ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«ΑΝΑΛΥΣΗ ΠΟΛΥΜΕΤΑΒΛΗΤΩΝ ΔΕΔΟΜΕΝΩΝ»

Πολυσόπουλος Δ. Χρήστος

Α.Μ –33107064

Καρλόβασι ,Ιούνιος 2013

Επιβλέπων καθηγητής: **Γεωργίου Στέλιος**
Αναπληρωτής Καθηγητής Σ.Α.Χ.Μ.

Μέλη τριμελούς επιτροπής:

Γεωργίου Στέλιος
Αναπληρωτής Καθηγητής Σ.Α.Χ.Μ.

Στυλιανού Στέλλα
Επίκουρη Καθηγήτρια Σ.Α.Χ.Μ.

Ζήμερας Στυλιανός
Επίκουρος Καθηγητής Σ.Α.Χ.Μ.

Καρλόβασι , Ιούνιος 2013

Πρόλογος

Οι μέθοδοι της πολυμεταβλητής στατιστικής ανάλυσης, όπως φανερώνει και η ονομασία τους, αναφέρονται σε διαδικασίες και μεθοδολογίες όπου προσπαθούμε να καταλήξουμε σε στατιστική συμπερασματολογία με την χρήση πολλών μεταβλητών. Στην πράξη τα δεδομένα ενός ερευνητή είναι από την φύση τους πολυμεταβλητά και ο σκοπός του ερευνητή σπάνια είναι να μελετήσει μια μεταβλητή ανεξάρτητα και απομονωμένα από τις υπόλοιπες. Συνεπώς, ουσιαστικά όλες οι στατιστικές μέθοδοι είναι από την φύση τους πολυμεταβλητές, ή τουλάχιστον τα δεδομένα που έχει ένας ερευνητής στη διάθεση του είναι σχεδόν πάντα πολυμεταβλητά και εξαρτάται πια από εκείνον το κατά πόσο θέλει να χρησιμοποιήσει όλα του τα δεδομένα για να αποκομίσει τη μεγαλύτερη πληροφορία από αυτά.

Από τα παραπάνω μπορεί να παρατηρήσει κανείς πως οι πολυμεταβλητές τεχνικές δεν αναπτύχθηκαν ξεχωριστά από τις μονομεταβλητές τεχνικές. Ο βασικός λόγος που δεν είναι κάποιες από αυτές τόσο διαδεδομένες έχει να κάνει κυρίως με την πολυπλοκότητα τους που οδήγησε σε σοβαρούς περιορισμούς στην πρακτική τους εφαρμογή. Κάτι τέτοιο δεν ισχύει πια με τη γενικευμένη χρήση υπολογιστών στη στατιστική, καθώς υπάρχει πάντα μια ποικιλία στατιστικών πακέτων που μπορούν να χρησιμοποιηθούν ακόμα και για ιδιαίτερα πολύπλοκες μεθόδους.

Οι λόγοι για τους οποίους οι πολυμεταβλητές τεχνικές είναι ιδιαίτερα χρήσιμες είναι οι παρακάτω:

- *Έχουμε περισσότερη πληροφορία* (περισσότερες μεταβλητές ερμηνεύουν καλύτερα ένα φαινόμενο). Συνήθως ο ερευνητής σκοπεύει με τα δεδομένα που έχει στα χέρια του να περιγράψει ή να ερμηνεύσει κάποιο φαινόμενο ή μηχανισμό. Είναι ευνόητο ότι όσο περισσότερη πληροφορία έχει κανείς, τόσο περισσότερο μπορεί να περιορίσει την αβεβαιότητά του και επομένως να εξαγάγει συμπεράσματα με μεγαλύτερη βαρύτητα.
- *Μελετάμε συσχετισμούς μεταξύ μεταβλητών και υποκειμένων*. Ο κόσμος μέσα στον οποίο ζούμε είναι ένας κόσμος γεμάτος από συσχετίσεις μεταξύ διαφορετικών πραγμάτων και οντοτήτων και θα ήταν απλοϊκό να τον μελετά κανείς, χωρίς να τις λαμβάνει υπόψη του. Από την άλλη, η ανακάλυψη τέτοιων συσχετίσεων ανάμεσα σε διαφορετικές μεταβλητές μπορεί από μόνη της να οδηγήσει σε καινούργιες ερμηνείες για τα υπό μελέτη φαινόμενα. Επομένως, μοιάζει καλή ιδέα να μελετήσουμε συγχρόνως ένα σύνολο μεταβλητών με σκοπό να αντλήσουμε όσο γίνεται περισσότερα από τα δεδομένα μας.

  Οι στόχοι των επιστημονικών ερευνών για τις οποίες οι πολυμεταβλητές μέθοδοι είναι εκ ΄φύσεως πιο κατάλληλες περιλαμβάνουν τα ακόλουθα:

1. ***Μείωση των δεδομένων ή δομική απλούστευση***. Το φαινόμενο που μελετάται παρουσιάζεται όσο πιο απλά γίνεται χωρίς να θυσιάζονται πολύτιμες πληροφορίες. Σκοπός είναι να να καταστήσει ευκολότερη την ερμηνεία.

2. ***Ταξινόμηση και ομαδοποίηση***. Δημιουργούνται ομάδες «παρόμοιων» αντικειμένων ή μεταβλητών, βασισμένες στα μετρηθέντα χαρακτηριστικά. Εναλλακτικά, μπορεί να απαιτούνται κανόνες για την ταξινόμηση των αντικειμένων σε πιο καλά καθορισμένες ομάδες.
3. ***Έρευνα της εξάρτησης μεταξύ των μεταβλητών***. Η φύση των σχέσεων μεταξύ των μεταβλητών είναι ένα βασικό ζήτημα. Είναι όλες οι μεταβλητές ανεξάρτητες μεταξύ τους ή είναι μία ή περισσότερες μεταβλητές εξαρτημένες από τις υπόλοιπες; Αν ναι, πώς;
4. ***Πρόβλεψη.*** Οι σχέσεις μεταξύ των μεταβλητών θα πρέπει να προσδιορίζονται με σκοπό την πρόβλεψη τιμών μίας ή περισσοτέρων μεταβλητών βάσει των παρατηρήσεων που προέρχονται από τις υπόλοιπες μεταβλητές.
5. ***Κατασκευή υποθέσεων και δοκιμές***. Εδώ διατυπώνονται ειδικές στατιστικές υποθέσεις όσον αφορά τις παραμέτρους των πολυμεταβλητών πληθυσμών και δοκιμάζονται. Αυτό γίνεται με σκοπό να επικυρώσει εικασίες ή να ενισχύσει ήδη υπάρχουσες πεποιθήσεις.

Σε αυτή την λίστα με τους στόχους της πολυμεταβλητής ανάλυσης θα μπορούσε κανείς να προσθέσει και άλλα σημεία, όπως για παράδειγμα την γραφική αναπαράσταση των δεδομένων (data visualization). Επίσης, πολλές φορές οι προσωπικοί σκοποί του ερευνητή είναι εξίσου σημαντικοί, αν και δεν μπορούν να μπουν σε αυτή την λίστα.

Table of Contents

Κεφάλαιο 1

1.Εισαγωγή

Για πρώτη φορά η Ελλάδα βρίσκεται στην δίνη μίας διεθνούς οικονομικής κρίσης. Διανύουμε μία δύσκολη εποχή. Η ακρίβεια εξελίχθηκε σε μείζον κοινωνικό πρόβλημα και η ανεργία απειλεί σοβαρά τις κοινωνίες. Τα πραγματικά όμως προβλήματα της κρίσης τα οποία επεκτάθηκαν και στο σύνολο των πολιτών, είναι η άνοδος των επιτοκίων, η δύσκολη λήψη δανείων, η άνοδος των τιμών των εμπορευμάτων και των καυσίμων που επέφεραν ακρίβεια και μείωση της αγοραστικής δύναμης των καταναλωτών. Η ελληνική οικονομία, βρίσκεται παγιδευμένη ανάμεσα στην οικονομική ύφεση και την δημοσιονομική κατάρρευση. (Κουφάρης, 2010)

Αντικείμενο της έρευνας

Συμπεριλαμβανομένων των παραπάνω, μπορούμε να δεχθούμε ότι ένα αδιαμφισβήτητο γεγονός είναι ότι η κρίση έχει επηρεάσει τις Ελληνικές οικογένειες και ως άμεσο αποτέλεσμα τους φοιτητές. Αντικείμενο της παρούσας έρευνας είναι η διερεύνηση του αντίκτυπου της οικονομικής κρίσης στην Ελλάδα τα τελευταία τρία χρόνια στην ποιότητα ζωής των φοιτητών που σπουδάζουν στα τρία τμήματα της σχολής στο Καρλόβασι της Σάμου. Η εξαγωγή συμπερασμάτων θα γίνει μέσω πολυμεταβλητών τεχνικών, ερμηνεύοντας προφίλ ομάδων φοιτητών, συσχετίσεις ανάμεσα σε μεταβλητές και την επίδραση τους στην ποιότητα ζωής των φοιτητών.

Συλλογή δεδομένων και ανάλυση

Τα δεδομένα τα οποία είναι διαθέσιμα συλλέχθηκαν με την συμπλήρωση ερωτηματολογίων από 200 φοιτητές οι οποίοι είναι πάνω από το τρίτο ακαδημαϊκο έτος σπουδών τους και διαμένουν μόνιμα στο Καρλόβασι Σάμου. Για την στατιστική ανάλυση των δεδομένων θα χρησιμοποιηθεί το στατιστικο πακέτο IBM SPSS Statistics19.

Περιγραφή των μεταβλητών

Το ερωτηματολόγιο που μοιράστηκε στους φοιτητές αποτελούνταν από τις παρακάτω μεταβλητές:

Ηλικία: Η ηλικία του φοιτητή σε χρόνια.

Προέλευση: Αν προέρχονται απο πόλη, νησί ή χωριό.

Στέγαση: Αν ο φοιτητής ενοικιάζει ή του παρέχεται εστία.

Αποταμίευση: Αν έχει καταφέρει να αποταμιεύσει χρήματα, αν ναι, τι ποσό.

Προ 3 ετών μηνιαίο εισόδημα: Τι ποσό έπαιρναν για την επιβίωση τους πριν από τρία ακαδημαϊκά χρόνια.

Τωρινό μηνιαίο εισόδημα: Τι ποσό παίρνουν κατά την διάρκεια της έρευνας.

Βραδυνή διασκέδαση: Πόσες φορές βγαίνουν για βραδυνή διασκέδαση τον μήνα.

Κλίμακα μείωσης: Από μία κλίμακα απο το 1 έως το 5 όπου το 1 δηλώνει ‘’καθόλου’’, το 2 ‘’λίγο’’ το 3 ‘’μέτρια’’ το 4 ‘’υψηλή μείωση’’ και το 5 ‘’πολύ υψηλή μείωση’’ να χαρακτηρίσουν την μείωση που έκαναν στις αγορές αγαθών πολυτελείας.

Κλίμακα ποιότητας ζωής: Από μία κλίμακα από το 1 έως το 5, όπου το 1 δηλώνει ‘’πολύ χαμηλή’’, το 2 ‘’χαμηλή’’, το 3 ‘’μέτρια’’, το 4 ‘’υψηλή’’ και το 5 ‘’πολύ υψηλή’’ να χαρακτηρίσει ο κάθε φοιτητής την ποιότητα ζωής του στο Καρλόβασι από οικονομική άποψη.

Κεφάλαιο 2

Περιγραφική Στατιστική

2.1 Εισαγωγή

Ο βασικός στόχος της περιγραφικής Στατιστικής είναι η ανάπτυξη μεθόδων για την συνοπτική και την αποτελεσματική παρουσίαση των δεδομένων. Γι'αυτό τον λόγο έχουν αναπτυχθεί διάφορες μεθόδοι γραφικής παρουσίασης των δεδομένων και αριθμητικών περιγραφικών μέτρων. Η επιλογή των κατάλληλων μεθόδων γίνεται με βάση τον τύπο της μεταβλητής που θέλουμε να παρουσιάσουμε.

2.2 Ποιοτικές Μεταβλητές

Τα δεδομένα μας αποτελούνται από 4 ποιοτικές μεταβλητές. Η μεταβλητή ''Προέλευση'' που αποτελείται από τρία επίπεδα, χωριό, πόλη νησί και η μεταβλητή ''Στέγαση'' που έχει 2 επίπεδα, Εστία, Ενοίκιο ανήκουν στην υποκατηγορία των κατηγορικών μεταβλητών. Οι μεταβλητες ''Κλίμακα μείωσης'' και ''Κλίμακα Ποιότητας Ζωής'' ανήκουν στις διατάξιμες ως κλίμακες Likert που μετρούν έναν χαρακτηρισμό από το 1 έως 5.

Πίνακας 2.1

origin

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Πόλη	140	70,0	70,0	70,0
	Νησί	19	9,5	9,5	79,5
	Χωριό	41	20,5	20,5	100,0
	Total	200	100,0	100,0	

Όπως διακρίνουμε το 70% των φοιτητών που συμπλήρωσαν τα ερωτηματολόγια προέρχονταν από Πόλη, το 9.5% από Νησί και το 20.5% από Χωριό. Συνολικά έχουμε 200 φοιτητές που πήραν μέρος στην διαδικασία συμπλήρωσης του ερωτηματολογίου.

Πίνακας 2.2

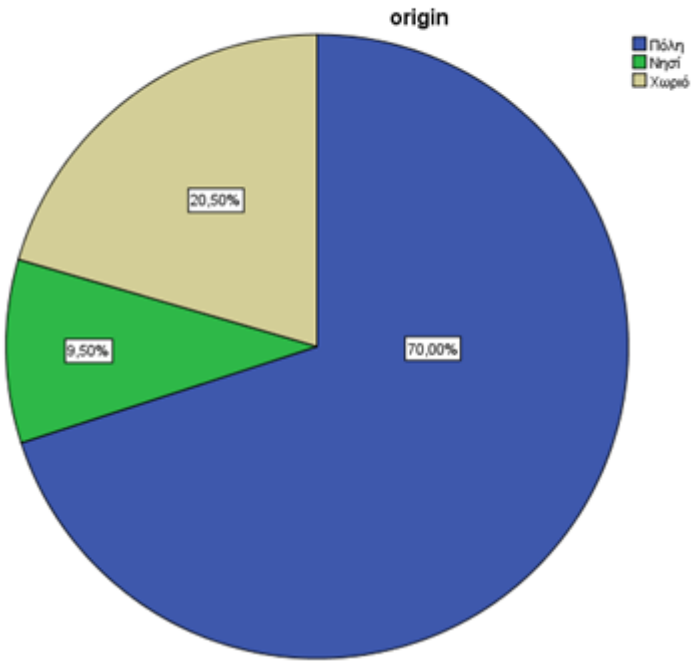
housing

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Εστία	32	16,0	16,0	16,0
	Ενοίκιο	168	84,0	84,0	100,0
	Total	200	100,0	100,0	

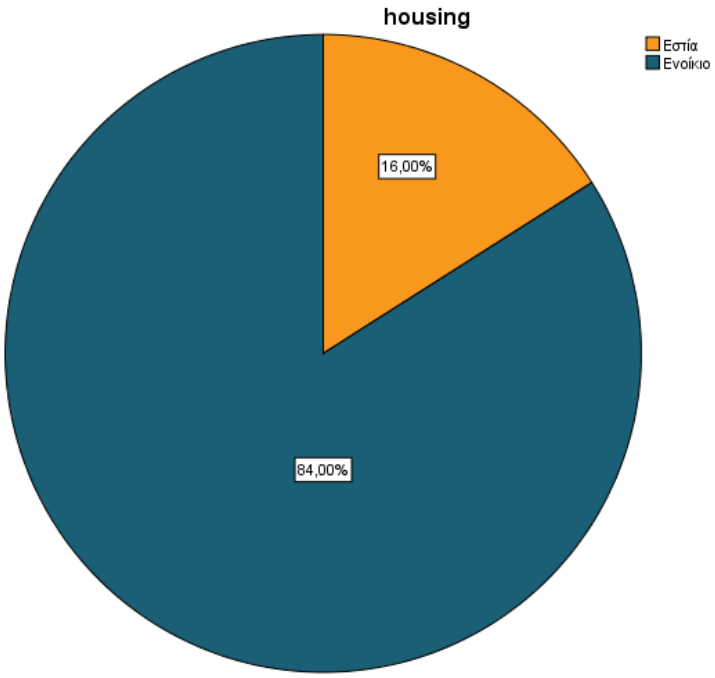
Επίσης, 32 φοιτητές ή 16% τους παρέχεται δωμάτιο στην εστία και οι υπόλοιποι 168 ή 84% νοικιάζουν σπίτι.

Τα αποτελέσματα μπορούμε να τα δούμε και διαγραμματικά κατασκευάζοντας διαγραμμα πίτας.

Διάγραμμα 2.1



Διάγραμμα 2.2



Στην συνέχεια θα δούμε τις συχνότητες για τις διατάξιμες μεταβλητές μας καθώς και τα ραβδογράμματα τους.

Πίνακας 2.3

reduction_likert

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	καθόλου	5	2,5	2,5	2,5
	λίγο	14	7,0	7,0	9,5
	μέτρια	59	29,5	29,5	39,0
	υψηλή μείωση	90	45,0	45,0	84,0
	πολύ υψηλή μείωση	32	16,0	16,0	100,0
	Total	200	100,0	100,0	

Στην ερώτηση να δωθεί ένας χαρακτηρισμός για την μείωση των αγορών σε αγαθά πολυτελείας βλέπουμε ότι μόνο το 2.5% απάντησε ότι δεν έχει μειώσει καθόλου τις αγορές του. Μεγάλες συχνότητες συναντάμε στην μέτρια μείωση και υψηλή με 29.5% και 45% αντίστοιχα. Πολύ υψηλές μειώσεις έχουν αναγκαστεί να κάνουν 32 φοιτητές ή το 16% και έχουν μειώσει λίγο τις αγορές τους το 7%.

Πίνακας 2.4

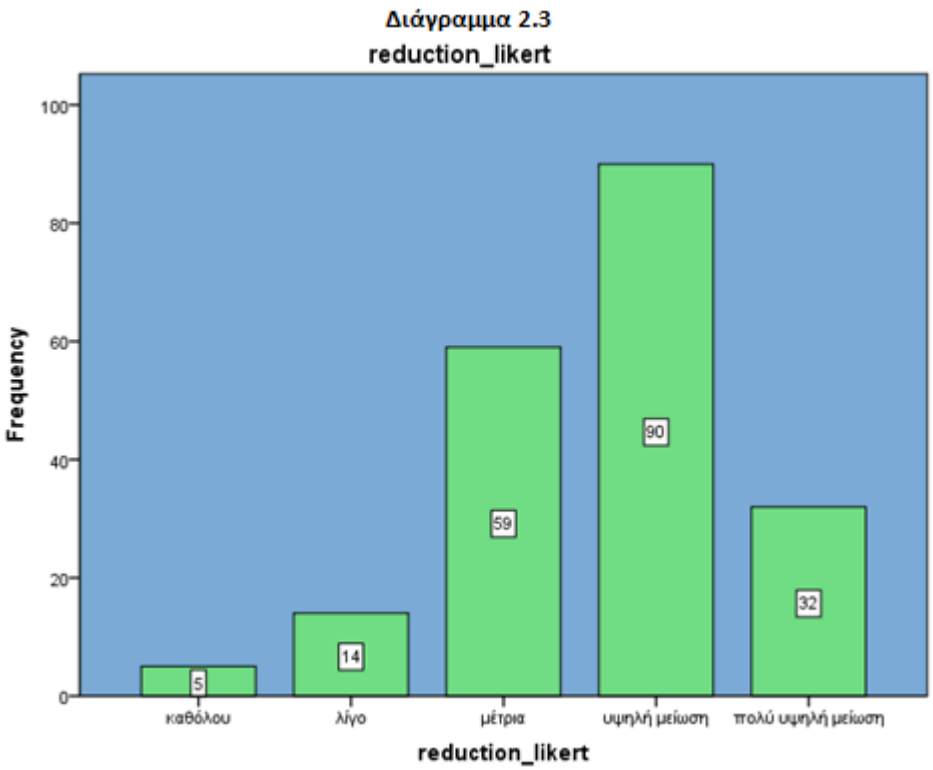
quality_likert

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	πολύ χαμηλή	8	4,0	4,0	4,0
	χαμηλή	21	10,5	10,5	14,5
	μέτρια	76	38,0	38,0	52,5
	υψηλή	87	43,5	43,5	96,0
	πολύ υψηλή	8	4,0	4,0	100,0
	Total	200	100,0	100,0	

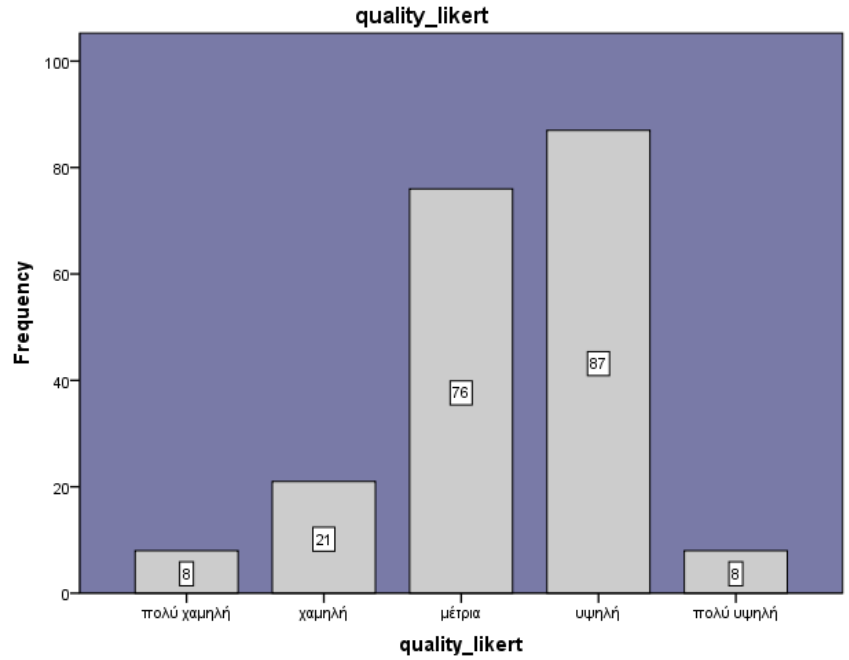
Στον χαρακτηρισμό της ποιότητας ζωής το 43.5% απάντησε ‘’υψηλή’’ ενώ τα δύο άκρα ‘’Πολύ χαμηλή’’ και ‘’Πολύ υψηλή’’ αποτελείται από 4% και 4% αντίστοιχα. Στον χαρακτηρισμό ‘’Χαμηλή’’ έχουμε το 10.5% των φοιτητών και στη ‘’Μέτρια’’ 38%.

Για να οπτικοποιήσουμε τις συχνότητες των μεταβλητών κατακευάσαμε δύο διαφορετικά ραβδογράμματα.

Διάγραμμα 2.3



Διάγραμμα 2.4



Έλεγχος ανεξαρτησίας X^2 του Pearson

Για να ελέγξουμε την ανεξαρτησία μεταξύ δύο κατηγορικών μεταβλητών χρησιμοποιούμε τον έλεγχο ανεξαρτησίας του Pearson. Σκοπός μας είναι να εξετάσουμε αν σχετίζονται οι μεταβλητές ‘’Στέγαση’’ και ‘’Κλίμακα Ποιότητας Ζωής’’.

Πίνακας 2.5

housing * quality_likert Crosstabulation

Count

		quality_likert					Total
		πολύ χαμηλή	χαμηλή	μέτρια	υψηλή	πολύ υψηλή	
housing	Εστία	2	5	17	7	1	32
	Ενοίκιο	6	16	59	80	7	168
Total		8	21	76	87	8	200

Πίνακας 2.6

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,897[a]	4	,095
Likelihood Ratio	8,296	4	,081
Linear-by-Linear Association	5,496	1	,019
N of Valid Cases	200		

a. 3 cells (30,0%) have expected count less than 5. The minimum expected count is 1,28.

