ΑΝΑΛΥΣΗ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ

- Η μέθοδος της Ανάλυσης σε Κύριες Συνιστώσες, (Analyse en Composantes Principales –ACP-) είναι μία μέθοδος καθαρά περιγραφική, συνίσταται δε σε μία τεχνική η οποία μειώνει την διάσταση ενός συνόλου ποσοτικών δεδομένων (το οποίο αποτελεί δείγμα ενός πληθυσμού), βρίσκοντας ένα νέο σύνολο μεταβλητών πιο μικρό σε αριθμό από εκείνο των αρχικών μεταβλητών, το οποίο όμως περιλαμβάνει το μεγαλύτερο μέρος της πληροφορίας που παρέχει το δείγμα.

- Πρέπει να τονιστεί ιδιαίτερα ότι όταν το πλήθος των μεταβλητών είναι αρκετά μικρό (4,5 ή 6) και το πλήθος των παρατηρήσεων είναι ισχνό (μικρότερο των 100), οφείλει ο ερευνητής να ακολουθήσει άλλη τεχνική π.χ το Alpha του Cronbach.

- Και τούτο διότι η μέθοδος δημιουργεί όπως προαναφέρθηκε νέες μεταβλητές που ονομάζονται «**κύριες συνιστώσες**» οι οποίες αντιστοιχούν σε διανύσματα που εξηγούν τις σχέσεις μεταξύ των μεταβλητών.

- Η ανάλυση στηρίζεται στον πίνακα συσχετίσεων μεταξύ του συνόλου των μεταβλητών, θεωρώντας την διακύμανση κάθε μεταβλητής ίση με 1, έτσι ώστε η συνολική διακύμανση προς ερμηνεία να ισούται με το σύνολο p των μεταβλητών.

- Η πρώτη κύρια συνιστώσα C^1 ερμηνεύει το μεγαλύτερο ποσοστό της συνολικής διακύμανσης, το οποίο προσδιορίζεται από την πρώτη χαρακτηριστική ρίζα, η οποία αντιστοιχεί μετά από διαγωνοποίηση της μήτρας των συσχετίσεων των μεταβλητών.

Διαδικασία κατασκευής και ανάλυσης του νέφους των στατιστικών μονάδων N(I) ενός πίνακα T(nxp)

Σε κάθε γραμμή i του πίνακα T(nxp) αντιστοιχούμε το διάνυσμα $T_j^i=(k_{i1},...k_{ip})$,όπου το k_{ij} παριστάνει την τιμή της μεταβλητής V_j για την i στατιστική μονάδα

Στη συνέχεια οι τιμές κάθε μεταβλητής V_j ομογενοποιείται χρησιμοποιώντας τον παρακάτω μετασχηματισμό

$$x_j^i = k_{ij} - \bar{k}_j$$

Επομένως ο πίνακας T(nxp) μετατρέπεται σ' ένα πίνακα X(nxp) του οποίου οι γραμμές δημιουργούν ένα νέφος σημείων N(I) που ανήκουν στο χώρο R^p. Η μετρική που χρησιμοποιείται είναι η D_p(nxn) και δίνεται από την σχέση

$$D_P = \begin{pmatrix} p_i & 0 & . & 0 & 0 \\ 0 & p_i & . & 0 & 0 \\ . & . & p_i & . & . \\ 0 & 0 & . & p_i & 0 \\ 0 & 0 & . & 0 & p_i \end{pmatrix} \qquad \text{όπου} \quad p_i = \frac{1}{n}$$

Ακολούθως υπολογίζουμε τον πίνακα V(pxp), ο οποίος αποτελεί τον πίνακα διακυμάνσεων - συνδιακυμάνσεων των p μεταβλητών χρησιμοποιώντας την σχέση

$$V(pxp) = X' \times D_p \times X$$

Μετά υπολογίζουμε τον πίνακα των συσχετίσεων μεταξύ των p μεταβλητών χρησιμοποιώντας τη σχέση

$$R(pxp) = D_{1/s} \times V \times D_{1/s}$$

Ο πίνακας $D_{1/s}$ διαστάσεων pxp είναι διαγώνιος πίνακας με στοιχεία τις τιμές $1/s_j$, όπου s_j είναι η τυπική απόκλιση των τιμών κάθε μεταβλητής j.

