

ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Οι μέθοδοι πολυμεταβλητής ανάλυσης όπως φανερώσει και η ονομασία τους αναφέρονται σε διαδικασίες και μεθοδολογίες όπου προσπαθούμε να καταλήτουμε σε συμπεράσματα με την χρήση πολλών μεταβλητών.

Στην πράξη τα δεδομένα ενός ερευνητή είναι από την φύση τους πολυμεταβλητά. Σπάνια κάποιος ερευνητής μελετά μια μεταβλητή ανεξάρτητα και απομονωμένα από τις υπόλοιπες.

Οι λόχοι για τους οποίους οι πολυμεταβλητές τεχνικές είναι ιδιαίτερα χρήσιμοι είναι:

- Έχουμε περισσότερη πληροφορία
- Μελετάμε συσχετισμούς (μεταξύ μεταβλητών και υποκειμένων)

Οι πολυμεταβλητές τεχνικές χρησιμοποιούνται για:

- Την εύρεση και ερμηνεία συσχετίσεων μεταξύ των μεταβλητών
- Την δημιουργία ομάδων είτε από παρατηρήσεις είτε από μεταβλητές σύμφωνα με κάποια χαρακτηριστικά.

- Την μείωση των διαστάσεων του προβλήματος
δηλαδή μείωση των μεταβλητών οι οποίες
είναι επιδραστικές
- Την πρόβλεψη νέων τιμών
- Μοντελοποίηση σε πολλές διαστάσεις
- Ποσοτικοποίηση μη παρατηρήσιμων ποσοτήτων.

Υπάρχουν διάφορες πολυμεταβλητές τεχνικές
όπως:

- Ανάλυση σε κύριες συνιστώσες (Principal components Analysis)
- Παραγοντική Ανάλυση (Factor Analysis)
- Ανάλυση σε ομάδες (cluster Analysis)
- Διακριτική ανάλυση (Discriminant Analysis)
- Ανάλυση αντιστοιχιών (Correspondence Analysis)
- Ανάλυση κανονικών συσχετίσεων (canonical correlation Analysis)
- Πολυδιάστατη κλιμακοποίηση (Multidimensional Scaling)
- πολυμεταβλητό γραμμικό μοντέλο (Multivariate Linear Model)
- Μέθοδοι για δεδομένα διεύθωνs (Directional data)

Γραφήματα

Κάθε στατιστική ανάλυση ξεκινά κατασκευάζοντας διάφορα γραφήματα με τα δεδομένα ώστε να αποκτήσει κανείς μια εικόνα για αυτά. Προφανώς μια εικόνα είναι κατά πολύ προτιμότερη από το να προσπαθήσουμε να εφάχουμε συμπεράσματα κοιτάζοντας απλά αριθμούς. Τα γραφήματα αποτελούν ένα πολύτιμο εργαλείο για τον ερευνητή.

Όταν έχουμε περισσότερες από μια μεταβλητή χρειαζόμαστε γραφήματα που να μπορούν να αναπαραστήσουν περισσότερες από μια μεταβλητές κάθε φορά. Γενικά κάτι τέτοιο δεν είναι καθόλου εύκολο γιατί διαγράμματα τριών διαστάσεων μπορούν να κατασκευαστούν αλλά δύσκολα να ερμηνευτούν.

2.1.1 Matrix Plot

Το Matrix Plot δεν είναι παρά ένας οργανωμένος πίνακας από απλά διαγράμματα σημείων για ζευγάρια μεταβλητών. Το πλεονέκτημα του είναι πως μπορούμε να δούμε όλα τα

δυνατά ζευγάρια αλλά και επειδή οι κλίμακες είναι σταθερές μπορούμε να συγκρίνουμε ζεύγη μεταβλητών μεταξύ τους.

Ας δούμε ένα παράδειγμα:

Παράδειγμα 2.1. Τα δεδομένα αφορούν τα 25 σημαντικότερα πανεπιστήμια και τα 25 κολλέγια κλασσικών σπουδών (liberal arts colleges) της Αμερικής. Τόσο τα πανεπιστήμια όσο και τα κολλέγια αυτά θεωρούνται ισάξια από την άποψη της σημαντικότητας των πτυχίων και το μόνο που αλλάζει είναι η κατεύθυνση των σπουδών που προσφέρουν. Για τις 50 παρατηρήσεις υπάρχουν 8 μεταβλητές και συγκεκριμένα οι μεταβλητές:

Τα δεδομένα υπάρχουν στον πίνακα 2.1

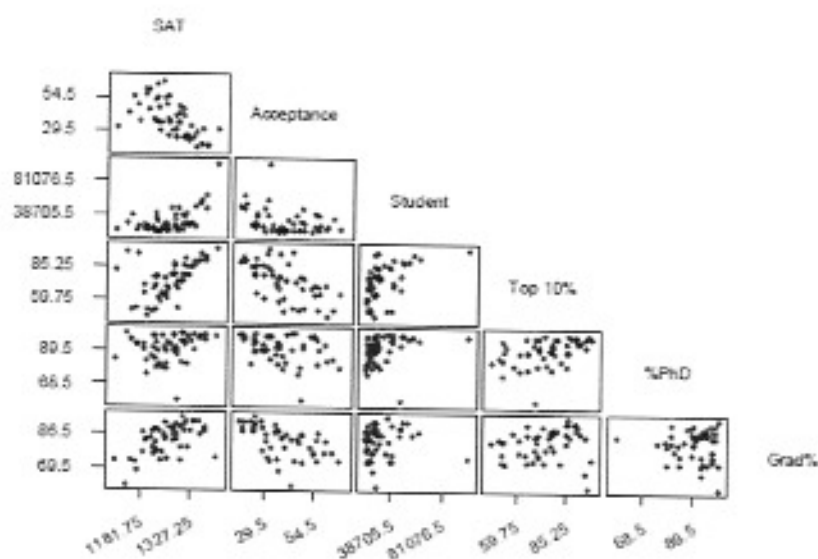
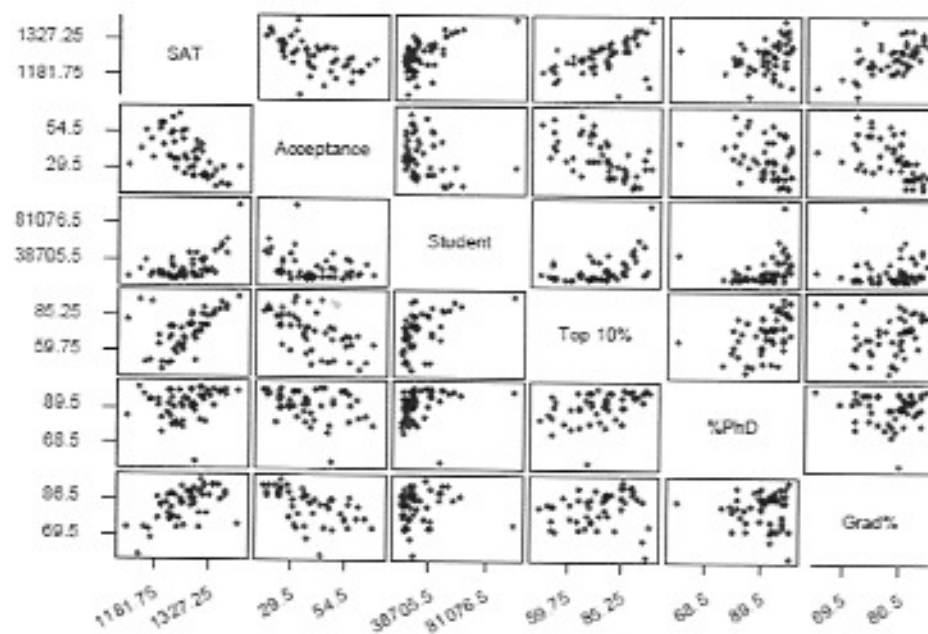
Όνομα σχολείου	School Type	SAT	Acceptance	Student	Top 10%	%PhD	Grad %
Ambert	Lib Arts	1315	22	26636	85	81	93
Swarthmore	Lib Arts	1310	24	27487	78	93	88
Williams	Lib Arts	1306	28	23772	86	90	93
Bowdoin	Lib Arts	1300	24	25703	78	95	90
Wellesley	Lib Arts	1290	49	27879	76	91	86
Pomona	Lib Arts	1320	33	26668	79	98	80
Wesleyan (CT)	Lib Arts	1290	35	19948	73	87	91
Middlebury	Lib Arts	1255	25	24718	65	89	92
Smith	Lib Arts	1195	57	25271	65	90	87
Davidson	Lib Arts	1230	36	17721	77	94	89
Vassar	Lib Arts	1287	43	20179	53	90	84
Carleton	Lib Arts	1300	40	19504	75	82	80
Claremont McKenna	Lib Arts	1260	36	20377	68	94	74
Oberlin	Lib Arts	1247	54	23591	64	98	77
Washington & Lee	Lib Arts	1234	29	17998	61	89	78
Grinnell	Lib Arts	1244	67	22301	65	79	73
Mount Holyoke	Lib Arts	1200	61	23358	47	83	83
Colby	Lib Arts	1200	46	18872	52	75	84
Hamilton	Lib Arts	1215	38	20722	51	86	85
Bates	Lib Arts	1240	36	17554	58	81	88
Haverford	Lib Arts	1285	35	19418	71	91	87
Colgate	Lib Arts	1258	38	17520	61	78	83
Bryn Mawr	Lib Arts	1255	56	18847	70	81	84
Occidental	Lib Arts	1170	49	20192	54	93	72
Barnard	Lib Arts	1220	53	17653	69	98	80
Harvard	Univ	1370	18	46918	90	99	90
Stanford	Univ	1370	18	61921	92	96	88
Yale	Univ	1350	19	52468	90	97	93
Princeton	Univ	1340	17	48123	89	99	93
Cal Tech	Univ	1400	31	102262	98	98	75
MIT	Univ	1357	30	56766	95	98	86
Duke	Univ	1310	25	39504	91	95	91

