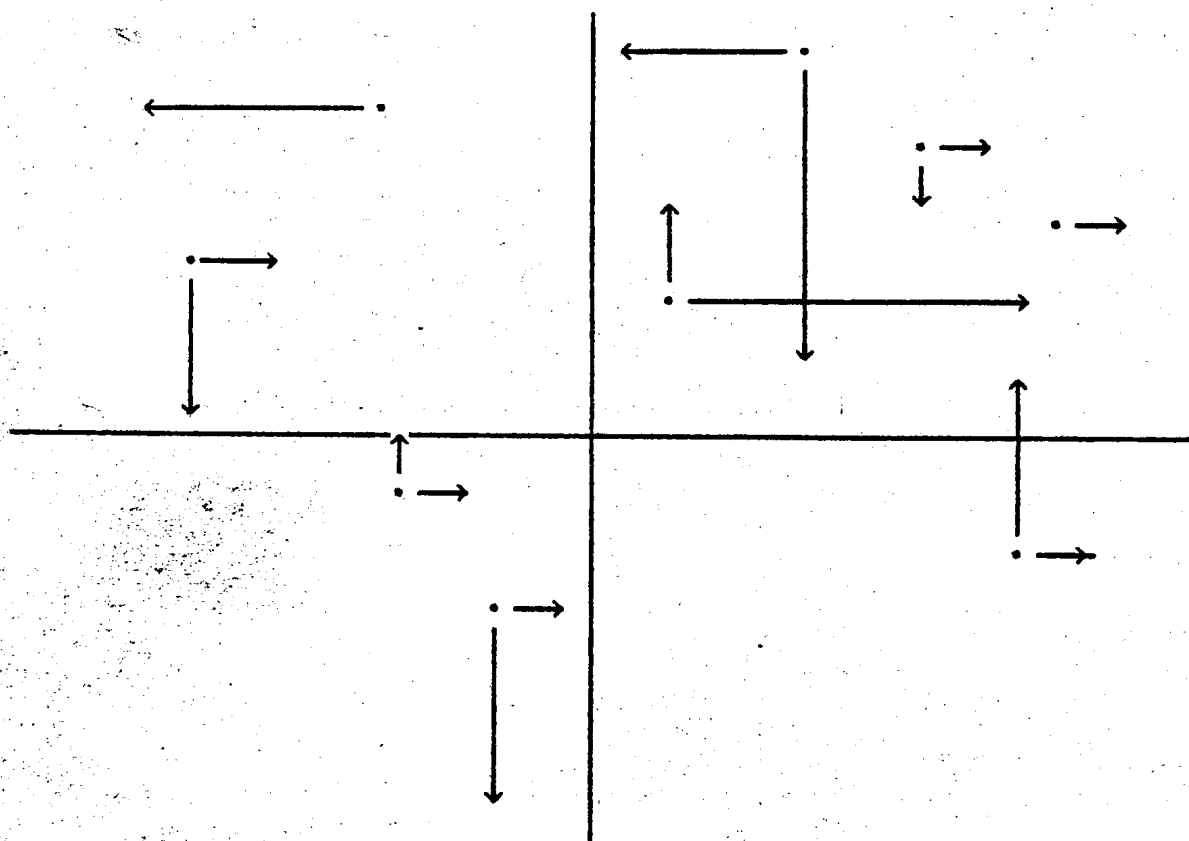
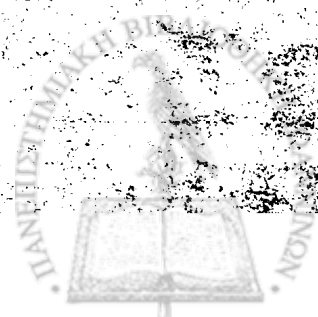


ΕΙΣΑΓΩΓΗ
ΣΤΗΝ
ΠΟΛΥΔΙΑΣΤΑΤΗ ΑΝΑΛΥΣΗ

Κ. Ξ. ΚΑΡΑΚΩΣΤΑΣ
ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ



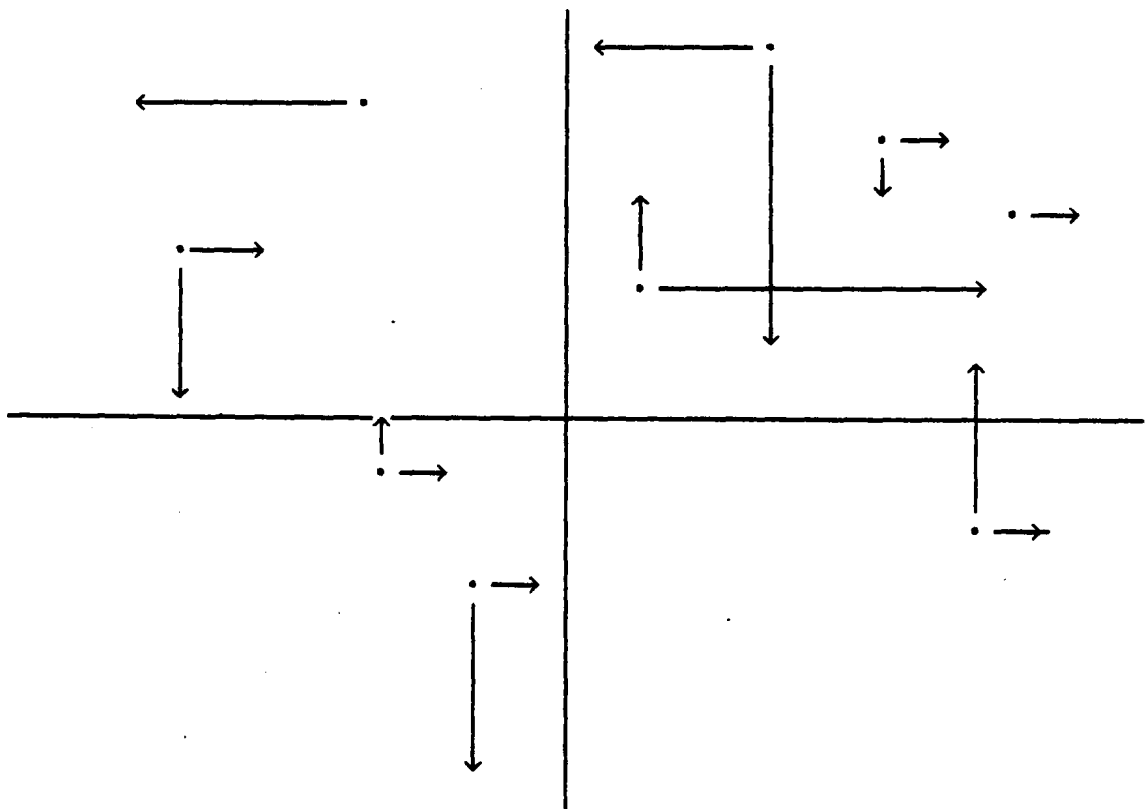
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΙΩΑΝΝΙΝΑ 1993



5 ε.

ΕΙΣΑΓΩΓΗ
ΣΤΗΝ
ΠΟΛΥΔΙΑΣΤΑΤΗ ΑΝΑΛΥΣΗ

Κ. Ξ. ΚΑΡΑΚΩΣΤΑΣ
ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ



ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΙΩΑΝΝΙΝΑ 1993



Στην σύζυγό μου **Στην κόρη μου**

Βούλα **Ναταλία**

Επειδὴ γενικεύονται, ἀπὸ τὴν ἐκτενέστερον ἀνάλυσιν πολλοὺς
ἀπαιτοῦν ὅπως : πολλὰ ποσότης ἐκτενέστερον καὶ περὶ ἑνὸς
ἐκτενέστερον, ἐκτενέστερον ἀνάλυσιν, ἐκτενέστερον καὶ ἀνάλυσιν
ἐκτενέστερον καὶ.

Οι πληροφορίες αυτές επιβεβαιώνουν μια συνήθη άποψη, η οποία
πρέπει να ληφθεί υπό την μορφοποιημένη ανάλυση του κ. Γουλιέλμου
Θεράπα της μορφοποιημένης ανάλυσης. Γράφοντας για να επιβεβαιωθεί τις
απόψεις του μεθυστικού "Εξέλιξη Θεράπα Στοιχειώδη" που εξετάζεται

Π Ρ Ο Λ Ο Γ Ο Σ

Η πολυδιάστατη ανάλυση είναι ο κλάδος εκείνος της στατιστικής που ασχολείται με την μελέτη και ανάλυση πολυδιάστατων δεδομένων. Σαν πολυδιάστατα δεδομένα εννοούμε την καταγραφή των τιμών πολλών διαφορετικών τυχαίων μεταβλητών σε κάθε πειραματική μονάδα. Οι μεταβλητές αυτές δεν είναι απαραίτητο να είναι όλες του ίδιου τύπου (π.χ. ποσοτικές ή ποιοτικές). Πολυδιάστατα δεδομένα συναντάμε συχνά στις κοινωνικές, ψυχολογικές και ιατρικές επιστήμες.

Τα θέματα της πολυδιάστατης ανάλυσης είναι αφ' ενός μεν γενικεύσεις αντιστοίχων θεμάτων της μονοδιάστατης ανάλυσης αφ' ετέρου δε η μελέτη και ανάπτυξη καινούργιων μεθόδων και τεχνικών. Οι μέθοδοι και οι τεχνικές αυτές μας είναι χρήσιμες για την καλύτερη κατανόηση της πολύπλοκης δομής των πολυδιάστατων δεδομένων.

Στις γενικεύσεις από την μονοδιάστατη ανάλυση ανήκουν θέματα όπως : πολυδιάστατες κατανομές και ροπές αυτών, εκτιμητική, έλεγχος υποθέσεων, παλινδρόμηση και ανάλυση διακύμανσης κ.ά.

Στις καινούργιες μεθόδους και τεχνικές ανήκουν θέματα όπως: κύριες συνιστώσες, διαχωριστική ανάλυση, παραγοντική ανάλυση, ταξινομική ανάλυση κ.ά.

Οι σημειώσεις αυτές επιχειρούν μια σύντομη εισαγωγή τόσο στις γενικεύσεις από την μονοδιάστατη ανάλυση όσο και στα ειδικά θέματα της πολυδιάστατης ανάλυσης. Γράφτηκαν για να καλύψουν τις ανάγκες του μαθήματος "Ειδικά θέματα Στατιστικής" που διδάσκεται



στο 8^ο εξάμηνο του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων. Αποτέλεσμα του σκοπού για τον οποίο γράφτηκαν οι σημειώσεις αυτές είναι να απευθύνονται, κατά κύριο λόγο, σε Μαθηματικούς. Ένας που χρησιμοποιεί μεθόδους της πολυδιάστατης ανάλυσης δεν θα μπορούσε να οφεληθεί από την ανάγνωση αυτών των σημειώσεων.

Συνολικά οι σημειώσεις αυτές αποτελούνται από έξι κεφάλαια. Τα τέσσερα πρώτα κεφάλαια αποτελούν γενίκευση αντίστοιχων θεμάτων στην μονοδιάστατη ανάλυση. Τα δύο τελευταία είναι στην κατεύθυνση των νέων μεθόδων και τεχνικών που αναφέραμε προηγουμένως.

Πιο συγκεκριμένα στο 1^ο κεφάλαιο δίνονται οι απαραίτητοι συμβολισμοί, η έννοια της πολυδιάστατης κατανομής και οι αντίστοιχες ροπές της. Το κεφάλαιο ολοκληρώνεται με την απόδειξη μερικών θεωρημάτων τα οποία αποτελούν γενικεύσεις αντίστοιχων θεωρημάτων στην μονοδιάστατη περίπτωση.

Στο 2^ο κεφάλαιο αναφέρονται οι πιο συχνά χρησιμοποιούμενες πολυδιάστατες κατανομές με μερικές από τις ιδιότητές τους.

Μέθοδοι για την εκτίμηση των παραμέτρων των πολυδιάστατων κατανομών με έμφαση, λόγου του ρόλου που παίζει στην πολυδιάστατη ανάλυση, την εκτίμηση των παραμέτρων της πολυδιάστατης κανονικής κατανομής, παρουσιάζονται στο 3^ο κεφάλαιο.

Οι διαφορές που υπάρχουν μεταξύ του ελέγχου υποθέσεων για την μονοδιάστατη και την πολυδιάστατη ανάλυση αναπτύσσονται στο 4^ο κεφάλαιο. Στο ίδιο κεφάλαιο παρουσιάζονται διάφοροι έλεγχοι υποθέσεων για τις παραμέτρους της πολυδιάστατης κανονικής κατανομής.

Από τα ειδικά θέματα της πολυδιάστατης ανάλυσης παρουσιάζ-



ζονται, συνοπτικά, μόνο δύο. Αυτά των κυρίων συνιστωσών, στο 5^ο κεφάλαιο, και της διαχωριστικής ανάλυσης, στο 6^ο κεφάλαιο.

Η επιμέλεια των κειμένων καθώς και οι διάφορες διορθώσεις στα αρχικά χειρόγραφα έγιναν από την συζυγό μου, Βούλα Νησσοπούλου, και την κόρη μου Ναταλία, τις οποίες και ευχαριστώ.

Δεν θα χρειαζόταν ίσως να αναφερθεί ότι ο μόνος υπεύθυνος για κάθε λάθος ή παράλειψη είναι ο συγγραφέας.

Ιωάννινα, Ιανουάριος 1993



Π Ε Ρ Ι Ε Χ Ο Μ Ε Ν Α

1. ΕΙΣΑΓΩΓΗ	1
1.1 ΓΕΝΙΚΑ	1
1.2 ΣΥΜΒΟΛΙΣΜΟΙ	3
1.3 ΡΟΠΕΣ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ	7
ΑΣΚΗΣΕΙΣ	14
2. ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ	17
2.1 ΠΟΛΥΔΙΑΣΤΑΤΗ ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ	17
2.1.1 Υποδιανύσματα, ανεξαρτησία, δεσμευμένες πιθανότητες	21
2.1.2 Τετραγωνικές μορφές	23
2.2 ΔΕΙΓΜΑΤΙΚΕΣ ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ	30
2.2.1 Κατανομή του Wishart	30
2.2.2 Κατανομή του Hotelling	35
2.2.3 Πολυδιάστατη t - κατανομή	41
2.3 ΜΕΡΙΚΕΣ ΑΛΛΕΣ ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ	42
2.3.1 Πολυωνυμική κατανομή	42
2.3.2 Βήτα κατανομή ή κατανομή του Dirichlet	42
ΑΣΚΗΣΕΙΣ	44
3. ΕΚΤΙΜΗΤΙΚΗ	49
3.1 ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΚΑΙ ΕΠΑΡΚΕΙΑ	49
3.1.1 Συνάρτηση πιθανοφάνειας	49
3.1.2 Επάρκεια και πληρότητα	53
3.1.3 Αμερόληπτοι εκτιμητές ελάχιστης διακύμανσης	57
3.2 ΕΚΤΙΜΗΤΕΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ	58



3.2.1 Γενικά	58
3.2.2 Πολυδιάστατη κανονική κατανομή	59
ΑΣΚΗΣΕΙΣ	62
4. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ	63
4.1 ΓΕΝΙΚΑ	63
4.2 ΜΕΘΟΔΟΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ	67
4.3 ΜΕΘΟΔΟΣ ΕΝΩΣΗΣ - ΤΟΜΗΣ	69
4.4 ΕΦΑΡΜΟΓΕΣ	71
4.4.1 Έλεγχος μέσων τιμών	71
4.4.1.1 Έλεγχος για τη μέση τιμή ενός κανονικού πληθυσμού	71
4.4.1.2 Έλεγχος για τη μέση τιμή δύο κανονικών πληθυσμών	76
4.4.1.3 Έλεγχος ζευγών	80
4.4.2 Πολλαπλές συγκρίσεις για τις μέσες τιμές	81
4.4.2.1 Πολλαπλές συγκρίσεις για τη μέση τιμή ενός κανονικού πληθυσμού	81
4.4.2.2 Πολλαπλές συγκρίσεις για τη διαφορά των μέσων τιμών δύο κανονικών πληθυσμών	82
4.4.3 Έλεγχος διακυμάνσεων	82
4.4.3.1 Έλεγχος για τον πίνακα Σ ενός κανονικού πληθυσμού	82
4.4.3.2 Έλεγχος για την ισότητα των πινάκων διακυμάνσεων-συνδιακυμάνσεων δύο ή περισσότερων κανονικών πληθυσμών	85
ΑΣΚΗΣΕΙΣ	88



5. ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ	91
5.1 ΓΕΝΙΚΑ	91
5.2 ΕΥΡΕΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ	93
5.2.1 Κύριες συνιστώσες από τον πίνακα διακυμάνσεων- συνδιακυμάνσεων	93
5.2.2 Κύριες συνιστώσες από τον πίνακα συσχέτισεως	97
5.3 ΕΚΤΙΜΗΣΗ ΤΩΝ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ	98
5.4 ΕΡΜΗΝΕΙΑ - ΧΡΗΣΕΙΣ - ΣΧΟΛΙΑ	100
5.4.1 Ερμηνεία	100
5.4.2 Χρήσεις	101
5.4.3 Σχόλια	107
5.5 ΕΛΕΓΧΟΣ ΓΙΑ ΤΗΝ ΠΟΛΥΔΙΑΣΤΑΤΗ ΚΑΝΟΝΙΚΗ ΑΣΚΗΣΕΙΣ	112 115
6. ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ	117
6.1 ΕΙΣΑΓΩΓΗ	117
6.2 ΠΕΡΙΠΤΩΣΗ ΔΥΟ ΠΛΗΘΥΣΜΩΝ	118
6.2.1 Ταξινόμηση σε δύο γνωστούς κανονικούς πληθυσμούς με κοινό πίνακα διακυμάνσεων-συνδιακυμάνσεων	118
6.2.1.1 Διαχωριστική συνάρτηση του <i>Fisher</i>	118
6.2.1.2 Μέθοδος του πηλίκου πιθανοφάνειας	121
6.2.2 Ταξινόμηση σε δύο κανονικούς πληθυσμούς με άγνωστες μέσες τιμές αλλά γνωστό τον κοινό πίνακα διακυμάνσεων-συνδιακυμάνσεων	121
6.2.3 Ταξινόμηση σε δύο κανονικούς πληθυσμούς με άγνωστες παραμέτρους αλλά κοινό πίνακα διακυμάνσεων - συνδιακυμάνσεων	127



6.2.4 Ταξινόμηση σε δύο γνωστούς κανονικούς πληθυσμούς με διαφορετικούς πίνακες διακυμάνσεων - συνδιακυμάνσεων	129
6.3 ΠΕΡΙΠΤΩΣΗ ΤΡΙΩΝ Ή ΠΕΡΙΣΣΟΤΕΡΩΝ ΠΛΗΘΥΣΜΩΝ	130
ΑΣΚΗΣΕΙΣ	133
ΠΑΡΑΡΤΗΜΑ	134
ΒΙΒΛΙΟΓΡΑΦΙΑ	137



1. ΕΙΣΑΓΩΓΗ

1.1 ΓΕΝΙΚΑ

Η πολυδιάστατη ή πολυμεταβλητή ανάλυση ασχολείται με δεδομένα τα οποία είναι παρατηρήσεις σε δύο ή περισσότερες τυχαίες μεταβλητές, όπου κάθε μεταβλητή μετρείται σε ένα σύνολο πειραματικών μονάδων.

Παράδειγμα 1.1 Ένα παράδειγμα πολυμεταβλητών δεδομένων είναι και οι βαθμοί των φοιτητών στα διάφορα μαθήματα όπως φαίνεται και στον πίνακα που ακολουθεί.

Πίνακας 1.1
Βαθμοί φοιτητών σε διάφορα μαθήματα

Όνομα	Ανάλυση	Υπολογιστές	Στατιστική	Άλγεβρα
Adams, E	58	43	45	61
Bloggs, J	43	58	59	48
Fisher, R	37	35	43	52

κ. λ. π.

Παράδειγμα 1.2 Τα παρακάτω δεδομένα είναι οι μετρήσεις μας σε πέντε φοιτητές (πειραματικές μονάδες). Για κάθε φοιτητή



καταγράφουμε τις τιμές τριών μεταβλητών. Της μεταβλητής X_1 που παριστάνει την ηλικία (σε χρόνια) του φοιτητή κατά την είσοδό του στο Πανεπιστήμιο. Της μεταβλητής X_2 που παριστάνει τη μέση βαθμολογία (με άριστα το 100) του φοιτητή στο τέλος του πρώτου έτους και τέλος την μεταβλητή X_3 που παριστάνει το φύλο του φοιτητή.

Πίνακας 1.2

	Φοιτητής	Μεταβλητές		
		X_1	X_2	$X_3^{(*)}$
Πειραματικές Μονάδες	1	18,45	70	1
	2	18,41	65	0
	3	18,39	71	0
	4	18,70	72	0
	5	18,34	94	1

(*) Το ένα (1) αντιστοιχεί σε αγόρι και το μηδέν (0) σε κορίτσι.

Από το τελευταίο παράδειγμα γίνεται φανερό ότι στην πολυμεταβλητή ανάλυση οι μεταβλητές που εξετάζουμε δεν είναι απαραίτητο να είναι όλες του ίδιου τύπου (π.χ. όλες ποσοτικές ή όλες ποιοτικές).

Η ανάγκη για χρήση πολυμεταβλητών μεθόδων προέκυψε από την διαπίστωση ότι αρκετές φορές οι υπό εξέταση μεταβλητές είναι συσχετισμένες μεταξύ τους. Αυτό έχει σαν αποτέλεσμα να μην είναι σωστή η μεμονομένη εξέταση κάθε μεταβλητής. Αλλά και στην περίπτωση που οι υπό εξέταση μεταβλητές είναι ασυσχέτιστες μεταξύ τους η αξία της πολυμεταβλητής ανάλυσης είναι μεγάλη διότι εξετάζει τις μετρήσεις σαν ένα ενιαίο σύστημα μετρήσεων.

Ο πίνακας που ακολουθεί μας δίνει αφ' ενός μεν τις διάφορες μεθόδους στην πολυμεταβλητή ανάλυση, αφ' ετέρου δε και το που μπορεί να χρησιμοποιειθεί η κάθε μία.



Πίνακας 1.3

Π Λ Η Θ Υ Σ Μ Ο Ι

Μ Ε Τ Α Β Λ Η Τ Ε Σ	Π Λ Η Θ Υ Σ Μ Ο Ι	
	Ενας Πληθυσμός	Δύο ή Περισσότεροι Πληθυσμοί
Ένα Σύνολο	Κύριες Συνιστώσες Παραγοντική Ανάλυση	Πολυδιάστατη Ανάλυση Διακύμανσης Διαχωριστική Ανάλυση Ταξινομική Ανάλυση
	Δύο ή Περισσότερα Σύνολα	Πολυδιάστατη Ανάλυση Συνδιακύμανσης

1.2 ΣΥΜΒΟΛΙΣΜΟΙ

Με n θα συμβολίζουμε τον αριθμό των πειραματικών μονάδων, ενώ με p τον αριθμό των υπό εξέταση μεταβλητών. X_j είναι η j ($j = 1, 2, \dots, p$) μεταβλητή και x_{ij} είναι η παρατήρηση στην i ($i = 1, 2, \dots, n$) πειραματική μονάδα για την j μεταβλητή. Με το συμβολισμό αυτό τα δεδομένα μας, σε ένα πρόβλημα πολυμεταβλητής ανάλυσης, θα έχουν την παρακάτω μορφή.

Πίνακας 1.4

	Μ Ε Τ Α Β Λ Η Τ Ε Σ				
	X_1	X_2	X_3	$\dots$	X_p
O_1	x_{11}	x_{12}	x_{13}	$\dots$	x_{1p}
O_2	x_{21}	x_{22}	x_{23}	$\dots$	x_{2p}
O_3	x_{31}	x_{32}	x_{33}	$\dots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
O_n	x_{n1}	x_{n2}	x_{n3}	$\dots$	x_{np}

ΠΕΙΡΑΜΑΤΙΚΕΣ
ΜΟΝΑΔΕΣ



Τον πίνακα των δεδομένων μας θα τον συμβολίζουμε με X , δηλαδή $X = (x_{ij})$. Προφανώς ο X είναι ένας $n \times p$ πίνακας.

Ας συμβολίσουμε με

$$\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

την δειγματική μέση τιμή για την j μεταβλητή, με

$$s_{jj}^2 = s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_{.j})^2$$

την δειγματική διακύμανση για την j μεταβλητή και με

$$s_{jj'} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_{.j})(x_{ij'} - \bar{x}_{.j'})$$

τη δειγματική συνδιακύμανση μεταξύ των μεταβλητών X_j και $X_{j'}$.

Αν τώρα θέσουμε $\mathbf{x} = (X_1, X_2, \dots, X_p)'$, τότε $\mathbf{x}$ είναι ένα $p \times 1$ διάνυσμα με μέση τιμή $\bar{\mathbf{x}} = (\bar{x}_{.1}, \bar{x}_{.2}, \dots, \bar{x}_{.p})'$ ή κάνοντας χρήση πινάκων

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{X}' \mathbf{1},$$

όπου $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})'$ και $\mathbf{1} = (1, 1, \dots, 1)'$.

Ο δειγματικός πίνακας διακυμάνσεων-συνδιακυμάνσεων του $\mathbf{x}$ θα είναι

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' - \bar{\mathbf{x}} \bar{\mathbf{x}}'$$

$$= \frac{1}{n} \mathbf{X}' \mathbf{X} - \bar{\mathbf{x}} \bar{\mathbf{x}}' = \frac{1}{n} \left(\mathbf{X}' \mathbf{X} - \frac{1}{n} \mathbf{X}' \mathbf{1} \mathbf{1}' \mathbf{X} \right) = \frac{1}{n} \mathbf{X}' \mathbf{H} \mathbf{X},$$

όπου $\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1} \mathbf{1}'$, και $\mathbf{I}_n$ είναι ο μοναδιαίος πίνακας τάξεως n .

Από τον ορισμό του πίνακα $\mathbf{H}$ προκύπτει ότι ο $\mathbf{H}$ είναι συμμετρικός και ταυτοδύναμος (idempotent). Άρα για κάθε διάνυσμα $\mathbf{a}$ θα έχουμε:

$$\mathbf{a}' \mathbf{S} \mathbf{a} = \frac{1}{n} \mathbf{a}' \mathbf{X}' \mathbf{H} \mathbf{X} \mathbf{a} = \frac{1}{n} \mathbf{a}' \mathbf{X}' \mathbf{H}' \mathbf{H} \mathbf{X} \mathbf{a} = \frac{1}{n} \mathbf{y}' \mathbf{y} \geq 0,$$



όπου $y = H\alpha$. Από αυτό συμπεραίνουμε ότι ο δειγματικός πίνακας διακυμάνσεων-συνδιακυμάνσεων S είναι θετικώς ημιορισμένος. Για συνεχή δεδομένα ο πίνακας S είναι θετικά ορισμένος εάν $n \geq p+1$

Σε αναλογία με την μονοδιάστατη περίπτωση ορίζουμε και τον αμερόληπτο δειγματικό πίνακα διακυμάνσεων - συνδιακυμάνσεων με τη σχέση

$$S_u = \frac{1}{n-1} X' H X = \frac{n}{n-1} S.$$

Από τον προηγούμενο συμβολισμό και από τον ορισμό του συντελεστή συσχέτισης r , μεταξύ δύο τυχαίων μεταβλητών X_j και $X_{j'}$, προκύπτει ότι $r_{jj'} = s_{jj'} / (s_j s_{j'})$. Ο $p \times p$ πίνακας $R = (r_{jj'})$ θα ονομάζεται δειγματικός πίνακας συσχέτισης. Εύκολα μπορούμε να αποδείξουμε ότι μεταξύ των πινάκων R και S ισχύουν οι σχέσεις:

$$i) R = D^{-1} S D^{-1}, \quad ii) S = D R D \quad \text{όπου } D = \text{diag}(s_j).$$

Παράδειγμα 1.3 Τα παρακάτω δεδομένα είναι οι ποσότητες φελλού, σε εκατοστά του γραμμαρίου, σε 28 δέντρα ως προς τις τέσσερις διευθύνσεις του ορίζοντα :

A = ανατολικά , Δ = δυτικά , Β = βόρεια , Ν = νότια

B	A	N	Δ	B	A	N	Δ	B	A	N	Δ	B	A	N	Δ
72	66	76	77	42	43	31	25	91	79	100	75	32	30	30	32
60	53	66	63	37	40	31	25	56	68	47	50	60	50	67	54
56	57	64	58	33	29	27	36	79	65	70	61	35	37	48	39
41	29	36	38	32	30	34	28	81	80	68	58	39	36	39	31
32	32	35	36	63	45	74	63	78	55	67	60	50	34	37	40
30	35	34	26	54	46	60	52	46	38	37	38	43	37	39	50
39	39	31	27	47	51	52	43	39	35	34	37	48	54	57	43

Με βάση τα δεδομένα αυτά θα έχουμε :

$\bar{x} = (50,536, 46,179, 49,679, 45,179)'$. Από τις τιμές των μέσων τιμών μπορούμε να συμπεράνουμε ότι υπάρχει μεγαλύτερη ποσότητα φελλού στην κατεύθυνση Β-Ν παρά στην κατεύθυνση Α-Δ.



Ο δειγματικός πίνακας διακυμάνσεων-συνδιακυμάνσεων S καθώς και ο δειγματικός πίνακας συσχέτισης R είναι

$$S = \begin{matrix} & \begin{matrix} B & A & N & \Delta \end{matrix} \\ \begin{matrix} B \\ A \\ N \\ \Delta \end{matrix} & \begin{pmatrix} 280,034 & 215,761 & 278,136 & 218,190 \\ & 212,075 & 220,879 & 165,254 \\ & & 337,504 & 250,272 \\ & & & 217,932 \end{pmatrix} \end{matrix}$$

$$R = \begin{matrix} & \begin{matrix} B & A & N & \Delta \end{matrix} \\ \begin{matrix} B \\ A \\ N \\ \Delta \end{matrix} & \begin{pmatrix} 1 & 0,885 & 0,905 & 0,883 \\ & 1 & 0,826 & 0,769 \\ & & 1 & 0,923 \\ & & & 1 \end{pmatrix} \end{matrix}$$

Επειδή οι πίνακες S και R είναι συμμετρικοί γι' αυτό έχουμε γράψει τον άνω τριγωνικό πίνακα σε κάθε περίπτωση.

Από τον πίνακα S μπορούμε να δούμε ότι η νότια κατεύθυνση έχει την μεγαλύτερη διακύμανση. Ακόμη, από τον πίνακα R , διαπιστώνουμε ότι υπάρχει μεγάλη συσχέτιση μεταξύ B και N ενώ η συσχέτιση μεταξύ A και Δ είναι μικρότερη.

Ο πίνακας S των δειγματικών διακυμάνσεων-συνδιακυμάνσεων είναι ένα μέτρο πολυδιάστατης διασποράς. Πολλές φορές όμως χρειαζόμαστε ένα μέτρο πολυδιάστατης διασποράς το οποίο να εκφράζεται με ένα και μόνο αριθμό. Τέτοια μέτρα είναι :

i) Η γενικευμένη διακύμανση , $\det(S)$ και

ii) Η ολική μεταβλητότητα , $\text{tr}(S)$,

όπου $\det(\cdot)$ και $\text{tr}(\cdot)$ συμβολίζουν, αντίστοιχα, την ορίζουσα και το ίχνος ενός πίνακα.



1.3 ΡΟΠΕΣ ΠΟΛΥΔΙΑΣΤΑΤΩΝ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Οι έννοιες, οι ορισμοί καθώς και οι ιδιότητες που θα αποδειχθούν σ' αυτήν την παράγραφο, αποτελούν όλα γενικεύσεις αντιστοιχών εννοιών, ορισμών και ιδιοτήτων για την μονοδιάστατη περίπτωση.

Έτσι, αν $X_1, X_2, \dots, X_p$ είναι τυχαίες μεταβλητές (τ. μ.) τότε το p -διάστατο διάνυσμα $\mathbf{x} = (X_1, X_2, \dots, X_p)'$ ονομάζεται p -διάστατο τυχαίο διάνυσμα (τ. δ.).

Για κάθε p -διάστατο τ. δ. ορίζεται η αθροιστική κατανομή από την σχέση :

$$F(\mathbf{x}^0) = \text{πιθ}(\mathbf{x} \leq \mathbf{x}^0) = \text{πιθ}(X_1 \leq x_1^0, X_2 \leq x_2^0, \dots, X_p \leq x_p^0)$$

όπου $\mathbf{x}^0 = (x_1^0, x_2^0, \dots, x_p^0)'$ είναι μια γνωστή τιμή στο πεδίο τιμών $D \subset \mathbb{R}^p$ του $\mathbf{x}$.

Αν όλες οι συνιστώσες του τ. δ. $\mathbf{x}$ είναι συνεχείς τ. μ. τότε η συνάρτηση πυκνότητας πιθανότητας $f(\mathbf{x})$ του $\mathbf{x}$ ικανοποιεί τις σχέσεις :

$$\text{i) } f(\mathbf{x}) \geq 0, \forall \mathbf{x} \in D \subset \mathbb{R}^p \quad \& \quad \text{ii) } \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(\mathbf{x}) d\mathbf{x} = 1.$$

Στην περίπτωση που όλες οι συνιστώσες του τ. δ. $\mathbf{x}$ είναι απαριθμητές (διακριτές) τ. μ. τότε η συνάρτηση πιθανότητας $p(\mathbf{x})$ του $\mathbf{x}$ ικανοποιεί τις σχέσεις :

$$\text{i) } p(\mathbf{x}) \geq 0, \forall \mathbf{x} \in D \subset \mathbb{R}^p \quad \& \quad \text{ii) } \sum_{x_1} \sum_{x_2} \dots \sum_{x_p} p(\mathbf{x}) = 1.$$

Σε ότι ακολουθεί θα μιλάμε για την κατανομή ενός τ. δ. $\mathbf{x}$ και θα την συμβολίζουμε με $f(\mathbf{x})$ ανεξάρτητα από τον τύπο των συνιστωσών του (συνεχείς ή διακριτές).



Αν $\mathbf{x}_1, \mathbf{x}_2$ είναι δύο τ. δ. με από κοινού κατανομή $f(\mathbf{x}_1, \mathbf{x}_2)$ τότε η κατανομή περιθωρίου του τ. δ. $\mathbf{x}_1$ ορίζεται από την σχέση

$$f_{\mathbf{x}_1}(\mathbf{x}_1) = \int_{-\infty}^{+\infty} f(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_2.$$

Τέλος η υπό συνθήκη κατανομή του τ. δ. $\mathbf{x}_1$ δοθέντος του τ. δ. $\mathbf{x}_2$ ορίζεται από την σχέση :

$$f_{\mathbf{x}_1|\mathbf{x}_2}(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f_{\mathbf{x}_2}(\mathbf{x}_2)}.$$

Ορισμός 1.1 Αν $\mathbf{x}$ είναι ένα pxl τ. δ. τότε η αναμενομένη ή μέση τιμή του $\mathbf{x}$ συμβολίζεται με μ και ορίζεται από την σχέση

$$\mu = E(\mathbf{x}) = [E(X_i)].$$

Για την αναμενομένη τιμή ενός τ. δ. $\mathbf{x}$ ισχύει η εξής ιδιότητα :

$$E(A\mathbf{x} + \mathbf{b}) = AE(\mathbf{x}) + \mathbf{b},$$

όπου A και $\mathbf{b}$ είναι κατάλληλοι σταθεροί πίνακας και διάνυσμα, αντίστοιχα.

Ορισμός 1.2 Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων ενός τ. δ. $\mathbf{x}$ συμβολίζεται με Σ και ορίζεται από την σχέση :

$$\Sigma = \text{Var}(\mathbf{x}) = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'].$$

Για την $\text{Var}(\mathbf{x})$ ισχύουν οι παρακάτω ιδιότητες :

- i) $\text{Var}(\mathbf{x}) = E(\mathbf{x}\mathbf{x}') - (E\mathbf{x})(E\mathbf{x})' = E(\mathbf{x}\mathbf{x}') - \mu\mu'$.
- ii) $\text{Var}(\mathbf{a}'\mathbf{x}) = \mathbf{a}'\text{Var}(\mathbf{x})\mathbf{a}$, για κάθε σταθερό διάνυσμα $\mathbf{a}$.
- iii) $\text{Var}(A\mathbf{x} + \mathbf{b}) = A \text{Var}(\mathbf{x}) A'$, για κάθε σταθερό πίνακα A και για κάθε σταθερό διάνυσμα $\mathbf{b}$.



Ορισμός 1.3 Η συνδιακύμανση δύο τ. δ. $\mathbf{x}_1$ και $\mathbf{x}_2$ με μέσες τιμές μ_1 και μ_2 , αντίστοιχα, ορίζεται από την σχέση :

$$\text{Cov}(\mathbf{x}_1, \mathbf{x}_2) = E[(\mathbf{x}_1 - \mu_1)(\mathbf{x}_2 - \mu_2)'] .$$

Για την συνδιακύμανση μπορούν εύκολα να δειχθούν οι παρακάτω ιδιότητες :

- i) $\text{Cov}(\mathbf{x}_1, \mathbf{x}_2) = E(\mathbf{x}_1 \mathbf{x}_2') - \mu_1 \mu_2'$.
- ii) $\text{Cov}(\mathbf{x}, \mathbf{x}) = \text{Var}(\mathbf{x})$.
- iii) $\text{Cov}(\mathbf{x}_1, \mathbf{x}_2) = [\text{Cov}(\mathbf{x}_2, \mathbf{x}_1)]'$.
- iv) $\text{Cov}(\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) = \text{Cov}(\mathbf{x}_1, \mathbf{y}) + \text{Cov}(\mathbf{x}_2, \mathbf{y})$.
- v) $\text{Cov}(\mathbf{A}\mathbf{x}_1, \mathbf{B}\mathbf{x}_2) = \mathbf{A}\text{Cov}(\mathbf{x}_1, \mathbf{x}_2)\mathbf{B}'$.
- vi) $\text{Cov}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{0}$, αν τα τ.δ. $\mathbf{x}_1$ και $\mathbf{x}_2$ είναι ανεξάρτητα.

Σε αντιστοιχία με τον δειγματικό πίνακα συσχέτισης $\mathbf{R}$ ορίζεται ο πίνακας συσχέτισης για τον πληθυσμό από την σχέση

$$\mathbf{P} = (\rho_{jj'}) , \quad \text{όπου} \quad \rho_{jj'} = \sigma_{jj'} / (\sigma_j \sigma_{j'}) .$$

Τα $\sigma_{jj'}$, σ_j & $\sigma_{j'}$, είναι τα ανάλογα των δειγματικών ποσοτήτων $s_{jj'}$, s_j & $s_{j'}$.

Αν $\mathbf{\Lambda} = \text{diag}(\sigma_i)$ τότε εύκολα μπορούμε να δούμε ότι

$$\mathbf{P} = \mathbf{\Lambda}^{-1} \mathbf{\Sigma} \mathbf{\Lambda}^{-1} \quad \text{και} \quad \mathbf{\Sigma} = \mathbf{\Lambda} \mathbf{P} \mathbf{\Lambda} .$$

Ακόμη, επειδή $\text{Var}(\mathbf{a}'\mathbf{x}) \geq 0$ για κάθε διάνυσμα $\mathbf{a}$, από την ιδιότητα (ii) της $\text{Var}(\mathbf{X}) = \mathbf{\Sigma}$, μπορούμε να συμπεράνουμε ότι ο πίνακας διακυμάνσεων - συνδιακυμάνσεων $\mathbf{\Sigma}$ είναι θετικώς ημιορισμένος. Επίσης από την σχέση $\mathbf{P} = \mathbf{\Lambda}^{-1} \mathbf{\Sigma} \mathbf{\Lambda}^{-1}$ και από γνωστή ιδιότητα της θεωρίας πινάκων έχουμε ότι και ο πίνακας $\mathbf{P}$ είναι θετικώς ημιορισμένος.



Ορισμός 1.4 Αν $\mathbf{x} = (X_1, X_2, \dots, X_p)'$ είναι ένα τ. δ. τότε η ροπογεννήτρια του $\mathbf{x}$ ή αλλιώς η από κοινού ροπογεννήτρια των $X_1, X_2, \dots, X_p$ ορίζεται από την σχέση

$$m_{\mathbf{x}}(\mathbf{t}) = E e^{\mathbf{x}'\mathbf{t}}, \quad \mathbf{t} \in \mathbb{R}^p. \quad (1.1)$$

Όπως και στην περίπτωση των μονοδιάστατων τ. μ., έτσι και η ροπογεννήτρια ενός τ. δ. μπορεί να χρησιμοποιηθεί για την εύρεση της $E(\mathbf{x})$, $\text{Var}(\mathbf{x})$, για την απόδειξη ανεξαρτησίας και την εύρεση της κατανομής τυχαίων διανυσμάτων.

Στο επόμενο θεώρημα αποδυνκνούνται μερικές από τις ιδιότητες της ροπογεννήτριας τυχαίου διανύσματος.

Θεώρημα 1.1 Έστω $\mathbf{x}$ ένα $p \times 1$ τυχαίο διάνυσμα, A ένας κατάλληλος σταθερός πίνακας και $\mathbf{b}$ ένα κατάλληλο σταθερό διάνυσμα.

i) Αν $\mathbf{y} = A\mathbf{x} + \mathbf{b}$, τότε $m_{\mathbf{y}}(\mathbf{t}) = e^{\mathbf{b}'\mathbf{t}} m_{\mathbf{x}}(A'\mathbf{t})$, $\mathbf{t} \in \mathbb{R}^p$.

ii) Αν $\mathbf{y} = c\mathbf{x}$, τότε $m_{\mathbf{y}}(\mathbf{t}) = m_{\mathbf{x}}(c\mathbf{t})$, $c \in \mathbb{R}$.

iii) $E(\mathbf{x}) = \left[\frac{\partial m_{\mathbf{x}}(\mathbf{t})}{\partial t_j} \Big|_{\mathbf{t}=\mathbf{0}} \right]'$ και

$$\text{Var}(\mathbf{x}) = \left[\frac{\partial^2 m_{\mathbf{x}}(\mathbf{t})}{\partial t_j \partial t_{j'}} \Big|_{\mathbf{t}=\mathbf{0}} - \frac{\partial m_{\mathbf{x}}(\mathbf{t})}{\partial t_j} \times \frac{\partial m_{\mathbf{x}}(\mathbf{t})}{\partial t_{j'}} \Big|_{\mathbf{t}=\mathbf{0}} \right],$$

όπου $j, j' = 1, 2, \dots, p$ $\mathbf{t} = (t_1, t_2, \dots, t_p)'$.

