.8}$$

Η **από κοινού (αθροιστική) συνάρτηση κατανομής** για δύο συνεχείς τ.μ. X και Y ορίζεται ως

$$F_{XY}(x, y) = \mathrm{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \mathrm{d}v \mathrm{d}u. \tag{2.9}$$

Δύο τ.μ. X και Y (συνεχείς ή διακριτές) είναι **ανεξάρτητες** (independent) αν για κάθε δυνατό ζεύγος τιμών τους (x, y) ισχύει

$$f_{XY}(x, y) = f_X(x) f_Y(y) \tag{2.10}$$

Με ανάλογο τρόπο ορίζονται οι συναρτήσεις από κοινού κατανομής για περισσότερες τ.μ. καθώς και η ανεξαρτησία πολλών τ.μ..

2.2 Παράμετροι κατανομής τυχαίων μεταβλητών

Η κατανομή πιθανότητας περιγράφει πλήρως τη συμπεριφορά της τ.μ., αλλά συνήθως στην πράξη δεν είναι γνωστή ή απαραίτητη. Όταν μελετάμε μια τ.μ. μας ενδιαφέρει κυρίως να προσδιορίσουμε κάποια βασικά χαρακτηριστικά της κατανομής της, όπως η *κεντρική τάση* και η *μεταβλητότητα* της τ.μ.. Αυτά τα χαρακτηριστικά είναι οι παράμετροι της κατανομής της τ.μ..

2.2.1 Μέση τιμή

Αν X είναι μια διακριτή τ.μ. που παίρνει m διακριτές τιμές $x_1, x_2, \ldots, x_m$, με σμπ $f_X(x)$, η μέση τιμή της, που συμβολίζεται $\mu_X \equiv \mathrm{E}[X]$ ή απλά μ, δίνεται ως

$$\mu \equiv \mathrm{E}[X] = \sum_{i=1}^{m} x_i f_X(x_i). \tag{2.11}$$

Αν η X είναι συνεχής τ.μ. με σπп $f_X(x)$, η μέση τιμή της δίνεται ως

$$\mu \equiv \mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x. \tag{2.12}$$

Κάποιες βασικές ιδιότητες της μέσης τιμής είναι:

1. Αν η τ.μ. X παίρνει μόνο μια σταθερή τιμή c είναι $\mathrm{E}[X] = c$.
2. Αν X είναι μια τ.μ. και c είναι μια σταθερά: $\mathrm{E}[cX] = c\mathrm{E}[X]$.
3. Αν X και Y είναι δύο τ.μ.: $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$.
4. Αν X και Y είναι δύο ανεξάρτητες τ.μ.: $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Οι ιδιότητες (2) και (3) δηλώνουν πως η μέση τιμή έχει τη γραμμική ιδιότητα, δηλαδή ισχύει $\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$.

Άλλα χαρακτηριστικά της τ.μ. X εκτός της μέσης τιμής είναι τα εκατοστιαία σημεία που μπορούν να προσδιοριστούν από την αθροιστική συνάρτηση κατανομής. Η **διάμεσος** (median) $\tilde{\mu}$ μιας τ.μ. X είναι το 50-εκατοστιαίο σημείο, δηλαδή η $\tilde{\mu}$ ικανοποιεί τη σχέση $F_X(\tilde{\mu}) = 0.5$.

2.2.2 Διασπορά

Η **διασπορά** ή **διακύμανση** (variance) μιας τ.μ. X και κυρίως η **τυπική απόκλιση** (standard deviation) (που είναι η τετραγωνική ρίζα της διασποράς), εκφράζουν τη μεταβλητότητα της τ.μ. X γύρω από τη μέση τιμή. Αν X είναι μια τ.μ. (διακριτή ή συνεχής) με μέση τιμή μ, τότε η διασπορά της που συμβολίζεται $\sigma_X^2 \equiv \mathrm{Var}[X]$ ή απλά σ^2 δίνεται ως

$$\sigma^2 \equiv \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mu^2. \tag{2.13}$$

Κάποιες βασικές ιδιότητες της διασποράς είναι:

1. Αν η τ.μ. X παίρνει μόνο μια σταθερή τιμή c είναι $\mathrm{Var}[c] = 0$.
2. Αν X είναι μια τ.μ. και c είναι μια σταθερά: $\mathrm{Var}[X + c] = \mathrm{Var}[X]$ και $\mathrm{Var}[cX] = c^2\mathrm{Var}[X]$.

3. Αν X και Y είναι δύο ανεξάρτητες τ.μ.: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Οι ιδιότητες (2) και (3) δηλώνουν πως η διασπορά δεν έχει τη γραμμική ιδιότητα. Όταν όμως οι δύο τ.μ. είναι ανεξάρτητες η διασπορά έχει τη ψευδογραμμική ιδιότητα, δηλαδή ισχύει $\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y]$.

