ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΣΤΟΙΧΕΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

1.1 Γενικά

Στην Στατιστική, η έννοια του πληθυσμού δεν έχει σχέση με την έννοια του πληθυσμού ανθρώπων που θεωρούμε στην καθημερινή ζωή. Έτσι, η έννοια του πληθυσμού αναφέρεται σε κάποιο πλήθος δεδομένων που μπορούν να ενταχθούν σε ποσοτικές ή ποιοτικές κατηγορίες. Τα δεδομένα αυτά μπορεί να είναι μετρήσεις μεγεθών όπως ημερομίσθια, μήκη πραγμάτων ή ποσοτητες υλικών ακόμα και ποσοτητες φυσικών μεγεθών όπως π.χ η θερμοκρασία, κατηγορίες γεγονότων ή ακόμη και εννοιών όπως η απόδοση φοιτητών στα μαθήματα χωρίς βαθμό π.χ Απορριπτέος, Μέτριος, Καλός, Πολύ Καλός, Άριστος ή και η προτίμηση ψηφοφορών πρός τα διάφορα κόμματα.

Σύμφωνα με τις παραπάνω μετρήσεις ή παρατηρήσεις, δημιουργούμε τις μεταβλητές οι οποίες σε κάθε δεδομένο (άτομο ή στοιχείο) αντιστοιχούν σε μία τιμή **[βιβλιογρ. 5].** Οι μεταβλητές αυτές λέγονται ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ, το σύνολο των ατόμων ή στοιχείων (ή στατιστικών μονάδων **[βιβλιογρ. 4]** επάνω στις οποίες αντιστοιχούν οι τιμές των τυχαίων μεταβλητών λέγεται ΠΛΗΘΥΣΜΟΣ. Το πλήθος των στοιχείων του πληθυσμού είναι το μέγεθος του πληθυσμού.

Ωστόσο, λόγω του μεγάλου μεγέθους του πληθυσμού που μελετάμε και το κόστος καταγραφής των τιμών που πάιρνει μία Τ.Μ., σ'όλα τα στοιχεία ενός πληθυσμού, μελετάμε δείγματα του πληθυσμού. ΔΕΙΓΜΑ είναι ένα υποσύνολο ενός πληθυσμού.

Επειδη ο σκοπός της Στατιστικής είναι η εξαγωγή συμπερασμάτων που έχουν βασιστεί σε δείγματα αλλά που αφορούν όλο τον πληθυσμό που μελετάμε, πρωταρχικό ρόλο παίζουν τα αντιπροσωπευτικά δείγματα ενός πληθυσμου.

ΑΝΤΙΠΡΟΣΩΠΕΥΤΙΚΟ είναι το δείγμα ενός πληθυσμού όταν τα συμπεράσματα που βγάζουμε απο το δείγμα αυτό είναι με πολυ μικρή προσέγγιση τα ίδια που θα βγάζαμε εάν μελετούσαμε ολόκληρο τον πληθυσμό. Κατα συνέπεια, το ΑΝΤΙΠΡΟΣΩΠΕΥΤΙΚΟ δείγμα έχει σύσταση και δομή όμοια και ανάλογη σε μικρότερο πλήθος μ'αυτές του πληθυσμού.

Ένα δείγμα λέγεται ΑΜΕΡΌΛΗΠΤΟ ή ΤΥΧΑΙΟ, εάν κάθε άτομα του πληθυσμού, στον οποίο ανήκει το δείγμα έχει ακριβός την ίδια πιθανότητα να περιλαμβάνεται σ'αυτό το δείγμα.[**βιβλιογρ. 2**]

1.2 Βασικά στάδια μιας δειγματοληψίας [βιβλιογρ.1]

Α) Σκοπός της δειγματοληψίας.

Εάν ο (ή οι) σκοπός για τον οποίο γίνεται η δειγματοληψία δεν είναι τελείως συγκεκριμένος υπάρχει κίνδυνος σημαντικής αποκλίσεως απο αυτόν μετά το τέλος της δειγματοληψίας.

Β) Σαφής καθορισμός του πληθυσμού από τον οποίο εκλέγονται τα δείγματα.

Είναι πολύ σημαντικό σε κάθε δειγματοληψία να πούμε με βεβαιότητα εάν ένα στοιχείο ανήκει ή όχι στον πληθυσμό που μελετάμε. Αυτό προϋποθέτει αυστηρό καθορισμό των χαρακτηριστικών, των στοιχείων του πληθυσμού που θέλουμε να μελετήσουμε, πριν φτάσουμε στην δειγματοληψία. Έτσι, για έναν πληθυσμό από κάποιο απλό μηχανικό εξάρτημα ενός αυτοκινήτου π.χ. δεν θα υπάρξει δυσκολία. Αντίθετα πρέπει να ορίσουμε με ακρίβεια τα πλάισια στα οποία ανήκει η "πολυκατοικία" ώστε να μπορούμε να την κατατάξουμε σε κάποιο δείγμα ή όχι, σε μία συγκεκριμένη δειγματοληψία κάποιου τύπου κτιρίων που χαρακτηρίσαμε σαν πολυκατοικίες.

Γ) Ακριβής συλλογή των απαραιτήτων δεδομένων

Λιγότερα, περισσότερα, άχρηστα ή ασαφή δεδομένα κατά την διάρκεια της δειγματοληψίας δίνουν λανθασμένα αποτελέσματα.

Δ) Προσδιορισμός της ακρίβειας των αποτελεσμάτων.

Επειδή πρόκειται για δειγματοληψία, τα συμπεράσματα υπόκεινται πάντα σε κάποιο ποσοστό σφάλματος. Το σφάλμα αυτό μειώνεται εάν χρησιμοποιηθούν μεγαλύτερα δείγματα ή και ακριβέστερα όργανα καταγραφής (μέτρησης). Ωστόσο, ο από την αρχή προσδιορισμός του μεγέθους του πιθανού σφάλματος προσδιορίζει σαφέστερα τα αποτελέσματα.

Ε) Μέθοδοι και ακρίβεια των μετρήσεων

Τα ερωτηματολόγια πρέπει να είναι καθολικά και ομοιόμορφα και προσεκτικά ελεγμένα μετα την συμπλήρωση τους. Σε μετρήσεις με όργανα, πρέπει να υπάρχει έλεγχος της καλής λειτουργίας και την αξιοπιστίας τους.

Και στις δύο περιπτώσεις, οι μετρήσεις πρέπει να καταγράφονται σε Η/Υ μεγάλης ισχύος ώστε, με το κατάλληλο λογισμικό, να είναι δυνατή η επεξεργασία τους.

ΣΤ) Το πλάισιο δειγματοληψίας

Απαραίτητη προϋπόθεση για την εκλογή ενός δείγματος, είναι ο διαχωρισμός του πληθυσμού σε μονάδες ή άτομα από τα οποία θα αποτελείται και το δείγμα. Πολλές φορές, τυχαίνει τα άτομα αυτά να είναι καθορισμένα σαν μονάδες - οντότητες αλλά αρκετα συχνά είναι αναγκαίο να προσδιοριστούν απο τον σκοπό της δειγματοληψίας.

Έτσι, για παράδειγμα, στην περίπτωση γεωργικών ιδιοκτησιών πρέπει να καθοριστούν η έκταση, η δυνατότητα και η χρήση της ιδιοκτησίας, ο αριθμός των ιδιοκτητών και ίσως άλλα χαρακτηριστικά χρήσιμα στόν καθορισμό της γεωργικής ιδιοκτησίας σαν μονάδα δειγματοληψίας.

Ζ) Η εκλογή του δείγματος

Ο τρόπος εκλογής του δείγματος καθώς και ο προσδιορισμός του μεγέθους του θα καθορίσουν και τα αποτελεσματά μιάς δειγματοληψίας. Ωστόσο, σε μία δειγματοληψία, πρέπει να λαμβάνεται υπ'όψη ο χρόνος και το απαιτούμενο ποσο χρημάτων για την διεκπεραίωση της.

Η) Προκαταρτική δειγματοληψία

Μία εικονική και σε μικρό μέγεθος του πληθυσμού προκαταρτική δειγματοληψία δείχνει ατέλειες που πιθανά θα δημιουργήσουν πρόβλημα εάν δεν ληφθούν υπ'όψη. Τέτοιες μπορεί να είναι ασαφείς μετρήσεις ή πολυ ψηλότερο από το εκτιμητέο κόστος.

Θ) Οργάνωση κατά την διάρκεια της δειγματοληψίας

Ο κατά μικρά χρονικά διαστήματα και συνεχής έλεγχος της συλλογής των δειγμάτων καθ'όλη την διάρκεια της δειγματοληψίας είναι απαραίτητος, όπως επίσης και η αντιμετώπιση της αδυναμίας λήψης δείγματος (ή αδυναμίας συγκέντρωσης ερωτηματολογίων). Τέλος, η κατά τον ταχύτερο και ακριβέστερο τρόπο συνεργασία των εργαζομένων για μια δειγματοληψία βοηθά πολύ στην σωστή διεξαγωγή της.

Ι) Συγκέντρωση και ανάλυση δεδομένων

Αν και σήμερα (1997) είναι πολύ δύσκολό οι εργαζόμενοι σε μία δειγματοληψία να καταγράφουν τα δείγματα σε ηλεκτρονικά μέσα (δισκέττες ή CD ηλεκτρονικών υπολογιστών) που επικοινωνούν μεταξύ τους μέσω κάποιου ηλεκτρονικού δικτύου (π.χ. INTERNET). Στο μέλλον, είναι πολύ πιθανό, να συμβεί αυτό, αλλά σήμερα θα πρέπει να συγκεντρώνονται τα δεδομένα, να ελέγχονται τα ερωτηματολόγια ή οι μετρήσεις, να καταγράφονται σε ηλεκτρονικούς υπολογιστές με προσοχή και να επεξεργάζονται από ειδικούς επιστήμονες με την βοήθεια του κατάλληλου λογισμικού που εκτός απο άλλες πληροφορίες σίγουρα θα πρέπει να μας δίνει και το ποσοστό λάθους για τους εκτιμητές των πιο αντιπροσωπευτικών μεθόδων του πληθυσμού.

ΙΑ) Πληροφορίες για μελλοντικές δειγματοληψίες

Λάθη στην όλη διαδικασία και πιθανά προβλήματα οδηγούν στην αποφυγή τους σε μια παρόμοια μελλοντική διαδικασία.

Συμπεράσματα για το μέγεθος του δείγματος και για τους εκτιμητές μεγεθών του πληθυσμού είναι οδηγοί για μια σωστότερη δειγματοληψία.

1.3 Εκτιμητες και κανονικη κατανομή

Έστω ότι θέλουμε να εκτιμήσουμε με δειγματοληψία την μέση τιμή του πληθυσμού μ. Αποφασίζουμε να πάρουμε μια σειρά δειγμάτων στό κάθε ένα από τα οποία υπολογίζουμε την μέση τιμή του έστω μ_i

Έτσι, δημιουργείται μια κατανομή των δειγματοληπτικών μέσων τιμών μ_i των δειγμάτων που πήραμε.

Η μέση τιμή του κάθε δείγματος αποτελεί έναν εκτιμητή της μέσης τιμής του πληθυσμού. Η καλύτερή προσέγγισή της μέσης τιμής του πληθυσμού είναι εάν βρούμε την μέση τιμή της κατανομής των δειγματοληπτικών μέσων.

Ένας εκτιμητής $\hat{\mu}$ της μέσης τιμής μ του πληθυσμού λέγεται **αμερόληπτος**, εάν η μέση τιμή του που προκύπτει απο όλα τα δυνατά δείγματα του πληθυσμού ισούται με την μέση τιμη του πληθυσμού. [βιβλιογρ. 5]

Εάν το κάθε δείγμα i απο τα n συνολικά δείγματα έχει πιθανότητα $\frac{1}{P_i}$ να εκλεγεί απο τον πληθυσμό, τότε η μαθηματική ελπίδα του αμερόληπτου εκτιμητή θα είναι

$$E(\hat{\mu}) = \sum_{i=1}^{N} \frac{1}{P_i} \hat{\mu}_i = \mu$$

αν δε όλα τα δείγματα είναι ισοπίθανα να εκλεγούν μεταξύ τους δηλ. $\frac{1}{P_i} = \frac{1}{\eta} \forall_i$ τότε

$$E(\hat{\mu}) = \frac{\sum \hat{\mu}_{\hat{i}}}{n} = \mu$$

Όταν τα δείγματα είναι αρκετά μεγάλα σε μέγεθος και αρκετά σε πλήθος τότε η κατανομή των εκτιμητών που προκύπτει απο αυτά ειναι προσεγγιστικά κανονική.

Σύμφωνα μ'αυτό το απόλυτο σφάλμα της εκτίμησης $|\hat{\mu} - \mu|$ της μέσης τιμής του πληθυσμού μ απο τον δειγματικό μέσο $\hat{\mu}$ (με διακύμανση $\sigma_{\hat{\mu}}$) έχει πιθανότητα:

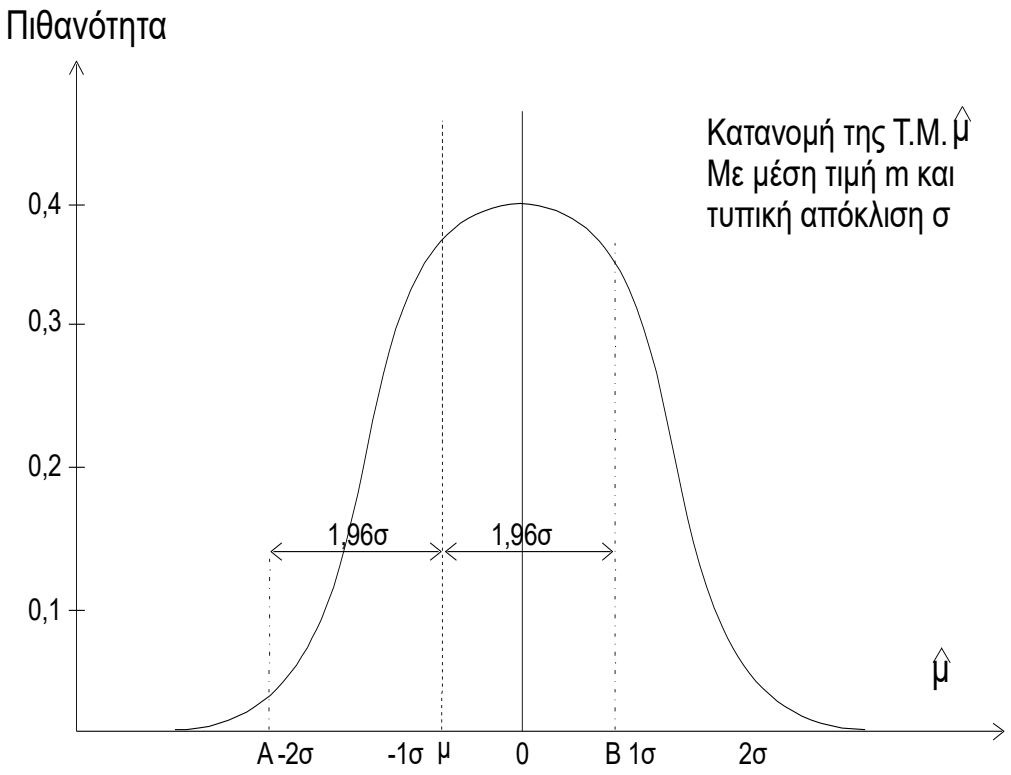
0,68 να μην υπερβαίνει την τιμή της $\sigma_{\hat{\mu}}$

0,95 να μην υπερβαίνει τιμή $\approx 2\sigma_{\hat{\mu}} \left(1{,}96\sigma_{\hat{\mu}}\right)$

0,99 να μην υπερβαίνει τιμή $= 2{,}58\sigma_{\hat{\mu}}$

1.4 Απόκλιση και μέσο τετραγωνικό σφάλμα ενός εκτιμητή [βιβλιογρ. 1]

Έστω ότι ο δειγματικός μέσος $\hat{\mu}$ κατανέμεται κανονικά με μέση τιμή m και μια απόκλιση $K = m - \mu$ απο τον πραγματικο μέσο μ του πληθυσμού, τον οποίο αγνοούμε και συνεπώς αγνοούμε και την ύπαρξη της απόκλισης K του πληθυσμού. Υπολογίζουμε την τυπική απόκλιση σ απο την μέση τιμή m



Σύμφωνα με την κανονική κατανομή, και θεωρώντας ότι δεν υπάρχει απόκλιση της μέσης τιμής m που θεωρούμε, απο την πραγματική μέση τιμή μ του πληθυσμού, θεωρούμε ότι η πιθανότητα το μέγεθος $\hat{\mu}$ να ξεπερνά την τιμή 1,966 είναι ίση με 0,05. Η πιθανότητα αυτή είναι στην δεξιά πλευρά της κατανομής και υπολογίζεται απο το εμβαδόν που εκφράζει το ολοκλήρωμα

$$\frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{\mu+1,966}^{\infty} e^{-\frac{(\hat{\mu}-m)^2}{2\cdot\sigma^2}} d\hat{\mu}$$

Τυποποιόντας την μεταβλητή $\hat{\mu}$ θέτουμε $t = \frac{\hat{\mu} - m}{\sigma} \Rightarrow \hat{\mu} = m + \sigma t \Rightarrow d\hat{\mu} = \sigma \cdot dt$. Έτσι το κάτω όριο της ολοκλήρωσης γίνεται $\frac{\mu + 1,966 - m}{\sigma} = \frac{\mu - m}{\sigma} + 1,96$ αλλά καθώς έχουμε την απόκλιση $K = m - \mu$ το τελευταίο ισούται με $1,96 - \frac{K}{\sigma}$ και έτσι, η πιθανότητα (δηλ. το εμβαδόν) είναι ίση με

$$\frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \int_{1,96-\frac{K}{\sigma}}^{\infty} e^{-\frac{t}{2}} \cdot \sigma \cdot dt = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{1,96-\frac{K}{\sigma}}^{\infty} e^{-\frac{t^2}{2}} dt$$

Στην αριστερή πλευρά της κατανομής, η αντίστοιχη πιθανότητα είναι

$$\frac{1}{2 \cdot \pi} \int_{-\infty}^{-1,96-\frac{K}{\sigma}} e^{-\frac{t^2}{2}} dt$$

Ο Cochran, **[1]** παραθέτει έναν πίνακα στο βιβλίο του "Sampling Techniques" που δίνει την πιθανότητα λάθους που αντιστοιχεί σε στάθμη σημαντικότητας 0,05 καθώς ο λόγος $\frac{K}{\sigma}$ αυξάνει. Μέρος του πίνακα αυτού είναι:

	Πιθανότητα Λάθους		
K/σ	<-1.96 σ	>1.96 σ	Συνολική
0.02	0.0238	0.0262	0.05
0.06	0.0217	0.0287	0.0504
0.1	0.0197	0.0314	0.0511
0.4	0.0091	0.0594	0.0685
0.8	0.0029	0.123	0.1259
1	0.0015	0.1685	0.17

1.5	0.0003	0.3228	0.3231

Θεωρώντας, τώρα ότι η μέση τιμή του εκτιμητή $\hat{\mu}$ είναι m και άρα $E(\hat{\mu} - m) = 0$, και ότι η απόκλιση της εκτιμούμενης μέσης τιμής απο την πραγματική τιμή του πληθυσμού είναι $|m - \mu|$, ορίζεται σαν μέτρο σύγκρισης 2 εκτιμητών το Μέσο Τετραγωνικό Σφάλμα του εκτιμητή που είναι:

$$M.S.E.(\hat{\mu}) = E(\hat{\mu} - \mu)^2 = E\left[(\hat{\mu} - m) + (m - \mu)\right]^2 =$$
$$= E(\hat{\mu} - m)^2 + 2 \cdot (m - \mu) \cdot E(\hat{\mu} - m) + (m - \mu)^2 = \sigma_{\hat{\mu}}^2 + \left(\text{αποκλ}_{\hat{\mu}}\right)^2$$

1.5 Μέθοδοι δειγματοληψίας

Οι μέθοδοι δειγματοληψίας χωρίζονται σε δύο μεγάλες κατηγορίες.

1) Δειγματοληψίες με πιθανότητα
2) Κατευθυνόμενες ή δειγματοληψίες σκοπιμοτητας

Εδώ, θα ασχοληθούμε πιο αναλυτικά με την πρώτη κατηγορία και θα αναφερθούμε συνοπτικά στα γενικά χαρακτηριστικά κάθε μεθόδου της δεύτερης.

Οι δειγματοληψίες με πιθανότητα προτιμούνται για τους παρακάτω λόγους:

1) Δεν υπόκεινται σε υποκειμενικά κριτήρια
2) Σ' αυτές, είναι δυνατόν, να βρεθεί ένα διάστημα εμπιστοσύνης για τον εκτιμητη που χρησιμοποιείται
3) Αυξανομένου του μεγέθους του δείγματος αυξάνουν το παρεχόμενο ποσο της πληροφορίας. [Φαρμάκης].

Αντίθετα, στις κατευθυνόμενες δειγματοληψίες, οι επιλογές των στοιχείων των δειγμάτων γίνεται σύμφωνα με τις αντιλήψεις των μελετητών.

Προτιμούνται όταν υπάρχει σαφής γνώση του πληθυσμού και όταν απαιτείται η εκλογή μικρού μεγέθους δειγματών. Επιπλέον πολλές απο αυτές έχουν σημαντικά μικρότερο κόστος από τις δειγματοληψίες με πιθανότητα.