Παρατήρηση : **Το γενικό στοιχείο r_{ij} του πίνακα R είναι ο συντελεστής συσχέτισης μεταξύ των μεταβλητών X_i και X_j.**

Διαγωνοποιώντας τον τετραγωνικό πίνακα R(pxp), βρίσκουμε τις χαρακτηριστικές ρίζες λ_a και τα αντίστοιχα χαρακτηριστικά διανύσματα U_a για a=1,...p. απ' όπου προκύπτουν οι κύριες συνιστώσες $C^J(i)$ (J=1,…p) των στατιστικών μονάδων σε κάθε παραγοντικό άξονα $\Delta_{\alpha,}$ σύμφωνα με την διανυσματική σχέση

$$\underset{(nxp)}{C^J(i)} = \underset{(nxp)}{X} \times \underset{(pxp)}{U_a}$$

Ο πίνακας Τ(nxp) των κύριων συνιστωσών των στατιστικών μονάδων έχει την παρακάτω μορφή

	C^1 ……. C^j ……. C^P
1	$c^1(1)$ …. $c^j(1)$ ….. $c^P(1)$
.	. . .
i	$c^1(i)$ ……. $c^j(i)$ ……. $c^P(i)$
.	. . .
n	$c^1(n)$ …. $c^j(n)$ …. $c^P(n)$

Η συνολική αδράνεια του νέφους I(N,G) είναι ίση με

$$I(N,G) = \sum_{a=1}^{p} \lambda_a = p$$

Για κάθε κύρια συνιστώσα των στατιστικών μονάδων και σε κάθε παραγοντικό άξονα Δ_a ισχύουν οι παρακάτω σχέσεις:

$$\sum_{i=1}^{n} c^j(i) = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} [c^j(i)]^2 = \lambda_a$$

$$E[c_s(i), c_r(i)] = 0 \quad \text{για } s<r$$

Η ΑΝΑΛΥΣΗ ΤΟΥ ΝΕΦΟΥΣ Ν(J) ΤΩΝ ΜΕΤΑΒΛΗΤΩΝ

Έχοντας βρει τα χαρακτηριστικά διανύσματα U_a του πίνακα R(pxp) έχουμε τη δυνατότητα να υπολογίσουμε τις κύριες συνιστώσες C^J (j=1,..p) των μεταβλητών σε κάθε παραγοντικού άξονα Δ_a, με την διανυσματική σχέση

$$C^J = \sqrt{\lambda_\alpha} \times U_a \qquad (a = 1,...p)$$

όπου U_a το χαρακτηριστικό διάνυσμα που συνδέεται με τον παραγοντικό άξονα Δ_a και λ_a η αντίστοιχη χαρακτηριστική ρίζα του παραγοντικού άξονα Δ_a.

Η εξαγωγή των κύριων συνιστωσών δημιουργεί τον επονομαζόμενο πίνακα κυρίων συνιστωσών. Η κάθε γραμμή αυτού του πίνακα εκφράζει την σχέση της μεταβλητής ως προς τις κύριες συνιστώσες, η δε αριθμητική τιμή του κάθε κελιού (i,j) δηλώνει το ποσοστό της κάθε συνιστώσα C^j που εξηγεί τη μεταβλητή X_i. Οι κύριες συνιστώσες C^J των μεταβλητών ως προς κάθε παραγοντικό άξονα Δ_α επαληθεύουν τη σχέση

$$\sum_{j=1}^{p} \{C_a^j\}^2 = \lambda_a$$

ενώ ως προς το σύνολο όλων των παραγοντικών αξόνων Δ_α (με α=1,…p) για κάθε μεταβλητή X_j ισχύει η σχέση

