ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ»

ΤΗΣ ΚΑΛΚΟΥΝΟΥ ΔΗΜΗΤΡΑΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΦΙΛΙΠΠΟΣ ΑΛΕΒΙΖΟΣ

ΕΤΟΣ 2013

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ 4

Πολυμεταβλητά Περιγραφικά Μέτρα 6

Πολυμεταβλητά Δεδομένα 6

Μέτρα Θέσης και Μεταβλητότητας στην Πολυμεταβλητή Ανάλυση 8

Μέτρα Θέσης 8

Μέτρα Μεταβλητότητας 9

Συντελεστής συσχέτισης 13

Μερικός Συντελεστής Συσχέτισης 17

Τρόποι Πολυμεταβλητής Ανάλυσης 19

Πολυδιάστατη Κατανομή Τυχαίων Μεταβλητών 20

Η πολυμεταβλητή Κανονική Πυκνότητα και οι Ιδιότητές της 26

Διάγραμμα Διμεταβλητής Κανονικής Πυκνότητας 33

Επιπλέον Ιδιότητες της Πολυμεταβλητής Κανονικής Κατανομής 37

Εκτίμηση παραμέτρων της πολυδιάστατης Κανονικής Κατανομής 54

Η δειγματική κατανομή των $\bar{x}$ και S, - κατανομή Wishart 54

Αμερόληπτοι εκτιμητές ελάχιστης διασποράς 61

Θεώρημα Lehmann και Scheffé 61

Παραγοντικό Κριτήριο Επάρκειας 62

Μέθοδος Μέγιστης Πιθανοφάνειας 64

Η συμπεριφορά των $\bar{x}$ και S σε μεγάλο δείγμα 70

Δειγματοληψία από μια Πολυμεταβλητή κανονική κατανομή και τη Μέγιστη Πιθανοφάνεια 74

Η πολυμεταβλητή κανονική πιθανότητα 74

Εκτίμηση Μέγιστης Πιθανοφάνειας του μ και Σ 78

Επαρκή Στατιστικά Στοιχεία..... 83

Δεσμευμένες και Περιθωριακές Κατανομές Πολυμεταβλητών Μεταβλητών

..... 85

Η δεσμευμένη πυκνότητα της πολυμεταβλητής κανονικής κατανομής..... 86

Έλεγχος υποθέσεων..... 90

Η αξιοπιστία του μ_0 ως αξία για μια Κανονική μέση τιμή του πληθυσμού

..... 92

Η κατανομή του Hotelling..... 95

Κατανομή T^2 Hotteling και λόγος Πιθανοφανειών..... 100

Γενική Μέθοδος Λόγου Πιθανότητας..... 105

Περιοχή Εμπιστοσύνης και ταυτόχρονες συγκρίσεις συνιστωσών..... 108

Ταυτόχρονες Δηλώσεις Εμπιστοσύνης..... 113

Έλεγχοι Διανυσμάτων Μέσων τιμών..... 118

Γνωστός πίνακας διακύμανσης..... 118

Άγνωστος Πίνακας Διακύμανσης..... 119

Συγκρίνοντας Διανύσματα Μέσης Τιμής από δυο Πληθυσμούς..... 127

Υπόθεση σχετικά με τη δομή των δεδομένων..... 128

Περισσότερες υποθέσεις όταν τα n_1 και n_2 είναι μικρά..... 128

Η κατάσταση του διδιάστατου δείγματος όταν $\Sigma_1 \neq \Sigma_2$ 135

Βιβλιογραφία..... 141

ΕΙΣΑΓΩΓΗ

Όλοι οι επιστήμονες οποιοδήποτε κι αν είναι το αντικείμενο με το οποίο ασχολούνται, όταν καλούνται να μελετήσουν για παράδειγμα ένα φυσικό φαινόμενο, ή έναν οργανισμό είναι υποχρεωμένοι να συλλέξουν ένα πλήθος δεδομένων πάνω στο οποίο θα δουλέψουν. Το τελικό προϊόν της έρευνας είναι συνήθως ένα σύνολο μετρήσεων που έγιναν σε ένα σύνολο πειραματικών μονάδων, ανεξάρτητα από το αντικείμενο προς μελέτη ή το πειραματικό σχέδιο που ακολουθήθηκε. Το σύνολο των πειραματικών μονάδων στον δικό μας κλάδο μπορεί να εμφανιστεί με τις ονομασίες όπως αντικείμενο, δείγματα, παρατηρήσεις, νομάδες. Υπάρχουν βέβαια και οι μεταβλητές, οι οποίες είναι ουσιαστικά οι διαφορετικές μετρήσεις που έχουν γίνει διαδοχικά στις λεγόμενες πειραματικές μονάδες, μέσα από τις οποίες ύστερα από διάφορες διαδικασίες θα εξαχθούν τα αποτελέσματα τα οποία θα μελετηθούν.

Γενικότερα, όταν πειραματιζόμαστε με ένα αντικείμενο κάθε φορά θα εξάγουμε μόνο ένα συμπέρασμα, χρησιμοποιώντας κάποιες στατιστικές μεθόδους. Αν η στατιστική μέθοδος που εφαρμόζεται σε μια μεταβλητή κάθε φορά, ονομάζεται μονομεταβλητή μέθοδος (univaviate method).

Αν χρειαστεί να μελετήσουμε την ταυτόχρονη δράση όλων των μεταβλητών μαζί; Αυτό δεν μπορεί να γίνει με τη βοήθεια της μονομεταβλητής μεθόδου, αλλά με τη βοήθεια της πολυμεταβλητής μεθόδου (multivariate method).

Τι γίνεται όμως όταν θα χρειαστεί να πειραματιστούμε με περισσότερα από ένα αντικείμενα;

Η πολυμεταβλητή μέθοδος (multivariate method) ονομάζεται η στατιστική ανάλυση που εφαρμόζεται ταυτόχρονα σε όλες τις εξεταζόμενες μεταβλητές ενός πειράματος.

Στην πολυμεταβλητή στατιστική το ενδιαφέρον δεν εστιάζεται στην κατανομή κάθε τυχαίας μεταβλητής $X_1,\ i=1,...,p$ ξεχωριστά, αλλά στη μελέτη πολλών τυχαίων μεταβλητών $X_1, X_2, ..., X_p$ συγχρόνως.

Δηλαδή, το ενδιαφέρον πια, εστιάζεται στη μελέτη της από κοινού κατανομής των τυχαίων μεταβλητών, για τις οποίες υποθέτουμε ότι υπάρχει εξάρτηση μεταξύ τους και αυτή την εξάρτηση ουσιαστικά θέλουμε να μελετήσουμε.

Η θεμελιώδης έννοια της πολυμεταβλητής στατιστικής, είναι η από κοινού συνάρτηση πυκνότητας πιθανότητας, για τις συνεχείς τυχαίες μεταβλητές και της από κοινού συνάρτησης πιθανότητας, για την περίπτωση των διακριτικών τυχαίων μεταβλητών.

Πολυμεταβλητά Περιγραφικά Μέτρα

Πολυμεταβλητά Δεδομένα

Συνήθως τα δεδομένα οργανώνονται υπό τη μορφή πινάκων δεδομένων. Οι στατιστικοί πίνακες και οι γραφικές παραστάσεις αποτελούν χρήσιμα μέσα για να παρουσιάσουμε τα δεδομένα καθαρά και με σαφήνεια. Επίσης, μπορούν να αποκαλύψουν σημαντικά χαρακτηριστικά των δεδομένων, όπως το εύρος τους, τη συμμετρικότητα τους ή την ύπαρξη ακραίων τιμών.

Παρακάτω εμφανίζεται ο πίνακας δεδομένων που έχει ως γραμμές τις παρατηρήσεις του δείγματος και ως στήλες τις μεταβλητές του.

	Χαρακτηριστικό ή Μεταβλητή 1	Χαρακτηριστικό ή Μεταβλητή 2	…	Χαρακτηριστικό ή Μεταβλητή p
Παρατήρηση 1	X_{11}	X_{12}	…	X_{1p}
Παρατήρηση 2	X_{21}	X_{22}	…	X_{2p}
⋮				
Παρατήρηση n	X_{n1}	X_{n2}	…	X_{np}

Όπου

n : Αριθμός παρατηρήσεων/αντικειμένων προς μελέτη

p : Αριθμός χαρακτηριστικών/μεταβλητών προς μελέτη

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

Ο πίνακας Δεδομένων (Data Matrix) έχει την εξής μορφή:

$$X = \begin{bmatrix} \mathrm{X}_{11} & \mathrm{X}_{12} & \dots & \mathrm{X}_{1p} \\ \mathrm{X}_{21} & \mathrm{X}_{22} & \dots & \mathrm{X}_{2p} \\ \vdots & \vdots & \dots & \vdots \\ \mathrm{X}_{n1} & \mathrm{X}_{n2} & \dots & \mathrm{X}_{np} \end{bmatrix}$$

και είναι ένας πίνακας διάστασης nxp.

Παρατηρούμε ότι, για κάθε παρατήρηση δεν έχουμε μόνο μια τιμή, αλλά έχουμε ένα διάνυσμα τιμών.

Συνήθως, από τη Γραμμική Άλγεβρα, κάθε παρατήρηση είναι ένα διάνυσμα $px1$, δηλαδή, το διάνυσμα στήλη (π.χ. το διάνυσμα X_1 αντιστοιχεί στην πρώτη παρατήρηση). Τελικά, ο πίνακας δεδομένων απαρτίζεται από τα διανύσματα γραμμής και είναι της μορφής:

$$\mathbf{X} = \begin{Bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{Bmatrix}$$

Συνεπώς, $\mathbf{X}'_{\mathrm{i}} = \left(x_{i1}, x_{i2}, ..., x_{ip} \right)$ είναι η i γραμμή του πίνακα και αντιστοιχεί στις τιμές των μεταβλητών για την i-παρατήρηση. Συνοπτικά, μπορούμε να πούμε, ότι ο πίνακας δεδομένων περιέχει τα διανύσματα γραμμής ως παρατηρήσεις.

Μέτρα Θέσης και Μεταβλητότητας στην Πολυμεταβλητή Ανάλυση

Όταν εξετάζουμε δεδομένα, είτε στην πρωτογενή τους μορφή, είτε ομαδοποιημένα, υπολογίζουμε και κάποια μέτρα, που σκοπό έχουν να περιγράψουν με περιληπτικό τρόπο, τα βασικά χαρακτηριστικά τους και γι' αυτό ονομάζονται περιληπτικά ή περιγραφικά μέτρα. Τα μέτρα θέσης και μεταβλητότητας, αποτελούν αριθμητικά περιγραφικά μέτρα και μας βοηθούν στο να επεξεργαστούμε στατιστικά δεδομένα, ενώ είμαστε βασισμένοι σε μικρό πλήθος αριθμητικών στοιχείων, σε σχέση με αυτά που είχαμε αρχικά. Έτσι, έχουμε τη δυνατότητα να συγκρίνουμε διαφορετικούς πληθυσμούς ή δείγματα μεταξύ τους.

Μέτρα Θέσης

Τα μέτρα θέσης δίνουν περιληπτικά τη θέση των δεδομένων επάνω στην ευθεία των πραγματικών αριθμών. Ως μέτρα θέσης εννοούμε κυρίως τα μέτρα κεντρικής τάσης που προσδιορίζουν ένα κεντρικό σημείο, γύρω από το οποίο τείνουν να συγκεντρώνονται τα δεδομένα.

Το πιο γνωστό και διαδεδομένο μέτρο θέσης είναι η μέση τιμή. Σε πολυδιάστατα δεδομένα η αντίστοιχη γενίκευση είναι το διάνυσμα των μέσων τιμών. Η δειγματική μέση τιμή συμβολίζεται με $\bar{x}$ και ορίζεται ως

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Για το διάνυσμα των μέσων ισχύουν όλα όσα ισχύουν και για την απλή μέση τιμή.

Άλλα μέτρα θέσης, όπως η διάμεσος ή ο περικεκομμένος μέσος, που είναι διαδεδομένα στη μονομεταβλητή περίπτωση, στην πολυμεταβλητή είναι δύσκολο να οριστούν, γι' αυτό και δεν χρησιμοποιούνται:

Για παράδειγμα, η διάμεσος στη μονομεταβλητή περίπτωση είναι η τιμή που χωρίζει το δείγμα στη μέση, όταν αυτό είναι διατεταγμένο. Στην πολυμεταβλητή,

όμως, αυτό δεν είναι εύκολο να γίνει και αυτό γιατί θα έπρεπε να έχουμε ένα διάνυσμα με τις διαμέσους κάθε μεταβλητής.

Από την άλλη, ο περιεκεκομμένος μέσος αποτελεί ένα εναλλακτικό μέτρο θέσης, ιδιαίτερα όταν υπάρχουν ακραίες τιμές που επηρεάζουν σημαντικά τη μέση τιμή. Ο μέσος αυτός προκύπτει, όταν αφαιρέσουμε από μερικές ακραίες τιμές και από τις δυο ουρές της κατανομής. Για να γίνει αυτό στην πολυμεταβλητή περίπτωση, είμαστε υποχρεωμένοι να αγνοήσουμε κάποιες παρατηρήσεις, δηλαδή, ποιες είναι οι ακραίες παρατηρήσεις. Αυτό, όμως, είναι δύσκολο, αφού δεν είναι εύκολο στην πολυμεταβλητή περίπτωση να ορίσουμε ακραίες τιμές στα πολυμεταβλητά δεδομένα. Η λύση είναι να βρούμε περικεκομμένους μέσους για κάθε μεταβλητή και μετά να συνθέσουμε το διάνυσμα των περικεκομμένων μέσων. Με αυτή, όμως, τη διαδικασία, ορίζουμε ξεχωριστά για κάθε μεταβλητή τις ακραίες τιμές και αγνοούμε οποιαδήποτε συσχέτιση.

Μέτρα Μεταβλητότητας

Εκτός από την κεντρική τάση μας ενδιαφέρει επίσης και η μεταβλητότητα ή διασπορά των παρατηρήσεων, επειδή, οι πληροφορίες που παρέχουν τα μέτρα θέσης για την κατανομή ενός δείγματος, δεν είναι επαρκή. Όταν τα δεδομένα είναι συγκεντρωμένα γύρω από μια κεντρική τιμή, δηλαδή, η διασπορά των δεδομένων είναι μικρή, τότε η κεντρική τιμή αντιπροσωπεύει ικανοποιητικά τα δεδομένα. Από την άλλη, όταν τα δεδομένα είναι πολύ σκορπισμένα, τα μέτρα κεντρικής τιμής δε δίνουν καλή περιληπτική περιγραφή των δεδομένων. Τα μέτρα μεταβλητότητας ελέγχουν, πόσο μακριά είναι οι τιμές των δεδομένων από τα μέσα κεντρικής τάσης, ή πόσο απλωμένα είναι. Επίσης, διαφορετικά δείγματα από τον ίδιο πληθυσμό μπορεί να έχουν το ίδιο μέτρο κεντρικής τάσης, αλλά, να διαφέρουν κατά κάποιο σημαντικό τρόπο ως προς τη διασπορά των παρατηρήσεων. Τα πιο συχνά χρησιμοποιούμενα μέτρα μεταβλητότητας είναι το εύρος, η διακύμανση, η τυπική απόκλιση και το ενδοτεταρτημοριακό εύρος.

Ξεκινάμε πρώτα απ’ όλα με τον ορισμό του εύρους.

Το εύρος είναι η διαφορά της ελάχιστης από τη μέγιστη τιμή του δείγματος.

Εκφράζεται με τον τύπο: $R = X\max - X\min$

Το εύρος υπολογίζεται εύκολα, αλλά, δεν είναι ανθεκτικό μέτρο μεταβλητότητας, γιατί βασίζεται μόνο στις ακραίες παρατηρήσεις και δεν επηρεάζεται από την κατανομή των ενδιάμεσων τιμών. Γενικά, το εύρος αυξάνει όταν μεγαλώνει το δείγμα, καθώς αναμένεται να συμπεριληφθούν πιο ακραίες τιμές.

Το βασικότερο μέτρο μεταβλητότητας στην μονοδιάστατη περίπτωση είναι η διακύμανση. Η διακύμανση είναι ο μέσος όρος των τετραγωνικών αποκλίσεων των τιμών από το μέσο αριθμητικό. Μετράει τη μεταβλητότητα των παρατηρήσεων γύρω από τη μέση τιμή.

Συμβολίζεται με s^2 και δίνεται από τον τύπο:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$

Η διακύμανση εκφράζεται σε τετραγωνικές μονάδες και τα δεδομένα που παίρνουμε δεν έχουν νόημα.

Γι' αυτό ορίζουμε τη δειγματική τυπική απόκλιση s, που είναι απλά η θετική ρίζα της δειγματικής διασποράς s^2. Η τυπική απόκλιση s μετριέται με τη μονάδα μέτρησης της τ. μ X και εκφράζει την τυπική απόκλιση των δεδομένων από τη δειγματική μέση τιμή, δηλαδή, μέχρι πόσο περίπου περιμένουμε μια τυπική τιμή της X να απέχει από τη μέση τιμή.

Η τυπική απόκλιση είναι η τετραγωνική ρίζα της διακύμανσης. Συμβολίζεται με s και δίνεται από τον τύπο:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Λαμβάνοντας την τυπική απόκλιση λοιπόν, θα πάρουμε μέτρο διασποράς, το οποίο θα έχει την ίδια μονάδα μέτρησης με το αρχικό δείγμα.

Για τα ζεύγη τιμών ορίζουμε τη συνδιακύμανση. Η συνδιακύμανση είναι ένα μέτρο μεταβλητότητας ως προς δυο μεταβλητές, που δείχνει πως συμμεταβάλλονται

αυτές. Ποιο συγκεκριμένα, ορίζουμε τη δειγματική συνδιακύμανση των μεταβλητών X_j και X_k που δίνεται από το παρακάτω τύπο:

$$\hat{Cov}\left(X_j, X_k\right) = S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$$

Στα πολυμεταβλητά δεδομένα ορίζουμε τον πίνακα διακύμανσης – συνδιακύμανσης. Είναι ένας συμμετρικός θετικά ημιορισμένος ή θετικά ορισμένος πίνακας. Στη διαγώνιο του έχει τις διακυμάνσεις των μεταβλητών και στα υπόλοιπα στοιχεία τις συνδιακυμάνσεις των μεταβλητών που αντιστοιχούν σε κάθε γραμμή και στήλη.

Το ενδοτεταρτημοριακό εύρος Q, ορίζεται ως η διαφορά του πρώτου τεταρτημόριου Q_1 από το τρίτο τεταρτημόριο Q_3, δηλαδή:

$$Q = Q_3 - Q_1$$

Το ενδοτεταρτημοριακό εύρος δίνει το εύρος που καλύπτουν τα μισά από τα δεδομένα που είναι πιο κοντά στην κεντρική τιμή (διάμεσο). Το Q είναι ένα άλλο μέτρο διασποράς των δεδομένων, που δεν ορίζεται ως προς τη δειγματική μέση τιμή και άρα δεν επηρεάζεται απ' αυτήν. Παρατηρούμε ότι στο μεταξύ τους διάστημα $Q_3 - Q_1$ περιλαμβάνεται το 50% των παρατηρήσεων. Όσο μικρότερο είναι το διάστημα αυτό, τόσο μεγαλύτερη θα είναι η συγκέντρωση των τιμών, επομένως, μικρότερη η διασπορά των τιμών της μεταβλητής Χ. Όμως, το πρόβλημα παραμένει το ίδιο με το εύρος, δηλαδή, το ότι αγνοούνται οι τιμές μεταξύ του Q_1 και του Q_3, όπως αγνοούνται και οι υπόλοιπες τιμές της X, εκτός από τις τιμές Q_1 και Q_3.

Έχει υπολογιστεί το ενδοτεταρτημοριακό εύρος $Q(X)$ και $Q(Y)$ δυο μεταβλητών X και Y, που δεν μετριούνται με τις ίδιες μονάδες μέτρησης. Για να γίνει η σύγκριση των διασπορών, καταφεύγουμε στο σχετικά ενδοτεταρτημοριακά εύρη διαιρώντας τα διά την αντίστοιχη διάμεσο.

$$CQ(X) = \frac{Q_3(X) - Q_1(X)}{M(X)}$$

όπου: $Q(X) = Q_3(X) - Q_1(X)$

και $CQ(Y) = \dfrac{Q_3(Y) - Q_1(Y)}{M(Y)}$

όπου: $Q(Y) = Q_3(Y) - Q_1(Y)$

Ως σχετικό ενδοτεταρτημοριακό εύρος ορίζεται ο λόγος του ενδοτεταρτημοριακού εύρους Q δια της διαμέσου τιμής M .

$$CQ = \frac{Q}{M} \quad \text{ή} \quad CQ = \frac{Q_3 - Q_1}{M}$$

Συντελεστής συσχέτισης

Η συσχέτιση είναι αυτή που μετρά την αλληλεπίδραση ανάμεσα σε δυο ή περισσότερες μεταβλητές.

Ο συντελεστής είναι ένας δείκτης και πρακτικά σημαίνει ότι από την τιμή ενός δείκτη κατανοούμε πόσο έντονη ή χαλαρή είναι η συσχέτιση των μεταβλητών.

Το είδος και το μέγεθος της γραμμικής σχέσης μεταξύ δυο ποσοτικών μεταβλητών που ακολουθούν την κανονική κατανομή εκτιμάται με τον συντελεστή συσχέτισης Pearson, ενώ όταν έστω και μια από τις ποσοτικές μεταβλητές δεν ακολουθεί την κανονική κατανομή, τότε χρησιμοποιείται ο συντελεστής συσχέτισης διατάξεων του Spearman.

Ο συντελεστής συσχέτισης του Pearson συμβολίζεται με p όταν αφορά ολόκληρο τον πληθυσμό και με r όταν υπολογίζεται για ένα συγκεκριμένο «δείγμα» (αποτελώντας έτσι εκτίμηση για τον πληθυσμό από τον οποίο προέρχεται το «δείγμα»). Ο συντελεστής συσχέτισης του Pearson δεν έχει μονάδες μέτρησης και λαμβάνει τιμές από -1 έως $+1$. Οι τιμές -1 έως $+1$ προκύπτουν όταν υπάρχει μια τέλεια (αρνητική ή θετική αντίστοιχα) γραμμική σχέση μεταξύ των δυο μεταβλητών. Όταν ο συντελεστής συσχέτισης πάρει την τιμή 0, σημαίνει ότι δεν υπάρχει σχέση γραμμική μεταξύ των δυο μεταβλητών, αλλά δεν αποκλείεται κιόλας η ύπαρξη μιας άλλης μη γραμμικής σχέσης.

Εάν το μέγεθος μιας μεταβλητής τείνει να αυξάνεται όπως αυξάνεται το μέγεθος και της άλλης μεταβλητής, τότε υπάρχει θετική συσχέτιση (positive correlation) μεταξύ των δυο μεταβλητών και ο συντελεστής συσχέτισης είναι μεγαλύτερος του μηδενός. Αντίθετα, αν το μέγεθος της μιας μεταβλητής τείνει να αυξάνεται, όπως ελαττώνεται το μέγεθος της άλλης μεταβλητής, τότε υπάρχει αρνητική συσχέτιση (negative correlation) μεταξύ των δυο μεταβλητών και ο συντελεστής συσχέτισης είναι μικρότερος του μηδενός.

Όμως, ο συντελεστής συσχέτισης του Pearson παρουσιάζει κάποια μειονεκτήματα. Ποσοτικοποιεί τη σχέση μεταξύ δυο μεταβλητών μόνο στην περίπτωση κατά την οποία η σχέση αυτή είναι γραμμική. Επίσης είναι πολύ ευαίσθητος σε ακραίες τιμές

και η ύπαρξη τους σε αρκετές περιπτώσεις οδηγεί σε λανθασμένα συμπεράσματα. Ακόμα η εκτιμηθείσα συσχέτιση δεν μπορεί να επεκταθεί πέρα από το παρατηρημένο εύρος των μεταβλητών. Τέλος, η ύπαρξη μιας ισχυρής συσχέτισης μεταξύ δυο μεταβλητών δεν υποδηλώνει απαραίτητα και την ύπαρξη μιας σχέσης μεταξύ αιτίας και έκβασης.

Ο συντελεστής συσχέτισης διατάξεων του Spearman όταν υπολογίζεται σε «δείγματα» συμβολίζεται με r_s και λαμβάνει τιμές από -1 έως $+1$, με την ερμηνεία να είναι ίδια με εκείνη στην περίπτωση του συντελεστή συσχέτισης του Pearson. Ο συντελεστής συσχέτισης διατάξεων του Spearman είναι πολύ λιγότερο ευαίσθητος σε ακραίες τιμές από το συντελεστή συσχέτισης του Pearson και μπορεί να χρησιμοποιηθεί όταν η μια ή και οι δύο μεταβλητές είναι σε διατεταγμένη κλίμακα, αλλά επειδή ως μη παραμετρική μέθοδος χρησιμοποιεί τις διατάξεις και όχι τις παρατηρούμενες τιμές, δεν λαμβάνει υπόψη όλη την πληροφορία που είναι γνωστή για μια κατανομή.

Ο *πίνακας συσχετίσεων* είναι ο πίνακας που περιέχει σαν στοιχεία του τους συντελεστές συσχέτισης του Pearson για κάθε ζευγάρι μεταβλητών. Έχει απαραίτητα τιμές ίσες με τη μονάδα στη διαγώνιο, είναι συμμετρικός και κανένα στοιχείο του δεν μπορεί να πάρει τιμή μεγαλύτερη σε απόλυτη τιμή από το 1. Τιμές -1 και 1 σημαίνουν απόλυτα γραμμική σχέση των δύο μεταβλητών (το πρόσημο υποδηλώνει την ύπαρξη θετικής ή αρνητικής σχέσης). Η θετική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής τόσο αυξάνει και η τιμή της άλλης, ενώ, η αρνητική σχέση ερμηνεύεται πως όσο αυξάνει η τιμή της μιας μεταβλητής μειώνεται η τιμή της άλλης.

Ο δειγματικός πίνακας συσχετίσεων συμβολίζεται με R και δίνεται από τον πιο κάτω τύπο:

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

όπου $r_{jk} = \dfrac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} = \dfrac{s_{jk}}{s_j \cdot s_k}$ ο συντελεστής γραμμικής συσχέτισης του Pearson.

Ο πίνακας R είναι συμμετρικός όπως ο πίνακας διακύμανσης-συνδιακύμανσης αφού:

$$\left.\begin{aligned} s_{jk} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)\left(x_{jk}-\overline{x_k}\right) \\ s_{kj} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{jk}-\overline{x_k}\right)\left(x_{ij}-\overline{x_j}\right) \end{aligned}\right\} s_{jk}=s_{kj} \ \kappa\alpha\iota \ r_{jk}=\frac{s_{jk}}{s_j\cdot s_k}=\frac{s_{kj}}{s_k\cdot s_j}=r_{kj}$$

Επίσης, $r_{jj}=1$ εφόσον $r_{jj}=\frac{s_{jj}}{s_j\cdot s_j}=\frac{s_{jj}}{s_j^{\,2}}=\frac{s_{jj}}{s_{jj}}=1.$

Παράλληλα, μπορούμε να πούμε ότι ο πίνακας διακύμανσης –συνδιακύμανσης S τυποποιημένων μεταβλητών, ταυτίζεται με τον πίνακα συσχετίσεων των αρχικών μεταβλητών πριν την τυποποίησή τους.

Παρατηρούμε ότι:

$$r_{jk}=\frac{s_{jk}}{s_j\cdot s_k}=\frac{\sum_{i=1}^{n}\left(x_{ij}-\overline{x_j}\right)\left(x_{jk}-\overline{x_k}\right)}{(n-1)\cdot s_j\cdot s_k}=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{ij}-\overline{x_j}}{s_j}\right)\left(\frac{x_{jk}-\overline{x_k}}{s_k}\right)=\hat{Cov}\left(X_j^*,X_k^*\right)$$

Όπου X_j^*, X_k^* είναι οι τυποποιημένες μεταβλητές.

Μπορούμε, και είναι χρήσιμο, να ελέγξουμε ποιες συσχετίσεις είναι στατιστικά διαφορετικές του μηδενός. Υποθέτοντας ότι X_j και X_k είναι κανονικές, μπορούμε να ελέγξουμε την υπόθεση

$$H_0: r_{jk}=0,$$

$$H_1: r_{jk}\neq 0$$

υπολογίζοντας τη συνάρτηση:

$$\mathrm{T}=\frac{r_{jk}}{\sqrt{\frac{1-r_{jk}^{\,2}}{n-2}}}\sim t_{n-2}$$

Αν $|T| > t_{n-2,1-a/2}$, τότε απορρίπτουμε την H_0 και υποθέτουμε ότι υπάρχει γραμμική συσχέτιση μεταξύ X_j και X_k, όπου $t_{n-2,1-a/2}$ είναι το $1-a/2$ ποσοστιαίο σημείο της κατανομής t με $n-2$ βαθμούς ελευθερίας, δηλαδή

$$P\left(|T| \le t_{n-2,1-a/2}\right) = 1-a/2$$

Αν το μέγεθος του δείγματος n είναι μεγάλο τότε, η $t-$κατανομή προσεγγίζει την κανονική και έχουμε $t_{n-2,1-a/2} \approx z_{1-a/2}$. Χρησιμοποιώντας τη σχέση του συντελεστή γραμμικής συσχέτισης r_{jk} με το συντελεστή R^2 της παλινδρόμησης, οι Chatfield και Collins (1992) προτείνουν να θεωρούμε μεγάλες τις τιμές για $r_{jk} > 0.70$. Σημειώνουμε επίσης ότι οι συντελεστές γραμμικής συσχέτισης δεν πρέπει να υπολογίζονται καθόλου αν το μέγεθος του δείγματος είναι πολύ μικρό (για παράδειγμα αν $n < 12$).

Για καλύτερη ερμηνεία, ειδικά όταν έχουμε πολλές μεταβλητές, μπορούμε αντί του πίνακα συσχετίσεων, να απεικονίσουμε τα δεδομένα με ένα πίνακα όπου θα φαίνονται μόνο οι σημαντικές συσχετίσεις ή έστω οι συσχετίσεις οι οποίες είναι μεγαλύτερες από μια τιμή πρακτικής σημαντικότητας. Ακόμη, για μια καλύτερη εικόνα, μπορούμε να απεικονίσουμε τις σχέσεις γραφικά. Επίσης, μπορούμε να μειώνουμε τα δεκαδικά στοιχεία σε ένα ή δύο για να μπορούμε να εντοπίσουμε καλύτερα σχέσεις μεταξύ των μεταβλητών.

Είναι χρήσιμο να τονιστεί ότι, για να είναι ο πίνακας συσχετίσεων πιο ευανάγνωστος, καλό είναι να παρουσιάζουμε μόνο την κάτω διαγώνιο. Εξάλλου, λόγω της συμμετρίας δεν χάνουμε καμιά πληροφορία. Επίσης, όταν είναι δυνατό, θα πρέπει να βάζουμε κάποια μορφή διάταξης στις συσχετίσεις, ώστε να μπορούμε σχετικά εύκολα να βρούμε αυτές που πραγματικά είναι μεγάλες και άρα χρήσιμες.

Εκτός απ' αυτά, ο συντελεστής συσχέτισης έχει επίσης ενδιαφέρουσα γεωμετρική ερμηνεία, καθώς, μπορεί να αναπαρασταθεί γραφικά. Είναι το συνημίτονο της γωνίας που σχηματίζουν οι δυο μεταβλητές (τυποποιημένες) αν παρουσιαστούν σαν διανύσματα στον χώρο των παρατηρήσεων. Για αυτό σε πολλές γραφικές

απεικονίσεις, αντί να παρουσιάζεται ο πίνακας συσχετίσεων, εμφανίζεται ένα γράφημα όπου κάθε συσχέτιση, αναπαρίσταται από μια γωνία. Αν η γωνία είναι 90 μοίρες, τότε οι δυο μεταβλητές είναι ασυσχέτιστες, ενώ, αν η γωνία έχει 0 μοίρες, τότε οι μεταβλητές είναι πολύ έντονα συσχετισμένες.

Μερικός Συντελεστής Συσχέτισης

Ο απλός συντελεστής συσχέτισης υπολογίζει την συσχέτιση μεταξύ δύο μεταβλητών αγνοώντας τις υπόλοιπες. Έτσι, μπορεί να εμφανίσει συσχετισμένες κάποιες μεταβλητές απλά και μόνο επειδή κάποιες άλλες έχουν μεγάλη συσχέτιση με αυτές και όταν ακυρώσουμε την επίδραση τους, οι αρχικές μεταβλητές να μην εμφανίσουν καμιά συσχέτιση. Για αυτό είναι χρήσιμος ένας συντελεστής, ο οποίος θα υπολογίζει την συσχέτιση, αφού αφαιρέσει την επίδραση των υπολοίπων μεταβλητών. Αυτή την ιδιότητα έχει ο *μερικός συντελεστής συσχέτισης*.

Στατιστικά, δεδομένου ότι η διαδικασία της απλής συσχέτισης μεταξύ των δύο μεταβλητών περιγράφει πλήρως τη γραμμική σχέση μεταξύ των δύο μεταβλητών, ο μερικός συντελεστής συσχέτισης υπολογίζεται με την γνώση του απλού συντελεστή συσχέτισης και μόνο, χωρίς τη χρήση οποιονδήποτε επιμέρους παρατηρήσεων. Άρα, ο μερικός συντελεστής συσχέτισης μιας εξαρτημένη μεταβλητής Y και μιας ανεξάρτητης μεταβλητής X_i, εκφράζει τη συσχέτιση της Y και της X_i, όταν οι επιδράσεις των υπολοίπων μεταβλητών επί της Y και της X_i έχουν απομακρυνθεί.

Έστω ότι έχω 3 μεταβλητές, τις Y, X_1, X_2. Ο συντελεστής μερικής συσχέτισης της εξαρτημένης μεταβλητής Y και της X_1, δίνεται από τον τύπο,

$$r_{y1,2} = \frac{r_{y1} - r_{y1}r_{12}}{\sqrt{\left(1 - r_{y2}^{2}\right)\left(1 - r_{12}^{2}\right)}}$$

Ο συντελεστής μερικής συσχέτισης της εξαρτημένης μεταβλητής Y και της X_2 δίνεται από τον τύπο,

$$r_{y2,1} = \frac{r_{y2} - r_{y1} r_{12}}{\sqrt{\left(1 - r_{y1}^{2}\right)\left(1 - r_{12}^{2}\right)}}$$

όπου,

r_{y1} : ο δειγματικός συντελεστής συσχέτισης του Pearson μεταξύ Y και X_1

r_{y2} : ο δειγματικός συντελεστής συσχέτισης του Pearson μεταξύ Y και X_2

r_{12} : ο δειγματικός συντελεστής συσχέτισης του Pearson μεταξύ X_1 και X_2

Ανάλογοι τύποι προκύπτουν για περισσότερες μεταβλητές. Για παράδειγμα, έστω 4 μεταβλητές Y, X_1, X_2, X_3. Ο μερικός συντελεστής συσχέτισης μεταξύ της εξαρτημένης μεταβλητής Y και της X_1 δίνεται από τον τύπο,

$$r_{y1,23} = \frac{r_{y1,2} - r_{y3,2} r_{13,2}}{\sqrt{\left(1 - r_{y3,2}^{2}\right)\left(1 - r_{13,2}^{2}\right)}}$$

όπου $r_{y1,2}$, $r_{y3,2}$ και $r_{13,2}$ είναι οι αντίστοιχοι μερικοί συντελεστές συσχέτισης.

Η έννοια της μερικής συσχέτισης είναι πολύ σημαντική στην πολλαπλή γραμμική παλινδρόμηση. Στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης, η πρόβλεψη των τιμών της εξαρτημένης μεταβλητής γίνεται με βάση τις συσχετίσεις της με τις τιμές δύο ή περισσοτέρων ανεξάρτητων μεταβλητών. Αυτές οι ανεξάρτητες μεταβλητές είναι (πολύ) πιθανόν να παρουσιάζουν και κάποια συσχέτιση μεταξύ τους. Είναι με άλλα λόγια δυνατόν η συσχέτιση μεταξύ δύο μεταβλητών (X_1 και Y) να επηρεάζεται από τη συσχέτισή τους με μια τρίτη μεταβλητή (X_2). Προκειμένου να εξετάσουμε τη συσχέτιση των δύο αρχικών μεταβλητών, μπορούμε να λάβουμε υπόψη και να εξουδετερώσουμε την επίδραση της τρίτης μεταβλητής, με τη διαδικασία της μερικής συσχέτισης.

