# ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

## ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
## ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ


# ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΠΟΛΥΔΙΑΣΤΑΤΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΦΑΡΜΟΓΕΣ


Βασιλική Ι. Βασιλειάδη


Διατριβή

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική


Πειραιάς
Μάιος 2015

# UNIVERSITY OF PIRAEUS

# DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE


# STATISTICAL TECHNIQUES FOR THE ANALYSIS OF MULTIVARIATE DATA AND APPLICATIONS


Vasiliki I. Vasileiadi


Thesis

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics


Piraeus, Greece
May 2015

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών.

Τα μέλη της Επιτροπής ήσαν:
- Καθηγητής Κούτρας Μάρκος (Επιβλέπων)
- Επίκουρος Καθηγητής Τήνιος Πλάτων
- Λέκτορας Μπερσίμης Σωτήριος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

*Στους γονείς μου*

*Ιωάννη και Κονδυλένια*

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω αρχικά τον επιβλέποντα καθηγητή μου, κύριο Μάρκο Κούτρα, για την πολύτιμη βοήθειά του, την καθοδήγησή του και τις συμβουλές του κατά τη διάρκεια εκπόνησης της διπλωματικής εργασίας μου. Θεωρώ πως είχαμε μια πολύ καλή συνεργασία στα πλαίσια της οποίας διεύρυνα τους ορίζοντες μου απέναντι στην επιστήμη της Στατιστικής και έμαθα να θέτω και να ολοκληρώνω τους στόχους μου. Τον ευχαριστώ επίσης για την διδασκαλία του καθ' όλη την διάρκεια των φοιτητικών μου ετών, η οποία συνέβαλε τόσο στο να κατανοήσω την Στατιστική όσο και στο να θελήσω να την τελειοποιήσω στα πλαίσια ενός μεταπτυχιακού προγράμματος.

Επιπλέον, θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή, κύριο Π. Τήνιο και τον Λέκτορα, κύριο Σ. Μπερσίμη τόσο για τις γνώσεις που μου έχουν προσφέρει κατά τη φοίτησή μου στο πανεπιστήμιο, όσο και για το χρόνο που αφιέρωσαν στη διόρθωση της διπλωματικής μου εργασίας.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τους γονείς μου που στέκονται πάντα δίπλα μου, πιστεύουν σε εμένα και μου δίνουν κίνητρο και ώθηση σε ό,τι αναλαμβάνω.

Τέλος, ευχαριστώ τους φίλους μου, που είναι πάντα στη ζωή μου και με την υπομονή και την συμπαράστασή τους κάνουν κάθε σκοπό μου να φαίνεται ευκολότερος. Αλλά και τους συμφοιτητές και φίλους μου Π. Χαρτά και Γ. Μπάρτζη που οι συμβουλές τους και η στήριξή τους είχαν καθοριστικό ρόλο στην ολοκλήρωση της εργασίας μου.

# Περίληψη

Η πολυμεταβλητή ανάλυση ασχολείται με τη συλλογή, περιγραφή και ανάλυση δεδομένων για τα οποία υπάρχουν μετρήσεις πολλών μεταβλητών. Η μελέτη των μεταβλητών αυτών μεμονωμένα για την καθεμία οδηγεί σε συμπεράσματα που δεν αποδίδουν πιστά την πραγματικότητα καθώς δεν λαμβάνουν καθόλου υπόψη την έννοια της αλληλεξάρτησης. Η πολυμεταβλητή ανάλυση εντοπίζει την εξάρτηση των μεταβλητών, τις εντάσσει σε ομάδες και μειώνει τον όγκο τους καθιστώντας τις κατάλληλες για περεταίρω επεξεργασία.

Οι τεχνικές που χρησιμοποιούνται στην πολυμεταβλητή ανάλυση είναι γραφικές και στατιστικές/μαθηματικές. Οι γραφικές μέθοδοι χρησιμεύουν στην απεικόνιση των πολυδιάστατων δεδομένων εξυπηρετώντας έτσι την εξαγωγή συμπερασμάτων. Αντίστοιχα, οι στατιστικές/μαθηματικές μέθοδοι αποσκοπούν τόσο στην ερμηνεία των συσχετίσεων μεταξύ των μεταβλητών όσο και στη μείωση των διαστάσεων του προβλήματος που μελετάται.

Η παρούσα διατριβή αρχικά παρέχει μια εισαγωγή στις βασικότερες τεχνικές της πολυμεταβλητής ανάλυσης. Στη συνέχεια, γίνεται μια εκτενής ανάλυση των σημαντικότερων γραφικών μεθόδων τόσο σε θεωρητικό επίπεδο όσο και σε επίπεδο εφαρμογής. Επιπροσθέτως, αναλύονται λεπτομερώς οι πιο γνωστές στατιστικές πολυμεταβλητές μέθοδοι. Τέλος, εφαρμόζονται αυτές οι μέθοδοι στα πλαίσια ενός πραγματικού δείγματος εξάγοντας και τα ανάλογα συμπεράσματα.

# Abstract

Multivariate analysis deals with the collection, description and inference of data containing measurements of a large number of variables. Studying these variables independently may lead to conclusions that do not reflect the true nature and structure of the data because the dependence is not taken into account. Multivariate analysis explores the dependence between variables, combines them into groups and reduces their dimensionality thereof making them suitable for additional processing.

The techniques that are used in multivariate analysis are graphical and statistical/mathematical. The graphical methods are used in the illustrative presentation of multivariate data thus facilitating the task of drawing some preliminary conclusions. Furthermore, the statistical/mathematical methods serve both the interpretation of the correlation between variables and the dimensionality reduction of the dataset being studied.

The present dissertation initially provides an introduction in the most popular techniques of Multivariate Analysis. An extensive analysis of the main graphical methods is presented both in theory and application level. Moreover, it describes in detail, the most common statistical/mathematical multivariate methods. Finally, these methods are applied on a real dataset and an illustration is offered how one can derive useful conclusions by practicing the techniques described in the thesis.

# Περιεχόμενα

# Κατάλογος Πινάκων

# Κατάλογος Σχημάτων

# ΚΕΦΑΛΑΙΟ 1

# Εισαγωγή

## 1.1 Πολυμεταβλητή Στατιστική Ανάλυση

Η πολυμεταβλητή ανάλυση ασχολείται με τη μελέτη και εφαρμογή στατιστικών μεθόδων συλλογής, περιγραφής και ανάλυσης δεδομένων που αποτελούνται από μετρήσεις πολλών μεταβλητών. Έτσι είναι δυνατόν να μελετηθούν ταυτόχρονα $p$ χαρακτηριστικά από ένα δείγμα $n$ δεδομένων, τοποθετώντας τα σε έναν πίνακα $X$ διαστάσεων $n \times p$. Κάθε γραμμή του πίνακα $X$ αποτελεί μια πολυμεταβλητή παρατήρηση, επομένως συνολικά μελετώνται $n$ πολυμεταβλητές παρατηρήσεις.

Η πολυμεταβλητή ανάλυση θα μπορούσε να θεωρηθεί ως επέκταση της μονομεταβλητής ανάλυσης, καθώς αναζητά και αναλύει την εξάρτηση που αναπτύσσεται μεταξύ των μεταβλητών. Τα περισσότερα φαινόμενα είναι από τη φύση τους πολυμεταβλητά. Αυτό σημαίνει πως στην πράξη ο αναλυτής έχει συνήθως στη διάθεσή του αφθονία δεδομένων και μεγάλο πλήθος μεταβλητών οι οποίες παρουσιάζουν κάποια μορφή εξάρτησης.

Η γρήγορη ανάπτυξη των υπολογιστών καθώς και η ραγδαία εξέλιξη των στατιστικών προγραμμάτων έχουν κάνει την πολυμεταβλητή ανάλυση μία από τις συχνότερα χρησιμοποιούμενες τεχνικές της στατιστικής. Επομένως η γνώση της είναι απαραίτητη σε πολλές επιστήμες όπως στην Ιατρική, στη Βιολογία, στις Πολιτικές, Οικονομικές και Κοινωνικές επιστήμες, στον Έλεγχο Ποιότητας κ.λ.π.

Τα επιστημονικά θέματα για τα οποία χρησιμοποιούνται κυρίως οι μέθοδοι της πολυμεταβλητής ανάλυσης είναι τα ακόλουθα (Πανάρετος και Ξεκαλάκη (1995)):

- Ελαχιστοποίηση των δεδομένων ή απλοποίηση της δομικής τους σχέσης: Στόχος είναι να περιγραφεί το υπό μελέτη φαινόμενο με τον απλούστερο δυνατό τρόπο χωρίς ταυτόχρονα να θυσιαστούν πολύτιμες πληροφορίες.
- Κατηγοριοποίηση και ομαδοποίηση: Δημιουργούνται ομάδες από παρόμοια αντικείμενα ή μεταβλητές με βάση τα χαρακτηριστικά των μετρήσεων.
- Μελέτη της εξάρτησης των μεταβλητών: Μελετάται η φύση της σχέσης που υφίσταται μεταξύ των μεταβλητών ώστε να εξετασθεί εάν οι μεταβλητές είναι ανεξάρτητες ή εάν υπάρχει κάποια είδους εξάρτηση μεταξύ αυτών.
- Πρόβλεψη: Καθορίζονται οι σχέσεις που υφίστανται μεταξύ μεταβλητών ώστε να δοθεί η δυνατότητα να προβλεφθεί η τιμή μιας ή περισσότερων μεταβλητών με βάση παρατηρήσεις στις υπόλοιπες μεταβλητές.

## 1.2 Ιστορική αναδρομή

Τα θεμέλια για την πολυμεταβλητή στατιστική ανάλυση τέθηκαν ήδη στις αρχές του 19$^{ου}$ αιώνα, όταν ο Robert Adrian εξέτασε την διμεταβλητή κανονική κατανομή και ο Francis Galton μελέτησε το φαινόμενο της συσχέτισης στα τέλη του ίδιου αιώνα. Ο Francis Galton (1889) στο βιβλίο του «Natural Inheritance», κατασκεύασε ένα διμεταβλητό πίνακα συχνοτήτων που περιείχε το ύψος των πατεράδων και των αρσενικών παιδιών σε μία οικογένεια. Με τη χρήση αυτών των δεδομένων μπορούσε να προβεί σε πρόβλεψη για το αναμενόμενο ύψος ενός παιδιού, γνωρίζοντας το ύψος του αδερφού του. Η διαδικασία που ακολούθησε βασίστηκε στην ομαδοποίηση όλων των ατόμων που έχουν ύψος αντίστοιχο του ατόμου που μελετάται και στην καταγραφή των αποκλίσεων τους από τον μέσο όρο.

Μαθητευόμενος του Galton και στη συνέχεια ένας από τους μεγαλύτερους στατιστικούς της γενιάς του, ήταν ο μαθηματικός Karl Pearson.Ο Pearson έβαλε τα θεμέλια για πολλές στατιστικές μεθόδους που χρησιμοποιούνται σήμερα. Μεταξύ άλλων, κατασκεύασε τον έλεγχο «$\chi^2$ κατά Pearson», μέτρο το οποίο επινοήθηκε για να ελεγχθεί η καλή προσαρμογή μιας καμπύλης δεδομένων αλλά στην πορεία γνώρισε ένα ευρύτερο φάσμα εφαρμογών. Παράλληλα ασχολήθηκε και με τη συσχέτιση των μεταβλητών. Στην δημοσίευσή του με

τίτλο «Regression, Heredity and Panmixia» (1896), γενίκευσε τις μεθόδους του Galton, ασχολήθηκε με τη συσχέτιση τριών μεταβλητών και ερμήνευσε τους συντελεστές πολλαπλής παλινδρόμησης. Το μέτρο που χρησιμοποίησε είναι γνωστό ως «Συντελεστής γραμμικής συσχέτισης *r* του Pearson» και αποτελεί απαραίτητο εργαλείο της Γραμμικής Παλινδρόμησης. Τέλος, αξίζει να σημειωθεί ότι ο Pearson στην δημοσίευσή του «Contribution to the mathematical theory of evolution» (1894), υπήρξε ο πρώτος που εισήγαγε τον όρο της «τυπικής απόκλισης» με τον συμβολισμό «*σ*».

Στις αρχές του 20ου αιώνα, όταν ο Pearson έκανε τις πρώτες του μελέτες στις καμπύλες συχνοτήτων και τη συσχέτιση των μεταβλητών, ο Άγγλος στατιστικός George Udny Yule, συνεχίζοντας με εξαιρετικό τρόπο την έρευνα που αφορούσε στη συσχέτιση, στη δημοσίευσή του (1907) «On the Theory of Correlation for any Number of Variables» έθεσε τα θεμέλια της θεωρίας της μερικής συσχέτισης και της γραμμικής παλινδρόμησης για οποιοδήποτε αριθμό μεταβλητών. Σημαντική ήταν και η προσφορά του στη θεωρία των χρονοσειρών. Ο ίδιος, με αφορμή τις υψηλές συσχετίσεις που παρατήρησε μεταξύ «μη συνδεδεμένων» ποσοτήτων κατά τη διάρκεια του χρόνου, έθεσε τις βάσεις της θεωρίας της «Αυτοπαλινδόμησης».

Ο Ronald Aylmer Fisher στις αρχές του 20ου αιώνα, οδήγησε τη στατιστική επιστήμη σε νέα πορεία. Ήδη από το 1912 και ενώ ήταν ακόμα φοιτητής στο Cambridge, εισήγαγε τη μέθοδο μέγιστης πιθανοφάνειας. Στη συνέχεια, ασχολήθηκε ιδιαίτερα με εφαρμογές της στατιστικής επιστήμης στον τομέα της γενετικής. Συγκεκριμένα, με το έργο του «The Genetical theory of natural selection» (1930), χρησιμοποίησε στατιστικά μοντέλα για να ελέγξει την θεωρία του Δαρβίνου. Μεταξύ άλλων ασχολήθηκε με την μέθοδο της Ανάλυσης Διακύμανσης (ANOVA) καθώς και την παραγωγή διαφόρων κατανομών δειγματοληψίας. Ο Fisher υπήρξε ιδιαιτέρως ενεργός στην ανάπτυξη της Πολυμεταβλητής Ανάλυσης. Ο ίδιος χρησιμοποίησε αποτελεσματικά τις συσχετίσεις μεταξύ πολλών μεταβλητών για να επιλύσει προβλήματα παλινδρόμησης ενώ το όνομά του έχει συνδεθεί με την Γραμμική Διαχωριστική Ανάλυση (Linear Discriminant Analysis) η οποία χρησιμοποιείται για την πρόβλεψη μιας κατηγορικής μεταβλητής απόκρισης, από μία ή περισσότερες συνεχείς ανεξάρτητες μεταβλητές.

Ένας στατιστικός που διδάχθηκε και επηρεάστηκε ιδιαίτερα από τον Fisher ήταν ο Harold Hotelling. Ο Hotelling θεωρήθηκε πρωτοπόρος στον τομέα της στατιστικής κατά τον 20ο αιώνα. Το 1931 δημοσίευσε την μελέτη του *«The Economics of Exhaustible Resources»* που αφορούσε στην κατανομή «Student's t» και χρησιμοποίησε για πρώτη φορά τον όρο

«διάστημα εμπιστοσύνης». Μελέτησε επίσης την θεωρία παιγνίων, αποσβέσεων και εξάντλησης πηγών. Ο τομέας όμως στον οποίο διέπρεψε ήταν η πολυμεταβλητή στατιστική ανάλυση. Το όνομά του είναι συνδεδεμένο με διάφορες μεθόδους και στατιστικές με πιο σημαντική την επονομαζόμενη «Γενίκευση της πολυδιάστατης κατανομής $T^2$». Στη συνέχεια, διατύπωσε με ακρίβεια την ιδέα της χρήσης των συνιστωσών, θέτοντας τα θεμέλια για μία νέα μέθοδο που οδηγεί στην στατιστική συμπερασματολογία, γνωστή ως «Μέθοδος των Κύριων Συνιστωσών». Το 1953, ο Hotelling δημοσίευσε ένα άρθρο σχετικό με την κατανομή του συντελεστή συσχέτισης. Η έρευνα του Hotelling αφορούσε στα μαθηματικά, οικονομικά και στη θεωρητική και εφαρμοσμένη στατιστική. Χρησιμοποίησε τις ήδη υπάρχουσες μεθόδους και πρωτοπόρησε με πολλές νέες τεχνικές. Αξίζει να σημειωθεί ότι τα περισσότερα κλασσικά συγγράμματα παρουσιάζουν το θέμα της πολυμεταβλητής ανάλυσης με την ίδια προσέγγιση που εισήγαγε ο Hotelling.

Με την πολυμεταβλητή ανάλυση ασχολήθηκε ακόμη ένας σημαντικός στατιστικός, ο Meyer Abraham Girshick , στη δημοσίευσή του (1939) «On the sampling theory of roots of determinantal equations». Σημαντικό επίτευγμα του Girshick ήταν η εύρεση της κατανομής των διανυσμάτων που χρησιμοποιούνται για τον έλεγχο της μηδενικής υπόθεσης που αφορά στην ανεξαρτησία δύο συνόλων μεταβλητών.

Ο Samuel Wilks, συνεισέφερε επίσης στην ανάπτυξη της πολυμεταβλητής ανάλυσης. Το 1932, στη δημοσίευσή του «Certain Generalizations in the Analysis of Variance», δημιούργησε μια γενίκευση του δείκτη συσχέτισης και του συντελεστή πολλαπλής συσχέτισης και μελέτησε τυχαία δείγματα από έναν κανονικό πληθυσμό πολλών μεταβλητών. Τέλος, δεν θα μπορούσε να μην γίνει μια αναφορά και στον καθηγητή C.R. Rao. Ο C.R. Rao είναι ίσως ένας από τους σημαντικότερους εν ζωή στατιστικούς. Η επίδρασή του σε κάθε κλάδο της στατιστικής είναι καθοριστική χαράζοντας το δρόμο και σε πολλές άλλες επιστήμες. Ο ίδιος είχε συνεργαστεί με τους πλέον αναγνωρισμένους στατιστικούς του $20^{ου}$ αιώνα όπως τον Fisher, τον Newman, τον Blackwell, τον Hamming και τον Rubin. Ο C.R. Rao σε ηλικία μόλις 25 ετών, μέσα σε ένα βράδυ απέδειξε το φράγμα διασποράς για αμερόληπτη εκτιμήτρια με μικρό δείγμα δημιουργώντας έτσι το γνωστό στις μέρες μας «φράγμα Cramer-Rao» και ταυτίζοντας το όνομα του με ένα από τα σημαντικότερα αποτελέσματα της στατιστικής θεωρίας. Τα περισσότερο γνωστά ευρήματά του, είναι τα αποτελέσματα που ανέπτυξε για την Εκτιμητική όπως το θεώρημα «Rao-Blackwell», το θεώρημα «Fisher-Rao» και η ανισότητα «Cramer-Rao». Από την άλλη, η πολυμεταβλητή

ανάλυση, θα ήταν πιο φτωχή χωρίς τη συμβολή του C.R. Rao. Οι έλεγχοι $U$ και $F$ του Rao, η χρήση του κριτηρίου $D^2$ του Mahalanobis και των κανονικών μεταβλητών στη μεθοδολογία της ανάλυσης συστάδων και στην ταξινομική θεωρία, η ανάλυση διασποράς είναι μόνο ένα μέρος της προσφοράς του στη Μαθηματική Στατιστική.

## 1.3 Πολυμεταβλητές Μέθοδοι

Οι τεχνικές της πολυμεταβλητής ανάλυσης, θα μπορούσαν να θεωρηθούν ότι αποτελούν γενίκευση των τεχνικών που χρησιμοποιούνται στην μονομεταβλητή ανάλυση. Στην πραγματικότητα, οι τεχνικές αυτές αποτελούν «εργαλείο» τόσο για τον εντοπισμό και την ερμηνεία των συσχετίσεων μεταξύ των μεταβλητών, βάσει των οποίων αποκτάται καλύτερη γνώση για το φαινόμενο που εξετάζεται, όσο και για τη μείωση των διαστάσεων του προβλήματος με την αφαίρεση των πλεονάζουσων μεταβλητών. Στο σημείο αυτό θα αναλύσουμε περιληπτικά κάποιες από τις πλέον δημοφιλείς μεθόδους πολυμεταβλητής ανάλυσης.

- Ανάλυση σε κύριες συνιστώσες (Principal Component Analysis)

Η ανάλυση σε κύριες συνιστώσες είναι περισσότερο ένα μέσον απλοποίησης των δεδομένων τα οποία θα υποστούν περαιτέρω ανάλυση, παρά ένας αυτοσκοπός. Πρόκειται για μία τεχνική που δίνει τη δυνατότητα να ερμηνευθούν τα αρχικά δεδομένα, μειώνοντας τις διαστάσεις τους. Δεδομένου ότι στην πολυμεταβλητή ανάλυση χρησιμοποιείται ένα σύνολο $p$ μεταβλητών που εκφράζουν τη συνολική διάσταση των δεδομένων, η ανάλυση σε κύριες συνιστώσες εξηγεί τη δομή διακύμανσης- συνδιακύμανσης των αρχικών μεταβλητών και εντοπίζοντας τις μεταξύ τους συσχετίσεις δημιουργεί ένα σύνολο καινούριων μεταβλητών οι οποίες αποτελούν γραμμικούς συνδυασμούς των αρχικών και οι οποίες περιέχουν όσο το δυνατόν μεγαλύτερο μέρος της αρχικής διακύμανσης. Οι νέες μεταβλητές είναι συνήθως πολύ λιγότερες από τις αρχικές, είναι ασυσχέτιστες και περιέχουν το μεγαλύτερο μέρος της αρχικής πληροφορίας. Οι καινούριες αυτές μεταβλητές ονομάζονται «Κύριες Συνιστώσες». Η μέθοδος των κύριων συνιστωσών είναι ιδιαίτερα χρήσιμη καθώς όπως φαίνεται και από όσα προαναφέρθηκαν έχει τα παρακάτω πλεονεκτήματα:

- Από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών, το οποίο είναι χρήσιμο σε πολλές στατιστικές μεθόδους.
- Μειώνεται ο όγκος των αρχικών μεταβλητών και επομένως διευκολύνει την επεξεργασία τους, με κόστος ότι χάνεται ένα μικρό ποσοστό της αρχικής μεταβλητότητας.
- Αποκαλύπτει σχέσεις μεταξύ των αρχικών μεταβλητών οι οποίες δεν ήταν από την αρχή αντιληπτές και επομένως επιτρέπει ερμηνείες που δεν ήταν δυνατόν να δοθούν με την αρχική μορφή των δεδομένων.

Η ευκολία στη χρήση της μεθόδου καθώς και στην ερμηνεία των αποτελεσμάτων της, την καθιστά μία ιδιαίτερα διαδεδομένη μέθοδο της πολυμεταβλητής ανάλυσης.

- Παραγοντική Ανάλυση ( Factor Analysis)

Η παραγοντική ανάλυση είναι μία στατιστική τεχνική που αποσκοπεί στον εντοπισμό «μοτίβων» σε σχετικά μεγάλα σύνολα δεδομένων με μεγάλους αριθμούς μεταβλητών. Η εισαγωγή της μεθόδου αυτής τοποθετείται στις αρχές του 20$^{ου}$ αιώνα και οφείλεται στον Karl Pearson και στον Charles Spearman. Ο στόχος της παραγοντικής ανάλυσης είναι να διερευνήσει τις συσχετίσεις μεταξύ των μεταβλητών και να περιγράψει τις σχέσεις συνδιακύμανσης μεταξύ τους μέσω κάποιων μη παρατηρούμενων τυχαίων ποσοτήτων που ονομάζονται παράγοντες. Η τεχνική αυτή βασίζεται επομένως, στις συσχετίσεις των μεταβλητών. Πιο συγκεκριμένα, θεωρεί πως οι αρχικές μεταβλητές μπορούν να ομαδοποιηθούν έτσι ώστε μέσα στις ομάδες να υπάρχουν ισχυρά συσχετισμένες μεταβλητές. Η ομαδοποίηση αποδίδεται στους παράγοντες που ουσιαστικά ευθύνονται για την ύπαρξη συσχετίσεων. Σύμφωνα με τα παραπάνω, ένας ερευνητής μπορεί να κατανοήσει την σημασία κάθε παράγοντα μελετώντας το σύνολο των μεταβλητών που εμφανίζουν υψηλή συσχέτιση με τον συγκεκριμένο παράγοντα. Η παραγοντική ανάλυση είναι ιδιαίτερα χρήσιμη σε επιστήμες που ενδιαφέρονται για ψυχομετρικές έρευνες και σε επιστήμες που σχετίζονται με την ανθρώπινη συμπεριφορά. Τα πλεονεκτήματά της είναι:

- Η μείωση των διαστάσεων του αρχικού προβλήματος και κατ' επέκταση η απλούστευσή του.
- Η αναγνώριση μη μετρήσιμων μεταβλητών.
- Η ερμηνεία των συσχετίσεων μεταξύ των μεταβλητών.

Παρόλα τα παραπάνω, η παραγοντική ανάλυση δέχεται αρκετές κριτικές καθώς οι παράγοντες οι οποίοι προκύπτουν επιδέχονται διαφορετικές ερμηνείες οι οποίες συχνά είναι αντικρουόμενες.

- Ανάλυση κατά συστάδες (Cluster Analysis)

Η ανάλυση κατά συστάδες είναι μία στατιστική μέθοδος που σκοπό έχει να ομαδοποιήσει τις υπάρχουσες παρατηρήσεις, χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες από αυτές. Σύμφωνα με τη μέθοδο αυτή, κάθε ομάδα συμπεριλαμβάνει δεδομένα ομοιογενή ενώ οι ομάδες μεταξύ τους διαφέρουν όσο γίνεται περισσότερο. Επομένως η τεχνική αυτή βασίζεται στον βαθμό ομοιότητας που παρατηρείται μεταξύ των παρατηρήσεων και τα εργαλεία που χρησιμοποιούνται είναι τα λεγόμενα «μέτρα ομοιότητας» και τα «μέτρα απόστασης». Κατά την ομαδοποίηση, το πλήθος των ομάδων καθώς και η δομή τους δεν είναι προκαθορισμένα από τον αναλυτή. Σε αντίθετη περίπτωση χρησιμοποιείται ο όρος «Ταξινόμηση» (Taxonomy), «Κατηγοριοποίηση» (Classification) ή «Διακρίνουσα» (Discriminant). Η ανάλυση κατά συστάδες είναι μια ιδιαίτερα χρήσιμη μέθοδος καθώς:

- Επιτυγχάνεται ευκολότερη και αποδοτικότερη επεξεργασία των δεδομένων.
- Είναι ένα μέσον που βοηθάει τον εντοπισμό ακραίων τιμών.

Η χρήση της μεθόδου είναι απαραίτητη στις επιστήμες που αποσκοπούν στην δημιουργία ομοειδών ομάδων. Κατά συνέπεια η ομαδοποίηση δεδομένων βρίσκει ευρεία εφαρμογή στην Ιατρική, στη Βιολογία, σε Κοινωνικές, Πολιτικές, Οικονομικές επιστήμες κ.λ.π.

- Ανάλυση Αντιστοιχιών (Correspondence Analysis)

Η ανάλυση αντιστοιχιών είναι μια μέθοδος ανάλυσης πολυδιάστατων κατηγορικών δεδομένων. Πρόκειται για μια τεχνική ανάλογη της «Ανάλυσης σε Κύριες Συνιστώσες» με τη διαφορά πως η ανάλυση αντιστοιχιών αφορά κατηγορικά δεδομένα. Στην περίπτωση αυτή, ο ερευνητής κατασκευάζει ένα πίνακα δύο (απλή ανάλυση αντιστοιχιών) ή περισσότερων (πολλαπλή ανάλυση αντιστοιχιών) διαστάσεων στον οποίο εμπεριέχεται κάποια «αντιστοιχία» μεταξύ γραμμών και στηλών. Σκοπός είναι η μετατροπή του πίνακα δεδομένων σε μια γραφική αναπαράσταση έτσι ώστε να γίνουν εμφανείς οι συσχετισμοί μεταξύ των

χαρακτηριστικών στα οποία στηρίζεται ο πίνακας. Οι πίνακες συχνοτήτων και συνάφειας είναι χαρακτηριστικό παράδειγμα πινάκων που μελετώνται με την μέθοδο αυτή. Η τεχνική της Ανάλυσης Αντιστοιχιών είναι ιδιαίτερα χρήσιμη καθώς:

- Εντοπίζει τον συσχετισμό και τη διάταξη μεταξύ γραμμών και στηλών των δεδομένων.
- Ελέγχει εάν τα ποσοστά στηλών διαφοροποιούνται μεταξύ των γραμμών και αντίστροφα. Επομένως εξετάζεται κατά πόσο οι γραμμές και οι στήλες είναι ανεξάρτητες.
- Δημιουργεί καινούριες μεταβλητές οι οποίες συνοψίζουν σημαντικό μέρος της αρχικής πληροφορίας.

Η ανάλυση αντιστοιχιών βρίσκει μεγάλη εφαρμογή στις κοινωνικές επιστήμες καθώς και στην ανάλυση εταιρικών χαρακτηριστικών.

- Ανάλυση κανονικών συσχετίσεων (Canonical Correlation Analysis)

Η ανάλυση κανονικών συσχετίσεων είναι επίσης μια μέθοδος παρόμοια με τη μέθοδο των κύριων συνιστωσών. Στην πραγματικότητα, αντί να διερευνάται ένα σύνολο μεταβλητών όπου οι αρχικές μεταβλητές αντικαθίστανται από τις κύριες συνιστώσες, διερευνώνται οι συσχετίσεις μεταξύ δύο συνόλων μεταβλητών. Επομένως, επικεντρώνεται στη συσχέτιση ανάμεσα σε ένα γραμμικό συνδυασμό των μεταβλητών του ενός συνόλου και σε ένα γραμμικό συνδυασμό του άλλου συνόλου με σκοπό να βρεθεί ο βέλτιστος αριθμός γραμμικών συνδυασμών που έχουν τη μεγαλύτερη δυνατή συσχέτιση. Από αυτή την άποψη θα μπορούσε κανείς να εντοπίσει πολλές ομοιότητες της μεθόδου Ανάλυσης Κανονικών Συσχετίσεων με την Πολλαπλή Παλινδρόμηση καθώς και οι δύο τεχνικές συσχετίζουν την εξαρτημένη μεταβλητή με ένα σύνολο ανεξάρτητων προσβλεπουσών μεταβλητών. Η διαφορά έγκειται στο γεγονός πως στην περίπτωση της ανάλυσης κανονικών συσχετίσεων υπάρχει ένα πλήθος εξαρτημένων μεταβλητών και κατ' επέκταση υπάρχουν πολλοί τρόποι να συνδυαστούν οι μεταβλητές του ενός συνόλου με τις μεταβλητές του άλλου. Η ανάλυση κανονικών συσχετίσεων έχει ένα ευρύ φάσμα εφαρμογών όπως στην επιστήμη της ψυχολογίας, τις κοινωνικές επιστήμες, την οικονομία και την βιολογία.

## 1.4 Γραφικές μέθοδοι στην Πολυμεταβλητή Ανάλυση

Όπως προαναφέρθηκε η πολυμεταβλητή ανάλυση μελετάει παρατηρήσεις καθεμία εκ των οποίων αφορά συγκεκριμένα χαρακτηριστικά. Επομένως η παρουσίαση των δεδομένων αυτών δεν είναι εύκολη και επιτυγχάνεται είτε με τη χρήση πινάκων, είτε με τη χρήση διαφόρων γραφικών μεθόδων.

Στη μονομεταβλητή περίπτωση η γραφική απεικόνιση των δεδομένων είναι ένα χρήσιμο εργαλείο το οποίο βοηθάει ιδιαίτερα στην εξαγωγή συμπερασμάτων. Αντίθετα, η γραφική απεικόνιση των πολυδιάστατων δεδομένων είναι αρκετά δύσκολη και ακόμα πιο περίπλοκη είναι η ερμηνεία τους. Στην περίπτωση αυτή τα δεδομένα παρουσιάζονται στον χώρο των δύο διαστάσεων έχοντας όμως το κόστος της απώλειας πληροφορίας. Υπάρχουν διάφορες τεχνικές ταξινόμησης των πολυμεταβλητών γραφικών μεθόδων, με επικρατέστερη όλων αυτή που πρότειναν οι Keim και Kriegel (1996) βάσει της οποίας γίνεται η εξής κατηγοριοποίηση:

- Τυπικές Τεχνικές (Standard Techniques)
    - Πολλαπλό διάγραμμα διασποράς (Multiple Scatter plots)
    - Διάγραμμα Profile (Profile plot)
- Γεωμετρικές Τεχνικές (Geometric Techniques)
    - Καμπύλες του Andrews (Andrew’s curves)
    - Διάγραμμα Παράλληλων Συντεταγμένων ( Parallel Coordinate Plot)
- Εικονογραφικές Τεχνικές (Icon-Based Techniques)
    - Πρόσωπα Chernoff (Chernoff Faces)
    - Διάγραμμα Αστέρα (Star plot)

Στο σημείο αυτό θα παρουσιάσουμε τα χαρακτηριστικά των παραπάνω γραφικών μεθόδων:

### Α) Τυπικές Τεχνικές

Οι Τυπικές τεχνικές αποτελούν γενίκευση των απλών γραφικών τεχνικών της μονομεταβλητής περίπτωσης. Πρόκειται για γραφήματα απλά τόσο στην κατασκευή τους όσο και στην κατανόησή τους. Δύο γνωστές τυπικές τεχνικές είναι οι ακόλουθες:

- Πολλαπλό διάγραμμα διασποράς

Το πολλαπλό διάγραμμα διασποράς, είναι ενδεχομένως η πιο απλή περίπτωση γραφήματος στην πολυμεταβλητή περίπτωση καθώς βασίζεται στη λογική του απλού διαγράμματος διασποράς. Ο αναγνώστης που είναι εξοικειωμένος με την ερμηνεία ενός απλού διαγράμματος διασποράς είναι αρκετά εύκολο να ερμηνεύσει τα αποτελέσματα του πολλαπλού διαγράμματος. Το πολλαπλό διάγραμμα διασποράς περιλαμβάνει την κατασκευή τόσων διαγραμμάτων όσα και τα ζεύγη των υπό μελέτη μεταβλητών. Επομένως συγκρίνοντας την γραφική απεικόνιση των σχέσεων που αναπτύσσονται ανά δύο μεταβλητές, προκύπτουν χρήσιμα συμπεράσματα που δίνουν πληροφορίες για το σύνολο των μεταβλητών.

- Διαγράμματα Profile

Το διάγραμμα Profile βασίζεται και αυτό στην τεχνική της απεικόνισης όλων των συνδυασμών των μεταβλητών που μελετώνται σε ένα γράφημα. Πιο αναλυτικά, κάθε παρατήρηση απεικονίζεται με ένα ιστόγραμμα (histogram) ή ένα γράφημα περιοχής (area plot) το οποίο εμφανίζει τις μετρήσεις της παρατήρησης για κάθε μεταβλητή. Στη συνέχεια, συγκρίνονται όλα τα γραφήματα και προκύπτουν συμπεράσματα για τις ομοιότητες και τις διαφορές των παρατηρήσεων.

**Β) Γεωμετρικές τεχνικές**

Στις Γεωμετρικές τεχνικές τα δεδομένα τοποθετούνται σε γεωμετρικά σχήματα ή γραφικές παραστάσεις έτσι ώστε να γίνονται φανερές οι σχέσεις μεταξύ των μεταβλητών. Δύο από τις πιο γνωστές Γεωμετρικές τεχνικές είναι οι ακόλουθες:

- Καμπύλες του Andrews:

Οι καμπύλες του Andrews είναι μια μέθοδος γραφικής απεικόνισης πολυμεταβλητών δεδομένων που δημιουργήθηκε από τον Andrews (1972). Ο Andrews πρότεινε μια τεχνική

που συμπεριλαμβάνει τον υπολογισμό περιοδικών συναρτήσεων και τη μεταξύ τους σύγκριση. Πιο αναλυτικά, για κάθε παρατήρηση απεικονίζεται γραφικά η συνάρτηση:

$$f_x(t) = x_1/\sqrt{2} + x_2 \cdot \sin(t) + x_3 \cdot \cos(t) + x_4 \cdot \sin(2t) + x_5 \cdot \cos(2t) + \cdots$$

όπου $x_1$ ,$x_2$,…,$x_p$ είναι το σύνολο των μεταβλητών και $t \epsilon [-\pi, \pi]$.

Προφανώς κάθε καμπύλη αποτελεί ένα γραμμικό συνδυασμό των μεταβλητών. Στη συνέχεια, όλες οι καμπύλες τοποθετούνται σε ένα γράφημα. Οι αποστάσεις μεταξύ των καμπύλων αλλά και η μορφή τους μπορούν να αναδείξουν ομαδοποιήσεις των παρατηρήσεων. Η μέθοδος είναι ιδιαιτέρως αποτελεσματική ενώ ταυτόχρονα αποτελεί ένα χρήσιμο εργαλείο για τον εντοπισμό ακραίων τιμών. Εν κατακλείδι, με τις καμπύλες Andrews είναι αρκετά εύκολο ένας μελετητής να έχει μια εικόνα που αφορά στην ύπαρξη ή μη ομοιότητας μεταξύ των παρατηρήσεων.

- Διάγραμμα Παράλληλων Συντεταγμένων

Το διάγραμμα παράλληλων συντεταγμένων είναι μια τεχνική απεικόνισης πολυμεταβλητών δεδομένων που βασίζεται στη μετατροπή του καρτεσιανού συστήματος συντεταγμένων σε σύστημα παράλληλων συντεταγμένων. Οι μεταβλητές αναπαρίστανται ως παράλληλοι άξονες πάνω στους οποίους τοποθετούνται με τη μορφή σημείων οι μετρήσεις για κάθε παρατήρηση. Η πολυδιάστατη παρατήρηση απεικονίζεται με μια τεθλασμένη γραμμή που συνδέει τις μεταβλητές. Η μέθοδος αυτή ενδείκνυται για την σύγκριση των παρατηρήσεων, για την ομαδοποίησή τους και τον εντοπισμό ακραίων τιμών. Πρόκειται επομένως για μια ιδιαίτερα χρήσιμη γραφική απεικόνιση που γίνεται περισσότερο αποτελεσματική όταν το δείγμα δεν αποτελείται από μεγάλο αριθμό μεταβλητών και δεδομένων.

**Γ) Εικονογραφικές Τεχνικές**

Στις Εικονογραφικές Τεχνικές η κάθε παρατήρηση απεικονίζεται με μια εικόνα, της οποίας κάθε χαρακτηριστικό αντιστοιχεί σε μία μεταβλητή. Τρεις από τις πιο γνωστές Εικονογραφικές Τεχνικές είναι οι ακόλουθες:

- Πρόσωπα του Chernoff

Μία ενδιαφέρουσα γραφική απεικόνιση πολυμεταβλητών δεδομένων γίνεται με τη χρήση προσώπων γνωστών και ως πρόσωπα του Chernoff. Τα πρόσωπα αυτά είναι μία τεχνική απεικόνισης πολυδιάστατων δεδομένων που δημιουργήθηκε από τον μαθηματικό Herman Chernoff (1973). Κάθε πρόσωπο που κατασκευάζεται, αντιστοιχεί σε ένα άτομο του πληθυσμού που μελετάται. Παράλληλα, τα χαρακτηριστικά του προσώπου αντιστοιχούν στις μεταβλητές που έχει ο μελετητής στη διάθεσή του. Επομένως διαφορετικές διαστάσεις των μεταβλητών αυτών αναλογούν σε διαφορετικές διαστάσεις των χαρακτηριστικών του προσώπου. Η τεχνική αυτή, δεν βοηθάει στον προσδιορισμό της ακριβούς τιμής της εκάστοτε μεταβλητής, αλλά δείχνει την «τάση» που έχει η μεταβλητή αυτή για το πρόσωπο που μελετάται. Συνεπώς, η γνώση των τάσεων, θα μπορούσε να καθορίσει τα τμήματα των δεδομένων που παρουσιάζουν περισσότερο ενδιαφέρον. Σύμφωνα με την πρόταση που έκανε ο Chernoff το μήκος της μύτης, το μέγεθος των ματιών, το σχήμα του προσώπου κ.λ.π. προσδιορίζουν τη διάσταση του χαρακτηριστικού που μελετάται. Τα πρόσωπα του Chernoff αντικαθιστούν τη χρήση μεγάλων και δυσανάγνωστων πινάκων δεδομένων παρουσιάζοντάς τα με αρκετά ενδιαφέροντα τρόπο. Παρότι η μέθοδος αποτελεί ένα χρήσιμο εργαλείο για τον διαχωρισμό των δεδομένων σε ομάδες, η υποκειμενική εκχώρηση των εκφράσεων του προσώπου σε μεταβλητές, οδηγεί συχνά σε μοιραία λάθη κατά την ταξινόμηση.

- Διάγραμμα Αστέρας

Το διάγραμμα Αστέρας είναι μια μέθοδος γραφικής απεικόνισης πολυδιάστατων δεδομένων που προτάθηκε από τον Chambers (1983). Πρόκειται για ένα γράφημα στο οποίο, κάθε παρατήρηση που μελετάται αναπαρίσταται με ένα αστέρι του οποίου κάθε ακτίνα αντιστοιχεί σε μία μεταβλητή. Το μέγεθος κάθε ακτίνας αποτελεί τη μέτρηση της παρατήρησης για τη μεταβλητή αυτή. Επομένως κρίνοντας από την εικόνα που προκύπτει είναι εύκολο να εντοπιστούν ομοιότητες ή διαφορές μεταξύ των παρατηρήσεων κι έτσι η τεχνική αυτή είναι ιδιαιτέρως χρήσιμη όταν μελετάται ο βαθμός συμφωνίας μεταξύ των δεδομένων. Παρόλα αυτά αξίζει να σημειωθεί πως οι γραφικές μέθοδοι είναι δύσκολο να ερμηνευθούν όταν ο ερευνητής έχει στην κατοχή του ένα μεγάλο πλήθος δεδομένων και μεταβλητών.

# ΚΕΦΑΛΑΙΟ 2

# Πολυμεταβλητές τεχνικές απεικόνισης

## 2.1 Εισαγωγή

Η μελέτη πολλών μεταβλητών απαιτεί την παρατήρηση της διάρθρωσης αυτών των μεταβλητών και των σχέσεων που αναπτύσσονται μεταξύ των δεδομένων. Όπως έχει προαναφερθεί, η πολυμεταβλητή ανάλυση αποτελεί επέκταση της μονομεταβλητής ανάλυσης. Η διαφορά τους είναι πως στην περίπτωση της πολυμεταβλητής ανάλυσης το δείγμα αποτελείται από ένα πλήθος δεδομένων κάθε ένα εκ των οποίων χαρακτηρίζεται από μία σειρά μετρήσεων. Εξαιτίας της ιδιαιτερότητας αυτής της πολυμεταβλητής ανάλυσης, είναι αναγκαία η χρήση πιο πολύπλοκων μεθόδων για την αναπαράσταση των σχέσεων που υπάρχουν μεταξύ των δεδομένων. Δύο διαφορετικές προσεγγίσεις έχουν αναπτυχθεί για το σκοπό αυτό. Η πρώτη προσέγγιση αφορά στην απεικόνιση όλων των μεταβλητών σε έναν πολυδιάστατο χώρο με τη χρήση μαθηματικών και γραφικών μεθόδων. Η δεύτερη προσέγγιση αφορά στην μείωση των διαστάσεων των αντικειμένων τα οποία μετατρέπονται σε διανύσματα με λιγότερες μεταβλητές. Η προσέγγιση αυτή θα αναλυθεί διεξοδικότερα στο επόμενο κεφάλαιο.

Η μαθηματική ή γραφική απεικόνιση των δεδομένων εξυπηρετείται με τη χρήση πινάκων διαστάσεων $n \times p$ , όπου $n$ είναι το πλήθος των ατόμων και $p$ το πλήθος των χαρακτηριστικών ανά άτομο. Τα περιγραφικά μέτρα που χρησιμοποιούνται στην πολυμεταβλητή ανάλυση θα μπορούσαν να θεωρηθούν γενίκευση των αντίστοιχων μέτρων της μονομεταβλητής ανάλυσης. Όμως, δεν είναι εύκολο να γενικευθούν όλα τα περιγραφικά μέτρα στην μελέτη των πολυμεταβλητών δεδομένων. Τα βασικότερα μέτρα που χρησιμοποιούνται είναι:

- Ο δειγματικός μέσος (Sample Mean)
- Η δειγματική διασπορά (Sample Variance)

- Ο συντελεστής συσχέτισης (Correlation Coefficient)

Τα μέτρα αυτά θα αναλυθούν διεξοδικότερα στη συνέχεια.

Η γραφική απεικόνιση των πολυμεταβλητών δεδομένων γίνεται με χρήση κατάλληλων γραφημάτων προσαρμοσμένα στον διδιάστατο χώρο, χάνοντας βέβαια ένα μέρος της πληροφορίας που θα γινόταν αντιληπτή σε ένα χώρο περισσότερων διαστάσεων. Όπως προαναφέρθηκε σε προηγούμενη ενότητα, υπάρχουν διάφορες πολυμεταβλητές γραφικές μέθοδοι, κυριότερες από τις οποίες είναι οι εξής:

- Τυπικές Τεχνικές (Standard Techniques)
  - Πολλαπλό διάγραμμα διασποράς (Multiple Scatter plots)
  - Διάγραμμα Profile (Profile Plot)
- Γεωμετρικές Τεχνικές (Geometric Techniques)
  - Καμπύλες του Andrews (Andrew's curves)
  - Διάγραμμα Παράλληλων Συντεταγμένων ( Parallel Coordinate Plot)
- Εικονογραφικές Τεχνικές (Icon-Based Techniques)
  - Πρόσωπα Chernoff (Chernoff Faces)
  - Διάγραμμα Αστέρα (Star plot)

Στο παρόν κεφάλαιο αναλύονται οι μαθηματικές και γραφικές πολυμεταβλητές μέθοδοι και γίνεται εφαρμογή κάθε μεθόδου σε ένα ενδεικτικό δείγμα δεδομένων.

## 2.2 Περιγραφικά μέτρα

Η μαθηματική ή γραφική απεικόνιση των πολυμεταβλητών δεδομένων, όπως έχει ήδη αναφερθεί, εξυπηρετείται με τη χρήση πινάκων διαστάσεων $n \times p$ , όπου $n$ είναι το πλήθος των ατόμων και $p$ το πλήθος των χαρακτηριστικών ανά άτομο. Ο πίνακας έχει τη μορφή:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Κάθε γραμμή του πίνακα $X$ αντιπροσωπεύει μια πολυμεταβλητή παρατήρηση. Επομένως τα δεδομένα είναι ένα δείγμα $n$ μετρήσεων κάθε μια από τις οποίες χαρακτηρίζεται από $p$ μεταβλητές.

