ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ

ΔΙΕΡΕΥΝΗΣΗ ΕΥΑΙΣΘΗΣΙΑΣ ΚΑΙ ΕΙΔΙΚΟΤΗΤΑΣ ΣΥΝΘΕΤΩΝ ΚΛΙΜΑΚΩΝ ΑΠΟΤΕΛΟΥΜΕΝΩΝ ΑΠΟ ΔΙΑΚΡΙΤΕΣ Η ΣΥΝΕΧΕΙΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ, ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΕ ΜΟΝΤΕΛΑ ΠΡΟΣΟΜΟΙΩΣΕΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Φραγκίσκος Γ. Μπερσίμης, *MSc*

Μαθηματικός - Στατιστικός

Αθήνα 2019

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ ΨΗΦΙΑΚΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ

Η Διδακτορική Διατριβή εξετάστηκε από την κάτωθι Επταμελή Επιτροπή:

Βαμβακάρη Μαλβίνα, Αναπληρώτρια Καθηγήτρια του Χαροκοπείου Πανεπιστημίου (Επιβλέπουσα) στο γνωστικό αντικείμενο «Στατιστική-Θεωρία Πιθανοτήτων και Εφαρμογές»,

Κυριακούσης Ανδρέας, Ομότιμος Καθηγητής του Χαροκοπείου Πανεπιστημίου

Παναγιωτάκος Δημοσθένης, Καθηγητής Χαροκοπείου Πανεπιστημίου στο γνωστικό αντικείμενο «Βιοστατιστική – Επιδημιολογία της Διατροφής»

Νικολαίδη Μάρα, Καθηγήτρια του Χαροκοπείου Πανεπιστημίου στο γνωστικό αντικείμενο «Κατανεμημένα Συστήματα, Μοντελοποίηση Συστημάτων και Ψηφιακές Βιβλιοθήκες»

Αναγνωστόπουλος Δημοσθένης, Καθηγητής του Χαροκοπείου Πανεπιστημίου στο γνωστικό αντικείμενο «Πληροφοριακά Συστήματα-Προσομοίωση»

Βαρλάμης Ηρακλής, Επίκουρος Καθηγητής του Χαροκοπείου Πανεπιστημίου στο γνωστικό αντικείμενο «Διαχείριση Δεδομένων»

Τζαβελάς Γεώργιος Επίκουρος Καθηγητής του Πανεπιστημίου Πειραιώς, στο γνωστικό αντικείμενο «Πιθανότητες και Στατιστική»

Η έγκριση της διδακτορικής διατριβής από το τμήμα Πληροφορικής & Τηλεματικής του Χαροκοπείου Πανεπιστημίου δεν υποδηλώνει και αποδοχή των απόψεων του συγγραφέα.

Για την εκπόνηση της παρούσας Διατριβής διατέθηκε χρηματοδότηση από το Τμήμα Πληροφορικής & Τηλεματικής του Χαροκοπείου Πανεπιστημίου για τη συμμετοχή σε 2 διεθνή συνέδρια (Τουρκία, Καππαδοκία, EMR IBS 2015 & Μάλτα, Βαλέττα, SMTDA 2016).

Στα πλαίσια της παρούσας διδακτορικής διατριβής έγινε χρήση δεδομένων από τη μελέτη ATTICA. Η μελέτη της ATTICA υποστηρίχθηκε από ερευνητικές επιχορηγήσεις της Ελληνικής Καρδιολογικής Εταιρείας (HCS2002) και της Ελληνικής Εταιρείας Αθηροσκλήρωσης (HAS2003).

Ο Φραγκίσκος Γ. Μπερσίμης δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΛΛΗΝΙΚΕΣ ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ 10

ΞΕΝΟΓΛΩΣΣΕΣ ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ 10

ΕΥΧΑΡΙΣΤΙΕΣ 11

ΠΕΡΙΛΗΨΗ 12

ABSTRACT 18

1. ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ - ΟΡΙΣΜΟΙ - ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ 28

1.1 Πληροφορική 35
1.1.1 Βιοπληροφορική 35
1.2 Στατιστική 35
1.2.1 Βιοστατιστική 39
1.3 Κλίμακες (Δείκτες) 42
1.3.1 Χρησιμότητα Κλιμάκων – Παραδείγματα Κλιμάκων 44
1.3.1.1 Κλίμακες στην Εκπαίδευση 46
1.3.1.2 Κλίμακες - Δείκτες στην Οικονομία 49
1.3.1.3 Κλίμακες στην Ιατρική 52
1.3.1.4 Κλίμακες στην Ψυχομετρία 56
1.3.1.5 Κλίμακες στη Διαιτολογία 58
1.3.2 Διακριτές & Συνεχείς Σύνθετες Κλίμακες 61
1.3.3 Μεθοδολογία Κατασκευής Σύνθετων Κλιμάκων 64
1.3.3.1 Αριθμός και Είδος Συνιστωσών Μεταβλητών 66
1.3.3.2 Εύρος Στηρίγματος Συνιστωσών Μεταβλητών των Κλιμάκων 66
1.3.3.3 Βαθμονόμηση Συνιστωσών Μεταβλητών 68
1.3.3.4 Στάθμιση Συνιστωσών Μεταβλητών 68
1.3.4 Τιμή Κλίμακας (Score) - Προβληματισμοί Κατασκευής Κλιμάκων 64
1.3.5 Αξιολόγηση Διαχωριστικής Ικανότητας Κλιμάκων 70
1.3.5.1 Ευαισθησία Κλιμάκων 72
1.3.5.2 Ειδικότητα Κλιμάκων 73
1.3.5.3 Ακρίβεια Κλιμάκων 73
1.3.5.4 Θετική Προγνωστική Αξία Κλιμάκων 74
1.3.5.5 Αρνητική Προγνωστική Αξία Κλιμάκων 74
1.3.5.6 Επιφάνεια κάτω από την Καμπύλη Λειτουργικών Χαρακτηριστικών 75
1.3.5.7 Μέθοδοι Επιλογής Βέλτιστου Διαχωριστικού Σημείου 77
1.4 Προσομοίωση – Αλγόριθμοι Monte Carlo & Bootstrap 779
1.5 Μέθοδοι Ταξινόμησης - Υπολογισμός Σταθμίσεων Κλιμάκων 85
1.5.1 Διαχωριστική Ανάλυση 86
1.5.2 Λογιστική Παλινδρόμηση 88
1.5.3 Μέθοδοι Στάθμισης Κλιμάκων 890
1.6 Διαγνωστική Ακρίβεια Σταθμισμένων Κλιμάκων 93
1.7 Διαγνωστική Ακρίβεια Κλιμάκων με Διαφορετικό Αριθμό Συνιστωσών 94

1.8 Εφαρμογή Σταθμισμένων Κλιμάκων σε Επιδημιολογικά Δεδομένα 94

2. ΣΚΟΠΟΣ ΔΙΑΤΡΙΒΗΣ.....96

3. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ 98

3.1 Διακριτές Κλίμακες & Αριθμός Διαμερίσεων.....98
3.1.1 Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Περίπτωση Άγνωστης κατανομής Διακριτής Κλίμακας 98
3.1.2 Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Περίπτωση Ομοιόμορφης κατανομής Κλίμακας 102
3.1.3 Σύνθετη Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Περίπτωση Άγνωστης κατανομής Κλίμακας.....105
3.1.4 Σύνθετη Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Ομοιόμορφη Κατανομή Συνιστωσών Μεταβλητών.....110
3.1.4.1 Ομοιόμορφη Κατανομή – Κοινό Διαχωριστικό Όριο 113
3.2 Συνεχείς Κλίμακες & Εύρος Στηρίγματος.....116
3.2.1 Συνεχής Κλίμακα & Εύρος Στηρίγματος - Περίπτωση Άγνωστης Κατανομής Κλίμακας 117
3.2.2 Συνεχής Κλίμακα & Εύρος Στηρίγματος - Περίπτωση Ομοιόμορφης Κατανομής Κλίμακας 121
3.2.3 Συνεχής Κλίμακα & Εύρος Στηρίγματος - Περίπτωση Κανονικής κατανομής Κλίμακας 125
3.2.4 Σύνθετη Συνεχής Κλίμακα & Εύρος Στηρίγματος Συνιστωσών - Περίπτωση άγνωστης κατανομής 130
3.2.5 Σύνθετη Συνεχής Κλίμακα & Εύρος Στηρίγματος Συνιστωσών - Περίπτωση Ομοιόμορφης κατανομής.....135
3.2.6 Σύνθετη Συνεχής Κλίμακα & Εύρος Στηρίγματος Συνιστωσών Μεταβλητών - Περίπτωση Κανονικής κατανομής 141
3.3. Αποτελέσματα Προσομοιώσεων 144
3.3.1 Ακρίβεια Σύνθετων Κλιμάκων ως Προς τον Αριθμό *m* των Συνιστωσών Μεταβλητών.....144
3.3.2 Διαγνωστική Ικανότητα Σταθμισμένων Σύνθετων Κλιμάκων 147
3.3.2.1 Κανονικά Κατανεμημένες Συνιστώσες Μεταβλητές.....147
3.3.2.2 Ομοιόμορφα Κατανεμημένες Συνιστώσες Μεταβλητές.....150
3.3.2.3 Αξιολόγηση Πρόβλεψης 152
3.3.2.4 Εφαρμογή σε δεδομένα από τη μελέτη ATTICA 154

4. ΣΥΖΗΤΗΣΗ.....157

4.1 Συμπεράσματα Διατριβής 157
4.1.1 Επίδραση Αύξησης Αριθμού Διαιρέσεων Διακριτών Συνιστωσών Μεταβλητών στην Ευαισθησία Κλιμάκων 157
4.1.2 Επίδραση Αύξησης Εύρους Στηρίγματος Συνιστωσών Μεταβλητών στην Ευαισθησία Κλιμάκων.....159
4.1.3 Επίδραση Σταθμίσεων στις Συνιστώσες Μεταβλητές στην Ακρίβεια Κλιμάκων 160
4.1.4 Επίδραση Αύξησης Αριθμού Συνιστωσών Μεταβλητών στη Διαχωριστική και Προβλεπτική Ικανότητα Κλιμάκων.....162
4.2 Περιορισμοί - Συνθήκες.....162

4.3 Γενίκευση των Συμπερασμάτων 164
4.4 Χρησιμότητα στη Δημόσια Υγεία 164
4.5 Μελλοντική Έρευνα..... 165

ΠΑΡΑΡΤΗΜΑ Α: Αλγόριθμοι σε περιβάλλον MatLab, R, SPSS, Mathematica..... 180

ΠΑΡΑΡΤΗΜΑ Β: Βιογραφικό Σημείωμα 213

ΠΑΡΑΡΤΗΜΑ Γ: Δημοσιεύσεις & Ανακοινώσεις που προέκυψαν από τη Διδακτορική Διατριβή 220

ΕΛΛΗΝΙΚΕΣ ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

ΑΘ: Αληθώς Θετικοί (TP: True Positive)
ΑΑ: Αληθώς Αρνητικοί (TN: True Negative)
ΨΘ: Ψευδώς Θετικοί (False Positive)
ΨΑ: Ψευδώς Αρνητικοί (False Negative)

ΞΕΝΟΓΛΩΣΣΕΣ ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

TR: True Ratio
TPR: True Positive Ratio
TNR: True Negative Ratio
PPV: Positive Predictive Value
NPV: Negative Predictive Value
ROC: Receiver Operating Characteristic Curve (Καμπύλη Λειτουργικών Χαρακτηριστικών)
AUC: Area Under Curve (Εμβαδόν Επιφάνειας κάτω από την Καμπύλη Λειτουργικών Χαρακτηριστικών)
CRM: Customer Relationship Management (Πληροφοριακό Σύστημα Διαχείρησης Πελατών)
CBT: Computer Based Training (Εκπαίδευση Βασιζόμενη σε Η/Υ)
BDI: Beck Depression Inventory
SDS: Self-Rating Depression Scale
HRSD: Hamilton Rating Scale for Depression

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω την επιβλέπουσα της διατριβής μου, Μαλβίνα Βαμβακάρη, αναπληρώτρια καθηγήτρια του Τμήματος Πληροφορικής & Τηλεματικής του Χαροκοπείου Πανεπιστημίου, για την ανάθεση του θέματος αλλά και για την πολύτιμη καθοδήγησή της καθ' όλη τη διάρκεια πραγματοποίησης της διατριβής.

Ιδιαιτέρως θα ήθελα να ευχαριστήσω τον Καθηγητή του Τμήματος Διαιτολογίας–Διατροφής του Χαροκοπείου Πανεπιστημίου Δημοσθένη Παναγιωτάκο, συνεπιβλέποντα της διδακτορικής μου διατριβής. Τον ευχαριστώ τόσο για την επιστημονική του καθοδήγηση, όσο και για τον χρόνο που αφιέρωσε, και τις ευκαιρίες που μου έδωσε να αποκτήσω διδακτική και ερευνητική εμπειρία όλα αυτά τα χρόνια.

Ένα μεγάλο ευχαριστώ οφείλω και στον Ομότιμο Καθηγητή Ανδρέα Κυριακούση του Τμήματος Πληροφορικής & Τηλεματικής του Χαροκοπείου Πανεπιστημίου, που ως μέλος της τριμελούς επιτροπής συνέβαλε στην ολοκλήρωσή της διατριβής μου. Επίσης, τον κ. Κυριακούση θα ήθελα να την ευχαριστήσω και για τις βάσεις που μου έδωσε στον τομέα της Συνδυαστικής και των Πιθανοτήτων στις προπτυχιακές μου σπουδές. Χωρίς αυτές τις βάσεις, η ολοκλήρωση της συγκεκριμένης διδακτορικής διατριβής θα ήταν πιο επίπονη.

Θέλω, επίσης, να ευχαριστήσω τους συντελεστές της μελέτης ΑΤΤΙΚΑ για την δυνατότητα που μου έδωσαν να χρησιμοποιήσω μέρος των δεδομένων της μελέτης για τις ανάγκες της παρούσας διατριβής. Ευχαριστώ τον υπ. διδάκτορα του Χαροκοπείου Πανεπιστημίου Παναγιώτη Ανδρεόπουλο για την αδιάκοπη συνεργασία που είχαμε και το συνάδελφο Γεώργιο Αντωνόπουλο για τη βοήθεια που μου παρείχε κατά την πραγματοποίηση της διατριβής. Ευχαριστώ τον Άγγελο Χαραλαμπίδη, διδάκτορα Πληροφορικής για την πολύτιμη βοήθεια του σε τεχνικές δυσκολίες που αντιμετώπισα κατά τη συγγραφή της διατριβής. Επίσης, θα ήθελα να ευχαριστήσω τον καλό μου φίλο Δημήτριο Μουρκογιάννη, υπ. διδάκτωρ της σχολής Ναυπηγών του ΕΜΠ, για την υποστήριξη που μου παρείχε όλα αυτά τα χρόνια, αλλά και τις πολύτιμες συμβουλές του γενικότερα. Τέλος, ένα μεγάλο ευχαριστώ θα ήθελα να εκφράσω στη Δέσποινα Σακέλλη για τον σεβασμό που έδειξε στις επιλογές μου αλλά και την υποστήριξη της όλα τα χρόνια των σπουδών μου.

Περίληψη

Εισαγωγή: Η ανάπτυξη της επιστήμης της πληροφορικής (Informatics), με ανακάλυψη τεχνικών και εργαλείων σε επίπεδο γλώσσας μηχανής, έχει συντελέσει στην περαιτέρω ανάπτυξη πολλών άλλων επιστημονικών κλάδων όπως είναι αυτοί της Ιατρικής, της Βιολογίας, της Βιομετρίας, της Στατιστικής κτλ (*Sim et al, 2012*; *Wong, 2016*). Σε αρκετές περιπτώσεις, η παράλληλη ανάπτυξη δύο διακριτών κλάδων οδήγησε στη δημιουργία νέων επιστημονικών περιοχών, όπως για παράδειγμα, η εφαρμογή της Πληροφορικής και της Στατιστικής σε Βιο-Ιατρικές επιστήμες οδήγησε στους διακριτούς αλλά άρρηκτα συνδεδεμένους επιστημονικούς κλάδους της Βιο-πληροφορικής (Bioinformatics) και της Βιο-στατιστικής (Biostatistics) (*Moody, 2004; Indrayan, 2012.*). Η Βιο-στατιστική και η Βιο-πληροφορική είναι σχετικά πρόσφατοι επιστημονικοί τομείς, όμως, είναι γοργά αναπτυσσόμενοι τα τελευταία χρόνια. Στις Βιο-ιατρικές επιστήμες γίνεται χρήση κλιμάκων/δεικτών που στόχο έχουν την ακριβή, έγκυρη και αξιόπιστη μέτρηση ενός κλινικού ή συμπεριφοριστικού χαρακτηριστικού ενός ατόμου (*McDowell, 2006*). Οι κλίμακες χρησιμοποιούνται σε διάφορους επιστημονικούς τομείς όπως είναι η εκπαίδευση, η οικονομία, η βιομετρία, η ψυχομετρία και η ιατρική και αποσκοπούν στη μέτρηση διαφόρων κλινικών χαρακτηριστικών, τα οποία πρακτικά είναι δύσκολο να μετρηθούν ποσοτικά (*Kant, 1996; Streiner & Norman, 2008*). Οι κλίμακες (Scales) είναι ποσοτικές τυχαίες μεταβλητές, συνεχείς ή διακριτές, που ακολουθούν γνωστή ή άγνωστη κατανομή, και στόχο έχουν να αποτιμούν βιοχημικά (π.χ. Επίπεδο Ορού Σακχάρου, Επίπεδο Χοληστερόλης, κτλ), κλινικά (π.χ. Αίσθηση Πόνου, Κατάθλιψη, κτλ) ή συμπεριφοριστικά χαρακτηριστικά (π.χ. Γενικό Διαιτητικό Πρότυπο, Βαθμός Υιοθέτησης Μεσογειακής Διατροφής, κτλ) των ατόμων ενός πληθυσμού (*Beck, 1961; Huskisson 2012; Bach, 2006*). Οι κλίμακες κατασκευάζονται από το μη σταθμισμένο ή σταθμισμένο αριθμητικό άθροισμα *m* μεταβλητών. Οι *m* μεταβλητές αντιπροσωπεύουν διαφορετικές πτυχές της κατάστασης ενός ατόμου (π.χ. εκπαιδευτικό επίπεδεο, οικονομική κατάσταση, πάθηση, υιοθέτηση διατροφικού προτύπου, κτλ) που στοχεύει να αποτιμήσει ο δείκτης με σκοπό τον ορθό και έγκυρο διαχωρισμό (classification) (*Bansal & Pepe, 2013*). Η μεθοδολογία κατασκευής των κλιμάκων δεν είναι μονοσήμαντα καθιερωμένη στη διεθνή βιβλιογραφία, καθώς δεν υπάρχει πλαίσιο που να περιλαμβάνει σαφείς οδηγίες για όλα τα στάδια κατασκευής

τους, όπως π.χ. ο αριθμός των μεταβλητών που απαιτούνται ή τον τρόπο βαθμονόμησης αυτών (*Panagiotakos, 2009*).

Σκοπός: Σκοπός της παρούσας διδακτορικής διατριβής είναι η ποσοτική διερεύνηση, με χρήση θεωρητικής ή/και αναλυτικής αποδεικτικής διαδικασίας, της σχέσης μεταξύ της ακρίβειας των κλιμάκων και συγκεκριμένων δομικών χαρακτηριστικών, όπως είναι:

α) Ο αριθμός k των διαμερίσεων/διαιρέσεων των m διακριτών συνιστωσών μεταβλητών, στην περίπτωση μιας διακριτής κλίμακας (μια κλίμακα που δημιουργείται από διακριτές συνιστώσες μεταβλητές), (*Bersimis F. et al, 2013*).

β) Το εύρος l του στηρίγματος/πεδίου ορισμού των m συνεχών συνιστωσών μεταβλητών, στην περίπτωση μιας συνεχούς κλίμακας (μια κλίμακα που δημιουργείται από συνεχείς συνιστώσες μεταβλητές), (*Bersimis F. et al, 2016*).

γ) Η απόδοση σταθμίσεων w_i στις συνιστώσες μεταβλητές μιας κλίμακας. Η τιμή της στάθμισης εκφράζει το βαθμό συσχέτισης ανάμεσα στις συνιστώσες μεταβλητές και το χαρακτηριστικό Y που η κλίμακα στοχεύει να αξιολογήσει (*Bersimis F. et al, 2017*).

δ) Ο αριθμός m των συνιστωσών μεταβλητών μιας κλίμακας (*Bersimis F. et al, 2017*).

Από τις προαναφερθείσες επιμέρους διερευνήσεις, ολιστικός σκοπός της διδακτορικής διατριβής αυτής είναι να σχηματιστεί ένα προτεινόμενο μεθοδολογικό υπόβαθρο για την κατασκευή κλιμάκων με υψηλή διαχωριστική/διακριτική ικανότητα και προβλεπτική ακρίβεια.

Υλικό-Μέθοδος: Στην παρούσα διατριβή χρησιμοποιούνται θεωρητικές αποδείξεις με χρήση στοιχείων άλγεβρας και ανάλυσης, καθώς επίσης και πλήθος σεναρίων με προσομοιωμένα δεδομένα. Πιο συγκεκριμένα, στην περίπτωση της διερεύνησης της ευαισθησίας ως προς τον αριθμό k των διαιρέσεων/διαμερίσεων των διακριτών συνιστωσών για μεταβλητές που ακολουθούν άγνωστη κατανομή ή την ομοιόμορφη, αποδεικνύονται θεωρητικά οι αντίστοιχες προτάσεις με χρήση στοιχείων άλγεβρας (Ακολουθίες και Σειρές). Θεωρητικές αποδείξεις χρησιμοποιούνται επίσης στην περίπτωση της διερεύνησης της ευαισθησίας ως προς το εύρος l του στηρίγματος των συνεχών συνιστωσών μεταβλητών, για ομοιόμορφα ή κανονικά κατανεμημένες μεταβλητές, με χρήση στοιχείων ανάλυσης (Ολοκληρωτικός Λογισμός). Στην περίπτωση της διερεύνησης της χρήσης σταθμίσεων σε κλίμακες, με χρήση προσομοίωσης (Monte Carlo) δημιουργήθηκε ένα σύνολο 1.000 αρχείων με

διαφοροποίηση ως προς τις παραμέτρους των χρησιμοποιούμενων κατανομών, ως προς τον αριθμό *m* των συνιστωσών μεταβλητών, ως προς το μέγεθος *n* των δειγμάτων (π.χ. 100, 1.000, 10.000 κτλ), ως προς την αναλογία μεταξύ ατόμων με κάποιο χαρακτηριστικό και ατόμων που δεν διαθέτουν το χαρακτηριστικό (1:3, 1:4, κτλ) (*Sheldon, 2006*). Πιο συγκεκριμένα, έγινε χρήση της ομοιόμορφης και της κανονικής κατανομής με ποικιλία παραμετροποίησης για κάθε σενάριο. Επίσης γίνεται χρήση τροποποιημένων κατανομών ώστε να προσεγγίζουν τις εμπειρικές κατανομές από διάφορα χαρακτηριστικά του εκπαιδευτικού, του οικονομικού και του ιατρο-βιολογικού χώρου. Στα πλαίσια αυτής της διατριβής κατασκευάστηκαν εννέα κλίμακες T_i, $i=0, 1,2,,\ldots,8$ με διαφορετική μεθοδολογία στάθμισης η κάθε μια. Για την κλίμακα T_0 δεν χρησιμοποιήθηκε στάθμιση, οπότε παράγεται από το απλό αλγεβρικό άθροισμα των μεταβλητών X_j, $j=1,2,..,m$, κατόπιν τυποποίησης τους. Οι σταθμίσεις προέκυψαν από την εφαρμογή της λογιστικής παλινδρόμησης και της διαχωριστικής ανάλυσης με εξαρτημένη μεταβλητή την *Y* που εκφράζει την κατάσταση ενός ατόμου και ανεξάρτητες τις επεξηγηματικές συνιστώσες μεταβλητές X_j, $j=1, 2,..,m$.. Η διαγνωστική ακρίβεια της εκάστοτε κλίμακας, αξιολογείται υπολογίζοντας το εμβαδόν (AUC) της επιφάνειας κάτω από την (ROC) καμπύλη λειτουργικών χαρακτηριστικών, την ευαισθησία (Sensitivity - True Positive Ratio), την ειδικότητα τους (Specificity - True Negative Ratio), την ακρίβεια (Accuracy - True Ratio), την θετική και αρνητική προγνωστική αξία (Positive/Negative Predictive Value). Με χρήση προσομοίωσης αξιολογείται επίσης η προβλεπτική τους ικανότητα και οι σταθμισμένες κλίμακες αξιολογούνται στην περίπτωση πραγματικών διαιτολογικών δεδομένων, από τη μελέτη ATTICA, και δημιουργούνται χρησιμοποιώντας 5 συνιστώσες, οι οποίες επιλέγονται τυχαία από το γενικό πρότυπο της Μεσογειακής διατροφής.

Αποτελέσματα: Η θεωρητική απόδειξη, τόσο στην περίπτωση των διακριτών κλιμάκων, όσο και στην περίπτωση των συνεχών κλιμάκων, παρουσιάζει ότι η διαγνωστική ακρίβεια μιας κλίμακας, μετρούμενη μέσω της συνάρτησης ευαισθησίας, βελτιώνεται όταν ο αριθμός *k* των διαμερίσεων (διαιρέσεων) των διακριτών μεταβλητών αυξάνει ή όταν το εύρος *l* του στηρίγματος (πεδίου ορισμού) των συνεχών μεταβλητών γίνεται πιο ευρύ. Ο βαθμός συσχέτισης των συνιστωσών μεταβλητών με την εξαρτημένη βρέθηκε να επηρεάζει τη διαγνωστική ακρίβεια μιας κλίμακας. Πιο συγκεκριμένα, χρησιμοποιήθηκαν αρχικά σετ προσομοιωμένων δεδομένων από συσχετισμένες μεταβλητές με διαφορετικό συντελεστή συσχέτισης με

την εξαρτημένη μεταβλητή. Η διαγνωστική ακρίβεια των σταθμισμένων κλιμάκων είναι υψηλότερη σε σχέση με αυτή της μη-σταθμισμένης κλίμακας. Η διαγνωστική ακρίβεια της κλίμακας βελτιώνεται επίσης στην περίπτωση που ο αριθμός *m* των συνιστωσών μεταβλητών αυξάνεται. Τα παραπάνω αποτελέσματα εφαρμόστηκαν και επαληθεύτηκαν σε κλίμακα που εκφράζει διατροφολογικό δείκτη, βασιζόμενο σε συνιστώσες μεταβλητές του MedDietScore, σε δείγμα 981 ανδρών και γυναικών, αναφορικά με τους παράγοντες καρδιαγγειακού κινδύνου (δηλαδή, παχυσαρκία, σακχαρώδη διαβήτη, υπέρταση, υπερχοληστερολαιμία), μετά από έλεγχο για πιθανούς συγχυτικούς παράγοντες.

Συμπεράσματα: Οι κλίμακες χρησιμοποιούνται ευρέως σε διαφόρους επιστημονικούς κλάδους, ειδικότερα στο χώρο της εκπαίδευσης, της οικονομίας και της υγείας, για την ανίχνευση χαρακτηριστικών ή στάσεων που είναι δύσκολο να ανιχνευθούν (*Kant, 1996*). Η ακριβής μέτρηση κάποιου χαρακτηριστικού, με τη χρήση μιας σύνθετης κλίμακας στο χώρο της υγείας, προσφέρει τη δυνατότητα έγκαιρης και ορθής διάγνωσης μιας νόσου, με συνέπεια την καταλληλότερη θεραπεία του ατόμου που νοσεί και κατ' επέκταση τη βελτίωση της ποιότητας της ζωής του. Σε αυτή τη διδακτορική διατριβή, προτείνονται μέθοδοι για τη βελτίωση της ακρίβειας κλιμάκων σχετικών με την υγεία που παρέχουν ωφελιμότητα σε επίπεδο δημόσιας υγείας με άμεσα αποτελέσματα τη χρήση λιγότερων κρατικών πόρων για τη θεραπεία των ασθενών ατόμων ενός πληθυσμού. Μια κλίμακα παρουσιάζει υψηλή διαχωριστική ικανότητα όταν ικανοποιούνται οι ακόλουθες συνθήκες:

α) Ο αριθμός *k* των διαμερίσεων/διαιρέσεων των *m* διακριτών συνιστωσών μεταβλητών λαμβάνει τη μέγιστη δυνατή τιμή, χωρίς να εμφανίζεται το πρόβλημα της δυασταξινόμησης των ατόμων ενός πληθυσμού, στην περίπτωση μιας διακριτής σύνθετης κλίμακας (*Bersimis F. et al, 2013*).

β) Το εύρος *l* του στηρίγματος/πεδίου ορισμού των *m* συνεχών συνιστωσών μεταβλητών γίνεται ευρύ, στην περίπτωση μιας συνεχούς σύνθετης κλίμακας, χωρίς να εμφανίζεται το πρόβλημα της δυσταξινόμησης των ατόμων ενός πληθυσμού (*Bersimis F. et al, 2016*).

γ) Η στάθμιση w_i στις συνιστώσες μεταβλητές μιας σύνθετης κλίμακας να προέρχεται από στατιστική μέθοδο ταξινόμησης, όπως είναι η Λογιστική Παλινδρόμηση ή η Διαχωριστική Ανάλυση (*Bersimis F. et al, 2017*).

δ) Ο αριθμός *m* των συνιστωσών μεταβλητών μιας σύνθετης κλίμακας που σχετίζονται με το χαρακτηριστικό που αποτιμά η κλίμακα να είναι ο μέγιστος,

δηλαδή να γίνεται χρήση όλων των συνιστωσών μεταβλητών που σχετίζονται με το αποτέλεσμα της υγείας (*Bersimis F. et al, 2017*).

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ ΠΕΡΙΛΗΨΗΣ

Bach, Anna, et al. "The use of indexes evaluating the adherence to the Mediterranean diet in epidemiological studies: a review." *Public health nutrition*9.1a (2006): 132-146.

Bansal, Aasthaa, and Margaret Sullivan Pepe. "When does combining markers improve classification performance and what are implications for practice?." *Statistics in medicine* 32.11 (2013): 1877-1892.

Beck, A., et al. "An inventory for measuring depression Arch Gen Psychiatry 4: 561–571." *Find this article online* (1961).

Bersimis, Fragkiskos, Demosthenes Panagiotakos, and Malvina Vamvakari. "Sensitivity of Health Related Indices is a Non-Decreasing Function of their Partitions." *Journal of Statistics Applications & Probability* 2.3 (2013): 183.

Bersimis, Fragkiskos, Demosthenes Panagiotakos, and Malvina Vamvakari. "Investigating the sensitivity function's monotony of a health-related index." *Journal of Applied Statistics* 44.9 (2016): 1680-1706.

Bersimis, Fragkiskos, Demosthenes Panagiotakos, and Malvina Vamvakari. " The use of components' weights improves the diagnostic accuracy of a health-related index." *Communication in Statistics: Theory and Methods* 44.9 (2017): 1680-1706.

Huskisson, E. C. "Measurement of pain." *The lancet* 304.7889 (1974): 1127-1131.

Indrayan, Abhaya. *Medical biostatistics*. CRC Press, 2012.

Kant, Ashima K. "Indexes of overall diet quality: a review." *Journal of the American Dietetic Association* 96.8 (1996): 785-791.

McDowell, Ian. *Measuring health: a guide to rating scales and questionnaires*. Oxford university press, 2006.

Moody, Glyn. *Digital code of life: how bioinformatics is revolutionizing science, medicine, and business*. John Wiley & Sons, 2004.

Panagiotakos, Demosthenes. "Health measurement scales: methodological issues." *The open cardiovascular medicine journal* 3 (2009): 160-165.

Pepe, Margaret Sullivan, and Mary Lou Thompson. "Combining diagnostic test results to increase accuracy." *Biostatistics* 1.2 (2000): 123-140.

Sheldon M. Ross, Simulation (4th edition), Simulation, 2006, Elsevier Academic Press

Streiner, David L., and Geoffrey R. Norman. "Health measurement scales: a practical guide to their development and use." *Health measurement scales: a practical guide to their development and use*. Oxford University Press, 2008.

Sim, Adelene Y. L., Peter Minary, and Michael Levitt. "Modeling Nucleic Acids." *Current opinion in structural biology* 22.3 (2012): 273–278. *PMC*. Web. 31 July 2017

Wong, K.C. (2016). Computational Biology and Bioinformatics: Gene Regulation. CRC Press, Taylor & Francis Group.

Abstract

Background/Introduction: The development of information science with the discovery of techniques and tools at the machine language level has contributed to the further development of many other disciplines such as Medicine, Biology, Biometrics, Statistics etc. (*Sim et al., 2012; Wong, 2016*). In several cases, the parallel development of two distinct branches has led to the creation of new scientific areas. For example, the application of Informatics and Statistics to Bio-Medical Sciences has led to the distinct but inextricably linked disciplines of Bioinformatics and Biostatistics *(Moody, 2004; Indrayan, 2012*). Bio-Statistics and Bio-Informatics are relatively recent scientific fields, however, they are rapidly growing in recent years. Biomedical sciences use mathematical scales to measure accurately and reliably a clinical or behavioral characteristic of a person (*McDowell, 2006*). Composite scales are used in a variety of scientific areas, such as biometrics, psychometry and medicine, and are designed to measure various clinical features, which are difficult to measure quantitatively (*Kant, 1996; Streiner & Norman, 2008*). Scales are quantitative, continuous or discrete, random variables that follow a known or unknown distribution, intending to evaluate biochemical (eg, Serum Serum Level, Cholesterol Level, etc), Clinical (eg Sense of Pain, Depression, etc.) or behavioral characteristics (eg General Dietary Standard, Degree of Adoption of Mediterranean Diet, etc.) of individuals of a population (*Beck, 1961; Huskisson 2012*). Scales are derived from the simple (unweighted) or weighted algebraic sum of a set of *m* variables. The *m* variables express distinct components of the feature that the scale aims to estimate. Advanced health indices are also used as healthy/patient segregation tools and are widely used in physic-biological, psychometric and social science research (*Bansal & Pepe, 2013*). The methodology for the construction of scales is not unambiguously established in the international literature but involves a high degree of subjectivity of the researcher and depends on the type of the characteristic that the scale aims to assess (Panagiotakos, 2009).

Aim: The purpose of this dissertation is to quantify the relationship between the diagnostic accuracy of composite scales and their specific structural characteristics, influencing their diagnostic accuracy, using a theoretical and/or analytical proofing process, such as:

(A) The k number of partitions/divisions of m discrete component variables, in the case of a discrete composite scale (i.e. an index created by discrete component variables), (*Bersimis F. et al, 2013*).

(B) The l range of the support/field of definition of m continuous component variables, in the case of a continuous composite scale (an index generated by continuous component variables), (*Bersimis F. et al., 2016*).

(C) The performance of weights w_i in the component variables of a composite scale. The weighting value expresses the degree of correlation between the component variables and the Y characteristic that the scale aims to evaluate (*Bersimis F. et al., 2017*).

(D) The number m of the composite scale's variables (*Bersimis F. et al, 2017*).

From the aforementioned individual investigations, the holistic goal of this dissertation is to form a proposed methodological background for the construction of scales with high diagnostic ability and predictive precision.

Material-Method: This thesis uses theoretical evidence using algebra and analysis elements, as well as a number of simulated data scenarios. More specifically, in the case of investigating the sensitivity of the number k of divisions/partitions of discrete components for variables that follow an unknown distribution or the uniform, the corresponding propositions using algebra elements (Sequences and Rows) are theoretically proved. Theoretical proof is also used in the case of exploration of the sensitivity of continuous indices for uniformly or normally distributed variables using analysis elements (Differential and Integral Calculus). In the case of investigating the use of weights in composite health indices using simulation (Monte Carlo), a set of 1,000 records were created with a differentiation in the parameters of the distributions used, in terms of the number m of the component variables, in terms of size of the samples (e.g., 100, 1,000, 10,000, etc.). More specifically, uniform and normal distribution were used with a variety of parameterization for each scenario. Also, modified distributions are used to approximate the empirical distributions of various characteristics of the medical-biological space. In the context of this thesis, nine scales T_i, $i = 0, 1,2, ..., 8$ were constructed with different weighting methodology each. For T_0 no weighting was used, so it is derived from the simple algebraic sum of the variables X_j, $j = 1,2, .., m$. Weights were derived from the application of logistic regression and discriminant analysis with a dependent variable Y expressing the clinical condition of individuals and independent of the explanatory component

variables X_j, $j = 1, 2, .., m$. The diagnostic accuracy of the scales was evaluated by calculating the surface area under the (ROC) receiver operating characteristic curve, Sensitivity (True Positive Ratio), Specificity (True Negative Ratio), Accuracy (True Ratio) Positive and Negative predictive value. Using simulation, their predictive capacity was also evaluated and weighted indices were evaluated in the case of actual dietary data from the ATTICA study and are generated using 5 components, which are randomly selected from the Mediterranean diet.

Results: The theoretical evidence, both in the case of discrete indices and in the case of continuous scales, shows that the diagnostic accuracy of a scale measured by the sensitivity function, improves when the number *k* of the divisions of discrete variables increases or when the width *l* of the support (field of definition) of continuous variables becomes broader. The degree of correlation between the component variables was found to affect the accuracy of a scale. More specifically, initially a set of simulated data was used from correlated variables with a different correlation coefficient, and then the Principal Component Analysis was used to produce rectangular (uncorrelated) variables. An increase in the capacity of the scales was observed when using the unrelated variables produced. Accuracy of the scale is also improved if the number *m* of component variables increase too. The accuracy of the weighted scales is higher than the diagnostic accuracy of a non-weighted scale. The above results were applied to a nutritional scale based on MedDieScore component variables in a sample of 981 men and women in terms of cardiovascular risk factors (i.e. obesity, diabetes mellitus, hypertension, hypercholesterolemia) after screening for potential confounders.

Conclusions: A composite scale presents high ability when the following conditions are met:

A) The number *k* of the partitions/divisions of m discrete component variables takes the maximum possible value, without the problem of the misclassification of individuals of a population in the case of a discrete composite scale (*Bersimis F. et al, 2013*).

B) The *l* range of the support/field of definition of the *m* continuous component variables, in the case of a continuous composite scale, without the problem of the misclassification of the individuals of a population (*Bersimis F. et al., 2016*).

C) The weighting w_i in the component variables of a composite scale is derived from a statistical classification method, such as Logistic Regression or Discriminant Analysis (*Bersimis F. et al., 2017*).

D) The number m of composite health indicator variables is the maximum, i.e. all the variables associated with the outcome are used (*Bersimis F. et al., 2017*).

Scales are widely used in various disciplines, particularly in the health sector, to detect features or attitudes that are difficult to be detected (*Kant, 1996*). Accurate measurement of a characteristic, using a composite scale, offers the possibility of early and accurate diagnosis of a disease, resulting in the most appropriate treatment of the diseased person who suffers and hence the improvement of the quality of his life. Extending past studies, this thesis proposes methods to improve the h scales' diagnostic capabilities that provide public health benefits, with direct results in the use of less government resources to treat the sick in a population.

Πρόλογος

Η Πληροφορική (Informatics) είναι η θετική και εφαρμοσμένη επιστήμη που σχετίζεται με την διαχείριση των πληροφοριών και με την ανάπτυξη τεχνολογικών εφαρμογών σε ψηφιακά υπολογιστικά συστήματα από τη σκοπιά της σχεδίασης, της ανάπτυξης, της υλοποίησης, της διερεύνησης και της ανάλυσης (*Zembylas & Vrasidas, 2005; Rapaport, 2005*). Το επιστημονικό πεδίο της Πληροφορικής αποτελείται από πλήθος μεθόδων και εφαρμογών που χρησιμοποιούνται σε πληθώρα επιστημονικών χώρων όπως είναι αυτοί της υγείας, της βιολογίας της κοινωνιολογίας, της στατιστικής κτλ (*ΕΠΕ, 2006*). Πιο συγκεκριμένα, η χρήση μεθόδων και εφαρμογών της Πληροφορικής σε βιο-ιατρικά δεδομένα, έχει αποκτήσει διακριτή οντότητα ως επιστημονικός χώρος και ονομάζεται Βιο-Πληροφορική (Bio-Informatics).

Η βελτίωση της ταχύτητας και της αποτελεσματικότητας των υπολογιστικών συστημάτων και των αντιστοίχων λογισμικών που χρησιμοποιούνται τα τελευταία χρόνια στον ευρύτερο επιστημονικό κλάδο της Πληροφορικής και της Βιολογίας, έδωσε τη δυνατότητα ανάπτυξης σε διεπιστημονικούς κλάδους όπως είναι η Βιο-πληροφορική (*Hogeweg, 2011*). Για παράδειγμα, στο χώρο της ιατρικής η τεχνητή νοημοσύνη ή η επικοινωνία ανθρώπου-μηχανής είναι απαραίτητα επιστημονικά πεδία για περεταίρω ανάπτυξη και τελειοποίηση διαφόρων ιατρικών μεθόδων και πρακτικών (Λαπαροσκοπική μέθοδος χειρουργείου με την χρήση των οπτικών ινών και της τεχνολογίας βίντεο κτλ). Το διεπιστημονικό πεδίο της Βιο-πληροφορικής αναπτύσσει εργαλεία λογισμικού της υπολογιστικής επιστήμης, υπολογιστικές μεθόδους της στατιστικής και των μαθηματικών για την κατανόηση, ανάλυση και ερμηνεία διάφορων βιολογικών δεδομένων όπως είναι η αλληλουχία γονίδιων (DNA, RNA), βιοχημικών χαρακτηριστικών (υπέρταση, σακχαρώδης διαβήτης) κτλ (*Joyce, 2014*).

Η χρήση μεθόδων ταξινόμησης και ανίχνευσης ατόμων με συγκεκριμένα χαρακτηριστικά συναντάται σε επιστήμες όπως είναι η Στατιστική και συγγενείς επιστημονικές περιοχές αυτής (*Schmidhuber, 2015*). Για παράδειγμα, για την κατάταξη ατόμων με συγκεκριμένα κοινωνικά χαρακτηριστικά ασχολείται η κοινωνική δημογραφία (*Gavrilov & Gavrilova, 2010*) και για την ορθή διάκριση ασθενών και υγειών ατόμων χρησιμοποιούνται τεχνικές της Βιο-στατιστικής και της Επιδημιολογίας (*Elliott & Shepherd, 2006*). Η Βιο-στατιστική είναι ο εξειδικευμένος

κλάδος της επιστήμης της Στατιστικής που εφαρμόζει στατιστικές τεχνικές στον ευρύτερο χώρο της ιατρικής και της βιολογικής έρευνας. Στο αντικείμενο της Βιο-στατιστικής εγκλείεται ο ορθός σχεδιασμός κλινικών δοκιμών και στην ανάλυση των δεδομένων με χρήση κατάλληλων ελέγχων υποθέσεων και επαρκούς ερμηνείας των παραγόμενων αποτελεσμάτων. Πιο συγκεκριμένα, η Βιο-στατιστική περιλαμβάνει την εξαγωγή έγκυρων συμπερασμάτων για διάφορους πληθυσμούς με χρήση μεθόδων από την Επαγωγική Στατιστική σε αντιπροσωπευτικά δείγματα, δηλαδή κατόπιν της συλλογής δεδομένων με ειδικές μεθόδους δειγματοληψίας, (*Μπερσίμης. 2016)*. Η Επιδημιολογία, βασίζεται σε μεθόδους της Βιοστατιστικής για την επιστημονική μελέτη των παραγόντων που επηρεάζουν την συχνότητα εμφάνισης ασθενειών σε έναν πληθυσμό (*Zhang, 2004*).

Η Βιο-πληροφορική και η Βιο-στατιστική αποτελούν πολύτιμους συγγενείς κλάδους των ιατροβιολογικών επιστημών με απώτερο στόχο την βελτίωση της ποιότητας και διάρκειας της ανθρώπινης ζωής (*Ewens & Grand, 2006*.). Πιο συγκεκριμένα, τις τελευταίες δεκαετίες η εφαρμογή στατιστικών μεθόδων έγινε πλέον αναντικατάστατο μέρος της ιατροβιολογικής έρευνας καθώς η επιστήμη της πληροφορικής επεκτάθηκε και δημιουργήθηκε ο κλάδος της Βιοπληροφορικής, ο οποίος είναι ένας ευρύτατος διεπιστημονικός κλάδος στον οποίο συνυπάρχουν και συνεργάζονται αρμονικά και αμοιβαία πολλοί επιστημονικοί κλάδοι, από τη Βιολογία μέχρι τα καθαρά μαθηματικά, τις πιθανότητες, τη στατιστική, τις διαφορικές εξισώσεις και τη στοχαστική ανάλυση. Η χρήση κατάλληλων στατιστικών και μαθηματικών λογισμικών/πακέτων (SPSS, R, STATA, Minitab, Matlab, κτλ) που ολοένα βελτιώνονται και προσφέρουν νέες τεχνικές παρέχοντας τη δυνατότητα για επεξεργασία δεδομένων από μεγάλου μεγέθους αποθήκες δεδομένων.

Στις επιστήμες υγείας, βασική επιδίωξη αποτελεί η έγκυρη και αξιόπιστη μέτρηση μεταβλητών, με στόχο την αδιάκοπη βελτίωση των παρεχόμενων υπηρεσιών υγείας στα μέλη της σύγχρονης κοινωνίας (*Greenland & Morgenstern, 2001*). Είναι προφανές ότι υπάρχουν ανθρώπινα χαρακτηριστικά (σωματομετρικά, βιολογικά, βιοχημικά, αιματολογικά, κτλ) όπως είναι το ύψος ενός ανθρώπου, η αρτηριακή πίεση ενός ανθρώπου, το επίπεδο σακχάρου στο αίμα ενός ανθρώπου, που μετρώνται με εξαιρετική ακρίβεια λόγω της ύπαρξης καθιερωμένων και αντικειμενικών κλιμάκων μέτρησης, υπό την προϋπόθεση χρήσης κοινών αντιδραστηρίων όπου αυτά απαιτούνται. Υπάρχουν όμως, επίσης, ανθρώπινα χαρακτηριστικά τα οποία δεν είναι εύκολο να μετρηθούν με απόλυτη ακρίβεια, όπως είναι ο βαθμός κατάθλιψης ή

άγχους (*Aitken, 1969*), η ένταση πόνου που αισθάνεται ένας ασθενής (*Huskisson, 1976*), ο βαθμός υιοθέτησης ενός συγκεκριμένου διατροφολογικού προτύπου ποσοτικά και ποιοτικά (*Bach et al, 2006*) και η συνολική ποιότητα ζωής (*Ferrans & Powers, 2010*), λόγω της έλλειψης μιας αντικειμενικής κλίμακας μέτρησης. υπάρχουν επίσης χαρακτηριστικά στην εκπαίδευση και στην οικονομία που δεν μπορούν με τρηθούν ποσοτικά και αντικειμενικά όπως είναι η ποιότητα της παρεχόμενης εκπαίδευσης σε πανεπιστημιακό ίδρυμα. Τα εργαλεία μέτρησης που χρησιμοποιούνται για την μέτρηση χαρακτηριστικών που δεν εκφράζονται ποσοτικά με αντικειμενικό τρόπο ονομάζονται «σύνθετες κλίμακες» ή «σύνθετοι δείκτες», και στοχεύουν στην ανίχνευση ή πρόβλεψη ενός συγκεκριμένου χαρακτηριστικού (π.χ. κατάθλιψη, άγχος, αίσθηση πόνου κτλ) (*Beck, 1961*).

Οι κλίμακες χρησιμοποιούνται ευρέως στη βιοιατρική έρευνα (*Kant, 1996; Streiner & Norman, 2008*) μετατρέποντας την πληροφορία από πλήθος διαφορετικών χαρακτηριστικών σε ένα, μονόμετρο συνήθως, μέγεθος που αποτιμά ολιστικά την κλινική κατάσταση ενός ατόμου. Παραδείγματα κλιμάκων, εδώ και αρκετές δεκαετίες, είναι, από τον επιστημονικό κλάδο της ψυχομετρίας, η κλίμακα μέτρησης έντασης της κατάθλιψης (*Jackson, 1970*) και από τον κλάδο της διαιτολογίας-διατροφολογίας η κλίμακα μέτρησης της επάρκειας και της ποιότητας της τροφής ενός ατόμου (*Kant, 1996*). Επομένως, οι κλίμακες είναι σημαντικά εργαλεία για την ανίχνευση διαφόρων παθήσεων ή συμπεριφοριστικών προτύπων σε άτομα γενικών ή ειδικών πληθυσμών. Οι δείκτες αυτοί λαμβάνουν υπόψη διαφορετικές μεταβλητές και συνήθως εκφράζονται μέσω του μη σταθμισμένου αθροίσματός τους σύμφωνα με τη σχέση $T_m = \sum_{i=1}^{m} X_i$. Η κλίμακα T_m κατατάσσει το εκάστοτε άτομο, σύμφωνα με την αριθμητική τιμή της, συγκρίνοντας την με συγκεκριμένη τιμή l που ονομάζεται διαγνωστικό κατώφλι ή διαγνωστικό όριο και έχει επιλεγεί κατάλληλα ώστε να διακρίνει τα άτομα ως ασθενείς και μη ασθενείς. Οι συνιστώσες μεταβλητές δύναται να είναι διακριτές ή συνεχείς, με αποτέλεσμα, η παραγόμενη κλίμακα να είναι διακριτή ή συνεχής, πεπερασμένου εύρους. Η κλίμακα MedDietScore (*Panagiotakos et al, 2007*) που αποτιμά τον βαθμό υιοθέτησης της μεσογειακής διατροφής έχει διακριτό πεδίο ορισμού, ενώ, η κλίμακα VAS που αξιολογεί την ένταση πόνου ασθενών (*Huskisson, 1974*) λαμβάνει τιμές σε συνεχές σύνολο τιμών. Για παράδειγμα, η κλίμακα μέτρησης της κατάθλιψης (*Zung, 1965*) με 20 συνιστώσες μεταβλητές έχει διακριτό πεδίο ορισμού καθώς κάθε μια συνιστώσα μεταβλητή

δύναται να λάβει 4 τιμές/απαντήσεις, οπότε κάθε μια συνιστώσα έχει 4 διαμερίσεις. Στην ουσία οι κλίμακες που σχετίζονται με κάποια πάθηση έχουν ως στόχο, αξιολογώντας την πληροφορία από τις συνιστώσες διακριτές ή συνεχείς μεταβλητές να κατατάξουν ένα άτομο ως ασθενή ή μη. Για παράδειγμα, δύναται να θεωρηθεί μια δίτιμη μεταβλητή Y που να εκφράζει την πραγματική κλινική κατάσταση ενός ατόμου και έστω ότι Y=1 σημαίνει ότι το άτομο φέρει ένα συγκεκριμένο χαρακτηριστικό (είναι ασθενής) και Y=0 σημαίνει ότι το άτομο δεν φέρει ένα συγκεκριμένο χαρακτηριστικό (είναι υγειές). Μια κλίμακα κατασκευάζεται από το άθροισμα m συνιστωσών μεταβλητών X_i, $i=1,2,\ldots,m$, συνήθως ασυσχέτιστων (ανεξάρτητων) μεταξύ τους (Ισχύει $Cor(X_i,X_j)\approx 0,\ i\neq j$), αλλά εξαρτώμενες με την πάθηση που στοχεύουν να ανιχνεύσουν ($Cor(X_i,Y)\neq 0,\ i=1,2,\ldots,m$).

Επομένως, ιδιαίτερο ενδιαφέρον παρουσιάζουν τα ποσοστά επιτυχίας και αποτυχίας των προαναφερθέντων κλιμάκων, καθώς, μια εσφαλμένη μέτρηση στο χώρο της υγείας, δύναται να οδηγήσει σε δυσμενείς, ακόμα και σε μη αναστρέψιμες συνθήκες υγείας (π.χ. θάνατος κάποιου ασθενούς εξαιτίας εσφαλμένης διάγνωσης). Επομένως, η χρήση έγκυρων εργαλείων μέτρησης, δηλαδή κλιμάκων υγείας, αποτελεί αναγκαιότητα για την ορθή διερεύνηση και ανίχνευση τυχόν ασθενειών (*Brandt & Gardner. 2000*), ειδικά σε περιπτώσεις που η ποσοτικοποίηση συγκεκριμένων χαρακτηριστικών όπως τα προαναφερθέντα, δεν εμπίπτουν σε καθιερωμένες μεθόδους αντικειμενικής μέτρησης.

Ως προς τη μεθοδολογία κατασκευής των κλιμάκων, και παρά το γεγονός ότι οι κλίμακες αποτελούν αναντικατάστατα εργαλεία μέτρησης που χρησιμοποιούνται σε πληθώρα επιστημονικών πεδίων όπως στην εκπαίδευση, στην οικονομία, στη βιολογία, στην ιατρική, στη διαιτολογία, στην ψυχομετρία στην κοινωνιολογία κτλ, δεν υπάρχει ένα σαφές πλαίσιο. Πιο συγκεκριμένα, το πλήθος m των συνιστωσών μεταβλητών που αθροίζονται και δημιουργούν την κλίμακα, ο αριθμός k των διαμερίσεων/διαιρέσεων κάθε διακριτής συνιστώσας ή το εύρος l του στηρίγματος κάθε συνεχούς συνιστώσας και ο συντελεστής βαρύτητας w κάθε συνιστώσας επιλέγονται με υποκειμενικό τρόπο, εξαρτώμενα από τον εκάστοτε ερευνητή και όχι με κάποιο αντικειμενικό πλαίσιο κριτήριων. Η πλειονότητα των χρησιμοποιούμενων κλιμάκων που χρησιμοποιούνται είναι κατασκευασμένες συνήθως με μικρό αριθμό διαμερίσεων/διαιρέσεων ανά συνιστώσα διακριτή μεταβλητή ή συγκεκριμένο εύρος ανά συνιστώσα συνεχή μεταβλητή. Επομένως, η αξιολόγηση του ίδιου συμπεριφοριστικού χαρακτηριστικού δύναται να παρέχει διαφορετικά αποτελέσματα

λόγω της χρήσης διαφορετικών κλιμάκων (*Bach, 2006*) και η εγκυρότητα τους κρίνεται υπό αμφισβήτηση όταν η ίδια ποσότητα μετριέται με διαφορετικά εργαλεία μέτρησης παρέχοντας διαφορετικά αριθμητικά αποτελέσματα.

Η παρούσα διατριβή αυτή στοχεύει να:

α) Αξιολογήσει θεωρητικώς κατά πόσον ο αριθμός *k* των διαιρέσεων (διαμερίσεων) των διακριτών συνιστωσών μεταβλητών επηρεάζει την ευαισθησία της ανίχνευσης ενός δυαδικού αποτελέσματος, δηλαδή εξετάζεται κατά πόσο η διαφοροποίηση του αριθμού των διαιρέσεων των διακριτών συνιστωσών μεταβλητών μιας κλίμακας επηρεάζει την ευαισθησία της. Έτσι, επιμέρους σκοπός αυτής της εργασίας ήταν να διερευνήσει την υπόθεση, η οποία ισχυρίζεται ότι η ευαισθησία μιας κλίμακας αυξάνεται όταν το πλήθος *k* των διαιρέσεων (διαμερίσεις) των *m* διακριτών συνιστωσών μεταβλητών της κλίμακας αυξάνουν επίσης, και να βασίζεται σε μια θεωρητική προσέγγιση (*Bersimis et al, 2013*).

β) Αξιολογήσει θεωρητικώς κατά πόσον το εύρος *l* του στηρίγματος των συνεχών συνιστωσών μεταβλητών επηρεάζει την ευαισθησία της ανίχνευσης ενός δυαδικού αποτελέσματος, δηλαδή εξετάζεται κατά πόσο η διαφοροποίηση του εύρους του στηρίγματος των συνεχών συνιστωσών μεταβλητών μιας κλίμακας επηρεάζει την ευαισθησία της. Έτσι, επιμέρους σκοπός αυτής της εργασίας ήταν να διερευνήσει την υπόθεση, η οποία ισχυρίζεται ότι η ευαισθησία μιας κλίμακας αυξάνεται όταν το στήριγμα *l* των *m* διακριτών συνιστωσών μεταβλητών της κλίμακας γίνεται πιο ευρύ επίσης, και να βασίζεται σε μια θεωρητική προσέγγιση (*Bersimis et al, 2016*).

γ) Αξιολογήσει κατά πόσον η χρήση σταθμίσεων *w* στις συνιστώσες μεταβλητές επηρεάζει την ακρίβεια της ανίχνευσης ενός δυαδικού αποτελέσματος, δηλαδή εξετάζεται κατά πόσο η απόδοση συγκεκριμένων βαρών στις συνιστώσες μεταβλητές μιας κλίμακας επηρεάζει την ακρίβεια της. Έτσι, επιμέρους σκοπός αυτής της εργασίας ήταν διερευνήσει την υπόθεση, η οποία ισχυρίζεται ότι η ακρίβεια μιας κλίμακας αυξάνεται όταν γίνεται χρήση σταθμίσεων στις συνιστώσες μεταβλητές της (*Bersimis et al, 2017*).

δ) Αξιολογήσει κατά πόσον ο αριθμός *m* των συνιστωσών μεταβλητών επηρεάζει τη διαγνωστική ικανότητα της ανίχνευσης ενός δυαδικού αποτελέσματος, δηλαδή εξετάζεται κατά πόσο η διαφοροποίηση του αριθμού *m* των συνιστωσών μεταβλητών μιας κλίμακας επηρεάζει την ακρίβεια της. Έτσι, επιμέρους σκοπός αυτής της εργασίας ήταν να διερευνήσει την υπόθεση, η οποία ισχυρίζεται ότι η

ακρίβεια μιας κλίμακας αυξάνεται όταν ο αριθμός *m* των συνιστωσών μεταβλητών της κλίμακας αυξάνουν επίσης (*Bersimis et al, 2017*).

Τα αποτελέσματα της διατριβής αυτής ευελπιστούμε να αποτελέσουν ένα επιπλέον βήμα προς την υλοποίηση ενός αυστηρού και αντικειμενικού μεθοδολογικού πλαισίου που διέπει γενικά την κατασκευή κλιμάκων που χρησιμοποιούνται σε διάφορα επιστημονικά πεδία.

Φραγκίσκος Γ. Μπερσίμης

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ ΠΡΟΛΟΓΟΥ

R. Aitken, A growing edge of measurement of feelings, Proc. R. Soc. Med. 62 (1969), pp. 989–993.

Bach, Anna, et al. "The use of indexes evaluating the adherence to the Mediterranean diet in epidemiological studies: a review." Public health nutrition 9.1a (2006): 132.

E.C. Huskisson, Measurement of pain, The Lancet 304 (1974), pp. 1127–1131. doi:10.1016/S0140-6736(74)90884-8. Available at http://www.sciencedirect.com/science/article/pii/S0140673674908848. Originally published as Volume 2, Issue 7889.

F. Bersimis, D. Panagiotakos, and M. Vamvakari, Sensitivity of health related indices is a non-decreasing function of their partitions, J. Stat. Appl. Probab. 2 (2013), pp. 183–194.

W. Zung, A self-rating depression scale, Arch. Gen. Psychiatr. 12 (1965), pp. 63–70. doi:10.1001/archpsyc.1965.01720310065008.

Carlsson, Anna Maria. "Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale." Pain 16.1 (1983): 87-101.

Panagiotakos, Demosthenes B., et al. "Adherence to the Mediterranean food pattern predicts the prevalence of hypertension, hypercholesterolemia, diabetes and obesity, among healthy adults; the accuracy of the MedDietScore." Preventive medicine 44.4 (2007): 335-340.

1. ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ - ΟΡΙΣΜΟΙ - ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Η επιστήμη της Πληροφορικής έχει εισχωρήσει σε πολλές επιστημονικές περιοχές, όπως είναι η Στατιστική, η Εκπαίδευση, η Οικονομία, η Βιολογία, η Ιατρική, η Δημογραφία κτλ. Η Πληροφορική είναι η επιστήμη που ασχολείται με την πληροφορία και την εφαρμογή τεχνολογικών εφαρμογών με στόχο την ανάπτυξη υπολογιστικών συστημάτων από τη σκοπιά κυρίως της σχεδίασης, ανάπτυξης, σύνθεσης και υλοποίησης λογισμικού (*Tucker, 2004*).

Η Στατιστική εκμεταλλεύεται σε μεγάλο βαθμό τις δυνατότητες που παρέχουν τα εργαλεία της πληροφορικής, κυρίως σε επίπεδο ταχύτητας και πολλαπλής επαναληψιμότητας, αλλά, αντίστροφα, θα μπορούσε να ειπωθεί ότι ο τομέας της Πληροφορικής χρησιμοποιεί τα εργαλεία της Στατιστικής προς όφελος της ταχύτητας και της αποτελεσματικότητας διαφόρων αλγορίθμων.

Η χρήση της Πληροφορικής και της Στατιστικής σε βιο-ιατρικές επιστήμες έχει καθιερώσει συγκεκριμένα επιστημονικά υπο-πεδία όπως είναι η Βιοπληροφορική και η Βιοστατιστική που συνυπάρχουν, συνεργάζονται και αλληλοσυμπληρώνονται. Πιο συγκεκριμένα, υπάρχουν κοινά εργαλεία μέτρησης και ταξινόμησης από τις περιοχές της Βιοστατιστικής και της Βιοπληροφορικής που χρησιμοποιούνται στο χώρο της υγείας προς όφελος των ατόμων γενικών ή/και ειδικών πληθυσμών (*Greenes, 1990*). Αυτά τα εργαλεία έχουν ως στόχο την ακριβή μέτρηση και υψηλή προβλεπτική ικανότητα καθώς υπάρχουν πολλά χαρακτηριστικά που είναι αρκετά δύσκολο ή ίσως και αδύνατο να μετρηθούν με ακρίβεια και εγκυρότητα λόγω έλλειψης μιας κοινώς αποδεκτής αντικειμενικής μέτρησης. Για παράδειγμα, οι κλίμακες / δείκτες χρησιμοποιούνται εκτεταμένα σε επιστημονικούς τομείς, όπως είναι η Οικονομία, η Εκπαίδευση, η Ιατρική, η Βιομετρία, η Βιοστατιστική, η Ψυχομετρία, η Διατροφολογία, η Κοινωνιολογία κ.α. (*Kant, 1996; Mcdowell I, 2002; Jackson, 1970*).

Οι κλίμακες / δείκτες, απλοί ή σύνθετοι, διακριτοί ή συνεχείς, αποσκοπούν στη μέτρηση διαφόρων χαρακτηριστικών, τα οποία, όπως αναφέρθηκε, είναι δύσκολο ή αδύνατο να μετρηθούν ποσοτικά (*Streiner & Norman, 2008; Panagiotakos, 2009*).

Για παράδειγμα, στον τομέα της ψυχομετρίας γίνεται χρήση σύνθετων διακριτών κλιμάκων για τη μέτρηση του άγχους ή της κατάθλιψης (*Zung, 1965; Jackson, 1970*), στον τομέα της διατροφολογίας γίνεται χρήση σύνθετων διακριτών κλιμάκων (π.χ. MedDietScore που εκφράζει το βαθμό προσήλωσης στο πρότυπο της Μεσογειακής διατροφής και συνδέεται με την πρόληψη γενικότερα του καρδιαγγειακού κινδύνου) που ως στόχο έχουν τη μέτρηση της επάρκειας και της ποιότητας της διατροφής και πως συνδέεται με την εμφάνιση ασθενειών (*Panagiotakos et al.; 2007, Kourlaba & Panagiotakos, 2009a*). Επίσης, στο χώρο των επιστήμων υγείας γίνεται χρήση συνεχών κλιμάκων / δεικτών όπως είναι η συνεχής κλίμακα VAS (Visual Analogous Scale) (*Carlsson M., 1983*) που παρέχει μια εκτίμηση χαρακτηριστικών, όπως είναι ο πόνος ενός ασθενούς.

Η χρήση μιας σύνθετης κλίμακας (δείκτη) υγείας προσφέρει τη δυνατότητα έγκαιρης διάγνωσης μιας νόσου με συνέπεια την έγκαιρη εφαρμογή θεραπείας στο άτομο που πιθανόν νοσεί (*McCullough et al, 2000*). Η ορθότητα της διάγνωσης έγκειται στην επάρκεια ανίχνευσης ασθενών και υγιών από το διαγνωστικό έλεγχο / κλίμακα, ή διαφορετικά, στον έγκυρο διαχωρισμό αυτών. Μια κλίμακα υγείας χρησιμοποιείται ως εξής: αν η τιμή της κλίμακας είναι μικρότερη ή μεγαλύτερη από μια δεδομένη τιμή (η τιμή αυτή αποτελεί το διαχωριστικό όριο μεταξύ υγιών και ασθενών, ή αλλιώς το διαγνωστικό κατώφλι), το εκάστοτε άτομο, κατατάσσεται ως υγιές ή μη. Επομένως, για την κάθε κλίμακα, υπάρχει μια τιμή, το διαχωριστικό όριο, που διακριτοποιεί δύο κατηγορίες ατόμων (ασθενείς – μη ασθενείς). Το διαχωριστικό όριο υπολογίζεται με διάφορες μεθόδους όπως είναι π.χ. η μέθοδος μεγιστοποίησης του αθροίσματος ευαισθησίας και ειδικότητας.

Μια σύνθετη κλίμακα T_m, συνήθως, δημιουργείται από το μη σταθμισμένο άθροισμα *m* διακριτών ή συνεχών συνιστωσών μεταβλητών που επιλέγονται κατάλληλα, όπως παρατίθεται ακολούθως:

$$\mathrm{T} = \sum_{j=1}^{m} X_j$$

Το διάνυσμα $\underline{\boldsymbol{X}}=(X_1,X_2, ..., X_m)$ των *m* συνιστωσών μεταβλητών εκφράζει *m* διαφορετικές διαστάσεις της εκπαιδευτικής – οικονομικής - κλινικής κατάστασης ενός ατόμου και το (μη σταθμισμένο) άθροισμα τους παρέχει την τιμή της αντίστοιχης κλίμακας. Για την κατασκευή μιας σύνθετης κλίμακας απαιτείται να επιλεχθούν οι *m* κατάλληλες μεταβλητές που να είναι σχετικές με το χαρακτηριστικό

που στοχεύει να αποτιμά/προβλέπει η κλίμακα. Επίσης, απαιτείται η επιλογή του κατάλληλου αριθμού *k* των διαμερίσεων (αριθμός πιθανών απαντήσεων) για κάθε διακριτή συνιστώσα μεταβλητή ή η επιλογή του κατάλληλου εύρους *l* του πεδίου ορισμού για κάθε συνεχή συνιστώσα μεταβλητή. Επιπλέον, απαραίτητο κρίνεται να γίνεται επιλογή ειδικών σταθμίσεων για την εκάστοτε συνιστώσα μεταβλητή.

Η επιλογή των *m* συνιστωσών μεταβλητών πραγματοποιείται κατά περίπτωση από τον ερευνητή, καθώς δεν υπάρχει θεσπισμένο μεθοδολογικό πλαίσιο, αλλά μια βασική αρχή είναι ότι η σύνθετη κλίμακα θα πρέπει να έχει την ικανότητα να διακρίνει, ή αλλιώς να ταξινομεί, τα υπό εξέταση υποκείμενα ή αντικείμενα σε δύο ή περισσότερες κατηγορίες, ως προς το χαρακτηριστικό (π.χ. κατάθλιψη) που η κλίμακα σκοπεύει να εκτιμήσει (Οι ασθενείς εμφανίζουν το/τα αντίστοιχο/α κλινικό/ά χαρακτηριστικό/ά ενώ οι υγιείς όχι ή σε χαμηλότερο βαθμό ή συχνότητα). Ένα άλλο στοιχείο που δεν εμπίπτει σε συγκεκριμένη μεθοδολογία είναι η εύρεση/επιλογή του αριθμού *k* των διαμερίσεων των διακριτών συνιστωσών μεταβλητών ή του εύρους *l* των συνεχών συνιστωσών μεταβλητών. Η απόδοση σταθμίσεων σε συγκεκριμένες συνιστώσες μιας κλίμακας έχει λογική καθώς, όλες οι συνιστώσες μεταβλητές δεν έχουν τον ίδιο βαθμό συσχέτισης με τη μεταβλητή που εκφράζει την κατάσταση, όμως, λόγω ευκολίας ίσως γίνεται χρήση αστάθμιστων κλιμάκων.

Η παρούσα διατριβή στοχεύει στη διερεύνηση και ανάπτυξη ενός μεθοδολογικού πλαισίου κατασκευής κλιμάκων ως προς βασικά χαρακτηριστικά τους, όπως είναι το πλήθος *m* των συνιστωσών μεταβλητών, το πλήθος *k* των διαμερίσεων των διακριτών συνιστωσών μεταβλητών ή το εύρος *l* του στηρίγματος των συνεχών συνιστωσών μεταβλητών, καθώς επίσης και η χρήση σταθμίσεων *w* στις συνιστώσες μεταβλητές.

1.1 Πληροφορική

Η Πληροφορική ορίζεται ως η θετική και εφαρμοσμένη επιστήμη που σκοπό έχει να εξετάζει τη σχεδίαση, την υλοποίηση, τη βελτιστοποίηση προγραμμάτων που αφορούν την επεξεργασία και ανταλλαγή συμβολικών ή αριθμητικών αναπαραστάσεων ενώ, παράλληλα, ερευνά την κωδικοποίηση, τη διαχείριση και τη μετάδοση πληροφοριών μεταξύ διατάξεων/συστημάτων συλλογής/αποθήκευσης δεδομένων. Η αυτοματοποιημένη εφαρμογή των τεχνικών της πληροφορικής βασίζεται θεμελιωδώς στους ηλεκτρονικούς υπολογιστές (Η/Υ). Συχνά, η Πληροφορική θεωρείται ταυτόσημος επιστημονικός κλάδος με αυτόν της επιστήμης

των υπολογιστών (https://sites.google.com/site/hplhroforikhsthnekpedeysh/ti-einai-e-plerophorike).

Ο όρος «Πληροφορική» (Informatik) πρωτοπαρουσιάστηκε το 1957 από το Γερμανό επιστήμονα Καρλ Στάινμπουχ (Karl Steinbuch) (*Simmons & Simmons, 1961*) και δύο έτη αργότερα, το 1959 ο όρος «Επιστήμη Υπολογιστών» (Computer Science) εμφανίστηκε σε επιστημονικό άρθρο (*Fein, 1959*). Το 1962 επινοήθηκε ο όρος informatique, από τον Γάλλο επιστήμον Φιλίπ Ντρέυφους ως συνδυασμός των λέξεων information (=πληροφορία) και automatic (=αυτοματισμός), προκειμένου να περιγραφεί η επιστήμη της αυτόματης επεξεργασίας πληροφοριών (*Schweitzer, 1999*) και καθιερώθηκε η ορολογία αυτή που χρησιμοποιείται μέχρι σήμερα.

Η Πληροφορική αποτελεί ένα διακριτό επιστημονικό κλάδο και μερικές από τις βασικότερες έννοιες της διδακτικής της Πληροφορικής (*Styliaras, et al., 2015*) περιλαμβάνουν τον διδακτικό μετασχηματισμό των εννοιών (didactic transposition) που διακρίνεται σε Εξωτερικό και Εσωτερικό μετασχηματισμό. Οι αναπαραστάσεις που χρησιμοποιούνται είναι δημιουργίες ανθρωπίνων δραστηριοτήτων και παρουσιάζονται υπό την μορφή συμβόλων έχοντας ως ιδιότητα να πάρουν την θέση άλλων οντοτήτων.

Οι τεχνολογίες της πληροφορικής έχουν διεισδύσει στην καθημερινή ζωή αλλά ίσως πολύ περισσότερο στον επαγγελματικό χώρο, του σύγχρονου ανθρώπου (*Κοίλιας, 2004*). Οι πρώτοι υπολογιστές κατασκευάστηκαν προκειμένου να καταστεί πιο εύκολη και σύντομη η λύση μαθηματικών προβλημάτων, και στις μέρες μας η χρήση τους έχει θέσει τον άνθρωπο ικανό να μοντελοποιήσει πολλά φυσικά φαινόμενα (Μετεωρολογικά φαινόμενα, φαινόμενα κατολισθήσεων, εξελικτική πορεία μικροοργανισμών, κτλ). Όπως αναφέρθηκε, η επίλυση μαθηματικών προβλημάτων ήταν ο πιο σημαντικός λόγος για την κατασκευή των πρώτων υπολογιστών και αυτό εξακολουθεί να ισχύει ακόμα και στις μέρες μας, αφού καθίσταται δυνατή η επεξεργασία μεγάλου όγκου δεδομένων σε σύντομο χρονικό διάστημα. Η εφαρμογή τεχνικών και μεθόδων της Πληροφορικής έχει καθιερωθεί σε όλους σχεδόν τους επιστημονικούς κλάδους, όπως επίσης και στην καθημερινότητα του ανθρώπου καθώς χρησιμοποιείται για εργασίες όπου απαιτούν αποθήκευση και διαχείριση τεράστιων ποσοτήτων δεδομένων, όταν δηλαδή, ο όγκος των πληροφοριών είναι υπερβολικά μεγάλος. Είτε σε κοινωνικό επίπεδο, είτε σε επιχειρηματικό επίπεδο, όταν υπάρχει πληθώρα πληροφοριών που συγκεντρώνονται σε τακτά χρονικά διαστήματα από επικαιροποιημένες αναφορές και από πολλές

διαφορετικές πηγές, απαιτούνται λογισμικά που έχουν την ικανότητα εύκολης αποθήκευσης και εύκολης ανάκλησης τους, όπως είναι οι βάσεις δεδομένων. Αυτού του είδους οι πληροφορίες, σε αρχικό στάδιο αποθηκεύονται σε κατάλληλη βάση δεδομένων και στη συνέχεια η πρόσβασή τους είναι δυνατή οποιαδήποτε χρονική στιγμή με το κατάλληλο ερώτημα (query) στην προαναφερθείσα βάση *(Μαστρακούλης, 2008*). Η χρήση του σύγχρονου ηλεκτρονικού υπολογιστή κρίνεται επιβεβλημένη και αναγκαία για την αποθήκευση αυτού του μεγάλου μεγέθους πληροφοριών.

Στις επόμενες παραγράφους δίδονται παραδείγματα χρήσης των τεχνολογιών της Πληροφορικής. Η χρήση του υπολογιστή είναι αυτονόητη πλέον για επεξεργασία δεδομένων οικονομικής φύσεως (π.χ. Δημόσιοι φορείς όπως οι εφορίες ή ιδιωτικές επιχειρήσεις όπως τα λογιστικά γραφεία παλαιότερα διαχειριζόντουσαν όλες τις πληροφορίες λογιστικού περιεχομένου σε χειρόγραφη μορφή ενώ σήμερα υπάρχουν ειδικές βάσεις δεδομένων που επιτρέπουν υπολογισμούς μεγάλης έκτασης με τα κατάλληλα λογισμικά χωρίς να είναι υπερβολικά χρονοβόρα η επεξεργασία τους) ή για επεξεργασία δεδομένων στο χώρο της βιομηχανίας όπου ένας αριθμός εργασιών που επαναλαμβάνεται σε τακτά χρονικά διαστήματα απαιτεί την καταγραφή και αποθήκευση τους για λόγους ποιότητας των παραγόμενων προϊόντων (*Παπαϊωάννου, 2011*).

Με τη χρήση της Πληροφορικής, ο προγραμματισμός αυτοματοποιημένων διαδικασιών είναι εφικτός (στη βιομηχανία ειδικές ρομποτικές διατάξεις μπορούν να αντικαταστήσουν με επιτυχία την ανθρώπινη παρουσία, ίσως με μεγαλύτερη ακρίβεια και εγκυρότητα) καθώς επίσης και η διάδραση μεταξύ μηχανής και ανθρώπου έχει περάσει σε ένα άλλο επίπεδο καθώς υπάρχει η δυνατότητα με συγκεκριμένα λογισμικά μια υπολογιστική διάταξη, να αυτο-εκπαιδεύεται (Data Mining – Neural Networks). Η ρομποτική είναι μια επιστημονική περιοχή που έχει στενή σχέση με την Πληροφορική και αποτελεί επιστημονικό κλάδο αιχμής που βρίσκει εφαρμογή σε διάφορους τομείς χάρη στην παραγωγική ταχύτητα και ευελιξία που παρέχει, η οποία εξασφαλίζεται από τη δυνατότητα προγραμματισμού μιας ρομποτικής κατασκευής (*Εμίρης κ.α., 2002*).

Σε επίπεδο επικοινωνίας, η υπηρεσία του ηλεκτρονικού ταχυδρομείου είναι ευρέως γνωστή και αποτελεί ένα ασύγχρονο τρόπο επικοινωνίας. Παρέχεται η δυνατότητα αποστολής μηνυμάτων σε πραγματικό χρόνο ανεξάρτητα του τόπου που βρίσκεται ο αποστολέας και ο παραλήπτης. Η μόνη, αλλά ουσιαστική προϋπόθεση

είναι η σύνδεση του υπολογιστή στο διαδίκτυο. Τα τελευταία χρόνια έχουν αναπτυχθεί πλατφόρμες σύγχρονης επικοινωνίας (Skype, viber, κτλ) που τις προτιμούν χρήστες Η/Υ από όλο τον κόσμο. Άλλη εφαρμογή της πληροφορικής αφορά τις υπηρεσίες τηλεφωνίας όπου ένας μεγάλος αριθμός εργασιών εξαρτιόταν από έμψυχο δυναμικό, ενώ, στις μέρες μας, στην εποχή ανάπτυξης της ψηφιακής τεχνολογίας σε κατάλληλα διαμορφωμένα κέντρα τηλεφωνίας με χρήση ειδικών λογισμικών, οι ενέργειες αυτές πραγματοποιούνται ψηφιακά.

Ο κλάδος της υγείας είναι άλλος ένας κλάδος που η εφαρμογή της Πληροφορικής και των υπολογιστών είναι χρήσιμη και απαραίτητη (*Μπαλής, 2007*). Πιο συγκεκριμένα, η εισχώρηση των υπολογιστικών συστημάτων στον κλάδο της υγείας είχε θετικά αποτελέσματα αφού βελτιώθηκαν πολλές λειτουργίες, όπως το σύστημα καταγραφής ασθενών, καθώς, η αρτιότερη διαχείριση του τρόπου αντιμετώπισης των διαφορετικών περιστατικών άλλαξε σε μεγάλο βαθμό κατά την εισαγωγή των πληροφοριακών συστημάτων (*Αγγελίδης, 2015*). Αυτό είχε ως αποτέλεσμα την προσβασιμότητα στα αρχεία του εκάστοτε ασθενούς από το ιατρικό προσωπικό σε οποιαδήποτε χρονική στιγμή και σε οποιοδήποτε τόπο. Για παράδειγμα οι ιατρικές εξετάσεις ενός ασθενούς είναι διαθέσιμες στους ιατρούς σύντομα, σε πραγματικό χρόνο, μέσω ενός e-mail, ενώ μέσω του παραδοσιακού ταχυδρομείου θα ήταν χρονοβόρα διαδικασία. Επομένως, η χρήση των τεχνολογιών της Πληροφορικής και των υπολογιστών γενικότερα δίνει το προβάδισμα στο ιατρικό προσωπικό με σκοπό την ίαση των ασθενών αφού θα μπορούν να λαμβάνουν την κατάλληλη θεραπεία νωρίτερα. Στα συστήματα ελέγχου της θέσης των ασθενοφόρων, η χρήση του υπολογιστή σε συνεργασία με τον κατάλληλο δορυφόρο είναι απαραίτητη βοηθώντας με τον τρόπο αυτό τη σύντομη μετάβασή τους στο σημείο όπου έχουν κληθεί. Προφανώς, η έγκαιρη προσέλευση τους στο χώρο κάποιου συμβάντος έχει ως αποτέλεσμα τη βέλτιστη αντιμετώπιση του περιστατικού.

Η εφαρμογή της Πληροφορικής και των Υπολογιστών στις Επιχειρήσεις είναι απαραίτητη ώστε να είναι ανταγωνιστικές και να αναπτύσσονται. Η ανάπτυξη λογισμικών τύπου CRM (Customer Relationship Management), δηλαδή, πληροφοριακών συστημάτων τα οποία ωφελούν στην ανάπτυξη σχέσεων μεταξύ επιχείρησης και πελάτη είναι επίσης μια αναγκαιότητα στη σύγχρονη εποχή για κάθε επιχείρηση (*Αλεξάνδρου, 2007*). Για παράδειγμα, στις αεροπορικές εταιρείες η χρήση εφαρμογών της Πληροφορικής είχε θετικές εξελίξεις για τις εταιρείες καθώς, μέσω του διαδικτύου ο ενδιαφερόμενος πελάτης είναι σε θέση να διεκπεραιώσει διάφορες

διαδικασίες (να αγοράσει το εισιτήριό του ηλεκτρονικά, να κρατήσει την θέση του, e-check in, κτλ). Σε εταιρικό διοικητικό επίπεδο, νέα συστήματα πληροφορικής, εισάγονται, όπως για παράδειγμα, τα Διοικητικά Συστήματα Πληροφορικής (Management Information Systems - MIS). Οι δυνατότητες τέτοιου είδους λογισμικών είναι η συγκέντρωση δεδομένων, η ασφαλής αποθήκευση δεδομένων, η ανάλυση δεδομένων, η ανάκτηση των δεδομένων σε περιπτώσεις που απαιτείται, η έκδοση οικονομικών στοιχείων και αποτελεσμάτων και η παρακολούθηση της πορείας των πωλήσεων της επιχείρησης. Για τη διαχείριση των προαναφερθέντων λογισμικών και ανάλυση των δεδομένων απαιτούνται στελέχη επιχειρήσεων που να συνδυάζουν γνώσεις πληροφορικής και γνώσεις μάνατζμεντ (*Αντωνίου, 2003*).

Ο χώρος της εκπαίδευσης είναι επίσης ένας κλάδος όπου η Πληροφορική χρησιμοποιείται ευρέως, σε όλες τις βαθμίδες. Η εκπαίδευση που βασίζεται στην Πληροφορική και σε Η/Υ (Computer Based Training, - CBT) είναι ήδη πραγματικότητα και πλήθος διαλέξεων πραγματοποιούνται με χρήση Η/Υ και προβολέων ή διαδραστικών πινάκων. Η εκπαίδευση με χρήση της Πληροφορικής πραγματοποιείται με ειδικά διαμορφωμένα προγράμματα που προσφέρουν την δυνατότητα ο κάθε εκπαιδευόμενος να χρησιμοποιεί το εκπαιδευτικό περιεχόμενο τους που ενδεχομένως να είναι σε μορφή κειμένου, ήχου και γραφημάτων. Για παράδειγμα, στο μάθημα των μαθηματικών γίνεται χρήση λογισμικών όπως είναι το Geogbra (https://www.geogebra.org/?lang=el) που είναι κατάλληλα για την οπτική αναπαράσταση συναρτήσεων και γεωμετρικών εννοιών. Επίσης, τα τελευταία χρόνια έχει αναπτυχθεί η επαίδευση μέσω παιχνιδιών Η/Υ (Bersimis et al, 2017). Η χρήση του Διαδικτύου στο οικιακό περιβάλλον αποτελεί πολύτιμη πηγή πληροφοριών γενικού και ειδικού περιεχομένου για φοιτητές και μαθητές ώστε να συλλέγουν το αναγκαίο υλικό που είναι απαραίτητο για τις υλοποίηση των εργασιών τους.

Βεβαίως, η Πληροφορική δεν είναι σε θέση να βοηθήσει το σύνολο των ανθρώπινων δραστηριοτήτων, όπως για παράδειγμα είναι οι εργασίες στις οποίες απαιτείται δημιουργική σκέψη. Πιο συγκεκριμένα, σε σύνθετες και συνδυαστικές εργασίες η δημιουργική, αλλά επίσης και η κριτική, σκέψη του ανθρώπου δεν μπορεί να αντικατασταθεί πάντοτε από μία μηχανή. Επίσης δεν μπορεί να αναγνωρίσει και να ερμηνεύσει τυχαίες καταστάσεις και σύμφωνα με αυτά να πάρει κάποια σημαντική απόφαση. Για παράδειγμα, η διάταξη των επίπλων σε ένα σπίτι απαιτεί εγκεφαλικές διεργασίες (τακτοποίησης, καλαισθησίας, κτλ) από πλευράς του ανθρώπου που δύσκολα μια υπολογιστική διάταξη είναι σε θέση να αναπαράγει ή η λήψη

αποφάσεων κατά τη διάρκεια πολεμικών επιχειρήσεων. Η εφαρμογή της Πληροφορικής σε συνιστώσες της ανθρώπινης ύπαρξης, όπως είναι λειτουργίες που απαιτούν κοινωνική επαφή επίσης δεν είναι δόκιμη καθώς η άμεση επικοινωνία μεταξύ των ατόμων είναι προτιμότερη σε αντίθεση με την χρήση των υπολογιστών. Για παράδειγμα, ο δάσκαλος προσεγγίζει με συναίσθημα τον μαθητή, ιδιαίτερα των πρώτων τάξεων του δημοτικού, όπου σε τέτοιες περιπτώσεις ο υπολογιστής δεν μπορεί να παίξει το ρόλο ενός ανθρώπου.

1.1.1 Βιοπληροφορική

Τις τελευταίες δεκαετίες η εφαρμογή μεθόδων της Πληροφορικής έγινε πλέον αναντικατάστατο μέρος της ιατροβιολογικής έρευνας και δημιουργήθηκε ο κλάδος της Βιο-πληροφορικής (Bioinformatics - Βιοπληροφορική), ο οποίος είναι ένας ευρύτατος διεπιστημονικός κλάδος που στόχο έχει τη μελέτη και ερμηνεία διάφορων βιολογικών φαινομένων. Εντός του επιστημονικού πεδίου της Βιο-πληροφορικής συνυπάρχουν και συνεργάζονται αρμονικά και αμοιβαία πολλοί επιστημονικοί κλάδοι, από τη Βιολογία μέχρι τα καθαρά μαθηματικά, τις πιθανότητες, τη στατιστική, τις διαφορικές εξισώσεις και τη στοχαστική ανάλυση. Η Βιοπληροφορική, με τη χρήση κατάλληλων στατιστικών πακέτων (SPSS, R, STATA, Minitab, Matlab, κτλ) που ολοένα βελτιώνονται και προσφέρουν νέες τεχνικές παρέχοντας τη δυνατότητα για επεξεργασία μεγάλου όγκου δεδομένων, έχει χρησιμοποιηθεί για υπολογιστικές αναλύσεις βιολογικών ερωτημάτων χρησιμοποιώντας μαθηματικές και στατιστικές τεχνικές. Επομένως, όπως αναφέρθηκε, ως Βιοπληροφορική θα μπορούσε να οριστεί ο κλάδος της επιστήμης που δημιουργήθηκε από το συνδυασμό των επιστημών της Βιολογίας, της Πληροφορικής, της Στατιστικής και των Μαθηματικών, που ασχολείται με την επεξεργασία βιολογικών δεδομένων και προσπαθεί να παράγει γενικότερους νόμους που διέπουν αυτά τα βιολογικά συστήματα. Συνεπώς, γίνεται λόγος για μια βιολογικής κατεύθυνσης επιστήμη ή ειδικότητα, με τις δικές της μεθοδολογίες (*Μπάγκος, 2015*). Η ανάπτυξη της Βιοπληροφορικής και η έντονη επίδραση στις μεταβολές που προκύπτουν στον επιστημονικό χώρο της υγείας και το ενδιαφέρον για επενδύσεις αυξάνεται όλο ένα και περισσότερο σε Ελλάδα και εξωτερικό (*ΣΕΒ, 2012*).