(Ο πίνακας συνεχίζεται στην επόμενη σελίδα)

Dartmouth	Univ	1306	25	35504	86	100	95
Cornell	Univ	1280	30	37137	85	90	83
Columbia	Univ	1268	29	45879	78	93	90
U of Chicago	Univ	1300	45	38937	74	100	73
Brown	Univ	1281	24	24201	80	98	90
U Penn	Univ	1280	41	30882	87	99	86
Berkeley	Univ	1176	37	23665	95	93	68
Johns Hopkins	Univ	1290	48	45460	69	58	86
Rice	Univ	1327	24	26730	85	95	88
UCLA	Univ	1142	43	26859	96	100	61
U Va	Univ	1218	37	19365	77	91	88
Georgetown	Univ	1278	24	23115	79	89	89
UNC	Univ	1109	32	19684	82	84	73
U Michigan	Univ	1195	60	21853	71	93	77
Carnegie Mellon	Univ	1225	64	33607	52	84	77
Northwestern	Univ	1230	47	28851	77	79	82
Washington U (MO)	Univ	1225	54	39883	71	98	76
U of Rochester	Univ	1155	56	38597	52	96	73

Πίνακας 2.1. Δεδομένα χρονοκρίσιμων πανεπιστημίων και κολεγίων

Όνομα μεταβλητής	Περιγραφή
School	Το όνομα του πανεπιστημίου
School_Type	Ο τύπος του πανεπιστημίου (Univ ή liberal arts college)
SAT	Η διάμεσος του βαθμού των φοιτητών στα διαγωνίσματα SAT
Acceptance	Το ποσοστό των φοιτητών που έγιναν δεκτοί στο πρόγραμμα του πανεπιστημίου
\$/Student	Το μέσο ποσό σε δολάρια που ξοδεύει κάθε φοιτητής
Top 10%	Το ποσοστό των τελειόφοιτων με βαθμό πάνω από κίτρινο όριο
%PhD	Ποσοστό διδασκόντων με διδακτορικό
Grad%	Το ποσοστό των φοιτητών που τελικά αποφοιτούν

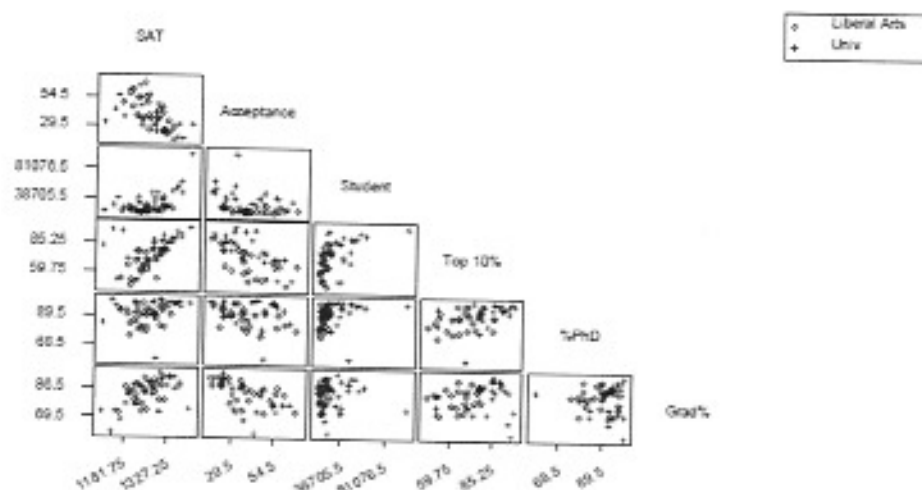


Γράφημα 2.2. Matrix plot, σε διαγώνια μορφή, για τα δεδομένα του Παραδείγματος 2.1

Από ένα matrix plot μπορεί κανείς να αποκτήσει γρήγορα μια εικόνα για το ποιες μεταβλητές συσχετίζονται με ποιες άλλες. Οι ονομασίες των μεταβλητών υπάρχουν στη διαγώνιο του πίνακα. Για να βρει κανείς ποιο είναι το ζεύγος των μεταβλητών για το οποίο έχει σχηματιστεί το διάγραμμα σημείων αρκεί να βρει ποια είναι η μεταβλητή που απεικονίζεται σε κάθε γραμμή και σε κάθε στήλη του πίνακα. Εναλλακτικά κάποια πακέτα έχουν στη διαγώνιο, αντί για τα ονόματα των μεταβλητών, κάποια μονοδιάστατα γραφήματα όπως boxplots, ή ιστογράμματα.

Ετσι από το γράφημα 2.1 μπορούμε να δούμε πως ανάμεσα στις μεταβλητές top 10% και SAT υπάρχει μια έντονη γραμμική σχέση ενώ αντίθετα η σχέση ανάμεσα στη μεταβλητή Acceptance και τη μεταβλητή Grad% φαίνεται να μην είναι ισχυρή.

Τα matrix plot είναι ιδιαίτερα χρήσιμα όταν θέλουμε να διαλέξουμε ζεύγη μεταβλητών με ισχυρή συσχέτιση, όπως πχ στην Γραμμική Παλινδρόμηση. Επίσης είναι πολύ χρήσιμα αν χρησιμοποιήσουμε διαφορετικά σύμβολα για διαφορετικές ομάδες. Για παράδειγμα, μπορεί κανείς να δει το γράφημα 2.3 όπου τα δύο διαφορετικά είδη πανεπιστημίων συμβολίζονται με διαφορετικά σύμβολα.



Γράφημα 2.3. Matrix plot για τα δεδομένα του Παραδείγματος 2.1, διακρίνοντας μεταξύ πανεπιστημίων και κολεγίων.

Μπορεί λοιπόν να δει κανείς από το γράφημα 2.3 πως τα πανεπιστήμια έχουν αρκετά διαφορετικά χαρακτηριστικά για κάποιες μεταβλητές. Για παράδειγμα από το διάγραμμα σημείων των μεταβλητών Top 10 % και Grad % βλέπουμε πως τα πανεπιστήμια ξεχωρίζουν από τα κολέγια κλασσικών σπουδών, γιατί τα '+' είναι τα περισσότερα σε διαφορετικό σημείο από ότι τα 'o'.

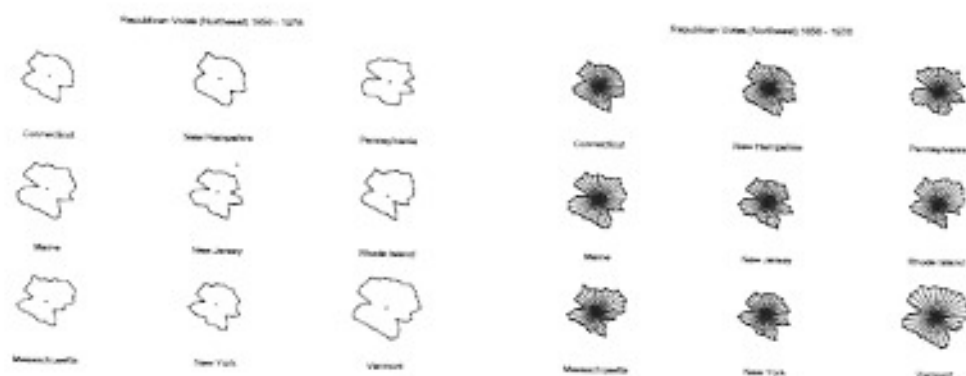
Θα πρέπει να τονιστεί πως ο ερευνητής θα μπορούσε να χρησιμοποιήσει και μια ακόμη κατηγορική μεταβλητή χρησιμοποιώντας το χρώμα για να ξεχωρίζει ανάμεσα από τις κατηγορίες αυτής της μεταβλητής. Αυτό αν και είναι ιδιαίτερα χρήσιμο σε έγχρωμες παρουσιάσεις θα μπορούσε να δημιουργήσει προβλήματα σε ασπρόμαυρες εκτυπώσεις των αποτελεσμάτων. Για αυτό η χρήση χρώματος για να υποδηλώνονται κατηγορικές μεταβλητές, αν και χρήσιμη οπτικά, μπορεί να οδηγήσει σε παρερμηνείες όταν τα χρώματα δεν διακρίνονται καλά (πχ σε μια εκτύπωση).

2.1.2 Starplots

Τα διαγράμματα-αστέρια (starplots) είναι ένας έξυπνος τρόπος να αναπαραστήσει κανείς γραφικά πολυδιάστατα γραφήματα. Για κάθε μια παρατήρηση ο ερευνητής κατασκευάζει ένα 'αστέρι' με τόσες ακτίνες όσες είναι και οι μεταβλητές, δηλαδή το μέγεθος κάθε ακτίνας αναπαριστά την τιμή της παρατήρησης για κάποια μεταβλητή. Συνεπώς κινώντας από το οπτικό αποτέλεσμα μπορεί κανείς να παρατηρήσει διαφορές ανάμεσα στις παρατηρήσεις καθώς αυτές θα έχουν διαφορετικά σχήματα. Τέτοια γραφήματα μπορούν να χρησιμοποιηθούν για μεγάλο αριθμό μεταβλητών. Ο περιορισμός είναι συνήθως πως επειδή κάθε παρατήρηση αντιστοιχεί σε ένα αστέρι, αν έχουμε ένα μεγάλο σετ δεδομένων με πολλές παρατηρήσεις δεν είναι εύκολο να χωρέσουμε σε μια σελίδα τόσα πολλά γραφήματα και άρα να κινούμε οπτικές συγκρίσεις για όλες τις παρατηρήσεις.

Στις περιπτώσεις που τα δεδομένα μας ανήκουν σε διαφορετικές κατηγορίες ένα starplot για τις μέσες τιμές μπορεί να μας δώσει μια εικόνα για το πως διαφέρουν οι μέσες τιμές.