Στο σημείο αυτό θα παρουσιαστούν και θα αναλυθούν οι πιο βασικές μαθηματικές μέθοδοι μελέτης και ερμηνείας των πολυμεταβλητών δεδομένων. Το ενδεικτικό δείγμα που

χρησιμοποιήθηκε για τον σκοπό αυτό προήλθε από έρευνα του SHARE (Survey of Health and Retirement in Europe) και συγκεκριμένα από την ιστοσελίδα *http://www.share-project.gr/*. Το SHARE είναι μια βάση δεδομένων η οποία αντλεί στοιχεία από διάφορους επιστημονικούς κλάδους και από διάφορες χώρες, ενώ ταυτόχρονα λειτουργεί διαχρονικά μέσω ενός επιλεγμένου δείγματος ατόμων. Τα δεδομένα του, αφορούν στοιχεία υγείας, κοινωνικά, οικονομικά και δημογραφικά για περισσότερα από 40000 άτομα ηλικίας 50 ετών και άνω. Η συλλογή και μελέτη των δεδομένων ξεκίνησε το 2004 με τη συμμετοχή έντεκα χωρών της Ευρώπης οι οποίες αποτέλεσαν το πρώτο κύμα. Στη συνέχεια προστέθηκαν άτομα από τέσσερις νέες χώρες κατά το έτος 2006 δημιουργώντας με αυτό τον τρόπο το δεύτερο κύμα δεδομένων, ενώ τέλος με την προσθήκη δεδομένων από την Σλοβενία κατά το 2008-09 ολοκληρώθηκε και το τρίτο κύμα δεδομένων. Στο παρόν κεφάλαιο, επιλέχθηκε ένα ενδεικτικό δείγμα 8 παρατηρήσεων από το 1$^{ο}$ κύμα συλλογής δεδομένων του SHARE. Πρόκειται για 4 χαρακτηριστικά (μεταβλητές) των παρατηρήσεων αυτών που σχετίζονται με την οικονομική κατάσταση της εκάστοτε οικογένειας. Οι πηγές εισοδήματος που συμμετέχουν στην διαμόρφωση του οικογενειακού εισοδήματος προέρχονται είτε από συντάξεις είτε από άλλες δραστηριότητες και αφορούν ολόκληρη την οικογένεια. Οι μεταβλητές που θα μελετηθούν είναι το «Ακαθάριστο Οικογενειακό εισόδημα» (Yhh_nir), το «Καθαρό Οικογενειακό» (Yhh_net), το «Οικογενειακό ετήσιο εισόδημα από σύνταξη γήρατος» (Hhold_Pensions), και το «Οικογενειακό ετήσιο εισόδημα από πηγές εκτός εργασίας και συντάξεων» (Other_source) και έχουν ως μονάδα μέτρησης το Ευρώ (€).

Στον πίνακα που ακολουθεί καταγράφονται τα περιουσιακά στοιχεία των 8 παρατηρήσεων του δείγματος. Ενδεικτικά θα μπορούσαμε να αναφέρουμε πως το στοιχείο $x_{21}$ του Πίνακα 2.2.1 αντιστοιχεί στο πρώτο χαρακτηριστικό της δεύτερης παρατήρησης δηλαδή στη μεταβλητή «Ακαθάριστο Οικογενειακό εισόδημα» της δεύτερης παρατήρησης και ανέρχεται στο ποσό των 105.136 €.

**ΠΙΝΑΚΑΣ 2.2.1**

Περιουσιακά στοιχεία νοικοκυριών

| Παρατήρηση | Ακαθάριστο Οικογενειακό εισόδημα | Καθαρό Οικογενειακό εισόδημα | Εισόδημα από σύνταξη γήρατος | Εισόδημα από πηγές εκτός εργασίας και συντάξεων |
|---|---|---|---|---|
| **1** | 81400 | 67071,73 | 23800 | 57600 |
| **2** | 105136 | 81349,21 | 10000 | 59646 |
| **3** | 102540 | 83451,64 | 7600 | 31240 |
| **4** | 144082 | 118720,3 | 0 | 144082 |
| **5** | 24685,7 | 21985,92 | 19600 | 5085,7 |
| **6** | 9100 | 9100 | 9100 | 0 |
| **7** | 12768 | 12522,05 | 12768 | 0 |
| **8** | 30612,3 | 28061,44 | 27300 | 3312,3 |

### 2.2.1 Δειγματικός Μέσος

Το πιο διαδεδομένο περιγραφικό μέτρο που χαρακτηρίζεται ως «μέτρο θέσης» είναι η μέση τιμή. Στην περίπτωση των πολυδιάστατων δεδομένων ο μέσος όρος είναι ένα διάνυσμα που περιέχει τις μέσες τιμές για κάθε μεταβλητή. Ο δειγματικός μέσος ενός δείγματος $n$ παρατηρήσεων που αφορά στην $j$ μεταβλητή βρίσκεται από τον τύπο:

$$\bar{x_j} = \frac{\sum_{i=1}^{n} x_{ij}}{n} \text{ όπου } j=1,2,3,..,p$$

Όπως στην μονομεταβλητή ανάλυση, έτσι και στην πολυμεταβλητή το πλεονέκτημα της μέσης τιμής είναι ο εύκολος υπολογισμός της. Αντίθετα, το μειονέκτημα της είναι ότι η τιμή της επηρεάζεται από τις ακραίες τιμές του δείγματος.

Με χρήση του στατιστικού πακέτου *IBM SPSS Statistics 20.0* υπολογίστηκε η μέση τιμή για καθεμία από τις τέσσερις μεταβλητές του δείγματος όπως φαίνεται ακολούθως:

**ΠΙΝΑΚΑΣ 2.2.2**

Μέση τιμή μεταβλητών

| | **Yhh_nir** | **Yhh_net** | **Hhold_Pensions** | **Other_source** |
|---|---|---|---|---|
| **Μέση τιμή** | 63790,5 | 52782,78 | 13771 | 37620,75 |

Το διάνυσμα των δειγματικών μέσων μπορεί να γραφεί ως διάνυσμα-στήλη με την ακόλουθη μορφή:

$$\bar{\boldsymbol{x}}=\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\\ \vdots\\ \bar{x}_p\end{bmatrix}$$

Το διάνυσμα αυτό μπορεί να θεωρηθεί ως το κέντρο του δείγματος.

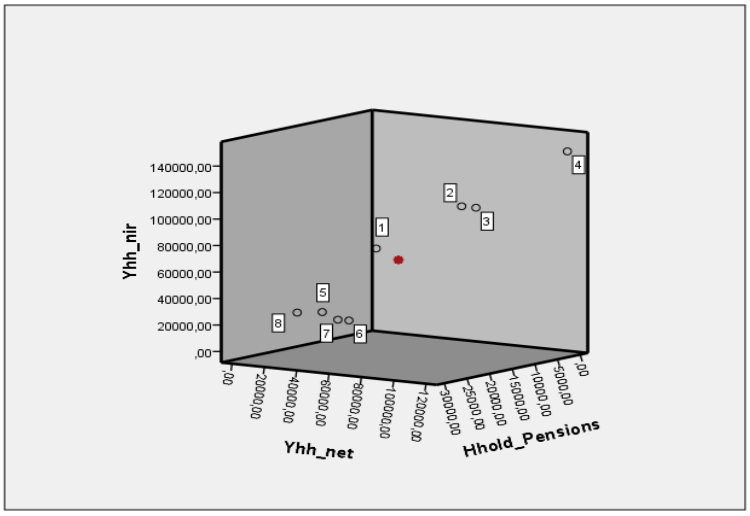
Σύμφωνα με τα παραπάνω, το διάνυσμα των δειγματικών μέσων των μεταβλητών της εφαρμογής μας είναι το ακόλουθο:

$$\bar{\boldsymbol{x}}=\begin{bmatrix}63790{,}5\\ 52782{,}78\\ 13771\\ 37620{,}75\end{bmatrix}$$

Στο σχήμα που ακολουθεί παρουσιάζεται η γραφική απεικόνιση των παρατηρήσεων, χρησιμοποιώντας τις μετρήσεις των τριών πρώτων μεταβλητών:

**ΣΧΗΜΑ 2.2.1**

Τρισδιάστατη απεικόνιση παρατηρήσεων και μέσης τιμής




Όπως φαίνεται το σχήμα απεικονίζει την θέση της κάθε παρατήρησης στον τρισδιάστατο χώρο, με βάση τις μετρήσεις της για τα τρία πρώτα οικονομικά χαρακτηριστικά. Ο κάθετος άξονας αντιστοιχεί στην μεταβλητή «Ακαθάριστο Οικογενειακό εισόδημα» και οι δύο

οριζόντιοι άξονες αντιστοιχούν στις μεταβλητές «Καθαρό Οικογενειακό» και «Οικογενειακό ετήσιο εισόδημα από σύνταξη γήρατος». Η κόκκινη κουκίδα που βρίσκεται στον κέντρο του σχήματος αντιστοιχεί στο διάνυσμα των μέσων τιμών των παρατηρήσεων για τις μεταβλητές αυτές.

### 2.2.2 Δειγματική Διασπορά

Στην μονομεταβλητή περίπτωση, η δειγματική διασπορά χρησιμοποιείται συχνά για να περιγράψει την μεταβλητότητα των μετρήσεων της εν λόγω μεταβλητής, γεγονός στο οποίο οφείλει την κατηγοριοποίησή της στα «μέτρα μεταβλητότητας». Αντίστοιχα, στην πολυμεταβλητή περίπτωση, η διακύμανση αποτελεί την συνδιακύμανση μιας μεταβλητής με τον εαυτό της, γι αυτό η μεταβλητότητα περιγράφεται από έναν πίνακα διακυμάνσεων-συνδιακυμάνσεων που έχει τη μορφή:

$$\boldsymbol{S}=\begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{pp} \end{bmatrix}$$

Ο πίνακας διακυμάνσεων – συνδιακυμάνσεων $S$ είναι ένας συμμετρικός πίνακας του οποίου τα διαγώνια στοιχεία είναι οι διακυμάνσεις των μεταβλητών και προκύπτουν από τον τύπο:

$$s_{ii}=\frac{1}{n}\sum_{j=1}^{n}\left(x_{ij}-\bar{x_{\iota}}\right)^2 \quad \text{όπου } i=1,2,\ldots,n$$

Τα υπόλοιπα στοιχεία εκτός της διαγωνίου ($s_{ij}\ με\ i \neq j$) είναι οι συνδιακυμάνσεις των μεταβλητών αυτών. Ο πίνακας αυτός είναι χρήσιμος αφενός για τη μελέτη της μεταβλητότητας του κάθε χαρακτηριστικού ξεχωριστά και αφετέρου για τη μελέτη της συνολικής μεταβλητότητας του δείγματος. Η συνολική μεταβλητότητα εντοπίζεται με τον υπολογισμό:

- Της συνολικής διακύμανσης (Total Sample Variance)
- Της γενικευμένης διακύμανσης (Generalized Sample Variance)

Η συνολική διακύμανση (TSV) είναι το άθροισμα των διακυμάνσεων των μεταβλητών οπότε αποτελεί το «ίχνος» του πίνακα διακυμάνσεων-συνδιακυμάνσεων, δηλαδή ισχύει ότι:

$$\text{TSV}= s_{11}+s_{22}+\cdots+s_{pp}=Tr(S)$$

Αν και είναι εύκολη στον υπολογισμό, η συνολική διακύμανση δεν είναι ακριβής καθώς δεν λαμβάνει υπόψη τις συσχετίσεις μεταξύ των μεταβλητών και χάνεται έτσι μεγάλο μέρος της πληροφορίας σχετικά με τη δομή των δεδομένων.

Η γενικευμένη διακύμανση (GSV) είναι η ορίζουσα του πίνακα $S$, δηλαδή ισχύει ότι:

$$\text{GSV}= |S|.$$

Το μέτρο αυτό αποτελεί ένα δείκτη που φανερώνει πόσο απομακρυσμένες είναι οι παρατηρήσεις από τον δειγματικό μέσο $\bar{X}$ στον $p$-διάστατο ευκλείδειο χώρο. Για να γίνει πιο κατανοητό αυτό, ας υποθέσουμε πως έχουμε ένα ελλειψοειδές με κέντρο $\bar{\boldsymbol{x}}$ στον $R^p$ χώρο το οποίο περιγράφει μια εξίσωση όπου όλα τα σημεία της ισαπέχουν από το μέσο $\bar{\boldsymbol{x}}$ και ορίζεται από τη σχέση:

$$(\boldsymbol{x}-\bar{\boldsymbol{x}})'S^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}) = c^2$$

Επομένως ισχύει ότι:

$$E = \{\boldsymbol{x} \in R^p : (\boldsymbol{x}-\bar{\boldsymbol{x}})'S^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}) \leq c^2\}$$

και κατ' επέκταση, ο όγκος του ελλειψοειδούς $E$ προκύπτει ως:

$$\text{Όγκος } E=k_p\sqrt{|S|}\,c^2,$$

όπου $k_p = \frac{2\pi^{p/2}}{(p\Gamma\left(\frac{p}{2}\right))}$

Συνεπώς, όσο πιο μικρή τιμή έχει η $|S|$, τόσο πιο μικρό Όγκο έχει το Ε και κατ' επέκταση, τόσο πιο κοντά είναι οι παρατηρήσεις $\boldsymbol{x_i}$ στο δειγματικό μέσο $\bar{x}$.

Η γενικευμένη διακύμανση, αν και είναι περισσότερο ακριβής από τη συνολική διακύμανση, δεν διευκολύνει τη στατιστική συμπερασματολογία.

Σύμφωνα με τα παραπάνω, υπολογίσθηκε ο πίνακας διακυμάνσεων- συνδιακυμάνσεων των δεδομένων του παραδείγματος:

$$S = \begin{bmatrix} 2598874031 & 1797885844 & -211787559 & 1993840468 \\ 1797885844 & 1628178118 & -166406086 & 1593363307 \\ -211787559 & -166406086 & 83194042{,}3 & -218716660 \\ 1993840468 & 1593363307 & -218716660 & 2475033524 \end{bmatrix}$$

Η Συνολική διακύμανση είναι:

$$\text{TSV}= tr\,(S) = 6785279716$$

και η Γενικευμένη διακύμανση προκύπτει ως εξής:

$$GSV = |S| = 5{,}15359 \cdot 10^{34}$$

### 2.2.3 Συντελεστής Συσχέτισης

Το σημαντικότερο ίσως μέτρο για τη μελέτη πολυμεταβλητών δεδομένων είναι ο συντελεστής συσχέτισης. Ο συντελεστής αυτός είναι απαραίτητος για να εντοπισθεί μια πιθανή συσχέτιση μεταξύ των μεταβλητών του δείγματος και να υπολογισθεί το μέγεθος της συσχέτισης αυτής. Έχει ορισθεί ένας μεγάλος αριθμός διαφορετικών συντελεστών συσχέτισης. Τρεις από τους πιο γνωστούς δείκτες συσχέτισης είναι:

- Ο Συντελεστής Γραμμικής Συσχέτισης του Pearson
- Ο Συντελεστής Συσχέτισης του Spearman
- Ο δείκτης Kendall W

Ο Συντελεστής Γραμμικής Συσχέτισης του Pearson είναι ο πιο διαδεδομένος συντελεστής συσχέτισης. Χρησιμοποιείται αποκλειστικά σε ποσοτικά δεδομένα και συμβολίζεται με το γράμμα *r*. Για να υπολογισθεί ο δείκτης *r* μεταξύ δύο μεταβλητών είναι απαραίτητο να έχει υπολογισθεί ο πίνακας διακυμάνσεων- συνδιακυμάνσεων αυτών:

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{pp} \end{bmatrix}$$

όπου $s_{ij}$ με $i \neq j$ είναι η συνδιακύμανση των μεταβλητών $x_i$ και $x_j$ και $s_{ii}$ είναι οι διακύμανση της μεταβλητής $x_i$.

Ο συντελεστής συσχέτισης του Pearson υπολογίζεται από τον τύπο:

$$r = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \text{ όπου } i, j = 1,2, \dots, p$$

και παίρνει τιμές στο διάστημα [-1,+1].

Εάν ο συντελεστής *r*=0 τότε οι μεταβλητές είναι μεταξύ τους ασυσχέτιστες. Ένας θετικός συντελεστής συσχέτισης, αποδεικνύει θετική συσχέτιση μεταξύ των μεταβλητών και κατ' επέκταση όταν *r*=+1 σημαίνει πως οι μεταβλητές που μελετώνται έχουν τέλεια θετική συσχέτιση. Στην περίπτωση αυτή, η αύξηση της τιμής της μίας εκ των δύο μεταβλητών αυξάνει αυτόματα και την τιμή της άλλης και η μείωση της τιμής της μίας μεταβλητής, μειώνει και την τιμή της άλλης. Αντίθετα, ένας αρνητικός συντελεστής συσχέτισης, αποδεικνύει αρνητική συσχέτιση μεταξύ των μεταβλητών και κατ' επέκταση όταν $r = -1$

σημαίνει πως οι μεταβλητές που μελετώνται έχουν τέλεια αρνητική συσχέτιση. Στην περίπτωση αυτή, η αύξηση της τιμής της μίας εκ των δύο μεταβλητών μειώνει αυτόματα κατά μέσο όρο και την τιμή της άλλης και η μείωση της τιμής της μίας μεταβλητής, αυξάνει κατά μέσο όρο και την τιμή της άλλης. Το γεγονός αυτό αποδεικνύει πως θετικές τιμές του *r* δεν υποδηλώνουν απαραίτητα μεγαλύτερο βαθμό γραμμικής συσχέτισης σε σχέση με τις αρνητικές τιμές του *r*. Το πρόσημο καθορίζει μόνο το είδος της συσχέτισης, ενώ ο βαθμός συσχέτισης καθορίζεται από την απόλυτη τιμή του συντελεστή.

Ο Συντελεστής Συσχέτισης του Spearman είναι ένα μη-παραμετρικό μέτρο της στατιστικής εξάρτησης δύο μεταβλητών και συμβολίζεται με το ελληνικό γράμμα *ρ*. Στην πραγματικότητα, το μέτρο αυτό δεν είναι άλλο από το μέτρο *r* του Pearson υπολογιζόμενο με βάση τις τάξεις μεγέθους των παρατηρήσεων και όχι αυτές καθαυτές τις παρατηρήσεις. Δηλαδή, σε ένα δείγμα *n* παρατηρήσεων μπορεί να ονομαστεί $R(x_i)$ η τάξη μεγέθους της μεταβλητής *X* όταν αυτή συγκρίνεται με τις υπόλοιπες τιμές της για $i = 1,2, \dots, n$. Επομένως $R(x_i)$=1 εάν $x_i$ είναι η μικρότερη από τις *X*-τιμές και $R(x_i)$=*n* εάν $x_i$ είναι η μεγαλύτερη από τις *X*-τιμές. Ομοίως γίνεται η διάταξη της μεταβλητής *Y* όταν συγκρίνεται με τις υπόλοιπες τιμές της. Σύμφωνα με τα παραπάνω ο συντελεστής *ρ* του Spearman προκύπτει από τον τύπο:

$$\rho = 1 - \frac{6T}{n(n^2 - 1)}$$

όπου $T = \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2$.

Όπως και στην περίπτωση του Συντελεστή Γραμμικής Συσχέτισης του Pearson, ο συντελεστής του Spearman παίρνει τιμές στο διάστημα [-1,+1]. Ο συντελεστής συσχέτισης Spearman είναι προτιμότερος όταν οι μεταβλητές είναι διατάξιμες, ή όταν δεν υπάρχουν ενδείξεις πως τα δεδομένα ακολουθούν κανονική κατανομή.

Ο δείκτης *W* του Kendall, είναι κι αυτός ένα μη-παραμετρικό μέτρο το οποίο υπολογίζει το βαθμό συμφωνίας των *p* χαρακτηριστικών ενός δείγματος *n* δεδομένων. Χρησιμοποιείται κυρίως όταν μελετώνται ταυτόχρονα περισσότερες από 3 μεταβλητές, δεδομένου ότι κατά την μελέτη των δύο μεταβλητών αξιοποιείται ο συντελεστής συσχέτισης του Spearman. Ο δείκτης βασίζεται στην ιδέα πως κάθε παρατήρηση μπορεί να ταξινομηθεί παίρνοντας τη θέση 0 έως *n* για κάθε μεταβλητή ξεχωριστά. Στη συνέχεια ο μελετητής υπολογίζει πόσο οι βαθμοί αυτοί ταιριάζουν μεταξύ των μεταβλητών. Ο δείκτης του Kendall παίρνει τιμές από 0 που δείχνει την πλήρη ασυμφωνία έως 1 που δείχνει την πλήρη συμφωνία. Με βάση τα παραπάνω, αν θεωρήσουμε πως μια παρατήρηση ταξινομείται στην θέση $r_{ij}$ ως προς ένα

χαρακτηριστικό, τότε το άθροισμα των βαθμών ταξινόμησης της παρατήρησης *i* για το σύνολο *p* των μεταβλητών δίνεται ως εξής:

$$R_i = \sum_{j=1}^{p} r_{ij}$$

και συμβολίζοντας ως $\bar{R}$ τον μέσο των ταξινομήσεων και *Sum* το άθροισμα των τετραγωνικών αποκλίσεων του βαθμού ταξινόμησης από τον μέσο των παρατηρήσεων, δηλαδή:

$$Sum = \sum_{i=1}^{n} (R_i - \bar{R})^2$$

ο δείκτης *W* υπολογίζεται με τον τύπο:

$$W = \frac{12Sum}{p^2(n^3 - n)}$$

όπου *n* είναι το πλήθος των παρατηρήσεων και *p* το πλήθος των μεταβλητών που τις χαρακτηρίζουν.

Στο σημείο αυτό, με χρήση του στατιστικού πακέτου *IBM SPSS Statistics 20.0* υπολογίστηκε ο Συντελεστής Συσχέτισης του Pearson και του Spearman καθώς και ο δείκτης του Kendall, *W* για καθεμιά από τις τέσσερις μεταβλητές του παραδείγματος όπως φαίνεται ακολούθως:

**ΠΙΝΑΚΑΣ 2.2.3**

Συντελεστής Γραμμικής Συσχέτισης του Pearson

| | **Yhh_nir** | **Yhh_net** | **Hhold_Pensions** | **Other_source** |
|---|---|---|---|---|
| **Yhh_nir r** | 1 | 0,999*** | -0,521$^{n.s.}$ | 0,898*** |
| **Yhh_net r** | 0,999*** | 1 | -0,517$^{n.s.}$ | 0,907*** |
| **Hhold_Pensions r** | -0,521$^{n.s.}$ | -0,517$^{n.s.}$ | 1 | -0,551$^{n.s.}$ |
| **Other_source r** | 0,898*** | 0,907*** | -0,551$^{n.s.}$ | 1 |

***p-value<0.01, **p-value<0.05, *p-value<0.10, n.s.: not significant

Στον παραπάνω πίνακα υπολογίστηκε η τιμή του συντελεστή *r* του Pearson. Ουσιαστικά εξετάζεται η μηδενική υπόθεση $H_0$: ο συντελεστής γραμμικής συσχέτισης είναι μηδενικός έναντι της εναλλακτικής $H_1$: ο συντελεστής γραμμικής συσχέτισης δεν είναι μηδενικός. Από τα αποτελέσματα του πίνακα είναι φανερό πως παρατηρείται τέλεια θετική συσχέτιση μεταξύ των μεταβλητών Yhh_nir και Yhh_net όπου η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson είναι $r = 0{,}999$. Το γεγονός αυτό είναι αναμενόμενο καθώς δείχνει πως όσο αυξάνεται το ακαθάριστο οικογενειακό εισόδημα τόσο αυξάνεται και το καθαρό οικογενειακό εισόδημα. Ομοίως, ισχυρή θετική συσχέτιση παρατηρείται μεταξύ των ποσοτήτων Yhh_net και Other_source όπου η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson είναι *r*= 0,907 αλλά και μεταξύ των Yhh_nir και Other_source όπου *r*= 0,898.

Επομένως όσο αυξάνεται είτε το ακαθάριστο είτε το καθαρό οικογενειακό εισόδημα, τόσο αυξάνεται το οικογενειακό εισόδημα από άλλες πηγές. Τέλος, οι μεταβλητή Hhold_Pensions σχετίζεται αρνητικά με τις υπόλοιπες. Η συσχέτιση όμως που αναπτύσσεται μεταξύ αυτής και των υπολοίπων οικονομικών μεταβλητών δεν είναι στατιστικά σημαντική καθώς ο συντελεστής *r* του Pearson έχει σχετικά χαμηλή τιμή.

**ΠΙΝΑΚΑΣ 2.2.4**

Συντελεστής Συσχέτισης του Spearman

| | **Yhh_nir** | **Yhh_net** | **Hhold_Pensions** | **Other_source** |
|---|---|---|---|---|
| **Yhh_nir ρ** | 1 | 0,976*** | -0,381$^{n.s.}$ | 0,946*** |
| **Yhh_net ρ** | 0,976*** | 1 | -0,429$^{n.s.}$ | 0,898*** |
| **Hhold_Pensions ρ** | -0,381$^{n.s.}$ | -0,429$^{n.s.}$ | 1 | -0,335$^{n.s.}$ |
| **Other_source ρ** | 0,946*** | 0,898*** | -0,335$^{n.s.}$ | 1 |

***p-value<0.01, **p-value<0.05, *p-value<0.10, n.s.: not significant

Τα συμπεράσματα που προκύπτουν από τον Πίνακα 2.2.4 ταυτίζονται με τα αντίστοιχα συμπεράσματα που προέκυψαν κατά τη μελέτη της συσχέτισης με βάση τον συντελεστή Pearson. Πιο αναλυτικά, από τον πίνακα είναι φανερό πως παρατηρείται τέλεια θετική συσχέτιση μεταξύ των ποσοτήτων Yhh_nir και Yhh_net όπου η τιμή του συντελεστή συσχέτισης του Spearman είναι $\rho$= 0,976. Το γεγονός αυτό δείχνει πως μια αύξηση ή μείωση στην τιμή του ακαθάριστου οικογενειακού εισοδήματος, προκαλεί αντίστοιχη μεταβολή στο καθαρό οικογενειακό εισόδημα. Ομοίως, τέλεια θετική συσχέτιση παρατηρείται μεταξύ των ποσοτήτων Yhh_nir και Other_source όπου η τιμή του συντελεστή γραμμικής συσχέτισης του Spearman είναι $\rho$= 0,946. Ισχυρή θετική συσχέτιση παρατηρείται και μεταξύ των ποσοτήτων Yhh_net και Other_source με $\rho$=0,892. Τέλος, όπως παρατηρήθηκε και με τον συντελεστή συσχέτισης Pearson, η μεταβλητή Hhold_Pensions δεν έχει στατιστικά σημαντική συσχέτιση με καμία άλλη μεταβλητή.

Η διαδικασία ολοκληρώνεται με τον δείκτη του Kendall, *W*:

**ΠΙΝΑΚΑΣ 2.2.5**

Δείκτης Kendall's W

| | |
|---|---|
| **Kendall's W** | 0,596 |
| **p-value** | 0,003 |

Ο δείκτης του Kendall δείχνει τον βαθμό συμφωνίας των τεσσάρων μεταβλητών του παραδείγματος. Όπως φαίνεται από τον Πίνακα 2.2.5, υπάρχει ισχυρή και στατιστικά σημαντική συμφωνία μεταξύ των μεταβλητών με τιμή *W*= 0,596 και *p-value*=0,003.

## 2.3 Γραφική Απεικόνιση των πολυμεταβλητών δεδομένων

Τα γραφήματα προσφέρουν σημαντική βοήθεια στη μελέτη πολυμεταβλητών δεδομένων. Αν και η προσφορά τους έχει υποτιμηθεί καθώς είναι αδύνατη η ταυτόχρονη απεικόνιση όλων των μετρήσεων που πραγματοποιούνται σε διαφορετικές μεταβλητές, εξακολουθούν να είναι μια χρήσιμη πηγή πληροφοριών για την ανακάλυψη των σχέσεων που αναπτύσσονται μεταξύ αυτών. Η ταχεία ανάπτυξη της τεχνολογίας έχει βοηθήσει την αύξηση και εξέλιξη των γραφημάτων για πολυμεταβλητά δεδομένα. Σε γενικές γραμμές, οι πολυδιάστατες παρατηρήσεις μπορούν να αναπαρασταθούν σε δύο διαστάσεις, δίνοντας την δυνατότητα να εντοπίζονται οπτικά οι διακριτές ομάδες και οι ακραίες τιμές των μετρήσεων. Στην ενότητα που ακολουθεί γίνεται μια εκτενής αναφορά των πιο διαδεδομένων γραφικών μεθόδων της πολυμεταβλητής ανάλυσης και εφαρμόζονται στο δείγμα των δεδομένων του Πίνακα 2.2.1. Για την υλοποίηση των γραφικών απεικονίσεων χρησιμοποιήθηκαν τα στατιστικά προγράμματα *IBM SPSS Statistics 20.0,* η γλώσσα προγραμματισμού *R* καθώς και το *Excel.*

### 2.3.1 Τυπικές Τεχνικές

Οι τυπικές τεχνικές συμπεριλαμβάνουν όλες τις βασικές μεθόδους γραφικής απεικόνισης πολυμεταβλητών δεδομένων. Πρόκειται για γραφήματα που διακρίνονται για την ευκολία τους στην κατασκευή και ερμηνεία των δεδομένων γεγονός που κάνει τις τυπικές τεχνικές ιδιαίτερα διαδεδομένες. Η χρήση τους συνήθως περιορίζεται στα αρχικά στάδια της ανάλυσης και για σχετικά μικρό όγκο δεδομένων. Παρακάτω αναλύονται δύο από τις πιο δημοφιλείς τυπικές τεχνικές:

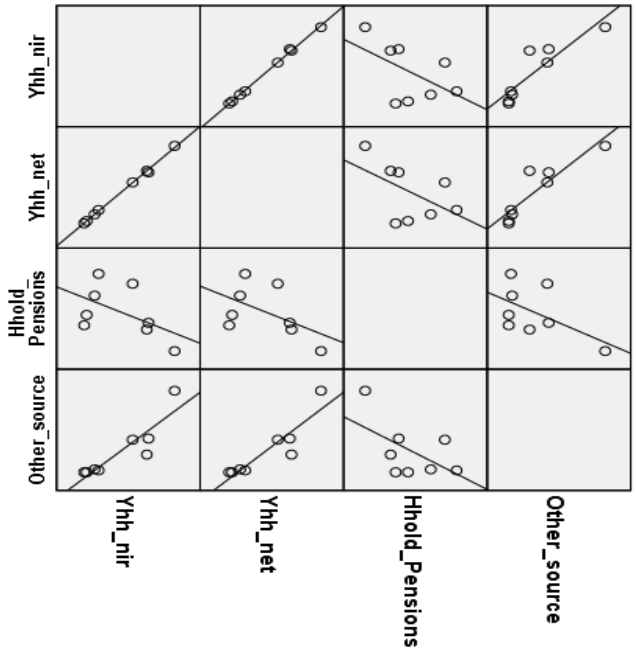
**Α) Πολλαπλά Διαγράμματα διασποράς**

Το διάγραμμα διασποράς πολυμεταβλητών δεδομένων αποτελεί μια γενίκευση του απλού διαγράμματος διασποράς μεταξύ δύο μεταβλητών. Σύμφωνα με την τεχνική αυτή, κάθε μεταβλητή συγκρίνεται γραφικά με κάθε άλλη μεταβλητή σε ένα πίνακα που αποτελείται από

τόσα διαγράμματα διασποράς όσα τα ζεύγη των μεταβλητών. Επομένως κατά την μελέτη $p$ μεταβλητών, το αποτέλεσμα συμπεριλαμβάνει $\frac{p(p-1)}{2}$ ξεχωριστά διαγράμματα, ενώ τα προσκείμενα γραφήματα έχουν πάντα ένα κοινό άξονα. Εφόσον το διάγραμμα διασποράς μιας μεταβλητής *x* ως προς μια μεταβλητή *y* είναι ίδιο με το διάγραμμα διασποράς της μεταβλητής *y* ως προς την *x*, ο πίνακας των γραφημάτων που δημιουργείται είναι προφανώς συμμετρικός. Επιπλέον, η διαγώνιος του διαγράμματος δεν παρέχει καμία πληροφορία αφού απεικονίζει τη διασπορά της κάθε μεταβλητής με την ίδια τη μεταβλητή. Ο πίνακας διαγραμμάτων διασποράς αποτελεί χρήσιμο εργαλείο στην περίπτωση της μελέτης λίγων μεταβλητών. Αντίθετα, όταν υπάρχει μεγάλος όγκος μεταβλητών είναι αρκετά πολύπλοκος και μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Με τη χρήση του πακέτου *IBM SPSS Statistics 20.0*, κατασκευάστηκε το πολλαπλό διάγραμμα διασποράς για τις τέσσερις μεταβλητές του Πίνακα 2.2.1 που αφορούν στα περιουσιακά στοιχεία 8 οικογενειών:

**ΣΧΗΜΑ 2.3.1**

Πολλαπλά διαγράμματα διασποράς




Από το Σχήμα 2.3.1 παρατηρείται μια ξεκάθαρα θετική συσχέτιση μεταξύ του Ακαθάριστου (Yhh_nir) και του Καθαρού (Yhh_net) οικογενειακού εισοδήματος. Στην περίπτωση αυτή οι τιμές των δύο μεταβλητών τείνουν να μεταβάλλονται προς την ίδια κατεύθυνση και ταυτόχρονα συμπίπτουν με την ευθεία γραμμικής συσχέτισης. Παράλληλα,

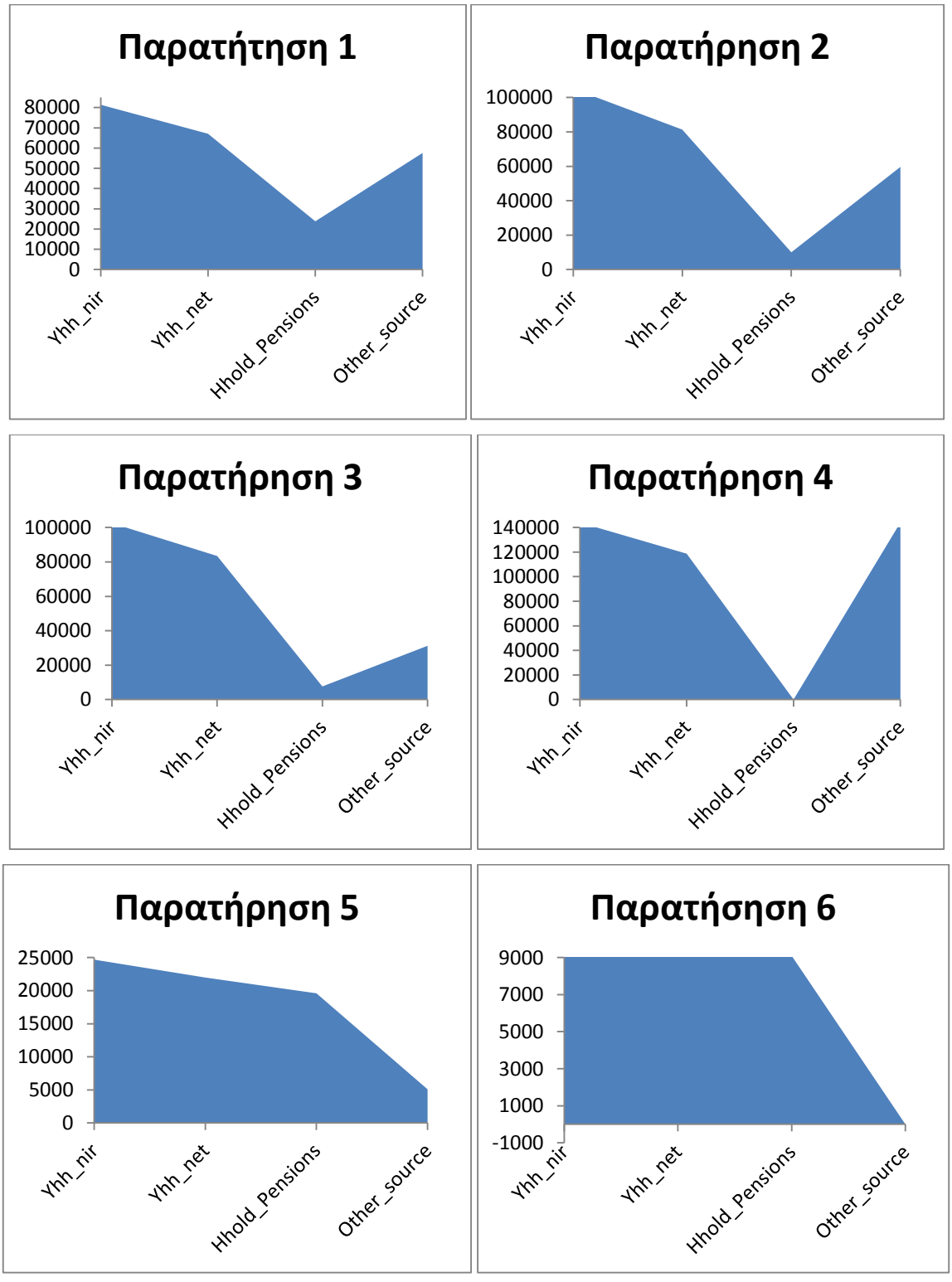
ισχυρή θετική συσχέτιση φαίνεται να παρουσιάζεται μεταξύ της μεταβλητής «οικογενειακό εισόδημα από άλλες πηγές» (Other_source) και των μεταβλητών Ακαθάριστο (Yhh_nir) και Καθαρό (Yhh_net) οικογενειακό εισόδημα. Αξίζει να σημειωθεί πως στις περιπτώσεις αυτές υπάρχουν δύο τιμές που είναι πιο απομακρυσμένες από την ευθεία. Σε αντίθετη περίπτωση, η μεταβλητή που δείχνει το οικογενειακό εισόδημα από σύνταξη (Hhold_Pensions) δείχνει να σχετίζεται αρνητικά με τις υπόλοιπες μεταβλητές καθώς σε κάθε περίπτωση, οι τιμές των μεταβλητών δείχνουν να μεταβάλλονται προς αντίθετη κατεύθυνση. Οι συσχετίσεις αυτές όμως δεν φαίνεται να είναι ισχυρές, καθώς αρκετές παρατηρήσεις αποκλίνουν από την ευθεία γραμμικής συσχέτισης.

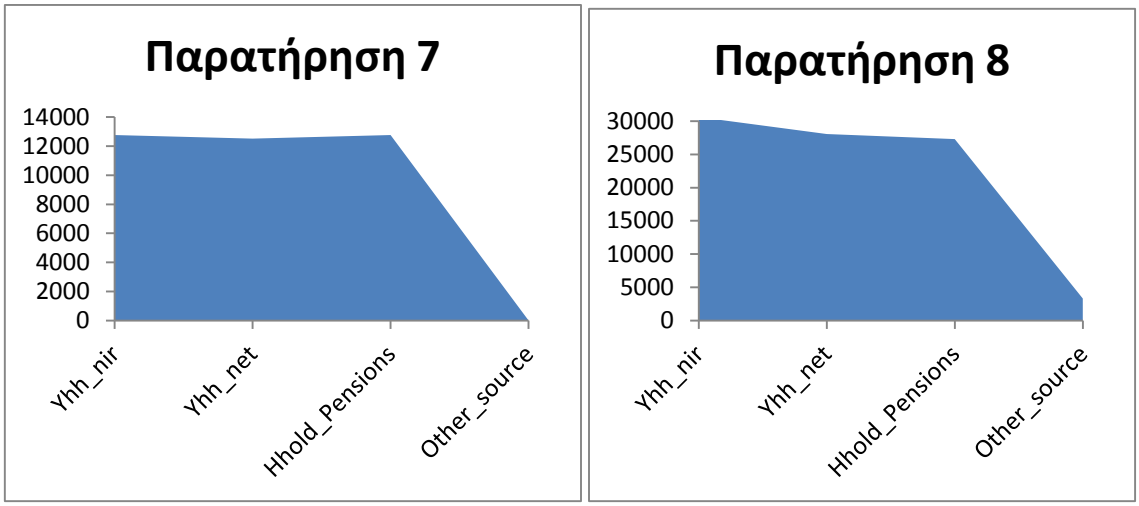
**Β) Διαγράμματα Profile**

Τα διαγράμματα Profile για πολυμεταβλητά δεδομένα βασίζονται στη λογική των πολλαπλών διαγραμμάτων διασποράς. Κάθε πολυδιάστατη παρατήρηση απεικονίζεται με ένα ιστόγραμμα (histogram) ή ένα γράφημα περιοχής (area plot). Οι κορυφές του κάθε γραφήματος δείχνουν την τιμή που παίρνει η παρατήρηση για την εκάστοτε μεταβλητή. Προφανώς για να συγκριθούν ταυτόχρονα πολλές παρατηρήσεις πρέπει να διατηρούν οι μεταβλητές μια σταθερή σειρά τοποθέτησής τους στο γράφημα. Αναπόφευκτα, εφόσον κάθε διάγραμμα αναλογεί σε μία παρατήρηση, δεν είναι εφικτή η χρήση των διαγραμμάτων αυτών για μεγάλο πλήθος παρατηρήσεων. Επιπλέον, ένας μεγάλος αριθμός μεταβλητών θα παραμόρφωνε το γράφημα δυσχεραίνοντας την ερμηνεία του. Κατ’ επέκταση, τα διαγράμματα Profile χρησιμοποιούνται κατά τη μελέτη μικρού όγκου δεδομένων με λίγες μεταβλητές. Η κατασκευή των διαγραμμάτων Profile, όπως και των περισσότερων τυπικών τεχνικών απεικόνισης δεδομένων, γίνεται εύκολα με τη βοήθεια απλών υπολογιστικών προγραμμάτων. Παρακάτω με τη βοήθεια του *Excel* κατασκευάστηκαν 8 διαγράμματα Profile που αφορούν στις παρατηρήσεις του Πίνακα 2.2.1. Για κάθε παρατήρηση καταγράφονται οι μετρήσεις των 4 μεταβλητών που αφορούν στα περιουσιακά στοιχεία κάθε νοικοκυριού.

**ΣΧΗΜΑ 2.3.2**

Διαγράμματα Profile






Όπως φαίνεται στο Σχήμα 2.3.2 κάθε παρατήρηση του δείγματος αντιστοιχεί σε ένα γράφημα Profile. Σε κάθε γράφημα χρωματίζεται με μπλε χρώμα η περιοχή που καλύπτουν οι μετρήσεις των τεσσάρων μεταβλητών. Μπορεί εύκολα κανείς να παρατηρήσει ότι οι 4 πρώτες παρατηρήσεις παρουσιάζουν παρόμοιες μετρήσεις στα 4 οικονομικά μεγέθη. Σύμφωνα με τα αντίστοιχα γραφήματα, οι τιμή του Ακαθάριστου οικογενειακού εισοδήματος ξεπερνάει τις 80.000€ για τις παρατηρήσεις αυτές και η τιμή του οικογενειακού εισοδήματος από σύνταξη είναι η ελάχιστη σε σχέση με τις υπόλοιπες μετρήσεις. Αντίθετα, οι 4 τελευταίες παρατηρήσεις συμφωνούν εξίσου μεταξύ τους. Στα τελευταία γραφήματα η μέγιστη τιμή του Ακαθάριστου οικογενειακού εισοδήματος είναι 30.000€ ενώ παράλληλα, η τιμή του οικογενειακού εισοδήματος από άλλες πηγές είναι η ελάχιστη σε σχέση με τις υπόλοιπες μετρήσεις.

### 2.3.2 Γεωμετρικές Τεχνικές

Στις Γεωμετρικές Τεχνικές γίνεται χρήση γεωμετρικών μετασχηματισμών των δεδομένων με σκοπό να εξετασθούν οι πιθανές σχέσεις και αλληλεπιδράσεις που υπάρχουν μεταξύ τους. Στην κατηγορία αυτή συγκαταλέγονται μερικές από τις πιο διαδεδομένες γραφικές μεθόδους ανάλυσης πολυμεταβλητών δεδομένων. Στο σημείο αυτό θα παρουσιαστούν δύο από τις πλέον διαδεδομένες τεχνικές αυτής της κατηγορίας.

#### Α) Καμπύλες Andrews

Μια από τις πλέον διαδεδομένες μεθόδους πολυμεταβλητών δεδομένων είναι οι καμπύλες Andrews οι οποίες πήραν το όνομά τους από τον κατασκευαστή τους Andrews (1972).

Σύμφωνα με τη μέθοδο αυτή, κάθε παρατήρηση που χαρακτηρίζεται από $p$ μεταβλητές, μπορεί να αναπαρασταθεί με τη μορφή μιας καμπύλης που προκύπτει από τη συνάρτηση:

$$f_x(t) = x_1/\sqrt{2} + x_2 \cdot \sin(t) + x_3 \cdot \cos(t) + x_4 \cdot \sin(2t) + x_5 \cdot \cos(2t) + \cdots$$

όπου $x_1$ ,$x_2$,…,$x_p$ είναι το σύνολο των μεταβλητών και $t \epsilon [-\pi, \pi]$.

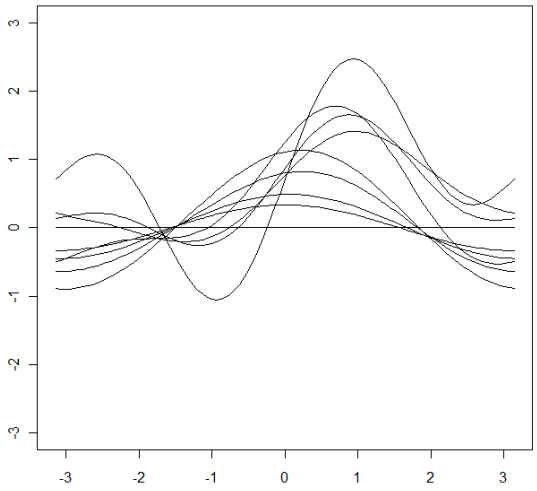
Επομένως, εφόσον κάθε πολυδιάστατη παρατήρηση αναπαριστάται από μία ξεχωριστή καμπύλη, είναι εύκολο παρατηρώντας τις καμπύλες καθώς και τις μεταξύ τους αποστάσεις, να προκύψουν συμπεράσματα σχετικά με τις ομοιότητες και τις διαφορές των παρατηρήσεων. Έτσι, καμπύλες με παρεμφερές σχήμα αναλογούν σε παρόμοιες τιμές των μεταβλητών στις αντίστοιχες παρατηρήσεις. Οι καμπύλες Andrews διακρίνονται από ιδιότητες που τις καθιστούν ως ένα ιδιαίτερα χρήσιμο εργαλείο για την εξαγωγή συμπερασμάτων. Δύο βασικές ιδιότητες είναι οι ακόλουθες:

- Η Ευκλείδεια απόσταση μεταξύ δύο σημείων, διατηρείται και μεταξύ των καμπυλών που αντιστοιχούν σε αυτά τα σημεία. Επομένως δύο σημεία που βρίσκονται κοντά στον $p$-διάστατο χώρο, θα έχουν καμπύλες που πλησιάζουν μεταξύ τους.
- Εάν οι μεταβλητές που μελετώνται είναι μεταξύ τους ασυσχέτιστες και έχουν σταθερή διακύμανση $\sigma^2$, τότε οι αντίστοιχες καμπύλες Andrews διατηρούν τη διακύμανση σταθερή.

