ΠΟΛΥΔΙΑΣΤΑΤΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

➢Ιστορική αναδρομή

✓Στην αρχή του 20ου αιώνα οι Ευρωπαίοι ψυχολόγοι έψαχναν με τα τέστ που υπέβαλαν στους αρρώστους και τους βαθμούς που συγκέντρωναν από διάφορες μεταβλητές που χρησιμοποιούσαν (μνήμη, ευφυΐα, κ.ά), να βρουν σύνθετες μεταβλητές οι οποίες όχι μόνο δεν παρατηρούνται απ' ευθείας από τα αρχικά δεδομένα, αλλά θα ερμήνευαν κατά τον καλύτερο τρόπο τη συμπεριφορά των αρρώστων όσων βέβαια παρουσίαζαν τα ίδια συμπτώματα. Τις μεταβλητές αυτές τις ονόμασαν **«παράγοντες»**.

✓Επιστήμονες που δημιούργησαν τις προϋποθέσεις ανάδυσης μιας οικογένειας στατιστικών μεθόδων, με την ονομασία **Ανάλυση Δεδομένων**

•C. Spearman (1904) Εργασίες πάνω στη μήτρα Διακυμάνσεων-Συνδιακυμάνσεων

•H. Hottelling (1933) Ανάλυση σε κύριες Συνιστώσες

•R.A. Fisher (1940) διατυπώνει τους διακριτικούς παράγοντες

•L.Guttman (1941) Το ομώνυμο φαινόμενο

•C. Burt (1950) Τον ομώνυμο πίνακα

✓Πάντως η πιο πρόσφατη και αποτελεσματικότερη μέθοδος της οικογένειας δημιουργήθηκε κατά τη δεκαετία του ’60 από το Γάλλο καθηγητή J. P Benzecri (1973) του Πανεπιστημίου Paris VI με την ονομασία **Παραγοντική Ανάλυση των Αντιστοιχών (**Analyse Factorielle des Correspondances -A.F.C-), η οποία επεξεργάζεται κυρίως ποιοτικά δεδομένα που παρουσιάζονται υπό μορφή πολυδιάστατων πινάκων συμπτώσεων

✓Στα τέλη της δεκαετίας του ’70, αρχές της δεκαετίας του ’80, μαθητές του J.P. Benzecri, όπως οι Γάλλοι Lebart, Roux, Escoffier, Morineau, Fenelon συντελέσαν όχι μόνο στη διάδοση αλλά και την καθιέρωση των μεθόδων αυτών στη συνείδηση πολλών ερευνητών σ’ ολόκληρο τον κόσμο, δημιουργώντας την Γαλλική Σχολή της Ανάλυσης Δεδομένων.

✓Αλλά και στην άλλη μεριά του Ατλαντικού δημιουργήθηκε η λεγόμενη Αμερικανική σχολή με τους J.D Carrol, J. B. Kruskal, R. S Sheppard, G. Yang κ.λ.π κάτω από το όνομα «multidimensional scaling», της οποίας η ευρηματικότητα δεν συγκρίνεται με εκείνη της Γαλλικής Σχολής.

Ανάλυση Δεδομένων. Στατιστική δίχως μοντέλα

- Κλασική αντιμετώπιση

Οικονομία-Μαθηματικά-Στατιστική-Οικονομετρία

Μοντελοποίηση-Κατάλοιπα

Αλληλεξαρτήσεις

Κανονικός νόμος

- Νέα αντίληψη

Μη παραμετρική θεώρηση του υπό μελέτη φαινόμενου

Εντοπισμός των αλληλεξαρτημένων παραγόντων

Καμία υπόθεση για τη συμπεριφορά των στοιχείων

Μελέτη και ποιοτικών μεταβλητών

➢Η Πολυδιάστατη Στατιστική Ανάλυση ή απλά Ανάλυση Δεδομένων συγκεντρώνει πολυάριθμες στατιστικές μεθόδους διαφορετικές μεταξύ τους. Μπορούμε όμως να διακρίνουμε δύο κύριες κατευθύνσεις:

✓Την **παραγοντική ανάλυση** η οποία επιτρέπει να αποκαλύψουμε τη δομή και τις αλληλεξαρτήσεις ενός συνόλου, εξετάζοντας κάποιο παραγοντικό επίπεδο.
Η παραγοντική ανάλυση είναι ένα εργαλείο που χρησιμοποιείται από τον ερευνητή για να παρατηρήσει αυτό που θέλει με τον ίδιο τρόπο που το τηλεσκόπιο ή το μικροσκόπιο χρησιμοποιείται σε άλλους επιστημονικούς χώρους.

✓Την **αυτόματη ταξινόμηση**, η οποία συνίσταται στο να κατατάξει τις στατιστικές μονάδες σε ομοιογενείς ομάδες, οι οποίες επηρεάζονται ταυτόχρονα από διάφορους παράγοντες, με την βοήθεια κάποιου Αλγορίθμου ταξινόμησης.

ΠΟΛΥΔΙΑΣΤΑΤΑ ΔΕΔΟΜΕΝΑ

- **Ορισμοί**

Στατιστική μονάδα

Χαρακτηριστικά-Ποσοτικά-Ποιοτικά

- **Μορφές ερωτήσεων**

✓**Κλειστές**

- Μονότιμες
- Δίτιμες
- Πολλαπλών απαντήσεων

✓**Ανοικτές**

Δίνουν τη δυνατότητα ελεύθερης διατύπωσης της απάντησης

ΚΛΙΜΑΚΕΣ ΜΕΤΡΗΣΗΣ

Κάθε χαρακτηριστικό που μετράμε, το οποίο αντιστοιχεί σε μία μεταβλητή, απαιτεί και διαφορετική μετρική για να αποδώσουμε τα αριθμητικά δεδομένα τα οποία θα επεξεργαστούμε.

Μία μετρική είναι μία ποσοτικοποίηση των παρατηρήσεων, με βάση μία συγκεκριμένη κλίμακα, του μεγέθους που μας ενδιαφέρει να ελέγξουμε.

- Όσον αφορά ποσοτικές μεταβλητές
  - ✓Αναλογικές κλίμακες (συνεχείς)
  - ✓Αναλογικές κλίμακες (ασυνεχείς)

- Όσον αφορά ποιοτικές μεταβλητές
  - ✓Τύπος Ονομαστικής/ Κατηγορικής κλίμακας
  - ✓Τύπος ιεραρχικής/Τακτικής κλίμακας
  - ✓Τύπος διαβαθμισμένης κλίμακας (Likert-Guttman-Thurstone

ΚΩΔΙΚΟΠΟΙΗΣΗ ΤΩΝ ΠΟΛΥΔΙΑΣΤΑΤΩΝ ΔΕΔΟΜΕΝΩΝ

Η Πολυπαραγοντική Ανάλυση Δεδομένων εφαρμόζεται στις περιπτώσεις, όπου ο χρήστης διαθέτει στοιχεία καταχωρημένα σε πίνακες, που περιλαμβάνουν N «αντικείμενα» τα οποία περιγράφονται από K>2 ποσοτικές ή ποιοτικές ιδιότητες (ή συνδυασμό αυτών).

Επιδίωξη κάθε ανάλυσης στρέφεται κυρίως στο να περιγράψει :

✓Τις σχέσεις μεταξύ των μεταβλητών (ή των διαβαθμίσεών τους)
✓Τις σχέσεις ενδεχομένως μεταξύ αντικειμένων
✓Τις σχέσεις μεταξύ αντικειμένων και μεταβλητών

Τα δεδομένα παρουσιάζονται συνήθως σε διδιάστατους πίνακες Τ(N,K) (πίνακες διπλής εισόδου).

Γενική μορφή πίνακα πολυδιάστατων δεδομένων

Στατιστικές Μονάδες	Μεταβλητές $X_1\ X_2 \ldots X_j \ldots X_p$
1	.
2	.
.	.
i	..... $X_j(i)$
.	
n	

Η αριθμητική τιμή $X_j(i)$ που αντιστοιχεί στην i-οστή γραμμή και j-οστή στήλη είναι η τιμή που πήρε η μεταβλητή X_j στη i-οστή στατιστική μονάδα.

Α) Πίνακες στατιστικών μονάδων και ποσοτικών μεταβλητών

Γενική μορφή πίνακα στατιστικών μονάδων και ποσοτικών μεταβλητών

Στατιστικές Μονάδες	Μεταβλητές X_1 X_2 ... X_j.....X_p
1	.
2	.
.	.
i	..... k_{ij}
.	.
n	.

Παράδειγμα

Ο πίνακας 1.4 παρουσιάζει πέντε μάρκες αυτοκινήτων ως προς τέσσερα ποσοτικά κριτήρια:τον κυβισμό, την ισχύ της μηχανής, την ανώτατη ταχύτητα και την τιμή κάθε αυτοκινήτου.