Για παράδειγμα δείτε τα starplots για ένα σετ δεδομένων, που αφορά τα ποσοστά του ρεπουμπλικανικού κόμματος στις Αμερικάνικες εκλογές για διάφορες πληθυσμιακές ομάδες. Παρουσιάζονται δύο διαφορετικές μορφές του γραφήματος, στη μια οι ακτίνες που συνδέουν το κέντρο με την περιφέρεια έχουν τυπωθεί ενώ στην άλλη περίπτωση μόνο το περίγραμμα εμφανίζεται. Από τη μορφή που παίρνει το αστέρι για κάθε παρατήρηση μπορεί κανείς να δει τις διαφορές που υπάρχουν ανάμεσα στις περιοχές. Θα πρέπει να τονιστεί πως το μήκος της ακτίνας υπολογίζεται με βάση τις τυποποιημένες τιμές.



Starplot με μόνο την περίμετρο

Starplot με ακτίνες

Γράφημα 2.4 Παράδειγματα γραφημάτων starplot

Παράδειγμα 2.1 (συνέχεια):

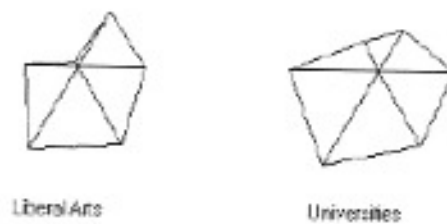
Ας επανέλθουμε στο παράδειγμα με τα αμερικάνικα πανεπιστήμια. Στα γραφήματα που ακολουθούν (2.4α και β αντίστοιχα) μπορείτε να δείτε τα starplots για τις δύο διαφορετικές κατηγορίες πανεπιστημίων. Είναι ενδιαφέροντες οι διαφορές και ανάμεσα σε πανεπιστήμια της ίδιας κατηγορίας. Γενικά τα πανεπιστήμια έχουν αρκετά μεγαλύτερα αστέρια από ότι τα κολέγια κλασικών σπουδών. Αυτό επαληθεύεται κοιτάζοντας τα αστέρια για τους μέσους κάθε ομάδας, στο γράφημα 2.4γ.



Γράφημα 2.4α. Starplot για τα κολέγια κλασικών σπουδών



Γράφημα 2.4β. Scatterplot για τα πανεπιστήμια



Γράφημα 2.4γ. Scatterplot για τους μέσους των δύο ομάδων

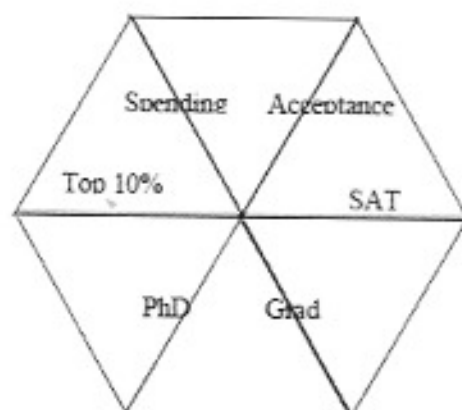
Κοιτάζοντας τις μέσες τιμές των μεταβλητών για τους δύο τύπους σχολείων (Πίνακας 2.2) βλέπουμε πως υπάρχει σημαντική διαφορά στη μεταβλητή που έχει να κάνει με τα έξοδα των φοιτητών. Επει παρατηρούμε και τη μεγαλύτερη διαφορά των αυτεχών στο γράφημα 2.4γ

SAT	Acceptance	Student	Top 10%	%PhD	Grad%	Σχολείο
1256.64	40.56	21755.60	67.24	88.24	84.12	Liberal Arts
1271.28	35.12	38738.90	81.64	92.88	82.84	Universities

Πίνακας 2.2. Οι μέσες τιμές για τις μεταβλητές μας για τους δύο τύπους πανεπιστημίων

Θα πρέπει εδώ να σημειώσουμε πως ο τρόπος που καθορίζεται η σειρά των μεταβλητών και άρα η ακτίνα που αντιστοιχεί σε κάθε μια από αυτές, είναι συνήθως ο εξής: η πρώτη σε σειρά μεταβλητή είναι παράλληλη με τον άξονα και προς τα δεξιά δηλαδή στη θέση 3 του ρολογιού. Οι υπόλοιπες τοποθετούνται διαδοχικά σε αντίστροφη σειρά από αυτή του ρολογιού.

Στο παράδειγμα μας η σειρά των μεταβλητών απεικονίζεται στο γράφημα 2.5. Δηλαδή ξεκινάμε από την πρώτη μεταβλητή SAT που αντιστοιχεί στη δεξιά μεριά του άξονα και στη συνέχεια συνεχίζουμε με τις μεταβλητές Acceptance κλπ.



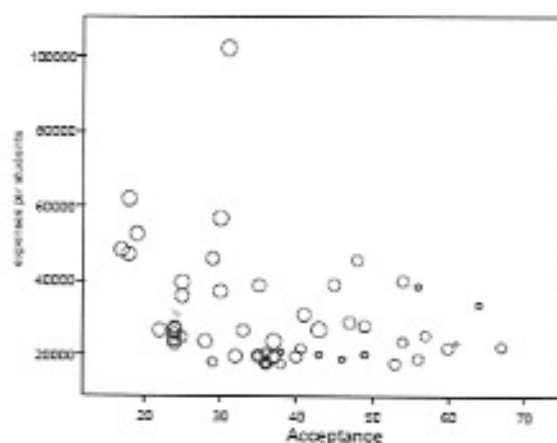
Γράφημα 2.5 Σειρά μεταβλητών σε starplot

2.1.3 Bubble Plots

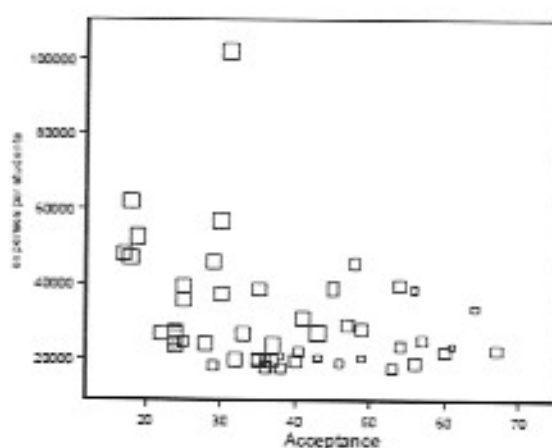
Τα bubble plots δεν είναι παρά απλά διαγράμματα σημείων για τα οποία μια τρίτη μεταβλητή απεικονίζεται ανάλογα με το μέγεθος των κύκλων που αναπαριστούν κάθε παρατήρηση. Δηλαδή ενώ στα απλά διαγράμματα σημείων κάθε παρατήρηση συμβολίζεται με ένα σύμβολο ίδιου μεγέθους για κάθε παρατήρηση, στα bubble plots, χρησιμοποιούμε διαφορετικό μέγεθος ανάλογα με μια τρίτη μεταβλητή. Συνήθως το σύμβολο είναι κύκλος, και για αυτό το γράφημα μοιάζει με σπουννόφουσκας, από όπου πηγάζει και το όνομα. Πολλά πακέτα επιτρέπουν τη χρήση άλλων σχημάτων όπως τρίγωνα ή τετράγωνα. Το κέντρο πάντα του συμβόλου τοποθετείται στο σημείο που ορίζουν οι δύο μεταβλητές και το μέγεθος του συμβόλου καθορίζεται από μια τρίτη μεταβλητή.

Για παράδειγμα στα γραφήματα που ακολουθούν μπορεί κανείς να δει Bubble plots με τη χρήση κύκλων και τετραγώνων για τις τριάδες μεταβλητών (Acceptance, Έξοδα φοίτησης και Top 10%).

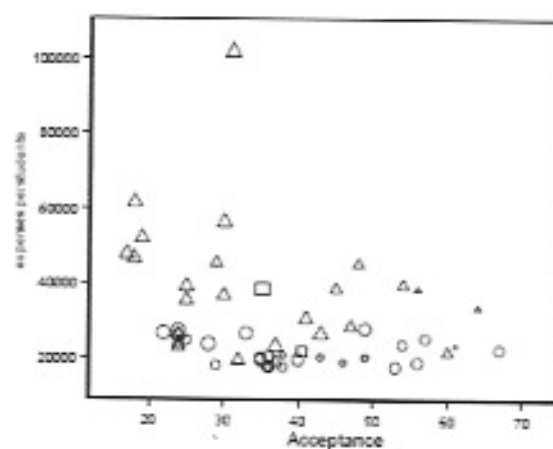
Μπορεί κανείς να εισάγει και κατηγορικές μεταβλητές στο γράφημα χρησιμοποιώντας άλλα σύμβολα ή διαφορετικά χρώματα. Για παράδειγμα στο γράφημα 2.6γ τα τριγωνάκια αντιστοιχούν σε παρατηρήσεις από πανεπιστήμια και τα κυκλάκια από κολέγια κλασσικών σπουδών.



Γράφημα 2.6α. Bubble plot για τις μεταβλητές 'ποσοστό αποδοχής' και 'έξοδα ανά φοιτητή'. Το μέγεθος κάθε κυκλίου καθορίζεται από την τιμή της μεταβλητής top 10%.



