

Πολυμεταβλητές κατανομές

Στην πολυμεταβλητή στατιστική το ενδιαφέρον δεν εστιάζεται στην κατανομή κάθε τυχαίας μεταβλητής X_i $i=1,2,\dots,p$ ξεχωριστά αλλά στην μελέτη πολλών τυχαίων μεταβλητών $X_1, X_2, \dots, X_p$ συγχρόνως. Δηλαδή το ενδιαφέρον 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 διακριτές τυχαίες μεταβλητές.

Στη συνέχεια για λόγους ευκολίας θα αναφερόμαστε μόνο σε συνεχείς τυχαίες μεταβλητές. Τα αποτελέσματα ισχύουν και στην περίπτωση διακριτών τυχαίων μεταβλητών αν αντί για ολοκλήρωμα χρησιμοποιήσουμε άθροισμα.

Υπενθυμίζουμε μερικά στοιχεία από την (αλ37)
θεωρία κατανομών που αφορούν την μελέτη
πολλών μεταβλητών συγχρόνως

• Συμβολίζουμε την από κοινού συνάρτηση
πυκνότητας πιθανότητας $f(x_1, x_2, \dots, x_p)$

• Άλλες κατανομές που προκύπτουν από την
από κοινού συνάρτηση πυκνότητας πιθανότητας

• Περιθωρια κατανομή μιας τυχαιάς μεταβλητής

$$f(x_i) = \int_{x_1} \int_{x_2} \dots \int_{x_{i-1}} \int_{x_{i+1}} \dots \int_{x_p} f(x_1, x_2, \dots, x_p) dx_p \dots dx_1$$

$i = 1, 2, \dots, p$

Ανλαδή ολοκληρώνουμε προς όλες τις υπόλοι-
πες μεταβλητές εκτός της x_i

• Περιθωρια κατανομή δύο ή περισσότερων
τυχαίων μεταβλητών

π.χ. $f(x_1, x_2) = \int_{x_3} \int_{x_4} \dots \int_{x_p} f(x_1, x_2, \dots, x_p) dx_p \dots dx_3$

• Δεσμευμένη κατανομή διανύσματος τυχαίων
μεταβλητών δοθέντων των τιμών άλλων μεταβλητών

π.χ $f(x_1, x_2 / x_3, x_4, \dots, x_p) = \frac{f(x_1, x_2, \dots, x_p)}{f(x_3, x_4, \dots, x_p)}$

Παράδειγμα

Εστω $f(x_1, x_2, x_3, x_4, x_5)$ η από κοινού συνάρτηση πυκνότητας πιθανότητας

$$f(x_1, x_3 / x_2, x_4, x_5) = \frac{f(x_1, x_2, x_3, x_4, x_5)}{f(x_2, x_4, x_5)} =$$

$$= \frac{f(x_1, x_2, x_3, x_4, x_5)}{\int_{x_1} \int_{x_3} f(x_1, x_2, x_3, x_4, x_5) dx_3 dx_2}$$

$$\int_{x_1} \int_{x_3} f(x_1, x_2, x_3, x_4, x_5) dx_3 dx_2$$

Εμπειρικά μπορεί να πει κανείς πως στον παρανομαστή έχουμε την πληροφορία που υπάρχει.

- Η αναμενόμενη τιμή μιας συνάρτησης τυχαίων μεταβλητών $g(x_1, x_2, \dots, x_p)$

$$E(g(x_1, x_2, \dots, x_p)) = \int_{x_1} \dots \int_{x_p} g(x_1, \dots, x_p) f(x_1, x_2, \dots, x_p) dx_p \dots dx_1$$

Ειδικές περιπτώσεις

- Αν $g(x_1, x_2, \dots, x_p) = x_1$

$$E(x_1) = \int_{x_1} \dots \int_{x_p} x_1 \cdot f(x_1, x_2, \dots, x_p) dx_p \dots dx_1 =$$

$$= \int_{x_1} x_1 \left[\int_{x_2} \dots \int_{x_p} f(x_1, x_2, \dots, x_p) dx_p \dots dx_2 \right] dx_1 = \int_{x_1} x_1 \cdot f(x_1) dx_1$$

• Αν $g(x_1, x_2) = (x_1 - E(x_1))(x_2 - E(x_2))$ (66)39)

$$\text{cov}(x_1, x_2) = E((x_1 - E(x_1))(x_2 - E(x_2))) = E(x_1 x_2) - E(x_1)E(x_2)$$

συνδιακύμανση δύο τυχαίων μεταβλητών

Βασικές ιδιότητες της συνδιακύμανσης

- Αν οι τ.μ. X_1, X_2 είναι ανεξάρτητες τότε $\text{cov}(x_1, x_2) = 0$
- Αν $\text{cov}(x_1, x_2) = 0$ δεν ισχύει απαραίτητα ότι οι τ.μ. X_1, X_2 είναι ανεξάρτητες
- $\text{cov}(x_1 + x_3, x_2) = \text{cov}(x_1, x_2) + \text{cov}(x_3, x_2)$
- $\text{cov}(ax_1 + b, \gamma x_2 + \delta) = a\gamma \cdot \text{cov}(x_1, x_2)$

Παρατήρηση: Η συνδιακύμανση μιας τυχαίας μεταβλητής με τον εαυτό της είναι η γνωστή μας διακύμανση. Η συνδιακύμανση μπορεί να πάρει και αρνητικές τιμές

$$\text{cov}(X_1, X_1) = E(X_1 \cdot X_1) - E(X_1) \cdot E(X_1) = E(X_1^2) - (E(X_1))^2$$

Ορισμός τυχαίο διάνυσμα $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ είναι το διάνυσμα όπου όλα τα στοιχεία του είναι τυχαίες μεταβλητές

Εστω το $X_{p \times 1} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$ τυχαίο διάνυσμα

Η αναμενόμενη τιμή $E(X) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{bmatrix} = \mu$

Επίσης με $Cov(X)$ τον πίνακα διακυμάνσεων του τυχαίου διανύσματος X

$$Cov(X) = \begin{bmatrix} Var(x_1) & cov(x_1, x_2) & \dots & cov(x_1, x_p) \\ cov(x_1, x_2) & Var(x_2) & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_1, x_p) & cov(x_2, x_p) & \dots & Var(x_p) \end{bmatrix}$$

Ο πίνακας $Cov(X)$ είναι συμμετρικός

Εστω τώρα ένας πίνακας $C_{m \times p}$ ο οποίος περιλαμβάνει γνωστές σταθερές και όχι τυχαίες μεταβλητές τότε ισχύει:

Το διάνυσμα $Y = C \cdot X$ είναι ένα τυχαίο

$$\begin{matrix} m \times 1 & m \times p & p \times 1 \end{matrix}$$

διάνυσμα και

$$E(Y) = E(C \cdot X) = C E(X)$$

$$Cov(Y) = Cov(C \cdot X) = C \underbrace{Cov(X)}_{\Sigma_X} C^T = C \Sigma_X C^T$$

$$\text{Αν } Y = Cx + b \quad \text{τότε ισχύει}$$

$\begin{matrix} m \times 1 & m \times p & p \times 1 & m \times 1 \end{matrix}$

(66741)

$$E(Y) = C \cdot E(X) + b \quad \text{και}$$

$$\text{Cov}(Y) = C \cdot \Sigma_X \cdot C^T$$

Ορισμός Εστω τ. δ. X_1, X_2 . Θα λέμε ότι τα διανύσματα αυτά είναι ανεξάρτητα όταν όλες οι μεταβλητές του πρώτου διανύσματος είναι ανεξάρτητες από όλες τις μεταβλητές του δεύτερου διανύσματος.

Ανάλογα με τον ορισμό του τυχαίου διανύσματος έχουμε και τον τυχαίο πίνακα

$$X = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & & & \\ X_{m1} & X_{m2} & & X_{mp} \end{bmatrix} \quad \begin{matrix} \text{όπου } X_{ij} \\ i=1,2,\dots,m \\ j=1,2,\dots,p \end{matrix}$$

είναι τυχαίες μεταβλητές

Τότε

$$E(X) = \begin{bmatrix} E(X_{11}) & \dots & E(X_{1p}) \\ E(X_{21}) & \dots & E(X_{2p}) \\ \vdots & \dots & \vdots \\ E(X_{m1}) & \dots & E(X_{mp}) \end{bmatrix}$$

$$E(X+Y) = E(X) + E(Y)$$

X, Y πίνακες
ίδιων διαστάσεων

Επίσης

$$E(AXB) = A E(X) B \quad \text{όπου } A, B \text{ κατάλ-}$$

-ήτοι πίνακες (οχι τυχαιοί)

Ο πίνακας ενδιακρίμανσης τυχαιου πίνακα είναι αρκετά δυσχρηστος και δεν τον χρησιμοποιούμε.

- Όταν αναφερόμαστε στη συνάρτηση πυκνότητας ενός διανύσματος μιζάρτε για την από κοινού συνάρτηση πυκνότητας των τυχαιων μεταβλητών που το απαρτίζουν

Παράδειγμα Έστω X_1, X_2 τ.μ και

ορίσουμε τις τ.μ $Z_1 = X_1 - X_2$ και $Z_2 = X_1 + X_2$

και το τ.δ. $Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} X_1 - X_2 \\ X_1 + X_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \leftarrow$

$$Z = \underbrace{\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}}_C \cdot X = C \cdot X \quad \text{Να βρεθούν τα}$$

$$E(Z), \quad \text{Cov}(Z)$$

Λύση

(6843)

$$E(Z) = E(CX) = C \cdot E(X) = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} E(x_1) \\ E(x_2) \end{bmatrix} = \begin{bmatrix} E(x_1) - E(x_2) \\ E(x_1) + E(x_2) \end{bmatrix}$$

$$\begin{aligned} \text{Cov}(Z) &= C \cdot \text{Cov}(X) \cdot C^T = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \\ &= \begin{bmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \sigma_1^2 - \sigma_2^2 \\ \sigma_1^2 - \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} \end{bmatrix} \end{aligned}$$

Πολυμεταβλητή Κανονική κατανομή

Αν η απλή κανονική κατανομή αποτεδεί την βάση για τις περισσότερες απλές στατιστικές εφαρμογές (π.χ. ελέγχους υποθέσεων, γραμμική παλινδρόμηση κλπ) ανάλογη είναι και η χρήση της πολυμεταβλητής κανονικής κατανομής στην πολυμεταβλητή στατιστική

Γνωρίζουμε ότι στη μονοδιάστατη

περίπτωση η συνάρτηση πυκνότητας

πιθανότητας είναι

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

$-\infty < x, \mu < +\infty$
και $\sigma > 0$

Η πολυμεταβλητή κανονική κατανομή που ορίζεται για το τυχαίο διάνυσμα $X_{p \times 1}$ έχει από κοινού συνάρτηση πυκνότητας πιθανότητας

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

όπου $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$, μ είναι ένα διάνυσμα

$p \times 1$ και Σ ένας πίνακας $p \times p$

Το μ περιέχει τις αναμενόμενες τιμές των μεταβλητών είναι δηλαδή το διάνυσμα των μέσων ενώ ο Σ είναι ο πίνακας με τις συνδιακυμάνσεις.

Δηλαδή $\mu = E(X)$

$\Sigma = \text{cov}(X)$

Γενικά συμβολίζουμε $X \sim N_p(\mu, \Sigma)$

Ο πίνακας Σ είναι θετικά ημιορισμένος συμμετρικός και τα διαχώνια στοιχεία του δεν μπορεί να είναι αρνητικά

Επίσης η πολυμεταβλητή κανονική κατανομή έχει $\frac{p(p+3)}{2}$ άγνωστες παραμέτρους τις p

τιμές που είναι οι αναμενόμενες τιμές των
 τυχαίων μεταβλητών και τις $\frac{P(P+1)}{2}$ διαδο-
 ρετικές τιμές του συμμετρικού πίνακα
 διακύμανσης Σ , $P + \frac{P(P+1)}{2} = \frac{P(P+3)}{2}$

Ας δούμε την ειδική περίπτωση των ανεξάρ-
 τητων μεταβλητών όπου ο πίνακας Σ

είναι $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \sigma_P^2 \end{bmatrix}$ τότε $|\Sigma| = \det \Sigma = \prod_{i=1}^P \sigma_i^2$

και $\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & \frac{1}{\sigma_P^2} \end{bmatrix}$

Τότε η πολυμεταβλητή κανονική κατα-
 νομή παίρνει την μορφή

$$f(x) = \prod_{i=1}^P \frac{1}{(2\pi)^{1/2} \sigma_i} e^{-\frac{1}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^2}} \quad \text{το οποίο}$$

είναι το χυνόμενο P μονοδιάστατων
 Gaussian πυκνότητας

Η από κοινού αθροιστική συνάρτηση (6.46)
κατανομής μιας ποσομεταβλητής κατανομής
κανονικής ορίζεται

$$F(x_1, x_2, \dots, x_p) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_p \leq x_p) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_p} f(y_1, \dots, y_p)$$

$dy_1, dy_2, \dots, dy_p$.