Η Βιοπληροφορική βρίσκει εφαρμογή στον χώρο των ιατροβιολογικών επιστημών αλλά και στο χώρο βιομηχανικής παραγωγής. Πιο συγκεκριμένα, οι νέες τεχνολογίες της Βιοπληροφορικής συμβάλουν στην ικανοποιητική ανάλυση και ερμηνεία της αλληλουχίας και της δομής του DNA. Απώτερος στόχος είναι να πραγματοποιηθεί μοντελοποίηση των πρωτεϊνών και να προσδιοριστούν οι ιδιότητες τους. Αυτό έχει ως αποτέλεσμα, ότι οι νέες τεχνολογίες επιδρούν σε διάφορους τομείς, όπως είναι ο τομέας της φαρμακοβιομηχανίας (Computer Aided Drug Design - CADD). Αυτή η επίδραση είναι αισθητή προς τον άνθρωπο και από οικονομικής σκοπιάς, μέσω της μείωσης του χρόνου και του κόστους παραγωγής. Επίσης, υπάρχουν διαδικασίες που μερικές φορές δεν συμβαδίζουν απόλυτα με τη σύγχρονη εποχή καθώς αποτελούν μια άυλη γέφυρα για το μέλλον αφού παρέχουν την δυνατότητα κατασκευής ενός προτύπου, όπου είναι δυνατόν να υλοποιηθεί η ανάλυση για νέες μεθόδους αντιμετώπισης ασθενειών και δημιουργία ισχυρών φαρμάκων. Έτσι στο χώρο της υγείας, ο ειδικός επιστήμονας αξιοποιώντας ειδικά υπολογιστικά ιατρικά πακέτα, που επιτρέπουν την επεξεργασία μεγάλου όγκου δεδομένων, εξάγει αποτελέσματα που αποσκοπούν στη βελτίωση των παρεχόμενων ιατροφαρμακευτικών υπηρεσιών. Πληθώρα μελετών, από τη διεθνή και την Ελληνική βιβλιογραφία, που έχουν πραγματοποιηθεί τα τελευταία χρόνια εμπίπτουν στον τομέα της Βιοπληροφορικής. Στις περισσότερες σχετικές έρευνες, έντονη βαρύτητα δίνεται στο ερευνητικό πλαίσιο που διέπει τον εκάστοτε επιστημονικό τομέα και σε κάποιες γίνεται προσπάθεια ταύτισης του κλάδου της Βιοπληροφορικής με άλλους κλάδους παρόμοιου αντικειμένου, όπως είναι ο κλάδος της Ιατρικής Πληροφορικής (*Μπάγκος, 2015*).

1.2 Στατιστική

Στην παράγραφο αυτή γίνεται μια προσπάθεια περιγραφής της επιστήμης της στατιστικής και των τεχνικών που περιλαμβάνει με απώτερο στόχο την περιγραφή και ανάλυση δεδομένων. Στατιστική είναι ο επιστημονικός κλάδος που ασχολείται με μεθόδους συλλογής, οργάνωσης, παρουσίασης και ανάλυσης δεδομένων, καθώς επίσης περιλαμβάνει πλήθος τεχνικών και μεθόδων που μετατρέπουν τα απλά δεδομένα σε αξιοποιήσιμη πληροφορία (*Κούτρας, 1998*). Η Στατιστική αποτελεί χρήσιμο εργαλείο σε διάφορους τομείς όπου εφαρμόζεται όπως είναι η Βιολογία, η Ιατρική, η Ψυχολογία, η Εκπαίδευση, η Κοινωνιολογία, η Πληροφορική, το Εμπόριο, η Γεωργία κτλ (*Φράγκος 1998*).

Ιστορικά, ο όρος «Στατιστική» και η έννοιά του πιθανολογείται ότι προέρχεται είτε από την αρχαία Ελλάδα και την αρχαία ελληνική λέξη «στατίζω» όπου η έννοιά της είναι «τοποθετώ, ταξινομώ, συμπεραίνω», είτε από τα λατινικά και τη λατινική λέξη «status» και η μετάφρασή της έχει την έννοια «της πολιτείας» ή του «κράτους» (*Τσαγρής, 2014*). Ο όρος «Στατιστική» σε αρχικό στάδιο χρησιμοποιήθηκε για τις διαδικασίες καταγραφής στοιχείων που αφορούσαν τα δημογραφικά στοιχεία του πληθυσμού ενός κράτους. Η Κίνα το 2238 π.Χ. διεξήγαγε την αρχαιότερη απογραφή πληθυσμού κατόπιν εντολής του αυτοκράτορα Yao, με στοιχεία που δεν αφορούσαν μόνο τον πληθυσμό αλλά και εμπορικά/γεωργικά στοιχεία. Το ίδιο είχε πραγματοποιηθεί νωρίτερα από άλλους λαούς (Αιγύπτιοι, Βαβυλώνιοι, Πέρσες κτλ) χωρίς να υπάρχει συστηματική συλλογή στοιχείων όμως. Μέχρι τα τέλη του 19^{ου} αιώνα η εξέλιξή της επιστήμης της στατιστικής ήταν σχετικά αργή, αλλά, μετά το 19^{ο} αιώνα κάνει αισθητή την παρουσία της ως διακριτός κλάδος των μαθηματικών και στις μέρες μας παρουσιάζει υψηλούς ρυθμούς ανάπτυξης, κυρίως χάρη στην παράλληλη ανάπτυξη της πληροφορικής (*Φράγκος, 1998*). Η επιστήμη της πληροφορικής διαδραματίζει σημαντικό ρόλο στην εξέλιξη της στατιστικής, και αυτή η σχέση είναι αμφίδρομη. Εξειδικευμένα στατιστικά πακέτα/λογισμικά ανάλυσης δεδομένων έχουν αναπτυχθεί που είναι σε θέση να διαχειρίζονται μεγάλο όγκο δεδομένων και να παράγουν ασφαλή και έγκυρα αποτελέσματα, σε διαγραμματική ή πινακοποιημένη μορφή (π.χ. STATGRAPHICS, MICROSTAT, MINITAB, GB-STAT, S-Plus, SAS, Numerical Algorithms Group for Statistics ή NAG, SSP). Τέτοια λογισμικά κυκλοφορούν σε μορφή κλειστού (Statistical Package for the Social Sciences, εν συντομία SPSS) ή ανοικτού κώδικα (R - The R Project for Statistical Computing) (*Κοίλιας, 2004*).

Η επιστήμη της Στατιστικής διακρίνεται σε δύο βασικούς κλάδους (*Χαλικιάς, 2010*):

- **Περιγραφική στατιστική (Descriptive Statistics)**: Περιέχει όλες τις μεθόδους που στοχεύουν στη συλλογή, στην παρουσίαση και στην ταξινόμηση ποσοτικών ή ποιοτικών δεδομένων.
- **Επαγωγική στατιστική (Inferential Statistics)**: Περιέχει όλες τις μεθόδους που στοχεύουν στην εκτίμηση των παραμέτρων ενός πληθυσμού από δειγματικά στοιχεία, δηλαδή τα συλλεχθέντα δεδομένα από ένα δείγμα.

Στο σημείο αυτό πρέπει να δοθούν οι ορισμοί του «Στατιστικού Πληθυσμού» & του «Στατιστικού Δείγματος».

- **Στατιστικός Πληθυσμός ή Πληθυσμός (Population)**: Το σύνολο των προσώπων (έμβιων όντων γενικότερα) ή πραγμάτων (φυσικών αντικειμένων) που παρατηρείται ως προς ένα ή περισσότερα χαρακτηριστικά γνωρίσματά από έναν ερευνητή (*Φράγκος, 2011*).
- **Στατιστικό Δείγμα ή Δείγμα (Sample)**: Το αντιπροσωπευτικό υποσύνολο ενός στατιστικού πληθυσμού (*Ζαχαροπούλου, 2010*).

Ένα δείγμα ή ένας πληθυσμός εξετάζονται ως προς κάποια χαρακτηριστικά τους που ονομάζονται μεταβλητές. Η μεταβλητή ουσιαστικά είναι ένα χαρακτηριστικό των μονάδων του πληθυσμού και λαμβάνει διαφορετικές τιμές για διαφορετικές πληθυσμιακές μονάδες, όπως για παράδειγμα η ηλικία ή το φύλο ενός ατόμου (*Ζαχαροπούλου, 2010*). Τυχαία Μεταβλητή είναι μία μεταβλητή για την οποία προσδιορίζεται πλήρως μία μαθηματική έκφραση που ονομάζεται συνάρτηση πιθανότητας, η οποία παρέχει τη σχετική συχνότητα με την οποία εμφανίζονται οι τιμές της μεταβλητής και είναι καλώς ορισμένη λαμβάνοντας θετικές τιμές. Οι τυχαίες μεταβλητές συνήθως συμβολίζονται με κεφαλαία γράμματα (πχ X, Y) και οι τιμές, που λαμβάνουν σε ένα δείγμα ή ένα πληθυσμό, συμβολίζονται με μικρά γράμματα (π.χ. x, y) (*Φράγκος, 2011*). Οι τυχαίες μεταβλητές προέρχονται από το τυχαίο δείγμα το οποίο είναι το δείγμα εκείνο όπου έχει εκλεγεί με τέτοιο τρόπο ώστε τα αποτελέσματα της επαγωγικής ανάλυσής του να μπορούν να χρησιμοποιηθούν για να εξάγουν στατιστικά συμπεράσματα για τον πληθυσμό από τον οποίο το δείγμα έχει συλλεχθεί (*Φράγκος, 2011*).

Οι μεταβλητές διακρίνονται σε δύο βασικές κατηγορίες, τις ποσοτικές και τις ποιοτικές. Οι ποσοτικές μεταβλητές λαμβάνουν αριθμητικές τιμές σε συνεχές ή διακριτό σύνολο τιμών. Παραδείγματα ποσοτικών μεταβλητών είναι το ύψος φοιτητών, το βάρος μαθητών, η θερμοκρασία κατά τη διάρκεια ενός μήνα κτλ (*Φράγκος, 2011*). Οι ποσοτικές μεταβλητές διακρίνονται σε δύο βασικές κατηγορίες ως προς το είδος του στηρίγματος των μεταβλητών τους. Πιο συγκεκριμένα υπάρχουν οι διακριτές και οι συνεχείς μεταβλητές.

- **Διακριτές Μεταβλητές**: είναι οι μεταβλητές που λαμβάνουν μόνο ορισμένες τιμές σε ένα διάστημα, όπως είναι ο αριθμός των παιδιών μια οικογένειας, ο αριθμός των πελατών μιας τράπεζας κτλ.

- **Συνεχείς Μεταβλητές**: είναι οι μεταβλητές: που λαμβάνουν όλες τις τιμές μέσα σε ένα διάστημα, όπως είναι το ύψος ενός ενήλικα, το βάρος ενός παιδιού κτλ,.

Εκτός των μεθόδων της στατιστικής αυτής καθεαυτής, σημαντικό ρόλο έχουν οι τεχνικές συλλογής των στοιχείων από τα αντίστοιχα δείγματα. Αυτές οι μέθοδοι εμπίπτουν στο ευρύτερο πεδίο της Δειγματοληψίας, που είναι η επιστήμη η οποία ασχολείται με τις τεχνικές συλλογής στοιχείων από γενικούς πληθυσμούς (*Δαμιανού, 1994*). Η Δειγματοληψία είναι η διαδικασία λήψεως ενός δείγματος, δηλαδή ενός μέρους του συνόλου των πληροφοριών που μας ενδιαφέρουν. Αυτό πραγματοποιείται με διάφορες μεθόδους, όπως είναι η απλή τυχαία δειγματοληψία, η στρωματοποιημένη δειγματοληψία, η δειγματοληψία κατά ποσοστά, κτλ (*Δαμιανού, 1994*). Η συλλογή παρατηρήσεων από όλα τα στοιχεία ενός πεπερασμένου πληθυσμού ονομάζεται απογραφή. Στην πράξη, η απογραφή εφαρμόζεται σπάνια (π.χ. απογραφή του πληθυσμού κάθε δέκα χρόνια) ενώ η δειγματοληψία πολύ πιο συχνά. Αυτό συμβαίνει διότι:

1. Ο πληθυσμός πιθανόν να μην είναι πεπερασμένος, οπότε είναι ανέφικτη η λήψη του συνόλου των πληθυσμιακών μονάδων.
2. Το μεγάλο χρονικό διάστημα το οποίο απαιτείται για την συλλογή και την επεξεργασία των παρατηρήσεων κάποιου μεγάλου πληθυσμού έχει ως αποτέλεσμα τη συντόμευση του χρόνου διενέργειας της έρευνας με χρήση δείγματος.
3. Τα σφάλματα μέτρησης και επεξεργασίας εντείνονται όσο το πλήθος τους αυξάνει, κατά συνέπεια, η προστιθέμενη πληροφορία πιθανόν να μην είναι ωφέλιμη αλλά παραπλανητική.
4. Δεν είναι εφικτό να γίνει απογράφη όταν οι παρατηρήσεις που συλλέγονται πρέπει να καταστρέφονται με κάποιο τρόπο. Για παράδειγμα η εξέταση της αντοχής μεταλλικών καρφιών κάποιας βιομηχανίας, συνεπάγεται την καταστροφή τους, οπότε αναγκαστικά περιοριζόμαστε στην εξέταση της αντοχής ενός δείγματος.
5. Το κόστος της παρατήρησης επί πλέον μονάδων είναι πολύ μεγάλο σχέση με τα διαθέσιμα μέσα ή τα αναμενόμενα πλεονεκτήματα. Αυτό σημαίνει ότι η δειγματοληψία μας επιτρέπει την συλλογή πληροφοριών γρήγορα και σε μικρό κόστος. Μπορεί να θεωρηθεί προφανές ότι η γενίκευση των πληροφοριών του δείγματος σε ολόκληρο τον πληθυσμό θα έχει ως συνέπεια την δημιουργία κάποιου σφάλματος το οποίο ελαχιστοποιείται όταν το δείγμα είναι

αντιπροσωπευτικό. Αντιπροσωπευτικό είναι το δείγμα εκείνο που αποτελεί μια μικρογραφία του αντίστοιχου πληθυσμού.

Όταν η δειγματοληψία πραγματοποιείται υπό ορισμένες συνθήκες πετυχαίνει ότι ορισμένα μετρήσιμα περιθώρια σφάλματος φθίνουν. Στα πλαίσια της επαγωγικής στατιστικής πραγματοποιούνται εκτιμήσεις των παραμέτρων διάφορων πληθυσμών με σημειακή εκτίμηση ή εκτίμηση διαστήματος. Η εκτίμηση διαστήματος περιλαμβάνει ένα σύνολο, το οποίο προκύπτει από τη γνώση του δείγματος και με συγκεκριμένη βεβαιότητα μπορεί να υποτεθεί ότι η αληθινή τιμή της παραμέτρου του πληθυσμού είναι εντός αυτού του διαστήματος. Επίσης, πραγματοποιούνται έλεγχοι υποθέσεων, παραμετρικοί ή μη παραμετρικοί, για τη διερεύνηση της τιμής μιας παραμέτρου του πληθυσμού με στοιχεία από ένα δείγμα. Παραμετρικοί έλεγχοι (π.χ. z-test, t-test, F-test) πραγματοποιούνται όταν τα διαθέσιμα δεδομένα ακολουθούν την κανονική κατανομή, ενώ οι μη παραμετρικοί έλεγχου (π.χ. Mann Whitney U test, Wilcoxon test) πραγματοποιούνται όταν τα διαθέσιμα δεδομένα δεν ακολουθούν την κανονική κατανομή. Οι έλεγχοι εφαρμόζονται κατάλληλα στην περίπτωση ενός δείγματος, δύο ή περισσοτέρων δειγμάτων. Στην περίπτωση που το μέγεθος ενός δείγματος είναι μεγαλύτερο του 30, μπορεί να γίνει η υπόθεση της κανονικής κατανομής των δεδομένων και να γίνει χρήση των παραμετρικών ελέγχων, που είναι ισχυρότεροι των μη παραμετρικών.

1.2.1 Βιοστατιστική

Η Βιοστατιστική είναι ο εξειδικευμένος επιστημονικός κλάδος της Στατιστικής που ασχολείται με την εφαρμογή στατιστικών μεθόδων, όπως η διαχείριση και ανάλυση αριθμητικών δεδομένων, στον ευρύτερο χώρο της ιατρικής και της βιολογικής έρευνας. Η συνεισφορά της Βιοστατιστικής συνίσταται στον ορθό σχεδιασμό μελετών ή κλινικών δοκιμών και στην ασφαλή εξαγωγή συμπερασμάτων για διάφορους πληθυσμούς με χρήση τεχνικών από την ευρύτερη Επαγωγική Στατιστική σε κατάλληλα επιλεγμένα δείγματα. Πολλές θετικές εξελίξεις στο χώρο της Ιατρικής πραγματοποιήθηκαν κατόπιν της συλλογής δεδομένων με ειδικές μεθόδους δειγματοληψίας, επεξεργασίας τους με χρήση κατάλληλων ελέγχων υποθέσεων και επαρκούς ερμηνείας των παραγόμενων αποτελεσμάτων, δηλαδή με μεθόδους που περιλαμβάνονται στο πεδίο της Βιοστατιστικής. Ιστορικά, εδώ και αιώνες η συλλογή και σωστή ερμηνεία των ποσοτικών κυρίως στοιχείων μέσω της

στατιστικής οδήγησε την Ιατρική σε μια συνεχή εξέλιξη. Συγκεκριμένα, τις τελευταίες δεκαετίες η εφαρμογή των στατιστικών μεθόδων έγινε πλέον αναπόσπαστο κομμάτι της ιατρικής έρευνας λαμβάνοντας υπόψη και τη συμβολή των στατιστικών πακέτων που συνεχώς προσφέρουν νέες δυνατότητες. Συνεπώς, η επιστήμη της Βιοστατιστικής μπορεί να καθοδηγήσει τους ερευνητές στην λήψη αποφάσεων όσον αφορά την προώθηση φαρμάκων, ιατρικών μεθόδων και θεραπειών. Οι μεταβλητές που μελετώνται στα πλαίσια της Βιοστατιστικής είναι ποσοτικές (διακριτές ή συνεχείς), ιεραρχικές και κατηγορικές. Οι γραφικές μέθοδοι που χρησιμοποιούνται για την περιγραφή ποιοτικών δεδομένων είναι τα κυκλικά διαγράμματα και τα ραβδογράμματα, και για τα ποσοτικά δεδομένα τα ιστογράμματα, τα θηκογράμματα και τα στικτά διαγράμματα (τα στικτά διαγράμματα αναπαριστούν γραφικά τη συμμεταβολή 2 ή περισσοτέρων μεταβλητών). Οι αριθμητικές μέθοδοι που χρησιμοποιούνται για την περιγραφή των ποιοτικών δεδομένων είναι οι πίνακες απόλυτων και σχετικών συχνοτήτων (ποσοστά), και για τα ποσοτικά δεδομένα τα μέτρα θέσης, διασποράς και μορφής. Άλλα εργαλεία της περιγραφικής έρευνας στον χώρο της υγείας είναι οι αναλογίες (π.χ. αναλογία υγιών και ασθενών), οι δείκτες (π.χ. δείκτης καρδιαγγειακού κινδύνου) και οι ρυθμοί μεταβολής (π.χ. ρυθμός επιπολασμού ασθένειας) (*Stavrinos & Panagiotakos, 2007*). Η αναλογία σε ένα δείγμα ορίζεται ως το πηλίκο του αριθμού των μελών του δείγματος με ένα χαρακτηριστικό προς το μέγεθος ν του δείγματος, ενώ στον πληθυσμό είναι ο αριθμός των μελών πληθυσμού με το χαρακτηριστικό προς το μέγεθος Ν του πληθυσμού. Ο επιπολασμός ενός χαρακτηριστικού στο δείγμα ή τον πληθυσμό είναι μια αναλογία η οποία εκφράζει τον αριθμό περιπτώσεων με το χαρακτηριστικό αυτό προς τον αριθμό των μελών του δείγματος ή του πληθυσμού την ίδια χρονική στιγμή. Συχνά το χαρακτηριστικό αυτό αντιπροσωπεύει κάποια νόσο. Η επίπτωση ενός χαρακτηριστικού στο δείγμα ή τον πληθυσμό είναι μια αναλογία η οποία εκφράζει τον αριθμό των νέων περιπτώσεων με το χαρακτηριστικό μέσα σε μια συγκεκριμένη χρονική περίοδο προς τον αριθμό μελών δείγματος ή πληθυσμού την ίδια χρονική περίοδο. (*Stavrinos & Panagiotakos, 2007*). Επιπρόσθετα, η επίπτωση μπορεί να εκφραστεί και ως ο κίνδυνος (risk) να υπάρξει το χαρακτηριστικό. Ο επιπολασμός του θανάτου σε ένα δείγμα ή πληθυσμό όταν δεν υπάρχει νόσος καλείται θνησιμότητα, ενώ όταν υπάρχει νόσος λέγεται θνητότητα. Ο ρυθμός μεταβολής αφορά την μέτρηση, δηλαδή τον επιπολσμό, ενός χαρακτηριστικού σε τουλάχιστον δύο χρονικές στιγμές και ορίζεται ως:

Επιπολασμός 2 – Επιπολασμός 1

Ο σχετικός ρυθμός μεταβολής ορίζεται ως ο ρυθμός μεταβολής προς τον επιπολασμό 1:

(Επιπολασμός 2 – Επιπολασμός 1)/ (Επιπολασμός 1)

Ένα άλλο σημαντικό μέρος της Βιοστατιστικής είναι οι διαγνωστικοί έλεγχοι (screening tests), οι οποίοι, είναι διάφοροι ιατρικοί έλεγχοι οι οποίοι αποσκοπούν στον εντοπισμό και την πρόληψη ασθενειών. Τα διαγνωστικά αυτά τεστ είναι πολύ σημαντικό να διακρίνονται από υψηλό ποσοστό επιτυχίας σε κάποια βασικά χαρακτηριστικά τους μεταξύ των οποίων είναι και η ευαισθησία (ποσοστό αληθώς θετικών αποτελεσμάτων) και η ειδικότητα (ποσοστό αληθώς αρνητικών αποτελεσμάτων). Η ευαισθησία και η ειδικότητα (όπως και οι καμπύλες ROC που αναφέρονται σε επόμενη παράγραφο) είναι δείκτες αξιολόγησης διαγνωστικών μεθόδων. Όταν ο διαγνωστικός έλεγχος είναι θετικός ή αρνητικός υπάρχει θετική ή αρνητική δήλωση αντίστοιχα, λόγω του αποτελέσματος του διαγνωστικού ελέγχου. Όταν η νόσος είναι θετική τότε υφίσταται η ασθένεια, δηλαδή, παρουσία της νόσου, ενώ στην περίπτωση που είναι αρνητική τότε δεν υφίσταται ασθένεια, δηλαδή, απουσία της νόσου.

Ευαισθησία (sensitivity) ενός διαγνωστικού ελέγχου είναι η πιθανότητα ο έλεγχος να παράγει θετικό αποτέλεσμα για κάποιο άτομο δεδομένου ότι αυτό το άτομο νοσεί. Η ευαισθησία δηλαδή εκφράζει την ικανότητα του τεστ να ανιχνεύει τους πραγματικά νοσούντες. Έτσι, ορίζεται το ποσοστό:

Ευαισθησία = P(Θετικός Έλεγχος|Ύπαρξη Νόσου)

Ικανοποιητική ευαισθησία, εμπειρικά άνω του 70%, σημαίνει ότι ο διαγνωστικός έλεγχος έχει αξιόπιστη ικανότητα στην διάκριση νοσούντων και μη νοσούντων. Ειδικότητα (specificity) ενός διαγνωστικού ελέγχου είναι η πιθανότητα το τεστ να είναι αρνητικό δεδομένου ότι κάποιος δεν νοσεί. Είναι, δηλαδή, η ικανότητα του τεστ να ανιχνεύει αυτούς που πραγματικά δεν φέρουν την νόσο. Έτσι, ορίζεται το ποσοστό:

Ειδικότητα = P(Αρνητικός Έλεγχος| Απουσία Νόσου)

Ικανοποιητική ειδικότητα, εμπειρικά άνω του 70%, σημαίνει ότι ο διαγνωστικός έλεγχος έχει αξιόπιστη ικανότητα στην διάκριση νοσούντων και μη νοσούντων. Όταν αναφερόμαστε στον γενικό πληθυσμό επιδιώκουμε την αύξηση της ευαισθησίας καθώς είναι πολύ σημαντικό να μην υπάρχουν απώλειες πραγματικών περιστατικών της εκάστοτε νόσου, ειδικά όταν η νόσος είναι σοβαρή. Από την άλλη

πλευρά, όταν ο διαγνωστικός έλεγχος εφαρμόζεται σε επιλεγμένο πληθυσμό, τότε στόχος μας είναι η αύξηση της ειδικότητας. Στην περίπτωση αυτή μας ενδιαφέρει να μην έχουμε απώλειες υγιών ώστε τα άτομα αυτά να αποφεύγουν άσκοπες και επίπονες περαιτέρω διαδικασίες. Για τις πιθανότητες εσφαλμένης διάγνωσης ισχύει:

P(Ψευδώς θετικού αποτελέσματος)=P(false positive) = 1-ειδικότητα

P(Ψευδώς αρνητικού αποτελέσματος)=P(false negative) = 1-ευαισθησία

Αξίζει να σημειωθεί ότι, η ευαισθησία και η ειδικότητα αντιστοιχούν στα σφάλματα τύπου Ι και ΙΙ, αντίστοιχα, των στατιστικών ελέγχων υποθέσεων θεωρώντας ως μηδενική υπόθεση την ύπαρξη ασθένειας. Δηλαδή,

Ευαισθησία = 1- P(σφάλμα τύπου Ι) = 1-α

Ειδικότητα = 1- P(σφάλμα τύπου ΙΙ) = 1-β

Η ευαισθησία και η ειδικότητα δεν είναι εφικτό να μεγιστοποιηθούν ταυτόχρονα λόγω του ότι τα α και β δεν μπορούν να ελαχιστοποιηθούν ταυτόχρονα. Η επιδημιολογία είναι άλλος ένας επιστημονικός κλάδος που βασίζεται σε μεθόδους της Βιοστατιστικής για την επιστημονική μελέτη των παραγόντων που επηρεάζουν την συχνότητα εμφάνισης ασθενειών σε έναν πληθυσμό, μέσω της εφαρμογής της στατιστικής στην ιατρική (*Stavrinos & Panagiotakos, 2007*). Η στατιστική έχει σημαντικό ρόλο σε όλα τα στάδια των κλινικών δοκιμών και στην περιγραφή των κατανομών ασθενειών στον ευρύτερο πληθυσμό. Επιπλέον, θεμελιώδης είναι και η συμβολή της στην διαμόρφωση και διαχείριση των κρατικών βάσεων δεδομένων ιατρικής φύσης που οδηγούν στον ανάλογο προγραμματισμό ζητημάτων υγείας.

1.3 Κλίμακες (Δείκτες)

Η κατανόηση, η βαθμολόγηση και ο συνδυασμός σε ένα ενιαίο εργαλείο ποικίλων κλινικών, βιολογικών ή συμπεριφοριστικών ανθρώπινων χαρακτηριστικών είναι μια σημαντική διαδικασία για την θέσπιση αποτελεσματικών στρατηγικών πρόληψης σε διάφορους τομείς, όπως η εκπαίδευση, η οικονομία, ο καρδιαγγειακός και ο καρκινικός κίνδυνος, διάφορες μεταβολικές διαταραχές, υποσιτισμός, κίνδυνος βρεφικής θνησιμότητας, κλπ. (*Mcdowell, 2002*). Για παράδειγμα, στις επιστήμες της υγείας υπάρχει μεγάλος αριθμός κλινικών χαρακτηριστικών που μπορούν να μετρηθούν με ακρίβεια χρησιμοποιώντας τα κατάλληλα εργαλεία μέτρησης, όπως είναι οι αιματολογικοί και οι βιοχημικοί δείκτες. Παρ'όλα αυτά, υπάρχει μια ποικιλία κλινικών καταστάσεων που δεν μπορούν να μετρηθούν με ακρίβεια, όπως για παράδειγμα η κατάθλιψη, το άγχος, η αίσθηση του πόνου και η ποιότητα των

διατροφικών συνηθειών (*Hamilton, 1960; Hawker et al, 2011; Harnack et al., 2002*). Συνεπώς, υπάρχει ανάγκη για εύρεση κατάλληλων εργαλείων προκειμένου να ποσοτικοποιηθούν χαρακτηριστικά, όπως τα προαναφερθέντα, που είναι αρκετά δύσκολο να μετρηθούν. Για το σκοπό αυτό, στον τομέα διαφόρων επιστημών, τα τελευταία χρόνια, έχουν προταθεί εργαλεία που ονομάζονται «κλίμακες» ή «δείκτες» και ταξινομούνται ανάλογα με το θέμα τους, το αντικείμενο τους, το σκοπό τους και το σχεδιασμό τους (*McDowell, 2006*).

Οι κλίμακες/δείκτες υγείας χρησιμοποιούνται ευρέως στη βιοιατρική έρευνα (*Kant, 1996, Streiner & Norman, 2008*) μετατρέποντας την πληροφορία από πλήθος διαφορετικών χαρακτηριστικών σε ένα, μονόμετρο συνήθως, μέγεθος, που αποτιμά ολιστικά την κλινική κατάσταση ενός ατόμου. Πιο συγκεκριμένα, οι κλίμακες που σχετίζονται με την υγεία χρησιμοποιούνται ευρέως σε ποικίλες επιστήμες, όπως η κλινική ιατρική, η ψυχομετρία, η ψυχολογία, η διατροφική αξιολόγηση, η κοινωνιολογία κλπ. (*Kant, 1996, Mcdowell, 2002, Jackson, 1970*). Οι περισσότερες από αυτές τις κλίμακες είναι σύνθετα εργαλεία, δηλαδή συνδυασμός μεμονωμένων χαρακτηριστικών, τα οποία έχουν σχεδιαστεί για τη μέτρηση των διαφορετικών κλινικών και συμπεριφοριστικών (λανθανόντων) χαρακτηριστικών, τα οποία είναι διφορούμενα ή σε ορισμένες περιπτώσεις ακόμη και αδύνατό να προσδιορίζονται άμεσα (*Streiner & Norman, 2008; D. Panagiotakos, 2009; Zung, 1965; Kourlaba & Panagiotakos, 2009a; Panagiotakos et al, 2007; Carlsson, 1983; Hawker, Mian et al, 2011*). Οι σύνθετες κλίμακες που σχετίζονται με την υγεία μπορεί να προέρχονται από λανθάνοντες συνεχείς μεταβλητές, με τη χρήση διάφορων σημείων αποκοπής, όπως η διάμεσος, συγκεκριμένα εκατοστημόρια σχετικά με τον επιπολασμό μιας ασθένειας, σημεία που είναι προικισμένα με ειδικές ιδιότητες και προτείνονται από διεθνείς οργανισμούς υγείας ή το βέλτιστο διαχωριστικό όριο που προκύπτει από το δείκτη του Youden από την καμπύλη ROC (*Youden, 1950*). Συνήθως, οι κλίμακες υγείας ταξινομούν τα υπό μελέτη άτομα σε δύο ή περισσότερες κατηγορίες, σύμφωνα με το χαρακτηριστικό που σκοπεύει να ανακαλύψει/προβλέψει/αποτιμήσει η κλίμακα.

Μια κλίμακα T μπορεί να αντιπροσωπεύεται από μία μοναδική μεταβλητή (T_1=X_1) η οποία δύναται να μετράει, είτε μια εκπαιδευτική παράμετρο, είτε μια οικονομική κατάσταση, είτε μια συγκεκριμένη βιοχημική ποσότητα, είτε την αναλογία ενός συγκεκριμένου συμπτώματος μιας ασθένειας (π.χ. η γλυκόζη του αίματος ή ο ρυθμός κρίσεων πανικού), είτε την ένταση του πόνου που αισθάνεται ένας ασθενής, αλλά μπορεί επίσης να συντίθεται από το άθροισμα m συνιστωσών

μεταβλητών, όπου κάθε μια από αυτές τις μεταβλητές X_i, $i=1, 2, ..., m$ αντανακλά μια διαφορετική πτυχή του εκπαιδευτικού επιπέδου ενός οργανισμού, της οικονομικής κατάστασης μιας εταιρείας ή της κλινικής κατάστασης ενός ατόμου. Αυτές οι m συνιστώσες είναι αθροισμένες, συνήθως χωρίς στάθμιση, για να παρέχουν μια συνολική βαθμολογία που περιγράφει καλύτερα κάποιο δύσκολα μετρήσιμο χαρακτηριστικό. Οι μεταβλητές X_i, $i=1, 2, ..., m$ μπορούν να είναι είτε διακριτές είτε συνεχείς.

Όπως αναφέρθηκε, μια σύνθετη κλίμακα, που συμβολίζεται ως T_m, παράγεται συνήθως από το μη σταθμισμένο άθροισμα των m συνιστωσών μεταβλητών που επιλέγονται κατάλληλα από τους ερευνητές, όπως παρατίθεται στη σχέση (1.3.1):

$$T_m = X_1 + X_2 + ... + X_m \qquad (1.3.1)$$

Σύμφωνα με τις τιμές των προαναφερθέντων κλιμάκων, οι αντίστοιχες οντότητες που εξετάζονται ταξινομούνται σε κάποια κατηγορία, ως προς το διαχωριστικό όριο για ένα συγκεκριμένο χαρακτηριστικό (*Mcdowell, 2002*). Ακρογωνιαίο λίθο αποτελεί το ζήτημα της εγκυρότητας, της ευαισθησίας, της ειδικότητας και της ακρίβειας αυτών των εργαλείων, των κλίμακων, ειδικά όταν σχετίζονται με ανθρώπινα χαρακτηριστικά.

1.3.1 Χρησιμότητα Κλιμάκων – Παραδείγματα Κλιμάκων

Για την περίπτωση των κλινικών χαρακτηριστικών που δεν μετριούνται άμεσα, απαιτείται ένα εργαλείο μέτρησης που θα είναι σε θέση να ποσοτικοποιήσει αυτά τα μη αντικειμενικώς ποσοτικά χαρακτηριστικά. Όπως αναφέρθηκε, τέτοια εργαλεία μέτρησης είναι οι κλίμακες και η χρησιμότητα τους είναι ιδιαίτερης σημασίας, καθώς, κατά τη διάρκεια των τελευταίων ετών, οι κλίμακες έχουν χρησιμοποιηθεί εκτενώς σε διάφορους τομείς όπως είναι η εκπαίδευση, η οικονομία, η ιατρική, η ψυχομετρία, η διατροφολογία, η βιομετρία, κτλ προκειμένου να μετρηθούν διάφορα παθολογικά χαρακτηριστικά ή μη, όπως είναι το εκπαιδευτικό επίπεδο, η οικονομική κατάσταση, η κατάθλιψη, το άγχος (*Jackson, 1970; Kant, 1996*), καθώς επίσης χρησιμοποιούνται και στην πρόληψη καρδιαγγειακού κινδύνου προκειμένου να μετρηθούν η ποιότητα και η επάρκεια της διατροφής (*Kourlaba & Panagiotakos, 2009*), μέσω συγκεκριμένων διατροφικών προτύπων. Με βάση αυτά τα εργαλεία, τις κλίμακες, τα υπό διερεύνηση άτομα ταξινομούνται ως σχετιζόμενα ή μη με κάποια

χαρακτηριστικά. Στις ακόλουθες παραγράφους δίνονται παραδείγματα τέτοιων κλιμάκων από διάφορα επιστημονικά πεδία.

1.3.1.1 Κλίμακες στην Εκπαίδευση

Στη σύγχρονη κοινωνία των γοργών οικονομικών, πολιτισμικών και τεχνολογικών εξελίξεων, στόχο αποτελεί η συνεχής και αδιάκοπη βελτίωση του εκπαιδευτικού έργου των εκπαιδευτικών μονάδων (Πρωτοβάθμια, Δευτεροβάθμια και Τριτόβάθμια) θέτοντας κατάλληλες πολιτικές στον τομέα αυτό. Η κύρια ευθύνη ανάληψης αποφάσεων με σκοπό τη βελτίωση ενός εκπαιδευτικού οργανισμού βαραίνει την ίδια τη διοίκηση αυτού καθώς και όλων των εμπλεκόμενων στις διεργασίες του εκπαιδευτικού έργου διότι όλοι οι προαναφερθέντες είναι άμεσα ενήμεροι για τα θετικά και αρνητικά της εκπαιδευτικής κοινότητας. Για να παρθούν οι κατάλληλες αποφάσεις σε κεντρικό επίπεδο με στόχο τη βελτίωση του εκπαιδευτικού έργου των προαναφερθέντων μονάδων, απαραίτητη είναι η αξιολόγηση και η αποτίμηση του εκπαιδευτικού έργου στο παρελθόν, ώστε να πραγματοποιηθούν τα κατάλληλα βήματα για την αναδιάρθρωση πιθανόν του συνόλου των εκπαιδευτικών οργανισμών. Τον σκοπό της αποτίμησης του εκπαιδευτικού έργου, εξυπηρετεί η εσωτερική (από όργανα που ανήκουν στον εκπαιδευτικό οργανισμό) ή εξωτερική αξιολόγηση (από όργανα που δεν ανήκουν στον εκπαιδευτικό οργανισμό) βασιζόμενη όχι σε υποκειμενικά κριτήρια κάποιων προσώπων αλλά στηρίζεται σε επιστημονικές μεθόδους και τεχνικές που εφαρμόζονται σε αναπτυγμένες χώρες. Συγκεκριμένα, η αξιολόγηση των εκπαιδευτικών μονάδων πρέπει να στοχεύει στην βελτίωση τους μέσα από ένα αυστηρό πλαίσιο διαδικασιών που περιλαμβάνουν αντικειμενικά κριτήρια αποτίμησης και αξιολόγησης τους. Η αξιολόγηση των εκπαιδευτικών μονάδων έχει ευρεία έννοια καθώς δεν περιορίζεται στην αξιολόγηση των σπουδαστών και των εκπαιδευτικών. Η αξιολόγηση ενός εκπαιδευτικού οργανισμού πρέπει να περιλαμβάνει πολλές διαφορετικές πτυχές και διαστάσεις αυτής και να παράγεται ένα συνολικό αποτέλεσμα που αποτελεί τη συνισταμένη όλων των παραπάνω. Οι επιμερους δείκτες ποιότητας του εκπαιδευτικού έργου που εκφράζουν ποικίλες διαστάσεις αυτού καθορίζονται με συστηματικό τρόπο ακολούθως:

- Η ποιότητα του παρεχόμενου εκπαιδευτικού έργου εκπαιδευτικών οργανισμών
- Οι μέθοδοι βελτίωσης της ποιότητας του εκπαιδευτικού έργου
- Η πρόοδος που επιτεύχθηκε από την τελευταία ακιολόγηση

Ο ρόλος των κλιμάκων / δεικτών ποιότητας είναι ουσιαστικός στην διαδικασία αξιολόγησης και αποτελούν προϊόν συλλογικής εργασίας και κριτικής ανατροφοδότησης με καταγραφή και χρήση κατάλληλων δεδομένων. Η επιλογή και η διαμόρφωση των κλιμάκων / δεικτών αξιολόγησης είναι «ένας τρόπος που έχει σκοπό την οργάνωση και διαχείριση της εκπαιδευτικής λειτουργίας» (*Παιδαγωγικό Ινστιτούτο,* 2000). Η διαμόρφωση των κλιμάκων / δεικτών αξιολόγησης της ποιότητας του εκπαιδευτικού έργου έχει σκοπό την αποτελεσματικότερη διαχείριση των μηχανισμών προόδου και ενδυνάμωσής τους σε σκοπό τη βελτίωση της ποιότητας στην εκπαιδευτική μονάδα. Η αξιολόγηση του εκπαιδευτικού έργου συνίσταται σε 5 κλίμακες που εκφράζουν 5 διακριτές θεματικές ενότητες που παρατίθενται ακολούθως και αναλύονται περεταίρω σε 20 μεταβλητές/χαρακτηριστικά ποιότητας:

- Διαθέσιμοι πόροι, αξιοποίηση και διαχείριση: Σε αυτό τον θεματικό άξονα περιλαμβάνονται:
    - Η ποιότητα και η τεχνική επάρκεια των κτιριακών υποδομών.
    - Η ποιότητα και η επάρκεια του οπτικοακουστικού και τεχνολογικού εξοπλισμού.
    - Οι ανθρώπινοι πόροι όπως είναι το διδακτικό προσωπικό, ο αριθμός αυτού, το ωράριο, οι σπουδές και η επιμόρφωση τους, το βοηθητικό προσωπικό κ.ά.
    - Οι οικονομικοί πόροι (τακτές και έκτατες επιχορηγήσεις από το τοπικό και το περιφερειακό περιβάλλον).
    - Το πρόγραμμα σπουδών, τα διδακτικά συγγράμματα, το εποπτικό υλικό, οι οργανωτικές δομές, το οργανόγραμμα.

- Διοίκηση (διοικητική οργάνωση σε επίπεδο εκπαιδευτικού οργανισμού αλλά και σε επίπεδο τάξης): Σε αυτό τον θεματικό άξονα περιλαμβάνονται:
    - Η Διοίκηση και η οικονομική διαχείριση που εκφράζονται μέσα από μέθοδους εργασίας διοικητικών οργάνων, τη φύση και την ποικιλία των διοικητικών υπηρεσιών, την κατανομή και τη χρήση οικονομικών πόρων, την υγιεινή και την κοινωνική φροντίδα κ.ά.).

- Η οργάνωση και η διοίκηση της τάξης μέσα από κοινωνικο-εκπαιδευτικές και αθλητικές δραστηριότητες, πολιτιστικά δρώμενα, κτλ.
- Η δυναμική της οργάνωσης μέσα από τον εκσυγχρονισμό, τις αλλαγές, τις προοπτικές, κτλ.
- Η δόμηση και η εφαρμογή του εκπαιδευτικού προγράμματος.
- Η αξιοποίηση μέσων και πόρων.

- Εκπαιδευτικές συνεργασίες και σχέσεις (οι σχέσεις που αναπτύσσονται μεταξύ των μελών της εκπαιδευτικής κοινότητας αλλά και μεταξύ αυτών και του θεσμικού περιβάλλοντος του εκπαιδευτικού οργανισμού): Σε αυτό τον θεματικό άξονα περιλαμβάνονται:
  - Σχέσεις εκπαιδευτικών (μελών ΔΕΠ) μεταξύ τους και με το υπόλοιπο προσωπικό.
  - Σχέσεις σπουδαστών μεταξύ τους και με τους εκπαιδευτικούς.
  - Σχέσεις με τοπικούς και περιφερειακούς εταίρους.

- Διδακτική και παιδαγωγική πρακτική (οι κυριότερες λειτουργίες του εκπαιδευτικού οργανισμού, που αποτελούν και το βασικό προσανατολισμό και σκοπό του εκπαιδευτικού έργου): Σε αυτό τον θεματικό άξονα περιλαμβάνονται:
  - Ποιότητα μαθημάτων και μάθησης.
  - Ποιότητα της παιδαγωγικής διδακτικής και επιστημονικής πράξης.
  - Αξιολόγηση.

- Εκπαιδευτικές επιτυχίες και αποτελέσματα (οι επιτυχίες και τα αποτελέσματα του εκπαιδευτικού έργου): Σε αυτό τον θεματικό άξονα περιλαμβάνονται:
  - Φοίτηση.
  - Επίδοση – εξέλιξη των σπουδαστών.
  - Συναισθηματική ανάπτυξη και ολοκλήρωση.
  - Επαγγελματικός προσανατολισμός και ακαδημαϊκή προοπτική.

Οι προαναφερθείσες θεματικές ενότητες που εκφράζονται μέσα από τις αντίστοιχες κλίμακες με τους επιμέρους δείκτες ποιότητας αλληλοσυμπληρώνονται και η αξιολόγηση γίνεται περισσότερο διεισδυτική, αναλυτική και αποτελεσματική (*Παιδαγωγικό Ινστιτούτο*, 2000). Όπως αναφέρεται χαρακτηριστικά στο κείμενο που αφορά την αξιολόγηση των εκπαιδευτικών οργανισμών από το Παιδαγωγικό Ινστιτούτο , οι κλίμακες / δείκτες ποιότητας συναρτώνται άμεσα και έχουν τα καλύτερα εκπαιδευτικά αποτελέσματα, όταν εξετάζονται σε σχέση με τα οικονομικά, κοινωνικά και πολιτισμικά χαρακτηριστικά των σπουδαστών της εκπαιδευτικής μονάδος αλλά και τις συνθήκες που επικρατούν σε τοπικό, περιφερειακό και εθνικό επίπεδο.

Τριτοβάθμια Εκπαίδευση

Η πολιτική και οι διαδικασίες Διασφάλισης Ποιότητας των ΑΕΙ, όπως και των ακαδημαϊκών τους μονάδων (Σχολών, Τμημάτων κλπ.), αξιολογούνται και πιστοποιούνται περιοδικά - τόσο εσωτερικά, με ευθύνη της ΜΟΔΙΠ, όσο και εξωτερικά, με ευθύνη της ΑΔΙΠ.

Σύμφωνα με τον Νόμο 4009/2011 , οι προϋποθέσεις, τα κριτήρια (ποιότητα των Προγραμμάτων Σπουδών, της Εκπαίδευσης, της Έρευνας, των Λοιπών Υπηρεσιών) και οι διαδικασίες της Εσωτερικής αλλά και της Εξωτερικής αξιολόγησης και Πιστοποίησης του εσωτερικού συστήματος διασφάλισης της ποιότητας των ΑΕΙ αναλύονται και εξειδικεύονται με ειδικές κατευθύνσεις και οδηγίες, που δημοσιοποιούνται από την ΑΔΙΠ.

Η διαδικασία πιστοποίησης περιλαμβάνει τέσσερις φάσεις:

Φάση 1η - Σχεδιασμός και ανάπτυξη του Εσωτερικού Συστήματος Διασφάλισης Ποιότητας του Ιδρύματος.

Φάση 2η - Αυτοαξιολόγηση του Ιδρύματος και του Συστήματος Διασφάλισης Ποιότητας.

Φάση 3η - Εξωτερική αξιολόγηση του Ιδρύματος και του Συστήματος Διασφάλισης Ποιότητας από επιτροπή ανεξάρτητων εξωτερικών εμπειρογνωμόνων.

Φάση 4η - Έκδοση Απόφασης πιστοποίησης από το Συμβούλιο της ΑΔΙΠ.

1.3.1.2 Κλίμακες - Δείκτες στην Οικονομία

Στο χώρο της οικονομίας γίνεται ευρεία χρήση δεικτών / κλιμάκων. Ενδεικτικά παρατίθενται οι ακόλουθες κλίμακες:

- **GDP (Gross Domestic Product)** – Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ): Το ΑΕΠ εκφράζει τον συνολικό όγκο των υλικών αγαθών που δημιούργησε η εθνική οικονομία, εκφραζόμενος σε τιμές. Η αύξηση του ΑΕΠ συνεπάγεται άνοδο της τιμής του συναλλάγματος της συγκεκριμένης χώρας.
- **Industrial Production (**Manufacturing*Output) – Όγκος Παραγωγής: Αυτή η κλίμακα εκφράζει τον όγκο της βιομηχανικής παραγωγής σε συγκεκριμένη περίοδο και η άνοδος της παραγωγικότητας συνεπάγεται άνοδο του συναλλάγματος.
- **Capacity Utilization** – Αξιοποίηση χωρητικότητας: Αυτή η κλίμακα εκφράζει το επίπεδο της ποσοστιαίας χρήσης της παραγωγικής ικανότητας και εξαρτάται από το επίπεδο της βιομηχανικής παραγωγής.
- **Retail Sales** – Λιανικές Πωλήσεις: Αυτή η κλίμακα εκφράζει τον όγκο των λιανικών πωλήσεων για συγκεκριμένη περίοδο. Η σταθερή ανάπτυξη των πωλήσεων συνεπάγεται εσωτερική ζήτηση και είναι ένας ισχυρός παράγοντας στο σχηματισμό της τιμής συναλλάγματος.
- **Business Inventories & Sales** – Αποθέματα και πωλήσεις επιχειρήσεων: Αυτή η κλίμακα εκφράζει τον όγκο των αποθεμάτων και των πωλήσεων των παραγωγών. Η αύξηση των αποθεμάτων και η μείωση των πωλήσεων μπορεί να προκαλέσει μείωση της παραγωγικότητας που είναι αρνητικός παράγοντας για το συνάλλαγμα. Αντίθετα, η αύξηση των αποθεμάτων, των πωλήσεων, αντίστοιχα συνεπάγεται αύξηση της επιχειρηματικής δραστηριότητας και αύξηση του κύκλου εργασιών, τα οποία είναι θετικός παράγοντας για το εθνικό νόμισμα.
- **Durable Goods** – Ανθεκτικά Είδη: Αυτή η κλίμακα εκφράζει την ποσότητα των παραγγελιών και μεταφερόμενων εμπορευμάτων και προορίζονται για μακροχρόνια χρήση. Αυτή η κλίμακα παρουσιάζει

σημαντική διακυμάνση, ανάλογα με την κατάσταση συγκεκριμένων βιομηχανιών.

- **New Home Sales** – Πωλήσεις Νέων Κατοικιών: Αυτή η κλίμακα εκφράζει τον όγκο των πωλήσεων νέων κατοικιών.
- **Existing Home Sales** – Πωλήσεις Παλαιών Κατοικιών: Αυτή η κλίμακα εκφράζει τον όγκο των πωλήσεων παλαιών κατοικιών. Οι δύο αυτές κλίμακες δείχνουν το επίπεδο της ευημερίας του πληθυσμού, το επίπεδο χρησιμοποίησης υποθήκης και το επίπεδο ικανοποίησης της ζήτησης, με βάση το υφιστάμενο επιτόκιο.
- **Factory Orders** – Παραγγελίες Εργοστασίων: Αυτή η κλίμακα εκφράζει την ποσότητα παραγγελιών των βασικών υλικών και ημι-προϊόντων στη διαδικασία παραγωγής. Η αύξηση της τιμής του δείκτη των παραγγελιών δείχνει το δυναμικό παραγωγής.
- **Budget** – Προυπολογισμός: Αυτή η κλίμακα εκφράζει την έγκριση και την εκτέλεση του προϋπολογισμού από τις κυβερνήσεις κρατών. Ο σχεδιασμός, η έγκριση και η εκτέλεση του προϋπολογισμού με πλεόνασμα δείχνει την αποτελεσματικότητα μιας κυβέρνησης αλλά και τη δυνατότητα αύξησης των δαπανών που συσχετίζονται με δημόσιες επενδύσεις, οι οποίες αποτελούν τη βάση της αύξησης του ΑΕΠ.
- **Construction Spending** – Κατασκευαστικές Δαπάνες: Αυτή η κλίμακα εκφράζει τον όγκο δαπανών για κατασκευές στον ιδιωτικό τομέα.
- **Leading Indicators** – Κύρια Κλίμακα: Αυτή η κλίμακα εκφράζει αρκετούς βασικούς δείκτες καθώς περιλαμβάνει 12 συνιστώσες μεταβλητές.
- **Personal Incomes** – Προσωπικά Εισοδήματα: Αυτή η κλίμακα εκφράζει το συνολικό εισόδημα ενός ατόμου.
- **Personal Spending** – Προσωπικές Δαπάνες: Αυτή η κλίμακα εκφράζει τις Προσωπικές Δαπάνες ενός ατόμου.
- **Real Earnings** – Πραγματικός Μισθός: Αυτή η κλίμακα εκφράζει το ωρομίσθιο ενός ατόμου, προσαρμοσμένο στο δείκτη πληθωρισμού.

- **Consumer Confidence** – Καταναλωτική Εμπιστοσύνη: Αυτή η κλίμακα εκφράζει τις τωρινές συνθηκες και των μελλοντικές προσδοκίες των καταναλωτών. Αυτή η κλίμακα έχει ιδιαίτερο ενδιαφέρον καθώς σε ένα δείγμα 10.000 νοικοκυριών δίνεται ερωτηματολόγιο με ερωτήσεις / μεταβλητές (Πίνακας 2.2) σχετικά με την τωρινή κατανάλωση και τις μελλοντικές καταναλωτικές προσδοκίες. Η κλίμακα αποτελείται από 5 συνιστώσες με 3 διαιρέσεις η κάθε μια (Θετική Στάση – Ουδέτερη Στάση – Αρνητική Στάση). OECD (2018), Consumer confidence index (CCI) (indicator). doi: 10.1787/46434d78-en.

Πίνακας 2.2: Κλίμακα καταναλωτικής εμπιστοσύνης (CCI)
Τρέχουσες επιχειρηματικές συνθήκες
Επιχειρηματικές συνθήκες για τους επόμενους έξι μήνες
Τρέχουσες συνθήκες απασχόλησης
Απαιτήσεις απασχόλησης για τους επόμενους έξι μήνες
Συνολικό οικογενειακό εισόδημα για τους επόμενους έξι μήνες

- **IFO**: Η κλίμακα αυτή εκφράζει τις πεποιθήσεις διεθυντικών στελεχών μεγάλων εταιρειών της Γερμανίας ως προς το επιχειρηματικό περιβάλλον και δημοσιεύεται από το Γερμανικό Ινστιτούτο Οικονομικών Ερευνών.
- **ISM** (Institute of Supply Management) - Δείκτης Ινστιτούτου Συντονισμού των Προμηθευτών: Η κλίμακα αυτή εκφράζει τις πεποιθήσεις διεθυντικών στελεχών 300 μεγάλων εταιρειών των ΗΠΑ ως προς την οικονομική κατάσταση. Από 0 έως 43.9 – η οικονομία στον τομέα της παραγωγής είναι σε παρακμή. Από 43.9 έως 50 – η οικονομία έχει ανοδική πορεία, όμως ο τομέας της παραγωγής έχει αρνητική πορεία. Πάνω από 50 – η οικονομία στον τομέα της παραγωγής έχει ανοδική πορεία.

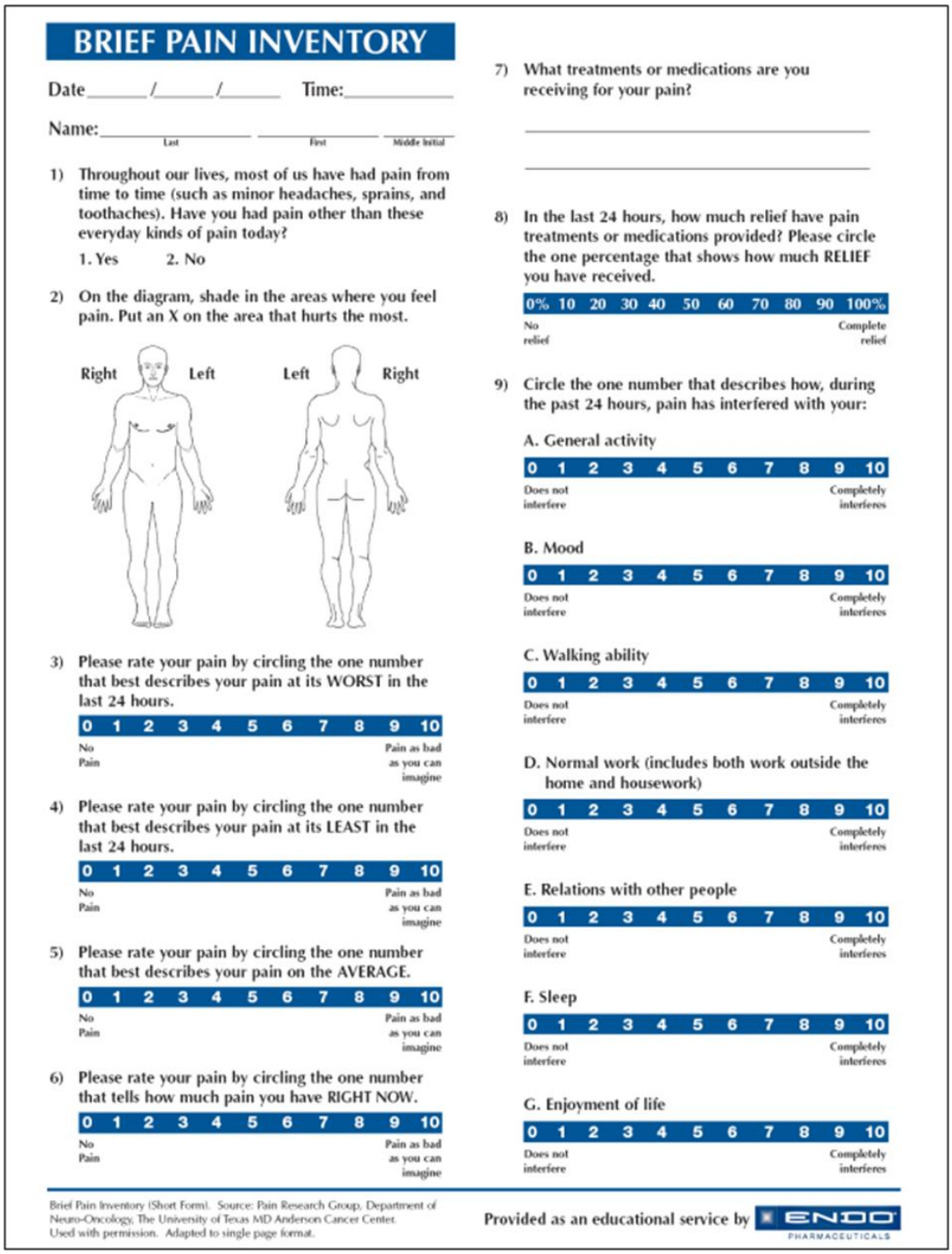
1.3.1.3 Κλίμακες στην Ιατρική

Στο χώρο της ιατρικής γίνεται ευρεία χρήση κλιμάκων με σκοπό την εκτίμηση, αξιολόγηση και κατάταξη ασθενών ή/και υγιών. Για παράδειγμα,

χρησιμοποιείται συνεχής κλίμακα μέτρησης που σκοπό έχει την αποτίμηση της έντασης του πόνου που αισθάνεται κάποιος ασθενής (*Huskisson, 1976*). Τέτοιες κλίμακες χρησιμοποιούνται για την εκτίμηση της ποσοτικοποίησης του αισθήματος του πόνου που προφανώς δεν μετράται ευθέως με αντικειμενικό τρόπο, αλλά, αποτελεί ένα ατομικό και υποκειμενικό συναίσθημα. Το αίσθημα του πόνου ενός ασθενούς δύναται να εκφραστεί σε κάποιο βαθμό από την απάντηση του ασθενούς σε σχετικές ερωτήσεις κατάλληλου ερωτηματολογίου, όταν γίνεται αυτοαναφορά από τον ίδιο των ασθενή, με προφορικό ή γραπτό τρόπο. Η κλίμακά δύναται να είναι διακριτή με χρήση επιθέτων όπως «ελαφριά αίσθηση πόνου», «μέτρια αίσθηση πόνου», «βαριά ή έντονη αίσθηση πόνου» κτλ ή συνεχής με χρήση κατάλληλης αριθμητικής/οπτικής αναλογικής κλίμακας. Η γνωστότερη ίσως κλίμακες μέτρησης της αίσθησης του πόνου είναι η BPI (Brief Pain Inventory) που δημιουργείται από τέσσερις δείκτες που αξιολογούν την ένταση του πόνου και επτά που αξιολογούν την επίδραση του και βαθμονομούνται από το 0 (καθόλου πόνος) μέχρι το 10 (τόσος πόνος όσο δεν μπορείς να φανταστείς) (*Cleeland & Ryan, 1994*). Το ερωτηματολόγιο με τις ερωτήσεις/συνιστώσες μεταβλητές παρατίθεται στο σχήμα 1.3.1.3.1.

Στο χώρο της ιατρικής χρησιμοποιούνται κλίμακες που στόχο έχουν την αποτίμηση της ποιότητας της ζωής ενός ατόμου γενικά, λαμβάνοντας υπόψη παράγοντες που επηρεάζουν την ποιότητα της ζωής, όπως είναι ο αισθητός πόνος, η κατάθλιψη ή το άγχος που βιώνει κάποιο άτομο, ο βαθμός εξάρτησης από άλλα άτομα στην καθημερινότητα κτλ. Μια τέτοια κλίμακα (δείκτης) είναι η QL Index (Quality of Life Index) (*Spitzer, Dobson et al. 1981*), η οποία στόχο έχει την αποτίμηση της συνολικής ευεξίας των ασθενών που βρίσκονται σε τελικό στάδιο χρόνιων παθήσεων (π.χ. καρκίνος) αξιολογώντας με αυτό τον τρόπο την επίδραση της θεραπείας και της αναλγητικής φροντίδας τους. Κάθε ερώτηση/συνιστώσα μεταβλητή της κλίμακας βαθμονομείται με τιμές 0, 1 και 2 και αθροιζόμενα τα επιμέρους σκορ προκύπτει ένα συνολικό σκορ που κυμαίνεται μεταξύ 0 και 10, όπου υψηλότερο σκορ υποδηλώνει αυξημένη ευεξία του ασθενούς. Μια εναλλακτική κλίμακα εκτίμησης της ποιότητας της ζωής συνολικά είναι η SF-36 (*Ware JE & Sherbourne CD, 1992*) που χρησιμοποιείται κυρίως σε επίπεδο πληθυσμιακών μελετών με στόχο την αξιολόγηση πολιτικών δημόσιας υγείας. Αυτή η κλίμακα δημιουργείται από συνιστώσες μεταβλητές που εκφράζουν χαρακτηριστικά όπως είναι η σωματική λειτουργία, οι περιορισμοί που βιώνονται λόγω προβλημάτων υγείας ή συναισθηματικών

προβλημάτων, οι σωματικοί πόνοι, η ένταση κοινωνικών επαφών, το επίπεδο πνευματικής διαύγειας και ο βαθμός κούρασης.



BRIEF PAIN INVENTORY

Date_____/_____/_____ Time:_____

Name:_____ (Last) _____ (First) _____ (Middle Initial)

1) Throughout our lives, most of us have had pain from time to time (such as minor headaches, sprains, and toothaches). Have you had pain other than these everyday kinds of pain today?
   1. Yes 2. No

2) On the diagram, shade in the areas where you feel pain. Put an X on the area that hurts the most.

Right Left Left Right

3) Please rate your pain by circling the one number that best describes your pain at its WORST in the last 24 hours.
   0 1 2 3 4 5 6 7 8 9 10
   No Pain — Pain as bad as you can imagine

4) Please rate your pain by circling the one number that best describes your pain at its LEAST in the last 24 hours.
   0 1 2 3 4 5 6 7 8 9 10
   No Pain — Pain as bad as you can imagine

5) Please rate your pain by circling the one number that best describes your pain on the AVERAGE.
   0 1 2 3 4 5 6 7 8 9 10
   No Pain — Pain as bad as you can imagine

6) Please rate your pain by circling the one number that tells how much pain you have RIGHT NOW.
   0 1 2 3 4 5 6 7 8 9 10
   No Pain — Pain as bad as you can imagine

7) What treatments or medications are you receiving for your pain?

8) In the last 24 hours, how much relief have pain treatments or medications provided? Please circle the one percentage that shows how much RELIEF you have received.
   0% 10 20 30 40 50 60 70 80 90 100%
   No relief — Complete relief

9) Circle the one number that describes how, during the past 24 hours, pain has interfered with your:

A. General activity
0 1 2 3 4 5 6 7 8 9 10
Does not interfere — Completely interferes

B. Mood
0 1 2 3 4 5 6 7 8 9 10
Does not interfere — Completely interferes

C. Walking ability
0 1 2 3 4 5 6 7 8 9 10
Does not interfere — Completely interferes

D. Normal work (includes both work outside the home and housework)
0 1 2 3 4 5 6 7 8 9 10
Does not interfere — Completely interferes

E. Relations with other people
0 1 2 3 4 5 6 7 8 9 10
Does not interfere — Completely interferes

F. Sleep
0 1 2 3 4 5 6 7 8 9 10
Does not interfere — Completely interferes

G. Enjoyment of life
0 1 2 3 4 5 6 7 8 9 10
Does not interfere — Completely interferes

Brief Pain Inventory (Short Form). Source: Pain Research Group, Department of Neuro-Oncology, The University of Texas MD Anderson Cancer Center. Used with permission. Adapted to single page format.

Provided as an educational service by ENDO PHARMACEUTICALS

Εικόνα 1.3.1.3.1: Το ερωτηματολόγιο με τις ερωτήσεις που συνθέτουν την κλίμακα Brief Pain Inventory
Πηγή: (Κουρλαμπά, 2007)

Για τον υπολογισμό του συνολικού σκορ, έχουν προταθεί δύο εναλλακτικές προσεγγίσεις υπολογισμού. Το αστάθμιστο άθροισμα που παράγει αποτέλεσμα με

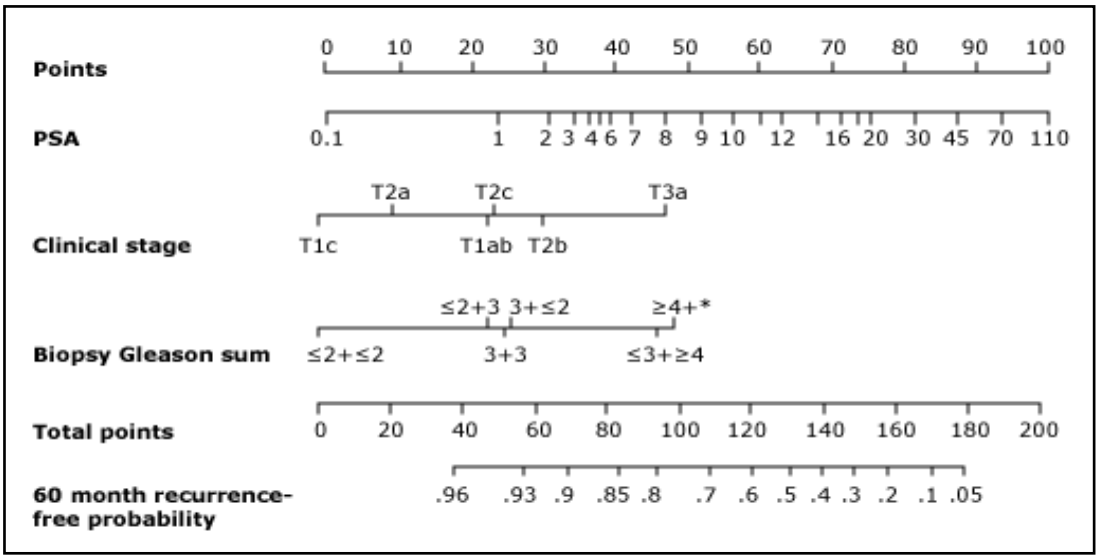
πεδίο ορισμού [0, 100] και το σταθμισμένο άθροισμα με μέση τιμή 50 και τυπική απόκλιση 10. Η κλίμακα αυτή έχει δοκιμαστεί ως προς την αξιοπιστία και η εγκυρότητα της σε ελληνικό πληθυσμό (*Pappa, Kontodimopoulos et al. 2005*). Κλίμακες έχουν χρησιμοποιηθεί για την πρόβλεψη κινδύνου από καρδιαγγειακές παθήσεις με χρήση μαθηματικών / στατιστικών μοντέλων που λαμβάνουν ως ερμηνευτικές μεταβλητές παράγοντες όπως είναι η ηλικία, το φύλο, το κάπνισμα, οι διατροφολογικές συνήθειες και η ύπαρξη κάποιου χρόνιου νοσήματος. Αυτού του είδους κλίμακες χρησιμοποιήθηκαν στην προοπτική επιδημιολογική μελέτη *Framingham Heart* (*Wilson, D'Agostino et al. 1998; Wang, Massaro et al. 2003*) που σχεδιάστηκε για την παρακολούθηση του καρδιαγγειακού κινδύνου του πληθυσμού της πόλης Framingham, των Η.Π.Α. Οι κλίμακες / δείκτες / μοντέλα πρόβλεψης που προέκυψαν προτείνουν εκτιμήσεις για την εκδήλωση καρδιαγγειακού συμβάντος (στηθάγχη, έμφραγμα του μυοκαρδίου, καρδιακή ανεπάρκεια, εγκεφαλικό επεισόδιο, κτλ) εντός συγκεκριμένου χρονικού διαστήματος. Η βαθμονόμηση των μεταβλητών/συνιστωσών κινδύνου όπου αθροιζόμενες παρέχουν το τελικό σκορ της κλίμακας για τον 10-ετη κίνδυνο εμφάνισης καρδιαγγειακού συμβάντος, στην περίπτωση του γυναικείου πληθυσμού, παρατίθεται στον πίνακα 1.3.1.3.1 (*D'Agostino, Vasan et al. 2008*).

Points	Age, y	HDL	Total Cholesterol	SBP Not Treated	SBP Treated	Smoker	Diabetic
−3				<120			
−2		60+					
−1		50–59			<120		
0	30–34	45–49	<160	120–129		No	No
1		35–44	160–199	130–139			
2	35–39	<35		140–149	120–129		
3			200–239		130–139	Yes	
4	40–44		240–279	150–159			Yes
5	45–49		280+	160+	140–149		
6					150–159		
7	50–54				160+		
8	55–59						
9	60–64						
10	65–69						
11	70–74						
12	75+						
Points allotted							Total

Πίνακας 1.3.1.3.1: Βαθμονόμηση παραγόντων κινδύνου στην περίπτωση του γυναικείου πληθυσμού. Πηγή: (Κουρλαμπά, 2007)

Τα τελευταία χρόνια, έχουν επίσης προταθεί, κλίμακες που αποτιμούν τον κίνδυνο εμφάνισης καρδιαγγειακής νόσου (στεφανιαία νόσο) κάνοντας χρήση συνιστωσών μεταβλητών που εκφράζουν γονιδιακούς πολυμορφισμούς, όπου η

βαθμονόμηση γίνεται εξαρτώμενη με το αν τα εξεταζόμενα άτομα ήταν ομόζυγα με το αλλήλιο υψηλού κινδύνου ή ετερόζυγα και το συνολικό σκορ λαμβάνει τιμές στο πεδίο ορισμού [0,11] (*Yiannakouris, Trichopoulou et al. 2006*). Κλίμακες χρησιμοποιούνται επίσης στην αντιμετώπιση/πρόληψη/εκτίμηση του καρκίνου. Για παράδειγμα, πριν την επέμβαση, η εκτίμηση κινδύνου υποτροπής του καρκίνου του προστάτη μετά την αφαίρεσή του (*Kattan, Eastham et al. 1998*) πραγματοποιείται με κατάλληλη κλίμακα που δημιουργείται από συνιστώσες που εκφράζουν την ποσόστωση του αντιγόνου του προστάτη (PSA) και του κλινικού σταδίου της νόσου. Στην εικόνα 1.3.1.3.2 παρουσιάζεται ο τρόπος βαθμονόμησης των συνιστωσών της κλίμακας. Για την πρόγνωση του κινδύνου θανάτου ασθενών που νοσηλεύονται σε μονάδες εντατικής θεραπείας (ΜΕΘ) χρησιμοποιούνται κλίμακες αξιολόγησης των ασθενών που νοσηλεύονται σε ΜΕΘ. Οι πιο συχνά χρησιμοποιούμενες κλίμακες είναι ο APACHE (Acute Physiology and Chronic Health Evaluation) (*Knaus, Zimmerman et al. 1981*) και ο SAPS (Simplified Acute Physiology Score) (*Le Gall, Lemeshow et al. 1993*). Η κλίμακα APACHE λαμβάνει υπόψη κριτήρια διατάραξης της κανονικής λειτουργίας ενός οργανισμού και τυχόν συνοδών νοσημάτων. Η κλίμακα SAPS δίνει μικρότερη βαρύτητα σε τυχόν συνοδά νοσήματα.



Εικόνα 1.3.1.3.2: Βαθμονόμηση παραγόντων κινδύνου στην περίπτωση του καρκίνου του προστάτη
Πηγή: (Κουρλαμπά, 2007)

1.3.1.4 Κλίμακες στην Ψυχομετρία

Στον τομέα της ψυχομετρίας, υπάρχουν ορισμένα χαρακτηριστικά γνωρίσματα όπως είναι η κατάθλιψη και το άγχος που εκτιμούνται από τους κατάλληλους ψυχομετρικές κλίμακες (*Jackson, 1970*), ενώ, ο κατάλληλος

επιστήμονας (ψυχίατρος) αποφαίνεται για την ακριβή κατάσταση της υγείας του (gold standard method). Ιστορικά, οι κλίμακες που σχετίζονται με την υγεία έχουν προταθεί από τις αρχές του περασμένου αιώνα, κυρίως στον ψυχομετρικό τομέα (*Aitken et al., 1969; Beck et al., 1961;, Hamilton, et al., 1960, Hayes, et al., 1921; Spielberger et al., 2010; Taylor et al., 1953, Zung et al., 1965*).

Συνιστώσες	**Σπάνια ή καθόλου λιγ. από 1 μέρα**	**Μερικές φορές για 1-2 μέρες**	**Αρκετές φορές για 3-4 μέρες**	**Τον περισσότερο καιρό για 5-7 μέρες**
1. Σας ενοχλούν πράγματα που συνήθως δεν σας ενοχλούν	0	2	2	3
2. Η όρεξη σας είναι κακή ώστε να μη θέλετε να τρώτε	0	2	2	3
3. Νιώθατε ότι δεν μπορούσατε να απαλλαγείτε από την κακοκεφιά αμέσως, ακόμα κι αν οι φίλοι ή η οικογένεια σας, σας βοηθούσαν	0	2	2	3
4. Νιώθετε ότι είσαστε εξίσου ικανός με τους άλλους	3	2	1	0
5. Έχετε δυσκολία να συγκεντρωθείτε σ' αυτό που κάνατε	0	2	2	3
6. Νιώθετε μελαγχολικός	0	2	2	3
7. Νιώθετε πως ότι κάνετε το κάνετε με μεγάλη προσπάθεια, πίεση	0	2	2	3
8. Νιώθετε ελπίδα για το μέλλον	3	2	1	0
9. Σκεφτόσαστε ότι η μέχρι τώρα ζωή σας ήταν μια αποτυχία	0	2	2	3
10. Νιώθετε φοβισμένος	0	2	2	3
11.Εχετε ανήσυχο ύπνο	0	2	2	3
12. Είσαστε ευτυχισμένος	3	2	1	0
13. Μιλάτε λιγότερο από το συνηθισμένο	0	2	2	3
14. Νιώθετε μοναξιά	0	2	2	3
15. Πιστεύετε ότι οι άνθρωποι δεν είναι φιλικοί μαζί σας	0	2	2	3
16. Ευχαριστιέστε (χαιρόσαστε τη ζωή)	3	2	1	0
17. Κλαίτε με λυγμούς	0	2	2	3
18. Νιώθετε λυπημένος	0	2	2	3
19. Νιώθετε ότι οι άνθρωποι δεν σας συμπαθούν	0	2	2	3
20. Δεν μπορείτε να κάνετε τις δουλειές σας (εργασία, σπίτι, σχολείο)	0	2	2	3

Πίνακας 1.3.1.4.1: Βαθμονόμηση συνιστωσών κλίμακας του Center for Epidemiologic Studies Depression Scale
Πηγή: (Κουρλαμπά, 2007)

Πιο συγκεκριμένα, για την εκτίμηση της κατάθλιψης υπάρχει πληθώρα κλιμάκων στη βιβλιογραφία, όπως η κλίμακα HRSD (Hamilton Rating Scale for Depression) (*Hamilton, et al., 1960*) που αποτελείται από 21 συνιστώσες, η κλίμακα BDI (Beck Depression Inventory) (*Beck et al., 1961*) που αποτελείται από 21

συνιστώσες, η κλίμακα Self-Rating Depression Scale (*Zung et al., 1965*) που αποτελείται από 20 συνιστώσες, η κλίμακα Depression Adjective Check-Lists (*Lubin et al., 1965*) που αποτελείται από 32 συνιστώσες, η κλίμακα Center for Epidemiologic Studies Depression Scale (*Radloff et al., 1977*) που αποτελείται από 20 συνιστώσες, η κλίμακα Carroll Rating Scale for Depression (*Carroll et al., 1981*) που αποτελείται από 52 συνιστώσες και η κλίμακα Geriatric Depression Scale (*Brink & Yesavage., 1982)* που αποτελείται από 30 συνιστώσες.

Ο τρόπος βαθμονόμησης των προαναφερθέντων κλιμάκων γίνεται με απόδοση υψηλών τιμών σε στάσεις που συνάδουν με την πάθηση της κατάθλιψης όταν αντιστοιχούν σε μεγάλη συχνότητα και αντιστρόφως. Για παράδειγμα, στον πίνακα 1.3.1.4.1 παρατίθεται ο τρόπος βαθμονόμησης των συνιστωσών που χρησιμοποιούνται για την κατασκευή της κλίμακας Center for Epidemiologic Studies Depression Scale (*Radloff, 1977*).

Στο χώρο της ψυχομετρίας, παρά το γεγονός ότι έχει δοθεί ιδιαίτερη σημασία στην ανίχνευση της νόσου της κατάθλιψης, υπάρχουν και άλλες ψυχολογικές καταστάσεις, όπως το άγχος, για τις οποίες υπάρχουν κλίμακες που σκοπό έχουν την αποτίμηση τους. Πιο συγκεκριμένα, για την εκτίμηση της έντασης του άγχους υπάρχει πληθώρα κλιμάκων στη βιβλιογραφία, όπως η κλίμακα MAS (Manifest Anxiety Scale) (*Taylor, et al., 1953*) που αποτελείται από 50 συνιστώσες, η κλίμακα HARSD (Hamilton Anxiety Rating Scale for Depression) (*Hamilton, et al., 1959*) που αποτελείται από 14 συνιστώσες, η κλίμακα STAI (State-Trait Anxiety Inventory) (*Spielberger, 1970*) που αποτελείται από 40 συνιστώσες, η κλίμακα HADS (Hospital Anxiety and Depression Scale) (*Zigmond & Snaith, 1983*) που αποτελείται από 14 συνιστώσες, η κλίμακα BAI (Beck Anxiety Inventory) (*Beck, 1988*) που αποτελείται από 21 συνιστώσες και η κλίμακα DASS (Depression Anxiety Stress Scales) (*Lovibond, 1995*) που αποτελείται από 42 συνιστώσες. Πρόσφατα, η κλίμακα STAI σταθμίστηκε σε ελληνικό πληθυσμό (*Fountoulakis, Papadopoulou et al. 2006*).

1.3.1.5 Κλίμακες στη Διαιτολογία

Στον τομέα της διατροφικής επιστήμης η επάρκεια, η ποιότητα και η συχνότητα της διατροφής μετρούνται από κατάλληλες διατροφικές κλίμακες (*Kant, 1996*). Στην πρόσφατη βιβλιογραφία, έχουν προταθεί συνδυαστικές κλίμακες για την υγεία που έχουν σχέση με τον τρόπο ζωής, οι οποίοι αποσκοπούν στην εκτίμηση της ποιότητας και ποσότητας διάφορων διατροφικών συστατικών (*Bach et al., 2006;*

Patterson, et al., 1994; Kennedy et al., 1995; Waijers, P et al., 2007). Για παράδειγμα, έχει διερευνηθεί ο βαθμός υιοθέτησης ενός ισορροπημένου διατροφικού προτύπου ως προς την ποιότητα και την ποσότητα (*Bach et al, 2006*) και επίσης η διατροφική κλίμακα «MedDietScore» (*Panagiotakos et al., 2007*) αποτιμά το βαθμό υιοθέτησης της Μεσογειακής διατροφής. Όσο πιο μεγάλες τιμές λαμβάνει η κλίμακα για ένα άτομο, τόσο πιο κοντά είναι αυτό στην υιοθέτηση στο παραδοσιακό πρότυπο διατροφής και διατρέχει μικρότερο κίνδυνο εμφάνισης καρδιαγγειακής νόσου.

Οι κλίμακες που χρησιμοποιούνται στην επιστήμη της Διαιτολογίας – Διατροφολογίας συνήθως κατασκευάζονται βάσει συστάσεων από οργανισμούς υγείας (π.χ. Παγκόσμιος Οργανισμός Υγείας - ΠΟΥ) ώστε να αποτελούν αξιόπιστα εργαλεία μέτρησης της διατροφικής ποιότητας. Παραδείγματα τέτοιων κλιμάκων αποτελούν, η κλίμακα DQI (Diet Quality Index) (*Patterson et al, 1994*) που αποτελείται από 8 συνιστώσες, με 3 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης θρεπτικών συστατικών και ομάδων τροφίμων που παραπέμπουν στο πρότυπο Μεσογειακής Διατροφής με βαθμονόμηση από 0 έως 2, η κλίμακα HEI (Healthy Eating Index) (*Kennedy et al, 1995*) που αποτελείται από 10 συνιστώσες, με 0-10 βαθμούς έκαστη, που εκφράζουν τη συχνότητα αλλά και την ποικιλία κατανάλωσης θρεπτικών συστατικών και ομάδων τροφίμων με πεδίο ορισμού [0, 100], η κλίμακα HDI (Healthy Diet Indicator) (*Huijbregts et al, 1997*) που αποτελείται από 9 συνιστώσες, με 2 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης θρεπτικών συστατικών και ομάδων τροφίμων με πεδίο ορισμού [0, 9], η κλίμακα DQI-R (Diet Quality Index Revised) (*Haines et al, 1999*) που αποτελείται από 10 συνιστώσες, με 10 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης και την ποικιλία θρεπτικών συστατικών και ομάδων τροφίμων με πεδίο ορισμού [0, 100], η κλίμακα DGI (Dietary Guideline Index) (*Harnack et al, 2002*) που αποτελείται από 9 συνιστώσες, με 3 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης θρεπτικών συστατικών και ομάδων τροφίμων με πεδίο ορισμού [0, 18], η κλίμακα AHEI (Alternative Healthy Eating Index) (*McCullough et al, 2002*) που αποτελείται από 9 συνιστώσες, που εκφράζουν τη συχνότητα κατανάλωσης θρεπτικών συστατικών και ομάδων τροφίμων με πεδίο ορισμού [2.5, 87.5] και η κλίμακα DQI-I (Diet Quality Index International) (*Kim et al, 2003*) που αποτελείται από 17 συνιστώσες, που εκφράζουν την ποικιλία, την επάρκεια, το μέτρο και την ισορροπία κατανάλωσης με πεδίο ορισμού [0, 100].

Εκτός από τις γενικές διατροφολογικές η κλίμακες υπάρχουν κλίμακες που εκφράζουν το βαθμό υιοθέτησης συγκεκριμένων διατροφικών προτύπων όπως είναι το πρότυπο της Μεσογειακής διατροφής. Τέτοιες κλίμακες έχουν μεγάλη ανάπτυξη τα τελευταία χρόνια, όπως είναι η κλίμακα MDQI (Mediterranean Diet Quality Index) (*Gerber et al, 2000*) που αποτελείται από 7 συνιστώσες, με 3 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης θρεπτικών συστατικών και ομάδων τροφίμων με πεδίο ορισμού [0, 14], η κλίμακα «A prior Mediterranean dietary pattern» (*Martinez-Gonzalez et al, 2002*) που αποτελείται από 8 συνιστώσες, με 5 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης ομάδων τροφίμων με πεδίο ορισμού [0, 40], η κλίμακα MMDS (Modified Mediterranean Diet Scale) (*Trichopoulou et al, 2003*) που αποτελείται από 9 συνιστώσες, με 2 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης ομάδων τροφίμων με πεδίο ορισμού [0, 9], η κλίμακα «Mediterranean Score» (*Martinez-Gonzalez et al, 2004*) που αποτελείται από 9 συνιστώσες, με 2 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης ομάδων τροφίμων με πεδίο ορισμού [0, 9] και η κλίμακα «MedDietScore» (*Panagiotakos et al, 2007*) που αποτελείται από 11 συνιστώσες, με 5 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης ομάδων τροφίμων με πεδίο ορισμού [0, 55].

Συχνότητα κατανάλωσης	**Πιθανές απαντήσεις**			
	Ποτέ **0 φορές/ εβδομάδα**	**Μερικές φορές** **1-2 φορές/ εβδομάδα**	**Συχνά** **3-5 φορές/ εβδομάδα**	**Πολύ συχνά** **6 ≥ φορές/ εβδομάδα**
Ψωμί	0	1	1	3
Δημητριακά	0	1	2	2
Φρούτα και φρουτοχυμούς	0	1	2	3
Λαχανικά	0	1	2	3
Όσπρια	0	1	3	3
Γάλα	0	1	2	3
Ψάρι και θαλασσινά	0	3	3	2
Κρέας	1	3	0	1
Αλμυρά & καπνιστά κρέατα	3	1	1	0
Γλυκά & σνακ	2	2	1	0
Αναψυκτικά	3	3	1	0
Τηγανιτά	2	3	1	0

Ψητά	0	3	3	2

Πίνακας 1.3.1.5.2: Βαθμονόμηση συνιστωσών της κλίμακας Foods e-KINDEX
Πηγή: (Κουρλαμπά, 2007)

Υψηλές τιμές των προαναφερθέντων κλιμάκων εκφράζουν πλήρη υιοθέτηση του προτύπου της Μεσογειακής Διατροφής και χαμηλές τιμές φτωχή υιοθέτηση του προτύπου της Μεσογειακής Διατροφής. Για παράδειγμα, στον πίνακα 1.3.1.5.2 παρατίθεται ο τρόπος βαθμονόμησης των συνιστωσών που χρησιμοποιούνται για την κατασκευή της η κλίμακας Foods e-KINDEX (*Lazarou, Panagiotakos et al. 2009*).

Οι κλίμακες που αναφέρθηκαν στις προηγούμενες παραγράφους εφαρμόζονται σε γενικούς πληθυσμούς. Στη διεθνή βιβλιογραφία εμφανίζονται κλίμακες που στοχεύουν σε ειδικές ομάδες όπως είναι τα παιδιά και οι έφηβοι, όπως είναι η κλίμακα KIDMED (Mediterranean Diet Quality Index in children and adolescents) (*Serra-Majem, Ribas et al. 2004*) που αποτελείται από 16 συνιστώσες, με 2 διαμερίσεις έκαστη, που εκφράζουν τη συχνότητα κατανάλωσης ομάδων τροφίμων με πεδίο ορισμού [0, 16], η κλίμακα YHEI (Youth Healthy Eating Index) (*Feskanich, Rockett et al. 2004*) που αποτελείται από 13 συνιστώσες, με μεταβλητό σύστημα βαθμονόμησης, που εκφράζουν τη συχνότητα κατανάλωσης ομάδων τροφίμων με πεδίο ορισμού [0, 100] και η κλίμακα Foods E-Kindex (*Lazarou, Panagiotakos et al. 2009*) που αποτελείται από 13 συνιστώσες, με μεταβλητό σύστημα βαθμονόμησης, που εκφράζουν τη συχνότητα κατανάλωσης τροφίμων με πεδίο ορισμού [0, 37].