2.2.3 Ροπές μιας τ.μ.

Συχνά για την περιγραφή των ιδιοτήτων μιας τ.μ. X δεν αρκεί μόνο η μέση τιμή και η διασπορά (που βασίζονται στην πρώτη και δεύτερη δύναμη της X), αλλά πρέπει να καταφύγουμε σε μεγαλύτερες δυνάμεις της X. Η μέση τιμή και η διασπορά αναφέρονται και ως ροπή πρώτης τάξης και (κεντρική) ροπή δεύτερης τάξης, αντίστοιχα. Γενικά ορίζονται οι ροπές $\mathrm{E}[X^n]$ και οι κεντρικές ροπές $\mu_n \equiv \mathrm{E}[(x-\mu_X)^n]$ για κάθε τάξη n.

Από τις σημαντικότερες ροπές είναι η κεντρική ροπή τρίτης τάξης που χρησιμοποιείται στον ορισμό του συντελεστή λοξότητας (coefficient of skewness) λ

$$\lambda = \frac{\mu_3}{\sigma^3} = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]. \tag{2.14}$$

Ο συντελεστής λοξότητας λ εκφράζει τη λοξότητα της κατανομής της τ.μ. X (δηλαδή της σμπ για διακριτή X ή της σππ για συνεχή X). Για $\lambda = 0$ η κατανομή είναι συμμετρική.

Από τη κεντρική ροπή τέταρτης τάξης ορίζεται ο συντελεστή κύρτωσης (coefficient of kurtosis) κ

$$\kappa = \frac{\mu_4}{\sigma^4} - 3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3. \tag{2.15}$$

Ο συντελεστής κύρτωσης κ δηλώνει τη σχέση του πλάτους της κατανομής γύρω από τη κεντρική τιμή με τις ουρές της. Πιο συγκεκριμένα δηλώνει κατά πόσο αυτή η σχέση αποκλίνει από αυτήν της τυπικής κανονικής κατανομής όπου ορίζεται να είναι ίση με 3 (και για αυτό αφαιρείται) (Η τυπική κανονική κατανομή παρουσιάζεται παρακάτω).

2.2.4 Συνδιασπορά και συντελεστής συσχέτισης

Όταν μελετάμε δύο τ.μ. X και Y που δεν είναι ανεξάρτητες, έχει ενδιαφέρον να προσδιορίσουμε πόσο ισχυρά συσχετίζεται η μια με την άλλη. Γι αυτό ορίζουμε τη **συνδιασπορά** ή **συνδιακύμανση** (covariance) των τ.μ. X και Y, που συμβολίζεται $\sigma_{XY} \equiv \mathrm{Cov}[X,Y]$, και ορίζεται ως

$$\sigma_{XY} \equiv \mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[XY] - \mu_X\mu_Y. \tag{2.16}$$

Αν οι δύο τ.μ. X και Y συσχετίζονται ισχυρά και θετικά, δηλαδή όταν αυξάνει η μία αυξάνει και η άλλη, τότε η συνδιασπορά παίρνει μεγάλη θετική τιμή σε σχέση με τις τιμές των X και Y. Αντίθετα αν οι δύο τ.μ. X και Y συσχετίζονται ισχυρά και αρνητικά, δηλαδή όταν αυξάνει η μία μειώνεται η άλλη, τότε η συνδιασπορά παίρνει μεγάλη αρνητική τιμή. Αν οι X και Y είναι ανεξάρτητες εύκολα μπορεί να δειχθεί από την (2.16) ότι $\sigma_{XY} = 0$.

Το μειονέκτημα της συνδιασποράς είναι ότι η τιμή της εξαρτάται από τις μονάδες μέτρησης των τ.μ. X και Y. Γι αυτό όταν θέλουμε να μετρήσουμε τη συσχέτιση δύο τ.μ. X και Y, χρησιμοποιούμε συνήθως το **συντελεστή συσχέτισης** (correlation coefficient), που συμβολίζεται $\rho_{XY} \equiv \mathrm{Corr}[X, Y]$ ή απλά ρ, και προκύπτει από την κανονικοποίηση της συνδιασποράς με το γινόμενο των τυπικών αποκλίσεων των X και Y

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.17}$$

Παραθέτουμε κάποιες ιδιότητες του συντελεστή συσχέτισης:

1. $-1 \leq \rho \leq 1$.

2. Αν οι τ.μ. X και Y είναι ανεξάρτητες είναι $\rho = 0$, αλλά $\rho = 0$ δε δηλώνει ότι οι X και Y είναι ανεξάρτητες αλλά απλά ότι δεν είναι γραμμικά συσχετισμένες (μπορεί δηλαδή να είναι μη-γραμμικά συσχετισμένες).