Απόδειξη Τα (i) και (ii) προκύπτουν εύκολα από τον ορισμό της ροπογεννήτριας συνάρτησης και τις ιδιότητες της αναμενόμενης τιμής.

iii) Αν υποθέσουμε ότι η $m_{\mathbf{x}}(\mathbf{t})$ είναι καλά ορισμένη - πεπερασμένη - σε μια περιοχή του μηδενός τότε μπορούμε να γράψουμε



$$\frac{\partial m_{\mathbf{x}}(t)}{\partial t_j} = \frac{\partial}{\partial t_j} E e^{\mathbf{x}'t} = E \frac{\partial}{\partial t_j} e^{\mathbf{x}'t} = E e^{\mathbf{x}'t} X_j.$$

$$\text{'Αρα } \left. \frac{\partial m_{\mathbf{x}}(t)}{\partial t_j} \right|_{t=0} = E X_j \Rightarrow E(\mathbf{x}) = (EX_j) = \left[\left. \frac{\partial m_{\mathbf{x}}(t)}{\partial t_j} \right|_{t=0} \right].$$

Για την $\text{Var}(\mathbf{x})$ θα έχουμε

$$\frac{\partial^2 m_{\mathbf{x}}(t)}{\partial t_j \partial t_{j'}} = E e^{\mathbf{x}'t} X_j X_{j'}, \quad \text{και} \quad \left. \frac{\partial^2 m_{\mathbf{x}}(t)}{\partial t_j \partial t_{j'}} \right|_{t=0} = E(X_j X_{j'}).$$

Επειδή τώρα $\text{Var}(\mathbf{x}) = [E(X_j X_{j'}) - (EX_j)(EX_{j'})]$, συμπεραίνουμε ότι η $\text{Var}(\mathbf{x})$ δίνεται από την σχέση στο (iii).

Το επόμενο θεώρημα δηλώνει ότι η ροπογεννήτρια ενός τ. δ. προσδιορίζει την κατανομή του τ. δ., υπό την προϋπόθεση ότι είναι καλά ορισμένη σε μια περιοχή του μηδενός.

Θεώρημα 1.2 Αν $\mathbf{x}$ και $\mathbf{y}$ είναι δύο $p \times 1$ τ. δ. και $m_{\mathbf{x}}(t) = m_{\mathbf{y}}(t)$, για κάθε t σε μια περιοχή του μηδενός, τότε τα τ. δ. $\mathbf{x}$ και $\mathbf{y}$ έχουν την ίδια κατανομή.

Απόδειξη Η απόδειξη στηρίζεται σε ιδιότητες μετασχηματισμών Laplace και παραλείπεται.

Αν $\mathbf{x}$ είναι ένα $p \times 1$ τ. δ. και γράψουμε $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2)'$, όπου $\mathbf{x}_k$ ($k = 1, 2$) είναι, αντίστοιχα, $p_k \times 1$ τ. δ. ($p_1 + p_2 = p$), τότε μπορούμε να δείξουμε το επόμενο θεώρημα.



Θεώρημα 1.3 Αν $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2)'$, τότε

$$i) m_{\mathbf{x}_1}(t_1) = m_{\mathbf{x}} \begin{pmatrix} t_1 \\ 0 \end{pmatrix}, \quad t_1 \in \mathbb{R}^{p_1} \text{ και } m_{\mathbf{x}_2}(t_2) = m_{\mathbf{x}} \begin{pmatrix} 0 \\ t_2 \end{pmatrix}, \quad t_2 \in \mathbb{R}^{p_2}.$$

ii) Τα τ. δ. $\mathbf{x}_1$ και $\mathbf{x}_2$ είναι ανεξάρτητα εάν και μόνον εάν

$$m_{\mathbf{x}} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = m_{\mathbf{x}}(t_1) m_{\mathbf{x}}(t_2)$$

για κάθε $t_1 \in \mathbb{R}^{p_1}$ και για κάθε $t_2 \in \mathbb{R}^{p_2}$.

$$\text{Απόδειξη } i) m_{\mathbf{x}} \begin{pmatrix} t_1 \\ 0 \end{pmatrix} = E e^{\mathbf{x}'_1 t_1 + \mathbf{x}'_2 0} = E e^{\mathbf{x}'_1 t_1} = m_{\mathbf{x}_1}(t_1).$$

Ανάλογη είναι η απόδειξη και για το $m_{\mathbf{x}_2}(t_2)$.

ii) Ας υποθέσουμε ότι τα τ. δ. $\mathbf{x}_1$ και $\mathbf{x}_2$ είναι ανεξάρτητα.

Τότε

$$\begin{aligned} m_{\mathbf{x}}(t) &= E e^{\mathbf{x}'_1 t_1 + \mathbf{x}'_2 t_2} = \left(E e^{\mathbf{x}'_1 t_1 + \mathbf{x}'_2 0} \right) \left(E e^{\mathbf{x}'_1 0 + \mathbf{x}'_2 t_2} \right) \\ &= m_{\mathbf{x}} \begin{pmatrix} t_1 \\ 0 \end{pmatrix} m_{\mathbf{x}} \begin{pmatrix} 0 \\ t_2 \end{pmatrix} = m_{\mathbf{x}_1}(t_1) m_{\mathbf{x}_2}(t_2). \end{aligned}$$

Αντίστροφα, ας υποθέσουμε ότι για τις ροπογεννήτριες των τ. δ. $\mathbf{x}_1$ και $\mathbf{x}_2$ ισχύει η σχέση του θεωρήματος στο (ii). Έστω ότι $\mathbf{y}_1$ και $\mathbf{y}_2$ είναι δύο ανεξάρτητα τ. δ. των οποίων οι κατανομές είναι οι ίδιες με αυτές των $\mathbf{x}_1$ και $\mathbf{x}_2$ αντίστοιχα.

Τότε από το προηγούμενο θεώρημα θα έχουμε :

$$m_{\mathbf{y}_1}(t_1) = m_{\mathbf{x}} \begin{pmatrix} t_1 \\ 0 \end{pmatrix} \text{ και } m_{\mathbf{y}_2}(t_2) = m_{\mathbf{x}} \begin{pmatrix} 0 \\ t_2 \end{pmatrix}.$$

Ακόμη, επειδή τα τ. δ. $\mathbf{y}_1$ και $\mathbf{y}_2$ τα πήραμε ανεξάρτητα θα ισχύει

$$m_{\mathbf{y}}(t) = m_{\mathbf{y}_1} \begin{pmatrix} t_1 \\ 0 \end{pmatrix} m_{\mathbf{y}_2} \begin{pmatrix} 0 \\ t_2 \end{pmatrix} = m_{\mathbf{x}} \begin{pmatrix} t_1 \\ 0 \end{pmatrix} m_{\mathbf{x}} \begin{pmatrix} 0 \\ t_2 \end{pmatrix} = m_{\mathbf{x}}(t),$$



όπου $y = (y'_1, y'_2)'$. Δηλαδή τα τ. δ. x και y έχουν την ίδια κατανομή, άρα και τα τ. δ. x_1 και x_2 είναι ανεξάρτητα. (Με άλλα λόγια ο παραπάνω συλλογισμός δείχνει ότι τα τ. δ. x_1 και x_2 είναι ανεξάρτητα επειδή έχουν την ίδια ροπογεννήτρια σαν να ήταν ανεξάρτητα).

Θεώρημα 1.4 Αν x_1 και x_2 είναι δύο ρχλ ανεξάρτητα τ. δ. , τότε για την ροπογεννήτρια του τ. δ. $y = x_1 + x_2$ ισχύει :

$$m_y(t) = m_{x_1}(t) m_{x_2}(t)$$

$$\begin{aligned} \text{Απόδειξη } m_y(t) &= E e^{y't} = E e^{x'_1 t + x'_2 t} = \left(E e^{x'_1 t} \right) \left(E e^{x'_2 t} \right) \\ &= m_{x_1}(t) m_{x_2}(t) . \end{aligned}$$



Α Σ Κ Η Σ Ε Ι Σ

1.1 Αν $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ είναι ένα τυχαίο δείγμα από την κατανομή $f(\mathbf{x}, \theta)$ με μέση τιμή μ και πίνακα διακυμάνσεων-συνδιακυμάνσεων Σ να δειχθεί ότι $E\bar{\mathbf{x}} = \mu$ και $\text{Var}(\bar{\mathbf{x}}) = n^{-1}\Sigma$.

1.2 Να δειχθεί ότι $E(S) = \frac{n-1}{n} \Sigma$.

1.3 Έστω $\mathbf{x} = (X_1, X_2, \dots, X_p)'$ ένα $p \times 1$ τυχαίο διάνυσμα και ας υποθέσουμε ότι όλες οι τ.μ. X_i , $i = 1, 2, \dots, p$, έχουν διακύμανση σ^2 , και όλα τα ζεύγη X_i, X_j , $i \neq j$, τον ίδιο συντελεστή συσχέτισης ρ . Να γραφεί ο πίνακας διακυμάνσεων-συνδιακυμάνσεων του $\mathbf{x}$ σαν συνάρτηση των ρ, σ^2 και να δειχθεί ότι $\rho \geq -\frac{1}{p-1}$, ($p \geq 2$).

1.4 Το τυχαίο διάνυσμα $\mathbf{x} = (X_1, X_2)'$ έχει $E(\mathbf{x}) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ και $\text{Var}(\mathbf{x}) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Να δοθούν $EX_1, EX_2, \text{Var}X_1, \text{Var}X_2, \text{Var}(\mathbf{x}), E\|\mathbf{x}\|^2$, και ο συντελεστής συσχέτισης των X_1 και X_2 .

1.5 Εάν $\mathbf{x} = (X_1, X_2, \dots, X_p)'$ είναι ένα τυχαίο διάνυσμα με $\text{Var}(\mathbf{x}) = \Sigma$ να δειχθεί ότι $\text{Var}(X_1 + X_2 + \dots + X_p) = \mathbf{1}' \Sigma \mathbf{1}$, όπου $\mathbf{1}$ είναι ένα διάνυσμα με όλα του τα στοιχεία μονάδες.

1.6 Αν $S(\mathbf{a}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{a})(\mathbf{x}_i - \mathbf{a})'$ είναι η δειγματική διακύμανση στο σημείο $\mathbf{x} = \mathbf{a}$ να δειχθεί ότι

$$S(\mathbf{a}) = S + (\bar{\mathbf{x}} - \mathbf{a})(\bar{\mathbf{x}} - \mathbf{a})'.$$

Στην συνέχεια χρησιμοποιώντας την σχέση

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} = \det(\mathbf{A}) \cdot \det(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})$$

να δειχθεί ότι



$$\det(S(a)) = \det(S) \cdot \left\{ 1 + (x - a)' S^{-1} (x - a) \right\}$$

και άρα

$$\min \det(S(a)) = \det(S) .$$

1.7 Αν $M = X'X$, όπου X είναι ο $n \times p$ πίνακας των δεδομένων, να δειχθεί ότι για τα στοιχεία του M ισχύει

$$m_{jj} = x_j' x_j = n(s_{jj} + \bar{x}_j^2) , \quad m_{jj'} = x_j' x_{j'} = n(s_{jj'} + \bar{x}_{.j} \bar{x}_{.j'})$$

1.8 Αν $f(x_1, x_2) = x_1 + x_2$, $0 \leq x_1, x_2 \leq 1$, να βρεθεί η μέση τιμή μ και ο πίνακας διακυμάνσεων-συνδιακυμάνσεων Σ .



2. ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ

Στο κεφάλαιο αυτό θα παρουσιάσουμε ορισμένες πολυδιάσφατες κατανόμες με μερικές από τις ιδιότητές τους. Όλες οι κατανόμες που θα παρουσιασθούν είναι γενικεύσεις αντίστοιχων μονοδιάσφατων κατανομών.

2.1 ΠΟΛΥΔΙΑΣΤΑΤΗ ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Σε όλους είναι γνωστή η έννοια της κανονικής κατανομής στην μονοδιάσφατη περίπτωση. Λέμε ότι η τυχαία μεταβλητή X έχει την κανονική κατανομή με μέση τιμή μ και διακύμανση σ^2 , και γράφουμε, $X \sim N(\mu, \sigma^2)$, όταν η συνάρτηση πυκνότητας πιθανότητας της τ.μ. X δίνεται από την σχέση

$$f_X(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} (x - \mu)^2}, \quad -\infty < x < +\infty.$$

Επίσης είναι γνωστό ότι αν $X \sim N(\mu, \sigma^2)$ τότε η κατανομή της τυχαίας μεταβλητής $Y = (X - \mu)/\sigma$ είναι η τυπική κανονική, και συμβολικά $Y \sim N(0, 1)$.



Σε αναλογία με τα παραπάνω η πολυδιάστατη κανονική κατανομή ορίζεται ως εξής .

Ορισμός 2.1 Έστω $\mathbf{x}$ ένα $p \times 1$ τ. δ. . Θα λέμε ότι η κατανομή του τ. δ. $\mathbf{x}$ είναι η p - διάστατη κανονική κατανομή όταν η κατανομή του $\mathbf{x}$ δίνεται από την σχέση

$$f_{\mathbf{x}}(\mathbf{x}) = \left[\det(2\pi \Sigma) \right]^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)' \Sigma^{-1}(\mathbf{x}-\mu)} , \quad (2.1)$$

όπου $\Sigma > 0$ (δηλ. Σ θετικώς ορισμένος) .

Θεώρημα 2.1 Ας υποθέσουμε ότι η κατανομή του τ. δ. $\mathbf{x}$ είναι η p - διάστατη κανονική. Έστω ακόμη ότι $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \mu)$. Τότε οι συνιστώσες $Y_1, Y_2, \dots, Y_p$ του τ. δ. $\mathbf{y}$ είναι ανεξάρτητες τ. μ. με $Y_j \sim N(0,1)$, $(j = 1, 2, \dots, p)$.

Απόδειξη Από την $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \mu)$ παίρνουμε $\mathbf{x} = \Sigma^{1/2}\mathbf{y} + \mu$. Αν με $J = (\partial x_j / \partial y_j)$ συμβολίσουμε την Ιακωβιανή του προηγούμενου μετασχηματισμού τότε θα έχουμε

$$\begin{aligned} f_{\mathbf{y}}(\mathbf{y}) &= f_{\mathbf{x}}(\mathbf{y}) \det(J) = \left[\det(2\pi \Sigma) \right]^{-1/2} e^{-\frac{1}{2}\mathbf{y}'\mathbf{y}} \left[\det(\Sigma) \right]^{1/2} \\ &= \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\mathbf{y}'\mathbf{y}} . \end{aligned}$$

Πόρισμα 2.1.1 Αν το $p \times 1$ τ. δ. $\mathbf{x}$ έχει την p - διάστατη κανονική κατανομή , τότε $E(\mathbf{x}) = \mu$ και $\text{Var}(\mathbf{x}) = \Sigma$.

Απόδειξη . Από το προηγούμενο θεώρημα θα έχουμε ότι $E(\mathbf{y}) = \mathbf{0}$ και $\text{Var}(\mathbf{y}) = \mathbf{I}$. Συνεπώς , από την σχέση $\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \mu)$ έχουμε το ζητούμενο.



Από το Πρόγραμμα 2.1.1 γίνεται φανερό ότι αν το τ. δ. $\mathbf{x}$ έχει την p -διάστατη κανονική κατανομή τότε μπορούμε να γράψουμε $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. (Το p δηλώνει την διάσταση της πολυδιάστατης κανονικής κατανομής).

Θεώρημα 2.2 Αν $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ τότε η ροπογεννήτρια συνάρτηση του τ. δ. $\mathbf{x}$ δίνεται από την σχέση

$$m_{\mathbf{x}}(\mathbf{t}) = e^{\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}, \mathbf{t} \in \mathbb{R}^p.$$

Απόδειξη. Αν θέσουμε $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ θα έχουμε

$$m_{\mathbf{x}}(\mathbf{t}) = E\left(e^{\mathbf{x}'\mathbf{t}}\right) = e^{\boldsymbol{\mu}'\mathbf{t}} E\left(e^{\mathbf{y}'\boldsymbol{\Sigma}^{1/2}\mathbf{t}}\right) = e^{\boldsymbol{\mu}'\mathbf{t}} E\left(e^{\mathbf{y}'\mathbf{u}}\right), \text{ όπου } \mathbf{u} = \boldsymbol{\Sigma}^{1/2}\mathbf{t}.$$

Από το Θεώρημα 2.1 έχουμε ότι οι τ. μ. $Y_1, Y_2, \dots, Y_p$ είναι ανεξάρτητες τ. μ.. Έτσι παίρνουμε ότι

$$E\left(e^{\mathbf{y}'\mathbf{u}}\right) = \prod_{j=1}^p m_{Y_j}(u_j) = \prod_{j=1}^p e^{\frac{1}{2}u_j^2} = e^{\frac{1}{2}\mathbf{u}'\mathbf{u}}.$$

Συνεπώς

$$m_{\mathbf{x}}(\mathbf{t}) = e^{\boldsymbol{\mu}'\mathbf{t}} e^{(1/2)\mathbf{u}'\mathbf{u}} = e^{\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}.$$

Παρατήρηση 2.1. Πολλοί συγγραφείς χρησιμοποιούν το Θεώρημα 2.2 σαν ορισμό της πολυδιάστατης κανονικής κατανομής. Δηλαδή λέμε ότι το $p \times 1$ τ. δ. $\mathbf{x}$ έχει την πολυδιάστατη κανονική κατανομή όταν η ροπογεννήτρια του τ. δ. $\mathbf{x}$ δίνεται από τη σχέση

$$m_{\mathbf{x}}(\mathbf{t}) = e^{\boldsymbol{\mu}'\mathbf{t} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}, \mathbf{t} \in \mathbb{R}^p.$$

Ο τρόπος αυτός ορισμού της πολυδιάστατης κανονικής κατανομής έχει το πλεονέκτημα ότι ο πίνακας $\boldsymbol{\Sigma}$ μπορεί να είναι θετικά



ημιορισμένος, ενώ στον ορισμό που δώσαμε (βλ. Ορισμό 2.1) ο Σ θα πρέπει να είναι θετικά ορισμένος.

Εφαρμογή του προηγούμενου θεωρήματος είναι το Πρόβλημα που ακολουθεί.

Πρόβλημα 2.2.1. Αν $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, τότε για κάθε σταθερό διάνυσμα $\mathbf{a} \neq \mathbf{0}$ η κατανομή του $\mathbf{a}'\mathbf{x}$ είναι η μονοδιάστατη κανονική.

Θεώρημα 2.3. Αν $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ και $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, όπου $\mathbf{A}$ είναι ένας $q \times p$ πίνακας με $\text{Rank}(\mathbf{A}) = q$ και $\mathbf{b}$ είναι ένα σταθερό διάνυσμα, τότε

$$\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}')$$

Απόδειξη. Από τα Θεωρήματα 1.1 και 2.2 θα έχουμε

$$\begin{aligned} m_{\mathbf{y}}(\mathbf{t}) &= e^{\mathbf{b}'\mathbf{t}} m_{\mathbf{x}}(\mathbf{A}'\mathbf{t}) = e^{\mathbf{b}'\mathbf{t}} e^{\boldsymbol{\mu}'(\mathbf{A}'\mathbf{t}) + \frac{1}{2}(\mathbf{A}'\mathbf{t})'\Sigma(\mathbf{A}'\mathbf{t})} \\ &= e^{(\mathbf{A}\boldsymbol{\mu} + \mathbf{b})'\mathbf{t} + \frac{1}{2}\mathbf{t}'(\mathbf{A}\Sigma\mathbf{A}')\mathbf{t}} \end{aligned}$$

Την απόδειξη συμπληρώνει το Θεώρημα μονοσημάντου των ροπογεννητριών (βλ. Θεώρημα 1.2).

Παρατήρηση 2.2. Η συνθήκη ότι $\text{Rank}(\mathbf{A}) = q$ χρειάζεται μόνο για να γράψουμε την συνάρτηση πυκνότητας πιθανότητας του τ. δ. $\mathbf{y}$. Διαφορετικά το q θα μπορούσε να είναι ο αριθμός των γραμμών του πίνακα $\mathbf{A}$.

Με ανάλογο τρόπο μπορούμε να αποδείξουμε το επόμενο Θεώρημα

Θεώρημα 2.4. i) Αν $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$ τότε $c\mathbf{x} \sim N_p(c\boldsymbol{\mu}, c^2\Sigma)$, $c \in \mathbb{R}$.

ii) Έστω ότι τα τ. δ. $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$ είναι ανεξάρτητα και $\mathbf{x}_k \sim N_{q_k}(\boldsymbol{\mu}_k, \Sigma_k)$. Αν $\mathbf{A}_k$ είναι $q_k \times p_k$ σταθεροί πίνακες τότε



$$y = \sum_{k=1}^K A_k x_k \sim Nq(\mu_y, \Sigma_y)$$

$$\text{όπου } \mu_y = \sum_{k=1}^k A_k \mu_k \quad \text{και} \quad \Sigma_y = \sum_{k=1}^k A_k \Sigma_k A_k'.$$

2.1.1. Υποδιανύσματα , ανεξαρτησία , δεσμευμένες πιθανότητες

Έστω x ένα $p \times 1$ τ. δ. με $x \sim Np(\mu, \Sigma)$. Στην παράγραφο αυτή θα εξετάσουμε τα εξής ερωτήματα :

i) ποια είναι η κατανομή μιας ή μερικών συνιστωσών του τ. δ. x ;

ii) κάτω από ποιες συνθήκες οι συνιστώσες του τ. δ. x (ή μερικές από αυτές) είναι ανεξάρτητες ;

iii) αν δοθεί η τιμή μιας ή μερικών συνιστωσών του τ. δ. x πως αυτή η πληροφορία επηρεάζει την κατανομή των υπολοίπων συνιστωσών ;

Για να απαντήσουμε στα ερωτήματα αυτά θα χρησιμοποιήσουμε τις παρακάτω διαμερίσεις των x , μ , και Σ . Έτσι αν το τ. δ. x το γράψουμε σαν

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \text{ τότε θα έχουμε } \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ και } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

όπου $\mu_k = E(x_k)$, $\Sigma_{kk} = \text{Var}(x_k)$ και $\Sigma_{kk'} = \text{Cov}(x_k, x_{k'})$ ($k \neq k' = 1, 2$)

Θεώρημα 2.5. Αν $x = (x_1', x_2')' \sim Np(\mu, \Sigma)$ τότε

i) $x_1 \sim Np_1(\mu_1, \Sigma_{11})$ & $x_2 \sim Np_2(\mu_2, \Sigma_{22})$, όπου p_1, p_2 ($p_1 + p_2 = p$) είναι οι διαστάσεις των τ. δ. x_1 και x_2 , αντίστοιχα .

ii) Τα τ. δ. x_1 και x_2 είναι ανεξάρτητα εάν και μόνον εάν $\Sigma_{12} = 0$.



Απόδειξη . i) Αν γράψουμε $\mathbf{x}_1 = \mathbf{A}\mathbf{x}$, όπου $\mathbf{A} = [\mathbf{I}_{p_1}, \mathbf{0}]$, τότε από το Θεώρημα 2.3 θα έχουμε ότι $\mathbf{x}_1 \sim N_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$. Ανάλογη είναι η απόδειξη και για το τ. δ. $\mathbf{x}_2$.

ii) Η απόδειξη της πρότασης αυτής του θεωρήματος βασίζεται στο Θεώρημα 2.9 που ακολουθεί .

Θεώρημα 2.6 . Αν $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2)'$ $\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και $\mathbf{x}_{2.1} = \mathbf{x}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1$ τότε

i) η κατανομή του τ. δ. $\mathbf{x}_{2.1}$ είναι η $N_{p_2}(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{2.1})$, όπου

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1 \quad \text{και} \quad \boldsymbol{\Sigma}_{2.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} .$$

ii) Τα τ. δ. $\mathbf{x}_1$ και $\mathbf{x}_{2.1}$ είναι ανεξάρτητα .

Απόδειξη. i) Θέτοντας $\mathbf{x}_{2.1} = \mathbf{B}\mathbf{x}$, όπου $\mathbf{B} = [-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}, \mathbf{I}]$, τότε από το Θεώρημα 2.3 παίρνουμε ότι

$$\mathbf{x}_{2.1} \sim N_{p_2}(\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) .$$

ii) Αν θέσουμε $\mathbf{x}_1 = \mathbf{A}\mathbf{x}$, όπου $\mathbf{A} = [\mathbf{I}_{p_1}, \mathbf{0}]$, (βλέπε και απόδειξη θεωρήματος 2.5), τότε έχουμε ότι $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$. Η απόδειξη συμπληρώνεται από το Θεώρημα 2.9(i) .

Θεώρημα 2.7. Αν $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2)'$ $\sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, τότε η υπό συνθήκη κατανομή του τ. δ. $\mathbf{x}_2$ δοθέντος του τ. δ. $\mathbf{x}_1$ είναι η p_2 -κανονική κατανομή με μέση τιμή $\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$ και διακύμανση $\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, ή συμβολικά

$$\mathbf{x}_2 | \mathbf{x}_1 \sim N_{p_2} \left(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) , \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \right) .$$

Απόδειξη. Από το προηγούμενο Θεώρημα (Θεώρημα 2.6) έχουμε ότι τα τ. δ. $\mathbf{x}_1$ και $\mathbf{x}_{2.1}$ είναι ανεξάρτητα . Τώρα μπορούμε να γράψουμε

$$\mathbf{x}_2 = \mathbf{x}_{2.1} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 ,$$



όπου ο δεύτερος όρος στα δεξιά της ισότητας είναι σταθερός. Αυτό σημαίνει ότι η κατανομή του τ. δ. $\mathbf{x}_2$ είναι η $N_{p_2}(\cdot, \cdot)$. Η μέση τιμή και ο πίνακας διακυμάνσεων - συνδιακυμάνσεων της κατανομής αυτής είναι

$$E(\mathbf{x}_2 | \mathbf{x}_1) = \mu_{2.1} + \Sigma_{21} \Sigma_{11}^{-1} \mathbf{x}_1 = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{x}_1 - \mu_1)$$

και

$$\text{Var}(\mathbf{x}_2 | \mathbf{x}_1) = \text{Var}(\mathbf{x}_{2.1}) = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}.$$

2.1.2 Τετραγωνικές μορφές

Στην παράγραφο αυτή θα δώσουμε μερικά χρήσιμα θεωρήματα για την κατανομή και την ανεξαρτησία τετραγωνικών μορφών.

Όπως είναι γνωστό, μια τετραγωνική μορφή είναι μια σχέση της μορφής

$$Q = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j = \mathbf{x}' \mathbf{A} \mathbf{x}.$$

Μπορεί ναδειχθεί ότι, σε μια οποιαδήποτε τετραγωνική μορφή, ο πίνακας $\mathbf{A}$ είναι πάντοτε συμμετρικός.

Λήμμα 2.1. Αν $\mathbf{x} \sim N_p(\mu, \mathbf{W})$, όπου $\mathbf{W}$ είναι ένας θετικά - ορισμένος πίνακας, τότε για κάθε σταθερό διάνυσμα $\mathbf{g}$ ισχύει :

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-\frac{1}{2} \mathbf{x}' \mathbf{W}^{-1} \mathbf{x} + \mathbf{g}' \mathbf{x}} d\mathbf{x} = (2\pi)^{p/2} [\det(\mathbf{W})]^{1/2} e^{\frac{1}{2} \mathbf{g}' \mathbf{W} \mathbf{g}}.$$

Απόδειξη. Από το ολοκλήρωμα της πολυδιάστατης κανονικής κατανομής θα έχουμε:



$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \mathbf{W}^{-1}(\mathbf{x}-\boldsymbol{\mu})} d\mathbf{x} = (2\pi)^{p/2} [\det(\mathbf{W})]^{-1/2}.$$

Το πρώτο μέλος στην προηγούμενη ισότητα μπορεί να γραφεί σαν

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-\frac{1}{2} \mathbf{x}' \mathbf{W}^{-1} \mathbf{x} + \boldsymbol{\mu}' \mathbf{W}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}' \mathbf{W}^{-1} \boldsymbol{\mu}} d\mathbf{x}$$

και θέτοντας $\mathbf{g}' = \boldsymbol{\mu}' \mathbf{W}^{-1}$ έχουμε το ζητούμενο.

Θεώρημα 2.8 . Αν $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ τότε

i) $E(\mathbf{x}' \mathbf{A} \mathbf{x}) = \text{tr}(\mathbf{A} \boldsymbol{\Sigma}) + \boldsymbol{\mu}' \mathbf{A} \boldsymbol{\mu}$. (Η σχέση αυτή ισχύει για κάθε τετραγωνική μορφή ανεξάρτητα αν το τ. δ. $\mathbf{x}$ ακολουθεί ή όχι την πολυδιάστατη κανονική κατανομή)

ii) Η ροπογεννήτρια της τετραγωνικής μορφής $\mathbf{x}' \mathbf{A} \mathbf{x}$ δίνεται από την σχέση

$$m_{\mathbf{x}' \mathbf{A} \mathbf{x}}(t) = \left[\det(\mathbf{I} - 2t \mathbf{A} \boldsymbol{\Sigma}) \right]^{-1/2} e^{-\frac{1}{2} \boldsymbol{\mu}' \left[\mathbf{I} - (\mathbf{I} - 2t \mathbf{A} \boldsymbol{\Sigma})^{-1} \right] \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}, t \in \mathbb{R}.$$

iii) Η συνδιακύμανση του τ. δ. $\mathbf{x}$ με την τετραγωνική μορφή $\mathbf{x}' \mathbf{A} \mathbf{x}$ είναι

$$\text{Cov}(\mathbf{x}, \mathbf{x}' \mathbf{A} \mathbf{x}) = 2 \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu}.$$

Απόδειξη . i) Επειδή $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ έπεται ότι $E(\mathbf{x}) = \boldsymbol{\mu}$ & $\text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$. Άρα $E(\mathbf{x} \mathbf{x}') = \boldsymbol{\Sigma} + \boldsymbol{\mu} \boldsymbol{\mu}'$. Συνεπώς

$$E(\mathbf{x}' \mathbf{A} \mathbf{x}) = E \left[\text{tr}(\mathbf{A} \mathbf{x} \mathbf{x}') \right] = \text{tr}(\mathbf{A} \boldsymbol{\Sigma} + \mathbf{A} \boldsymbol{\mu} \boldsymbol{\mu}') = \text{tr}(\mathbf{A} \boldsymbol{\Sigma}) + \boldsymbol{\mu}' \mathbf{A} \boldsymbol{\mu}.$$

$$\text{ii) } m_{\mathbf{x}' \mathbf{A} \mathbf{x}}(t) = \left(E e^{t \mathbf{x}' \mathbf{A} \mathbf{x}} \right) =$$



$$= (2\pi)^{-p/2} [\det(\Sigma)]^{-1/2} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{t\mathbf{x}'\mathbf{A}\mathbf{x} - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} d\mathbf{x}.$$

Το πολλαπλό ολοκλήρωμα στην τελευταία σχέση μπορεί να γραφεί σαν

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{t\mathbf{x}'\mathbf{A}\mathbf{x} - \frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}'\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}} dx_1 dx_2 \dots dx_p =$$

$$\int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\mathbf{x}'(\mathbf{I}-2t\mathbf{A}\Sigma)\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}'\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}} dx_1 dx_2 \dots dx_p.$$

Άρα θα είναι

$$m_{\mathbf{x}'\mathbf{A}\mathbf{x}}(t) = \frac{e^{-\frac{1}{2}\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}}}{(2\pi)^{p/2} [\det(\Sigma)]^{1/2}} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\mathbf{x}'(\mathbf{I}-2t\mathbf{A}\Sigma)\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}'\Sigma^{-1}\mathbf{x}} d\mathbf{x}$$

και κάνοντας χρήση του Λήμματος 2.1 με $\mathbf{g}' = \boldsymbol{\mu}'\Sigma^{-1}$ και $\mathbf{W} = [(\mathbf{I} - 2t\mathbf{A}\Sigma)\Sigma^{-1}]^{-1} = \Sigma(\mathbf{I} - 2t\mathbf{A}\Sigma)^{-1}$ έχουμε το ζητούμενο.

iii) Για την συνδιακύμανση μεταξύ $\mathbf{x}$ και $\mathbf{x}'\mathbf{A}\mathbf{x}$ θα έχουμε

$$\begin{aligned} \text{Cov}(\mathbf{x}, \mathbf{x}'\mathbf{A}\mathbf{x}) &= \left\{ E(\mathbf{x}-\boldsymbol{\mu}) \left[\mathbf{x}'\mathbf{A}\mathbf{x} - E(\mathbf{x}'\mathbf{A}\mathbf{x}) \right] \right\} = \\ &= E \left\{ (\mathbf{x}-\boldsymbol{\mu}) \left[\mathbf{x}'\mathbf{A}\mathbf{x} - \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} - \text{tr}(\mathbf{A}\Sigma) \right] \right\} = \\ &= E \left\{ (\mathbf{x}-\boldsymbol{\mu}) \left[(\mathbf{x}-\boldsymbol{\mu})'\mathbf{A}(\mathbf{x}-\boldsymbol{\mu}) + 2(\mathbf{x}-\boldsymbol{\mu})'\mathbf{A}\boldsymbol{\mu} - \text{tr}(\mathbf{A}\Sigma) \right] \right\} \end{aligned}$$



$$\begin{aligned}
&= E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})' \mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right] + \\
&\quad + 2E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})' \mathbf{A}\right] - E(\mathbf{x}-\boldsymbol{\mu})\text{tr}(\mathbf{A}\boldsymbol{\Sigma}) \\
&= 0 + 2 \text{Var}(\mathbf{x}) \mathbf{A}\boldsymbol{\mu} - 0 = 2 \boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} .
\end{aligned}$$

Πόρισμα 2.8.1. Αν $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ τότε η διακύμανση της τετραγωνικής μορφής $\mathbf{x}'\mathbf{A}\mathbf{x}$ δίνεται από την σχέση

$$\text{Var}(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2 \text{tr}(\mathbf{A}\boldsymbol{\Sigma})^2 + 4 \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu} .$$

Απόδειξη. Η απόδειξη στηρίζεται στα θεωρήματα 1.1(iii) και 2.8(ii).

Το επόμενο θεώρημα μας λέει πότε δύο γραμμικές ή/και τετραγωνικές μορφές είναι ανεξάρτητες.

Θεώρημα 2.9. Αν $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και $\mathbf{A}, \mathbf{B}$ είναι δύο σταθεροί πίνακες με $\text{Rank}(\mathbf{A}) = q_1$ και $\text{Rank}(\mathbf{B}) = q_2$, τότε

i) οι γραμμικές μορφές $\mathbf{A}\mathbf{x}$ και $\mathbf{B}\mathbf{x}$ είναι ανεξάρτητες εάν και μόνον εάν $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = 0$.

ii) η γραμμική μορφή $\mathbf{B}\mathbf{x}$ είναι ανεξάρτητη από την τετραγωνική μορφή $\mathbf{x}'\mathbf{A}\mathbf{x}$ εάν και μόνον εάν $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = 0$.

iii) οι τετραγωνικές μορφές $\mathbf{x}'\mathbf{A}\mathbf{x}$ και $\mathbf{x}'\mathbf{B}\mathbf{x}$ είναι ανεξάρτητες εάν και μόνο εάν $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B} = 0$ (ή ισοδύναμα $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = 0$).

Απόδειξη. i) Από το Θεώρημα 2.3 και την Παρατήρηση 2.2 θα έχουμε ότι $\mathbf{A}\mathbf{x} \sim N_{q_1}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ και $\mathbf{B}\mathbf{x} \sim N_{q_2}(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$.

Ακόμη, από την ιδιότητα (v) της συνδιακύμανσης ενός τ. δ. θα είναι $\text{Cov}(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}'$.

ii) Θα αποδείξουμε πρώτα ότι εάν $\mathbf{B}\boldsymbol{\Sigma}\mathbf{A} = 0$ τότε $\mathbf{x}'\mathbf{A}\mathbf{x}$ και $\mathbf{B}\mathbf{x}$ είναι ανεξάρτητα. Επειδή ο πίνακας $\mathbf{A}$ είναι πίνακας



μιας τετραγωνικής μορφής, έχουμε ότι ο A θα είναι συμμετρικός. Κατά συνέπεια θα υπάρχει ένας πίνακας L , με $\text{Rank}(L) = q_1$, τέτοιος ώστε $A = LL'$. Ακόμη, επειδή $\text{Rank}(L) = q_1$, ο $(LL')^{-1}$ υπάρχει. Συνεπώς

$$B\bar{\Sigma}A = 0 \Rightarrow B\bar{\Sigma}LL' = 0 \Rightarrow B\bar{\Sigma}LL'L(L'L)^{-1} = 0 \Rightarrow B\bar{\Sigma}L = 0.$$

Δηλαδή $\text{Cov}(Bx, L'x) = 0$. Αυτό σημαίνει ότι Bx και $L'x$ είναι ανεξάρτητα. Κατά συνέπεια και Bx είναι ανεξάρτητο από το $x'LL'x = x'Ax$.

Αντίστροφα, εάν $x'Ax$ είναι ανεξάρτητο από το Bx , τότε $\text{Cov}(Bx, x'Ax) = 0$. Αλλά από το θεώρημα 2.8(ii) έχουμε ότι $\text{Cov}(Bx, x'Ax) = 2B\bar{\Sigma}A\mu$. Άρα $2B\bar{\Sigma}A\mu = 0$, για κάθε μ , και συνεπώς $B\bar{\Sigma}A = 0$.

iii) Η συνθήκη $A\bar{\Sigma}B = 0$ είναι ισοδύναμη με την $B\bar{\Sigma}A = 0$, επειδή και οι τρεις πίνακες $A, \bar{\Sigma}, B$ είναι συμμετρικοί. Έστω τώρα ότι $A\bar{\Sigma}B = 0$. Τότε, όπως και στο (ii), θα υπάρχουν πίνακες L και M , με $\text{Rank}(L) = q_1$, και $\text{Rank}(M) = q_2$, τέτοιοι ώστε $A = LL'$ και $B = MM'$. Συνεπώς, από την σχέση $A\bar{\Sigma}B = 0$ παίρνουμε, επειδή οι πίνακες $(LL')^{-1}$ και $(MM')^{-1}$ υπάρχουν, ότι

$$\begin{aligned} A\bar{\Sigma}B = 0 &\Rightarrow LL'\bar{\Sigma}MM' = 0 \Rightarrow (L'L)^{-1}L'LL'\bar{\Sigma}MM'M(M'M)^{-1} = 0 \\ &\Rightarrow L'\bar{\Sigma}M = 0 \Rightarrow \text{Cov}(L'x, M'x) = 0 \Rightarrow L'x \text{ και } M'x \text{ ανεξάρτητα.} \\ &\text{Άρα και } x'Ax = x'LL'x \text{ ανεξάρτητο από το } x'Bx = x'MM'x. \end{aligned}$$

Αντίστροφα, έστω ότι $x'Ax$ και $x'Bx$ είναι ανεξάρτητα. Τότε $\text{Cov}(x'Ax, x'Bx) = 0$ και κατά συνέπεια

$$\begin{aligned} \text{Var}(x'Ax + x'Bx) &= \text{Var}(x'Ax) + \text{Var}(x'Bx) \\ \text{ή} \quad \text{Var}[x'(A+B)x] &= \text{Var}(x'Ax) + \text{Var}(x'Bx). \end{aligned}$$

Εφαρμόζοντας το Πρόγραμμα 2.8.1 και στους τρεις όρους της τελευταίας σχέσης παίρνουμε ότι

$$\text{tr}(\bar{\Sigma}A\bar{\Sigma}B) + 2\mu'A\bar{\Sigma}B\mu = 0.$$

Η σχέση αυτή ισχύει για κάθε μ , συνεπώς και για $\mu = 0$. Στην



περίπτωση αυτή $\text{tr}(\Sigma\Lambda\Sigma\mathbf{B}) = 0$, και αντικαθιστώντας στην προηγούμενη σχέση παίρνουμε ότι $2\mu'\Lambda\Sigma\mathbf{B}\mu = 0$. Αλλά και η σχέση αυτή ισχύει για κάθε μ και συνεπώς $\Lambda\Sigma\mathbf{B} = 0$.

Είναι γνωστό ότι αν $\mathbf{x} \sim N_p(\mathbf{0}, I_p)$, τότε η κατανομή του $\mathbf{x}'\mathbf{x} = \sum x_i^2$ είναι η κεντρική χ_p^2 -κατανομή. Αν τώρα $\mathbf{x} \sim N_p(\mu, I_p)$, τότε η κατανομή του $\mathbf{x}'\mathbf{x}$ είναι η μη-κεντρική $\chi_p^2(\lambda)$ -κατανομή, όπου $\lambda = (1/2)\mu'\mu$ είναι η μη κεντρική παράμετρος της χ^2 -κατανομής.

Μπορεί να αποδειχθεί ότι αν $U \sim \chi_p^2(\lambda)$ τότε η ροπογεννήτρια της U δίνεται από την σχέση

$$m_U(t) = (1 - 2t)^{-p/2} e^{-\lambda[1-(1-2t)^{-1}]}, \quad t \in \mathbb{R}. \quad (2.2)$$

Θεώρημα 2.10. Αν $\mathbf{x} \sim N_p(\mu, \Sigma)$, τότε $\mathbf{x}'\Lambda\mathbf{x} \sim \chi_r^2(\frac{1}{2}\mu'\Lambda\mu)$, όπου $r = \text{Rank}(\Lambda)$, εάν και μόνον εάν ο πίνακας $\Lambda\Sigma$ είναι ταυτοδύναμος.