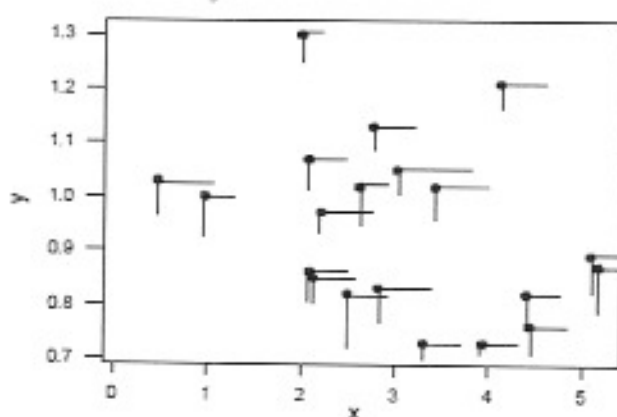
Γράφημα 2.6β. Bubble plot για τις μεταβλητές 'ποσοστό αποδοχής' και 'έξοδα ανά φοιτητή'. Το μέγεθος κάθε τετραγώνου καθορίζεται από την τιμή της μεταβλητής top 10%.



Γράφημα 2.6γ Το ίδιο Bubble plot για τις μεταβλητές 'ποσοστό αποδοχής' και 'έξοδα ανά φοιτητή' αλλά με διαφορετικά σύμβολα για τα κολλέγια και τα πανεπιστήμια. Το μέγεθος κάθε τετραγώνου/τριγώνου καθορίζεται από την τιμή της μεταβλητής top 10%.

2.1.4 Glyph Plot

Το Glyph plot είναι ένας έξυπνος τρόπος να αναπαραστήσουμε τέσσερις μεταβλητές σε ένα μόνο γράφημα. Οι δύο πρώτες τοποθετούνται στο γράφημα όπως και σε ένα απλό διάγραμμα σημείων, ενώ οι άλλες δύο τοποθετούνται ως γραμμές από το σημείο που ορίζουν οι δύο πρώτες. Το μήκος αυτών των γραμμών καθορίζεται από την τιμή των παρατηρήσεων για αυτές τις δύο μεταβλητές.



Γράφημα 2.7. Παράδειγμα glyph plot. Οι δύο μεταβλητές x και y αναπαρίστανται από τη θέση του σημείου όπως σε κάθε διάγραμμα σημείων, ενώ οι άλλες δύο μεταβλητές με το μήκος των δύο κάθετων γραμμών

2.1.5 Τα Πρόσωπα του Chernoff

Η μέθοδος αυτή είναι ένας εντυπωσιακός αλλά συνάμα αρκετά αποτελεσματικός τρόπος να απεικονίσει κινείς πολυμεταβλητά δεδομένα και είναι ευρέως διαδεδομένη στις κοινωνικές επιστήμες και την ψυχολογία.

Η μέθοδος πρωτοεμφανίστηκε από τον Αμερικάνο Στατιστικό Chernoff το 1973. Στην πραγματικότητα για κάθε παρατήρηση δημιουργούμε ένα πρόσωπο του οποίου τα χαρακτηριστικά (πχ μέγεθος, θέση και μέγεθος μύτης και αυτιών κλπ) καθορίζονται από την τιμή της παρατήρησης για κάποια μεταβλητή. Έτσι ανάλογα με τις τιμές των παρατηρήσεων προκύπτουν διαφορετικά πρόσωπα. Φυσικά ανάλογα με το πόσες μεταβλητές χρησιμοποιούμε χρειάζομαστε ανάλογο αριθμό χαρακτηριστικών του προσώπου για να το κατασκευάσουμε.

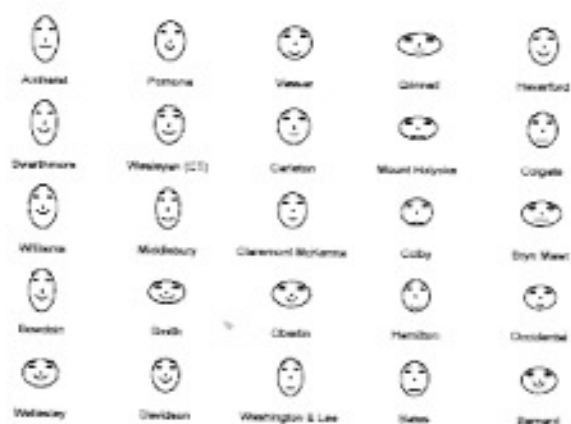
Στα γραφήματα (2.8 και 2.9) που ακολουθούν βλέπουμε τα πρόσωπα του Chernoff για τα σχολεία κλαστικών σπουδών, τα πανεπιστήμια και για τους μέσους των δύο ομάδων αντίστοιχα. Παρατηρείστε πως είναι εμφανείς οι διαφορές ως προς τα χαρακτηριστικά των προσώπων και πως υπάρχει μια έντονη ποικιλία προσώπων. Στο τρίτο γράφημα όπου πια συγκρίνουμε διαφορετικές ομάδες μεταξύ τους είναι εμφανής η διαφορά ανάμεσα στα δύο είδη σχολείων.

Βέβαια για να μπορέσει κανείς να μελετήσει τα πρόσωπα αυτά πρέπει να γνωρίζει τι αντιστοιχεί σε κάθε μεταβλητή. Έτσι στην περίπτωση μας έχουμε πως

1. *Περιοχή προσώπου* = βαθμός SAT
2. *Σχήμα προσώπου* = Ποσοστό φοιτητών που έγιναν δεικτοί
3. *Μήκος μύτης* = Ποσό που ξοδεύει κάθε φοιτητής
4. *Τοποθεσία στόματος* = Ποσοστό τελειόφοιτων με μεγάλο βαθμό
5. *Καμπύλη χαμόγελου* = Ποσοστό διδασκόντων με διδακτορικό
6. *Μήκος στόματος* = Ποσοστό φοιτητών που αποφοιτούν

Παρατηρείστε πως όπως είχαμε δει και στα starplots η διαφορά στα έξοδα των φοιτητών είναι μεγάλη, το μέγεθος της μύτης είναι πολύ διαφορετικό ανάμεσα στα δύο είδη σχολείων. Θα πρέπει πάντως να παρατηρήσει κανείς πως επειδή οι διαφορές που γίνονται πιο εύκολα και άμεσα κατανοητές είναι αυτές που έχουν να κάνουν με το μέγεθος του προσώπου, συνήθως οι δύο πρώτες μεταβλητές που αντιστοιχούν σε αυτά τα χαρακτηριστικά θα πρέπει να φροντίζουμε να είναι και αυτές που έχουν μεγαλύτερη σημασία, ώστε να είναι εύκολο να αποκαλύπτουν διαφορές.

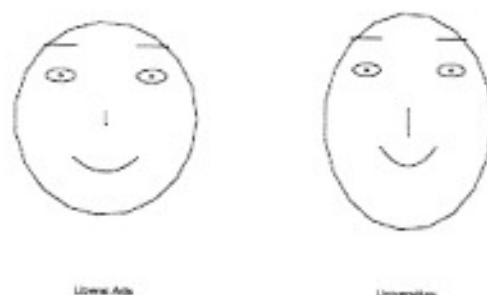
Τα διαγράμματα του Chernoff είναι πολύ χρήσιμα και για την απεικόνιση πολυμεταβλητών χρονοσειρών, στα οποία μπορούμε να δούμε πώς αλλάζει το πρόσωπο με το χρόνο.



Γράφημα 2.8α. Τα πρόσωπα του Chernoff για τα καλύτερα κλασικών σπορτών



Γράφημα 2.8β. Τα πρόσωπα του Chernoff για τα πανεπιστήμια



Γράφημα 2.9. Τα πρόσωπα του Chernoff για τους δύο τύπους σχολείων

Αξίζει να αναφερθεί πως τα πρόσωπα του Chernoff στηρίζονται στη λογική πως ο ανθρώπινος εγκέφαλος είναι καλά εκπαιδευμένος να αναγνωρίζει διαφορές και ομοιότητες σε πρόσωπα, μέρος ίσως μιας υποσυνείδητης ικανότητας του. Αυτό προέτρεψε τον Chernoff να προτείνει αυτόν τον τύπο γραφημάτων. Θα πρέπει όμως να τονιστεί πως για να μπορούμε να ξεχωρίσουμε δύο παρατηρήσεις θα πρέπει να υπάρχουν έντονες διαφορές σε ευδιάκριτα χαρακτηριστικά όπως πχ το σχήμα του προσώπου. Δηλαδή έχει σημασία ο τρόπος με τον οποίο αναπαριστούμε κάθε μεταβλητή, ποία μεταβλητή αντιστοιχεί σε ποιο χαρακτηριστικό του προσώπου.

2.1.6 Οι Καμπύλες του Andrews

Οι καμπύλες του Andrews (Andrews' curves) εισήχθηκαν από τον Andrews με σκοπό να απεικονιστούν πολυμεταβλητά δεδομένα.

Για κάθε παρατήρηση σχηματίζουμε την καμπύλη της συνάρτησης

$$f_i(t) = X_{i1}/\sqrt{2} + X_{i2} \sin t + X_{i3} \cos t + X_{i4} \sin(2t) + X_{i5} \cos(2t) + \dots, \quad t \in (-\pi, \pi)$$

για διαφορετικές τιμές του t και στη συνέχεια φτιάχνουμε το γράφημα $(t, f(t))$ για το διάστημα $(-\pi, \pi)$.

Όπως μπορείτε να δείτε η καμπύλη αυτή αποτελείται από ημίτονα και συνημίτονα έχει δηλαδή μια περιοδικότητα. Αυτή η περιοδικότητα όμως εξαρτάται από τις τιμές των μεταβλητών και επομένως για διαφορετικές παρατηρήσεις περιμένουμε και διαφορετικές καμπύλες. Επομένως με τη χρήση αυτών των καμπυλών ελπίζουμε πως απεικονίζοντας πολυμεταβλητές παρατηρήσεις θα μπορούσαμε να δούμε πόσο και ποιες από αυτές διαφέρουν, καθώς παρατηρήσεις με διαφορετικές τιμές σε κάθε μεταβλητή θα έχουν διαφορετικά χαρακτηριστικά (πχ διαφορετική περιοδικότητα, μεγαλύτερη συχνότητα, κλπ).