$$\sum_{a=1}^{p} \{C_a^j\}^2 = 1 \qquad \text{για} \quad j = 1,..p$$

Η ποιότητα προβολής κάθε μεταβλητής σε ένα παραγοντικό άξονα προσδιορίζεται από τη σχέση

$$COR_a(j) = (c_a^j)^2 \qquad \text{για} \quad a = 1,..p$$

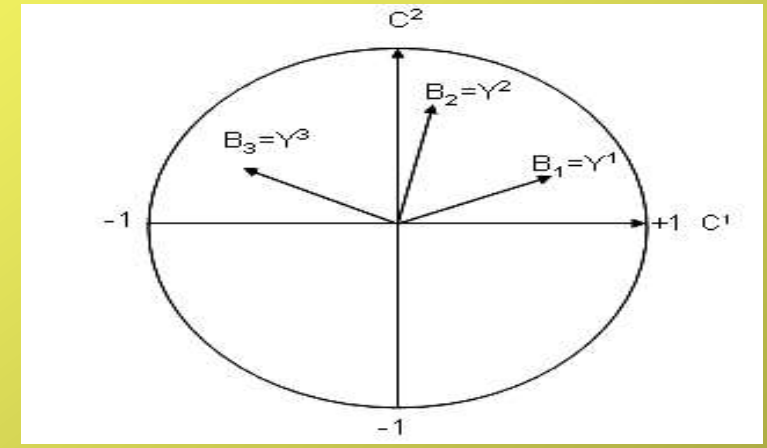
Τέλος επειδή η χαρακτηριστική ρίζα λ_a είναι η αδράνεια που εξηγείται από τον άξονα Δ_a, η συμβολή κάθε μεταβλητής στη διαμόρφωση του παραγοντικού άξονα $_{\Delta a}$ προκύπτει από τη σχέση

$$CTR_a(j) = \frac{(c_a^j)^2}{\lambda_\alpha} \qquad \text{για } \alpha=1,\ldots p$$

Ο κύκλος συσχέτισης

Ορίζοντας με N(J) το νέφος των σημείων-μεταβλητών Y^j, αυτό βρίσκεται εντός υπερσφαίρας ακτίνας r=1,καθόσον η κάθε συντεταγμένη των σημείων αυτών είναι ένας **συντελεστής συσχέτισης**.

Συνεπώς όταν προβάλουμε το νέφος N(J) σε οποιοδήποτε παραγοντικό επίπεδο οι προβολές των σημείων θα βρίσκονται στο εσωτερικό ενός κύκλου (0,1),που καλείται **κύκλος συσχέτισης**



Ο κύκλος συσχέτισης των κυρίων συνιστωσών 1 και 2

Το μήκος R_j των διανυσμάτων-μεταβλητών B_j αντιπροσωπεύει το **συντελεστή πολλαπλής συσχέτισης** $R(Y^j,C^1,C^2)$ μεταξύ της μεταβλητής Y^j με τις δύο πρώτες κύριες συνιστώσες C^1,C^2.

Τα σημεία του νέφους N(J) όσο πλησιάζουν την περιφέρεια του κύκλου, τόσο καλύτερα αντιπροσωπεύονται στο παραγοντικό επίπεδο.

Τα παραγοντικά επίπεδα

Για την ταυτόχρονη παρουσίαση των νεφών N(I) και N(J) στο ίδιο παραγοντικό επίπεδο, πρέπει να ληφθούν υπόψη οι εξής παρατηρήσεις:

- Οι συντεταγμένες των μεταβλητών επειδή προσδιορίζουν συντελεστές συσχέτισης η αριθμητική τιμή της κάθε μιας είναι μικρότερη της μονάδος. Για τον λόγο αυτό είναι σκόπιμο να εξετάζεται αρχικά το παραγοντικό επίπεδο μόνο με τις μεταβλητές, ο οποίος ενέχει θέση κύκλου συσχέτισης.