Σύμφωνα με τα παραπάνω, τα δεδομένα που αντιστοιχούν σε καμπύλες με παρόμοια μορφή μπορούν να ομαδοποιηθούν διότι θεωρητικά πρόκειται για συσχετισμένες παρατηρήσεις. Αντίστοιχα, όσες καμπύλες διαφέρουν είναι ακραίες τιμές. Όμως η κατηγοριοποίηση αυτή των καμπυλών βασίζεται στην υποκειμενική κρίση του εκάστοτε ερευνητή και επομένως μπορεί να οδηγήσει σε αναληθή συμπεράσματα. Επιπλέον, η μορφή του γραφήματος είναι τέτοια που μπορεί να προκαλέσει σύγχυση όταν οι παρατηρήσεις είναι πολλές. Για τον λόγο αυτό είναι προτιμότερο το πλήθος των παρατηρήσεων που απεικονίζονται να μην ξεπερνάει τις 30. Στο σημείο αυτό κατασκευάστηκε με τη βοήθεια της γλώσσας *R*, ένα γράφημα καμπυλών Andrews που αφορά στα 4 χαρακτηριστικά του Πίνακα 2.2.1 για ένα πλήθος 8 παρατηρήσεων:

**ΣΧΗΜΑ 2.3.3**

Καμπύλες Andrews




Από το παραπάνω γράφημα είναι φανερό πως υπάρχουν 4 παρατηρήσεις που κυμαίνονται μεταξύ -1 και +1 και θα μπορούσαν να κατηγοριοποιηθούν σε μία ομάδα. Από την άλλη, μια δεύτερη ομάδα διαμορφώνεται από τις 3 παρατηρήσεις που κυμαίνονται μεταξύ -1 και 2. Τέλος, στην τελευταία ομάδα θα μπορούσε να ενταχθεί και η παρατήρηση που κυμαίνεται μεταξύ -2 και 3 αν και οι τιμές της είναι αρκετά μεγαλύτερες από τις τιμές των παρατηρήσεων της ομάδας αυτής.

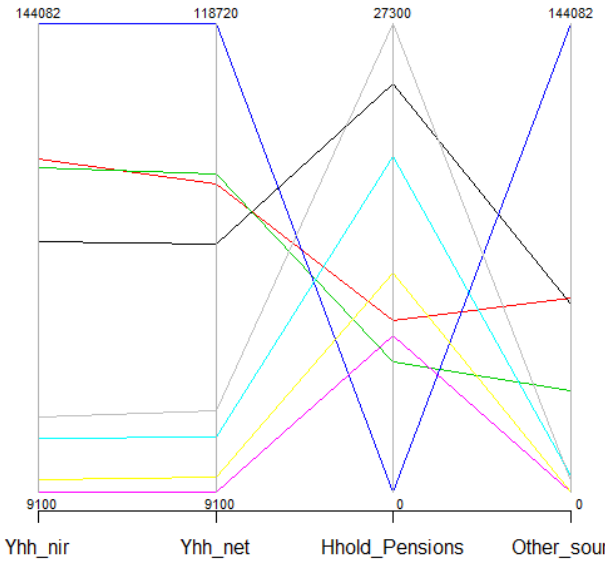
**Β) Διάγραμμα Παράλληλων Συντεταγμένων**

Το διάγραμμα παράλληλων συντεταγμένων είναι επίσης ένα εργαλείο οπτικοποίησης πολυδιάστατων δεδομένων. Αν και η μορφή του φαίνεται αρκετά περίπλοκη, στην πραγματικότητα πρόκειται για ένα γράφημα με αρκετά εύκολη κατασκευή που παρέχει σημαντικές πληροφορίες στον αναγνώστη. Κάθε μεταβλητή αναπαρίσταται με μια ευθεία παράλληλη προς τις υπόλοιπες. Οι παρατηρήσεις παίρνουν τιμές πάνω σε κάθε ευθεία ανάλογα με τις μετρήσεις τους για την εκάστοτε μεταβλητή. Στη συνέχεια, οι τιμές κάθε παρατήρησης συνδέονται με μια τεθλασμένη γραμμή. Με τον τρόπο αυτό κάθε πολυδιάστατη παρατήρηση δεν έχει το κόστος απώλειας πληροφορίας. Το διάγραμμα παράλληλων συντεταγμένων δίνει τη δυνατότητα να εντοπισθούν οι ακραίες τιμές καθώς και οι ομοιότητες μεταξύ των παρατηρήσεων. Επιπλέον, κατά τη σύγκριση των τιμών των παρατηρήσεων σε γειτονικές παράλληλες ευθείες, μπορεί να αναδείξει το βαθμό ανεξαρτησίας των μεταβλητών αυτών. Για τον λόγο αυτό, είναι προτιμότερο να εξετασθούν όλες οι πιθανές ταξινομήσεις των μεταβλητών ώστε να μπορούν να προκύψουν τα αντίστοιχα συμπεράσματα. Προφανώς

μια τέτοια διαδικασία είναι δύσκολη όταν υπάρχει μεγάλο πλήθος μεταβλητών, έτσι η χρήση του διαγράμματος προτείνεται κυρίως για μικρό αριθμό μεταβλητών. Επίσης, η μορφή του γραφήματος είναι τέτοια που δυσκολεύει την ύπαρξη πολλών παρατηρήσεων. Σε αυτή την περίπτωση, αφού εξετασθούν οι πιθανές σχέσεις που αναπτύσσονται μεταξύ των παρατηρήσεων, είναι δυνατόν να ομαδοποιηθούν από τον μελετητή πριν προχωρήσει στη κατασκευή των παράλληλων συντεταγμένων. Στο σημείο αυτό εφαρμόστηκε ένα διάγραμμα παράλληλων συντεταγμένων για τις 8 παρατηρήσεις και τις 4 μεταβλητές του Πίνακα 2.2.1. Το γράφημα κατασκευάστηκε με τη βοήθεια της γλώσσας *R*.

**ΣΧΗΜΑ 2.3.4**

Διάγραμμα Παράλληλων Συντεταγμένων




Από το Διάγραμμα Παράλληλων Συντεταγμένων, είναι φανερό πως οι τρεις πρώτες παρατηρήσεις οι οποίες απεικονίζονται με τις γραμμές χρώματος μπλε, κόκκινο και πράσινο, μπορούν να κατηγοριοποιηθούν στην ίδια ομάδα καθώς έχουν υψηλή τιμή στις μεταβλητές Yhh_nir και Yhh_net, χαμηλή τιμή στην μεταβλητή Hhold_Pensions και ξανά υψηλή τιμή στην μεταβλητή Other_source. Αντίθετα, οι παρατηρήσεις που απεικονίζονται με τα χρώματα γκρι, γαλάζιο, κίτρινο και ροζ έχουν χαμηλή τιμή στις μεταβλητές Yhh_nir και Yhh_net, υψηλή τιμή στην Hhold_Pensions και χαμηλή στην Other_source με αποτέλεσμα να κατηγοριοποιούνται στην ίδια ομάδα. Τέλος, η παρατήρηση με το μαύρο χρώμα δεν μπορεί να προσδιορισθεί πλήρως σε ποια ομάδα ανήκει καθώς παρουσιάζει σχετικά υψηλή τιμή στις μεταβλητές Yhh_nir και Yhh_net, υψηλή τιμή στην μεταβλητή Hhold_Pensions και σχετικά υψηλή τιμή στη μεταβλητή Other_source.

### 2.3.3 Εικονογραφικές Τεχνικές

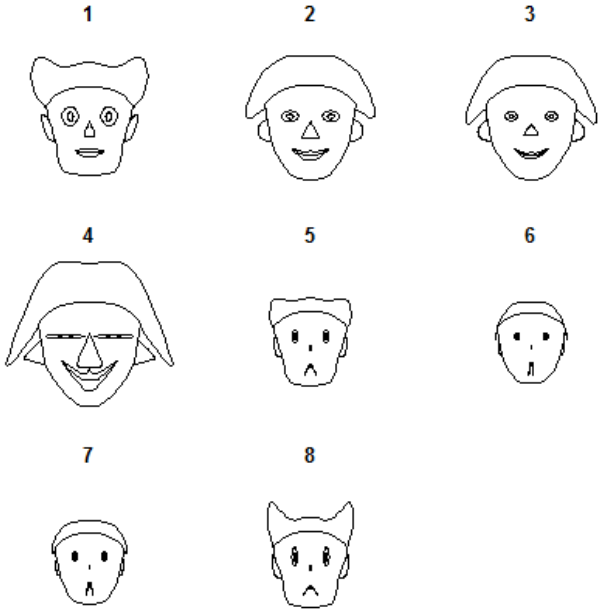
Στις Εικονογραφικές Τεχνικές γίνεται η χρήση εικονογραφημάτων για την απεικόνιση κάθε πολυμεταβλητής παρατήρησης. Το σχήμα, το μέγεθος, το χρώμα ή η θέση των εικονογραφημάτων δίνουν στοιχεία σχετικά με τις μετρήσεις των τιμών για κάθε μεταβλητή. Ο μελετητής είναι με αυτόν τον τρόπο σε θέση να συγκρίνει τις παρατηρήσεις μεταξύ τους. Παρακάτω αναλύονται οι πιο γνωστές εικονογραφικές τεχνικές.

**Α) Πρόσωπα του Chernoff**

Τα πρόσωπα του Chernoff είναι μια γραφική τεχνική απεικόνισης πολυμεταβλητών δεδομένων που εισήγαγε ο Herman Chernoff (1973). Η τεχνική αυτή βασίζεται στον ενδιαφέροντα τρόπο παρουσίασης των παρατηρήσεων με τη μορφή εικονογραφημάτων που απεικονίζουν πρόσωπα. Τα χαρακτηριστικά κάθε προσώπου αντιστοιχούν σε κάθε μεταβλητή. Επομένως το γράφημα αποτελείται από τόσα πρόσωπα όσα είναι οι προς μελέτη παρατηρήσεις και κάθε πρόσωπο έχει χαρακτηριστικά ανάλογα με το πλήθος των μεταβλητών. Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη καθώς βασίζεται στην ικανότητα του ανθρώπου να εντοπίζει εύκολα ομοιότητες και διαφορές μεταξύ διαφορετικών προσώπων. Ένα πρόβλημα της μεθόδου είναι πως περιορίζεται στα χαρακτηριστικά του προσώπου τα οποία όμως είναι σχετικά λίγα και κατ' επέκταση οι μεταβλητές που μπορούν να χρησιμοποιηθούν είναι λίγες. Για να αντιμετωπιστεί το πρόβλημα αυτό, υπάρχει η δυνατότητα να καταργηθεί η συμμετρία του προσώπου και επομένως να διπλασιαστούν οι μεταβλητές που μπορούν να χρησιμοποιούνται. Διαπιστώθηκε έτσι, ότι με την κατάλληλη τεχνική υποστήριξη, είναι δυνατόν να κατασκευαστούν πρόσωπα με τα οποία μελετώνται ταυτόχρονα έως και 18 μεταβλητές. Ένα επίσης μειονέκτημα της μεθόδου το οποίο αφορά στις περισσότερες γραφικές μεθόδους είναι η υποκειμενική αντίληψη των αποτελεσμάτων. Συγκεκριμένα, έχει αποδειχθεί ότι ο άνθρωπος έχει την τάση να παρατηρεί περισσότερο κάποια σημεία του προσώπου όπως τα μάτια και λιγότερο κάποια άλλα σημεία όπως τα φρύδια. Επομένως ανάλογα με το χαρακτηριστικό που αντιπροσωπεύει κάθε μεταβλητή, δίνεται και η αντίστοιχη προσοχή στην μεταβλητή αυτή. Στο σημείο αυτό με τη βοήθεια της γλώσσας *R*, κατασκευάστηκε ένα γράφημα με 8 πρόσωπα του Chernoff όσα δηλαδή τα νοικοκυριά μου μελετώνται στο παράδειγμά μας. Κάθε πρόσωπο έχει 4 χαρακτηριστικά (σχήμα κεφαλιού, μάτια, μύτη-στόμα και μαλλιά) που αντιστοιχούν στις 4 μεταβλητές που σχετίζονται με την οικονομική τους κατάσταση.

**ΣΧΗΜΑ 2.3.5**

Πρόσωπα του Chernoff




Τα πρόσωπα του Chernoff για το ενδεικτικό δείγμα που μελετάμε μας οδηγούν στο συμπέρασμα να θεωρήσουμε ότι έχουν παρόμοιες μετρήσεις οι 4 πρώτες παρατηρήσεις κατατάσσοντάς τις σε μία ομάδα και ομοίως οι 4 τελευταίες εντάσσονται σε μια δεύτερη ομάδα. Το συμπέρασμα αυτό βασίζεται στο γεγονός πως τα 4 πρώτα πρόσωπα έχουν σχετικά μεγαλύτερο σχήμα κεφαλιού, καπέλο με κλήση προς τα κάτω (με εξαίρεση την πρώτη παρατήρηση), μεγάλα μάτια με ξεκάθαρες κόρες, τριγωνική μύτη και στόμα με κλήση προς τα πάνω. Αντιθέτως, τα 4 τελευταία πρόσωπα έχουν μικρότερο σχήμα κεφαλιού, μικρά μάτια και μύτη και μικρό στόμα με κλήση προς τα κάτω.

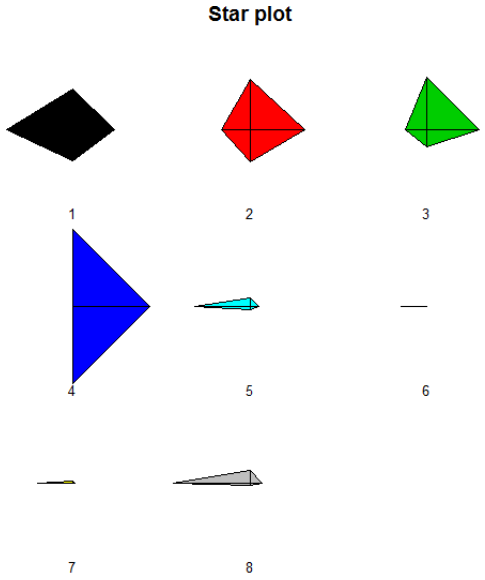
**Γ) Διάγραμμα Αστέρας**

Το διάγραμμα αστέρα συγκαταλέγεται σε μια ευρύτερη κατηγορία γραφικών μεθόδων απεικόνισης πολυδιάστατων δεδομένων, γνωστή ως «Διαγράμματα Ραντάρ» (Radar plots). Η τεχνική του διαγράμματος αυτού παρουσιάστηκε από τον Chambers (1983). Η τεχνική αυτή βασίζεται στην απεικόνιση των παρατηρήσεων με τη μορφή εικονογραφημάτων που απεικονίζουν αστέρια. Συγκεκριμένα, κάθε πολυμεταβλητή παρατήρηση έχει τη μορφή αστέρα ο οποίος αποτελείται από τόσες ακτίνες όσες και οι μεταβλητές που μελετώνται. Για κάθε παρατήρηση, το μήκος της ακτίνας διαφέρει ανάλογα με την τιμή που έχει δοθεί στην εκάστοτε μεταβλητή. Οι ακτίνες ισαπέχουν μεταξύ τους και τα άκρα τους ενώνονται

σχηματίζοντας ένα πολύγωνο. Προφανώς, η μορφή του διαγράμματος αυτού δεν βοηθάει στην εξαγωγή συμπερασμάτων όταν μελετώνται πολλές μεταβλητές. Αντίθετα, με ένα σχετικά μικρό πλήθος μεταβλητών είναι εύκολος ο εντοπισμός ομοιοτήτων και διαφορών μεταξύ των δεδομένων. Το διάγραμμα Αστέρα το συναντάμε συχνά και με την ονομασία «Διάγραμμα Ακτινών» (Sun Ray plot), «Διάγραμμα Αράχνη» (Spider chart), «Διάγραμμα Ιστός» (Web chart) κ.λ.π. Σε κάθε περίπτωση το γράφημα έχει την ίδια τεχνική κατασκευής με κάποιες πιθανές αλλαγές στη μορφοποίηση του. Στο σημείο αυτό, με τη βοήθεια της γλώσσας *R*, κατασκευάστηκε ένα γράφημα με 8 διαγράμματα Αστέρα όσες δηλαδή οι παρατηρήσεις μου μελετώνται στο παράδειγμά μας. Κάθε Αστέρι έχει 4 ακτίνες που αντιστοιχούν στις 4 μεταβλητές που σχετίζονται με την οικονομική τους κατάσταση.

**ΣΧΗΜΑ 2.3.6**

Διάγραμμα Αστέρας




Από το γράφημα Αστέρα είναι δυνατόν να καταταχθούν στην ίδια ομάδα οι τρεις πρώτες παρατηρήσεις του δείγματος που μελετάμε καθώς οι αντίστοιχοι αστέρες παρουσιάζουν μεγάλες ομοιότητες στην μορφή τους. Από την άλλη, οι παρατηρήσεις 5,6,7 και 8 μπορούν να ενταχθούν σε μια δεύτερη ομάδα καθώς οι αστέρες που τις απεικονίζουν είναι μικροί με έντονη κλήση προς τα αριστερά. Η παρατήρηση 4 δεν ταιριάζει απόλυτα σε καμία από τις δύο ομάδες και είναι στην ευχέρεια του εκάστοτε μελετητή να την εντάξει. Εξαιτίας του μεγέθους του αστέρα και της ελάχιστης τιμής που έχει η μέτρηση της μεταβλητής που αναλογεί στην αριστερή ακτίνα, προσωπικά θα επέλεγα να την εντάξω στην πρώτη ομάδα.

### 2.3.4 Άλλες γραφικές τεχνικές

Οι γραφικές μέθοδοι απεικόνισης πολυδιάστατων δεδομένων είναι πολλές και αυξάνονται διαρκώς όσο εξελίσσεται η επιστήμη της στατιστικής παράλληλα με την τεχνολογία. Επιπλέον, οι δυνατότητες που προσφέρουν έχουν τόσο μεγάλο εύρος που καθίσταται αναγκαία η ανάπτυξη και εφαρμογή οπτικών μεθόδων. Παραπάνω έγινε μια εκτενής αναφορά κάποιων γραφικών μεθόδων εφαρμόζοντας τις σε ένα δείγμα 8 νοικοκυριών κάθε ένα από τα οποία χαρακτηριζόταν από 4 μεταβλητές. Ολοκληρώνοντας την ενότητα αυτή θα ήταν καλό να γίνει μια ενδεικτική αναφορά και σε κάποιες επιπλέον γνωστές γραφικές μεθόδους, οι οποίες είναι αρκετά δημοφιλείς εξαιτίας της απλότητας στον τρόπο κατασκευής τους. Τα γραφήματα που ακολουθούν κατασκευάστηκαν με τα προγράμματα *Excel* αλλά είναι δυνατό να κατασκευαστούν με όλα τα στατιστικά πακέτα.

**Α) Διάγραμμα Φυσαλίδα**

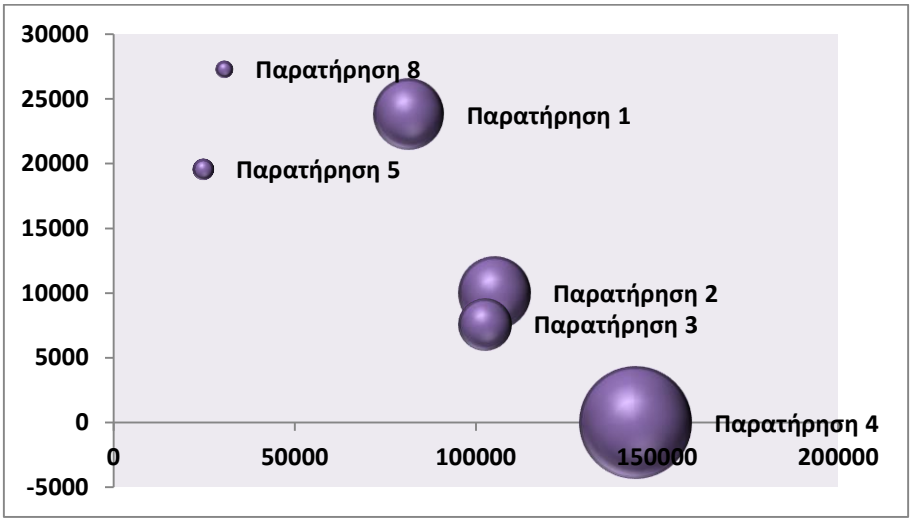
Το διάγραμμα φυσαλίδα, είναι ένα γράφημα 3 διαστάσεων κατασκευασμένο σε ένα σύστημα δύο αξόνων. Ο κάθε άξονας αντιπροσωπεύει μια μεταβλητή και κάθε παρατήρηση έχει τη μορφή μιας σφαίρας. Το μέγεθος της σφαίρας αντιστοιχεί στις μετρήσεις της παρατήρησης που αφορούν στην τελευταία μεταβλητή. Το διάγραμμα Φυσαλίδα που ακολουθεί αφορά στις μετρήσεις των 8 παρατηρήσεων του δείγματος για τις μεταβλητές Yhh_nir, Hhold_Pensions και Other_source οι οποίες, όπως έχει ειπωθεί, συμβολίζουν το Ακαθάριστο οικογενειακό εισόδημα, το οικογενειακό εισόδημα από σύνταξη και το οικογενειακό εισόδημα από άλλες πηγές αντίστοιχα.

Το γράφημα που ακολουθεί κατασκευάστηκε με τη χρήση δύο αξόνων, ο οριζόντιος αντιστοιχεί στην μεταβλητή Yhh_nir και ο κάθετος στην μεταβλητή Hhold_Pensions. Κάθε φυσαλίδα αναλογεί σε μία παρατήρηση το μέγεθος της οποίας οφείλεται στην τιμή που έχει για την μεταβλητή Other_source. Παρατηρώντας το Σχήμα 2.3.6, δεν είναι απολύτως ξεκάθαρη η ομαδοποίηση των παρατηρήσεων με βάση τις τρεις μεταβλητές. Οι παρατηρήσεις 5 και 8 δείχνουν να έχουν υψηλή τιμή στην μεταβλητή Hhold_Pensions (κάθετος άξονας) και χαμηλή τιμή τόσο για τη μεταβλητή Yhh_nir (οριζόντιος άξονας) όσο και για τη μεταβλητή Other_source (μέγεθος σφαίρας). Το γεγονός αυτό είναι αρκετό για να ενταχθούν οι παρατηρήσεις αυτές στην ίδια ομάδα. Από την άλλη, οι παρατηρήσεις 1,2 και 3 έχουν μέτρια τιμή στον οριζόντιο άξονα και μέτριο μέγεθος σφαίρας, ενώ διαφέρουν ως προς τον κάθετο

άξονα. Θα μπορούσαν έτσι να ενταχθούν σε μια δεύτερη ομάδα. Αξίζει επίσης να σημειωθεί ότι η παρατηρήσεις 6 και 7 αποκρύπτονται από τις παρατηρήσεις 2 και 3 γεγονός που καθιστά αδύνατο να βγάλουμε συμπεράσματα για το μέγεθός τους, ενώ τέλος, η παρατήρηση 4 δείχνει να διαφέρει έντονα από το σύνολο ως προς όλα τα οικονομικά χαρακτηριστικά που μελετάμε στο διάγραμμα αυτό.

**ΣΧΗΜΑ 2.3.7**

Διάγραμμα Φυσαλίδα



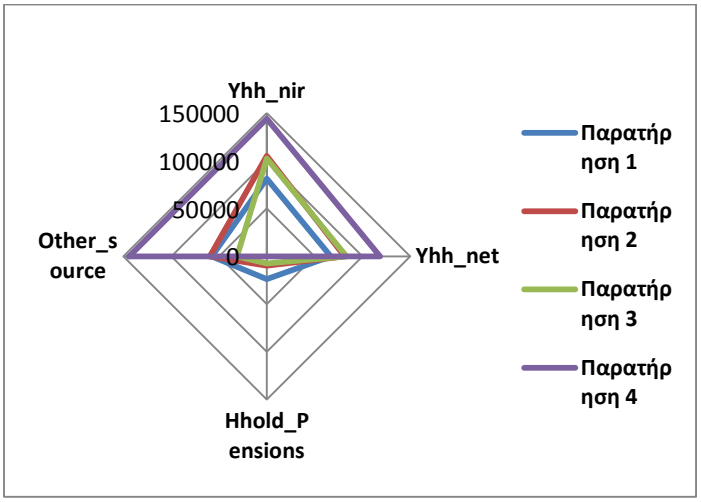

**Β) Διάγραμμα Αραχνοειδές**

Το Αραχνοειδές διάγραμμα βασίζεται στην λογική του διαγράμματος Αστέρα που αναφέρθηκε παραπάνω. Στο γράφημα αυτό, κάθε μεταβλητή αντιστοιχεί σε έναν άξονα-ακτίνα και κάθε παρατήρηση παίρνει τιμές πάνω σε αυτόν. Οι τιμές κάθε παρατήρησης συνδέονται μεταξύ τους με μια τεθλασμένη γραμμή. Οι τεθλασμένες γραμμές έχουν διαφορετικό χρώμα μεταξύ τους ώστε να ξεχωρίζουν οι παρατηρήσεις. Στο γράφημα που ακολουθεί κατασκευάσθηκαν δύο Αραχνοειδή διαγράμματα, εκ των οποίων το πρώτο συμπεριλαμβάνει τις μετρήσεις των τεσσάρων πρώτων παρατηρήσεων για τις οικονομικές μεταβλητές που μελετάμε και το δεύτερο συμπεριλαμβάνει τις μετρήσεις των υπόλοιπων παρατηρήσεων για τις ίδιες μεταβλητές.

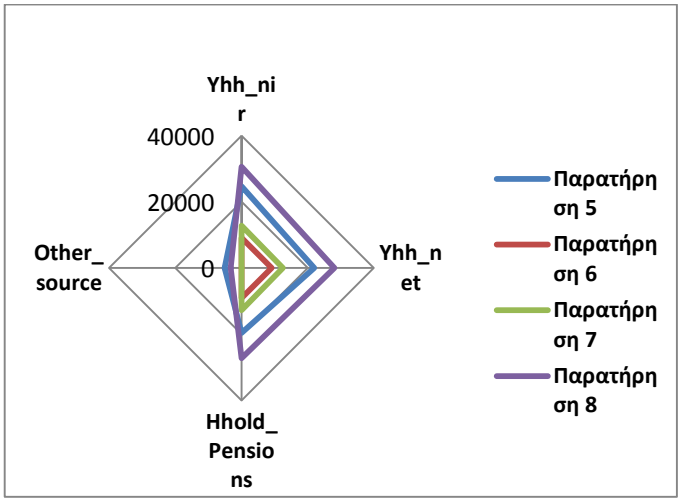
**ΣΧΗΜΑ 2.3.8**

Διάγραμμα Αραχνοειδές

**Παρατηρήσεις 1,2,3,4**




**Παρατηρήσεις 5,6,7,8**




Το κάθε γράφημα του Σχήματος 2.3.8 απεικονίζει τις παρατηρήσεις που με βάση προηγούμενα διαγράμματα, έχουν κατηγοριοποιηθεί στην ίδια ομάδα. Τα διαγράμματα αραχνοειδή επιβεβαιώνουν την ομαδοποίηση αυτή καθώς οι μετρήσεις των παρατηρήσεων σε κάθε ομάδα είναι παρεμφερείς μεταξύ τους. Πιο συγκεκριμένα, από το διάγραμμα φαίνεται ότι στην πρώτη ομάδα οι παρατηρήσεις εμφανίζουν υψηλές τιμές στις μεταβλητές Yhh_nir και Other_source και χαμηλές τιμές στην μεταβλητή Hhold_Pensions. Αντίθετα, η δεύτερη ομάδα διακρίνεται από υψηλές τιμές στις μεταβλητές Yhh_net και Hhold_Pensions και χαμηλές τιμές στην μεταβλητή Other_source.

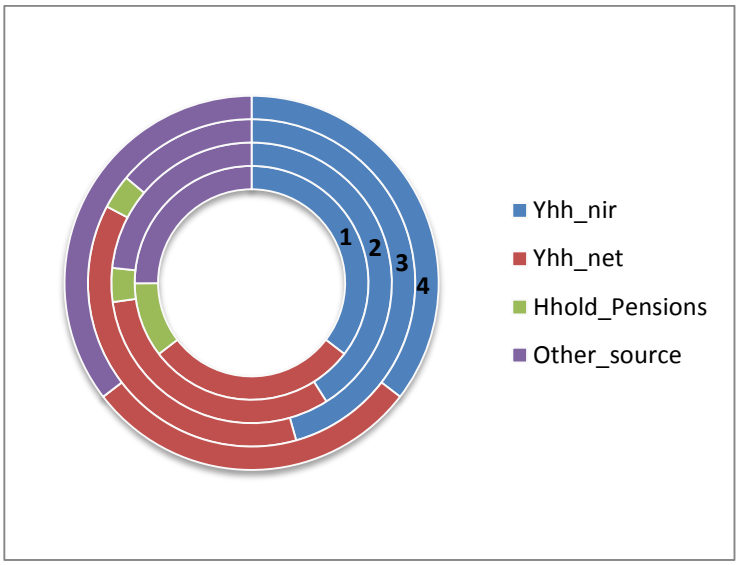
**Γ) Διάγραμμα Δακτύλιος**

Το διάγραμμα Δακτύλιος είναι ένα γράφημα αντίστοιχο του κυκλικού διαγράμματος. Κάθε παρατήρηση αντιστοιχεί σε έναν δακτύλιο και κάθε μεταβλητή καταλαμβάνει ένα μέρος του δακτυλίου αυτού ανάλογα με την τιμή που έχει. Επιπλέον, για να γίνεται ο διαχωρισμός των μεταβλητών αντιστοιχίζεται στην καθεμία ένα διαφορετικό χρώμα. Όπως έγινε και στην περίπτωση του αραχνοειδούς διαγράμματος, κατασκευάσθηκαν δύο διαγράμματα δακτύλιοι. Στο πρώτο απεικονίζονται οι παρατηρήσεις 1,2,3,4 και στο δεύτερο οι παρατηρήσεις 5,6,7,8. Οι παρατηρήσεις τοποθετούνται ξεκινώντας από τον εσωτερικό δακτύλιο του κάθε γραφήματος προς τον εξωτερικό. Κάθε χρώμα των δακτυλίων αντιστοιχίζεται σε μία από τις τέσσερις μεταβλητές της μελέτης.

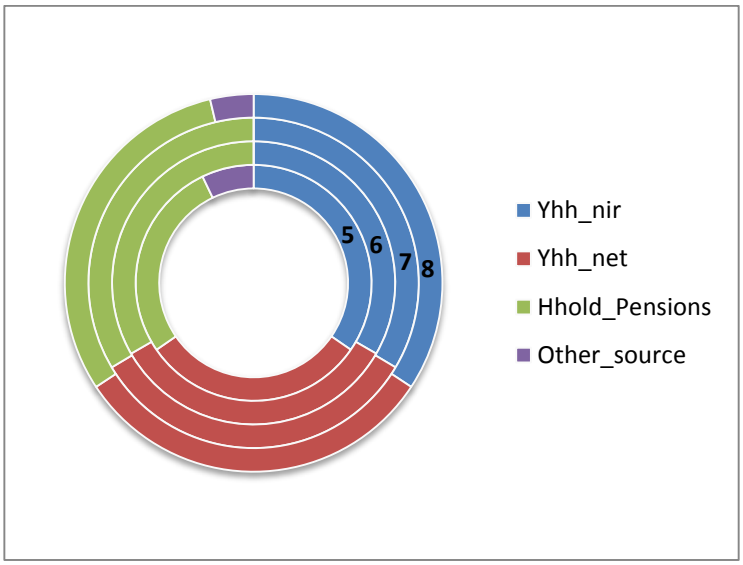
**ΣΧΗΜΑ 2.3.9**

Διάγραμμα Δακτύλιος

**Παρατηρήσεις 1,2,3,4**




**Παρατηρήσεις 5,6,7,8**




Όπως αναφέρθηκε και παραπάνω, κάθε δακτύλιος του Σχήματος 2.3.9 αντιστοιχεί σε μια παρατήρηση του δείγματος και κάθε χρώμα αντιστοιχεί σε μία μεταβλητή. Επομένως είναι φανερό πως οι παρατηρήσεις 1 έως 4 παρουσιάζουν πολύ χαμηλές τιμές στην μεταβλητή Hhold_Pensions (πράσινο χρώμα), υψηλές τιμές στην μεταβλητές Yhh_nir (μπλε χρώμα) και Yhh_net (κόκκινο χρώμα) και μέτρια τιμή στην μεταβλητή Other_source (μωβ χρώμα). Αντίθετα, οι παρατηρήσεις 5 έως 8 παρουσιάζουν σχετικά υψηλές τιμές για τις μεταβλητές Yhh_nir (μπλε χρώμα), Yhh_net (κόκκινο χρώμα) και Hhold_Pensions (πράσινο χρώμα) και ελάχιστες τιμές στην μεταβλητή Other_source (μωβ χρώμα).

# ΚΕΦΑΛΑΙΟ 3

# Μαθηματικές πολυμεταβλητές μέθοδοι

## 3.1 Εισαγωγή

Η Πολυμεταβλητή Ανάλυση ασχολείται με στατιστικές μεθόδους συλλογής, περιγραφής και ανάλυσης δεδομένων που αποτελούνται από μετρήσεις πολλών μεταβλητών σε ένα δείγμα παρατηρήσεων. Ένα βασικό χαρακτηριστικό που ξεχωρίζει την Πολυμεταβλητή Ανάλυση από την Μονομεταβλητή είναι πως η πρώτη αναζητά και αναλύει την εξάρτηση μεταξύ των μεταβλητών ή των ομάδων μεταβλητών. Για την επεξεργασία, την ανάλυση, την ομαδοποίηση και την κατάλληλη μορφοποίηση των πολυμεταβλητών δεδομένων υπάρχουν πολλές μαθηματικές πολυμεταβλητές μέθοδοι, πιο δημοφιλείς από τις οποίες είναι οι εξής:

1. Ανάλυση σε κύριες συνιστώσες (Principal Component Analysis)
2. Παραγοντική Ανάλυση (Factor Analysis)
3. Ανάλυση κατά συστάδες (Cluster Analysis)
4. Ανάλυση Αντιστοιχιών (Correspondence Analysis)

Οι μέθοδοι αυτές θα αναλυθούν εκτενέστερα στο παρόν κεφάλαιο.

## 3.2 Ανάλυση σε κύριες συνιστώσες

Η Ανάλυση σε κύριες συνιστώσες είναι μια στατιστική μέθοδος η οποία μετασχηματίζει γραμμικά ένα σύνολο συσχετισμένων μεταβλητών σε ένα σύνολο νέων ασυσχέτιστων μεταβλητών. Η μέθοδος αυτή εισήχθηκε το 1901 από τον Karl Pearson και αναπτύχθηκε περισσότερο το 1933 από τον Hotelling. Όπως αναφέρθηκε, οι αρχικές μεταβλητές οι οποίες στη συνέχεια θα μετασχηματιστούν σε γραμμικούς συνδυασμούς, είναι μεταξύ τους συσχετισμένες. Οι γραμμικοί συνδυασμοί που προκύπτουν, οι οποίοι είναι μεταξύ τους

ασυσχέτιστοι και διατηρούν το μεγαλύτερο μέρος της αρχικής πληροφορίας, ονομάζονται «Κύριες Συνιστώσες». Για να δημιουργηθούν επομένως οι κύριες συνιστώσες, είναι απαραίτητη η χρήση είτε του απλού πίνακα διακυμάνσεων-συνδιακυμάνσεων είτε του πίνακα διακυμάνσεων-συνδιακυμάνσεων των τυποποιημένων δεδομένων, δηλαδή του πίνακα συσχετίσεων.

Έστω λοιπόν πως διαθέτουμε ένα σύνολο $p$ μεταβλητών $(X_1, X_2, \dots, X_p)$ οι οποίες προσδιορίζουν τα χαρακτηριστικά $n$ ατόμων όπως φαίνεται στον πίνακα $X=(x_{ij})$:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Οι κύριες συνιστώσες $(Y_1, Y_2, \dots, Y_p)$ θα είναι της μορφής:

$$Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p$$
$$Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p$$
$$\vdots$$
$$Y_n = \alpha_{p1}X_1 + \alpha_{p2}X_2 + \cdots + \alpha_{pp}X_p$$

και αποτελούν γραμμικούς συνδυασμούς των αρχικών.

Οι τελευταίες ισότητες θα μπορούσαν να γραφτούν στη μορφή $\mathbf{Y} = A\mathbf{X}$ όπου $A$ είναι ο πίνακας:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \dots & \boldsymbol{\alpha}_p \end{bmatrix}$$

Και $\mathbf{X} = (X_1, X_2, \dots, X_p)'$, $\mathbf{Y} = \left(Y_1, Y_2, \dots, Y_p\right)'$.

Συνεπώς το πρόβλημα εύρεσης των κύριων συνιστωσών περιορίζεται στην εύρεση των στοιχείων του πίνακα $A$. Οι περιορισμοί που θέτονται για την επίλυση του προβλήματος αυτού είναι:

- Το διάνυσμα $\boldsymbol{\alpha}$ να έχει μέτρο ίσο με τη μονάδα, δηλαδή, $\|\boldsymbol{\alpha}\| = 1 \Leftrightarrow \sum \alpha_i^2 = 1$
- Οι κύριες συνιστώσες πρέπει να είναι σε φθίνουσα σειρά ως προς τη διακύμανσή τους, δηλαδή η πρώτη να έχει τη μεγαλύτερη διακύμανση, η δεύτερη τη δεύτερη μεγαλύτερη κ.ο.κ.

Αν συμβολίσουμε με $\Sigma$ τον πίνακα διακυμάνσεων-συνδιακυμάνσεων του τυχαίου διανύσματος $\mathbf{X}$ , τότε η διακύμανση της πρώτης συνιστώσας $Y_1$ προκύπτει από τον τύπο:

$$Var(Y_1) = \boldsymbol{\alpha}_1{}'\Sigma\boldsymbol{\alpha}_1$$

Η διακύμανση αυτή θα λέγεται «διασπορά του συνόλου $N$ κατά μήκος του διανύσματος $\boldsymbol{\alpha}$» και θα συμβολίζεται $Dis_{\boldsymbol{\alpha}}(N) = Dis(Y)$. Επομένως, με βάση τους δύο περιορισμούς, για να βρούμε το $\boldsymbol{\alpha}_1$, θα πρέπει να μεγιστοποιήσουμε την $Dis(Y)$ λαμβάνοντας υπόψη τον περιορισμό $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1$=1. Η λύση του προβλήματος αυτού γίνεται με χρήση των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα $\Sigma$.

Πιο συγκεκριμένα έχουμε τα εξής αποτελέσματα: Ο πίνακας $\Sigma$ έχει μη αρνητικές ιδιοτιμές, έστω $\lambda_1, \lambda_2, \dots, \lambda_p$. Ας θεωρήσουμε ότι $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ και ας συμβολίσουμε $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ τα αντίστοιχα μοναδιαία ιδιοδιανύσματα. Τότε:

- Το διάνυσμα $\boldsymbol{\alpha}$ που μεγιστοποιεί την $Dis(Y)$ είναι το μοναδιαίο διάνυσμα $\mathbf{u}_1$ που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή $\lambda_1$
- Η μέγιστη τιμή της τετραγωνικής μορφής $Dis(Y)$ είναι ίση με $\lambda_1$, δηλαδή ισχύει ότι: $max_{\|\mathbf{a}\|=1} Dis_{\boldsymbol{\alpha}}(N) = Dis_{\mathbf{u}_i}(N) = \lambda_1$.

Η μεταβλητή $Y$ για την οποία επιτυγχάνεται η προαναφερθείσα μεγιστοποίηση λέγεται πρώτη κύρια συνιστώσα (first principal component). Με παρόμοιο τρόπο μπορούμε να δούμε πως για όλες τις κύριες συνιστώσες τα διανύσματα $\boldsymbol{\alpha}_j$ αντιστοιχούν στα ιδιοδιανύσματα της $j$ σε φθίνουσα σειρά ιδιοτιμής. Φυσικά, για την εύρεση των υπολοίπων κύριων συνιστωσών χρειάζεται να προσθέσουμε τον περιορισμό ότι οι κύριες συνιστώσες είναι ασυσχέτιστες με τις προηγούμενες τους και κατ' επέκταση ο πίνακας διακυμάνσεων-συνδιακυμάνσεων τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές $\lambda_j$. Οι κύριες συνιστώσες, θα έχουν συνολική διακύμανση ίδια με τη συνολική διακύμανση των αρχικών μεταβλητών (ιδιότητα ίχνους συμμετρικού και τετραγωνικού πίνακα), δηλαδή αν συμβολίσουμε τη συνολική διακύμανση ως $Dis(N)$, θα ισχύει ότι:

$$Dis(N) = tr(\Sigma) = \sum_{j=1}^{p} \lambda_j.$$

Τέλος, για να υπολογιστεί πόσο συμβάλει η $j$ συνιστώσα στη συνολική διακύμανση, δηλαδή τι ποσοστό της συνολικής διακύμανσης ερμηνεύει η $j$ συνιστώσα, χρησιμοποιείται ο τύπος:

$$\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$$

Ολοκληρώνοντας τη μελέτη της ανάλυσης των κύριων συνιστωσών, θα μπορούσαμε να συνοψίσουμε τη διαδικασία στα παρακάτω τρία βήματα:

1. Εύρεση ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα $\Sigma$.

2. Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμά της αντιστοιχεί στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή αντιστοιχεί στη δεύτερη κύρια συνιστώσα κ.ο.κ.
3. Η διακύμανση της κάθε κύριας συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί.

Ένα από τα μειονεκτήματα της ανάλυσης σε κύριες συνιστώσες, χρησιμοποιώντας τον πίνακα διακύμανσης $\Sigma$, είναι πως εάν αλλάξει η κλίμακα μέτρησης των δεδομένων, τότε αλλάζουν και οι κύριες συνιστώσες και η ερμηνεία τους. Παράλληλα, εάν μια μεταβλητή έχει πολύ μεγαλύτερη διακύμανση από τις υπόλοιπες, αυτή τείνει να ταυτίζεται με την πρώτη κύρια συνιστώσα. Για να αντιμετωπιστεί αυτό το πρόβλημα, όταν θέλουμε να δημιουργήσουμε ένα γραμμικό συνδυασμό που δεν θα επηρεάζεται από τις μονάδες μέτρησης των μεταβλητών $X_1, X_2, \dots, X_p$ θα πρέπει να προβούμε σε κανονικοποίηση της κάθε μεταβλητής. Με άλλα λόγια, στην περίπτωση αυτή, ενδείκνυται η χρήση του πίνακα συσχετίσεων έναντι του πίνακα διακυμάνσεων-συνδιακυμάνσεων. Με τον τρόπο αυτό, οι συσχετίσεις δεν αλλάζουν όταν αλλάζει η κλίμακα και ταυτόχρονα δίνεται ίδιο βάρος σε όλες τις μεταβλητές εφόσον όλα τα στοιχεία της διαγωνίου είναι ίσα με τη μονάδα. Από την άλλη, υπάρχουν περιπτώσεις που κάποιες μεταβλητές πρέπει να θεωρηθούν πως έχουν μεγαλύτερο βάρος. Κατά συνέπεια, συχνά δεν είναι ξεκάθαρο ποιον πίνακα πρέπει να επιλέγουμε. Μια καλή στρατηγική είναι να αποφεύγουμε τον πίνακα διακυμάνσεων-συνδιακυμάνσεων όταν υπάρχουν κάποιες μεταβλητές με πολύ μεγαλύτερη διακύμανση.

### 3.2.1 Επιλογή πλήθους κυρίων συνιστωσών

Η επιλογή του πλήθους των κύριων συνιστωσών είναι το σημαντικότερο κομμάτι της ανάλυσης καθώς ο βέλτιστος αριθμός συνιστωσών πρέπει να είναι αρκετά μικρός για λόγους οικονομίας και ευκολίας, και αρκετά μεγάλος ώστε να διατηρηθεί όσο το δυνατόν περισσότερη πληροφορία. Στη βιβλιογραφία έχουν καταγραφεί αρκετά κριτήρια που βοηθούν στην επιλογή αυτή. Κάποια από αυτά αναφέρονται στη συνέχεια:

- <u>Ποσοστό συνολικής διακύμανσης που εξηγούν οι συνιστώσες:</u>

Σύμφωνα με το κριτήριο αυτό, το βέλτιστο πλήθος συνιστωσών είναι αυτό σύμφωνα με το οποίο αθροιστικά εξηγούν ένα επιθυμητό στόχο (π.χ. 80%). Ο στόχος αυτός επιλέγεται με υποκειμενικά κριτήρια.

- Κριτήριο Kaiser:

Σύμφωνα με το κριτήριο αυτό, επιλέγονται τόσες ιδιοτιμές όσες είναι μεγαλύτερες από τη μέση τιμή των ιδιοτιμών, δηλαδή μεγαλύτερες από $\bar{\lambda}=\frac{\sum_{j=1}^{p}\lambda_j}{p}$, όπου $\lambda_j$ είναι οι ιδιοτιμές.

- Ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται:

Σύμφωνα με το κριτήριο αυτό, επιλέγεται κάθε συνιστώσα που ερμηνεύει ένα συγκεκριμένο στόχο. Και σε αυτή την περίπτωση ο στόχος επιλέγεται με υποκειμενικά κριτήρια.