Παρά την απλή σχετική ιδέα υπάρχουν κάποια προβλήματα που έχει να αντιμετωπίσει ο ερευνητής. Μερικά ενδιαφέροντα σημεία για τις καμπύλες αυτές είναι τα εξής:

- Η επιλογή της σειράς με την οποία οι μεταβλητές θα χρησιμοποιηθούν. Δεδομένου πως η σειρά καθορίζει τη σημαντικότητα κάθε μεταβλητής στη δημιουργία της καμπύλης, οι μεταβλητές τοποθετούνται με φθίνουσα σειρά διακύμανσης, δηλαδή η X_1 είναι η μεταβλητή με τη μεγαλύτερη διακύμανση, η X_2 η μεταβλητή με τη δεύτερη μεγαλύτερη διακύμανση και ούτω καθεξής. Εναλλακτικά, θα μπορούσε να χρησιμοποιήσει κανείς τις κύριες συνιστώσες που θα δούμε όμως αργότερα.
- Παρατηρήσεις που είναι σχετικά ίδιες μεταξύ τους θα έχουν σχετικά ίδιες καμπύλες και επομένως μπορούμε να διακρίνουμε ομάδες παρατηρήσεων. Αντίθετα παρατηρήσεις που διαφέρουν θα έχουν πολύ διαφορετικές καμπύλες.
- Επομένως η μέθοδος είναι ικανή να βρει ακραίες παρατηρήσεις (outliers) από ένα σύνολο παρατηρήσεων. Αν δηλαδή υπάρχει κάποια(ες) παρατήρηση(εις) για την οποία η καμπύλη είναι ολόκληρα διαφορετική αυτό σημαίνει πως η παρατήρηση αυτή είναι πολύ διαφορετική από τις άλλες και άρα είναι πιθανό outlier.
- Δυστυχώς αν έχουμε ένα μεγάλο αριθμό παρατηρήσεων, και δεδομένου πως για κάθε παρατήρηση έχουμε μια καμπύλη το γράφημα που θα πάρουμε δεν θα είναι καθόλου χρήσιμο, εκτός και αν κάποιες παρατηρήσεις είναι πολύ μακριά από τις άλλες και ξεχωρίζουν εμφανώς.
- Είναι ευνόητο πως εναλλακτικά κανείς μπορεί να σχηματίσει τις καμπύλες για τις μέσες τιμές υποομάδων και να συγκρίνει πια υποομάδες. Ουσιαστικά οι καμπύλες αυτές είναι ένα ισχυρό εργαλείο για την περιγραφή πολυμεταβλητών δεδομένων.
- Είναι πολύ σημαντικό πως οι καμπύλες Andrews έχουν την εξής ιδιότητα: η απόσταση ανάμεσα σε δύο καμπύλες είναι ανάλογη της ευθείας απόστασης ανάμεσα στις παρατηρήσεις. Επομένως οι καμπύλες Andrews αναπαριστούν τις διαφορές που υπάρχουν ανάμεσα στις παρατηρήσεις.

Εν κατακλείδι, οι καμπύλες αυτές μας προσφέρουν έναν εύκολο τρόπο να πάρουμε μια πρώτη εικόνα από τα δεδομένα και κυρίως να δούμε κατά πόσο οι παρατηρήσεις μας μοιάζουν μεταξύ τους και αν υπάρχουν outliers στα δεδομένα μας.

Παράδειγμα 2.2.

Στο παράδειγμα που ακολουθεί για 20 αμερικανικές πόλεις έχουν καταγραφεί μια σειρά από κοινωνικοοικονομικά χαρακτηριστικά και συγκεκριμένα, ο αριθμός καταστημάτων (σε χιλιάδες), ο αριθμός πρατηρίων χονδρικής (σε εκατοντάδες), το μέσο οικογενειακό εισόδημα (σε χιλιάδες δολάρια), ο δείκτης εξόδων και ο πληθυσμός (σε εκατομμύρια). Τα δεδομένα υπάρχουν στον πίνακα 2.3.

Αριθμός καταστημάτων (σε χιλιάδες)	Αριθμός πρατηρίων χονδρικής (σε χιλιάδες)	Μέσο οικογενειακό εισόδημα (σε χιλιάδες δολάρια)	Δείκτης εξόδων	Πληθυσμός
0.73	3.94	6.30	1.24	0.91
1.05	3.03	4.69	0.76	1.06
0.82	4.41	3.67	1.01	0.90
1.13	2.77	9.36	1.37	1.02
0.89	5.10	9.38	1.14	0.95
0.86	2.09	5.88	1.38	1.04
0.97	2.22	5.77	0.76	0.93
0.85	2.13	8.29	0.97	0.91
0.83	2.83	4.43	1.70	1.02
1.00	0.97	8.56	1.11	0.95
1.21	4.12	7.23	0.99	1.09
1.02	2.63	4.69	1.42	1.05
0.76	4.45	8.49	1.06	0.90
1.30	2.00	11.71	1.33	1.17
1.03	0.48	6.07	1.20	0.99
1.07	2.08	9.12	2.08	1.11
0.82	2.50	6.66	0.86	0.72
0.73	3.31	8.62	1.36	0.72
0.87	5.18	10.38	0.57	0.75
1.02	3.43	6.44	1.89	1.04

Πίνακας 2.3. Δεδομένα παραδείγματος 2.2

Για τα δεδομένα του πίνακα 2.3. βρήκαμε πως οι δοκιμασίες των μεταβλητών είναι

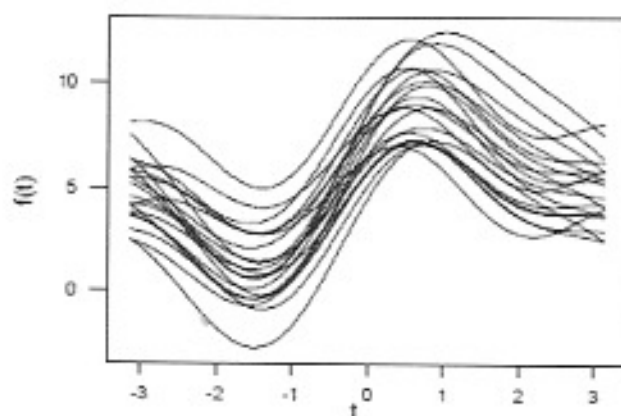
Μεταβλητή	Αριθμός καταστημάτων (σε χιλιάδες)	Αριθμός πρατηρίων χονδρικής (σε χιλιάδες)	Μέσο οικογενειακό εισόδημα (σε χιλιάδες δολάρια)	Δείκτης εξόδων	Πληθυσμός
	X_0	X_1	X_2	X_3	X_4
Διακύμανση	0.0250063	1.61200	4.71420	0.142968	0.0155292

Πίνακας 2.4. Διακυμάνσεις δεδομένων πίνακα 2.3

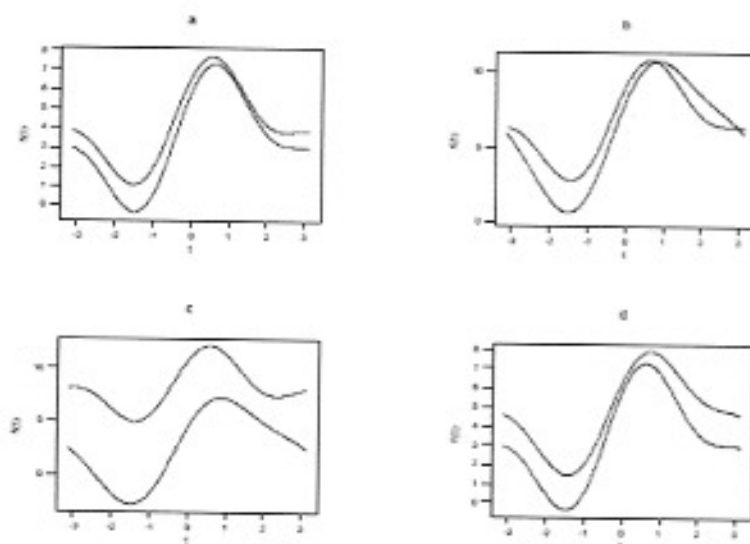
Επομένως η σειρά τους με φθίνουσα διακύμανση είναι X_0 , X_1 , X_2 , X_3 και X_4 και επομένως η καμπύλη για οποιαδήποτε παρατήρηση υπολογίζεται ως

$$f_X(t) = X_0 / \sqrt{2} + X_1 \sin t + X_2 \cos t + X_3 \sin(2t) + X_4 \cos(2t), \quad t \in (-\pi, \pi)$$

Στο γράφημα 2.10 μπορεί κανείς να δει τις 20 καμπύλες για όλες τις παρατηρήσεις. Παρατηρείστε το ενδιαφέρον σημείο πως δεν φαίνεται να υπάρχουν απενός πόλεις που μοιάζουν πολύ μεταξύ τους και αφετέρου δεν φαίνεται να υπάρχει κάποια πόλη με μεγάλη διαφορά από τις υπόλοιπες. Παρατηρείστε πως η πόλη 14 που αποκλίνει περισσότερο από τις άλλες στην περιοχή $(-2, -1)$ σε άλλα σημεία της καμπύλης ταυτίζεται με τις υπόλοιπες.