1.5.1 Δειγματοληψίες με πιθανότητες

1.5.1.1 Απλή Τυχαία Δειγματοληψία (ΑΤΔ) [Simple Random Sampling]

Προϋποθέσεις

1. Κάθε δείγμα μεγέθους n έχει την ίδια πιθανότητα να εκλεγεί με οποιοδήποτε άλλο δείγμα ίδίου μεγέθους
2. Κάθε στοιχείο του ενεργού πληθυσμού, έχει την ίδια πιθανότητα να εκλεγεί στο δείγμα που λαμβάνεται απο τον πληθυσμό αυτό.

Πιθανότητα Εκλογής Δειγματος Μεγέθους n

Η πιθανότητα επιλογής του πρώτου στοιχείου δείγματος μεγέθους n απο τον πληθυσμό μεγέθους N είναι $\frac{\eta}{N}$.Του δευτέρου στοιχείου και εφ'όσον ξέρουμε ότι το πρώτο έχει επιλεγεί και άρα έχουν μείνει η-1 στοιχεία για την συμπλήρωση του δείγματος και N-1 στοιχεία του πληθυσμού είναι $\frac{\eta-1}{N-1}$. Άρα, για παράδειγμα, για ένα 2-μελές δείγμα η πιθανότητα εκλογής του θα ήταν $\frac{\eta}{N}\times\frac{\eta-1}{N-1}$, ενώ για ένα n-μελές η πιθανότητα εκλογής του είναι

$$P=\frac{\eta}{N}\cdot\frac{\eta-1}{N-1}\cdots\frac{\eta-(\eta-1)}{N-(\eta-1)}=\frac{\eta}{N}\cdots\frac{1}{N-\eta+1}=\frac{\eta!}{\frac{N!}{(N-\eta)!}}=\frac{1}{\frac{N!}{\eta!\cdot(N-\eta)!}}=\frac{1}{\binom{N}{\eta}}$$

Τρόπος Εκλογής Δείγματος Μεγέθους n

Α. Η Μέθοδος της Κάλπης (Παλία Μέθοδος) [βιβλιογρ. 4]

1ο Βήμα

Αριθμούμε, χωρίς παραλήψεις ή επαναλήψεις, τα Ν στοιχεία του πληθυσμού απο 1 έως Ν

2ο Βήμα

Δημιουργούμε και αριθμούμε Ν κλήρους ή Ν μπαλάκια τα οποία τοποθετούμε σε μία κάλπη

3ο Βήμα

Τραβάμε από την κάλπη τόσους κλήρους όσους και το μέγεθος του δείγματος δηλαδή n

4ο Βήμα

Αφού σημειώσουμε την αρίθμηση που είχαν οι κλήροι που επιλέχτηκαν απο την κάλπη, βρίσκουμε και ελέγχουμε απο τον πληθυσμό τα άτομα που έχουν την ίδια αρίθμηση με τους n κλήρους.

Η μέθοδος αυτή επειδή είναι ιδιαιτέρα δύσχρηστη και έχει εγκαταλειφθεί, εκτός απο ειδικές περιπτώσεις τυχερών παιγνιδιών (λόττο, λαχεία κ.τ.λ).

Β. Μέθοδος των Τυχαίων Αριθμών με την Βοήθεια Ηλεκτρονικών Υπολογιστών.

Σήμερα, οι αξιόπιστοι ηλεκτρονικοί υπολογιστές τσέπης (κομπιουτεράκια) καθώς και τα προγράμματα ηλεκτρονικών υπολογιστών (PC’s) που χρησιμοποιούν τυχαίους αριθμούς είναι εφοδιασμένοι με υποπρογράμματα (Functions ή Subroutines) παραγωγής ψευδοτυχαίων αριθμών (πλήκτρα ή προγράμματα που η ονομασία τους περιέχει μέρος ή ολόκληρη την λέξη random). Έτσι, ο χρήστης έχει στην διάθεση του πίνακες τυχαίων αριθμών για την εκλογή τυχαίου δείγματος βάσει των πινάκων αυτών, και ακολούθει την διαδικασία που περιγράφεται παρακάτω.

Γ. Μέθοδος Επιλογής των Στοιχείων του Δείγματος Βάσει του Πίνακα των Τυχαίων αριθμών.

Οι πίνακες των τυχαίων αριθμών, είναι πίνακες των μονοψηφίων ακεραίων αριθμών 0 έως 9 όπου ο κάθε ένας απο αυτούς έχει την ίδια πιθανότητα επιλογής με τους υπόλοιπους.

Παρακάτω παραθέτουμε μέρος ένος πίνακα τυχαίων αριθμών που δίνει ο Cochran στο βιβλίο του **[1]** 'Sampling Methods', είναι μονοψήφιοι αριθμοί τοποθετημένοι σε ομάδες απο 5 γραμμές και 5 στήλες και το σύνολο του πίνακα περιέχει 20 γραμμές και 50 στήλες.

Η διαδικασία επιλογής τυχαίου δείγματος με την βοήθεια του πίνακα τυχαίων αριθμών περιγράφεται με το παρακάτω παράδειγμα:

	00–04	05–09	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067

Έστω ότι θέλουμε να επιλέξουμε ένα δεκαμελές δείγμα απο μία ομογενή περιοχή που αποτελείται απο 100 pixels (pixel: στοιχειώδες κομμάτι εικόνας) σε μία δορυφορική εικόνα που είναι σε ψηφιακή μορφή. Κάθε pixel έχει μια τιμή απο 0 έως 255 και η δορυφορική εικόνα απεικονίζει κάποια περιοχή.

1ο Βήμα

Αριθμούμε τα (pixels) στοιχεία όλου του πληθυσμού μας που στο συγκεκριμένο παράδειγμα είναι το σύνολο των N=100 pixels και σχηματίζουμε τον ακόλουθο πίνακα.

Πληθυσμός 100 ραδιομετρικών τιμών μιας ομογενούς περιοχής ως προς την κάλυψη εδάφους μιας περιοχής (N=100)									
1	58	21	46	41	63	61	76	81	40
2	47	22	57	42	58	62	70	82	47
3	49	23	53	43	48	63	37	83	37
4	56	24	61	44	44	64	45	84	53
5	47	25	63	45	49	65	49	85	68
6	39	26	58	46	64	66	58	86	67
7	56	27	59	47	39	67	53	87	49
8	47	28	71	48	59	68	54	88	40
9	58	29	54	49	57	69	68	89	51
10	67	30	48	50	67	70	62	90	54
11	66	31	49	51	62	71	66	91	72
12	64	32	50	52	60	72	54	92	65
13	52	33	40	53	67	73	59	93	64
14	39	34	58	54	58	74	67	94	68
15	40	35	63	55	71	75	68	95	38
16	58	36	57	56	67	76	67	96	47
17	70	37	63	57	46	77	38	97	55
18	59	38	62	58	49	78	55	98	56
19	67	39	48	59	54	79	47	99	49
20	48	40	47	60	47	80	46	100	48

2ο Βήμα

Εκλέγουμε στην τύχη 2 αριθμούς (για τον συγκεκριμένο πίνακα απο 1 έως 20 και απο 1 έως 50, γενικά απο 1 εώς του αριθμού των γραμμών και από το 1 έως τον αριθμό των στηλών) π.χ. 7,36 και βρίσκουμε τον τυχαίο αριθμό του πίνακα που αντιστοιχεί στην γραμμή 7 και στήν στήλη 36, είναι ο αριθμός 2.

Επίσης, σημειώνουμε τον αριθμό των ψηφίων που έχει το πλήθος του πληθυσμού μας, εδώ K=3.

3ο Βήμα

Ακολουθούμε την γραμμή ή την στήλη που κατέχει ο τριψήφιος αριθμός που με πρώτο ψηφίο το 2, που βρήκαμε στο 2ο βήμα εδώ για παράδειγμα ο 248 και συνεχίζοντας κάθετα ή οριζόντια, διαλέγουμε όλους τους τριψήφιους αριθμούς που

βρίσκονται μεταξύ 1 και 100 (ουσιαστικά εκτός από το 100 όλους όσους έχουν για πρώτο ψηφίο 0)

Παρατηρούμε ότι, η μέθοδος αυτή έχει ΜΕΓΑΛΟ ΠΟΣΟΣΤΟ ΑΠΟΡΡΙΨΗΣ ΤΥΧΑΙΩΝ ΑΡΙΘΜΩΝ. Γι'αυτό, στην περίπτωση που ο Ν είναι Κ-ψήφιος και μικρότερος απο $\frac{10^{\kappa}}{2}$ όπως εδώ που είναι $100 < \frac{1000}{2} = 500$, εφαρμόζουμε διάφορες μεθόδους για να μειώσουμε το μεγάλο ποσοστό απόρριψης εάν βέβαια $N > \frac{10^{\kappa}}{2}$ ακολουθούμε την αρχική διαδικασία του 3ο βήματος.

1η μέθοδος

Χωρίζουμε όσο το δυνατόν με μικρότερη απώλεια[1] το διάστημα $10^{\kappa} - N$ (εδώ $10^3 - 100 = 900$) σε ίσα διαστήματα (εδώ ανά 100), ακολουθώντας όπως και πρίν την διαδικασία έως το βήμα 3 δεν απορρίπτουμε τους τριψήφιους >100 αλλά:

Αφαιρούμε 100 απο τους αριθμούς μεταξύ 101 και 200
Αφαιρούμε 200 απο τους αριθμούς μεταξύ 201 και 300
Αφαιρούμε 300 απο τους αριθμούς μεταξύ 301 και 400

Αφαιρούμε 900 απο τους αριθμούς μεταξύ 901 και 999

Σημειώνουμε τους 10 πρώτους αριθμούς που προκύπτουν και είναι μεταξύ 0 και 101 και που θα αποτελέσουν τους αύξοντες αριθμούς των στοιχείων του δείγματος που θα εκλεγεί απο τον πληθυσμό μας (N=100)

2η μέθοδος

Διαλέγουμε τόσους αριθμούς όσους το δείγμα (εδώ 10) κατά στήλη ή σειρά σύμφωνα με το 3ο βήμα. Όσοι αριθμοί είναι >100 τους διαιρούμε με 100 και στην θέση τους, κρατάμε το υπόλοιπό της διαίρεσης.

4ο Βήμα

[1] Εάν N=137 θα αρχίζαμε από το 201 απορρίπτοντας τους ενδιάμεσους από το 138 έως το 200

Από τον πληθυσμό μας, επιλέγουμε τους αριθμούς που έχουν αριθμηθεί σύμφωνα με τους 10 τυχαίους αριθμούς που εκλέξαμε στο 3ο βήμα.

Εφαρμογή: 2 μέθοδοι στο παράδειγμα του πλήθυσμού των 100 ραδιομετρικών τιμών

Άσκηση

Να εφαρμοστεί η 1η μέθοδος του 3ου βήματος για N=145 και 10-μελές δείγμα και να βρεθεί το ποσοστό απόρριψης στον πίνακα των τυχαίων αριθμών και να σιγκριθεί με το θεωρητικό ποσοστό απόρριψης της μεθόδου. [Cochran, σελ 20]

Θεώρημα 1

Η δειγματική μέση τιμή μ είναι ένας αμερόληπτος εκτιμητης της μέσης τιμής του πληθυσμού.

Απόδειξη:

$$\text{Έχουμε } \mathrm{E}\left(\hat{\mu}\right) = \frac{\sum_{i=1}^{\binom{N}{\mu}} \hat{\mu}_i}{\binom{N}{\eta}} = \frac{\left(\hat{\mu}_1 + \hat{\mu}_2 + \cdots + \hat{\mu}_{\binom{N}{\eta}}\right)}{\binom{N}{\eta}} \qquad (1)$$

έστω $\chi_i^1, \chi_i^2, \cdots, \chi_i^\eta$ οι μονάδες του πληθυσμού που ανήκουν στο τυχόν δείγμα μεγέθους n

Έτσι, η σχέση (1):

$$= \frac{1}{\binom{N}{\eta}} \cdot \left[\frac{1}{\eta}\left(\chi_1^1 + \chi_1^2 + \cdots + \chi_1^\eta\right) + \frac{1}{\eta}\left(\chi_2^1 + \chi_2^2 + \cdots + \chi_2^\eta\right) + \frac{1}{\eta}\left(\chi_{\binom{N}{\eta}}^1 + \chi_{\binom{N}{\eta}}^2 + \cdots + \chi_{\binom{N}{\eta}}^\eta\right) \right]$$

όπου τα διάφορα χ_i^j μπορεί να συμπίπτουν μεταξύ τους.

Έτσι, για να βρεθεί το παραπάνω άθροισμα, πρέπει να βρούμε για κάθε μονάδα του πληθυσμού, σε πόσα δείγματα μπορεί να ανήκει. Το βρίσκουμε με τον παρακάτω συλλογισμό:

Παίρνουμε μια μονάδα του πληθυσμού και την τοποθετούμε για να σχηματίσουμε ένα τυχόν δείγμα μεγεθούς n. Επειδή οι μονάδες που μένουν απο τον πληθυσμό είναι N-1 και οι μονάδες που μένουν για να συμπληρώσουμε το δείγμα n-1 υπάρχουν $\binom{N-1}{\eta-1}$ δυνατοί συνδυασμοί για να συμπληρώθει το δείγμα. Άρα, κάθε μονάδα του πληθυσμού* μπορεί να ανήκει σε $\binom{N-1}{\eta-1}$ δείγματα και το κάθε στοιχείο χ_i^j στο παραπάνω άθροισμα εμφανίζεται $\binom{N-1}{\eta-1}$ φορές.

Άρα η (1) γράφεται:

$$E(\hat{\mu}) = \frac{1}{\frac{N!}{\eta!\cdot(N-\eta)!}} \cdot \frac{1}{\eta} \cdot \frac{(N-1)!}{(\eta-1)!\cdot(N-\eta)!} \cdot \sum_{i=1}^{N} \chi_i =$$

$$= \frac{\eta!\cdot(N-\eta)!}{N!} \cdot \frac{(N-1)!}{(\eta-1)!\cdot \eta \cdot (N-\eta)!} \cdot \sum_{i=1}^{\eta} \chi_i = \frac{1}{N} \cdot \sum_{i=1}^{N} \chi_i = \mu$$

όπου οι άνω δείκτες των χ_i^j διαγράφηκαν λόγο σύμπτωσης.

Παρακάτω, αναφέρουμε χωρίς απόδειξη μερικά σημαντικά θεωρήματα στην ΑΤΔ

Θεώρημα 2

Η διακύμανση της δειγματικής μέσης τιμής $\hat{\mu}$ στην ΑΤΔ είναι:

$$Var(\hat{\mu}) = E(\hat{\mu} - \mu)^2 = \frac{\sigma^2}{\eta}\left(1 - \frac{\eta}{N}\right)$$

όπου σ η διακύμανση του πληθυσμού.

Θεώρημα 3

Η διακύμανση του δείγματος

$$\hat{S}^2 = \frac{1}{\eta-1} \cdot \sum_{i=1}^{\eta} (\chi_i - \hat{\mu})^2$$

είναι αμερόληπτος εκτιμητής της διακύμανσης σ^2 του πληθυσμού

$$\sigma^2 = \frac{\sum_{i=1}^{N}\left(\chi_i - \hat{\mu}\right)^2}{N-1}$$

$$E\left(\hat{S}^2\right) = \sigma^2$$

Σημείωση 1

Γνωρίζουμε από την βασική Στατιστική ότι, για ένα δείγμα μεγέθους n απο ΑΠΕΙΡΟ πληθυσμό, η διακύμανση της δειγματικής μέσης τιμής είναι $\frac{\sigma^2}{\eta}$

Το θεώρημα 2 που αναφέρεται σε πληθυσμό μεγέθους N, διαφέρει λίγο στο παραπάνω αποτέλεσμα, δηλ. κατά τον πολλαπλασιαστικό παράγοντα $\left(1-\frac{\eta}{N}\right)$

Όταν λοιπόν ο λόγος $\frac{\eta}{N}$ είναι πολύ μικρός, το μέγεθος του πληθυσμού δεν παίζει ουσιαστικό ρόλο στόν υπολογισμό της διακύμανσης (της δειγματικής μέσης τιμής) του δείγματος.

Θεώρημα 4

Η διακύμανση του εκτιμητή του συνολικού αριθμού του πληθυσμού $\hat{\chi} = N\hat{\mu}$ είναι:

$$\mathrm{Var}\left(\hat{\chi}\right) = E\left(\hat{\chi} - \sum_{i=1}^{N}\chi_i\right)^2 = \frac{N^2\cdot\sigma^2}{\eta}\cdot\frac{\left(N-\eta\right)}{N} = \frac{N^2\cdot\sigma^2}{\eta}\cdot\left(1-\frac{\eta}{N}\right)$$

Πόρισμα 1

Το μέγεθός $N\hat{\mu}$ είναι ένας αμερόληπτος εκτιμητής του αθροίσματος των στοιχείων του πληθυσμού $\sum_{i=1}^{N}\chi_i$

Απόδειξη

$$E\left(N\hat{\mu}\right) = N\cdot E\left(\hat{\mu}\right) = N\mu = N\cdot\frac{\sum_{i=1}^{N}\chi_i}{N} = \sum_{i=1}^{N}\chi_i$$

Πόρισμα 2

Από το θεώρημα 2 και απο το θεώρημα 4 αμέσως προκύπτει ότι, η διακύμανση του συνολικού πληθυσμού είναι N^2 φορές η διακύμανση της δειγματικής μέσης τιμής δηλ.

$$\mathrm{Var}\left(\hat{\chi}\right) = N^2 \mathrm{Var}\left(\hat{\mu}\right)$$

1.5.1.2 Στρωματοποιημένη τυχαία δειγματοληψία (Startified Random Sampling) (Σ.Τ.Δ.)

Γενικά

Στην γενική της μορφή, η Σ.Τ.Δ. αποτελείται από επιμέρους Α.Τ.Δ. σε διάφορα στρώματα που εμφανίζει ένας πληθυσμός. Μία βασική ένδειξη για τη χρήση της Σ.Τ.Δ είναι ο υπό μελέτη πληθυσμός να εμφανίζεται σαν την ένωση επιμέρους πληθυσμών δηλ. $\Pi = \Pi_1 \bigcup \Pi_2 \bigcup \cdots \Pi_\kappa$ όπου Π ο συνολικός πληθυσμός και Πι, ι=1,2,...,κ οι επιμέρους υποπληθυσμοί οι οποίοι είναι ξένοι μεταξύ τους και στο σύνολο τους αποτελούν τον συνολικό πληθυσμό. $\left(\Pi_i \bigcap \Pi_j = \varnothing, i \neq j\right)$ (Stratum)

Εάν τα μεγέθη των ν υποπληθυσμών (στρωμάτων) αυτών είναι $N_1, N_2, \cdots N_k$ και του συνολικού πληθυσμού Ν τότε:

$$N = N_1 + N_2 + \cdots N_k = \sum_{i=1}^{k} N_i \quad (1)$$

Μόλις προσδιοριστούν τα συγκεκριμένα αυτά στρώματα (υποπληθυσμοί) του συνολικού πληθυσμού, εκλέγεται χωριστά σε κάθε στρώμα ένα δείγμα μεγέθους $\eta_i \; i = 1,2,\cdots,k$. Επειδή ακριβώς η εκλογή καθε δείγματος σε κάθε στρώμα γίνεται με Α.Τ.Δ γι'αυτό και η δειγματοληψία αυτή λέγεται Σ.Τυχαία.Δ.

Η Σ.Τ.Δ χρησιμοποιείται όταν:

1. Απαιτείται ο διαχωρισμός του μελετουμενου πληθυσμού σε στρώματα διαφορετικών μεγεθών τα οποία έχουν δημιουργηθεί για άλους λόγους απο

την δειγματοληψία π.χ. Η μελέτη του μέσου όρου κατανάλωσης ενός προιόντος σ'έναν νομό της χώρας μπορεί να γίνει ανά κοινότητα η οποία διοικητική μονάδα χωρισμού και στην Σ.Τ.Δ μπορεί να αποτελέσει το "στρώμα".

2. Υπάρχει προκαθορισμένη ετερογένεια ενός πληθυσμού μεταξύ υποπληθυσμών απο τους οποίους φαίνεται να αποτελείται και οι οποίοι στο εσωτερικό τους είναι ομογενείς.
3. Υπάρχει ανάγκη εκτιμήσεων στατιστικών μεγεθών συγκεκριμένων στρωμάτων του πληθυσμού, π.χ. για την τάξη των δημοσίων υπαλλήλων σ'ένα συγκεκριμένο μεγάλο νησί της χώρας.

Εάν θεωρήσουμε Χ την Τ.Μ. που μελετάμε ορίζουμε τα παρακάτω μεγέθη.

N: μέγεθος πληθυσμού.

Ni ,i=1,2,3...k: μέγεθος (υποπληθυσμού) στρώματος i.

K: πλήθος στρώματων του πληθυσμού.

Ni : μέγεθος δείγματος στρώματος Ni.

X_{ij}: j-οστή τιμή της Τ.Μ.Χ στο δείγμα του στρώματος Ni.

$B_i = \frac{N_i}{N}$: ποσοστό βάρους του στρώματος Ni ως πρός τον συνολικό πληθυσμό.

$\beta_i = \frac{\eta_i}{\eta}$: ποσοστό βάρους του μεγέθους του δείγματος ni του στρώματος Ni ως πρός το μέγεθος n του συνολικού δείγματος.