Απόδειξη. Ας υποθέσουμε ότι ο πίνακας $\Lambda\Sigma$ είναι ταυτοδύναμος. Επειδή $\text{Rank}(\Lambda) = r$ και $\text{Rank}(\Sigma) = p$ έχουμε ότι $\text{Rank}(\Lambda\Sigma) = r$. Ας συμβολίσουμε με λ_i ($i = 1, 2, \dots, p$) τις ιδιοτιμές του πίνακα $\Lambda\Sigma$. Το γεγονός ότι ο $\Lambda\Sigma$ είναι ταυτοδύναμος με $\text{Rank}(\Lambda\Sigma) = r$, συνεπάγεται ότι r από τις ιδιοτιμές λ_i θα είναι ίσες με τη μονάδα και $p - r$ ίσες με το μηδέν. Με βάση αυτά και από το θεώρημα 2.8 (ii) παίρνουμε

$$m_{\mathbf{x}'\Lambda\mathbf{x}}(t) = \left[\det(I - 2t\Lambda\Sigma) \right]^{-1/2} e^{-\frac{1}{2}\mu'[I - (I - 2t\Lambda\Sigma)^{-1}]\Sigma^{-1}\mu}$$

$$= \prod_{i=1}^p (1 - 2t\lambda_i)^{-1/2} e^{-\frac{1}{2}\mu' \left[-\sum_{k=1}^{\infty} (2t)^k (\Lambda\Sigma)^k \right] \Sigma^{-1}\mu}$$



$$\begin{aligned}
&= \prod_{i=1}^p (1-2t)^{-1/2} e^{-\frac{1}{2} \mu' \left[-\sum_{k=1}^{\infty} (2t)^k \right] \Lambda \Sigma \Sigma^{-1} \mu} \\
&= (1-2t)^{-r/2} e^{-\frac{1}{2} \mu' [1-(1-2t)^{-1}] \Lambda \mu} \\
&= (1-2t)^{-r/2} e^{-\frac{1}{2} \mu' \Lambda \mu [1-(1-2t)^{-1}]} , t \in \mathbb{R}. \quad (2.3)
\end{aligned}$$

Συγκρίνοντας την τελευταία σχέση με την (2.2) έχουμε ότι

$$\mathbf{x}' \Lambda \mathbf{x} \sim \kappa_r^2 \left(\frac{1}{2} \mu' \Lambda \mu \right).$$

Αντίστροφα, ως υποθέσουμε ότι $\mathbf{x}' \Lambda \mathbf{x} \sim \kappa_r^2 \left(\frac{1}{2} \mu' \Lambda \mu \right)$.

Στην περίπτωση αυτή η ροπογεννήτρια της τετραγωνικής μορφής θα δίνεται από την σχέση (2.3). Η σχέση (2.3) πρέπει να ισούται με την σχέση στο θεώρημα 2.8 (ii), για κάθε τιμή του μ και συνεπώς και για $\mu = 0$. Αυτό μας δίνει

$$(1-2t)^{-r/2} = \left[\det(I - 2t\Lambda\Sigma) \right]^{-1/2}$$

ή

$$(1-w)^r = \det(I - 2t\Lambda\Sigma), \quad (w = 2t).$$

Αν συμβολίσουμε με λ_i ($i = 1, 2, \dots, p$) τις ιδιοτιμές του πίνακα $\Lambda\Sigma$, τότε η τελευταία σχέση γράφεται

$$(1-w)^r = \prod_{i=1}^p (1 - w \lambda_i).$$

Η σχέση αυτή είναι μια ταυτότητα ως προς w , όπου το δεξιό μέλος της δεν μπορεί να έχει δυνάμεις του w μεγαλύτερες από r . Αυτό σημαίνει ότι τουλάχιστον ένα από τα λ_i είναι μηδέν.



Επαναλαμβάνοντας τον ίδιο συλλογισμό παίρνουμε ότι $p-r$ από τα λ_i είναι μηδέν και συνεπώς

$$(1-w)^r = \prod_{i=1}^r (1 - w \lambda_i).$$

Αν στην τελευταία σχέση εξισώσουμε τους συντελεστές του w , και στα δύο μέλη της ισότητας, θα πάρουμε ένα σύστημα r εξισώσεων με r αγνώστους τα λ_i . Η λύση του συστήματος αυτού μας δίνει $\lambda_i = 1$, για $i = 1, 2, \dots, r$. Συνεπώς για τον πίνακα $A\Sigma$ με $\text{Rank}(A\Sigma) = r$, έχουμε ότι r από τις ιδιοτιμές του ισούνται με τη μονάδα ενώ οι υπόλοιπες $p - r$ ισούνται με μηδέν. Άρα ο πίνακας $A\Sigma$ είναι ταυτοδύναμος.

2.2 ΔΕΙΓΜΑΤΙΚΕΣ ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ

2.2.1 Κατανομή του Wishart

Η κατανομή του Wishart είναι η πολυδιάστατη γενίκευση της χ^2 -κατανομής.

Ορισμός 2.2 Έστω ότι $x_1, x_2, \dots, x_n$ είναι ανεξάρτητα τυχαία διανύσματα με $x_i \sim N_p(\mu_i, \Sigma)$, ($i = 1, 2, \dots, n$). Τότε ο τυχαίος

πίνακας $W = \sum_{i=1}^n x_i x_i'$ λέμε ότι έχει την μη - κεντρική κατανομή

Wishart με n βαθμούς ελευθερίας. Συμβολικά θα γράφουμε $W \sim W_p(n, \Sigma; M)$, όπου $M' = [\mu_1, \mu_2, \dots, \mu_n]$.

Αν $x_i \sim N_p(0, \Sigma)$ ($i = 1, 2, \dots, n$), τότε η κατανομή του πίνακα W είναι η κεντρική κατανομή Wishart με n βαθμούς ελευθερίας. Συμβολικά $W \sim W_p(n, \Sigma)$.



Από τον παραπάνω ορισμό προκύπτει ότι αν ο $p \times p$ πίνακας W μπορεί να γραφεί σαν $W = X'X$, όπου X είναι ένας $n \times p$ πίνακας του οποίου η i γραμμή ακολουθεί την $N_p(\mu_i, \Sigma)$ ($i = 1, 2, \dots, n$) [ή την $N_p(0, \Sigma)$] κατανομή, τότε η κατανομή του πίνακα W είναι η μη - κεντρική Wishart (ή η κεντρική Wishart).

Αν $\Sigma > 0$ τότε μιλάμε για μη ιδιάζων κατανομή Wishart, ενώ αν $\Sigma \geq 0$ μιλάμε για την ιδιάζων κατανομή Wishart. Ακόμη αν $p = 1$ τότε η κεντρική κατανομή Wishart αντιστοιχεί στην κεντρική χ^2 - κατανομή. Δηλαδή η $W_1(n, \sigma^2)$ είναι ισοδύναμη με την $\sigma^2 \chi_n^2$ - κατανομή. Προφανώς η μη - κεντρική κατανομή Wishart αντιστοιχεί στην μη - κεντρική χ^2 - κατανομή.

Μπορεί να αποδειχθεί ότι αν $n \geq p$ και $\Sigma \geq 0$, τότε $W > 0$.

Σε ότι ακολουθεί θα ασχοληθούμε μόνο με την κεντρική κατανομή Wishart την οποία για συντομία θα αναφέρουμε απλώς σαν κατανομή Wishart. Η αναλυτική μορφή της κατανομής αυτής είναι η ακόλουθη.

$$f(W) = W_p(n, \Sigma) = \begin{cases} \frac{\det(W)^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}\text{tr}(W\Sigma^{-1})}}{\det(\Sigma)^{n/2} \prod_{j=1}^p \Gamma\left[\frac{1}{2}(n+1-j)\right]} & n \geq p, \Sigma \geq 0 \\ 0, & \text{αλλού} \end{cases}$$

$$\text{όπου } c_p = 2^{np/2} \pi^{p(p-1)/4}.$$

Στην συνέχεια θα δώσουμε μερικές από τις ιδιότητες της κατανομής Wishart.

Θεώρημα 2.11 Αν $W \sim W_p(n, \Sigma)$, τότε $E(W) = n \Sigma$.



Απόδειξη Από τον ορισμό της κατανομής Wishart έχουμε ότι

$$W = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' . \text{ Συνεπώς } E(W) = \sum_{i=1}^n E(\mathbf{x}_i \mathbf{x}_i') . \text{ Αλλά } E(\mathbf{x}_i \mathbf{x}_i') = \Sigma .$$

Θεώρημα 2.12 Αν $W \sim W_p(n, \Sigma)$ και B είναι ένας σταθερός $p \times q$ πίνακας , τότε

$$B'WB \sim W_q(n, B'\Sigma B) .$$

Απόδειξη Έστω ότι X είναι ένας $n \times p$ πίνακας με γραμμές τα τ. δ. $\mathbf{x}_i'$ ($i = 1, 2, \dots, n$) . (Να θυμήσουμε εδώ ότι $\mathbf{x}_i$ είναι το διάνυσμα στήλης με συνιστώσες τα στοιχεία της i γραμμής του X). Τότε $W = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' = X'X$. Άρα $B'WB = B'X'XB = Y'Y$, όπου $Y = XB$. Οι γραμμές όμως του X είναι ανεξάρτητα τ. δ. με κατανομή την $N_p(0, \Sigma)$. Συνεπώς $Y \sim N_q(0, B'\Sigma B)$. Ο Ορισμός 2.2 συμπληρώνει την απόδειξη.

Πόρισμα 2.12.1 Διαγώνιοι υποπίνακες του W έχουν μια κατανομή Wishart .

Πόρισμα 2.12.2 $\Sigma^{-1/2} W \Sigma^{-1/2} \sim W_p(n, I)$.

Πόρισμα 2.12.3 Αν $W \sim W_p(n, I)$ και B είναι ένας σταθερός $p \times q$ πίνακας τέτοιος ώστε $B'B = I_q$, τότε $B'WB \sim W_q(n, I)$.

Θεώρημα 2.13 Αν $W \sim W_p(n, \Sigma)$ και $\mathbf{a}$ είναι ένα σταθερό $p \times 1$ διάνυσμα τέτοιο ώστε $\mathbf{a}'\Sigma\mathbf{a} \neq 0$, τότε

$$\frac{\mathbf{a}'W\mathbf{a}}{\mathbf{a}'\Sigma\mathbf{a}} \sim \chi^2_n .$$

Απόδειξη Από το προηγούμενο Θεώρημα $\mathbf{a}'W\mathbf{a} \sim W_1(n, \mathbf{a}'\Sigma\mathbf{a})$.



Το επόμενο θεώρημα μας λέει ότι το άθροισμα ανεξάρτητων Wishart είναι πάλι κατανομή Wishart.

Θεώρημα 2.14 Αν $W_1 \sim W_p(n_1, \Sigma)$, $W_2 \sim W_p(n_2, \Sigma)$ και W_1 και W_2 είναι ανεξάρτητα, τότε $W = W_1 + W_2 \sim W_p(n_1 + n_2, \Sigma)$.

Απόδειξη Αν X_i ($i = 1, 2$) είναι ένας $n_i \times p$ πίνακας του οποίου οι n_i γραμμές είναι ανεξάρτητα τ. δ. με κατανομή την $N_p(0, \Sigma)$, τότε $W_i = X_i' X_i$. Έτσι $W = W_1 + W_2 = X_1' X_1 + X_2' X_2 = X' X$ όπου $X = (X_1', X_2')'$. Τώρα οι X_1 και X_2 μπορούν να εκλεγούν κατά τέτοιο τρόπο ώστε να είναι ανεξάρτητοι μεταξύ τους. Αυτό σημαίνει ότι οι γραμμές του πίνακα X είναι ανεξάρτητες μεταξύ τους με κατανομή την $N_p(0, \Sigma)$. Ο Ορισμός της κατανομής Wishart συμπληρώνει την απόδειξη.

Μέχρι τώρα η κατανομή του Wishart βασίσθηκε σε ένα $n \times p$ πίνακα X του οποίου οι γραμμές είναι ανεξάρτητες και ακολουθούν την $N_p(0, \Sigma)$ κατανομή. Προφανώς και άλλες συναρτήσεις του X μπορεί να ακολουθούν την κατανομή Wishart. Στη συνέχεια θα δώσουμε, χωρίς απόδειξη, τρία θεωρήματα τα οποία εκτός του ότι θα χρησιμοποιηθούν στις επόμενες σελίδες, παρουσιάζουν και γενικώτερο ενδιαφέρον. (Για τις αποδείξεις βλ. Mardia, Kent & Bibby (1982)).

Θεώρημα 2.15 Έστω X ένας $n \times p$ πίνακας του οποίου οι γραμμές είναι ανεξάρτητες με κατανομή την $N_p(\mu, \Sigma)$. Αν $Y = AXB$ και $Z = CXD$, τότε τα στοιχεία του πίνακα Y είναι ανεξάρτητα από τα στοιχεία του πίνακα Z εάν και μόνον εάν

$$i) B' \Sigma D = 0 \text{ ή } ii) AC = 0$$

Το επόμενο θεώρημα είναι μέρος ενός γενικώτερου θεωρήματος γνωστού σαν θεώρημα του Cochran.



Θεώρημα 2.16 Έστω X ένας $n \times p$ πίνακας του οποίου οι γραμμές είναι ανεξάρτητες με κατανομή την $N_p(\mathbf{0}, \Sigma)$, και C ένας σταθερός $n \times n$ συμμετρικός πίνακας. Τότε $X'CX$ έχει την κατανομή Wishart εάν και μόνον εάν C είναι ταυτοδύναμος. Στην περίπτωση αυτή $X'CX \sim W_p(r, \Sigma)$, όπου $r = \text{tr}(C) = \text{Rank}(C)$.

Ας υποθέσουμε ότι οι $p \times p$ πίνακες W και Σ διαμερίζονται ως εξής :

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \quad \& \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

όπου η τάξη των πινάκων W_{11} , Σ_{11} , W_{22} , Σ_{22} είναι $p_1 \times p_1$ και $p_2 \times p_2$, αντίστοιχα, με $p_1 + p_2 = p$. Τότε μπορούμε να δείξουμε το εξής

Θεώρημα 2.17 Αν $W \sim W_p(n, \Sigma)$ και $W_{22.1} = W_{22} - W_{21}W_{11}^{-1}W_{12}$, τότε για $n > p_1$

- i) $W_{21.1} \sim W_{p_2}(n - p_1, \Sigma_{22.1})$ και είναι ανεξάρτητο από το (W_{11}, W_{12}) όπου $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.
- ii) Αν $\Sigma_{12} = 0$ τότε ο $W_{22} - W_{22.1} \sim W_{p_2}(p_1, \Sigma_{22})$ και είναι ανεξάρτητο από το $W_{22.1}$.

Ας υποθέσουμε τώρα ότι X είναι ένας $n \times p$ πίνακας δεδομένων οι γραμμές του οποίου είναι ανεξάρτητα τ. δ. με κοινή κατανομή την $N_p(\mu, \Sigma)$. Έστω ακόμη ότι $\bar{X}$ είναι το διάνυσμα των δειγματικών μέσων και S ο πίνακας των δειγματικών διακυμάνσεων - συνδιακυμάνσεων. Από το Κεφάλαιο 1 ξέρουμε ότι $nS = X'HX$ όπου $H = I - (1/n)1'1$. Ο πίνακας H είναι συμμετρικός και ταυτοδύναμος με $\text{Rank}(H) = n - 1$. Σύμφωνα λοιπόν με το Θεώρημα



του Cochran (βλ. Θεώρημα 2.16) $X'HX \sim W_p(n-1, \Sigma)$, δηλαδή $nS \sim W_p(n-1, \Sigma)$.

Γράφοντας (βλ. Κεφάλαιο 1) $\bar{X}' = (1/n)1'X$ και θέτοντας $A = (1/n)1'$ και $C = H$ παίρνουμε ότι $AC' = 0$. Συνεπώς, από το Θεώρημα 2.15, έχουμε ότι $\bar{X}'$ είναι ανεξάρτητο από το HX και κατ' επέκταση είναι ανεξάρτητο από το S . Έχουμε λοιπόν αποδείξει το εξής Θεώρημα.

Θεώρημα 2.18 Έστω X ένας $n \times p$ πίνακας του οποίου οι γραμμές είναι ανεξάρτητα τ. δ. με κοινή κατανομή την $N_p(\mu, \Sigma)$. Αν $\bar{X}$ είναι το διάνυσμα των δειγματικών μέσων και S ο πίνακας των δειγματικών διακυμάνσεων - συνδιακυμάνσεων τότε

i) $nS \sim W_p(n-1, \Sigma)$ και ii) $\bar{X}$ & S είναι ανεξάρτητα.

Παρατήρηση 2.3 Το Θεώρημα 2.18 αποτελεί γενίκευση αντίστοιχου θεωρήματος στην μονοδιάστατη περίπτωση.

2.2.2 Κατανομή του Hotelling

Η κατανομή αυτή, όπως και η κατανομή του Wishart, είναι μια δειγματική κατανομή. Αποτελεί δε μια γενίκευση της μονοδιάστατης t - κατανομής.

Ορισμός 2.3 Αν η τυχαία μεταβλητή X μπορεί να γραφεί σαν $nd'W^{-1}d$, όπου $d \sim N_p(0, I)$, $W \sim W_p(n, I)$ και d & W ανεξάρτητα, τότε λέμε ότι η κατανομή της τυχαίας μεταβλητής X είναι η Hotelling T^2 με παραμέτρους p και n , και συμβολικά $X \sim T^2(p, n)$.



Θεώρημα 2.19 Αν $\mathbf{x} \sim N_p(\mu, \Sigma)$, $\mathbf{M} \sim W_p(m, \Sigma)$ και $\mathbf{x}$ & $\mathbf{M}$ είναι ανεξάρτητα, τότε $m(\mathbf{x}-\mu)' \mathbf{M}^{-1}(\mathbf{x}-\mu) \sim T^2(p, m)$.

Απόδειξη . Αν θέσουμε $\mathbf{d} = \Sigma^{-1/2}(\mathbf{x} - \mu)$ και $\mathbf{W} = \Sigma^{1/2} \mathbf{M} \Sigma^{1/2}$, τότε από το Θεώρημα 2.1 και το Πρόγραμμα 2.12.2 παίρνουμε, αντίστοιχα, ότι $\mathbf{d} \sim N_p(0, \mathbf{I})$ και $\mathbf{W} \sim W_p(m, \mathbf{I})$. Η ανεξαρτησία των $\mathbf{x}$ και $\mathbf{M}$ και ο Ορισμός 2.3 ολοκληρώνουν την απόδειξη.

Πρόγραμμα 2.19.1 Αν $\bar{\mathbf{x}}$ και $\mathbf{S}$ είναι, αντίστοιχα, το διάνυσμα των δειγματικών μέσων και ο πίνακας των δειγματικών διακυμάνσεων - συνδιακυμάνσεων, για ένα τυχαίο δείγμα μεγέθους n από την $N_p(\mu, \Sigma)$, τότε

$$(n-1)(\bar{\mathbf{x}}-\mu)' \mathbf{S}^{-1}(\bar{\mathbf{x}}-\mu) = n(\bar{\mathbf{x}}-\mu)' \mathbf{S}_u^{-1}(\bar{\mathbf{x}}-\mu) \sim T^2(p, n-1),$$

όπου $\mathbf{S}_u = n \mathbf{S} / (n - 1)$.

Απόδειξη . Θέτοντας, στο προηγούμενο θεώρημα, $\mathbf{M} = n \mathbf{S}$, $m = n - 1$ και $\mathbf{x} - \mu = n^{1/2}(\bar{\mathbf{x}} - \mu)$ έχουμε το ζητούμενο.

Το επόμενο θεώρημα εξυπηρετεί δύο σκοπούς. Αφ' ενός μεν μας δίνει την αναλυτική μορφή της T^2 - κατανομής , αφ' ετέρου δε γενικεύει το γνωστό μας , από την μονοδιάστατη περίπτωση, αποτέλεσμα, ότι δηλαδή

$$t_n^2 = F_{1, n} ,$$

όπου $F_{1, n}$ είναι η F κατανομή με 1 και n βαθμούς ελευθερίας.

Θεώρημα 2.20. $T^2(n, p) = \frac{np}{n - p + 1} F_{p, n-p+1}$



Πριν από την απόδειξη του θεωρήματος θα αποδείξουμε το εξής Λήμμα.

Λήμμα 2.2 Αν $W \sim W_p(n, \Sigma)$ με $n > p$, τότε

$$a' \Sigma^{-1} a / a' W^{-1} a \sim \chi^2_{n-p+1},$$

για κάθε σταθερό διάνυσμα a .

Απόδειξη . Γράφοντας

$$W(p \times p) = \begin{bmatrix} W_{11}(p_1 \times p_1) & W_{12}(p_1 \times p_2) \\ W_{21}(p_2 \times p_1) & W_{22}(p_2 \times p_2) \end{bmatrix} \quad \text{παίρνουμε} \quad W^{-1} = \begin{bmatrix} W^{11} & W^{12} \\ W^{21} & W^{22} \end{bmatrix},$$

όπου οι εντός παρενθέσεως ποσότητες δηλώνουν την τάξη των αντίστοιχων πινάκων και $W^{22} = W_{22.1}^{-1}$. Συνεπώς, από το θεώρημα 2.17(i), έχουμε ότι

$$\left(W^{22}\right)^{-1} \sim W_{p_2}\left(n-p_1, \left(\Sigma^{22}\right)^{-1}\right) \quad \text{και ακόμη ότι} \quad \left(W^{22}\right)^{-1} \text{ είναι}$$

ανεξάρτητο από το (W_{11}, W_{22}) . Αν τώρα στην διαμέριση του πίνακα W θέσουμε $p_2 = 1$, οπότε $p_1 = p-1$, τότε η σχέση του Λήμματος γράφεται σαν

$$\left(W^{pp}\right)^{-1} \sim \left(\sigma^{pp}\right)^{-1} \chi^2_{n-p+1}. \quad (2.4)$$

Συνεπώς το προς απόδειξη Λήμμα ισχύει για $a = (0, 0, \dots, 0, 1)'$. Για να αποδείξουμε τώρα ότι το Λήμμα ισχύει για κάθε διάνυσμα a εργαζόμαστε ως εξής. Έστω A ένας μη ιδιάζων πίνακας του οποίου η τελευταία στήλη είναι ένα διάνυσμα a . Θέτοντας $M = A^{-1}W(A^{-1})'$ παίρνουμε, από το θεώρημα 2.12, ότι

$M \sim W_p[n, A^{-1}\Sigma(A^{-1})']$. Όμως $M^{-1} = A'W^{-1}A$ και από την σχέση (2.4) έχουμε

$$\left(m^{pp}\right)^{-1} = \left(a'W^{-1}a\right)^{-1} \sim \left[\left(A'\Sigma^{-1}A\right)_{pp}\right]^{-1} \chi^2_{n-p+1} = \left(a'\Sigma^{-1}a\right)^{-1} \chi^2_{n-p+1}.$$



Παρατήρηση 2.4 Μια γενίκευση του παραπάνω Λήμματος δίνεται από τον Eaton (1972) . Η γενίκευση αυτή έχει ως εξής. "Αν $W \sim W_p(n, \Sigma)$ και A είναι ένας $q \times p$ σταθερός πίνακας με $\text{Rank}(A) = q$, τότε

$$\left(A W^{-1} A' \right)^{-1} \sim W_q \left(n-p-q, (A \Sigma^{-1} A')^{-1} \right).$$

Παρατήρηση 2.5 Μερικοί συγγραφείς (βλ. π.χ Arnold (1981)) χρησιμοποιούν το θεώρημα 2.20 σαν ορισμό της κατανομής Hotelling

Απόδειξη του Θεωρήματος 2.20 Από τον Ορισμό 2.3 μπορούμε να γράψουμε $X = n(d' W^{-1} d / d' d) d' d$. Από την ανεξαρτησία των d και W και με βάση το προηγούμενο Λήμμα έχουμε ότι η υπό συνθήκη κατανομή της τυχαίας μεταβλητής $Y = d' d / d' W^{-1} d$, δεθέντος d , είναι η χ^2_{n-p+1} . Από το Θεώρημα 2.10 έχουμε ότι $d' d \sim \chi^2_p$, άρα μπορούμε να γράψουμε

$$X = n \frac{\chi^2_p}{\chi^2_{n-p+1}} = \frac{n-p}{n-p+1} F_{p, n-p+1}.$$

Πόρισμα 2.20.1 Αν $\bar{X}$ και S είναι, αντίστοιχα, το διάνυσμα των δειγματικών μέσων και ο πίνακας των δειγματικών διακυμάνσεων - συνδιακυμάνσεων ενός τυχαίου δείγματος μεγέθους n από την $N_p(\mu, \Sigma)$, τότε

$$\frac{n-p}{p} (\bar{X} - \mu)' S^{-1} (\bar{X} - \mu) \sim F_{p, n-p}.$$

Απόδειξη . Άμεση από το Πόρισμα 2.19.1 .

Για την μονοδιάστατη περίπτωση είναι γνωστό ότι αν $X \sim$



$N(0,1)$, $Y^2 \sim \chi^2$ και οι τυχαίες μεταβλητές X και Y^2 είναι ανεξάρτητες, τότε το X^2+Y^2 είναι ανεξάρτητο από το X^2/Y^2 . Μια γενίκευση αυτού του αποτελέσματος είναι το επόμενο θεώρημα.

Θεώρημα 2.21 Αν d και W είναι ανεξάρτητα με κατανομές $N_p(0,I)$ και $W_p(n,I)$, αντίστοιχα, τότε για την τυχαία μεταβλητή

$$Z = d'd \left(1 + \frac{1}{d'W^{-1}d} \right) ,$$

ισχύει ότι $Z \sim \chi^2_{n+1}$ και είναι ανεξάρτητη από την τ. μ. $d'W^{-1}d$.

Απόδειξη . Από την απόδειξη του θεωρήματος 2.20 έχουμε ότι οι τυχαίες μεταβλητές $d'd$ και Y είναι ανεξάρτητες ενώ κάθε μία έχει την χ^2 - κατανομή . Από το αποτέλεσμα στην μονοδιάστατη περίπτωση θα έχουμε ότι το άθροισμα αυτών των τ. μ. θα ακολουθεί την χ^2 - κατανομή και θα είναι ανεξάρτητο από το πηλίκο τους. Αλλά το άθροισμα των μεταβλητών αυτών είναι

$$d'd + Y = d'd \left(1 + \frac{1}{d'W^{-1}d} \right) \sim \chi^2_{n+1} ,$$

το δε πηλίκο τους είναι

$$\frac{d'd}{Y} = d'W^{-1}d.$$

Παρατήρηση 2.6 Ένα αποτέλεσμα γενικώτερο από αυτό του προηγούμενου θεωρήματος είναι το εξής. " Αν d και W είναι στατιστικώς ανεξάρτητα και $d \sim N_p(0,I)$ και $W \sim W_p(n,I)$, τότε $d'W^{-1}d$ είναι ανεξάρτητο από το $W + d d'$." Για μια απόδειξη βλ. Mardia, Kent & Bibby (1979).



Θεώρημα 2.22 Έστω $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ και $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ δύο τυχαία δείγματα από τους πληθυσμούς $N_p(\mu_{\mathbf{x}}, \Sigma_{\mathbf{x}})$ και $N_p(\mu_{\mathbf{y}}, \Sigma_{\mathbf{y}})$, αντίστοιχα. Αν $\mu_{\mathbf{x}} = \mu_{\mathbf{y}}$ και $\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{y}}$, τότε

$$\frac{n \cdot m}{n + m} D^2 \sim T^2(p, N-2),$$

όπου $D^2 = (\bar{\mathbf{x}} - \bar{\mathbf{y}})' \mathbf{S}_u^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$, $\mathbf{S}_u = \frac{n \mathbf{S}_{\mathbf{x}} + m \mathbf{S}_{\mathbf{y}}}{n + m - 2}$, και $N = n + m$.

Απόδειξη. Επειδή $\bar{\mathbf{x}} \sim N_p(\mu_{\mathbf{x}}, n^{-1} \Sigma_{\mathbf{x}})$, $\bar{\mathbf{y}} \sim N_p(\mu_{\mathbf{y}}, m^{-1} \Sigma_{\mathbf{y}})$ και τα δύο δείγματα είναι ανεξάρτητα παίρνουμε ότι η κατανομή του $\mathbf{d} = \bar{\mathbf{x}} - \bar{\mathbf{y}}$ είναι η $N_p(\mu_{\mathbf{x}} - \mu_{\mathbf{y}}, n^{-1} \Sigma_{\mathbf{x}} + m^{-1} \Sigma_{\mathbf{y}})$. Συνεπώς αν

$$\mu_{\mathbf{x}} = \mu_{\mathbf{y}} \text{ και } \Sigma_{\mathbf{x}} = \Sigma_{\mathbf{y}} = \Sigma, \text{ τότε } \mathbf{d} \sim N_p\left(0, \frac{n + m}{n \cdot m} \Sigma\right).$$

Αν θέσουμε $\mathbf{W}_{\mathbf{x}} = n \cdot \mathbf{S}_{\mathbf{x}}$ και $\mathbf{W}_{\mathbf{y}} = m \cdot \mathbf{S}_{\mathbf{y}}$, θα έχουμε ότι $\mathbf{W}_{\mathbf{x}} \sim W_p(n-1, \Sigma_{\mathbf{x}})$ και $\mathbf{W}_{\mathbf{y}} \sim W_p(m-1, \Sigma_{\mathbf{y}})$. Συνεπώς όταν $\Sigma_{\mathbf{x}} = \Sigma_{\mathbf{y}} = \Sigma$ θα είναι $\mathbf{W} = (N-2)\mathbf{S} = \mathbf{W}_{\mathbf{x}} + \mathbf{W}_{\mathbf{y}} \sim W_p(N-2, \Sigma)$. Άρα αν $c = \frac{n + m}{n \cdot m}$, τότε $c \cdot \mathbf{W} \sim W_p(N-2, c \cdot \Sigma)$. Αλλά $\mathbf{W}$ είναι ανεξάρτητο από το $\mathbf{d}$ επειδή τα $\bar{\mathbf{x}}$ και $\bar{\mathbf{y}}$ είναι ανεξάρτητα από τα $\mathbf{S}_{\mathbf{x}}$ και $\mathbf{S}_{\mathbf{y}}$, αντίστοιχα, και τα δύο δείγματα είναι ανεξάρτητα. Συνέπεια αυτού είναι ότι $(N-2)\mathbf{d}'(c \cdot \mathbf{W})^{-1}\mathbf{d} \sim T^2(p, N-2)$, ή αντικαθιστώντας

$$\begin{aligned} (N-2)(\bar{\mathbf{x}} - \bar{\mathbf{y}})' \frac{n \cdot m}{n + m} \cdot \frac{1}{n \cdot \mathbf{S}_{\mathbf{x}} + m \cdot \mathbf{S}_{\mathbf{y}}} (\bar{\mathbf{x}} - \bar{\mathbf{y}}) &= \\ &= \frac{n \cdot m}{n + m} (\bar{\mathbf{x}} - \bar{\mathbf{y}})' \frac{n + m - 2}{n \cdot \mathbf{S}_{\mathbf{x}} + m \cdot \mathbf{S}_{\mathbf{y}}} (\bar{\mathbf{x}} - \bar{\mathbf{y}}) \\ &= \frac{n \cdot m}{n + m} (\bar{\mathbf{x}} - \bar{\mathbf{y}})' \mathbf{S}_u^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}}) = \frac{n \cdot m}{n + m} D^2 \sim T^2(p, N-2) \end{aligned}$$



Παρατήρηση 2.7 Η ποσότητα $\Delta^2 = (\mu_{\mathbf{x}} - \mu_{\mathbf{y}})' \Sigma^{-1} (\mu_{\mathbf{x}} - \mu_{\mathbf{y}})$ είναι γνωστή σαν απόσταση του Mahalanobis (Mahalanobis distance) μεταξύ δύο πληθυσμών. Η δειγματική απόσταση του Mahalanobis ορίζεται από την σχέση $D^2 = (\bar{\mathbf{x}} - \bar{\mathbf{y}})' S_u^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$, όπου $S_u = (n \cdot S_{\mathbf{x}} + m \cdot S_{\mathbf{y}}) / (N-2)$. Η ποσότητα S_u είναι ένας αμερόληπτος εκτιμητής του Σ .

2.2.3 Πολυδιάστατη t - κατανομή

Ορισμός 2.4 Έστω ότι $\mathbf{z} \sim N_p(\mathbf{0}, \mathbf{I})$, $Y \sim \chi_m^2$ και $\mathbf{z}$, Y ανεξάρτητα. Τότε η κατανομή του τ. δ. $\mathbf{x} = \mathbf{z} / \sqrt{Y/m}$ είναι η p - διάστατη t - κατανομή με m βαθμούς ελευθερίας.

Χρησιμοποιώντας τη μέθοδο της αλλαγής μεταβλητών μπορούμε να δείξουμε ότι η μαθηματική έκφραση της κατανομής του $\mathbf{x}$ (δηλαδή ο τύπος της p - διάστατης t - κατανομής) είναι ο εξής.

$$f(\mathbf{x}) = \frac{\Gamma\left(\frac{m+p}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \cdot (m\pi)^{p/2}} \cdot \frac{1}{\left(1 + \frac{1}{m} \mathbf{x}' \mathbf{x}\right)^{(m+p)/2}}, \quad \mathbf{x} \in \mathbb{R}^p.$$

Για $m = 1$ η p - διάστατη t - κατανομή λέγεται p - διάστατη κατανομή του Cauchy.

Για την p - διάστατη t - κατανομή με m βαθμούς ελευθερίας μπερεί ναδειχθεί ότι i) $E(\mathbf{x}) = \mathbf{0}$, για $m > 1$ και ii) $\text{Var}(\mathbf{x}) = [m/(m-2)]\mathbf{I}$, για $m > 2$.



2.3 ΜΕΡΙΚΕΣ ΑΛΛΕΣ ΠΟΛΥΔΙΑΣΤΑΤΕΣ ΚΑΤΑΝΟΜΕΣ

2.3.1 Πολυωνυμική κατανομή

Ορισμός 2.5 Έστω $\mathbf{x} = (x_1, x_2, \dots, x_k)'$ ένα διακριτό τ. δ. . Θα λέμε ότι η κατανομή του $\mathbf{x}$ είναι η k - διάστατη πολυωνυμική αν η συνάρτηση πιθανότητας του $\mathbf{x}$ είναι η

$$p(\mathbf{x}) = \begin{cases} \frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i} & , x \geq 0, p \geq 0, (i=1,2,\dots,k) \\ 0 & , \text{αλλού} \end{cases}$$

$$\text{με } \sum_{i=1}^k x_i = n \text{ και } \sum_{i=1}^k p_i = 1 .$$

Μερικές από τις ιδιότητες της k - διάστατης πολυωνυμικής κατανομής είναι:

i) η κατανομή της τυχαίας μεταβλητής X_i είναι η $B(n, p_i)$, ($i = 1, 2, \dots, k$)

ii) $\mu = E(\mathbf{x}) = (np_1, np_2, \dots, np_k)'$.

iii) $\Sigma = \text{Var}(\mathbf{x}) = (\sigma_{ij})$, όπου $\sigma_{ij} = \text{Cov}(X_i, X_j) = -np_i p_j$, για κάθε $i \neq j$ και $\sigma_{ii} = \text{Var}(X_i) = np_i(1-p_i)$. Να σημειωθεί ότι ο πίνακας Σ δεν μπορεί να είναι θετικά ορισμένος.

2.3.2 Βήτα κατανομή ή κατανομή του Dirichlet

Ορισμός 2.6 Αν η συνάρτηση πυκνότητας πιθανότητας ενός τ. δ. $\mathbf{x} = (X_1, X_2, \dots, X_p)'$ γράφεται στην μορφή



$$f(\mathbf{x}) = \begin{cases} \frac{\Gamma\left(\sum_{i=1}^{p+1} \alpha_i\right)}{\Gamma\left(\prod_{i=1}^{p+1} \alpha_i\right)} \prod_{i=1}^p x_i^{\alpha_i-1} \cdot \left(1 - \sum_{i=1}^p x_i\right)^{\alpha_{p+1}-1}, & \mathbf{x} \in A \\ 0, & \mathbf{x} \notin A \end{cases}$$

όπου $A = \left\{ \mathbf{x} = (x_1, x_2, \dots, x_p)' : x_i \geq 0, \sum_{i=1}^p x_i \leq 1 \right\}$, τότε λέμε ότι η κατανομή του $\mathbf{x}$ είναι η πολυδιάστατη Βήτα ή η κατανομή του Dirichlet.

Η κατανομή του Dirichlet αποτελεί γενίκευση της γνωστής κατανομής Βήτα στην μονοδιάστατη περίπτωση.



ΑΣΚΗΣΕΙΣ

2.1 Έστω ότι

$$\mathbf{x} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left(\begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \right)$$

α) Να βρεθεί η κατανομή του τυχαίου διανύσματος $\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$,όπου $Y_1 = X_1 + X_2 + X_3 + X_4$ και $Y_2 = X_1 - X_2 + X_3 + X_4$.β) Είναι οι τυχαίες μεταβλητές Y_1 και Y_2 ανεξάρτητες; Ποια είναι η κατανομή της Y_1 ;γ) Να βρεθεί η κατανομή του τυχαίου διανύσματος $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.δ) Να βρεθεί η δεσμευμένη κατανομή του $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ δοθέντος ότι $X_3 = x_3$.

2.2 Έστω $\mathbf{X} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$, $\mathbf{A}$ και $\mathbf{B}$ $p \times n$ και $s \times n$ πίνακες αντίστοιχα. α) Να δειχθεί ότι το τυχαίο διάνυσμα $\begin{pmatrix} \mathbf{AX} \\ \mathbf{BX} \end{pmatrix}$ έχει την $(p+s)$ -διάστατη κανονική κατανομή και να δοθούν η μέση τιμή και ο πίνακας διακυμάνσεων-συνδιακυμάνσεων. β) Να δειχθεί ότι $\mathbf{AX}$ και $\mathbf{BX}$ είναι ανεξάρτητα εάν και μόνον εάν $\mathbf{AB}' = \mathbf{0}$.

2.3 Έστω $\mathbf{x} \sim N_3(\mathbf{0}, \boldsymbol{\Sigma})$ όπου $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & p & 0 \\ p & 1 & p \\ 0 & p & 1 \end{pmatrix}$, $|p| \leq 1$. Υπάρχει τιμή του p για την οποία οι τυχαίες μεταβλητές $X_1 + X_2 + X_3$ και $X_1 - X_2 - X_3$ είναι ανεξάρτητες; $[X_i \ (i=1,2,3)]$ είναι οι συνιστώσες



του διανύσματος $\mathbf{x}$]

2.4 Αν x_1, x_2, x_3 είναι ανεξάρτητα τυχαία διανύσματα από την $N_p(\mu, \Sigma)$ και $y_1 = x_1 + x_2$, $y_2 = x_2 + x_3$, $y_3 = x_1 + x_3$ να βρεθεί

α) η υπό συνθήκη κατανομή της y_1 δοθέντος y_2 .

β) η υπο συνθήκη κατανομή του y_1 δοθέντος y_2 και y_3 .

2.5 Αν $\mathbf{x} \sim N_p(\mu, \Sigma)$ ναδειχθεί ότι $x_i \sim N(\mu_i, \sigma_{ii})$, ($i = 1, 2, \dots, p$) όπου x_i και μ είναι οι συνιστώσες των διανυσμάτων $\mathbf{x}$ και μ αντίστοιχα και σ_{ii} είναι το i -στό διαγώνιο στοιχείο του πίνακα Σ .

2.6 Έστω $\mathbf{x} \sim N_3(\mu, \Sigma)$ με $\mu = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$ και $\Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix}$.

να βρεθεί η από κοινού κατανομή των τυχαίων μεταβλητών

$$Y_1 = X_1 + X_2 + X_3 \quad \text{και} \quad Y_2 = X_1 - X_2.$$

2.7 Αν (x_1, y_1) , (x_2, y_2) , $\dots$, (x_n, y_n) είναι ένα τυχαίο δείγμα από την $N_2(\mu, \Sigma)$ να βρεθεί η από κοινού κατανομή των δειγματικών μέσων $\bar{X}$, $\bar{Y}$.

2.8 Αν X και Y είναι δύο τυχαίες μεταβλητές με από κοινού συνάρτηση κατανομής την

$$f(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \left(1 - \frac{xy}{(1+x^2)(1+y^2)} \right), \quad \begin{matrix} x \in \mathbb{R} \\ y \in \mathbb{R} \end{matrix},$$

ναδειχθεί ότι οι κατανομές περιθωρίου των X και Y είναι κανονικές. Να σχολιάσετε το γεγονός αυτό.



2.9 Αν $y \sim N_n(0, I_n)$, $y'y = y'Ay + y'By$ και $y'Ay \sim \chi_r^2$, τότε $y'By \sim \chi_{n-r}^2$.

2.10 Έστω $y \sim N_p(0, I_n)$ και $Q_i = y'P_i y$ ($i = 1, 2$). Αν καθ'ένα από τα Q_i έχει την χ^2 -κατανομή να δειχθεί ότι τα Q_1 και Q_2 είναι ανεξάρτητα εάν και μόνον εάν $P_1 P_2 = 0$.

2.11 (Θεώρημα των Hogg & Graig) Έστω ότι $y \sim N_n(\mu, \sigma^2 I_n)$ και $Q_i = (y-\mu)'P_i(y-\mu)/\sigma^2$, ($i = 1, 2$). αν $Q_i \sim \chi_{r_i}^2$ και $Q_1 - Q_2 \geq 0$, τότε $Q_1 - Q_2$ είναι ανεξάρτητο από το Q_2 με κατανομές $\chi_{r_1}^2 - \chi_{r_2}^2$, και $\chi_{r_2}^2$, αντίστοιχα.

2.12 Αν $W \sim W_p(m, \Sigma)$, να δειχθεί ότι οι μεταβλητές $b'Wb$ και $d'Wd$ είναι στατιστικώς ανεξάρτητες εάν $b'\Sigma d = 0$.

2.13 Αν $W \sim W_p(m, \Sigma)$ να δειχθεί ότι τα w_{ii} και w_{jj} είναι ανεξάρτητα εάν $\sigma_{ij} = 0$. (w_{ii} και w_{jj} είναι τα διαγώνια στοιχεία του πίνακα W στις θέσεις (i, i) και (j, j) , αντίστοιχα, ενώ σ_{ij} είναι το στοιχείο του πίνακα Σ στη θέση (i, j)).

2.14 Αν $W \sim W_p(m, I)$ να δειχθεί ότι $\text{tr}W \sim \chi_{mp}^2$.

2.15 Έστω ότι $W \sim W_2(m, \Sigma)$. i) Να δειχθεί ότι

$$W = m \begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{bmatrix}$$

ii) να βρεθεί η κατανομή του $ms_2^2(1 - r^2)$ και να δειχθεί ότι είναι ανεξάρτητη από την από κοινού κατανομή των s_1^2 και rs_1s_2 .



2.16 Αν $W \sim W_p(n, \Sigma)$ να δειχθεί ότι $EW^{-1} = \Sigma^{-1}/(n-p-1)$.