Τρόποι Πολυμεταβλητής Ανάλυσης

Παρατηρούμε ότι οι γραμμές του πίνακα δεδομένων αντιστοιχούν σε διαφορετικές παρατηρήσεις και οι στήλες σε διαφορετικές μεταβλητές. Σε πολλές εφαρμογές το ενδιαφέρον κινείται γύρω από τις μεταβλητές και τις σχέσεις που υπάρχουν μεταξύ τους. Ο πίνακας R παίζει μεγάλο ρόλο σε αυτήν την προσέγγιση και έτσι πολλές μέθοδοι μεταβλητών ονομάζονται R*-τεχνικές*. Αυτές οι τεχνικές μπορούν να εξηγήσουν τις σχέσεις που υπάρχουν ανάμεσα στις μεταβλητές. Τέτοιες τεχνικές είναι *η μέθοδος των κύριων συνιστωσών, η παραγοντική ανάλυση* και *η ανάλυση κανονικών συσχετίσεων*.

Πολλές φορές, στόχος είναι να βρεθούν σχέσεις ανάμεσα στις γραμμές του πίνακα και όχι ανάμεσα στις στήλες, σχέσεις δηλαδή ανάμεσα στις παρατηρήσεις και όχι ανάμεσα στις μεταβλητές. Οι τεχνικές που χρησιμοποιούνται σ' αυτήν την περίπτωση είναι γνωστές ως Q*-τεχνικές*. Τέτοιες τεχνικές είναι *η διακριτική ανάλυση, η ανάλυση κατά συστάδες και η πολυδιάστατη κλιμακοποίηση.*

Τέλος, υπάρχουν μέθοδοι που ενδιαφέρονται και για τις γραμμές και για τις στήλες ταυτόχρονα. Προσπαθούν δηλαδή, να αναλύσουν τα δεδομένα και ως προς τις γραμμές και ως προς τις στήλες την ίδια στιγμή. Τέτοια τεχνική είναι η *ανάλυση αντιστοιχιών.*

Πρέπει να έχει κανείς υπόψη του ότι κάθε παρατήρηση είναι ένα διάνυσμα στο χώρο p-διαστάσεων που ορίζουν οι p μεταβλητές που έχουμε.

Συγχρόνως, κάθε μεταβλητή μπορεί να ιδωθεί σαν ένα διάνυσμα στο χώρο n-διαστάσεων που ορίζουν οι παρατηρήσεις μας. Πολλές φορές, όταν έχουμε ένα πίνακα δεδομένων, δεν μπορούμε εύκολα να ξεχωρίσουμε τις μεταβλητές από τις παρατηρήσεις. Γι' αυτό και χρειάζεται ιδιαίτερη προσοχή αφού ο ερευνητής θα πρέπει να ξεκαθαρίσει τι ακριβώς θέλει να μελετήσεις για να μπορέσει να προχωρήσει.

Πολυδιάστατη Κατανομή Τυχαίων Μεταβλητών

Στην πολυμεταβλητή στατιστική ανάλυση το ενδιαφέρον εστιάζεται στη μελέτη πολλών μεταβλητών $X_1, X_2, ..., X_n$ συγχρόνως και όχι ξεχωριστά. Έτσι, μας ενδιαφέρει η μελέτη της από κοινού κατανομής των τυχαίων μεταβλητών. Αυτό που υποθέτουμε είναι ότι υπάρχει εξάρτηση μεταξύ των μεταβλητών αυτών και ο στόχος μας είναι να την μελετήσουμε.

Όπως αναφέραμε και αρχικά, στην πολυμεταβλητή στατιστική μας ενδιαφέρει η από κοινού συνάρτηση πυκνότητας πιθανότητας για τις συνεχείς τυχαίες μεταβλητές και η από κοινού συνάρτηση πιθανότητας για τις διακριτές μεταβλητές.

Έστω Ω ένας δειγματικός χώρος και $X_1, X_2, ..., X_n$, μια συλλογή από n το πλήθος διακριτές τυχαίες μεταβλητές ορισμένες στο Ω.

Αν $R_{X_1, X_2, ..., X_n}$ είναι το σύνολο τιμών ή πεδίο τιμών του n-διάστατου τυχαίου διανύσματος $(X_1, X_2, ..., X_n)$, τότε η συνάρτηση ορίζεται από τον τύπο:

$$f(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) =$$

$$= P\left(\{\omega \in \Omega : X_1(\omega) = x_1, X_2(\omega) = x_2, ..., X_n(\omega) = x_n\}\right)$$

για τα $(x_1, x_2, ..., x_n) \in R_{X_1, X_2, ..., X_n}$

και λέγεται από κοινού συνάρτηση πιθανότητας των διακριτών τυχαίων μεταβλητών $X_1, X_2, ..., X_n$.

Για την από κοινού συνάρτηση πιθανότητας $f(x_1, x_2, ..., x_n)$ του n-διάστατου τυχαίου διανύσματος $(X_1, X_2, ..., X_n)$ με σύνολο τιμών $R_{X_1, X_2, ..., X_n}$ ικανοποιεί τις επόμενες συνθήκες.

1. $f(x_1, x_2, ..., x_n) = 0$, αν $(x_1, x_2, ..., x_n) \notin R_{X_1, X_2, ..., X_n}$
2. $f(x_1, x_2, ..., x_n) \geq 0$ για κάθε $(x_1, x_2, ..., x_n) \in R_{X_1, X_2, ..., X_n}$
3. $\sum_{(x_1, x_2, ..., x_n) \in R_{X_1, X_2, ..., X_n}} f(x_1, x_2, ..., x_n) = 1.$

Δίνεται, ότι ο τύπος της περιθώριας συνάρτησης πιθανότητας της X_i, με $1 \leq i \leq n$ είναι:

$$f_{X_i}(x_i) = P(X_i = x_i) = \sum_{X_j, j \neq i} f(x_1, x_2, ..., x_n)$$

όπου η άθροιση γίνεται ως προς όλες τις άλλες μεταβλητές εκτός από την X_i.

Αν αθροίσουμε τις τιμές της συνάρτησης $f(x_1, x_2, ..., x_n)$ ως προς k, όπου $1 \leq k \leq n-1$ από τις μεταβλητές $X_1, X_2, ..., X_n$, τότε θα πάρουμε την από κοινού περιθώρια συνάρτηση πιθανότητας των υπολοίπων $n-k$.

Αν πάρουμε $f(x, y, z)$ να είναι η από κοινού συνάρτηση πιθανότητας των τυχαίων μεταβλητών X, Y, Z, τότε η περιθώρια συνάρτηση πιθανότητας των Y, Z θα δίνεται από τον τύπο:

$$f_{Y,Z}(y, z) = P(Y = y, Z = z) = \sum_{Y} f(x, y, z), \text{ με } (y, z) \in R_{Y,Z}$$

Ενώ για τις αντίστοιχες περιθώριες των X, Y και X, Z θα έχουμε

$$f_{X,Y}(x, y) = P(X = y, Y = y) = \sum_{X} f(x, y, z), \text{ με } (x, y) \in R_{X,Y}$$

και $f_{X,Z}(x, z) = P(\mathrm{X} = x, Z = z) = \sum_{X} f(x, y, z)$ με $(x, z) \in R_{X,Z}$

Έτσι, λοιπόν, είδαμε ότι το X έχει μια από κοινού συνάρτηση πιθανότητας

$$f(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

Η συνάρτηση που ορίζεται από τον τύπο

$$f(x_1, x_2, ..., x_n) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n)$$

ονομάζεται από κοινού (αθροιστική) συνάρτηση κατανομής του n-διάστατου τυχαίου διανύσματος $X_1, X_2, ..., X_n$

Έστω $X_1, X_2, \ldots X_n$ διακριτές τυχαίες μεταβλητές με από κοινού συνάρτηση πιθανότητας $f(x_1, x_2, ..., x_n)$ και περιθώριες συναρτήσεις

$f_{X_1}(x_1), f_{X_2}(x_2), \ldots, f_{X_n}(x_n)$

Τότε, οι τυχαίες μεταβλητές $X_1, X_2, \ldots X_n$ θα είναι ανεξάρτητες αν και μόνο αν ισχύει: $f(x_1, x_2, ..., x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \ldots \cdot f_{X_n}(x_n)$

για κάθε $x_1, \ldots, x_n \in R_{X_1, \ldots, X_n,}$

Εάν τώρα αυτή η συνάρτηση πιθανότητας είναι συνεχής θα γραφεί ως εξής:

$$f(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(u_1, \ldots, u_n) du_1 \ldots du_n$$

όπου $f\left(x_1, x_2, ..., x_n\right)$ είναι η από κοινού συνάρτηση πυκνότητας των στοιχείων του X.

Όμως, στα επόμενα στατιστικά μοντέλα που θα συναντήσουμε, θα υποθέσουμε, ότι τα στοιχεία του τυχαίου διανύσματος είναι εξαρτημένα.

Χαρακτηριστικές ιδιότητες της από κοινού συνάρτησης πυκνότητας του n-διάστατου τυχαίου διανύσματος $X_1, X_2, ..., X_n$ είναι:

1. $f\left(x_1, x_2, ..., x_n\right) \geq 0$ για κάθε $\left(x_1, x_2, ..., x_n\right) \in R$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\left(x_1, ..., x_n\right) dx_1 ... dx_n = 1$

Πιο συγκεκριμένα αν για μια συνάρτηση f n μεταβλητών ισχύουν οι δυο ιδιότητες που αναφέραμε, τότε η f μπορεί να χρησιμοποιηθεί ως από κοινού συνάρτηση πυκνότητας ενός n-διάστατου συνεχούς τυχαίου διανύσματος.

Η περιθώρια συνάρτηση πυκνότητας της X_1 θα δίνεται από τον τύπο:

$$f_{X_1}(x_1) = \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{\text{n-1 όροι}} f\left(x_1, x_2, ..., x_n\right) dx_2 dx_3 ... dx_n$$

Της X_n από τον τύπο:

$$f_{X_n}(x_n) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f\left(x_1, x_2, ..., x_n\right) dx_1 dx_2 dx_3 ... dx_{n-1}$$

Ενώ γενικά η περιθώρια συνάρτηση πυκνότητας της X_i για $1 < i < n$ είναι ίση με

$$f_{X_i}(x_i) = \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{\text{n-1 όροι}} f\left(x_1, ..., x_{i-1}, x_i, ..., x_{i+1}, ..., x_n\right) dx_1 ... dx_{i-1} dx_{i+1} dx_n$$

Στις σχέσεις αυτές που περιέχουν πολλαπλά ολοκληρώματα είναι σημαντικό να αναγνωρίσουμε, ότι οι λειτουργίες πυκνότητας έχουν οριστεί σωστά, έτσι ώστε τα όρια του ολοκληρώματος να απλώνονται στο $-\infty$ και στο ∞.

Αν ολοκληρώσουμε την από κοινού συνάρτηση πυκνότητας ως προς k $(1 \leq k \leq n-1)$ από τις μεταβλητές $X_1, X_2, ..., X_n$ τότε θα προκύψει η περιθώρια συνάρτηση πυκνότητας των υπολοίπων $n-k$.

Έτσι, αν για παράδειγμα $f(x, y, z, w)$ είναι η από κοινού συνάρτηση πυκνότητας της πολυδιάστατης τυχαίας μεταβλητής (X, Y, Z, W), τότε η περιθώρια συνάρτηση πυκνότητας των X, Y θα δίνεται από τον τύπο

$$f_{X,Z}(x, z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z,W}(x, y, z, w) dy dw$$

ενώ των X, Y, Z από τον τύπο

$$f_{X,Y,W}(x,y,z) = \int_{-\infty}^{\infty} f_{X,Y,Z,W}(x,y,z,w)dw$$

Τέλος η από κοινού (αθροιστική) συνάρτηση κατανομής τυχαίων μεταβλητών $X_1, X_2, ..., X_n$ (διακριτών ή συνεχών) ορίζεται από τον τύπο:

$$f\left(x_1, x_2, ..., x_n\right) = P\left(X_1 \le x_1, X_2 \le x_2, ..., X_n \le x_n\right), \quad x_1, x_2, ..., x_n \in R$$

Έστω $X_1,\ldots X_n$ συνεχείς τυχαίες μεταβλητές με από κοινού συνάρτηση πυκνότητας $f\left(x_1, ..., x_n\right)$ και περιθώριες συναρτήσεις πυκνότητας $f_{X_1}(x_1), f_{X_2}(x_2), \ldots, f_X(x_n)$

Τότε, οι τυχαίες μεταβλητές $X_1, X_2, \ldots X_n$ θα είναι ανεξάρτητες αν και μόνο αν ισχύει: $f\left(x_1, x_2, ..., x_n\right) = f_{X_1}(x_1) \cdot f_X(x_2) \cdot \ldots \cdot f_{X_n}(x_n)$

για κάθε $x_1, x_2, \ldots, x_n \in R_{X_1, X_2, \ldots, X_n,}$

Κλείνοντας αυτή την παράγραφο, αναφέρουμε ότι στην περίπτωση που έχουμε $n > 2$ τυχαίες μεταβλητές, μπορούμε να ορίσουμε διάφορες δεσμευμένες κατανομές χρησιμοποιώντας μια ή περισσότερες τυχαίες μεταβλητές για καθένα από τα δύο τμήματα (μη δεσμευμένο, δεσμευμένο).

Πριν αναλύσουμε με σχέσεις τη δεσμευμένη κατανομή, θα πρέπει να πούμε κάποια πράγματα πάνω στην πολυμεταβλητή ανάλυση. Είναι, λοιπόν, συχνά απαραίτητο στην πολυμεταβλητή ανάλυση να γνωρίσουμε την κατανομή του ενός συνόλου των τυχαίων μεταβλητών, δοθέντος ότι οι μεταβλητές μιας δεύτερης ομάδας ορίζονται το ίδιο με σταθερές τιμές ή αναγκάζονται να βρίσκονται σε ορισμένες υποπεριοχές των χώρων τους. Αυτές οι λειτουργίες της πυκνότητας και της κατανομής λέγονται δεσμευμένες.

Έτσι, λοιπόν, για να ερμηνεύσουμε τη δεσμευμένη κατανομή, θα πάρουμε τέσσερις τυχαίες μεταβλητές X_1, X_2, X_3, X_4 με από κοινού συνάρτηση πιθανότητας την $f\left(x_1, x_2, x_3, x_4\right)$, τότε:

- Η δεσμευμένη κατανομή του X_1, δοθέντος ότι $X_2 = x_2, X_3 = x_3, X_4 = x_4$, ορίζεται από τον τύπο:

$$f_{X_1/X_2,X_3,X_4}\left(x_1 / x_2, x_3, x_4\right) = \frac{f\left(x_1, x_2, x_3, x_4\right)}{f_{X_2,X_3,X_4}\left(x_2, x_3, x_4\right)} \qquad (1)$$

- Η δεσμευμένη κατανομή του X_2, δοθέντος ότι $X_3 = x_3$ και $X_4 = x_4$, ορίζεται από τον τύπο:

$$f_{X_2/X_3,X_4}\left(x_2/x_3,x_4\right)=\frac{f_{X_2,X_3,X_4}\left(x_2,x_3,x_4\right)}{f_{X_3,X_4}\left(x_3,x_4\right)} \qquad (2)$$

- Η δεσμευμένη κατανομή των X_1,X_2,X_3, δοθέντος ότι $X_4=x_4$ έχει από κοινού συνάρτηση πιθανότητας ή πυκνότητας πιθανότητας την:

$$f_{X_1,X_2,X_3/X_4}\left(x_1,x_2,x_3/x_4\right)=\frac{f\left(x_1,x_2,x_3,x_4\right)}{f_{X_4}\left(x_4\right)} \qquad (3)$$

- Η δεσμευμένη κατανομή των X_2,X_4, δοθέντος ότι $X_1=x_1,X_3=x_3$, έχει από κοινού συνάρτηση πιθανότητας ή πυκνότητας πιθανότητας την:

$$f_{X_2,X_4/X_1,X_3}\left(x_2,x_4/x_1,x_3\right)=\frac{f\left(x_1,x_2,x_3,x_4\right)}{f_{X_1,X_3}\left(x_1,x_3\right)} \qquad (4)$$

Με βάση τα παραπάνω μπορούμε να ορίσουμε και τις αντίστοιχες μέσες τιμές:
Έτσι, για τις περιπτώσεις των δεσμευμένων των σχέσεων (1) και (2) θα έχουμε, ότι για συνεχείς τυχαίες μεταβλητές ο τύπος της μέσης τιμής της δεσμευμένης κατανομής είναι:

$$E\left(X_1/X_2=x_2,X_3=x_3,X_4=x_4\right)=\int_{-\infty}^{\infty}x_1 f_{X_1/X_2,X_3,X_4}\left(x_1/x_2,x_3,x_4\right)dx_1$$

$$E\left(X_2/X_3=x_3,X_4=x_4\right)=\int_{-\infty}^{\infty}x_2 f_{X_2/X_3,X_4}\left(x_2/x_3,x_4\right)dx_2$$

Η μη δεσμευμένη μέση τιμή μιας συνάρτησης $h\left(X_1,X_2,...,X_n\right)$ των n μεταβλητών $X_1,X_2,...,X_n$ υπολογίζεται από τον τύπο (για διακριτές τυχαίες μεταβλητές)

$$E\left[h\left(X_1,X_2,...,X_n\right)\right]=$$

$$\sum_{(x_1,x_2,\ldots,x_n)\in R_{X_1,X_2,\cdots X_n}}h\left(x_1,x_2,...,x_n\right)f\left(x_1,x_2,\ldots,x_n\right)$$

ενώ, για τις συνεχείς τυχαίες μεταβλητές, ο τύπος είναι:

$$E\left[h\left(X_1,X_2,...,X_n\right)\right]=$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}h\left(x_1,x_2,...,x_n\right)f\left(x_1,x_2,...,x_n\right)dx_1dx_2\ldots dx_n$$

Οι ιδιότητες της μέσης τιμής εδώ είναι εντελώς ανάλογες με τις ιδιότητες που γνωρίσαμε στην περίπτωση των δισδιάστατων τυχαίων μεταβλητών.
Ιδιαίτερα χρήσιμος είναι ο τύπος:

$$E\left(a_1X_1+a_2X_2+...+a_nX_n\right)=a_1E\left(X_1\right)+a_2E\left(X_2\right)+...+a_nE\left(X_n\right)$$

Έστω ότι $h\left(X_1,X_2\right)=\left(X_1-E\left(X_1\right)\right)\cdot\left(X_2-E\left(X_2\right)\right)$.

Η συνάρτηση αυτή οδηγεί στον ορισμό της συνδιακύμανσης δυο τυχαίων μεταβλητών, της οποίας ο τύπος είναι:

$$Cov(X_1, X_2) = E\left[\left(X_1 - E(X_1)\right)\cdot\left(X_2 - E(X_2)\right)\right] =$$
$$= E(X_1X_2) - E(X_1)\cdot E(X_2).$$

Δυο ακόμη βασικές ιδιότητες της συνδιακύμανσης είναι οι εξής:

- $Cov(X_1 + X_3 + X_2) = Cov(X_1, X_2) + Cov(X_3, X_2)$
- $Cov(aX_1 + \beta, \gamma X_2 + \delta) = \alpha\cdot\gamma\cdot Cov(X_1, X_2)$

Η πολυδιάστατη Κανονική Κατανομή και οι Ιδιότητές της

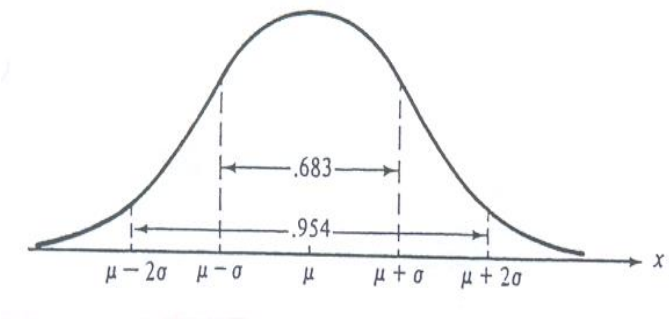
Η πολυδιάστατη κανονική κατανομή είναι η γενίκευση της μονομεταβλητής κανονικής κατανομής για διαστάσεις $p \geq 2$. Αν η απλή κανονική κατανομή αποτελεί τη βάση για τις περισσότερες απλές στατιστικές εφαρμογές π.χ. ελέγχους υποθέσεων, γραμμικές παλινδρομήσεις κ.τ.λ. ανάλογη είναι και η χρήση της πολυδιάστατης κανονικής κατανομής στην πολυμεταβλητή στατιστική. Στην πράξη, οι περισσότερες μέθοδοι αναπτύχθηκαν με βάση την κατανομή αυτή.

Υπενθυμίζουμε ότι η μονομεταβλητή συνάρτηση πυκνότητας με μέση τιμή μ και διακύμανση σ^2 είναι:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2}, \quad -\infty < x < \infty \qquad (1)$$

Το γνώριμο σχήμα είναι αυτό της καμπάνας. Παρατηρούμε ότι στο διάστημα $[\mu - 2\sigma, \mu + 2\sigma]$ βρίσκεται το 95,4% και στο διάστημα $[\mu - \sigma, \mu + \sigma]$ παρατηρείται το 68,3% . Οι περιοχές αυτές αντιπροσωπεύουν πιθανότητες και συνεπώς για την κανονική τυχαία μεταβλητή X έχουμε

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$$



Σχήμα 1 A normal density with mean μ and variance σ^2 and selected areas under the curve.

Έτσι δηλώνεται η κανονική πυκνότητα με μέση τιμή μ και διακύμανση σ^2 με κανονική κατανομή $N(\mu,\sigma^2)$. Ως εκ τούτου το $N(10,4)$ παραπέμπει στο Σχήμα 1 με μέση τιμή 10 και διασπορά το 2.

Ο όρος $$\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu) \qquad (2)$$

όπου x είναι η τιμή της τυχαίας μεταβλητής X και στον εκθέτη της μονομεταβλητής κανονικής συνάρτησης πυκνότητας, μετρά το τετράγωνο της απόστασης, του x από τη μέση τιμή μ σε μονάδες τυπικής απόκλισης. Μπορεί να γενικευτεί για ένα $px1$ διάνυσμα των παρατηρήσεων σε διάφορες μεταβλητές όπως

$$(x-\mu)'\Sigma^{-1}(x-\mu) \qquad (3)$$

Το px1 διάνυσμα μ, όπου μ΄=($\mu_1,\mu_2,\ldots,\mu_n$) αντιπροσωπεύει την αναμενόμενη τιμή του τυχαίου διανύσματος X=[$X_1,X_2,\ldots,X_p$]΄ και ο pxp πίνακας Σ είναι ο πίνακας διακύμανσης – συνδιακύμανσης.

Η πολυδιάστατη κανονική κατανομή που ορίζεται για το τυχαίο διάνυσμα X διαστάσεων $px1$ έχει από κοινού συνάρτηση πυκνότητας πιθανότητας.

$$f(x) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \qquad e^{-(x-\mu)'\Sigma^{-1}(x-\mu)/2} \qquad (4)$$

Όπου έχουμε $-\infty < x_i < \infty$, i=1,2,…,p

Το διάνυσμα μ αναφέραμε ότι περιέχει τις αναμενόμενες τιμές κάθε μεταβλητής. Είναι δηλαδή το διάνυσμα των μέσων, ενώ ο πίνακας Σ είναι ο πίνακας με τις συνδιακυμάνσεις των μεταβλητών του τυχαίου διανύσματος X δηλαδή

$$\mu = \mathrm{E}(X)$$
$$\Sigma = COV(X)$$

Γενικά, θα συμβολίζουμε $X \sim N_p(\mu,\Sigma)$ και εννοούμε ότι το τυχαίο διάνυσμα X ακολουθεί p-διάσταση κανονική κατανομή, με διάνυσμα μέσων μ και πίνακα διακυμάνσεων Σ.

Παράδειγμα 1 (Διμεταβλητή Κανονική Κατανομή)

Έστω X_1,X_2 δυο τυχαίες μεταβλητές που ακολουθούν από κοινού την κανονική κατανομή $X_1,X_2 \sim M_2(\mu,\Sigma)$ όπου $\boldsymbol{\mu}= (\mu_1,\mu_2)$ για τα οποία ισχύει $\mu_1=E(X_1)$, $\mu_2=E(X_2)$, $\sigma_{11} = Var(X_1)$, $\sigma_{22} = Var(X_2)$ και $p_{12} = \sigma_{12} / \left(\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\right) = Corr(X_1, X_2)$

Ο πίνακας συνδιακύμανσης αναγράφεται ως εξής:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma_{11}} & \boldsymbol{\sigma_{12}} \\ \boldsymbol{\sigma_{12}} & \boldsymbol{\sigma_{22}} \end{bmatrix}$$

Είναι:

$$\boldsymbol{\Sigma^{-1}} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^{\,2}} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

Δημιουργώντας τον συντελεστή συσχέτισης p_{12} και γράφοντας $\sigma_{12} = p_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{12}}$ παίρνουμε $\sigma_{11}\sigma_{22} - \sigma_{12}^{\,2} = \sigma_{11}\sigma_{22}(1 - p_{12}^{\,2})$ οπότε η τετραγωνική απόσταση $(x-\mu)'\Sigma^{-1}(x-\mu)$ γίνεται:

$$\left[x_1 - \mu_{1,} x_2 - \mu_2\right] \frac{1}{\sigma_{11}\sigma_{22}(1 - p_{12}^{\,2})} \begin{bmatrix} \sigma_{22} & -p_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{12}} \\ -p_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{12}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

$$= \frac{\sigma_{22}(x_1 - \mu_1)^2 + \sigma_{11}\left(x_2 - \mu_2\right)^2 - 2p_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_{11}\sigma_{22} \cdot (1 - p_{12}^{\,2})} =$$

$$\frac{1}{1 - p_{12}^2}\left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2p_{12}\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)\right]$$

Η τελευταία έκφραση γράφεται όσον αφορά τις τυποποιημένες τιμές

$\left(x_1 - \mu_1\right)/\sqrt{\sigma_{11}}$ και $\left(x_2 - \mu_2\right)/\sqrt{\sigma_{22}}$

Στη συνέχεια, αν $|\Sigma| = \sigma_{11}\sigma_{22} - \sigma_{12}^{\,2} = \sigma_{11}\sigma_{22}\left(1 - p_{12}^2\right)$, αντικαθιστούμε τα Σ^{-1} και $|\Sigma|$ στη σχέση 4 ώστε να πάρουμε την έκφραση για τη διμεταβλητή $\left(p = 2\right)$

κανονική πυκνότητα όπου συμπεριλαμβανομένου και των παραμέτρων $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$ και p_{12} έχουμε

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}\left(1-p_{12}^2\right)}} \cdot \exp(Q)$$

όπου

$$Q = \left\{ -\frac{1}{2\left(1-p_{12}^2\right)} \left[\left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right)^2 + \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right)^2 - 2p_{12} \left(\frac{x_1-\mu_1}{\sqrt{\sigma_{11}}} \right) \left(\frac{x_2-\mu_2}{\sqrt{\sigma_{22}}} \right) \right] \right\}$$

Στην περίπτωση που $p_{12}=0$, τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας της διμεταβλητής κατανομής είναι απλά το γινόμενο δυο συναρτήσεων πιθανοτήτων απλών κανονικών κατανομών, δηλαδή οι δυο μεταβλητές είναι ανεξάρτητες και ισχύει $f(x_1, x_2) = f\left(x_1\right) \cdot f\left(x_2\right)$

Για τη δισδιάστατη κανονική κατανομή είναι εύκολο να φτιάξουμε γραφήματα της από κοινού συνάρτησης πυκνότητας. Αυτά μπορούν να μας δώσουν ενδιαφέρουσες ερμηνείες για τις παραμέτρους. Στη μονοδιάστατη περίπτωση έχουμε μια συμμετρική κατανομή. Το ίδιο ισχύει και στη δισδιάστατη περίπτωση της οποίας τα στοιχεία καθορίζονται από τις παραμέτρους.

Παρατηρώντας το γράφημα που έχουμε δώσει παρακάτω βλέπουμε ότι η κορυφή της καμπάνας βρίσκεται πάντα στο σημείο μ, δηλαδή ακριβώς στο σημείο που υποδεικνύει το διάνυσμα των μέσων.

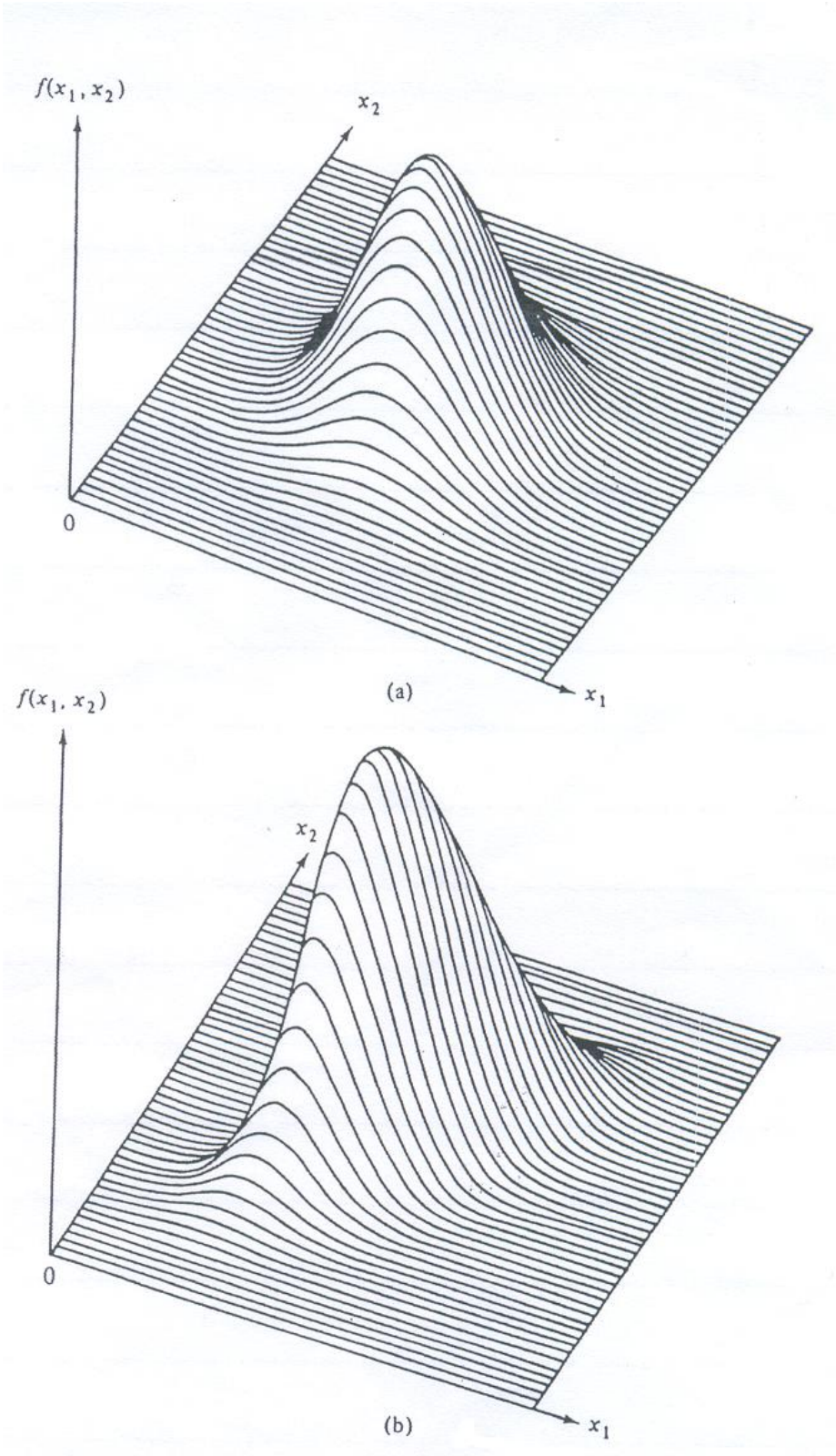
Από τον τύπο της πυκνότητας πιθανότητας της πολυδιάστατης κανονικής κατανομής $f(x) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} e^{-(x-\mu)'\Sigma^{-1}(x-\mu)/2}$

παρατηρούμε ότι όλα τα διανύσματα **x** που ικανοποιούν τις σχέσεις $(x-\mu)'\Sigma^{-1}(x-\mu)=c^2$ όπου c σταθερά, βρίσκονται στην επιφάνεια ενός ελλειψοειδούς (έλλειψη σε δυο διαστάσεις) και έχουν σταθερή πυκνότητα πιθανότητας.

Τα ελλειψοειδή αυτά έχουν κέντρο το μ.

Οι διευθύνσεις των αξόνων των ελλειψοειδών καθορίζονται από τα ιδιοδιανύσματα του πίνακα Σ^{-1}, ενώ τα μήκη τους είναι ανάλογα των τετραγωνικών ριζών των ιδιοτιμών του Σ^{-1}. Ονομάζονται δε ελλειψοειδή σταθερής πυκνότητας.

Έτσι έχουμε το ακόλουθο αποτέλεσμα.



Σχέδιο 2 Δυο διμεταβλητές κανονικές κατανομές (α) $\sigma_{11} = \sigma_{22}$ και $p_{12} = 0$ (b) $\sigma_{11} = \sigma_{22}$ και $p_{12} = 0.75$

Αποτέλεσμα 1

Σαν αποτέλεσμα έχουμε ότι εάν ο Σ είναι θετικά ορισμένος τότε ο Σ^{-1} υπάρχει και έχουμε ότι $\boldsymbol{\Sigma e = \lambda e}$ συνεπάγεται $\boldsymbol{\Sigma^{-1} e = \left(\frac{1}{\lambda}\right) e}$

Έτσι το (λ, e) είναι ένα χαρακτηριστικό ζεύγος του Σ και το αντίστοιχο του Σ^{-1} είναι (1/λ,e). Άρα καταλήγουμε ότι και ο Σ^{-1} είναι θετικά ορισμένος.

Απόδειξη

Για θετικά ορισμένο Σ και ένα ιδιοδιάνυσμα $e \neq 0$ έχουμε ότι $e'(\Sigma e) = e'(\lambda e) = \lambda e' e = \lambda$. Επιπλέον, ισχύει ότι $e = \Sigma^{-1}(\Sigma e) = \Sigma^{-1}(\lambda e)$ ή $e = \lambda \Sigma^{-1} e$ και η διαίρεση με $\lambda > 0$ δίνει $\Sigma^{-1} e = \left(1/\lambda\right) \cdot e$. Έτσι $\left(1/\lambda, e\right)$ είναι ένα ζευγάρι ιδιοτιμών – ιδιοδιανυσμάτων για το Σ^{-1}.

Για κάθε $px1$ διάνυσμα x έχω

$$x'\Sigma^{-1}x = x'\left(\sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\right)e_i e_i'\right)x = \sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\right)(x'e_i)^2 \geq 0$$

Κάθε όρος $\lambda_i^{-1}(x'e_i)^2$ είναι μη αρνητικός. Επίσης, ισχύει $x'e_i = 0$ για όλα τα i μόνο αν $x = 0$

Έτσι για $x \neq 0$ συνεπάγεται $\sum_{i=1}^{p}\left(1/\lambda_i\right)(x'e_i)^2 > 0$ και Σ^{-1} είναι θετικά ορισμένος.

Συνοψίζοντας, η καμπύλη της σταθερής πυκνότητας για την ρ-διαστάσεων κανονική κατανομή είναι ελλειψοειδής και ορίζεται από τα x που ικανοποιούν τη σχέση $(x-\mu)'\Sigma^{-1}(x-\mu) = c^2$

Η εξίσωση $(x-\mu)'\Sigma^{-1}(x-\mu)=c^2$ είναι μια εξίσωση ελλειψοειδούς. Μάλιστα η ποσότητα αυτή καθορίζει τις υψομετρικές καμπύλες σταθερής πυκνότητας, δηλαδή κάθε σημείο πάνω σε αυτό το ελλειψοειδές (ή της έλλειψης για δυο διαστάσεις) έχει την ίδια ακριβώς πυκνότητα.

Το διάνυσμα μ είναι το κέντρο των καμπύλων που έχουν ημιάξονες του οποίου είναι $c\sqrt{\lambda_i}\boldsymbol{e}_i$, όπου $\Sigma\boldsymbol{e}_i=\lambda_i\boldsymbol{e}_i$, $i=1,2,\ldots,p$

Παράδειγμα 2

Διάγραμμα Διμεταβλητής Κανονικής Πυκνότητας

Θα μελετήσουμε τους άξονες του διαγράμματος σταθερής πυκνότητας πιθανότητας για μια διμεταβλητή κατανομή όταν οι συνδιακυμάνσεις είναι ίσες, δηλαδή $\sigma_{11}=\sigma_{22}$.