3. $\rho = -1$ ή $\rho = 1$ αν και μόνο αν $Y = a + \beta X$ για κάποιους αριθμούς a και β.

2.3 Γνωστές κατανομές μιας τ.μ.

Στη μελέτη μιας τ.μ. μας βοηθάει να έχουμε κάποια πρότυπα για την κατανομή της, δηλαδή κάποιες γνωστές συναρτήσεις $f_X(x)$ της τ.μ. X με γνωστές παραμέτρους. Επίσης σε πολλά πραγματικά προβλήματα η κατανομή μιας τ.μ. μπορεί να περιγραφεί ικανοποιητικά από κάποια γνωστή κατανομή. Θα παραθέσουμε εδώ δύο τέτοια παραδείγματα πολύ γνωστών κατανομών, μια για διακριτή και μια για συνεχή τ.μ..

2.3.1 Διωνυμική κατανομή

Πολλά προβλήματα και κυρίως πειράματα εμπεριέχουν **επαναλαμβανόμενες δοκιμές** (repeated trials). Για παράδειγμα μπορεί να θέλουμε να γνωρίζουμε την πιθανότητα η μια στις 5 βελόνες χαρακτικής να σπάσει σε ένα πείραμα αντοχής τάνυσης. Σε κάθε δοκιμή ορίζουμε δύο μόνο δυνατά

αποτελέσματα που συνήθως τα χαρακτηρίζουμε συμβολικά ως 'επιτυχία' και 'αποτυχία', χωρίς το όνομα να έχει απαραίτητα πραγματική σημασία ('επιτυχία' μπορεί να είναι το σπάσιμο της βελόνας). Υποθέτουμε ότι έχουμε κάνει n δοκιμές και η πιθανότητα 'επιτυχίας' p σε κάθε προσπάθεια είναι ίδια. Δοκιμές που τηρούν αυτές τις προϋποθέσεις λέγονται **δοκιμές Bernoulli**.

Ορίζουμε την τ.μ. X ως τον αριθμό των επιτυχίων σε n δοκιμές. Η X παίρνει τιμές στο σύνολο $\{0, 1, \ldots, n\}$. Η πιθανότητα να έχουμε x 'επιτυχίες' δίνεται από τη **διωνυμική** (binomial) σμπ, που συμβολίζεται $\mathrm{B}(n, p)$, και ορίζεται ως

$$f_X(x) = \mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{2.18}$$

όπου $\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}$ είναι ο *διωνυμικός συντελεστής* (binomial coefficient). Τα n και p είναι οι παράμετροι που ορίζουν τη διωνυμική κατανομή. Δηλώνουμε ότι μια τ.μ. X ακολουθεί διωνυμική κατανομή ως $X \sim \mathrm{B}(n, p)$. Η μέση τιμή και η διασπορά της X είναι

$$\mu = \mathrm{E}[X] = np \qquad \text{και} \qquad \sigma^2 = \mathrm{Var}[X] = np(1-p). \tag{2.19}$$

Παράδειγμα 2.1. Σε ένα πείραμα αντοχής τάνυσης δοκιμάζουμε 4 βελόνες χαρακτικής σε ένα συγκεκριμένο όριο τάνυσης. Η πιθανότητα να σπάσει η βελόνα σε μια δοκιμή είναι $p = 0.2$. Οι δοκιμές είναι τύπου Bernoulli. Μπορούμε να ορίσουμε την πιθανότητα να μη σπάσει καμιά βελόνα στις 4 δοκιμές από τη διωνυμική κατανομή ως

$$f_X(0) = \mathrm{P}(X = 0) = \binom{4}{0} 0.2^0 0.8^4 = 0.4096$$

Όμοια υπολογίζονται οι πιθανότητες όταν στις 4 δοκιμές σπάζει μια βελόνα κι όταν σπάζουν 2, 3 και 4 βελόνες

$$f_X(1) = 0.4096 \quad f_X(2) = 0.1536 \quad f_X(3) = 0.0256 \quad f_X(4) = 0.0016.$$