- Λόγω της ιδιαιτερότητας αυτής για να επιτύχουμε καλύτερη γραφική απεικόνιση των νεφών N(J) και N(I) στα διάφορα παραγοντικά επίπεδα, πολλαπλασιάζουμε είτε τις συντεταγμένες του νέφους N(J) με τον αριθμό $\sqrt{n/p}$ ώστε να εξασφαλίσουμε στις προβολές των σημείων των δύο νεφών μία σχετική συμβατότητα. Και αυτό επειδή η ανάλυση του νέφους N(I) γίνεται σε σχέση με το κέντρο βάρους του νέφους, ενώ δε συμβαίνει το ίδιο με την ανάλυση του νέφους N(J). Οι κύριες συνιστώσες των μεταβλητών επειδή είναι συντελεστές συσχέτισης, πολύ πιθανόν όλες οι μεταβλητές να παρουσιάζουν θετική ή αρνητική συσχέτιση μ' ένα από τους άξονες, οπότε θα βρίσκονται προς την ίδια μεριά του άξονα και όχι εκατέρωθεν της αρχής των αξόνων όπως συμβαίνει με τις συνιστώσες των στατιστικών μονάδων.

- Με την ανάλυση σε κύριες συνιστώσες ερμηνεύονται οι γραμμικές σχέσεις μεταξύ των μεταβλητών. Έτσι ένας μικρός συντελεστής συσχέτισης μεταξύ δύο μεταβλητών παρέχει την ένδειξη ότι οι μεταβλητές αυτές είναι **ανεξάρτητες**, μπορεί βέβαια να υφίσταται μεταξύ τους μία **μη γραμμική σχέση**, η οποία δεν αποκαλύπτεται στη παρούσα φάση της ανάλυσης.

ΕΡΜΗΝΕΙΑ ΤΟΥ ΠΑΡΑΓΟΝΤΙΚΟΥ **ΕΠΙΠΕΔΟΥ**

Οι γραφικές παραστάσεις μεταξύ γραμμών και στηλών ενός πίνακα δεδομένων πάνω στο παραγοντικό επίπεδο επιτρέπουν να οπτικοποιήσουμε τις προσεγγίσεις μεταξύ αντικειμένων (γραμμών) και μεταβλητών (στηλών).
Έτσι στον χώρο R^p των σημείων-αντικειμένων $\{I_k, (k=1,2,\ldots n)\}$ όταν δύο σημεία-αντικείμενα είναι γειτονικά, σημαίνει ότι τα δύο αυτά αντικείμενα χαρακτηρίζονται από τις ίδιες περίπου τιμές για κάθε μία από τις p μεταβλητές, οι δε αποστάσεις μεταξύ των σημείων υπολογίζονται με βάση την Ευκλείδεια απόσταση.
Στον χώρο Rn των σημείων-μεταβλητών $\{J_m, (m=1,2,\ldots p)\}$ όταν δύο σημεία-μεταβλητές είναι κοντά το ένα από το άλλο, αυτό σημαίνει πως το σύνολο των αντικειμένων έδωσε στις μεταβλητές αυτές περίπου τις ίδιες τιμές. Βέβαια οι μονάδες μέτρησης των μεταβλητών μπορεί να διαφέρουν μεταξύ τους, για τον λόγο αυτό είναι απαραίτητο να υπάρξουν μετατροπές των στοιχείων του πίνακα δεδομένων.