Γράφημα 2.10. Οι καμπύλες του Andrews για τα δεδομένα των αμερικάνικων πόλεων



Γράφημα 2.11. Σύγκριση των καμπυλών του Andrews για επιλεγμένα ζεύγη παρατηρήσεων

Στο γράφημα μπορεί κανείς να δει 4 ζεύγη παρατηρήσεων και συγκεκριμένα τα ζεύγη (7,13, γράφημα a), (5,14, γράφημα b), (4,15, γράφημα c), (13,18, γράφημα d). Από το γράφημα c μπορούμε να δούμε τις 2 παρατηρήσεις που διαφέρουν περισσότερο. Παρατηρείστε ότι σχεδόν πουθενά οι καμπύλες δεν πλησιάζουν ή μια την άλλη αρκετά κοντά. Από την άλλη στο γράφημα a βλέπουμε τις 2 παρατηρήσεις που είναι πιο κοντά ή μια στην άλλη, οι καμπύλες τους κινούνται πολύ κοντά. Το γράφημα b έχει το ενδιαφέρον σημείο πως οι καμπύλες τέμνονται σε κάποιο σημείο κοντά στο $t=f$, ενώ στο γράφημα d παρατηρούμε πως οι καμπύλες έρχονται κοντά και μετά απομακρύνονται. Η ποικιλία των καμπυλών που μπορούμε να πάρουμε είναι τεράστια και το γράφημα μπορεί να μας δώσει μια καλή εικόνα του πόσο μοιάζουν οι παρατηρήσεις μεταξύ τους.

ΠΟΛΥΜΕΤΑΒΛΗΤΑ ΔΕΔΟΜΕΝΑ

Όταν έχουμε πολυμεταβλητά δεδομένα οργανώνονται υπο την μορφή πινάκων

Για τα περισσότερα στατιστικά πακέτα τα δεδομένα έχουν την εξής μορφή

	χαρακτηριστικό ή μεταβλητή 1	χαρακτηριστικό μεταβλητή 2	...	χαρακτηριστικό μεταβλητή p
Παρατήρηση 1	x_{11}	x_{12}	...	x_{1p}
Παρατήρηση 2	x_{21}	x_{22}	...	x_{2p}
...	...	...	...	...
Παρατήρηση n	x_{n1}	x_{n2}	...	x_{np}

n: Αριθμός παρατηρήσεων / αντικειμένων προς μελέτη

p: -||- χαρακτηριστικών / μεταβλητών προς μελέτη

x_{ij} : Τιμή του i αντικειμένου στο j χαρακτηριστικό μεταβλητή

Ο πίνακας $\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$ λέγεται πίνακας δεδομένων και είναι διαστάσεων $n \times p$

Παράδειγμα: Εστω πως έχουμε παρατηρήσεις για 4 φοιτητές σχετικά με το ύψος, το βάρος και την ηλικία τους

		Υψος σε cm	Βάρος σε κιλά	Ηλικία σε χρόνια	(σε > 21)
φοιτητής 1		167	67	21	
- II -	2	178	75	22	
- II -	3	162	52	20	
- II	4	190	85	21	

Χρησιμοποιώντας την προηγούμενη ορολογία έχουμε $n=4$ (αριθμός παρατηρήσεων) και $p=3$ μεταβλητές

Ο πίνακας είναι $X = \begin{bmatrix} 167 & 67 & 21 \\ 178 & 75 & 22 \\ 162 & 52 & 20 \\ 190 & 85 & 21 \end{bmatrix}$

σε κάθε παρατήρηση αντιστοιχεί ένα διάνυσμα. Δηλαδή αν X_1 είναι ο φοιτητής 1 αντιστοιχεί σε αυτόν το διάνυσμα στήλη

$X_1 = \begin{bmatrix} 167 \\ 67 \\ 21 \end{bmatrix}$ που είναι η πρώτη γραμμή

του πίνακα X . Επίσης $X_2 = \begin{bmatrix} 178 \\ 75 \\ 22 \end{bmatrix}$ $X_3 = \begin{bmatrix} 162 \\ 52 \\ 20 \end{bmatrix}$
 $X_4 = \begin{bmatrix} 190 \\ 85 \\ 21 \end{bmatrix}$

Αρα ο πίνακας $X = \begin{bmatrix} X_1^T \\ X_2^T \\ X_3^T \\ X_4^T \end{bmatrix}$

(Υπενθύμιση ΑΤ ο ανάστροφος του πίνακα A)

Δηλαδή εδώ πχ $X_1^T = [167 \ 67 \ 21]$ η πρώτη γραμμή του πίνακα X κ.ο.κ.

Επομένως ο πίνακας δεδομένων όπως αυτός εμφανίζεται σε όλα τα στατιστικά πακέτα περιέχει τα διανύσματα γραμμή ως παρατηρήσεις. Όταν όμως μιλάμε για μεμονωμένες παρατηρήσεις θα εννοούμε τα διανύσματα στήλη που είναι οι ανάστροφοι των γραμμών του πίνακα δεδομένων.

Μέτρα Θέσης

Μέση τιμή Διάνυσμα μέσης τιμής είναι το

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \quad \text{όπου} \quad X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \quad \text{γενικά} \quad X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \quad j = 1, 2, \dots, p$$

$$\text{και} \quad \bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n} \quad \text{είναι ο μέσος της}$$

μεταβλητής $j \quad j = 1, 2, \dots, p$

Στο παράδειγμα με τους φοιτητές το διάνυσμα μέσης τιμής είναι

$$\bar{X} = \frac{1}{4} \begin{bmatrix} 167 + 178 + 162 + 190 \\ 67 + 75 + 52 + 85 \\ 21 + 22 + 20 + 21 \end{bmatrix} = \begin{bmatrix} 174,25 \\ 69,75 \\ 21 \end{bmatrix}$$

(6693)

Για το διάνυσμα των μέσων ισχύουν όλα χωρίζουμε για την μέση τιμή. Παράδειγμα ότι επηρεάζεται από τις ακραίες τιμές. Στις μονομετάβλητες περιπτώσεις έχουμε και άλλα μέτρα θέσης όπως π.χ η διάμεσος τα οποία όμως δεν μπορούν να οριστούν στα πολυμετάβλητα δεδομένα. Συγκεκριμένα η διάμεσος στα μονομετάβλητα δεδομένα είναι ένας αριθμός που διατάσσεται στη μέση των δεδομένων Δηλαδή 50% των δεδομένων έχουν τιμή μικρότερη από την διάμεσο και 50% μεγαλύτερη. Στα πολυμετάβλητα δεδομένα η διάμεσος θα ήταν διάνυσμα στην $\mathbb{R}^n$. Όμως τα διανύσματα (vectors $\mathbb{R}^n$) δεν έχουν διατάξη ώστε να μπορούμε να βρούμε κάποιο στη μέση.

Μέτρα μεταβλητότητας

(66) 24)

Είναι γνωστό πως η διακύμανση αποτελεί το βασικό μέτρο μεταβλητότητας στην μονοδιάστατη περίπτωση.

Αν έχουμε δύο μεταβλητές 'έστω X_j και X_k τότε ορίζουμε την συνδιακύμανση δύο μεταβλητών ως εξής

$$s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \quad j, k = 1, 2, \dots, p$$

Προσοχή Εδώ $X_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}$ j στήλη του πίνακα X

και $X_k = \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{bmatrix}$ k στήλη του πίνακα X

Στα ποσμετάβλητα δεδομένα ορίζουμε τον πίνακα διακυμάνσεων συνδιακυμάνσεων

$$S = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{bmatrix} \quad \text{ο οποίος}$$

είναι συμμετρικός (δηλαδή $s_{k1} = s_{1k}$

για $k, 1 = 1, 2, \dots, p$) και θετικά ημιορισμένος (δηλαδή όλες οι ιδιοτιμές του είναι μη αρνητικές)

Ένας συμμετρικός πίνακας A είναι θετικά ^{(σε 25) ημι-}
 -ορισμένος αν κάθε υποπίνακας του A_m της
 μορφής $A_m = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix} \quad m=1,2,\dots,n$ έχει
 ορίζουσα $\det A_m \geq 0 \quad \forall m=1,2,\dots,n$
 Επίσης τα στοιχεία της κύριας διαγωνίου
 του πίνακα S είναι θετικοί αριθμοί
 γιατί είναι διακυμάνσεις

$$s_{jj} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j) = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = s_j^2$$
 _{$j=1,2,\dots,p$}
 ενώ τα υπόλοιπα στοιχεία
 είναι συνδιακυμάνσεις και μπορεί να είναι
 και αρνητικοί αριθμοί
 Για παράδειγμα οι επόμενοι πίνακες
 δεν είναι πίνακες διακύμανσης

$$\begin{bmatrix} -3 & 1 \\ 1 & -2 \end{bmatrix} \quad \begin{bmatrix} 2 & -3 \\ -3 & 2 \end{bmatrix} \quad \begin{bmatrix} 2 & 0,5 \\ 0,7 & 1 \end{bmatrix}$$

• Το βασικό μειονέκτημα του πίνακα S ως
 μέτρο μεταβλητότητας είναι πως έχοντας
 πολλές μεταβλητές χρειαζόμαστε πολλές τιμές

για να μελετήσουμε την συνολική μεταβλητότητα (σε 26)
-τητα. Δηλαδή δεν μπορούμε να ποσοτικο-
-ποιήσουμε την μεταβλητότητα με έναν
αριθμό. Για να το επιτύχουμε αυτό έχου-
-με δυο μέτρα

- Συνολική διακύμανση $tr(S) = \sum_{i=1}^p s_{ii}$
(ίχνος του πίνακα S δηλαδή άθροισμα
των στοιχείων της κύριας διαγωνίου
του πίνακα)

- Γενικευμένη διακύμανση $|S|$ ορίζουσα του S
Είναι προφανές ότι όσο πιο μεγάλη τιμή
βρούμε και για τα δυο μέτρα τόσο
μεγαλύτερη είναι η μεταβλητότητα στα
δεδομένα. Και τα δυο μέτρα μετρώνε το
πόσο απλωμένες από το διάνυσμα των
μέσων είναι οι παρατηρήσεις μας.

Για την γενικευμένη διακύμανση μηδενική
τιμή ορίζουσας υποδηλώνει ότι κάποιες μεταβλη-
τές είναι γραμμικά εξαρτημένες μεταξύ τους.