1.3.2 Διακριτές & Συνεχείς Σύνθετες Κλίμακες

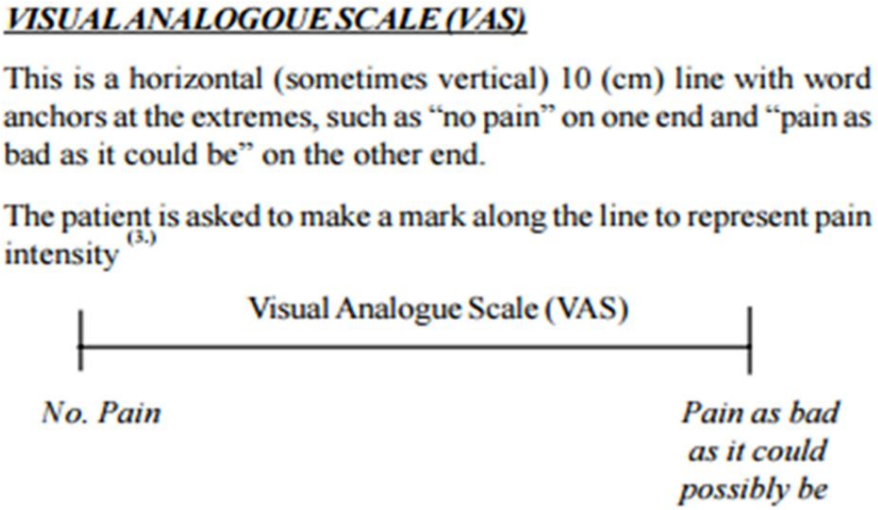
Οι κλίμακες είναι σύνθετα εργαλεία, που κατασκευάζονται από το σταθμισμένο ή αστάθμιστο άθροισμα διακριτών ή συνεχών μεταβλητών. Αυτός ο τύπος κλιμάκων στοχεύει να μετρήσει με τη μέγιστη δυνατή ακρίβεια μια ποικιλία κλινικών χαρακτηριστικών, συμπεριφορές, νοοτροπίες και πεποιθήσεις, οι οποίες είναι δύσκολο ή ακόμη και αδύνατο να μετρηθούν ποσοτικά και άμεσα (π.χ. συναισθήματα, άγχος, κατάθλιψη, πόνος, ποιότητα διατροφής, κλπ) (*Panagiotakos. 2009; Streiner & Norman. 2008*).

Ο στόχος των κλιμάκων που σχετίζονται με την υγεία, όπως ήδη αναφέρθηκε, είναι ο ποσοτικός προσδιορισμός και η μέτρηση των κλινικών χαρακτηριστικών ενός ατόμου, όπως είναι διάφορα ψυχολογικά χαρακτηριστικά του, οι διατροφικές συνήθειες του και η συσχέτιση με κάποια νόσο που είναι δύσκολο ή και αδύνατο να

μετρηθούν εξαιτίας της έλλειψης μιας αντικειμενικής κλίμακας. Όπως κάθε εργαλείο μέτρησης, έτσι και οι κλίμακες, δύναται να είναι διακριτές ή συνεχείς, εξαρτώμενη από τη φύση του προς εκτίμηση χαρακτηριστικού.

Οι διακριτές κλίμακες δημιουργούνται από διακριτές ή ιεραρχικές (διατακτικές) ή κατηγορικές (μη διατακτικές) μεταβλητές, οι οποίες εκφράζονται μέσω συγκεκριμένων απαντήσεων σε κατάλληλες ερωτήσεις ερωτηματολογίων (με αυτοαναφορά ή παρουσία ερευνητή). Αυτές οι μεταβλητές χρησιμοποιούνται ως συνιστώσες μεταβλητές για τη δημιουργία σύνθετων κλιμάκων υγείας. Μια κατηγορική μεταβλητή δύναται να είναι δίτιμη (διχοτομική) (π.χ. μεταβλητή φύλο με κατηγορίες «άνδρας» και «γυναίκα») ή με περισσότερες κατηγορίες (π.χ. μεταβλητή οικογενειακή κατάσταση με κατηγορίες «ανύπαντρος», «παντρεμένος» και «διαζευγμένος») αλλά συνήθως εμφανίζεται η αδυναμία του να περιγραφεί ένα συνεχές χαρακτηριστικό με ποσοτική συνεχή κλίμακα, όπως είναι ο δείκτης μάζας σώματος. Ο δείκτης μάζας σώματος είναι μια συνεχής μεταβλητή, αλλά για λόγους ευκολίας κατηγοριοποιείται κατάλληλα στις κατηγορίες «λιποβαρής», «κανονικός», «υπέρβαρος» και «παχύσαρκος» με αντίστοιχα όρια που προτείνονται από επίσημους οργανισμούς υγείας, δημιουργώντας μια ιεραρχική μεταβλητή. Σε άλλο παράδειγμα, για την αξιολόγηση της συχνότητας και διάρκειας εκγύμνασης απαιτείται μια συνεχή κλίμακα που να περιγράφει πλήρως τη συχνότητα και διάρκεια εκγύμνασης, σε εβδομαδιαία βάση, και δεν είναι απόλυτα δόκιμη η χρήση ερωτήσεων της μορφής «Πόσο συχνά γυμνάζεστε ανά εβδομάδα;» με πιθανές απαντήσεις «Ποτέ», «Μια – τρεις φορές την εβδομάδα», «τέσσερις φορές την εβδομάδα» και «Κάθε μέρα», καθώς υπάρχει απώλεια μέρους της αρχικής πληροφορίας (διότι υπάρχει διαφορά μεταξύ της απάντησης «1 φορά την εβδομάδα» και της απάντησης «3 φορές την εβδομάδα»), και επίσης, είναι σαν να αγνοούμε το γεγονός ότι υπάρχουν ποικίλες μορφές εκγύμνασης, με διαφορετικά αποτελέσματα στην απώλεια θερμίδων ανά μονάδα χρόνου. Λόγω της ευκολίας όμως αποτίμησης ενός χαρακτηριστικού με διακριτό τρόπο, χρησιμοποιούνται ευρέως οι διακριτές κλίμακες (π.χ. είναι δύσκολο κάποιο άτομο να μετρά με ακρίβεια τα λεπτά εκγύμνασης του ανά ημέρα), παρότι δίνουν λιγότερο έγκυρα αποτελέσματα σε σχέση με τις συνεχείς κλίμακες. Λόγω του γεγονότος ότι πολλά βιοχημικά ή/και βιολογικά χαρακτηριστικά προέρχονται από μέτρηση συνεχών μεταβλητών, έχουν προταθεί δείκτες που κάνουν χρήση ποσοτικής κλίμακας όπως είναι η οπτική αναλογική συνεχής κλίμακα VAS (Visual Analogous Scale) και η διακριτή κλίμακα Likert (*Huskisson, 1974*). Η κλίμακα VAS

οπτικοποιείται με τη χρήση ενός προσανατολισμένου ευθύγραμμου τμήματος με συγκεκριμένο μήκος, συνήθως 10 cm όπου στα δύο άκρα του αναφέρονται η απόλυτη απουσία και η απόλυτη παρουσία ενός χαρακτηριστικού υγείας (Εικόνα 1.3.2.1).



Εικόνα 1.3.2.2: Κλίμακα μέτρησης πόνου (Huskisson, 1965)

Η χρήση της κλίμακας VAS πραγματοποιείται με τους συμμετέχοντες να εκφράζουν την άποψη τους, σε ερώτηση ενός ερωτηματολογίου, σημειώνοντας πάνω στο ευθύγραμμο τμήμα ένα κατάλληλο σημείο ώστε να αντιστοιχεί στη θέση που τους εκφράζει. Η κλίμακα VAS πρωτοεμφανίστηκε πολλά χρόνια πριν (*Hayes, 1921*) και χρησιμοποιείται κυρίως στο χώρο της ιατρικής για την ποσοτικοποίηση μη αντικειμενικά μετρήσιμων χαρακτηριστικών, όπως ο πόνος (*Huskisson 1974*) και η διάθεση *(Aitken 1969*). Η κλίμακα VAS χαρακτηρίζεται από την απλότητα της και συμπεριλαμβάνοντας πολλές συνιστώσες που να μετρώνται με κλίμακα VAS, η κάθε μια αξιολογώντας διαφορετικές πτυχές του χαρακτηριστικού που η σύνθετη κλίμακα στοχεύει να αποτιμήσει, δημιουργούνται σύνθετες συνεχείς κλίμακες. Σε κάθε περίπτωση, η κλίμακα VAS πλεονεκτεί λόγω της απλής μορφής της, όμως υπάρχουν άλλες κλίμακες μέτρησης που θεωρείται ότι είναι εγκυρότερες (*Streiner & Norman 2008*), όπως είναι η κλίμακα μέτρησης Likert (*Likert 1952*) που συνήθως αξιολογεί τον βαθμό συμφωνίας ή διαφωνίας ενός ατόμου με μια στάση. Για παράδειγμα, στην πενταβάθμια κλίμακα Likert, οι πέντε πιθανές απαντήσεις (διαμερίσεις) που θα αντιστοιχούσαν σε κάποια ερώτηση που διερευνά τον βαθμό συμφωνίας με κάποια συγκεκριμένη στάση είναι «Διαφωνώ απόλυτα», «Διαφωνώ λίγο», «Ούτε Διαφωνώ, ούτε Συμφωνώ», «Συμφωνώ Λίγο» και «Συμφωνώ απόλυτα».

Παράδειγμα διακριτής κλίμακας είναι ο δείκτης MedDietScore που σχετίζεται με την υγεία (*Panagiotakos, et al., 2007*) που μετράει το βαθμό προσκόλλησης στο πρότυπο της Μεσογειακής Διατροφής, Ακολούθως παρατίθεται η κλίμακα μέτρησης της κατάθλιψης (*Zung, 1965*) με 20 διακριτές/ιεραρχικές συνιστώσες μεταβλητές και 4 διαμερίσεις η κάθε μια που εκφράζουν τη συχνότητα εμφάνισης ενός συμπτώματος (Πίνακας 1.3.2.1). Το συνολικό σκορ της κλίμακας αυτής κυμαίνεται από 20 μέχρι 80 μονάδες.

Δείκτης Μέτρησης Κατάθλιψης

Δείκτης Μέτρησης Κατάθλιψης T_{20}^{4} (Zung 1965)

	Συμπτώματα	Συχνότητα Εμφάνισης Συμπτώματος - k=4 διαμερίσεις			
	m=20 διακριτές μεταβλητές	ΚΑΘΟΛΟΥ	ΜΕΡΙΚΕΣ ΦΟΡΕΣ	ΣΥΧΝΑ	ΠΑΝΤΟΤΕ
X_1^4	Αισθάνομαι αποκαρδιωμένος ή λυπημένος	1	2	3	4
X_2^4	Το πρωί νοιώθω καλύτερα από κάθε ώρα της ημέρας	4	3	2	1
X_3^4	Κλαίω εύκολα ή νοιώθω έτοιμος να κλάψω	1	2	3	4
X_4^4	Έχω προβλήματα ύπνου το βράδυ	1	2	3	4
X_5^4	Τρώω όσο συνήθιζα προηγουμένως	4	3	2	1
x_6^4	Εξακολουθώ να ενδιαφέρομαι για το σεξ	4	3	2	1
x_7^4	Παρατηρώ ότι χάνω βάρος	1	2	3	4
X_8^4	Έχω δυσκοιλιότητα	1	2	3	4
x_9^4	Έχω ταχυπαλμίες	1	2	3	4
x_{10}^4	Κουράζομαι χωρίς ιδιαίτερο λόγο	1	2	3	4
x_{11}^4	Το μυαλό μου είναι τόσο καθαρό όσο και πριν	4	3	2	1
x_{12}^4	Μου είναι εύκολο να κάνω τα πράγματα που συνήθιζα να κάνω και πριν	4	3	2	1
x_{13}^4	Έχω ανησυχία και δεν μπορώ να ηρεμήσω	1	2	3	4
X_{14}^4	Έχω αισιοδοξία για το μέλλον μου	4	3	2	1
X_{15}^4	Έχω περισσότερη νευρικότητα από πριν	1	2	3	4
x_{16}^4	Παίρνω αποφάσεις το ίδιο εύκολα όπως και πριν	4	3	2	1
X_{17}^4	Αισθάνομαι χρήσιμος και απαραίτητος	4	3	2	1
x_{18}^4	Η ζωή μου είναι αρκετά «γεμάτη»	4	3	2	1
x_{19}^4	Αισθάνομαι ότι θα ήταν καλύτερα για τους άλλους αν πέθαινα	1	2	3	4
x_{20}^4	Με ευχαριστούν ακόμα και τα πράγματα που με ευχαριστούσαν και προηγουμένως	4	3	2	1
Συνολικό σκορ:	$T_{20}^{4} = \sum_{i=1}^{20} X_i^4$	Ελάχιστη Τιμή Δείκτη:	$\min\left(T_{20}^{4}\right) = 20$		
		Μέγιστη Τιμή Δείκτη:	$\max\left(T_{20}^{4}\right) = 80$		

Πίνακας 1.3.2.1: Κλίμακα (αυτοαξιολόγησης) μέτρησης κατάθλιψης (Zung, 1965)

1.3.3 Μεθοδολογία Κατασκευής Σύνθετων Κλιμάκων

Στην πράξη, μια σύνθετη διακριτή κλίμακα T_m^k δημιουργείται από το άθροισμα των *m* συνιστωσών μεταβλητών $\boldsymbol{X^k}$. Έτσι, μια σύνθετη (μη σταθμισμένος) κλίμακα T_m^k έχει την ακόλουθη μαθηματική μορφή:

$$T_m^k = \sum_{j=1}^{m} X_j^k \qquad (1.3.2)$$

Οι *m* συνιστώσες μεταβλητές αντανακλούν διαφορετικές πτυχές της κατάστασης ενός ατόμου. Κάθε μία από τις *m* συνιστώσες X_i, $i=1, 2,\ldots, m$ του διανύσματος $\underline{\boldsymbol{X}}=(X_1, X_2, \ldots, X_m)$ εκφράζει διαφορετικές διαστάσεις χαρακτηριστικών ενός ατόμου και η τιμή τους παρέχει αθροιστικά την τιμή της αντίστοιχης σύνθετης κλίμακας. Κάθε συνιστώσα μεταβλητή ακολουθεί συνήθως μια άγνωστη κατανομή. Επομένως, για κάθε μία από τις *m* συνιστώσες X_i, $i=1, 2,\ldots, m$ ισχύει $X_i \sim F_i$,

$i=1,2,\ldots,m$ όπου F_i συμβολίζει την κατανομή της εκάστοτε συνιστώσας μεταβλητής X_i. Η πραγματική κατάσταση ενός ατόμου χαρακτηρίζεται από μια τυχαία διτιμη μεταβλητή Y που θεωρείται ότι ακολουθεί μια κατανομή Bernoulli ($Y \sim B(n,p)$), όπου το n αντιπροσωπεύει το πλήθος των ατόμων και το p αντιπροσωπεύει την πιθανότητα παρουσίας μιας νόσου (επιπολασμός) και το 1-p αντιπροσωπεύει την πιθανότητα της απουσίας της νόσου (η τιμή Y=0 εκφράζει μια υγιή κλινική κατάσταση και η τιμή Y=1 μια μη υγιή κλινική κατάσταση). Μια επέκταση του προαναφερθέντος σκεπτικού θα μπορούσε να χρησιμοποιηθεί εάν χρειάζονταν περισσότερες από δύο κλινικές καταστάσεις για να περιγραφεί πλήρως ένα άτομο με χρήση της πολυωνυμικής κατανομής (π.χ., όπως στην περίπτωση σακχαρώδους διαβήτη, όπου υπάρχει μια ενδιάμεση κατάσταση που ονομάζεται προ-διαβήτης (*Lars Rydén Et al., 2013*).

Η κατασκευή αυτών των εργαλείων, των κλιμάκων είναι κατά κύριο λόγο αυθαίρετη και από μεθοδολογικών σκοπιάς υπάρχουν αρκετές πτυχές που προβληματίζουν ένα ερευνητή κατά τη διαδικασία κατασκευής τους, όπως είναι οι ακόλουθες:

- Η επιλογή των κατάλληλων συνιστωσών μεταβλητών (Variables' selection), οι οποίες λογικά πρέπει να σχετίζονται με την έκβαση ώστε να παρέχουν κάποιο είδος ποσοτικής ή/και ποιοτικής πληροφορίας.
- Η κατάλληλη βαθμονόμηση των συνιστωσών μεταβλητών (Variables' scoring) ώστε να λαμβάνουν κατάλληλο σκορ τιμές διαμερίσεων με ευθεία ή αντίστροφη σχέση με το αποτέλεσμα υγείας που η κλίμακα αποτιμά.
- Η εύρεση του κατάλληλου εύρους στηρίγματος των συνιστωσών μεταβλητών, ώστε σε αυτό το μέρος της αντίστοιχη κλίμακας να γίνεται δυνατή η διακριτική ιδιότητα των συνιστωσών μεταβλητών.
- Η κατάλληλη στάθμιση των συνιστωσών μεταβλητών (Variables' weighting) ώστε να αποδίδεται μεγαλύτερη βαρύτητα σε εκείνες τις μεταβλητές που σχετίζονται εντονότερα με το χαρακτηριστικό.

Το τελικό αριθμητικό αποτέλεσμα (σκορ) αποτελεί μια σύνθετη βαθμολογία (*Panagiotakos, 2009; Streiner & Norman, 2008*) όπου συγκρινόμενο με το διαγνωστικό όριο παρέχει την επιθυμητή πληροφορία. Παρά την ευρεία χρήση των σύνθετων κλιμάκων, δεν υπάρχει σαφές και συγκεκριμένο μεθοδολογικό πλαίσιο για την κατασκευή τους και βασίζεται κυρίως στον υποκειμενισμό του εκάστοτε ερευνητή καθώς επίσης βέβαια και στη φύση του προς εκτίμηση χαρακτηριστικού. Ως

αποτέλεσμα των παραπάνω αυθαίρετων επιλογών, αμφισβητείται η ακρίβεια αυτών των εργαλείων από την άποψη της ευαισθησίας, δηλαδή του ορθού εντοπισμού π.χ. ενός ασθενούς από τον πληθυσμό των ασθενών και από την άποψη της ειδικότητας, δηλαδή του ορθού εντοπισμού του υγιούς από τον πληθυσμό των υγιών, με ορθό εγκλεισμό ή αποκλεισμό υγιών ή ασθενών από ένα γενικό πληθυσμό. Όπως αναφέρθηκε σε προηγούμενη παράγραφο, δεν υπάρχει μια σταθερή και συνεπής μεθοδολογία για την κατασκευή σύνθετων κλιμάκων προκειμένου να μεγιστοποιηθεί η ακρίβεια και η διαγνωστική τους ικανότητα, οπότε ένα γενικότερο πλαίσιο με κανόνες που διέπουν την κατασκευή σύνθετων κλιμάκων, είναι απαραίτητο.

1.3.3.1 Αριθμός και Είδος Συνιστωσών Μεταβλητών

Η επιλογή των συνιστωσών του δείκτη πραγματοποιείται, κατά περίπτωση, από τους κατάλληλους ερευνητές ανάλογα με τη φύση του χαρακτηριστικού που η κλίμακα στοχεύει να ερευνήσει και επιλέγονται προσεκτικά μετά από κατάλληλη ανασκόπηση της βιβλιογραφίας (*Streiner & Norman, 2008*). Σύμφωνα με τους Kourlaba και Panagiotakos (2009), οι συνιστώσες μεταβλητές μιας κλίμακας θα πρέπει να συσχετιστούν σημαντικά με την μεταβλητή που εκφράζει την πραγματική κατάσταση/έκβαση (*Kourlaba & Panagiotakos, 2009b*) και να έχουν τη χαμηλότερη δυνατή συσχέτιση μεταξύ τους.

Βασική αρχή είναι ότι οι μεταβλητές που θα χρησιμοποιηθούν ως συνιστώσες των κλιμάκων, θα πρέπει να είναι σε θέση να διακρίνουν, σε κάποιο βαθμό, π.χ. τους ασθενείς από τους υγιείς, όσον αφορά την πραγματική κατάσταση της υγείας που η κλίμακα σκοπεύει να αξιολογήσει. Αυτό σημαίνει ότι η επιλογή των συνιστωσών μεταβλητών πρέπει να γίνεται με μεγάλη προσοχή και να εφαρμόζεται πιλοτικά ώστε να αναθεωρείται η χρήση κάποιων μεταβλητών ή/και να επιβεβαιώνεται η χρήση κάποιων άλλων. Επομένως, ανάλογα με το τι η κλίμακα σκοπεύει να εκτιμήσει, διερευνήσει και μετρήσει, πρέπει οι αντίστοιχες συνιστώσες μεταβλητές να παρέχουν διακριτή πληροφορία που αθροιστικά να είναι αξιοποιήσιμη.

1.3.3.2 Εύρος Στηρίγματος Συνιστωσών Μεταβλητών των Κλιμάκων

Επόμενο βήμα στην κατασκευή μιας κλίμακας είναι η εύρεση του κατάλληλου εύρους του στηρίγματος των συνιστωσών μεταβλητών. Ένα ερώτημα που έχει τεθεί παραπάνω και πρέπει να διερευνηθεί είναι εάν το εύρος του στηρίγματος σχετίζεται θετικά με την ευαισθησία μιας κλίμακας. Πιο συγκεκριμένα, στην περίπτωση των

διακριτών συνιστωσών μεταβλητών, αν ο αριθμός των διαμερίσεων των συνιστωσών μεταβλητών μιας διακριτής κλίμακας σχετίζεται θετικά με τη διαγνωστική ικανότητα του, και στην περίπτωση των συνεχών μεταβλητών, αν το εύρος του πεδίου ορισμού των συνιστωσών μεταβλητών μιας συνεχούς κλίμακας σχετίζεται θετικά με τη διαγνωστική ικανότητα του. Για παράδειγμα, στην περίπτωση της κλίμακας αυτοαξιολόγησης της κατάθλιψης του Zung, αν ο αριθμός των διαμερίσεων ήταν μεγαλύτερος, η ικανότητα ανίχνευσης ασθενών, από πλευράς του δείκτη θα αυξάνονταν; Επίσης, στην περίπτωση της κλίμακας VAS, αν το μήκος του ευθυγράμμου τμήματος, στο οποίο απαντά ένας ασθενής, αυξηθεί, τι συμβαίνει με την ικανότητα του εργαλείου για ανίχνευση του χαρακτηριστικού που αποτιμά; Το ερώτημα αυτό επιχειρείται να επιλυθεί σε αυτή την εργασία. Οι Kourlaba και Panagiotakos έχουν καταλήξει σε κατάλληλο συμπέρασμα για το παραπάνω ερώτημα, χρησιμοποιώντας προσομοιωμένα δεδομένα. Πιο συγκεκριμένα, έχει αποδειχθεί ότι για τη μεγιστοποίηση της ακρίβειας μιας κλίμακας πρέπει να χρησιμοποιούνται συνεχείς συνιστώσες μεταβλητές (*Kourlaba & Panagiotakos, 2010*). Σε πρόσφατη δημοσίευση, η οποία βασίστηκε σε προσομοιώσεις, προέκυψε ότι η ευαισθησία μιας κλίμακας είναι μια αύξουσα (μη φθίνουσα) συνάρτηση των διαμερίσεων της κλίμακας του (*Kourlaba & Panagiotakos,2009*). Επιπροσθέτως, βασιζόμενοι σε θεωρητικές αποδείξεις, έχει αποδειχθεί ότι η ευαισθησία μιας κλίμακας είναι αύξουσα συνάρτηση ως προς τον αριθμό των διαμερίσεων των συνιστωσών του (*Bersimis et al,2013*). Στο σημείο αυτό είναι απαραίτητο να δοθεί ο ορισμός της διαμέρισης του στηρίγματος μιας διακριτής μεταβλητής.

Ορισμός 1.1: Ως *k*-διαμέριση ενός συνόλου *A* θεωρείται ένα σύνολο *k* συνόλων $\{A_1, A_2, ..., A_k\}$ ανά δύο ξένα μεταξύ τους, μη μηδενικά υποσύνολα του *A* και η ένωσή τους είναι το *A* (*Charalambides, 2002*). Η μαθηματική έκφραση της διαμέρισης είναι η ακόλουθη:

$$A_i \subseteq A,\ A_i \neq \varnothing,\ i = 1,2,...,k, \quad A_i \bigcap A_j = \varnothing, \quad i, j = 1,2,...,k,\ i \neq j\ , \bigcup_{i=1}^{k} A_i = A.$$

Στην περίπτωση των συνεχών συνιστωσών μεταβλητών, έχει διερευνηθεί αν το εύρος του στηρίγματος των συνιστωσών μεταβλητών μιας συνεχούς μιας κλίμακας σχετίζεται θετικά με την ακρίβεια του, από τους Bersimis, Panagiotakos και Vamvakari (*Bersimis et al,2015*) και το συμπέρασμα εκφράζει ότι η ευαισθησία μιας κλίμακας είναι μια αύξουσα συνάρτηση ως προς το εύρος του στηρίγματος των συνιστωσών μεταβλητών. Οι συνεχείς κλίμακες παρέχουν μεγαλύτερη ακρίβεια σε

σχέση με τις διακριτές, όμως οι διακριτές κλίμακες προτιμούνται στην καθημερινή κλινική πράξη λόγω της ευκολίας χρήσης της.

1.3.3.3 Βαθμονόμηση Συνιστωσών Μεταβλητών

Η βαθμονόμηση των διαμερίσεων των συνιστωσών μεταβλητών χρησιμοποιείται στην περίπτωση των διακριτών κλιμάκων όπου συγκεκριμένοι βαθμοί αποδίδονται στην κάθε διαμέριση. Η βαθμονόμηση πολλές φορές γίνεται με αυθαίρετο τρόπο, λαμβάνοντας όμως ο κατάλληλος ερευνητής υπόψη του την ένταση της σχέσης μεταξύ της συνιστώσας μεταβλητής με τη μεταβλητή που αντιπροσωπεύει το χαρακτηριστικό που αποτιμά η κλίμακα. Η βαθμονόμηση των διαμερίσεων μπορεί να γίνεται με γραμμικό ή μη γραμμικό τρόπο. Για παράδειγμα, στην περίπτωση μιας ψυχομετρικής κλίμακας που αποσκοπεί στην αξιολόγηση της κατάθλιψης θεωρούμε δύο υποθετικές συνιστώσες που εκφράζονται μέσω των ερωτήσεων: «Πόσο συχνά γελάτε;» και «Πόσο συχνά κλαίτε;» με 5 διαμερίσεις για την κάθε μια, που εκφράζονται μέσω των απαντήσεων «Ποτέ», «1-3 φορές/μήνα», 1-6 φορές/εβδομάδα», «1 φορά/ημέρα» και «>1φορά/ημέρα». Στην κάθε μια απάντηση αποδίδεται ένας βαθμός/σκορ που κυμαίνεται από 1 έως 5 με τη φιλοσοφία ότι μεγάλοι βαθμοί εκφράζουν καταστάσεις που σχετίζονται με τη νόσο της κατάθλιψης. Δηλαδή, στην ερώτηση «Πόσο συχνά γελάτε;», η απάντηση «Ποτέ» θα λάμβανε την τιμή 5 και η απάντηση «>1 φορά τη μέρα» την τιμή 1, ενώ για την ερώτηση «Πόσο συχνά κλαίτε;» θα αντιστοιχούσαμε αντίστροφη βαθμονόμηση. Έτσι, αθροίζοντας τους βαθμούς των επιμέρους συνιστωσών με την προαναφερθείσα φιλοσοφία, προκύπτει το συνολικό σκορ της κλίμακας, του οποίου μεγάλες τιμές αποτελούν ένδειξη ύπαρξης κάποιας μορφής κατάθλιψης.

Υπάρχει περίπτωση βαθμονόμησης σε διακριτές συνιστώσες μεταβλητές που να μην ακολουθεί γραμμική ομόρροπη ή αντίρροπη σχέση με την συχνότητα που εκφράζεται μέσω των δυνατών απαντήσεων σε μια ερώτηση. Για παράδειγμα, η κατανάλωση κόκκινου κρασιού είναι ευεργετικότερη από την καθόλου κατανάλωση αλλά μόνο σε περιπτώσεις που δεν γίνεται κατάχρηση. Αυτό σημαίνει ότι δεν υπάρχει γραμμική σχέση, αλλά καμπυλόγραμμη, μεταξύ της κατανάλωσης κόκκινου κρασιού και της ωφέλειας στην υγεία του καταναλωτή. Σε αυτή την περίπτωση θα μπορούσε να γίνει η ακόλουθη βαθμονόμηση: «κανένα ποτήρι/ημέρα» να λάβει βαθμό 1, «1-2

ποτήρια/ημέρα» να λάβει βαθμό 2 και «>3 ποτήρια/ημέρα» να λάβει βαθμό 0, με υψηλές τιμές να εκφράζουν θετικό αποτέλεσμα υγείας.

1.3.3.4 Στάθμιση Συνιστωσών Μεταβλητών

Στη βιβλιογραφία έχουν προταθεί ποικίλες μέθοδοι που συνδυάζουν μεμονωμένες μεταβλητές σε μια σύνθετη κλίμακα με βάση στατιστικές μεθόδους προκειμένου να αυξηθεί η διαγνωστική ακρίβεια των δεικτών (*Pepe & Thompson, 2000*), όπως η μεγιστοποίηση της συνάρτησης πιθανοφάνειας που αντιστοιχεί στα νοσούντα και υγιή άτομα (*Qin & Zhang, 2010; Peuc, Cai, & Longton, 2006*), η μεγιστοποίηση της συνάρτησης που αντιστοιχεί στην περιοχή (AUC) κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (ROC) *(Pepe, Cai, & Longton, 2006*), η χρήση κατάλληλης γραμμικής διακριτικής συνάρτησης (*Su & Liu, 1993*) και η χρήση τροποποιημένης λογιστικής παλινδρόμησης βασιζόμενη στη θεωρία αποφάσεων (*Yuan & Ghosh, 2008*). Επίσης, εφαρμόζονται τεχνικές μείωσης της διαστασιμότητας χωρίς να γίνεται υπόθεση για συγκεκριμένη κατανομή (*Pfeiffer & Bur, 2008*) ή τεχνικές βελτιστοποίησης της συνάρτησης κινδύνου (*McIntosh & Pepe, 2002*).

Στην παρούσα διδακτορική διατριβή οι συντελεστές στάθμισης *w* που χρησιμοποιήθηκαν για την κατασκευή των προτεινόμενων σταθμισμένων κλιμάκων προέρχονται από δύο ευρέως χρησιμοποιούμενες στατιστικές τεχνικές που ως στόχο έχουν το διαχωρισμό παρατηρήσεων σε δύο ομάδες. Οι τεχνικές αυτές είναι η Λογιστική Παλινδρόμηση και η Διαχωριστική Ανάλυση. Με τη χρήση ειδικών σταθμίσεων παρήχθησαν δείκτες που δίνονται από τον παρακάτω τύπο:

$$\mathrm{T}_i = \sum_{j=1}^{m} w_{ij} X_j, \quad j = 1, 2, ..., m, \quad i = 0, 1, 2, ..., 8$$

όπου w_{ij}, είναι, για παράδειγμα, οι σχετικοί λόγοι πιθανοτήτων από τη λογιστική παλινδρόμηση ή οι μη τυποποιημένοι συντελεστές της διαχωριστικής συνάρτησης από τη διαχωριστική ανάλυση. Η λογική που κρύβεται πίσω από τους προτεινόμενους σταθμισμένους δείκτες είναι ότι οι μεταβλητές που διαχωρίζουν με μεγαλύτερη σαφήνεια τους υγιείς από τους ασθενείς, θα συνεισφέρουν περισσότερο στο τελικό αποτέλεσμα (σκορ) του δείκτη. Οι σταθμισμένες κλίμακες που χρησιμοποιήθηκαν στην παρούσα διατριβή παρουσιάζονται αναλυτικά σε επόμενη παράγραφο στο κεφάλαιο της μεθοδολογίας. Παρουσιάζεται η μεθοδολογία εύρεσης των σταθμίσεων με χρήση παραδείγματος από προσομοιωμένα και από πραγματικά δεδομένα έρευνας που διενεργήθηκε στην περιοχή της Αττικής (ATTICA Study).

1.3.4 Τιμή Κλίμακας (Score) - Προβληματισμοί Κατασκευής Κλιμάκων

Η αριθμητική τιμή που λαμβάνει μια κλίμακα προκύπτει από το αστάθμιστο ή σταθμισμένο άθροισμα των συνιστωσών μεταβλητών του όπως παρατίθεται στις σχέσεις (1.3.4.1) και (1.3.4.2):

$$T_m = \sum_{i=1}^{m} X_i \text{ , } i=1,2,\ldots,m \qquad (1.3.4.1)$$

$$T_m = \sum_{i=1}^{m} w_i \cdot X_i \text{ , } i=1,2,\ldots,m \qquad (1.3.4.2)$$

Στο παράδειγμα της σύνθετης διατροφικής κλίμακας «MedDietScore» (*Panagiotakos et al, 2007*) που αποτιμά το βαθμό υιοθέτησης της Μεσογειακής διατροφής υπάρχουν 11 συνιστώσες μεταβλητές που εκφράζουν την κατανάλωση 11 ομάδων τροφίμων (φρούτα, λαχανικά, κρέας, ψάρια, γαλακτοκομικά προϊόντα, κ.λ.π). Όπως αναφέρθηκε σε προηγούμενη παράγραφο, η βαθμονόμηση της εκάστοτε μεταβλητής είναι ανάλογη/αντίστοιχη με τη συχνότητα κατανάλωσης (π.χ. ποτέ, σπάνια, 1-2 φορές το μήνα κ.λπ.) και εάν η ομάδα των τροφίμων είναι ευεργετική για την υγεία έχει ομόρροπη κλίμακα από 0 έως 5 με τη συχνότητα κατανάλωσης και από 5 έως 0 εάν συμβαίνει το αντίθετο (π.χ. κόκκινο κρέας). Στο παράδειγμα του «MedDietScore», υψηλή τιμή αυτής της διατροφικής κλίμακας για ένα άτομο σημαίνει ότι, αυτό το άτομο είναι κοντά στην τήρηση του παραδοσιακού μοτίβου της μεσογειακής διατροφής και έτσι διατρέχει μικρότερο κίνδυνο καρδιαγγειακής πάθησης. Συνεπώς, η «MedDietScore» είναι μια σύνθετη κλίμακα σχετικός με την υγεία με m=11 συνιστώσες κάθε μία από τις οποίες έχει k=6 διαιρέσεις και αθροιζόμενα τα επιμέρους σκορ της κάθε συνιστώσας παράγουν το συνολικό σκορ της σύνθετης κλίμακας.

Άλλο παράδειγμα αποτελεί η διακριτή κλίμακα του Zung (*Zung, 1965*) που είναι μια κλίμακα που αποτιμά το βαθμό κατάθλιψης ενός ατόμου (με αυτοαναφορά) και αποτελείται από 20 ερωτήσεις/στάσεις (διακριτές διατακτικές μεταβλητές) όπως π.χ. «Αισθάνομαι αποκαρδιωμένος ή λυπημένος», κλπ). Ο αριθμός των δυνατών απαντήσεων, που στην περίπτωση αυτή είναι 4, εκφράζει το πλήθος των διαμερίσεων της αντίστοιχης μεταβλητής. Η απάντηση καθεμιάς συνιστώσας μεταβλητής βαθμονομείται κατάλληλα με τιμή (σκορ) από 1 έως 4, ανάλογα με τη συχνότητα εμφάνισης του αντίστοιχου συμπτώματος (καθόλου, μερικές φορές, συχνά, πάντοτε).

Η συνολική τιμή της κλίμακας προκύπτει από το άθροισμα των επιμέρους τιμών που λαμβάνουν οι συνιστώσες μεταβλητές. Μεγάλες τιμές της κλίμακας για ένα άτομο, εκφράζουν ότι το άτομο αυτό πιθανόν να πάσχει από κατάθλιψη. Επομένως, το συνολικό σκορ που προκύπτει από τις επιμέρους μεταβλητές που εκφράζουν διαφορετικές διαστάσεις ενός χαρακτηριστικού, παράγει ποσοτική ή/και ποιοτική πληροφορία για το χαρακτηριστικό που η αντίστοιχη της κλίμακας αποσκοπεί να αποτιμήσει.

Υπάρχουν παραδείγματα κλιμάκων υγείας, όπως είναι η οπτική αναλογική κλίμακα μέτρησης του πόνου VAS (Visual Analogous Scale) όπου η κλίμακα είναι συνεχής και αναπαρίσταται από ένα ευθύγραμμο τμήμα (*Hawker et al, 2011*) ώστε το άτομο που συμπληρώνει τη φόρμα να κατατάσσει εντελώς ελεύθερα την κατάσταση του σε ένα αδιάστατο ευθύγραμμο τμήμα. Σε αυτή την περίπτωση, γίνεται ομοίως η άθροιση των επιμέρους σκορ, με τη διαφορά ότι οι επιμέρους τιμές δύναται να είναι πραγματικοί αριθμοί, δηλαδή να λαμβάνουν δεκαδικές τιμές. Η τελική τιμή μιας κλίμακας είναι πιθανόν δεκαδικός αριθμός, που ναι μεν παρέχει μεγάλη ακρίβεια, αλλά πολύ συχνά είναι δύσκολα χρηστική λόγω της δυσκολίας υπολογισμών με δεκαδικούς αριθμούς.

Το σκορ της κλίμακας είναι καθοριστικής σημασίας για λόγους πρόβλεψης ή/και ανίχνευσης κάποιου χαρακτηριστικού, επομένως είναι πολύ σημαντικό να υπολογίζεται με έγκυρα μαθηματικό τρόπο μέσω ενός αλγεβρικού αθροίσματος των επιμέρους τιμών των συνιστωσών μεταβλητών. Το σκορ της κλίμακας εξαρτάται από τον αριθμό και το είδος των συνιστωσών μεταβλητών, από το εύρος του στηρίγματος και τη βαθμονόμηση των συνιστωσών μεταβλητών και από την πιθανή χρήση σταθμίσεων, δηλαδή εξαρτάται από όλους εκείνους τους παράγοντες που εξετάζονται σε αυτή τη διατριβή.

1.3.5 Αξιολόγηση Διαχωριστικής Ικανότητας Κλιμάκων

Σε διχοτομικά/δυαδικά προβλήματα ταξινόμησης ενός ατόμου ως προς την κλινική του κατάσταση, η διαγνωστική ικανότητα μιας κλίμακας, αξιολογείται χρησιμοποιώντας διάφορα μέτρα/κριτήρια όπως είναι το αληθώς θετικό ποσοστό (TPR), δηλαδή η ευαισθησία S_e, το αληθώς αρνητικό ποσοστό (TNR), δηλαδή η ειδικότητα S_p, τη θετική προβλεπτική αξία (PPV), την αρνητική προβλεπτική αξία (NPV), και το πραγματικό ποσοστό (TR), δηλαδή την ακρίβεια A_c, σε ένα

συγκεκριμένο διαχωριστικό σημείο (σημείο αποκοπής) (*Baron, 1994*). Επιπλέον, χρησιμοποιείται η καμπύλη λειτουργικών χαρακτηριστικών (ROC) που λαμβάνει υπόψη τόσο την ευαισθησία (TPR) όσο και την ειδικότητα (TNR) χρησιμοποιώντας την περιοχή (AUC) κάτω από την καμπύλη (*Hanley, 1982*). Όπως είναι γνωστό, η ευαισθησία μιας κλίμακας υγείας υποδηλώνει την ικανότητα της κλίμακας να εντοπίζει αληθώς παθολογικά περιστατικά, η ειδικότητα υποδηλώνει την ικανότητα της κλίμακας να εντοπίζει πραγματικά φυσιολογικά (υγιή) περιστατικά και η ακρίβεια υποδηλώνει την ικανότητα της κλίμακας να εντοπίζει αληθινά περιστατικά γενικά (*Wayne, 1995, Fletcher & Fletcher, 2005*).

Έστω, τυχαίο μικτό δείγμα από ασθενείς και μη ασθενείς που ταξινομούνται από μια κλίμακα υγείας με τη βοήθεια του πίνακα 1.3.5.1. Ο 2x2 πίνακας συνάφειας περιλαμβάνει:

• Αληθώς χαρακτηρισμένες περιπτώσεις από την κλίμακα: Οι αληθώς θετικές περιπτώσεις (ΑΘ) και οι αληθώς αρνητικές περιπτώσεις (ΑΑ).

• Ψευδώς (εσφαλμένα) χαρακτηρισμένες περιπτώσεις από την κλίμακα: Οι ψευδώς θετικές περιπτώσεις (ΨΘ) και οι ψευδώς αρνητικές περιπτώσεις (ΨΑ).

		Πραγματική Κλινική Κατάσταση		
		Θετικός (Ασθενής: Y=1)	Αρνητικός (Υγιής: Y=0)	
Ταξινόμηση ατόμου με χρήση της κλίμακας	Θετικός (Ασθενής)	α Αληθώς Θετικές Περιπτώσεις (ΑΘ) True Positive (TP)	β Ψευδώς Θετικές Περιπτώσεις (ΨΘ) False Positive (FP)	α+β
	Αρνητικός (Υγιής)	γ Ψευδώς Αρνητικές Περιπτώσεις (ΨΑ) False Negative (FN)	δ Αληθώς Αρνητικές Περιπτώσεις (ΑΑ) True Negative (TN)	γ+δ
		α+γ	β+δ	α+β+γ+δ=N

Πίνακας 1.3.5.1: Ταξινόμηση υγιών και ασθενών με χρήση του δείκτη ως προς τυχαία ασθένεια

Βασιζόμενοι στον πίνακα 1.3.5.1, θα δοθούν στις επόμενες παραγράφους οι ορισμοί της Ευαισθησίας, της Ειδικότητας, της Ακρίβειας, της Θετικής Προγνωστικής Αξίας, της Αρνητικής Προγνωστικής Αξίας και του Εμβαδού της επιφάνειας κάτω από την καμπύλη λειτουργικών χαρακτηριστικών.

1.3.5.1 Ευαισθησία Κλιμάκων

Η ευαισθησία σε μια κλινική δοκιμή ή μιας κλίμακας σχετικής με την υγεία είναι η πιθανότητα ενός θετικού αποτελέσματος μιας εξέτασης, με δεδομένη την παρουσία του συμπτώματος μιας ασθένειας (*Wayne, 1995*). Σε γενικές γραμμές, η

ευαισθησία μετρά το ποσοστό των θετικών πραγματικών (άρρωστων) ατόμων που προσδιορίζονται ως τέτοια από το αποτέλεσμα της κλινικής δοκιμής και εκφράζει την πιθανότητα ορθής ταξινόμησης ενός ασθενούς ατόμου στον πληθυσμό των ασθενών.

Το αληθές θετικό ποσοστό (ευαισθησία) δίνεται από την ακόλουθη δεσμευμένη πιθανότητα ως εξής:

$$S_e\left(T_m\right) = TPR = P\left(T_m > c \mid Y = 1\right) = \frac{TP}{TP + FN} = \frac{\text{ΑΘ}}{\text{ΑΘ} + \text{ΨΑ}} \qquad (1.3.5.1)$$

Η συνάρτηση ευαισθησίας είναι ένα μέτρο της διαγνωστικής ικανότητας της κλίμακας που σχετίζεται με την υγεία, για την ανίχνευση των ασθενών (αληθώς θετικών). Όταν αυξάνει η συνάρτηση ευαισθησίας, οδηγεί στη μείωση της συνάρτησης ειδικότητας, αλλάζοντας τη θέση του κατάλληλου διαγνωστικού κατωφλίου.

1.3.5.2 Ειδικότητα Κλιμάκων

Η ειδικότητα σε μια κλινική δοκιμή ή μιας κλίμακας υγείας είναι η πιθανότητα ενός αρνητικού αποτελέσματος μιας εξέτασης, με δεδομένη την απουσία του συμπτώματος μιας ασθένειας (*Wayne, 1995*). Σε γενικές γραμμές, η ειδικότητα μετρά το ποσοστό των αρνητικών πραγματικών (υγιών) ατόμων που προσδιορίζονται ως τέτοια από το αποτέλεσμα της κλινικής δοκιμής και εκφράζει την πιθανότητα ορθής ταξινόμησης ενός υγιούς ατόμου στον πληθυσμό των υγιών.

Το αληθές αρνητικό ποσοστό (ειδικότητα) δίνεται από τη δεσμευμένη πιθανότητα ως εξής:

$$S_p\left(T_m\right) = TNR = P\left(T_m < c \mid Y = 0\right) = \frac{TN}{TN + FP} = \frac{\text{ΑΑ}}{\text{ΑΑ} + \text{ΨΘ}} \qquad (1.3.5.2)$$

Η συνάρτηση ειδικότητας είναι ένα μέτρο της διαγνωστικής ικανότητας της κλίμακας που σχετίζεται με την υγεία, για την ανίχνευση των υγιών (αληθώς αρνητικών). Όταν αυξάνει η συνάρτηση ειδικότητας, οδηγεί στη μείωση της συνάρτησης ευαισθησίας, αλλάζοντας τη θέση του κατάλληλου διαγνωστικού κατωφλίου.

1.3.5.3 Ακρίβεια Κλιμάκων

Η ακρίβεια σε μια κλινική δοκιμή είναι η πιθανότητα ενός ορθού (θετικού ή αρνητικού) αποτελέσματος μιας εξέτασης, με δεδομένη την παρουσία ή απουσία του συμπτώματος μιας ασθένειας αντίστοιχα (*Wayne, 1995*). Σε γενικές γραμμές, η

ακρίβεια μετρά το ποσοστό των πραγματικών περιπτώσεων που προσδιορίζονται ως τέτοια από το αποτέλεσμα της κλινικής δοκιμής και εκφράζει το πιθανότητα ορθής ταξινόμησης του ατόμου από την κλίμακα υγείας.

Το αληθές ποσοστό δίνεται από την πιθανότητα της ένωσης δεσμευμένων ενδεχομένων ως εξής:

$$A_c(T_m) = TR = P\left[(T_m > c \mid Y=1) \cup (T_m < c \mid Y=0)\right] = \frac{TP+TN}{TP+FP+TN+FN} = \frac{\text{ΑΘ}+\text{ΑΑ}}{\text{ΑΘ}+\text{ΨΘ}+\text{ΑΑ}+\text{ΨΑ}} \quad (1.3.5.3)$$

Η συνάρτηση ακρίβειας είναι ένα μέτρο της διαγνωστικής ικανότητας της κλίμακας που σχετίζεται με την υγεία, για την ανίχνευση των αληθινών περιπτώσεων (αληθώς θετικών ή αληθώς αρνητικών).

1.3.5.4 Θετική Προγνωστική Αξία Κλιμάκων

Η θετική προγνωστική αξία σε μια κλινική δοκιμή ή μιας κλίμακας υγείας είναι η πιθανότητα παρουσίας του συμπτώματος μιας ασθένειας με δεδομένο το θετικό αποτέλεσμα της κλίμακας ή της εξέτασης (*Wayne, 1995*). Σε γενικές γραμμές, η θετική προγνωστική αξία μετρά το ποσοστό των θετικών πραγματικών (ασθενών) ατόμων προς το σύνολο των ατόμων που προσδιορίζονται ως τέτοια από το αποτέλεσμα της κλινικής δοκιμής και εκφράζει την πιθανότητα ορθής ταξινόμησης ενός ασθενούς ατόμου στον πληθυσμό των ασθενών που ορίζει η κλίμακα.

Η θετική προγνωστική αξία δίνεται από την παρακάτω δεσμευμένη πιθανότητα ως εξής:

$$P_r(T_m) = PPV = P(Y=1 \mid T_m > c) = \frac{TP}{TP+FP} = \frac{\text{ΑΘ}}{\text{ΑΘ}+\text{ΨΑ}} \quad (1.3.5.4.1)$$

.

1.3.5.5 Αρνητική Προγνωστική Αξία Κλιμάκων

Η αρνητική προγνωστική αξία σε μια κλινική δοκιμή ή μιας κλίμακας υγείας είναι η πιθανότητα απουσίας του συμπτώματος μιας ασθένειας με δεδομένο το αρνητικό αποτέλεσμα της κλίμακας ή της εξέτασης (Wayne, 1995). Σε γενικές γραμμές, η αρνητική προγνωστική αξία μετρά το ποσοστό των αρνητικών πραγματικών (υγιών) ατόμων προς το σύνολο των ατόμων που προσδιορίζονται ως τέτοια από το αποτέλεσμα της κλινικής δοκιμής και εκφράζει την πιθανότητα ορθής ταξινόμησης ενός υγιούς ατόμου στον πληθυσμό των υγιών που ορίζει η κλίμακα.

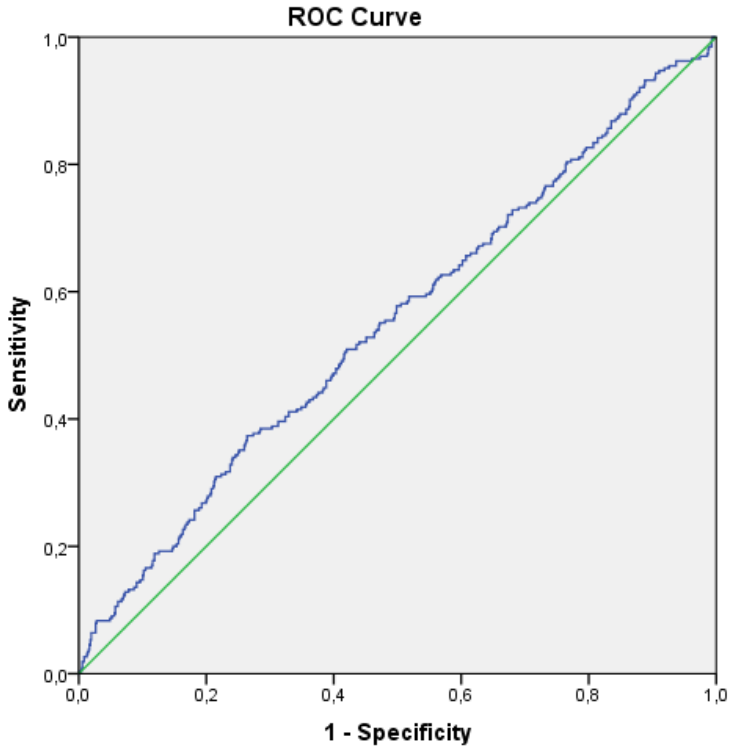
Η αρνητική προγνωστική αξία (αντίστροφη ακρίβεια) δίνεται από την παρακάτω τη δεσμευμένη πιθανότητα ως εξής:

$$IP_r\left(T_m\right) = NPV = P\left(Y = 0 \mid T_m < c\right) = \frac{TN}{TN + FN} = \frac{\text{AA}}{\text{AA} + \Psi\text{A}} \qquad (1.3.5.5.1)$$

1.3.5.6 Επιφάνεια κάτω από την Καμπύλη Λειτουργικών Χαρακτηριστικών

Η καμπύλη λειτουργικών χαρακτηριστικών (Receiver Operating Curve - ROC) χρησιμοποιείται ευρέως για την αξιολόγηση κλιμάκων και διαγνωστικών ελέγχων λαμβάνοντας υπόψη τόσο την ευαισθησία (TPR) όσο και την ειδικότητα (TNR), δηλαδή, η καμπύλη αυτή εκφράζει την επίδοση μιας κλίμακας ή ενός διαγνωστικού ελέγχου και διχοτομεί περιστατικά σε κατηγορίες που περιγράφονται με διωνυμικό τρόπο. Η καμπύλη ROC χρησιμοποιείται για την αξιολόγηση της ικανότητας ταξινόμησης μιας κλίμακας και ορίζεται στο μοναδιαίου εμβαδού τετράγωνο [0,1]x[0,1] του καρτεσιανού επιπέδου με τη γραφική παράσταση της "Ευαισθησίας" (TPR) έναντι του συμπληρώματος της "Ειδικότητας" (1-TNR), δηλαδή αποτελεί τη διαγραμματική απεικόνιση των χαρακτηριστικών ενός ποσοτικού διαγνωστικού τεστ, όπως φαίνεται στο σχήμα 1.3.5.6.1. Επομένως, καμπύλη ROC σχηματίζεται από τα σημεία με συντεταγμένες (TPR, 1-TNR), δηλαδή από τα σημεία με τετμημένη την ευαισθησία και τεταγμένη τη συμπληρωματική πιθανότητα της ειδικότητας, για όλα τα δυνατά διαγνωστικά κατώφλια/όρια (cutoff points). Με άλλα λόγια, μια καμπύλη ROC παρουσιάζει διαγραμματικά το πώς διαφοροποιείται η ευαισθησία όταν μεταβάλλεται η ειδικότητα, και αντιστρόφως. Τα σημεία με συντεταγμένες (TPR, 1-TNR) της καμπύλης ROC που βρίσκονται κοντά στην άνω αριστερή γωνία του διαγράμματος θεωρούνται τα πιο κατάλληλα σημεία διαχωρισμού καθώς αντιστοιχούν σε υψηλό αριθμό αληθώς θετικών περιπτώσεων και υψηλό αριθμό αληθώς αρνητικών περιπτώσεων (χαμηλό αριθμό ψευδών θετικών περιπτώσεων). Όσο μικρότερη είναι η απόσταση της καμπύλης στην άνω αριστερή γωνία ή όσο μεγαλύτερη είναι η απόσταση της από την διχοτόμο της κάτω αριστερής γωνίας, τόσο ικανότερος είναι ο δείκτης υγείας ή ο διαγνωστικός έλεγχος. Στην υποθετική περίπτωση που εμφανίζεται ορθή γωνία, τότε επιτυγχάνεται τέλεια διακριτική ικανότητα της κλίμακας ή του διαγνωστικού ελέγχου (τέλεια δοκιμασία) και ευαισθησία – ειδικότητα λαμβάνουν τιμή 100%. Σημαντικό στοιχείο στην ανάλυση της καμπύλης λειτουργικών χαρακτηριστικών (ROC) είναι η περιοχή κάτω από την καμπύλη αυτή, η οποία είναι γνωστή ως "Περιοχή/εμβαδόν κάτω από την καμπύλη λειτουργικών χαρακτηριστικών " (Area Under Curve - AUC) και είναι μια ποσοτική έκφραση του βαθμού στον οποίο η κλίμακα διακρίνει τα νοσούντα και μη

νοσούντα άτομα (*Powers, 2011*). Επομένως, το εμβαδόν (AUC) αποτελεί κριτήριο για τη διακριτική ικανότητα της κλίμακας μεταξύ ασθενών – υγειών όσο μεγαλύτερη είναι η τιμή του, τόσο μεγαλύτερη είναι η ακρίβεια του διαγνωστικού ελέγχου ως προς την ανίχνευση ασθενών, ή μη, ατόμων. Η μέγιστη τιμή AUC είναι επιθυμητή διότι όσο μεγαλύτερη είναι η τιμή του εμβαδού κάτω από την καμπύλη, τόσο μεγαλύτερη είναι η ακρίβεια του διαγνωστικού ελέγχου ως προς την ταξινόμηση ασθενών και υγειών ατόμων.



Σχήμα 1.3.5.6.1: Παράδειγμα καμπύλης λειτουργικών χαρακτηριστικών

Η υποθετική κλίμακα υγείας που θα διέκρινε τέλεια υγιείς και ασθενείς θα αντιστοιχούσε σε εμβαδόν 1 (οι κατανομές ασθενών και υγειών δεν συμπίπτουν) ενώ μια «άχρηστη» κλίμακα υγείας θα αντιστοιχούσε σε εμβαδόν 0,5 (στην περίπτωση αυτή οι κατανομές ασθενών και υγειών συμπίπτουν). Ο υπολογισμός του AUC γίνεται με χρήση παραμετρικής προσέγγισης με αναλυτικό υπολογισμό του εμβαδού της καμπύλης με χρήση του αντίστοιχου ολοκληρώματος ή με χρήση μη παραμετρικής προσέγγισης της πιθανότητας $P(T_{A\Theta}>T_{AA})$, όπου $T_{A\Theta}$ είναι οι τιμές του δείκτη για τις αληθώς θετικές περιπτώσεις, δηλαδή τους ασθενείς, και T_{AA} είναι οι τιμές της κλίμακας για τις αληθώς αρνητικές περιπτώσεις, δηλαδή τους μη ασθενείς με χρήση του τύπου:

$$\frac{1}{n_{\text{ΑΘ}} \cdot n_{\text{ΑΑ}}} \cdot \sum S\left(T_{\text{ΑΘ}}, T_{\text{ΑΑ}}\right) \qquad (1.3.5.6.1)$$

για όλους τους δυνατούς συνδυασμούς ($T_{\text{ΑΘ}}$,$T_{\text{ΑΑ}}$), όπου

$$S\left(T_{\text{ΑΘ}}, T_{\text{ΑΑ}}\right) = \begin{cases} 1, & \text{αν } T_{\text{ΑΘ}} > T_{\text{ΑΑ}} \\ 1/2, & \text{αν } T_{\text{ΑΘ}} = T_{\text{ΑΑ}} \\ 0, & \text{αν } T_{\text{ΑΘ}} < T_{\text{ΑΑ}} \end{cases} \text{ και } n_{\text{ΑΘ}}, n_{\text{ΑΑ}}$$

είναι τα πλήθη ασθενών και υγειών αντιστοίχως (*Hanley & McNeil, 1982*). Το εμβαδόν κάτω από την καμπύλη λειτουργικών χαρακτηριστικών ισοδυναμεί με την πιθανότητα της κλίμακας υγείας ή του διαγνωστικού ελέγχου να ταξινομεί ένα τυχαία επιλεγμένο άτομο από τους ασθενείς υψηλότερα σε σχέση με ένα τυχαία επιλεγμένο άτομο από τους υγιείς το οποίο ισοδυναμεί με το Wilcoxon rank test. Επομένως, το εμβαδό AUC συνδέεται με τον έλεγχο του Wilcoxon test και επίσης με το στατιστικό U του Mann-Whitney (*Pepe, 2003*).

1.3.5.7 Μέθοδοι Επιλογής Βέλτιστου Διαχωριστικού Σημείου

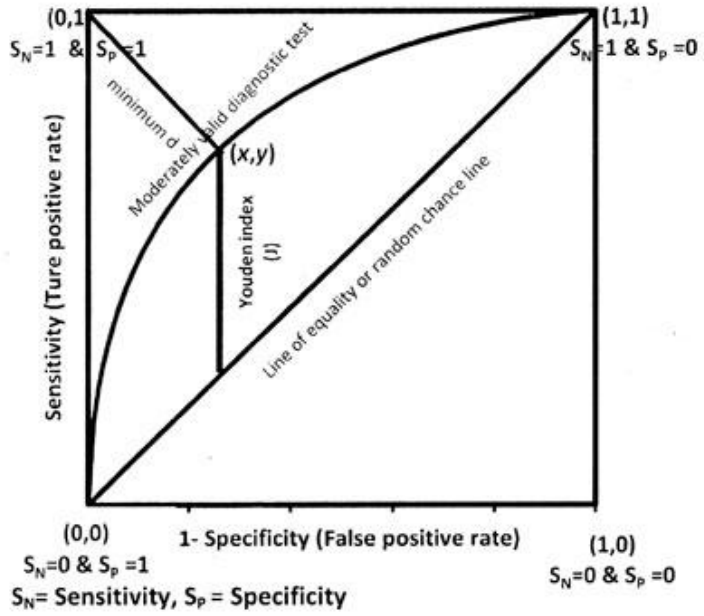
Στο σύνολο των διαγνωστικών ελέγχων γίνεται χρήση των μεταβλητών που αποσκοπούν στη βέλτιστη διακριτοποίηση μεταξύ υγιών και ασθενών. Εκτός από τις μεταβλητές αυτές, που είναι ιδιαίτερης σημασίας καθώς πρέπει να σχετίζονται με το αποτέλεσμα υγείας που η κλίμακα αποσκοπεί να εκτιμήσει, σημαντικό ρόλο έχει η επιλογή του διαγνωστικού κατωφλιού ή βέλτιστου σημείου απόφασης ή σημείο διαχωρισμού ή διαχωριστικού ορίου (επίσης ονομάζεται και κριτήριο θετικότητας) (*Egan, 1975*).

Η επιλογή του βέλτιστου διαγνωστικού κατωφλιού εξαρτάται από τη φύση της διαγνωστικής διαδικασίας με χρήση της καμπύλης ROC. Πιο συγκεκριμένα, τα κυριότερα κριτήρια απόφασης που χρησιμοποιούνται (*Νάκας, 2002*) είναι το αυστηρό κριτήριο, που αντιστοιχεί σε επιλογή σφάλματος τύπου I ίσο με 0.01 ή 0.05 (σε αντιστοιχία με τους ελέγχους υποθέσεων), το επιεικές κριτήριο, που αντιστοιχεί σε επιλογή του TPR να είναι περίπου ίσο με 1, και το δημοφιλέστερο κριτήριο που αντιστοιχεί στη συν-μεγιστοποίηση των ορθών αποφάσεων, δηλαδή στη μεγιστοποίηση ορθών θετικών και ορθών αρνητικών αποφάσεων. Η τιμή της στατιστικής συνάρτησης Youden's J (*Youden, 1950*) είναι το δημοφιλές κριτήριο για την επιλογή του βέλτιστου διαγνωστικού κατωφλίου ιατροβιολογικού ελέγχου ή κλίμακας που σχετίζεται με την υγεία. Το σημείο αυτό εντοπίζεται μεγιστοποιώντας

το άθροισμα της ευαισθησίας (TPR) και της ειδικότητας (TNR). Οπότε, στο βέλτιστο διαγνωστικό κατώφλι *c* θα ισχύει:

$$\boldsymbol{max}(\boldsymbol{TPR}+\boldsymbol{TNR}) \qquad (1.3.5.7.1)$$

Το σημείο αυτό, συνήθως συμπίπτει, με αυτό που είναι το πλησιέστερο στην πάνω αριστερή γωνία της καμπύλης ROC και είναι αυτό που επιλέγεται επίσης ως σημείο αποκοπής καθώς μεγιστοποιεί την ευαισθησία και την ειδικότητα του ελέγχου ταυτόχρονα (closest topleft & max accuracy methods). Για να εντοπιστεί το βέλτιστο διαχωριστικό σημείο του διαγνωστικού κριτηρίου, υπολογίζεται η μέγιστη απόσταση από την διαγώνιο, όπως παρατίθεται οπτικά στο σχήμα 1.3.5.7.1.



Σχήμα 1.3.5.7.1: Εύρεση βέλτιστου διαχωριστικού ορίου ως προς τη μεγιστοποίηση του αθροίσματος ευαισθησίας και ειδικότητας (Πηγή: http://www.indianpediatrics.net/apr2011/apr-277-287.htm)

Άλλος τρόπος επιλογής του βέλτιστου διαγνωστικού κατωφλιού αποτελεί η ελαχιστοποίηση του ποσοστού εσφαλμένης ταξινόμησης (misclassification rate) και σχετίζεται με τον επιπολασμό μιας ασθένειας σε γενικούς πληθυσμούς και δίνεται από τον ακόλουθο τύπο:

$$c = \frac{1-\Pr}{\Pr} \qquad (1.3.5.7.2)$$

όπου Pr συμβολίζει τον επιπολασμό μιας ασθένειας στον αντίστοιχο γενικό πληθυσμό.

Στην περίπτωση συνυπολογισμού του κέρδους της ορθής κατάταξης ή του κόστους της εσφαλμένης κατάταξης, τότε το βέλτιστο διαγνωστικό κατώφλι διαφοροποιείται και δίνεται από τον τύπο:

$$c_{opt} = \left(\frac{1-\Pr}{\Pr}\right)\left(\frac{B_{TN} - C_{FP}}{B_{TP} - C_{FN}}\right) \qquad (1.3.5.7.3)$$

όπου B_{TN}, B_{TP} είναι τα κέρδη ορθής ταξινόμησης και C_{FP}, C_{FN} είναι τα κόστη εσφαλμένης ταξινόμησης.

1.4 Προσομοίωση – Αλγόριθμοι Monte Carlo & Bootstrap

Ο αλγόριθμος «Monte Carlo» αποτελεί μια μέθοδο στοχαστικής διαδικασίας όπου με χρήση τυχαίων αριθμών και μεθόδους της στατιστικής επιχειρείται η λύση ενός πιθανοθεωρητικού προβλήματος (*Eckhardt, 1987*). Ετυμολογικά, η ονομασία «Monte Carlo» προέρχεται από την ομώνυμη πόλη του Μονακό όπου υπάρχει το αντίστοιχο διάσημο καζίνο. Ιστορικά, η μέθοδος «Monte Carlo» πρωτοπαρουσιάστηκε το 1949 από τους N. Metropolis και S. Ulam (*Metropolis & Ulam, 1949*). Το κλασικότερο παράδειγμα χρήσης της μεθόδου «Monte Carlo» είναι αυτό του μεθυσμένου ναύτη που προσπαθεί να βρει την έξοδο από την πόλη στην οποία βρίσκεται και σε κάθε σταυροδρόμι η πιθανότητα επιλογής του οποιοδήποτε από τους τέσσερις δρόμους είναι ίση με 25%. Στο πρόβλημα αυτό, αν η πόλη απαρτίζεται από λίγα οικοδομικά τετράγωνα, τότε η εύρεση της πιθανότητας που αντιστοιχεί στο ενδεχόμενο εύρεσης της εξόδου είναι αρκετά εύκολη, αλλά στην περίπτωση μιας κανονικού μεγέθους πόλης, το πρόβλημα είναι εξαιρετικά δύσκολο. Στο παραπάνω πρόβλημα βρίσκει εφαρμογή ο αλγόριθμος «Monte Carlo» σε συνδυασμό με κατάλληλο λογισμικό και χρήση Η/Υ. Με χρήση μιας γεννήτριας αριθμών, γίνεται προσομοίωση της κίνησης του ναύτη μέχρι να λυθεί το πρόβλημα, δηλαδή ο ναύτης να βρει την έξοδο. Αυτή η διαδικασία επαναλαμβάνεται πάρα πολλές φορές και στο τέλος υπολογίζεται η πιθανότητα κάθε λύσης αναλυτικά και όχι πιθανοθεωρητικά. Με την ίδια λογική, στην παρούσα διατριβή, γίνεται παραγωγή τυχαίων αριθμών με συγκεκριμένα χαρακτηριστικά ώστε να προσομοιάζουν σε χαρακτηριστικά ασθενών ή υγιών ατόμων. Με αυτή τη λογική δημιουργήθηκαν πολύ μεγάλα δείγματα και υπολογίστηκαν τα κριτήρια αξιολόγησης για την εκάστοτε παραγόμενη σταθμισμένη κλίμακα.

Η επιτυχία του αλγόριθμου «Monte Carlo» εντοπίζεται στον πολύ μεγάλο αριθμό επαναλήψεων που μπορεί να επιτευχθεί, με αποτέλεσμα τον υπολογισμό κάποιας παραμέτρου με ταυτόχρονη εύρεση μιας κατανομής της που προσεγγίζει τη θεωρητική (πραγματική).

Ο αλγόριθμος Bootstrap είναι μια αναλυτική διαδικασία που περιλαμβάνει τυχαία δειγματοληψία παρατηρήσεων με επανάθεση από το διαθέσιμο δείγμα, με σκοπό την εκτίμηση παραμέτρων της αντίστοιχης κατανομής (*Ross, 2006*). Αυτή η τεχνική επιτρέπει την εκτίμηση της δειγματοληπτικής κατανομής μιας παραμέτρου, καθώς παράγεται μια κατανομή (Bootstrap Distribution) που τείνει στην εμπειρική κατανομή της αντίστοιχης μεταβλητής και δίνει τη δυνατότητα παραγωγής Διαστημάτων Εμπιστοσύνης, με απώτερο σκοπό την ενίσχυση ενός αποτελέσματος με την αντίστοιχη στατιστική σημαντικότητα (*Efron, 2003*). Ιστορικά, η μέθοδος «Bootstrap» πρωτοπαρουσιάστηκε από τον Efron B. το 1979 (*Efron, 1979*) ο οποίος βασίστηκε σε παλαιότερη εργασία-έρευνα στην τεχνική «jackknife». Στην παρούσα διατριβή, η τεχνική χρησιμοποιήθηκε για την εκτίμηση των μέτρων της ευαισθησίας, της ειδικότητας, της ακρίβειας, της θετικής και αρνητικής προγνωστικής αξίας, καθώς επίσης και του εμβαδού της επιφάνειας κάτω από την καμπύλη λειτουργικών χαρακτηριστικών που προκύπτουν από τα δεδομένα της έρευνας ΑΤΤΙΚΑ.

Ένα μεγάλο πλεονέκτημα της μεθόδου «Monte Carlo» είναι η απλότητα της εφαρμογή της σε επίπεδο γέννησης/δημιουργίας τυχαίων αριθμών και η μέθοδος «Bootstrap» χρησιμοποιείται για την επαναδειγματοληψία από ήδη υπάρχοντα πραγματικά δεδομένα. Ειδικότερα η μέθοδος «Bootstrap» προσφέρει εκτιμήσεις για ποικιλία παραμέτρων ενός πληθυσμού, με αντίστοιχα διαστήματα εμπιστοσύνης. Ένα μειονέκτημα που έγκειται στη χρήση δεδομένων προσομοίωσης, είναι ότι δεν υπάρχει αρκετή ποικιλία για να θεωρηθεί μαθηματική απόδειξη ώστε να τεκμηριωθεί μια αξίωση μιας ερευνητικής υπόθεσης. Επομένως, από μια θεωρητική άποψη, οι προσομοιώσεις δεν είναι τόσο ισχυρές, ώστε να καθιερώσουν μια μεθοδολογία, επειδή δεν μπορούν να καλυφθούν όλες οι πιθανές περιπτώσεις, που μπορούν να υπάρξουν σε μια μαθηματική λύση μιας ερευνητικής υπόθεσης. Επομένως, η μέθοδος προσομοίωσης δεν θεωρείται ως θεωρητική μαθηματική απόδειξη, αν και είναι μια μορφή αναλυτικής απόδειξης.

Σε αυτή τη διατριβή, κατασκευάστηκαν πολλαπλά σενάρια σύμφωνα με διάφορα χαρακτηριστικά, όπως η φύση των δεδομένων που χρησιμοποιήθηκαν για τη δημιουργία των συνιστωσών μεταβλητών των σταθμισμένων κλιμάκων, χρησιμοποιώντας μεθόδους προσομοίωσης. Ως εκ τούτου, αναπτύχθηκαν σενάρια ανάλογα με την κατανομή των κλιμάκων, όπως είναι η Ομοιόμορφη κατανομή (Uniform distribution) και η Κανονική κατανομή (Normal distribution), τις παραμέτρους κατανομής των συνιστωσών, καθώς και τον αριθμό *m* των συνιστωσών

μεταβλητών. Επιπρόσθετα, διάφορα σενάρια δημιουργήθηκαν σύμφωνα με την υποθετική αναλογία πληθυσμού μεταξύ ατόμων που διαθέτουν ένα χαρακτηριστικό και αυτών που δεν το διαθέτουν για την αξιολόγηση της ακρίβειας των προβλέψεων των κλιμάκων. Η κανονική κατανομή αντιστοιχεί σε μια πιο ρεαλιστική αντίληψη για άτομα που κατατάσσονται σε δύο κατηγορίες σε αντίθεση με την ομοιόμορφη κατανομή. Τα προαναφερθέντα προσομοιωμένα δεδομένα παρήχθησαν χρησιμοποιώντας κατάλληλα λογισμικά, όπως το ***MatLab R2012b*** και το ***IBM SPSS 21***. Πιο συγκεκριμένα, σε αυτή την διατριβή επιλέχθηκαν να χρησιμοποιηθούν οι προαναφερθείσες γνωστές στατιστικές διαδικασίες, όπως η δημιουργία δεδομένων με προσομοιώσεις και μεθόδους ταξινόμησης. Τα καινοτόμα στοιχεία είναι κατά κύριο λόγο οι μέθοδοι που χρησιμοποιούνται για την ταξινόμηση των ατόμων καθώς και ο τρόπος με τον οποίο οι διαδικασίες αυτές συνδυάζονται αλγοριθμικά.

Το μέγεθος του εκάστοτε δείγματος ήταν ίσο με 100, 1.000, 10.000, 100.000 ή 1.000.000 και, για παράδειγμα, ο γενικός κανόνας που ακολουθήθηκε αναφέρει ότι η μέση τιμή των ασθενών ήταν μεγαλύτερη από τις μέσες τιμές υγιών κάθε συνιστώσας μεταβλητής. Αυτό το φαινόμενο συμβαίνει σε πολλά ανθρώπινα βιοχημικά χαρακτηριστικά όπως η υπερχοληστερολαιμία ή η υπέρταση. Η ομοιόμορφη κατανομή και η κανονική κατανομή χρησιμοποιήθηκαν για τη δημιουργία 100, 1.000, 10.000, 100.000 ή 1.000.000 δειγμάτων κάθε φορά (*Kroese, 2014*). Ως εκ τούτου, δημιουργήθηκαν 100, 1.000, 10.000, 100.000 ή 1.000.000 πίνακες δεδομένων μεγέθους (100, 1.000, 10.000, 100.000 ή 1.000.000)×*m*. Πιο συγκεκριμένα, με χρήση της μεθόδου προσομοίωσης Monte Carlo δημιουργήθηκαν τα προαναφερθέντα δείγματα και υπολογίστηκαν διαστήματα εμπιστοσύνης για όλα τα μέτρα/κριτήρια αξιολόγησης, όπως το εμβαδόν της περιοχής (AUC) (*Parikh et al., 2008*) κάτω από την καμπύλη Λειτουργικών Χαρακτηριστικών του Δέκτη (ROC), την ευαισθησία (True Positive Ratio - TPR), την ειδικότητα (True Negative Ratio - TNR), την ακρίβεια (True Ratio - TR)), τη θετική προγνωστική αξία (Positive Predictive Value - PPV) και την αρνητική προγνωστική αξία (Negative Predictive Value - NPV) στο βέλτιστο διαγνωστικό όριο/σημείο/κατώφλι (*Kourlaba & Panagiotakos, 2010*).

Ο Πίνακας 1.4.1 παρουσιάζει ενδεικτικές παραμέτρους των κατανομών που χρησιμοποιούνται για τη διεξαγωγή της προαναφερθείσας παραγωγής δεδομένων για τις κλίμακες που σχετίζονται με την υγεία και αποτελούνται από έξι συνιστώσες μεταβλητές. Αυτές οι ενδεικτικές παράμετροι των κατανομών επιλέχθηκαν τυχαία. Σε αυτό το σημείο θα πρέπει να σημειωθεί ότι εκτός από τα αποτελέσματα

προσομοίωσης που παρουσιάζονται σε αυτή τη διατριβή, όλα τα άλλα αποτελέσματα προσομοίωσης για διαφορετικές τιμές παραμέτρων των κατανομών, διαφορετικό αριθμό δειγμάτων ή διαφορετική αναλογία ασθενών και υγιών που ελέγχθηκαν ήταν σε πλήρη συμφωνία με τα αποτελέσματα που παρουσιάζονται.

Σενάριο 1: Κανονική κατανομή		Σενάριο 2: Ομοιόμορφη κατανομή	
Υγιής	Ασθενής	Υγιής	Ασθενής
X_1~N(110, 30)	**X_1~N(120, 30)**	**X_1~U(50, 150)**	**X_1~U(70, 170)**
X_2~N(108, 28)	**X_2~N(122, 28)**	**X_2~U(48, 148)**	**X_2~U(72, 172)**
X_3~N(106, 26)	**X_3~N(124, 26)**	**X_3~U(46, 146)**	**X_3~U(74, 174)**
X_4~N(104, 24)	**X_4~N(126, 24)**	**X_4~U(44, 144)**	**X_4~U(76, 176)**
X_5~N(102, 22)	**X_5~N(128, 22)**	**X_5~U(42, 142)**	**X_5~U(78, 178)**
X_6~N(100, 20)	**X_6~N(130, 20)**	**X_6~U(40, 140)**	**X_6~U(80, 180)**

Πίνακας 1.4.1: Ενδεικτικές παράμετροι κατανομής των *m* = 6 συνιστωσών μεταβλητών

Στην περίπτωση της κανονικής κατανομής, οι συνιστώσες μεταβλητές X_i, $i=1,2,\ldots,m$, κατανέμονται όπως φαίνεται στη σχέση (1.4.1)

$$X_i \sim N\left(\mu_i, \sigma_i^2\right), i = 1, 2, \ldots, m \tag{1.4.1}$$

Στην περίπτωση που μια κλίμακα αποτελείται από το αλγεβρικό άθροισμα *m* ανεξάρτητων μεταβλητών που ακολουθούν την κανονική κατανομή, τότε η τυχαία μεταβλητή *t* που εκφράζει το αλγεβρικό άθροισμα *m* ανεξάρτητων μεταβλητών ακολουθεί ομοίως την κανονική κατανομή με τις παραμέτρους (μέση τιμή & διακύμανση) όπως φαίνεται στη σχέση (*Feller, 1966*):

$$t \sim N\left(\sum_{i=1}^{m} \mu_i, \sum_{i=1}^{m} \sigma_i^2\right) \tag{1.4.2}$$

Στην περίπτωση που μια κλίμακα αποτελείται από το αλγεβρικό σταθμισμένο άθροισμα *m* ανεξάρτητων μεταβλητών που ακολουθούν την κανονική κατανομή, τότε η μεταβλητή t ακολουθεί την κανονική κατανομή με τις παραμέτρους (μέση τιμή & διακύμανση) όπως φαίνεται στη σχέση (1.4.3) (*Feller, 1966*):

$$t \sim N\left(\sum_{i=1}^{m} w_i \mu_i, \sum_{i=1}^{m} \left(w_i \sigma_i\right)^2\right) \tag{1.4.3}$$

Στην περίπτωση της ομοιόμορφης κατανομής, οι συνιστώσες μεταβλητές X_i, $i=1,2,\ldots, m$ κατανέμονται όπως η σχέση (1.4.4) παρουσιάζει:

$$X \sim U\left(a, b\right), i = 1, 2, \ldots, m \tag{1.4.4}$$

Στην περίπτωση που μια κλίμακα αποτελείται από το αλγεβρικό άθροισμα *m* ανεξάρτητων μεταβλητών που ακολουθούν την ομοιόμορφη κατανομή με το ίδιο πεδίο ορισμού X_i~$U(a,b)$, $i=1,2,\ldots,m$, τότε, η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας του *t*, που εκφράζει το το αλγεβρικό άθροισμα *m* ανεξάρτητων

μεταβλητών που ακολουθούν την ομοιόμορφη κατανομή, δίνεται από την εξίσωση (1.4.5) (*Bradley and Gupta, 2002*):

$$f_m(t)=\begin{cases}\dfrac{1}{(m-1)!(b-a)^m}\sum\limits_{r=0}^{\left\lfloor\frac{t-ma}{b-a}\right\rfloor}(-1)^r\binom{m}{r}\big(t-ma-r(b-a)\big)^{m-1}, \text{ when } t\in[m\cdot a,m\cdot b]\\ 0, \text{ when } t\notin[m\cdot a,m\cdot b]\end{cases} \quad (1.4.5)$$

Στην περίπτωση που μια κλίμακα αποτελείται από το σταθμισμένο άθροισμα *m* ανεξάρτητων τυχαίων μεταβλητών X_i ομοιόμορφα κατανεμημένων, τότε μπορεί να υποτεθεί για κάθε μεταβλητή ότι ισχύει $t_i=w_iX_i\sim U_i(a_i,b_i),\ i=1,2,\ldots m$ (*Potuschak and Mller, 2009; Buonocore et al., 2009*). Δίχως απώλεια της γενικότητας, αν υποθέσουμε ένα συμμετρικό στήριγμα, τότε συμβολικά έχουμε $y_i\sim U(-\alpha_i,\ \alpha_i),\ i=1,\ 2,\ \ldots,\ m$ (Οποιοδήποτε διάστημα θα μπορούσε να θεωρηθεί με κατάλληλη μετατόπιση του α_i). Η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας του *t* δίνεται από την εξίσωση (1.4.6) (*Potuschak & Muller, 2009*):

$$f_m(t)=\frac{\sum\limits_{j=1}^{2^m}\sigma_j\cdot sign\left(\underline{a}\cdot\underline{e}_j-t\right)\cdot\left(\underline{a}\cdot\underline{e}_j-t\right)^{m-1}}{2^{m+1}\cdot(m-1)!\left(\prod a_i\right)\cdot}, \text{ when } t\in\left[-\sum_{i=1}^{m}a_i,\sum_{i=1}^{m}a_i\right] \quad (1.4.6)$$

όπου το e_j συμβολίζει μια από τις 2^m πιθανές ακολουθίες μήκους m των πρόσημων −1 και +1 σε μορφή διανύσματος, το α εκφράζει το διάνυσμα ($a_1,\ a_2,\ \ldots,\ a_m$) και τα πρόσημα σ_j είναι θετικά εάν ο αριθμός των αρνητικών μονάδων (-1) είναι άρτιος αριθμός στην ακολουθία e_j και αρνητικός διαφορετικά (*Potuschak & Muller, 2009*).

Επιπλέον, η γνωστή μέθοδος τυχαίας επαναδειγματοληψίας με επανάθεση, η μέθοδος Bootstrap, χρησιμοποιήθηκε για τη δημιουργία διαστημάτων εμπιστοσύνης για τα μέτρα αξιολόγησης στα δεδομένα του δείγματος της επιδημιολογικής μελέτης ΑΤΤΙΚΑ που χρησιμοποιήθηκαν στην εφαρμογή αυτής της διατριβής (*Efron and Tibshirani, 1993*). Για την αξιολόγηση της στατιστικής σημαντικότητας των διαφορών της ευαισθησίας και ειδικότητας μεταξύ των κλιμάκων έγινε χρήση του αλγόριθμου bootstrap (bootstrap method) (*Janitza κ.ά., 2016*). Με αυτό τον τρόπο παρήχθησαν διαστήματα εμπιστοσύνης για την ευαισθησία και την ειδικότητα της κάθε κλίμακας στο βέλτιστο διαγνωστικό κατώφλι (cutoff point), με παραγωγή 10, 100, 1000 ή 10.000 bootstrap δειγμάτων, ιδίου μεγέθους με το αρχικό δείγμα. Πιο συγκεκριμένα, η επαναδειγματοληψία (resampling) των Bootstrap δειγμάτων έγινε σε κάθε σενάριο χωριστά επίσης. Αυτή η διαδικασία πραγματοποιήθηκε παράγοντας νέες μεταβλητές που εκφράζουν την πρόβλεψη της εκάστοτε κλίμακας, βασιζόμενες

στις αντίστοιχες μεταβλητές και στη συνέχεια κάνοντας χρήση του αντίστοιχου αλγορίθμου bootstrap υπολογίστηκαν κατάλληλα τα διαστήματα εμπιστοσύνης. Η χρήση της τεχνικής Bootstrap για παραγωγή νέων δειγμάτων εφαρμόζεται κυρίως για την εκτίμηση παραμέτρων (σημειακή και διαστήματος) από πραγματικά δεδομένα.

Στα πλαίσια της διερεύνησης της χρήσης σταθμίσεων σε κλίμακες, δημιουργήθηκε πρόγραμμα σε περιβάλλον MatLab και παρατίθεται ακολούθως ο αντίστοιχος ψευδοκώδικας (Ο πλήρης κώδικας παρατίθεται στο παράρτημα Α). Πιο συγκεκριμένα, ο ψευδοκώδικας διακρίνεται σε δύο μέρη. Στο 1ο μέρος παρατίθεται η αξιολόγηση των παραγόμενων κλιμάκων (Composite Index Evaluation) και στο 2ο μέρος παρατίθενται οι υποστηρικτικές συναρτήσεις παραγωγής δεδομένων από κατάλληλες κατανομές, η τυποποίηση των αρχικών δεδομένων και ο υπολογισμός των σταθμίσεων με χρήση κατάλληλων ρουτινών (Logistic Regression - Discriminant Analysis).

Algorithm 1 Composite Index Evaluation

1: **procedure** EVALUATE COMPOSITE INDEX(i_{max}, k_{max}, n_1, n_2,$dist$)
2: $Input$:i_{max} - # repetitions, k_{max} - # component variables,
3: n_1 - # non diseased cases, n_2 - # diseased cases, dist - Normal or Uniform distribution
4: $Output$:$avgAUC$
5: **for** $i = 1 : i_{max}$ **do**
6: $Z, Y \leftarrow Generate\ Sample\ Dataset(n_1, n_2, k_{max}, dist)$
7: $Z' = Standardize\ Sample\ Dataset(Z)$
8: $W = Calculate\ Weights'\ Matrix(Z')$
9: $T = Z_{n_1+n_2,k_{max}} \bullet W$
10: $AUC_i \leftarrow ROC(T, Y)$
11: **end for**
12: $avgAUC = average(AUC)$
13: **end procedure**

Στη γραμμή 6 του Αλγορίθμου 1, το $Y_{n1+n2,1}$ εκφράζει την τελευταία στήλη της μήτρας *Z* και περιέχει τη σωστή κλάση ανά περίπτωση. Οι πρώτες στήλες k_{max} του *Z* περιέχουν τις τιμές δείγματος για τις αντίστοιχες μεταβλητές συνιστωσών. Στη γραμμή 9 του Αλγορίθμου 1, ο *T* είναι ένας πίνακας που περιέχει τις προβλεπόμενες τιμές χρησιμοποιώντας τη σχέση 11. Η μήτρα AUC_{imax}, περιέχει τις 9 τιμές AUC (μια τιμή για κάθε κλίμακα) για κάθε επανάληψη *i* και το *avgAUC* περιέχει τις μέσες τιμές AUC για τις 9 κλίμακες. Μια λεπτομερής ανάλυση των λειτουργιών/διαδικασιών που χρησιμοποιούνται στα βήματα 6, 7, 8 και 10 του Αλγορίθμου 1 δίνονται στον Αλγόριθμο 2. Ο και J είναι ο μηδενικός και ο μοναδιαίος πίνακας αντίστοιχα.

Algorithm 2 Supportive Functions

function GENERATE SAMPLE DATASET(n_1,n_2,k_{max},$dist$)
Input:n_1 -# non diseased cases, n_2 -# diseased cases, k_{max} - # component variables, dist - flag for Normal or Uniform distribution
Output:$Z_{n_1+n_2,k_{max}+1}$
  $Y_1 \leftarrow O_{n_1,1}$
  $Y_2 \leftarrow J_{n_2,1}$
  $Y_{n_1+n_2,1} \leftarrow (Y_1|Y_2)$
IF (dist = Normal)
  for $k = 1 : k_{max}$ **do**
    $X_{1:n_1,k} \leftarrow random(dist, mean_{healthy}, stdev_{healthy}, n_1, 1)$
    $X_{n_1+1:n_1+n_2,k} \leftarrow random(dist, mean_{diseased}, stdev_{diseased}, n_2, 1)$
  end for
ELSE IF (dist = Uniform)
  for $k = 1 : k_{max}$ **do**
    $X_{1:n_1,k} \leftarrow random(dist, lower_{healthy}, upper_{healthy}, n_1, 1)$
    $X_{n_1+1:n_1+n_2,k} \leftarrow random(dist, lower_{diseased}, upper_{diseased}, n_2, 1)$
  end for
ENDIF
  return $Z_{n_1+n_2,k_{max}+1} \leftarrow (X_{n_1+n_2,k_{max}}|Y_{n_1+n_2,1})$
end function

---

function STANDARDIZE SAMPLE DATASET(Z)
  for $k = 1 : k_{max}$ **do**
    $mean_{healthy}(k) \leftarrow mean(Z_{1:n_1,k})$
    $stdev(k) \leftarrow stdev(Z_{:,k})$
    $Z_{:,k} \leftarrow Z_{:,k} - \frac{mean_{healthy}(k)}{stdev(k)}$
  end for
  return Z
end function

---

function CALCULATE WEIGHTS' MATRIX(Z)
  $w_{ulr} \leftarrow UBLR\ OR(Z)$
  $w_{mlr} \leftarrow MBLR\ OR(Z)$
  $w_{ulr\_ds} \leftarrow UBLR\ OR/DS(Z)$
  $w_{mlr\_ds} \leftarrow MBLR\ OR/DS(Z)$
  $w_{ulda} \leftarrow ULDA(Z)$
  $w_{mlda} \leftarrow MLDA(Z)$
  $w_{ulda\_lw} \leftarrow ULDA\ FC/LW(Z)$
  $w_{mlda\_lw} \leftarrow MLDA\ FC/LW(Z)$
  $W \leftarrow (J_1|w_{ulr}|w_{mlr}|w_{ulr\_ds}|w_{mlr\_ds}|w_{ulda}|w_{mlda}|w_{ulda\_lw}|w_{mlda\_lw})$
  return W
end function

1.5 Μέθοδοι Ταξινόμησης - Υπολογισμός Σταθμίσεων Κλιμάκων

Σε αυτή τη διατριβή προτείνονται καινοτόμες κλίμακες, χρησιμοποιώντας κατάλληλες σταθμίσεις που υπολογίζονται με μεθόδους ταξινόμησης, όπως η δυαδική λογιστική παλινδρόμηση (Binary Logistic Regression - BLR) και η γραμμική διαχωριστική ανάλυση (Linear Discriminant Analysis - LDA), προκειμένου να επιτευχθεί υψηλότερη διαγνωστική ακρίβεια σε σχέση με μη σταθμισμένες κλίμακες (*Pepe & Thompson, 2000*). Οι προτεινόμενες σταθμίσεις είναι δυναμικές και όχι σταθερές, δηλαδή η τιμή τους διαμορφώνεται από τα δεδομένα του κάθε πληθυσμού που εφαρμόζεται. Στον τομέα των επιστημών της υγείας συνεχίζονται οι έρευνες για την κατασκευή συνδυαστικών/αθροιστικών κλιμάκων που έχουν καλύτερες επιδόσεις

κατάταξης ατόμων ενός γενικού πληθυσμού σε ασθενείς και υγιείς (*Bansal & Pepe, 2013*).