2.17 Αν οι γραμμές του X είναι ανεξάρτητες $N_p(\mu, \Sigma)$ και C είναι ένας συμμετρικός πίνακας, τότε $X'CX$ έχει την $W_p(r, \Sigma)$ εάν και μόνον εάν i) C είναι ταυτοδύναμος και είτε α) $\mu = 0$ ή b) $C1 = 0$. Σε οποιαδήποτε περίπτωση $r = \text{tr}C$.

2.18 Έστω ότι $x \sim N_p(\mu, \Sigma)$ και $W \sim W_p(m, \Sigma)$ με x και W ανεξάρτητα. Αν η τυχαία μεταβλητή X μπορεί να γραφεί σαν $X = md'W^{-1}d$, όπου $d = x - \mu$, να δειχθεί ότι

$$i) \frac{d'W^{-1}d}{1 + d'W^{-1}d} \sim B\left(\frac{1}{2}p, \frac{1}{2}(m-p+1)\right)$$

$$ii) \frac{1}{1 + d'W^{-1}d} \sim B\left(\frac{1}{2}(m-p+1), \frac{1}{2}p\right)$$



3. ΕΚΤΙΜΗΤΙΚΗ

3.1 ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΚΑΙ ΕΠΑΡΚΕΙΑ

3.1.1 Συνάρτηση πιθανοφάνειας

Έστω ότι $x_1, x_2, \dots, x_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με κατανομή $f(x; \theta)$ όπου θ είναι το διάνυσμα των παραμέτρων. Τότε η συνάρτηση πιθανοφάνειας ορίζεται σαν

$$L(X; \theta) = \prod_{i=1}^n f(x_i; \theta) \quad , \quad (3.1)$$

όπου X είναι ο πίνακας με γραμμές τα διανύσματα x_i ($i = 1, 2, \dots, n$).

Ο λογάριθμος της συνάρτησης πιθανοφάνειας συμβολίζεται με $l(X; \theta)$ και ισούται με

$$l(X; \theta) = \log [L(X; \theta)] = \sum_{i=1}^n \log [f(x_i; \theta)] \quad . \quad (3.2)$$

Δοθέντος του τ. δ. $x_1, x_2, \dots, x_n$ τόσο η $L(X; \theta)$ όσο και $l(X; \theta)$ θεωρούνται συναρτήσεις του θ .

Η σκορ συνάρτηση (score function) $s = s(X; \theta)$ ορίζεται σαν



$$s(X; \theta) = \frac{\partial}{\partial \theta} \ln L(X; \theta) = \frac{1}{L(X; \theta)} \cdot \frac{\partial}{\partial \theta} L(X; \theta) . \quad (3.3)$$

Προφανώς $s(X; \theta)$ είναι ένα τυχαίο διάνυσμα και συνεπώς η $\text{Var}(s)$ θα υπάρχει. Ο πίνακας αυτός ονομάζεται *πίνακας πληροφορίας* του Fisher και συμβολίζεται με F . Δηλαδή $F = \text{Var}(s)$.

Θεώρημα 3.1 Έστω ότι $s(X; \theta)$ είναι η σκορ συνάρτηση που αντιστοιχεί στην συνάρτηση πιθανοφάνειας $L(X; \theta)$. Αν $t = t(X, \theta)$ είναι μια οποιαδήποτε συνάρτηση των X και θ , τότε, κάτω από κάποιες συνθήκες κανονικότητας, ισχύει

$$E(st') = \frac{\partial}{\partial \theta} E(t') - E\left(\frac{\partial t'}{\partial \theta}\right) .$$

Απόδειξη. Είναι γνωστό ότι $E(t') = \int t' L(X; \theta) dX$. Παραγωγίζοντας και τα δύο μέλη ως προς θ και υποθέτοντας ότι

$$\frac{\partial}{\partial \theta} \int t' L(X; \theta) dX = \int \frac{\partial}{\partial \theta} [t' L(X; \theta)] dX ,$$

παίρνουμε

$$\frac{\partial}{\partial \theta} E(t') = \int \frac{\partial t'}{\partial \theta} L(X; \theta) dX + \int \frac{\partial \ln L(X; \theta)}{\partial \theta} t' L(X; \theta) dX .$$

Παρατηρώντας ότι

$$E\left(\frac{\partial t'}{\partial \theta}\right) = \int \frac{\partial t'}{\partial \theta} L(X; \theta) dX \quad \text{και} \quad E(st') = \int \frac{\partial \ln L(X; \theta)}{\partial \theta} t' L(X; \theta) dX$$

έχουμε το ζητούμενο.

Πόρισμα 3.1.1 Αν $s(X; \theta)$ είναι η σκορ συνάρτηση που αντιστοιχεί στην συνάρτηση πιθανοφάνειας $L(X; \theta)$, τότε $E(s) = 0$.



Απόδειξη . Διαλέγοντας t να είναι ένα οποιοδήποτε σταθερό διάνυσμα θα έχουμε , από το προηγούμενο θεώρημα , ότι $E(s)t' = 0$, για κάθε σταθερό διάνυσμα t . Άρα $E(s) = 0$.

Πόρισμα 3.1.2 Αν $s(X;\theta)$ είναι η σκορ συνάρτηση που αντιστοιχεί στην συνάρτηση πιθανοφάνειας $L(X;\theta)$ και t είναι ένας αμερόληπτος εκτιμητής του θ , τότε $E(st') = I$.

Απόδειξη . Επειδή t είναι αμερόληπτος εκτιμητής του θ συνεπάγεται ότι $E(t) = \theta$ και το t δεν εξαρτάται από το θ δηλ. $\partial t' / \partial \theta = 0$.

Πόρισμα 3.1.3 Αν $s(X;\theta)$ είναι η σκορ συνάρτηση που αντιστοιχεί στην συνάρτηση πιθανοφάνειας $L(X;\theta)$ και t είναι ένας εκτιμητής τέτοιος ώστε $E(t) = \theta + b(\theta)$, τότε $E(st') = I + B$, όπου $B = B(\partial b_j / \partial \theta_i)$.

Απόδειξη . Επειδή το t είναι εκτιμητής θα είναι ανεξάρτητο από το θ και συνεπώς $\partial t' / \partial \theta = 0$.

Πόρισμα 3.1.4 Αν $s(X;\theta)$ είναι η σκορ συνάρτηση που αντιστοιχεί στην συνάρτηση πιθανοφάνειας $L(X;\theta)$ και $t = s$, τότε

$$F = \text{Var}(s) = - E \left(\frac{\partial t'}{\partial \theta} \right) = - E \left(\frac{\partial^2 l(X;\theta)}{\partial \theta \partial \theta'} \right) .$$

Απόδειξη . Η απόδειξη βασίζεται στο ότι $\text{Var}(s) = E(ss')$ και $E(s) = 0$ (βλ. Πόρισμα 3.1.1) .

Παράδειγμα 3.1 Έστω ένα τυχαίο δείγμα x_i ($i = 1, 2, \dots, n$) από την κανονική κατανομή $N(\mu, \Sigma)$. Τότε θα έχουμε



$$L(X; \theta) = L(X; \mu, \Sigma) = \frac{\exp\left\{-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu)\right\}}{\left[\det(2\pi\Sigma)\right]^{n/2}}$$

και

$$l(X; \mu, \Sigma) = -\frac{n}{2} \log_e \left[\det(2\pi \Sigma) \right] - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) .$$

Κάνοντας χρήση της ταυτότητας

$$\begin{aligned} (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) &= (\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) \\ &\quad + 2(\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) , \end{aligned}$$

μπορούμε να γράψουμε

$$\begin{aligned} \sum_{i=1}^n (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) &= \\ &= \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) + n (\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) , \end{aligned}$$

ή ακόμη επειδή

$$\begin{aligned} (\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) &= \text{tr} \left[(\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \right] = \\ &= \text{tr} \left[\Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})' \right] , \end{aligned}$$

έχουμε

$$\begin{aligned} \sum_{i=1}^n (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) &= \text{tr} \left\{ \Sigma^{-1} \left[\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})' \right] \right\} \\ &\quad + n (\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) . \end{aligned}$$



θέτοντας $n S = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ παίρνουμε τελικά

$$l(\mathbf{X}; \mu, \Sigma) = - \frac{n}{2} \log_e [\det(2\pi\Sigma)] - \frac{n}{2} \text{tr}(\Sigma^{-1}S) - \frac{n}{2} (\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) \quad (3.4)$$

Στην ειδική περίπτωση όπου $\Sigma = I$ η (3.4) γράφεται

$$l(\mathbf{X}; \mu, \Sigma) = - \frac{np}{2} \log_e (2\pi) - \frac{n}{2} \text{tr}(S) - \frac{n}{2} (\bar{\mathbf{x}} - \mu)' (\bar{\mathbf{x}} - \mu) .$$

Στην ίδια περίπτωση (δηλ. $\Sigma = I$) για την σκορ συνάρτηση θα έχουμε

$$s(\mathbf{X}; \mu) = \frac{\partial}{\partial \mu} l(\mathbf{X}; \mu) = n(\bar{\mathbf{x}} - \mu) , \quad \frac{\partial s'}{\partial \mu} = - nI$$

και συνεπώς $F = nI$.

3.1.2 Επάρκεια και πληρότητα

Έστω $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)'$ ένα τυχαίο δείγμα από μια κατανομή $f(\mathbf{x}; \theta)$ και $\mathbf{t} = \mathbf{t}(\mathbf{X})$ ένα στατιστικό .

Ορισμός 3.1 Θα λέμε ότι το στατιστικό $\mathbf{t} = \mathbf{t}(\mathbf{X})$ είναι επαρκές για την παράμετρο θ αν η υπό συνθήκη κατανομή του $\mathbf{X}$ δοθέντος $\mathbf{t}$ δεν εξαρτάται από το θ .

Στην μονοδιάστατη περίπτωση η εύρεση ενός επαρκούς στατιστικού γίνεται με το θεώρημα παραγοντοποίησews του Fisher. Γενίκευση του θεωρήματος αυτού είναι το επόμενο θεώρημα .



Θεώρημα 3.2 (Fisher - Neyman) 'Εστω το τυχαίο δείγμα $X = (x_1, x_2, \dots, x_n)$ από την κατανομή $f(x; \theta)$ και ένα στατιστικό $t = t(X)$. Το στατιστικό t είναι επαρκές για την παράμετρο θ εάν και μόνον εάν $L(X; \theta) = g[t(X); \theta] \cdot h(X)$ όπου g είναι μια συνάρτηση των t και θ και h είναι μια μη αρνητική συνάρτηση η οποία εξαρτάται μόνο από το X .

Απόδειξη . Για μια απόδειξη του Θεωρήματος παραπέμπουμε στο βιβλίο του Giri (1975) ή για μια πιο γενική απόδειξη στην εργασία των Halmos & Savage (1949) .

Ορισμός 3.2 'Ενα στατιστικό t ονομάζεται πλήρες αν για κάθε συνάρτηση $h(t)$, η σχέση $E[h(t)] = 0$ συνεπάγεται $h(t) = 0$ με πιθανότητα μονάδα .

Το επόμενο θεώρημα μας λέει ποιες συναρτήσεις, ενός πλήρους και επαρκούς στατιστικού είναι πλήρη και επαρκεί στατιστικά .

Θεώρημα 3.3 Αν $t(X)$ είναι ένα επαρκές και πλήρες στατιστικό και $f(t)$ είναι μια αντιστρέψιμη συνάρτηση του t , τότε η $f(t)$ είναι ένα επαρκές και πλήρες στατιστικό.

Απόδειξη . Η επάρκεια του $f(t)$ προκύπτει άμεσα από το Θεώρημα 3.2 . Η πληρότητα του $f(t)$ προκύπτει από τον Ορισμό του πλήρους στατιστικού (Ορισμός 3.2) .

Παράδειγμα 3.2 Αν x_i ($i = 1, 2, \dots, n$) είναι ένα τυχαίο δείγμα από την $N(\mu, \Sigma)$, τότε, όπως είδαμε στο Παράδειγμα 3.1



$$L(\mathbf{X}; \theta) = L(\mathbf{X}; \mu, \Sigma) = \frac{\exp\left\{-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu)\right\}}{\left[\det(2\pi\Sigma)\right]^{n/2}}.$$

Πάιρνοντας $h(\mathbf{X}) = 1$, βλέπουμε ότι τα $\bar{\mathbf{x}}$ και S είναι επαρκή στατιστικά για τα μ και Σ . Ακόμη μπορούμε να δείξουμε (βλ. Giri (1977) ή Arnold (1981)) ότι το επαρκές στατιστικό $(\bar{\mathbf{x}}, S)$ είναι πλήρες για το (μ, Σ) .

Παράδειγμα 3.3 Ένα πείραμα, από την γεννητική, έχει τέσσερα αποτελέσματα με πιθανότητες $(2+\theta)/4$, $(1-\theta)/4$, $(1-\theta)/4$ και $\theta/4$, αντίστοιχα. Σε n δοκιμές παρατηρούμε x_i εμφανίσεις του i αποτελέσματος ($i = 1, 2, 3, 4$), με $\sum x_i = n$. Αν $\mathbf{x} = (x_1, x_2, x_3, x_4)'$, τότε το τ. δ. $\mathbf{x}$ έχει την πολυωνυμική κατανομή με συνάρτηση πιθανοφάνειας

$$L(\mathbf{x}; \theta) = \frac{n!}{x_1! x_2! x_3! x_4!} \cdot \frac{(2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4}}{4^n}.$$

Ακόμη

$$l(\mathbf{x}; \theta) = \log_e c - n \log_e 4 + x_1 \log_e (2+\theta) + (x_2+x_3) \log_e (1-\theta) + x_4 \log_e \theta,$$

όπου $c = n!/(x_1! x_2! x_3! x_4!)$.

Η σκορ συνάρτηση θα δίνεται από την σχέση

$$s(\mathbf{x}; \theta) = \frac{\partial}{\partial \theta} l(\mathbf{x}; \theta) = \frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}$$

και συνεπώς, από το Πόρισμα 3.1.4,

$$F = -E\left[-\frac{x_1}{(2+\theta)^2} - \frac{x_2+x_3}{(1-\theta)^2} - \frac{x_4}{\theta^2}\right] = \frac{n(1+2\theta)}{\{2\theta(1-\theta)(2+\theta)\}}.$$

Από την μορφή της $L(\mathbf{x}; \theta)$, παίρνοντας

$$g(\mathbf{t}; \theta) = (2+\theta)^{x_1} (1-\theta)^{x_2+x_3} \theta^{x_4} \quad \text{και} \quad \mathbf{t} = (x_1, x_2+x_3, x_4) \quad \text{βλέπουμε}$$



ότι το t είναι ένα επαρκές στατιστικό για το θ . Ακόμη, επειδή $n = x_1 + x_2 + x_3 + x_4$ είναι σταθερό, το $t^* = (x_1, x_4)$ είναι επίσης ένα επαρκές στατιστικό για το θ .

Παράδειγμα 3.4 Αν x είναι ένα τ. δ. από την κατανομή $f(x; \theta)$ θα λέμε ότι η κατανομή του x ανήκει στην γενική εκθετική οικογένεια κατανομών αν

$$f(x; \theta) = \exp \left[a_0(\theta) + b_0(x) + \sum_{i=1}^q a_i(\theta) b_i(x) \right], \quad (3.5)$$

όπου $\theta = (\theta_1, \theta_2, \dots, \theta_r)'$. Στην περίπτωση όπου $r = q$ και $a_i(\theta) = \theta_i$ ($i \neq 0$), τότε η κατανομή του x ανήκει στην απλή εκθετική οικογένεια κατανομών.

Έστω τώρα $X = (x_1, x_2, \dots, x_n)'$ ένα τυχαίο δείγμα από την (3.5). Τότε η συνάρτηση πιθανοφάνειας του X θα είναι

$$\begin{aligned} L(X; \theta) &= \exp \left[n a_0(\theta) + \sum_{j=1}^n b_0(x_j) + \sum_{i=1}^q a_i(\theta) \cdot \sum_{j=1}^n b_i(x_j) \right] \\ &= \exp \left[n a_0(\theta) + \sum_{i=1}^q a_i(\theta) \cdot \sum_{j=1}^n b_i(x_j) \right] \cdot \exp \left[\sum_{j=1}^n b_0(x_j) \right]. \end{aligned}$$

Από την μορφή της $L(X; \theta)$ έχουμε ότι το διάνυσμα t με συνιστώσες $t_i = \sum_{j=1}^n b_i(x_j)$ είναι επαρκές στατιστικό για το θ .

Αν θέσουμε $a(\theta) = [a_1(\theta), a_2(\theta), \dots, a_q(\theta)]'$ με $a(\Omega) \neq \emptyset$, όπου Ω είναι ο παραμετρικός δειγματικός χώρος, τότε μπορεί να δειχθεί ότι το στατιστικό t είναι και πλήρες.



3.1.3 Αμερόληπτοι εκτιμητές ελάχιστης διακύμανσης

Έστω $\mathbf{X} = (x_1, x_2, \dots, x_n)'$ ένα τυχαίο δείγμα από την κατανομή $f(\mathbf{x}; \theta)$. Τότε το στατιστικό $t(\mathbf{X})$ είναι ένας αμερόληπτος εκτιμητής του $\tau(\theta)$ αν $E_{\theta}[t(\mathbf{X})] = \tau(\theta)$.

Θα λέμε ότι το στατιστικό $t(\mathbf{X})$ είναι ένας αμερόληπτος εκτιμητής ελάχιστης διακύμανσης (α.ε.ε.δ.) αν ο $t(\mathbf{X})$ είναι αμερόληπτος και για κάθε άλλο αμερόληπτο εκτιμητή $t^*(\mathbf{X})$ του $\tau(\theta)$ ισχύει

$$\text{Var}(t_i) \leq \text{Var}(t_i^*)$$

όπου t_i και t_i^* είναι, αντίστοιχα, οι συνιστώσες των t και t^* .

Από την μονοδιάστατη περίπτωση είναι γνωστό ότι η εύρεση αμερόληπτου εκτιμητή ελάχιστης διασποράς γίνεται με τη βοήθεια των θεωρημάτων των Cramér - Rao και Lehmann - Scheffé. Στη συνέχεια δίνουμε τα θεωρήματα αυτά (χωρίς αποδείξεις) στην πολυδιάστατη μορφή τους.

Θεώρημα 3.4 (Cramér - Rao) Έστω $\mathbf{X} = (x_1, x_2, \dots, x_n)'$ ένα τυχαίο δείγμα από την κατανομή $f(\mathbf{x}; \theta)$. Αν $t(\mathbf{X})$ είναι ένας αμερόληπτος εκτιμητής του $\tau(\theta)$, τότε ο $t(\mathbf{X})$ είναι ελάχιστης διακύμανσης εάν

$$\text{Var}[t(\mathbf{X})] \geq \Delta F^{-1} \Delta',$$

όπου $\Delta = \left[\frac{\partial \tau_i(\theta)}{\partial \theta_j} \right]$ και τ_i & θ_j είναι, αντίστοιχα, οι συνιστώσες των τ και θ .

Απόδειξη. Βλέπε Giri (1977).



Θεώρημα 3.5 (Lehmann - Scheffé) 'Εστω $X = (x_1, x_2, \dots, x_n)'$ ένα τυχαίο δείγμα από την κατανομή $f(x; \theta)$ και $t_1(X)$ ένα πλήρες και επαρκές στατιστικό. Αν $t(t_1)$ είναι ένας αμερόληπτος εκτιμητής του $\tau(\theta)$, τότε ο t είναι αμερόληπτος ελάχιστης διασποράς.

Απόδειξη. Βλέπε Giri (1977).

3.2 ΕΚΤΙΜΗΤΕΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

3.2.1 Γενικά

Ο εκτιμητής μέγιστης πιθανοφάνειας (ε.μ.π.) μιας άγνωστης παραμέτρου είναι εκείνη η τιμή της παραμέτρου η οποία μεγιστοποιεί την συνάρτηση πιθανοφάνειας των παρατηρήσεων.

Κατά κανόνα η εύρεση αυτού του μέγιστου γίνεται με παραγωγή της $L(X; \theta)$. Επειδή όμως η $l(X; \theta)$ μεγιστοποιείται όταν η $L(X; \theta)$ μεγιστοποιείται, γι' αυτό η εύρεση του εκτιμητή μέγιστης πιθανοφάνειας γίνεται με την λύση της εξίσωσης

$$\frac{\partial}{\partial \theta} l(X; \theta) = s = 0. \quad (3.6)$$

Από την (3.6), επειδή η s είναι συνάρτηση ενός επαρκούς στατιστικού, συνεπάγεται ότι οι εκτιμητές μέγιστης πιθανοφάνειας είναι συναρτήσεις κάποιου επαρκούς στατιστικού.

Για τους εκτιμητές μέγιστης πιθανοφάνειας ισχύουν, μεταξύ άλλων, οι ιδιότητες.

i) Ασυμπτωτικώς έχουν την πολυδιάστατη κανονική κατανομή με μέση τιμή $\mu = \theta$ και πίνακα διακυμάνσεων - συνδιακυμάνσεων

$\Sigma = (nF)^{-1}$ όπου n είναι το μέγεθος του δείγματος και F ο πίνακας πληροφορίας του Fisher.



ii) Αν $\hat{\theta}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του θ , τότε $\tau(\hat{\theta})$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας του $\tau(\theta)$.

Παράδειγμα 3.5 Για το Παράδειγμα 3.3 ο ε. μ. π. $\hat{\theta}$ του θ θα προκύψει από την λύση της εξίσωσης $s = 0$, η οποία στην περίπτωση μας είναι η

$$n\theta^2 + (2x_2 + 2x_3 - x_1 + x_4)\theta - 2x_4 = 0,$$

ή, θέτοντας $p_i = x_i/n$ ($i = 1, 2, 3, 4$),

$$\theta^2 + (2 - 3p_1 - p_4)\theta - 2p_4 = 0.$$

Η εξίσωση αυτή έχει δύο ρίζες ως προς θ , μία θετική και μία αρνητική. Απορρίπτοντας την αρνητική ρίζα βρίσκουμε ότι

$$\hat{\theta} = \lambda + (\lambda^2 + 2p_4)^{1/2} \quad \text{με} \quad \lambda = \frac{1}{2}(3p_1 + p_4) - 1.$$

3.2.2 Πολυδιάστατη κανονική κατανομή

Αν $x_1, x_2, \dots, x_n$ είναι ένα τυχαίο δείγμα από την $N_p(\mu, \Sigma)$ με $n \geq p+1$, τότε από την (3.4) θα έχουμε

$$l(X; \mu, \Sigma) = -\frac{n}{2} \log_e [\det(2\pi\Sigma)] - \frac{n}{2} \text{tr}(\Sigma^{-1}S) - \frac{n}{2} (\bar{x} - \mu)' \Sigma^{-1} (\bar{x} - \mu).$$

Θέτοντας $V = \Sigma^{-1}$ η προηγούμενη σχέση γράφεται

$$l(X; \mu, \Sigma) = -\frac{n}{2} \log_e (2\pi) + \frac{n}{2} \log_e [\det(V)] - \frac{n}{2} \text{tr}(VS) - \frac{n}{2} \text{tr} \left[V(\bar{x} - \mu)(\bar{x} - \mu)' \right]. \quad (3.7)$$

Οι εκτιμητές μέγιστης πιθανοφάνειας, $\hat{\mu}$ και $\hat{V}$, των μ και V , αντίστοιχα, θα προκύψουν από τη λύση του συστήματος

$$\frac{\partial}{\partial \mu} l(X; \mu, V) = 0 \quad \& \quad \frac{\partial}{\partial V} l(X; \mu, V) = 0$$



Αλλά (βλ. Παράρτημα) $\frac{\partial}{\partial \mu} l(X; \mu, V) = nV(\bar{x} - \mu)$.

Για να βρούμε την $\partial l(X; \mu, V) / \partial V$ θα υπολογίσουμε πρώτα την μερική μερική παράγωγο για κάθε όρο στα δεξιά της ισότητας (3.2). Έτσι από το Παράρτημα θα έχουμε

$$\frac{\partial}{\partial v_{ij}} \log_e [\det(V)] = \begin{cases} 2v_{ij} / \det(V) & , i \neq j \\ v_{ij} / \det(V) & , i = j , \end{cases}$$

όπου v_{ij} είναι ο ij συμπαράγων του V . Επειδή ο V είναι συμμετρικός, ο πίνακας με στοιχεία τα $v_{ij} / \det(V)$ είναι ο $V^{-1} = \Sigma$. Συνεπώς

$$\frac{\partial \log_e [\det(V)]}{\partial V} = 2\Sigma - \text{diag}(\Sigma) .$$

Ακόμη (βλ. Παράρτημα)

$$\frac{\partial \text{tr}(VS)}{\partial V} = 2S - \text{diag}(S) ,$$

και

$$\frac{\partial \text{tr}[V(\bar{x} - \mu)(\bar{x} - \mu)']}{\partial V} = 2(\bar{x} - \mu)(\bar{x} - \mu)' - \text{diag}[(\bar{x} - \mu)(\bar{x} - \mu)'] .$$

Συνδυάζοντας τις τρεις τελευταίες εξισώσεις παίρνουμε

$$\frac{\partial}{\partial V} l(X; \mu, V) = \frac{n}{2} [2M - \text{diag}(M)] ,$$

όπου $M = \Sigma - S - (\bar{x} - \mu)(\bar{x} - \mu)'$.

Από την εξίσωση $\partial l(X; \mu, V) / \partial \mu = 0$ παίρνουμε ότι $V(\bar{x} - \mu) = 0$ και συνεπώς ο εκτιμητής μέγιστης πιθανοφάνειας του μ θα είναι ο $\hat{\mu} = \bar{x}$. Από την $\partial l(X; \mu, V) / \partial V = 0$ παίρνουμε $2M - \text{diag}(M) = 0$ το οποίο συνεπάγεται ότι $M = 0$. Δηλαδή ο εκτιμητής μέγιστης πιθανοφάνειας του Σ θα είναι ο

$\hat{\Sigma} = S + (\bar{x} - \mu)(\bar{x} - \mu)'$. Από την τελευταία σχέση και επειδή $\hat{\mu} = \bar{x}$



παιρνουμε ότι $\hat{\Sigma} = S$.

Από αυστηρή μαθηματική σκοπιά το προηγούμενο αποτέλεσμα δεν μας λέει τίποτε παραπάνω από το ότι το σημείο $(\hat{\mu}, \hat{\Sigma})$ είναι ένα σημείο στάσεως για την $L(X; \mu, \Sigma)$. Για να δείξουμε ότι το σημείο στάσεως πράγματι μεγιστοποιεί την $L(X; \mu, \Sigma)$ χρειαζόμαστε το επόμενο Λήμμα το οποίο δίνουμε χωρίς απόδειξη . (Για μια απόδειξη βλέπε Mardia , Kent and Bibby (1979)) .

Λήμμα 3.1 Για κάθε σταθερό $p \times p$ πίνακα A η συνάρτηση

$$f(\Sigma) = [\det(\Sigma)]^{-n/2} \exp\left[-\frac{1}{2} \text{tr}(\Sigma^{-1}A)\right]$$

μεγιστοποιείται, ως προς $\Sigma > 0$, για $\Sigma = n^{-1}A$ και η μέγιστη τιμή της είναι

$$f(n^{-1}A) = [\det(n^{-1}A)]^{-n/2} \cdot e^{-np/2} .$$



Α Σ Κ Η Σ Ε Ι Σ

3.1 Έστω $x_1, x_2, \dots, x_n$ ένα τυχαίο δείγμα από την $N_p(\mu, I)$. Να βρεθεί ένας αμερόληπτος εκτιμητής ελάχιστης διακύμανσης για την Παράμετρο $\vartheta = \mu' \mu + 1' \mu$.

3.2 Για την πολυδιάστατη κανονική κατανομή να βρεθεί ο εκτιμητής μέγιστης πιθανοφάνειας του μ χωρίς παραγωγή.

3.3 Έστω ότι το τυχαίο διάνυσμα x ακολουθεί την γενική εκθετική οικογένεια κατανομών με συνάρτηση πυκνότητας πιθανότητας

$$f(x; \vartheta) = \left\{ \alpha_0(\vartheta) + b_0(x) + \sum_{i=1}^q \alpha_i(\vartheta) b_i(x) \right\}.$$

i) Να δειχθεί ότι για την σκορ συνάρτηση ισχύει

$$s = \frac{\partial \alpha_0}{\partial \vartheta} + A b,$$

όπου η i -στη στήλη του πίνακα A είναι $\partial \alpha_i / \partial \vartheta$ και το διάνυσμα b' είναι το $(b_1(x), \dots, b_q(x))$.

ii) Αν ο A είναι μη-ιδιάζων πίνακας να δειχθεί ότι

$$E(b) = - A^{-1} \frac{\partial \alpha_0}{\partial \vartheta}.$$

3.4 Να δειχθεί ότι οι εκτιμητές μέγιστης πιθανοφάνειας μεγιστοποιούν την συνάρτηση πιθανοφάνειας.



4. ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

4.1 ΓΕΝΙΚΑ

Στο Κεφάλαιο αυτό θα γενικεύσουμε γνωστούς, από την μονοδιάστατη περίπτωση, ελέγχους υποθέσεων για τη μέση τιμή και την διακύμανση, σε ελέγχους για πολυδιάστατες μέσες τιμές και πίνακες διακυμάνσεων-συνδιακυμάνσεων .

Ο έλεγχος υποθέσεων στην πολυδιάστατη περίπτωση παρουσιάζει πιο πολλές δυσκολίες απ' ότι ο έλεγχος υποθέσεων στην μονοδιάστατη περίπτωση .

Δύο είναι τα βασικά προβλήματα στον έλεγχο υποθέσεων στην πολυδιάστατη περίπτωση : i) ο μεγάλος αριθμός των παραμέτρων και ii) η δυσκολία στην εκλογή ενός κατάλληλου στατιστικού.

Ο μεγάλος αριθμός των παραμέτρων έχει σαν αποτέλεσμα την δυνατότητα διατύπωσης πολλών υποθέσεων προς έλεγχο. Για την p -διάστατη κανονική κατανομή π.χ έχουμε $[p(p+3)/2]$ παραμέτρους και συνεπώς τουλάχιστον $2^{p(p+3)/2}$ υποθέσεις οι οποίες ορίζουν διάφορες σχέσεις για ένα υποσύνολο από αυτές τις παραμέτρους. Ο αριθμός αυτός είναι αρκετά μεγάλος ακόμη και για $p = 3$. Φυσικά



στην πράξη ο αριθμός των υποθέσεων που μας ενδιαφέρει είναι αρκετά μικρότερος.

Το δεύτερο πρόβλημα, όπως αναφέραμε και προηγουμένως, είναι η δυσκολία εκλογής ενός στατιστικού. Δηλαδή θα χρησιμοποιήσουμε μια σειρά από μονοδιάστατα τεστ ή κάποιο παλυδιάστατο τεστ θα ήταν καλύτερο; Και αν είμαστε στην πρώτη περίπτωση, ποιο από τα πολλά πιθανώς τεστ που θα υπάρχουν θα χρησιμοποιήσουμε; Θα προσπαθήσουμε να εξηγήσουμε τα προβλήματα αυτά με ένα παράδειγμα.

Παράδειγμα 4.1 Ας θεωρήσουμε τα δεδομένα του πίνακα 4.1.

Πίνακας 4.1

Μετρήσεις (σε mm) του μήκους και του πλάτους της κεφαλής για το πρώτο και δεύτερο ενήλικο αγόρι σε 25 οικογένειες.

Α'	Μήκος	191	195	181	183	176	208	189	197	188	192	179	183
Α	Κεφαλής (X_1)	174	190	188	163	195	186	181	175	192	174	176	197
δ		190											
ο													
ρ	Πλάτος	155	149	148	153	144	157	150	159	152	150	158	147
ι	Κεφαλής (X_2)	150	159	151	137	155	153	145	140	154	143	139	167
		163											
Β'	Μήκος	179	201	185	188	171	192	190	189	197	187	186	174
Α	Κεφαλής (X_3)	185	195	187	161	183	173	182	165	185	178	176	200
δ		187											
ο													
ρ	Πλάτος	145	152	149	149	142	152	149	152	159	151	148	147
ι	Κεφαλής (X_4)	152	157	158	130	158	148	146	137	152	147	143	158
		150											

Με βάση τα δεδομένα αυτά βρίσκουμε ότι :

$$\bar{\bar{X}} = \begin{bmatrix} 185,72 \\ 151,12 \\ 183,84 \\ 149,24 \end{bmatrix} \quad \text{και} \quad S = \begin{bmatrix} 91,481 & 50,753 & 66,875 & 44,267 \\ & 52,186 & 49,259 & 33,651 \\ & & 96,775 & 54,278 \\ & & & 43,222 \end{bmatrix}$$



Για να γίνει αντιληπτή η διαφορά μεταξύ πολυδιάστατων και μονοδιάστατων τεστ ως θεωρήσουμε μόνο τις τ. μ. X_1 και X_3 . Ας υποθέσουμε, αρχικά, ότι οι τ. μ. X_1 και X_3 είναι ανεξάρτητες και ότι η κάθε μία ακολουθεί μία κανονική κατανομή με μέσες τιμές μ_1 και μ_3 , αντίστοιχα, και κοινή τυπική απόκλιση $\sigma = 10$. (Η υπόθεση της ανεξαρτησίας των τ. μ. X_1 και X_3 στην πραγματικότητα δεν ισχύει. Γίνεται μόνο και μόνο για παιδαγωγικούς σκοπούς). Στην περίπτωση αυτή ο έλεγχος των υποθέσεων

$$H_0^1: \mu_1 = 182 \quad \text{και} \quad H_0^3: \mu_3 = 182$$

γίνεται με τα στατιστικά

$$Z_1 = \frac{\bar{x}_1 - 182}{10/\sqrt{25}} = 1,86 \quad \text{και} \quad Z_3 = \frac{\bar{x}_3 - 182}{10/\sqrt{25}} = 0,92.$$

Επειδή τώρα $|Z_i| < 1,96$ ($i = 1, 2$) συμπεραίνουμε ότι με επίπεδο σημαντικότητας $\alpha = 5\%$ καμία από τις H_0^1 και H_0^3 δεν μπορεί να απορριφθεί.

Ας θεωρήσουμε τώρα προς έλεγχο την διδιάστατη υπόθεση

$$H_0: \mu_1 = 182 \text{ \& } \mu_3 = 182.$$

Για τον έλεγχο της H_0 υπάρχουν πολλοί τρόποι. Κατ' αρχήν η H_0 δεν θα μπορεί να απορριφθεί εάν τα προηγούμενα στατιστικά δεν απορρίπτουν τις υποθέσεις H_0^1 και H_0^3 . Ένας άλλος τρόπος για τον έλεγχο της H_0 είναι ο εξής: εάν τόσο η H_0^1 όσο και η H_0^3 ισχύουν, τότε

$$Z = \frac{1}{\sqrt{2}} (Z_1 + Z_3) \sim N(0, 1).$$

Η περιοχή αποδοχής της H_0 , για το στατιστικό Z θα είναι το ορθογώνιο που περικλείεται μεταξύ των ευθειών $Z_1 + Z_3 = \pm c$. Παρατηρήστε ότι η τιμή του Z είναι 1,966, δηλαδή μόλις και απορρίπτεται η H_0 , για $\alpha = 5\%$. Αυτό όμως έρχεται σε αντίθεση με το προηγούμενο συμπέρασμα.



Ένας τρίτος τρόπος για τον έλεγχο της H_0 προκύπτει από το γεγονός ότι

$$Z_1^2 + Z_3^2 \sim \chi_2^2 .$$

Στην περίπτωση αυτή η περιοχή αποδοχής της H_0 είναι ένας κύκλος. Για το παράδειγμά μας $Z_1^2 + Z_3^2 = 4,306$ και η H_0 δεν μπορεί να απορριφθεί .

Από τα προηγούμενα γίνεται φανερό ότι από την εκλογή του στατιστικού εξαρτάται και η περιοχή αποδοχής . Η περιοχή αποδοχής μπορεί να είναι π.χ. είτε γραμμική, είτε ερθογώνια, είτε κυκλική .

Στην συνέχεια θα παρουσιάσουμε δύο τεχνικές για την εύρεση στατιστικών για τον έλεγχο πολυδιάστατων υποθέσεων . Η μία είναι η τεχνική του Πηλίκου Μέγιστης Πιθανοφάνειας (Π.Μ.Π) και η άλλη είναι η τεχνική της Ένωσης - Τομής (E-T).

Θα πρέπει να σημειωθεί ότι για τον έλεγχο κάποιων υποθέσεων και οι δύο τεχνικές οδηγούν στο ίδιο στατιστικό . Για κάποιες άλλες υποθέσεις οι δύο τεχνικές οδηγούν σε διαφορετικά στατιστικά . Τέλος υπάρχουν υποθέσεις για τις οποίες η τεχνική της Ένωσης - Τομής δεν μπορεί να εφαρμοσθεί.



4.2 ΜΕΘΟΔΟΣ ΜΕΓΙΣΤΗΣ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Η τεχνική αυτή, η οποία είναι γνωστή από την μονοδιάστατη περίπτωση, συνίσταται στο να μεγιστοποιήσουμε την συνάρτηση πιθανοφάνειας τόσο όταν η μηδενική υπόθεση ισχύει, όσο και όταν η εναλλακτική υπόθεση ισχύει. Τα κύρια σημεία αυτής της τεχνικής δίνονται στους ορισμούς και το θεώρημα που ακολουθούν.

Ορισμός 4.1 Έστω $X = (x_1, x_2, \dots, x_n)'$ ένα τυχαίο δείγμα από ένα πληθυσμό με κατανομή $f(x; \theta)$. Αν $H_0: \theta \in \Omega_0 \subseteq \mathbb{R}^q$ και $H_a: \theta \in \Omega_a \subseteq \mathbb{R}^q$ ($r < q$) είναι δύο οποιοσδήποτε υποθέσεις, τότε το στατιστικό του Π.Μ.Π. για τον έλεγχο της H_0 ως προς την H_a είναι το

$$\lambda(x) = \frac{\max_{\theta \in \Omega_0} L(X; \theta)}{\max_{\theta \in \Omega_a} L(X; \theta)} \quad (4.1)$$

Αντί για το στατιστικό (4.1) μπορούμε να χρησιμοποιήσουμε το ισοδύναμο του στατιστικό

$$-2 \log_e \lambda(x) = 2(l_a - l_0), \quad (4.2)$$

όπου $l_0 = \max_{\theta \in \Omega_0} L(X; \theta)$ και $l_a = \max_{\theta \in \Omega_a} L(X; \theta)$.

Όταν οι υποθέσεις H_0 και H_a είναι απλές, δηλαδή καθ' ένα από τα Ω_0 και Ω_a περιέχουν ένα μόνο σημείο, τότε οι βέλτιστες ιδιότητες του στατιστικού (4.1) ή (4.2) περιγράφονται από το γνωστό (βλ. Παπαιωάννου - Φερεντίνος (1983) σελ. 136) θεώρημα των Neyman - Pearson. Στην περίπτωση που οι H_0 και H_a είναι σύνθετες υποθέσεις, τότε οι βέλτιστες ιδιότητες των στατιστικών (4.1) και (4.2) δίνονται από τους Kendall & Stuart (1973).

Η χρήση του στατιστικού (4.1) έχει ως εξής: δεχόμαστε την



H_a όταν η τιμή του $\lambda(x)$ είναι μικρή ή δεχόμαστε την H_0 όταν η τιμή του $\lambda(x)$ είναι μεγάλη. Πιο συγκεκριμένα μπορούμε να δώσουμε τον επόμενο ορισμό .

Ορισμός 4.2 Για το πηλίκο μέγιστης πιθανοφάνειας η περιοχή απόρριψης της H_0 ως προς την H_a , με επίπεδο σημαντικότητας α , είναι

$$\text{Περιοχή Απόρριψης (Π. Α.)} = \{x : \lambda(x) < c\}$$

όπου η τιμή της σταθεράς c είναι τέτοια ώστε

$$\sup_{\theta \in \Omega_0} P_{\theta}(x \in \text{Π.Α.}) = \alpha .$$

Για τις υποθέσεις για τις οποίες ενδιαφερόμαστε η κατανομή του $\lambda(x)$, στην πράξη, δεν εξαρτάται από το θ έτσι ώστε το προηγούμενο *supremum* να μην είναι απαραίτητο .

Μια σπουδαία ασυμπτωτική ιδιότητα του πηλίκου μέγιστης πιθανοφάνειας μας δίνει το επόμενο θεώρημα .

Θεώρημα 4.1 Κάτω από κατάλληλες συνθήκες κανονικότητας και όταν $n \rightarrow \infty$ η κατανομή του $-2 \log_e \lambda$, όταν η H_0 ισχύει, είναι η χ^2_{q-r} - κατανομή.

Απόδειξη . Βλέπε Silvey (1970) .



4.3 ΜΕΘΟΔΟΣ ΕΝΩΣΗΣ - ΤΟΜΗΣ

Η μέθοδος αυτή οφείλεται στον Roy (1957) . Για να γίνει καλύτερα αντιληπτή θα την παρουσιάσουμε με ένα παράδειγμα . Ας θεωρήσουμε ένα τυχαίο διάνυσμα x με κατανομή την $N_p(\mu, I)$ και έστω a ένα $p \times 1$ σταθερό διάνυσμα . Αν θέσουμε $y_a = a'x$, τότε γνωρίζουμε ότι $y_a \sim N(a'\mu, a'a)$ και αυτό θα ισχύει για κάθε σταθερό $p \times 1$ διάνυσμα a .

Ας υποθέσουμε τώρα ότι θέλουμε να ελέγξουμε την υπόθεση $H_0: \mu = 0$. Όταν η H_0 ισχύει τότε $y_a \sim N(0, a'a)$. Ας συμβολί - σουμε με $H_0(a)$ την υπόθεση ότι $y_a \sim N(0, a'a)$. Προφανώς όταν η H_0 ισχύει τότε η $H_0(a)$ ισχύει για κάθε p -διάστατο διάνυσμα a . Με άλλα λόγια η πολυδιάστατη υπόθεση H_0 μπορεί να γραφεί σαν η τομή του συνόλου των μονοδιάστατων υποθέσεων $H_0(a)$. Δηλαδή

$$H_0 = \bigcap H_0(a) \quad (4.3)$$

Στην (4.3) χρησιμοποιήσαμε το σύμβολο της τομής γιατί για να αληθεύει η H_0 θα πρέπει να αληθεύουν όλες οι $H_0(a)$. Την $H_0(a)$ την ονομάζουμε *συνιστώσα* της H_0 .