Η γενικευμένη διακύμανση μετρά τον "όγκο"^(σε 24)
του πίνακα S

Ισχύει $|S| = \frac{(\text{όγκος})^2}{n^p}$

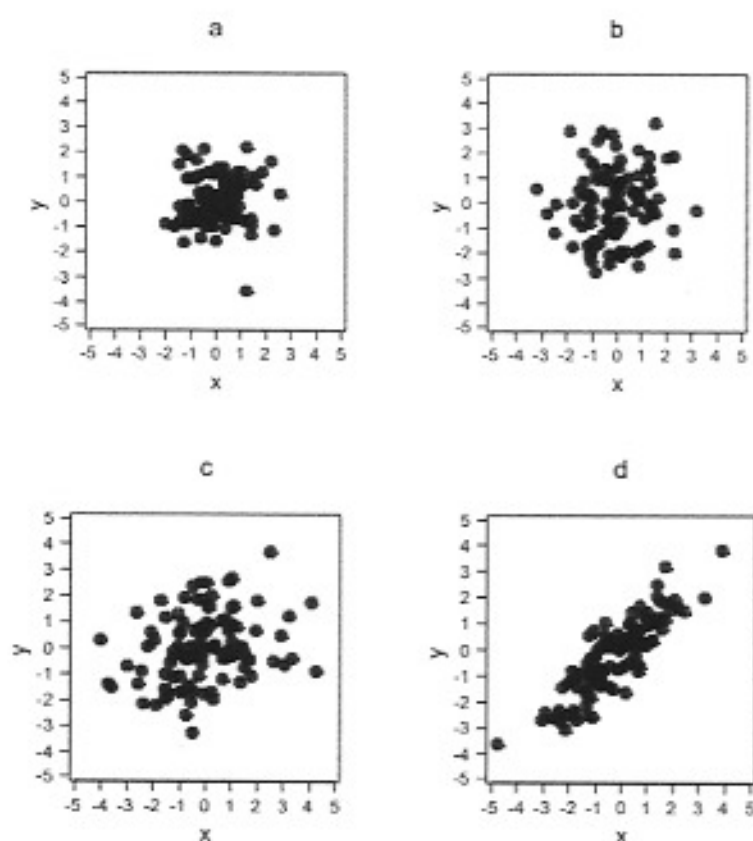
Στην μονομετάβλητη περίπτωση όγκος είναι το άθροισμα των τιμών. Στη διμετάβλητη περίπτωση το εμβαδόν στις τρεις μεταβλητές ο όγκος στο χώρο για περισσότερες από τρεις δεν είναι εύκολο να δούμε

Για να πάρουμε μια ιδέα τα σημεία που εμφανίζονται στα γραφήματα 2.12 αποτελούν δείγματα μεγέθους 100. Οι πίνακες συνδιακυμάνσεων που χρησιμοποιήθηκαν ήταν οι

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1.75 & 0 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0.5 \\ 0.5 & 1 \end{bmatrix} \text{ και } \begin{bmatrix} 2 & 1.65831 \\ 1.65831 & 2 \end{bmatrix}$$

Παρατηρείται πως από την κατασκευή τους οι 3 τελευταίοι πίνακες έχουν την ίδια οριζουσα και άρα την ίδια γενικευμένη διακύμανση, παρόλα αυτά έχουν ολότελα διαφορετική δομή, ο δεύτερος πίνακας υποθέτει την απουσία συσχέτισης ανάμεσα στις μεταβλητές.

Στο διάγραμμα σημείων βλέπουμε πως αν και το εμβαδόν για αυτούς τους τρεις πίνακες είναι παρόμοιο το σχήμα είναι πολύ διαφορετικό, αφού για τον πίνακα 4 υπάρχει έντονη σχέση ανάμεσα στις δύο μεταβλητές και το σχήμα μοιάζει με έλλειψη.



Γράφημα 2.12. Διαγράμματα σημείων για δείγματα μεγέθους 100. Ο πίνακας διακύμανσης του πληθυσμού είχε την ίδια γενικευμένη διακύμανση για τα γραφήματα b,c,d

Δηλαδή το μειονέκτημα της γενικευμένης διακύμανσης είναι πως αγνοεί τη δομή του πίνακα. Αξίζει να παρατηρήσει κανείς πως γενικά είναι πολύ δύσκολο να αναπαρασταθεί όλη η πολυμεταβλητή μεταβλητότητα με έναν μόνο αριθμό. Επίσης δεδομένου πως βασίζεται στην ορίζουσα, η στατιστική συμπερασματολογία σχετικά με τη γενικευμένη διακύμανση είναι δύσκολη.

Εναλλακτικά μέτρα πολυμεταβλητής μεταβλητότητας μπορούν να οριστούν με τη βοήθεια των ιδιοτιμών του πίνακα διακύμανσης.

2.2.4 Πίνακας Συσχετίσεων R

Ο πίνακας συσχέτισεων είναι ο πίνακας που περιέχει σαν στοιχεία του τους συντελεστές συσχέτισης του Pearson για κάθε ζευγάρι μεταβλητών. Σε αυτό το σημείο θα πρέπει να τονίσουμε τα εξής:

- Ο συντελεστής συσχέτισης του Pearson μετράει μόνο γραμμική συσχέτιση ανάμεσα στις μεταβλητές και επομένως δεν μπορεί να μας δώσει πληροφορίες για άλλης μορφής συσχέτιση.
- Υπάρχουν πολλοί άλλοι συντελεστές συσχέτισης στη βιβλιογραφία που είτε μετρούν άλλης μορφής συσχέτιση, πχ ο συντελεστής του Spearman είναι κατάλληλος για κάθε μορφή μονότονης συσχέτισης, είτε για άλλης μορφής δεδομένα. Ο συντελεστής συσχέτισης του Pearson είναι κατάλληλος μόνο για ζεύγη ποσοτικών μεταβλητών.

Επομένως στη γενική του μορφή ο πίνακας συσχέτισεων ορίζεται ως

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{bmatrix},$$

$$\text{όπου } r_{jk} = \frac{S_{jk}}{S_j \cdot S_k} = \frac{S_{jk}}{\sqrt{S_j^2} \cdot \sqrt{S_k^2}}, j, k=1, 2, \dots, p.$$

δηλαδή, το στοιχείο r_{jk} είναι, ο συντελεστής συσχέτισης του Pearson μεταξύ των μεταβλητών X_j και X_k .

Ο πίνακας έχει απαραίτητα τιμές ίσες με τη μονάδα στη διαγώνιο, είναι συμμετρικός και κανένα στοιχείο του δεν μπορεί να πάρει τιμή μεγαλύτερη σε απόλυτη τιμή από το 1. Τιμές -1 και 1 σημαίνουν απόλυτα γραμμική σχέση των δύο μεταβλητών, το πρόσημο υποδηλώνει

την ύπαρξη θετικής ή αρνητικής σχέσης. Η θετική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής τόσο αυξάνει και η τιμή της άλλης ενώ η αρνητική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής μειώνεται η τιμή της άλλης.

Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων $\mathbf{S}$ τυποποιημένων μεταβλητών ταυτίζεται με τον πίνακα συσχετίσεων των αρχικών μεταβλητών πριν την τυποποίηση τους. Εύκολα βλέπει κανείς πως

$$r_{ij} = \frac{S_{ij}}{S_i \cdot S_j} = \frac{\sum_{t=1}^n (X_{it} - \bar{X}_i)(X_{jt} - \bar{X}_j)}{n \cdot S_i \cdot S_j} = \frac{1}{n} \sum_{t=1}^n \left(\frac{X_{it} - \bar{X}_i}{S_i} \right) \left(\frac{X_{jt} - \bar{X}_j}{S_j} \right) = \text{Cov}(X_i^*, X_j^*),$$

όπου X_i^*, X_j^* είναι οι τυποποιημένες μεταβλητές.

Σχετικά με ελέγχους υποθέσεων για συντελεστές συσχέτισης θα μιλήσουμε σε επόμενο κεφάλαιο. Αυτό που πρέπει να τονιστεί είναι πως γενικά μη μηδενικές συσχετίσεις δεν σημαίνουν πως υπάρχει και κάποια αξιοσημείωτη σχέση ανάμεσα στις μεταβλητές. Για παράδειγμα η ελεγχυσμένη σχέση που ελέγχει σε ένα δείγμα μεγέθους n αν η συσχέτιση του πληθυσμού είναι

στατιστικά σημαντικά διαφορετική από το 0 είναι η $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ η οποία ακολουθεί κάτω από

τη μηδενική υπόθεση (και την υπόθεση πως τα δεδομένα προέρχονται από διμεταβλητή κανονική κατανομή) κατανομή t με $n-2$ βαθμούς ελευθερίας. Συνεπώς σε επίπεδο στατιστικής

σημαντικότητας 5% απορρίπτω τη μηδενική υπόθεση αν $\left| \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \right| \geq t_{n-2, 1-\alpha/2}$ ή, λύνοντας ως

προς r , απορρίπτω όταν

$$r^2 \geq \frac{t_{n-2, 1-\alpha/2}^2}{n-2+t_{n-2, 1-\alpha/2}^2}, \text{ όπου } t_{n-2, 1-\alpha/2} \text{ είναι το } 1-\alpha/2 \text{ ποσοστιαίο σημείο της κατανομής } t$$

με n βαθμούς ελευθερίας.

Από την παραπάνω σχέση μπορεί κανείς να υπολογίσει για κάθε μέγεθος δείγματος ποια είναι η τιμή πάνω από την οποία απορρίπτουμε τη μηδενική υπόθεση περί μηδενικής συσχέτισης. Στον πίνακα 2.5 μπορεί κανείς να δει την απόλυτη τιμή του συντελεστή συσχέτισης που μας οδηγεί σε απόρριψη της μηδενικής υπόθεσης πως τα δεδομένα μας είναι ασυσχέτιστα σε επίπεδο στατιστικής σημαντικότητας 5%.