Όπως βλέπουμε το παρατηρούμενο επίπεδο σημαντικότητας είναι μεγαλύτερο του 0.05 σε επίπεδο σημαντικότητας 5% και έτσι αποδεχόμαστε την μηδενική υπόθεση που δηλώνει ότι δεν υπάρχει συσχέτιση μεταξύ του να μένει κάποιος στην εστία και στην ποιότητα ζωής του.

Ένας ακόμα έλεγχος θα γίνει μεταξύ της ‘’Στέγασης’’ και της ‘’Μείωσης των αγορών’’

Πίνακας 2.7

housing * reduction_likert Crosstabulation

Count

		reduction_likert					Total
		καθόλου	λίγο	μέτρια	υψηλή μείωση	πολύ υψηλή μείωση	
housing	Εστία	0	5	4	13	10	32
	Ενοίκιο	5	9	55	77	22	168
Total		5	14	59	90	32	200

Πίνακας 2.8

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	14,432[a]	4	,006
Likelihood Ratio	14,286	4	,006
Linear-by-Linear Association	2,291	1	,130
N of Valid Cases	200		

a. 3 cells (30,0%) have expected count less than 5. The minimum expected count is ,80.

Σε αυτή την περίπτωση το p-value είναι μικρότερο του 0.05 οπότε οι δύο μεταβλητές είναι εξαρτημένες μεταξύ τους.

2.3 Ποσοτικές μεταβλητές

Το δεδομένα μας περιέχουν επίσης και 5 ποσοτικές μεταβλητές, που είναι η ηλικία του ερωτώμενου,τωρινό εισόδημα, εισόδημα πριν 3 ακαδημαικά έτη, το ποσό που έχει αποταμιεύσει και τον αριθμό βραδυνών εξόδων για διασκέδαση.

Πίνακας 2.9

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean
age in years	200	7	20	27	22,42
Bank_savings	200	4500	0	4500	131,48
Montlhy income 3 years ago	200	900	100	1000	402,50
Monthly Current income	200	800	0	800	320,50
Monthly nightlife	200	19	1	20	5,00
Valid N (listwise)	200				

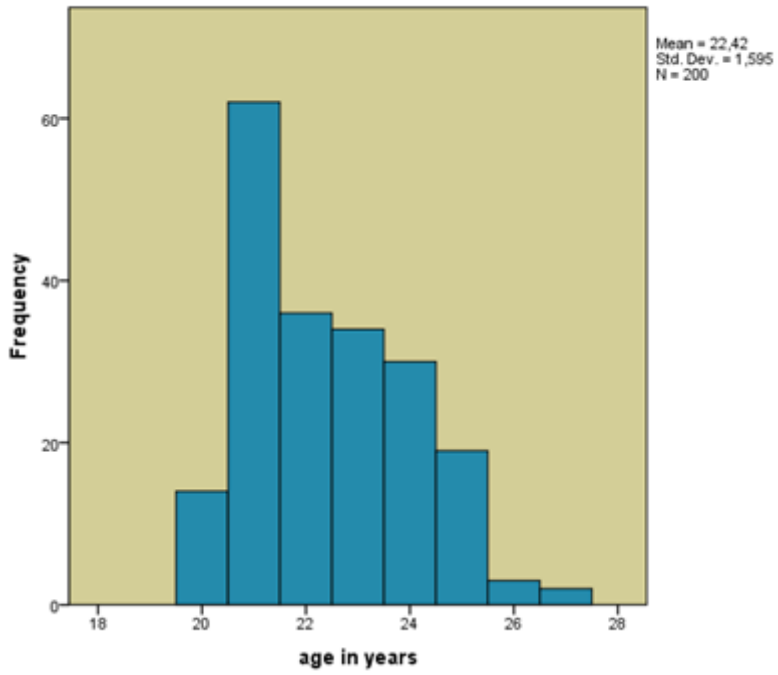
Ο παραπάνω πίνακας μας δίνει αρκετά στοιχεία για τα χαρακτηριστικά των φοιτητών που συμπλήρωσαν τα ερωτηματολόγια. Η μέση ηλικία των ερωτώμενων είναι 22.42 χρόνια, ενώ η μικρότερη και η μεγαλύτερη ηλικία που συναντάται είναι 20 και 27 χρόνια αντίστοιχα. Η μέση τιμή των αποταμιευμένων ποσών είναι 131.48 με μέγιστο τα 4500 ευρώ και ελάχιστο το 0 αφού κάποιοι δεν κατάφεραν να κάνουν οικονομία.

Επίσης παρατηρούμε ότι το μέσο ποσό χρημάτων που έπαιρναν οι φοιτητές μειώθηκε κατά 82 ευρώ τα τελευταία 3 χρόνια. Παρατηρούμε ότι η ελάχιστη τιμή του τωρινού εισοδήματος είναι 0 , πράγμα που σημαίνει δεν υπάρχει καμία οικονομική βοήθεια και ορισμένοι φοιτητές αναγκάζονται να δουλεύουν για να τα βγάλουν πέρα.

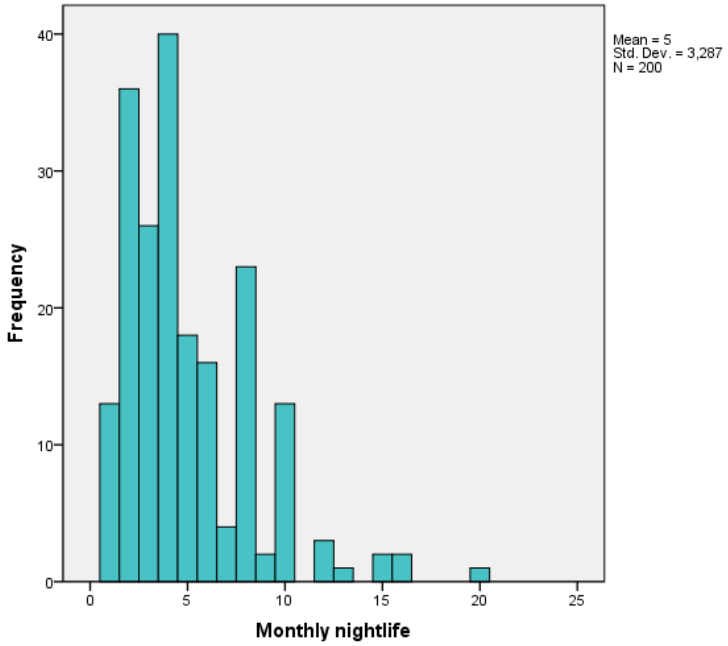
Οι φοιτητές για βραδινή διασκέδαση βγαίνουν κατά μέσο όρο 5 φορές τον μήνα με τιμές που κυμαίνονται από 1 φορά μέχρι και 20.

Παρακάτω ακολουθούν τα διαγράμματα για τις ποσοτικές μεταβλητές.

Διάγραμμα 2.5



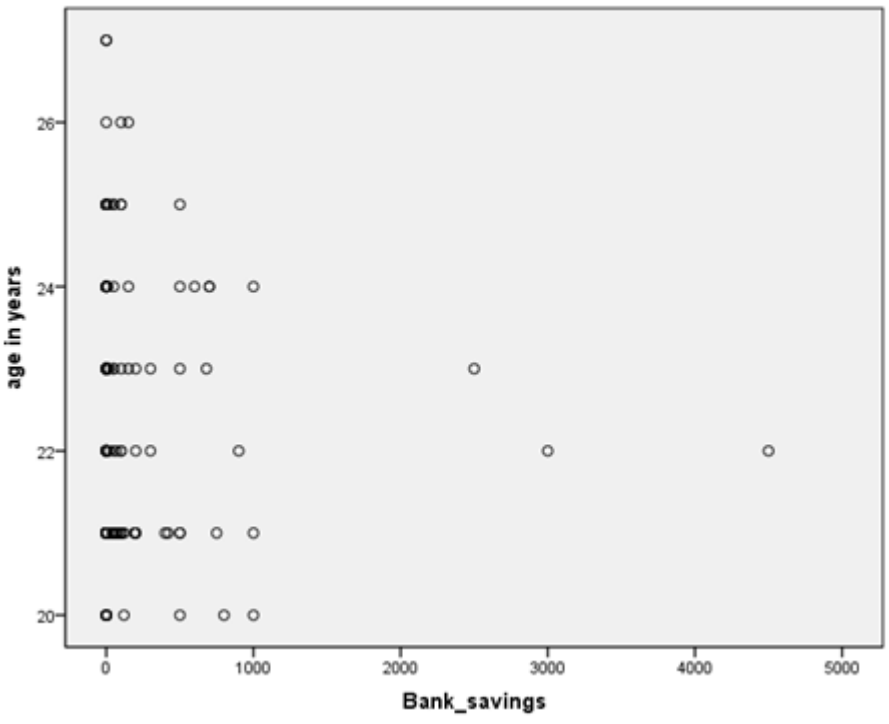
Διάγραμμα 2.6



Η επικρατούσα τιμή της ηλικίας είναι τα 21 χρόνια ενώ η τιμή που συναντάται περισσότερες φορές στον αριθμό των βραδινών εξόδων είναι οι 4 φορές.

Θα συνεχίσουμε με 2 διαγράμματα σημείων που μας δείχνουν αν υπάρχει γραμμική σχέση μεταξύ 2 μεταβλητών.

Διάγραμμα 2.7



Δεν παρατηρούμε καμία γραμμικότητα μεταξύ της ηλικίας και των αποταμιευμένων χρημάτων.

Έλεγχος ύπαρξης γραμμικής συσχέτισης

Θα χρησιμοποιήσουμε τον έλεγχο του Pearson για να καθορίσουμε άν υπάρχει συσχέτιση μεταξύ των ποσοτικών μεταβλητών μας. Εξετάζουμε την μηδενική υπόθεση ότι η συσχέτιση είναι 0 έναντι της εναλλακτικής ότι είναι διάφορη του 0.

Πινακας 2.10

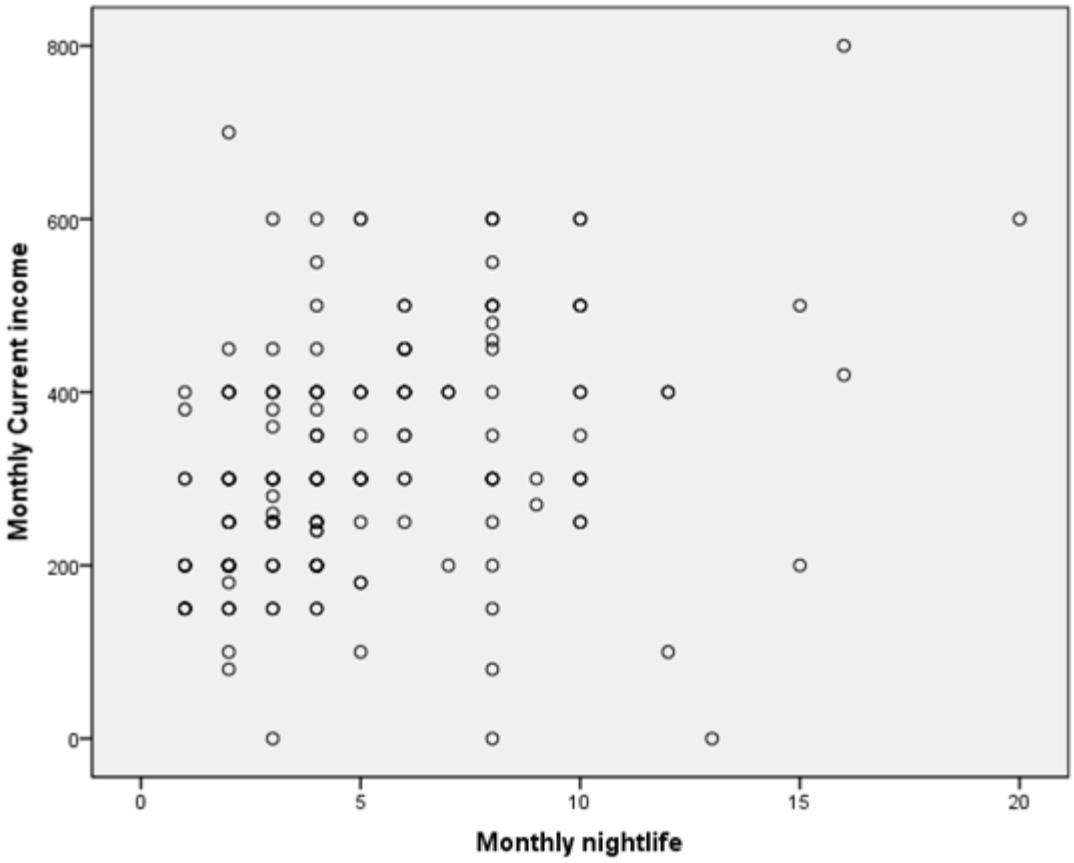
Correlations

		Bank_savings	Monthly Current income	Monthly nightlife	age in years
Bank_savings	Pearson Correlation	1	,104	,044	-,039
	Sig. (2-tailed)		,144	,541	,586
	N	200	200	200	200
Monthly Current income	Pearson Correlation	,104	1	,359**	,058
	Sig. (2-tailed)	,144		,000	,412
	N	200	200	200	200
Monthly nightlife	Pearson Correlation	,044	,359**	1	-,031
	Sig. (2-tailed)	,541	,000		,666
	N	200	200	200	200
age in years	Pearson Correlation	-,039	,058	-,031	1
	Sig. (2-tailed)	,586	,412	,666	
	N	200	200	200	200

**. Correlation is significant at the 0.01 level (2-tailed).

Παρατηρούμε ότι η μόνη γραμμική συσχέτιση βρίσκεται μεταξύ των χρημάτων που λαμβάνουν οι φοιτητές και του αριθμού των φορών που βγαίνουν για βραδυνή διασκέδαση. Ο συντελεστής του Pearson είναι της τάξεως του 35.9% που δηλώνει μία ασθενής θετική συσχέτιση. Το αποτέλεσμα αυτο μπορούμε να το διασταυρώσουμε και με το παρακάτω διάγραμμα σημείων.

Διάγραμμα 2.8



Κεφάλαιο 3

Ανάλυση κατά συστάδες

Εισαγωγή 3.1

Η ανάλυση κατά συστάδες είναι μία μέθοδος που έχει σκοπό να κατατάξει σε ομάδες τις υπάρχουσες παρατηρήσεις, χρησιμοποιώντας την πληροφορία που υπάρχει σε κάποιες μεταβλητές. Με άλλα λόγια η ανάλυση κατά συστάδες εξετάζει πόσο όμοιες είναι κάποιες παρατηρήσεις ως προς κάποιον αριθμό μεταβλητών με σκοπό να δημιουργήσει ομάδες από παρατηρήσεις που μοιάζουν μεταξύ τους.

Μία επιτυχημένη ανάλυση θα πρέπει να καταλήξει σε ομάδες για τις οποίες οι παρατηρήσεις μέσα σε κάθε ομάδα να είναι όσο γίνεται πιο ομοιογενείς, αλλά παρατηρήσεις διαφορετικών ομάδων να διαφέρουν όσο γίνεται περισσότερο. Η ανάλυση σε συστάδες βρίσκει πληθώρα εφαρμογών σχεδόν σε κάθε επιστήμη και επομένως αποτελεί ένα πολύτιμο εργαλείο στα χέρια όλων των επιστημονικών κλάδων.

Δύο βασικές έννοιες για την ανάλυση κατά συστάδες, αλλά όχι μόνο, είναι οι έννοιες της απόστασης και της ομοιότητας. Μπορούμε εύκολα να διαπιστώσουμε ότι αυτές οι δύο έννοιες είναι αντίθετες, παρατηρήσεις που είναι θα έχουν μεγάλη ομοιότητα και μικρή απόσταση. Οι έννοιες αυτές ουσιαστικά ποσοτικοποιούν αυτό που στην καθημερινή γλώσσα εννοούν. Δηλαδή παρατηρήσεις που μοιάζουν πολύ μεταξύ τους έχουν με απλά λόγια σχετικά όμοιες τιμές, θα πρέπει να έχουν πολύ μεγάλη τιμή για το μέτρο της ομοιότητας που θα χρησιμοποιήσουμε και πολύ μικρή απόσταση. Οι έννοιες αυτές είναι πολύ χρήσιμες, καθώς μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους και επομένως να τις τοποθετήσουμε στην ίδια ομάδα. Επομένως, σκοπός της ανάλυσης σε συστάδες είναι να δημιουργήσουμε ομάδες μέσα στις οποίες οι παρατηρήσεις απέχουν λίγο, ενώ παρατηρήσεις διαφορετικών ομάδων να απέχουν μεταξύ τους αρκετά.

Οι βασικότερες και πιο διαδεδομένες προσεγγίσεις για την ομαδοποίηση των δεδομένων μας είναι:

- **Ιεραρχικές μέθοδοι:** Ξεκινάμε με κάθε παρατήρηση αν είναι από μόνη της μια ομάδα. Σε κάθε βήμα ενώνουμε τις 2 παρατηρήσεις που έχουν πιο μικρή απόσταση. Αν 2 παρατηρήσεις έχουν ενωθεί σε προηγούμενο βήμα, ενώνουμε μια προϋπάρχουσα ομάδα με μια παρατήρηση μέχρι να φτιάξουμε μια ομάδα. Κοιτώντας τα αποτελέσματα, διαλέγουμε πόσες ομάδες τελικά προκύπτουν.
- **K-Means:** Ο αριθμός των ομάδων είναι γνωστός από πριν. Με έναν επαναληπτικό αλγόριθμο μοιράζουμε τις παρατηρήσεις στις ομάδες ανάλογα με το ποια ομάδα είναι πιο κοντά στην παρατήρηση.
- **Two-Step ομαδοποίηση:** Ο αλγόριθμος της Two-Step Cluster analysis είναι σχεδιασμένος στο να χειρίζεται πολύ μεγάλα σετ δεδομένων. Υπάρχει η δυνατότητα να χρησιμοποιηθούν και συνεχείς και κατηγορικές μεταβλητές. Η

επιλογή του αριθμού των ομάδων μπορεί να γίνει αυτόματα από το SPSS ή να επιλέξουμε εμείς τον αριθμό.

Τελειώνοντας αυτή την εισαγωγή, θα πρέπει να τονιστεί ότι μερικές φορές η ανάλυση σε συστάδες μπορεί να έχει και άλλους σκοπούς εκτός από την απλή ομαδοποίηση των δεδομένων. Έτσι, η ανάλυση σε συστάδες μπορεί να χρησιμοποιηθεί για :

- Να αποκτηθεί κάποια γνώση σχετικά με τα δεδομένα, αν για παράδειγμα παρουσιάζουν ομοιότητες, ποιες μεταβλητές μοιάζουν να έχουν διακριτική ικανότητα κ.λπ.
- Τη διερεύνηση σχέσεων στα δεδομένα, συνήθως έχοντας ένα σετ δεδομένων στα χέρια μας έχουμε μια πολύ ασαφή εικόνα για το τι περιέχουν τα δεδομένα και τι είδους σχέσεις υπάρχουν.
- Τη μείωση των διαστάσεων του προβλήματος. Ειδικά στη σύγχρονη εποχή το πλήθος των δεδομένων που συγκεντρώνεται είναι τεράστιο, χωρίς αυτό να σημαίνει ότι και η πληροφορία που περιέχεται είναι εξίσου τεράστια. Υπάρχουν επικαλύψεις, μεταβλητές χωρίς ιδιαίτερο ενδιαφέρον κ.λπ. Επομένως, ομαδοποιώντας τα δεδομένα αποκτούμε μια εικόνα σχετικά με τις μεταβλητές που παρουσιάζουν ενδιαφέρον και επικεντρωνόμαστε σε αυτές.
- Δημιουργία και έλεγχο υποθέσεων σχετικά με τα δεδομένα. Πολλές φορές ο ερευνητής υποψιάζεται την ύπαρξη κάποιων ομάδων με βάση κάποιο θεωρητικό μοντέλο που έχει στο μυαλό του (π.χ. κάποια είδη του ζωικού βασιλείου μοιάζουν μεταξύ τους, επομένως ο ερευνητής θέλει να διαπιστώσει κατά πόσο μπορεί να τα κατατάξει στην ίδια ομάδα).
- Πρόβλεψη καινούργιων τιμών. Έχοντας δημιουργήσει ομάδες από παρατηρήσεις σε πολλές εφαρμογές, ενδιαφερόμαστε να κατατάξουμε καινούργιες παρατηρήσεις. Για παράδειγμα, μια τράπεζα έχει κατατάξει τους πελάτες της σε καλούς, μέτριους και κακούς και θέλει να κατατάσσει καινούργιους πελάτες σε μία από αυτές τις κατηγορίες με βάση τα χαρακτηριστικά τους.

3.2 Η έννοια της απόστασης

Η απόσταση είναι μια θεμελιώδες έννοια στην πολυμεταβλητή ανάλυση και όχι μόνο για την ανάλυση δεδομένων. Σκοπός της απόστασης είναι να μετρήσει πόσο απέχουν δύο παρατηρήσεις, να ποσοτικοποιήσει δηλαδή αν μοιάζουν ή όχι οι παρατηρήσεις.

Για παράδειγμα, ας υποθέσουμε ότι ενδιαφερόμαστε για δύο μεταβλητές, το βάρος και το ύψος, δηλαδή για κάθε παρατήρηση έχουμε μετρήσεις για αυτές τις δύο μεταβλητές. Αν συμβολίσουμε τις δύο παρατηρήσεις ως $y = (y_{1,} y_2)$ και $x = (x_1, x_2)$, τότε μια πρώτη προσέγγιση για την επιλογή μιας απόστασης ανάμεσα στις δύο παρατηρήσεις, θα ήταν η ευκλείδεια απόσταση

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

την οποία μπορούμε να γενικεύσουμε για την περίπτωση που έχουμε παρατηρήσεις σε p μεταβλητές, δηλαδή $y=(y_{1,}y_2,....,y_p)$ και $x=(x_1,x_2,....,x_p)$. Τότε η αντίστοιχη απόσταση μπορεί να οριστεί ως

$$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}\ .$$

Αν και η ευκλείδεια απόσταση είναι θεμελιώδης στα μαθηματικά και την καθημερινή μας ζωή, στη στατιστική τα πράγματα είναι λίγο διαφορετικά. Η ευκλείδεια απόσταση δεν είναι στατιστικά επαρκής και αυτό συμβαίνει εξαιτίας του ότι οι μεταβλητές μπορεί να μην είναι σε συγκρίσιμη κλίμακα. Ένας τρόπος ώστε να φέρουμε κάθε μεταβλητή σε συγκρίσιμη κλίμακα είναι να διαιρέσουμε κάθε μεταβλητή με την τυπική απόκλιση της και επομένως , αφού όλες οι μεταβλητές πια θα αναφέρονται σε μονάδες τυπικής απόκλισης, έχουμε εξάλειψη το πρόβλημα. Έτσι, η απόσταση που θα χρησιμοποιήσουμε θα έχει την μορφή

$$d(x,y)=\sqrt{\sum_{i=1}^{p}\left(\frac{x_i-y_i}{s_i}\right)^2}\ .$$

Από στατιστική άποψη η απόσταση αυτή είναι πιο ενδιαφέρουσα και επιτρέπει πιο καλές συγκρίσεις ανάμεσα στις μεταβλητές. Το μόνο μειονέκτημα όμως που έχει είναι πως δεν λαμβάνει υπόψη της τις συνδιακύμανσεις ανάμεσα στις μεταβλητές. Αν δύο μεταβλητές είναι πολύ συσχετισμένες, τότε η απόσταση των παρατηρήσεων ουσιαστικά οφείλεται μόνο σε μία από αυτές, αφού η άλλη μεταβλητή απλώς ακολουθεί την πρώτη εξαιτίας της συσχέτισης. Μια απόσταση που λαμβάνει υπόψη τις συνδιακυμάνσεις είναι η απόσταση του Mahalanobis που υπολογίζεται ως εξής

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

Όπου S ο δειγματικός πίνακας διακυμάνσεων.