Επειδή ο υπολογισμός των πολλαπλών ολο-
κληρωμάτων είναι εξαιρετικά δύσκολος
και χρειάζονται ειδικές αριθμητικές μέ-
θοδοι, αυτό κάνει την από κοινού
αθροιστική συνάρτηση κατανομής εξαιρετικά
δυσχρήστη και η χρησιμότητά της είναι
πολύ μικρή.

3.3 Διδιάστατη κανονική κατανομή

Τότε πως έχουμε μόνο 2 τυχαίες μεταβλητές X και Y . Τότε η από κοινού τους κατανομή έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 + 2\rho \left(\frac{x-\mu_x}{\sigma_x} \right) \left(\frac{y-\mu_y}{\sigma_y} \right) + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\}$$

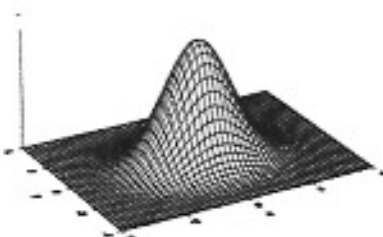
όπου $\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$ είναι η συσχέτιση των τ.μ. X και Y .

Το διάνυσμα των μέσων είναι το $\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ και ο πίνακας διασποράς $\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}$.

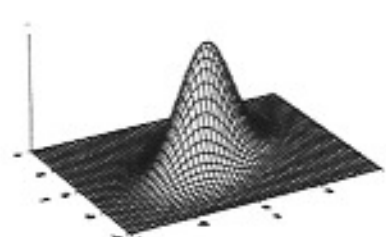
Παρατηρείστε πως αν $\rho=0$ τότε η από κοινού συνάρτηση πυκνότητας πιθανότητας της διμεταβλητής κατανομής είναι απλά το γινόμενο δύο συναρτήσεων πιθανοτήτων απλών κανονικών κατανομών, δηλαδή οι δύο μεταβλητές είναι ανεξάρτητες.

Για τη διδιάστατη 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 συνδέσει τη μορφή της κανονικής κατανομής με μια καμπάνα, δηλαδή μια συμμετρική κατανομή. Το ίδιο ισχύει και στη διδιάστατη περίπτωση, όπου τώρα έχουμε μια τρισδιάστατη εικόνα. Σε όλα τα γραφήματα 3.1-3.4 βλέπουμε πως η μορφή της κατανομής είναι συμμετρική και μοιάζει με μια καμπάνα της οποίας τα χαρακτηριστικά καθορίζονται από τις παραμέτρους. Η κορυφή της καμπάνας βρίσκεται πάντα

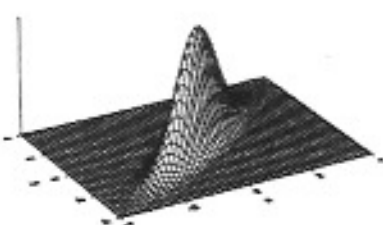
στο σημείο μ , δηλαδή ακριβώς στο σημείο που υποδεικνύει το διάνυσμα των μέσων.



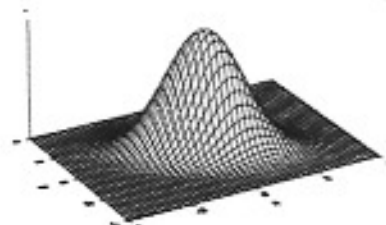
$\rho=0$



$\rho=0.5$

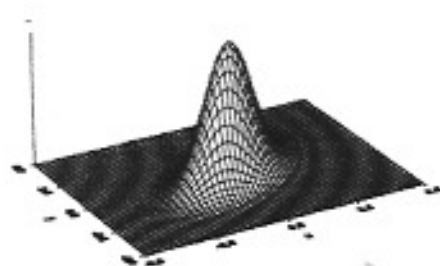


$\rho=0.9$

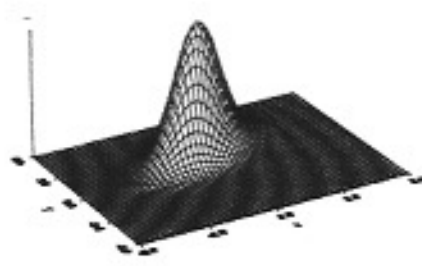


$\rho=0.5$

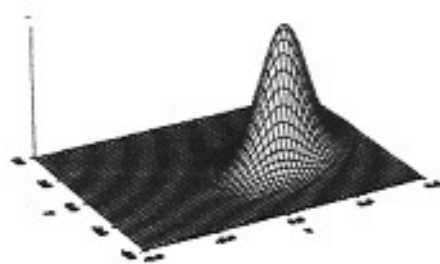
Εικόνα 3.1 Διδιάστατη κανονική κατανομή, με ίδιο διάνυσμα μέσων αλλά διαφορετική συσχέτιση ρ .



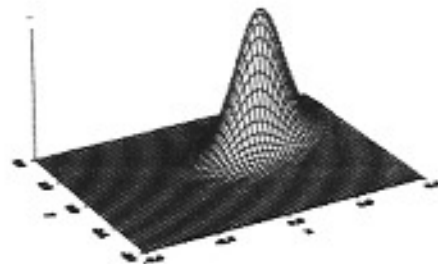
$$m=[0 \ 0]$$



$$m=[0 \ 2]$$



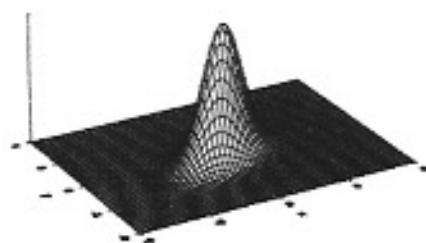
$$m=[2 \ 0]$$



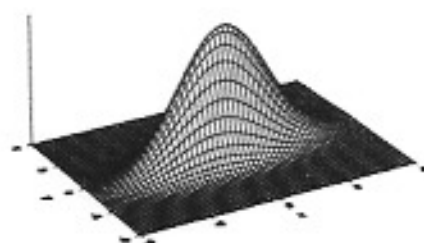
$$m=[2 \ 2]$$

Επίπεδα 3.2 Διόριστα, αντιστοιχία με ίδιες σφαιρικές, αλλά διαφορετικές θέσεις μέσης

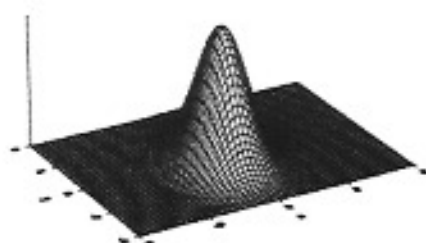
56



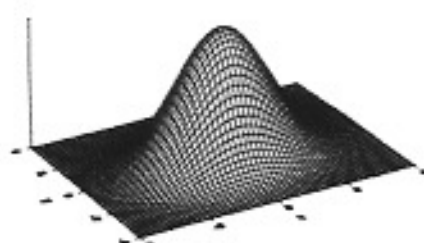
$$var=[4 \ 0.5 / \ 0.5 \ 1]$$



$$var=[4 \ 0.5 / \ 0.5 \ 1]$$

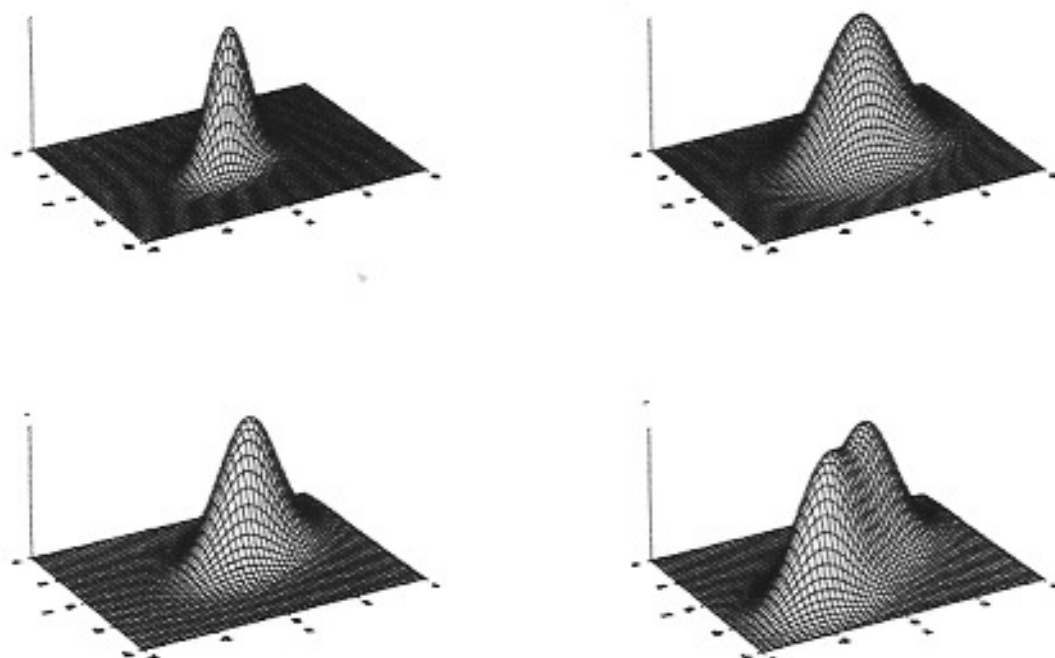


$$var=[4 \ 0.5 / \ 0.5 \ 1]$$



$$var=[4 \ 0.5 / \ 0.5 \ 1]$$

Επίπεδα 3.3 Διόριστα, αντιστοιχία με ίδιες σφαιρικές, αλλά διαφορετικές θέσεις μέσης



Γράφημα 3.4. Διάφορες μορφές διαταραχής κυματικής

Τότε

- Από το γράφημα 3.1 βλέπουμε την ερμηνεία του συντελεστή συσχέτισης ρ . Το δείνισμα των μέσων είναι το ίδιο για όλα τα γραφήματα και οι διακυμάνσεις είναι επίσης ίσες με 1 και για τις δύο διαστάσεις. Αυτό που βλέπουμε είναι πως το ρ καθορίζει απενός τον προσανατολισμό της από κοινού συνάμνησης πυκνότητας πιθανότητας, δηλαδή για θετικές τιμές η κατανομή τοποθετείται ως προς τη διαγώνιο από κάτω αριστερά προς το πάνω δεξιά σημείο, ενώ για αρνητικές τιμές από πάνω αριστερά προς κάτω δεξιά. Αρκετόσο η τιμή του συντελεστή καθορίζει πόσο μεγάλη βάση θα έχει η κατανομή ως προς τη διαγώνιο. Έτσι για τιμές του $\rho=0.9$ παρατηρούμε πως η κατανομή είναι πολύ συγκεντρωμένη πλησίον της διαγώνιου, ενώ για $\rho=0$ αυτό δεν συμβαίνει.

- Όπως βλέπουμε στο γράφημα 3.2 το δείνισμα των μέσων καθορίζει το σημείο κορυφής δηλαδή που θα τοποθετήσουμε την καμπάνια. Στο γράφημα έχουμε χρησιμοποιήσει τον ίδιο πίνακα διακύμανσης και συγκεκριμένα $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, ενώ το δείνισμα των μέσων αλλάζει. Ανάλογα με τις πρής μετακινούμεστε πάνω στους άξονες οι οποίες είναι οι ίδιες για όλα τα γραφήματα.

- Από το γράφημα 3.3 μπορούμε να δούμε την ερμηνεία των διακυμάνσεων. Αυτές λοιπόν καθορίζουν πόση έκταση ως προς τον αντίστοιχο άξονα θα έχουν οι καμπάνιες. Αν δηλαδή είναι ίσες οι διακυμάνσεις (τα διαγώνια στοιχεία του Σ) τότε η κατανομή θα απλώνεται το ίδιο και προς τους δύο άξονες.

