

# Στατιστική Μεγάλων Διαστάσεων και Μεγάλα Δεδομένα (High Dimensional Statistics and Big Data)

**ΔΙΔΑΣΚΩΝ: Π.ΠΑΠΑΣΤΑΜΟΥΛΗΣ**

## Γενικά Στοιχεία Μαθήματος

Κωδικός: 61227

Τύπος: Επιλογής

Επίπεδο: Μεταπτυχιακό

Έτος σπουδών: Β'

Εξάμηνο σπουδών: 4<sup>ο</sup>

ECTS: 5

Γλώσσα διδασκαλίας: Ελληνική

## Περιεχόμενο Μαθήματος

Σύγχρονες εφαρμογές της Στατιστικής απαιτούν τη διαχείριση και ανάλυση μεγάλου όγκου δεδομένων. Παρ' αυτά, η αύξηση της διάστασης των δεδομένων δεν συνεπάγεται πάντα αύξηση της χρήσιμης πληροφορίας, αλλά και αύξηση του θορύβου που εμπεριέχεται σε αυτά. Έτσι, τίθενται νέες προκλήσεις στη Συμπερασματολογία. Το μάθημα της Στατιστικής Μεγάλων Διαστάσεων θα παρουσιάσει σύγχρονες Στατιστικές τεχνικές για την αντιμετώπιση τέτοιων προβλημάτων, όπως:

- Προβλήματα πολλαπλών συγκρίσεων και έλεγχος του ρυθμού των ψευδώς θετικών αποτελεσμάτων
- Στατιστικές εφαρμογές σε προβλήματα βιοπληροφορικής
- Εφαρμογές (γενικευμένων) γραμμικών μοντέλων σε προβλήματα μεγάλων διαστάσεων
- Τεχνικές κανονικοποίησης (LASSO, Ridge regression)
- Διαχωριστική ανάλυση και ανάλυση συστάδων σε πολυμεταβλητά δεδομένα με τεχνικές μείωσης διάστασης
- Στατιστική ανάλυση δικτύων

## Προαπαιτούμενα

- Καλή γνώση
  - Εκτιμητικής και Ελέγχων Υποθέσεων
  - (Γενικευμένων) Γραμμικών Μοντέλων
- Βασικές γνώσεις
  - Υπολογιστικής Στατιστικής
  - Προγραμματισμού σε R
- Σχετικά προσφερόμενα μαθήματα (δεν είναι προαπαιτούμενα)
  - Εφαρμοσμένη Στατιστική κατά Bayes
  - Στατιστική Μάθηση

## Επιδιωκόμενα Μαθησιακά Αποτελέσματα

Μετά την επιτυχή ολοκλήρωση του μαθήματος, οι φοιτητές θα είναι σε θέση

- να γνωρίζουν τις προκλήσεις που θέτουν τα δεδομένα μεγάλων διαστάσεων στη Στατιστική Συμπερασματολογία
- να αντιμετωπίζουν σωστά προβλήματα ελέγχου πολλαπλών υποθέσεων
- να έρθουν σε επαφή με σύνολα δεδομένων από σύγχρονες ερευνητικές περιοχές όπως η Βιοπληροφορική και η ανάλυση δικτύων
- να μάθουν νέες Στατιστικές μεθοδολογίες που αναπτύχθηκαν τα τελευταία 20 χρόνια ειδικά σχεδιασμένες για την αντιμετώπιση προβλημάτων μεγάλων διαστάσεων
- να υλοποιούν κατάλληλους αλγόριθμους στην R και Bioconductor.

## **Συνιστώμενη Βιβλιογραφία**

- [1] Christophe Giraud (2015). Introduction to High-Dimensional Statistics. Philadelphia: Chapman and Hall/CRC
- [2] Tony Cai, Xiaotong Shen, eds. 2011. High-dimensional data analysis. Frontiers of Statistics. Singapore: World Scientific
- [3] Hastie, R. Tibshirani and R. Friedman (2009). Elements of Statistical Learning. Springer
- [4] Efron, Bradley. Large scale inference: Empirical Bayes Methods for Estimation, Testing and Prediction. Cambridge University Press, 2010
- [5] McLachlan, G. and Peel, D (2000). Finite Mixture Models. Cambridge University Press, 2010
- [6] Wasserman, S. and G. Robins (2005). An introduction to random graphs, dependence graphs, and  $p^*$  . Models and methods in social network analysis.
- [7] Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. Journal of the American Statistical Association.

## **Διδακτικές και Μαθησιακές Μέθοδοι**

Μια διάλεξη τριών ωρών εβδομαδιαίως, ασκήσεις μελέτης και προγραμματισμού στο σπίτι (ορισμένες προς παράδοση).

## **Μέθοδοι Αξιολόγησης και Βαθμολόγησης**

Ο τελικός βαθμός θα προκύψει ως σταθμισμένος μέσος όρος 2 ή 3 εργασιών και μιας τελικής παρουσίασης/προφορικής εξέτασης.