3.3 Μέτρα απόστασης

Τα μέτρα απόστασης είναι χωρισμένα στις παρακάτω κατηγορίες ανάλογα με το είδος των δεδομένων στα οποία μπορούν να εφαρμοσθούν. Εμείς θα αναφέρουμε όσα χρησιμοποιούνται στην παρούσα εργασία.

Συνεχή δεδομένα

Η περίπτωση των συνεχών δεδομένων είναι ίσως η απλούστερη αλλά και περισσότερο διαδεδομένη. Υπάρχουν πολλές αποστάσεις που έχουν χρησιμοποιηθεί για να μετρήσουν την απόσταση ανάμεσα σε συνεχή δεδομένα.

- Ευκλείδεια απόσταση

Η ευκλείδεια απόσταση αποτελεί την πιο απλή και την πιο γνωστή απόσταση ανάμεσα σε συνεχή δεδομένα. Εξαρτάται πολύ από την κλίμακα μέτρησης και επομένως αλλάζοντας την κλίμακα μπορούμε να πάρουμε ολότελα διαφορετικές αποστάσεις. Επίσης, μεταβλητές με μεγάλες απόλυτες τιμές έχουν πολύ μεγαλύτερο βάρος και σχεδόν καθορίζουν την απόσταση ανάμεσα σε παρατηρήσεις. Δεδομένου ότι παίρνουμε τετραγωνικές αποκλίσεις τα outliers έχουν μεγάλη επίδραση στον υπολογισμό της απόστασης.

- City-block (Manhattan) distance

Η απόσταση Manhattan μοιάζει πολύ με την ευκλείδεια απόσταση με την διαφορά ότι αντί για τετραγωνικές αποκλίσεις χρησιμοποιούμε απόλυτες αποκλίσεις. Επομένως η απόσταση ορίζεται ως

$$d(x,y)=\sum_{i=1}^{p}\left|x_i-y_i\right|.$$

Συνήθως, λόγω της ομοιότητας της με την ευκλείδεια απόσταση, δίνει περίπου ίδια αποτελέσματα εκτός από την περίπτωση που υπάρχουν outliers όπου επειδή τους δίνει μικρότερο βάρος (εξαιτίας της απόλυτης τιμής) μπορεί να οδηγήσει σε πιο ανθεκτικά αποτελέσματα.

- Απόσταση Minkowski (or Lq norm)

Η απόσταση Minkowski κατά κάποιον τρόπο γενικεύει την Ευκλείδεια απόσταση και την απόσταση Manhattan. Η απόσταση ορίζεται ως εξής:

$$d(x,y)=\left[\sum_{i=1}^{p}\left(\left|x_i-y_i\right|\right)^{q}\right]^{1/q}.$$

Η τιμή της παραμέτρου q μπορεί να χρησιμοποιηθεί για να δώσει ιδιαίτερο βάρος σε κάποιες αποκλίσεις. Προφανώς, αν q=1, προκύπτει η απόσταση Manhattan ,ενώ αν q=2, προκύπτει η ευκλείδεια.

Μια γενίκευση είναι η Power Distance , που ορίζεται ως εξής:

$$d(x,y)=\left[\sum_{i=1}^{p}\left(\left|x_i-y_i\right|\right)^{q}\right]^{1/r}$$

όπου τα r και q είναι παράμετροι που ορίζει ο ερευνητής.

- Chebychev distance

Η απόσταση Chebychev, σε αντίθεση με τις υπόλοιπες αποστάσεις που έχουμε αναφέρει, δεν χρησιμοποιεί όλες τις αποκλίσεις αλλά μόνο την μεγαλύτερη εξ αυτών. Συγκεκριμένα , η απόσταση ορίζεται ως εξής:

$$d(x, y) = \max\left\{\left(x_i - y_i\right), i = 1,...p\right\}.$$

Η απόσταση αυτή είναι χρήσιμη, όταν κανείς θέλει να εξετάσει δύο διαφορετικές παρατηρήσεις για το αν διαφέρουν τουλάχιστον ως προς μία μεταβλητή. Επειδή η απόσταση χρησιμοποιεί μόνο την μεγαλύτερη απόκλιση, εξαρτάται πολύ από τις διαφορές στην κλίμακα των μεταβλητών και επομένως αν οι κλίμακες είναι διαφορετικές, ουσιαστικά θα αντικατοπτρίζει τη διαφορά στη μεταβλητή με την μεγαλύτερη κλίμακα.

Δεδομένα σε ονομαστική κλίμακα

Όταν οι μεταβλητές αναφέρονται σε ονομαστική κλίμακα τότε είναι σχετικά δύσκολο να υπολογίσουμε απόσταση. Αυτό οφείλεται στο γεγονός ότι ουσιαστικά, αν οι δύο παρατηρήσεις έχουν ίδια τιμή, τότε αυτό είναι χρήσιμη πληροφορία, αλλά αν δεν έχουν την ίδια τιμή , τότε υπάρχουν πολλοί τρόποι που αυτό μπορεί να συμβεί και όλοι αυτοί έχουν το ίδιο βάρος.

Συνήθως η απόσταση που χρησιμοποιείται είναι ο συντελεστής ομοιότητας (simple matching approach) ο οποίος ορίζεται ως:

$$s(x, y) = \frac{p}{u} \quad \text{και} \quad d(x, y) = \frac{p - u}{p},$$

Όπου u είναι ο αριθμός των μεταβλητών που έχουν την ίδια τιμή και p ο συνολικός αριθμός μεταβλητών.

Μεταβλητές σε κλίμακα κατάταξης

Στην περίπτωση κατηγορικών μεταβλητών σε κλίμακα κατάταξης συνήθως αυτό που γίνεται είναι να θεωρήσουμε τις μεταβλητές ως συνεχείς και να χρησιμοποιήσουμε μία κατάλληλη απόσταση. Συνήθως τέτοιας μορφής δεδομένα χρησιμοποιούνται στην ψυχομετρία όπου ζητείται σε κάποιον να απαντήσει σε ερωτήσεις, αποδίδοντας βαθμό σε κάποια κλίμακα. Σε τέτοιες περιπτώσεις πρέπει να φροντίζει κανείς να χρησιμοποιεί την ίδια κλίμακα, ώστε να μην υπάρχει πρόβλημα. Εναλλακτικά αυτό που μπορεί να γίνει είναι να μετασχηματίσουμε την κλίμακα για να παίρνει τιμές στο διάστημα (0,1).

Μεταβλητές διαφόρων τύπων

Στην περίπτωση αυτή μιλάμε για δεδομένου μικτού τύπου (mixed model variables). Τα δεδομένα μας αποτελούνται από διαφορετικούς τύπους μεταβλητών,

όπως για παράδειγμα ηλικία(συνεχής), καταγωγή (ονομαστική). Υπάρχουν αρκετοί τρόποι να αντιμετωπίσουμε τέτοιας μορφής δεδομένα.

- Στην πρώτη περίπτωση προχωράμε στην ανάλυση, χρησιμοποιώντας ομοειδής μεταβλητές. Με αυτή την προσέγγιση κάνουμε ομαδοποιήσεις για κάθε τύπο μεταβλητών ξεχωριστά.
- Ένας άλλος τρόπος είναι να ορίσουμε ψευδομεταβλητές για όλους τους τύπους των δεδομένων, με αποτέλεσμα να καταλήξει σε ένα σύνολο από δυαδικές μεταβλητές. Στην περίπτωση συνεχών μεταβλητών αυτό προϋποθέτει μία διακριτοποίηση τους, δηλαδή αρχικά τις μετατρέπουμε σε μικρότερα διακριτά διαστήματα και στη συνέχεια ορίζουμε ψευδομεταβλητές για τη διακριτοποιημένη μεταβλητή. Προφανώς, αυτή η προσέγγιση μπορεί να οδηγήσει σε σοβαρό χάσιμο πληροφορίας κατά την διάρκεια των μετασχηματισμών.

3.4 Η μέθοδος K-means

Ο αλγόριθμος K-means ανήκει σε μια μεγάλη κατηγορία αλγορίθμων ομαδοποίησης που είναι γνωστοί ως αλγόριθμοι διαμέρισης (partitioning algorithms). Ουσιαστικά οι αλγόριθμοι είναι έτσι φτιαγμένοι, ώστε να διαμερίζουν το πολυεπίπεδο που δημιουργούν τα δεδομένα σε περιοχές και να αντιστοιχεί μια περιοχή σε κάθε ομάδα.

3.4.1 Ο αλγόριθμος

Η μέθοδος θεωρεί πως ο αριθμός των ομάδων που θα προκύψουν είναι γνωστός εκ των προτέρων. Αυτό αποτελεί έναν περιορισμό της μεθόδου, καθώς είτε πρέπει να τρέξουμε τον αλγόριθμο με διαφορετικές επιλογές ως προς το πλήθος των ομάδων είτε πρέπει με κάποιον άλλο τρόπο να έχουμε καταλήξει στον αριθμό των ομάδων.

Η μέθοδος δουλεύει επαναληπτικά. Χρησιμοποιεί την έννοια του κέντρου της ομάδας και στην συνέχεια κατατάσσει τις παρατηρήσεις ανάλογα με την απόσταση τους από τα κέντρα όλων των ομάδων. Το κέντρο της ομάδας δεν είναι τίποτα άλλο από την μέση τιμή για κάθε μεταβλητή όλων των παρατηρήσεων της ομάδας, δηλαδή αντιστοιχεί στο διάνυσμα των μέσων. Στην συνέχεια για κάθε παρατήρηση υπολογίζουμε την ευκλείδεια απόσταση της από τα κέντρα των ομάδων που έχουμε και κατατάσσουμε κάθε παρατήρηση στην ομάδα που είναι πιο κοντά, για την ακρίβεια στην ομάδα με κέντρο πιο κοντά στην παρατήρηση. Αφού κατατάξουμε όλες τις παρατηρήσεις, τότε υπολογίζουμε εκ νέου τα κέντρα, απλώς ως τα διανύσματα των μέσων για τις παρατηρήσεις που ανήκουν σε κάθε ομάδα. Η διαδικασία επαναλαμβάνεται μέχρις ότου δεν υπάρχουν διαφορές ανάμεσα σε δύο διαδοχικές επαναλήψεις.

Αλγοριθμικά έχουμε:

- **Βήμα 1ο**. Βρες τα αρχικά κέντρα.
- **Βήμα 2ο**. Κατάταξε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση.
- **Βήμα 3ο**. Από τις παρατηρήσεις που είναι μέσα στην ομάδα υπολόγισε τα νέα κέντρα.
- **Βήμα 4ο**. Αν τα νέα κέντρα δε διαφέρουν από τα παλιά σταμάτα αλλιώς πήγαινε στο **βήμα 2ο**.

Θα πρέπει να παρατηρήσουμε πως η δυναμική του αλγορίθμου είναι ότι με τις πρώτες λίγες επαναλήψεις πλησιάζει πολύ κοντά στην τελική λύση και στις υπόλοιπες επαναλήψεις, οι οποίες διαφορές οφείλονται σε μετακίνηση κάποιων λίγων παρατηρήσεων που βρίσκονται ουσιαστικά ανάμεσα σε δύο ομάδες. Επομένως δεν είναι απαραίτητος ένας μεγάλος αριθμός επαναλήψεων, καθώς η βασική δομή θα σχηματιστεί πολύ γρήγορα.

3.4.2 K-Means στο SPSS

Ποιες μεταβλητές πρέπει να χρησιμοποιηθούν

Στην πραγματικότητα δεν υπάρχει κάποιος τρόπος για να οδηγήσει στην επιλογή μεταβλητών πριν την ανάλυση. Αν, λοιπόν, δεν υπάρχει κάποια εμπειρία ή κάποιος θεωρητικός λόγος για να επιλέξουμε κάποιες συγκεκριμένες μεταβλητές για την ανάλυση, καταφεύγουμε στην χρήση όλων των διαθέσιμων μεταβλητών. Εναλλακτικά θα μπορούσε κανείς να διαλέξει μόνο εκείνες τις μεταβλητές που πιστεύουμε για κάποιους λόγους ότι έχουν τη δυνατότητα να δημιουργήσουν ομοιογενείς ομάδες.

Αφού κάνουμε την ανάλυση, μπορούμε εκ των υστέρων αν δούμε αν κάποιες μεταβλητές τελικά ήταν αδιάφορες με την έννοια ότι η τιμή τους είναι ίδια για όλες τις ομάδες που δημιουργήσαμε και επομένως δεν έχουν καμία διακριτική ικανότητα. Αν μάλιστα πιστεύουμε ότι δεν υπάρχει λόγος η μεταβλητή αυτή να παραμείνει στην ανάλυση μπορούμε να την αφαιρέσουμε. Εμείς θα χρησιμοποιήσουμε όλες τις συνεχείς μας μεταβλητές (ηλικία, εισόδημα, κ.λπ.) και τις μεταβλητές που ανήκουν σε κλίμακα κατάταξης (ordinal), καθώς γνωρίζουμε ότι μπορούμε να τις χρησιμοποιήσουμε ως συνεχείς με την προϋπόθεση ότι βρίσκονται στην ίδια κλίμακα.

Τυποποίηση μεταβλητών

Αν οι μεταβλητές μετρώνται σε διαφορετικά μεγέθη, οι μεταβλητές με μεγάλες τιμές θα συμβάλλουν περισσότερο στην απόσταση από τις μεταβλητές με μικρές τιμές. Έτσι το βασικό πριν προχωρήσουμε σε οποιαδήποτε ανάλυση είναι να βγάλουμε περιγραφικά μέτρα για κάθε μία από τις μεταβλητές που θα χρησιμοποιήσουμε. Αν δούμε μεγάλες αποκλίσεις στα περιγραφικά χαρακτηριστικά των δεδομένων μας όπως για παράδειγμα η μέση τιμή της μεταβλητής ηλικία είναι 22,42 ενώ η μέση τιμή του εισοδήματος είναι 402,50 τότε μία καλή τεχνική για να

απαλείψουμε αυτό το πρόβλημα είναι να τυποποιήσουμε τα δεδομένα μας. Αυτό μπορεί να γίνει μέσω της εντολής του SPSS Transform →Compute Variable, αφαιρούμε την μέση τιμή και διαιρούμε με την τυπική απόκλιση την κάθε μεταβλητή.

(Δημήτρης Καρλής. *Πολυμεταβλητή Στατιστική Ανάλυση. Εκδόσεις Σταμούλη, Αθήνα 2005)*

Εφαρμογή της μεθόδου

Για να ξεκινήσουμε επιλέγουμε Analyse→Classify→K-means. Επιλέγουμε όλες τις μεταβλητές που μας ενδιαφέρουν και θέλουμε να τις κατατάξουμε σε 3 διαφορετικές ομάδες, στην συνέχεια καθορίζουμε τον αριθμό των επαναλήψεων που θα σταματήσει ο αλγόριθμος, στην περίπτωση μας 10, και θα αποθηκεύσουμε μία καινούργια μεταβλητή όπου θα δίνεται η τιμή της ομάδας στην οποία κατατάξαμε κάθε παρατήρηση. Τέλος επιλέγουμε να μας εμφανίσει έναν πίνακα με τα αρχικά κέντρα των ομάδων. Τα αποτελέσματα είναι τα εξής.

Πίνακας 3.1

Initial Cluster Centers

	Cluster		
	1	2	3
Z_age	,991	-,263	-1,517
Z_savings	-,267	9,586	-,289
Z_Income3yo	3,462	-,017	-1,757
Z_Income	3,568	,443	-1,269
Z_nightlife	3,348	-,308	-,612
Z_reduction	-2,890	,382	1,472
Z_quality	1,922	,771	-1,530

Ο πίνακας 3.1 περιέχει τα αρχικά κέντρα των ομάδων, αυτά δηλαδή από όπου ξεκίνησε ο αλγόριθμος.

Πίνακας 3.2

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	4,073	2,827	3,147
2	,624	,000	,172
3	,426	,000	,179
4	,214	,000	,091
5	,118	,000	,057
6	,127	,000	,066
7	,078	,000	,046
8	,043	,000	,025
9	,000	,000	,000

Ο πίνακας 3.2 περιέχει πληροφορίες για το πως μετακινείται ο αλγόριθμος σε κάθε επανάληψη. Η τιμή που εμφανίζεται είναι η απόσταση ανάμεσα στο κέντρο της ομάδας στην τρέχουσα επανάληψη με το κέντρο της ομάδας κατά την προηγούμενη. Βλέπουμε ότι η διαδικασία σταματά στην ένατη επανάληψη, επειδή η απόσταση έχει μηδενιστεί.

Πίνακας 3.3

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
Z_age	7,246	2	,937	197	7,732	,001
Z_savings	75,969	2	,239	197	318,002	,000
Z_Income3yo	38,435	2	,620	197	61,997	,000
Z_Income	50,440	2	,498	197	101,271	,000
Z_nightlife	12,779	2	,880	197	14,514	,000
Z_reduction	22,127	2	,786	197	28,133	,000
Z_quality	14,810	2	,859	197	17,233	,000

Τα F test θα πρέπει να χρησιμοποιηθούν μόνο για περιγραφικούς σκοπούς επειδή οι ομάδες έχουν επιλεχθεί ώστε να μεγιστοποιούνται οι διαφορές ανάμεσα στις διαφορετικές ομάδες. Τα παρατηρούμενα επίπεδα σημαντικότητας δεν είναι κατάλληλα διορθωμένα και έτσι δεν μπορούν αν χρησιμοποιηθούν ως τεστ για τον έλεγχο της υπόθεσης ότι οι μέσες τιμές των ομάδων είναι ίσες. Ως συμπέρασμα από αυτόν τον πίνακα μπορούμε να πάρουμε βλέποντας τα p-value, ότι οι μεταβλητές είναι σημαντικές για την ανάλυση μας και έχουν διακριτική ικανότητα. Δηλαδή συμβάλλουν όλες στην διαχώριση και κατασκευή των ομάδων. Αν συνέβαινε το αντίθετο, τότε θα αφαιρούσαμε όποια μεταβλητή δεν είχε διακριτική ικανότητα.

Πίνακας 3.4

Final Cluster Centers

	Cluster		
	1	2	3
Z_age	,355	-,054	-,208
Z_savings	,011	7,026	-,175
Z_Income3yo	,826	-,365	-,467
Z_Income	,940	,170	-,545
Z_nightlife	,475	-,206	-,271
Z_reduction	-,603	-,709	,364
Z_quality	,483	,771	-,297

Ο πίνακας 3.4 παρουσιάζει τα κέντρα των ομάδων που βρέθηκαν, αφού σταμάτησε ο αλγόριθμος. Χρησιμοποιώντας τα τελικά κέντρα των ομάδων, ας αναλύσουμε τώρα τα προφίλ των ομάδων που δημιούργησε η μέθοδος.

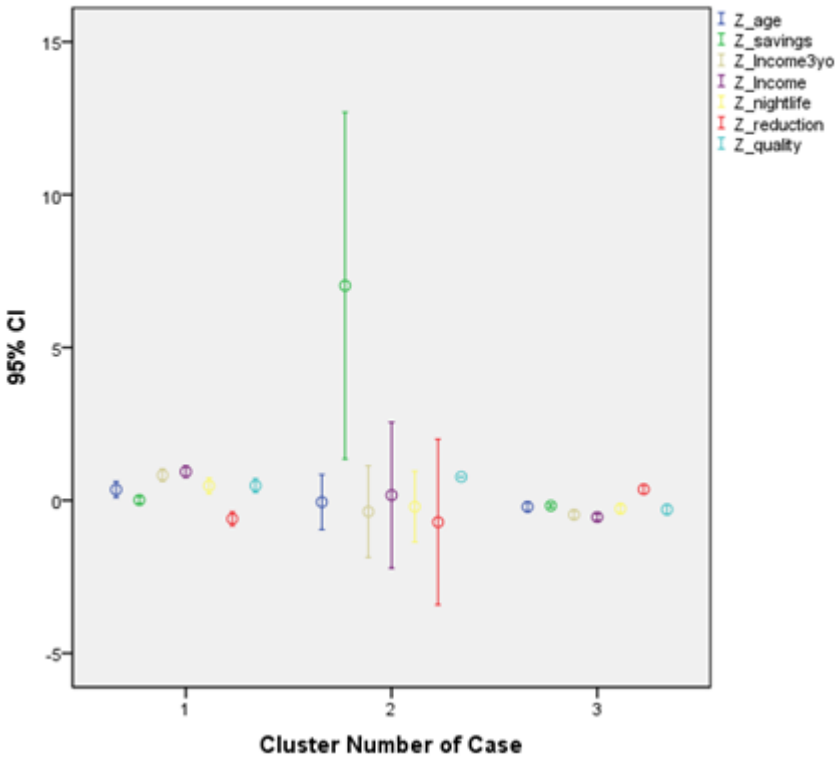
Ομάδα 1: Αποτελείται από φοιτητές που είναι σε μεγαλύτερη ηλικία. Το εισόδημα τους, τωρινό και πριν 3 ακαδημαϊκά έτη είναι αρκετά υψηλό σε σύγκριση με τις υπόλοιπες ομάδες. Τα χρήματα που έχουν αποταμιεύσει κυμαίνονται γύρω από το μέσο όρο. Η μηνιαία βραδινή διασκέδαση τους είναι υψηλή και χαρακτηρίζουν την ποιότητα ζωής τους στο Καρλόβασι από οικονομικής απόψεως αρκετά ικανοποιητική. Τέλος, όπως είναι λογικό η μείωση σε αγαθά πολυτελείας είναι χαμηλή σε αυτή την ομάδα.

Ομάδα 2: Σε αυτή την ομάδα βρίσκονται φοιτητές ηλικίας λίγο κάτω από το μέσο όρο όπου το εισόδημα τους πριν από 3 χρόνια ήταν κάτω από το μέσο όρο αλλά αυξήθηκε το τωρινό τους εισόδημα. Μήπως αυτό εξηγεί και τα υψηλά ποσά που έχουν στην τράπεζα; Η βραδινή διασκέδαση τους μπορεί να χαρακτηρισθεί ως κάτω από το μέσο όρο. Επίσης δεν έχουν μειώσει τις αγορές του σε αγαθά πολυτελείας και έχουν χαρακτηρίσει την ποιότητα ζωής τους πολύ υψηλή κατά μέσο όρο.

Ομάδα 3: Η ομάδα 3 αποτελείται από τους νεότερους φοιτητές και μπορούμε να δούμε ότι σε όλες τις μεταβλητές έχουν κάτω από το μέσο όρο. Η μείωση στις αγορές τους χαρακτηρίζεται και αυτή σχετικά υψηλή.

Όπως περιγράψαμε τις ομάδες παραπάνω με λόγια μπορούμε να το δούμε και διαγραμματικά. Θα μας δώσει και μία εικόνα γραφικά πως χωρίζονται ομάδες. Έτσι πάμε να δημιουργήσουμε ένα Error Bar.

