ΣΤΑΤΙΣΤΙΚΗ 2

Σάμης Τρέβεζας

ΣΑΜΗΣ ΤΡΕΒΕΖΑΣ
Λέκτορας
Τμήμα Μαθηματικών
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνων

Στατιστική ΙΙ

Σημειώσεις σε εξέλιξη (09/03/19)

Περιεχόμενα

1 Εισαγωγή σε Τυχαία Διανύσματα και Πολυδιάστατες Κατανομές **1**

1.1 Εισαγωγή 1
1.2 Προκαταρκτικά 1
1.3 Τυχαία Διανύσματα 2
1.4 Πολυωνυμική Κατανομή 5
1.5 Πολυδιάστατη Κανονική Κατανομή 8
1.6 Σύγκλιση ακολουθίας τυχαίων διανυσμάτων και οριακά θεωρήματα 10

2 Απαραμετρική Στατιστική **14**

2.1 Εισαγωγικά 14
2.2 Συνάρτηση Κατανομής - Γενικευμένη αντίστροφη 17
2.3 Εμπειρική Συνάρτηση Κατανομής 20

1

Εισαγωγή σε Τυχαία Διανύσματα και Πολυδιάστατες Κατανομές

1.1 Εισαγωγή

Τα προβλήματα που αντιμετωπίζει ένας στατιστικός στην πράξη, στη συντριπτική τους πλειοψηφία, εμπεριέχουν την περιγραφή πολυδιάστατων χαρακτηριστικών που υπόκεινται σε πολλαπλών ειδών στοχαστικές εξαρτήσεις. Σε ένα πρακτικό πρόβλημα, υπάρχουν διαθέσιμα δεδομένα και καλείται ο Στατιστικός να κάνει μία μοντελοποίηση που να εξηγεί ικανοποιητικά την παραγωγή των δεδομένων, να εξάγει κάποια ορθά στατιστικά συμπεράσματα και αν είναι δυνατόν να μπορεί να προβλέπει στα όρια του στατιστικού σφάλματος καινούριες παρατηρήσεις που μπορούν να προκύψουν από το υπό μελέτη πρόβλημα. Βασική προϋπόθεση κατανόησης στατιστικών προβλημάτων που εμφανίζονται σε πραγματικές εφαρμογές είναι μία καλή γνώση του πιθανοθεωρητικού πλαισίου πάνω στο οποίο βασίζεται ένα στατιστικό μοντέλο. Ξεκινάμε με μία υπενθύμιση των βασικότερων στοιχείων που θα χρειαστούμε από τυχαίες μεταβλητές και βασικές κατανομές και στη συνέχεια θα κάνουμε μία επέκταση για πολυδιάστατες τυχαίες μεταβλητές και πολυδιάστατες κατανομές, στοιχεία απαραίτητα για τη μελέτη ενός προβλήματος που εμφανίζεται στην πράξη.

1.2 Προκαταρκτικά

Ξεκινάμε με μία σύντομη περιγραφική υπενθύμιση του χώρου πιθανότητας και της τυχαίας μεταβλητής. Η μοντελοποίηση ενός στοχαστικού φαινομένου ή ενός πειράματος τύχης γίνεται με την έννοια του χώρου πιθανότητας $(\Omega, \mathcal{A}, \mathbf{P})$. Το σύνολο Ω περιλαμβάνει τα δυνατά εξαγόμενα (αποτελέσματα) ω ενός πειράματος τύχης, αυτά που λέμε δειγματικά σημεία. Η συλλογή $\mathcal{A}$ αποτελείται από όλα τα σύνολα που μας ενδιαφέρει να αποδώσουμε πιθανότητα, αυτά που λέμε ενδεχόμενα, με την απαίτηση να ικανοποιούν κάποιες ιδιότητες κλειστότητας που είναι συμβατές με συνήθεις συνολοθεωρητικές πράξεις. Ο χώρος πιθανότητας συμπληρώνεται με μία βασική συνολοσυνάρτηση $\mathbf{P}$ που με το δικό της τρόπο θα αποδώσει τελικά πιθανότητα, δηλαδή μέτρο αβεβαιότητας, στα διάφορα ενδεχόμενα του πειράματος τύχης. Οι απαιτήσεις που έχει είναι μικρές, να δίνει πιθανότητα 1 στο βέβαιο ενδεχόμενο Ω και να είναι αθροιστική σε ξένα ενδεχόμενα, ακόμα και άπειρα το πλήθος, αρκεί να μπορούμε να τα απαριθμήσουμε (σ-προσθετικότητα).

Σε κάθε δειγματικό σημείο ω αποδίδουμε διάφορα ποσοτικά ή ποιοτικά χαρακτηριστικά ενδιαφέροντος. Κάθε τέτοιο χαρακτηριστικό X μπορούμε να το δούμε ως μία συνάρτηση. Αν οι τιμές της είναι μέσα στο σύνολο $\mathbb{R}$, τότε οδηγούμαστε στην έννοια της τυχαίας μεταβλητής. Απαιτεί βέβαια κάτι παραπάνω. Ο τυπικός ορισμός μιας τυχαίας μεταβλητής είναι μία εξειδίκευση της έννοιας της μετρήσιμης συνάρτησης (έννοια από τη θεωρία μέτρου) όταν αυτή παίρνει τιμές στο $\mathbb{R}$. Η ουσία είναι ότι η συνάρτηση X δημιουργεί έναν καινούριο δειγματικό χώρο, δεν ενδιαφερόμαστε πλέον για τα δειγματικά σημεία ω, αλλά για τις τιμές του χαρακτηριστικού X, δηλ., τα καινούρια σημεία $X(\omega) = x$. Για να μπορέσουμε να δούμε μέσα σε ένα κοινό πλαίσιο, όλα τα χαρακτηριστικά που παίρνουν πραγματικές τιμές μπορούμε να θεωρήσουμε ότι ο καινούριος δειγματικός χώρος είναι το $\mathbb{R}$. Όπως και πριν, έτσι και τώρα, χρειαζόμαστε και μία συλλογή ενδεχομένων που θέλουμε και μπορούμε να τους αποδώσουμε πιθανότητα. Για διάφορους λόγους που εξηγούνται στο μάθημα των Πιθανοτήτων ΙΙ, μία συλλογή που μας καλύπτει είναι τα σύνολα Borel του $\mathbb{R}$, που τα συμβολίζουμε $\mathcal{B}(\mathbb{R})$. Είναι αρκετά

πλούσια για να περιλαμβάνει σχεδόν όλα τα σύνολα που εμφανίζονται σε πρακτικές εφαρμογές, αλλά και για να σέβεται τις ιδιότητες κλειστότητας που αναφερθήκαμε πριν. Είναι ταυτόχρονα όσο πρέπει μικρή για να είναι συμβατή με τα αξιώματα που θέλουμε να ικανοποιούν οι συνολοσυναρτήσεις πιθανότητας πάνω σε υποσύνολα του $\mathbb{R}$.

Υπενθυμίζουμε λοιπόν ότι αν $(\Omega, \mathcal{A}, \mathbf{P})$ είναι ένας χώρος πιθανότητας, τότε πραγματική τ.μ. λέμε οποιαδήποτε πραγματική συνάρτηση $X : \Omega \to \mathbb{R}$ ικανοποιεί $\{X \in B\} := X^{-1}(B) \in \mathcal{A}$ για κάθε $B \in \mathcal{B}(\mathbb{R})$. Η απαίτηση να επιστρέφει όλα τα Borel υποσύνολα του $\mathbb{R}$ μέσα στην σ-άλγεβρα $\mathcal{A}$ είναι αναγκαία από την άποψη ότι μόνο σε αυτά τα σύνολα (τα ενδεχόμενα) αποδίδουμε πιθανότητα μέσω της συνολοσυνάρτησης $\mathbf{P}$. Με αυτήν τη διαδικασία κάθε τ.μ. καθορίζει με το δικό της τρόπο πιθανότητα σε όλα τα σύνολα Borel του $\mathbb{R}$. Πιο συγκεκριμένα, η συνολοσυνάρτηση $\mathbf{P}_X : \mathcal{B}(\mathbb{R}) \to [0, 1]$ που καθορίζεται μέσω της σχέσης $\mathbf{P}_X(B) = \mathbf{P}(X \in B)$ μπορεί να δειχθεί ότι ικανοποιεί τα αξιώματα της πιθανότητας. Ο όρος κατανομή της τ.μ. X αναφέρεται ακριβώς σε αυτήν τη συνολοσυνάρτηση και είναι αντικείμενο μελέτης της θεωρίας Πιθανοτήτων. Έτσι μπορούμε να αντιληφθούμε το χώρο $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbf{P}_X)$ ως έναν καινούριο χώρο πιθανότητας με δειγματικά σημεία τα $x \in \mathbb{R}$. Αυτός ο χώρος συνήθως μας καλύπτει για να πάρουμε ότι πληροφορία θέλουμε για τη X, καθώς η φύση του Ω συνήθως είναι αδιάφορη για το πρόβλημα που μελετάμε. Μάλιστα μας αρκεί πολλές φορές να περιοριστούμε μόνο στη γνώση της κατανομής της X. Έτσι φυσιολογικά οδηγούμαστε στην έννοια της ισονομίας τυχαίων μεταβλητών. Δύο τυχαίες μεταβλητές είναι ισόνομες αν $\mathbf{P}_X = \mathbf{P}_Y$, χωρίς να μας ενδιαφέρει καν σε ποιούς χώρους ορίζονται. Η μελέτη των κατανομών διευκολύνεται με τη χρήση απλούστερων συναρτήσεων, πιο βολικών αντικειμένων που χαρακτηρίζουν την κατανομή μιας τυχαίας μεταβλητής. Ενδεικτικά αναφέρουμε τη συνάρτηση κατανομής F_X, που ορίζεται από τη σχέση $F_X(x) = \mathbf{P}(X \leq x)$ για κάθε $x \in \mathbb{R}$, η οποία είναι προφανώς αύξουσα συνάρτηση. Θα αναφερθούμε διεξοδικά στις ιδιότητες της F_X στο κεφάλαιο της Απαραμετρικής Στατιστικής, καθώς αποτελεί το βασικό αντικείμενο ενδιαφέροντος. Θα θυμηθούμε μόνο ότι υπάρχει μία βασική διάκριση των τυχαίων μεταβλητών X σε διακριτές και συνεχείς τυχαίες μεταβλητές. Διακριτή είναι μία τ.μ. όταν η σ.κ. F_X αυξάνει μόνο με άλματα ασυνέχειας, και τα σημεία ασυνέχειας είναι ακριβώς εκείνα τα σημεία που παίρνει τιμές με θετική πιθανότητα. Αν αντίθετα η F_X είναι παντού συνεχής, τότε η τ.μ. λέγεται συνεχής. Από τις συνεχείς τ.μ. ξεχωρίζουμε εκείνες που έχουν συνάρτηση πυκνότητας πιθανότητας και είναι αυτές που καταχρηστικά λέμε συνεχείς. Το ορθό βέβαια είναι να λέμε απόλυτα συνεχείς τ.μ., αφού χαρακτηρίζονται από το γεγονός ότι η σ.κ. τους F_X είναι απόλυτα συνεχής συνάρτηση. Με άλλα λόγια, η F_X είναι τουλάχιστον τόσο ομαλή ώστε να επιτρέπεται η αναπαράσταση της μεταβολής της μεταξύ δύο σημείων μέσω ολοκληρώματος κάποιας (θετικής) συνάρτησης στο αντίστοιχο διάστημα που αυτά καθορίζουν. Είναι ακριβώς αυτή η συνάρτηση που χαρακτηρίζεται ως συνάρτηση πυκνότητας πιθανότητας f_X.

1.3 Τυχαία Διανύσματα

Η χρήση τυχαίων διανυσμάτων (τ.δ.) ή ισοδύναμα διανυσματικών τυχαίων μεταβλητών (δ.τ.μ.) είναι απαραίτητη για την περιγραφή περισσότερων του ενός χαρακτηριστικού ενδιαφέροντος σε ένα πείραμα τύχης και σχεδόν αποκλειστικά ασχολούμαστε με τέτοιες τ.μ. όταν χρησιμοποιούμε εκτιμήτριες σε πολυπαραμετρικά προβλήματα. Ο τυπικός ορισμός ενός τ.δ. είναι μία εξειδίκευση της έννοιας της μετρήσιμης συνάρτησης (έννοια από τη θεωρία μέτρου) όταν αυτή παίρνει τιμές στον $\mathbb{R}^d$. Όπως και στη μονοδιάστατη περίπτωση, έτσι και εδώ εμπλέκεται η έννοια των συνόλων Borel. Στο μάθημα αυτό δεν θα μας απασχολήσει τίποτα περισσότερο από το να θυμόμαστε ότι η συλλογή των συνόλων Borel, συμβολικά $\mathcal{B}(\mathbb{R}^d)$, αφορούν τα 'καλά' υποσύνολα του $\mathbb{R}^d$ στα οποία κάποιος θέλει και μπορεί να αποδώσει πιθανότητα για να μπορέσει να απαντήσει σχεδόν σε όλα τα προβλήματα που τον απασχολούν στην πράξη. Ξεκινάμε λοιπόν με τον τυπικό ορισμό ενός τυχαίου διανύσματος.

Ορισμός 1.1. (τυχαίο διάνυσμα-κατανομή-ισονομία)

Έστω $(\Omega, \mathcal{A}, \mathbf{P})$ ένας χώρος πιθανότητας και $d \geq 2$.

(i) Μία συνάρτηση $X : \Omega \to \mathbb{R}^d$ λέγεται *τυχαίο διάννυσμα* (τ.δ.) ή *διανυσματική τ.μ.* (δ.τ.μ.), αν $X^{-1}(B) \in \mathcal{A}$ για κάθε $B \in \mathcal{B}(\mathbb{R}^d)$.

(ii) Η συνολοσυνάρτηση $\mathbf{P}_X : \mathcal{B}(\mathbb{R}^d) \to [0,1]$ που καθορίζεται μέσω της σχέσης $\mathbf{P}_X(B) = \mathbf{P}(X \in B) = \mathbf{P}\left(X^{-1}(B)\right)$ λέγεται *κατανομή* της X, όπως στη μονοδιάστατη περίπτωση.