Από αυτούς εδώ τους άξονες δίνονται οι ιδιότιμες και τα ιδιοδιανύσματα του Σ

Εδώ το $|\Sigma-\lambda\mathrm{I}|=0$ γίνεται:

$$0=\begin{vmatrix}\sigma_{11}-\lambda & \sigma_{12}\\ \sigma_{12} & \sigma_{11}-\lambda\end{vmatrix}=(\sigma_{11}-\lambda)^2-\sigma_{12}^2=(\lambda-\sigma_{11}-\sigma_{12})(\lambda-\sigma_{11}+\sigma_{12})$$

Αντίστοιχα, οι ιδιοτιμές είναι $\lambda=\sigma_{11}+\sigma_{12}$ και $\lambda=\sigma_{11}-\sigma_{12}$

Το ιδιοδιάνυσμα $\boldsymbol{e}_1$ προσδιορίζεται από:

$$\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{11}\end{bmatrix}\begin{bmatrix}\boldsymbol{e}_1\\ \boldsymbol{e}_2\end{bmatrix}=(\sigma_{11}+\sigma_{12})\begin{bmatrix}\boldsymbol{e}_1\\ \boldsymbol{e}_2\end{bmatrix}$$

ή

$$\sigma_{11}\boldsymbol{e}_1+\sigma_{12}\boldsymbol{e}_2=(\sigma_{11}+\sigma_{12})\ \ \boldsymbol{e}_1$$

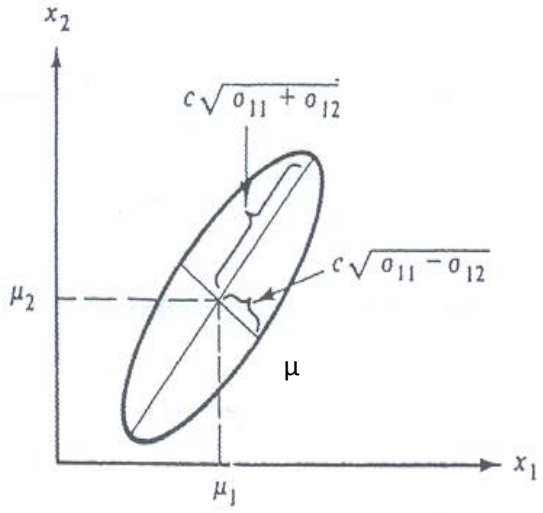
$$\sigma_{12}\boldsymbol{e}_1+\sigma_{11}\boldsymbol{e}_2=(\sigma_{11}+\sigma_{12})\ \ \boldsymbol{e}_2$$

Από τη λύση του συστήματος έχουμε

$\lambda_1 = \sigma_{11} + \sigma_{12}$, $\boldsymbol{e_1} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

Ομοίως για το $\lambda_2 = \sigma_{11} - \sigma_{12}$ το ιδιοδιάνυσμα $\boldsymbol{e'}_2$ είναι: $\boldsymbol{e'}_2 = \left[1/\sqrt{2}, -1/\sqrt{2}\right]$

Όταν η συνδιασπορά σ_{12} είναι θετική και $\lambda_1 = \sigma_{11} + \sigma_{12}$ είναι η μεγαλύτερη ιδιοτιμή τότε το ιδιοδιάνυσμα $\boldsymbol{e'}_1 = \left[1/\sqrt{2}, 1/\sqrt{2}\right]$ βρίσκεται κατά μήκος της γραμμής που σχηματίζει 45° με τον οριζόντιο άξονα στο σημείο $\mu' = [\mu_1, \mu_2]$ (βλέπε σχήμα 3). Αυτό γίνεται για κάθε θετική τιμή της συνδιασποράς. Επιπλέον, οι άξονες της σταθερής έλλειψης πυκνότητας δίνονται από τους τύπους $\pm c\sqrt{\lambda_1}\,\mathbf{e}_1$ και $\pm c\sqrt{\lambda_2}\,\mathbf{e}_2$, και ο κύριος άξονας πα πρέπει να συνδέεται με την μεγαλύτερη ιδιοτιμή.



Σχήμα 3 A constant density contour for a bivariate normal distribution with $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} > 0$ (or $\rho_{12} > 0$).

Για να συνοψίσουμε, οι άξονες της έλλειψης της σταθερής πυκνότητα για διμεταβλητή κανονική κατανομή με $\sigma_{11} = \sigma_{22}$ καθορίζονται από:

$$\pm c\sqrt{\sigma_{11} + \sigma_{12}}\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \text{και} \quad \pm c\sqrt{\sigma_{11} - \sigma_{12}}\begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

όπου η επιλογή $c^2 = x_p^2(a)$, με $x_p^2(a)$ να είναι το ανώτερο ποσοστιαίο εκατοστημόριο της κατανομής x^2 με ρ βαθμούς.

Συγκεκριμένα, το ακόλουθο είναι αληθές για μια ρ-διαστάσεων κανονική κατανομή.

Από τη θεωρία γνωρίζουμε ότι η κατανομή x^2 προκύπτει αν υψώσουμε μια τυποποιημένη κανονική τυχαία μεταβλητή στο τετράγωνο. Δηλαδή αν $X_i \sim N(0,1)$ ανεξάρτητες τυχαίες μεταβλητές, όπου $i = 1,..., p$ τότε $\sum_{i=1}^{p} x_i^2 \sim x_p^2$, η οποία ονομάζεται x^2 - κατανομή με p βαθμούς ελευθερίας.

Το στερεό ελλειψοειδές των τιμών x ικανοποιεί τη σχέση

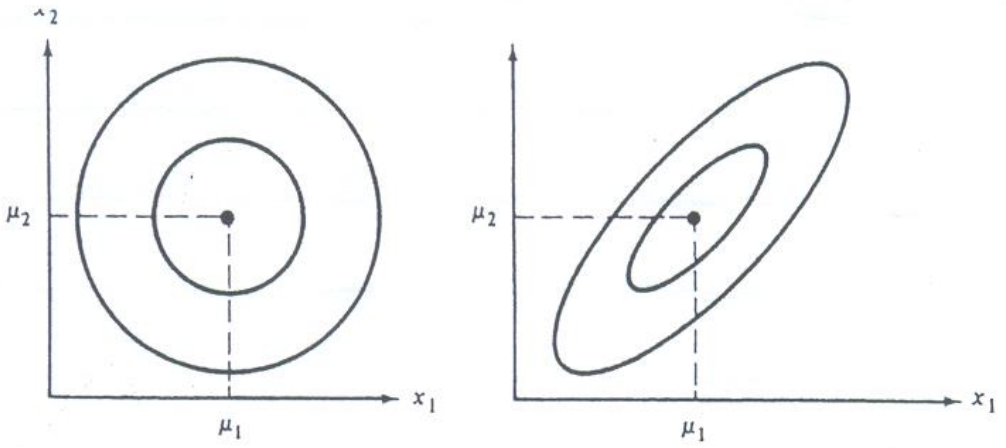
$$(x-\mu)'\Sigma^{-1}(x-\mu) \le x_p^2(\alpha) \qquad (8)$$

με πιθανότητα $1-\alpha$.

Στο παρακάτω σχήμα βλέπουμε ότι τα περιγράμματα της σταθερής πυκνότητας περιέχουν το 50% και το 90% της πιθανότητας κάτω από τις επιφάνειες της διμεταβλητής κανονικής.

Η ρ-μεταβλητή κανονική πυκνότητα έχει μια μέγιστη τιμή όταν η τετραγωνική απόσταση του $(x-\mu)'\Sigma^{-1}(x-\mu)$ είναι μηδέν και αυτό συμβαίνει όταν $x = \mu$. Έτσι το μ είναι το σημείο της μέγιστης πυκνότητας όπως η αναμενόμενη τιμή του X. Το γεγονός ότι το μ είναι η μέση τιμή της πολυμεταβλητής κανονικής κατανομής

δείχνει ότι ακολουθεί τη συμμετρία που παρουσιάστηκε από τα περιγράμματα των συνεχών καμπυλών πυκνότητας. Αυτά τα περιγράμματα έχουν ως κέντρο το μ.



(a) $\sigma_{11} = \sigma_{22}, \rho_{12} = 0$ (b) $\sigma_{11} = \sigma_{22}, \rho_{12} = .75$

Σχήμα 4 The 50% and 90% contours for the bivariate normal distributions in σχήμα 2

Αν έχουμε περισσότερες από δυο μεταβλητές είναι δύσκολο να αναπαραστήσουμε γραφικά την παράσταση η οποία θα μοιάζει με έλλειψη, στις τρείς θα μοιάζει με μπάλα του ράγκμπι ενώ για περισσότερες διαστάσεις δεν μπορεί να αναπαρασταθεί με κάποιο σχήμα.

Κάτι πού πρέπει να σημειώσουμε είναι ότι υπολογίζοντας την εξίσωση της έλλειψης μπορούμε να ελέγξουμε αν ένα σημείο είναι μέσα ή έξω από την περιοχή αυτή. Επίσης αξιοσημείωτο είναι ότι αν τα δεδομένα είναι ασυσχέτιστα $(p=0)$, τότε η έλλειψη είναι κύκλος. Συγκεκριμένα, στις περισσότερες διαστάσεις, όταν οι μεταβλητές είναι ασυσχέτιστες το υπερελλειψοειδές γίνεται υπέρσφαιρα, γι' αυτό και πολλές φορές την έλλειψη συσχέτισης την ονομάζουμε και σφαιρικότητα των δεδομένων. Εκτός απ' αυτά πρέπει να πούμε ότι γενικά το πρόσημο του συντελεστή

συσχέτισης καθορίζει την κλίση ελλείψεων. Ακόμα, η απόλυτη τιμή του καθορίζει τα χαρακτηριστικά της έλλειψης.

Επιπλέον Ιδιότητες της Πολυμεταβλητής Κανονικής Κατανομής

Στη μονοδιάστατη περίπτωση οι γραμμικοί μετασχηματισμοί κανονικής κατανομής αλλά και οι γραμμικοί συνδυασμοί ανεξαρτήτων κανονικών κατανομών ακολουθούν κανονικές κατανομές. Οι ιδιότητες αυτές γενικεύονται στη πολυδιάστατη περίπτωση.

Οι ιδιότητες της πολυμεταβλητής κανονικής κατανομής είναι οι εξής:

1. Οι Γραμμικοί Συνδυασμοί των συνιστωσών του τ.δ. Χ όπου $X \sim N_p(\mu, \Sigma)$ είναι κανονικές τυχαίες μεταβλητές.
2. Όλα τα υποσύνολα των συνιστωσών του Χ είναι τ.δ. με πολυδιάστατη κανονική κατανομή.
3. Μηδενική Συνδιακύμανση μεταξύ των συνιστωσών του Χ συνεπάγεται ότι οι συνιστώσες αυτές είναι ανεξάρτητα κατανεμημένες.
4. Οι δεσμευμένες κατανομές των συνιστωσών του Χ είναι πολυμεταβλητές κανονικές.

Αποτέλεσμα 2

Εάν $X' = (X_1, X_2, \ldots, X_p) \sim N_p(\boldsymbol{\mu}, \Sigma)$ και $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ είναι διάνυσμα σταθερών

τότε $\alpha' X = \alpha_1 X_1 + \ldots + \alpha_p X_p \sim N(\boldsymbol{\alpha'\mu}, \boldsymbol{\alpha'\Sigma\alpha})$

επίσης αν $\alpha' X$ έχει την κατανομή στο $N(\boldsymbol{\alpha'\mu}, \boldsymbol{\alpha'\Sigma\alpha})$ για κάθε α, τότε το Χ έχει την κατανομή $N_p(\boldsymbol{\mu}, \Sigma)$

Παράδειγμα 3

Θεωρούμε τον γραμμικό συνδυασμό $\alpha' X$ ενός πολυμεταβλητού κανονικού τυχαίου διανύσματος που καθορίζεται από την επιλογή $\alpha' = [1, 0, ..., 0]$. Αν

$$\boldsymbol{\alpha}'\mathbf{X} = [1, 0, ..., 0] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\rho \end{bmatrix} = X_1$$

και $\boldsymbol{\alpha}'\boldsymbol{\mu} = [1, 0, ..., 0] \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_\rho \end{bmatrix} = \mu_1$

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha} = [1, 0, ..., 0] \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1\rho} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2\rho} \\ \vdots & \vdots & & \vdots \\ \sigma_{1\rho} & \sigma_{2\rho} & \dots & \sigma_{\rho\rho} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \sigma_{11}$$

Το X_1 λοιπόν ακολουθεί κανονική κατανομή $\mathrm{N}(\mu_1, \sigma_{11})$. Γενικότερα, η οριακή κατανομή για κάθε συνιστώσα X_i του X είναι $\mathrm{N}(\mu_i, \sigma_{ii})$

Το επόμενο αποτέλεσμα θεωρεί αρκετούς γραμμικούς συνδυασμούς για το πολυμεταβλητό κανονικό διάνυσμα.

Αποτέλεσμα 3

Εάν $X \sim N_p(\boldsymbol{\mu},\Sigma)$ και A πίνακας qxp, τότε

$$\underset{(qxp)}{A} \cdot \underset{(px1)}{X} = \begin{bmatrix} a_{11}X_1 + ... + a_{1p}Xp \\ a_{21}X_1 + ... + a_{2p}Xp \\ . \\ . \\ a_{q1}X_1 + ... + a_{qp}Xp \end{bmatrix} \sim N_q(A\boldsymbol{\mu}, A\Sigma A')$$

Αν d ένα διάνυσμα σταθερών τότε το τ.δ. $X+d \sim N_p(\mu+d,\Sigma)$

Απόδειξη

Κάθε γραμμικός συνδυασμός $b'(AX)$ είναι γραμμικός συνδυασμός του πίνακα X διαστάσεων (px1)

Θα εξετάσουμε τη μορφή $a' \cdot X$ όπου $a = A'b$.

Γνωρίζουμε από το προηγούμενο αποτέλεσμα ότι το $\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ είναι διάνυσμα σταθερών όρων.

Ισχύει ότι $\alpha' X = \alpha_1 X_1 + \ldots + \alpha_p X_p \sim N(\alpha'\mu, \alpha'\Sigma\alpha)$

Άρα $AX \sim N_q(A\mu, A\Sigma A')$

Το δεύτερο μέρος του αποτελέσματος μπορεί να επιτευχθεί εξετάζοντας τη μορφή $\alpha'(x+d) = \alpha' X + (\alpha' d)$, όπου $\alpha' X \sim N(\alpha'\mu, \alpha'\Sigma\alpha)$.

Στη μονομεταβλητή περίπτωση προσθέτοντας τη σταθερά α΄d, όπου d ένα διάνυσμα σταθερών, στην τυχαία μεταβλητή α΄Χ παρατηρούμε ότι η διακύμανση θα παραμείνει αμετάβλητη.

Η μέση τιμή που προκύπτει είναι: $\alpha'\mu + \alpha' d = \alpha'(\mu+d)$

Επομένως $X+d \sim N_p(\mu+d, \Sigma)$

Παράδειγμα 4

Για $X \sim N_3(\mu, \Sigma)$, να βρείτε την κατανομή του

$$\begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = AX$$

Από το Αποτέλεσμα 3, η κατανομή ΑΧ είναι πολυμεταβλητή κανονική κατανομή με μέση τιμή

$$\mathbf{A\mu} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_2 - \mu_3 \end{bmatrix}$$

Πίνακα συνδιασποράς – συνδιακύμανσης

$$\mathbf{A\Sigma A'} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ \sigma_{12} - \sigma_{13} & \sigma_{22} - \sigma_{23} & \sigma_{23} - \sigma_{33} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} =$$

$$= \begin{bmatrix} \sigma_{11} - 2\sigma_{12} + \sigma_{22} & \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} \\ \sigma_{12} + \sigma_{23} - \sigma_{22} - \sigma_{13} & \sigma_{22} - 2\sigma_{23} + \sigma_{33} \end{bmatrix}$$

Αποτέλεσμα 4

Όλα τα υποσύνολα του X κατανέμονται κανονικά με $X_2 \sim N(\mu, \Sigma)$. Αν διαμερίσουμε το X για κάθε έννοια του μέσου μ και του πίνακα συνδυασποράς Σ σαν:

$$\underset{(p\text{x}1)}{X} = \begin{bmatrix} \underset{(q\text{x}1)}{X_1} \\ \underset{(p-q)\text{x}1}{X_2} \end{bmatrix} \qquad \underset{(p\text{x}1)}{\mu} = \begin{bmatrix} \underset{(q\text{x}1)}{\mu_1} \\ \underset{(p-q)\text{x}1}{\mu_2} \end{bmatrix}$$

$$\Sigma = \left[\begin{array}{c|c} \underset{(qxq)}{\Sigma_{11}} & \underset{(qx(p-q))}{\Sigma_{12}} \\ \hline \underset{(p-q)xq}{\Sigma_{21}} & \underset{((p-q)x(p-q))}{\Sigma_{22}} \end{array}\right].$$

Όπου Σ_{11},Σ_{12} είναι συμμετρικοί πίνακες και $\Sigma'_{21}=\Sigma_{12}$

Παράδειγμα 5

Εάν $X\sim N_5(\mu,\Sigma)$, να βρείτε την κατανομή του $\begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$

Θέτουμε $X_1 = \begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$, $\mu_1 = \begin{bmatrix} \mu_2 \\ \mu_4 \end{bmatrix}$, $\sigma_{11} = \begin{bmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{bmatrix}$

Έτσι λοιπόν με αυτή την ανάθεση και σύμφωνα με το αποτέλεσμα 4 έχουμε

$$X = \left[\begin{array}{c} X_2 \\ X_4 \\ \hline X_1 \\ X_3 \\ X_5 \end{array}\right], \quad \mu = \left[\begin{array}{c} \mu_2 \\ \mu_4 \\ \hline \mu_1 \\ \mu_3 \\ \mu_5 \end{array}\right], \quad \Sigma = \left[\begin{array}{cc|ccc} \sigma_{22} & \sigma_{24} & \sigma_{12} & \sigma_{23} & \sigma_{25} \\ \sigma_{24} & \sigma_{44} & \sigma_{14} & \sigma_{34} & \sigma_{45} \\ \hline \sigma_{12} & \sigma_{14} & \sigma_{11} & \sigma_{13} & \sigma_{15} \\ \sigma_{23} & \sigma_{34} & \sigma_{13} & \sigma_{33} & \sigma_{35} \\ \sigma_{25} & \sigma_{45} & \sigma_{15} & \sigma_{35} & \sigma_{55} \end{array}\right]$$

ή

$$X = \left[\begin{array}{c} \underset{(2x1)}{X_1} \\ \hline \underset{(3x1)}{X_2} \end{array}\right], \quad \mu = \left[\begin{array}{c} \underset{(2x1)}{\mu_1} \\ \hline \underset{(3x1)}{\mu_2} \end{array}\right], \quad \Sigma = \left[\begin{array}{c|c} \underset{(2x2)}{\Sigma_{11}} & \underset{(2x3)}{\Sigma_{12}} \\ \hline \underset{(3x2)}{\Sigma_{21}} & \underset{(3x3)}{\Sigma_{22}} \end{array}\right]$$

Οπότε για $X_1 = \begin{bmatrix} X_2 \\ X_4 \end{bmatrix}$

έχοντας την εξής κατανομή $N_2 \sim (\mu_1,\Sigma_{11})$ προκύπτει $N_2 \sim \left[\begin{bmatrix} \mu_2 \\ \mu_4 \end{bmatrix}, \begin{bmatrix} \sigma_{22} & \sigma_{24} \\ \sigma_{24} & \sigma_{44} \end{bmatrix}\right]$

Από το Παράδειγμα είναι ξεκάθαρο, ότι η κανονική κατανομή για οποιοδήποτε υποσύνολο μπορεί να εκφραστεί αν επιλέξουμε το κατάλληλο μέσο (μ) και την συνδιακύμανση (Σ) κάθε φορά.

Αποτέλεσμα 5

Κατανοούμε πλέον ότι η μηδενική συσχέτιση της κανονικής κατανομής είναι ισοδύναμη με την στατιστική ανεξαρτησία.

Παραθέτουμε τα εξής αποτελέσματα

a. Εάν τα τυχαία διανύσματα $\underset{(q_1x1)}{X_1}$ και $\underset{(q_2x1)}{X_2}$ είναι ανεξάρτητα, τότε ισχύει ότι

$\text{cov}(X_1, X_2) = 0$

b. Εάν το $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ είναι $N_{q_1+q_2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]\right)$

τότε τα X_1 και X_2 είναι ανεξάρτητα μόνο στην περίπτωση που ισχύει $\Sigma_{12} = 0$

c. Εάν X_1 και X_2 είναι ανεξάρτητα και κατανέμονται ως $N_{q_1}(\mu_1, \Sigma_{11})$ και $N_{q_2}(\mu_2, \Sigma_{22})$ αντίστοιχα, τότε το $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ ακολουθεί την πολυμεταβλητή κατανομή

$$N_{q_1+q_2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \left[\begin{array}{c|c} \Sigma_{11} & 0 \\ \hline 0' & \Sigma_{22} \end{array}\right]\right)$$

Παράδειγμα 6

Έχουμε ότι $\underset{(3X1)}{X} \sim N_3(\mu, \Sigma)$ με $\Sigma = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$

Είναι τα X_1 και X_2 ανεξάρτητα; Τι συμβαίνει με τα (X_1, X_2) και X_3;

Εφόσον τα X_1 και X_2 έχουν συνδιακύμανση $\sigma_{12} = 1$, αυτό σημαίνει ότι δεν είναι ανεξάρτητα.

Όμως τα X και Σ διαμερίζονται ως εξής:

$$X=\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix},\quad \Sigma=\left[\begin{array}{cc:c} 4 & 1 & 0 \\ 1 & 3 & 0 \\ \hdashline 0 & 0 & 2 \end{array}\right]=\left[\begin{array}{c:c} \underset{(2x2)}{\Sigma_{11}} & \underset{(2x1)}{\Sigma_{12}} \\ \hdashline \underset{(1x2)}{\Sigma_{21}} & \underset{(1x1)}{\Sigma_{22}} \end{array}\right]$$

Οι Σ_{11} και Σ_{22} είναι πίνακες διακύμανσης – συνδιακύμανσης, αλλά οι πίνακες $\Sigma_{12},\Sigma_{21},$ δεν είναι πίνακες διακύμανσης – συνδιακύμανσης.

Παρατηρούμε λοιπόν ότι $X_1=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ και το X_3 έχουν πίνακα συνδυασποράς

$$\Sigma_{12}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Όμως, τα (X_1,X_2) και X_3 είναι ανεξάρτητα με βάση το Αποτέλεσμα 5b. Αυτό συνεπάγεται ότι το X_3 είναι ανεξάρτητο του X_1 και ανεξάρτητο επίσης του X_2.

Αποτέλεσμα 6

Έστω $X=\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ που κατανέμεται ως $N_p(\mu,\Sigma)$ με $\mu=\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},\ \Sigma=\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ και

$|\Sigma_{22}|>0$

Τότε για τη δεσμευμένη κατανομή του X_1 δοθέντος ότι $X_2=x_2$

η μέση τιμή = $\mu'=\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(\chi_2-\mu_2)$

και η συνδιακύμανση $=\Sigma'=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Να σημειώσουμε ότι η συνδιακύμανση δεν εξαρτάται από το X_2 της δεσμευμένης μεταβλητής.

Απόδειξη

Θα δοθεί μια έμμεση απόδειξη

Παίρνουμε: $\underset{(pxp)}{\mathrm{A}} = \begin{bmatrix} \underset{(qxq)}{I} & \underset{qx(p-q)}{-\Sigma_{12}\Sigma_{22}^{-1}} \\ \underset{(p-q)xq}{0} & \underset{(p-q)x(p-q)}{I} \end{bmatrix}$

Ώστε

$$A(\mathrm{X}-\mu) = A\begin{bmatrix} X_1-\mu_1 \\ \hdashline X_2-\mu_2 \end{bmatrix} = \begin{bmatrix} X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2) \\ \hdashline X_2-\mu_2 \end{bmatrix}$$

που είναι από κοινού κανονική με πίνακα συνδιακύμανσης $A\Sigma A'$ που δίνεται:

$$\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ \hdashline 0 & I \end{bmatrix} \cdot \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \mathrm{I} & 0' \\ \hdashline \left(-\Sigma_{12}\Sigma_{22}^{-1}\right)' & \mathrm{I} \end{bmatrix} =$$

$$\begin{bmatrix} \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0' \\ \hdashline 0 & \Sigma_{22} \end{bmatrix}$$

Αφού τα $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$ και $X_2-\mu_2$ έχουν μηδενική διακύμανση, αυτό σημαίνει ότι είναι ανεξάρτητα. Επιπλέον, η ποσότητα $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$ κατανέμεται ως $\mathrm{N}_q(0,\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.

Δίνεται ότι το $X_2 = X_2$, $\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$ είναι μια σταθερά.

Επειδή τα $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$ και $X_2-\mu_2$ είναι ανεξάρτητα, η δεσμευμένη κατανομή του $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$ είναι ίδια με την κατανομή του $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$.

Το $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$ κατανέμεται ως $\mathrm{N}_q(0,\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$, έτσι είναι το τυχαίο διάνυσμα $X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$ όταν το X_2 έχει τιμή x_2. Ισοδύναμα, δοθέντος ότι έχουμε $X_2 = x_2$ το X_1 κατανέμεται ως $\mathrm{N}_q\left(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2),\ \ \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$

Παρατήρηση

Μπορούμε να χρησιμοποιήσουμε το παραπάνω αποτέλεσμα, για να βρούμε την απλή δεσμευμένη κατανομή μιας και μόνο τυχαίας μεταβλητής, όπως επίσης και να βρούμε δεσμευμένη κατανομή, αγνοώντας κάποιες άλλες μεταβλητές. Για παράδειγμα, αν έχουμε ένα αρχικό διάνυσμα $X' = (X_1, X_2, X_3)$ μπορούμε να βρούμε τις δεσμευμένες κατανομές των $X_1 X_2 \mid X_3$, $X_1 \mid X_2 X_3$ ή ακόμα και την κατανομή του $X_1 \mid X_3$

Είναι συχνά απαραίτητο στην πολυμεταβλητή ανάλυση να γνωρίζουμε την κατανομή του ενός συνόλου των τυχαίων μεταβλητών, όταν δίνεται ότι οι μεταβλητές μιας δεύτερης ομάδας ορίζονται το ίδιο.

Αυτές οι λειτουργίες πυκνότητας και της κατανομής λέγονται δεσμευμένες.

Η πυκνότητα της δεσμευμένης κατανομής των $x_1, ..., x_p$ δίνεται $x_{p+1} = x_{p+1}, ..., x_{p+q} = x_{p+q}$ και μπορούμε να το δούμε από :

$$h(x_1, ..., x_p \mid x_{p+1}, ..., x_{p+q}) = \frac{f(x_1, ..., x_{p+q})}{g(x_{p+1}, ..., x_{p+q})} \qquad (12)$$

Όπου το $f(x_1, ..., x_{p+q})$ είναι η από κοινού πυκνότητα του συνόλου των $p+q$ παρατηρήσεων και το $g(x_{p+1}, ..., x_{p+q})$ είναι η θετική από κοινού πυκνότητα των q σταθερών μεταβλητών. Εάν τα δυο σύνολα των μεταβλητών είναι ανεξάρτητα, η παραγοντοποίηση της από κοινού πυκνότητας συνεπάγεται ότι η δεσμευμένη πυκνότητα του πρώτου συνόλου είναι απλώς η από κοινού πυκνότητα αυτών των τυχαίων μεταβλητών.

Παράδειγμα 7

Σε αυτό το Παράδειγμα θα δούμε ότι η πυκνότητα εξαρτάται από μια διμεταβλητή κατανομή. Έτσι η δεσμευμένη πυκνότητα του X_1, δοθέντος $X_2 = x_2$ για κάθε διμεταβλητή κατανομή ορίζεται από:

$$f(x_1 \mid x_2) = \{\text{η δεσμευμένη πυκνότητα του } X_1, \text{ δοθέντος } X_2 = x_2\} = \frac{f(x_1, x_2)}{f(x_2)}$$

όπου το $f(x_2)$ είναι η περιθώρια κατανομή του X_2. Εάν, $f(x_1, x_2)$ είναι η διμεταβλητή κανονική κατανομή, πρέπει να δείξουμε ότι

$$f(x_1 \mid x_2) \sim \mathrm{N}\left(\mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2), \quad \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} \right)$$

Ισχύει ότι $\sigma_{11} - \sigma_{12}^2 / \sigma_{22} = \sigma_{11}(1 - p_{12}^2)$

Οι δύο όροι $x_1 - \mu_1$ που αφορούν τον εκθέτη της διμεταβλητής κανονικής πυκνότητας γίνεται εκτός από την πολλαπλασιαστική σταθερά $-\frac{1}{2}(1 - p_{12}^2)$

$$\frac{(x_1 - \mu_1)^2}{\sigma_{11}} - 2p_{12}\frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} = \frac{1}{\sigma_{11}}\left[x_1 - \mu_1 - p_{12}\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(x_2 - \mu_2) \right]^2 - \frac{p_{12}^2}{\sigma_{22}}(x_2 - \mu_2)^2$$

Επειδή $p_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$ ή $\frac{p_{12}\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}} = \frac{\sigma_{12}}{\sigma_{22}}$, ο πλήρης εκθέτης είναι:

$$\frac{-1}{2(1 - p_{12}^2)}\left[\frac{(x_1 - \mu_1)^2}{\sigma_{11}} - 2p_{12}\frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} + \frac{(x_2 - \mu_2)^2}{\sigma_{22}} \right] =$$

$$\frac{-1}{2\sigma_{11}(1 - p_{12}^2)}\left(x_1 - \mu_1 - p_{12}\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(x_2 - \mu_2) \right)^2 - \frac{1}{2(1 - p_{12}^2)}\left(\frac{1}{\sigma_{22}} - \frac{p_{12}^2}{\sigma_{22}} \right)(x_2 - \mu_2)^2 =$$

$$\frac{1}{2\sigma_{11}(1 - p_{12}^2)}\left(x_1 - \mu_1 - \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2) \right)^2 - \frac{1}{2}\frac{(x_2 - \mu_2)^2}{\sigma_{22}}$$

Ο σταθερός όρος $2\pi\sqrt{\sigma_{11}\sigma_{22}\left(1-p_{12}^2\right)}$ γίνεται:

$$\sqrt{2\pi}\sqrt{\sigma_{22}} \quad x \quad \sqrt{2\pi}\sqrt{\sigma_{11}\left(1-p_{12}^2\right)}$$

Διαιρώντας την από κοινού κατανομή των X_1 και X_2 από την περιθώρια πυκνότητα:

$$f(x_2)=\frac{1}{\sqrt{2\pi}\sqrt{\sigma_{22}}}e^{-(x_2-\mu_2)^2/2\sigma_{22}}$$

και ακυρώνοντας την άποψη των αποδόσεων της δεσμευμένης πυκνότητα:

$$f(x_1\mid x_2)=\frac{f\left(x_1,x_2\right)}{f\left(x_2\right)}=\frac{1}{\sqrt{2\pi}\sqrt{\sigma_{11}\left(1-p_{12}^2\right)}}e^{-\left[x_1-\mu_1-(\sigma_{12}/\sigma_{22})(x_2-\mu_2)\right]^2/2\sigma_{11}\left(1-p_{12}^2\right)}$$

$$-\infty<x_1<\infty$$

Επιπλέον, η δεσμευμένη κατανομή του X_1 δοθέντος του $X_2=x_2$ είναι

$$N\left(\mu_1+\left(\sigma_{12}/\sigma_{22}\right)\left(x_2-\mu_2\right),\quad \sigma_{11}\left(1-p_{12}^2\right)\right)$$

Τώρα $\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}=\sigma_{11}-\sigma_{12}^2/\sigma_{22}=\sigma_{11}\left(1-p_{12}^2\right)$

και $\Sigma_{12}\Sigma_{22}^{-1}=\sigma_{12}/\sigma_{22}$ συμφωνεί με το Αποτέλεσμα 6

Όταν ασχολούμαστε με την πολυμεταβλητή κανονική κατάσταση καλό είναι να δίνουμε έμφαση στα ακόλουθα:

1. Όλες οι δεσμευμένες κατανομές είναι κανονικές
2. Η δεσμευμένη μέση τιμή είναι της μορφής

$$\mu_1+\beta_{1,q+1}\left(x_{q+1}-\mu_{q+1}\right)+\ldots+\beta_{1,q}\left(x_p-\mu_p\right)$$

$$\mu_q+\beta_{q,q+1}\left(x_{q+1}-\mu_{q+1}\right)+\ldots+\beta_{q,p}\left(x_p-\mu_p\right)$$

όπου τα β’s ορίζονται από:

$$\Sigma_{12}\Sigma_{22}^{-1}=\begin{bmatrix}\beta_{1,q+1} & \beta_{1,q+2} & \cdots & \beta_{1,p}\\ \beta_{2,q+1} & \beta_{2,q+2} & \cdots & \beta_{2,p}\\ \vdots & \vdots & \ddots & \vdots\\ \beta_{q,q+1} & \beta_{q,q+2} & \cdots & \beta_{q,p}\end{bmatrix}$$

3. Η δεσμευμένη συνδυακύμανση $\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ δεν εξαρτάται από τις τιμές των δεσμευμένων μεταβλητών.

Καταλήγουμε ότι, σε αυτή την ενότητα παρουσιάζονται δυο ιδιότητες της πολυμεταβλητής κανονικής με τυχαία διανύσματα. Η πρώτη αναφέρεται στα ελλειψοειδή σταθερής πυκνότητα και η δεύτερη στην κατανομή των γραμμικών συνδυασμών τυχαίων μεταβλητών.

Αποτέλεσμα 7

Έστω $X\sim N_p(\mu,\Sigma)$ με $|\Sigma|>0$. Τότε:

(a) Το $(X-\mu)'\Sigma^{-1}(X-\mu)$ έχει την x_p^2 όπου το x_p^2 κατανομή

(b) $P\left[(x-\mu)'\Sigma^{-1}(x-\mu)\le x_p^2;a\right]=1-\alpha$ όπου το $x_p^2;\alpha$ υποδηλώνει το άνω α-εκατοστημόριο της x_p^2 κατανομής.

Απόδειξη

Γνωρίζουμε ότι η x_p^2 ορίζεται ως η κατανομή του αθροίσματος $Z_1^2+Z_2^2+...+Z_p^2$ όπου τα $Z_1,Z_2,...,Z_p$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ακολουθούν την τυπική κανονικά κατανομή $N(0,1)$

Στη συνέχεια από τη φασματική ανάλυση έχουμε ότι:

$\Sigma^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_i} e_i e_i'$ όπου $\Sigma ei = \lambda iei$, έτσι $\Sigma^{-1} ei = \left(1/\lambda i\right) ei$

Κατά συνέπεια αντικαθιστώντας έχουμε

$$\left(X-\mu\right)'\Sigma^{-1}\left(X-\mu\right) = \sum_{i=1}^{p}\left(1/\lambda_i\right)\left(X-\mu\right)'e_i e'_i\left(X-\mu\right) =$$

$$= \sum_{i=1}^{p}\left(1/\lambda_i\right)\left(e'_i\left(X-\mu\right)\right)^2 = \sum_{i=1}^{p}\left[\left(1/\sqrt{\lambda_i}\right)e'_i\left(X-\mu\right)\right]^2 = \sum_{i=1}^{p} Z_i^2$$

Τώρα έχουμε $Z = A\left(X-\mu\right)$ όπου

$$\underset{(px1)}{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} \quad , \quad \underset{(pxp)}{A} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_i}} e'_1 \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}} e'_p \end{bmatrix}$$

με **(X-μ)**~$N_p(0,\Sigma)$ τέτοιο ώστε να προκύπτει ότι και ο γραμμικός συνδυασμός $Z=A(\mathbf{X-\mu}) \sim N_p(0,A\Sigma A')$

όπου:

$$\underset{(pxp)}{A}\ \underset{(pxp)}{\Sigma}\ \underset{(pxp)}{A'} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} e'_1 \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}} e'_p \end{bmatrix} \left[\sum_{i=1}^{p} \lambda_i e_i e'_i\right] \left[\frac{1}{\sqrt{\lambda_1}} e_1 \ldots \frac{1}{\sqrt{\lambda_p}} e_p\right] =$$

$$\begin{bmatrix} \sqrt{\lambda_1} e'_1 \\ \vdots \\ \sqrt{\lambda_p} e'_p \end{bmatrix} \left[\frac{1}{\sqrt{\lambda_1}} e_1 \ldots . \frac{1}{\sqrt{\lambda_p}} e_p\right]$$

Τα $z_1, z_2, \ldots, z_p$ είναι ανεξάρτητες τυπικές κανονικές μεταβλητές οπότε καταλήγουμε ότι το $\left(X-\mu\right)'\Sigma^{-1}\left(X-\mu\right)$ ακολουθεί μια x_p^2 κατανομή.