- Scree plot:

Το Scree plot είναι μια γραφική μέθοδος σύμφωνα με την οποία δημιουργείται ένα σχήμα όπου στον οριζόντιο άξονα καταγράφεται η σειρά και στον κατακόρυφο άξονα οι ιδιοτιμές. Το κριτήριο αυτό προτείνει να πάρουμε τόσες συνιστώσες μέχρι το γράφημα να αρχίσει να γίνεται περίπου επίπεδο, στην ουσία μέχρι να διαπιστώσουμε ότι αρχίζει να αλλάζει η κλίση.

- Μέθοδος Velicer:

Η μέθοδος στηρίζεται στους συντελεστές μερικής συσχέτισης ανάμεσα στις αρχικές μεταβλητές, όταν παραλείψουμε κάποιες συνιστώσες. Με τη μέθοδο αυτή αρχίζουμε να διώχνουμε τις συνιστώσες, μέχρι να φτάσουμε στο σημείο που δεν πρέπει να διώξουμε άλλη.

Ολοκληρώνοντας την περιγραφή της μεθόδου της Ανάλυσης σε κύριες συνιστώσες, αξίζει να επισημανθεί πως η μέθοδος αυτή είναι ιδιαίτερα σημαντική καθώς όχι μόνο μειώνει τις αρχικές διαστάσεις του προβλήματος, αλλά από ένα σύνολο συσχετισμένων μεταβλητών καταλήγει σε ένα σύνολο ασυσχέτιστων γεγονός αρκετά χρήσιμο στην περεταίρω στατιστική ανάλυση.

## 3.3 Παραγοντική Ανάλυση

Η Παραγοντική Ανάλυση είναι μια μέθοδος η οποία έχει σκοπό να εντοπίσει την ύπαρξη κοινών παραγόντων ανάμεσα σε μια ομάδα μεταβλητών. Η τεχνική αυτή της ανάλυσης χρησιμοποιήθηκε για πρώτη φορά από τον Charles Spearman το 1904, ο οποίος έκανε την υπόθεση πως οι συσχετίσεις των υπό μελέτη μεταβλητών οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων που είναι άγνωστοι στον μελετητή, αλλά μπορούν να εκτιμηθούν. Απώτερος σκοπός της μεθόδου είναι να εντοπίσει μέσα σε ομάδες που περιλαμβάνουν υψηλά συσχετισμένες μεταβλητές, τον παράγοντα εκείνο που είναι υπεύθυνος

για τις παρατηρούμενες συσχετίσεις. Έτσι εάν θεωρήσουμε πως οι συσχετίσεις μεταξύ των μεταβλητών οφείλονται αποκλειστικά στην ύπαρξη κάποιων κοινών παραγόντων $F_1, F_2, \dots, F_k$ τότε οι $p$ μεταβλητές που διαθέτουμε γράφονται εύκολα ως γραμμικός συνδυασμός των $k$ παραγόντων χρησιμοποιώντας το ορθογώνιο μοντέλο που είναι το πιο διαδεδομένο μοντέλο της παραγοντικής ανάλυσης. Δηλαδή εισάγοντας τους συμβολισμούς:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}, L = \begin{bmatrix} L_{11} & \cdots & L_{1k} \\ \vdots & \ddots & \vdots \\ L_{p1} & \cdots & L_{pk} \end{bmatrix}, \mathbf{F} = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

όπου

**X** είναι το διάνυσμα των αρχικών μεταβλητών

**μ** είναι το διάνυσμα των μέσων

$L$ είναι ένας πίνακας $p \times k$ όπου το $L_{ij}$ είναι η επιβάρυνση του παράγοντα $F_j$ στη μεταβλητή $X_i$

**F** είναι το διάνυσμα με τους παράγοντες

**ε** είναι το σφάλμα, δηλαδή το μέρος της μεταβλητής το οποίο δεν μπορεί να εξηγηθεί από τους παράγοντες.

Τότε:

$$\mathbf{X} - \boldsymbol{\mu} = L\mathbf{F} + \boldsymbol{\varepsilon}.$$

Εάν υποθέσουμε πως όλες οι μεταβλητές έχουν μέσο 0 και γνωρίζοντας από τον ορισμό της μεθόδου πως το πλήθος $k$ των παραγόντων είναι μικρότερο από το πλήθος $p$ των μεταβλητών, η σχέση που αναφέραμε παίρνει τη μορφή:

$$X_1 = L_{11}F_1 + L_{12}F_2 + \cdots + L_{1k}F_k + \varepsilon_1$$
$$X_2 = L_{21}F_1 + L_{22}F_2 + \cdots + L_{2k}F_k + \varepsilon_2$$
$$\dots$$
$$X_p = L_{p1}F_1 + L_{p2}F_2 + \cdots + L_{pk}F_k + \varepsilon_p$$

Προφανώς το μοντέλο που αναπτύχθηκε παραπάνω διαφέρει από το γραμμικό μοντέλο παλινδρόμησης καθώς αφενός τα $X_i$ δεν είναι παρατηρήσεις αλλά μεταβλητές και αφετέρου το δεξί μέλος της εξίσωσης δεν είναι παρατηρήσιμο και πρέπει να εκτιμηθεί. Όμως είναι φανερό πως με τη χρήση του μοντέλου, το **X** εξαρτάται γραμμικά από μια σειρά τυχαίων

μεταβλητών $F_i$ οι οποίες δεν είναι δυνατόν να παρατηρηθούν, αλλά και από $p$ πρόσθετες πηγές μεταβλητότητας $\varepsilon_i$ που ονομάζονται σφάλματα.

Για να εφαρμοσθεί το παραγοντικό μοντέλο $\mathbf{X} - \boldsymbol{\mu} = L\mathbf{F} + \boldsymbol{\varepsilon}$ , θα πρέπει να ισχύουν ορισμένες υποθέσεις:

- $E(\mathbf{F}) = 0$
- $Cov(\mathbf{F}) = I$
- $E(\boldsymbol{\varepsilon}) = 0$
- $Cov(\boldsymbol{\varepsilon}) = \Psi$, όπου $\Psi$ είναι ο πίνακας:

$$\Psi = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \psi_p \end{bmatrix}$$

- $Cov(\boldsymbol{\varepsilon}, \mathbf{F}) = 0$

Σύμφωνα λοιπόν με τις παραπάνω υποθέσεις, είναι φανερό πως οι παράγοντες που προκύπτουν είναι μεταξύ τους ασυσχέτιστοι. Αξίζει να σημειωθεί πως η δεύτερη προϋπόθεση του μοντέλου είναι αυτή που οδήγησε στον χαρακτηρισμό του ως «Ορθογώνιο Μοντέλο».

Από τις παραπάνω υποθέσεις μπορεί εύκολα να δειχθεί ότι ο πίνακας διακύμανσης $\Sigma$ είναι:

$$\Sigma = Cov(\mathbf{X}) = Cov(L\mathbf{F} + \boldsymbol{\varepsilon}) = LCov(\mathbf{F})L' + Cov(\boldsymbol{\varepsilon}) = LL' + \Psi$$

και συνεπώς μπορεί να διασπαστεί σε δύο μέρη, στο κομμάτι που ερμηνεύουν οι κοινοί παράγοντες και ονομάζεται «εταιρικότητα» (communality) και στο κομμάτι που οφείλεται στα σφάλματα και ονομάζεται «ιδιαιτερότητα» (specificity). Κατά συνέπεια, το ενδιαφέρον της παραγοντικής ανάλυσης περιορίζεται στην εκτίμηση των πινάκων $L$ και $\Psi$ ώστε να μπορεί να αναπαρασταθεί ο πίνακας διακύμανσης $\Sigma$. Τα βήματα που θα ακολουθηθούν είναι:

- Έλεγχος ύπαρξης συσχετίσεων μεταξύ των αρχικών μεταβλητών, καθώς όπως έχει αναφερθεί παραπάνω, οι αρχικές μεταβλητές πρέπει να είναι ισχυρά συσχετισμένες.
- Εύρεση του αριθμού παραγόντων και εκτίμηση των παραμέτρων του μοντέλου.
- Περιστροφή του μοντέλου εάν χρειάζεται.

Η Ανάλυση Παραγόντων υποθέτει ότι υπάρχουν κάποιοι κοινοί παράγοντες που εξηγούν τα δεδομένα. Επομένως, οι μεταβλητές πρέπει να έχουν μεγάλες συσχετίσεις αφού αυτές τις συσχετίσεις προσπαθεί να εξηγήσει η μέθοδος. Εάν υπάρχουν κάποιες μεταβλητές που είναι

ασυσχέτιστες με τις υπόλοιπες θα πρέπει να αγνοηθούν από τον μελετητή. Μια συσχέτιση μεγαλύτερη από 40% κατ' απόλυτη τιμή, είναι ικανοποιητική.

Ένα μέτρο για να συγκρίνουμε το σχετικό μέγεθος των συντελεστών συσχέτισης είναι το *Kaiser-Meyer-Olkin* στατιστικό που υπολογίζεται ως:

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j\neq i}\rho_{ij}^2}{\sum_{i=1}^{p}\sum_{j\neq i}\rho_{ij}^2 + \sum_{i=1}^{p}\sum_{j\neq i}a_{ij}^2}$$

όπου $\rho_{ij}$ και $\alpha_{ij}$ είναι οι δειγματικοί συντελεστές συσχέτισης και μερικής συσχέτισης αντίστοιχα μεταξύ των μεταβλητών $X_i$ και $X_j$. Αν η τιμή *KMO* είναι μεγαλύτερη από 0,5 τότε τα δεδομένα είναι κατάλληλα για παραγοντική ανάλυση.

Ένα άλλο μέτρο που επιτρέπει να εξετάσουμε μία-μία τις μεταβλητές και την καταλληλότητα τους για περαιτέρω επεξεργασία με Παραγοντική Ανάλυση είναι το *μέτρο δειγματικής καταλληλότητας* (measure of sampling adequacy) το οποίο υπολογίζεται για την *i* μεταβλητή:

$$MSA_i = \frac{\sum_{j=1}^{p}\rho_{ij}^2}{\sum_{j=1}^{p}\rho_{ij}^2+\sum_{j=1}^{p}a_{ij}^2}, i = 1,2,\dots,p$$

Αν το $MSA_i$ πάρει τιμές κοντά στο 1 τότε υπάρχουν ενδείξεις καταλληλότητας της *i* μεταβλητής για Παραγοντική Ανάλυση.

Αφού έχει ολοκληρωθεί η διαδικασία κατά την οποία ελέγχθηκε η ύπαρξη συσχετίσεων μεταξύ των αρχικών μεταβλητών, στη συνέχεια πρέπει να καθοριστεί ο κατάλληλος αριθμός παραγόντων και να εκτιμηθούν οι παράμετροι του μοντέλου. Δηλαδή θέλουμε να εκτιμήσουμε τα στοιχεία του πίνακα επιβαρύνσεων *L* και τα στοιχεία της διαγωνίου του πίνακα *Ψ*. Για να βρεθεί ο αριθμός των παραγόντων, ο ερευνητής θα μπορούσε να χρησιμοποιήσει παρόμοιες τεχνικές με αυτές που χρησιμοποιούνται στην ανάλυση κύριων συνιστωσών. Εφόσον όμως, η επιλογή των παραγόντων προηγείται της εκτίμησης των παραμέτρων, θα μπορούσε κανείς να δουλέψει διαδοχικά με αυξανόμενο πλήθος παραγόντων και να κρατήσει το μοντέλο με βάση κάποιο κριτήριο «καλής προσαρμογής». Τέτοιου είδους κριτήρια είναι (Κούτρας, (2013)):

- Από τον πίνακα επιβαρύνσεων μπορεί κάποιος να εκτιμήσει τον πίνακα *Σ*. Οι αποκλίσεις του πραγματικού πίνακα με τον εκτιμημένο (συνήθως ονομάζεται

*reproduced matrix*) θα πρέπει να είναι μικρές. Δυστυχώς δεν υπάρχει ένα κριτήριο του πόσο μικρές.

- Έλεγχος λόγου πιθανοφανειών αν οι εκτιμήσεις έχουν γίνει με τη μέθοδο μεγίστης πιθανοφάνειας. Τέτοιοι έλεγχοι στηρίζονται σε υποθέσεις για την κατανομή του πληθυσμού.

Οι δύο βασικές μέθοδοι εκτίμησης των παραμέτρων είναι η μέθοδος των κύριων συνιστωσών και η μέθοδος της μέγιστης πιθανοφάνειας.

**i. Εκτίμηση με τη μέθοδο των κύριων συνιστωσών**

Μια πολύ βασική μέθοδος στην εκτίμηση των παραμέτρων είναι η μέθοδος των κύριων συνιστωσών. Όπως έχει αναφερθεί, σκοπός μας είναι να βρεθούν εκτιμημένοι πίνακες $\hat{L}$ και $\hat{\Psi}$ για τους οποίους ο πίνακας $\hat{L}\hat{L}' + \hat{\Psi}$ είναι όσο γίνεται πιο κοντά στον πίνακα διακύμανσης $\Sigma$. Η εκτίμηση με τη μέθοδο αυτή βασίζεται στη φασματική ανάλυση του πίνακα $\Sigma$. Επομένως, με βάση τις ιδιότητες της φασματικής ανάλυσης αν χρησιμοποιήσουμε ως εκτίμηση του πίνακα $L$ τον τύπο $\hat{L} = \Pi\Lambda^{1/2}$, όπου $\Lambda$ είναι ο διαγώνιος πίνακας που περιέχει στη διαγώνιο τις ιδιοτιμές και $\Pi$ είναι ο πίνακας με στήλες τα ιδιοδιανύσματα του πίνακα $\Sigma$, τότε μπορούμε να αναπαραστήσουμε πλήρως τον πίνακα $\Sigma$, αφού $\hat{L}\hat{L}' = \Pi\Lambda^{\frac{1}{2}}\left(\Lambda^{\frac{1}{2}}\Pi\right)' = \Sigma$. Στην πράξη δουλεύουμε με τον δειγματικό πίνακα διακύμανσης $S$. Στη συνέχεια προκύπτουν δύο περιπτώσεις σχετικές με το πλήθος $k$ των παραγόντων:

- Αν το πλήθος των παραγόντων είναι ίσο με το πλήθος των μεταβλητών, δηλαδή $k = p$, επιτυγχάνεται η πλήρης αναπαράσταση του πίνακα $\Sigma$ και επομένως οι εκτιμήσεις των ιδιαιτεροτήτων, δηλαδή ο πίνακας $\widehat{\Psi}$ είναι 0. Στην περίπτωση αυτή η διακύμανση εξηγείται εξ ολοκλήρου από τους παράγοντες.
- Αν το πλήθος των παραγόντων είναι μικρότερο από το πλήθος των μεταβλητών, δηλαδή $k < p$, τότε δεν επιτυγχάνεται πλήρης αναπαράσταση του πίνακα $\Sigma$. Στην περίπτωση αυτή οι εκτιμήσεις των ιδιαιτεροτήτων προκύπτουν από τον τύπο:

$$\widehat{\psi_i} = s_i^2 - \sum_{j=1}^{p} \widehat{L_{ij}^2}$$

όπου $\widehat{L_{ij}}$ είναι το ij-στοιχείο του πίνακα $\hat{L}$, δηλαδή η επιβάρυνση του $j$ παράγοντα στην $i$ μεταβλητή.

Ολοκληρώνοντας τη μέθοδο αυτή, αξίζει να σημειωθεί πως δεν υπάρχει περιορισμός στον αριθμό των παραγόντων που μπορούμε να εκτιμήσουμε, δεδομένου ότι το πλήθος τους θα είναι μικρότερο ή ίσο από το πλήθος των μεταβλητών. Επιπλέον, δεν πρέπει να συγχέουμε την πολυμεταβλητή μέθοδο ανάλυσης σε κύριες συνιστώσες με την μέθοδο κύριων συνιστωσών που χρησιμοποιούμε για να εκτιμήσουμε το παραγοντικό μοντέλο. Η πρώτη είναι μια αυτοτελής μέθοδος ανάλυσης, ενώ η δεύτερη είναι ένα εργαλείο εκτίμησης παραμέτρων.

**ii. Εκτίμηση με τη μέθοδο μέγιστης πιθανοφάνειας**

Μια δεύτερη πολύ βασική μέθοδος στην εκτίμηση των παραμέτρων είναι η μέθοδος μέγιστης πιθανοφάνειας. Για την εκτίμηση των παραμέτρων με τη μέθοδο αυτή χρειάζεται να υποθέσουμε, αφενός ότι τα σφάλματα ακολουθούν πολυμεταβλητή κανονική κατανομή, με διάνυσμα μέσων το μηδενικό διάνυσμα και πίνακα διακύμανσης τον πίνακα $\Psi$, δηλαδή $\varepsilon \sim N_p(0, \Psi)$ και αφετέρου ότι το διάνυσμα των τυχαίων μεταβλητών $X$, δοθέντος του διανύσματος των παραγόντων $F$, ακολουθεί πολυδιάστατη κανονική κατανομή. Έτσι αν υποθέσουμε ότι και οι παράγοντες προέρχονται από πολυδιάστατη κανονική κατανομή, προκύπτει πως $X \sim N_p(LF, LL' + \Psi)$. Παράλληλα, με τη μέθοδο μέγιστης πιθανοφάνειας υπάρχει περιορισμός στο πλήθος των παραγόντων που μπορούμε να εκτιμήσουμε και πιο συγκεκριμένα, το μέγιστο πλήθος αυτών είναι ίσο με το ακέραιο μέρος του αριθμού $\frac{p}{2}$, όπου $p$ είναι ο αριθμός των μεταβλητών. Για να μεγιστοποιηθεί η πιθανοφάνεια, έχοντας αρκετούς περιορισμούς, χρειάζονται κατάλληλες αριθμητικές μέθοδοι καθώς και ένα κριτήριο τερματισμού αυτών των μεθόδων. Όμως μας δίνεται η δυνατότητα στη μέθοδο αυτή, να δουλέψουμε είτε με τον πίνακα διακύμανσης είτε με τον πίνακα συσχετίσεων, καθώς η λύση δεν επηρεάζεται από τις μονάδες μέτρησης.

Ολοκληρώνοντας τις μεθόδους εκτίμησης παραμέτρων, αξίζει να σημειωθεί πως υπάρχουν και άλλες ανάλογες μέθοδοι όπως:

**iii. Η μέθοδος ελαχίστων τετραγώνων:**

Προσπαθεί να ελαχιστοποιήσει το άθροισμα των τετραγωνικών διαφορών των πραγματικών συνδιακυμάνσεων με αυτές που το μοντέλο εκτιμά.

**iv. Η γενικευμένη μέθοδος ελαχίστων τετραγώνων:**

Αποτελεί παραλλαγή της προηγούμενης και χρησιμοποιεί ως βάρη τις αντίστροφες τιμές των μοναδικών διακυμάνσεων.

**v. Η μέθοδος των κυρίων αξόνων:**

Αποτελεί παραλλαγή της μεθόδου των κυρίων συνιστωσών και αντικαθιστά τις μονάδες στη διαγώνιο του πίνακα συσχέτισης με εκτιμήσεις εταιρικότητας.

Η παραπάνω εκτίμηση των παραμέτρων έγινε με σκοπό να αναπαρασταθεί ο πίνακας $\Sigma$ σε μορφή κατάλληλη για την ανάλυση παραγόντων. Το ερώτημα που προκύπτει είναι εάν οι κατάλληλοι πίνακες $L$ και $\Psi$ που ικανοποιούν τη σχέση $\Sigma = LL' + \Psi$, είναι μοναδικοί. Στην πραγματικότητα, το πρόβλημα της ανάλυσης παραγόντων επιδέχεται πολλές λύσεις και ο πίνακας των φορτίων $L$ δεν είναι μοναδικός. Προκειμένου λοιπόν να δημιουργηθούν παράγοντες που επιδέχονται καλύτερη ερμηνεία, θα μπορούσαμε να πολλαπλασιάσουμε τον πίνακα των φορτίων με κάποιο ορθογώνιο πίνακα. Η μέθοδος αυτή κατά την οποία προσπαθούμε να κάνουμε τους παράγοντες πιο ερμηνεύσιμους ονομάζεται «περιστροφή». Με την περιστροφή δεν αλλάζουν τα χαρακτηριστικά του μοντέλου όπως η καλή προσαρμοστικότητα του, ούτε ο πίνακας $\Psi$. Το μόνο που αλλάζει είναι οι τιμές των επιβαρύνσεων. Επομένως, με τον τρόπο αυτό, επιδιώκουμε οι επιβαρύνσεις κάποιων παραγόντων να είναι μεγάλες σε απόλυτη κλίμακα για μερικές μεταβλητές έτσι ώστε να γίνεται ξεκάθαρο ποιες μεταβλητές εξαρτώνται από αυτούς τους παράγοντες. Οι βασικές μέθοδοι περιστροφής είναι:

- Varimax: Ελαχιστοποιεί τον αριθμό των μεταβλητών που έχουν μεγάλες επιβαρύνσεις για όλους τους παράγοντες.
- Quartimax: Ελαχιστοποιεί τον αριθμό των παραγόντων που εξηγούν μια μεταβλητή.
- Equimax: Συνδυάζει τις δύο παραπάνω μεθόδους.
- Oblique: Μη ορθογώνια περιστροφή κι έτσι οι παράγοντες που προκύπτουν δεν είναι ασυσχέτιστοι.

## 3.4 Ανάλυση κατά συστάδες

Η ανάλυση κατά συστάδες είναι μια μέθοδος που χρησιμοποιείται για την τμηματοποίηση των δεδομένων και την δημιουργία ομάδων χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Εξετάζει έτσι το βαθμό ομοιότητας των

παρατηρήσεων ώστε αυτές στη συνέχεια να καταταχθούν στις εκάστοτε ομάδες. Συνεπώς οι βασικές έννοιες που πρέπει να χρησιμοποιηθούν για την ανάλυση κατά συστάδες είναι η έννοια της *απόστασης* και η έννοια της *ομοιότητας*. Παρατηρήσεις που παρέχουν παρόμοιες πληροφορίες έχουν μικρή απόσταση μεταξύ τους ενώ οι παρατηρήσεις που διαφέρουν έχουν μεγάλη απόσταση. Είναι έτσι φανερό πως οι έννοιες απόσταση και ομοιότητα είναι αντίθετες μεταξύ τους. Προφανώς, η μέθοδος είναι αποτελεσματική όταν οι παρατηρήσεις που βρίσκονται στην ίδια ομάδα έχουν μικρές αποστάσεις, ενώ παρατηρήσεις που βρίσκονται σε διαφορετικές ομάδες διακρίνονται από μεγάλες αποστάσεις. Από την άλλη, ίδιες παρατηρήσεις έχουν μεγάλο βαθμό ομοιότητας και διαφορετικές παρατηρήσεις έχουν μικρό βαθμό ομοιότητας. Πέρα όμως από την ομαδοποίηση των δεδομένων, με την Ανάλυση κατά συστάδες παρέχει κι άλλες δυνατότητες:

- ✓ Βοηθά στην διερεύνηση των σχέσεων μεταξύ των δεδομένων δημιουργώντας μια πιο σαφή εικόνα για αυτά.
- ✓ Μειώνει τις διαστάσεις του προβλήματος, καθώς με την ομαδοποίηση των δεδομένων ο μελετητής επικεντρώνεται στις μεταβλητές που παρουσιάζουν το μεγαλύτερο ενδιαφέρον.
- ✓ Επιτυγχάνει την περαιτέρω στατιστική ανάλυση των δεδομένων όπως την πρόβλεψη νέων τιμών και τον έλεγχο υποθέσεων.

Υπάρχουν πολλές διαφορετικές προσεγγίσεις για την ομαδοποίηση των παρατηρήσεων, οι πιο διαδεδομένες εκ των οποίων είναι οι ακόλουθες:

- Ιεραρχικές μέθοδοι: Οι ομάδες σχηματίζονται σταδιακά είτε με συνένωση των παρατηρήσεων σε μικρές ομάδες, έπειτα σε μεγαλύτερες και καταλήγοντας σε μία μεγάλη ομάδα, είτε με τη διαίρεση μιας ομάδας σε μικρότερες καταλήγοντας σε τόσες ομάδες όσες και το πλήθος των παρατηρήσεων.
- Μη ιεραρχικές μέθοδοι: Ο αριθμός των ομάδων είναι προκαθορισμένος. Με τη χρήση ενός αλγορίθμου μοιράζονται οι παρατηρήσεις στις εκάστοτε ομάδες.

Στο σημείο αυτό κρίνεται σκόπιμο να γίνει εκτενέστερη αναφορά στις δύο έννοιες που είναι απαραίτητες στην ανάλυση κατά συστάδες, την απόσταση και την ομοιότητα.

### 3.4.1 Απόσταση και Ομοιότητα

Η απόσταση είναι ένα βασικό μέτρο στην πολυμεταβλητή ανάλυση. Σκοπός της είναι να μετρήσει πόσο απέχουν δύο παρατηρήσεις και κατ' επέκταση να δείξει τον βαθμό ομοιότητας

τους. Έτσι, ένα μέτρο απόστασης μετασχηματίζεται εύκολα σε μέτρο ομοιότητας και αντίστροφα.

Έστω ότι διαθέτουμε δύο παρατηρήσεις και για κάθε παρατήρηση δίνονται οι μετρήσεις τους για δύο συγκεκριμένες μεταβλητές. Ας συμβολίσουμε τις παρατηρήσεις μας ως $\mathbf{x}=(x_1, x_2)'$ και $\mathbf{y}=(y_1, y_2)'$. Μια συνάρτηση $f(\mathbf{x}, \mathbf{y})$ είναι η απόσταση των παρατηρήσεων εάν ισχύουν οι παρακάτω ιδιότητες (Καρλής, (2005)):

1. $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$ (συμμετρική ιδιότητα)
2. $f(\mathbf{x}, \mathbf{y}) \leq f(\mathbf{x}, \mathbf{z}) + f(\mathbf{z}, \mathbf{y})$ (τριγωνική ιδιότητα)
3. $f(\mathbf{x}, \mathbf{y}) \neq 0 \Leftrightarrow \mathbf{x} \neq \mathbf{y}$
4. $f(\mathbf{x}, \mathbf{x}) = 0$

Η πιο γνωστή προσέγγιση για τον υπολογισμό μιας απόστασης *d* μεταξύ δύο παρατηρήσεων *x* και *y* είναι η «Ευκλείδεια απόσταση» η οποία προκύπτει από τον τύπο:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$

Η Ευκλείδεια απόσταση αν και έχει αρκετά εύκολο υπολογισμό, ωστόσο δεν είναι επαρκής καθώς επηρεάζεται από την αλλαγή κλίμακας των μεταβλητών. Είναι όμως γνωστό, πως για να φέρει κανείς κάθε μεταβλητή σε συγκρίσιμη κλίμακα μπορεί να τις «τυποποιήσει», δηλαδή να διαιρέσει κάθε μεταβλητή με την τυπική της απόκλιση. Έτσι, συμβολίζοντας με $s_i$ τη διακύμανση της *i* μεταβλητής, η απόσταση δύο παρατηρήσεων προκύπτει πλέον από τον τύπο:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\left(\frac{x_1 - y_1}{s_1}\right)^2 + \left(\frac{x_2 - y_2}{s_2}\right)^2}$$

Η απόσταση αυτή είναι πιο ενδιαφέρουσα και δεν επηρεάζεται από τη διαφορά κλίμακας. Όμως, φαίνεται να μην λαμβάνει υπόψη τις συνδιακυμάνσεις μεταξύ των μεταβλητών γεγονός που προκαλεί προβλήματα κυρίως όταν οι μεταβλητές είναι πολύ συσχετισμένες. Ένας τύπος που λαμβάνει υπόψη τις συνδιακυμάνσεις των μεταβλητών είναι η απόσταση Mahalanobis που υπολογίζεται ως εξής:

$$d^2(x, y) = (x - y)' S^{-1} (x - y)$$

όπου *S* είναι ο δειγματικός πίνακας διακυμάνσεων και *x, y* είναι διανύσματα.

Είναι φανερό πως μπορεί κανείς εύκολα να κατασκευάσει πολλά μέτρα απόστασης, αρκεί να ικανοποιούνται οι προϋποθέσεις που αναφέρθηκαν παραπάνω γι' αυτό και στην βιβλιογραφία υπάρχει μια μεγάλη ποικιλία τέτοιων μέτρων.

Ένα άλλο είδος μέτρων που χρησιμοποιούνται για να δείξουν εάν μοιάζουν ή διαφέρουν δύο παρατηρήσεις είναι τα μέτρα ομοιότητας. Στην περίπτωση αυτή, τα δεδομένα αφορούν μια σειρά μεταβλητών για τις οποίες παίρνουν την τιμή 1 εάν έχουν το χαρακτηριστικό και 0 εάν δεν το έχουν. Έπειτα κατασκευάζεται ο πίνακας ομοιότητας. Όπως έχει ειπωθεί και παραπάνω, παρατηρήσεις που μοιάζουν πολύ μεταξύ τους έχουν μεγάλη τιμή στο μέτρο ομοιότητας ενώ εάν διαφέρουν έχουν μικρή τιμή. Στη συνέχεια μπορούν να ομαδοποιηθούν τα δεδομένα με βάση το μέτρο αυτό. Όμοιες παρατηρήσεις μπαίνουν στην ίδια ομάδα ενώ ανόμοιες παρατηρήσεις βρίσκονται σε διαφορετική ομάδα. Όπως γίνεται και με τα μέτρα απόστασης, έτσι και η κατασκευή των μέτρων ομοιότητας βασίζεται σε κάποιες προϋποθέσεις. Έτσι, εάν υποθέσουμε ότι για κάθε ζεύγος παρατηρήσεων $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ και $\mathbf{x}_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ ορίζεται ένας πραγματικός αριθμός $s_{ij} = s(x_i, x_j)$ έτσι ώστε να ισχύουν οι επόμενες τρεις ιδιότητες:

1. $s_{ij} \geq 0$ για κάθε $i,j$ και $i=j \Rightarrow s_{ij} = 1$
2. $s_{ij} \leq 1$
3. $s_{ij} = s_{ji}$ (συμμετρική ιδιότητα)

τότε λέμε ότι η συνάρτηση $s_{ij} = s(\mathbf{x}_i, \mathbf{x}_j)$ δίνει ένα μέτρο ομοιότητας.

Στη συνέχεια κατασκευάζεται ο ακόλουθος πίνακας συνάφειας 2×2 όπου σε κάθε κελί καταγράφεται το πλήθος των συνδυασμών (1,1),(1,0),(0,1),(0,0) για κάθε συνδυασμό παρατηρήσεων $i,j$. Δηλαδή:

| | | Παρατήρηση $j$ | | |
|---|---|---|---|---|
| | | 1 | 0 | |
| Παρατήρηση $i$ | 1 | $\boldsymbol{a}$ | $\boldsymbol{b}$ | $a+b$ |
| | 0 | $\boldsymbol{c}$ | $\boldsymbol{d}$ | $c+d$ |
| | | $a+c$ | $b+d$ | $p$ |

Το κελί $a$=(1,1) υποδηλώνει πως τα χαρακτηριστικά είναι παρόντα και στις δύο παρατηρήσεις, το κελί $b$=(1,0) υποδηλώνει πως η παρατήρηση $i$ έχει τα συγκεκριμένα χαρακτηριστικά ενώ η παρατήρηση $j$ δεν τα έχει κ.ο.κ. Προφανώς το σύνολο των συμφωνιών για τις δύο παρατηρήσεις βρίσκεται με το άθροισμα $a+d$ ενώ $p = a + b + c + d$ είναι το σύνολο όλων των συμφωνιών και ασυμφωνιών των παρατηρήσεων.

Όπως και στην περίπτωση των αποστάσεων είναι εύκολο να κατασκευαστούν πολλά μέτρα ομοιότητας, αρκεί να πληρούνται οι προϋποθέσεις που αναφέρθηκαν παραπάνω. Κάποια γνωστά μέτρα είναι τα ακόλουθα:

- Simple matching coefficient (Sokal and Michener, 1958): Πρόκειται για το σύνολο των συμφωνιών προς το συνολικό πλήθος συμφωνιών και ασυμφωνιών

$$s_{ij} = \frac{a + d}{a + b + c + d}$$

- Rogers and Tarimoto (1960): Πρόκειται για το σύνολο των συμφωνιών προς το συνολικό πλήθος συμφωνιών και ασυμφωνιών, δίνοντας διπλάσιο βάρος στις ασυμφωνίες

$$s_{ij} = \frac{a + d}{(a + d) + 2(b + c)}$$

- Jaccard coefficient (1908): Πρόκειται για το σύνολο των θετικών συμφωνιών προς των σύνολο των θετικών συμφωνιών και ασυμφωνιών

$$s_{ij} = \frac{a}{a + b + c}$$

Εφόσον έχουμε αναφέρει πως οι έννοιες της ομοιότητας και της απόστασης είναι αντίθετες, μπορούμε να αντιληφθούμε πως ένα μέτρο ομοιότητας μπορεί πολύ εύκολα να μετατραπεί σε μέτρο ανομοιότητας δηλαδή σε μέτρο απόστασης. Έτσι για παράδειγμα, το μέτρο απόστασης δίτιμων μεταβλητών που βασίζεται στο μέτρο ομοιότητας «Simple matching» είναι:

$$d_{ij} = \frac{b + c}{a + b + c + d}$$

Ομοίως, το μέτρο απόστασης δίτιμων μεταβλητών που βασίζεται στο μέτρο ομοιότητας των «Rogers και Tarimoto» είναι:

$$d_{ij} = \frac{2(b + c)}{(a + d) + 2(b + c)}$$

Γενικά ισχύει πως στην περίπτωση που μελετώνται δίτιμες μεταβλητές για τις οποίες έχει ορισθεί ένα μέτρο απόστασης $d_{ij}$ τότε είναι δυνατόν να δημιουργηθεί το αντίστοιχο μέτρο ομοιότητας μέσω του τύπου:

$$s_{ij} = \frac{1}{1 + d_{ij}}$$

Ομοίως, εάν έχει ορισθεί ένα μέτρο ομοιότητας $s_{ij}$, τότε είναι δυνατόν να δημιουργηθεί το αντίστοιχο μέτρο απόστασης μέσω του τύπου:

$$d_{ij} = \sqrt{2(1 - s_{ij})}$$

### 3.4.2 Μέθοδοι Ομαδοποίησης

Όπως έχει ήδη αναφερθεί η Ανάλυση κατά Συστάδες έχει σκοπό να διαχωρίσει τα δεδομένα σε ομάδες ανάλογα με το βαθμό ομοιότητας των παρατηρήσεων. Η διαμόρφωση των ομάδων μπορεί να γίνει είτε προκαθορίζοντας το πλήθος τους, είτε με τη σταδιακή τους εξέλιξη. Με βάση το παραπάνω κριτήριο οι μέθοδοι ομαδοποίησης χωρίζονται σε «Ιεραρχικές» και «Μη Ιεραρχικές».

**Α) Ιεραρχική Ομαδοποίηση**

Στις Ιεραρχικές μεθόδους οι ομάδες δεν έχουν καθοριστεί εξ αρχής και σχηματίζονται σταδιακά. Οι παρατηρήσεις είτε είναι αρχικά χωρισμένες σε τόσες ομάδες όσα και τα δεδομένα και σταδιακά συνενώνονται μέχρι να καταλήξουν σε μία ομάδα, είτε αρχικά είναι ενωμένες σε μία ομάδα και σταδιακά διαχωρίζονται μέχρι να καταλήξουν σε τόσες ομάδες όσα και τα δεδομένα. Στην πρώτη περίπτωση η κατηγορία της ιεραρχικής ομαδοποίησης λέγεται «Συσσωρευτική Μέθοδος» (Agglomerative method) ενώ στην δεύτερη περίπτωση μιλάμε για την «Διαιρετική Μέθοδο» (Divisive Method).

**i. Συσσωρευτική Μέθοδος**

Η διαδικασία ξεκινάει θεωρώντας πως κάθε παρατήρηση αποτελεί μία ομάδα. Επομένως εάν διαθέτουμε $n$ παρατηρήσεις, έχουμε εξ αρχής $n$ ομάδες. Στη συνέχεια, υπολογίζονται όλες οι αποστάσεις $d_{ij}$ μεταξύ των παρατηρήσεων και καταγράφονται σε έναν $n \times n$ πίνακα αποστάσεων $D$. Έπειτα ενώνονται οι παρατηρήσεις με τη μικρότερη απόσταση δημιουργώντας μια ομάδα. Η διαδικασία αυτή συνεχίζεται ενώνοντας σταδιακά τις κοντινότερες ομάδες. Ο αλγόριθμος ολοκληρώνεται όταν όλες οι παρατηρήσεις έχουν μπει σε μία ομάδα. Αξίζει να σημειωθεί πως εναλλακτικά είναι δυνατόν να υπολογισθούν στην αρχή τα μέτρα ομοιότητας $s_{ij}$ μεταξύ των παρατηρήσεων αντί οι αποστάσεις τους και στη συνέχεια να συνενώνονται οι παρατηρήσεις που διακρίνονται από τον μεγαλύτερο βαθμό ομοιότητας.
Αν και έχουν οριστεί ήδη μέτρα αποστάσεων μεταξύ των στοιχείων, δεν έχουν οριστεί μέτρα αποστάσεων μεταξύ των ομάδων. Υπάρχουν πολλές μέθοδοι, μερικές από τις οποίες είναι (Καρλής, (2005)):

- Η μέθοδος της απλής συνένωσης (Single Linkage Method/ Nearest Neighbor):

Στην μέθοδο αυτή υπολογίζεται η απόσταση μεταξύ των ομάδων ως η μικρότερη απόσταση από μια παρατήρηση στην μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Η μέθοδος έχει κάποιες χρήσιμες μαθηματικές ιδιότητες, αλλά παράγει ομάδες που δεν είναι συμπαγείς και συνήθως δημιουργεί μερικές πολύ μεγάλες ομάδες και κάποιες πολύ μικρές.

- Η μέθοδος της πλήρους συνένωσης (Complete Linkage Method/ Furthest Neighbor):

Στην περίπτωση αυτή υπολογίζεται η απόσταση μεταξύ των ομάδων ως η μεγαλύτερη απόσταση από μια παρατήρηση μέσα στη μια ομάδα με κάποια παρατήρηση στην άλλη ομάδα. Οι ομάδες αυτή τη φορά είναι συμπαγείς αλλά αποτυγχάνει να ξεχωρίσει κάποιες πολύ συμπαγείς μικρές ομάδες.

- Η μέθοδος των σταθμισμένων μέσων (Weighted Average Linkage Method/ Average between groups):

Σε αυτή την περίπτωση η απόσταση είναι ο μέσος της απόστασης ανάμεσα σε όλες τις αποστάσεις της μιας ομάδας με τα στοιχεία της άλλης. Δηλαδή εάν η μία ομάδα περιλαμβάνει τις παρατηρήσεις {1,2} και η άλλη τις παρατηρήσεις {3,4,5}, τότε η απόσταση μεταξύ των δύο ομάδων βρίσκεται ως ο μέσος των αποστάσεων $d(1,3)$, $d(1,4)$, $d(1,5)$, $d(2,3)$, $d(2,4)$, $d(2,5)$.

- Μέθοδος των κέντρων βάρους (Centroid Method):

Η απόσταση υπολογίζεται ως η απόσταση των κέντρων των ομάδων. Πρόκειται για μια πολύ καλή μέθοδο καθώς παράγει συνήθως ομάδες συμπαγείς και ελλειπτικές. Όμως εφαρμόζεται μόνο όταν τα δεδομένα είναι ποσοτικά.

### ii. Διαιρετική Μέθοδος

Οι διαιρετικές μέθοδοι είναι επίσης μια κατηγορία Ιεραρχικής Ομαδοποίησης αν και λιγότερο διαδεδομένες. Όπως έχει ήδη αναφερθεί, η λογική τους είναι η αντίστροφη ακριβώς διαδικασία από τις συσσωρευτικές. Επομένως, ξεκινούν με μία ομάδα που περιέχει και τις $n$ παρατηρήσεις και στη συνέχεια διαιρούνται σε όλο και μικρότερες υποομάδες. Η διαδικασία ολοκληρώνεται όταν κάθε παρατήρηση αποτελεί μια ομάδα. Ο κύριος λόγος που οι διαιρετικές μέθοδοι δεν προτιμώνται ιδιαίτερα είναι διότι απαιτούν πολύ περισσότερους υπολογισμούς από τις συσσωρευτικές.

Οι αλγόριθμοι Ιεραρχικής ομαδοποίησης είναι εύκολοι στην κατανόηση και δεν απαιτούν τον προσδιορισμό των ομάδων εξ αρχής. Από την άλλη χαρακτηρίζονται και από κάποια βασικά μειονεκτήματα:

- Από άποψη υπολογιστικού φόρτου είναι ασύμφοροι για μεγάλα σετ δεδομένων.
- Οι ομάδες που φτιάχνονται στα αρχικά βήματα, δεν μπορούν να χωρισθούν και επομένως οι αντίστοιχες παρατηρήσεις μένουν πάντα μαζί.
- Δίνουν μια ποικιλία πιθανών λύσεων, επομένως απαιτείται ένα κριτήριο που θα βοηθάει στην επιλογή της βέλτιστης λύσης.

**Β) Μη Ιεραρχική Ομαδοποίηση**

Στην Μη Ιεραρχική Ομαδοποίηση, ο στόχος είναι να καταταχθούν οι *n* παρατηρήσεις σε *k* ομάδες ο αριθμός των οποίων είναι προκαθορισμένος. Η λειτουργία βασίζεται είτε στην επιλογή *k* ατόμων τα οποία αποτελούν τον «πυρήνα» των μετέπειτα ομάδων και γύρω από αυτά ταξινομούνται τα υπόλοιπα στοιχεία, είτε εξ αρχής δημιουργούνται *k* ομάδες μεταξύ των οποίων στη συνέχεια οι παρατηρήσεις ανακατανέμονται. Οι αλγόριθμοι που χρησιμοποιούνται στηρίζονται στην έννοια του κέντρου βάρους μιας ομάδας, δηλαδή τη μέση τιμή κάθε μεταβλητής, γύρω από το οποίο κατατάσσονται οι παρατηρήσεις ανάλογα με την απόσταση που έχουν από αυτό. Με τον όρο απόσταση στην μέθοδο αυτή εννοούμε την Ευκλείδεια απόσταση, χωρίς όμως να αποκλείεται η χρήση και των άλλων μέτρων απόστασης. Στο σημείο αυτό θα αναλυθούν δύο από τις πιο γνωστές μεθόδους μη ιεραρχικής ομαδοποίησης:

**i. Η μέθοδος Forgy**

Για την υλοποίηση της μεθόδου αυτής, καθορίζονται στην αρχή *k*-ομάδες για τις οποίες υπολογίζεται το κέντρο βέρους. Στη συνέχεια κατατάσσονται οι υπόλοιπες *n-k* παρατηρήσεις στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την εκάστοτε παρατήρηση. Έπειτα υπολογίζονται ξανά τα νέα κέντρα βάρους των ομάδων. Οι παρατηρήσεις ανακατατάσσονται μεταξύ των ομάδων με βάση τα νέα κέντρα βάρους. Η διαδικασία επαναλαμβάνεται μέχρι να σταματήσουν να υπάρχουν παρατηρήσεις που αλλάζουν ομάδα και κατ' επέκταση η διαδικασία οδηγείται σε ένα «σημείο ισορροπίας». Συνήθως απαιτούνται το πολύ 10 επαναλήψεις.

### ii. Η μέθοδος K-Means

Όμοια με τη μέθοδο Forgy, για την υλοποίηση της μεθόδου K-Means απαιτείται στην αρχή ο καθορισμός *k*-ομάδων και ο υπολογισμός των κέντρων βάρους τους. Στη συνέχεια κατατάσσεται καθεμία από τις υπόλοιπες *n-k* παρατηρήσεις στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση αυτή. Μετά από κάθε τοποθέτηση επαναϋπολογίζεται το κέντρο βάρους των ομάδων. Η διαδικασία ολοκληρώνεται όταν όλα τα άτομα έχουν τοποθετηθεί σε ομάδες και γίνει και η τελευταία σάρωση των δεδομένων.

Η διαφορά των δύο μεθόδων έγκειται στο σημείο όπου γίνεται ο υπολογισμός του κέντρου βάρους των ομάδων. Στην μεν μέθοδο Forgy γίνεται με την ολοκλήρωση της ταξινόμησης όλων των ατόμων, στη δε μέθοδο K-Means υπολογίζεται με την προκύπτουσα σύνθεση των ομάδων μετά από κάθε ταξινόμηση ενός ατόμου. Το γεγονός αυτό κάνει τη μέθοδο K-Means να υπερέχει σε ταχύτητα σε σχέση με τη μέθοδο Forgy καθώς απαιτεί λίγες σαρώσεις των δεδομένων, και συνεπώς λίγες επαναλήψεις του αλγορίθμου.

Βασικό μειονέκτημα των Μη Ιεραρχικών μεθόδων είναι η εξάρτηση του αλγορίθμου από τα αρχικά σημεία που έχουν επιλεγεί. Μια λάθος αρχική επιλογή παρατηρήσεων μπορεί να οδηγήσει σε εντελώς διαφορετικές ομάδες από τη φυσική ομαδοποίηση των δεδομένων. Επομένως είναι χρήσιμή η επανάληψη της μεθόδου με επιλογή διαφορετικών σημείων κάθε φορά.

## 3.5 Ανάλυση Αντιστοιχιών

Η Ανάλυση Αντιστοιχιών είναι μια τεχνική ανάλυσης δεδομένων η οποία σχεδιάστηκε για την ανάλυση διδιάστατων και πολυδιάστατων πινάκων που περιέχουν ένα μέτρο αντιστοιχίας μεταξύ γραμμών και στηλών. Πρόκειται για μια μέθοδο που εφαρμόζεται σε κατηγορικά δεδομένα και δίνει τη δυνατότητα να εντοπισθούν οι συστηματικές σχέσεις μεταξύ των μεταβλητών όταν δεν υπάρχει εκ των προτέρων προσδοκία ως προς τη φύση των σχέσεων αυτών. Η μέθοδος αυτή αναλύει πίνακες 2 διαστάσεων (Απλή Ανάλυση Αντιστοιχιών), ή περισσότερων διαστάσεων (Πολλαπλή Ανάλυση Αντιστοιχιών). Στον τομέα της πολυμεταβλητής ανάλυσης, η Ανάλυση Αντιστοιχιών δίνει τη δυνατότητα απλοποίησης εξαιρετικά σύνθετων δεδομένων παρέχοντας μια λεπτομερή περιγραφή για κάθε στοιχείο που

δίνεται από τα δεδομένα. Με τον τρόπο αυτό δεν χάνει μεγάλο μέρος της πληροφορίας που αυτά παρέχουν. Η μέθοδος αυτή διαφοροποιείται από άλλες τεχνικές ανάλυσης δεδομένων εξαιτίας:

- ✓ Της εφαρμογής της σε πολυμεταβλητά δεδομένα
- ✓ Του τρόπου γραφικής απεικόνισης της μεθόδου
- ✓ Των ελάχιστων απαιτήσεων που έχει ως προς τα δεδομένα.

