Κεφάλαιο 1

Εισαγωγή

Έννοιες των πιθανοτήτων και της στατιστικής καθώς και μέθοδοι που στηρίζονται σε αυτές είναι χρήσιμα και απαραίτητα εργαλεία για να καταλάβουμε τον κόσμο γύρω μας, να μελετήσουμε φυσικά μεγέθη, φαινόμενα και διαδικασίες. Αν για ένα μέγεθος ή σύστημα που μελετάμε δεν υπάρχει καθόλου αβεβαιότητα ή τυχαιότητα δε χρειάζεται η πιθανοκρατική και στατιστική προσέγγιση αφού **καθοριστικά μοντέλα** (deterministic models) μπορούν να περιγράψουν το φαινόμενο επακριβώς. Για παράδειγμα τέτοιο σύστημα είναι αυτό των κινήσεων των πλανητών του ηλιακού συστήματος. Μπορούμε με μεγάλη ακρίβεια και χρησιμοποιώντας καθοριστικά μοντέλα να προσδιορίσουμε τη θέση των πλανητών που συμφωνεί με την πραγματική παρατήρηση.

Όμως τα δεδομένα που συλλέγουμε από τα περισσότερα φυσικά φαινόμενα και πραγματικές διαδικασίες δε μπορούν να εξηγηθούν ικανοποιητικά με μαθηματικά καθοριστικά μοντέλα. Αυτό συμβαίνει γιατί υπάρχει ο παράγοντας της αβεβαιότητας ή τυχαιότητας και για αυτό χρειάζεται να περιγράψουμε το σύστημα **πιθανοκρατικά** (probabilistically) και να καταφύγουμε σε **στατιστικές μεθόδους και μοντέλα** (statistical methods and models) για να λύσουμε προβλήματα εκτίμησης και πρόβλεψης σε τέτοια συστήματα.

Η **θεωρία πιθανοτήτων** μας επιτρέπει να μελετήσουμε τη μεταβλητότητα του αποτελέσματος ενός πειράματος (ή γενικά μιας πραγματοποίησης ενός φαινομένου ή μιας διαδικασίας) για το οποίο το ακριβές αποτέλεσμα δεν είναι δυνατόν να προβλεφθεί με ακρίβεια. Από την άλλη μεριά, η **στατιστική** συνίσταται στη συλλογή δεδομένων που λέγεται *δειγματοληψία* (sampling), στην περιγραφή τους, που αναφέρεται ως *περιγραφική στατιστική* (descriptive statistics) και κυρίως στην ανάλυση των δεδομένων που οδηγεί και στην απόκτηση συμπερασμάτων και για αυτό αναφέρεται ως *στατιστική συμπερασματολογία* (statistical inference). Ο συνδυασμός των εννοιών και ιδιοτήτων από τη θεωρία πιθανοτήτων με τις τεχνικές της στατιστικής είναι το αντικείμενο της **ανάλυσης δεδομένων** (data analysis). Σκοπός της ανάλυσης δεδομέ-

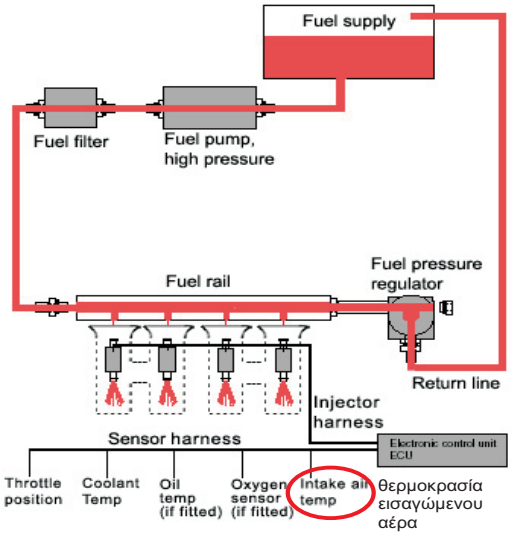
νων είναι η σύνοψη σε λίγες παραμέτρους της πληροφορίας από ένα σύνολο δεδομένων, το οποίο μπορεί να αποτελείται από πολλές παρατηρήσεις και να αφορά περισσότερα από ένα μεγέθη. Αυτή είναι η προσέγγιση που ακολουθείται στο πλαίσιο του μαθήματος με τίτλο 'Ανάλυση Δεδομένων' στο 7ο εξάμηνο του Νέου Προγράμματος Σπουδών του ΤΗΜΜΥ, ΑΠΘ.