- Τέλος, στο γράφημα 3.4 μπορεί κανείς να δει άλλες δεμεταβλητές κατανομές, μη κανονικές. Παρατηρούμε την ποικιλία σχημάτων που μπορούν να σχηματιστούν, δηλαδή κατανομές με πολύ πιο παχιά βάση από ότι η κανονική, ή ακόμα και με δύο κορυφές. Όλες οι κατανομές που εμφανίζονται στο γράφημα είναι μίγματα δεμεταβλητών κανονικών κατανομών. Η ποικιλία σχημάτων που μπορεί κανείς να σχηματίσει είναι ουσιαστικά άπειρη.

Το ακόμα ενδιαφέρον αποτέλεσμα είναι το εξής

Για τη δεμεταβλητή κοινική κατανομή ισχύει πως αν οι τοιχίες μεταβλητές X και Y είναι αποσχέτιστες τότε είναι και ανεξάρτητες, ισχύει και το αντίστροφο.

Θυμηθείτε πως γενικά το αντίστροφο δεν ισχύει.

3.4 Ιδιότητες της πολυμεταβλητής κανονικής κατανομής

Ας δούμε τώρα μερικές χρήσιμες ιδιότητες της πολυμεταβλητής κανονικής κατανομής.

Γραμμικοί Μετασχηματισμοί

Θεώρημα Έστω $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και έστω $\boldsymbol{\alpha}$ ένα διάνυσμα διαστάσεων $(p \times 1)$ από σταθερούς αριθμούς (όχι τυχαίες μεταβλητές). Τότε για την τυχαία μεταβλητή $y = \boldsymbol{\alpha}'\mathbf{x}$ ισχύει πως : $y \sim N(\boldsymbol{\alpha}'\boldsymbol{\mu}, \boldsymbol{\alpha}'\boldsymbol{\Sigma}\boldsymbol{\alpha})$

Η απόδειξη του θεωρήματος δεν θα δοθεί αλλά απλά πρόκειται για τη χρήση της γνωστής μεθόδου μετασχηματισμού τυχαίων μεταβλητών. Αυτό που έχει περισσότερη σημασία είναι η ερμηνεία του παραπάνω αποτελέσματος. Το θεώρημα λοιπόν μας επιτρέπει να δούμε πως η κατανομή γραμμικού συνδυασμού μεταβλητών με από κοινού κατανομή την πολυδιάστατη κανονική ακολουθεί την μονοδιάστατη κανονική κατανομή. Είναι γνωστό πως το αποτέλεσμα ισχύει για την περίπτωση ανεξάρτητων κανονικών τυχαίων μεταβλητών, τώρα παρατηρούμε πως το ίδιο ισχύει και για την περίπτωση που οι τιμές έχουν συσχέτιση μεταξύ τους.

Εφαρμογή: Κατανομή Δειγματοκού Μέσου για Συσχετισμένες Παρατηρήσεις

Μια ενδιαφέρουσα εφαρμογή των ιδιοτήτων της πολυμεταβλητής κανονικής κατανομής είναι να εξετάσουμε την κατανομή της μέσης τιμής από ένα δείγμα που οι τιμές του δεν είναι ανεξάρτητες. Συγκεκριμένα, έστω $X_1, \dots, X_n$ συσχετισμένες παρατηρήσεις από κανονικές κατανομές και έστω πως η από κοινού τους κατανομή είναι πολυμεταβλητή κανονική. Δηλαδή υποθέτουμε πως το διάνυσμα $\mathbf{x}$ που περιέχει όλες τις τυχαίες μεταβλητές ακολουθεί την πολυμεταβλητή κανονική κατανομή. Έχουμε, δηλαδή πως

$$\mathbf{x} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ όπου } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_{nn} \end{bmatrix} \text{ και } \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}.$$

Προσέχουμε το $\boldsymbol{\alpha}' = \boldsymbol{\alpha}^T$

Μπορεί κάποιος εύκολα να δει πως η μέση τιμή προκύπτει ως ένας γραμμικός συνδυασμός των τυχαίων μας μεταβλητών. Συγκεκριμένα, ορίζουμε το διάνυσμα

$$\alpha_{n \times 1} : \alpha' = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{pmatrix}$$

Τότε είναι $y = \alpha' x \sim N(\alpha' \mu, \alpha' \Sigma \alpha)$ και συνεπώς $\bar{x} = \frac{\sum_{i=1}^n X_i}{n} \sim N(\alpha' \mu, \alpha' \Sigma \alpha)$. Κάνοντας τις πράξεις βρίσκουμε πως

$$\alpha' \mu = \frac{1}{n} \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{bmatrix} = \frac{\sum_{i=1}^n \mu_i}{n} \text{ και}$$

$$\begin{aligned} \alpha' \Sigma \alpha &= \frac{1}{n} \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{12} & \sigma_{22} & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_{nn} \end{bmatrix} \frac{1}{n} \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} = \\ &= \frac{1}{n^2} \begin{bmatrix} \sum_{i=1}^n \sigma_{i1} & \sum_{i=1}^n \sigma_{i2} & \dots & \sum_{i=1}^n \sigma_{in} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} = \frac{\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}}{n^2} \end{aligned}$$

Επομένως $\bar{x} \sim N\left(\frac{\sum_{i=1}^n \mu_i}{n}, \frac{\sum_{i=1}^n \sum_{j=1}^n \sigma_{ij}}{n^2}\right)$. Δηλαδή η κατανομή του μέσου είναι κανονική αλλά τώρα έχει αλλάξει η διακύμανση.

Ας δούμε κάποιες πιο ειδικές περιπτώσεις.

Στην περίπτωση ανεξάρτητων και ισόνομων τ.μ. δηλαδή στην περίπτωση που κάθε

$$X_i \sim N(\mu, \sigma^2), \quad i=1, \dots, n \quad \text{έχουμε πως} \quad \Sigma = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \text{ και άρα } \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right),$$

το γνωστό δηλαδή αποτέλεσμα για ανεξάρτητα δείγματα.

Για λόγους ευκολίας και χωρίς να χάνουμε τίποτα στη γενικότητα των αποτελεσμάτων ας υποθέσουμε πως $\sigma^2=1$. Έστω τώρα πως ξέρουμε πως κάθε μεταβλητή έχει συσχέτιση με κάθε άλλη ίση με ρ . (Παρατήρηση: αν $\rho>0$, τότε ο πίνακας Σ που προκύπτει είναι πίνακας διακύμανσης, αυτό όμως δεν ισχύει πάντα αν $\rho<0$).

Δηλαδή υποθέτουμε πως $\Sigma = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{bmatrix}$. Χρησιμοποιώντας το γενικό αποτέλεσμα

βρίσκουμε πως $\bar{\mathbf{x}} \sim N\left(\mu, \frac{n+n(n-1)\rho}{n^2}\right)$.

Αν αναλογιστούμε πως η περίπτωση ανεξάρτητου δείγματος αντιστοιχεί στην τιμή $\rho=0$, προκύπτει πολύ εύκολα πως αν $\rho>0$ η διακύμανση της μέσης τιμής από συσχετισμένα δεδομένα θα είναι μεγαλύτερη από αυτήν όταν τα δεδομένα ήταν ασυσχέτιστα. Αυτό είναι λογικό αν αναλογιστούμε πως θετικά συσχετισμένες μεταβλητές ουσιαστικά περιέχουν την ίδια πληροφορία. Συνεπώς η συνολική πληροφορία που έχουμε δεν είναι τόση όσο θα περιμέναμε γιατί κάποια πληροφορία επαναλαμβάνετε και μας οδηγεί στο να αυξάνει τη διακύμανση της μέσης τιμής.

Μπορεί κάποιος να δει πως ισχύει και το αντίθετο. Δηλαδή ας υποθέσουμε πως ο συντελεστής συσχέτισης είναι αρνητικός. Μπορεί εύκολα κανείς να επιβεβαιώσει πως για να είναι θετικά ορισμένος ο πίνακας Σ θα πρέπει να ισχύει πως $\rho > -\frac{1}{n-1}$. Όχι τυχαία αυτή είναι και η τιμή που μας εξασφαλίζει πως η μέση τιμή έχει θετική διακύμανση. Αν λοιπόν η τιμή του συντελεστή είναι αρνητική τότε η διακύμανση της μέσης τιμής μειώνεται και μάλιστα όσο πιο έντονη (αλλά επιτρεπτή) αρνητική συσχέτιση έχουμε τόσο μεγαλύτερη μείωση της διακύμανσης πετυχαίνουμε.

Αυτό συμβαίνει γιατί αρνητικά συσχετισμένες μεταβλητές περιέχουν περισσότερη πληροφορία. Επομένως μας συμφέρει να έχουμε αρνητικά συσχετισμένες μεταβλητές.

Σημείωση: Η τεχνική να χρησιμοποιούμε αρνητικά συσχετισμένες μεταβλητές χρησιμοποιείται στην προσομοίωση για να μειώσει κανείς την διακύμανση. Η μέθοδος ονομάζεται 'μέθοδος αντιθετικών μεταβλητών' (antithetic variables). $\square$

Εφαρμογή: Γραμμικό μοντέλο.

Έστω το απλό γραμμικό μοντέλο $y_i = \alpha + \beta x_i + \varepsilon_i$. Εκτιμούμε τις παραμέτρους α και β και συνήθως στη συνέχεια θέλουμε να βρούμε την κατανομή της πρόβλεψης σε ένα συγκεκριμένο σημείο, έστω x . Αυτό γίνεται συνήθως με τη χρήση του τύπου

$$\hat{y} = \hat{\alpha} + \hat{\beta}x,$$

$\hat{\alpha}, \hat{\beta}$ είναι οι εκτιμήσεις των α και β αντίστοιχα. Επομένως η πρόβλεψη είναι ένας γραμμικός συνδυασμός των τυχαίων μεταβλητών $\hat{\alpha}, \hat{\beta}$. Οι υποθέσεις του γραμμικού μοντέλου είναι πως

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{και πως}$$

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$

Οι υποθέσεις αυτές μπορούν να περιγραφούν ισοδύναμα ως εξής. Αν συμβολίσουμε με ε το τυχαίο διάνυσμα $\varepsilon' = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$. Οι παραπάνω υποθέσεις γράφονται ισοδύναμα ως

$$\varepsilon \sim N_n(0, \sigma^2 I)$$

Παρατηρείστε πως ο πίνακας διακύμανσης είναι διαγώνιος, άρα οι τμ ανεξάρτητες, η διακύμανση σταθερή και όλοι οι μέσοι ίσοι με 0.

Γνωρίζουμε πως αν ισχύουν οι υποθέσεις του γραμμικού μοντέλου τότε το διάνυσμα

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \sigma^2 (X'X)^{-1} \right)$$

δηλαδή η από κοινού κατανομή των παραμέτρων είναι η διδιάστατη κανονική κατανομή, όπου σ^2 είναι η (ας υποθέσουμε γνωστή) διακύμανση του μοντέλου και X ο γνωστός πίνακας σχεδιασμού. $\square$

Επομένως η κατανομή του $\hat{y}$ θα είναι κανονική κατανομή με μέσο $\begin{bmatrix} 1 & x \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \alpha + \beta x$ και

διακύμανση $\begin{bmatrix} 1 & x \end{bmatrix} \sigma^2 (X'X)^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$.

Θεώρημα Έστω $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και έστω $\mathbf{A}_{q \times p}$ ένας πίνακας διαστάσεων $(q \times p)$ από σταθερούς αριθμούς (όχι τυχαίες μεταβλητές) και $\mathbf{b}$ ένα διάνυσμα $(q \times 1)$. Τότε για το τυχαίο διάνυσμα $\mathbf{y} = \mathbf{Ax} + \mathbf{b}$ ($q \times 1$) ισχύει πως: $\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$

Εφαρμογή: Έστω οι τ.μ. X_1, X_2, X_3 που αποτελούν το τυχαίο διάνυσμα $\mathbf{x}' = (X_1, X_2, X_3)$ και έστω πως $\mathbf{x} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Ορίζουμε νέες μεταβλητές $Y_1 = X_1 - X_2$ και $Y_2 = X_2 - X_3$. Θέλουμε να βρούμε τη συνδιακύμανση των Y_1, Y_2 .