$\sum_{i=1}^{\kappa} \eta_i$ =n: Μέγεθος συνολικού δείγματος.

$\overline{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i}$: Μέση τιμή της Τ.Μ X_{ij} στο στρώμα Ni

$\overline{\chi}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{\eta_i}$: Μέση τιμή δείγματός n_i στό στό στρώμα Ni

$\sigma_i^2 = \frac{\sum_{j=1}^{Ni} \left(\overline{X}_i - X_{ij}\right)^2}{N_i - 1}$: Διακύμανση της Τ.Μ X_{ij} στο στρώμα Ni

Η μέση τιμή του πληθυσμού μ μπορεί να εκτιμηθεί με 2 ειδών εκτιμητές **[7]**

1. Τον αναλογικό εκτιμητή ο οποίος εκτιμά ουσιαστικά την ανά στρώμα μέση τιμή του πληθυσμού και είναι

$$\overline{X}_{\sigma\tau} = \frac{\sum_{i=1}^{K} N_i \cdot \overline{X}_i}{N} = \frac{\sum_{i=1}^{K} N_i \cdot \left[\sum_{j=1}^{\eta_i} \frac{X_{ij}}{\eta_i}\right]}{N} = \sum_{i=1}^{K} B_i \cdot \overline{X}_i \quad (2)$$

2. Τον γενικό εκτιμητή της μέσης τιμής του πληθυσμού μ που είναι ο δειγματικός μέσος $\overline{X}$

$$\overline{X} = \frac{\sum_{i=1}^{K} \eta_i \cdot \overline{\chi}_i}{\eta} = \sum_{i=1}^{K} \beta_i \cdot \overline{\chi}_i \quad (3)$$

Παρατηρήσεις

1η. Υπάρχει σύμπτωση του αναλογικού με τον γενικό εκτιμητη όταν ισχύει η ισότητα:

$$B_i = \beta_i \Rightarrow \frac{N_i}{N} = \frac{\eta_i}{\eta} \Rightarrow \frac{\eta_i}{N_i} = \frac{\eta}{N} \quad (4)$$

δηλαδή σύμφωνα με την τελευταία ισότητα όταν η αναλογία του μεγέθους του δείγματος ως προς το μέγεθος του στρώματος που ανήκει, είναι ίδια σε κάθε στρώμα και ίση με $\frac{\eta}{N}$.

2η. το αντίστοιχο δείγμα με την Α.Τ.Δ. στην Σ.Τ.Δ. είναι αυτό που προκύπτει απο την συνένωση των επι μέρους δειγμάτων n_i κέθε στρώματος και έχει μέγεθος $\eta = \sum_{i=1}^{K} \eta_i$

Το κύριο φυσικά ερώτημα που τίθεται στην Σ.Τ.Δ. είναι πόσο μεγάλο πρέπει να είναι το μέγεθος του συνολικού δείγματος που είναι συνάρτηση του πόσο μεγάλα πρέπει να είναι τα δείγματα μεγεθών n_i σε κάθε στρώμα;

Την απάντηση εν μέρει στο παραπάνω ερώτημα δίνει η σχέση (4) που μας ορίζει απο την αρχή της δειγματοληψίας τον λόγο $\frac{\eta_i}{\eta}$ (που έιναι ίσος με τον σταθερό λογό $\frac{N_i}{N}$ εφ'όσον είναι καθορισμένα τα στρώματα N_i).

Αλλά πάλι, πρέπει να καθοριστεί το μέγεθος n η καθ'ένα απο τα n_i. Σαν παράδειγμα, έστω ότι έχουμε έναν πληθυσμό N=1000 και έστω ότι το τέταρτο στρώμα του πληθυσμού έχει μέγεθος $N_4 = 100$. Έτσι έχουμε $\frac{\eta_4}{\eta} = \frac{1}{10}$ πράγμα που μπορεί να σημαίνει ότι $\eta_4 = 5$ και $\eta = 50$ ή ότι $\eta_4 = 10$ και $\eta = 100$ ή οτιδήποτε αριθμοί διατηρούν την αναλογία αυτή.

Άρα, ο εκ των προτέρων προσδιορισμός του μεγέθους του συνολικού δείγματος n προσδιορίζει και τα εκάστοτε μεγεθη καθενός απο τα μερικά δείγματα n_i.

Παρακάτω αναφέρουμε μερικά θεωρήματα της Σ.Τ.Δ. χωρίς απόδειξη.

Θεώρημα 1

Εάν για κάθε στρώμα του πληθυσμού N_i η μέση δειγματική τιμή $\overline{\chi}_i$ είναι αμερόληπτος εκτιμητής της μέσης τιμής του στώματος N_i δηλαδή της $\overline{X}_i$, τότε ο αναλογικός εκτιμητής $\chi_{\sigma\tau}$ είναι αμερόληπτος εκτιμητής της μέσης τιμής του πληθυσμού μ.

Λόγο σχετικής ευκολίας, αναφέρουμε παρακάτω την απόδειξη του θεωρήματος 1

Η μέση τιμή μ του πληθυσμού είναι

$$\mu = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i}\chi_{ij}}{N} = \frac{\sum_{i=1}^{K} N_i \cdot \overline{X}_i}{N} \quad (5)$$

γιατί η μέση τιμή της τυχαίας μεταβλητής Xij στο στρώμα i είναι

$$\overline{X}_i = \frac{\sum_{j=1}^{N_i} X_{ij}}{N_i}$$

οπότε η (5) γράφεται $\mu = \sum_{i=1}^{K} B_i \cdot \overline{X}_i$ και επειδή δεχθήκαμε ότι η μέση δειγματική τιμή $\overline{\chi}_i$ είναι αμερόληπτος εκτιμητής της μέσης τιμής $\overline{X}_i$ του στρώματος Ni δηλαδή $E(\overline{\chi}_i) = \overline{X}_i$ θα έχουμε

$$\mu = \sum_{i=1}^{K} B_i \cdot E(\overline{\chi}_i) = E\left(\sum_{i=1}^{K} B_i \cdot \overline{\chi}_i \right) = E(\overline{\chi}_{\sigma\tau})$$

Θεώρημα 2

Για μία ανεξάρτητη εκλογή δειγμάτων στα διάφορα στρώματα και αν ισχύει το Θεωρ. 1 τότε

$$\mathrm{Var}(\overline{\chi}_{\sigma\tau}) = \sum_{i=1}^{K} B_i^2 \cdot \mathrm{Var}(\overline{\chi}_i)$$

όπου $\mathrm{Var}(\overline{\chi}_{\sigma\tau})$ η ανά στρώμα διακύμανση του αναλογικού εκτιμητή $\overline{\chi}_{\sigma\tau}$ και $\mathrm{Var}(\overline{\chi}_i)$ η διακύμανση της δειγματικής μέσης τιμής σε διάφορα δείγματα του ίδιου στρώματος i.

Θεώρημα 3

Στην Σ.Τ.Δ. η διακύμανση του εκτιμητή $\overline{\chi}_{\sigma\tau}$ είναι:

$$\mathrm{Var}(\overline{\chi}_{\sigma\tau}) = \frac{1}{N^2} \cdot \sum_{i=1}^{K} N_i \cdot (N_i - \eta_i) \cdot \frac{\sigma_i^2}{\eta_i} = \sum_{i=1}^{K} B_i^2 \cdot \frac{\sigma_i^2}{\eta_i} \left(1 - \frac{\eta_i}{N_i} \right)$$

α) Εάν $\frac{\eta_i}{N_i} \mapsto 0$ τότε $\mathrm{Var}(\overline{\chi}_{\sigma\tau}) = \sum_{i=1}^{K} B_i^2 \cdot \frac{\sigma_i^2}{\eta_i}$

β) Εάν $\frac{\eta_i}{N_i} \mapsto \frac{\eta}{N}$ τότε $\mathrm{Var}(\overline{\chi}_{\sigma\tau}) = \frac{1}{\eta} \cdot \left(1 - \frac{\eta}{N} \right) \cdot \sum_{i=1}^{K} B_i^2 \cdot \frac{\sigma_i^2}{\eta_i}$

γ) Εάν η δειγματοληψία είναι αναλογική και η διακύμανση όλων των στρωμάτων ίσες μεταξύ τους π.χ. σ_k τότε

$$\mathrm{Var}(\overline{\chi}_{\sigma\tau}) = \frac{1}{\eta} \cdot \left(1 - \frac{\eta}{N} \right) \cdot \sigma_\kappa^2$$

Θεώρημα 4

Η διακύμανση $\mathrm{Var}\left(\hat{\mathrm{X}}_{\sigma\tau}\right)$ του εκτιμητή του πληθυσμού του συνολικού αριθμού των ατόμων του πληθυσμού $\hat{\mathrm{X}}_{\sigma\tau} = \mathrm{N} \cdot \overline{\chi}_{\sigma\tau}$ είναι:

$$\mathrm{Var}\left(\hat{\mathrm{X}}_{\sigma\tau}\right) = \sum_{i=1}^{K} \mathrm{N}_i \cdot \left(\mathrm{N}_i - \eta_i\right) \frac{\sigma_i^2}{\eta_i} = \mathrm{N}^2 \cdot \mathrm{Var}\left(\overline{\chi}_{\sigma\tau}\right)$$

Για να δούμε καλύτερα την διαφορά των εκτιμητών μεταξύ Α.Τ.Δ. και Σ.Τ.Δ. ο Cochran **[βιβλιογρ. 1]** παραθέτει το πιο κάτω παράδειγμα:

Ο παρακάτω πίνακας, δείχνει το 1930, τον αριθμό των κατοίκων σε χιλιάδες από 64 μεγάλες πόλεις των Η.Π.Α. Το σύνολο χωρίστηκε σε 2 στρώματα, το πρώτο με 16 μεγάλες πόλεις και το δεύτερο με τις υπόλοιπές 48.

Από ένα δείγμα 24 πόλεων θα εκτιμηθεί ο συνολικός πληθυσμός των 64 πόλεων. Ζητείται να βρεθεί η τυπική απόκλιση του συνολικού πληθυσμού (η Τ.Μ του παραδείγματος)

α) Με Α.Τ.Δ.

β) Με Σ.Τ.Δ.με αναλογικό εκτιμητή και,

γ) Με Σ.Τ.Δ. που το δείγμα αποτελείται απο 12 πόλεις απο κάθε στρώμα.

Για όλον τον πληθυσμό, το 1930 βρέθηκε, ο συνολικός πληθυσμός $\sum_{i=1}^{N} \chi_i = 19586$ και η διακύμανση $\sigma^2 = 52448$

Οι εκτιμητές για τις τρείς περιπτώσεις που ζητούνται θα συμβολίζονται αντίστοιχα με $\hat{\mathrm{X}}_1, \hat{\mathrm{X}}_2, \hat{\mathrm{X}}_3$.

Στρώμα 1	Στρώμα2		
900	364	209	113
822	317	183	115
781	328	163	123
805	302	253	154
670	288	232	140
1238	291	260	119
573	253	201	130
634	291	147	127
578	308	292	100
487	272	164	107
442	284	143	114
451	255	169	111
459	270	139	163
464	214	170	116
400	195	150	122
366	260	143	134

1) Με Α.Τ.Δ.

$$\mathrm{Var}\left(\hat{X}_1\right)=\frac{N^2\cdot\sigma^2}{\eta}\cdot\frac{N-\eta}{N}=\frac{64^2\cdot 52448}{24}\cdot\frac{40}{64}=5594453\xrightarrow{\text{Θεωρ.4}}\sigma\left(\hat{X}_1\right)=\sqrt{5594453}=2365.26$$

2) Με Σ.Τ.Δ

Οι διακυμάνσεις των δύο στρωμάτων είναι :

$\sigma_1^2=53843$ και $\sigma_2^2=5584$ (τύπος σ_i^2 σελ .86549-67 για κάθε στρώμα)

Επειδή για την μέθοδο των αναλογικών εκτιμητών πρέπει: $\frac{\eta_1}{N_1}=\frac{\eta_2}{N_2}=\frac{\eta}{N}$

είναι $N=64,\eta=24,$ και $N_1=1,N_2=48$. Οπότε τα η_1,η_2 υπολογίζονται ως εξής:

$\frac{\eta_1}{16}=\frac{24}{64}\Rightarrow\eta_1=6$ ομοίως: $\frac{\eta_2}{48}=\frac{24}{64}\Rightarrow\eta_2=18$

Από το Πόρισμα 2 της Α.Τ.Δ. και την προυπόθεση $\frac{\eta_1}{\eta}=\frac{N_1}{N}$ (θεωρημα 3 σελ

21) έχουμε ότι $\mathrm{Var}\left(\hat{X}_2\right)=N^2\cdot\mathrm{Var}\left(X_{\sigma\tau}\right)$ άρα

$$\mathrm{Var}\left(X_2\right)=N^2\cdot\left[\frac{1}{\eta}\cdot\left(1-\frac{\eta}{N}\right)\cdot\sum_{i=1}^{2}\frac{N_i}{N}\cdot\sigma_i^2\right]=$$

$$= N^2 \cdot \frac{N-\eta}{\eta} \cdot \frac{1}{N^2} \cdot \sum_{i=1}^{2} N_i \cdot \sigma_i^2 = \frac{N-\eta}{\eta} \cdot \sum_{i=1}^{2} N_i \cdot \sigma_i^2 = \frac{60-24}{24} \cdot \left[N_1 \cdot \sigma_1^2 + N_2 \cdot \sigma_2^2 \right] =$$

$$= \frac{40}{24} \cdot [16 \cdot 53843 + 48 \cdot 5581] = \frac{3}{5} \cdot [861488 + 267888] = 1882293{,}3 \text{ και}$$

$$\sigma\left(\hat{X}_1\right) = \sqrt{\text{Var}\left(\hat{X}_1\right)} = 1371{,}966 \approx 1372$$

3) Εάν $\eta_1 = \eta_2 = 12$ χρησιμοποιώντας το θεώρημα 4 έχουμε:

$$\text{Var}\left(\hat{X}_3\right) = \frac{1}{12} \cdot \sum_{i=1}^{2} N_i \cdot \left(N_i - \eta_i\right) \cdot \sigma_i^2 = \frac{1}{12} \cdot \left[N_1 \cdot \left(N_1 - \eta_1\right) \cdot \sigma_1^2 + N_2 \cdot \left(N_2 - \eta_2\right) \cdot \sigma_2^2 \right] =$$

$$= \frac{1}{12} \cdot [16 \cdot (16-12) \cdot 53843 + 48 \cdot (48-12) \cdot 5581] = 1090827 \text{ και}$$

$$\sigma\left(\hat{X}_3\right) = \sqrt{1090827} = 1044{,}426 \approx 1044$$

Τελικά παρατηρούμε ότι: $\sigma\left(\hat{X}_3\right) < \sigma\left(\hat{X}_2\right) < \sigma\left(\hat{X}_1\right)$

1.5.1.3 Συστηματική δειγματοληψία [Systematic Sampling] (Σ.Δ.)

Η εκλογή δείγματος είναι ευκολότερη, ταχύτερη και με μικρότερη πιθανότητα σφάλματος απο εκείνη της τυχαίας δειγματοληψίας. Ειδικά, όταν η Σ.Δ διενεργείται σε κάποιο χωρικό πλαίσιο όπως δειγματοληψίες εδάφους χρήσης γης ή πληθυσμιακές στον χώρο είναι πιο εύκολη απο την Τ.Δ λόγω της χωρικής περιοδικότητας των στοιχείων εκλογής του δείγματος.

Έστω ότι έχουμε να εκλέξουμε ένα δείγμα n στοιχείων απο έναν πληθυσμό N στοιχείων αριθμημένων απο 1 έως N.

Για να είναι ομοιόμορφα κατανεμημένο και να πιάνει όλον τον πληθυσμό, το δείγμα αρκεί να χωρίσουμε τον πλήθυσμό σε n υποσύνολα κάθε ένα απο τα οποία θα περιέχει K=N/n στοιχεία.

Εάν K: ακέραιος, τότε δημιουργούνται πλήρεις ομάδες απο K στοιχεία και όλα τα δυνατά δείγματα που έχουν ένα στοιχείο απο κάθε ομάδα είναι μεγέθους n.

Εάν $N < \eta \cdot K$ όπου $K = \left[\frac{N}{\eta}\right] + 1$, με $\left[\frac{N}{\eta}\right]$ το ακέραιο μέρος οτυ αριθμού N/n, δημιουργούμε πάλι n ομάδες απο τις οποίες η n-οστή είναι ελλιπής. Κατά συνέπεια, μερικά από τα δείγματα θα περιέχουν n-1 στοιχεία, άρα έχουμε κάποια ομοιογένεια ως προς το μέγεθος των δειγμάτων. Η ανομοιογένεια αυτή εξαλείφεται σύμφωνα με τον κυκλικό νομό που περιγράφεται παρακάτω στο 2ο παράδειγμα.

1ο Παράδειγμα

Έστω ότι έχουμε να εκλέξουμε δείγματα n=5 ατόμων απο έναν πληθυσμό N=35 ατόμων έχουμε $K = \frac{N}{\eta} = \frac{35}{5} = 7$ άρα ο πληθυσμός θα χωριστεί σε 5 ομάδες των 7 ατόμων σύμφωνα με τον παρακάτω πίνακα:

	Δείγμα 1	Δείγμα 2	Δείγμα 3	Δείγμα 4	Δείγμα 5	Δείγμα 6	Δείγμα 7
Ομάδα 1	1	2	3	4	5	6	7
Ομάδα 2	8	9	10	11	12	13	14
Ομάδα 3	15	16	17	18	19	20	21
Ομάδα 4	22	23	24	25	26	27	28
Ομάδα 5	29	30	31	32	33	34	35

Παρατηρούμε ότι, ο δυνατός αριθμός των δειγμάτων είναι ίσος με τον αριθμό τον ατόμων κάθε ομάδας που δημιουργήθηκε.

2ο Παράδειγμα

Έστω ότι έχουμε πάλι να εκλέξουμε δείγματα n=5 απο έναν πληθυσμο N=32 ατόμων, εδώ βλέπουμε ότι $\frac{N}{\eta} = \frac{35}{5} = 6{,}4$ και $\left[\frac{32}{5}\right] = 6$ και $K = 6 + 1 = 7$ άρα, δημιουργούμε πάλι 5 ομάδες των 7 ατόμων εκ των οποίων η 5η είναι ελλιπής.

Για να βρούμε πόσα απο αυτά θα είναι άρτια (με 5 στοιχεία), βρίσκουμε το υπόλοιπο της διαίρεσης του N=32 με το K=7 δηλ. $N = 4 \cdot 7 + 4$, άρα 4. Το υπόλοιπο

της αφαίρεσης του 4 από το K=7 μας δίνει τον αριθμό των ελλειπτικών δειγμάτων. Αυτά φαίνονται στον παρακάτω πίνακα:

	Δείγμα 1	Δείγμα 2	Δείγμα 3	Δείγμα 4	Δείγμα 5	Δείγμα 6	Δείγμα 7
Ομάδα 1	1	2	3	4	5	6	7
Ομάδα 2	8	9	10	11	12	13	14
Ομάδα 3	15	16	17	18	19	20	21
Ομάδα 4	22	23	24	25	26	27	28
Ομάδα 5	29	30	31	32			

Το σημαντικότερο βέβαια στην παραπάνω περίπτωση που το μέγεθος του πληθυσμού δεν είναι ακέραιο πολλαπλάσιο του μεγέθους του πληθυσμού των δειγμάτων που θέλουμε, είναι η συμπλήρωση των ελλιπών δειγμάτων.

Μία μέθοδος που προτείνεται για την λύση του προβλήματος αυτού είναι:
Εκλέγουμε έναν τυχαίο αριθμό απο το 1 έως N. Με αφετηρία τον αριθμό αυτό και βήμα ίσο με K επιλέγουμε n αριθμούς απο τους N, τοποθετημένους σε κύκλο έτσι ώστε ο επόμενος του N να είναι ο 1. Δηλαδή στο συγκεκριμένο παράδειγμα εάν εκλεγεί τυχαία ο αριθμός 29 τότε το 5-μελές δείγμα μας θα είναι: 29, 2, 7, 12, 17
Η γενίκευση του πίνακα του πρώτου παραδείγματος στην περίπτωση που ο πληθυσμός $N = \eta \cdot K$ μπορεί να χωριστεί σε K διαφορετικά δείγματα μεγέθους n το καθ'ένα (δηλαδή είναι ακέραιο πολλαπλάσιο του μεγέθους του δείγματος n) δίνεται απο τον παρακάτω πίνακα (Cochran).

$\frac{\text{Αριθμος\_δειγματος}\longrightarrow}{\text{Ομαδες\_απο\_κ\_στοιχεια}}$	1	…	ι	…	κ
1η_Ομαδα	x_1	…	x_i	…	x_κ
2η_Ομαδα	$x_{\kappa+1}$	…	⋮		⋮
⋮	⋮		⋮		⋮
n − οστη_ομαδα	$x_{(n-1)\kappa+1}$	…	$x_{(n-1)\kappa+i}$	…	$x_{n\kappa}$
Μεσες_Τιμες	$\overline{x}_1$	…	$\overline{x}_i$	…	$\overline{x}_\kappa$

Εάν συμβολίσουμε με $\hat{X}_{συσ}$ την δειγματική μέση τιμή στην συστηματική δειγματοληψία και αν $N = \eta \cdot K$ τότε ισχύουν τα εξείς θεωρήματα.