Πίνακας 1.4

Μάρκες	ΚΡΙΤΗΡΙΑ			
	κυβισμός	ισχύς	ταχύτητα	τιμή σε €
Α	1396	90	174	13.950
Β	1721	92	180	15.780
Γ	1580	83	170	14.957
Δ	1769	90	180	15.134
Ε	1116	58	145	13.135

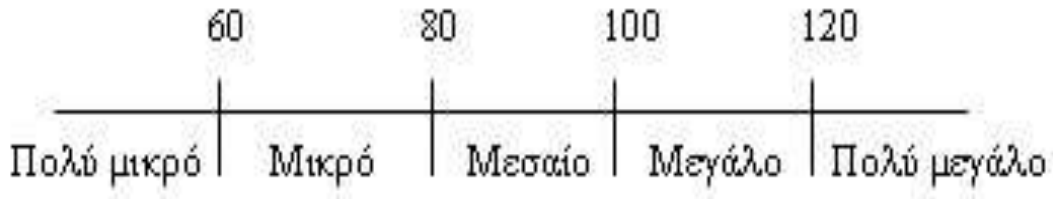
Β) Πίνακες συμπτώσεων

Ένας πίνακας συμπτώσεων παρουσιάζει την κατανομή των απολύτων συχνοτήτων nij των στατιστικών μονάδων ενός δείγματος σύμφωνα με δυο ποιοτικά χαρακτηριστικά ή ομογενοποιημένα ποσοτικά.

Ομογενοποίηση

Με τον όρο ομογενοποίηση μιας ποσοτικής μεταβλητής εννοούμε ότι θα πρέπει να χωρίσουμε τις τιμές της μεταβλητής σε διαβαθμίσεις (κλάσεις).

Έτσι αν λχ σε μία έρευνα ο ερευνητής ενδιαφέρεται για την επιφάνεια των διαμερισμάτων των κατοίκων μίας συγκεκριμένης περιοχής, αντί να δημιουργήσει μία κατανομή με καθορισμένες τιμές εμβαδών, π.χ 60τ.μ, 80 τ.μ, 100τ.μ, κ.λ.π μπορεί να χωρίσει τις τιμές σε τάξεις (κλάσεις) σε μικρότερο από 60, από 61-80, από 81-100, από 100-120, από 121 και άνω και στην συνέχεια να θεωρήσει τις αντίστοιχες διαβαθμίσεις «Πολύ Μικρό», «Μικρό», «Μεσαίο», «Μεγάλο», «Πολύ μεγάλο» διαμέρισμα.



Σχήμα 1.7: Ομογενοποίηση μιας ποσοτικής μεταβλητής

Παράδειγμα

Υποθέτουμε ότι μελετάμε 500 επιχειρήσεις στις οποίες τέθηκε το ερώτημα αν ο ανταγωνισμός που υφίστανται είναι: αμελητέος, μέτριος ή έντονος. Τις επιχειρήσεις τις κατατάξαμε σε τρεις κατηγορίες σε μικρές, μεσαίες και μεγάλες. Ο συνδυασμός των δύο ποιοτικών μεταβλητών, ανταγωνισμός και κατηγορία επιχειρήσεων παρέχει την δυνατότητα δημιουργίας ενός πίνακα διπλής εισόδου, ο οποίος αποτελεί ένα πίνακα συμπτώσεων.

Πίνακας 1.6

Περιγραφή 500 επιχειρήσεων με βάση το μέγεθος που παρουσιάζουν και τον ανταγωνισμό που δέχονται

	Αμελητέος	**Μέτριος**	**Έντονος**	**Σύνολο**
Μικρή	15	30	15	**60**
Μεσαία	10	150	70	**230**
Μεγάλη	20	130	60	**210**
Σύνολο	**45**	**310**	**145**	**500**

όπου

I={Μέγεθος εταιρείας} με τρεις διαβαθμίσεις i_1=Μικρή, i_2=Μεσαία και i_3=Μεγάλη
J={Ανταγωνισμός}με τρεις διαβαθμίσεις j_1=Αμελητέος, j_2=Μέτριος και j_3=Έντονος

Γ) Πίνακες διατεταγμένων δεδομένων

Ένα σύνολο δεδομένων S λέγεται διατεταγμένο όταν ισχύει για τα στοιχεία του η σχέση

$$(n_1,n_2,...n_i,n_j,..,n_m)\neq(n_1,n_2,...n_j,n_i,..,n_m)$$

Δηλαδή η αντιμετάθεση δύο στοιχείων n_i,n_j του συνόλου S δημιουργεί δύο διαφορετικές ν-ιάδες.

Οι πίνακες διατεταγμένων δεδομένων διακρίνονται σε :

- πίνακες **κατάταξης στατιστικών μονάδων**
- πίνακες **κατάταξης μεταβλητών**.

i) Πίνακες κατάταξης στατιστικών μονάδων

Μερικές φορές, κατά τη μελέτη p ποσοτικών χαρακτηριστικών, συμβαίνει οι τιμές κάποιων μεταβλητών να παρουσιάζουν μεγάλη μεταβλητακότητα (CV>0.6) που οφείλονται σε μερικές πολύ υψηλές (ή και χαμηλές) τιμές, αλλοιώνοντας την γενική εικόνα του συνόλου των παρατηρήσεων.

Για να αμβλυθούν αυτές οι ανεπιθύμητες επιδράσεις, έχουμε τη δυνατότητα να αντικαταστήσουμε τις τιμές με τις αντίστοιχες **τάξεις μεγέθους** που παρουσιάζουν, μετά την κατάταξη τους σε φθίνουσα σειρά.

Παράδειγμα

Δίνονται οι τιμές πέντε μετοχών ως προς τρία κριτήρια. Στον πίνακα 1.8 παρουσιάζονται οι τιμές, ενώ στον πίνακα 1.8α οι κατατάξεις τους.

Πίνακας 1.8

Τιμές

	x_1	x_2	x_3
i1	30	4	-1,9
i2	9	0	-0,1
i3	10	13	0,8
i4	7	20	1,1
i5	200	120	0,0

Πίνακας 1.8α

Κατατάξεις

	x_1	x_2	x_3
i1	2	4	5
i2	4	5	4
i3	3	3	2
i4	5	5	1
i5	1	1	3

ii) Πίνακες κατάταξης μεταβλητών

Κατά τη μελέτη p ποσοτικών χαρακτηριστικών, ή p διαβαθμίσεων ενός ποιοτικού χαρακτηριστικού συχνά συμβαίνει να ζητείται η κατάταξή τους κατ' αύξουσα σειρά, δηλαδή ποια μεταβλητή (ή διαβάθμιση) τοποθετείται 1η στη κατάταξη, ποια 2η , ποια 3η κ.λ.π.

Κατ' αυτόν τον τρόπο δημιουργείται ο πίνακας κατάταξης των μεταβλητών ο οποίος χρήζει ιδιαίτερης επεξεργασίας. Για την ανάλυση τέτοιων πινάκων χρησιμοποιείται η μέθοδος της Ανάλυσης των Τάξεων (Δ. Καραπιστόλης 2004).

Παράδειγμα

Σε μία μελέτη που περιλαμβάνει 20 επιχειρήσεις πωλήσεων τίτλων CD-ROM τέθηκε το ερώτημα:

«Ιεραρχήστε κατά σειρά σπουδαιότητας τους παρακάτω παράγοντες που πιστεύετε ότι επηρεάζουν θετικά τις πωλήσεις τίτλων CD-ROM της εταιρείας.»

α) Επίπεδο τιμών,
β) Οργάνωση service
γ) Καλύτερη εξυπηρέτηση
δ) Ποιότητα προϊόντων
ε) Ευρεία σειρά προϊόντων
στ) Χρηματικές διευκολύνσεις
ζ) Σημαντικές εκπτώσεις

Σχηματίζεται ο παρακάτω πίνακας κατάταξης των επτά μεταβλητών.

Πίνακας 1.10

	ΤΙΜΗ	ΟΡΓΑΝ	ΕΞΥΠΗΡ	ΠΟΙΟΤ.	ΣΕΙΡΑ	ΔΙΕΥΚ	ΕΚΠΤ.
Ι01	1	5	6	7	2	3	4
Ι02	2	6	1	4	3	7	5
Ι03	3	5	4	1	2	7	6
Ι04	2	7	5	3	4	6	1
Ι05	1	5	3	2	4	6	7

Δ) Λογικοί πίνακες

Με τους λογικούς πίνακες περιγράφονται στατιστικές μονάδες οι οποίες χαρακτηρίζονται από ποιοτικές ή ποσοτικές μεταβλητές χωρισμένες σε διαβαθμίσεις.

Αν συμβολίσουμε με J_i (i=1,..,p) το σύνολο των μεταβλητών, με J_{ij} (j=1,..,m) το σύνολο των διαβαθμίσεων της μεταβλητής J_i και θέσουμε σε κάθε στατιστική μονάδα που εμφανίζει την διαβάθμιση J_{ij} τον αριθμό 1,ενώ όταν δεν την εμφανίζει το 0,δημιουργείται ένας ομογενοποιημένος πίνακας T(I,J)={I=1,..,n και J=1,..,k}, του οποίου κάθε γραμμή i αποτελείται από τόσα 0 και 1 όσες είναι οι διαβαθμίσεις όλων των μεταβλητών, το σύνολο των οποίων συμβολίζεται με k.

Ένας τέτοιος πίνακας ονομάζεται **λογικός πίνακας 0-1 ή διαζευκτικός πίνακας.**

Πίνακας 1.11

Γενική μορφή ενός λογικού πίνακα 0-1

	J_1				J_2					J_p			
	J_{11}	J_{12}	....	J_{1r}	J_{21}	J_{22}	....	J_{2s}	....	J_{p1}	J_{p2}	....	J_{pt}
I_1	1	0	....	0	0	1	....	0	....	0	0	....	1
I_2	0	1	....	0	1	0	....	0	....	1	0	....	0
....	....	....	....	....	....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....	....	....	....	....	....
I_n	0	0	....	1	0	1	....	0	....	0	1	....	0

Με την δημιουργία του λογικού πίνακα αναδεικνύονται οι ιδιότητες των μεταβλητών που χαρακτηρίζουν τα αντικείμενα (στατιστικές μονάδες) σε αντίθεση με τον αρχικό πίνακα όπου τα αντικείμενα χαρακτηρίζονται από τις τιμές των μεταβλητές.