Έστω το διάνυσμα $\mathbf{y} = \mathbf{Ax}$, όπου $\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ και $\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$. Μπορεί κανείς εύκολα να δει πως η μορφή του $\mathbf{A}$ είναι η κατάλληλη για να μετασχηματίσουμε το αρχικό διάνυσμα $\mathbf{x}$ στο διάνυσμα $\mathbf{y}$.

Ο πίνακας συνδιακύμανσης θα είναι ο

$$\begin{aligned} \text{Cov}(\mathbf{y}) &= \mathbf{ACov}(\mathbf{x})\mathbf{A}' = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 - \sigma_{12} & \sigma_{12} - \sigma_2^2 & \sigma_{13} - \sigma_{23} \\ \sigma_{12} - \sigma_{13} & \sigma_2^2 - \sigma_{23} & \sigma_{23} - \sigma_3^2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \\ &= \begin{bmatrix} \sigma_1^2 - 2\sigma_{12} + \sigma_2^2 & \sigma_{12} + \sigma_{23} - \sigma_{13} - \sigma_2^2 \\ \sigma_{12} + \sigma_{23} - \sigma_{13} - \sigma_2^2 & \sigma_2^2 - 2\sigma_{23} + \sigma_3^2 \end{bmatrix} \end{aligned}$$

Συνεπώς η συνδιακύμανση θα είναι $\sigma_{12} + \sigma_{23} - \sigma_{13} - \sigma_2^2$ όπου σ_i^2 είναι η διακύμανση της τ.μ. X_i .

Ένα ακόμα ενδιαφέρον αποτέλεσμα είναι το εξής:

Γνωρίζουμε από τη θεωρία πως η κατανομή χ^2 προκύπτει αν υψώσουμε μια τυποποιημένη κανονική τ.μ. στο τετράγωνο, ή στη γενική μορφή

Αν $X_i \sim N(0,1)$, $i = 1, \dots, n$ και τα X_i είναι ανεξάρτητες τ.μ., τότε ισχύει πως $\sum_{i=1}^n X_i^2 \sim \chi_n^2$.

Αν κανείς ορίσει το διάνυσμα $\mathbf{x}' = [X_1, X_2, \dots, X_p]$ τότε $\mathbf{x}'\mathbf{x} = \sum_{i=1}^p X_i^2$, δηλαδή η κατανομή χ^2 προκύπτει από ένα τυχαίο διάνυσμα από ανεξάρτητες τυχαίες μεταβλητές.

Θα δείξουμε πως

Θεώρημα: Αν $\mathbf{x}_{p \times 1} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, τότε $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_p^2$.

Απόδειξη:

Επειδή $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, μπορεί κανείς εύκολα να δει από το προηγούμενο θεώρημα πως

$$\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{I}),$$

και επομένως όλα τα στοιχεία του $\mathbf{z}$ είναι ανεξάρτητα μεταξύ τους.

Όμως

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= \\ (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu}) &= \\ [\boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu})]' [\boldsymbol{\Sigma}^{-1/2} (\mathbf{x} - \boldsymbol{\mu})] &= \\ \mathbf{z}' \mathbf{z} = \sum_{i=1}^p Z_i^2 \end{aligned}$$

και επομένως αφού το διάνυσμα $\mathbf{z}$ αποτελείται από ανεξάρτητες τυποποιημένες κανονικές τ.μ (προσέξτε πως ο πίνακας διακύμανσης είναι ο μοναδιαίος πίνακας) προκύπτει πως

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_p^2$$

Ελλειψοειδή ίσης πιθανότητας

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

Επομένως μπορούμε να ορίσουμε πως

$$P\left[(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi_{p,\alpha}^2\right] = \alpha$$

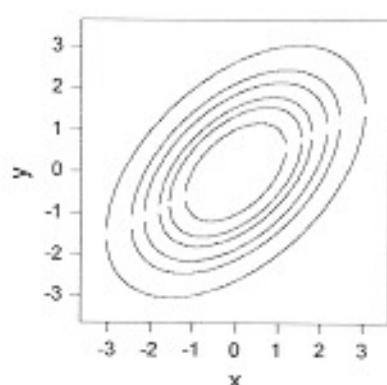
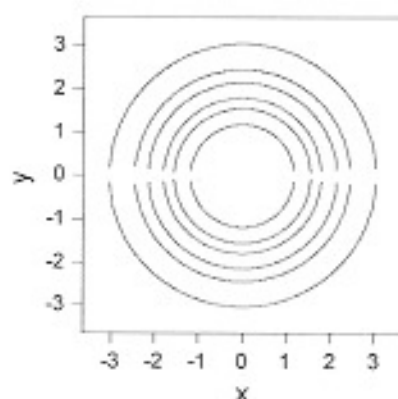
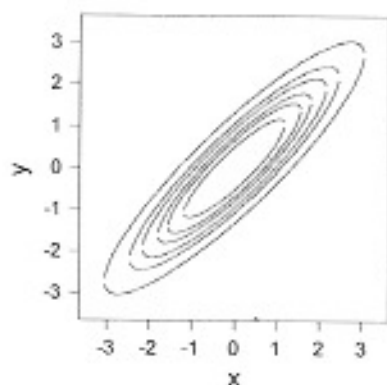
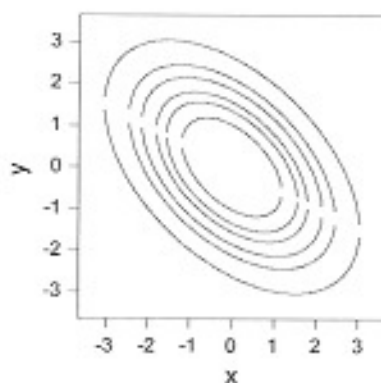
όπου $\chi_{p,\alpha}^2$ είναι το α -ποσοστιαίο σημείο της κατανομής χ^2 με p βαθμούς ελευθερίας.

Αυτό μας επιτρέπει να κατασκευάζουμε διαστήματα εμπιστοσύνης, ή καλύτερα περιοχές εμπιστοσύνης (confidence regions) για τα τυχαία διανύσματα. Οι περιοχές εμπιστοσύνης είναι τα αντίστοιχα των διαστημάτων εμπιστοσύνης για τη μονομεταβλητή περίπτωση. Το διάνυσμα $\boldsymbol{\mu}$ είναι το κέντρο όλων των ισοϋψών καμπυλών. Βέβαια για περισσότερες από 2 μεταβλητές δεν είναι καθόλου εύκολο να αναπαραστήσουμε γραφικά την περιοχή εμπιστοσύνης. Στις 2 διαστάσεις αυτή θα είναι μια έλλειψη, στις 3 διαστάσεις θα μοιάζει με μπάλα του ράγκμπι ενώ για περισσότερες δεν μπορεί να αναπαρασταθεί με κάποιο σχήμα.

Βέβαια κανείς μπορεί υπολογίζοντας την εξίσωση της έλλειψης να δει αν το σημείο που τον ενδιαφέρει είναι μέσα ή έξω από την περιοχή αυτή. Επίσης ο παραπάνω ορισμός έχει περιορισμένη χρησιμότητα καθώς υποθέτει πως το διάνυσμα των μέσων και ο πίνακας

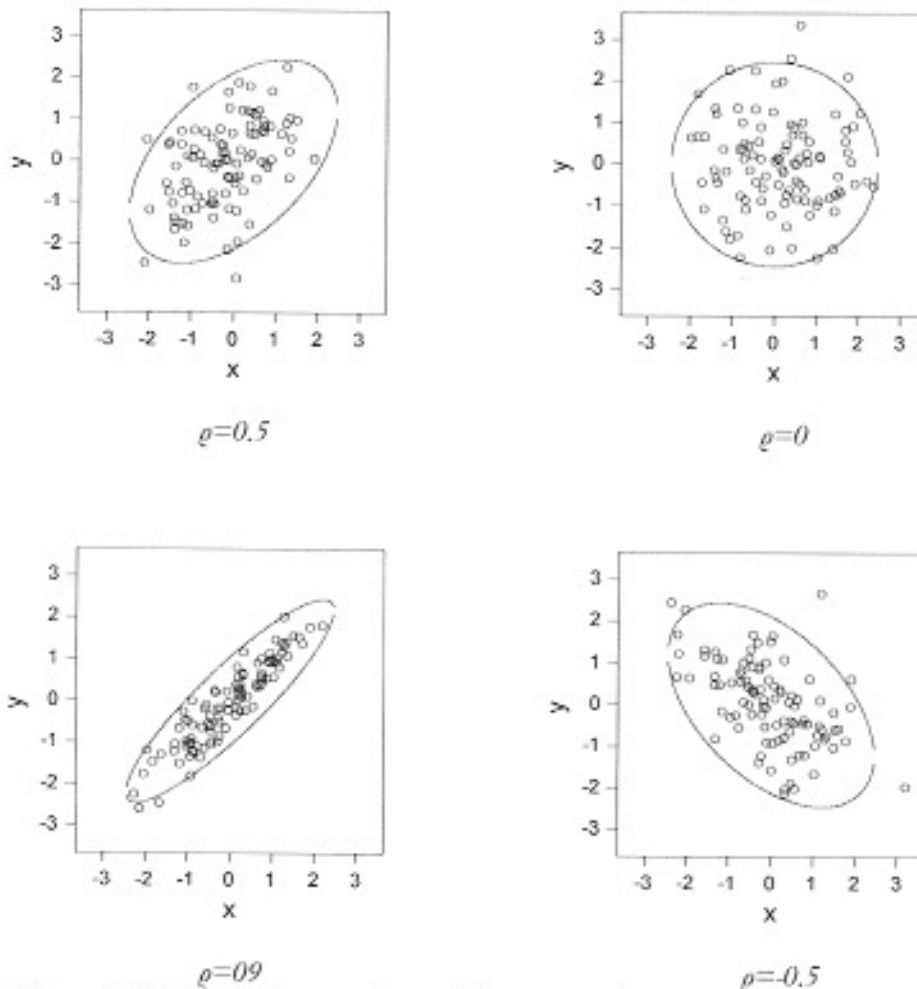
διακύμανσης είναι γνωστά. Θα δούμε αργότερα τι γίνεται με την περίπτωση που αυτά είναι άγνωστα και πρέπει να εκτιμηθούν.

Στο γράφημα 3.5 μπορεί να δει κανείς τα ελλειψοειδή για μια σειρά από τιμές για την πιθανότητα. Έτσι η εσωτερική έλλειψη αντιστοιχεί σε πιθανότητα 0.50 (με απλά λόγια η πιθανότητα μια παρατήρηση να είναι μέσα σε αυτή την έλλειψη είναι 0.5, ενώ με σειρά εμφάνισης από μέσα προς τα έξω οι ελλείψεις αντιστοιχούν σε πιθανότητες 0.5, 0.7, 0.8, 0.9, 0.95, 0.99. Μπορεί να παρατηρήσει κανείς πως αν τα δεδομένα είναι ασυσχέτιστα ($\rho=0$) τότε η έλλειψη εκφυλίζεται σε κύκλο.

 $\rho=0.5$  $\rho=0$  $\rho=0.9$  $\rho=-0.5$

Γράφημα 3.5. Ελλειψοειδή ίσης πυκνότητας για διάφορες τιμές του ρ . Το διάγραμμα των μέσων ήταν $[0, 0]$ για όλα τα γραφήματα, το ίδιο και οι διακυμάνσεις, ήταν όλες ίσες με 1

Στις περισσότερες διαστάσεις όταν οι μεταβλητές είναι ασυσχέτιστες προκύπτει πως το υπερελλειψοειδές γίνεται υπερσφαίρα, για αυτό και πολλές φορές την έλλειψη συσχέτισης την ονομάζουμε και σφαιρικότητα των δεδομένων. Το πρόσημο του συντελεστή συσχέτισης καθορίζει την κλίση των ελλείψεων. Τέλος η απόλυτη τιμή του συντελεστή καθορίζει τα χαρακτηριστικά της έλλειψης. Για την τιμή $\rho=0.9$ η έλλειψη είναι πολύ στενή, όσο η τιμή τείνει στο 1 η έλλειψη τείνει να εκφυλιστεί σε γραμμή.