Στην πρόσφατη βιβλιογραφία χρησιμοποιούνται διάφορες στατιστικές μέθοδοι ταξινόμησης, όπως η Λογιστική Παλινδρόμηση (Logistic Regression - LR) (*Stavrinos & Panagiotakos, 2007*), η Διαχωριστική ή Διακρίνουσα Ανάλυση (Discriminant Analysis - DA) (*Karlis, 2005*), η Ανάλυση Ταξινόμησης Δέντρων (Classification Trees - CT) (*Worth & Cronin, 2003*), τα Νευρωνικά Δίκτυα (Neural Networks - NN) (*McCulloch &P itts, 1943)* και στοιχεία από εξόρυξη πληροφορίας όπως είναι η μηχανική μάθηση και οι μηχανές διανυσματικής υποστήριξης (Support Vector Machines - SVM) (*Kruppa et al., 2014*) που στοχεύουν στη διάκριση δύο ή περισσοτέρων διαφορετικών ομάδων σε σύνολα δεδομένων που έχουν ένα συγκεκριμένο χαρακτηριστικό ή όχι. Σε αυτή τη διατριβή, επιλέχθηκαν οι μέθοδοι της λογιστικής παλινδρόμησης (LR) και διακρίνουσας ανάλυσης (DA) για να αναπτυχθούν μοντέλα ταξινόμησης που να εφαρμόζονται σε προβλήματα υγείας.

Αμφότερες οι στατιστικές μέθοδοι της λογιστικής παλινδρόμησης και της διαχωριστικής ανάλυσης χρησιμοποιούνται για να κατατάξουν παρατηρήσεις σε γνωστές ομάδες και συναντώνται συχνά στη βιβλιογραφία σε προβλήματα διαχωρισμού πληθυσμών (*Worth & Cronin, 2003; Pohar et al, 2004*). Στη λογιστική παλινδρόμηση δεν είναι απαραίτητες οι υποθέσεις της διαχωριστικής ανάλυσης που αναφέρονται στην κανονική κατανομή των ανεξάρτητων μεταβλητών καθώς επίσης και στην υπόθεση περί ισότητας διασπορών μεταξύ των πληθυσμών. Βασική διαφορά των δύο μεθόδων είναι ότι στη διαχωριστική ανάλυση γίνεται χρήση αποκλειστικά ποσοτικών μεταβλητών ως ανεξάρτητων, ενώ, στην περίπτωση της λογιστικής παλινδρόμησής δύναται να χρησιμοποιηθούν και ποιοτικές μεταβλητές. Συνεπώς, θα μπορούσε να υποτεθεί ότι η λογιστική παλινδρόμηση είναι πιο εύχρηστη μέθοδος από αυτή της διαχωριστικής ανάλυσης, αλλά στην περίπτωση που ικανοποιούνται οι υποθέσεις περί κανονικότητας και ισότητας των πινάκων διακύμανσης, τότε η διαχωριστική ανάλυση δίνει καλύτερα αποτελέσματος προς τη διάκριση και πρόβλεψη ασθενών και υγιών.

1.5.1 Διαχωριστική Ανάλυση

Η διαχωριστική ανάλυση είναι μια γνωστή στατιστική τεχνική η οποία ως στόχο έχει την ταξινόμηση/κατάταξη παρατηρήσεων σε δύο ή περισσότερους ήδη

γνωστούς πληθυσμούς (*Karlis, 2005*) και εφαρμόζεται συχνά σε ιατρικά ή/και βιολογικά δεδομένα για τη διάκριση ατόμων που έχουν ένα συγκεκριμένο χαρακτηριστικό ή όχι. Βασική διαφορά της διαχωριστικής ανάλυσης από την ανάλυση κατά συστάδες είναι ότι στην τελευταία οι ομάδες δεν είναι γνωστές ενώ στην πρώτη είναι γνωστές. Για την υλοποίηση της μεθόδου της γραμμικής διαχωριστικής ανάλυσης (Linear Discriminant Analysis) απαιτείται η κατασκευή ενός κανόνα διαχωρισμού των ομάδων (π.χ. Μέγιστης Πιθανοφάνειας, Μέθοδος Bayes, Ελαχιστοποίηση του κόστους εσφαλμένης κατάταξης) με τη βοήθεια του οποίου, οι παρατηρήσεις κατατάσσονται στις διακριτές ομάδες μέσω μιας γραμμικής διαχωριστικής συνάρτησης (διαχωριστική συνάρτηση του Fisher). Ο κανόνας διαίρεσης από τον Fisher μετατρέπει τις αρχικές μεταβλητές σε τιμές μιας διαστάσεως που παράγονται από την αντίστοιχη διαχωριστική συνάρτηση (*Karlis, 2005*). Δηλαδή, η διάκριση των ατόμων πραγματοποιείται με την ανάπτυξη μιας κατάλληλης συνάρτησης, βασισμένης στον κανόνα διαχωρισμού του Fisher (*Fisher, 1936*). Στη γραμμική διαχωριστική ανάλυση (LDA), η αντίστοιχη διαχωριστική συνάρτηση είναι ένας γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών και έχει ως στόχο να διαιρέσει έναν πληθυσμό σε δύο ή περισσότερες υποομάδες και να ταξινομήσει όσο το δυνατόν περισσότερες νέες περιπτώσεις σωστά. Σε αυτή τη διατριβή, χρησιμοποιήθηκε δυαδική γραμμική διαχωριστική ανάλυση (BLDA) με στόχο τη δημιουργία γραμμικών συνδυασμών που διακρίνονται με ακρίβεια σε άτομα που κατείχαν εκ των προτέρων (a priori) ένα σύμπτωμα ασθένειας ή όχι (*Kachigan, 1991*).

Υποθέτοντας κανονικότητα των πληθυσμών και ισότητα των πινάκων συνδιακύμανσης, τότε για τυχαία παρατήρηση $x \sim N(\mu_k, \Sigma)$ για την αντίστοιχη πυκνότητα της πολυδιάστατης κανονικής κατανομής ισχύει:

$$f_k(x \mid \mu_k, \Sigma) = (2\pi)^{-p/2} \cdot |\Sigma|^{-1/2} \cdot \exp\left(-\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)\right), \quad k = 1,2 \quad (1.5.1.1)$$

όπου η ποσότητα $(x-\mu_k)'\Sigma^{-1}(x-\mu_k)$ είναι η απόσταση Mahalanobis της παρατήρησης *x* από τον μέσο της *k* ομάδας. Στην περίπτωση αυτή ο αντίστοιχος διαχωριστικός κανόνας δίνεται από τη σχέση (1.5.1.2):

$$\log\left(\frac{f_1(x \mid \mu_1, \Sigma)}{f_2(x \mid \mu_2, \Sigma)}\right) = (\mu_1 - \mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) \geq k_0 \qquad (1.5.1.2)$$

όπου η τιμή k_0 εξαρτάται από τον κανόνα διαχωρισμού των ομάδων (π.χ. στην περίπτωση του κανόνα μέγιστης πιθανοφάνειας ισχύει k_0=1).

Ο διαχωριστικός κανόνας του Fisher μετατρέπει τις αρχικές μεταβλητές σε μονοδιάστατες τιμές μέσω της παραγόμενης διαχωριστικής συνάρτησης (discriminant function) (*Karlis, 2005*). Αρχικά, παράγονται δύο γραμμικές συναρτήσεις με τη μέθοδο του Fisher, W_1 και W_2, μία για τους ασθενείς και μια για τους υγιείς, που ονομάζονται διαχωριστικές συναρτήσεις του Fisher (Classification Functions). Θεωρώντας τη διαφορά $Z=W_1-W_2$, μεταξύ των διαχωριστικών συναρτήσεων του Fisher, λαμβάνουμε μια συνάρτηση, της οποίας οι συντελεστές είναι ανάλογοι των μη τυποποιημένων συντελεστών της κανονικοποιημένης διαχωριστικής συνάρτησης (Canonical Discriminant Function Coefficients). Στη δυαδική γραμμική διαχωριστική ανάλυση (BLDA), η αντίστοιχη γραμμική διαχωριστική συνάρτηση (LDF) που διαχωρίζει τα κεντροειδή των δύο ομάδων δίνεται από τη γενική μορφή της εξίσωσης (1.5.1.3):

$$LDF\left(X_1, X_2, ..., X_m\right) = b_0 + b_1X_1 + b_2X_2 + ... + b_mX_m \qquad (1.5.1.3)$$

όπου b_0 είναι η σταθερά και $b_1, b_2, ..., b_m$ είναι οι διακριτικοί (διαχωριστικοί) συντελεστές για τις αντίστοιχες m μεταβλητές $X_1, X_2, ..., X_m$ (*Worth & Cronin, 2003*). Έτσι, αν ισχύει ότι $Z>0$, τότε κατατάσσουμε την αντίστοιχη παρατήρηση, στην 1η ομάδα, διαφορετικά στη 2η. Οι μη τυποποιημένοι συντελεστές της κανονικοποιημένης διαχωριστικής συνάρτησης προκύπτουν από την εύρεση των ιδιοτιμών της σχέσης (T-W)V=λWV, όπου V η μήτρα των συντελεστών της διαχωριστικής συνάρτησης, T είναι ο πίνακας τετραγωνικών συνολικών αθροισμάτων, W είναι ο πίνακας τετραγωνικών αθροισμάτων εντός των ομάδων και το λ είναι ένας διαγώνιος πίνακας των ιδιοτιμών. Ένα μέτρο που εκφράζει τη διαχωριστική ικανότητα της προαναφερθείσας συνάρτησης είναι το Λ του Wilks, όπου εκφράζει το ποσοστό της ανερμήνευτης διακύμανσης από το αντίστοιχο μοντέλο της ανάλυσης διακύμανσης κατά ένα παράγοντα, ή διαφορετικά, το ποσοστό μεταβλητότητας που δεν εξηγείται από τη διαχωριστική συνάρτηση και λαμβάνει τιμές από το 0 έως το 1, συνεπώς μικρές τιμές του Λ είναι επιθυμητές (Καρλής, 2005).

Για την ορθή εφαρμογή της μεθόδου της γραμμικής διαχωριστικής ανάλυσης (LDA) απαιτούνται δύο βασικές προϋποθέσεις: Πρώτον, οι επεξηγηματικές μεταβλητές θα πρέπει να μετρούνται σε κλίμακα διαστήματος ή κλίμακας και να

ακολουθούν μια πολυμεταβλητή κανονική κατανομή και, δευτερευόντως, να ισχύει η ισότητα μεταξύ των πινάκων διακύμανσης/συνδιακύμανσης μεταξύ των δύο διαφορετικών και διακριτών ομάδων.

1.5.2 Λογιστική Παλινδρόμηση

Το λογιστικό μοντέλο παλινδρόμησης είναι ένα μη γραμμικό μοντέλο παλινδρόμησης που εφαρμόζεται στην περίπτωση που η εξαρτημένη μεταβλητή απόκρισης *Y* δεν είναι ποσοτική, αλλά κατηγορική με δύο ή περισσότερες κατηγορίες. Στην παρούσα εργασία, εφαρμόστηκε λογιστική παλινδρόμηση με δίτιμη εξαρτημένη μεταβλητή (Binary Logistic Regression - BLR) που εκφράζει παρουσία (*Y*=1) ή απουσία νόσου (*Y*=0). Το απλό λογιστικό μοντέλο είναι της μορφής (*Stavrinos & Panagiotakos, 2007*) που δίνεται από τη σχέση (1.5.2.1):

$$P\left(Y=1|\,X\right)=\pi\left(X\right)=\frac{e^{a+\beta X}}{1+e^{a+\beta X}}=\left(1+e^{-\left(a+\beta X\right)}\right)^{-1} \qquad (1.5.2.1)$$

όπου *P(Y=1/X)=π(X)* εκφράζει το λογαριθμικό υπόδειγμα και είναι η πιθανότητα παρουσίας της νόσου δεδομένου του χαρακτηριστικού *X*. Επομένως η πιθανότητα απουσίας της νόσου δίνεται από την μαθηματική έκφραση της σχέσης (1.5.2.2):

$$P\left(Y=0\,|\,X\right)=1-\pi\left(X\right)=\frac{1}{1+e^{a+\beta X}}=\left(1+e^{\left(a+\beta X\right)}\right)^{-1} \qquad (1.5.2.2)$$

Το πηλίκο *π(X)/(1-π(X))* εκφράζει το λόγο των συμπληρωματικών πιθανοτήτων (odds) και δίνεται από τον τύπο της σχέσης (1.5.2.3) (*Stavrinos & Panagiotakos, 2007*) λύνοντας τη σχέση (3.3.2.1) ως προς τον όρο *exp*($\alpha+\beta X_j$):

$$e^{a+\beta \mathrm{X}}=\frac{\pi\left(\mathrm{X}\right)}{1-\pi\left(\mathrm{X}\right)} \qquad (1.5.2.3)$$

Λαμβάνοντας το λογάριθμο κατά μέλη στη σχέση (1.5.2.3), ο γραμμικός όρος $\alpha+\beta X$ δίνεται από τη σχέση (1.5.2.4):

$$a+\beta\mathrm{X}=\ln\left(\frac{\pi\left(\mathrm{X}\right)}{1-\pi\left(\mathrm{X}\right)}\right) \qquad (1.5.2.4)$$

Ο όρος e^{β} ισούται με το σχετικό λόγο συμπληρωματικών πιθανοτήτων να εμφανιστεί η νόσος σε όσους έχουν ένα δίτιμο χαρακτηριστικό *X* σε σύγκριση με αυτούς που δεν το έχουν, οπότε ισχύει:

$$OR=\frac{\pi\left(1\right)\cdot\left[1-\pi\left(0\right)\right]}{\pi\left(0\right)\cdot\left[1-\pi\left(1\right)\right]}=e^{\beta} \qquad (1.5.2.5)$$

Στο ίδιο αποτέλεσμα καταλήγουμε αν η ανεξάρτητη μεταβλητή X είναι ποσοτική. Η διαφοροποίηση εντοπίζεται στον τρόπο ερμηνείας του αντίστοιχου σχετικού λόγου συμπληρωματικών πιθανοτήτων. Το πολλαπλό λογιστικό μοντέλο παλινδρόμησης είναι της μορφής:

$$P\left(Y=1 \mid \tilde{X}\right)=\pi\left(\tilde{X}\right)=\frac{e^{g\left(\tilde{X}\right)}}{1+e^{g\left(\tilde{X}\right)}} \qquad (1.5.2.6)$$

όπου $\tilde{\mathrm{X}}=\left(X_1, X_2, X_3, ..., X_m\right)$ το διάνυσμα των ανεξάρτητων κανονικών μεταβλητών X_j, $j=1,2,...m$ και $g\left(\tilde{\mathrm{X}}\right)=a+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\beta_3\mathrm{X}_3+....+\beta_m\mathrm{X}_m$.

Η απόκλιση (Deviance) μεταξύ παρατηρούμενων και προβλεπόμενων τιμών του λογιστικού υποδείγματος εκφράζει την απόκλιση μεταξύ του πλήρους (θεωρητικού) υποδείγματος (Full) και του υποδείγματος που εκτιμήθηκε συναρτήσει των διαθέσιμων ανεξάρτητων μεταβλητών (Estimated). Το επιθυμητό αποτέλεσμα είναι το εκτιμώμενο υπόδειγμα να «πλησιάζει» το πλήρες μοντέλο και τότε η απόκλιση λαμβάνει μικρές τιμές. Η απόκλιση (Deviance) εκφράζεται με τη βοήθεια του λόγου πιθανοφανειών (*-2LL*) μεταξύ των δυο προαναφερθέντων υποδειγμάτων, δηλαδή είναι ένα μέτρο της καλής προσαρμοστικότητας του μοντέλου λογιστικής παλινδρόμησης και δίνεται από τη μαθηματική έκφραση της σχέσης (1.5.2.7):

$$D=-2\ln l\left(estimated\right)-\left[-2\ln l\left(full\right)\right] \qquad (1.5.2.7)$$

Η στατιστική συνάρτηση D ακολουθεί χ^2 κατανομή με n-p βαθμούς ελευθερίας, όπου n το μέγεθος του δείγματος και p ο αριθμός των ερμηνευτικών μεταβλητών που συμμετέχουν στο μοντέλο λογιστικής παλινδρόμησης.

1.5.3 Μέθοδοι Στάθμισης Κλιμάκων

Η παράγραφος αυτή παρέχει μια περιγραφή της χρησιμοποιούμενης μεθοδολογίας κατασκευής των σταθμισμένων κλιμάκων, καθώς και των επιλεγμένων μεθόδων στάθμισης. Σε γενικούς πληθυσμούς, η απόκλιση των ατόμων από την «κανονικότητα», π.χ. η απόκλιση από μια υγιή κλινική κατάσταση, μια φυσιολογική συμπεριφορά, τη μέση κατανάλωση τροφίμων κλπ. παρουσιάζει μεγάλο ενδιαφέρον, λόγω του γεγονότος ότι τα άτομα που παρουσιάζουν μεγάλες αποκλίσεις πρέπει να διερευνηθούν περαιτέρω ώστε να χαρακτηρίζονται με ορθό και έγκυρο τρόπο ως ασθενή ή υγιή. Στα πλαίσια αυτής της λογικής, κρίθηκε απαραίτητη η μετατροπή των αρχικών μεταβλητών με κατάλληλο μετασχηματισμό. Ειδικότερα, ο προαναφερθείς

μετασχηματισμός διεξήχθη χρησιμοποιώντας τη μέση τιμή των υγιών ατόμων, για κάθε συνιστώσα μεταβλητή, και την τυπική απόκλιση του δείγματος ή του συνολικού πληθυσμού, στην περίπτωση που ήταν γνωστή. Με αυτόν τον τρόπο, αποφεύχθηκε η επίδραση του μεγέθους των μεταβλητών μεγάλης κλίμακας καθώς όλες οι μεταβλητές μετασχηματιστήκαν έχοντας μια κοινή βάση μεγέθους. Ο μετασχηματισμός που χρησιμοποιείται αντιπροσωπεύει την απόκλιση των παρατηρήσεων από τον μέσο του πληθυσμού, υπολογιζόμενη σε μονάδες τυπικής απόκλισης, για κάθε μεταβλητή ξεχωριστά, οπότε θα μπορούσε να ειπωθεί ότι πρόκειται για μια τροποποιημένη μορφή τυποποίησης. Αυτός ο μετασχηματισμός, ωφελεί επίσης όταν οι μεταβλητές έχουν τελείως διαφορετικές διασπορές, ειδικά όταν μερικές από αυτές έχουν μεγάλη τυπική απόκλιση και μερικές άλλες είναι μικρές. Επομένως, θα πρέπει να προτιμάται η τυποποίηση των μεταβλητών μέσω του γενικού μετασχηματισμού των z-σημείων. Ο τύπος μετασχηματισμού που χρησιμοποιείται δίνεται παρακάτω στην εξίσωση (1.5.3.1):

$$Z = \frac{X - \mu_{healthy}}{\sigma_{total}} \qquad (1.5.3.1)$$

Ο αλγόριθμος που κατασκευάστηκε σε περιβάλλον MatLab παρείχε τη δυνατότητα εφαρμογής των προτεινόμενων σταθμισμένων κλιμάκων με ή χωρίς αυτή την τυποποίηση, αν και τα τελικά αποτελέσματα ήταν εγκυρότερα στην περίπτωση τυποποιημένων δεδομένων. Σε αυτή την διδακτορική διατριβή, αναπτύχθηκαν οκτώ διαφορετικές μέθοδοι στάθμισης και δοκιμάστηκαν σε προσομοιωμένα καθώς και σε πραγματικά δεδομένα. Η μη σταθμισμένη κλίμακα (T_0) κατασκευάστηκε χρησιμοποιώντας το απλό αλγεβρικό άθροισμα των τυποποιημένων τιμών των συνιστωσών Z_i, $i=1,2,\ldots m$. Επιπλέον, οι οκτώ διακριτές σταθμισμένες κλίμακες με διαφορετικά βάρη w_{ij}, $i=1,2,\ldots m$, $j= 1, 2,\ldots, 8$ για κάθε συνιστώσα, αναπτύχθηκαν με χρήση των λόγων συμπληρωματικών πιθανοτήτων από τη λογιστική παλινδρόμηση και με χρήση των συντελεστών των διαχωριστικών συναρτήσεων του Fisher από τη διαχωριστική ανάλυση. Τέσσερις από τις σταθμισμένες κλίμακες (T_1, T_2, T_3, T_4), κατασκευάστηκαν με χρήση των σχετικών λόγων πιθανοτήτων (OR) κάθε συνιστώσας που λαμβάνεται από δυαδικά μονοπαραγοντικά και πολυπαραγοντικά μοντέλα λογικής παλινδρόμησης. Πιο συγκεκριμένα, η σταθμισμένη κλίμακα T_1 (UBLR OR) κατασκευάστηκε χρησιμοποιώντας τους σχετικούς λόγους συμπληρωματικών πιθανοτήτων από την μονοπαραγοντική διωνυμική λογιστική

παλινδρόμηση, ενώ η σταθμισμένη κλίμακα T_2 κατασκευάστηκε χρησιμοποιώντας τους σχετικούς λόγους συμπληρωματικών πιθανοτήτων κάθε συνιστώσας που λαμβάνεται από το πολυμεταβλητό μοντέλο δυαδικής λογιστικής παλινδρόμησης (MBLR OR). Οι σταθμισμένες κλίμακες T_3 και T_4 κατασκευάστηκαν χρησιμοποιώντας τους προαναφερθέντες συντελεστές πιθανότητας σε συνδυασμό με τη στατιστική συνάρτηση που εκφράζει την απόκλιση (DS) που αποκτήθηκε από την αντίστοιχη μονομεταβλητή λογιστική παλινδρόμηση. Η στατιστική συνάρτηση της απόκλισης (DS -2LL) χρησιμοποιήθηκε για την ενίσχυση των σταθμίσεων για τις μεταβλητές που αντιστοιχούν στις χαμηλότερες τιμές απόκλισης. Επιπλέον, άλλες τέσσερις σταθμισμένες κλίμακες (T_5, T_6, T_7, T_8) κατασκευάστηκαν με τη χρήση των συντελεστών της διαχωριστικής συνάρτησης του Fisher (FDFC) κάθε συνιστώσας που λαμβάνεται από μοντέλα μονοπαραγοντικής και πολυπαραγοντικής γραμμικής διαχωριστικής ανάλυσης. Ειδικότερα, η σταθμισμένη κλίμακα T_5, κατασκευάσθηκε χρησιμοποιώντας ως βάρη τους συντελεστές της διαχωριστικής συνάρτησης του Fisher (ULDA FDFC) κάθε συνιστώσας που λαμβάνεται από το αντίστοιχο μοντέλο διαχωριστικής ανάλυσης. Για την σταθμισμένη κλίμακα (T_6), χρησιμοποιήθηκαν (MLDA FDFC) οι συντελεστές διακριτικής συνάρτησης του Fisher που ελήφθησαν από το μοντέλο απόφασης της πολυμεταβλητής διαχωριστικής ανάλυσης (*Karlis, 2005*). Οι σταθμισμένες κλίμακα T_7 και T_8 κατασκευάστηκαν χρησιμοποιώντας τους προαναφερθέντες συντελεστές της διαχωριστικής συνάρτησης του Fisher (FDFC) σε συνδυασμό με το Λάμδα (Λ) του Wilk (LW) που αποκτήθηκε από την αντίστοιχη διαχωριστική ανάλυση. Το μέτρο Wilks 'L χρησιμοποιήθηκε επειδή περιγράφει το ποσοστό μεταβλητότητας που δεν εξηγείται από την αντίστοιχη διαχωριστική συνάρτηση, έτσι αντισταθμίζεται με το αντίστοιχο πηλίκο και στηρίζονται εκείνες οι μεταβλητές που αντιστοιχούν στις χαμηλότερες τιμές του L. Η μαθηματική έκφραση των σταθμισμένων κλιμάκων δίνεται από τη σχέση (1.5.3.2):

$$\mathrm{T}_i = \sum_{j=1}^{m} w_{ij} Z_j, \quad j = 1, 2, ..., m, \quad i = 0, 1, 2, ..., 8 \qquad (1.5.3.2)$$

με τις σταθμίσεις που περιγράφονται παρακάτω:

$$w_{0j} = 1,\ w_{1j} = \frac{(UBLR\ OR)_j}{\sum_{j=1}^{m}(UBLR\ OR)_j},\ w_{2j} = \frac{(MBLR\ OR)_j}{\sum_{j=1}^{m}(MBLR\ OR)_j}$$

$$w_{3j} = \frac{(UBLR\,OR)_j / (DS)_j}{\sum_{j=1}^{m}\left((UBLR\,OR)_j / (DS)_j\right)}, w_{4j} = \frac{(MBLR\,OR)_j / (DS)_j}{\sum_{j=1}^{m}\left((MBLR\,OR)_j / (DS)_j\right)},$$

$$w_{5j} = \frac{(ULDA\,FDFC)_j}{\sum_{j=1}^{m}(ULDA\,FDFC)_j}, w_{6j} = \frac{(MLDA\,FDFC)_j}{\sum_{j=1}^{m}(MLDA\,FDFC)_j},$$

$$w_{7j} = \frac{(ULDA\,FDFC)_j / (LW)_j}{\sum_{j=1}^{m}\left((ULDA\,FDFC)_j / (LW)_j\right)}, w_{8j} = \frac{(MLDA\,FDFC)_j / (LW)_j}{\sum_{j=1}^{m}\left((MLDA\,FDFC)_j / (LW)_j\right)}, \quad j = 1, 2, ..., m$$

(Ο πλήρης αλγόριθμος παρατίθεται στο παράρτημα Α).

1.6 Διαγνωστική Ακρίβεια Σταθμισμένων Κλιμάκων

Η διαχωριστική ικανότητα των σταθμισμένων κλιμάκων αξιολογήθηκε με την χρήση της περιοχής (AUC) (*Parikh et al., 2008*) κάτω από την καμπύλη Λειτουργικών Χαρακτηριστικών του Δέκτη (ROC), με τον υπολογισμό της ευαισθησίας (True Positive Ratio - TPR), της ειδικότητας (True Negative Ratio - TNR), της ακρίβειας (True Ratio - TR)), της θετικής προγνωστικής αξίας (Positive Predictive Value - PPV) και της αρνητικής προγνωστικής αξίας (Negative Predictive Value - NPV) στο βέλτιστο διαγνωστικό όριο/σημείο/κατώφλι (*Kourlaba & Panagiotakos, 2010*).

Εκτός από την προαναφερθείσα αξιολόγηση της ταξινόμησης, διεξήχθη κατάλληλη δοκιμασία αξιολόγησης της πρόβλεψης για κάθε κλίμακα σε ανεξάρτητο δείγμα. Πιο συγκεκριμένα, διαχωρίζοντας τα αρχικά δεδομένα σε σύνολο εκπαίδευσης (training set) και σύνολο δοκιμής/ελέγχου (test set) με τη χρήση τεχνικής διαχωρισμού (*Fielding & Bell, 1997*). Το αρχικό δείγμα χωρίστηκε τυχαία σε δύο υποδείγματα, το δείγμα εκπαίδευσης περιείχε το 90% του αρχικού δείγματος και το δείγμα δοκιμής/ελέγχου περιείχε το 10% του αρχικού δείγματος. Η διαδικασία αυτή επαναλήφθηκε για 1.000 φορές και παρουσιάζονται οι μέσες τιμές του εμβαδού κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (AUC). Οι προτεινόμενες σταθμίσεις παρήχθησαν χρησιμοποιώντας το δείγμα εκπαίδευσης και οι σταθμισμένες κλίμακες δοκιμάστηκαν στο δείγμα ελέγχου/δοκιμής χρησιμοποιώντας τα προαναφερθέντα μέτρα. Ο λόγος μεταξύ υγιών και μη υγιών ατόμων ορίστηκε τυχαία σε 4:1. Η εφαρμογή των προτεινόμενων σταθμισμένων κλιμάκων σε ένα ανεξάρτητο δείγμα παρέχει μια ποσοτική αποτύπωση της προγνωστικής τους ικανότητας. Οι

προσομοιώσεις και τα πραγματικά δεδομένα από την επιδημιολογική μελέτη της ATTICA χρησιμοποιήθηκαν για να εξεταστεί το ερευνητικό ερώτημα περί χρήσης σταθμίσεων σε κλίμακες υγείας και προτείνονται λύσεις για τον προσδιορισμό των μεθόδων στάθμισης κατά την κατασκευή κλιμάκων που σχετίζονται με την υγεία.

1.7 Διαγνωστική Ακρίβεια Κλιμάκων με Διαφορετικό Αριθμό Συνιστωσών

Οι σταθμισμένες κλίμακες και η αστάθμητη κλίμακα παρήχθησαν από διάφορους αριθμούς *m* των συνιστωσών, προκειμένου να εξεταστεί η πιθανή επίπτωση του αριθμού *m* στη διαγνωστική ακρίβεια της κλίμακας. Πιο συγκεκριμένα, δημιουργήθηκαν κλίμακες από διαφορετικούς αριθμούς συνιστωσών μεταβλητών διατηρώντας τις ίδιες παραμέτρους των ακραίων (αριθμητικά) συνιστωσών μεταβλητών. Πιο συγκεκριμένα, αρχικά χρησιμοποιήθηκαν 2 μεταβλητές με συγκεκριμένη παραμετροποίηση (X_1 και X_2) και στη συνέχεια ενσωματώθηκαν μεταβλητές με παραμετροποίηση που άνηκε στο σύνολο μεταξύ των παραμέτρων των αρχικών μεταβλητών X_1 και X_2. Πιο συγκεκριμένα, στην περίπτωση της κανονικής κατανομής, οι μεταβλητές X_1 και X_2 κατανέμονταν ως εξής:

$$X_1 \sim N(30,5) \text{ και } X_2 \sim N(100,5)$$

Οι ενδιάμεσες μεταβλητές που εισήχθησαν στο μοντέλο πρόβλεψης κατανέμονταν ως εξής:

$$X_3 \sim N(35,5),\ X_4 \sim N(40,5),\ X_5 \sim N(45,5),\ \ldots\ X_{.} \sim N(95,5) \text{ και } X_2 \sim N(100,5)$$

Επομένως, για το σύνολο των περιπτώσεων που δοκιμάστηκαν με χρήση προσομοίωσης, όλες οι μεταβλητές είχαν κατάλληλη κωδικοποίηση ώστε η μόνη διαφοροποίηση να εντοπίζεται στον αριθμό τους *m*.

1.8 Εφαρμογή Σταθμισμένων Κλιμάκων σε Επιδημιολογικά Δεδομένα

Τα αποτελέσματα των προσομοιώσεων στη συνέχεια εφαρμόστηκαν σε πραγματικά δεδομένα από την επιδημιολογική μελέτη της ATTICA. Η μελέτη ATTIKA είναι μια έρευνα για την υγεία σχετιζόμενη με τη διατροφή που διεξήχθη στην περιφέρεια Αττικής (78% αστικές και 22% αγροτικές περιοχές) από τον Μάιο του 2001 έως το Δεκέμβριο του 2002. Η επιλεγμένη μέθοδος δειγματοληψίας ήταν

τυχαία και πολυεπίπεδη βασισμένη σε διάφορες ιδιότητες του όπως η πόλη, το φύλο, η ηλικία, κτλ. της περιφέρειας Αττικής που παρείχε η Εθνική Στατιστική Υπηρεσία (απογραφή του 2000) και τελικά συμφώνησαν να συμμετάσχουν 3042 άτομα (*Pitsavos et al., 2003*). Όλοι οι συμμετέχοντες ήταν αρχικά απαλλαγμένοι από οποιοδήποτε σύμπτωμα καρδιαγγειακής νόσου ή χρόνιες λοιμώξεις από ιούς. Κατά την εφαρμογή αυτής της έρευνας, η συχνότητα κατανάλωσης των ομάδων τροφίμων χρησιμοποιήθηκε για να ταξινομηθούν τα άτομα ως υγιή ή μη υγιή όσον αφορά τα αποτελέσματα καρδιαγγειακής νόσου εντός 10 ετών (2002-2012). Οι ομάδες τροφίμων που χρησιμοποιήθηκαν για την παρούσα διατριβή ήταν οι εξής: ψάρια, ξηροί καρποί, όσπρια, γαλακτοκομικά και λαχανικά (μερίδες: μετρούμενες σε χρόνο / εβδομάδα που καταναλώθηκε)**.**

2. ΣΚΟΠΟΣ ΔΙΑΤΡΙΒΗΣ

Σκοπός αυτής της διδακτορικής διατριβής ήταν να διερευνήσει τρόπους/μεθόδους βελτίωσης της διαγνωστικής ακρίβειας και προβλεπτικής ικανότητας κλιμάκων, θεωρητικά με χρήση πιθανοθεωρητικών τεχνικών από την επιστήμη της Στατιστικής και με τεχνικές προσομοίωσης (Monte Carlo - Bootstrap) παράγοντας κατάλληλο αλγόριθμο/πρόγραμμα με χρήση κατάλληλων λογισμικών από την επιστήμη της Πληροφορικής. Στην περίπτωση των διακριτών κλιμάκων, στόχος ήταν να αξιολογηθεί η υπόθεση, η οποία, ισχυρίζεται ότι η ευαισθησία μιας διακριτής κλίμακας αυξάνεται όταν οι διαμερίσεις/διαιρέσεις των συνιστωσών μεταβλητών της διακριτής κλίμακας επίσης αυξάνουν, βασιζόμενος σε μια θεωρητική προσέγγιση (*Bersimis et al., 2013*; *Maroulas & Panagiotakos, 2010*) για την περίπτωση κλίμακας που ακολουθεί άγνωστη κατανομή. Στην περίπτωση των συνεχών κλιμάκων, στόχος ήταν να αξιολογηθεί η υπόθεση, η οποία, ισχυρίζεται ότι η ευαισθησία μιας συνεχούς κλίμακας αυξάνεται όταν το εύρος του στηρίγματος/πεδίου ορισμού των συνιστωσών μεταβλητών του επίσης αυξάνει, βασιζόμενος σε μια θεωρητική προσέγγιση (*Bersimis et al., 2016*). Επιπλέον, στόχος της παρούσας διατριβής ήταν να αξιολογήσει την επίδραση συγκεκριμένων συντελεστών στάθμισης στις συνιστώσες μεταβλητές μιας, διακριτής ή συνεχούς, σύνθετης κλίμακας στην ακρίβεια του (*Bersimis et al., 2017*), δηλαδή, αν οι σταθμισμένες κλίμακες παρουσιάζουν μεγαλύτερη ακρίβεια στην ταξινόμηση των ατόμων σε σύγκριση με μια μη σταθμισμένη δείκτη κλίμακα, επομένως, επιμέρους στόχος της παρούσας διατριβής είναι η αξιολόγηση της τυχούσας επίδρασης προτεινόμενων σταθμίσεων που αποδίδονται στις *m* συνιστώσες μεταβλητές μιας σταθμισμένης σύνθετης κλίμακας, στην ακρίβεια αυτής, με χρήση του εμβαδού (AUC) κάτω από την Καμπύλη Λειτουργικών Χαρακτηριστικών (ROC), της ευαισθησίας και της ειδικότητας (*Kourlaba, G., Panagiotakos D.B., 2009b*) σε συγκεκριμένα διαχωριστικά σημεία (cutoff points). Πιο συγκεκριμένα γίνεται χρήση του βέλτιστου διαχωριστικού σημείου (optimum cutoff point - δηλαδή του σημείου της καμπύλης ROC όπου ισχύει $max(S_e+S_p)$). Επιπροσθέτως, σε αυτή τη διατριβή εξετάστηκε κατά πόσο ο αριθμός *m* των συνιστωσών που αθροίζονται και κατασκευάζουν μια σύνθετη κλίμακα έχουν επίδραση στην ακρίβεια της (*Bersimis et al., 2017*).

Συνοπτικά, λαμβάνοντας υπόψη τα κενά της βιβλιογραφίας (*Panagiotakos, 2009*) αναφορικά με τη μεθοδολογία κατασκευής των σύνθετων κλιμάκων, σκοπός της παρούσας διατριβής είναι να διερευνηθεί αν:

- Ο αριθμός *k* των διαμερίσεων/διαιρέσεων των συνιστωσών μεταβλητών μιας σύνθετης διακριτης κλίμακας επηρεάζει την ακρίβεια της.
- Το εύρος *l* του στηρίγματος/πεδίου ορισμού των συνιστωσών μεταβλητών μιας σύνθετης συνεχούς κλίμακας επηρεάζει την ακρίβεια της.
- Ο αριθμός *m* των συνιστωσών μεταβλητών μιας σύνθετης συνεχούς κλίμακας επηρεάζει την ακρίβεια της.
- Η απόδοση ειδικών σταθμίσεων *w* στις συνιστώσες μεταβλητές μιας σύνθετης συνεχούς κλίμακας επηρεάζει την ακρίβεια της.
- Ο βαθμός συσχέτισης των συνιστωσών μεταβλητών με το προβλεπόμενο αποτέλεσμα υγείας μιας σύνθετης συνεχούς κλίμακας επηρεάζει την ακρίβεια της.

Οι ερευνητικές υποθέσεις της διατριβής διατυπώνονται με σαφήνεια ακολούθως ως εξής:

- Η ακρίβεια μιας σύνθετης διακριτης κλίμακας βελτιώνεται με την αύξηση του αριθμού *k* των διαμερίσεων/διαιρέσεων των συνιστωσών μεταβλητών της.
- Η ακρίβεια μιας σύνθετης συνεχούς κλίμακας βελτιώνεται με την αύξηση του εύρους *l* του στηρίγματος/πεδίου ορισμού των συνιστωσών μεταβλητών της.
- Η ακρίβεια μιας σύνθετης κλίμακας βελτιώνεται με τη χρήση ειδικών σταθμίσεων *w* στις συνιστώσες μεταβλητές της, που υπολογίζονται από στατιστικές μεθόδους.
- Η ακρίβεια μιας σύνθετης κλίμακας βελτιώνεται με την αύξηση του αριθμού *m* των συνιστωσών μεταβλητών της.
- Η ακρίβεια μιας σύνθετης κλίμακας βελτιώνεται με χρήση συνιστωσών μεταβλητών που είναι μέτρια ή ισχυρά συσχετισμένες με την έκβαση υγείας που αποσκοπεί να προβλέψει η κλίμακα.

3. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1 Διακριτές Κλίμακες & Αριθμός Διαμερίσεων

Στην περίπτωση των διακριτών κλιμάκων, ενδιαφέρον παρουσιάζει η σχέση της ευαισθησίας τους ως προς τον αριθμό k των διαμερίσεων/διαιρέσεων των συνιστωσών μεταβλητών τους. Προκειμένου να εξεταστεί τόσο η φύση όσο και η ένταση της σχέσης μεταξύ του αριθμού k των διαμερίσεων/διαιρέσεων της κάθε συνιστώσας που έχει μια σύνθετη κλίμακα, ως προς την ακρίβεια της, γίνεται χρήση της ευαισθησίας. Η συνάρτηση ευαισθησίας μιας σύνθετης κλίμακας είναι κατάλληλη για να χρησιμοποιηθεί ως ένα μέτρο της ακρίβειας της κλίμακας διότι εκφράζει την πιθανότητα ορθής ταξινόμησης ενός ατόμου με ένα χαρακτηριστικό στον πληθυσμό των ατόμων με το χαρακτηριστικό αυτό.

3.1.1 Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Περίπτωση Άγνωστης κατανομής Διακριτής Κλίμακας

Στην περίπτωση μιας διακριτής κλίμακας (X^k), που κατανέμεται με άγνωστο τρόπο, διερευνάται, με τη χρήση της συνάρτησης ευαισθησίας, εάν ο αριθμός k των διαμερίσεων μιας διακριτής μεταβλητής X^k (που εκφράζει την κλίμακα) επηρεάζει την προβλεπτική ικανότητα του για την πραγματική κατάσταση ενός ατόμου. Μεθοδολογικά, έγινε χρήση, δύο διακριτών τυχαίων μεταβλητών X^{k} και $X^{\mathrm{k+1}}$ με k και k+1 διαμερίσεις αντιστοίχως που αναπαριστούν δύο κλίμακες T_1 και T_2, αντιστοίχως. Οι διακριτές μεταβλητές X^k και X^{k+1} λαμβάνουν τιμές από τα ακόλουθα σύνολα:

$$R_{X^k} = \{1, 2, \ldots, k\} \text{ και } R_{X^{k+1}} = \{1, 2, \ldots, k, k+1\}$$

Επίσης, έστω μια δίτιμη τυχαία μεταβλητή Y που ακολουθεί την κατανομή Bernoulli και εκφράζει την πραγματική κατάσταση ενός ατόμου, καθώς δείχνει την παρουσία ή την απουσία ενός χαρακτηριστικό (π.χ. στην περίπτωση μιας νόσου: Υγιείς: $Y=0$ & Ασθενείς: $Y=1$). Οι συναρτήσεις ευαισθησίας των μεταβλητών X^k και X^{k+1} σε σχέση με το Y, ορίζονται σύμφωνα με τις ακόλουθες δεσμευμένες πιθανότητες των τύπων (3.1.1.1) και (3.1.1.2) (*Wayne, 1995*):

$$S_e\left(X^k\right) = P\left(X^k > c_0 \mid Y=1\right) \qquad (3.1.1.1)$$

και

$$S_e\left(X^{k+1}\right)=P\left(X^{k+1}>c_0' \mid Y=1\right) \qquad (3.1.1.2)$$

όπου c_0 και c_0' είναι τα διαχωριστικά όρια, επιλεγμένα με κατάλληλη στατιστική μέθοδο (π.χ. από την καμπύλη λειτουργικών χαρακτηριστικών της κλίμακας) και $c_0, c_0' \in \mathbb{R}$ (*Youden, 1950*). Αν π_i^k και π_i^{k+1} συμβολίζουν τις δεσμευμένες πιθανότητες (conditional probabilities), όπου οι διακριτές τυχαίες μεταβλητές X^k και X^{k+1} λαμβάνουν μια σταθερή τιμή *i*, για ένα άτομο του πληθυσμού με το χαρακτηριστικό $(Y=1)$, τότε οι δεσμευμένες πιθανότητες π_i^k και π_i^{k+1} δίνονται από τους ακόλουθους τύπους:

$$\pi_i^k = P\left(X^k = i \mid Y=1\right),\ i=1,2,\ldots,k\text{, και } k\in\mathbb{N}, k>1 \qquad (3.1.1.3)$$

και

$$\pi_i^{k+1} = P\left(X^{k+1} = i \mid Y=1\right),\ i=1,2,\ldots,k,k+1\text{, και } k\in\mathbb{N}, k>1 \qquad (3.1.1.4)$$

Ορίζουμε την ακολουθία

$$a_i = \pi_i^k - \pi_i^{k+1},\ i=1,\ 2,\ \ldots,\ k,\ k+1 \qquad (3.1.1.5)$$

όπου τα π_i^k και π_i^{k+1} ορίζονται από τις σχέσεις (3.1.1.3) και (3.1.1.4).

■ **<u>Λήμμα 1.1</u> ($\mathbf{L_{D1}}$)**: Για την ακολουθία των διαφορών των πιθανοτήτων α_i, $i=1,2,\ldots,k$ ισχύει:

$$\sum_{i=1}^{k} a_i = \pi_{k+1}^{k+1} \quad \left(L_{D1}\right)$$

<u>Απόδειξη</u>: Αθροίζοντας από $i=1$ έως k τα δύο μέλη της σχέσης (3.1.1.5) έχουμε ότι:

$$\sum_{i=1}^{k} a_i = \sum_{i=1}^{k}\left(\pi_i^k - \pi_i^{k+1}\right) = \sum_{i=1}^{k}\pi_i^k - \sum_{i=1}^{k}\pi_i^{k+1} = 1 - \sum_{i=1}^{k}\pi_i^{k+1} = \sum_{i=1}^{k+1}\pi_i^{k+1} - \sum_{i=1}^{k}\pi_i^{k+1} = \pi_{k+1}^{k+1} \qquad ■$$

Συνδυάζοντας τις σχέσεις (3.1.1.1) και (3.1.1.2) με τις (3.1.1.3) και (3.1.1.4) οι συναρτήσεις ευαισθησίας των μεταβλητών X^k και X^{k+1} δίνονται από τους τύπους:

$$S_e\left(X^k\right) = \sum_{i=c}^{k}\pi_i^k \qquad (3.1.1.6)$$

και

$$S_e\left(X^{k+1}\right) = \sum_{i=c'}^{k+1}\pi_i^{k+1} \qquad (3.1.1.7)$$

όπου $c \in R_{X^k} = \{1, 2, \ldots, k\}$ είναι η ελάχιστη ακέραια τιμή της μεταβλητής X^k που είναι μεγαλύτερη από την τιμή c_0 και $c' \in R_{X^{k+1}} = \{1, 2, \ldots, k, k+1\}$ είναι η ελάχιστη ακέραια τιμή της μεταβλητής X^{k+1} που είναι μεγαλύτερη από την τιμή c'_0. Στο σημείο αυτό πρέπει να τονιστεί ότι δεν έχει πρακτική αξία το διαχωριστικό όριο να λαμβάνει την ελάχιστη τιμή.

<u>Συνθήκη 1.1.1</u> (C_{D11})

- Για $c \geq c' > 1$ η ακολουθία των διαφορών των πιθανοτήτων α_i, $i=1,2,\ldots,k$ ικανοποιεί την ακόλουθη ανισότητα:

$$\sum_{i=1}^{c-1} a_i \geq 0 \qquad \textbf{(C}_{\textbf{D11}}\textbf{)}$$

<u>Συνθήκη 1.1.2</u> (C_{D12})

- Για $c' > c > 1$, η ακολουθία των διαφορών των πιθανοτήτων α_i, $i=1,2,\ldots,k$, $k+1$ ικανοποιεί την ακόλουθη ανισότητα:

$$\sum_{i=1}^{c-1} a_i \geq \sum_{i=c}^{c'-1} \pi_i^{k+1} \qquad \textbf{(C}_{\textbf{D12}}\textbf{)}$$

Ερμηνεία Συνθήκης 1.1.1 (C_{D11}): Η συνθήκη 1.1.1 (**C_{D11}**), εκφράζει ότι το άθροισμα των διαφορών των πιθανοτήτων των ψευδώς αρνητικών τιμών, μεταξύ της κλίμακας με τις λιγότερες διαμερίσεις και της κλίμακας με τις περισσότερες διαμερίσεις, θα πρέπει να είναι θετικό ή ίσο του μηδέν.

Ερμηνεία Συνθήκης 1.1.2 (C_{D12}): Το άθροισμα στο αριστερό μέλος της συνθήκης **C_{D12}** αναφέρεται στη διαφορά μεταξύ των πιθανοτήτων των ψευδών αρνητικών τιμών των κλιμάκων X^k και X^{k+1}. Στο δεξιό μέλος, το άθροισμα αναφέρεται στις πιθανότητες των ψευδών αρνητικών τιμών των κλιμάκων X^{k+1}, μεταξύ των διαχωριστικών ορίων c και $c'-1$. Πρακτικά, η συνθήκη 1.1.2 (**C_{D12}**), δηλώνει ότι οι διαφορές των πιθανοτήτων των ψευδώς αρνητικών τιμών, μεταξύ της κλίμακας με τις λιγότερες διαμερίσεις και της κλίμακας με τις περισσότερες διαμερίσεις, αθροιστικά, θα πρέπει να είναι μεγαλύτερο ή ίσο, από το άθροισμα των πιθανοτήτων των ψευδώς αρνητικών τιμών της κλίμακας X^{k+1}, μεταξύ των διαχωριστικών ορίων c και $c'-1$.

<u>Πρόταση 1.1</u> (P_{D1}). Έστω $T_1 = X^k$ και $T_2 = X^{k+1}$ κλίμακες που κατανέμονται με άγνωστο τρόπο με στήριγμα $R_{X^k} = \{1, 2, \ldots, k\}$ και $R_{X^{k+1}} = \{1, 2, \ldots, k, k+1\}$, αντιστοίχως. Η ευαισθησία των κλιμάκων $T_1 = X^k$ και $T_2 = X^{k+1}$ είναι

$S_e\left(T_1\right)=S_e\left(X^k\right)$ και $S_e\left(T_2\right)=S_e\left(X^{k+1}\right)$, αντιστοίχως. Τότε, για όλες τις δυνατές τιμές των διαχωριστικών ορίων c και c' ισχύει $S_e\left(X^k\right)\le S_e\left(X^{k+1}\right)$, υπό τις συνθήκες 1.1.1 και 1.1.2.

► ***<u>Απόδειξη</u>***: Προκειμένου να αποδειχθεί η πρόταση, αρκεί να δειχθεί ότι για οποιαδήποτε $k\in\mathbb{N}$ ισχύει η ακόλουθη ανισότητα:

$$S_e\left(X^k\right)\le S_e\left(X^{k+1}\right) \tag{3.1.1.8}$$

Σύμφωνα με τις σχέσεις (3.1.1.6) και (3.1.1.7), η σχέση (3.1.1.8) γίνεται:

$$\sum_{i=c}^{k}\pi_i^k\le\sum_{i=c'}^{k+1}\pi_i^{k+1} \tag{3.1.1.9}$$

Ως εκ τούτου, οι περιπτώσεις που διακρίνονται, σύμφωνα με τη σχετική θέση μεταξύ των διαχωριστικών ορίων c και c' είναι οι ακόλουθες:

- **Περίπτωση 1**: $c=c'>1$

Σύμφωνα με η σχέση (3.1.1.6), η ευαισθησία της μεταβλητής X^k γίνεται:

$$S_e\left(X^k\right)=\sum_{i=c}^{k}\pi_i^k=\sum_{i=c}^{k}\left(\pi_i^{k+1}+a_i\right)=\sum_{i=c}^{k}\pi_i^{k+1}+\sum_{i=c}^{k}a_i\overset{(C_{D11})}{\le}\sum_{i=c}^{k}\pi_i^{k+1}+\sum_{i=1}^{c-1}a_i+\sum_{i=c}^{k}a_i$$

$$=\sum_{i=c}^{k}\pi_i^{k+1}+\sum_{i=1}^{k}a_i\overset{(L_{D1})}{=}\sum_{i=c}^{k}\pi_i^{k+1}+\pi_{k+1}^{k+1}=\sum_{i=c}^{k+1}\pi_i^{k+1}=S_e\left(X^{k+1}\right)$$

Επομένως προκύπτει η ανισότητα

$$S_e\left(X^k\right)\le S_e\left(X^{k+1}\right).$$

Συνεπώς η (3.1.1.8) ισχύει.

- **Περίπτωση 2**: $c>c'>1$

Σύμφωνα με την σχέση (3.1.1.6), η ευαισθησία της μεταβλητής X^k γίνεται:

$$S_e\left(X^k\right)=\sum_{i=c}^{k}\pi_i^k=\sum_{i=c}^{k}\left(\pi_i^k+a_i\right)=\sum_{i=c}^{k}\pi_i^{k+1}+\sum_{i=c}^{k}a_i\overset{(C_{D11})}{\le}\sum_{i=c'}^{k}\pi_i^{k+1}+\sum_{i=1}^{k}a_i\overset{(L_{D1})}{=}\sum_{i=c'}^{k}\pi_i^{k+1}+\pi_{k+1}^{k+1}=$$

$$=\sum_{i=c'}^{k+1}\pi_i^{k+1}=S_e\left(X^{k+1}\right)$$

Επομένως προκύπτει η ανισότητα:

$$S_e\left(X^k\right)\le S_e\left(X^{k+1}\right)$$

Συνεπώς η (3.1.1.8) ισχύει.

- **Περίπτωση 3**: $c'>c>1$

Έστω η ακόλουθη διαφορά:

$$S_e\left(X^{k+1}\right)-S_e\left(X^{k}\right)=\sum_{i=c'}^{k+1}\pi_i^{k+1}-\sum_{i=c}^{k}\pi_i^{k}=\sum_{i=c'}^{k}\pi_i^{k+1}+\pi_{k+1}^{k+1}-\left(\sum_{i=c}^{c'-1}\pi_i^{k}+\sum_{i=c'}^{k}\pi_i^{k}\right)$$

$$=\sum_{i=c'}^{k}\left(\pi_i^{k+1}-\pi_i^{k}\right)+\pi_{k+1}^{k+1}-\sum_{i=c}^{c'-1}\pi_i^{k}=\sum_{i=c'}^{k}\left(-a_i\right)+\sum_{i=1}^{k}a_i-\sum_{i=c}^{c'-1}\pi_i^{k}=\sum_{i=1}^{c'-1}a_i-\sum_{i=c}^{c'-1}\pi_i^{k}$$

$$=\sum_{i=1}^{c-1}a_i+\sum_{i=c}^{c'-1}\left(a_i-\pi_i^{k}\right)=\sum_{i=1}^{c-1}a_i+\sum_{i=c}^{c'-1}\left(-\pi_i^{k+1}\right)=\sum_{i=1}^{c-1}a_i-\sum_{i=c}^{c'-1}\pi_i^{k+1}\overset{(C_{D12})}{\geq}0$$

Επομένως προκύπτει η ανισότητα

$$S_e\left(X^{k}\right)\leq S_e\left(X^{k+1}\right)$$

Συνεπώς η (3.1.1.8) ισχύει. ▲

<u>Πόρισμα 1.1</u> (C_{D1}): Για οποιαδήποτε $k\in\mathbb{N}$, η ευαισθησία της κλίμακας X^k, είναι μια μη φθίνουσα συνάρτηση του k.

Παρατήρηση: Η συνάρτηση ευαισθησίας είναι μια συνάρτηση μιας μεταβλητής, του αριθμού των διαιρέσεων *k* και το διαχωριστικό όριο *c* θεωρείται μια παράμετρος. Ο μαθηματικός ορισμός της αύξουσας μονοτονίας μιας συνάρτησης μιας μεταβλητής εκφράζεται από την περίπτωση 1, δηλαδή την περίπτωση του κοινού διαχωριστικού ορίου ($c=c'>1$).

3.1.2 Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Περίπτωση Ομοιόμορφης κατανομής Κλίμακας

Στην παράγραφο αυτή παρατίθεται η θεωρητική απόδειξη της πρότασης (Πρόταση 1.2 - (P_{DU1})) που εκφράζει την αύξουσα μονοτονία μιας μη σύνθετης κλίμακας ως προς τον αριθμό των διαμερίσεων της, στην περίπτωση που ακολουθεί την Ομοιόμορφη κατανομή (Uniform Distribution). Δηλαδή, περιγράφεται η περίπτωση, όπου μια κλίμακα εκφράζεται από μια διακριτή μεταβλητή X^k που ακολουθεί την ομοιόμορφη κατανομή ($X^k\sim U_D\left(1,k\right)$). Όμοια με την προηγούμενη παράγραφο γίνεται χρήση δύο διακριτών τυχαίων μεταβλητών X^k και X^{k+1} με *k* και *k*+1 διαμερίσεις, αντίστοιχα. Στην περίπτωση της ομοιόμορφης κατανομής, οι δεσμευμένες πιθανότητες π_i^k και π_i^{k+1} των μαθηματικών εκφράσεων (3.1.1.3) και (3.1.1.4), όπου οι τυχαίες διακριτές μεταβλητές X^k και X^{k+1} λαμβάνουν μια σταθερή τιμή *i*, για τα άτομα του πληθυσμού (*Y*=1) των ασθενών, δίνονται από τους ακόλουθους τύπους (3.1.2.1) και (3.1.2.2):

$$\pi_i^k = P\left(X^k = i \mid Y=1\right) = \frac{1}{k}, \quad i=1,2,\ldots,k, \quad k\in\mathbb{N},\ k>1 \qquad (3.1.2.1)$$

και

$$\pi_i^{k+1} = P\left(X^{k+1} = i \mid Y=1\right) = \frac{1}{k+1}, \quad i=1,2,\ldots,k,k+1, \quad k\in\mathbb{N},\ k>1 \qquad (3.1.2.2)$$

Ορίζουμε την ακολουθία α_i, $i=1,2,\ldots,k$:

$$a_i = \pi_i^k - \pi_i^{k+1} = \frac{1}{k} - \frac{1}{k+1} = \frac{1}{k\cdot(k+1)} \qquad (3.1.2.3)$$

όπου τα π_i^k και π_i^{k+1} ορίζονται από τις σχέσεις (3.1.2.1) και (3.1.2.2).

■ **<u>Λήμμα 1.2</u> (L_{DU1})**: Για την ακολουθία των διαφορών των πιθανοτήτων α_i, $i=1,2,\ldots,k$ ισχύει:

$$\sum_{i=1}^{k} a_i = \pi_{k+1}^{k+1} \quad \left(L_{DU1}\right)$$

<u>Απόδειξη</u>: Αθροίζοντας από $i=1$ έως k τα δύο μέλη της σχέσης (3.1.1.5) έχουμε ότι:

$$\sum_{i=1}^{k} a_i = \sum_{i=1}^{k}\left(\pi_i^k - \pi_i^{k+1}\right) = \sum_{i=1}^{k}\left(\frac{1}{k} - \frac{1}{k+1}\right) = \sum_{i=1}^{k}\left(\frac{1}{k(k+1)}\right) = \frac{k}{k(k+1)} = \frac{1}{(k+1)} = \pi_{k+1}^{k+1} \qquad ■$$

Στην περίπτωση αυτή, σύμφωνα με τον ορισμό της ευαισθησίας, θα έχουμε για τους δείκτες X^k και X^{k+1}:

$$S_e\left(X^k\right) = \sum_{i=c}^{k} \pi_i^k = \frac{k-c+1}{k} \qquad (3.1.2.4)$$

και

$$S_e\left(X^{k+1}\right) = \sum_{i=c'}^{k+1} \pi_i^{k+1} = \frac{k-c'+2}{k+1} \qquad (3.1.2.5)$$

Όπου $c\in R_{X^k} = \{1,2,\ldots,k\}$ και $c'\in R_{X^{k+1}} = \{1,2,\ldots,k,k+1\}$.

<u>Σημείωση:</u> Στην περίπτωση της ομοιόμορφης κατανομής των κλιμάκων, δεν είναι απαραίτητη η Συνθήκη 1.1.1 (CD_{11}), καθώς η ανισότητα $\sum_{i=1}^{c-1} a_i \geq 0$ ισχύει για $c\geq c'>1$.

<u>Συνθήκη 1.2.1</u> (C_{DU1})

- Για $c'>c>1$, η ακολουθία των διαφορών των πιθανοτήτων α_i, $i=1,2,\ldots,k$, $k+1$ ικανοποιεί την ακόλουθη ανισότητα:

$$c(k+1) - c'k \geq 1 \qquad \textbf{(}C_{DU1}\textbf{)}$$

Η συνθήκη **(C_{DU1})** αποτελεί μια εναλλακτική έκφραση της συνθήκης **(C_{D12})** στην περίπτωση που η μεταβλητή X^k κατανέμεται ομοιόμορφα.

<u>Πρόταση 1.2</u>. (P_{DU1}). Έστω $T_1 = X^k$ και $T_2 = X^{k+1}$ κλίμακες που κατανέμονται Ομοιόμορφα με στήριγμα $R_{X^k} = \{1, 2, \ldots, k\}$ και $R_{X^{k+1}} = \{1, 2, \ldots, k, k+1\}$, αντιστοίχως. Η ευαισθησία των κλιμάκων $T_1 = X^k$ και $T_2 = X^{k+1}$ είναι $S_e(T_1) = S_e(X^k)$ και $S_e(T_2) = S_e(X^{k+1})$, αντιστοίχως. Τότε, για όλες τις δυνατές τιμές των διαχωριστικών ορίων c και c' ισχύει $S_e(X^k) \le S_e(X^{k+1})$, υπό τη συνθήκη 1.2.1.

Απόδειξη: Σύμφωνα με τις σχέσεις (3.1.2.3) και (3.1.2.4) αρκεί να αποδειχθεί ότι:

$$\frac{k-c+1}{k} \le \frac{k-c'+2}{k+1} \qquad (3.1.2.6)$$

Έστω η διαφορά:

$$S_e(X^{k+1}) - S_e(X^k) = \sum_{i=c'}^{k+1} \pi_i^{k+1} - \sum_{i=c}^{k} \pi_i^k = \frac{k-c'+2}{k+1} - \frac{k-c+1}{k}$$

$$= \frac{(k-c'+2)\cdot k - (k-c+1)\cdot(k+1)}{k\cdot(k+1)} = \frac{k^2 - kc' + 2k - k^2 - k + kc + c - k - 1}{k\cdot(k+1)}$$

$$S_e(X^{k+1}) - S_e(X^k) = \frac{-kc' + kc + c - 1}{k\cdot(k+1)} \qquad (3.1.2.7)$$

Οι περιπτώσεις οι οποίες διακρίνονται, σύμφωνα με τη σχετική θέση μεταξύ c και c' είναι οι εξής:

- **Περίπτωση 1**: $c = c' > 1$

Σύμφωνα με τη σχέση (3.1.2.7) για αυτή την περίπτωση, η διαφορά $S_e(X^{k+1}) - S_e(X^k)$ γίνεται:

$$S_e(X^{k+1}) - S_e(X^k) = \frac{c-1}{k\cdot(k+1)} > 0$$

επειδή $c > 1$ και $k\cdot(k+1) > 0$. Έτσι, για $c = c' > 1$, η σχέση (3.1.2.5) αποδείχτηκε, δηλαδή η ευαισθησία είναι μια αύξουσα (μη φθίνουσα) συνάρτηση του k.

- Περίπτωση 2: $c > c' > 1$

Σύμφωνα με τη σχέση (3.1.2.7) για αυτή την περίπτωση, η διαφορά $S_e(X^{k+1}) - S_e(X^k)$ γίνεται:

$$S_e\left(X^{k+1}\right)-S_e\left(X^k\right)=\frac{-k\cdot c'+k\cdot c+c-1}{k\cdot(k+1)}=\frac{(-c'+c)\cdot k+c-1}{k\cdot(k+1)}>0$$

επειδή $c>1$, $c-c'>0$ και $k\cdot(k+1)>0$. Έτσι, η σχέση (3.1.2.5) έχει αποδειχθεί, δηλαδή η ευαισθησία είναι μια αύξουσα (μη φθίνουσα) συνάρτηση ως προς k.

- Περίπτωση 3: $c'>c>1$

Σύμφωνα με τη σχέση (3.1.2.7) για αυτή την περίπτωση, η διαφορά $S_e\left(X^{k+1}\right)-S_e\left(X^k\right)$ γίνεται:

$$S_e\left(X^{k+1}\right)-S_e\left(X^k\right)=\frac{-k\cdot c'+k\cdot c+c-1}{k\cdot(k+1)}=\frac{(k+1)\cdot c+k\cdot c'-1}{k\cdot(k+1)}>0$$

Όπου $\frac{(k+1)\cdot c+k\cdot c'-1}{k\cdot(k+1)}$ είναι μία μη αρνητική ποσότητα λόγω της συνθήκης 2.

Ο αριθμητής του κλάσματος αυτού γίνεται:

$$(k+1)\cdot c+k\cdot c'\geq 1$$

και στη συνέχεια γίνεται:

$$\frac{c}{k}-\frac{c'}{k+1}\geq\frac{1}{k\cdot(k+1)}$$

Επομένως, η σχέση (3.1.2.6) έχει αποδειχθεί.

<u>Πόρισμα 1.2</u> (C_{DU1}): Για οποιαδήποτε $k\in\mathbb{N}$, η ευαισθησία της κλίμακας X^k που κατανέμεται ομοιόμορφα, είναι μια μη φθίνουσα συνάρτηση του k.

3.1.3 Σύνθετη Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Περίπτωση Άγνωστης κατανομής Κλίμακας

Οι σύνθετες κλίμακες που χρησιμοποιούνται στην πράξη για την ανίχνευση ή πρόβλεψη της τιμής ενός δυαδικού αποτελέσματος *Y*, που εκφράζει την κατάσταση ενός ατόμου, είναι συνήθως, το άθροισμα άλλων διακριτών συνιστωσών μεταβλητών X_j^k, $j=1,2,\ldots,m$. Για το λόγο αυτό, η ευαισθησία μιας σύνθετης κλίμακας $T_m^k=\sum_{j=1}^m X_j^k$, είναι συχνά ενδιαφέρον να εξεταστεί επειδή πολλά χαρακτηριστικά σε διάφορους τομείς, εξαρτώνται από πολλούς παράγοντες και η ορθή ανίχνευση τους είναι ιδιαίτερα κρίσιμης σημασίας. Σε αυτή την παράγραφο παρατίθεται το μεθοδολογικό πλαίσιο της διερεύνησης για το εάν ο αριθμός των διαμερίσεων κάθε

διακριτής μεταβλητής X_j^k επηρεάζει τη συνάρτηση ευαισθησίας S_e, ή όχι της σύνθετης κλίμακας T_m^k, στην περίπτωση που η κατανομή των μεταβλητών X_j^k, $j=1,2,\ldots,m$ είναι άγνωστη. Για αυτό το σκοπό, δύο μεταβλητές T_m^k και T_m^{k+1} θεωρούνται ως ακολούθως:

$$T_m^k = X_1^k + X_2^k + \ldots + X_m^k = \sum_{j=1}^{m} X_j^k \qquad (3.1.3.1)$$

και

$$T_m^{k+1} = X_1^{k+1} + X_2^{k+1} + \ldots + X_m^{k+1} = \sum_{j=1}^{m} X_j^{k+1} \qquad (3.1.3.2)$$

Οι μεταβλητές $X_1^k, X_2^k, \ldots, X_m^k$ και $X_1^{k+1}, X_2^{k+1}, \ldots, X_m^{k+1}$ θεωρούνται ως ανεξάρτητες και ακολουθούν την ίδια κατανομή. Οι μεταβλητές T_m^k και T_m^{k+1} λαμβάνουν τιμές από τα ακόλουθα σύνολα:

$$R_{T_m^k} = \{m, m+1, \ldots, mk\} \quad \text{και} \quad R_{T_m^{k+1}} = \{m, m+1, \ldots, mk, \ldots, m(k+1)\}$$

Οι συναρτήσεις ευαισθησίας των κλιμάκων T_m^k και T_m^{k+1}, σε σχέση με το *Y*, ορίζονται ως οι δεσμευμένες πιθανότητες από τους ακόλουθους τύπους (3.1.3.3) και (3.1.3.4) (*Wayne. 1995*)*:*

$$S_e\left(T_m^k\right) = P\left(T_m^k > l_{0,m}^k \,|\, Y=1\right) \qquad (3.1.3.3)$$

και

$$S_e\left(T_m^{k+1}\right) = P\left(T_m^{k+1} > l_{0,m}^{k+1} \,|\, Y=1\right) \qquad (3.1.3.4)$$

όπου $l_{0,m}^k$ και $l_{0,m}^{k+1}$ είναι τα διαγνωστικά κατώφλια που επιλέγονται με κατάλληλη στατιστική μέθοδο και $l_{0,m}^k, l_{0,m}^{k+1} \in \mathbb{R}$ (*Youden, 1950*). Εάν $\pi_{m,t}^k$ και $\pi_{m,t}^{k+1}$ εκφράζουν τις δεσμευμένες πιθανότητες που οι τυχαίες μεταβλητές T_m^k και T_m^{k+1} λαμβάνουν μια σταθερή τιμή *t* για ένα άτομο του πληθυσμού που έχει ένα συγκεκριμένο χαρακτηριστικό $(Y=1)$ τότε οι πιθανότητες $\pi_{m,t}^k$ και $\pi_{m,t}^{k+1}$ δίνονται από τους ακόλουθους τύπους (3.1.3.5) και (3.1.3.6):

$$\pi_{m,t}^k = P\left(T_m^k = t \,|\, Y=1\right),\ t = m, m+1, \ldots, mk, \text{ και } k \in \mathbb{N},\ k > 1 \qquad (3.1.3.5)$$

και

$$\pi_{m,t}^{k+1} = P\left(T_m^{k+1} = t \,|\, Y=1\right),\ t = m, m+1, \ldots, mk, \ldots, m(k+1),\ k \in \mathbb{N},\ k > 1 \quad (3.1.3.6).$$

Ορίζουμε την ακολουθία:

$$a_{m,t} = \pi_{m,t}^{k} - \pi_{m,t}^{k+1} \qquad (3.1.3.7)$$

όπου τα $\pi_{m,t}^{k}$ και $\pi_{m,t}^{k+1}$ ορίζονται από τις σχέσεις (3.1.3.5) και (3.1.3.6).

■ **<u>Λήμμα 1.2</u> (L_{Dm})**. Για την ακολουθία των διαφορών των πιθανοτήτων $a_{m,t}$, $t = m, m+1, \ldots, mk$ ισχύει:

$$\sum_{t=m}^{mk} a_{m,t} = \sum_{t=mk+1}^{m(k+1)} \pi_{m,t}^{k+1} \quad \left(L_{D2}\right)$$

<u>Απόδειξη</u>: Αθροίζοντας από $t = m$ έως mk κατά μέρη τη σχέση (3.1.3.7) έχουμε ότι:

$$\sum_{t=m}^{mk} a_{m,t} = \sum_{t=m}^{mk}\left(\pi_{m,t}^{k} - \pi_{m,t}^{k+1}\right) = \sum_{t=m}^{mk} \pi_{m,t}^{k} - \sum_{t=m}^{mk} \pi_{m,t}^{k+1} = 1 - \sum_{t=m}^{mk} \pi_{m,t}^{k+1} = \sum_{t=m}^{m(k+1)} \pi_{m,t}^{k+1} - \sum_{t=m}^{mk} \pi_{m,t}^{k+1} = \sum_{t=mk+1}^{m(k+1)} \pi_{m,t}^{k+1} \quad ■$$

Επομένως, η συνάρτηση ευαισθησία, για τις σύνθετες κλίμακες T_m^k και T_m^{k+1} δίνονται από τους ακόλουθους τύπους (3.1.3.8) και (3.1.3.9):

$$S_e\left(T_m^k\right) = P\left(T_m^k \geq l_m^k \,|\, Y = 1\right) = \sum_{t=l_m^k}^{mk} \pi_{m,t}^{k} \qquad (3.1.3.8)$$

και

$$S_e\left(T_m^{k+1}\right) = P\left(T_m^{k+1} \geq l_m^{k+1} \,|\, Y = 1\right) = \sum_{t=l_m^{k+1}}^{m(k+1)} \pi_{m,t}^{k+1} \qquad (3.1.3.9)$$

όπου $l_m^k \in R_{T_m^k} = \{m, m+1, \ldots, mk\}$ είναι η ελάχιστη ακέραια τιμή της κλίμακας T_m^k μεγαλύτερη από $l_{0,m}^k$ και $l_m^{k+1} \in R_{T_m^{k+1}} = \{m, m+1, \ldots, mk, \ldots, m(k+1)\}$ είναι η ελάχιστη ακέραια τιμή της κλίμακας T_m^{k+1} μεγαλύτερη από $l_{0,m}^{k+1}$.

<u>Συνθήκη 1.3.1</u> (C_{Dm1})

- Για $l_m^k \geq l_m^{k+1} > m$ η ακολουθία των διαφορών των πιθανοτήτων $a_{m,t}$, $t = m, m+1, \ldots, mk$ ικανοποιεί την ακόλουθη ανισότητα:

$$\sum_{t=m}^{l_m^k - 1} a_{m,t} \geq 0 \qquad \textbf{(C}_{\textbf{D11}}\textbf{)}$$

<u>Συνθήκη 1.3.2</u> (C_{Dm2}).

- Για $l_m^k < l_m^{k+1}$, η ακολουθία των διαφορών των πιθανοτήτων $a_{m,t}, t = m, m+1, \ldots, mk$ ικανοποιεί την ακόλουθη ανισότητα:

$$\sum_{t=m}^{l_m^k - 1} a_{m,t} \geq \sum_{t=l_m^k}^{l_m^{k+1} - 1} \pi_{m,t}^{k+1} \qquad \textbf{(C}_{\textbf{Dm}}\textbf{)}$$

Ερμηνεία Συνθήκης 1.3.1 (C_{Dm1}): Η συνθήκη 1.3.1 (**C_{Dm1}**), εκφράζει ότι το άθροισμα των διαφορών των πιθανοτήτων των ψευδώς αρνητικών τιμών, μεταξύ της

κλίμακας με τις λιγότερες διαμερίσεις και της κλίμακας με τις περισσότερες διαμερίσεις, θα πρέπει να είναι θετικό ή ίσο του μηδέν.

Ερμηνεία Συνθήκης 1.3.2 (C_{Dm2}): Το άθροισμα στο αριστερό μέλος της συνθήκης C_{D12} αναφέρεται στη διαφορά μεταξύ των πιθανοτήτων των ψευδών αρνητικών τιμών των κλιμάκων T_m^k και T_m^{k+1}. Στο δεξιό μέλος, το άθροισμα αναφέρεται στις πιθανότητες των ψευδών αρνητικών τιμών των κλιμάκων T_m^{k+1}, μεταξύ των διαχωριστικών ορίων l_m^k και $l_m^{k+1}-1$. Πρακτικά, η συνθήκη 1.3.2 (C_{Dm2}), δηλώνει ότι οι διαφορές των πιθανοτήτων των ψευδώς αρνητικών τιμών, μεταξύ της κλίμακας με τις λιγότερες διαμερίσεις και της κλίμακας με τις περισσότερες διαμερίσεις, αθροιστικά, θα πρέπει να είναι μεγαλύτερο ή ίσο, από το άθροισμα των πιθανοτήτων των ψευδώς αρνητικών τιμών της κλίμακας T_m^{k+1}, μεταξύ των διαχωριστικών ορίων l_m^k και $l_m^{k+1}-1$.

<u>Πρόταση 1.3</u>. (P_{Dm}). Έστω T_m^k και T_m^{k+1} κλίμακες που κατανέμονται με άγνωστο τρόπο με πεδίο ορισμού $R_{T_m^k}=\{m,m+1,\ldots,mk\}$ και $R_{T_m^{k+1}}=\{m,m+1,\ldots,mk,\ldots,m(k+1)\}$, αντιστοίχως. Η ευαισθησία των κλιμάκων T_m^k και T_m^{k+1} είναι $S_e\left(T_m^k\right)$ και $S_e\left(T_m^{k+1}\right)$, αντιστοίχως. Τότε, για όλες τις δυνατές τιμές των διαχωριστικών ορίων l_m^k και l_m^{k+1} ισχύει $S_e\left(T_m^k\right)\le S_e\left(T_m^{k+1}\right)$, υπό τις συνθήκες 1.3.1 και 1.3.2.

► ***Απόδειξη***. Προκειμένου να αποδειχθεί η προαναφερθείσα πρόταση, αρκεί να δειχθεί ότι για κάθε $k\in\mathbb{N}$ ισχύει η ακόλουθη ανισότητα (3.1.3.10):

$$S_e\left(T_m^k\right)\le S_e\left(T_m^{k+1}\right) \qquad (3.1.3.10)$$

Σύμφωνα με τον ορισμό της συνάρτησης ευαισθησίας η προηγούμενη σχέση γίνεται όπως η (3.1.3.11) δείχνει:

$$\sum_{t=l_m^k}^{mk}\pi_{m,t}^k \le \sum_{t=l_m^{k+1}}^{m(k+1)}\pi_{m,t}^{k+1} \qquad (3.1.3.11)$$

Ως εκ τούτου, οι περιπτώσεις που διακρίνονται, σύμφωνα με τη σχετική θέση μεταξύ των διαχωριστικών ορίων l_m^k και l_m^{k+1} είναι οι ακόλουθες:

- **Περίπτωση 1**: $l_m^k=l_m^{k+1}>m$

Η ευαισθησία της μεταβλητής T_m^k γίνεται:

$$S_e\left(T_m^k\right)=\sum_{t=l_m^k}^{mk}\pi_{m,t}^k=\sum_{t=l_m^k}^{mk}\left(\pi_{m,t}^{k+1}+a_{m,t}\right)=\sum_{t=l_m^k}^{mk}\pi_{m,t}^{k+1}+\sum_{t=l_m^k}^{mk}a_{m,t}\overset{(C_{Dm1})}{\leq}\sum_{t=l_m^k}^{mk}\pi_{m,t}^{k+1}+\sum_{t=m}^{l_m^k-1}a_{m,t}+\sum_{t=l_m^k}^{mk}a_{m,t}=$$

$$=\sum_{t=l_m^k}^{mk}\pi_{m,t}^{k+1}+\sum_{t=m}^{mk}a_{m,t}\overset{(L_{Dm})}{=}\sum_{t=l_m^k}^{mk}\pi_{m,t}^{k+1}+\sum_{t=mk+1}^{m(k+1)}\pi_{m,t}^{k+1}=\sum_{t=l_m^k}^{m(k+1)}\pi_{m,t}^{k+1}=S_e\left(T_m^{k+1}\right)$$

Επομένως προκύπτει η ανισότητα:

$$S_e\left(T_m^k\right)\leq S_e\left(T_m^{k+1}\right)$$

Συνεπώς η (3.1.13.10) ισχύει.

- **Περίπτωση 2**: $l_m^k>l_m^{k+1}>m$

Η ευαισθησία της μεταβλητής T_m^k γίνεται:

$$S_e\left(T_m^k\right)=\sum_{t=l_m^k}^{mk}\pi_{m,t}^k=\sum_{t=l_m^k}^{mk}\left(\pi_{m,t}^{k+1}+a_{m,t}\right)=\sum_{t=l_m^k}^{mk}\pi_{m,t}^{k+1}+\sum_{t=l_m^k}^{mk}a_{m,t}\overset{(C_{Dm1})}{\leq}\sum_{t=l_m^{k+1}}^{mk}\pi_{m,t}^{k+1}+\sum_{t=m}^{l_m^k-1}a_{m,t}+\sum_{t=l_m^k}^{mk}a_{m,t}=$$

$$=\sum_{t=l_m^{k+1}}^{mk}\pi_{m,t}^{k+1}+\sum_{t=m}^{mk}a_{m,t}\overset{(L_{Dm})}{=}\sum_{t=l_m^{k+1}}^{mk}\pi_{m,t}^{k+1}+\sum_{t=mk+1}^{m(k+1)}\pi_{m,t}^{k+1}=\sum_{t=l_m^{k+1}}^{m(k+1)}\pi_{m,t}^{k+1}=S_e\left(T_m^{k+1}\right)$$

Επομένως προκύπτει η ανισότητα:

$$S_e\left(T_m^k\right)\leq S_e\left(T_m^{k+1}\right)$$

Δεδομένου ότι η ποσότητα $\pi_{m,t}^{k+1}$ είναι ένας μη αρνητικός αριθμός, ως πιθανότητα, η ακόλουθη ανισότητα ισχύει:

$$\sum_{t=l_m^k}^{mk}\pi_{m,t}^{k+1}\leq\sum_{t=l_m^{k+1}}^{m(k+1)}\pi_{m,t}^{k+1}$$

Συνεπώς η (3.1.13.10) ισχύει.

- **Περίπτωση 3**: $l_m^{k+1}>l_m^k>m$

Η ευαισθησία της μεταβλητής T_m^k γίνεται:

$$S_e\left(T_m^k\right)=\sum_{t=l_m^k}^{mk}\pi_{m,t}^k=\sum_{t=l_m^k}^{mk}\left(\pi_{m,t}^{k+1}+a_{m,t}\right)=\sum_{t=l_m^k}^{mk}\pi_{m,t}^{k+1}+\sum_{t=l_m^k}^{mk}a_{m,t}=\sum_{t=l_m^k}^{l_m^{k+1}-1}\pi_{m,t}^{k+1}+\sum_{t=l_m^{k+1}}^{mk}\pi_{m,t}^{k+1}+\sum_{t=l_m^k}^{mk}a_{m,t}$$

$$\overset{(C_{Dm2})}{\leq}\sum_{t=l_m^{k+1}}^{mk}\pi_{m,t}^{k+1}+\sum_{t=m}^{l_m^k-1}a_{m,t}+\sum_{t=l_m^k}^{mk}a_{m,t}=\sum_{t=l_m^{k+1}}^{mk}\pi_{m,t}^{k+1}+\sum_{t=m}^{mk}a_{m,t}\overset{(L_{Dm})}{=}\sum_{t=l_m^{k+1}}^{mk}\pi_{m,t}^{k+1}+\sum_{t=mk+1}^{m(k+1)}\pi_{m,t}^{k+1}=\sum_{t=l_m^{k+1}}^{m(k+1)}\pi_{m,t}^{k+1}=S_e\left(T_m^{k+1}\right)$$

Επομένως προκύπτει η ανισότητα:

$$S_e\left(T_m^k\right)\leq S_e\left(T_m^{k+1}\right)$$

Συνεπώς η (3.1.13.10) ισχύει. ▲

<u>Πόρισμα 1.3</u> (C_{Dm}): Για οποιαδήποτε $k \in \mathbb{N}$, η ευαισθησία της σύνθετης κλίμακας T_m^k, είναι μια μη φθίνουσα συνάρτηση του k.

3.1.4 Σύνθετη Διακριτή Κλίμακα & Αριθμός Διαμερίσεων - Ομοιόμορφη Κατανομή Συνιστωσών Μεταβλητών

Σε αυτήν την παράγραφο παρουσιάζεται η διερεύνηση του εάν ο αριθμός των διαμερίσεων της εκάστοτε διακριτής μεταβλητής X_j^k, $j=1,2,\ldots,m$ επηρεάζει την ευαισθησία μιας σύνθετης κλίμακας $T_m^k = \sum_{j=1}^m X_j^k$, στην περίπτωση που οι διακριτές μεταβλητές $X_1^k, X_2^k, \ldots, X_m^k$ ακολουθούν τη διακριτή ομοιόμορφη κατανομή, καθώς επίσης και η εφαρμογή της πρότασης 1.3 στην περίπτωση αυτή. Οι δεσμευμένες πιθανότητες $\pi_{m,t}^k = P\left(T_m^k = t \mid Y=1\right)$ και $\pi_{m,t}^{k+1} = P\left(T_m^{k+1} = t \mid Y=1\right)$ των μεταβλητών T_m^k και T_m^{k+1} δίνονται από τους τύπους (3.1.4.1) και (3.1.4.2) (*Charalambides. 2005*):

$$\pi_{m,t}^k = P\left(T_m^k = t \mid Y=1\right) = \frac{1}{k^m} \sum_{r=0}^{\left[\frac{t-m}{k}\right]} (-1)^r \binom{m}{r} \binom{t-kr-1}{m-1} \tag{3.1.4.1}$$

και

$$\pi_{m,t}^{k+1} = P\left(T_m^{k+1} = t \mid Y=1\right) = \frac{1}{(k+1)^m} \sum_{r=0}^{\left[\frac{t-m}{k+1}\right]} (-1)^r \binom{m}{r} \binom{t-(k+1)r-1}{m-1} \tag{3.1.4.2}$$

Ορίζουμε την ακολουθία $a_{m,t}$, $t=m, m+1, \ldots, mk$

$$a_{m,t} = \pi_{m,t}^k - \pi_{m,t}^{k+1} \tag{3.1.4.3}$$

Η ακολουθία των διαφορών των πιθανοτήτων $a_{m,t}$ ικανοποιεί τις συνθήκες (C_{Dm1}) και (C_{Dm2}), από την προηγούμενη ενότητα, στην οποία οι m μεταβλητές ακολουθούν μια άγνωστη κατανομή. Στην περίπτωση που η κατανομή των μεταβλητών $X_1^k, X_2^k, \ldots, X_m^k$ είναι η διακριτή ομοιόμορφη, τότε οι συναρτήσεις ευαισθησίας των μεταβλητών T_m^k και T_m^{k+1} δίνονται από τους τύπους (3.1.4.4) και (3.1.4.5):

$$S_e\left(T_m^k\right) = \sum_{t=l_m^k}^{mk} \pi_{m,t}^k = \sum_{t=l_m^k}^{mk} \left\{ \frac{1}{k^m} \sum_{r=0}^{\left[\frac{t-m}{k}\right]} (-1)^r \binom{m}{r} \binom{t-kr-1}{m-1} \right\} \tag{3.1.4.4}$$

και

$$S_e\left(T_m^{k+1}\right)=\sum_{t=l_m^{k+1}}^{m(k+1)}\pi_{m,t}^{k+1}=\sum_{t=l_m^{k+1}}^{m(k+1)}\left\{\frac{1}{(k+1)^m}\sum_{r=0}^{\left[\frac{t-m}{k+1}\right]}(-1)^r\binom{m}{r}\binom{t-(k+1)r-1}{m-1}\right\} \quad (3.1.4.5)$$

Επομένως, προκειμένου να διερευνηθεί η μονοτονία της συνάρτησης ευαισθησίας, πρέπει να εξεταστεί το πρόσημο της ακόλουθης διαφοράς:

$$S_e\left(T_m^{k+1}\right)-S_e\left(T_m^{k}\right)=\sum_{t=l_m^{k+1}}^{m(k+1)}\pi_{m,t}^{k+1}-\sum_{t=l_m^{k}}^{mk}\pi_{m,t}^{k}=$$

$$\sum_{t=l_m^{k+1}}^{m(k+1)}\left\{\frac{1}{(k+1)^m}\sum_{r=0}^{\left[\frac{t-m}{k+1}\right]}(-1)^r\binom{m}{r}\binom{t-(k+1)r-1}{m-1}\right\}-\sum_{t=l_m^{k}}^{mk}\left\{\frac{1}{k^m}\sum_{r=0}^{\left[\frac{t-m}{k}\right]}(-1)^r\binom{m}{r}\binom{t-kr-1}{m-1}\right\} \quad (3.1.4.6).$$

Οι πίνακες (3.1.4.1), (3.1.4.2) και (3.1.4.3) παρουσιάζουν τις τιμές της διαφοράς $S_e\left(T_m^{k+1}\right)-S_e\left(T_m^{k}\right)$ για τρεις περιπτώσεις ως προς τον αριθμό των μεταβλητών (m) και των διαμερίσεων (k). Για μεγαλύτερες τιμές του m και k, οι αντίστοιχοι πίνακες δεν είναι αρκετά πρακτικό να παρουσιαστούν. Οι περιπτώσεις "$m=2$ & $k=2$", "$m=3$ & $k=2$" καθώς επίσης και "$m=2$ & $k=3$" έχουν επιλεγεί, οπότε παρουσιάζονται οι διαφορές $S_e\left(T_2^3\right)-S_e\left(T_2^2\right)$, $S_e\left(T_3^3\right)-S_e\left(T_3^2\right)$ και $S_e\left(T_2^4\right)-S_e\left(T_2^3\right)$.