Το άθροισμα των μονάδων πρέπει να είναι ίσο με τον αριθμό p που προσδιορίζει το πλήθος των μεταβλητών Jp, αφού κάθε στατιστική μονάδα δε μπορεί να πάρει την τιμή 1 παρά μόνο σε μία κλάση (ιδιότητα ή διαβάθμιση) της ίδιας μεταβλητής.

Συνεπώς για τα στοιχεία t_{ij} του πίνακα T(I,J) {όπου i=1,..n και j=1,..k} ισχύει η σχέση

$$\sum_{j=1}^{k} t_{ij} = p \quad \text{για κάθε } i = 1, 2, \ldots . n$$

Έτσι εφόσον το σύνολο των διαβαθμίσεων των μεταβλητών Jp είναι k τότε κάθε στατιστική μονάδα I μπορεί να θεωρηθεί ως διάνυσμα του χώρου των k διαστάσεων με συντεταγμένες 0 και 1.

Παράδειγμα

Εξετάζουμε 9 επιχειρήσεις ανάλογα με τα ποιοτικά χαρακτηριστικά Μέγεθος εταιρείας και Χρησιμοποιούμενη Τεχνολογία (πίνακας 1.12).

Η μεταβλητή J_1={μέγεθος της εταιρείας} εξετάζεται ως προς τρεις διαβαθμίσεις τις (J_{11}=Μικρό,J_{12}=Μεσαίο,J_{13}=Μεγάλο μέγεθος)
ενώ
η μεταβλητή J_2={Χρησιμοποιούμενη τεχνολογία} ως προς δύο διαβαθμίσεις τις (J_{21} =Παλαιά, J_{22}=Σύγχρονη).

Πίνακας 1.12

Κατανομή εννέα επιχειρήσεων ανάλογα με το μέγεθος και την χρησιμοποιούμενη τεχνολογία

	Μέγεθος Εταιρείας			Τεχνολογία	
	Μικρό	Μεσαίο	Μεγάλο	Παλαιά	Σύγχρονη
I1	1	0	0	1	0
I2	0	1	0	0	1
I3	1	0	0	1	0
I4	0	0	1	1	0
I5	0	1	0	0	1
I6	0	0	1	1	0
I7	1	0	0	0	1
I8	0	1	0	0	1
I9	1	0	0	1	0

Στην περίπτωση αυτή έχουμε δύο μεταβλητές με τρεις και δύο αντίστοιχα διαβαθμίσεις. Το σύνολο των διαβαθμίσεων είναι 5, άρα κάθε στατιστική μονάδα I_n περιγράφεται μ' ένα διάνυσμα που ανήκει στο χώρο των 5 διαστάσεων. Έτσι οι συντεταγμένες της στατιστικής μονάδας I_4 είναι I_4={0,0,1,1,0}.

Ε) Πίνακες Burt

Μία ειδική περίπτωση γενικευμένου πίνακα συμπτώσεων απολύτων συχνοτήτων αποτελεί ο πίνακας Burt.
Ο πίνακας Burt παράγεται από ένα διαζευκτικό πίνακα Χ(nxp) ο οποίος διασταυρώνει τις κλάσεις κάθε μεταβλητής με το σύνολο των κλάσεων των μεταβλητών του πίνακα, χρησιμοποιώντας τη παρακάτω διανυσματική εξίσωση

$$\underset{(pxp)}{B} = \underset{(pxn)}{X'} \cdot \underset{(nxp)}{X} \qquad (1.2)$$

όπου Χ' ο ανάστροφος πίνακας του Χ.

Η σχέση 1.2 παράγει ένα συμμετρικό τετραγωνικό πίνακα διπλής εισόδου με τόσες γραμμές και στήλες, όσες το άθροισμα k των κλάσεων των μεταβλητών του διαζευκτικού πίνακα.
Ο πίνακας Burt αποτελεί "μωσαϊκό" απλών πινάκων συμπτώσεων, το πλήθος των οποίων ανέρχεται σε p^2.

ΠΑΡΑΔΕΙΓΜΑ

(δεδομένα στη διαφάνεια 17)

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 4 & 0 & 0 & 3 & 1 \\ 0 & 3 & 0 & 0 & 3 \\ 0 & 0 & 2 & 2 & 0 \\ 3 & 0 & 2 & 5 & 0 \\ 1 & 3 & 0 & 0 & 4 \end{pmatrix}$$

Ο πίνακας 1.13 αποτυπώνει τον πίνακα συμπτώσεων που δημιουργήθηκε.

Πίνακας 1.13

Ο γενικευμένος πίνακας Burt

	Μικρό μεγεθος	**Μεσαίο μεγεθος**	**Μεγάλο μεγεθος**	**Παλαιά τεχνολογία**	**Σύγχρονη τεχνολογία**
Μικρό μεγεθος	4	0	0	3	1
Μεσαίο μεγεθος	0	3	0	0	3
Μεγάλο μεγεθος	0	0	2	2	0
Παλαιά τεχνολογία	3	0	2	5	0
Σύγχρονη τεχνολογία	1	3	0	0	4

Πίνακας 1.14

Πίνακας Burt που διασταυρώνει 5 μεταβλητές με τις διαβαθμίσεις τους που περιγράφουν 567 στατιστικές μονάδες

	X1 X_{11} X_{12}	**X2** X_{21} X_{22} X_{23}	**X3**	**X4** X_{41} X_{42} X_{43} X_{44} X_{45} X_{46}	**X5** .....
X11 **X12** **X13** **X14**					
X21 **X22** **X23**		386 0 0 0 67 0 0 0 114		97 73 70 67 21 58 2 9 12 18 20 6 5 20 25 0 44 20	
X31 **X32**					
X41 **X42** **X43** **X44** **X45** **X46**		97 2 5 73 9 20 70 12 25 67 18 0 21 20 44 58 6 20		104 0 0 0 0 0 0 102 0 0 0 0 0 0 107 0 0 0 0 0 0 85 0 0 0 0 0 0 85 0 0 0 0 0 0 84	
X51 **X52**					

Ιδιότητα α) 114=5+20+25+0+44+20 Ιδιότητα β) Ιδιότητα γ) 386+67+114=567

Ιδιότητα δ) 97+2+5=104

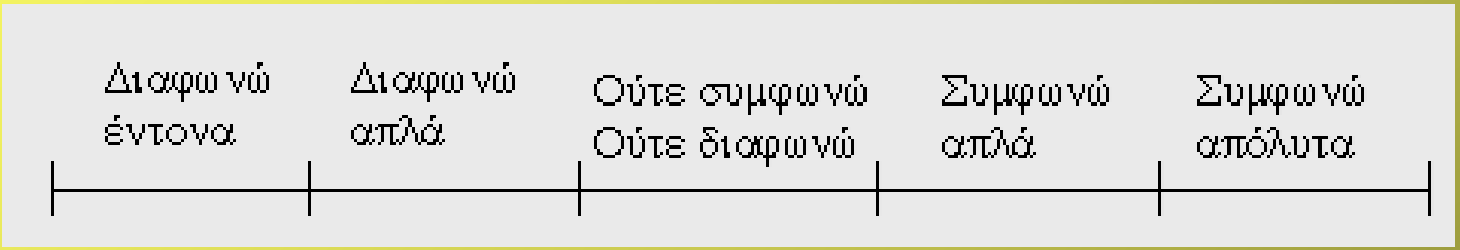
Ιδιότητα ε)

Ιδιότητα X_2 **Διαβαθμίσεις**	**Ιδιότητα X_4** **Σχετική συχνότητα ως προς X_{41}**
X_{21}	97/104=0,933
X_{22}	2/104=0,019
X_{23}	5/104=0,048

Ιδιότητα στ) 58/386 ≠ 58/84

ΣΤ) Πίνακες αξιολόγησης με την χρήση διαβαθμισμένης κλίμακας

Ένας πίνακας αξιολόγησης περιλαμβάνει απαντήσεις ερωτηματολογίων όπου τέθηκαν ερωτήσεις στις οποίες τα άτομα καλούνται να δηλώσουν τον βαθμό συμφωνίας τους, δίνοντας ένα βαθμό αποδοχής ή απόρριψης, για μια σειρά απόψεων στη βάση μιας αριθμητικής κλίμακας, η οποία καλείται κλίμακα Likert.



Παράδειγμα :Ερωτώνται κατά πόσο είναι ικανοποιημένοι οι πελάτες ενός εστιατορίου από τις παρακάτω υπηρεσίες που προσφέρει. Έστω μία απάντηση

ΕΝΔ	Καθόλου	Λίγο	Μέτρια	Καλή	Πολύ καλή
Καθαριότητα			✓		
Εξυπηρέτηση		✓			
Ποικιλία				✓	

Ζ) Δυαδικοί πίνακες δεδομένων

Έστω ότι έχουμε p ερωτήματα, όπου οι απαντήσεις μπορεί να είναι θετικές ή αρνητικές. Τότε εφόσον κωδικοποιήσουμε την θετική απάντηση με 1 και την αρνητική με 0, αυτό έχει ως αποτέλεσμα να υπάρχουν 2^p πιθανές απαντήσεις. Έτσι στην περίπτωση όπου έχουμε p=3 τότε οι $2^3=8$ απαντήσεις είναι οι εξής:

000,100,010,001,110,101,011,111

Τότε στην περίπτωση όπου έχουμε n=30 ερωτώμενους, ο δυαδικός πίνακας δεδομένων θα έχει την παρακάτω μορφή :

Πίνακας 1.17

Απαντήσεις	n_i
000	2
100	3
010	5
001	4
110	6
101	3
011	2
111	5
Σύνολο	30

Η) Πίνακες από ανοικτές ερωτήσεις

Έστω η ερώτηση σε μία έρευνα: Με ποια προβλήματα κατά την άποψή σας δεν ασχολείται όσο θα έπρεπε το κόμμα της προτίμησή σας; (Γράψτε όσα προβλήματα επιθυμείτε, ενώ ελήφθησαν οι απαντήσεις μόνο των ψηφοφόρων των τριών μεγάλων κομμάτων ΠΑΣΟΚ,ΝΔ και ΚΚΕ)

Μετά την συγκέντρωση των ερωτηματολογίων σε μία βάση δεδομένων, ακολούθησε η κατηγοριοποίηση των απαντήσεων, όπως ενδεικτικά παρουσιάζεται στον παρακάτω πίνακα.