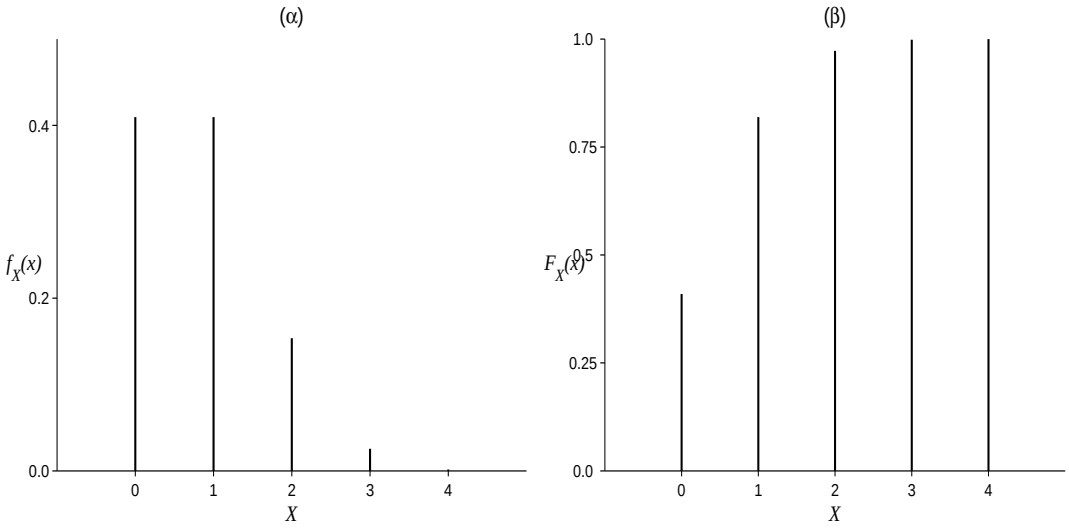
Με αυτόν τον τρόπο έχουμε ορίσει τη συνάρτηση σμπ $f_X(x)$ γι αυτό το παράδειγμα. Από την $f_X(x)$ ορίζεται εύκολα και η αθροιστική συνάρτηση $F_X(x) \equiv \mathrm{P}(X \leq x)$. Η γραφική παράσταση των συναρτήσεων $f_X(x)$ και $F_X(x)$ δίνονται στο Σχήμα 2.1.

Η πιθανότητα να σπάσει η βελόνα τουλάχιστον μια φορά είναι

$$\mathrm{P}(X \geq 1) = 1 - \mathrm{P}(X = 0) = 1 - 0.4096 = 0.5904$$

ενώ η πιθανότητα να σπάσει η βελόνα το πολύ δύο φορές δίνεται από την αθροιστική συνάρτηση κατανομής

$$F_X(2) \equiv \mathrm{P}(X \leq 2) = \sum_{x=0}^{2} \mathrm{P}(X = x) = 0.4096 + 0.4096 + 0.1536 = 0.9728.$$



Σχήμα 2.1: Γραφική παράσταση της συνάρτησης $f_X(x)$ στο (α) και της συνάρτησης $F_X(x)$ στο (β) για τον αριθμό βελόνων που σπάζουν σε n δοκιμές.

Η μέση τιμή για τον αριθμό 'επιτυχιών' (όπου επιτυχία είναι το σπάσιμο της βελόνας στη δοκιμή) είναι $\mathrm{E}[X] = 4 \cdot 0.2 = 0.8$ δηλαδή στις 4 δοκιμές περίπου μια φορά θα σπάζει η βελόνα. Η τυπική απόκλιση από αυτήν την τιμή είναι

$$\sigma = \sqrt{\mathrm{VarX}} = \sqrt{4 \cdot 0.2 \cdot 0.8} = 0.8.$$

2.3.2 Ομοιόμορφη κατανομή

Η πιο απλή συνεχής κατανομή είναι η ομοιόμορφη κατανομή που ορίζεται σε πεπερασμένο διάστημα $[a, b]$ και έχει σππ (δες Σχήμα 2.2)

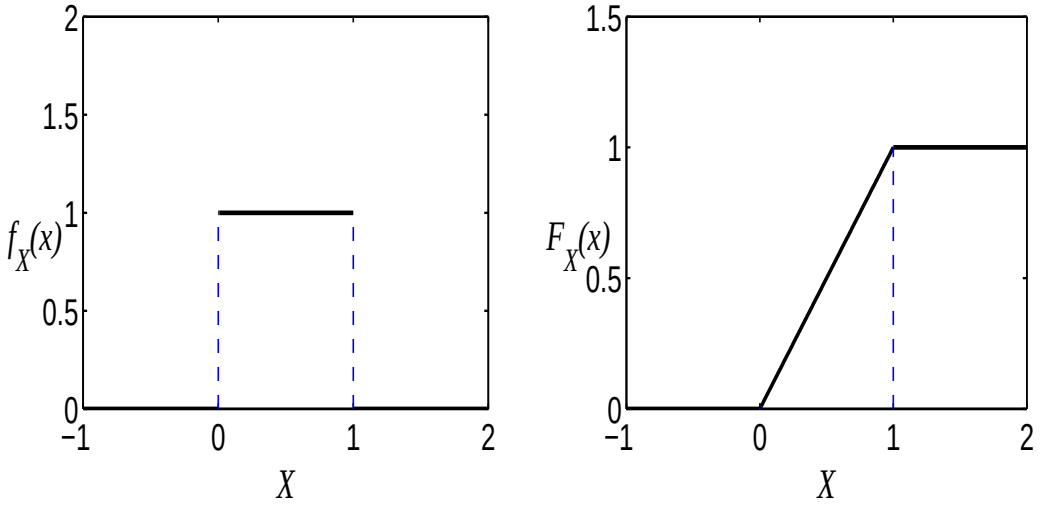
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{αλλού} \end{cases} \tag{2.20}$$