Πίνακας 4.1.1.4.1: Η διαφορά των τιμών $S_e\left(T_2^3\right)-S_e\left(T_2^2\right)$ $(m=2$ και $k=2)$ για όλες τις δυνατές τιμές των ορίων l_2^2 και l_2^3 (Για τα όρια ισχύει $l_2^2, l_2^3>2$).

$l_2^2=l_2^3$	$l_2^2=3$	$l_2^3=3$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=0.1389$
	$l_2^2=4$	$l_2^3=4$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=0.4167$
$l_2^2>l_2^3$	$l_2^2=4$	$l_2^3=3$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=0.6389$
$l_2^2<l_2^3$	$l_2^2=3$	$l_2^3=4$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=-0.0833$
	$l_2^2=3$	$l_2^3=5$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=-0.4167$
	$l_2^2=3$	$l_2^3=6$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=-0.6389$
	$l_2^2=4$	$l_2^3=5$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=0.0833>0$
	$l_2^2=4$	$l_2^3=6$	$S_e\left(T_2^3\right)-S_e\left(T_2^2\right)=-0.1389$

Πίνακας 4.1.1.4.2: Η διαφορά των συναρτήσεων ευαισθησίας $S_e\left(T_3^3\right)-S_e\left(T_3^2\right)$ (m=3 & k=2) για όλες τις δυνατές τιμές των ορίων l_3^2 & l_3^3 (Ισχύει $l_3^2, l_3^3>3$).

$l_3^2=l_3^3$	$l_3^2=4$	$l_3^3=4$	$S_e\left(T_3^3\right)-S_e\left(T_3^2\right)=0.0880$
	$l_3^2=5$	$l_3^3=5$	$S_e\left(T_3^3\right)-S_e\left(T_3^2\right)=0.3519$
	$l_3^2=6$	$l_3^3=6$	$S_e\left(T_3^3\right)-S_e\left(T_3^2\right)=0.5046$

$l_3^2 > l_3^3$	$l_3^2 = 5$	$l_3^3 = 4$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = 0.4630$
	$l_3^2 = 6$	$l_3^3 = 4$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = 0.8380$
	$l_3^2 = 6$	$l_3^3 = 5$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = 0.7269$
$l_3^2 < l_3^3$	$l_3^2 = 4$	$l_3^3 = 5$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.0231$
	$l_3^2 = 4$	$l_3^3 = 6$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.2454$
	$l_3^2 = 4$	$l_3^3 = 7$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.5046$
	$l_3^2 = 4$	$l_3^3 = 8$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.7269$
	$l_3^2 = 4$	$l_3^3 = 9$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.8380$
	$l_3^2 = 5$	$l_3^3 = 6$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = 0.1296 > 0$
	$l_3^2 = 5$	$l_3^3 = 7$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.1296$
	$l_3^2 = 5$	$l_3^3 = 8$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.3519$
	$l_3^2 = 5$	$l_3^3 = 9$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.4630$
	$l_3^2 = 6$	$l_3^3 = 7$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = 0.2454 > 0$
	$l_3^2 = 6$	$l_3^3 = 8$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = 0.0231 > 0$
	$l_3^2 = 6$	$l_3^3 = 9$	$S_e\left(T_3^3\right) - S_e\left(T_3^2\right) = -0.0880$

Πίνακας 4.1.1.4.3 :Η διαφορά των συναρτήσεων $S_e\left(T_2^4\right) - S_e\left(T_2^3\right)$($m$=2 και k=3για όλες τις δυνατές τιμές των ορίων l_2^3 και l_2^4 (Για τα όρια ισχύει $l_2^3, l_2^4 > 2$).

$l_2^3 = l_2^4$	$l_2^3 = 3$	$l_2^4 = 3$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.0486$
	$l_2^3 = 4$	$l_2^4 = 4$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.1458$
	$l_2^3 = 5$	$l_2^4 = 5$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.2917$
	$l_2^3 = 6$	$l_2^4 = 6$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.2639$
$l_2^3 > l_2^4$	$l_2^3 = 4$	$l_2^4 = 3$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.2708$
	$l_2^3 = 5$	$l_2^4 = 3$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.6042$
	$l_2^3 = 6$	$l_2^4 = 3$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.8264$
	$l_2^3 = 5$	$l_2^4 = 4$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.4792$
	$l_2^3 = 6$	$l_2^4 = 4$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.7014$
	$l_2^3 = 6$	$l_2^4 = 5$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = 0.5139$
$l_2^3 < l_2^4$	$l_2^3 = 3$	$l_2^4 = 4$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = -0.0764$
	$l_2^3 = 3$	$l_2^4 = 5$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = -0.2639$
	$l_2^3 = 3$	$l_2^4 = 6$	$S_e\left(T_2^4\right) - S_e\left(T_2^3\right) = -0.5139$

	$l_2^3=3$	$l_2^4=7$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.7014$
	$l_2^3=3$	$l_2^4=8$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.8264$
	$l_2^3=4$	$l_2^4=5$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.0417$
	$l_2^3=4$	$l_2^4=6$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.2917$
	$l_2^3=4$	$l_2^4=7$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.4792$
	$l_2^3=4$	$l_2^4=8$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.6042$
	$l_2^3=5$	$l_2^4=6$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=0.0417>0$
	$l_2^3=5$	$l_2^4=7$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.1458$
	$l_2^3=5$	$l_2^4=8$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.2708$
	$l_2^3=6$	$l_2^4=7$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=0.0764>0$
	$l_2^3=6$	$l_2^4=8$	$S_e\left(T_2^4\right)-S_e\left(T_2^3\right)=-0.0486$

Στους πίνακες (3.1.4.1), (3.1.4.2) και (3.1.4.3), όπου τα όρια l_m^k και l_m^{k+1} ικανοποιούν τις σχέσεις $l_m^k=l_m^{k+1}$ και $l_m^k>l_m^{k+1}$, όλες οι τιμές των διαφορών $S_e\left(T_m^{k+1}\right)-S_e\left(T_m^k\right)$ είναι μη αρνητικές. Όταν τα όρια l_m^k και l_m^{k+1} ικανοποιούν την σχέση $l_m^k<l_m^{k+1}$, οι τιμές των διαφορών $S_e\left(T_m^{k+1}\right)-S_e\left(T_m^k\right)$ εμφανίζονται τόσο αρνητικές όσο και θετικές, αλλά κυρίως αρνητικές. Στην περίπτωση που η παραπάνω διαφορά είναι αρνητική, τότε τα όρια l_m^k και l_m^{k+1} δεν ικανοποιούν τις συνθήκες (**1.3.1)** και **(1.3.2)**, ενώ εμφανίζεται η θετική τιμή όταν τα όρια l_m^k και l_m^{k+1} ικανοποιούν τις παραπάνω συνθήκες. Αυτή η παρατήρηση είναι σε συμφωνία με τΙς συνθήκες (**1.3.1)** και **(1.3.2)** της παραγράφου 3.1.1.3. Αυτό σημαίνει, ότι στα τρία αυτά παραδείγματα, όπως και στα υπόλοιπα παραδείγματα που ελέχθησαν, στα οποία οι μεταβλητές $X_1^k, X_2^k, \ldots, X_m^k$ ακολουθούν την ομοιόμορφη κατανομή, η πρόταση 1.3 και το πόρισμα 1.3 ικανοποιούνται. Όλες οι τιμές των πινάκων (3.1.4.1), (31.4.2) και (3.1.1.3), έχουν παραχθεί με τη χρήση του μαθηματικού λογισμικού (**Matlab v2012**).

3.1.4.1 Ομοιόμορφη Κατανομή – Κοινό Διαχωριστικό Όριο

Στην παράγραφο αυτή, παρουσιάζεται η τέταρτη πρόταση της διατριβής αυτής, καθώς επίσης και η απόδειξή της, στην περίπτωση, όπου οι συνιστώσες μεταβλητές μιας διακριτής κλίμακας T_k, ακολουθούν Ομοιόμορφη κατανομή. με πεπερασμένο

πεδίο ορισμού και κοινό διαχωριστικό όριο, στην οποία αποδεικνύεται ότι η ευαισθησία είναι γνησίως αύξουσα συνάρτηση του αριθμού των διαμερίσεων των συνιστωσών μεταβλητών, χωρίς να είναι αναγκαίες οι συνθήκες **(1.3.1)** και **(1.3.2)**. Στην ειδική περίπτωση που γίνεται χρήση κοινού διαχωριστικού όριου *l*, η διαφορά της ευαισθησίας μεταξύ κλιμάκων με αριθμό διαμερίσεων που διαφέρουν κατά μια μονάδα, δίνεται από τον τύπο (3.1.4.1):

$$S_e\left(T_m^{k+1}\right)-S_e\left(T_m^{k}\right)=\sum_{t=l}^{m(k+1)}\left\{\frac{1}{(k+1)^m}\sum_{r=0}^{\left[\frac{t-m}{k+1}\right]}(-1)^r\binom{m}{r}\binom{t-(k+1)r-1}{m-1}\right\}-\sum_{t=l}^{mk}\left\{\frac{1}{k^m}\sum_{r=0}^{\left[\frac{t-m}{k}\right]}(-1)^r\binom{m}{r}\binom{t-kr-1}{m-1}\right\} \quad (3.1.4.1.1)$$

όπου η ευαισθησία των κλιμάκων T_m^k και T_m^{k+1} μπορεί να γραφεί με τη βοήθεια των συμπληρωματικών πιθανοτήτων σύμφωνα με τις σχέσεις (3.1.4.1.2) και (3.1.4.1.3):

$$S_e\left(T_m^{k}\right)=\sum_{t=l}^{m\cdot k}\left\{\frac{1}{k^m}\sum_{r=0}^{\left[\frac{t-m}{k}\right]}\left\{(-1)^r\binom{m}{r}\binom{t-k\cdot r-1}{m-1}\right\}\right\}=1-\sum_{t=m}^{l-1}\left\{\frac{1}{k^m}\sum_{r=0}^{\left[\frac{t-m}{k}\right]}\left\{(-1)^r\binom{m}{r}\binom{t-k\cdot r-1}{m-1}\right\}\right\} \quad (3.1.4.1.2)$$

και

$$S_e\left(T_m^{k+1}\right)=\sum_{t=l}^{m\cdot(k+1)}\left\{\frac{1}{(k+1)^m}\sum_{r=0}^{\left[\frac{t-m}{k+1}\right]}\left\{(-1)^r\cdot\binom{m}{r}\cdot\binom{t-(k+1)\cdot r-1}{m-1}\right\}\right\}=1-\sum_{t=m}^{l-1}\left\{\frac{1}{(k+1)^m}\sum_{r=0}^{\left[\frac{t-m}{k+1}\right]}\left\{(-1)^r\cdot\binom{m}{r}\cdot\binom{t-(k+1)\cdot r-1}{m-1}\right\}\right\} \quad (3.1.4.1.3)$$

Αντικαθιστώντας στη σχέση (3.1.4.1.1) τους τύπους (3.1.4.1.2) και (3.1.4.1.3) για την ευαισθησία των κλιμάκων T_m^k και T_m^{k+1}, η διαφορά της σχέσης (3.1.4.1.1) γίνεται σύμφωνα με τη σχέση (4.1.4.1.4):

$$S_e\left(T_m^{k+1}\right)-S_e\left(T_m^{k}\right)=\sum_{t=m}^{l-1}\left\{\frac{1}{k^m}\sum_{r=0}^{\left[\frac{t-m}{k}\right]}\left\{(-1)^r\cdot\binom{m}{r}\cdot\binom{t-k\cdot r-1}{m-1}\right\}\right\}-\sum_{t=m}^{l-1}\left\{\frac{1}{(k+1)^m}\sum_{r=0}^{\left[\frac{t-m}{k+1}\right]}\left\{(-1)^r\cdot\binom{m}{r}\cdot\binom{t-(k+1)\cdot r-1}{m-1}\right\}\right\} \quad (3.1.4.1.4)$$

όπου $\left[\frac{t-m}{k}\right]$ είναι το ακέραιο μέρος του πηλίκου $\frac{t-m}{k}$.

<u>Πρόταση 1.4</u> ($\mathbf{P_{DUm}}$). Έστω T_m^k και T_m^{k+1} κλίμακες που κατανέμονται με άγνωστο τρόπο με πεδίο ορισμού $R_{T_m^k}=\{m,m+1,\ldots,mk\}$ και $R_{T_m^{k+1}}=\{m,m+1,\ldots,mk,\ldots,m(k+1)\}$, αντιστοίχως. Η ευαισθησία των κλιμάκων T_m^k και T_m^{k+1} είναι $S_e\left(T_m^k\right)$ και $S_e\left(T_m^{k+1}\right)$, αντιστοίχως. Τότε, για κοινό διαχωριστικό όριο l με $m<l<m+k+1$ ισχύει $S_e\left(T_m^k\right)\le S_e\left(T_m^{k+1}\right)$.

<u>Απόδειξη</u>:

Αρκεί η διαφορά των ευαισθησιών των κλιμάκων T_m^k και T_m^{k+1}, να είναι γνησίως θετική, ώστε, η συνάρτηση ευαισθησίας να είναι αύξουσα. Η σχέση (3.1.4.1.4) για $l < m+k+1$ γίνεται:

$$S_e\left(T_m^{k+1}\right) - S_e\left(T_m^k\right) = \frac{1}{\left[k\cdot(k+1)\right]^m}\cdot\left[(k+1)^m\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1} - k^m\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1}\right] \quad (3.1.4.1.5)$$

όπου

$$(k+1)^m = \sum_{s=o}^{m}\binom{m}{s}\cdot k^s = \sum_{s=o}^{m-1}\binom{m}{s}\cdot k^s + \binom{m}{m}\cdot k^m = \sum_{s=o}^{m-1}\binom{m}{s}\cdot k^s + k^m \quad (3.1.4.1.6)$$

Η σχέση (3.1.4.1.4) γίνεται λόγω της (3.1.4.1.5):

$$S_e\left(T_m^{k+1}\right) - S_e\left(T_m^k\right) = \frac{1}{\left[k\cdot(k+1)\right]^m}\cdot\left[\left(\sum_{s=o}^{m-1}\binom{m}{s}\cdot k^s + k^m\right)\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1} - k^m\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1}\right] =$$

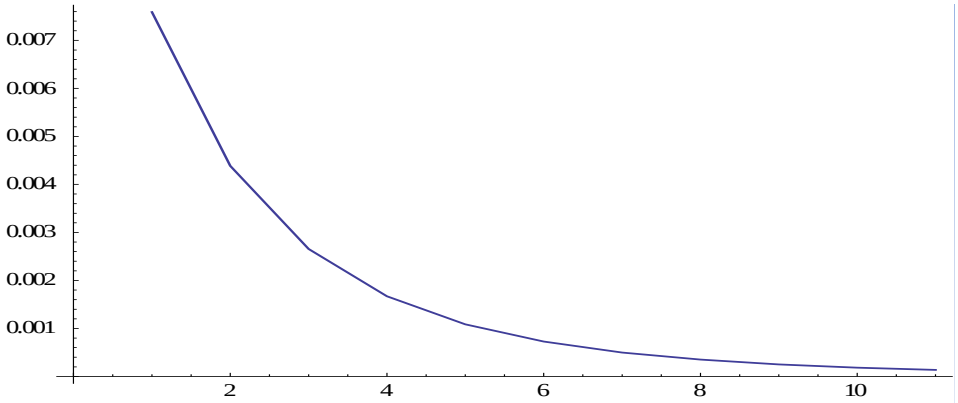
$$= \frac{1}{\left[k\cdot(k+1)\right]^m}\cdot\left[\sum_{s=o}^{m-1}\binom{m}{s}\cdot k^s\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1} + k^m\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1} - k^m\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1}\right] =$$

$$= \frac{1}{\left[k\cdot(k+1)\right]^m}\cdot\left[\sum_{s=o}^{m-1}\binom{m}{s}\cdot k^s\cdot\sum_{t=m}^{l-1}\binom{t-1}{m-1} + k^m\cdot\left(\sum_{t=m}^{l-1}\binom{t-1}{m-1} - \sum_{t=m}^{l-1}\binom{t-1}{m-1}\right)\right] =$$

$$\frac{\left[\sum_{t=m}^{l-1}\binom{t-1}{m-1}\right]\cdot\left[\sum_{s=o}^{m-1}\binom{m}{s}\cdot k^s\right]}{\left[k\cdot(k+1)\right]^m} \geq 0 \quad (3.1.4.1.7)$$

διότι $\left[k\cdot(k+1)\right]^m > 0$, $\left[\sum_{t=m}^{l-1}\binom{t-1}{m-1}\right] \geq 0$ και $\sum_{s=o}^{m-1}\binom{m}{s}\cdot k^s > 0$.

Συνεπώς ισχύει $S_e\left(T_m^{k+1}\right) > S_e\left(T_m^k\right)$, οπότε η συνάρτηση ευαισθησίας είναι γνησίως αύξουσα ως προς τον αριθμό *k* των διαμερίσεων για διαχωριστικό όριο *l* που ανήκει στο διάστημα (*m*,*m*+*k*]. Επομένως η πρόταση της εργασίας είναι αληθής.



Σχήμα 3.1.4.1.1: Διάγραμμα της διαφοράς $S_e\left(T_m^{k+1}\right) - S_e\left(T_m^k\right)$ για εύρος αριθμού

συνιστωσών $10 \le k \le 20$.

Η προηγούμενη θεωρητική απόδειξη εφαρμόστηκε σε περίπτωση κλίμακας που αποτελείται από *m=5* συνιστώσες μεταβλητές με τυχαία επιλεγμένο διαγνωστικό κατώφλι *l=15,* με χρήση τεχνητών δεδομένων που κατανέμονται ομοιόμορφα. Στο σχήμα (**3.1.4.1.1)** παρατίθεται γραφικά η διαφορά των ευαισθησιών των κλιμάκων T_m^k και T_m^{k+1}, ως προς το *k* με χρήση του λογισμικού Wolfram Mathematica. Από το σχήμα, φαίνεται ότι όταν αυξάνεται η τιμή του *k*, η συνάρτηση ευαισθησίας είναι αύξουσα συνάρτηση, καθώς το πρόσημο της διαφοράς $S_e\left(T_m^{k+1}\right) - S_e\left(T_m^k\right)$ παραμένει θετικό και ο ρυθμός μεταβολής της διαφοράς $S_e\left(T_m^{k+1}\right) - S_e\left(T_m^k\right)$ είναι μειούμενος.

3.2 Συνεχείς Κλίμακες & Εύρος Στηρίγματος

Σε αυτή την παράγραφο παρουσιάζεται η γενική περίπτωση μιας κλίμακας T_m που αποτελείται από συνεχείς μεταβλητές και συγκεκριμένα διερευνάται κατά πόσον το στήριγμα (πεδίο ορισμού) των *m* συνιστωσών μεταβλητών X_i, *i*=1, 2, ..., *m* σχετίζεται με την ακρίβεια της κλίμακας, με τη χρήση της συνάρτησης ευαισθησίας. Πιο συγκεκριμένα, σε αυτή τη διατριβή, το εύρος *l* του στηρίγματος μιας σύνθετης κλίμακας διαφοροποιείται προκειμένου να διερευνηθεί αν διαφοροποιείται ομοίως η τιμή της συνάρτησης ευαισθησίας.

Στις ακόλουθες ενότητες παρατίθενται αναλυτικά τα αποτελέσματα που αφορούν την αύξουσα μονοτονία της συνάρτησης ευαισθησίας συνεχών κλιμάκων, ως προς το εύρος *l* του πεδίου ορισμού τους, που προήλθαν από θεωρητικές αποδείξεις με χρήση στοιχείων μαθηματικής ανάλυσης (ολοκληρωτικού και διαφορικού λογισμού). Πιο συγκεκριμένα, η παράγραφος (3.2) αποτελείται από έξι ενότητες. Στην ενότητα 3.2.1 παρουσιάζεται η πρόταση περί αύξουσας μονοτονίας, καθώς και η μαθηματική απόδειξη της στην περίπτωση μίας συνεχούς μεταβλητής *X*, η οποία αντιπροσωπεύει μια κλίμακα T_1 που ακολουθεί άγνωστη κατανομή. Η ενότητα 3.2.2 παρουσιάζει την ειδική περίπτωση κατά την οποία η κλίμακα T_1 ακολουθεί τη συνεχή ομοιόμορφη κατανομή, ενώ η ενότητα 3.2.3 παρουσιάζει την ειδική περίπτωση στην οποία κλίμακα T_1 ακολουθεί την κανονική κατανομή. Στην ενότητα 3.2.4 παρουσιάζεται η πρόταση περί αύξουσας μονοτονίας, στην περίπτωση που κλίμακα T_m κατασκευάζεται από το άθροισμα των συνεχών μεταβλητών X_1 , X_2 ,

..., X_m και η κατανομή τους είναι άγνωστη. Η ενότητα 3.2.5 αναφέρεται στην εφαρμογή της προαναφερθείσας πρότασης και παρουσιάζει συγκεκριμένα παραδείγματα της συνάρτησης ευαισθησίας μιας σύνθετης κλίμακας T_m, στην περίπτωση που οι μεταβλητές X_i, i=1, 2, ..., m κατανέμονται ομοιόμορφα και επιπλέον η ενότητα 3.2.6 παρουσιάζει την ειδική περίπτωση που η κλίμακα T_m κατασκευάζεται από το άθροισμα των m κανονικά κατανεμημένων μεταβλητών X_1 , X_2 , ..., X_m.

3.2.1 Συνεχής Κλίμακα & Εύρος Στηρίγματος - Περίπτωση Άγνωστης Κατανομής Κλίμακας

Στην περίπτωση μιας απλής συνεχούς κλίμακας (X), που κατανέμεται με άγνωστο τρόπο, διερευνάται, με τη χρήση της συνάρτησης ευαισθησίας, εάν εάν το μήκος l του στηρίγματος της συνεχούς μεταβλητής X (που εκφράζει την απλή κλίμακα) επηρεάζει την προβλεπτική ικανότητα του για την πραγματική κατάσταση ενός ατόμου. Μεθοδολογικά, έγινε χρήση, δύο συνεχών τυχαίων μεταβλητών $X^{(1)}$ και $X^{(2)}$ που αναπαριστούν δύο κλίμακας T_1 και T_2 , αντιστοίχως. Οι συνεχείς μεταβλητές $X^{(1)}$ και $X^{(2)}$ λαμβάνουν τιμές από τα ακόλουθα σύνολα/πεδία ορισμού/στηρίγματα $R_{X^{(1)}} = [a, b_1]$ και $R_{X^{(2)}} = [a, b_2]$ αντίστοιχα, όπου $a, b_1, b_2 \in \mathbb{R}$ και $b_1 < b_2$.

Έστω $X^{(1)}$ και $X^{(2)}$ τυχαίες μεταβλητές με συναρτήσεις πυκνότητας πιθανότητας $f^{(1)}(x)$ και $f^{(2)}(x)$ καθώς επίσης και, με αθροιστικές συναρτήσεις κατανομής $F^{(1)}(x)$ και $F^{(2)}(x)$. Επίσης, έστω Y είναι μια τυχαία δίτιμη μεταβλητή που διακρίνει τα άτομα σε ασθενείς και υγιείς με την ένδειξη της παρουσίας ή όχι μιας ασθένειας (υγιείς Y=0, ασθενείς Y=1).

Σύμφωνα με τον ορισμό της ευαισθησίας (*Altman & Bland, 1994*), οι συναρτήσεις ευαισθησίας των μεταβλητών $X^{(1)}$ και $X^{(2)}$ σε σχέση με το Y, δίνονται από τις δεσμευμένες πιθανότητες των ακόλουθων τύπων (3.2.1.1) και (3.2.1.2) (*Bersimis et al, 2013*):

$$S_e\left(X^{(1)}\right) = P\left(X^{(1)} > c_1 \middle| Y = 1\right) \tag{3.2.1.1}$$

και

$$S_e\left(X^{(2)}\right) = P\left(X^{(2)} > c_2 \middle| Y = 1\right) \tag{3.2.1.2}$$

όπου $c_1 \in (a, b_1]$ και $c_2 \in (a, b_2]$ είναι τα διαχωριστικά όρια που αντιστοιχούν στη μέγιστη διακριτική ικανότητα μεταξύ των ατόμων, σε σχέση με την αξιολόγηση κάποιου χαρακτηριστικού (*Fawcett, 2006*). Τα διαχωριστικά όρια συμβολίζονται με μικρά γράμματα στην περίπτωση των απλών κλιμάκων. Στην περίπτωση που ένας πληθυσμός ερευνάται και περιλαμβάνει άτομα τα οποία έχουν ένα χαρακτηριστικό, τότε οι συναρτήσεις ευαισθησίας των μεταβλητών $X^{(1)}$ και $X^{(2)}$, ορίζονται σύμφωνα με τις ακόλουθες εξισώσεις (3.2.1.3) και (3.2.1.4):

$$S_e\left(X^{(1)}\right) = \int_{c_1}^{b_1} f^{(1)}(x)dx = 1 - F^{(1)}(c_1) \qquad (3.2.1.3)$$

και

$$S_e\left(X^{(2)}\right) = \int_{c_2}^{b_2} f^{(2)}(x)dx = 1 - F^{(2)}(c_2) \qquad (3.2.1.4)$$

Επομένως, κάνοντας χρήση του ορισμού ευαισθησίας, αποτυπώνονται οι μαθηματικές εκφράσεις των αντίστοιχων δεσμευμένων πιθανοτήτων απλών κλιμάκων με διαφορετικό πεδίο ορισμού, με χρήση της πιθανότητας συνεχούς άγνωστης κατανομής.

Η διαφορά $m(x)$ μεταξύ των συναρτήσεων πυκνότητας πιθανότητας των μεταβλητών $X^{(1)}$ και $X^{(2)}$ θεωρείται ως εξής:

$$m(x) = f^{(1)}(x) - f^{(2)}(x) \qquad (3.2.1.5)$$

■ **<u>Λήμμα 2.1</u> (L_{C1})**. Για τη διαφορά $m(x), x \in R_{X^{(1)}} \cup R_{X^{(2)}}$, ισχύει

$$\int_a^{b_1} m(x)dx = \int_{b_1}^{b_2} f^{(2)}(x)dx \quad (L_{C1})$$

***<u>Απόδειξη</u>*:** Ολοκληρώνοντας κατά μέλη, από το a στο b_1, στη σχέση (3.2.1.5) συνεπάγεται ότι:

$$\int_a^{b_1} m(x)dx = \int_a^{b_1} f^{(1)}(x)dx - \int_a^{b_1} f^{(2)}(x)dx = 1 - \int_a^{b_1} f^{(2)}(x)dx =$$

$$= \int_a^{b_2} f^{(2)}(x)dx - \int_a^{b_1} f^{(2)}(x)dx = \int_{b_1}^{b_2} f^{(2)}(x)dx \qquad ■$$

<u>Συνθήκη 2.1.1</u> (C_{C11})

- Για $c_1 \geq c_2 > a$ η διαφορά $m(x)$ ($x \in R_{X^{(1)}} \cup R_{X^{(2)}}$), ικανοποιεί την ακόλουθη ανισότητα:

$$\int_a^{c_1} m(x)dx \geq 0 \qquad \textbf{(C}_{\textbf{C11}}\textbf{)}$$

<u>Συνθήκη 2.1.2</u> (C_{C12})

- Για $c_2 > c_1 > a$ η διαφορά $m(x)$ ($x \in R_{X^{(1)}} \cup R_{X^{(2)}}$), ικανοποιεί την ακόλουθη ανισότητα:

$$\int_a^{c_1} m(x)dx \geq \int_{c_1}^{c_2} f^{(2)}(x)dx \qquad \textbf{(C}_{\textbf{C12}}\textbf{)}$$

Ερμηνεία Συνθήκης 2.1.1 (C$_{C11}$): Η συνθήκη (**C$_{C11}$**) δηλώνει ότι η τιμή του αριστερού μέλους, το οποίο εκφράζει την διαφορά μεταξύ των αθροιστικών συναρτήσεων κατανομής των ψευδών αρνητικών τιμών μεταξύ της κλίμακας με το στενότερο εύρος στηρίγματος και της κλίμακας με το πλατύτερο εύρος στηρίγματος θα πρέπει να είναι μεγαλύτερη από ή ίση του μηδεν.

Ερμηνεία Συνθήκης 2.1.2 (C$_{C12}$): Η συνθήκη (**C$_{C12}$**) δηλώνει ότι η τιμή του αριστερού μέλους, το οποίο εκφράζει την διαφορά μεταξύ των αθροιστικών συναρτήσεων κατανομής των ψευδών αρνητικών τιμών μεταξύ της κλίμακας με το στενότερο εύρος στηρίγματος και της κλίμακας με το πλατύτερο εύρος στηρίγματος θα πρέπει να είναι μεγαλύτερη από την τιμή του δεξιού μέλους που εκφράζει την αθροιστική συνάρτηση κατανομής των ψεύτικων αρνητικών τιμών της κλίμακας $X^{(2)}$, για τιμές μεγαλύτερες από το διαγνωστικό κατώφλι c_1.

<u>Πόρισμα 2.1.1</u>. Δεδομένου ότι το $f^{(2)}(x)$ είναι συνάρτηση της πυκνότητας πιθανότητας, η ανισότητα (3.2.1.9) ισχύει, για $c_2 < c_1$:

$$\int_{c_2}^{c_1} f^{(2)}(x)dx \geq 0 \qquad (3.2.1.6)$$

<u>Πρόταση 2.1</u> (P$_{C1}$). Έστω $T_1 = X^{(1)}$ και $T_2 = X^{(2)}$ κλίμακες που κατανέμονται με άγνωστο τρόπο με στήριγμα $R_{X^{(1)}} = [a, b_1]$ και $R_{X^{(2)}} = [a, b_2]$, αντιστοίχως. Η ευαισθησία των κλιμάκων $T_1 = X^{(1)}$ και $T_2 = X^{(2)}$ είναι $S_e(T_1) = S_e(X^{(1)})$ και $S_e(T_2) = S_e(X^{(2)})$, αντιστοίχως. Τότε, για όλες τις δυνατές τιμές των διαχωριστικών ορίων c_1 και c_2 ισχύει $S_e(X^{(1)}) \leq S_e(X^{(2)})$, υπό τις συνθήκες 2.1.1 και 2.1.2.

<u>Απόδειξη</u> Προκειμένου να αποδειχθεί η πρόταση 2.1, αρκεί να αποδειχθεί ότι για τους πραγματικούς αριθμούς $a, b_1, b_2, \in \mathrm{R}$ με $b_1 < b_2$, ισχύει η ακόλουθη ανισότητα:

$$S_e(X^{(1)}) \leq S_e(X^{(2)}) \qquad (3.2.1.7)$$

Ειδικότερα, σύμφωνα με τους ορισμούς (3.2.1.3) και (3.2.1.4), η σχέση (3.2.1.7) γίνεται:

$$\int_{c_1}^{b_1} f^{(1)}(x)\,dx \le \int_{c_2}^{b_2} f^{(2)}(x)\,dx \qquad (3.2.1.8)$$

Ως εκ τούτου, τρεις περιπτώσεις διακρίνονται όσον αφορά τη σχέση μεταξύ των κατωφλιών c_1 και c_2:

- **<u>Περίπτωση 1</u>:** $a < c_1 = c_2 = c \le b_1 < b_2$

Σύμφωνα με τις σχέσεις (3.2.1.3) και (3.2.1.4), με τη χρήση του λήμματος (2.1) και του πορίσματος (2.1), για τη συνάρτηση ευαισθησίας της μεταβλητής $X^{(1)}$ ισχύει:

$$S_e\left(X^{(1)}\right) = \int_c^{b_1} f^{(1)}(x)dx = \int_c^{b_1}\left(f^{(2)}(x) + m(x)\right)dx = \int_c^{b_1} f^{(2)}(x)dx + \int_c^{b_1} m(x)dx \le$$

$$\overset{(C_{C11})}{\le} \int_c^{b_1} f^{(2)}(x)dx + \int_c^{b_1} m(x)dx + \int_a^{c} m(x)dx = \int_c^{b_1} f^{(2)}(x)dx + \int_a^{b_1} m(x)dx =$$

$$\overset{(3.2.1.6)}{=} \int_c^{b_1} f^{(2)}(x)dx + \int_{b_1}^{b_2} f^{(2)}(x)dx = \int_c^{b_2} f^{(2)}(x)dx = S_e\left(X^{(2)}\right)$$

Συνεπώς, η ανισότητα (3.2.1.7) ισχύει.

- **<u>Περίπτωση 2</u>:** $a < c_2 < c_1 \le b_1 < b_2$

Σύμφωνα με τις σχέσεις (3.2.1.3) και (3.2.1.4) καθώς επίσης και με τη χρήση του λήμματος (2.1), των πορισμάτων (2.1) και (2.2), η συνάρτηση ευαισθησίας της μεταβλητής $X^{(1)}$ γίνεται:

$$S_e\left(X^{(1)}\right) = \int_{c_1}^{b_1} f^{(1)}(x)dx = \int_{c_1}^{b_1}\left(f^{(2)}(x) + m(x)\right)dx = \int_{c_1}^{b_1} f^{(2)}(x)dx + \int_{c_1}^{b_1} m(x)dx \le$$

$$\overset{(C_{C11})}{\le} \int_{c_1}^{b_1} f^{(2)}(x)dx + \int_{c_1}^{b_1} m(x)dx + \int_a^{c_1} m(x)dx = \int_{c_1}^{b_1} f^{(2)}(x)dx + \int_a^{b_1} m(x)dx =$$

$$= \int_{c_1}^{b_1} f^{(2)}(x)dx + \int_{b_1}^{b_2} f^{(2)}(x)dx = \int_{c_1}^{b_2} f^{(2)}(x)dx \overset{(3.2.1.6)}{\le} \int_{c_1}^{b_2} f^{(2)}(x)dx + \int_{c_2}^{c_1} f^{(2)}(x)dx =$$

$$= \int_{c_2}^{b_2} f^{(2)}(x)dx = S_e\left(X^{(2)}\right)$$

Συνεπώς, η ανισότητα (3.2.1.7) ισχύει.

- **<u>Περίπτωση 3</u>:** $a < c_1 < c_2 \le b_1 < b_2$

Σύμφωνα με τις σχέσεις (3.2.1.3) και (3.2.1.4), καθώς επίσης και με τη χρήση του λήμματος (2.1) και της συνθήκης (2.1), η συνάρτηση ευαισθησίας της μεταβλητής $X^{(1)}$ γίνεται:

$$S_e\left(X^{(1)}\right) = \int_{c_1}^{b_1} f^{(1)}(x)dx = \int_{c_1}^{b_1}\left(f^{(2)}(x) + m(x)\right)dx = \int_{c_1}^{b_1} f^{(2)}(x)dx + \int_{c_1}^{b_1} m(x)dx =$$

$$= \int_{c_1}^{c_2} f^{(2)}(x)dx + \int_{c_2}^{b_1} f^{(2)}(x)dx + \int_{c_1}^{b_1} m(x)dx \overset{(Cc12)}{\le} \int_a^{c_1} m(x)dx + \int_{c_2}^{b_1} f^{(2)}(x)dx + \int_{c_1}^{b_1} m(x)dx =$$

$$=\int_{c_2}^{b_1} f^{(2)}(x)dx+\int_{a}^{b_1} m(x)dx=\int_{c_2}^{b_1} f^{(2)}(x)dx+\int_{b_1}^{b_2} f^{(2)}(x)dx=\int_{c_2}^{b_2} f^{(2)}(x)dx=S_e\left(X^{(2)}\right)$$

Κατά συνέπεια, η ισοδυναμία (3.2.1.7) αποδείχθηκε και η πρόταση **(P_{C1})** ισχύει.

Σε αυτό το σημείο πρέπει να σημειωθεί ότι η πρόταση (2.1) ισχύει στην περίπτωση των δεικτών που σχετίζονται με την υγεία, οι οποίοι λαμβάνουν τιμές σε φραγμένα πεδία ορισμού.

<u>Πόρισμα 2.1.2</u> (C_{C21}): Η ευαισθησία της συνεχούς κλίμακας $T_1 \in [a,b]$ είναι μια αύξουσα (μη φθίνουσα) συνάρτηση του μήκους l του στηρίγματος της, υπό ορισμένες συνθήκες (C_{C11} - C_{C12}), στην περίπτωση που ακολουθεί μια άγνωστη κατανομή.

3.2.2 Συνεχής Κλίμακα & Εύρος Στηρίγματος - Περίπτωση Ομοιόμορφης Κατανομής Κλίμακας

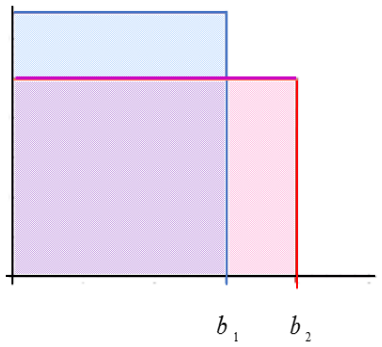
Στην παράγραφο αυτή παρατίθεται η θεωρητική απόδειξη της πρότασης (Πρόταση 2.2 - (P_{CU1})) που εκφράζει την αύξουσα μονοτονία μιας συνεχούς κλίμακας ως προς το εύρος του στηρίγματος της, στην περίπτωση που ακολουθεί την Ομοιόμορφη κατανομή (Uniform Distribution). Δηλαδή, παρουσιάζεται η ειδική περίπτωση μιας κλίμακας T_1 η οποία ακολουθεί την ομοιόμορφη κατανομή. Ομοίως με την προηγούμενη παράγραφο, παρατίθεται η μεθοδολογία διερεύνησης του κατά πόσον το πεπερασμένο πεδίο ορισμού (στήριγμα) της μεταβλητής T_1 επηρεάζει την προβλεπτική ικανότητα της. Για τον σκοπό αυτό, έστω δύο συνεχείς τυχαίες μεταβλητές $X^{(1)}=T_1^1$ και $X^{(2)}=T_1^2$ που ακολουθούν την ομοιόμορφη κατανομή με στηρίγματα $[a,b_1]$ και $[a,b_2]$ αντίστοιχα, με $b_1<b_2$. Οι συναρτήσεις πυκνότητας πιθανότητας $f^{(1)}(x)$ και $f^{(2)}(x)$ των μεταβλητών $X^{(1)}$ και $X^{(2)}$, αντιστοίχως, δίνονται από τους ακόλουθους τύπους (3.2.2.1) και (3.2.2.2) (*Frank & Collani, 2001*):

$$f^{(1)}(x)=\begin{cases}\dfrac{1}{b_1-a}, & x\in[a,b_1]\\ 0, & x\notin[a,b_1]\end{cases} \qquad (3.2.2.1)$$

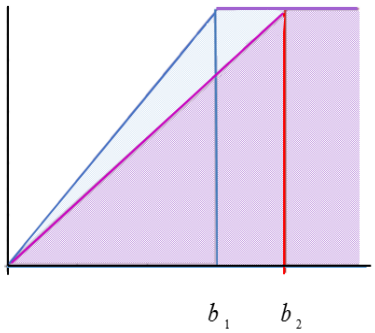
και

$$f^{(2)}(x)=\begin{cases}\dfrac{1}{b_2-a}, & x\in[a,b_2]\\ 0, & x\notin[a,b_2]\end{cases} \qquad (3.2.2.2)$$

Στα σχήματα (3.2.2.1) και (3.2.2.2), οι συναρτήσεις πυκνότητας πιθανότητας $f^{(1)}(x)$ και $f^{(2)}(x)$ και αθροιστικές συναρτήσεις κατανομής $F^{(1)}(x)$ και $F^{(2)}(x)$ των μεταβλητών $X^{(1)}$ και $X^{(2)}$ παρουσιάζονται γραφικά, υπό τον περιορισμό $b_1 < b_2$.



Σχήμα 3.2.2.1: Συνάρτηση πυκνότητας πιθανότητας των μεταβλητών $X^{(1)}$ και $X^{(2)}$



Σχήμα 3.2.2.2: Αθροιστική συνάρτηση πιθανότητας των μεταβλητών $X^{(1)}$ και $X^{(2)}$

Επιπλέον, οι συναρτήσεις της αθροιστικής κατανομής $F^{(1)}(x)$ και $F^{(2)}(x)$ των μεταβλητών $X^{(1)}$ και $X^{(2)}$ δίνονται από τους τύπους (3.2.2.3) και (3.2.2.4) (*Weisstein, 2003*):

$$F^{(1)}(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b_1 - a}, & a \le x \le b_1 \\ 1, & x > b_1 \end{cases} \tag{3.2.2.3}$$

και

$$F^{(2)}(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b_2 - a}, & a \le x \le b_2 \\ 1, & x > b_2 \end{cases} \tag{3.2.2.4}$$

Σύμφωνα με τον ορισμό της συνάρτησης ευαισθησίας, οι συναρτήσεις ευαισθησίας των μεταβλητών $X^{(1)}$ και $X^{(2)}$ δίνονται από τους ακόλουθους τύπους (3.2.2.5) και (3.2.2.6):

$$S_e\left(X^{(1)}\right) = \int_{c_1}^{b_1} \frac{1}{b_1 - a} dx = \frac{b_1 - c_1}{b_1 - a} = \frac{b_1 - c_1}{l_1} \tag{3.2.2.5}$$

και

$$S_e\left(X^{(2)}\right) = \int_{c_2}^{b_2} \frac{1}{b_2 - a} dx = \frac{b_2 - c_2}{b_2 - a} = \frac{b_2 - c_2}{l_2} \tag{3.2.2.6}$$

όπου $c_1 \in (a, b_1]$ και $c_2 \in (a, b_2]$ είναι τα διαχωριστικά όρια.

Επομένως, κάνοντας χρήση του ορισμού ευαισθησίας, αποτυπώνονται οι μαθηματικές εκφράσεις των αντίστοιχων δεσμευμένων πιθανοτήτων απλών κλιμάκων με διαφορετικό πεδίο ορισμού, με χρήση της κατανομής πιθανότητας συνεχούς ομοιόμορφης κατανομής.

■ **<u>Λήμμα 2.2.</u> (L $_{CU1}$)** Για τη διαφορά $m(x)$, $x \in [a, b_2]$ ισχύει:

$$\int_a^{b_1} m(x)\,dx = \frac{b_2 - b_1}{b_2 - a} \quad (L_{CU1})$$

<u>Απόδειξη</u>: Ολοκληρώνοντας κατά μέλη, από *α* έως b_1, έχουμε ότι:

$$\int_a^{b_1} m(x)\,dx = \int_a^{b_1} \frac{b_2 - b_1}{(b_2 - a)\cdot(b_1 - a)}\,dx = \frac{(b_1 - a)\cdot(b_2 - b_1)}{(b_2 - a)\cdot(b_1 - a)} = \frac{b_2 - b_1}{b_2 - a} \qquad ■$$

<u>Σημείωση:</u> Στην περίπτωση της ομοιόμορφης κατανομής των κλιμάκων, δεν είναι απαραίτητη η Συνθήκη 2.1.1 (C_{C11}), καθώς η ανισότητα $\int_a^{c_1} m(x)dx \geq 0$ ισχύει για $c \geq c' > 1$.

<u>Συνθήκη 2.2.1</u> (C$_{CU1}$)

- Για $c_1 < c_2$, η διαφορά $m(x)$ ($x \in R_{X^{(1)}} \cup R_{X^{(2)}}$), ικανοποιεί την ακόλουθη ανισότητα:

$$(b_2 - b_1)\cdot(c_1 - a) \geq (c_2 - c_1)\cdot(b_1 - a) \qquad \textbf{(C}_{\textbf{CU1}}\textbf{)}$$

Η συνθήκη **(C$_{CU1}$)** αποτελεί μια εναλλακτική έκφραση της συνθήκης **(C$_{C12}$)** στην περίπτωση που η μεταβλητή *X* κατανέμεται ομοιόμορφα.

<u>Πόρισμα 2.2.1</u> Η συνθήκη (**C$_{C12}$**) είναι μια ειδική έκφραση της συνθήκης (**C$_{C1}$**) στην περίπτωση που οι μεταβλητές $X^{(1)}$ και $X^{(2)}$ κατανέμονται ομοιόμορφα.

<u>Απόδειξη</u>: Η συνθήκη (**C$_{C12}$**) γίνεται:

$$(c_1 - a)\cdot(b_2 - b_1) \geq (b_1 - a)\cdot(c_2 - c_1) \qquad (3.2.2.7)$$

<u>Πρόταση 2.2</u> (P$_{CU1}$). Έστω $T_1 = X^{(1)}$ και $T_2 = X^{(2)}$ δείκτες που κατανέμονται ομοιόμορφα με στήριγμα $R_{X^{(1)}} = [a, b_1]$ και $R_{X^{(2)}} = [a, b_2]$, αντιστοίχως. Η ευαισθησία των κλιμάκων $T_1 = X^{(1)}$ και $T_2 = X^{(2)}$ είναι $S_e(T_1) = S_e(X^{(1)})$ και $S_e(T_2) = S_e(X^{(2)})$, αντιστοίχως. Τότε, για όλες τις δυνατές τιμές των διαχωριστικών ορίων c_1 και c_2 ισχύει $S_e(X^{(1)}) \leq S_e(X^{(2)})$, υπό τη συνθήκη 2.2.1.

Απόδειξη: Για να αποδειχθεί η πρόταση (2.2) αρκεί να δειχθεί ότι για τους πραγματικούς αριθμούς $a, b_1, b_2 \in \mathrm{R}$ με $b_1 < b_2$, ισχύει η ανισότητα (3.2.2.8). Σύμφωνα με τις σχέσεις (3.2.2.5) και (3.2.2.6), ισχύει:

$$\frac{b_1 - c_1}{b_1 - a} \leq \frac{b_2 - c_2}{b_2 - a} \tag{3.2.2.8}$$

Η διαφορά $S_e\left(X^{(2)}\right) - S_e\left(X^{(1)}\right)$ θεωρείται ως ακολούθως:

$$S_e\left(X^{(2)}\right) - S_e\left(X^{(1)}\right) = \frac{b_2 - c_2}{b_2 - a} - \frac{b_1 - c_1}{b_1 - a} = \frac{(b_1 - a)\cdot(b_2 - c_2) - (b_2 - a)\cdot(b_1 - c_1)}{(b_1 - a)\cdot(b_2 - a)} =$$

$$= \frac{b_2 \cdot (c_1 - a) + b_1 \cdot (a - c_2) + a \cdot (c_2 - c_1)}{(b_1 - a)\cdot(b_2 - a)} = \frac{c_1 \cdot (b_2 - a) - c_2 \cdot (b_1 - a) - a \cdot (b_2 - b_1)}{(b_1 - a)\cdot(b_2 - a)}$$

Επομένως ισχύει:

$$S_e\left(X^{(2)}\right) - S_e\left(X^{(1)}\right) = \frac{c_1 \cdot (b_2 - a) - c_2 \cdot (b_1 - a) - a \cdot (b_2 - b_1)}{(b_1 - a)\cdot(b_2 - a)} \tag{3.2.2.9}$$

Η διαφορά $m(x) = f^{(1)}(x) - f^{(2)}(x)$ στην περίπτωση που οι μεταβλητές $X^{(1)}$ και $X^{(2)}$ κατανέμονται ομοιόμορφα, λαμβάνει την ακόλουθη μορφή (3.2.2.10):

$$m(x) = \frac{b_2 - b_1}{(b_1 - a)\cdot(b_2 - a)} \tag{3.2.2.10}$$

Για να αποδειχθεί η πρόταση 2.2, τρεις περιπτώσεις διακρίνονται αναφορικά με την σχέση θέση των c_1 και c_2:

- **<u>Περίπτωση 1</u>**: $a < c_1 = c_2 = c \leq b_1 < b_2$

Σύμφωνα με τη σχέση (3.2.2.9) η διαφορά $S_e\left(X^{(2)}\right) - S_e\left(X^{(1)}\right)$ γίνεται:

$$S_e\left(X^{(2)}\right) - S_e\left(X^{(1)}\right) = \frac{b_2 \cdot (c - a) + b_1 \cdot (a - c)}{(b_1 - a)\cdot(b_2 - a)} = \frac{(b_2 - b_1)\cdot(c - a)}{(b_1 - a)\cdot(b_2 - a)} > 0$$

Κατά συνέπεια, η ανισότητα (3.2.2.8) ισχύει

- **<u>Περίπτωση 2</u>**: $a < c_2 < c_1 \leq b_1 < b_2$.

Σύμφωνα με τη σχέση (3.2.2.9), η διαφορά $S_e\left(X^{(2)}\right) - S_e\left(X^{(1)}\right)$ γίνεται:

$$S_e\left(X^{(2)}\right) - S_e\left(X^{(1)}\right) = \frac{c_1 \cdot (b_2 - a) - c_2 \cdot (b_1 - a) - a \cdot (b_2 - b_1)}{(b_1 - a)\cdot(b_2 - a)} >$$

$$>\frac{c_2\cdot(b_2-a)-c_2\cdot(b_1-a)-a\cdot(b_2-b_1)}{(b_1-a)\cdot(b_2-a)}=\frac{c_2\cdot(b_2-a-b_1+a)-a\cdot(b_2-b_1)}{(b_1-a)\cdot(b_2-a)}=$$

$$=\frac{c_2\cdot(b_2-b_1)-a\cdot(b_2-b_1)}{(b_1-a)\cdot(b_2-a)}=\frac{(c_2-a)\cdot(b_2-b_1)}{(b_1-a)\cdot(b_2-a)}>0$$

Κατά συνέπεια, η ανισότητα (3.2.2.8) αποδείχθηκε.

- **<u>Περίπτωση 3</u>**: $a<c_1<c_2\le b_1<b_2$.

Σύμφωνα με τη σχέση (3.2.2.9) και τη συνθήκη **(C_{CU1})**, η διαφορά $S_e\left(X^{(2)}\right)-S_e\left(X^{(1)}\right)$ γίνεται:

$$S_e\left(X^{(2)}\right)-S_e\left(X^{(1)}\right)=\frac{b_2\cdot(c_1-a)+b_1\cdot(a-c_2)+a\cdot(c_2-c_1)}{(b_1-a)\cdot(b_2-a)}=\frac{b_2\cdot(c_1-a)+b_1\cdot a-b_1\cdot c_2+a\cdot(c_2-c_1)}{(b_1-a)\cdot(b_2-a)}=$$

$$=\frac{b_2\cdot(c_1-a)+b_1\cdot a-b_1\cdot c_1+b_1\cdot c_1-b_1\cdot c_2+a\cdot(c_2-c_1)}{(b_1-a)\cdot(b_2-a)}=\frac{b_2\cdot(c_1-a)-b_1\cdot(c_1-a)-b_1\cdot(c_2-c_1)+a\cdot(c_2-c_1)}{(b_1-a)\cdot(b_2-a)}=$$

$$=\frac{(b_2-b_1)\cdot(c_1-a)+(a-b_1)\cdot(c_2-c_1)}{(b_1-a)\cdot(b_2-a)}\ge\frac{(c_2-c_1)\cdot(b_1-a)-(b_1-a)\cdot(c_2-c_1)}{(b_1-a)\cdot(b_2-a)}=0$$

Κατά συνέπεια, η ανισότητα (3.2.2.8) αποδείχθηκε και η πρόταση (P_{CU1}) ισχύει.

<u>Πόρισμα 2.2.2</u> (C_{CU1}): Η ευαισθησία της συνεχούς κλίμακας $T_1\in[a,b]$ είναι μια αύξουσα (μη φθίνουσα) συνάρτηση του μήκους l του στηρίγματος της, υπό ορισμένη συνθήκη **(C_{C11} - C_{C12})**, στην περίπτωση που ακολουθεί την ομοιόμορφη κατανομή.

3.2.3 Συνεχής Κλίμακα & Εύρος Στηρίγματος - Περίπτωση Κανονικής κατανομής Κλίμακας

Η παράγραφος αυτή παρουσιάζει την ειδική περίπτωση μιας κλίμακας T_1 που ακολουθεί την κανονική κατανομή και διερευνάται εάν οι παράμετροι της κανονικής κατανομής μ και σ επηρεάζουν την ακρίβεια της. Η περίπτωση μιας κλίμακας T_1 που κατανέμεται κανονικά είναι αρκετά διαφορετική από την περίπτωση της προηγούμενης παραγράφου στην οποία η κλίμακα T_1 ακολουθεί την ομοιόμορφη κατανομή. Η διαφορά μεταξύ των δύο περιπτώσεων εντοπίζεται στο μήκος του πεδίου ορισμού (στηρίγματος) τους. Όταν μια μεταβλητή X ακολουθεί την ομοιόμορφη κατανομή τότε το πεδίο ορισμού της περιορίζεται (π.χ. $x\in[a,b]$), ως εκ τούτου, το μήκος l της στηρίγματος είναι ένας πραγματικός πεπερασμένος αριθμός, αλλά εάν μια μεταβλητή X ακολουθεί την κανονική κατανομή, το πεδίο ορισμού της

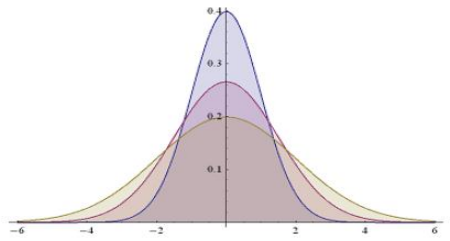
είναι μη πεπερασμένο, επομένως το μήκος του στηρίγματος του τείνει στο άπειρο αν και η πλειοψηφία των παρατηρούμενων τιμών είναι εντός του συνόλου [*μ-3σ, μ+3σ*]. Ως εκ τούτου, δύο συνεχείς τυχαίες μεταβλητές $X^{(1)} = T_1^{(1)} \sim N\left(\mu_1, \sigma_1^2\right)$ και $X^{(2)} = T_1^{(2)} \sim N\left(\mu_2, \sigma_2^2\right)$ θεωρούνται με τις συναρτήσεις κατανομής πιθανότητας $f^{(1)}(x)$ και $f^{(2)}(x)$, που δίνονται από τους ακόλουθους τύπους (3.2.3.1) και (3.2.3.2) (*Feller, 1966*):

$$f^{(1)}(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_1} \cdot \exp\left(\frac{-\left(x-\mu_1\right)^2}{2\sigma_1^2}\right) \tag{3.2.3.1}$$

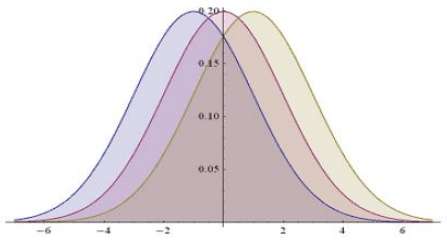
και

$$f^{(2)}(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma_2} \cdot \exp\left(\frac{-\left(x-\mu_2\right)^2}{2\sigma_2^2}\right) \tag{3.2.3.2}$$

Στα σχήματα (3.2.3.1) και (3.2.3.2), παρουσιάζονται οι συναρτήσεις πυκνότητας πιθανότητας της μεταβλητής $X \sim N\left(\mu, \sigma^2\right)$ για διάφορες τιμές των παραμέτρων *μ* και *σ*. Τα σχήματα παρουσιάζουν τη γνωστή συμμετρική και σε σχήμα καμπάνας καμπύλη της κανονικής κατανομής.



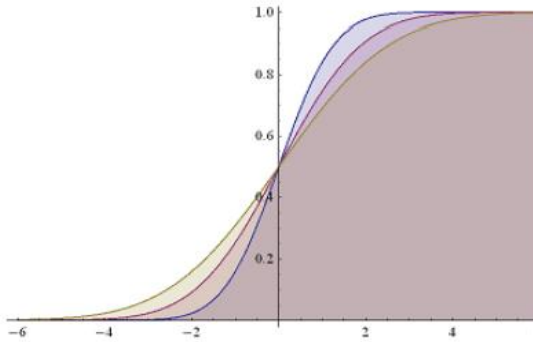
Σχήμα 3.2.3.1:Κανονική Συνάρτηση Πυκνότητας της μεταβλητής X για *σ*=1, 1.5 & 2



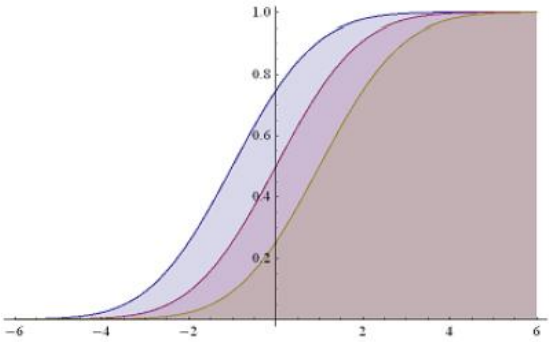
Σχήμα 3.2.3.2:Κανονική Συνάρτηση Πυκνότητας της μεταβλητής X για *μ*=-1, 0 & 1

Πιο συγκεκριμένα, στο σχήμα (3.2.3.1) παρουσιάζονται τρείς καμπύλες για σταθερή μέση τιμή *μ*=0 και για τρεις διαφορετικές αριθμητικές τιμές της τυπικής απόκλισης, *σ* =1, 1.5 και 2. Στο σχήμα (3.2.3.2) παρουσιάζονται τρεις καμπύλες για τη σταθερή τυπική απόκλιση *σ*=2 και τρεις διαφορετικές αριθμητικές τιμές για την μέση τιμή, *μ*=-1, 0 και 1. Στα σχήματα (3.2.3.3) και (3.2.3.4) παρουσιάζονται οι αθροιστικές συναρτήσεις κατανομής της μεταβλητής $X \sim N\left(\mu, \sigma^2\right)$ για τις διαφορετικές τιμές των παραμέτρων *μ* και *σ*. Πιο συγκεκριμένα, στο σχήμα (3.2.3.3) παρουσιάζονται τρεις καμπύλες για στα σταθερή μέση τιμή *μ*=0 και τρεις

διαφορετικές τιμές για την τυπική απόκλιση, σ=1, 1.5 και 2, καθώς επίσης, στο σχήμα (3.2.3.4) παρουσιάζονται τρεις καμπύλες για σταθερή τυπική απόκλιση σ=1 και τρεις διαφορετικές αριθμητικές τιμές για μέση τιμή μ =-1, 0 και 1.



Σχήμα 3.2.3.3: Κανονική Αθροιστική Συνάρτηση της μεταβλητής X για σ=1, 1.5 & 2



Σχήμα 3.2.3.4: Κανονική Αθροιστική Συνάρτηση τη μεταβλητής X για μ=-1, 0 & 1

Ως εκ τούτου, με την παραδοχή του πληθυσμού ατόμων που διαθέτουν ένα συγκεκριμένο χαρακτηριστικό, σύμφωνα με τον ορισμό της συνάρτησης ευαισθησίας, οι συναρτήσεις ευαισθησίας, των μεταβλητών $X^{(1)}$ και $X^{(2)}$, δίνονται από τις ακόλουθες σχέσεις (3.2.3.3) και (3.2.3.4) (*Feller, 1966*):

$$S_e\left(X^{(1)}\right)=\frac{1}{\sqrt{2\pi}\cdot\sigma_1}\cdot\int_{c_1}^{\infty}\exp\left(\frac{-\left(x-\mu_1\right)^2}{2\cdot\sigma_1^2}\right)dx \qquad (3.2.3.3)$$

και

$$S_e\left(X^{(2)}\right)=\frac{1}{\sqrt{2\pi}\cdot\sigma_2}\cdot\int_{c_2}^{\infty}\exp\left(\frac{-\left(x-\mu_2\right)^2}{2\cdot\sigma_2^2}\right)dx \qquad (3.2.3.4)$$

Επιπροσθέτως, χρησιμοποιώντας την αθροιστική συνάρτηση κατανομής, οι συναρτήσεις ευαισθησίας για τις μεταβλητές $X^{(1)}$ και $X^{(2)}$, δίνονται από τους τύπους (3.2.3.5) και (3.2.3.6) (*Feller, 1966*):

$$S_e\left(X^{(1)}\right)=P\left(\frac{X^{(1)}-\mu_1}{\sigma_1}>\frac{c_1-\mu_1}{\sigma_1}\right)=1-\Phi\left(\frac{c_1-\mu_1}{\sigma_1}\right) \qquad (3.2.3.5)$$

και

$$S_e\left(X^{(2)}\right)=P\left(\frac{X^{(2)}-\mu_2}{\sigma_2}>\frac{c_2-\mu_2}{\sigma_2}\right)=1-\Phi\left(\frac{c_2-\mu_2}{\sigma_2}\right) \qquad (3.2.3.6)$$

όπου $c_1, c_2 \in \mathrm{R}$ το διαχωριστικό όριο και Φ η γνωστή αθροιστική συνάρτηση της κανονικής κατανομής. Ορίζουμε τους πίνακες ***A*** και ***B*** ως ακολούθως:

$$A=\begin{bmatrix} c_1 & c_2 \\ \sigma_1 & \sigma_2 \end{bmatrix} \text{ και } B=\begin{bmatrix} \mu_1 & \mu_2 \\ \sigma_1 & \sigma_2 \end{bmatrix}$$

Ο πίνακας ***B*** περιέχει ανά στήλη τις παραμέτρους των κατανομών των κλιμάκων $X^{(1)}$ και $X^{(2)}$ και ο πίνακας ***A*** περιέχει τις τυπικές αποκλίσεις των κλιμάκων, καθώς επίσης και τα διαχωριστικά όρια. Οι ορίζουσες των πινάκων ***A*** και ***B***, δίνονται από τις ακόλουθες σχέσεις:

$$\det(A)=\begin{vmatrix} c_1 & c_2 \\ \sigma_1 & \sigma_2 \end{vmatrix} \text{ και } \det(B)=\begin{vmatrix} \mu_1 & \mu_2 \\ \sigma_1 & \sigma_2 \end{vmatrix}$$

Επομένως, κάνοντας χρήση του ορισμού ευαισθησίας, αποτυπώνονται οι μαθηματικές εκφράσεις των αντίστοιχων δεσμευμένων πιθανοτήτων απλών κλιμάκων που ακολουθούν κανονική κατανομή.

<u>Συνθήκη 2.3</u> (C_{CN1}). Για τους ορίζουσες των πινάκων $\det(A)$ και $\det(B)$ ικανοποιείται η ακόλουθη ανισότητα (3.2.3.7):

$$\det(A)\geq\det(B) \tag{3.2.3.7}$$

<u>Πρόταση 2.3</u> (P_{CN1}). Στην περίπτωση που η κλίμακα T_1 κατανέμεται κανονικά, η συνάρτηση ευαισθησίας της κλίμακας T_1 είναι μια αύξουσα (μη φθίνουσα) συνάρτηση των παραμέτρων μ και σ, υπό ορισμένη συνθήκη (3.2.3.8).

<u>Απόδειξη</u>: Σύμφωνα με τις σχέσεις (3.2.3.5) και (3.2.3.6), η ανισότητα $S_e\left(X^{(2)}\right)\geq S_e\left(X^{(1)}\right)$ γίνεται:

$$\Phi\left(\frac{c_1-\mu_1}{\sigma_1}\right)\geq\Phi\left(\frac{c_2-\mu_2}{\sigma_2}\right) \tag{3.2.3.8}$$

Είναι γνωστό ότι η συνάρτηση Φ είναι μια αύξουσα συνάρτηση, επομένως η ανισότητα (3.2.3.8) γίνεται με χρήση:

$$\frac{c_1-\mu_1}{\sigma_1}\geq\frac{c_2-\mu_2}{\sigma_2}$$

με χρήση των οριζουσών των πινάκων A και B

$$\det(A)\geq\det(B) \tag{3.2.3.9}$$

Επομένως η ανισότητα (3.2.3.8), αποδείχθηκε υπό την συνθήκη (3.2.3.7).

Στη συνέχεια παρατίθενται ειδικές περιπτώσεις όσον αφορά στις αριθμητικές τιμές των παραμέτρων της κατανομής, καθώς επίσης και την τιμή των διαχωριστικών ορίων.

- **<u>Ειδική περίπτωση 3.2.3.1</u>**: $c_1 = c_2 = c$

Σύμφωνα με τη σχέση (3.2.3.8)), η ανισότητα $S_e\left(X^{(2)}\right) \geq S_e\left(X^{(1)}\right)$ γίνεται

$$\sigma_2 \cdot c - \sigma_1 \cdot c \geq \sigma_2 \cdot \mu_1 - \sigma_1 \cdot \mu_2$$

οπότε λαμβάνουμε την παρακάτω ανισότητα

$$\left(\sigma_2 - \sigma_1\right) \cdot c \geq \det\left(B\right) \qquad (3.2.3.10)$$

Επομένως, η συνάρτηση ευαισθησίας της κλίμακας είναι αύξουσα (μη φθίνουσα) ως προς το διαγνωστικό κατώφλι c, όταν ισχύει:

- $c \geq \dfrac{\det\left(B\right)}{\sigma_2 - \sigma_1}$ εάν $\sigma_2 > \sigma_1$
- $c \leq \dfrac{\det\left(B\right)}{\sigma_2 - \sigma_1}$ εάν $\sigma_2 < \sigma_1$,

τότε η ανισότητα (3.2.3.11) ισχύει:

$$\mu_2 \geq \mu_1 \qquad (3.2.3.11)$$

<u>Πρόταση 2.3.1</u> ($P_{CN1.1}$). Στην περίπτωση που η κλίμακα T_1 κατανέμεται κανονικά, η ευαισθησία της είναι μια αύξουσα (μη φθίνουσα) συνάρτηση της αναμενόμενης τιμής μ, όταν η τυπική απόκλιση σ είναι σταθερή.

- **<u>Ειδική περίπτωση 3.2.3.2</u>**: $\mu_1 = \mu_2 = \mu$.

Σύμφωνα με τη σχέση (3.2.3.8), η ανισότητα $S_e\left(X^{(2)}\right) \geq S_e\left(X^{(1)}\right)$ γίνεται:

$$\sigma_2 \cdot c_1 - \sigma_1 \cdot c_2 \geq \sigma_2 \cdot \mu - \sigma_1 \cdot \mu$$

Επομένως προκύπτει η ανισότητα:

$$\det\left(A\right) \geq \left(\sigma_2 - \sigma_1\right) \cdot \mu \qquad (3.2.3.12)$$

Επομένως, η συνάρτηση ευαισθησίας της κλίμακας είναι μη φθίνουσα ως προς την αναμενόμενη τιμή μ, όταν ισχύει:

- $\mu \leq \dfrac{\det\left(A\right)}{\sigma_2 - \sigma_1}$ εάν $\sigma_2 > \sigma_1$
- $\mu \geq \dfrac{\det\left(A\right)}{\sigma_2 - \sigma_1}$ εάν $\sigma_2 < \sigma_1$

Εάν $\sigma_2 = \sigma_1$ τότε η παρακάτω ανισότητα ισχύει:

$$c_1 \geq c_2 \qquad (3.2.3.13)$$

Είναι γνωστό ότι μείωση της τιμής του διαχωριστικού ορίου *c* οδηγεί σε αύξηση της ευαισθησίας της κλίμακας. Σε αυτήν την περίπτωση η πρόταση αυτή αποδείχθηκε όταν η κλίμακα κατανέμεται κανονικά.

- **<u>Ειδική περίπτωση 3.2.3.3</u>**: $\sigma_1 = \sigma_2 = \sigma$.

Σύμφωνα με τη σχέση (3.2.3.8), η ανισότητα $S_e\left(X^{(2)}\right) \geq S_e\left(X^{(1)}\right)$ γίνεται:

$$\sigma \cdot c_1 - \sigma \cdot c_2 \geq \sigma \cdot \mu_1 - \sigma \cdot \mu_2$$

Επομένως καταλήγουμε στην ανισότητα (3.2.3.14)

$$d(\mu) \geq d(c) \qquad (3.2.3.14)$$

όπου $d(\mu)$ είναι η μεταβολή της αναμενόμενης τιμής μ και $d(c)$ είναι η μεταβολή του διαχωριστικού ορίου.

<u>Πρόταση 2.3.2</u> ($P_{CN1.2}$). Στην περίπτωση που η κλίμακα T_1 κατανέμεται κανονικά, η ευαισθησία της είναι μια αύξουσα (μη φθίνουσα) συνάρτηση αν η αύξηση της αναμενόμενης τιμής είναι μεγαλύτερη από την αύξηση των διαχωριστικών ορίων, όταν η τυπική απόκλιση σ είναι σταθερή.

Η υπόθεση ότι η κλίμακα κατανέμεται κανονικά έχει πλεονεκτήματα αλλά και μειονεκτήματα. Ένα κρίσιμο πλεονέκτημα είναι ότι η χρήση της κανονικής προσεγγίσεως προσφέρει πολλές δυνατότητες που έχουν ερευνηθεί επί σειράς ετών στη βιβλιογραφία και βασίζεται στις μοναδικές ιδιότητες της κανονικής κατανομής (*Feller, 1966*). Ένα σημαντικό μειονέκτημα είναι ότι το στήριγμα της κλίμακας δεν μπορεί να θεωρηθεί ως πεπερασμένο, παρά το γεγονός ότι ένα διάστημα της μορφής (μ-$k\sigma$,μ+$k\sigma$) με k>3, εμπεριέχεται σχεδόν το σύνολο των αντίστοιχων παρατηρήσεων.

3.2.4 Σύνθετη Συνεχής Κλίμακα & Εύρος Στηρίγματος Συνιστωσών - Περίπτωση άγνωστης κατανομής

Η παράγραφος αυτή παρουσιάζει τη γενική περίπτωση μιας σύνθετης κλίμακας T_m που κατασκευάζεται από το άθροισμα m ανεξάρτητων και ταυτόσημα κατανεμημένων μεταβλητών X_i, $i = 1, 2, ..., m$, των οποίων η κατανομή είναι άγνωστη. Πιο συγκεκριμένα, ερευνάται εάν το πεδίο ορισμού (στήριγμα) των μεταβλητών X_i, $i = 1, 2, ..., m$, έχει επιπτώσεις στην ακρίβεια της κλίμακας. Δύο συνεχείς κλίμακες $T_m^{(1)}$ και $T_m^{(2)}$ θεωρούνται ως εξής:

$$T_m^{(1)} = \sum_{i=1}^{m} X_i^{(1)} \quad \text{και} \quad T_m^{(2)} = \sum_{i=1}^{m} X_i^{(2)}$$

όπου $X_i^{(1)} \in [a, b_1]$ και $X_i^{(2)} \in [a, b_2]$, $i = 1, 2, ..., m$ αντίστοιχα με $a, b_1, b_2 \in \mathrm{R}$, $m \in N$ και $b_1 < b_2$. Ως εκ τούτου, το πεδίο ορισμού (στήριγμα) των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$ είναι $R_{T_m^{(1)}} = [m \cdot a, m \cdot b_1]$ και $R_{T_m^{(2)}} = [m \cdot a, m \cdot b_2]$ αντιστοίχως.

Ως μήκος του στηρίγματος $R_{T_m^{(j)}} = [m \cdot a, m \cdot b_j]$ της κλίμακας $T_m^{(j)}$, $j = 1, 2$ θεωρείται ο θετικός αριθμός L_j, $j = 1, 2$ ο οποίος ορίζεται από την ακόλουθη ισότητα: $L_j = m \cdot (b_j - a) = m \cdot l_j$, $j = 1, 2$. Επομένως, αν ισχύει $b_1 < b_2$ τότε ισχύει $L_1 < L_2$.

Έστω $T_m^{(1)}$ και $T_m^{(2)}$ με τις συναρτήσεις πυκνότητας πιθανότητας $f_m^{(1)}(t)$ και $f_m^{(2)}(t)$, καθώς επίσης, με τις αθροιστικές συναρτήσεις κατανομής $F_m^{(1)}(t)$ και $F_m^{(2)}(t)$. Σύμφωνα με τον ορισμό της ευαισθησίας οι συναρτήσεις ευαισθησίας των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$, σε σχέση με το *Y*, δίνονται από τις δεσμευμένες πιθανότητες των ακόλουθων τύπων (3.2.4.1) και (3.2.4.2):

$$S_e\left(T_m^{(1)}\right) = P\left(T_m^{(1)} > C_1 \mid Y = 1\right) \tag{3.2.4.1}$$

και

$$S_e\left(T_m^{(2)}\right) = P\left(T_m^{(2)} > C_2 \mid Y = 1\right) \tag{3.2.4.2}$$

όπου $C_1 \in (m \cdot a, m \cdot b_1]$ και $C_2 \in (m \cdot a, m \cdot b_2]$ είναι τα διαχωριστικά όρια. Στην περίπτωση των σύνθετων κλιμάκων, τα διαχωριστικά όρια συμβολίζονται με τα κεφαλαία γράμματα C_1 και C_2. Στην περίπτωση που ένας ερευνούμενος πληθυσμός περιλαμβάνει άτομα που διαθέτουν ένα συγκεκριμένο χαρακτηριστικό, τότε οι συναρτήσεις ευαισθησίας των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$, ορίζονται σύμφωνα με τις ισότητες (3.2.4.3) και (3.2.4.4):

$$S_e\left(T_m^{(1)}\right) = \int_{C_1}^{m \cdot b_1} f_m^{(1)}(t)\, dt = 1 - F_m^{(1)}(C_1) \tag{3.2.4.3}$$

και

$$S_e\left(T_m^{(2)}\right) = \int_{C_2}^{m \cdot b_2} f_m^{(2)}(t)\, dt = 1 - F_m^{(2)}(C_2) \tag{3.2.4.4}$$

Επομένως, κάνοντας χρήση του ορισμού ευαισθησίας, αποτυπώνονται οι μαθηματικές εκφράσεις των αντίστοιχων δεσμευμένων πιθανοτήτων σύνθετων κλιμάκων με διαφορετικό πεδίο ορισμού, με χρήση άγνωστης συνεχούς κατανομής.

Η διαφορά $m(t)$ μεταξύ των συναρτήσεων της πυκνότητας πιθανότητας των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$ θεωρείται ως εξής:

$$m(t) = f_m^{(1)}(t) - f_m^{(2)}(t) \qquad (3.2.4.5)$$

■ **<u>Λήμμα 2.3</u>** (**L**$_{C2}$). Για την διαφορά $m(t)$, $t \in R_{T_m^{(1)}} \cup R_{T_m^{(2)}}$, ισχύει:

$$\int_{m\cdot a}^{m\cdot b_1} m(t)\,dt = \int_{m\cdot b_1}^{m\cdot b_2} f_m^{(2)}(t)\,dt \qquad (\mathbf{L_{Cm}})$$

<u>Απόδειξη</u>: Ολοκληρώνοντας κατά μέλη από $m\cdot a$ στο $m\cdot b_1$ στη σχέση (3.2.4.5), ισχύει ότι:

$$\int_{m\cdot a}^{m\cdot b_1} m(t)\,dt = \int_{m\cdot a}^{m\cdot b_1} f_m^{(1)}(t)\,dt - \int_{m\cdot a}^{m\cdot b_1} f_m^{(2)}(t)\,dt = 1 - \int_{m\cdot a}^{m\cdot b_1} f_m^{(2)}(t)\,dt =$$

$$= \int_{m\cdot a}^{m\cdot b_2} f_m^{(2)}(t)\,dt - \int_{m\cdot a}^{m\cdot b_1} f_m^{(2)}(t)\,dt = \int_{m\cdot b_1}^{m\cdot b_2} f_m^{(2)}(t)\,dt \qquad ■$$

<u>Συνθήκη 2.4.1</u> (**C**$_{Cm11}$)

- Για $C_1 \geq C_2$ η διαφορά $m(t)$ για $t \in R_{T_m^{(1)}} \cup R_{T_m^{(2)}}$, ικανοποιεί την ανισότητα:

$$\int_{m\cdot a}^{C_1} m(t)\,dt \geq 0 \qquad (\mathbf{C_{Cm11}})$$

<u>Συνθήκη 2.4.2</u> (**C**$_{Cm12}$)

- Για $C_1 < C_2$ η διαφορά $m(t)$ για $t \in R_{T_m^{(1)}} \cup R_{T_m^{(2)}}$, ικανοποιεί την ανισότητα:

$$\int_{m\cdot a}^{C_1} m(t)\,dt \geq \int_{C_1}^{C_2} f_m^{(1)}(t)\,dt \qquad (\mathbf{C_{Cm12}})$$

Ερμηνεία Συνθήκης 2.4.1 (C$_{Cm11}$): Η συνθήκη (**C**$_{Cm11}$) δηλώνει ότι η τιμή του αριστερού μέλους, το οποίο εκφράζει την διαφορά μεταξύ των αθροιστικών συναρτήσεων κατανομής των ψευδών αρνητικών τιμών μεταξύ της κλίμακας με τις συνιστώσες με στενότερο εύρος στηρίγματος και της κλίμακας με τις συνιστώσες με το πλατύτερο εύρος στηρίγματος θα πρέπει να είναι μεγαλύτερη από ή ίση του μηδεν.

Ερμηνεία Συνθήκης 2.4.2 (C$_{Cm12}$): Η συνθήκη (C_{Cm12}) εκφράζει ότι το αριστερό μέλος της ανισότητας, η οποία εκφράζει τη διαφορά μεταξύ των αθροιστικών συναρτήσεων κατανομής για τις ψευδώς αρνητικές τιμές μεταξύ της κλίμακας με τις συνιστώσες στενότερου μήκους στηρίγματος και της κλίμακας με τις συνιστώσες μεγαλύτερου μήκους στηρίγματος θα πρέπει να είναι μεγαλύτερο από το δεξιό μέλος

που εκφράζει την αθροιστική συνάρτηση κατανομής για τις ψευδώς αρνητικές τιμές της κλίμακας $T_m^{(2)}$ για τιμές μεγαλύτερες από το διαχωριστικό όριο C_1.

<u>Πρόταση 2.4</u> ($\mathbf{P_{Cm}}$). Έστω $T_m^{(1)} = \sum_{i=1}^{m} X_i^{(1)}$ και $T_m^{(2)} = \sum_{i=1}^{m} X_i^{(2)}$ κλίμακες με συνιστώσες $X_i^{(1)}, X_i^{(2)}$, $i = 1, 2, ..., m$, που κατανέμονται με άγνωστο τρόπο και στήριγμα $R_{X^{(1)}} = [a, b_1]$ και $R_{X^{(2)}} = [a, b_2]$, αντιστοίχως. Η ευαισθησία των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$ είναι $S_e\left(T_m^{(1)}\right)$ και $S_e\left(T_m^{(2)}\right)$, αντιστοίχως. Τότε, για όλες τις δυνατές τιμές των διαχωριστικών ορίων C_1 και C_2 ισχύει $S_e\left(T_m^{(1)}\right) \le S_e\left(T_m^{(2)}\right)$, υπό τις συνθήκες 2.4.1 και 2.4.2.

<u>Απόδειξη</u> Αρκεί να αποδειχθεί ότι για τους πραγματικούς αριθμούς $a, b_1, b_2 \in \mathrm{R}$ και $m \in N$ με $b_1 < b_2$ ισχύει η ακόλουθη ανισότητα (3.2.4.6):

$$S_e\left(T_m^{(1)}\right) \le S_e\left(T_m^{(2)}\right) \qquad (3.2.4.6)$$

Επιπλέον, σύμφωνα με τους ορισμούς (3.2.4.3) και (3.2.4.4), η ανισότητα (3.2.4.6) γίνεται:

$$\int_{C_1}^{m \cdot b_1} f_m^{(1)}(t)\,dt \le \int_{C_2}^{m \cdot b_2} f_m^{(2)}(t)\,dt \qquad (3.2.4.6)$$

Τρεις περιπτώσεις διακρίνονται όσον αφορά τη σχετική θέση μεταξύ των ορίων C_1 και C_2:

- **<u>Περίπτωση 1</u>**: $m \cdot a < C_1 = C_2 = C \le m \cdot b_1 < m \cdot b_2$

Με τη χρήση του λήμματος (2.3) και της συνθήκης $\mathbf{C_{Cm11}}$ η συνάρτηση ευαισθησίας της κλίμακας $T_m^{(1)}$ γίνεται:

$$S_e\left(T_m^{(1)}\right) = \int_C^{m \cdot b_1} f_m^{(1)}(t)\,dt = \int_C^{m \cdot b_1} \left(f_m^{(2)}(t) + m(t) \right) dt = \int_C^{m \cdot b_1} f_m^{(2)}(t)\,dt + \int_C^{m \cdot b_1} m(t)\,dt \overset{(\mathbf{C_{Cm11}})}{\le}$$

$$\le \int_C^{m \cdot b_1} f_m^{(2)}(t)\,dt + \int_C^{m \cdot b_1} m(t)\,dt + \int_{m \cdot a}^{C} m(t)\,dt = \int_C^{m \cdot b_1} f_m^{(2)}(t)\,dt + \int_{m \cdot a}^{m \cdot b_1} m(t)\,dt =$$

$$= \int_C^{m \cdot b_1} f_m^{(2)}(t)\,dt + \int_{m \cdot b_1}^{m \cdot b_2} f_m^{(2)}(t)\,dt = \int_C^{m \cdot b_2} f_m^{(2)}(t)\,dt = S_e\left(T_m^{(2)}\right)$$

Επομένως λαμβάνουμε την ανισότητα:

$$S_e\left(T_m^{(1)}\right) \le S_e\left(T_m^{(2)}\right)$$

Κατά συνέπεια, η ανισότητα (3.2.4.6) ισχύει.

- **<u>Περίπτωση 2</u>**: $m \cdot a < C_2 < C_1 \le m \cdot b_1 < m \cdot b_2$

Με τη χρήση του λήμματος (2.3) και της συνθήκης **C$_{\mathbf{Cm11}}$**, η συνάρτηση ευαισθησίας της κλίμακας $T_m^{(1)}$ γίνεται:

$$S_e\left(T_m^{(1)}\right)=\int_{C_1}^{m\cdot b_1} f_m^{(1)}(t)\,dt=\int_{C_1}^{m\cdot b_1}\left(f_m^{(2)}(t)+m(t)\right)dt=\int_{C_1}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{C_1}^{m\cdot b_1} m(t)\,dt \overset{(\mathbf{C_{Cm11}})}{\leq}$$

$$\leq\int_{C_1}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{C_1}^{m\cdot b_1} m(t)\,dt+\int_{m\cdot a}^{C_1} m(t)\,dt=\int_{C_1}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{m\cdot a}^{m\cdot b_1} m(t)\,dt=$$

$$=\int_{C_1}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{m\cdot b_1}^{m\cdot b_2} f_m^{(2)}(t)\,dt=\int_{C_1}^{m\cdot b_2} f_m^{(2)}(t)\,dt\leq\int_{C_1}^{m\cdot b_2} f_m^{(2)}(t)\,dt+\int_{C_2}^{C_1} f_m^{(2)}(t)\,dt=$$

$$=\int_{C_2}^{m\cdot b_2} f_m^{(2)}(t)\,dt=S_e\left(T_m^{(2)}\right)$$

Επομένως λαμβάνουμε την ανισότητα:

$$S_e\left(T_m^{(1)}\right)\leq S_e\left(T_m^{(2)}\right)$$

Κατά συνέπεια, η ανισότητα (3.2.4.6) ισχύει.

- **<u>Περίπτωση 3</u>**: $m\cdot a<C_1<C_2\leq m\cdot b_1<m\cdot b_2$

Με τη χρήση του λήμματος (2.3) και της συνθήκης **C$_{\mathbf{Cm12}}$**, η συνάρτηση ευαισθησίας της κλίμακας $T_m^{(1)}$ γίνεται:

$$S_e\left(T_m^{(1)}\right)=\int_{C_1}^{m\cdot b_1} f_m^{(1)}(t)\,dt=\int_{C_1}^{m\cdot b_1}\left(f_m^{(2)}(t)+m(t)\right)dt=\int_{C_1}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{C_1}^{m\cdot b_1} m(t)\,dt=$$

$$=\int_{C_1}^{C_2} f_m^{(2)}(t)\,dt+\int_{C_2}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{C_1}^{m\cdot b_1} m(t)\,dt \overset{(\mathbf{C_{Cm12}})}{\leq}$$

$$\leq\int_{m\cdot a}^{C_1} m(t)\,dt+\int_{C_2}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{C_1}^{m\cdot b_1} m(t)\,dt=\int_{C_2}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{m\cdot a}^{m\cdot b_1} m(t)\,dt=$$

$$=\int_{C_2}^{m\cdot b_1} f_m^{(2)}(t)\,dt+\int_{m\cdot b_1}^{m\cdot b_2} f_m^{(2)}(t)\,dt=\int_{C_2}^{m\cdot b_2} f_m^{(2)}(t)\,dt=S_e\left(T_m^{(2)}\right)\Rightarrow S_e\left(T_m^{(1)}\right)\leq S_e\left(T_m^{(2)}\right)$$

Επομένως λαμβάνουμε την ανισότητα:

$$S_e\left(T_m^{(1)}\right)\leq S_e\left(T_m^{(2)}\right)$$

Κατά συνέπεια, η ανισότητα (3.2.4.6) ισχύει οπότε η πρόταση 2.3 ισχύει.

<u>Πόρισμα 2.4</u> (C$_{\mathbf{Cm}}$): Η ευαισθησία της σύνθετης συνεχούς κλίμακας $T_m\in[ma,mb]$ είναι μια αύξουσα (μη φθίνουσα) συνάρτηση του μήκους l του στηρίγματος των συνιστωσών μεταβλητών του, υπό ορισμένες συνθήκες (**C$_{\mathbf{Cm11}}$** - **C$_{\mathbf{Cm12}}$**), στην περίπτωση που οι συνιστώσες μεταβλητές ακολουθούν μια άγνωστη κατανομή.

3.2.5 Σύνθετη Συνεχής Κλίμακα & Εύρος Στηρίγματος Συνιστωσών - Περίπτωση Ομοιόμορφης κατανομής

Αυτή η παράγραφος παρουσιάζει την ειδική περίπτωση μιας σύνθετης κλίμακας T_m, η οποία είναι κατασκευασμένη από το άθροισμα των *m* ανεξάρτητων μεταβλητών που ακολουθούν την ομοιόμορφη κατανομή X_i, $i=1,2,...,m$. Επομένως, για κάθε μία συνιστώσα των μεταβλητών ισχύει $X_i \sim \bigcup_C [a,b]$, $i=1,2,...,m$ και διερευνάται κατά πόσο το εύρος l του στηρίγματος τους επηρεάζει την ακρίβεια της σύνθετης κλίμακας.

Η κατανομή του αθροίσματος των *m* ανεξάρτητων, ταυτόσημα, κατανεμημένων ομοιόμορφων συνεχών μεταβλητών έχει ερευνηθεί στην πρόσφατη βιβλιογραφία (*Bradley & Gupta, 2002; Potuschak & Mller, 2009; Sadooghi-Alvandi et al., 2009).* Δεδομένου ότι οι τυχαίες μεταβλητές, X_i, $i=1,2,...,m$ είναι ανεξάρτητες, με βάση τη συνέλιξη των ταυτόσημα, κατανεμημένων μεταβλητών, η συνάρτηση πυκνότητας πιθανότητας, καθώς επίσης και, η αθροιστική συνάρτηση κατανομής της μεταβλητής T_m που εκφράζει το άθροισμα τους, δίνεται από τους ακόλουθους τύπους (3.2.5.1) και (3.2.5.2) (*Buonocore et al., 2009*):

$$f_m(t)=\begin{cases}\sum_{r=0}^{\left\lfloor \frac{t-m\cdot a}{b-a}\right\rfloor}(-1)^r\binom{m}{r}\left(t-m\cdot a-r(b-a)\right)^{m-1}, & t\in[ma,mb]\\ 0, & ,t\notin[ma,mb]\end{cases} \tag{3.2.5.1}$$

και

$$F_m(t)=\begin{cases}0, & t<ma\\ \sum_{r=0}^{\left\lfloor \frac{t-m\cdot a}{b-a}\right\rfloor}(-1)^r\binom{m}{r}\left(t-m\cdot a-r(b-a)\right)^{m}, & t\in[ma,mb]\\ 1, & t>mb\end{cases} \tag{3.2.5.2}$$

όπου $\left\lfloor \frac{t-m\cdot a}{b-a}\right\rfloor$ υποδεικνύει το ακέραιο μέρος του αριθμού $\frac{t-m\cdot a}{b-a}$. Στα σχήματα (3.2.5.1) και (3.2.5.2), η συνάρτηση πυκνότητας πιθανότητας $f_m(t)$ και η αθροιστική συνάρτηση πιθανότητας $F_m(t)$ παρουσιάζονται γραφικά για $a=0$, $b=1$ και $m=20$, αντιστοίχως. Δύο συνεχείς τυχαίες μεταβλητές $T_m^{(1)}$ και $T_m^{(2)}$ θεωρούνται, με

αντίστοιχα στηρίγματα (πεδία ορισμού) $R_{T_m^{(1)}} = [m \cdot a, m \cdot b_1]$ και $R_{T_m^{(2)}} = [m \cdot a, m \cdot b_2]$ αντιστοίχως, με $b_1 < b_2$. Επιπλέον η συνάρτηση πυκνότητας πιθανότητας, καθώς επίσης και, η αθροιστική συνάρτηση κατανομής των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$ δίνονται από τις σχέσεις (3.2.5.1) και (3.2.5.2) αντικαθιστώντας κατάλληλα το άνω όριο του στηρίγματος με b_1 ή b_2.