Πίνακας 1.18: Διαβαθμίσεις και αντίστοιχοι λεξικοί τύποι ή προτάσεις της ερώτησης που αφορούσε προβλήματα όπως εκτέθηκαν από τους ερωτώμενους

Κοινωνικά	Κοινωνικές διακρίσεις, ναρκωτικά, κοινωνική πρόνοια, κοινωνική ανισότητα, αλκοολισμός, aids, άστεγοι, σεισμοπαθείς, ρατσισμός, άνθρωποι με ειδικές ανάγκες.
Οικονομικά	Εθνική οικονομία, προβλήματα μικρομεσαίων επιχειρήσεων, βιομηχανική ανάπτυξη, επενδύσεις, καπιταλισμός, πληθωρισμός, ΣΔΟΕ, χρηματιστήριο, οικογενειακός προϋπολογισμός, ευρώ,ΟΝΕ, καταναλωτική κοινωνία.
Παιδείας	Εκπαίδευση, παιδεία
Μεταναστών	Λαθρομετανάστες, αλλοδαποί, διακίνηση ξένων, μετανάστευση, πρόσφυγες, παράνομες ελληνοποιήσεις, άνοιγμα συνόρων
Αγροτικά	Γεωργία, αγροτική ανάπτυξη, αγροτική πολιτική, επιδότηση αγροτικών προϊόντων.

Πίνακας 1.19: Πίνακας συχνοτήτων των διαβαθμίσεων των προβλημάτων με τα οποία δεν ασχολούνται επαρκώς το ΠΑΣΟΚ, η ΝΔ και το ΚΚΕ με βάση τις απαντήσεις των ψηφοφόρων τους

	ΠΡΟΒΛΗΜΑ	ΠΑΣΟΚ		ΝΔ		ΚΚΕ	
		n_i	%	n_i	%	n_i	%
1	ΑΓΡΟΤΙΚΟ	99	21,81	62	11,69	3	2,59
3	ΑΝΑΠΤΥΞΗΣ	20	4,4	17	3,20	8	6,89
4	ΑΝΕΡΓΙΑΣ	115	25,33	89	16,79	18	15,52
5	ΑΣΦΑΛΙΣΤΙΚΟ	68	14,97	59	11,53	8	6,89
10	ΕΞΩΤΕΡΙΚΗΣ ΠΟΛΙΤΙΚΗΣ	16	3,52	70	13,2	23	19,82
13	ΚΟΙΝΩΝΙΚΑ	75	16,51	51	9,62	12	10,34
15	ΟΙΚΟΝΟΜΙΑΣ	28	6,16	46	8,68	25	21,55
16	ΠΑΙΔΕΙΑΣ	125	27,53	123	23,2	20	17,24
19	ΜΕΤΑΝΑΣΤΩΝ	12	2,64	20	3,77	3	2,59
20	ΥΓΕΙΑΣ	123	27,09	92	17,35	21	18,10
21	ΑΛΛΟ ΠΡΟΒΛΗΜΑ	10	2,2	7	1,32	6	5,17

Θ) Τριδιάστατοι πίνακες (*)

Ο παρακάτω πίνακας είναι τριδιάστατος διότι εξετάζει ταυτόχρονα τρία χαρακτηριστικά
α) τις εμπορικές συναλλαγές της Ελλάδος (Εισαγωγές-Εξαγωγές) σε εκατ. ευρώ
β) τις περιοχές όπου πραγματοποιούνται οι συναλλαγές αυτές
γ) τη χρονική περίοδο εντός της οποίας έχουμε τα συγκεκριμένα αποτελέσματα.

Τα κράτη του πλανήτη χωρίστηκαν σε οκτώ περιοχές. Την Ευρωπαϊκή ένωση των 15 (ΕΕ-15), την Β.Αμερική (Β_ΑΜ), τις Υπόλοιπες χώρες του ΟΟΣΑ (Υ_ΟΟS), την Κ&Α.Ευρώπη (ΚΑ_ΕU),την Β. Αφρική & Μ. Ανατολή (ΒΑ_ΜΑ),την ΝΑ Ασία (ΝΑΑSΙ), την Λατινική Αμερική (LAT_A) και τις Λοιπές χώρες (LOIPE).

Πίνακας 1.20: Εμπορικές συναλλαγές της Ελλάδος σε εκατ. Ευρώ

	Εισαγωγές σε εκατ. €			Εξαγωγές σε εκατ. €		
Περιοχές	2002	2003	2004	2002	2003	2004
ΕΕ-15	14560	16028	17462	4708	4319	4281
Β. Αμερική	920	1276	1635	730	624	516
Υπόλοιπες ΟΟΣΑ	2559	2392	2196	524	558	602
Κ & Α. Ευρώπη	3592	3454	3319	1980	1998	2008
Β.Αφρική & Μ. Ανατολή	1790	2237	2765	425	483	560
ΝΑ Ασία	2240	2037	1819	165	154	141
Λατινική Αμερική	158	222	291	41	37	32
Λοιπές χώρες	1282	1600	1913	360	310	264
ΣΥΝΟΛΟ	27166	29325	31562	9389	8483	8404

*Πολλοί χρησιμοποιούν τον όρο τρισδιάστατο, λανθασμένα βέβαια, αφού σκοπός του συγκεκριμένου όρου είναι να πληροφορεί ότι υπάρχουν τρεις διαφορετικές διαστάσεις και όχι τρεις φορές η ίδια διάσταση που αποδίδεται με τον όρο τρις. Κατά τη λανθασμένη αντίληψη, θα έπρεπε να λέγαμε τρίσποδο κάθισμα αντί τρίποδο, δίσκυκλο ποδήλατο αντί δίκυκλο, δίστομο έργο αντί δίτομο, ενώ σωστά χαρακτηρίζεται κάποιος τρισευτυχισμένος, επειδή θεωρείται πολλαπλά ευτυχισμένος. Ορθοί όροι είναι και οι όροι δισκελής, τρισυπόστατος κ.λπ. Βέβαια δεν πρέπει να παρασύρεται κάποιος από τους όρους δισέλιδο, δίστηλο, τρίστηλο, επειδή το γράμμα σ είναι αρχικό των απλών λέξεων σελίδα, στήλη.

1.5 Έκτροπες τιμές-Ύποπτες ακραίες τιμές (Outliers data)

Κατά τις επαναλαμβανόμενες μετρήσεις ενός μεγέθους, κάποιες από τις αριθμητικές τιμές μπορεί να φαίνεται ότι απέχουν πολύ από την κύρια μάζα μετρήσεων. Στην περίπτωση αυτή υπάρχει μια φυσική προδιάθεση να "απορρίψουμε" και να μην συμπεριλάβουμε τις "ύποπτες" αυτές τιμές στους υπολογισμούς που θα ακολουθήσουν (π.χ. για τον υπολογισμό της μέσης τιμής και της τυπικής απόκλισης των μετρήσεων). Αυτό επιτρέπεται να γίνει μόνο αν αυτές οι τιμές χαρακτηρισθούν ως **έκτροπες** (outlier).

1.5.1 Δοκιμασία Q (Q-test) του Dixon

Για δείγματα μεγέθους 3 έως 12 πειραματικών τιμών

1.5.2 Δοκιμασία των Iglewicz και Hoaglin

Με το τεστ αυτό υπολογίζεται η παρακάτω ποσότητα

$$M_i = \frac{0,675 \times (x_i - \bar{x})}{MAD}$$

Όπου το MAD δηλώνει τη διάμεση απόλυτη απόκλιση

Για κάθε απόλυτη τιμή M_i = {1,…v} εφόσον προκύψει μεγαλύτερη από 3,5 επισημαίνεται ότι η i τιμή είναι ύποπτη ακραία τιμή.

Παράδειγμα

Έστω οι παρακάτω τιμές 4,5 4,9 5,6 4,2 6,2 5,2 9,9

Στη συνέχεια τοποθετούνται κατ' αύξουσα σειρά 4,2 4,5 4,9 **5,2** 5,6 6,2 9,9

Η διάμεση τιμή M=είναι ίση με 5,2

Στη συνέχεια παίρνουμε τις απόλυτες διαφορές των τιμών από τη διάμεσο τιμή M

|4,2-5,2|=1,0 |4,5-5,2|=0,7 |4,9-5,2|=0,3 |5,2-5,2|=0,0 |5,6-5,2|=0,4 |6,2-5,2|=1,0

|9,9-5,2|=4,7

και τις τακτοποιούμε κατ' αύξουσα σειρά. Ήτοι

0,0 0,3 0,4 **0.7** 1,0 1,0 4,7

Στη συνέχεια βρίσκουμε τη διάμεσο τιμή αυτής της σειράς που είναι

MAD=0,7

Η μέση τιμή της αρχικής σειράς είναι είναι ίση με $\bar{x} = 5,786$

Εφαρμόζουμε τον τύπο (1.5) για κάθε μία από τις επτά τιμές Οπότε

$$M_1 = \left| \frac{0,675 \cdot (4,2 - 5,786)}{0,7} \right| = 1,699$$

.....