Για το (b) έχουμε να αναφέρουμε τα εξής:

Η εξίσωση $(x-\mu)'\Sigma^{-1}(x-\mu)=c^2$ είναι μια εξίσωση ελλειψοειδούς. Μάλιστα η ποσότητα αυτή καθορίζει τις υψομετρικές καμπύλες σταθερής πυκνότητας, δηλαδή κάθε σημείο πάνω σε αυτό το ελλειψοειδές (ή η έλλειψη για δυο μεταβλητές) έχει την ίδια ακριβώς πυκνότητα.

Έτσι λοιπόν μπορούμε να σημειώσουμε ότι το:

$$P\left[(X-\mu)'\Sigma^{-1}(X-\mu)\le \mathrm{X}_p^2;a\right]=1-a$$

Στη συνέχεια θα θελήσουμε να εξετάσουμε τον γραμμικό συνδυασμό των διανυσμάτων των τυχαίων μεταβλητών.

$$c_1\,\mathbf{X}_1 + c_2\,\mathbf{X}_2 + \ldots + c_n\,\mathbf{X}_n = [\mathbf{X}_1 \vdots \mathbf{X}_2 \vdots \ldots \vdots \mathbf{X}_n]\cdot \mathbf{c} \qquad (10)$$

(pxn) (nx1)

Αυτός ο γραμμικός συνδυασμός διαφέρει από τους γραμμικούς συνδυασμούς που εξετάσαμε νωρίτερα.

Η εξίσωση (10) ορίζει ένα $px1$ τυχαίο διάνυσμα που είναι γραμμικός συνδυασμός τυχαίων διανυσμάτων.

Προηγουμένως, μιλήσαμε για μια μόνο τυχαία μεταβλητή η οποία θα μπορούσε να γραφτεί ως ένας γραμμικός συνδυασμός των άλλων μονομεταβλητών τυχαίων μεταβλητών.

Αποτέλεσμα 8

Έστω τα $X_1,...,X_n$ διανύσματα είναι αμοιβαίως ανεξάρτητα με το X_j να κατανέμεται στο $N_p\left(\mu_j,\Sigma\right)$. Να σημειώσουμε ότι καθένα X_j έχει τον ίδιο πίνακα συνδυασποράς Σ

Τότε: $V_1 = c_1\,\mathbf{X}_1 + c_2\,\mathbf{X}_2 + ... + c_n\,\mathbf{X}_n \sim N_p\left(\sum_{j=1}^{n} c_j\mu_j,\ \left(\sum_{j=1}^{n} c_j^2\right)\Sigma\right)$

Εξάλλου τα V_1 και $V_2 = b_1X_1 + b_2X_2 + ... + b_nX_n$ είναι από κοινού κανονικές πολυμεταβλητές με πίνακα συνδυασποράς

$$\begin{bmatrix} \left(\sum_{j=1}^{n} c_j^2\right)\Sigma & (\boldsymbol{b'c})\boldsymbol{\Sigma} \\ (\boldsymbol{b'c})\boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} b_j^2\right)\Sigma \end{bmatrix}$$

Κατά συνέπεια τα V_1 και V_2 είναι ανεξάρτητα εάν ισχύει:

$$\boldsymbol{b'c} = \sum_{j=1}^{n} c_j b_j = 0$$

Απόδειξη

Αφού γνωρίζουμε ότι εάν τα X_1 και X_2 είναι ανεξάρτητα και κατανέμονται σε $N_{q1}\left(\mu_1,\Sigma_{11}\right)$ και $N_{q2}\left(\mu_2,\Sigma_{22}\right)$ αντίστοιχα, τότε τα $\begin{bmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \end{bmatrix}$ είναι πολυμεταβλητή κανονική.

$$N_{q1+q2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right)$$

Έτσι η p_n συνιστώσα είναι :

$$\left[X_{11},...,X_{p1},X_{12},...,X_{p2},...,X_{pn}\right]=[\mathbf{X}_1',\mathbf{X}_2',\ldots,\ \mathbf{X}_n']=\ \mathbf{X}'$$

(1 x np)

είναι πολυμεταβλητή κανονική.

Ιδίως όταν το $X\sim N_{np}(\mu,\Sigma x)$ όπου

$$\underset{(np\ x\ 1)}{\mu}=\begin{bmatrix}\mu_1\\ \vdots\\ \mu_n\end{bmatrix} \text{ και } \quad \underset{np\ x\ np}{\Sigma_x}=\begin{bmatrix}\Sigma & 0 & \cdots & 0\\ 0 & \Sigma & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \Sigma\end{bmatrix}$$

Η επιλογή $\underset{(2p\ x\ np)}{A}=\begin{bmatrix}c_1I & c_2I & \ldots & c_nI\\ b_1I & b_2I & \ldots & b_nI\end{bmatrix}$

όπου ο I είναι ο Ταυτοτικός πίνακας, δίνει

$$AX=\begin{bmatrix}\sum_{j=1}^{n}c_jX_j\\ \sum_{j=1}^{n}b_jX_j\end{bmatrix}=\begin{bmatrix}V_1\\ V_2\end{bmatrix}$$

και το AX ακολουθεί την κανονική κατανομή $N_{2p}\left(A\mu,A\Sigma_xA'\right)$

Το $A\Sigma_xA'$ έχει τον πρώτο διαγώνιο όρο:

$$[c_1\mathbf{\Sigma},\ c_2\ \mathbf{\Sigma}\ ,...,c_n\mathbf{\Sigma}]\ x\ [c_1\mathbf{I},c_2\mathbf{I}\ ,...,c_n\mathbf{I}]'=\left(\sum_{j=1}^{n}c_j^2\right)\mathbf{\Sigma}$$

Ο εκτός της διαγωνίου όρος είναι:

$$[c_1\mathbf{\Sigma},c_2\mathbf{\Sigma}\ ,...,c_n\mathbf{\Sigma}]\ x\ [b_1\ \mathbf{I},b_2\mathbf{I}\ ,...,b_n\mathbf{I}]'=\left(\sum_{j=1}^{n}c_jb_j\right)\mathbf{\Sigma}$$

Αυτός ο όρος είναι ο πίνακας συνδυασποράς για τα V_1,V_2

Κατά συνέπεια όταν το $\sum_{j=1}^{n}c_jb_j=b'c=0$ έτσι ώστε

$\left(\sum_{j=1}^{n} c_j b_j\right)\Sigma = \underset{(pxp)}{0}$, τα V_1 και V_2 να είναι ανεξάρτητα και αυτό αφού, εάν $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ είναι $N_{q1+q2}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$, τότε X_1 και X_2 είναι ανεξάρτητα μόνο εάν $\Sigma_{12} = 0$

Εκτίμηση παραμέτρων της πολυδιάστατης Κανονικής Κατανομής

Έστω ότι $X_1, X_2, ..., X_n$ τυχαίο δείγμα από την κ-διάστατη κανονική κατανομή $N_\kappa(\mu, \Sigma)$ με R^κ και $\Sigma > 0$. Θα υποθέσουμε ότι οι δυο αυτοί παράμετροι είναι άγνωστοι και άρα ο παραμετρικός χώρος είναι το σύνολο,

$$\Theta = \left\{ \theta - (\mu, \Sigma) : \ με\ R^\kappa,\ \Sigma > 0\ kxk\ \ πίνακα \right\}$$

Εδώ λοιπόν είμαστε σε θέση να ορίσουμε ορισμένες κλασσικές μεθόδους εκτίμησης των παραμέτρων μ και Σ (της παραμέτρου $\Theta = (\mu, \Sigma)$). Χρησιμοποιώντας το δείγμα μας θα βρούμε τους αμερόληπτους εκτιμητές ελάχιστης διασποράς αλλά και εκτιμητές με τη χρήση της μέγιστης πιθανοφάνειας. Με αυτό τον τρόπο θα ορίσουμε την κατανομή Wishart που θα χρειαστεί στους αμερόληπτους εκτιμητές.

Η δειγματική κατανομή των $\bar{x}$ και S, - κατανομή Wishart

Η υπόθεση ότι οι στήλες του πίνακα

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

αποτελούν ένα τυχαίο δείγμα από έναν κανονικό πληθυσμό με μέση τιμή μ , και συνδυασπορά Σ είναι αυτά που μας προσδιορίζουν την κατανομή δειγματοληψίας των $\bar{x}$ και S.

Αυτό που θα κάνουμε εδώ είναι να προσπαθήσουμε να παρουσιάσουμε με τη βοήθεια των μονομεταβλητών συμπερασμάτων τα αποτελέσματα της δειγματοληψίας των $\bar{x}$ και S.

Έτσι στη μονομεταβλητή περίπτωση για την οποία ισχύει ότι $p=1$ γνωρίζουμε ότι το $\bar{x}$ είναι κανονικό με μέση τιμή μ και διακύμανση $\frac{1}{n}\sigma^2$

Το αποτέλεσμα τώρα για την πολυμεταβλητή περίπτωση όπου έχουμε $p \geq 2$ είναι ανάλογο με τη μονομεταβλητή περίπτωση, αφού κι εδώ το $\bar{x}$ ακολουθεί κανονική κατανομή με μέση τιμή μ και πίνακα συνδιακύμανσης $\frac{1}{n}\Sigma$

Για τη διακύμανση, τώρα του δείγματος αξίζει να θυμηθούμε ότι:

$$(n-1)s^2 = \sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2$$

κατανέμεται ως σ^2 φορές μιας x^2 μεταβλητής που έχει $n-1$ βαθμούς ελευθερίας.

Η x^2 μεταβλητή είναι μια κατανομή ενός αθροίσματος τετραγώνων ανεξαρτήτων τυποποιημένων κανονικών τυχαίων μεταβλητών. Έτσι το $(n-1)s^2$ κατανέμεται όταν

$$\sigma^2 Z_1^2 + \ldots + Z_{n-1}^2 = \left(\sigma Z_1\right)^2 + \ldots + \left(\sigma Z_{n-1}\right)^2$$

στο οποίο οι επιμέρους όροι σZ_1 είναι ανεξάρτητα κατανεμημένοιι σε κανονική κατανομή $N\left(0,\sigma^2\right)$

Καταλήγουμε στο ότι: η κατανομή λοιπόν δειγματοληψίας του πίνακα συνδιακύμανσης του δείγματος καλείται κατανομή Wishart.

Η κατανομή αυτή ορίζεται ως το σύνολο από ανεξάρτητα προϊόντα της πολυμεταβλητής κανονικής με τυχαία δείγματα.

Ειδικά $W_m\left(\cdot/\Sigma\right)$ κατανομή Wishart με m βαθμούς ελευθερίας = κατανομή του

$$\sum_{j=1}^{m} Z_j Z'_j$$

Δηλαδή $W = \sum_{j=1}^{m} Z_j Z'_j$

Όπου τα Z_j είναι το καθένα ξεχωριστά ανεξάρτητα κατανεμημένα σε κανονική κατανομή $N_p(0,\Sigma)$

Επομένως λέμε ότι ακολουθεί την κατανομή Wishart με παραμέτρους n και Z. Θα συμβολίζουμε την κατανομή Wishart με $W_p(n,\Sigma)$

Η συνάρτηση πυκνότητας πιθανότητας του θετικά ορισμένου πίνακα Α είναι:

$$W_{n-1}(A/\Sigma) = \frac{|A|^{(n-p-2)/2} \cdot e^{-trA\Sigma^{-1}/2}}{2^{p(p-1)/2} \cdot \pi^{p(p-1)/4} \cdot |\Sigma|^{(n-1)/2} \prod_{\Pi=1}^{p} \Gamma\frac{1}{2}(n-i)}$$

Η μέση τιμή δίνεται από τον τύπο $E(w) = n\Sigma$ ενώ η διακύμανση ενός τυχαίου πίνακα συνήθως δεν γίνεται γιατί είναι ιδιαίτερα πολύπλοκη. Να σημειώσουμε ότι η κατανομή Wishart ορίζεται μόνο για συμμετρικούς θετικούς πίνακες των τετραγωνικών και των διαγραμμικών σχημάτων, οι οποίοι μπορούν να μεταμορφωθούν στο άθροισμα που αναφέραμε και προηγουμένως.

Δηλαδή $\sum_{j=1}^{m} Z_j Z_j^{'}$

Ενδιαφέρον, παρουσιάζει και ο αντίστροφος ενός πίνακα w που ακολουθεί ην κατανομή Wishart. Η κατανομή λοιπόν του αντίστροφου ακολουθεί την αντίστροφη Wishart κατανομή (Wishart distribution) και έχει αρκετή χρησιμότητα στη Μπεϋζιανή στατιστική.

Συνοψίζοντας, τα αποτελέσματα όσων αναφέραμε παραπάνω, παρατηρήσαμε ότι:

Ας πάρουμε ότι, τα $X_1, X_2, ..., X_n$ είναι ένα τυχαίο δείγμα μεγέθους n από μια p-μεταβλητή κανονική κατανομή με μέση τιμή μ και πίνακα συνδιακύμανσης Σ τότε:

1. Ο $\bar{x}$ κατανέμεται $N_p\left(\mu, \left(\frac{1}{n}\right)\Sigma\right)$

2. $(n-1)\cdot S$ κατανέμεται σαν ένας τυχαίος πίνακας Wishart με $(n-1)$ βαθμούς ελευθερίας
3. Τα $\bar{x}$ και S είναι ανεξάρτητα

Η κατανομή Wishart μπορεί να θεωρηθεί ως πολυδιάστατη γενίκευση της κατανομής x^2. Πράγματι για $p=1$ και $|\Sigma|=1$ είναι η κατανομή x^2 με n βαθμούς ελευθερίας. Από την άλλη αν στο μοντέλο γέννησης της κατανομής αφήσουμε τα τυχαία διανύσματα να μην έχουν μέσα διανύσματα O δηλαδή αν $x_i \sim N_p(\mu_i,\Sigma)$ $i=1,...,n$, τότε προκύπτει η μη κεντρική κατανομή Wishart $W_p(n,\Sigma,\mu)$ όπου $\mu=\sum_{i=1}^{n}\mu_i$ είναι η παράμετρος μη κεντρικότητας.

Η μεγάλη χρησιμότητα, την οποία παρουσιάζει η κατανομή Wishart στη πολυμεταβλητή στατιστική στηρίζεται στο γεγονός, πως είναι η κατανομή δειγματοληψίας του δειγματικού πίνακα διακύμανσης.

Συγκεκριμένα ισχύει ότι:

<u>Θεώρημα</u>:

Έστω $X_1,...,X_n$ ανεξάρτητα και ισόνομα με κατανομή $\mathrm{N}_\kappa(\mu,\Sigma)$, $\Sigma>0$ και $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, $S=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})'$ Τότε έχουμε τα εξής:

1) $\overline{X} \sim \mathrm{N}_\kappa(\mu,\Sigma/n)$
2) $S \sim W_\kappa\left(n-1,\frac{1}{n-1}\Sigma\right)$
3) $\overline{X}$ και S είναι ανεξάρτητα

Τώρα λοιπόν μπορούμε να δείξουμε ότι,

1) ο $\overline{X}$ είναι ο ΑΟΕΔ εκτιμητής της παραμέτρου μ και
2) ο S είναι ο ΑΟΕΔ εκτιμητής της παραμέτρου Σ.

Απόδειξη:

1) $\overline{X} \sim \mathrm{N}_{\kappa}\left(\mu,\Sigma/n\right)$ συνεπώς $E\left(\overline{X}\right)=\mu$. Άρα ο $\overline{X}$ είναι αμερόληπτος εκτιμητής της παραμέτρου μ. Είναι όμως και συνάρτηση της επαρκούς και πλήρους στατιστικής συνάρτησης $\left(\overline{X},S\right)$. Άρα από το θεώρημα Lehmann – Scheffé είναι ο ΑΟΕΔ εκτιμητής του μ.

2) $S \sim W_{\kappa}\left(n-1,\frac{1}{n-1}\Sigma\right)$ και άρα $E\left(S\right)=\left(n-1\right)\left(\frac{1}{n-1}\Sigma\right)=\Sigma$. Άρα S είναι αμερόληπτος εκτιμητής της παραμέτρου Σ. Είναι όμως και συνάρτηση της επαρκούς και πλήρους στατιστικής συνάρτησης$\left(\overline{X},S\right)$. Άρα από το θεώρημα Lehmann – Scheffé είναι ο ΑΟΕΔ εκτιμητής του Σ.

Το θεώρημα αυτό είναι η αιτία για να χρησιμοποιήσουμε τον αμερόληπτο δειγματικό πίνακα διακύμανσης.
Παρατηρούμε δε ότι η αναμενόμενη τιμή του δειγματικού πίνακα διακύμανσης δεν είναι αμερόληπτη και γι' αυτό εμφανίζεται το $\left(n-1\right)$ στον παρανομαστή ώστε η εκτιμήτρια να είναι αμερόληπτη.

Από τα τρία αποτελέσματα που αναφέραμε προηγουμένως συμπεραίνουμε ότι επειδή το Σ είναι άγνωστο η κατανομή του $\bar{x}$ δεν μπορεί να χρησιμοποιηθεί άμεσα για να βγάλουμε συμπεράσματα για τη μέση τιμή μ.
Ομοίως το S παρέχει ανεξάρτητες πληροφορίες για το Σ και η κατανομή του S δεν εξαρτάται από το μ. Έτσι λοιπόν μας δίνεται η δυνατότητα να κατασκευάσουμε ένα στατιστικό στοιχείο για να βγάλουμε συμπεράσματα για το μ.

Μερικές ιδιότητες της κατανομής Wishart εμφανίζονται παρακάτω:

(1) Εάν $A_1 \sim W_{m1}(A_1 | \Sigma)$ και $A_2 \sim W_{m2}(A_2 | \Sigma)$, τότε $A_1 + A_2 \sim W_{m1} + W_{m2}(A_1 + A_2 | \Sigma)$.

(2) Παρατηρούμε ότι οι δυο πίνακες πρέπει να έχουν τον ίδιο πίνακα Σ για να ισχύει το αποτέλεσμα. Προφανώς, το αποτέλεσμα γενικεύεται και για περισσότερους από 2 πίνακες.

(3) Αν ο A ακολουθεί την κατανομή Wishart με $W_m(A|\Sigma)$, τότε CAC' κατανέμεται σε $W_m(CAC'|C\Sigma C')$

Το αποτέλεσμα έχει το εξής ενδιαφέρον. Αν ξεκινήσουμε από ένα τυχαίο διάνυσμα το οποίο μετασχηματίζουμε πολλαπλασιάζοντάς το με έναν κατάλληλο πίνακα τότε η διακύμανση του καινούργιου διανύσματος θα δίνεται από μια έκφραση που βλέπουμε στο δεξί μέλος της παραπάνω ιδιότητας. Αυτό σημαίνει ότι η διακύμανση του καινούργιου διανύσματος θα ακολουθεί και αυτή κατανομή Wishart, αν αυτό συνέβαινε για τη διακύμανση του αρχικού διανύσματος.

Αν και δεν έχουμε κάποια ιδιαίτερη ανάγκη για τη λειτουργία της πυκνότητας πιθανότητας της κατανομής Wishart, μπορεί να έχει κάποιο ενδιαφέρον να δούμε περαιτέρω περίπλοκη μορφή. Η πυκνότητα δεν υπάρχει εκτός αν το μέγεθος n του δείγματος είναι μεγαλύτερο από τον αριθμό p των μεταβλητών. Όταν η πυκνότητα υπάρχει, η τιμή κατά τον θετικά ορισμένο πίνακα A είναι:

$$W_{n-1}(A|\Sigma) = \frac{|A|^{(n-p-2)/2} e^{-trA\Sigma^{-1}/2}}{2^{P(n-1)/2} \pi^{p(p-1)/4} |\Sigma|^{(n-1)/2} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n-i)\right)}$$

όπου $\Gamma(\cdot)$ είναι γάμμα κατανομή.

Αν αντικαταστήσουμε όπου $p=1$ και $|\Sigma|=1$, η κατανομή είναι η x^2 με n βαθμούς ελευθερίας. Μπορεί κανείς να παρατηρήσει ότι και από τον τρόπο που προκύπτει η κατανομή Wishart είναι γενίκευση της x^2 κατανομής.

Τέλος, η κατανομή Wishart ορίζεται μόνο από συμμετρικούς πίνακες. Επομένως και ο πίνακας Σ να είναι συμμετρικός.

Αμερόληπτοι εκτιμητές ελάχιστης διασποράς

Θα αναζητήσουμε να βρούμε τους αμερόληπτους εκτιμητές ελάχιστης διασποράς των παραμέτρων μ και Σ, μέσω της επάρκειας και της πληρότητας. Η μεθοδολογία αυτή αναπτύχθηκε από τους Lehmann και Scheffé

Ορισμός

Η στατιστική συνάρτηση $\overline{T} = T(\mathrm{X})$ ονομάζεται πλήρης, αν ισχύει η ακόλουθη σχέση:

$$E_\theta\left(\Phi(\mathrm{T})\right) = 0, \quad \forall t \in \Theta \Rightarrow \Phi(\mathrm{T}) = 0,$$

Για κάθε δυνατή τιμή t της T, δηλαδή $\Phi(\mathrm{T})$

Θεώρημα Lehmann και Scheffé

Έστω διάνυσμα X και $T = T(\mathrm{X})$ είναι επαρκής και πλήρης στατιστική συνάρτηση και S είναι ένας αμερόληπτος εκτιμητής της $g(\theta)$. Τότε $S^* = E_\theta(S / \mathrm{T})$ είναι ο μοναδικός Αμερόληπτος Εκτιμητής Ελάχιστης Διασποράς (ΑΟΕΔ Εκτιμητής) της $g(\theta)$.

Η έννοια της επάρκειας είναι μια από τις πιο σημαντικές έννοιες της Στατιστικής Συμπερασματολογίας. Εισήχθηκε από τον Fisher στις αρχές του 20ου αιώνα και χαρακτηρίζει τις στατιστικές συναρτήσεις $T(\mathrm{X})$ που έχουν την ιδιότητα να περιέχουν όλες τις πληροφορίες για την άγνωστη παράμετρο θ που περιέχει και τα δείγματα $X_1, X_2, ..., X_n$.

Παραγοντικό Κριτήριο Επάρκειας

Αν Y είναι τυχαίο διάνυσμα με πυκνότητα $f\left(Y/\theta\right)$ όπου θ είναι κάποια παράμετρος.

Αν $f\left(Y/\theta\right)=h\left(Y\right)g\left(T\left(Y\right),\theta\right)$ τότε η $T\left(Y\right)$ είναι επαρκής.

Απ' την άλλη, σχετικά με την πληρότητα, λέμε ότι η στατιστική συνάρτηση $T\left(Y\right)$ είναι πλήρης αν η σχέση,

$$E_\theta\phi\left(\mathrm{T}\right)=0,\ \forall\theta\in\Theta,$$

Συνεπάγεται ότι $\phi\left(\mathrm{T}\right)=0$, δηλαδή $\phi\left(t\right)=0$ για κάθε τιμή t της T.

Από τη Στατιστική γνωρίζουμε ότι αν $T=T\left(Y\right)$ επαρκής και $T^*=h\left(T\right)$ είναι 1-1 στατιστική συνάρτηση του T, τότε η T^* είναι επίσης επαρκής. Ανάλογα, εάν η $T=T\left(Y\right)$ είναι πλήρης, τότε η $T_1=g\left(T\right)$ είναι πλήρης.

Ο δειγματικός μέσος που θα χρησιμοποιηθεί για την εκτίμηση του μ είναι ο $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ ενώ ο δειγματικός πίνακας συνδιακύμανσης $S=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)'$ θα χρησιμοποιηθεί για την εκτίμηση του Σ.

Σ' αυτό το σημείο είναι σημαντικό να δείξουμε ότι το $\left(\overline{X},S\right)$ είναι επαρκής και πλήρης στατιστική συνάρτηση για την παράμετρο $\theta=\left(\mu,\Sigma\right)$. Μπορεί εύκολα να δειχθεί ότι συνδιακύμανσης $S=\frac{1}{n-1}\left[\sum_{i=1}^{n}X_iX_i'-n\left(\overline{XX'}\right)\right]$. Απ' αυτή τη σχέση συμπεραίνουμε ότι $\left(\overline{X},S\right)$ είναι 1-1 συνάρτηση του $\left(\sum_{i=1}^{n}X_i,\sum_{i=1}^{n}X_iX_i'\right)$.

Συνεχίζουμε την απόδειξη και τελικά με τη βοήθεια του παραγοντικού κριτηρίου καταλήγουμε στο ότι $\left(\sum_{i=1}^{n}X_i,\sum_{i=1}^{n}X_iX_i'\right)$ είναι επαρκής στατιστική

συνάρτηση. Άρα, και η $\left(\overline{X},S\right)$ είναι επαρκής στατιστική συνάρτηση για την παράμετρο $\theta=\left(\mu,\Sigma\right)$.

Για την πληρότητα αρκεί να αποδείξουμε ότι η από κοινού πυκνότητα των $X_1,...,X_n$ ανήκει στην εκθετική οικογένεια και κατόπιν χρησιμοποιούμε και την πιο κάτω πρόταση.

Πρόταση: (Εκθετική οικογένεια) Εάν το Y έχει πυκνότητα $f\left(y/\theta\right)$ στη μορφή,

$$f\left(y/\theta\right)=A\left(\theta\right)h\left(y\right)e^{Q^{'}(\theta)T(y)}.$$

Όπου $Q^{'}\left(\theta\right)$ συμβολίζει το εσωτερικό γινόμενο (και το πεδίο τιμών του διανύσματος $Q\left(\theta\right)$ έχει κάποιο εσωτερικό σημείο) τότε $T\left(Y\right)$ είναι επαρκής και πλήρης.

Καταλήγουμε λοιπόν στο ότι η $\left(\overline{X},S\right)$ είναι επαρκής και πλήρης στατιστική συνάρτηση για την παράμετρο $\theta=\left(\mu,\Sigma\right)$. Σ' αυτό το σημείο θα πρέπει να βρούμε τις κατανομές των $\overline{X}$ και S έτσι ώστε να βρούμε τους αμερόληπτους εκτιμητές. Η κατανομή του $\overline{X}$ είναι εύκολο να βρεθεί αφού είναι γραμμικός συνδυασμός ανεξάρτητων πολυδιάστατων κανονικών κατανομών. Για την κατανομή του S όμως, θα χρησιμοποιήσουμε την κατανομή Wishart. Έχουμε λοιπόν το πιο κάτω θεώρημα.

Μέθοδος Μέγιστης Πιθανοφάνειας

Θα ξεκινήσουμε με το παρακάτω παράδειγμα. Έστω ότι η τυχαία μεταβλητή X ακολουθεί $B(1, p)$ κατανομή με άγνωστη την παράμετρο p.

Η συνάρτηση πιθανότητας X έχει τη μορφή:

$$f(x; p) = p^x \cdot (1-p)^{1-x}, \quad x = 0,1, \quad 0 \le p \le 1.$$

Το πεδίο τιμών της παραμέτρου p είναι:

$$\Omega = \{ p : 0 \le p \le 1 \}.$$

Το πρόβλημά μας είναι δοθέντος ενός τυχαίου δείγματος $X_1, X_2, ..., X_n$ να υπολογίσουμε έναν εκτιμητή $T(X_1, X_2, ..., X_n)$ τέτοιο ώστε $T(x_1, x_2, ..., x_n)$ να είναι ένας καλός εκτιμητής σε σημείο του p, δηλαδή να είναι μια τιμή που να πλησιάζει όσο το δυνατόν περισσότερο τον πραγματικό p, όπου $x_1, x_2, ..., x_n$ είναι οι παρατηρούμενες τιμές του τυχαίου δείγματος.

Η κοινή συνάρτηση πιθανότητας του δείγματος

$X' = (X_1, ..., X_n)$ είναι:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \prod_{i=1}^{n} p^x \cdot (1-p)^{1-x} = p^{\Sigma x} \cdot (1-p)^{n-\Sigma x}$$

Θεωρώντας τη συνάρτηση αυτή ως συνάρτηση της παραμέτρου p, θα μπορούμε σαν εκτιμητή p να λάβουμε εκείνη την τιμή της p, που μεγιστοποιεί την κοινή συνάρτηση πιθανότητας.

Τη συνάρτηση $p^{\Sigma x} \cdot (1-p)^{n-\Sigma x}$ την ονομάζουμε συνάρτηση πιθανοφάνειας και τη συμβολίζουμε με $L(p)$.

Η $L(p)$ μεγιστοποιείται για την τιμή του p, για την οποία

$\frac{dL(p)}{dp} = 0$ ή

$$\frac{dL(p)}{dp} = (\Sigma x_i)\cdot p^{\Sigma x_i - 1}(1-p)^{n-\Sigma x_i} - (n-\Sigma x_i)\, p^{\Sigma x_i}(1-p)^{n-\Sigma x_i - 1} = 0$$

ή $\frac{\Sigma x_i}{p} - \frac{n-\Sigma x_i}{1-p} = 0$ ή $\Sigma x_i - np = 0$ ή $p = \frac{\Sigma x_i}{n} = \bar{x}$

Η στατιστική συνάρτηση $\frac{\Sigma x_i}{n}$ λέγεται εκτιμητής μέγιστης πιθανοφάνειας και το συμβολίζουμε με $\hat{p}$.

Θα δώσουμε, τώρα, τον ορισμό της συνάρτησης πιθανοφάνειας:
Έστω $X_1, X_2, ..., X_n$ ένα τυχαίο δείγμα από μια κατανομή που εξαρτάται από μια άγνωστη παράμετρο θ. Εάν $f(x,\theta)$ η συνάρτηση πιθανότητας ή πυκνότητας πιθανότητας της κατανομής και $\theta \in \Omega$, τότε την κοινή συνάρτηση πιθανότητας ή πυκνότητας πιθανότητας του δείγματος.

$f(x_1;\theta)\cdot f(x_2;\theta)\cdot ...\cdot f(x_n;\theta)$

τη συμβολίζουμε με $L(\theta)$ και την ονομάζουμε συνάρτηση πιθανοφάνειας.

Δηλαδή $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$

<u>Εκτιμητής μέγιστης Πιθανοφάνειας (ΕΜΠ)</u> της παραμέτρου θ, ονομάζεται η συνάρτηση $\hat{\theta} = T(x_1, x_2, ..., x_n)$ που μεγιστοποιεί την $L(\theta)$.

Έτσι, λοιπόν, συνοψίζοντας έχουμε, ότι για την εύρεση της εκτιμήτριας μέγιστης πιθανοφάνειας (εμπ) μιας παραμέτρου θ, μεγιστοποιούμε ως προς το θ τη συνάρτηση πιθανοφάνειας. Η τιμή του θ, έστω $\hat{\theta}$, που μεγιστοποιεί αυτή τη συνάρτηση, λέγεται εκτιμήτρια μέγιστης πιθανοφάνειας του θ. Η συνήθης διαδικασία για την εύρεση της ε.μ.π του θ είναι να παραγωγίσουμε τη συνάρτηση πιθανοφάνειας ως προς θ και να θέσουμε την παράγωγο ίση με μηδέν. Η λύση $\hat{\theta}$ της προκύπτουσας εξίσωσης είναι η ε.μ.π του θ.

Στην περίπτωση, όμως, της $\kappa - διάστατης$ κανονικής κατανομής $N_\kappa(\mu,\Sigma)$ με μ και Σ άγνωστα, η παράμετρος θ γίνεται $\theta=(\mu,\Sigma)$. δηλαδή, έχουμε ένα διάνυσμα και έναν πίνακα.

Δυσκολία, όμως, παρουσιάζει η διαδικασία να ορίσουμε παραγώγιση ως προς μεταβλητή πίνακα για να έχουμε τη δυνατότητα να εφαρμόσουμε την παραπάνω διαδικασία.

Επομένως, θα χρησιμοποιήσουμε μια άλλη μέθοδο για την εύρεση της ε.μ.π. των μ και Σ.

Έστω ότι έχουμε ένα δείγμα μεγέθους n από πολυμεταβλητή κατανομή, δηλαδή, υποθέτουμε ότι $X \sim N_p(\mu,\Sigma)$, $i=1,2,...,n$ και είναι ανεξάρτητα.

Τότε, η πιθανοφάνεια του δείγματος δίνεται από τη σχέση:

$$L = f(x_1,x_1,...,x_n) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}\cdot|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]$$

και άρα:

$$L = (2\pi)^{-np/2}\cdot|\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\mu)'\Sigma^{-1}(x_i-\mu)'\right]\right\}$$

Λογαριθμίζοντας έχουμε ότι:

$$l = \log L = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{n}{2}\left[\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right]$$

$$= cons\tan t - \frac{n}{2}\log|\Sigma| - \frac{n}{2}tr\left(\Sigma^{-1}S\right) - \frac{n}{2}\left(\bar{x}-\mu\right)\Sigma^{-1}\left(\bar{x}-\mu\right)$$

Όμως, από τη γραμμική άλγεβρα ξέρουμε ότι, αν τα x_i είναι διανύσματα $px1$ και Α πίνακας pxp, τότε ισχύει $\sum_{i=1}^{n} x_i' A x_i = tr(AT)$

Όπου $T = \sum_{i=1}^{n} x_i \cdot x_i'$ είναι ένας πίνακας διαστάσεων pxp

Άρα

$$\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)'\Sigma^{-1}\left(x_i-\bar{x}\right)\right]=$$

$$\sum_{i=1}^{n}tr\left[\Sigma^{-1}\left(x_i-\bar{x}\right)'\left(x_i-\bar{x}\right)\right]=$$

$$tr\left\{\Sigma^{-1}\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)'\left(x_i-\bar{x}\right)\right]\right\}=$$

$$tr\left\{\Sigma^{-1}\left(nS\right)\right\}=ntr\left(\Sigma^{-1}S\right)$$

$$\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)'\Sigma^{-1}\left(x_i-\bar{x}\right)\right]+n\left(\bar{x}-\mu\right)'\Sigma^{-1}\left(\bar{x}-\mu\right)=$$

$$ntr\left(\Sigma^{-1}S\right)+n\sum_{i=1}^{n}\left[\left(x_i-\mu\right)'\Sigma^{-1}\left(x_i-\mu\right)\right]$$

Άρα για να μεγιστοποιήσει κανείς τη πιθανοφάνεια ως προς μ, ουσιαστικά ενδιαφέρεται για την ποσότητα

$$-\frac{n}{2}\left(\bar{x}-\mu\right)'\Sigma^{-1}\left(\bar{x}-\mu\right)$$

η οποία είναι οποιοσδήποτε αρνητικός αριθμός ως το αρνητικό μιας τετραγωνικής μορφής και άρα για $\mu=\bar{x}$ η συνάρτηση μεγιστοποιείται.

Θα δώσουμε τώρα δυο παραδείγματα για να παρακολουθήσουμε πως υπολογίζεται ο εκτιμητής μέγιστης πιθανοφάνειας για το θ και στο δεύτερο παράδειγμα ο εκτιμητής άγνωστης παραμέτρου μ.

Παράδειγμα 1

Έστω τυχαίο δείγμα $X_1, X_2, ..., X_n$ από εκθετική κατανομή με συνάρτηση πυκνότητας πιθανότητας

$$f(x;\theta) = \frac{1}{\theta} e^{\frac{x}{\theta}}, \quad 0 < x < \infty, \quad \theta \in \Theta = \{\theta : 0 < \theta < \infty\}$$

α) να υπολογιστεί ένας εκτιμητής μέγιστης πιθανοφάνειας για την παράμετρο θ

β) εάν από ένα δείγμα μεγέθους s πήραμε τις παρακάτω μετρήσεις: 3.5, 8.1, 0.9, 4.4, και 0.5 να δοθεί τιμή του εκτιμητή που υπολογίσατε.

Λύση

α) Η συνάρτηση πιθανοφάνειας του δείγματος είναι:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \left(\frac{1}{\theta} e^{-\frac{x}{\theta}}\right) \cdot \left(\frac{1}{\theta} e^{-\frac{x}{\theta}}\right) \cdot \ldots \cdot \left(\frac{1}{\theta} e^{-\frac{x}{\theta}}\right) =$$

$$\frac{1}{\theta^n} \cdot e^{\frac{\Sigma x}{\theta}}$$

Ο λογάριθμος της συνάρτησης πιθανοφάνειας είναι:

$$ln\left(L(\theta)\right) = ln(\theta) = \frac{1}{\theta} \sum_{i=1}^{n} x_i, \quad 0 < \theta < \infty$$

Η παραπάνω συνάρτηση μεγιστοποιείται στα σημεία που μηδενίζεται η παράγωγος της ως προς την παράμετρο Θ.