Γενικότερα η ανάλυση δεδομένων, όπως ορίζεται στη Wikipedia, είναι η διαδικασία της επιθεώρησης, καθαρισμού, μετασχηματισμού και μοντελοποίησης των δεδομένων που έχει ως σκοπό να ανακαλύψει χρήσιμη πληροφορία, να δώσει συμπεράσματα και να υποστηρίξει τη λήψη αποφάσεων. Τα τελευταία έτη η ανάλυση δεδομένων έχει γνωρίσει ιδιαίτερο ενδιαφέρον και μέσα από τη χρήση νέων όρων όπως 'αναλύσεις δεδομένων' (data analytics) και 'αναλύσεις δεδομένων μεγάλης κλίμακας' (big data analytics), 'εξόρυξη δεδομένων' (data mining) και 'επιστήμη δεδομένων' (data science). Οι όροι αυτοί σχετίζονται άμεσα ή λιγότερα άμεσα με την ανάλυση δεδομένων. Δεν είναι ξεκάθαρη η διαφορά του όρου 'αναλύσεις δεδομένων' από την ανάλυση δεδομένων και περισσότερο χρησιμοποιείται στο χώρο της επιχειρηματικότητας και αγοράς και έχει περισσότερο επικοινωνιακό χαρακτήρα, δηλαδή να παρουσιάζει, οπτικοποιεί και επικοινωνεί τα αποτελέσματα της ανάλυσης. Ο όρος 'αναλύσεις δεδομένων μεγάλης κλίμακας' αναφέρεται σε δεδομένα μεγάλης κλίμακας, όπου πέρα από την ανάγκη της ανάλυσης δεδομένων (με την έννοια της χρήσης εργαλείων κυρίως της στατιστικής για την ανάλυση των δεδομένων), επεκτείνεται και σε άλλα θέματα που προκύπτουν λόγω του μεγάλου όγκου και πολυ-τροπικότητας των δεδομένων, όπως η απόκτηση και αποθήκευση τους, η αναζήτηση και μεταφορά στοιχείων από το σύνολο των δεδομένων, η ενημέρωση και οπτικοποίηση τους, καθώς και θέματα ιδιωτικότητας και προσβασιμότητας. Ο όρος 'εξόρυξη δεδομένων' έχει επικάλυψη με την ανάλυση δεδομένων, αλλά επικεντρώνεται περισσότερο στην ανακάλυψη προτύπων στα δεδομένα και χρησιμοποιεί μεθόδους της στατιστικής και πληροφορικής (μηχανική μάθηση). Τέλος ο όρος 'επιστήμη δεδομένων' είναι γενικότερος όρος, συμπεριλαμβάνει την ανάλυση δεδομένων, αλλά επεκτείνεται στη διαχείριση δεδομένων που μπορεί να είναι από διαφορετικές πηγές και διαφορετικών τύπων. Στο μάθημα αυτό, θα θεωρήσουμε ότι τα δεδομένα προς ανάλυση είναι καλά ορισμένα και αναφέρονται σε συγκεκριμένες μεταβλητές ενδιαφέροντος, χωρίς να μας απασχολεί αν είναι μικρού ή μεγάλου όγκου.

Ας δούμε κάποια παραδείγματα που αναδεικνύουν τα προβλήματα που αφορούν την ανάλυση δεδομένων, τις έννοιες και τα θέματα που θα μας απασχολήσουν στη συνέχεια.

Παράδειγμα 1.1. Μια μηχανή με αυτόματο ψεκασμό καυσίμου έχει τη μονάδα ελέγχου της μηχανής που συμπεριλαμβάνει αισθητήρες και μετρητές,

όπως αισθητήρα για τη θέση της βαλβίδας εισαγωγής ατμοποιημένου καυσίμου, τη θερμοκρασία αέρα, το οξυγόνο και το χρόνο ανάφλεξης (δες Σχήμα 1.1). Μπορούμε λοιπόν να συλλέξουμε μετρήσεις για όλα αυτά τα μεγέθη



Σχήμα 1.1: Διάγραμμα συστήματος ψεκασμού καυσίμου (αντιγραφή από τη διεύθυνση `http://www.twminduction.com`).

