

ΕΙΣΑΓΩΓΗ

Περίπου δύο δεκαετίες πριν, οι προσωπικοί υπολογιστές, ξεκίνησαν να αποκτούν ευρεία διάδοση. Πραγματοποιούσαν δε, σε ελάχιστα λεπτά, υπολογισμούς που για τα υλισμικά της δεκαετίας του '60, απαιτούνταν πολυήμερες εκτελέσεις. Οι μέθοδοι της πολυμεταβλητής ανάλυσης δεδομένων αρχικά υιοθετήθηκαν από μικρό αριθμό στατιστικών. Κατά τα τελευταία όμως χρόνια η αποδοχή τους έχει αυξηθεί σημαντικά. Μεγάλο μέρος αυτής της αποδοχής οφείλεται στην αξιοσημείωτη διαθεσιμότητα των ηλεκτρονικών υπολογιστών και την έκρηξη του πλήθους των δεδομένων για τα οποία άλλες περιγραφικές στατιστικές μέθοδοι δεν εμφανίζουν ανάλογη χρηστικότητα. Επιπλέον, νέες υπολογιστικές εφαρμογές έδωσαν τη δυνατότητα για ανάπτυξη πρωτοπόρων γραφικών μεθόδων που υποστηρίζουν τη φιλοσοφία της ανάλυσης δεδομένων κατά τρόπο εξαιρετικά αποτελεσματικό. Η εφαρμογή της ανάλυσης δεδομένων σε μεγάλα σύνολα δεδομένων θα ήταν δυσχερής, αν όχι αδύνατη, χωρίς την υποστήριξη των σύγχρονων ηλεκτρονικών υπολογιστών.

Πέρα από τα γνωστά εμπορικά πακέτα στατιστικής επεξεργασίας δεδομένων (MINITAB, SPSS, S-Plus, SAS) - που από τη δεκαετία του '80 και μετά συμπεριέλαβαν και ρουτίνες πολυμεταβλητής ανάλυσης δεδομένων- πολλοί ερευνητές έχουν αναπτύξει εξειδικευμένα λογισμικά πακέτα διαδραστικού περιβάλλοντος γραφικών που εφαρμόζουν τις μεθόδους αυτές σε ευρύ επιστημονικό πεδίο.

Στον ελληνικό χώρο, αναφέρουμε το λογισμικό ανάλυσης δεδομένων MAD του Δ. Καραπιστόλη (1996) και το λογισμικό S-Pro του συγγραφέα του παρόντος (Κουτσουπιάς 2002), το Chic Analysis του Α. Μάρκου (2010), καθώς και το δικτυακό πρόγραμμα αξιοποίησης τεχνικών ανάλυσης δεδομένων PRAXITELE του Α. Καράκου (2012).

Στο παρόν συγγραφικό εγχείρημα περιγράφονται οι δυνατότητες, οι συναρτήσεις, τα πακέτα¹ κατασκευής και επεξεργασίας πινάκων συχνοτήτων και αυτά της υλοποίησης μεθόδων της Ανάλυσης Δεδομένων και συγκεκριμένα της Ανάλυσης Αντιστοιχιών και Ταξινόμησης κατ' Αύξουσα Ιεραρχία για την επεξεργασία πινάκων συχνοτήτων και τη βοήθεια στην ερμηνεία των αποτελεσμάτων

¹ R packages - βιβλιοθήκες κώδικα της R.

των μεθόδων αυτών, μέσω της γλώσσας R. Αναφορές στις παραπάνω εφαρμογές γίνονται και από τους Κουτσουπιά & Παπαδημητρίου (2018).

Εκτός από την περιγραφή των χαρακτηριστικών και των ιδιοτήτων των πινάκων συχνοτήτων, το ανά χείρας πόνημα περιέχει σύντομες αναφορές και εφαρμογές της Παραγοντικής Ανάλυσης Αντιστοιχιών και της Αυτόματης Ταξινόμησης κατ' Αύξουσα Ιεραρχία και τις στατιστικές μεθόδους Πολυμεταβλητής Ανάλυσης Δεδομένων που συνιστούν δύο σημαντικά εργαλεία τόσο για την συνοπτική αναπαράσταση, όσο και για την ερμηνεία των παραπάνω δεδομένων. Οι μέθοδοι αυτές κατά τον καθηγητή Γ. Παπαδημητρίου (2007) αποτελούν σύνθεση και συστηματοποίηση μεθόδων της περιγραφικής στατιστικής.