και ασκ

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x \ge b \end{cases} \tag{2.21}$$

Ο συμβολισμός που χρησιμοποιείται για να δείξουμε ότι μια τ.μ. X ακολουθεί ομοιόμορφη κατανομή στο διάστημα $[a, b]$ είναι $X \sim \mathrm{U}[a, b]$. Η μέση τιμή και διασπορά της X είναι

$$\mu = \mathrm{E}[X] = \frac{a+b}{2} \qquad \text{και} \qquad \sigma^2 = \mathrm{Var}[X] = \frac{(b-a)^2}{12}. \tag{2.22}$$



Σχήμα 2.2: Η συνάρτηση πυκνότητας πιθανότητας στο (α) και η αθροιστική συνάρτηση στο (β) της ομοιόμορφης κατανομής στο διάστημα $[a, b]$.

2.3.3 Δημιουργία τυχαίων αριθμών από δεδομένη κατανομή μέσω της ομοιόμορφης κατανομής

Συχνά σε προσομοιώσεις χρειάζεται να δημιουργήσουμε τυχαίους αριθμούς από κάποια δεδομένη κατανομή συνεχούς τ.μ., δηλαδή κατανομή που ορίζεται με κάποια γνωστή σππ ή εναλλακτικά ασκ και για ορισμένες τιμές των παραμέτρων της. Ένα χρήσιμο αποτέλεσμα που χρησιμοποιείται στη δημιουργία τυχαίων αριθμών από δεδομένη κατανομή είναι το παρακάτω θεώρημα.

Θεώρημα 2.1. Αν $X \sim \mathrm{U}[0, 1]$ τότε η τ.μ. $Y = F_Y^{-1}(X)$ έχει ασκ $F_Y(y)$.

Αν λοιπόν γνωρίζουμε την ασκ $F_Y(y)$ μιας τ.μ. Y για την οποία θέλουμε να παράγουμε τυχαίους αριθμούς, μπορούμε να υπολογίσουμε την αντίστροφη ασκ $F_Y^{-1}(.)$, που είναι πάντα εφικτό αφού η $F_Y(y)$ είναι μονότονη. Θεωρούμε πως έχουμε κάποια γεννήτρια συνάρτηση ψευδο-τυχαίων αριθμών για να παράγουμε τυχαίους αριθμούς x στο διάστημα $[0, 1]$, δηλαδή να παράγουμε τιμές x της $X \sim \mathrm{U}[0, 1]$ (δες άσκηση 1). Εφαρμόζοντας την $F_Y^{-1}(.)$ σε κάθε τιμή x της $X \sim \mathrm{U}[0, 1]$ θα μας δώσει την αντίστοιχη τιμή y της Y που ακολουθεί τη δεδομένη κατανομή με ασκ $F_Y(y)$.

Παράδειγμα 2.2. Η εκθετική κατανομή δίνεται από την σππ $f_Y(y) = \lambda e^{-\lambda y}$ και ασκ $F_Y(y) = 1 - e^{-\lambda y}$, όπου λ η παράμετρος της εκθετικής κατανομής (που είναι και η μέση τιμή). Θέτοντας $X \equiv F_Y(y)$, έχουμε $X \sim \mathrm{U}[0, 1]$ και

μπορούμε να δημιουργήσουμε τυχαίους αριθμούς από ομοιόμορφη κατανομή. Τότε για κάθε τέτοια τιμή x υπολογίζουμε την αντίστοιχη τιμή y από εκθετική κατανομή με παράμετρο λ από την αντίστροφη της $F_Y(y)$

$$y = -\frac{1}{\lambda}\ln(1-x).$$

2.3.4 Κανονική κατανομή

Η κανονική κατανομή είναι η σπουδαιότερη συνεχής κατανομή και αποτελεί τη βάση για πολλά στατιστικά μοντέλα και συμπεράσματα. Η σπουδαιότητα της οφείλεται κυρίως στο ότι περιγράφει ικανοποιητικά την κατανομή πολλών τυχαίων πραγματικών μεγεθών που παίρνουν συνεχείς αριθμητικές τιμές, αλλά προσεγγίζει ικανοποιητικά και πολλές διακριτές κατανομές. Σε πολλά τυχαία μεγέθη που μελετάμε παρατηρούμε ότι οι τιμές τους 'μαζεύονται' συμμετρικά γύρω από μια κεντρική τιμή και 'αραιώνουν' καθώς απομακρύνονται από αυτήν την κεντρική τιμή. Η κατάλληλη συνάρτηση πυκνότητας πιθανότητας για μια κατανομή τέτοιου τύπου 'τομή καμπάνας' είναι αυτή της **κανονικής κατανομής** (normal distribution) που ορίζεται ως

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty, \tag{2.23}$$

όπου οι παράμετροι μ και σ^2 που ορίζουν την κανονική κατανομή είναι η μέση τιμή και η διασπορά αντίστοιχα και η κατανομή συμβολίζεται ως $N(\mu, \sigma^2)$. Η αθροιστική συνάρτηση κατανομής $F_X(x)$ δίνεται από το ολοκλήρωμα της $f_X(x)$ όπως ορίστηκε στην (2.5). Στο Σχήμα 2.3 δίνονται σχηματικά η $f_X(x)$ και η $F_X(x)$.