και ας σταθούμε για παράδειγμα στη θερμοκρασία του αέρα που εισάγεται στη μηχανή. Η θερμοκρασία του αέρα είναι *τυχαίο μέγεθος* που μπορεί να αλλάζει σε διάφορες χρονικές στιγμές ή σε διαφορετικές καταστάσεις λειτουργίας της μηχανής. Για τις μετρήσεις της θερμοκρασίας αέρα μπορεί να μας ενδιαφέρουν δύο βασικά χαρακτηριστικά. Το πρώτο είναι η *ακρίβεια επανάληψης* (precision), δηλαδή αν μεταβάλλονται πολύ ή λίγο οι μετρήσεις θερμοκρασίας αέρα (για τις ίδιες συνθήκες λειτουργίας). Το δεύτερο είναι η *ακρίβεια (ορθότητα)* (accuracy), δηλαδή κατά πόσο οι μετρήσεις θερμοκρασίας αέρα είναι κοντά στην επιθυμητή τιμή ή υπάρχουν συστηματικές αποκλίσεις. Η περιγραφή της τυχαίας μεταβολής της θερμοκρασίας αέρα είναι ένα θέμα της θεωρίας των πιθανοτήτων, που αναφέρεται ως *κατανομή μιας τυχαίας μεταβλητής*. Όταν έχουμε ένα πλήθος μετρήσεων της θερμοκρασίας αέρα μπορούμε να εκτιμήσουμε συγκεκριμένα χαρακτηριστικά της, όπως η μέση τιμή και η διασπορά της. Επίσης μπορούμε να συγκρίνουμε τα χαρακτηριστικά της θερμοκρασίας αέρα σε διαφορετικές συνθήκες λειτουργίας ή σε διαφορετικές μηχανές (κάτω από τις ίδιες συνθήκες λειτουργίας), ή ακόμα να διορθώσουμε τη μηχανή για να πετύχουμε καλύτερη ορθότητα στην τιμή της θερμοκρασίας αέρα.

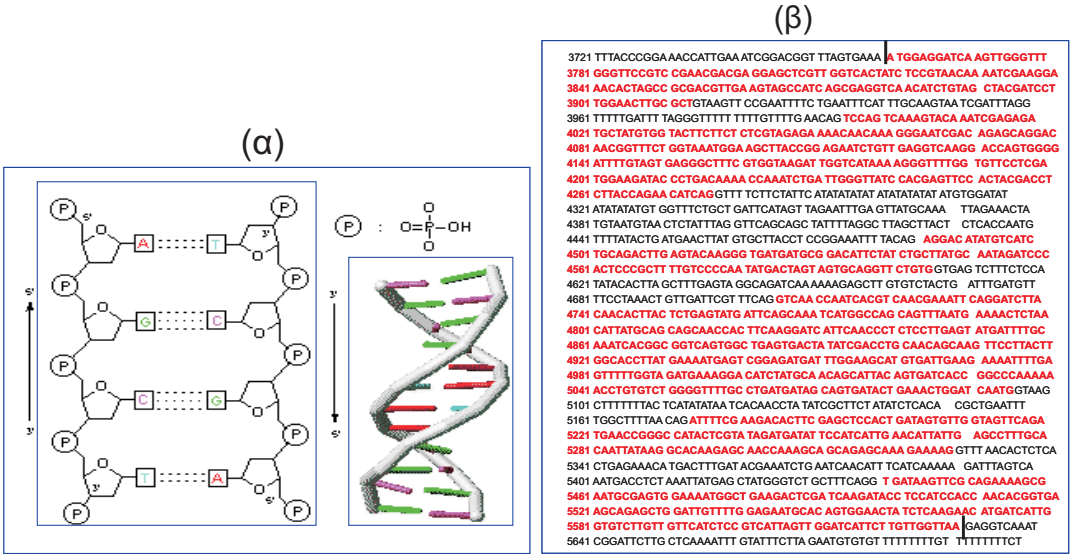
Σε συνέχεια του παραδείγματος, θα θέλαμε επίσης να προσδιορίσουμε την ποσότητα καυσίμου που χρησιμοποιείται γνωρίζοντας άλλα μεγέθη σχετικά

με τη λειτουργία της μηχανής την ίδια χρονική στιγμή ή διάρκεια, όπως είναι το άνοιγμα της βαλβίδας (ή ο ρυθμός ανοίγματος), ο χρόνος ανάφλεξης και η πίεση ψεκασμού. Ένα τέτοιο μοντέλο λέγεται *μοντέλο παλινδρόμησης* (regression model).