Ας δούμε τώρα πως μπορούμε να ελέγξουμε τις μονοδιάστατες υποθέσεις $H_0(a)$. Προφανώς ένας τρόπος είναι η χρήση του στατιστικού $z_a = y_a / \sqrt{a'a}$ η κατανομή του οποίου είναι η τυπική κανονική . Με βάση το z_a η περιοχή απόρριψης της $H_0(a)$ είναι η $A_a = \{z_a : |z_a| > c\}$, όπου c είναι μια κάποια τιμή, έστω η 1,96 . Η προηγούμενη περιοχή απόρριψης μπορεί να γραφεί σαν

$$A_a = \{z_a : z_a^2 > c^2\} \quad (4.4)$$

Συνεπώς για κάθε μια από τις μονοδιάστατες υποθέσεις έχουμε μια περιοχή απόρριψής τους, την (4.4) .

Ας γυρίσουμε τώρα στον έλεγχο της πολυδιάστατης υπόθεσης H_0 και ας προσπαθήσουμε να βρούμε μια λογική περιοχή απόρριψής της.



Επειδή η H_0 αληθεύει εάν και μόνον εάν κάθε συνιστώσα υπόθεση $H_0(a)$ αληθεύει, φαίνεται λογικό να δεχθούμε την H_0 εάν κάθε συνιστώσα υπόθεση $H_0(a)$ γίνεται δεκτή. Με άλλα λόγια απορρίπτουμε την H_0 εάν μια οποιαδήποτε συνιστώσα υπόθεση $H_0(a)$ απορρίπτεται. Συνεπώς η περιοχή απόρριψης της H_0 θα είναι η ένωση των περιοχών απόρριψης των $H_0(a)$, δηλαδή

$$A = \bigcup_a A_a \quad (4.5)$$

Η ένωση των περιοχών απόρριψης, όπως δίνεται από την (4.5), και η τομή των συνιστωσών υποθέσεων, όπως δίνεται από την (4.3), αποτελούν την βάση αυτής της τεχνικής.

Εφαρμόζοντας την τεχνική αυτή στο παραδειγμά μας παίρνουμε ότι η H_0 απορρίπτεται εάν και μόνον εάν ένα οποιοδήποτε z_a^2 είναι μεγαλύτερο από το c^2 . Ή διαφορετικά, η H_0 γίνεται δεκτή εάν και μόνον εάν $z_a^2 \leq c^2$ για κάθε z_a . Αυτό είναι το ίδιο σαν να λέμε ότι η H_0 γίνεται δεκτή εάν και μόνον εάν

$$\max_a z_a^2 \leq c^2.$$

Κατά κανόνα η μέθοδος της Ένωσης - Τομής συχνά οδηγεί σε προβλήματα μεγιστοποίησης ή ελαχιστοποίησης κάποιων σύνθετων στατιστικών όπως π. χ. το z_a . Στην περίπτωση μας το z_a^2 μπορεί να γραφεί σαν $z_a^2 = y_a^2 / a'a = a'xx'a$. Παρατηρήστε ότι η τιμή του a δεν μπορεί να προσδιορισθεί μονοσήμαντα εκτός αν βάλουμε τον περιορισμό ότι $a'a = 1$. Τότε (βλ. Παράρτημα) το z_a^2 μεγιστοποιείται για $a = x/x'x$, οπότε $\max_a z_a^2 = x'x$. Άρα, για τη μέθοδο της Ένωσης - Τομής, το στατιστικό για τον έλεγχο της H_0 είναι το $x'x$ η κατανομή του οποίου, όταν η H_0 ισχύει, είναι η χ_p^2 .

Ο τρόπος με τον οποίο βρίσκουμε ένα στατιστικό με την μέθοδο της Ένωσης - Τομής έχει σπουδαία πρακτική εφαρμογή. Έτσι αν η H_0 απορριφθεί, τότε μπορούμε να βρούμε ποια συνιστώσα είναι υπεύθυνη για την απόρριψη αυτή. Αυτό έχει σαν αποτέλεσμα



να γνωρίζουμε καλύτερα τον τρόπο της απόκλισης από την H_0 . Δυστυχώς η ιδιότητα αυτή δεν ισχύει για την μέθοδο του πηλίκου μέγιστης πιθανοφάνειας.

4.4 ΕΦΑΡΜΟΓΕΣ

Στην παράγραφο αυτή θα δώσουμε μερικούς βασικούς πολυδιάστατους ελέγχους υποθέσεων. Οι έλεγχοι αυτοί είναι γενικεύσεις αντίστοιχων ελέγχων στην μονοδιάστατη περίπτωση. Η εύρεση των στατιστικών, για κάθε έλεγχο υπόθεσης, θα γίνει τόσο με την μέθοδο του πηλίκου μέγιστης πιθανοφάνειας όσο και με τη μέθοδο της Ένωσης - Τομής.

4.4.1 Έλεγχος μέσω τιμών

4.4.1.1 Έλεγχος για τη μέση τιμή ενός κανονικού πληθυσμού

Έστω $X = (x_1, x_2, \dots, x_n)'$ ένα τυχαίο δείγμα από την κανονική κατανομή $N_p(\mu, \Sigma)$ και ας υποθέσουμε ότι θέλουμε να ελέγξουμε την υπόθεση $H_0: \mu = \mu_0$, όπου μ_0 είναι ένα γνωστό διάνυσμα. Διακρίνουμε δύο περιπτώσεις: Σ γνωστό και Σ άγνωστο.

I) $H_0: \mu = \mu_0$, $H_a: \mu \neq \mu_0$ με Σ γνωστό.

Πηλίκο Μέγιστης Πιθανοφάνειας: Όταν το Σ είναι γνωστό τότε η H_0 είναι μια απλή υπόθεση. Από την (3.4), όταν η H_0 ισχύει, παίρνουμε.

$$l_0 = l(X; \mu_0, \Sigma) = -\frac{n}{2} \log_e [\det(2\pi\Sigma)] - \frac{n}{2} \text{tr}(\Sigma^{-1}S) - \frac{n}{2} (\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0).$$



Επειδή η $H_a: \mu \neq \mu_0$ δεν θέτει κανένα περιορισμό για το μ , γι' αυτό ο εκτιμητής μέγιστης πιθανοφάνειας του μ , όταν η H_a ισχύει, θα είναι το $\bar{\mathbf{x}}$. Συνεπώς

$$l_1 = l(\mathbf{X}; \bar{\mathbf{x}}, \Sigma) = -\frac{n}{2} \log_e \left[\det(2\pi\Sigma) \right] - \frac{n}{2} \text{tr}(\Sigma^{-1}\mathbf{S}).$$

Εφαρμόζοντας την (4.2) παίρνουμε

$$-2\log_e \lambda = 2(l_1 - l_0) = n(\bar{\mathbf{x}} - \mu_0)' \Sigma^{-1}(\bar{\mathbf{x}} - \mu_0).$$

Από το θεώρημα 2.10 έχουμε ότι η ακριβής κατανομή του

$n(\bar{\mathbf{x}} - \mu_0)' \Sigma^{-1}(\bar{\mathbf{x}} - \mu_0)$ είναι η χ_p^2 . Συνεπώς η H_0 απορρίπτεται εάν

$$n(\bar{\mathbf{x}} - \mu_0)' \Sigma^{-1}(\bar{\mathbf{x}} - \mu_0) \geq \chi_{\alpha; p}^2.$$

Ένωση - Τομή : Θέτοντας $\mathbf{y} = \mathbf{X}\mathbf{a}$ έχουμε ότι οι συνιστώσες του $\mathbf{y}$ είναι ανεξάρτητες με κατανομή $N(\mathbf{a}'\mu, \mathbf{a}'\Sigma\mathbf{a})$. Όταν η H_0 ισχύει τότε η κατανομή κάθε συνιστώσας του $\mathbf{y}$ θα είναι η $N(\mathbf{a}'\mu_0, \mathbf{a}'\Sigma\mathbf{a})$. Ο έλεγχος της μονοδιάστατης υπόθεσης $H_0(\mathbf{a}): \mathbf{a}'\mu = \mathbf{a}'\mu_0$ γίνεται με στατιστικό

$$z_{\mathbf{a}} = \sqrt{n}(\bar{\mathbf{y}} - \mathbf{a}'\mu_0) / \sqrt{\mathbf{a}'\Sigma\mathbf{a}}.$$

Από το στατιστικό αυτό προκύπτει, επειδή $\bar{\mathbf{y}} = \mathbf{a}'\bar{\mathbf{x}}$, ότι

$$z_{\mathbf{a}}^2 = n\mathbf{a}'(\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \mu_0)'\mathbf{a} / \mathbf{a}'\Sigma\mathbf{a}.$$

Σύμφωνα τώρα με τα όσα έχουμε προαναφέρει για τη μέθοδο αυτή, το στατιστικό για τον έλεγχο της $H_0: \mu = \mu_0$, Σ γνωστό, θα προκύψει από την (4.5), δηλαδή από την μεγιστοποίηση του $z_{\mathbf{a}}^2$ ως προς $\mathbf{a}$. Για να βρούμε την τιμή του $\mathbf{a}$ για την οποία το $z_{\mathbf{a}}^2$ μεγιστοποιείται θα πρέπει να θέσουμε τον περιορισμό $\mathbf{a}'\Sigma\mathbf{a} = 1$. Στην περίπτωση αυτή (βλ. Παράρτημα) το $z_{\mathbf{a}}^2$ μεγιστοποιείται για

$$\mathbf{a} = \Sigma^{-1}(\bar{\mathbf{x}} - \mu_0) / \left[(\bar{\mathbf{x}} - \mu_0)' \Sigma^{-1}(\bar{\mathbf{x}} - \mu_0) \right],$$

οπότε

$$\max_{\mathbf{a}} z_{\mathbf{a}}^2 = n(\bar{\mathbf{x}} - \mu_0)' \Sigma^{-1}(\bar{\mathbf{x}} - \mu_0).$$



Η κατανομή του στατιστικού αυτού είναι, όπως αναφέραμε και στη μέθοδο του Πηλίκου Μέγιστης Πιθανοφάνειας, η χ^2_p κατανομή.

Στην περίπτωση αυτή τόσο η μέθοδος του Πηλίκου Μέγιστης Πιθανοφάνειας όσο και η μέθοδος της Ένωσης - Τομής οδηγούν στο ίδιο στατιστικό. Αυτό δεν συμβαίνει πάντοτε όπως θα δούμε στην συνέχεια.

Παράδειγμα 4.2 Ας θεωρήσουμε τις μεταβλητές X_1 και X_3 από το Παράδειγμα 4.1 και ας υποθέσουμε ότι $(X_1, X_3) \sim N_2(\mu, \Sigma)$. Τότε για τον έλεγχο της $H_0: \mu = \mu_0$, Σ γνωστό, όπου

$$\mu_0 = \begin{bmatrix} 182 \\ 182 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix},$$

το στατιστικό $n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0)$ μας δίνει

$$25(3,721,84) \begin{bmatrix} 0,01 & 0 \\ 0 & 0,01 \end{bmatrix} \begin{pmatrix} 3,72 \\ 1,84 \end{pmatrix} = 4,31.$$

Συγκρίνοντας την τιμή αυτή με την τιμή της $\chi^2_{0,05;2} = 5,99$ βλέπουμε ότι η H_0 δεν μπορεί να απορριφθεί.

II) $H_0: \mu = \mu_0$, $H_a: \mu \neq \mu_0$ με Σ άγνωστο.

Πηλίκου Μέγιστης Πιθανοφάνειας : Στην περίπτωση αυτή ο πίνακας διακυμάνσεων-συνδιακυμάνσεων Σ θα πρέπει να εκτιμηθεί τόσο όταν η H_0 ισχύει, όσο και όταν η H_a ισχύει.

Από την § 3.2.2 παίρνουμε ότι οι εκτιμητές μέγιστης πιθανοφάνειας των μ και Σ όταν η H_0 ισχύει είναι :

$$\hat{\mu} = \mu_0 \quad \text{και} \quad \hat{\Sigma} = S + dd',$$

όπου $d = \bar{x} - \mu_0$ και S είναι ο δειγματικός πίνακας διακυμάνσεων-συνδιακυμάνσεων.

Όταν η H_a ισχύει τότε οι εκτιμητές μέγιστης πιθανοφάνειας των μ και Σ είναι :

$$\hat{\mu} = \bar{x} \quad \text{και} \quad \hat{\Sigma} = S.$$



Από την σχέση (3.4) θα έχουμε ότι

$$l_0 = l(\mathbf{X}; \mu_0, \mathbf{S} + \mathbf{d}\mathbf{d}') = -\frac{n}{2} \left\{ p \log_e(2\pi) + \log_e [\det(\mathbf{S})] + \log_e (1 + \mathbf{d}' \mathbf{S}^{-1} \mathbf{d}) + p \right\}.$$

Θέτοντας στην τελευταία σχέση $\mathbf{d} = \mathbf{0}$ παίρνουμε

$$l_1 = l(\mathbf{X}; \bar{\mathbf{x}}, \mathbf{S}) = -\frac{n}{2} \left\{ p \log_e(2\pi) + \log_e [\det(\mathbf{S})] + p \right\}.$$

Άρα $-2 \log_e \lambda = 2(l_1 - l_0) = n \log_e (1 + \mathbf{d}' \mathbf{S}^{-1} \mathbf{d}).$

Το τελευταίο στατιστικό εξαρτάται από την ποσότητα $(n-1)\mathbf{d}' \mathbf{S}^{-1} \mathbf{d}$ η κατανομή της οποίας (βλ. Πρόγραμμα 2.19.1) είναι η $T^2(p, n-1)$. Η κατανομή αυτή είναι γνωστή σαν η κατανομή του Hotelling για ένα πληθυσμό. Επιπλέον, από το θεώρημα 2.20, η ποσότητα $\{(n-p)/p\} \mathbf{d}' \mathbf{S}^{-1} \mathbf{d}$ έχει την $F_{p, n-p}$ -κατανομή.

Ένωση - Τομή : Θέτοντας $\mathbf{y} = \mathbf{X}\mathbf{a}$ έχουμε ότι οι συνιστώσες του $\mathbf{y}$ είναι ανεξάρτητες και κάθε μία έχει κατανομή την $N(\mu_y, \sigma_y^2)$ όπου $\mu_y = \mathbf{a}'\boldsymbol{\mu}$ και $\sigma_y^2 = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$.

Όταν η H_0 ισχύει τότε η κατανομή των συνιστωσών του $\mathbf{y}$ είναι η $N(\mathbf{a}'\boldsymbol{\mu}_0, \sigma_y^2)$. Το σ_y^2 όμως είναι άγνωστο και θα πρέπει να εκτιμηθεί από το δείγμα. Ένας εκτιμητής του σ_y^2 είναι

$$s_y^2 = \frac{1}{n} \sum (y_i - \bar{y})^2 = \frac{1}{n} \sum (\mathbf{x}_i' \mathbf{a} - \bar{\mathbf{x}}' \mathbf{a})^2 = \mathbf{a}' \mathbf{S} \mathbf{a}.$$

Ο έλεγχος της μονοδιάστατης υπόθεσης $H_0(\mathbf{a}): \mu_i = \mu_{0i}, \sigma_y^2$ άγνωστο γίνεται με το στατιστικό

$$t(\mathbf{a}) = \frac{\bar{y} - \mu_y}{\sqrt{s_y^2 / (n-1)}}.$$



Συνεπώς το

$$t^2(\mathbf{a}) = (n-1) \frac{\mathbf{a}'(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{a}}{\mathbf{a}'\mathbf{S}\mathbf{a}}$$

είναι το πηλίκο δύο τετραγωνικών μορφών και το στατιστικό για τον έλεγχο της H_0 θα προκύψει από την μεγιστοποίηση του $t^2(\mathbf{a})$, ως προς $\mathbf{a}$, με τον περιορισμό ότι $\mathbf{a}'\mathbf{S}\mathbf{a} = 1$. Το $t^2(\mathbf{a})$ μεγιστοποιείται (βλ. Παράρτημα) για

$$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) / [(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)]$$

και συνεπώς

$$\max_{\mathbf{a}} t^2(\mathbf{a}) = (n-1)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0). \quad (4.6)$$

Βλέπουμε δηλαδή ότι και σε αυτή την περίπτωση οι δύο μέθοδοι δίνουν το ίδιο στατιστικό.

Παράδειγμα 4.3 Θα χρησιμοποιήσουμε ξανά τις μεταβλητές X_1 και X_3 από το Παράδειγμα 4.1. Ας υποθέσουμε ότι η προς έλεγχο υπόθεση είναι η $H_0: \mu = \mu_0$, $H_a: \mu \neq \mu_0$, Σ άγνωστο, όπου $\mu_0 = (182 \ 182)'$. Το στατιστικό για τον έλεγχο της H_0 είναι το $\{(n-p)/p\} \mathbf{d}'\mathbf{S}^{-1}\mathbf{d}$ η αριθμητική τιμή του οποίου είναι

$$\frac{n-p}{p} \mathbf{d}'\mathbf{S}^{-1}\mathbf{d} = \frac{23}{3} (3,72 \ 1,84) \begin{bmatrix} 91,481 & 66,875 \\ 66,875 & 96,775 \end{bmatrix} \begin{pmatrix} 3,72 \\ 1,84 \end{pmatrix} = 1,28.$$

Συγκρίνοντας την τιμή αυτή με την τιμή, που βρίσκουμε από τους πίνακες, της $F_{0,05;2,23} = 3,44$ συμπεραίνουμε ότι η H_0 δεν μπορεί να απορριφθεί.



4.4.1.2 Έλεγχος για τη μέση τιμή δύο κανονικών πληθυσμών

Έστω ότι $X = (x_1, x_2, \dots, x_n)'$ και $Y = (y_1, y_2, \dots, y_m)'$ είναι δύο ανεξάρτητα τυχαία δείγματα από τις κανονικές κατανομές $N_p(\mu_x, \Sigma_x)$ και $N_p(\mu_y, \Sigma_y)$, αντίστοιχα, και ας υποθέσουμε ότι θέλουμε να ελέγξουμε την υπόθεση $H_0: \mu_x = \mu_y$. Θα διακρίνουμε τις επόμενες δύο περιπτώσεις.

I) $H_0: \mu_x = \mu_y$, $H_a: \mu_x \neq \mu_y$, με Σ_x και Σ_y γνωστά.

Αν $n = m$ τότε θέτοντας $\mu_x - \mu_y = \delta$ και από το γεγονός ότι $\bar{x} - \bar{y} \sim N_p(\delta, \Sigma)$ με $\Sigma = n^{-1}\Sigma_x + m^{-1}\Sigma_y$ βλέπουμε ότι το πρόβλημά μας ανάγεται σε πρόβλημα ελέγχου υποθέσεων για τη μέση τιμή ενός πληθυσμού (βλ. § 4.4.1.1 (II)).

Στην περίπτωση που $n \neq m$ μια μέθοδος ελέγχου της αρχικής υπόθεσης δίνεται από τον Anderson (1984).

II) $H_0: \mu_x = \mu_y$, $H_a: \mu_x \neq \mu_y$, Σ_x , Σ_y άγνωστα, με $\Sigma_x = \Sigma_y$

Πηλίκο Μέγιστης Πιθανοφάνειας : Έστω Σ η κοινή τιμή των πινάκων Σ_x και Σ_y . Τότε ο εκτιμητής μέγιστης πιθανοφάνειας του Σ θα είναι ο $S = (nS_x + mS_y)/(n+m)$, όπου S_x και S_y είναι οι εκτιμητές μέγιστης πιθανοφάνειας των Σ_x και Σ_y , αντίστοιχα.

Όταν η H_0 ισχύει, τότε ο εκτιμητής μέγιστης πιθανοφάνειας της κοινής τιμής μ , των μ_x και μ_y , είναι ο $\hat{\mu} = (n\bar{x} + m\bar{y})/(n+m)$.

Όταν η H_a ισχύει τότε οι εκτιμητές μέγιστης πιθανοφάνειας των μ_x και μ_y είναι, αντίστοιχα, οι $\hat{\mu}_x = \bar{x}$ και $\hat{\mu}_y = \bar{y}$.

Συνεπώς από τις (3.4) και (4.2) θα έχουμε ότι



$$l_0 = l(X, Y; \hat{\mu}, S) = -\frac{n}{2} \log_e [\det(2\pi S)] - \frac{n}{2} \text{tr}(S^{-1} S_x) \\ - \frac{n}{2} (\bar{x} - \hat{\mu})' S^{-1} (\bar{x} - \hat{\mu}) \\ - \frac{m}{2} \log_e [\det(2\pi S)] - \frac{m}{2} \text{tr}(S^{-1} S_y) \\ - \frac{m}{2} (\bar{y} - \hat{\mu})' S^{-1} (\bar{y} - \hat{\mu})$$

και

$$l_1 = l(X, Y; \hat{\mu}_x, \hat{\mu}_y, S) = -\frac{n}{2} \log_e [\det(2\pi S)] - \frac{n}{2} \text{tr}(S^{-1} S_x) \\ - \frac{n}{2} (\bar{x} - \hat{\mu}_x)' S^{-1} (\bar{x} - \hat{\mu}_x) \\ - \frac{m}{2} \log_e [\det(2\pi S)] - \frac{m}{2} \text{tr}(S^{-1} S_y) \\ - \frac{m}{2} (\bar{y} - \hat{\mu}_y)' S^{-1} (\bar{y} - \hat{\mu}_y).$$

Συνεπώς

$$-2 \log_e \lambda = 2(l_1 - l_0) = n(\bar{x} - \hat{\mu})' S^{-1} (\bar{x} - \hat{\mu}) \\ + m(\bar{y} - \hat{\mu})' S^{-1} (\bar{y} - \hat{\mu})$$

ή, επειδή $\hat{\mu} = (n\bar{x} + m\bar{y})/(n+m)$,

$$-2 \log_e \lambda = 2(l_1 - l_0) = \frac{n \cdot m}{n+m} (\bar{x} - \bar{y})' S^{-1} (\bar{x} - \bar{y}).$$

Θέτοντας $S_u = [(n+m)/(n+m-2)] S$ παίρνουμε

$$-2 \log_e \lambda = \frac{n+m}{n+m-2} \cdot \frac{n \cdot m}{n+m} (\bar{x} - \bar{y})' S_u^{-1} (\bar{x} - \bar{y}).$$

Το στατιστικό αυτό είναι ανάλογο του στατιστικού του Hotelling για δύο δείγματα. (βλ. Θεώρημα 2.2.2)



Ένωση = Τομή : Αν θέσουμε $z(a) = Xa$ και $w(a) = Ya$, τότε για τις συνιστώσες του $z(a)$, $z_1, z_2, \dots, z_n$, θα έχουμε ότι, $z_i \sim N(\mu_z, \sigma_z^2)$ ($i = 1, 2, \dots, n$) ενώ για τις συνιστώσες του $w(a)$, $w_1, w_2, \dots, w_m$, θα έχουμε ότι $w_j \sim N(\mu_w, \sigma_w^2)$ ($j = 1, 2, \dots, m$), όπου $\mu_z = a' \mu_x$, $\sigma_z^2 = a' \Sigma a$, $\mu_w = a' \mu_y$ και $\sigma_w^2 = a' \Sigma a$. Έπειδή $\Sigma_x = \Sigma_y = \Sigma$ συνεπάγεται ότι $\sigma_z^2 = \sigma_w^2$. Ακόμη, ένας αμερόληπτος εκτιμητής του σ_z^2 (ή του σ_w^2) θα είναι ο $s^2 = a' S_u a$, όπου S_u είναι ο αμερόληπτος εκτιμητής του κοινού πίνακα Σ , δηλαδή

$$S_u = \frac{n \cdot S_x + m \cdot S_y}{n+m-2}.$$

Η H_0 γίνεται δεκτή όταν η $H_0(a): \mu_z = \mu_w$ γίνεται δεκτή για κάθε διάνυσμα a . Το στατιστικό για τον έλεγχο της $H_0(a)$ είναι το

$$t(a) = \frac{\bar{z} - \bar{w}}{s \sqrt{\frac{1}{n} + \frac{1}{m}}},$$

από το οποίο παίρνουμε

$$t^2(a) = \frac{(\bar{z} - \bar{w})^2}{s^2 \cdot \frac{n+m}{n \cdot m}} = \frac{n \cdot m}{n+m} \cdot \frac{a' (\bar{x} - \bar{y})(\bar{x} - \bar{y})' a}{a' S_u a}.$$

Το στατιστικό για τον έλεγχο της H_0 θα προκύψει από την μεγιστοποίηση του $t^2(a)$ ως προς a με τον περιορισμό ότι $a' S_u a = 1$. Το $t^2(a)$ είναι το πηλίκο δύο τετραγωνικών μορφών και συνεπώς (βλ. Παράρτημα) το μέγιστο θα λαμβάνεται για

$$a = \frac{S_u^{-1} (\bar{x} - \bar{y})}{(\bar{x} - \bar{y})' S_u^{-1} (\bar{x} - \bar{y})}.$$

Συνεπώς

$$\max_a t^2(a) = \frac{n \cdot m}{n+m} (\bar{x} - \bar{y})' S_u^{-1} (\bar{x} - \bar{y}). \quad (4.7)$$

Δηλαδή για μια ακόμη φορά οι δύο τεχνικές οδηγούν στο ίδιο στατιστικό.



Παρατήρηση 4.1 Το πρόβλημα του ελέγχου της υπόθεσης $H_0: \mu_x = \mu_y$ χωρίς την υπόθεση της ισότητας των αντίστοιχων πινάκων διακυμάνσεων-συνδιακυμάνσεων, είναι μια γενίκευση του γνωστού, από την μονοδιάστατη περίπτωση, προβλήματος των Behrens-Fisher. Για περισσότερα στοιχεία βλέπε Mardia, Kent & Bibby (1979).

Παράδειγμα 4.4 Σε ένα πείραμα μετρήσαμε το μήκος, το πλάτος και το ύψος του κελύφους σε 48 χελώνες. Από αυτές 24 ήταν αρσενικές και 24 θηλυκές. Τα δειγματικά διανύσματα των μέσων τιμών καθώς και οι δειγματικοί πίνακες διακυμάνσεων - συνδιακυμάνσεων, για τις αρσενικές και θηλυκές χελώνες, ήταν αντίστοιχα :

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 113,38 \\ 88,29 \\ 40,71 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 132,9975,8535,82 \\ & 47,9620,75 \\ & & 10,79 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 136,00 \\ 102,58 \\ 51,96 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 432,58259,87161,67 \\ & 164,57 & 98,99 \\ & & 63,87 \end{bmatrix}.$$

Υποθέτοντας ότι τα δύο δείγματα προέρχονται, αντίστοιχα, από τις κατανομές $N_3(\mu_1, \Sigma)$ και $N_3(\mu_2, \Sigma)$ ο έλεγχος της $H_0: \mu_1 = \mu_2$ γίνεται ως εξής : από τις τιμές των $\mathbf{S}_1$ και $\mathbf{S}_2$ παίρνουμε

$$\mathbf{S} = \frac{n \cdot \mathbf{S}_x + m \cdot \mathbf{S}_y}{n+m} = \begin{bmatrix} 282,79167,8698,75 \\ & 106,2459,87 \\ & & 37,84 \end{bmatrix}.$$

Συνεπώς

$$\mathbf{S}_u = \frac{n+m}{n+m-2} \mathbf{S} = \begin{bmatrix} 294,1174,57102,7 \\ & 110,5 & 62,26 \\ & & 39,35 \end{bmatrix}$$

και

$$\mathbf{S}_u^{-1} = \begin{bmatrix} 0,07 & -0,072 & -0,073 \\ & 0,156 & -0,058 \\ & & 0,308 \end{bmatrix}.$$



Η τιμή του στατιστικού $T^2 = [n \cdot m / (n+m)] (\bar{x}_1 - \bar{x}_2)' S_u^{-1} (\bar{x}_1 - \bar{x}_2)$ είναι $T^2 = 12 \cdot 3,292 = 39,504$. Άρα η $H_0: \mu_1 = \mu_2$ θα πρέπει να απορριφθεί αν (βλ. Θεώρημα 2.20)

$$\frac{n-2-p+1}{(n-2)p} T^2 \geq F_{\alpha; p, n-2-p+1}$$

Στην ανισότητα αυτή η τιμή του αριστερού μέλους είναι 12,59 ενώ από τους πίνακες της F- κατανομής και για $\alpha = 5\%$ βρίσκουμε ότι $F_{\alpha; n-2-p+1} = F_{0,05; 3,44} = 2,76$. Συνεπώς η H_0 θα πρέπει να απορριφθεί.

4.4.1.3 Έλεγχος ζευγών

Στην § 4.4.1.2 ασχοληθήκαμε με τον έλεγχο της ισοτιμίας των μέσων τιμών δύο κανονικών πληθυσμών, με την προϋπόθεση ότι τα δύο τυχαία δείγματα $x_1, x_2, \dots, x_n$ και $y_1, y_2, \dots, y_m$ είναι ανεξάρτητα. Πολλές φορές όμως, στην πράξη, η υπόθεση αυτή δεν ισχύει. Για παράδειγμα οι παρατηρήσεις $x_1, x_2, \dots, x_n$ μπορεί να είναι οι παρατηρήσεις σε n άτομα πριν από το γεύμα, ενώ οι παρατηρήσεις $y_1, y_2, \dots, y_m$ μπορεί να είναι οι παρατηρήσεις στα ίδια άτομα π.χ. δύο ώρες αργότερα.

Στην περίπτωση αυτή, υποθέτοντας ότι $n = m = N$, ο έλεγχος της $H_0: \mu_x = \mu_y$ γίνεται ως εξής. Θέτοντας $d_i = x_i - y_i$ και υποθέτοντας ότι $d_i \sim N_p(\delta, \Sigma)$, όπου $\delta = \mu_x - \mu_y$ ($i = 1, 2, \dots, N$), έχουμε ότι η $H_0: \mu_x = \mu_y$ είναι ισοδύναμη με την $H_0: \delta = 0$. Σύμφωνα τώρα με τα όσα αναφέραμε στην § 4.4.1.1(II) η H_0 απορρίπτεται εάν

$$\frac{n-p}{p} \bar{d}' S_d^{-1} \bar{d} \geq F_{\alpha; p, n-p},$$

όπου $\bar{d} = \bar{x} - \bar{y}$ και $S_d = N^{-1} \sum_{i=1}^N (d_i - \bar{d})(d_i - \bar{d})'$.



4.4.2 Πολλαπλές συγκρίσεις για τις μέσες τιμές

Όπως και στην περίπτωση της ανάλυσης διακύμανσης (βλ. Κ. Καρακώστας (1992)) έτσι και τώρα η απόρριψη της $H_0^1: \mu = \mu_0$ ή της $H_0^2: \mu_x = \mu_y$ δεν μας απαντάει στο ερώτημα για το ποια ή ποιες από τις συνιστώσες είναι υπεύθυνες για την απόρριψη της H_0^1 ή της H_0^2 . Στο ερώτημα αυτό μπορούμε να απαντήσουμε με την βοήθεια της μεθόδου Ένωση - Τομή που αναφέραμε στην § 4.3 .

4.4.2.1 Πολλαπλές συγκρίσεις για τη μέση τιμή ενός κανονικού πληθυσμού

Έστω $x_1, x_2, \dots, x_n$ ένα τυχαίο δείγμα από την $N_p(\mu, \Sigma)$ με Σ άγνωστο. Τότε από την σχέση (4.6) και από το Πρόγραμμα 2.19.1 παίρνουμε ότι , για όλα τα p -διάστατα διανύσματα a ,

$$t^2(a) \leq T^2 \quad (4.8)$$

$$\text{όπου} \quad t(a) = a'(\bar{x} - \mu) / [a'Sa/(n-1)]^{1/2} \quad (4.9)$$

και T^2 είναι το στατιστικό του Hotelling για τον έλεγχο υποθέσεων για τη μέση τιμή ενός κανονικού πληθυσμού. Κάνοντας χρήση του θεωρήματος 2.20 βρίσκουμε ότι το άνω 100α εκατοστιαίο σημείο του T^2 , το T_a^2 , είναι το σημείο εκείνο για το οποίο

$$\text{πιθ}(T^2 \geq T_a^2) = \alpha , \quad (4.10)$$

$$\text{όπου} \quad T_a^2 = \{(n-1)p/(n-p)\}F_{\alpha; p, n-p} , \quad (4.11)$$

και $F_{\alpha; p, n-p}$ είναι το άνω 100α εκατοστιαίο σημείο της $F_{p, n-p}$ κατανομής. Συνεπώς αν $T^2 \leq T_a^2$, τότε από την (4.8) , θα έχουμε ότι για κάθε p -διάστατο διάνυσμα a

$$t^2(a) \leq T_a^2 . \quad (4.12)$$

Αντικαθιστώντας τις (4.9) και (4.11) στην (4.12) και αντιστρέφοντας την πιθανότητα (4.10) παίρνουμε ότι



$\text{πιθ}\{a'\mu \in (a'\bar{x} - b, a'\bar{x} + b) \text{ για όλα τα } a\} = 1 - \alpha,$

$$\text{όπου } b = \left[\left\{ a' S a \frac{p}{(n-p)} \right\} F_{\alpha; p, n-p} \right]^{1/2}.$$

Συνεπώς με πιθανότητα $100(1-\alpha)\%$ ισχύει ότι

$$a'\bar{x} - \left(a' S a \frac{n}{n-p} F_{\alpha, p, n-p} \right)^{1/2} \leq a'\mu \leq a'\bar{x} + \left(a' S a \frac{n}{n-p} F_{\alpha, p, n-p} \right)^{1/2} \quad (4.13)$$

4.4.2.2 Πολλαπλές συγκρίσεις για τη διαφορά των μέσων τιμών δύο κανονικών πληθυσμών

Έστω $x_1, x_2, \dots, x_n$ και $y_1, y_2, \dots, y_m$ δύο ανεξάρτητα τυχαία δείγματα από τις κατανομές $N_p(\mu_x, \Sigma)$ και $N_p(\mu_y, \Sigma)$, αντίστοιχα, με Σ άγνωστο. Εργαζόμενοι όπως και στην προηγούμενη παράγραφο (§ 4.4.2.1) μπορούμε να δείξουμε ότι για κάθε σταθερό διάνυσμα a ισχύει:

$$\text{πιθ} \left\{ a' \delta \in \left[a' (\bar{x} - \bar{y}) - \left[a' (nS_x + mS_y) a \frac{p(n+m)}{nm(n+m-p-1)} F_{\alpha; p, p'} \right]^{1/2} \right] \right\} = 1 - \alpha$$

όπου $\delta = \mu_x - \mu_y$ και $p' = nm(n+m-p-1)$.

4.4.3 Έλεγχος διακυμάνσεων

4.4.3.1 Έλεγχος για τον πίνακα Σ ενός κανονικού πληθυσμού

Αν $x_1, x_2, \dots, x_n$ είναι ένα τυχαίο δείγμα από την $N_p(\mu, \Sigma)$, τότε ένας έλεγχος υπόθεσης για το Σ είναι ο

$$H_0: \Sigma = \Sigma_0, \text{ με } \mu \text{ άγνωστο,}$$



όπου Σ_0 είναι ένας γνωστός πίνακας διακυμάνσεων - συνδιακυμάνσεων.

Πηλίκo Μέγιστης Πιθανοφάνειας : Οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων μ και Σ όταν η H_0 ισχύει : $\hat{\mu} = \bar{x}$ και $\hat{\Sigma} = \Sigma_0$, ενώ όταν η H_a ισχύει : $\hat{\mu} = \bar{x}$ και $\hat{\Sigma} = S$. Συνεπώς

$$l_0(X; \bar{x}, \Sigma_0) = -\frac{n}{2} \log_e [\det(2\pi\Sigma)] - \frac{n}{2} \text{tr}(\Sigma_0^{-1}S)$$

και

$$l_1(X; \bar{x}, S) = -\frac{n}{2} \log_e [\det(2\pi S)] - \frac{n}{2} p.$$

Άρα

$$-2\log_e \lambda = 2(l_1 - l_0) = n \text{tr}(\Sigma_0^{-1}S) - n \log_e [\det(\Sigma_0^{-1}S)] - np.$$

Αν συμβολίσουμε με α και g τον αριθμητικό και γεωμετρικό μέσο, αντίστοιχα, των ιδιοτιμών του πίνακα $\Sigma_0^{-1}S$, τότε

$$\text{tr}(\Sigma_0^{-1}S) = p\alpha \quad \text{και} \quad \det(\Sigma_0^{-1}S) = g^p,$$

οπότε $-2\log_e \lambda = np(\alpha - \log_e p - 1)$. Η κατανομή του στατιστικού αυτού είναι αρκετά πολύπλοκη. (Για κάποιες προσπάθειες προσδιορισμού της ακριβούς κατανομής του βλ. Anderson (1958), Korin (1968) και Pearson & Hartley (1972)). Για το λόγο αυτό καταφεύγουμε στην ασυμπτωτική κατανομή του $-2\log_e \lambda$ όπως δίνεται από το θεώρημα 4.1.

Παράδειγμα 4.5 Χρησιμοποιώντας και πάλι τις μεταβλητές X_1 και X_3 από το παράδειγμα 4.1, ας ελέγξουμε την υπόθεση

$$H_0: \Sigma = \Sigma_0, \quad \Sigma_0 = \text{diag}(100 \quad 100), \quad \mu \text{ άγνωστο.}$$

Από τις τιμές των μεταβλητών X_1 και X_3 (βλ. Πίνακα 4.1) παίρνουμε

$$\Sigma_0^{-1}S = \begin{bmatrix} 0,01 & 0 \\ 0 & 0,01 \end{bmatrix} \begin{bmatrix} 91,48166,875 \\ 66,87596,775 \end{bmatrix} = \begin{bmatrix} 0,91480,6688 \\ 0,66880,9678 \end{bmatrix}.$$

Οι ιδιοτιμές του πίνακα αυτού είναι $\lambda_1 = 1,611$ και $\lambda_2 = 0,272$.



Συνεπώς $\alpha = 0,9415$ και $g = 0,6619$ και $-2\log_e \lambda = 17,707$.

Συγκρίνοντας την τιμή αυτή με την τιμή της $\chi^2_{0,05;3} = 9,49$ βλέπουμε ότι η H_0 απορρίπτεται.

Έχοντας απορρίψει ότι ο Σ είναι διαγώνιος ας ελέγξουμε την υπόθεση $H_0: \Sigma = \Sigma_0$, με $\Sigma_0 = \begin{bmatrix} 100 & 50 \\ 50 & 100 \end{bmatrix}$ και μ άγνωστο.

Στην περίπτωση αυτή

$$\Sigma_0^{-1} S = \begin{bmatrix} 0,77390, 2465 \\ 0,28180, 8445 \end{bmatrix}.$$

Οι ιδιοτιμές του πίνακα αυτού είναι $\lambda_1 = 1,0753$ και $\lambda_2 = 0,5431$ έτσι ώστε $\alpha = 0,8092$ και $g = 0,7642$. Συνεπώς $-2\log_e \lambda = 3,9065$. Συγκρίνοντας την τιμή αυτή με την τιμή της $\chi^2_{0,05;3} = 9,49$ βλέπουμε ότι η H_0 δεν μπορεί να απορριφθεί.

Ένωση - Τομή: θέτοντας ξανά $y = Xa$ έχουμε ότι οι συνιστώσες του $y, y_1, y_2, \dots, y_n$, είναι ανεξάρτητες με κατανομή την $N(\mu_y, \sigma_y^2)$, όπου $\mu_y = a'\mu$ και $\sigma_y^2 = a'\Sigma a$. Όταν η $H_0: \Sigma = \Sigma_0$ ισχύει, τότε ο έλεγχος της μονοδιάστατης υπόθεσης $H_0(a): \sigma_y^2 = \sigma_{oy}^2$, $\sigma_{oy}^2 = a'\Sigma_0 a$, γίνεται με το στατιστικό

$$U_a = \frac{n \cdot s_y^2}{\sigma_{oy}^2} = \frac{na'Sa}{a'\Sigma_0 a}.$$

Απορρίπτουμε την $H_0(a)$ αν $U_a \leq c_{1a}$ ή $U_a \geq c_{2a}$, όπου c_{1a} και c_{2a} εκλέγονται κατά τέτοιο τρόπο ούτως ώστε να έχουμε ένα τεστ ίσων ουρών. Επειδή (βλ. Παράρτημα)

$$\max_a \frac{n a'Sa}{a'\Sigma_0 a} = n\lambda_1 \quad \text{και} \quad \min_a \frac{n a'Sa}{a'\Sigma_0 a} = n\lambda_p$$

όπου λ_1 και λ_p είναι, αντίστοιχα, η μεγαλύτερη και η μικρότερη ιδιοτιμή του πίνακα $\Sigma_0^{-1} S$. Συνεπώς η H_0 απορρίπτεται όταν $\lambda_1 \leq c_1$ ή $\lambda_p \geq c_2$, όπου οι τιμές των c_1 και c_2 προσδιορίζονται έτσι ώστε να έχουμε τεστ ίσων ουρών. Η από κοινού κατανομή των ιδιοτιμών του πίνακα $\Sigma_0^{-1} S$ είναι αρκετά πολύπλοκη και δεν υπάρχουν πίνακες για διάφορα επίπεδα σημαντικότητας.



Παρατηρήστε ότι αυτή τη φορά οι δύο μέθοδοι, πηλίκo μέγιστης πιθανοφάνειας και ένωση - τομή, δεν έδωσαν τα ίδια στατιστικά.

4.4.3.2 Έλεγχος για την ισότητα των πινάκων διακυμάνσεων-συνδιακυμάνσεων δύο ή περισσότερων κανονικών πληθυσμών

Έστω ότι $\mathbf{x}_{k1}, \mathbf{x}_{k2}, \dots, \mathbf{x}_{kn_k}$ ($k = 1, 2, \dots, K$) είναι K τυχαία ανεξάρτητα μεταξύ τους διανύσματα από την $N_p(\mu_k, \Sigma_k)$. Τότε ο έλεγχος ομοιογένειας των K πληθυσμών, δηλαδή ο έλεγχος της υπόθεσης $H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_K$, ως προς την εναλλακτική $H_a: \Sigma_k \neq \Sigma_{k'}$, για τουλάχιστον ένα ζευγάρι k, k' ($k \neq k'$), με τη μέθοδο του πηλίκου μέγιστης πιθανοφάνειας, γίνεται ως εξής.