Άλλες μέθοδοι ανάλυσης δεδομένων είναι η Ανάλυση σε Κύριες Συνιστώσες (ΑΚΣ)² για τη μελέτη ποσοτικών μεταβλητών (Hotelling 1933) που χτίζονται ανάλογα με κάποιες συντεταγμένες σε ένα σύνολο αξόνων (*composantes principales*) αντιπροσωπευτικά επιλεγμένων (Bertier & Bouroche 1975, Παπαδημητρίου 1994), η Ανάλυση Πολλαπλών Αντιστοιχιών, γνωστή και σαν Homogeneity Analysis (Gifi 1990) ή Dual Scaling (Nishisato 1980) για την μελέτη ενός συνόλου παρατηρήσεων ως προς ένα σύνολο κατηγορικών μεταβλητών (Escofier & Pagès 1988 & 1992), η Διακρίνουσα Ανάλυση (Porebski 1966, Tatsuoka 1970) που αποτελεί σύνολο μεθόδων για την ταξινόμηση ενός πλήθους παρατηρήσεων σε προϋπάρχουσες κατηγορίες, καθώς και η Πολυδιάστατη Κλιμάκωση (Multidimensional Scaling) που πρωτοπαρουσιάστηκε από τον Torgerson (1952) και εξελίχθηκε από τους Shepard (1961) και Kruskal (1963).

Εκτενής ιστορική αναδρομή της προέλευσης και της εξέλιξης των μεθόδων της ανάλυσης δεδομένων γίνεται από τον καθηγητή J.-P. Benzécri (1982a), τον κυριότερο θεωρητικό στατιστικό, που από την δεκαετία του '70 και μετά θεμελίωσε με τα συγγράμματά του (Benzécri 1973 & 1980) την μαθηματική βάση των μεθόδων. Τις ανέδειξε από την αφάνεια και πέτυχε να βρεθούν σε θέση αντάξια της σημαντικότητάς τους. Η Παραγοντική Ανάλυση Αντιστοιχιών και η Ταξινόμηση παρουσιάζονται συνοπτικά από τους J.-P. Benzécri (1980, 1981, 1986), J.-P., Benzécri & Maiti (1992) και F. Benzecri (1980a, 1980b). Στην Ελληνική βιβλιογραφία οι δυο μέθοδοι περιγράφονται και από τους Παπαδημητρίου (1994,

² δες και στην ενότητα Η ΑΚΣ με το FactoMineR (Κεφ. 3).

2007), Αθανασιάδη (1995), Καραπιστόλη (1999), Κουτσοπιιά (1999, 2005), Μπεχράκη (1999), Μάρκο (2006) και Μενεξέ (2006).

Με τις μεθόδους της ανάλυσης δεδομένων ο ερευνητής ξεκινά από τα δεδομένα και ερευνά για ομαδοποιήσεις, σχέσεις ή τάσεις που μπορούν αυτά να εμπεριέχουν, δίνει έμφαση στη γραφική απεικόνιση τους και ανακαλύπτει απλές, αλλά λειτουργικές περιγραφές των χαρακτηριστικών των δεδομένων. Ο αναλυτής δεδομένων αναζητά ενδείξεις τυχόν εξαιρετικών περιπτώσεων αναγνωρίζοντας ότι τα δεδομένα συχνά είναι ετερογενή και έχει στη διάθεσή του εργαλεία για τον εντοπισμό και την απομόνωση ιδιόμορφων τιμών και ξεχωριστών υποομάδων.