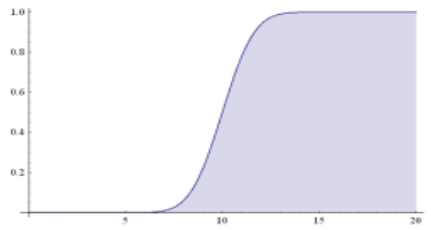
Σύμφωνα με τους ορισμό οι συναρτήσεις της ευαισθησίας των μεταβλητών $T_m^{(1)}$ και $T_m^{(2)}$ δίνονται από τους ακόλουθους τύπους (3.2.5.3) και (3.2.5.4):

$$S_e\left(T_m^{(1)}\right) = \int_{C_1}^{m \cdot b_1} f_m^{(1)}(t)\,dt = 1 - F_m^{(1)}\left(C_1\right) \tag{3.2.5.3}$$

και

$$S_e\left(T_m^{(2)}\right) = \int_{C_2}^{m \cdot b_2} f_m^{(2)}(t)\,dt = 1 - F_m^{(2)}\left(C_2\right) \tag{3.2.5.4}$$

Σχήμα 3.2.5.1: Συνάρτηση Πιθανότητας της μεταβλητής X για σ=1, 1.5 & 2



Σχήμα 3.2.5.2: Αθροιστική Κατανομή της μεταβλητής X για μ=-1, 0 & 1

όπου $C_1 \in (m \cdot a, m \cdot b_1]$ και $C_2 \in (m \cdot a, m \cdot b_2]$ είναι τα διαχωριστικά όρια που υπολογίζονται από τη μεγιστοποίηση του αθροίσματος των συναρτήσεων ευαισθησίας και ειδικότητας. Επομένως, L_1 και L_2 είναι τα μήκη των στηριγμάτων των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$. Η διαφορά $m(t)$ μεταξύ των συναρτήσεων ευαισθησίας μεταξύ των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$ όταν σχηματίζονται από συνιστώσες που κατανέμονται ομοιόμορφα λαμβάνει την ακόλουθη μορφή (3.2.5.5):

$$m(t) = \frac{\dfrac{\sum_{r=0}^{\left\lfloor \frac{t-ma}{b_1-a} \right\rfloor} (-1)^r \binom{m}{r} \left(t - ma - r\left(b_1 - a\right)\right)^{m-1}}{\left(b_1 - a\right)^m} - \dfrac{\sum_{r=0}^{\left\lfloor \frac{t-ma}{b_2-a} \right\rfloor} (-1)^r \binom{m}{r} \left(t - ma - r\left(b_2 - a\right)\right)^{m-1}}{\left(b_2 - a\right)^m}}{(m-1)!} \tag{3.2.5.5}$$

Χωρίς βλάβη της γενικότητας, αν θέσουμε το *α* ίσο με μηδέν (το κάτω όριο του πεδίου ορισμού (στηρίγματος) *α* θεωρείται ίσο με το 0), τότε η προηγούμενη σχέση θα απλοποιηθεί ως εξής:

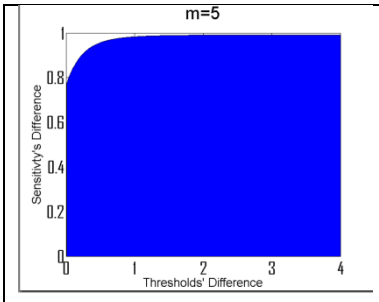
$$m(t)=\frac{\dfrac{\sum_{r=0}^{\left\lfloor \frac{t}{b_1}\right\rfloor}(-1)^r\binom{m}{r}(t-rb_1)^{m-1}}{b_1^{\,m}}-\dfrac{\sum_{r=0}^{\left\lfloor \frac{t}{b_2}\right\rfloor}(-1)^r\binom{m}{r}(t-rb_2)^{m-1}}{b_2^{\,m}}}{(m-1)!} \qquad (3.2.5.6)$$

Στη συνέχεια, σύμφωνα με τις σχέσεις (3.2.5.3) και (3.2.5.4), επίσης χωρίς απώλεια της γενικότητας, εάν το κατώτερο όριο *α* το θέσουμε ως μηδέν, η διαφορά $S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)$ γίνεται (3.2.5.7):

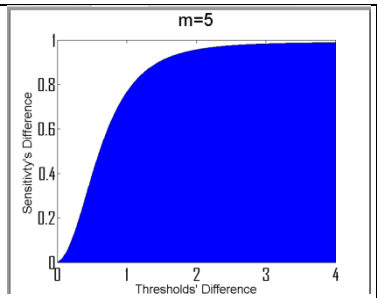
$$S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)=\frac{\dfrac{\sum_{r=0}^{\left\lfloor \frac{C_1}{b_1}\right\rfloor}(-1)^r\binom{m}{r}(C_1-rb_1)^{m}}{b_1^{\,m}}-\dfrac{\sum_{r=0}^{\left\lfloor \frac{C_2}{b_2}\right\rfloor}(-1)^r\binom{m}{r}(C_2-rb_2)^{m}}{b_2^{\,m}}}{m!} \qquad (3.2.5.7)$$

Προκειμένου να ερευνηθεί η μονοτονία της συνάρτησης ευαισθησίας, πρέπει να εξεταστεί το πρόσημο της διαφοράς της σχέσης (3.2.5.7). Σε καθένα από τα σχήματα (3.2.5.3) – (3.2.5.11), παρουσιάζονται οι γραφικές παραστάσεις για συγκεκριμένες αριθμητικές τιμές των *m* συνιστωσών μεταβλητών όσον αφορά την σχετική θέση μεταξύ των διαχωριστικών ορίων C_1 και C_2. Ο άξονας $Y'Y$ κάθε γραφικής παράστασης αντιπροσωπεύει τις τιμές της διαφοράς $S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)$ και ο άξονας $X'X$ αντιπροσωπεύει τη διαφορά του μήκους των στηριγμάτων b_2-b_1 των κλιμάκων, για το συγκεκριμένο εύρος. Ορίζουμε το b_1 να λαμβάνει μια σταθερή τιμή (την τιμή 1) και ορίζουμε το b_2 να αυξάνεται μέχρι την τιμή f=3. Πιο συγκεκριμένα, στα σχήματα (3.2.5.3), (3.2.5.4) και (3.2.5.5) παρουσιάζονται τρία γραφήματα τα οποία δείχνουν τη διαφορά $S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)$ λαμβάνοντας μη αρνητικές τιμές, καθώς το μήκος της διαφοράς μεταξύ b_1 και b_2 μεγαλώνει, στην περίπτωση που για τα διαγνωστικά κατώφλια ισχύει $C_1>C_2$. Επιπροσθέτως, στα σχήματα (3.2.5.6), (3.2.5.7) και (3.2.5.8) παρουσιάζονται τρία γραφήματα τα οποία δείχνουν τη διαφορά $S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)$ λαμβάνοντας τόσο μη αρνητικές όσο και αρνητικές τιμές, καθώς

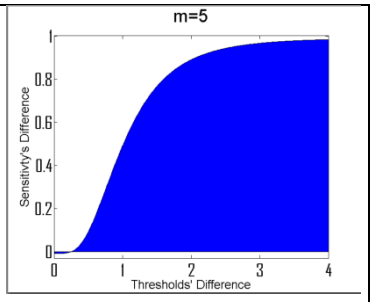
το εύρος της διαφοράς μεταξύ b_1 και b_2 μεγαλώνει στην περίπτωση που για τα διαχωριστικά όρια ισχύει $C_1 = C_2$. Στα σχήματα (3.2.5.9), (3.2.5.10) και (3.2.5.11) παρουσιάζονται τρία γραφήματα που δείχνουν τη διαφορά $S_e\left(T_m^{(2)}\right) - S_e\left(T_m^{(1)}\right)$ λαμβάνοντας τόσο μη αρνητικές τιμές όσο και αρνητικές τιμές καθώς το μήκος της διαφοράς μεταξύ b_1 και b_2 μεγαλώνει στην περίπτωση που για τα διαχωριστικά όρια ισχύει $C_1 < C_2$.



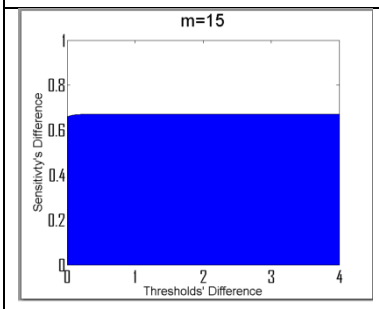
Σχήμα 3.2.5.3: C_1=4 & C_2=2



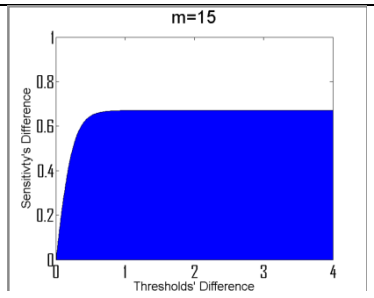
Σχήμα 3.2.5.6: C_1=4 & C_2=4



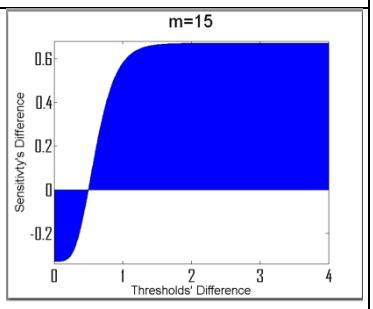
Σχήμα 3.2.5.9: C_1=4 & C_2=5



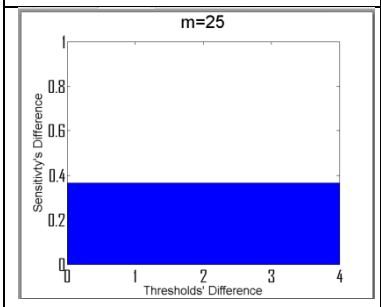
Σχήμα 3.2.5.4: C_1=8 & C_2=5



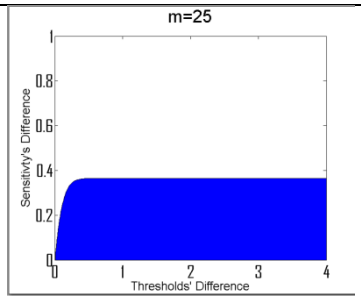
Σχήμα 3.2.5.7: C_1=8 & C_2=8



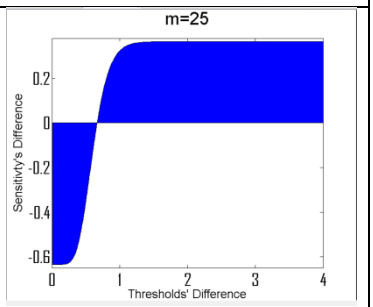
Σχήμα 3.2.5.10: C_1=8 & C_2=12



Σχήμα 3.2.5.5: C_1=12 & C_2=8



Σχήμα 3.2.5.8: C_1=12 & C_2=12



Σχήμα 3.2.5.11: C_1=12 & C_2=20

Ένα γενικό σχόλιο είναι ότι για όλα τα σχήματα η διαφορά τείνει ασυμπτωτικά σε έναν θετικό αριθμό, που εξαρτάται κυρίως από τον αριθμό m του πλήθους των συνιστωσών μεταβλητών, καθώς η διαφορά μεταξύ b_1 και b_2 αυξάνεται. Αυτό σημαίνει ότι η ευαισθησία της κλίμακας με το ευρύτερο στήριγμα αυξάνεται ενώ η απόσταση μεταξύ b_1 και b_2 αυξάνεται επίσης και καταλήγει σε ένα ορισμένο όριο που

εξαρτάται από τον αριθμό *m* των συνιστωσών καθώς και από τις τιμές διαχωριστικών κατωφλίων C_1 και C_2. Όταν αυξάνεται ο αριθμός *m* των συνιστωσών μεταβλητών, η ευαισθησία αυξάνεται επίσης, με μειούμενο ρυθμό αύξησης.

Ειδικότερα, για την κάθε περίπτωση ισχύει:

- . Στην περίπτωση που ισχύει $C_1 > C_2$, η διαφορά $S_e\left(T_m^{(2)}\right) - S_e\left(T_m^{(1)}\right)$ λαμβάνει μη αρνητικές τιμές, ακόμα και όταν η απόσταση μεταξύ του b_1 και b_2 είναι εξαιρετικά μικρή. Σε αυτή την περίπτωση η συνθήκη 2.4 (C_{Cm}) δεν είναι απαραίτητη.
- Στην περίπτωση που ισχύει $C_1 = C_2 = C$, η διαφορά $S_e\left(T_m^{(2)}\right) - S_e\left(T_m^{(1)}\right)$ λαμβάνει ως επί το πλείστον μη αρνητικές τιμές. Αρνητικές τιμές εμφανίζονται για υψηλές τιμές του *m* και του *C*, καθώς επίσης όταν η απόσταση μεταξύ του b_1 και του b_2 είναι σχεδόν αμελητέα. Όταν η διαφορά $S_e\left(T_m^{(2)}\right) - S_e\left(T_m^{(1)}\right)$ λαμβάνει αρνητικές τιμές, η συνθήκη 2.4 (C_{Cm}) δεν ισχύει. Η περίπτωση αυτή αντικατοπτρίζει τη σύνδεση των θεωρητικών τεκμηρίων αυτής της εργασίας και των πραγματικών δεδομένων, λόγω του γεγονότος ότι για ένα καθορισμένο διαχωριστικό όριο, η ευαισθησία μιας κλίμακας αυξάνεται όταν αυξάνεται και το εύρος του αντίστοιχου στηρίγματος των συνιστωσών. Αυτή η θετική σχέση μεταξύ της ευαισθησίας και του εύρους του στηρίγματος δύναται να είναι χρήσιμη και ωφέλιμη σε ένα πραγματικό περιβάλλον.
- Στην περίπτωση που ισχύει $C_1 < C_2$, η διαφορά $S_e\left(T_m^{(2)}\right) - S_e\left(T_m^{(1)}\right)$ λαμβάνει τόσο αρνητικές όσο και μη αρνητικές τιμές. Αρνητικές τιμές εμφανίζονται όταν η απόσταση μεταξύ του b_1 και b_2 είναι μικρή και όταν συμβαίνει αυτό, η συνθήκη 2.4 (C_{Cm}), δεν ισχύει. Όταν το b_2 είναι αρκούντως μεγάλο, η διαφορά $S_e\left(T_m^{(2)}\right) - S_e\left(T_m^{(1)}\right)$ γίνεται θετική και τείνει ασυμπτωτικά σε όριο που εξαρτάται από το *m* τη σχετική θέση των C_1 και C_2

Στη συνέχεια, ειδικές περιπτώσεις παρουσιάζονται με χρήση θεωρητικής απόδειξης σύμφωνα με τις σχετικές τιμές μεταξύ C_1, C_2, b_1 και b_2.

- **<u>1η Ειδική περίπτωση</u>**: $C_1 < b_1$ και $C_2 < b_2$

Όταν ισχύουν οι σχέσεις $C_1 < b_1$ και $C_2 < b_2$, η διαφορά $S_e\left(T_m^{(2)}\right) - S_e\left(T_m^{(1)}\right)$ γίνεται:

$$S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)=\frac{\left(b_2\cdot C_1\right)^m-\left(b_1\cdot C_2\right)^m}{b_1^m\cdot b_2^m\cdot m!} \qquad (3.2.5.8)$$

Επομένως, η διάφορα των ευαισθησιών $S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)$ για να είναι μη αρνητική, αρκεί να ισχύει:

$$\frac{b_2}{b_1}\geq\frac{C_2}{C_1} \qquad (3.2.5.9)$$

Η ανισότητα (3.2.5.9) παρέχει μία συνθήκη προκειμένου η συνάρτηση ευαισθησίας να είναι αύξουσα (μη φθίνουσα) όταν ισχύει $C_1<b_1$ και $C_2<b_2$. Επιπλέον, εάν $C_1=C_2=C$ τότε η διαφορά της σχέση (3.2.5.9) γίνεται:

$$S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)=\frac{C^m\cdot\left(b_2^m-b_1^m\right)}{b_1^m\cdot b_2^m\cdot m!}\overset{(b_1<b_2)}{>}0 \qquad (3.2.5.10)$$

Έτσι, στην περίπτωση που $C_1=C_2=C<b_1<b_2$, η συνάρτηση ευαισθησίας μιας σύνθετης κλίμακας είναι αύξουσα (μη φθίνουσα), δίχως κάποια συνθήκη να είναι απαραίτητη.

- **<u>2η Ειδική περίπτωση</u>**: $C_1=C_2=C$

Η πιο ενδιαφέρουσα περίπτωση είναι όταν για τα διαχωριστικά όρια ισχύει $C_1=C_2=C$, διότι στην περίπτωση αυτή η μονοτονία της συνάρτησης της ευαισθησίας μπορεί να διερευνηθεί στο πλαίσιο μιας κοινής βάσης όσον αφορά το διαχωριστικό όριο C. Επομένως, αν υποτεθεί:

$$m\cdot a<C_1=C_2=C\leq m\cdot b_1<m\cdot b_2$$

τότε η διαφορά $S_e\left(T_m^{(2)}\right)-S_e\left(T_m^{(1)}\right)$ από τη σχέση (3.2.5.5) γίνεται:

$$\frac{b_2^m\sum_{r=0}^{\left\lfloor\frac{C}{b_1}\right\rfloor}(-1)^r\binom{m}{r}\left(C-rb_1\right)^m-b_1^m\sum_{r=0}^{\left\lfloor\frac{C}{b_2}\right\rfloor}(-1)^r\binom{m}{r}\left(C-rb_2\right)^m}{b_1^m\cdot b_2^m\cdot m!} \qquad (3.2.5.10)$$

Για τα όρια b_1 και b_2 ισχύει:

$$b_1<b_2\Leftrightarrow\begin{cases}(-1)^r\binom{m}{r}\left(C-rb_1\right)^m\geq(-1)^r\binom{m}{r}\left(C-rb_2\right)^m,\ r=2k\\(-1)^r\binom{m}{r}\left(C-rb_1\right)^m\leq(-1)^r\binom{m}{r}\left(C-rb_2\right)^m,\ r=2k+1\end{cases}$$

Το παραπάνω συμπέρασμα αναφέρεται σε κάθε έναν από τους όρους του αθροίσματος. Το συνολικό άθροισμα εξαρτάται επίσης από τη σχετική θέση μεταξύ του b_1 και b_2, καθώς επίσης και από την τιμή του διαχωριστικού ορίου C.

3.2.6 Σύνθετη Συνεχής Κλίμακα & Εύρος Στηρίγματος Συνιστωσών Μεταβλητών - Περίπτωση Κανονικής κατανομής

Η παράγραφος αυτή παρουσιάζει την ειδική περίπτωση μιας σύνθετης κλίμακας T_m που κατασκευάζεται από το άθροισμα των m ανεξάρτητων ταυτόσημα κανονικά κατανεμημένων μεταβλητών X_i, $i=1,2,\ldots,m$. Έτσι, για κάθε μία από τις συνιστώσες μεταβλητές ισχύει $X_i \sim N\left(\mu,\sigma^2\right)$, $i=1,2,\ldots,m$ και ερευνάται εάν οι παράμετροι μ και σ της κατανομής τους, έχουν επίπτωση στην ακρίβεια μιας σύνθετης κλίμακας, αντί του εύρους του στηρίγματος, όπως στην προηγούμενη ενότητα. Όπως είναι γνωστό, το άθροισμα T_m, των m ανεξάρτητων κανονικά κατανεμημένων τυχαίων μεταβλητών ακολουθεί την κανονική κατανομή (*Feller, 1966*). Επομένως, δύο συνεχείς τυχαίες μεταβλητές $T_m^{(1)}=\sum_{i=1}^{m} X_i^{(1)}$ με $X_i^{(1)} \sim N\left(\mu_1,\sigma_1^2\right)$, $i=1,2,\ldots,m$, και $T_m^{(2)}=\sum_{i=1}^{m} X_i^{(2)}$ με $X_i^{(2)} \sim N\left(\mu_2,\sigma_2^2\right)$, $i=1,2,\ldots,m$ εξετάζεται. Για τις κλίμακες $T_m^{(1)}$ και $T_m^{(2)}$ εφαρμόζεται $T_m^{(1)} \sim N\left(m\mu_1, m\sigma_1^2\right)$ και $T_m^{(2)} \sim N\left(m\mu_2, m\sigma_2^2\right)$. Η συνάρτηση πυκνότητας πιθανότητας $f_m^{(1)}(t)$ και $f_m^{(2)}(t)$ για τις κλίμακες $T_m^{(1)}$ και $T_m^{(2)}$ δίνεται από τους ακόλουθους τύπους (3.2.6.1) και (3.2.6.2):

$$f_m^{(1)}(t)=\frac{1}{\sqrt{2\pi m}\cdot\sigma_1}\cdot\exp\left(\frac{-\left(t-m\mu_1\right)^2}{2m\sigma_1^2}\right) \tag{3.2.6.1}$$

και

$$f_m^{(2)}(t)=\frac{1}{\sqrt{2\pi m}\cdot\sigma_2}\cdot\exp\left(\frac{-\left(t-m\mu_2\right)^2}{2m\sigma_2^2}\right) \tag{3.2.6.2}$$

Έτσι, σύμφωνα με τον ορισμό της ευαισθησίας, υποθέτοντας ότι ο πληθυσμός που ερευνήθηκε περιλαμβάνει άτομα που διαθέτουν ένα συγκεκριμένο χαρακτηριστικό, η συνάρτηση ευαισθησίας των κλιμάκων $T_m^{(1)}$ και $T_m^{(2)}$ δίνεται από τις ακόλουθες σχέσεις (3.2.6.3) και (3.2.6.4):

$$S_e\left(T_m^{(1)}\right)=\frac{1}{\sqrt{2\pi m}\cdot\sigma_1}\int_{C_1}^{\infty}\exp\left(\frac{-\left(t-m\mu_1\right)^2}{2m\sigma_1^2}\right)dt=1-\Phi\left(\frac{C_1-m\mu_1}{\sqrt{m}\cdot\sigma_1}\right) \quad (3.2.6.3)$$

και

$$S_e\left(T_m^{(2)}\right)=\frac{1}{\sqrt{2\pi m}\cdot\sigma_2}\int_{C_2}^{\infty}\exp\left(\frac{-\left(t-m\mu_2\right)^2}{2m\sigma_2^2}\right)dt=1-\Phi\left(\frac{C_2-m\mu_2}{\sqrt{m}\cdot\sigma_2}\right) \quad (3.2.6.4)$$

όπου $C_1, C_2 \in \mathbb{R}$ είναι τα διαχωριστικά όρια που προκύπτουν από τη μεγιστοποίηση του αθροίσματος ευαισθησίας και ειδικότητας. Οι πίνακες A' και B ορίζονται ως ακολούθως:

$$A'=\begin{bmatrix} C_1 & C_2 \\ \sigma_1 & \sigma_2 \end{bmatrix} \text{ και } B=\begin{bmatrix} \mu_1 & \mu_2 \\ \sigma_1 & \sigma_2 \end{bmatrix}$$

<u>Συνθήκη 2.5</u> ($\mathbf{P_{CNm}}$). Για τις ορίζουσες $\det\left(A'\right)$ και $\det\left(B\right)$ ικανοποιείται η ακόλουθη ανισότητα:

$$\det\left(A'\right)\geq m\cdot\det\left(B\right) \quad (3.2.6.5)$$

<u>Πρόταση 2.5</u> ($\mathbf{P_{CNm}}$). Στην περίπτωση που οι συνιστώσες μεταβλητές X_i, $i=1,2,\ldots,m$ μιας σύνθετης κλίμακας, είναι ταυτόσημα κανονικά κατανεμημένες, η ευαισθησία της κλίμακας T_m είναι μια αύξουσα (μη φθίνουσα) συνάρτηση των παραμέτρων μ και σ, υπό συγκεκριμένη συνθήκη (P_{CNm}).

<u>Απόδειξη</u> Σύμφωνα με τις σχέσεις (3.2.6.3) και (3.2.6.4), η ανισότητα $S_e\left(T_m^{(2)}\right)\geq S_e\left(T_m^{(1)}\right)$ γίνεται:

$$\Phi\left(\frac{C_1-m\mu_1}{\sqrt{m}\cdot\sigma_1}\right)\geq\Phi\left(\frac{C_2-m\mu_2}{\sqrt{m}\cdot\sigma_2}\right) \quad (3.2.6.6)$$

Είναι γνωστό ότι η συνάρτηση $\boldsymbol{\Phi}$ είναι αύξουσα (μη φθίνουσα), επομένως η ανισότητα της σχέσης (3.2.6.6), γίνεται:

$$\frac{C_1-m\cdot\mu_1}{\sqrt{m}\cdot\sigma_1}\geq\frac{C_2-m\cdot\mu_2}{\sqrt{m}\cdot\sigma_2}$$

οπότε έχουμε

$$\sigma_2\cdot\left(C_1-m\cdot\mu_1\right)\geq\sigma_1\cdot\left(C_2-m\cdot\mu_2\right)$$

με χρήση των οριζουσών των πινάκων Α και Β έχουμε:

$$\det\left(A'\right)\geq m\cdot\det\left(B\right) \quad (3.2.6.7)$$

Επομένως, η ανισότητα (3.2.6.6) ισχύει υπό την συνθήκη (3.2.6.5).

Στη συνέχεια παρουσιάζονται, ειδικές περιπτώσεις όσον αφορά τις αριθμητικές τιμές των παραμέτρων μ και σ της κατανομής, καθώς επίσης και οι τιμές των διαχωριστικών ορίων.

- **<u>Ειδική περίπτωση 3.2.6.1</u>**: $C_1 = C_2 = C$.

Σύμφωνα με την σχέση (3.2.6.7), η ανισότητα $S_e\left(T_m^{(2)}\right) \geq S_e\left(T_m^{(1)}\right)$ γίνεται:

$$\sigma_2 \cdot C - \sigma_1 \cdot C \geq m \cdot \left(\sigma_2 \cdot \mu_1 - \sigma_1 \cdot \mu_2\right)$$

που μπορεί να γραφεί με την μορφή της ανισότητας (3.2.6.8):

$$\left(\sigma_2 - \sigma_1\right) \cdot C \geq m \cdot \det\left(B\right) \qquad (3.2.6.8)$$

Επομένως, η συνάρτηση ευαισθησίας της σύνθετης κλίμακας είναι αύξουσα (μη φθίνουσα) ως προς το διαχωριστικό όριο C, όταν ισχύει:

- $C \geq \dfrac{m \cdot \det\left(B\right)}{\sigma_2 - \sigma_1}$ εάν $\sigma_2 > \sigma_1$
- $C \leq \dfrac{m \cdot \det\left(B\right)}{\sigma_2 - \sigma_1}$ εάν $\sigma_2 < \sigma_1$.

Εάν $\sigma_2 = \sigma_1$, τότε η ανισότητα $S_e\left(T_m^{(2)}\right) \geq S_e\left(T_m^{(1)}\right)$ γίνεται:

$$\mu_2 \geq \mu_1 \qquad (3.2.6.9)$$

<u>Πρόταση 2.5.1</u> ($\mathbf{P_{CNm.1}}$). Στην περίπτωση που η σύνθετη κλίμακα T_m κατανέμεται κανονικά, η ευαισθησία είναι μια αύξουσα (μη φθίνουσα) συνάρτηση της αναμενόμενης τιμής μ των συνιστωσών μεταβλητών όταν η τυπική απόκλιση σ είναι σταθερή.

- **<u>Ειδική περίπτωση 3.2.6.2</u>**: $\mu_1 = \mu_2 = \mu$.

Σύμφωνα με τη σχέση (3.2.6.7) η ανισότητα $S_e\left(T_m^{(2)}\right) \geq S_e\left(T_m^{(1)}\right)$ γίνεται:

$$\sigma_2 \cdot C_1 - \sigma_1 \cdot C_2 \geq m \cdot \left(\sigma_2 \cdot \mu - \sigma_1 \cdot \mu\right)$$

που μπορεί να γραφεί με την μορφή της ανισότητας (3.2.6.10):

$$\det\left(A'\right) \geq m \cdot \mu \cdot \left(\sigma_2 - \sigma_1\right) \qquad (3.2.6.10)$$

Επομένως, η συνάρτηση ευαισθησίας της σύνθετης κλίμακας είναι αύξουσα (μη φθίνουσα), ως προς τη μέση τιμή μ των συνιστωσών μεταβλητών, όταν:

- $\mu \leq \dfrac{\det\left(A'\right)}{m \cdot \left(\sigma_2 - \sigma_1\right)}$ εάν $\sigma_2 > \sigma_1$

- $\mu \geq \dfrac{\det(A')}{m \cdot (\sigma_2 - \sigma_1)}$ εάν $\sigma_2 < \sigma_1$

Εάν $\sigma_2 = \sigma_1$, τότε η ανισότητα $S_e\left(T_m^{(2)}\right) \geq S_e\left(T_m^{(1)}\right)$ γίνεται:

$$C_1 \geq C_2 \qquad (3.2.6.11)$$

- **<u>Ειδική περίπτωση 3.2.6.3</u>**: $\sigma_1 = \sigma_2 = \sigma$.

Σύμφωνα με την σχέση (3.2.6.7), η ανισότητα $S_e\left(T_m^{(2)}\right) \geq S_e\left(T_m^{(1)}\right)$ γίνεται:

$$\sigma \cdot C_1 - \sigma \cdot C_2 \geq m \cdot (\sigma \cdot \mu_1 - \sigma \cdot \mu_2)$$

που μπορεί να γραφεί με την μορφή της ανισότητας (3.2.6.12):

$$m \cdot (\mu_2 - \mu_1) \geq C_2 - C_1 \Rightarrow m \cdot d(\mu) \geq d(C) \qquad (3.2.6.12)$$

<u>Πρόταση 2.5.2</u> ($P_{CNm.2}$). Στην περίπτωση που η σύνθετη κλίμακα T_m κατανέμεται κανονικά, η συνάρτηση ευαισθησίας είναι αύξουσα (μη φθίνουσα), εάν η αύξηση της αναμενόμενης τιμής, των συνιστωσών μεταβλητών πολλαπλασιαζόμενη με το *m* να είναι μεγαλύτερο από την αύξηση των διαχωριστικών ορίων, όταν η τυπική απόκλιση *σ* είναι σταθερή.

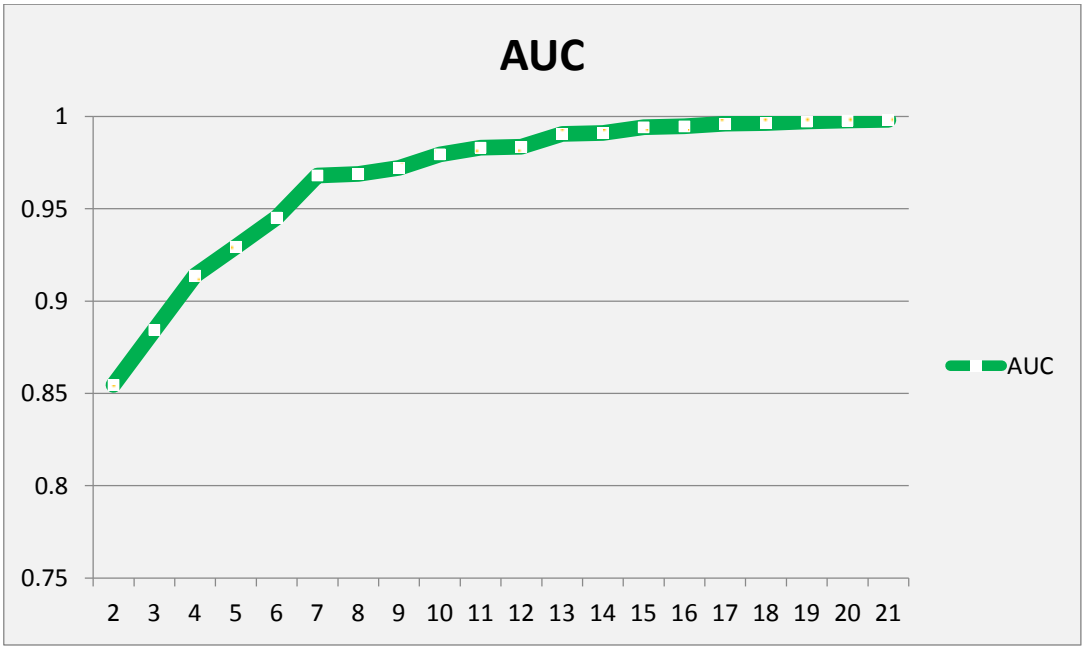
3.3. Αποτελέσματα Προσομοιώσεων

3.3.1 Ακρίβεια Σύνθετων Κλιμάκων ως Προς τον Αριθμό *m* των Συνιστωσών Μεταβλητών

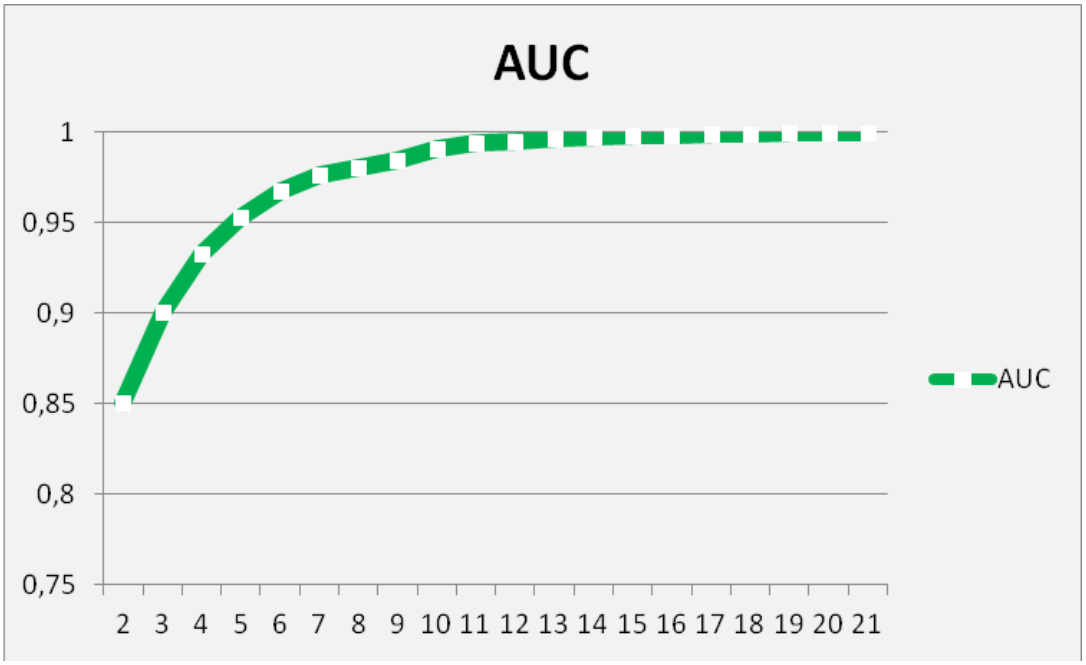
Όσον αφορά το ερευνητικό ερώτημα που εκφράζει το εάν το πλήθος των συνιστωσών της κλίμακας επηρεάζει τη διαγνωστική ακρίβειά της, ανεξάρτητα από την κατανομή των συνιστωσών του, τα γραφήματα (3.3.1.1) και (3.3.1.2) κατασκευάστηκαν για να παρουσιάσουν τη γραφική αναπαράσταση της σχέσης του εμβαδού κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (AUC) σε συνδυασμό με τον αριθμό *m* των συνιστωσών μεταβλητών, σε σχέση με ένα συγκεκριμένο χαρακτηριστικό, μιας σύνθετης κλίμακας. Από τα σχήματα (3.3.1.1) και (3.3.1.2) είναι προφανές ότι όταν αυξήθηκε ο αριθμός των συνιστωσών, που σχετίζονται με το χαρακτηριστικό που η κλίμακα αποσκοπεί να εκτιμήσει, επίσης αυξήθηκε ο μέσος όρος του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) των εννέα κλιμάκων, αλλά ο ρυθμός αύξησης του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) μειώθηκε (η κλίση της

πολυγωνικής γραμμής μειώθηκε καθώς αυξανόταν το πλήθος των συνιστωσών). Σε κάποιο βαθμό, το παραπάνω αποτέλεσμα είναι αναμενόμενο, καθώς όταν εισάγονται νέες μεταβλητές σε ένα προγνωστικό μοντέλο, η πληροφόρηση αυξάνει, οπότε βελτιώνεται το μοντέλο, με αποτέλεσμα να αυξάνει και η τιμή του AUC. Επειδή όμως, η μέγιστη τιμή του AUC είναι το 1, επίσης είναι λογικό να μειώνεται ο ρυθμός αύξησης καθώς τείνει ασυμπτωτικά στη μέγιστη τιμή του (Αλγόριθμος 1 - Παράρτημα Α).

Στην περίπτωση ενσωμάτωσης μεταβλητών που έχουν μηδενική ή εξαιρετικά ασθενή (θετική ή αρνητική) συσχέτιση, η μεταβολή του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) ήταν σχεδόν αμελητέα. Συγκεκριμένα, υπάρχει κάποιο όριο για τον αριθμό των συνιστωσών (σχετιζόμενων με το χαρακτηριστικό από μέτρια έως ισχυρά) μεταβλητών *m*, πέρα από αυτόν, η τιμή του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) δεν φαίνεται να αυξάνεται σημαντικά. Αυτό συμβαδίζει με το γεγονός ότι, για παράδειγμα, σε μοντέλα γραμμικής παλινδρόμησης, η αύξηση του αριθμού των ανεξάρτητων μεταβλητών δεν παρέχει απαραίτητα πολλές περισσότερες πρόσθετες ποιοτικές πληροφορίες, καθώς υπάρχει μια ποινή (penalty) για κάθε πρόσθετη επεξηγηματική μεταβλητή που εισάγεται στο μοντέλο (*Theil, 1961*).



Σχήμα 3.3.1.1: Αποτελέσματα προσομοίωσης των κλιμάκων σε 1.000 επαναλήψεις - Κανονική κατανομή



Σχήμα 3.3.1.2: Αποτελέσματα προσομοίωσης των κλιμάκων σε 1.000 επαναλήψεις - Ομοιόμορφη κατανομή

Για όλα τα σενάρια που χρησιμοποιήθηκαν, οι αρχικές και οι τελικές μεταβλητές ακολούθησαν ακριβώς την ίδια κατανομή και οι πρόσθετες/ενδιάμεσες συνιστώσες μεταβλητές, είτε μετά την κανονική είτε την ομοιόμορφη κατανομή, προστέθηκαν μεταξύ του πρώτου και του τελευταίου στοιχείου για λόγους αντικειμενικής σύγκρισης. Το αποτέλεσμα αυτό ισχύει στην περίπτωση των κλιμάκων με μεταβλητές που ακολουθούν κανονική ή ομοιόμορφη κατανομή.

3.3.2 Διαγνωστική Ικανότητα Σταθμισμένων Σύνθετων Κλιμάκων

Στην παράγραφο αυτή παρουσιάζονται τα αποτελέσματα τόσο από προσομοιωμένα όσο και από πραγματικά δεδομένα. Στις παραγράφους 3.3.2.1 και 3.3.2.2 προκύπτουν αποτελέσματα από προσομοιωμένα δεδομένα, όσον αφορά την ακρίβεια ταξινόμησης ατόμων που διαθέτουν συγκεκριμένα χαρακτηριστικά ή όχι από δεδομένα κανονικής και ομοιόμορφης κατανομής, αντιστοίχως (Αλγόριθμος (1) Ταξινόμησης – Δεδομένα Προσομοίωσης – Παράρτημα Α). Στην παράγραφο 3.3.2.3 αξιολογείται η ικανότητα πρόβλεψης των προτεινόμενων κλιμάκων (Αλγόριθμος (2) Πρόβλεψης – Δεδομένα Προσομοίωσης – Παράρτημα Α). Επιπλέον, στην παράγραφο 3.3.2.4 παρουσιάζονται τα αποτελέσματα της εφαρμογής των σταθμισμένων κλιμάκων σε μέρος των επιδημιολογικών δεδομένων υγείας της ΑΤΤΙΚΑ (Αλγόριθμος (3) Ταξινόμησης – Πραγματικά Δεδομένα – Παράρτημα Α).

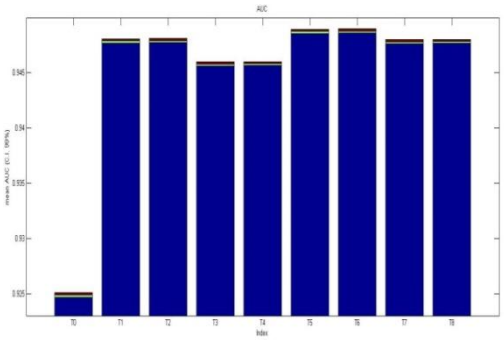
3.3.2.1 Κανονικά Κατανεμημένες Συνιστώσες Μεταβλητές

Σε αυτή την παράγραφο παρουσιάζονται προσομοιωμένα αποτελέσματα με χρήση συνιστωσών μεταβλητών που δημιουργούν ένα μια κλίμακα και ακολουθούν την κανονική κατανομή. Πιο συγκεκριμένα, στην περίπτωση που όλες οι κλίμακες αποτελούνται από m=6 συνιστώσες μεταβλητές που ακολουθούν την κανονική κατανομή με διαφορετικές παραμέτρους, τα αποτελέσματα των προσομοιώσεων δίδονται γραφικά στα σχήματα 3.3.2.1 – 3.3.2.6. Στο σημείο αυτό τονίζεται ότι παρουσιάζονται αποτελέσματα για την περίπτωση των m=6 συνιστωσών μεταβλητών για πρακτικούς λόγους συμπύκνωσης της έκτασης των αποτελεσμάτων καθώς, για όλες τις τιμές του m που εξετάστηκαν, τα αποτελέσματα είχαν όμοια ποιοτική αποτύπωση (Παράρτημα Α).

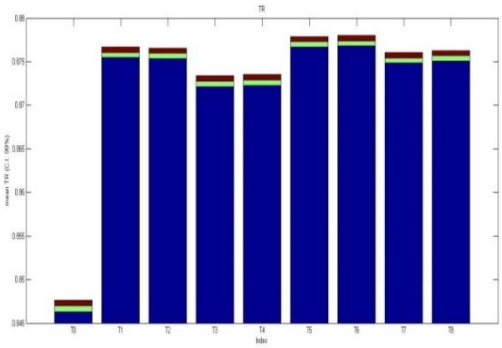
Πίνακας 3.3.2.1 Μέσες τιμές των μέτρων αξιολόγησης για όλους τους δείκτες - λόγος 1: 3.

Index	AUC	TR (%)	TPR (%)	TNR (%)	PPV (%)	NPV (%)
T_0	**0.92487***	**84.696***	**84.692***	**84.711***	**64.948***	**94.314***
T_1	**0.94789**	**87.608**	**87.579**	**87.631**	**70.311**	**95.481**
T_2	**0.94792**	**87.595**	**87.615**	**87.602**	**70.268**	**95.492**
T_3	**0.94580**	**87.275**	**87.403**	**87.245**	**69.624**	**95.402**
T_4	**0.94582**	**87.289**	**87.374**	**87.274**	**69.661**	**95.392**
T_5	**0.94875**	**87.728**	**87.698**	**87.751**	**70.537**	**95.528**
T_6	**0.94879**	**87.739**	**87.689**	**87.770**	**70.572**	**95.526**
T_7	**0.94781**	**87.542**	**87.697**	**87.504**	**70.119**	**95.516**
T_8	**0.94785**	**87.567**	**87.662**	**87.548**	**70.186**	**95.506**

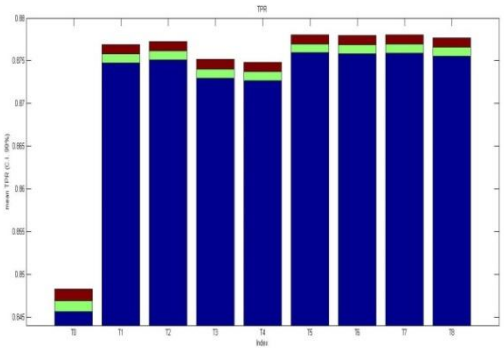
* Στατιστικώς σημαντική διαφοροποίηση στις τιμές των κριτηρίων μεταξύ του T_0 και των υπολοίπων κλιμάκων, $p_v<0.01$



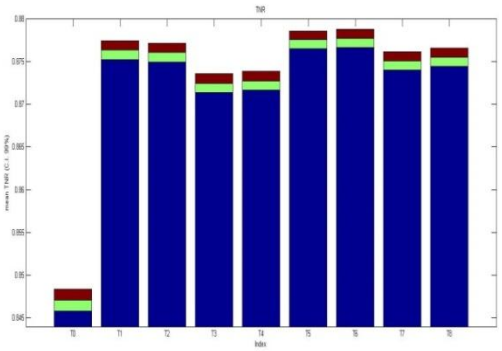
Σχήμα 3.3.2.1: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς AUC



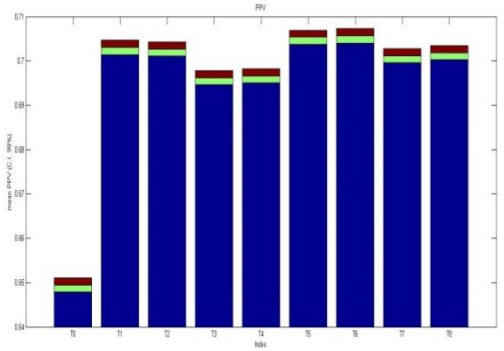
Σχήμα 3.3.2.2: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς TR



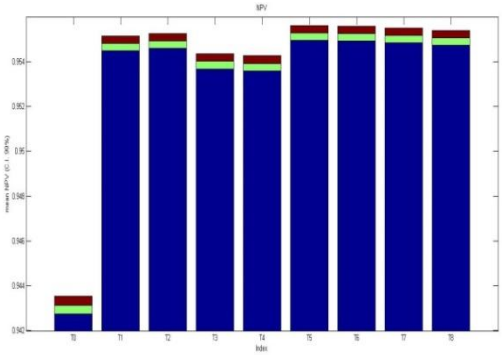
Σχήμα 3.3.2.3: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς TPR



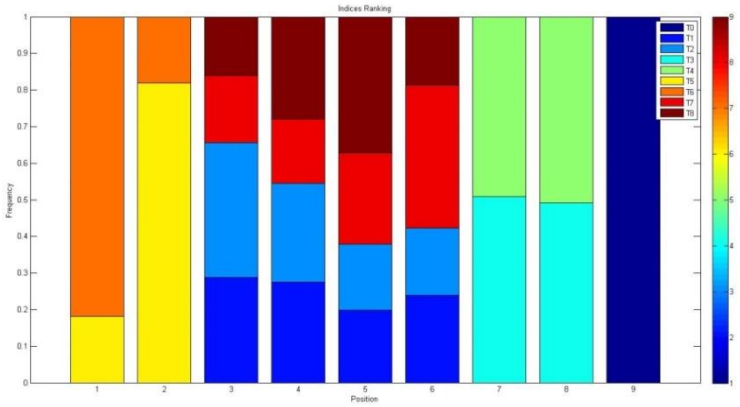
Σχήμα 3.3.2.4: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς TNR



Σχήμα 3.3.2.5: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς PPV



Σχήμα 3.3.2.6: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς NPV



Σχήμα 3.3.2.7: Αποτελέσματα προσομοίωσης για τη συχνότητα της θέσης ταξινόμησης των κλιμάκων σε 1.000 επαναλήψεις

Πιο συγκεκριμένα, παρουσιάζονται: το εμβαδόν κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (AUC), το αληθές ποσοστό (TR), το αληθές θετικό ποσοστό (TPR), το αληθές αρνητικό ποσοστό (TNR), η θετική προγνωστική αξία (PPV), η αρνητική προγνωστική αξία (NPV) με τα αντίστοιχα διαστήματα εμπιστοσύνης 99%. Επιπρόσθετα, στον πίνακα 3.3.2.1 παρουσιάζονται οι μέσες τιμές των κριτηρίων που χρησιμοποιούνται για τις εννέα κλίμακες των 10.000 παραγόμενων δειγμάτων.

Οι τιμές με μπλε χρώμα αντιστοιχούν στις υψηλότερες τιμές των κριτηρίων και οι τιμές με το κόκκινο χρώμα γραμματοσειράς αντιστοιχούν στις χαμηλότερες τιμές των χρησιμοποιούμενων κριτηρίων. Η μη σταθμισμένη κλίμακα T_0 βρέθηκε συνεχώς στη χαμηλότερη θέση όλων των κριτηρίων που χρησιμοποιήθηκαν (p_v<0,01). Οι σταθμισμένες κλίμακες T_5 και T_6 που παράχθησαν με τη χρήση σταθμίσεων από τη γραμμική διαχωριστική ανάλυση (LDA) πέτυχαν τις μεγαλύτερες τιμές. Όπως φαίνεται από τα σχήματα 3.3..2.1 - 3.3.2.6, οι σταθμισμένες κλίμακες T_1 – T_8 πέτυχαν υψηλότερες τιμές έναντι της μη σταθμισμένης κλίμακας T_0 σε όλα τα μέτρα αξιολόγησης.

Η σταθμισμένη κλίμακα T_6 αντιστοιχούσε στην υψηλότερη μέση τιμή των μέτρων/κριτηρίων "εμβαδόν υπό της καμπύλης λειτουργικών χαρακτηριστικών" (AUC), "αληθές ποσοστό" (TR), "αληθώς αρνητικό ποσοστό" (TNR) και "θετική προγνωστική αξία" (PPV), ενώ η σταθμισμένη κλίμακα T_5 αντιστοιχούσε στην υψηλότερη μέση τιμή των μέτρων/κριτηρίων "αληθώς θετικό ποσοστό" (TPR) και "αρνητική προγνωστική αξία" (NPV). Αυτή η παρατήρηση οδηγεί στο συμπέρασμα ότι πιθανόν είναι καλύτερη η χρήση διαφορετικών κλιμάκων στην περίπτωση γενικών και ειδικών πληθυσμών. Οι κλίμακες που δημιουργήθηκαν από τη μέθοδο γραμμικής διαχωριστικής ανάλυσης (LDA) πέτυχαν υψηλότερες τιμές σε όλα σχεδόν τα κριτήρια που χρησιμοποιήθηκαν σε σχέση με τις κλίμακες που δημιουργήθηκαν από τη μέθοδο λογιστικής παλινδρόμησης (BLR). Επιπλέον, η κατάταξη των κλιμάκων ως προς τη συχνότητα των θέσεών τους, όσον αφορά το εμβαδόν υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC), παρουσιάζεται στο σχήμα 4.2.2.7, στο οποίο τα διαφορετικά χρώματα στις ράβδους αντιστοιχούν στις εννέα κλίμακες. Η σταθμισμένη κλίμακα T_6 (πορτοκαλί χρώμα) αντιστοιχεί στην υψηλότερη τιμή εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC σε περίπου 82% των συνολικών πειραμάτων και στο υπόλοιπο 18%, την πρώτη θέση καταλαμβάνει η κλίμακα T_5 (κίτρινο χρώμα). Η μη σταθμισμένη κλίμακα T_0 (μπλε χρώμα) είναι

συνεχώς στην τελευταία θέση με μέση τιμή AUC 0,92487 (Πίνακας 3.3.2.1). Στην τρίτη θέση οι κλίμακες T_2 και T_1 εμφανίζονται ως επί το πλείστον με ποσοστά περίπου 34% και 27% αντίστοιχα. Οι κλίμακες T_7 και T_8 εμφανίζονται κυρίως στις θέσεις πέντε και έξι αντίστοιχα, με ποσοστά περίπου 40%. Οι κλίμακες T_4 και T_3 μοιράζονται τις θέσεις επτά και οκτώ με παρόμοια ποσοστά. Ένα άλλο αποτέλεσμα που παρατηρήθηκε είναι ότι οι κλίμακες που παράγονται με την ενίσχυση των βαρών τους διαιρώντας με τις στατιστικές συναρτήσεις του Λάμδα του Wilk (Wilk's Lambda) και την Απόκλιση (Deviance) παρουσιάζουν χαμηλότερη ακρίβεια έναντι των αρχικών κλιμάκων χρησιμοποιώντας τις συμπληρωματικές πιθανότητες από τη λογιστική παλινδρόμηση (OR) και του συντελεστές της διαχωριστικής συνάρτησης από τη διακρίνουσα ανάλυση (CDFC).

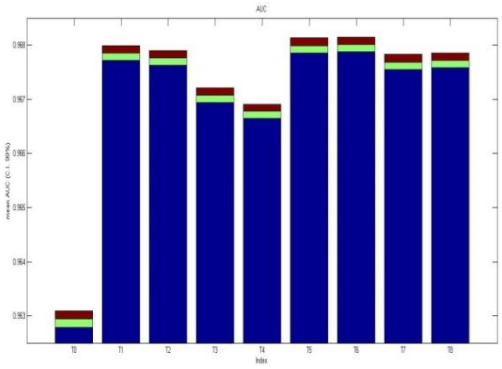
3.3.2.2 Ομοιόμορφα Κατανεμημένες Συνιστώσες Μεταβλητές

Σε αυτή την παράγραφο παρουσιάζονται τα αποτελέσματα προσομοίωσης από σύνθετες κλίμακες με συνιστώσες μεταβλητές που ακολουθούν την Ομοιόμορφη Κατανομή. Το εμβαδόν κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (AUC), το αληθές ποσοστό (TR), το αληθές θετικό ποσοστό (TPR), το αληθές αρνητικό ποσοστό (TNR), η θετική προγνωστική αξία (PPV), η αρνητική προγνωστική αξία (NPV) παρουσιάζονται στα σχήματα (3.3.2.8) - (3.3.2.13). Στον πίνακα 3.3.2.2 παρουσιάζονται οι μέσες τιμές των μέτρων αξιολόγησης που χρησιμοποιήθηκαν. Για άλλη μια φορά, η μη σταθμισμένη κλίμακα T_0 ήταν στη χαμηλότερη θέση όλων των χρησιμοποιούμενων κριτηρίων (pv <0,01) έναντι όλων των σταθμισμένων κλιμάκων. Συγκεκριμένα, οι σταθμισμένες κλίμακες T_5 και T_6 πέτυχαν τις μεγαλύτερες τιμές, όμοια με την περίπτωση της Κανονικής κατανομής (Παράρτημα Α). Το γεγονός ότι οι συνιστώσες μεταβλητές δεν ακολουθούν την κανονική κατανομή φαίνεται να επηρεάζει ελάχιστα τα προηγούμενα αποτελέσματα προσομοίωσης.

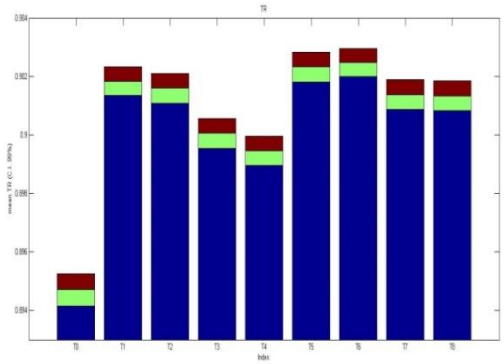
Πίνακας 3.3.2.2 Μέσες τιμές των μέτρων αξιολόγησης για όλους τις κλίμακες - λόγος 1: 3.

Δείκτης	AUC	TR (%)	TPR (%)	TNR (%)	PPV (%)	NPV (%)
T_0	0.96294*	89.471*	89.506*	89.472*	73.988*	96.230*
T_1	0.96785	90.184	90.177	90.199	75.471	96.488
T_2	0.96776	90.159	90.150	90.175	75.427	96.479
T_3	0.96707	90.005	89.951	90.037	75.119	96.404
T_4	0.96678	89.946	89.867	89.985	75.008	96.374
T_5	0.96799	90.232	**90.250**	90.239	75.562	**96.515**
T_6	**0.96801**	**90.247**	90.236	**90.264**	**75.601**	96.510
T_7	0.96769	90.138	90.137	90.151	75.376	96.473
T_8	0.96771	90.134	90.160	90.138	75.357	96.480

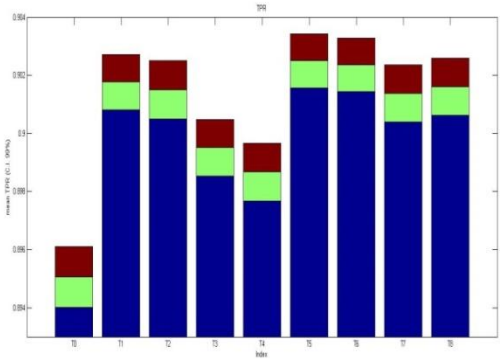
* Στατιστικώς σημαντική διαφοροποίηση στις τιμές των κριτηρίων μεταξύ του T_0 και των υπολοίπων κλιμάκων, $p_v<0.01$



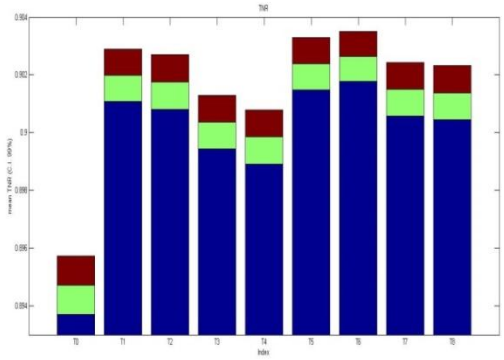
Σχήμα 3.3.2.8: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς AUC



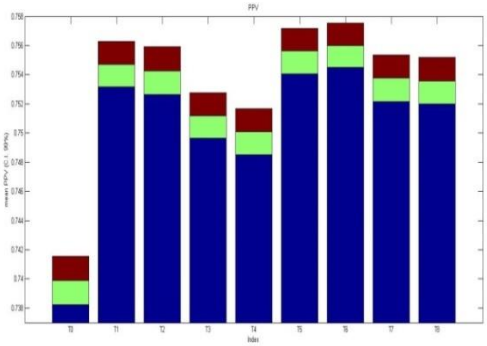
Σχήμα 3.3.2.9: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς TR



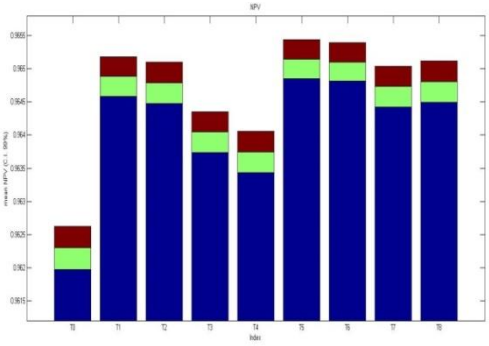
Σχήμα 4.2.2.10: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς TPR



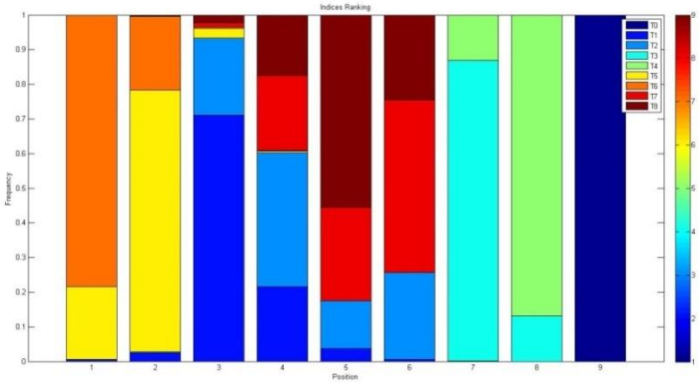
Σχήμα 4.2.2.11: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς TNR



Σχήμα 3.3.2.12: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς PPV



Σχήμα 3.3.2.13: Αποτελέσματα προσομοίωσης για την ταξινόμηση των κλιμάκων ως προς NPV



Σχήμα 3.3.2.14: Αποτελέσματα προσομοίωσης για τη συχνότητα ταξινόμησης θέσης των

κλιμάκων σε 1.000 επαναλήψεις

Το είδος της κατανομής φάνηκε να επηρεάζει στην περίπτωση που τα μεγέθη των προσομοιωμένων δειγμάτων ήταν αρκετά μικρότερα. Πιο συγκεκριμένα, για μικρότερου μεγέθους δείγματα ($n<100$) τα αποτελέσματα των σταθμισμένων κλιμάκων από τις μεθόδους της Λογιστικής Παλινδρόμησης και της Διακρίνουσας Ανάλυσης ήταν σχετικά όμοια. Οι σταθμισμένες κλίμακες T_1 – T_8 πέτυχαν υψηλότερες τιμές ($p_v<0,01$) σε όλα τα μέτρα σε σύγκριση με τις τιμές της κλίμακας T_0. Η κλίμακα T_6 αντιστοιχούσε στην υψηλότερη μέση τιμή των μέτρων/κριτηρίων: "εμβαδόν υπό της καμπύλης λειτουργικών χαρακτηριστικών" (AUC), "αληθές ποσοστό" (TR), "αληθώς αρνητικό ποσοστό" (TNR) και "θετική προγνωστική αξία" (PPV), ενώ η σταθμισμένη κλίμακα T_5 αντιστοιχούσε στην υψηλότερη μέση τιμή των μέτρων/κριτηρίων "αληθώς θετικό ποσοστό" (TPR) και "αρνητική προγνωστική αξία" (NPV). Οι κλίμακες που προέκυψαν από τη μέθοδο διαχωριστικής ανάλυσης (LDA) πέτυχαν υψηλότερες τιμές σε όλα σχεδόν τα κριτήρια που χρησιμοποιήθηκαν έναντι των δεικτών που δημιουργήθηκαν από τη μέθοδο λογιστικής παλινδρόμησης BLR, αλλά η διαφορά στα μέτρα αξιολόγησης ήταν χαμηλότερη από ότι στην περίπτωση της κανονικής κατανομής. Επιπλέον, η κατάταξη των κλιμάκων ως προς τη συχνότητα των θέσεών τους, όσον αφορά το εμβαδόν υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC), παρουσιάζεται στο σχήμα 3.3.2.2.14. Η σταθμισμένη κλίμακα T_6 αντιστοιχούσε στην υψηλότερη τιμή του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) για περίπου το 78% των συνολικών πειραμάτων, η σταθμισμένη κλίμακα T_5 έλαβε την πρώτη θέση στο 21% και η σταθμισμένη κλίμακα T_1 πήρε την πρώτη θέση στο 1% των συνολικών πειραμάτων. Η μη σταθμισμένη κλίμακα T_0 ήταν συνεχώς στην τελευταία θέση. Στη δεύτερη θέση η σειρά ήταν αντίθετη σε σχέση με αυτή της πρώτης θέσης. Στην τρίτη θέση οι κλίμακες T_2 και T_1 εμφανίστηκαν ως επί το πλείστον με ποσοστά περίπου 21% και 72% αντίστοιχα. Οι κλίμακες T_8 και T_7 εμφανίστηκαν κυρίως στις θέσεις πέντε και έξι, αντίστοιχα. Οι κλίμακες T_3 και T_4 εμφανίστηκαν κυρίως στις θέσεις επτά και οκτώ, αντίστοιχα.

3.3.2.3 Αξιολόγηση Πρόβλεψης

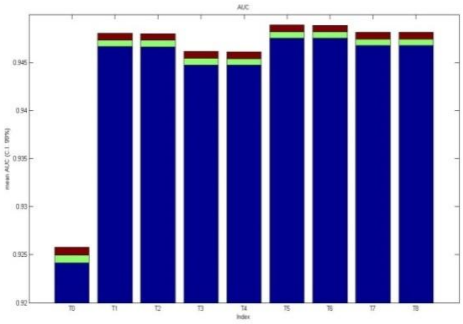
Τα αποτελέσματα από προσομοιωμένα δεδομένα όσον αφορά την απόδοση πρόβλεψης των προτεινόμενων κλιμάκων έναντι της μη σταθμισμένης σταθμισμένης

κλίμακας ελέγχθηκαν με τη χρήση τεχνικής διαχωρισμού του αρχικού δείγματος και πολλαπλής δειγματοληψίας.

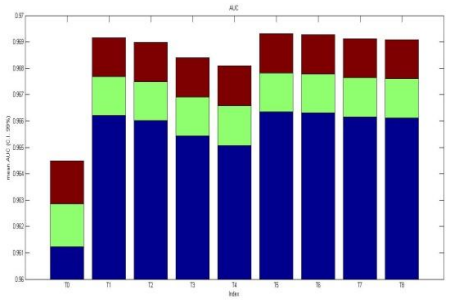
Πίνακας 3.3.2.3: Μέτρο AUC του δείγματος ελέγχου για T_0 -T_8

Κατανομή \ Δείκτης	T_0	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8
Κανονική	0.92501	0.94729	0.94726	0.94538	0.94534	0.94816	0.94813	0.94739	0.94738
Ομοιόμορφη	0.96286	0.96769	0.96750	0.96692	0.96658	0.96783	0.96779	0.96764	0.96760

Τα αποτελέσματα των προσομοιώσεων όσον αφορά το μέτρο του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) στο δείγμα δοκιμής, το οποίο αποτελεί μέτρο της καλής πρόβλεψης ή της καλής αναγνώρισης νέων περιπτώσεων, τόσο για τις κανονικές όσο και για τις ομοιόμορφες κατανομές, δίνονται γραφικά στα σχήματα (3.3.2.15) και (3.3.2.16). Το αρχικό δείγμα περιελάμβανε 10.000 περιπτώσεις και η αναλογία μεταξύ υγιών και μη υγιών ήταν 4:1. Στον πίνακα 3.3.2.3 παρουσιάζονται οι τιμές του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών AUC για κάθε σενάριο που πραγματοποιήθηκε (Παράρτημα Α).



Σχήμα 3.3.2.15: Αποτελέσματα πρόβλεψης βάσει AUC – Κανονική Κατανομή (Αλγόριθμος 2 - Παράρτημα Α)



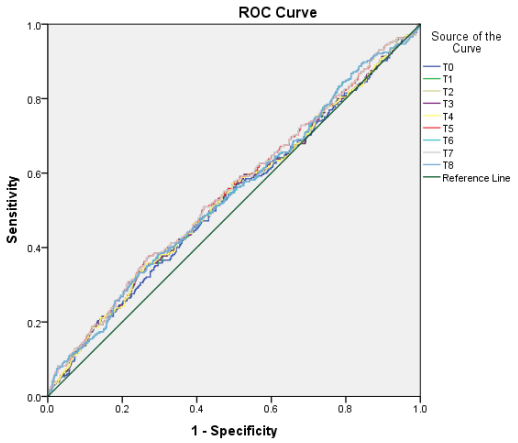
Σχήμα 4.2.2.1.16: Αποτελέσματα πρόβλεψης βάσει AUC – Ομοιόμορφη Κατανομή (Αλγόριθμος 2 - Παράρτημα Α)

Με βάση το κριτήριο του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC), στην περίπτωση πρόβλεψης ατόμων με ή δίχως ένα συγκεκριμένο χαρακτηριστικό, οι σταθμισμένες κλίμακες πέτυχαν υψηλότερη προβλεπτική ακρίβεια σε σύγκριση με τη μη σταθμισμένη κλίμακα. Επομένως, τα αποτελέσματα που παρουσιάστηκαν στα προηγούμενα κεφάλαια σχετικά με την ταξινόμηση ατόμων με ή δίχως ένα συγκεκριμένο χαρακτηριστικό, είναι σύμφωνα με τα αποτελέσματα αυτού του κεφαλαίου σχετικά με την πρόβλεψη νέων περιπτώσεων. Συγκεκριμένα, οι κλίμακες T_6 και T_5 αντιστοιχούσαν στις μεγαλύτερες τιμές του AUC. Τα αποτελέσματα, ως προς την προβλεπτική ικανότητα των κλιμάκων, ήταν όμοια ως προς τη χρήση και των υπόλοιπων χρησιμοποιηθέντων κριτηρίων, όπως η ευαισθησία, η ειδικότητα, η ακρίβεια, η θετική και η αρνητική προγνωστική αξία,

όπου, σε όλες τις περιπτώσεις, μεγαλύτερες τιμές πέτυχαν οι σταθμισμένες κλίμακες που προέρχονται από τη διαχωριστική ανάλυση.

3.3.2.4 Εφαρμογή σε δεδομένα από τη μελέτη ATTICA

Οι προτεινόμενες μέθοδοι στάθμισης για την κατασκευή κλιμάκων εφαρμόστηκαν σε ένα υποσύνολο πραγματικών δεδομένων της μελέτης ATTICA και επαληθεύτηκαν τα αποτελέσματα προσομοίωσης της προηγούμενης παραγράφου. Συγκεκριμένα, η εμφάνιση υπέρτασης (hypertension - HTN) χρησιμοποιήθηκε ως προβλεπτική μεταβλητή/αποτέλεσμα της υγείας και χρησιμοποιήθηκαν πέντε, επιλεγμένες τυχαία, επεξηγηματικές διατροφικές μεταβλητές (ψάρια, ξηροί καρποί, όσπρια, γαλακτοκομικά και λαχανικά) για την ταξινόμηση ατόμων όσον αφορά την υπέρταση (HTN). Σε αυτή την περίπτωση, οι κατανομές των αντίστοιχων συνιστωσών μεταβλητών ήταν άγνωστες και όλες απέτυχαν στους ελέγχους κανονικής και ομοιόμορφης κατανομής (οι τιμές p από τη δοκιμασία Kolmogorov-Smirnov <0,01). Ειδικότερα, το εμβαδόν υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) με το αντίστοιχο Δ.Ε. παρουσιάζεται στον πίνακα (3.3.2.4) και η καμπύλη ROC για τις εννέα κλίμακες παρουσιάζεται στο σχήμα (4.2.2.1.17).



Σχήμα 3.3.2.1.17: Καμπύλη ROC για τις εννέα κλίμακες

Πίνακας 3.3.2.4: Τιμή AUC με Δ.Ε.95% H_0: true area = 0.5

Δείκτης	AUC	95% Διάστημα Εμπιστοσύνης		Σχετική Αύξηση AUC
T_0	0.5289	0.4873	0.5705	-
T_1	0.5358	0.4942	0.5775	1.31%
T_2	0.5365	0.4948	0.5781	1.43%
T_3	0.5360	0.4944	0.5777	1.35%
T_4	0.5367	0.4950	0.5783	1.47%
T_5	0.5500	0.5087	0.5913	3.99%
T_6	0.5419	0.5006	0.5831	2.45%
T_7	0.5503	0.5090	0.5916	4.05%

T_8	0.5421	0.5009	0.5834	2.50%

Οι τιμές του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) δεν είναι υψηλές (σχήμα 3.3.2.1.17) σε αυτό το παράδειγμα λόγω του γεγονότος ότι η υπέρταση δεν εξαρτάται αποκλειστικά από τη διατροφική συμπεριφορά, αλλά και από άλλα χαρακτηριστικά όπως συχνότητα φυσικής άσκησης, επίπεδα στρες ή γενετικούς παράγοντες. Το ενδιαφέρον στοιχείο είναι ότι, όπως και στα αποτελέσματα προσομοίωσης, η μη σταθμισμένη κλίμακα πέτυχε χαμηλότερες τιμές έναντι όλων των σταθμισμένων κλιμάκων, στο κριτήριο AUC.

Η σταθμισμένη κλίμακα T_7 πέτυχε τη μέγιστη τιμή του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC), ενώ η μη σταθμισμένη κλίμακα T_0 πέτυχε τη χαμηλότερη τιμή του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC). Η τελευταία στήλη του πίνακα (3.3.2.4) παρουσιάζει τη σχετική αύξηση του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) για τις σταθμισμένες κλίμακες έναντι της μη σταθμισμένης κλίμακας. Όλες οι εκατοστιαίες τιμές αύξησης είναι πάνω από 1%, πράγμα που σημαίνει ότι σε γενικούς πληθυσμούς ένας αρκετά μεγάλος αριθμός ατόμων μπορεί να χαρακτηριστεί σωστά ως υγιής ή το αντίθετο, σε σχέση με την κλίμακα T_0. Η μέγιστη τιμή του αληθώς θετικού ποσοστού (TPR) επιτεύχθηκε με την κλίμακα T_8 και η κλίμακα T_0 παρουσίασε μικρή ικανότητα διάκρισης για ασθενή άτομα στο βέλτιστο διαγνωστικό όριο (Πίνακας 3.3.2.4). Αντίθετα, η μη σταθμισμένη κλίμακα πέτυχε την υψηλότερη τιμή του αληθώς αρνητικών ποσοστού (TNR) λόγω του γεγονότος ότι απέτυχε πλήρως στο ποσοστό των αληθώς θετικών (TPR). Όπως αναφέρθηκε, το κριτήριο του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) τοποθετεί την κλίμακα T_7 στην πρώτη θέση της σχετικής κατάταξης. Τα αποτελέσματα αυτά μπορούν να οδηγήσουν στη χρήση συνδυασμού κλιμάκων σε διάφορους πληθυσμούς, ειδικούς ή γενικούς, όπου η ευαισθησία ή η ειδικότητα παρουσιάζουν μεγάλο ενδιαφέρον. Δηλαδή, στην περίπτωση που υψηλή τιμή ευαισθησίας απαιτείται, να γίνεται χρήση μιας συγκεκριμένης κλίμακας και αντιστοίχως, στην περίπτωση που υψηλή τιμή ειδικότητας απαιτείται, να γίνεται χρήση ενός άλλου δείκτη.

Πίνακας 3.3.2.5 TPR and TNR values with 99% Bootstrap C.I.

Δείκτης	TPR (%)	99% Bootstrap C.I. for TPR		TNR (%)	99% Bootstrap C.I. for TNR	
T_0	20.38	14.17	26.94	86.59	83.17	89.72%
T_1	33.59	26.49	41.47	75.42	71.02%	79.48%
T_2	33.21	25.93	40.85	76.26	72.03%	80.36%
T_3	35.85*	28.30	43.65	73.18	68.88%	77.37%
T_4	33.58	26.27	41.33	75.98	71.73%	80.06%
T_5	37.36*	29.80	45.14	73.46	69.09%	77.78%
T_6	32.45	25.40	40.07	76.40	72.25%	80.44%
T_7	37.36*	29.80	45.14	73.60	69.22%	77.84%

T_8	**37.74***	**30.30**	**45.39**	**70.53**	**66.16%**	**74.86%**

(* statistical significance in criteria values between T_0 and the corresponding index, pv<0.01)

4. ΣΥΖΗΤΗΣΗ

4.1 Συμπεράσματα Διατριβής

Οι κλίμακες έχουν χρησιμοποιηθεί εκτενώς για την έρευνα στον τομέα της οικονομίας, της εκπαίδευσης και της βιο-ϊατρικής, ειδικά σε περιπτώσεις όπου τα χαρακτηριστικά των ατόμων είναι δύσκολο να μετρηθούν άμεσα, όπως π.χ. τα συναισθήματα και οι συνήθειες (*Kant. 1996*). Παρά το γεγονός, ότι αυτές οι κλίμακες είναι εξαιρετικά σημαντικές σε ένα ευρύ επιστημονικό πεδίο, δεν έχει δοθεί αρκετή προσοχή στο παρελθόν ώστε να καθιερωθεί μια μεθοδολογία για την κατασκευή τους (*Panagiotakos, 2009*). Ιδιαιτέρως, στην περίπτωση διακριτών κλιμάκων, δεν υπάρχει σαφής μέθοδος επιλογής του βέλτιστου αριθμού *k* των διαμερίσεων της κλίμακας των συνιστωσών (*Maroulas & Panagiotakos, 2010*). Στην περίπτωση συνεχών κλιμάκων, δεν υπάρχει σαφής μέθοδος επιλογής του βέλτιστου εύρους *l* του στηρίγματος των συνιστωσών (*Huskisson, 1974*). Επίσης, στη βιβλιογραφία δεν προτείνονται συγκεκριμένες μέθοδοι στάθμισης *w* των συνιστωσών (*Pepe et al, 2006*), αλλά ούτε υπάρχει σαφές πλαίσιο που να αφορά των αριθμό *m* των σχετιζόμενων μεταβλητών που απαρτίζουν μια κλίμακα.

Στην παρούσα εργασία, με βάση τις θεωρητικές και αναλυτικές αποδείξεις, συνιστάται η χρήση του μέγιστου αριθμού διαμερίσεων που δύναται να χρησιμοποιηθούν στην περίπτωση μιας διακριτής κλίμακας, επειδή σχετίζεται θετικά με την ακρίβεια της κλίμακας (*Bersimis et al, 2013*). Επίσης συνίσταται η χρήση ευρέως πεδίου ορισμού των συνιστωσών μεταβλητών της κλίμακας, επειδή σχετίζεται θετικά με την προβλεπτική ικανότητα της κλίμακας (*Bersimis et al, 2016*). Επιπροσθέτως, συνίσταται η χρήση σταθμίσεων που προέρχονται από στατιστικές μεθόδους ταξινόμησης, καθώς επίσης, η χρήση μεγάλου αριθμού συνιστωσών μεταβλητών, σχετιζόμενων με το χαρακτηριστικό που αποσκοπεί να εκτιμήσει ο δείκτης (*Bersimis et al, 2017*).

4.1.1 Επίδραση Αύξησης Αριθμού Διαιρέσεων Διακριτών Συνιστωσών Μεταβλητών στην Ευαισθησία Κλιμάκων

Οι κλίμακες αποτελούν πολύτιμο εργαλείο πολλών επιστημονικών κλάδων, όπως είναι η οικονομία, η εκπαίδευση, ψυχομετρία, η βιομετρία, κτλ. Για παράδειγμα, στο χώρο της υγείας χρησιμοποιούνται κλίμακες υγείας για την

ανίχνευση κλινικών χαρακτηριστικών ή στάσεων συμπεριφοράς (Kant, 1996). Συνεπώς, παρότι χρησιμοποιούνται εκτεταμένα στο χώρο της υγείας, στην πλειονότητα των περιπτώσεων οι *m* διακριτές συνιστώσες μεταβλητές σχηματίζουν τη σύνθετο κλίμακα υγείας χωρίς να λαμβάνεται υπόψη η επίδραση του αριθμού *k* των διαμερίσεων τους.

Σε αυτή τη διδακτορική εργασία, αποδείχθηκε θεωρητικά ότι η ευαισθησία μιας κλίμακας είναι μια αύξουσα συνάρτηση ως προς τον αριθμό των διαμερίσεων της (*Bersimis, et al., 2013*). Το εύρημα δίνει έναν επιπλέον λόγο για περεταίρω σκέψη στην περίπτωση που πρέπει να επιλεγεί ο κατάλληλος αριθμός *k* των διαμερίσεων μιας κλίμακας και είναι ιδιαίτερης μεθοδολογικής σημασίας για τη δημιουργία αξιόπιστων κλιμάκων με ακρίβεια, οι οποίοι έχουν σχεδιαστεί για να προβλέψουν διάφορα χαρακτηριστικά ατόμων (οικονομικό επίπεδο, εκπαιδευτικό επίπεδο, κλινική διάγνωση ασθενειών χωρίς συμπτώματα, κλίμακες για ψυχολογικές διαταραχές, κλίμακες διατροφικής κατάστασης) (*Harnack et al., 2002; McCullough et al, 2000),* καθώς επίσης και κλίμακες που χρησιμοποιούνται για διάφορους σκοπούς στις κοινωνικές επιστήμες, βιοεπιστήμες κλπ (*Kant. 1996; Mcdowell. 2002; kranz et al, 2006*). Σε πρόσφατη δημοσίευση τους οι Bersimis et al (*Bersimis et al, 2013*), απέδειξαν ότι η ευαισθησία μιας κλίμακας είναι αύξουσα (δεν μειώνεται) σε σχέση με τον αριθμό *k* των διαμερίσεων των συνιστωσών του. Επομένως, ο αριθμός των διαμερίσεων κάθε μεταβλητής συνιστώσας μιας κλίμακας, σχετίζεται θετικά με την ευαισθησία του δείκτη (*Bersimis et al, 2013*). Το εύρημα αυτό παρέχει ένα βασικό στοιχείο στη δημιουργία μιας ακριβούς διακριτής κλίμακας, προκειμένου να διαχωριστούν οι αληθώς έχοντες κάποιο χαρακτηριστικό σε σχέση με αυτούς που δεν το διαθέτουν.

Παρά το γεγονός, ότι οι κλίμακες με έναν μικρό αριθμό διαμερίσεων είναι εύκολες στην καθημερινή χρήση, σύμφωνα με το αποτέλεσμα της μαθηματικής απόδειξης αυτής της διδακτορικής εργασίας, καθώς επίσης και τα αποτελέσματα της αντίστοιχης προσομοίωσης, δεν επιτυγχάνεται υψηλή διαγνωστική ακρίβεια. Επομένως, είναι προτιμότερο να χρησιμοποιηθούν κλίμακες με μεγάλο αριθμό διαμερίσεων, προκειμένου να επιτευχθεί υψηλή ακρίβεια (*Kourlaba & Panagiotakos, 2009*).

4.1.2 Επίδραση Αύξησης Εύρους Στηρίγματος Συνιστωσών Μεταβλητών στην Ευαισθησία Κλιμάκων

Οι κλίμακες χρησιμοποιούνται για την ανίχνευση και πρόγνωση διαφόρων κλινικών καταστάσεων (παθήσεις χωρίς φανερά συμπτώματα, ψυχολογικές διαταραχές, διατροφική κατάσταση, κτλ) (*Harnack et al., 2002; McCullought et al., 2000*) ή χρησιμοποιείται σε κοινωνικές επιστήμες και στις βιοεπιστήμες (*Mcdowell, 2009; Kranz et al., 2006*).

Σε αυτή τη διδακτορική εργασία, αποδείχθηκε θεωρητικά ότι η ευαισθησία μιας κλίμακας είναι μια αύξουσα συνάρτηση ως προς το εύρος του στηρίγματος των συνιστωσών του (*Bersimis, et al., 2016*). Το εύρημα δίνει έναν επιπλέον λόγο για περεταίρω σκέψη στην περίπτωση που πρέπει να επιλεγεί το κατάλληλο εύρος l του πεδίου ορισμού των συνιστωσών μεταβλητών μιας σύνθετης κλίμακας και είναι ιδιαίτερης μεθοδολογικής σημασίας για τη δημιουργία αξιόπιστων κλιμάκων με διαχωριστική ακρίβεια (*Kant, 1996*).

Επεκτείνοντας τα προαναφερθέντα συμπεράσματα, σε αυτή την διατριβή αποδείχθηκε ότι η ευαισθησία ενός σχετιζόμενου με την υγεία δείκτη που σχηματίζεται από μία συνεχή μεταβλητή είναι μια αύξουσα συνάρτηση σε σχέση με το εύρος του στηρίγματος l της μεταβλητής υπό ορισμένη συνθήκη, καθώς επίσης, και η ευαισθησία μιας κλίμακας που σχηματίζεται από m συνεχείς μεταβλητές είναι μια αύξουσα συνάρτηση σε σχέση με το μήκος του στηρίγματος l των μεταβλητών υπό ορισμένη συνθήκη, όταν η κατανομή των συνιστωσών είναι άγνωστη. Το εύρημα αυτό επαληθεύθηκε στην περίπτωση που η κλίμακα κατανέμεται ομοιόμορφα ενώ εφαρμόστηκε στην περίπτωση που η κλίμακα σχηματίζεται από το άθροισμα των m συνεχών μεταβλητών που ακολουθούν την ομοιόμορφη κατανομή.

Επιπλέον, σε αυτή την εργασία μελετήθηκε η επίδραση των παραμέτρων της κανονικής κατανομής στην περίπτωση μιας κλίμακας, όταν σχηματίζεται από μία μεταβλητή που ακολουθεί την κανονική κατανομή είτε στην περίπτωση μιας σύνθετης κλίμακας, που σχηματίζεται από m συνεχείς μεταβλητές που ακολουθούν την κανονική κατανομή. Πιο συγκεκριμένα, βρέθηκαν κατάλληλες συνθήκες για το μ και το σ, ώστε η αντίστοιχη συνάρτηση ευαισθησίας να έχει αύξουσα μονοτονία.

Στην περίπτωση συνεχών σύνθετων κλιμάκων, δηλαδή κλιμάκων που δημιουργούνται από την άθροιση συνεχών συνιστωσών μεταβλητών (συνήθως τυποποιημένης κλίμακας) με πεπερασμένο εύρος, έχει αποδειχθεί ότι η αύξηση του

εύρους του στηρίγματος της εκάστοτε συνιστώσας μεταβλητής οδηγεί σε αύξηση της ευαισθησίας της κλίμακας (*Bersimis F. et al, 2016*). Αυτό το συμπέρασμα σημαίνει ότι για κάθε κλίμακα που χρησιμοποιείται, αν αυξηθεί η κλίμακα των συνιστωσών μεταβλητών, τότε αυξάνει και η ακρίβεια της.

4.1.3 Επίδραση Σταθμίσεων στις Συνιστώσες Μεταβλητές στην Ακρίβεια Κλιμάκων

Οι κλίμακες έχουν γίνει πολύτιμα εργαλεία σε πολλά επιστημονικά πεδία, συμπεριλαμβανομένης της ιατρικής έρευνας. Ωστόσο, αν και οι σύνθετες κλίμακες χρησιμοποιούνται ευρέως στην έρευνα για την υγεία, στις περισσότερες περιπτώσεις, οι συνιστώσες τους μεταβλητές αθροίζονται μέσω ενός απλού (μη σταθμισμένου) αθροίσματος των τιμών τους (*Patterson, Haines, & Popkin, 1994, Zung, 1965*). Η εργασία αυτή προτείνει, εκτός των άλλων, μια μέθοδο για την επιλογή μιας αποτελεσματικής σύνθετης κλίμακας χρησιμοποιώντας κατάλληλους διορθωτικούς συντελεστές σε σχέση με τα αποτελέσματα που επιδιώκει να ανιχνεύσει η κλίμακα, όπως είναι οι ψυχολογικές διαταραχές ή η κατάσταση της θρεπτικής επάρκειας κλπ. (*McCullough et al., 2000*). Η επιλογή κλιμάκων χωρίς στάθμιση είναι μια συνήθης πρακτική, καθώς μπορεί να είναι ευκολότερο να υπολογιστεί η βαθμολογία τους (score).