Από την (3.4) παίρνουμε ότι όταν η H_0 ισχύει τότε οι εκτιμητές μέγιστης πιθανοφάνειας των μ_k και Σ (όπου Σ είναι η κοινή τιμή των Σ_k) είναι οι :

$$\hat{\mu}_k = \bar{\mathbf{x}}_k \quad \text{και} \quad \hat{\Sigma} = \frac{n_1 \mathbf{S}_1 + n_2 \mathbf{S}_2 + \dots + n_K \mathbf{S}_K}{n_1 + n_2 + \dots + n_K},$$

όπου $\mathbf{S}_k$ είναι ο πίνακας των δειγματικών διακυμάνσεων-συνδιακυμάνσεων για τον k πληθυσμό ($k = 1, 2, \dots, K$)

Όταν η H_a ισχύει τότε οι εκτιμητές μέγιστης πιθανοφάνειας των μ_k και Σ_k είναι οι : $\hat{\mu}_k = \bar{\mathbf{x}}_k$ και $\hat{\Sigma}_k = \mathbf{S}_k$.

Συνεπώς θα είναι

$$\begin{aligned} -2 \log_e \lambda &= n \log_e [\det(\mathbf{S})] - \sum n_k \log_e [\det(\mathbf{S}_k)] \\ &= \sum n_k \log_e [\det(\mathbf{S}_k^{-1} \mathbf{S})]. \end{aligned} \quad (4.14)$$

Ή χρησιμοποιώντας το θεώρημα 4.1 θα έχουμε ότι ασυμπτωτικά

$$\sum n_k \log_e [\det(\mathbf{S}_k^{-1} \mathbf{S})] \sim \chi^2_{(1/2)p(p+1)(K-1)}.$$

Αν δούμε την σχέση (4.14) πιο προσεκτικά θα παρατηρήσουμε ότι, όταν τα n_k είναι μικρά, τότε ο κύριος παράγων που επηρεάζει



το $-2\log_e \lambda$ είναι ο S . Αυτή η παρατήρηση οδήγησε τον Box (1949) στην τροποποίηση του προηγούμενου στατιστικού. Η τροποποίηση αυτή είναι γνωστή σαν το M - στατιστικό του Box και δίνεται από την επόμενη σχέση

$$M = \gamma \sum (n_k - 1) \log_e \left[\det \left(S_{uk}^{-1} S_u \right) \right], \quad (4.15)$$

όπου

$$\gamma = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(K-1)} \left(\sum \frac{1}{n_k - 1} - \frac{1}{n - K} \right),$$

$$S_{uk} = \frac{n_k}{n_k - 1} S_k, \quad S_u = \frac{n}{n - K} S \quad \text{και} \quad n = \sum n_k.$$

Ασυμπτωτικά η κατανομή του στατιστικού (4.15) είναι η $\chi^2_{(1/2)p(p+1)(K-1)}$ -κατανομή. Η προσέγγιση αυτή είναι ικανοποιητική εάν κάθε n_k είναι μεγαλύτερο του 20 και εάν τα K και p δεν είναι μεγαλύτερα από 5.

Παράδειγμα 4.6 Στο Παράδειγμα 4.4 είχαμε να κάνουμε τον έλεγχο για την ισότητα των μέσων τιμών δύο πληθυσμών. Για τον έλεγχο αυτό μια από τις υποθέσεις είναι η ισότητα των πινάκων διακυμάνσεων - συνδιακυμάνσεων των δύο πληθυσμών. Στο παράδειγμα αυτό θα ελέγξουμε αυτήν ακριβώς την υπόθεση.

Οι τιμές των πινάκων S, S_1, S_2, S_u δίνονται στο Παράδειγμα 4.4. Οι τιμές των S_{u1} και S_{u2} είναι

$$S_{u1} = \frac{n_1}{n_1 - 1} S_1 = \begin{bmatrix} 138,31 & 78,88 & 37,25 \\ & 49,88 & 21,58 \\ & & 11,22 \end{bmatrix}$$

$$S_{u2} = \frac{n_2}{n_2 - 1} S_2 = \begin{bmatrix} 449,88 & 270,26 & 168,14 \\ & 171,15 & 102,95 \\ & & 66,42 \end{bmatrix}.$$

Ακόμη $\det(S_1) = 698,87$, $\det(S_2) = 11124,82$, $\det(S) = 4899,74$, $\det(S_{u1}) = 794,05$, $\det(S_{u2}) = 12639,89$, $\det(S_u) = 5567,03$ ενώ για την περίπτωση μας $K = 2$ και $p = 3$.



Από τις τιμές αυτές βρίσκουμε ότι η τιμή του στατιστικού $\sum n_k \log_e [\det(S_k^{-1}S)]$ είναι 27,06. Η τιμή αυτή είναι μεγαλύτερη από την τιμή της $\chi^2_{0,05;6} = 12,59$ και συνεπώς η υπόθεση ότι οι πίνακες διακυμάνσεων-συνδιακυμάνσεων των δύο πληθυσμών είναι ίσοι απορρίπτεται.

Για την τιμή του M - στατιστικού βρίσκουμε $\gamma = 0,9293$ και συνεπώς $M = 24,099$. Και η τιμή αυτή είναι μεγαλύτερη της τιμής του $\chi^2_{0,05;6} = 12,59$ άρα το συμπέρασμά μας είναι το ίδιο όπως και προηγουμένως.



Α Σ Κ Η Σ Ε Ι Σ

4.1 Έστω $X = (x_1, x_2, \dots, x_n)'$ ένα τυχαίο δείγμα από την $N_p(\mu, \Sigma)$ με Σ άγνωστο και ως υποθέσουμε ότι θέλουμε να ελέγξουμε την υπόθεση $H_0: \mu = \mu_0$. Να δειχθεί ότι ο εκτιμητής μέγιστης πιθανοφάνειας του Σ , $\hat{\Sigma} = S + dd'$, όπου $S = n^{-1} \sum (x_i - \bar{x})(x_i - \bar{x})'$ και $d = \bar{x} - \mu_0$ μπορεί να γραφεί σαν

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)(x_i - \mu_0)'.$$

4.2 Με τις προϋποθέσεις της άσκησης 4.1 να δειχθεί ότι

$$l_0 = l(X; \mu_0, S + dd') = -\frac{n}{2} \left\{ p \log(2\pi) + \log[\det(S)] + \log(1 + d'S^{-1}d) + p \right\}.$$

(Υπόδειξη: Αν $A(p \times p)$, $B(p \times n)$, $C(n \times p)$ και $D(n \times p)$ είναι πίνακες με τις διαστάσεις που αναφέρονται για τον καθένα εντός παρενθέσεως και όλοι οι απαραίτητοι αντίστροφοι υπάρχουν τότε

$$i) (A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

$$ii) \det(A + bb') = \det(A)(1 + b'S^{-1}b) \quad).$$

4.3 Στην περίπτωση ελέγχου ισότητας των μέσων τιμών δύο πολυδιάστατων κανονικών πληθυσμών με άγνωστες αλλά ίσες διακυμάνσεις να δειχθεί ότι

$$-2 \log \lambda = \frac{n \cdot m}{n+m} (\bar{x} - \bar{y})' S^{-1} (\bar{x} - \bar{y})$$

όπου n και m είναι, αντίστοιχα, τα μεγέθη των δειγμάτων των δύο πληθυσμών, $\bar{x}$ και $\bar{y}$ οι δειγματικές μέσες τιμές τους και S είναι ο εκτιμητής μέγιστης πιθανοφάνειας του κοινού πίνακα διακυμάνσεων



- συνδιακυμάνσεων.

4.4 Στην περίπτωση ενός τυχαίου δείγματος από κανονικό πληθυσμό με γνωστό Σ να βρεθεί ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την ποσότητα $a'\mu$.

Επίσης να δειχθεί ότι το $a'\mu$ ανήκει στο διάστημα που βρήκατε εάν και μόνον εάν το μ ανήκει στο $100(1-\alpha)\%$ ελλειψοειδές εμπιστοσύνης

$$n(\bar{x} - \mu)' \Sigma^{-1} (\bar{x} - \mu) < \chi^2_{\alpha; p}.$$

4.5 i) Μετρήσεις του μήκους (X_1) και του πλάτους (X_2) της κεφαλής 35 θυληκών βατράχων έδωσαν τις εξής τιμές

$$\bar{x}_1 = \begin{bmatrix} 22,860 \\ 24,397 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 17,178 & 19,710 \\ & 23,710 \end{bmatrix}.$$

Να ελεχθεί η υπόθεση ότι $\mu_1^\theta = \mu_2^\theta$ ($\mu_1^\theta = E(X_1^\theta)$ και $\mu_2^\theta = E(X_2^\theta)$).

ii) Παρόμοιες μετρήσεις σε 14 αρσενικούς βάτραχους έδωσαν

$$\bar{x}_2 = \begin{bmatrix} 21,821 \\ 22,843 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 17,159 & 17,731 \\ & 19,273 \end{bmatrix}.$$

Να ελεχθεί η υπόθεση ότι $\mu_1^a = \mu_2^a$ ($\mu_1^a = E(X_1^a)$ και $\mu_2^a = E(X_2^a)$).

iii) Να ελεχθεί η υπόθεση ότι οι διακυμάνσεις των δύο πληθυσμών είναι ίσες.

iv) Να ελεχθεί η υπόθεση ότι $\mu_1 = \mu_2$ ($\mu_1 = (\mu_1^\theta, \mu_2^\theta)'$ και $\mu_2 = (\mu_1^a, \mu_2^a)'$).

Σε όλα τα παραπάνω να αναφερθούν προσεκτικά όλες οι απαραίτητες υποθέσεις για την ισχύ των στατιστικών τεστ που θα χρησιμοποιείτε.

4.6 Ο παρακάτω πίνακας μας δίνει, για καθ' ένα από τους δύο πληθυσμούς, το μέγεθος του δείγματος και την μέση τιμή για



κάθε μία από τις δύο μεταβλητές X_1 και X_2 .

Πληθυσμός	Μέγεθος δείγματος	Δειγματική μέση τιμή	
		X_1	X_2
1	27	460,4	335,1
2	20	444,3	323,1

Ακόμη, από τα αρχικά δεδομένα, βρήκαμε ότι ο αμερόληπτος εκτιμητής του κοινού πίνακα διακυμάνσεων - συνδιακυμάνσεων είναι

$$S_u = \begin{bmatrix} 561,8 & 380,7 \\ & 343,2 \end{bmatrix}.$$

Αν συμβολίσουμε με $\mu_1 = [\mu_{11}, \mu_{21}]'$ και με $\mu_2 = [\mu_{12}, \mu_{22}]'$ τις μέσες τιμές των δύο πληθυσμών, να ελεγχθεί

$$i) H_0^1: \mu_{11} = \mu_{12} \quad \text{και} \quad H_0^2: \mu_{21} = \mu_{22}$$

$$ii) H_0^3: \mu_1 = \mu_2.$$

Να σχολιάσεται τα συμπεράσματά σας.

4.7 Με βάση την πληροφορία μιας παρατήρησης x από την $Np(\mu, \sigma^2 I)$, με σ^2 γνωστό, να βρεθεί το τεστ του πηλίκου μέγιστης πιθανοφάνειας για τον έλεγχο της $H_0: \mu' \mu = 1$.



5. ΚΥΡΙΕΣ ΣΥΝΙΣΤΩΣΕΣ

5.1 ΕΙΣΑΓΩΓΗ

Ας υποθέσουμε ότι θέλουμε να βρούμε το βαθμό επίδοσης στα Μαθηματικά των μαθητών της Γ' τάξης ενός Γυμνασίου. Στην περίπτωση αυτή θα μπορούσαμε να καταγράψουμε τους βαθμούς των μαθητών στα Μαθηματικά, δηλαδή να θεωρήσουμε ένα μόνο χαρακτηριστικό για κάθε μαθητή. Ας υποθέσουμε τώρα ότι αντί για τον βαθμό επίδοσης στα Μαθηματικά ενδιαφερόμαστε για την γενική επίδοση των μαθητών. Για το σκοπό αυτό καταγράφουμε τους βαθμούς των μαθητών σε p αντικείμενα (χαρακτηριστικά) π.χ. Μαθηματικά, Ελληνικά, Ιστορία, Θρησκευτικά κ.τ.λ. . Τα αντικείμενα αυτά , μολονότι συσχετίζονται μεταξύ τους, δεν μας δίνουν όλα το ίδιο ποσό πληροφορίας και μάλιστα μερικά από αυτά μπορεί να είναι τελείως άχρηστα. Προφανώς αυτό έχει σαν αποτέλεσμα το χάσιμο πληροφορίας κατά την ανάλυση των δεδομένων μας. Αυτό που θα έπρεπε να κάνουμε θα ήταν να εκλέξουμε εκείνα τα αντικείμενα (μαθήματα) τα οποία πράγματι μας διαχωρίζουν την απόδοση ενός μαθητή από αυτή ενός άλλου και να αγνοήσουμε τα υπόλοιπα αντικείμενα. Συχνά όμως ένα τέτοιο έργο δεν είναι και τόσο εύκολη υπόθεση.

Ένας άλλος τρόπος, για να αντιμετωπίσουμε το πρόβλημα



αυτό, θα ήταν να θεωρήσουμε p γραμμικές συναρτήσεις των p αντικειμένων και να διαλέξουμε εκείνες οι οποίες μας διαχωρίζουν την απόδοση ενός μαθητή από αυτή ενός άλλου κατά τον καλύτερο τρόπο. Η διαδικασία αυτή γίνεται πιο απλή αν αυτές τις p γραμμικές συναρτήσεις τις φτιάξουμε έτσι ώστε να είναι ασυσχέτιστες μεταξύ τους. Στην περίπτωση αυτή θα μπορούμε να κρατήσουμε εκείνες τις γραμμικές συναρτήσεις, από τις p , οι οποίες παρουσιάζουν μεγάλη μεταβλητότητα (διακύμανση) και να αγνοήσουμε εκείνες με μικρή μεταβλητότητα. Αυτές οι p γραμμικές συναρτήσεις ονομάζονται *κύριες συνιστώσες* (principal components).

Από τη μαθηματική του πλευρά το πρόβλημα των κυρίων συνιστωσών έχει ως εξής : δοθέντων p συσχετισμένων μεταβλητών να βρούμε ένα νέο σύνολο p ασυσχέτιστων μεταβλητών οι οποίες να είναι γραμμικές συναρτήσεις των αρχικών μεταβλητών. Οι νέες αυτές μεταβλητές βρίσκονται κατά φθίνουσα τάξη ενδιαφέροντος έτσι ώστε, για παράδειγμα, η πρώτη κύρια συνιστώσα να μας εξηγεί το μεγαλύτερο ποσοστό από τη μεταβλητότητα των αρχικών δεδομένων μας, η δεύτερη το αμέσως επόμενο ποσοστό κ.ο.κ .

Η χρήση των κυρίων συνιστωσών έχει σαν σκοπό να εξετάσουμε αν κάποιες από τις πρώτες κύριες συνιστώσες μας εξηγούν το μεγαλύτερο ποσοστό από την μεταβλητότητα των αρχικών μας δεδομένων. Αν αυτό συμβαίνει τότε λέμε ότι η πραγματική διάσταση του προβλήματός μας δεν είναι p , αλλά μικρότερη από p .

Η ιδέα των κυρίων συνιστωσών οφείλεται στον Pearson (1901), ενώ η συστηματική μελέτη τους στον Hotelling (1933) . Για περισσότερες λεπτομέρειες και διάφορες εφαρμογές της τεχνικής των κυρίων συνιστωσών βλέπε Rao (1964) .



5.2 ΕΥΡΕΣΗ ΤΩΝ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

5.2.1 Κύριες συνιστώσες από τον πίνακα διακυμάνσεων - συνδιακυμάνσεων

Έστω ότι $\mathbf{x} = (X_1, X_2, \dots, X_p)'$ είναι μια p -διάστατη τυχαία μεταβλητή με πίνακα διακυμάνσεων - συνδιακυμάνσεων Σ . Ας υποθέσουμε, για απλότητα, μολονότι από μαθηματικής πλευράς δεν είναι απαραίτητο, ότι ο πίνακας Σ είναι θετικά ορισμένος. Ας υποθέσουμε ακόμη, προς το παρόν, ότι για τις ιδιοτιμές του Σ ισχύει $\lambda_1 > \lambda_2 > \dots > \lambda_p \geq 0$.

Σκόπος μας είναι να βρούμε ένα νέο σύνολο τυχαίων μεταβλητών, έστω τις $Y_1, Y_2, \dots, Y_p$, οι οποίες να είναι ασυσχέτιστες μεταξύ τους και των οποίων η διακύμανση να φθίνει από την πρώτη στην τελευταία. Κάθε Y_j ($j = 1, 2, \dots, p$) θεωρείται σαν ένας γραμμικός συνδιασμός των X_j , έτσι ώστε

$$Y_j = \alpha_{1j} X_1 + \alpha_{2j} X_2 + \dots + \alpha_{pj} X_p = \alpha'_j \mathbf{x}, \quad (5.1)$$

όπου $\alpha_j = (\alpha_{1j}, \alpha_{2j}, \dots, \alpha_{pj})'$ είναι ένα σταθερό διάνυσμα.

Η πρώτη κύρια συνιστώσα, Y_1 , θα βρεθεί αν εκλέξουμε το α_1 έτσι ώστε η διακύμανση της Y_1 να είναι όσο το δυνατόν μεγαλύτερη. Από την (5.1) παίρνουμε ότι

$$\text{Var}(Y_1) = \text{Var}(\alpha'_1 \mathbf{x}) = \alpha'_1 \Sigma \alpha_1. \quad (5.2)$$

Επειδή τώρα η (5.2) περιέχει έναν αυθαίρετο παράγοντα, το α_1 , γι' αυτό βάζουμε τον περιορισμό $\alpha'_1 \alpha_1 = 1$. Ο περιορισμός αυτός μας προφυλάσσει από το να αυξήσουμε την διακύμανση του $\alpha'_1 \mathbf{x}$ αυθαίρετα με το να αυξήσουμε τις συνιστώσες του α_1 όσο εμείς θέλουμε. Αυτό σημαίνει ότι για κάθε διάνυσμα α_1 μπορούμε πάντοτε να βρούμε ένα άλλο διάνυσμα α_1^* έτσι ώστε

$$\text{Var}(\alpha_1^* \mathbf{x}) \geq \text{Var}(\alpha_1 \mathbf{x}).$$



Αυτό π.χ. μπορεί να γίνει με το να εκλέξουμε το διάνυσμα α_1^* έτσι ώστε να έχει την ίδια διεύθυνση με το α_1 αλλά μεγαλύτερο μήκος. Δηλαδή το νέο διάνυσμα θα είναι ένα πολλαπλάσιο του α_1 και κατά συνέπεια δεν θα αλλάζει το βασικό χαρακτηριστικό του $\alpha'x$. Συνεπώς αυτό που θα πρέπει να προσδιορίσουμε, για να έχουμε μια ικανοποιητική λύση στο πρόβλημά μας, είναι η διεύθυνση του α_1 και όχι το μήκος του. Το μήκος του α_1 θα πρέπει να είναι σταθερό και για ευκολία το παίρνουμε ίσο με 1.

Η ελαχιστοποίηση της (5.2), με τον περιορισμό ότι $\alpha_1' \alpha_1 = 1$, θα γίνει με την μέθοδο των πολλαπλασιαστών του Lagrange. Η συνάρτηση του Lagrange για την περίπτωση μας είναι :

$$L(\alpha_1) = \alpha_1' \Sigma \alpha_1 - \lambda(\alpha_1' \alpha_1 - 1), \quad (5.3)$$

όπου λ είναι ο πολλαπλασιαστής του Lagrange. Παραγωγίζοντας την (5.3) , ως προς α_1 , παίρνουμε

$$\frac{\partial}{\partial \alpha_1} L(\alpha_1) = 2 \Sigma \alpha_1 - 2 \lambda \alpha_1 .$$

Θέτοντας $\partial L(\alpha_1) / \partial \alpha_1 = 0$ παίρνουμε την εξίσωση

$$(\Sigma - \lambda I_p) \alpha_1 = 0, \quad (5.4)$$

όπου I_p είναι ο μοναδιαίος πίνακας τάξεως p . Η εξίσωση (5.4) παριστάνει ένα ομογενές σύστημα pxp εξισώσεων με p αγνώστους, τις συνιστώσες του διανύσματος α_1 . Έτσι για να βρούμε μια μη μηδενική λύση του συστήματος αυτού θα πρέπει ο πίνακας $\Sigma - \lambda I_p$ να είναι ιδιάζων. Άρα η τιμή του λ θα πρέπει να εκλεγεί έτσι ώστε $\det(\Sigma - \lambda I_p) = 0$. Η σχέση αυτή μας λέει ότι μια μη μηδενική λύση της (5.4) υπάρχει εάν και μόνον εάν το λ είναι μια από τις ιδιοτιμές του πίνακα Σ . Αλλά ο πίνακας Σ έχει, γενικώς, p ιδιοτιμές. Με το ποια από τις p αυτές ιδιοτιμές θα πρέπει να ισούται το λ θα το βρούμε από τις σχέσεις (5.2) και (5.4) . Από τις σχέσεις αυτές έχουμε ότι



$$\text{Var}(\alpha_1' \mathbf{x}) = \alpha_1' \Sigma \alpha_1 = \alpha_1' \lambda \mathbf{I}_p \alpha_1 = \lambda .$$

Η σχέση αυτή μαζί με το γεγονός ότι θέλουμε η διακύμανση να γίνεται μέγιστη μας οδηγεί στο συμπέρασμα ότι το λ θα πρέπει να είναι η μεγαλύτερη ιδιοτιμή του πίνακα Σ , δηλαδή $\lambda = \lambda_1$. Συνεπώς το α_1 θα πρέπει να είναι το ιδιοδιάνυσμα του Σ που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή.

Η δεύτερη κύρια συνιστώσα, δηλαδή η $Y_2 = \alpha_2' \mathbf{x}$, θα βρεθεί με τον ίδιο συλλογισμό όπως και η πρώτη κύρια συνιστώσα Y_1 , με τον επιπλέον περιορισμό ότι η Y_2 θα πρέπει να είναι ασυσχέτιστη με την Y_1 . Αλλά

$$\text{Cov}(Y_1, Y_2) = \text{Cov}(\alpha_1' \mathbf{x}, \alpha_2' \mathbf{x}) = \alpha_1' \text{Var}(\mathbf{x}) \alpha_2 = \alpha_1' \Sigma \alpha_2 . \quad (5.5)$$

Για να είναι η Y_2 ασυσχέτιστη με την Y_1 θα πρέπει $\alpha_1' \Sigma \alpha_2 = 0$. Επειδή όμως $\alpha_1' \Sigma = \alpha_1' \lambda_1$ παίρνουμε ότι η συνθήκη $\alpha_1' \alpha_2 = 0$ είναι ισοδύναμη της $\alpha_1' \Sigma \alpha_2 = 0$. Με άλλα λόγια θα πρέπει τα δύο διανύσματα α_1 και α_2 να είναι ορθογώνια μεταξύ τους.

Για να μεγιστοποιήσουμε την διακύμανση της Y_2 , δηλαδή την $\alpha_2' \Sigma \alpha_2$, με τους περιορισμούς $\alpha_2' \alpha_2 = 1$ και $\alpha_1' \alpha_2 = 0$ θα πρέπει να χρησιμοποιήσουμε και πάλι τους πολλαπλασιαστές του Lagrange. Έτσι θα έχουμε

$$L(\alpha_2) = \alpha_2' \Sigma \alpha_2 - \lambda(\alpha_2' \alpha_2 - 1) - \delta \alpha_1' \alpha_2 . \quad (5.6)$$

Παραγωγίζοντας ως προς α_2 και εξισώνοντας με την παράγωγο με το μηδέν θα έχουμε

$$\frac{\partial}{\partial \alpha_2} L(\alpha_2) = 2(\Sigma - \lambda \mathbf{I}_p) \alpha_2 - \delta \alpha_1 = 0 . \quad (5.7)$$

Πολλαπλασιάζοντας την τελευταία σχέση από αριστερά με α_1' και επειδή $\alpha_1' \alpha_2 = 0$ και $\alpha_1' \alpha_1 = 1$, παίρνουμε ότι

$$2\alpha_1' \Sigma \alpha_2 - \delta = 0 .$$

Αλλά, από την εξίσωση (5.5) έχουμε ότι $\alpha_1' \Sigma \alpha_2 = 0$ άρα το δ θα πρέπει να ισούται με μηδέν. Συνεπώς η εξίσωση (5.7) γράφεται



$$(\Sigma - \lambda I_p) \alpha_2 = 0$$

Σκεφτόμενοι όπως και για την πρώτη κύρια συνιστώσα μπορούμε να δούμε ότι η τιμή του λ θα πρέπει να είναι η δεύτερη μεγαλύτερη ιδιοτιμή του πίνακα Σ και το α_2 το ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή αυτή.

Συνεχίζοντας με αυτό τον τρόπο μπορούμε να δούμε ότι η j -στη κύρια συνιστώσα είναι το ιδιοδιάνυσμα που αντιστοιχεί στην j -στη μεγαλύτερη ιδιοτιμή του πίνακα Σ .

Όταν κάποιες από τις ιδιοτιμές του πίνακα Σ είναι ίσες μεταξύ τους τότε μπορούμε ακόμη να χρησιμοποιήσουμε τον προηγούμενο συλλογισμό, για την εύρεση των κυρίων συνιστωσών. Στην περίπτωση αυτή όμως τα ιδιοδιανύσματα που σχετίζονται με τις πολλαπλές ιδιοτιμές δεν είναι μοναδικά, αλλά από την στιγμή που εκλέγονται να είναι ορθογώνια μεταξύ τους, δεν υπάρχει πρόβλημα να εργασθούμε όπως και προηγουμένως.

Από τον τρόπο που βρήκαμε τις κύριες συνιστώσες θα έχουμε ότι

$$\sum_{j=1}^p \text{Var}(Y_j) = \sum_{j=1}^p \lambda_j = \text{tr}(\Lambda), \quad (5.8)$$

όπου Λ είναι ο διαγώνιος πίνακας $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$. Αν τώρα συμβολίσουμε με Λ τον πίνακα του οποίου στήλες είναι τα ιδιοδιανύσματα του πίνακα Σ , δηλαδή τα $\alpha_1, \alpha_2, \dots, \alpha_p$, τότε θα είναι

$$\text{tr}(\Lambda) = \text{tr}(A' \Sigma A) = \text{tr}(\Sigma A A') = \text{tr}(\Sigma) = \sum_{j=1}^p \text{Var}(X_j). \quad (5.9)$$

Από τις (5.8) και (5.9) παίρνουμε

$$\sum_{j=1}^p \text{Var}(Y_j) = \sum_{j=1}^p \text{Var}(X_j). \quad (5.10)$$

Η (5.10) μας λέει ότι το άθροισμα των διακυμάνσεων των αρχικών μεταβλητών ισούται με το άθροισμα των διακυμάνσεων των κυρίων συνιστωσών. Μπορούμε λοιπόν να πούμε ότι "η j -στη κύρια



συνιστώσα μας εξηγεί το $\lambda_j / \sum_{i=1}^p \lambda_i$ τις εκατό της ολικής μεταβλητότητας των αρχικών παρατηρήσεων." Ακόμη μπορούμε να πούμε ότι "οι m πρώτες κύριες συνιστώσες εξηγούν το $\sum_{i=1}^m \lambda_i / \sum_{i=1}^p \lambda_i$ τις εκατό της ολικής μεταβλητότητας των αρχικών δεδομένων." Θα πρέπει να τονισθεί εδώ ότι οι προηγούμενες προτάσεις δεν έχουν καμία σχέση με τους ανάλογους όρους της ανάλυσης διακύμανσης.

5.2.2 Κύριες συνιστώσες από τον πίνακα συσχέτισης

Στην πράξη πολλές φορές, και ιδίως όταν τα δεδομένα μας δεν μετρώνται με την ίδια μονάδα μέτρησης, μετασχηματίζουμε τις αρχικές μας μεταβλητές έτσι ώστε να έχουν μέση τιμή μηδέν και διακύμανση μονάδα. Ο μετασχηματισμός αυτός των αρχικών μεταβλητών είναι γνωστός με το όνομα μετασχηματισμός τυποποίησης. Η εύρεση των κυρίων συνιστωσών από τον πίνακα διακυμάνσεων - συνδιακυμάνσεων των μετασχηματισμένων μεταβλητών είναι ισοδύναμη με την εύρεση των κυρίων συνιστωσών από τον πίνακα συσχέτισης P των αρχικών μεταβλητών.

Από μαθηματικής πλευράς η εύρεση των κυρίων συνιστωσών με βάση τον πίνακα συσχέτισης P είναι η ίδια με αυτή για τον πίνακα Σ (βλ. § 5.2.1). Όπως και στην περίπτωση του πίνακα Σ έτσι και τώρα οι κύριες συνιστώσες, με βάση τον πίνακα P , είναι τα ιδιοδιανύσματα του P . Στο σημείο αυτό θα πρέπει να τονισθεί ότι οι ιδιοτιμές και τα ιδιοδιανύσματα των πινάκων Σ και P είναι, κατά κανόνα, διαφορετικά μεταξύ τους (βλ. και § 5.4).

Στον πίνακα συσχέτισης P όλα τα διαγώνια στοιχεία του είναι μονάδες. Αυτό σημαίνει ότι το άθροισμα των διαγώνιων στοιχείων του θα ισούται με p . Συνεπώς, σύμφωνα με την σχέση (5.8), το άθροισμα των ιδιοτιμών του P θα είναι επίσης ίσο με p . Άρα το



ποσοστό της ολικής μεταβλητότητας των αρχικών μεταβλητών που εξηγείται από την j -στη κύρια συνιστώσα θα είναι λ_j/p .

5.3 ΕΚΤΙΜΗΣΗ ΤΩΝ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ

Η εξαγωγή των κυρίων συνιστωσών, όπως παρουσιάσθηκε στις § 5.2.1 και 5.2.2, προϋποθέτει ότι οι πίνακες Σ και P είναι γνωστοί. Στην πραγματικότητα αυτό σπανίως συμβαίνει. Στις περιπτώσεις αυτές συνηθίζεται να αντικαθιστούμε τον πίνακα Σ (ή τον P) με τον αντίστοιχο δειγματικό S (ή τον R). Η εύρεση των κυρίων συνιστωσών με βάση τον πίνακα S (ή τον R) γίνεται όπως και στην § 5.2.1. Εάν συμβολίσουμε με $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$ τις ιδιοτιμές του πίνακα S , σε φθίνουσα τάξη και με $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_p$ τα αντίστοιχα ιδιοδιανύσματα, τότε τα $\{\hat{\lambda}_j\}$ και $\{\hat{\alpha}_j\}$ ($j = 1, 2, \dots, p$), μπορούν να θεωρηθούν σαν οι εκτιμητές των $\{\lambda_j\}$ και $\{\alpha_j\}$, αντίστοιχα.

Για να μελετήσουμε τις ιδιότητες των εκτιμητών αυτών θα πρέπει να υποθέσουμε ότι τα δεδομένα μας ακολουθούν πολυδιάστατη κανονική κατανομή. Στην περίπτωση αυτή μπορούμε να διατυπώσουμε μερικά ασυμπτωτικά συμπεράσματα (δηλ. όταν $n \rightarrow \infty$). Τα συμπεράσματα αυτά όμως (βλ. Morrison (1976) § 8.7, Mardia, Kend & Bibby (1982) § 8.3 & 8.4) έχουν περιορισμένο πρακτικό ενδιαφέρον επειδή i) είναι ασυμπτωτικά και ii) η υπόθεση της κανονικότητας των δεδομένων συνήθως δεν ισχύει. Για τους λόγους αυτούς η σύγχρονη τάση είναι να θεωρούνται οι κύριες συνιστώσες σαν μια μαθηματική τεχνική, χωρίς δηλαδή την υπόθεση ενός στατιστικού μοντέλου και ακόμη χωρίς την ανάγκη οι μεταβλητές $X_1, X_2, \dots, X_p$ και $Y_1, Y_2, \dots, Y_p$ να είναι τυχαίες μεταβλητές. Ακόμη οι κύριες συνιστώσες που βρίσκονται με βάση τον πίνακα S (ή R) θεωρούνται σαν οι κύριες συνιστώσες και όχι σαν οι εκτιμητές των αντίστοιχων ποσοτήτων του πίνακα Σ (ή P).



Παράδειγμα 5.1 Ας υποθέσουμε ότι έχουμε δύο τυποποιημένες μεταβλητές, τις X_1 και X_2 με πίνακα συσχέτισης

$$P = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Για να βρούμε τις κύριες συνιστώσες θα πρέπει να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα P . Οι ιδιοτιμές θα προκύψουν από την εξίσωση $\det(P - \lambda I_2) = 0$ η οποία μας δίνει $(1 - \lambda)^2 - \rho^2 = 0$. Συνεπώς οι ιδιοτιμές θα είναι $(1 + \rho)$ και $(1 - \rho)$.

Αν $\rho > 0$ τότε οι ιδιοτιμές σε φθίνουσα τάξη θα είναι $\lambda_1 = 1 + \rho$ και $\lambda_2 = 1 - \rho$. Το ιδιοδιάνυσμα που αντιστοιχεί στην ιδιοτιμή λ_1 , δηλαδή το $\alpha_1 = (\alpha_{11}, \alpha_{21})'$, θα βρεθεί από την σχέση $P\alpha_1 = \lambda_1 \alpha_1$ η οποία είναι ισοδύναμη με το σύστημα των δύο εξισώσεων

$$\alpha_{11} + \rho\alpha_{21} = (1 + \rho)\alpha_{11} \quad \text{και} \quad \rho\alpha_{11} + \alpha_{21} = (1 + \rho)\alpha_{21}.$$

Οι δύο αυτές εξισώσεις είναι προφανώς ίδιες και μας δίνουν

$\alpha_{11} = \alpha_{21}$. Αυτό σημαίνει ότι δεν έχουμε μοναδική λύση. Για να βρούμε την πρώτη κύρια συνιστώσα κάνουμε χρήση του περιορισμού $\alpha_1' \alpha_1 = 1$. Στην περίπτωση αυτή βρίσκουμε ότι $\alpha_{11} = \alpha_{21} = 1/\sqrt{2}$. Εργαζόμενοι ανάλογα για τη δεύτερη κύρια συνιστώσα βρίσκουμε ότι $\alpha_2 = (1/\sqrt{2}, -1/\sqrt{2})'$. Συνεπώς οι κύριες συνιστώσες θα είναι

$$\begin{aligned} Y_1 &= (X_1 + X_2)/\sqrt{2} \\ \text{και} \quad Y_2 &= (X_1 - X_2)/\sqrt{2}. \end{aligned}$$

Αν $\rho < 0$ τότε η σειρά των ιδιοτιμών καθώς και αυτή των κυρίων συνιστωσών θα είναι αντίστροφη από την περίπτωση $\rho > 0$.

Αν $\rho = 0$ τότε και οι δύο ιδιοτιμές θα είναι ίσες με τη μονάδα. Στην περίπτωση αυτή μπορούμε να εκλέξουμε οποιεσδήποτε συνιστώσες, σαν κύριες συνιστώσες, αρκεί να είναι κάθετες μεταξύ τους.



5.4 ΕΡΜΗΝΕΙΑ - ΧΡΗΣΕΙΣ - ΣΧΟΛΙΑ

5.4.1 Ερμηνεία

Όπως έχουμε αναφέρει, κάθε κύρια συνιστώσα είναι γραμμικός συνδιασμός των αρχικών μεταβλητών. Αυτό έχει σαν αποτέλεσμα οι κύριες συνιστώσες να μην έχουν κάποια φυσική σημασία, σχετική με το υπό εξέταση πρόβλημα. Μια διέξοδος στη δυσκολία αυτή δίνεται στην συνέχεια.

Από την § 5.2.1 θα έχουμε ότι $Y = A'X$. Αυτό σημαίνει ότι η j -στη κύρια συνιστώσα θα εξαρτάται από την j -στη γραμμή του πίνακα A' (ή ισοδύναμα, από την j -στη στήλη του πίνακα A). Επειδή ο πίνακας A είναι ορθογώνιος παίρνουμε ότι $X = AY$. Δηλαδή η i -στη αρχική μεταβλητή θα είναι ένας γραμμικός συνδιασμός των κύριων συνιστωσών με συντελεστές τα στοιχεία της i -στης γραμμής του πίνακα A . Έτσι

$$\text{Cov}(Y_j, X_i) = \text{Cov}\left(Y_j, \sum_{k=1}^p \alpha_{ik} Y_k\right) = \alpha_{ij} \text{Var}(Y_j) = \alpha_{ij} \lambda_j$$

και συνεπώς

$$\rho(Y_j, X_i) = \frac{\text{Cov}(Y_j, X_i)}{\sqrt{\text{Var}(Y_j) \cdot \text{Var}(X_i)}} = \frac{\alpha_{ij} \cdot \lambda_j}{\sigma_{ii}^2 \cdot \sqrt{\lambda_j}}.$$

Άρα

$$\rho(Y_j, X_i) = \frac{\alpha_{ij}^2 \cdot \lambda_j^{1/2}}{\sigma_{ii}^2}. \quad (5.11)$$

Όταν οι αρχικές μας μεταβλητές έχουν μετασχηματισθεί ώστε να έχουν διακύμανση μονάδα, τότε

$$\rho(Y_j, X_i) = \alpha_{ij} \lambda_j^{1/2}.$$

Οι συσχετίσεις μεταξύ των Y_j και X_i ονομάζονται *φορτία* ή *βαρύτητα συνιστωσών* (component loadings). Ανάλογα με την βαρύτητα κάθε αρχικής μεταβλητής με την κύρια συνιστώσα Y_j



($j = 1, 2, \dots, p$) προσπαθούμε να δώσουμε κάποια φυσική σημασία στις κύριες συνιστώσες. (Βλ. και Παραδείγματα 5.2 και 5.3) .

5.4.2 Χρήσεις

Πριν μιλήσουμε για τις χρήσεις των κυρίων συνιστωσών θα πρέπει να τονίσουμε τα εξής:

1) Οι κύριες συνιστώσες που προκύπτουν από τον πίνακα Σ (ή S) δεν είναι ίδιες με αυτές που προκύπτουν από τον πίνακα P (ή R). Ακόμη, δεν υπάρχει καμία σχέση που να μας επιτρέπει την μετάβαση από το σύνολο των κυρίων συνιστωσών για τον πίνακα Σ (ή S) στο σύνολο των κυρίων συνιστωσών για τον πίνακα P (ή R).

2) Αν δούμε τις κύριες συνιστώσες που προκύπτουν από τον πίνακα S (ή R) σαν τους εκτιμητές των αντίστοιχων κυρίων συνιστωσών που προκύπτουν από τον πίνακα Σ (ή P), τότε οι κύριες συνιστώσες του πίνακα S (ή R) μπορούν να θεωρηθούν σαν τυχαίες μεταβλητές. Στην περίπτωση αυτή υπάρχουν διάφορα αποτελέσματα τόσο για την κατανομή των κυρίων συνιστωσών, όσο και για τον έλεγχο υποθέσεων σχετικά με αυτές. (βλ. Morrison (1976)) , Anderson (1984)) Όλα αυτά τα αποτελέσματα είναι ασυμπτωτικά και βασίζονται στην υπόθεση ότι η από κοινού κατανομή των αρχικών μας μεταβλητών είναι η πολυδιάστατη κανονική. Επιπλέον τα συμπεράσματα αυτά είναι πιο δύσκολα στην περίπτωση που οι κύριες συνιστώσες εξάγονται από τον πίνακα R .

3) Η εξαγωγή των κυρίων συνιστωσών θα πρέπει να γίνεται από τον πίνακα Σ (ή S) , όταν όλες οι μεταβλητές μας έχουν την ίδια διάσταση. Στην αντίθετη περίπτωση θα πρέπει να χρησιμοποιείται ο πίνακας P (ή R).



ι) *Ελάττωση της διάστασης ενός προβλήματος*

Μια από τις πρώτες χρήσεις των κυρίων συνιστωσών είναι η ελάττωση της διάστασης ενός προβλήματος (βλ. § 5.1) . Για να ελαττώσουμε την διάσταση ενός προβλήματος είναι φανερό ότι θα πρέπει να κρατήσουμε τις πρώτες k ($k < p$) κύριες συνιστώσες. Το πρόβλημα, στην περίπτωση αυτή , είναι ο προσδιορισμός της τιμής του k .

Στην περίπτωση που οι κύριες συνιστώσες εξαγώνται από τον πίνακα Σ (ή S) τότε ο προσδιορισμός της τιμής του k γίνεται είτε με την βοήθεια στατιστικών τεστ (βλ. Mardia, Kent & Bibby (1982)) είτε με την βοήθεια της σχέσης (βλ. § 5.2.1)

$$\sum_{i=1}^k \lambda_i / \sum_{i=1}^p \lambda_i . \quad (k < p) \quad (5.12)$$

Δηλαδή η τιμή του k προσδιορίζεται έτσι ώστε οι k πρώτες κύριες συνιστώσες να μας εξηγούν ένα "σημαντικό" ποσοστό από την ολική μεταβλητότητα των αρχικών μας μεταβλητών . Προφανώς ο τρόπος είναι υποκειμενικός.

Στην περίπτωση που η εύρεση των κυρίων συνιστωσών γίνεται με βάση τον πίνακα P (ή R) τότε αφ' ενός μεν k δεν υπάρχει στατιστικό τεστ αφ' ετέρου δε η χρήση της σχέσης $\sum_{i=1}^k \lambda_i / p$ δεν έχει την σημασία που έχει η (5.12). Για τους λόγους αυτούς έχουν προταθεί κατά καιρούς διάφορες τεχνικές για τον προσδιορισμό της τιμής του k . Μερικές από τις τεχνικές αυτές είναι οι εξής :

1) Τεχνική του Kaiser : Η τεχνική του Kaiser (1958), είναι πιθανότατα η πιο συχνά χρησιμοποιούμενη τεχνική. Η τεχνική αυτή λέει ότι " κρατάμε όλες εκείνες τις κύριες συνιστώσες των οποίων η ιδιοτιμή είναι μεγαλύτερη του ένα (1) ". Η λογική πίσω από την τεχνική αυτή είναι ότι κάποιες κύριες συνιστώσες θα πρέπει να εξηγούν περισσότερη μεταβλητότητα από οποιαδήποτε αρχική



μεταβλητή. (Θα πρέπει να έχουμε στο μυαλό μας ότι όταν κάνουμε χρήση του πίνακα P (ή R) οι αρχικές μας μεταβλητές είναι τυποποιημένες (βλ. § 5.2.2)) .