(iii) Δύο τυχαία διανύσματα X, Y, όχι κατ'ανάγκη ορισμένα στον ίδιο χώρο πιθανότητας, λέγονται *ισόνομα* αν $\mathbf{P}_X = \mathbf{P}_Y$ και τότε γράφουμε $X \overset{d}{=} Y$.

Η έννοια της συνάρτησης κατανομής επεκτείνεται φυσιολογικά και για τυχαία διανύσματα. Σε ότι ακολουθεί θεωρούμε το $\mathbb{R}^d$ εφοδιασμένο με τη σχέση μερικής διάταξης που προκύπτει μέσω της σχέσης $x \leq y$, αν και μόνο αν, $x_i \leq y_i$, για κάθε $1 \leq i \leq d$.

Ορισμός 1.2. Έστω X ένα τυχαίο διάνυσμα με τιμές στον $\mathbb{R}^d$. Η συνάρτηση $F_X : \mathbb{R}^d \to [0,1]$, όπου

$$F_X(x) := \mathbf{P}(X \leq x) = \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d) = \mathbf{P}\left(\cap_{i=1}^d \{X_i \leq x_i\}\right),$$

λέγεται *συνάρτηση κατανομής* του τ.δ. X ή *από κοινού συνάρτηση κατανομής* των $X_1, X_2, \ldots, X_d$. Όταν θέλουμε να δώσουμε έμφαση στις συνιστώσες τ.μ. γράφεται και ως $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$.

Δίνουμε παρακάτω χωρίς απόδειξη τρεις χρήσιμους χαρακτηρισμούς της ισονομίας τυχαίων διανυσμάτων. Η απόδειξη γίνεται στο μάθημα των Πιθανοτήτων ΙΙ.

Πρόταση 1.3. *Δύο τυχαία διανύσματα X,Y είναι ισόνομα, αν και μόνο αν,*

(i) $F_X = F_Y$, *δηλ., έχουν κοινή συνάρτηση κατανομής.*

(ii) $u^\top X \overset{d}{=} u^\top Y$, *για κάθε* $u \in \mathbb{R}^d$, *δηλ., κάθε γραμμικός συνδιασμός των συνιστωσών της X είναι ισόνομος με τον αντίστοιχο γραμμικό συνδιασμό των συνιστωσών της Y.*

(iii) $\mathbf{E}[f(X)] = \mathbf{E}[f(Y)]$, *για κάθε* $f : \mathbb{R}^d \to \mathbb{R}$ *συνεχή και φραγμένη συνάρτηση.*

Ο πρώτος χαρακτηρισμός είναι φυσιολογική γενίκευση του αντίστοιχου χαρακτηρισμού για μονοδιάστατες τ.μ.. Ο δεύτερος είναι αρκετά χρήσιμος σε αποδείξεις και δίνει τη δυνατότητα χαρακτηρισμού της κατανομής ενός τ.δ. μέσω της γνώσης μιας (άπειρης) οικογένειας κατανομών μονοδιάστατων τ.μ. που προκύπτουν μάλιστα μόνο μέσω γραμμικών μετασχηματισμών. Είναι μάλιστα θεωρητικά πιο γόνιμος, αφού επεκτείνεται και σε απειροδιάστατους διανυσματικούς χώρους μέσω της έννοιας του γραμμικού φραγμένου τελεστή (αντικείμενο μελέτης της συναρτησιακής ανάλυσης). Ο τελευταίος χαρακτηρισμός είναι επίσης αρκετά χρήσιμος και μπορεί να χρησιμοποιηθεί σε πιο αφηρημένους χώρους.

Οι έννοιες της μέσης τιμής και της διασποράς, όπως και της ροπής οποιασδήποτε τάξης, επεκτείνονται σε τυχαία διανύσματα. Θα περιοριστούμε εδώ μόνο στη μέση τιμή και τη διασπορά που είναι και οι πιο χρήσιμες. Είναι φανερό από τον χαρακτηρισμό της κατανομής που δίνεται στην Πρόταση 1.3–(ii) ότι μία επέκταση κάποιας έννοιας στην πολυδιάστατη περίπτωση θα μπορούσε να εμπνέεται απο την εφαρμογή της αντίστοιχης έννοιας στη μονοδιάστατη περίπτωση σε όλους τους γραμμικούς συνδιασμούς $u^\top X$. Λόγω της γραμμικότητας της μέσης τιμής, η μέση τιμή $\mathbf{E}(u^\top X)$ καθορίζεται πλήρως από το διάνυσμα των μέσων τιμών $(\mathbf{E}(X_1), \ldots, \mathbf{E}(X_d))$. Στην περίπτωση της διασποράς, δεν αρκεί η γνώση των συνιστωσών διασπορών, αλλά χρειάζεται η γνώση και των συνδιακυμάνσεων $\mathrm{Cov}(X_i, X_j)$. Παρ'όλα αυτά, η γνώση του πίνακα $\left(\mathrm{Cov}(X_i, X_j)\right)_{i,j}$ είναι αρκετή για τον καθορισμό της διασποράς $\mathbf{V}(u^\top X)$, δηλ., της διασποράς όλων των γραμμικών συνδιασμών της X. Οδηγούμαστε λοιπόν στους επόμενους ορισμούς.

Ορισμός 1.4. Έστω X ένα τυχαίο διάνυσμα με τιμές στον $\mathbb{R}^d$.

(i) Αν $\mathbf{E}\,|X_i| < +\infty$ για κάθε $1 \le i \le d$, τότε λέμε ότι το X *έχει μέση τιμή* $\mathbf{E}(X)$ που ορίζεται από τη σχέση:

$$\mu_X \equiv \mathbf{E}(X) := (\mathbf{E}(X_1), \mathbf{E}(X_2), \ldots, \mathbf{E}(X_d))\,.$$

Η συνθήκη ύπαρξης μέσης τιμής είναι ισοδύναμη με τη συνθήκη $\mathbf{E}\|X\| < +\infty$, όπου $\|\cdot\|$ είναι η ευκλείδια νόρμα στον $\mathbb{R}^d$, την οποία και θα χρησιμοποιούμε για απλότητα.

(ii) Αν $\mathbf{E}(X_i^2) < +\infty$ για κάθε $1 \le i \le d$, τότε ορίζουμε ως *διασπορά* του X τον πίνακα:

$$\Sigma_X \equiv \mathbf{V}(X) := \Big(\mathrm{Cov}(X_i, X_j)\Big)_{1\le i,j\le d}\,.$$

Ο πίνακας $\mathbf{V}(X)$ λέγεται και *πίνακας διασπορών–συνδιασπορών* (για εμφανείς λόγους) ή *πίνακας συνδιακύμανσης* και συμβολίζεται και με $\mathrm{Cov}(X)$. Η συνθήκη ύπαρξης της διασποράς είναι ισοδύναμη με τη συνθήκη $\mathbf{E}\|X\|^2 < +\infty$, την οποία και θα χρησιμοποιούμε για απλότητα.

(iii) Αν Y είναι ένα ακόμα τυχαίο διάνυσμα με τιμές στον $\mathbb{R}^s$ και $\mathbf{E}\|Y\|^2 < +\infty$, τότε ορίζουμε ως (σταυρωτή) *συνδιακύμανση* των τ.δ. X και Y, τον πίνακα

$$\Sigma_{X,Y} \equiv \mathbf{C}(X,Y) := \begin{pmatrix} \mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \ldots & \mathrm{Cov}(X_1,Y_s) \\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \ldots & \mathrm{Cov}(X_2,Y_s) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d,Y_1) & \mathrm{Cov}(X_d,Y_2) & \ldots & \mathrm{Cov}(X_d,Y_s) \end{pmatrix}$$

Τα τ.δ. X και Y λέγονται *ασυσχέτιστα*, αν ο $C(X,Y)$ είναι ο μηδενικός πίνακας.

Παρατήρηση 1.5. Από τους παραπάνω ορισμούς είναι φανερό ότι $\mathbf{C}(X,X) = \mathbf{V}(X)$ και $\mathbf{C}(Y,X) = \mathbf{C}(X,Y)^\top$. Παρατηρούμε λοιπόν εδώ ότι ο πίνακας συνδιακύμανσης μεταξύ δύο τ.δ. (έτσι όπως τον ορίσαμε) δεν είναι γενικά συμμετρικός, εκτός αν $X = Y$, οπότε ταυτίζεται με τον συμμετρικό πίνακα $\mathbf{V}(X)$.

Στην επόμενη πρόταση δίνουμε κάποιες στοιχειώδεις ιδιότητες της μέσης τιμής τυχαίων διανυσμάτων.

Πρόταση 1.6. *Έστω X, Y τυχαία διανύσματα με τιμές στον $\mathbb{R}^d$ με $\mathbf{E}\|X\|, \mathbf{E}\|Y\| < +\infty$ και A, b πίνακας και διάνυσμα κατάλληλης διάστασης. Τότε ισχύουν τα εξής:*

$$\begin{aligned} \mathbf{E}(AX+b) &= A\,\mathbf{E}(X) + b \\ \mathbf{E}(X+Y) &= \mathbf{E}(X) + \mathbf{E}(Y) \end{aligned}$$

Στην επόμενη πρόταση δίνουμε στοιχειώδεις ιδιότητες της διασποράς τυχαίων διανυσμάτων.

Πρόταση 1.7. *Έστω $X, Y \in \mathbb{R}^d$ τυχαία διανύσματα με $\mathbf{E}\|X\|^2, \mathbf{E}\|Y\|^2 < +\infty$ και A, b πίνακας και διάνυσμα κατάλληλης διάστασης. Τότε ισχύουν τα εξής:*

$$\begin{aligned} \mathbf{V}(X) &= \mathbf{E}\left((X-\mu_X)(X-\mu_X)^\top\right) = \mathbf{E}(XX^\top) - \mathbf{E}(X)\,\mathbf{E}(X^\top) \\ \mathbf{V}(AX+b) &= A\,\mathbf{V}(X)\,A^\top \\ \mathbf{V}(X+Y) &= \mathbf{V}(X) + \mathbf{V}(Y) + \mathbf{C}(X,Y) + \mathbf{C}(Y,X). \end{aligned}$$

Ακολουθούν ιδιότητες που συνδέονται με τη συνδιακύμανση δύο τυχαίων διανυσμάτων.

Πρόταση 1.8. *Έστω X, Y τυχαία διανύσματα με τιμές στον $\mathbb{R}^d$ και $\mathbb{R}^s$ αντίστοιχα. Αν υποθέσουμε ότι $\mathbf{E}\|X\|^2, \mathbf{E}\|Y\|^2 < +\infty$, τότε ισχύουν τα εξής:*

$$\begin{aligned} \mathbf{C}(X,Y) &= \mathbf{E}\left((X-\mu_X)(Y-\mu_Y)^\top\right) = \mathbf{E}(XY^\top) - \mathbf{E}(X)\,\mathbf{E}(Y^\top) \\ \mathbf{C}(AX+b, BY+c) &= A\,\mathbf{C}(X,Y)\,B^\top \end{aligned}$$

Ορισμός 1.9. (ανεξαρτησία τ.δ.)
Δύο τ.δ. $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^s$ λέγονται *ανεξάρτητα*, αν για κάθε $A \in \mathcal{B}(\mathbb{R}^d)$ και $B \in \mathcal{B}(\mathbb{R}^s)$

$$\mathbf{P}(X \in A,\ Y \in B) = \mathbf{P}(X \in A)\,\mathbf{P}(Y \in B). \tag{1.1}$$

Χρησιμοποιείται αρκετά ο συμβολισμός $X \perp\!\!\!\perp Y$ για να δηλώσουμε την ανεξαρτησία των X και Y.

Μπορούν να αποδειχθούν και οι επόμενοι χρήσιμοι χαρακτηρισμοί της ανεξαρτησίας τ.δ..

Πρόταση 1.10. *Δύο τ.δ. $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^s$ είναι ανεξάρτητα, αν και μόνο αν,*

(i) για κάθε $x \in \mathbb{R}^d$, $y \in \mathbb{R}^s$

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

όπου ως $F_{X,Y}$ συμβολίζουμε την από κοινού συνάρτηση κατανομής των τ.δ. X, Y.

(ii) $u^tX \perp\!\!\!\perp v^tY$, για κάθε $u \in \mathbb{R}^d$, $v \in \mathbb{R}^s$, δηλ., κάθε γραμμικός συνδιασμός των συνιστωσών της X είναι ανεξάρτητος από κάθε γραμμικό συνδιασμό των συνιστωσών της Y.

(iii) για κάθε $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^s \to \mathbb{R}$ συνεχείς και φραγμένες

$$\mathbf{E}[f(X)g(U)] = \mathbf{E}[f(X)]\,\mathbf{E}[g(U)]. \tag{1.2}$$

Η παραπάνω σχέση ισχύει και για οποιεσδήποτε f, g (μετρήσιμες), αρκεί να υπάρχουν οι μέσες τιμές.

Από τα παραπάνω συμπεραίνουμε τα εξής:

Πόρισμα 1.11. *Αν τα τ.δ. $X \in \mathbb{R}^d$, $Y \in \mathbb{R}^s$ είναι ανεξάρτητα, τότε όταν ορίζονται καλά*

$$\begin{aligned}
\mathbf{E}(XY^\top) &= \mathbf{E}(X)\,\mathbf{E}(Y^\top) \\
\mathbf{C}(X, Y) &= 0_{d\times s} \qquad (\Rightarrow\ X, Y \text{ ασυσχέτιστες}) \\
\mathbf{V}(X + Y) &= \mathbf{V}(X) + \mathbf{V}(Y), \qquad \text{για } s = d,
\end{aligned}$$

όπου $0_{d\times s}$ είναι ο μηδενικός πίνακας διάστασης $d \times s$.

Στη συνέχεια θα αναφέρουμε και θα μελετήσουμε τις στοιχειώδεις ιδιότητες της πιο σημαντικής διακριτής και της πιο σημαντικής συνεχούς δ.τ.μ. που είναι η πολυωνυμική και η πολυδιάστατη κανονική τ.μ. αντίστοιχα.