Οι προϋποθέσεις πάνω στις οποίες βασίζεται είναι:

- ➢ Ο πίνακας δεδομένων είναι αρκετά μεγάλος ώστε να μην είναι εύκολη η κατανόηση του με τη χρήση των απλών στατιστικών μεθόδων.
- ➢ Υπάρχει ομοιογένεια μεταξύ των μεταβλητών έτσι ώστε να είναι λογικός ο υπολογισμός των μεταξύ τους αποστάσεων.

Η Ανάλυση Αντιστοιχιών θα μπορούσε να οριστεί ως ειδική περίπτωση της Ανάλυσης Κυρίων Συνιστωσών. Ωστόσο είναι γεγονός πως οι δύο μέθοδοι χρησιμοποιούν τελείως διαφορετικό εύρος δεδομένων καθώς η δεύτερη βασίζεται σε πίνακες με συνεχείς μετρήσεις. Από την άλλη, η γραφική απεικόνιση της μεθόδου γίνεται αναπαριστώντας τις γραμμές και τις στήλες με σημεία στο χώρο λίγων διαστάσεων. Γειτονικά σημεία πάνω στο γράφημα υποδηλώνουν εκτός των άλλων, συσχετισμό ανάμεσα στις αντίστοιχες γραμμές και στήλες. Ο έλεγχος πάνω στον οποίο βασίζεται η επιβεβαίωση της συσχέτισης των μεταβλητών είναι το $\chi^2$ τεστ. Με τη μέθοδο αυτή, ο μελετητής δεν βλέπει μόνο εάν υπάρχει συσχέτιση μεταξύ των μεταβλητών αλλά και τον βαθμό αυτής της συσχέτισης. Η Ανάλυση Αντιστοιχιών σε πολυμεταβλητά δεδομένα απαιτεί τη χρήση του πίνακα αρχικών παρατηρήσεων έπειτα από κατάλληλη μετατροπή των κατηγορικών μεταβλητών. Για τον λόγο αυτό, είναι απαραίτητη η χρήση του «πίνακα δείκτη» (*disjunctive matrix*) ή εναλλακτικά του πίνακα «Burt».

### 3.5.1 Πίνακας δείκτης και Πίνακας Burt

Τα πολυδιάστατα δεδομένα, όπως είναι γνωστό, αναπαριστώνται σε πίνακες των οποίων οι γραμμές αντιστοιχούν στις παρατηρήσεις και οι στήλες στις μεταβλητές. Ομοίως, η ανάλυση αντιστοιχιών βασίζεται στην αναπαράσταση ενός τέτοιου πίνακα με τη δημιουργία ενός «πίνακα δείκτη» $\boldsymbol{Z}$. Ο πίνακας $\boldsymbol{Z}$ αποτελείται από ψευδομεταβλητές για όλες τις κατηγορίες των μεταβλητών του αρχικού πίνακα δεδομένων και από τόσες γραμμές όσες είναι οι αρχικές παρατηρήσεις. Η ψευδομεταβλητή αντιστοιχίζεται με τον αριθμό 1 εάν η τιμή της μεταβλητής ανήκει στην συγκεκριμένη κατηγορία και 0 εάν δεν ανήκει. Για κάθε

κατηγορική μεταβλητή κατασκευάζονται τόσες ψευδομεταβλητές, όσες και τα επίπεδά της. Σύμφωνα με τα παραπάνω, σε κάθε «block» ψευδομεταβλητών μπορεί να υπάρχει μόνο ένα 1 και όλες οι υπόλοιπες τιμές είναι 0. Έτσι εάν έχουμε $p$ κατηγορικές μεταβλητές με $c$ το πλήθος όλων των διαφορετικών κατηγοριών των μεταβλητών και $n$ συνολικές παρατηρήσεις, τότε ο πίνακας δείκτης μπορεί να γραφτεί ως $Z = [Z_1, \dots, Z_p]$ με διαστάσεις $n \times c$ και μπορεί να αναλυθεί με τη μέθοδο της ανάλυσης αντιστοιχιών όπως και κάθε άλλος πίνακας δύο διαστάσεων.

Εναλλακτικά, από τον πίνακα δείκτη $Z$ που αναφέρθηκε, μπορεί να κατασκευαστεί ο λεγόμενος πίνακας «Burt». Πρόκειται για το αποτέλεσμα του γραμμικού γινομένου ενός πίνακα δείκτη, δηλαδή είναι ο πίνακας $B = Z'Z$ . Ο πίνακας «Burt» είναι ένας συμμετρικός πίνακας διάστασης $c \times c$ στην διαγώνιο του οποίου βρίσκονται οι πίνακες συνάφειας κάθε μεταβλητής με τον εαυτό της. Επομένως η μορφή του πίνακα «Burt» είναι η εξής:

$$B = \begin{bmatrix} Z_1'Z_1 & \cdots & Z_1'Z_p \\ \vdots & \ddots & \vdots \\ Z_p'Z_1 & \cdots & Z_p'Z_p \end{bmatrix}$$

Λόγω της συμμετρίας του πίνακα $B$, η ανάλυση καταλήγει σε μια απλή ανάλυση στηλών.
Η πολλαπλή ανάλυση αντιστοιχιών είναι η ανάλυση αντιστοιχιών του πίνακα δείκτη ή του πίνακα Burt.

# ΚΕΦΑΛΑΙΟ 4

# Εφαρμογή σε πραγματικά δεδομένα

## 4.1 Εισαγωγή

Το κεφάλαιο που ακολουθεί θα εξετάσει στοιχεία του πρώτου κύματος της έρευνας SHARE (Survey on Health Ageing and Retirement in Europe). Στόχος του κεφαλαίου είναι να διαπιστωθεί κατά πόσο οι ευρωπαϊκές χώρες μπορούν να χωριστούν σε ομάδες βάσει των χαρακτηριστικών που προσδιορίζουν το οικονομικό και το ψυχολογικό επίπεδο των κατοίκων τους, αλλά και κατά πόσο οι ομάδες αυτές παρουσιάζουν ομοιότητες μεταξύ τους.

Το πρώτο κύμα της έρευνας του SHARE συμπεριλαμβάνει ένα σύνολο δεδομένων από έντεκα ευρωπαϊκές χώρες. Στην πορεία της έρευνας θα μελετηθούν και θα αναλυθούν στατιστικά 16537 παρατηρήσεις της έρευνας, οι οποίες αποτελούν το σύνολο των ατόμων μετά την αφαίρεση των ελλειπουσών και των ακραίων τιμών στα υπό μελέτη χαρακτηριστικά. Τα άτομα της μελέτης είναι κάτοικοι των έντεκα ευρωπαϊκών χωρών, ηλικίας 50 ετών και άνω και διακρίνονται ως προς κάποια βασικά δημογραφικά, οικονομικά και ψυχολογικά στοιχεία. Στην πρώτη ενότητα του κεφαλαίου θα χρησιμοποιηθεί ένα δείγμα οικονομικών μετρήσεων των έντεκα αυτών χωρών βάσει των οποίων θα γίνει μια πρώτη ομαδοποίηση τους. Το δείγμα αυτό επιλέγεται τυχαία και αντιπροσωπεύει το σύνολο των παρατηρήσεων με σκοπό την διευκόλυνση της γραφικής απεικόνισης των δεδομένων. Στη συνέχεια, η έρευνα θα ξεκινήσει με μια απλή περιγραφική αναφορά των δεδομένων καθώς κι έναν έλεγχο ύπαρξης συσχετίσεων των χαρακτηριστικών με τον παράγοντα χώρα. Έπειτα, θα εξετασθεί με «Πολλαπλή ανάλυση αντιστοιχιών», πως διαμορφώνονται τα στοιχεία που συνθέτουν την ψυχική κατάσταση της υγείας των ατόμων σε σχέση με την χώρα διαμονής και ποιες διαφορές παρουσιάζονται μεταξύ των χωρών. Κατά τον ίδιο τρόπο, με χρήση της μεθόδου «Ανάλυση σε κύριες συνιστώσες», θα μειωθούν σε διάσταση τα στοιχεία που συνθέτουν την οικονομική κατάσταση των κατοίκων και θα ομαδοποιηθούν οι ευρωπαϊκές

χώρες ως προς το οικονομικό επίπεδο των κατοίκων τους. Η διαδικασία θα ολοκληρωθεί με παρουσίαση των συμπερασμάτων που θα εξαχθούν.

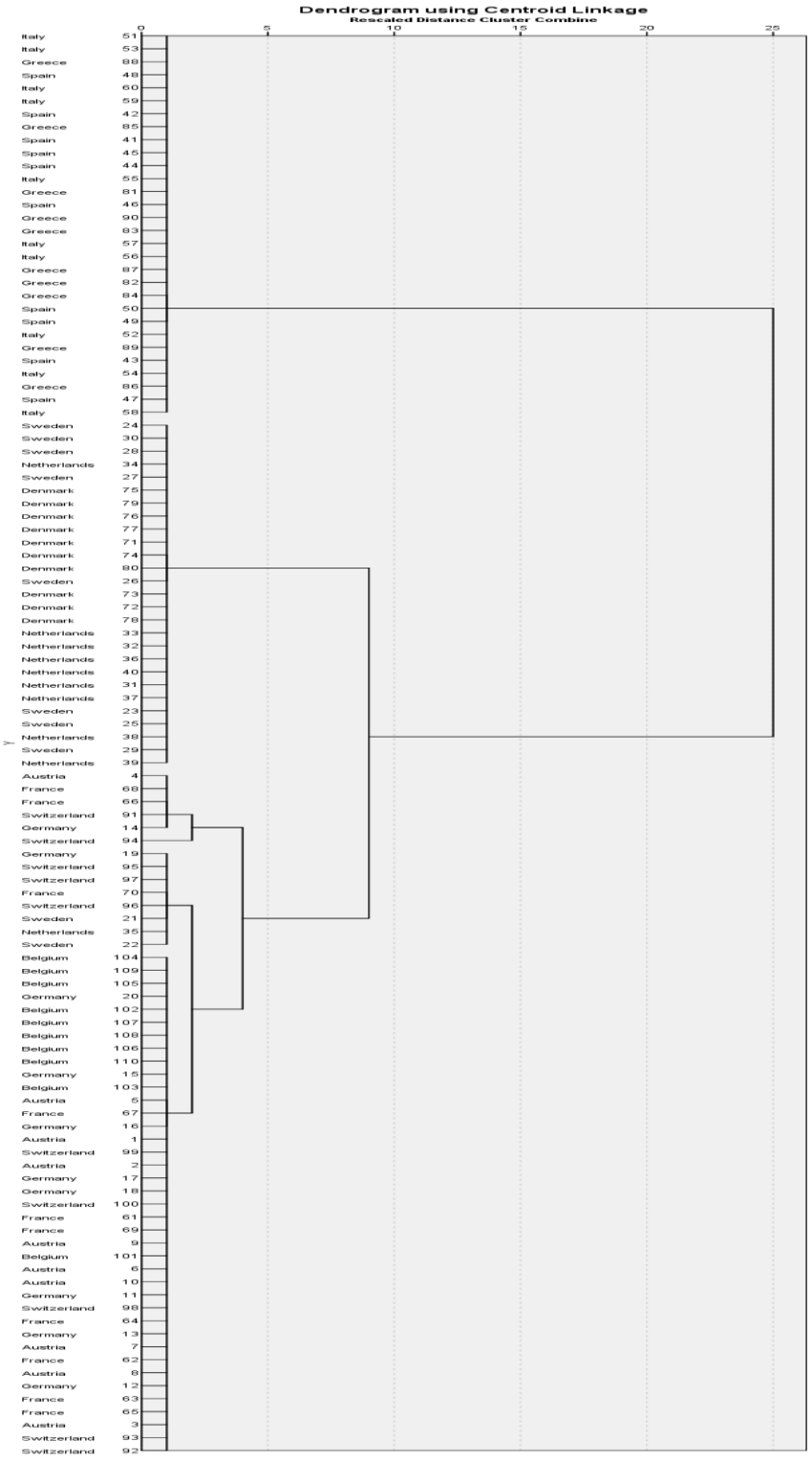
## 4.2 Δενδρόγραμμα και Ανάλυση κατά συστάδες των Ευρωπαϊκών χωρών

Στην παρούσα ενότητα θα χρησιμοποιηθεί ένα τυχαίο δείγμα 110 παρατηρήσεων από τα δεδομένα του πρώτου κύματος της έρευνας SHARE. Όπως αναφέρθηκε και παραπάνω, ο λόγος που επιλέγεται μόλις ένα μικρό μέρος του δείγματος είναι για την επίτευξη μιας πιο εύκολης ανάγνωσης τόσο του γραφήματος όσο και των πινάκων που θα χρησιμοποιηθούν κατά την ομαδοποίηση. Οι παρατηρήσεις είναι ισομερώς κατανεμημένες και προέρχονται από τις 11 χώρες που περιέχονται στο πρώτο κύμα και συγκεκριμένα από την Αυστρία, το Βέλγιο, τη Γαλλία, τη Γερμανία, την Δανία, την Ελβετία, την Ελλάδα, την Ισπανία, την Ιταλία, την Ολλανδία και την Σουηδία. Οι μετρήσεις που έχουν καταγραφεί για κάθε παρατήρηση αφορούν στο Ακαθάριστο οικογενειακό εισόδημα (Yhh_nir) και στο Καθαρό οικογενειακό εισόδημα (Yhh_net) με τη λογική πως τα οικονομικά αυτά χαρακτηριστικά καθορίζουν σε μεγάλο βαθμό την οικονομική κατάσταση του νοικοκυριού. Στόχος της ενότητας αυτής είναι η ομαδοποίηση των παρατηρήσεων. Κάθε ομάδα πρέπει να συμπεριλαμβάνει παρατηρήσεις με παρόμοιο οικογενειακό εισόδημα ενώ οι τιμές μεταξύ των ομάδων θα απέχουν πολύ. Σύμφωνα με τα παραπάνω, αρχικά απαιτείται η κατασκευή ενός δενδρογράμματος (Ιεραρχική μέθοδος ομαδοποίησης) η απεικόνιση του οποίου θα βοηθήσει στην επιλογή των ομάδων. Στη συνέχεια, η καταλληλότερη μέθοδος είναι μια μη ιεραρχική μέθοδος και συγκεκριμένα η μέθοδος «K-Means».

Το δενδρόγραμμα που ακολουθεί κατασκευάστηκε με χρήση της μεθόδου «Centroid» σύμφωνα με την οποία, η απόσταση μεταξύ των ομάδων υπολογίζεται ως απόσταση των κέντρων των ομάδων. Αποτέλεσμα της μεθόδου είναι η παραγωγή συμπαγών ομάδων. Αξίζει να σημειωθεί επίσης, ότι ως μέτρο απόστασης μεταξύ των παρατηρήσεων χρησιμοποιήθηκε η «Τετραγωνική Ευκλείδεια απόσταση» καθώς είναι απλή μέθοδος και δεν υπάρχει διαφορετική κλίμακα μέτρησης μεταξύ των παρατηρήσεων ώστε να την επηρεάσει.

**ΣΧΗΜΑ 4.2.1**

Δενδρόγραμμα χωρών σε σχέση με Ακαθάριστο και Καθαρό εισόδημα



Στο παραπάνω σχήμα απεικονίζονται οι ομάδες που σχηματίζονται σταδιακά μεταξύ των παρατηρήσεων. Τα χαρακτηριστικά στα οποία βασίζεται η ομαδοποίηση είναι το Καθαρό και Ακαθάριστο εισόδημα των νοικοκυριών. Αριστερά καταγράφεται η χώρα στην οποία ανήκει η κάθε παρατήρηση και δεξιά των χωρών είναι η σειρά με την οποία επιλέχθηκαν οι παρατηρήσεις στην ομαδοποίηση. Με βάση το δενδρόγραμμα θα μπορούσε κανείς να διακρίνει ότι προκύπτουν τρεις ομάδες παρατηρήσεων αφού παρατηρώντας το επόμενο επίπεδο συνένωσης (2 ομάδες) η απόσταση στην οποία πραγματοποιείται φθάνει περίπου το 9 όταν πριν άγγιζε το 5. Έτσι θα μπορούσαμε να δούμε ότι το γράφημα κατατάσσει στην ίδια ομάδα τις πρώτες 30 παρατηρήσεις οι οποίες συμπεριλαμβάνουν κατοίκους της Ελλάδας, της Ισπανίας και της Ιταλίας. Έπειτα οι επόμενες 27 παρατηρήσεις ανήκουν σε μια δεύτερη ομάδα και αποτελούνται από κατοίκους της Δανίας, της Ολλανδίας και της Σουηδίας. Τέλος, όλες οι υπόλοιπες παρατηρήσεις που προέρχονται κυρίως από κατοίκους της Αυστρίας, Ελβετίας, Γερμανίας, Γαλλίας και του Βελγίου αποτελούν μια τρίτη ομάδα.
Τα παραπάνω συμπεράσματα μπορούν να επαληθευθούν χρησιμοποιώντας τη μη Ιεραρχική μέθοδο K-Means για τις 110 παρατηρήσεις και επιλέγοντας τη δημιουργία τριών ομάδων. Στον πίνακα που ακολουθεί παρουσιάζεται το πλήθος των ατόμων που αποτελούν την κάθε ομάδα.

**ΠΙΝΑΚΑΣ 4.2.1**

Πλήθος ατόμων σε κάθε ομάδα

| | Άτομα |
|---|---|
| Ομάδα 1 | 48 |
| Ομάδα 2 | 32 |
| Ομάδα 3 | 30 |
| Σύνολο | 110 |

Όπως φαίνεται από τον παραπάνω πίνακα, από τα 110 άτομα του δείγματος, τα 48 εντάσσονται στην πρώτη ομάδα, τα 32 στην δεύτερη και 30 στην τρίτη.

Η μέθοδος K-Means, όπως έχει αναφερθεί, χρησιμοποιεί κάποια αρχικά κέντρα βάρους με βάση τις τιμές των μεταβλητών, ενώ με την ολοκλήρωση των επαναλήψεων ορίζονται και τα τελικά κέντρα των ομάδων. Στην παρούσα εφαρμογή χρειάσθηκαν 5 επαναλήψεις για την ολοκλήρωση του αλγορίθμου ενώ τα αρχικά και τα τελικά κέντρα των ομάδων παρουσιάζονται στον ακόλουθο πίνακα:

**ΠΙΝΑΚΑΣ 4.2.2**

Αρχικά και τελικά κέντρα ομάδων

| | 1η ομάδα | | 2η ομάδα | | 3η ομάδα | |
|---|---|---|---|---|---|---|
| | Αρχικό κέντρο | Τελικό κέντρο | Αρχικό κέντρο | Τελικό κέντρο | Αρχικό κέντρο | Τελικό κέντρο |
| Yhh_nir | 196273,63 | 156808,68 | 106680,51 | 91603,62 | 900,00 | 8456,75 |
| Yhh_net | 180369,38 | 121560,14 | 62853,80 | 67052,19 | 900,00 | 7863,66 |

Σύμφωνα με τον Πίνακα 4.2.2, το αρχικό κέντρο της ομάδας 1 για το ακαθάριστο οικογενειακό εισόδημα ήταν 196273,63 και το τελικό κέντρο ήταν 156808,68. Για την ίδια ομάδα, το αρχικό κέντρο για το καθαρό οικογενειακό εισόδημα ήταν 180369,38 και το τελικό ήταν 121560,14. Αντίστοιχα, για την ομάδα 2 το αρχικό κέντρο για το ακαθάριστο οικογενειακό εισόδημα ήταν 106680,51 και το τελικό κέντρο ήταν 91603,62. Ομοίως για το καθαρό οικογενειακό εισόδημα το αρχικό κέντρο ήταν 62853,80 και το τελικό κέντρο ήταν 67052,19. Τέλος, για την ομάδα 3, το αρχικό κέντρο για το ακαθάριστο οικογενειακό εισόδημα ήταν 900 και το τελικό κέντρο ήταν 8456,75. Το καθαρό οικογενειακό εισόδημα για την τρίτη ομάδα είχε αρχικό κέντρο 900 και τελικό κέντρο ήταν 7863,66. Τα παραπάνω οδηγούν στο συμπέρασμα πως οι χώρες που ανήκουν στην πρώτη ομάδα παρουσιάζουν οικονομική ευημερία καθώς τα τελικά κέντρα έχουν τις υψηλότερες τιμές. Στη συνέχεια, οι χώρες της δεύτερης ομάδας χαρακτηρίζονται από μέτρια οικονομία. Τέλος στην τρίτη ομάδα, οι χώρες δείχνουν να έχουν οικονομική δυσχέρεια καθώς τε τελικά κέντρα έχουν τις χαμηλότερες τιμές.

Ολοκληρώνοντας τη διαδικασία παρουσιάζεται η διαμόρφωση των ομάδων σε σχέση με τη χώρα διαμονής. Στον πίνακα που ακολουθεί καταγράφεται στην πρώτη στήλη η χώρα, στην δεύτερη στήλη η ομάδα στην οποία εντάσσονται τα άτομα και στην τρίτη στήλη, το πλήθος των ατόμων που ανήκουν σε αυτή την ομάδα. Στόχος είναι να εξετασθεί αφενός εάν τα άτομα που κατοικούν στην ίδια χώρα τυγχάνει να εντάσσονται στην πλειοψηφία τους στην ίδια ομάδα και αφετέρου ποιες χώρες παρουσιάζουν ομοιότητες ως προς τα οικονομικά τους χαρακτηριστικά και επομένως ανήκουν στην ίδια ομάδα.

**ΠΙΝΑΚΑΣ 4.2.3**

Χώρα ατόμου και συχνότητα ομάδας κατάταξης του

| Χώρα | Ομάδα | Συχνότητα |
|---|---|---|
| Αυστρία | 1 | 10 |
| Βέλγιο | 1 | 10 |
| Δανία | 2 | 10 |
| Γαλλία | 1 | 10 |
| Γερμανία | 1 | 9 |
| | 2 | 1 |
| Ελλάδα | 3 | 10 |
| Ιταλία | 3 | 10 |
| Ολλανδία | 1 | 1 |
| | 2 | 9 |
| Ισπανία | 3 | 10 |
| Σουηδία | 2 | 10 |
| Ελβετία | 1 | 8 |
| | 2 | 2 |

Από τον Πίνακα 4.2.3, προκύπτει το συμπέρασμα ότι οι κάτοικοι της Αυστρίας, του Βελγίου, της Γαλλίας, της Γερμανίας και της Ελβετίας ανήκουν στην πλειοψηφία τους στην πρώτη ομάδα. Αντίστοιχα, οι κάτοικοι της Δανίας, της Ολλανδίας και της Σουηδίας ανήκουν στη δεύτερη ομάδα. Τέλος, οι κάτοικοι της Ελλάδας, της Ιταλίας και της Ισπανίας ανήκουν στην τρίτη ομάδα. Είναι φανερό ότι τα συμπεράσματα που προκύπτουν επιβεβαιώνουν το δενδρόγραμμα που προηγήθηκε. Σύμφωνα με τα παραπάνω, μπορούμε να θεωρήσουμε ότι οι χώρες που ανήκουν σε κάθε ομάδα έχουν κοινά οικονομικά χαρακτηριστικά. Οι χώρες της πρώτης ομάδας θεωρούνται «Ηπειρωτικές χώρες», οι χώρες της δεύτερης ομάδας είναι οι «Βόρειες χώρες» και τέλος, οι κάτοικοι της τρίτης ομάδας είναι οι «Νότιες χώρες». Αξίζει να σημειωθεί ότι την ίδια ομαδοποίηση έχει κάνει και το SHARE χρησιμοποιώντας τους ίδιους τίτλους των ομάδων βάσει της γεωγραφικής θέσης των χωρών.

## 4.3 Ανάλυση των διαθέσιμων δεδομένων

Σκοπός της περιγραφικής ανάλυσης είναι να δώσει μια συνοπτική παρουσίαση του δείγματος, καθώς και να ελέγξει την ορθότητα των τιμών του. Αυτό γίνεται με τη βοήθεια απλών περιγραφικών πινάκων ή γραφημάτων. Η επιλογή των κατάλληλων αριθμητικών και περιγραφικών μεθόδων γίνεται με βάση τον τύπο της μεταβλητής που θέλουμε να παρουσιάσουμε.

**ΠΙΝΑΚΑΣ 4.3.1**

Πίνακας συχνοτήτων για τη μεταβλητή «χώρα»

| Χώρα | Συχνότητα | Ποσοστό |
|---|---|---|
| Αυστρία | 1587 | 9,6 |
| Βέλγιο | 2440 | 14,8 |
| Γαλλία | 933 | 5,6 |
| Γερμανία | 1779 | 10,8 |
| Δανία | 1107 | 6,7 |
| Ελβετία | 649 | 3,9 |
| Ελλάδα | 1602 | 9,7 |
| Ισπανία | 1170 | 7,1 |
| Ιταλία | 1393 | 8,4 |
| Ολλανδία | 1900 | 11,5 |
| Σουηδία | 1977 | 12,0 |
| Σύνολο | 16537 | 100,0 |

Όπως φαίνεται από τον Πίνακα 4.3.1, το δείγμα αποτελείται από 16537 άτομα διαχωρισμένα σε έντεκα ευρωπαϊκές χώρες. Από τον πίνακα που ακολουθεί προκύπτει ότι στην έρευνα συμμετέχουν 1587 Αυστριακοί (9,6%), 2440 Βέλγοι (14,8%), 933 Γάλλοι (5,6%), 1779 Γερμανοί (10,8%), 1107 Δανοί (6,7%), 649 Ελβετοί (3,9%), 1602 Έλληνες (9,7%), 1170 Ισπανοί (7,1%), 1393 Ιταλοί (8,4%), 1900 Ολλανδοί (11,5%) και 2977 Σουηδοί (12%).

Επιπλέον, σύμφωνα με τον ακόλουθο πίνακα, οι γυναίκες υπερτερούν σε σχέση με τους άνδρες.

**ΠΙΝΑΚΑΣ 4.3.2**

Πίνακας συχνοτήτων για τη μεταβλητή «φύλο»

| Φύλο | Συχνότητα | Ποσοστό |
|---|---|---|
| Γυναίκα | 8917 | 53,9 |
| Άνδρας | 7620 | 46,1 |
| Σύνολο | 16537 | 100% |

Οι γυναίκες αναλογούν στο δείγμα σε ποσοστό 53,9% ενώ το αντίστοιχο ποσοστό των ανδρών ισοδυναμεί με 46,1% του δείγματος.

Ο μέσος όρος ηλικίας του δείγματος υπολογίζεται περίπου 64,19 έτη και η διάμεσος είναι 63 έτη. Επιπλέον η μικρότερη τιμή που παρατηρείται είναι η ηλικία των 50 ετών και η μεγαλύτερη τιμή είναι η ηλικία των 99 ετών.

**ΠΙΝΑΚΑΣ 4.3.3**

Πίνακας συχνοτήτων για τη μεταβλητή «ηλικία»

| Μεταβλητή | Μέγεθος δείγματος | Ελάχιστο | Μέγιστο | Μέση τιμή | Διάμεσος | Διασπορά |
|---|---|---|---|---|---|---|
| Ηλικία | 16537 | 50 | 99 | 64,19 | 63,00 | 94,356 |

Το μεγαλύτερο μέρος του δείγματος κατοικεί σε ένα μεγάλο αστικό κέντρο (a large town) και το μικρότερο ποσοστό διαμένει σε μια αγροτική περιοχή ή χωριό (a rural area or village). Τα αποτελέσματα αυτά απεικονίζονται τόσο στον ακόλουθο πίνακα όσο και στο Σχήμα 4.3.1:

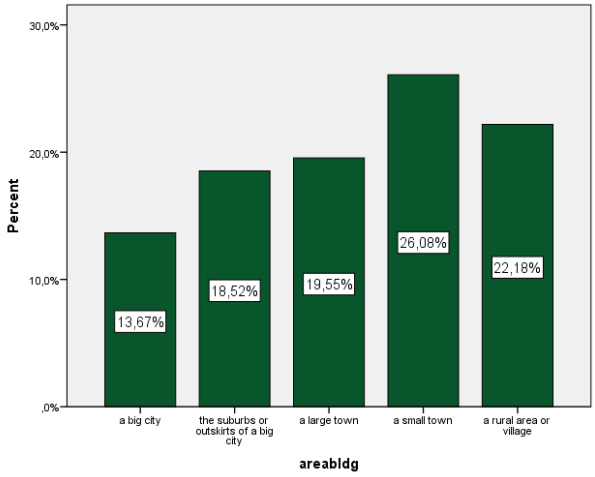
**ΠΙΝΑΚΑΣ 4.3.4**

Πίνακας συχνοτήτων για τη μεταβλητή «περιοχή διαμονής»

| Περιοχή διαμονής | Συχνότητα | Ποσοστό |
|---|---|---|
| Μεγάλη πόλη | 2260 | 13,7% |
| Προάστιο μεγάλης πόλης | 3063 | 18,5% |
| Μεγάλο αστικό κέντρο | 3233 | 19,6% |
| Μικρό αστικό κέντρο | 4313 | 26,1% |
| Αγροτική περιοχή ή χωριό | 3668 | 22,2% |
| Σύνολο | 16537 | 100% |

**ΣΧΗΜΑ 4.3.1**

Ραβδόγραμμα συχνοτήτων για τη μεταβλητή «περιοχή διαμονής»



Από τον παραπάνω πίνακα και το σχήμα, φαίνεται πως το δείγμα αποτελείται από 2260 άτομα (13,7%) που κατοικούν σε μια μεγάλη πόλη (a big city), 3063 άτομα (18,5%) που κατοικούν σε ένα προάστιο μεγάλης πόλης (the suburbs or outskirts of a big city), 3233 άτομα (19,6%) που κατοικούν σε ένα μεγάλο αστικό κέντρο (a large town), 4313 άτομα (26,1%) που κατοικούν σε ένα μικρό αστικό κέντρο (a small town) και 3668 άτομα (22,2%) που κατοικούν σε μια αγροτική περιοχή ή χωριό (a rural area or village).

Το εκπαιδευτικό επίπεδο του δείγματος έχει ένα μεγάλο εύρος τιμών καθώς κυμαίνεται από 2 έτη εκπαίδευσης, που σημαίνει πως το εκάστοτε άτομο δεν έχει ολοκληρώσει την πρωτοβάθμια εκπαίδευση, έως 21 έτη εκπαίδευσης, που σημαίνει πως το άτομο έχει συνεχίσει τις σπουδές του στην τριτοβάθμια εκπαίδευση. Ο μέσος όρος του δείγματος είναι περίπου 10,53 έτη εκπαίδευσης και η διάμεσος είναι 11 έτη.

**ΠΙΝΑΚΑΣ 4.3.5**

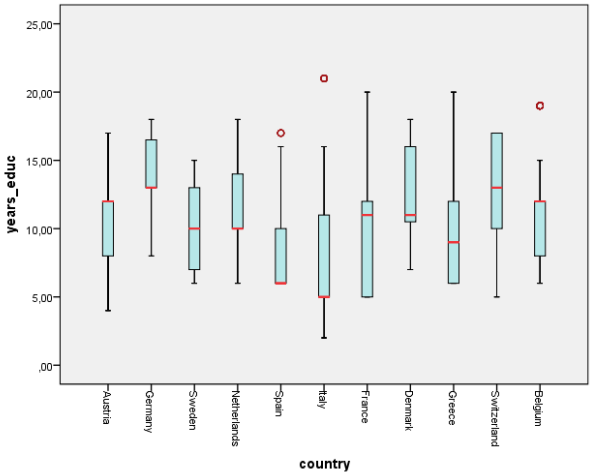
Πίνακας συχνοτήτων για τη μεταβλητή «έτη εκπαίδευσης»

| Μεταβλητή | Μέγεθος δείγματος | Ελάχιστο | Μέγιστο | Μέση τιμή | Διάμεσος | Διασπορά |
|---|---|---|---|---|---|---|
| Έτη εκπαίδευσης | 16537 | 2,00 | 21,00 | 10,53 | 11,00 | 15,106 |

Στο σημείο αυτό θα μπορούσαμε να δούμε πως διαμορφώνεται το επίπεδο εκπαίδευσης ανά χώρα γραφικά. Έτσι κατασκευάστηκαν τα παρακάτω 11 θηκογράμματα, ένα για κάθε χώρα:

**ΣΧΗΜΑ 4.3.2**

Θηκογράμματα ετών εκπαίδευσης ανά χώρα




Η κόκκινη γραμμή κάθε θηκογράμματος δείχνει τη διάμεσο των ετών εκπαίδευσης για κάθε χώρα. Τα άκρα κάθε σχήματος αντιστοιχεί στη μέγιστη και ελάχιστη τιμή ετών

εκπαίδευσης και κατ' επέκταση η απόσταση των άκρων αντιστοιχεί στο εύρος τιμών τους. Οι μικροί κόκκινοι κύκλοι που έχουν σημειωθεί πάνω από τα θηκογράμματα της Ισπανίας, της Ιταλίας και του Βελγίου αντιστοιχούν σε έκτροπες παρατηρήσεις των χωρών αυτών. Σύμφωνα με όσα προαναφέρθηκαν, η μικρότερη διάμεσος των ετών εκπαίδευσης εντοπίζεται στην Ιταλία και είναι ίση με 5 έτη. Στην ίδια χώρα εντοπίζεται και η μικρότερη από τις ελάχιστες τιμές ετών εκπαίδευσης, όπου είναι τα 2 έτη. Τόσο η Γαλλία, όσο και η Ελλάδα είναι οι χώρες με τα περισσότερα έτη εκπαίδευσης. Και οι δύο χώρες έχουν μέγιστο αριθμό ετών 20 χρόνια, η διάμεσος και η ελάχιστη τιμή της Ελλάδας είναι 9 και 6 χρόνια αντίστοιχα και οι ανάλογες τιμές της Γαλλίας είναι 11 και 5 χρόνια αντίστοιχα. Η μεγαλύτερη τιμή διαμέσου παρατηρείται στην Γερμανία και στην Ελβετία και είναι 13 έτη. Στην Γερμανία, οι κάτοικοι έχουν τουλάχιστον 8 έτη σπουδών και το πολύ 18 έτη. Κατά συνέπεια παρατηρείται επίσης ένα υψηλό ποσοστό εκπαίδευσης. Στην Ελβετία το εύρος τιμών είναι μεγαλύτερο, καθώς η ελάχιστη τιμή ετών εκπαίδευσης είναι 5 έτη και η μέγιστη 17. Η Σουηδία και η Ολλανδία έχουν παρόμοιο εκπαιδευτικό επίπεδο καθώς και στις δύο χώρες δεν υπάρχει κάτοικος που να σημειώνει λιγότερα από 6 χρόνια εκπαίδευσης και η διάμεσος είναι 10 έτη. Η Αυστρία παρουσιάζει ένα σχετικά μεγάλο εύρος τιμών, σημειώνοντας ως ελάχιστη τιμή τα 4 έτη και ως μέγιστη τα 17 ενώ τέλος, το Βέλγιο κυμαίνεται από 6 έως 15 έτη σπουδών, σημειώνοντας όμως μια έκτροπη παρατήρηση στα 19 έτη σπουδών.

Όπως προαναφέρθηκε το δείγμα έχει απαντήσει σε ένα σύνολο ερωτήσεων που αφορούν στην οικονομική του κατάσταση. Επομένως, για την παρούσα έρευνα συλλέχθηκαν οι μετρήσεις που σημειώθηκαν για συγκεκριμένα οικονομικά μεγέθη. Πιο αναλυτικά, οι μεταβλητές που χρησιμοποιήθηκαν είναι οι ακόλουθες:

Ακαθάριστο οικογενειακό εισόδημα (Yhh_nir)

Καθαρό οικογενειακό εισόδημα (Yhh_net)

Οικογενειακό ετήσιο εισόδημα από σύνταξη γήρατος (Hhold_Pensions)

Οικογενειακό ετήσιο εισόδημα από εργασία (Hhold_labour)

Οικογενειακό ετήσιο εισόδημα από άλλες πηγές (Other_source)

Συνολική αξία των χρηματικών περιουσιακών στοιχείων του νοικοκυριού (hgfinv)

Συνολική αξία της ιδιόκτητης κατοικίας του νοικοκυριού (homev)

Συνολική αξία της άλλης ακίνητης περιουσίας του νοικοκυριού (horesv)

Συνολική αξία των παγίων περιουσιακών στοιχείων του νοικοκυριού (hrav)

Στον πίνακα που ακολουθεί καταγράφεται για κάθε οικονομικό στοιχείο, η μέση και η μέγιστη τιμή, η διασπορά, η διάμεσος καθώς και το πρώτο και τρίτο τεταρτημόριο. Στην συνέχεια θα παραταθεί ένας πιο αναλυτικός πίνακας με τις τιμές των παραπάνω στοιχείων για κάθε χώρα ξεχωριστά.

**ΠΙΝΑΚΑΣ 4.3.6**

Πίνακας συχνοτήτων για τα οικονομικά μεγέθη όλων των χωρών

| Μεταβλητή | Μέση Τιμή | Μέγιστη Τιμή | Διασπορά | Διάμεσος | $Q_1$ | $Q_3$ |
|---|---|---|---|---|---|---|
| Yhh_nir | 45490,98 | 5293357,10 | $2{,}09 \cdot 10^{10}$ | 28866,70 | 15906,81 | 52451,10 |
| Yhh_net | 36633,13 | 4883136,00 | $1{,}69 \cdot 10^{10}$ | 24088,40 | 14285,09 | 40797,82 |
| Hhold_Pensions | 15101,69 | 267850,41 | $3{,}52 \cdot 10^{8}$ | 11281,20 | 0,00 | 21600,00 |
| Hhold_labour | 17373,30 | 608500,00 | $1{,}29 \cdot 10^{9}$ | 0,00 | 0,00 | 23964,36 |
| Other_source | 13015,99 | 5203875,00 | $1{,}91 \cdot 10^{10}$ | 708,04 | 66,20 | 7225,65 |
| Hgfinv | 68075,62 | 13401712,20 | $5{,}64 \cdot 10^{10}$ | 15046,30 | 2010,92 | 60215,75 |
| Homev | 174055,60 | 15000000,0 | $2{,}01 \cdot 10^{11}$ | 112520,7 | 0,00 | 201645,4 |
| Horesv | 40056,23 | 600000,00 | $4{,}03 \cdot 10^{10}$ | 0,00 | 0,00 | 0,00 |
| Hrav | 256769,67 | 27980000,00 | $7{,}12 \cdot 10^{11}$ | 126369,30 | 26142,94 | 249420,75 |

Από τον πίνακα συχνοτήτων για τα οικονομικά μεγέθη, φαίνεται πως το μέσο ακαθάριστο οικογενειακό εισόδημα που παρουσιάζεται στο δείγμα είναι 45490,98€, όταν η μέγιστη τιμή για το ίδιο χαρακτηριστικό είναι 5293357,1€ και η αντίστοιχη διασπορά είναι περίπου 20.900.000.000€. Ομοίως, το μέσο καθαρό εισόδημα του δείγματος είναι 36633,13€, η διάμεσος είναι 24088,4€ και το πρώτο και τρίτο τεταρτημόριο έχουν τις τιμές 14285,09 και 40797,82 αντίστοιχα. Με τον ίδιο τρόπο, παρατηρείται ότι η μέση τιμή του οικογενειακού ετήσιου εισοδήματος από σύνταξη γήρατος είναι 15101,69€, η μέγιστη τιμή του οικογενειακού ετήσιου εισοδήματος από εργασία είναι 608500€, η διασπορά του οικογενειακού ετήσιου εισοδήματος από άλλες πηγές είναι 19.100.000.000€, το πρώτο τεταρτημόριο της συνολικής αξίας των χρηματικών περιουσιακών στοιχείων του νοικοκυριού είναι 2010,92€ και το τρίτο τεταρτημόριο της συνολικής αξίας της ιδιόκτητης κατοικίας του νοικοκυριού είναι 201645,4€, η μέση τιμή της συνολικής αξίας της άλλης ακίνητης περιουσίας του νοικοκυριού είναι 40056,23€ και τέλος, η διάμεσος της συνολικής αξίας των παγίων περιουσιακών στοιχείων του νοικοκυριού είναι 126369,30€.