Διάγραμμα 3.1



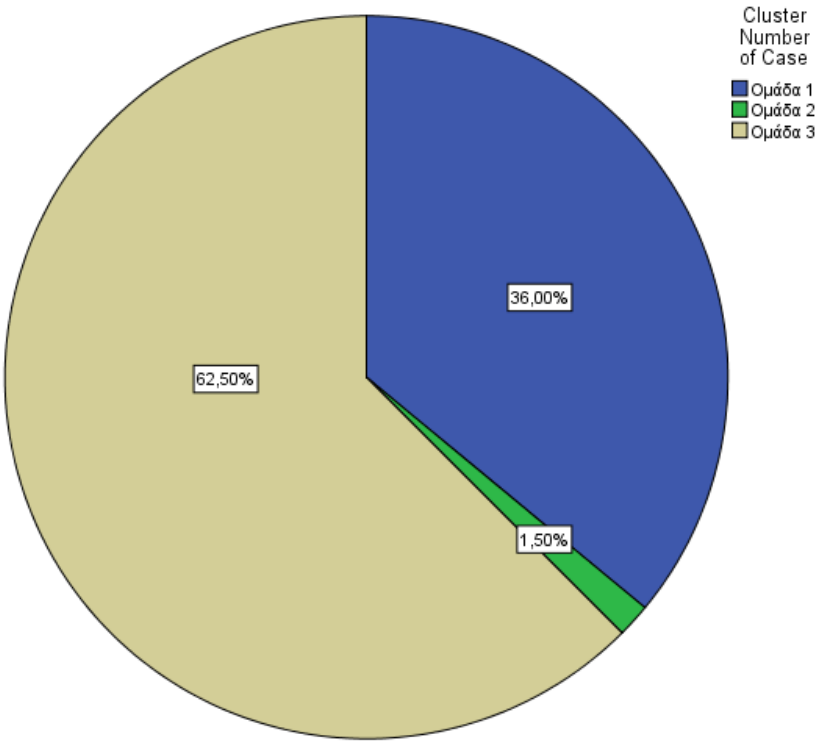
Το SPSS μας παρέχει με έναν ακόμα πίνακα. Αυτός μας δίνει τις συχνότητες στην κάθε ομάδα.

Πίνακας 3.5

Number of Cases in each Cluster

Cluster	1	72,000
	2	3,000
	3	125,000
Valid		200,000
Missing		,000

Έτσι λοιπόν βλέπουμε ότι στην πρώτη ομάδα βρίσκονται 72 φοιτητές, στην δεύτερη 3 φοιτητές και στην τρίτη 125. Ας δημιουργήσουμε ένα Pie chart ώστε να δούμε οπτικά πως η μέθοδος κατένειμε τους φοιτητές.



3.4.3 Συμπεράσματα της μεθόδου

Από τις ομάδες που δημιούργησε ο αλγόριθμος K-Means έχουμε την δυνατότητα να αντλήσουμε ορισμένες πληροφορίες. Όπως βλέπουμε η οικονομική κρίση έχει επηρεάσει σε μεγαλύτερο βαθμό την ομάδα 3 που αποτελείται από τους νεότερους φοιτητές, και αντιπροσωπεύουν το 36% του δείγματος μας, σε σχέση με τους υπόλοιπους. Τα εισοδήματα τους ήταν χαμηλά και απήλθε περεταίρω μείωση. Όπως είναι φυσικό, αυτό έχει αντίκτυπο στην ποιότητα ζωής τους, την αποταμίευση, στην υψηλή μείωση των αγορών τους και της διασκέδασης τους.

Η ομάδα 1 δεν δείχνει να φαίνεται να έχει επηρεαστεί σε μεγάλο βαθμό απο την κρίση. Τα εισοδήματα τους πριν από τρία ακαδημαϊκά έτη ήταν υψηλά και αυξήθηκαν ελάχιστα. Οι αποταμιεύσεις κυμαίνονται σε μία μεσαία κατάσταση, ενώ η ποιότητα ζωής τους χαρακτηρίζεται υψηλή, όπως και η διασκέδαση τους. Η μείωση των αγορών τους δεν έχει μειωθεί αισθητά.

Τέλος , η δεύτερη ομάδα αποτελείται από ηλικίες λίγο κάτω από τον μέσο όρο, και θα την χαρακτηρίσουμε ως «ιδιαίτερη». Αυτή η ομάδα αντιπροσωπεύει το 1,5% του δείγματος των οποίων τα εισοδήματα έχουν αυξηθεί σε πολύ μεγάλο βαθμό

τα τελευταία 3 χρόνια. Σε συνδυασμό με την βραδινή τους διασκέδαση που είναι αρκετά κάτω από το μέσο όρο, από ότι φαίνεται έχουν καταφέρει να αποταμιεύσουν σε μεγάλο βαθμό χρήματα. Έχουν χαρακτηρίσει πολύ καλή την ποιότητα ζωής τους στο Καρλόβασι και η μείωση των αγορών τους χαρακτηρίζεται από πολύ χαμηλές τιμές.

3.5 Two-Step ομαδοποίηση

3.5.1 Εισαγωγή

Η Two-Step ανάλυση συστάδων σε σύγκριση με την Ιεραρχική Ανάλυση και την K-Means παρέχει περισσότερα πλεονεκτήματα. Με την μέθοδο αυτή, μπορούμε να επιλέξουμε εμείς των αριθμό συστάδων που επιθυμούμε ή μπορεί να γίνει αυτόματα από το στατιστικό πακέτο, βάση στατιστικών κριτηρίων. Απαιτεί μόνο ένα «σάρωμα» των δεδομένων ,το οποίο είναι πολύ χρήσιμο για πολύ μεγάλα σετ δεδομένων, και μπορεί να παρέχει λύσεις βασισμένες σε ένα μείγμα συνεχών και κατηγορικών μεταβλητών.

Ο αλγόριθμος ομαδοποίησης βασίζεται σε ένα μέτρο απόστασης που δίνει τα καλύτερα αποτελέσματα αν όλες οι μεταβλητές είναι ανεξάρτητες, με τις συνεχείς να ακολουθούν την κανονική κατανομή και τις κατηγορικές μεταβλητές την πολυωνυμική. Αυτό συμβαίνει σπάνια στην πράξη ,αλλά ο αλγόριθμος θεωρείται ότι συμπεριφέρεται αρκετά καλά, ακόμη και όταν δεν πληρούνται οι προϋποθέσεις. Επειδή η ανάλυση συστάδων δεν περιλαμβάνει έλεγχο υποθέσεων και υπολογισμό επιπέδων σημαντικότητας, παρά μόνο για περιγραφικούς σκοπούς, είναι απολύτως αποδεκτά τα αποτελέσματα της ομαδοποίησης να μην πληρούν τις υποθέσεις για την βέλτιστη λύση.

3.5.2 Τα βήματα

Βήμα 1^{ο} : ***Preclustering: Δημιουργία πρωταρχικών ομάδων (Preclusters)***

Το πρώτο στάδιο της διαδικασίας Two-Step είναι ο σχηματισμός πρωταρχικών ομάδων (Theodoridis & Koutroumbas, 1999). Ο στόχος του Preclustering είναι να η μείωση του μεγέθους του πίνακα που περιέχει τις αποστάσεις μεταξύ όλων των πιθανών ζευγών περιπτώσεων. Καθώς διαβάζεται μία παρατήρηση ο αλγόριθμος αποφασίζει, βασιζόμενος σε ένα μέτρο απόστασης, αν η συγκεκριμένη παρατήρηση θα πρέπει να ομαδοποιηθεί σε κάποια ήδη υπάρχουσα πρωταρχική ομάδα ή θα δημιουργήσει μία καινούργια. Όταν η διαδικασία του Preclustering ολοκληρωθεί, όλες οι παρατηρήσεις την ίδια πρωταρχική ομάδα αντιμετωπίζονται ως μία οντότητα. Το μέγεθος του πίνακα αποστάσεων δεν είναι πια

εξαρτημένο από των αριθμό των παρατηρήσεων αλλά από τον αριθμό των πρωταρχικών ομάδων.

Βήμα 2^{ο} : ***Ιεραρχική ομαδοποίηση των πρωταρχικών ομάδων***

Στο δεύτερο βήμα, το SPSS χρησιμοποιεί τον κλασσικό αλγόριθμο ιεραρχικής ομαδοποίησης. Η δημιουργία συστάδων ιεραρχικά μας δίνει τη δυνατότητα να εξερευνήσουμε ένα εύρος λύσεων με διαφορετικούς αριθμούς συστάδων.

Επιπλέον πληροφορίες της μεθόδου

Τυποποίηση: Ο αλγόριθμος αυτόματα θα τυποποιήσει όλες τις μεταβλητές εκτός και αν δεν το επιθυμούμε.

Μέτρα απόστασης: Αν τα δεδομένα μας είναι μείγμα συνεχών και κατηγορικών μεταβλητών, μπορούμε να χρησιμοποιήσουμε το κριτήριο log-likelihood. Η απόσταση μεταξύ δύο συστάδων εξαρτάται από την μείωση του log-likelihood όταν ενώνονται σε μία ομάδα. Αν τα δεδομένα αποτελούνται μόνο από συνεχείς μεταβλητές, μπορούμε να χρησιμοποιήσουμε και την Ευκλείδεια απόσταση μεταξύ των δύο κέντρων των συστάδων. Ανάλογα το μέτρο απόστασης που θα χρησιμοποιήσουμε οι παρατηρήσεις εκχωρούνται στις ομάδες που οδηγούν στο μεγαλύτερο log-likelihood ή στην ομάδα η οποία έχει την μικρότερη Ευκλείδεια απόσταση.

Αριθμός συστάδων: Μπορούμε να ορίσουμε τον αριθμό των συστάδων που θέλουμε να δημιουργηθούν, ή να αφήσουμε τον αλγόριθμο να επιλέξει τον βέλτιστο αριθμό βάση του Μπεϋζιανού κριτηρίου του Schwartz ή του κριτηρίου πληροφορίας του Akaike.

Ακραίες τιμές: Έχουμε την επιλογή αν δημιουργήσουμε ξεχωριστή ομάδα για τις παρατηρήσεις οι οποίες δεν προσαρμόζονται καλά σε καμία άλλη συστάδα.

3.5.3 Two-Step ομαδοποίηση στο SPSS

Για να ξεκινήσουμε την διαδικασία στο Spss επιλέγουμε Analyse→ Classify→ Two-Step Cluster.

Μεταβλητές

Ως κατηγορικές μεταβλητές που θα χρησιμοποιήσουμε για την ομαδοποίηση θα είναι η προέλευση (πόλη, νησί, χωριό), η στέγαση (εστία, ενοίκιο), η κλίμακα ποιότητας ζωής και η κλίμακα μείωσης των αγορών. Ως συνεχείς θα χρησιμοποιηθεί το τωρινό μηνιαίο εισόδημα των φοιτητών. Η Two-Step μέθοδος μας δίνει την δυνατότητα να επιλέξουμε μεταβλητή ως Evaluation Variable. Αυτό σημαίνει ότι η ίδια δεν θα επηρεάσει καθόλου την διαδικασία της ομαδοποίησης αλλά θα μπορούμε να διακρίνουμε τις τιμές της σε κάθε ομάδα. Οι μεταβλητές που θα χρησιμοποιήσουμε είναι συνεχείς και κατηγορικές άρα θα επιλέξουμε ως μέτρο απόστασης την Log-likelihood.

Αριθμός συστάδων

Για την βέλτιστη λύση του αριθμού των συστάδων θα αφήσουμε το Spss να επιλέξει βάση του Μπεϋζιανού Κριτηρίου του Schwartz.

Αποτελέσματα

Πίνακας 3.5.1

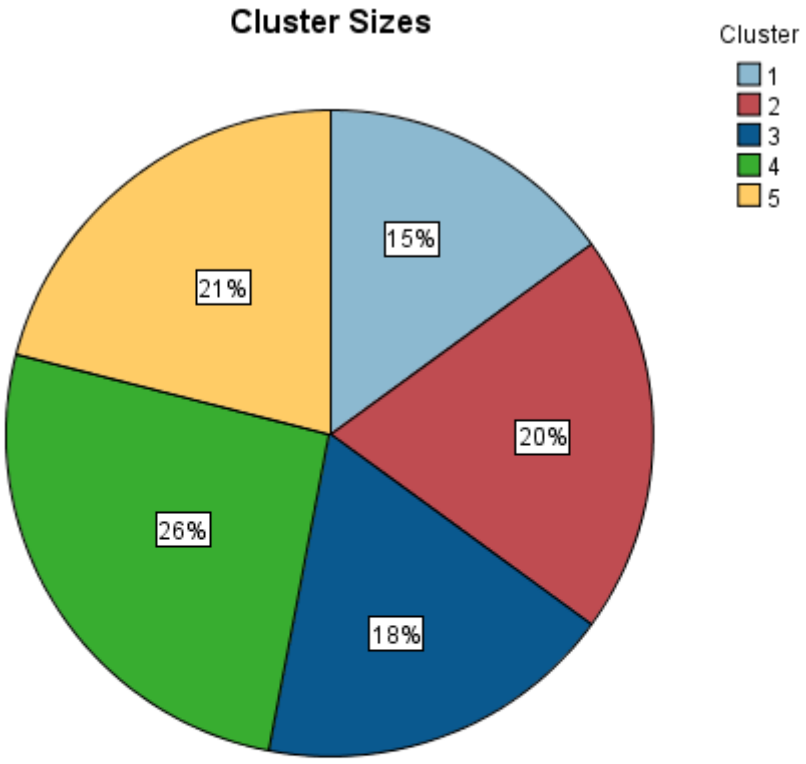
Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	5

Cluster Quality

Ο πρώτος πίνακας που μας εμφανίζει το Spss μας δίνει την τελική λύση του αλγορίθμου και την ποιότητα της ομαδοποίησης. Στη περίπτωση μας δημιουργήθηκαν 5 ομάδες, βάση των 5 μεταβλητών που δώσαμε και η ποιότητα χαρακτηρίζεται αρκετά ικανοποιητική.

Διάγραμμα 3.5.1



Size of Smallest Cluster	30 (15%)
Size of Largest Cluster	52 (26%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.73

Στην συνέχεια βλέπουμε με ένα διάγραμμα πίτας που μας δείχνει τα ποσοστά που αντιπροσωπεύει η κάθε συστάδα καθώς και το μέγεθος της μεγαλύτερης και μικρότερης. Επίσης αναγράφει και την αναλογία μεταξύ τους που είναι 1.73, που είναι πολύ ικανοποιητικό.

Πίνακας 3.5.2

Clusters

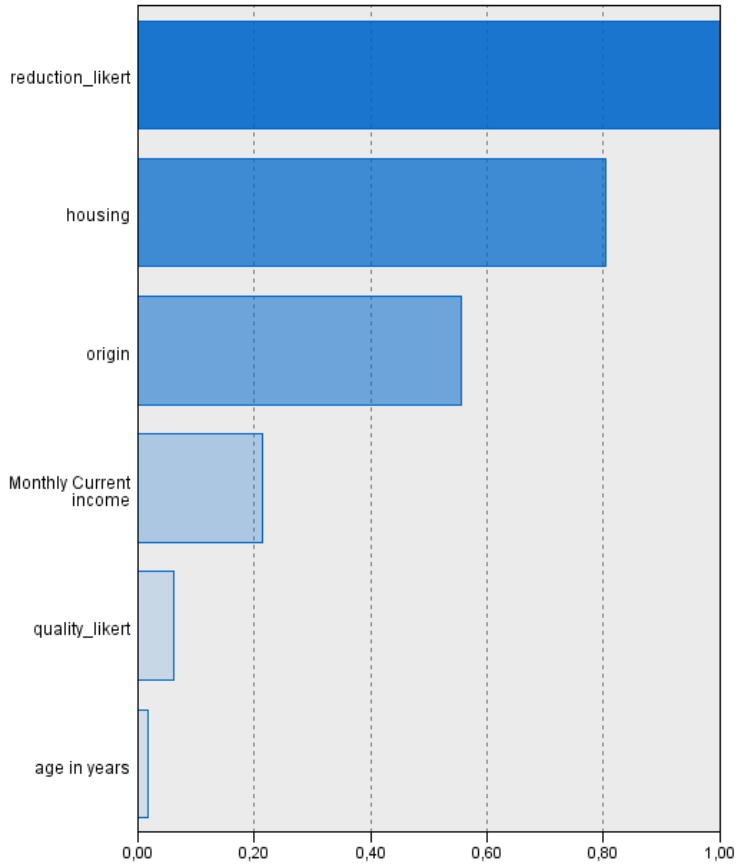
Input (Predictor) Importance
1,0 0,8 0,6 0,4 0,2 0,0

Cluster	4	5	2	3	1
Label					
Description					
Size	26.0% (52)	21.0% (42)	20.0% (40)	18.0% (36)	15.0% (30)
Inputs	reduction_likert 4 (100.0%)	reduction_likert 3 (100.0%)	reduction_likert 4 (60.0%)	reduction_likert 5 (58.3%)	reduction_likert 4 (43.3%)
	housing Ενοίκιο (100.0%)	housing Ενοίκιο (100.0%)	housing Ενοίκιο (100.0%)	housing Ενοίκιο (94.4%)	housing Εστία (100.0%)
	origin Πόλη (100.0%)	origin Πόλη (100.0%)	origin Χωριό (60.0%)	origin Πόλη (77.8%)	origin Πόλη (60.0%)
	Monthly Current income 297.12	Monthly Current income 398.10	Monthly Current income 327.75	Monthly Current income 364.17	Monthly Current income 190.33
	quality_likert 4 (48.1%)	quality_likert 4 (52.4%)	quality_likert 4 (50.0%)	quality_likert 4 (36.1%)	quality_likert 3 (56.7%)
Evaluation Fields	age in years 22.75	age in years 22.64	age in years 22.30	age in years 22.22	age in years 21.90

Ο πίνακας Clusters (πίνακας 3.5.2) παρουσιάζει τα αρχικά αποτελέσματα της μεθόδου συγκεντρωτικά. Βλέπουμε τα ποσοστά που αντιστοιχούν σε κάθε κατηγορική μεταβλητή για κάθε ομάδα και τις μέσες τιμές αντίστοιχα για τις συνεχείς. Η διαφορά στο χρώμα του μπλε αντιπροσωπεύει κλίμακα [0,1] που δηλώνει το κατά πόσο η κάθε μεταβλητή συνέβαλλε στον προσδιορισμό των ομάδων. Μία ιδεατή κατάσταση θα ήταν η κάθε μεταβλητή να είχε την ίδια βαρύτητα. Τέλος, στο κάτω μέρος βλέπουμε και την μεταβλητή ηλικία που δεν συνέβαλλε στην ομαδοποίηση, αλλά μας παρουσιάζει τις μέσες τιμές της για κάθε συστάδα. Δεν θα προχωρήσουμε στην ανάλυση των συστάδων από αυτόν τον πίνακα, αλλά παρακάτω για κάθε ομάδα ξεχωριστά.

Πίνακας 3.5.3

Predictor Importance



Το διάγραμμα Predictor Importance (Πίνακας 3.5.3) μας παρουσιάζει αναλυτικά πόσο και ποιες μεταβλητές συνέβαλλαν στην ομαδοποίηση. Όπως βλέπουμε η

μείωση των αγορών επηρέασε περισσότερο από τις υπόλοιπες μεταβλητές την ομαδοποίηση, λίγο λιγότερο η στέγαση και η προέλευση.

Ανάλυση συστάδων

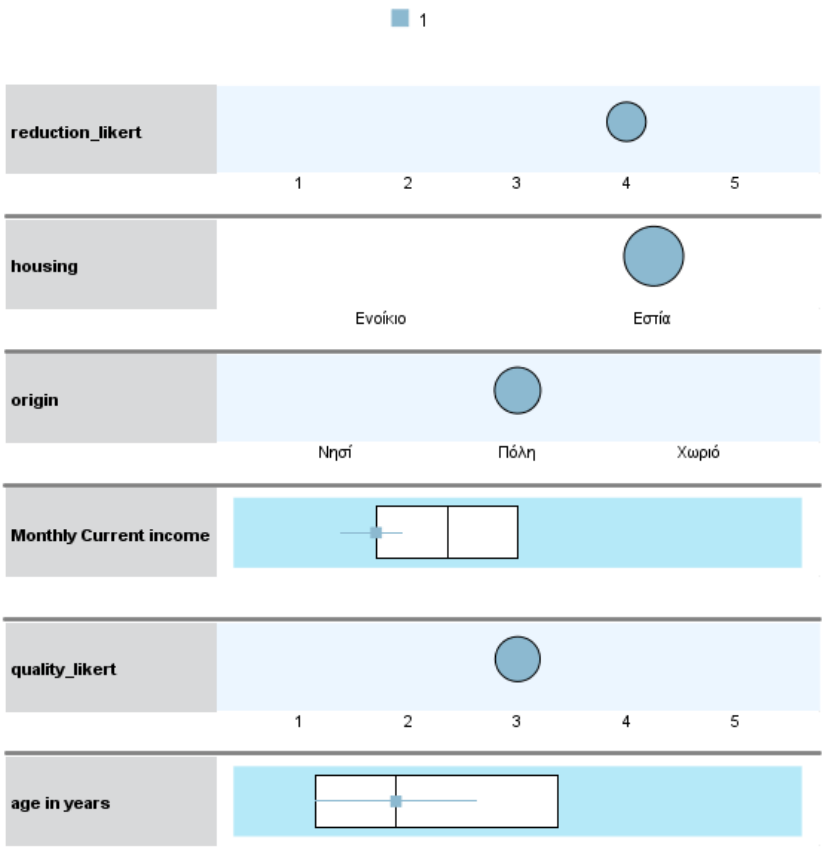
Επιλέγοντας μία συστάδα, το output μας εμφανίζει περισσότερους πίνακες που μπορούμε να εκμεταλλευτούμε για να περιγράψουμε κάθε ομάδα ξεχωριστά.

1η ομάδα

Η πρώτη ομάδα περιέχει 30 παρατηρήσεις που αποτελούν το 15% του συνολικού δείγματος. Η ομάδα περιέχει μόνο φοιτητές που μένουν στην εστία. Το 60% των φοιτητών προέρχονται από πόλη. Όπως βλέπουμε στον πίνακα 3.5.4 η επικρατούσα τιμή της μείωσης των αγορών τους είναι 4 που δηλώνει ότι είναι υψηλή. Το 56.7% των φοιτητών έχει χαρακτηρίσει την ποιότητα ζωής του μέτρια. Από το Box-plot της εικόνας 4.3.5 βλέπουμε ότι το εισόδημα των φοιτητών που μένουν στην εστία είναι πολύ χαμηλό. Συγκεκριμένα, η διάμεσος του δείγματος είναι 301.17 ευρώ, ενώ στην συστάδα 1 είναι μόλις 200.82. Η ομάδα χαρακτηρίζεται από ηλικίες πολύ κοντά στην διάμεσο. Η μέση τιμή των ηλικιών είναι 21.90.