.....

$$M_7 = \left| \frac{0,675 \cdot (9,9 - 5,786)}{0,7} \right| = 3,967$$

Συνεπώς η τιμή M_7=9,9 δυνητικά είναι ακραία τιμή.

1.6 Διαγωνοποίηση ενός τετραγωνικού πίνακα

Ο J.P Benzecri στην αρχή του μαθήματος περί παραγοντικής ανάλυσης ανέφερε τα εξής: **«Στη καρδιά κάθε ανάλυσης δεδομένων υπάρχει διαγωνοποίηση ενός συμμετρικού τετραγωνικού πίνακα».**

Ένας τετραγωνικός πίνακας Μ είναι ένας πίνακας όπου το πλήθος των γραμμών είναι ίσο με το πλήθος των στηλών του. Γράφεται Μ(n,n).

Για παράδειγμα ο παρακάτω πίνακας:

$$M(3,3) = \begin{pmatrix} 3 & -1 & 4 \\ 5 & 3 & 0 \\ 2 & 1 & 2 \end{pmatrix}$$

Ένας τετραγωνικός πίνακας είναι **συμμετρικός** ως προς την κύρια διαγώνιό του όταν δύο συμμετρικά στοιχεία ως προς την κύρια διαγώνιο είναι ίσα. Ήτοι:

$$M(3,3) = \begin{pmatrix} 30 & 40 & 50 \\ 40 & 54 & 68 \\ 50 & 68 & 86 \end{pmatrix}$$

Σημείωση 1: Ορίζεται ως **ίχνος** ενός πίνακα το άθροισμα των στοιχείων της κύριας διαγωνίου.

Με τα στοιχεία του προηγούμενου παραδείγματος έχουμε: 30+54+86=170

ΠΑΡΑΤΗΡΗΣΕΙΣ

Έστω ένας πίνακας Μ(m,n) και ο ανάστροφός του Μ'(n,m). Αν πάρουμε το γινόμενο Μ'(n,m)·Μ(m,n) τότε θα προκύψει ένας πίνακας Χ(n,n) ο οποίος είναι συμμετρικός και τετραγωνικός.

Έστω λοιπόν ο πίνακας

$$M(4,3)=\begin{pmatrix}1&2&3\\2&3&4\\3&4&5\\4&5&6\end{pmatrix}$$ και ο ανάστροφός του $$M'(3,4)=\begin{pmatrix}1&2&3&4\\2&3&4&5\\3&4&5&6\end{pmatrix}$$

τότε

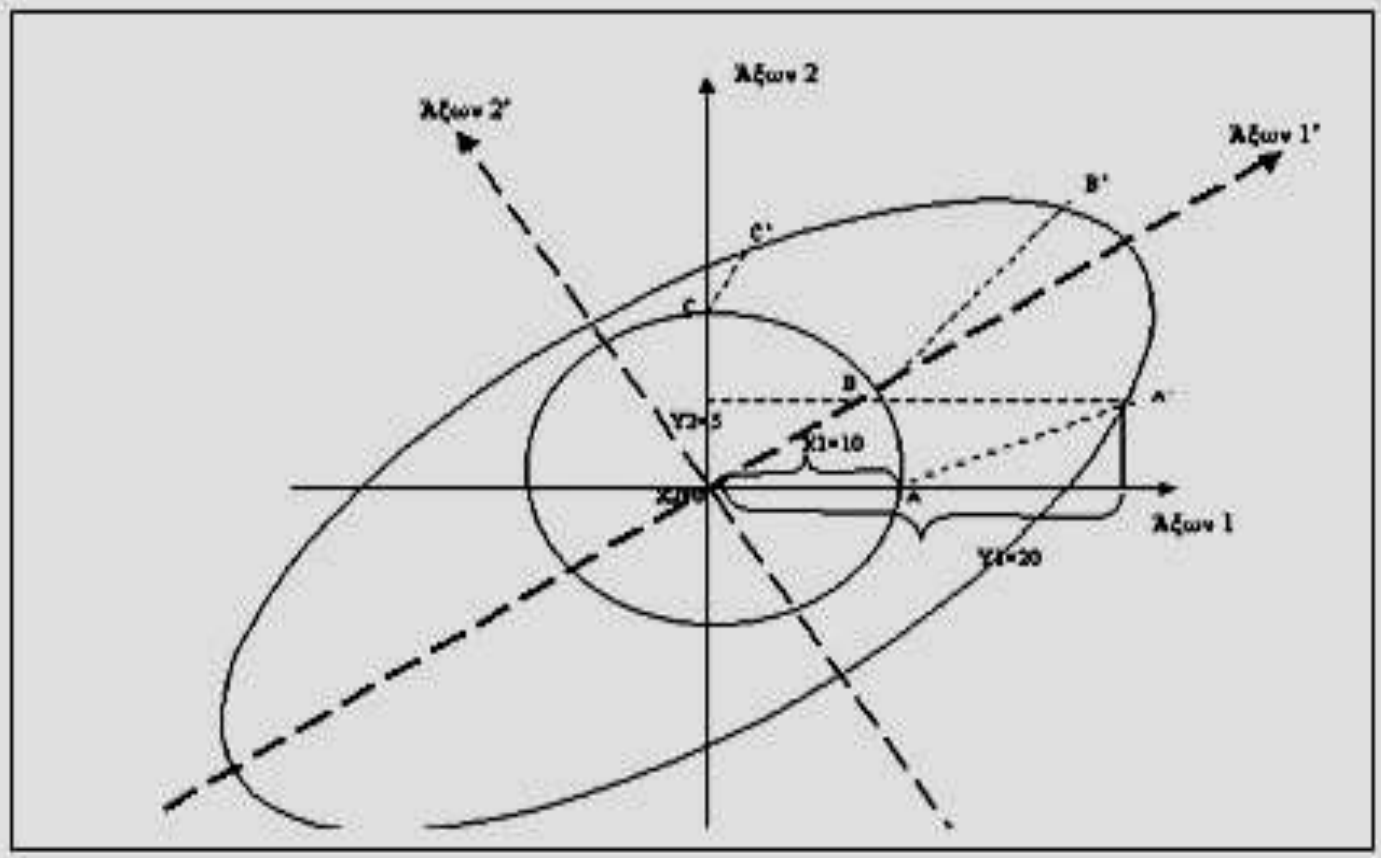
$$\begin{pmatrix}1&2&3&4\\2&3&4&5\\3&4&5&6\end{pmatrix}\times\begin{pmatrix}1&2&3\\2&3&4\\3&4&5\\4&5&6\end{pmatrix}=\begin{pmatrix}30&40&50\\40&54&68\\50&68&86\end{pmatrix}$$

$$M'(3,4)\times M(4,3)=X(3,3)$$

Παρατήρηση 1: Καθώς τα στοιχεία της διακύμανσης είναι πάντοτε ένα άθροισμα τετραγώνων, δεν είναι καθόλου παράξενο ότι τα στοιχεία της κύριας διαγωνίου ενός συμμετρικού τετραγωνικού πίνακα ονομάζονται **στοιχεία διακύμανσης**.

Παρατήρηση 2: Σχετικά με τους δύο πίνακες Μ'. Μ και Μ. Μ' τους ονομάζουμε είτε **πίνακα αδράνειας** είτε **πίνακα διακύμανσης- συνδιακύμανσης**, παρότι είναι διαφορετικοί μεταξύ τους αποκαλύπτουν την ίδια πραγματικότητα, δηλαδή ένα **ελλειψοειδές αδράνειας**.

Παρατήρηση 3: Η διαγωνοποίηση ενός τετραγωνικού πίνακα είναι η εργασία εύρεσης των αξόνων συμμετρίας αυτού του ελλειψοειδούς.



Σχήμα 1.9: *Οι άξονες 1 και 2 είναι οι άξονες του κύκλου ενώ οι άξονες 1' και 2' είναι οι άξονες της έλλειψης*

Παράδειγμα

Έστω ο τετραγωνικός συμμετρικός πίνακας Μ(2,2):

$$M(2,2) = \begin{pmatrix} 2 & 0,5 \\ 0,5 & 1,5 \end{pmatrix}$$

Αν τώρα περιστρέψουμε τους άξονες συντεταγμένων 1 και 2 του κύκλου ώστε να συμπέσουν με τους άξονες συμμετρίας 1’ και 2’ της έλλειψης, τότε είναι φανερό ότι μετατράπηκε στο νέο αυτό σύστημα συντεταγμένων ο κύκλος σε έλλειψη.

Με την μετατροπή αυτή γενικώς οι νέες συντεταγμένες y_1 και y_2 βρίσκονται βάσει των σχέσεων

$$y_1=\lambda_1 \cdot x_1+0 \cdot x_2$$
$$y_2=0 \cdot x_2+\lambda_2 \cdot x_2$$

Ο πίνακας Μ στο νέο σύστημα συντεταγμένων έχει την παρακάτω μορφή:

$$M(2,2) = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

Τα στοιχεία του πίνακα Μ(2,2) είναι μηδέν εκτός εκείνων που βρίσκονται στην κύρια διαγώνιο. Λέμε τότε ότι ο πίνακας Μ **διαγωνοποιήθηκε**.