1.4 Πολυωνυμική Κατανομή

Η πολυωνυμική κατανομή γενικεύει τη διωνυμική κατανομή. Είναι γνωστό ότι μία τ.μ. που ακολουθεί τη διωνυμική κατανομή ερμηνεύεται ως το πλήθος των επιτυχιών σε ένα δεδομένο αριθμό ανεξάρτητων δοκιμών Bernoulli με σταθερή πιθανότητα επιτυχίας σε κάθε δοκιμή. Αν η επιτυχία δεν είναι μονοσήμαντα ορισμένη, αλλά θεωρήσουμε για παράδειγμα ότι υπάρχουν d είδη επιτυχιών, τότε είναι φυσιολογικό να αναζητήσουμε την κατανομή της δ.τ.μ. που εκφράζει την από κοινού καταγραφή του πλήθους των επιτυχιών $i-$είδους σε ένα δεδομένο αριθμό ανεξάρτητων δοκιμών, όπου σε κάθε δοκιμή έχουμε το πολύ μία επιτυχία $i-$είδους με σταθερή πιθανότητα. Για παράδειγμα, σε 20 ρίψεις ενός ζαριού μπορεί κάποιος να ενδιαφέρεται για την απο κοινού κατανομή του πλήθους των ρίψεων που ήρθαν 1 μαζί με αυτές που ήρθαν 6. Εδώ, σε κάθε ρίψη μπορούμε να θεωρήσουμε την ένδειξη 1 ως επιτυχία πρώτου είδους και την ένδειξη 6 ως επιτυχία δεύτερου είδους. Αν το ζάρι είναι δίκαιο τότε και τα δύο είδη επιτυχιών έχουν σταθερή πιθανότητα $1/6$ σε κάθε δοκιμή. Αν στο ίδιο τυχαίο πείραμα σε κάθε δοκιμή αντιστοιχίσουμε την αποτυχία, δηλ. την εμφάνιση κάποιας από τις έδρες 2, 3, 4 ή 5, σε επιτυχία τρίτου είδους και συμπεριλάβουμε την τ.μ. που καταγράφει το πλήθος των επιτυχιών τρίτου

είδους στις δύο προηγούμενες, τότε είναι φανερό ότι δεν κερδίζουμε παραπάνω πληροφορία αφού η τελευταία καθορίζεται πλήρως ως η διαφορά του πλήθους των συνολικών δοκιμών από το άθροισμα των επιτυχιών πρώτου και δεύτερου είδους. Παρ'όλα αυτά συναντάμε στη βιβλιογραφία και τις δύο αυτές ισοδύναμες μορφές της πολυωνυμικής κατανομής. Πριν δώσουμε τον τυπικό ορισμό τους, ας δούμε μία χρήσιμη επέκταση του διωνυμικού συντελεστή.

Υπενθυμίζουμε ότι ο διωνυμικός συντελεστής $\binom{n}{k}$ φέρει το όνομα αυτό λόγω του ότι προκύπτει ως συντελεστής στο διωνυμικό ανάπτυγμα. Συγκεκριμένα,

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \quad \text{όπου } \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Εύκολα δείχνεται (με επαγωγή στο n) η ισχύς του πολυωνυμικού θεωρήματος, το οποίο μας επιτρέπει να γενικεύσουμε το διωνυμικό ανάπτυγμα από πολυώνυμα 2 μεταβλητών σε πολυώνυμα d μεταβλητών:

$$(x_1 + x_2 + \ldots + x_d)^n = \sum_{k=0}^{n} \frac{n!}{n_1! n_2! \ldots n_d!} x_1^{n_1} x_2^{n_2} \ldots x_d^{n_d}, \quad \text{όπου } n_1 + n_2 + \ldots n_d = n. \tag{1.3}$$

Αν κάνουμε μία αρίθμηση (διάκριση) των n παραγόντων που εμφανίζονται στο αριστερό μέλος:

$$\underbrace{(x_1 + x_2 + \ldots + x_d)}_{1} * \underbrace{(x_1 + x_2 + \ldots + x_d)}_{2} * \ldots * \underbrace{(x_1 + x_2 + \ldots + x_d)}_{n},$$

τότε είναι φανερό ότι το ανάπτυγμα συνίσταται σε ένα άθροισμα διακεκριμένων μονωνύμων βαθμού n (επιλέγοντας μία μεταβλητή από κάθε παράγοντα) της μορφής $x_1^{n_1} x_2^{n_2} \ldots x_d^{n_d}$, όπου κάθε εκθέτης αντιστοιχεί στο πλήθος των φορών που επιλέχθηκε η αντίστοιχη μεταβλητή, και σε κάθε μονώνυμο αντιστοιχεί ένας συντελεστής που εκφράζει το πλήθος των διαφορετικών τρόπων με τους οποίους μπορούμε να επιλέξουμε μέσα από το σύνολο των n διακεκριμένων παραγόντων (θέσεων), τις n_1 θέσεις για την επιλογή του x_1, τις n_2 θέσεις για την επιλογή του x_2 και τις υπολοιπόμενες n_d θέσεις για την επιλογή του x_d. Η σκέψη αυτή μας οδηγεί σε μία δεύτερη συνδιαστική απόδειξη (χωρίς επαγωγή) του πολυωνυμικού αναπτύγματος (1.3) κάνοντας χρήση της πολλαπλασιαστικής αρχής (ολοκληρώστε την απόδειξη). Από τα παραπάνω

Ορισμός 1.12. Έστω $n, n_i \in \mathbb{N}$, $i = 1, 2, \ldots, d$ με $n_1 + n_2 + \ldots n_d = n$. Λέμε *πολυωνυμικό συντελεστή* τον αριθμό

$$\binom{n}{n_1, n_2, \ldots, n_d} := \frac{n!}{n_1! n_2! \ldots n_d!}, \tag{1.4}$$

που εμφανίζεται ως συντελεστής στο πολυωνυμικό ανάπτυγμα (1.3).

Παρατήρηση 1.13. (i) Στην ειδική περίπτωση που $d = 2$, ο πολυωνυμικός συντελεστής αντιστοιχεί στο διωνυμικό συντελεστή, έτσι έχουμε $\binom{n}{k,n-k} = \binom{n}{k}$. Παραδοσιακά προτιμάται ο τελευταίος συμβολισμός που είναι και ο πιο σύντομος.

(ii) Κάποιες φορές ο πολυωνυμικός συντελεστής (1.6) εμφανίζεται στη μορφή

$$\binom{n}{n_1, n_2, \ldots, n_{d-1}} := \frac{n!}{n_1! \ldots n_{d-1}! (n - n_1 - \ldots - n_{d-1})!}. \tag{1.5}$$

Αυτό δεν θα πρέπει να προκαλεί σύγχυση αφού οι 2 εκφράσεις διακρίνονται από το αν οι δείκτες που εκφράζουν το πλήθος των επιλογών αθροίζουν ή όχι στο συνολικό πλήθος των στοιχείων n.

(iii) Είναι φανερό από τα σχόλια που ακολουθούν το πολυωνυμικό ανάπτυγμα ότι ο πολυωνυμικός συντελεστής εκφράζει το πλήθος των διαφορετικών τρόπων με τους οποίους μπορούμε να διαμερίσουμε ένα σύνολο n στοιχείων σε d υποσύνολα με n_i στοιχεία το καθένα, $1 \le i \le d$.

Είμαστε τώρα σε θέση να ορίσουμε την πολυωνυμική κατανομή που αντιπροσωπεύει την κατανομή ενός τυχαίου διανύσματος που καταγράφει το πλήθος των επιτυχιών όλων των ειδών σε n ανεξάρτητες δοκιμές με σταθερές πιθανότητες επιτυχίας κάθε είδους σε κάθε δοκιμή, όταν σε κάθε μία από αυτές έχουμε επιτυχία μόνο ενός είδους ή το πολύ ενός. Για να αποφύγουμε ενδεχόμενη σύγχυση θα διακρίνουμε αυτές τις δύο ισοδύναμες μορφές, οι οποίες διαφέρουν μόνο στο αν θα ονομάσουμε την αποτυχία ως επιτυχία ενός επιπρόσθετου είδους με ταυτόχρονη καταγραφή του πλήθους τους ή απλά θα τις εξαιρέσουμε από την ανάλυσή μας. Πρακτικά, αυτό φαίνεται από το αν τα σημεία που έχουν θετική πιθανότητα έχουν σταθερό άθροισμα ή όχι αντίστοιχα.

Ορισμός 1.14. Έστω $X = (X_1, X_2, \ldots, X_d)$ ένα τ.δ. με $d \geq 2$. Θα λέμε ότι η X ακολουθεί την *πολυωνυμική κατανομή* $\mathcal{M}(n, p_1, p_2, \ldots, p_d)$ με παραμέτρους $n \geq 1$ και $p_1, p_2, \ldots, p_d > 0$,
όπου $\sum_{i=1}^{d} p_i = 1$, αν

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_d = x_d) = \binom{n}{x_1, x_2, \ldots, x_d} p_1^{x_1} p_2^{x_2} \ldots p_d^{x_d}, \text{ όπου } x_i \in \mathbb{N},\ 1 \leq i \leq d, \text{ και } \sum_{i=1}^{d} x_i = n. \tag{1.6}$$

Αν η X ορίζεται όπως παραπάνω με $d \geq 1$, θετικές παραμέτρους που ικανοποιούν $0 < \sum_{i=1}^{d} p_i < 1$ και στήριγμα $S_X = \left\{x \in \mathbb{N}^d : 0 \leq \sum_{i=1}^{d} x_i \leq n\right\}$, τότε θα λέμε ότι η X ακολουθεί την *πολυωνυμική κατανομή* $\mathcal{M}^*(n, p_1, p_2, \ldots, p_d)$.

Παρατήρηση 1.15. (i) Με τον παραπάνω ορισμό είναι φανερό ότι η διωνυμική κατανομή $\mathrm{Bin}(n, p)$ ταυτίζεται με την $\mathcal{M}^*(n, p)$

(ii) Για $n = 1$, η πολυωνυμική κατανομή $\mathcal{M}(1, p_1, p_2, \ldots, p_d)$ ή $\mathcal{M}^*(1, p_1, p_2, \ldots, p_d)$, αναφέρεται ως *κατηγορική* και θα λέμε ότι το τ.δ. $C \sim Cat(p_1, p_2, \ldots, p_d)$ $(d \geq 2)$ ή $C \sim Cat^*(p_1, p_2, \ldots, p_d)$ $(d \geq 1)$ αντίστοιχα. Η κατηγορική κατανομή γενικεύει την κατανομή Bernoulli, όπου αντί για τον καθορισμό μόνο μιας πιθανότητας επιτυχίας, προσδιορίζει τις πιθανότητες όλων των επιτυχιών i-είδους. Ένα τ.δ. $C \sim Cat(p_1, p_2, \ldots, p_d)$ παίρνει τιμές στο σύνολο $\{e_1, e_2, \ldots, e_d\}$, δηλ. στα διανύσματα της ορθοκανονικής βάσης του $\mathbb{R}^d$ και $\mathbf{P}(C = e_i) = p_i$ για κάθε $i = 1, \ldots, d$.

(iii) Είναι φανερό από τον ορισμό και την ερμηνεία της πολυωνυμικής κατανομής ότι αν $X \sim \mathcal{M}(n, p_1, p_2, \ldots, p_d)$, τότε μπορούμε να έχουμε την εξής χρήσιμη αναπαράσταση

$$X = \sum_{i=1}^{n} C_i, \tag{1.7}$$

όπου $C_i \sim Cat(p_1, p_2, \ldots, p_d)$, $i = 1, 2, \ldots, n$, και είναι μεταξύ τους ανεξάρτητα τ.δ.. Η έκφραση αυτή του X που ακολουθεί πολυωνυμική κατανομή ως άθροισμα ανεξάρτητων και ισόνομων κατηγορικών τ.δ. είναι η γενίκευση στα πολυδιάστατα της αντίστοιχης αναπαράστασης μιας τυχαίας μεταβλητής που ακολουθεί τη διωνυμική κατανομή ως άθροισμα ανεξάρτητων και ισόνομων τ.μ. που ακολουθούν κατανομή Bernoulli. Οι αναπαραστάσεις αυτές είναι χρήσιμες, όπως γνωρίζουμε και από τις Πιθανότητες Ι, για τον υπολογισμό μέσων τιμών και διασπορών.

Πρόταση 1.16. *Αν ένα τ.δ. $X \sim \mathcal{M}(n, p_1, p_2, \ldots, p_d)$, τότε η μέση τιμή του είναι*

$$\mathbf{E}(X) = np = n\,(p_1, p_2, \ldots, p_d)$$

και η διασπορά του (ο πίνακας διασποράς) είναι

$$\mathbf{V}(X) = n[diag(p) - pp^\top] = n \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \ldots & -p_1p_d \\ -p_2p_1 & p_2(1-p_2) & \ldots & -p_2p_d \\ \vdots & \vdots & \ddots & \vdots \\ -p_dp_1 & -p_dp_2 & \ldots & p_d(1-p_d) \end{pmatrix}$$

Απόδειξη:
Χρησιμοποιώντας την αναπαράσταση (1.7) έχουμε

$$\mathbf{E}(X) = \mathbf{E}\left(\sum_{i=1}^{n} C_i\right) = \sum_{i=1}^{n} \mathbf{E}(C_i) = n\ \mathbf{E}(C),$$

όπου $C \sim Cat(p)$. Όμως $C \in \{e_1, \ldots, e_d\}$ και $\mathbf{P}(C = e_i) = p_i$, $1 \le i \le d$. Επειδή $\{C_i = 1\} = \{C = e_i\}$ συμπεραίνουμε ότι

$$\mathbf{E}(C) = \Big(\mathbf{E}(C_1), \ldots, \mathbf{E}(C_d)\Big) = \Big(\mathbf{P}(C_1 = 1), \ldots, \mathbf{P}(C_d = 1)\Big) = (p_1, \ldots, p_d) = p.$$

Επίσης, λόγω της ανεξαρτησίας, για τον υπολογισμό της διασποράς έχουμε

$$\mathbf{V}(X) = \mathbf{V}\left(\sum_{i=1}^{n} C_i\right) = \sum_{i=1}^{n} \mathbf{V}(C_i) = n\,\mathbf{V}(C).$$

Για τον τελευταίο πίνακα έχουμε

$$\mathbf{V}(C) = \mathbf{E}(CC^\top) - \mathbf{E}(C)\,\mathbf{E}(C^\top) = \mathrm{diag}(p) - pp^\top,$$

χρησιμοποιώντας ότι $C_i \sim \mathrm{Be}(p_i)$, ότι $C_iC_j = 0$ για $i \ne j$, και άρα

$$\mathbf{E}(CC^\top) = \begin{pmatrix} \mathbf{E}(C_1^2) & \mathbf{E}(C_1C_2) & \ldots & \mathbf{E}(C_1C_d) \\ \mathbf{E}(C_2C_1) & \mathbf{E}(C_2^2) & \ldots & \mathbf{E}(C_2C_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}(C_dC_1) & \mathbf{E}(C_dC_2) & \ldots & \mathbf{E}(C_d^2) \end{pmatrix} = \begin{pmatrix} p_1 & 0 & \ldots & 0 \\ 0 & p_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_d \end{pmatrix}.$$

□

1.5 Πολυδιάστατη Κανονική Κατανομή

Η πιο σημαντική συνεχής πολυδιάστατη κατανομή είναι η κανονική κατανομή για τους ίδιους λόγους που είναι και στη μονοδιάστη περίπτωση. Θα στηρίξουμε τον επόμενο ορισμό της κανονικής κατανομής στην Πρόταση 1.3-(ii), χρησιμοποιώντας τις κατανομές όλων των γραμμικών συνδιασμών των συνιστωσών ενός τ.δ..