Θεώρημα 1

Η $\hat{X}_{συσ}$ είναι ένας αμερόληπτος εκτιμητής της μέσης τιμής μ του πληθυσμού

Θεώρήμα 2

Η διακύμανση της δειγματικής μέσης τιμής $\hat{X}_{συσ}$ στην συστηματική δειγματοληψία είναι:

$$\mathrm{Var}\left(\hat{X}_{συσ}\right) = \frac{N-1}{N}\cdot\sigma^2 - \frac{K\cdot(\eta-1)}{N}\cdot\sigma_\delta^2$$

$$\text{όπου } \sigma_\delta^2 = \frac{1}{K\cdot(\eta-1)}\sum_{i=1}^{K}\sum_{j=1}^{\eta}\left(\chi_{ij}-\hat{\chi}_i\right)^2$$

είναι η ενδο-δειγματική διακύμανση και σ^2 η διακύμανση του πληθυσμού.

Επειδή όμως στην Α.Τ.Δ. ειχαμε (Θεώρημα 2ο)

$$\mathrm{Var}\left(\hat{\mu}\right) = \frac{\sigma^2}{\eta}\cdot\left(1-\frac{\eta}{N}\right)$$

εύκολα προκύπτει το παρακάτω πόρισμα

Πόρισμα

Η δειγματική μέση τιμή στην Σ.Δ είναι πιο ακριβής εκτιμητής της μέσης τιμής του πληθυσμού, απο την δειγματική μέση τιμή της Α.Τ.Δ. εάν $\sigma_\delta^2 > \sigma^2$

Το παραπάνω πόρισμα μας οδηγεί στο παρακάτω συμπέρασμα:
Η Σ.Δ προτιμάται απο την Α.Τ.Δ όταν στο εσωτερικό των δειγμάτων (Στην Σ.Δ.) υπάρχει μεγάλη ετερογένεια μεταξύ των στοιχείων τους.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Cochran G. William 'Sampling Techniques' John Willey & Sons, 1977.

[2] Κάτος Β. Αναστάσιος 'Στατιστική', Παρατηρητής Θεσ/νική1986.

[3] Κουτσόπουλος Κωστής 'Γεωγραφία: μεθοδολογία και μέθοδοι ανάλυσης χώρου' Εκδόσεις Συμμετρία, Αθήνα 1990.

[4] Λιάκη Π. Ιωάννου 'Στοιχεία Στατιστικής' Τόμος Ι Β΄ έκδοση Εκδόσεις Ζήτη Θεσ/νική 1976.

[5] Περάκης Κ., "Εισαγωγική Στατιστική για Χωροτάκτες", Πανεπιστημιακές Σημειώσεις, Πανεπιστήμιο Θεσσαλίας, ΤΜΧΠΑ 1997.

[6] Yamane Taro 'Statistics: An Introductory Analysis' Harper & Row and John Weatherhill Publishers, New York, 1964.

[7] Φαρμάκης Νίκος 'Εισαγωγή στη δειγματοληψία' Εκδόσεις Κ. Χριστοδουλίδη, Θεσσαλονίκη 1992.

[8] Τσέρπε Ν., Αλεβίζος Φ.,'Εισαγωγή στην θεωρία δειγματοληψίας' Εκδόσεις Παν/μιου Πατρών, Πάτρα 1993.

ΚΕΦΑΛΑΙΟ 2

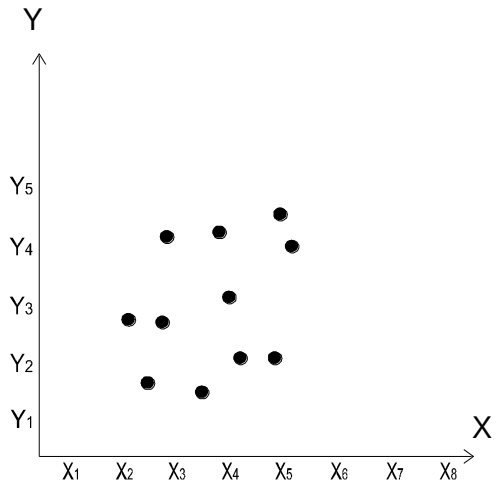
ΠΑΛΙΝΔΡΟΜΗΣΗ

1. Το μοντέλο της παλινδρόμησης

Σπάνια θα μπορούσαμε να σκεφτούμε κάποια μεταβλητή που να μην εξαρτάται κατά κάποιον τρόπο από μία ή περισσότερες άλλες.

Στην περίπτωση τυχαίων μεταβλητών (Τ.Μ.), θεωρούμε ότι η μία από τις 2 Τ.Μ. παίζει τον ρόλο της ανεξάρτητης και η άλλη της εξαρτημένης μεταβλητής. Θα συμβολίζουμε στο εξής Χ την ανεξάρτητη και Υ την εξαρτημένη μεταβλητή.

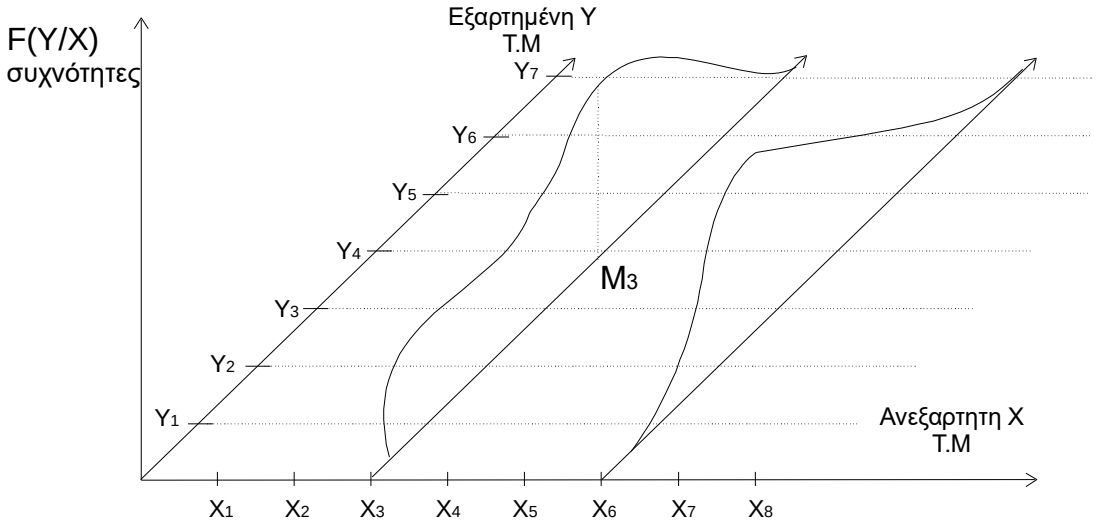
Η απλή περίπτωση του σε κάθε τιμή της ανεξάρτητης μεταβλητής αντιστοιχούν μία ή περισσότερες τιμές της εξαρτημένης μεταβλητής, αλλά εμφανιζόμενες η κάθε μία από αυτές μία φορά (δηλαδή με συχνότητα ίση με 1), μπορεί να παρασταθεί στο επίπεδο ΧΥ σαν ένα νέφος σημείων (σχ. 1)



Σχήμα 1

Όμως, στην πραγματικότητα, σε κάθε τιμή της ανεξάρτητης μεταβλητής, αντιστοιχεί μια κατανομή τιμών της εξαρτημένης (κάθε μία από τις τιμές της εξαρτημένης μεταβλητής εμφανίζεται περισσότερες από μία φορές)

Έτσι, αν υποθέσουμε ότι, εκτός από το επίπεδο ΧΥ έχουμε και μια τρίτη διάσταση που εμφανίζει τις συχνότητες εμφάνισης των τιμών της μεταβλητής Υ για κάθε μία από τις τιμές της μεταβλητής Χ, στην πράξη θα έχουμε διάφορες κατανομές των τιμών της Υ για κάθε μία από τις τιμές της ανεξάρτητης μεταβλητής Χ.



Ωστόσο, για την μελέτη των μεταβολών των δύο μεταβλητών συγχρόνως, πρέπει να γίνουν ορισμένες υποθέσεις που απλοποιούν το δύσκολο πρόβλημα των τυχαίων κατανομών:

1ο. Όλες οι Υ κατανομές έχουν την ίδια διακύμανση σ^2 (δηλαδή για κάθε τιμή της Χ).

2ο. Όλες οι μέσες τιμές των Υ κατανομών βρίσκονται στην ίδια ευθεία γραμμή στο επίπεδο ΧΥ την $E(Y) = \mu = \alpha + \beta\chi$, όπου οι συντελεστές α και β εκτιμούνται από το δείγμα που λαμβάνεται.

3ο. Οι τυχαίες μεταβλητές Υι είναι ανεξάρτητες μεταξύ τους. **[βιβλιογρ. 1].**

Έτσι, συνοπτικά υποθέτουμε ότι, οι Υι είναι ανεξάρτητες μεταξύ τους με μέση τιμή $\alpha + \beta\chi$ και διακύμανση σ^2.

Εαν θεωρήσουμε ως e_i την απόκλιση της Y_i από την αναμενόμενη τιμή της (μαθ. ελπίδα) E(Y) δηλαδή $Y_i = \alpha + \beta X_i + e_i$ **[1],** η παραπάνω υπόθεση γράφεται $\mu_{ei} = 0$ και $Var(e_i) = \sigma^2$.

2.2 Προσδιορισμός των συντελεστών α και β της ευθείας παλινδρόμησης

Εαν, όπως και πριν, ονομάσουμε e_i τις αποκλίσεις των προβολών των τιμών της Y_i στο επίπεδο XY, από την πραγματική ευθεία παλινδρόμησης που ελαχιστοποιεί το άθροισμα των τετραγώνων των σφαλμάτων αυτών από την εξίσωση **(1)** έχουμε $e_i = Y_i - \alpha - \beta X_i$ και υψώνοντας στο τετράγωνο και αθροίζοντας ως προς i ονομαζουμε

$$S = \sum_{i=1}^{n} e_i^2 = \sum \left(Y_i - a - \beta x_i \right)^2$$

και θεωρώντας το S συνάρτηση των μεγεθών α και β, δηλαδή S(α,β).

Παραγωγίζουμε, μηδενίζουμε τις πρώτες παραγώγους και λύνοντας το γραμμικό σύστημα προκύπτουν οι τιμές των α και β **(2)**

Αναλυτικότερα:

$$\frac{\partial S}{\partial a} = \frac{\partial}{\partial a} \sum_{i-1}^{n} \left(Y_i^2 + a^2 + \beta^2 X_i^2 - 2 Y_i a - 2 Y_i \beta X_i + 2 a \beta X_i \right) =$$

και

$$= -2 \sum \left(Y_i - a - \beta X_i \right)$$

$$\frac{\partial^2 S}{\partial a^2} = -2 \sum_{i=1}^{n} (-1) = 2n > 0$$

$$\frac{\partial S}{\partial b} = -2 \sum \left(Y_i - a - \beta X_i \right) \cdot X_i \quad \text{και}$$

$$\frac{\partial^2 S}{\partial b^2} = -2 \sum \left(- X_i^2 \right) = 2 \sum X_i^2 > 0$$

και $\quad \frac{\partial^2 S}{\partial a \partial b} = -2 \sum \left(- X_i \right) = 2 \sum X_i \quad$ και

$$\frac{\partial^2 S}{\partial a^2}\cdot\frac{\partial^2 S}{\partial b^2}=\left(\frac{\partial^2 S}{\partial a\partial b}\right)^2=2n\cdot 2\sum X_i^2-4\left(\sum X_i\right)^2=4n\sum X_i^2-4\left(\sum X_i\right)^2=$$

$$=4n\sum\left(X_i-\mu_x\right)^2>0$$

Έτσι, επειδή οι τρεις παραπάνω ποσότητες είναι θετικές, το σύστημα $\frac{\partial S}{\partial a}=0$ και $\frac{\partial S}{\partial \beta}=0$ **(2)** μας ελαχιστοποιεί την συνάρτηση S, δηλαδή την ευθεία των ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ γι αυτό και η μέθοδος αυτή λέγεται **μέθοδος των ελάχιστων τετραγώνων (least squares)**.

Η λύση του συστήματος **(2)**, δηλαδή των εξισώσεων

$$-2\sum\left(\mathrm{Y_i}-\alpha-\beta\mathrm{x_i}\right)=0\ \Big\}$$

ή

$$-2\sum\left(\mathrm{Y_\iota}-\alpha-\beta\mathrm{x_\iota}\right)\cdot\mathrm{x_\iota}=0\ \Big\}$$

$$\sum\mathrm{Y_i}=\alpha\mathrm{n}+\beta\sum\mathrm{X_I}\quad\textbf{(3)}$$

και

$$\sum\mathrm{Y_iX_i}=\alpha\sum\mathrm{X_i}+\beta\sum\mathrm{X_i^2}\quad\textbf{(4)}$$

μας δίνει $\beta=\dfrac{\mathrm{n}\sum\mathrm{X_iY_i}-\sum\mathrm{X_i}\sum\mathrm{Y_i}}{\mathrm{n}\sum\mathrm{X_i^2}-\left(\sum\mathrm{X_i}\right)^2}=\dfrac{\mathrm{Cov(X,Y)}}{\mathrm{Var(X)}}$ **(5)**

και $\alpha=\dfrac{\sum\mathrm{Y_i}}{\mathrm{n}}-\beta\dfrac{\sum\mathrm{X_i}}{\mathrm{n}}=\mu_\gamma-\beta\mu_\mathrm{x}$ **(6)**

Ένας πιο εύκολος τρόπος για να προσδιορίσουμε τις τιμές των συντελεστών α και β είναι να αθροίσουμε για όλα τα ι την εξίσωση $\mathrm{Y_i}=\alpha+\beta\mathrm{X_i}$, οπότε, καταλήγουμε στην εξίσωση **(3),** δηλαδή $\sum\mathrm{Y_i}=\mathrm{n}\alpha+\beta\sum\mathrm{X_i}$ και για την εξίσωση **(4)** να πολ/σουμε και τα 2 μέλη της $\mathrm{Y_i}=\alpha+\beta\mathrm{X_i}$ επί $\mathrm{X_i}$ και μετά να αθροίσουμε για όλα τα ι.

Εάν τώρα στην εξίσωση **(3)** διαιρέσουμε και τα δύο μέλη της με n έχουμε των εξίσωση

$$\overline{Y} = \alpha + \beta\overline{X} \qquad \textbf{(7)}$$

Έτσι, αν αφαιρέσουμε κατά μέλη από την εξίσωση της ευθείας των ελάχιστων τετραγώνων την **(7),** βρίσκουμε με την βοήθεια της **(6)** ότι

$$Y - \overline{Y} = \frac{\sum\left(X - \overline{X}\right)\cdot\left(Y - \overline{Y}\right)}{\sum\left(X - \overline{X}\right)^2}\cdot\left(X - \overline{X}\right) \quad \textbf{(8)}$$

Έτσι, η σταθερά β είναι η ΚΛΙΣΗ της ευθείας των ελάχιστων τετραγώνων και ακόμη, η ευθεία αυτή περνάει από το σημείο $\left(\overline{X}, \overline{Y}\right)$ που λέγεται και κέντρο βάρους των δεδομένων.

Μια μεταφορά των αρχικών αξόνων **[βιβλιογρ.4]** Χ και Ψ, δηλαδή ένας ΓΡΑΜΜΙΚΟΣ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ της μορφής $X = x' + \kappa$ και $Y = \psi' + \lambda$ δεν μεταβάλλει την μορφή της κλίσης β, που με τις νέες μεταβλητές γράφεται

$$\beta = \frac{\sum\left(X' - \overline{X'}\right)\cdot\left(Y' - \overline{Y'}\right)}{\sum\left(X' - \overline{X}\right)^2}$$

2.3 Διασπορές των μεταβλητών Χ και Ψ και συντελεστής συσχέτισης

Ακριβώς με τον ίδιο τρόπο που υπολογίστηκαν οι συντελεστές α και β της ευθείας παλινδρόμησης της μεταβλητής Y ως προς την X, υπολογίζονται και της ευθείας παλινδρόμησης της X ως προς την Y.

Έτσι, έχουμε 2 ευθείες παλινδρόμησης

της Ψ ως προς την X: $Y = \alpha + \beta X$

της X ως προς την Y: $X = \gamma + \delta Y$

που στην μορφή της εξίσωσης **(8)** γράφονται:

$$Y - \overline{Y} = \frac{\sum\left(X - \overline{X}\right)\cdot\left(Y - \overline{Y}\right)}{\sum\left(X - \overline{X}\right)^2}\cdot\left(X - \overline{X}\right) \quad \textbf{(9)}$$

$$X - \overline{X} = \frac{\sum\left(X - \overline{X}\right)\cdot\left(Y - \overline{Y}\right)}{\sum\left(Y - \overline{Y}\right)^2}\cdot\left(Y - \overline{Y}\right) \quad \textbf{(10)}$$

Εάν τώρα, ορίσουμε, όπως είναι γνωστό, ως δειγματική διασπορά του Χ $S_x^2 = \frac{\sum (X - \overline{X})^2}{n}$ και του Ψ $S_\psi^2 = \frac{\sum (Y - \overline{Y})^2}{n}$ δειγματική συνδιασπορά $S_{x\psi} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n}$ και τέλος, ως **δειγματικό συντελεστή συσχέτισης r** το μέγεθος $r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2} = \frac{S_{x\psi}}{S_x S_\psi}$

οι εξισώσεις (9) και (10) γράφονται

$Y - \overline{Y} = \frac{S_{x\psi}}{S_x^2}(X - \overline{X})$ και $X - \overline{X} = \frac{S_{x\psi}}{S_\psi^2}(Y - \overline{Y})$ ή ακόμη

$\frac{Y - \overline{Y}}{S_\psi} = r \cdot \frac{X - \overline{X}}{S_x}$ και $\frac{X - \overline{X}}{S_x} = r \cdot \frac{Y - \overline{Y}}{S_\psi}$

Τέλος, ενδιαφέρουσα είναι η σχέση που συνδέει τους συντελεστές κλίσης των ευθειών παλινδρόμησης της Ψ ως προς την Χ, δηλαδή το β με της Χ ως προς Ψ, δηλαδή το δ είναι **$\beta\delta = r^2$.**

2.4 Παραβολική παλινδρόμηση

Εδώ, η σχέση που μας δίνει την καλύτερη προσέγγιση στο νέφος των σημείων μας είναι μια παραβολή της μορφής

$$Y = \alpha + \beta x + \gamma x^2 \quad \textbf{(1)}$$

Αθροίζοντας την (1) για όλα τα ι έχουμε

$$\sum^n Y_i = n\alpha + \beta \sum X_i + \gamma \sum X_i^2 \quad \textbf{(2)}$$

και πολαπλασιάζοντας την (1) επί χ και αθροίζοντας για όλα τα ι

$$\sum X_i Y_i = \alpha \sum X_i + \beta \sum X_i^2 + \gamma \sum X_i^3 \quad \textbf{(3)}$$

και πολαπλασιάζοντας την **(1)** επί X^2 και αθροίζοντας για όλα τα ι

$$\sum X_i^2 Y_i = \alpha \sum X_i^2 + \beta \sum X_i^3 + \gamma \sum X_i^4 \quad \textbf{(4)}$$

δημιουργείται ένα γραμμικό σύστημα 3 εξισώσεων, των **(2), (3)** και **(4)** με τρεις αγνώστους τα α, β και γ που επιλυόμενο μας δίνει την παραβολή που προσεγγίζει καλύτερα τα δεδομένα μας.

2.5 Εφαρμογή της εκθετικής παλινδρόμησης

Ο R. Baxter **[2]** δίνει ένα ωραίο παράδειγμα μοντέλου εκθετικής παλινδρόμησης, μεταξύ του συντελεστή δόμησης και της απόστασης μιάς κατοικίας από το κέντρο της πόλης του Reading στην Αγγλία. Η εκλογή της συγκεκριμένης πόλης έγινε διότι υπήρχαν τα δεδομένα.

Εάν θεωρήσουμε σαν ανεξάρτητη μεταβλητή X την απόσταση από το κέντρο της πόλης και εξαρτημένη Ψ, τον συντελεστή δόμησης, έχουμε δεδομένα που δίνονται από τον παρακάτω πίνακα:

X_i	0,5	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0	5,5	6,0	6,5	7,0	7,5	8,0
Y_i	0,790	0,505	0,390	0,320	0,260	0,157	0,156	0,098	0,080	0,076	0,082	0,074	0,05	0,025	0,012	0,010

Αυτά τα ζεύγη τιμών, αποτυπώθηκαν σε σχεδιάγραμμα σε χαρτί μιλιμετρέ και διαπιστώθηκε ότι ακολουθούν μια σχέση της μορφής

$$Y = \lambda e^{bx}$$

όπου λ και b είναι κάποιες παράμετροι. Εάν γράψουμε τώρα, τον συντελεστή λ στην μορφή e^{α}, τότε η παραπάνω σχέση γίνεται

$$Y = e^{\alpha+bx} \Rightarrow \ell nY = \alpha = +bx$$

Έτσι, ο συντελεστής συσχέτισης μεταξύ της μεταβλητής ℓnY και X, που βρέθηκε, ήταν ίσος με -0,9740.