Τα στοιχεία λ_1 και λ_2 της κυρίας διαγωνίου ονομάζονται **χαρακτηριστικές ρίζες** του πίνακα Μ.

Σημειωτέον η διαγωνοποίηση ενός πίνακα δεν αλλάζει την τιμή του ίχνους του, οπότε στο παράδειγμά μας έχουμε

$$\lambda_1+\lambda_2=2+1,5=3,5$$

Με άλλα λόγια η διαγωνοποίηση δεν αλλάζει την διακύμανση, απλώς απαλείφει τις συνδιακυμάνσεις.

1.6.2 Δυϊκός χώρος

Έστω ένας πίνακας δεδομένων M(n,3). Σχηματίζουμε τον πίνακα X(3,3)=M'(3,n).M(n,3). Με την διαγωνοποίηση του συμμετρικού τετραγωνικού πίνακα X(3,3) βρίσκουμε τρεις χαρακτηριστικές ρίζες $\lambda_1,\lambda_2,\lambda_3$.Αν τώρα σχηματίσουμε τον πίνακα W(n,n)=M(n,3).M'(3,n) και τον διαγωνοποιήσουμε θα βρούμε και σ' αυτή την περίπτωση τις ίδιες χαρακτηριστικές ρίζες $\lambda_1,\lambda_2,\lambda_3$ ενώ οι υπόλοιπες n-3 θα είναι όλες ίσες με μηδέν.

Οι πίνακες X και W στο νέο σύστημα συντεταγμένων είναι της μορφής:

$$X=\begin{pmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{pmatrix} \qquad W=\begin{pmatrix}\lambda_1 & 0 & 0 & 0 & . & 0\\ 0 & \lambda_2 & 0 & 0 & . & 0\\ 0 & 0 & \lambda_3 & 0 & . & 0\\ 0 & 0 & 0 & 0 & . & 0\\ . & . & . & . & . & .\\ 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}$$

Ο χώρος των n διαστάσεων στον οποίο μεταφερόμαστε με τον πίνακα M'(3,n) ονομάζεται **δυϊκός χώρος** του χώρου των 3 διαστάσεων

Οι πίνακες M(n,3) και M'(3,n) προφανώς είναι τελείως διαφορετικοί αφού ο μεν πρώτος απεικονίζει n σημεία στο χώρο των τριών διαστάσεων ενώ ο δεύτερος 3 σημεία στον χώρο των n διαστάσεων, αλλά όταν διαγωνοποιηθούν παρουσιάζουν τις ίδιες χαρακτηριστικές ρίζες, συνεπώς τους ίδιους άξονες συμμετρίας των ελλειψοειδών αδράνειας.

Η αξιόλογη αυτή ιδιότητα των συνόλων M και M' επιτρέπει να θέτουμε τα σημεία των δύο χώρων στο ίδιο σύστημα συντεταγμένων, αφού διατηρούνται μόνο οι τρεις άξονες του δυϊκού χώρου ως τους μόνους χρήσιμους.

Το στοιχείο αυτό επιτρέπει στα σημεία των δύο χώρων R^3 και R^n να απεικονίζονται ταυτόχρονα στο ίδιο διάγραμμα.

Μια πρώτη προσέγγιση των εννοιών της παραγοντικής ανάλυσης

Ως γνωστόν από τις πρώτες γνώσεις που αποκτά κανείς στα Μαθηματικά, είναι η ανάλυση της παράστασης $\alpha^2 - \beta^2$ σε γινόμενο πρώτων παραγόντων.
Ήτοι

$$\alpha^2 - \beta^2 = (\alpha+\beta).(\alpha-\beta)$$

Παράδειγμα

Δίνεται ένας πίνακας συμπτώσεων ο οποίος παρουσιάζει τη κατανομή 100 θεατών τεσσάρων ταινιών I_1, I_2, I_3, I_4 οι οποίες αξιολογήθηκαν σύμφωνα με την παρακάτω διαβαθμισμένη κλίμακα: "Μέτρια» (Μ)", "Καλή (Κ)", "Άριστη (Α)".

Πίνακας 1.22

Αξιολόγηση τεσσάρων ταινιών από 100 θεατές

	Μ	**Κ**	**Α**	**ΠΓ**
I_1	13	2	5	**20**
I_2	20	2	8	**30**
I_3	10	5	5	**20**
I_4	7	1	22	**30**
ΠΣ	**50**	**10**	**40**	**100**

Πίνακας 1.22

	M	K	A		M	K	A		M	K	A
I_1	13	2	5	-	10	2	8	=	3	0	-3
I_2	20	2	8	-	15	3	12	=	5	-1	-4
I_3	10	5	5	-	10	2	8	=	0	3	-3
I_4	7	1	22	-	15	3	12	=	-8	-2	10

T	**-**	**T_o**	**=**	**R_1**

Έτσι

Εντοπίζουμε πως η διαβάθμιση Μέτριο θέαμα χαρακτηρίζει θετικά τις ταινίες I_1 και I_2 (εντονότερα την I_2),ενώ η διαβάθμιση Καλό θέαμα την I_3 και το Άριστο θέαμα την ταινία I_4.

Γενικά η απόκλιση από την **κατάσταση ισορροπίας** ενός συστήματος (από άποψη Θερμοδυναμικής) μας παρέχει την πληροφορία κατά πόσο είναι απομακρυσμένο το σύστημα από την κατάσταση **μέγιστης εντροπίας**, ενώ από άποψη Μηχανικής, μας επιτρέπει να εντοπίσουμε αν υπάρχει έλξη, ισορροπία (ανεξαρτησία) ή άπωση μεταξύ μιας " γραμμής " και μιας " στήλης« του πίνακα δεδομένων.

Ο πίνακας R_1 διασπάται σε δύο πίνακες T_1 και T_2 οι οποίοι αναπαράγονται εφόσον πολλαπλασιάσουμε ένα διάνυσμα στήλη με ένα διάνυσμα γραμμή για τον καθένα, διαδικασία γνωστή ως **διαγωνοποίηση**. Τα διανύσματα αυτά είναι περιθωριακές στήλες και γραμμές των πινάκων και καλούνται **χαρακτηριστικά διανύσματα**

$$R_1 = \begin{bmatrix} -1 \\ -1 \\ -2 \\ 4 \end{bmatrix} \cdot \begin{pmatrix} -1 & -1 & 2 \end{pmatrix} + \begin{bmatrix} -1 \\ -2 \\ 1 \\ 2 \end{bmatrix} \cdot \begin{pmatrix} -2 & 1 & 1 \end{pmatrix}$$

$$R_1 = \quad T_1 \quad + \quad T_2$$

Οπότε

με την διαγωγοποίηση προκύπτει ότι η αρχική τοποθέτηση η οποία προέβλεπε την ανάλυση του πίνακα Τ σε άθροισμα απλούστερων πινάκων πραγματοποιήθηκε.

Ήτοι:

$$T = T_0 + T_1 + T_2$$

Παραγοντικοί άξονες. Παραγοντικό επίπεδο

Είναι σαφές πως η κατάλληλη πληροφορία δεν προκύπτει από τον πίνακα ανεξαρτησίας T_0, αφού όπως προαναφέραμε το ενδιαφέρον μας επικεντρώνεται κυρίως στις αποκλίσεις από την κατάσταση ισορροπίας και στους λόγους που επιβάλλουν τις αποκλίσεις αυτές.

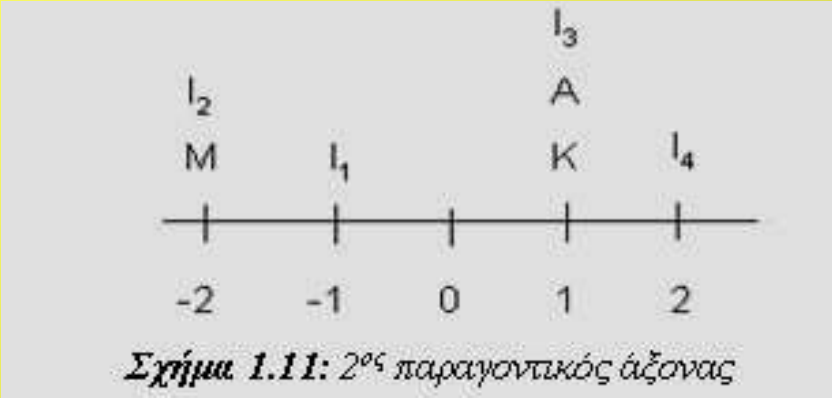
Κατασκευάζουμε στο καρτεσιανό επίπεδο ένα σύστημα ορθογωνίων συντεταγμένων, όπου ο καθένας άξονας συνδέεται με τα χαρακτηριστικά διανύσματα των πινάκων T_1 και T_2, ο οποίος καλείται **παραγοντικός άξονας**. Αυτό σημαίνει πως οι συντεταγμένες κάθε ταινίας και των κλάσεων των προτιμήσεων των θεατών που προκύπτουν με βάση κάθε χαρακτηριστικό διάνυσμα, τοποθετούνται στους αντίστοιχους άξονες, δημιουργώντας αυτό που ονομάζουμε **παραγοντικό επίπεδο**.