Ορισμός 1.17. Έστω $X = (X_1, X_2, \ldots, X_d)$ ένα τ.δ. με $d \ge 1$, $\mu \in \mathbb{R}^d$ ένα διάνυσμα και Σ ένας συμμετρικός και θετικά ημιορισμένος πίνακας. Θα λέμε ότι το X ακολουθεί την *κανονική κατανομή* $\mathcal{N}_d(\mu, \Sigma)$ με παραμέτρους μ και Σ, αν για κάθε $u \in \mathbb{R}^d$

$$u^\top X \sim \mathcal{N}_1(u^\top \mu,\ u^\top \Sigma\, u). \tag{1.8}$$

Πρόταση 1.18. *Αν ένα τ.δ.* $X \sim \mathcal{N}_d(\mu, \Sigma)$*, τότε* $\mathbf{E}(X) = \mu$ *και* $\mathbf{V}(X) = \Sigma$.

Απόδειξη:
Από την (1.8) για $u = e_i$, $i = 1, \ldots, d$, έχουμε

$$X_i = e_i^\top X \sim \mathcal{N}_1\left(e_i^\top \mu,\ e_i^\top \Sigma\, e_i\right) = \mathcal{N}_1(\mu_i,\ \Sigma_{ii}),$$

από το οποίο συμπεραίνουμε ότι $\mathbf{E}(X_i) = \mu_i$ και $\mathbf{V}(X_i) = \Sigma_{ii}$. Από τις συνιστώσες μέσες τιμές συμπεραίνουμε ότι $\mathbf{E}(X) = \mu$ και από τις διασπορές ότι τα διαγώνια στοιχεία του Σ συμπίπτουν με αυτές. Θέτοντας τώρα $u = e_i + e_j$, $i, j = 1, \ldots, d$, $i \ne j$ έχουμε

$$\mathbf{V}(X_i + X_j) = \mathbf{V}[(e_i + e_j)^\top X] = (e_i^\top + e_j^\top)\,\Sigma\,(e_i + e_j) = \Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij}.$$

Από τη σχέση

$$\mathbf{V}(X_i + X_j) = \mathbf{V}(X_i) + \mathbf{V}(X_j) + 2\,\mathrm{Cov}(X_i, X_j),$$

και τα παραπάνω καταλήγουμε στο ότι $\mathrm{Cov}(X_i, X_j) = \Sigma_{ij}$ και έτσι φτάνουμε στο ζητούμενο και για τον πίνακα διασποράς. □

Παρατήρηση 1.19. Γίνεται σαφές από την παραπάνω απόδειξη ότι αν για ένα οποιοδήποτε τ.δ. X ισχύει ότι για κάθε $u \in \mathbb{R}^d$, η μέση τιμή $\mathbf{E}(u^\top X) = u^\top \mu$ και η διασπορά $\mathbf{V}(u^\top X) = u^\top \Sigma\, u$, τότε $\mathbf{E}(X) = \mu$ και $\mathbf{V}(X) = \Sigma$. Σε αλγεβρική γλώσσα θα λέγαμε ότι η μέση τιμή μ καθορίζει μία γραμμική μορφή $f : \mathbb{R}^d \to \mathbb{R}$, όπου $f(u) = u^\top \mu$. Αντίστοιχα, ο πίνακας διασποράς Σ καθορίζει μία θετικά ημιορισμένη τετραγωνική μορφή $Q : \mathbb{R}^d \to \mathbb{R}$, όπου $Q(u) = u^\top \Sigma\, u$. Θα αναφερθούμε εκτενέστερα σε αυτήν σε επόμενο κεφάλαιο, όταν θα ασχοληθούμε με την εφαρμογή της στην Ανάλυση Παλινδρόμησης.

Η πιο απλή περίπτωση πολυδιάστατης κανονικής είναι βέβαια ένα τ.δ. Z που αποτελείται από ανεξάρτητες τ.μ. που ακολουθούν την τυπική κανονική.

Λήμμα 1.20. *Αν $Z \sim \mathcal{N}_d(0_d, I_d)$, όπου 0_d είναι το μηδενικό διάνυσμα και I_d είναι ο ταυτοτικός πίνακας d-τάξης, τότε οι συνιστώσες του Z είναι ανεξάρτητες και ισόνομες τ.μ. που ακολουθούν την τυπική κανονική. Λέμε ότι το Z ακολουθεί την τυπική κανονική διάστασης d.*

Απόδειξη:
Ας υποθέσουμε ότι $Z \sim \mathcal{N}_d(0_d, I_d)$. Τότε από τον ορισμό της πολυδιάστατης κανονικής έχουμε $u^\top Z \sim \mathcal{N}_1(0, u^\top u)$. Ας υποθέσουμε τώρα ότι παίρνουμε ένα τ.δ. $W = (W_1, \ldots, W_d)$ που αποτελείται από τ.μ. W_i ανεξάρτητες τυπικές κανονικές. Τότε $u^\top W \sim \mathcal{N}_1(0, u^\top u)$, από γνωστό αποτέλεσμα για γραμμικούς συνδιασμούς ανεξάρτητων τ.μ. που ακολουθούν κανονική κατανομή. Καταλήγουμε στο συμπέρασμα ότι $u^\top Z \overset{d}{=} u^\top W$ για κάθε $u \in \mathbb{R}^d$ και άρα $Z \overset{d}{=} W$. □

Η έννοια της συνάρτησης πυκνότητας πιθανότητας (σ.π.π.) για απόλυτα συνεχείς τ.μ. είναι ήδη γνωστή από τις Πιθανότητες Ι, όπως επίσης και η από κοινού σ.π.π. για τ.δ.. Αν γνωρίζουμε τη σ.π.π. για ένα απόλυτα συνεχές τ.δ. X, τότε μπορούμε να χαρακτηρίσουμε την κατανομή της X. Η εύρεση της σ.π.π. για μετασχηματισμούς της μορφής $AX + b$ (αφφινικούς μετασχηματισμούς) είναι εντελώς ανάλογη της περίπτωσης για πραγματικές τ.μ. και το μόνο που αλλάζει είναι η υπολογιστική πολυπλοκότητα. Υπενθυμίζουμε ότι στη μονοδιάστατη περίπτωση αν $Y = aX + b$ και $f_X(x)$ είναι η σ.π.π. της X, τότε για $a \neq 0$,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right).$$

Δίνουμε παρακάτω το πολυδιάστατο ανάλογο της παραπάνω σχέσης.

Πρόταση 1.21. *Αν $X \in \mathbb{R}^d$ είναι ένα τ.δ. που έχει σ.π.π. $f_X(x)$, A είναι ένας αντιστρέψιμος πίνακας διάστασης $d \times d$ και $b \in \mathbb{R}^d$ διάνυσμα, τότε το τ.δ. $Y = AX + b$ έχει σ.π.π. στον $\mathbb{R}^d$ που δίνεται από τη σχέση*

$$f_Y(y) = \frac{1}{|\det(A)|} f_X\left(A^{-1}(y - b)\right).$$

Στην επόμενη πρόταση δίνονται κάποιες σημαντικές ιδιότητες της πολυδιάστατης κανονικής κατανομής.

Πρόταση 1.22. *Έστω $X \sim \mathcal{N}_d(\mu, \Sigma)$, με μέση τιμή $\mu \in \mathbb{R}^d$ και πίνακα διασποράς Σ.*

(i) Αν $A \in \mathbb{R}^{s \times d}$ και $b \in \mathbb{R}^s$, τότε

$$AX + b \sim \mathcal{N}_s\left(A\mu + b,\, A\,\Sigma\, A^\top\right) \tag{1.9}$$

(ii) Αν ο Σ είναι θετικά ορισμένος (αντιστρέψιμος), τότε το τ.δ. X έχει συνάρτηση πυκνότητας πιθανότητας στον $\mathbb{R}^d$ της μορφής

$$f_X(x) = (2\pi)^{-d/2} \left(\det(\Sigma)\right)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\} \quad \forall x \in \mathbb{R}^d. \tag{1.10}$$

Απόδειξη:

(i) Έστω $u \in \mathbb{R}^s$. Τότε,

$$u^\top(AX+b) = (u^\top A)X + u^\top b = (A^\top u)^\top X + u^\top b \sim \mathcal{N}_1\left((A^\top u)^\top \mu,\ (A^\top u)^\top \Sigma A^\top u\right) + u^\top b,$$

όπου η τελευταία σχέση έπεται από την υπόθεση ότι $X \sim \mathcal{N}_d(\mu, \Sigma)$ και άρα $v^\top X \sim \mathcal{N}_1\left(v^\top \mu,\ v^\top \Sigma v\right)$ για $v = A^\top u \in \mathbb{R}^d$. Τελικά, η παραπάνω σχέση ξαναγράφεται στη μορφή

$$u^\top(AX+b) \sim \mathcal{N}_1\left(u^\top(A\mu+b),\ u^\top A\Sigma A^\top u\right),$$

από την οποία συμπεραίνουμε την (1.9).

(ii) Αν ο Σ είναι θετικά ορισμένος, τότε γράφεται στη μορφή $\Sigma = AA^\top$, με A αντιστρέψιμο. Ο πίνακας A παίζει το ρόλο μιας τετραγωνικής ρίζας του Σ. Συμπεραίνουμε λοιπόν από το ερώτημα (i) ότι το τ.δ.

$$Z = A^{-1}(X-\mu) \sim \mathcal{N}_d\left(0_d,\ A^{-1}\Sigma(A^{-1})^\top\right)$$

και επειδή $A^{-1}\Sigma(A^{-1})^\top = A^{-1}AA^\top(A^{-1})^\top = I_d$, καταλήγουμε στο ότι $Z \sim \mathcal{N}_d(0_d, I_d)$. Η σ.π.π. του Z είναι

$$f_Z(z) = \prod_{i=1}^d f_{Z_i}(z_i) = \prod_{i=1}^d (2\pi)^{-1/2} e^{-z_i^2/2} = (2\pi)^{-d/2} \exp\left\{-\frac{1}{2}z^\top z\right\} \quad \forall z \in \mathbb{R}^d.$$

Από την Πρόταση 1.21, συμπεραίνουμε ότι το τ.δ. $X = AZ + \mu$ έχει σ.π.π. στον $\mathbb{R}^d$ της μορφής

$$f_X(x) = (2\pi)^{-d/2} |\det(A)|^{-1} \exp\left\{-\frac{1}{2}\left(A^{-1}(x-\mu)\right)^\top \left(A^{-1}(x-\mu)\right)\right\}.$$

Η παραπάνω έκφραση συμπίπτει με την (1.10). □

1.6 Σύγκλιση ακολουθίας τυχαίων διανυσμάτων και οριακά θεωρήματα

Σε αυτήν την ενότητα υπενθυμίζουμε τους σημαντικότερους τρόπους σύγκλισης ακολουθίας τ.μ., δίνουμε τις πολυδιάστατες επεκτάσεις τους και καταλήγουμε σε νόμους μεγάλων αριθμών και ένα πολυδιάστατο κεντρικό οριακό θεώρημα.

Ορισμός 1.23. Μία ακολουθία τ.μ. (X_n) λέμε ότι συγκλίνει *με πιθανότητα 1* ή *σχεδόν βεβαίως*, ή *ισχυρά*, σε μία τ.μ. X, και γράφουμε $X_n \xrightarrow{a.s.} X$, αν

$$\mathbf{P}\left(\left\{\omega \in \Omega : \exists \lim X_n(\omega) \text{ και } \lim X_n(\omega) = X(\omega)\right\}\right) = 1.$$

Το παραπάνω ενδεχόμενο γράφεται και συνοπτικά ως $\{\lim X_n = X\}$ και έτσι σύντομα γράφουμε $\mathbf{P}(\lim X_n = X) = 1$ για να δηλώσουμε την παραπάνω σύγκλιση. Από τον ορισμό αυτό είναι φανερό ότι το μόνο που χρειάζεται να ορίσει κανείς είναι η σύγκλιση στο 0, αφού

$$X_n \xrightarrow{a.s.} X \quad \Longleftrightarrow \quad |X_n - X| \xrightarrow{a.s.} 0.$$

Ο τρόπος αυτός σύγκλισης επεκτείνεται εντελώς φυσιολογικά για ακολουθίες τ.δ. στον $\mathbb{R}^d$. Ο τυπικός ορισμός είναι όπως και στη μονοδιάστατη περίπτωση, απλά η έννοια του ορίου νοείται ως προς την ευκλείδια νόρμα.

Ορισμός 1.24. Μία ακολουθία τ.δ. (X_n) λέμε ότι συγκλίνει *με πιθανότητα 1* ή *σχεδόν βεβαίως*, ή *ισχυρά*, σε ένα τ.δ. X, και γράφουμε $X_n \xrightarrow{a.s.} X$, αν

$$\|X_n - X\| \xrightarrow{a.s.} 0 \quad \text{ή ισοδύναμα} \quad \mathbf{P}\left(\lim X_n = X\right) = 1.$$

Ασθενέστερη της παραπάνω σύγκλισης είναι η στοχαστική σύγκλιση.