Έτσι έχουμε

$$\frac{dL(\theta)}{d\theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i = 0$$

$$\hat{\theta} = \frac{\sum x_i}{n} = \bar{x}$$

Ο εκτιμητής λοιπόν μέγιστης πιθανοφάνειας για την παράμετρο Θ είναι η δειγματική μέση τιμή.

β) Εφόσον, πλέον καταλήξαμε στη σχέση $\hat{\theta} = \frac{\Sigma xi}{n} = \bar{x}$ αντικαθιστώντας τις τιμές τις οποίες έχει δώσει η άσκηση θα προκύψει η εξής σχέση:

$$\bar{x} = \frac{3.5 + 8.1 + 0.9 + 4.4 + 0.5}{5} = 3.48$$

Συνεπώς, για συγκεκριμένες τιμές του δείγματος έχουμε ότι ο Εκτιμητής Μέγιστης Πιθανοφάνειας (ΕΜΠ) θα είναι $\hat{\theta} = 3.48$

Παράδειγμα 2

Υποθέτουμε, ότι έχουμε τυχαίο δείγμα από πληθυσμό με κανονική κατανομή, του οποίου η μέση τιμή είναι μ και η διασπορά 1. Με τη μέθοδο μέγιστης πιθανοφάνειας, να υπολογιστεί ένας εκτιμητής της άγνωστης παραμέτρου μ.

Έστω το τυχαίο δείγμα $X_1, X_2, ..., X_n$ από $\mathrm{N}(\mu, 1)$

Η συνάρτηση πιθανοφάνειας του δείγματος είναι:

$$L(\mu) = (2\pi)^{-n/2} exp\left[-\frac{1}{2} \sum_{j=1}^{n} (x_j - \mu)^2 \right]$$

$$\frac{\partial \ln L}{\partial \mu} = \sum_{j=1}^{n} (x_j - \mu) = 0$$

Συνεπώς, προκύπτει ότι $\mu = \bar{x}$ η τιμή που μεγιστοποιεί τη συνάρτηση πιθανοφάνειας, αφού μάλιστα ισχύει ότι:

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -n < 0$$

Η συμπεριφορά των $\bar{x}$ και S σε μεγάλο δείγμα

Ας υποθέσουμε ότι η ποσότητα X που καθορίζεται από ένα μεγάλο αριθμό ανεξάρτητων αιτιών $V_1, V_2, ..., V_n$, όπου οι τυχαίες μεταβλητές V_i εκπροσωπούν τα αίτια που έχουν περίπου την ίδια μεταβλητότητα.

Εάν πάρουμε ότι το X είναι το άθροισμα:

$$X = V_1 + V_2 + ... + V_n$$

τότε το κεντρικό οριακό θεωρημένο ισχύει και καταλήγουμε ότι το X έχει μια κατανομή που είναι σχεδόν κανονική.

Αυτό ισχύει για σχεδόν οποιαδήποτε μητρική κατανομή των V_i

Το κεντρικό οριακό θεώρημα με μια μεταβλητή μας λέει επίσης ότι η κατανομή δειγματοληψίας της μέσης τιμής $\overline{X}$, για μεγάλο μέγεθος δείγματος είναι σχεδόν κανονική ανεξάρτητα από την κατανομή του πληθυσμού. Ένα παρόμοιο αποτέλεσμα ισχύει και για πολλές άλλες σημαντικές μονομεταβλητές περιπτώσεις.

Συμπερασματικά, έχουμε ότι ορισμένες πολυμεταβλητές στατιστικές όπως τα $\bar{x}$ και S, έχουν ιδιότητες ενός μεγάλου δείγματος ανάλογες με τα μονομεταβλητά. Όπως το μέγεθος του δείγματος αυξάνεται, χωρίς να δεσμεύεται, ορισμένες κανονικότητες διέπουν τη διακύμανση δειγματοληψίας των $\bar{x}$ και S, ανεξάρτητα από το πια είναι η μορφή του αρχικού πληθυσμού.

Άρα, δεν χρειάζεται να έχουμε μόνο πολυμεταβλητό πληθυσμό, αλλά ο πληθυσμός που παίρνουμε είναι να έχει μια μέση τιμή μ και μια πεπερασμένη συνδιακύμανση S.

Αποτέλεσμα 12 Ο Νόμος των Μεγάλων Αριθμών

Ας πάρουμε $Y_1, Y_2, ..., Y_n$ να είναι ανεξάρτητες παρατηρήσεις από έναν πληθυσμό με μέση τιμή $E(Y_i) = \mu$ Τότε ισχύει ότι:

$$\overline{Y} = \frac{Y_1 + Y_2 + ... + Y_n}{n}$$

που συγκλίνει στη πιθανότητα του μ σαν n αυξήσεις χωρίς περιορισμούς. Αυτό γίνεται για κάθε προκαθορισμένο $\varepsilon > 0$, όπου η πιθανότητα $P\left[-\varepsilon < \overline{Y} - \mu < \varepsilon\right]$ προσεγγίζει το 1 όταν $n \to \infty$.

Απόδειξη

Σαν μια ευθεία συνέπεια του νόμου των μεγάλων αριθμών, ο οποίος λέει ότι για κάθε μέση τιμή $\overline{X_i}$, η οποία συγκλίνει στην πιθανότητα μ_i με $i = 1, 2, ..., p$ τότε το $\overline{X}$ συγκλίνει στην πιθανότητα μ.

Επιπλέον κάθε συνδυακύμανση $S_{i,k}$ δείγματος συγκλίνει στην πιθανότητα $\sigma_{i,k}$ για $i, k = 1, 2, ..., p$ και

$S\left(n \text{ εκτιμητής } \hat{\Sigma} = S_n\right)$ συγκλίνει στην πιθανότητα Σ.

Έτσι έχουμε ότι:

$$(n-1)s_{ik} = \sum_{j=1}^{n} (X_{ij} - \overline{X}_i)(X_{kj} - \overline{X}_k) =$$

$$\sum_{j=1}^{n} (X_{ij} - \mu_i + \mu_i - \overline{X}_i)(X_{kj} - \mu_k + \mu_k - \overline{X}_k) =$$

$$\sum_{j=1}^{n} (X_{ij} - \mu_i) + (X_{kj} - \mu_k) + n(\overline{X}_i - \mu_i)(\overline{X}_k - \mu_k)$$

Ας πάρουμε το $Y_j = (X_{ij} - \mu_i)(X_{kj} - \mu_k)$ με μέση τιμή για το Y_j να είναι $E(Y_{jk}) = \sigma_{ik}$ Εδώ παρατηρούμε ότι ο πρώτος όρος του $s_{i,k}$ συγκλίνει στο $\sigma_{i,k}$ και ο δεύτερος όρος συγκλίνει στο μηδέν, εφαρμόζοντας τον Νόμο των Μεγάλων Αριθμών.

Η πρακτική ερμηνεία όλων αυτών είναι ότι όταν έχουμε μεγάλη πιθανότητα, το $\overline{X}$ θα συγκλίνει στο μ και το S θα συγκλίνει στο Σ, κάθε φορά που το μέγεθος του δείγματος είναι μεγάλο. Όλο αυτό αφορά το $\overline{X}$ έγινε πιο ακριβής για μια πολυμεταβλητή έκδοση του κεντρικού οριακού θεωρήματος.

Αποτέλεσμα 13 Κεντρικό Οριακό Θεώρημα

Ας πάρουμε ότι $X_1, X_2, ..., X_n$ είναι ανεξάρτητες από ένα πληθυσμό που έχει μέση τιμή μ και πεπερασμένη συνδιακύμανση Σ.

Τότε το $\sqrt{n}(\bar{x} - \mu)$ έχει μια κατά προσέγγιση κανονική κατανομή $Np(0, \Sigma)$ για μεγάλα μεγέθη δείγματος. Εδώ το n θα έπρεπε να είναι μεγάλο σε σχέση με το p.

Η προσέγγιση που γίνεται από το Κεντρικό Οριακό Θεώρημα ισχύει για διακριτούς πληθυσμούς, καθώς και για n συνεχής πολυμεταβλητούς πληθυσμούς. Μαθηματικά, το όριο είναι ακριβές και η προσέγγιση στην ομαλότητα είναι αρκετά γρήγορη. Επιπλέον, γνωρίζουμε ότι η μέση τιμή $\overline{X}$ κατανέμεται κανονικά. Έτσι θα περιμέναμε ότι το Κεντρικό Οριακό Θεώρημα να είναι αρκετά καλό για n, όταν ο πληθυσμός είναι κανονικός.

Όπως έχουμε παρατηρήσει όταν το n είναι μεγάλο, το S είναι κοντά στο Σ με μεγάλη πιθανότητα. Κατά συνέπεια αντικαθιστώντας το Σ από το S στην κατά προσέγγιση κανονική κατανομή, θα έχουμε μια αμελητέα επίδραση στους μετέπειτα υπολογισμούς των πιθανοτήτων.

Το X_p^2 είναι περίπου η κατανομή δειγματοληψίας του $n(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)$ όταν το $\overline{X}$ κατανέμεται $Np(\mu,\frac{1}{n}\Sigma)$ ή ισοδύναμα όταν $\sqrt{n}(\bar{x}-\mu)$ κατανέμεται στο $Np(0,\Sigma)$. Αντικαθιστώντας το Σ^{-1} με τον εκτιμητή S^{-1}, δεν επηρεάζεται η παρούσα προσέγγιση για μεγάλο n και πολύ μεγαλύτερη από το p.

Συνοψίζοντας έχουμε τα εξής:

Ας πάρουμε ότι $X_1, X_2, ..., Xn$ είναι ανεξάρτητες παρατηρήσεις από ένα πληθυσμό με μέση τιμή μ και πεπερασμένη συνδιασπορά Σ Τότε έχουμε ότι:

το $\sqrt{n}(\bar{x}-\mu)$ προσεγγίζει την κανονική κατανομή $Np(0,\Sigma)$ και $n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu)$ είναι περίπου X_p^2.

Δειγματοληψία από μια Πολυμεταβλητή κανονική κατανομή και τη Μέγιστη Πιθανοφάνεια

Σε αυτή την ενότητα θα ασχοληθούμε με τα δείγματα από έναν πολυμεταβλητό κανονικό πληθυσμό και ιδίως με την κατανομή δειγματοληψίας του $\overline{X}$ και του S.

Η πολυμεταβλητή κανονική πιθανότητα

Οι άγνωστες παράμετροι της πολυμεταβλητής κανονικής κατανομής είναι το διάνυσμα των μέσων μ (διάνυσμα $px1$) και ο πίνακας Σ διαστάσεων pxp.

Ας υποθέσουμε λοιπόν ότι τα $px1$ διανύσματα $X_1,...,X_n$ τα οποία αποτελούν ένα τυχαίο δείγμα από έναν πολυμεταβλητό κανονικό πληθυσμό με μέση τιμή το διάνυσμα μ και πίνακα συνδυασποράς Σ όπως προείπαμε.

Δεδομένου πως ο πίνακας Σ είναι συμμετρικός δεν χρειάζεται κανείς να εκτιμήσει όλα του τα στοιχεία.

Στην πραγματικότητα χρειάζεται να εκτιμήσει $p(p+1)/2$ παραμέτρους (διακυμάνσεις και συνδυακυμάνσεις) και άλλες p παραμέτρους από το διάνυσμα των μέσων, άρα συνολικά $p(p+1)/2+p$ άγνωστες παραμέτρους. Καταλαβαίνει κανείς πως το πρόβλημα της εκτίμησης είναι αρκετά πολύπλοκο καθώς χρειαζόμαστε $p(p+3)/2$ εξισώσεις.

Έτσι λοιπόν δεδομένου ότι τα $X_1,...,X_n$ είναι ανεξάρτητα μεταξύ τους και το καθένα έχει κατανομή $N_p(\mu,\Sigma)$, η από κοινού συνάρτηση πυκνότητας

απ΄ όλες τις παρατηρήσεις αποτελεί το προϊόν των οριακών κανονικών πυκνοτήτων.

Η από κοινού συνάρτηση πυκνότητας:

$$L = f\left(X_1, X_2, ..., X_n\right) = \prod_{j=1}^{n}\left\{\frac{1}{\left(2\pi\right)^{p/2}\left|\Sigma\right|^{1/2}} e^{-\left(X_j-\mu\right)'\Sigma^{-1}\left(X_j-\mu\right)/2}\right\} =$$

$$\frac{1}{\left(2\pi\right)^{\pi p/2}\left|\Sigma\right|^{\pi/2}} e^{-\sum_{j=1}^{n}\left(X_j-\mu\right)'\Sigma^{-1}\left(X_j-\mu\right)/2}$$

ισχύει ότι $tr\left(BC\right) = \sum_{i=1}^{m}\left(\sum_{j=1}^{\kappa} b_{ij}c_{ji}\right)$

Ομοίως ισχύει και για το $j-$στοιχείο της διαγωνίου του BC που είναι το $\sum_{i=1}^{m} c_{ji}b_{ij}$, όπου ισχύει ότι:

$$tr\left(CB\right) = \sum_{j=1}^{\kappa}\left(\sum_{i=1}^{m} c_{ji}b_{ij}\right) = \sum_{i=1}^{m}\left(\sum_{j=1}^{\kappa} b_{ij}c_{ji}\right) = tr\left(BC\right)$$

Ας πάρουμε το **x΄** να είναι ο πίνακας B Με $m=1$ και ας πάρουμε το Ax να παίξει το ρόλο του πίνακα C.

Όταν ισχύει $tr\left(x'\left(\mathrm{A}x\right)\right) = tr\left(\left(Ax\right)x'\right)$

(a) Αποδεικνύεται χρησιμοποιώντας την φασματική ανάλυση.

Θα γράψουμε ότι $A = P'\Lambda P$, όπου $PP' = I$ και Λ είναι ένας διαγώνιος πίνακας με καταχωρήσεις τα $\lambda_1, \lambda_2, ..., \lambda_\kappa$. Ως εκ τούτου, έχουμε ότι

$$tr\left(A\right) = tr\left(P'\Lambda P\right) = tr\left(\Lambda PP'\right) = tr\left(\Lambda\right) = \lambda_1 + \lambda_2 + ... + \lambda_\kappa$$

Τώρα, ο εκθέτης της από κοινού πυκνότητας μπορεί να απλοποιηθεί:

$$\left(x_j-\mu\right)'\sum^{-1}\left(x_j-\mu\right)=tr\left[\left(x_j-\mu\right)'\sum^{-1}\left(x_j-\mu\right)\right]=$$

$$tr\left[\sum^{-1}\left(x_j-\mu\right)\left(x_j-\mu\right)'\right]$$

Άρα έχουμε:

$$\sum_{j=1}^{n}\left(x_j-\mu\right)'\sum^{-1}\left(x_j-\mu\right)=\sum_{j=1}^{n}tr\left[\left(x_j-\mu\right)'\sum^{-1}\left(x_j-\mu\right)\right]=$$

$$\sum_{j=1}^{n}tr\left[\sum^{-1}\left(x_j-\mu\right)\left(x_j-\mu\right)'\right]=$$

$$=tr\left[\sum^{-1}\sum_{j=1}^{n}\left(x_j-\mu\right)\left(x_j-\mu\right)'\right]$$

Όπου το ίχνος του αθροίσματος των πινάκων είναι ίσο με τα αθροίσματα των ιχνών των πινάκων. Μπορούμε να προσθέσουμε και να αφαιρέσουμε το $\bar{x}=(1/n)\sum_{j=1}^{n}x_j$ σε καθένα όρο $\left(x_j-\mu\right)$ στο $\sum_{j=1}^{n}\left(x_j-\mu\right)\left(x_j-\mu\right)'$ για να δώσει:

$$\sum_{j=1}^{n}\left(x_j-\bar{x}+\bar{x}-\mu\right)\left(x_j-\bar{x}+\bar{x}-\mu\right)'=\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'+\sum_{j=1}^{n}\left(\bar{x}-\mu\right)\left(\bar{x}-\mu\right)'=$$

$$=\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'+n\left(\bar{x}-\mu\right)\left(\bar{x}-\mu\right)'$$

Επειδή ο λόγος των διαγωνίων $=\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(\bar{x}-\mu\right)'$ και $=\sum_{j=1}^{n}\left(\bar{x}-\mu\right)\left(x_j-\bar{x}\right)'$, είναι και οι δυο μήτρες της μηδενικής.

Κατά συνέπεια, η από κοινού πυκνότητα ενός τυχαίου δείγματος από μια πολυμεταβλητή κανονικού πληθυσμού γράφεται ως εξής:

$$\{\text{Από κοινού πυκνότητα } x_1, x_2, ..., x_n\} =$$

$$(2\pi)^{np/2} |\Sigma|^{-n/2} \, exp\left\{-tr\left[\sum^{-1}\left(\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)' + n\left(\bar{x}-\mu\right)\left(\bar{x}-\mu\right)'\right)\right]/2\right\}$$

Αντικαθιστώντας τις παρατηρούμενες τιμές $x_1, x_2, ..., x_n$ στην από κοινού πυκνότητα απόδοσης της λειτουργίας της πιθανότητας θα δείξουμε με αυτή τη λειτουργία με $L(\mu, \Sigma)$ να τονίσουμε το γεγονός ότι είναι μια λειτουργία στις παραμέτρους του πληθυσμού μ και Σ. Επιπλέον, όταν τα διανύσματα x_j περιέχουν τους συγκεκριμένους αριθμούς, παρατηρήθηκε ότι έχουμε:

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-tr\left[\sum^{-1}\left(\Sigma\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)' + n\left(\bar{x}-\mu\right)\left(\bar{x}-\mu\right)'\right)\right]/2}$$

Ειδικότερα θα κάνουμε χρήση του τύπου

$$tr\left[\sum^{-1}\left(\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)' + n\left(\bar{x}-\mu\right)\left(\bar{x}-\mu\right)'\right)\right] =$$

$$tr\left[\sum^{-1}\left(\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'\right)\right] + n\ tr\left[\sum^{-1}\left(\bar{x}-\mu\right)\left(\bar{x}-\mu\right)'\right]$$

$$tr\left[\sum^{-1}\left(\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'\right)\right] + n\left(\bar{x}-\mu\right)'\sum^{-1}\left(\bar{x}-\mu\right)$$

Εκτίμηση Μέγιστης Πιθανοφάνειας του μ και Σ

Το επόμενο αποτέλεσμα θα μας επιτρέψει να πετύχουμε την μέγιστη πιθανοφάνεια των μ και Σ

Αποτέλεσμα 10

Δίνεται ένας pxp συμμετρικός θετικά ορισμένος πίνακας $\mathbf{B}$ και βαθμωτό $\beta > 0$ όπου:

$$\frac{1}{|\boldsymbol{\Sigma}|^b} e^{-tr(\Sigma^{-1}\boldsymbol{B})/2} \leq \frac{1}{|\boldsymbol{B}|^b} (2b)^{pb} e^{-bp}$$

για όλα τα θετικά ορισμένα $\underset{(pxp)}{\boldsymbol{\Sigma}}$, με ισότητα $\boldsymbol{\Sigma} = (1/2b)\mathbf{B}$

Απόδειξη

Έστω ότι το $\mathbf{B}^{1/2}$ είναι συμμετρική τετραγωνική ρίζα του **B**, τέτοια ώστε

$\mathbf{B}^{1/2}\mathbf{B}^{1/2} = \mathbf{B}$, $\mathbf{B}^{1/2}\mathbf{B}^{-1/2} = I$, και $\mathbf{B}^{-1/2}\mathbf{B}^{-1/2} = \mathbf{B}^{-1}$

Όταν $\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B}) = \text{tr}[\Sigma^{-1}\mathbf{B}^{1/2})\mathbf{B}^{1/2}] = \text{tr}[\mathbf{B}^{1/2}(\Sigma^{-1}\mathbf{B}^{1/2})]$

Ας πάρουμε το n να είναι η ιδιοτιμή του $\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}$

Ο πίνακας λοιπόν αυτός είναι θετικά ορισμένος επειδή:

$y'\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}y = (\mathbf{B}^{1/2}y)'\boldsymbol{\Sigma}^{-1}(\mathbf{B}^{1/2}y) > 0$

εάν $\mathbf{B}^{1/2}y \neq 0$ ή ισοδύναμα $y \neq 0$. Επιπλέον, οι ιδιοτιμές n_i του $\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}$ είναι θετικές και με βάση τη γενική μορφή που ισχύει δηλαδή $tr(A) = \sum_{i=1}^{\kappa} \lambda_i$, όπου λ_i είναι ιδιοτιμές του A πίνακα, τότε έχουμε ότι $\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B}) = \text{tr}(\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}) = \sum_{i=1}^{p} n_i$

και επιπλέον $|\mathbf{B}^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{1/2}| = \prod_{i=1}^{p} n_i$

Έτσι λοιπόν μπορούμε να γράψουμε ότι:

$|\mathbf{B}^{1/2}\mathbf{\Sigma}^{-1}\ \mathbf{B}^{1/2}|=|\mathbf{B}^{1/2}||\mathbf{\Sigma}^{-1}||\ \mathbf{B}^{1/2}|=|\mathbf{\Sigma}^{-1}||\mathbf{B}^{1/2}||\ \mathbf{B}^{1/2}|=|\mathbf{\Sigma}^{-1}||\mathbf{B}|=\frac{1}{|\boldsymbol{\Sigma}|}|\mathbf{B}|$

ή $\quad \frac{1}{|\boldsymbol{\Sigma}|}=\frac{|\mathbf{B}^{1/2}\mathbf{\Sigma}^{-1}\ \mathbf{B}^{1/2}|}{|\boldsymbol{B}|}=\frac{\prod_{i=1}^{p} n_i}{|\boldsymbol{B}|}$

Συνδυάζοντας τα αποτελέσματα έχουμε ότι:

$$\frac{1}{|\boldsymbol{\Sigma}|^b}e^{-tr[\boldsymbol{\Sigma}^{-1}\boldsymbol{B}]/2}=\frac{\left(\prod_{i=1}^{p} n_i\right)^b}{|\boldsymbol{B}|}\ e^{-\sum_{i=1}^{p} n_i/2}=\frac{1}{|\boldsymbol{B}|^b}\ \prod_{i=1}^{p} n_i^b e^{-n_i/2}$$

Αλλά η λειτουργία $n^b e^{-n/2}$ έχει μέγιστη τιμή, σε σχέση με το n του $(2b)^b e^{-b}$ που συμβαίνουν στο $n=2b$

Η επιλογή $n_i=2b$ για κάθε i, ως εκ τούτου δίνει

$\frac{1}{|\boldsymbol{\Sigma}|^b}e^{-tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{B})/2}\leq\frac{1}{|\boldsymbol{B}|^b}(2b)^{pb}\ e^{-bp}$

Το ανώτερο όριο έχει επιτευχθεί όταν $\mathbf{\Sigma}=(1/2b)\mathbf{B}$ δεδομένου ότι:

$\mathbf{B}^{1/2}\mathbf{\Sigma}^{-1}\ \mathbf{B}^{1/2}=\mathbf{B}^{1/2}(2b)\mathbf{B}^{-1}\mathbf{B}^{1/2}=(2b)\ \underset{pXp}{I}$

και $\quad \text{tr}(\mathbf{\Sigma}^{-1}\mathbf{B})=\text{tr}\ [\mathbf{B}^{1/2}\ \mathbf{\Sigma}^{-1}\ )\mathbf{B}^{1/2}]=tr[(2b)\ \boldsymbol{I}]=2bp$

Εξάλλου:

$$\frac{1}{|\boldsymbol{\Sigma}|}=\frac{\left|\boldsymbol{B}^{1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}^{1/2}\right|}{|\boldsymbol{B}|}=\frac{|(2b)\boldsymbol{I}|}{|\boldsymbol{B}|}\ \frac{\left|(2b)I\right|}{|\boldsymbol{B}|}=\frac{(2b)^p}{|\boldsymbol{B}|}$$

Με απλή αντικατάσταση για $\text{tr}[\mathbf{\Sigma}^{-1}\mathbf{B}]$ και $1/|\mathbf{\Sigma}|^b$ το δεσμευμένο ισχύει.

Ο Εκτιμητής Μέγιστης Πιθανοφάνειας των $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ είναι εκείνες οι τιμές για τις οποίες έχουμε ότι $\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}$ και $\boldsymbol{\Sigma}=\hat{\boldsymbol{\Sigma}}$, οι οποίες προκύπτουν εύκολα αν αναλογιστούμε την ερμηνεία των παραμέτρων της πολυμεταβλητής κανονικής

κατανομής, (δειγματικό διάνυσμα μέσων και πίνακας διακύμανσης – συνδυακύμανσης) και για τις οποίες η τιμή της $L(\mu,\Sigma)$ μεγιστοποιείται.

Οι εκτιμήσεις πλέον $\widehat{\boldsymbol{\mu}}$ και $\widehat{\boldsymbol{\Sigma}}$ βασίζονται στις παρατηρηθέντες τιμές $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ μέσω των στατιστικών $\overline{\boldsymbol{x}}$ και $\boldsymbol{S}$.

Αποτέλεσμα 11

Έστω ότι $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ είναι ένα τυχαίο δείγμα για κανονικό πληθυσμό με μέση τιμή $\boldsymbol{\mu}$, και συνδυακύμανση $\boldsymbol{\Sigma}$.

Τότε ισχύουν τα εξής για τους εκτιμητές τους:

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}} \quad \text{και} \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n} (\boldsymbol{X_j} - \overline{\boldsymbol{X}})(\boldsymbol{X_j} - \overline{\boldsymbol{X}})' = \frac{(n-1)}{n}\,\boldsymbol{S}$$

Οι οποίοι αποτελούν τους εκτιμητές μέγιστης πιθανοφάνειας των $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$ αντίστοιχα.

Παρατηρείται ότι οι τιμές των $\overline{\boldsymbol{x}}$ και $\left(\frac{1}{n}\right)\sum_{j=1}^{n} (\boldsymbol{x_j} - \overline{\boldsymbol{x}})(\boldsymbol{x_j} - \overline{\boldsymbol{x}})'$ καλούνται εκτιμητές μέγιστης πιθανοφάνειας των $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$.

Απόδειξη

Ο εκθέτης της συνάρτησης πιθανότητας εκτός από πολυσύνθετο παράγοντα $-\frac{1}{2}$, τον οποίο έχει είναι:

$$\mathrm{tr}[\boldsymbol{\Sigma}^{-1}(\sum_{j=1}^{n} (\boldsymbol{x_j} - \overline{\boldsymbol{x}})(\boldsymbol{x_j} - \overline{\boldsymbol{x}})')] + n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})$$

Νωρίτερα αποδείξαμε (Αποτέλεσμα 1) ότι το $\boldsymbol{\Sigma^{-1}}$ είναι θετικά ορισμένο, έτσι το $(\overline{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma^{-1}}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) > \mathbf{0}$ εκτός εάν $\boldsymbol{\mu} = \overline{\boldsymbol{x}}$ που σημαίνει ότι είναι ίσο με το μηδέν. Έτσι βλέπουμε ότι η πιθανότητα μεγιστοποιείται για $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}$

Παραμένει η μεγιστοποίηση:

$$L(\widehat{\boldsymbol{\mu}} = \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n} (\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})'\right)\right]/2}$$

πάνω από $\boldsymbol{\Sigma}$.

Για $b = n/2$ και $\mathbf{B} = \sum_{j=1}^{n} (\boldsymbol{x}_j - \overline{\boldsymbol{x}})(\boldsymbol{x}_j - \overline{\boldsymbol{x}})'$ το μέγιστο εμφανίζεται στο $\widehat{\boldsymbol{\Sigma}} = (1/n)(\boldsymbol{x_j} - \overline{\boldsymbol{x}})(\boldsymbol{x_j} - \overline{\boldsymbol{x}})'$ όπως αναφέρεται.

Οι εκτιμητές μέγιστης πιθανοφάνειας αποτελούν τυχαίες ποσότητες. Έχουν παραχθεί με την αντικατάσταση των $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ των παρατηρήσεων για πλέον $\widehat{\boldsymbol{\mu}}$ και $\widehat{\boldsymbol{\Sigma}}$ με τα αντίστοιχα τυχαία διανύσματα $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.

Σημειώνουμε επίσης ότι ο εκτιμητής μέγιστης πιθανοφάνειας $\overline{\boldsymbol{X}}$ είναι ένα τυχαίο διάνυσμα και ο εκτιμητής $\overline{\boldsymbol{\Sigma}}$ είναι ένας τυχαίος πίνακας.

Επιπλέον, έχουμε ότι η μέγιστη πιθανότητα είναι:

$$L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \frac{1}{(2\pi)^{np/2}} e^{-np/2} \frac{1}{|\widehat{\boldsymbol{\Sigma}}|^{n/2}}$$

ή από το $|\widehat{\boldsymbol{\Sigma}}| = \left[\frac{n-1}{n}\right]^p$

<u>Γενικά</u>

$L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \sigma\tau\alpha\theta\epsilon\rho\acute{\alpha}\ X\ (\gamma\epsilon\nu\iota\kappa\epsilon\upsilon\mu\acute{\epsilon}\nu\eta\ \delta\iota\alpha\kappa\acute{\upsilon}\mu\alpha\nu\sigma\eta)^{-n/2}$

Η γενικευμένη διακύμανση είναι ένα φυσικό μέτρο της μεταβλητότητας όταν είναι πολυμεταβλητή κανονική.

Από την άλλη οι εκτιμητές μέγιστης πιθανοφάνειας παραμένουν αναλλοίωτοι. Ας πάρουμε ότι ο $\widehat{\boldsymbol{\theta}}$ είναι ο εκτιμητής του $\boldsymbol{\theta}$ και θεωρούμε την εκτίμηση της παραμέτρου h($\boldsymbol{\theta}$), η οποία είναι μια λειτουργία του **θ**.

Τότε η μέγιστη εκτίμηση της πιθανότητας

h($\boldsymbol{\theta}$) δίνεται από $h(\widehat{\boldsymbol{\theta}})$

(λειτουργία του **θ**) (λειτουργία του $\widehat{\boldsymbol{\theta}}$)

Θα δούμε δυο μικρά παραδείγματα:

(1) Ο εκτιμητής Μέγιστης Πιθανοφάνειας του $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ είναι ο $\widehat{\boldsymbol{\mu}}'\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}$, όπου $\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{X}}$ και $\widehat{\boldsymbol{\Sigma}} = \left((n-1)/n\right)\mathbf{S}$, τα οποία είναι οι εκτιμητές των **μ** και **Σ** αντίστοιχα.

(2) Ο εκτιμητής μέγιστης πιθανοφάνειας του $\sqrt{\sigma_{ii}}$ είναι ο $\sqrt{\hat{\sigma}_{ii}}$ όπου $\hat{\sigma}_{ii} = \frac{1}{n}\sum_{j=1}^{n}\left(X_{ij} - \overline{X_i}\right)^2$, ο οποίος είναι ο εκτιμητής του $\sigma_{ii} = Var\left(X_i\right)$ που είναι ίσο με τη διασπορά.

Επαρκή Στατιστικά Στοιχεία

Από τον τύπο της από κοινού πυκνότητας των $x_1,...,x_n$ όπου:

$$\{\text{Από κοινού πυκνότητα } x_{1,}...x_n\} =$$

$$(2\pi^{-np/2}|\Sigma|^{-\nu/2}\exp\left\{-tr\left[\Sigma^{-1}\left(\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'+n(\bar{x}-\mu)(\bar{x}-\mu)'\right)\right]/2\right\}$$

ο οποίος βασίζεται σε όλο το σύνολο των παρατηρήσεων $\mathrm{X}_1,\mathrm{X}_2,...,\mathrm{X}_n$, μόνο για τη μέση τιμή δείγματος $\bar{x}$ και τον λόγο των διαγωνίων πινάκων $\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'=(n-1)S$.

Εκφράζουμε αυτό το γεγονός λέγοντας ότι τα $\bar{x}$ και το $(n-1)S$ $(\acute{\eta}\ S)$ είναι Επαρκή Στατιστικά Στοιχεία.

Γενικά

Ας πάρουμε ότι $X_1,X_2,...,X_n$ είναι ένα τυχαίο δείγμα από έναν πολυμεταβλητό κανονικό πληθυσμό με μέση τιμή μ και συνδυασπορά Σ. Τότε λέμε ότι τα $\bar{\boldsymbol{X}}$ και S είναι Επαρκή Στατιστικά Στοιχεία.

Η σημασία των Επαρκούς Στατιστικών Στοιχείων για κανονικούς πληθυσμούς είναι ότι τα στοιχεία του δείγματος για τα δεδομένα του πίνακα Χ περιέχονται στα $\bar{x}$ και S, ανεξάρτητα από το πιο θα είναι το μέγεθος του δείγματος n.

Αυτό γενικά δεν ισχύει για μη κανονικούς πληθυσμούς.

Δεδομένου ότι πολλές πολυμεταβλητές τεχνικές αρχίζουν με μέσο δείγμα και συνδυακύμανση, είναι καλύτερα να ελέγξουμε για το πόσο επαρκεί η υπόθεση της πολυμεταβλητής κανονικής.

Εάν καταλήξουμε ότι τα δεδομένα δεν μπορούν να θεωρηθούν ως πολυμεταβλητές κανονικές, τότε οι τεχνικές που εξαρτώνται αποκλειστικά από τα $\bar{x}$ και S μπορούν να αγνοήσουν άλλες χρήσιμες πληροφορίες του δείγματος.

Δεσμευμένες και Περιθωριακές Κατανομές Πολυμεταβλητών Μεταβλητών

Εδώ θα παρουσιάσουμε ένα αριθμό ουσιωδών αποτελεσμάτων παραγόμενων από πολυκανονική κατανομή. Αυτό που θα κάνουμε θα είναι να περιορίσουμε την προσοχή μας στη μη ιδιάζουσα πολυμεταβλητή κατανομή ή σε ένα πίνακα συνδιακύμανση πλήρους βαθμού. Ξεκινάμε με αυτό το βασικό αποτέλεσμα.

Ιδιότητα 1

Ας πάρουμε το $p-διάστατο$ τυχαίο διάνυσμα X κατανεμημένο σύμφωνα με την πολυκανονική κατανομή με μέση τιμή διανύσματος μ και πίνακα συνδιακύμανσης Σ τάξης p. Εάν A είναι κάθε mxp πίνακας πραγματικών αριθμών με βαθμό $m \le p$, η νέα m συνιστώσα τυχαίου διανύσματος είναι:

$$Y = AX$$

Είναι μια πολυκανονική τυχαία μεταβλητή με μέση τιμή $E(Y) = A\mu$ και πίνακα συνδιακύμανσης $A\Sigma A'$.

Αυτή η ιδιότητα έχει κάποιες συνέπειες:

1. Εάν $m = 1,$ η βαθμωτή μεταβλητή $a'X$ ακολουθεί μονομεταβλητή κανονική κατανομή με μέση τιμή $a'\mu$ και διακύμανση $a'\Sigma\alpha$
2. Η κοινή κατανομή για κάθε σύνολο στοιχείων X είναι πολυκανονική με μέση τιμή διανύσματος και πίνακα συνδιακύμανσης που δίνονται από τα κατάλληλα στοιχεία μ και Σ. Η περιθώρια κατανομή για κάθε στοιχείο είναι μονομεταβλητή κανονική. Όμως, κανονικές περιθώριες δεν αναφέρουν απαραίτητα μια κοινή πολυμεταβλητή κατανομή: παραδείγματα για τέτοιες κατανομές έχουν συζητηθεί από τον Kawaski (1973).

Η ιδιότητα 1 αναφέρει ότι οι μεταβλητές του κύριου άξονα ορίζονται από τη μεταφορά της προηγούμενης ενότητας, ακολουθεί την πολυμεταβλητή κανονική κατανομή και χάρη στις μηδενικές συσχετίσεις κατανέμεται ανεξάρτητα.

Η δεσμευμένη πυκνότητα της πολυμεταβλητής κανονικής κατανομής

Θα ερμηνεύσουμε τις ποσότητες των παραμέτρων της δεσμευμένης πολυκανονικής πυκνότητας.