Στον 1ο άξονα οι νέες συντεταγμένες των ταινιών και των διαβαθμίσεων είναι αντιστοίχως:

I_1=-1 I_2=-1 I_3=-2 I_4=4 M=-1 K=-1 A=2

Σχήμα 1.10: 1ος παραγοντικός άξονας

Όσον αφορά στον 2ο παραγοντικό άξονα έχουμε



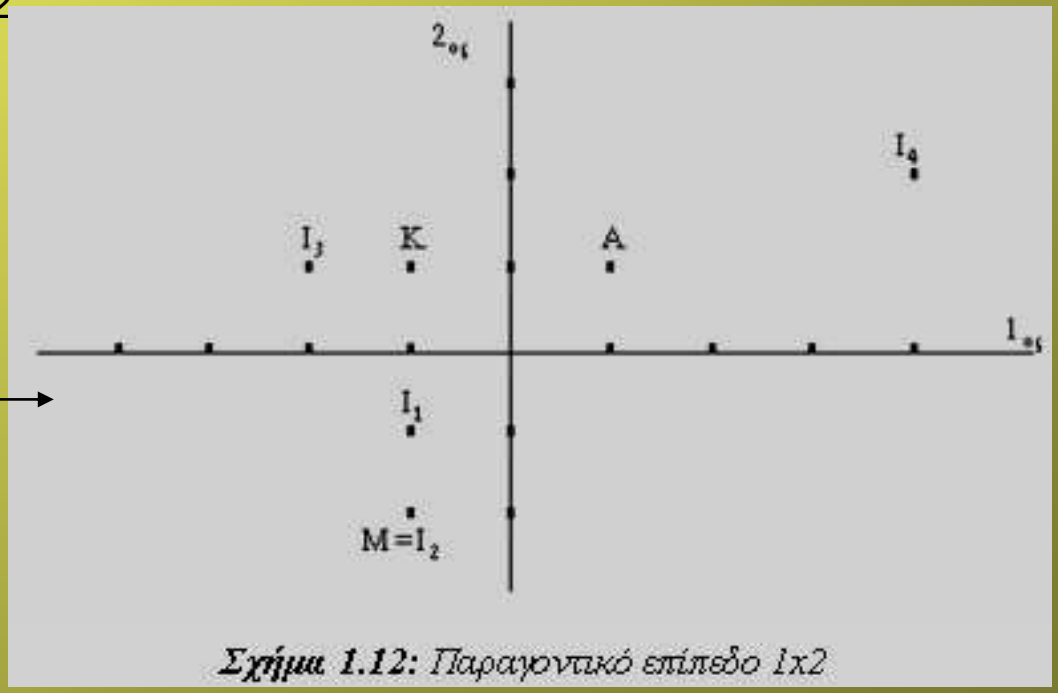
Σχήμα 1.11: 2ος παραγοντικός άξονας

Όσον αφορά στο παραγοντικό επίπεδο 1x2

Πίνακας 1.24

Συντεταγμένες των ταινιών και των διαβαθμίσεων στους δύο πρώτους παραγοντικούς άξονες

	1ος άξονας	2ος άξονας
I_1	-1	-1
I_2	-1	-2
I_3	-2	1
I_4	4	2
M	-1	-2
K	-1	1
A	2	1



Σχήμα 1.12: Παραγοντικό επίπεδο 1x2

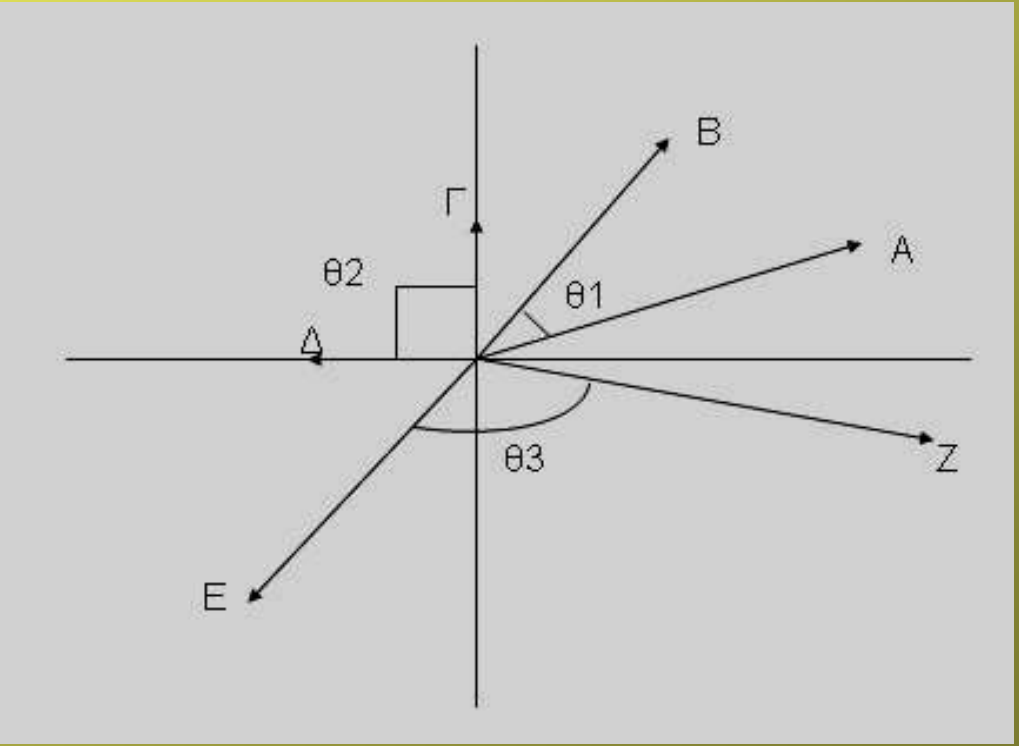
ΘΕΣΕΙΣ ΤΩΝ ΣΗΜΕΙΩΝ ΣΤΟ ΠΑΡΑΓΟΤΙΚΟ ΕΠΙΠΕΔΟ

Οι δυνατές **"θέσεις"** που μπορεί να παρουσιάσουν στο παραγοντικό επίπεδο τα διανύσματα τα οποία αντιπροσωπεύουν είτε στατιστικές μονάδες (γραμμές) είτε μεταβλητές (στήλες), όπως προαναφέραμε, είναι τρεις.

ι) η θέση της συζυγίας (0<θ1<90)

ιι) η θέση της καθετότητας (θ2=90)

ιιι) η θέση της αντίθεσης (90<θ3<180)



ΠΛΗΘΟΣ ΔΙΑΣΠΟΜΕΝΩΝ ΠΙΝΑΚΩΝ

ΕΡΩΤΗΜΑ:

«Ποιος είναι ο ελάχιστος αριθμός πινάκων στους οποίους μπορεί να διασπαστεί ένας πίνακας δεδομένων Τ διαστάσεων (nxp), δίχως να υπάρχει απώλεια της πληροφορίας που παρέχει; »

ΑΠΑΝΤΗΣΗ:

«Ο ελάχιστος αριθμός διασπομένων πινάκων που απαιτείται για να ανασυσταθεί ένας πίνακας Τ είναι ίσος τουλάχιστον με την μικρότερη διάστασή του και οπωσδήποτε ίσος με τη τάξη του πίνακα»

Τι συμβαίνει όμως όταν ο πίνακας έχει διάσταση μεγαλύτερη του 3 και πρέπει ο πίνακας R_1 να διασπαστεί σε περισσότερους από δύο πίνακες;

Στην περίπτωση αυτή διατηρούμε όσους πίνακες παρουσιάζουν μία απώλεια προσέγγισης (ανασύστασης) του αρχικού πίνακα, η οποία να θεωρείται ικανοποιητική.

Η ΕΝΝΟΙΑ ΤΗΣ ΔΙΑΣΤΑΣΗΣ ΕΝΟΣ ΧΩΡΟΥ

Όταν λέμε ότι η **διάσταση του χώρου** στον οποίο ζούμε είναι **τριών** διαστάσεων, ενώ η διάσταση του επιπέδου είναι **δύο** και της ευθείας είναι **μία**, μοιάζει να αναφερόμαστε σε διαστάσεις προσωποποιημένες όπου κάθε μια έχει την δική της σημασία.

Παραδείγματος χάριν ένα ορθογώνιο παραληλλεπίπεδο έχει τρεις διαστάσεις, το μήκος, το πλάτος και το ύψος, ενώ ένα ευθύγραμμα τμήμα είναι απολύτως μετρήσιμο από το μήκος του.

Στην **Αναλυτική Γεωμετρία**, αντίθετα με ότι είχε ορίσει ο **Ευκλείδης**, οι διαστάσεις έχουν ως κύριο ρόλο να προσδιορίζουν την θέση ενός σημείου μέσα σε οποιοδήποτε χώρο.

Παραδείγματος χάριν στο επίπεδο αντιστοιχούν δύο άξονες, ο άξονας ΟΧ και ο άξονας ΟΥ (ορθογώνιοι ή πλάγιοι), οι οποίοι επιτρέπουν να συνδέσουμε κάθε σημείο Μ του επιπέδου μ' ένα ζεύγος τιμών (x,y) που ονομάζονται **συντεταγμένες**.

Με βάση αυτή την αντίληψη η έννοια της **διάστασης** ορίζεται ως η δυνατότητα να προσδιοριστούν χωρίς αμφιβολία τα σημεία ενός χώρου R^n από ένα πλήθος n συντεταγμένων.

Την ακολουθία των n αριθμών $(x_1,x_2,.....x_n)$ την ορίζουμε ως **ν-ιάδα**.