Ορισμός 1.25. Μία ακολουθία τ.μ. (X_n) λέμε ότι συγκλίνει *κατά πιθανότητα* ή *στοχαστικά* σε μία τ.μ. X, και γράφουμε $X_n \xrightarrow{p} X$, αν για κάθε $\varepsilon > 0$

$$\mathbf{P}\left(|X_n - X| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0 \quad \text{ή ισοδύναμα} \quad \mathbf{P}\left(|X_n - X| \le \varepsilon\right) \xrightarrow[n\to\infty]{} 1.$$

Όπως για την ισχυρή σύγκλιση, έτσι και για τη στοχαστική, έχουμε

$$X_n \xrightarrow{p} X \quad \Longleftrightarrow \quad |X_n - X| \xrightarrow{p} 0.$$

Έτσι οδηγούμαστε στην ακόλουθη επέκταση.

Ορισμός 1.26. Μία ακολουθία τ.δ. (X_n) λέμε ότι συγκλίνει *κατά πιθανότητα* ή *στοχαστικά* σε ένα τ.δ. X, και γράφουμε $X_n \xrightarrow{p} X$, αν

$$\|X_n - X\| \xrightarrow{p} 0.$$

Παρόλο που η ισχυρή και η στοχαστική σύγκλιση εμπλέκουν τη συμπεριφορά της νόρμας στον $\mathbb{R}^d$, η μελέτη τους απλοποιείται και είναι ισοδύναμη με την εξέταση των αντίστοιχων ορίων κατά συνιστώσα. Συγκεκριμένα, είναι εύκολο να δειχθεί το επόμενο αποτέλεσμα.

Πρόταση 1.27. *Έστω $(X_n) = (X_{n,1}, X_{n,2}, \ldots, X_{n,d})$ ακολουθία τ.δ. και $X = (X_1, X_2, \ldots, X_d)$ ένα ακόμα τ.δ.. Τότε,*

$$X_n \xrightarrow{a.s./p} X \quad \Longleftrightarrow \quad X_{n,i} \xrightarrow{a.s./p} X_i, \quad \text{για κάθε } i = 1, 2, \ldots, d,$$

δηλ., και οι δύο συγκλίσεις είναι ισοδύναμες με τις αντίστοιχες συγκλίσεις κατά συνιστώσα.

Με βάση την παραπάνω πρόταση καταλήγουμε άμεσα σε πολυδιάστατες επεκτάσεις του Ισχυρού Νόμου των Μεγάλων Αριθμών (I.N.M.A.), αλλά και του Ασθενή Νόμου των Μεγάλων Αριθμών (A.N.M.A.).

Πόρισμα 1.28. *Έστω (X_n) μία ακολουθία ανεξάρτητων και ισόνομων τ.δ. με $\mathbf{E}\|X_1\| < +\infty$. Αν $\overline{X_n}$ είναι ο δειγματικός μέσος, τότε*

$$\overline{X_n} \xrightarrow{a.s.} \mathbf{E}(X_1) \qquad (I.N.M.A)$$

και

$$\overline{X_n} \xrightarrow{p} \mathbf{E}(X_1) \qquad (A.N.M.A)$$

Όπως βέβαια και στη μονοδιάστατη περίπτωση ο A.N.M.A. είναι συνέπεια του I.N.M.A., καθώς η στοχαστική σύγκλιση προκύπτει άμεσα από την ισχυρή (δείτε Πρόταση 1.32).

Πιο σύνθετη είναι η κατάσταση στη σύγκλιση κατά κατανομή. Η σημασία βέβαια της σύγκλισης αυτής είναι τεράστια και έχει πολλές εφαρμογές, ιδιαίτερα δε στη Στατιστική. Έτσι έχουν δοθεί αρκετοί ισοδύναμοι χαρακτηρισμοί, κάποιους από τους οποίους θα δούμε σε λίγο. Υπενθυμίζουμε πρώτα τον ορισμό που συνήθως δίνεται για ακολουθίες πραγματικών τ.μ. και είναι και αυτός που παρουσιάζεται ως απλούστερος και στις Πιθανότητες Ι.

Ορισμός 1.29. Μία ακολουθία τ.μ. (X_n) λέμε ότι συγκλίνει *κατά κατανομή* ή *ασθενώς* σε ένα τ.δ. X, και γράφουμε $X_n \xrightarrow{d} X$, αν για κάθε $x \in \mathbb{R}$ με $F(x-) = F(x)$ (σημεία συνεχείας της F),

$$F_n(x) \xrightarrow[n\to\infty]{} F(x),$$

όπου F_n είναι η συνάρτηση κατανομής της X_n και F είναι η σ.κ. της X.

Όπως ήδη βέβαια προαναφέραμε η συνάρτηση κατανομής είναι ένα βολικό αντικείμενο περιγραφής της κατανομής $\mathbf{P}_X$ μιας πραγματικής, αλλά και μίας διανυσματικής τ.μ.. Δίνουμε τώρα έναν γενικότερο ορισμό, που ισχύει και σε απειροδιάστατους χώρους, αλλάζοντας βέβαια κατάλληλα τους χώρους που οι τ.μ. παίρνουν τιμές.

Ορισμός 1.30. Μία ακολουθία τ.δ. (X_n) λέμε ότι συγκλίνει *κατά κατανομή* ή *ασθενώς* σε ένα τ.δ. X, και γράφουμε $X_n \xrightarrow{d} X$, αν για κάθε $B \in \mathcal{B}(R^d)$ με $\mathbf{P}(X \in \partial B) = 0$,

$$\mathbf{P}(X_n \in B) \xrightarrow[n\to\infty]{} \mathbf{P}(X \in B),$$

όπου $\partial B = \overline{B} \setminus B^\circ$ είναι το τοπολογικό σύνορο του B.

Στην ισχυρή και στη στοχαστική σύγκλιση, όλες οι τ.μ. πρέπει να είναι ορισμένες στον ίδιο χώρο πιθανότητας. Στην ασθενή σύγκλιση αυτό δεν είναι απαραίτητο, όπως φαίνεται εξάλλου και από τον ορισμό. Τονίζουμε επίσης το γεγονός ότι η ασθενής σύγκλιση μιας ακολουθίας τ.δ. κατά συνιστώσες, σε αντίθεση με τους άλλους δύο τρόπους σύγκλισης, δεν αρκεί για να συμπεράνουμε ολικά τη σύγκλιση. Υπάρχουν αρκετοί ισοδύναμοι χαρακτηρισμοί, βολικοί κατά περίπτωση και δίνουμε τώρα μερικούς από αυτούς σε αντιστοιχία με το Θεώρημα 1.3.

Θεώρημα 1.31. *Μία ακολουθία τ.δ. (X_n) συγκλίνει κατά κατανομή σε ένα τ.δ. $X \in \mathbb{R}^d$, αν και μόνο αν,*

(i) για κάθε σημείο συνεχείας $x \in \mathbb{R}^d$ της F

$$F_n(x) \xrightarrow[n\to\infty]{} F(x),$$

όπου F_n, F είναι οι σ.κ. των X_n, X.

(ii) κάθε γραμμικός συνδιασμός των συνιστωσών της X_n συγκλίνει κατά κατανομή στον αντίστοιχο γραμμικό συνδιασμό των συνιστωσών της X, δηλ., για κάθε $u \in \mathbb{R}^d$

$$u^t X_n \xrightarrow{d} u^t X$$

(iii) για κάθε $f : \mathbb{R}^d \to \mathbb{R}$ συνεχή και φραγμένη συνάρτηση

$$\mathbf{E}[f(X_n)] \xrightarrow[n\to\infty]{} \mathbf{E}[f(X)].$$

Στην παρακάτω πρόταση υπενθυμίζουμε τη σχέση που έχουν οι διάφοροι τρόποι σύγκλισης μεταξύ τους.

Πρόταση 1.32. *Έστω (X_n) μία ακολουθία τ.δ. και $X \in \mathbb{R}^d$ ένα ακόμα τ.δ. ορισμένα σε κοινό χώρο πιθανότητας. Τότε*

$$X_n \xrightarrow{a.s.} X \;\Rightarrow\; X_n \xrightarrow{p} X \;\Rightarrow\; X_n \xrightarrow{d} X$$

Στην ειδική περίπτωση που $X \overset{a.s.}{=} c$, όπου c σταθερά, τότε

$$X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c.$$

Είμαστε τώρα σε θέση να παρουσιάσουμε το πολυδιάστατο Κεντρικό Οριακό Θεώρημα σε δύο ισοδύναμες μορφές.

Πρόταση 1.33. *Έστω (X_n) μία ακολουθία ανεξάρτητων και ισόνομων τ.δ. με $\mathbf{E}\|X_1\|^2 < +\infty$. Θέτουμε $\mu_X = \mathbf{E}(X)$ και $\Sigma_X = \mathbf{V}(X)$. Αν $\left(\overline{X_n}\right)$ είναι η ακολουθία των δειγματικών μέσων και (S_n) η ακολουθία των μερικών αθροισμάτων της (X_n), τότε*

$$\sqrt{n}\left(\overline{X_n} - \mu_X\right) \xrightarrow{d} \mathcal{N}_d\left(0_d, \Sigma_X\right)$$

και

$$\frac{S_n - n\mu_X}{\sqrt{n}} \xrightarrow{d} \mathcal{N}_d\left(0_d, \Sigma_X\right)$$

Απόδειξη:
Θα αποδείξουμε την πρώτη από τις δύο ισοδύναμες μορφές. Η δεύτερη έπεται από την παρατήρηση ότι $S_n = n\overline{X_n}$. Έστω λοιπόν U ένα τ.δ. με $U \sim \mathcal{N}_d(0_d, \Sigma_X)$ και $u \in \mathbb{R}^d$. Από το Θεώρημα 1.31 αρκεί να δείξουμε ότι

$$u^\top \left\{ \sqrt{n}\left(\overline{X_n} - \mu_X\right)\right\} \xrightarrow{d} u^\top U \sim \mathcal{N}_1(0,\, u^\top \Sigma_X u) \tag{1.11}$$

Παρατηρούμε όμως ότι

$$u^\top \left\{ \sqrt{n}\left(\overline{X_n} - \mu_X\right)\right\} = \sqrt{n}\left(\overline{u^\top X_n} - u^\top \mu_X\right) = \sqrt{n}\left(\overline{Y_n} - \mu_Y\right),$$

όπου (Y_n) είναι ακολουθία ανεξάρτητων και ισόνομων τ.μ. με $Y_n = u^\top X_n$ και $\mu_Y = \mathbf{E}(Y_1) = u^\top \mu_X$. Από το κλασσικό Κεντρικό Οριακό Θεώρημα έχουμε

$$\sqrt{n}\left(\overline{Y_n} - \mu_Y\right) \xrightarrow{d} \mathcal{N}_1(0,\, \mathbf{V}(Y_1)) = \mathcal{N}_1(0,\, u^\top \Sigma_X u),$$

που είναι η οριακή κατανομή (1.11) που θέλαμε να φτάσουμε. □

Παρατήρηση 1.34. Αν ο πίνακας Σ_X είναι αντιστρέψιμος, τότε μπορούμε να γράψουμε τη σύγκλιση στην πολυμεταβλητή κανονική στη μορφή

$$\sqrt{n}\ \Sigma_X^{-1/2}\left(\overline{X_n} - \mu_X\right) \xrightarrow{d} \mathcal{N}_d(0_d,\, I_d),$$

όπου $\Sigma_X^{1/2}$ είναι μία οποιαδήποτε τετραγωνική ρίζα του Σ_X, δηλ., οποιαδήποτε επιλογή αντιστρέψιμου πίνακα A που ικανοποιεί τη σχέση $\Sigma_X = AA^\top$.

To be continued...

2

Απαραμετρική Στατιστική

2.1 Εισαγωγικά

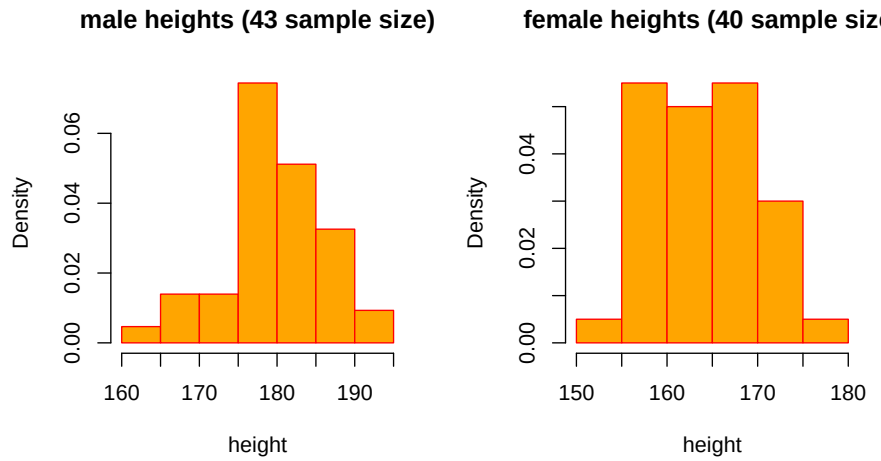
Στο εισαγωγικό μάθημα της Στατιστικής μία από τις βασικές υποθέσεις μοντελοποίησης είναι ότι οι παρατηρήσεις $X_1, X_2, \ldots, X_n$ αποτελούν ένα τυχαίο δείγμα (τ.δ.) από μία παραμετρική οικογένεια κατανομών. Οι παραπάνω τυχαίες μεταβλητές (τ.μ.), συνήθως προέρχονται από κλασικές οικογένειες κατανομών, είτε διακριτές, π.χ., Bernoulli $Be(p)$, διωνυμική $Bin(n, p)$, Poisson $\mathcal{P}(\lambda)$, είτε συνεχείς, π.χ., κανονική $\mathcal{N}(\mu, \sigma^2)$, εκθετική $Exp(\theta)$, Γάμμα $G(\alpha, \theta)$. Το κοινό τους χαρακτηριστικό είναι ότι ο ακριβής προσδιορισμός της κατανομής μέσα στην αντίστοιχη οικογένεια εξαρτάται από μία παράμετρο πεπερασμένης διάστασης. Σε μία πρώτη επαφή με τη Στατιστική συνήθως ασχολούμαστε με προβλήματα που εμπλέκουν την εκτίμηση μίας ή το πολύ δύο παραμέτρων, θεωρώντας ότι η αντίστοιχη παραμετρική οικογένεια περιγράφει καλά την αβεβαιότητα που συνδέεται με το υπό μελέτη στοχαστικό φαινόμενο. Το πολύ ουσιώδες θέμα της καταλληλότητας ενός μοντέλου ελάχιστα θίγεται, αν και ίσως αποτελεί το σημαντικότερο στοιχείο μιας καλής μοντελοποίησης. Διάφορα ερωτήματα ανακύπτουν φυσιολογικά.