ΣΥΜΠΛΗΡΩΜΑΤΙΚΕΣ ΣΤΑΤΙΣΤΙΚΕΣ ΜΟΝΑΔΕΣ

Θεωρούμε ως συμπληρωματική στατιστική μονάδα κάθε μονάδα I_s η οποία δεν συμμετείχε στην ανάλυση, δηλαδή δεν συμπεριλαμβανόταν ως γραμμή του αρχικού πίνακα δεδομένων A(nxp).
Επιθυμία μας είναι να προσδιορίσουμε τις ιδιότητες της I_s σε συνδυασμό με τις ιδιότητες των στατιστικών μονάδων που συμμετείχαν στην ανάλυση. Προς τούτο τοποθετούμε την I_s στα διάφορα παραγοντικά επίπεδα που δημιουργήθηκαν στα προηγούμενα στάδια της ανάλυσης ώστε να προσδιορίσουμε τη συμπεριφορά της.

Προφανώς γνωρίζουμε το διάνυσμα $A(i_s)=\{V_1(i_s),\ldots\ldots,V_p(i_s)\}$

όπου $V_j(i_s)$ είναι η μέτρηση της μεταβλητής Vj στη στατιστική μονάδα i_s

Δημιουργώντας το κανονικοποιημένο διάνυσμα $X(i_s)=\{x_1(i_s),.,x_p(i_s)\}$ πολλαπλασιάζουμε τις συντεταγμένες του διανύσματος $X(i_s)$ με τους συντελεστές κάθε χαρακτηριστικού διανύσματος u_j και βρίσκουμε τις αντίστοιχες κύριες συνιστώσες $C^J(i_s)$, οι οποίες προσδιορίζουν τη θέση της στατιστικής μονάδας i_s στα παραγοντικά επίπεδα που δημιουργούνται από την ανάλυση.

Πλήθος διατηρητέων συνιστωσών

Πίνακας 3.3

Κύριες Συνιστώσες	Χαρακτηριστική ρίζα	Ποσοστό ερμηνείας της διακύμανσης	Αθροιστικό ποσοστό
1	4,384	62,633	62,633
2	1,060	15,140	77,773
3	0,477	6,814	84,587
4	0,368	5,261	89,849
5	0,277	3,959	93,808
6	0,234	3,340	97,148
7	0,200	2,852	100,00
	7,000		

Παρατηρούμε ότι σ' ένα πίνακα με 7 μεταβλητές, το άθροισμα των χαρακτηριστικών ριζών είναι ίσο με 7. Επίσης η πρώτη χαρακτηριστική ρίζα έχει τιμή 4,384, η δεύτερη έχει τιμή 1,060 και η τρίτη 0,477.
Αυτό σημαίνει ότι η πρώτη κύρια συνιστώσα «εξηγεί» 4,384 μονάδες διακύμανσης, ενώ η δεύτερη 1,06 και η τρίτη μόνο 0,477.

Συνεπώς είναι ανώφελο να χρησιμοποιηθούν περισσότερες των δύο πρώτων κυρίων συνιστωσών, οι οποίες ερμηνεύουν το 77,77% της συνολικής διακύμανσης.

Ο Παραγοντικός δείκτης απλότητας (IFS) του H. F. Kaiser

Ο παραγοντικός δείκτης απλότητας (IFS) υπολογίζεται βάση του παρακάτω τύπου

$$IFS_i = \frac{p\left[\sum_{k=1}^{p} a_{ik}^4 - \left(\sum_{k=1}^{p} a_{ik}^2\right)^2\right]}{(p-1)\times\left(\sum_{k=1}^{p} a_{ik}^2\right)^2}$$

Ο δείκτης αυτός υπολογίζεται για κάθε μία μεταβλητή και παίρνει τιμές από 0 έως 1. Για να διατηρηθεί μία μεταβλητή σε μία ACP πρέπει η μεταβλητή να έχει δείκτη με τιμή μεγαλύτερη του 0,5. Ο H. F. Kaiser πρότεινε την παρακάτω διαβάθμιση

Πίνακας 3.4:Πίνακας διαβαθμίσεων του δείκτη IFS

Τιμές	Χαρακτηρισμός
IFS<0,5	Απαράδεκτη
0,50<IFS≤0,60	Άθλια
0,60<IFS≤0,70	Μέτρια
0,70<IFS≤0,80	Ενδιάμεση
0,80<IFS≤0,90	Άξια
0,90<IFS	Εξαιρετική