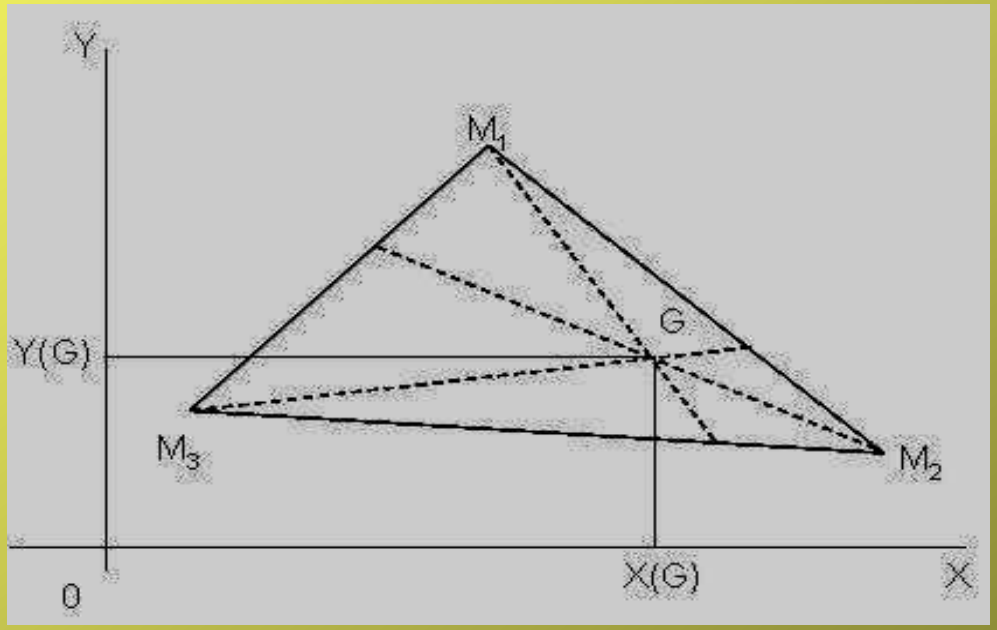
Οπότε το σύνολο των ν-ιάδων λέμε ότι ορίζουν ένα **χώρο n διαστάσεων**.

Κέντρο βάρους ενός συστήματος

Όταν οι μάζες είναι ίσες μεταξύ τους τότε το κέντρο βάρους βρίσκεται στην τομή των διαμέσων του τριγώνου που σχηματίζουν τα τρία σημεία στο επίπεδο.

Αν όμως οι μάζες m_i είναι άνισες μεταξύ τους (σχήμα 1.16), τότε το κέντρο βάρους έλκεται προς το μέρος του σημείου με τη μεγαλύτερη μάζα.

Για τον λόγο αυτό και για να μην υπάρξει σύγχυση το κέντρο βάρους ονομάζεται και **βαρύκεντρο**.



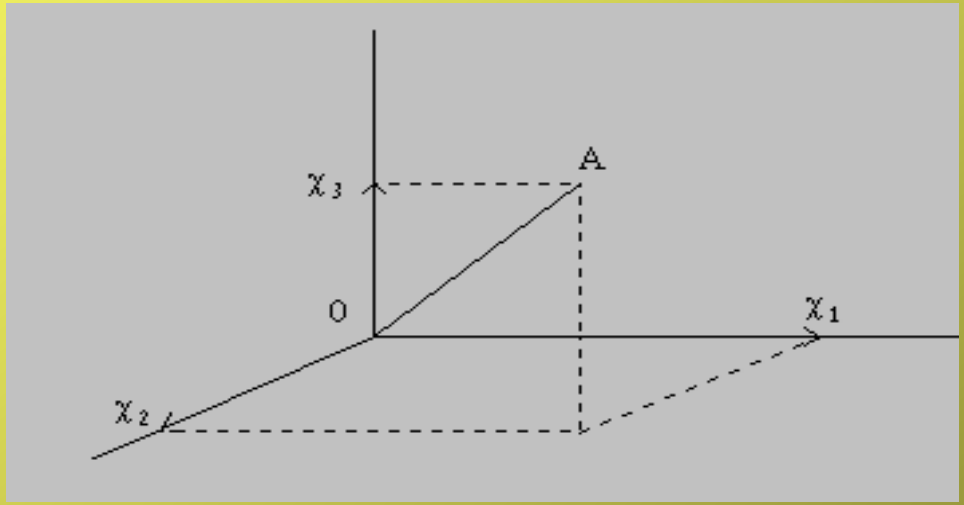
Σχήμα 1.16: *Γραφική παράσταση του βαρύκεντρου*

Ευκλείδεια απόσταση

Στην **Αναλυτική Γεωμετρία** ένα ζεύγος πραγματικών αριθμών (X_1,X_2) ορίζει στο **Καρτεσιανό επίπεδο** ένα σημείο, ενώ μία τριάδα πραγματικών αριθμών προσδιορίζει ένα σημείο στον **τριδιάστατο χώρο**.

Επεκτείνοντας την αντίληψη αυτή μπορούμε να καθορίσουμε ότι μία σειρά ν πραγματικών αριθμών, αποτελεί τις συντεταγμένες ενός σημείου στο **ν-διάστατο χώρο**.

Ένα παράδειγμα στο τριδιάστατο χώρο παρουσιάζει το σχήμα 1.17



Στον ν-διάστατο χώρο ισχύει το γενικευμένο Πυθαγόρειο θεώρημα

$$d^2(M_1, M_2) = \sum_{i=1}^{\nu} (x_2^i - x_1^i)^2$$

Το νέφος N(I) των σημείων που αντιπροσωπεύουν οι γραμμές ενός πίνακα

Στο σύνολο λοιπόν των διανυσμάτων I_i (i=1,2,3,4) (γραμμές του πίνακα 1.22) που καθορίζουν τις τέσσερις ταινίες (οι οποίες αποτελούν τις στατιστικές μονάδες) αντιστοιχούμε ένα σύνολο τεσσάρων σημείων του χώρου R^3. Ήτοι

I_1=(13,2,5), I_2=(20,2,8), I_3=(10,5,5) και I_4=(7,1,22).

Σε κάθε σημείο I_i αντιστοιχούμε την μάζα του m_i, η οποία είναι ίση με την περιθωριακή συχνότητα .Ήτοι

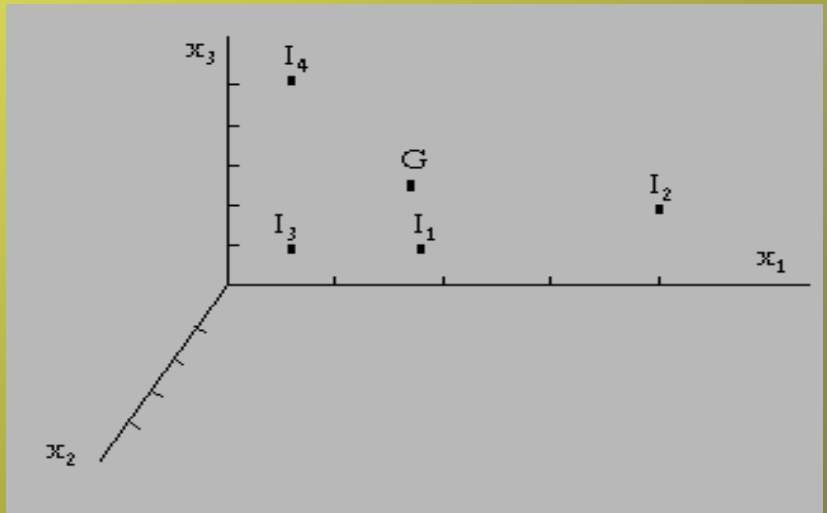
Για το I_1 έχουμε m_1=20, για το I_2 έχουμε m_2=30 για το I_3 έχουμε m_3=20 για το I_4 έχουμε m_4=30

Το βαρύκεντρο βρίσκεται με βάση τον παρακάτω τύπο

$$G_j = \frac{\sum_i m_i \times k_{ij}}{\sum_i m_i} (j = 1,....p)$$

Με τα δεδομένα του πίνακα 1.22 και για j=1,2,3 έχουμε G_1=12.7, G_2=2.3,G_3=11.

Το G_1 λ.χ προέκυψε ως εξής: έχουμε (13x20+20x30+10x20+7x30)/100= 1270/100=12,7



Αδράνεια του νέφους των στατιστικών μονάδων N(I)

Η παρουσιαζόμενη αδράνεια **I(i,G)** κάθε σημείου i (i=1,...n), υπολογίζεται με το άθροισμα των τετραγώνων των αποκλίσεων των συντεταγμένων του σημείου i, από τις αντίστοιχες συντεταγμένες του βαρύκεντρου G, σταθμίζοντας κάθε απόκλιση με το βάρος m_i

$$I(i,G) = \sum_i m_i \| i - G \|^2$$

όπου $$\| i - G \|^2 = \sum_j (i_j - G_j)^2 \text{ για } j = 1,.....p$$

Η **συνολική αδράνεια** του νέφους N(I) ισούται με το άθροισμα όλων των επί μέρους αδρανειών.

$$I(N,G) = I_{ολ} = \sum_i I(i,G)$$

Ενώ η **διασπορά (διακύμανση)** του νέφους N(I) υπολογίζεται από τη σχέση

$$Var(N) = \frac{I(N,G)}{\sum_i m_i}$$

Π.χ Η αδράνεια του I_1 ως προς το βαρύκεντρο ισούται με:

$I(i_1,G) = 20(13-12,7)^2+20(2-2,3)^2+20(5-11)^2 = 723,6$