- Πώς δικαιολογείται η χρήση μίας συγκεκριμένης παραμετρικής οικογένειας κατανομών;
- Αν έχουμε αμφιβολίες, τι εναλλακτικές υπάρχουν;
- η φύση των δεδομένων μπορεί να μην επιτρέπει τη χρήση παραμετρικών μοντέλων, π.χ., σειρά προτίμησης (διατάξιμα δεδομένα) ή κατηγορικά δεδομένα. Πώς αντιμετωπίζουμε αυτές τις περιπτώσεις;

Σε αυτό το κεφάλαιο θα κάνουμε μία εισαγωγή στη μη παραμετρική (ή απαραμετρική) Στατιστική. Ο στόχος είναι να ρίξουμε μία κλεφτή ματιά σε ένα πολύ σημαντικό κλάδο της Στατιστικής που προσπαθεί να απαντήσει σε τέτοια ερωτήματα και συμπληρώνει την εικόνα που σχηματίζεται σε ένα πρώτο εισαγωγικό μάθημα.

Στο παραπάνω σχήμα απεικονίζονται δύο ιστογράμματα δεδομένων που αφορούν τα ύψη 43 φοιτητών και 40 φοιτητριών του μαθήματος της Στατιστικής Ι αντίστοιχα. Θα ήταν ίσως λογικό να υποθέσουμε ότι τα δεδομένα αυτά αποτελούν ένα τυχαίο δείγμα του ύψους δύο διαφορετικών πληθυσμών (φοιτητών και φοιτητριών) κανονικά κατανεμημένων με τη δική τους μέση τιμή και διασπορά τα οποία θέλουμε να εκτιμήσουμε. Παρ'ολα αυτά θα θέλαμε να ελέγξουμε στατιστικά αυτόν τον ισχυρισμό στη βάση του τυχαίου δείγματος που διαθέτουμε και η μη παραμετρική Στατιστική μας δίνει κατάλληλα εργαλεία. Πριν φτάσουμε όμως σε αυτό, ας θυμηθούμε με αυτά που ήδη γνωρίζουμε από τη Στατιστική Ι κάποια ερωτήματα που μπορούμε να απαντήσουμε.

Υποθέτοντας λοιπόν ότι τα δεδομένα, και στις δύο περιπτώσεις, αποτελούν τυχαίο δείγμα από κανονική κατανομή με άγνωστη μέση τιμή και διασπορά μπορούμε (i) να εκτιμήσουμε μέσες τιμές και διασπορές, (ii) να κατασκευάσουμε $(1-\alpha)-$Δ.Ε. (διαστήματα εμπιστοσύνης) και (iii) να κάνουμε ελέγχους υποθέσεων αναζητώντας α-Κ.Π. (κρίσιμες περιοχές). Για παράδειγμα θα μπορούσαμε να ελέγξουμε αν η μέση τιμή του ύψους του μέσου Έλληνα και της μέσης Ελληνίδας που είναι ηλικιακά παρόμοια με τα άτομα στο δείγμα είναι συμβατή με τα δεδομένα που έχουμε για τους φοιτητές και



Σχήμα 2.1: Δεδομένα ύψους από φοιτητές της Στατιστικής Ι

τις φοιτήτριες αντίστοιχα ή αν έχουμε ενδείξεις για μία στατιστικά σημαντική διαφορά. Ας κάνουμε τώρα μία πρώτη προσέγγιση σε αυτό το θέμα.

Είναι γνωστό ότι σε ένα τυχαίο δείγμα από κανονική κατανομή $\mathcal{N}(\mu, \sigma^2)$ με μέση τιμή μ και διασπορά σ^2 άγνωστα, αμερόληπτες εκτιμήτριες των αντίστοιχων παραμέτρων είναι ο δειγματικός μέσος $\overline{X}$ και η αμερόληπτη δειγματική διασπορά $S^2 = (n-1)^{-1} \sum_{1 \le i \le n} (X_i - \overline{X})^2$ αντίστοιχα. Στον παρακάτω πίνακα συνοψίζουμε τις εκτιμήτριες, τα $(1-\alpha)$-Δ.Ε. για τις δύο παραμέτρους και τις α-Κ.Π. (περιοχές απόρριψης) για τους αμφίπλευρους ελέγχους $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ και $H_0 : \sigma^2 = \sigma_0$ vs $H_1 : \sigma^2 \neq \sigma_0^2$. Τα παραπάνω θεωρητικά αποτελέσματα μπορούν να εφαρμοστούν στα δεδομένα

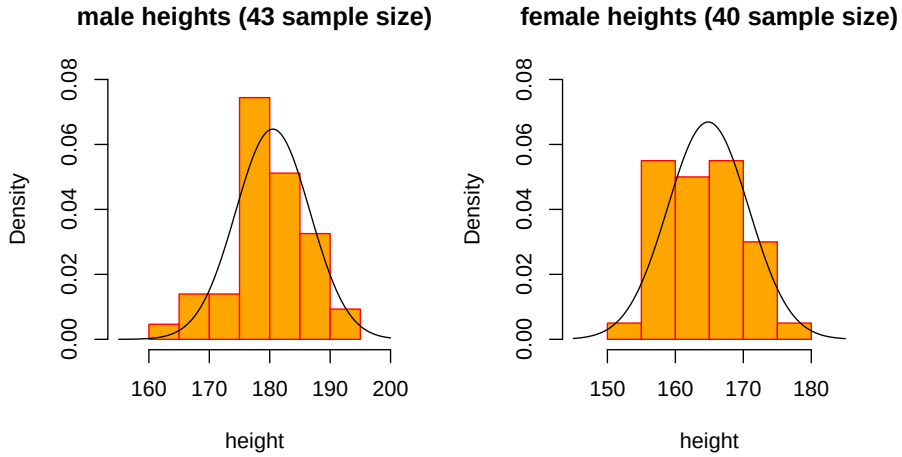
Πίνακας 2.1: Συνοπτικά στοιχεία συμπερασματολογίας για τ.δ. $\mathcal{N}(\mu, \sigma^2)$

παράμετρος	εκτιμήτρια	$(1-\alpha)-$Δ.Ε.	$\alpha-$Κ.Π.
μ	$\overline{X}$	$\overline{X} \pm t_{n-1,\alpha/2} \dfrac{S}{\sqrt{n}}$	$\dfrac{\left\lvert \overline{X} - \mu_0 \right\rvert}{S/\sqrt{n}} \ge t_{n-1,\alpha/2}$
σ^2	S^2	$\left(\dfrac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \dfrac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \right)$	$\dfrac{(n-1)S^2}{\sigma_0^2} \notin \left(\chi^2_{n-1,1-\alpha/2}, \chi^2_{n-1,\alpha/2} \right)$

ύψους των φοιτητών της Στατιστικής Ι, ξεχωριστά σε κάθε ένα από τα δύο δείγματα, των φοιτητών (μεγέθους $n_1 = 43$) και των φοιτητριών (μεγέθους $n_2 = 40$). Υποθέτουμε αρχικά ότι τα δείγματα αυτά προέρχονται από $\mathcal{N}(\mu_1, \sigma_1^2)$ και $\mathcal{N}(\mu_2, \sigma_2^2)$ αντίστοιχα με άγνωστες μέσες τιμές και διασπορές. Με τη βοήθεια του Πίνακα 2.1, αν $\overline{X}_i$ και S_i^2, $i = 1, 2$ είναι οι δειγματικοί μέσοι και οι αμερόληπτες δειγματικές διασπορές που αντιστοιχούν στα δύο δείγματα, τότε θέτοντας $\alpha = 0.05$ παίρνουμε τα αποτελέσματα που δίνονται στον Πίνακα 2.1. Σημειώνουμε ότι λόγω της δυϊκότητας διαστημάτων εμπιστοσύνης και ελέγχων υποθέσεων όταν έχουμε αμφίπλευρες εναλλακτικές, μπορούμε άμεσα να πραγματοποιήσουμε τους ελέγχους με μόνη γνώση των διαστημάτων εμπιστοσύνης. Με την παραπάνω στοιχειώδη στατιστική ανάλυση μπορούμε να έχουμε και μία εικόνα της προσαρμογής των μοντέλων αυτών στα δύο σύνολα δεδομένων. Στο παρακάτω γράφημα οι μαύρες καμπύλες αντιστοιχούν στα γραφήματα των συναρτήσεων πυκνότητας πιθανότητας των κανονικών κατανομών με μέσες τιμές και διασπορές τις εκτιμούμενες στα δύο δείγματα (δεύτερη στήλη του Πίνακα 2.1).

Πίνακας 2.2: Εφαρμογή στο ύψος φοιτητών/φοιτητριών με $\alpha = 0.05$

παράμετρος	εκτιμήτρια	95%−Δ.Ε.	0.05−Κ.Π.
μ_1	$\overline{X}_1 = 180.51$	(178.62 , 182.41)	$\mu_{1,0} \notin$ (178.62 , 182.41)
μ_2	$\overline{X}_2 = 164.83$	(162.92 , 166.73)	$\mu_{2,0} \notin$ (162.92 , 166.73)
σ_1^2	$S_1^2 = 37.97$	(25.81 , 61.34)	$\sigma_{1,0}^2 \notin$ (25.81 , 61.34)
σ_2^2	$S_2^2 = 35.53$	(23.84 , 58.58)	$\sigma_{2,0}^2 \notin$ (23.84 , 58.58)



Σχήμα 2.2: Προσαρμοσμένες καμπύλες των δεδομένων του Σχήματος 2.1

Σε μία πρόσφατη έρευνα (2016) εκτιμήθηκε το μέσο ύψος ανδρών και γυναικών στην Ελλάδα σε κάθε ηλιακή ομάδα. Αν για χάρη απλότητας δεχθούμε ότι τα δεδομένα αυτά της έρευνας είναι ακριβή (δηλαδή τα θεωρητικά) για τον αντίστοιχο πληθυσμό, τότε για ηλικίες 20-22 ετών αντιστοιχούν μέσα ύψη $\mu_1 = 177.4$ και $\mu_2 = 164.8$ για άντρες και γυναίκες αντίστοιχα. Αν θέλουμε λοιπόν να ελέγξουμε τη μη ύπαρξη στατιστικά σημαντικών διαφορών των μέσων τιμών $\overline{X}_1$ και $\overline{X}_2$ από τις αντίστοιχες θεωρητικές σε επίπεδο στατιστικής σημαντικότητας (ε.σ.σ.) α, τότε αρκεί να ελέγξουμε αν οι τελευταίες βρίσκονται μέσα στα $(1-\alpha)$-Δ.Ε.. Για $\alpha = 0.05$ παρατηρούμε στον πίνακα 2.1 ότι $\mu_{1,0} = 177.4 \notin (178.62 , 182.41)$ και $\mu_{2,0} = 164.8 \in (162.92 , 166.73)$. Το συμπέρασμα λοιπόν είναι ότι ενώ το μέσο ύψος των φοιτητριών φαίνεται να συμφωνεί με τα στοιχεία της πρόσφατης έρευνας, δεν φαίνεται να συμβαίνει το ίδιο με το μέσο ύψος των φοιτητών, αφού εντοπίζουμε μία στατιστικά σημαντική διαφορά. Πριν όμως βιαστούμε να καταλήξουμε σε κάποια οριστικά συμπεράσματα, θα πρέπει να τονίσουμε ότι το αποτέλεσμα αυτό θα πρέπει να το βλέπουμε ως ένα κίνητρο περαιτέρω διερεύνησης για τους πραγματικούς λόγους που οδήγησαν σε ένα τέτοιο συμπέρασμα. Για παράδειγμα, θα μπορούσε να υπάρχει μη αμελητέα απόκλιση μεταξύ πραγματικών μετρήσεων και του ύψους που δηλώθηκε ή ακόμα και κάποιες ακραίες παρατηρήσεις που επηρεάζουν το αποτέλεσμα. Εξάλλου οι μέσες τιμές που χρησιμοποιήθηκαν στη μηδενική υπόθεση είναι στην ουσία εκτιμούμενες και όχι πραγματικές του πληθυσμού. Σε κάθε περίπτωση, το μέγεθος του δείγματος και οι συνθήκες λήψης των δεδομένων μπορούν να επηρεάσουν σημαντικά το αποτέλεσμα. Θα πρέπει λοιπόν να είμαστε πάντα επιφυλακτικοί στα συμπεράσματα τα οποία φτάνουμε. Από την άλλη μεριά, η αδυναμία να φτάσουμε σε οριστικά συμπεράσματα ακόμα και αν κάνουμε την καλύτερη δυνατή μοντελοποίηση

δεν θα πρέπει να εκλαμβάνεται ως μειονέκτημα, αλλά ως μία φυσική συνέπεια του ρόλου της Στατιστικής ως συμβουλευτικής επιστήμης με στόχο κοινωνικά και επιστημονικά οφέλη. Η συνεισφορά της είναι μεγάλη και γίνεται ακόμα μεγαλύτερη στο να βάλει μία τάξη σε ένα εκρηκτικά αυξανόμενο πλήθος γνώσεων και πληροφοριών, με λιγότερο ή περισσότερο κρυμμένες κανονικότητες και εξαρτήσεις τις οποίες έχει αναλάβει το ιερό χρέος να αναδείξει μέσα σε ένα σύμπαν εγγενούς αβεβαιότητας. Τα στατιστικά συμπεράσματα πρέπει πάντα να ερμηνεύονται με πολύ μεγάλη προσοχή και να λειτουργούν μόνο συμβουλευτικά, συνειδητοποιώντας κάθε φορά τα όρια του στατιστικού σφάλματος.