Γράφημα 3.6. 95% περιοχή εμπιστοσύνης και διάγραμμα σημείων για προσομοιωμένα δείγματα από την διμεταβλητή κανονική κατανομή που υποθέσαμε. Το δείγμα των μέσων ήταν $[0 \ 0]$ για όλα τα γραφήματα, το ίδιο και οι διακυμάνσεις, ήταν όλες ίσες με 1. Από τον τρόπο κατασκευής της έλλειψης περιμένουμε περίπου το 95% των παρατηρήσεων να είναι μέσα στην έλλειψη

Για να εξετάσουμε λίγο την ιδέα των περιοχών εμπιστοσύνης για ένα τυχαίο διάνυσμα, προσομοιώσαμε δείγματα μεγέθους 100 από συγκεκριμένες διμεταβλητές κανονικές κατανομές. Για τον τρόπο προσομοίωσης θα μιλήσουμε αργότερα. Στο γράφημα 3.6 μπορεί κανείς να δει το διάγραμμα σημείων για το προσομοιωμένο δείγμα καθώς και μια 95% περιοχή

εμπιστοσύνης για κάθε παρατήρηση. Οι διμεταβλητές κατανομές είχαν διάνυσμα μέσων

$\mu' = [0 \ 0]$ και πίνακα διακύμανσης $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ για διάφορες τιμές του ρ . Θα πρέπει να

τονιστεί πως η επιλογή αυτή των παραμέτρων δεν επηρεάζει τα όποια συμπεράσματα καθώς το διάνυσμα των μέσων απλά μετακινεί το κέντρο της έλλειψης, ενώ ο πίνακας διακύμανσης απλά καθορίζει το μήκος των αξόνων. Επειδή οι περιοχές εμπιστοσύνης ήταν 95% αυτό σημαίνει πως περιμένουμε περίπου το 95% των παρατηρήσεων να είναι μέσα στην έλλειψη. Παρατηρείστε πως γενικά αυτό συμβαίνει. (Για λόγους τυχαιότητας σε κάποιες περιπτώσεις υπάρχουν λιγότερα ή περισσότερα από 5 σημεία έξω από τις ελλείψεις, αυτό ήταν αναμενόμενο)

3.5 Περιθώριες και δεσμευμένες κατανομές

Περιθώριες κατανομές

Έστω $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και $\mathbf{x}' = (X_1, X_2, \dots, X_p)$. Έστω πως θέλουμε να βρούμε την από κοινού περιθώρια κατανομή των q πρώτων μεταβλητών, δηλαδή του $\mathbf{y}$ όπου $\mathbf{y}' = (X_1, X_2, \dots, X_q)$, $q < p$. Θα πρέπει να πούμε πως η σειρά των μεταβλητών δεν παίζει καμιά σημασία και απλά θεωρούμε τις q πρώτες για λόγους ευκολίας.

$$\begin{aligned} \mathbf{y} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_q \end{bmatrix} &= \begin{bmatrix} X_1 + 0 \cdot X_2 + \dots + 0 \cdot X_q + \dots + 0 \cdot X_p \\ 0 \cdot X_1 + X_2 + \dots + 0 \cdot X_q + \dots + 0 \cdot X_p \\ \dots \\ 0 \cdot X_1 + 0 \cdot X_2 + \dots + X_q + \dots + 0 \cdot X_p \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & \dots & & 0 \\ \dots & \dots & \dots & \dots & & \ddots & \ddots & \\ 0 & \dots & \dots & 1 & 0 & \dots & \dots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_p \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{I}_{q \times q} & \mathbf{0}_{q \times (p-q)} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_p \end{bmatrix} \end{aligned}$$

όπου $\mathbf{0}_{m \times n}$ είναι ένας πίνακας με όλα του τα στοιχεία 0 και $\mathbf{I}$ είναι ο μοναδιαίος πίνακας. Δηλαδή

$\mathbf{y}_{q \times 1} = \mathbf{A}_{q \times p} \mathbf{x}_{p \times 1}$, όπου $\mathbf{A} = \begin{bmatrix} \mathbf{I}_{q \times q} & \mathbf{0}_{q \times (p-q)} \end{bmatrix}$ και επομένως κάνοντας χρήση του προηγούμενου

θεωρήματος προκύπτει πως $\mathbf{y} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$

Χρησιμοποιώντας διαμερισμένους (ή μερικούς) πίνακες (partitioned matrices) και διανύσματα μπορούμε να γράψουμε

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_q \\ x_{q+1} \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \quad \text{όπου } \mathbf{y} = \mathbf{x}_1 \text{ και}$$

$\begin{matrix} q \times 1 \\ (p-q) \times 1 \end{matrix}$

$$\text{Όπου } \boldsymbol{\mu}_x = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_q \\ \mu_{q+1} \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_x = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$\begin{matrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{matrix}$

Δηλαδή δημιουργούμε τους μερικούς πίνακες παίρνοντας συγκεκριμένα κομμάτια των πινάκων. Παρατηρήστε πως έτσι όπως έχουμε διαμερίσει τον πίνακα συνδιακύμανσης ο υποπίνακας $\boldsymbol{\Sigma}_{11}$ περιέχει τις διακυμάνσεις και τις συνδιακυμάνσεις όλων των μεταβλητών που περιέχονται στο διάνυσμα $\mathbf{x}_1$ και αντίστοιχη ερμηνεία μπορεί να δοθεί στον πίνακα $\boldsymbol{\Sigma}_{22}$. Μόνο αυτοί οι δύο πίνακες από τη διαμέριση του $\boldsymbol{\Sigma}$ είναι πίνακες διακυμάνσεις καθώς οι πίνακες $\boldsymbol{\Sigma}_{12}$ και $\boldsymbol{\Sigma}_{21}$ δεν είναι συμμετρικοί και απλά περιέχουν τις συνδιακυμάνσεις των μεταβλητών του ενός διανύσματος με τις μεταβλητές του άλλου διανύσματος. Τότε βρίσκουμε πως

$$\mathbf{A}\boldsymbol{\mu}_x = \begin{bmatrix} \underbrace{1 \ 0 \ \dots \ 0}_{q \times q} & \underbrace{0 \ \dots \ 0}_{q \times (p-q)} \\ \underbrace{0 \ 1 \ \dots \ 0}_{\vdots} & \underbrace{0 \ \dots \ 0}_{\vdots} \\ \underbrace{\dots \ \dots \ \dots \ \dots}_{\vdots} & \underbrace{0 \ \dots \ 0}_{\vdots} \\ \underbrace{0 \ \dots \ \dots \ 1}_{(p-q) \times q} & \underbrace{0 \ \dots \ 0}_{(p-q) \times (p-q)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix}$$

και

$$\mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}' = \begin{bmatrix} \mathbf{I}_{q \times q} & \mathbf{0}_{q \times (p-q)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{q \times q} \\ \mathbf{0}_{(p-q) \times q} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{q \times q} \\ \mathbf{0}_{(p-q) \times (p-q)} \end{bmatrix} = \boldsymbol{\Sigma}_{11}$$

Επομένως

$$\mathbf{y} \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}).$$

Δηλαδή κάθε περιθώρια (είτε απλή περιθώρια, δηλαδή για καθεμία μεταβλητή είτε από κοινού περιθώρια, δηλαδή για ζεύγη, τριάδες κλπ μεταβλητών) είναι πολυμεταβλητή κανονική κατανομή. Επομένως και κάθε απλή περιθώρια κατανομή είναι μια απλή κανονική κατανομή.

Δεσμευμένες κατανομές

Θα δούμε τώρα πως και όλες οι δεσμευμένες κατανομές είναι και αυτές πολυμεταβλητές κανονικές κατανομές.

Ας υποθέσουμε τη διαμέριση των πινάκων και των διανυσμάτων ως εξής:

$$\mathbf{x}_{p \times 1} : \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \text{ όπου } q+m=p, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ και } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}'$$

όπου $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{22}$ είναι πίνακες διακύμανσης-συνδιακύμανσης αλλά οι πίνακες $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$ δεν είναι πίνακες διακύμανσης-συνδιακύμανσης. Επίσης αν όλα τα στοιχεία του πίνακα $\boldsymbol{\Sigma}_{12}$ είναι 0, δηλαδή αν $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ τότε τα διανύσματα $\mathbf{x}_1$ και $\mathbf{x}_2$ είναι ανεξάρτητα.

Για να γίνει πιο κατανοητή η διαμέριση ας δούμε ένα παράδειγμα.

Παράδειγμα 3.4: Έστω το διάνυσμα $\mathbf{x} = \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix}$ το οποίο διαμερίζουμε σε δύο μικρότερα

διανύσματα $\mathbf{x}_1 = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$ και $\mathbf{x}_2 = \begin{bmatrix} Z_3 \\ Z_4 \end{bmatrix}$. Επομένως διαμερίζουμε και το διάνυσμα των μέσων και τον πίνακα διακύμανσης ως

$$\boldsymbol{\mu}_1 = \begin{bmatrix} E(Z_1) \\ E(Z_2) \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} E(Z_3) \\ E(Z_4) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) \\ \text{Cov}(Z_1, Z_2) & \text{Var}(Z_2) \end{bmatrix}, \boldsymbol{\Sigma}_{22} = \begin{bmatrix} \text{Var}(Z_3) & \text{Cov}(Z_3, Z_4) \\ \text{Cov}(Z_3, Z_4) & \text{Var}(Z_4) \end{bmatrix}$$

$$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} \text{Cov}(Z_1, Z_3) & \text{Cov}(Z_1, Z_4) \\ \text{Cov}(Z_2, Z_3) & \text{Cov}(Z_2, Z_4) \end{bmatrix}$$

Στην περίπτωση που οι μεταβλητές δεν είναι στη σειρά πρέπει να προσέξουμε πως θα κατασκευάσουμε τους υποπίνακες που θέλουμε ώστε να είμαστε σίγουροι ότι αντιστοιχούν στην διαμέριση που ζητάμε

Μπορούμε τώρα να δούμε το θεώρημα που αφορά τις δεσμευμένες κατανομές

Θεώρημα

Έστω ένα τυχαίο διάνυσμα $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ το οποίο μπορούμε να διαμερίσουμε όπως δείξαμε πριν. Τότε για τη δεσμευμένη κατανομή θα έχουμε πως

$$\mathbf{x}_1 | \mathbf{x}_2 \sim N_q(\boldsymbol{\mu}', \boldsymbol{\Sigma}'),$$

$$\text{όπου } \boldsymbol{\mu}' = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2), \quad \boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}$$

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

Αυτό που είναι πολύ ενδιαφέρον στο παραπάνω αποτέλεσμα είναι πως ο δεσμευμένος πίνακας διακύμανσης-συνδιακύμανσης δεν εξαρτάται από τις τιμές των μεταβλητών ως προς τις οποίες έχουμε δεσμεύσει. Αυτό σημαίνει πως οποιαδήποτε και αν είναι η τιμή τους η δεσμευμένη διακύμανση είναι η ίδια.