Πίνακας 3.5.4

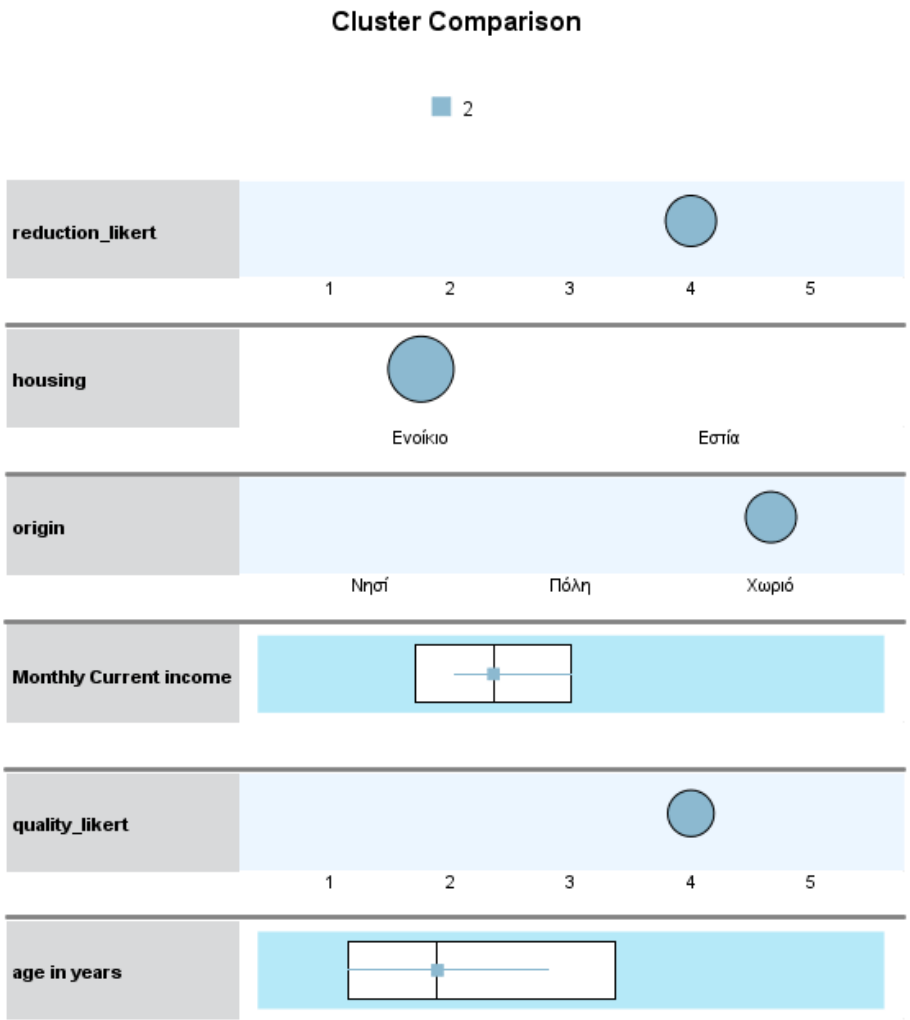
Cluster Comparison



2η ομάδα

Στην δεύτερη ομάδα ανήκουν 40 φοιτητές που αποτελούν το 20% του συνολικού δείγματος οι οποίοι ενοικιάζουν όλοι. Το 60% αυτών προέρχονται από χωριό και το υπόλοιπο 40% από νησί. Όπως βλέπουμε και αυτοί έχουν χαρακτηρίσει υψηλή την μείωση των αγορών τους. Υψηλή επίσης είναι και η ποιότητα ζωής τους στην κλίμακα μέτρησης. Το μηνιαίο εισόδημά τους έχει μέση τιμή 327.75 και η διάμεσος της ομάδας βρίσκεται σχεδόν πάνω στην ολική διάμεσο. Η ηλικία που παρατηρείται σε αυτή την συστάδα έχει μέση τιμή 22.30.

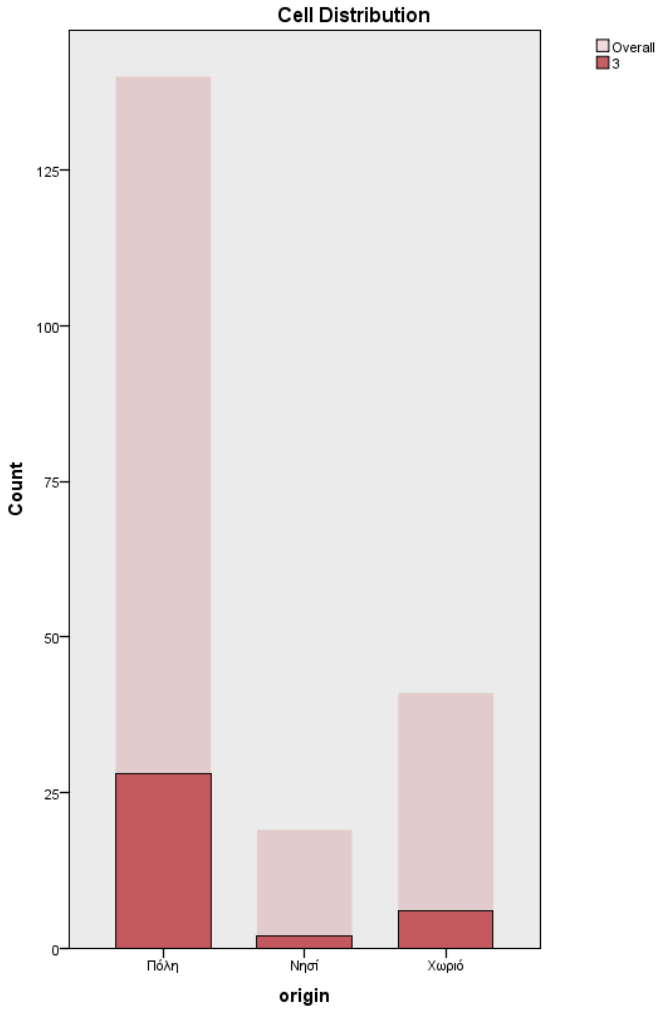
Πίνακας 3.5.5



3η ομάδα

Η ομάδα 3 περιέχει 36 παρατηρήσεις (18% του συνολικού) και το 94.4% ενοικιάζει σπίτι ενώ το υπόλοιπο 5.6% μένει στην εστία. Το 77.8% είναι από πόλη ενώ το υπόλοιπο 22.2% χωρίζεται σε χωριό και νησί όπως δείχνει το διάγραμμα 3.5.2.

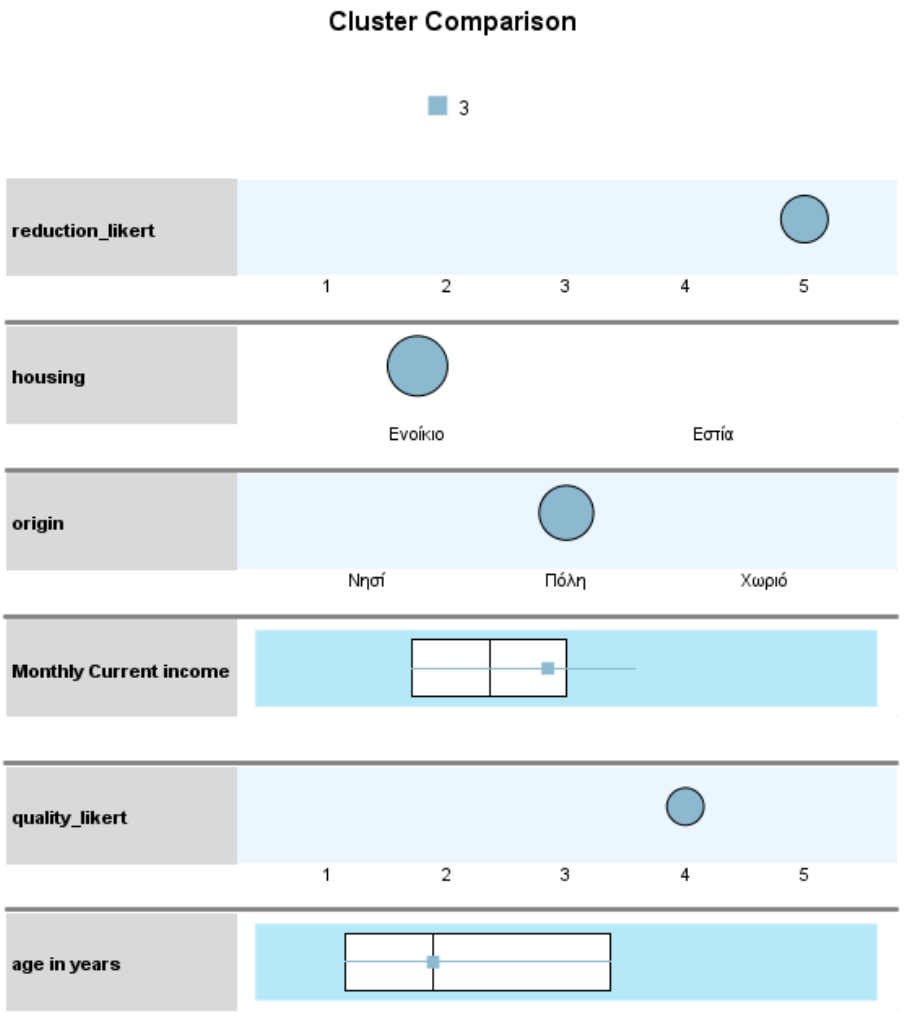
Διάγραμμα 3.5.2



Όπως βλέπουμε στον πίνακα 3.5.6 η ποιότητα ζωής των φοιτητών στην ομάδα 3 είναι υψηλή όπως πολύ υψηλό είναι και το μηνιαίο εισόδημα τους με μέση τιμή 364.17, με διάμεσο πολύ πιο πάνω από την ολική διάμεσο. Όπως φαίνεται έχουν

κόψει αρκετά τις ακριβές αγορές τους, αφού η επικρατούσα τιμή της μείωσης είναι 5, όπου δηλώνει πολύ μεγάλη μείωση. Έχουν χαρακτηρίσει υψηλή την ποιότητα ζωής τους και η μέση τιμή των ηλικιών τους είναι 22.22.

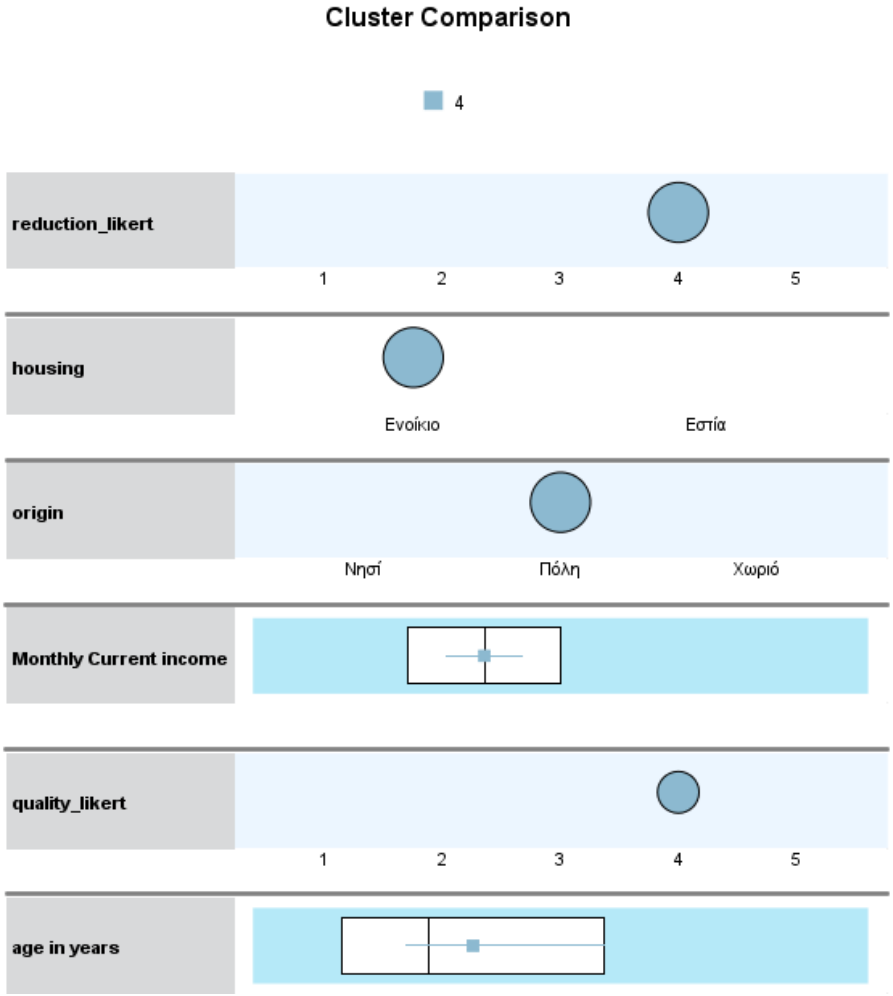
Πίνακας 3.5.6



4η ομάδα

Η τέταρτη ομάδα περιέχει 52 φοιτητές όπου είναι και η μεγαλύτερη ομάδα. Όλοι τους ενοικιάζουν σπίτι και προέρχονται από κάποια πόλη όπως φαίνεται στον πίνακα 3.5.7.Η μέση ηλικία αυτής της ομάδας είναι η μεγαλύτερη από τις υπόλοιπες με τιμή 22.75. Έχουν χαρακτηρίσει την ποιότητα ζωής τους υψηλή αν και έχουν υψηλές μειώσεις στις αγορές τους. Η συστάδα αυτή οικονομικά χαρακτηρίζεται ως μία μεσαία κατάσταση με μέση τιμή εισοδήματος 297.12.

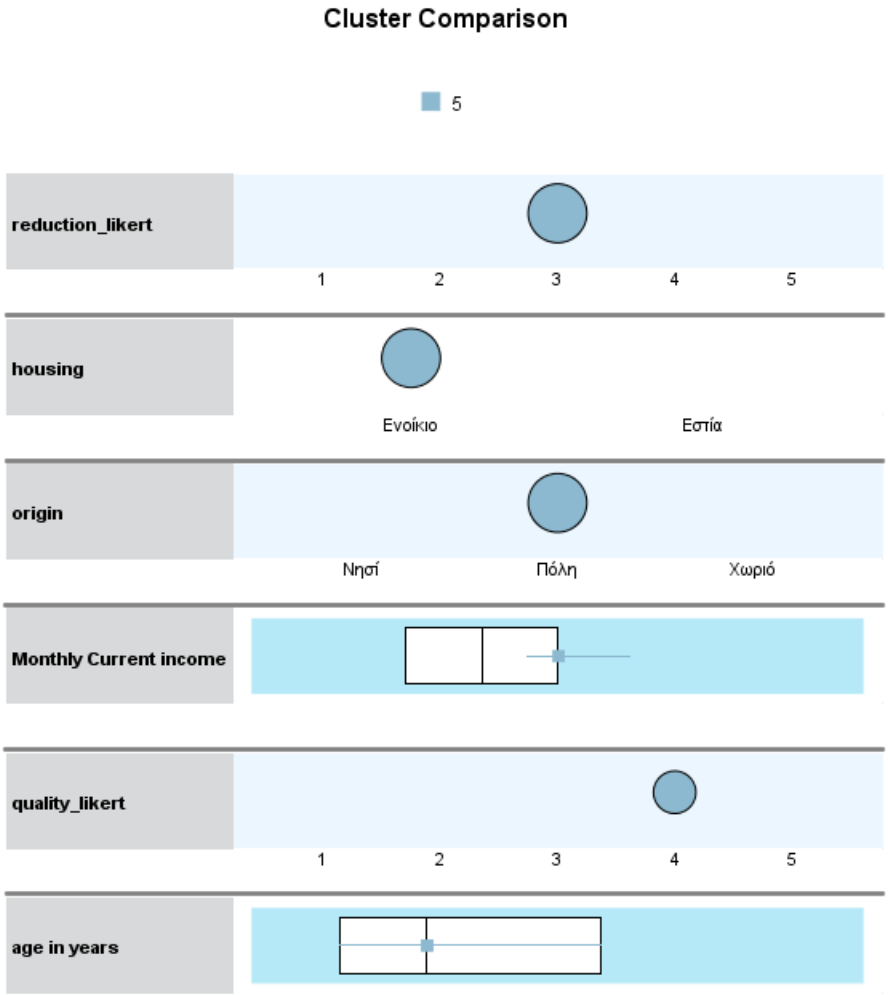
Πίνακας 3.5.7



5η ομάδα

Η ομάδα 5 περιέχει 42 φοιτητές όπου όλοι προέρχονται από πόλη και ενοικιάζουν σπίτι. Αυτή ομάδα θα μπορούσε να χαρακτηριστεί ως η πιο οικονομικά άνετη. Όπως βλέπουμε στον πίνακα 3.5.9 η διάμεσος του εισοδήματος της ομάδας είναι πάρα πολύ ψηλά σε σχέση με το δείγμα. Η μέση τιμή του είναι 398.10. Φυσικό είναι να έχουν δηλώσει μία μέτρια μείωση των αγορών τους, και χαρακτηρίζουν την ποιότητα ζωής τους υψηλή. Με μέση τιμή 22.64 οι φοιτητές αυτοί είναι οι δεύτεροι μεγαλύτεροι εκ των ομάδων.

Πίνακας 3.5.9



3.5.4 Γενικά συμπεράσματα της μεθόδου

Τα συμπεράσματα που μπορούμε να αντλήσουμε από αυτή την μέθοδο είναι τα εξής:

Ο αλγόριθμος Two-Step χώρισε τις ομάδες δίνοντας την μεγαλύτερη βαρύτητα στην μείωση των αγορών των φοιτητών. Όπως είδαμε, η ακρίβεια έχει επηρεάζει όλους τους φοιτητές, μερικούς περισσότερο, μερικούς λιγότερο. Ανεξάρτητα την οικονομική κατάσταση του καθενός όλοι αναγκάστηκαν να μειώσουν έστω και σε λίγο βαθμό τις αγορές τους. Η Two Step μέθοδος έρχεται σε συμφωνία με την K-Means μέθοδο, στο πόρισμα ότι η οικονομική κρίση έχει επηρεάσει περισσότερο τους φοιτητές μικρότερης ηλικίας. Τέλος, σημαντικό αποτέλεσμα της μεθόδου είναι ότι το γραφείο φοιτητικής μέριμνας παρέχει τις εστίες στα άτομα που το έχουν μεγαλύτερη ανάγκη.

Κεφάλαιο 4

Πολυμεταβλητή ανάλυση διακύμανσης

4.1 Εισαγωγή

Η μέθοδος της πολυμεταβλητής ανάλυσης διακύμανσης (Manova) αποτελεί γενίκευση της μονομεταβλητής ανάλυσης διακύμανσης, όταν εξετάζουμε συγχρόνως περισσότερες από μία μεταβλητές. Αποτελεί επομένως μία μέθοδο ελέγχου του αν οι μέσοι δύο ή περισσότερων ομάδων διαφέρουν και, γενικεύοντας σε περιπτώσεις πολλών παραγόντων, αν οι παράγοντες αυτοί επιδρούν στη μέση τιμή. Πολλά πράγματα που ισχύουν στην μονομεταβλητή περίπτωση μεταφέρονται με ανάλογο τρόπο και στην πολυμεταβλητή περίπτωση όπως για παράδειγμα η διάσπαση της συνολικής διακύμανσης στη μεταξύ των γκρουπ και εντός των γκρουπ διακύμανση (between και within).

4.2 MANOVA ως προς ένα Παράγοντα

Θα αρχίσουμε εξετάζοντας την απλούστερη περίπτωση ανάλυσης με έναν παράγοντα (one way).

Έστω ότι έχουμε ένα δείγμα X_{ij} , $j=1,.....,k$, και $i=1,....,n_j$, όπου γενικά j συμβολίζει το γκρουπ (μεταβλητή) και i την παρατήρηση μέσα στο γκρουπ. Η παρατήρηση X_{ij} , είναι πια ένα διάνυσμα με p στοιχεία και όχι μία απλή τιμή. Το μέγεθος του δείγματος δεν είναι αναγκαστικά το ίδιο για όλα τα γκρουπ. Υποθέτουμε ότι ο πληθυσμός για το j γκρουπ ακολουθεί την κατανομή $N_p(\mu_j,\Sigma)$, όπου p είναι η διάσταση και μ_j και Σ είναι το διάνυσμα των μέσων και ο πίνακας διακύμανσης συνδιακύμανσης αντίστοιχα. Υποθέτουμε επίσης πως ο πίνακας διακύμανσης συνδιακύμανσης είναι ο ίδιος για όλα τα γκρουπ. Επομένως, οι δύο υποθέσεις που χρειάζεται να κάνουμε στη MANOVA είναι:

- Κανονικότητα των δεδομένων
- Ίδιος πίνακας διακύμανσης συνδιακύμανσης για όλα τα γκρουπ

Η μηδενική υπόθεση που θέλουμε να ελέγξουμε είναι :

$H_0 : \mu_1 = \mu_2 = ... = \mu_k$ έναντι της εναλλακτικής

H_1 : τουλάχιστον δύο μέσοι διαφέρουν

Προχωρώντας με το συνηθισμένο τρόπο κατασκευής ελέγχων με την μέθοδο του λόγου των πιθανοφανειών, μπορούμε να βρούμε την πιθανοφάνεια κάτω από τις δύο υποθέσεις.

Όπως και στη μονομεταβλητή περίπτωση $\bar{x}_{\bullet j} = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$, $\bar{x}_{\bullet\bullet} = \frac{\sum_j \sum_i x_{ij}}{n}$, δηλαδή ο μέσος του j γκρουπ και ο γενικός μέσος αντίστοιχα.

Κάτω από την μηδενική υπόθεση η εκτίμηση για το μέσο κάθε ομάδας είναι η ίδια και συγκεκριμένα ο μέσος όλων των παρατηρήσεων. Επομένως, υποθέτοντας ότι $n_j = n$, ο λογάριθμος της πιθανοφάνειας των δεδομένων είναι

$-\frac{nkp}{2}\log(2\pi) - \frac{nkp}{2}\log|\Sigma| - \frac{1}{2}\sum_i\sum_j \left(x_{ij}-\mu_j\right)' \Sigma^{-1}\left(x_{ij}-\mu_j\right)$ το οποίο αγνοώντας τον πρώτο όρο και αντικαθιστώντας γράφεται ως

$-\frac{nkp}{2}\log|\Sigma| - \frac{1}{2} tr\Sigma^{-1}\left[\sum_j\sum_i \left(x_{ij}-\bar{x}_{\bullet j}\right)\left(x_{ij}-\bar{x}_{\bullet j}\right)' + n\sum_j \left(\bar{x}_{\bullet j}-\bar{x}_{\bullet j}\right)\left(\bar{x}_{\bullet j}-\bar{x}_{\bullet j}\right)'\right]$, όπου ο δεύτερος όρος στην αγκύλη είναι μηδέν.

Δουλεύοντας ομοίως κάτω από την εναλλακτική υπόθεση βρίσκουμε πως ο λογάριθμος της πιθανοφάνειας είναι:

$-\frac{nkp}{2}\log|\Sigma| - \frac{1}{2} tr\Sigma^{-1}\left[\sum_j\sum_i \left(x_{ij}-\bar{x}_{\bullet j}\right)\left(x_{ij}-\bar{x}_{\bullet j}\right)' + n\sum_j \left(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet}\right)\left(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet}\right)'\right]$.

Μπορεί να γραφτεί στη μορφή

$-\frac{nkp}{2}\log|\Sigma| - \frac{1}{2} tr\Sigma^{-1}\left[B+W\right]$,όπου

$W = \sum_j\sum_i \left(x_{ij}-\bar{x}_{\bullet j}\right)\left(x_{ij}-\bar{x}_{\bullet j}\right)'$, $B = n\sum_j \left(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet}\right)\left(\bar{x}_{\bullet j}-\bar{x}_{\bullet\bullet}\right)'$.

Όπως είδαμε σπάσαμε σε δύο όρους τις τετραγωνικές αποκλίσεις, όπως κάναμε και στη μονομεταβλητή περίπτωση. Ο πρώτος όρος που τον αποτελεί ο πίνακας W μας δίνει τις αποκλίσεις μέσα στο κάθε γκρουπ και αντιστοιχεί στο within sum of squares. Ο δεύτερος όρος μας δίνει τις αποκλίσεις ανάμεσα στους μέσους των γκρουπ από τον γενικό μέσο και αντιστοιχεί στο between sum of squares. Να σημειώσουμε ότι τα στοιχεία της διαγωνίου είναι τα γνωστά από την μονομεταβλητή

ανάλυση διακύμανσης αθροίσματα τετραγώνων που χρησιμοποιούνται για τους μονομεταβλητούς ελέγχους.

Χρησιμοποιώντας τον παραπάνω συμβολισμό βρίσκουμε πως η λογαριθμική πιθανοφάνεια για τη μηδενική υπόθεση γίνεται

$$-\frac{nkp}{2}\log|\Sigma|-\frac{1}{2}tr\Sigma^{-1}[W].$$

Στην πρώτη περίπτωση, επειδή υποθέτουμε πως όλοι οι μέσοι είναι ίδιοι, ο εκτιμητής μέγιστης πιθανοφάνειας του κοινού μέσου είναι ο μέσος όλου του δείγματος. Στη δεύτερη περίπτωση όμως, που τέτοια υπόθεση δεν κάνουμε, χρησιμοποιούμε τον εκτιμητή για κάθε γκρουπ και έτσι οδηγούμαστε στο αποτέλεσμα που είδαμε.