2) Τεχνική του Cattell : Η τεχνική του Cattell (1966) είναι γνωστή και με το όνομα τεχνική του αγκώνα (scree test) . Σύμφωνα με την τεχνική αυτή εργαζόμαστε ως εξής :

Βήμα 1._ Κάνουμε γραφική παράσταση των κυρίων συνιστωσών κατά σειρά μεγέθους.

Βήμα 2._ Κατόπιν βρίσκουμε, με το μάτι, την ευθεία πάνω ή κοντά στην οποία βρίσκονται οι τελευταίες ιδιοτιμές .

Βήμα 3._ Η ευθεία αυτή σχηματίζει μια γωνία, ένα αγκώνα, με τις υπόλοιπες ιδιοτιμές .

Βήμα 4._ Η τιμή του k θα είναι ο αριθμός των ιδιοτιμών που βρίσκονται πάνω από την ευθεία στο Βήμα 2. (βλ. Cattell & Jaspers (1967)).

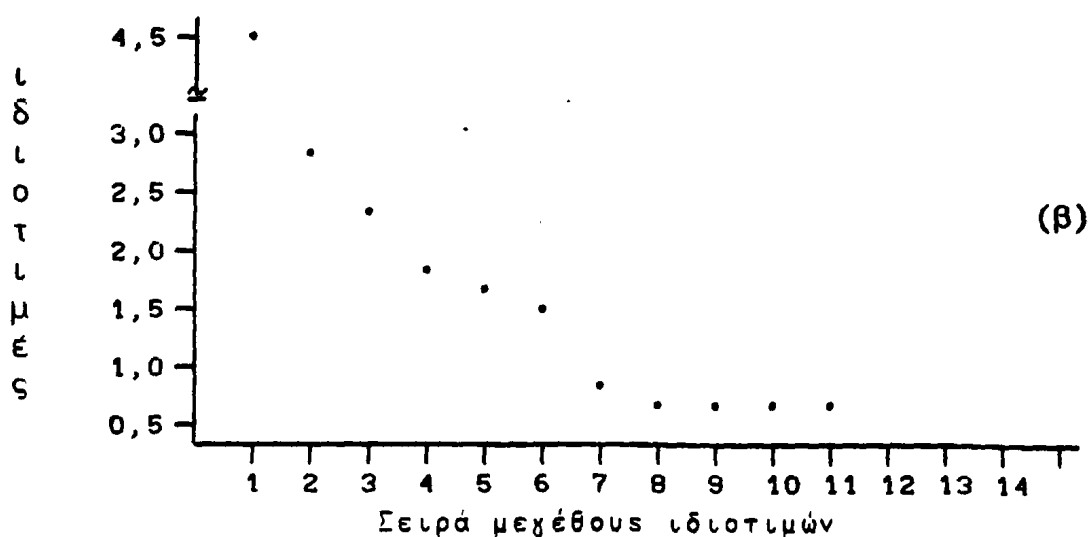
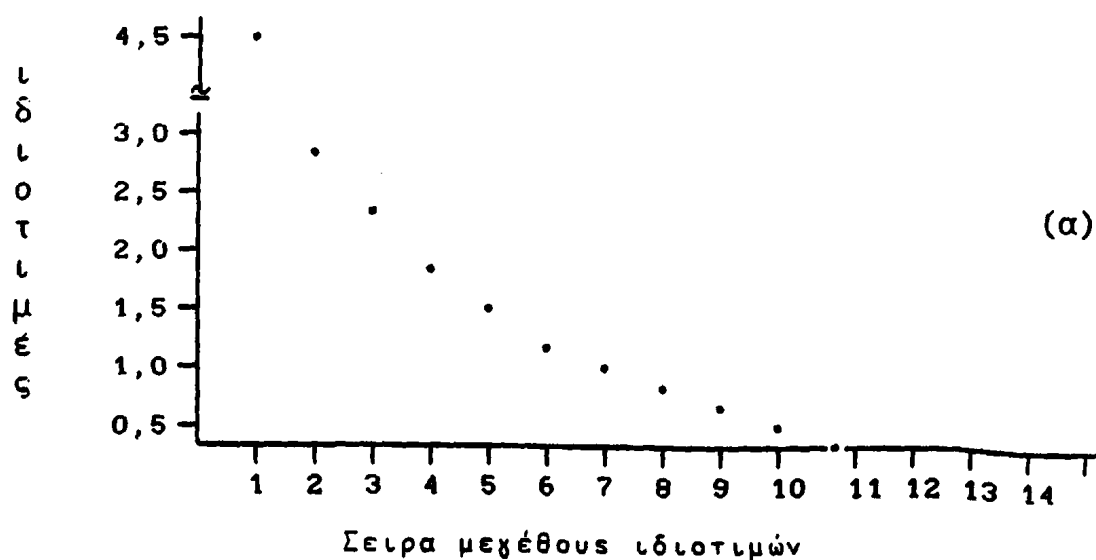
Τα προηγούμενα βήματα επεξηγούνται στο επόμενο σχήμα .



Σχ. 5.1. Επεξήγηση της Τεχνικής του αγκώνα.

Η τεχνική αυτή μολονότι είναι σχετικά απλή εν τούτοις δεν είναι πάντοτε αποτελεσματική. Κατ' αρχήν μπορεί να μην υπάρχει μια προφανής ευθεία (βλ. π.χ. Σχήμα 5.2 (α)) ή αντίθετα, μπορεί να υπάρχουν δύο ή περισσότερες προφανείς γωνίες (βλ. π.χ. Σχήμα 5.2 (β)) . Στις περιπτώσεις αυτές η υπό εξέταση τεχνική δεν είναι συμπερασματική.





Σχ. 5.2. Επεξήγηση των δυσκολιών της τεχνικής του αγκώνα.

Μια άλλη τεχνική, την οποία όμως δεν θα αναπτύξουμε εδώ, είναι η τεχνική του Horn (1965). Η τεχνική αυτή αντιμετωπίζει με επιτυχία τα μεμονωμένα της προηγούμενης τεχνικής.



ii) Προσδιορισμός συγγραμμικών δεδομένων.

Ως γνωστόν, (βλ. Κ. Καρακώστας (1992)), το μοντέλο της γραμμικής παλινδρόμησης μπορεί να γραφεί υπό μορφή πινάκων ως

$$y = X\beta + \varepsilon ,$$

όπου y είναι ένα το $n \times 1$ διάνυσμα της εξηρτημένης μεταβλητής, X είναι ο $n \times p$ πίνακας των ανεξάρτητων μεταβλητών (συμπεριλαμβανομένου και του σταθερού όρου), β είναι το $p \times 1$ διάνυσμα των αγνώστων παραμέτρων και ε είναι το $n \times 1$ διάνυσμα των σφαλμάτων. Ο εκτιμητής ελαχίστων τετραγώνων του β , με την προ- υπόθεση ότι ο $(X'X)^{-1}$ υπάρχει, δίνεται από την σχέση

$$\hat{\beta} = (X'X)^{-1}X'y . \quad (5.13)$$

Στην περίπτωση που υπάρχει γραμμική σχέση μεταξύ όλων ή κάποιων ανεξάρτητων μεταβλητών τότε ο $(X'X)^{-1}$ δεν υπάρχει. Αυτό έχει σαν αποτέλεσμα (από την (5.13)) να μην μπορούμε να βρούμε τους εκτιμητές των β . Το πρόβλημα αυτό είναι γνωστό με το όνομα *συγγραμμικά δεδομένα* (collinear data).

Ένας τρόπος αντιμετώπισης του προβλήματος αυτού είναι με την βοήθεια των κυρίων συνιστωσών. Η ανάπτυξη του θέματος αυτού είναι εκτός του σκοπού των σημειώσεων αυτών και γι' αυτό δεν θα παρουσιασθεί εδώ. Ο αναγνώστης μπορεί να ανατρέξει στα βιβλία των Chatterjee & Price (1977) και Mardia, Kent & Bibby (1982).

iii) Απόρριψη μεταβλητών

Η ελάττωση της διάστασης ενός προβλήματος μπορεί να γίνει και στις αρχικές μεταβλητές. Αυτό μπορεί να επιτευχθεί με την απόρριψη εκείνων, από τις αρχικές μεταβλητές, οι οποίες δεν τόσο πληροφοριακές για το υπό εξέταση πρόβλημα.

Ας υποθέσουμε ότι οι κύριες συνιστώσες βρίσκονται με βάση



τον πίνακα συσχέτισεως των p αρχικών μεταβλητών. Ας υποθέσουμε ακόμη ότι πρόκειται να κρατήσουμε k από τις αρχικές μεταβλητές, όπου k είναι γνωστός αριθμός, ενώ οι υπόλοιπες $p-k$ πρόκειται να απορριφθούν. Για την απόρριψη των $p-k$ αρχικών μεταβλητών χρησιμοποιούμε την μέθοδο των Beal et al (1967) η οποία έχει ως εξής:

Βήμα 1.- Βρίσκουμε το ιδιοδιάνυσμα (κύρια συνιστώσα) που αντιστοιχεί στην μικρότερη ιδιοτιμή.

Βήμα 2.- Στην συνέχεια η μεταβλητή με τον μεγαλύτερο, κατ' απόλυτον τιμή, συντελεστή απορρίπτεται.

Βήμα 3.- Επαναλαμβάνουμε τα Βήματα 1 και 2 έως ότου απορρίψουμε $p-k$ από τις αρχικές μεταβλητές.

Η λογική πίσω από την διαδικασία αυτή είναι ότι μια κύρια συνιστώσα με μικρή ιδιοτιμή δεν είναι και τόσο πληροφοριακή. Κατά συνέπεια και η μεταβλητή που χαρακτηρίζει την συνιστώσα αυτή θα πρέπει να είναι περιορισμένου ενδιαφέροντος.

Προφανώς η τιμή του k είναι τελείως υποκειμενική. Μια πιο αντικειμενική εκλογή της τιμής του k είναι να θέσουμε ένα άνω φράγμα, έστω λ_0 , έτσι ώστε όλες οι ιδιοτιμές λ οι οποίες είναι μικρότερες της τιμής λ_0 θα θεωρούνται ως περιορισμένου ενδιαφέροντος. Ο Jolliffe (1972) συνιστά όπως η τιμή του λ_0 είναι ίση με 0,7.

iv) Χρησιμοποίηση σε επόμενη ανάλυση.

Ένα σπουδαίο πλεονέκτημα των κυρίων συνιστωσών είναι ότι μας παρέχουν ένα γρήγορο τρόπο προσδιορισμού της διάστασης ενός προβλήματος. Αν μερικές από τις πρώτες κύριες συνιστώσες εξηγούν το μεγαλύτερο ποσοστό της μεταβλητότητας των αρχικών μας μεταβλητών, τότε μια καλή ιδέα θα ήταν να χρησιμοποιήσουμε τις



πρώτες αυτές κύριες συνιστώσες στην συνέχεια της ανάλυσης των δεδομένων μας.

5.4.3 Σχόλια

Στην § 5.4.2(i) τονίσαμε ότι οι κύριες συνιστώσες που βρίσκουμε όταν χρησιμοποιούμε τον πίνακα Σ (ή S) είναι διαφορετικές από αυτές που βρίσκουμε όταν χρησιμοποιούμε τον πίνακα P (ή R).

Ένα άλλο πρόβλημα με τις κύριες συνιστώσες, όταν χρησιμοποιούμε τον πίνακα Σ (ή S) είναι η εξάρτησή τους από την κλίμακα με την οποία μετράμε τις παρατηρήσεις μας. Για παράδειγμα ας υποθέσουμε ότι σε n άτομα μετράμε το βάρος σε λίμβρες το ύψος σε πόδια και την ηλικία τους σε χρόνια. Τα μεγέθη αυτά αποτελούν το διάνυσμα x , δηλαδή

$x = (\text{Βάρος σε λίμβρες} , \text{'Υψος σε πόδια} , \text{Ηλικία σε χρόνια})'$.
Ας συμβολίσουμε με S_x τον δειγματικό πίνακα διακυμάνσεων - συνδιακυμάνσεων του διανύσματος x , και έστω $\{\lambda_i\}$ οι ιδιοτιμές και $\{\alpha_i\}$ και τα ιδιοδιανύσματα του S_x .

Αν τώρα θελήσουμε να αλλάξουμε μονάδες μέτρησης και να χρησιμοποιήσουμε π. χ. το διάνυσμα

$z = (\text{Βαρος σε κιλα} , \text{'Υψος σε μέτρα} , \text{Ηλικία σε μήνες})'$
τότε $z = Kx$ όπου $K = \text{diag}(1/2, 2 , 1/3, 28 , 12)$. Ο δειγματικός πίνακας των διακυμάνσεων - συνδιακυμάνσεων για το διάνυσμα z θα είναι $S_z = KS_xK$, επειδή $K' = K$. Οι ιδιοτιμές και τα ιδιοδιανύσματα του S_z θα είναι, κατά κανόνα, διαφορετικά από τις αντίστοιχες ιδιοτιμές και ιδιοδιανύσματα του S_x . Το πρακτικό συμπέρασμα από τα παραπάνω είναι ότι οι κύριες συνιστώσες δεν αποτελούν ένα μοναδικό χαρακτηριστικό για τα δεδομένα μας.

Ακόμη εάν μία από τις μεταβλητές $X_1, X_2, \dots, X_p$, έχει



διακύμανση αρκετά μεγαλύτερη από τις διακυμάνσεις των υπόλοιπων μεταβλητών, τότε η μεταβλητή αυτή θα χαρακτηρίζει την πρώτη κύρια συνιστώσα, ανεξάρτητα από τις συσχετίσεις που υπάρχουν μεταξύ των $X_1, X_2, \dots, X_p$. Αντίθετα, εάν μετασχηματίσουμε τις αρχικές μας μεταβλητές έτσι ώστε να έχουν διακύμανση μονάδα, τότε η πρώτη κύρια συνιστώσα θα είναι αρκετά διαφορετική.

Οι παραπάνω λόγοι περιορίζουν τη χρήση των κυρίων συνιστωσών, με βάση τον πίνακα Σ (ή S), εκτός και αν οι μεταβλητές μας έχουν περίπου την ίδια διακύμανση, όπως π. χ. μπορεί να συμβεί αν όλες οι μεταβλητές μας εκφράζουν ποσοστά ή μετρώνται με την ίδια κλίμακα.

Ο συνήθης τρόπος αντιμετώπισης των προβλημάτων αυτών είναι η χρήση του πίνακα συσχέτισης P (ή R). Ο τρόπος αυτός όμως δεν λύνει, αλλά μάλλον αποφεύγει το πρόβλημα, διότι με τον τρόπο αυτό όλες οι μεταβλητές μας αποκτούν την ίδια σπουδαιότητα. Συνεπώς αν κάποια ή κάποιες από τις μεταβλητές μας έχουν κάποια πρακτική σπουδαιότητα μεγαλύτερη από τις υπόλοιπες μεταβλητές, η χρήση του πίνακα συσχέτισης θα πρέπει να αποφεύγεται.

Θα κλείσουμε το κομμάτι αυτό λέγοντας ότι σήμερα, με την ευρεία διάδοση και χρήση των μικρουπολογιστών και των διάφορων στατιστικών πακέτων, η εύρεση των κυρίων συνιστωσών είναι μια πολύ εύκολη υπόθεση. Αυτό το οποίο είναι δύσκολο είναι η ερμηνεία τους.

Στη συνέχεια θα δώσουμε μερικά παραδείγματα πάνω στις κύριες συνιστώσες και στην ερμηνεία τους.

Παράδειγμα 5.2 Από τις μετρήσεις σε εννέα μεταβλητές $X_1, X_2, \dots, X_9$ βρήκαμε ότι ο πίνακας των δειγματικών διακυμάνσεων - συνδιακυμάνσεων S είναι



$$S = \begin{bmatrix} 177 & 179 & 95 & 96 & 53 & 32 & -7 & -4 & -3 \\ & 419 & 245 & 131 & 181 & 127 & -2 & 1 & 4 \\ & & 302 & 60 & 109 & 142 & 4 & 4 & 11 \\ & & & 158 & 102 & 42 & 4 & 3 & 2 \\ & & & & 137 & 96 & 4 & 5 & 6 \\ & & & & & 128 & 2 & 2 & 8 \\ & & & & & & 34 & 31 & 33 \\ & & & & & & & 39 & 39 \\ & & & & & & & & 48 \end{bmatrix}$$

Οι ιδιοτιμές του S μαζί με το ποσοστό της μεταβλητότητας των αρχικών μεταβλητών που εξηγεί η κάθε μία, καθώς και το αθροιστικό ποσοστό δίνονται στον επόμενο πίνακα.

Πίνακας 5.1

A/A Κύριας Συνιστώσας	Ιδιοτιμή	Ποσοστό Μεταβλητότητας	Αθροιστικό Ποσοστό Μεταβλητότητας
1	878.52	0,6092	0,6092
2	196.10	0,1360	0,7452
3	128,64	0,0892	0,8344
4	103,43	0,0717	0,9061
5	81.26	0,0564	0,9625
6	37,85	0,0262	0,9887
7	6,98	0,0048	0,9935
8	5,71	0,0040	0,9975
9	3,52	0,0024	0,9999

Από τον πίνακα 5.1 βλέπουμε ότι το 83% της ολικής μεταβλητότητας των αρχικών μας μεταβλητών εξηγείται από τις τρεις πρώτες κύριες συνιστώσες, ενώ οι δύο πρώτες κύριες συνιστώσες εξηγούν το 75% .

Στον επόμενο πίνακα δίνουμε τα ιδιοδιανύσματα για τις τρεις πρώτες κύριες συνιστώσες.



Πίνακας 5.2

Κύρια Συν/σα	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	0,305	0,654	0,482	0,261	0,323	0,271	0,002	0,006	0,014
2	-0,485	-0,150	0,587	-0,491	-0,038	0,376	0,057	0,053	0,088
3	-0,412	-0,182	-0,235	0,457	0,495	0,268	0,256	0,266	0,282

Με βάση τον Πίνακα 5.2 μπορούμε να πούμε ότι i) η πρώτη κύρια συνιστώσα εκφράζει το μέσο όρο των έξι πρώτων αρχικών μεταβλητών, ii) η δεύτερη κύρια συνιστώσα συγκρίνει τις μεταβλητές X_1 και X_2 με την μεταβλητή X_3 και τις μεταβλητές X_4 και X_5 με την μεταβλητή X_6 και iii) η τρίτη κύρια συνιστώσα συγκρίνει τις τρεις πρώτες μεταβλητές με τις υπόλοιπες.

Στον Πίνακα 5.3 δίνουμε τα φορτία συνιστωσών για τις τρεις πρώτες κύριες συνιστώσες.

Πίνακας 5.3

Κύρια Συν/σα	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	0,679	0,946	0,824	0,616	0,819	0,710	0,012	0,028	0,062
2	-0,511	-0,103	0,473	-0,574	-0,045	0,462	0,137	0,120	0,179
3	-0,351	-0,101	-0,154	0,412	0,480	0,264	0,497	0,483	0,462

Σαν συμπέρασμα από τον τελευταίο Πίνακα μπορούμε να πούμε ότι η συνεισφορά των αρχικών μεταβλητών X_7 , X_8 , X_9 στις δύο πρώτες κύριες συνιστώσες είναι σχεδόν μηδαμινή.

Παράδειγμα 5.3 Σε ένα σύνολο πουλιών μετρήσαμε έξι μεταβλητές : το μήκος και το πλάτος της κεφαλής, το μήκος και το πλάτος του φτερού και το μήκος και το πλάτος του ποδιού. Ο πίνακας των δειγματικών συσχετίσεων βρέθηκε ίσος με



$$R = \begin{bmatrix} 1 & 0,584 & 0,615 & 0,604 & 0,570 & 0,600 \\ & 1 & 0,576 & 0,530 & 0,526 & 0,555 \\ & & 1 & 0,940 & 0,875 & 0,878 \\ & & & 1 & 0,877 & 0,886 \\ & & & & 1 & 0,924 \\ & & & & & 1 \end{bmatrix}$$

Οι ιδιοτιμές (δισκυμάνσεις) και τα ιδιοδιανύσματα (κύριες συνιστώσες) του R φαίνονται στον Πίνακα 5.4 .

Πίνακας 5.4

Διάσταση	Κύρια Συνιστώσα					
	1	2	3	4	5	6
Κεφαλή :						
Μήκος	0,74	0,45	0,49	-0,02	0,01	0,00
Πλάτος	0,70	0,59	-0,41	0,00	0,00	-0,01
Φτερό :						
Μήκος	0,95	-0,16	-0,03	0,22	0,05	0,16
Πλάτος	0,94	-0,21	0,01	0,20	-0,04	-0,17
Πόδι :						
Μήκος	0,93	-0,24	-0,04	-0,21	0,18	-0,03
Πλάτος	0,94	-0,19	-0,03	-0,20	-0,19	0,04
Ιδιοτιμή	4,568	0,714	0,412	0,173	0,076	0,057
% Ολικής Διακύμανσης	76,1	11,9	6,9	2,9	1,3	0,9

Με βάση τον πίνακα 5.4 μπορούμε να πούμε ότι η φυσική σημασία των κυρίων συνιστωσών είναι η εξής:

- 1^η Κύρια συνιστώσα : Μέσος όρος όλων των διαστάσεων.
- 2^η Κύρια συνιστώσα : Σύγκριση μεταξύ των διαστάσεων της κεφαλής και των διαστάσεων του φτερού και του ποδιού.
- 3^η Κύρια συνιστώσα : Σύγκριση μεταξύ των διαστάσεων της κεφαλής.
- 4^η Κύρια συνιστώσα : Σύγκριση μεταξύ των διαστάσεων του φτερού και του ποδιού.
- 5^η Κύρια συνιστώσα : Σύγκριση μεταξύ των διαστάσεων του ποδιού.
- 6^η Κύρια συνιστώσα : Σύγκριση μεταξύ των διαστάσεων του φτερού.



Παρατήρηση 5.1 Στο σημείο αυτό θα πρέπει να τονισθεί ότι η ερμηνεία των κυρίων συνιστωσών δεν είναι μια εύκολη υπόθεση, όπως φάνηκε στα Παραδείγματα 5.2 και 5.3. Αντίθετα τα Παραδείγματα αυτά αποτελούν μάλλον την εξαίρεση. Στην πράξη η ερμηνεία των κυρίων συνιστωσών είναι ένα αρκετά δύσκολο έργο. Πολλές φορές χρειάζεται να περιστρέψουμε τους άξονες των κυρίων συνιστωσών, κατα κάποια γωνία, για να πετύχουμε μια καλύτερη ερμηνεία.

5.5 ΕΛΕΓΧΟΣ ΓΙΑ ΤΗΝ ΠΟΛΥΔΙΑΣΤΑΤΗ ΚΑΝΟΝΙΚΗ

Όπως και στην μονοδιάστατη ανάλυση έτσι και στην πολυδιάστατη, βασικό ρόλο στην αντιμετώπιση των διαφόρων προβλημάτων παίζει η πολυδιάστατη κανονική κατανομή. Θα πρέπει δε να τονισθεί ότι στην πολυδιάστατη ανάλυση η υπόθεση της πολυδιάστατης κανονικής κατανομής είναι πιο απαραίτητη από ότι η υπόθεση της κανονικής κατανομής στην μονοδιάστατη περίπτωση. Και αυτό διότι στην μονοδιάστατη περίπτωση η υπόθεση της κανονικής κατανομής μπορεί να αντικατασταθεί με την υπόθεση ότι το δείγμα μας είναι αρκετά μεγάλο. Στην πολυδιάστατη ανάλυση κάτι τέτοιο δεν μπορεί να συμβεί.

Για τον έλεγχο της υπόθεσης ότι τα δεδομένα μας ακολουθούν την μονοδιάστατη κανονική κατανομή υπάρχουν πολλά και καλά τεστ (π.χ τεστ λοξότητας, τεστ κύρτωσης, τεστ Shapiro - Wilk, γραφική παράσταση σε κανονικό χαρτί πιθανοτήτων κ.τ.λ.). Αυτό δεν ισχύει όταν θέλουμε να ελέγξουμε ότι τα δεδομένα μας ακολουθούν την πολυδιάστατη κανονική κατανομή. Γενικεύσεις των μονοδιάστατων τεστ δίνονται από τους Mardia (1970), Malkovich & Afifi (1973) και Cox & Small (1978). Η αξιοπιστία όμως αυτών των τεστ δεν είναι ακόμη γνωστή. Ακόμη οι γραφικές μέθοδοι ελέγχου της πολυδιάστατης κανονικής που δίνονται από τους



Healy (1968), Small (1978) και Andrews et al (1973) χρειάζονται περαιτέρω μελέτη.

Στην παράγραφο αυτή θα παρουσιάσουμε μερικούς τρόπους ελέγχου για το αν τα δεδομένα μας ακολουθούν την πολυδιάστατη κανονική κατανομή.

Έστω ότι $x_1, x_2, \dots, x_n$ είναι ένα τυχαίο δείγμα από την $N_p(\mu, \Sigma)$. Τότε η ποσότητα

$$c_i = n(n-1)^{-2} (x_i - \bar{x})' S_u^{-1} (x_i - \bar{x}), \quad (i = 1, 2, \dots, n)$$

έχει την Βήτα κατανομή με παραμέτρους $\alpha = p/2$ και $\beta = (n-p-1)/2$. Αν με $c_{(1)} < c_{(2)} < \dots < c_{(n)}$ συμβολίσουμε το διατεταγμένο δείγμα των c_i , τότε ο έλεγχος ότι τα δεδομένα μας ακολουθούν την πολυδιάστατη κανονική κατανομή γίνεται με την γραφική παράσταση των c_i ως προς τα d_i . Αν U έχει την Βήτα κατανομή με παραμέτρους α και β , τότε τα d_i θα βρεθούν από την σχέση

$$\text{πιθ}(U \leq d_i) = (i-\alpha)/(n-\alpha-\beta+1),$$

όπου $a = (\alpha-1)/2$ και $b = (\beta-1)/(2\beta)$. Τα a και b εκλέγονται κατά τέτοιο τρόπο ώστε η αναμενόμενη τιμή των $c_{(i)}$ να είναι όσο πιο κοντά γίνεται στην τιμή των d_i . Αν η γραφική αυτή παράσταση είναι "περίπου" μια ευθεία που περνά από την αρχή των αξόνων και έχει κλίση μονάδα, τότε μπορούμε να πούμε ότι τα δεδομένα μας, προσεγγιστικά, ακολουθούν την πολυδιάστατη κανονική κατανομή.

Στην μονοδιάστατη περίπτωση ένα γνωστό τεστ για τον έλεγχο της κανονικότητας των δεδομένων μας είναι και το τεστ των Shapiro - Wilk (1965). Το τεστ αυτό έχει ως εξής : έστω ότι $x_1, x_2, \dots, x_n$ είναι ένα τυχαίο δείγμα με μέση τιμή $\bar{x}$ και διακύμανση s^2 . Το στατιστικό των Shapiro - Wilk είναι το



$$W = \sum_{i=1}^n \alpha_i x_{(i)} / s ,$$

όπου α_i είναι σταθερές που δίνονται από πίνακες των Shapiro - Wilk, και $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ είναι οι διατεταγμένες τιμές του τυχαίου δείγματος $x_1, x_2, \dots, x_n$. Η μηδενική υπόθεση της κανονικότητας των δεδομένων μας απορρίπτεται όταν $W \leq W_\alpha$, όπου οι τιμές του W_α , για επίπεδο σημαντικότητας α , δίνονται και πάλι από πίνακες των Shapiro - Wilk.

Για την πολυδιάστατη περίπτωση έστω $x_1, x_2, \dots, x_n$ ένα τυχαίο δείγμα και $y_1, y_2, \dots, y_n$ τα διανύσματα των κυρίων συνιστωσών που αντιστοιχούν στα $x_1, x_2, \dots, x_n$. Τότε το στατιστικό των Shapiro - Wilk είναι το

$$M = \sum_{j=1}^p \lambda_j^{-1} \left[\sum_{i=1}^n \alpha_i (y_{(i)j} - \bar{y}_j) \right] ,$$

όπου $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ είναι οι ιδιοτιμές του πίνακα των δειγματικών διακυμάνσεων - συνδιακυμάνσεων, α_i είναι σταθερές όπως και στην μονοδιάστατη περίπτωση και $y_{(i)j}$ είναι το i -στο διατεταγμένο στατιστικό στην j μεταβλητή. Οι παρατηρήσεις διατάσσονται χωριστά για κάθε μεταβλητή.

Ένας άλλος γραφικός τρόπος για τον έλεγχο της πολυδιάστατης κανονικής κατανομής βασίζεται στο γεγονός ότι αν οι αρχικές μας μεταβλητές ακολουθούν την πολυδιάστατη κανονική κατανομή, τότε η κατανομή κάθε κύριας συνιστώσας θα είναι η μονοδιάστατη κανονική. Επιπλέον η ανεξαρτησία των κυρίων συνιστωσών μας επιτρέπει την χρήση μονοδιάστατων τεχνικών ελέγχου κανονικής κατανομής.

Περαιτέρω στο θέμα αυτό ο αναγνώστης μπορεί να βρεί στους Kozíol (1986) και Gsörgö (1986,1989).



Α Σ Κ Η Σ Ε Ι Σ

5.1 Έστω x ένα p -διάστατο τυχαίο διάνυσμα με πίνακα διακυμάνσεων - συνδιακυμάνσεων Σ . Αν $\{\lambda_i\}$ είναι οι ιδιοτιμές και $\{a_i\}$ ($i = 1, 2, \dots, p$) τα ορθοκανονικά διανύσματα του Σ , τότε να δειχθεί ότι

$$I = \sum_{i=1}^p a_i a_i' \quad \text{και} \quad \Sigma = \sum_{i=1}^p \lambda_i a_i a_i'.$$

5.2 Έστω X_1 και X_2 δύο τυχαίες μεταβλητές με πίνακα διακυμάνσεων - συνδιακυμάνσεων

$$\Sigma = \begin{bmatrix} 9 & \sqrt{6} \\ \sqrt{6} & 4 \end{bmatrix}.$$

Να βρεθούν οι κύριες συνιστώσες καθώς και το ποσοστό της ολικής μεταβλητότητας που εξηγεί η κάθε μία. Να επαναλάβεται τα ίδια για τον πίνακα συσχέτισης και να διατυπωσεται τα συμπεράσματά σας.

5.3 Τα παρακάτω δεδομένα είναι η βαθμολογία 8 ατόμων σε δύο τεστ. Με βάση τα δεδομένα αυτά να δεχθεί ότι ο δειγματικός πίνακας διακυμάνσεων - συνδιακυμάνσεων είναι

$$S = \frac{1}{7} \begin{bmatrix} 208 & 144 \\ 144 & 292 \end{bmatrix}.$$

Στη συνέχεια να βρεθούν οι κύριες συνιστώσες του πίνακα S .

ΔΕΔΟΜΕΝΑ

α/α Ατόμου	1	2	3	4	5	6	7	8
Τεστ 1	2	16	8	18	10	4	10	12
Τεστ 2	9	20	0	11	13	4	7	16



5.4 Σε κάθε ένα από 500 ζώα μετράμε 4 μεγέθη. Οι ιδιοτιμές του δειγματικού πίνακα διακυμάνσεων - συνδιακυμάνσεων είναι 14,1 , 4,3 , 1,2 και 0,4 , ενώ τα δύο πρώτα ιδιοδιανύσματα είναι

$$\alpha'_1 = (0,39 \ 0,42 \ 0,44 \ 0,69) \text{ και } \alpha'_2 = (0,4 \ 0,39 \ 0,42 \ -0,72) .$$

Με βάση τα στοιχεία αυτά να διατυπώστε τα συμπεράσματά σας.

Εάν σας έδιναν μόνον τις ιδιοτιμές και τα ιδιοδιανύσματα από τα προηγούμενα δεδομένα να δείξετε πως μπορούσατε να προσδιορίσετε τον αρχικό πίνακα των δειγματικών διακυμάνσεων - συνδιακυμάνσεων.



6. - ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

6.1 ΕΙΣΑΓΩΓΗ

Οι στατιστικές μέθοδοι που αναπτύξαμε στα προηγούμενα κεφάλαια για την εκτίμηση παραμέτρων, έλεγχο υποθέσεων και διαστήματα εμπιστοσύνης βασίζονταν σε συγκεκριμένες υποθέσεις για τον πληθυσμό από τον οποίο προήλθε το δείγμα μας. Ένα άλλο πρόβλημα στην πολυδιάστατη ανάλυση, το οποίο συναντάται συχνά στις εφαρμοσμένες επιστήμες, είναι αυτό της ταξινόμησης μιας παρατήρησης σε έναν από δύο ή περισσότερους πληθυσμούς.

Για παράδειγμα ένας βοτανολόγος θέλει να ταξινομήσει ένα καινούργιο είδος λουλουδιού σε ένα από τα ήδη υπάρχοντα είδη μιας συνομοταξίας λουλουδιών. Στην εκπαίδευση ο υποψήφιος σε ένα πρόγραμμα θα πρέπει, με βάση τους βαθμούς του σε κάποια μαθήματα, να καταχωρηθεί σε μία από τις τρεις κατηγορίες "Δεκτός", "Δεκτός με περιορισμούς", "Απορριπτέος".

Η εύρεση κανόνων βέλτιστης ταξινόμησης καινούργιων παρατηρήσεων σε έναν από δύο ή περισσότερους καλώς διαχωρισμένους πληθυσμούς είναι το αντικείμενο της *διαχωριστικής ανάλυσης*.

Στο κεφάλαιο αυτό θα ασχοληθούμε κυρίως με κανόνες ταξινόμησης βασισμένους στην γραμμική διαχωριστική συνάρτηση (linear discriminant function).



6.2 ΠΕΡΙΠΤΩΣΗ ΔΥΟ ΠΛΗΘΥΣΜΩΝ

6.2.1 Ταξινόμηση σε δύο γνωστούς κανονικούς πληθυσμούς με κοινό πίνακα διακυμάνσεων-συνδιακυμάνσεων

Το πρόβλημα της ταξινόμησης μιας παρατήρησης σε έναν από δύο πληθυσμούς μπορεί να περιγραφεί ως εξής. Έστω Π_0 , Π_1 , Π_2 τρεις πληθυσμοί. Είναι γνωστό ότι $\Pi_0 = \Pi_i$ για ένα και μόνον τόνον ένα $i = 1, 2$. Το πρόβλημά μας είναι να εξετάσουμε για ποιο i αυτό ισχύει. Στη συνέχεια θα υποθέσουμε ότι οι πληθυσμοί μας ακολουθούν την p -διάστατη κανονική κατανομή με γνωστές μέσες τιμές μ_1 και μ_2 , αντίστοιχα, και γνωστό κοινό πίνακα διακυμάνσεων-συνδιακυμάνσεων Σ .

6.2.1.1 Διαχωριστική συνάρτηση του Fisher

Η πρώτη αντιμετώπιση του προηγούμενου προβλήματος οφείλεται στον Fisher (1936). Έστω ότι x_0 είναι μια p -διάστατη παρατήρηση την οποία θέλουμε να ταξινομήσουμε σε έναν από τους δύο πληθυσμούς Π_1 , Π_2 . Η πρόταση του Fisher για την ταξινόμηση αυτή είναι να βρούμε ένα γραμμικό συνδιασμό της παρατήρησης x_0 ο οποίος να μεγιστοποιεί το πηλίκον της διαφοράς των μέσων τιμών των γραμμικών συνδιασμών για τους πληθυσμούς Π_1 και Π_2 , ως προς την διακυμανσή του. Με άλλα λόγια έστω $y = a'x_0$ ο ζητούμενος γραμμικός συνδιασμός, όπου a είναι σταθερό, αλλά άγνωστο προς το παρών διάνυσμα. Αν τώρα η x_0 ανήκει στον πληθυσμό Π_1 , τότε η μέση τιμή του y θα είναι $a'\mu_1$, ενώ αν η x_0 ανήκει στον πληθυσμό Π_2 τότε η μέση τιμή του y θα είναι $a'\mu_2$. Επειδή οι δύο πληθυσμοί έχουν τον ίδιο πίνακα διακυμάνσεων-συνδιακυμάνσεων γι'



αυτό η διακύμανση του y θα είναι $a' \Sigma a$. Σύμφωνα με την πρόταση του Fisher η εύρεση της γραμμικής διαχωριστικής συνάρτησης, δηλαδή ο προσδιορισμός του διανύσματος a , θα γίνει από την μεγιστοποίηση της ποσότητας

$$\frac{(a' \mu_1 - a' \mu_2)^2}{a' \Sigma a} \quad . \quad (6.1)$$

Από την ανισότητα των Cauchy - Schwarz (βλ. Παράρτημα) παίρνουμε ότι

$$\frac{(a' \mu_1 - a' \mu_2)^2}{a' \Sigma a} \leq (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = \Delta^2, \quad (6.2)$$

όπου η ποσότητα Δ^2 είναι γνωστή με το όνομα *απόσταση του Mahalanobis* (Mahalanobis distance). Η ισότητα στην (6.2) ισχύει για

$$a = \lambda \Sigma^{-1} (\mu_1 - \mu_2),$$

όπου το λ είναι μια σταθερά αναλογίας. Χωρίς βλάβη της γενικότητας μπορούμε να θέσουμε $\lambda = 1$. Στην περίπτωση αυτή

$$a' x_0 = (\mu_1 - \mu_2)' \Sigma^{-1} x_0 \equiv L(x_0). \quad (6.3)$$

Η συνάρτηση $L(x_0)$ στην (6.3) είναι γνωστή με το όνομα *γραμμική διαχωριστική συνάρτηση του Fisher*.

Κανόνας ταξινόμησης : Με την βοήθεια της (6.3) η παρατήρηση x_0 ταξινομείται στον πληθυσμό Π_1 αν

$$L(x_0) > K, \quad (6.4)$$

διαφορετικά ταξινομείται στον πληθυσμό Π_2 . Το K στην (6.4) είναι μια σταθερά η οποία εκλέγεται κατά τέτοιο τρόπο ώστε να έχει κάποιες επιθυμητές ιδιότητες. Αν το K εκλεγεί έτσι ώστε

$$K = \frac{a' \mu_1 + a' \mu_2}{2} = \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2) \quad (6.5)$$

τότε, από την (6.4), σαν κανόνα ταξινόμησης μπορούμε να χρησιμο-



ποιήσουμε τον εξής: ταξινομούμε την παρατήρηση $\mathbf{x}_0$ στον πληθυσμό

$$\Pi_1 \text{ αν } \left(\mu_1 - \mu_2 \right)' \Sigma^{-1} \mathbf{x}_0 - \frac{1}{2} \left(\mu_1 - \mu_2 \right)' \Sigma^{-1} \left(\mu_1 + \mu_2 \right) > 0, \quad (6.6)$$

διαφορετικά η παρατήρηση $\mathbf{x}_0$ ταξινομείται στον πληθυσμό Π_2 .

Σφάλματα λανθασμένης ταξινόμησης : Η διαχωριστική ικανότητα της (6.3) μετράται από τα σφάλματα λανθασμένης ταξινόμησης. Σαν σφάλμα λανθασμένης ταξινόμησης θεωρούμε την πιθανότητα η παρατήρηση $\mathbf{x}_0$ να ταξινομηθεί στον πληθυσμό Π_2 ενώ $\mathbf{x}_0 \in \Pi_1$ και αντίστροφα. Έτσι αν με e_{ij} ($i \neq j = 1, 2$) συμβολίσουμε την πιθανότητα να ταξινομήσουμε την παρατήρηση $\mathbf{x}_0$ στον πληθυσμό Π_i , δοθέντος ότι $\mathbf{x}_0 \in \Pi_j$, τότε θα έχουμε

$$e_{12} = P \left[\left(\mu_1 - \mu_2 \right)' \Sigma^{-1} \mathbf{x}_0 - \frac{1}{2} \left(\mu_1 - \mu_2 \right)' \Sigma^{-1} \left(\mu_1 + \mu_2 \right) > 0 \mid \mathbf{x}_0 \in \Pi_2 \right], \quad (6.7\alpha)$$

και

$$e_{21} = P \left[\left(\mu_1 - \mu_2 \right)' \Sigma^{-1} \mathbf{x}_0 - \frac{1}{2} \left(\mu_1 - \mu_2 \right)' \Sigma^{-1} \left(\mu_1 + \mu_2 \right) < 0 \mid \mathbf{x}_0 \in \Pi_1 \right]. \quad (6.7\beta)$$

Εάν εκλέξουμε το K όπως στην (6.5) τότε εύκολα προκύπτει ότι

$e_{12} = e_{21}$. Ακόμη μπορούμε να δείξουμε (η απόδειξη αφήνεται σαν άσκηση) ότι

$$e_{12} = e_{21} = \Phi \left(-\frac{1}{2} \Delta^2 \right), \quad (6.8)$$

όπου Φ συμβολίζει την αθροιστική συνάρτηση της τυπικής κανονικής κατανομής, δηλαδή

$$\Phi(x) = \int_{-\infty}^x (2\pi)^{-1/2} e^{(-1/2)t^2} dt.$$

Παρατήρηση 6.1 Η σχέση (6.6) μπορεί να τροποποιηθεί έτσι ώστε να λάβουμε υπ' όψιν μας το τυχόν κόστος από μια λάθος ταξινόμηση καθώς και τις τυχόν πληροφορίες σχετικά με την πιθανότητα η



παρατήρησή μας να ανήκει σε έναν από τους δύο πληθυσμούς. Για τα θέματα αυτά καθώς και άλλα σχετικά με την γραμμική διαχωριστική συνάρτηση βλέπε Srivastava & Carter (1983) καθώς και Dillon & Goldstein (1984).

6.2.1.2 Μέθοδος του πηλίκου πιθανοφάνειας

Η μέθοδος αυτή οφείλεται στον Welch (1939) και έχει ως εξής. Έστω ότι $f_1(\mathbf{x}; \theta_1)$ είναι η κατανομή του πληθυσμού Π_1 και $f_2(\mathbf{x}; \theta_2)$ η κατανομή του πληθυσμού Π_2 . Έστω ακόμη ότι $\mathbf{x}_0$ είναι μια καινούργια παρατήρηση προς ταξινόμηση. Αν συμβολίσουμε με

$$\lambda = \frac{f_1(\mathbf{x}_0; \theta_1)}{f_2(\mathbf{x}_0; \theta_2)} \quad (6.9)$$

το πηλίκο πιθανοφάνειας των δύο κατανομών τότε η παρατήρηση $\mathbf{x}_0$ θα ανήκει στην πληθυσμό Π_1 αν $\lambda > 1$ διαφορετικά θα ανήκει στον πληθυσμό Π_2 .