Παράδειγμα 3.5: Έστω πως έχουμε μόλις 2 τμ X και Y και πως η από κοινού κατανομή τους είναι διμεταβλητή κανονική με διάνυσμα μέσων $\boldsymbol{\mu}' = [\mu_x \quad \mu_y]$ και πίνακα διακύμανσης

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}. \text{ Επομένως με χρήση του παραπάνω θεωρήματος προκύπτει πως η}$$

δεσμευμένη κατανομή της τμ Y δοθέντος ότι $X=x$ θα είναι κανονική κατανομή με

$$E(Y | X = x) = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2} (x - \mu_x) = \left[\mu_y - \frac{\sigma_{xy}}{\sigma_x^2} \mu_x \right] + \frac{\sigma_{xy}}{\sigma_x^2} x \text{ και}$$

$$Var(Y | X = x) = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}$$

Αν χρησιμοποιήσουμε το συντελεστή συσχέτισης $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, προκύπτει πως

$$\text{Var}(Y|X=x) = \sigma_y^2(1-\rho^2)$$

Από τους παραπάνω τύπους υπάρχουν μερικά πολύ χρήσιμα συμπεράσματα:

- Κατ' αρχάς είναι προφανές πως η δεσμευμένη διακύμανση δεν εξαρτάται από την τιμή x της τμ X που έχουμε ως δέσμευση, δηλαδή την πληροφορία. Γενικά αυτό δεν συμβαίνει συχνά, δηλαδή η δεσμευμένη διακύμανση να μην εξαρτάται από τη δέσμευση.
- Επίσης μπορεί κανείς να δει πως η δεσμευμένη αναμενόμενη τιμή είναι γραμμική ως προς την τιμή της τμ X , δηλαδή ως προς τη δέσμευση. Στη στατιστική η δεσμευμένη αναμενόμενη τιμή ονομάζεται παλινδρόμηση (regression). Ο όρος παλινδρόμηση είναι γνωστός από το γραμμικό μοντέλο αλλά αυτό είναι λογικό αφού το γραμμικό μοντέλο στηρίζεται στις υποθέσεις που μόλις τώρα είδαμε, δηλαδή πως η αναμενόμενη τιμή του Y δοθέντος του $X=x$ είναι γραμμική ως προς το x και πως η διακύμανση είναι σταθερή και ανεξάρτητη από το x . Επίσης η κατανομή του Y είναι η κανονική. Οι υποθέσεις της γραμμικής παλινδρόμησης είναι αυτές που βλέπουμε για τη δεσμευμένη κατανομή (Παρατήρηση: στο γραμμικό μοντέλο το X δεν είναι τμ αλλά γνωστό εκ των προτέρων, αυτό δηλαδή που ουσιαστικά δηλώνει μια δεσμευμένη κατανομή)).
- Οι ομοιότητες δεν σταματούν εδώ καθώς αν κοιτάξει κανείς τη δεσμευμένη αναμενόμενη τιμή θα αναγνωρίζει πως τα α και β της γραμμικής παλινδρόμησης εκτιμούνται με τα δειγματικά αντίστοιχα των τύπων που βλέπετε στον τύπο της. Ανατρέχοντας δηλαδή σε οποιοδήποτε βιβλίο γραμμικής παλινδρόμησης θα δει κανείς πως Οι εκτιμήτριες ελαχίστων τετραγώνων για τα α και β δίνονται από τις σχέσεις
$$\hat{\alpha} = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x}, \quad \hat{\beta} = \frac{s_{xy}}{s_x^2}$$
 οι οποίες είναι τα δειγματικά αντίστοιχα αυτών που βλέπουμε στη δεσμευμένη αναμενόμενη τιμή.

- Επίσης πολύ ενδιαφέρον είναι να δει κανείς πως η δεσμευμένη διακύμανση του Y είναι σε κάθε περίπτωση μικρότερη ή ίση της διακύμανσης του Y , δηλαδή $\text{Var}(Y|X=x) \leq \text{Var}(Y)$ με την ισότητα να ισχύει μόνο αν $\rho=0$, δηλαδή οι δύο μεταβλητές είναι ασυσχέτιστες. Αυτό σημαίνει πως αν υπάρχει έστω και μικρή συσχέτιση ανάμεσα σε δύο μεταβλητές μπορούμε να μειώσουμε τη διακύμανση της μιας έχοντας κάποια πληροφορία για την άλλη. Ουσιαστικά αυτό είναι το βασικό συστατικό κάθε πολυμεταβλητής ανάλυσης, να χρησιμοποιήσουμε δηλαδή πληροφορίες από άλλες μεταβλητές ώστε να μειώσουμε τη διακύμανση (και άρα να βελτιώσουμε τη δυνατότητα μας να ελέγξουμε την αβεβαιότητα) της μεταβλητής που μας ενδιαφέρει.
- Το πρόσημο της συσχέτισης ανάμεσα σε δύο μεταβλητές δεν παίζει ρόλο.

Παράδειγμα 3.6: Έστω το τυχαίο διάνυσμα $\mathbf{x}' = (X_1, X_2, X_3, X_4)$ και έστω πως $\mathbf{x} \sim N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ όπου

$$\boldsymbol{\mu}' = [0 \quad 1 \quad 0 \quad 2] \text{ και } \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 & 0.5 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}. \text{ Να βρεθεί η δεσμευμένη κατανομή των}$$

$$X_1, X_2 \mid X_3 = x, X_4 = y.$$

Γνωρίζουμε πως η δεσμευμένη κατανομή θα είναι πολυμεταβλητή κανονική. Για να βρούμε τις παραμέτρους χρειάζεται να διαμερίσουμε τόσο το διάνυσμα των μέσων όσο και τον πίνακα διακύμανσης. Έτσι θα έχουμε πως

$$\boldsymbol{\mu}_1' = [0 \quad 1], \quad \boldsymbol{\mu}_2' = [0 \quad 2]$$

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{22} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}' = \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix}. \text{ Παρατηρείστε πως καθώς δεν}$$

είναι όλα τα στοιχεία του πίνακα $\boldsymbol{\Sigma}_{12}$ ίσα με 0, και άρα τα δύο διανύσματα (X_1, X_2) και (X_3, X_4) δεν είναι ανεξάρτητα.

Χρησιμοποιώντας τους τύπους βρίσκουμε πως η δεσμευμένη αναμενόμενη τιμή θα είναι

$$(X_1, X_2 \mid X_3 = x, X_4 = y) \sim N_2(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$$

όπου

$$\begin{aligned} \boldsymbol{\mu}' &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right) = \\ &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y-2 \end{bmatrix} = \\ &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} y/2-1 \\ 0 \end{bmatrix} = \begin{bmatrix} y/2-1 \\ 1 \end{bmatrix} \end{aligned}$$

και

$$\begin{aligned} \boldsymbol{\Sigma}' &= \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 0.5 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0.25 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.75 & 0 \\ 0 & 2 \end{bmatrix} \end{aligned}$$

Συνεπώς αν $x=1$ και $y=1$ προκύπτει πως

$$x_1 | x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.75 & 0 \\ 0 & 2 \end{bmatrix} \right)$$

ενώ στη γενική μορφή η δεσμευμένη κατανομή θα είναι η

$$x_1 | x_2 = \begin{bmatrix} x \\ y \end{bmatrix} \sim N_2 \left(\begin{bmatrix} y/2 - 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.75 & 0 \\ 0 & 2 \end{bmatrix} \right)$$

Αν δούμε τα αποτελέσματα παρατηρούμε πως τα X_1 και X_2 είναι και πάλι ανεξάρτητα αλλά τώρα η διακύμανση της τμ X_1 έχει μειωθεί.

$$L = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \prod_{i=1}^n f(\mathbf{x}_i) = \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu) \right)$$

και άρα

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

Λογαριθμίζοντας βρούμε πως

$$\begin{aligned} l = \ln L &= -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \sum_{i=1}^n [(\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu)] = \\ &= \text{constant} - \frac{n}{2} \ln |\Sigma| - \frac{n}{2} \text{tr}(\Sigma^{-1} S) - \frac{n}{2} (\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) \end{aligned}$$

γιατί

$$\begin{aligned} \sum_{i=1}^n [(\mathbf{x}_i - \mu)' \Sigma^{-1} (\mathbf{x}_i - \mu)] &= \sum_{i=1}^n [((\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu))' \Sigma^{-1} ((\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu))] \\ &= \sum_{i=1}^n [(\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) + 2(\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})] = \\ &= \sum_{i=1}^n [(\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})] + n(\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) \end{aligned}$$

επειδή

$$\sum_{i=1}^n [2(\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})] = 2(\bar{\mathbf{x}} - \mu)' \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}}) = 0$$

3.6 Εκτίμηση παραμέτρων

Οι άγνωστες παράμετροι της πολυμεταβλητής κανονικής κατανομής είναι το διάνυσμα των μέσων μ (διάνυσμα $p \times 1$) και ο πίνακας Σ διαστάσεων $p \times p$. Δεδομένου πως ο πίνακας Σ είναι συμμετρικός, δεν χρειάζεται να εκτιμήσει κανείς όλα του τα στοιχεία. Στην πραγματικότητα χρειάζεται να εκτιμήσει $p(p+1)/2$ παραμέτρους (διακυμάνσεις και συνδιακυμάνσεις) 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.

Μέθοδος των ροπών

Οι εκτιμήσεις με την μέθοδο των ροπών είναι:

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

και προσέχουμε εύκολα αν αναλυριστούμε την ερμηνεία των παραμέτρων της πολυμεταβλητής κανονικής κατανομής, (δευγματικό διάνυσμα μέσων και πίνακας διακύμανσης-συνδιακύμανσης).

Μέθοδος μέτρησης πιθανοφάνειας

Τίτω πως έχουμε ένα δείγμα μεγέθους n από πολυμεταβλητή κανονική κατανομή, δηλαδή υποθέτουμε πως $X_i \sim N_p(\mu, \Sigma)$, $i=1, \dots, n$.

Τότε η πιθανοφάνειας του δείγματος δίνεται ως

Όμως από τη γραμμική άλγεβρα ξέρουμε ότι αν τα x_i είναι διανύσματα $p \times 1$ και A πίνακας $p \times p$ τότε ισχύει πως

$$\sum_{i=1}^n x_i' A x_i = tr(AT)$$

όπου $T = \sum_{i=1}^n x_i x_i'$ είναι ένας πίνακας διαστάσεων $p \times p$ και άρα

$$\begin{aligned} \sum_{i=1}^n \left[(x_i - \bar{x})' \Sigma^{-1} (x_i - \bar{x}) \right] &= \sum_{i=1}^n tr \left[\Sigma^{-1} (x_i - \bar{x}) (x_i - \bar{x})' \right] = \\ tr \Sigma^{-1} \sum_{i=1}^n \left[(x_i - \bar{x}) (x_i - \bar{x})' \right] &= tr \Sigma^{-1} (nS) = n \, tr(\Sigma^{-1} S) \end{aligned}$$

Έτσι,

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

Επομένως για να μεγιστοποιήσει κανείς την πιθανοφάνεια ως προς το $\boldsymbol{\mu}$ ουσιαστικά ενδιαφέρεται για την ποσότητα

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

η οποία είναι οπωσδήποτε αρνητικός αριθμός ως το αρνητικό μιας τετραγωνικής μορφής, και επομένως για $\boldsymbol{\mu} = \bar{\mathbf{x}}$ η συνάρτηση μεγιστοποιείται.

Για να δείξουμε πως προκύπτει η αντίστοιχη εκτιμήτρια μέγιστης πιθανοφάνειας για το $\boldsymbol{\Sigma}$ δεν είναι τόσο απλό και για αυτό η απόδειξη δεν δίνεται σε αυτές τις σημειώσεις. Προκύπτει όμως πως $\hat{\boldsymbol{\Sigma}} = S$, δηλαδή οι εκτιμήτριες μέγιστης πιθανοφάνειας ταυτίζονται με τις εκτιμήτριες της μεθόδου των ροπών. Όπως θα δούμε αργότερα η εκτιμήτρια για τη διακύμανση είναι μεροληπτική.

3.7 Προσομοίωση δεδομένων από πολυμεταβλητή κανονική κατανομή

Όπως είδαμε προηγουμένως ισχύει πως αν $\mathbf{x} \sim N_p(\mathbf{0}, \mathbf{I})$ τότε ο μετασχηματισμός $\mathbf{y} = \boldsymbol{\Sigma}^{1/2} \mathbf{x} + \boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Επομένως μπορεί κανείς να προσομοιώσει από τη ζητούμενη πολυμεταβλητή κανονική κατανομή ξεκινώντας με ένα δείγμα ανεξάρτητων τυποποιημένων κανονικών μεταβλητών και χρησιμοποιώντας στη συνέχεια το μετασχηματισμό που περιγράψαμε. Ουσιαστικά το πρόβλημα είναι η εύρεση του πίνακα $\boldsymbol{\Sigma}^{1/2}$, ο οποίος μπορεί να βρεθεί με την ανάλυση Cholesky (Cholesky decomposition) ενός συμμετρικού πίνακα. Επομένως ο αλγόριθμος προσομοίωσης είναι ο εξής:

Βήμα 1: Προσομοιώσει p τυποποιημένες κανονικές τυχαίες μεταβλητές, $X_1, \dots, X_p$. Με αυτές όρισε το διάνυσμα $\mathbf{x}' = (X_1, \dots, X_p)$

Βήμα 2: Θέσει $\mathbf{y} = \boldsymbol{\Sigma}^{1/2} \mathbf{x} + \boldsymbol{\mu}$

Το y είναι ένα τυχαίο διάνυσμα που ακολουθεί τη ζητούμενη $N_p(\mu, \Sigma)$ κατανομή

Εναλλακτικά μπορεί κανείς να προσομοιώσει από μια πολυμεταβλητή κανονική κατανομή χρησιμοποιώντας την ιδιότητα πως όλες οι περιθώριες και όλες οι δεσμευμένες κατανομές είναι και αυτές πολυμεταβλητές κανονικές. Δηλαδή ένας αλγόριθμος είναι ο εξής:

- Παρήγαγε μια τμ X_1 από την περιθώρια κατανομή της X_1
- Παρήγαγε μια τμ από την κατανομή της $X_2 | X_1 = x_1$ (η οποία είναι μια κανονική κατανομή με κάποιες παραμέτρους που μπορεί κανείς σχετικά εύκολα να υπολογίσει)
- Παρήγαγε μια τμ από την κατανομή της $X_3 | X_1 = x_1, X_2 = x_2$ (η οποία και πάλι είναι μια κανονική κατανομή με κάποιες παραμέτρους που μπορεί κανείς σχετικά εύκολα να υπολογίσει)
-
- Παρήγαγε μια τμ από την κατανομή της $X_p | X_1 = x_1, X_2 = x_2, \dots, X_{p-1} = x_{p-1}$ (η οποία και πάλι είναι μια κανονική κατανομή με κάποιες παραμέτρους που μπορεί κανείς σχετικά εύκολα να υπολογίσει)

Στην πραγματικότητα αυτός ο αλγόριθμος δεν είναι ιδιαίτερα αποδοτικός καθώς χρειάζεται να υπολογίζεται σε κάθε βήμα η δεσμευμένη μέση τιμή και η αντίστοιχη διακύμανση. Παρόλα αυτά στη δεμεταβλητή περίπτωση τα πράγματα είναι πιο εύκολα και ο αλγόριθμος αρκετά γρήγορος. Έτσι για αυτή την περίπτωση έχουμε:

Έστω ότι θέλουμε να γεννήσουμε ένα τυχαίο διάνυσμα από μια δεμεταβλητή κανονική κατανομή με διάνυσμα μέσων τιμών $\mu' = (\mu_1, \mu_2)$ και πίνακα διακύμανσης $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$.