Η χαρακτηριστική ιδιότητα που ξεχωρίζει τις μεθόδους αυτές από την κλασική στατιστική ανάλυση είναι η απουσία τεχνικών υποθέσεων. Κατά τον καθηγητή Μ. Μειμάρη, ο ερευνητής αξιοποιεί τις μεθόδους αυτές με "πίστη στην αλήθεια των δεδομένων" (Μειμάρης, 2002). Η ανάλυση δεδομένων αφήνει στον ερευνητή την φροντίδα και την υπευθυνότητα να εξάγει τις ερμηνείες των γεγονότων και τις συνέπειές τους (Lagarde, 1983). Ο καθηγητής J.-P. Benzécri (1980, t.1) αναφέρει σχετικά: *"κάτω από τον όρο μαθηματική στατιστική, κάποιον συγγραφείς έχτισαν μια πομπώδη επιστήμη, πλούσια σε υποθέσεις που ποτέ στην πράξη δεν ικανοποιούνται"*. Ένας από τους σημαντικότερους αγγλόφωνους στατιστικούς, ο Tuckey (1977), θεωρεί ότι η ανάλυση δεδομένων συνιστά φιλοσοφία που επιστρέφει την στατιστική στους αρχικούς της στόχους, όπου η αξία της στατιστικής περιγραφής των δεδομένων με χρήση σύγχρονων μεθόδων είναι εξέχουσας σημασίας. Υπογραμμίζει αναφορικά με τη χρησιμότητα των μεθόδων αυτών στη στατιστική έρευνα ότι: *"Η περιγραφική ανάλυση δεδομένων δεν είναι σε θέση να αποδώσει ολόκληρη την εικόνα, αλλά τίποτε άλλο δεν μπορεί να αποτελέσει τον ακρογωνιαίο λίθο, το πρώτο βήμα"* και επισημαίνει ότι *"η επιστήμη και τέχνη της ανάλυσης δεδομένων αφορά τη διαδικασία εκμάθησης από ποιοτικές καταγραφές εμπειριών"* (Tukey & Wilk, 1966). Κατά τον Good (1990) η περιγραφική ανάλυση δεδομένων ασχολείται με πραγματικές παρατηρήσεις και όχι με δεδομένα που έχουν συγκεντρωθεί με βάση κάποιο προσχεδιασμένο πείραμα, όπου οι παρατηρήσεις εξαρτώνται από γνωστές και ελεγχόμενες μεταβλητές.

Σκοπός της ανάλυσης δεδομένων είναι να αναπτύξει ένα μοντέλο –κάθε φορά εξαρτώμενο μόνο από τα ίδια τα δεδομένα - των επιρροών που σχηματοποιούν γραφικά τα δεδομένα. Οι Lubinsky

& Pregibon (1988) αναφέρουν ότι αυτή η διαδικασία απορρέει από τον κύκλο αποτύπωσης/ενέργειας, όπου ο ερευνητής, μετά την εφαρμογή των μεθόδων αυτών, εξετάζει τη γραφική αποτύπωση των δεδομένων και επιλέγει να ενεργήσει με βάση την ερμηνεία αυτής της αποτύπωσης. Σε κάθε κύκλο αυτής της διαδικασίας ο αναλυτής αποκτά περισσότερες πληροφορίες και εστιάζει το ενδιαφέρον σε νέα ή βαθύτερα σημεία της έρευνας. Περιπτώσεις εφαρμογών των μεθόδων ανάλυσης δεδομένων δημοσιεύονται και στα τεύχη του περιοδικού Τετράδια Ανάλυσης Δεδομένων της Ελληνικής Εταιρίας Ανάλυσης Δεδομένων.

Όπως προαναφέραμε, στο πρώτο μέρος του βιβλίου περιγράφονται οι κύριες μέθοδοι και εφαρμογές τους από τις Ανθρωπιστικές, Κοινωνικές και Οικονομικές επιστήμες πάνω σε πραγματικά δεδομένα με τη χρήση του πακέτου **FactoMineR** (Le et al. 2008), στο δεύτερο μέρος παρουσιάζονται υλοποιήσεις εφαρμογών με εργαλεία της R, ενώ στο παράρτημα περιέχονται οι κύριες διαδικασίες χρήσης του γραφικού περιβάλλοντος εργασίας R-Studio της R, συμπεριλαμβάνοντας σύντομα παραδείγματα, αλλά και αναφορές στη χρήση δημοφιλών πακέτων και προσθέτων, όπως το **dplyr** (Wickham & Francois, 2015), το **factoextra** (Kassambara & Mundt 2016), **Factoshiny** (Vaissie et al. 2015), **ggplot2** (Whickham 2016), **Rcmdr** (Fox 2005) και το **RcmdrPlugin.Factominer** (Husson et al. 2012).