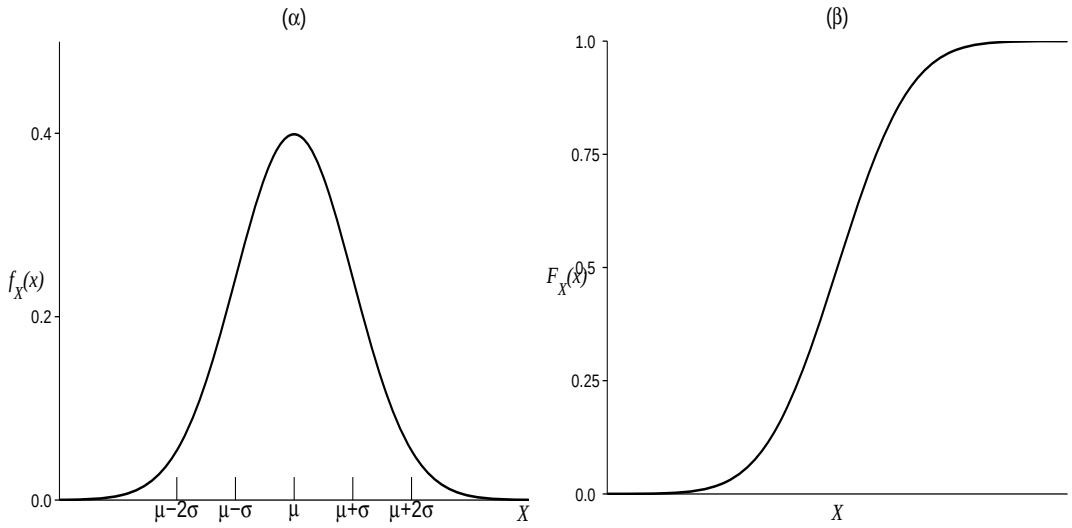
Φαίνεται ότι περίπου το 70% των τιμών της X βρίσκονται στο διάστημα $[\mu-\sigma, \mu+\sigma]$ και περίπου το 95% των τιμών της X βρίσκονται στο διάστημα $[\mu-2\,\sigma, \mu+2\,\sigma]$.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η **τυπική ή τυποποιημένη κανονική κατανομή** (standard normal distribution) που είναι η πιο απλή μορφή της κανονικής κατανομής, δηλαδή για $\mu = 0$ και $\sigma = 1$. Για να ξεχωρίσουμε την τ.μ. που ακολουθεί τυπική κανονική κατανομή τη συμβολίζουμε με Z ή z και είναι $Z \sim N(0,1)$. Η σππ είναι

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad -\infty < z < \infty. \tag{2.24}$$

Η αθροιστική συνάρτηση συμβολίζεται $\Phi(z)$ και είναι

$$\Phi(z) \equiv F_Z(z) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\,du \qquad -\infty < z < \infty. \tag{2.25}$$



Σχήμα 2.3: Η συνάρτηση πυκνότητας πιθανότητας στο (α) και η αθροιστική συνάρτηση στο (β) της κανονικής κατανομής.

Οι τιμές της $\Phi(z)$ για διάφορες τιμές του z είναι πολύ χρήσιμες στη στατιστική γι αυτό και δίνονται σε στατιστικό πίνακα σε κάθε βιβλίο στατιστικής. Κάθε τ.μ. X που ακολουθεί κανονική κατανομή μπορεί να μετασχηματιστεί στη Z με τον απλό μετασχηματισμό

$$X \sim \mathrm{N}(\mu, \sigma^2) \quad \Longrightarrow \quad Z \equiv \frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1). \tag{2.26}$$

Μπορούμε λοιπόν να υπολογίσουμε οποιαδήποτε πιθανότητα για τη X από την $\Phi(z)$. Γενικά η πιθανότητα για $X \in [a,b]$ είναι

$$\mathrm{P}(a \le X \le b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{2.27}$$

Παράδειγμα 2.3. Το πάχος ενός κυλινδρικού σωλήνα είναι σχεδιασμένο από το εργοστάσιο να είναι μ, αλλά παρατηρείται ότι το πάχος δεν είναι σταθερό σε κάθε παραγόμενο σωλήνα αλλά αποκλίνει από το μ με τυπική απόκλιση $\sigma = 0.1\,\mathrm{mm}$. Υποθέτουμε λοιπόν ότι το πάχος του κυλινδρικού σωλήνα είναι τυχαία μεταβλητή X που ακολουθεί κανονική κατανομή, δηλαδή $X \sim \mathrm{N}(\mu, 0.1^2)$ (σε mm).