Ένα άλλο πρόβλημα, ίσως πιο δύσκολο να το διερευνήσουμε είναι η περιγραφή της χρονικής εξέλιξης της ποσότητας καυσίμου που καταναλώνεται κατά τη λειτουργία της μηχανής. Μια προσέγγιση είναι να θεωρήσουμε την εξέλιξη αυτή ως μια *καθοριστική διαδικασία* (deterministic process), δηλαδή να αποκλείσουμε την παρουσία τυχαίων διακυμάνσεων και παρεμβολών, που συνήθως αναφέρονται ως *θόρυβος* (noise), ή πιο ρεαλιστικά, να αγνοήσουμε την επίδραση τους. Αλλά αν θα θέλαμε να μελετήσουμε τη χρονική μεταβολή της θερμοκρασίας του αέρα στο σύστημα ψεκασμού της μηχανής θα ήταν πιο κατάλληλο να θεωρήσουμε τη διαδικασία ως *στοχαστική* (stochastic process), δηλαδή να συμπεριλάβουμε στην περιγραφή και το θόρυβο. Και οι δύο προσεγγίσεις αφορούν την *ανάλυση χρονοσειράς* (time series analysis) ενός μεγέθους (την ποσότητα καυσίμου ή τη θερμοκρασία αέρα), αποτελούν το αντικείμενο άλλου μαθήματος και για αυτό δε θα επεκταθούμε σε ανάλυση δεδομένων από χρονοσειρές.

Για τη θερμοκρασία του αέρα σε μια πολύπλοκη μηχανή θα μπορούσαν να υπάρχει πληθώρα άλλων παρατηρούμενων μεγεθών που μπορεί να επηρεάζει τη θερμοκρασία αέρα. Σε μια τέτοια περίπτωση θα θέλαμε να επιλέξουμε τα πιο σχετικά μεγέθη, είτε με απευθείας επιλογή από το σύνολο των μεγεθών ή μέσω κάποιου μετασχηματισμού. Γενικά η μείωση διάστασης σε δεδομένα από μεγάλο πλήθος μεγεθών είναι ένα θέμα που θα μας απασχολήσει στο τέλος του μαθήματος.

Παράδειγμα 1.2. Ένα άλλο παράδειγμα είναι η περιγραφή της δομής των στοιχείων στη σειρά του DNA (δες Σχήμα 1.2α). Είναι γνωστό πως η σειρά του DNA αποτελείται από τέσσερα αμινοξέα, την αδενίνη (A), κυτοσίνη (C), γουανίνη (G) και θυμίνη (T). Η μεταβλητή ενδιαφέροντος είναι το στοιχείο της σειράς DNA που δεν παίρνει αριθμητικές τιμές αλλά ένα από τα τέσσερα αυτά σύμβολα (δες Σχήμα 1.2β). Τμήματα της σειράς ορίζουν τα γονίδια (genes) που αποτελούνται από τις λεγόμενες κωδικοποιημένες περιοχές (με κόκκινο είναι οι κωδικοποιημένες περιοχές του γονιδίου που δίνεται στο Σχήμα 1.2β μεταξύ των κάθετων γραμμών) ενώ άλλα δεν έχουν κάποια γνωστή κωδικοποίηση (το μεγαλύτερο τμήμα των ανθρωπίνων χρωμοσωμάτων). Η περιγραφή της θέσης, της συχνότητας εμφάνισης και γενικά της δομής των τεσσάρων βάσεων όπως και συνδυασμών αυτών στη σειρά του DNA μπορεί να γίνει με τη βοήθεια πιθανοκρατικών (στατιστικών) μεθόδων για την ανάλυση διακριτών (κατηγορηματικών) δεδομένων ή συμβολοσειρών. Και εδώ υπάρχουν δύο προσεγγίσεις, η πρώτη θεωρώντας τη συμβολοσειρά ως πραγματοποίηση κά-



Σχήμα 1.2: (α) Η δομή του DNA. (β) Ένα τμήμα της σειράς DNA που αποτελεί ένα γονίδιο.

ποιας στοχαστικής αλυσίδας (διακριτής διαδικασίας) ή κάποιου δυναμικού συστήματος ορισμένο σε σύμβολα.