Οι συντελεστές της ευθείας παλινδρόμησης μεταξύ του ℓnY και X βρέθηκαν ίσοι με α = -0,0935 και β = -0,5136, δηλαδή

$$\ell nY = -0,0935-0,5136X \qquad \text{ή} \qquad Y = e^{-0,0935-0,5136X}$$

ή βρίσκοντας την τιμή του $e^{-0,0935}$ καταλήγουμε στο εκθετικό μοντέλο που υποθέσαμε στην αρχή και έχει την συγκεκριμένη μορφή

$$Y = 0,9107 \cdot e^{-0,5136X}$$

Η τυπική απόκλιση του Y βρέθηκε $\sigma_\psi = 0,2756$ και έτσι, μεταξύ των καμπύλων $Y = e^{-0,0935-0,5136X-0,2756}$ και $Y = e^{-0,0935-0,5136X+0,2756}$ περιέχονται περίπου το 68 % των μετρήσεων.

Μια άλλα μορφή της εκθετικής παλινδρόμησης θα είναι της μορφής $Y = \alpha X^{\beta} \Rightarrow \ell nY = \ell n\alpha + \beta \ell nX$, που σε λογαριθμικό διάγραμμα είναι ευθεία. **[5]**

Αθροίζοντας ως προς όλες τις μετρήσεις του δείγματός μας, καταλήγουμε στο γραμμικό σύστημα με αγνώστους τα α και β:

$$\begin{cases} \sum_{i=1}^{n}\left(\ell nY_i\right) = n\ell n\alpha + \beta\sum\left(\ell nX_i\right) \\ \sum\left(\ell nX\right)\left(\ell nY\right) = \left(\ell n\alpha\right)\sum\left(\ell nX_i\right) + \beta\sum\left(\ell nX_i\right)^2 \end{cases}$$

Άρα, για δύο μεταβλητές Χ και Ψ που μας δίνονται οι τιμές τους, αρκεί να κατασκευάσουμε ένα πίνακα που θα έχει την παρακάτω μορφή:

X_i	Y_i	ℓnX_i	ℓnY_i	$\left(\ell nX_i\right)\left(\ell nY_i\right)$	$\left(\ell nX_i\right)^2$
⋮	⋮	⋮	⋮	⋮	⋮
		$\sum \ell nX_i$	$\sum \ell nY_i$	$\sum \ell nX_i \cdot \ell nY_i$	$\sum\left(\ell nX_i\right)^2$

2.6 Πολλαπλή παλινδρόμηση

Σε περισσότερες από δύο μεταβλητές, εάν θεωρήσουμε π.χ. για τρείς μεταβλητές ότι οι δύο, Χ και Ψ είναι ανεξάρτητες και ότι η τρίτη, Z είναι η εξαρτημένη, τότε ζητάμε να βρούμε ένα **επίπεδο παλινδρόμησης** που προσεγγίζει καλύτερα τα σημεία μας στον χώρο των τριών διαστάσεων.

Η εξίσωση του επιπέδου αυτού θα είναι

$$Z = \alpha + \beta X + \gamma Y \quad \textbf{(1)}$$

και οι συντελεστές α, β και γ προσδιορίζονται εύκολα από την επίλυση του παρακάτω γραμμικού συστήματος που δημιουργείται με πολ/σμό της **(1)** επί 1, X και Y αντίστοιχα για τις τρεις εξισώσεις και άθροιση κατά μέλη γιά όλα τα ι των μετρήσεών μας.

Το σύστημα αυτό είναι:

$$\begin{cases} \sum Z = n\alpha + \beta \sum X + \gamma \sum Y \\ \sum ZX = \alpha \sum X + \beta \sum X^2 + \gamma \sum XY \\ \sum ZY = \alpha \sum Y + \beta \sum XY + \gamma \sum Y^2 \end{cases}$$

Στη γενική μορφή, το υπερ-επίπεδο της πολλαπλής παλινδρόμησης γράφεται:

$$y = b_o + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + e \qquad \textbf{(2)}$$

Αν θεωρήσουμε τώρα ότι έχουμε n τιμές της εξαρτημένης μεταβλητής y που η κάθε μία από αυτές συνδέεται με την παραπάνω σχέση **(2),** με κ τιμές των μεγεθών $x_1,\ldots,x_\kappa$. Εάν θέσουμε

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ b = \begin{bmatrix} b_o \\ b_1 \\ \vdots \\ b_k \end{bmatrix},\ e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \text{ και } X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & & & & \\ \vdots & & & & \\ 1 & x_{n1} & & & x_{nk} \end{bmatrix}$$

και με $\hat{b}_o, \hat{b}_1, \ldots, \hat{b}_k$ τους εκτιμητές των $b_o, b_1, \ldots, b_k$ η σχέση **(2)** γενικεύεται για n παρατηρήσεις και γράφεται υπό μορφή πινάκων

$$Y = XB + e$$

Αποδεικνύεται ότι το διάνυσμα $\hat{b}$ (με τους εκτιμητές), που ελαχιστοποιεί την ποσότητα $\sum e_i^2$ (δηλ. οι συντελεστές που προσδιορίζουν το υπέρ-επίπεδο ως προς τους άξονές $x_1,\ldots,x_\kappa$) δηλαδή, οι εκτιμητές των ελάχιστων τετραγώνων είναι $\hat{b} = (X'X)^{-1} X'Y$ με την προϋπόθεση ύπαρξης του πίνακα $(X'X)$

2.7 Μελέτη της ευθείας παλινδρόμησης [βιβλιογρ. 7,8,9]

Είδαμε στην παράγραφο 2, εξισ. **(5)** και **(6)** ότι οι συντελεστές α και β της ευθείας παλινδρόμησης $Y = \alpha + \beta X$ είναι ίσοι αντίστοιχα με

$$\alpha = \frac{\sum Y_i}{n} - \beta \frac{\sum X_i}{n} = \mu_Y - \beta \mu_X$$

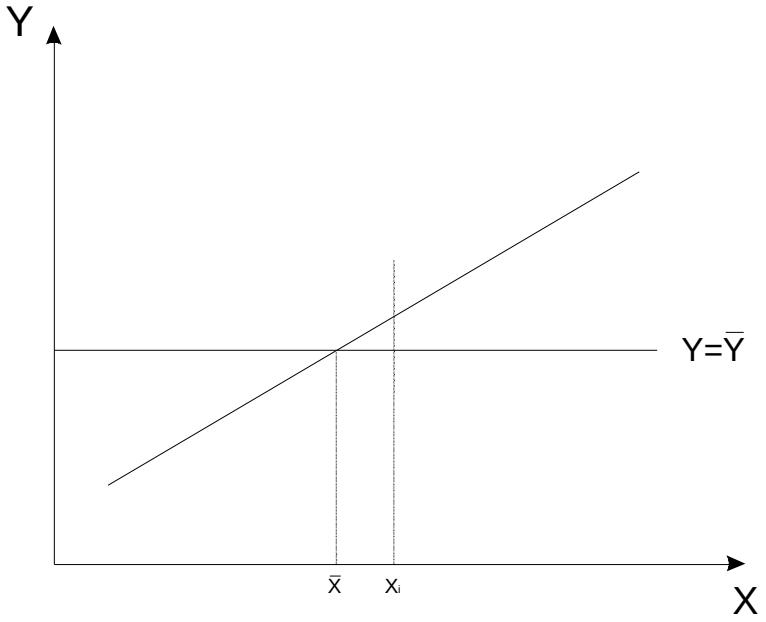
όπου το

$$\beta = \frac{\sum \left(X_i - \overline{X}\right) \cdot \left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} = \frac{n \cdot \sum X_i Y_i - \sum X_i \cdot \sum Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} = \frac{Cov(X,Y)}{Var(X)}$$

όπου συμβολίσαμε με Cov (x,y) την συνδιακύμανση των X και Y και με Var (X) την διακύμανση του X, που γράφονται S_{XY} και S_X^2 αντίστοιχα και έτσι $\beta = \frac{S_{XY}}{S_X^2}$

Έαν στο παρακάτω σχήμα παραστήσουμε με $\left(X_i, Y_i\right)$ το ζευγάρι των τιμών που βρίσκεται επάνω στην ευθεία παλινδρόμησης, δηλ. το ζευγάρι κάθε τιμής της ανεξάρτητης μεταβλητής με την αντίστοιχη προβλεπόμενη τιμή μετά την εύρεση της ευθείας παλινδρόμησης, με $\left(X_i, \overline{Y}\right)$ τα ζευγάρια του X_i με την μέση τιμή των Y_i και που βρίσκονται πάνω στην παράλληλη ευθεία προς τον άξονα των X σε ύψος $\overline{Y}$ και $\left(X_i, Y_i\right)$ τα αρχικά ζευγάρια τιμών, παρατηρούμε σύμφωνα με το σχήμα ότι ισχύει η σχέση

$$\left(Y_i - \overline{Y}\right) = \left(Y_i - \hat{Y}\right) + \left(\hat{Y} - \overline{Y}_i\right) \qquad \textbf{(1)}$$



Σύμφωνα με την **(1)** η **συνολική απόκλιση** $Y_i - \overline{Y}$ της τιμής Y_i από τον μέσο Υ αποτελείται από την $Y_i - \hat{Y}_i$ που υπάρχει μεταξύ της αρχικής τιμής Y_i και αυτής που βρέθηκε με την παλινδρόμηση $\hat{Y}_i$ και λέγεται **απόκλιση που δεν εξηγήθηκε** από το μοντέλο της παλινδρόμησης και της $\hat{Y}_i - \overline{Y}$ που είναι το ποσο που μειώθηκε η συνολική απόκλιση με την εύρεση της ευθείας παλινδρόμησης και λέγεται απόκλιση που εξηγήθηκε από το μοντέλο.

Η **(1)** αποδεικνύεται ότι ισχύει και για τα αθροίσματα τετραγώνων των όρων της, δηλ. για όλα τα $\hat{Y}_i$ Y_i και άρα ισχύει η παρακάτω σχέση:

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(Y_i - \hat{Y}\right)^2 + \sum\left(\hat{Y}_i - \overline{Y}\right)^2$$

που αν θέσουμε με την διεθνή ορολογία :

$\sum\left(Y_i - \overline{Y}\right)^2 = SS_{total}$: συνολικό άθροισμα τετραγώνων (Total sum of squares)

$\sum\left(Y_i - \hat{Y}_i\right)^2 = SS_{resid}$: άθροισμα τετραγώνων που δεν ερμηνεύτηκε από την παλινδρόμηση (Sum of squares residual)

$\sum\left(\hat{Y}_i - \overline{Y}\ \right)^2 = SS_{reg}$: άθροισμα των τετραγώνων που ερμηνεύτηκε με την παλινδρόμηση (Sum of squares of the regression)

θα έχουμε:

$$\mathbf{SS_{total} = SS_{resid} + SS_{reg}}$$

και έτσι, για τον υπολογισμό και των συντελεστών της ευθείας παλινδρόμησης και των παραπάνω μεγεθών δημιουργούμε τον παρακάτω πίνακα:

X_i	Y_i	$X_i Y_i$	$\hat{Y}_i$	$Y_i - \hat{Y}_I$	$\left(Yi - \hat{Y}_i\right)^2$	$Y_i - \overline{Y}$	$\left(Y_i - \overline{Y}\right)^2$	$\hat{Y}_i - \overline{Y}$	$\left(\hat{Y}_i - \overline{Y}\right)^2$	X_i^2
$\sum X_i$	$\sum Y_i$	…	…	…						$\sum X_i^2$

Γνωρίζοντας τώρα ότι οι βαθμοί ελευθερίας για την ευθεία παλινδρόμησης είναι n-1, δηλ. ο αριθμός των σημείων -1 για την εκτίμηση της μέσης

τιμής $\overline{Y} = \alpha + \beta\overline{X}$, ο βαθμός ελευθερίας για το άθροισμα των τετραγώνων της παλινδρόμησης SS_{reg} είναι 1, έχουμε ότι οι βαθμοί ελευθερίας που αντιστοιχούν στο SS_{resid} είναι n-2.

Αποδεικνύεται ότι ο λόγος $\dfrac{SS_{reg} / 1}{SS_{resid} / n - 2}$

ακολουθεί την $F_{a(1,n-2)}$ κατανομή

Δηλαδή, αν το μέγεθος F που θα βρεθεί τελικά από τον παρακάτω πίνακα διακύμανσης της γραμμής παλινδρόμησης είναι μεγαλύτερο από το αντίστοιχο $F_{a(1,n-2)}$ του πίνακα της F-κατανομής, τότε δεχόμαστε ως αξιόπιστο το μοντέλο της παλινδρόμησης που βρέθηκε.

Ο πίνακας της ανάλυσης διακύμανσης της παλινδρόμησης είναι ο:

ANOVA της παλινδρόμησης

	Αθροίσματα τετραγώνων	**Βαθμοί ελευθερίας**	**Μέσα αθροίσματα τετραγώνων**	**F**
Παλινδρόμηση που εξηγήθηκε	SS_{reg}	1	$SS_{reg} / 1$	
Παλινδρόμηση που δεν εξηγήθηκε	SS_{resid}	n-2	SS_{resid} / n-2	$F = \dfrac{SS_{reg} / 1}{SS_{resid} / n - 2}$
Σύνολο παλινδρόμησης	SS_{Total}	n-1		

7.α. Κατανομές των εκτιμητών $\hat{\alpha}$ και $\hat{\beta}$ της ευθείας των ελαχίστων τετραγώνων.

Αποδεικνύεται, όταν ισχύουν οι προϋποθέσεις της παραγράφου 1, ότι οι εκτιμητές $\hat{\alpha}$ και $\hat{\beta}$ που βρίσκονται από την ευθεία της γραμμικής παλινδρόμησης ακολουθούν την κανονική κατανομή με

$$\alpha \to N(\alpha, \frac{\sigma^2 \sum X_i^2}{n\sum (X_i - \overline{X})^2}) \qquad \textbf{(1)}$$

$$\beta \to N(\beta, \frac{\sigma^2}{\sum (X_i - \overline{X})^2}) \qquad \textbf{(2)}$$

Στους παραπάνω τύπους σ^2 είναι η κοινή διακύμανση των κατανομών Y που συχνά δεν γνωρίζουμε. Γι αυτό, χρησιμοποιούμε έναν αμερόληπτο εκτιμητή της σ^2 που είναι η δειγματική διασπορά S^2 των $Y_ι$ ή με τον συμβολισμό της προηγούμενης παραγράφου $SSresid/n-2$.

Έτσι, αντικαθιστώντας τους τύπους **(1)** και **(2)** όπου σ^2 των εκτιμητή του $S^2 = SSresid/n-2$ έχουμε ότι οι τυχαίες μεταβλητές $t_\alpha = \frac{\hat{\alpha} - \alpha}{S_\alpha}$ και $t_\beta = \frac{\hat{\beta} - \beta}{S_\beta}$ ακολουθούν την t_{n-2} κατανομή του Student, εδώ, s_a και s_β είναι οι τυπικές αποκλίσεις που προκύπτουν από τους τύπους **(1)** και **(2)** εάν στις διακυμάνσεις θέσουμε όπου σ^2 το s^2.

Μπορούμε τώρα να δημιουργήσουμε:

1ο. ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΤΑ α ΚΑΙ β ΤΗΣ ΕΥΘΕΙΑΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Για το α

Ένα διαστημα εμπιστοσύνης σε στάθμη σημαντικότητας a/o είναι το

$$\left[\hat{\alpha} - s_\alpha \cdot t_{n-2, a/2}, \alpha + s_\alpha \cdot t_{n-2, a/2}\right]$$

Για το β

είναι το $\left[\hat{\beta} - s_\beta \cdot t_{n-2,\frac{a}{2}}, \beta + s_\beta \cdot t_{n-2,\frac{a}{2}}\right]$

2ο. ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΗΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΔΙΑΦΟΡΑΣ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ α ΚΑΙ β ΑΠΟ ΣΥΓΚΕΚΡΙΜΕΝΕΣ ΤΙΜΕΣ α_o ΚΑΙ β_o.

Για το α έχω ότι

H_o: $\alpha = \alpha_o$ με στάθμη σημαντικότητας a

H_1: $\alpha \neq \alpha_o$

Η H_o απορρίπτεται εάν $t_{\alpha_o} \neq t_{n-2,a(n\frac{a}{2})}$, για δίπλευρο έλεγχο.

Όμοια για το β

2.8 Ο συντελεστής προσδιορισμού r^2 και συντελεστής συσχέτισης r. Ελεγχοι για μεγάλα και μικρά δείγματα για τον r.

Ο συντελεστής προσδιορισμού r^2 είναι το ποσοστό της συνολικής απόκλισης (από την μέση τιμή $\overline{Y}$ της ευθείας παλινδρόμησης) που εξηγείται από την ευθεία των ελαχίστων τετραγώνων.

Έτσι, $\hat{r}^2 = \frac{SS_{REG}}{SS_{TOTAL}} = \frac{\sum\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2}$

Ο $\hat{r}^2$ δεν είναι αμερόληπτος εκτιμητής του πληθυσμού. Ο αντίστοιχος αμερόληπτος εκτιμητής του πληθυσμού είναι ο:

$$r^2 = 1 - \left(1 - \hat{r}^2\right)\frac{n-1}{n-2} = 1 - \frac{SS_{reg}/n-2}{SS_{total}/n-1}$$

Έτσι, η αξιολόγηση της ευθείας παλινδρόμησης πρέπει να γίνεται πρώτα από την τιμή του F της ANOVA της γραμμικής παλινδρόμησης (μεγαλύτερο από πίνακες) δεύτερον, από τους συντελεστές της ευθείας και τα διαστήματα εμπιστοσύνης τους (όσο το δυνατόν μικρότερα) και από τον συντελεστή προσδιορισμού r^2 που κυμαίνεται από 0 έως 1 (με τιμή 1 όταν όλα τα σημεία μου συμπίπτουν στην ευθεία παλινδρόμησης) και που όσο κοντα στο 1 βρίσκεται τόσο είναι καλύτερη η ευθεία παλινδρόμησης.

Αποδεικνύεται ότι, ο συντελεστής προσδιορισμού είναι το τετράγωνο του συντελεστή συσχέτισης.

Εάν ισχύουν οι προϋποθέσεις της παραγράφου 1 και αν οι κατανομές των $\hat{X}$ και $\hat{Y}$ είναι κανονικές και η διασπορά των $\hat{X}$ κατανομών κοινή, ισχύουν οι παρακάτω έλεγχοι υποθέσεων για τον συντελεστή συσχέτισης r

(ρ: σ.σ. του πληθυσμού)

ΓΙΑ ΜΙΚΡΑ ΔΕΙΓΜΑΤΑ και Ho: $\rho = 0$ $(n<25)$

Αποδεικνύεται ότι η ποσότητα $\dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} \rightarrow t_{n-2}$

(ρ_o: συγκεκριμένη τιμή του ρ)

ΓΙΑ ΠΟΛΥ ΜΕΓΑΛΑ ΔΕΙΓΜΑΤΑ και Ho: $\rho = \rho_o$ $(n>500)$

Αποδεικνύεται ότι η ποσότητα $r \rightarrow N\left(\rho, \left(1 - \dfrac{\rho^2}{\sqrt{n}}\right)^2\right)$ όπου εξετάζεται εάν ο σ.σ. του πληθυσμού ρ ισούται με μια συγκεκριμένη τιμή ρ_o.

<u>Μετασχηματισμοί Fischer σε μεγάλα και μικρά δείγματα</u>

ΜΕΓΑΛΑ ΔΕΙΓΜΑΤΑ Ho: $\rho = \rho_o$

$$z' = \frac{1}{2}\ln\frac{1+r}{1-r}, \sigma^2_{z'} = \frac{1}{n-3}$$

$$\text{και } z = \frac{z' - z_{\rho_o}}{\sigma_{z'}} \rightarrow N(0,1)$$

ΜΙΚΡΑ ΔΕΙΓΜΑΤΑ $(n<25)$

$$z'' = z' - \frac{3z'+r}{4n} \text{ και } \sigma^2_{z''} = \frac{1}{n-1}$$

$$\text{με } z = \frac{z'' - z''_{\rho_o}}{\sigma_{z''}} \rightarrow N(0,1).$$

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Wonnacott R., Wonnacott T., «Introductory Statistics» 4th Edition, John Wiley & Sons, 1985.

[2] Baxter R., «Computer and Statistical Techniques for Planners», Metuen & Co Ltd, London 1976.

[3] Darlington R., «Regression and Linear modelw», McGraw Hill, 1990.

[4] Spiegel M., «Πιθανότητες και Στατιστική», μετάφραση Σ. Περσίδης, Schaums outline series, McGraw Hill, ΕΣΠΙ Αθήνα 1977.

[5] Δρόσος Γ., Καραμπιστόλης Δ., «Στατιστική Επιχειρήσεων» Εκδόσεις Έλλην, 1994.

[6] Χρίστουλας Κ., «Στοιχεία πολλαπλής παλινδρομήσεως», Υπουργείο Γεωργίας, ΓΕΝ. ΔΙΕΥΘΥΝΣΙΣ Γ/ΚΗΣ ΑΝΑΠΤΥΞΕΩΣ ΚΑΙ ΕΡΕΥΝΩΝ, ΥΠΗΡ. Γ/ΚΩΝ ΕΡΕΥΝΩΝ, ΑΘΗΝΑΙ 1974.

[7] Κάτος Αναστάσιος, "Στατιστική", Παρατηρητής, Θεσ/νίκη 1986.

[8] Draper N.R., Smith H., "Applied Regression Analysis", 2nd Edition, J. Wiley & Sons, 1981.