Με το παρόν επιχειρείται μια σύντομη προβολή των δυνατοτήτων εφαρμογής μεθόδων Ανάλυσης Δεδομένων σε γραφικό περιβάλλον προσωπικών υπολογιστών. Το σύνολο του κώδικα που περιγράφεται έχει μορφή σεναρίων κώδικα (R scripts), βρίσκεται στον συνοδευτικό ιστότοπο του βιβλίου (<http://bit.do/DAinR>) και κατασκευάστηκε χρησιμοποιώντας τη γλώσσα R, με σκοπό την ευχερή αναπαραγωγή του κατά τη μελέτη του παρόντος.

Βιβλιογραφία

Ελληνική

Αθανασιαδης Η., Π. (1995). Παραγοντική Ανάλυση Αντιστοιχιών και Ιεραρχική Ταξινόμηση, Παραδείγματα και Εφαρμογές, Εκδ. Νέων Τεχνολογιών, Αθήνα.

- Καραπιστόλης, Δ. (2002). Το Λογισμικό MAD. Τετράδια Ανάλυσης Δεδομένων, 2, 133-147.
- Καραπιστόλης, Δ. (1999). *Ανάλυση Δεδομένων και Έρευνα Αγοράς*, Εκδ. Ανικούλα, Θεσσαλονίκη.
- Κουτσουπιάς, Ν. & Παπαδημητρίου Ι. (2018). Μέθοδοι Πολυδιάστατης Ανάλυσης Δεδομένων με τη γλώσσα R, *Τετράδια Ανάλυσης Δεδομένων* (υπό δημοσίευση).
- Κουτσουπιάς, Ν. (2002). S-PRO: Εφαρμογή Υλοποίησης Μεθόδων της Ανάλυσης Δεδομένων. *Τετράδια Ανάλυσης Δεδομένων*, 1, 118-128.
- Μάρκος, Α. (2006). Βοήθεια στην ερμηνεία των αποτελεσμάτων της παραγοντικής ανάλυσης των αντιστοιχιών και αλγόριθμοι κατασκευής και ανάλυσης ειδικών πινάκων εισόδου. Διδακτορική Διατριβή. Θεσσαλονίκη: Πανεπιστήμιο Μακεδονίας.
- Μεϊμάρης, Μ. (2002). Εξισώσεις Μεταφοράς και Μεταφορά Εξισώσεων, *Τετράδια Ανάλυσης Δεδομένων*, 1, 12-19.
- Μενεξές, Γ. Χ. (2006). *Πειραματικοί σχεδιασμοί στην ανάλυση δεδομένων*. Διδακτορική Διατριβή. Θεσσαλονίκη: Πανεπιστήμιο Μακεδονίας.
- Παπαδημητρίου, Γ. (1994). *Μέθοδοι Ανάλυσης Δεδομένων* (Πανεπιστημιακές Παραδόσεις), Εκδ. Πανεπιστημίου Μακεδονίας, Θεσσαλονίκη.
- Παπαδημητρίου, Γ. (2007). *Η Ανάλυση Δεδομένων*, Τυπωθήτω.

Ξενογλωσση

- Benzécri, F. (1980a). "Introduction á la Analyse des Correspondances d'après un exemple de données médicales", *Les Cahiers de l'Analyse des Données*, 5(3), Dunod, Paris.
- Benzécri, F. (1980b). "Introduction á la Classification Automatique d'après un exemple de données médicales", *Les Cahiers de l'Analyse des Données*, 5(3), Dunod, Paris, p311-340.
- Benzécri, J.-P. & Maiti, G. D. (1992). *Pratique ed l'Analyse des Données en médecine*, Statmatic, Paris.
- Benzécri, J.-P. (1973). *Analyse des Données (1: La Taxonomie, 2: Correspondances)*, Dunod, Paris.
- Benzécri, J.-P. (1980). *Pratique ed l'Analyse des Données (1: Analyse des Correspondances, exposé élémentaire, 2: Abrégé théorique)*, Dunod, Paris.