Έτσι οι μέγιστες τιμές των λογαριθμοποιημένων πιθανοφανείων είναι

$$-\frac{nkp}{2}\log|B+W|-\frac{np}{2} \quad \text{και} \quad -\frac{nkp}{2}\log|W|-\frac{np}{2}$$

για την εναλλακτική και την μηδενική υπόθεση αντίστοιχα, οπότε καταλήγουμε πως ο λόγος των πιθανοφανειών LR είναι

$$LR=kn\log\frac{|W|}{|B+W|}.$$

Είναι επομένως λογικό να βασίσουμε τη συμπερασματολογία μας στο λόγο των δύο οριζουσών και συγκεκριμένα στην ποσότητα

$$\Lambda=\frac{|W|}{|B+W|}.$$

4.3 'Ελεγχοι υποθέσεων

Αν συμβολίσουμε με $T=B+W$ και γνωρίζουμε ότι η ορίζουσα ενός πίνακα διακύμανσης συνδιακύμανσης είναι ένα μέτρο της γενικής μεταβλητότητας, μπορούμε να δούμε πως η ποσότητα αυτή είναι παρόμοια με τον λόγο F που χρησιμοποιούμε στην μονομεταβλητή περίπτωση.

Βασισμένοι στην υπόθεση ότι τα δεδομένα προέρχονται από μία πολυμεταβλητή κανονική κατανομή και χρησιμοποιώντας τις ιδιότητες της κατανομής Wishart και κυρίως του τρόπου που αυτή προκύπτει από την πολυμεταβλητή κανονική κατανομή, έχει δειχτεί ότι η στατιστική συνάρτηση Λ ακολουθεί την κατανομή Λ-Wilks. Πιο συγκεκριμένα, μπορεί να δειχτεί

$$W\sim W_p(n-k,\Sigma)$$

$$B\sim W_p(k-1,\Sigma).$$

Αν οι πίνακες B, W είναι ανεξάρτητοι, τότε η συνάρτηση Λ ακολουθεί την κατανομή $\Lambda\ (p, n-k, k-1)$, και ο έλεγχος θα στηριχθεί στην κατανομή Λ Wilks και τις σχέσεις που την συνδέουν με γνωστές κατανομές.

Έτσι λοιπόν απορρίπτουμε τη μηδενική υπόθεση, όταν η τιμή του Λ είναι κοντά στο 0. Αυτό είναι λογικό, αν αναλογισθεί κανείς ότι μικρή τιμή Λ συνεπάγεται για τον πίνακα W των within διαφορών ότι αυτές είναι μικρές σε σχέση με τις between διαφορές του πίνακα B και άρα οι διαφορές ανάμεσα στις ομάδες είναι μεγάλες.

Έχουμε

$$\Lambda = \frac{1}{\left|W^{-1}\right|\left|B+W\right|} = \frac{1}{\left|I+W^{-1}B\right|} = \prod_{i=1}^{p} \frac{1}{\lambda_i}$$

όπου λ_i είναι οι ιδιοτιμές του πίνακα $I+W^{-1}B$.

Όμως

$$\left|I+W^{-1}B-\lambda I\right| = 0 \Rightarrow \left|W^{-1}B-(\lambda-1)I\right| = 0$$

$$\Rightarrow l = (\lambda-1) \Rightarrow \lambda = l+1$$

όπου l_i είναι οι ιδιοτιμές του πίνακα $W^{-1}B$. Συνεπώς μπορούμε να γράψουμε πως $\Lambda = \prod_{i=1}^{p} \frac{1}{l_i+1}$ και να βασιστούμε στις ιδιοτιμές αυτές για την κατασκευή εναλλακτικών ελέγχων. Άλλοι έλεγχοι είναι :

1. ‘Ελεγχος Roy. Θεωρούμε τη στατιστική συνάρτηση $F = \frac{n-k-p+1}{p} l_1$,όπου l_1 είναι η μεγαλύτερη ιδιοτιμή του πίνακα $W^{-1}B$. Η στατιστική συνάρτηση F ακολουθεί την κατανομή F με βαθμούς ελευθερίας $\left|k-p-1\right|+1$ και $n-k-p+1$, αντίστοιχα.Απορρίπτουμε για μεγάλες τιμές της F .
2. Έλεγχος Pilai. Η στατιστική συνάρτηση είναι η $tr(B+W)^{-1}B = \sum_{j=1}^{p} \frac{1}{\lambda_j}$.
3. ‘Ελεγχος ίχνους Lawley-Hotteling. Η στατιστική συνάρτηση είναι η $tr(W^{-1}B) = \sum_{j=1}^{p} \frac{1}{l_j}$.

Σημειώνουμε πως, αν οι υποθέσεις του μοντέλου μας είναι γενικά σωστές (κανονικότητα και σταθερή διακύμανση), όλοι οι έλεγχοι θα πρέπει να δίνουν το ίδιο αποτέλεσμα , και P-value τα οποία να είναι κοντά. Γι’αυτό σε περίπτωση που παρατηρούμε μεγάλες αποκλίσεις στα αποτελέσματα, χρησιμοποιώντας τους ελέγχους, αυτό είναι μια ισχυρή ένδειξη πως θα πρέπει πως θα πρέπει να

κοιτάξουμε τα δεδομένα για το κατά πόσο η MANOVA μπορεί να εφαρμοσθεί. Ο έλεγχος Pilai είναι πιο ανθεκτικός ,όταν οι υποθέσεις δεν τηρούνται. Σε περίπτωση που έχουμε 2 γκρουπ όλοι οι έλεγχοι είναι ισοδύναμοι με τον έλεγχο του Hotelling.

Ας τονίσουμε ότι, όπως και στην απλή ANOVA, χρειαζόμαστε την υπόθεση ότι οι πίνακες διακυμάνσεων είναι ίσοι για να είναι αξιόπιστα τα αποτελέσματα. Χωρίς αυτή την υπόθεση η μεθοδολογία δεν ισχύει πια. Για να ελέγξουμε την υπόθεση ότι όλοι οι πίνακες είναι ίσοι, μπορούμε να χρησιμοποιήσουμε τον έλεγχο Box-M που μας παρέχει το Spss στα αποτελέσματα. (Δημήτρης Καρλής. *Πολυμεταβλητή Στατιστική Ανάλυση. Εκδόσεις Σταμούλη, Αθήνα 2005.)*

4.4 Εφαρμογή της MANOVA στο Spss

Για να ξεκινήσουμε την διαδικασία ακολουθούμε την διαδρομή Analyze→General Linear model→Multivariate.

Μεταβλητές που θα χρησιμοποιήσουμε

Στην εφαρμογή της MANOVA θα εξετάσουμε την υπόθεση ότι οι μέσες τιμές των εξαρτημένων μας μεταβλητών θα διαφέρουν σε κάθε επίπεδο της ανεξάρτητης ποσοτικής μας. Ως ανεξάρτητη έχουμε την μεταβλητή ‘’Κλίμακα ποιότητας ζωής’’ με 5 επίπεδα χαρακτηρισμού της, και ως εξαρτημένες θα χρησιμοποιήσουμε τις μεταβλητές ‘’Τωρινό μηνιαίο εισόδημα’’ και ‘’Βραδινή διασκέδαση’’. Υποθετικά, περιμένουμε οι φοιτητές που έχουν χαρακτηρίσει την ποιότητα ζωής τους υψηλή, να διακρίνουμε μεγαλύτερα μέσα ποσά χρημάτων και περισσότερες βραδινές εξόδους.

Εφόσον παρατηρήσουμε στατιστικά σημαντική διαφορά στους μέσους μπορούμε να εντοπίσουμε που βρίσκονται αυτές οι διαφορές κάνοντας εκ των υστέρων ανάλυση (Post-Hocs tests). Εφόσον πιστεύουμε ότι θα υπάρχουν στατιστικά σημαντικές διαφορές θα επιλέξουμε τον έλεγχο LSD μιας και είναι ο πιο ισχυρός σε αυτή την περίπτωση.

Κανονικότητα των δεδομένων

Πριν ξεκινήσουμε με την MANOVA θα εξετάσουμε την κανονικότητα των εξαρτημένων μεταβλητών μιας και είναι η μία από τις δύο προϋποθέσεις που πρέπει να ακολουθούν τα δεδομένα μας. Έτσι με ένα Kolmogorov - Smirnoff test έχουμε:

Πινακας 4.1

One-Sample Kolmogorov-Smirnov Test

		Monthly Current income	Monthly nightlife
N		200	200
Normal Parameters[a,b]	Mean	320,50	5,00
	Std. Deviation	134,384	3,287
Most Extreme Differences	Absolute	,151	,195
	Positive	,151	,195
	Negative	-,073	-,116
Kolmogorov-Smirnov Z		2,130	2,751
Asymp. Sig. (2-tailed)		,000	,000

a. Test distribution is Normal.
b. Calculated from data.

Όπως διακρίνουμε το παρατηρούμενο επίπεδο σημαντικότητας είναι μικρότερο από 0.05 έτσι απορρίπτουμε την μηδενική υπόθεση, δηλαδή οι μεταβλητές δεν ακολουθούν την κανονική κατανομή. Η μέθοδος της MANOVA είναι αρκετά ανεκτική στην παραβίαση της υπόθεσης της κανονικότητας έτσι θα συνεχίσουμε την ανάλυση μας.

Εφαρμογή

Ο πρώτος πίνακας των αποτελεσμάτων της μεθόδου 4.2 είναι μερικά περιγραφικά στοιχεία των εξαρτημένων μεταβλητών μας για κάθε γκρουπ της ανεξάρτητης μεταβλητής. Όπως βλέπουμε οι μέσες τιμές και στα χρηματικά ποσά και στις βραδυνές εξόδους είναι διαφορετικές για κάθε γκρουπ, και όπως περιμέναμε όσο αυξάνεται η ποιότητα ζωής αυξάνουν και οι τιμές των εξαρτημένων μεταβλητών, εκτός από το γκρουπ που χαρακτήρισε την ποιότητα ζωής ‘’Πολύ χαμηλή’’, ενώ η μέση τιμή των χρημάτων και της διασκέδασης είναι αρκετά ψηλές.

Πίνακας 4.2

Descriptive Statistics

	quality_likert	Mean	Std. Deviation	N
Monthly Current income	πολύ χαμηλή	396,25	101,972	8
	χαμηλή	271,90	157,690	21
	μέτρια	271,84	104,380	76
	υψηλή	350,11	127,183	87
	πολύ υψηλή	512,50	155,265	8
	Total	320,50	134,384	200
Monthly nightlife	πολύ χαμηλή	8,50	4,140	8
	χαμηλή	4,71	3,862	21
	μέτρια	3,93	2,537	76
	υψηλή	5,29	3,129	87
	πολύ υψηλή	9,25	3,151	8
	Total	5,00	3,287	200

Στην συνέχεια το Spss μας παρέχει τον πίνακα (4.3) Box's Test of Equality of Covariance Matrices. Εδώ ελέγχεται η προϋπόθεση του ίδιου πίνακα διακύμανσης συνδιακύμανσης για όλα τα γκρουπ. Αποδεχόμαστε την μηδενική υπόθεση αφού το P-value >0.05 και δεν παραβιάζουμε την δεύτερη προϋπόθεση. Μπορούμε να συνεχίσουμε.

Πίνακας 4.3

Box's Test of Equality of Covariance Matrices[a]

Box's M	19,704
F	1,532
df1	12
df2	4450,043
Sig.	,105

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + quality_likert

Στην συνέχεια θα δούμε τον πίνακα Multivariate Tests (4.4) που δίνει τα αποτελέσματα της μεθόδου. Έχοντας παραβιάσει την υπόθεση της κανονικότητας θα κοιτάξουμε τον έλεγχο Pilai που είναι πιο ανθεκτικός σε αυτή την περίπτωση. Το παρατηρούμενο επίπεδο σημαντικότητας είναι κάτω από 0.05 επομένως απορρίπτουμε την μηδενική υπόθεση ότι οι μέσες τιμές των εξαρτημένων μεταβλητών είναι ίσες σε όλα τα γκρουπ.

Πίνακας 4.4

Multivariate Tests[d]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Intercept	Pillai's Trace	,808	407,576[a]	2,000	194,000	,000	,808	815,152	1,000
	Wilks' Lambda	,192	407,576[a]	2,000	194,000	,000	,808	815,152	1,000
	Hotelling's Trace	4,202	407,576[a]	2,000	194,000	,000	,808	815,152	1,000
	Roy's Largest Root	4,202	407,576[a]	2,000	194,000	,000	,808	815,152	1,000
quality_likert	Pillai's Trace	,259	7,255	8,000	390,000	,000	,130	58,042	1,000
	Wilks' Lambda	,746	7,661[a]	8,000	388,000	,000	,136	61,288	1,000
	Hotelling's Trace	,334	8,066	8,000	386,000	,000	,143	64,525	1,000
	Roy's Largest Root	,313	15,281[c]	4,000	195,000	,000	,239	61,125	1,000

a. Exact statistic
b. Computed using alpha = ,05
c. The statistic is an upper bound on F that yields a lower bound on the significance level.
d. Design: Intercept + quality_likert

Η τιμή 13% του Partial Eta Square δηλώνει πόση διακύμανση εξηγείται από την ανεξάρτητη μεταβλητή. Είναι το αντίστοιχο του R^2 στην απλή ANOVA. Η ένδειξη Observed Power μας δίνει την πιθανότητα να αποδεχτούμε σωστά την μηδενική υπόθεση, δεν μας παρέχει καμία ιδιαίτερη πληροφορία μιας και ο έλεγχος βγήκε στατιστικά σημαντικός.

Μόλις είδαμε ότι ως συνδυασμός οι δύο εξαρτημένες μεταβλητές έχουν στατιστική σημαντική διαφορά στις μέση τιμή για τα γκρουπ. Στην συνέχεια, το Spss μας παρέχει με τους μεμονωμένους ελέγχους για να εξετάσει την κάθε μεταβλητή ξεχωριστά. Πριν όμως αυτό το βήμα , ελέγχει την προϋπόθεση της απλής ανάλυσης

διασποράς, δηλαδή την ισότητα των διακύμανσεων. Τον έλεγχο αυτόν τον βλέπουμε στον πίνακα Levene's test of Equality of Error Variances (4.5) .

Πίνακας 4.5

Levene's Test of Equality of Error Variances[a]

	F	df1	df2	Sig.
Monthly Current income	2,434	4	195	,049
Monthly nightlife	1,743	4	195	,142

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + quality_likert

Στην περίπτωση της μεταβλητής Διασκέδαση η προϋπόθεση πληρείται (p-value >0.05) ενώ στο χρηματικό ποσό είναι στο όριο, παρόλα αυτά, δεν δεχόμαστε την μηδενική.

Ο Πίνακας 4.6 μας παρουσιάζει τις απλές ANOVA ξεχωριστά για κάθε μεταβλητή. Και εδώ υπάρχουν στατιστικά σημαντικές διαφορές ,βλέποντας το p-value για ακόμα μία φορά κάτω από το 0.05.

Πίνακας 4.6

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[b]
Corrected Model	Monthly Current income	646647,735[a]	4	161661,934	10,697	,000	,180	42,787	1,000
	Monthly nightlife	337,727[c]	4	84,432	9,085	,000	,157	36,339	,999
Intercept	Monthly Current income	10082841,36	1	10082841,36	667,148	,000	,774	667,148	1,000
	Monthly nightlife	3115,368	1	3115,368	335,213	,000	,632	335,213	1,000
quality_likert	Monthly Current income	646647,735	4	161661,934	10,697	,000	,180	42,787	1,000
	Monthly nightlife	337,727	4	84,432	9,085	,000	,157	36,339	,999
Error	Monthly Current income	2947102,265	195	15113,345					
	Monthly nightlife	1812,273	195	9,294					
Total	Monthly Current income	24137800,00	200						
	Monthly nightlife	7150,000	200						
Corrected Total	Monthly Current income	3593750,000	199						
	Monthly nightlife	2150,000	199						

a. R Squared = ,180 (Adjusted R Squared = ,163)
b. Computed using alpha = ,05
c. R Squared = ,157 (Adjusted R Squared = ,140)

Σχόλιο : Όπως είδαμε και στην MANOVA και στις δύο ανεξάρτητες ANOVA το αποτέλεσμα βγήκε το ίδιο, οι διαφορές είναι στατιστικά σημαντικές. Αυτό δεν συμβαίνει πάντοτε. Το Spss στην πολυμεταβλητή ανάλυση διακύμανσης δημιουργεί μία καινούργια μεταβλητή, (canonical variate) ως γραμμικό συνδυασμό 2 ή περισσότερων ανεξάρτητων μεταβλητών. Θα μπορούσαμε κάλλιστα να μην βρίσκαμε διαφορές των μέσων σε μονομεταβλητό επίπεδο, αλλά ο γραμμικός συνδυασμός τους να μας έδινε στατιστικά σημαντικό αποτέλεσμα.

Εδώ τελειώνει ο έλεγχος της υπόθεσης ότι οι μέσες τιμές διαφέρουν σε κάθε γκρουπ για κάθε εξαρτημένη μεταβλητή. Αφού πήραμε αποτέλεσμα στατιστικά σημαντικό, το output θα εκτελέσει την διαδικασία Post-hoc testing με σκοπό να

γίνουν πολλαπλές πολλαπλές συγκρίσεις. Με αυτό τον τρόπο θα διακρίνουμε που υπάρχουν αυτές οι διαφορές και πόσο μεγάλες είναι.

Το στατιστικό που χρησιμοποιεί ο LSD έλεγχος είναι:

$$t = \frac{\left|\left(Y_j - Y_{j'}\right)\right|}{\sqrt{\left[MS_w / \left(1/n_j + 1/n_{j'}\right)\right]}}$$

όπου MS_w το μέσο τετραγωνικό σφάλμα , Y_j η μέση τιμή του γκρουπ j , $Y_{j'}$ η μέση τιμή του γκρουπ j', n_j αριθμός των αντικειμένων στο γκρουπ j και αντίστοιχα $n_{j'}$ ο αριθμός των αντικειμένων στο γκρουπ j'.

Ο πίνακας 4.7 Multiple Comparisons μας δείχνει τις πολλαπλές συγκρίσεις μεταξύ των γκρουπ, δηλαδή πόσο είναι οι διαφορές και αν είναι στατιστικά σημαντικές.

Πινακας 4.7

Multiple Comparisons

LSD

Dependent Variable	(I) quality_likert	(J) quality_likert	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Monthly Current income	πολύ χαμηλή	χαμηλή	124,35*	51,077	,016	23,61	225,08
		μέτρια	124,41*	45,695	,007	34,29	214,53
		υψηλή	46,14	45,419	,311	-43,44	135,71
		πολύ υψηλή	-116,25	61,468	,060	-237,48	4,98
	χαμηλή	πολύ χαμηλή	-124,35*	51,077	,016	-225,08	-23,61
		μέτρια	,06	30,307	,998	-59,71	59,84
		υψηλή	-78,21*	29,890	,010	-137,16	-19,26
		πολύ υψηλή	-240,60*	51,077	,000	-341,33	-139,86
	μέτρια	πολύ χαμηλή	-124,41*	45,695	,007	-214,53	-34,29
		χαμηλή	-,06	30,307	,998	-59,84	59,71
		υψηλή	-78,27*	19,302	,000	-116,34	-40,20
		πολύ υψηλή	-240,66*	45,695	,000	-330,78	-150,54
	υψηλή	πολύ χαμηλή	-46,14	45,419	,311	-135,71	43,44
		χαμηλή	78,21*	29,890	,010	19,26	137,16
		μέτρια	78,27*	19,302	,000	40,20	116,34
		πολύ υψηλή	-162,39*	45,419	,000	-251,96	-72,81
	πολύ υψηλή	πολύ χαμηλή	116,25	61,468	,060	-4,98	237,48
		χαμηλή	240,60*	51,077	,000	139,86	341,33
		μέτρια	240,66*	45,695	,000	150,54	330,78
		υψηλή	162,39*	45,419	,000	72,81	251,96
Monthly nightlife	πολύ χαμηλή	χαμηλή	3,79*	1,267	,003	1,29	6,28
		μέτρια	4,57*	1,133	,000	2,33	6,80
		υψηλή	3,21*	1,126	,005	,99	5,43
		πολύ υψηλή	-,75	1,524	,623	-3,76	2,26
	χαμηλή	πολύ χαμηλή	-3,79*	1,267	,003	-6,28	-1,29
		μέτρια	,78	,752	,301	-,70	2,26
		υψηλή	-,57	,741	,440	-2,03	,89
		πολύ υψηλή	-4,54*	1,267	,000	-7,03	-2,04
	μέτρια	πολύ χαμηλή	-4,57*	1,133	,000	-6,80	-2,33
		χαμηλή	-,78	,752	,301	-2,26	,70
		υψηλή	-1,35*	,479	,005	-2,30	-,41
		πολύ υψηλή	-5,32*	1,133	,000	-7,55	-3,08
	υψηλή	πολύ χαμηλή	-3,21*	1,126	,005	-5,43	-,99
		χαμηλή	,57	,741	,440	-,89	2,03
		μέτρια	1,35*	,479	,005	,41	2,30
		πολύ υψηλή	-3,96*	1,126	,001	-6,18	-1,74
	πολύ υψηλή	πολύ χαμηλή	,75	1,524	,623	-2,26	3,76
		χαμηλή	4,54*	1,267	,000	2,04	7,03
		μέτρια	5,32*	1,133	,000	3,08	7,55
		υψηλή	3,96*	1,126	,001	1,74	6,18

Based on observed means.
The error term is Mean Square(Error) = 9,294.

*. The mean difference is significant at the ,05 level.

Όπως βλέπουμε δεν υπάρχουν σε όλες τις συγκρίσεις στατιστικά σημαντικές διαφορές. Ας δούμε για παράδειγμα την σύγκριση μεταξύ των ατόμων που δήλωσαν ότι η ποιότητα ζωής τους είναι πολύ υψηλή σε σχέση με τα υπόλοιπα γκρουπ στην εξαρτημένη μεταβλητή ‘’τωρινό μηνιαίο εισόδημα’’.

Η μέση διαφορά στο ποσό χρημάτων μεταξύ των γκρουπ ‘’Πολύ υψηλή’’ και ‘’υψηλή’’ είναι 162.39 ευρώ και θεωρείται στατιστικά σημαντική σε επίπεδο σημαντικότητας 0.05. Η διαφορές είναι μεγαλύτερες καθώς η ποιότητα ζωής μειώνεται, όπως περιμέναμε. Μεταξύ ‘’Πολύ υψηλή’’ και ‘’Μέτρια’’ και ‘’Πολύ υψηλή’’ και ‘’Χαμηλή’’ υπάρχει μία μέση διαφορά των 240 ευρώ μηνιαίως. Και σε αυτές τις διαφορές υπάρχει στατιστικά σημαντική διαφορά σε επίπεδο σημαντικότητας 0.05.

Η τελευταία σύγκριση μεταξύ ‘’Πολύ υψηλή’’ και ‘’Πολύ χαμηλή’’ δεν είναι αυτό που περιμέναμε. Δεν υπάρχει στατιστικά σημαντική διαφορά στις μέσες τιμές, η μέση διαφορά εντοπίζεται μόλις στα 116.25 ευρώ. Σε αυτή την περίπτωση μπορούμε να υποθέσουμε ότι τα 8 άτομα που βρίσκονται στο γκρουπ που απάντησαν ‘’Πολύ χαμηλή’’, δεν απάντησαν στην ερώτηση ‘’Χαρακτηρίστε την ποιότητα ζωής σας στο Καρλόβασι από οικονομική άποψη’’ αφού το μηνιαίο εισόδημα τους είναι πολύ υψηλό και η βραδινές τους έξοδοι είναι πολλές.