[9] Myers R., "Classical and modern Regression with applications" THE DUXBURY ADVANCED SERIES IN STATISTICS AND DECISION SCIENCES, PWS-KENT Publishing Comnpany, 1990.

ΚΕΦΑΛΑΙΟ 3

ΑΝΑΛΥΣΗ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ

3.1 Η έννοια της απόστασης

Η έννοια της απόστασης μεταξύ ατόμων είναι θεμελιώδης στην Ανάλυση σε Κύριες Συνιστώσες (Α.Κ.Σ.) και πιο γενικά στην Ανάλυση Δεδομένων γιατί:

1. ΠΟΛΥ ΣΥΧΝΑ, η ανάλυση ενός πίνακα ΑΤΟΜΩΝ-ΜΕΤΑΒΛΗΤΩΝ, οδηγεί σε συμπεράσματα που αφορούν την προσέγγιση ή την απομάκρυνση 2 ατόμων, ως προς μία ή περισσότερες μεταβλητές.
2. ΓΕΝΙΚΑ, ο αριθμός των ατόμων είναι μεγάλος (συνήθως μεγαλύτερος από τον αριθμό των μεταβλητών) και ενδιαφέρει εάν μπορούμε να τα ομαδοποιήσουμε σε κλάσεις ομογενείς ή όχι.
3. Τα αποτελέσματα μιας Α.Κ.Σ. περιέχουν σχήματα που μας παρουσιάζουν τα άτομα ή τις μεταβλητές ΄΄οπυ το μάτι «διαβάζει» τις αποστάσεις με τον ίδιο τρόπο που θα έβλεπε και κάποιον χάρτη.

Παράδειγμα

Έστω ο πίνακας δεδομένων:

Μεταβλητη	X_1	X_2	X_3
Ατομο	1	2	3
1	2	1	2
2	3	2	1
3	2	2	7
4	2	1	1
5	1	2	2
6	2	1	2

με $\overline{X_1} = \frac{12}{6} = 2, \overline{X_2} = 1{,}5, \overline{X_3} = 2{,}5$

Η απόσταση ανάμεσα από 2 άτομα είναι μια συνάρτηση των διαφορών των τιμών για όλες τις τιμές όλων των μεταβλητών για τα δύο αυτά άτομα . Έστω ι και ι′ και j ο αριθμός των μεταβλητών τότε:

$$d_{ii'}^2 = \frac{1}{p}\left(\sum_j \left(x_{ij} - x_{i'j}\right)^2\right)$$ όπου p ο αριθμός των μεταβλητών.

Στο παραπάνω παράδειγμα, η απόσταση μεταξύ των ατόμων 2 και 3 είναι ίση με p=3

$$d_{23} = \frac{1}{3}\left(\left(x_{21} - x_{31}\right)^2 + \left(x_{22} - x_{32}\right)^2 + \left(x_{23} - x_{33}\right)^2\right)^{\frac{1}{2}} =$$

$$= \frac{1}{3}\left((3-2)^2 + (2-2)^2 + (1-7)^2\right)^{\frac{1}{2}} = \frac{1}{3}(1+0+36)^{\frac{1}{2}} = \frac{\sqrt{37}}{3}$$

Βλέπουμε ότι, η απόσταση ορισμένη έτσι εξαρτάται πολύ από τις μεταβλητές που έχουν μεγάλη διακύμανση. Στη συγκεκριμένη περίπτωση, η μεταβλητή 3 συμμετέχει κατά $6^2/3$ στο προηγούμενο άθροισμα και ουσιαστικά, ορίζει αυτή την απόσταση.

Για να απαλειφθεί η εξάρτηση αυτή από την διακύμανση κάθε μεταβλητής, θεωρούμε ως απόσταση την

$$d_{ii'} = \frac{1}{p}\left(\sum_j \left(x_{ij} - x_{i'j}\right)^2 \Big/ \sigma_j^2\right) \quad \textbf{(1)}$$

όπου $\sigma_j^2 = \sum_i \left(x_{ij} - \overline{x_j}\right)^2 \Big/ n$ είναι η διακύμανση της μεταβλητής x_j και $\overline{x_j} = \frac{1}{n}\sum_i x_{ij}$,

η μέση τιμή της μεταβλητής x_j.

Στο παραπάνω παράδειγμα, οι μέσες τιμές των τριών μεταβλητών είναι:

$\overline{x_1} = 2, \overline{x_2} = 1{,}5, \overline{x_3} = 2{,}5$ και οι διακυμάνσεις τους:

$$\sigma_1^2 = \left[(2-2)^2 + (2-3)^2 + (2-2)^2 + (2-2)^2 + (1-2)^2 + (2-2)^2\right]/6 = \frac{2}{6} = 0{,}33$$

$$\sigma_2^2 = \left[(1-1{,}5)^2 + (2-1{,}5)^2 + (2-1{,}5)^2 + (1-1{,}5)^2 + (2-1{,}5)^2 + (1-1{,}5)^2\right]/6 = \frac{6\cdot(0{,}5)^2}{6} = 0{,}5^2 = 0{,}25$$

$$\sigma_3^2 = \left[(2-2{,}5)^2 + (1-2{,}5)^2 + (7-2{,}5)^2 + (1-2{,}5)^2 + (2-2{,}5)^2 + (2-2{,}5)^2\right]/6 = \frac{25{,}5}{6} = 4{,}25$$

Η μ'αυτόν τον τρόπο ΤΥΠΙΚΗ απόσταση (απόσταση που προκύπτει από την τυποποίηση των μεταβλητών) έχει την ιδιότητα να δίνει το ίδιο βάρος σε όλες τις μεταβλητές που υπεισέρχονται στον υπολογισμό της απόστασης.

Ο τύπος (1) γράφεται επίσης:

$$d_{ii'} = \frac{1}{p} \cdot \frac{\sum_{j} \left(x_{ij} - x_{i'j} \right)^2}{\sum_{i=1}^{n} \left(x_{ij} - \frac{1}{n} \sum x_{ij} \right)^2}$$

Έτσι, η τυποποιημένη ή ΤΥΠΙΚΗ απόσταση μεταξύ των ατόμων 2 και 3 στο παραπάνω παράδειγμα θα είναι:

$$d_{23}^2 = \frac{1}{3} \cdot \frac{\sum_{j=1}^{3} \left(x_{2j} - x_{3j} \right)^2}{\sum_{i=1}^{6} \left(x_{ij} - \frac{1}{n} \sum x_{ij} \right)^2} =$$

$$= \frac{1}{3} \left(\frac{\left(x_{21} - x_{31} \right)^2}{\sigma_1^2} + \frac{\left(x_{22} - x_{32} \right)^2}{\sigma_2^2} + \frac{\left(x_{23} - x_{33} \right)^2}{\sigma_3^2} \right) = \frac{1}{3} \left(\frac{1}{0,33} + \frac{0}{0,25} + \frac{36}{4,25} \right) = \frac{3 + 0 + 8,47}{3} \Rightarrow$$

$$\Rightarrow d_{23} = \frac{11,47}{3}$$

Βλέπουμε δηλαδή, ότι η μεταβλητή 3 συμμετέχει κατά $\frac{8,47}{3}$ στα $\frac{11,47}{3}$ του d_{23}^2 ενώ, πριν την τυποποίηση, συμμετείχε κατά $\frac{36}{3}$ στα $\frac{37}{3}$.

Η τυπική απόσταση στις αρχικές μεταβλητές γίνεται μη τυπική με κατάλληλο μετασχηματισμό των αρχικών μεταβλητών.

Έτσι, αν θέσουμε $\psi_{ij} = \frac{x_{ij}}{\sigma_j}$, τότε θα έχουμε:

$$d_{ii'} = \frac{1}{p} \cdot \left(\sum_{j} \left(x_{ij} - x_{i'j} \right)^2 \Big/ \sigma_j^2 \right) = \frac{1}{p} \cdot \left(\sum_{j} \left(\frac{x_{ij}}{\sigma_j} - \frac{x_{i'j}}{\sigma_j} \right)^2 \right) = \frac{1}{p} \cdot \left(\sum_{j} \left(\psi_{ij} - \psi_{i'j} \right)^2 \right)$$

Γενικά, μια ΕΥΚΛΕΙΔΙΑ ΑΠΟΣΤΑΣΗ μεταξύ δύο ατόμων ι και ι' είναι:

$$d_{ii'} = \frac{1}{p} \cdot \left(\sum_{j} \sum_{j'} \left(x_{ij} - x_{i'j} \right) \left(x_{ij'} - x_{i'j'} \right) \cdot \alpha_{jj'} \right) = \frac{1}{p} \left(x_i - x_{i'} \right) \cdot A \cdot \left(x_i - x_{i'} \right)' \quad \textbf{(2)}$$

όπου $\left(x_i - x_{i'} \right)' = \begin{bmatrix} x_{i1} - x_{i'1} \\ \vdots \\ x_{ip} - x_{i'p} \end{bmatrix}$ είναι ο ανάστροφος πίνακας του $\left(x_\iota - x_{\iota'} \right)$ και

Α ο τετραγωνικός, συμμετρικός πίνακας γενικού όρου $\alpha_{jj'}$.

Έτσι, όταν $\begin{cases} \alpha_{jj'} = 0, j \neq j' \\ \alpha_{jj'} = 1, j = j' \end{cases}$ έχουμε την **ΜΗ ΤΥΠΙΚΗ ΕΥΚΛΕΙΔΙΑ ΑΠΟΣΤΑΣΗ** και όταν $\begin{cases} \alpha_{jj'} = 0, j \neq j' \\ \alpha_{jj'} = \frac{1}{\sigma_j}, j = j' \end{cases}$, έχουμε την **ΤΥΠΙΚΗ ΕΥΚΛΕΙΔΙΑ ΑΠΟΣΤΑΣΗ.**

Στο συγκεκριμένο παράδειγμα των τριών μεταβλητών, οι δείκτες j και j′ μεταβάλλονται από 1 έως 3 και έτσι, το μέλος με τα αθροίσματα του τύπου 2 γράφεται:

$$d_{ii'} = \frac{1}{p} \left[\sum_{j'} \left(x_{i1} - x_{i'1} \right) \left(x_{ij'} - x_{i'j'} \right) \alpha_{1j'} + \sum_{j'} \left(x_{i2} - x_{i'2} \right) \left(x_{ij'} - x_{i'j'} \right) \alpha_{2j'} + \sum_{j'} \left(x_{i3} - x_{i'3} \right) \left(x_{ij'} - x_{i'j'} \right) \alpha_{3j'} \right] =$$

$= \frac{1}{p}$ * 3 όροι για το κάθε $\sum_{j'}$ με j′ από το 1 έως το 3 (δηλ. σύνολο 9 όροι).

Αλλά, εάν διαλέξουμε την τυπική Ευκλείδια απόσταση που αναφέρθηκε προηγούμενα και θελήσουμε να βρούμε την απόσταση μεταξύ των ατόμων ι = 1 και ι′ = 3 όπως και προηγούμενα από το πρώτο $\sum_{j'}$ j′ = 1,3 θα μείνει ο όρος που αντιστοιχεί στο j′ = 1 από το δεύτερο, στο j′ = 2 και από το τρίτο στο j′ = 3.

Έτσι, τελικά θα έχουμε:

$$d_{23}^2 = \frac{1}{p} \left[\left(x_{21} - x_{31} \right) \left(x_{21} x_{31} \right) \alpha_{11} + \left(x_{22} - x_{32} \right) \left(x_{22} - x_{32} \right) \alpha_{22} + \left(x_{23} - x_{33} \right) \left(x_{23} - x_{33} \right) \alpha_{33} \right]$$

και όπως $\alpha_{jj'} = \frac{1}{\sigma_j}$ θα έχουμε

$$d_{23}^2 = \frac{1}{3}\left[\frac{\left(x_{21}-x_{31}\right)^2}{\sigma_1^2} + \frac{\left(x_{22}-x_{32}\right)^2}{\sigma_2^2} + \frac{\left(x_{23}-x_{33}\right)^2}{\sigma_3^2}\right]$$ που βρήκαμε και προηγούμενα.

3.2 Ανάλυση σε κύριες συνιστώσες

Συχνά, έχουμε να αντιμετωπίσουμε πίνακες αριθμών που εμφανίζουν την παρακάτω δομή: Οι κολώνες του εμφανίζονται σαν ορισμένες μεταβλητές των οποίων οι τιμές μετρήθηκαν σε ορισμένα άτομα, κάθε ένα από τα οποία αντιστοιχεί σε μία γραμμή του πίνακα.

Έτσι, κάθε γραμμή του πίνακα, εκφράζει τις τιμές που παίρνει η κάθεμία από τις μεταβλητές (κολώνες) στο συγκεκριμένο άτομο που χαρακτηρίζεται από αυτή τη γραμμή.

Συγκεκριμένα, η μορφή των πινάκων αυτών, είναι συνήθως η παρακάτω:

ΜΕΤΑΒΛΗΤΕΣ / ΑΤΟΜΑ	ψ_{o1}	$\cdots$	ψ_{oj}	$\cdots$	ψ_{op}
ψ_{1o}	ψ_{11}	$\cdots$	$\cdots$		$\vdots$
$\vdots$	$\vdots$	$\ddots$			$\vdots$
ψ_{io}	ψ_{i1}		ψ_{ij}		$\vdots$
$\vdots$				$\ddots$	$\vdots$
ψ_{no}	$\cdots$	$\cdots$	$\cdots$	$\cdots$	ψ_{np}

όπου ψ_i το διάνυσμα γραμμή $\left(\psi_{i1},\ldots,\psi_{ij},\ldots,\psi_{ip}\right)$

και ψ_j το διάνυσμα στήλη $\begin{pmatrix}\psi_{1j}\\ \vdots \\ \psi_{ij}\\ \vdots \\ \psi_{ip}\end{pmatrix}$

Το μεγάλο μέγεθος των πινάκων αυτών (π.χ. 200*40) καθώς και οι σχέσεις μεταξύ διαφόρων μεταβλητών ή ακόμα και ατόμων που δεν είναι καθόλου εμφανής, γέννησε την ανάγκη ύπαρξης μιας μεθόδου ανάλυσης δεδομένων που να λύνει τα

προβλήματα αυτά. Η μέθοδος αυτή ονομάστηκε ΑΝΑΛΥΣΗ ΣΕ ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ (principal componet analysis ή analyse en composantes principales-γαλ.)

Χαρακτηρίζεται σαν μια από τις βασικές **ΠΑΡΑΓΟΝΤΙΚΕΣ** ΜΕΘΟΔΟΥΣ γιατί, η μείωση των δεδομένων, διατηρώντας κατά το μέγιστο δυνατό το ποσό της αρχικής πληροφορίας, γίνεται μέσω ενός **γραμμικού** συνδιασμού των αρχικών μεταβλητών κάθε μιας από αυτές πολ/μένης με έναν ΠΑΡΑΓΟΝΤΑ. Οι γραμμικές σχέσεις που θα βρεθούν μετά την εφαρμογή της μεθόδου και που συνδέουν κάθε μία από τις καινούριες μεταβλητές με τιε παλιές, χαρακτηρίζουν την μέθοδο σαν **ΓΡΑΜΜΙΚΗ.**

Διαχωρίζεται σαφώς από την ΑΝΑΛΥΣΗ ΑΝΤΙΣΤΟΙΧΙΩΝ (Correspondance Analysis), η οποία επεξεργάζεται μόνο δεδομένα σε ποιοτική και όχι σε ποσοτική μορφή και από τις

1ο ΚΑΝΟΝΙΚΗ ΑΝΑΛΥΣΗ (canonical analysis)

2ο ΑΝΑΛΥΣΗ ΔΙΑΧΩΡΙΣΜΟΥ (discriminant analysis)

γιατί, οι δύο τελευταίες επεξεργάζονται δεδομένα που μπορούν, σαφώς, να χωριστούν κατά ομάδες (μεταβλητών) **[βιβλιογρ. 3].**

Το συχνότερο πρόβλημα που εμφανίζεται στην πράξη, κατά την επεξεργασία ενός τέτοιου πίνακα δεδομένων είναι η ανομοιογένεια των μονάδων μετρήσεων μεταξύ των μεταβλητών.

Έτσι, μπορεί μία μεταβλητή να εκφράζει δρχ, άλλη cm και άλλη βαθμούς Κελσίου. **[βιβλιογρ. 8]** Σε τέτοιες συνθήκες είναι πολύ σύσκολο να υπολογίσει κανείς την απόσταση μεταξύ 2 σημείων.

Στον χώρο των 2 διαστάσεων και με δύο μεταβλητές χ και ψ, ορθογώνιες μεταξύ τους και με τις ίδιες μονάδες μέτρησης, η κλασσικά χρησιμοποιούμενη απόσταση είναι η Ευκλείδεια, δηλαδή $d^2 = (x_2 - x_1)^2 + (\psi_2 - \psi_1)^2$.

Όμως, αυτή δεν είναι η γενική περίπτωση. Πριν αναφερθούμε στην Α.Κ.Σ. σαν μια μεθοδολογία που χωρίζεται σε διάφορα στάδια, για την εφαρμογή της πρέπει να θυμίσουμε μερικές βασικές μαθηματικές-στατιστικές έννοιες, απαραίτητες στην χρησιμοποίησή της.

Η στοιχειώδης αδράνεια ενός σημειακού σώματος Μ, μάζας m, απόστασης ΟΜ, ως προς ένα σημείο Ο στην Φυσική, είναι ίση με $m*OM^2$.

Στην Στατιστική, η έννοια της αδράνειας είναι ακριβώς ίδια υπό την συνθήκη να αντικαταστήσουμε την λέξη μάζα με την συχνότητα και την λέξη απόσταση με την απόκλιση από το μέγεθος κ (συνήθως η μέση τιμή):

$$f(x-\kappa)^2$$

Βλέπουμε ότι, για ένα νέφος σημείων, η συνολική αδράνεια ταυτίζεται με την **διακύμανση** $\sigma^2 = \frac{\sum_i f_i}{f}(x_i - \bar{x})^2$ **,** όπου $\frac{f_i}{f}$ η σχετική συχνότητα του x_i .

Το κέντρο βάρους ενός νέφους σημείων, είναι η **μέση τιμή** τους.

Ένα διάνυσμα $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ καλείται **ιδιοδιάνυσμα** ή **χαρακτηριστικό διάνυσμα** ενός τετραγωνικού πίνακα A = [α_{ij}], n x n, εάν υπάρχει $\lambda \in \Re$ τέτοιο ώστε:

AX=λX $\Rightarrow AX - \lambda X = \bar{0} \Rightarrow (A - \lambda I)X = \bar{0}$ **,** όπου I ο μοναδιαίος πίνακας n x n.

Η εξίσωση $|A - \lambda I| = 0$ λέγεται **χαρακτηριστική** εξίσωση του πίνακα Α. Οι ρίζες της εξίσωσης αυτής λέγονται **ιδιοτιμές** του πίνακα Α.

Οι μη μηδενικές λύσεις του ομογενούς συστήματος $(A - \lambda I)X = \bar{0}$ για κάθε ιδιοτιμή λ λέγονται **ιδιοδιανύσματα** του πίνακα Α.

Παράδειγμα: Να βρεθούν οι ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα $A = \begin{bmatrix} -3 & -2 \\ -2 & 4 \end{bmatrix}$.

Έχω ότι $|A - \lambda I| = 0 \Rightarrow \begin{vmatrix} -3\lambda & -2 \\ 3 & 4-\lambda \end{vmatrix} = 0 \Rightarrow \lambda^2 - \lambda - 6 = 0 \Rightarrow \lambda_1 = -2, \lambda_2 = 3$

άρα για $\lambda_1 = -2$ το αντίστοιχο ομογενές σύστημα γράφεται:

$$\left.\begin{matrix} (-3+2)x_1 - 2x_2 = 0 \\ 3x_1 + (4+2)x_2 = 0 \end{matrix}\right\} \Rightarrow x_1 = -2\kappa, x_2 = \kappa, \in \Re$$

Όλα τα ιδιοδιανύσματα $x_1 = \kappa\begin{pmatrix} -2 \\ 1 \end{pmatrix}$ αντιστοιχούν στην τιμή λ_1=-2 και όμοια, στην τιμή λ_2=3 αντιστοιχούν όλα τα ιδιοδιανύσματα $x_2 = \kappa\begin{pmatrix} 1 \\ -3 \end{pmatrix}$ **[βιβλιογρ. 9]**

ΣΚΟΠΟΙ ΤΗΣ Α.Κ.Σ.

Δύο είναι οι βασικοί σκοποί της μεθόδου:

1. **Μείωση του όγκου των δεδομένων.**

Με την εύρεση νέων μεταβλητών που είναι γραμμικοί συνδιασμοί των παλιών, επιτυγχάνεται η προβολή του συνόλου (νέφους) των σημείων, σε χώρους μικρότερων διαστάσεων από τον αρχικό. Π.χ. στο επίπεδο των δύο πρώτων παραγοντικών αξόνων (νέων μεταβλητών, ενώ αρχικά είχανε 3).

Έτσι, ουσιαστικά επιτυγχάνεται η μείωση του μεγάλου όγκου των αρχικών δεδομένων, με την μικρότερη δυνατή τροποποίηση των αποστάσεων μεταξύ των σημείων.

2. **Εμφάνιση σχέσεων μεταξύ των μεταβλητών και των σημείων των αρχικών δεδομένων που δεν είναι εμφανής**.