Η συσχέτιση επομένως είναι $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$. Έτσι για να προσομοιώσουμε από αυτή την κατανομή

Βήμα 1: Προσομοιώσε μια τμ $x_1 \sim N(\mu_1, \sigma_1^2)$

Βήμα 2: Προσομοιώσε μια τμ $x_2 \sim N\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)\right)$

4 ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Πριν ξεκινήσουμε την περιγραφή δειγματοληπτικών κατανομών, καθώς και πολυμεταβλητές κατανομές σχετικά με αυτές θα παρουσιάσουμε μερικές ενδιαφέρουσες μονοδιάστατες κατανομές με σκοπό να δούμε αμέσως μετά τη γενίκευσή τους στην πολυμεταβλητή περίπτωση.

4.1 Μη κεντρικές κατανομές

Οι μη κεντρικές κατανομές είναι γενικεύσεις γνωστών κατανομών και προκύπτουν με ανάλογο τρόπο. Συνήθως η διαφορά τους είναι πως περιέχουν άλλη μια παράμετρο που συνήθως ονομάζεται παράμετρος μη κεντρικότητας. Συνήθως οι περισσότερες κατανομές ξεκινούν από τυποποιημένες κανονικές τυχαιές μεταβλητές. Οι μη κεντρικές κατανομές προκύπτουν συνήθως όταν οι τυχαιές μεταβλητές δεν είναι τυποποιημένες δηλαδή δεν έχουν μέση τιμή 0.

Μη κεντρική χ^2 κατανομή

Από τη θεωρία κατανομών είναι γνωστό πως η κατανομή χ^2 προκύπτει ως το άθροισμα τετραγώνων ανεξάρτητων κανονικών κατανομών. Δηλαδή

$$X_i \sim N(0,1), \quad i=1,\dots,n$$

$$Z = \sum_{i=1}^n X_i^2 \sim \chi_n^2$$

ή ισοδύναμα

$$X_i \sim N(0,\sigma_i^2), \quad i=1,\dots,n$$

$$Z = \sum_{i=1}^n \frac{X_i^2}{\sigma_i^2} \sim \chi_n^2$$

Στην περίπτωση που οι κανονικές τιμ. δεν ακολουθούν κανονική κατανομή με μέση τιμή 0, τότε οδηγούμαστε στην μη κεντρική χ^2 κατανομή, και συγκεκριμένα

$$X_i \sim N(\mu_i, \sigma_i^2), \quad i = 1, \dots, n$$

$$Z = \sum_{i=1}^n \frac{X_i^2}{\sigma_i^2} \sim \chi_n^2(\delta)$$

όπου $\chi_n^2(\delta)$ είναι η μη κεντρική χ^2 κατανομή με παράμετρο μη κεντρικότητας $\delta = \sum_{i=1}^n \mu_i^2$.

Μπορεί εύκολα κανείς να δει πως αν $\mu_i = 0$ τότε και $\delta = 0$ και επομένως οδηγούμαστε στη γνωστή χ^2 κατανομή. Είναι ενδιαφέρον πως αν $Z \sim \chi_n^2(\delta)$ τότε $E(Z) = n + \delta$, $Var(Z) = 2n + 4\delta$.

Μη κεντρική F κατανομή

Η κατανομή F προκύπτει ως ο λόγος δύο χ^2 τυχαίων μεταβλητών. Δηλαδή

$$X \sim \chi_n^2, \quad Y \sim \chi_m^2$$

$$Z = \frac{m}{n} \frac{X}{Y} \sim F(n, m)$$

Αν τώρα η τυχαία μεταβλητή του αριθμητή δεν είναι κεντρική χ^2 αλλά μη κεντρική προκύπτει η μη κεντρική κατανομή F, δηλαδή

$$X \sim \chi_n^2(\delta), \quad Y \sim \chi_m^2$$

$$Z = \frac{m}{n} \frac{X}{Y} \sim F(n, m; \delta)$$

όπου και πάλι δ είναι η παράμετρος μη κεντρικότητας

Μη κεντρική t κατανομή

Η κατανομή t-student (ή απλά κατανομή t) προκύπτει ως ο λόγος μιας τυποποιημένης κανονικής μεταβλητής και της τετραγωνικής ρίζας μιας χ^2 τυχαίας μεταβλητής. Δηλαδή

$$X \sim N(0, 1), \quad Y \sim \chi_n^2$$

$$Z = \frac{X}{\sqrt{Y/n}} = \frac{\sqrt{n}X}{\sqrt{Y}} \sim t_n$$

Η μη κεντρική κατανομή t προκύπτει αν η κανονική τυχαία μεταβλητή δεν έχει μέσο 0, δηλαδή προκύπτει ως

$$X \sim N(\mu, 1), Y \sim \chi_n^2$$

$$Z = \frac{X}{\sqrt{Y/n}} = \frac{\sqrt{n}X}{\sqrt{Y}} \sim t_n(\mu)$$

όπου το μ είναι η παράμετρος μη κεντρικότητας

Όλες αυτές οι μη κεντρικές κατανομές προκύπτουν ως οι κατανομές γνωστών μας ελεγχουσυναρτήσεων κάτω από την εναλλακτική υπόθεση. Θυμηθείτε πως συνήθως για να κάνουμε τους ελέγχους μας αρκεί η κατανομή της ελεγχουσυνάρτησης κάτω από τη μηδενική υπόθεση και για αυτό συνήθως δεν ασχολούμαστε με το τι γίνεται όταν δεν ισχύει η μηδενική υπόθεση.

4.2 Η κατανομή Wishart

Έστω τα ανεξάρτητα τυχαία διανύσματα $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ μεγέθους p όπου $\mathbf{x}_i \sim N_p(\mathbf{0}, \Sigma)$, $i=1, \dots, n$. Ο τυχαίος πίνακας $\mathbf{W}$ διαστάσεων $p \times p$ που προκύπτει ως

$$\mathbf{W} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^n \mathbf{W}_i \text{ λέμε πως ακολουθεί την κατανομή Wishart με παραμέτρους } n \text{ και } \Sigma.$$

Θα συμβολίζουμε την κατανομή Wishart με παραμέτρους n και Σ ως $W_p(n, \Sigma)$. Μερικά σημαντικά σημεία για την κατανομή αυτή είναι τα εξής

- Αν $\mathbf{X} \sim \text{Wishart}(m, \Sigma)$, η συνάρτηση πυκνότητας πιθανότητας του πίνακα $\mathbf{X}$ είναι η

$$f(\mathbf{X}) = \frac{|\mathbf{X}|^{(m-p-1)/2} \exp\left\{-\frac{1}{2} \text{tr}(\Sigma^{-1}\mathbf{X})\right\}}{\pi^{p(p-1)/4} 2^{mp/2} |\Sigma|^{m/2} \prod_{i=1}^p \Gamma\left(\frac{m-i+1}{2}\right)}.$$

- Για $p=1$, $|\Sigma|=1$ η κατανομή είναι η χ^2 με m βαθμούς ελευθερίας. Μπορεί κανείς να παρατηρήσει ότι και από τον τρόπο που προκύπτει η κατανομή Wishart είναι γενίκευση της χ^2 κατανομής
- Αν στο μοντέλο γέννησης της κατανομής αφήσουμε τα τυχαία διανύσματα να μην έχουν μέσα διανύσματα 0, δηλαδή αν $\mathbf{x}_i \sim N_p(\mu_i, \Sigma)$, $i=1, \dots, n$ τότε προκύπτει η μη κεντρική

κατανομή Wishart, (συμβολισμός $Wishart(n, \Sigma, \mu)$) όπου $\mu = \sum_{i=1}^n \mu_i$ είναι η παράμετρος μη κεντρικότητας.

- Ισχύει πως $E(W) = n\Sigma$ ενώ για τη διακύμανση ενός τυχαίου πίνακα η σχέση είναι ιδιαίτερα πολύπλοκη και δεν θα δοθεί.
- Ενδιαφέρον παρουσιάζει ο αντίστροφος ενός πίνακα W που ακολουθεί την κατανομή Wishart. Η κατανομή λοιπόν του αντιστρόφου ακολουθεί την αντίστροφη Wishart κατανομή (inverted Wishart distribution) και έχει αρκετή χρησιμότητα στην Μπευζιανή στατιστική.
- Η κατανομή Wishart ορίζεται μόνο για συμμετρικούς πίνακες. Επομένως και ο πίνακας Σ πρέπει να είναι συμμετρικός

Η μεγάλη χρησιμότητα της κατανομής Wishart στην πολυμεταβλητή στατιστική στηρίζεται στο γεγονός πως είναι η κατανομή δειγματοληψίας του δειγματικού πίνακα διακύμανσης. Συγκεκριμένα ισχύει πως

Θεώρημα: Έστω ένα τυχαίο δείγμα από ανεξάρτητα τυχαία διανύσματα $x_1, \dots, x_n$ (διαστάσεων $p \times 1$). Αν $x_i \sim N_p(\mu, \Sigma)$, τότε $nS = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})' \sim W_p(n-1, \Sigma)$.

Το παραπάνω θεώρημα είναι και η αιτία για να χρησιμοποιούμε τον αμερόληπτο δειγματικό πίνακα διακύμανσης. Παρατηρείστε πως η αναμενόμενη τιμή του δειγματικού πίνακα διακύμανσης δεν είναι αμερόληπτη και για αυτό εμφανίζεται το $(n-1)$ στον παρονομαστή ώστε η εκτιμήτρια να είναι αμερόληπτη.

Επίσης γνωρίζουμε από πριν πως για το διάνυσμα των μέσων ισχύει πως $\bar{X} \sim N_p(\mu, \frac{1}{n}\Sigma)$. Είναι ενδιαφέρον ότι ισχύει και το επόμενο θεώρημα:

Θεώρημα: Το δειγματικό διάνυσμα των μέσων ($\bar{X}$) είναι ανεξάρτητο από τον δειγματικό πίνακα διακύμανσης-συνδιακύμανσης (S).

Το παραπάνω θεώρημα γενικεύει την ιδιότητα που ξέρουμε από την μονομεταβλητή περίπτωση πως ο δειγματικός μέσος είναι ανεξάρτητος από τη δειγματική διακύμανση.