Παραθέτονται στη συνέχεια κάποιοι βασικοί ορισμοί που χρησιμοποιούνται στη θεωρία πιθανοτήτων και στη στατιστική ανάλυση:

τυχαία μεταβλητή (τ.μ.) (random variable): οποιοδήποτε χαρακτηριστικό του οποίου η τιμή αλλάζει στα διάφορα στοιχεία του πληθυσμού. Η τ.μ. μπορεί να είναι:

- **συνεχής (continuous):** να παίρνει τιμές σ' ένα διάστημα, όπως είναι η θερμοκρασία αέρα σε μια μηχανή ψεκασμού,
- **διακριτή (discrete):** να παίρνει μια τιμή σε ένα αριθμήσιμο σύνολο διακριτών τιμών, όπως ένα στοιχείο της σειράς DNA.

δεδομένα (data): ένα σύνολο τιμών μιας τ.μ. που έχουμε στη διάθεση μας, π.χ. μετρήσεις της θερμοκρασίας αέρα σε διάφορες χρονικές στιγμές ή σε διάφορες μηχανές ψεκασμού για τις ίδιες συνθήκες λειτουργίας, ή ένα κομμάτι της σειράς DNA.

πληθυσμός (population): μια ομάδα ή μια κατηγορία στην οποία αναφέρεται η τ.μ., το χρωμόσωμα No 22 ή ένας τύπος αυτόματης μηχανής με σύστημα ψεκασμού.

δείγμα (sample): ένα υποσύνολο του πληθυσμού που μελετάμε, π.χ. ένα κομμάτι της σειράς DNA του χρωμοσώματος Νο 22 ή 20 αυτόματες μηχανές με σύστημα ψεκασμού ίδιου τύπου.

παράμετρος (parameter): ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στον πληθυσμό, π.χ. το ποσοστό εμφάνισης του στοιχείου Α στο χρωμόσωμα Νο 22, ή η μέση θερμοκρασία αέρα κάτω από κάποιες συνθήκες για έναν τύπο μηχανής με σύστημα ψεκασμού.

στατιστικό (statistic): ένα μέγεθος που συνοψίζει με κάποιο τρόπο τις τιμές της τ.μ. στο δείγμα, π.χ. το ποσοστό εμφάνισης του στοιχείου Α σ' ένα κομμάτι 1000 στοιχείων της σειράς DNA από το χρωμόσωμα Νο 22, ή ο μέσος όρος της θερμοκρασίας αέρα που υπολογίσαμε σε 20 αυτόματες μηχανές με σύστημα ψεκασμού κάτω από τις ίδιες συνθήκες.

Η μελέτη μιας τ.μ. με τη βοήθεια της πιθανοθεωρίας προϋποθέτει ότι γνωρίζουμε (ή υποθέτουμε) την κατανομή της τ.μ., και άρα και τον πληθυσμό στον οποίο παίρνει τιμές, καθώς και τις παραμέτρους της. Στην πράξη βέβαια κάτι τέτοιο δε συμβαίνει, αλλά σε κάποια προβλήματα μπορούμε να υποθέτουμε γνωστές κατανομές. Σε κάθε περίπτωση, ένα από τα κύρια προβλήματα στην ανάλυση δεδομένων είναι να εκτιμήσουμε τις άγνωστες (αλλά σταθερές) παραμέτρους του πληθυσμού από τα γνωστά αλλά μεταβλητά στατιστικά του δείγματος δεδομένων που έχουμε στη διάθεση μας.

Στο γενικό πλαίσιο της στοχαστικής προσέγγισης που χρησιμοποιείται στην ανάλυση δεδομένων, υπάρχουν δύο κύριες κατευθύνσεις. Η πρώτη αναφέρεται ως ανάλυση κατά Bayes ή Μπεϋζιανή ανάλυση ή προσέγγιση (Bayesian approach) και υποθέτει από πριν κάποια κατανομή ή πιθανότητες σχετικά με το πρόβλημα που μελετάμε. Η δεύτερη κατεύθυνση δε χρησιμοποιεί κάποια υπόθεση κατανομής αλλά υπολογίζει τις πιθανότητες από τη συχνότητα εμφάνισης που υπολογίζεται απευθείας από τα δεδομένα και αναφέρεται ως ανάλυση με βάση τις συχνότητες (frequentist approach). Στα θέματα που θα μελετήσουμε δε θα εμβαθύνουμε στη Μπεϋζιανή προσέγγιση.