Έτσι, στους γραμμικούς συνδιασμούς, ένας συντελεστής κοντά στο 0, μας δείχνει ότι η μεταβλητή αυτή δεν λαμβάνει ενεργό μέρος στην δημιουργία της νέας μεταβλητής.

3.2.1 Στάδια της εφαρμογής της μεθόδου

3.2.1.1 Τυποποίηση του αρχικού πίνακα δεδομένων $Y = \left(\psi_{ij}\right)$

Μετασχηματίζουμε κάθε στοιχείο ψ_{ij} αφαιρώντας την μέση τιμή $\overline{\psi}_{oj}$ της μεταβλητής ψ_{oj} (στήλης) στην οποία ανήκει και διαιρώντας με την τυπική απόκλιση $S_{\psi_{oj}}$ επίσης της μεταβλητής που ανήκει και διαιρώντας με $\sqrt{n}$, όπου n, ο αριθμός των σημείων (γραμμών) του αρχικού πίνακα ΔΗΛΑΔΗ:

Τα νέα στοιχεία $x_{ij} = \frac{\psi_{ij} - \overline{\psi_{oj}}}{\sqrt{n} \cdot S_{\psi_{oj}}}$,δημιουργούν τον ΤΥΠΙΚΟ ΠΙΝΑΚΑ Χ.

3.2.1.2 Δημιουργία του πίνακα συσχετισεων

Ο πίνακας συσχετίσεων των μεταβλητών (correlation matrix) είναι ο $C = X^t X$ και το τυχόν στοιχείο του $c_{ικ}$ δίνει την συσχέτιση κάθε μιας μεταβλητής με τις υπόλοιπες.

Έχουμε: $c_{jk} = \sum_{i=1}^{n} \left(\frac{\psi_{ij} - \overline{\psi}_{oj}}{\sqrt{n} \cdot S_{\psi_{oj}}} \right) \left(\frac{\psi_{ik} - \overline{\psi}_{ok}}{\sqrt{n} \cdot S_{\psi_{ok}}} \right)$ και είναι ένας τετραγωνικός πίνακας p x p. (και ο αρχικός πίνακας δεδομένων Ψ και ο τυπικός Χ είναι διαστάσεων n x p)

3.2.1.3 Εύρεση των ιδιοτιμών και των ιδιοδιανύσματων του πίνακα συσχετίσεων

Οι ιδιοτιμές είναι οι λ_i, ι =1,..., p και ο πίνακας των ιδιοδιανυσμάτων u_i, ι = 1,...p είναι ο $u = \left[\left(u_1\right) \ldots \left(u_p\right) \right]$, p x p

3.2.1.4 Υπολογισμός του ποσοστού αδράνειας (διασποράς) του νέφους των σημείων (γραμμών) στον κάθε έναν από τους νέους παραγοντικούς άξονες ι=1,...p

Η **συνολική διασπορά Δ** δίνεται από το ίχνος της διαγωνίου του πίνακα των συσχετίσεων C, δηλαδή

$$\Delta = \text{Ιχν.}\left(X^t X\right) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$$

Άρα, το ποσοστό διασποράς σε κάθε έναν από τους νέους παραγοντικούς άξονες είναι

$$\Pi_i = \frac{\lambda_i}{\Delta} \quad , \iota=1,\ldots,p$$

Οι δύο πρώτοι άξονες συγκεντρώνουν το $\lambda_1+\lambda_2$ ποσό της συνολικής αδράνειας του νέφους, δηλαδή το 1ο ΠΑΡΑΓΟΝΤΙΚΟ ΕΠΙΠΕΔΟ συγκεντρώνει το $\pi_1+\pi_2$ ποσοστό διασποράς του αρχικού νέφους, και δίνει την λιγότερο παραμορφωμένη προβολή του αρχικού νέφους, σε σύγκριση με όλα τα άλλα παραγοντικά επίπεδα.

Επαλήθευση της ορθότητας των υπολογισμών μέχρι αυτό το σημείο

Πρέπει $$\sum_{i=1}^{p}\lambda_i = \text{Ιχνος}\left(X^t X\right),\left(= p\right)$$

(με μια πολύ καλή προσέγγιση)

3.2.1.5 Υπολογισμός των συντεταγμένων των σημείων στους νέους άξονες

Επειδή ο πίνακας X αυτο κατασκευής, είναι τυπικός, η διαγωνοποίηση του επιτυγχάνεται με τον πίνακα των συσχετίσεων C και έτσι:

Ο 1ος παραγοντικός άξονας οπίζεται (κατά διεύθυνση) από το 1ο ιδιοδιάνυσμα του πίνακα των συσχετίσεων που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή του πίνακα, ο 2ος στην αμέσως μικρότερη ιδιοτιμή και ούτω καθ' εξής.

Έτσι, οι συντεταγμένες των αρχικών σημείων στους νέους άξονες, ενώ π.χ. για το σημείο 1 (1η γραμμή) ήταν $(\psi_{11},\psi_{12},...,\psi_{1p})$, τώρα είναι $\left(\psi_{11}\cdot u_{11}+\psi_{12}\cdot u_{12}+\psi_{13}\cdot u_{13}\right)$, όπου $u_1=\begin{pmatrix}u_{11}\\u_{12}\\u_{13}\end{pmatrix}$ δηλαδή

ΟΙ ΣΥΝΤΕΤΑΓΜΕΝΕΣ ΤΩΝ ΠΡΟΒΟΛΩΝ ΤΩΝ ΣΗΜΕΙΩΝ ΣΤΟΥΣ ΠΑΡΑΓΟΝΤΙΚΟΥΣ ΑΞΟΝΕΣ ΔΙΝΟΝΤΑΙ ΑΠΟ ΤΟΝ ΠΙΝΑΚΑ ΓΙΝΟΜΕΝΟ XU (n x p) x (p x p),

δηλαδή, η συντεταγμένη του σημείου ι του αρχικού πίνακα στο νέο άξονα j είναι:

$$d_{ij}=\sum_{j=1}^{p}\psi_{ij}u_{ij}$$

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Daudin J., “Analyse en Composantes Principales” MATHEMATIQUE ET INFORMATIQUE, INSITUT NATIONAL AGRONOMIQUE, PARIS-GRIGNON, Deuxieme annee, 1981-1982.

[2] Cebois P., “L’ Analyse factorielle”, Que sais-je? PRESSES UNIVERSITAIRES DE FRANCE, 1983.

[3] Bouroche J-M., Saporta G., “L’ Analyse des donnees”,Que sais-je? PRESSES UNIVERSITAIRES DE FRANCE, 1983.

[4] Nacache J-P., Chevalier A., Morice V., “Exercises commentes de mathematiques pour d’ analyse statistique des donnees” DUNOD DECISION, Bordeaux, Paris 1981.

[5] Fennetau H., Biales C., “Analyse statistique des donnes-Applications et cas pour le marketing” Enseignement Superieur Tertiaire, Ellipses 1993.

[6] Kim J-O., Mueller C., “Introduction to factor analysis-What it is and how to do it” Series: Quantitavive Applications in the Social Sciences, A SAGE UNIVERSITY PAPER 13, SAGE PUBLICATIONS 1982

[7] Lagarde J., “Initation a l’ analyse der donnees” DUNOD, Bordas, Paris 1983.

[8] Grange D., Lebart L., “ Traitements statistiques des enquetes” DUNOD, Paris 1994.

[9] Γεωργίου Δ., “Σημειώσεις Γ. Μαθηματικών” Μέρος Α. ΣΤΟΙΧΕΙΑ ΓΡΑΜΜΙΚΗΣ ΑΛΓΕΒΡΑΣ Π.Θ. Υπό Έκδοση 1997.

ΚΕΦΑΛΑΙΟ 4

ΜΕΘΟΔΟΙ ΑΥΤΟΜΑΤΗΣ ΟΜΑΔΟΠΟΙΗΣΗΣ (CLUSTERING METHODS)

4.1 Γενικά

Με την γενική έννοια "μέθοδοι ομαδοποίησης" ή ακριβέστερα "δημιουργίας συστάδων (clustering)" εννοούμε το σύνολο των μαθηματικών-στατιστικών αλγορίθμων που ΑΥΤΟΜΑΤΑ διαχωρίζουν τα άτομα κάποιου υποπληθυσμού ή δείγματος σε ομάδες.

Το σύνολο των μεθόδων που έχουν ερευνηθεί με τον σκοπό αυτό είναι αρκετά ευρύ και ποικίλο. Έτσι, έγκειται στην γνώση του ερευνητή καθώς και τις ιδιαίτερες συνθήκες που διέπουν την μελέτη του, η εκλογή της ή των μεθόδων ομαδοποίησης για την επίλυση του προβλήματος.

Ώστόσο, σε κάθε μέθοδο, **δύο** είναι τα απαραίτητα **κριτήρια** για τον σχηματισμό ομάδων:

α) Η **μεγαλύτερη** δυνατή **συνοχή** και ομοιογένεια **στο εσωτερικό** κάθε ομάδας σε συνδιασμό με

β) την **μεγαλύτερη** δυνατόν **διαφοροποίηση** (ετερογένεια) **μεταξύ** των ομάδων. **[βιβλιογρ. 4]**

Αριθμητικά μεταφρασμένα τα κριτήρια αυτά καταλήγουν το μεν πρώτο στους **δείκτες ομοιότητας** το δε δεύτερο στα **μέτρα απόστασης** μεταξύ των ατόμων του δείγματος. Η βασική διαφορά τους είναι ότι οι μεν δείκτες ομοιότητας είναι μεγέθη που παίρνουν τιμές μεταξύ 0 και 1, ενώ τα μέτρα απόστασης μπορούν να πάρουν οποιαδήποτε θετική τιμή. **[βιβλιογρ. 1]**

4.2 Κατηγορίες μεθόδων ομαδοποίησης

Αν και στο μεγαλύτερο μέρος της βιβλιογραφίας συναντάμε ένα διαχωρισμό των μεθόδων ομαδοποίησης σε δύο κυρίως κατηγορίες, θα αναφερθούμε γενικά εδώ και σε κάποιες άλλες μεθόδους που, είτε προκύπτουν ως παράγωγες των παραπάνω,

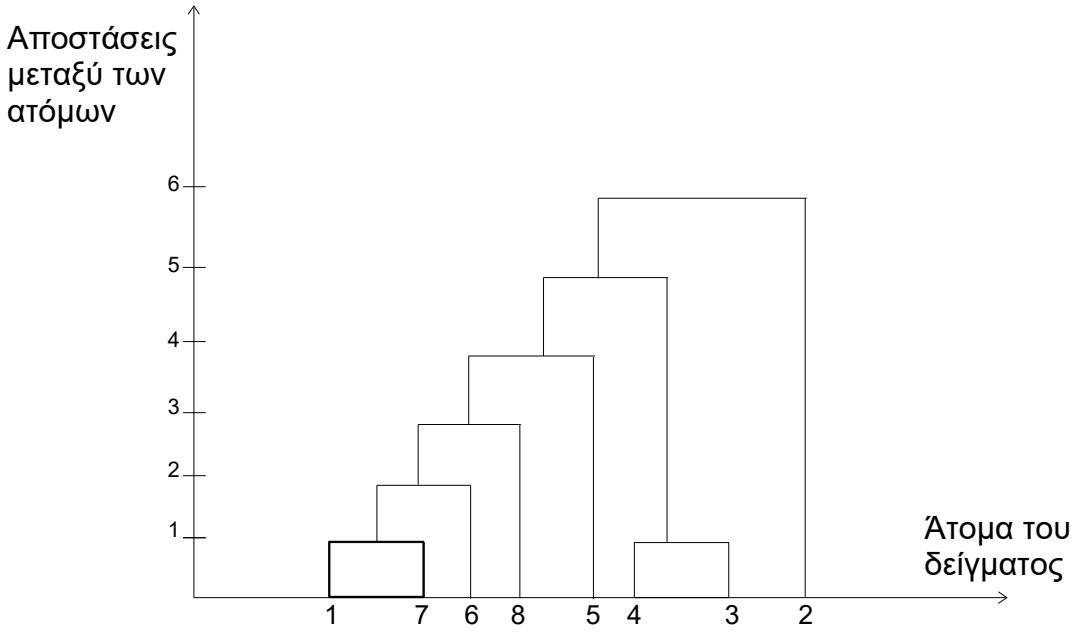
είτε μπορούν να αποτελέσουν ξεχωριστή κατηγορία λόγω των ιδιομορφιών τους ή του πεδίου εφαρμογής τους.

Οι κατηγορίες που υπάρχουν είναι οι εξής:

4.2.1 Οι ιεραρχικές μέθοδοι

1. Οι **ΙΕΡΑΡΧΙΚΕΣ ΜΕΘΟΔΟΙ** παράγουν έναν συνεχόμενο διαμερισμό σε ομάδες, αρχίζοντας από τις πιο χονδρικές (λίγες και με μεγάλο αριθμό ατόμων) και διαδοχικά φτάνοντας στις πιο μικρές και λεπτομερείς ομάδες. **[βιβλιογρ. 5]**

Η διαγραμματική μορφή των μεθόδων αυτών έχει συνήθως την μορφή ορθογωνίου δενδρογράμματος π.χ.



Με τον τρόπο που δημιουργούνται οι κλάσεις που αναφέρθηκαν παραπάνω, διακρίνουμε τις ιεραρχικές μεθόδους σε μεθόδους ΣΥΜΠΤΥΞΗΣ από μικρότερες σε μεγαλύτερες ομάδες και σε μεθόδους ΔΙΑΙΡΕΣΗΣ από μεγαλύτερες σε μικρότερες ομάδες.

Δύο είναι οι απαραίτητες προεργασίες στις μεθόδους ΣΥΜΠΤΥΞΗΣ:

1ο. Η **εκλογή** ενός **μέτρου απόστασης μεταξύ των ατόμων** του δείγματος στο οποίο εφαρμόζουμε την ταξινόμηση και

2ο. Η **εκλογή** ενός **κριτηρίου προσάρτησης** ενός **ατόμου** σε μια **ομάδα** (ή ενός κριτηρίου απόστασης μεταξύ ατόμου και ομάδας).

Οι Fennetau H. και Bialer C. **[βιβλιογρ. 6]** προτείνουν το παρακάτω οργανόγραμμα που περιγράφει την μεθοδολογία σε βήματα μιας μεθόδου (σύμπτυξης) ομαδοποίησης:

1ο. Αρχικός διαχωρισμός του δείγματος σε τόσες ομάδες όσα και τα άτομα του δείγματος

2ο. Υπολογισμός του πίνακα των αποστάσεων μεταξύ των (ομάδων) ατόμων

3ο. Αφού καθοριστεί το μέτρο απόστασης μεταξύ των ομάδων (ατόμων), έχουμε την δημιουργία ομάδων που βρίσκονται στην κοντινότερη απόσταση ως προς το καθορισμένο μέτρο

4ο. Υπολογισμός των αποστάσεων μεταξύ των ομάδων στην ομαδοποίηση του βήματος 3

5ο. Σύμπτυξη των κοντινότερων ομάδων (ως προς την απόσταση που θεωρήθηκε)

6ο. Εαν ο αριθμός των ομάδων είναι 1 σημαίνει το τέλος της μεθόδου. Εάν είναι μεγαλύτερος από 1 τότε έχουμε επανάληψη από το βήμα 4 και μετά.

Από τις ιεραρχικές μεθόδους ομαδοποίησης αναφέρουμε με παραδείγματα στη συνέχεια δύο από τις γνωστότερες μεθόδους:

4.2.1.1 Η μέθοδος του πλησιέστερου γειτονικού σημείου ή μέθοδος της απλής σύνδεσης (Nearest Neighbour ή Single Linkage Method)

Μερικοί συγγραφείς **[βιβλιογρ. 2]** μετασχηματίζουν τον πίνακα των αρχικών δεδομένων x_{ij} i=1,...,N (αριθμός ατόμων) και j=1,...p (αριθμός μεταβλητών) σε πίνακα τυπικών δεδομένων

$$z_{ij} = (x_{ij} - \bar{x}_j) \Big/ S_j$$

και στην συνέχεια δημιουργούν τον πίνακα των ευκλείδειων αποστάσεων μεταξύ των τοπικών μεγεθών που προκύπτουν έτσι, δηλαδή:

$$d_{ih} = \left[\sum_{j=1}^{p} (z_{ij} - z_{hj})^2 \right]^{\frac{1}{2}}$$

οπότε καταλήγουν σ' έναν πίνακα αποστάσεων της παρακάτω μορφής:

$$D = \begin{bmatrix} 0 & d_{12} & \dots & d_{1N} \\ d_{21} & 0 & & \\ \vdots & & \ddots & \\ d_{N1} & & & 0 \end{bmatrix}$$

Αυτό γίνεται στις περισσότερες των περιπτώσεων όπου οι μεταβλητές κολώνες του πίνακά μας δεν είναι μετρημένες στις ίδιες μονάδες μέτρησης.

Όταν όμως οι μεταβλητές ανήκουν στο ίδιο σύστημα μέτρησης μονάδων, τότε χρησιμοποιούμε τον τύπο της απόστασης που αναφέρθηκε παραπάνω στα αρχικά δεδομένα δηλαδή:

$$d_{ih} = \left[\sum_{j=1}^{p} (z_{ij} - z_{hj})^2 \right]^{\frac{1}{2}}$$

Παράδειγμα

Έστω ότι από τον αρχικό πίνακα δεδομένων, που οι μεταβλητές είναι μετρημένες στις ίδιες μονάδες μέτρησης και που είναι της μορφής

ΜΕΤΑΒΛΗΤΕΣ / ΑΤΟΜΑ	1	2	3
1	2	3	1
2	6	3	4
3	.	.	.
4	.	.	.

καταλήγουμε στον πίνακα αποστάσεων 4 x 4 (επειδή έχουμε 4 άτομα), ο οποίος είναι τριγωνικά συμμετρικός και είναι ο εξής:

$$D_1 = \begin{bmatrix} 0 & 5 & 2 & 3 \\ 5 & 0 & 1 & 2 \\ 2 & 1 & 0 & 3 \\ 3 & 2 & 3 & 0 \end{bmatrix}$$

(Η απόσταση π.χ. 5 μεταξύ των ατόμων 1 και 2 βρέθηκε από τον αρχικό πίνακα γιατι

$$d_{12} = \left((6-2)^2 + (3-3)^2 + (4-1)^2\right)^{\frac{1}{2}} = \sqrt{25} = 5$$

Στο πρώτο βήμα της μεθόδου, τα άτομα 2 και 3 συγχωνεύονται σε μια ομάδα αφού, η απόστασή τους είναι η μικρότερη στον πίνακα D_1 δηλαδή $d_{23}=1$.

Η απόσταση της ομάδας 23 που τώρα θεωρείται σαν ένα άτομα από τα άλλα άτομα του πίνακα βρίσκεται ως εξής:

$$d_{(23)1} = \min\left\{d_{21,} d_{31}\right\} = \min\{5,2\} = 2$$

$$d_{(23)4} = \min\left\{d_{24}, d_{34}\right\} = \min\{2,3\} = 2$$

Έτσι, ο πίνακας D_1 μετασχηματίζεται στον

$$D_2 = \begin{bmatrix} 0 & 2 & 3 \\ 2 & 0 & 2 \\ 3 & 2 & 0 \end{bmatrix}$$

όπου $d_{14}=3$ από τον D_1.

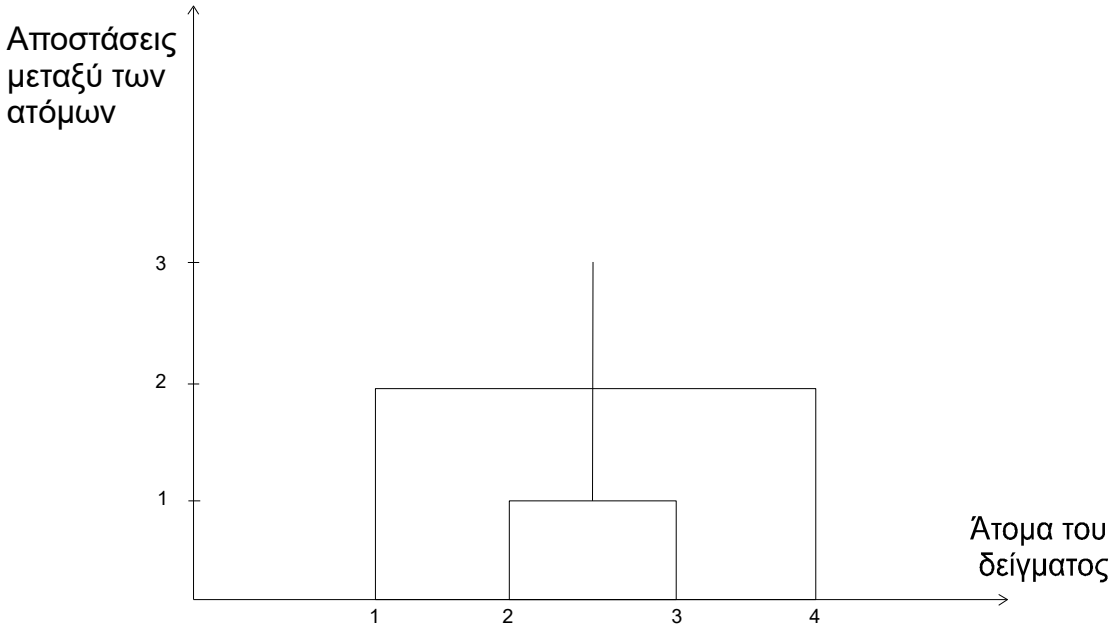
Η μικρότερη απόσταση τώρα είναι η $d_{(23)4}$ και $d_{(23)1}=2$ άρα τα άτομα (23) και 1 συγχωνεύονται εκ νέου όπως και τα (23) και 4.