Πίνακας 3.5:Τιμές δείκτη IFS των επτά μεταβλητών του παραδείγματος

Μεταβλητή	IFS
Τιμή	0,78512
Ποιότητα	0,55894
Αλκοόλ	0,64103
Κύρος	0,74289
Χρώμα	0,59006
Άρωμα	0,79444
Γεύση	0,67012

Ο συντελεστής Alpha του Lee J. Cronbach

Ο συντελεστής Alpha του Lee J. Cronbach υπολογίζεται από τη παρακάτω σχέση:

$$a = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} s_i^2}{\left(\sum_{I=1}^{n} s_i\right)^2}\right)$$

Όπου

n είναι το πλήθος των μεταβλητών Y_i

s_i^2 η διακύμανση των τιμών κάθε μεταβλητής Y

S_i η τυπική απόκλιση κάθε μεταβλητής Y_i

Το a του Cronbach μεταβάλλεται συνήθως μεταξύ 0≤a≤1. Όταν το a ισούται με 1 τότε υφίσταται μεγάλη εσωτερική συνέπεια στις απαντήσεις μεταξύ των μεταβλητών

Πίνακας 3.15: Πίνακας διαβαθμίσεων του δείκτη a

Τιμές	Χαρακτηρισμός
a <0,5	Απαράδεκτη
0,50< a ≤0,60	Φτωχή
0,60< a ≤0,70	Αμφισβητήσιμη
0,70< a ≤0,80	Αποδεκτή
0,80< a ≤0,90	Καλή
0,90<a	Εξαιρετική

Κανόνες διατηρητέων συνιστωσών

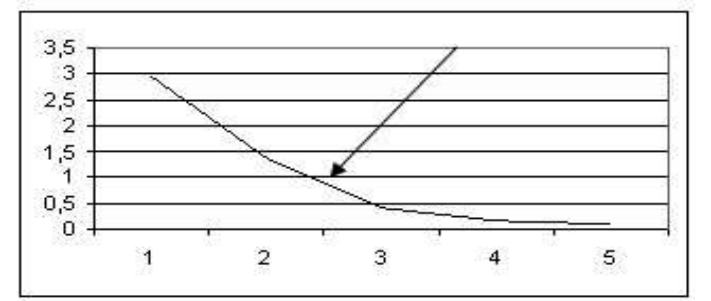
➢**Το κριτήριο του Kaiser**

Το κριτήριο του Kaiser λέει ότι πρέπει να διατηρήσουμε στην ανάλυση όσες κύριες συνιστώσες έχουν χαρακτηριστική ρίζα μεγαλύτερη του 1.

➢**Το κριτήριο του Cattell**

Σχεδιάζουμε το γράφημα των χαρακτηριστικών ριζών σε συνάρτηση της τάξης των συνιστωσών (1η,2η,.. κ.λ.π)

Η ένδειξη μέχρι πόσες κύριες συνιστώσες θα διατηρήσουμε στην ανάλυση, αποτελεί το σημείο όπου εκδηλώνεται έντονη αλλαγή της κλίσης του γραφήματος.



Σχήμα 3.5: Γραφική λύση του κριτηρίου Cattell

➢**Το κριτήριο του Horn**

Σύμφωνα με τον Horn πρέπει να διατηρηθούν τόσες κύριες συνιστώσες όσες έχουν τιμή μεγαλύτερη από την αντίστοιχη κύρια συνιστώσα που προέκυψε με τυχαίο τρόπο

Πίνακας 3.18: Χαρακτηριστικές τιμές κανονικές λ_i και τυχαίες λ_T

λ_i	λ_T
2,941952	2,167485
1,377971	1,325574
0,413693	0,830216
0,177038	0,458513
0,089347	0,218212