Για να αποκτήσουμε την πυκνότητα ξεκινάμε από τον μη ιδιάζοντα $(p+q)-διάστατο$ πολυκανονικό πληθυσμό, του οποίου οι μεταβλητές γράφονται ως χωρισμένο το διάνυσμα $X'=[X'_1,\ \ X'_2]$, όπου το $X_1^{'}$ περιέχει p στοιχεία και το $X_2^{'}$ τα υπόλοιπα q. Η μέση τιμή διανύσματος και ο πίνακας συνδιακύμανσης είναι σχετικά διχοτομημένοι ως

(1) $\mu=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\quad \Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^{'} & \Sigma_{22}\end{bmatrix}$

Οι υποπίνακες Σ_{11},Σ_{12} και Σ_{22} έχουν διαστάσεις pxp, pxq και qxq. Κατά συνέπεια τα X_1 και X_2 είναι περιθώρια πολυκανονικά τυχαία διανύσματα με κατανομές $N(\mu_1,\Sigma_{11})$ και $N(\mu_2,\Sigma_{22})$. Οι τιμές των στοιχείων του Σ_{12}, σχετικά με εκείνες των Σ_{11} και Σ_{22} καθορίζουν το βαθμό και τον τρόπο της εξαρτησίας μεταξύ των δυο συνόλων των μεταβλητών.

Έτσι, λοιπόν, έχουμε:

Ιδιότητα 2

Τα τυχαία διανύσματα X_1 και X_2 με την πολυμεταβλητή κανονική κατανομή περιγράφεται από τις παραμέτρους μ και Σ που κατανέμονται ανεξάρτητα μόνο εάν $\boxed{\Sigma_{12}=0}$

Η λειτουργία της δεσμευμένης πυκνότητας της X_1 ας την πούμε x_1 δοθέντος ότι τα στοιχεία του x_2 δόθηκαν για γενικές πολυδιάστατες κατανομές ως:

$$g(x_1/x_2)=\frac{f(x_1,x_2)}{f(x_2)}$$

Στην παρούσα περίπτωση

(2) $$h(x_2)=\frac{1}{(2\pi)^{q/2}|\Sigma_{22}|^{1/2}}\exp\left[-\frac{1}{2}(x_2-\mu_2)'\Sigma_{22}^{-1}(x_2-\mu_2)\right]$$

Για τον υπολογισμό του $g(x_1/x_2)$, χρειαζόμαστε την κοινή πυκνότητα $f(x_1,x_2)$ εφαρμοσμένη σε όρους των υποπινάκων του Σ. Ο αντίστροφος του πίνακα προκύπτει από:

(3) $$\Sigma^{-1}=\begin{bmatrix}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1} & \left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ -\Sigma_{22}^{-1}\Sigma'_{12}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1} & \Sigma_{22}^{-1}\Sigma_{22}^{-1}\Sigma'_{12}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}\end{bmatrix}$$

και

(4) $$|\Sigma|=|\Sigma_{22}|\cdot|\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}|$$

Η κοινή πυκνότητα μπορεί να γραφεί ως:

(5) $$f(x_1,x_2)=\frac{1}{(2\pi)^{(p+q)/2}|\Sigma_{22}|^{1/2}\cdot|\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}|^{1/2}}$$

$$\cdot\exp\{-\frac{1}{2}[(x_1-\mu_1)'\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1}(x_1-\mu_1)$$

$$-(x_2-\mu_2)'\Sigma_{22}^{-1}\Sigma'_{12}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1}(x_1-\mu_1)$$

$$-(x_1-\mu_1)'\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)$$

$$+(x_2-\mu_2)'\Sigma_{22}^{-1}\Sigma'_{12}\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right)^{-1}\Sigma_{12}^{-1}\Sigma'_{22}(x_2-\mu_2)$$

$$(x_2-\mu_2)'_{1.2}\Sigma_{22}^{-1}(x_2-\mu_2)]\}$$

Με το διαχωρισμό από το $h(x_2)$, ο όρος $(2\pi)^{q/2}|\Sigma_{22}|^{1/2}$ στην ομαλοποιημένη σταθερά και η τελική τετραγωνική μορφή του εκθέτη εξαφανίζεται. Η δεσμευμένη συνάρτηση γίνεται:

(6) $$g\left(x_1 / x_2\right)=\frac{1}{\left(2\pi\right)^{p/2}\left|\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}\right|^{1/2}}$$

$$\cdot\exp\{-\frac{1}{2}[x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}\left(x_2-\mu_2\right)]'$$

$$\cdot\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{22}\right)^{-1}\left[x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}\left(x_2-\mu_2\right)\right]\}$$

Αυτή είναι η πυκνότητα μια πολυκανονικής κατανομής με μέση τιμή διανύσματος

(7) $$\mu_1+\Sigma_{12}\Sigma_{22}^{-1}\left(x_2-\mu_2\right)$$

και τον πίνακα συνδιακύμανσης

(8) $$\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}$$

Από τα στοιχεία των X_1 και X_2 που θα μπορούσαν να επιλεγούν σε συνδυασμό με ένα αρχικό τυχαίο διάνυσμα, θα πρέπει να συνοψίσουμε το προηγούμενο αποτέλεσμα ως:

<u>Ιδιότητα 3</u>

Η δεσμευμένη κατανομή για κάθε σύνολο των p μεταβλητών από ένα πολυκανονικό πληθυσμό w με q άλλες μεταβλητές του πληθυσμού να παραμένουν σταθερές είναι η πολυκανονική με μέση τιμή διανύσματος και πίνακα συνδιακύμανσης να δίνονται από τις εκφράσεις (7) και (8).

Το τυχαίο διάνυσμα

(9) $$X_{12}=X_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}\left(X_2-\mu_2\right)$$

καλείται το σύνολο των υπολειπόμενων μεταβλητών. Αντιπροσωπεύει τις διαφορές των στοιχείων του X και τις τιμές τους, όπως έχει προβλεφθεί από την γραμμική σχέση της μέσης τιμής της δεσμευμένης κατανομής με τις X_2 μεταβλητές. Οι συνδιακυμάνσεις των στοιχείων του X_1 με εκείνες του X_{12} είναι:

(10) $$E\left[\left(X_1-\mu_1\right)X'_{12}\right]=E\left(X_1X'_{12}\right)$$
$$=E\left\{X_1\left[X'_1-\mu_1-\left(X_2-\mu_2\right)'\Sigma_{22}^{-1}\Sigma'_{12}\right]\right\}$$
$$=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma'_{12}$$

ή μερικώς ο πίνακας συνδιακύμανσης της δεσμευμένης κατανομής.

(11) $$E\left[\left(\mathrm{X}_2-\mu_2\right)X'_{12}\right]=\mathrm{E}\left\{\left(\mathrm{X}_2-\mu_2\right)\left[X'_1-\mu'_1-\left(\mathrm{X}_2-\mu_2\right)'\Sigma_{22}^{-1}\Sigma'_{12}\right]\right\}$$
$$=\Sigma'_{12}-\Sigma_{22}\Sigma_{22}^{-1}\Sigma'_{12}$$
$$=0$$

Όμως, οι διακυμάνσεις των υπολειπόμενων και οι σταθερές μεταβλητές κατανέμονται ανεξάρτητα.

Έλεγχος υποθέσεων

Πρέπει να μάθουμε να χρησιμοποιούμε τις έννοιες και τα αποτελέσματα, ώστε να αναπτύξουμε τεχνικές για την ανάλυση δεδομένων.
Ο στόχος μας είναι να παρουσιάσουμε μια πλήρη στατιστική ανάλυση του στοιχείου που σημαίνει ότι βασίζεται σε ταυτόχρονες δηλώσεις εμπιστοσύνης.
Ένα από τα κεντρικά μηνύματα της πολυμεταβλητής ανάλυσης είναι ότι οι ρ-μεταβλητές πρέπει να αναλυθούν από κοινού.
Έτσι λοιπόν το μεγάλο πλεονέκτημα της πολυμεταβλητής ανάλυσης είναι ότι μας επιτρέπει να εξετάσουμε πολλές μεταβλητές συγχρόνως λαμβάνοντας υπόψη και τις όποιες συσχετίσεις υπάρχουν μεταξύ τους.

Σε αυτό το κεφάλαιο θα περιγράψουμε ελέγχους που αφορούν παρατηρήσεις από πολυμεταβλητούς κανονικούς πληθυσμούς. Συγκεκριμένα θα μελετήσουμε προβλήματα ελέγχου υποθέσεων για τις παραμέτρους μ και Σ
Για παράδειγμα έχουμε 5 μεταβλητές, οι οποίες μας περιγράφουν κάποιο χαρακτηριστικό. Αν θέλουμε να εξετάσουμε αν οι παρατηρήσεις μας δεν συμφωνούν σε κάποια εκ των προτέρων γνώση για τα 5 αυτά χαρακτηριστικά, θα μπορούσε κανείς να δουλέψει ξεχωριστά με τις πέντε μεταβλητές, να κάνει δηλαδή έναν έλεγχο για κάθε μεταβλητή και με βάση αυτούς τους ελέγχους να αποφανθεί για το σύνολο.

Αυτή η προσέγγιση έχει <u>*δυο μειονεκτήματα*</u>:

- Αν ο έλεγχος για κάθε χαρακτηριστικό γίνει σε επίπεδο 5%, τότε, δεδομένου πως κάνουμε 5 ελέγχους ξεχωριστά, το συνολικό επίπεδο του πολύπλοκου (πενταπλού στην περίπτωσή μας) ελέγχου, δεν θα είναι 5% αλλά $1\text{-}0.95^5$, δηλαδή, μόλις 22%, δηλαδή, η πιθανότητα σφάλματος εδώ είναι πολύ μεγάλη.

- Επειδή είναι πολύ πιθανό οι μεταβλητές μας να είναι μεταξύ τους σχετισμένες κάνοντας πέντε ξεχωριστούς ελέγχους, αγνοούμε αυτή τη συσχέτιση, δηλαδή, αγνοούμε την πληροφορία που έχουμε για κάθε μεταβλητή από τις υπόλοιπες. Επομένως, χρειαζόμαστε πολυμεταβλητούς

ελέγχους υποθέσεων, οι οποίοι θα εξετάσουν όλο το διάνυσμα κάθε παρατήρησης και όχι μεμονωμένες μεταβλητές.

Οι έλεγχοι που θα περιγράψουμε παρακάτω αφορούν παρατηρήσεις από πολυμεταβλητούς κανονικούς πληθυσμούς.

Η αξιοπιστία του μ_0 ως αξία για μια Κανονική μέση τιμή του πληθυσμού

Ας ξεκινήσουμε με την πιο απλή περίπτωση του ελέγχου για ένα διάνυσμα μέσων με γνωστή διακύμανση. Θυμόμαστε ότι για τον έλεγχο υποθέτουμε, πως οι μεταβλητές μας είναι κανονικές, αλλά αν δεν ισχύει η υπόθεση της κανονικότητας με επίκληση του κεντρικού οριακού θεωρήματος, το αποτέλεσμα ισχύει για μεγάλα δείγματα.

Στην πολυμεταβλητή περίπτωση υποθέτουμε πως τα δεδομένα προέρχονται από την πολυμεταβλητή κανονική κατανομή.

Έχουμε την εξής υπόθεση:

$$H_0 : \mu = \mu_0 \text{ έναντι της εναλλακτικής } H_1 : \mu \neq \mu_0$$

Εδώ, το H_0 είναι η μηδενική υπόθεση και η H_1 είναι εναλλακτική υπόθεση. Η εναλλακτική H_1 είναι δίπλευρη καθώς σε ένα διάνυσμα δεν είναι εύκολο να ορίσουμε μονόπλευρους ελέγχους. Παρόλα αυτά, μονόπλευροι έλεγχοι θα μπορούσαν να οριστούν χρησιμοποιώντας ελέγχους πιθανοφάνειας.

Εάν $X_1, X_2, ..., X_n$ υποδηλώνουν ένα τυχαίο δείγμα από έναν κανονικό πληθυσμό, το κατάλληλο στατιστικό αποτέλεσμα είναι το εξής:

$$t = \frac{\left(\overline{\mathrm{X}} - \mu_0\right)}{s/\sqrt{n}}, \quad \text{όπου} \quad \overline{\mathrm{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j \quad \text{και} \quad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{\mathrm{X}}\right)^2$$

Αυτή η στατιστική δοκιμή έχει $t-$κατανομές με $n-1$ βαθμούς ελευθερίας.

Απορρίπτουμε το H_0 όταν το μ_0 είναι εύλογη τιμή του μ, εάν η παρατηρούμενη τιμή $|t|$ υπερβαίνει την $t-$κατανομή με $n-1$ βαθμούς ελευθερίας.

Απορρίπτοντας την H_0, όταν το $|t|$ είναι μεγάλο, ισοδυναμεί με την απόρριψη του H_0 όταν,

$$t^2 = \frac{\left(\overline{X} - \mu_0\right)^2}{s^2/n} = n\left(\overline{X} - \mu_0\right)\left(s^2\right)^{-1}\left(\overline{X} - \mu_0\right) \quad \text{είναι μεγάλο.}$$

Η μεταβλητή t^2 είναι το τετράγωνο της απόστασης από τη μέση τιμή δείγματος $\left|\overline{\mathrm{X}}\right|$ στην τιμή δοκιμής μ_0. Οι μονάδες της απόστασης εκφράζονται με όρους του $s/\sqrt{n}$ ή εκτιμώντας την τυπική απόκλιση $\left|\bar{x}\right|$.

Όταν παρατηρούνται τα $\left|\bar{x}\right|$ και s^2 προκύπτει ότι:

Απορρίπτοντας το H_0 υπέρ του H_1 σε επίπεδο σημαντικότητας a έχουμε ότι:

$$n\left(\bar{x}-\mu_0\right)(s^2)^{-1}\left(\bar{x}-\mu_0\right) > t^2_{n-1}(a/2)$$

όπου το $t_{n-1}(a/2)$ δηλώνει το ανώτερο εκατοστημόριο της $t-$κατανομής με $n-1$ βαθμούς ελευθερίας.

Εάν το H_0 δεν απορρίπτεται, καταλήγουμε στο συμπέρασμα, ότι το μ_0 είναι μια εύλογη τιμή για την κανονική μέση τιμή του πληθυσμού.

Ερώτηση

Υπάρχουν άλλες τιμές του μ που να είναι συμβατές με τα δεδομένα; Η απάντηση είναι ναι. Στην πραγματικότητα υπάρχει πάντα ένα σύνολο εύλογων τιμών για μια κανονική μέση τιμή του πληθυσμού. Παίρνοντας τον έλεγχο και λαμβάνοντας υπόψιν την αποδεκτή περιοχή

$H_0 : \mu = \mu_0$ σε σχέση με $H_1 : \mu \neq \mu_0$

και διαστήματα εμπιστοσύνης για το μ έχουμε:

Δεν απορρίπτεται το $H_0 : \mu = \mu_0$ στο επίπεδο α δηλαδή όταν

$$\left|\frac{\bar{x}-\mu_0}{s\sqrt{n}}\right| \leq t_{n-1}(\alpha/2)$$

Είναι ισοδύναμο με το:

το μ_0 βρίσκεται σε $100(1-\alpha)\%$ διαστήματα εμπιστοσύνης $\bar{x} \pm t_{n-1}(\alpha/2)\dfrac{s}{\sqrt{n}}$

ή $\quad \bar{x}-t_{n-1}(\alpha/2)\dfrac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x}+t_{n-1}(\alpha/2)\dfrac{s}{\sqrt{n}}$

Τα διαστήματα εμπιστοσύνης αποτελούνται από όλες εκείνες τις τιμές του μ_0 για τις οποίες απορρίπτουμε το $H_0 : \mu = \mu_0$.

Πριν το δείγμα επιλέγει το $100(1-a)\%$ διάστημα εμπιστοσύνης του

$$\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} \le \mu_0 \le \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}$$

είναι ένα τυχαίο διάστημα, γιατί τα τελικά σημεία εξαρτώνται από τις τυχαίες μεταβλητές $\bar{x}$ και s .

Η κατανομή του Hotelling

Μια φυσική γενίκευση των τετραγώνων των αποστάσεων στην πολυμεταβλητή είναι:

$$T^2 = (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \left(\frac{\boldsymbol{S}}{n}\right)^{-1} . (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

Όπου έχουμε ότι:

$$\bar{\boldsymbol{X}} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{X}_j \text{ για έναν } px1 \text{ πίνακα}$$

$$\mathbf{S} = \frac{1}{n-1} \sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})' \text{ για έναν } pxp \text{ πίνακα}$$

$$\text{Και } \boldsymbol{\mu}_0 = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix} \text{ ένας } px1 \text{ πίνακας}$$

Η κατανομή T^2 καλείται Hotelling προς τιμήν του Harold Hotelling, ο οποίος ήταν πρωτοπόρος στην πολυπαραγοντική ανάλυση. Εδώ το $\left(\frac{1}{n}\right)\mathbf{S}$ είναι ο εκτιμώμενος πίνακας συνδιακύμανσης του $|\bar{\boldsymbol{X}}|$.

Αν η παρατηρούμενη γενικευμένη απόσταση T^2 είναι πολύ μεγάλη που είναι το $|\bar{x}|$, είναι πολύ μακριά από το $\boldsymbol{\mu}_0$, τότε παρατηρούμε ότι η υπόθεση $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ απορρίπτεται.

Αποδεικνύεται ότι δεν είναι οι ειδικοί πίνακες του T^2 που απαιτούνται για τις δοκιμές των υποθέσεων.

Αυτό συμβαίνει γιατί το:

T^2 κατανέμεται ως $\frac{(n-1)p}{n-p}F_{p,n-p}$

Η κατανομή T^2 συνδέεται με την κατανομή F και επομένως τα ποσοστιαία σημεία της μπορούν να υπολογιστούν πολύ εύκολα.

Εδώ λοιπόν το $F_{p,n-p}$ δηλώνει μια τυχαία μεταβλητή με μια $F-$κατανομή με p και $n-p$ βαθμούς ελευθερίας.

Συνοψίζοντας έχουμε τα εξής:

Ας πάρουμε $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ να είναι ένα τυχαίο δείγμα από ένα κανονικό πληθυσμό $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$

Τότε, με $\bar{\boldsymbol{X}} = = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j$ και $\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'$

έχουμε $\alpha = P\left[T^2 > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right]$

$=P\left[n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right]$

ανεξάρτητα από τα $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$.

Αυτό λοιπόν μας οδηγεί αμέσως στο να εξετάσουμε την υπόθεση: $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ έναντι $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$.

Στο επίπεδο σημαντικότητας α, απορρίπτουμε το H_0 υπέρ του H_1 εάν

$$T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$

Θα είναι χρήσιμο να εξετάσουμε τη φύση της T^2-κατανομής.

Θα πάρουμε, λοιπόν, με τη βοήθεια της κατανομής Wishart τη γενίκευση της t^2, όπου

$$T^2 = \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)' \left(\frac{\sum_{j=1}^{n} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'}{n-1} \right)^{-1} \sqrt{n}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

η οποία είναι της μορφής

$$\begin{pmatrix}\text{πολυμεταβλητή κανονική}\\ \text{τυχαίου διανύσματος}\end{pmatrix}'\left(\frac{\text{wishart τυχαίος πίνακας}}{\text{βαθμού ελευθερίας}}\right).\begin{pmatrix}\text{πολυμεταβλητή κανονική}\\ \text{τυχαίου διανύσματος}\end{pmatrix}$$

Αυτό είναι ανάλογο του

$$T^2 = \sqrt{n}(\bar{\boldsymbol{X}} - \mu_0)\left(s^2\right)^{-1}\sqrt{n}(\bar{\boldsymbol{X}} - \mu_0)$$

<u>ή</u>

$$\begin{pmatrix}\text{κανονική τυχαία}\\ \text{μεταβλητή}\end{pmatrix}\left(\frac{\text{τυχαία μεταβλητή } s^2}{\text{βαθμούς ελευθερίας}}\right)^{-1}\begin{pmatrix}\text{κανονική τυχαία}\\ \text{μεταβλητή}\end{pmatrix}$$

για την μονομεταβλητή περίπτωση. Από τότε που η πολυμεταβλητή κανονική και οι Wishart τυχαίες μεταβλητές κατανέμονται ανεξάρτητα, η κοινή τους πυκνότητα είναι προϊόν της οριακής κανονικής και των κατανομών Wishart.

Χρησιμοποιώντας, λοιπόν, την κατανομή του T^2, όπως αναφέραμε και παραπάνω, μπορούν να προκύψουν από αυτήν την κοινή κατανομή.

Κάποιες ενδιαφέρουσες παρατηρήσεις είναι οι εξής:

- Η συνάρτηση T^2 έχει την πολύ χρήσιμη ιδιότητα πως είναι αμετάβλητη από γραμμικούς μετασχηματισμούς, δηλαδή, αν μετασχηματίσουμε τα δεδομένα, η τιμή της δεν αλλάζει.
- Στην ουσία η συνάρτηση T^2 είναι μια γενίκευση του τρόπου που προκύπτει η κατανομή t στη μονομεταβλητή περίπτωση. Αν μάλιστα θυμηθούμε πως το τετράγωνο μιας t_n τυχαίας μεταβλητής ακολουθεί $F_{n,1}$ κατανομή, η κατανομή T^2 είναι μια γενίκευση της κατανομής t.
- Αν έχω ότι $p=1$, τότε, η συνάρτηση T^2 δεν είναι τίποτα άλλο παρά η συνάρτηση $t=\dfrac{\left(\bar{x}-\mu\right)}{s/\sqrt{n}}$, που τόσο πολύ χρησιμοποιούμε στη μονομεταβλητή περίπτωση.
- Τέλος, τα διαστήματα εμπιστοσύνης θα είναι υπερελλειψοειδή.

Παράδειγμα

Ας πάρουμε έναν πίνακα δεδομένων ενός τυχαίου δείγματος μεγέθους $n=3$ από τον δισδιάστατο κανονικό πληθυσμό να είναι:

$$X=\begin{bmatrix}6 & 10 & 8\\ 9 & 6 & 3\end{bmatrix}$$

Αξιολογώ την παρατηρούμενη T^2 για $\mu_0{}'=[9,5]$.

Ποια είναι η κατανομή δειγματοληψίας της T^2 σε αυτή την περίπτωση;

Βρίσκουμε τη μέση τιμή:

$$\bar{x}=\begin{bmatrix}\bar{x}_1\\ \bar{x}_2\end{bmatrix}=\begin{bmatrix}\frac{6+10+8}{3}\\ \frac{9+6+3}{3}\end{bmatrix}=\begin{bmatrix}8\\ 6\end{bmatrix}$$

και $$s_{11}\frac{(6-8)^2+(10-8)^2+(8-8)^2}{2}=4$$

$$s_{12}\frac{(6-8)(9-6)+(10-8)(6-6)+(8-8)(3-6)}{2}=-3$$

και $$s_{22}\frac{(9-6)^2+(6-6)^2+(3-6)^2}{2}=9$$

Έτσι, προκύπτει ότι:

$$S=\begin{bmatrix}4 & -3\\ -3 & 9\end{bmatrix}$$

Άρα $S^{-1}=\frac{1}{4\cdot 9-(-3)(-3)}\begin{bmatrix}9 & 3\\ 3 & 4\end{bmatrix}=\begin{bmatrix}1/3 & 1/9\\ 1/9 & 4/27\end{bmatrix}$

και από τον τύπο: $T^2=(\bar{\boldsymbol{X}}-\mu_0)'\left(\frac{\boldsymbol{S}}{n}\right)^{-1}(\bar{\boldsymbol{X}}-\mu_0)=n(\bar{\boldsymbol{X}}-\mu_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)$

προκύπτει: $T^2 = 3\begin{bmatrix}8-9, & 6-5\end{bmatrix}\begin{bmatrix}1/3 & 1/9\\ 1/9 & 4/27\end{bmatrix}\begin{bmatrix}8-9\\ 6-5\end{bmatrix} = 3\begin{bmatrix}-1, & 1\end{bmatrix}\begin{bmatrix}-2/9\\ 1/27\end{bmatrix} = \frac{7}{9}$

Πριν το δείγμα επιλεγεί, η T^2 έχει την κατανομή

$$\frac{(3-1)2}{3-2}F_{2,3-2} = 4F_{2,1}$$

Κατανομή T^2 Hotteling και λόγος Πιθανοφανειών

Παρουσιάζουμε την στατιστική T^2 κατ' αναλογία με την μονομεταβλητή απόσταση t^2.

Υπάρχει μια γενική αρχή για την κατασκευή των δοκιμών που ονομάζεται μέθοδος του λόγου πιθανοφάνειας και το T^2 μπορεί να προκύψει, όπως το τεστ του λόγου πιθανοφάνειας του $H_0 : \mu = \mu_0$

Τα τεστ του λόγου πιθανοφάνειας έχουν διάφορες ιδιότητες που είναι ιδανικές για μεγάλα δείγματα και είναι ιδιαίτερα καλό για τις υποθέσεις που διατυπώνονται ως προς τις μεταβλητές των πολυμεταβλητών κανονικών κατανομών. Η μέγιστη πιθανοφάνεια της πολυμεταβλητής κανονικής κατανομής ως προς τα μ και Σ είναι διαφορετική σε σχέση με πιθανές τιμές και δίνεται από τον τύπο:

$$\max_{\mu,\Sigma} L(\mu,\Sigma) = \frac{1}{(2\pi)^{np/2}\left|\hat{\Sigma}\right|^{n/2}} e^{-np/2}$$

Όπου $\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)\left(x_j - \bar{x}\right)'$ και $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$

οι οποίες είναι οι μέγιστες εκτιμήσεις πιθανοφάνειας.

Θα πρέπει να θυμίσουμε ότι οι μέγιστες εκτιμήσεις πιθανοφάνειας $\hat{\mu}$ και $\hat{\Sigma}$ είναι αυτές οι επιλογές για τα μ και Σ, τα οποία εξηγούν καλύτερα τις παρατηρούμενες τιμές του τυχαίου δείγματος.

Στα πλαίσια της υπόθεσης $H_0 : \mu = \mu_0$ η κανονική πιθανοφάνεια ειδικεύεται στο:

$$L(\mu_0,\Sigma) = \frac{1}{(2\pi)^{np/2}\left|\Sigma\right|^{n/2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(x_j - \mu_0\right)'\Sigma^{-1}\left(x_j - \mu_0\right)\right)$$

Η μέση τιμή μ_0 είναι πλέον σταθερή, αλλά, η συνδιακύμανση Σ μπορεί να μεταβάλλεται για να βρει την τιμή, που είναι πιο πιθανό να έχουν οδηγηθεί με μ_0 σταθερό στο παρατηθηθέν δείγμα.

Η τιμή αυτή λαμβάνεται με τη μεγιστοποίηση του $L(\mu_0, \Sigma)$ σε σχέση με το Σ.

Ο εκθέτης, λοιπόν, του $L(\mu_0, \Sigma)$ πρέπει να γραφτεί ως εξής:

$$-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu_0)'\Sigma^{-1}(x_j-\mu_0) = -\frac{1}{2}\sum_{j=1}^{n} tr\left[\Sigma^{-1}(x_j-\mu_0)(x_j-\mu_0)'\right]$$

$$= -\frac{1}{2} tr\left[\Sigma^{-1}\sum_{j=1}^{n}(x_j-\mu_0)(x_j-\mu_0)'\right]$$

Εφαρμόζοντας τώρα το:

$$\frac{1}{|\Sigma|^b} e^{-tr\left(\Sigma^{-1}B\right)/2} \le \frac{1}{|B|^b}(2b)^{pb} e^{-pb}$$

για πίνακα B έχουμε τα εξής:

για $B = \sum_{j=1}^{n}(x_j-\mu_0)(x_j-\mu_0)'$ και $b = n/2$ Προκύπτει ότι:

$$\max_{\Sigma} L(\mu_0, \Sigma) = \frac{1}{(2\pi)^{np/2}\left|\hat{\Sigma}_0\right|^{n/2}} e^{-np/2}$$

με $\hat{\Sigma}_0 = \frac{1}{n}\sum_{j=1}^{n}(x_j-\mu_0)(x_j-\mu_0)'$

Για να προσδιορίσουμε εάν το μ_0 είναι η κατάλληλη τιμή για το μ, το μέγιστο όριο του $L(\mu_0, \Sigma)$ είναι σε σύγκριση με το $L(\mu, \Sigma)$. Ο λόγος ονομάζεται στατιστική του λόγου της πιθανοφάνειας.

Έτσι έχουμε:

$$\text{Λόγος Πιθανοφάνειας} = \Lambda = \frac{\max_{\Sigma} L(\mu_0, \Sigma)}{\max_{\mu,\Sigma} L(\mu, \Sigma)} = \left(\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_0\right|}\right)^{n/2}$$

Το αντίστοιχο στατιστικό στοιχείο $\Lambda^{2/n} = \dfrac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_0\right|}$ ονομάζεται

κατανομή Λάμδα του Wilks.

Η κατανομή Λάμδα είναι μονοδιάστατη. Η συνάρτηση πυκνότητας πιθανότητας της είναι αρκετά πολύπλοκη.

Αν η παρατηρούμενη τιμή αυτή του λόγου πιθανότητας είναι πολύ μικρή, έτσι η υπόθεση $H_0 : \mu = \mu_0$ είναι πιθανό να είναι αληθινή και ως εκ τούτου απορρίφθηκαν. Συγκεκριμένα, ο έλεγχος του λόγου πιθανοφάνειας $H_0 : \mu = \mu_0$ έναντι $H_1 : \mu \neq \mu_0$, απορρίπτει το H_0, στην περίπτωση που:

$$\Lambda = \left(\frac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_0\right|} \right)^{n/2} = \left(\frac{\left| \sum_{j=1}^{n} \left(x_j - \bar{x}\right)\left(x_j - \bar{x}\right)' \right|}{\left| \sum_{j=1}^{n} \left(x_j - \mu_0\right)\left(x_j - \mu_0\right)' \right|} \right)^{n/2} < c_a$$

όπου το C_a είναι το κατώτερο όριο της κατανομής Λ. Πρέπει να σημειώσουμε ότι το στατιστικό αποτέλεσμα του λόγου πιθανοφάνειας είναι μια δύναμη του λόγου της γενικευμένης διακύμανσης.

Ευτυχώς, επειδή από την παρακάτω σχέση μεταξύ των T^2 και Λ, δεν χρειαζόμαστε την κατανομή του Λ για τη διεξαγωγή των δοκιμών.

Αποτέλεσμα

Ας πάρουμε ότι $X_1, X_2, ..., X_n$ να είναι ένα τυχαίο δείγμα από μια κανονική κατανομή $N_p\left(\mu, \Sigma\right)$. Τότε, από τον τύπο $T^2 = n\left(\bar{x} - \mu_0\right)' S^{-1}\left(\bar{x} - \mu_0\right)$ που βασίζεται στο T^2 είναι ισοδύναμο με το τεστ του λόγου πιθανοφάνειας του:

$H_0 : \mu = \mu_0$ έναντι $H_1 : \mu \neq \mu_0$ επειδή

$$\Lambda^{2/n} = \left(1 + \frac{T^2}{(n-1)}\right)^{-1}$$

Απόδειξη

Ας πάρουμε τον $(p+1)x(p+1)$ πίνακα A που καθορίζεται από:

$$A=\left[\begin{array}{c|c} \sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)' & \sqrt{n}\left(\bar{x}-\mu_0\right) \\ \hline \sqrt{n}\left(\bar{x}-\mu_0\right)' & -1 \end{array}\right]=\left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]$$

Ισχύει ότι $\left|A\right|=\left|A_{22}\right|\left|A_{11}-A_{12}A_{22}^{-1}A_{21}\right|=\left|A_{11}\right|\left|A_{22}-A_{21}A_{11}^{-1}A_{12}\right|$

από το οποίο παίρνουμε

$$(-1)\left|\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'+n\left(\bar{x}-\mu_0\right)\left(\bar{x}-\mu_0\right)'\right|=$$

$$\left|\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'\right|\left|-1-n\left(\bar{x}-\mu_0\right)'\left(\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'\right)^{-1}\left(\bar{x}-\mu_0\right)\right|$$

Έχουμε ότι:

$$\sum_{j=1}^{n}\left(x_j-\mu_0\right)\left(x_j-\mu_0\right)'=\sum_{j=1}^{n}\left(x_j-\bar{x}+\bar{x}-\mu_0\right)\left(x_j-\bar{x}+\bar{x}-\mu_0\right)'$$

$$=\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'+n\left(\bar{x}-\mu_0\right)\left(\bar{x}-\mu_0\right)'$$

Η παραπάνω ισότητα μπορεί να γραφτεί με βάση τους καθοριστικούς παράγοντες

$$(-1)\left|\sum_{j=1}^{n}\left(x_j-\mu_0\right)\left(x_j-\mu_0\right)'\right|=\left|\sum_{j=1}^{n}\left(x_j-\bar{x}\right)\left(x_j-\bar{x}\right)'\right|(-1)\left(1+\frac{T^2}{(n-1)}\right)$$

ή $$\left|n\hat{\Sigma}_0\right|=\left|n\hat{\Sigma}\right|\left(1+\frac{T^2}{(n-1)}\right)$$

Έτσι, έχουμε ότι: $\Lambda^{2/n} = \dfrac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_0\right|} = \left(1 + \dfrac{T^2}{(n-1)}\right)^{-1}$

Εδώ το H_0 απορρίπτεται για μικρές τιμές του $\Lambda^{2/n}$ ή ισοδύναμα για μεγάλες τιμές του T^2. Καθορίζονται οι κρίσιμες τιμές των T^2 από τον τύπο:

$$a = P\left[n\left(\bar{x} - \mu\right)' S^{-1}\left(\bar{x} - \mu\right) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \right]$$

Η σχέση $\Lambda^{2/n} = \dfrac{\left|\hat{\Sigma}\right|}{\left|\hat{\Sigma}_0\right|} = \left(1 + \dfrac{T^2}{(n-1)}\right)^{-1}$ δείχνει ότι το T^2 μπορεί να υπολογιστεί από δυο καθοριστικούς παράγοντες αποφεύγοντας έτσι τον υπολογισμό του S^{-1}.

Λύνοντας, λοιπόν, αυτή τη σχέση ως προς T^2, θα έχουμε ότι:

$$T^2 = \frac{(n-1)\left|\hat{\Sigma}_0\right|}{\left|\hat{\Sigma}\right|} - (n-1) = (n-1)\frac{\left|\sum_{j=1}^{n}\left(x_j - \mu_0\right)\left(x_j - \mu_0\right)'\right|}{\left|\sum_{j=1}^{n}\left(x_j - \bar{x}\right)\left(x_j - \bar{x}\right)'\right|} - (n-1)$$

Τα τεστ του λόγου πιθανοφάνειας είναι κοινά στην πολυμεταβλητή ανάλυση.

Γενική Μέθοδος Λόγου Πιθανότητας

Ας πάρουμε το Θ να είναι ένα διάνυσμα που αποτελείται απ' όλες τις άγνωστες παραμέτρους του πληθυσμού και ας είναι το $L(\theta)$ η λειτουργία πιθανοφάνειας, που προέρχεται από την αξιολόγηση της κοινής πυκνότητας των $X_1, X_2, ..., X_n$ σε παρατηρούμενες τιμές των $X_1, X_2, ..., X_n$.

Το διάνυσμα των παραμέτρων θ παίρνει την τιμή του στο σύνολο παραμέτρων του Θ. Για παράδειγμα, στην $p-$διαστάσεων πολυμεταβλητή κανονική περίπτωση, έχουμε ότι:

$\theta' = \left[\mu_1, ..., \mu_{p,} \quad \sigma_{11}, ..., \sigma_{1p,} \quad \sigma_{22}, ..., \sigma_{2p,} \ ..., \ \sigma_{p-1,p}, \quad \sigma_{pp} \quad \right]$

και το Θ αποτελείται από την ένωση του $p-$διάστατου χώρου, όπου ισχύει $-\infty < \mu_1 < \infty, \ ..., \ -\infty < \mu_p < \infty$ και $\left[p(p+1)/2 \right]$-διάστατος χώρος των διακυμάνσεων και των συνδυακυμάνσεων όπου το Σ είναι θετικά ορισμένο.

Ως εκ τούτου το Θ έχει διάσταση $v = p + p(p+1)/2$. Με βάση τη μηδενική υπόθεση $H_0 : \Theta = \Theta_0$, το Θ περιορίζεται να βρίσκεται σε ένα υποσύνολο Θ_0 του Θ. Για την πολυμεταβλητή κανονική κατάσταση με $\mu = \mu_0$ και Σ απροσδιόριστο, το

$\Theta_0 = \left\{ \mu_1 = \mu_{10,} \mu_2 = \mu_{20,} ..., \mu_p = \mu_{p0}, \sigma_{11j} ..., \sigma_{1p}, \sigma_{22}, ..., \sigma_{2p}, ..., \sigma_{p-1,p}, \sigma_{pp} \text{ με } \Sigma \text{ θετικά ορισμένο} \right\}$

Έτσι, το Θ_0 έχει διάσταση $v_0 = 0 + p(p+1)/2 = (p(p+1)/2$

Ένα τεστ του λόγου πιθανοφάνειας με:

$H_0 : \Theta \in \Theta_0$ απορρίπτει το H_0 έναντι $H_1 : \Theta \notin \Theta_0$ εάν:

$$\Lambda = \frac{\max\limits_{\Theta \in \Theta_0} L(\Theta)}{\max\limits_{\Theta \in \Theta} L(\Theta)} < c$$

όπου c είναι μια σταθερά κατάλληλα επιλεγμένη.