- Benzécri, J.-P. (1981). *Pratique ed l' Analyse des Données* , 3: Linguistique et Lexicologie, Dunod, Paris
- Benzécri, J.-P. (1982a). Histoire et Prehistoire ed l' Analyse des Données, Dunod, Paris.
- Benzécri, J.-P., et coll. (1986). *Pratique ed l' Analyse des Données en économie*, Dunod, Paris.
- Bertier, P. & Bouroche, J.-M. (1975). *Analyse des données multidimensionnelles*, Presses Universitaires de France, Paris.
- Escofier, B. & Pages, J. (1988). *Analyses factorielles simples et multiples*, Dunod, Paris.
- Escofier, B. & Pagès, J. (1992). Análisis Factoriales Simples y Múltiples: Objetivos, Métodos e Interpretación, Bilbao, Universidad del Pais Vasco.
- Fox, J. (2005). Getting started with the R commander: a basic-statistics graphical user interface to R. *J Stat Softw*, 14(9), 1-42.
- Gifi, A. (1990). *Nonlinear Multivariate Analysis*, J. Wiley, Chichester.
- Good, I. J. (1990). "Exploratory data analysis", *Journal of Statistical Computation and Simulation*, 37, pp. 243-245.
- Hotelling, H. (1933). "Analysis of a complex of statistical variables into Principal Components", *Journal of Educational Psychology*, 24, pp. 417-441, pp.498-520.
- Husson, F., Josse, J., & Lê, S. (2012). RcmdrPlugin. FactoMineR: Graphical User Interface for FactoMineR (R package version 1.02)[Software].
- Karakos, A., (2012). PRAXITELE: The New Generation of Data Analysis Tools. *Software Engineering*, 2(5), 186-194.
- Kassambara, A. & Mundt, F. (2016). *Factoextra: extract and visualize the results of multivariate data analyses*. R package version, 1 (3).
- Kruskal, J. B. (1963). Nonmetric multidimensional scaling: A numerical method, *Psychometrika*, 29, 115-129.
- Lagarde, J. (1983). *Initiation a l' Analyse des Donees*, Dunod, Paris.
- Le, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1-18.
- Lubinsky, D. & Pregibon, D. (1988). "Data analysis as search", *Journal of Econometrics*, 38, pp.247-268.
- Markos, A., Menexes, G. & Papadimitriou, I. (2010). The CHIC Analysis Software v1.0. In H. Loracek-Junge & C. Weihs (eds.), *Classification as a Tool for Research*, Proceedings of the 11th IFCS Conference. Berlin: Springer, 409-416.

- Nishisato, S. (1980). Analysis of Categorical Data, *Dual Scaling and its Applications*, University of Toronto Press, Toronto.
- Porebski, O. R. (1966). Discriminatory and canonical analysis of technical college data. *British Journal of Mathematical and Statistical Psychology*, 19, 215-236.
- Shepard, R. N. (1961). The analysis of proximities: Multidimensional scaling with an unknown distance function. II, *Psychometrika*, 27, 219-246.
- Tatsuoka, M. M. (1970). *Discriminant analysis*. Champaign, IL: Institute for Personality and Ability Testing.
- Torgerson, S. W. (1952). Multidimensional Scaling: I. Theory and Method, *Psychometrika*, 17, 401-419.
- Tuckey, J. W. (1977). *Exploratory data analysis*, Addison-Wesley.
- Tukey, J. W. & Wilk, M. B. (1966). Data analysis and statistics: an expository overview. In Proceedings of the November 7-10, 1966, fall joint computer conference (pp. 695-709). ACM.
- Vaissie, P., Monge, A., & Husson, F. (2015). Factoshiny: perform factorial analysis from FactoMineR with a shiny application. R package version, 1.
- Wickham, H. (2016). ggplot2: elegant graphics for data analysis. Springer.
- Wickham, H., & Francois, R. (2015). dplyr: A grammar of data manipulation. *R package version 0.4*, 1, 20.