Σε αυτή τη διδακτορική διατριβή δημιουργήθηκαν προσομοιωμένα σενάρια 1.000 δειγμάτων, μεγέθους 10.000 έκαστο, που αποκάλυψαν σημαντική αύξηση του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC) των σταθμισμένων κλιμάκων έναντι του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών της μη σταθμισμένης κλίμακας. Το ίδιο συμπέρασμα λήφθηκε για τα υπόλοιπα αξιολογικά κριτήρια που χρησιμοποιήθηκαν από όλα τα εξεταζόμενα σενάρια. Οι σταθμισμένες κλίμακες, πέτυχαν τις υψηλότερες τιμές στα μέτρα "εμβαδόν υπό της καμπύλης λειτουργικών χαρακτηριστικών" (AUC), "αληθές ποσοστό" (TR), "αληθώς αρνητικό ποσοστό" (TNR), "θετική προγνωστική αξία" (PPV), "αληθώς θετικό ποσοστό" (TPR) και "αρνητική προγνωστική αξία" (NPV). Κατά την εφαρμογή της διαγνωστικής διαδικασίας σε γενικό ή ειδικό πληθυσμό, είναι επιθυμητή η αύξηση της ευαισθησίας ή της ειδικότητας, αναλόγως. Επομένως αυτό θα μπορούσε να επιτευχθεί με τη χρήση της κατάλληλης κλίμακας σε ανάλογες περιπτώσεις. Ένα γενικό αποτέλεσμα δηλώνει ότι όταν η απόσταση μεταξύ της μέσης τιμής του θεωρητικού πληθυσμού των ατόμων με ένα συγκεκριμένο χαρακτηριστικό και της μέσης τιμής του θεωρητικού πληθυσμού των ατόμων δίχως το συγκεκριμένο

χαρακτηριστικό είναι σχετικά μεγάλη τότε η διακριτική ικανότητα των προτεινόμενων σταθμισμένων κλιμάκων αυξάνεται. Σε όλα σχεδόν τα σενάρια που δοκιμάστηκαν, οι κλίμακες που παρήχθησαν με χρήση σταθμίσεων που ελήφθησαν από τη γραμμική διαχωριστική ανάλυση ήταν ακριβέστεροι στην ταξινόμηση ατόμων έναντι της μη σταθμισμένης, καθώς και έναντι άλλων σταθμισμένων κλιμάκων, ειδικά σε μεγάλα δείγματα. Σε μικρότερα δείγματα, οι κλίμακες που παράγονται από την δίτιμη λογιστική παλινδρόμηση ήταν κοντά στις κλίμακες από την γραμμική διαχωριστική ανάλυση, όσον αφορά την ακρίβειά τους και τις τιμές των αντίστοιχων κριτηρίων. Παρατηρήθηκε επίσης ότι όταν ο αριθμός των μεταβλητών συνιστωσών ήταν σχετικά μεγάλος, υπήρξε μια τάση των κλιμάκων που παράγονται από πολυπαραγοντικές διαδικασίες να έχουν μεγαλύτερη τιμή εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών, όπως η κλίμακα που παράγεται από την πολυμεταβλητή δίτιμη λογιστική παλινδρόμηση.

Η προβλεπτική ικανότητα των κλιμάκων δοκιμάστηκε με τη χρήση προσομοιωμένων δειγμάτων εκπαίδευσης και εφαρμογής. Συμπερασματικά, για μια σωστή πρόβλεψη, το δείγμα εκπαίδευσης πρέπει να συλλέγεται από τον πληθυσμό στόχο. Στην περίπτωση αυτή, οι παράμετροι του δείγματος εκπαίδευσης συγκλίνουν στις παραμέτρους του πληθυσμού. Επιπροσθέτως, τα αποτελέσματα προσομοίωσης επικυρώθηκαν κατά την εφαρμογή των προτεινόμενων σταθμισμένων κλιμάκων σε πραγματικά διατροφικά δεδομένα. Ειδικότερα, ανεξαρτήτως της εμπειρικής κατανομής της συχνότητας κατανάλωσης των ομάδων τροφίμων, οι σταθμισμένες κλίμακες πέτυχαν μεγαλύτερες τιμές στα προαναφερθέντα μέτρα.

Επομένως, για συνεχείς κλίμακες, η χρήση σταθμίσεων που προκύπτουν από μαθηματικές και στατιστικές μεθόδους, όπως η Διακρίνουσα Ανάλυση και η Λογιστική Παλινδρόμηση, αυξάνει τη διαγνωστική ικανότητα των κλιμάκων. Αυτό συμβαίνει λόγω του γεγονότος ότι αποδίδεται μεγαλύτερη στάθμιση σε εκείνες τις συνιστώσες μεταβλητές που διαθέτουν μεγαλύτερη διακριτική ικανότητα, βασιζόμενοι σε στοχαστικές και μαθηματικές μεθόδους (*Μπερσίμης, Παναγιωτάκος & Βαμβακάρη, 2015*). Συμπερασματικά, πρέπει να προταθεί η χρήση σταθμίσεων για την κατασκευή ενός εργαλείου το οποίο να είναι ακριβές για τη διάκριση των ατόμων με ή χωρίς κάποιο συγκεκριμένο χαρακτηριστικό.

4.1.4 Επίδραση Αύξησης Αριθμού Συνιστωσών Μεταβλητών στη Διαχωριστική και Προβλεπτική Ικανότητα Κλιμάκων

Η αύξηση του πλήθους *m* των συνιστωσών μεταβλητών, που σχετίζονται με το χαρακτηριστικό που η κλίμακα στοχεύει να ανιχνεύσει, οδήγησε στην αύξηση του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών ανεξάρτητα από το αν χρησιμοποιήθηκε στάθμιση ή όχι. Το ίδιο αποτέλεσμα παρατηρήθηκε ομοίως στην περίπτωση των υπόλοιπων κριτηρίων που χρησιμοποιήθηκαν (ευαισθησία, ειδικότητα, ακρίβεια, θετική και αρνητική προγνωστική αξία). Ο αντίστοιχος ρυθμός αύξησης των τιμών των προαναφερθέντων μέτρων μειωνόταν καθώς ο αριθμός *m* των συνιστωσών μεταβλητών αυξανόταν. Ο βαθμός συσχέτισης των συνιστωσών μεταβλητών, με τη μεταβλητή που εκφράζει το χαρακτηριστικό, πρέπει να είναι τουλάχιστον μέτρια. Τα προαναφερθέντα αποτελέσματα ήταν ισχυρά ακόμη και στην περίπτωση πρόβλεψης ατόμων σε νέους πληθυσμούς (test populations) με τη χρήση εκπαιδευτικών πληθυσμών (training populations).

Συγκεκριμένα, οι κλίμακες με μεγαλύτερο αριθμό συνιστωσών μεταβλητών πέτυχαν υψηλότερες τιμές του μέτρου του εμβαδού υπό της καμπύλης λειτουργικών χαρακτηριστικών (AUC), καθώς και τις τιμές των υπόλοιπων κριτηρίων σε σχέση με τις κλίμακες με λιγότερες συνιστώσες. Επομένως, συνιστάται η χρήση μεγαλύτερου αριθμού συνιστωσών κατά τον σχεδιασμό μιας κλίμακας, ενώ τα στοιχεία αυτά συσχετίζονται με το χαρακτηριστικό το οποίο η κλίμακα σκοπεύει να αξιολογήσει.

4.2 Περιορισμοί - Συνθήκες

Στην εργασία αυτή παρουσιάστηκε μια περίπτωση σύνθετης διακριτής κλίμακας, όπου η αύξηση του αριθμού των διαμερίσεων των συνιστωσών μεταβλητών του οδηγεί σε αύξηση της ευαισθησίας της. Το προαναφερθέν αποτέλεσμα παρουσιάζει πρακτικό ενδιαφέρον και εφαρμογή για οποιοδήποτε αριθμό συνιστωσών μεταβλητών *m* και αριθμό διαιρέσεων *k*. Όμως, η υπερβολική αύξηση του αριθμού *k* των διαμερίσεων μιας κλίμακας, μπορεί να προκαλέσει πρακτικές δυσκολίες κατά τη χρήση του καθώς μπορεί επίσης να προκαλέσει το σημαντικό πρόβλημα της εσφαλμένης ταξινόμησης (δυσταξινόμησης). Σε κάθε περίπτωση, η παρουσία εσφαλμένης ταξινόμησης δεν επηρεάζει την προαναφερθείσα σχέση μεταξύ του αριθμού των διαμερίσεων και την ευαισθησία της κλίμακας. Γενικά, το ζήτημα εσφαλμένης ταξινόμησης μπορεί να μειωθεί εάν ο ερευνητής χρησιμοποιεί ορθά την

εμπειρία του και επιλέγει τον βέλτιστο αριθμό k των διαμερίσεων σύμφωνα με την φύση του χαρακτηριστικού που η κλίμακα έχει σχεδιαστεί για να αξιολογήσει.

Επίσης, στην διατριβή αυτή παρουσιάστηκε μια περίπτωση σύνθετης συνεχούς κλίμακας, όπου η αύξηση του εύρους του στηρίγματος των συνιστωσών μεταβλητών της οδηγεί σε αύξηση της ευαισθησίας της. Είναι εύλογο να υποθέσουμε ότι μια απεριόριστη ανάπτυξη του μήκους του στηρίγματος (του πεδίου ορισμού) ίσως να μην είναι πραγματικά χρήσιμη σε ένα πραγματικό περιβάλλον, όσον αφορά την επαναληψιμότητα και την αναπαραγωγικότητα του εργαλείου. Παρ 'όλα αυτά, το πιθανό κόστος με την αύξηση του μήκους στήριξης μέσα σε ορισμένα όρια μετριάζεται με την επίτευξη μεγαλύτερης ακρίβειας, χωρίς οποιαδήποτε προβλήματα εσφαλμένης ταξινόμησης. Επομένως, θα πρέπει, σύμφωνα με τη φύση του εκτιμούμενου χαρακτηριστικού, να λαμβάνεται το μέγιστο δυνατό εύρος του στηρίγματος των συνιστωσών μεταβλητών. Στην περίπτωση της εφαρμογής της κανονικής κατανομής παρουσιάζεται ένα σημαντικό μειονέκτημα το οποίο συνίσταται στο ότι το στήριγμα της κλίμακας δεν οριοθετείται. Σε μια τέτοια περίπτωση, η χρήση της κανονικής κατανομής μπορεί να οδηγήσει στο γνωστό πρόβλημα της εσφαλμένης ταξινόμησης (δυσταξινόμησης) εξαιτίας της έλλειψης μιας ελάχιστης και μιας μέγιστης τιμής στην αντίστοιχη κλίμακα. Για το λόγο αυτό, θα μπορούσε να χρησιμοποιηθεί μια ειδική περικεκομένη μορφή της κανονικής προσέγγισης, με πεπερασμένο πεδίο ορισμού της μορφής $\left[\mu - k\cdot\sigma, \mu + k\cdot\sigma\right]$, $k \in \mathrm{N}$ με $k > 3$.

Επίσης, όπως αναφέρθηκε στο κεφάλαιο της εισαγωγής, η συνάρτηση ευαισθησίας μιας σύνθετης κλίμακας αποτελεί ένα μέτρο της προβλεπτικής ικανότητας αυτής και κρίθηκε κατάλληλη να χρησιμοποιηθεί για τη διερεύνηση των αρχικών ερευνητικών ερωτημάτων της παρούσας διατριβής. Είναι όμως γνωστό ότι από μόνη της η ευαισθησία δεν αρκεί για την αξιολόγηση της διαγνωστικής ακρίβειας, καθώς η αύξηση της ευαισθησίας μέσω της μετακίνησης του διαγνωστικού κατωφλιού, συνεπάγεται μείωση της ειδικότητάς του, και αντιστρόφως. Στην παρούσα διατριβή όμως δεν επηρεάζεται η τιμή του διαγνωστικού κατωφλιού, αλλά μόνο ο αριθμός των διαμερίσεων της εκάστοτε συνιστώσας.

4.3 Γενίκευση των Συμπερασμάτων

Τα συμπεράσματα αυτής της διδακτορικής διατριβής αναφέρουν ότι ο αριθμός των διαμερίσεων στην περίπτωση των διακριτών κλιμάκων συνίσταται να είναι ο μεγαλύτερος που μπορεί να χρησιμοποιηθεί (*Bersimis et al., 2013*), καθώς και στην περίπτωση των συνεχών κλιμάκων, το μήκος του αντίστοιχου στηρίγματος θα πρέπει να είναι ευρύ (*Bersimis et al., 2016*). Προς βελτίωση της ακρίβειας των σύνθετων κλιμάκων συνίσταται η χρήση σταθμίσεων στους συντελεστές των κλιμάκων ώστε να λαμβάνουν μεγαλύτερη βαρύτητα αυτές οι μεταβλητές που έχουν εντονότερη σχέση με το χαρακτηριστικό που η κλίμακα αποσκοπεί να προβλέψει (*Bersimis et al., 2017*). Επίσης, ο αριθμός των συνιστωσών μεταβλητών επηρεάζει την ακρίβεια, υπό την προϋπόθεση ότι αυτές οι μεταβλητές έχουν από μέτρια προς ισχυρή συσχέτιση με το χαρακτηριστικό που η κλίμακα αποσκοπεί να εκτιμήσει (*Bersimis et al., 2017*).

4.4 Χρησιμότητα στη Δημόσια Υγεία

Η χρήση μιας σύνθετης κλίμακας υγείας προσφέρει την δυνατότητα της πρώιμης διάγνωσης της νόσου, με αποτέλεσμα την έγκαιρη εφαρμογή της κατάλληλης θεραπείας στο πιθανώς νοσούν άτομο (*Mc Cullough et al., 2000*). Η ακρίβεια μιας διάγνωσης έγκειται στην ικανότητα της κλίμακας που χρησιμοποιείται να διακρίνει τα νοσούντα και υγιή άτομα. Συγκεκριμένα, σύμφωνα με την αξία της προαναφερθείσας κλίμακας, ένα άτομο ταξινομείται ως υγιές ή όχι. Συνεπώς, είναι επιθυμητό, η σύνθετη κλίμακα υγείας να ταξινομήσει σωστά κάθε άτομο ως υγιές ή όχι σύμφωνα με την πραγματική κλινική του κατάσταση. Η χρήση κλιμάκων υγείας που χαρακτηρίζονται από ακρίβεια μπορεί να οδηγήσει στην έγκαιρη διάγνωση μιας ασθένειας και με κατάλληλη θεραπεία να παραταθεί το προσδόκιμο ζωής και να βελτιωθεί η ποιότητα της ζωής ενός ασθενούς.

Τα ευρήματα που παρουσιάζονται σε αυτή τη διδακτορική διατριβή μπορούν να ασκήσουν ιδιαίτερη επίδραση στην εκτίμηση των συμπεριφορών που σχετίζονται με την υγεία και να ερευνήσουν τους παθοφυσιολογικούς μηχανισμούς μιας ασθένειας καλύτερα με την ανάπτυξη της ακρίβειας των κλιμάκων που περιγράφουν τα ανθρώπινα χαρακτηριστικά. Κατά συνέπεια, πιο αποτελεσματική δημόσια υγεία και άλλα κοινωνικά προγράμματα της πολιτείας μπορούν να διαμορφωθούν με μια καλύτερη διαχείριση των κρατικών πόρων στο τομέα της υγείας.

Τα προαναφερθέντα αποτελέσματα θα μπορούσαν να είναι χρήσιμα σε διάφορους τομείς της δημόσιας υγειονομικής περίθαλψης, όπου όλοι οι σχετικοί επιστήμονες προσπαθούν να αποκτήσουν ακριβείς κλίμακες υγείας, δηλαδή διαγνωστικά εργαλεία που μπορούν να διακρίνουν τα υγιή άτομα από τα μη υγιή. Ειδικότερα, ως επακόλουθο των διαπιστώσεων αυτής της διατριβής, τα ακριβή εργαλεία μέτρησης, δηλαδή οι κλίμακες που σχετίζονται με την υγεία, μπορούν να βοηθήσουν στον έγκαιρο εντοπισμό ενός ασθενούς και συνεπώς στην έναρξη εφαρμογής θεραπευτικών μέσων που μπορούν να παρατείνουν τη ζωή των ασθενών και να βελτιώσουν την ποιότητά της. Επιπλέον, τα ακριβή διαγνωστικά εργαλεία ελέγχου της υγείας των ατόμων ενός πληθυσμού μπορούν να βοηθήσουν στην οργάνωση κατάλληλων προγραμμάτων δημόσιας υγείας για την αξιόπιστη πρόληψη των διαφόρων ασθενειών και να ανταποκριθούν στις πιο πιεστικές ανάγκες της σύγχρονης εποχής για βασικές υπηρεσίες υγείας. Οι ακριβείς κλίμακες που σχετίζονται με την υγεία μπορεί να είναι χρήσιμοι για την ορθολογιστική χρήση των δημόσιων οικονομικών πόρων στον τομέα της υγείας με τη χρησιμοποίηση λιγότερων βασικών ιατρικών προμηθειών και με τον τρόπο αυτό να ελαχιστοποιηθούν τα ιατρικά έξοδα όσον αφορά τις επαναλαμβανόμενες κλινικές εξετάσεις λόγω εσφαλμένης διάγνωσης.

Κατά συνέπεια, τα συμπεράσματα αυτής της διατριβής είναι σημαντικά, καθώς προσφέρουν μεθοδολογίες όσον αφορά την κατασκευή εγκυρότερων και αποτελεσματικότερων σύνθετων κλιμάκων υγείας, προκειμένου να δημιουργηθούν ακριβέστερες κλίμακες για την πρόβλεψη της κλινικής κατάστασης ενός ατόμου.

4.5 Μελλοντική Έρευνα

Η χρήση κλιμάκων στις επιστήμες έχει καθιερωθεί πολλές δεκαετίες αλλά απαιτείται περαιτέρω έρευνα στον τομέα μεθοδολογίας κατασκευής τους, καθώς η θεωρητική προσέγγιση του εύρους του στηρίγματος των συνιστωσών μεταβλητών ή στάθμισης των κλιμάκων δεν έχει μελετηθεί επαρκώς.

Στο μέλλον, ένα στοιχείο που αφορά την κατασκευή των κλιμάκων και πρέπει να ερευνηθεί, θεωρητικά, είναι το αν υπάρχει ένας κατάλληλος αριθμός των διαμερίσεων πέρα του οποίου, η αύξηση της ακρίβειας (δηλαδή ευαισθησίας) της κλίμακας, είναι πολύ μικρή ή αμελητέα. Επίσης, σε μελλοντική έρευνα, θα μπορούσε

να εξεταστεί η συμπεριφορά της ειδικότητας της κλίμακας, καθώς ο αριθμός των διαμερίσεων αυξάνεται.

Ένα άλλο ζήτημα που πρέπει να διερευνηθεί θεωρητικά είναι ότι επειδή ορισμένες συνιστώσες της κλίμακας συσχετίζονται με το χαρακτηριστικό που εκτιμά η κλίμακα πιο δυναμικά από κάποιες άλλες, θα μπορούσαν να δοκιμαστούν διαφορετικές μέθοδοι στάθμισης, εκτός από τις προτεινόμενες σε αυτή τη διατριβή (*Kranz et al. 2006*).

ΞΕΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ (MLA)

Aitken, R. C. B. "A Growing Edge of Measurement of Feelings [Abridged] Measurement of Feelings Using Visual Analogue Scales." (1969): 989-993.

Altman, Douglas G., and J. Martin Bland. "Diagnostic tests. 1: Sensitivity and specificity." *BMJ: British Medical Journal* 308.6943 (1994): 1552.

Andrew P. Worth, Mark T.D. Cronin, The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects, Journal of Molecular Structure (Theochem) 622 (2003) 97–111

Bach, Anna, et al. "The use of indexes evaluating the adherence to the Mediterranean diet in epidemiological studies: a review." *Public health nutrition*9.1a (2006): 132-146.

Bansal, Aasthaa, and Margaret Sullivan Pepe. "When does combining markers improve classification performance and what are implications for practice?." *Statistics in medicine* 32.11 (2013): 1877-1892.

Baron, John A. "Too Bad It Isn't True." *Medical Decision Making* 14.2 (1994): 107-107.

Beck, A., et al. "An inventory for measuring depression Arch Gen Psychiatry 4: 561–571." *Find this article online* (1961).

Bersimis, Fragkiskos, Demosthenes Panagiotakos, and Malvina Vamvakari. "Sensitivity of Health Related Indices is a Non-Decreasing Function of their Partitions." *Journal of Statistics Applications & Probability* 2.3 (2013): 183.

Bersimis, Fragkiskos, Demosthenes Panagiotakos, and Malvina Vamvakari. "Investigating the sensitivity function's monotony of a health-related index." *Journal of Applied Statistics* 44.9 (2016): 1680-1706.

Bersimis, Fragkiskos, Demosthenes Panagiotakos, and Malvina Vamvakari. " The use of components’ weights improves the diagnostic accuracy of a health-related index." *Communication in Statistics: Theory and Methods*. (2017): 1-24.

Bersimis G. Fragiskos, Lygiaris M. Michalis, André Thomas. "Learn Math by Playing» An Alternative, Game-Based Teaching Proposal". 3rd International Conference for the Promotion of Educational Innovationprocedings 2017.

Bradley, David M., and Ramesh C. Gupta. "On the distribution of the sum of n non-identically distributed uniform random variables." *Annals of the Institute of Statistical Mathematics* 54.3 (2002): 689-700.

Brandt, Allan M., and Martha Gardner. "Antagonism and accommodation: interpreting the relationship between public health and medicine in the United States during the 20th century." *American journal of public health* 90.5 (2000): 707.

Buonocore, Aniello, Enrica Pirozzi, and Luigia Caputo. "A note on the sum of uniform random variables." *Statistics & Probability Letters* 79.19 (2009): 2092-2097.

Carlsson, Anna Maria. "Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale." *Pain* 16.1 (1983): 87-101.

Charalambides, Charalambos A. "On the enumeration of certain compositions and related sequences of numbers." *FIBONACCI QUARTERLY* 20.2 (1982): 132-146.

Charalambides, Charalambos A. *Enumerative combinatorics*. CRC Press, 2002.

Charalambides, Charalambos A. *Combinatorial methods in discrete distributions*. Vol. 600. John Wiley & Sons, 2005.

Cleeland, C. S. and K. M. Ryan (1994). "Pain assessment: global use of the Brief Pain Inventory." Ann Acad Med Singapore **23**(2): 129-38.

Cox, David R. "The regression analysis of binary sequences." *Journal of the Royal Statistical Society. Series B (Methodological)* (1958): 215-242.

D'Agostino, R. B., Sr., R. S. Vasan, et al. (2008). "General cardiovascular risk profile for use in primary care: the Framingham Heart Study." *Circulation* **117** (6): 743-53.

Daniel, Wayne W., and Chad Lee Cross. "*Biostatistics: a foundation for analysis in the health sciences*." (1995).

Eckhardt, Roger. "Stan ulam, john von neumann, and the monte carlo method." *Los Alamos Science* 15.131-136 (1987): 30.

Bootstrap, Efron B. "methods: Another look at the Jackknife [J]." *Ann. Statist* 7.1 (1979): 1-26.

Efron, Bradley, and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Efron, Bradley. "Second thoughts on the bootstrap." *Statistical Science* 18.2 (2003): 135-140.

Egan, James P. "Signal detection theory and {ROC} analysis." (1975).

Elliott, Jane, and Peter Shepherd. "Cohort profile: 1970 British birth cohort (BCS70)." International Journal of Epidemiology 35.4 (2006): 836-843.

Ewens, Warren J., and Gregory R. Grant. *Statistical methods in bioinformatics: an introduction*. Springer Science & Business Media, 2006.

Fan, Jianqing, and Irene Gijbels. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Vol. 66. CRC Press, 1996.

Fawcett, Tom. "An introduction to ROC analysis." *Pattern recognition letters* 27.8 (2006): 861-874.

Fein, Louis. "The role of the university in computers, data processing, and related fields." *Communications of the ACM* 2.9 (1959): 7-14.

Feller, William. *Introduction to Probality Theory and Its Applications*. Wiley Eastern, 1966.

Feller, William. *An introduction to probability theory and its applications: volume I*. Vol. 3. New York: John Wiley & Sons, 1968.

Ferrans, C. A. R. O. L., and M. Powers. "Quality of life index." (2010).

Feskanich, D., H. R. Rockett, et al. (2004). "Modifying the Healthy Eating Index to assess diet quality in children and adolescents." *J Am Diet Assoc* **104**(9): 1375-83.

Fielding, Alan H., and John F. Bell. "A review of methods for the assessment of prediction errors in conservation presence/absence models." *Environmental conservation* 24.1 (1997): 38-49..

Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." *Annals of human genetics* 7.2 (1936): 179-188.

Fletcher, Robert H., Suzanne W. Fletcher, and Grant S. Fletcher. *Clinical epidemiology: the essentials*. Lippincott Williams & Wilkins, 2012.

Fountoulakis, K. N., M. Papadopoulou, et al. (2006). "Reliability and psychometric properties of the Greek translation of the State-Trait Anxiety Inventory form Y: preliminary data." *Ann Gen Psychiatry* **5**: 2.

Gavrilov, Leonid A., and Natalia S. Gavrilova. "Demographic consequences of defeating aging." Rejuvenation research 13.2-3 (2010): 329-334.

Gerber, M. J., J. D. Scali, et al. (2000). "Profiles of a healthful diet and its relationship to biomarkers in a population sample from Mediterranean southern France." J Am Diet Assoc **100**(10): 1164-71.

Greenland, Sander, and Hal Morgenstern. "Confounding in health research." *Annual review of public health* 22.1 (2001): 189-212.

Greenes, Robert A., and Edward H. Shortliffe. "Medical informatics: an emerging academic discipline and institutional priority." *Jama* 263.8 (1990): 1114-1120.

Haines, P. S., A. M. Siega-Riz, et al. (1999). "The Diet Quality Index revised: a measurement instrument for populations*." J Am Diet Assoc* ***99****(6): 697-704.*

Hamilton, Max. "A rating scale for depression." *Journal of neurology, neurosurgery, and psychiatry* 23.1 (1960): 56.

Hanley, James A., and Barbara J. McNeil. "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology* 143.1 (1982): 29-36.

Harnack, Lisa, et al. "An evaluation of the Dietary Guidelines for Americans in relation to cancer occurrence." *The American journal of clinical nutrition* 76.4 (2002): 889-896.

Hawker, Gillian A., et al. "Measures of adult pain: Visual analog scale for pain (vas pain), numeric rating scale for pain (nrs pain), mcgill pain questionnaire (mpq), short form mcgill pain questionnaire (sf- mpq), chronic pain grade scale (cpgs), short form 36 bodily pain scale (sf- 36 bps), and measure of intermittent and constant osteoarthritis pain (icoap)." *Arthritis care & research* 63.S11 (2011).

Hayes, M. H. "Experimental developement of the graphics rating method." *Psychol Bull* 18 (1921): 98-99.

Henrard, Séverine, Niko Speybroeck, and Cédric Hermans. "Classification and regression tree analysis vs. multivariable linear and logistic regression methods as statistical tools for studying haemophilia." *Haemophilia* 21.6 (2015): 715-722.

Hogeweg, Paulien. "The roots of bioinformatics in theoretical biology." PLoS computational biology 7.3 (2011): e1002021

Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

Huskisson, E. C. "Measurement of pain." *The lancet* 304.7889 (1974): 1127-1131.

Indrayan, Abhaya. *Medical biostatistics*. CRC Press, 2012.

Jackson, Douglas N. "A sequential system for personality scale development." *Current topics in clinical and community psychology* 2 (1970): 61-96.

Joyce, Adam P., et al. "Structure-based modeling of protein: DNA specificity." *Briefings in functional genomics* 14.1 (2014): 39-49.

Janitza, Silke, Harald Binder, and Anne Laure Boulesteix. "Pitfalls of hypothesis tests and model selection on bootstrap samples: causes and consequences in biometrical applications." *Biometrical Journal* 58.3 (2016): 447-473.

Kachigan, Sam Kash. *Multivariate statistical analysis: A conceptual introduction*. Radius Press, 1991.

Kant, Ashima K. "Indexes of overall diet quality: a review." *Journal of the American Dietetic Association* 96.8 (1996): 785-791.

Karlis, D. (2005). Multivariate statistical analysis. Stamoulis Publications.

Kattan, M. W., J. A. Eastham, et al. (1998). "A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer." *J Natl Cancer Inst* **90**(10): 766-71.

Kennedy, T. E. I. L. E. E. N., et al. "The healthy eating index: design and applications." *Journal of the American Dietetic Association* 95.10 (1995): 1103-1108.

Knaus, W. A., J. E. Zimmerman, et al. (1981). "APACHE-acute physiology and chronic health evaluation: a physiologically based classification system." *Crit Care Med* **9**(8): 591-7.

Killmann, Frank, and Elart von Collani. "A note on the convolution of the uniform and related distributions and their use in quality control." *Economic Quality Control* 16.1 (2001): 17-41.

Kourlaba, Georgia, and Demosthenes B. Panagiotakos. "Dietary quality indices and human health: a review." *Maturitas* 62.1 (2009a): 1-8.

Kourlaba, Georgia, and Demosthenes Panagiotakos. "The number of index components affects the diagnostic accuracy of a diet quality index: the role of intracorrelation and intercorrelation structure of the components." *Annals of epidemiology* 19.10 (2009b): 692-700.

Kourlaba, Georgia, and Demosthenes Panagiotakos. "The diagnostic accuracy of a composite index increases as the number of partitions of the components increases and when specific weights are assigned to each component." *Journal of Applied Statistics* 37.4 (2010): 537-554.

Kranz, Sibylle, et al. "A diet quality index for American preschoolers based on current dietary intake recommendations and an indicator of energy balance." *Journal of the American Dietetic Association* 106.10 (2006): 1594-1604.

Kroese, Dirk P., et al. "Why the Monte Carlo method is so important today." *Wiley Interdisciplinary Reviews: Computational Statistics* 6.6 (2014): 386-392.

Kruppa, Jochen, et al. "Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory." *Biometrical Journal* 56.4 (2014): 534-563.

Le Gall, J. R., S. Lemeshow, et al. (1993). "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicentre study." *Jama* **270**(24): 2957-63.

Likert, R. A. (1952). "A technique for the development of attitude scales." Educ Psychol Measur **12**: 313-315.

Maroulas, Vasileios, and Demosthenes B. Panagiotakos. "Sensitivity of health-related scales is a non-decreasing function of their classes." *arXiv preprint arXiv:1005.3737* (2010).

Martinez-Gonzalez, M. A., E. Fernandez-Jarne, et al. (2002). "Mediterranean diet and reduction in the risk of a first acute myocardial infarction: an operational healthy dietary score." Eur J Nutr **41**(4): 153-60.

Martinez-Gonzalez, M. A., E. Fernandez-Jarne, et al. (2004). "Development of a short dietary intake questionnaire for the quantitative estimation of adherence to a cardioprotective Mediterranean diet." Eur J Clin Nutr **58**(11): 1550-2.

McCulloch, Warren; Walter Pitts. "A Logical Calculus of Ideas Immanent in Nervous Activity*". Bulletin of Mathematical Biophysics. 5 (4):* (1943) *115 –133*.

McCullough, Marjorie L., et al. "Adherence to the Dietary Guidelines for Americans and risk of major chronic disease in men." *The American journal of clinical nutrition* 72.5 (2000): 1223-1231.

McCullough, Marjorie L., et al. "Adherence to the Dietary Guidelines for Americans and risk of major chronic disease in women." *The American journal of clinical nutrition* 72.5 (2000): 1214-1222.

McDowell, I. "“Health Measurement Scales.” Encyclopedia of Public Health." (2002).

McDowell, Ian. *Measuring health: a guide to rating scales and questionnaires*. Oxford university press, 2006.

McIntosh, Martin W., and Margaret Sullivan Pepe. "Combining several screening tests: optimality of the risk score." *Biometrics* 58.3 (2002): 657-664.

Metropolis, Nicholas, and Stanislaw Ulam. "The monte carlo method." *Journal of the American statistical association* 44.247 (1949): 335-341.

Moody, Glyn. *Digital code of life: how bioinformatics is revolutionizing science, medicine, and business*. John Wiley & Sons, 2004.

Parikh, Rajul et al. “Understanding and Using Sensitivity, Specificity and Predictive Values.” Indian Journal of Ophthalmology 56.1 (2008): 45–50. Print.

Panagiotakos, Demosthenes B., et al. "Adherence to the Mediterranean food pattern predicts the prevalence of hypertension, hypercholesterolemia, diabetes and obesity, among healthy adults; the accuracy of the MedDietScore." *Preventive medicine* 44.4 (2007): 335-340.

Panagiotakos, Demosthenes. "Health measurement scales: methodological issues." *The open cardiovascular medicine journal* 3 (2009): 160-165.

Pappa, E., N. Kontodimopoulos, et al. (2005). "Validating and norming of the Greek SF-36 Health Survey." Qual Life Res **14**(5): 1433-8.

Patterson, Ruth E., Pamela S. Haines, and Barry M. Popkin. "Diet quality index: capturing a multidimensional behavior." *Journal of the American Dietetic Association* 94.1 (1994): 57-64.

Pepe, Margaret Sullivan, and Mary Lou Thompson. "Combining diagnostic test results to increase accuracy." *Biostatistics* 1.2 (2000): 123-140.

Pepe, Margaret Sullivan. *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

Pepe, Margaret Sullivan, Tianxi Cai, and Gary Longton. "Combining predictors for classification using the area under the receiver operating characteristic curve." *Biometrics* 62.1 (2006): 221-229

Pfeiffer, Ruth M., and Efstathia Bur. "A model free approach to combining biomarkers." *Biometrical journal* 50.4 (2008): 558-570.

Pitsavos, Christos, et al. "Epidemiology of cardiovascular risk factors in Greece: aims, design and baseline characteristics of the ATTICA study." *BMC public health* 3.1 (2003): 32.

Pohar, Maja, Mateja Blas, and Sandra Turk. "Comparison of logistic regression and linear discriminant analysis: a simulation study." Metodoloski zvezki 1.1 (2004): 143.

Potuschak, Heinrich, and Werner G. Müller. "More on the distribution of the sum of uniform random variables." *Statistical papers* 50.1 (2009): 177-183.

Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

Qin, Jing, and Biao Zhang. "Best combination of multiple diagnostic tests for screening purposes." *Statistics in medicine* 29.28 (2010): 2905-2919.

Radloff, Lenore Sawyer. "The CES-D scale: A self-report depression scale for research in the general population." Applied psychological measurement 1.3 (1977): 385-401.

Rapaport, William J. "Philosophy of computer science: An introductory course." *Teaching Philosophy* 28.4 (2005): 319-341.

Rydén, Lars, et al. "ESC guidelines on diabetes, pre-diabetes, and cardiovascular diseases developed in collaboration with the EASD-summary." *Diabetes &* vascular disease research 11.3 (2014): 133-73.

Sadooghi-Alvandi, S. M., A. R. Nematollahi, and R. Habibi. "On the distribution of the sum of independent uniform random variables." Statistical Papers 50.1 (2009): 171.

Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." Neural networks 61 (2015): 85-117

Schweitzer, Sylvie. "Lieux et acteurs de l'industrialisation, XVIIIe-XXe siècles. Centre Pierre Léon d'histoire économique et sociale UMR 5599 du CNRS." *Cahiers d'histoire* 44-1 (1999).

Sheldon M. Ross, Simulation (4th edition), Simulation, 2006, Elsevier Academic Press

Sim, Adelene Y. L., Peter Minary, and Michael Levitt. “Modeling Nucleic Acids.” *Current opinion in structural biology* 22.3 (2012): 273–278. *PMC*. Web. 31 July 2017

Simmons, Paul L., and Robert F. Simmons. "The simulation of cognitive processes: an annotated bibliography." *IRE Transactions on Electronic Computers* 3 (1961): 462-483.

Spielberger, Charles D. *State Trait anxiety inventory*. John Wiley & Sons, Inc., 2010.

Spitzer, W. O., A. J. Dobson, et al. (1981). "Measuring the quality of life of cancer patients: a concise QL-index for use by physicians." *J Chronic Dis* **34**(12): 585-97.

Stavrinos, V., & Panagiotakos, D. (2007). Biostatistics. Gudenberg Publications

Streiner, David L., and Geoffrey R. Norman. "Health measurement scales: a practical guide to their development and use." *Health measurement scales: a practical guide to their development and use*. Oxford University Press, 2008.

Streiner, David L., Geoffrey R. Norman, and John Cairney. *Health measurement scales: a practical guide to their development and use. Oxford University Press, USA, 2015.*

Styliaras, Georgios, *et al. "Teaching of Informatics*." (2015).

Su, John Q., and Jun S. Liu. *"Linear combinations of multiple diagnostic markers."* Journal of the American Statistical Association 88.424 (1993): 1350-1355.

Taylor, Janet A. "A personality scale of manifest anxiety." *The Journal of abnormal and social psychology* 48.2 (1953): 285.

Theil, Henri. "Economic Forecasts and Policy (2nd revised edn.) North Holland Publishing Co." *Amsterdam.[Links]* (1961).

Trichopoulou, A., T. Costacou, et al. (2003). "Adherence to a Mediterranean diet and survival in a Greek population." *N Engl J Med* **348**(26): 2599-608.

Tucker, Allen B., ed. *Computer science handbook*. CRC press, 2004.

Wang, T. J., J. M. Massaro, et al. (2003). "A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study." Jama **290**(8): 1049-56.

Ware JE and Sherbourne CD (1992). "The MOS 36- Item Short-Form Health Survey (SF-36). Conceptual framework and item selection." Med Care **30**: Med Care.

Wayne, W.D. (1995). Biostatistics: a foundation for analysis in the health sciences. Wiley, New York.

Waijers, Patricia MCM, Edith JM Feskens, and Marga C. Ocké. "A critical review of predefined diet quality scores." *British Journal of Nutrition* 97.2 (2007): 219-231.

Weisstein, Eric W. "Uniform distribution." *moment* 6 (2003): 7.

Wilson, P. W., R. B. D'Agostino, et al. (1998). "Prediction of coronary heart disease using risk factor categories." *Circulation* **97**(18): 1837-47.

Wong, K.C. (2016). Computational Biology and Bioinformatics: Gene Regulation. CRC Press, Taylor & Francis Group.

Worth, Andrew P., and Mark TD Cronin. "The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects." *Journal of Molecular Structure: THEOCHEM* 622.1 (2003): 97-111.

Yiannakouris, N., A. Trichopoulou, et al. (2006). "A direct assessment of genetic contribution to the incidence of coronary infarct in the general population Greek EPIC cohort." *Eur J Epidemiol* **21**(12): 859-67.

Youden, William J. "Index for rating diagnostic tests." *Cancer* 3.1 (1950): 32-35.

Yuan, Zheng, and Debashis Ghosh. "Combining multiple biomarker models in logistic regression." *Biometrics* 64.2 (2008): 431-439.

Zembylas, Michalinos, and Charalambos Vrasidas. "Globalization, information and communication technologies, and the prospect of a 'global village': promises of inclusion or electronic colonization?." *Journal of Curriculum Studies* 37.1 (2005): 65-83.

Zhang, Fang F., et al. "Evolution of epidemiologic methods and concepts in selected textbooks of the 20th century." *A History of Epidemiologic Methods and Concepts*. Birkhäuser Basel, 2004. 351-362.

Zung, William WK. "A self-rating depression scale." *Archives of general psychiatry* 12.1 (1965): 63-70.

ΕΛΛΗΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

Αγγελίδης, Π., 2015. Ηλεκτρονική Υγεία. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/5994

Αλεξάνδρου Ε., (2007). *Η ανάπτυξη ενός συστήματος CRM για ένα Πληροφοριακό Σύστημα σε μια βιομηχανική επιχείρηση.* Μεταπτυχιακή Εργασία, http://dione.lib.unipi.gr/xmlui/handle/unipi/1359.

Ανδρέου, Α. (1999). Θέματα Οργάνωσης και Διοίκησης της Εκπαίδευσης και της Σχολικής Μονάδας. Αθήνα: Νέα Σύνορα-Α. Λιβάνη.

Αντωνίου Β., (2003). *Ελλειψη υβριδικών στελεχών στην σύγχρονη ελληνική επιχείρηση: δεξιότητες (skills) και ικανότητες (competencies) που πρέπει να διαθέτει ένα στέλεχος του τμήματος πληροφοριακής τεχνολογίας και το ρόλο που έχει στα πλαίσια μιας σύγχρονης επιχείρησης..* Μεταπτυχιακή Εργασία, http://dione.lib.unipi.gr/xmlui/handle/unipi/376.

Δαμιανού Χ., (1999). Μεθοδολογια Δειγματοληψιας, Τεχνικες και Εφαρμογες. Αίθρα

Δημητρόπουλος, Ε. (2002). Εκπαιδευτική αξιολόγηση. Η αξιολόγηση της εκπαίδευσης και του εκπαιδευτικού έργου. Αθήνα: Γρηγόρης.

Εμίρης Δ.Μ., Κουλουριώτης Δ.Ε., Αντωνιδάκης Ε.Ν. και Καλιακάτσος Ι.Ν. «Βελτίωση της Σχεδιαστικής Αποτελεσματικότητας βάσει Κινηματικής Αξιολόγησης του Ρομποτικού Χώρου Εργασίας- Μέρος Α: Δείκτες και Κριτήρια», 4ο Συνέδριο Αυτοματισμού και Τεχνολογίας, Σύλλογος Επιστημόνων Αυτοματισμού Ελλάδας, 5-6 Οκτωβρίου 2002, Θεσσαλονίκη, Πρακτικά σελ. 203-210, Εκδ. Τζιόλα, 2002. (ISBN: 960-8050-98-7)

Ένωση Πληροφορικών Ελλάδος, *Μελετη επισκοπησης της πληροφορικης στην Ελλαδα* (2006). https://www.epe.org.gr/meleth/final/MEP2006.pdf

Ζαχαροπούλου Χρυσούλα, *Στατιστική, Μέθοδοι – Εφαρμογές, Τόμος Α'*, 2010

Κοίλιας Χρ., Στρ. Καλαφατούδης, Ευαγ. Μπανδήλα. *Χρήση Υπολογιστή*, Εκδόσεις νέων Τεχνολογιών, 2004

Κουτούζης, Μ. – Χατζηευστρατίου, Ι., «Αξιολόγηση στην εκπαιδευτική μονάδα», στο Αθανασούλα-Ρέππα, Α. – Κουτούζη, Μ. – Χατζηευστρατίου, Ι. (1999). Κοινωνική και ευρωπαϊκή διάσταση της εκπαιδευτικής πράξης. Πάτρα: ΕΑΠ.

Κούτρας Μ., *Εισαγωγή στις Πιθανότητες*, Εκδόσεις Σταμούλη, 2005

Κυριακούσης Α., *Στατιστικές Μέθοδοι. Χαροκόπειο* Πανεπιστήμιο (2000)

Μαστρακούλης Γ., (2008). *Σύγχρονα τραπεζικά προϊόντα και η λογιστική και πληροφοριακή τους διαχείριση από τις επιχειρήσεις.* Μεταπτυχιακή Εργασία, https://dspace.lib.uom.gr/bitstream/2159/3780/2/MastrakoulisMsc2008.pdf.

Μπάγκος Π., *Βιοπληροφορική*. Πανεπιστήμιο Θεσσαλίας (2015)

Μπαλής Χ., (2007). *Σύστημα διαχείρισης γνώσης σε οργανισμούς φροντίδας υγείας.* Διδακτορική Διατριβή, 10.12681/eadd/18439.

Μπερσίμης Φ., (2016). *Βελτιωση διαγνωστικης ικανοτητας δεικτων υγειας με χρηση μαθηματικων & στατιστικων μεθοδων*. Πρακτικά 33ου Πανελληνίου Συνεδρίου Μαθηματικής Παιδείας, σελ.714-724.

Μπερσίμης Φ., Παναγιωτάκος Δ., Βαμβακάρη Μ. (2012). *Η συνάρτηση ευαισθησίας ενός σύνθετου δείκτη υγείας είναι αύξουσα ως προς τον αριθμό των διαμερισεων των συνιστωσών του*. Πρακτικά 25ου Πανελληνίου Συνεδρίου Στατιστικής, σελ.141-150.

Μπερσίμης Φ., Παναγιωτάκος Δ., Βαμβακάρη Μ. (2015). *Η χρήση ειδικών σταθμίσεων στις συνιστώσες ενός σύνθετου δείκτη υγείας αυξάνει τη διαγνωστική του ικανότητα.* Πρακτικά 28ου Πανελληνίου Συνεδρίου Στατιστικής, σελ.237-251.

Νάκας, Χρήστος. *Προσαρμογή καμπύλης, στατιστική συμπερασματολογία, επεκτάσεις και εφαρμογές στην ανάλυση των καμπυλών λειτουργικού χαρακτηριστικού δέκτη (ROC)*. Diss. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ). Σχολή Θετικών Επιστημών. Τμήμα Μαθηματικών, 2002.

Παιδαγωγικό Ινστιτούτο (2000). Δείκτες ποιότητας εκπαιδευτικού έργου, τεύχος Β’, Αθήνα.

Παιδαγωγικό Ινστιτούτο (2000). Διαδικασίες εσωτερικής αξιολόγησης και σχεδιασμού του εκπαιδευτικού έργου στη σχολική μονάδα, τεύχος Γ’, Αθήνα.

Παιδαγωγικό Ινστιτούτο (2000). Εισαγωγή στην αξιολόγηση της ποιότητας του εκπαιδευτικού έργου. Βασικές έννοιες και παραδοχές, τεύχος Α’, Αθήνα.

Παναγιωτάκος, Δ. & Σταυρινός, Γ. (2007). *Βιοστατιστική.* Αθήνα: Gutenberg

Παπαδηματρακόπουλος, Β. (2006). «Για μια εσωτερική αξιολόγηση της σχολικής μονάδας» στο Επιστημονικό Βήμα, τεύχος 6, σ. 43-59.

Παπαϊωάννου, Β., 2011. *Σύγχρονα τραπεζικά προϊόντα και η λογιστική και πληροφορική τους διαχείριση από τις επιχειρήσεις*. Πτυχιακή εργασία. Πρέβεζα. Τ.Ε.Ι. Ηπείρου. Σχολή Διοίκησης και Οικονομίας. Τμήμα Λογιστικής.

ΣΕΒ, Σύγχρονες Επιχειρήσεις, Σύγχρονη Ελλάδα, Τεχνολογία & Καινοτομία Τεχνολογίες Υγείας Ενημερωτικό Δελτίο 2012

Σολομών, Ι. (1999). Εσωτερική αξιολόγηση και προγραμματισμός του εκπαιδευτικού έργου στη σχολική μονάδα. Αθήνα: Παιδαγωγικό Ινστιτούτο.

Τσαγρής Μιχαήλ, *Στατιστική με τη χρήση του IBM SPSS 22, STATISTICAL PACKAGE for the SOCIAL SCIENCES* , ΑΘΗΝΑ και Nottingham, Μάρτιος 2014

Φράγκος Χρήστος Κ. *Στατιστική Επιχειρήσεων για τις Οικονομικές και Κοινωνικές Επιστήμες,* Εκδόσεις Α. Σταμούλη, Αθήνα 1998

Χαλικιάς Γ. Ιωάννης, *Στατιστική- Μέθοδοι ανάλυσης για επιχειρηματικές αποφάσεις,* Rosili, Αθήνα, 2010

Χρήστος Κων. Φράγκος, *Στατιστική ΙΙ – Βασικές Μέθοδοι Πολυμεταβλητής Στατιστικής Ανάλυσης, Λύση Παραδειγμάτων Με το Στατιστικό Πακέτο PASW Statistics 18.0* (όπως μετανομάστηκε το SPSS), Εκδόσεις Νέων Τεχνολογιών, 2011

Ιστιακές Πηγές – Αναφορές

https://sites.google.com/site/hplhroforikhsthnekpedeysh/ti-einai-e-plerophorike

https://www.geogebra.org/?lang=el

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα Α1: Αλγόριθμοι σε περιβάλλον MatLab

Algorithm 1: Data Generation & Indices' Evaluation

```
%%% This algorithm corresponds to paragraph 4.1. "Results from simulated
data - Classification evaluation."%%%
%%% Figures 1-16, as well as tables 2-3 are produced by using simulation
code %%%
%%% This algorithm with core i7 pc takes approximatelly 3 hours to be
%%% accomplshed - You could get less iterations to reduce time e.g.
ixmax=100 or ixmax=500%%%
%%%%% Data Construction %%%%%
clear all % erase previous data & results
format long % 16 Decimals
ixmax=10; % "ix" corresponds to the number of repeated experimenets
jflag=1; % jflag=1 indicates Data standardization
pflag=1; % pflag=1 indicates Data conversion from data set X to data set
from PCA (Varimax Method - Regression Scores)
for ix=1:ixmax % Iterations' loop
    i_iter=ix
    tic % Time counter
    n1=7500; % Number of non diseased cases
    n2=2500; % Number of diseased cases
    kmax=5; % Number of m component variables
    iflag=1; % Determination of distribution - iflag=1 means Normal -
iflag=2 means Uniform
    X=zeros(n1+n2, kmax); % X matrix Dimensions (n1+n2)*kmax
    Y=zeros(n1+n2, 1); % Y matrix Dimensions (n1+n2)*1)
    Y1=zeros(n1,1); % Healthy cases - Value zero (0)
    Y2=ones(n2,1); % Diseased cases - Value ones (1)
    Y=[Y1;Y2];  % Vector Y - Dependent Variable
    if (iflag==1) % Normal distribution
        mu_h=[110 108 106 104 102 100]; % Mean values for healthy cases
        mu_d=[120 122 124 126 128 130]; % Mean values for diseased cases
        sigma_h=[30 28 26 24 22 20]; % Standard deviation for healthy cases
        sigma_d=[30 28 26 24 22 20]; % Standard deviation for diseased cases
        for k=1:kmax % Loop for the m components
            X(1:n1,k)=random('Normal',mu_h(k),sigma_h(k),n1,1); % Matrix X
is filled up with data for healthy cases
            X(n1+1:n1+n2,k)=random('Normal',mu_d(k),sigma_d(k),n2,1); %
Matrix X is filled up with data for diseased cases
        end
    else if (iflag==2) % Uniform distribution (continuous)
            a_h=[50 48 46 44 42 40]; % Lower bound for healthy cases
            a_d=[70 72 74 76 78 80]; % lower bound for diseased cases
            b_h=[150 148 146 144 142 140]; % Upper bound for healthy cases
            b_d=[170 172 174 176 178 180]; % Upper bound for diseased cases
            for k=1:kmax % Loop for all the m components
                X(1:n1,k)=random('unif',a_h(k),b_h(k),n1,1); % Matrix X is
filled up with data for healthy cases
                X(n1+1:n1+n2,k)=random('unif',a_d(k),b_d(k),n2,1); % Matrix
X is filled up with data for diseased cases
            end
        else  if (iflag==3) % Uniform distribution (discrete)
                a_d_h=[1 1 1]; % Distribution lower bound in case of
discrete uniform distribution for healthy individuals
                a_d_d=[30 40 4]; % Distribution lower bound in case of
discrete uniform distribution for diseased individuals
                b_d_h=[80 60 9]; % Distribution upper bound in case of
discrete uniform distribution for healthy individuals
                b_d_d=[100 100 10]; % Distribution upper bound in case of
discrete uniform distribution for diseased individuals
```

```
  for k=1:kmax % k is the variable that runs from 1 until kmax that has been
idendified before
    X(1:n1,k)= randi([a_d_h(k) b_d_h(k)], n1, 1); % This command creates
integers in the range from x to y, where x < y.
    %random('unid',a_d_h(k),b_d_h(k),n1,1); % Matrix X is filled up from the
first line until n1 line and from first column until kmax column with data
from uniform distribution for healthy cases
    X(n1+1:n1+n2,k)=randi([a_d_d(k) b_d_d(k)], n2, 1); % This command
creates integers in the range from x to y, where x < y.
    % random('unid',a_d_d(k),b_d_d(k),n2,1); % Matrix X is filled up from
the n1+1 line until n1+n2 line and from first column until kmax column with
data from uniform distribution for diseased cases
  end % End of the aforementioned loop
            else  if (iflag==4) % Normalized distribution (discrete)
                a_d_h=[1 2 3 4 5 6]; % Distribution lower bound in case of
discrete uniform distribution for healthy individuals
                a_d_d=[5 6 7 8 9 10]; % Distribution lower bound in case of
discrete uniform distribution for diseased individuals
                b_d_h=[15 15 15 15 15 15]; % Distribution upper bound in
case of discrete uniform distribution for healthy individuals
                b_d_d=[20 20 20 20 20 20]; % Distribution upper bound in
case of discrete uniform distribution for diseased individuals
                end
            end
        end
    end
    % [coefs,scores,variances,t2] = princomp(); %

    [pc, zscores, pcvars] = princomp(yeastvalues)
    cumsum(pcvars./sum(pcvars) * 100)
    figure
scatter(zscores(:,1),zscores(:,2));
xlabel('First Principal Component');
ylabel('Second Principal Component');
title('Principal Component Scatter Plot');

    %%%%% Data Standardization %%%%%
    if (jflag==1) % Data standardization if jflag=1
        for j=1:kmax % Loop for all the m components
            mu(j)=mean(X(:,j)); % Calculation of mean value of the total
sample
            muh(j)=mean(X(1:n1,j)); % Calculation of mean value of the
healthy individuals
            sdev(j)=std(X(:,j)); % Calculation of St. Dev. of the total
sample
            X(:,j)=(X(:,j)-muh(j))/sdev(j); % Matrix X is converted to
standardized X
        end
    end
%%%%% Binary Logistic Regression (BLR) %%%%%%
    [BLR,dev,stats] = glmfit(X, Y, 'binomial','logit'); % Multivariate
Binary Logistic Regression
    coef_m(1:kmax)=BLR(2:kmax+1);
    coef_m=exp(coef_m); % Odds Ratio from Multivariate Binary Logistic
Regression
    DEV_m=dev; % Deviance from Multivariate Binary Logistic Regression
    for k=1:kmax % Loop for the m components  - k is the variable that runs
from 1 until kmax
        [BLR,dev,stats] = glmfit(X(:,k), Y, 'binomial','logit'); %
Univariate Binary Logistic Regression
        coef_u(k)=exp(BLR(2)); % Odds Ratio from Univariate Binary Logistic
Regression
        DEV_u(k)=dev; %  Deviance from Univariate Binary Logistic Regression
        coef_ds_u(k)=coef_u(k)/DEV_u(k); % Odds Ratio from Univariate Binary
Logistic Regression divided by Deviance
        coef_ds_m(k)=coef_m(k)/DEV_u(k); % Odds Ratio from Multivariate
Binary Logistic Regression divided by Deviance
    end
```

```
    sumcoef_u=sum(coef_u)/kmax; % Sum of UBLR OR divided by kmax
    w_ulr=coef_u/sumcoef_u; % Weights from UBLR OR
    sumcoef_m=sum(coef_m)/kmax; % Sum of MBLR OR divided by kmax
    w_mlr=coef_m/sumcoef_m; % Weights from MBLR OR
    sumcoef_ds_u=sum(coef_ds_u)/kmax; % Sum of (UBLR OR divided by Deviance
Statistic) divided by kmax
    w_ulr_ds=coef_ds_u/sumcoef_ds_u; % Weights from UBLR OR/DS
    sumcoef_ds_m=sum(coef_ds_m)/kmax; % Sum of (MBLR OR divided by Deviance
Statistic) divided by kmax
    w_mlr_ds=coef_ds_m/sumcoef_ds_m; % Weights from MBLR OR/DS
%%%%% Linear Discriminant Analysis (LDA) %%%%%%
    [LDA,err,POSTERIOR,logp,coeff] = classify([X],[X],Y); % Multivariate
Linear Discriminant Analysis
    [A,B,r,U,V,stats1] = canoncorr(X,Y); % Statistical Output from
Multivariate Linear Discriminant Analysis
    coeff_m=coeff(2,1).linear; % % Fisher Coefficient from multivariate
Linear Discriminant Analysis
    coeff_m=abs(coeff_m); % Absolute values of Fisher coefficients - MLDA
    LW_m=stats1.Wilks;  % Wilks Lambda from multivariate
    for k=1:kmax % Loop for the m components
        [LDA,err,POSTERIOR,logp,coeff] = classify([X(:,k)],[X(:,k)],Y); %
Univariate Linear Discrimianant Analysis
        [A,B,r,U,V,stats1] = canoncorr(X(:,k),Y); % Statistical Output from
univariate Linear Discriminant Analysis
        coeff_u(k)=coeff(2,1).linear; % Fisher Coefficient from Univariate
Linear Discriminant Analysis
        coeff_u_sign(k)=coeff_u(k);  % Absolute values of Fisher
coefficients - ULDA
        coeff_u(k)=abs(coeff_u(k)); % Absolute values of Fisher coefficients
- ULDA
        LW_u(k)=stats1.Wilks; % Wilks Lamda from Univariate
        coeff_lw_u(k)=coeff_u(k)/LW_u(k); %  Fisher Coefficient from
univariate Linear Discriminant Analysis divided by Wilks Lamda
        coeff_lw_m(k)=coeff_m(k)/LW_u(k); % Fisher Coefficient from
multivariate Linear Discriminant Analysis divided by Wilks Lamda
    end
    sumcoeff_u=sum(coeff_u)/kmax; % Sum of ULDA FC divided by kmax
    w_ulda=coeff_u/sumcoeff_u; % Weights from ULDA
    sumcoeff_m=sum(coeff_m)/kmax; % Sum of MLDA FC divided by kmax
    w_mlda=coeff_m/sumcoeff_m; % Weights from MLDA
    sumcoeff_lw_u=sum(coeff_lw_u)/kmax; % Sum of (ULDA FC divided by Lambda
Wilks) divided by kmax
    w_ulda_lw=coeff_lw_u/sumcoeff_lw_u; % Weights from ULDA FC/LW
    sumcoeff_lw_m=sum(coeff_lw_m)/kmax; % Sum of (MLDA FC divided by Lambda
Wilks) divided by kmax
    w_mlda_lw=coeff_lw_m/sumcoeff_lw_m; % Weights from MLDA FC/LW
%%%% Support Vector Machines %%%%
 % svmStruct = svmtrain(X,Y,'showplot',true);
  % Group = svmclassify(SVMStruct,X);
%%%% Classification Trees %%%%
    ClassificationTree.fit(X,Y)
%%%% Neural Network %%%%


%%%%% Construction of Weights Matrix %%%%%%
    w=zeros(kmax,9);% construction of weight matrix
    ww=zeros(kmax,9,ixmax);
    w(:,1)=ones(1,kmax); ww(:,1,ix)=w(:,1);% first column ones ;
    w(:,2)=w_ulr; ww(:,2,ix)=w(:,2); % second column: UBLR weights
    w(:,3)=w_mlr; ww(:,3,ix)=w(:,3); % third column: MBLR weights
    w(:,4)=w_ulr_ds; ww(:,4,ix)=w(:,4); % fourth column: UBLR/DS weights
    w(:,5)=w_mlr_ds; ww(:,5,ix)=w(:,5); % fifth column: UBLR/DS weights
    w(:,6)=w_ulda; ww(:,6,ix)=w(:,6); % sixth column: ULDA weights
    w(:,7)=w_mlda; ww(:,7,ix)=w(:,7); % seventh column: MLDA weights
    w(:,8)=w_ulda_lw; ww(:,8,ix)=w(:,8); % eighth column: ULDA/LW (Lambda of
Wilks) weights
    w(:,9)=w_mlda_lw; ww(:,9,ix)=w(:,9); % ninth column: MLDA/LW (Lambda of
Wilks)weights
```

```
%%%%% Construction of Indices' Matrix %%%%%
    T=zeros(n1+n2,9); % Initialazation of indices' matrix - Assigning
dimensions (n1+n2)*9
    T=X*w; %  Multiplication of matrices X and w in order to produce the
indices T0-T8
    [FPR,TPR,Thresholds,AUC(ix,j),OPTROCPT] = perfcurve(Y,T(:,1),1); %
Function of the ROC curve.
        tnr=zeros(length(FPR),9);
        tpr=zeros(length(FPR),9);
        fnr=zeros(length(FPR),9);
        fpr=zeros(length(FPR),9);
    for j=1:9 % % Loop for the 9 indices
        [FPR,TPR,Thresholds,AUC(ix,j),OPTROCPT] = perfcurve(Y,T(:,j),1); %
Function of the ROC curve.
        tnr_tpr=1-FPR+TPR; % Sum of True Positive Rate and 1 minus False
Positive Rate
        TNR=1-FPR; % True Negative Rate (complementary possibility of FPR)
        tnr(:,j)=TNR(:); % True Negative Ratio for all Indices
        fpr(:,j)=1-tnr(:,j); % False Positive Ratio for all Indices
        tpr(:,j)=TPR; % True Positive Ratio for all Indices
        max1(j)=max(tnr_tpr); % Maximum from Matlab
        max2(j)=tnr_tpr(1); % Maximum Customized for validation
        imax(j)=1; % Initialazation of imax
        for i=2:length(tnr_tpr); % Loop for finding maximum and it's
position
            if(tnr_tpr(i)>max2(j));
                max2(j)=tnr_tpr(i);
                imax(j)=i; % Position of Indices that corresponds to the
Youden criterion
            end;
        end;
        TMAX(ix,j)=Thresholds(imax(j));% Value of the index that corresponds
to the threshold that becomes from maximization of Se+Sp
        TNRMAX(ix,j)=TNR(imax(j));%  TNR that corresponds to the threshold
that becomes from maximization of Se+Sp
        FPRMAX(ix,j)=FPR(imax(j));%  FPR that corresponds to the threshold
that becomes from maximization of Se+Sp
        TPRMAX(ix,j)=TPR(imax(j));%  TPR that corresponds to the threshold
that becomes from maximization of Se+Sp
        idh1=0; % Counter for accuracy from criterion 1
        for idh=1:n1; % Loop for finding True Negative Cases
            if (T(idh,j)< TMAX(ix,j)); % Inequality - criterion
                idh1=idh1+1; % Counter for True Negative Cases
            end;
        end
        TrN(ix,j)=idh1; % True Negative Cases
        for idh=n1+1:n1+n2; % Loop for finding True Positive Cases
            if (T(idh,j)> TMAX(ix,j));
                idh1=idh1+1; % Counter for True Positive Cases
            end;
        end;
        TrP(ix,j)=idh1-TrN(ix,j); % True Positive Cases
        TR(ix,j)=(TrP(ix,j)+TrN(ix,j))/(n1+n2); % True Classified Cases
        FlsP(ix,j)=n1-TrN(ix,j); %  False Positive Cases
        FlsN(ix,j)=n2-TrP(ix,j); %  False Negative Cases
        PPV(ix,j)=TrP(ix,j)/(TrP(ix,j)+FlsP(ix,j)); % Positive Predictive
Value
        NPV(ix,j)=TrN(ix,j)/(TrN(ix,j)+FlsN(ix,j)); % Negative Predictive
Value
        F1(ix,j)=2*TR(ix,j)*tpr(ix,j)/(TR(ix,j)+tpr(ix,j)); % F1 score: the
harmonic mean of precision and sensitivity
        G(ix,j)=sqrt(TR(ix,j)*tpr(ix,j));  % G score is the geometric mean
of precision and sensitivity: G=( TR*TPR)1/2
        MCC(ix,j)=((TrP(ix,j)*TrN(ix,j))-
(FlsP(ix,j)*FlsN(ix,j)))/sqrt((TrP(ix,j)+FlsP(ix,j))*(TrP(ix,j)+
FlsN(ix,j))*(TrN(ix,j)+FlsP(ix,j))*(TrN(ix,j)+ FlsN(ix,j))); % o  Matthews
correlation
    end
```

```
        for j=1:9 % % Loop for the 9 indices
            for ai=1:length(TPR)-1 % % Loop for the 9 indices

ppv(ai,j)=tpr(ai,j)*((n2)/(n1+n2))/(((tpr(ai,j))*((n2)/(n1+n2)))+(1-
tnr(ai,j))*((n1)/(n1+n2)));
            npv(ai,j)=tnr(ai,j)*((n1)/(n1+n2))/(((1-
tpr(ai,j))*((n2)/(n1+n2)))+tnr(ai,j)*((n1)/(n1+n2)));
            end;
        end;
    Cr1=[TNRMAX' TPRMAX']; % Criterion 1: Maximization of the sum (TPR+TNR)
Sensitivity + Specificity
%%%%% Construction of RANK AUC matrix - Indices' ranking as to their
position frequency %%%%%
for j=1:9; % Ranking of AUC
        RANK_AUC(ix,j)=1; % AUC Position/Ranking - Construction of RANK AUC
table
        for jj=1:9; % New index variable jj for running inside j=1:9
            if(AUC(ix,j)<AUC(ix,jj));  % Condition
                RANK_AUC(ix,j)=RANK_AUC(ix,j)+1; % Transfer 1 position lower
            end
        end
    end
    t_iter(ix)=toc; % Time counter ending
end % End of ix loop

% svmStruct = svmtrain(X,Y,'showplot',true)

%%%%% Ranking of AUC %%%%%
for j=1:9 % Health related indices T0, T1, ..., T8
    for ipos=1:9 % Position/Final Ranking of Indices
        ifreq(j,ipos)=0; % Frequency of j index for every position from 1 to
9
        for ix=1:ixmax  % Loop for ix
            if (RANK_AUC(ix,j)==ipos) %
                ifreq(j,ipos)=ifreq(j,ipos)+1; % 1 more unit in frequency
            end % End of if
        end % End of ix loop
    end % End of ipos loop
end % End of j loop
mean_iter_time=mean(t_iter); % Mean time
%%%%%%%%%%%%%%%%%%%  AUC2  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
ii=0;
for ix=1:ixmax %
   icount(ix)=0;
   icount2(ix)=0;
       for it=1:8; %
        for itt=it+1:9;  %
            if AUC(ix,it)==AUC(ix,itt); %
             icount(ix)=icount(ix)+1;
             icount2(ix)=1;
            end %
        end %
       end %
    if(icount(ix)==0);
        ii=ii+1;
        AUC2(ii,:)=AUC(ix,:);
    end % End of if
end % End of ix loop
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%% Construction of RANK AUC matrix - Indices' ranking as to their
position frequency %%%%%
 ix2max=size(AUC2,1);
 for ix=1:ix2max % Iterations' loop
 for j=1:9; % Ranking of AUC2
        RANK_AUC2(ix,j)=1; % AUC2 Position/Ranking - Construction of RANK
AUC2 table
        for jj=1:9; % New index variable jj for running inside j=1:9
            if(AUC2(ix,j)<AUC2(ix,jj));  % Condition
```

```
                RANK_AUC2(ix,j)=RANK_AUC2(ix,j)+1; % Transfer 1 position
lower
            end
        end
      end
end %

%%%%% Ranking of AUC2 %%%%%
for j=1:9 % Health related indices T0, T1, ..., T8
    for ipos=1:9 % Position/Final Ranking of Indices
        ifreq2(j,ipos)=0; % Frequency of j index for every position from 1
to 9
        for ix=1:ix2max  % Loop for ix
            if (RANK_AUC2(ix,j)==ipos) %
                ifreq2(j,ipos)=ifreq2(j,ipos)+1; % 1 more unit in frequency
            end % End of if
        end % End of ix loop
    end % End of ipos loop
 end % End of j loop

%%%%% Descriptive Statistics of Evaluation Measures - Mean Value - Std.
Deviation %%%%%
for j=1:9
    mean_AUC(j)= mean(AUC(:,j)); % Mean value of AUC for iterations
performed
    mean_TPR(j)= mean(TPRMAX(:,j)); % Mean value of TPR for iterations
performed
    mean_TNR(j)= mean(TNRMAX(:,j)); % Mean value of TNR for iterations
performed
    mean_TR(j)= mean(TR(:,j)); % Mean value of TR for iterations performed
    mean_PPV(j)= mean(PPV(:,j)); % Mean value of PPV for iterations
performed
    mean_NPV(j)= mean(NPV(:,j)); % Mean value of NPV for iterations
performed
    mean_F1(j)= mean(F1(:,j)); % Mean value of F1 for iterations performed
    mean_G(j)= mean(G(:,j)); % Mean value of G for iterations performed
    mean_MCC(j)= mean(MCC(:,j)); % Mean value of G for iterations performed
    std_AUC(j)= std(AUC(:,j)); % St. Dev. of AUC for iterations performed
    std_TPR(j)= std(TPRMAX(:,j)); % St. Dev. of TPR for iterations performed
    std_TNR(j)= std(TNRMAX(:,j)); % St. Dev. of TNR for iterations performed
    std_TR(j)= std(TR(:,j)); % St. Dev. of TR for iterations performed
    std_PPV(j)= std(PPV(:,j)); % St. Dev. of PPV for iterations performed
    std_NPV(j)= std(NPV(:,j)); % St. Dev. of NPV for iterations performed
    std_F1(j)= std(F1(:,j)); % St. Dev. of F1 for iterations performed
    std_G(j)= std(G(:,j)); % St. Dev. of G for iterations performed
    std_MCC(j)= std(MCC(:,j)); % St. Dev. of G for iterations performed
end
for it=1:kmax % Loop for finding mean weights from the proposed methods
    for j=1:9
        meanw(it,j)=mean(ww(it,j,:))
    end
end
%%%%% Frequencies of indices in each position %%%%%
ifreq_stack(:,1)=ifreq(1,:)/ixmax; % Frequencies of indices in first
position
ifreq_stack(:,2)=ifreq(2,:)/ixmax; % Frequencies of indices in second
position
ifreq_stack(:,3)=ifreq(3,:)/ixmax; % Frequencies of indices in third
position
ifreq_stack(:,4)=ifreq(4,:)/ixmax; % Frequencies of indices in fourth
position
ifreq_stack(:,5)=ifreq(5,:)/ixmax; % Frequencies of indices in fifth
position
ifreq_stack(:,6)=ifreq(6,:)/ixmax; % Frequencies of indices in sixth
position
ifreq_stack(:,7)=ifreq(7,:)/ixmax; % Frequencies of indices in seventh
position
```

```
ifreq_stack(:,8)=ifreq(8,:)/ixmax; % Frequencies of indices in eight
position
ifreq_stack(:,9)=ifreq(9,:)/ixmax; % Frequencies of indices in ninth
position

%%%%% Frequencies of indices in each position %%%%%
 ifreq2_stack(:,1)=ifreq2(1,:)/ix2max; % Frequencies of indices in first
position
 ifreq2_stack(:,2)=ifreq2(2,:)/ix2max; % Frequencies of indices in second
position
 ifreq2_stack(:,3)=ifreq2(3,:)/ix2max; % Frequencies of indices in third
position
 ifreq2_stack(:,4)=ifreq2(4,:)/ix2max; % Frequencies of indices in fourth
position
 ifreq2_stack(:,5)=ifreq2(5,:)/ix2max; % Frequencies of indices in fifth
position
 ifreq2_stack(:,6)=ifreq2(6,:)/ix2max; % Frequencies of indices in sixth
position
 ifreq2_stack(:,7)=ifreq2(7,:)/ix2max; % Frequencies of indices in seventh
position
 ifreq2_stack(:,8)=ifreq2(8,:)/ix2max; % Frequencies of indices in eight
position
 ifreq2_stack(:,9)=ifreq2(9,:)/ix2max; % Frequencies of indices in ninth
position

%%%%% Indices AUC values for stack plot with 99% confidence interval %%%%%
AUC_stack1(:,1)=mean_AUC'-2.576*std_AUC'/sqrt(ixmax); % Lower bound of C.I.
99% for AUC
AUC_stack1(:,2)=2.576*std_AUC'/sqrt(ixmax); % Mean AUC
AUC_stack1(:,3)=2.576*std_AUC'/sqrt(ixmax); % Upper bound of C.I. 99% for
AUC
%%%%% Indices PPV values for stack plot with 99% confidence interval %%%%%
PPV_stack1(:,1)=mean_PPV'-2.576*std_PPV'/sqrt(ixmax); % Lower bound of C.I.
99% for PPV
PPV_stack1(:,2)=2.576*std_PPV'/sqrt(ixmax); % Mean PPV
PPV_stack1(:,3)=2.576*std_PPV'/sqrt(ixmax); % Upper bound of C.I. 99% for
PPV
%%%%% Indices NPV values for stack plot with 99% confidence interval %%%%%
NPV_stack1(:,1)=mean_NPV'-2.576*std_NPV'/sqrt(ixmax); % Lower bound of C.I.
99% for NPV
NPV_stack1(:,2)=2.576*std_NPV'/sqrt(ixmax); % Mean NPV
NPV_stack1(:,3)=2.576*std_NPV'/sqrt(ixmax); % Upper bound of C.I. 99% for
NPV
%%%%% Indices TR values for stack plot with 99% confidence interval %%%%%
TR_stack1(:,1)=mean_TR'-2.576*std_TR'/sqrt(ixmax); % Lower bound of C.I. 99%
for TR
TR_stack1(:,2)=2.576*std_TR'/sqrt(ixmax); % Mean TR
TR_stack1(:,3)=2.576*std_TR'/sqrt(ixmax); % Upper bound of C.I. 99% for TR
%%%%% Indices TPR values for stack plot with 99% confidence interval %%%%%
TPR_stack1(:,1)=mean_TPR'-2.576*std_TPR'/sqrt(ixmax); % Lower bound of C.I.
99% for TPR
TPR_stack1(:,2)=2.576*std_TPR'/sqrt(ixmax); % Mean TPR
TPR_stack1(:,3)=2.576*std_TPR'/sqrt(ixmax); % Upper bound of C.I. 99% for
TPR
%%%%% Indices TNR values for stack plot with 99% confidence interval %%%%%
TNR_stack1(:,1)=mean_TNR'-2.576*std_TNR'/sqrt(ixmax); % Lower bound of C.I.
99% for TNR
TNR_stack1(:,2)=2.576*std_TNR'/sqrt(ixmax); % Mean TNR
TNR_stack1(:,3)=2.576*std_TNR'/sqrt(ixmax); % Upper bound of C.I. 99% for
TNR
%%%%% Indices F1 values for stack plot with 99% confidence interval %%%%%
F1_stack1(:,1)=mean_F1'-2.576*std_F1'/sqrt(ixmax); % Lower bound of C.I. 99%
for F1
F1_stack1(:,2)=2.576*std_F1'/sqrt(ixmax); % Mean F1
F1_stack1(:,3)=2.576*std_F1'/sqrt(ixmax); % Upper bound of C.I. 99% for F1
%%%%% Indices G values for stack plot with 99% confidence interval %%%%%
G_stack1(:,1)=mean_G'-2.576*std_G'/sqrt(ixmax); % Lower bound of C.I. 99%
for G
```

```
G_stack1(:,2)=2.576*std_G'/sqrt(ixmax); % Mean G
G_stack1(:,3)=2.576*std_G'/sqrt(ixmax); % Upper bound of C.I. 99% for G
%%%%% Indices MCC values for stack plot with 99% confidence interval %%%%%
MCC_stack1(:,1)=mean_MCC'-2.576*std_MCC'/sqrt(ixmax); % Lower bound of C.I.
99% for G
MCC_stack1(:,2)=2.576*std_MCC'/sqrt(ixmax); % Mean G
MCC_stack1(:,3)=2.576*std_MCC'/sqrt(ixmax); % Upper bound of C.I. 99% for G

%%%%% Figure with 99% confidence interval for AUC values %%%%%
figure (100) % Figure - Barchart of indices' AUC with C.I.99%
bar(AUC_stack1,'stacked')
title('AUC')
ylabel('mean AUC (C.I. 99%)')
xlabel('Index')
%%%%% Figure with 99% confidence interval for TR values %%%%%
figure (200) % Figure - Barchart of indices' TR with C.I.99%
bar(TR_stack1,'stacked')
title('TR')
ylabel('mean TR (C.I. 99%)')
xlabel('Index')
%%%%% Figure with 99% confidence interval for TPR values %%%%%
figure (300) % Figure - Barchart of indices' TPR with C.I.99%
bar(TPR_stack1,'stacked')
title('TPR')
ylabel('mean TPR (C.I. 99%)')
xlabel('Index')
%%%%% Figure with 99% confidence interval for TNR values %%%%%
figure (400) % Figure - Barchart of indices' TNR with C.I.99%
bar(TNR_stack1,'stacked')
title('TNR')
ylabel('mean TNR (C.I. 99%)')
xlabel('Index')
%%%%% Figure with 99% confidence interval for PPV values %%%%%
figure (900) % Figure - Barchart of indices' PPV with C.I.99%
bar(PPV_stack1,'stacked')
title('PPV')
ylabel('mean PPV (C.I. 99%)')
xlabel('Index')
%%%%% Figure with 99% confidence interval for NPV values %%%%%
figure (1000) % Figure - Barchart of indices' NPV with C.I.99%
bar(NPV_stack1,'stacked')
title('NPV')
ylabel('mean NPV (C.I. 99%)')
xlabel('Index')
%%%%% Figure with 99% confidence interval for F1 values %%%%%
figure (1100) % Figure - Barchart of indices' F1 with C.I.99%
bar(F1_stack1,'stacked')
title('F1')
ylabel('mean F1 (C.I. 99%)')
xlabel('Index')

%%%%% Figure with 99% confidence interval for G values %%%%%
figure (1200) % Figure - Barchart of indices' G with C.I.99%
bar(G_stack1,'stacked')
title('G')
ylabel('mean G (C.I. 99%)')
xlabel('Index')

%%%%% Figure with 99% confidence interval for MCC values %%%%%
figure (1300) % Figure - Barchart of indices' MCC with C.I.99%
bar(MCC_stack1,'stacked')
title('MCC')
ylabel('mean MCC (C.I. 99%)')
xlabel('Index')

%%%%% Figure for AUC values frequency %%%%%
figure (11) % Figure - Barchart stacked of indices' AUC frequency
bar(ifreq_stack,'stacked')
```

```
title('Indices Ranking')
ylabel('Frequency')
xlabel('Position')

%%%%% Figure for AUC values frequency %%%%%
figure (12) % Figure - Barchart stacked of indices' AUC frequency
bar(ifreq2_stack,'stacked')
title('Indices Ranking 2')
ylabel('Frequency')
xlabel('Position')

%%%%%% Figure - ROC Curve%%%%%%
color=['r' 'g' 'b' 'c' 'm' 'y' 'r' 'g' 'b']
figure (14) % Figure - ROC Plots of the last iteration
for j=1:9
  plot(fpr(:,j),tpr(:,j), color(j)) %
    hold on
    xlabel('False positive rate')
    ylabel('True positive rate')
    title('ROC for Classification health indices')
end

%%%%%% Figure -  PPV vs 1- NPV %%%%%%
color=['r' 'g' 'b' 'c' 'm' 'y' 'r' 'g' 'b']
figure (15) % Figure - PPV vs 1- NPV Plots of the last iteration
for j=1:9
  plot(ppv(:,j),1-npv(:,j),color(j))
    hold on
    xlabel('PPV')
    ylabel('1-NPV')
    title('PPV vs 1- NPV')
end

%%%%%% Figure -  NPV vs 1- PPV %%%%%%
color=['r' 'g' 'b' 'c' 'm' 'y' 'r' 'g' 'b']
figure (16) % Figure - NPV vs 1- PPV Plots of the last iteration
for j=1:9
  plot(1-ppv(:,j),npv(:,j),color(j))
    hold on
    xlabel('1-PPV')
    ylabel('NPV')
    title('NPV vs 1- PPV')
end
```