Διαισθητικά, λοιπόν, απορρίπτουμε το H_0, εάν το μέγιστο της πιθανοφάνειας προέρχεται από επιτρεπόμενο θ που μεταβάλλεται στο σύνολο Θ_0 είναι πολύ μικρότερο από το μέγιστο της πιθανοφάνειας που προέρχεται από τη μεταβολή θ πάνω απ' όλες τις τιμές του Θ. Όταν η μέγιστη τιμή της πιθανοφάνειας είναι μικρότερη. Από το μέγιστο στο αριθμητή του τύπου:

$$\Lambda = \frac{\max\limits_{\Theta\in\Theta_0} L(\Theta)}{\max\limits_{\Theta\in\Theta} L(\Theta)} < c$$

είναι πολύ μικρότερη απ' ότι ο παρανομαστής Θ_0, που δεν περιέχει τις κατάλληλες τιμές για το θ.

Για κάθε εφαρμογή της μεθόδου του λόγου πιθανοφάνειας, θα πρέπει να λάβουμε την κατανομή δειγματοληψίας του λόγου πιθανοφάνειας Λ ως στατιστικό αποτέλεσμα της δοκιμής. Τότε, το c μπορεί να επιλεγεί για να παράγουν μια δοκιμή με ένα συγκεκριμένο επίπεδο σημαντικότητας a. Ωστόσο, όταν το μέγεθος του δείγματος είναι μεγάλο και πληρούνται ορισμένες προϋποθέσεις κανονικότητας, η κατανομή δειγματοληψίας του $-2l_n\Lambda$ έχει καλά προσεγγιστεί από την κατανομή x^2.

Αποτέλεσμα

Όταν το μέγεθος του δείγματος n είναι μεγάλο, τότε:

$-2\ln\Lambda = -2\ln\dfrac{\max\limits_{\Theta\in\Theta_0} L(\Theta)}{\max\limits_{\Theta\in\Theta} L(\Theta)}$ είναι κατά προσέγγιση μια $x^2_{v-v_0}$ τυχαία μεταβλητή. Εδώ, οι βαθμοί ελευθερίας είναι:

$v - v_0 = (\text{διάσταση του } \Theta) - (\text{διάσταση του } \Theta_0)$

Οι στατιστικές δοκιμές συγκρίνονται με βάση την ισχύ που είναι ορισμένη η καμπύλη της επιφάνειας, των οποίων το ύψος εκτιμάται από το P (δοκιμάζει απόρριψη H_0 / θ) σε κάθε παράμετρο διάνυσμα. Η ισχύς μετράει την ικανότητα η δοκιμή να απορρίψει το H_0 όταν αυτό δεν ισχύει. Στη σπάνια περίπτωση όπου έχουμε ότι $\theta = \theta_0$ που καθορίζεται από την H_0 με την εναλλακτική H_1, η οποία

αποτελείται από το $\theta = \theta_1$, τότε το τεστ του λόγου πιθανοφάνειας έχει την υψηλότερη ισχύ, μεταξύ απ' όλες τις δοκιμές με το ίδιο επίπεδο σημαντικότητας a.

Σε πολλές περιπτώσεις, όπου Θ έχει μια συνιστώσα, ο έλεγχος του λόγου πιθανοφάνειας είναι ομοιόμορφα πιο ισχυρός από όλες τις εναλλακτικές λύσεις του $H_0 : \theta = \theta_0$. Σε άλλες περιπτώσεις, κατέχει για μεγάλα δείγματα.

Η γενική, λοιπόν, εισαγωγή είναι ότι περιέχουν την υψηλότερη δυνατή ισχύ (μέσος όρος), όταν το μέγεθος του δείγματος είναι μεγάλο.

Περιοχή Εμπιστοσύνης και ταυτόχρονες συγκρίσεις συνιστωσών

Για να αποκτήσουμε την κύρια μέθοδο για την κατασκευή συμπερασμάτων από ένα δείγμα, χρειαζόμαστε την έννοια του διαστήματος εμπιστοσύνης μιας μονομεταβλητής σε μια πολυμεταβλητή περιοχή εμπιστοσύνης. Ας πάρουμε, λοιπόν, το θ να είναι ένα διάνυσμα αγνώστων παραμέτρων και Θ το σύνολο όλων των πιθανών τιμών του θ. Η περιοχή εμπιστοσύνης είναι μια περιοχή θ πιθανών τιμών. Αυτή η περιοχή προσδιορίζεται από τα δεδομένα, όπου πρέπει να προσδιορίσουμε με $R(X)$, όπου $X=[X_1,...,X_n]$ να είναι ο πίνακας δεδομένων.

Η περιοχή $R(X)$ είναι η $100(1-a)\%$ περιοχή εμπιστοσύνης πριν το δείγμα επιλεγεί:

$$P[R(X) \text{ καλύπτει το πραγματικό } \theta]=1-\alpha$$

Η πιθανότητα υπολογίζεται από το πραγματικό αλλά άγνωστο θ.

Η περιοχή εμπιστοσύνης για τη μέση τιμή μ ενός $p-$διάστατου κανονικού πληθυσμού είναι διαθέσιμη από τον τύπο:

$$P\left[n\left(\overline{X}-\mu\right)'S^{-1}\left(\overline{X}-\mu\right)>\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\right]$$

Πριν το δείγμα επιλεγεί, έχουμε:

$$P\left[n\left(\bar{x}-\mu\right)'S^{-1}\left(\bar{x}-\mu\right)>\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\right]=1-\alpha$$

Ανεξάρτητα από τις άγνωστες τιμές των μ και Σ.

Το $\overline{X}$ θα είναι μέσα στο:

$$\left[(n-1)pF_{p,n-p}(\alpha)/(n-p)\right]^{1/2}$$

του μ με πιθανότητα $1-a$, προερχόμενη απόσταση με όρους το $(S/n)^{-1}$.

Ειδικότερα, για ένα δείγμα, τα $\bar{x}$ και s μπορούν να υπολογιστούν και η ανισότητα

$$n\left(\bar{x}-\mu\right)'S^{-1}\left(\bar{x}-\mu\right)\le(n-1)p\ F_{p,n-p}(\alpha)/(n-p)$$

θα καθορίσει μια περιοχή $R(X)$, εντός του χώρου όλων των πιθανών παραμέτρων. Σε αυτήν την περίπτωση η περιοχή θα είναι μια έλλειψη με κέντρο το $\bar{x}$ και με ιδιοδιανύσματα και ιδιοτιμές από τον πίνακα S^{-1}. Αυτή η έλλειψη έχει $100(1-a)\%$ περιοχή εμπιστοσύνης για το μ.

Μια $100(1-a)\%$ περιοχή εμπιστοσύνης για το μέσο μιας $p-$διάστασης κανονικής κατανομής είναι ένα σύνολο από όλα τα μ, όπως:

$$n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu) \le \frac{p(n-1)}{(n-p)}\ F_{p,n-p}(\alpha)$$

όπου $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, $S = \frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})'$ και τα $x_1, x_2, ..., x_n$ είναι οι παρατηρήσεις του δείγματος.

Για να εξακριβωθεί αν για οποιαδήποτε μ_0, πέφτει μέσα στην περιοχή εμπιστοσύνης (είναι μια κατάλληλη τιμή για το μ) χρειαζόμαστε τη γενικευμένη τετραγωνική απόσταση $n(\bar{x}-\mu_0)'S^{-1}(\bar{x}-\mu_0)$ και να την συγκρίνουμε με το $[p(n-1)/(n-p)]\ F_{p,n-p}(\alpha)$. Εάν η τετραγωνική απόσταση είναι μεγαλύτερη από το $[p(n-1)/(n-p)]\ F_{p,n-p}(\alpha)$, το μ_0 δεν βρίσκεται μέσα στην περιοχή εμπιστοσύνης.

Δεδομένου ότι πρόκειται για μια αναλογία της δοκιμής $H_0 : \mu = \mu_0$ έναντι $H_1 : \mu \neq \mu_0$, παρατηρούμε ότι η περιοχή εμπιστοσύνης του:

$n(\bar{x}-\mu)'S^{-1}(\bar{x}-\mu) \le \frac{p(n-1)}{(n-p)}\ F_{p,n-p}(\alpha)$ αποτελείται από όλα τα διανύσματα μ_0, για τα οποία η δοκιμή T^2 δεν θα απέρριπτε την H_0 έναντι της H_1, σε σημαντικό επίπεδο a.

Για $p \ge 4$ δεν μπορούμε να αναπαραστήσουμε της κοινής περιοχής εμπιστοσύνης. Μπορούμε, όμως, να υπολογίσουμε τους άξονες της έλλειψης εμπιστοσύνης και το σχετικό του μήκος. Αυτό γίνεται με τη βοήθεια των ιδιοτιμών λ_i και των ιδιοδιανυσμάτων e_i του S.

Έτσι, έχουμε ότι η κατεύθυνση και τα μήκη των

$$n\left(\bar{x}-\mu\right)'S^{-1}\left(\bar{x}-\mu\right)\le c^2=\frac{p(n-1)}{(n-p)}\ \ F_{p,n-p}(\alpha)$$

καθορίζονται με τη μετάβαση

$$\sqrt{\lambda_i}\,\mathrm{c}/\sqrt{\mathrm{n}}=\sqrt{\lambda_i}\cdot\sqrt{p(n-1)F_{p,n-p}(\alpha)/\mathrm{n(n\text{-}p)}}$$

στις μονάδες κατά μήκος των ιδιοδιανυσμάτων e_i. Ξεκινώντας από το κέντρο $\bar{x}$, οι άξονες της έλλειψης εμπιστοσύνης είναι:

$$\pm\sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}\mathrm{e_i}\quad \text{όπου}\quad \mathrm{Se_i}=\lambda_i\mathrm{e_i},\quad i=1,2,...,p$$

Οι αναλογίες των λ_i θα βοηθήσουν να εντοπιστούν σχετικά ποσά της επιμήκυνσης κατά μήκος των ζευγών των αξόνων.

Παράδειγμα

Τα δεδομένα για την ακτινοβολία από τους φούρνους μικροκυμάτων είναι:

ας πάρουμε $x_1=\sqrt[4]{\text{μετρήσιμες ακτινοβολίες με την πόρτα κλειστή}}$

και $x_2=\sqrt[4]{\text{μετρήσιμες ακτινοβολίες με την πόρτα ανοικτή}}$

Για $n=2$ ζευγάρια των μετατρεπόμενων παρατηρήσεων, υπολογίζουμε:

$$\bar{x}=\begin{bmatrix}0.564\\0.603\end{bmatrix},\quad s=\begin{bmatrix}0.0144 & 0.0117\\0.0117 & 0.0146\end{bmatrix},\quad S^{-1}=\begin{bmatrix}203.018 & -163.391\\-163.391 & 200.228\end{bmatrix}$$

Τα ζευγάρια των ιδιοτιμών και των ιδιοδιανυσμάτων είναι:

$$\lambda_1=0.026\qquad \mathrm{e_1'}=\left[0.704,\quad 0.710\right]$$
$$\lambda_2=0.002\qquad \mathrm{e_2'}=\left[-0.710,\quad 0.704\right]$$

Το 95% της έλλειψης εμπιστοσύνης για το μ αποτελείται από όλες τις τιμές (μ_1, μ_2) ικανοποιώντας τη γνωστή σχέση

$$42\left[0.564-\mu_1,\quad 0.603-\mu_2\right]\begin{bmatrix} 203.018 & -163.391 \\ -163.391 & 200.228 \end{bmatrix}\begin{bmatrix} 0.564-\mu_1 \\ 0.603-\mu_2 \end{bmatrix} \le \frac{2(41)}{40}\quad \mathrm{F}_{2,40}(0.05)$$

ή από το $\mathrm{F}_{2,40}(0.05) = 3.23$

$$42(203.018)(0.564-\mu_1)^2 + 42(200.228)(0.603-\mu_2)^2 - 84(163.391)(0.564-\mu_1)(0.603-\mu_2)^2 \le 6.62$$

Για να δούμε εάν το $\mu' = \left[0.562,\quad 0.582\right]$ είναι μέσα στην περιοχή εμπιστοσύνης, υπολογίζουμε:

$$42(203.018)(0.564-0.562)^2 + 42(200.228)(0.603-\mu_2) \le 6.62$$

και καταλήγουμε ότι είναι μέσα στην περιοχή.

Ισοδύναμα, μια δοκιμή του $H_0 : \mu = \begin{bmatrix} 0,562 \\ 0,589 \end{bmatrix}$ θα μπορούσε να μην απορριφθεί στην εύνοια του $H_1 : \mu \neq \begin{bmatrix} 0,562 \\ 0,589 \end{bmatrix}$ στο επίπεδο $a = 0,05$.

Η κοινή έλλειψη εμπιστοσύνης αναπαριστάται από τη σχέση

$$t^2 = \frac{\left(\bar{x}-\mu_0\right)^2}{s^2/n} = n\left(\bar{x}-\mu_0\right)\left(s^2\right)^{-1}\left(\bar{x}-\mu_0\right)$$

Το κέντρο της έλλειψης βρίσκεται στο $\bar{x}' = [0.564, 0.603]$ και δίνονται τα χρονικά μήκη των μεγάλων και μικρών αξόνων

από $\sqrt{\lambda_1}\sqrt{\frac{p(n-1)}{n(n-p)}}F_{p,n-p},(\alpha)=\sqrt{0.026}\sqrt{\frac{2(41)}{42(40)}(3.23)}=0.064$

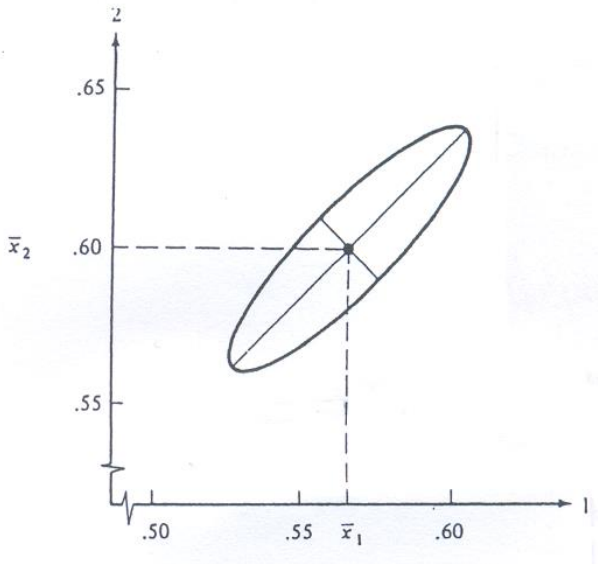
και $\sqrt{\lambda_2}\sqrt{\frac{p(n-1)}{n(n-p)}}F_{p,n-p},(\alpha)=\sqrt{0.002}\sqrt{\frac{2(41)}{42(40)}(3.23)}=0.018$

αντίστοιχα. Βρίσκονται κατά μήκος των $e'_1=[0.704,\ 0.710]$ και $e'_2=[-0.710,\ 0.704]$ όταν αυτά τα διανύσματα αναπαριστώνται με το $\bar{x}$. Η ένδειξη της έλλειψης εμπιστοσύνης δίνεται από τον λόγο του μήκους των μέγιστων και των ελάχιστων αξόνων.

Αυτή η σχέση είναι:

$$\frac{2\sqrt{\lambda_1}\sqrt{\frac{p(n-1)}{n(n-p)}}F_{p,n-p},(\alpha)}{2\sqrt{\lambda_2}\sqrt{\frac{p(n-1)}{n(n-p)}}F_{p,n-p},(\alpha)}=\frac{\sqrt{\lambda_1}}{\sqrt{\lambda_2}}=\frac{0.161}{0.045}=3.6$$

Το μήκος του κύριου άξονα είναι 3,6 φορές το μήκος του μικρού άξονα.



Ταυτόχρονες Δηλώσεις Εμπιστοσύνης

Ενώ η περιοχή εμπιστοσύνης $n\left(\bar{x}-\mu\right)' S^{-1}\left(\bar{x}-\mu\right)\le c^2$, όπου c είναι μια σταθερά, αξιολογεί σωστά την κοινή γνώση σχετικά με τις πιθανές τιμές για το μ, οποιαδήποτε περίληψη των συμπερασμάτων περιλαμβάνει συνήθως δηλώσεις εμπιστοσύνης για τα ιδιωτικά μέσα. Έτσι έχουμε υιοθετήσει τη στάση ότι όλες οι χωριστές δηλώσεις εμπιστοσύνης θα πρέπει να κατέχουν ταυτόχρονα με μια υψηλή πιθανότητα. Είναι η εγγύηση για μια δεδομένη πιθανότητα ενάντια σε οποιαδήποτε δήλωση ανακρίβειας που κινητοποιεί τα διαστήματα εμπιστοσύνης αρχίζοντας με την εξέταση των δηλώσεων εμπιστοσύνης, τα οποία είναι αλληλένδετα με την κοινή περιοχή εμπιστοσύνης που βασίζεται στη στατιστική T^2.

Ας πάρουμε το X να έχει $N_p\left(\mu,\Sigma\right)$ κατανομή και η μορφή του γραμμικού συνδυασμού είναι

$$Z=l_1x_1+l_2x_2+\cdots+l_px_p=l'x$$

Έχουμε $\mu_z=E(Z)=l'\mu$ και $\sigma_z^2=Var(Z)=l'\Sigma l$

Επιπλέον, το Z ακολουθεί κανονική κατανομή $N(l'\mu,l'\Sigma l)$

Εάν ένα τυχαίο δείγμα $X_1,X_2,...,X_n$ από έναν κανονικό πληθυσμό $N_p\left(\mu,\Sigma\right)$ είναι διαθέσιμο, υπάρχει ένα αντίστοιχο δείγμα του Z το οποίο μπορεί να δημιουργηθεί παίρνοντας τον γραμμικό συνδυασμό

$Z_j=l_1X_{1j}+l_2X_{2j}+\cdots+l_pX_{pj}=l'X_j \quad j=1,2,\ldots,n$

Η μέση τιμή δείγματος και η διακύμανση των $z_1,z_2,...,z_n$ είναι $\bar{z}=l'\bar{x}$ και $s_z^2=l'Sl$

όπου $\bar{x}$ και S είναι η μέση τιμή δείγματος και ο πίνακας συνδυακύμανσης του x_j αντίστοιχα.

Ταυτόχρονα διαστήματα εμπιστοσύνης μπορούν να αναπτυχθούν από την εξέταση των διαστημάτων εμπιστοσύνης για το $l'\cdot\mu$ για διαφορετικές επιλογές του l.

Για σταθερό l και άγνωστο σ_z^2 ένα $\left(1-a\right)$ διάστημα εμπιστοσύνης για $\mu_z=l'\mu$ βασίζεται στον λόγο t, όπου:

$$t = \frac{\bar{z} - \mu_z}{s_z / \sqrt{n}} = \frac{\sqrt{n}(l'\bar{x} - l'\mu)}{\sqrt{l'Sl}}$$

και οδηγεί στη δήλωση

$$\bar{z} - t_{n-1}(a/2)\frac{s_z}{\sqrt{n}} \le \mu_z \le \bar{z} + t_{n-1}(a/2)\frac{s_z}{\sqrt{n}}$$

ή

$$l'\bar{x} - t_{n-1}(a/2)\frac{\sqrt{l'Sl}}{\sqrt{n}} \le l'\mu \le l'\bar{x} + t_{n-1}(a/2)\frac{\sqrt{l'Sl}}{\sqrt{n}}$$

όπου το $t_{n-1}(a/2)$ είναι το ανώτερο όριο της $t-$κατανομής με $n-1$ βαθμούς ελευθερίας.

Η τελευταία ανισότητα μπορεί να ερμηνευτεί ως δήλωση σχετικά με τα στοιχεία του διανύσματος της μέσης τιμής μ. Για παράδειγμα, με $l' = [1, 0, ..., 0]$ το $l'\mu = \mu_1$ και η ανισότητα γίνεται το σύνηθες διάστημα εμπιστοσύνης για τη μέση τιμή ενός κανονικού πληθυσμού.

Σαφώς θα μπορούσαμε να κάνουμε πολλές δηλώσεις σχετικά με τα στοιχεία της εμπιστοσύνης μ, το καθένα έχοντας συντελεστή εμπιστοσύνης $1-a$, επιλέγοντας διαφορετικό συντελεστή διανύσματος του l. Ωστόσο, ο συντελεστής που σχετίζεται με το σύνολο των δηλώσεων που λαμβάνονται από κοινού δεν είναι $1-a$.

Διαισθητικά, θα ήταν καλύτερα να συνδεθεί ο συντελεστής $1-a$ με τα διανύσματα εμπιστοσύνης που μπορεί να προκύψουν από τις επιλογές του l.

Η τιμή είναι με τη μορφή των διαστημάτων που είναι λιγότερο ακριβή σε σχέση με το διάστημα της σχέσης

$$l'\bar{x} - t_{n-1}(a/2)\frac{\sqrt{l'Sl}}{\sqrt{n}} \le l'\mu \le l'\bar{x} + t_{n-1}(a/2)\frac{\sqrt{l'Sl}}{\sqrt{n}}$$

για μια συγκεκριμένη επιλογή του l.

Δίνεται το σύνολο των δεδομένων $X_1, X_2, ..., X_n$ και μια τιμή l και το διάστημα εμπιστοσύνης του

$$l'\bar{x} - t_{n-1}(a/2)\frac{\sqrt{l'Sl}}{\sqrt{n}} \le l\mu \le l'\bar{x} + t_{n-1}(a/2)\frac{\sqrt{l'Sl}}{\sqrt{n}}$$

είναι εκείνο το σύνολο των $l'\mu$ τιμών για το οποίο

$$|t| = \frac{\sqrt{n}\left(l'\bar{x} - l'\mu\right)}{\sqrt{l'Sl}} \le t_{n-1}(a/2)$$

ή ισοδύναμα

$$t^2 = \frac{n\left(l'\bar{x} - l'\mu\right)^2}{\sqrt{l'Sl}} = \frac{n\left(l'\left(\bar{x} - \mu\right)\right)^2}{l'Sl} \le t^2{}_{n-1}(a/2)$$

Μια ταυτόχρονη περιοχή εμπιστοσύνης δίνεται από ένα σύνολο τιμών $l'\mu$ έτσι ώστε το t^2 να είναι σχετικά μικρό για όλες τις επιλογές του l. Φαίνεται λογικό να περιμένουμε ότι η συνεχής σχέση $t^2_{n-1}(\alpha/2)$ στον προηγούμενο τύπο, θα αντικατασταθεί από μεγαλύτερη τιμή c^2, όταν οι δηλώσεις αναπτύσσονται για πολλές επιλογές του l.

Θεωρώντας τις τιμές του l, για τις οποίες $t^2 \le c^2$, οδηγούμαστε στον περιορισμό του:

$$\max_l t^2 = \max_l n = \frac{\left(l'\left(\bar{x} - \mu\right)\right)^2}{l'Sl}$$

Χρησιμοποιώντας το Λήμμα για την μεγιστοποίηση με $x = l$, $d = \left(\bar{x} - \mu\right)$ και $B = S$, προκύπτει ότι:

$$\max_{l}\frac{n\left(l'\left(\bar{x}-\mu\right)\right)^2}{l'Sl}=n\left[\max_{l}\frac{\left(l'\left(\bar{x}-\mu\right)\right)^2}{l'Sl}\right]=n\left(\bar{x}-\mu\right)'S^{-1}\left(\bar{x}-\mu\right)=T^2$$

με το μέγιστο ποσοστό που συνεπάγεται για l του $S^{-1}\left(\bar{x}-\mu\right)$

Αποτέλεσμα

Ας πάρουμε τα $X_1, X_2, ..., X_n$ να είναι ένα τυχαίο δείγμα από έναν κανονικό πληθυσμό $N_p\left(\mu,\Sigma\right)$ με Σ θετικά ορισμένο. Τότε, ταυτόχρονα για όλα τα l, το χρονικό διάστημα είναι:

$$l'\overline{X}-\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p},(\alpha)l'Sl}\ ,\ l'\overline{X}+\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p},(\alpha)l'Sl}$$

θα περιοριστεί το $l'\mu$ με πιθανότητα $1-a$.

Απόδειξη

Από τη σχέση 1 έχουμε ότι:

$T^2=n\left(\bar{x}-\mu\right)^{'}S^{-1}\left(\bar{x}-\mu\right)\le c^2$ συνεπάγεται $\frac{n\left(l^{'}\bar{x}-l^{'}\mu\right)^2}{l^{'}Sl}\le c^2$

για κάθε l, n

$$l^{'}\bar{x}-c\sqrt{\frac{l^{'}Sl}{n}}\le l^{'}\mu\le l^{'}\bar{x}+c\sqrt{\frac{l^{'}Sl}{n}}$$

για κάθε l. Διαλέγοντας $c^2=p(n-1)F_{p,n-p},(\alpha)/(n-p)$ δίνει χρονικά διαστήματα που θα περιέχουν το $l'\mu$ για όλα τα l με πιθανότητα $1-a=P[T^2\le c^2]$.

Είναι βολικό να αναφερθούμε στα ταυτόχρονα διαστήματα του αποτελέσματος των διαστημάτων του T^2, αφού η πιθανότητα κάλυψης καθορίζεται από την κατανομή της T^2. Οι καλύτερες επιλογές είναι $l^{'}=[1,0,...,0]$, $l^{'}=[0,1,...,0]$ και ούτω καθεξής μέσα από το $l^{'}=[0,0,...,1]$ για τα διαστήματα του T^2 που μας επιτρέπουν να συμπεριληφθούν:

$$\overline{x_1} - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p},(\alpha)} \sqrt{\frac{s_{11}}{n}} \le \mu_1 \le \overline{x_1} + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p},(\alpha)} \sqrt{\frac{s_{11}}{n}}$$

$$\overline{x_2} - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p},(\alpha)} \sqrt{\frac{s_{22}}{n}} \le \mu_2 \le \overline{x_2} + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p},(\alpha)} \sqrt{\frac{s_{22}}{n}}$$

$$\vdots$$

$$\overline{x_p} - \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p},(\alpha)} \sqrt{\frac{s_{pp}}{n}} \le \mu_p \le \overline{x_p} + \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p},(\alpha)} \sqrt{\frac{s_{pp}}{n}}$$

όλα διαθέτουν ταυτόχρονα, με συντελεστή εμπιστοσύνης $1-a$.

Να σημειώσουμε ότι χωρίς μεταβολή συντελεστή $1-a$, μπορούμε να κάνουμε δηλώσεις για τις διαφορές $\mu_i - \mu_k$ αντίστοιχα του $l' = [0,...,0,l_i,0,...,0,l_k,0,...,0]$, όπου $l_i = 1$ και $l_k = -1$. Σε αυτή την περίπτωση $l'Sl = S_{ii} - 2S_{ik} + S_{kk}$ και έχουμε την εξής δήλωση:

$$\overline{x}_1 - \overline{x}_k - \sqrt{\frac{p(n-1)}{(n-p)} \; F_p, \text{n-p}(\alpha)} \sqrt{\frac{S_{ii} - 2S_{ik} + S_{kk}}{n}} \le \mu_1 - \mu_k$$

$$\le \overline{x}_1 - \overline{x}_k - \sqrt{\frac{p(n-1)}{(n-p)} \; F_p, \text{n-p}(\alpha)} \sqrt{\frac{S_{ii} - 2S_{ik} + S_{kk}}{n}}$$

Τα ταυτόχρονα T^2 διαστήματα εμπιστοσύνης είναι ιδανικά για δεδομένα "snooping". Ο συντελεστής εμπιστοσύνης $1-a$ παραμένει αμετάβλητος για κάθε επιλογή του l, έτσι οι γραμμικοί συνδυασμοί των μέσων τιμών μ_i, η αξία βασίζεται στην επιθεώρηση της εξέτασης των στοιχείων που μπορούν να εκτιμηθούν.

Έλεγχοι Διανυσμάτων Μέσων τιμών

Γνωστός πίνακας διακύμανσης

Ξεκινάμε με την πιο απλή περίπτωση του ελέγχου για ένα διάνυσμα μέσων, στο οποίο είναι γνωστή η διακύμανση. Στη μονομεταβλητή περίπτωση γίνεται μέσω της ελέγχοσυνάρτησης Z και της κανονικής κατανομής, δηλαδή έχουμε τον έλεγχο:

$H_0 : \mu = \mu_0$ έναντι της $H_1 : \mu \neq \mu_0$.

Η ελεγχοσυνάρτηση είναι η:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Στην πολυμεταβλητη περίπτωση έχουμε να ελέγξουμε την υπόθεση
$H_0 : \mu = \mu_0$ έναντι της εναλλακτικής $H_1 : \mu \neq \mu_0$

Εδώ η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε θα είναι:

$$Z^2 = n\left(\bar{x} - \mu_0\right)' \Sigma^{-1} \left(\bar{x} - \mu_0\right)$$

η οποία ακολουθεί την $x^2_{k-κατανομή}$. Επομένως, παρατηρούμε, ότι ο έλεγχος μπορεί να γίνει είτε μέσω του υπολογισμού της p-value, είτε χρησιμοποιώντας την κριτική τιμή της κατανομής αυτής.
Σε ένα διάνυσμα όμως δεν είναι εύκολο να χρησιμοποιήσουμε μονόπλευρους ελέγχους, γι' αυτό το λόγο και παίρνουμε δίπλευρο έλεγχο. Οι μονόπλευροι βέβαια έλεγχοι θα μπορούσαν να οριστούν χρησιμοποιώντας ελέγχους πιθανοφάνειας.

Παρατηρούμε ότι μπορούμε να χρησιμοποιήσουμε την ελεγχοσυνάρτηση Z^2 για να κατασκευάσουμε από κοινού διαστήματα εμπιστοσύνης για όλο το διάνυσμα των μέσων.

Πράγματι ισχύει ότι:

$$P(Z^2 < x^2_{1-a,\kappa}) = P\left(n(\bar{x}-\mu_0)'\Sigma^{-1}(\bar{x}-\mu_0) < x^2_{1-a,\kappa}\right) = 1-a$$

με διανύσματα διαφορών για τα παρατηρηθέντα διανύσματα, όπου $x^2_{1-a,\kappa}$ είναι το $1-a$ ποσοστιαίο σημείο της x^2-κατανομής με κ βαθμούς ελευθερίας.

Άρα η εξίσωση

$$n(\bar{x}-\mu_0)'\Sigma^{-1}(\bar{x}-\mu_0) = x^2_{1-a,\kappa}$$

θα ορίζει ένα υπερελλειψοειδές, το οποίο με πιθανότητα $1-a$ περιλαμβάνει το άγνωστο διάνυσμα μέσων τιμών του πληθυσμού.

Άγνωστος Πίνακας Διακύμανσης

Υποθέτουμε, ότι ο πίνακας διακύμανσης είναι άγνωστος. Γνωρίζουμε, ότι οι μετρήσεις συχνά καταγράφονται κάτω από διαφορετικές πειραματικές συνθήκες, ώστε να δούμε αν οι απαντήσεις διαφέρουν σημαντικά μέσα σε αυτά τα σύνολα. Για παράδειγμα η δραστικότητα ενός καινούργιου φαρμάκου ή ο κορεσμός μιας διαφημιστικής καμπάνιας, μπορούν να καθοριστούν με το να συγκρίνουμε τις μετρήσεις πριν την «θεραπεία» (ναρκωτικό ή διαφήμιση) με αυτές μετά τη «θεραπεία». Μια λογική προσέγγιση, ώστε να συγκρίνουμε δυο θεραπείες ή την απουσία μιας θεραπείας, είναι να αναθέσουμε και τις δυο «θεραπείες» σε μια ενότητα. Οι διπλές απαντήσεις μπορούν να αναλυθούν με το να υπολογίσουμε τις διαφορές τους εξαλείφοντας πολλές από τις επιρροές των ξένων παραλλαγών ανά ενότητα.

Στη μονομεταβλητή περίπτωση χρησιμοποιούμε τη δειγματική διακύμανση, αντί για την διακύμανση του πληθυσμού και κάνουμε τον έλεγχο με την κατανομή t, δηλαδή η ελεγχοσυνάρτηση είναι η:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

όπου $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ και $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2$

ακολουθεί λοιπόν μια t -κατανομή με $n-1$ βαθμούς ελευθερίας.

Στην πολυμεταβλητή περίπτωση αντικαθιστούμε τον πίνακα διακύμανσης του πληθυσμού με το δειγματικό πίνακα διακύμανσης και έτσι προκύπτει η συνάρτηση του Hotelling με τύπο:

$$T^2 = n \cdot \left(\bar{x} - \mu_0\right)' S^{-1} \left(\bar{x} - \mu_0\right)$$

Με τη βοήθεια του πηλίκου πιθανοφάνειας ο Hotelling κατέληξε στο παρακάτω θεώρημα:

Θεώρημα

Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητα διανύσματα από τυχαίο δείγμα από έναν κανονικό πληθυσμό $N_k\left(\mu, \Sigma\right)$, όπου $n > k$ και Σ ένας θετικά ορισμένος πίνακας, με μ και Σ άγνωστα.

Τότε, για τον έλεγχο $H_0 : \mu = \mu_0$ έναντι της εναλλακτικής υπόθεσης $H_1 : \mu \neq \mu_0$, ο έλεγχος πηλίκου πιθανοφάνειας απορρίπτει την μηδενική υπόθεση H_0 όταν:

$$T^2 = n \cdot (\bar{X} - \mu_0)' S^{-1} \cdot (\bar{X} - \mu_0) \geq \frac{(n-1)k}{n-k} F_{k,n-k,a}$$

Όπου $F_{k,n-k,a}$ είναι το ανώτερο όριο μιας F -κατανομής με k και n-k τους βαθμούς ελευθερίας.

Ο έλεγχος αυτός ονομάστηκε test του Hotelling και αποτελεί γενίκευση του ελέγχου t .

Απόδειξη

Η ακριβής κατανομή του T^2 είναι:

$$P\left[n \cdot \left(\bar{x} - \mu_0\right)' S^{-1} \left(\bar{x} - \mu_0\right) \right] > \frac{(n-1)k}{n-k} \quad F_{k,n-k,a}$$

Η κατά προσέγγιση T^2 για την n-k ακολουθία είναι:

$n \cdot \left(\bar{x} - \mu_0\right)' S^{-1} \left(\bar{x} - \mu_0\right)$ για την κατανομή x_k^2.

Παίρνοντας τον έλεγχο με μηδενική υπόθεση $H_0 : \mu = 0$ έναντι της εναλλακτικής υπόθεσης $H_1 : \mu \neq 0$ για κανονικό πληθυσμό $N_k(\mu, \Sigma)$, που απορρίπτει τη μηδενική υπόθεση H_0, εάν ισχύει ότι:

$$T^2 = n \cdot \left(\bar{x} - \mu_0\right)' S^{-1} \left(\bar{x} - \mu_0\right) \geq \frac{(n-1)k}{n-k} \quad F_{k,n-k,a}$$

όπου $F_{k,n-k,a}$ συμβολίζει το a-ποσοστιαίο σημείο της κατανομής $F_{k,n-k}$, με k και n-k βαθμούς ελευθερίας. Εδώ οι $\bar{x}$ και S δίνονται από τους εξής τύπους:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j \text{ και } S = \frac{1}{n-1} \sum_{j=1}^{n} \left(x_j - \bar{x}\right)\left(x_j - \bar{x}\right)'$$

Η $1-a$ περιοχή εμπιστοσύνης για το μ αποτελείται από όλα τα μ, για τα οποία ισχύει ότι:

$$\left(\bar{x} - \mu_0\right)' S^{-1} \left(\bar{x} - \mu_0\right) \leq \frac{(n-1)k}{n-k} \quad F_{k,n-k,a}$$

Έτσι λοιπόν έχουμε ότι τα $1-a$ ταυτόχρονα διαστήματα εμπιστοσύνης για τα $\bar{x}$ είναι:

$$\overline{x_i} \pm \sqrt{\frac{(n-1)\cdot k \cdot F_{k,n-k,a}}{n-k}} \cdot \sqrt{\frac{s^2}{n}}$$

Επομένως, τα Bonferroni $1-a$ ταυτόχρονα διαστήματα εμπιστοσύνης για ατομικές διαφορές είναι:

$$\overline{x_i} \pm t_{n-1}\left(\frac{a}{2k}\right)\sqrt{\frac{s^2}{n}}$$

Όπου, τα $t_{n-1}\left(\frac{a}{2k}\right)$ είναι το ανώτερο όριο για την t -κατανομή με $n-1$ βαθμούς ελευθερίας.