Μέγεθος δείγματος	5	10	15	20	30	50	100	200	500	1000
Τιμή συντελεστή	0.754	0.576	0.482	0.422	0.349	0.273	0.195	0.138	0.088	0.06

Πίνακας 2.5 Η τιμή του συντελεστή συσχέτισης πάνω από την οποία απορρίπτουμε τη μηδενική υπόθεση περί μηδενικής συσχέτισης σε επίπεδο σημαντικότητας 5%.

Από τον πίνακα γίνεται σαφές πως ακόμα και με σχετικά μικρά δείγματα μικρές συσχετίσεις είναι στατιστικά σημαντικές αν και ουσιαστικά αδιάφορες από στατιστικής απόψεως. Θυμηθείτε πως ο συντελεστής συσχέτισης σχετίζεται με τον συντελεστή προσδιορισμού της γραμμικής παλινδρόμησης. Έτσι μια συσχέτιση της τάξης του 0.20, (που για δείγμα μεγέθους 100 είναι στατιστικά σημαντική) σημαίνει πως ο συντελεστής προσδιορισμού σε μια παλινδρόμηση ανάμεσα στις δύο μεταβλητές θα είναι 4%, δηλαδή πάρα πολύ μικρός για οποιδήποτε στατιστική χρήση. Συνεπώς μας ενδιαφέρουν μεγάλες σε απόλυτη τιμή συσχετίσεις και όχι απαραίτητα στατιστικά σημαντικές συσχετίσεις.

Επίσης είναι χρήσιμο να τονιστεί πως όταν παρουσιάζουμε τον πίνακα συσχετίσεων καλό είναι να παρουσιάζουμε μόνο την κάτω διαγώνιο, ώστε ο πίνακας να είναι πιο ευανάγνωστος. Λόγω της συμμετρίας δεν χάνουμε κάποια πληροφορία. Επίσης όταν είναι δυνατόν θα πρέπει να βάζουμε κάποια μορφή διάταξης στις συσχετίσεις ώστε να μπορούμε σχετικά εύκολα να βρούμε αυτές που πραγματικά είναι μεγάλες και άρα χρήσιμες.

Για καλύτερη ερμηνεία, ειδικά όταν έχουμε πολλές μεταβλητές μπορούμε αντί του πίνακα συσχετίσεων να απεικονίσουμε τα δεδομένα με ένα πίνακα όπου θα φαίνονται μόνο οι σημαντικές συσχετίσεις ή έστω οι συσχετίσεις οι οποίες είναι μεγαλύτερες από μία τιμή πρακτικής σημαντικότητας (όπως η τιμή 0.70 που προτείνουν οι Chatfield και Collins, 1992). Έπειτα μπορούμε να απεικονίσουμε τις σχέσεις γραφικά για να έχουμε μια πιο άμεση εικόνα. Άλλη τεχνική είναι να μειώσουμε τα δεκαδικά στοιχεία σε ένα ή δύο για να μπορούμε να εντοπίσουμε καλύτερα σχέσεις μεταξύ των μεταβλητών.

Ο συντελεστής συσχέτισης έχει επίσης ενδιαφέρουσα γεωμετρική ερμηνεία, καθώς μπορεί να αναπαρασταθεί γραφικά. Έτσι ο συντελεστής συσχέτισης είναι το συνημίτονο της γωνίας που προκύπτει από τα δύο διανύσματα με τις παρατηρήσεις για κάθε μεταβλητή. Για αυτό σε πολλές γραφικές απεικονίσεις αντί να παρουσιάζεται ο πίνακας συσχετίσεων, εμφανίζεται ένα γράφημα όπου κάθε συσχέτιση αναπαρίσταται από μια γωνία. Αν η γωνία είναι 90 μοίρες τότε οι δύο μεταβλητές είναι ασυσχέτιστες, ενώ αν η γωνία έχει 0 μοίρες τότε οι μεταβλητές είναι πολύ έντονα συσχετισμένες.

Τέλος θα πρέπει να σημειωθεί πως η έννοια της συσχέτισης και της ανεξαρτησίας δεν ταυτίζονται. Οι ανεξάρτητες μεταβλητές είναι και ασυσχέτιστες αλλά οι ασυσχέτιστες δεν είναι και ανάγκη και ανεξάρτητες. Όπως θα δούμε αργότερα υπάρχουν κάποιες υποθέσεις που μας εξασφαλίζουν πως οι ασυσχέτιστες μεταβλητές είναι και ανεξάρτητες, αλλά γενικά κάτι τέτοιο δεν ισχύει.

Παράδειγμα 2.2. (συνέχεια)

Με βάση τα δεδομένα του παραδείγματος 2.2 που αφορούσαν τις 20 αμερικάνικες πόλεις υπολογίζουμε τα ακόλουθα περιγραφικά μέτρα:

Το διάνυσμα των μέσων είναι

$$\bar{X} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \\ \bar{x}_5 \end{bmatrix} = \begin{bmatrix} 0.948 \\ 2.984 \\ 7.286 \\ 1.210 \\ 0.962 \end{bmatrix}$$

Για τον πίνακα διακύμανσης βρούμε πως είναι ο

$$S = \begin{bmatrix} 0.024 & -0.065 & 0.090 & 0.010 & 0.015 \\ -0.065 & 1.531 & 0.198 & -0.112 & -0.047 \\ 0.090 & 0.198 & 4.479 & 0.000 & -0.011 \\ 0.010 & -0.112 & 0.000 & 0.136 & 0.022 \\ 0.015 & -0.047 & -0.011 & 0.022 & 0.015 \end{bmatrix}$$

Εδώ θα πρέπει να σημειώσουμε το εξής: τα περισσότερα στατιστικά πακέτα υπολογίζουν έναν λίγο διαφορετικό πίνακα διακύμανσης. Συγκεκριμένα ο παρονομαστής δεν είναι n αλλά $n-1$, για τον ίδιο λόγο και στην μονομεταβλητή περίπτωση χρησιμοποιούμε την αμερόληπτη διακύμανση. Δηλαδή έχει αποδειχτεί πως ο δειγματικός πίνακας διακύμανσης δεν είναι αμερόληπτος και για αυτό διαιρούμε με $n-1$ για να πάρουμε τον αμερόληπτο πίνακα διακύμανσης S^* . Είναι εμφανές πως η σχέση που συνδέει τους δύο πίνακες είναι η

$$S^* = \frac{nS}{n-1}$$

Συνεπώς για τα δεδομένα μας βρούμε πως

$$S^* = \begin{bmatrix} 0.025 & -0.068 & 0.094 & 0.010 & 0.015 \\ -0.068 & 1.612 & 0.206 & -0.118 & -0.049 \\ 0.094 & 0.206 & 4.714 & 0.000 & -0.012 \\ 0.010 & -0.118 & 0.000 & 0.143 & 0.022 \\ 0.015 & -0.049 & -0.012 & 0.022 & 0.015 \end{bmatrix}$$

Παρατηρείται πως μόνο εκεί που είχαμε μεγάλες τιμές υπάρχουν σαφείς διαφορές ενώ σε μικρότερες τιμές λόγω στρογγυλοποίησης οι τιμές είναι ίδιες.

Ο πίνακας συσχετίσεων είναι ο ίδιος άσχετα αν πάρουμε τον μεροληπτικό ή τον αμεροληπτο πίνακα διακύμανσης. Για τα δεδομένα μας είναι

$$R = \begin{bmatrix} 1 & -0.342 & 0.277 & 0.164 & 0.752 \\ -0.342 & 1 & 0.076 & -0.246 & -0.311 \\ 0.277 & 0.076 & 1 & 0.000 & -0.043 \\ 0.164 & -0.246 & 0.000 & 1 & 0.488 \\ 0.752 & -0.311 & -0.043 & 0.488 & 1 \end{bmatrix}.$$

Παρατηρείστε λοιπόν πως υπάρχουν κάποιες αρκετά μεγάλες συσχετίσεις όπως αυτή του πληθυσμού με τον αριθμό καταστημάτων που είναι 0.756. Επίσης υπάρχουν και μερικές αρνητικές συσχετίσεις, όπως η συσχέτιση ανάμεσα στον αριθμό καταστημάτων και τον αριθμό πρατηρίων χονδρικής.

Χρησιμοποιώντας τον πίνακα S' βρίσκουμε πως η συνολική διακύμανση είναι το άθροισμα των διαγώνιων στοιχείων του πίνακα δηλαδή 6.51 και η γενικευμένη διακύμανση είναι 0.00007245. Θα πρέπει να τονιστεί πως τόσο η συνολική όσο και η γενικευμένη διακύμανση ελάχιστα μπορούν να μας φανερώσουν για τα δεδομένα παρά μόνο αν τις χρησιμοποιήσουμε συγκριτικά με άλλα σετ δεδομένων για να δούμε πιο σετ έχει μεγαλύτερη μεταβλητότητα.

Ας δούμε πως μπορούμε να αναπαράγουμε με αλγεβρα πινάκων το διάνυσμα των μέσων και τον πίνακα διακύμανσης και συσχέτισης

Εστω $X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$ ο πίνακας δεδομένων

Η i παρατήρηση έχει την μορφή

$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \text{ και σαν γραμμή}$$

$$X_i^T = [x_{i1} \quad x_{i2} \quad \dots \quad x_{ip}]$$

Ισχύει $X_i^T \cdot X_i = \sum_{j=1}^p x_{ij}^2$ και

$$X_i X_i^T = \begin{bmatrix} x_{i1}^2 & x_{i1}x_{i2} & \dots & x_{i1}x_{ip} \\ x_{i1}x_{i2} & x_{i2}^2 & & \\ \vdots & & \ddots & \\ x_{i1}x_{ip} & \dots & & x_{ip}^2 \end{bmatrix}$$

Επομένως το διάνυσμα των μέσων

αμών

$$\bar{X} = \frac{1}{n} X_{p \times n}^T \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1} = \frac{1}{n} X^T \cdot \mathbf{1}_{1 \times n}$$