Στο σημείο αυτό ακολουθεί ο Πίνακας 4.3.7 στον οποίο καταγράφονται τα βασικά στατιστικά μεγέθη των οικονομικών χαρακτηριστικών για κάθε χώρα της έρευνας:

**ΠΙΝΑΚΑΣ 4.3.7**

Πίνακας συχνοτήτων για οικονομικά μεγέθη των χωρών ξεχωριστά

| | | Yhh_nir | Yhh_net | Hhold_ Pensions | Hhold_ labour | Other_ source | Hgfinv | Homev | Horesv | Hrav |
|---|---|---|---|---|---|---|---|---|---|---|
| **Αυστρία** | Μέση Τιμή | 39201,05 | 33952,85 | 20758,34 | 8349,42 | 10093,28 | 44420,11 | 127188,44 | 19383,84 | 157539,92 |
| | Διάμεσος | 28228,60 | 24706,21 | 15596,00 | 0,00 | 1012,30 | 8200,00 | 75000,00 | 0,00 | 90000,00 |
| | Διασπορά | $1{,}5\cdot10^{9}$ | $1{,}06\cdot10^{9}$ | $6{,}1\cdot10^{8}$ | $4{,}4\cdot10^{8}$ | $5{,}2\cdot10^{8}$ | $2{,}1\cdot10^{10}$ | $7{,}7\cdot10^{10}$ | $6{,}7\cdot10^{9}$ | $9{,}2\cdot10^{10}$ |
| | $Q_1$ | 15517,40 | 14836,89 | 6720,00 | 0,00 | 40,00 | 800,00 | 0,00 | 0,00 | 3000,00 |
| | $Q_3$ | 47511,10 | 40645,23 | 27160,00 | 3500,00 | 11955,39 | 33410,80 | 180000,00 | 0,00 | 205000,00 |
| **Βέλγιο** | Μέση Τιμή | 38646,85 | 28718,13 | 13898,81 | 14384,10 | 10363,92 | 110355,18 | 204899,97 | 42944,62 | 264253,09 |
| | Διάμεσος | 24704,10 | 20366,44 | 12000,00 | 0,00 | 1017,60 | 26141,45 | 150000,00 | 0,00 | 171382,60 |
| | Διασπορά | $2{,}9\cdot10^{9}$ | $1{,}3\cdot10^{9}$ | $2{,}3\cdot10^{8}$ | $1{,}2\cdot10^{9}$ | $1{,}5\cdot10^{9}$ | $7{,}02\cdot10^{10}$ | $2{,}7\cdot10^{11}$ | $6{,}9\cdot10^{10}$ | $3{,}7\cdot10^{11}$ |
| | $Q_1$ | 14834,25 | 13055,16 | 459,45 | 0,00 | 124,77 | 4251,42 | 75000,00 | 0,00 | 92478,90 |
| | $Q_3$ | 43801,75 | 31791,80 | 19418,09 | 16000,00 | 7433,20 | 105158,70 | 247893,50 | 0,00 | 264004,95 |
| **Γαλλία** | Μέση Τιμή | 59738,02 | 51891,97 | 17798,51 | 21327,58 | 20611,91 | 41195,70 | 192168,10 | 68466,43 | 279836,58 |
| | Διάμεσος | 31843,00 | 27404,49 | 13111,20 | 0,00 | 350,00 | 16434,20 | 150000,00 | 0,0000 | 177393,90 |
| | Διασπορά | $8{,}8\cdot10^{10}$ | $7{,}3\cdot10^{10}$ | $4{,}4\cdot10^{8}$ | $1{,}9\cdot10^{9}$ | $8{,}5\cdot10^{10}$ | $3{,}9\cdot10^{9}$ | $9{,}1\cdot10^{10}$ | $1{,}2\cdot10^{11}$ | $3{,}6\cdot10^{11}$ |
| | $Q_1$ | 20177,00 | 17844,47 | 0,00 | 0,00 | 66,20 | 3058,80 | 72593,55 | 0,00 | 80601,10 |
| | $Q_3$ | 52561,65 | 44597,78 | 25543,79 | 29300,00 | 2402,55 | 52667,15 | 228673,50 | 22867,40 | 300000,00 |
| **Γερμανία** | Μέση Τιμή | 50050,99 | 39214,19 | 13611,67 | 24081,27 | 12358,04 | 59409,11 | 132848,15 | 33345,39 | 170555,39 |
| | Διάμεσος | 32738,00 | 27965,79 | 11300,00 | 0,00 | 1444,80 | 25000,00 | 100000,00 | 0,00 | 100500,00 |
| | Διασπορά | $2{,}8\cdot10^{9}$ | $1{,}3\cdot10^{9}$ | $2{,}3\cdot10^{8}$ | $2{,}3\cdot10^{9}$ | $7{,}7\cdot10^{8}$ | $2{,}1\cdot10^{10}$ | $5{,}5\cdot10^{10}$ | $2{,}9\cdot10^{10}$ | $1{,}2\cdot10^{11}$ |
| | $Q_1$ | 19362,80 | 18096,83 | 0,00 | 0,00 | 173,80 | 5000,00 | 0,00 | 0,00 | 4642,50 |
| | $Q_3$ | 63600,00 | 47961,27 | 21192,00 | 36000,00 | 10924,29 | 72333,60 | 200000,00 | 0,00 | 216679,00 |
| **Δανία** | Μέση Τιμή | 58639,91 | 39003,01 | 16532,52 | 34780,30 | 7327,08 | 120100,01 | 189067,03 | 30797,68 | 347579,79 |
| | Διάμεσος | 47105,44 | 32825,18 | 14518,73 | 15056,19 | 1037,80 | 36296,17 | 134430,28 | 0,00 | 111577,13 |
| | Διασπορά | $2{,}4\cdot10^{9}$ | $9{,}1\cdot10^{8}$ | $3{,}4\cdot10^{8}$ | $2{,}4\cdot10^{9}$ | $3{,}5\cdot10^{8}$ | $3{,}5\cdot10^{11}$ | $1{,}9\cdot10^{11}$ | $2{,}8\cdot10^{10}$ | $1{,}2\cdot10^{12}$ |
| | $Q_1$ | 24718,23 | 18300,05 | 0,00 | 0,00 | 188,20 | 5205,67 | 14787,33 | 0,00 | 26886,05 |
| | $Q_3$ | 78173,49 | 49999,47 | 24196,77 | 63182,23 | 4129,02 | 107544,22 | 201645,42 | 0,00 | 215088,45 |
| **Ελβετία** | Μέση Τιμή | 68137,60 | 59195,30 | 20607,63 | 27978,61 | 19551,35 | 206139,00 | 349296,37 | 77402,03 | 395114,01 |

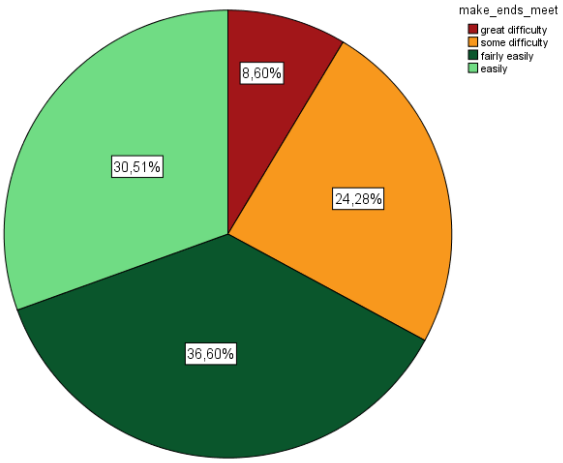
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Διάμεσος | 49853,34 | 45074,21 | 15909,92 | 1766,39 | 1877,85 | 65876,02 | 228131,92 | 0,00 | 157709,55 |
| | Διασπορά | $4{,}3\cdot10^{9}$ | $3{,}1\cdot10^{9}$ | $5{,}8\cdot10^{8}$ | $2{,}5\cdot10^{9}$ | $1{,}4\cdot10^{9}$ | $1{,}8\cdot10^{11}$ | $6{,}2\cdot10^{11}$ | $6{,}05\cdot10^{10}$ | $1{,}2\cdot10^{12}$ |
| | $Q_1$ | 25350,01 | 23945,31 | 0,00 | 0,00 | 228,13 | 19098,22 | 0,00 | 0,00 | 6375,63 |
| | $Q_3$ | 89570,13 | 74392,93 | 30385,21 | 45626,38 | 22982,66 | 194746,77 | 456263,85 | 0,00 | 394994,13 |
| **Ελλάδα** | Μέση Τιμή | 19156,18 | 17546,97 | 7665,16 | 8864,02 | 2626,99 | 11737,02 | 92044,78 | 54599,77 | 174212,92 |
| | Διάμεσος | 14000,00 | 12902,72 | 5600,00 | 0,00 | 90,99 | 3000,00 | 75000,00 | 0,00 | 100000,00 |
| | Διασπορά | $2{,}8\cdot10^{8}$ | $2{,}2\cdot10^{8}$ | $1{,}1\cdot10^{8}$ | $2{,}1\cdot10^{8}$ | $4{,}3\cdot10^{7}$ | $8{,}03\cdot10^{8}$ | $1{,}97\cdot10^{10}$ | $1{,}7\cdot10^{10}$ | $6{,}6\cdot10^{10}$ |
| | $Q_1$ | 8400,00 | 8080,00 | 0,00 | 0,00 | 0,00 | 0,00 | 39000,00 | 0,00 | 53798,95 |
| | $Q_3$ | 25000,00 | 22400,15 | 10769,50 | 14000,00 | 1265,72 | 11065,57 | 120000,00 | 51000,00 | 190000,00 |
| **Ισπανία** | Μέση Τιμή | 64957,63 | 58439,49 | 8488,98 | 10634,08 | 45834,56 | 26957,65 | 202836,00 | 48008,62 | 275465,44 |
| | Διάμεσος | 15480,95 | 14257,52 | 5600,00 | 0,00 | 141,40 | 2713,40 | 116336,00 | 0,00 | 131909,50 |
| | Διασπορά | $1{,}99\cdot10^{11}$ | $1{,}6\cdot10^{11}$ | $2{,}3\cdot10^{8}$ | $1{,}4\cdot10^{9}$ | $1{,}9\cdot10^{11}$ | $3{,}3\cdot10^{10}$ | $6.1\cdot10^{11}$ | $4{,}8\cdot10^{10}$ | $7{,}5\cdot10^{11}$ |
| | $Q_1$ | 9053,02 | 8722,77 | 0,00 | 0,00 | 0,00 | 567,72 | 60000,00 | 0,00 | 68625,00 |
| | $Q_3$ | 28000,00 | 23920,93 | 11200,00 | 10371,80 | 8219,94 | 13730,70 | 180000,00 | 6000,00 | 220250,00 |
| **Ιταλία** | Μέση Τιμή | 31114,76 | 25785,92 | 12610,28 | 8067,99 | 10436,47 | 21556,10 | 217663,65 | 47136,17 | 301534,08 |
| | Διάμεσος | 22800,00 | 18877,19 | 10920,00 | 0,00 | 751,29 | 6000,00 | 130000,00 | 0,00 | 151500,00 |
| | Διασπορά | $1{,}2\cdot10^{9}$ | $7{,}7\cdot10^{8}$ | $1{,}9\cdot10^{8}$ | $5{,}01\cdot10^{8}$ | $4{,}9\cdot10^{8}$ | $1{,}7\cdot10^{9}$ | $4{,}3\cdot10^{11}$ | $6{,}6\cdot10^{10}$ | $1{,}08\cdot10^{12}$ |
| | $Q_1$ | 13147,05 | 12538,79 | 0,00 | 0,00 | 0,00 | 0,00 | 50000,00 | 0,00 | 65432,25 |
| | $Q_3$ | 35660,60 | 29159,36 | 18200,00 | 8000,00 | 13810,50 | 23852,20 | 250000,00 | 0,00 | 276617,10 |
| **Ολλανδία** | Μέση Τιμή | 47873,22 | 40059,20 | 17773,53 | 19628,92 | 10470,76 | 73541,16 | 203088,64 | 12700,30 | 236322,52 |
| | Διάμεσος | 37574,05 | 31103,29 | 14311,79 | 0,00 | 830,00 | 20000,00 | 166386,50 | 0,00 | 118115,00 |
| | Διασπορά | $1{,}8\cdot10^{9}$ | $1{,}2\cdot10^{9}$ | $4{,}05\cdot10^{8}$ | $1{,}01\cdot10^{9}$ | $8{,}7\cdot10^{8}$ | $1{,}8\cdot10^{10}$ | $1{,}2\cdot10^{11}$ | $5{,}06\cdot10^{9}$ | $1{,}03\cdot10^{12}$ |
| | $Q_1$ | 20851,40 | 19768,00 | 0,00 | 0,00 | 139,55 | 3663,60 | 0,00 | 0,00 | 5000,00 |
| | $Q_3$ | 61181,80 | 48656,14 | 25431,59 | 32725,69 | 6255,40 | 77500,00 | 300000,00 | 0,00 | 256500,00 |
| **Σουηδία** | Μέση Τιμή | 51022,40 | 37207,16 | 18631,94 | 22448,00 | 9942,45 | 78425,29 | 127003,09 | 43452,20 | 341561,87 |
| | Διάμεσος | 42654,27 | 31642,32 | 15032,40 | 9803,60 | 740,71 | 35787,61 | 78973,45 | 0,00 | 95312,78 |
| | Διασπορά | $1{,}2\cdot10^{9}$ | $6{,}3\cdot10^{8}$ | $4{,}7\cdot10^{8}$ | $8{,}9\cdot10^{8}$ | $4{,}8\cdot10^{8}$ | $3{,}5\cdot10^{10}$ | $6{,}6\cdot10^{10}$ | $3{,}4\cdot10^{10}$ | $1{,}8\cdot10^{12}$ |
| | $Q_1$ | 27132,93 | 20290,02 | 0,00 | 0,00 | 112,41 | 10892,89 | 10892,89 | 0,00 | 32678,67 |
| | $Q_3$ | 64572,72 | 46253,90 | 27449,42 | 37796,36 | 5882,15 | 83590,40 | 163393,35 | 32678,67 | 204931,40 |

Από τον παραπάνω πίνακα ενδεικτικά μπορούμε να αναφέρουμε ότι το υψηλότερο μέσο ακαθάριστο οικογενειακό εισόδημα παρατηρείται στην Ελβετία (68137,60€), η μικρότερη διάμεσος καθαρού οικογενειακού εισοδήματος ανήκει στην Ελλάδα (12902,72€), το υψηλότερο μέσο οικογενειακό εισόδημα από σύνταξη γήρατος είναι της Αυστρίας (20758,34€), η μεγαλύτερη διάμεσος οικογενειακού εισοδήματος από εργασία είναι της Δανίας (15056,19€) και ελάχιστη τιμή μέσου εισοδήματος από άλλες πηγές εντοπίζεται στην Ελλάδα (2626,99€). Επίσης, η Ισπανία έχει την μικρότερη διάμεσο συνολικής αξίας των χρηματικών περιουσιακών στοιχείων του νοικοκυριού (2713,40€) ενώ η Ελβετία έχει τη μέγιστη συνολική αξία της ιδιόκτητης κατοικίας του νοικοκυριού (349296,37€), τη μέγιστη συνολική αξία της άλλης ακίνητης περιουσίας του νοικοκυριού (77402,03€) και τη μέγιστη συνολική αξία των παγίων περιουσιακών στοιχείων του νοικοκυριού (395114,01€).

Στη συνέχεια το δείγμα της έρευνας ρωτήθηκε για την «οικονομική του δυνατότητα» δηλαδή για την δυνατότητα που έχει το νοικοκυριό να «τα βγάζει πέρα» οικονομικώς. Η μεταβλητή που χρησιμοποιήθηκε έχει κωδικοποιηθεί ως «make ends meet» και απεικονίζεται γραφικά με το ακόλουθο Pie chart:

**ΣΧΗΜΑ 4.3.3**

Pie-chart συχνοτήτων για τη μεταβλητή «οικονομική δυνατότητα νοικοκυριού»




Από το γράφημα είναι φανερό πως το μεγαλύτερο ποσοστό του δείγματος ισχυρίζεται πως «τα βγάζει πέρα» σχετικά εύκολα (36,6%), ενώ ακολουθεί ένα ισχυρό ποσοστό του δείγματος το οποίο θεωρεί πως «τα βγάζει πέρα» εύκολα (30,51%). Στη συνέχεια ένα λίγο

μικρότερο ποσοστό πιστεύει πως «τα βγάζει πέρα» σχετικά δύσκολα (24,28%) και το ελάχιστο ποσοστό του δείγματος θεωρεί πως «τα βγάζει πέρα» πολύ δύσκολα (8,60%).

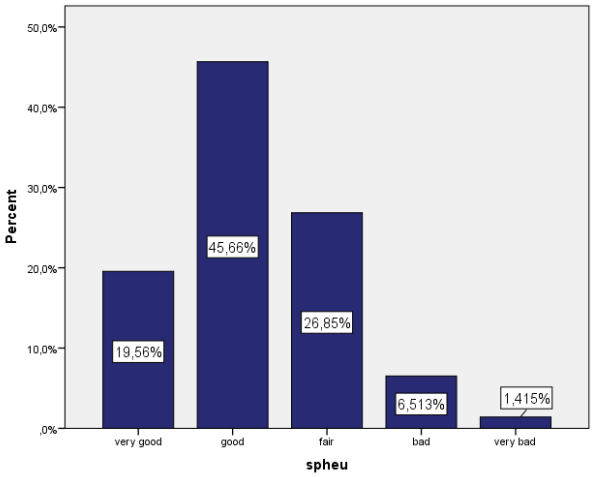
Εκτός από τις ερωτήσεις που αφορούν στην οικονομική κατάσταση του εκάστοτε νοικοκυριού, συλλέχθηκαν για το δείγμα της έρευνας κι ένα σύνολο μετρήσεων που αφορούν στην ατομική ψυχολογική κατάσταση. Πιο αναλυτικά οι μεταβλητές που χρησιμοποιήθηκαν είναι οι ακόλουθες:

- Κατάσταση σωματικής υγείας (Spheu)
- Κατάσταση ψυχικής υγείας (eurodcat)
- Ικανοποίηση από ζωή (q1)

Σύμφωνα με το γράφημα που ακολουθεί, παρατηρήθηκε ότι το μεγαλύτερο μέρος του δείγματος έχει καλή κατάσταση σωματικής υγείας:

**ΣΧΗΜΑ 4.3.4**

Ραβδόγραμμα συχνοτήτων για τη μεταβλητή «κατάσταση σωματικής υγείας»

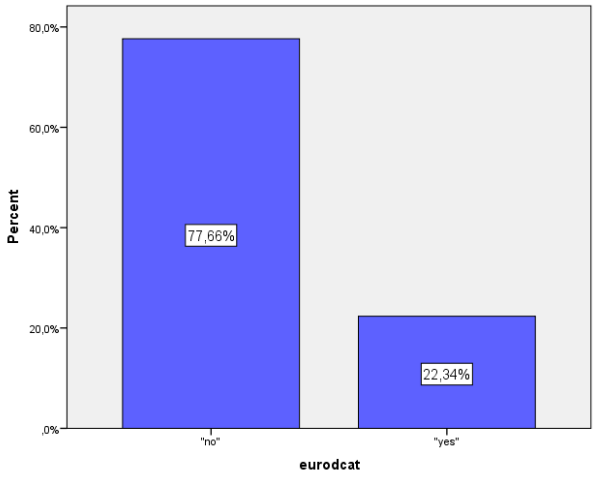



Πιο αναλυτικά, από το Σχήμα 4.3.4 είναι φανερό πως σε ποσοστό μόλις 1,415% των παρατηρήσεων σημειώνεται πολύ άσχημη κατάσταση υγείας ενώ αντίθετα το 19,56% των παρατηρήσεων έχουν πολύ καλή κατάσταση υγείας. Όπως αναφέρθηκε, οι περισσότερες παρατηρήσεις παρουσιάζουν καλή κατάσταση υγείας (45,66%), ενώ ακολουθεί ένα ποσοστό της τάξεως του 26,85% με μέτρια κατάσταση υγείας και ένα της τάξεως του 6,513% με άσχημη κατάσταση υγείας.

Με τον ίδιο τρόπο υπολογίστηκε η ψυχική κατάσταση υγείας των ατόμων του δείγματος. Πιο αναλυτικά, έχει δημιουργηθεί η ψευδομεταβλητή *eurodcat*, η οποία λαμβάνει την τιμή 1 εάν το άτομο δεν αντιμετωπίζει προβλήματα σε τρεις ή περισσότερες διαστάσεις που αφορούν την ψυχική υγεία, ή την τιμή 2 διαφορετικά.

**ΣΧΗΜΑ 4.3.5**

Ραβδόγραμμα συχνοτήτων για τη μεταβλητή «κατάσταση ψυχικής υγείας»



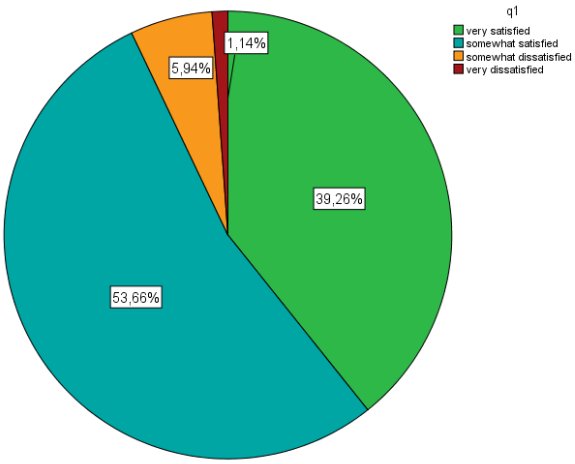

Από το παραπάνω σχήμα είναι φανερό πως το μεγαλύτερο ποσοστό του δείγματος (77,66%) δεν αντιμετωπίζει προβλήματα που αφορούν την ψυχική υγεία, σε τρεις ή περισσότερες διαστάσεις και κατ' επέκταση το ποσοστό αυτό μπορεί να χαρακτηριστεί πως παρουσιάζει καλή ψυχική υγεία. Αντίθετα, υπάρχει ένα ποσοστό των παρατηρήσεων της τάξεως του 22,34% το οποίο φαίνεται να παρουσιάζει αρκετά συμπτώματα κακής ψυχικής υγείας.

Ολοκληρώνοντας την περιγραφική ανάλυση, σημειώθηκαν οι μετρήσεις που έλαβε η μεταβλητή που σχετίζεται με την ικανοποίηση που κάθε άτομο λαμβάνει από τη ζωή. Η μεταβλητή αυτή κωδικοποιήθηκε με την ονομασία «q1» και κυμαίνεται από 1 που δείχνει πλήρη ικανοποίηση από τη ζωή, έως 4 που δείχνει πλήρη δυσαρέσκεια από τη ζωή. Τα σχετικά αποτελέσματα απεικονίζονται στο ακόλουθο pie-chart, σύμφωνα με το οποίο τα περισσότερα άτομα δηλώνουν σχετικά ικανοποιημένοι από τη ζωή (53,66%). Επίσης, το 39,26% των παρατηρήσεων αισθάνονται πολύ ικανοποιημένοι από τη ζωή, το 5,94% των

ατόμων νιώθουν σχετικά δυσαρεστημένοι από τη ζωή, ενώ τέλος μόλις 1,14% του δείγματος αισθάνονται πολύ δυσαρεστημένοι από τη ζωή.

**ΣΧΗΜΑ 4.3.6**

Pie-chart συχνοτήτων για τη μεταβλητή «ικανοποίηση από ζωή»




## 4.4 Συνάφεια μεταβλητών με τον παράγοντα χώρα

Στην ενότητα αυτή η έρευνα θα επικεντρωθεί στην μελέτη της ύπαρξης ή μη, σχέσης μεταξύ ορισμένων μεταβλητών και του παράγοντα χώρα. Στόχος της ενότητας είναι να διαπιστωθεί κατά πόσο βασικά χαρακτηριστικά που ορίζουν την συμπεριφορά των ανθρώπων σχετίζονται με την χώρα στην οποία αυτοί διαμένουν. Η παραπάνω σχέση, μεταξύ δύο ποιοτικών μεταβλητών, ονομάζεται συνάφεια (association ή contiguity). Πιο αναλυτικά, αντικείμενο της ενότητας αυτής θα είναι η ταυτόχρονη περιγραφή δύο μεταβλητών, η σύγκρισή τους καθώς και ο έλεγχος ύπαρξης εξάρτησης μεταξύ τους.

Αρχικά θα εξετασθεί εάν υπάρχει σχέση ανάμεσα σε βασικά χαρακτηριστικά που συνθέτουν την ψυχολογία ενός ανθρώπου και στην χώρα στην οποία διαμένει. Όπως έχει ήδη αναφερθεί, η μεταβλητή «*eurodcat*» που έχει κατασκευαστεί είναι μια ψευδομεταβλητή η οποία λαμβάνει την τιμή 1 εάν το άτομο δεν αντιμετωπίζει προβλήματα σε τρεις ή περισσότερες διαστάσεις που αφορούν την ψυχική υγεία, ή την τιμή 2 διαφορετικά. Θα

εξετάσουμε λοιπόν εάν η κατάσταση της ψυχικής υγείας του ατόμου εξαρτάται από τον παράγοντα χώρα. Για τον λόγο αυτό, θα γίνει ένας έλεγχος ανεξαρτησίας $\chi^2$ (chi-square test). Ο έλεγχος αυτός εξετάζει την μηδενική υπόθεση $H_0$: οι μεταβλητές είναι ανεξάρτητες, έναντι της $H_1$: οι μεταβλητές είναι εξαρτημένες.

**ΠΙΝΑΚΑΣ 4.4.1**

Συσχέτιση χώρας- κατάσταση ψυχικής υγείας

| | | eurodcat | | |
|---|---|---|---|---|
| | | Δεν υπάρχουν συμπτώματα κακής ψυχικής υγείας | Υπάρχουν συμπτώματα κακής ψυχικής υγείας | Σύνολο |
| Χώρα | Αυστρία | 1280 | 307 | 1587 |
| | Βέλγιο | 1878 | 562 | 2440 |
| | Γαλλία | 664 | 269 | 933 |
| | Γερμανία | 1441 | 338 | 1779 |
| | Δανία | 943 | 164 | 1107 |
| | Ελβετία | 539 | 110 | 649 |
| | Ελλάδα | 1226 | 376 | 1602 |
| | Ισπανία | 760 | 410 | 1170 |
| | Ιταλία | 971 | 422 | 1393 |
| | Ολλανδία | 1532 | 368 | 1900 |
| | Σουηδία | 1609 | 368 | 1977 |
| Σύνολο | | 12843 | 3694 | 16537 |

Στον Πίνακα διπλής εισόδου 4.4.1, παρατηρείται ότι σε όλες τις χώρες τα άτομα που δεν παρουσιάζουν συμπτώματα κακής ψυχικής υγείας είναι κατά πολύ περισσότερα από τα άτομα που παρουσιάζουν συμπτώματα κακής ψυχικής υγείας, σε σχέση πάντα με το συνολικό πληθυσμό της εκάστοτε χώρας. Οι κάτοικοι της Αυστρίας, της Γερμανίας, της Δανίας, της Ελβετίας, της Ολλανδίας και της Σουηδίας φαίνεται να παρουσιάζουν πολύ καλή ψυχική υγεία. Στις χώρες αυτές τα άτομα που δεν παρουσιάζουν συμπτώματα κακής ψυχικής υγείας υπερέχουν σε μεγάλο βαθμό σε σχέση με αυτά που παρουσιάζουν ξεπερνώντας το 80% του συνολικού πληθυσμού της χώρας. Αντίθετα οι κάτοικοι του Βελγίου, της Γαλλίας, της Ελλάδας, της Ισπανίας και της Ιταλίας φαίνεται ως επί το πλείστον να μην παρουσιάζουν επαρκή συμπτώματα κακής ψυχικής υγείας, όμως η αναλογία των ατόμων αυτών σε σχέση με όσους παρουσιάζουν κακή ψυχική υγεία φαίνεται να είναι μικρότερη, συγκριτικά με τις προηγούμενες χώρες. Πιο συγκεκριμένα, στο Βέλγιο και στην Ελλάδα, περίπου το 77% του πληθυσμού δεν παρουσιάζει συμπτώματα κακής ψυχικής υγείας. Στην Ιταλία και στην

Γαλλία το ποσοστό αυτό ανέρχεται περίπου στο 70% ενώ τέλος στην Ισπανία, μόλις το 65% των ατόμων δεν παρουσιάζουν κακή ψυχική υγεία.

Τα αποτελέσματα του ελέγχου ανεξαρτησίας $\chi^2$ παρουσιάζονται στον πίνακα 4.4.2:

**ΠΙΝΑΚΑΣ 4.4.2**

Έλεγχος $\chi^2$, για τη σχέση χώρας-κατάσταση ψυχικής υγείας

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 276,301 | 10 | 0,000 |

Από τον παραπάνω πίνακα, είναι φανερό πως η στατιστική συνάρτηση έχει τιμή 286,301 και το p-value του ελέγχου είναι περίπου ίσο με 0. Εφόσον το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%. Αυτό σημαίνει ότι υπάρχει εξάρτηση μεταξύ των δύο μεταβλητών. Επομένως, με βάσει το υπό μελέτη δείγμα, μπορούμε να πούμε πως η κατάσταση της ψυχικής υγείας του ατόμου εξαρτάται από τη χώρα στην οποία κατοικεί.

Στο σημείο αυτό θα εξετασθεί εάν η σωματική υγεία του ατόμου εξαρτάται από την χώρα στην οποία διαμένει. Όμοια με προηγουμένως, ο έλεγχος που εφαρμόζεται ορίζεται από την μηδενική υπόθεση $H_0$: οι μεταβλητές είναι ανεξάρτητες, έναντι της $H_1$: διαφορετικά.

**ΠΙΝΑΚΑΣ 4.4.3**

Σχέση χώρας- σωματικής υγείας ατόμου

| | | Κατάσταση σωματικής υγείας | | | | | |
|---|---|---|---|---|---|---|---|
| | | Πολύ καλή | Καλή | Μέτρια | Κακή | Πολύ κακή | Σύνολο |
| Χώρα | Αυστρία | 279 | 720 | 451 | 109 | 28 | 1587 |
| | Βέλγιο | 504 | 1223 | 557 | 127 | 29 | 2440 |
| | Γαλλία | 158 | 518 | 207 | 39 | 11 | 933 |
| | Γερμανία | 213 | 779 | 558 | 189 | 40 | 1779 |
| | Δανία | 281 | 511 | 240 | 51 | 24 | 1107 |
| | Ελβετία | 235 | 295 | 102 | 16 | 1 | 649 |
| | Ελλάδα | 374 | 699 | 441 | 78 | 10 | 1602 |
| | Ισπανία | 119 | 529 | 378 | 114 | 30 | 1170 |
| | Ιταλία | 123 | 609 | 524 | 122 | 15 | 1393 |
| | Ολλανδία | 362 | 962 | 470 | 92 | 14 | 1900 |
| | Σουηδία | 587 | 706 | 512 | 140 | 32 | 1977 |
| Σύνολο | | 3235 | 7551 | 4440 | 1077 | 234 | 16537 |

Από τον παραπάνω πίνακα φαίνεται ότι σε όλες τις χώρες το μεγαλύτερο μέρος του πληθυσμού δήλωσε πως έχει «καλή» σωματική υγεία. Παρατηρώντας όμως τις απαντήσεις «κακή» και «πολύ κακή» υγεία, μπορούμε να διαπιστώσουμε ότι η Γερμανία και η Ισπανία συγκεντρώνει ένα μεγάλο αριθμό πληθυσμού σε αυτές τις κατηγορίες που συνολικά ξεπερνάει το 12% σε κάθε χώρα. Αντίθετα στην Ελβετία, μόλις 17 άτομα από τα 649 της έρευνας, απάντησαν πως έχουν κακή ή πολύ κακή υγεία. Το ποσοστό τους είναι το μικρότερο που σημειώνεται στο δείγμα και φθάνει το 2,6%. Τέλος, αξίζει να αναφερθεί ότι στην Ελλάδα ρωτήθηκαν 1602 άτομα, εκ των οποίων οι 88 δήλωσαν πως έχουν άσχημη υγεία. Το ποσοστό αυτό είναι σχετικά μικρό και αντιστοιχεί στο 5,5% του δείγματος.

**ΠΙΝΑΚΑΣ 4.4.4**

Έλεγχος $\chi^2$, για τη σχέση χώρας-κατάσταση υγείας

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 842,543 | 40 | 0,000 |

Η στατιστική συνάρτηση του ελέγχου $\chi^2$ έχει τιμή 842,543 ενώ το p-value του ελέγχου ισούται περίπου με 0. Επειδή το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική υπόθεση. Επομένως τα δεδομένα φανερώνουν ισχυρή ένδειξη εξάρτησης των ατόμων που αναλογούν σε κάθε χώρα και της κατάστασης της υγείας τους.

Επιπλέον θα μελετηθεί η σχέση μεταξύ της ικανοποίησης που νιώθει ένα άτομο για τη ζωή και της χώρας στην οποία κατοικεί. Πιο αναλυτικά, από τον παρακάτω πίνακα φαίνεται ότι σε όλες τις χώρες του δείγματος, το μικρότερο ποσοστό των κατοίκων δήλωσε «πολύ δυσαρεστημένος» από τη ζωή. Αντίθετα, στην Δανία, στην Ελβετία και στην Ολλανδία, οι περισσότεροι κάτοικοι δήλωσαν «πολύ ικανοποιημένοι» από τη ζωή. Συγκεκριμένα στην Ολλανδία, η απάντηση αυτή δόθηκε από το 60% του πληθυσμού. Στο Βέλγιο, στη Γαλλία, στη Γερμανία, στην Ελλάδα, στην Ισπανία και στην Σουηδία, οι περισσότεροι κάτοικοι δήλωσαν «σχετικά ικανοποιημένοι» από τη ζωή. Αξίζει να σημειωθεί ότι στην Γερμανία, την Ιταλία και τη Σουηδία η απάντηση αυτή δόθηκε σε ποσοστό μεγαλύτερο του 60% του συνολικού πληθυσμού ενώ στο Βέλγιο και τη Γαλλία, το ποσοστό αυτό αγγίζει το 73%. Τέλος, στην Αυστρία 100 άτομα δήλωσαν «σχετικά δυσαρεστημένοι» από τη ζωή.

**ΠΙΝΑΚΑΣ 4.4.5**

Σχέση χώρας-ικανοποίηση από ζωή

| | | Ικανοποίηση από ζωή | | | | |
|---|---|---|---|---|---|---|
| | | Πολύ ικανοποιημένος | Σχετικά ικανοποιημένος | Σχετικά δυσαρεστημένος | Πολύ δυσαρεστημένος | Σύνολο |
| Χώρα | Αυστρία | 518 | 957 | 100 | 12 | 1587 |
| | Βέλγιο | 963 | 1357 | 98 | 22 | 2440 |
| | Γαλλία | 153 | 682 | 87 | 11 | 933 |
| | Γερμανία | 546 | 1055 | 154 | 24 | 1779 |
| | Δανία | 756 | 319 | 29 | 3 | 1107 |
| | Ελβετία | 315 | 305 | 26 | 3 | 649 |
| | Ελλάδα | 608 | 826 | 131 | 37 | 1602 |
| | Ισπανία | 520 | 530 | 102 | 18 | 1170 |
| | Ιταλία | 246 | 956 | 158 | 33 | 1393 |
| | Ολλανδία | 1136 | 711 | 40 | 13 | 1900 |
| | Σουηδία | 732 | 1175 | 58 | 12 | 1977 |
| Σύνολο | | 6493 | 8873 | 983 | 188 | 16537 |

Ακολουθεί ο έλεγχος $\chi^2$ για την σχέση της μεταβλητής «χώρα» με την «ικανοποίηση από τη ζωή»:

**ΠΙΝΑΚΑΣ 4.4.6**

Έλεγχος $\chi^2$, για τη σχέση χώρας-ικανοποίηση από ζωή

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 1528,050 | 30 | 0,000 |

Η στατιστική συνάρτηση του ελέγχου $\chi^2$ έχει τιμή 1528,050 ενώ το p-value του ελέγχου ισούται περίπου με 0. Επειδή το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική υπόθεση. Επομένως τα δεδομένα φανερώνουν εξάρτηση των ατόμων που αναλογούν σε κάθε χώρα και της ικανοποίησης που νιώθουν για τη ζωή.

Σε αυτό το σημείο θα εξετασθεί εάν το χαρακτηριστικό που δείχνει κατά πόσο ο εκπρόσωπος του νοικοκυριού δήλωσε πως έχει οικονομική δυνατότητα (τα βγάζει πέρα οικονομικά) εξαρτάται από την χώρα στην οποία κατοικεί.

**ΠΙΝΑΚΑΣ 4.4.7**

Σχέση χώρας-οικονομική δυνατότητα

| | | Οικονομική δυνατότητα | | | | |
|---|---|---|---|---|---|---|
| | | Μεγάλη οικονομική δυσκολία | Σχετική οικονομική δυσκολία | Σχετική οικονομική ευκολία | Οικονομική ευκολία | Σύνολο |
| Χώρα | Αυστρία | 80 | 334 | 785 | 388 | 1587 |
| | Βέλγιο | 152 | 504 | 852 | 932 | 2440 |
| | Γαλλία | 68 | 209 | 408 | 248 | 933 |
| | Γερμανία | 113 | 338 | 652 | 676 | 1779 |
| | Δανία | 30 | 162 | 409 | 506 | 1107 |
| | Ελβετία | 18 | 82 | 234 | 315 | 649 |
| | Ελλάδα | 377 | 723 | 294 | 208 | 1602 |
| | Ισπανία | 168 | 469 | 409 | 124 | 1170 |
| | Ιταλία | 263 | 600 | 432 | 98 | 1393 |
| | Ολλανδία | 89 | 277 | 791 | 743 | 1900 |
| | Σουηδία | 64 | 318 | 787 | 808 | 1977 |
| Σύνολο | | 1422 | 4016 | 6053 | 5046 | 16537 |

Σύμφωνα με τον Πίνακα 4.4.7 το μεγαλύτερο μέρος του πληθυσμού του Βελγίου, της Γερμανίας, της Δανίας, της Ελβετίας και της Σουηδίας δήλωσαν πως έχουν οικονομική ευκολία. Αξίζει να σημειωθεί ότι συγκεκριμένα στην Ελβετία, το ποσοστό των ατόμων που δήλωσε οικονομική ευημερία αγγίζει το 48,5% του πληθυσμού της ενώ μόλις 15% του ίδιου πληθυσμού αντιμετωπίζει σχετική ή μεγάλη δυσκολία. Στην Αυστρία, την Γαλλία και την Ολλανδία, το μεγαλύτερο μέρος του δείγματος δήλωσε πως έχει «σχετική οικονομική ευκολία». Τέλος, στην Ελλάδα, την Ισπανία και την Ιταλία, οι περισσότεροι κάτοικοι φαίνεται να παρουσιάζουν «σχετική οικονομική δυσκολία». Συγκεκριμένα στην Ελλάδα, από τα 1602 άτομα του δείγματος, οι 723 δήλωσαν «σχετική οικονομική δυσκολία» και οι 377 δήλωσαν «μεγάλη οικονομική δυσκολία», συνεπώς το 68,6% των Ελλήνων φαίνεται να αντιμετωπίζουν οικονομικά προβλήματα.

**ΠΙΝΑΚΑΣ 4.4.8**

Έλεγχος $\chi^2$, για τη σχέση χώρας-οικονομική αυτοπεποίθηση

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 2913,188 | 30 | 0,000 |

Η στατιστική συνάρτηση του ελέγχου $\chi^2$ έχει τιμή 2913,188 ενώ το p-value του ελέγχου ισούται περίπου με 0. Επειδή το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική

υπόθεση. Επομένως τα δεδομένα φανερώνουν ισχυρή ένδειξη εξάρτησης των ατόμων που αναλογούν σε κάθε χώρα και της οικονομικής αυτοπεποίθησης.

Σύμφωνα με τα παραπάνω, εύκολα μπορεί κανείς να αντιληφθεί πως ο παράγοντας «χώρα» είναι ένας αρκετά ισχυρός παράγοντας που επηρεάζει τόσο την ψυχική όσο και την σωματική υγεία του ανθρώπου. Συνοπτικά αποδείχθηκε με βάση το δείγμα των τριών ευρωπαϊκών χωρών, πως η κατάσταση της ψυχικής υγείας, η κατάσταση της σωματικής υγείας, η ικανοποίηση που αντλεί κανείς από τη ζωή καθώς και η οικονομική αυτοπεποίθηση του ατόμου είναι χαρακτηριστικά που σχετίζονται με την χώρα κατοικίας του.

## 4.5 Εφαρμογή της Πολλαπλής Ανάλυσης Αντιστοιχιών

Στην προηγούμενη ενότητα έγινε φανερό πως στοιχεία που συνθέτουν την ψυχική και σωματική υγεία των ατόμων έχουν άμεση σχέση με την χώρα στην οποία κατοικούν. Στην παρούσα ενότητα θα εφαρμοσθεί Πολλαπλή Ανάλυση Αντιστοιχιών στα δεδομένα του ερωτηματολογίου και συγκεκριμένα θα ελεγχθεί η σχέση που συνδέει τα χαρακτηριστικά:

- Οικονομική αυτοπεποίθηση (make_ends_meet)
- Κατάσταση υγείας (spheu)
- Ικανοποίηση από τη ζωή (q1)
- Ύπαρξη ή μη, συμπτωμάτων κακής ψυχικής υγείας (eurodcat)

με τον παράγοντα χώρα (country).

Αξίζει να σημειωθεί ότι κατά την εκτέλεση της μεθόδου χρειάσθηκαν 34 επαναλήψεις καθώς κατά την 34$^{η}$ επανάληψη, η διαφορά μεταξύ των δύο τελευταίων επαναλήψεων ήταν μικρότερη από το σημείο σύγκλισης που ορίσαμε (τιμή σύγκλισης=0,00001).

Η Ανάλυση Πολλαπλών Αντιστοιχιών μπορεί να ομαδοποιήσει τα δεδομένα ανάλογα με τις διαστάσεις που ορίζονται. Το πλήθος των διαστάσεων που μπορεί να χρησιμοποιηθεί εξαρτάται από τον αριθμό των μεταβλητών και των παρατηρήσεων, ωστόσο είθισται να μην χρησιμοποιούνται παραπάνω από 3 διαστάσεις. Στην περίπτωσή μας χρησιμοποιήθηκαν δύο διαστάσεις για τη λύση όπως φαίνεται στον ακόλουθο πίνακα στην στήλη Dimension.

**ΠΙΝΑΚΑΣ 4.5.1**

Μοντέλο Ανάλυσης Αντιστοιχιών

| Dimension | Variance Accounted for | |
|---|---|---|
| | Total (Eigenvalue) | Inertia |
| 1 | 1,809 | 0,452 |
| 2 | 1,194 | 0,298 |
| Total | 3,002 | 0,751 |
| Mean | 1,501 | 0,375 |

Στην στήλη Eigenvalue του Πίνακα 4.5.1 δίνονται οι ιδιοτιμές. Οι ιδιοτιμές μετρούν το μέγεθος της πληροφορίας που αποδίδεται σε κάθε διάσταση. Η πρώτη διάσταση ερμηνεύει την παρατηρούμενη διακύμανση των τιμών κατά 1,809 ενώ η δεύτερη διάσταση ερμηνεύει την διακύμανση κατά 1,194. Σύμφωνα με την τελευταία στήλη (Inertia) προκύπτει το συμπέρασμα ότι η πρώτη διάσταση ερμηνεύει το 45,2% της διακύμανσης και η δεύτερη το 29,8%. Συνολικά, οι δύο διαστάσεις ερμηνεύουν το 75,1% της διακύμανσης. Το ποσοστό αυτό είναι αρκετό ώστε να συνεχιστεί η μελέτη του μοντέλου χωρίς την προσθήκη τρίτης διάστασης. Αξίζει να σημειωθεί ότι τα ποσοστά αυτά προκύπτουν σε σχέση με τη συνολική διακύμανση του μοντέλου η οποία δεν ερμηνεύεται μόνο από τις δύο πρώτες διαστάσεις. Τέλος, από την τελευταία γραμμή του πίνακα φαίνεται ότι καθεμία εκ των δύο διαστάσεων αναμένεται να ερμηνεύει την παρατηρούμενη διακύμανση των τιμών κατά 1,501 γεγονός που αντιστοιχεί στο 37,5% της διακύμανσης.

Στη συνέχεια θα μελετηθεί η «διακριτική ικανότητα» των μεταβλητών. Ως μέτρο διάκρισης (Discrimination Measure) προσδιορίζεται ποσοτικά το τετράγωνο της συσχέτισης μεταξύ της μεταβλητής και του μετασχηματισμού της και ερμηνεύεται ως η τετραγωνική επιβάρυνση μιας συνιστώσας. Η μέγιστη τιμή της είναι η μονάδα, η οποία επιτυγχάνεται εάν δεν υπάρχει καμία διαφορά μεταξύ του μετασχηματισμού και της αρχικής μεταβλητής. Εάν μία μεταβλητή έχει υψηλές μετρήσεις διάκρισης τότε υποδηλώνεται υψηλός βαθμός διάκρισης ανάμεσα στις κατηγορίες της μεταβλητής και της αντίστοιχης διάστασης. Ο μέσος όρος των μέτρων διάκρισης για κάθε μεταβλητή, ισοδυναμεί με τη διακύμανση που αναμένεται να ερμηνεύει κάθε διάσταση για τη συγκεκριμένη μεταβλητή. Τα παραπάνω γίνονται αντιληπτά με την βοήθεια του ακόλουθου πίνακα:

**ΠΙΝΑΚΑΣ 4.5.2**

Discrimination Measures

| | Dimension | | Mean |
|---|---|---|---|
| | 1 | 2 | |
| make_ends_meet | 0,315 | 0,253 | 0,284 |
| spheu | 0,512 | 0,431 | 0,471 |
| q1 | 0,514 | 0,505 | 0,509 |
| eurodcat | 0,468 | 0,004 | 0,236 |
| Country | 0,079 | 0,024 | 0,052 |
| Active Total | 1,809 | 1,194 | 1,501 |

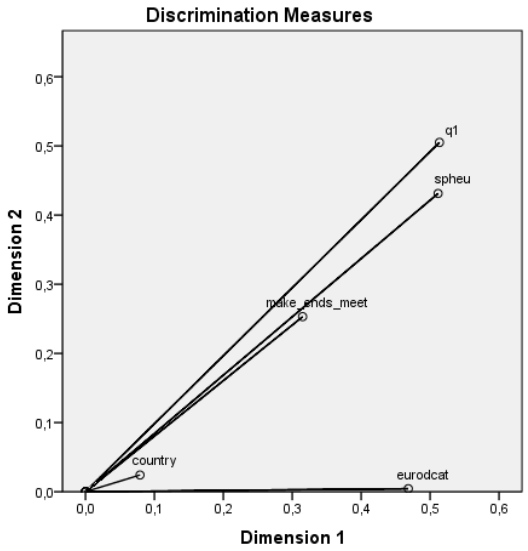
Από τον Πίνακα 4.5.2 είναι φανερό πως όλες οι μεταβλητές διακρίνονται σε μικρότερο ή μεγαλύτερο βαθμό περισσότερο στην πρώτη διάσταση (Dimension 1). Πιο αναλυτικά, η οικονομική δυνατότητα (make_ends_meet) καθώς και η σωματική υγεία (spheu), διακρίνονται περισσότερο στην πρώτη διάσταση σε σχέση με την δεύτερη με τιμές 0,315 (έναντι 0,253) και 0,512 (έναντι 0,431) αντίστοιχα. Η ικανοποίηση από τη ζωή (q1) φαίνεται να τείνει και αυτή προς την πρώτη διάσταση με τιμή 0,514 (έναντι 0,505). Η ύπαρξη ή μη συμπτωμάτων κακής ψυχικής υγείας (eurodcat) διακρίνεται φανερά στην πρώτη διάσταση έχοντας τιμή 0,468 έναντι 0,004 που έχει στη δεύτερη διάσταση. Τέλος, η χώρα (country), η οποία αποτελεί την βοηθητική μεταβλητή, θα μπορούσαμε να πούμε πως τείνει προς την πρώτη διάσταση σημειώνοντας την τιμή 0,079 έναντι 0,024 που έχει στη δεύτερη. Η τελευταία στήλη του πίνακα υπολογίζει τη μέση τιμή των διαστάσεων για κάθε μεταβλητή και ισούται με την αναμενόμενη διακύμανση που ερμηνεύεται για τη συγκεκριμένη διάσταση. Επιπλέον, η τελευταία γραμμή του πίνακα δείχνει την διακύμανση που ερμηνεύει κάθε διάσταση για το σύνολο των μεταβλητών καθώς και τη μέση διακύμανση. Αξίζει να επισημανθεί ότι τα αποτελέσματα αυτά ταυτίζονται με τα αποτελέσματα του Πίνακα 4.5.1 και επιπλέον ότι οι τιμές των μέτρων διάκρισης δεν είναι κατά βάση υψηλές επομένως συμπεραίνουμε ότι δεν θα υφίσταται ικανοποιητικός βαθμός διάκρισης μεταξύ των κατηγοριών των μεταβλητών.

Το γράφημα που ακολουθεί, αναπαριστά τις μετρήσεις διάκρισης. Συγκεκριμένα απεικονίζει στον οριζόντιο άξονα την πρώτη διάσταση και στον κάθετο άξονα τη δεύτερη διάσταση. Κάθε μεταβλητή απεικονίζεται με ένα σημείο στον διδιάστατο χώρο. Η γωνία που σχηματίζει κάθε χαρακτηριστικό με τους άξονες, είναι αυτή που ορίζει και την σχέση του με την αντίστοιχη διάσταση. Στην πραγματικότητα, όσο μικρότερη είναι η γωνία που σχηματίζει

ένα χαρακτηριστικό με την πρώτη διάσταση, τόσο το αντίστοιχο συνημίτονο τείνει στην μονάδα που σημαίνει πως η διάσταση τείνει να ταυτιστεί με την απόσταση του χαρακτηριστικού από την αρχή των αξόνων. Η ίδια ερμηνεία σημειώνεται και για την δεύτερη διάσταση. Παράλληλα η απόσταση κάθε μεταβλητής από την αρχή των αξόνων ερμηνεύει την συσχέτιση της μεταβλητής αυτής με τον εκάστοτε άξονα. Επομένως, η «διακριτική ικανότητα» των μεταβλητών προκύπτει αφενός από το συνημίτονο της γωνίας που σχηματίζει κάθε μεταβλητή με τον άξονα και αφετέρου από την απόσταση της μεταβλητής από την αρχή των αξόνων.

**ΣΧΗΜΑ 4.5.1**

Discrimination Measures




Variable Principal Normalization.

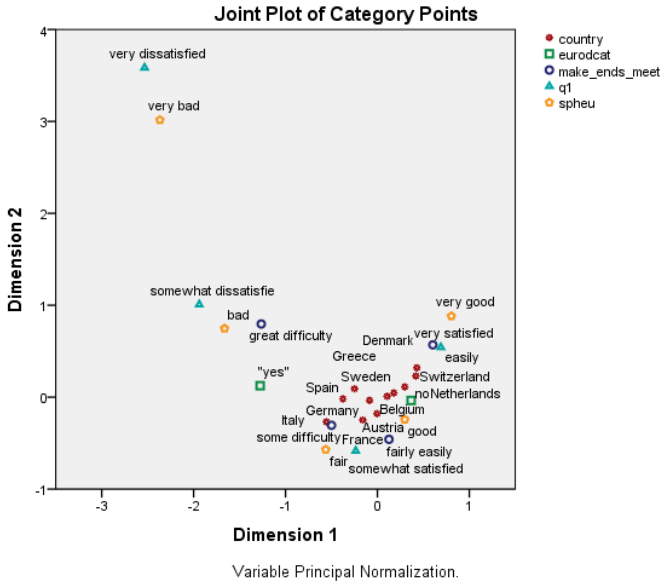
Σύμφωνα με τα παραπάνω, είναι φανερό, πως η μεταβλητή *eurodcat* σχηματίζει την ελάχιστη δυνατή γωνία με την πρώτη διάσταση γεγονός που δείχνει πως το χαρακτηριστικό αυτό ταυτίζεται με τον άξονα. Στη συνέχεια ακολουθεί η χώρα, η οποία είναι η βοηθητική μεταβλητή, και φαίνεται να σχηματίζει μικρή γωνία με τον οριζόντιο άξονα. Τέλος, οι γωνίες που σχηματίζουν οι υπόλοιπες τρεις μεταβλητές με την πρώτη διάσταση δεν είναι αρκετά μικρές, αλλά είναι μικρότερες από τις αντίστοιχες γωνίες που σχηματίζουν με τη δεύτερη. Το γεγονός αυτό μας επιτρέπει να συμπεράνουμε πως οι μετρήσεις των μεταβλητών αυτών είναι μεγαλύτερες στον πρώτο άξονα. Σε συνέχεια των παραπάνω, η μεταβλητή *make_ends_meet* σημειώνει σχετικά μικρή απόσταση από την αρχή των αξόνων και συνεπώς χαρακτηρίζεται από μικρή διακριτική ικανότητα, η *eurodcat* έχει μέτρια απόσταση και επομένως μέτρια

διακριτική ικανότητα και τέλος οι *q1* και *spheu* σημειώνουν σχετικά μεγάλη απόσταση και κατ' επέκταση καλή διακριτική ικανότητα.