Έστω ότι θέλουμε να υπολογίσουμε την πιθανότητα η απόκλιση του πάχους του κυλινδρικού σωλήνα από το προδιαγεγραμμένο πάχος να μην είναι μεγαλύτερη από 0.1 mm. Έχουμε σύμφωνα με την (2.27)

$$\begin{aligned} \mathrm{P}(\mu - 0.1 \le X \le \mu + 0.1) &= \mathrm{P}(-1 \le Z \le 1) = \Phi(1) - \Phi(-1) \\ &= 0.8413 - 0.1587 = 0.6826, \end{aligned}$$

που συμφώνει με το αποτέλεσμα που αναφέρθηκε παραπάνω, δηλαδή ότι περίπου το 70% των τιμών της X βρίσκονται στο διάστημα $[\mu - \sigma, \mu + \sigma]$. Αντίστροφα, μπορούμε να προσδιορίσουμε ένα όριο για το σφάλμα (πάνω και κάτω από την προδιαγεγραμμένη τιμή μ) που αντιστοιχεί σε κάποια πιθανότητα, ας πούμε 0.05. Αν ονομάσουμε το σφάλμα ϵ έχουμε

$$\begin{aligned} \mathrm{P}(X \leq \mu - \epsilon \text{ ή } X \geq \mu + \epsilon) &= 0.05 \Rightarrow \\ \mathrm{P}(\mu - \epsilon \leq X \leq \mu + \epsilon) &= 0.95 \Rightarrow \\ \Phi\left(\frac{\epsilon}{0.1}\right) - \Phi\left(-\frac{\epsilon}{0.1}\right) &= 0.95 \Rightarrow \\ 2\Phi\left(\frac{\epsilon}{0.1}\right) - 1 &= 0.95 \Rightarrow \\ \Phi\left(\frac{\epsilon}{0.1}\right) &= 0.975. \end{aligned}$$

Από τον πίνακα της τυπικής κανονικής κατανομής βρίσκουμε πως η τιμή z που αντιστοιχεί για $\Phi(z) = 0.975$ είναι $z = 1.96$. Άρα με πιθανότητα 0.95 το πάχος του κυλινδρικού σωλήνα δεν αποκλίνει από τη μέση τιμή μ περισσότερο από $0.196\,\mathrm{mm}$.

2.3.5 Κανονικότητα

Στους λόγους που αναφέρθηκαν παραπάνω για τη σπουδαιότητα της κανονικής κατανομής, θα πρέπει να προστεθεί και η ιδιότητα της κανονικής κατανομής να 'έλκει' τα αθροίσματα τυχαίων μεταβλητών, που δεν είναι απαραίτητα κανονικές, δηλαδή η κατανομή των αθροισμάτων τ.μ. να προσεγγίζει την κανονική κατανομή. Οι περιορισμοί για να ισχύει αυτό είναι οι τυχαίες μεταβλητές να είναι ανεξάρτητες μεταξύ τους, να έχουν πεπερασμένη διασπορά και το άθροισμα να είναι αρκετά μεγάλο. Κάτω από αυτές τις συνθήκες ισχύει το *κεντρικό οριακό θεώρημα, ΚΟΘ,* (central limit theorem, CLT):

Θεώρημα 2.2. Έστω οι τ.μ. X_i, $i = 1, \ldots, n$, για n μεγάλο (συνήθως θεωρούμε $n > 30$) που έχουν κατανομές με μέσες τιμές μ_i και διασπορές σ_i^2 για $i = 1, \ldots, n$, αντίστοιχα. Τότε ισχύει

$$Y = \sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu_Y, \sigma_Y^2), \tag{2.28}$$

όπου η μέση τιμή της τ.μ. του αθροίσματος Y είναι $\mu_Y = \sum_{i=1}^{n} \mu_i$ και η διασπορά είναι $\sigma_Y^2 = \sum_{i=1}^{n} \sigma_i^2$.

Προφανώς όταν οι τ.μ. X_i έχουν την ίδια κατανομή με μέση τιμή μ και διασπορά σ^2, τότε ισχύει $\mu_Y = n\mu$ και $\sigma_Y^2 = n\sigma^2$.