Algorithm 2: Training & Test Sample

```
clear all % Erase the previous data & results
format long
ixmax=10; % ("ix" is the variable that corresponds to the total amount of
repeated experimenets - therefore if I construct 1000 samples, ix=1000, each
one is of size 10.000)
jflag=1; % (jflag is the variable that indicates Data standardization)
for ix=1:ixmax
% (; is used for avoiding printing outputs - whatever I write after % is not
getting in the algorith, is just a comment)
tic % Time counter beggining
n1=7500; % Number of non diseased sample cases - From this number I adjust
the ratio between diseased and healthy individuals
n2=2500; % Number of diseased sample cases - From this number I adjust the
ration between diseased and healthy individuals
kmax=6; % Number of component variables that can be customized - I shall
produce different scenarios with various numbers of components
iflag=2; % Determination of measurement & distribution - iflag is an
auxiliary/indexing variable that determines from which distribution shall
the primary data comes from
X=zeros(n1+n2, kmax); % Initialazation of variables' matrix - Assigning
dimensions (n1+n2*kmax) - Produce a matrix full of zeroes for the components
matrix
Y=zeros(n1+n2, 1); % Initialazation of dependent variable matrix - Assigning
dimensions (n1+n2*1) - Produce a matrix full of zeroes for the dependents
matrix
Y1=zeros(n1,1); % Constraction of binomial dependent variable - healthy
cases - Value zero (0) in this column
Y2=ones(n2,1);% Constraction of binomial dependent variable - diseased cases
- Value ones (1) in this column
Y=[Y1;Y2];  %; is used for changing line - I constract the vector Y by
adding Y1 and Y2 one after another
 if (iflag==1) % Normal distribution
 mu_h=[110 108 106 104 102 100]; % [109.5 109 108.5 108 107.5 107 106.5 106
105.5 105 104.5 104 103.5 103 102.5 102 101.5 101 100.5] Distribution mean
values in case of normal distribution for healthy individuals
 mu_d=[120 122 124 126 128 130]; % [120.5 121 121.5 122 122.5 123 123.5 124
124.5 125 125.5 126 126.5 127 127.5 128 128.5 129 129.5] Distribution mean
values in case of normal distribution for diseased individuals
 sigma_h=[30 28 26 24 22 20]; % [29.5 29 28.5 28 27.5 27 26.5 26 25.5 25
24.5 24 23.5 23 22.5 22 21.5 21 20.5] Distribution standard deviation in
case of normal distribution for healthy individuals
 sigma_d=[30 28 26 24 22 20]; % [29.5 29 28.5 28 27.5 27 26.5 26 25.5 25
24.5 24 23.5 23 22.5 22 21.5 21 20.5] Distribution standard deviation in
case of normal distribution for diseased individuals
  for k=1:kmax % k is the variable that runs from 1 until kmax that has been
idendified before
   X(1:n1,k)=random('Normal',mu_h(k),sigma_h(k),n1,1); % Matrix X is filled
up from the first line until n1 line and from first column until kmax column
with data from normal distribution for healthy cases
   X(n1+1:n1+n2,k)=random('Normal',mu_d(k),sigma_d(k),n2,1); % Matrix X is
filled up from the n1+1 line until n1+n2 line and from first column until
kmax column with data from normal distribution for diseased cases
  end % End of the aforementioned loop
 else if (iflag==2) % Uniform distribution (continuous)
 a_h=[50 48 46 44 42 40]; % Distribution lower bound in case of continuous
uniform distribution for healthy individuals
 a_d=[70 72 74 76 78 80]; % Distribution lower bound in case of continuous
uniform distribution for diseased individuals
 b_h=[150 148 146 144 142 140]; % Distribution upper bound in case of
continuous uniform distribution for healthy individuals
 b_d=[170 172 174 176 178 180]; % Distribution upper bound in case of
continuous uniform distribution for diseased individuals
  for k=1:kmax % k is the variable that runs from 1 until kmax that has been
idendified before
```

```
    X(1:n1,k)=random('unif',a_h(k),b_h(k),n1,1); % Matrix X is filled up
from the first line until n1 line and from first column until kmax column
with data from uniform distribution for healthy cases
    X(n1+1:n1+n2,k)=random('unif',a_d(k),b_d(k),n2,1); % Matrix X is filled
up from the n1+1 line until n1+n2 line and from first column until kmax
column with data from uniform distribution for diseased cases
  end % End of the aforementioned loop
  else  if (iflag==3) % Uniform distribution (discrete)
 a_d_h=[1 1]; % Distribution lower bound in case of discrete uniform
distribution for healthy individuals
 a_d_d=[2 3]; % Distribution lower bound in case of discrete uniform
distribution for diseased individuals
 b_d_h=[4 3]; % Distribution upper bound in case of discrete uniform
distribution for healthy individuals
 b_d_d=[5 5]; % Distribution upper bound in case of discrete uniform
distribution for diseased individuals
  for k=1:kmax % k is the variable that runs from 1 until kmax that has been
idendified before
    X(1:n1,k)= randi([a_d_h(k) b_d_h(k)], n1, 1); % This command creates
integers in the range from x to y, where x < y.
    %random('unid',a_d_h(k),b_d_h(k),n1,1); % Matrix X is filled up from the
first line until n1 line and from first column until kmax column with data
from uniform distribution for healthy cases
    X(n1+1:n1+n2,k)=randi([a_d_d(k) b_d_d(k)], n2, 1); % This command
creates integers in the range from x to y, where x < y.
    % random('unid',a_d_d(k),b_d_d(k),n2,1); % Matrix X is filled up from
the n1+1 line until n1+n2 line and from first column until kmax column with
data from uniform distribution for diseased cases
  end % End of the aforementioned loop
  else  if (iflag==4) % Normalized distribution (discrete)
  for k=1:kmax % k is the variable that runs from 1 until kmax that has been
idendified before
   X(1:n1,k)=random('Normal',mu_h(k),sigma_h(k),n1,1); % Matrix X is filled
up from the first line until n1 line and from first column until kmax column
with data from normal distribution for healthy cases
   X(n1+1:n1+n2,k)=random('Normal',mu_d(k),sigma_d(k),n2,1); % Matrix X is
filled up from the n1+1 line until n1+n2 line and from first column until
kmax column with data from normal distribution for diseased cases
  end % End of the aforementioned loop
        end
      end
    end
 end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%5
 if (jflag==1) % Data standardization
 for j=1:kmax
 mu(j)=mean(X(:,j));
 muh(j)=mean(X(1:n1,j));
 sdev(j)=std(X(:,j));
  X(:,j)=(X(:,j)-muh(j))/sdev(j); % Matrix X is converted to standardized X
 end
 end
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
TRAINING & TEST SAMPLE
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%
 X01=X(1:nearest(0.9*n1),:); % Constraction of the training subsample X of
healthy cases - 90% of the initial sample
 X02=X(n1+1:n1+nearest(0.9*n2),:); %  Constraction of the training subsample
X of non healthy cases - 90% of the initial sample
 Y01=Y(1:nearest(0.9*n1),1); % Constraction of the training subsample of
healthy cases Y - 90% of the initial sample
 Y02=Y(n1+1:n1+nearest(0.9*n2),1); %  Constraction of the training subsample
X of non healthy cases - 90% of the initial sample
 X11=X(nearest(0.9*n1)+1:n1,:); % Constraction of the test subsample of
healthy cases - 10% of the initial sample
```

```
 X12=X(n1+nearest(0.9*n2)+1:n1+n2,:); %  Constraction of the test subsample
of non healthy cases - 10% of the initial sample
 L11=length(X11);
 L12=length(X12);
 Y11=Y(nearest(0.9*n1)+1:n1,1); % Constraction of the test subsample of Y
healthy cases - 10% of the initial sample
 Y12=Y(n1+nearest(0.9*n2)+1:n1+n2,:); %  Constraction of the test subsample
of non healthy cases Y - 10% of the initial sample
 XTR=[X01 ; X02];
 XSA=[X11 ; X12];
 YTR=[Y01 ; Y02];
 YSA=[Y11 ; Y12];
 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% BLR
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%
[BLR,dev,stats] = glmfit(XTR, YTR, 'binomial','logit'); % Calculating Odds
Ratio and Deviance from Multivariate Binary Logistic Regression by using the
function "glmfit" for each columnn/variable
    coef_m(1:kmax)=BLR(2:kmax+1);
    coef_m=exp(coef_m); % Calculating Odds Ratio from Multivariate Binary
Logistic Regression  - saved in coef_m
    DEV_m=dev; % Calculating  Deviance from Multivariate Binary Logistic
Regression
for k=1:kmax % k is the variable that runs from 1 until kmax that has been
idendified before
    [BLR,dev,stats] = glmfit(XTR(:,k), YTR, 'binomial','logit'); %
Calculating Odds Ratio and Deviance from Univariate Binary Logistic
Regression by using the function "glmfit" for each columnn/variable
    coef_u(k)=exp(BLR(2)); % Calculating Odds Ratio from Univariate Binary
Logistic Regression - saved in coef_u
    DEV_u(k)=dev; % Calculating  Deviance from Univariate Binary Logistic
Regression
    coef_ds_u(k)=coef_u(k)/DEV_u(k); % Calculating Odds Ratio from
Univariate Binary Logistic Regression divided by Deviance - saved in
coef_ds_u
    coef_ds_m(k)=coef_m(k)/DEV_u(k); % Calculating Odds Ratio from
Multivariate Binary Logistic Regression divided by Deviance - saved in
coef_ds_m
end
sumcoef_u=sum(coef_u)/kmax; % Sum of UBLR OR divided by kmax
   w_ulr=coef_u/sumcoef_u; % Weights from UBLR - to be multiplied with
corresponding values of each variable
sumcoef_m=sum(coef_m)/kmax; % Sum of MBLR OR divided by kmax
   w_mlr=coef_m/sumcoef_m; % Weights from MBLR - to be multiplied with
corresponding values of each variable
sumcoef_ds_u=sum(coef_ds_u)/kmax; % Sum of (UBLR OR divided by Deviance
Statistic) divided by kmax
   w_ulr_ds=coef_ds_u/sumcoef_ds_u; % Weights from UBLR/DS - to be
multiplied with corresponding values of each variable
sumcoef_ds_m=sum(coef_ds_m)/kmax; % Sum of (MBLR OR divided by Deviance
Statistic) divided by kmax
   w_mlr_ds=coef_ds_m/sumcoef_ds_m; % Weights from MBLR/DS - to be
multiplied with corresponding values of each variable
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%5    LDA
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[LDA,err,POSTERIOR,logp,coeff] = classify([XTR],[XTR],YTR); % Calculating
Fiscer Coefficient from Multivariate Linear Discrimianant Analysis by using
the function "classify" for each columnn/variable
[A,B,r,U,V,stats1] = canoncorr(XTR,YTR);
coeff_m=coeff(2,1).linear; % Calculating Fiscer Coefficient from
Multivariate Linear Discrimianant Analysis - saved in coeff_u
coeff_m_sign=coeff_m;
coeff_m=abs(coeff_m);% Absolute values of Fisher coefficients - MLDA
LW_m=stats1.Wilks;
for k=1:kmax % k is the variable that runs from 1 until kmax that has been
idendified before
```

```
    [LDA,err,POSTERIOR,logp,coeff] = classify([XTR(:,k)],[XTR(:,k)],YTR); %
Calculating Fiscer Coefficient from Univariate Linear Discrimianant Analysis
by using the function "classify" for each columnn/variable
    [A,B,r,U,V,stats1] = canoncorr(XTR(:,k),YTR);
    coeff_u(k)=coeff(2,1).linear; % Calculating Fisher Coefficient from
Univariate Linear Discriminant Analysis - saved in coeff_u
    coeff_u_sign(k)=coeff_u(k);
    coeff_u(k)=abs(coeff_u(k)); % Absolute values of Fisher coefficients -
ULDA
    LW_u(k)=stats1.Wilks;
    coeff_lw_u(k)=coeff_u(k)/LW_u(k); % Calculating Odds Ratio from
Univariate Binary Logistic Regression divided by Deviance - saved in
coef_ds_u
    coeff_lw_m(k)=coeff_m(k)/LW_u(k); % Calculating Odds Ratio from
Multivariate Binary Logistic Regression divided by Deviance - saved in
coef_ds_m
end
sumcoeff_u=sum(coeff_u)/kmax; % Sum of ULDA FC divided by kmax
   w_ulda=coeff_u/sumcoeff_u; % Weights from ULDA - to be multiplied with
corresponding values of each variable
sumcoeff_m=sum(coeff_m)/kmax; % Sum of MLDA FC divided by kmax
   w_mlda=coeff_m/sumcoeff_m; % Weights from MLDA - to be multiplied with
corresponding values of each variable
sumcoeff_lw_u=sum(coeff_lw_u)/kmax; % Sum of (ULDA FC divided by Lambda
Wilks) divided by kmax
   w_ulda_lw=coeff_lw_u/sumcoeff_lw_u; % Weights from ULDA/LW - to be
multiplied with corresponding values of each variable
sumcoeff_lw_m=sum(coeff_lw_m)/kmax; % Sum of (MLDA FC divided by Lambda
Wilks) divided by kmax
   w_mlda_lw=coeff_lw_m/sumcoeff_lw_m; % Weights from MLDA/LW - to be
multiplied with corresponding values of each variable
   %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Weights
Matrix
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    w=zeros(kmax,9);% construction of weight matrix
  ww=zeros(kmax,9,ixmax);
w(:,1)=ones(1,kmax); ww(:,1,ix)=w(:,1);% first column ones ;
 w(:,2)=w_ulr; ww(:,2,ix)=w(:,2); % second column of UBLR weights
  w(:,3)=w_mlr; ww(:,3,ix)=w(:,3); % third column of MBLR weights
   w(:,4)=w_ulr_ds; ww(:,4,ix)=w(:,4); % fourth column of UBLR/DS weights
    w(:,5)=w_mlr_ds; ww(:,5,ix)=w(:,5); % fifth column of UBLR/DS weights
     w(:,6)=w_ulda; ww(:,6,ix)=w(:,6); % sixth column of ULDA weights
      w(:,7)=w_mlda; ww(:,7,ix)=w(:,7); % seventh column of MLDA weights
       w(:,8)=w_ulda_lw; ww(:,8,ix)=w(:,8); % eighth column of ULDA/LW
(Lambda of Wilks) weights
        w(:,9)=w_mlda_lw; ww(:,9,ix)=w(:,9); % ninth column of MLDA/LW
(Lambda of Wilks)weights
        %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%5%%%%%%%%%%%%%%%%%%%% Index
Matrix
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%5%%%
l1=length(XSA);
T=zeros(l1,9); % Initialazation of indices matrix - Assigning dimensions l1
- because 9 indices are constructed
T=XSA*w; %  Multiplication of matrices X () and w (weights) in order to
produce the indices
opt=zeros(9,2); %
for j=1:9 % i is the variable that runs from 1 until 9 that represents the
amount of indices
 [FPR,TPR,Thresholds,AUC(ix,j),OPTROCPT] = perfcurve(YSA,T(:,j),1); %
returns the optimal operating point of the ROC curve.
opt(j,1)=1-OPTROCPT(1); %  coordinates - abscissa
opt(j,2)=OPTROCPT(2); % coordinates - ordinate
Cr3=opt; % Criterion 3: MATLAB Criterion
tnr_tpr=1-FPR+TPR; % Sum of True Positive Rate and 1 minus False Positive
Rate
TNR=1-FPR; % True Negative Rate is equal to 1 minus False Positive Ratio
(complementary possibility of FPR)
tnr(:,j)=TNR(:); % True Negative Ratio for all Indices
```

```
fpr(:,j)=1-tnr(:,j); % False Positive Ratio for all Indices
tpr(:,j)=TPR; % True Positive Ratio for all Indices
Distance=FPR.^2+(TPR-1).^2; % Distance from the upper left point
min1(j)=min(Distance);
min2(j)=min(Distance(1));
imin(j)=1;
max1(j)=max(tnr_tpr); % Maximum from Matlab
max2(j)=tnr_tpr(1); % Maximum Customized by Fragy
imax(j)=1;
for i=2:length(tnr_tpr);
    if(tnr_tpr(i)>max2(j));
        max2(j)=tnr_tpr(i);
        imax(j)=i; % Location that corresponds to the Youden criterion
    end;
    if(Distance(i)<min2(j));
        min2(j)=Distance(i);
        imin(j)=i; % Location that corresponds to this criterion
    end;
end;
TMAX(ix,j)=Thresholds(imax(j));% Value of the index that corresponds to the
threshold that becomes from maximization of Se+Sp (ERROR)
TNRMAX(ix,j)=TNR(imax(j));%  TNR that corresponds to the threshold that
becomes from maximization of Se+Sp
FPRMAX(ix,j)=FPR(imax(j));%  FPR that corresponds to the threshold that
becomes from maximization of Se+Sp
TPRMAX(ix,j)=TPR(imax(j));%  TPR that corresponds to the threshold that
becomes from maximization of Se+Sp
TMIN(ix,j)=T(imin(j),j);% Value of the index that corresponds to the
threshold that becomes from minimization of Distance
TNRMIN(ix,j)=TNR(imin(j));% TNR that corresponds to the  threshold that
becomes from minimization of Distance
FPRMIN(ix,j)=FPR(imin(j));% FPR that corresponds to the  threshold that
becomes from minimization of Distance
TPRMIN(ix,j)=TPR(imin(j));% TPR that corresponds to the  threshold that
becomes from minimization of Distance
TNROPT(ix,j)=opt(j,1);% TNR that corresponds maybe to the threshold from
MATLAB
TPROPT(ix,j)=opt(j,2);% TPR that corresponds maybe to the threshold from
MATLAB
idh1=0; % Counter for accuracy from criterion 1
idh2=0; % Counter for accuracy from criterion 2
idh3=0; % Counter for accuracy from criterion 3
for idh=1:L11;
   if (T(idh,j)< TMAX(ix,j));
       idh1=idh1+1;
    end;
end
TrN(ix,j)=idh1;

for idh=L11+1:L11+L12;
    if (T(idh,j)> TMAX(ix,j));
       idh1=idh1+1; %
    end;
    end;
    TrP(ix,j)=idh1-TrN(ix,j);
TR(ix,j)=(TrP(ix,j)+TrN(ix,j))/(L11+L12);
FlsP(ix,j)=L11-TrN(ix,j);
FlsN(ix,j)=L12-TrP(ix,j);
PPV(ix,j)=TrP(ix,j)/(TrP(ix,j)+FlsP(ix,j));
NPV(ix,j)=TrN(ix,j)/(TrN(ix,j)+FlsN(ix,j));
end
Cr1=[TNRMAX' TPRMAX']; % Criterion 1: Maximization of the sum (TPR+TNR)
Sensitivity + Specificity
Cr2=[TNRMIN' TPRMIN']; % Criterion 2: Minimization of distance from upper
left point

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Construction of RANK AUC
matrix %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
for j=1:9; % Ranking of AUC
RANK_AUC(ix,j)=1; % Initialazation of AUC Position/Ranking - Construction of
RANK AUC table
for jj=1:9; % New index variable jj for running inside j=1:9
    if(AUC(ix,j)<AUC(ix,jj));  %
       RANK_AUC(ix,j)=RANK_AUC(ix,j)+1; % oso isxuei to apo panw if tote
metefere ti thesi tou j index kata mia thesi pio katw / xeiroteri thesi
    end % End of if
end % End of jj loop
end % End of j loop
t_iter(ix)=toc % Time counter ending
end % End of ix loop

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% Ranking of AUC
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%5
for j=1:9 % Health related indices T0, T1, ..., T8
for ipos=1:9 % Position/Final Ranking of Indices
    ifreq(j,ipos)=0; % Frequency of j index for every position from 1 to 9
    for ix=1:ixmax
    if (RANK_AUC(ix,j)==ipos) %
       ifreq(j,ipos)=ifreq(j,ipos)+1; %
    end % End of if
    end % End of ix loop
end % End of ipos loop
end % End of j loop

 mean_iter_time=mean(t_iter)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for j=1:9
    mean_AUC(j)= mean(AUC(:,j));
    mean_TPR(j)= mean(TPRMAX(:,j));
    mean_TNR(j)= mean(TNRMAX(:,j));
    mean_TR(j)= mean(TR(:,j));
    mean_PPV(j)= mean(PPV(:,j));
    mean_NPV(j)= mean(NPV(:,j));
    std_AUC(j)= std(AUC(:,j));
    std_TPR(j)= std(TPRMAX(:,j));
    std_TNR(j)= std(TNRMAX(:,j));
    std_TR(j)= std(TR(:,j));
    std_PPV(j)= std(PPV(:,j));
    std_NPV(j)= std(NPV(:,j));
    mean_TPRmin(j)= mean(TPRMIN(:,j));
    mean_TNRmin(j)= mean(TNRMIN(:,j));
    mean_TPRopt(j)= mean(TPROPT(:,j));
    mean_TNRopt(j)= mean(TNROPT(:,j));
    std_TPRmin(j)= std(TPRMIN(:,j));
    std_TNRmin(j)= std(TNRMIN(:,j));
    std_TPRopt(j)= std(TPROPT(:,j));
    std_TNRopt(j)= std(TNROPT(:,j));
end
for it=1:kmax
    for j=1:9
        meanw(it,j)=mean(ww(it,j,:));
    end
end

ifreq_stack(:,1)=ifreq(1,:)/ixmax;
ifreq_stack(:,2)=ifreq(2,:)/ixmax;
ifreq_stack(:,3)=ifreq(3,:)/ixmax;
ifreq_stack(:,4)=ifreq(4,:)/ixmax;
ifreq_stack(:,5)=ifreq(5,:)/ixmax;
ifreq_stack(:,6)=ifreq(6,:)/ixmax;
ifreq_stack(:,7)=ifreq(7,:)/ixmax;
ifreq_stack(:,8)=ifreq(8,:)/ixmax;
ifreq_stack(:,9)=ifreq(9,:)/ixmax;
```

```
AUC_stack(:,1)=mean_AUC'-1.96*std_AUC'/sqrt(ixmax);
AUC_stack(:,2)=1.96*std_AUC'/sqrt(ixmax);
AUC_stack(:,3)=1.96*std_AUC'/sqrt(ixmax);

PPV_stack(:,1)=mean_PPV'-1.96*std_PPV'/sqrt(ixmax);
PPV_stack(:,2)=1.96*std_PPV'/sqrt(ixmax);
PPV_stack(:,3)=1.96*std_PPV'/sqrt(ixmax);

NPV_stack(:,1)=mean_NPV'-1.96*std_NPV'/sqrt(ixmax);
NPV_stack(:,2)=1.96*std_NPV'/sqrt(ixmax);
NPV_stack(:,3)=1.96*std_NPV'/sqrt(ixmax);

TR_stack(:,1)=mean_TR'-1.96*std_TR'/sqrt(ixmax);
TR_stack(:,2)=1.96*std_TR'/sqrt(ixmax);
TR_stack(:,3)=1.96*std_TR'/sqrt(ixmax);

TPR_stack(:,1)=mean_TPR'-1.96*std_TPR'/sqrt(ixmax);
TPR_stack(:,2)=1.96*std_TPR'/sqrt(ixmax);
TPR_stack(:,3)=1.96*std_TPR'/sqrt(ixmax);

TNR_stack(:,1)=mean_TNR'-1.96*std_TNR'/sqrt(ixmax);
TNR_stack(:,2)=1.96*std_TNR'/sqrt(ixmax);
TNR_stack(:,3)=1.96*std_TNR'/sqrt(ixmax);

TPRmin_stack(:,1)=mean_TPRmin'-1.96*std_TPRmin'/sqrt(ixmax);
TPRmin_stack(:,2)=1.96*std_TPRmin'/sqrt(ixmax);
TPRmin_stack(:,3)=1.96*std_TPRmin'/sqrt(ixmax);

TNRmin_stack(:,1)=mean_TNRmin'-1.96*std_TNRmin'/sqrt(ixmax);
TNRmin_stack(:,2)=1.96*std_TNRmin'/sqrt(ixmax);
TNRmin_stack(:,3)=1.96*std_TNRmin'/sqrt(ixmax);

TPRopt_stack(:,1)=mean_TPRopt'-1.96*std_TPRopt'/sqrt(ixmax);
TPRopt_stack(:,2)=1.96*std_TPRopt'/sqrt(ixmax);
TPRopt_stack(:,3)=1.96*std_TPRopt'/sqrt(ixmax);

TNRopt_stack(:,1)=mean_TNRopt'-1.96*std_TNRopt'/sqrt(ixmax);
TNRopt_stack(:,2)=1.96*std_TNRopt'/sqrt(ixmax);
TNRopt_stack(:,3)=1.96*std_TNRopt'/sqrt(ixmax);

AUC_stack1(:,1)=mean_AUC'-2.575*std_AUC'/sqrt(ixmax);
AUC_stack1(:,2)=2.575*std_AUC'/sqrt(ixmax);
AUC_stack1(:,3)=2.575*std_AUC'/sqrt(ixmax);

PPV_stack1(:,1)=mean_PPV'-2.575*std_PPV'/sqrt(ixmax);
PPV_stack1(:,2)=2.575*std_PPV'/sqrt(ixmax);
PPV_stack1(:,3)=2.575*std_PPV'/sqrt(ixmax);

NPV_stack1(:,1)=mean_NPV'-2.575*std_NPV'/sqrt(ixmax);
NPV_stack1(:,2)=2.575*std_NPV'/sqrt(ixmax);
NPV_stack1(:,3)=2.575*std_NPV'/sqrt(ixmax);

TR_stack1(:,1)=mean_TR'-2.575*std_TR'/sqrt(ixmax);
TR_stack1(:,2)=2.575*std_TR'/sqrt(ixmax);
TR_stack1(:,3)=2.575*std_TR'/sqrt(ixmax);

TPR_stack1(:,1)=mean_TPR'-2.575*std_TPR'/sqrt(ixmax);
TPR_stack1(:,2)=2.575*std_TPR'/sqrt(ixmax);
TPR_stack1(:,3)=2.575*std_TPR'/sqrt(ixmax);

TNR_stack1(:,1)=mean_TNR'-2.575*std_TNR'/sqrt(ixmax);
TNR_stack1(:,2)=2.575*std_TNR'/sqrt(ixmax);
TNR_stack1(:,3)=2.575*std_TNR'/sqrt(ixmax);

TPRmin_stack1(:,1)=mean_TPRmin'-2.575*std_TPRmin'/sqrt(ixmax);
TPRmin_stack1(:,2)=2.575*std_TPRmin'/sqrt(ixmax);
TPRmin_stack1(:,3)=2.575*std_TPRmin'/sqrt(ixmax);
```

```
TNRmin_stack1(:,1)=mean_TNRmin'-2.575*std_TNRmin'/sqrt(ixmax);
TNRmin_stack1(:,2)=2.575*std_TNRmin'/sqrt(ixmax);
TNRmin_stack1(:,3)=2.575*std_TNRmin'/sqrt(ixmax);

TPRopt_stack1(:,1)=mean_TPRopt'-2.575*std_TPRopt'/sqrt(ixmax);
TPRopt_stack1(:,2)=2.575*std_TPRopt'/sqrt(ixmax);
TPRopt_stack1(:,3)=2.575*std_TPRopt'/sqrt(ixmax);

TNRopt_stack1(:,1)=mean_TNRopt'-2.575*std_TNRopt'/sqrt(ixmax);
TNRopt_stack1(:,2)=2.575*std_TNRopt'/sqrt(ixmax);
TNRopt_stack1(:,3)=2.575*std_TNRopt'/sqrt(ixmax);

 figure (1)
 bar(AUC_stack,'stacked')
title('AUC')
ylabel('mean AUC (C.I. 95%)')
xlabel('Index')

figure (100)
 bar(AUC_stack1,'stacked')
title('AUC')
ylabel('mean AUC (C.I. 99%)')
xlabel('Index')

figure (2)
 bar(TR_stack,'stacked')
 title('TR')
ylabel('mean TR (C.I. 95%)')
xlabel('Index')

figure (200)
 bar(TR_stack1,'stacked')
 title('TR')
ylabel('mean TR (C.I. 99%)')
xlabel('Index')

  figure (3)
 bar(TPR_stack,'stacked')
 title('TPR')
ylabel('mean TPR (C.I. 95%)')
xlabel('Index')

 figure (300)
 bar(TPR_stack1,'stacked')
 title('TPR')
ylabel('mean TPR (C.I. 99%)')
xlabel('Index')

  figure (4)
 bar(TNR_stack,'stacked')
title('TNR')
ylabel('mean TNR (C.I. 95%)')
xlabel('Index')

figure (400)
 bar(TNR_stack1,'stacked')
title('TNR')
ylabel('mean TNR (C.I. 99%)')
xlabel('Index')

  figure (9)
  bar(PPV_stack,'stacked')
  title('PPV')
  ylabel('mean PPV (C.I. 95%)')
  xlabel('Index')

   figure (900)
  bar(PPV_stack1,'stacked')
```

```
title('PPV')
ylabel('mean PPV (C.I. 99%)')
xlabel('Index')

figure (10)
bar(NPV_stack,'stacked')
title('NPV')
ylabel('mean NPV (C.I. 95%)')
xlabel('Index')

figure (1000)
bar(NPV_stack1,'stacked')
title('NPV')
ylabel('mean NPV (C.I. 99%)')
xlabel('Index')

figure (11)
bar(ifreq_stack,'stacked')
title('Indices Ranking')
ylabel('Frequency')
xlabel('Position')
```

Algorithm 3: Application in imported real data

```
%%% This algorithm corresponds to paragraph 4.3. Results from Application in
Data from ATTICA Study ."%%%
%%% Figure 19 and tables 5-6 are produced by using application code %%%
clear all % Erase the previous data & results
format long % Decimals=16
ix=1; % In application, algorithms are executed once
jflag=1; % ("jflag" is the variable that corresponds to standardization)
n1=716; % Number of non diseased sample cases
n2=265; % Number of diseased sample cases
kmax=5; % Number of component variables
SPSS_MAT=load('htn_paper.txt') % Load of the data - data in txt file - first
comlumn is the vector Y in ascending ordering
Y=SPSS_MAT(:,1) % Vector Y
X=SPSS_MAT(:,2:kmax+1) % Matrix X
if (jflag==1) % Data standardization
    for j=1:kmax % Loop for calculating mean of healthy individuals (Y=0) &
st.dev of total sample
        mu(j)=mean(X(:,j)); % calculating mean of total sample for each
column
        muh(j)=mean(X(1:n1,j)); % calculating mean of healthy individuals
(Y=0) for each column
        sdev(j)=std(X(:,j)); % calculating st.dev of total sample for each
column
        X(:,j)=(X(:,j)-muh(j))/sdev(j); % Matrix X is converted to
standardized X
    end
end
%%%%% Binary Logistic Regression (BLR) %%%%%%
[BLR,dev,stats] = glmfit(X, Y, 'binomial','logit'); % Multivariate Binary
Logistic Regression
coef_m(1:kmax)=BLR(2:kmax+1);
coef_m=exp(coef_m); % Odds Ratio from Multivariate Binary Logistic
Regression
DEV_m=dev; % Deviance from Multivariate Binary Logistic Regression
for k=1:kmax % Loop for the m components
    [BLR,dev,stats] = glmfit(X(:,k), Y, 'binomial','logit'); % Univariate
Binary Logistic Regression
    coef_u(k)=exp(BLR(2)); % Odds Ratio from Univariate Binary Logistic
Regression
    DEV_u(k)=dev; % Deviance from Univariate Binary Logistic Regression
    coef_ds_u(k)=coef_u(k)/DEV_u(k); % Odds Ratio from Univariate Binary
Logistic Regression divided by Deviance
    coef_ds_m(k)=coef_m(k)/DEV_u(k); % Odds Ratio from Multivariate Binary
Logistic Regression divided by Deviance
end
sumcoef_u=sum(coef_u)/kmax; % Sum of UBLR OR divided by kmax
w_ulr=coef_u/sumcoef_u; % Weights from UBLR
sumcoef_m=sum(coef_m)/kmax; % Sum of MBLR OR divided by kmax
w_mlr=coef_m/sumcoef_m; % Weights from MBLR
sumcoef_ds_u=sum(coef_ds_u)/kmax; % Sum of (UBLR OR divided by Deviance
Statistic) divided by kmax
w_ulr_ds=coef_ds_u/sumcoef_ds_u; % Weights from UBLR OR/DS
sumcoef_ds_m=sum(coef_ds_m)/kmax; % Sum of (MBLR OR divided by Deviance
Statistic) divided by kmax
w_mlr_ds=coef_ds_m/sumcoef_ds_m; % Weights from MBLR OR/DS
%%%%% Linear Discriminant Analysis (LDA) %%%%%%
[LDA,err,POSTERIOR,logp,coeff] = classify([X],[X],Y); % Linear Discrimianant
Analysis
[A,B,r,U,V,stats1] = canoncorr(X,Y); % Statistical Output from Multivariate
Linear Discriminant Analysis
coeff_m=coeff(2,1).linear; % Fiscer Coefficient from Linear Discrimianant
Analysis
coeff_m_sign=coeff_m;
coeff_m=abs(coeff_m);% Absolute values of Fisher coefficients - MLDA
LW_m=stats1.Wilks; % Wilks Lambda from multivariate
for k=1:kmax % Loop for the m components
```

```
    [LDA,err,POSTERIOR,logp,coeff] = classify([X(:,k)],[X(:,k)],Y); % Fisher
Coefficient from Linear Discrimianant Analysis
    [A,B,r,U,V,stats1] = canoncorr(X(:,k),Y); % Statistical Output from
Univariate Linear Discriminant Analysis
    coeff_u(k)=coeff(2,1).linear; % Calculating Fisher Coefficient from
Univariate Linear Discrimianant Analysis - saved in coeff_u
    coeff_u_sign(k)=coeff_u(k);
    coeff_u(k)=abs(coeff_u(k)); % Absolute values of Fisher coefficients -
ULDA
    LW_u(k)=stats1.Wilks; % Wilks Lambda for each component
    coeff_lw_u(k)=coeff_u(k)/LW_u(k); % Fisher Coefficient from univariate
Linear Discriminant Analysis divided by Wilks Lamda
    coeff_lw_m(k)=coeff_m(k)/LW_u(k); % Fisher Coefficient from multivariate
Linear Discriminant Analysis divided by Wilks Lamda
end
sumcoeff_u=sum(coeff_u)/kmax; % Sum of ULDA FC divided by kmax
w_ulda=coeff_u/sumcoeff_u; % Weights from ULDA
sumcoeff_m=sum(coeff_m)/kmax; % Sum of MLDA FC divided by kmax
w_mlda=coeff_m/sumcoeff_m; % Weights from MLDA
sumcoeff_lw_u=sum(coeff_lw_u)/kmax; % Sum of (ULDA FC divided by Lambda
Wilks) divided by kmax
w_ulda_lw=coeff_lw_u/sumcoeff_lw_u; % Weights from ULDA FC/LW
sumcoeff_lw_m=sum(coeff_lw_m)/kmax; % Sum of (MLDA FC divided by Lambda
Wilks) divided by kmax
w_mlda_lw=coeff_lw_m/sumcoeff_lw_m; % Weights from MLDA FC/LW
%%%%% Construction of Weights Matrix %%%%%%
w=zeros(kmax,9);% construction of weight matrix
w(:,1)=ones(1,kmax);% first column: ones
w(:,2)=w_ulr; % second column: UBLR weights
w(:,3)=w_mlr; % third column: MBLR weights
w(:,4)=w_ulr_ds; % fourth column: UBLR/DS weights
w(:,5)=w_mlr_ds; % fifth column: UBLR/DS weights
w(:,6)=w_ulda; % sixth column: ULDA weights
w(:,7)=w_mlda; % seventh column: MLDA weights
w(:,8)=w_ulda_lw; % eighth column: ULDA/LW (Lambda of Wilks) weights
w(:,9)=w_mlda_lw; % ninth column: MLDA/LW (Lambda of Wilks)weights
%%%%% Construction of Indices' Matrix %%%%%
T=zeros(n1+n2,9); % Initialazation of indices matrix - Assigning dimensions
(n1+n2)*9
T=X*w; %  Multiplication of matrices X () and w (weights) in order to
produce the indices
for j=1:9 % i is the variable that runs from 1 until 9 that represents the
amount of indices
    [FPR,TPR,Thresholds,AUC(ix,j),OPTROCPT] = perfcurve(Y,T(:,j),1); %
returns the optimal operating point of the ROC curve.
    TNR=1-FPR;
    tpr(:,j)=TPR; % True Positive Ratio for all Indices
    tnr(:,j)=TNR; % True Negative Ratio for all Indices
    fpr(:,j)=FPR; % False Positive Ratio for all Indices
    tnr_tpr=1-FPR+TPR; % One column - sum of TPR and TNR
    max1(j)=max(tnr_tpr); % maximum of Youden with matlab
    max2(j)=tnr_tpr(1); % maximum of Youden - validation
    imax(j)=1;
    for i=2:length(tnr_tpr); % Loop in order to find the location of the
maximum
        if(tnr_tpr(i)>max2(j));
            max2(j)=tnr_tpr(i);
            imax(j)=i; % Position of Indices that corresponds to the Youden
criterion
        end;
    end;
    TMAX(ix,j)=Thresholds(imax(j));% Value of the index that corresponds to
the Jouden threshold
    TNRMAX(ix,j)=TNR(imax(j));% Location of the TNR that corresponds to the
Jouden threshold
    FPRMAX(ix,j)=FPR(imax(j));% Location of the FNR that corresponds to the
Jouden threshold
```

```
    TPRMAX(ix,j)=TPR(imax(j));% Location of the TPR that corresponds to the
Jouden threshold
    idh1=0; % Counter for accuracy from criterion 1
    for idh=1:n1; % Loop for finding True Negative Cases
        if (T(idh,j)< TMAX(ix,j)); % Inequality - criterion
            idh1=idh1+1; % Counter for True Negative Cases
        end;
    end
    TrN(j)=idh1; % Loop for finding True Positive Cases
    for idh=n1+1:n1+n2;
        if (T(idh,j)> TMAX(ix,j));
            idh1=idh1+1; % Counter for True Positive Cases
        end;
    end;
end
Cr1=[TNRMAX' TPRMAX']; % Criterion 1: Maximization of the sum (TPR+TNR)
Sensitivity + Specificity
% Figure - Barchart of indices' TNR %
figure (11) % Figure - Barchart of indices' TNR
bar(TNRMAX,'DisplayName','TNRMAX');figure(gcf)
title('TNRmax')
ylabel('TNR')
xlabel('Index')
% Figure - Barchart of indices' TPR %
figure (12) % Figure - Barchart of indices' TPR
bar(TPRMAX,'DisplayName','TPRMAX');figure(gcf)
title('TPRmax')
ylabel('TPR')
xlabel('Index')
% Figure - Barchart of indices' AUC %
figure (3) % Figure - Barchart of indices' AUC
bar(AUC,'DisplayName','AUC');figure(gcf)
title('AUC')
ylabel('AUC')
xlabel('Index')
% Figure - ROC Curve%
color=['r' 'g' 'b' 'c' 'm' 'y' 'r' 'g' 'b']
figure (4) % Figure - ROC Plots
for j=1:9
    plot(fpr(:,j),tpr(:,j),color(j))
    hold on
    xlabel('False positive rate')
    ylabel('True positive rate')
    title('ROC for Classification by health indices')
end
```

Παράρτημα Α2: Αλγόριθμοι σε περιβάλλον R

Algorithm 1: Data Geaneration & Indices' Evaluation

require(MASS)

library(CCP)

```
# generation methods (either Normal or Unif)

genNorm <- function (n1, n2, kmax) {
  mu_h <- c(110,108,106,104,102,100)
  mu_d <- c(120,122,124,126,128,130)
  sigma_h <- c(30,28,26,24,22,20)
  sigma_d <- sigma_h
  sapply(1:kmax, function(k) {
        X_h <- rnorm(n1, mu_h[k], sigma_h[k])
        X_d <- rnorm(n2, mu_d[k], sigma_d[k])
        X_k <- c(X_h, X_d)
        (X_k - mean(X_h))/sd(X_k) # Data normalization (maybe use scale instead?)
     })
}

genUnif <- function(n1, n2, kmax) {
  r <- 100 # range
  a_h <- c(50,48,46,44,42,40)
  a_d <- c(70,72,74,76,78,80)
  b_h <- a_h + r
  b_d <- a_d + r
  X_h <- sapply(1:kmax, function(i) { rusnif(n1, a_h[i], b_h[i]) })
  X_d <- sapply(1:kmax, function(i) { runif(n2, a_d[i], b_d[i]) })
  rbind(X_h, X_d)
}

generateY <- function(n1,n2,kmax) {
  rbind(matrix(0, n1, 1), matrix(1, n2, 1))
}

# selection method to generate X
generateX <- genNorm
# generate <- genUnif

n1 <- 75
n2 <- 25
kmax <- 2
```

```
X <- generateX(n1,n2,kmax)
Y <- generateY(n1, n2, kmax)

#BLR
blr <- function(X, Y) {
  model <- glm(Y~X, family=binomial(link=logit))
  coeff_m <- exp(coefficients(model)[-1])
  dev_m <- residuals(model, type="deviance")

  model_u <- apply(X, 2, function(X_k) { glm(Y~X_k, family=binomial(link=logit)) })
  coeff_u <- sapply(model_u, function(m) { exp(coefficients(m)[2])})
  dev_u   <- sapply(model_u, function(m) { residuals(m, type="deviance")})

  coeff_ds_u <- coeff_u / dev_u
  coeff_ds_m <- coeff_m / dev_u

  w_ulr    <- coeff_u/(sum(coeff_u)/kmax)        # Weights from UBLR OR
  w_mlr    <- coeff_m/(sum(coeff_m)/kmax)        # Weights from MBLR OR
  w_ulr_ds <- coeff_ds_u/(sum(coeff_ds_u)/kmax)    # Weights from UBLR OR/DS
  w_mlr_ds <- coeff_ds_m/(sum(coeff_ds_m)/kmax)    # Weights from MBLR OR/DS

  list(w_ulr=w_ulr, w_mlr=w_mlr, w_ulr_ds=w_ulr_ds, w_mlr_ds=w_mlr_ds)
}

blr <- blr(X,Y)

lda_model <- lda(Y~X)
coeff_m <- abs(coefficients(lda_model)[2,1])
LW_m <- p.asym(cancor(X,Y)$cor, dim(X)[1], dim(X)[2], dim(Y)[2], tstat = "Wilks")$stat
```

Παράρτημα Α3: Αλγόριθμοι στο λογισμικό SPSS (Syntax)

```
COMPUTE Z1=(X1-25.0889)/13.65355.
EXECUTE.
COMPUTE Z2=(X2-32.9073)/13.96035.
EXECUTE.
COMPUTE Z3=(X3-11.5132)/7.32894.
EXECUTE.
COMPUTE Z4=(X4-5.0171)/2.26793.
EXECUTE.
COMPUTE Z5=(X5-4.6006)/2.54201.
EXECUTE.
COMPUTE T0=X1+X2+X3+X4+X5.
EXECUTE.
COMPUTE T1=1.08943907*X1+1.03735806*X2+1.08136719*X3+0.84983073*X4+0.94200495*X5.
EXECUTE.
COMPUTE T2=1.15376728*X1+1.03250676*X2+1.16020932*X3+0.73115472*X4+0.92236192*X5.
EXECUTE.
COMPUTE T3=1.11470926*X1+1.02459200*X2+1.07365323*X3+0.85332297*X4+0.93372253*X5.
EXECUTE.
COMPUTE T4=1.18037039*X1+1.01966286*X2+1.15177757*X3+0.73406027*X4+0.91412891*X5.
EXECUTE.
COMPUTE T5=-1.277870293*X1+0.268161465*X2+-1.051115692*X3+4.769503044*X4+2.291321476*X5.
EXECUTE.
COMPUTE T6=-1.100262998*X1+0.595335654*X2+-1.12428489*X3+4.540014575*X4+2.089197659*X5.
EXECUTE.
COMPUTE T7=-1.289036882*X1+0.267963939*X2+-1.051908594*X3+4.782114773*X4+2.290866763*X5.
EXECUTE.
COMPUTE T8=-1.109670204*X1+0.594786071*X2+-1.124922764*X3+4.55116897*X4+2.088637928*X5.
EXECUTE.
COMPUTE Tz0=Z1+Z2+Z3+Z4+Z5.
EXECUTE.
COMPUTE Tz1=1.78306653*Z1+0.91047646*Z2+1.32052060*Z3+0.65175426*Z4+0.79870043*Z5.
EXECUTE.
COMPUTE Tz2=2.12687839*Z1+0.46043674*Z2+1.46806018*Z3+0.36033800*Z4+0.58428670*Z5.
EXECUTE.
COMPUTE Tz3=1.82442583*Z1+0.89927185*Z2+1.31110063*Z3+0.65443254*Z4+0.79167799*Z5.
EXECUTE.
COMPUTE Tz4=2.163431494*Z1+0.452099539*Z2+1.449027165*Z3+0.35969375*Z4+0.575748052*Z5.
EXECUTE.
COMPUTE Tz5=18.30432665*Z1-3.927480584*Z2+8.081892262*Z3-11.34812594*Z4-6.110612381*Z5.
EXECUTE.
```

```
COMPUTE Tz6=-1.02620892425552*Z1+0.56774313783707*Z2-0.56287405975476*Z3+0.70336473053580*Z4+0.36278596688663*Z5.
EXECUTE.
COMPUTE Tz7=17.96084921*Z1-3.817583731*Z2+7.867469729*Z3-11.0679083*Z4-5.942826915*Z5.
EXECUTE.
COMPUTE Tz8=-140.5300951*Z1+77.01705821*Z2-76.47045986*Z3+95.73757788*Z4+49.24591884*Z5.
EXECUTE.

COMPUTE Tz5NN=1.915783188*Z1+0.411061353*Z2+0.845873963*Z3+1.187727324*Z4+0.639554172*Z5.
EXECUTE.
COMPUTE Tz6NN=1.592020331*Z1+0.880774467*Z2+0.87322077*Z3+1.091172494*Z4+0.562811939*Z5.
EXECUTE.
COMPUTE Tz7NN=1.924790344*Z1+0.409114748*Z2+0.843124375*Z3+1.186102214*Z4+0.63686832*Z5.
EXECUTE.
COMPUTE Tz8NN=1.600566512*Z1+0.877185234*Z2+0.870959755*Z3+1.090402458*Z4+0.56088604*Z5.
EXECUTE.

FREQUENCIES VARIABLES=T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9
  /FORMAT=NOTABLE
  /STATISTICS=MEDIAN
  /ORDER=ANALYSIS.

RECODE T_1 (Lowest thru 69.3777965=0) (ELSE=1) INTO Sen_T_1_Q3.
VARIABLE LABELS  Sen_T_1_Q3 'Sensitivity T_1_Q3'.
EXECUTE.
RECODE T_2 (Lowest thru 69.2518854=0) (ELSE=1) INTO Sen_T_2_Q3.
VARIABLE LABELS  Sen_T_2_Q3 'Sensitivity T_2_Q3'.
EXECUTE.
RECODE T_3 (Lowest thru 69.2671255=0) (ELSE=1) INTO Sen_T_3_Q3.
VARIABLE LABELS  Sen_T_3_Q3 'Sensitivity T_3_Q3'.
EXECUTE.
RECODE T_4 (Lowest thru 69.2892847=0) (ELSE=1) INTO Sen_T_4_Q3.
VARIABLE LABELS  Sen_T_4_Q3 'Sensitivity T_4_Q3'.
EXECUTE.
RECODE T_5 (Lowest thru 69.2862027=0) (ELSE=1) INTO Sen_T_5_Q3.
VARIABLE LABELS  Sen_T_5_Q3 'Sensitivity T_5_Q3'.
EXECUTE.
RECODE T_6 (Lowest thru 69.3027961=0) (ELSE=1) INTO Sen_T_6_Q3.
VARIABLE LABELS  Sen_T_6_Q3 'Sensitivity T_6_Q3'.
EXECUTE.
RECODE T_7 (Lowest thru 69.3277735=0) (ELSE=1) INTO Sen_T_7_Q3.
VARIABLE LABELS  Sen_T_7_Q3 'Sensitivity T_7_Q3'.
EXECUTE.
RECODE T_8 (Lowest thru 69.2492887=0) (ELSE=1) INTO Sen_T_8_Q3.
VARIABLE LABELS  Sen_T_8_Q3 'Sensitivity T_8_Q3'.
```

```
EXECUTE.
RECODE T_9 (Lowest thru 69.2603880=0) (ELSE=1) INTO Sen_T_9_Q3.
VARIABLE LABELS  Sen_T_9_Q3 'Sensitivity T_9_Q3'.
EXECUTE.

RECODE T_1 (Lowest thru 69.3777965=1) (ELSE=0) INTO Spe_T_1_Q3.
VARIABLE LABELS  Spe_T_1_Q3 'Specificity T_1_Q3'.
EXECUTE.
RECODE T_2 (Lowest thru 69.2518854=1) (ELSE=0) INTO Spe_T_2_Q3.
VARIABLE LABELS  Spe_T_2_Q3 'Specificity T_2_Q3'.
EXECUTE.
RECODE T_3 (Lowest thru 69.2671255=1) (ELSE=0) INTO Spe_T_3_Q3.
VARIABLE LABELS  Spe_T_3_Q3 'Specificity T_3_Q3'.
EXECUTE.
RECODE T_4 (Lowest thru 69.2892847=1) (ELSE=0) INTO Spe_T_4_Q3.
VARIABLE LABELS  Spe_T_4_Q3 'Specificity T_4_Q3'.
EXECUTE.
RECODE T_5 (Lowest thru 69.2862027=1) (ELSE=0) INTO Spe_T_5_Q3.
VARIABLE LABELS  Spe_T_5_Q3 'Specificity T_5_Q3'.
EXECUTE.
RECODE T_6 (Lowest thru 69.3027961=1) (ELSE=0) INTO Spe_T_6_Q3.
VARIABLE LABELS  Spe_T_6_Q3 'Specificity T_6_Q3'.
EXECUTE.
RECODE T_7 (Lowest thru 69.3277735=1) (ELSE=0) INTO Spe_T_7_Q3.
VARIABLE LABELS  Spe_T_7_Q3 'Specificity T_7_Q3'.
EXECUTE.
RECODE T_8 (Lowest thru 69.2492887=1) (ELSE=0) INTO Spe_T_8_Q3.
VARIABLE LABELS  Spe_T_8_Q3 'Specificity T_8_Q3'.
EXECUTE.
RECODE T_9 (Lowest thru 69.2603880=1) (ELSE=0) INTO Spe_T_9_Q3.
VARIABLE LABELS  Spe_T_9_Q3 'Specificity T_9_Q3'.
EXECUTE.

RECODE T_1 (Lowest thru 67.8669963=0) (ELSE=1) INTO Sen_T_1_opt.
VARIABLE LABELS  Sen_T_1_opt 'Sensitivity T_1_opt'.
EXECUTE.
RECODE T_2 (Lowest thru 67.6353042=0) (ELSE=1) INTO Sen_T_2_opt.
VARIABLE LABELS  Sen_T_2_opt 'Sensitivity T_2_opt'.
EXECUTE.
RECODE T_3 (Lowest thru 68.1175961=0) (ELSE=1) INTO Sen_T_3_opt.
VARIABLE LABELS  Sen_T_3_opt 'Sensitivity T_3_opt'.
EXECUTE.
RECODE T_4 (Lowest thru 67.7746397=0) (ELSE=1) INTO Sen_T_4_opt.
VARIABLE LABELS  Sen_T_4_opt 'Sensitivity T_4_opt'.
EXECUTE.
```

```
RECODE T_5 (Lowest thru 67.8124823=0) (ELSE=1) INTO Sen_T_5_opt.
VARIABLE LABELS  Sen_T_5_opt 'Sensitivity T_5_opt'.
EXECUTE.
RECODE T_6 (Lowest thru 67.8789820=0) (ELSE=1) INTO Sen_T_6_opt.
VARIABLE LABELS  Sen_T_6_opt 'Sensitivity T_6_opt'.
EXECUTE.
RECODE T_7 (Lowest thru 67.7669335=0) (ELSE=1) INTO Sen_T_7_opt.
VARIABLE LABELS  Sen_T_7_opt 'Sensitivity T_7_opt'.
EXECUTE.
RECODE T_8 (Lowest thru 68.2513100=0) (ELSE=1) INTO Sen_T_8_opt.
VARIABLE LABELS  Sen_T_8_opt 'Sensitivity T_8_opt'.
EXECUTE.
RECODE T_9 (Lowest thru 68.3848841=0) (ELSE=1) INTO Sen_T_9_opt.
VARIABLE LABELS  Sen_T_9_opt 'Sensitivity T_9_opt'.
EXECUTE.

RECODE T_1 (Lowest thru 67.8669963=1) (ELSE=0) INTO Spe_T_1_opt.
VARIABLE LABELS  Spe_T_1_opt 'Specificity T_1_opt'.
EXECUTE.
RECODE T_2 (Lowest thru 67.6353042=1) (ELSE=0) INTO Spe_T_2_opt.
VARIABLE LABELS  Spe_T_2_opt 'Specificity T_2_opt'.
EXECUTE.
RECODE T_3 (Lowest thru 68.1175961=1) (ELSE=0) INTO Spe_T_3_opt.
VARIABLE LABELS  Spe_T_3_opt 'Specificity T_3_opt'.
EXECUTE.
RECODE T_4 (Lowest thru 67.7746397=1) (ELSE=0) INTO Spe_T_4_opt.
VARIABLE LABELS  Spe_T_4_opt 'Specificity T_4_opt'.
EXECUTE.
RECODE T_5 (Lowest thru 67.8124823=1) (ELSE=0) INTO Spe_T_5_opt.
VARIABLE LABELS  Spe_T_5_opt 'Specificity T_5_opt'.
EXECUTE.
RECODE T_6 (Lowest thru 67.8789820=1) (ELSE=0) INTO Spe_T_6_opt.
VARIABLE LABELS  Spe_T_6_opt 'Specificity T_6_opt'.
EXECUTE.
RECODE T_7 (Lowest thru 677669335=1) (ELSE=0) INTO Spe_T_7_opt.
VARIABLE LABELS  Spe_T_7_opt 'Specificity T_7_opt'.
EXECUTE.
RECODE T_8 (Lowest thru 68.2513100=1) (ELSE=0) INTO Spe_T_8_opt.
VARIABLE LABELS  Spe_T_8_opt 'Specificity T_8_opt'.
EXECUTE.
RECODE T_9 (Lowest thru 68.,3848841=1) (ELSE=0) INTO Spe_T_9_opt.
VARIABLE LABELS  Spe_T_9_opt 'Specificity T_9_opt'.
EXECUTE.

COMPUTE filter_$=(Clinical_Status = 1).
```

```
VARIABLE LABELS filter_$ 'Clinical_Status = 1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_1
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_2
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_3
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_4
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .
```

```
NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_5
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_6
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_7
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_8
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .

NOMREG
Clinical_status (BASE=FIRST ORDER=ASCENDING) WITH T_9
/CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20) LCONVERGE(0)
PCONVERGE(0.000001) SINGULAR(0.00000001)
```

```
/MODEL
/STEPWISE = PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
/INTERCEPT =INCLUDE
/PRINT = PARAMETER SUMMARY LRT CPS STEP MFI association .
```

Παράρτημα Α4: Αλγόριθμοι στο λογισμικό Mathematica

Ευαισθησία σύνθετου δείκτη με συνιστώσες μεταβλητές που ακολουθούν ομοιόμορφη κατανομή

```
Table[1 − (1/((b^m) ∗ (m)!)) ∗ Sum[((−1)^r) ∗ Binomial[m, r] ∗ (c − b
∗ r)^(m), {r, 0, Floor[c/b]}], {m, 2,2}, {b, 1,10}, {c, 2,2,0.5}]
```

Διαφορά Ευαισθησίας μεταξύ 2 σύνθετων δεικτών με συνιστώσες μεταβλητές που ακολουθούν ομοιόμορφη κατανομή

```
Table[1/(m)! ∗ ((1/b1^m) ∗ Sum[(−1)^r ∗ Binomial[m, r] ∗ (c − b1 ∗ r)^(m), {r, 0, Floor[c/b1]}]
− (1/b2^m) ∗ Sum[(−1)^r ∗ Binomial[m, r] ∗ (c − b2
∗ r)^(m), {r, 0, Floor[c/b2]}]), {c, 1, m = 15,0.5}, {b1,1,1}, {b2,1.1,1.1}]
```

ΠΑΡΑΡΤΗΜΑ Β

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

ΠΡΟΣΩΠΙΚΑ ΣΤΟΙΧΕΙΑ:

ΕΠΩΝΥΜΟ	: Μπερσίμης
ΟΝΟΜΑ	: Φραγκίσκος
ΗΜΕΡΟΜΗΝΙΑ ΓΕΝΝΗΣΗΣ	: 18/02/1978
ΤΟΠΟΣ ΓΕΝΝΗΣΗΣ	: Αθήνα Αττικής
ΟΙΚ. ΚΑΤΑΣΤΑΣΗ	: Έγγαμος
ΟΝΟΜΑ ΠΑΤΡΟΣ	: Γεράσιμος
ΟΝΟΜΑ ΜΗΤΡΟΣ	: Μαρία
ΔΙΕΥΘΥΝΣΗ ΚΑΤΟΙΚΙΑΣ	: Μαυρομιχάλη 17 Καματερό, ΤΚ 13451
ΤΗΛ	: 210 2320683, 697 4015238
E-mail	: fbersim@hua.gr

ΣΠΟΥΔΕΣ & ΤΙΤΛΟΙ

2019 Υποψήφιος Διδάκτωρ, Τμήμα Πληροφορικής και Τηλεματικής Χαροκόπειου Πανεπιστημίου Αθηνών. Θέμα Διδακτορικής Διατριβής: «ΔΙΕΡΥΝΗΣΗ ΕΥΑΙΣΘΗΣΙΑΣ ΚΑΙ ΕΙΔΙΚΟΤΗΤΑΣ ΣΥΝΘΕΤΩΝ ΚΛΙΜΑΚΩΝ ΑΠΟΤΕΛΟΥΜΕΝΩΝ ΑΠΟ ΔΙΑΚΡΙΤΕΣ Ή ΣΥΝΕΧΕΙΣ ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΕ ΜΟΝΤΕΛΑ ΠΡΟΣΟΜΟΙΩΣΕΩΝ»

Τριμελής Συμβουλευτική Επιτροπή:

- Βαμβακάρη Μ., Αναπληρώτρια Καθηγήτρια στο γνωστικό αντικείμενο Στατιστική-Θεωρία Πιθανοτήτων και Εφαρμογές (επιβλέπουσα), Τμήμα Πληροφορικής & Τηλεματικής, Χαροκόπειο Πανεπιστήμιο
- Ανδρέας Κυριακούσης, Ομότιμος Καθηγητής, Τμήμα Πληροφορικής & Τηλεματικής, Χαροκόπειο Πανεπιστήμιο
- Παναγιωτάκος Δ., Αναπληρωτής Καθηγητής Βιοστατιστικής – Επιδημιολογίας, Τμήμα Επιστήμης Διαιτολογίας – Διατροφής, Χαροκόπειο Πανεπιστήμιο

2009 Μεταπτυχιακό Δίπλωμα Ειδίκευσης:
«Εφαρμοσμένη Στατιστική»
Τμήμα Στατιστικής, Οικονομικό Πανεπιστήμιο Αθηνών, κατεύθυνση «Εφαρμοσμένη Στατιστική για Εκπαιδευτικούς & Στελέχη Επιχειρήσεων - Οργανισμών»
Βαθμός: 9,14 - Άριστα

- *Θέμα Μεταπτυχιακής Διατριβής*: «Στατιστικός Έλεγχος Διεργασιών».
  Επιβλέπον μέλος: Ψαράκης Στέλιος

2003 Πτυχίο Μαθηματικού.
Τμήμα Θετικών Επιστημών, ΕΚΠΑ. Βαθμός: 6,83 - Καλώς

ΞΕΝΕΣ ΓΛΩΣΣΕΣ

- ΑΓΓΛΙΚΑ: Πολύ καλή γνώση, Κρατικό Πιστοποιητικό Γ1 (Μάιος 2012) & First certificate in English, university of Cambridge (Δεκέμβριος 1993).
- ΙΤΑΛΙΚΑ; Certificato Intermedio Di Lingua Italiana –Palso (Ιούλ.1992).
- ΓΕΡΜΑΝΙΚΑ: Μέτρια γνώση, Πιστοποιήσεις Παρακολούθησης ΚΔΒΜ (2015)
-

ΗΛΕΚΤΡΟΝΙΚΟΙ ΥΠΟΛΟΓΙΣΤΕΣ

Πολύ καλή γνώση:

- Παρακολούθηση Επίπεδου Β (5-6/2017).
- Πιστοποίηση Δεξιοτήτων & Γνώσεων στις Τεχνολογίες Πληροφορίας & Επικοινωνιών (Επίπεδο Α) (14/10/2011).
- Πιστοποιητικό επιμόρφωσης του εκπαιδευτικού προγράμματος «Εισαγωγή σε Υπολογιστικά Φύλλα – Παρουσιάσεις – Βάσεις Δεοδομένων» (25/06/2010).
- Windows XP, Word, Excel, Powerpoint, Internet
- Στατιστικά προγράμματα: SPSS, Minitab, R
- Μαθηματικά προγράμματα: Mathematica, Matlab, Geogebra

ΒΡΑΒΕΥΣΕΙΣ

- Τίτλος Υψηλής Ακαδημαϊκής Επίδοσης στο Μεταπτυχιακό Πρόγραμμα Ειδίκευσης στη Στατιστική Επιστήμη – Τμήμα Στατιστικής – Οικονομικό Πανεπιστήμιο Αθήνας (2004-2005)

ΥΠΟΤΡΟΦΙΕΣ / ΔΙΑΚΡΙΣΕΙΣ

- Υπότροφος του Ιδρύματος Κρατικών Υποτροφιών (ΙΚΥ) για την απόκτηση Μεταπτυχιακού Διπλώματος – Χρηματικό Έπαθλο (2003–2004)

ΣΥΜΜΕΤΟΧΗ ΣΕ ΣΥΝΕΔΡΙΑ

- 5-7/05/2017: 3ο Πανελλήνιο Επιστημονικό Συνέδριο με Διεθνή Συμμετοχή: «Ανθρωπιστικές Επιστήμες, Εκπαίδευση, Κοινωνία και Πολιτική Παιδεία», Ηράκλειο, Κρήτη , Ελλάδα
- 4-6/11/2016: 33ο Πανελλήνιο Συνέδριο Μαθηματικών: «Τα Μαθηματικά στην Εκπαίδευση, στη Τεχνολογία και στην Κοινωνία», Χανιά, Κρήτη, Ελλάδα
- 21-23/10/16: 2ο Πανελλήνιο Συνέδριο για την Προώθηση της Εκπαιδευτικής Καινοτομίας, Λάρισα, Ελλάδα
- 17-20/09/16: The 38th ESPEN Conference - Nutritional Epidemiology , Copenhagen, Denmark
- 1-3/07/16: 1ο Πανελλήνιο Συνέδριο με θέμα «Ελλάδα - Ευρώπη 2020: Εκπαίδευση, Δια Βίου Μάθηση, Έρευνα, Καινοτομία & Οικονομία», Αθήνα, Ελλάδα
- 1-4/06/16: The 4th Stochastic Modeling Techniques and Data Analysis International Conference and 5th Demographics 2016 Workshop, Valletta, Malta
- 27/05/16: 1ο Πανελλήνιο Συνέδριο νέων κατόχων μεταπτυχιακών διπλωμάτων ειδίκευσης και διδακτορικών διπλωμάτων στα μαθηματικά, Ιωάννινα, Ελλάδα
- 5-6/12/16: 1ο Πανελλήνιο προ-Συνέδριο με θέμα «Ελλάδα - Ευρώπη 2020: Εκπαίδευση, Δια Βίου Μάθηση, Έρευνα, Καινοτομία & Οικονομία», Λαμία, Ελλάδα
- 5-7/11/15: International Conference "Science in Technology" 2015, Athens, Greece
- 30/10-1/11/15: 32ο Πανελλήνιο Συνέδριο Μαθηματικών με θέμα «Τα Μαθηματικά στην Εκπαίδευση, στη Τεχνολογία και στην Κοινωνία», Καστοριά, Ελλάδα
- 13-15/11/15: 1ο Πανελλήνιο Συνέδριο για την Προώθηση της Εκπαιδευτικής Καινοτομίας, Λάρισα, Ελλάδα
- 11-15/5/15 The 8th Conference of Eastern Mediterranean Region International Biometric Society, Nevsehir, Cappadocia, Turkey
- 5-8/04/15: 28° Πανελλήνιο Συνέδριο Στατιστικής, με θέμα «Χωρική Στατιστική και εφαρμογές της Στατιστικής στη Γενετική και την Πληροφορική», Αθήνα, Ελλάδα
- 7-9/11/14: 31ο Πανελλήνιο Συνέδριο Μαθηματικών με θέμα «Προκλήσεις και Προοπτικές της Μαθηματικής Εκπαίδευσης και Έρευνας στη διεθνοποιημένη δικτυακή εποχή», Βέροια, Ελλάδα
- 18-22/04/14: 27° Πανελλήνιο Συνέδριο Στατιστικής, με θέμα «Ανάλυση Μεγάλων Δεδομένων – Αναλυτική των Επιχειρήσεων», Θεσσαλονίκη, Ελλάδα
- 11-12/04/14: Πανελλήνιο Συνέδριο Πρότυπων Γυμνασίων – Λυκείων Αθήνα με θέμα «Ανάλυση Μεγάλων Δεδομένων – Αναλυτική των Επιχειρήσεων» Αθήνα, Ελλάδα
- 8-10/10/13: 30ο Πανελλήνιο Συνέδριο Μαθηματικών με θέμα «Τα Μαθηματικά στην Εκπαίδευση, στη Τεχνολογία και στην Κοινωνία», Καρδίτσα, Ελλάδα
- 9-11/07/13: 3rd International Symposium on Statistical Process Control, Piraeus, Greece
- 17/06/13: One Day Workshop on Probability and Statistics, Department of Mathematics in University of Athens
- 5-8/06/12: The 2nd Stochastic Modeling Techniques and Data Analysis International Conference, Chania, Crete, Greece
- 18-22/04/12: 25° Πανελλήνιο Συνέδριο Στατιστικής, με θέμα «Στατιστική στις Βιοεπιστήμες», Βόλος, Ελλάδα
- 8-10/10/12: 29ο Πανελλήνιο Συνέδριο Μαθηματικών με θέμα «Μαθηματική Μοντελοποίηση: Εφαρμογές στις Επιστήμες, την Τεχνολογία και την Εκπαίδευση», Αθήνα, Ελλάδα

ΣΥΓΓΡΑΦΙΚΟ ΕΡΓΟ

Άρθρα δημοσιευμένα σε Διεθνή ή Ελληνικά επιστημονικά περιοδικά (με κριτές)

1. Michalis Matthaios Lygkiaris, **Fragiskos Gerasimos Bersimis**, André Thomas. ARTe Mecenas: In the Shoes of a Medici, Interactive Mobile Communication Technologies and Learning

Educational games aimed at teaching human beings are often structured as quiz games, not differing from a typical examination. However, such as their mechanics, interactive environments, character portrayal and story. This paper explores the concepts above, as applied in the educational game "ARTé: Mecenas" by Triseum, as well as their effects in a small sample of people who played the game. The guided pilot study focused on the avatar identification process, which was found to be positive

2. Kalliopi Anna Poulia, Meropi Kontogianni, **Fragkiskos G. Bersimis**, S. Kosak, Karin Schindler, M. Hiesmayr, Michael Chourdakis. Region based Differences in Nutritional Support Practices in Eastern Europe. Preliminary Results from analyzing nutritionday's Database, MON-P191. 2016 Sep

Rationale: Hospital food consumption and practices of nutritional support (NS) can affect patients' outcome. In the present study the nutritionDay (nD) database was used to explore differences in hospital food and NS practices between 8 different geographical regions, participating in nD worldwide. Methods: Data from 113,930 adult patients (male 49.9%; mean age 64.0 ± 18.1 yrs, mean BMI 25.7 ± 6.0 kg/m^2) between 2006–2013 (in 4329 units, 1358 hospitals, 54 countries) were included. The nD methodology has been previously described [1]. Results: 14.3% ate nothing (9.3% although allowed), with the highest % recorded in Eastern Europe(EE) and the lowest in Central Europe(CE) ($p < 0.01$). The highest % of patients "not allowed to eat" was reported in EE (8.6% vs 5.0% in the whole sample), as well as the highest response in "missed my meal to examination/surgery". Being in EE or Middle East was also associated with higher odds of not consuming the whole meal in multivariate logistic regression model compared to CE (reference region) ($p < 0.001$). Finally, the highest % of food consumption apart from that served at hospital was also recorded in EE (73.4% vs 44.2% in the whole sample, $p < 0.01$), along with the lowest number of dietitians per patient per unit. Conclusion: Lowest consumption of hospital food and highest consumption of non hospital food were reported in EE. This could be indicative of a different organization model, allowing the provision of food other than the one served at hospital. As these practices could intervene to the NS plan, enforcement o he role of NS teams and an increase in the personnel dedicated to nutrition, seems to be necessary for the improvement of nutritional care in hospitals

3. Michael Chourdakis, Meropi Kontogianni, Kalliopi Anna Poulia, **Fragkiskos G. Bersimis**, Karin Schindler, S. Kosak, Demosthenes Panagiotakos, M. Hiesmayr,. Institutional Factors Affecting Food Consumption during Hospitalization. Preliminary Results from analyzing nutritionday's Database, MON-P193. 2016 Sep

Rationale:Disease related malnutrition is associated with prolonged hospital stay, increased morbidity and mortality and subsequent increase in the cost of health care. Low food intake during hospitalization, is an independent factor for mortality and prolonged hospital stay. The nutritionDay database was used to explore prevalence of decreased food consumption and institutional factors affecting it during hospitalization. Methods: Data from 113,930 adult patients (females:50.1%; mean age 64.0 ± 18.1 yrs, mean BMI 25.7 ± 6.0 kg/m2) obtained between 2006– 2013 (4329 units, 1358 hospitals, 54 countries) were analyzed. The nutritionDay methodology has been previously described. Results: 41.6% reported to have consumed all their served meal, 29.5% half of it, 14.6% a quarter and 9.3% ate nothing although allowed to eat. Common reasons for reduced meal consumption were "bad taste/smell of the served food" (17.1%) and "skipping a meal due to examination/surgery" (7%). These reasons remained important in binary logistic regression models and after adjustment for several confounders (all $p < 0.001$). Furthermore, 44.2% of the patients (ranging from 34.6% in N. Europe and America to 73.4% in E. Europe) reported eating food apart from that served in the hospital. Conclusion: Skipping a meal due to medical procedures and the provision of low quality hospital food were detected as important factors for low food intake during hospitalization. These findings support the importance of both protected meal times and of the improved palatability of hospital food for enhancing dietary intake during hospital stay.

4. Eleni Karfopoulou, Dora Brikou, Eirini Mamalaki, **Fragkiskos G. Bersimis**, Costas A. Anastasiou, James O. Hill, Mary Yannakoulia. Dietary patterns in weight loss maintenance: results from the MedWeight study, European Journal of Nutrition. 2016 Jan 07

The dietary habits contributing to weight loss maintenance are not sufficiently understood. We studied weight loss maintainers in comparison with regainers, to identify the differentiating behaviors. The MedWeight study is a Greek

registry of weight loss maintainers and regainers. Participants had intentionally lost ≥10 % of their weight and either had maintained this loss for over a year, or had regained weight. Questionnaires on demographics and lifestyle habits were completed online. Dietary assessment was carried out by two telephone 24-h recalls. Present analysis focused on 361 participants (32 years old, 39 % men): 264 maintainers and 97 regainers. Energy and macronutrient intake did not differ by maintenance status (1770 ± 651 kcal in maintainers vs. 1845 ± 678 kcal in regainers, p = 0.338), although protein intake per kg of body weight was higher in maintainers (1.02 ± 0.39 vs. 0.83 ± 0.28 g/kg in regainers, p < 0.001). Physical activity energy expenditure was greater for maintainers in men (by 1380 kcal per week, p = 0.016), but not women. Salty snacks, alcohol and regular soda were more frequently consumed by men regainers. Principal component analysis identified a healthy dietary pattern featuring mainly unprocessed cereal, fruit, vegetables, olive oil and low-fat dairy. Male maintainers were 4.6 times more likely to follow this healthy pattern compared to regainers (OR 4.6, 95 % CI 2.0–11.0). No similar finding was revealed in women. Other characteristics of maintainers but not of regainers were: involvement in meal preparation and eating at home for men, and a higher eating frequency and slower eating rate for women. Men maintaining weight loss were much more likely to adhere to a healthy eating pattern. Eating at home, involvement in meal preparation, higher eating frequency and slower eating rate were also associated with maintenance. These lifestyle habits of successful maintainers provide target behaviors to improve obesity treatment.

5. Sakelli M. Despoina, **Fragkiskos G. Bersimis**. Attitudes of students and teachers towards the course "Ancient Greek Language" of the first grade of Secondary school, MIBES Journal. 2015 Oct 23

The contribution of teaching ancient Greek language to humanitarian education constituted the reason why we started studying the effectiveness of the way of teaching ancient Greek language in the first grade of Secondary School. During the first phase of the study (preliminary work) mismatch points were identified between the textbook and the Curriculum rendering the investigation of the students' and teachers' attitudes towards the textbook and the course as a whole, with a main goal to draw conclusions and formulate a proposal aimed at improving the effectiveness of the way of teaching the ancient Greek language. This paper presents part of the initial indicative results of the research which is being carried out that confirms to a certain degree the findings of the preliminary study on the correspondence between the Curriculum and the textbook of the course. The results of the research will offer the basis for formulating an alternative proposal to upgrade the way of teaching the course.

6. Βασίλης Καρκάνης, Γιώργος Κόσυβας, **Φραγκίσκος Γ. Μπερσίμης**. Αξιολόγηση και βελτίωση της βαθμολογικής διαδικασίας κατά τη διόρθωση μαθηματικών γραπτών στις πανελλήνιες εξετάσεις, ΕΡΚΥΝΑ. 2015 (7):5-24

Σε αυτή την εργασία παρουσιάζεται έρευνα, η οποία πραγματοποιήθηκε στο 31ο Βαθμολογικό Κέντρο στο πλαίσιο των πανελλήνιων εξετάσεων του έτους 2014. Στην έρευνα αυτή συμμετείχαν 37 βαθμολογητές ειδικότητας μαθηματικού, οι οποίοι μετά τις αρχικές συζητήσεις πάνω στην πειραματική βαθμολόγηση, την παρουσίαση κρίσιμων λαθών και την κατάρτιση του Οδηγού Αναλυτικής Μοριοδότησης (Ο.Α.Μ.), διόρθωσαν γραπτά Μαθηματικών και στοτέλος απάντησαν στις ερωτήσεις ενός κατάλληλα δομημένου ερωτηματολογίου. Στην παρούσα έρευνα ανιχνεύτηκαν τέσσερις παράγοντες οι οποίοι εκφράζουν τις κύριες πεποιθήσεις των βαθμολογητών. Τα ευρήματα παρέχουν ενδείξεις ότι η πειραματική βαθμολόγηση και η συλλογική διαμόρφωση του Ο.Α.Μ. είναι σημαντικοί παράγοντες για τη βελτίωση της διαδικασίας βαθμολόγησης των μαθηματικών καθώς επίσης ότι συμβάλλουν θετικά στην αναβάθμιση του εκπαιδευτικού έργου των βαθμολογητών.

7. Panagiotakos Demosthenes, **Bersimis G. Fragkiskos**. An index that rates relevance of scientific work in bio-sciences: The scientific relevance-index, Journal of Statistics Applications & Probability. 2014 Oct 25

This work aims to present an objective index that rates scientific relevance (SR) of scientists′ published work, the SR index, in a specific thematic field. The proposed index is calculated based on equally or weighted individual parameters that measure scientist′s publication record and recognition, using easily accessible and unbiased data from existing bibliometric databases. The application of the SR-index could be in any scientific field; an example is given here on cell and molecular biology field.

Ανακοινώσεις σε Διεθνή ή Ελληνικά συνέδρια (με κριτές) & δημοσίευση εργασιών στα πρακτικά κατόπιν κρίσης

3ο Διεθνές Συνέδριο Εκπαίδευτικής Καινοτομίας (Λάρισα, 2017)

1. **Bersimis G. Fragiskos**, Lygiaris M. Michalis, André Thomas. «Learn Math by Playing» An Alternative, Game-Based Teaching Proposal»

   The use of Information and Communication Technologies (ICT) in the educational process is an essential tool for establishing learning skills for students of all levels, through their engagement with appropriate educational technologies. An innovation that has emerged in recent years is interactive learning by using educational video games, the digital environment of which is governed by basic computer game principles. Thus, the Gamification of Education, in addition to the intended learning outcomes, provides entertainment and cultivates the student's competing spirit, contributing to engagement-attachment in the teaching subject. This paper proposes an alternative teaching approach for Calculus and more specifically the concept of function's limits, using Triseum's Variant: Limits™ software, while following the instructions of the corresponding study program and presents the advantages of this approach over traditional teaching methods. In particular, a series of challenges or puzzles were selected from different levels or zones of Variant: Limits, which describes and teaches the concept of Function's Limits and its key attributes, with appropriate matching of audiovisual media in the three-dimensional (3D) digital environment of the Variant: Limits Educational Game. Within the framework of the game, students are given the opportunity to learn the above-mentioned concepts of mathematics and to deeply understand/100% acquire this knowledge through their active participation.

3ο Πανελλήνιο Συνέδριο Συνέδριο Ι.Α.Κ.Ε (Ηράκλειο, 2017)

2. Καμπέρης Νίκος, Βερεβή Άλκηστις, **Μπερσίμης Γ. Φραγκίσκος**. «Όψεις κοινωνικης αλληλεγγυης και εθελοντισμου στο μητροπολιτικο κοινωνικο ιατρείο του δήμου Ελληνικού (περιφέρεια Αττικής)».

   Η έρευνά στο Μητροπολιτικό Κοινωνικό Ιατρείο του Ελληνικού μελετά τα κίνητρα και τα αντικίνητρα για τη συμμετοχή των εθελοντών γιατρών και των εθελοντών της γραμματείας, τα προβλήματα που αντιμετωπίζουν και τις προτάσεις τους για τη βελτίωση της δομής. Εξετάζεται μέσα από συνεντεύξεις με εθελοντές το πώς αντιλαμβάνονται το μέλλον των κοινωνικών ιατρείων και του εθνικού συστήματος υγείας και πώς επιδρούν τα περιστατικά που αντιμετωπίζουν και τα αισθήματα των επωφελουμένων στη διάθεσή τους να συνεχίσουν να προσφέρουν. Το υλικό των συνεντεύξεων αξιοποιείται για έναν ουσιαστικό προβληματισμό σχετικά με τις δυνατότητες και τα όρια που θέτουν οι σύγχρονες μορφές εθελοντισμού στο πλαίσιο των συνθηκών υγειονομικής διακινδύνευσης..

33ο Πανελλήνιο Συνέδριο Συνέδριο Μαθηματικών (Χανιά, 2016)

3. **Μπερσίμης Γ. Φραγκίσκος**. «Βελτίωση διαγνωστικής ικανότητας δεικτών υγείας με χρήση μαθηματικών & στατιστικών μεθόδων».

   Η εργασία περιλαμβάνει γενικά στοιχεία της επιστήμης της Βιοστατιστικής και παρουσιάζει αναλυτικά στοιχεία για δείκτες που χρησιμοποιούνται ευρέως στο χώρο των βιολογικών επιστημών. Οι δείκτες υγείας σχετίζονται με κάποια πάθηση, με απώτερο στόχο κυρίως, την ταξινόμηση ενός ατόμου ως ασθενή ή μη. Σε αυτή την εργασία πραγματοποιείται μια προσπάθεια αποτύπωσης μαθηματικών και στατιστικών μεθόδων που βελτιώνουν τη διαγνωστική ικανότητα δεικτών υγείας. Επίσης, παρουσιάζεται ειδική περίπτωση διακριτών δεικτών στην οποία η ευαισθησία είναι γνησίως αύξουσα συνάρτηση του αριθμού των διαμερίσεων των συνιστωσών.

4. **Μπερσίμης Γ. Φραγκίσκος**, Παναγιώτης Ανδρεόπουλος. «Ανάπτυξη πιθανοθεωρητικών κατανομών από μαθηματικά μοντέλα θνησιμότητας».

   Στην παρούσα εργασία παρουσιάζονται μαθηματικά μοντέλα εξομάλυνσης δεδομένων θνησιμότητας με χρήση ειδικών κατανομών (Gompertz, Makeham), που είναι κατάλληλες για την περιγραφή τέτοιων δεδομένων στην επιστήμη της δημογραφίας και της χωρικής στατιστικής. Κάνοντας μίξη των προαναφερθέντων κατανομών με τις ήδη γνωστές πιθανοθεωρητικές κατανομές Βήτα και Γάμμα, αναπτύσσονται κατανομές που έχουν ικανοποιητική προσαρμογή και επαρκή προβλεπτική ικανότητα για διαφορετικά σετ δεδομένων θνησιμότητας. Ειδικότερα, οι νέες πιθανοθεωρητικές κατανομές (**ANBE distribution models)**, δυνητικά θα μπορούν να εφαρμοστούν σε ευρύτερο φάσμα επιστημονικών τομέων (π.χ. δημογραφία, βιοστατιστική, αναλογιστική επιστήμη). Αποδεικνύεται με χρήση θεωρίας πιθανοτήτων, ότι τα μοντέλα **B-ANBE** και **ANBE-G** κατανομές ικανοποιούν βασικές ιδιότητες πιθανοθεωρητικών κατανομών.

5. **Μπερσίμης Γ. Φραγκίσκος**, Καρκάνης Βασίλης, Κόσυβας Γιώργος «Πανελλαδικές Εξετάσεις 2016: «Η εκδίκηση του γνωστού;»».

Σε αυτή την εργασία παρουσιάζονται τα ευρήματα της στατιστικής επεξεργασίας, της βαθμολογίας των γραπτών εξεταστικών δοκιμίων των Μαθηματικών, που συγκεντρώθηκαν στο 31ο Β.Κ. Ειδικότερα μελετήθηκαν οι επιδόσεις των υποψηφίων στη θεωρία και σε ερωτήματα των ασκήσεων που έχουν δομηθεί με βάση «γνωστές» ασκήσεις του σχολικού βιβλίου. Σύμφωνα με τα ευρήματα της έρευνας η αποτυχία των υποψηφίων είναι ιδιαίτερα ανησυχητική και χρήζει περαιτέρω μελέτης.

2ο Πανελλήνιο Συνέδριο Εκπαίδευτικής Καινοτομίας (Λάρισα, 2016)

6. **Μπερσίμης Γ. Φραγκίσκος**, Σακέλλη Μ. Δέσποινα, «Διδάσκοντας Γεωμετρία στα Αρχαία Ελληνικά: Μια διδακτική πρόταση»
   Η έννοια της διαθεματικότητας αποτελεί έναν από τους βασικούς άξονες της σύγχρονης φυσιογνωμίας της εκπαίδευσης, καθώς, όπως επισημαίνεται στα Προγράμματα Σπουδών, είναι σημαντικό να δίνεται σε αυτήν ουσιαστική βαρύτητα μέσω αμφίδρομων συνδέσεων ανάμεσα στα μαθησιακά-διδακτικά αντικείμενα. Το παρόν διδακτικό σενάριο αποτυπώνει την προσπάθεια «διαθεματικής» προσέγγισης του μαθήματος της Αρχαίας Ελληνικής Γλώσσας, με παράλληλη αξιοποίηση των γνώσεων από τη Γεωμετρία. Συγκεκριμένα, επιλέχθηκε ένα απόσπασμα κειμένου από το έργο «Στοιχεία» του Ευκλείδη, στο οποίο περιγράφεται η κατασκευή ισόπλευρου τριγώνου. Στο πλαίσιο της επεξεργασίας του εν λόγω αποσπάσματος δίνεται η δυνατότητα στους μαθητές αφενός να διδαχθούν τα επίθετα της β' κλίσης της αρχαίας ελληνικής και αφετέρου να κατανοήσουν τη συγκεκριμένη γεωμετρική κατασκευή ισόπλευρου τριγώνου.

7. **Μπερσίμης Γ. Φραγκίσκος**, Αντωνόπουλος Γιώργος, «Πρόγραμμα Ρομποτικής
   Η εκπαιδευτική ρομποτική αποτελεί ένα πεδίο της εκπαίδευσης που παρουσιάζει ιδιαιτέρως μεγάλο ενδιαφέρον ως προς τα παιδαγωγικά εργαλεία που χρησιμοποιούνται καθώς αξιοποιούνται κλασσικές φυσικές επιστήμες και νέες τεχνολογίες αιχμής του 21^{ου} αιώνα. Η δυνατότητα κατασκευής ρομποτικών διατάξεων, σχεδιασμού αλγορίθμων και ο προγραμματισμός τους ώστε να αποτελούν αυτόνομα συστήματα αποτελεί μια πρόκληση για τους μαθητές που παύουν να είναι παθητικοί δέκτες στην εκπαιδευτική διαδικασία αλλά γίνονται ενεργά μέλη αναπτύσσοντας τις ικανότητες τους σε επίπεδο δημιουργικότητας και εφευρετικότητας και συγχρόνως ανακαλύπτουν την χρησιμότητα των μαθηματικών και της πληροφορικής στον πραγματικό κόσμο. Στην παρούσα εργασία η διδασκαλία της ρομποτικής πραγματοποιήθηκε με χρήση της ομαδοσυνεργατικής εκπαιδευτικής διαδικασίας.

38ο Διεθνές Συνέδριο Διαιτολογίας (ESPEN, Danmark 2016)

8. K.A. Poulia, M. Kontogianni, **F. Bersimis**, S. Kosak, on behalf of nutritionDay, K. Schindler, on behalf of nutritionDay, M. Hiesmayr, on behalf of nutritionDay, M. Chourdakis, «Region Based Differences in Nutritional Support Practices in Eastern Europe. Preliminary Results from analyzing nutritionday's Database (2006-13)»

9. K.A. Poulia, M. Kontogianni**, F. Bersimis**, S. Kosak, on behalf of nutritionDay, K. Schindler, on behalf of nutritionDay, M. Hiesmayr, on behalf of nutritionDay, M. Chourdakis, «Institutional Factors Affecting Food Consumption during Hospitalization. Preliminary Results from Analyzing Nutritionday's Database (2006-13)»

ΠΡΟΫΠΗΡΕΣΙΑ / ΕΜΠΕΙΡΙΑ:

Διδασκαλία: Τριτοβάθμια Εκπαίδευση:

2015-2016 AMC, Key Investigative Skills I & Key Investigative Skills II

2012-2015 Χαροκόπειο Πανεπιστήμιο
Τμήμα Διαιτολογίας: Βιοστατική Ι, , Βιοστατική ΙΙ, Ειδικά Θέματα Ανάλυσης Δεδομένων
Τμήμα Οικιακής Οικονομίας: Στατιστική με χρήση Η/Υ (SPSS) Ι, Στατιστική με χρήση Η/Υ (SPSS) ΙΙ

2010-2012 Τεχνολογικό Εκπαιδευτικό Ίδρυμα Ιονίων Νήσων (Παράρτημα Λευκάδας)
Τμήμα Εφαρμογών Πληροφορικής στη Διοίκηση και στην Οικονομία: Στατιστική Ανάλυση Δεδομένων (SPSS)

Τμήμα Πληροφορικής & Τηλεπικοινωνιών: Γραμμική Άλγεβρα, Στατιστική & Θεωρία Πιθανοτήτων (SPSS)

Διδασκαλία: Δευτεροβάθμια Εκπαίδευση:
2007-2018 Γυμνάσιο – Λύκειο, Μαθηματικά

Συμμετοχή σε Ερευνητικά Προγράμματα:

2015-2016 Συνεργάτης στο ερευνητικό πρόγραμμα του Χαροκοπείου Πανεπιστημίου: «Exploring factors influencing dietary intake during hospitalization and their impact on patients' outcome – a country based analysis». (Συνεδρία 315/23.12.2015) Επιστημονικός Υπεύθυνος: Δρ. Δ. Παναγιωτάκος.

ΜΕΛΟΣ ΣΕ ΕΠΙΣΤΗΜΟΝΙΚΕΣ ΕΝΩΣΕΙΣ

- Ελληνικό Στατιστικό Ινστιτούτο
- Ελληνική Μαθηματική Εταιρεία
- Stochastic Modeling Techniques and Data Analysis International Network
- Επιστημονική Ένωση για την Προώθηση της Εκπαιδευτικής Καινοτομίας
- Πανελληνια Παιδαγωγικη Εταιρεια Δευτεροβαθμιας Εκπαιδευσης

ΠΑΡΑΡΤΗΜΑ Γ

Δημοσιεύσεις και Ανακοινώσεις που προέκυψαν από τη διατριβή

Δημοσιευμένες Εργασίες σε επιστημονικά περιοδικά

1. **Bersimis G. Fragkiskos**, Panagiotakos Demosthenes, Vamvakari Malvina. The use of components' weights improves the diagnostic accuracy of a health-related index

   This work aims to evaluate whether the use of specific weights in each component of a composite health-related index, is associated with the diagnostic accuracy of the index. In addition, the impact of a composite health related index's components multitude is examined in relation to its classification ability. An un-weighted and various weighted indices were constructed using different weighting methods and were assessed as to their diagnostic performance. The indices' diagnostic ability was evaluated by using true positive rate (TPR), true negative rate (TNR), true rate (TR), positive predictive value (PPV), negative predictive value (NPV) and the area (AUC) under the receiver operating characteristic curve (ROC). Weights used in this study were obtained from linear discriminant analysis (LDA) and binary logistic regression (BLR). These indices were applied in both simulated and actual data; and a variety of scenarios was applied based on the distribution's parameters of the component variables and on the number of components used. Results indicate that AUC, TR, TPR, TNR, PPV and NPV of the weighted indices were significantly higher compared to the un-weighted one; whereas AUC was positively associated with the number of components of each index that were correlated with the outcome. Weighting of index's components, as well as greater number of components related to the investigated outcome should be recommended for the construction of accurate indices.

2. **Bersimis G. Fragkiskos**, Panagiotakos Demosthenes, Vamvakari Malvina. Investigating the sensitivity function's monotony of a health-related index, Journal of Applied Statistics. 2016 Sep 06

   In this work it is investigated theoretically whether the support's length of a continuous variable, which represents a simple health-related index, affects the index's diagnostic ability of a binary health outcome. The aforementioned is attempted by studying the monotony of the index's sensitivity function, which is a measure of its diagnostic ability, in the cases that the index's distribution was either unknown or the uniform. The case of a composite health- related index which is formed by the sum of m component variables is also presented when the distribution of its component variables was either unknown or the uniform. It is proved that a health-related index's sensitivity is a non-decreasing function as to the finite length of its components' support, under certain condition. In addition, similar propositions are presented in the case that a health-related index is distributed normally according to its distribution parameters.

3. Panagiotakos Demosthenes & **Bersimis G. Fragkiskos**. An index that rates relevance of scientific work in bio-sciences: The scientific relevance-index. 2014 Jan

   This work aims to present an objective index that rates scientific relevance (SR) of scientists' published work, the SR index, in a specific thematic field. The proposed index is calculated based on equally or weighted individual parameters that measure scientist's publication record and recognition, using easily accessible and unbiased data from existing bibliometric databases. The application of the SR‑index could be in any scientific field; an example is given here on cell and molecular biology fiel

4. **Bersimis G. Fragkiskos**, Panagiotakos Demosthenes, Vamvakari Malvina. Sensitivity of health related indices is a non-decreasing function of their partitions, Journal of Statistics Applications & Probability. 2013 Jun 25

   This work aims to theoretically evaluate whether the number k of partitions of a discrete variable X affects the sensitivity $S_e(X)$ of a binary health outcome Y. The distribution of variable X was either unknown or the uniform. Thus, two discrete random variables X_k and X_{k+1} with k and k+1 partitions, respectively, were considered. In addition, a random variable Y that indicates the actual health status of an individual was also considered. The case of the composite index T_k which is formed by the sum of *m* variables X_k^j, j=1,2,...,m either when the distribution of each variable X_k^j is unknown or the uniform was also investigated. This work suggests that the sensitivity of an index is a non-decreasing function of the number of partitions, under certain conditions.

Παρουσιάσεις σε Συνέδρια & Δημοσιευμένες Εργασίες σε πρακτικά:

1. **F.G. Bersimis** (Invited Speaker) «New findings in the health state of populations and implications in the Health Systems: Indices in health related scientific fields», ISBIS 2018. Greece, Piraeus.
2. **Bersimis, F. G**., Panagiotakos, D., Varlamis I. & Vamvakari, M. "*Calculation Methods for Binary Classification based on Discrete Data. An Application in both Simulated and Real Data*." **5th Stochastic Modeling Techniques and Data Analysis International Conference. 1 June – 4 June, Chania, Crete, Grece**.

3. **Μπερσίμης Γ. Φραγκίσκος**. «***ΒΕΛΤΙΩΣΗ ΔΙΑΓΝΩΣΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ ΔΕΙΚΤΩΝ ΥΓΕΙΑΣ ΜΕ ΧΡΗΣΗ ΜΑΘΗΜΑΤΙΚΩΝ & ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΘΟΔΩΝ*».** 33ο Πανελλήνιο Συνέδριο Μαθηματικών, 11 – 15 Νοεμβρίου, 2016, Χανιά – Ελλάδα.
4. **Bersimis G. Fragkiskos, Panagiotakos Demosthenes, Vamvakari Malvina.** ***"WEIGHTING AS A METHOD OF OPTIMIZING AN INDEX'S DIAGNOSTIC PERFORMANCE. THE CASE OF ATTICA STUDY IN GREECE.*** 4th Stochastic Modeling Techniques and Data Analysis International Conference and 5th Demographics 2016 Workshop. 1 June – 4 June, Valletta, Malta.
5. **Bersimis G. Fragkiskos, Panagiotakos Demosthenes, Vamvakari Malvina**. ***"A COMPOSITE HEALTH RELATED INDEX'S DIAGNOSTIC ACCURACY INCREASES WHEN ITS COMPONENTS' SUPPORT LENGTH INCREASES AND WHEN SPECIFIC WEIGHTS ARE ASSIGNED TO THEM"***. 8th conference of the Eastern Mediterranean Region of the International Biometric Society (IBS-EMR). 11-15 May, 2015. Nevsehir, Cappadocia – Turkey.
6. **Φ. Γ. Μπερσίμης, Μ. Βαμβακάρη, Δ. Β. Παναγιωτάκος**. ***"Η ΧΡΗΣΗ ΕΙΔΙΚΩΝ ΣΤΑΘΜΙΣΕΩΝ ΣΤΙΣ ΣΥΝΙΣΤΩΣΕΣ ΕΝΟΣ ΣΥΝΘΕΤΟΥ ΔΕΙΚΤΗ ΥΓΕΙΑΣ ΑΥΞΑΝΕΙ ΤΗ ΔΙΑΓΝΩΣΤΙΚΗ ΤΟΥ ΙΚΑΝΟΤΗΤΑ"***. 28ο Πανελλήνιο Συνέδριο Στατιστικής, με θέμα «Χωρική Στατιστική και εφαρμογές της Στατιστικής στη Γενετική και την Πληροφορική», 5-8 Απριλίου, 2015, Αθήνα, Ελλάδα.
7. **Bersimis G. Fragkiskos, Panagiotakos Demosthenes, Vamvakari Malvina.** ***"THE SENSITIVITY FUNCTION OF HEALTH RELATED SCALES IS A NON DECREASING FUNCTION OF THEIR PARTITIONS.*** 2nd Stochastic Modeling Techniques and Data Analysis International Conference, 5-8 June, 2013, Chania, Crete, Greece.
8. **Φ.Γ. Μπερσίμης, Μ. Βαμβακάρη, Δ.Β. Παναγιωτάκος**. ***"Η ΧΡΗΣΗ ΕΙΔΙΚΩΝ ΣΤΑΘΜΙΣΕΩΝ ΣΤΙΣ ΣΥΝΙΣΤΩΣΕΣ ΕΝΟΣ ΣΥΝΘΕΤΟΥ ΔΕΙΚΤΗ***

ΥΓΕΙΑΣ ΑΥΞΑΝΕΙ ΤΗ ΔΙΑΓΝΩΣΤΙΚΗ ΤΟΥ ΙΚΑΝΟΤΗΤΑ". 25° Πανελλήνιο Συνέδριο Στατιστικής, με θέμα «Στατιστική στις Βιοεπιστήμες», 18 - 22 Απριλίου, Βόλος, Ελλάδα.