Οι παραπάνω μετρήσεις διάκρισης και το σχετικό διάγραμμα φανερώνουν ποιες μεταβλητές διακρίνονται καλύτερα ως προς κάθε διάσταση. Ωστόσο δεν μας πληροφορούν για το πώς κατανέμονται οι κατηγορίες της κάθε μεταβλητής. Το γράφημα που ακολουθεί παρουσιάζει αυτή τη δυνατότητα οπτικοποίησης των μεταβλητών σε σχέση με τις κατηγορίες τους. Επομένως θα γίνει φανερό ποιες κατηγορίες μοιάζουν μεταξύ τους και ποιες διαφέρουν.

**ΣΧΗΜΑ 4.5.2α**

Διάγραμμα Joint



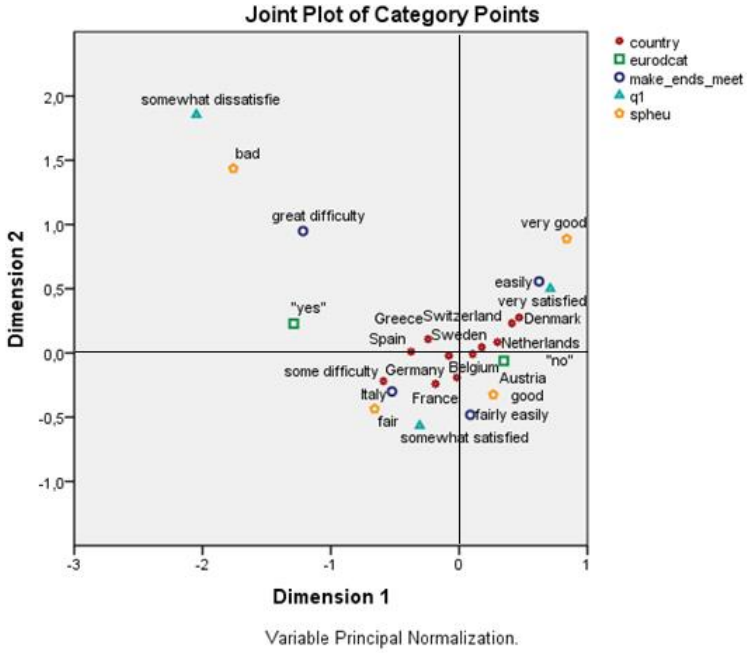

Variable Principal Normalization.

Το διάγραμμα «Joint» κατασκευάσθηκε από δύο άξονες κάθε ένας εκ των οποίων αντιστοιχεί σε μία διάσταση. Μέσα στο διάγραμμα απεικονίζονται οι κατηγορίες των μεταβλητών οι οποίες σημειώνονται με το χρώμα και το σχήμα που έχει ορισθεί για την κάθε μεταβλητή αντίστοιχα. Οι έντεκα χώρες απεικονίζονται με ένα κόκκινο αστερίσκο η καθεμία ώστε να ξεχωρίζουν από τα υπόλοιπα χαρακτηριστικά. Επιπλέον η ύπαρξη ή μη συμπτωμάτων κακής ψυχικής υγείας (eurodcat) απεικονίζεται με πράσινο τετράγωνο, η οικονομική δυνατότητα (make ends meet) σημειώνεται με μπλε κύκλο, η ικανοποίηση από τη ζωή (q1) έχει γαλάζιο τρίγωνο και η σωματική υγεία (spheu) απεικονίζεται με κίτρινο πεντάγωνο. Είναι φανερό πως οι κάτοικοι καμιάς χώρας δεν φαίνεται να χαρακτηρίζονται από πλήρη δυσαρέσκεια (very dissatisfied) από τη ζωή ούτε από πολύ κακή υγεία (very bad). Έτσι κατασκευάστηκε ένα δεύτερο διάγραμμα «Joint» το οποίο εστιάζει στις υπόλοιπες

κατηγορίες των χαρακτηριστικών. Παρατηρώντας το παρακάτω γράφημα, θα έλεγε κανείς πως οι Ευρωπαϊκές χώρες δεν παρουσιάζουν ιδιαίτερα μεγάλες διαφορές μεταξύ τους σε θέματα που αφορούν στον ψυχισμό των κατοίκων. Όπως και στο προηγούμενο διάγραμμα, έτσι και τώρα, υπάρχουν τρεις κατηγορίες μεταβλητών που δεν χαρακτηρίζουν τους κατοίκους καμίας χώρας.

**ΣΧΗΜΑ 4.5.2β**

Διάγραμμα Joint μετά την αφαίρεση δύο κατηγοριών




Variable Principal Normalization.

Για τη διευκόλυνση της μελέτης, κατασκευάστηκε και ένα τρίτο γράφημα, μετά την αφαίρεση των κατηγοριών «somewhat dissatisfied» για την μεταβλητή *q1*, «bad» για τη μεταβλητή *spheu* και «great difficulty» για την *make_ends_meet*. Στόχος του τελικού γραφήματος είναι να αποσυρθούν οι κατηγορίες που δεν φαίνεται να χαρακτηρίζουν τους κατοίκους των χωρών, ώστε τελικά να υπάρξει μια πιο ξεκάθαρη εικόνα.

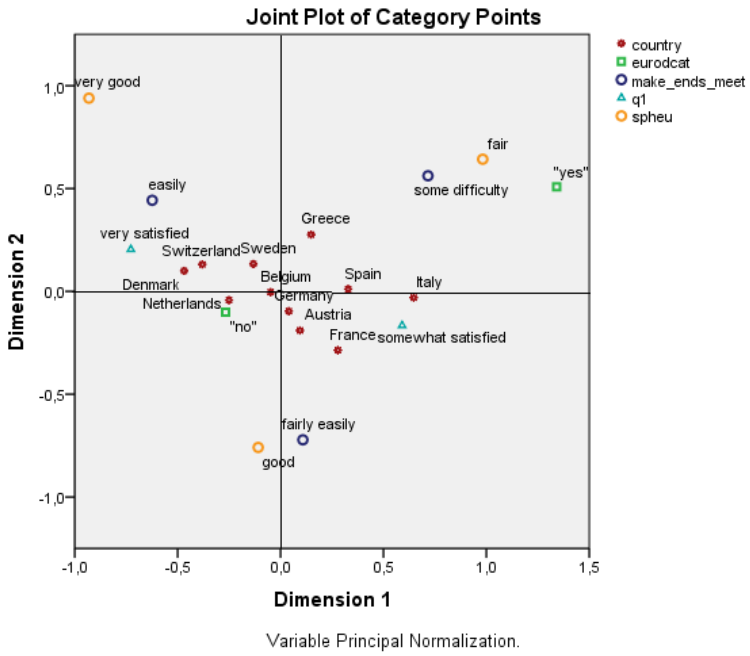
Το διάγραμμα Joint του σχήματος 4.5.2γ αποτελείται από 5 μεταβλητές, εκ των οποίων η μία είναι βοηθητική, οι κατηγορίες των οποίων είναι χωρισμένες ως εξής:

Αναφορικά με την μεταβλητή *eurodcat* οι δύο κατηγορίες της είναι εκ διαμέτρου αντίθετες ως προς την πρώτη διάσταση, γεγονός που υποδεικνύει μια σχετικά μεγάλη τιμή του μέτρου διάκρισης επιβεβαιώνοντας έτσι τόσο τα αποτελέσματα του Πίνακα 4.5.2, όσο και του Σχήματος 4.5.1. Το ίδιο φαινόμενο παρατηρείται επίσης για τις δύο κατηγορίες που

απέμειναν (οι άλλες δύο αφαιρέθηκαν σε προηγούμενο στάδιο) της μεταβλητής *q1*. Στη συνέχεια, ως προς την μεταβλητή *make_ends_meet*, απεικονίζονται οι τρεις από τις τέσσερις κατηγορίες της (η μία αφαιρέθηκε σε προηγούμενο στάδιο) οι οποίες διαφοροποιούνται και ως προς τις δύο διαστάσεις. Ομοίως, οι τρεις κατηγορίες (μετά την αφαίρεση της τέταρτης και της πέμπτης) της μεταβλητής *spheu* είναι διασκορπισμένες στα τρία από τα τέσσερα τεταρτημόρια του γραφήματος παρουσιάζοντας με αυτό τον τρόπο αρκετά μεγάλη διαφοροποίηση. Τέλος, οι έντεκα κατηγορίες της βοηθητικής μεταβλητής *country* είναι χωρισμένες στα τέσσερα τεταρτημόρια και συγκεντρώνονται κυρίως κοντά στην αρχή των αξόνων.

**ΣΧΗΜΑ 4.5.2γ**

Διάγραμμα Joint μετά την αφαίρεση τριών κατηγοριών




Με βάση το γράφημα Joint όπου παρουσιάζονται οι προβολές των διαφορετικών κατηγοριών των πέντε μεταβλητών, αλλά και των πινάκων που προηγήθηκαν, γνωρίζουμε ότι ο πρώτος άξονας (οριζόντιος) ερμηνεύει το 45,2% της συνολικής διακύμανσης των δεδομένων και ο δεύτερος άξονας (κατακόρυφος) ερμηνεύει το 29,8%. Και οι δυο μαζί ερμηνεύουν το 75,1% της συνολικής διακύμανσης.

Ο πρώτος άξονας διαφοροποιεί τους κατοίκους των χωρών που δήλωσαν πολύ ικανοποιημένοι από τη ζωή τους (αριστερά στον πρώτο άξονα), από εκείνους που δήλωσαν

σχετικά ευχαριστημένοι (δεξιά στον πρώτο άξονα). Στην πρώτη κατηγορία οι κάτοικοι των χωρών δεν εμφανίζουν συμπτώματα κακής ψυχικής υγείας, έχουν καλή ή πολύ καλή σωματική υγεία και οικονομικά τα βγάζουν εύκολα πέρα. Στην δεύτερη κατηγορία κάποιοι κάτοικοι των χωρών εμφανίζουν συμπτώματα κακής ψυχικής υγείας, έχουν μέτρια σωματική υγεία και οικονομικά τα βγάζουν πέρα είτε σχετικά εύκολα είτε σχετικά δύσκολα. Σύμφωνα με τον πρώτο άξονα, στην πρώτη κατηγορία ανήκουν η Ελβετία, η Δανία, η Ολλανδία και η Σουηδία ενώ στην δεύτερη κατηγορία ανήκουν η Ελλάδα, η Ιταλία, η Ισπανία και η Γαλλία. Τέλος, το Βέλγιο, η Γερμανία και η Αυστρία βρίσκονται σχεδόν στο σημείο 0 ως προς τον πρώτο άξονα και κατ' επέκταση αποτελούν μια τρίτη ομάδα.

Ο δεύτερος άξονας (κατακόρυφος) διαφοροποιεί τους κατοίκους που έχουν καλή σωματική υγεία, τα βγάζουν πέρα σχετικά εύκολα, είναι σχετικά ικανοποιημένοι από τη ζωή και δεν εμφανίζουν συμπτώματα κακής ψυχικής υγείας από αυτούς που έχουν μέτρια ή πολύ καλή σωματική υγεία, είναι πολύ ικανοποιημένοι από τη ζωή, τα βγάζουν πέρα είτε εύκολα είτε σχετικά δύσκολα και παρουσιάζουν συμπτώματα κακής ψυχικής υγείας. Σύμφωνα με τον δεύτερο άξονα, στην πρώτη κατηγορία εντάσσονται η Γαλλία, η Αυστρία, η Γερμανία και η Ολλανδία ενώ η Δανία, η Ελβετία, η Σουηδία και η Ελλάδα εντάσσονται στην δεύτερη. Το Βέλγιο, η Ισπανία και η Ιταλία βρίσκονται στο σημείο 0 ως προς τον κατακόρυφο άξονα και κατ' επέκταση αποτελούν μια τρίτη ομάδα ως προς αυτόν τον άξονα.

Για να διαπιστωθεί κατά πόσο οι απεικονίσεις των χωρών στο διδιάστατο σύστημα αξόνων αντικατοπτρίζουν την πραγματικότητα πρέπει να υπολογισθεί η «ποιότητα αναπαράστασης» κάθε χώρας. Ένας γενικός τύπος για να ορισθεί η ποιότητα αναπαράστασης μιας χώρας $X$ σε έναν άξονα είναι να υπολογισθεί το κλάσμα του τετραγώνου της απόστασης της απεικόνισης $X'$ πάνω στον άξονα από την αρχή των αξόνων προς το τετράγωνο της απόστασης της $X$ στον διδιάστατο χώρο.

Επομένως, σύμφωνα με το Σχήμα 4.5.3 που ακολουθεί, η ποιότητα αναπαράστασης της χώρας $X$ ως προς τον οριζόντιο άξονα συμβολίζεται ως «sqcorr», παίρνει τιμές από 0 έως 1 και βρίσκεται ως εξής:

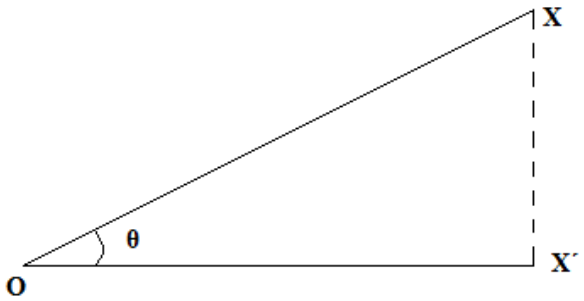
$$sqcorr = \frac{OX'^2}{OX^2} = cos^2\theta$$

όπου,

$\cos\theta$: το συνημίτονο της γωνίας $\theta$.

**ΣΧΗΜΑ 4.5.3**

Ποιότητα αναπαράστασης χώρας *X*




Βάσει των παραπάνω κατασκευάσθηκε ο ακόλουθος πίνακας για τις δύο πρώτες διαστάσεις της ανάλυσης αντιστοιχιών:

**ΠΙΝΑΚΑΣ 4.5.3**

Discrimination Measures για τα επίπεδα της μεταβλητής *country*

| | | overall | | Dimension 1 | | Dimension 2 | |
|---|---|---|---|---|---|---|---|
| Country | mass | quality | %inert | coord | sqcorr | coord | sqcorr |
| Austria | 0,096 | 0,007 | 1,350 | -0,026 | 0,000 | -1,039 | 0,007 |
| Germany | 0,108 | 0,005 | 1,309 | -0,245 | 0,004 | -0,328 | 0,001 |
| Sweden | 0,120 | 0,013 | 1,344 | 0,406 | 0,013 | 0,057 | 0,000 |
| Netherlands | 0,115 | 0,044 | 1,391 | 0,713 | 0,038 | 0,927 | 0,006 |
| Spain | 0,071 | 0,042 | 1,176 | -0,877 | 0,041 | 0,222 | 0,000 |
| Italy | 0,084 | 0,116 | 1,297 | -1,315 | 0,101 | -1,658 | 0,016 |
| France | 0,056 | 0,016 | 1,371 | -0,399 | 0,006 | -1,672 | 0,010 |
| Denmark | 0,067 | 0,062 | 1,465 | 1,030 | 0,043 | 2,173 | 0,019 |
| Greece | 0,097 | 0,021 | 1,243 | -0,529 | 0,020 | 0,512 | 0,002 |
| Switzerland | 0,039 | 0,029 | 1,413 | 1,001 | 0,025 | 1,279 | 0,004 |
| Belgium | 0,148 | 0,006 | 1,305 | 0,249 | 0,006 | -0,037 | 0,000 |

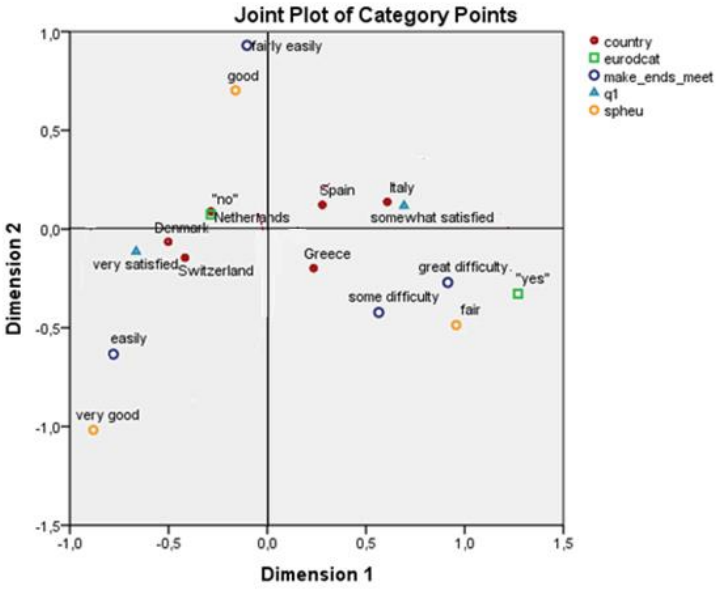
Ο Πίνακας 4.5.3 απεικονίζει, μεταξύ άλλων, την ποιότητα αναπαράστασης κάθε χώρας. Επομένως δίνεται για κάθε χώρα μια τιμή *sqcorr* η οποία δείχνει κατά πόσο ταυτίζεται η κάθε χώρα με κάθε διάσταση του μοντέλου. Με δεδομένο ότι όσο η τιμή *sqcorr* τείνει στην μονάδα τόσο καλύτερη είναι η ποιότητα αναπαράστασης, συμπεραίνουμε πως η ποιότητα αναπαράστασης κάθε χώρας είναι πολύ χαμηλή. Το γεγονός αυτό φανερώνει πως η απεικόνιση των χωρών στον διδιάστατο χώρο δεν είναι αντιπροσωπευτική.

Είναι φανερό πως οι χώρες που παρουσιάζουν την χαμηλότερη ποιότητα αναπαράστασης και για τους δύο άξονες, είναι η Αυστρία, η Γερμανία, η Σουηδία, η Γαλλία και το Βέλγιο. Οι χώρες αυτές έχουν ποιότητα αναπαράστασης μικρότερη από 2% (*overall quality*< 2%) και αποσκοπώντας σε μια ομαδοποίηση που αντικατοπτρίζει την πραγματικότητα, θα μπορούσαν να αφαιρεθούν.

Σύμφωνα με τα παραπάνω, θα ήταν καλό με χρήση της Πολλαπλής Ανάλυσης Αντιστοιχιών να ομαδοποιηθούν μόνο οι χώρες Δανία, Ελβετία, Ελλάδα, Ισπανία, Ιταλία και Ολλανδία βάσει των στοιχείων που συνθέτουν την ψυχική και σωματική υγεία των κατοίκων τους. Στο γράφημα «Joint» που ακολουθεί, παρουσιάζονται οι προβολές των κατηγοριών των πέντε μεταβλητών:

**ΣΧΗΜΑ 4.5.4**

Διάγραμμα Joint για τις έξι χώρες




Από το παραπάνω διάγραμμα έχουν αφαιρεθεί η ακραίες κατηγορίες των μεταβλητών *q1* και *spheu* επιδιώκοντας την ευκολότερη κατανόηση του. Τα συμπεράσματα σχετικά με την ομαδοποίηση των χωρών θα εξαχθούν μετά την μελέτη της ποιότητας αναπαράστασης που έχουν σε σχέση με τους άξονες.

Στο σημείο αυτό υπολογίσθηκε ο πίνακας με την ποιότητα αναπαράστασης κάθε χώρας:

**ΠΙΝΑΚΑΣ 4.5.4**

Discrimination Measures για τα επίπεδα της μεταβλητής *country*

| | | overall | | Dimension 1 | | Dimension 2 | |
|---|---|---|---|---|---|---|---|
| Country | mass | quality | %inert | coord | sqcorr | coord | sqcorr |
| Netherlands | 0,243 | 0,107 | 1,251 | 0,737 | 0,097 | 0,913 | 0,010 |
| Spain | 0,150 | 0,072 | 1,058 | -0,738 | 0,071 | -0,236 | 0,000 |
| Italy | 0,178 | 0,261 | 1,167 | -0,180 | 0,196 | -2,619 | 0,065 |
| Denmark | 0,142 | 0,144 | 1,318 | 1,045 | 0,108 | 2,323 | 0,036 |
| Greece | 0,205 | 0,037 | 1,118 | -0,429 | 0,031 | 0,737 | 0,006 |
| Switzerland | 0,083 | 0,064 | 1,271 | 0,981 | 0,058 | 1,231 | 0,006 |

Όπως είναι φανερό από τον πίνακα η ποιότητα αναπαράστασης των χωρών είναι μεγαλύτερη στον πρώτο άξονα σε σχέση με τον δεύτερο. Κατά συνέπεια, η ομαδοποίηση τους, θα γίνει με βάση αυτόν τον άξονα. Σύμφωνα με το Σχήμα 4.5.4 οι ομάδες με βάση τον πρώτο άξονα χωρίζονται ως εξής:

- Στην πρώτη ομάδα ανήκουν η Δανία, η Ελβετία και η Ολλανδία. Οι κάτοικοι των χωρών αυτών παρουσιάζουν καλή σωματική υγεία και δεν εμφανίζουν σοβαρά συμπτώματα κακής ψυχικής υγείας, δηλώνουν πολύ ικανοποιημένοι από τη ζωή και «τα βγάζουν πέρα» εύκολα.
- Στην δεύτερη ομάδα ανήκουν η Ελλάδα, η Ισπανία και η Ιταλία. Οι κάτοικοι των χωρών αυτών παρουσιάζουν μέτρια σωματική υγεία και δεν εμφανίζουν σοβαρά συμπτώματα κακής ψυχικής υγείας, δηλώνουν σχετικά ικανοποιημένοι από τη ζωή και «τα βγάζουν πέρα» με σχετική δυσκολία.

Συνοψίζοντας τις δύο ομάδες που προκύπτουν με βάση τα χαρακτηριστικά που σχετίζονται με την ψυχική διάθεση των κατοίκων των χωρών προκύπτει ο ακόλουθος πίνακας:

**ΠΙΝΑΚΑΣ 4.5.5**

Ομαδοποίηση χωρών με βάση την ψυχική διάθεση

| **Ομάδα 1** | **Ομάδα 2** |
|---|---|
| Δανία | Ελλάδα |
| Ελβετία | Ισπανία |
| Ολλανδία | Ιταλία |

## 4.6 Μελέτη ψυχολογικών μεταβλητών ανά ομάδα χωρών

Στην ενότητα αυτή η έρευνα θα επικεντρωθεί στην περεταίρω μελέτη των μεταβλητών που ορίζουν την κατάσταση της ψυχικής υγείας σε σχέση με τις ομάδες των χωρών που δημιουργήθηκαν μετά την Ανάλυση Πολλαπλών Αντιστοιχιών. Επομένως θα χρησιμοποιηθούν μόνο οι μετρήσεις των έξι χωρών που συμμετείχαν στην τελική ομαδοποίηση του Πίνακα 4.5.5. Στόχος της ενότητας είναι η εξαγωγή συμπερασμάτων που αφορούν στην διαμόρφωση των δύο ομάδων. Αρχικά θα ελεγχθεί η ύπαρξη ή μη, ισχυρής σχέσης μεταξύ των ψυχολογικών μεταβλητών και των δύο ομάδων. Στη συνέχεια, θα υπάρξει η κατάλληλη γραφική απεικόνιση και η ερμηνεία των ομάδων αυτών.

Ξεκινώντας τη μελέτη, γίνεται ταυτόχρονη περιγραφή, σύγκρισή καθώς και έλεγχος ύπαρξης εξάρτησης των ομάδων σε σχέση με κάθε μεταβλητή ξεχωριστά.

Αρχικά εξετάζεται εάν η κατάσταση της ψυχικής υγείας του ατόμου εξαρτάται από την ομάδα στην οποία εντάσσεται η χώρα του. Για τον λόγο αυτό, θα γίνει ένας έλεγχος ανεξαρτησίας $\chi^2$ (chi-square test). Ο έλεγχος αυτός εξετάζει την μηδενική υπόθεση $H_0$: οι μεταβλητές είναι ανεξάρτητες, έναντι της $H_1$: οι μεταβλητές είναι εξαρτημένες.

**ΠΙΝΑΚΑΣ 4.6.1**

Συσχέτιση ομάδας χώρας- κατάσταση ψυχικής υγείας

| | | eurodcat | | |
|---|---|---|---|---|
| | | Δεν υπάρχουν συμπτώματα κακής ψυχικής υγείας | Υπάρχουν συμπτώματα κακής ψυχικής υγείας | Σύνολο |
| Ομάδα χωρών | Ομάδα 1 | 3014 | 642 | 3656 |
| | Ομάδα 2 | 2957 | 1208 | 4165 |
| Σύνολο | | 5971 | 1850 | 7821 |

Στον Πίνακα διπλής εισόδου 4.6.1, παρατηρείται ότι σε όλες τις χώρες τα άτομα που δεν παρουσιάζουν συμπτώματα κακής ψυχικής υγείας είναι κατά πολύ περισσότερα από τα άτομα που παρουσιάζουν συμπτώματα κακής ψυχικής υγείας, σε σχέση πάντα με το συνολικό πληθυσμό της εκάστοτε χώρας. Οι κάτοικοι της Δανίας, της Ελβετίας και της Ολλανδίας (Ομάδα 1) φαίνεται να παρουσιάζουν πολύ καλή ψυχική υγεία. Στις χώρες αυτές τα άτομα που παρουσιάζουν συμπτώματα κακής ψυχικής υγείας αποτελούν μόλις το 17,5 του

συνολικού πληθυσμού της ομάδας. Αντίθετα οι κάτοικοι της Ελλάδας, της Ισπανίας και της Ιταλίας (Ομάδα 2) φαίνεται ως επί το πλείστον να μην παρουσιάζουν επαρκή συμπτώματα κακής ψυχικής υγείας, όμως η αναλογία των ατόμων αυτών σε σχέση με όσους παρουσιάζουν κακή ψυχική υγεία φαίνεται να είναι μικρότερη, συγκριτικά με τις προηγούμενες χώρες.

Τα αποτελέσματα του ελέγχου ανεξαρτησίας $\chi^2$ παρουσιάζονται στον πίνακα 4.6.2:

**ΠΙΝΑΚΑΣ 4.6.2**

Έλεγχος ανεξαρτησίας $\chi^2$, για τη σχέση ομάδας χώρας-κατάσταση ψυχικής υγείας

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 141,181 | 1 | 0,000 |

Από τον παραπάνω πίνακα, είναι φανερό πως η στατιστική συνάρτηση έχει τιμή 141,181 και το p-value του ελέγχου είναι περίπου ίσο με 0. Εφόσον το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική υπόθεση σε επίπεδο σημαντικότητας 5%. Αυτό σημαίνει ότι υπάρχει εξάρτηση μεταξύ των δύο μεταβλητών. Επομένως, με βάσει το υπό μελέτη δείγμα, μπορούμε να πούμε πως η κατάσταση της ψυχικής υγείας του ατόμου εξαρτάται από την ομάδα στην οποία εντάσσεται η χώρα που κατοικεί.

Στο σημείο αυτό θα εξετασθεί εάν η σωματική υγεία του ατόμου εξαρτάται από την ομάδα στην οποία εντάσσεται η χώρα στην οποία διαμένει. Όμοια με προηγουμένως, ο έλεγχος που εφαρμόζεται ορίζεται από την μηδενική υπόθεση $H_0$: οι μεταβλητές είναι ανεξάρτητες, έναντι της $H_1$: οι μεταβλητές είναι εξαρτημένες.

**ΠΙΝΑΚΑΣ 4.6.3**

Σχέση ομάδας χώρας- σωματικής υγείας ατόμου

| | | Κατάσταση σωματικής υγείας | | | | | |
|---|---|---|---|---|---|---|---|
| | | Πολύ καλή | Καλή | Μέτρια | Κακή | Πολύ κακή | Σύνολο |
| Ομάδα χωρών | Ομάδα 1 | 878 | 1768 | 812 | 159 | 39 | 3656 |
| | Ομάδα 2 | 616 | 1837 | 1343 | 314 | 55 | 4165 |
| Σύνολο | | 1494 | 3605 | 2155 | 473 | 94 | 7821 |

Από τον παραπάνω πίνακα παρατηρείται ότι οι κάτοικοι των ευρωπαϊκών χωρών παρουσιάζουν σχετικά καλή υγεία στο σύνολό τους. Οι χώρες που εντάσσονται στην πρώτη ομάδα, δηλαδή η Δανία, η Ελβετία και η Ολλανδία, φαίνεται πως χαρακτηρίζονται από «πολύ

καλή» υγεία σε μεγαλύτερο ποσοστό σε σχέση με την δεύτερη ομάδα και σε μικρότερο ποσοστό από «πολύ κακή» υγεία. Αντίθετα, οι κάτοικοι της δεύτερης ομάδας, δηλαδή της Ελλάδας, της Ισπανίας και της Ιταλίας, χαρακτηρίζονται σε μεγαλύτερο ποσοστό από «καλή» υγεία σε σχέση με την πρώτη.

**ΠΙΝΑΚΑΣ 4.6.4**

Έλεγχος $\chi^2$, για τη σχέση ομάδας χώρας-κατάσταση υγείας

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 199,342 | 4 | 0,000 |

Η στατιστική συνάρτηση του ελέγχου $\chi^2$ έχει τιμή 199,342 ενώ το p-value του ελέγχου ισούται περίπου με 0. Επειδή το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική υπόθεση. Επομένως τα δεδομένα φανερώνουν ισχυρή ένδειξη εξάρτησης των ατόμων που αναλογούν σε κάθε ομάδα χωρών και της κατάστασης της υγείας τους.

Επιπλέον θα μελετηθεί η σχέση μεταξύ της ικανοποίησης που νιώθει ένα άτομο για τη ζωή και της ομάδας στην οποία εντάσσεται η χώρα στην οποία κατοικεί.

**ΠΙΝΑΚΑΣ 4.6.5**

Σχέση ομάδας χώρας-ικανοποίηση από ζωή

| | | Ικανοποίηση από ζωή | | | | |
|---|---|---|---|---|---|---|
| | | Πολύ ικανοποιημένος | Σχετικά ικανοποιημένος | Σχετικά δυσαρεστημένος | Πολύ δυσαρεστημένος | Σύνολο |
| Ομάδα χωρών | Ομάδα 1 | 2207 | 1335 | 95 | 19 | 3656 |
| | Ομάδα 2 | 1374 | 2312 | 391 | 88 | 4165 |
| Σύνολο | | 3581 | 3647 | 486 | 107 | 7821 |

Από τον παραπάνω πίνακα φαίνεται ότι σε όλες τις χώρες του δείγματος, ένα πολύ μικρό ποσοστό των κατοίκων δήλωσε «πολύ δυσαρεστημένος» από τη ζωή. Στην Δανία, την Ελβετία και την Ολλανδία (Ομάδα 1) είναι φανερό πως οι περισσότεροι κάτοικοι δήλωσαν «πολύ ικανοποιημένοι» από τη ζωή. Αντίθετα, στην Ελλάδα, την Ισπανία και την Ιταλία (Ομάδα 2), οι περισσότεροι κάτοικοι δήλωσαν «σχετικά ικανοποιημένοι» από τη ζωή. Αξιοσημείωτη είναι και η διαφορά που υπάρχει μεταξύ των δύο ομάδων ως προς τις κατηγορίες που δείχνουν δυσαρέσκεια για τη ζωή. Πιο αναλυτικά, στην πρώτη ομάδα

παρατηρούνται μόλις 114 από τους 3656 κατοίκους των χωρών που δήλωσαν «σχετικά» ή «πολύ δυσαρεστημένοι», σε αντίθεση με την δεύτερη ομάδα όπου στις ίδιες κατηγορίες παρατηρούνται 479 από τους 4165 κατοίκους.

**ΠΙΝΑΚΑΣ 4.6.6**

Έλεγχος $\chi^2$, για τη σχέση ομάδας χώρας-ικανοποίηση από ζωή

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 649,901 | 3 | 0,000 |

Η στατιστική συνάρτηση του ελέγχου $\chi^2$ έχει τιμή 649,901 ενώ το p-value του ελέγχου ισούται περίπου με 0. Επειδή το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική υπόθεση. Επομένως τα δεδομένα φανερώνουν εξάρτηση των ατόμων που αναλογούν σε κάθε ομάδα χωρών και της ικανοποίησης που νιώθουν για τη ζωή.

Σε αυτό το σημείο θα εξετασθεί εάν το χαρακτηριστικό που δείχνει κατά πόσο ο εκπρόσωπος του νοικοκυριού δήλωσε πως έχει οικονομική δυνατότητα (τα βγάζει πέρα οικονομικά) εξαρτάται από την ομάδα στην οποία εντάσσεται η χώρα που κατοικεί.

**ΠΙΝΑΚΑΣ 4.6.7**

Σχέση ομάδας χώρας-οικονομική δυνατότητα

| | | Οικονομική δυνατότητα | | | | |
|---|---|---|---|---|---|---|
| | | Μεγάλη οικονομική δυσκολία | Σχετική οικονομική δυσκολία | Σχετική οικονομική ευκολία | Οικονομική ευκολία | Σύνολο |
| Ομάδα χωρών | Ομάδα 1 | 137 | 521 | 1434 | 1564 | 3656 |
| | Ομάδα 2 | 808 | 1792 | 1135 | 430 | 4165 |
| Σύνολο | | 945 | 2313 | 2569 | 1994 | 7821 |

Σύμφωνα με τον Πίνακα 4.6.7, οι περισσότεροι κάτοικοι της πρώτης ομάδας (Δανία, Ελβετία, Ολλανδία) δήλωσαν πως έχουν «οικονομική ευκολία» ενώ οι λιγότεροι κάτοικοι της ίδιας ομάδας χαρακτηρίζονται από «μεγάλη οικονομική δυσκολία». Αντίθετα, στην δεύτερη ομάδα (Ελλάδα, Ισπανία, Ιταλία), οι περισσότεροι κάτοικοι αντιμετωπίζουν «σχετική οικονομική δυσκολία» και οι λιγότεροι κάτοικοι χαρακτηρίζονται από «οικονομική ευκολία».

**ΠΙΝΑΚΑΣ 4.6.8**

Έλεγχος $\chi^2$, για τη σχέση ομάδας χώρας-οικονομική δυνατότητα

| | Τιμή | Βαθμοί ελευθερίας | p-value |
|---|---|---|---|
| $\chi^2$ του Pearson | 1829,198 | 3 | 0,000 |

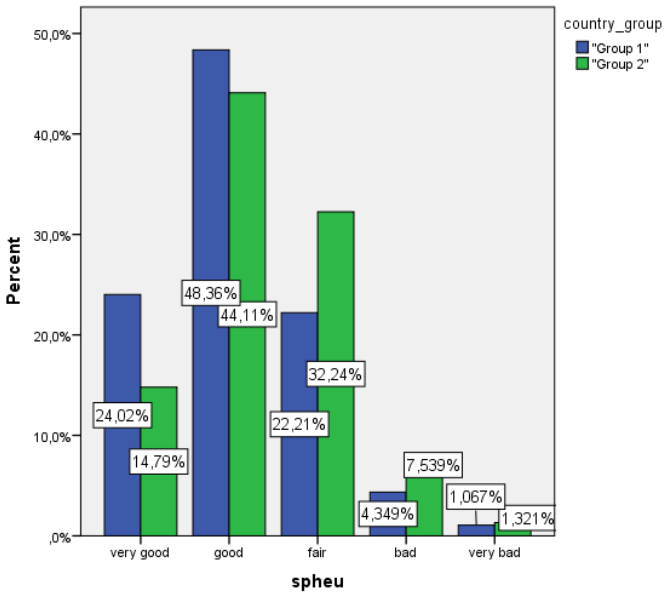
Η στατιστική συνάρτηση του ελέγχου $\chi^2$ έχει τιμή 1829,198 ενώ το p-value του ελέγχου ισούται περίπου με 0. Επειδή το p-value είναι μικρότερο από 5%, απορρίπτουμε την μηδενική υπόθεση. Επομένως τα δεδομένα φανερώνουν ισχυρή ένδειξη εξάρτησης της οικονομικής δυνατότητας των ατόμων σε σχέση με την ομάδα στην οποία εντάσσεται η χώρα που κατοικούν.

Ολοκληρώνοντας τη μελέτη των ομάδων στις οποίες χωρίστηκαν οι ευρωπαϊκές χώρες μετά την Ανάλυση Αντιστοιχιών, θα ήταν χρήσιμο για την εξαγωγή συμπερασμάτων να απεικονιστούν γραφικά τα χαρακτηριστικά που μελετήθηκαν στην Ενότητα 4.5 ξεχωριστά για κάθε ομάδα.

Αρχικά κατασκευάστηκε ένα ραβδόγραμμα για την μεταβλητή που δείχνει την κατάσταση της σωματικής υγείας (*spheu*) στις δύο ομάδες:

**ΣΧΗΜΑ 4.6.1**

Ραβδόγραμμα ομάδων χωρών για την κατάσταση της σωματικής υγείας

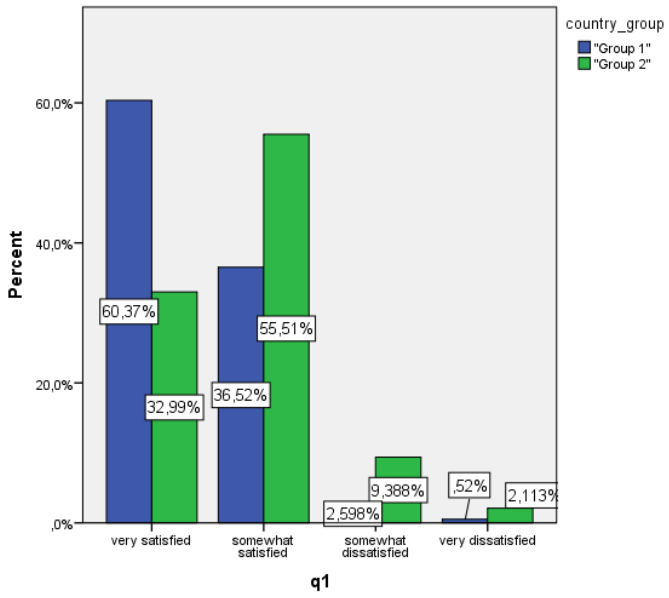


Στο παραπάνω σχήμα, απεικονίζονται με μπλε χρώμα οι χώρες της πρώτης ομάδας και με πράσινο χρώμα οι χώρες της δεύτερης. Σύμφωνα με το γράφημα μπορούμε να παρατηρήσουμε ότι οι κάτοικοι των χωρών της πρώτης ομάδας παρουσιάζουν καλή σωματική υγεία καθώς στην κατηγορία «very good» της μεταβλητής η πρώτη ομάδα σημειώνει μεγαλύτερο ποσοστό έναντι της δεύτερης ενώ στις κατηγορίες «bad» και «very bad» η ίδια ομάδα σημειώνει μικρότερο ποσοστό. Τέλος, αξίζει να σημειωθεί ότι σε γενικές γραμμές δεν παρατηρείται ιδιαίτερα κακή σωματική υγεία στους κατοίκους των χωρών καθώς τα ποσοστά στις αντίστοιχες κατηγορίες είναι αρκετά χαμηλά.

Ακολουθεί ένα ραβδόγραμμα της μεταβλητής που δείχνει την ικανοποίηση από τη ζωή (*q1*) για τις δύο ομάδες.

**ΣΧΗΜΑ 4.6.2**

Ραβδόγραμμα ομάδων χωρών για την μεταβλητή «ικανοποίηση από τη ζωή»

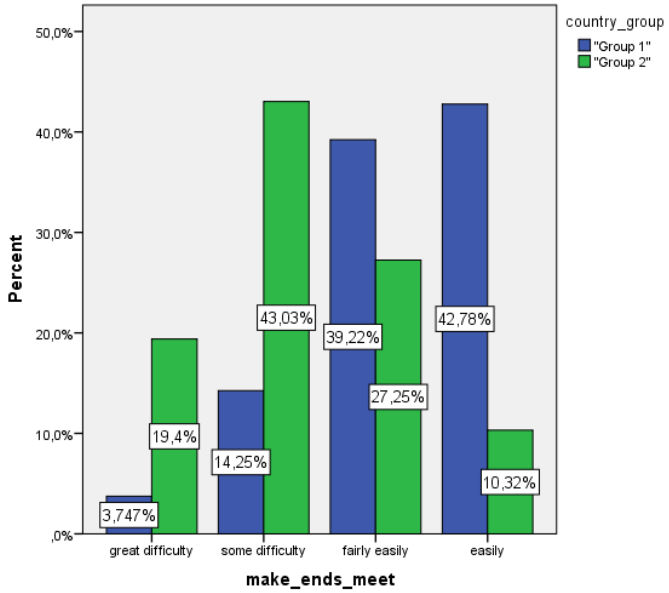



Το παραπάνω γράφημα δεν μας δίνει μια ξεκάθαρη εικόνα για τον τρόπο που οι κάτοικοι των δύο ομάδων βλέπουν τη ζωή καθώς στην κατηγορία «very satisfied» υπερέχουν σαφώς οι κάτοικοι της πρώτης ομάδας, όμως στη δεύτερη κατηγορία «somewhat satisfied» υπερέχει έντονα η δεύτερη. Παρολ' αυτά, πρέπει να σημειωθεί ότι στις κατηγορίες που δείχνουν μια σχετική απογοήτευση από τη ζωή (somewhat και very dissatisfied) το ποσοστό που σημειώνουν οι κάτοικοι της δεύτερης ομάδας είναι σαφώς μεγαλύτερο έναντι της πρώτης.

Στη συνέχεια, κατασκευάστηκε ένα ραβδόγραμμα για την μεταβλητή που ορίζει την οικονομική δυνατότητα (*make_ends_meet*) στις δύο ομάδες:

**ΣΧΗΜΑ 4.6.3**

Ραβδόγραμμα ομάδων χωρών για την μεταβλητή «οικονομική δυνατότητα»

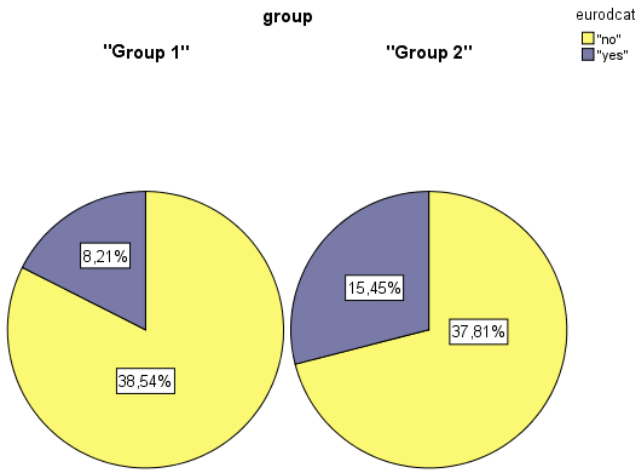



Από το γράφημα είναι φανερό πως οι κάτοικοι των χωρών της πρώτης ομάδας είναι πιο αναπτυγμένοι οικονομικά σε σχέση με τις υπόλοιπες καθώς στις κατηγορίες που δείχνουν οικονομική ευκολία (fairly easily και easily) οι μπλε ράβδοι υπερέχουν έναντι των πράσινων ενώ στις κατηγορίες που δείχνουν οικονομική δυσκολία (great και some difficulty) γίνεται το αντίστροφο.

Τέλος, ακολουθεί ένα διπλό pie-chart για την μεταβλητή που σχετίζεται με την ύπαρξη ή μη, συμπτωμάτων κακής ψυχικής υγείας (*eurodcat*). Στο Σχήμα 4.6.4, κάθε κυκλικό γράφημα αντιστοιχεί σε κάθε ομάδα χωρών. Με κίτρινο χρώμα, απεικονίζεται το ποσοστό των ατόμων που δεν παρουσιάζουν συμπτώματα κακής ψυχικής υγείας και με μωβ χρώμα το ποσοστό των ατόμων που παρουσιάζουν συμπτώματα. Όπως είναι φανερό, ένα μικρό ποσοστό ατόμων παρουσιάζει συμπτώματα κακής ψυχικής υγείας. Όμως, συγκρίνοντας τις δύο ομάδες μεταξύ τους, παρατηρείται πως το ποσοστό αυτό είναι μεγαλύτερο στις χώρες της δεύτερης ομάδας σε σχέση με αυτές της πρώτης κατά 7,24%.

**ΣΧΗΜΑ 4.6.4**

Pie-chart ομάδων χωρών για την μεταβλητή «eurodcat»




## 4.7 Εφαρμογή της Ανάλυσης Κύριων Συνιστωσών

Στην προηγούμενη ενότητα διαπιστώθηκε ότι κάποιες μεταβλητές που ορίζουν την ψυχολογική συμπεριφορά είναι συσχετισμένες τόσο μεταξύ τους όσο και με την χώρα διαμονής και έτσι εφαρμόστηκε πολλαπλή ανάλυση αντιστοιχιών για την μελέτη των μεταβλητών αυτών. Σε αυτή την ενότητα, θα εξετασθεί εάν οι μεταβλητές που ορίζουν την οικονομική κατάσταση των νοικοκυριών είναι μεταξύ τους συσχετισμένες και στην περίπτωση αυτή θα εφαρμοστεί Ανάλυση Κύριων Συνιστωσών. Η μέθοδος των Κύριων Συνιστωσών είναι μια τεχνική ανάλυσης δεδομένων που στοχεύει στην δημιουργία καινούριων μεταβλητών, οι οποίες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών, έτσι ώστε να είναι ασυσχέτιστες μεταξύ τους και να περιέχουν όσο το δυνατόν μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Οι νέες αυτές μεταβλητές που παράγονται ονομάζονται «Κύριες Συνιστώσες».