Για το μέσο όρο των τ.μ. X_i, $i = 1, \ldots, n$, με ίδια κατανομή, από το ΚΟΘ ισχύει

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathrm{N}(\mu, \sigma^2/n), \tag{2.29}$$

δηλαδή ο μέσος όρος ακολουθεί κανονική κατανομή με την ίδια μέση τιμή όπως οι τ.μ. X_i και διασπορά $\sigma^2_{\bar{X}} = \sigma^2/n$.

Ασκήσεις Κεφαλαίου 2

1. Επιβεβαίωσε τον ορισμό της πιθανότητας ως το όριο της σχετικής συχνότητας για αριθμό επαναλήψεων να τείνει στο άπειρο. Προσομοίωσε τη ρίψη ενός νομίσματος n φορές χρησιμοποιώντας τη γενέτειρα συνάρτηση τυχαίων αριθμών, είτε από ομοιόμορφη διακριτή κατανομή (δίτιμη για 'κορώνα' και 'γράμματα'), ή από ομοιόμορφη συνεχή κατανομή στο διάστημα $[0, 1]$ χρησιμοποιώντας κατώφλι 0.5 (π.χ. αριθμός μικρότερος του 0.5 είναι 'κορώνα' και μεγαλύτερος 'γράμματα'). Επανέλαβε το πείραμα για αυξανόμενα n και υπολόγισε κάθε φορά την αναλογία των 'γραμμάτων' στις n επαναλήψεις. Κάνε την αντίστοιχη γραφική παράσταση της αναλογίας για τα διαφορετικά n.

   Βοήθεια (`matlab`*):* Για τη δημιουργία των τυχαίων αριθμών χρησιμοποίησε τη συνάρτηση `rand` ή `unidrnd`.

2. Δημιούργησε 1000 τυχαίους αριθμούς από εκθετική κατανομή με παράμετρο $\lambda = 1$ χρησιμοποιώντας την τεχνική που δίνεται στην Παρ. 2.3.3. Κάνε το ιστόγραμμα των τιμών και στο ίδιο σχήμα την καμπύλη της εκθετικής σππ $f_X(x) = \lambda e^{-\lambda x}$.

   Βοήθεια (`matlab`*):* Το ιστόγραμμα δίνεται με τη συνάρτηση `hist`.

3. Δείξε με προσομοίωση ότι όταν δύο τ.μ. X και Y δεν είναι ανεξάρτητες δεν ισχύει η ιδιότητα $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$. Για να το δείξεις θεώρησε μεγάλο πλήθος τιμών n από X και Y που ακολουθούν τη διμεταβλητή κανονική κατανομή.

   Βοήθεια (`matlab`*):* Για τον υπολογισμό της διασποράς από n παρατηρήσεις, χρησιμοποίησε τη συνάρτηση `var`. Για να δημιουργήσεις παρατηρήσεις από διμεταβλητή κανονική κατανομή χρησιμοποίησε τη συνάρτηση `mvnrnd`.

4. Ισχύει $\mathrm{E}[1/X] = 1/\mathrm{E}[X]$; Διερεύνησε το υπολογιστικά για X από ομοιόμορφη συνεχή κατανομή στο διάστημα $[1, 2]$ υπολογίζοντας τους αντίστοιχους μέσους όρους για αυξανόμενο μέγεθος επαναλήψεων n. Κάνε κατάλληλη γραφική παράσταση για τις δύο μέσες τιμές και τα διαφορετικά n. Τι συμβαίνει αν το διάστημα της ομοιόμορφης κατανομής είναι $[0, 1]$ ή $[-1, 1]$;

5. Το μήκος X των σιδηροδοκών που παράγονται από μια μηχανή, είναι γνωστό ότι κατανέμεται κανονικά $X \sim \mathrm{N}(4, 0.01)$. Στον ποιοτικό έλεγχο που ακολουθεί αμέσως μετά την παραγωγή απορρίπτονται όσοι σιδηροδοκοί έχουν μήκος λιγότερο από 3.9. Ποια είναι η πιθανότητα

μια σιδηροδοκός να καταστραφεί; Που πρέπει να μπει το όριο για να καταστρέφονται το πολύ το 1% των σιδηροδοκών;

Βοήθεια (`matlab`*):* Η αθροιστική συνάρτηση κανονικής κατανομής δίνεται με τη συνάρτηση `normcdf`. Η αντίστροφη της δίνεται με τη συνάρτηση `norminv`.

6. Δείξε ότι ισχύει το ΚΟΘ με προσομοίωση. Έστω $n = 100$ τ.μ. από ομοιόμορφη κατανομή στο διάστημα $[0, 1]$ και έστω Y η μέση τιμή τους. Υπολόγισε $N = 10000$ τιμές της Y και σχημάτισε το ιστόγραμμα των τιμών μαζί με την καμπύλη της κανονικής κατανομής.

   Βοήθεια (`matlab`*):* Το ιστόγραμμα δίνεται με τη συνάρτηση `hist`.