Παράδειγμα

Στα εργοστάσια επεξεργασίας λυμάτων απαιτούνται από το νόμο να καταγράψουν τις αποβολές λυμάτων μέσα στα ποτάμια σε τακτική βάση. Η ανησυχία για την αξιοπιστία των δεδομένων οδήγησε σε μια μελέτη στην οποία δείγματα διαιρέθηκαν και στάλθηκαν σε δυο εργοστάσια για δοκιμή. Το μισό μέρος από το δείγμα στάλθηκε στο Wisconsin State Laboratory of Hygiene και το άλλο μισό σε ιδιωτικό εργαστήριο που χρησιμοποιείται σε προγράμματα παρακολούθησης.

Μετρήσεις (BOD) και ανασταλθέντα στερεά αποκτήθηκαν από $n=11$ χωριζόμενα δείγματα από τα δυο εργαστήρια.

Τα δεδομένα βρίσκονται στον πίνακα 1

Πίνακας 1 ΔΕΔΟΜΕΝΑ ΛΥΜΜΑΤΩΝ				
Δείγμα j	**Εμπορικό Εργαστήριο**		**Κρατικό εργαστήριο υγιεινής**	
	$x_{11j}(BOD)$	$x_{12j}(SS)$	$x_{21j}(BOD)$	$x_{22j}(SS)$
1	6	27	25	15
2	6	23	28	13
3	18	64	36	22
4	8	44	35	29
5	11	30	15	31
6	34	75	44	64
7	28	26	42	30
8	71	124	54	64
9	43	54	34	56
10	33	30	29	20
11	20	12	39	21

Συμφωνούν οι χημικές αναλύσεις των 2 εργοστασίων; Αν υπάρχουν διαφορές, ποια είναι η φύση τους;

Η T^2 για δοκιμή $H_0 : \delta' = [\delta_1, \delta_2] = [0,0]$ κατασκευάστηκε από τις διαφορές των ζευγών των παρατηρήσεων.

$d_{1j} = x_{11}j - x_{21}j$	-19	-22	-18	-27	-4	-10	-14	17	9	4	-19
$d_{2j} = x_{12}j - x_{22}j$	12	10	42	15	-1	11	-4	60	-2	10	-1

Εδώ έχουμε $\bar{d} = \begin{bmatrix} \overline{d_1} \\ \overline{d_2} \end{bmatrix} = \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix}$, $S_d = \begin{bmatrix} 199.26 & 88.38 \\ 88.38 & 418.61 \end{bmatrix}$

και $T^2 = \begin{bmatrix} -9.36, & 13.27 \end{bmatrix} = \begin{bmatrix} 0.0055 & 88.38 - 0.0012 \\ -0.0012 & 418.610.0026 \end{bmatrix} \begin{bmatrix} -9.36 \\ 13.27 \end{bmatrix} = 13.6$

Παίρνοντας $a = 0.05$ βρίσκουμε

$$\left[p(n-1)(n-p) \right] F_{p,n-p(0.05)} = \left[2(10)/9 \right] F_{2,9(0.05)} = 9.47$$

Με $T^2 = 13.6 > 9.47$ απορρίπτουμε την H_0 συνοψίζοντας ότι υπάρχει μη μηδενική μέση διαφορά μεταξύ των μετρήσεων των δυο εργαστηρίων. Αυτό εμφανίζεται από την επιθεώρηση των δεδομένων, για το οποίο το εμπορικό εργαστήριο τείνει να παράγει χαμηλότερες μετρήσεις BOD και υψηλότερες SS μετρήσεις από το δημοκρατικό εργαστήριο του Hygiene. Το 95% των ταυτόχρονων διαστημάτων εμπιστοσύνης για τις διαφορές των μέσων δ1 και δ2 μπορεί να υπολογιστεί χρησιμοποιώντας τον τύπο:

$$\delta_1 = d_i \pm \sqrt{\frac{(n-1)p \cdot F_{p(n-p)}(a)}{(n-p)}} \sqrt{\frac{s_{d_i}^2}{n}}$$

Αυτά τα διαστήματα είναι:

$$\delta_1 = d_i \pm \sqrt{\frac{(n-1)p \cdot F_{p(n-p)}(a)}{(n-p)}} \sqrt{\frac{s_{d_i}^2}{n}} = -9.36 \pm \sqrt{9.47} \sqrt{\frac{199.26}{11}}$$

ή (-22.46, 3.74)

$$\delta_2 : 13.27 \pm \sqrt{9.47} \sqrt{\frac{418.61}{11}}$$

ή (-5.71, 32.25)

Το 95% των ταυτόχρονων διαστημάτων εμπιστοσύνης συμπεριλαμβανομένου και του μηδενός, όμως η υπόθεση $H_0 : \delta = 0$ απορρίπτεται στο επίπεδο 5%.

Οι αποδείξεις δείχνουν πραγματικές διαφορές. Το δ=0 πέφτει έξω από το 95% της περιοχής εμπιστοσύνης για τα δ και αυτό είναι σύμφωνα με τη δοκιμή T^2. Το 95% των ταυτόχρονων διαστημάτων εμπιστοσύνης ισχύει για ολόκληρο το σύνολο που θα μπορούσε να δημιουργηθεί για το πιθανό γραμμικό συνδυασμό της φόρμας: $l_1\delta_1 + l_2\delta_2$. Τα χρονικά διαστήματα που αντιστοιχούν στις επιλογές $(l_1 = 1,\ l_2 = 0)$ και $(l_1 = 0,\ l_2 = 1)$ περιέχουν το μηδέν.

Άλλες επιλογές των l_1 *και* l_2 θα παράγουν ταυτόχρονα παύσεις που δεν θα περιέχουν το μηδέν. Εάν η υπόθεση $H_0 : \delta = 0$ δεν απορρίπτονταν τότε όλα τα ταυτόχρονα διαστήματα θα περιλάμβαναν το μηδέν.

Οι ταυτόχρονες παύσεις που συμβαίνουν στην περίπτωση Bonferroni επίσης καλύπτουν το μηδέν.

Η ανάλυση μας υπέθεσε μια κανονική κατανομή για το D_j.

Τα αριθμητικά αποτελέσματα του παραδείγματος δείγματος δείχνουν μια ασυνήθιστη περίσταση που μπορεί να συμβεί όταν βγάλουμε τα συμπεράσματα. Ευτυχώς με αυτά τα δεδομένα ένας λογάριθμος οδηγεί σε ξεκάθαρα συμπεράσματα, ενώ η υπόθεση της κανονικής κατανομής είναι λογική.

Ο πειραματιστής στο παράδειγμα μας διαίρεσε ένα δείγμα με το να κουνάει και μετά να ρίχνει γρήγορα μεταξύ των δυο μπουκαλιών για χημική ανάλυση. Αυτό ήταν λογικό, διότι μια απλή διαίρεση του δείγματος σε δυο κομμάτια που αποκτήθηκε με το να ρίξεις το πάνω μισό σε ένα μπουκάλι και το υπόλοιπο σ' ένα άλλο μπουκάλι, ίσως να έχει σαν αποτέλεσμα αιωρούμενα στερεά στο χαμηλότερο μισό εξαιτίας της καθίζησης. Τα δυο εργαστήρια τότε δεν θα δούλευαν με την ίδια τεχνική ή σαν πειραματικές ενότητες και τα αποτελέσματα δεν θα εξαρτώνται από την ικανότητα των εργαστηρίων μετρώντας τις τεχνικές.

Οπότε ένας ερευνητής μπορεί να ελέγξει την ανάθεση των θεραπειών σε πειραματικές ενότητες, ένα κατάλληλο ζευγάρωμα των ενοτήτων και μια τυχαία ανάθεση μπορεί να ενισχύσει την στατιστική ανάλυση. Διαφορές, αν υπάρχουν

μεταξύ υποτιθέμενων ιδίων ενοτήτων πρέπει να εντοπιστούν και οι ενότητες που μοιάζουν περισσότερο να ζευγαρωθούν. Επιπλέον, μια τυχαία ανάθεση μιας θεραπείας 1 σε μια ενότητα και μιας θεραπείας 2 σε μια άλλη ενότητα, θα βοηθήσει να εξαλειφθούν οι συστηματικές επιδράσεις μη ελεγχόμενων πηγών παραλλαγής. Τυχαία μπορεί να εφαρμοστεί καθορίζοντας αν η πρώτη ενότητα λαμβάνει θεραπεία κορώνα για τη θεραπεία 1 και γράμματα για τη θεραπεία 2. Μια ξεχωριστή ανεξαρτησία γίνεται για κάθε ζευγάρι.

Συγκρίνοντας Διανύσματα Μέσης Τιμής από δυο Πληθυσμούς

Χρησιμοποιώντας τη στατιστική T^2, για να δοκιμάσουμε την ισότητα των μέσων τιμών από δυο πολυμεταβλητούς πληθυσμούς συνειδητοποιούμε ότι μπορούν να αναπτυχθούν αναλογικά με τη μονομεταβλητή διαδικασία. Η στατιστική T^2 είναι κατάλληλη για να συγκρίνουμε απαντήσεις από ένα σύνολο πειραματικών ρυθμίσεων (πληθυσμός 1) με ανεξάρτητες απαντήσεις από ένα άλλο σύνολο πειραματικών ρυθμίσεων (πληθυσμός 2). Αυτό μπορεί να γίνει χωρίς να ελέγξουμε μονάδα σε μονάδα τη μεταβλητότητα, όπως στη ζευγαρωτή περίπτωση. Έτσι, εάν είναι δυνατόν οι πειραματικές μονάδες θα έπρεπε να ανατεθούν τυχαία στα σύνολα των πειραματικών συνθηκών. Η τυχαιοποίηση αυτή μειώνει την επίδραση της μεταβλητότητας ανά μονάδα. Αν και κάποια ακρίβεια χάνεται σε σχέση με τα ζευγάρια σύγκρισης, δεν απαιτείται ομοιογένεια, αφού τα συμπεράσματα ισχύουν για μια πιο γενική συλλογή από πειραματικές μονάδες.

Θεωρούμε ένα τυχαίο δείγμα με μέγεθος n_1 από πληθυσμό 1 και ένα δείγμα μεγέθους n_2 από πληθυσμό 2.

Οι παρατηρήσεις στις p μεταβλητές μπορεί να ποικίλουν ως εξής:

Πληθυσμός 1 n_1

$x_{11}, x_{12}, \ldots, x_{1n1}$ $\quad \bar{x}_1 = \frac{1}{n}\sum_{j=1}^{n_1} x_{1j} \quad S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}\left(x_{1j} - \bar{x}_1\right)\left(x_{1j} - \bar{x}_1\right)'$

Πληθυσμός 2 n_2

$x_{21}, x_{22}, \ldots, x_{2n2}$ $\quad \bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} \quad S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}\left(x_{2j} - \bar{x}_2\right)\left(x_{2j} - \bar{x}_2\right)'$

Οι δείκτες 1 και 2 μας δηλώνουν τον πληθυσμό.

Υπόθεση σχετικά με τη δομή των δεδομένων

1. Το δείγμα $X_{11},X_{12},\ldots,X_{1n1}$ είναι ένα τυχαίο δείγμα μεγέθους n_1 από ένα p - μεταβλητό πληθυσμό με μέση τιμή μ_1 και πίνακα συνδιακύμανσης Σ_1
2. Το δείγμα $X_{21},X_{22},\ldots,X_{2n2}$ είναι ένα τυχαίο δείγμα μεγέθους n_2 από p - μεταβλητό πληθυσμό με μέση τιμή μ_2 και πίνακα συνδιακύμανσης Σ_2
3. Επίσης, $X_{11},X_{12},\ldots,X_{1n1}$ είναι ανεξάρτητα των $X_{21},X_{22},\ldots,X_{2n2}$

Τι γίνεται στην περίπτωση που έχουμε δύο δείγματα; Υπάρχουν διαφορές ανάμεσα στα διαστήματα των δύο μέσων; Θα δούμε την περίπτωση που η διακύμανση είναι άγνωστη. Παράλληλα υποθέτουμε ότι η διακύμανση είναι κοινή.
Στη μονομεταβλητή περίπτωση έχουμε τις υποθέσεις
$H_0 :\mu_1=\mu_2$ $H_1 :\mu_1\neq\mu_2$

Η Ελεγχοσυνάρτηση είναι η: $t=\dfrac{\left(\bar{x}_1-\bar{x}_2\right)-(\mu_1-\mu_2)}{S_{pooled}/\sqrt{n_1+n_2-2}}\sim t_{n_1+n_2-2}$

που ακολουθεί κάτω από την μηδενική υπόθεση, την κατανομή t.
Η ποσότητα S^2_{pooled} είναι μια εκτίμηση της κοινής διακύμανσης και υπολογίζεται ως σταθμικός μέσος των διακυμάνσεων των δυο ομάδων.
Για μεγάλα δείγματα τώρα αυτή η δομή είναι επαρκής για να δημιουργήσουμε αναφορές για ένα $px1$ διάνυσμα διαφοράς $\mu_1-\mu_2$. Ωστόσο, όταν τα δείγματα μεγέθους n_1 και n_2 είναι μικρά, χρειάζεται να γίνουν περισσότερες υποθέσεις.

Περισσότερες υποθέσεις όταν τα n_1 και n_2 είναι μικρά

1. Και οι δυο πληθυσμοί είναι πολυμεταβλητοί κανονικοί.
2. Επίσης, ισχύει $\Sigma_1=\Sigma_2$ δηλαδή, έχουμε ίδιους πίνακες συνδιασποράς

Η δεύτερη υπόθεση $\Sigma_1=\Sigma_2$ είναι πιο δυνατή από κάθε μία μονομεταβλητή.
Εδώ υποθέτουμε ότι αρκετά ζεύγη διακυμάνσεως και συδιακυμάνσεων είναι σχεδόν ίσα.

Όταν $\Sigma_1 = \Sigma_2 = \Sigma$, $\sum_{j=1}^{n_1}\left(x_{1j} - \overline{x_1}\right)\left(x_{1j} - \overline{x_1}\right)'$ είναι ο υπολογισμός του $(n_1 - 1)\Sigma$

και $\sum_{j=1}^{n_2}\left(x_{2j} - \overline{x_2}\right)\left(x_{2j} - \overline{x_2}\right)'$ είναι ένας υπολογισμός του $(n_2 - 1)\Sigma$.

Συνεπώς, μπορούμε να συγκεντρώσουμε τις πληροφορίες και για τα δυο δείγματα, ώστε να υπολογίσουμε την κοινή συνδιασπορά Σ.

Έχουμε ότι:

$$S_{pooled} = \frac{\sum_{j=1}^{n_1}\left(x_{1j} - \bar{x}_1\right)\left(x_{1j} - \bar{x}_1\right)' + \sum_{j=1}^{n_2}\left(x_{2j} - \overline{x_2}\right)\left(x_{1j} - \overline{x_2}\right)'}{n_1 + n_2 - 2}$$

$$= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Η $\sum_{j=1}^{n_1}\left(x_{1j} - \bar{x}_1\right)\left(x_{1j} - \bar{x}_1\right)'$ έχει $n_1 - 1$ βαθμούς ελευθερίας και $\sum_{j=1}^{n_2}\left(x_{2j} - \bar{x}_2\right)\left(x_{2j} - \bar{x}_2\right)'$ έχει $n_2 - 1$ βαθμούς ελευθερίας, τότε ο διαιρέτης $(n_1 - 1) + (n_2 - 1)$ στον τύπο παραπάνω δημιουργείται συνδυάζοντας δυο μέρη από βαθμούς ελευθερίας.

Για να δοκιμάσουμε την υπόθεση ότι $\mu_1 - \mu_2 = \delta_0$, ενός συγκεκριμένου διανύσματος, υπολογίζουμε την τετραγωνική απόσταση από τα $\overline{x_1} - \overline{x_2}$ στο δ_0. Τώρα, έχουμε ότι:

$E(\bar{X}_1 - \bar{X}_2) = \mathrm{E}(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$

Τα $\bar{X}_1$ και $\bar{X}_2$ είναι ανεξάρτητα και έτσι έχουμε ότι η συνδιακύμανση
$cov(\bar{X}_1 - \bar{X}_2) = 0$

Άρα:

$$cov(\bar{X}_1 - \bar{X}_2) = cov(\bar{X}_1) + cov(\bar{X}_2) = \frac{1}{n_1}\Sigma + \frac{1}{n_2}\Sigma = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma$$

Επειδή το Spooled υπολογίζει το Σ, βλέπουμε ότι:

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pooled}$$

είναι ένας εκτιμητής του $cov(\bar{X}_1 - \bar{X}_2)$

Ο λόγος πιθανοφάνειας της δοκιμής είναι:

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

ο οποίος βασίζεται, στην τετραγωνική στατιστική απόσταση, T^2.

Έχουμε απόρριψη του H_0, εάν:

$$\mathrm{T}^2 = \left(\overline{x_1} - \overline{x_2} - \delta_0\right)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) spooled\right]^{-1} \left(\overline{x_1} - \overline{x_2} - \delta_0\right) > c^2$$

Όπου η κρίσιμη απόσταση c^2, καθορίζεται από την κατανομή των δυο δειγμάτων της στατιστικής T^2. Για τους ελέγχους των μέσων τιμών να αναφέρουμε ότι όταν απορριφθεί η μηδενική υπόθεση, μπορούμε να εστιάσουμε το ενδιαφέρον μας στο ποιες μεταβλητές είναι υπεύθυνες για την απόρριψη της υπόθεσης.

Αποτέλεσμα

Όταν $X_{11}, X_{12}, \ldots, X_{1n1}$ είναι ένα τυχαίο δείγμα μεγέθους n_1 από $N_p\left(\mu_1, \Sigma\right)$ και $X_{21}, X_{22}, \ldots, X_{2n1}$ είναι ένα ανεξάρτητο τυχαίο δείγμα μεγέθους n_2 από κανονικό πληθυσμό $N_p\left(\mu_2, \Sigma\right)$

Τότε:

$$T^2 = \left[\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)\right]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{pooled}\right]^{-1} \left[\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)\right]$$

κατανέμεται στην:

$$\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} \quad F_{p,n_1+n_2-p-1}$$

Συνεπώς,

$$P=\left[\left(\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)\right)'\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{pooled}\right]^{-1}\left(\overline{X}_1-\overline{X}_2-(\mu_1-\mu_2)\right)\le c^2\right]=1-a$$

όπου, $c^2=\frac{(n_1+n_2-2)p}{(n_1+n_2-p-1)} \quad F_{p,n_1+n_2-p-1}$ (α)

<u>Απόδειξη</u>

Πρώτα σημειώνουμε ότι:

$$\overline{X}_1-\overline{X}_2=\frac{1}{n_1}X_{11}+\frac{1}{n_1}X_{12}+\ldots+\frac{1}{n_1}X_{1n1}-\frac{1}{n_2}X_{21}-\frac{1}{n_2}X_{22}-\ldots-\frac{1}{n_2}X_{2n2}$$

κατανέμεται σε κανονική κατανομή

$$N_p\left(\mu_1-\mu_2,\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\Sigma\right)$$

έχοντας ότι $c_1=c_2=\ldots=c_{n_1}=1/n_1$ και $c_{n_1+1}=c_{n_2+2}=\ldots=c_{n_1+n_2}=-1/n_2$

Έχουμε επιπλέον ότι:

$(n_1-1)S_1$ κατανέμεται $W_{n_1-1}(\cdot|\Sigma)$ και $(n_2-1)S_2$ ως $W_{n_2-1}(\cdot|\Sigma_2)$

Υποθέτοντας ότι τα x_{1j} και τα x_{2j} είναι ανεξάρτητα, έτσι και τα $(n_1-1)S_1$ και $(n_2-1)S_2$ είναι ανεξάρτητα.

Τα $(n_1-1)S_1+(n_2-1)S_2$ κατανέμονται στο $W_{n_1+n_2-2}(\cdot|\Sigma)$

Έτσι:

$$T^2 = \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-\frac{1}{2}} \left(\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)\right)' S_{pooled}^{-1} \left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-\frac{1}{2}} \left(\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)\right)$$

$$= \begin{pmatrix} \textit{πολυμεταβλητό} \\ \textit{κανονικό τυχαίο} \\ \textit{διάνυσμα} \end{pmatrix}' \left(\frac{\textit{τυχαίος πίνακας Wishart}}{\textit{βαθμούς ελευθερίας}}\right)^{-1} \begin{pmatrix} \textit{πολυμεταβλητό} \\ \textit{κανονικό τυχαίο} \\ \textit{διάνυσμα} \end{pmatrix}$$

$$= N_p(0,\Sigma)' \left[\frac{W_{n_1+n_2-2}(\cdot|\Sigma)}{n_1+n_2-2}\right]^{-1} N_p(0,\Sigma)$$

Αρχικά ενδιαφερόμαστε για τις περιοχές εμπιστοσύνης που αφορούν τη διαφορά των μέσων τιμών $\mu_1 - \mu_2$. Όλα τα $\mu_1 - \mu_2$ κοντά στο $\overline{x_1} - \overline{x_2}$ αποτελούν την περιοχή εμπιστοσύνης. Είναι ένα ελλειψοειδές που επικεντρώνεται στις παρατηρημένες διαφορές $\overline{x_1} - \overline{x_2}$, των οποίων οι άξονες καθορίζονται από τις ιδιοτιμές και τα ιδιοδιανύσματα του spooled (ή S_{pooled}^{-1})

Παράδειγμα

Πενήντα σαπούνια κατασκευάζονται το καθένα με δυο τρόπους. Έχουμε δυο χαρακτηριστικά, $x_1 =$ σαπουνάδα αφρός και $x_2 =$ ήπιο, μαλακό.

Παράγονται λοιπόν με τις μεθόδους 1 και 2, όπου:

$$\overline{x_1} = \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix} \qquad S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$$

$$\overline{x}_2 = \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix} \qquad S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Αποκτώ το 95% περιοχή εμπιστοσύνης για $\mu_1 - \mu_2$

Σημειώνουμε ότι τα S_1 και S_2 είναι περίπου ίσα, έτσι είναι λογικό να συγκεντρώνονται. Έτσι:

$$S_{pooled} = \frac{(50-1)S_1 + (50-1)S_2}{50+50-2} = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$$

Επίσης,

$$\overline{x}_1 - \overline{x}_2 = \begin{bmatrix} -1.9 \\ 0.2 \end{bmatrix}$$

Έτσι, η έλλειψη εμπιστοσύνης επικεντρώνεται στο $\begin{bmatrix} -1.9, & 0.2 \end{bmatrix}'$ οι ιδιοτιμές και τα ιδιοδιανύσματα του S_{pooled} αποκτώνται από την εξίσωση:

$$0 = |S_{pooled} - \lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 5-\lambda \end{vmatrix} = \lambda^2 - 7\lambda + 9$$

Έτσι $\lambda_{1,2} = \dfrac{7 \pm \sqrt{49-36}}{2} \Leftrightarrow \begin{cases} \lambda_1 = 5.303 \\ \lambda_2 = 1.697 \end{cases}$

και τα αντίστοιχα ιδιοδιανύσματα e_1 και e_2 που δίνονται

$$S_{pooled} e_i = \lambda_i e_i, \qquad i = 1,2$$

είναι $e_1 = \begin{bmatrix} 0.290 \\ 0.957 \end{bmatrix}$ και $e_2 = \begin{bmatrix} 0.957 \\ -0.290 \end{bmatrix}$

έχουμε:

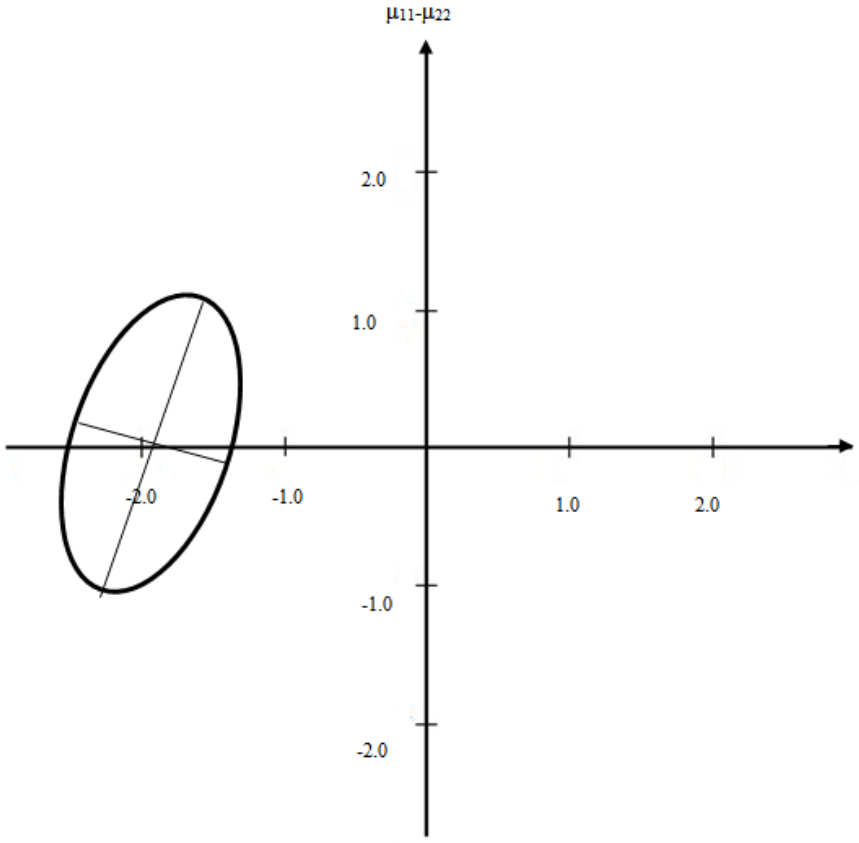
$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2 = \left(\frac{1}{50} + \frac{1}{50}\right) = \frac{98 \cdot 2}{97} F_{2,97}(0,05) = 0,25$$

Όπου $F_{2,97}(0,05) = 3.1$

Η έλλειψη εμπιστοσύνης εκτείνεται

$$\sqrt{\lambda_i}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)c^2}=\sqrt{\lambda_i}\sqrt{0.25}$$

κατά μήκος του ιδιοδιανύσματος e_1 ή 1.15 μονάδες στην e_1 κατεύθυνση και 0.65 μονάδες στην κατεύθυνση e_2 καθαρά τα $\mu_1-\mu_2=0$ δεν είναι στην έλλειψη και κλείνουμε με τις δυο μεθόδους κατασκευής του σαπουνιού παράγοντας διαφορετικά αποτελέσματα.



Σχήμα: 95% έλλειψη εμπιστοσύνης για $\mu_1-\mu_2$

Φαίνεται σαν δυο διαδικασίες που παράγουν σαπούνι με την ίδια ηπιότητα (X_2), αλλά εκείνες που παρήχθησαν από τη δεύτερη διαδικασία έχουν περισσότερο αφρό (X_1)

Η κατάσταση του διδιάστατου δείγματος όταν $\Sigma_1 \neq \Sigma_2$

Όταν έχουμε $\Sigma_1 \neq \Sigma_2$, τότε δεν είμαστε σε θέση να βρούμε ένα μέτρο απόστασης όπως είναι το T^2 του οποίου η κατανομή δεν ορίζεται, όταν τα Σ_1 και Σ_2 είναι άγνωστα. Για να ελέγξουμε την ισότητα των Σ_1 και Σ_2 σε σχέση με γενικευμένες διακυμάνσεις, χρησιμοποιούμε τον έλεγχο Bartlett. Όταν, όμως οι πληθυσμοί είναι μη κανονικοί υπάρχει η πιθανότητα τα συμπεράσματα να είναι παραπλανητικά. Οι μη κανονικές και άνισες συνδιακυμάνσεις δεν μπορούν να διαχωριστούν με τον έλεγχο Bartlett.

Μια μέθοδος για να ελέγξουμε την ισότητα δυο πινάκων συνδιακύμανσης, η οποία είναι λιγότερο ευαίσθητη στην υπόθεση της πολυδιάστατης κανονικότητας έχει προταθεί από τους Tiku και Balakrishnan.

Δυστυχώς, όμως χρειάζεται μεγαλύτερη πρακτική εμπειρία με αυτόν τον έλεγχο.

Από την άλλη ένας μετασχηματισμός μπορεί να βελτιώσει τα πράγματα στους ελέγχους, όταν οι οριακές διακυμάνσεις είναι αρκετά διαφορετικές.

Έτσι για n_1 και n_2 μεγάλα, μπορούμε να αγνοήσουμε τις πολυπλοκότητες που δημιουργούνται εξαιτίας των άνισων πινάκων συνδιακύμανσης.

Αποτέλεσμα

Ας πάρουμε τα μεγέθη δυο δειγμάτων n_1-p και n_2-p να είναι μεγάλα. Η κατά προσέγγιση 100(1-α)% έλλειψη εμπιστοσύνης για μ_1-μ_2 δίνεται από όλα τα μ_1-μ_2 τα οποία ικανοποιούν τη σχέση:

$$[\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)]' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right]^{-1} [\overline{x_1} - \overline{x_2} - (\mu_1 - \mu_2)] \leq x_p^2(\alpha)$$

Όπου το $x_p^2(\alpha)$ είναι το ανώτερο ποσοστιαίο εκατοστημόριο μιας X^2-κατανομής με p βαθμούς εμπιστοσύνης.

Επιπλέον, τα 100(1-α)% ταυτόχρονα διαστήματα εμπιστοσύνης για όλους τους γραμμικούς συνδυασμούς $l'(\mu_1 - \mu_2)$ παρέχονται από:

Το $l'(\mu_1 - \mu_2)$ ανήκει στο $l'(\overline{x_1} - \overline{x_2}) \pm \sqrt{x_p^2(\alpha)}\sqrt{l'\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)l}$

Απόδειξη

Παίρνουμε ότι

$$Cov(\overline{X_1} - \overline{X_2}) = cov(\overline{X_1}) + cov(\overline{X_2}) = \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$$

και $E(\overline{X_1} - \overline{X_2}) = \mu_1 - \mu_2$

Από το Κεντρικό Οριακό Θεώρημα τα

$$\overline{X_1} - \overline{X_2} \sim N_p\left[\mu_1 - \mu_2, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right]$$

Εάν τα Σ_1 και Σ_2 είναι γνωστά, η τετραγωνική απόσταση των $\overline{X_1} - \overline{X_2}$ από τα μ_1-μ_2 θα ήταν

$$[\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)]'\left[\frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right]^{-1}[\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)]$$

Η τετραγωνική αυτή απόσταση προσεγγίζει την $x_p^2(\alpha)$ κατανομή.

Όταν τα n_1 και n_2 είναι μεγάλα με μεγάλη πιθανότητα, τότε το S_1 θα είναι πολύ κοντά στο Σ_1 και το S_2 θα είναι πολύ κοντά στο Σ_2 αντίστοιχα.

Παρατήρηση

Εάν $n_1=n_2=n$ και $(n-1)/(n+n-2)=1/2$, τότε

$$\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2 = \frac{1}{n_1}(S_1 + S_2) =$$

$$= \frac{(n-1)S_1 + (n-1)S_2}{n+n-2}\left(\frac{1}{n} + \frac{1}{n}\right) =$$

$=S_{pooled}\left(\frac{1}{n} + \frac{1}{n}\right)$

Σε ισομεγέθη δείγματα, η διαδικασία του μεγάλου δείγματος είναι βασικά η ίδια, με τη διαδικασία που βασίζεται σε συγκεντρωτικό πίνακα συνδιακύμανσης. Όταν, έχουμε μια διάσταση είναι γνωστό, ότι η επίδραση ανόμοιων διακυμάνσεων είναι ελάχιστη όταν τα μεγέθη είναι ίσα δηλαδή $n_1=n_2$, ενώ είναι μέγιστη, όταν το n_1 είναι πολύ μικρότερο από το n_2.

Παράδειγμα

Θα αναλύσουμε την ηλεκτρική κατανάλωση χρησιμοποιώντας μεγάλο δείγμα.

Δίνεται ότι $n_1=45$ και $n_2=45$

$$S_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix}, \quad \overline{x_1} = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 8632.0 & 19619.7 \\ 1916.7 & 55964.5 \end{bmatrix}, \quad \overline{x_2} = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix}$$

Υπολογίζουμε

$$\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2 =$$

$$\frac{1}{45}\begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix} + \frac{1}{45}\begin{bmatrix} 8632.0 & 19619.7 \\ 1916.7 & 55964,5 \end{bmatrix} = \begin{bmatrix} 464.17 & 886.08 \\ 886.08 & 3642.15 \end{bmatrix}$$

Τα 95% ταυτόχρονα διαστήματα εμπιστοσύνης για γραμμικούς συνδυασμούς είναι:

$$l'\,(\mu_1 - \mu_2) = [1,0]\begin{bmatrix} \mu_{11}-\mu_{21} \\ \mu_{12}-\mu_{22} \end{bmatrix} = \mu_{11} - \mu_{21}$$

και

$$l'\,(\mu_1 - \mu_2) = [0,1]\begin{bmatrix} \mu_{11}-\mu_{21} \\ \mu_{12}-\mu_{22} \end{bmatrix} = \mu_{12} - \mu_{22}$$

Με τύπο $l'(\overline{x_1} - \overline{x_2}) \pm \sqrt{x_p^2(\alpha)}\sqrt{l'\left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)l}$

Προκύπτει ότι:

$\mu_{11} - \mu_{21}: 74.4 \pm \sqrt{5.99}\sqrt{464.17}$ ή $(21.7, 127.1)$

$\mu_{12} - \mu_{22}: 201.6 \pm \sqrt{5.99}\sqrt{4642.15}$ ή $(75.8, 327.4)$

Εφαρμόζοντας την T^2-στατιστική για τον έλεγχο H_0:μ_1-μ_2=0 έναντι της εναλλακτικής H_1:μ_1-μ_2≠0 προκύπτει ότι:

$$T^2 = [\overline{x_1} - \overline{x_2}]'\left[\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right]^{-1}[\overline{x_1} - \overline{x_2}] =$$

$$\begin{bmatrix}204.4 - 130.0\\556.6 - 355.0\end{bmatrix}'\begin{bmatrix}464.17 & 886.08\\886.08 & 3642.15\end{bmatrix}^{-1}\begin{bmatrix}204.4 - 130.0\\556.6 - 355.0\end{bmatrix} =$$

$$[74.4 \quad 201.6](10^{-4})\begin{bmatrix}59.874 & 20.080\\-20.080 & 10.519\end{bmatrix}^{-1} = 15.66$$

Για α=0.05, η κρίσιμη τιμή είναι $x_2^2(0.05) = 5.99$

Απορρίπτεται η υπόθεση H_0, αφού

T_2=15.66>$x_2^2(0.05)$ (0.05)=5.99

Ο κρίσιμος γραμμικός συνδυασμός που οδηγεί στην απόρριψη της H_0, έχει ως συντελεστή διανύσματος το

$\hat{l} \propto \left(\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right)^{-1}(\overline{x_1} - \overline{x_2})$=

$(10^{-4})\begin{bmatrix}59.874 & 20.080\\-20.080 & 10.519\end{bmatrix}\begin{bmatrix}74.4\\201.6\end{bmatrix} =$

$\begin{bmatrix}0.041\\0.063\end{bmatrix}$

Η διαφορά όταν δεν είμαστε σε ώρα αιχμής για την κατανάλωση ρεύματος μεταξύ εκείνων που έχουν κλιματισμό και αυτών που δεν έχουν, συμφέρει περισσότερο από την αντίστοιχη διαφορά όταν βρισκόμαστε σε ώρα αιχμής, όταν αφορά την απόρριψη της H_0:μ_1-μ_2=0.

Παρόμοια στατιστική με αυτή της T^2, η οποία είναι λιγότερο ευαίσθητη για παρατηρήσεις μικρών ή μεσαίων σε μέγεθος δειγμάτων, αναπτύχτηκε από τους Tiku και Singh.

Όμως, αν το μέγεθος του δείγματος τείνει να είναι μεγάλο, τότε χρησιμοποιούμε την κατανομή Hotteling T^2 διότι δεν επηρεάζεται από μικρές αποκλίσεις.

ΒΙΒΛΙΟΓΡΑΦΙΑ

“Εισαγωγή στις πιθανότητες. Θεωρία και εφαρμογές”, Μέρος Ι, Β΄ Έκδοση, Κούτρας Μάρκος Β.

“Πολυμεταβλητή Στατιστική ανάλυση”, Προπτυχιακή Διπλωματική Εργασία, Μ. Φιλίππου, Τμήμα Μαθηματικών, Πανεπιστημίου Πατρών.

“Applied Multivariate Statistical Analysis”, Richard. A. Johnson – Dean W. Wichern

“Multivariate Statistical methods”, Donald F. Morrison