Οι μεταβλητές που θα χρησιμοποιηθούν για την ενότητα αυτή, στην ανάλυση κύριων συνιστωσών είναι οι εξής:

- Ακαθάριστο οικογενειακό εισόδημα (Yhh_nir)
- Καθαρό οικογενειακό εισόδημα (Yhh_net)
- Οικογενειακό εισόδημα από σύνταξη (Hhold_pensions)

- Οικογενειακό εισόδημα από εργασία (Hhold_labour)
- Οικογενειακό εισόδημα από άλλες πηγές (Other_source)
- Συνολική αξία των χρηματικών περιουσιακών στοιχείων του νοικοκυριού (hgfinv)
- Συνολική αξία της ιδιόκτητης κατοικίας του νοικοκυριού (homev)
- Συνολική αξία της άλλης ακίνητης περιουσίας (εκτός της κύριας κατοικίας) του νοικοκυριού (horesv)
- Συνολική αξία των παγίων περιουσιακών στοιχείων του νοικοκυριού (hrav)

Η μέθοδος της Ανάλυσης Κύριων Συνιστωσών υποθέτει πως υπάρχουν κάποιοι γραμμικοί συνδυασμοί των αρχικών μεταβλητών που περιέχουν το μεγαλύτερο μέρος της πληροφορίας τους. Συνεπώς, εάν οι αρχικές μεταβλητές είναι ασυσχέτιστες, δεν θα υπάρχουν κύριες συνιστώσες και άρα η μέθοδος δεν μπορεί να εφαρμοστεί. Αρχικά θα κατασκευασθεί ο πίνακας διακυμάνσεων για να υπάρξει μια πρώτη εικόνα.

Ο Πίνακας 4.7.1 που ακολουθεί είναι ένας δισδιάστατος πίνακας διασπορών-συνδιασπορών των οικονομικών χαρακτηριστικών του δείγματος. Πιο συγκεκριμένα, η διαγώνιός του δείχνει τις τιμές των διασπορών των μεταβλητών και κατ' επέκταση τη μεταβλητότητα τους και όλα τα υπόλοιπα κελιά έχουν συμπληρωθεί με τη συνδιασπορά κάθε ζεύγους μεταβλητών. Παρατηρώντας την διαγώνιο βλέπουμε σημαντικές διαφορές μεταξύ των διακυμάνσεων. Ενδεικτικά αναφέρουμε πως $Var(Yhh\,nir) = 20909830040$, $Var(Hhold\ pensions) = 351797919$ και $Var(hrav) = 713127702005$. Το γεγονός αυτό, σε συνδυασμό με το δεδομένο πως η κλίμακα μέτρησης είναι ίδια (€) θα μας αποτρέψει να χρησιμοποιήσουμε τον πίνακα διακυμάνσεων στη συνέχεια της μελέτης. Στην περίπτωση αυτή, θα ήταν απαραίτητο να κατασκευασθεί ο πίνακας συσχετίσεων. Πρόκειται για τον τυποποιημένο πίνακα διασπορών και είναι κατάλληλος για να επιβεβαιωθεί η υπόθεση ύπαρξης συσχετίσεων μεταξύ όλων των μεταβλητών.

Ο Πίνακας 4.7.2 αποτελεί έναν πίνακα συσχετίσεων μεταξύ των μεταβλητών. Για κάθε ζεύγος μεταβλητών καταγράφεται ο βαθμός συσχέτισης. Μελετώντας τις τιμές των συντελεστών συσχέτισης είναι φανερό πως υπάρχει υψηλή συσχέτιση κάθε μεταβλητής με τουλάχιστον μία ακόμα και κατ' επέκταση δεν είναι ασυσχέτιστες. Συμπερασματικά, εφόσον κάθε μεταβλητή συσχετίζεται με τουλάχιστον μία, είναι δυνατόν να εφαρμοστεί η μέθοδος των Κυρίων Συνιστωσών.

**ΠΙΝΑΚΑΣ 4.7.1**

Πίνακας διακυμάνσεων

| | Yhh_nir | Yhh_net | Hhold_pensions | Hhold_labour | Other_source | hgfinv | homev | horesv | hrav |
|---|---|---|---|---|---|---|---|---|---|
| Yhh_nir | 20909830040 | | | | | | | | |
| Yhh_net | 18750970161 | 16920413527 | | | | | | | |
| Hhold_pensions | 201093479 | 195027523 | 351797919 | | | | | | |
| Hhold_labour | 1363568209 | 979122852 | -160995417 | 1294148499 | | | | | |
| Other_source | 19345168472 | 17576819894 | 10290976 | 230415127 | 19104462494 | | | | |
| hgfinv | 2573150301 | 1869294458 | 167741785 | 1177742028 | 1227666482 | 56441731423 | | | |
| homev | 2912859493 | 2170732775 | 119299418 | 1871801862 | 921758193 | 9512732850 | 200988484471 | | |
| horesv | 2149688718 | 1769424569 | 104041829 | 693642148 | 1352004757 | 3422040501 | 11431211867 | 40318535535 | |
| hrav | 7404368814 | 5441786713 | 111986051 | 4357529228 | 2934853485 | 22525181720 | 211819171793 | 65449952016 | 713127702005 |

**ΠΙΝΑΚΑΣ 4.7.2**

Πίνακας συσχετίσεων για τις αρχικές μεταβλητές

| | Yhh_nir | Yhh_net | Hhold_pensions | Hhold_labour | Other_source | hgfinv | homev | horesv |
|---|---|---|---|---|---|---|---|---|
| Yhh_net | 0,997*** | | | | | | | |
| Hhold_pensions | 0,074*** | 0,080*** | | | | | | |
| Hhold_labour | 0,262*** | 0,209*** | -0,239*** | | | | | |
| Other_source | 0,968*** | 0,978*** | 0,004*** | 0,046*** | | | | |
| hgfinv | 0,075*** | 0,060*** | 0,038*** | 0,138*** | 0,037*** | | | |
| homev | 0,045*** | 0,037*** | 0,014* | 0,116*** | 0,015* | 0,089*** | | |
| horesv | 0,074*** | 0,068*** | 0,028*** | 0,096*** | 0,049*** | 0,072*** | 0,127*** | |
| hrav | 0,061*** | 0,050*** | 0,007 n.s. | 0,143*** | 0,025*** | 0,112*** | 0,559*** | 0,386*** |

***p-value<0.01, **p-value<0.05, *p-value<0.10, n.s.: not significant

Για να γίνει η επιλογή των κύριων συνιστωσών κατασκευάστηκε ο παρακάτω πίνακας:

**ΠΙΝΑΚΑΣ 4.7.3**

Ιδιοτιμές των συνιστωσών

| | Eigenanalysis of the Correlation Matrix | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $PC1$ | $PC2$ | $PC3$ | $PC4$ | $PC5$ | $PC6$ | $PC7$ | $PC8$ | $PC9$ |
| Eigenvalue | 3,0408 | 1,8033 | 1,2104 | 0,9856 | 0,8802 | 0,7028 | 0,3748 | 0,0021 | 0,000 |
| Proportion | 0,338 | 0,200 | 0,134 | 0,110 | 0,098 | 0,078 | 0,042 | 0,000 | 0,000 |
| Cumulative | 0,338 | 0,538 | 0,673 | 0,782 | 0,880 | 0,958 | 1,000 | 1,000 | 1,000 |

Στην πρώτη γραμμή του Πίνακα 4.7.3 (Eigenvalue) δίνονται οι ιδιοτιμές των συνιστωσών. Η πρώτη κύρια συνιστώσα είναι αυτή με τη μέγιστη ιδιοτιμή (3,0408). Στην δεύτερη γραμμή (Proportion) δίνεται το ποσοστό της μεταβλητότητας που ερμηνεύει ξεχωριστά κάθε κύρια συνιστώσα. Επομένως φαίνεται πως η πρώτη κύρια συνιστώσα ερμηνεύει το 33,8% της μεταβλητότητας, η δεύτερη κύρια συνιστώσα ερμηνεύει το 20%, η τρίτη συνιστώσα ερμηνεύει το 13,4% κ.ο.κ. Στην τρίτη γραμμή του πίνακα (Cumulative), δίνεται το αθροιστικό ποσοστό που ερμηνεύεται από το εκάστοτε πλήθος των συνιστωσών. Έτσι, η δύο πρώτες κύριες συνιστώσες ερμηνεύουν αθροιστικά το 53,8% της συνολικής μεταβλητότητας, οι τρεις πρώτες κύριες συνιστώσες ερμηνεύουν το 67,3% της συνολικής μεταβλητότητας κ.ο.κ.

Για την επιλογή του βέλτιστου πλήθους συνιστωσών μπορούμε να χρησιμοποιήσουμε το κριτήριο του Kaiser σύμφωνα με το οποίο, διατηρούμε μόνο τις συνιστώσες που είναι μεγαλύτερες από τη μέση τιμή των ιδιοτιμών $\lambda_i$ .Επομένως εφόσον εργαζόμαστε με τον πίνακα συσχετίσεων, βάσει του παραπάνω κριτηρίου, σημαντικές είναι όσες συνιστώσες έχουν ιδιοτιμή μεγαλύτερη της μονάδας (eigenvalue>1). Επομένως, με βάση το κριτήριο αυτό επιλέγονται οι τρεις πρώτες κύριες συνιστώσες οι οποίες ερμηνεύουν αθροιστικά το 67,3% της συνολικής μεταβλητότητας. Καθώς το ποσοστό αυτό πλησιάζει το 70%, μπορούμε να δεχτούμε ότι οι τρεις πρώτες συνιστώσες έχουν μια καλή ερμηνευτική ικανότητα.

Στο σημείο αυτό δίνεται ο πίνακας με τους συντελεστές που αντιστοιχούν στις αρχικές μεταβλητές για την κατασκευή της πρώτης έως και της τρίτης κύριας συνιστώσας.

**ΠΙΝΑΚΑΣ 4.7.4**

Συντελεστές αρχικών μεταβλητών

| | $PC1$ | $PC2$ | $PC3$ |
|---|---|---|---|
| Yhh_nir | 0,570 | -0,080 | 0,001 |
| Yhh_net | 0,568 | -0,097 | 0,031 |
| Hhold pensions | 0,029 | -0,047 | 0,738 |
| Hhold labour | 0,159 | 0,231 | -0,626 |
| Other source | 0,551 | -0,137 | 0,064 |
| hgfinv | 0,070 | 0,202 | -0,087 |
| homev | 0,068 | 0,549 | 0,123 |
| horesv | 0,084 | 0,411 | 0,133 |
| hrav | 0,087 | 0,632 | 0,137 |

Σύμφωνα με τον Πίνακα 4.7.4, οι 3 κύριες συνιστώσες έχουν τη μορφή:

$$PC1 = 0{,}570Yhhnir + 0{,}568Yhhnet + 0{,}029Hholdpensions + 0{,}159Hholdlabour + 0{,}551Othersource + 0{,}070hgfinv + 0{,}068homev + 0{,}084horesv + 0{,}087hrav$$

$$PC2 = -0{,}080Yhhnir - 0{,}097Yhhnet - 0{,}047Hholdpensions + 0{,}231Hholdlabour - 0{,}137Othersource + 0{,}202hgfinv + 0{,}549homev + 0{,}411horesv + 0{,}632hrav$$

$$PC3 = 0{,}001Yhhnir + 0{,}031Yhhnet + 0{,}738Hholdpensions - 0{,}626Hholdlabour + 0{,}064Othersource - 0{,}087hgfinv + 0{,}123homev + 0{,}133horesv + 0{,}137hrav$$

Επομένως για κάθε στοιχείο που προστίθεται, είναι δυνατόν να υπολογισθούν οι αντίστοιχες τιμές των συνιστωσών αντικαθιστώντας κάθε μεταβλητή με την αντίστοιχη μέτρησή της.

Ολοκληρώνοντας τη διαδικασία θα μπορούσε να δοθεί κάποια ερμηνεία στις τρεις κύριες συνιστώσες που κατασκευάστηκαν. Για να γίνει αυτό θα πρέπει να εξετασθεί η συσχέτιση που έχει κάθε συνιστώσα με καθεμία από τις αρχικές μεταβλητές. Επομένως κατασκευάστηκε ο Πίνακας 4.7.5 που αποτελεί έναν πίνακα συσχετίσεων ο οποίος συμπεριλαμβάνει τις αρχικές μεταβλητές και τις 3 κύριες συνιστώσες.

**ΠΙΝΑΚΑΣ 4.7.5**

Πίνακας συσχετίσεων αρχικών μεταβλητών και κύριων συνιστωσών

| | Yhh_nir | Yhh_net | Hhold pensions | Hhold labour | Other source | hgfinv | homev | horesv | hrav |
|---|---|---|---|---|---|---|---|---|---|
| $PC1$ | 0,993*** | 0,990*** | 0,050*** | 0,278*** | 0,960*** | 0,122*** | 0,119*** | 0,146*** | 0,152*** |
| $PC2$ | -0,108*** | -0,13*** | -0,064*** | 0,310*** | -0,185*** | 0,271*** | 0,737*** | 0,551*** | 0,849*** |
| $PC3$ | 0,001 n.s. | 0,034*** | 0,812*** | -0,688*** | 0,070*** | -0,095*** | 0,135*** | 0,146 n.s. | 0,151*** |

***p-value<0.01, **p-value<0.05, *p-value<0.10, n.s.: not significant

Σύμφωνα με τα παραπάνω, προκύπτουν τα ακόλουθα συμπεράσματα:

- η πρώτη συνιστώσα συσχετίζεται πολύ ισχυρά με τις μεταβλητές Yhh_nir με συντελεστή συσχέτισης $r = 0{,}993$ , Yhh_net με συντελεστή συσχέτισης $r = 0{,}990$ και Other_source με συντελεστή συσχέτισης $r = 0{,}960$. Επομένως, μπορούμε να θεωρήσουμε πως η πρώτη κύρια συνιστώσα αφορά στο ακαθάριστο και καθαρό οικογενειακό εισόδημα καθώς και στο οικογενειακό εισόδημα από πηγές εκτός εργασίας και σύνταξης.
- Η δεύτερη κύρια συνιστώσα συσχετίζεται ισχυρά με τις μεταβλητές homev με συντελεστή συσχέτισης $r = 0{,}737$ , hrav με συντελεστή συσχέτισης $r = 0{,}849$ και horesv με $r = 0{,}551$. Συνεπώς μπορούμε να πούμε ότι αφορά στην συνολική αξία της ιδιόκτητης κατοικίας και των παγίων περιουσιακών στοιχείων του νοικοκυριού καθώς και στη συνολική αξία της άλλης ακίνητης περιουσίας (εκτός της κύριας κατοικίας) του νοικοκυριού. Θα πρέπει βέβαια να σημειωθεί και η συσχέτιση που έχει η συνιστώσα με την μεταβλητή Hhold_labour ( $r = 0{,}310$) αλλά και με την hgfinv ( $r = 0{,}271$) και κατ' επέκταση θα μπορούσαμε επίσης να συμπεράνουμε ότι η δεύτερη συνιστώσα αφορά και το οικογενειακό εισόδημα από εργασία καθώς και τη συνολική αξία των χρηματικών περιουσιακών στοιχείων του νοικοκυριού.
- Η τρίτη κύρια συνιστώσα συσχετίζεται ισχυρά με τις μεταβλητές Hhold_pensions με συντελεστή συσχέτισης $r = 0{,}812$ και Hhold_labour με συντελεστή $r = -0{,}688$. Επομένως αφορά στο οικογενειακό εισόδημα από σύνταξη και εργασία.

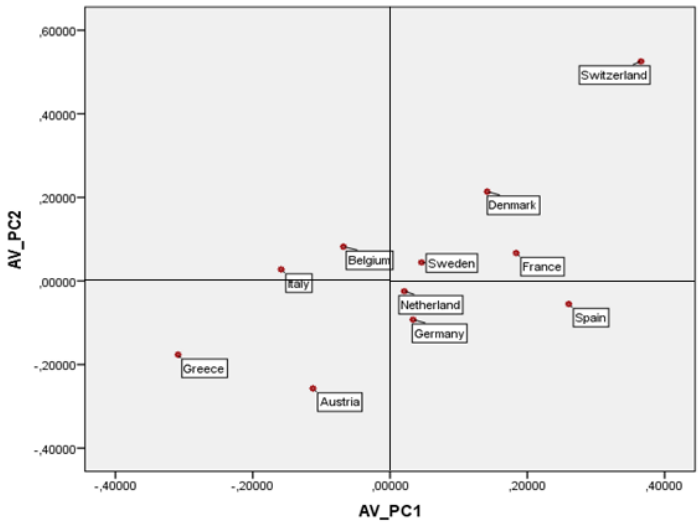
## 4.8 Γραφική απεικόνιση οικονομικών παραγόντων ανά χώρα

Στην ενότητα αυτή, θα γίνει γραφική απεικόνιση των μέσων οικονομικών στοιχείων των χωρών. Πιο αναλυτικά, οι οικονομικές μεταβλητές έχουν αντικατασταθεί πλέον από τις τρεις κύριες συνιστώσες. Συνεπώς το δείγμα αποτελείται από 16537 κατοίκους 11 ευρωπαϊκών χωρών οι οποίοι στην παρούσα φάση μελετώνται ως προς 3 οικονομικές συνιστώσες. Για την ευκολότερη ανάγνωση των γραφημάτων που πρόκειται να κατασκευασθούν, υπολογίσθηκε η μέση τιμή κάθε συνιστώσας για κάθε χώρα. Αποτέλεσμα του υπολογισμού αυτού, είναι να υπάρχουν τελικά 11 μέσες τιμές για κάθε οικονομική συνιστώσα. Η ενότητα ξεκινάει με απεικόνιση των τιμών σε διαγράμματα διασποράς. Στη συνέχεια, ακολουθεί ένα τρισδιάστατο γραφήματα που απεικονίζει ταυτόχρονα τις τρεις συνιστώσες. Τέλος, ολοκληρώνουμε με ένα δενδρόγραμμα των μέσων τιμών των συνιστωσών που δίνει μια άποψη της ομαδοποίησης των οικονομικών στοιχείων που αντιπροσωπεύουν οι συνιστώσες, σε σχέση με τη χώρα διαμονής.

Αρχικά κατασκευάσθηκε ένα διάγραμμα διασποράς που απεικονίζει τη μέση τιμή της πρώτης και της δεύτερης κύριας συνιστώσας για κάθε ευρωπαϊκή χώρα του δείγματος:

**ΣΧΗΜΑ 4.8.1α**

Διάγραμμα διασποράς μεταξύ των μέσων τιμών των συνιστωσών $PC1$ και $PC2$



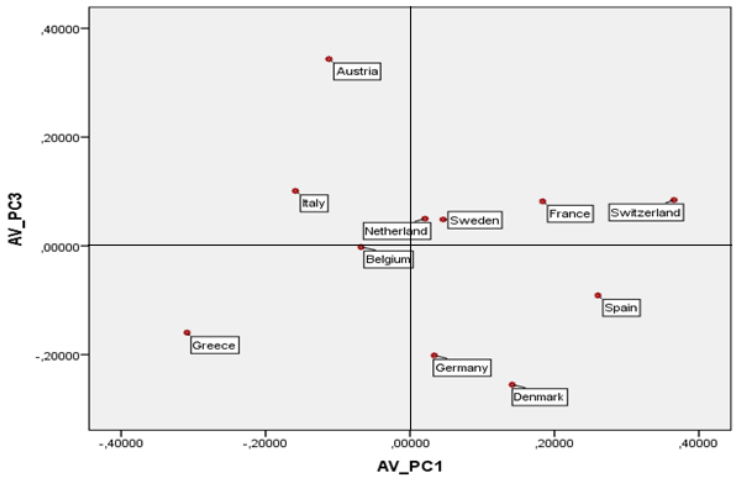
Στο παραπάνω γράφημα, ο οριζόντιος άξονας αντιστοιχεί στην μέση τιμή της πρώτης κύριας συνιστώσας και ο κάθετος, στην μέση τιμή της δεύτερης κύριας συνιστώσας. Για κάθε σημείο του διαγράμματος διασποράς, καταγράφεται και η χώρα στην οποία αντιστοιχεί.

Σύμφωνα με το διάγραμμα, η Ελλάδα και η Ελβετία παρουσιάζουν εκ διαμέτρου αντίθετες τιμές ως προς τις δύο πρώτες συνιστώσες. Το γεγονός αυτό τις εντάσσει σε δύο ξεχωριστές ομάδες καθεμία εκ των οποίων περιλαμβάνει μόνο μία χώρα. Παράλληλα, το Βέλγιο, η Σουηδία, η Ολλανδία και η Γερμανία συσσωρεύονται γύρω από το 0 δημιουργώντας κατά αυτόν τον τρόπο μια τρίτη ομάδα. Η Δανία, η Γαλλία και η Ισπανία δείχνουν να έχουν κοντινές τιμές ως προς τις δύο πρώτες συνιστώσες κι έτσι μια τέταρτη ομάδα συμπεριλαμβάνει τις χώρες αυτές. Τέλος, η Αυστρία και η Ιταλία απέχουν από κάθε άλλη ομάδα με αποτέλεσμα να αποτελούν δύο ξεχωριστές ομάδες.

Με τον ίδιο τρόπο κατασκευάσθηκε το διάγραμμα διασποράς μεταξύ της μέσης τιμής της πρώτης και της τρίτης κύριας συνιστώσας για κάθε ευρωπαϊκή χώρα του δείγματος:

**ΣΧΗΜΑ 4.8.1β**

Διάγραμμα διασποράς μεταξύ των μέσων τιμών των συνιστωσών $PC1$ και $PC3$



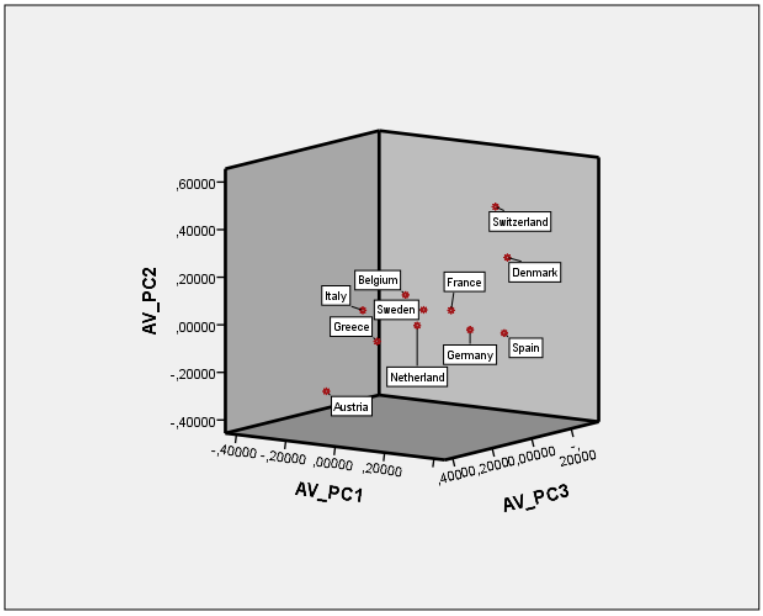

Όμοια με παραπάνω, στο Σχήμα 4.8.1β ο οριζόντιος άξονας αντιστοιχεί στην μέση τιμή της πρώτης κύριας συνιστώσας και ο κάθετος, στην μέση τιμή της τρίτης κύριας συνιστώσας. Για κάθε σημείο του διαγράμματος διασποράς, καταγράφεται και η χώρα στην οποία αντιστοιχεί.

Παρατηρώντας το γράφημα με την πρώτη και την τρίτη κύρια συνιστώσα, φαίνεται ότι η Ολλανδία, η Σουηδία και το Βέλγιο συσσωρεύονται γύρω από το 0 δημιουργώντας μια ομάδα (όπως στο Σχήμα 4.8.1α) αλλά πλέον η Γερμανία απέχει από την ομάδα αυτή και προσεγγίζει περισσότερο τη Δανία. Ομοίως η Γαλλία, η Ισπανία και η Δανία φαίνεται να έχουν μεγάλες αποστάσεις μεταξύ τους και κατά συνέπεια θα ήταν λάθος να ενταχθούν στην ίδια ομάδα όπως προαναφέρθηκε. Τέλος, η Ελβετία, η Ελλάδα αλλά και η Αυστρία αποτελούν τρεις διαφορετικές ομάδες καθώς δεν προσεγγίζουν καμία άλλη τιμή.

Τα παραπάνω αποτελέσματα μπορούν να αναπαρασταθούν ταυτόχρονα για τις τρεις κύριες συνιστώσες, σε ένα τρισδιάστατο διάγραμμα διασποράς ώστε να γίνει πιο ξεκάθαρη η ομαδοποίηση των χωρών.

**ΣΧΗΜΑ 4.8.2**

Τρισδιάστατο διάγραμμα διασποράς μεταξύ των μέσων τιμών των συνιστωσών $PC1$ ,$PC2$ και $PC3$




Όπως φαίνεται από το Σχήμα 4.8.2, η Ελβετία συγκεντρώνει υψηλές τιμές και στις τρεις συνιστώσες ενώ η Ελλάδα παρουσιάζει χαμηλές τιμές και στις τρεις συνιστώσες. Η Γερμανία, η Γαλλία και η Ισπανία φαίνεται να έχουν χαμηλές τιμές στην δεύτερη και τρίτη συνιστώσα αλλά πιο υψηλές στην πρώτη, αντίθετα η Αυστρία έχει υψηλές τιμές στην τρίτη συνιστώσα και χαμηλές στις άλλες δύο. Η Σουηδία, το Βέλγιο και η Ολλανδία έχουν μέσες

τιμές που προσεγγίζουν το 0 σε όλες τις συνιστώσες. Η Δανία έχει υψηλές τιμές σε πρώτη και δεύτερη συνιστώσα και χαμηλή τιμή στην τρίτη ενώ αντίθετα η Ιταλία έχει χαμηλές τιμές στις δύο πρώτες συνιστώσες και υψηλή στην τρίτη.

Ολοκληρώνοντας την μελέτη σχετικά με την ομαδοποίηση των χωρών με βάση τις κύριες συνιστώσες θα εφαρμοσθεί η μη Ιεραρχική μέθοδος K-Means με σκοπό την κατασκευή τεσσάρων ομάδων. Στον πίνακα που ακολουθεί παρουσιάζεται το πλήθος των χωρών που αποτελούν την καθεμία εκ των τεσσάρων ομάδων.

**ΠΙΝΑΚΑΣ 4.8.1**

Πλήθος χωρών για τις τέσσερεις ομάδες

| | Πλήθος χωρών |
|---|---|
| Ομάδα 1 | 1 |
| Ομάδα 2 | 1 |
| Ομάδα 3 | 2 |
| Ομάδα 4 | 7 |
| Σύνολο | 11 |

Όπως φαίνεται από τον παραπάνω πίνακα, από τις 11 χώρες του δείγματος, οι 7 εντάσσονται στην τέταρτη ομάδα, οι 2 στην τρίτη ενώ οι ομάδες 1 και 2 έχουν από μία χώρα η καθεμία.

Η διαμόρφωση των τεσσάρων ομάδων ανάλογα με την χώρα διαμονής παρουσιάζεται στον Πίνακα 4.8.2:

**ΠΙΝΑΚΑΣ 4.8.2**

Ομαδοποίηση χωρών σε τέσσερεις ομάδες με βάση την οικονομική κατάσταση των κατοίκων

| **Ομάδα 1** | **Ομάδα 2** | **Ομάδα 3** | **Ομάδα 4** |
|---|---|---|---|
| Αυστρία | Ελβετία | Ελλάδα | Βέλγιο |
| | | Ιταλία | Γαλλία |
| | | | Γερμανία |
| | | | Δανία |
| | | | Ισπανία |
| | | | Ολλανδία |
| | | | Σουηδία |

Από τα παραπάνω προκύπτουν τα εξής συμπεράσματα:

- Η Αυστρία παρουσιάζει μέτρια τιμή στην πρώτη κύρια συνιστώσα, χαμηλή στην δεύτερη και υψηλή στην τρίτη. Το γεγονός αυτό σημαίνει ότι οι κάτοικοι της Αυστρίας χαρακτηρίζονται από μέτριο ακαθάριστο και καθαρό εισόδημα καθώς και οικογενειακό εισόδημα από πηγές εκτός εργασίας και σύνταξης, χαμηλή συνολική αξία των περιουσιακών στοιχείων του νοικοκυριού και υψηλό οικογενειακό εισόδημα από σύνταξη και εργασία.
- Η Ελβετία παρουσιάζει υψηλή τιμή σε όλες τις κύριες συνιστώσες. Το γεγονός αυτό σημαίνει ότι οι κάτοικοι της Ελβετίας χαρακτηρίζονται από υψηλό οικογενειακό εισόδημα ανεξαρτήτως πηγής, αλλά και από υψηλά περιουσιακά στοιχεία.
- Η Ελλάδα και η Ιταλία παρουσιάζουν σχετικά χαμηλές τιμές σε όλες τις κύριες συνιστώσες. Το γεγονός αυτό σημαίνει ότι οι κάτοικοι των χωρών αυτών χαρακτηρίζονται από χαμηλό ακαθάριστο και καθαρό οικογενειακό εισόδημα σε σχέση με τις υπόλοιπες ευρωπαϊκές χώρες, χαμηλό οικογενειακό εισόδημα ανεξαρτήτως πηγής αλλά και από χαμηλά περιουσιακά στοιχεία.
- Οι υπόλοιπες χώρες, δηλαδή το Βέλγιο, η Γαλλία, η Γερμανία, η Δανία, η Ισπανία, η Ολλανδία και η Σουηδία αποτελούν τον «πυρήνα» των ευρωπαϊκών χωρών. Χαρακτηρίζονται κυρίως από μέσες τιμές στις τρεις κύριες συνιστώσες γεγονός που δείχνει πως οι κάτοικοι τους έχουν μέσα οικογενειακά εισοδήματα.

## 4.9 Συμπεράσματα

Στην ενότητα αυτή θα παρουσιαστούν τα συμπεράσματα που προέκυψαν κατά την εφαρμογή των πολυμεταβλητών μεθόδων ομαδοποίησης των δεδομένων. Πιο αναλυτικά, η έρευνα πραγματοποιήθηκε μετά από συλλογή δεδομένων 11 Ευρωπαϊκών χωρών. Τα χαρακτηριστικά που μελετήθηκαν αφορούσαν σε μετρήσεις στοιχείων που συνθέτουν την ψυχολογία των ατόμων αλλά και στοιχείων που αφορούν στην οικονομική τους κατάσταση. Απώτερος σκοπός είναι να διαπιστωθεί κατά πόσο οι ευρωπαϊκές χώρες μπορούν να χωριστούν σε ομάδες με βάσει τα προαναφερθέντα στοιχεία, αλλά και κατά πόσο οι ομάδες αυτές παρουσιάζουν ομοιότητες μεταξύ τους.

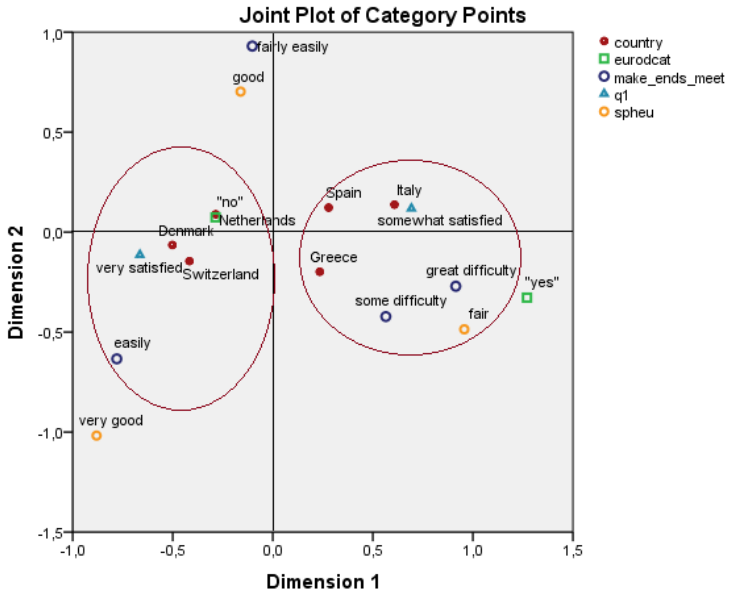
Στην αρχή της έρευνας έγινε μια πρώτη ομαδοποίηση των χωρών με χρήση δενδρογράμματος (ΣΧΗΜΑ 4.2.1) το οποίο κατασκευάσθηκε για ένα μικρό δείγμα δεδομένων βάσει του Καθαρού και του Ακαθάριστου οικογενειακού εισοδήματος των κατοίκων. Διαπιστώθηκε έτσι (ΠΙΝΑΚΑΣ 4.2.3) ότι οι Νότιες χώρες, δηλαδή η Ελλάδα, η Ιταλία και η Ισπανία παρουσιάζουν χαμηλότερο Καθαρό και Ακαθάριστο οικογενειακό εισόδημα σε σύγκριση με τις υπόλοιπες χώρες. Οι Βόρειες χώρες, δηλαδή η Δανία, η Σουηδία και η Ολλανδία φαίνεται να δημιουργούν μια δεύτερη ομάδα με παρεμφερές, σχετικά υψηλό Καθαρό και Ακαθάριστο οικογενειακό εισόδημα. Τέλος, οι Ηπειρωτικές χώρες, δηλαδή η Αυστρία, το Βέλγιο, η Γαλλία, η Γερμανία και η Ελβετία ανήκουν σε μια τρίτη ομάδα εξαιτίας του αρκετά υψηλού οικογενειακού εισοδήματος των κατοίκων.

Στην συνέχεια, μελετήθηκαν τα στοιχεία που συνθέτουν την ψυχολογία των κατοίκων και με χρήση της μεθόδου «Πολλαπλή Ανάλυση Αντιστοιχιών» ομαδοποιήθηκαν οι χώρες που παρουσίαζαν ομοιότητες. Οι μετρήσεις των χαρακτηριστικών που χρησιμοποιήθηκαν αφορούσαν την οικονομική αυτοπεποίθηση, την κατάσταση της υγείας, την ικανοποίηση από τη ζωή και την ύπαρξη ή μη συμπτωμάτων κακής ψυχικής υγείας όλων των ατόμων του δείγματος. Οι χώρες που παρουσίασαν καλή ποιότητα αναπαράστασης και κατ' επέκταση μπορούσαν να ομαδοποιηθούν ήταν η Δανία, η Ελβετία, η Ελλάδα, η Ισπανία, η Ιταλία και η Ολλανδία.

Το γράφημα «Joint» που απεικονίζει τις προβολές των κατηγοριών των πέντε μεταβλητών στους δύο άξονες και οδηγεί στον σχετικό διαχωρισμό των χωρών είναι το ακόλουθο:

**ΣΧΗΜΑ 4.9.1**

Ομαδοποίηση χωρών με Πολλαπλή Ανάλυση Αντιστοιχιών




Από την Πολλαπλή Ανάλυση Αντιστοιχιών προέκυψαν δύο ομάδες. Στην πρώτη ομάδα ανήκουν η Δανία, η Ελβετία και η Ολλανδία. Παρατηρείται ότι στις χώρες αυτές, οι κάτοικοι παρουσιάζουν καλή σωματική υγεία και δεν εμφανίζουν σοβαρά συμπτώματα κακής ψυχικής υγείας, δηλώνουν πολύ ικανοποιημένοι από τη ζωή και «τα βγάζουν πέρα» εύκολα. Στην δεύτερη ομάδα ανήκουν η Ελλάδα, η Ισπανία και η Ιταλία. Αντίθετα από την πρώτη ομάδα, οι κάτοικοι των χωρών της δεύτερης ομάδας παρουσιάζουν μέτρια σωματική υγεία και δεν εμφανίζουν σοβαρά συμπτώματα κακής ψυχικής υγείας, δηλώνουν σχετικά ικανοποιημένοι από τη ζωή και «τα βγάζουν πέρα» με σχετική δυσκολία.

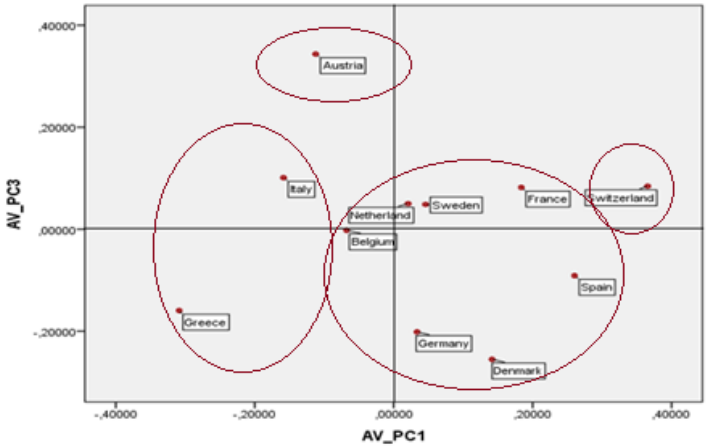
Αξίζει να παρατηρηθεί πως αν και η παραπάνω ομαδοποίηση βασίστηκε σε έξι χώρες προσεγγίζει αρκετά την ομαδοποίηση που προέκυψε λόγω του Ακαθάριστου και Καθαρού οικογενειακού εισοδήματος.

Τέλος, συγκεντρώθηκαν οι μετρήσεις όλων των οικονομικών χαρακτηριστικών της έρευνας και με «Ανάλυση κύριων συνιστωσών» κατασκευάσθηκαν τρεις κύριες συνιστώσες (ΠΙΝΑΚΑΣ 4.7.4). Η πρώτη κύρια συνιστώσα αφορά στο ακαθάριστο και καθαρό οικογενειακό εισόδημα καθώς και στο οικογενειακό εισόδημα από πηγές εκτός εργασίας και σύνταξης. Η δεύτερη κύρια συνιστώσα αφορά στην συνολική αξία της ιδιόκτητης κατοικίας, των παγίων περιουσιακών στοιχείων και άλλης ακίνητης περιουσίας του νοικοκυριού και στο οικογενειακό εισόδημα από εργασία καθώς και στη συνολική αξία των χρηματικών περιουσιακών στοιχείων του νοικοκυριού. Τέλος, η τρίτη κύρια συνιστώσα αφορά στο

οικογενειακό εισόδημα από σύνταξη και εργασία. Οι ευρωπαϊκές χώρες ομαδοποιηθήκαν με βάση τα οικονομικά τους στοιχεία, χρησιμοποιώντας τις κύριες συνιστώσες. Η ομαδοποίηση τους προέκυψε αρχικά με γραφική απεικόνιση των χωρών και στη συνέχεια με τη μη Ιεραρχική μέθοδο «K-Means». Ο διαχωρισμός των χωρών έγινε σε τέσσερεις ομάδες και απεικονίζεται γραφικά στα παρακάτω σχήματα:

**ΣΧΗΜΑ 4.9.2**

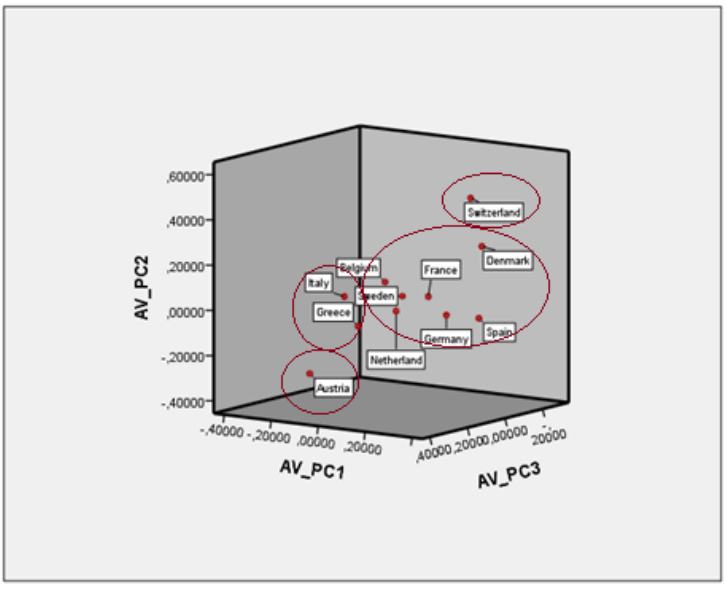
Ομαδοποίηση χωρών με μέθοδο κύριων συνιστωσών




**ΣΧΗΜΑ 4.9.3**

Ομαδοποίηση χωρών με PCA-

Τρισδιάστατο διάγραμμα διασποράς



Από την Ανάλυση σε Κύριες Συνιστώσες προέκυψαν τέσσερεις ομάδες. Η Αυστρία αποτελεί μία ομάδα μόνη της. Σε αυτή την χώρα, οι κάτοικοι χαρακτηρίζονται από μέτριο ακαθάριστο και καθαρό εισόδημα καθώς και οικογενειακό εισόδημα από πηγές εκτός εργασίας και σύνταξης, χαμηλή συνολική αξία των περιουσιακών στοιχείων του νοικοκυριού και υψηλό οικογενειακό εισόδημα από σύνταξη και εργασία. Η Ελβετία είναι μια δεύτερη ομάδα. Οι κάτοικοι της Ελβετίας χαρακτηρίζονται από υψηλό ακαθάριστο και καθαρό οικογενειακό εισόδημα, οικογενειακό εισόδημα από σύνταξη, εργασία και άλλες πηγές αλλά και από υψηλά περιουσιακά στοιχεία. Η τρίτη ομάδα αποτελείται από την Ελλάδα και την Ιταλία. Στις χώρες αυτές, οι κάτοικοι χαρακτηρίζονται από χαμηλό ακαθάριστο και καθαρό οικογενειακό εισόδημα, χαμηλό οικογενειακό εισόδημα ανεξαρτήτως πηγής αλλά και από χαμηλά περιουσιακά στοιχεία. Η τέταρτη ομάδα συμπεριλαμβάνει το Βέλγιο, τη Γαλλία, τη Γερμανία, τη Δανία, την Ισπανία, την Ολλανδία και τη Σουηδία. Στις χώρες αυτές δεν παρουσιάζονται ακραίες τιμές ως προς το οικογενειακό εισόδημα. Παρατηρείται πως τόσο το εισόδημα από εργασία, σύνταξη και άλλες πηγές, όσο και τα περιουσιακά στοιχεία των κατοίκων χαρακτηρίζονται μέτρια και σε κάποιες χώρες σχετικά υψηλά. Η τελευταία ομάδα αποτελεί το «κέντρο βάσης» γύρω από το οποίο αυξάνονται ή μειώνονται οι οικονομικές τιμές των υπόλοιπων χωρών.

# ACKNOWLEDGEMENT

*"This paper uses data from SHARE wave 5 release 1.0.0, as of March 31st 2015 (DOI: 10.6103/SHARE.w5.100) or SHARE wave 4 release 1.1.1, as of March 28th 2013 (DOI: 10.6103/SHARE.w4.111) or SHARE waves 1 and 2 release 2.6.0, as of November 29th 2013 (DOIs: 10.6103/SHARE.w1.260 and 10.6103/SHARE.w2.260) or SHARELIFE release 1.0.0, as of November 24th 2010 (DOI: 10.6103/SHARE.w3.100). The SHARE data collection has been primarily funded by the European Commission through the 5th Framework Programme (project QLK6-CT2001-00360 in the thematic programme Quality of Life), through the 6th Framework Programme (projects SHARE-I3, RII-CT-2006-062193, COMPARE, CIT5- CT-2005-028857, and SHARELIFE, CIT4-CT-2006-028812) and through the 7th Framework Programme (SHARE-PREP, N° 211909, SHARE-LEAP, N° 227822 and SHARE M4, N° 261982). Additional funding from the U.S. National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, R21 AG025169, Y1-AG-4553-01, IAG BSR06-11 and OGHA 04-064) and the German Ministry of Education and Research as well as from various national sources is gratefully acknowledged (see www.share-project.org for a full list of funding institutions)."*

# ΒΙΒΛΙΟΓΡΑΦΙΑ

## Ελληνική Βιβλιογραφία


Καρλής, Δ. (2005). *Πολυμεταβλητή στατιστική ανάλυση*, Εκδόσεις Σταμούλης, Αθήνα.

Κούτρας, Μ. (2008). *Εφαρμοσμένη Πολυμεταβλητή Ανάλυση-Ανάλυση κατά συστάδες*, Πανεπιστήμιο Πειραιώς, Πειραιάς.

Πανάρετος, Ι. και Ξεκαλάκη, Ε. (1995). *Εισαγωγή στην Πολυμεταβλητή Ανάλυση*, Εκδόσεις Πανάρετος Ι., Αθήνα.

Παπαδημητρίου, Γ. (2007). *Η ανάλυση δεδομένων: Παραγοντική Ανάλυση των αντιστοιχιών ιεραρχική ταξινόμηση και άλλες μέθοδοι*, Εκδόσεις Τυπωθήτω, Αθήνα.

Σιάρδος, Γ. (2005). *Μέθοδοι πολυμεταβλητής στατιστικής ανάλυσης: με την επίλυση ασκήσεων μέσω του στατιστικού προγράμματος SPSS*, Εκδόσεις Σταμούλης.

Τήνιος, Π. (2009). *Ζωή 50+: υγεία, γήρανση και σύνταξη στην Ελλάδα και στην Ευρώπη*, Εκδόσεις Κριτική, Αθήνα.


## Ξένη Βιβλιογραφία


Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis, third edition*, John Wiley & Sons, Inc.

Andrews, David F. (1972). Plots of High-Dimensional Data, *International Biometric Society*, **18**, 125-136.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth.

Chernoff, Herman (1973). The Use of Faces to Represent Points in K-Dimensional Space Graphically, *Journal of the American Statistical Association*, **68**, 361-368.

Galton, F. (1889). *Natural Inheritance*, MacMillan, London.

Girshick, M.A. (1939). On the Sampling Theory of Roots of Determinantal Equations, *The Annals of Mathematical Statistics*, **10**, 203-224.

Greenacre, M. (2007). *Correspondence analysis in practice, second edition*, Chapman & Hall/CRC.

Greenacre, M., Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC.

Jackson, E. (1991). *A User's Guide to Principal Components*, John Wiley & Sons, Inc., New York.

Johnson, R.A. and Wichern, D.W. (1998). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Manly, Bryan F.J. (1994). *Multivariate Statistical Methods: a primer, second edition*, Chapman & Hall/CRC, U.S. of America.

Paul E.Green and J.Douglas Carroll (1976). *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York.

Pearson, K. (1894). Contribution to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society of London*, **A 185**, 71-110.

Wilks, S.S. (1932). Certain Generalizations in the Analysis of Variance, *Biometrika*, **24**, 471-494.

Yules, G. U. (1896). On the Significance of Bravais' Formulae for Regression, &c., in the Case of Skew Correlation, *Royal Society of London*, **60**, 477-489.

Yule, G. U. (1907). On the Theory of Correlation for any Number of Variables, Treated by a New System of Notation, *Royal Society of London*, **A 79**, 182-193.