Είναι φανερό ότι ένα παραμετρικό μοντέλο θέτει πολύ αυστηρούς περιορισμούς στη φύση της άγνωστης κατανομής. Διαισθητικά, αν το μοντέλο είναι προσεγγιστικά σωστό, τότε αυτό δεν είναι προβληματικό, αντίθετα ισχυροποιεί τα οποιαδήποτε στατιστικά συμπεράσματα. Σε αρκετές περιπτώσεις όμως, υπάρχουν πολλές αμφιβολίες για τη φύση της άγνωστης κατανομής και ιδιαίτερα όταν αντιμετωπίζουμε ένα πρόβλημα για πρώτη φορά. Για το λόγο αυτό, χαλαρώνοντας αρκετά τις υποθέσεις μας, θα προσεγγίσουμε τα δεδομένα με μεγαλύτερη ασφάλεια. Αυξάνοντας κατά πολύ τους βαθμούς ελευθερίας μας, μπορούμε μέσα από τη μη παραμετρική Στατιστική να εξηγήσουμε μία πολύ μεγαλύτερη γκάμα συμπεριφοράς των δεδομένων. Θα πρέπει βέβαια να γίνει κατανοητό από την αρχή ότι αυτό επιφέρει κάποιο κόστος (μικρότερο ή μεγαλύτερο) που συνδέεται γενικά με την απώλεια ισχύος των στατιστικών συμπερασμάτων που εξάγουμε.

Η πιο σημαντική διάκριση μεταξύ παραμετρικής και μη παραμετρικής Στατιστικής είναι στο πλήθος των παραμέτρων. Γενικά, στη μη παραμετρική Στατιστική το πλήθος των άγνωστων παραμέτρων είναι άπειρο, έτσι λοιπόν τοποθετούμαστε σε απειροδιάστατους χώρους. Η πιο ακραία περίπτωση, την οποία θα εξετάσουμε τώρα είναι όταν υποθέτουμε ότι η κατανομή του υπο μελέτη χαρακτηριστικού είναι εντελώς άγνωστη. Εφόσον η κατανομή μίας τ.μ. χαρακτηρίζεται πλήρως από τη συνάρτηση κατανομής της (που ορίζεται και για διακριτές, αλλά και για συνεχείς τ.μ.) μπορούμε λοιπόν να δουλεύουμε με συναρτήσεις κατανομής (σ.κ.). Θα υποθέσουμε λοιπόν τώρα ότι παράμετρος είναι η ίδια η σ.κ. F. Ας θυμηθούμε λοιπόν τον χαρακτηρισμό μίας συνάρτησης ως συνάρτησης κατανομής.

2.2 Συνάρτηση Κατανομής - Γενικευμένη αντίστροφη

Το πρώτο πράγμα που έρχεται στο νου όταν αναφερόμαστε σε συνάρτηση κατανομής είναι η συνάρτηση κατανομής τυχαίας μεταβλητής. Λόγω της μεγαλύτερης δυσκολίας μελέτης που παρουσιάζουν οι συνολοσυναρτήσεις, αναζητούνται πιο εύχρηστες συναρτήσεις που χαρακτηρίζουν την κατανομή μιας τ.μ.. Η σημασία της συνάρτησης κατανομής μιας τ.μ. έγκειται στο ότι χαρακτηρίζει πλήρως την κατανομή της με σχετικά απλό τρόπο. Αν X είναι η τ.μ., τότε τη συμβολίζουμε F_X και ορίζεται μέσω της σχέσης $F_X(x) := \mathbf{P}(X \leq x)$. Εύκολα αποδεικνύεται ότι η F_X ικανοποιεί τις ιδιότητες που περιγράφονται στον παρακάτω ορισμό. Είναι σκόπιμο να απομονώσουμε αυτές τις ιδιότητες και να ορίσουμε την έννοια της συνάρτησης κατανομής χωρίς να αντιστοιχεί εξ'αρχής σε κάποια τ.μ. X.

Ορισμός 2.1. (συνάρτηση κατανομής)
Μία συνάρτηση $F : \mathbb{R} \to [0, 1]$ λέγεται *συνάρτηση κατανομής*, αν (i) είναι αύξουσα, (ii) δεξιά συνεχής και (iii) ικανοποεί $F(-\infty) = 0$ και $F(+\infty) = 1$.

Λόγω της μονοτονίας της, μπορούμε να δείξουμε ότι η F έχει αριστερό όριο σε κάθε σημείο και μάλιστα έχει αριθμήσιμο πλήθος σημείων ασυνέχειας. Πρώτα όμως χρειαζόμαστε το εξής γενικότερο αποτέλεσμα.

Λήμμα 2.2. *Κάθε μονότονη συνάρτηση $f : \mathbb{R} \to \mathbb{R}$*

(i) έχει πλευρικά όρια σε κάθε σημείο της και μάλιστα αν η f είναι αύξουσα έχει δεξί όριο

$$f(x+) := \lim_{y \to x^+} f(y) = \inf\{f(y) : y > x\}, \tag{2.1}$$

και αριστερό όριο

$$f(x-) := \lim_{y \to x^-} f(y) = \sup\{f(y) : y < x\}. \tag{2.2}$$

Αν η f είναι φθίνουσα γίνεται εναλλαγή των infimum και supremum στις παραπάνω σχέσεις.

(ii) έχει αριθμήσιμο πλήθος σημείων ασυνέχειας.

Απόδειξη:

(i) Αρκεί να αποδειχθεί για f αύξουσα και μόνο για αριστερό όριο. Το δεξί όριο προκύπτει από την παρατήρηση ότι $\lim_{y \to x^+} f(y) = \lim_{y \to (-x)^-} f(-y)$ και τα αποτελέσματα για f φθίνουσα από το ότι η $-f$ είναι αύξουσα. Στην περίπτωση λοιπόν που η f είναι αύξουσα έχουμε ότι το σύνολο $\{f(y) : y < x\}$ είναι προφανώς μη κενό και άνω φραγμένο αφού $f(y) \leq f(x)$ για κάθε $y < x$. Άρα το $s := \sup\{f(y) : y < x\} \in \mathbb{R}$. Είναι άμεσο ότι είναι και το όριο της f από τα αριστερά στο x. Πράγματι, για κάθε $\varepsilon > 0$ έχουμε ότι υπάρχει y^* με $y^* < x$ ώστε $s - \varepsilon < f(y^*)$. Τότε αν πάρουμε οποιοδήποτε y με $y^* < y < x$, έχουμε $s - \varepsilon < f(y^*) \leq f(y)$ ή $s - f(y) < \varepsilon$. Άρα η f έχει αριστερό όριο στο x το s.

(ii) Αρκεί να το δείξουμε για f αύξουσα. Ένα σημείο x είναι σημείο ασυνέχειας της f αν και μόνο αν $f(x-) < f(x)$ ή $f(x) < f(x+)$ και σε κάθε τέτοιο σημείο μπορούμε να αντιστοιχίσουμε το διάστημα $I_x \in \{(f(x-), f(x)),\ (f(x), f(x+)),\ (f(x-), f(x+))\}$ ανάλογα με το αν η f είναι μόνο αριστερά ασυνεχής, μόνο δεξιά ασυνεχής ή ταυτόχρονα αριστερά και δεξιά ασυνεχής αντίστοιχα. Λόγω της μονοτονίας της f προκύπτει λοιπόν μία οικογένεια μη κενών ανοικτών διαστημάτων που είναι ξένα ανά δύο. Αν το πλήθος των σημείων ασυνέχειας ήταν υπεραριθμήσιμο, τότε το $\mathbb{R}$ θα περιείχε ένα υπεραριθμήσιμο πλήθος ξένων ανά δύο ανοικτών διαστημάτων, το οποίο είναι άτοπο (π.χ., με επιλογή ενός ρητού από κάθε διάστημα καταλήγουμε στο άτοπο συμπέρασμα ότι το $\mathbb{Q}$ είναι υπεραριθμήσιμο σύνολο).

□

Από το παραπάνω λήμμα προκύπτουν άμεσα οι ιδιότητες για την F.

Πόρισμα 2.3. *Κάθε συνάρτηση κατανομής F*

(i) έχει αριστερό όριο για κάθε $x \in \mathbb{R}$ και ισχύει ότι

$$F(x-) = \sup\{F(y) : y < x\}. \tag{2.3}$$

(ii) είναι συνεχής στα x για τα οποία $F(x-) = F(x)$ και ασυνεχής όταν $F(x-) < F(x)$ με άλμα ασυνέχειας $F(x) - F(x-)$.

(iii) έχει αριθμήσιμο πλήθος σημείων ασυνέχειας.

Όπως θα δούμε παρακάτω, αν μία συνάρτηση F είναι συνάρτηση κατανομής, τότε είναι και συνάρτηση κατανομής κάποιας τ.μ. X (όχι κατ' ανάγκη μοναδικής), δηλ., υπάρχει κατάλληλος χώρος πιθανότητας και τ.μ. X που ορίζεται σε αυτόν έτσι ώστε $F_X = F$. Πριν το αποδείξουμε (Πρόταση 2.7) θα χρειαστούμε την έννοια της γενικευμένης αντίστροφης.

Η έννοια της γενικευμένης αντίστροφης παίζει σημαντικό ρόλο στη θεωρία Πιθανοτήτων και τη Στατιστική.

Ορισμός 2.4. (γ.α.σ.κ.)
Έστω $F : \mathbb{R} \to [0, 1]$ μία συνάρτηση κατανομής. Η συνάρτηση $F^- : (0, 1) \to \mathbb{R}$, όπου

$$F^-(u) = \inf\{x : F(x) \geq u\} \tag{2.4}$$

λέγεται *γενικευμένη αντίστροφη* της F.

Παρατήρηση 2.5. Η συνάρτηση F^- είναι καλά ορισμένη, δηλ. $F^-(u) \in \mathbb{R}$ για κάθε $u \in (0,1)$. Πράγματι, έστω $u \in (0,1)$ και $I_u = \{x : F(x) \geq u\}$. Επειδή $F(+\infty) = 1$, υπάρχει $x_r \in \mathbb{R}$ με $F(x_r) \geq u$ και άρα $A_u \neq \emptyset$. Επιπλέον, αφού $F(-\infty) = 0$, υπάρχει $x_l \in \mathbb{R}$ με $x_l \notin I_u$, δηλ. $x_l \in I_u^c$ (διαφορετικά το $u > 0$ θα ήταν κάτω φράγμα του συνόλου τιμών της F). Άρα $F(x_l) < u \leq F(x)$, για κάθε $x \in I_u$. Επειδή η F είναι αύξουσα συμπεραίνουμε ότι $x_l < x$ για κάθε $x \in I_u$ και άρα το I_u είναι κάτω φραγμένο. Ως μη κενό και κάτω φραγμένο έχει μέγιστο κάτω φράγμα πραγματικό αριθμό που τον ονομάζουμε $F^-(u)$. Μάλιστα παρατηρούμε ότι για κάθε $x_r \in I_u$ και $x_l \in I_u^c$ θα ισχύει $F(x_l) < u \leq F(x_r)$ και άρα $x_l < x_r$. Συμπεραίνουμε ότι για κάθε $u \in (0,1)$ θα έχουμε $I_u = (F^-(u), +\infty)$ ή $[F^-(u), +\infty)$. Είναι τώρα φανερό από τη δεξιά συνέχεια της F ότι $F(F^-(u)) = \lim_{x \to (F^-(u))^+} F(x) \geq u$ αφού $F(x) \geq u$ για κάθε $x > F^-(u)$. Τελικά,

$$I_u := \{x : F(x) \geq u\} \; = \; [F^-(u), +\infty). \tag{2.5}$$

Στην παρακάτω πρόταση δίνουμε κάποιες στοιχειώδεις ιδιότητες της F^-.

Πρόταση 2.6. *Η γενικευμένη αντίστροφη $F^- : (0,1) \to \mathbb{R}$ ικανοποιεί τις ιδιότητες:*

(i) $F^-(u) = \min\{x : F(x) \geq u\}$, *για κάθε* $u \in (0,1)$,

(ii) $F(F^-(u)) \geq u$, *για κάθε* $u \in (0,1)$,

(iii) $F^-(u) \leq x \Leftrightarrow u \leq F(x)$, *για κάθε* $x \in \mathbb{R}$ *και* $u \in (0,1)$,

(iv) $F^-(F(x)) \leq x$, *για κάθε* $x \in \mathbb{R}$,

(v) είναι αύξουσα,

(vi) είναι αριστερά συνεχής[1]

Απόδειξη:
Τα (i), (ii), (iii) προκύπτουν άμεσα από την (2.5).
(iv) Προκύπτει από το ($\Leftarrow$) της (iii) αν θέσουμε $u = F(x)$.
(v) Έστω $v < u$. Αν $F(x) \geq u$, τότε $F(x) \geq v$ και άρα $\{x : F(x) \geq u\} \subset \{x : F(x) \geq v\}$. Παίρνοντας infimum και στα δύο μέλη συμπεραίνουμε ότι $F^-(v) \leq F^-(u)$.
(vi) Έστω $u \in (0,1)$. Από το (iv) και το Λήμμα 2.2 υπάρχει το αριστερό όριο της F^- στο u και $F^-(u-) \leq F^-(u)$. Έστω τώρα $v < u$. Προφανώς, $F^-(u-) \geq F^-(v)$, αφού η F^- είναι αύξουσα από το (iv). Όμως και η F είναι αύξουσα, άρα $F(F^-(u-)) \geq F(F^-(v)) \geq v$, όπου η τελευταία ανισότητα έπεται από το (ii). Συμπεραίνουμε ότι $F(F^-(u-)) \geq v$ για κάθε $v < u$, και άρα $F(F^-(u-)) \geq u$ ή ισοδύναμα $F^-(u) \leq F^-(u-)$, λόγω του (iii). Τελικά λοιπόν ισχύει η ισότητα. □

Στην επόμενη πρόταση εξασφαλίζεται ότι για οποιαδήποτε συνάρτηση κατανομής F υπάρχει τυχαία μεταβλητή X με $F_X = F$.