Κεφάλαιο 5

Ανάλυση Σε Κύριες Συνιστώσες

5.1 Εισαγωγή

Η μέθοδος των κύριων συνιστωσών (Principal Component Analysis) είναι μία μέθοδος η οποία έχει σκοπό να δημιουργήσει γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε οι γραμμικοί αυτοί συνδυασμοί να είναι ασυσχέτιστοι μεταξύ τους, αλλά να περιέχουν όσο γίνεται μεγαλύτερο μέρος της διακύμανσης των αρχικών μεταβλητών. Το κέρδος από μία τέτοια διαδικασία είναι πως:

- Από ένα σύνολο συσχετισμένων μεταβλητών καταλήγουμε σε ένα σύνολο ασυσχέτιστων μεταβλητών, κάτι το οποίο για ορισμένες στατιστικές μεθόδους είναι περισσότερο χρήσιμο.
- Αν οι κύριες συνιστώσες που θα προκύψουν μπορούν να ερμηνεύσουν ένα μεγάλο ποσοστό της διακύμανσης, τότε αυτό σημαίνει πως αντί να έχουμε p μεταβλητές, όπως είχαμε αρχικά , έχουμε λιγότερες, με κόστος βέβαια ότι χάνουμε κάποιο ποσοστό της συνολικής μεταβλητότητας.
- Ένα άλλο πλεονέκτημα είναι πως με την μέθοδο των κύριων συνιστωσών μπορούμε να εξετάσουμε τις συσχετίσεις ανάμεσα στις μεταβλητές και να διαπιστώσουμε πόσο οι μεταβλητές μοιάζουν ή όχι. Επίσης , η μέθοδος μας επιτρέπει να αναγνωρίσουμε δίνοντας ονόματα στις καινούργιες μεταβλητές (συνιστώσες) παρατηρώντας ποιες από τις αρχικές μεταβλητές έχουν μεγάλη επίδραση σε αυτές.

5.2 Η βασική ιδέα

Έστω ένας τετραγωνικός συμμετρικός πίνακας A διαστάσεων . Ο πίνακας αυτός μπορεί να αναπαρασταθεί ως

$$A = P\Lambda P'$$

όπου Λ είναι ένας $p \times p$ διαγώνιος πίνακας όπου τα στοιχεία της διαγωνίου είναι οι ιδιοτιμές του πίνακα A , δηλαδή

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & \\ \dots & & & \\ 0 & \dots & \dots & \lambda_p \end{bmatrix}$$

και P είναι ένας ορθογώνιος $p \times p$ πίνακας ο οποίος αποτελείται από τα κανονικοποιημένα ιδιοδιανύσματα των αντίστοιχων ιδιοτιμών. Η παραπάνω

αναπαράσταση του πίνακα A ονομάζεται φασματική ανάλυση του πίνακα A. Επομένως, αφού ο πίνακας P είναι ορθογώνιος, θα ισχύει $P^{-1} = P^{'}$.

Μπορεί κάποιος να δείξει με βάση τις παραπάνω ιδιότητες πως ισχύει

$$\Lambda = P^{'} A P$$

καθώς $$A = P\Lambda P^{'} \Leftrightarrow P^{-1}A = P^{-1}P\Lambda P^{-1} \Leftrightarrow$$

$$\Leftrightarrow P^{-1}AP = \Lambda P^{'}P = \Lambda .$$

Δηλαδή ,ξεκινήσαμε από ένα τετραγωνικό πίνακα A και καταλήξαμε σε έναν διαγώνιο πίνακα Λ. Αυτό είναι σημαντικό επειδή αν ο τετραγωνικός πίνακας είναι πίνακας διακύμανσης, καταλήγουμε σε ένα διαγώνιο πίνακα διακύμανσης. Δηλαδή το τυχαίο διάνυσμα που αντιστοιχεί στον πίνακα αυτόν είναι ασυσχέτιστο.

5.3 Εύρεση των Κύριων Συνιστωσών

Η μέθοδος στηρίζεται στην φασματική ανάλυση ενός τετραγωνικού πίνακα. Αυτό σημαίνει πως μπορούμε να χρησιμοποιήσουμε είτε τον πίνακα διακύμανσεων είτε τον πίνακα συσχετίσεων που είναι στην ουσία ο πίνακας διακυμάνσεων των τυποποιημένων δεδομένων.

Έστω λοιπόν, πως έχουμε ένα σύνολο από p μεταβλητές $\left(X_1, X_2, ..., X_p\right)$ και θέλουμε να δημιουργήσουμε τις κύριες συνιστώσες $\left(Y_1, Y_2, ..., Y_p\right)$ οι οποίες να είναι γραμμικός συνδυασμός των αρχικών μεταβλητών, δηλαδή

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\ &.... \\ Y_p &= a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p \end{aligned}$$

Υπό μορφή πινάκων μπορεί να γραφεί ως $Y = AX$ όπου Y, X είναι διανύσματα $p \times 1$ και A είναι $p \times p$ πίνακας με στοιχεία

$$A = \begin{bmatrix} a_{11} & a_{12} & ... & a_{1p} \\ a_{21} & a_{22} & ... & a_{2p} \\ ... & & ... & \\ a_{p1} & a_{p2} & ... & a_{pp} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & ... & a_p \end{bmatrix}$$

όπου a_j είναι το διάνυσμα στήλη με στοιχεία $a_j^{'} = \left[a_{j1} \quad a_{j2} \quad ... \quad a_{jp} \right], j = 1,..., p$ και για να μην υπάρχουν προβλήματα ταυτοποίησης θέτουμε $\sum_{i=1}^{p} a_{ji}^{\ 2} = a_j^{'} a_j = 1$.

Οπότε, το πρόβλημα εύρεσης των κύριων συνιστωσών είναι το πρόβλημα εύρεσης των στοιχείων του πίνακα A. Έχουμε έναν ακόμα περιορισμό, ότι δηλαδή οι κύριες συνιστώσες πρέπει να είναι σε φθίνουσα σειρά ως προς την διακύμανση τους, δηλαδή η πρώτη να έχει μεγαλύτερη διακύμανση, η δεύτερη να έχει την δεύτερη μεγαλύτερη διακύμανση και ούτω καθεξής.

Αν δουλέψουμε για την πρώτη συνιστώσα $Y_1 = a_1^{'} X$, είναι σαφές πως $Var(Y_1) = a_1^{'} \Sigma a_1$, όπου Σ ο πίνακας διακυμάνσεων του τυχαίου διανύσματος X. Επομένως για να βρούμε το a_1 θα πρέπει να μεγιστοποιήσουμε την $Var(Y_1)$ με τον περιορισμό πως $a_1^{'} a_1 = 1$, δηλαδή θα μεγιστοποιήσουμε την συνάρτηση

$$L(a_1) = a_1^{'} \Sigma a_1 - \lambda \left(a_1^{'} a_1 - 1 \right)$$

όπου λ είναι ο πολλαπλασιαστής Lagrange.

Χρησιμοποιώντας παραγώγους διανυσμάτων, βρίσκουμε πως

$$\frac{\partial L(a_1)}{\partial a_1} = 2 \left(\Sigma - \lambda I \right) a_1 = 0$$

και επομένως αντιστοιχεί στο να λύσουμε την εξίσωση

$$\Sigma a_1 = \lambda a_1$$

η οποία είναι η εξίσωση των ιδιοδιανυσμάτων του πίνακα Σ όπου λ είναι η ιδιοτιμή.

Δηλαδή κάθε ζεύγος ιδιοτιμής και του ιδιοδιανύσματος που την συνοδεύει είναι λύση της εξίσωσης, και άρα έχει p δυνατές λύσεις. Από αυτές πρέπει να διαλέξουμε ποια οδηγεί σε μεγαλύτερη διακύμανση. Για την εύρεση των υπόλοιπων κύριων συνιστωσών χρειάζεται να προσθέσουμε ακόμη έναν περιορισμό, ότι οι κύριες συνιστώσες είναι ασυσχέτιστες με τις προηγούμενες τους.

Καταλήγοντας:

- Για να κατασκευάσουμε τις κύριες συνιστώσες, χρειάζεται να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα Σ που χρησιμοποιούμε.
- Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμα της αντιστοιχούν στην πρώτη κύρια συνιστώσα, η δεύτερη μεγαλύτερη ιδιοτιμή στην δεύτερη κύρια συνιστώσα κ.λ.π.
- Η διακύμανση της κάθε συνιστώσας είναι ίση με την ιδιοτιμή που της αντιστοιχεί. Έτσι αν συμβολίσουμε με λ_j την j μεγαλύτερη ιδιοτιμή, τότε έχουμε πως $Var(Y_1) = \lambda_j$.
- Όπως είπαμε πριν, οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους και άρα ο πίνακας διακύμανσης τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές λ_j .

- Η συνολική διακύμανση των κύριων συνιστωσών θα είναι η ίδια με την συνολική διακύμανση των αρχικών μεταβλητών εξαιτίας των ιδιοτήτων του ίχνους συμμετρικού και τετραγωνικού πίνακα. Δηλαδή, θα ισχύει $tr(\Sigma) = tr(\Lambda)$ και άρα η συνολική διακύμανση διατηρείται.
- Η ποσότητα $\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$ μας δείχνει το ποσοστό της συνολικής διακύμανσης που εξηγεί η j συνιστώσα.

5.4 Αλλαγή κλίμακας

Ένα από τα μειονεκτήματα της ανάλυσης σε κύριες συνιστώσες, χρησιμοποιώντας τον πίνακα διακύμανσης, είναι πως, αν η κλίμακα μέτρησης των δεδομένων είναι διαφορετική, τότε αλλάζουν και οι κύριες συνιστώσες και η ερμηνεία τους. Αν μία μεταβλητή έχει πολύ μεγαλύτερη διακύμανση από τις υπόλοιπες, αυτή τείνει να ταυτιστεί με την πρώτη κύρια συνιστώσα.

Ένας τρόπος να ξεπεράσουμε τις κακές αυτές ιδιότητες της ανάλυσης σε κύριες συνιστώσες στον πίνακα διακύμανσης είναι να χρησιμοποιήσουμε τον πίνακα συσχετίσεων. Οι συσχετίσεις δεν αλλάζουν, όταν αλλάξουν οι μονάδες μέτρησης ή η κλίμακα. Επίσης, στην ουσία δίνουν ίδιο βάρος σε όλες τις μεταβλητές, καθώς όλα τα στοιχεία της διαγωνίου είναι 1, και άρα τα προβλήματα που δημιουργούσε ο πίνακας διακύμανσης μπορούν να ξεπεραστούν.

Από την άλλη μεριά, η γενικευμένη χρήση του πίνακα συσχετίσεων δεν ενδείκνυται, καθώς η διαφορά στις διακύμανσεις ενδέχεται να περιέχει πληροφορία πολύτιμη για το θέμα που εξετάζουμε. Ίσως δηλαδή κάποιες μεταβλητές να πρέπει να θεωρηθούν πως έχουν μεγαλύτερο βάρος εξαιτίας της και επομένως θέτοντας όλες τις μεταβλητές να έχουν το ίδιο βάρος χάνουμε χρήσιμη πληροφορία. Μία καλή στρατηγική είναι να αποφεύγουμε τον πίνακα διακύμανσης, όταν υπάρχουν κάποιες μεταβλητές με πολύ μεγαλύτερη διακύμανση από ότι οι υπόλοιπες.

5.5 Εφαρμογή της ανάλυσης κύριων συνιστωσών στο SPSS

Για να ξεκινήσουμε την διαδικασία επιλέγουμε στο Spss διαδρομή Analyze → Dimension Reduction → Factor Analysis. Από την επιλογή Extraction επιλέγουμε Principal Component Analysis.

Ο πρώτος πίνακας μας δίνει ορισμένα περιγραφικά στοιχεία για τις μεταβλητές που θα χρησιμοποιήσουμε όπως την μέση τιμή και την τυπική απόκλιση και τον αριθμό των παρατηρήσεων που θα χρησιμοποιήσει η μέθοδος.

Πίνακας 5.1

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Monthly income 3 years ago	402,50	143,687	200
Monthly Current income	320,50	134,384	200
Monthly nightlife	5,00	3,287	200
Bank_savings	131,48	455,718	200
age in years	22,42	1,595	200

Ο παρακάτω πίνακας μας δίνει τις συσχετίσεις ανάμεσα στις αρχικές μεταβλητές που θα χρησιμοποιήσουμε. Πριν διεξάγουμε την ανάλυση κύριων συνιστωσών πρέπει να ελέγξουμε τις συσχετίσεις μεταξύ των μεταβλητών. Αν υπάρχουν τιμές πολύ υψηλές (μεγαλύτερες του 0.9) θα χρειαστεί να αφαιρέσουμε μία από τις μεταβλητές, καθώς οι δύο μεταβλητές θα φαίνονται να μετρούν το ίδιο πράγμα. Σε περίπτωση που οι συσχετίσεις είναι πολύ μικρές (μικρότερες του 0.1) τότε μπορεί μία ή περισσότερες μεταβλητές να προσαρμόζονται σε μία μόνο συνιστώσα. Με άλλα λόγια , να δημιουργήσει δική της συνιστώσα. Αυτό δεν είναι χρήσιμο γιατί σκοπός μας είναι να μειώσουμε τον αριθμό των μεταβλητών.

Πίνακας 5.2

Correlation Matrix

		Monthly income 3 years ago	Monthly Current income	Monthly nightlife	Bank_savings	age in years
Correlation	Monthly income 3 years ago	1,000	,755	,291	,003	,161
	Monthly Current income	,755	1,000	,359	,104	,058
	Monthly nightlife	,291	,359	1,000	,044	-,031
	Bank_savings	,003	,104	,044	1,000	-,039
	age in years	,161	,058	-,031	-,039	1,000

Ποιες όμως θεωρούνται ικανοποιητικές συσχετίσεις ώστε να προχωρήσουμε στην ανάλυση; Δεν χρειάζεται απαραίτητα μια συσχέτιση να είναι στατιστικά σημαντική, σύμφωνα με το αποτέλεσμα κάποιου ελέγχου υποθέσεων. όπως βλέπουμε υπάρχουν ορισμένες ισχυρές, μέτριες και χαμηλές συσχετίσεις μεταξύ των μεταβλητών μας.

Ο πίνακας 5.3 θα μας δώσει την τελική απόφαση αν οι μεταβλητές είναι κατάλληλες για την ανάλυση.

Πίνακας 5.3

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,557
Bartlett's Test of Sphericity	Approx. Chi-Square	205,690
	df	10
	Sig.	,000

Η τιμή του Kaiser –Meyer-Olkin Measure of Sampling Adequacy κυμαίνεται μεταξύ του 0 και 1. Τιμές κοντά στο 1 είναι καλύτερες. Με ελάχιστη τιμή το 0.5 οι μεταβλητές μας είναι κατάλληλες για την ανάλυση και μπορούμε να συνεχίσουμε.

Το Bartlett’s test of Sphericity ελέγχει την μηδενική υπόθεση ότι ο πίνακας συσχετίσεων είναι ο μοναδιαίος πίνακας. Ένας μοναδιαίος πίνακας είναι ο πίνακας που έχει όλα τα στοιχεία της διαγωνίου μονάδες και τα υπόλοιπα 0. Θέλουμε να απορρίψουμε την μηδενική υπόθεση.

Παίρνοντας υπόψη και τα δύο παραπάνω τεστ, μας παρέχουν ένα ελάχιστο όριο που πρέπει να περάσουν οι μεταβλητές ώστε να είναι κατάλληλες για την ανάλυση.

Πίνακας 5.4

Communalities

	Initial	Extraction
Monthly income 3 years ago	1,000	,810
Monthly Current income	1,000	,823
Monthly nightlife	1,000	,419
age in years	1,000	,613
Bank_savings	1,000	,402

Extraction Method: Principal Component Analysis.

Ο παραπάνω πίνακας Communalities μας δείχνει στην στήλη Extraction το ποσοστό της κάθε μεταβλητής που μπορεί να εξηγηθεί από τις κύριες συνιστώσες. Μεταβλητές με μεγάλες τιμές προσαρμόζονται ικανοποιητικά στον κοινό χώρο των παραγόντων σε αντίθεση με τις χαμηλές τιμές.

Πίνακας 5.5

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1,995	39,907	39,907	1,995	39,907	39,907	1,994	39,872	39,872
2	1,072	21,447	61,353	1,072	21,447	61,353	1,074	21,481	61,353
3	,956	19,129	80,482						
4	,746	14,918	95,400						
5	,230	4,600	100,000						

Extraction Method: Principal Component Analysis.

Η στήλη Component περιέχει τόσες συνιστώσες όσες και οι μεταβλητές μας. Βάλαμε δηλαδή 5 μεταβλητές για ανάλυση και πήραμε 5 συνιστώσες.

Initial Eigenvalues- οι ιδιοτιμές είναι οι διακυμάνσεις των κύριων συνιστωσών. Επειδή για την ανάλυση χρησιμοποιήσαμε τον πίνακα συσχετίσεων, οι μεταβλητές τυποποιήθηκαν, το οποίο σημαίνει ότι κάθε μεταβλητή έχει διακύμανση 1 και η συνολική διακύμανση ισούται με τον αριθμό των μεταβλητών που χρησιμοποιήθηκαν για την ανάλυση.

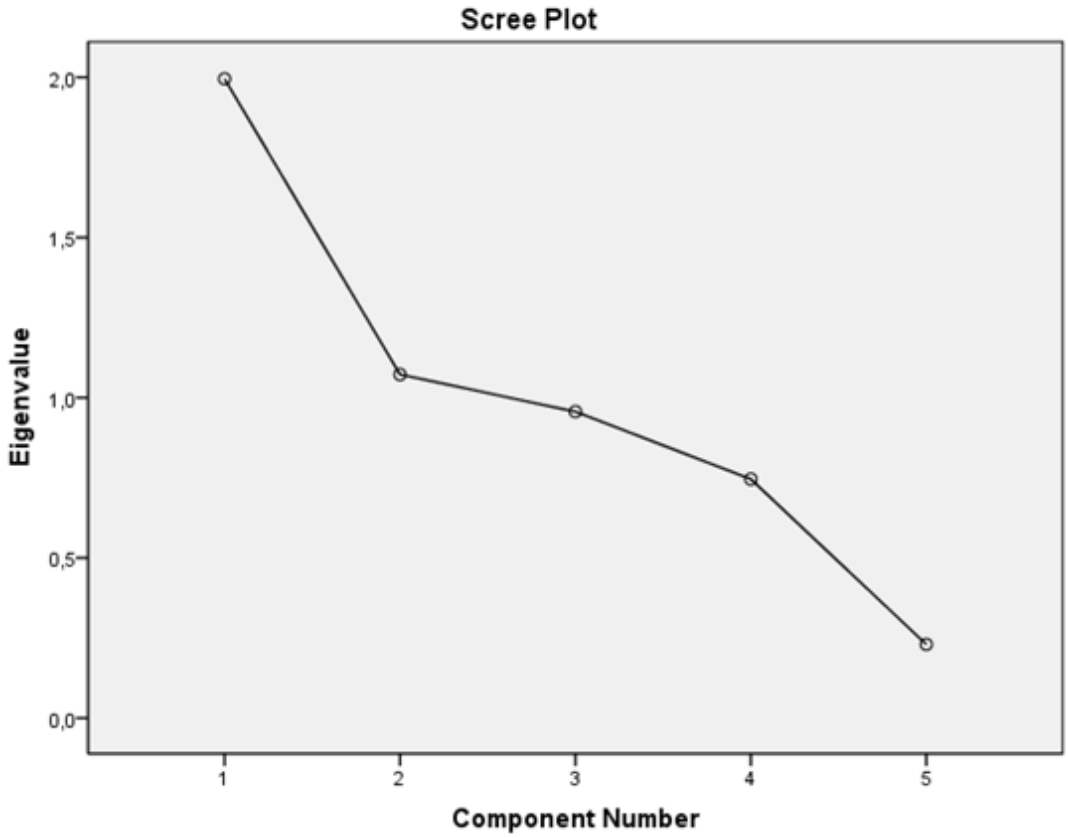
Η στήλη Total περιέχει τις ιδιοτιμές . Η πρώτη συνιστώσα πάντα θα αντιπροσωπεύει την περισσότερη διακύμανση (ως εκ τούτου θα έχει και την μεγαλύτερη ιδιοτιμή), η δεύτερη θα περιέχει όσο το δυνατόν περισσότερο διακύμανση απέμεινε και ούτω καθεξής.

Οι τρεις στήλες κάτω από το Extraction Sums of Squared Loadings δίνει τα ίδια αποτελέσματα με την αριστερή μεριά του πίνακα μόνο που έχει κρατήσει τις ιδιοτιμές που είναι μεγαλύτερες της μονάδας.

Τέλος , η δεξιά μεριά του πίνακα παρουσιάζει τα ίδια αποτελέσματα μετά την περιστροφή. Επιλέξαμε την περιστροφή Varimax που προσπαθεί να ελαχιστοποιήσει τον αριθμό των μεταβλητών που έχουν μεγάλες επιβαρύνσεις για κάθε παράγοντα.

Όπως βλέπουμε οι δύο συνιστώσες που έχουν ιδιοτιμές μεγαλύτερες της μονάδας εξηγούν το 61.353% της συνολικής διακύμανσης.

Διάγραμμα 5.1



Το Scree Plot δείχνει γραφικά στον Υ άξονα τις ιδιοτιμές και στον Χ τις συνιστώσες. Όπως βλέπουμε από την δεύτερη συνιστώσα και μετά η γραμμή τείνει να γίνει επίπεδη, αυτό σημαίνει ότι κάθε συνιστώσα εξηγεί όλο και λιγότερη διακύμανση από την συνολική. Γενικά, μας ενδιαφέρει να κρατήσουμε μόνο τις συνιστώσες που έχουν ιδιοτιμές μεγαλύτερες της μονάδας. Εκείνες που έχουν μικρότερη ιδιοτιμή, εξηγούν λιγότερη διακύμανση από την αρχική που είχε 1 και έτσι δεν είναι χρήσιμες.

πίνακας 5.6

Component Matrix[a]

	Component	
	1	2
Monthly income 3 years ago	,887	-,154
Monthly Current income	,906	,039
Monthly nightlife	,586	,276
age in years	,174	-,763
Bank_savings	,116	,624

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Ο πίνακας 5.6 μας πληροφορεί για το πόσες συνιστώσες τελικά κρατήσαμε για την ανάλυση. Είναι ένας πίνακας συσχέτισης μεταξύ της μεταβλητής και της συνιστώσας και για αυτό οι τιμές κυμαίνονται από το -1 έως 1. Συσχετίσεις μικρότερες του 0.30 δεν είναι χρήσιμες. Συνήθως δεν προσπαθούμε να ερμηνεύσουμε τις συνιστώσες όπως στην παραγοντική ανάλυση, αλλά ενδιαφερόμαστε περισσότερο για τα score των συνιστωσών που έχουμε αποθηκεύσει.

Πίνακας 5.7

Reproduced Correlations

		Monthly income 3 years ago	Monthly Current income	Monthly nightlife	age in years	Bank_savings
Reproduced Correlation	Monthly income 3 years ago	,810[a]	,798	,477	,272	,007
	Monthly Current income	,798	,823[a]	,542	,128	,130
	Monthly nightlife	,477	,542	,419[a]	-,108	,240
	age in years	,272	,128	-,108	,613[a]	-,456
	Bank_savings	,007	,130	,240	-,456	,402[a]
Residual[b]	Monthly income 3 years ago		-,043	-,187	-,111	-,004
	Monthly Current income	-,043		-,183	-,070	-,026
	Monthly nightlife	-,187	-,183		,077	-,196
	age in years	-,111	-,070	,077		,417
	Bank_savings	-,004	-,026	-,196	,417	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities
b. Residuals are computed between observed and reproduced correlations. There are 7 (70,0%) nonredundant residuals with absolute values greater than 0.05.