Έχουμε δηλαδή στο συγκεκριμένο παράδειγμα μία προσάρτηση των σημείων 1 και 4 στην ομάδα (23) συγχρόνως, οπότε δημιουργείται και η τελική ομάδα.

ΠΡΟΣΟΧΗ !!

Οι διαδοχικοί κλάδοι του δενδρογράμματος να σχεδιάζονται διαδοχικά και κάθε φορά που δημιουργείται νέος πίνακας αποστάσεων.

Το δενδρόγραμμα της παραπάνω ιεραρχικής ταξινόμησης είναι το παρακάτω:



Φυσικά, το δενδρόγραμμα τελειώνει ακριβώς μόλις όλα τα άτομα ή ομάδες προσαρτηθούν σε μία και μοναδική ομάδα.

4.2.1.2 Η εύκαμπτη μέθοδος των Lance και Williams (Lance & Williams, Flexible Method)

Ένας γενικευμένος τύπος απόστασης μεταξύ μιας ομάδας ι και μιας ομάδας (κλ) που προήλθε από την ένωση των ομάδων (στοιχείων) κ και λ, βρέθηκε από τους Lance και Williams και είναι ο

$$d_{i(\kappa\lambda)} = a_\kappa d_{i\kappa} + a_\lambda d_{i\lambda} + \beta d_{\kappa\lambda} + \gamma \left| d_{i\kappa} - d_{i\lambda} \right| \qquad \textbf{(1)}$$

όπου οι συντελεστές α_κ, α_λ, β και γ παίρνουν τιμές ανάλογα με την μέθοδο που εφαρμόζεται.

Δυστυχώς, ο παραπάνω τύπος **δεν εφαρμόζεται** στις μεθόδους που χρησιμοποιούν **δείκτες ομοιότητας**.

Οι Lance και Williams χρησιμοποιούν τις εξής 4 συνθήκες που πρέπει να πληρούν οι 4 συντελεστές α_κ, α_λ, β και γ

$\alpha_\kappa+\alpha_\lambda+\beta=1$

$\alpha\kappa=\alpha\lambda$

$\beta<1$

$\gamma=0$

και πιο συγκεκριμένα, για το β την τιμή -0,25, οπότε η παραπάνω γενικευμένη απόσταση γίνεται

$$d_{i(\kappa\lambda)} = 0{,}625\ d_{i\kappa} + 0{,}625 d_{i\lambda} - 0{,}25 d_{\kappa\lambda}$$

Οι παραπάνω συντελεστές για μερικές από τις πιο γνωστές μεθόδους είναι:

α) Στην μέθοδο **του πλησιέστερου γειτονικού σημείου**

$\alpha_\kappa = \alpha_\lambda = \frac{1}{2}$, $\beta = 0$, $\gamma = -\frac{1}{2}$

Έτσι, στο συγκεκριμένο παράδειγμα που αναφέρθηκε σ' αυτήν την μέθοδο και επειδή $d_{(\kappa\lambda)i} = d_{i(\kappa\lambda)}$ ή για ι=1 (κλ)=(23) θα έχουμε:

$$d_{(23)1} = d_{1(23)} = \frac{1}{2}\left(d_{12} + d_{13}\right) - \frac{1}{2}\left|d_{12} - d_{13}\right| = \frac{1}{2}(5+2) - \frac{1}{2}|5-2| = \frac{7}{2} - \frac{3}{2} = 2$$

και για ι = 4 (κλ)=(23) και έχουμε:

$$d_{(23)4} = d_{4(23)} = \frac{1}{2}\left(d_{42} + d_{43}\right) - \frac{1}{2}\left|d_{42} - d_{43}\right| = \frac{1}{2}(2+3) - \frac{1}{2}|2-3| = \frac{5}{2} - \frac{1}{2} = 2$$

αποτελέσματα που συμφωνούν με αυτά που υπολογίστηκαν σύμφωνα με τον αρχικό ορισμό της απόστασης, έτσι όπως ορίστηκε στην μέθοδο του nearest neighbouhr (πλησιέστερου γειτονικού σημείου).

β) Στη μέθοδο της **μέσης τιμής των ομάδων** (group average)

$$a_\kappa = \frac{n_\kappa}{n_\kappa + n_\lambda}, a_\lambda = \frac{n_\lambda}{n_\kappa + n_\lambda}\textbf{, } \boldsymbol{\beta = \gamma = 0} \qquad \textbf{(2)}$$

όπου n_i είναι η συχνότητα (βάρος ή αριθμός των στοιχείων) της ομάδας i. Φυσικά, στην πρώτη ομαδοποίηση όπου κάθε στοιχείο συμμετέχει μόνο του $n_i = 1$ $\forall i$.

Έτσι, στο αρχικό παράδειγμα, εάν επιχειρήσουμε μία ομαδοποίηση με την μέθοδο της μέσης τιμής, θα έχουμε:

Αρχικός πίνακας:

$$D = \begin{bmatrix} 0 & 5 & 2 & 3 \\ 5 & 0 & 1 & 2 \\ 2 & 1 & 0 & 3 \\ 3 & 2 & 3 & 0 \end{bmatrix}$$

Τα στοιχεία 2 και 3 συγχονεύονται γιατί έχουν την μικρότερη μεταξύ τους απόσταση και ίση με 1 και προκύπτει όπως και πριν ο πίνακας D_1

$$D_1 = \begin{bmatrix} 0 & 3{,}5 & 3 \\ 3{,}5 & 0 & 2{,}5 \\ 3 & 2{,}5 & 0 \end{bmatrix}$$

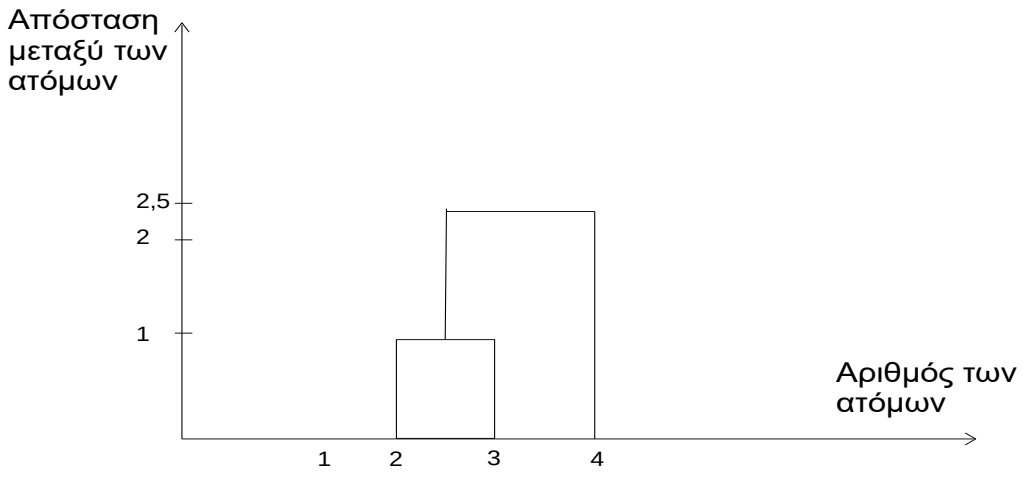
όπου οι αποστάσεις $d_{1(23)}$ και $d_{4(23)}$ υπολογίστηκαν από τον γενικό τύπο (1) των Lance και Williams και από τους τύπους (2) της μεθόδου της μέσης τιμής των ομάδων ως εξής:

για ι = 1, κ = 2 και λ = 3

$$d_{1(23)} = a_2 d_{12} + a_3 d_{13} = \frac{n_2}{n_2 + n_3} d_{12} + \frac{n_3}{n_3 + n_2} d_{13} = \frac{1}{2} \cdot 5 + \frac{1}{2} \cdot 2 = \frac{7}{2} = 3{,}5$$

για ι = 4, κ = 2 και λ = 3

$d_{4(23)} = a_2 d_{42} + a_3 d_{43} = a_2 d_{42} + a_3 d_{43} = \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 3 = \frac{5}{2} = 2{,}5$ οπότε το δενδρόγραμμα έχει ήδη διαμορφωθεί ως εξής:

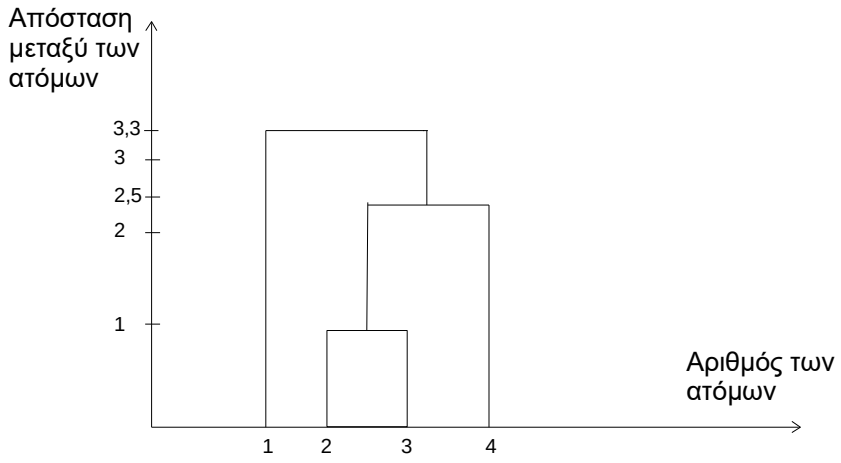


Αμέσως μετά προκύπτει ο πίνακας D_2

$$D_2 = \begin{bmatrix} 0 & 3{,}33 \\ 3{,}33 & 0 \end{bmatrix} \quad \text{όπου}$$

$$d_{[(23)4]} = \frac{n_{23}}{n_{23}+n_4} d_{1(23)} + \frac{n_4}{n_{23}+n_4} d_{14} = \frac{2}{2+1} \cdot 3{,}5 + \frac{1}{1+2} \cdot 3 = \frac{10}{3} = 3{,}33.....$$

και τελικά, το δενδρόγραμμα με αυτήν την **ιεραρχική μέθοδο σύμπτυξης**, διαμορφώνεται όπως παρακάτω



γ. Στην μέθοδο του **Ward**

όπου οι συντελεστές του γενικού τύπου (1) των Lance και Williams είναι

$$a_\kappa = \frac{n_\iota + n_\kappa}{n_\iota + n_\kappa + n_\lambda}, a_\lambda = \frac{n_\iota + n_\lambda}{n_\iota + n_\kappa + n_\lambda} \qquad \text{και}$$

$$\beta = \frac{-n_\iota}{n_\iota + n_\kappa + n_\lambda}, \gamma = 0$$

4.2.2 Οι ΜΗ ΙΕΡΑΡΧΙΚΕΣ μέθοδοι ομαδοποίησης (Nonierarhical clustering methods)

Σ' αυτές, ο αριθμός των ομάδων (ή συστάδων-cluster) είναι δυνατόν να είναι προκαθορισμένος πριν την εφαρμογή της μεθόδου ή να προσδιορίζεται κατά την διαδικασία της ομαδοποίησης.

Εδώ, δεν υπάρχει η ανάγκη δημιουργίας του πίνακα αποστάσεων μεταξύ των ατόμων όπως γίνεται στις ιεραρχικές μεθόδους.

Το βασικό πλεονέκτημα των μη ιεραρχικών μεθόδων είναι ότι επειδή δεν υπάρχει ανάγκη αποθήκευσης των αρχικών δεδομένων καθ' όλη την διάρκεια εκτέλεσης της μεθόδου (συνήθως από ηλ. υπολογιστές) μπορούν να επεξεργαστούν πολύ μεγαλύτερα σε όγκο δεδομένα από ότι οι άλλες μέθοδοι, λόγω μη αναγκαιότητας μεγάλου μεγέθους ενεργής μνήμης του ηλεκτρονικού υπολογιστή (RAM: Random Access Memory).

Η πιο γνωστή από αυτές τις μεθόδους είναι:

4.2.2.1 Η μέθοδος ομαδοποίησης γύρω από κινητά κέντρα [βιβλιογρ. 5] (K-Means method)

Σύμφωνα με την μέθοδο αυτή ομαδοποιούνται όλα τα άτομα γύρω από κ αυθαίρετα εκλεγμένα άτομα-κέντρα.

Κάθε μία από τις ομάδες αποτελείται από εκείνα τα άτομα τα οποία βρίσκονται πιο κοντά στο κέντρο της σύμφωνα με την απόσταση που ορίστηκε, και είναι συνήθως η Ευκλείδια.

Έτσι, οι αρχικές αυτές ομάδες είναι αυθαίρετες, αφού δημιουργούνται γύρω από αυθαίρετα εκλεγμένα κέντρα. Η μέχρι εδώ διαδικασία μπορεί να αντικατασταθεί και με τελείως αυθαίρετη επιλογή ομάδων, χωρίς την χρήση ατόμων κέντρων.

Στην συνέχεια, υπολογίζεται το κέντρο βάρους (η μέση τιμή) των ομάδων αυτών και οι αποστάσεις των σημείων από τα νέα κέντρα των ομάδων.

Έτσι, αλλάζοντας πάλι την σύσταση των ομάδων, προσαρτούνται σε κάθε νέο κέντρο-βάρους τα κοντινότερα ως προς την απόσταση που ορίστηκε, σημεία-άτομα.

Υπολογίζονται, πάλι, οι αποστάσεις των ατόμων από τα Κ.Β. και προσαρτούνται σε κάθε Κ.Β. τα κοντινότερα σημεία.

Με αυτόν τον τρόπο, υπάρχουν ολοένα και λιγότερα σημεία-άτομα που αλλάζουν ομάδα σε κάθε επανάληψη του αλγορίθμου προσάρτησης.

Η μέθοδος συγλίνει ή κατά σύμβαση όταν δεν αλλάζουν πια ομάδες κ % άτομα ή απόλυτα όταν όλα τα άτομα σταθεροποιηθούν σε κάποια ομάδα.

Παράδειγμα:

Έστω ότι έχουμε τα παρακάτω δεδομένα:

ΜΕΤΑΒΛΗΤΕΣ / ΑΤΟΜΑ	x_1	x_2
a	3	-2
β	-1	2
γ	2	4
δ	2	2

και θέλουμε να δημιουργήσουμε 2 ομάδες.

Η **διαδικασία** που ακολουθεί είναι σύμφωνα με το στατιστικό λογισμικό SPSS 9.0 και τα αποτελέσματα δίνονται υπό μορφή πινάκων και εκτίθενται οι αντίστοιχοι υπολογισμοί.

Διαδικασία στο SPSS 9.0

Analyse

Classify

K-means cluster

Variables : Var00001

Var00002

Options

- **Initial cluster centers**
- **Cluster information for each case**

Α Π Ο Τ Ε Λ Ε Σ Μ Α Τ Α

1ος ΠΙΝΑΚΑΣ

Initial Cluster Centers

	Cluster	
	1	2
VAR00001	3.00	2.00
VAR00002	-2.00	4.00

Η επιλογή των αρχικών κέντρων των clusters γίνεται τυχαία. Ετσι επιλέγονται γιά το 1ο cluster το 1ο στοιχείο (3,-2) και για το 2ο cluster το 3ο στοιχείο (2,4).

2ος ΠΙΝΑΚΑΣ

Iteration History[a]

	Change in Cluster Centers	
Iteration	1	2
1	.000	1.667
2	.000	.000

a. Convergence achieved due to no or small distance change. The maximum distance by which any center has changed is .000. The current iteration is 2. The minimum distance between initial centers is 6.083.

Για να βρεθεί ο παραπάνω πίνακας των επαναλήψεων βρίσκονται οι ευκλείδειες αποστάσεις κάθε σημείου από τα αρχικά επιλεγμένα κέντρα. Ονομάζοντας dij την απόσταση του I-στοιχείου από το j-cluster έχουμε:

1η ΕΠΑΝΑΛΗΨΗ

$d^2_{11}= (3-3)^2+[-2-(-2)]^2=0$ άρα $d_{11}=0$

$d^2_{12}= (3-2)^2+(-2-4)^2=1+36$ άρα $d12=\sqrt{37}$

$\Rightarrow d_{11}<d_{12}$

Άρα το 1ο στοιχείο ανήκει στο 1ο cluster.

$d^2_{21}=(-1-3)^2+[2-(-2)]^2=16+16=32$ άρα $d_{21}=\sqrt{32}$

$d^2_{22}=(-1-2)^2+(2-4)^2=9+4=13$ άρα $d_{22}=\sqrt{13}$

$\Rightarrow d_{22}<d_{21}$

Άρα το 2ο στοιχείο ανήκει στο 2ο cluster.

Ομοίως $0=d_{32}<d_{31}=\sqrt{37}$

Άρα το 3ο στοιχείο ανήκει στο 2ο cluster.
Ομοίως $2=d_{42}<d_41=\sqrt{17}$
Άρα το 4ο στοιχείο ανήκει στο 2ο cluster.

ΕΤΣΙ: Το κέντρο του 1ου cluster που έχει μόνο το 1ο στοιχείο παραμένει αμετάβλητο δηλ. παραμένει το 1ο στοιχείο. Το κέντρο του 2ου cluster είναι στην Var00001: (-1+2+2)/3=1 και στην Var00002: (2+4+2)/3=8/3=2,667 δηλ. το κέντρο του 2ου cluster είναι το σημείο $B_2(1,2,667)$ ενώ το αρχικό επιλεγμένο ήταν το $B_1(2,4)$ 4ος πίνακας.

Η αλλαγή του κέντρου του cluster που μας δίνει ο 2ος πίνακας είναι η απόσταση B_1B_2 δηλ. $d^2=(2-1)^2+(4-2,667)^2= 1^2+1,33^2=1+1,777=2,77$ άρα $d=\sqrt{2,77}=1,667$.

2η ΕΠΑΝΑΛΗΨΗ

Βρίσκουμε πάλι τις αποστάσεις όλων των σημείων από τα καινούργια κέντρα των cluster μετά την 1η επανάληψη.

$0=d^2_{11}<d^2_{12}=(3-1)^2+(-2-2,667)^2$
Άρα το 1ο στοιχείο παραμένει στο 1ο cluster.

Ομοίως $d_{21}=\sqrt{32}>d22=\sqrt{(-1-1)^2+(2-2,667)^2}$
Άρα το 2ο στοιχείο παραμένει στο 2ο cluster και

Το 3ο στοιχείο παραμένει στο 2ο cluster
Το 4ο στοιχείο παραμένει στο 2ο cluster

Η ελάχιστη απόσταση των αρχικών κέντρων των cluster ήταν:

$d^2=(2-3)^2+[4-(-2)]^2=37$ άρα $d=\sqrt{37}=6,083$

Στο τέλος της 1ης επανάληψης εξηγήθηκε η δημιουργία του μεθεπόμενου 4ου πίνακα.
3ος ΠΙΝΑΚΑΣ

Cluster Membership

Case Number	Cluster	Distance
1	1	.000
2	2	2.108
3	2	1.667
4	2	1.202

4ος ΠΙΝΑΚΑΣ

Final Cluster Centers

	Cluster	
	1	2
VAR00001	3.00	1.00
VAR00002	-2.00	2.67

Ο 3ος πίνακας μας δίνει σε ποιο cluster ανήκει κάθε στοιχείο και την απόσταση του από το κέντρο του cluster που ανήκει

Π.χ. το 2ο στοιχείο ανήκει στο 2ο cluster και
$d^2_{22}=(-1-1)^2+(2-2,67)^2=22+0,672=2,108$ άρα $d_{22}=2,108$

5ος ΠΙΝΑΚΑΣ

Distances between Final Cluster Centers

Cluster	1	2
1		5.077
2	5.077	

Ο 5ος πίνακας μας δίνει τη απόσταση μεταξύ των 2 clusters
$d^2_{A2B2}=(3-1)^2+(-2-2,67)^2=4+21,8089$ άρα $d_{A2B2}=5,077$

6ος ΠΙΝΑΚΑΣ

Number of Cases in each Cluster

Cluster	1	1.000
	2	3.000
Valid		4.000
Missing		.000

Ο 6ος πίνακας μας δίνει πληροφορίες για την τελική ταξινόμηση.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Everitt Br., “Cluster Analysis” Halsted Press a Division of J. Wiley & Sons, New York, 1981.

[2] Morisson D., “Multivariate Statistical Methods” McGraw-Hill Series in Probability and Statistics, 1990.

[3] Krzanowski W., “Principles of Multivariate Analysis. A user perpective” Oxford Science Publications,1990.

[4] Danzart M., “Classification Automatique” Institut National Agronomique, Paris-Grignon 1981-82.

[5] Bouroche J.M., Saporta G., “L’ analyse des donnees” P.U.F. 2^{eme} edition, Paris 1983.

[6] Fennetaeu H., Biales C., “Analyse Statistique des donnees”, Ellipses, 1993

[7] Johnson R., Wichern D., “Applied Multivariate Statistical Analysis” , Prentice-Hall, 1982.

[8] MacQueen J.B., “Some Methods for Classification and Analysis of Multivariate Observations”, Proccedings of 5^{th} Berkeley Symposium on Mathematical Statistics and Probability 1, Calif: University of California Press (1967), 281-297