Μπορεί να δειχθεί (βλ. άσκηση 6.3) ότι στην περίπτωση της πολυδιάστατης κανονικής κατανομής οι μέθοδοι των Welch και Fisher μας δίνουν την ίδια διαχωριστική συνάρτηση.

6.2.2 Ταξινόμηση σε δύο κανονικούς πληθυσμούς με άγνωστες μέσες τιμές αλλά γνωστό τον κοινό πίνακα διακυμάνσεων-συνδιακυμάνσεων

Στην παράγραφο αυτή θα υποθέσουμε ότι οι μέσες τιμές μ_1 και μ_2 των δύο πληθυσμών Π_1 και Π_2 , αντίστοιχα, είναι άγνωστες αλλά ο κοινός πίνακας των διακυμάνσεων-συνδιακυμάνσεων Σ είναι γνωστός. Στην περίπτωση αυτή θα πρέπει να εκτιμήσουμε τα μ_1 και μ_2 . Έστω λοιπόν ένα τυχαίο δείγμα $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+m}$, όπου οι πρώτες n παρατηρήσεις προέρχονται από τον πληθυσμό Π_1 ,



ενώ οι υπόλοιπες m παρατηρήσεις από τον πληθυσμό Π_2 .

Αν με

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{και} \quad \bar{x}_2 = \frac{1}{m} \sum_{i=n+1}^m x_i$$

συμβολίσουμε τις δειγματικές μέσες τιμές των πληθυσμών Π_1 και Π_2 αντίστοιχα, τότε, σύμφωνα με την (6.6), μια παρατήρηση x_0 θα ταξινομείται στον πληθυσμό Π_1 αν

$$(\bar{x}_1 - \bar{x}_2)' \Sigma^{-1} x_0 - \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' \Sigma^{-1} (\bar{x}_1 + \bar{x}_2) > 0 \quad (6.10)$$

ή με βάση την άσκηση 6.1, αν

$$(x_0 - \bar{x}_1)' \Sigma^{-1} (x_0 - \bar{x}_1) < (x_0 - \bar{x}_2)' \Sigma^{-1} (x_0 - \bar{x}_2). \quad (6.11)$$

Τα ασυμπτωτικά σφάλματα λανθασμένης ταξινόμησης e_{12} και e_{21} (βλ. (6.7α) και (6.7β)) δίνονται από τον Okamoto (1963) ως εξής:

$$e_{21} = \Phi(-\Delta/2) + \alpha_1 n^{-1} + \alpha_2 m^{-1} \quad (6.12\alpha)$$

και

$$e_{12} = \Phi(-\Delta/2) + \alpha_1 m^{-1} + \alpha_2 n^{-1} \quad (6.12\beta)$$

όπου

$$\alpha_1 = (2\Delta^2)^{-1} (d_4 + 3pd_2) \quad \text{και} \quad \alpha_2 = (2\Delta^2)^{-1} [d_4 - (p-4)d_2]$$

με
$$d_2 = (2\pi)^{-1/2} \frac{1}{2} \Delta \exp(-\Delta^2/2)$$

και
$$d_4 = (2\pi)^{-1/2} \left(\frac{1}{8} \Delta^3 - \frac{3}{2} \Delta \right) \exp(-\Delta^2/2).$$

Είναι προφανές ότι οι τιμές των e_{12} και e_{21} όπως δίνονται από τις σχέσεις (6.12αβ) δεν μπορούν να υπολογισθούν διότι εξαρτώνται από την ποσότητα Δ η οποία στην συνέχεια εξαρτάται από την ποσότητα $\delta = \mu_1 - \mu_2$. Ένας εκτιμητής του Δ είναι ο

$$\hat{\Delta}^2 = (\bar{x}_1 - \bar{x}_2)' \Sigma^{-1} (\bar{x}_1 - \bar{x}_2).$$

Για την εκτίμηση των σφαλμάτων e_{12} και e_{21} , όπως αυτά δίνονται από τις σχέσεις (6.12αβ), έχουν κατά καιρούς προταθεί διάφορες μέθοδοι. Δυστυχώς οι μόνες συγκρίσεις που υπάρχουν για τις μεθόδους αυτές είναι υπολογιστικής υφής (π.χ. Monte Carlo).



Αυτό έχει σαν αποτέλεσμα να μην είναι δυνατή η επιλογή της "καλύτερης" μεθόδου σε κάθε περίπτωση. Στη συνέχεια θα περιγράψουμε τρεις τέτοιες μεθόδους. Οι μέθοδοι αυτές έχουν το πλεονέκτημα να μην εξαρτώνται από την υπόθεση της κανονικής κατανομής.

Μέθοδος της αντικατάστασης (Resubstitution method)

Σύμφωνα με τη μέθοδο αυτή, οφείλεται στον Hills (1966), κάθε παρατήρηση, στους δύο πληθυσμούς, παίρνει το ρόλο της x_0 και ταξινομείται σύμφωνα με τον κανόνα ταξινόμησης (π.χ. τον κανόνα (6.9)) σε έναν από τους πληθυσμούς Π_1 ή Π_2 . Γνωρίζοντας τον πληθυσμό από τον οποίο προέρχεται κάθε παρατήρηση είναι εύκολο να υπολογίσουμε το ποσοστό των παρατηρήσεων που ενώ ανήκουν σε ένα πληθυσμό ταξινομούνται στον άλλο. Τα ποσοστά αυτά χρησιμο-ποιούνται σαν εκτιμητές $\hat{e}_{12}$ και $\hat{e}_{21}$ των e_{12} και e_{21} , αντίστοιχα. Οι εκτιμητές αυτοί είναι συνεπείς (δηλ. $\text{Var}(\hat{e}_{12}) \rightarrow 0$ και $\text{Var}(\hat{e}_{21}) \rightarrow 0$ όταν το μέγεθος του δείγματος τείνει στο ∞) αλλά το ποσόν της μεροληψίας είναι αρκετά μεγάλο.

Μέθοδος της ενδοεπικύρωσης (Cross - Validation method).

Η ιδέα της μεθόδου αυτής οφείλεται στον Lachenbruch (1967) και έχει ως εξής. Αφαιρούμε μια παρατήρηση από έναν οποιοδήποτε πληθυσμό. Με τις υπόλοιπες $n+m-1$ παρατηρήσεις προσδιορίζουμε την διαχωριστική συνάρτηση. Στην συνέχεια χρησιμοποιώντας αυτή την διαχωριστική συνάρτηση ταξινομούμε την παρατήρηση που αφαιρέσαμε. Η διαδικασία αυτή επαναλαμβάνεται έως ότου όλες οι παρατηρήσεις μας ταξινομηθούν. (Εξυπακούεται ότι κάθε φορά που αφαιρούμε μια παρατήρηση επανατοποθετούμε την προηγούμενη).

Αν με m_{ij} ($i \neq j = 1, 2$) συμβολίσουμε τον αριθμό των παρατηρήσεων που ταξινομούνται στον πληθυσμό Π_i ενώ ανήκει στον πληθυσμό Π_j



τότε οι εκτιμητές των σφαλμάτων e_{12} και e_{21} θα είναι αντίστοιχα $\hat{e}_{12} = m_{12} / n$ και $e_{21} = m_{21} / m$. Οι εκτιμητές των e_{12} και e_{21} που προκύπτουν από την μέθοδο αυτή είναι σχεδόν αμερόληπτοι. Παρ' όλο που η μέθοδος αυτή χρησιμοποιείται ευρύτατα στις εφαρμογές εν τούτοις μια σειρά από εργασίες (Toussant (1974), McLachlan (1974) και Glick (1978)) έχουν αναφερθεί με αυστηρότητα στην διακύμανση των εν λόγω εκτιμητών. Ιδιαίτερα ο Glick στην εργασία του δείχνει αφ' ενός μεν την μεγάλη ελάττωση του ποσού της μεροληψίας των εκτιμητών $\hat{e}_{12}$ και $\hat{e}_{21}$ που επιτυγχάνει η μέθοδος αυτή, αφ' ετέρου δε την μεγάλη διακύμανση (ή μέσο τετραγωνικό σφάλμα) των εκτιμητών αυτών.

Μέθοδος του διαχωρισμού (Holdout method).

Σύμφωνα με τη μέθοδο αυτή το αρχικό δείγμα χωρίζεται σε δύο επί μέρους δείγματα. Το ένα δείγμα χρησιμοποιείται για τον προσδιορισμό του κανόνα ταξινόμησης ενώ το άλλο για την εύρεση του ποσοστού των παρατηρήσεων που προέρχονται από ένα πληθυσμό και ταξινομούνται στον άλλο. Τα ποσοστά αυτά χρησιμοποιούνται, όπως στην και στις προηγούμενες μεθόδους, σαν εκτιμητές των e_{12} και e_{21} .

Οι εκτιμητές αυτοί είναι συνεπείς και αμερόληπτοι. Το μειονέκτημα της μεθόδου αυτής είναι ότι χρειάζεται μεγάλο μέγεθος δείγματος. Όταν το δείγμα μας δεν είναι μεγάλο τότε η χρησιμοποίηση της μεθόδου αυτής έχει σαν αποτέλεσμα ή λάθος προσδιορισμό του κανόνα ταξινόμησης ή λάθος εκτίμηση των e_{12} και e_{21} .

Αν χρησιμοποιήσουμε την μέθοδο του πηλίκου πιθανοφάνειας (βλ. σχέση (6.9)) τότε μπορούμε να δείξουμε (βλ. άσκηση 6.5)) ότι ο κανόνας ταξινόμησης θα είναι ο :



παρατήρηση $\mathbf{x}_0$ στον πληθυσμό Π_1 αν

$$(1+n^{-1})^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)\Sigma^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1) < (1+m^{-1})^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2)\Sigma^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2), \quad (6.13)$$

διαφορετικά την ταξινομούμε στον πληθυσμό Π_2 .

Παρατηρήστε ότι η διαχωριστική συνάρτηση (6.13)

(i) λαμβάνει υπ' όψιν της τα διαφορετικά μεγέθη των δειγμάτων και (ii) είναι δευτέρου βαθμού ως προς $\mathbf{x}_0$. Η ιδιότητα (ii) της (6.13) έχει σαν αποτέλεσμα να είναι δύσκολος ο υπολογισμός των πιθανοτήτων e_{12} και e_{21} . Σύμφωνα με τον John (1961) τα e_{12} και e_{21} δίνονται από τις σχέσεις

$$e_{21} = P\left[(1+n^{-1})^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)\Sigma^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1) > (1+m^{-1})^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2)\Sigma^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2)\right]$$

$$|\mathbf{x}_0 \in \Pi_1| = P\left[F_{p,p}(\lambda_1, \lambda_2) > 1\right],$$

όπου $F_{p,p}(\lambda_1, \lambda_2)$ είναι η διπλώς μη - κεντρική F κατανομή με p και p ελευθερίας και μη κεντρικές παραμέτρους λ_1 και λ_2 . Οι τιμές των λ_1 και λ_2 είναι

$$\lambda_1 = \Delta^2(1-c^{1/2})/2c(1+m^{-1}) \quad \text{και} \quad \lambda_2 = \Delta^2(1+c^{1/2})/2c(1+m^{-1})$$

με

$$c = 1 - \left[(1+n^{-1})(1+m^{-1})\right]^{-1} \quad \text{και} \quad \Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2).$$

$$e_{12} = P\left[(1+n^{-1})^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1)\Sigma^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_1) < (1+m^{-1})^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2)\Sigma^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_2)\right]$$

$$|\mathbf{x}_0 \in \Pi_2| = P\left[F_{p,p}(\lambda_1^*, \lambda_2^*) > 1\right],$$

με $\lambda_1^* = b\lambda_1$, $\lambda_2^* = b\lambda_2$ και $b = (1+n^{-1})^{-1}(1+m)^{-1}$.

Ένας πίνακας της διπλώς μη-κεντρικής F - κατανομής δίνεται από τον Tiku (1975). Η εκτίμηση των e_{12} και e_{21} γίνεται με μία από



τις μεθόδους που αναφέραμε προηγουμένως.

Παρατήρηση 6.2 Αν, κάνοντας χρήση της διαχωριστικής συνάρτησης (6.13), έχουμε n_0 παρατηρήσεις για ταυτόχρονη ταξινόμηση τότε στην (6.13) το $\mathbf{x}_0$ θα πρέπει να αντικατασταθεί με το $\bar{\mathbf{x}}_0$, τη μέση των n_0 παρατηρήσεων, το $(1+n^{-1})$ με το $(n_0^{-1}+n^{-1})$ και το $(1+m^{-1})$ με το $(n_0^{-1}+m^{-1})$.

Παράδειγμα 6.1 Οι μέσες τιμές από τα αποτελέσματα τριών ψυχολογικών τεστ σε ασθενείς που πάσχουν από άγχος ή υστερία φαίνονται στον επόμενο πίνακα.

Ασθένεια	Τεστ 1	Τεστ 2	Τεστ 3	Μέγεθος Δείγματος
Άγχος	2,9298	1,667	0,7281	114
Υστερία	3,0303	1,2424	0,5455	33

Με την βοήθεια τόσο των παραπάνω μετρήσεων όσο και άλλων βρήκαμε ότι ο από κοινού πίνακας των δειγματικών διακυμάνσεων-συνδιακυμάνσεων είναι ο

$$\mathbf{S} = \begin{bmatrix} 2,3008 & 0,2516 & 0,4742 \\ 0,2516 & 0,6075 & 0,0358 \\ 0,4742 & 0,0358 & 0,5951 \end{bmatrix}$$

με 250 βαθμούς ελευθερίας. Ο μεγάλος αριθμός των βαθμών ελευθερίας του πίνακα $\mathbf{S}$ μας επιτρέπει να χρησιμοποιήσουμε τον $\mathbf{S}$ σαν τον αληθινό πίνακα διακυμάνσεων-συνδιακυμάνσεων Σ .

Αν με $\mathbf{x}' = (x_1, x_2, x_3)$ συμβολίσουμε τις τιμές των τριών τεστ για κάποιον ασθενή, τότε, σύμφωνα με την διαχωριστική συνάρτηση του Fisher, ο ασθενής αυτός θα ταξινομείται σαν αγχωτικός αν

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) > 0$$

διαφορετικά θα ταξινομείται σαν υστερικός. Από τον πίνακα των δεδομένων έχουμε ότι



$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 2,9298 \\ 1,6670 \\ 0,7281 \end{bmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{bmatrix} 3,0303 \\ 1,2424 \\ 0,5455 \end{bmatrix}.$$

Αντικαθιστώντας στην τελευταία σχέση παίρνουμε ότι ένας ασθενής θα χαρακτηρίζεται σαν αγχοτικός αν

$$-0,217 x_1 + 0,764 x_2 + 0,069 x_3 > 0,512$$

διαφορετικά θα χαρακτηρίζεται σαν υστερικός.

Για να βρούμε το κατά πόσον η διαχωριστική αυτή συνάρτηση είναι αξιόπιστη θα πρέπει να υπολογίσουμε τις πιθανότητες λανθασμένης ταξινόμησης όπως αυτές δίνονται από τις σχέσεις (6.12αβ). Ο εκτιμητής του $\hat{\Delta}^2$ θα είναι ο $\hat{\Delta}^2 = 0,359$, $d_2 = 0,10$, $d_4 = -0,29$, $\alpha_1 = 0,85$, $\alpha_2 = -0,26$, $n = 114$ και $m = 33$. Συνεπώς

$$\begin{aligned} e_{21} &= \Phi(-\hat{\Delta}^2/2) + \alpha_1 n^{-1} + \alpha_2 m^{-1} = \\ &= 0,4032 + 0,0074 - 0,0079 = 0,4027, \\ e_{12} &= \Phi(-\hat{\Delta}^2/2) + \alpha_2 n^{-1} + \alpha_1 m^{-1} = \\ &= 0,4032 - 0,0022 - 0,0257 = 0,4267. \end{aligned}$$

Οι προηγούμενες τιμές λανθασμένης ταξινόμησης είναι αρκετά μεγάλες. Αυτό σημαίνει ότι δεν θα είναι σωστό να χαρακτηρίζουμε έναν ασθενή σαν αγχοτικό ή υστερικό με μόνο στοιχείο τις τιμές στα τρία προηγούμενα ψυχολογικά τεστ.

6.2.3 Ταξινόμηση σε δύο κανονικούς πληθυσμούς με άγνωστες παραμέτρους αλλά κοινό πίνακα διακυμάνσεων-συνδιακυμάνσεων

Στην περίπτωση αυτή και οι τρεις άγνωστοι παράμετροι μ_1, μ_2 και Σ θα πρέπει να εκτιμηθούν από το δείγμα (βλ. § 6.2.2). Τότε ο κανόνας ταξινόμησης του Fisher μας δίνει: η παρατήρηση $\mathbf{x}_0$ θα ταξινομείται στον πληθυσμό Π_1 εάν

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} \mathbf{x}_0 - \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) > 0$$



ή ισοδύναμα εάν

$$(\mathbf{x}_0 - \bar{\mathbf{x}}_1)' \mathbf{S}_u^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_1) < (\mathbf{x}_0 - \bar{\mathbf{x}}_2)' \mathbf{S}_u^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_2) \quad (6.14)$$

διαφορετικά θα ταξινομείται στον πληθυσμό Π_2 .

Τα ασυμπτωτικά σφάλματα λανθασμένης ταξινόμησης οφείλονται στον Okamoto (1963) και δίνονται από τις επόμενες σχέσεις :

$$e_{21} = \Phi(-\Delta^2/2) + \alpha_1 n^{-1} + \alpha_2 m^{-1} + \alpha_3 (n+m-2)^{-1} \quad (6.15\alpha)$$

$$e_{12} = \Phi(-\Delta^2/2) + \alpha_2 n^{-1} + \alpha_1 m^{-1} + \alpha_3 (n+m-2)^{-1} \quad (6.15\beta)$$

όπου $\alpha_3 = p(p-1)d_2$ και οι τιμές των α_1 , α_2 και d_2 είναι όπως και στην παράγραφο 6.2.2 (βλ. σχέσεις 6.12αβ). Η εκτίμηση των σφαλμάτων αυτών γίνεται με την αντικατάσταση του Δ^2 με την δειγματική απόσταση του Mahalanobis $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_u^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. Ακόμη μπορούν να χρησιμοποιηθούν οι μέθοδοι που περιγράψαμε στην § 6.2.2 και οι οποίες δεν εξαρτώνται από την υπόθεση της κανονικής κατανομής.

Κάνοντας χρήση του πηλίκου πιθανοφάνειας βρίσκουμε τον εξής κανόνα ταξινόμησης : η παρατήρηση $\mathbf{x}_0$ θα ταξινομείται στον πληθυσμό Π_1 αν

$$(1+n^{-1})^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_1)' \mathbf{S}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_1) < (1+m^{-1})^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_1)' \mathbf{S}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_1) \quad (6.16)$$

διαφορετικά θα ταξινομείται στον πληθυσμό Π_2 .

Στην περίπτωση αυτή τα σφάλματα λανθασμένης ταξινόμησης δεν είναι εύκολο να υπολογισθούν και έτσι για την εκτίμησή τους καταφεύγουμε σε μία από τις μεθόδους της παραγράφου 6.2.2 .

Παράδειγμα 6.2 Σε καθ' ένα από δύο τύπους λουλουδιών, Α και Β, πήραμε τρεις μετρήσεις X_1, X_2, X_3 , όλες με την ίδια μονάδα μέτρησης (το μικρό, μ). Τα αποτελέσματα παρουσιάζονται στον επόμενο πίνακα.



Τύπος Α			Τύπος Β		
X_1	X_2	X_3	X_1	X_2	X_3
191	131	53	186	107	49
185	134	50	211	122	49
200	137	52	201	144	47
173	127	50	242	131	54
171	128	49	184	108	43
160	118	47	211	118	51
188	134	54	217	122	49
186	129	51	223	127	51
174	131	52	208	125	50
163	115	47	199	124	46

Από τον πίνακα αυτό βρίσκουμε $\mathbf{x}'_A = (179,1 \quad 127,4 \quad 50,5)$,
 $\mathbf{x}'_B = (208,2 \quad 119,8 \quad 48,9)$ ενώ ο από κοινού πίνακας των
 δειγματικών διακυμάνσεων-συνδιακυμάνσεων είναι ο

$$S = \begin{bmatrix} 231,25 & 103,83 & 34,15 \\ 103,83 & 61,67 & 15,66 \\ 34,15 & 15,66 & 7,41 \end{bmatrix}.$$

Επειδή $n = m$ ο κανόνας ταξινόμησης του Fisher (βλ.σχέση (6.14)) συμπίπτει με τον κανόνα ταξινόμησης του πηλίκου μέγιστης πιθανοφάνειας (βλ. σχέση (6.16)). Αντικαθιστώντας τις τιμές των $\bar{\mathbf{x}}'_A$, $\bar{\mathbf{x}}'_B$ και S στην (6.14) ή στην (6.16) παίρνουμε ότι ένα λουλούδι με μετρήσεις $\mathbf{x}'_0 = (x_1, x_2, x_3)$ θα χαρακτηρίζεται σαν τύπου Α αν

$$-1,06 x_1 + 1,34 x_2 + 2,31 x_3 > 72,42.$$

Ο εκτιμητής του Δ^2 είναι ο $\hat{\Delta}^2 = 50,23$ ενώ από τις σχέσεις (6.15αβ) βρίσκουμε ότι $e_{21} = e_{12} = 0,0003$.

6.2.4 Ταξινόμηση σε δύο γνωστούς κανονικούς πληθυσμούς με διαφορετικούς πίνακες διακυμάνσεων-συνδιακυμάνσεων

Έστω $\mathbf{x}_0$ μια παρατήρηση από κάποιον πληθυσμό. Είναι γνωστό ότι η $\mathbf{x}_0$ προέρχεται είτε από τον πληθυσμό Π_1 , με κατανομή την



$Np(\mu_1, \Sigma_1)$, ή από τον πληθυσμό Π_2 , με κατανομή την $Np(\mu_2, \Sigma_2)$. Αν όλες οι παράμετροι είναι γνωστές τότε ο κανόνας ταξινόμησης που παίρνουμε από το πηλίκο πιθανοφάνειας (βλ. και άσκηση (6.5)) είναι ο: η παρατήρηση x_0 ταξινομείται στον πληθυσμό Π_1 αν

$$(x_0 - \mu_1)' \Sigma_1^{-1} (x_0 - \mu_1) < (x_0 - \mu_2)' \Sigma_2^{-1} (x_0 - \mu_2) + C, \quad (6.17)$$

διαφορετικά ταξινομείται στον πληθυσμό Π_2 . Η τιμή της σταθεράς C , στην (6.17) είναι

$$C = \log_e [\det(\Sigma_2)] - \log_e [\det(\Sigma_1)].$$

Αν οι παράμετροι $\mu_1, \mu_2, \Sigma_1, \Sigma_2$ είναι άγνωστοι τότε η συνηθισμένη πρακτική είναι να χρησιμοποιούμε την (6.17) αφού πρώτα αντικαταστήσουμε τις άγνωστες παραμέτρους με τους αντίστοιχους εκτιμητές τους.

Η εκτίμηση των σφαλμάτων λανθασμένης ταξινόμησης γίνεται με μία από τις μεθόδους της § 6.2.2.

6.3 ΠΕΡΙΠΤΩΣΗ ΤΡΙΩΝ Ή ΠΕΡΙΣΣΟΤΕΡΩΝ ΠΛΗΘΥΣΜΩΝ

Στην περίπτωση των δύο πληθυσμών η προσπάθειά μας ήταν να χωρίσουμε το δειγματικό χώρο σε δύο ξένους μεταξύ τους χώρους. Το ρόλο του διαχωρισμού έπαιξε η διαχωριστική συνάρτηση. Γίνεται αμέσως φανερό ότι όταν έχουμε τρεις ή περισσότερους πληθυσμούς τότε μια διαχωριστική συνάρτηση δεν είναι αρκετή. Μπορεί να δειχθεί (βλ. Dilon & Goldstein, 1984) ότι ο αριθμός των διαχωριστικών συναρτήσεων που είναι απαραίτητες για τον διαχωρισμό K ($K \geq 2$) πληθυσμών ισούται με το $\min(p, K-1)$, όπου p είναι ο αριθμός των μεταβλητών του προβλήματός μας.



Έστω ότι $Np(\mu_i, \Sigma)$ είναι η κατανομή του i ($i = 1, 2, \dots, K$) πληθυσμού. Αν συμβολίσουμε με $\bar{x}_i$ τη δειγματική μέση τιμή του i πληθυσμού και με S_u τον από κοινού αμερόληπτο εκτιμητή του Σ , τότε η παρατήρηση x_0 ταξινομείται στον πληθυσμό i εάν και μόνον εάν

$$T_i = \min_{1 \leq j \leq K} T_j,$$

όπου $T_j = (x_0 - \bar{x}_j)' S_u^{-1} (x_0 - \bar{x}_j)$. Γράφοντας $c'_j = \bar{x}'_j S_u^{-1}$ και $c_j = -\frac{1}{2} \bar{x}'_j S_u^{-1} \bar{x}_j$, ($j = 1, 2, \dots, K$), τότε ο κανόνας ταξινόμησης (6.16) μπορεί να γραφεί σαν

$$c'_i x_0 + c_i = \max_{1 \leq j \leq K} (c'_j x_0 + c_j) \quad (6.19)$$

Τύποι για τον υπολογισμό των πιθανοτήτων λανθασμένης ταξινόμησης για την περίπτωση αυτή δεν υπάρχουν. Η εκτίμησή τους γίνεται με μία από τις μεθόδους της παραγράφου 6.2.2.

Παράδειγμα 6.3 Ας θεωρήσουμε τα δεδομένα του παραδείγματος 6.2 μαζί με τις μετρήσεις σε ένα τρίτο τύπο λουλουδιών, Γ , όπως φαίνονται στον επόμενο πίνακα.

Τ Υ Π Ο Σ Γ	X_1	158	146	151	122	138	132	131	135	125	130
	X_2	141	119	130	113	121	115	127	123	119	120
	X_3	58	51	51	45	53	49	51	50	51	48

Ο από κοινού πίνακας των δειγματικών διακυμάνσεων - συνδιακυμάνσεων, και για τους τρεις τύπους λουλουδιών, είναι

$$S = \begin{bmatrix} 198,67 & 94,13 & 32,82 \\ 94,13 & 63,24 & 18,08 \\ 32,82 & 18,08 & 8,72 \end{bmatrix}.$$

Οι τιμές των διανυσμάτων c_i ($i = 1, 2, 3$) και σταθερών c_i είναι



$$c'_1 = (-0,0655 \quad 1,496 \quad 5,150) \quad , \quad c_1 = -166,7$$

$$c'_2 = (0,129 \quad 0,562 \quad 3,962) \quad , \quad c_2 = -144$$

$$c'_3 = (-1,426 \quad 2,070 \quad 6,892) \quad , \quad c_3 = -205,1$$

Με βάση τις παραπάνω τιμές η ταξινόμηση μιας νέας παρατήρησης, $x'_0 = (x_1, x_2, x_3)$, σε έναν από τους τρεις πληθυσμούς θα γίνει σύμφωνα με τον κανόνα ταξινόμησης (6.19)

Παρατήρηση 6.3 Το κεφάλαιο της διαχωριστικής ή ταξινομικής ανάλυσης είναι αρκετά μεγάλο. Εδώ καλύψαμε ένα ελάχιστο τμήμα του. Θα πρέπει όμως να τονίσουμε ότι εκτός από τις μεθόδους που παρουσιάσαμε, και που όλες έχουν σαν βασική τους υπόθεση ότι οι κατανομές των πληθυσμών είναι κανονικές, υπάρχουν και μέθοδοι που δεν βασίζονται σε καμία υπόθεση για την κατανομή των πληθυσμών. Οι μέθοδοι αυτές είναι γνωστές με το όνομα *μη παραμετρικές*. (Βλ. π.χ. Hawkins (1982), Mardia, Kent & Bibby (1979), Puri & Sen (1971)).



Α Σ Κ Η Σ Ε Ι Σ

6.1 Να δικαιολογηθεί ο κανόνας ταξινόμησης (6.6)

6.2 Να δειχθεί ότι

$$(\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{x}_0 - \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)' = \frac{1}{2} (T_2 - T_1),$$

όπου $T_i = (\mathbf{x}_0 - \mu_i)' \Sigma^{-1} (\mathbf{x}_0 - \mu_i)$ ($i = 1, 2$).

Στην συνέχεια να διατυπώσετε τον κανόνα ταξινόμησης.

6.3 Να δειχθεί ότι στην περίπτωση της πολυδιάστατης κανονικής κατανομής με γνωστά μ και Σ η μέθοδος του Welch και του Fisher μας δίνουν την ίδια διαχωριστική συνάρτηση.

6.4 Με βάση τις σχέσεις (6.7 αβ) να αποδειχθεί η σχέση (6.8).

6.5 Να αποδειχθεί η σχέση (6.11).



ΠΑΡΑΡΤΗΜΑ

Στο παράρτημα αυτό θα παρουσιάσουμε, χωρίς αποδείξεις, μερικά αποτελέσματα από την θεωρία των πινάκων τα οποία είναι απαραίτητα για τις ανάγκες των σημειώσεων αυτών. Ο αναγνώστης μπορεί να βρει τις αποδείξεις στο βιβλίο των Mardia, Kent και Bibby (1982).

Έστω X ένας $n \times p$ πίνακας και $f(X)$ μια συνάρτηση του X . Η μερική παράγωγος της $f(X)$ ως προς X ορίζεται σαν ο πίνακας του οποίου το (i, j) στοιχείο είναι το $\partial f(X) / \partial x_{ij}$, όπου x_{ij} είναι το (i, j) στοιχείο του πίνακα X . Δηλαδή

$$\frac{\partial f(X)}{\partial X} = \left(\frac{\partial f(X)}{\partial x_{ij}} \right). \quad (\Pi.1)$$

Με βάση την σχέση (Π.1) μπορούμε να δείξουμε ότι

- i) $\frac{\partial a'x}{\partial x} = a$.
- ii) $\frac{\partial x'x}{\partial x} = 2x$, $\frac{\partial x'Ax}{\partial x} = 2Ax$, $\frac{\partial x'Ay}{\partial x} = Ay$.
- iii) $\frac{\partial \det(X)}{\partial x_{ij}} = X_{ij}$, εάν όλα τα στοιχεία του $n \times n$ πίνακα X είναι διακεκριμένα.

$$= \begin{cases} X_{ii} & , i = j \\ 2X_{ij} & , i \neq j \end{cases} \text{ εάν ο } X \text{ είναι συμμετρικός.}$$

Ο πίνακας X_{ij} είναι ο συμπαράγων του (i, j) στοιχείου του πίνακα X .



$$\text{iv)} \quad \frac{\partial \text{tr}(XY)}{\partial X} = Y', \quad \text{εάν όλα τα στοιχεία του } n \times n \text{ πίνακα } X \text{ είναι διακεκριμένα.}$$

$$= \dot{Y} + Y' - \text{diag}(Y), \quad \text{εάν ο } n \times n \text{ πίνακας } X \text{ είναι συμμετρικός}$$

$$\text{v)} \quad \frac{\partial X^{-1}}{\partial x_{ij}} = -X^{-1} J_{ij} X^{-1}, \quad \text{εάν όλα τα στοιχεία του } n \times n \text{ πίνακα } X \text{ είναι διακεκριμένα.}$$

$$= \begin{cases} -X^{-1} J_{ii} X^{-1}, & i = j \\ -X^{-1} (J_{ij} + J_{ji}) X^{-1}, & i \neq j \end{cases} \quad \text{εάν ο πίνακας } X \text{ είναι συμμετρικός}$$

Με J_{ij} συμβολίζουμε τον πίνακα με το στοιχείο 1 στην (i, j) θέση και 0 οπουδήποτε αλλού.

Θεώρημα. Έστω A και B δύο $n \times n$ συμμετρικοί πίνακες με τον πίνακα B θετικά ορισμένο ($B > 0$). Τότε το μέγιστο (ελάχιστο) του $x'Ax$ δοθέντος ότι

$$x'Bx = 1 \quad (\Pi.2)$$

επιτυγχάνεται όταν το x είναι το ιδιοδιάνυσμα του πίνακα $B^{-1}A$ που αντιστοιχεί στην μεγαλύτερη (μικρότερη) ιδιοτιμή του πίνακα $B^{-1}A$. Συνεπώς αν με λ_1 και με λ_p συμβολίσουμε, αντίστοιχα, την μικρότερη και την μεγαλύτερη ιδιοτιμή του πίνακα $B^{-1}A$, τότε, με τον περιορισμό (Π.2) θα έχουμε ότι

$$\max_x x'Ax = \lambda_1 \quad \text{και} \quad \min_x x'Ax = \lambda_p.$$

Πόρισμα 1. Αν $R(x) = x'Ax/x'Bx$, τότε, για κάθε $x \neq 0$,

$$\lambda_p \leq R(x) \leq \lambda_1.$$



Πόρισμα 2. Το μέγιστο του $\mathbf{a}'\mathbf{x}$, με τον περιορισμό (Π.2), είναι το

$$\left(\mathbf{a}'\mathbf{B}^{-1}\mathbf{a}\right)^{1/2}.$$

Επιπλέον

$$\max_{\mathbf{x}} \frac{(\mathbf{a}'\mathbf{x})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a},$$

και το μέγιστο επιτυγχάνεται για

$$\mathbf{x} = \frac{\mathbf{B}^{-1}\mathbf{a}}{(\mathbf{a}'\mathbf{B}^{-1}\mathbf{a})}.$$



B I B Λ Ι Ο Γ Ρ Α Φ Ι Α

- Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis*. 2^η Έκδοση, John Wiley & Sons.
- Andrews, D. F. Gnanadesekan, R. Warner, J. L. (1973) *Methods for Assesing Multivariate Normality*, Multivariate Analysis (P. R. Krishnaiah, Editor), Academic Press.
- Arnold, S. F. (1981) *The Theory of Linear Models and Multivariate Analysis*. John Wiley & Sons.
- Beal, E. M. L. Kendall, M. G. Mann, D. W. (1967) *The discarding of variables in multivariate analysis*, Biometrika, Τόμος 54, σελ. 357 - 366.
- Box, G. E. P. (1949) *A general distribution theory for a class of likelihood criteria*, Biometrika, Τόμος 36, σελ. 317 - 346.
- Cattell, R. B. (1966) *The scree test for the numbers of factors*, Multivariate Behav. Res., Τόμος 1, σελ. 245 - 276.
- Cattell, R. B. Jaspers, J. (1967) *A general plasmode for factor analytic exercises and research*, Multivariate Behavioral Research Monographs, Τόμος 67, σελ. 1 - 212 .
- Chatterjee, S. Price, B. (1977) *Regression Analysis by Examples*, John Wiley & Sons.
- Cox, D. R. Small, N. J.H. (1978) *Testing multivariate normality*, Biometrika, Τόμος 65, σελ. 263 - 272 .
- Dillon, W. R. Goldstein, M. (1984) *Multivariate Analysis. Methods and Applications*, John Wiley & Sons .



- Eaton, M. L. (1972) *Multivariate Statistical Analysis* ,
Institute of Mathematics and Statistics, University of
Copenhagen, Denmark.
- Fisher, R. A. (1936) *The use of multiple measurments in
taxonomic problems*, Ann. Eugen., Τόμος 7, σελ. 107 - 116 .
- Giri, N. (1977) *Multivariate Statistical Inference*, Academic
Press.
- Glick, N. (1978) *Additive estimators of probabilities of correct
classification*, Pattern Recognition, Τόμος 10, σελ.211-222.
- Gsörgö, S. (1986) *Testing for normality in arbitrary dimension*,
Annals of Statistics, Τόμος 14, σελ. 708 - 723 .
- Gsörgö, S. (1989) *Consistency of some tests for multivaraitaite
normality*, Metrika, Τόμος 36, σελ. 107 - 116 .
- Halmos, P. L. Savege, L. J. (1949) *Application of Radon-Nicodym
theorem of the sufficient statistics*, Ann. Math. Statist. ,
Τόμος 20, σελ. 225 - 241 .
- Hawkins, D. M. (1963) *Topics in Multivariate Analysis*, Cambridge
University Press .
- Healy, M. J. R. (1968) *Multivariate normal plotting*, Appl.
Statist. , Τόμος 17, σελ. 157 - 161 .
- Hills, M. (1966) *Allocation rules and their error rates*. Journal
of the Royal Statistical Society - B, Τόμος 28, σελ. 1 .
- Horn, J. L. (1965) *A rationale and test for the number of factors
in factor analysis*, Psychometrika, Τόμος 30, σελ. 179 .
- Hotelling, H. (1933) *Analysis of a complex of statistical
variables into principal components*, Journal of Educ.
Psych. , Τόμος 24, σελ. 417 - 441 & 498 - 520 .
- John, S. (1961) *Errors in disrimination*, Ann. Math. Statist.,
Τόμος 32, σελ. 1125 - 1144 .



- Jolliffe, I. T. (1972) *Discarding variables in principal components analysis I : artificial data*, Applied Statist., Τόμος 22, σελ. 160 - 173 .
- Jolliffe, I. T. (1973) *Discarding variables in principal components analysis II : real data*, Applied Statistics, Τόμος 23, σελ. 21 - 31 .
- Kaiser, H. F. (1958) *The varimax criterion for analytic rotation in factor analysis*, Psychometrika, Τόμος 23, σελ. 187 - 200
- Καρακώστας, Κ. Ε. (1992) *Παλινδρόμηση και Ανάλυση Διακύμανσης*, Παν/μιο Ιωαννίνων .
- Kendall, M. G. , Stuart, A. (1973) *The Advanced Theory of Statistics* , Τόμος 3, Griffin .
- Korin, B. P. (1968) *On the distribution of a statistic used for testing a covariance matrix*, Biometrika, Τόμος 55, σελ. 171 - 178 .
- Koziol, J. A. (1986) *Assesing multivariate normality : a compendium*, Communications in Statistics - Theory and Methods, Τόμος 15, σελ. 2763 - 2783 .
- Lachenbruch, P. A. (1967) *An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis*, Biometrics, Τόμος 23, σελ. 639 - 645 .
- MacLachlan, G. J. (1974) *An asymptotic unbiased technique for estimating the error rates in discriminant analysis*, Biometrics, Τόμος 20, σελ. 472 - 479 .
- Malkovich, J. R. Afifi, A. A. (1973) *On tests for multivariate normality*, Journal of Amer. Statist. Soc., Τόμος 68, σελ. 176 - 179 .



- Mardia, K. V. (1970) *Measures of multivariate skewness and kurtosis with applications*, Biometrika, Τόμος 57, σελ. 519 - 530 .
- Mardia, K. V. Kent, J. T. Bibby, J. M. (1982) *Multivariate Analysis*, Academic Press.
- Morrison, D. F. (1976) *Multivariate Statistical Methods*, 2^η Έκδοση , McGraw - Hill .
- Okamoto, M. (1963) *An asymptotic expansion for the linear discriminant function*, Ann. Math. Statist., Τόμος 39, σελ. 1358 .
- Παπαϊωάννου, Τ. Φερεντίνος, Κ. (1983) *Μαθηματική Στατιστική*, Παν/μιο Ιωαννίνων.
- Pearson, K. (1901) *On lines and planes of closest fit to systems of points in space*, Phil. Mag., Τόμος 2, σελ. 559 - 572 .
- Pearson, E. S. Hartley, H. O. (1972) *Biometrika Tables for Statisticians*, Τόμος 2, Cambridge University Press.
- Puri, M. L. Sen, P. K. (1971) *Nonparametric Methods in Multivariate Analysis*, John Wiley & Sons.
- Rao, C. R. (1964) *The use and interpretation of principal components analysis in applied research*, Sankhya - A, Τόμος 26, σελ. 329 - 358 .
- Roy, S. N. (1957) *Some Aspects of Multivariate Analysis*, John Wiley.
- Shapiro, S. S. Wilk, M. B. (1965) *An analysis of variance test for normality (complete samples)*, Biometrika, Τόμος 52, σελ. 591 - 611 .
- Silvey, S. D. (1970) *Statistical Inference*, Penguin.
- Small, N. J. H. (1978) *Plotting square radii*, Biometrika, Τόμος 65, σελ. 657 - 658 .



- Srivastava, M. S. Carter, E. M. (1983) *An Introduction to Applied Multivariate Statistics*, North - Holland .
- Tiuu, M. L. (1963) *Selected Tables in Mathematical Statistics*, Τόμος 2, Inst. Math. Statist., Amer. Math. Soc., Providence, R. I.
- Toussant, G. T. (1974) *Bibliography on estimation of misclassification*, IEEE Transactions on Information Theory, Τόμος 20, σελ. 472 - 479 .
- Welch, B. L: (1939) *Note on discriminant functions*, Biometrika, τόμος 31, σελ. 218 - 220 .



Τυπώθηκε στο Πανεπιστημιακό Τυπογραφείο με δαπάνη
του Πανεπιστημίου Ιωαννίνων

Κ.Α. Πανεπιστημιακού Τυπογραφείου.

Ο.Τ.

Copyright : Πανεπιστήμιο Ιωαννίνων

Απαγορεύεται η μερική ή ολική ανατύπωση, καθώς και η
λήψη φωτοαντιγράφων από το βιβλίο χωρίς τη γραπτή
άδεια του Τμήματος Δημοσιευμάτων του Πανεπιστημίου
Ιωαννίνων και του συγγραφέα.

Διατίθεται και στο Βιβλιοπωλείο του Πανεπιστημίου
Ιωαννίνων, Δομπόλη, 451 10 Ιωάννινα τηλ. 21801.

ΔΙΑΝΕΜΕΤΑΙ ΔΩΡΕΑΝ στους φοιτητές.



Το παρόν έγγραφο αποτελεί μέρος της συλλογής των βιβλίων της βιβλιοθήκης του Πανεπιστημίου Ιωαννίνων.

Η Α. Γραμματεία του Πανεπιστημίου Ιωαννίνων

Τ.Ο.

Οι υπεύθυνοι της βιβλιοθήκης

Αναγράφεται η ημερομηνία η οποία αφορά την απόκτηση του βιβλίου από την βιβλιοθήκη. Το βιβλίο αυτό ανήκει στην βιβλιοθήκη του Πανεπιστημίου Ιωαννίνων και είναι υπό την προστασία του νόμου. Οποιαδήποτε χρήση του βιβλίου χωρίς την άδεια της βιβλιοθήκης αποτελεί παράβαση του νόμου.