Ο πίνακας 5.7 περιέχει δύο υποπίνακες, τις αναπαραγόμενες συσχετίσεις και τα κατάλοιπα.

Οι αναπαραγόμενες συσχετίσεις ή Reproduced Correlations είναι ο πίνακας συσχέτισης βασιζόμενος στην εξαγόμενες συνιστώσες. Θέλουμε αυτές τις τιμές να είναι όσο πιο κοντά γίνεται στον αρχικό πίνακα συσχετίσεων. Αυτό σημαίνει ότι θέλουμε τον πίνακα με τα κατάλοιπα, ο οποίος περιέχει τις διαφορές μεταξύ των αρχικών και αναπαραγόμενων συσχετίσεων να είναι κοντά στο μηδέν.

Ο πίνακας 5.8 περιέχει τις συνιστώσες μετά την περιστροφή Varimax. Με την περιστροφή των παραγόντων προσπαθούμε να κάνουμε του παράγοντες πιο ερμηνεύσιμους. Με την περιστροφή δεν αλλάζουν κάποια από τα χαρακτηριστικά του μοντέλου, όπως η καλή προσαρμοστικότητα και το ποσό της διακύμανσης συνδιακύμανσης που εξηγεί το μοντέλο, παρά μόνο τις τιμές των επιβαρύνσεων. Κάνοντας λοιπόν την περιστροφή, ελπίζουμε ότι οι επιβαρύνσεις κάποιων παραγόντων θα είναι μεγάλες σε απόλυτη κλίμακα, μόνο για κάποιες από τις μεταβλητές και έτσι, βλέποντας ποιες μεταβλητές εξαρτώνται με ποιους παράγοντες να δώσουμε μία ερμηνεία σε αυτούς.

Πίνακας 5.8

Rotated Component Matrix[a]

	Component	
	1	2
Montlhy income 3 years ago	,879	,192
Monthly Current income	,907	,000
Monthly nightlife	,597	-,250
age in years	,141	,770
Bank_savings	,143	-,618

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Με την μέθοδο της ανάλυσης κύριων Συνιστωσών καταφέραμε να μειώσουμε τις μεταβλητές μας από 5 σε 2. Αποθηκεύσαμε τις καινούργιες μεταβλητές ως Factor_1 και Factor_2 και μπορούμε να τιε χρησιμοποιήσουμε στο μέλλον για περεταίρω ανάλυση.

FAC1_1	FAC2_1
-,56044	,18342
-,24691	1,27712
1,07293	,55875
1,47476	-,27814
-,29472	-,58494
-,50654	-,31152
,12698	,11794
-1,51463	-,65394
-1,28663	-,60779
-1,00720	1,38978
,06153	-,20142
-,42980	-,24478
-1,01546	-,61826
-,43831	-,06735
,82035	1,43143
-,72298	,98017
-,33614	-,32445

ΚΕΦΑΛΑΙΟ 6

Διακριτική ανάλυση

6.1 Εισαγωγή

Η βασική ιδέα της διαχωριστικής ανάλυσης είναι να κατατάξει παρατηρήσεις σε γνωστούς πληθυσμούς με γνωστές κατανομές για κάθε πληθυσμό. Η διαχωριστική ή διακριτική ανάλυση (discriminant analysis) αποτελεί μια μέθοδο με πλήθος εφαρμογών σε πολλές επιστήμες.

Ας υποθέσουμε ότι έχουμε K πληθυσμούς (ομάδες) $\Pi_1, \Pi_2, ..., \Pi_k$ με $K \geq 2$. Τότε για κάθε πληθυσμό Π_k έχουμε μία κατανομή $f_k(x)$. Σκοπός της διαχωριστικής συνάρτησης είναι να «διαχωρίσει» ή να κατανείμει κάθε παρατήρηση στους K γνωστούς πληθυσμούς- ομάδες. Προφανώς, ψάχνουμε για ένα διαχωριστικό κανόνα που μπορεί να κατατάξει σωστά όσο το δυνατόν περισσότερες παρατηρήσεις.

Οι εφαρμογές είναι πάρα πολλές. Θα αναφέρουμε μερικά από τα παραδείγματα παρακάτω:

- Στην ιατρική συνήθως το ενδιαφέρον είναι να διαγνώσουμε την ασθένεια κάποιου ασθενή με βάση κάποια συμπτώματα που αυτός έχει. Δεδομένου πως για κάθε αρρώστια είναι γνωστά τα συμπτώματα της, θέλουμε να κατασκευάσουμε έναν κανόνα, ο οποίος λαμβάνοντας υπόψη τα συμπτώματα αλλά και την γνώση μας για τα συμπτώματα ενός συνόλου ασθενειών να κάνει διάγνωση για τον καινούργιο ασθενή.
- Στα χρηματοοικονομικά οι τράπεζες ενδιαφέρονται να εντοπίσουν «καλούς» ή «κακούς» πελάτες πριν την χορήγηση δανείου ή πιστωτικής κάρτας. Ως «καλούς» και «κακούς» μπορούμε να θεωρήσουμε αυτούς που πληρώνουν αυτούς που πληρώνουν κανονικά τις δόσεις του και αυτούς που δεν πληρώνουν αντίστοιχα. Συνεπώς με την χρήση ιστορικών στοιχείων σχετικά με τα άτομα που έλαβαν δάνειο από την τράπεζα μπορεί να σχηματίσει κανόνες, ώστε να κατατάξει έναν καινούργιο πελάτη σε μία από τις δύο κατηγορίες, και πιθανότατα, να αρνηθεί τη χορήγηση ενός δανείου.

Είναι ενδιαφέρον να δούμε ότι η κατάταξη γίνεται είτε σε δύο ομάδες είτε σε περισσότερες ομάδες. Τέλος, να αναφέρουμε ότι η διαχωριστική ανάλυση μοιάζει αρκετά με την ανάλυση σε συστάδες, που κάναμε παραπάνω, αλλά έχει σημαντικές διαφορές από αυτή. Η πρώτη και πιο σημαντική είναι ότι στην διαχωριστική ανάλυση οι ομάδες είναι γνωστές, ενώ στην ανάλυση κατά συστάδες δεν είναι γνωστές και σκοπός μας είναι να τις βρούμε. Για τον λόγο αυτόν ο στόχος είναι διαφορετικός. Στην διαχωριστική ανάλυση κύριο μέλημα μας είναι η κατασκευή ενός κανόνα που θα μας βοηθήσει να λάβουμε αποφάσεις στο μέλλον ενώ στην ανάλυση σε συστάδες ο κύριος στόχος μας είναι να δημιουργήσουμε ομοειδής ομάδες με σκοπό την κατανόηση των ήδη υπάρχοντων στοιχείων και την μείωση της διασποράς σε επιμέρους ομάδες.

6.2 Η λογική της διαχωριστικής Συνάρτησης του Fisher

Ο διαχωριστικός κανόνας του Fisher βασίζεται στη μετατροπή των χαρακτηριστικών x σε μονοδιάστατα σκορ μέσω μιας συνάρτησης, η οποία λέγεται διαχωριστική συνάρτηση (discriminant function). Τα σκορ των δύο ομάδων θα πρέπει να είναι όσο το πιο δυνατόν πιο απομακρυσμένα, έτσι ώστε να μπορούμε εύκολα με βάση αυτά τα σκορ να κάνουμε διαχωρισμό και ταξινόμηση των δύο ομάδων. Έτσι λοιπόν, ο Fisher πρότεινε την χρήση γραμμικών συνδυασμών για τη δημιουργία αυτών των σκορ, χωρίς να γίνει κάποια υπόθεση για την κατανομή των ομάδων. Η γραμμικότητα υιοθετήθηκε για λόγους ευκολίας. Παρόλα αυτά, υπέθεσε ισότητα των πινάκων διακύμανσης, αφού χρησιμοποίησε την συνδυασμένη κοινή εκτίμηση S_p.

Έστω λοιπόν, ότι τα σκορ δίνονται ως U_1 για την πρώτη ομάδα και ως U_2 για την δεύτερη ομάδα. Τότε ένα μέτρο του κατά πόσο κοντά είναι τα σκορ των δύο ομάδων δίνεται από την απόσταση των μέσων τιμών $\bar{U}_1-\bar{U}_2$. Ο Fisher μέτρησε αποστάσεις σε τυπικές αποκλίσεις και κατά απόλυτες τιμές , δηλαδή πήρε σας μέτρο απόστασης την ποσότητα $D=\dfrac{\left|\bar{U}_1-\bar{U}_2\right|}{S_U}$ με $S_U=\dfrac{\sum_{i\in G_1}\left(U_i-\bar{U}_1\right)^2+\sum_{i\in G_2}\left(U_i-\bar{U}_2\right)^2}{n_1+n_2-2}$ όπου $i\in G_k$ σημαίνει ότι λαμβάνουμε υπόψη τις παρατηρήσεις που ανήκουν στην k ομάδα. Σκοπός είναι να μεγιστοποιήσουμε την ποσότητα D ή αντίστοιχα, την απόσταση D^2, καθώς αυτό σημαίνει ότι τα σκορ των δύο ομάδων θα είναι όσο γίνεται πιο διαφορετικά μεταξύ τους.

Έστω ο γραμμικός συνδυασμός $L'x$ τότε πρέπει να μεγιστοποιήσουμε την ποσότητα $D^2=\dfrac{\left|(\bar{x}_1-\bar{x}_2)\right|^2}{L'S_pL}$.

Από την ανισότητα Cauchy- Schwartz έχουμε ότι για κάθε $p\times 1$ διανύσματα a και b ισχύει ότι $\left(a'b\right)\le\left(a'a\right)\left(b'b\right)$. Εφόσον ο πίνακας συνδιακυμάνσεων είναι θετικά ορισμένος, μπορούμε να θέσουμε $a=S_p^{1/2}L$ και $b=S_p^{-1/2}(\bar{x}_1-\bar{x}_2)$. Τότε έχουμε

$$\left|L'(\bar{x}_1-\bar{x}_2)\right|^2\le\left(L'S_p^{1/2}S_p^{1/2}L\right)\left[(\bar{x}_1-\bar{x}_2)'S_p^{-1/2}S_p^{-1/2}(\bar{x}_1-\bar{x}_2)\right]\Leftrightarrow$$

$$\left|L'(\bar{x}_1-\bar{x}_2)\right|^2\le\left(L'S_pL\right)\left[(\bar{x}_1-\bar{x}_2)'S_p^{-1/2}(\bar{x}_1-\bar{x}_2)\right]\Leftrightarrow$$

$$D^2 = \frac{\left[L'(\bar{x}_1 - \bar{x}_2) \right]^2}{L' S_p L} \leq (\bar{x}_1 - \bar{x}_2)' S_p^{-1} (\bar{x}_1 - \bar{x}_2).$$

Άρα για $L = cS_p^{-1}(\bar{x}_1 - \bar{x}_2)$, όπου $c > 0$, έχουμε $D^2 = (\bar{x}_1 - \bar{x}_2)' S_p^{-1} (\bar{x}_1 - \bar{x}_2)$ δηλαδή την μέγιστη απόσταση μεταξύ των μέσων και τον καλύτερο δυνατό διαχωρισμό. Το c είναι μία σταθερά και συνήθως παίρνουμε c=1. Ο διαχωριστικός κανόνας ολοκληρώνεται ορίζοντας την κρίσιμη τιμή, η οποία δεν είναι άλλη από τη μέση τιμή των $\bar{U}_1$ και $\bar{U}_2$, δηλαδή η ποσότητα $m = \left(\bar{U}_1 - \bar{U}_2 \right) / 2 = L'(\bar{x}_1 - \bar{x}_2) / 2$ η οποία ισαπέχει από τα $\bar{U}_1$ και $\bar{U}_2$. Έτσι ο διαχωριστικός κανόνας γίνεται: αν $L'x \geq m$ ή ισοδύναμα $L'x - m \geq 0$, τότε κατατάσσουμε στην 1η ομάδα αλλιώς στην 2η.

Σχόλια: Υπάρχουν και άλλοι κανόνες διαχωρισμού των ομάδων εκτός του Fisher. Στο βιβλίο Πολυμεταβλητή Στατιστική Ανάλυση του Καρλή Δημήτριου μπορούμε να βρούμε τον Κανόνα Μέγιστης πιθανοφάνειας, τον Κανόνα του Bayes και τον κανόνα Διαχωρισμού δύο ομάδων με την χρήση της Κανονικής Κατανομής.

6.3 Εφαρμογή της διαχωριστικής ανάλυσης στο SPSS

Κοιτάζοντας τα δεδομένα που έχουμε διαθέσιμα ενδιαφερόμαστε να κατασκευάσουμε έναν διαχωριστικό κανόνα με σκοπό να ταξινομήσουμε τον φοιτητικό πληθυσμό αν έχει ανάγκη την παροχή φοιτητικής εστίας του πανεπιστημίου. Ο διαχωρισμός θα λάβει υπόψη του το χρηματικό ποσό που λαμβάνει κάθε μήνα και των αριθμό των βραδινών του εξόδων για διασκέδαση. Με αυτές τις δύο ανεξάρτητες μεταβλητές θα μπορούμε να προβλέψουμε και να κατατάξουμε σε ομάδες φοιτητές που έχουν ανάγκη ή όχι.

Για να ξεκινήσουμε επιλέγουμε Analyze→ Classify→ Discriminant Analysis.

Ο πρώτος πίνακας μας δίνει περιγραφικούς δείκτες, όπως μέγεθος του δείγματος, μέση τιμή και τυπική απόκλιση.

Πίνακας 6.1

Group Statistics

housing		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Εστία	Monthly Current income	204,06	92,313	32	32,000
	Monthly nightlife	3,75	3,417	32	32,000
Ενοίκιο	Monthly Current income	342,68	129,788	168	168,000
	Monthly nightlife	5,24	3,203	168	168,000
Total	Monthly Current income	320,50	134,384	200	200,000
	Monthly nightlife	5,01	3,275	200	200,000

Είναι εμφανές ότι υπάρχουν διαφορές στις μέσες τιμές των δύο ομάδων Εστία και Ενοίκιο. Η επιλογή στις ρυθμίσεις της μεθόδου για ανάλυση διακύμανσης παράγει τον ακόλουθο πίνακα 6.2

Πίνακας 6.2

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Monthly Current income	,856	33,232	1	198	,000
Monthly nightlife	,972	5,725	1	198	,018

από τον οποίο βλέπουμε ότι και για τις δύο μεταβλητές μας, οι μέσες τιμές στις δύο ομάδες διαφοροποιούνται σημαντικά, τα p-value είναι μικρότερα του 0.05. Ο Fisher υπέθεσε ισότητα των πινάκων συνδιακύμανσης. Ο έλεγχος γίνεται με την βοήθεια του τεστ Box M που βλέπουμε στον παρακάτω πίνακα.

Πίνακας 6.3

Test Results

Box's M		6,195
F	Approx.	2,015
	df1	3
	df2	39732,942
	Sig.	,109

Tests null hypothesis of equal population covariance matrices.

Το συγκεκριμένο τεστ είναι πολύ ευαίσθητο σε αποκλίσεις από την κανονική κατανομή. Έτσι βλέπουμε ότι p-value είναι μεγαλύτερο του 0.05 επομένως δεν απορρίπτουμε την μηδενική υπόθεση που υποθέτει ίσους πίνακες συνδιακύμανσης.

Συνεχίζουμε με τον υπολογισμό και την εμφάνιση των συντελεστών της διαχωριστικής συνάρτησης.

Πίνακας 6.4

Classification Function Coefficients

	housing	
	Εστία	Ενοίκιο
Monthly Current income	,011	,020
Monthly nightlife	,205	,230
(Constant)	-2,226	-4,713

Fisher's linear discriminant functions

Για κάθε ομάδα υπολογίζουμε ένα σκορ με βάση μια συνάρτηση. Οι συναρτήσεις αυτές είναι γραμμικές ως προς τις ανεξάρτητες μεταβλητές. Άρα σύμφωνα με την μέθοδο του Fisher έχουμε για την ομάδα **Εστία**:

$$w_1 = -2.226 + 0.11 \times Current\_Income + 0.205 \times nightlife$$

ενώ για την ομάδα **Ενοίκιο**:

$$w_2 = -0.713 + 0.020 \times Current\_Income + 0.230 \times nightlife \ .$$

Αν $w_1 > w_2$, κατατάσσουμε την παρατήρηση στην ομάδα Εστία, αλλιώς στην ομάδα Ενοίκιο. Διαφορετικά, αν $Z = w_1 - w_2$, τότε για $Z > 0$ κατατάσσουμε στην ομάδα Εστία , αλλιώς στην ομάδα Ενοίκιο.

Πίνακας 6.5

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Monthly Current income	,973
Monthly nightlife	,071

Ό παραπάνω πίνακας 6.5 μας δίνει μία ένδειξη της συνεισφοράς της κάθε μεταβλητής στην διαχωριστική συνάρτηση.

Όταν έχουμε παραπάνω από δύο ομάδες, οι ιδιοτιμές είναι χρήσιμες ως δείκτες μέτρησης της διασποράς των κέντρο-ειδών στον αντίστοιχο πολυμεταβλητό χώρο. Ο πίνακας 7.6 μας δίνει τον δείκτη κανονικής συσχέτισης που δηλώνει πόσο συσχέτιση υπάρχει μεταξύ των ομάδων και των σκορ της διαχωριστικής συνάρτησης.

Πίνακας 6.6

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,169[a]	100,0	100,0	,380

a. First 1 canonical discriminant functions were used in the analysis.

Από τον πίνακα που ακολουθεί μπορούμε να ελέγξουμε την υπόθεση ότι οι μέσοι όλων των μεταβλητών είναι ίδιοι ανά ομάδα είναι ίσοι. Όπως βλέπουμε απορρίπτουμε την ισότητα των μέσων, άρα δεν φαίνεται να υπάρχει πρόβλημα με την εφαρμογή της διαχωριστικής ανάλυσης.

Πίνακας 6.7

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	,856	30,693	2	,000

Ο πίνακας Structure matrix 6.8 μας δίνει τους δείκτες συσχέτισης κάθε ανεξάρτητης μεταβλητής με τις διαχωριστικές συναρτήσεις και μπορούν να χρησιμοποιήθουν για να αξιολογήσουμε πόσο σημαντική είναι κάθε μεταβλητή για την κατασκευή της διαχωριστικής συνάρτησης.

Πίνακας 6.8

Structure Matrix

	Function
	1
Monthly Current income	,998
Monthly nightlife	,414

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

Οι μεταβλητές μας δεν είναι το ίδιο σημαντικές, αφού οι συσχετίσεις τους διαφέρουν αρκετά.

Πίνακας 6.9

Classification Results[b,c]

			Predicted Group Membership		
		housing	Εστία	Ενοίκιο	Total
Original	Count	Εστία	27	5	32
		Ενοίκιο	47	121	168
	%	Εστία	84,4	15,6	100,0
		Ενοίκιο	28,0	72,0	100,0
Cross-validated[a]	Count	Εστία	27	5	32
		Ενοίκιο	47	121	168
	%	Εστία	84,4	15,6	100,0
		Ενοίκιο	28,0	72,0	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b. 74,0% of original grouped cases correctly classified.
c. 74,0% of cross-validated grouped cases correctly classified.

Ο πίνακας 6.9 είναι χρήσιμος για τον υπολογισμό της επιτυχίας της διαχωριστικής ανάλυσης. Συγκεκριμένα , το ποσοστό του σωστού διαχωρισμού είναι 74% για την συνολική διαχωριστική ανάλυση.

Όπως βλέπουμε από τους 32 συνολικά που μένουν στην Εστία εκτιμήσαμε σωστά τους 27 και τους 5 λάθος. Παρόμοια, στο Ενοίκιο 121 σωστά και 47 λάθος.

Έχοντας αποθηκεύσει την εκτίμηση μας ως κατηγορική μεταβλητή για την πρόβλεψη των ομάδων μπορούμε να υπολογίσουμε τις συχνότητες και να την συγκρίνουμε τις πραγματικές συχνότητες.

Πίνακας 6.10

Predicted Group for Analysis 1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Εστία	73	36,5	36,5	36,5
	Ενοίκιο	127	63,5	63,5	100,0
	Total	200	100,0	100,0	

Όπως βλέπουμε στο επόμενο πίνακα υπάρχει μία διαφορά της τάξης των 41 φοιτητών που χρειάζονται παροχή εστίας ενώ αναγκάζονται να νοικιάσουν σπίτι.

Πίνακας 6.11

housing

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Εστία	32	16,0	16,0	16,0
	Ενοίκιο	168	84,0	84,0	100,0
	Total	200	100,0	100,0	

6.4 Συμπεράσματα της μεθόδου

Εφαρμόζοντας την Διαχωριστική Ανάλυση στο δείγμα 200 φοιτητών που είχαμε διαθέσιμο καταφέραμε να κατασκευάσουμε έναν κανόνα ώστε να εκτιμήσουμε τον αριθμό τον ατόμων που χρειάζονται δωμάτιο στην εστία ή των το αριθμό των μελλοντικών φοιτητών που θα έχουν ανάγκη δωμάτιο στην εστία όταν έρθουν να σπουδάσουν στο Καρλόβασι.

Σύμφωνα με τις εξαρτημένες μεταβλητές που χρησιμοποιήσαμε, Μηνιαίο Εισόδημα και Βραδυνή Διασκέδαση είδαμε ότι υπάρχουν 41 φοιτητές που σύμφωνα με τις απαντήσεις τους στο ερωτηματολόγιο έχουν ανάγκη το δωμάτιο στην εστία αλλά αναγκάζονται να νοικιάσουν, με ποσοστό επιτυχίας 74% σωστής διαχώρισης.

Βέβαια το ποσοστό επιτυχίας δεν είναι μεγαλύτερο λόγω του ότι η φοιτητική μέριμνα του πανεπιστημίου δεν χρησιμοποιεί τα ίδια κριτήρια για να κάνει την διαχώριση των φοιτητών. Υπάρχουν τα κριτήρια της φορολογικής δήλωσης, των πολύτεκνων οικογενειών, αδέλφια που να σπουδάζουν σε διαφορετικά μέρη κ.τ.λ. Οι μεταβλητές που χρησιμοποιήσαμε αντικατοπτρίζουν σε κάποιο βαθμό τα παραπάνω.

Εκτίμησαμε τελικά ότι το 36.5% των φοιτητών στο Καρλόβασι, παίρνοντας δείγμα 200 ατόμων χρειάζονται δωμάτιο στην εστια. Αν θέλουμε να το μεταφέρουμε στο πληθυσμό (1400 ενεργοί φοιτητές) αυτό το ποσοστό αντιστοιχεί σε 511 άτομα, πολύ παραπάνω από ότι μπορεί να προσφέρει το πανεπιστήμιο Αιγαίου.

Βιβλιογραφία

Δημήτρης Καρλής. *Πολυμεταβλητή Στατιστική Ανάλυση. Εκδόσεις Σταμούλη, Αθήνα 2005*

Πανάρετος, Ι. Και Ξεκαλάκη,Ε (1995). Εισαγωγή στην Πολυμεταβλητή Στατιστική Ανάλυση. Εκδόσεις Πανάρετος, Αθήνα.

Joseph F. Hair , Willian C. Black, Barry J. Babin , Rolph E. Anderson. *Multivariate Data Analysis. Pearson Prentice Hall 2010*

Richard A. Johnson , Dean W. Wichern. *Applied Multivariate Statistical Analysis. Pearson Prentice Hall ,2007.*

Χαράλαμπος Γναρδέλλης, *Ανάλυση Δεδομένων με το Spss14.0 for Windows.*

http://www.statsoft.com

http://statistics.ats.ucla.edu

http://www.spss.ch/upload/1122644952_The%20SPSS%20TwoStep%20Cluster%20Component.pdf