4.3 Ιδιότητες της κατανομής Wishart

- 1) Αν $\mathbf{M}_1 \sim W(m_1, \Sigma)$, $\mathbf{M}_2 \sim W(m_2, \Sigma)$ και οι δύο πίνακες είναι ανεξάρτητοι μεταξύ τους τότε $\mathbf{M}_1 + \mathbf{M}_2 \sim W(m_1 + m_2, \Sigma)$

Παρατηρείστε πως οι δύο πίνακες πρέπει να έχουν τον ίδιο πίνακα Σ για να ισχύει το αποτέλεσμα. Προφανώς το αποτέλεσμα γενικεύεται και για περισσότερους από 2 πίνακες

- 2) Αν $\mathbf{M} \sim W(m, \Sigma)$ και πίνακας $\mathbf{C}_{p \times p}$ τότε $\mathbf{CMC}' \sim W(m, \mathbf{C}\Sigma\mathbf{C}')$

Το αποτέλεσμα έχει το εξής ενδιαφέρον. Αν ξεκινήσουμε από ένα τυχαίο διάνυσμα το οποίο μετασχηματίσουμε πολλαπλασιάζοντας το με έναν κατάλληλο πίνακα τότε η διακύμανση του καινούριου διανύσματος θα δίνεται από μια έκφραση αντιστοιχη με αυτές που βλέπουμε στο δεξιό μέλος της παραπάνω ιδιότητας. Αυτό σημαίνει πως η διακύμανση του καινούριου διανύσματος θα ακολουθεί και αυτή κατανομή Wishart αν αυτό συνέβαινε για τη διακύμανση του αρχικού διανύσματος.

- 3) Αν $\mathbf{M} \sim W(m, \Sigma)$ και διάνυσμα $\mathbf{C}_{p \times 1}$ τότε $\frac{\mathbf{c}'\mathbf{M}\mathbf{c}}{\sigma^2} \sim \chi_m^2$ όπου $\sigma^2 = \mathbf{c}'\Sigma\mathbf{c}$

Η παραπάνω ιδιότητα μας επιτρέπει να δούμε πως κάθε στοιχείο του πίνακα $\mathbf{M}$ ακολουθεί μια χ^2 κατανομή. Αυτό προκύπτει εύκολα αν για παράδειγμα ορίσει κανείς το διάνυσμα $\mathbf{c}$ να έχει όλο μηδενικά και μόνο μια μονάδα. Για παράδειγμα, αν $\mathbf{c}' = [1, 0, 0, \dots, 0]$ τότε θα ισχύει για το στοιχείο M_{11} του πίνακα $\mathbf{M}$ πως

$$\frac{M_{11}}{\sigma_1^2} \sim \chi_m^2 \text{ όπου } \sigma_1^2 \text{ είναι το } (1,1) \text{ στοιχείο του πίνακα } \Sigma.$$

4.4 Η κατανομή T^2 του Hotelling

Θεώρημα: Έστω ένα τυχαίο διάνυσμα $\mathbf{x}$ και ένας τυχαίος πίνακας $\mathbf{M}$ που ακολουθούν $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ και $W_p(m, \boldsymbol{\Sigma})$ αντίστοιχα τότε η ποσότητα $m(\mathbf{x} - \boldsymbol{\mu})' \mathbf{M}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ ακολουθεί την κατανομή T^2 του Hotelling με παραμέτρους p και m (συμβολικά $T^2(p, m)$) $\square$

Το παραπάνω θεώρημα έχει μεγάλη σημασία αν θυμηθούμε πως το διάνυσμα των δειγματικών μέσων και ο δειγματικός πίνακας διακύμανσης ικανοποιούν τις συνθήκες του παραπάνω θεωρήματος. Συγκεκριμένα έχουμε πως

$$\bar{\mathbf{x}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n} \boldsymbol{\Sigma}\right) \text{ και } n\mathbf{S} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \sim W_p(n-1, \boldsymbol{\Sigma})$$

και συνδυάζοντας τα με το παραπάνω θεώρημα προκύπτει πως

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

ή ισοδύναμα αν χρησιμοποιήσουμε τον αμερόληπτο πίνακα διακύμανσης $\mathbf{S}_*$

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}_*^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$$

Η κατανομή T^2 συνδέεται με την κατανομή F και επομένως ποσοστιαία σημεία της μπορούν να υπολογιστούν πολύ εύκολα. Συγκεκριμένα για μια t_m που ακολουθεί την κατανομή $T^2(p, m)$ ισχύει πως

$$\frac{m-p+1}{mp} T^2(p, m) \sim F(p, m-p+1)$$

και άρα μπορεί κανείς να χρησιμοποιήσει τους πίνακες της F κατανομής.

Μερικές ενδιαφέρουσες παρατηρήσεις είναι οι εξής:

- Η συνάρτηση T^2 έχει την πολύ χρήσιμη ιδιότητα πως είναι αμετάβλητη (invariant) κάτω 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 $F_{1,n}$ κατανομή η κατανομή T^2 είναι μια γενίκευση της κατανομής t .

- Αν $p=1$ τότε η συνάρτηση T^2 δεν είναι τίποτα άλλο παρά η συνάρτηση $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$ που τόσο πολύ χρησιμοποιούμε στη μονομεταβλητή περίπτωση.

Το γεγονός πως η ποσότητα

$$\frac{n-p}{(n-1)p} n(\bar{x} - \mu)' S_*^{-1} (\bar{x} - \mu) \sim F(p, n-p)$$

μπορεί να χρησιμοποιηθεί για την κατασκευή διαστημάτων εμπιστοσύνης για τη μέση τιμή του πληθυσμού. Τα διαστήματα εμπιστοσύνης θα είναι υπερελλειψοειδή όπως και στην περίπτωση της γνωστής διακύμανσης που είδαμε προηγούμενα.

Γενικεύοντας η σχέση που το θεώρημα αποδεικνύει μπορεί να χρησιμοποιηθεί για διαστήματα εμπιστοσύνης σε πολλές περιπτώσεις κατά αναλογία των ελλειψοειδών που κατασκευάσαμε όταν ο πίνακας διακύμανσης ήταν γνωστός.

Με τα δεδομένα του θεωρήματος η ποσότητα $m(x - \mu_0)' M^{-1} (x - \mu_0)$ θα ακολουθεί μια μη κεντρική $T^2(p, m, \delta)$ κατανομή όπου δ είναι μια παράμετρος μη κεντρικότητας και δίνεται από τον τύπο

$$\delta = m(\mu - \mu_0)' M^{-1} (\mu - \mu_0).$$

Η μη κεντρική κατανομή T^2 συνδέεται με τη μη κεντρική κατανομή F ως εξής

$$\frac{m-p+1}{mp} T^2(p, m, \delta) \sim F(p, m-p+1, \delta)$$

4.5 Η κατανομή Λάμδα του Wilks

Αν $\mathbf{A}, \mathbf{B}$ δύο ανεξάρτητοι πίνακες, $\mathbf{A} \sim W_p(m, \Sigma)$ και $\mathbf{B} \sim W_p(n, \Sigma)$ ($m \geq p$), τότε η ποσότητα

$\Lambda = \frac{|\mathbf{A}|}{|\mathbf{A} + \mathbf{B}|}$ ακολουθεί την κατανομή Λάμδα του Wilks με παραμέτρους p, m και n (συμβολισμός $\Lambda \sim \Lambda(p, m, n)$).

Η κατανομή Λάμδα είναι μονοδιάστατη. Η συνάρτηση πυκνότητας πιθανότητας της είναι αρκετά πολύπλοκη. Εναλλακτικά προκύπτει ως το γινόμενο τυχαίων μεταβλητών από τη Βήτα κατανομή, δηλαδή

$$\Lambda(p, m, n) = \prod_{i=1}^n u_i$$

$$u_i \sim B((m+i-p)/2, p/2), \quad i=1, \dots, n$$

Παρατηρείστε πως η κατανομή Λ είναι ανεξάρτητη από τον κοινό πίνακα Σ .

Ποσοστιαία σημεία της κατανομής έχουν πινακοποιηθεί. Παρόλα αυτά σε κάποιες περιπτώσεις, για συγκεκριμένες τιμές των παραμέτρων, μπορούμε με μετασχηματισμούς να καταλήξουμε σε μια F κατανομή. Συγκεκριμένα ισχύουν τα ακόλουθα

Η κατανομή $\Lambda(p, m, n)$ είναι ισοδύναμη με την κατανομή $\Lambda(n, m+n-p, p)$

Για συγκεκριμένες τιμές των παραμέτρων ισχύουν τα ακόλουθα

$$\left(\frac{m-p+1}{p} \right) \frac{1-\Lambda(p, m, 1)}{\Lambda(p, m, 1)} \sim F_{p, m-p+1}$$

$$\left(\frac{m}{n} \right) \frac{1-\Lambda(1, m, n)}{\Lambda(1, m, n)} \sim F_{n, m}$$

$$\left(\frac{m-p+1}{p} \right) \frac{1-\sqrt{\Lambda(p, m, 2)}}{\sqrt{\Lambda(p, m, 2)}} \sim F_{2p, 2(m-p+1)}$$

$$\left(\frac{m-1}{n} \right) \frac{1-\sqrt{\Lambda(2, m, n)}}{\sqrt{\Lambda(2, m, n)}} \sim F_{2n, 2(m-1)}$$

Μπορεί να δει κανείς πως επειδή η κατανομή προκύπτει ως το γινόμενο βήτα τυχαίων μεταβλητών, ο λογάριθμος μια τυχαίας μεταβλητής από τη Λ κατανομή θα είναι το άθροισμα ανεξάρτητων τυχαίων μεταβλητών από βήτα κατανομές μετά από λογαριθμικό μετασχηματισμό των τυχαίων μεταβλητών. Αυτό διευκολύνει κάπως τον υπολογισμό της συνάρτησης πυκνότητας πιθανότητας σε περιπτώσεις όπου οι παραπάνω τύποι δεν μπορούν να χρησιμοποιηθούν. Μια προσέγγιση για το λογάριθμο μιας Λ τυχαίας μεταβλητής είναι η εξής

$$\left[\frac{p-n+1}{2} - m \right] \ln \Lambda(p, m, n) \sim \chi_{np}^2$$

Θα πρέπει να παρατηρήσουμε πως αυτό το αποτέλεσμα στηρίζεται στο γεγονός πως η κατανομή του Λ προκύπτει ως η κατανομή μιας ελεγχουσυνάρτησης που είναι λόγος πιθανοφανειών και άρα ασυμπτωτικά ακολουθεί τη χ^2 κατανομή.

Ενδιαφέρον είναι και το εξής. Αν $p=1$ δηλαδή έχουμε απλές τυχαίες μεταβλητές, τότε με βάση όσα είπαμε πριν αυτές θα ακολουθούν χ^2 κατανομή και επομένως ο λόγος που παίρνουμε θα είναι μια F κατανομή εκτός ίσως από μια σταθερά που θα είναι οι βαθμοί ελευθερίας (αφού θα έχουμε μια χ^2 στον αριθμητή και μια χ^2 στον παρονομαστή).

Τέλος πρέπει να σημειωθεί πως η συνάρτηση Λ μπορεί να γραφεί μέσω των ιδιοτιμών του πίνακα $\mathbf{A}^{-1}\mathbf{B}$. Αν λοιπόν συμβολίσουμε με λ_i τις ιδιοτιμές του πίνακα αυτού τότε

$$\Lambda = \prod_{i=1}^p \frac{1}{1 + \lambda_i}$$

Από αυτό τον τύπο διαπιστώνει κανείς εύκολα πως η τιμή του Λ ανήκει αναγκαστικά στο διάστημα $(0,1)$.

Αν $\mathbf{A}, \mathbf{B}$ δύο ανεξάρτητοι πίνακες, $\mathbf{A} \sim W_p(m, \Sigma)$ και $\mathbf{B} \sim W_p(n, \Sigma)$ ($m \geq p$), τότε η μεγαλύτερη ιδιοτιμή θ_i του πίνακα $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$ ακολουθεί την κατανομή Θ ήτα με παραμέτρους p, m και n (συμβολισμός $\theta_i \sim \Theta(p, m, n)$).

Η κατανομή είναι και πάλι ανεξάρτητη του πίνακα Σ . Η συνάρτηση πυκνότητας πιθανότητας είναι αρκετά πολύπλοκη. Και πάλι για συγκεκριμένες τιμές των παραμέτρων η κατανομή μπορεί να μετασχηματιστεί σε μια F κατανομή. Ισχύει ότι

$$\frac{m-p+1}{p} \frac{\Theta(p, m, 1)}{1 - \Theta(p, m, 1)} \sim F(p, m-p+1)$$

Είναι απλό να δει κανείς πως αν έχω $p=1$ τότε η Θ ήτα κατανομή ταυτίζεται με την Λ και επομένως και με την F .