Πρόταση 2.7. *Αν $U \sim Unif(0,1)$, τότε η τ.μ. $X = F^-(U) \sim F$.*

Απόδειξη: Κατ'αρχήν χ.β.γ. θεωρούμε ότι $U \in (0,1)$ και άρα η $X = F^-(U)$ είναι καλά ορισμένη. Αν $x \in \mathbb{R}$, τότε

$$F_X(x) = \mathbf{P}(X \leq x) = \mathbf{P}(F^-(U) \leq x) = \mathbf{P}(U \leq F(x)) = F(x), \tag{2.6}$$

όπου η τρίτη ισότητα έπεται από την Πρόταση 2.6-(iii) και η τελευταία από τη σχέση $F_U(u) = u$ που ισχύει για τη σ.κ. της ομοιόμορφης για κάθε $u \in [0,1]$. □

Αν η X είναι συνεχής τ.μ., δηλ. έχει συνεχή συνάρτηση κατανομής F, τότε ισχύει το εξής ενδιαφέρον αποτέλεσμα.

[1] (iii) σε ποιά περίπτωση ισχύει ότι $F^-(0+) = -\infty$ και $F^-(1-) = +\infty$;

Πρόταση 2.8. *Αν η τ.μ. $X \sim F$ και η X είναι συνεχής τ.μ., τότε*
(i) για κάθε $u \in (0,1)$, ισχύει $F(F^{-}(u)) = u$,
(ii) η τ.μ. $U = F(X) \sim Unif(0,1)$.

Απόδειξη:
(i) Έστω $u \in (0,1)$ και $x_u := F^{-}(u)$. Αν $x < x_u$ τότε $F(x) < u$. Συμπεραίνουμε ότι

$$F(x_u-) = \lim_{x \to (x_u)^-} F(x) \leq u.$$

Από υπόθεση η X είναι συνεχής τ.μ. και άρα η F είναι συνεχής συνάρτηση και $F(x_u) = F(x_u-)$. Αντικαθιστώντας στην παραπάνω σχέση έχουμε $F(x_u) \leq u$. Από την Πρόταση 2.6-(ii) ισχύει και η αντίστροφη ανισότητα, άρα τελικά $F(x_u) = u$.
(ii) Έστω $u \in (0,1)$. Τότε,

$$\mathbf{P}(U \leq u) = \mathbf{P}(F(X) \leq u) = \mathbf{P}(F(X) < u) + \mathbf{P}(F(X) = u). \tag{2.7}$$

Όταν η X είναι συνεχής τ.μ., τότε η F είναι συνεχής συνάρτηση και επειδή $F(-\infty) = 0$ και $F(+\infty) = 1$ συμπεραίνουμε ότι η εξίσωση $F(x) = u$ έχει πάντα λύση. Επιπλέον η F είναι αύξουσα και άρα $F^{-1}(\{u\}) = [a,b]$ για κάποια $a, b \in \mathbb{R}$ με $a \leq b$ (μάλιστα $a = F^{-}(u)$). Από τα παραπάνω συμπεραίνουμε ότι

$$\mathbf{P}(F(X) = u) = \mathbf{P}\left(X \in F^{-1}(\{u\})\right) = \mathbf{P}(X \in [a,b]) = F(b) - F(a-) = F(b) - F(a) = 0. \tag{2.8}$$

Από την Πρόταση 2.6-(iii) έχουμε για $u \in (0,1)$ ότι $F(x) < u \Leftrightarrow x < F^{-}(u)$ και άρα

$$\mathbf{P}(F(X) < u) = \mathbf{P}(X < F^{-}(u)) = \mathbf{P}(X \leq F^{-}(u)) = F(F^{-}(u)) = u, \tag{2.9}$$

όπου χρησιμοποιήσαμε ότι η X είναι συνεχής τ.μ., αλλά και το (i). Αντικαθιστώντας τις (2.8) και (2.9) στην (2.7) φτάνουμε στο ζητούμενο. □

Παρατήρηση 2.9. Αν η $X \sim F$ δεν είναι συνεχής τ.μ., δηλ. η F είναι ασυνεχής σε κάποιο $x \in \mathbb{R}$, τότε δεν ισχύει το παραπάνω αποτέλεσμα. Πράγματι, σε αυτήν την περίπτωση $F(x) - F(x-) > 0$. Συμπεραίνουμε ότι $F(X) \notin (F(x-), F(x))$, άρα $\mathbf{P}\left(F(X) \in (F(x-), F(x))\right) = 0$. Έτσι η $F(X)$ δεν ακολουθεί την ομοιόμορφη κατανομή.

2.3 Εμπειρική Συνάρτηση Κατανομής

Θα συμβολίζουμε το χώρο όλων των σ.κ. με $\mathcal{F}$ και θα παίζει το ρόλο του παραμετρικού χώρου, όπως έπαιζε το Θ στην παραμετρική Στατιστική. Έτσι η έκφραση $\theta \in \Theta$ μετατρέπεται εδώ σε $F \in \mathcal{F}$. Υπάρχει βέβαια μία μικρή δυσκολία. Ενώ στην παραμετρική Στατιστική $\Theta \subset \mathbb{R}^d$ για κατάλληλο πεπερασμένο d και έτσι φυσιολογικά μετράμε αποστάσεις με την ευκλείδια μετρική ή με άλλη ισοδύναμή της, η κλάση $\mathcal{F}$ είναι απειροδιάστατη και το πρόβλημα επιλογής κατάλληλου μετρικού χώρου είναι πιο λεπτό. Θα επανέρθουμε σε αυτό το πρόβλημα αργότερα. Ας χτίσουμε πρώτα διαισθητικά μία πρώτη μη παραμετρική προσέγγιση.

Έστω λοιπόν ότι διαθέτουμε n παρατηρήσεις $x_1, x_2, \ldots, x_n$ που είναι n ανεξάρτητες πραγματοποιήσεις μίας τ.μ. $X \sim F$. Η μόνη πληροφορία που έχουμε είναι ότι μπορεί να πάρει αυτές τις τιμές. Αν κάποιος τώρα μας ζητήσει μία τιμή αυτής της τ.μ. δεν θα είχαμε κάποιο λόγο να ευνοήσουμε κάποια από αυτές. Θα μπορούσαμε να τις βάλουμε σε μία κληρωτίδα και να τραβήξουμε στην τύχη μία από αυτές. Με άλλα λόγια θα επιλέγαμε κάποια απο αυτές ισοπίθανα. Αυτή η απλή σκέψη μας οδηγεί στην κατασκευή μίας καινούριας τ.μ. X_n^* που έχει τη διακριτή ομοιόμορφη κατανομή στο σύνολο $\{x_1, x_2, \ldots x_n\}$. Θα ήταν βέβαια ιδιαίτερα επιθυμητό η συνάρτηση κατανομής F_n^* της X_n^* να προσεγγίζει

(με κάποια έννοια) την άγνωστη συνάρτηση κατανομής F για μεγάλο n. Εφόσον βέβαια η X_n^* είναι διακριτή τ.μ. έχουμε ότι $\forall x \in \mathbb{R}$,

$$F_n^*(x) = \mathbf{P}(X_n^* \le x) = \sum_{i:x_i \le x} \mathbf{P}(X_n^* = x_i) = \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{x_i \le x} = \frac{\sum_{i=1}^{n} \mathbf{1}_{x_i \le x}}{n}. \tag{2.10}$$

Όπως κάναμε και στη Στατιστική Ι για την κατασκευή μίας εκτιμήτριας, είναι φανερό ότι η παραπάνω διαδικασία μας οδηγεί σε μία εκτιμήτρια $\widehat{F}_n(x)$ του $F(x)$. Συγκεκριμένα, αντικαθιστώντας με τ.μ. έχουμε

$$\widehat{F}_n(x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{X_i \le x}}{n} \qquad \forall x \in \mathbb{R}. \tag{2.11}$$

Η $\widehat{F}_n$ λέγεται *εμπειρική συνάρτηση κατανομής*. Κάθε πραγματοποίηση F_n^* της $\widehat{F}_n$ αντιστοιχεί σε μία συνάρτηση κατανομής όπως προαναφέραμε και άρα παίρνει τιμές στο χώρο συναρτήσεων $\mathcal{F}$. Μπορούμε λοιπόν να σκεφτόμαστε την $\widehat{F}_n$ ως μία τυχαία συνάρτηση του x ή ισοδύναμα ως μία τυχαία μεταβλητή με τιμές συναρτήσεις στον χώρο $\mathcal{F}$. και αξίζει βέβαια να μελετηθούν οι οριακές ιδιότητες της $\widehat{F}_n$. Μας ενδιαφέρει λοιπόν να δούμε αν και με ποιό τρόπο συγκλίνει η $\widehat{F}_n$ στην F.

Λήμμα 2.10. *Αν για μία θετική τ.μ. X ισχύει ότι $X \overset{a.s.}{\le} \varepsilon$ για κάθε $\varepsilon > 0$, τότε $X \overset{a.s.}{=} 0$.*

Απόδειξη:
Αρκεί να δείξουμε ότι $\mathbf{P}(X = 0) = 1$ ή ισοδύναμα ότι $F(0) = 1$, αφού $\mathbf{P}(X < 0) = 0$ από υπόθεση (X θετική τ.μ.) και έτσι $F(0) = \mathbf{P}(X = 0)$. Από την υπόθεση έχουμε επίσης ότι για κάθε $\varepsilon > 0$, $F(\varepsilon) = \mathbf{P}(X \le \varepsilon) = 1$. Όμως, η σ.κ. F είναι δεξιά συνεχής (πάντα) και άρα $F(0) = F(0+) = \lim_{\varepsilon \to 0^+} F(\varepsilon) = 1$. □

Θεώρημα 2.11. *(Glivenko-Cantelli)*
Η εμπειρική συνάρτηση κατανομής $\widehat{F}_n$ συγκλίνει με πιθανότητα 1 ομοιόμορφα στην F, δηλ.,

$$\left\| \widehat{F}_n - F \right\|_\infty \xrightarrow{a.s.} 0. \tag{2.12}$$

Απόδειξη:
Θα δείξουμε ότι $\limsup_n \left\| \widehat{F}_n - F \right\|_\infty \overset{a.s.}{\le} \varepsilon$ για κάθε $\varepsilon > 0$ και λόγω του Λήμματος 2.10 και της προφανούς σχέσης $0 \le \liminf_n \left\| \widehat{F}_n - F \right\|_\infty \le \limsup_n \left\| \widehat{F}_n - F \right\|_\infty$ οδηγούμαστε στο ότι $\lim_n \left\| \widehat{F}_n - F \right\|_\infty \overset{a.s.}{=} 0$ και άρα το ζητούμενο. Έστω λοιπόν $\varepsilon > 0$. Επειδή η F είναι αύξουσα και παίνει τιμές στο $[0, 1]$ είναι φανερό ότι θα έχει το πολύ ένα πεπερασμένο πλήθος σημείων ασυνέχειας με άλμα ασυνέχειας $\ge \varepsilon$. Συμπεραίνουμε ότι υπάρχει ένα πεπερασμένο k και σημεία $x_0 = -\infty, x_1, \ldots, x_k, x_{k+1} = +\infty$ (εδώ περιλαμβάνονται τα παραπάνω σημεία ασυνέχειας), έτσι ώστε $F(x_{i+1}^-) - F(x_i) < \varepsilon$ για κάθε $0 \le i \le k$. Παρατηρούμε τώρα ότι αν $x \in [x_i, x_{i+1})$,

$$\begin{aligned} \widehat{F}_n(x) - F(x) &\le \widehat{F}_n(x_{i+1}^-) - F(x_i) \le \widehat{F}_n(x_{i+1}^-) - F(x_{i+1}^-) + \varepsilon \\ F(x) - \widehat{F}_n(x) &\le F(x_{i+1}^-) - \widehat{F}_n(x_i) \le F(x_i) - \widehat{F}_n(x_i) + \varepsilon \end{aligned}$$

Επειδή τα φράγματα στις παραπάνω σχέσεις είναι ανεξάρτητα του x, για $x \in [x_i, x_{i+1})$, έχουμε

$$\sup_{x \in [x_i, x_{i+1})} |\widehat{F}_n(x) - F(x)| \le \max\{\widehat{F}_n(x_{i+1}^-) - F(x_{i+1}^-), F(x_i) - \widehat{F}_n(x_i)\} + \varepsilon$$

Άρα

$$\left\| \widehat{F}_n - F \right\|_\infty = \max_{0 \le i \le k} \sup_{x \in [x_i, x_{i+1})} |\widehat{F}_n(x) - F(x)| \le \max_{0 \le i \le k} \max\{\widehat{F}_n(x_{i+1}^-) - F(x_{i+1}^-), F(x_i) - \widehat{F}_n(x_i)\} + \varepsilon \tag{2.13}$$

Από τον Ισχυρό Νόμο των Μεγάλων Αριθμών έχουμε

$$\frac{\sum_{i=1}^{n} \mathbf{1}_{X_i \in A}}{n} \xrightarrow[n\to\infty]{a.s.} \mathbf{E}\, \mathbf{1}_{X\in A} = \mathbf{P}(X \in A). \tag{2.14}$$

Επιλέγοντας $A = (-\infty, x_i)$ και $(-\infty, x_i]$ έχουμε ότι $\widehat{F}_n(x_i^-) \overset{a.s}{\to} F(x_i^-)$ και $\widehat{F}_n(x_i) \overset{a.s}{\to} F(x_i)$. Απο τα παραπάνω συμπεραίνουμε ότι όλες οι ακολουθίες τ.μ. που υπεισέρχονται στο max της σχέσης (2.13) συγκλίνουν με πιθ. 1 στο 0 και άρα $\limsup_n \left\|\widehat{F}_n - F\right\|_\infty \overset{a.s.}{\leq} \varepsilon$ □

Ασκήσεις

2.1 Έστω X συνεχής τ.μ.. Να δείξετε ότι για κάθε $u \in (0,1)$

$$F^-(u) = \min\{x : F(x) = u\}$$

2.2 Έστω X συνεχής τ.μ. και $f : \mathbb{R} \to [0,1]$ συνεχής συνάρτηση.

(i) Να δείξετε ότι η $f(X)$ δεν είναι απαραίτητα συνεχής τ.μ..

(ii) Αλλάζει το παραπάνω συμπέρασμα αν υποθέσουμε επιπλέον ότι $(0,1) \subset f(\mathbb{R})$;